TECHNICAL NOTE

Illumina Complete Long Read Prep with Enrichment, Human을 위한 커스텀 패널 디자인

인간 유전체 연구를 위한 유연성 높은 표적 롱 리드 인리치먼트

illumina®

소개

전장 유전체 시퀀싱(Whole-genome sequencing, WGS)을 수행할 때 쇼트 리드(short read)만으로는 매핑하기 어려운 영역이 일부 있을 수 있습니다. 롱 리드(Long read) 시퀀싱은 표준 쇼트 리드 WGS 데이터를 상호 보완하여 연구자가 이러한 분석이 어려운 영역을 분석할 수 있도록 해 줍니다. Illumina Complete Long Reads 기술은 일반적인 차세대 시퀀싱(next-generation sequencing, NGS) 워크플로우를 이용하며, Illumina 시퀀싱 시스템에서 하나의 분석 파이프라인을 통해 contiguous 롱 리드 시퀀스를 생성합니다(그림 1).[1-3] Illumina Complete Long Read Prep with Enrichment, Human은 특정 영역을 표적으로 하는 비용 대비 효과적인 롱 리드 시퀀싱 방법을 제시합니다.* Illumina Complete Long Read의 인리치먼트 chemistry는 매우 유연한 표적 및 프로브 디자인을 지원하므로 매핑이 어려운(difficult-to-map) 영역의 분석을 지원하거나, 페이징(phasing)을 통해 더 깊은 통찰력을 제공할 수 있습니다.

롱 리드를 위한 인리치먼트 프로브 패널 디자인

Illumina Complete Long Read Prep with Enrichment, Human은 일반적으로 짧은 절편(약 200~500 bp)을 포획할 때 사용하는 방법과는 다른 프로브 디자인 전략을 이용해 더 긴 절편(약 7~10 kb)을 포획합니다. Illumina DesignStudio™ 소프트웨어는 사용이 용이한 무료 인리치먼트 프로브 패널 디자인 도구입니다. DesignStudio 알고리즘은 GC 함량, 표적 특이도(specificity), 프로브 간격(즉, 표적 영역 내 프로브 개수)을 고려합니다. 120-mer 쇼트 리드 인리치먼트 패널의 표준 간격은 250~350 bp 프로브 윈도우입니다. 롱 리드 인리치먼트 패널 디자인의 경우, 다양한 길이에서 프로브 배치 간격을 테스트해 본 결과, 비용 대비 효과적이며 효율성이 높은 포획이 가능한 최적의 윈도우는 1 킬로베이스(kilobase, kb)인 것으로 확인되었습니다.

하이브리드화(Hybridization) 인리치먼트의 효과는 프로브 특이도에 따라 크게 달라집니다. 온타겟(On-target) 인리치먼트의 %는 표적 커버리지 뎁스(coverage depth)를 달성하는 데 요구되는 시퀀싱 양에 직접적인 영향을 줍니다. 반복적인 영역(Repetitive region)에서는 높은 특이도를 달성하는 것이 더 어려운데, 이 경우 더 큰 프로브 윈도우를 사용하면 성능이 좋지 않은 프로브를 배제하고, 반복 영역(윈도우 크기 최대 1 kb)을 피하여 더 적은 수의 프로브를 사용해 향상된 인리치먼트 효율성을 유지할 수 있습니다(그림 2). DesignStudio 알고리즘은 이러한 고려 사항들을 반영해 적절한 프로브 배치를 권장해 줄 수 있습니다. 타사 패널도 최상의 성능과 비용 효율성을 달성하기 위해 비슷한 가이드라인을 따라야 합니다. 표준 인리치먼트 프로브 배치 간격도 완벽하게 호환됩니다.

유연한 프로브 디자인 및 표적 전략

Illumina Complete Long Read Prep with Enrichment, Human은 연구자가 연구 목표에 따라 커스텀 프로브 패널을 선택하고 디자인할 수 있는 뛰어난 유연성을 제공합니다. 표적 영역은 개별적으로 단일 염기(base)부터 최대 수백 kb에 달할

* 분석 시 동일한 샘플로 얻은 ≥ 30× 표준 쇼트 리드 WGS 데이터 필요. 이전 런(run) 샘플로 생성한 FASTQ 파일 사용 가능.

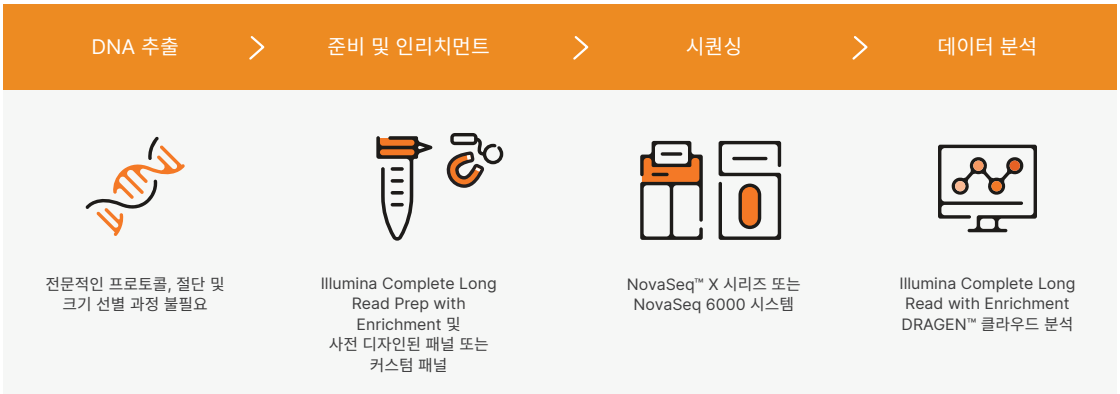

그림 1: 통합 워크플로우의 한 구성 요소 — 연구 규모에 따라 조정이 가능한 최적화된 라이브러리 준비 및 인리치먼트 프로토콜, 입증된 Illumina의 시퀀싱 chemistry 및 DRAGEN Secondary Analysis를 활용해 비용 대비 효과적으로 표적 롱 리드 WGS 데이터를 생성하는 워크플로우. 분석 시 동일한 샘플로 얻은 ≥ 30× 표준 쇼트 리드 WGS 데이터 필요. 이전 런 샘플로 생성한 FASTQ 파일 사용 가능.

수 있습니다. 패널의 크기는 커스텀 패널을 기준으로 작게는 2.5 Mb에서 최대 > 95 Mb의 수준입니다. 연구자는 표적 롱 리드를 사용해 쇼트 리드 데이터에서 매핑률(mappability)이 낮은 것으로 알려진 특정 영역에 대한 커버리지를 향상시킬 수 있습니다. 또는 롱 리드를 표적화하여 긴 다중 유전자(multigene) 영역까지 전체 유전자를 커버하면 변이를 페이징하고 하플로타입(haplotype)을 검출할 수 있습니다.

연구자는 DesignStudio 도구를 통해 다양한 사전 디자인된 패널을 사용해 볼 수 있습니다(표 1). 이러한 패널은 의학적 관련이 있는 분석이 어려운 유전자(challenging medically relevant gene, CMRG),[4] 약물유전학(pharmacogenetic, PGx) 검사 assay가 일반적으로 표적화하는 유전자,[5-7] 미국의학유전학회(American College of Medical Genetics and Genomics, ACMG)의 2차 소견(ACMG SF v3.1) 목록에 제시된 유전자[8] 또는 주조직 적합성 복합체(major histocompatibility complex, MHC) 영역 전체[9]를 표적화합니다. Illumina Human Comprehensive Panel은 단백질 코딩 유전자(protein-coding gene) 내 커버리지가 낮은 별개의 영역들을 주로 표적으로 하며, 사전 디자인된 패널 또는 즉시 배송이 가능한 사전 제작된 패널의 형태로 제공됩니다(Illumina, 카탈로그 번호: 20113836).[10,11] DesignStudio 소프트웨어는 BED 파일을 활용한 커스텀 패널의 디자인†이나 기존 사전 디자인된 패널의 편집을 지원합니다.

커스텀 프로브 패널에 권장되는 시퀀싱 뎁스

Illumina Complete Long Read Prep with Enrichment, Human은 매우 일관적이고 우수한 성능을 제공합니다. 테스트를 거친 사전 디자인된 패널의 경우, 1 Mb의 표적 패널 영역당 약 1.5 Gb의 시퀀스 데이터(약 5M 개의 페어드 엔드 리드(paired-end read))를 적용했을 때 최적의 성능을 달성할 수 있었습니다(그림3). 성능이 아직 알려지지 않은 새롭게 디자인된 패널의 경우, 우선 1 Mb의 표적 패널 영역당 3 Gb의 시퀀스 데이터(약 10M 개의 페어드 엔드 리드)로 시작하는 것을 권장하고 최적화를 통해 줄이는 것이 가능합니다.

분석이 어려운 영역의 정확도 높은 커버리지 및 페이징

Illumina Human Comprehensive Panel이나 CMRG 패널과 같이 커버리지가 낮은 특정 영역에 대한 검출력 향상에 초점을 맞춘 롱 리드 인리치먼트 프로브 패널은 분석이 어려운 표적 영역에 대한 변이 검출 정확도를 높입니다(그림 4). 또한 롱 리드 인리치먼트와 CMRG 패널을 함께 사용하면 단백질 코딩 영역에 걸쳐 커버리지의 완전성(completeness)을 높이고 변이 검출력을 향상시킬 수 있습니다(그림 5, 그림 6).

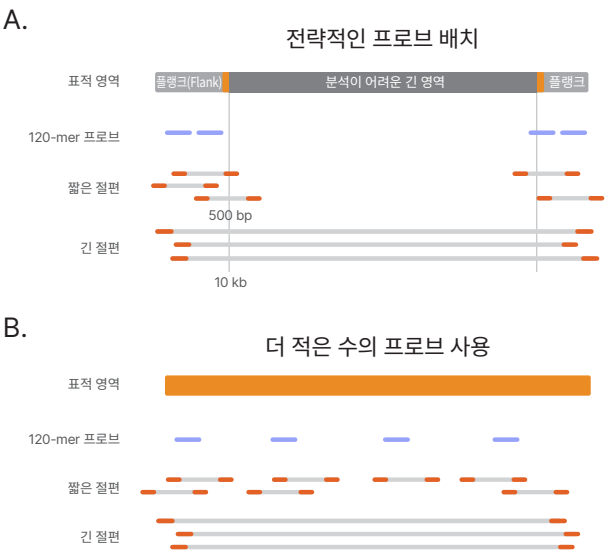

그림 2: 긴 절편의 하이브리드화로 향상된 인리치먼트 효율성 — 긴 절편의 하이브리드화는 짧은 절편의 포획에 비해 (A) 지나치게 높은 GC 함량, 낮은 복잡성(complexity), 반복 등 프로브 디자인에서 분석이 어려운 영역 바깥쪽에 프로브를 전략적으로 배치하고, (B) 각 표적 영역을 포획하는 데 필요한 프로브 수가 더 적다는 장점을 제공함. DesignStudio 알고리즘은 1 kb 섹션 내 표적 영역에서 프로브 배치를 위해 최적의 GC 함량을 지니고 특이도가 가장 높은 영역을 찾음.

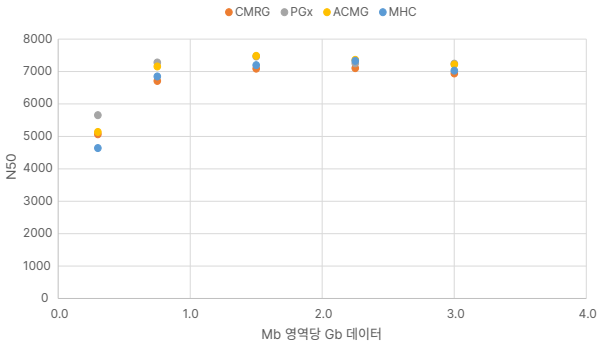

그림 3: 커스텀 프로브 패널의 시퀀싱 요구 사항 — 최장 길이의 N50에 요구되는 시퀀싱 데이터를 적정(titration)한 결과, Mb 표적 영역당 1.5 Gb(약 5M 개의 페어드 엔드 리드)로 Illumina Complete Long Read 데이터 생성에 필요한 표적 영역이 효율적으로 분석됨을 확인함.

† BED = browser extensible data(브라우저 확장형 데이터)

표 1: Illumina Complete Long Read Prep with Enrichment, Human에 맞게 사전 디자인된 인리치먼트 프로브 패널

패널[a]	CMRG 패널	PGx 패널	ACMG 패널	MHC 패널
표적 유전자	쇼트 리드만으로는 분석이 어려운 것으로 알려진 의학적 관련이 있는 391개의 유전자[5]	약물유전학 검사 assay에서 일반적으로 표적으로 하는 98개의 유전자[6~8]	ACMG 2차 소견 목록의 특정 유전자 78개(ACMG SF v3.1)[9]	GRCh38.p14 assembly의 전체 MHC 영역 (140개 이상의 유전자)[10]
표적 영역 크기[b]	22.5 Mb	8.1 Mb	7 Mb	4.9 Mb
샘플당 데이터 아웃풋[c]	약 67.5 Gb	약 24.3 Gb	약 21 Gb	약 14.7 Gb
프로브 수	약 22,500개	약 8,200개	약 6,900개	약 5,000개
N50[d]	6.1 kb	7.3 kb	7.3 kb	7.3 kb
페이즈 블록(Phase block) N50[d,e]	82.8 kb	94.4 kb	94.4 kb	357 kb
평균 표적 영역 크기[e]	58 kb	83 kb	88 kb	5,000 kb
균일성(Uniformity)[d,f]	97.9%	99.0%	99.5%	97.8%
패딩된 리드 인리치먼트(PRE)[d,f]	80.1%	79.3%	66.3%	67.5%
페이징된 이질접합(heterozygous) SNV의 %[d]	98.9%	98.9%	99.6%	98.6%

a. CMRG = challenging medically relevant gene(의학적 관련이 있는 분석이 어려운 유전자), PGx = pharmacogenomics(약물유전체학), ACMG = American College of Medical Genetics and Genomics(미국의학유전학회), MHC = major histocompatibility complex(주조직 적합성 복합체), SNV = single nucleotide variant(단일 염기서열 변이)
b. 표적 영역의 크기는 패딩된 프로브 위치 길이(padded probe location length)의 합으로, 겹치는 부분은 병합됨.
c. 2 × 150 bp 시퀀싱 런 및 Mb 표적 영역당 5M~10M 개의 페어드 엔드 리드(약 1.5~3 Gb의 데이터) 조건에서 약 30×의 최종 Illumina Complete Long Reads 커버리지 확보. 단, 커스텀 패널의 샘플당 요구되는 데이터는 권장 사항이며, 패널 성능에 따라 사용자가 할당되는 데이터를 최적화할 수 있음.
d. 50 ng의 HG002 유전체 DNA(Coriell, 카탈로그 번호: NA24385)를 사용해 생성한 데이터. 실제 성능은 DNA 사용량과 샘플의 품질에 따라 상이할 수 있음.
e. 페이즈 블록의 크기는 개별 contiguous 표적 영역의 크기로 제한됨.
f. % > 0.2 * 평균값으로 계산된 커버리지 균일성. 패딩된 리드 인리치먼트(Padded read enrichment, PRE) 값은 100 * (패딩된 표적 정렬된 리드의 수/총 정렬된 리드의 수)로 계산됨.

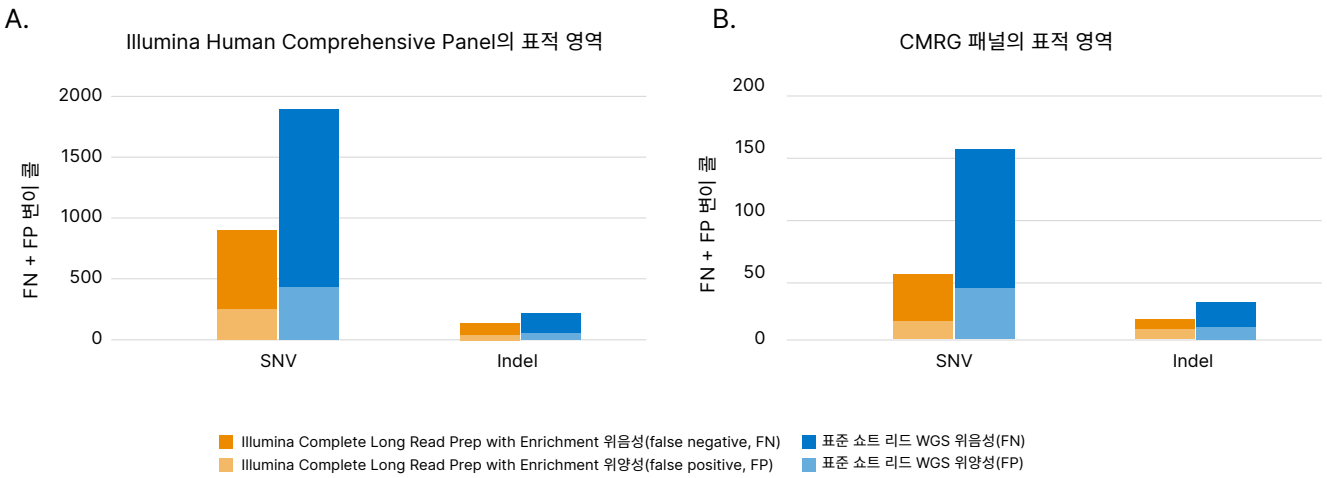

그림 4: 분석이 어려운 영역에서의 변이 검출 정확도를 향상시키는 표적 롱 리드 — Illumina Complete Long Read Prep with Enrichment(주황색)와 표준 쇼트 리드 WGS(파란색)를 사용해 SNV 및 삽입/결실(insertion/deletion, Indel) 검출 시 (A) Human Comprehensive Panel 또는 (B) CMRG 패널이 표적으로 하는 HG002 유전자 영역에서의 위음성(FN) + 위양성(FP) 결과

그림 5: 커버리지가 낮은 영역의 검출력을 향상시키는 표적 롱 리드 — Illumina Complete Long Read Prep, Human WGS(위), Illumina Complete Long Read Prep with Enrichment, Human 및 CMRG 패널(가운데), 그리고 표준 쇼트 리드 WGS(아래)를 사용한 *HBG1* 유전자의 롱 리드 시퀀싱 데이터를 Integrative Genomics Viewer(IGV) 플롯으로 나타냄.

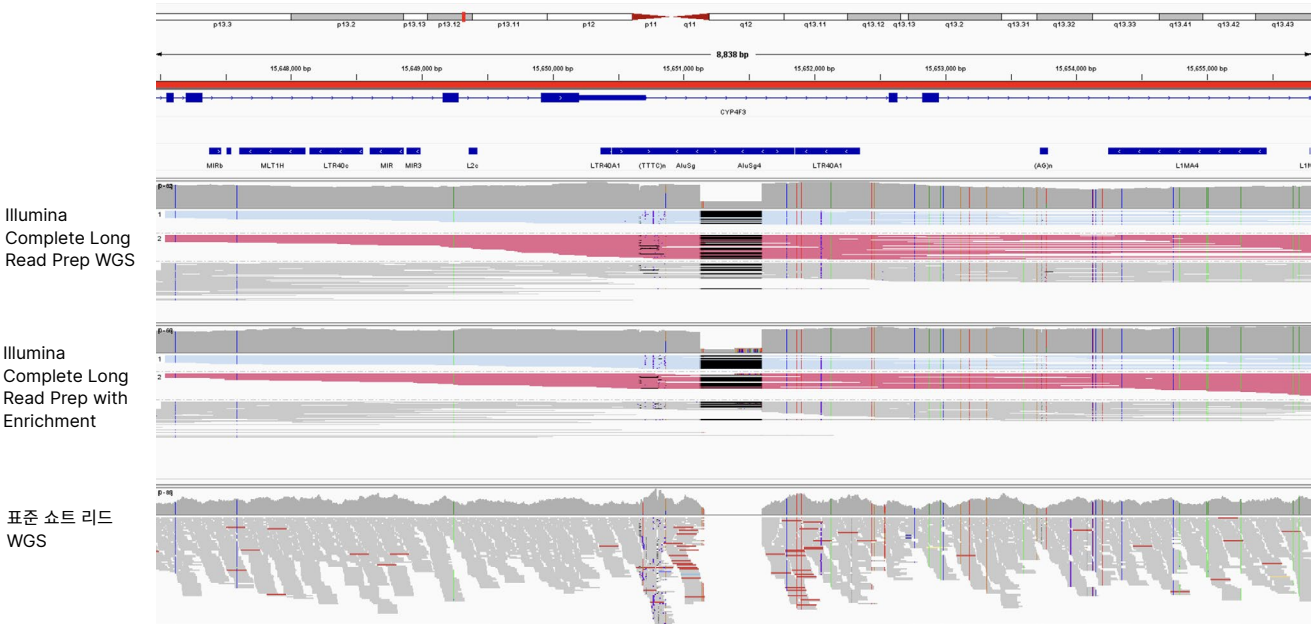

그림 6: 표적 롱 리드로 얻은 선명한 결실 경계 해상도 — Illumina Complete Long Read Prep, Human WGS(위), Illumina Complete Long Read Prep with Enrichment, Human 및 CMRG 패널(가운데), 그리고 표준 쇼트 리드 WGS(아래)를 사용한 *CYP4F3* 유전자의 롱 리드 시퀀싱 및 페이징 데이터를 IGV 플롯으로 나타낸 것으로, 대립유전자(allele) 1은 파란색, 대립유전자 2는 분홍색으로 표시되어 있음.

하플로타입의 분석을 위한 긴 페이즈 블록

각 패널의 페이즈 블록 N50‡은 표적 영역의 contiguous 길이와 관련이 있습니다(그림 7, 표 1). CMRG 패널, PGx 패널 및 ACMG 패널은 관심 있는 전장 유전자를 표적화하도록 디자인되었으며 평균 약 80~95 kb의 페이즈 블록 N50를 생성하므로 전체 이질접합 대립유전자의 페이징이 가능합니다(그림 8). MHC 패널은 약 4.9 Mb의 단일 contiguous 영역을 표적화하며 평균 350 kb가 넘는 페이즈 블록 N50를 생성하므로 전장 유전체 영역을 커버하는 해상도를 제공합니다(그림 9).

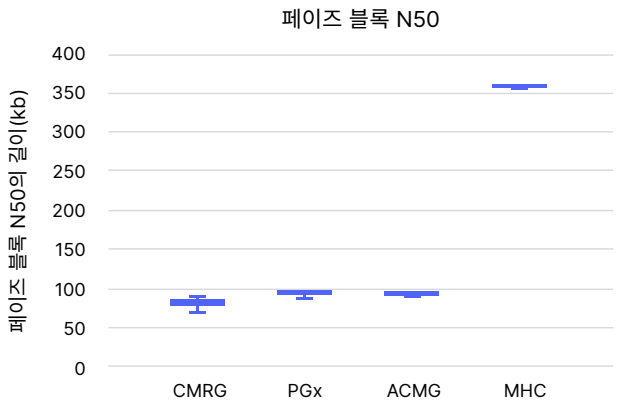

그림 7: Contiguous 표적 영역의 길이에 따른 페이즈 블록 N50의 차이 — CMRG 패널, PGx 패널 및 ACMG 패널은 관심 있는 전장 유전자를 표적화하여 평균 약 80~95 kb의 페이즈 블록 N50를 생성함. MHC 패널은 전체 MHC 유전자 영역을 표적화하여 평균 350 kb가 넘는 페이즈 블록 N50를 생성함. 평균 표적 영역의 크기는 CMRG 패널의 경우 58 kb, PGx 패널의 경우 83 kb, ACMG 패널의 경우 88 kb, MHC 패널의 경우 5,000 kb임.

‡ 페이즈 블록 N50은 표적 영역의 전체 어셈블리(assembly) 길이의 50%일 때 가장 짧은 contiguous 시퀀스 블록을 반영함.

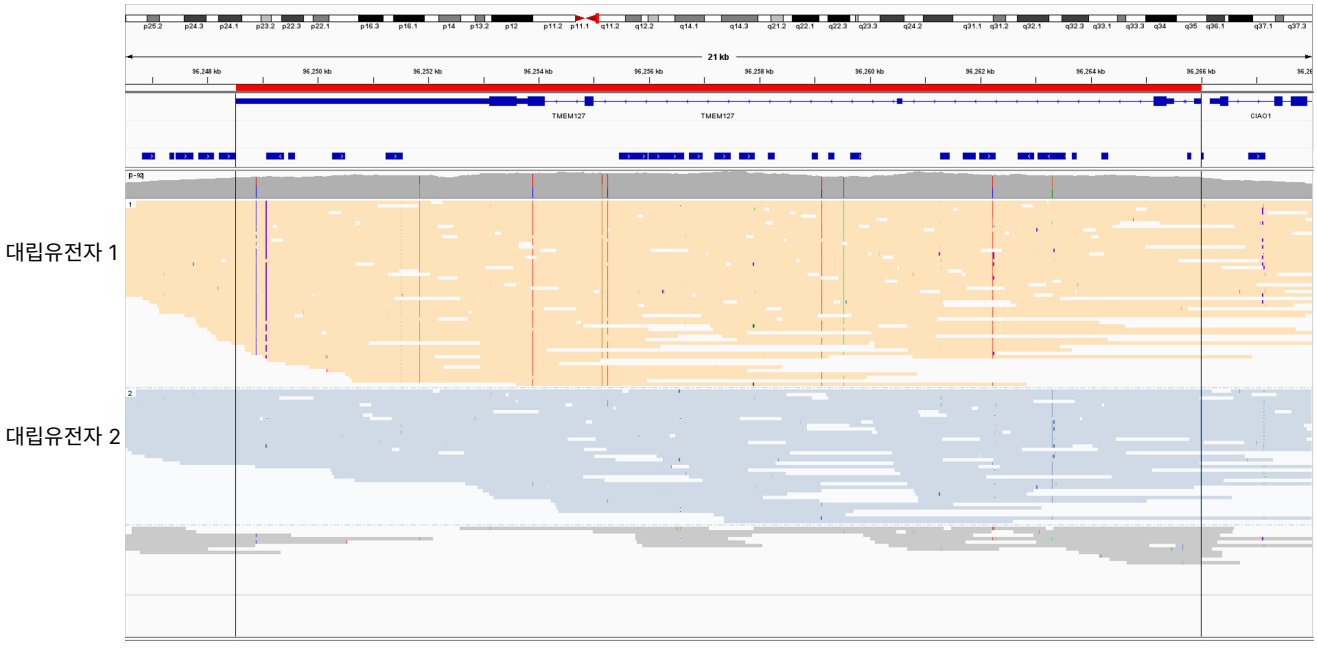

그림 8: 이질접합 SNV 영역의 페이징을 지원하는 표적 롱 리드 — Illumina Complete Long Read Prep with Enrichment, Human 및 ACMG 패널을 사용한 21 kb *TMEM17* 유전자의 롱 리드 시퀀싱 데이터를 IGV 플롯으로 보면, 하나의 페이즈 블록에서의 전체 페이징을 확인할 수 있음. 대립유전자 1은 노란색, 대립유전자 2는 파란색으로 표시되어 있음.

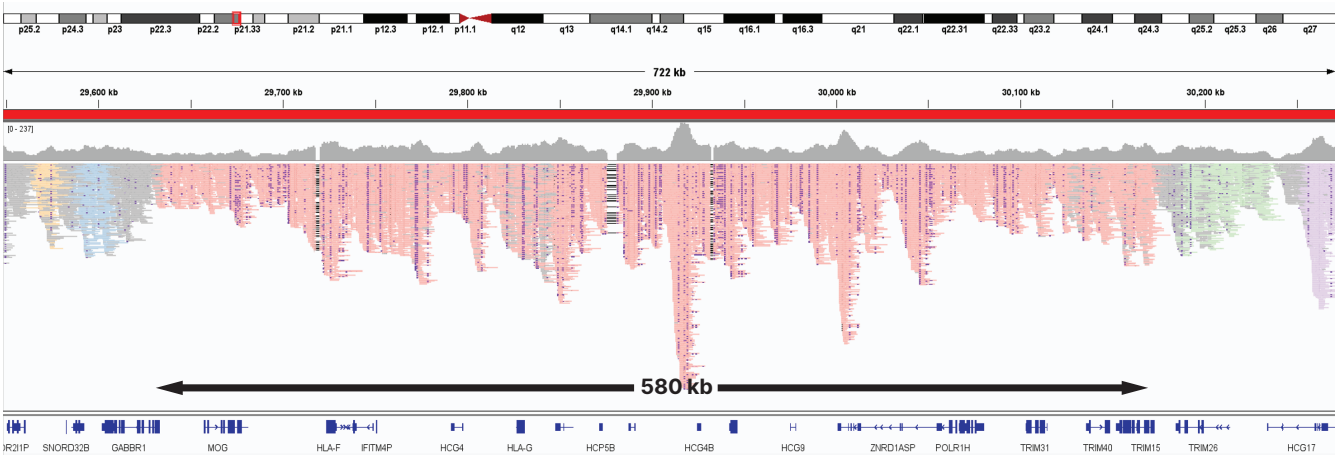

그림 9: 다형성(Polymorphic) 유전자의 하플로타입 분석을 지원하는 표적 롱 리드 — Illumina Complete Long Read Prep with Enrichment, Human을 사용한 롱 리드 시퀀싱의 IGV 플롯을 보면, MHC 유전좌위의 722 kb 영역에 걸쳐 페이징을 진행했을 때 580 kb 영역(분홍색)이 하나의 페이즈 블록으로 압축됨.

요약

Illumina Complete Long Read Prep with Enrichment, Human은 입증된 Illumina의 쇼트 리드 WGS 기술을 상호 보완하고 롱 리드 시퀀싱으로 가장 유용한 정보를 얻을 수 있는 영역에 초점을 맞춥니다. 연구자는 유연하게 사전 디자인된 패널을 선택하거나, DesignStudio 알고리즘을 사용해 롱 리드 표적 인리치먼트에 맞는 커스텀 패널을 디자인할 수 있습니다. 인리치먼트 프로브 패널을 표적화하면 전체 유전자를 페이징함으로써 커버리지를 향상시키거나 더 깊은 통찰력을 얻을 수 있으며, 완전한 워크플로우 솔루션을 갖춘 비용 대비 효과적이고 정확한 WGS를 수행할 수 있습니다.

상세 정보

Illumina Complete Long Read Prep with Enrichment, Human

DesignStudio Assay Design Tool

롱 리드 시퀀싱 기술

참고 문헌

1. Illumina. Illumina Complete Long Read Prep, Human data sheet. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-data-sheet-m-gl-01420/illumina-long-read-prep-data-sheet-m-gl-01420.pdf. Published 2022. Accessed September 22, 2023.
2. Illumina. Comprehensive whole-genome sequencing with Illumina Complete Long Read Prep, Human technical note. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-tech-note-m-gl-01421/ilmn-long-read-hu-tech-note-m-gl-01421.pdf. Published 2022. Accessed September 22, 2023.
3. Roessler K. Illumina Complete Long Reads software analysis workflow for human WGS. https://www.illumina.com/science/genomics-research/articles/complete-long-read-software-analysis.html. Published 2023. Accessed September 22, 2023.
4. Wagner J, Olson ND, Harris L, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol*. 2022;40(5):672-680. doi:10.1038/s41587-021-01158-1
5. PharmGKB. VIPs: Very Important Pharmacogenes. pharmgkb.org/vips. Accessed September 22, 2023.
6. National Library of Medicine. GTR: Genetic Testing Registry. Precision HealthPGx Panel (25 Genes). ncbi.nlm.nih.gov/gtr/tests/593428/. Updated November 29, 2022. Accessed September 22, 2023.

7. Pratt VM, Everts RE, Aggarwal P, et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. *J Mol Diagn*. 2016;18(1):109-123. doi:10.1016/j.jmoldx.2015.08.005
8. Miller DT, Lee K, Abul-Husn NS, et al. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2022;24(7):1407-1414. doi:10.1016/j.gim.2022.04.006
9. Kulski JK, Suzuki S, Shiina T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum Genome Var*. 2022;9(1):49. doi:10.1038/s41439-022-00226-5
10. Bekritsky MA, Colombo C, Eberle MA. Identifying genomic regions with high quality single nucleotide variant calling. Published 2021. Accessed August 30, 2023.
11. Illumina. Illumina Human Comprehensive Panel data sheet. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-enrich-hu-comp-panel-data-sheet-m-gl-02191/long-read-hu-comp-panel-data-sheet-m-gl-02191.pdf. Published 2024. Accessed January 26, 2024.

illumina®

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

M-GL-02189 v1.0 KOR