KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (오프라인)

Genome assembly tutorial

김준 _ 충남대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 : 2월 24일 (토)

시간	강 의 (자연과학대학 28동 101호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:30	**의료빅데이터/인공지능 총론** 김헌성 교수(가톨릭대학교)
14:30-14:45	휴식
14:45-16:15	**의료영상 인공지능의 이해 및 의료영상 레이블링 실습** 백서연 교수(연석대학교)
16:15-16:30	휴식
16:30-18:00	**의료 정보처리 자동화 실습 / 독자적인 어플리케이션 만들기** 김선근 대표(원닥 주식회사), 서사도 조교

시간	강 의 (자연과학대학 28동 102호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:20	**EMR 데이터를 활용한 머신러닝 기반 예후예측:** **Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)** 고태훈 교수(가톨릭대학교)
14:20-14:40	휴식
14:40-16:00	**Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단:** **Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)** 고태훈 교수(가톨릭대학교)
16:00-16:20	휴식
16:20-17:40	**심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network +** **Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)** 고태훈 교수(가톨릭대학교)

DAY1 : 2월 26일 (월)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**DNN (이론)** 이상근 교수(고려대학교)
10:50-11:00	휴식
11:00-12:10	**CNN (이론)** 이상근 교수(고려대학교)
12:10-13:40	점심
13:40-15:10	**RNN, ChatGPT, XAI (이론)** 이상근 교수(고려대학교)
15:10-15:20	휴식
15:20-16:50	**CNN/RNN 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)** 이정현 조교, 한성민 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Best practice for single-cell data analysis** 박종은 교수(KAIST)
11:00-11:10	휴식
11:10-12:40	**Practice1: Scanpy basic workflow** 정성민 조교, 고용준 조교
12:40-14:10	점심
14:10-15:30	**Public database, data integration, reference mapping, multiomics** 박종은 교수(KAIST)
15:30-15:40	휴식
15:40-16:50	**Practice2: Advanced single-cell analysis (siVI universe)** 정성민 조교, 고용준 조교

DAY1 : 2월 27일 (화)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**AI-based protein structure prediction** **- Intro to protein structure prediction** **- Early AI-based approaches** **- AlphaFold and RoseTTAFold** 백민경 교수(서울대학교)
10:50-11:00	휴식
11:00-12:10	**단백질 구조 예측 실습** **- ColabFold를 활용한 단백질 구조 및 상호작용 예측** **- Tips &Tricks for better structure modeling** 백민경 교수(서울대학교)
12:10-13:40	점심
13:40-15:10	**AI-based protein design** **- Intro to protein design** **- Protein backbone design using RFdiffusion** **- Protein sequence design using ProteinMPNN** 백민경 교수(서울대학교)
15:10-15:20	휴식
15:20-16:50	**단백질 디자인 실습** **- RFdiffusion 및 ProteinMPNN의 활용법 실습** 백민경 교수(서울대학교)

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Single-cell biology** 최정민 교수(고려대학교)
11:00-11:10	휴식
11:10-12:40	**i. Unsupervised Spatial transcriptome analysis** **ii. Tumor Boundary Determination in Spatial Transcriptomics** 유광민 조교, 이문영 조교
12:40-14:10	점심
14:10-15:30	**i. Deconvolution Analysis Using Single-cell RNA Sequencing and Spatial Transcriptomics** **ii. Cell-Cell Interaction Analysis in Spatial Transcriptomics** 김지현 조교, 최승지 조교
15:30-15:40	휴식
15:40-16:50	**i. Open Chromatin Region Analysis and Biological Interpretation of Using scATAC-seq Dataset** **ii. Construction of Gene Regulatory Networks Based on Integrated Analysis of scATAC-seq and scRNA-seq Datasets** 천하림 조교, 이호진 조교

DAY1 : 2월 28일 (수)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Transformers (이론)** 전민지 교수 (고려대학교)
11:00-11:10	휴식
11:10-12:40	**Introduction to Transformers (실습)** 봉현수 조교, 임우택 조교
12:40-14:10	점심
14:10-15:40	**Deep learning in Bioinformatics** 노미나 교수(한양대학교)
15:40-15:50	휴식
15:50-17:20	**Deep learning model을 이용한 실습** 박예솔 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**마이크로바이옴 기본 이론** 이선재 교수(GIST)
10:50-11:00	휴식
11:00-12:10	**16S rRNA amplicon seq. - DADA2** 조준우 조교, 백재우 조교
12:10-13:40	점심
13:40-14:40	**최신 메타지놈 분석 기법의 현황** 이선재 교수(GIST)
14:40-14:50	휴식
14:50-16:50	**Shotgun metagenome 분석 (Linux)** 조준우 조교, 백재우 조교

DAY1 : 2월 29일 (목)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**화학정보학 기초(Cheminformatics) / 약물특성 및 약물다움(druglikeness)** **Molecular Notations &Descriptors / AI 신약개발을 위한 Databases** **AI 신약개발을 위한 Programming 기초** 김동섭 교수(KAIST)
10:50-11:00	휴식
11:00-12:10	**Google Colab에 RDKit 설치 / 화합물 정보 읽기 실습** **Bioactivity database 검색 및 정보 읽기 실습** **Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습** 정수재 조교, 나민주 조교
12:10-13:40	점심
13:40-15:10	**AI 신약개발을 위한 기계학습법 기초 / QSAR 모델링 기초 / AI 신약개발을 위한 딥러닝 모델** **Virtual screening (ligand-based, structure-based) 및 de novo design** 김동섭 교수(KAIST)
15:10-15:20	휴식
15:20-16:50	**QSAR modeling 전체 과정 실습/ 화합물의 Bioactivity 예측 모델 개발** **Virtual screening 과정을 통한 신약후보물질 발굴 실습** 정수재 조교, 나민주 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Single cell multiomics 이론 / Gene regulatory network 이론** 김준일 교수(숭실대학교)
11:00-11:10	휴식
11:10-12:40	**Seurat/Signac, ArchR, TENET+ 실습** 김현규 조교, 정회빈 조교
12:40-14:10	점심
14:10-15:40	**롱리드 시퀀싱 소개 및 유전체 조립 실습** 김준 교수(충남대학교)
15:40-15:50	휴식
15:50-17:20	**변이 분석 및 시각화 실습** 김준 교수(충남대학교)

강의개요

Genome assembly tutorial

2000년대 초반, 약 3조 원 가량의 연구비가 투입된 결과 최초의 고품질 인간 유전체 지도가 확보되었다. 20여 년이 흐르는 사이 롱리드 시퀀싱(long-read sequencing)이라 불리는 기술이 빠르게 발전함에 따라 그 비용은 수십만 배 가량 저렴해지면서 매우 다양한 활용이 가능해졌다. 특히 유전체 지도가 없던 생물을 연구하기 위한 유전체 지도 작성 사업(genome project)이 가능해졌으며, 암세포처럼 엄청나게 큰 돌연변이를 지니고 있는 경우에도 효과적인 변이 분석이 가능해지고 있다.

본 강의에서는 이러한 롱리드 시퀀싱 기법을 활용해 유전체 지도를 작성하는 과정인 유전체 조립 분석법을 실습하고자 한다. 이를 통해 다양한 생물의 유전체를 조립하고 분석하는 것이 가능해질 것이다. 이에 더해 암세포에 존재하는 수많은 돌연변이를 보다 정교하게, 그리고 보다 포괄적으로 분석할 수 있는 방식에 대해 알아보고자 한다. 해당 강의에서는 최근 대중적으로 공개된 암 세포주 대상 롱리드 시퀀싱 데이터를 활용하고자 한다.

강의는 다음의 내용을 포함한다:

- 롱리드 시퀀싱 개요 및 데이터 소개
- 유전체 조립 실습
- 유전체 지도 기반 변이 분석 실습
- 전좌를 비롯한 거대한 구조 변이 시각화

* 참고강의교재:

  Kim, J.† and Kim, C.† (2022). A beginner's guide to assembling a draft genome and analyzing structural variants with long-read sequencing technologies. STAR Protocols 101506. (†Co-corresponding)

* 교육생준비물: 인터넷 접속 가능한 노트북

* 강의 난이도: 중급

* 강의: 김준 교수 (충남대학교 생명정보융합학과)

Curriculum Vitae

Speaker Name: Jun Kim, Ph.D.

▶ **Personal Info**

Name	Jun Kim
Title	Assistant professor
Affiliation	Chungnam National University

▶ **Contact Information**

Address	99, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
Email	junkim@cnu.ac.kr
Phone Number	010-3360-1208

Research Interest

Structural variant, Splicing

Educational Experience

2015	B.S. in Life Sciences, POSTECH, Republic of Korea
2020	Ph.D. in Biological Sciences, Seoul National University, Republic of Korea

Professional Experience

2018-2020	Researcher, Research Institute of Basic Sciences, SNU, Korea
2020-2022	Postdoctoral Researcher, Research Institute of Basic Sciences, SNU, Korea
2022-2023	Researcher, KRIBB, Korea
2023-	Assistant professor, Chungnam National University, Korea

Selected Publications (5 maximum)

1. Lim, J., Kim, W., **Kim, J.†** and Lee, J.† (2023). Telomeric repeat evolution in the phylum Nematoda revealed by high-quality genome assemblies and subtelomere structures. ***Genome Research***. (†Co-corresponding)
2. Kim, E.*, **Kim, J.*†**, Kim, C., and Lee, J.† (2021). Long-read sequencing and de novo genome assemblies reveal complex chromosome end structures caused by telomere dysfunction at the single nucleotide level. ***Nucleic Acids Research*** 49, 3338-3353. (*Co-first; †Co-corresponding)
3. Kim, C.*, **Kim, J.***, Kim, S.*, Cook, D.E., Evans, K.S., Andersen, E.C., and Lee, J. (2019). Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in C. elegans. ***Genome Research*** 29, 1023-1035. (*Co-first)
4. Lee, H., **Kim, J.†** and Lee, J.† (2023). Benchmarking datasets for assembly-based variant calling using high-fidelity long reads. ***BMC Genomics*** 24. (†Co-corresponding)
5. Kim, C.*, Sung, S.*, **Kim, J.***, and Lee, J. (2020). Repair and Reconstruction of Telomeric and Subtelomeric Regions and Genesis of New Telomeres: Implications for Chromosome Evolution. ***BioEssays*** 42, 1900177. (*Co-first)

KSBI | 한국생명정보학회

KSBi-BIML 2024

Genome assembly tutorial

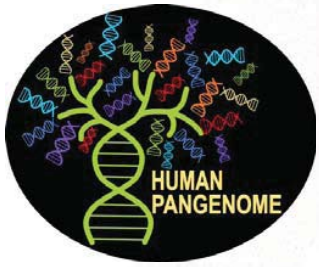
Genomics, as a big data science

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Stephens et al., 2015

Large-scale sequencing projects



30 million USD
350 human genomes

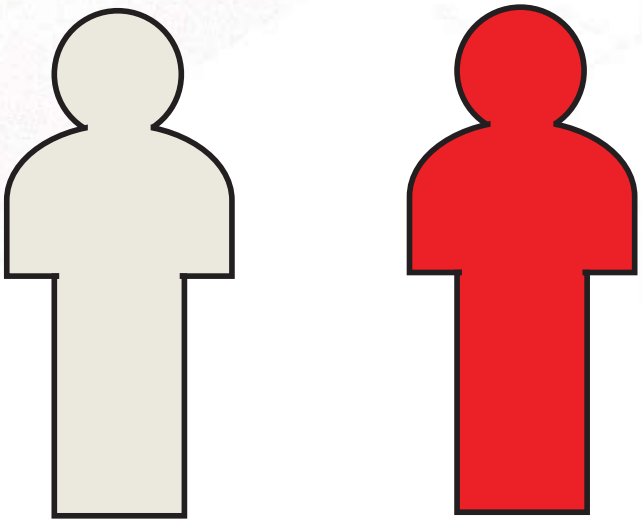
5 billion USD
>9,000 lives

국가 바이오
빅데이터 구축

1.5 trillion KRW
1 million Koreans
~100 PB

3

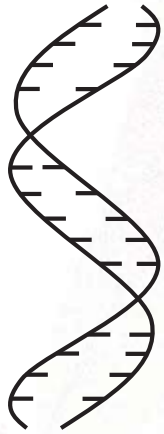
Intra-specific variations



Disease susceptibility

4

Sequencing



DNA

Sequencing

Normal ATGCACGTCAGT

Patient ATGCAC**C**TCAGT

Variant

5

Sequencing technologies

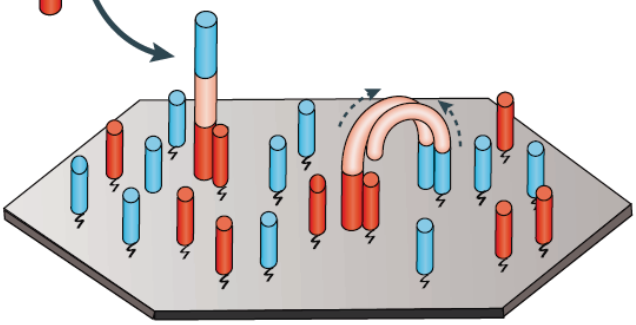
Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60	87–92	0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100		5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/ GridION	Long	10–60	>1,000	87–98	2-20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500		0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15	>99.9	16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25		65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

Logsdon et al., 2020

6

Illumina

Template binding
Free templates hybridize with slide-bound adapters

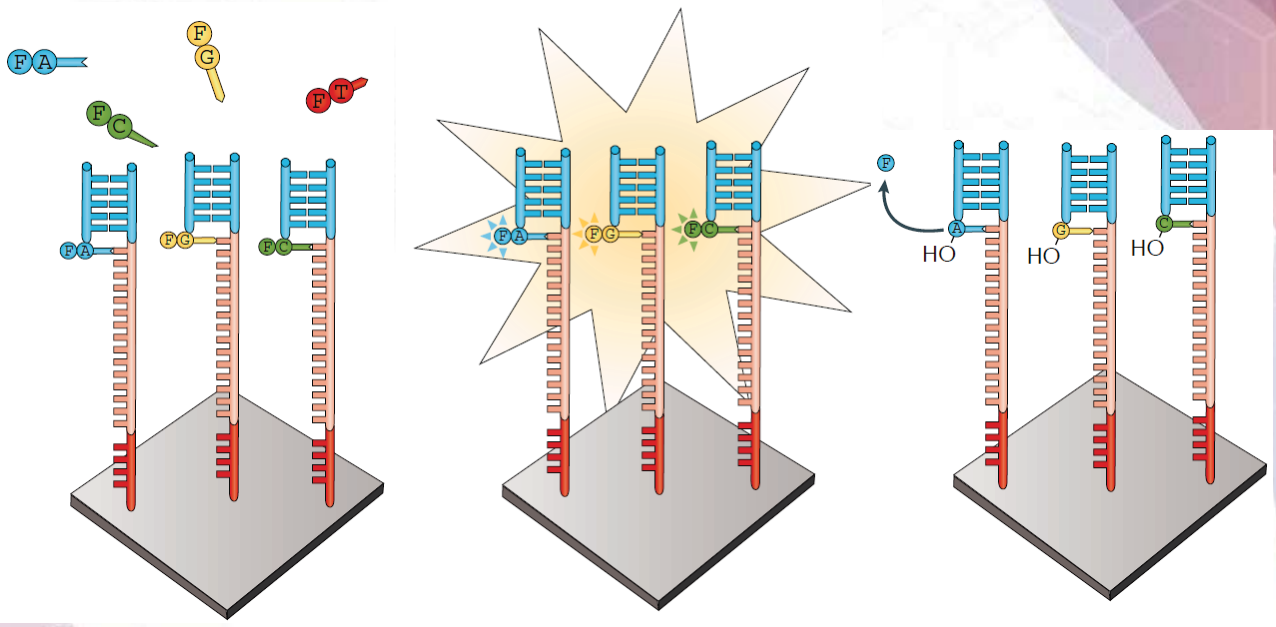


Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place

Goodwin et al., 2016

7

Illumina



Nucleotide addition

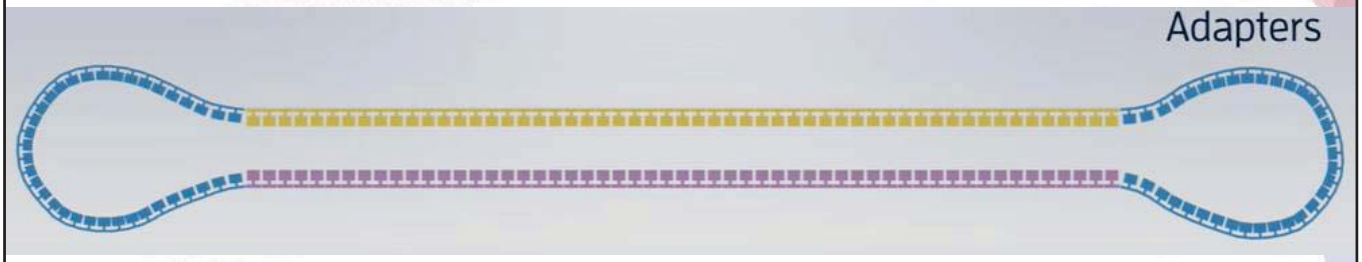
Imaging

Cleavage

Goodwin et al., 2016

8

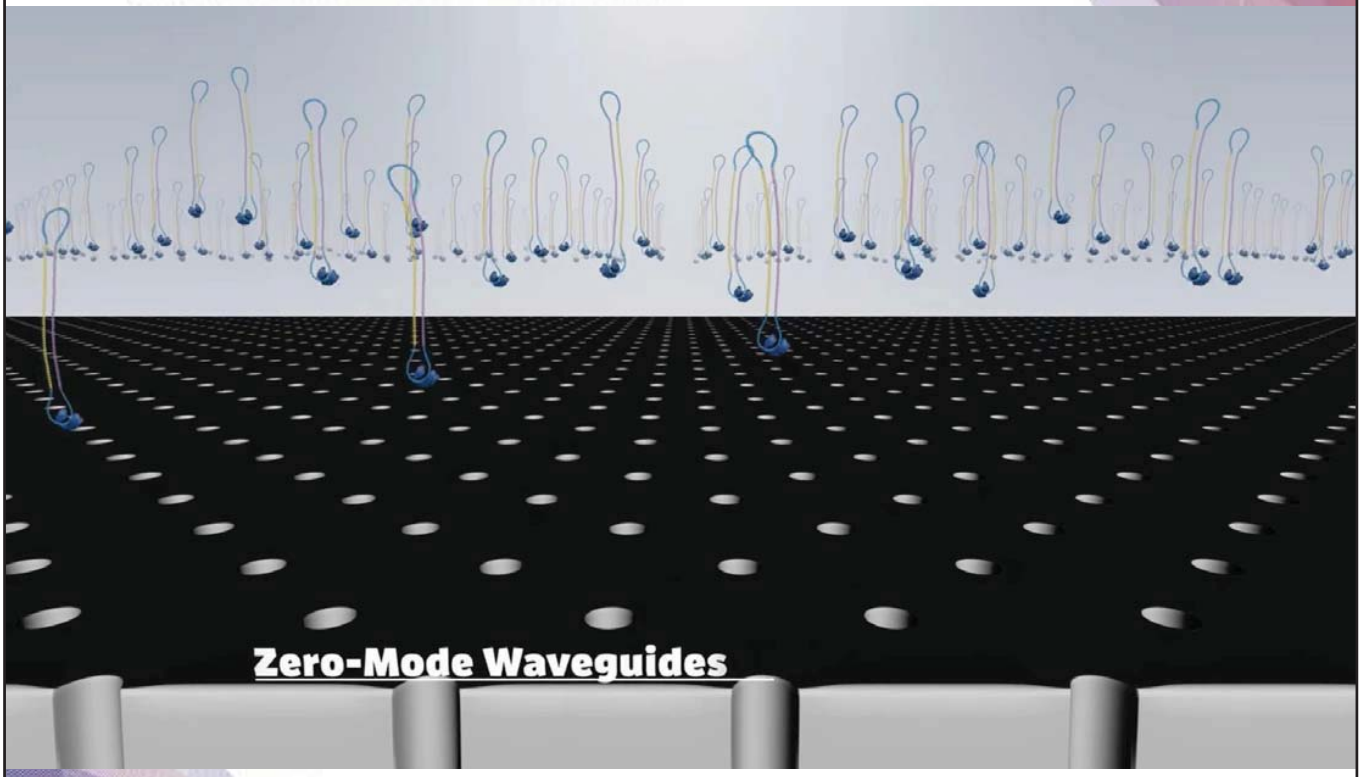
PacBio



Goodwin et al., 2016

9

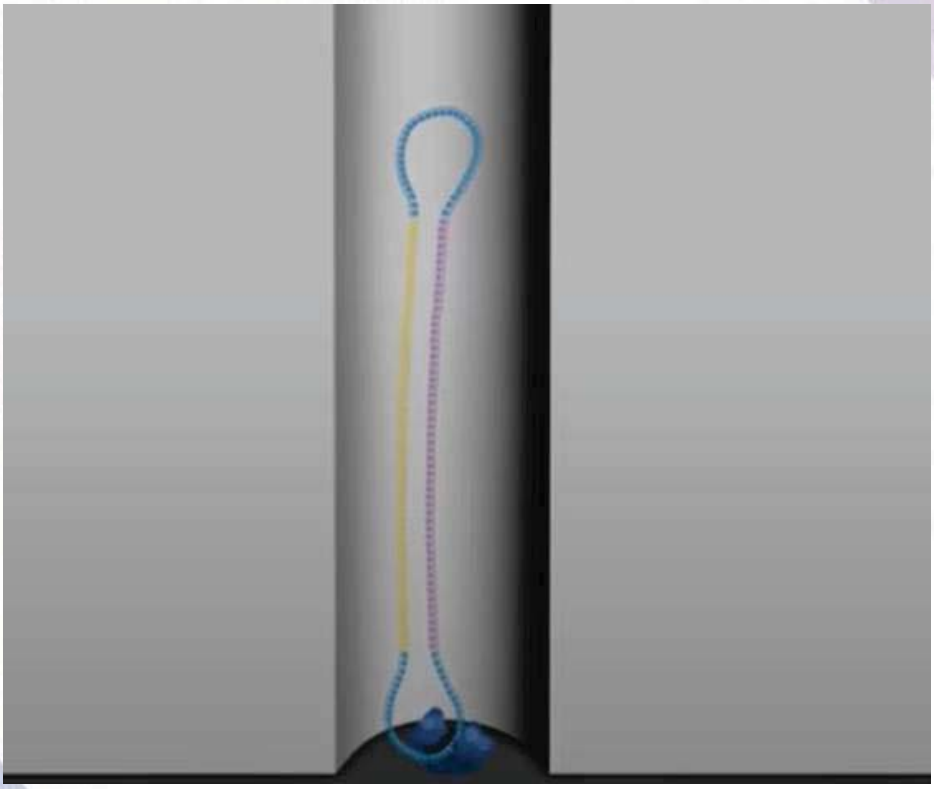
PacBio



https://www.youtube.com/watch?v=_lD8JyAbwEo

10

PacBio



https://www.youtube.com/watch?v=_ID8JyAbwEo

11

PacBio

https://www.youtube.com/watch?v=_ID8JyAbwEo

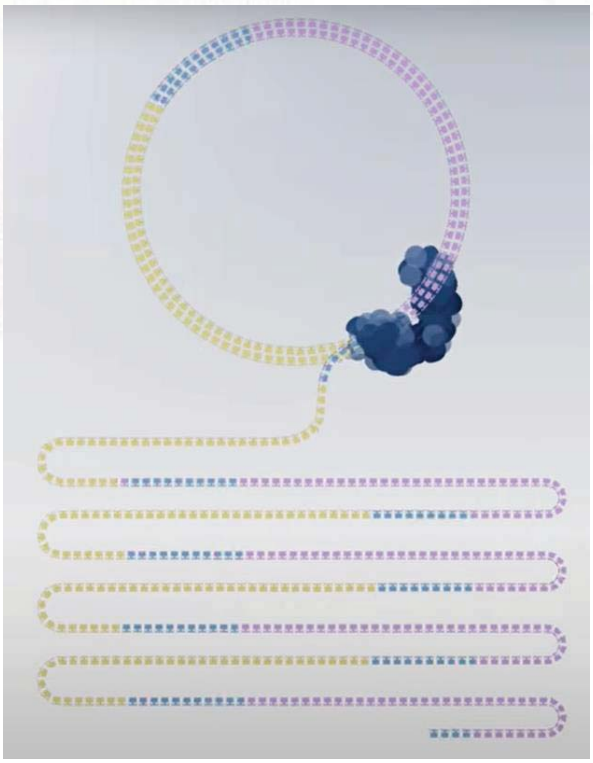
12

PacBio

https://www.youtube.com/watch?v=_ID8JyAbwEo

13

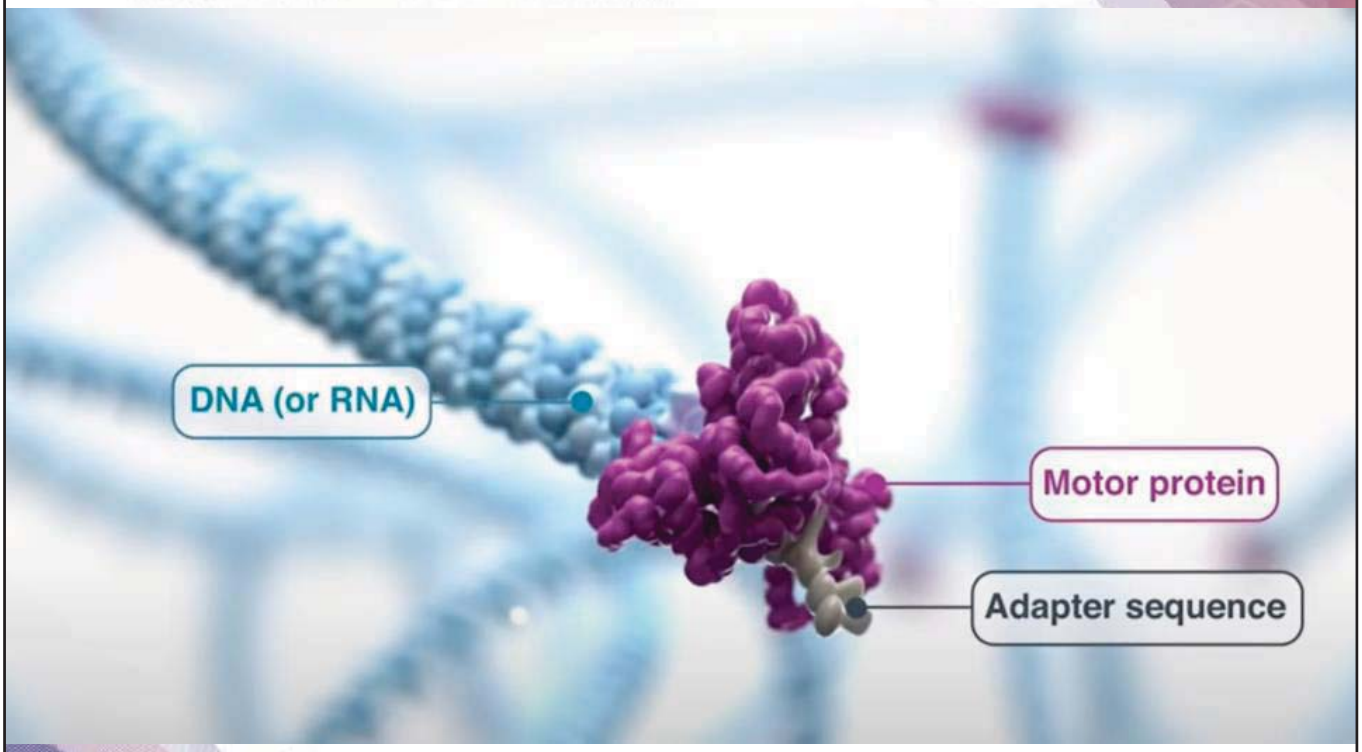
PacBio, high-fidelity reads (HiFi)



https://www.youtube.com/watch?v=_ID8JyAbwEo

14

ONT



https://www.youtube.com/watch?v=RcP85JHLmnI

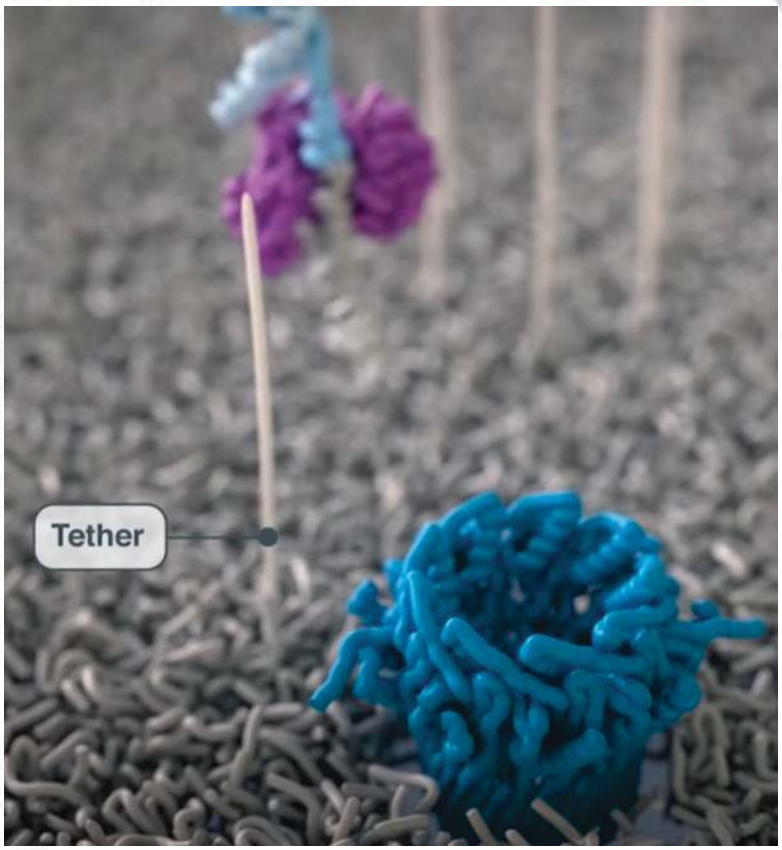
15

ONT

https://www.youtube.com/watch?v=RcP85JHLmnI

16

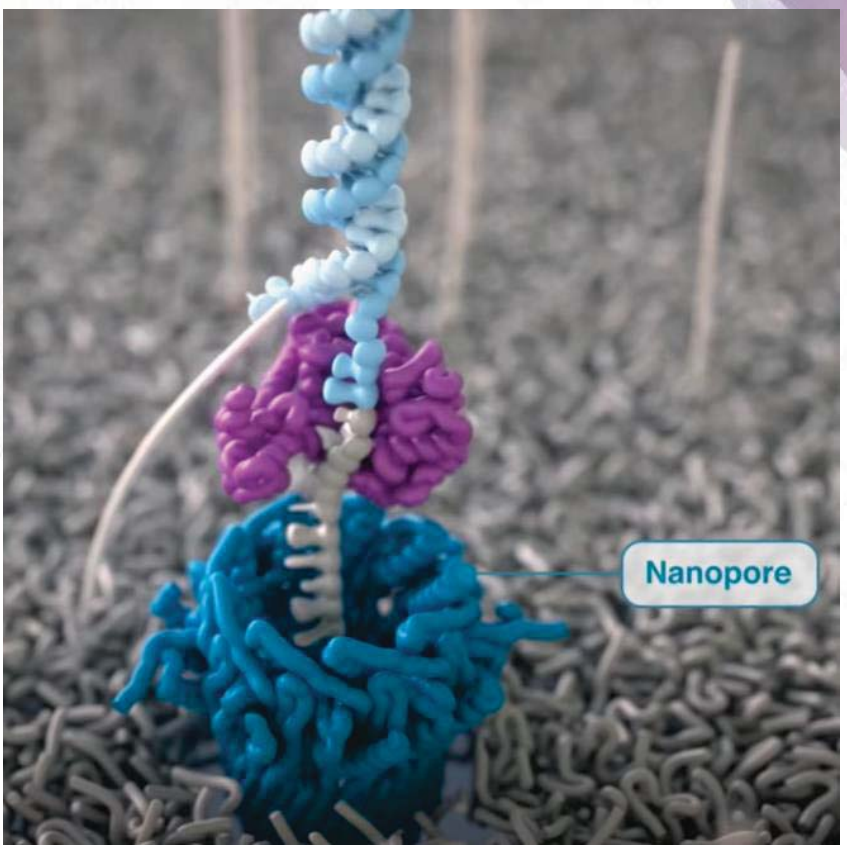
ONT



https://www.youtube.com/watch?v=RcP85JHLmnI

17

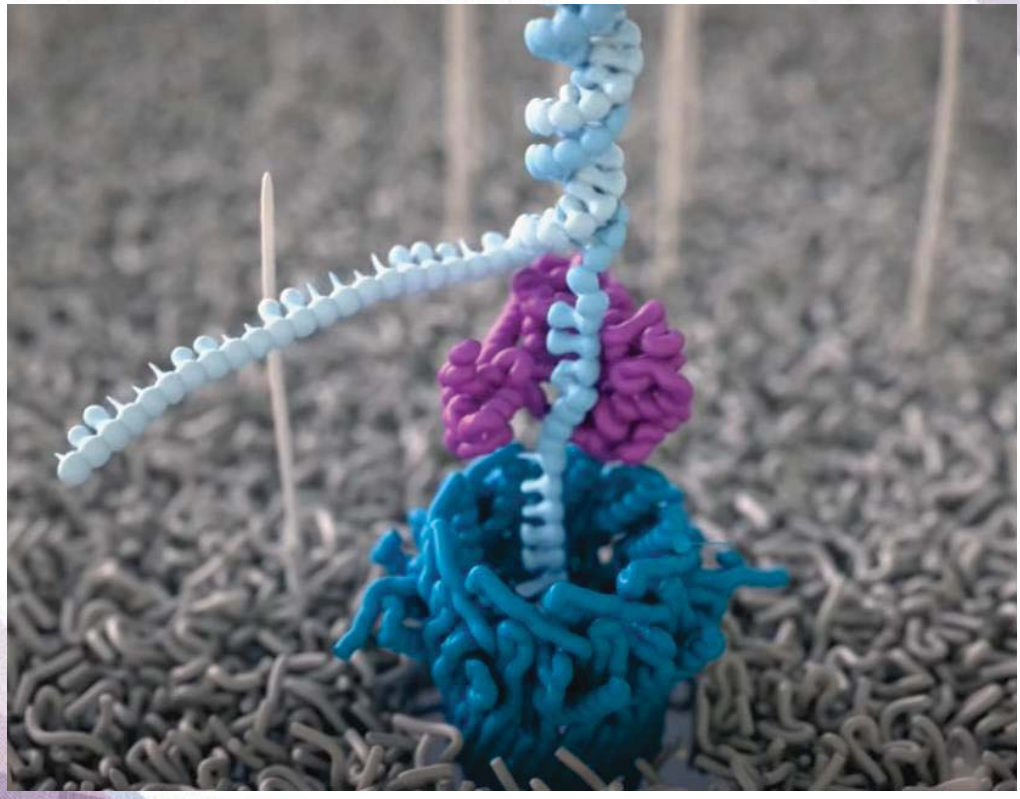
ONT



https://www.youtube.com/watch?v=RcP85JHLmnI

18

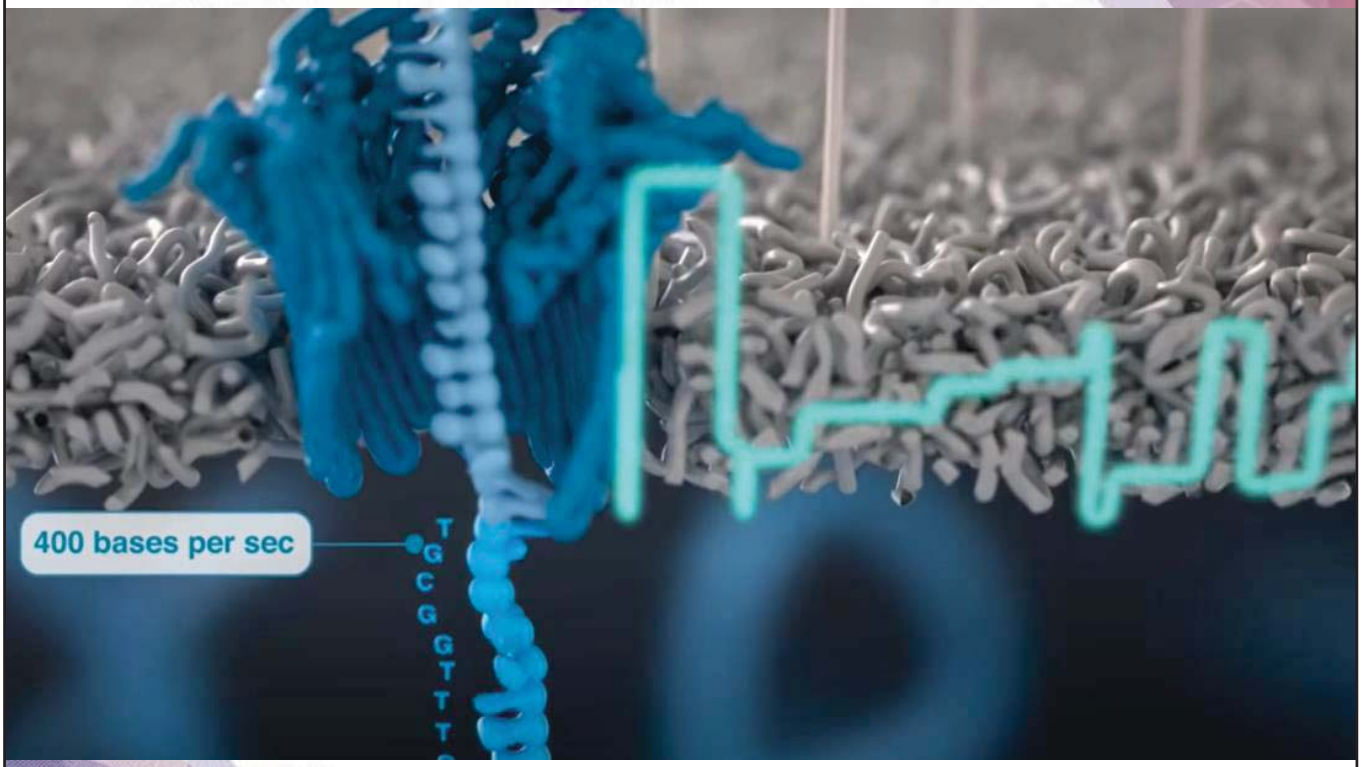
ONT



https://www.youtube.com/watch?v=RcP85JHLmnI

19

ONT



https://www.youtube.com/watch?v=RcP85JHLmnI

20

Summary

Commercial sequencing technologies provide high-throughput, high-quality read sequences.

Illumina
Cheap; fragmented DNA; amplification; highly accurate; very short (~250 bp)

PacBio
Expensive; high-molecular-weight DNA; abundant DNA; long & accurate (~20 kb)

ONT
Expensive, high-molecular-weight DNA or RNA; abundant DNA or RNA; ultra-long (~1 Mb)

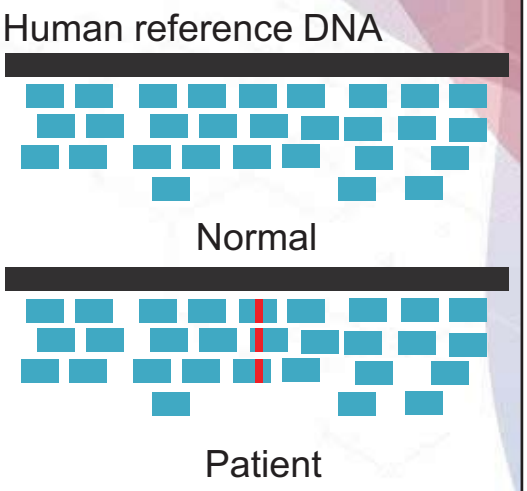
21

Variant detection

Human reference DNA (3 Gb)
ACAGTCAGTCAGTCAG
CTAGCATGCATCGATC
GTAGCATGCTAATCGA
.....
AGCTCGATCGATCGAT
CGATCGATCGATCGAT
CGATGACGTACGTGCG

Sequenced DNA (100 bp/each)
ACAGTCAGTCAGTCA
.....
ATCGATGACGTACGT

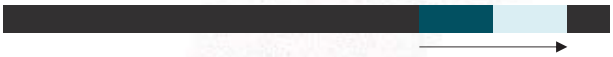
× 10^9



22

Structural variant (SV) detection

Human reference



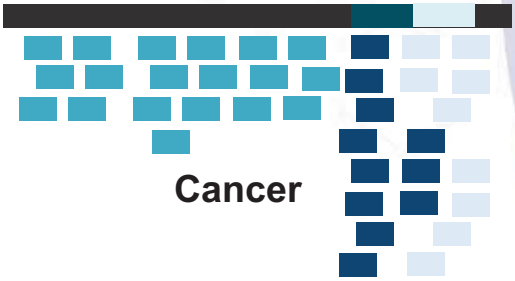
Cancer (DNA damage)

Cancer (repaired DNA)

Increased cancer survival

Human reference DNA

Normal



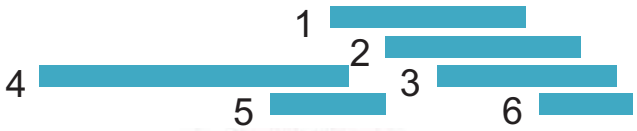
Cancer

Duplicated

23

Genome assembly and SV calling

Cancer



Human reference

Cancer patient

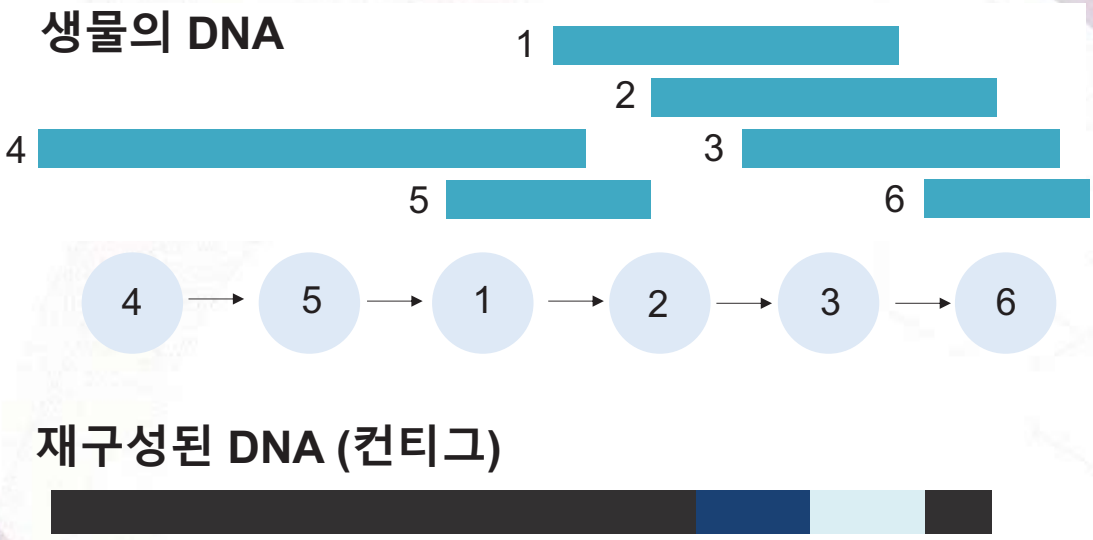
Inversion

Duplication

Reconstructed cancer DNA sequences

24

유전체 조립(컨티그 형성)



롱리드 시퀀싱 기법이 발전하면서, 매우 높은 정확도를 지닌 DNA (HiFi 리드)를 활용해 더 큰 DNA로 재구성할 수 있게 됐다.

25

DNA 시퀀싱 기법 – 컨티그 형성용

PacBio

Long & accurate (~20 kb)

Error rate

Raw read: 10^{-3} – 10^{-4}

Assembly: 10^{-5} – 10^{-6}

ONT

Ultra-long (~200 kb, up to ~1 Mb)

Error rate

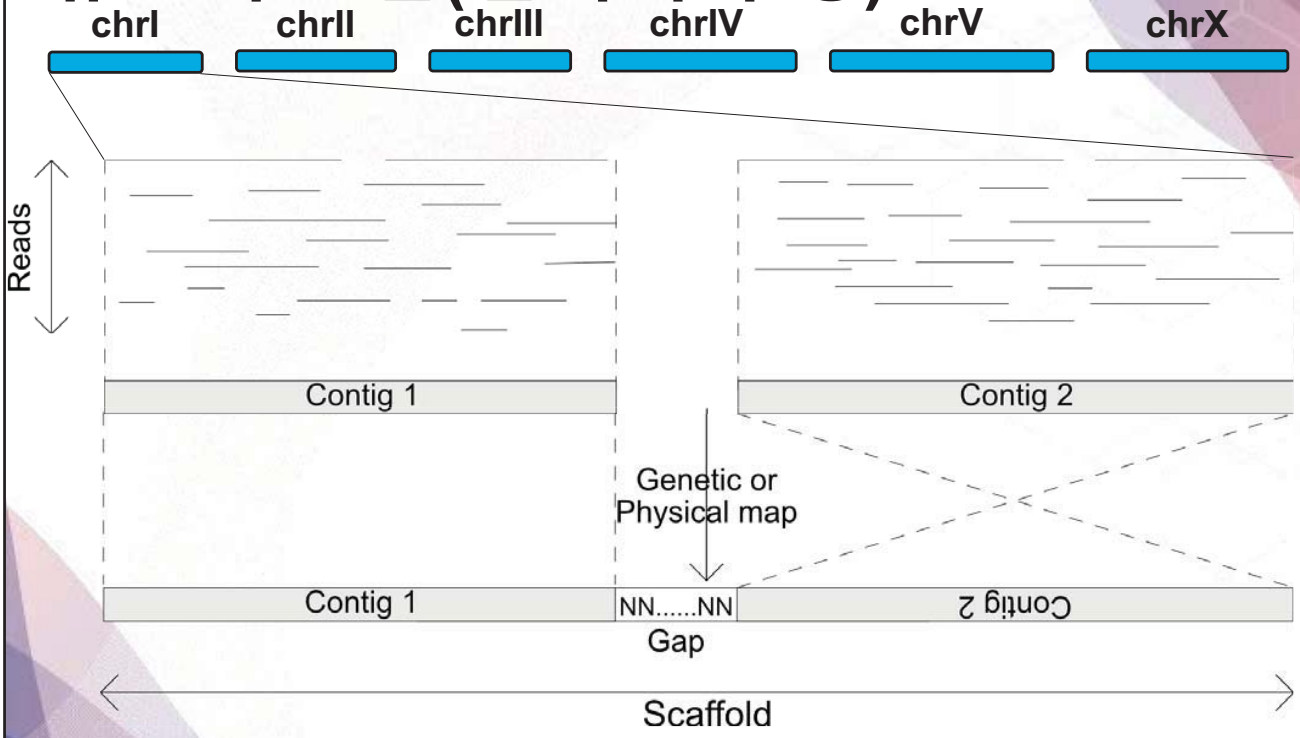
Raw read: 10^{-1} – 10^{-2}

Assembly: 10^{-3} – 10^{-4} → 유전자에도 오류 존재 가능

현재는 HiFi 기법을 이용한 컨티그 형성이 전세계 표준으로 활용되고 있다. (고품질 DNA 필요함) ONT는 상반기 개선 예정

26

유전체 조립(염색체 구성)



컨티그에 더해, Hi-C 및 UL 데이터 등을 활용해 염색체 수준의 표준 유전체 지도를 확보하는 것이 가능해졌다.

27

DNA 시퀀싱 기법 – 염색체 구성용

Hi-C

- 컨티그의 순서 및 방향을 결정함
- Noise가 심해 실험이 실패하는 경우가 잦음(= 비쌈)

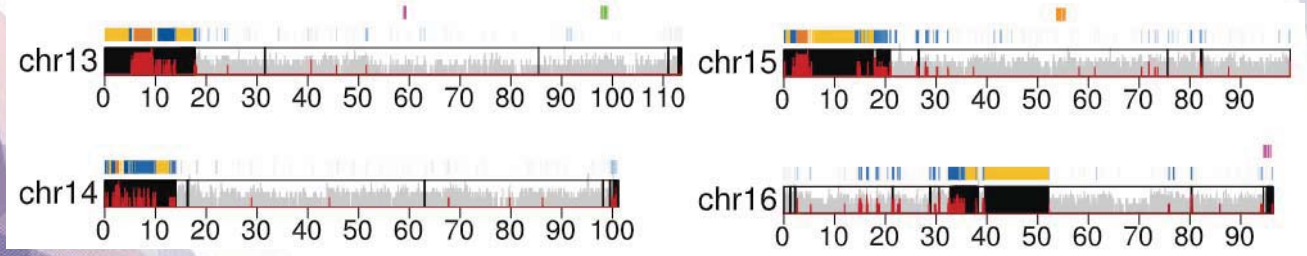
ONT UL

- 컨티그 사이의 빈틈을 메워 고품질 염색체 제작
- 품질이 매우매우 높은 DNA가 필요함(= 비쌈)

28

최초로 완성된 빈틈 없는 인간 유전체

STATISTICS	GRCH38	T2T-CHM13	DIFFERENCE (±%)
	Summary		
Assembled bases (Gbp)	2.92	3.05	+4.5
Unplaced bases (Mbp)	11.42	0	−100.0
Gap bases (Mbp)	120.31	0	−100.0
Number of contigs	949	24	−97.5
Contig NG50 (Mbp)	56.41	154.26	+173.5
Number of issues	230	46	−80.0
Issues (Mbp)	230.43	8.18	−96.5



Nurk et al., 2022 (Nature)

29

DNA 시퀀싱 기법 – 범유전체 구축

PacBio HiFi

- 다수 개체의 유전체 정보를 고품질로 확보 가능
- 염색체 정보가 있을 경우 저렴하게 분석 가능
- 변이 정보가 필요할 경우에만 유용

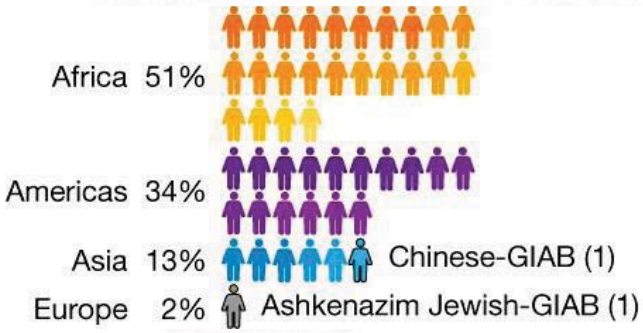
Trio

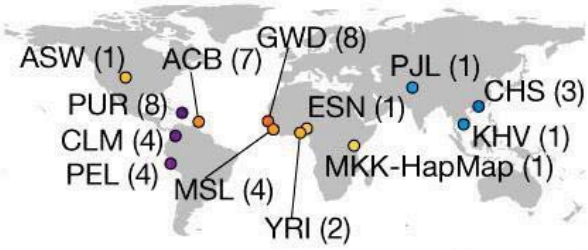
- 매우 저렴하고 정확하게 부/모 양쪽의 염색체를 분리
- 부/모 개체를 모두 구할 수 있는 경우에만 활용 가능

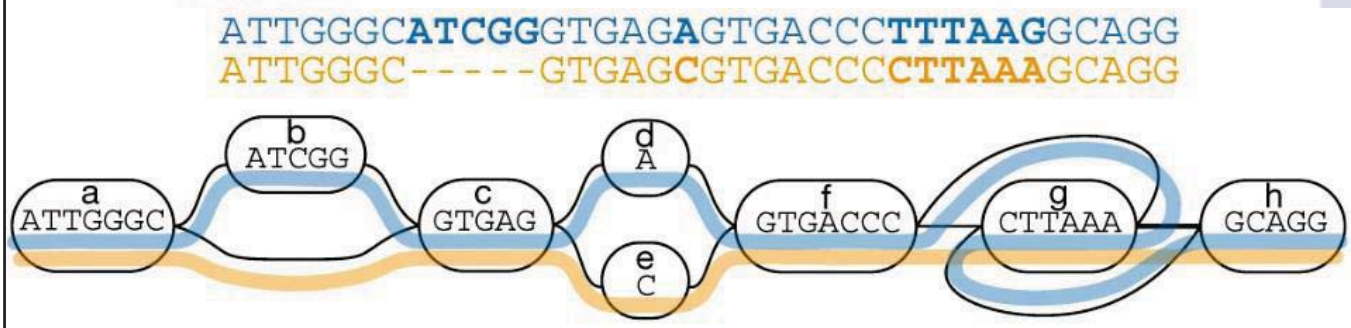
30

전세계 인구를 대표할 범유전체 지도

Liao et al., 2023 (Nature)







집단에 존재하는 모든 변이를 분석함으로써, 육종 효율을 개선하는 사례가 점차 늘고 있다.

31

요약

- 롱리드 시퀀싱을 통해 유전체 지도를 확보할 수 있다.

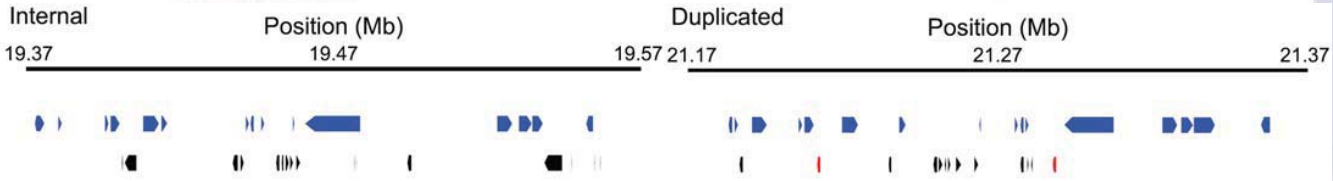
- 집단에 존재하는 모든 변이를 확인할 수 있다.

32

Segmental duplication (*C. elegans*)

Bristol, **Reference strain**

Hawaiian, **Divergent strain**



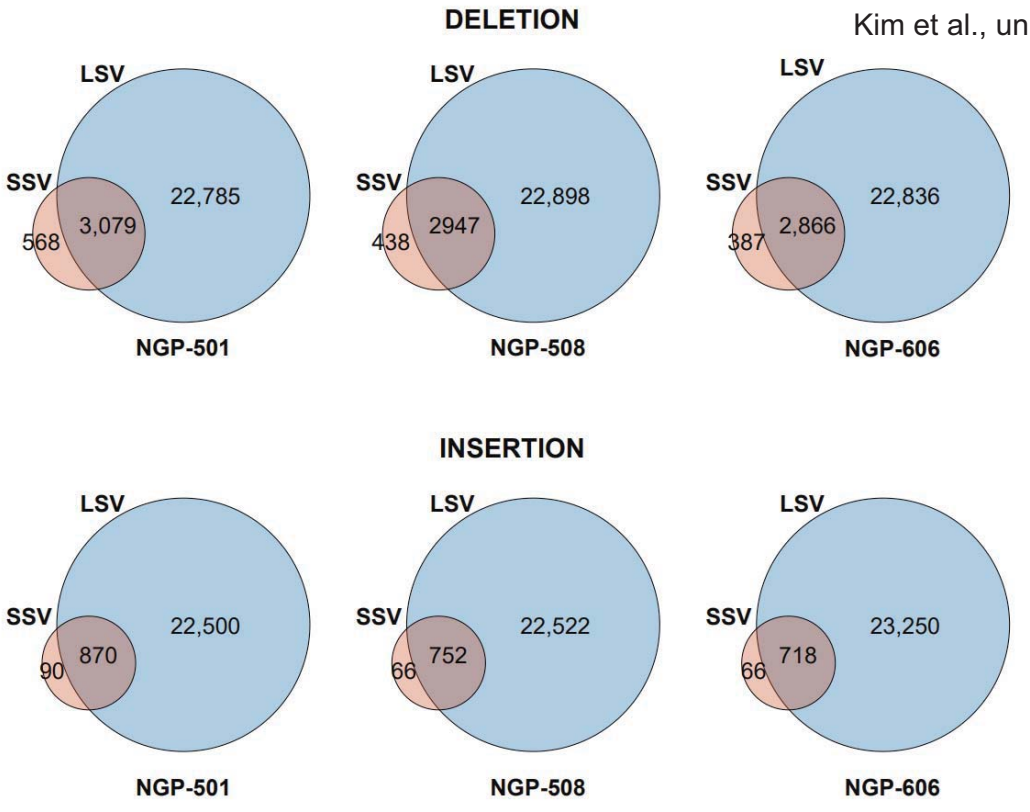
- Common genes
- Specific genes
- Homolog in other chromosomes

Kim et al., 2019
Kim et al., 2020

33

SV detection using two technologies

Kim et al., unpublished



34

Application in human genetic diseases

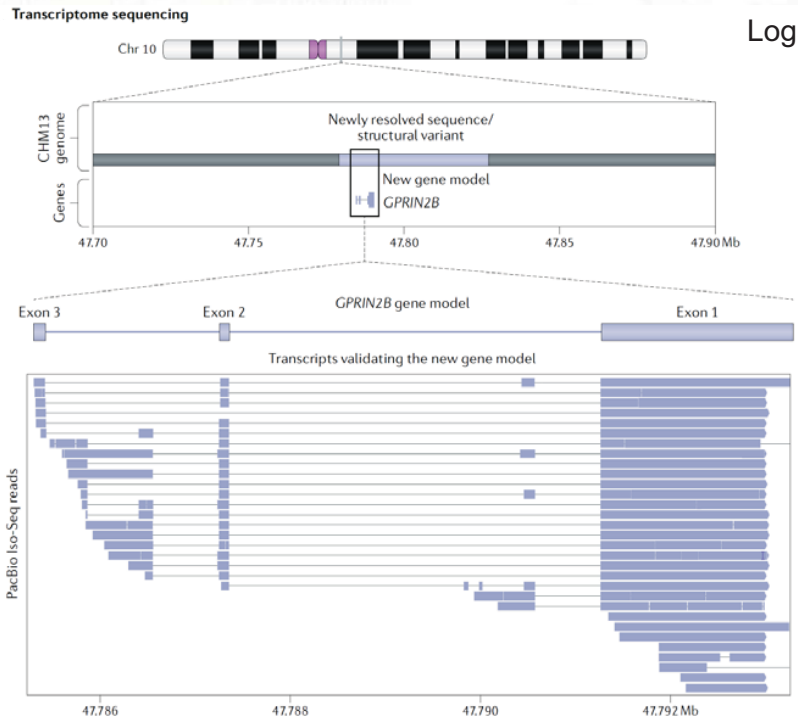
Disease	Mutation type	Gene/Locus	Method	Sequencing platform	Mean read length	Predicted cogerage	Reference
Benign adult familial myoclonic epilepsy	Tandem repeat	*SAMD12*	Whole genome	PacBio/Sequel	7705	7×	Mizuguchi et al. [9]
Benign adult familial myoclonic epilepsy	Tandem repeat	*SAMD12*	Whole genome	PacBio/Sequel and Nanopore	>10 k	~10×	Zeng et al. [10]
Benign adult familial myoclonic epilepsy	Tandem repeat	*SAMD12*	BAC clone or whole genome	PacBio/RSII or Nanopore/MinION	N.A.	N.A.	Ishiura et al. [25]
Neuronal intranuclear inclusion disease	Tandem repeat	*NOTCH2NLC*	Whole genome	PacBio/RSII or Nanopore/PromethION(R9.4)	1269–20,204	6×–25×	Sone et al. [6]
Congenital abnormalities	Chromothripsis	–	Whole genome	Nanopore/MinION(R7.3,R9,R9.4)	(Distribution plot is available)	16×, 11 ×	Stancu et al. [32]
Intellectual disability and seizure	Chromosome translocation	*ARFGEF9*	Whole genome	Nanopore/MinION(R9.4)	9 k	1.46× (chr20, chrX)	Dutta et al. [33]
Progressive myoclonic epilepsy	12 kb deletion	*CLN6*	Whole genome	PacBio/Sequel	9744	6×	Mizuguchi et al. [8]
Carney complex	2k deletion	*PRKAR1A*	Whole genome	PacBio/Sequel	>9 k	9×	Merker et al. [7]
Bardet-Biedl syndrome	72.8-Kb deletion	7q14.3	Whole genome	PacBio/RSII	N.A.	539,118 bases	Reiner et al. [34]
Neonatal hypoxic-ischaemic encephalopathy	Duplication-inversion-duplication	*CDKL5*	Whole genome	Nanopore/R9	8136 (median)	3× (minimum)	Sanchis-Juan et al. [11]
Amyotrophic lateral sclerosis and fromtotemporal dementia	Tandem repeat	*C9orf72*	whole genome and No-Amp	PacBio/Sequel or Nanopore/MinION	N.A.	7× or 3×	Ebbert et al. [42]
Glycogen storage disease	7.1 kb deletion	*G6PC*	whole genome sequencing	Nanopore/GridION(R9.4)	16,579	10×	Miao et al. [35]
Myotonic dystrophy	Tandem repeat	*DMPK*	PCR amplicon	PacBio/	N.A.	N.A.	Cumming et al. [20]
Fragile X syndrome	Tandem repeat	*FMR1*	PCR amplicon	PacBio/RSII	N.A.	N.A.	Ardui et al. [24]
Fragile X syndrome	Tandem repeat	*FMR1*	PCR amplicon	PacBio/RS	N.A.	N.A.	Loomis et al. [23]
Spinocerebellar atasia type 10	Tandem repeat	*ATXN10*	PCR amplicon	PacBio/RS	N.A.	N.A.	McFarland et al. [37]
Autosomal dominant tubulointerstitial kidney disease	SNV in tandem repeats	*MUC1*	PCR amplicon	PacBio/RSII	N.A.	N.A.	Wenzel et al. [38]
Huntington's disease	Tandem repeat	*HTT*	No-Amp	PacBio	N.A.	N.A.	Höijer et al. [41]
Parkinson's disease	Tandem repeat	*ATXN10*	No-Amp	PacBio	N.A.	N.A.	Schüle et al. [44]
Fuchs endothelial corneal dystrophy	Tandem repeat	*TCF4*	No-Amp	PacBio/RSII	N.A.	N.A.	Hafford-Tear et al. [43]
Neuronal intranuclear inclusion disease	Tandem repeat	*NOTCH2NLC*	Cas9-mediated target sequence	Nanopore/MinION	N.A.	N.A.	Sone et al. [6]
Becker muscular dystrophy	L1 insertion	*DMD*	PCR amplicon	PacBio/RSII	N.A.	N.A.	Gonçalves et al. [47]
X-Linked Dystonia-Parkinsonism	SVA insertion	*TAF1*	BAC clone	PacBio/RSII	N.A.	N.A.	Aneichyk et al. [48]

Mitsuhashi and Matsumoto, 2020

35

Full-length RNA sequencing

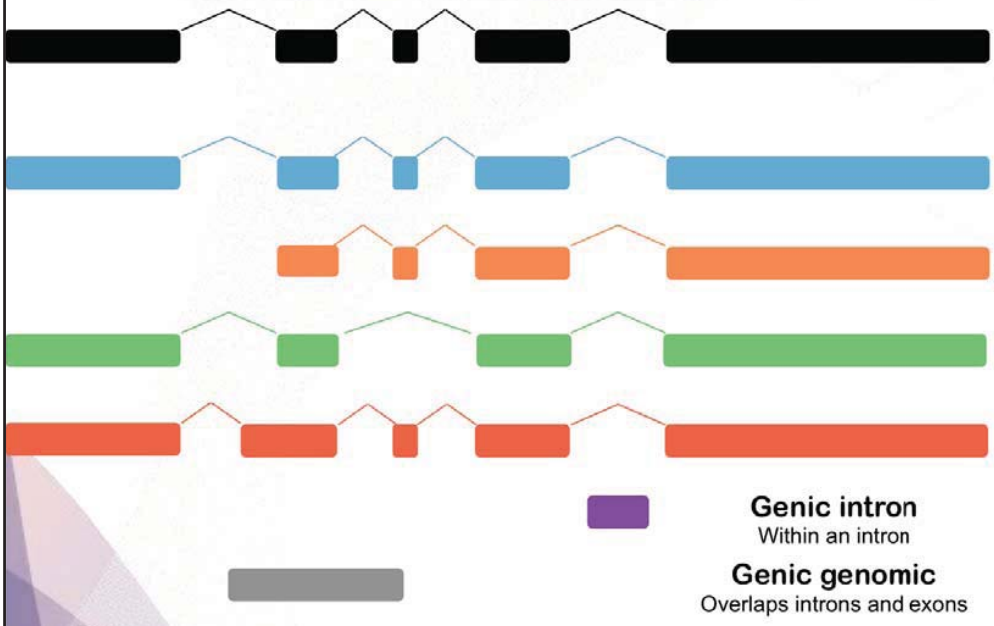
Logsdon et al., 2020



Even full-length transcripts can be sequences at once, so any isoform information can be fully interpreted.

36

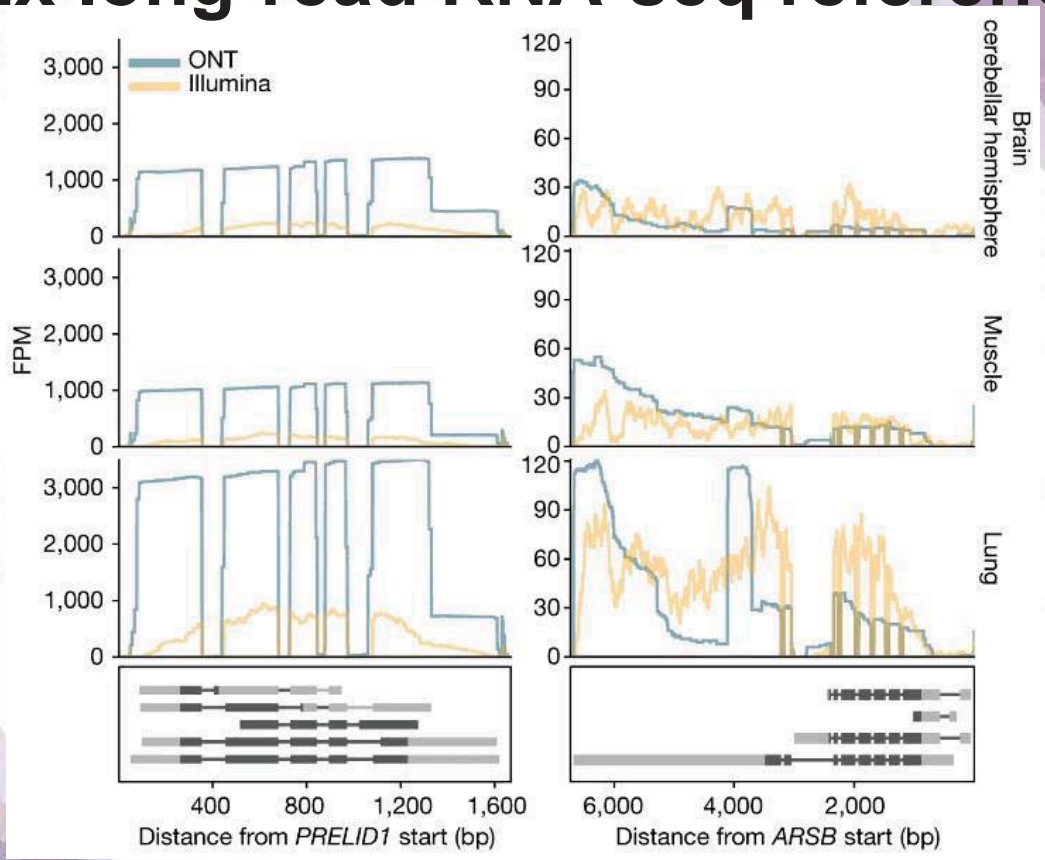
Novel isoform detection



Tardaguila et al., 2018
https://github.com/ConesaLab/SQANTI3

37

GTEx long-read RNA-seq reference

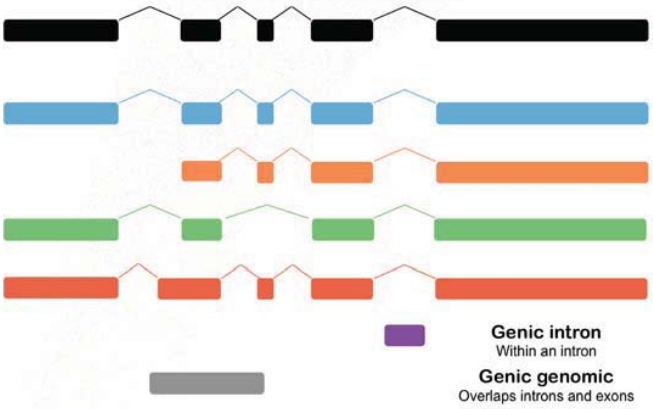


Glinos et al., 2022

38

Novel isoform detection

Kim et al., unpublished



Human tissue	
FSM	21,385
ISM	2,299
NIC	8,381
NNC	2,845

Tardaguila et al., 2018
https://github.com/ConesaLab/SQANTI3

39

STAR Protocols

PROTOCOL | JULY 1, 2022

A beginner's guide to assembling a draft genome and analyzing structural variants with long-read sequencing technologies

Jun Kim[1,3,*] and Chuna Kim[2,4,**]

[1]Research Institute of Basic Sciences, Seoul National University, Seoul 08826, Korea

[2]Aging Convergence Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Korea

[3]Technical contact

[4]Lead contact

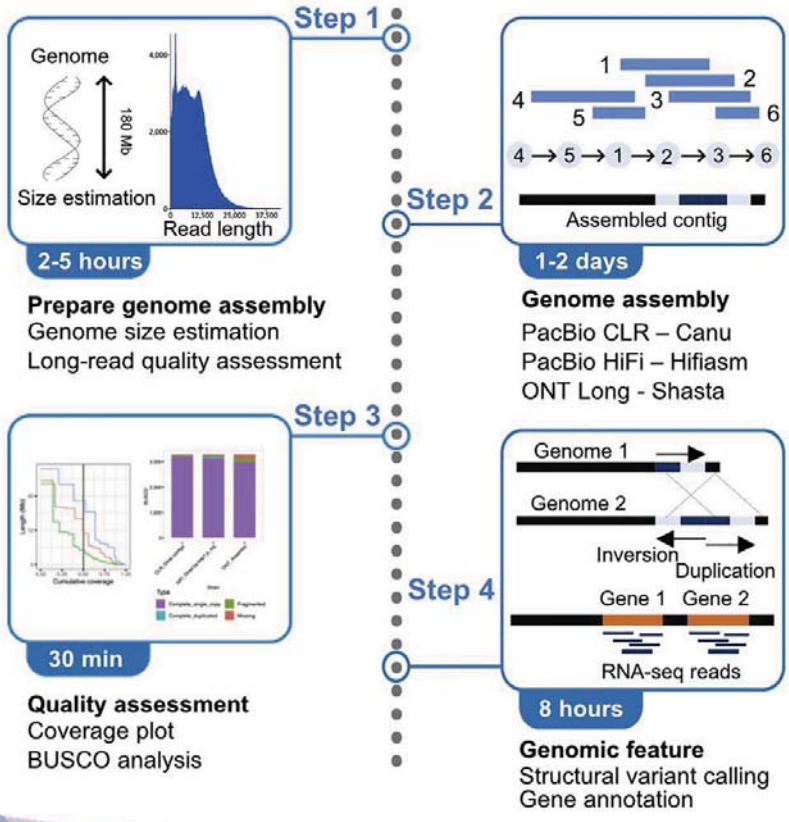
*Correspondence: dauer@snu.ac.kr

**Correspondence: kimchuna@kribb.re.kr

Open Access • DOI: 10.1016/j.xpro.2022.101506

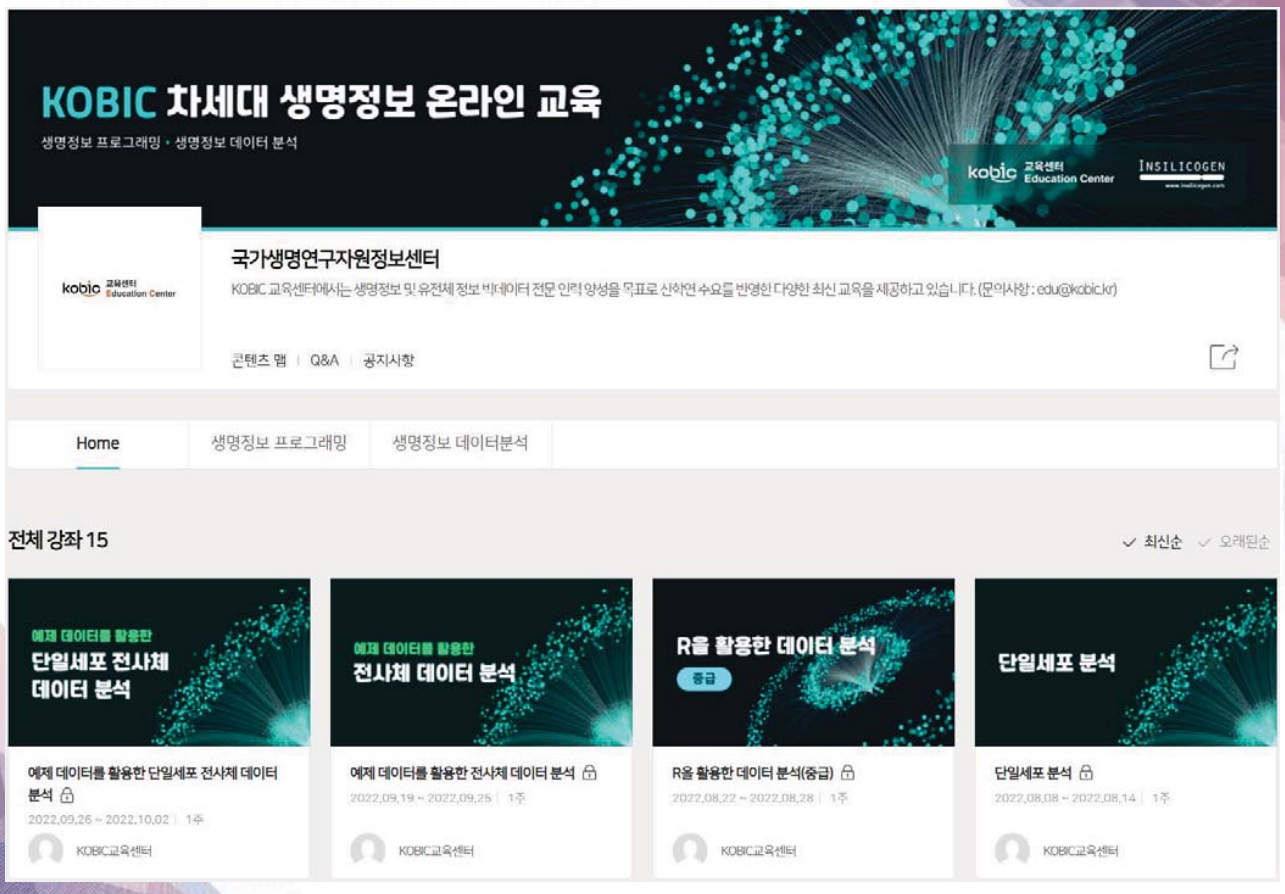
Kim and Kim 2022
https://star-protocols.cell.com/protocols/1799

40



Kim and Kim 2022
https://star-protocols.cell.com/protocols/1799

41



국가생명연구자원정보센터 : edwith
https://www.edwith.org/ptnr/kobic

42

암 유전체 연구

In collaboration with Professor Hyunho Han (YUHS)

- 6개 종양 세포주에 대한 HiFi data 확보

- Read depth 이상, 5-methylcytosine 위치 확인

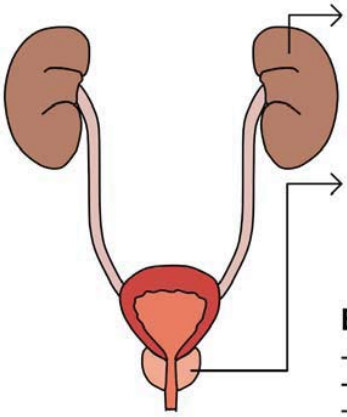
- Genome assembly 수행

- DNA rearrangement 등 조사

43

Fig. 1

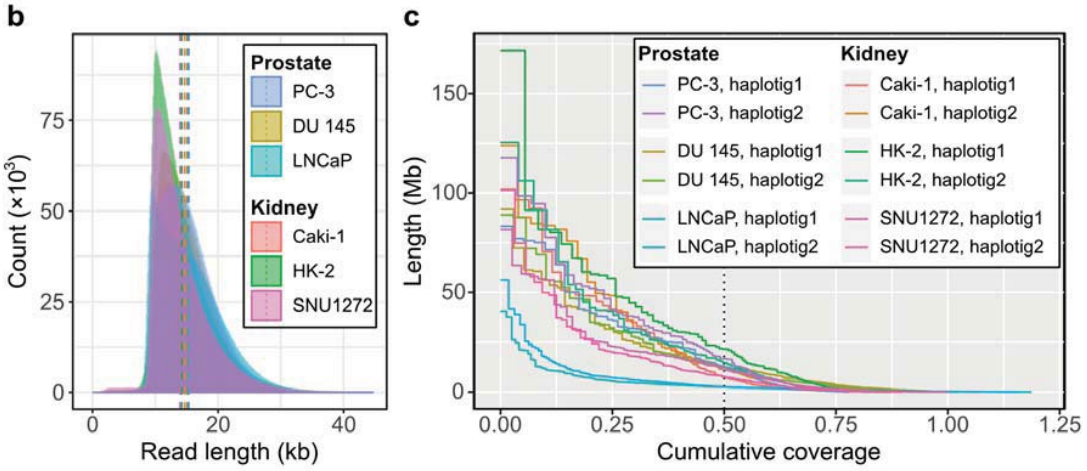
a



Kidney
Caki-1: Cancer, clear cell carcinoma
SNU1272: Cancer, clear cell renal carcinoma
HK-2: Benign tumour, proximal tubule epithelial, papilloma

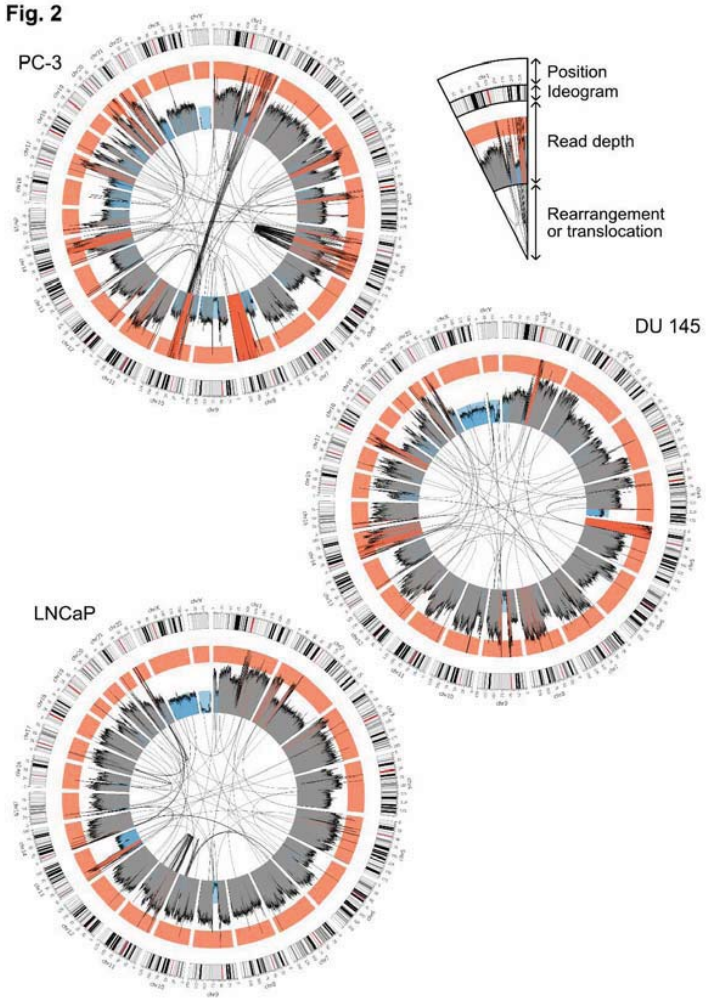
Prostate
PC-3: Prostate cancer, adenocarcinoma (Grade IV)
DU 145: Prostate cancer, carcinoma
LNCaP: Prostate cancer, carcinoma

Experimental scheme
- 25–30× High-fidelity long-read sequencing data generation
- *De novo* genome assembly
- Analyzing DNA rearrangements, translocations, ...



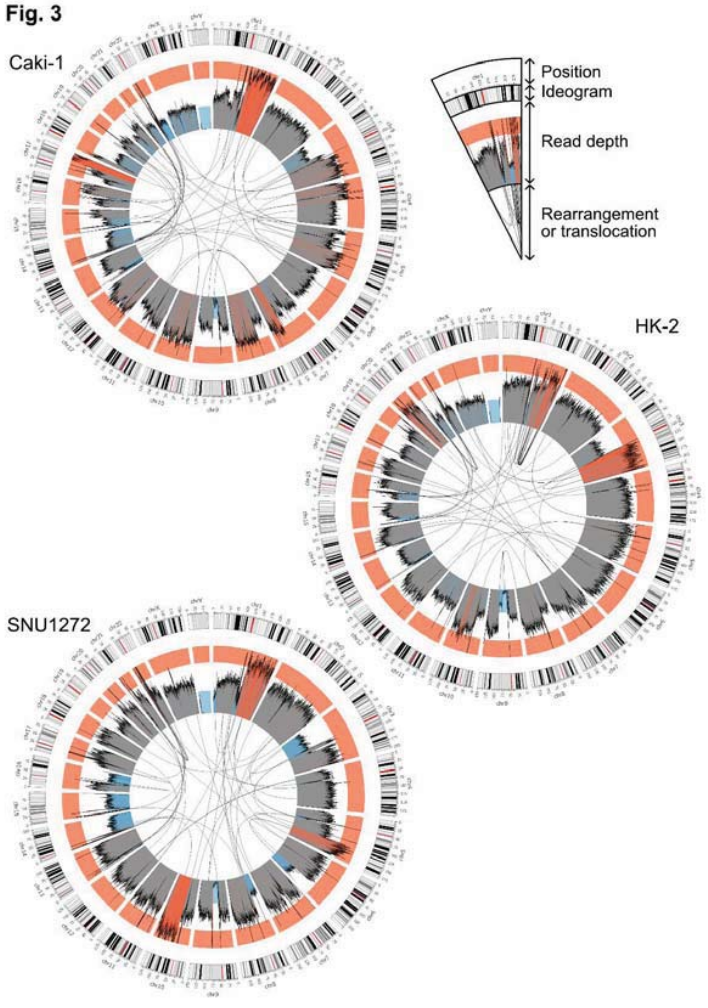
14

Fig. 2



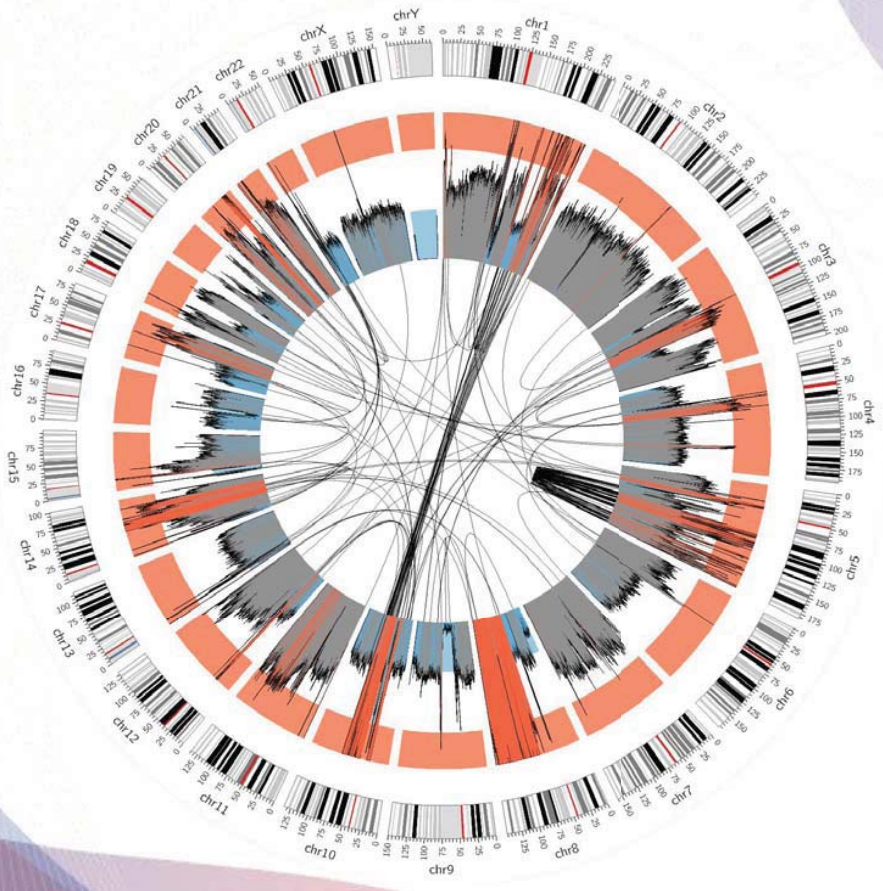
45

Fig. 3



46

PC-3



47

요약

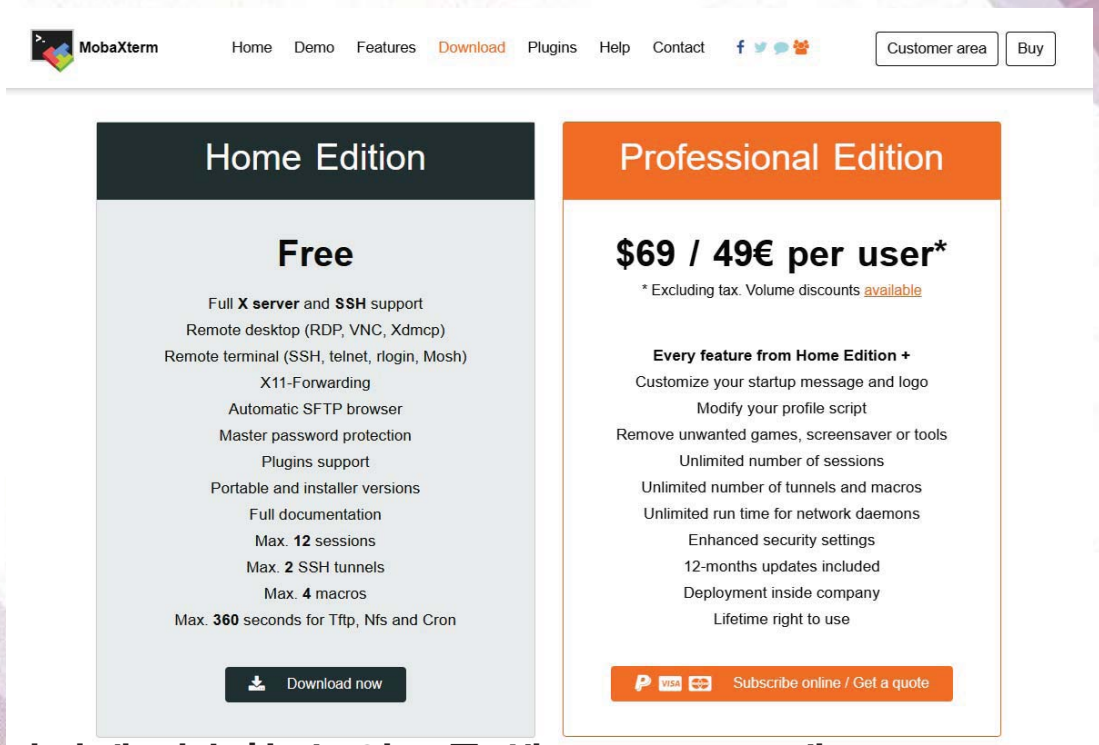
- 롱리드 시퀀싱으로 암 유전체 지도를 확보할 수 있다.

- Copy number variation은 물론이거니와, 5-mC 등도 한번에 분석할 수 있다.

- 복잡한 구조 변이 및 translocation 등을 분석할 수 있다.

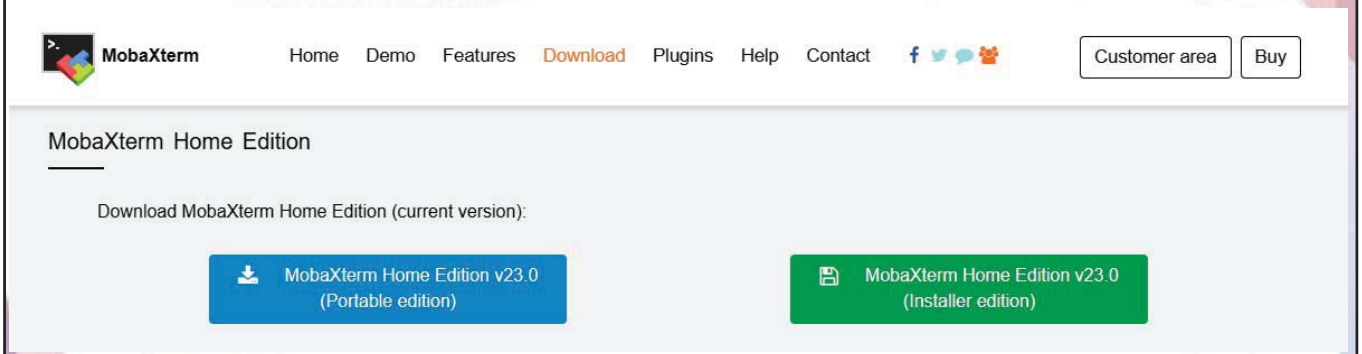
48

서버 접속하기 MobaXterm



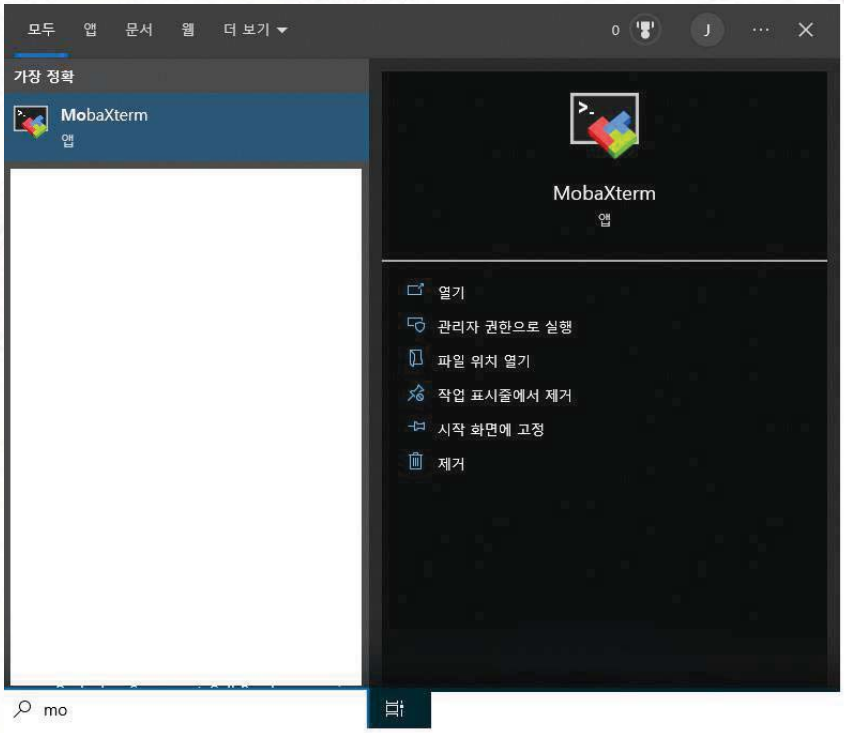
외부 서버에 접속할 수 있도록 해주는 프로그램

서버 접속하기 MobaXterm



인스톨러 에디션 다운 받고 실행해서 설치하시면 됨

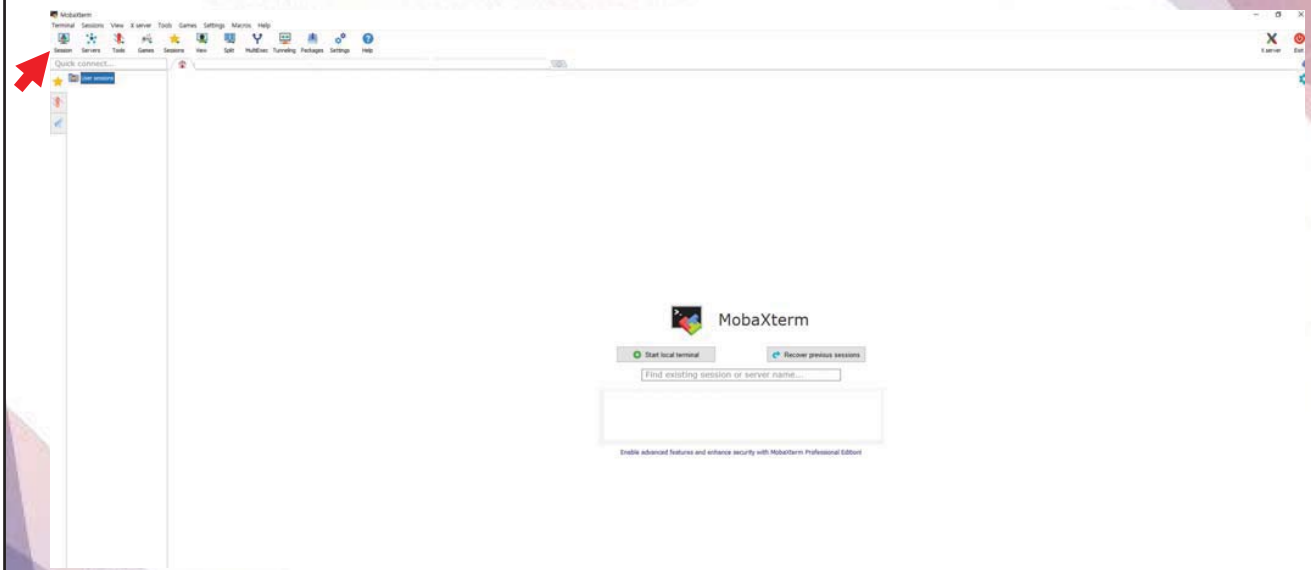
서버 접속하기 MobaXterm



검색 후 실행

51

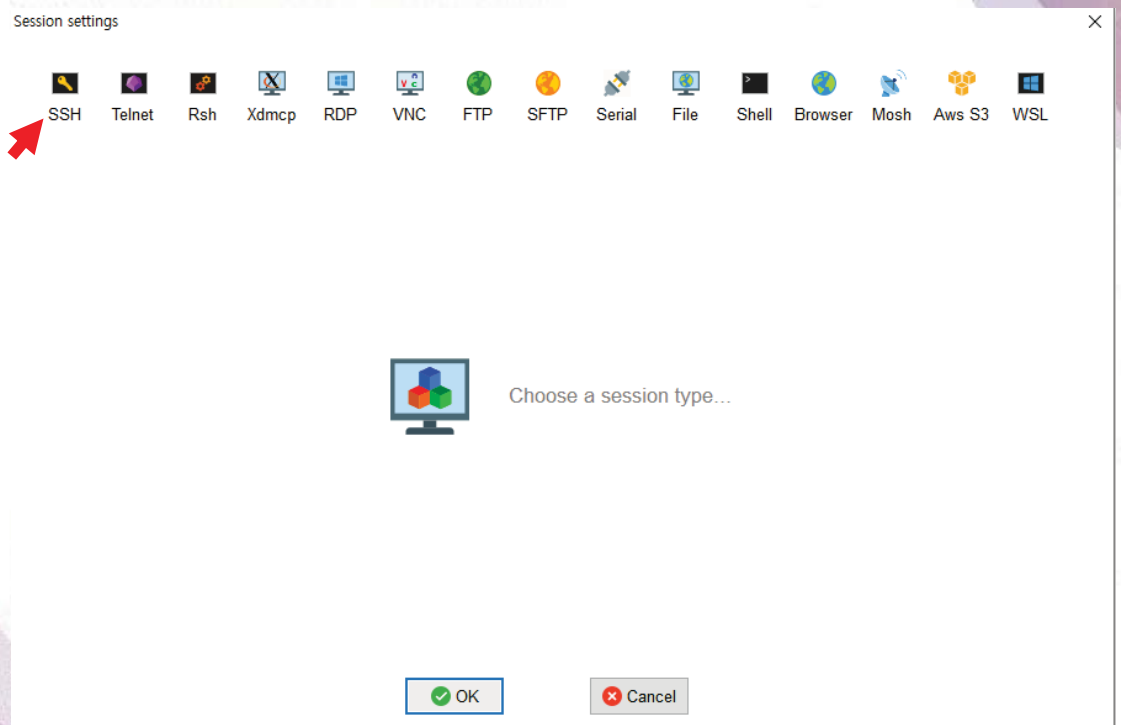
서버 접속하기 MobaXterm



왼쪽 위 “Session” 클릭

52

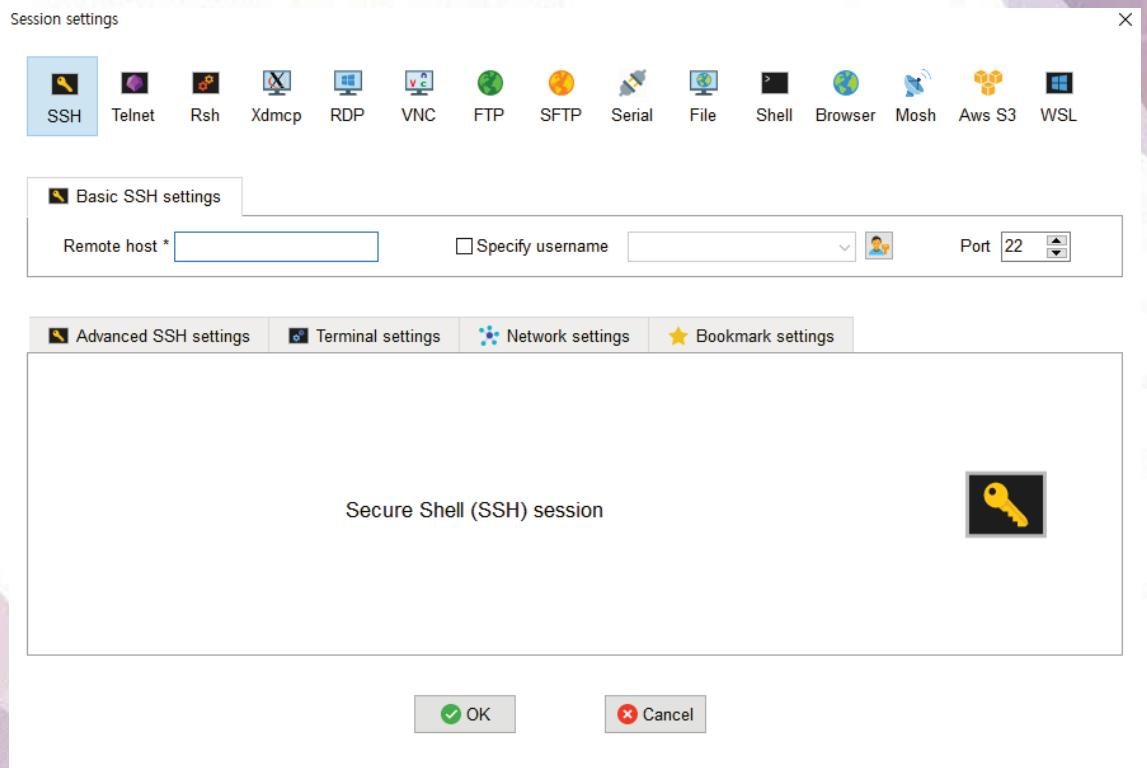
서버 접속하기 MobaXterm



SSH 클릭

53

서버 접속하기 MobaXterm



Specify username 클릭해서 체크 표시하기
나눠드린 종이에 적힌 Remote host, Port 입력 후 OK 클릭

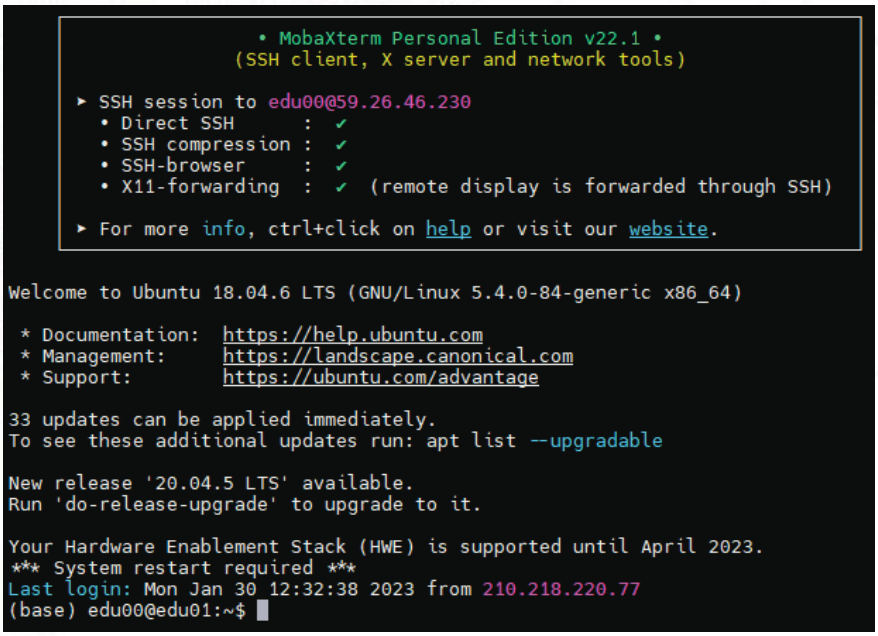
54

교육 서버 정보(KOBIC 제공)

비밀번호 입력 후 접속
비밀번호: kribb!23$kogo

55

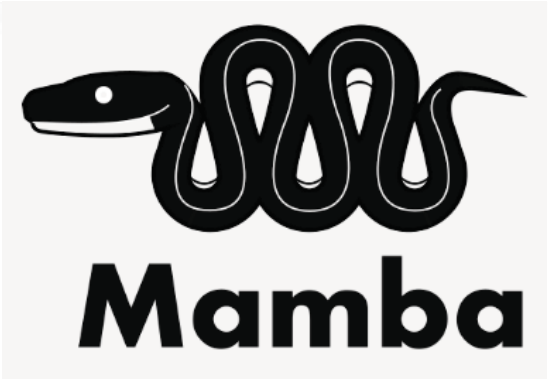
교육 서버 정보(KOBIC 제공)



여기까지 나오면 성공

56

Conda & mamba



https://docs.conda.io/en/latest/
https://github.com/mamba-org/mamba

패키지 매니저의 일종으로, 프로그램 설치를 아주 쉽고 빠르게 해결할 수 있도록 도와준다. (엔터 한 번이면 끝)

57

Conda 설치하기

 # 터미널에 아래와 같이 입력하기

wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
 # conda installer 다운로드 됨

chmod +x Miniconda3-latest-Linux-x86_64.sh
bash Miniconda3-latest-Linux-x86_64.sh
 # 실행 권한 부여 및 설치 실행

</path_to_your_conda>/bin/conda init
source ~/.bashrc
 # 어디서든 conda 사용할 수 있게 해줌

58

Conda 환경 설정

환경 = 격리된 프로그램 설치 공간
Restriction enzyme도 쓸 때마다 buffer 바꿔야 하는 것처럼,
프로그램들도 상황마다 필요로 하는 것들이 달라 환경 재설정 필요

 # 터미널에 아래와 같이 입력하기

conda update -n base -c defaults conda
conda config --add channels conda-forge
conda config --add channels bioconda
 # 다운 받을 채널에 연결해줌

conda install mamba -n base -c conda-forge
 # conda 대신 mamba 쓸 수 있도록 mamba 설치하기

59

Mamba 설치 및 프로그램 설치

conda create -n assembly
conda activate assembly
 # 이름이 assembly 인 새로운 conda 환경 형성 및 활성화

sudo apt-get install datamash

mamba install -c bioconda -c conda-forge assembly-stats bioawk hifiasm svim-asm circos minimap2 samtools sniffles
 # 다양한 genome assembler와 구조변이 분석 프로그램 설치하기

60

Public data download

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...], AND ADAM M. PHILLIPPY +90 authors Authors Info & Affiliations

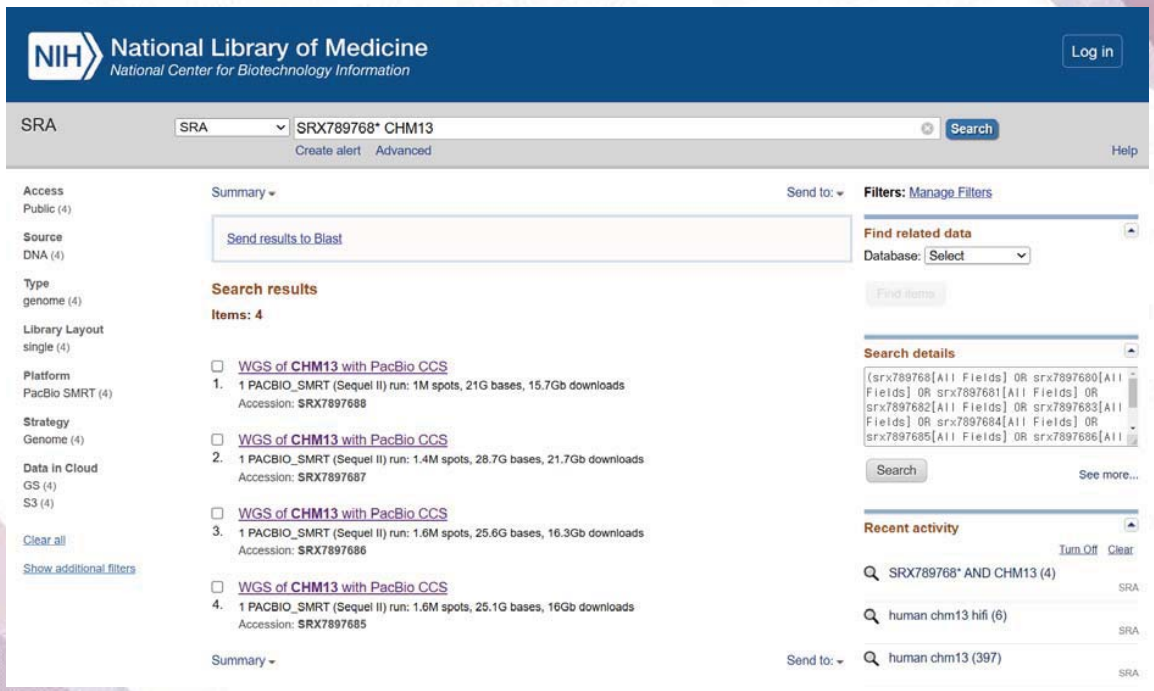
SCIENCE · 31 Mar 2022 · Vol 376, Issue 6588 · pp. 44-53 · DOI: 10.1126/science.abj6987

29,052 34

[The complete sequence of a human genome | Science]

61

Public data download



62

Public data download

SRX7897688: WGS of CHM13 with PacBio CCS
1 PACBIO_SMRT (Sequel II) run: 1M spots, 21G bases, 15.7Gb downloads

Design: DNA extracted from cultured adherent cells with modified Gentra Puregene protocol (C Baker, KM Munson & EE Eichler, Univ of Wash); sheared on Megaruptor to 25kb, library constructed with SMRTbell Template Prep Kit 1.0 (PN: 1000-222-300), size selection with SageELF using the 0.75% Agarose Cassette (ELD7510) and the SageELF 1-18kb version2 Cassette Definition Protocol to generate 15kb and 20kb libraries (P Peluso, PacBio).

Submitted by: Pacific Biosciences

Study: WGS of CHM13 with PacBio CCS
PRJNA530776 • SRP190633 • All experiments • All runs
show Abstract

Sample: Human sample from CHM13htert cell line from Homo sapiens
SAMN03255769 • SRS798661 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: CHM13-CCS-20kb-m64062_190806_063919
Instrument: Sequel II
Strategy: WGS
Source: GENOMIC
Selection: size fractionation
Layout: SINGLE

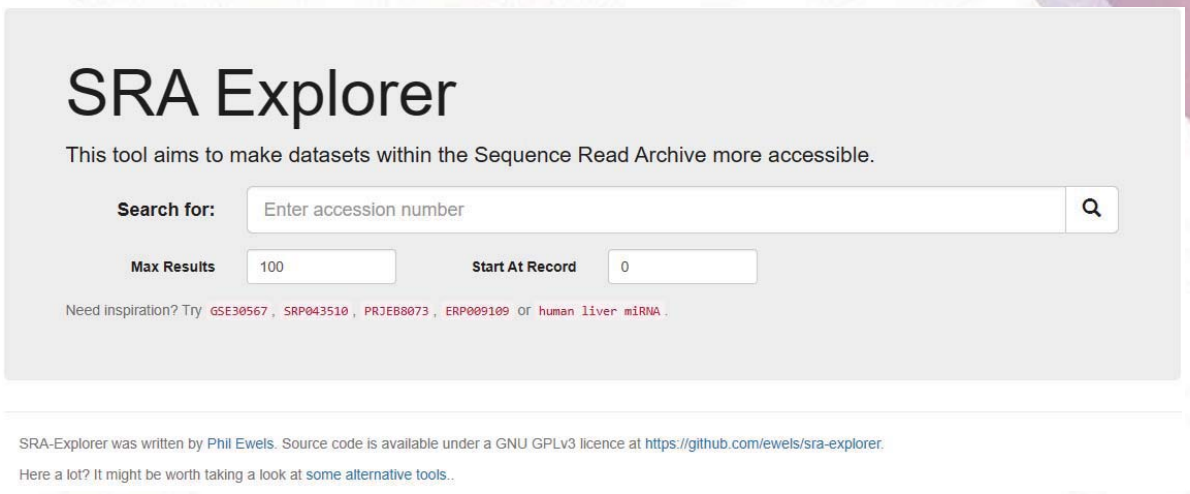
Runs: 1 run, 1M spots, 21G bases, 15.7Gb

Run	# of Spots	# of Bases	Size	Published
SRR11292120	1,012,393	21G	15.7Gb	2020-03-12

ID: 10331994

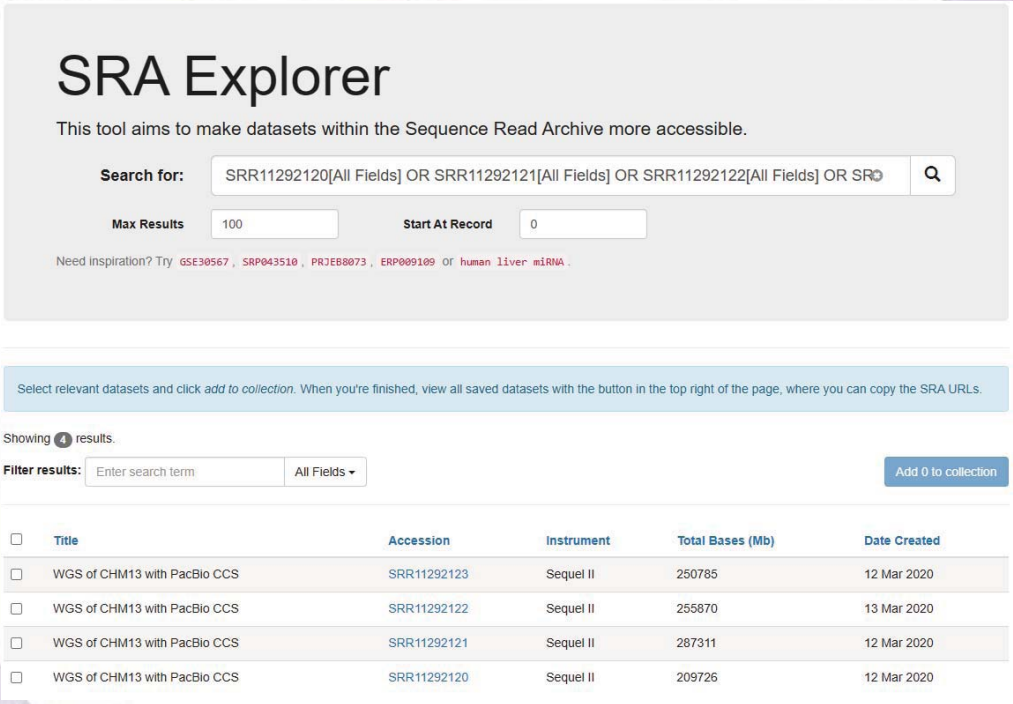
63

Public data download



64

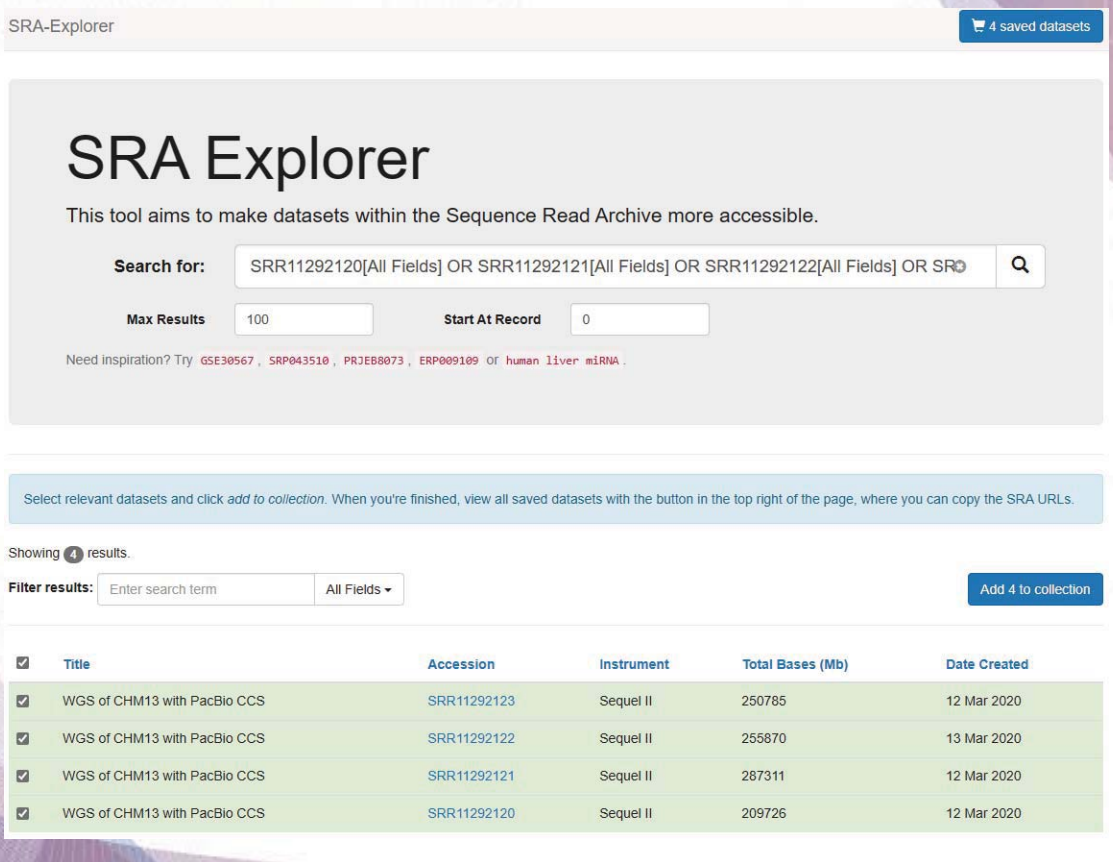
Public data download



SRR11292120 or SRR11292121 or SRR11292122 or SRR11292123

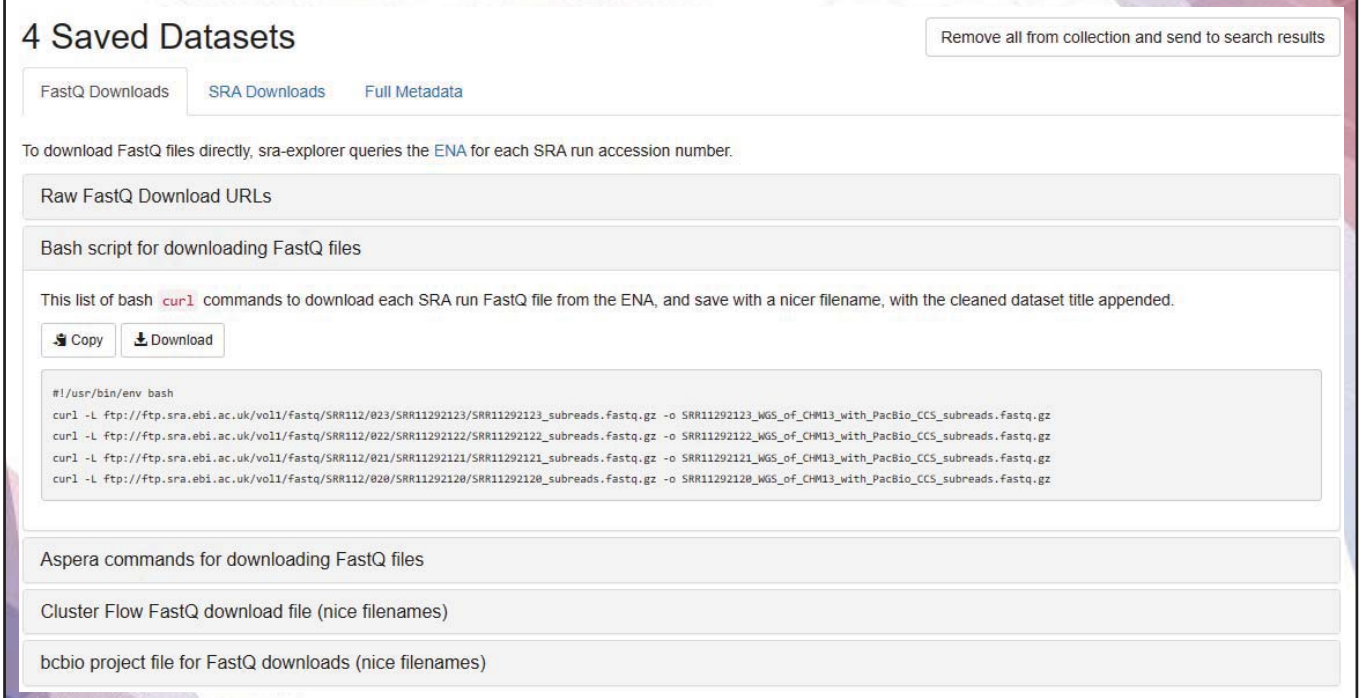
65

Public data download



66

Public data download



Command line에 복붙하면 됨

67

Genome assembly 실습(HiFi data)

Public data 활용(Han et al., submitted)
cd 쳐서 홈 디렉토리로 이동
cp /BiO/home/edu001/.bashrc .
source .bashrc
mamba activate assembly

68

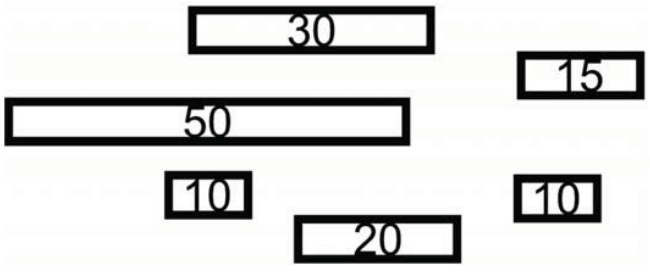
N50 & NG50

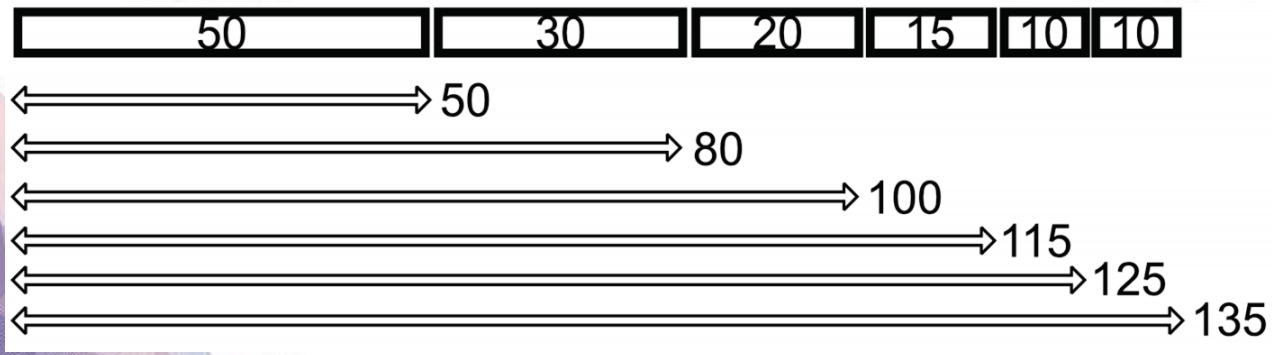
Read/contig/scaffold를 길이가 긴 것부터 정렬하고,
그 길이를 순차적으로 하나씩 더했을 때,
전체 길이의 딱 절반이 넘는 순간,
그에 해당하는 read/contig/scaffold의 길이
기타 metric은
https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics
참조

NG50는 전체 길이 대신 **알려진 genome size**를 활용함.

69

N50 & NG50





70

Quality 확인하기

assembly-stats ${PREFIX}.h1.fa ${PREFIX}.h2.fa

71

BUSCO

Benchmarking Universal Single-Copy Orthologs

각 lineage별로 알려진 single-copy 유전자가 얼마나 잘 assemble되었는지 확인해서 assembly quality 확인

요샌 다 높게 나온다지만, 그래도 논문에 쓰는 필수 stat 중 하나

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs | Bioinformatics | Oxford Academic (oup.com)

72

Quality 확인하기

```
busco --list-datasets
busco -i ${PREFIX}.h1.fa -c 3 -o ${PREFIX} -m genome -l
nematoda_odb10
# or
# busco -i ${PREFIX}.h1.fa -c 3 -o ${PREFIX} -m genome --auto-
lineage
# busco -i ${PREFIX}.h1.fa -c 3 -o ${PREFIX} -m genome --auto-
lineage-euk
# busco -i ${PREFIX}.h1.fa -c 3 -o ${PREFIX} -m genome --auto-
lineage-prok

# -i : 인풋 파일 이름
# -c : 동시 연산 개수
# -o : 아웃풋 파일 이름
# -m : 모드(유전체, 전사체 등등)
# -l : 리니지 정보
```

73

Visualization

```
STRAIN1=ALT1.chr5
REF1=Celegans.chr5.fa
TYPE1=contig
echo "line,length,type,coverage" > length.csv
LEN1=`bioawk -c fastx '{sum+=length($seq)}END{print sum}'
$REF1`
cat $REF1 | bioawk -c fastx -v line="$STRAIN1" '{print
line","length($seq)","length($seq)}' | sort -k3rV -t "," | awk -F "," -v
len="$LEN1" -v type="$TYPE1" 'OFS=","{ print
$1,$2,type,(sum+0)/len; sum+=$3 }' >> length.csv
```

74

Mamba 설치 및 프로그램 설치

conda create -n assembly2
conda activate assembly2
mamba install -c bioconda -c conda-forge -c r assembly-stats bioawk shasta canu hifiasm busco svim svim-asm r-ggplot2 r-reshape2 r-tidyverse
 # 실제로 사용한 스크립트

conda activate assembly2
 # 입력 후 왼쪽의 (base)*가* (assembly)*로 변경된 걸 확인*

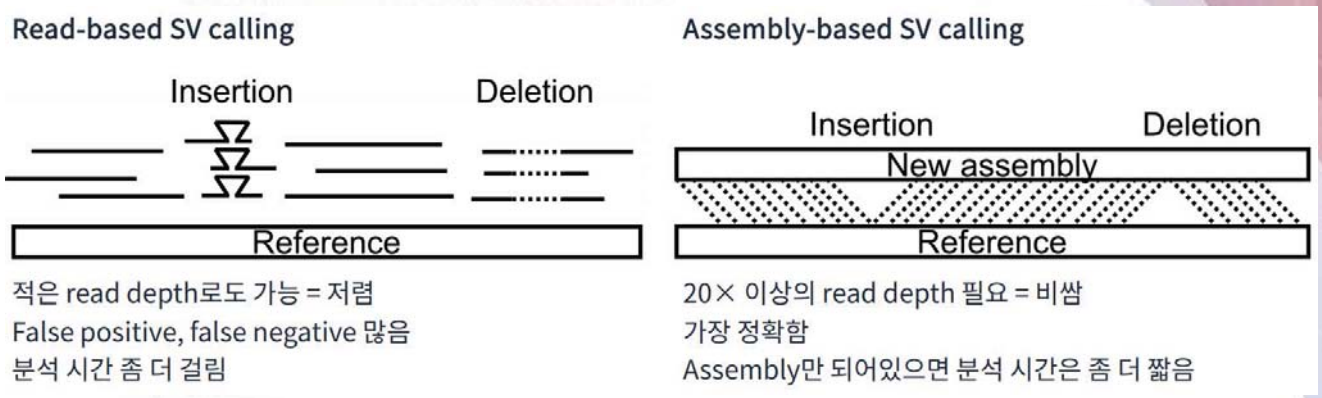
75

Visualization

# R 에서 작업하면 됨. 터미널에 R 입력.
setwd("length.csv")
library(ggplot2)
contig_cumulative_sum_df <- read.csv("length.csv", header = TRUE)
contig_cumulative_sum_df$type <- factor(contig_cumulative_sum_df$type, levels=c("scaffold", "contig")) # or any
other assembly types
plot <- ggplot(data=contig_cumulative_sum_df, aes(x=coverage, y=length/1000000, color=line)) +
geom_vline(xintercept = 0.5, linetype="dotted", size=0.5) +
xlim(0, 1) +
geom_step(aes(linetype=type)) +
labs(x = "Cumulative coverage", y = "Length (Mb)")
pdf("coverage.pdf",width=4,height=3,paper='special')
print(plot)
dev.off()

76

SV calling



최근 전략: population study라면, 가장 유전적으로 거리가 먼 개체들의 assembly를 얻어 backbone 만들고, 그 뒤에 다수 개체의 low-depth long-read data 확보해 집단 분석에 사용함

77

SV calling 실습

```
minimap2 -a -x asm5 --cs -r2k -t 3 reference_genome.fa ${PREFIX}.h1.fa | samtools sort -m4G -@ 3 -O BAM -o ${PREFIX}.h1_to_reference_genome.bam

samtools index ${PREFIX}.h1_to_reference_genome.bam

svim-asm haploid ${PREFIX} ${PREFIX}.h1_to_reference_genome.bam reference_genome.fa
```

78

SV calling 실습

grep "DEL"
grep "INS"
grep "INV"
grep "DUP"
grep "DUP:TANDEM"
grep "BND"
등을 활용해 다양한 variant 정보 확인 가능

79

BED format

가장 기본 틀:
1번 컬럼: chromosome (다른 파일에 있는 이름과 동일해야 함)
2번 컬럼: 시작 위치(숫자)
3번 컬럼: 끝 위치(숫자)
컬럼과 컬럼 사이는 tab으로 구분되어야 하며, 정렬 되어 있어야 bedtools 적용 가능

80

BED format

가장 기본 틀:
1번 컬럼: chromosome (다른 파일에 있는 이름과 동일해야 함)
2번 컬럼: 시작 위치(숫자)
3번 컬럼: 끝 위치(숫자)
컬럼과 컬럼 사이는 tab으로 구분되어야 하며, 정렬 되어 있어야 bedtools 적용 가능

81

경청해주셔서 감사합니다

충남대학교 생명정보융합학과
조교수 김준
junkim@cnu.ac.kr

82