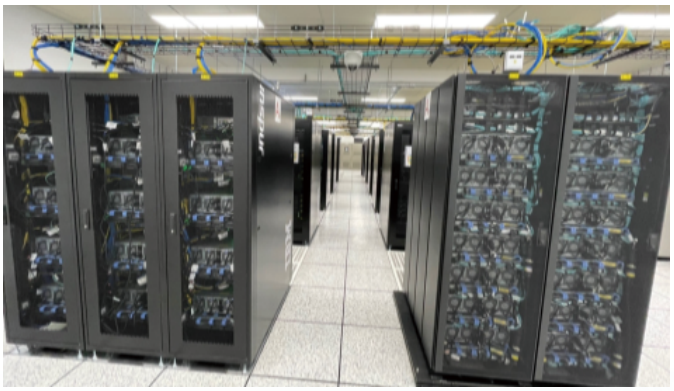
바이오데이터팜 개요

- **목　　적 |** 바이오 빅데이터의 안전한 보관 및 분양 / 활용성 제고를 위한 사용자 분석 지원
- **구　　성 |**

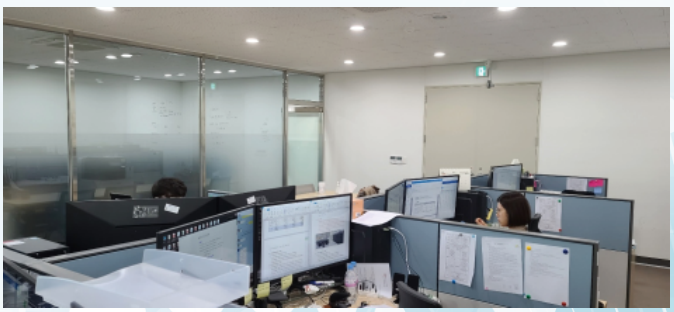
HPC	데이터분석을 위한 컴퓨팅 환경, 연산서버, 스토리지 제공
분석플랫폼	규제자유특구사업 실증 R&D로 개발되는 오믹스 분석 플랫폼 제공
데이터	울산만명 게놈데이터 보유 및 분양신청기업에게 기초분석결과 제공
의사결정기구	인체유래물은행 운영(데이터 분양심사), 기관생명윤리위원회 운영

- **위　　치 |** 울산광역시 테크노산업로 55번길 14
- **현　　황 |**

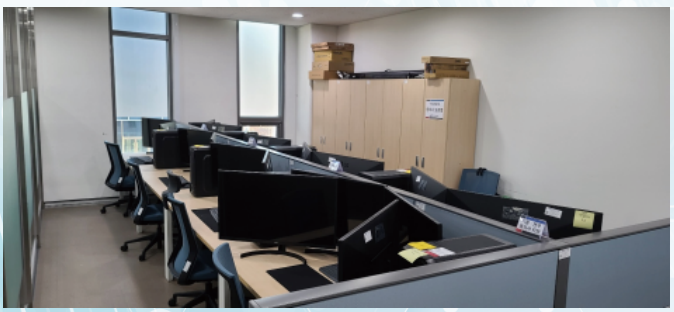
- 바이오데이터팜 장비도입, 시설구축 완료(‘22.1.)
- 인체유래물은행 설립, 바이오데이터팜 개소(‘22.3.)
- 울산만명 게놈데이터 이관(‘22.4.)
- 실증R&D사업 착수, 데이터분양심사(‘22.7.~)



데이터 연산 및 저장장비



인체유래물은행 사무국



데이터프리존

바이오데이터팜 주요 장비 및 데이터

바이오데이터팜 주요 장비

장 비	용 도
연산클러스터 마스터노드	연산 클러스터의 계정과 작업 관리용 매니지먼트서버
연산클러스터 계산노드서버(2way,8way)	대용량 바이오빅데이터의 초고속 생정보 분석 연산 이론상 총 성능 : 663.4TFLops
고가용성 GPU 서버	*A100 칩셋 GPU서버 GPU 전용 프로그램 기반 바이오빅데이터 분석 이론상 총 성능 : 7,488TFLops
고성능 메모리 연산서버	대용량 게놈데이터의 대량 비교를 위한 고용량 메모리 연산
연산클러스터 파일서버	RawData 및 결과 데이터 저장 및 연산 스토리지
네트워크	클러스터 노드와 클러스터 파일 서버간 200G 네트워크 연동으로 네트워크 I/O 최소화

데이터 정보

구 분	참여자 수(명)
공개된 일반인	3,300
공개된 질환군[1]	107
규제특례적용공개가능 일반인[2]	1,002
규제특례적용공개가능 질환군[1,2]	5,525
합 계	**10,033**

* 만명 게놈 프로젝트 참여자 게놈 정보 10,033건
(건강인, 질환자의 전장 유전체)

* 만명 게놈 프로젝트 참여자의 임상정보 3,382건
(신체계측, 심폐기능, 혈액검사, 소변검사 등)

* 만명 게놈 프로젝트 참여자의 생활습관정보 3,465건
(질병력, 가족력, 음주, 흡연, 스트레스 등)

[1] 한 사람이 2개 이상의 질병을 가진 경우 각각 나누어 분류함
[1] 질환군 분류: 암, 대사질환, 심혈관질환, 신경계질환, 희귀질환

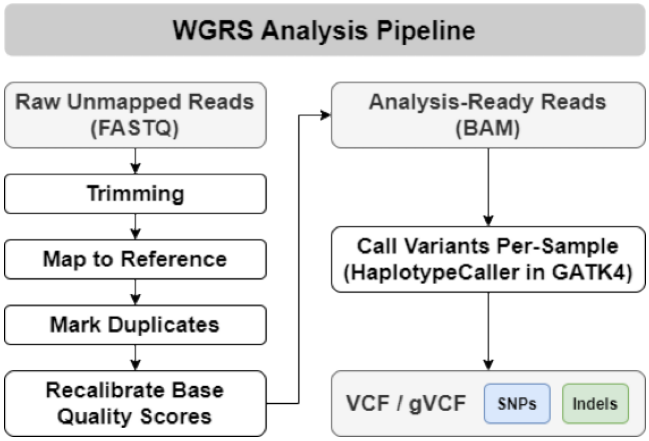
[2] 규제 자유 특구 후 해당 질환군 정보 공개 예정
[2] 질환군 분류에 따른 샘플 수는 개별 문의 필요

파일 정보

파일 종류	수 량	크 기	용 량
전장게놈서열 Raw FASTQ data	10,033	140GB	1,400TB
표준게놈에 매핑된 BAM file	10,033	120GB	1,200TB
개인별 유전변이 gVCF file	10,033	58GB	580TB
개인별 유전변이 VCF file	10,033	1.7GB	17TB
유전서열,맵핑,유전변이 QC file	10,033	1MB	10GB
설문결과 및 건강검진결과 Excel file	1	100MB	100MB
총용량			**3,200TB**

바이오데이터팜 분석 서비스 플랫폼

전장게놈 분석 파이프라인

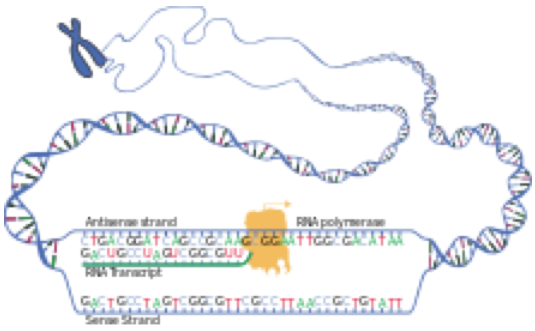


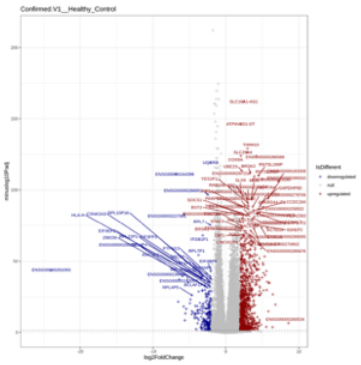
NGS 기술을 통해 얻은 개인의 시퀀싱 데이터를 기반으로 각 개인의 변이 정보(gVCF, VCF)를 추출. 개인의 변이 정보를 통해 희귀병 및 복합 질환을 유발하는 변이들을 확인하여 진단 키트 및 질병 치료제 개발 등에 도움을 줌.

Input/Output	Data	Format
Input	시퀀싱데이터	FASTA
Output	리드맵핑데이터	BAM
Output	변이데이터	gVCF, VCF

WGRS, Homogeneity, Heterogenity, GATK, Short Read, Variant Calling, FASTQ, VCF, gVCF, HaplotypeCaller, Picard, BWA

전령 RNA 분석 파이프라인





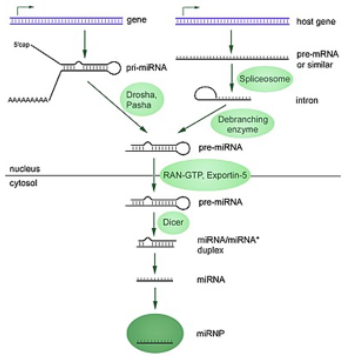
출처 위 | https://en.wikipedia.org/wiki/Messenger_RNA

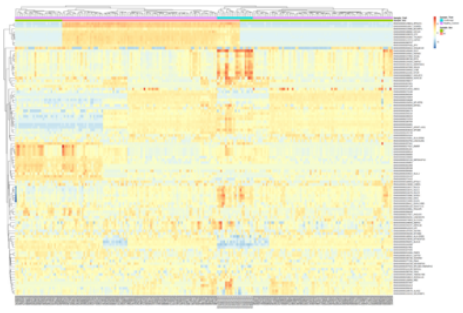
mRNA 정보를 기반으로 유전자 발현량을 분석함. 수집한 샘플들을 대상으로 그룹별 유전자 발현량을 측정, 차등으로 발현되는 유전자들을 제시. 또한, 차등 발현 유전자 세트에 대한 기능들을 함께 제공하여, 연구자가 생물학적 기전을 밝히거나 약물 타겟을 선정하는데 도움을 줌.

Input/Output	Data	Format
Input	mRNA 시퀀싱데이터	FASTQ
Output	차등 발현 유전자 정보 및 그림	TSV, PNG
Output	유전자 세트 기능 정보 및 그림	TSV, PNG

mRNA, Transcriptome, STAR, RSEM, DEG, DESeq2, GO, KEGG

miRNA 분석 파이프라인





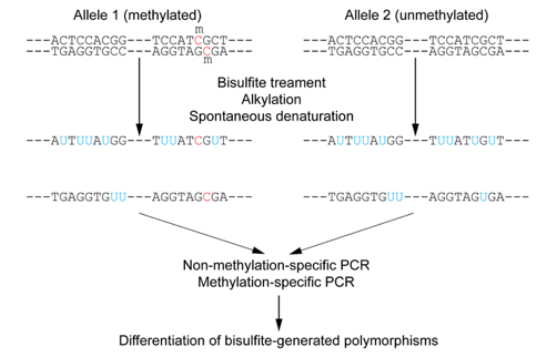
출처 위 | https://en.wikipedia.org/wiki/MicroRNA

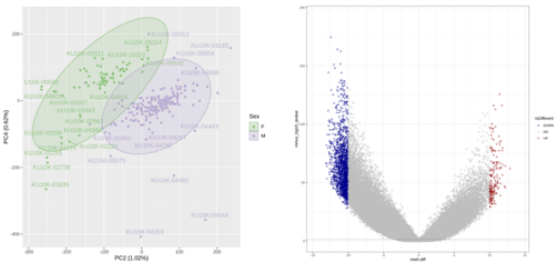
miRNA는 전령 RNA이 단백질로 번역되는 과정을 억제를 하여 전령 RNA의 발현량을 조절함. total RNA 또는 targeted miRNA 시퀀싱 데이터로부터 그룹별 miRNA 발현량을 측정하여 차등으로 발현되는 miRNA들을 제시. miRNA의 기능을 함께 제공하여, 연구자가 생물학적 기전을 밝히거나 약물 타겟을 선정하는데 도움을 줌.

Input/Output	Data	Format
Input	miRNA 시퀀싱데이터	FASTQ
Output	차등 발현 miRNA 정보 및 그림	TSV, PNG
Output	miRNA 기능 정보 및 그림	TSV, PNG

miRNA, Transcriptome, non-coding RNA, STAR, RSEM, DEG, DESeq2, GO, KEGG, miRBase

메틸레이션 분석 파이프라인





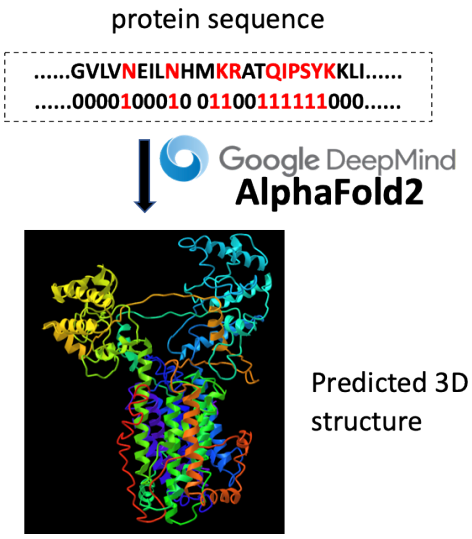
출처 위 | https://en.wikipedia.org/wiki/File:Wiki_Bisulfite_sequencing_Figure_1_small.pngpek160114_273

DNA 상의 CpG site에 붙어 있는 메틸기는 유전자 발현을 조절하는 후성유전학적 요소로 나이, 환경, 습관에 따라 변화됨. 샘플들의 그룹별 메틸화 정도를 측정하여 차등으로 메틸화된 CpG site를 제시. 이 분석을 통해 특정 질병에 대한 후성유전학적 매커니즘 분석함.

Input/Output	Data	Format
Input	Bisulfite 시퀀싱데이터	FASTQ
Output	차등 메틸화 CpG site 정보 및 그림	TSV, PNG

Methylation, CpG site, Epigenetics, Bisulfite Sequencing, Bismark, Bowtie2, Methyl kit, DMR

단백질 3차 구조 예측 파이프라인



Protein, 3D structure, Stability, Homology modeling, PDB, Alphafold2, Mutation

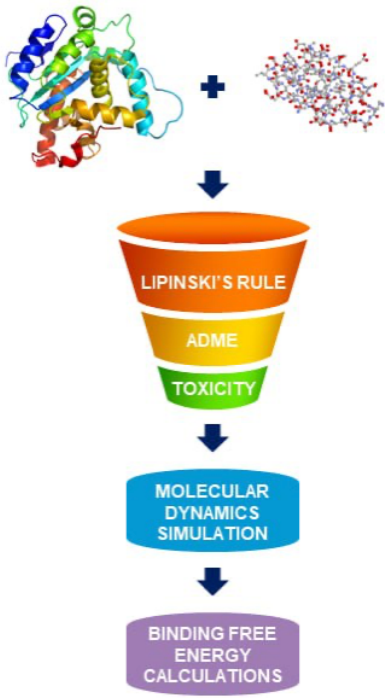
단백질의 아미노산 서열을 기반으로 구글 딥마인드에서 개발한 alphafold2를 통해 3차원 구조를 예측함. 또한, 해당 구조의 에너지 안정성, 구조적 타당성 등에 대한 여러 분석 결과들을(Modeller, DFIRE, Molprobity, Procheck 등) 제공함.

Input/Output	Data	Format
Input	단백질 서열	TXT
Input	돌연변이 리스트	Program argument
Output	예측된 3차원 구조	PDB
Output	구조 관련 스코어	TSV

출처 **단백질 그림 |** A proposed 3D structure for the protein TBC1D30, 17 November 2021, auther : 87poatoes, https://commons.wikimedia.org/wiki/File:TBC1D30_3D_structure.png

로고 | The Official Logo Of Google DeepMind, auther : google, http://www.dicetowernews.com/tag/deep-mind, https://commons.wikimedia.org/wiki/File:GoogleDeepMindLogo.png

단백질/단백질, 단백질/약물 도킹 시뮬레이션 파이프라인

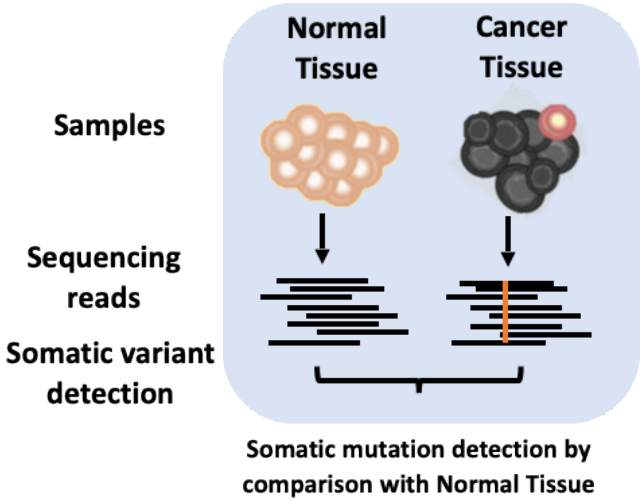


분자 도킹(docking) 기술은 분자 모델링 기법을 활용하여 최적의단백질/단백질와 단백질/약물 결합 형태를 계산 및 예측함. 또한, 도킹 기술에는 분자간 결합 형태를 예측하는 과정에서 열역학적으로 안정한 단백질/단백질와 단백체/약물 복합체를 형성하기 위한 결합친화도(affinity score)를 제시함.

Input/Output	Data	Format
Input	단백질 구조, 화합물 정보	PDB, SDF
Output (Optional)	에피토프 리스트	TSV
Output	단백질-단백질 복합체	PDB
Output	단백질-단백질 복합체	PDB
Output	결합친화도	TSV

Protein, Ligand, Protein-Ligand Docking, Protein-Protein Interaction, HADDOCK, PDB, Drug Development

Somatic 돌연변이 분석 파이프라인



암 조직 샘플에서 낮은 변이율로 존재하는 Somatic 돌연변이들을 확인 하는 파이프라인 입니다. 이 파이프라인은 정상 및 암샘플을 비교하거나 암샘플 하나로도 분석이 가능합니다. 이 분석을 통해 암을 유발하는 SNV와 같은 Pathogenic 변이들을 확인 할 수 있습니다.

Input/Output	Data	Format
Input	DNA 시퀀싱데이터	FASTQ
Output	Somatic 돌연변이 정보	VCF

WGS, Normal Tissue, Cancer Tissue, Heterogeneity, Somatic mutation, Mutect2, VarScan2

• 바이오데이터팜 클라우드 및 포털 서비스 |

서비스	분석 서비스	용 도
AI 클라우드 서비스	질병 발현 가능성 예측	만게놈 데이터를 활용해 당뇨, 고혈압, 고지혈증 등 각 질환에 대한 진단 및 예측
	생체 나이 추정 인공지능	바이오 빅데이터 기반 유전자 생체나이 추정 인공지능 모델을 사용하여 현재 나이와 유전자 생체 나이를 비교하여 건강 및 질병 위험도 분석
게놈 웹 포탈 분석 서비스	공개 데이터베이스 검색	기공개 데이터베이스를 대량 수집하여 사용자가 검색하고자 하는 바이오 Entity(예 : snp, gene)와 연관 된 바이오 정보를 제공
	공개 논문 검색	Pubmed 기반으로 수집된 대량의 논문을 데이터베이스화 하여 사용자가 빠르게 문헌 정보를 찾을 수 있는 검색 시스템
	공개 논문 기반 관련 유전자 검색 시스템	대량 수집된 논문 데이터베이스 기반으로 사용자가 입력한 유전자와 관련된 대량의 유전자를 관련성 기준으로 순위 별로 보여주는 시스템

울산 게놈서비스산업 규제자유특구사업

- **목　　적 |** 게놈기반 바이오빅데이터 구축·활용을 통한 정밀의료, 헬스케어산업육성
- **특　　례 |**

  ❶ 연구자가 연구결과 얻은 유전정보의 바이오데이터팜 제공 특례 허용(생안법 근거부재)
  ❷ 바이오데이터팜이 수집한 유전정보의 연구목적으로 기업활용 특례 허용(생안법 42조 1항)

- **사업기간 |** '21. 1. ~ '23. 3.
- **사 업 비 |** 37,503백만원(국19,960, 시15,494, 민2,048)
- **전담/주관 |** 한국산업기술진흥원/울산정보산업진흥원
- **참여기관 |**

  기업 솔트룩스, 테이크오프, 오상헬스케어, 누리바이오, 원드롭, 메디에이아이, 에이치앤비지노믹스, 클리노믹스, 힐릭스코, 타이로스코프, 에이테크

  병원 울산대학교병원, 울산병원

  대학 울산과학기술원

- **사업내용 |**

  인프라구축 바이오빅데이터 수집, 저장을 위한 컴퓨팅인프라 구축, 게놈분석 플랫폼개발

  실증R&D ❶ 질환 진단 및 치료를 위한 진단마커 개발(심혈관, 우울증, 복합만성질환),
  ❷ 감염성질환 팬데믹 대응 플랫폼 구축

  사업화 시제품제작, 특허/인증, 사업화/마케팅 등 상용화를 위한 기업지원