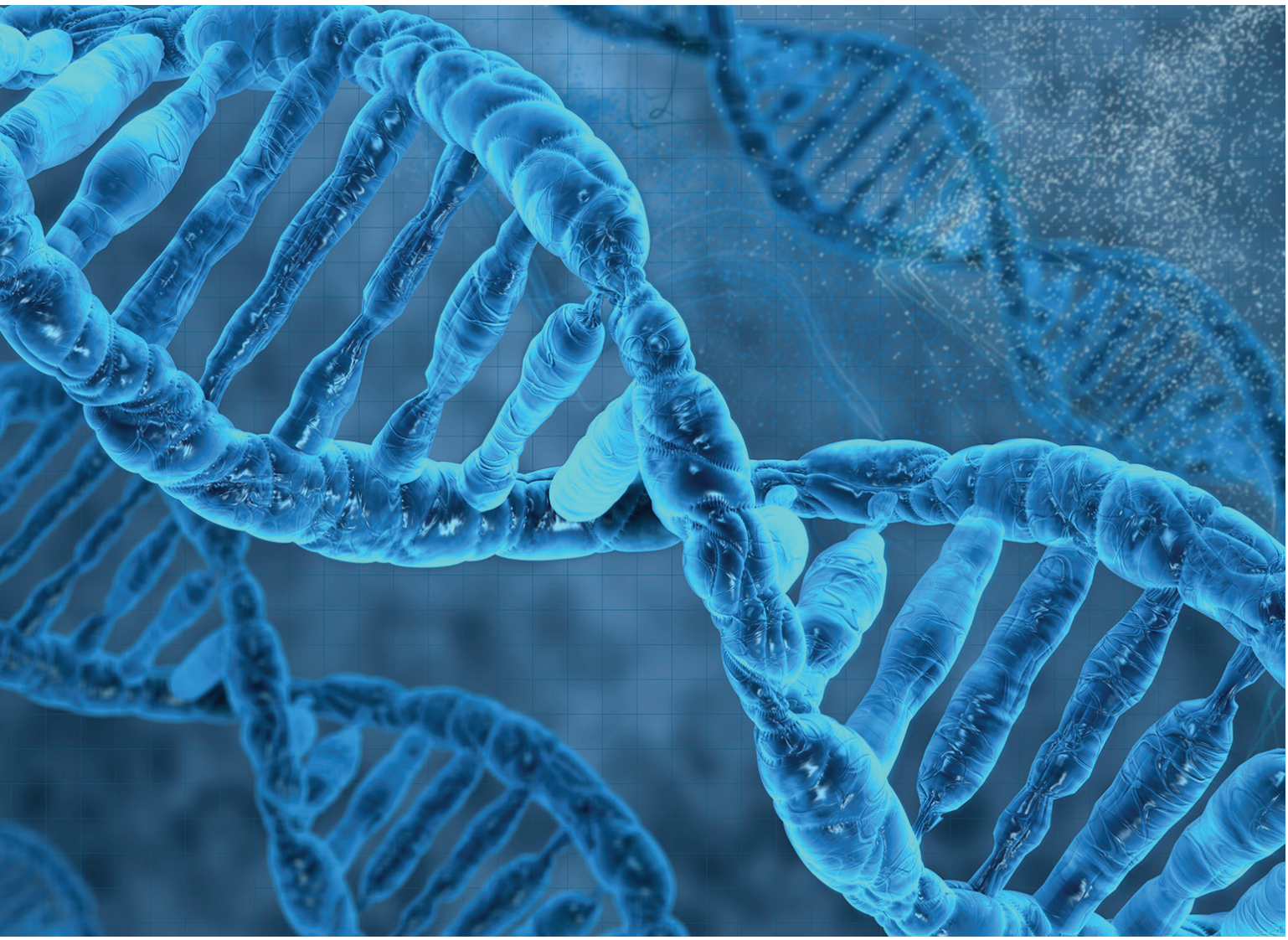
illumina®

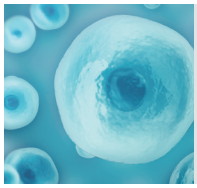
유전자 발현 및 조절 연구의 분석: 주요 고려 사항



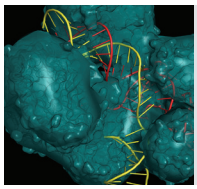
연구 전용입니다. 진단 절차에는 사용할 수 없습니다.

목차

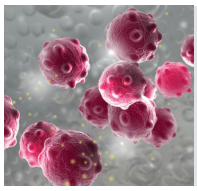
머리말



유전자 발현은 어떻게 제어되는가?



다양한 종류의 세포가 어떻게 유전자 발현과 단백질 생산을 조절하는가?



유전자 발현 이상은 다양한 질병 상태에 어떠한 기여를 하는가?

우리는 유전자 발현 연구를 통해 이를 비롯한 다양한 질문에 대한 해답을 찾을 수 있습니다.

유전자 발현(gene expression)은 DNA가 생물학적으로 활성이 있는 기능적 단위로 번역(translation)되는 복잡한 과정입니다. 유전자는 단백질을 암호화할 뿐만 아니라 생물의 발생이나 분화와 같은 필수적인 생물학적 과정 그리고 유전자 발현 이상(dysregulation)이 있을 경우 복합 질환의 발생에 관여하는 광범위한 단백질 비코딩(non-protein coding) RNA 요소도 추가로 암호화합니다.[1,2]

차세대 시퀀싱(next-generation sequencing, NGS) 기반의 RNA 시퀀싱(RNA-Seq)은 과학자들이 기존 연구 방법의 한계를 극복할 수 있도록 한 단계 더 진보된 시퀀싱 기술입니다. RNA-Seq은 전체 전사체(whole transcriptome)에 대한 유전자 발현 수준을 측정할 수 있도록 염기(base)별 코딩 및 비코딩 RNA 활성을 고해상도로 보여주는 민감도와 정확도가 높은 분석 도구입니다. 이 시퀀싱 방법은 우리가 이전에는 파악할 수 없었던 여러 질병 상태, 치료제에 대한 반응, 다양한 환경 조건, 광범위한 연구 설계 유형에서 발생하는 변화를 설명하는 데 도움을 줄 수 있습니다.

NGS 방법은 유전자 조절(gene regulation)과 단백질 발현을 이해하기 위해 더욱 발전하였고, 그 결과 ATAC-Seq(Assay for Transposase-Accessible Chromatin using Sequencing), CITE- Seq(Cellular Indexing of Transcriptomes and Epitopes by Sequencing), 메틸레이션 시퀀싱(methylation sequencing), ChIP-Seq(chromatin immunoprecipitation sequencing) 등 많은 시퀀싱 방법이 개발되었습니다.

한편 세포 분리와 시퀀싱 기술의 발전은 단일세포 시퀀싱(single-cell sequencing)의 개발과 활발한 활용으로 이어졌습니다. 이 NGS 방법은 개별 세포의 유전체, 후성유전체(epigenome), 전사체 또는 단백체(proteome)를 분석하여 복잡한 생물학적 시스템을 주도하는 세포의 이질성(heterogeneity)을 파악할 수 있도록 고해상도의 세포 간 변이(cell-to-cell variation) 데이터를 제공합니다.

데이터의 분석 그리고 시각화 및 해석 단계는 대량 샘플 시퀀싱(bulk sequencing) 연구에서나 단일세포 시퀀싱 연구에서나 통찰력 있는 결론에 도달하는 데에 핵심적인 역할을 합니다. 유전자 발현 및 조절의 연구 방법이 점차 늘어나면서 바이오인포매틱스 파이프라인(bioinformatics pipeline) 또한 매우 다양해졌습니다. 각각 뚜렷한 장단점을 지니고 있으므로 성공적인 연구를 위해서는 분석 워크플로우 전체에 걸쳐 신중하게 파이프라인을 설계하고 최적화하는 것이 매우 중요합니다.

본 e-book은 현재 제공 중인 전산 해석 파이프라인에 중점을 두고 대량 샘플 및 단일세포 시퀀싱 데이터 분석에 적용 가능한 NGS 워크플로우를 단계별로 구분하여 개략적으로 설명하고 있습니다. 나아가 연구 수행 시 반드시 고려해야 할 사항과 겪을 수 있는 어려움을 기술하고, 상용 제품을 제안하며, 성공적인 NGS 기반의 유전자 발현 및 조절 연구의 설계와 수행에 필요한 정보도 제공하고 있습니다.

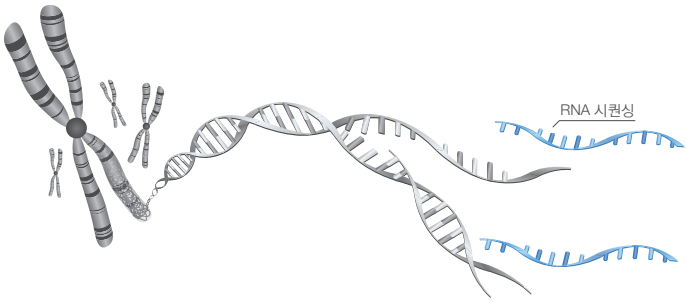
제1장: NGS 기반 유전자 발현 및 조절 연구에서 실험 설계 시 고려할 사항

정상적인 세포 발생과 질병 기전을 이해하기 위한 유전자 발현과 조절의 연구에는 다양한 대량 샘플 시퀀싱 방법을 활용해 볼 수 있습니다. 고해상도의 분석이 필요하다면 단일세포 시퀀싱 방법을 이용하여 개별 세포의 기여도와 복잡한 조직(tissue)의 전반적인 기능을 파악할 수 있습니다.

대량 유전자 발현 솔루션

Illumina의 RNA Library Prep Kit 제품군

한층 더 강화된 Illumina의 RNA Library Prep Kit 제품 포트폴리오는 일반적으로 하루 안에 완료 가능한 간소화된 워크플로우를 통해 연구자가 요구하는 고품질의 데이터를 제공합니다. Illumina는 아래와 같은 세 가지 RNA Library Prep Kit 제품을 제공하고 있습니다(표 1).



- **Illumina Stranded Total RNA Prep**은 포괄적인 생물학적 정보의 수집을 위해 Ribo-Zero™ Plus를 통한 전체 전사체 분석을 지원하며 코딩 RNA와 여러 형태의 비코딩 RNA를 포획합니다.
- **Illumina Stranded mRNA Prep**은 코딩 RNA 중심의 분석을 위한 비용 효율적인 라이브러리 준비 옵션을 제공합니다.
- **Illumina RNA Prep with Enrichment**는 RNA-Seq에 비드 결합 트랜스포좀(bead-linked transposome, BLT) 기술을 적용한 제품으로, 여러 유전체 위치에 대한 정보를 제공하기 위해 하루 안에 완료 가능한 신속한 RNA Enrichment 워크플로우를 지원합니다.

표 1: Illumina의 RNA 라이브러리 준비 솔루션

	Illumina Stranded Total RNA Prep with Ribo-Zero Plus	Illumina Stranded mRNA Prep	Illumina RNA Prep with Enrichment
설명	다양한 종류의 샘플이 사용되는 전체 전사체 시퀀싱 연구에 적합한 솔루션. 여러 종에 풍부하게 존재하는 전사물을 하나의 튜브에서 제거(single-tube depletion)할 수 있는 Ribo-Zero Plus 모듈 포함.	정밀한 가닥 정보를 활용한 코딩 전사체 분석에 적합한 간단하고 비용 대비 효율적인 솔루션.	포르말린 고정 파라핀 포매(formalin-fixed, paraffin-embedded, FFPE) 조직 샘플, 저품질 샘플 등 다양한 종류의 샘플과 샘플 사용량을 지원함으로써 표적 전사물의 검출 및 발견을 돕는 재현 가능하고 경제적인 솔루션.
분석 대상 RNA	코딩 RNA 및 다양한 형태의 비코딩 RNA 포획.	가닥 정보를 활용한 코딩 전사체 포획.	Illumina의 Exome Panel과 함께 사용 시 코딩 전사체 포획.
방법	전체 전사체 시퀀싱	mRNA 시퀀싱	mRNA 시퀀싱, 표적 농축(target enrichment)
Assay 소요 시간	약 7시간	6.5시간	< 9시간
수작업 시간	< 3시간	< 3시간	< 2시간
FFPE 지원 여부	지원	미지원	지원
자동화 지원 여부	지원	지원	지원
RNA 사용량	표준 품질의 total RNA 1~1000 ng. 최적의 성능을 위해/FFPE 샘플 사용 시 최소 10 ng의 total RNA 사용 권장.	표준 품질의 total RNA 25~1000 ng.	신선/냉동 샘플에서 얻은 total RNA 10 ng, FFPE 샘플에서 얻은 total RNA 20 ng.
작용 기전	효소를 이용한 rRNA depletion, ligation 기반 어댑터/인덱스 추가.	PolyA 포획, ligation 기반 어댑터/인덱스 추가.	BLT를 이용한 tagmentation.
상세 정보	더 알아보기	더 알아보기	더 알아보기

AmpliSeq™ for Illumina 표적 패널 제품군

AmpliSeq for Illumina는 Illumina의 NGS 플랫폼과 호환되는 AmpliSeq chemistry 제품군의 명칭입니다. AmpliSeq for Illumina 솔루션은 1회의 런(run)으로 한두 개에서 수백 개에 이르는 표적을 지정해 시퀀싱을 수행할 수 있도록 해 주는 지극히 다중의 PCR 기반 워크플로우를 지원합니다. AmpliSeq for Illumina는 Illumina의 NGS 기술을 결합하여 다양한 응용 분야에 신뢰도가 높은 데이터를 제공합니다. AmpliSeq for Illumina는 RNA 샘플도 지원하며, 적게는 1 ng의 RNA 샘플만 사용해도 분석이 가능합니다. 사용자는 다양한 RNA 기반 응용 분야에 적합한 미리 설계된 패널이나 맞춤형 콘텐츠를 선택할 수 있습니다(표 2).

- **AmpliSeq for Illumina Transcriptome Human Gene Expression Panel**은 2만 개가 넘는 인간 RefSeq 유전자의 발현 수준을 측정합니다.
- **AmpliSeq for Illumina Immune Response Panel**은 종양과 면역 체계의 상호작용에 관여하는 유전자 395개의 발현을 조사합니다.
- **AmpliSeq for Illumina Immune Repertoire Plus, TCR beta Panel**은 T세포 수용체의 베타 사슬 재배치 (T-cell receptor(TCR) beta chain rearrangement)를 시퀀싱하여 T세포의 다양성 및 클론 확장(clonal expansion)을 조사합니다.
- **AmpliSeq for Illumina Custom RNA Panel**은 2만 개가 넘는 인간 RefSeq 유전자로 구성된 메뉴를 바탕으로 설계된 하나의 assay를 통해 12~1,200개의 원하는 유전자를 대상으로 유전자 발현을 측정합니다.

표 2: AmpliSeq for Illumina의 표적 RNA 패널

	AmpliSeq for Illumina Transcriptome Human Gene Expression Panel	AmpliSeq for Illumina Immune Response Panel	AmpliSeq for Illumina Immune Repertoire Plus, TCR beta Panel	AmpliSeq for Illumina Custom RNA Panel
분석 대상 RNA	> 95%의 인간 RefSeq 유전자. 20,802개의 유전자.	면역 반응과 관련된 395개의 유전자.	CDR1, CDR2, CDR3 포함 TCRβ 사슬 재배치.	12~1,200개의 맞춤 유전자.
방법	표적 RNA-Seq			
Assay 소요 시간	6시간	6시간	5.5~7.5시간	5.5~7.5시간
수작업 시간	< 1.5시간	< 1.5시간	< 1.5시간	< 1.5시간
FFPE 지원 여부	지원	지원	미지원	지원
사용량	1~100 ng의 RNA (10 ng 권장)	1~100 ng의 RNA (10 ng 권장)	10~1000 ng	1 ng
작용 기전	멀티플렉스 PCR			
상세 정보	더 알아보기	더 알아보기	더 알아보기	더 알아보기

대량 유전자 조절 솔루션

현재 DNA/RNA와 단백질의 상호작용, 메틸화 상태(methylation state), 단백질 생산 수준에서의 유전자 조절을 연구할 수 있는 다양한 시퀀싱 방법이 마련되어 있습니다(표 3).

DNA/RNA와 단백질의 상호작용

ChIP-Seq은 단백질, DNA, RNA 사이의 상호작용을 정확하게 조사함으로써 연구자가 여러 생물학적 과정과 질병 상태의 중심이 되는 유전자 조절 이벤트를 해석할 수 있도록 해 줍니다. ChIP-Seq을 활용하면 전사 인자 결합 부위(transcription factor binding site)를 식별하고, 유전체 전체에 걸쳐 히스톤 변형(histone modification)을 추적하며, 크로마틴(chromatin)의 구조와 기능을 집중적으로 파악할 수 있습니다.

- **TruSeq™ ChIP Library Prep Kit**는 지극히 다중의 고품질 ChIP-Seq을 비용 대비 효율적으로 지원하기 위해 ChIP에서 유래된 DNA로 라이브러리를 생성하는 간단하고 가성비가 뛰어난 솔루션을 제공합니다.

메틸레이션 시퀀싱

사이토신 메틸화(cytosine methylation)는 시공간적 유전자 발현과 크로마틴 리모델링을 크게 변화시킬 수 있습니다. 연구자는 전장 유전체(whole-genome) 분석 및 표적 NGS 접근법을 통해 단일 뉴클레오티드 수준에서의 메틸화 패턴에 대한 정보를 얻을 수 있습니다.

- **전장 유전체 바이설파이트 시퀀싱(whole-genome bisulfite sequencing, WGBS)**은 포괄적인 DNA 메틸화 연구 방법으로, DNA를 바이설파이트로 변환하여 메틸화되지 않은 사이토신을 검출하며 리드(read) 수를 기준으로 전체 유전체 중 메틸화된 사이토신의 백분율을 측정합니다.
- **TruSeq Methyl Capture EPIC Library Prep Kit**는 유전체 부분군에서 관심 있는 특정 영역을 분석하는 표적 메틸레이션 시퀀싱 솔루션으로, WGBS보다 관리가 더 용이한 데이터 세트를 생성하고 더 빠른 워크플로우를 제공합니다.

크로마틴 접근성

ATAC-Seq은 전체 유전체의 크로마틴 접근성(chromatin accessibility)을 측정하는 데 널리 사용되는 NGS 방법입니다. ATAC-Seq은 열린 크로마틴 영역을 시퀀싱하여 크로마틴 포장(packaging) 및 기타 요인이 유전자 발현에 어떠한 영향을 주는지 알아낼 수 있습니다.[3] ATAC-Seq은 대량의 세포 집단 혹은 고해상도의 단일 세포를 대상으로 수행 가능합니다.

표 3: 유전자 조절 프로파일링 방법

방법	설명	상용 제품/방법 예시	상세 정보
ChIP-Seq	DNA–단백질 복합체의 면역침강 및 관련 핵산의 시퀀싱을 통해 단백질, DNA, RNA 사이의 상호작용 조사.	TruSeq ChIP Library Prep Kit	더 알아보기
WGBS	메틸화되지 않은 사이토신의 바이설파이트로의 변환을 통해 유전체 전체에 대한 DNA 메틸화 분석.	Lister R, et al. Human DNA methylomes at base resolution show widesread epigenomic differences. *Nature*. 2009;462(7271):315-322.	더 알아보기
표적 메틸레이션 시퀀싱	바이설파이트 변환 및 라이브러리 준비 단계 후 표적 농축 과정을 통해 관심 있는 유전체의 특정 영역 내 DNA 메틸화 분석.	TruSeq Methyl Capture EPIC Library Prep Kit	더 알아보기
ATAC-Seq	트랜스포사제를 사용하여 시퀀싱 어댑터를 열린 크로마틴 영역에 삽입함으로써 유전체 전체에 대한 크로마틴 접근성 평가.	Buenrostro J, et al. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29-9	더 알아보기

단일세포 시퀀싱 솔루션

조직 준비 및 세포 분리

단일세포 시퀀싱 워크플로우에서 가장 중요한 단계는 라이브러리 준비 전 단계인 첫 조직 샘플 준비와 세포 분리입니다. 초창기 단일세포 분리 방법은 처리량이 낮아 한 실험에서 단지 수십에서 수천 개의 세포만을 처리할 수 있었습니다. 반면에 최근 떠오르고 있는 처리량이 높은 미세유체(microfluidic) 기반의 세포 분리 방법은 연구자가 실험당 수백에서 수만 개의 세포를 비용 대비 효율적으로 연구할 수 있게 해 줍니다. 또한 광범위한 조직 준비, 단일세포 분리 및 라이브러리 준비 옵션이 마련되어 있어, 연구자가 다양한 종(species), 조직/세포 유형 및 방법에 따라 알맞은 옵션을 선택해 연구를 수행할 수 있습니다.

라이브러리 준비

대부분 라이브러리 준비 방법을 비롯한 세포 프로파일링 접근법은 주어진 실험 문제에 따라 결정됩니다(표 4).

전사체 프로파일링

단일세포 RNA 시퀀싱(single-cell RNA sequencing, 이하 scRNA-Seq)을 통한 개별 세포의 전사 프로파일링은 대량으로 시퀀싱된 세포 집단의 평균적인 발현 신호로는 알 수 없었던 정보를 제공할 수 있습니다. 전사체 프로파일링에는 전체 전사체나 표적 유전자 발현 패널을 이용할 수 있습니다(표 4). 면역검출(immunodetection)법의 발전으로 전사체와 단백체 프로파일링을 결합한 조합적인 멀티오믹스(multiomics) 연구도 가능해 졌습니다.

scATAC-Seq

앞서 설명한 바와 같이 ATAC-Seq을 통해 연구자는 유전체 전체에 대한 크로마틴 접근성을 살펴볼 수 있습니다. 단일세포 ATAC-Seq(single-cell ATAC-Seq, 이하 scATAC-Seq)은 단일세포의 구획화(compartmentalization) 및 바코딩과 Tn5(highly active transposase, 과활동성 트랜스포사제) tagmentation 기술을 결합한 시퀀싱 방법입니다. Tn5 트랜스포사제는 시퀀싱 어댑터를 이용해 열린 크로마틴 영역에 태그를 부착합니다. 이후 태그가 부착된 DNA 절편은 정제, 증폭 및 시퀀싱 과정을 거치게 됩니다.

단백질 프로파일링

통상적으로 단백질 발현의 분석에는 형광물질이 결합된 항체(fluorophore-conjugated antibody)를 이용하는 유세포 분석법(flow cytometry method)이 활용되어 왔습니다. 다만 이 기술에는 사용 가능한 형광물질의 수가 상대적으로 적고 형광물질과 관련해 스펙트럼이 중복(spectrum overlap)된다는 한계가 있었습니다. 이후 기술의 발전으로 단백질 검출과 정량화에 사용되는 형광 표지가 올리고뉴클레오티드(oligonucleotide) 표지로 대체되었고, 단백질 발현의 판독에 NGS를 적용할 수 있게 되었습니다.[4-6] 현재는 AbSeq, CITE-Seq, REAP-Seq(RNA expression and protein sequencing) 등 DNA로 표지된 항체를 통합한 다양한 방법을 연구에 활용할 수 있습니다.

자세한 정보는 단일세포 시퀀싱 워크플로우: 주요 단계 및 고려 사항 eBook을 통해 확인하실 수 있습니다.

표 4: 단일세포 시퀀싱에 적합한 증폭 기법

방법	설명	상용 제품/방법 예시
유전자 발현		
전장 RNA-Seq	SMART(Switching Mechanism at 5′ End of RNA Template) 기술로 전장 cDNA 증폭	• Takara SMARTer cDNA Synthesis Kits
mRNA 말단 태그 증폭 (3′ WTA 또는 5′ WTA)	가닥 특이적 정보를 바탕으로 코딩 전사체를 시퀀싱하기 위해 3′ 말단의 폴리아데닐화된 poly(A) 꼬리로 mRNA 포획	• 10x Genomics Chromium Single Cell Gene Expression Solution (3′ WTA) • 10x Genomics Chromium Single Cell Immune Profiling Solution (5′ WTA) • SureCell WTA 3′ Library Prep Kit for the ddSEQ System
표적 패널	다양한 종류의 사전 설계된 단일세포 표적 RNA 시퀀싱 패널을 면역 수용체(immune receptor: IR), T세포, 유방암 프로파일링 등에 활용	• BD Rhapsody Single-Cell Analysis
유전자 조절		
ATAC-Seq	트랜스포사제를 사용하여 시퀀싱 어댑터를 열린 크로마틴 영역에 삽입함으로써 유전체 전체에 대한 크로마틴 접근성 평가	• 10X Genomics Chromium Single Cell ATAC Solution • Abcam ATAC-Seq protocol • Bio-Rad SureCell ATAC-Seq Library Prep Kit
단백질 프로파일링		
AbSeq	DNA 태그가 부착된 항체와 NGS를 활용한 단백질 프로파일링	• BD AbSeq antibody-oligonucleotide conjugates
CITE-Seq	올리고뉴클레오티드로 표지된 항체를 사용해 세포 단백질 및 전사체 측정을 하나의 assay로 통합	• Stoeckius M, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14:865-868. • cite-seq.com
REAP-Seq	DNA-항체 접합체로 유전자 및 단백질 발현 수준을 정량화	• Peterson VM, et al. Mulitplexed quantification of proteins and transcripts in single cells. *Nat Biotech*. 2017;35:936-939.

NGS 기기 선택 방법

NGS 시스템의 선택은 모든 연구실에 중요한 과제입니다. Illumina는 대량 샘플 시퀀싱과 단일세포 시퀀싱 실험에 모두 활용 가능한 우수한 데이터 품질과 정확성을 갖춘 대규모의 혁신적인 NGS 플랫폼을 제공하고 있습니다. 연구의 예산이나 성격, 실험 목적에 따라 선택이 가능한 다양한 시스템이 준비되어 있으므로 어느 연구실에서나 요구 조건에 부합하는 시스템을 찾을 수 있습니다(표 5).

표 5: Illumina의 시퀀싱 시스템

						
시스템	**iSeq 100 시스템**	**MiniSeq 시스템**	**MiSeq 시스템**	**NextSeq 550 시스템**	**NextSeq 1000 및 NextSeq 2000 시스템**	**NovaSeq 6000 시스템**
가장 중요한 부분	가격 적정성 및 효율성	간편성 및 기기의 가격 적정성	속도, 정확성 및 간편성	엑솜, 전사체, 전장 유전체 시퀀싱을 수행할 유연한 데스크톱 시퀀싱 시스템	대용량 처리 시퀀싱을 수행하는 능력과 프로젝트/워크플로우의 요구 사항에 따라 규모 조정을 가능하게 해 주는 유연성	프로덕션 규모의 유전체 연구를 위해 대용량 처리 시퀀싱을 적은 비용으로 수행하는 능력
기기 제어 소프트웨어	Local Run Manager	Local Run Manager	Local Run Manager	Local Run Manager	Local Run Manager	Illumina Experiment Manager
온보드 인포매틱스	○	○	○	X	○	X
벤치탑 시스템	○	○	○	○	○	X
프로덕션 규모 처리 용량	X	X	X	○	○	○
플로우 셀 옵션	Standard	Mid-output, High-output	Standard v2, Micro v2, Nano v2, Standard v3	Mid-output, High-output	P2, P3	SP, S1, S2, S4
처리되는 플로우 셀의 수/런	1	1	1	1	1	1 또는 2

저용량 처리 벤치탑 시퀀싱 시스템

저용량 처리(low-throughput) 벤치탑 시퀀싱 시스템은 초기 구매 비용 측면에서 가장 합리적인 NGS 기기로, 중소 규모의 유전자 발현 및 조절 연구에 이상적이며 빠른 표적 샘플 처리 워크플로우를 제공합니다. 일반적인 응용 분야로는 mRNA 전사물(transcript)의 부분군이나 관심 있는 특정 유전체 영역을 효율적이고 비용 대비 효과적으로 시퀀싱할 수 있도록 해 주는 표적 유전자 발현이 있습니다. 일부 규모가 큰 검사실에서는 고용량 처리 시퀀싱 기기 외에도 벤치탑 시스템을 따로 마련해 두고 작은 규모의 추적관찰 연구나 라이브러리 품질 관리(QC) 등에 사용하기도 합니다.

iSeq™ 100 시스템

iSeq 100 시스템은 Illumina 기기 중 가격이 가장 저렴하고, 풋프린트(footprint)는 가장 작으며, 런 소요 시간은 가장 짧아 합리적인 가격으로 연구에 NGS를 활용할 수 있는 옵션을 제공합니다. 버튼 하나만 누르면 표적 유전자, RNA 전사물 등의 시퀀싱이 가능합니다.[7]

자세한 정보는 www.illumina.com/iseq에서 확인하시기 바랍니다.

iSeq 100 시스템의 플로우 셀 옵션 및 사양	
플로우 셀 종류	i1
데이터 아웃풋/런	144 Mb~1.2 Gb
리드 수/런	4M
최대 리드 길이	2 × 150 bp
사이클 수	300

MiniSeq™ 시스템

MiniSeq 시스템은 단일 유전자 또는 전체 경로(pathway)를 검사하는 광범위한 표적 DNA 및 RNA 시퀀싱 애플리케이션을 지원하는 간단하고 적정한 가격의 솔루션을 제공합니다. 직관적인 사용자 인터페이스, 로딩만 하면 즉시 작동하는 간단한 조작 방식, 온보드 데이터 분석이 지원되므로 사용법을 익히기 쉽고 사용하기 편리합니다.[8]

자세한 정보는 www.illumina.com/miniseq에서 확인하시기 바랍니다.

MiniSeq 시스템의 플로우 셀 옵션 및 사양				
플로우 셀 종류	Mid-output	High-output		
데이터 아웃풋/런	2.1~2.4 Gb	1.7~1.9 Gb	3.3~3.8 Gb	6.6~7.5 Gb
리드 수/런	8M	25M	25M	25M
최대 리드 길이	2 × 150 bp	1 × 75 bp	2 × 75 bp	2 × 150 bp
사이클 수	300	75	150	300

MiSeq™ 시스템

빠른 속도, 고품질의 데이터, Illumina에서 가장 긴 리드 길이(read length)를 제공하는 MiSeq 시스템은 다양한 표적 RNA 유전자 발현 및 조절 연구에 이상적이며 중간 크기의 샘플 배치(batch)에 적합합니다.[9]

자세한 정보는 www.illumina.com/miseq에서 확인하시기 바랍니다.

MiSeq 시스템의 플로우 셀 옵션 및 사양								
플로우 셀 종류	Standard v2			Micro v2	Nano v2		Standard v3	
데이터 아웃풋/런	0.75~0.85 Gb	4.5~5.1 Gb	7.5~8.5 Gb	1.2 Gb	0.3 Gb	0.5 Gb	3.3~3.8 Gb	13.2~15 Gb
리드 수/런	15M	15M	15M	4M	1M	1M	25M	25M
최대 리드 길이	2 × 25 bp	2 × 150 bp	2 × 250 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp	2 × 75 bp	2 × 300 bp
사이클 수	50	300	500	300	300	500	150	600

대용량 처리 벤치탑 시퀀싱 시스템

대용량 처리(high-throughput) 벤치탑 NGS 기기는 중간 정도의 가격대이면서도 접근성과 편의성을 제공합니다. 대용량 처리 시퀀싱 시스템은 풋프린트가 작은 벤치탑 시스템의 형태를 그대로 유지하며 일반적으로 엑솜, mRNA, 단일세포 시퀀싱 연구에 사용됩니다.

NextSeq™ 550 시스템

대용량 처리 시퀀싱 시스템의 성능과 벤치탑 NGS 시스템의 빠른 속도, 간편한 사용성, 가격 적정성을 모두 갖추고 있는 NextSeq 550 시스템은 중간 내지 높은 처리량의 시퀀싱 애플리케이션을 지원하며 작은 규모의 단일세포 시퀀싱 연구에 가장 적합합니다. NextSeq 550 시스템은 다양한 종류의 프로젝트에 효율적이고 유연하게 적용이 가능하여 하루 안에 빠른 속도로 대량 샘플 시퀀싱이나 단일세포 시퀀싱을 통해 주문형 전사체 분석을 완료할 수 있습니다.[10]

자세한 정보는 www.illumina.com/nextseq550에서 확인하시기 바랍니다.

NextSeq 550 시스템의 플로우 셀 옵션 및 사양					
플로우 셀 종류	Mid-output v2.5		High-output v2.5		
데이터 아웃풋/런	16~19 Gb	32~39 Gb	25~30 Gb	50~60 Gb	100~120 Gb
리드 수/런	130M	130M	400M	400M	400M
최대 리드 길이	2 × 75 bp	2 × 150 bp	1 × 75 bp	2 × 75 bp	2 × 150 bp
사이클 수	150	300	75	150	300

NextSeq 1000 및 NextSeq 2000 시스템

NextSeq 1000 및 NextSeq 2000 시퀀싱 시스템은 시퀀싱 반응 부피(reaction volume)의 소형화(miniaturization)와 데이터 아웃풋의 향상 및 런당 소요 비용 절감을 위해 최신 기술을 채택하였습니다. 연구자는 벤치탑 시퀀싱 시스템을 사용함으로써 연구의 규모와 범위를 확장하는 데 필요한 처리량, 데이터 품질 및 비용을 확보할 수 있습니다. 또한 향상된 기술, 진보된 chemistry, 간소화된 워크플로우, 온보드 2차 분석 기능을 통해 지금까지 경험해 보지 못한 유연한 옵션을 활용하여 연구를 진행하고 과학적 발견을 이룰 수 있습니다. NextSeq 2000 시스템은 기존의 애플리케이션에서 런 경제성을 높여주고 새롭게 떠오르는 애플리케이션의 요구에 부합하는 고처리량 옵션도 제공합니다. NextSeq 1000 시스템은 NextSeq 2000 시스템보다 처리량이 적고 가격이 저렴합니다. NextSeq 1000 시스템을 구매한 고객은 향후 실험 규모의 변경에 따라 유연하게 조정하기 위해 NextSeq 2000 시스템으로 간단하게 업그레이드를 받는 것도 가능합니다.[11]

자세한 정보는 www.illumina.com/nextseq2000에서 확인하시기 바랍니다.

NextSeq 2000 시스템의 플로우 셀 옵션 및 사양						
플로우 셀 종류	NextSeq 1000/2000 P2 Reagents			NextSeq 2000 P3 Reagents		
데이터 아웃풋/런	40 Gb	80 Gb	120 Gb	110 Gb	220 Gb	330 Gb
리드 수/런	400 M	400 M	400 M	1.1 B	1.1 B	1.1 B
최대 리드 길이	2 × 50 bp	2 × 100 bp	2 × 150 bp	2 × 50 bp	2 × 100 bp	2 × 150 bp
사이클 수	100	200	300	100	200	300

대용량 처리/고용량 시퀀싱 시스템

분주하게 돌아가는 검사실에는 샘플당 비용이 가장 적게 드는 대용량 처리/고용량(high-volume) 시퀀싱 시스템이 필요합니다. 고처리량 시스템은 검사실이 최상의 운영 효율성을 확보할 수 있도록 최소의 런만을 수행하여 프로젝트를 완료할 수 있게 해 줍니다. 단일세포 및 대량 샘플 시퀀싱 애플리케이션으로는 전장 전사체 시퀀싱, 후성유전학적 조절(epigenetic regulation) 관련 시퀀싱, 엑솜 시퀀싱, mRNA 시퀀싱 등이 있습니다.

NovaSeq™ 6000 시스템

NovaSeq 6000 시스템은 신뢰성 있고 규모 조정이 가능한 강력한 성능을 지닌 Illumina의 대용량 처리 시퀀싱 플랫폼으로, 뛰어난 데이터 품질을 제공합니다. 프로덕션 규모의 플랫폼 중 가장 높은 처리량을 지원하는 NovaSeq 6000 시스템은 연구자가 더 많은 샘플을 더 높은 커버리지 뎁스(coverage depth)로 손쉽게 연구를 수행할 수 있게 해 주므로 의약품 스크리닝, 단일세포 지도(single-cell atlas) 연구, 기타 대규모 실험과 같은 광범위한 스크리닝 연구에 가장 적합합니다.[12]

자세한 정보는 www.illumina.com/novaseq에서 확인하시기 바랍니다.

NovaSeq 6000 v1.5 Reagent Kit

Illumina는 한층 더 경제적이고 유연한 시퀀싱 워크플로우를 지원하고 시약의 유통 기한을 더 연장한 NovaSeq 6000 v1.5 Reagent Kit를 제공하고 있습니다. 검사실에서는 NovaSeq6000 v1.5 Reagent Kit를 사용하여 기존 v1.0 시약과 동일한 수준의 높은 데이터 품질을 유지하면서 검사실의 NGS 역량도 높일 수 있습니다.

자세한 정보는 *Enhanced sequencing capabilities with the NovaSeq 6000 v1.5 Reagent Kit* Technical Note를 통해 확인하실 수 있습니다.

NovaSeq 6000 시스템의 플로우 셀 옵션 및 사양												
플로우 셀 종류	SP(v1.5 시약)			S1(v1.5 시약)			S2(v1.5 시약)			S4(v1.5 시약)		
데이터 아웃풋/런	80 Gb	250 Gb	400 Gb	167 Gb	333 Gb	500 Gb	417 Gb	833 Gb	1250 Gb	350 Gb	2000 Gb	3000 Gb
리드 수/런	800M	800M	800M	1600M	1600M	1600M	4100M	4100M	4100M	8000~10,000M		
사이클 수	100	300	500	100	200	300	100	200	300	35	200	300
플로우 셀당 아웃풋												
1 × 35 bp	해당 없음			해당 없음			해당 없음			280~350 Gb		
2 × 50 bp	65~80 Gb			134~167 Gb			333~417 Gb			해당 없음		
2 × 100 bp	134~167 Gb			266~333 Gb			667~833 Gb			1600~2000 Gb		
2 × 150 bp	200~250 Gb			400~500 Gb			1000~1250 Gb			2400~3000 Gb		
2 × 250 bp	325~400 Gb			해당 없음			해당 없음			해당 없음		

시퀀싱 수행 시 기술적으로 고려할 사항

NGS 시스템을 선택하기에 앞서 리드 뎁스, 싱글 리드(single-read, 단방향 리드) 또는 페어드 엔드 리드(paired-end read, 양방향 리드), 품질 메트릭스(quality metrics), 기기 제어 소프트웨어(control software) 등 몇 가지 기술적인 요소를 고려해 봐야 합니다.

시퀀싱 커버리지(리드) 뎁스

일반 샘플 또는 대량 샘플의 시퀀싱 커버리지는 알려진 참조 염기(reference base)에 정렬(alignment)되는, 즉 참조 염기를 커버하는 평균 리드 수를 의미합니다. 시퀀싱 커버리지 수준(즉, 리드 뎁스)은 보통 특정 염기 위치에서 발견되는 변이의 신뢰 수준을 결정합니다.

요구되는 시퀀싱 커버리지는 애플리케이션마다 차이가 있습니다. 커버리지가 높으면 각각의 염기가 더 많은 수의 정렬된 시퀀스 리드로 커버되므로 신뢰 수준이 더 높은 베이스 콜(base call)을 얻을 수 있습니다(그림 1).[13]

다양한 단일세포 시퀀싱 애플리케이션에서 리드 뎁스는 베이스당 리드 수가 아닌 세포당 리드 수를 기준으로 설명합니다. 1회의 단일세포 시퀀싱 런에 요구되는 시퀀싱 뎁스는 샘플의 종류, 분석하고자 하는 세포의 수, 실험 목적 등 여러 요인에 따라 달라집니다. 요구되는 시퀀싱 뎁스는 대개 샘플의 종류와 실험의 목적에 따라 결정되며 각 연구에 맞게 최적화해야 합니다.

자세한 정보는 www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html에서 확인하시기 바랍니다.

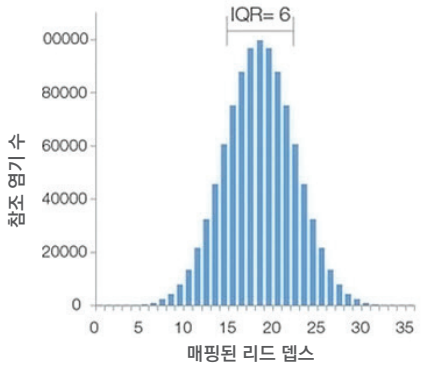


그림 1: 시퀀싱 커버리지 히스토그램 — 표준 편차가 작은 포아송 분포와 유사한 형태를 나타내는 그래프 예시(리드가 유전체 전체에 걸쳐 무작위로 분포되어 있고 1회의 시퀀싱 런에서 리드 간 일관된 참(true) 중복 검출 능력을 보인다는 가정에 적합함).

시퀀싱 리드 옵션

시퀀싱 수행 시 싱글 리드 또는 페어드 엔드 리드 옵션을 선택할 수 있습니다. DNA를 한 쪽 끝에서만 시퀀싱하는 싱글 리드 시퀀싱은 가장 간단하게 Illumina의 시퀀싱 기술을 활용할 수 있는 방법입니다. 싱글 리드 시퀀싱은 대량의 고품질 데이터를 페어드 엔드 시퀀싱보다 빠르고 저렴하게 제공합니다.[14] 싱글 리드 런은 small RNA 시퀀싱과 같은 일부 시퀀싱 방법에 적합합니다. 반면 페어드 엔드 시퀀싱은 하나의 라이브러리에서 DNA 절편의 양쪽 끝을 시퀀싱하고 순방향(forward) 리드와 역방향(reverse) 리드를 리드 페어(read pair)로 정렬합니다. 페어드 엔드 시퀀싱은 특히 시퀀싱이 어려운 반복 영역에서 리드의 정렬을 향상시켜 줍니다(그림 2). Illumina의 모든 NGS 시스템은 페어드 엔드 시퀀싱을 지원합니다.

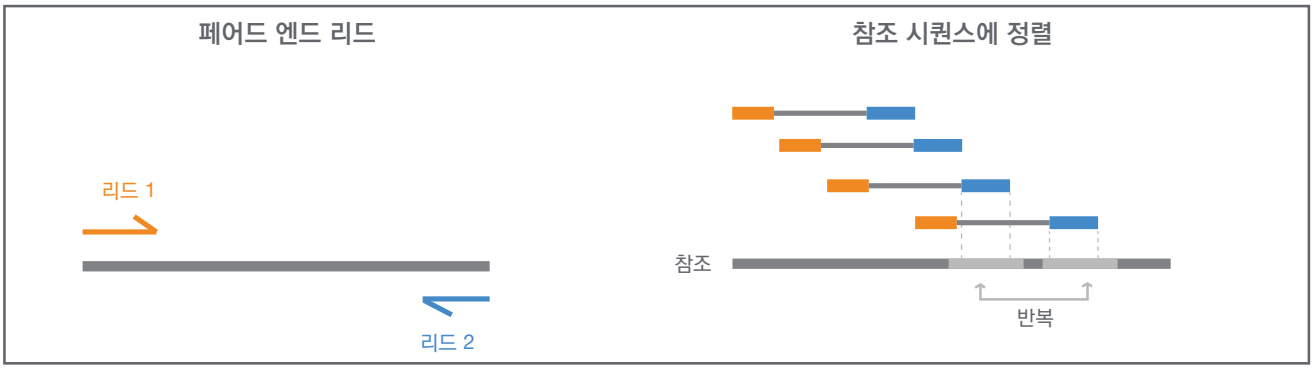


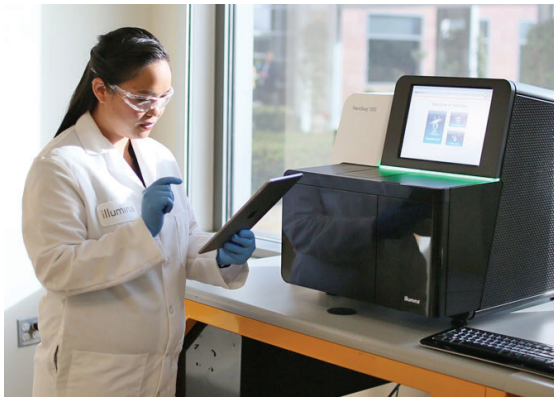
그림 2: 페어드 엔드 시퀀싱 및 정렬 — 페어드 엔드 시퀀싱은 DNA 절편 양 끝의 시퀀싱 가능. 정렬 알고리즘은 알려져 있는 각 리드 사이의 거리를 참조해 싱글 리드 시퀀싱보다 정확하게 반복 영역의 리드를 매핑.

연구자는 리드 페어로 정렬된 시퀀스 정보를 활용하여 라이브러리 준비 단계에서 같은 시간과 노력을 들여 두 배 많은 리드를 얻을 수 있을 뿐만 아니라 싱글 리드 데이터로는 검출이 불가능한 indel(insertion-deletion, 삽입-결실) 변이도 검출할 수 있습니다.[15] 페어드 엔드 RNA 시퀀싱은 gene fusion(유전자 융합), 새로운 전사물, 새로운 splice isoform(동형 접합) 등의 발견에도 응용해 볼 수 있습니다.[16]

NGS 품질 메트릭스

연구자는 시퀀싱 품질 메트릭스를 통해 NGS 워크플로우 각 단계의 정확도에 관한 중요한 정보를 얻을 수 있습니다. Phred quality score(Q-score)로 측정되는 베이스 콜링 정확도(base calling accuracy)는 시퀀싱 플랫폼의 정확도를 평가하는 데 사용되는 일반적인 메트릭이며, 시퀀싱 시스템이 특정 염기를 부정확하게 검출했을 확률을 보여줍니다.

Q30 점수는 Q-score가 30점 이상인 염기의 비율을 나타내며, Q-score 30은 베이스 콜링에 오류가 있을 확률이 1/1000, 즉 베이스 콜링의 정확도가 99.9%임을 의미합니다. Illumina의 모든 시퀀싱 시스템의 염기 Q-score는 대부분 Q30 이상(% ≥ Q30)으로 측정되므로 Illumina의 sequencing by synthesis(SBS) chemistry가 매우 높은 비율로 오류 없는 리드를 생성한다는 사실을 확인할 수 있습니다.



기기 제어 소프트웨어

대부분의 Illumina의 시퀀싱 시스템에는 런 생성, 상태 모니터링, 시퀀싱 데이터 분석 및 결과 확인에 사용되는 내장된 통합형 솔루션인 Local Run Manager가 설치되어 있습니다. Local Run Manager는 설정 단계를 간소화하고 사용자로 인한 오류의 발생을 최소화해 줍니다. 연구자는 이 소프트웨어를 통해 실시간 데이터 및 성능 메트릭스를 기기에서 바로 확인할 수 있습니다. Local Run Manager는 사용이 용이한 내장형 인터페이스 또는 웹 브라우저를 통해 이용할 수 있습니다.

NovaSeq 6000 시스템은 Illumina 시퀀싱 런을 수행하기에 앞서 Illumina Experiment Manager를 이용하여 실험을 설계합니다. Illumina Experiment Manager는 사용자에게 단계별 샘플 시트 생성 및 설정 지침을 제공합니다. 또 자체적인 유효성 검사(validation check)를 통해 차선의 인덱스 조합을 감지하고 이를 경고하여 오류를 최소화할 수 있도록 해 줍니다.

BaseSpace™ Clarity LIMS는 검사실에서 최적화된 효율적인 시퀀싱을 수행할 수 있도록 샘플의 추적, 워크플로우의 관리, 작업의 간소화를 도와주는 검사실 정보 관리 시스템(laboratory information management system, LIMS)입니다. BaseSpace Clarity LIMS는 Illumina의 모든 기기와 쉽게 통합됩니다.

요약

대량 샘플 실험이든 단일세포 실험이든, 유전자 발현 및 조절 실험 유형에 따라 특별히 요구되는 라이브러리 준비 및 시퀀싱 조건이 있기 마련입니다. Illumina는 다양한 대량 샘플 RNA-Seq 애플리케이션을 지원하는 강력한 성능의 라이브러리 준비 솔루션을 제공하고 있습니다. Illumina는 또한 높은 데이터 정확도와 유연한 처리량 옵션을 제공하는 시퀀싱 시스템과 호환이 가능한 타사의 scRNA-Seq 라이브러리 준비 솔루션도 지원합니다. 이를 토대로 Illumina는 어떠한 규모의 유전자 발현 및 조절 연구에도 활용할 수 있는 입증된 NGS 솔루션을 제공합니다. 연구자는 신뢰성 있는 고품질의 RNA-Seq 데이터를 확보한 후에 데이터 시각화, 분석 및 해석 단계를 진행할 수 있습니다.

제2장: NGS 기반 유전자 발현 및 조절 연구용 바이오인포매틱스 워크플로우

NGS 기반의 유전자 발현 및 조절 실험은 대용량의 raw data를 생성합니다. 이렇게 방대한 양의 데이터를 분석하는 것은 벅차고 어려운 일일 수 있습니다. 다행히 오늘날 사용 가능한 데이터 저장 및 분석 도구들을 통해 기존의 데이터 분석 과정에서 수동으로 처리하던 작업의 많은 부분을 생략할 수 있습니다. Illumina의 바이오인포매틱스(bioinformatics) 도구는 사용이 용이하여 이러한 분석 과정을 더욱 원활하게 만들어 주고 연구자가 유의미한 생물학적 정보를 신속하게 얻게 해 줍니다.

NGS 데이터 분석 워크플로우는 1차, 2차, 3차 분석의 3가지 주요 분석 단계를 거칩니다(그림 3). 1차 데이터 분석 단계에서는 유전자 정보를 뉴클레오티드 시퀀스 리드로 디지털화하고 미가공 시퀀싱 파일을 다양하게 활용 가능한 파일 형식으로 변환합니다. 2차 분석 단계에서는 1개의 샘플 내 유전자 및/또는 전사물의 개수 측정 등 유전적 특성을 규명하는 작업을 수행하기 위해 시퀀싱 리드를 디멀티플렉싱(demultiplexing, 필요시)한 후 주석(annotation)이 달린 참조 유전체(참조 유전체가 없는 경우 *de novo* 조립 유전체)에 정렬합니다. 연구자의 실험 설계 목표에 따라 차이가 있지만, 3차 분석 단계에서는 일반적으로 변이 주석, 필터링, 우선순위화, 데이터 시각화, 보고 등 더욱 광범위한 생물학적 맥락에서 앞서 얻은 데이터를 해석하는 작업을 수행합니다.

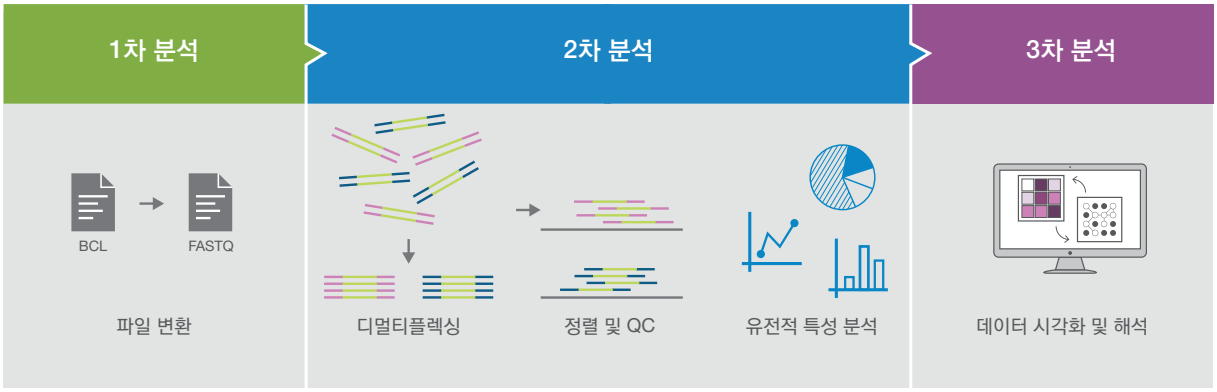


그림 3: NGS 데이터 분석 워크플로우 — 미가공 베이스 콜(BCL) 파일을 텍스트 기반의 시퀀스(FASTQ) 파일로 변환하는 1차 분석, 리드를 디멀티플렉싱하고 참조 유전체에 정렬한 후 시퀀스의 기본적인 유전적 특성을 규명하는 2차 분석, 그리고 고급 데이터 시각화 및 생물학적 해석을 수행하는 3차 분석의 3가지 주요 단계로 구성된 NGS 데이터 분석 과정.

1차 분석 — 파일 변환

모든 Illumina 시퀀싱 시스템은 SBS chemistry 및 이미징 사이클이 진행되는 동안 실행되는 내장형 Real-Time Analysis(RTA) 소프트웨어를 통해 1차 분석 단계를 시작합니다. RTA 소프트웨어는 DNA나 RNA 가닥의 기본 구조를 나타내는 베이스 콜과 이에 대한 품질 점수를 제공합니다. 시퀀싱 런 후 생성되는 raw data는 각각 BCL 형식의 베이스 콜 파일로 저장됩니다. 시퀀싱 완료 후 BCL 파일은 후속 분석(downstream analysis) 도구에서 사용될 수 있는 FASTQ 형식으로 변환되어야 합니다. FASTQ는 미가공 시퀀스 데이터와 품질 점수를 모두 저장하는 텍스트 기반의 시퀀스 파일 형식입니다. FASTQ 파일에는 후속 분석 과정에서 기준 미달의 리드를 필터링하고 폐기하는 데 사용되는 품질 점수가 포함되어 있기 때문에 시퀀싱 QC에서 중요한 부분을 차지합니다. FASTQ 파일은 또한 Illumina의 시퀀싱 시스템에서 생성된 NGS 데이터를 저장할 때 사용하는 기본 파일 형식이며 다양한 2차 데이터 분석 솔루션에서 인풋 파일로도 사용할 수 있습니다.

2차 분석 — 디멀티플렉싱, 정렬 및 QC, 유전적 특성 분석

시퀀싱 런 완료 후 미가공 리드 파일이 FASTQ 파일로 변환되면 연구자는 다음 단계인 2차 분석을 진행할 수 있습니다. NGS 데이터는 디멀티플렉싱 과정을 거친 뒤 참조 유전체에 정렬됩니다. 후속 분석을 실시하기 전에 다양한 데이터 QC 메트릭스를 적용하여 데이터의 품질을 평가해 볼 수 있습니다.

디멀티플렉싱

Illumina 시퀀싱 시스템의 용량 증대에 큰 기여를 한 것은 바로 라이브러리 준비 단계에서 각 DNA 절편에 인덱스(index)라는 고유한 시퀀스를 추가하는 멀티플렉싱(multiplexing) 기술입니다. 멀티플렉싱을 통해 단 1회의 시퀀싱 런으로 많은 수의 라이브러리를 동시에 풀링(pooling) 및 시퀀싱할 수 있습니다. 한편 디멀티플렉싱(demultiplexing)이란 인덱스 시퀀스를 바탕으로 풀링된 라이브러리의 시퀀싱 리드를 개별 라이브러리로 분리하는 기술을 의미합니다. 단일세포 시퀀싱의 경우에는 디멀티플렉싱을 통해 세포 분리 과정에서 추가되었던 세포 바코드 혹은 고유한 분자 식별자(unique molecular identifier, UMI)를 기준으로 풀링된 샘플의 리드를 개별 세포로 분리합니다.

정렬

정렬은 FASTQ 파일을 참조 유전체에 매핑하는 작업입니다. 관련 참조 유전체가 없는 경우 리드는 콘티그(contig)라는 더 긴 연속적인 분절로 조립됩니다. BWA Aligner BaseSpace App에 사용된 Burrows-Wheeler Alignment(BWA)[17] 알고리즘, RNA-Seq Alignment BaseSpace App에 포함되어 있는 Spliced Transcripts Alignment to a Reference(STAR)[18] 알고리즘 등 다양한 소프트웨어 앱을 이용하여 시퀀스를 정렬할 수 있습니다(표 6).

표 6: 시퀀스 정렬 BaseSpace 앱

BaseSpace 앱	설명	상제 정보
 BWA Aligner	BWA-MEM Aligner를 이용해 FASTQ 파일로 구성된 샘플을 참조 유전체 (예: 가져온 FASTA 파일로 생성한 사용자 맞춤형 참조 유전체)에 정렬.	더 알아보기
 RNA-Seq Alignment	STAR Aligner를 이용한 리드 매핑, Salmon을 이용한 참조 유전자 및 전사물 정량화, Strelka Variant Caller를 이용한 변이 검출(SNV 및 small indel), Manta를 통한 fusion 검출, Picard 및 기타 출처의 QC 메트릭스 수행.	더 알아보기

QC 메트릭스

QC는 실험 결과의 정확성과 재현성에 대한 신뢰도를 높여 주므로 모든 NGS 실험의 핵심적인 구성요소라 할 수 있습니다. QC 메트릭스로는 사용한 RNA의 품질, 시퀀싱 시스템이 생성한 미가공 리드 데이터, 시퀀싱 리드의 매핑 및 정렬 등이 있습니다.

미가공 리드 데이터의 QC

미가공 시퀀싱 리드 데이터를 평가하는 데 일반적으로 사용되는 QC 메트릭스로는 GC 함량, 시퀀싱 Q-Score, 시퀀스의 표현, K-mer가 있습니다. GC 함량은 하나의 참조 시퀀스에서의 구아닌(G)과 사이토신(C)의 백분율을 의미하며 생성된 시퀀스의 예상 GC 함량의 근사치를 구하는 데 사용할 수 있습니다. 시퀀스 데이터의 GC 함량이 참조 시퀀스를 크게 벗어난다면 샘플이 오염되었다는 뜻일 수 있습니다. RNA-Seq의 경우 GC 함량은 RNA 종류에 따라 차이가 있습니다(표 7). RNA-Seq 방법에 따라 예상 GC 함량을 하나의 QC 메트릭으로 사용해 볼 수 있습니다(예: total RNA-Seq의 경우 39.7~48.9%).[21]

표 7: 인간의 RNA 종류별 예상 GC 함량

RNA 종류	예상 GC 함량
코딩 RNA	48.9%
긴 비코딩 RNA	39.7%
rRNA	50.2%
miRNA	51.5%
tRNA(Transfer RNA, 운반 RNA)	55.7%
기타 small RNA	46.7%

제1장에서 설명한 바와 같이, Q-Score는 기기의 베이스 콜링 정확도를 평가하는 데 활용할 수 있으며, Q30 점수는 정확도가 99.9%임을 의미합니다. Illumina의 모든 시퀀싱 시스템은 1회의 시퀀싱 런 전반에 걸쳐 평균 75% 이상의 염기가 Q30을 넘는 정확도를 보입니다. RNA-Seq 수행 시 유전자 발현 수준과 관련해 미가공 리드 데이터를 평가할 때는 추가적인 메트릭스가 있으면 유용합니다. 고품질 시퀀싱 라이브러리는 다양한 RNA 시퀀스로 구성되어 있으므로 특정 시퀀스가 과하게 표현된다면

시퀀싱 어댑터나 다른 소스로 인한 오염을 의심해 볼 수 있습니다. 시퀀스 표현(sequence representation) 분석에서 비교적 짧은 시퀀스(뉴클레오티드 10개 미만)는 간과될 수도 있습니다. K-mer 분석에서는 짧고 중복된 시퀀스를 조사하기 위해 특정 길이(k)의 모든 가능한 뉴클레오티드 조합을 검사합니다.

정렬 QC

RNA-Seq의 경우 원하는 표적에 매핑되는 시퀀싱된 리드의 총 개수의 백분율인 포획 효율(capture efficiency)이 리드 정렬에서 중요한 부분을 차지합니다. NGS 방법의 포획 효율은 100%가 아니며, 엑솜 시퀀싱이나 RNA-Seq과 같은 시퀀싱 방법의 포획 효율은 50%에서 80% 사이입니다.[22] 일반적으로 샘플의 품질이 높을수록 포획 효율이 높습니다. RNA-Seq 데이터에서 낮은 포획 효율이 관찰될 경우 샘플의 품질이 낮거나 다른 샘플 RNA에 의한 오염을 의심해 볼 수 있으며, total RNA-Seq의 경우에는 비효율적인 rRNA의 제거를 의미할 수도 있습니다.

RNA 품질 및 무결성

RNA의 품질과 무결성(integrity)은 고품질 RNA-Seq 결과를 성공적으로 도출하는 데 매우 중요합니다. RNA의 무결성은 두 가지 리보솜 RNA(rRNA)의 비를 나타내는 28s:18s를 평가하여 측정할 수 있습니다. 일반적으로 28s:18s의 비가 높을수록 RNA의 품질과 무결성이 높습니다. RNA 샘플을 Agilent사의 Bioanalyzer 시스템과 Agilent사의 Fragment Analyzer 시스템으로 분석하면 RNA의 무결성을 평가하는 두 가지 동등한 메트릭스인 RNA 무결성값(RNA Integrity Number, RIN)과 RNA 품질값(RNA Quality Number, RQN)을 각각 얻을 수 있습니다.[19,20]

RNA 무결성 평가에 대한 더 자세한 정보는 *Scalable Nucleic Acid Quality Assessments for Illumina NGS Preparation* Application Note를 통해 확인하실 수 있습니다.

유전적 특성 분석

연구자는 Illumina의 유전체학 클라우드 컴퓨팅 플랫폼인 BaseSpace Sequence Hub를 통해 NGS 데이터를 즉시 안전하게 전송, 저장, 분석할 수 있습니다. 또한 Illumina의 DRAGEN™ Bio-IT Platform은 BaseSpace Sequence Hub나 온프레미스(on-premise) 방식을 통해 NGS 데이터의 정확한 초고속 2차 분석을 제공합니다.

BaseSpace Sequence Hub

BaseSpace Sequence Hub는 인터넷에 연결된 컴퓨터 또는 모바일 기기를 통해 사용할 수 있으며 모든 연구자가 런을 설정하고 기기의 런 품질을 모니터링할 수 있도록 보안을 최우선으로 하는 환경을 만들어 줍니다. BaseSpace Sequence Hub는 시퀀싱 데이터의 저장 및 분석 절차를 간소화함으로써 검사실의 자체적인 컴퓨팅 인프라 구축을 위한 자본적 지출의 필요성을 줄여 줍니다.[23] 숙련된 사용자라면 확장성 애플리케이션 프로그래밍 인터페이스(extensible API)를 통해 BaseSpace Sequence Hub를 검사실의 자체적인 인포매틱스(informatics) 시스템과 통합하여 사용해 볼 수도 있습니다.

BaseSpace Sequence Hub는 간단히 버튼만 누르면 실행되는 다양한 데이터 분석 솔루션과 버전 제어를 지원하는 완성된 워크플로우를 제공합니다. 또한 개인 맞춤형 분석 워크플로우를 위한 간단하고 안전하며 효율적인 파이프라인 구성을 지원합니다. BaseSpace Apps Store에서 Illumina가 개발하고 최적화한 다양한 도구나 여러 외부 앱 제공 업체가 개발한 도구를 선택해 이용 가능합니다(그림 4).

자세한 정보는 www.illumina.com/basespace에서 확인하시기 바랍니다.

DRAGEN Bio-IT Platform

DRAGEN(Dynamic Read Analysis for GENomics) Bio-IT Platform은 재구성이 용이한 필드 프로그래밍 가능 게이트 어레이(field-programmable gate array technology, FPGA)기술을 채택하여 BCL 변환, 매핑, 정렬, 분류, 중복 표시, 하플로타입(haploytype) 변이 검출과 같은 하드웨어 가속화(hardware-accelerated)된 유전체 분석 알고리즘의 구현을 지원합니다. DRAGEN Platform은 또한 긴 컴퓨팅 시간, 방대한 데이터 양 등 유전체학 분석 시 흔히 발생하는 문제를 해결해 줄 수 있는 기본 기능을 탑재하고 있습니다.[24] DRAGEN Platform은 재프로그래밍도 가능하기 때문에 Illumina는 이를 활용해 맞춤형 알고리즘을 개발하고 미래의 애플리케이션을 수용하기 위한 기술을 더욱 향상할 수 있습니다.

자세한 정보는 www.illumina.com/DRAGEN에서 확인하시기 바랍니다.

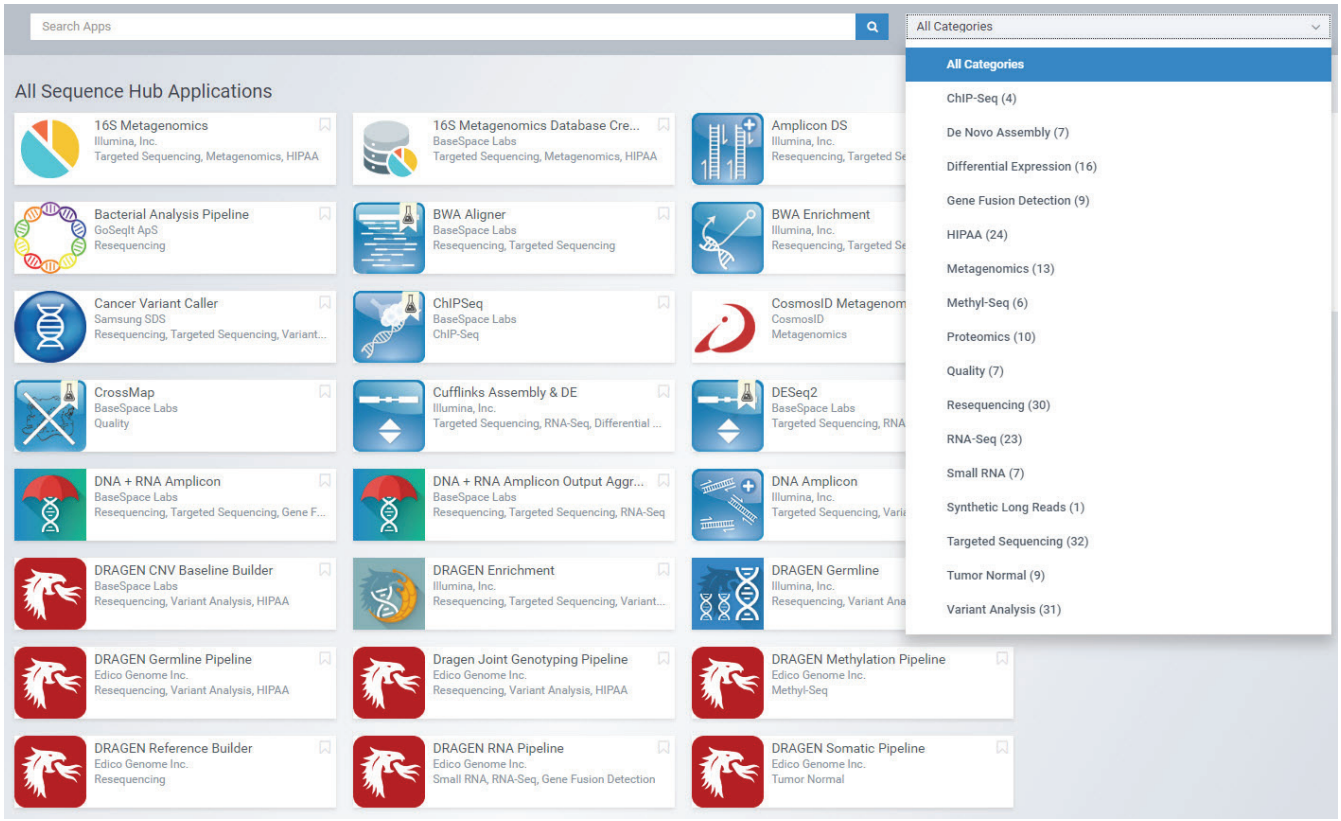


그림 4: 주문형 분석 도구 — BaseSpace Apps Store 내 Bioinformatics Community에서 DRAGEN 파이프라인을 비롯한 다양한 앱을 살펴보고 원하는 앱을 선택한 후 클릭 한 번으로 해당 앱을 데이터 세트에서 바로 실행 가능.

3차 분석 — 데이터 시각화 및 해석

3차 분석의 목표는 2차 분석을 통해 얻은 결과에 생물학적 맥락을 추가하는 것입니다. 3차 분석 단계에서는 생물학적 질문과 시퀀싱 방법에 따라 해당 연구에 적합한 다양한 후속 조사를 따로 나누어 진행할 수 있습니다. 예를 들어, 차등(differential) 유전자 발현의 분석에서는 먼저 각 샘플의 유전체 발현을 정량화한 후 통계적으로 풍부도(abundance) 수준이 다른 유전자나 전사물을 식별하기 위해 그룹 간 비교를 수행합니다. 또 이러한 데이터는 2차 분석 결과의 광범위한 영향에 대한 이해를 높이기 위해 표현형(phenotype) 정보, 기능 유전체학(functional genomics)의 생물학적 정보, 혹은 다른 출처의 데이터와 결합해 볼 수 있습니다. 후성유전학적 접근법, 단백체적 접근법 등 멀티오믹스 관점에서 생물을 연구하려면 다른 종류의 유전체 데이터를 사용하여 2차 분석 결과를 부가적인 실험 데이터와 결합해 볼 수도 있습니다. 이미 대화식 고급 데이터 시각화 및 탐색 기능을 통해 데이터의 해석을 돕는 다양한 소프트웨어 프로그램과 도구가 마련되어 있습니다(그림 5).

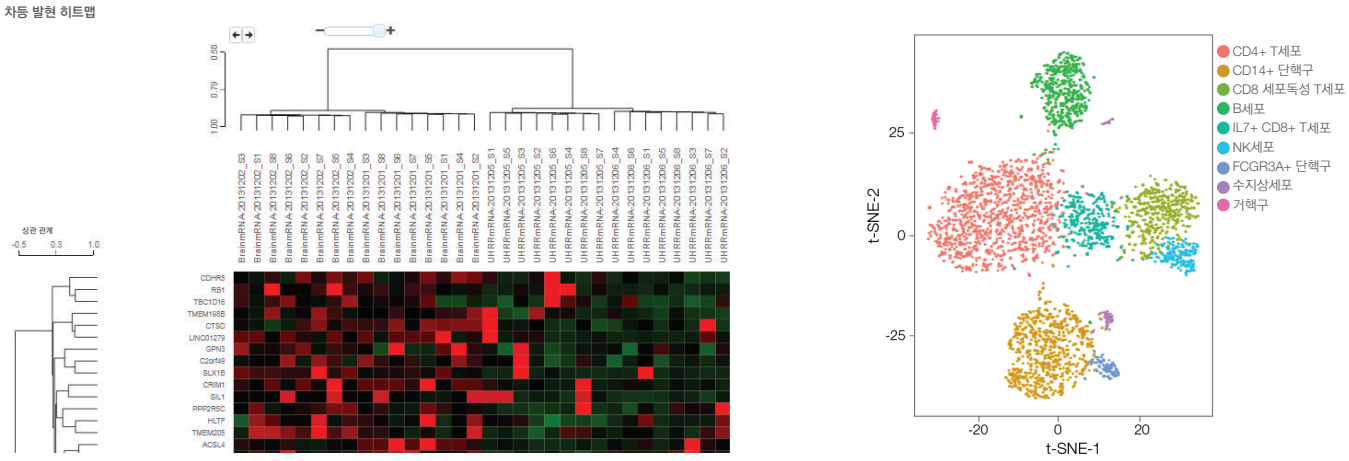


그림 5: 고급 데이터 시각화 도구 — 특정 연구 유형과 NGS 방법에 적합한 다양한 NGS 데이터 시각화 및 탐색 소프트웨어 도구 선택 가능. 차등 유전자 발현을 시각화한 히트맵(좌)과 단일세포 시퀀싱 데이터의 탐색을 지원하는 세포 클러스터 식별 도구(우) 참조. 히트맵은 RNA-Seq Differential Expression BaseSpace App으로, 세포 클러스터 플롯은 Seurat 소프트웨어로 생성.

연구자는 생물학적인 해석 방법을 적용하여 기본적인 생물학적 과정과 유전 질환의 원인에 대한 이해를 높일 수 있습니다. 바이오인포매틱스 소프트웨어와 생물학적 데이터 마이닝(data mining) 접근법은 시퀀스 데이터를 지식으로 전환해 줍니다. Illumina는 새로운 실험 데이터 세트를 오픈 액세스(open-access) 및 통제된 접근을 허용하는(controlled-access) 공용 NGS 데이터와 마이크로어레이(microarray) 데이터가 있는 큐레이션된 대규모 저장소와 연결해 주는 바이오인포매틱스 솔루션인 BaseSpace Correlation Engine을 제공하고 있습니다(그림 6). 연구자는 BaseSpace Correlation Engine의 대화식 데이터 분석 환경에서 데이터 기반 접근법을 사용하여 새로운 데이터와 관련성이 있는 연구를 찾아 질병, 조직, 문헌과의 새로운 연관성을 발견할 수 있습니다.[25]

BaseSpace Correlation Engine에 대한 자세한 정보는 www.illumina.com/products/by-type/informatics-products/basespace-correlation-engine.html에서 확인하시기 바랍니다.

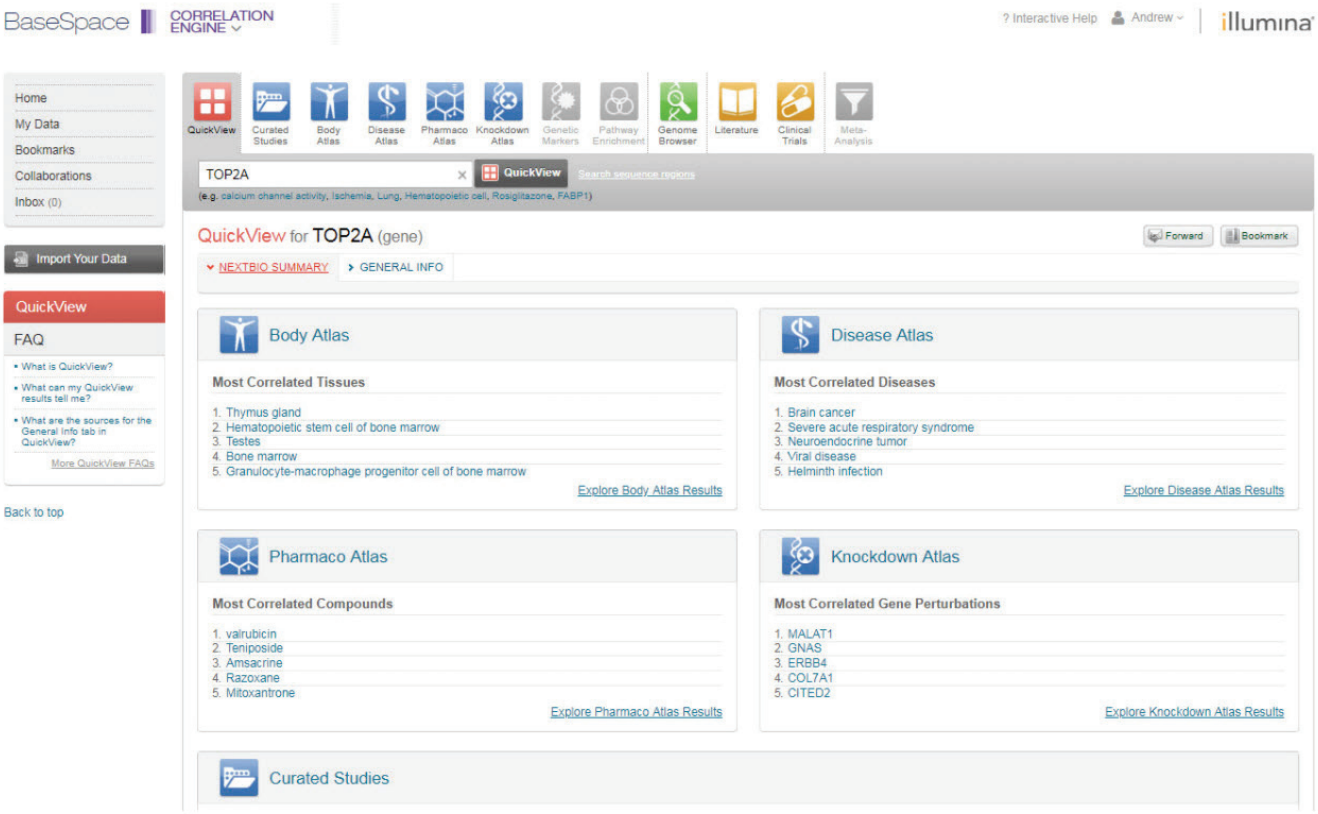


그림 6: BaseSpace Correlation Engine — 소프트웨어 인터페이스를 통해 특정 쿼리에 대한 새로운 상관관계와 연관성을 신속하게 파악하여 데이터 기반의 유전자, 질병, 화합물, 조직, 경로, 문헌 간 관련성을 발견 가능.

요약

연구자는 NGS 기술을 활용한 유전자 발현 및 조절 실험을 수행할 때 여러 종류의 샘플과 생물학적 질문에 알맞은 강력한 성능의 다양한 바이오인포매틱스 분석 도구를 선택해 사용해 볼 수 있습니다. 기본적인 유전자 정량화 및 차등 발현 분석에서 *de novo* 전사체 조립 방법에 이르기까지, NGS는 유전자 발현, 조절, 단백질로의 전사(transcription)에 대한 생물학적 이해를 높이기 위해 지속적으로 발전하고 있습니다. NGS 분석 워크플로우는 파일 변환(1차 분석), 데이터 정렬 및 유전적 특성 분석(2차 분석), 고급 데이터 시각화 및 생물학적 해석(3차 분석)의 순서로 진행됩니다. 실험의 목적과 선택한 시퀀싱 방법에 따라 2차 및 3차 분석을 용이하게 해 줄 매우 다양한 바이오인포매틱스 도구와 파이프라인이 준비되어 있습니다.

제3장: NGS 기반 유전자 발현 및 조절 연구용 바이오인포매틱스 파이프라인

대량 샘플을 사용한 NGS 기반의 유전자 발현 및 조절 연구는 복잡한 대용량의 데이터 세트를 생성합니다. 이러한 데이터의 분석에는 근본적인 생물학적 절차에 대한 유의미한 정보를 제공할 수 있는 빠르고 규모 조정이 가능하며 사용이 용이한 분석 소프트웨어가 필요합니다. 현재 제공되는 바이오인포매틱스 도구와 파이프라인은 연구자라면 누구나 바이오인포매틱 경험 유무에 관계없이 데이터 분석에 활용할 수 있도록 개발되었습니다. 제3장에서는 대량 샘플 RNA-Seq 및 ATAC-Seq 연구에서 2차 및 3차 분석 수행 시 적용 가능한 NGS 바이오인포매틱스 워크플로우에 대해 설명합니다.

대량 유전자 발현 연구용 바이오인포매틱스 솔루션

RNA-Seq 유전자 발현 연구의 유전적 특성 분석 단계에서는 유전자 위치를 확인하고 전사물을 식별하기 위해 주석이 달린 참조 유전체를 사용하여 각 유전자 위치에서의 전사물의 풍부도를 정량화합니다. 차등 발현 분석에서 생성되는 결과 파일(output file)에는 각 유전자나 전사물의 위치(행), 각 샘플이나 그룹(열), 발현 차이의 통계적 유의성을 나타내는 p-값 또는 q-값이 포함되어 있습니다. 분석에 정렬이 포함되어 있으면 각 리드가 참조 유전체의 어느 위치에 정렬되어 있는지 알려주는 BAM 파일이 생성됩니다. 샘플에 관련 참조 유전체가 없는 경우의 2차 분석은 스캐폴드(scaffold) 유전체 제공을 위한 유전체 조립과 이를 이용한 전사물 정량화 등의 몇 가지 단계로 이루어질 수 있습니다. 전사물 풍부도의 정량화뿐 아니라 변이의 식별과 정량화도 수행 가능합니다. SNV와 indel은 VCF 또는 genome VCF(gVCF) 형식의 파일로 보고되며, splice 변이도 별도의 VCF 파일에 포함될 수 있습니다.

유전자 발현의 2차 분석에는 상대적인 장점과 단점에 따라 접근 방식이 상이한 여러가지 분석 도구를 활용해 볼 수 있습니다(표 8). 예를 들어, 대부분의 분석 파이프라인은 첫 단계를 정렬로 시작하지만(예: TopHat, STAR, Strelka), 컴퓨팅 효율성은 우수한 반면 풍부도가 낮은 전사물에 대한 정확도가 낮을 수 있는 정렬 단계를 포함하지 않는 접근법(예: Sailfish, Salmon)도 선택이 가능합니다.[26] 또한 접근법에 따라 도구가 하나의 유전체의 두 곳 이상의 위치에 모호하게 정렬되는 리드를 매핑하는 방법이나 동형 전사물(transcript isoform)이 존재할 때 풍부도를 정량화하는 방법이 상이할 수 있습니다.

TopHat2 & Cufflinks

TopHat2와 Cufflinks는 전장 시퀀싱 방법을 사용하는 단일세포 실험을 통해 생성된 데이터를 비롯한 NGS 데이터의 유전자 발견과 포괄적인 발현 분석에 사용되는 소프트웨어 도구입니다.[27] TopHat2는 시퀀싱 리드를 정렬하고 splice 부위를 발견하며, Cufflinks는 TopHat2의 리드 정렬 결과를 전사물로 조립합니다. 생물학자는 이 두 앱을 사용하여 새로운 유전자와 알려진 유전자 내의 새로운 splice 변이를 식별하고 여러 조건에서 유전자와 전사물의 발현을 비교할 수 있습니다. 또한 두 앱은 시퀀싱 시스템에서 생성된 파일을 선택하는 시작 단계부터 NGS 데이터를 필터링하고 분석하는 마지막 단계까지 전 과정에 걸쳐 사용자가 쉽게 따라할 수 있는 프롬프트도 제공합니다.

STAR

STAR Aligner는 splice junction(접합 경계) 및 fusion 리드의 검출을 지원하는 고속 RNA-Seq 리드 매핑 도구입니다. 정렬이 진행되는 동안 리드 1개의 여러 부분이 splicing 또는 RNA-fusion에 상응하는 여러 유전체 위치에 매핑됩니다. STAR에는 주석이 달린 유전자 모델로부터 얻은 알려져 있는 splice junction에 대한 정보가 포함되어 있어 접합된 리드를 민감하게 검출할 수 있습니다.

Salmon

Salmon은 유전자 발현을 정량화하는 소프트웨어 도구로, 새로운 알고리즘과 편향(bias) 인식 모델을 결합하여 전사체 전체에 걸쳐 전사물의 풍부도를 정확하게 측정합니다.

Sailfish

Sailfish는 이전에 RNA-Seq 데이터에서 주석이 달렸던 동형 RNA를 정량화하는 데에 사용되는 소프트웨어 프로그램입니다. Sailfish는 리드 매핑 단계를 생략하기 때문에 시퀀싱 데이터 재분석에 소요되는 시간을 크게 줄여 줍니다.

Strelka

Strelka Somatic Variant Caller는 정렬된 시퀀싱 리드를 분석하여 SNV 및 indel을 식별하며 종양/정상 샘플 분석에서 변이를 검출하기에 적합한 도구입니다. Strelka는 순도가 낮은(low-purity) 종양 샘플에서 높은 민감도로 변이를 검출할 수 있도록 해 주는 알고리즘을 사용합니다.

표 8: RNA-Seq 분석 도구

소프트웨어 프로그램	설명	상세 정보
Bowtie2	짧은 리드의 정렬	더 알아보기
TopHat2	RNA-Seq 리드의 정렬, splice 부위의 발견	더 알아보기
TopHat-Fusion	Gene fusion의 발견	더 알아보기
Cufflinks	전사물의 조립	더 알아보기
Cuffcompare	조립물과 참조 유전체 간의 비교	
Cuffmerge	복수 조립물의 결합	
Cuffdiff	차등 발현 분석 수행	
Strelka	작은 변이 검출	더 알아보기
STAR	RNA-Seq 리드의 매핑, splice junction 및 gene fusion의 검출	더 알아보기
Salmon	전사물 발현의 정량화	더 알아보기
Sailfish	동형 RNA의 정량화	더 알아보기

RNA-Seq 분석용 BaseSpace 앱

Illumina는 앞서 기술한 RNA-Seq 도구와 추가적인 도구를 하나의 패키지로 묶어 BaseSpace Sequence Hub에서 실행 가능한 앱으로 제공합니다(표 9). 예를 들어, RNA-Seq Alignment 앱과 RNA-Seq Differential Expression 앱은 리드 정렬, 유전자 및 전사물 풍부도 정량화, SNV 및 small indel 변이 검출, gene fusion 후보 검출 기능과 QC 메트릭스를 제공합니다. 또한 RNA 전사물의 빠르고 정확도 높은 2차 분석을 위해 BaseSpace Sequence Hub 및 로컬 DRAGEN Server에서 DRAGEN 파이프라인을 지원하고 있습니다.

AmpliSeq for Illumina를 통한 표적 RNA-Seq 분석

AmpliSeq for Illumina 패널을 이용하여 생성된 시퀀싱 데이터는 RNA Amplicon BaseSpace 앱으로 분석할 수 있습니다(표 9). RNA Amplicon은 DESeq2를 이용한 간소화된 유전자 발현 분석 기능과 BWA를 활용한 정렬 기능을 제공합니다. 또한 맞춤형 manifest 파일을 불러와 맞춤형 앰플리콘 패널 분석을 수행할 수 있는 기능도 지원합니다. 특정 시퀀싱 시스템에서는 Local Run Manager 소프트웨어를 통해 동일한 분석 워크플로우를 온프레미스 방식으로 이용하는 것도 가능합니다. 연구자는 분석 결과를 저장하거나 다른 연구자와 손쉽게 공유할 수 있습니다. 아울러 작업 효율성을 극대화하고 유전체 데이터에서 생물학적인 정보를 얻는 데 소요되는 시간과 노력을 줄여 주는 BaseSpace Variant Interpreter라는 데이터 해석 및 보고 플랫폼을 사용하여 변이 검출 데이터를 추가적으로 분석해 볼 수도 있습니다(그림 7).

BaseSpace Variant Interpreter에 대한 자세한 정보는 www.illumina.com/products/by-type/informatics-products/basespace-variant-interpreter.html에서 확인하시기 바랍니다.

표 9: RNA-Seq 분석용 BaseSpace 앱

BaseSpace 앱	설명	상세 정보
유전자 발현		
Cufflinks Assembly & DE	RNA-Seq Alignment 앱의 분석 결과를 바탕으로 신속하게 새로운 동형 전사물 및 유전자 발현 수준을 평가하며 Cuffmerge, Cuffquant, Cuffnorm, Cufflinks와 같은 도구를 통해 새로운 전사물 병합과 차등 발현 분석을 수행하는 앱.	더 알아보기
RNA-Seq Alignment	STAR를 이용한 리드 매핑, Salmon을 이용한 참조 유전체 및 전사물 정량화, Strelka를 이용한 변이 검출, Manta를 통한 fusion 검출을 수행하는 앱.	더 알아보기
RNA-Seq Differential Expression	정렬, 유전자 및 전사물 개수 측정, 주석 작업, 변이 검출, fusion 검출, 새로운 전사물 조립을 위해 STAR나 TopHat을 이용한 참조 유전자의 차등 발현 분석 및 전사체 마이닝을 수행하는 앱.	더 알아보기
DESeq2	정렬된 샘플에서 참조 유전자의 차등 발현 분석을 수행하여 유전자 개수, 유전자 FPKM, 주성분 분석, 대조군 대 비교군 결과를 제공하는 앱.	더 알아보기
RNA Express	STAR Aligner 및 DESeq2 분석 도구를 하나의 간단한 워크플로우로 통합하여 가장 흔히 사용되는 몇 가지 RNA 분석 기능을 편리하고 신속한 분석 패키지로 묶어서 제공하는 앱.	더 알아보기
DRAGEN RNA Pipeline	RNA 전사물의 2차 분석을 수행하며, reference-only alignment, annotation-assisted alignment with gene fusion detection 등 복수의 작업 모드를 지원. Gene fusion 모듈은 절단점(breakpoint)일 가능성이 있는 부위를 검출하기 위해 DRAGEN RNA Spliced Aligner를 통해 전체 파이프라인에 최소한의 처리 시간만을 추가하여 보충적인 정렬(chimera read, 키메라 리드)에 대한 스플릿 리드(split-read) 분석을 수행하는 앱.	더 알아보기
DRAGEN Differential Expression	RNA 전사물의 2차 분석을 수행하며 Salmon 정량화 파일에 DESeq2 알고리즘을 실행하여 두 샘플 그룹 간에 차등적으로 발현되는 유전자 및 전사물 데이터를 생성하는 앱.	더 알아보기
RNA Amplicon	정렬, 차등 발현 분석 등 NGS 앰플리콘 패널의 간소화된 유전자 발현 분석을 제공하며 맞춤형 manifest 파일을 불러와 맞춤형 앰플리콘 패널 분석을 수행할 수 있도록 지원하는 앱.	더 알아보기
유전자 조절		
ChIPSeq	ChIP으로 풀다운(pull-down)된 농축(enriched) 영역을 식별하고 해당 영역에서 모티프(motif)를 발견하는 앱. ChIPSeq이 생성하는 raw data는 다운로드 가능하며 대화식 표의 형식으로도 제공 가능.	더 알아보기
MethylSeq	신속하게 전장 유전체 및 표적 바이설파이트 DNA 시퀀스 데이터를 분석하며, 정렬 및 메틸 검출을 수행하고, 정렬 및 메틸화 메트릭스를 계산하는 앱으로, TruSeq DNA Methylation 및 TruSeq Methyl Capture Library Prep Kit로 준비된 라이브러리 지원 가능.	더 알아보기
MethylKit	대용량 처리 바이설파이트 시퀀싱을 통해 얻은 DNA 메틸화 분석과 주석을 제공하는 앱으로, 메틸화 통계와 같은 기초 통계뿐 아니라 두 조건(예: 실험군 대 참조군) 간 차등 메틸화 영역 계산 가능.	더 알아보기
DRAGEN Methylation Pipeline	신속하게 전장 유전체 및 표적 바이설파이트 DNA 시퀀스 데이터를 분석하며, 정렬 및 메틸 검출을 수행하고, 정렬 및 메틸화 메트릭스를 계산하는 앱으로, TruSeq DNA Methylation 및 TruSeq Methyl Capture Library Prep Kit로 준비된 라이브러리 지원 가능.	더 알아보기

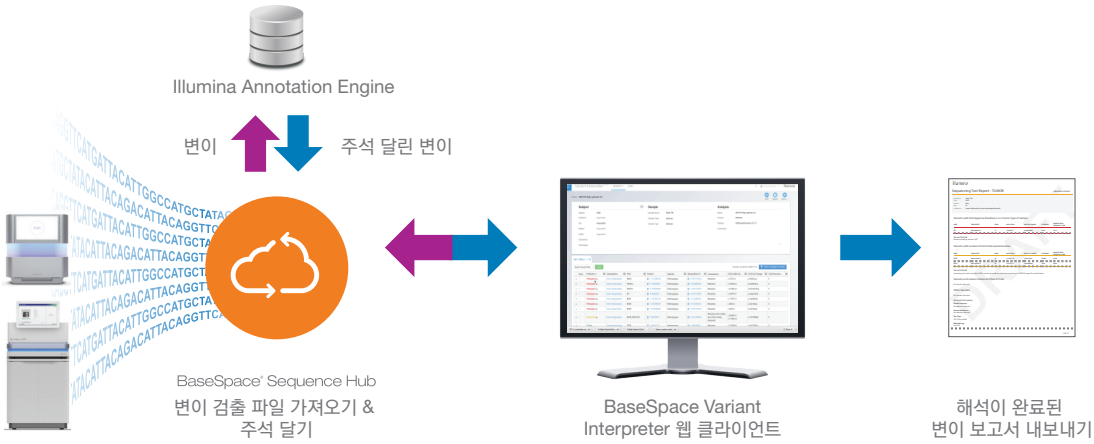


그림 7: BaseSpace Variant Interpreter — 변이 데이터 분석과 해석에 사용되는 고성능 데이터 보고 솔루션으로, 여러 데이터베이스로부터 수집한 정보를 통합하여 주석 작업을 간소화하며 생물학적으로 관련이 있는 변이의 분류와 보고를 위한 유연한 변이 데이터 분석용 필터링 옵션과 도구 제공.

대량 유전자 조절 연구용 바이오인포매틱스 솔루션

유전자 조절은 특정 유전자가 생물학적 과정에서 활성화되거나 비활성화되는 과정을 말합니다. 유전자 조절 연구에서 유전적 특성을 파악하려면 전체 유전체에 대한 메틸화 패턴의 탐지 및 특징 파악, DNA와 단백질 간의 상호작용 확인, 열린 크로마틴 영역의 평가 등이 필요합니다. 메틸화 분석의 결과 파일에는 메틸화 통계 플롯, 메틸화 상관관계 플롯, 차등 메틸화 영역 및 요약표, 메틸화 통계 요약이 포함됩니다. DNA-단백질 분석 결과에는 주석, 피크(peak), Interactive Annotated Peak/Motif Explorer로 시각화할 수 있는 모티프 파일, 그리고 정렬 파일(BAM 형식)이 포함됩니다. ATAC-Seq을 통한 열린 크로마틴 분석 결과에는 피크 검출(peak calling), 피크 주석, 피크 차등 분석, 뉴클레오솜 위치 결정(nucleosome positioning) 등이 포함됩니다.[28]

메틸레이션 시퀀싱

메틸레이션 시퀀싱 데이터는 MethylSeq 앱과 MethylKit 앱을 포함한 다양한 BaseSpace 앱을 이용해 분석할 수 있습니다 (표 9). MethylSeq 앱은 Bismark[29] 및 Bowtie2[30]를 사용하여 바이설파이트로 처리한 시퀀싱 리드를 원하는 유전체에 매핑하며 메틸화 검출을 수행합니다. MethylKit 앱은 샘플 간 존재하는 메틸화의 차이를 확인하기 위해 시퀀싱 데이터를 분석합니다. BaseSpace MethylSeq 앱과 MethylKit 앱은 표적 데이터를 지원하며 메틸화 검출, 샘플 간 차등 메틸화 분석, 메틸화 정도가 현저하게 높은 영역의 범주화 등 일반적인 후성유전체 분석 작업을 실행합니다. 또한 BaseSpace MethylSeq 앱과 MethylKit 앱은 연구자라면 누구나 바이오인포매틱스 경험 유무에 관계없이 버튼만 누르면 쉽게 사용이 가능하도록 설계되어 있습니다.

ChIP-Seq

ChIP-Seq 데이터는 BaseSpace ChIPSeq 앱을 통해 분석할 수 있습니다. ChIPSeq 앱은 Model-based Analysis of ChIP-Seq(MACS)[31] 방법으로 크로마틴 면역침강을 통해 풀다운된 농축된 영역을 찾아내고 HOMER[32]를 이용해 해당 영역 내 모티프를 발견합니다. 이들 도구가 생성하는 raw data는 다운로드가 가능하며 대화식 표(interactive table)의 형식으로도 제공됩니다. 샘플 그룹별로 분석이 따로 수행됩니다. 각 샘플 그룹은 처리 샘플(treatment sample, 즉 풀다운이 수행된 샘플)과 대조(control) 샘플로 구성되며, 두 샘플 모두 반복실험(replicate)에 따라 복수의 FASTQ 파일에 걸쳐 나뉘어질 수 있습니다.

ATAC-Seq

처음 처리 및 정렬 단계가 완료되면 다양한 상용 알고리즘과 및 오픈소스 알고리즘 그리고 ENCODE ATAC-Seq, chromVAR, Partek과 같은 소프트웨어 프로그램을 사용하여 ATAC-Seq 데이터를 분석할 수 있습니다(표 10). 이들 도구는 주로 피크 검출 단계에서 실행되어 접근 가능한 크로마틴 영역을 찾아 냅니다. ATAC-Seq 분석 시 피크 검출 작업에는 ChIP-Seq 분석을 위해 전사 인자 결합 부위의 식별을 목적으로 설계된 MACS를 사용할 수 있습니다. 결과 파일로는 검출된 피크의 유전체 좌표와 관련 통계(배수 변화, p-값, q-값 등)를 포함하는 평문 브라우저 확장성 데이터(browser extensible data, BED) 파일이 있습니다.

ATAC-Seq 데이터 분석 파이프라인에 대한 자세한 정보는 yiweiniu.github.io/blog/2019/03/ATAC-seq-data-analysis-from-FASTQ-to-peaks/에서 확인하시기 바랍니다.

표 10: ATAC-Seq 분석 도구

소프트웨어 프로그램	설명	상세 정보
ENCODE ATAC-Seq pipeline	FASTQ 파일에서 피크 검출에 이르기까지, 자동화된 ATAC-Seq 데이터 분석을 위해 설계된 오픈소스 파이프라인.	더 알아보기
chromVAR	관련이 있는 모티프나 유전체 주석의 식별을 위해 크로마틴 접근성의 변화를 분석할 때 사용하는 오픈소스 R 패키지.	더 알아보기
Partek	ATAC- Seq 데이터 분석에 사용되는 상용 통계 분석 소프트웨어로, 지침이 제공되는 워크플로우와 사용이 용이한 인터페이스 제공.	더 알아보기

요약

유전자 발현 및 조절 NGS 실험에서 생물학적 의미를 알아내려면 사용이 용이한 바이오인포매틱스 도구가 매우 중요합니다. 오늘날 이러한 도구는 손쉽게 구할 수 있으며, 바이오인포매틱스 기술은 관련 데이터를 제공하고 다양한 연구 환경에서 수많은 유전자 발현 및 조절 관련 질문에 대한 해답을 찾는 데에 도움이 될 수 있습니다. 여기에 Illumina의 NGS 시스템을 더하면 연구자는 원하는 수준의 세포 해상도로 유전자 발현 및 조절 정보를 탐색할 수 있습니다.

제4장: 단일세포 분석용 바이오인포매틱스 파이프라인

일반적으로 scRNA-Seq, scATAC-Seq, 단일세포 단백체학(single-cell proteomics) 등에 적용되는 단일세포 시퀀싱 데이터 분석 워크플로우는 각각의 상보적인(complementary) 대량 샘플 시퀀싱 프로토콜과 유사합니다. 그러나 대량 샘플 시퀀싱 데이터 분석의 디자인을 뒷받침하는 근거는 종종 단일세포 데이터에는 유효하지 않기 때문에 단일세포 시퀀싱 데이터의 분석에는 특화된 바이오인포매틱스 솔루션과 새로운 방법이 필요합니다.[33-35]

scRNA-Seq 데이터 분석의 어려움

scRNA-Seq 데이터 분석 시 연구자는 대량 샘플 시퀀싱의 데이터 분석 때와는 다르게 몇 가지 어려움에 맞닥뜨리게 됩니다. 대용량 단일세포 분리에 이용되는 미세유체 기술의 발전으로 분석이 가능한 세포의 개수가 증가함에 따라 규모 조정이 가능한 분석 방법을 요구하는 데이터 포인트의 수도 덩달아 크게 증가했습니다. scRNA-Seq 데이터는 증가된 생물학적 복잡성(biological complexity)을 반영하므로 대량 샘플 RNA-Seq과 비교했을 때 가변성(variability)이 높다는 특징이 있습니다. scRNA-Seq 데이터 분석을 더욱 어렵게 만드는 것은 특정 세포에 UMI나 시퀀싱 리드가 매핑되지 않아 발현된 유전자가 없는 상황(observed zero), 즉 드롭아웃(dropout)입니다. 이러한 현상은 두 가지 원인으로 인해 발생하는데, 기술적인 문제로 인해 유전자가 발현되지만 검출은 되지 않는 경우가 있고, 특정 세포에서 유전자가 실제로 발현되지 않는 경우가 있습니다. 게다가 유전자 발현은 일시적으로 나타나므로(세포는 일정 시간 동안 또는 특정 조건 하에만 활발한 전사 활동을 보일 수 있기 때문에) 데이터 분석은 더욱 힘들어질 수 있습니다. scRNA-Seq 데이터의 1차, 2차, 3차 분석에는 이를 포함한 여러 요인들을 반드시 고려해야 합니다.[36-41]

단일세포 시퀀싱 데이터의 1차 및 2차 분석

단일세포 시퀀싱 데이터의 1차 분석에는 대량 샘플 시퀀싱의 1차 분석과 동일한 방법이 적용되며, 주로 BCL 형식에서 FASTQ 형식으로 파일을 변환하는 작업이 진행됩니다. 파일 변환 후 2차 분석을 수행할 수 있습니다.

디멀티플렉싱

디멀티플렉싱은 단일세포 시퀀싱 데이터 분석에서 중요한 단계입니다. 대량 샘플 시퀀싱에서는 디멀티플렉싱을 통해 풀링된 라이브러리의 리드를 개별 라이브러리로 분리하는 반면, 단일세포 시퀀싱에서는 세포 분리 단계에서 추가되었던 세포 바코드를 기준으로 풀링된 샘플의 리드를 개별 세포로 분리합니다(그림 8). 대부분의 단일세포 분석 플랫폼이 디멀티플렉싱 기능을 제공하지만, 필요하다면 zUMIs와 같은 오픈소스 파이프라인을 별도로 사용해 볼 수도 있습니다.

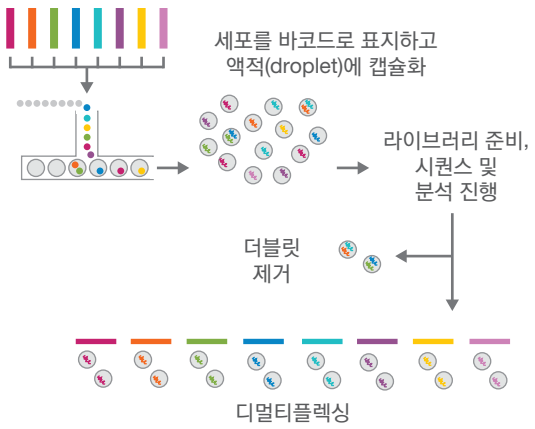


그림 8: 단일세포 시퀀싱의 디멀티플렉싱 과정 — 세포 분리 단계에서 추가되었던 세포 바코드를 이용해 2차 분석에서 시퀀싱 리드를 개별 세포로 파싱(parsing)하고 세포 더블릿(doublet) 제거.

QC 메트릭스

후속 분석 진행에 앞서 단일세포 시퀀싱 데이터 세트의 품질을 확인하기 위해 몇 가지 QC 메트릭스를 활용해 볼 수 있습니다. 일반적으로 사용되는 QC 메트릭스로는 추정된 세포 수, 유전자 간/인트론/엑손 함량, 세포 내 리드의 분율, 예상 라이브러리 크기, 발현된 유전자 개수가 있습니다. 이러한 메트릭스는 형광 정량화나 qPCR과 같은 종래의 정량화 방법으로는 평가할 수 없습니다. 따라서 세포 지도(cell atlas) 연구와 같은 대용량 단일세포 실험의 효율성을 극대화하거나 여러 단일세포 라이브러리를 합쳐야 할 경우, 높은 뎁스로 NGS 런을 수행하기 전에 먼저 iSeq 100 시스템에서 낮은 뎁스로 시퀀싱을 수행하여 주요 메트릭스의 특성을 파악한 후 양적 균형을 재조정(rebalancing)해 볼 수 있습니다. 또 라이브러리 QC를 통해 더 일관성 있는 결과를 얻고 데이터 분석 및 해석 단계도 간소화할 수 있습니다.

자세한 내용은 *QC and rebalancing of single-cell gene expression libraries using the iSeq 100 System* Application Note를 통해 확인하실 수 있습니다.

단일세포 시퀀싱 데이터의 3차 분석

일반적으로 단일세포 시퀀싱 데이터의 분석 파이프라인은 리드 정렬, 특징-바코드 매트릭스(feature-barcode matrix) 생성 그리고 3차 분석에 사용됩니다. 차원 축소(dimensionality reduction)는 데이터 포인트의 클러스터링에 필요하며 scRNA-Seq, sc-ATAC-Seq, 단일세포 단백질 프로파일링 분석에도 중요한 단계입니다.

차원 축소 알고리즘

수백에서 수천 개의 세포에 걸쳐 수천 개의 데이터 포인트를 동시에 분석할 수 있다면 고차원성(high dimensionality)을 가진 단일세포 시퀀싱 데이터를 얻을 수 있습니다.[39] 일반적으로 고차원 데이터 세트는 후속 분석의 계산량 감축, 데이터 노이즈 감소 및 데이터 시각화를 위해 데이터 세트의 근본적인 구조를 2차원 또는 3차원으로 캡처하는 차원 축소 과정을 거치게 됩니다.[43] 차원 축소에 이용할 수 있는 다음과 같은 몇 가지 알고리즘이 마련되어 있습니다.

PCA

PCA(principal component analysis, 주성분 분석)은 선형(linear) 차원 축소에 흔히 사용되는 알고리즘으로, 최대 분산(maximal variance) 캡처를 위해 데이터를 더욱 적은 수의 독립적인 차원으로 투영합니다(그림 9A). PCA는 데이터가 대략 정규적으로 분포되었다고 가정하는데, 이러한 가정은 단일세포 시퀀싱 데이터에 항상 적용되지는 않습니다.[39]

t-SNE

t-SNE(t-distributed stochastic neighborhood embedding, t-분포 확률적 이웃 임베딩)은 비선형(non-linear) 차원 축소 기법입니다. t-SNE는 수학적으로 차원의 수를 2차원 또는 3차원으로 축소 표현함으로써 다차원 데이터의 이해를 돕고 흔히 단일세포 시퀀싱 데이터로 아집단(subpopulation)을 시각화하는 데 활용됩니다. 그러나 t-SNE 알고리즘은 계산에 오랜 시간이 필요하며, 대용량 데이터 세트를 표현하는 능력에 한계가 있고, 전체 구조(global structure)를 보존하지 않습니다. 이는 곧 클러스터 내 데이터 포인트 간 거리는 유의미하고 유용한 정보를 제공하지만, 클러스터 간 거리는 그러한 정보를 제공하지 못한다는 것을 의미합니다(그림 9B).[43]

UMAP

UMAP(uniform manifold approximation and projection, 균일 매니폴드 근사 및 추정)은 생물학적 데이터 세트와 단일세포 시퀀싱 데이터 세트에 적용 가능한 모든 종류의 고차원 데이터에 사용되는 비선형 차원 축소 기법입니다. UMAP은 비교적 적은 계산 시간으로 대량의 데이터 세트 내에서 지역 구조(local structure)와 전체 구조를 보존합니다.[44]

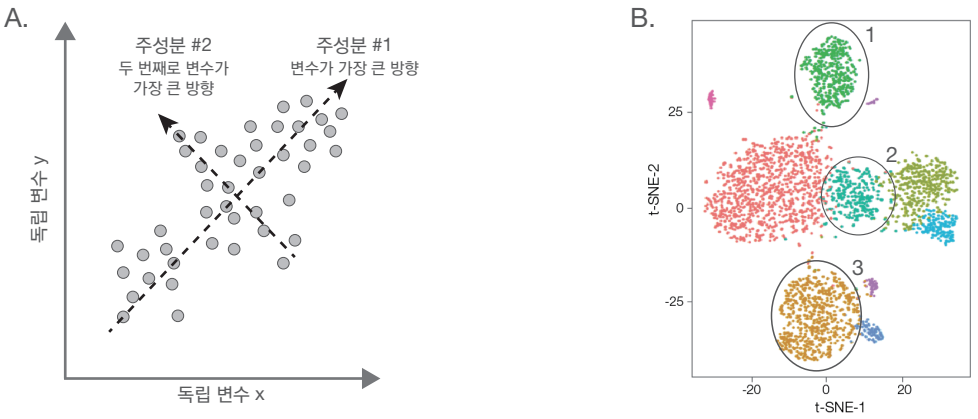


그림 9: PCA 및 t-SNE의 클러스터링 과정 — (A) PCA는 최대 분산 캡처를 위해 고차원 데이터를 더 적은 수의 차원으로 투영하지만 축소 대상이 선형 차원으로 한정되며 데이터가 정규적으로 분포되었다고 가정함. (B) t-SNE는 차원을 축소하여 2차원 또는 3차원 플롯에 데이터 포인트를 표현함. 참고: 클러스터들은 각 클러스터 내 데이터 포인트 간의 유사성을 표시하지만, 클러스터 간 거리는 유사성을 의미하지 않음(즉, 클러스터 1과 2가 반드시 클러스터 1과 3보다 유사성이 더 높지는 않음).

단일세포 시퀀싱 데이터의 후속(3차) 분석

추가적인 후속 분석 단계는 시퀀싱 방법에 따라 결정됩니다. 예를 들어, scRNA-Seq 분석에는 차등 유전자 발현 분석이 포함됩니다.[42] scATAC-Seq을 활용한 후성유전학적 연구는 scRNA-Seq을 이용한 유전자 발현 연구를 보완해 주고 유전자 조절에 관한 정보를 제공할 수 있습니다.[47] scATAC-Seq 실험의 경우 데이터 세트가 희소(sparse)하며 양분법적(binary)이고 크기가 크기 때문에 분석이 쉽지 않습니다. 하지만 크로마틴 접근성 데이터로부터 생물학적 정보를 얻을 수 있도록 해 주는 몇 가지 분석 솔루션이 있습니다.[28]

NGS를 활용한 단일세포 단백체 분석으로 단일세포 내 단백질 발현을 판독할 때는 AbSeq[4], CITE-Seq[5], REAP-Seq[6] 이렇게 세 가지 접근 방식이 주로 사용됩니다. 이러한 접근 방식을 통해 얻어지는 데이터 세트는 멀티모달(multimodal) 데이터이기 때문에 분석이 쉽지는 않지만, 전사물의 정량 시 UMI의 개수를 세는 것과 거의 동일한 방식으로 항체 태그(antibody-tag, ADT) 시퀀스를 정량화할 수 있습니다. 이러한 데이터 세트를 더욱 편리하게 분석하려면 기존의 scRNA-Seq 분석 도구가 멀티모달 데이터를 지원할 수 있도록 업데이트를 실행하거나 새로 개발된 도구를 사용해 보는 것이 좋습니다.

현재 연구자들은 학술 검사실에서 R과 Python 같은 대중적인 프로그래밍 언어를 사용해 개발한 오픈소스 분석 도구, 미리 구성되어 있는 분석 워크플로우를 지원하는 'plug-and-play' 패키지, 그리고 상용 제품 등 다양한 도구를 단일세포 시퀀싱 데이터의 3차 분석에 활용해 볼 수 있습니다. 분석 도구는 연구 목표와 실험 목적에 따라 선택하면 됩니다.

오픈소스 바이오인포매틱스 도구(프리웨어)

Seurat

Seurat는 정규화(normalization), 차원 축소 접근법, 그래프, 히트맵, 데이터 통합 도구를 통해 세포 간 이질성을 평가하도록 설계된 R 기반의 단일세포 RNA-Seq 분석 소프트웨어입니다.[35] Seurat는 차원 축소 기능을 이용하여 세포를 독특한 특징을 지닌 세포의 상태 또는 유형에 해당하는 그룹으로 군집화합니다.

Seurat 사용자는 다양한 조건, 기술 또는 종에 걸쳐 생성된 단일세포 데이터 세트를 통합할 수 있습니다. 또한 해당 앱으로 멀티모달 데이터(예: scRNA-Seq 데이터와 scATAC-Seq 및 단일세포 단백체학 분석 데이터)도 탐색할 수도 있습니다.

자세한 정보는 satijalab.org/seurat/v3.0/integration.html에서 확인하시기 바랍니다.

Monocle

Monocle은 세포의 발달적 궤적(cell developmental trajectory)을 확인하기 위한 목적으로 개발된 R 기반의 scRNA-Seq 분석 소프트웨어입니다. Monocle은 세포의 초기 상태와 최종 상태가 알려져 있는 실험에 적합합니다. 이 소프트웨어는 머신 러닝 기법을 사용하여 분화 패턴에 따라 각 세포를 정렬(ordering)하고, t-SNE로 세포의 클러스터링을 실행하며, 차등 유전자 발현 분석을 수행합니다.

자세한 정보는 cole-trapnell-lab.github.io/monocle-release/에서 확인하시기 바랍니다.

www.illumina.com/science/customer-stories/icommunity-customer-interviews-case-studies/trapnell-uwash-interview-single-cell.html에서 Cole Trapnell의 인터뷰를 통해 단일세포 분석 기술의 발전에 대해 알아보시기 바랍니다.

velocyto

RNA 속도(velocity)는 스플라이싱된 mRNA와 스플라이싱되지 않은 mRNA의 비를 기준으로 특정 시점에서 특정 유전자에 대한 유전자 발현 변화 속도를 나타냅니다. scRNA-Seq은 단일세포 내에서의 RNA 속도를 추정하여 몇 시간 내지 며칠에 걸쳐 진행되는 분화의 궤적을 예측하고 세포 계통(lineage)을 분석할 수 있습니다. velocyto 소프트웨어 패키지를 사용하면 RNA 속도를 추정하기 위해 scRNA-Seq 데이터를 바탕으로 유전자 발현의 역학을 분석할 수 있습니다.[45,46]

자세한 정보는 velocyto.org에서 확인하시기 바랍니다.

chromVAR

chromVAR는 관련 모티프나 유전체 주석을 파악하기 위해 대량 샘플 시퀀싱 데이터 또는 scATAC-Seq 데이터에서 크로마틴 접근성의 차이를 분석하는 데 사용되는 오픈소스 R 패키지입니다. 이 소프트웨어를 이용하면 크로마틴 접근성과 관련이 있는 알려진 시퀀스 모티프 및 *de novo* 시퀀스 모티프를 식별할 수 있습니다.[48]

자세한 정보는 github.com/GreenleafLab/chromVAR에서 확인하시기 바랍니다.

Human Cell Atlas Data Coordination Platform

Human Cell Atlas Data Coordination Platform(HCA DCP)은 데이터 제출 접수 서비스, 여러 클라우드의 동기화된 데이터 보관, 표준화된 2차 분석 파이프라인, 데이터 접근, 그리고 3차 분석 및 시각화를 위한 포털 이렇게 4가지 주요 기능을 제공하기 위해 개발된 데이터 조정용 오픈소스 소프트웨어입니다. 검사실과 연구자는 전 세계 어느 곳에서나 단일세포 시퀀싱 데이터를 HCA DCP에 제출할 수 있습니다.

자세한 정보는 data.humancellatlas.org에서 확인하시기 바랍니다.

CITE-Seq-Count & CITEFuse

데이터 분석 및 시각화의 지원을 위해 CITE-seq-Count와 CiteFuse[49] 같은 바이오인포매틱스 도구를 다양한 단일세포 단백질 프로파일링 방법에 활용해 볼 수 있습니다.

자세한 정보는 cite-seq.com에서 확인하시기 바랍니다.

플랫폼별 상용 바이오인포매틱스 도구

상용 서비스 제공 업체가 개발한 다양한 소프트웨어 도구는 특정 단일세포 분리 및 분석 플랫폼(예: 10x Genomics, MissionBio)을 지원하고 있습니다(표 11).

표 11: 특정 플랫폼을 지원하는 단일세포 시퀀싱 데이터 분석 소프트웨어

	설명	시퀀싱 데이터 유형	상세 정보
10x Loupe Browser	10x Chromium scRNA-Seq 데이터, scATAC-Seq 데이터 등 단일세포 시퀀싱 데이터의 시각화 및 분석 기능을 제공하는 데스크톱 애플리케이션. 이 소프트웨어로 scRNA-Seq 데이터 내의 유의미한 유전자, 세포 유형 및 하부 구조를 대화식으로 빠르게 탐색 가능. scATAC-Seq 데이터 분석을 통해 차등 크로마틴 접근성 확인, 유의한 조절 특징(regulatory feature) 탐색, 전사 인자 모티프 구별 등 다양한 작업 수행 가능.	scRNA-Seq sc-ATAC-Seq	더 알아보기
10x Cell Ranger ATAC	10x Chromium Single Cell ATAC 데이터의 처리를 위해 개발된 다양한 분석 파이프라인을 제공하는 소프트웨어. 파일 변환, 디멀티플렉싱, 리드 필터링 및 정렬 등의 1차 분석 수행. 트랜스포사제 절단 부위의 식별, 접근 가능한 크로마틴 피크의 검출, 세포 검출, 피크 및 전사 인자에 대한 카운트 매트릭스 생성 등의 2차 분석 수행. 차원 축소, 세포 클러스터링, 차등 접근성 영역의 클러스터링 등의 3차 분석 수행.	sc-ATAC-Seq	더 알아보기
Tapestri Insight	Tapestri 단일세포 DNA 시퀀싱 데이터 분석에 사용되는 소프트웨어 솔루션으로, 시퀀스 불러오기, 데이터 분석 및 시각화 기능을 포함하며, 클론 및 서브클론 수준에서 SNV 및 유전자 복제수 변이(copy number variation, CNV) 등의 변이를 식별 가능.	single-cell DNA sequencing	더 알아보기
Bio-Rad SureCell ATAC-Seq Analysis Toolkit	SureCell ATAC-Seq Library Prep Kit와 함께 사용 시 유전체 전체에 대한 단일세포의 후성유전학적 분석 기능 제공. 이 소프트웨어로 피크 내 크로마틴 접근성의 획득 및 손실(gain and loss) 추정, scATAC-Seq 프로필의 클러스터링, 유전자 발현과 관련 시퀀스 모티프의 특성 규명, 다양한 세포 표현형의 출처인 시스-작용 요소 및 트랜스-작용 요소 식별 가능.	sc-ATAC-Seq	더 알아보기

플랫폼 중립적인 상용 바이오인포매틱스 도구

SeqGeq™ v1.6 Software

FlowJo 소프트웨어의 제조사에서 개발한 SeqGeq 소프트웨어는 단일세포 실험 분석을 목적으로 개발되었으며 플랫폼에 구애받지 않는 데스크톱 바이오인포매틱스 플랫폼입니다. SeqGeq은 V(D)J 분석, Seurat 클러스터링, Monocle 궤적 추론 등 다양한 인포매틱스 기능을 제공합니다. 이들 도구는 모두 사용된 기기나 시퀀싱 파이프라인에 상관없이 어떠한 데이터와도 호환이 가능하도록 설계되었습니다.

SeqGeq은 FlowJo 사용자에게 친숙한 사용하기 쉬운 드래그 앤 드롭 인터페이스를 통해 고급 분석, 데이터 탐색 및 시각화 작업을 수행할 수 있도록 해 줍니다. SeqGeq은 출판 및 협업을 위해 손쉽게 공유할 수 있는 고품질의 그림을 생성하며 (그림 10) BaseSpace Sequence Hub에 통합되어 있어 데이터 분석 워크플로우를 바로 완료할 수 있습니다.

자세한 정보는 www.flowjo.com/solutions/seqgeq에서 확인하시기 바랍니다.

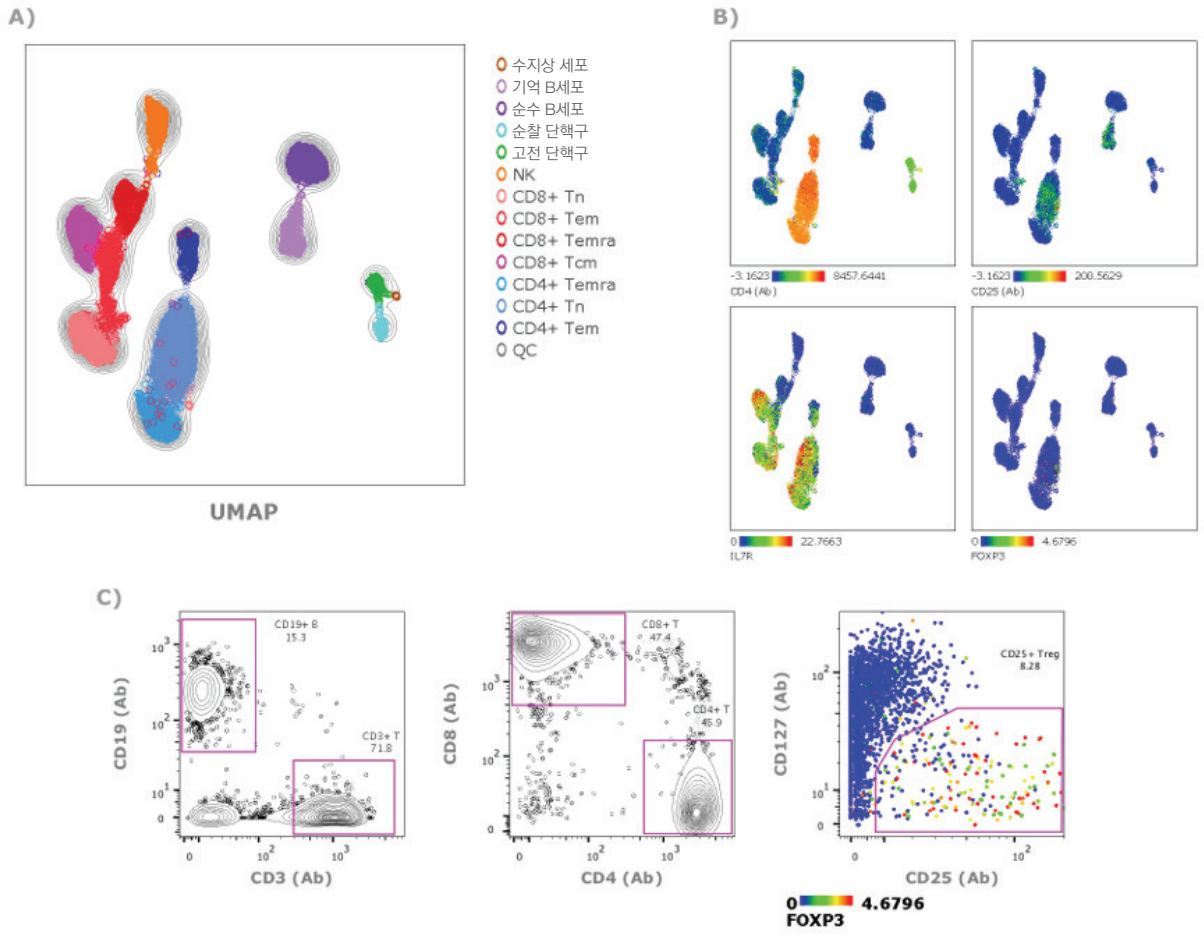


그림 10: SeqGeq의 표준 말초혈액 샘플 분석 결과 — (A) 사용이 용이한 GUI를 통해 생성된 차원 축소, 클러스터링 및 주석 결과, (B) 유전자 및 항체 발현을 표시한 세포의 히트맵, (C) 수동 게이팅(gating) 분석을 통해 관심 있는 Treg 집단으로 이어지는 전통적 게이팅 구조. BD Biosciences에서 SeqGeq Software를 사용하여 생성 및 제공된 데이터.

Partek Flow

견고한 통계 및 대화식 시각화 도구를 제공하는 Partek Flow는 연구자가 고급 바이오인포매틱스 기술 없이도 대량 샘플 및 단일세포 유전체 분석 데이터에서 빠르고 안정적으로 생물학적 정보를 발견할 수 있도록 해 줍니다. 전 세계 많은 연구자들이 데이터 분석에 사용해 온 Partek은 동료 평가(peer-review)를 마친 8천 건이 넘는 출판물에 인용되기도 하였습니다. Partek은 scRNA-Seq, CITE-Seq, 세포 해싱(hashing), 공간 전사체학(spatial transcriptomics) 등 다수의 단일세포 애플리케이션과 호환 가능합니다(그림 11). Partek은 다음과 같은 분석 기능을 제공합니다.

- 세포를 알려진 세포 유형과 새로운 세포 유형으로 구분
- 세포 집단을 정의하는 바이오마커(biomarker) 발견
- 차등적으로 발현된 유전자, 단백질, 경로 찾기
- 실험군 간/표현형 간 세포 유형 집단 비교
- 멀티오믹스 실험을 통해 얻은 유전자 및 단백질 발현 데이터 통합

자세한 정보는 www.partek.com/partek-flow/에서 확인하시기 바랍니다.

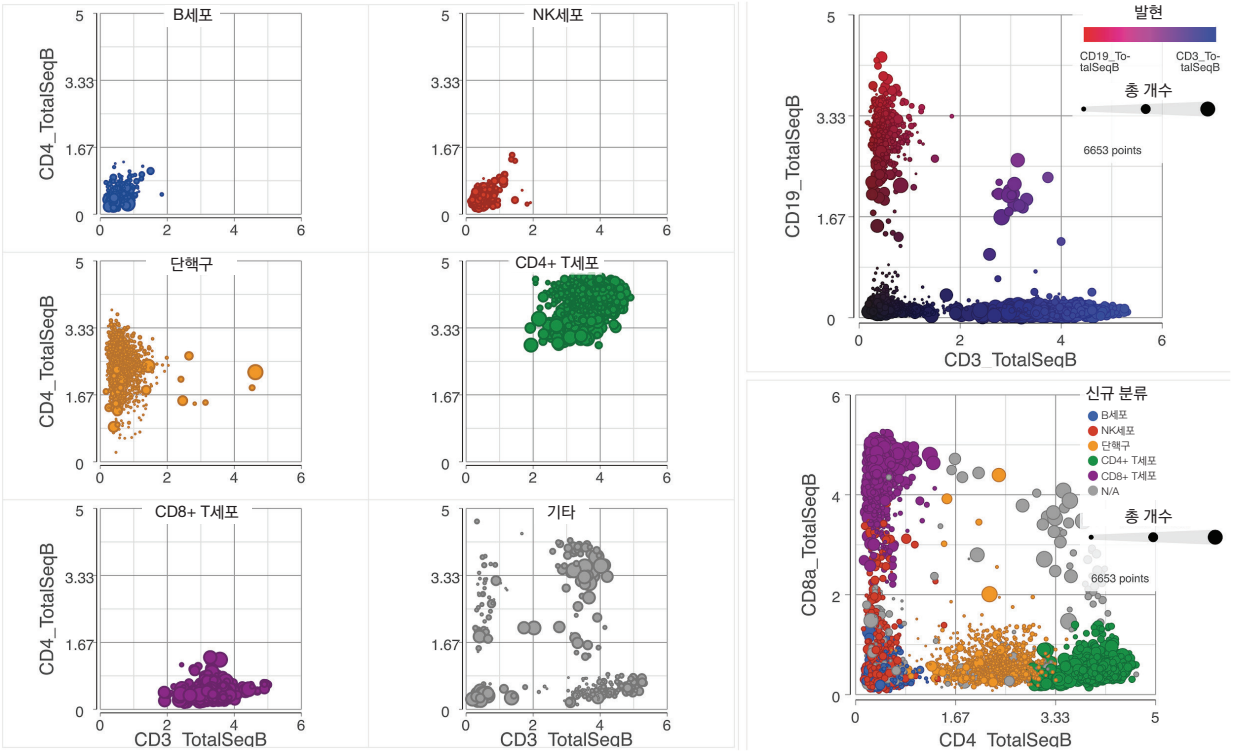


그림 11: Partek Flow의 단일세포 샘플 분석 기능 — Partek Flow는 강력한 통계 알고리즘, 풍부한 정보를 포함하는 시각화 기능, 최첨단 유전체학 도구를 제공하며, BCL, FASTQ, BAM, FCS, H5, TXT, CSV 또는 CBCL 형식의 파일로 분석을 시작할 수 있어 풍부한 대화식 시각화 기능을 지원함. Partek에서 Partek Flow 소프트웨어를 사용하여 생성 및 제공된 데이터.

요약

대량 샘플 시퀀싱 연구와는 달리, 단일세포 시퀀싱 연구에는 새로운 데이터 분석 방법이 필요합니다. 다양한 오픈소스 및 상용 바이오인포매틱스 도구를 활용하면 scRNA-Seq, scATAC-Seq 및 단백질 프로파일링 연구를 위한 데이터 분석, 시각화 및 해석이 가능합니다. 세포 및 분자생물학적 복잡성을 이해하기 위해서는 이러한 방법을 통해 유의미한 정보를 도출하는 능력이 중요합니다.

맺음말

대량 샘플 또는 단일세포 샘플을 사용하는 NGS 기반의 유전자 발현 및 조절 연구는 복잡한 생물학적 시스템에 대한 정보를 제공할 수 있는 잠재력을 지니고 있습니다. 유전자 발현 및 조절의 조사 옵션이 늘어남에 따라 강점과 약점을 내재하고 있는 바이오인포매틱스 파이프라인 또한 매우 다양해졌습니다. 신중한 파이프라인 디자인과 분석 워크플로우 전체에 대한 최적화가 연구의 성패를 좌우한다 해도 과언이 아닙니다.

이 문서에서는 대량 샘플 시퀀싱 데이터와 단일세포 시퀀싱 데이터의 분석에 이용할 수 있는 전산 분석 파이프라인에 중점을 두고 NGS 워크플로우를 단계별로 설명하였습니다. 또한 주요 고려 사항과 잠재적인 어려움에 대해 논의하고, 상용 제품을 소개하였으며, 성공적인 NGS 기반 유전자 발현 및 조절 연구를 위해 분석 워크플로우를 수행 시 도움이 될 만한 내용을 요약해 드렸습니다.

참고 문헌

1. Ozsolak F, Milos PM. RNA Sequencing: advances, challenges and opportunities. *Nat Rev Genet*. 2011;12:87–98.
2. Wang W, Niu Z, Wang Y, et al. Comparative transcriptome analysis of atrial septal defect identifies dysregulated genes during heart septum morphogenesis. *Gene*. 2016;575:303–312.
3. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29–39.
4. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci Reports*. 2017;7:44447.
5. Stoeckius M, Hafemeister C, Stephenson W, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14:865–868.
6. Peterson VM, Zhang KX, Kumar N, et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotech*. 2017;35:936–939.
7. Illumina (2018). iSeq 100 Sequencing System specification sheet. Accessed July 16, 2020.
8. Illumina (2019). MiniSeq Sequencing System specification sheet. Accessed July 16, 2020.
9. Illumina (2018). MiSeq System specification sheet. Accessed July 16, 2020.
10. Illumina (2019). NextSeq 550 Sequencing System specification sheet. Accessed July 16, 2020.
11. Illumina (2020). NextSeq 1000 and NextSeq 2000 Sequencing Systems specification sheet. Accessed July 16, 2020.
12. Illumina (2019). NovaSeq 6000 Sequencing System specification sheet. Accessed July 16, 2020.
13. Sims D, Sudbery I, Ilot NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121–132.
14. Corley SM, MacKenzie KL, Beverdam A, Roddam LF, Wilkins MR. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC Genomics*. 2017;18:399.
15. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*. 2013;8(10):e77910.
16. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
18. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
19. Agilent Technologies. (2016). RNA Integrity Number (RIN) – Standardization of RNA Quality Control Publication. PN-5989-1165EN. Accessed August 4, 2020.
20. Wong KS, Pang H. Simplifying HT RNA Quality & Quantity Analysis. *Genet Eng & Biotech News*. 2013;33(2):4688.
21. Sheng Q, Vickers K, Wang J, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. *Brief Func Genomics*. 2017;16(4):194–204.
22. Guo Y, Long J, He J, et al. Exome sequencing generates high quality data in non-target regions. *BMC Genomics*. 2012;13:194.
23. Illumina (2020). BaseSpace Sequence Hub data sheet. Accessed August 4, 2020.
24. Illumina (2019). Illumina DRAGEN Bio-IT Platform data sheet. Accessed August 4, 2020.
25. Illumina (2018). BaseSpace Correlation Engine data sheet. Accessed August 4, 2020.
26. Sahraeian SME, Mohiyuddin M, Sebra R, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun*. 2017;8(1):59.
27. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562–578.
28. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol*. 2020;21(1):22.
29. Krueger F and Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011; 27:1571–1572.
30. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–359.
31. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
32. Heinz S, Benner C, Spann N, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* 2010 May 28;38(4):576-589.
33. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50(8):96.
34. Eberwine J, Sul JY, Bartfai, T, Kim J. The promise of single-cell sequencing. *Nat Methods*. 2014;11(1):25–27.
35. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. 2017;9(1):75.
36. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019;15(6):e8746.
37. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21(31):doi.org/10.1186/s13059-020-1926-6.
38. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131(4):281–285.
39. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet*. 2019;10:317.
40. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32(9):896–902.
41. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381–386.
42. Gao M, Ling M, Tang X, et al. Comparison of High-Throughput Single-Cell RNA Sequencing Data Processing Pipelines. *bioRxiv*. 2020;02.09.940221.
43. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. *Mol Aspects Med*. 2018;59:114–122.
44. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2019;37:38–44.
45. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. *Nature*. 2018;60(7719):494–498.
46. Bergen V, Lange M, Peidli S, Wolf A, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020; doi: 10.1038/s41587-020-0591-3.
47. Buenrostro J, Wu B, Litzenburger U, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015;523(7561)486–490.
48. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods*. 2017;14:975–978.
49. Kim HJ, Lin Y, Geddes TA, Hwa Yang JY, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*. 2020;36(14):4137–4143.

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

연구 전용입니다. 진단 절차에는 사용할 수 없습니다.

illumina®