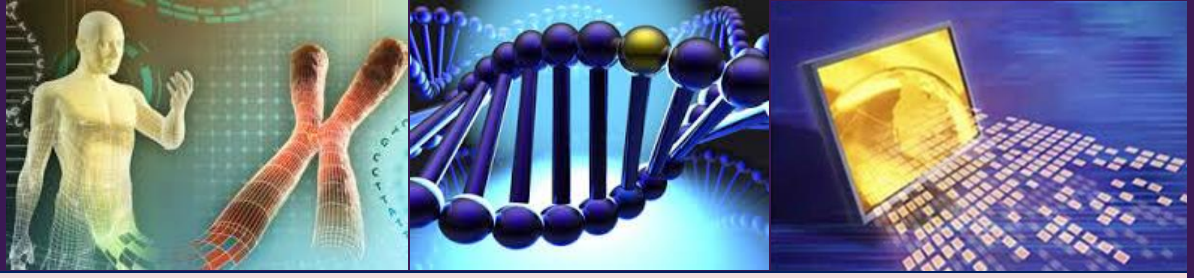
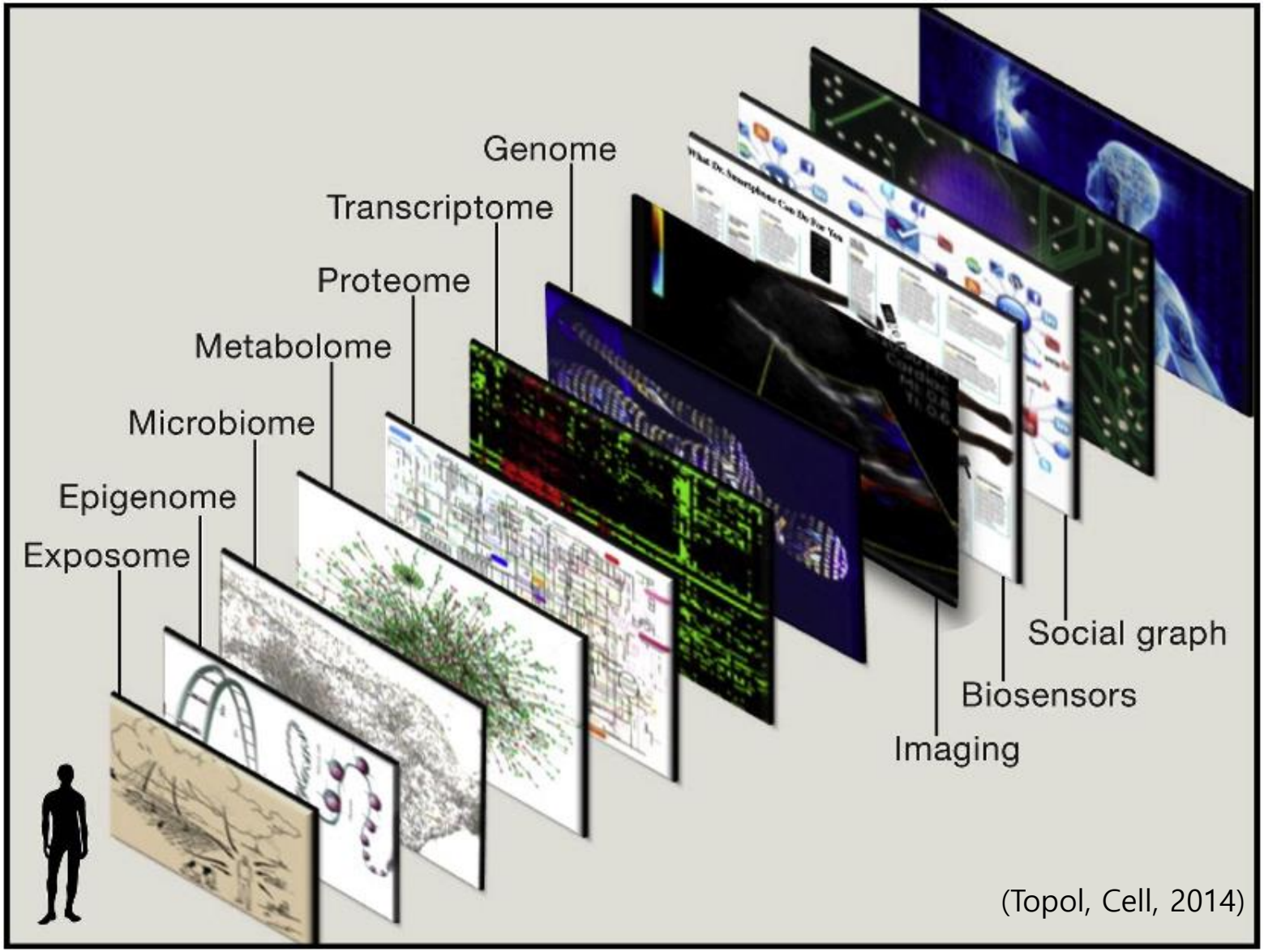


유전체 빅데이터 현황 및 응용사례

차의과학대학교 분당차병원

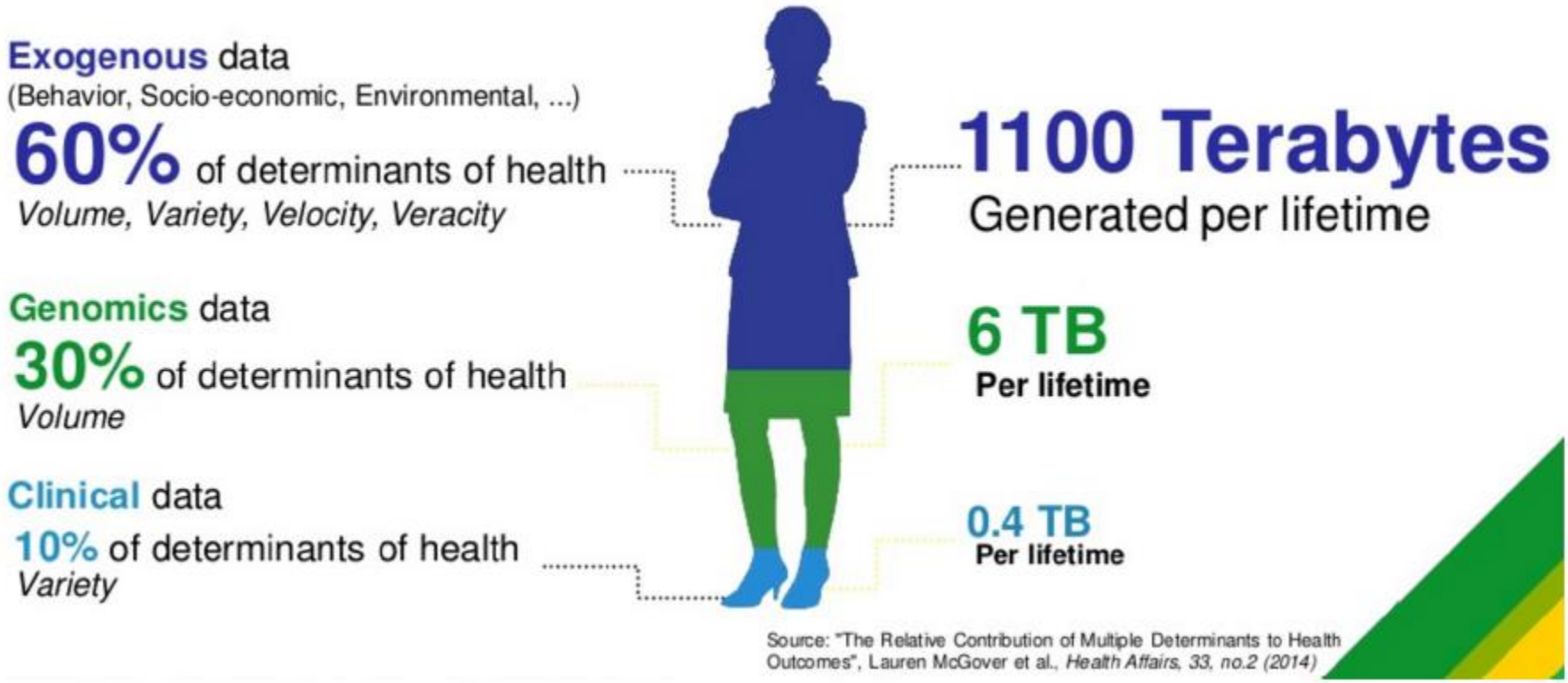
정제균

Individualized Medicine from Prewomb to Tomb

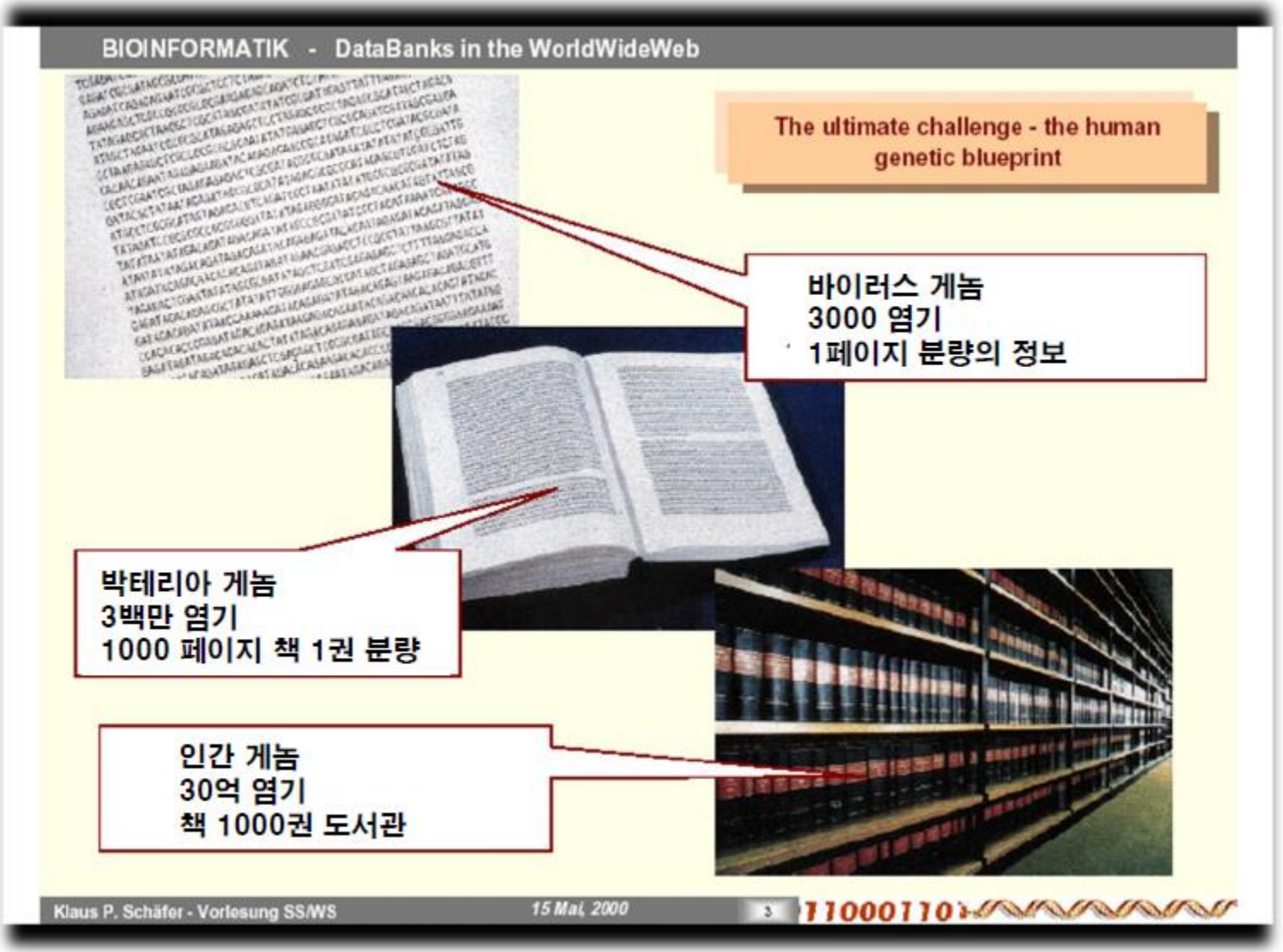


(Topol, Cell, 2014)

인간이 평생 만들어내는 데이터의 종류와 크기



유전체 지도


BIOINFORMATIK - DataBanks in the WorldWideWeb
The ultimate challenge - the human genetic blueprint
바이러스 게놈
3000 염기
1페이지 분량의 정보
박테리아 게놈
3백만 염기
1000 페이지 책 1권 분량
인간 게놈
30억 염기
책 1000권 도서관
Klaus P. Schäfer - Vorlesung SS/WS
15 Mai, 2000
3
11000110

DNA Chip vs. NGS

DNA Chip (microarray)

IT·BT의 발전

IT

1990 년 2001 년 2014 년

인간 유전체 프로젝트

$2.7 billion

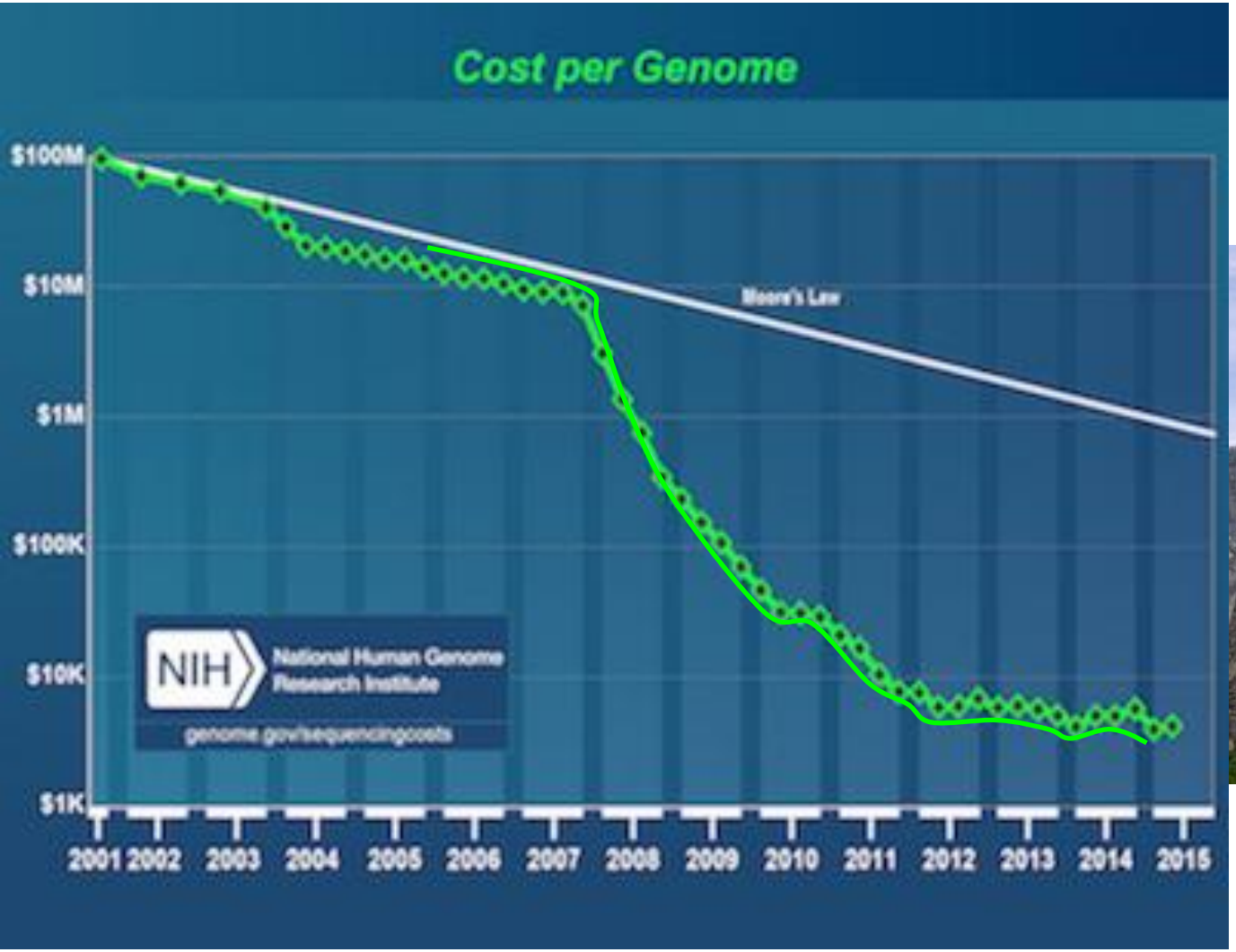
10년 소요

BT

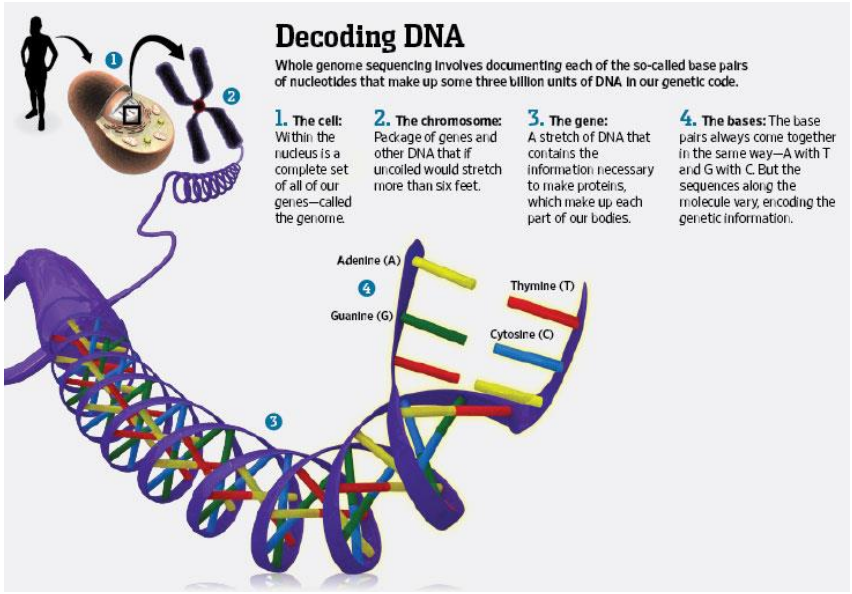
$1,000

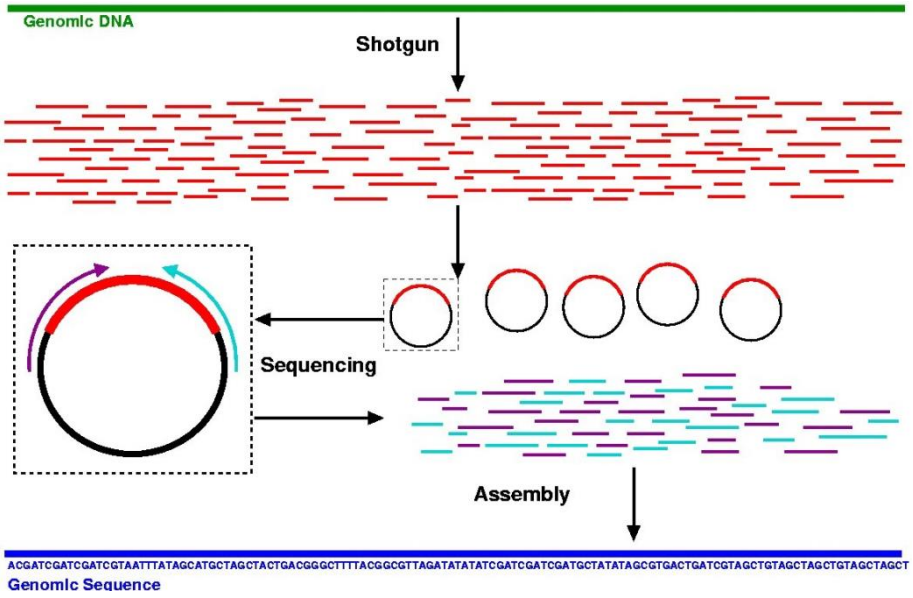
수시간 소요

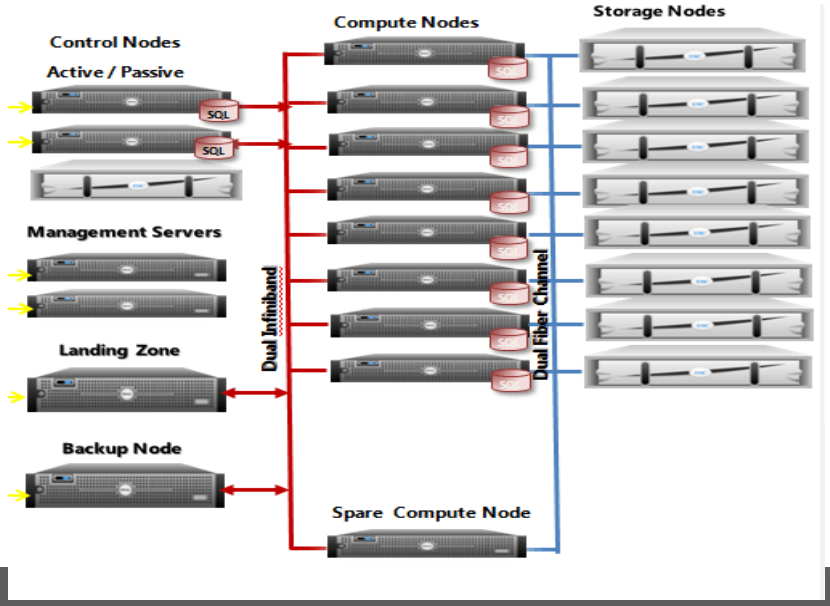
유전체 시퀀싱 가격 추이


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

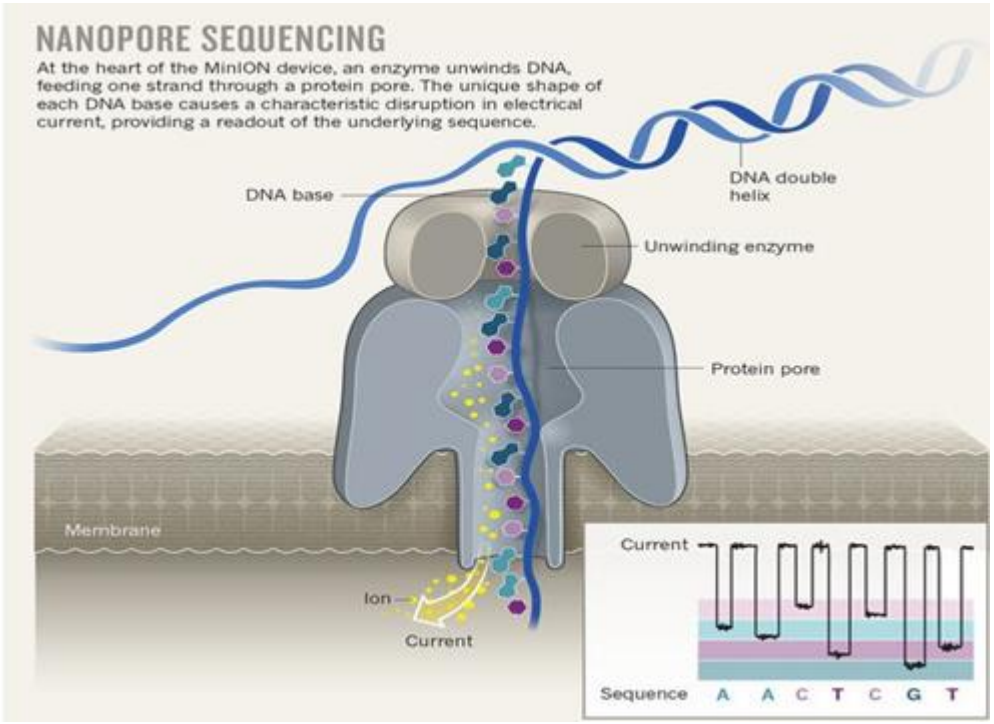
컴퓨팅


Decoding DNA
Whole genome sequencing involves documenting each of the so-called base pairs of nucleotides that make up some three billion units of DNA in our genetic code.
1. The cell: Within the nucleus is a complete set of all of our genes—called the genome.
2. The chromosome: Package of genes and other DNA that if uncoiled would stretch more than six feet.
3. The gene: A stretch of DNA that contains the information necessary to make proteins, which make up each part of our bodies.
4. The bases: The base pairs always come together in the same way—A with T and G with C. But the sequences along the molecule vary, encoding the genetic information.
Adenine (A)
Thymine (T)
Guanine (G)
Cytosine (C)


Genomic DNA
Shotgun
Sequencing
Assembly
ACGATCGATCGATCGTAATTTATAGCATGCTAGCTACTGACGGGCTTTTACGGCGTTAGATATATATCGATCGATCGATGCTATATAGCGTGACTGATCGTAGCTGTAGCTAGCTGTAGCTAGCT
Genomic Sequence


Control Nodes
Active / Passive
SQL
SQL
Management Servers
Landing Zone
Backup Node
Compute Nodes
Dual Infiniband
Dual Fiber Channel
Storage Nodes
Spare Compute Node

나노포어 시퀀서



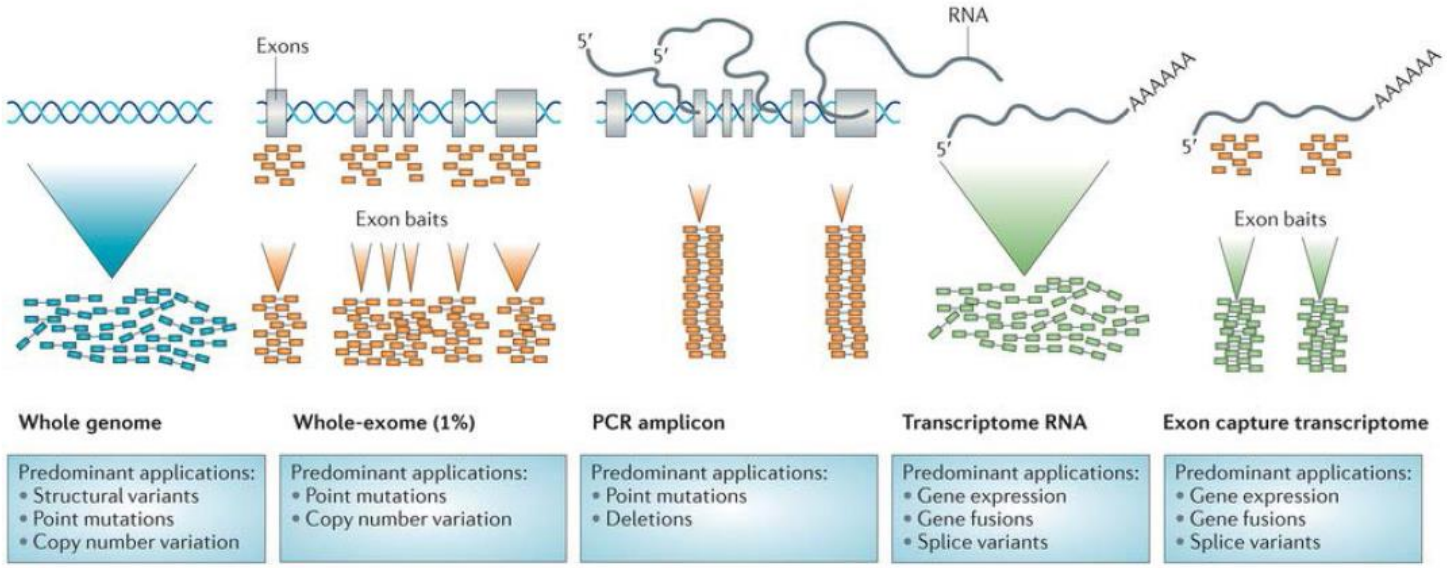
MinION Mkl: portable, real time biological analyses

MinION

나노포어 막단백질을 이용 전류 신호 측정을 통해서 ATGC 염기서열을 알아내는 방식

연구 대상별 유전체 분석 방법

WGS **WES** **Targeted Exome-seq**



	Whole genome	Whole-exome (1%)	PCR amplicon	Transcriptome RNA	Exon capture transcriptome
	Predominant applications: • Structural variants • Point mutations • Copy number variation	Predominant applications: • Point mutations • Copy number variation	Predominant applications: • Point mutations • Deletions	Predominant applications: • Gene expression • Gene fusions • Splice variants	Predominant applications: • Gene expression • Gene fusions • Splice variants
Target Size:	3Gb	30Mb	1Mb (200 genes)	~30Mb	3Mb
Scale:	1	0.01	0.0003		

File size for human germline WGS (30X)

- Image Data 16TB
- BaseCall/Quality score data 200GB
- Final Alignment output 1 TB

(Nature Reviews Drug Discovery, 2013)

유전자를 통한 질병 예측?

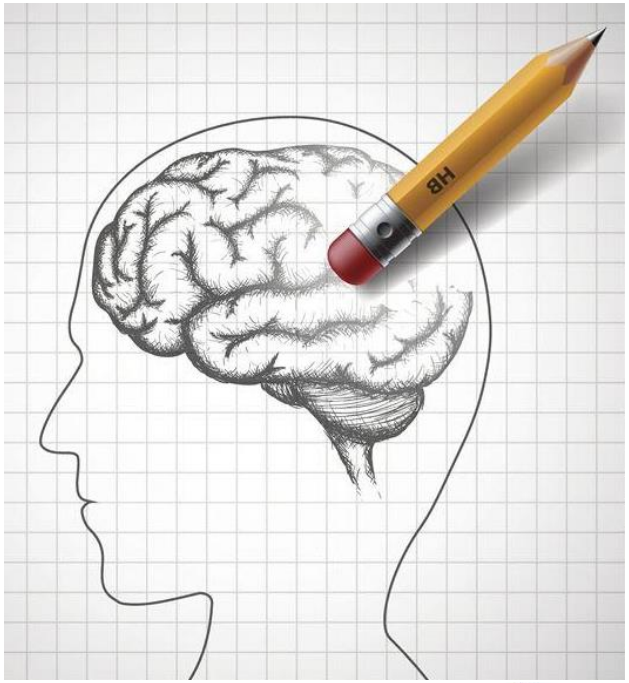
체온조절과 섭식조절을
담당하는
뇌의 시상하부에 위치

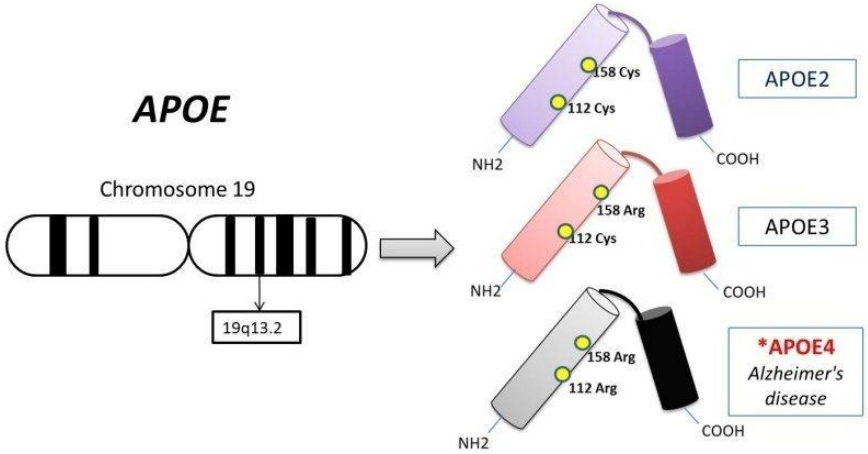
FTO 유전자는 크로모좀 16q12.1에 위치한 유전자입니다.
여러 비만 관련 유전자 중 가장 먼저 발견되어, '비만 유전자'로 불립니다.
BMI(Body Mass Index), 비만위험도, 제2형 당뇨병과 관련이 있습니다.

특정 변이가 생길시
식욕증가, 포만감 감소, 지방세포 에너지 소모 비율이 감소하고
충동적 성격이 강해지며, 지방이 많은 음식을 더 많이 섭취하게 됩니다.

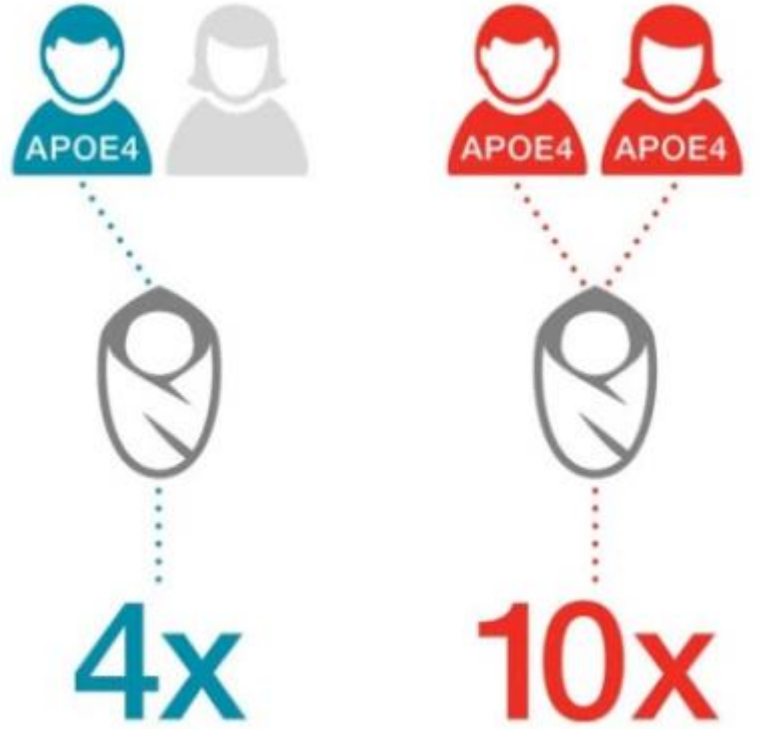
(이원다이애그노믹스 제공)

유전자를 통한 질병 예측?



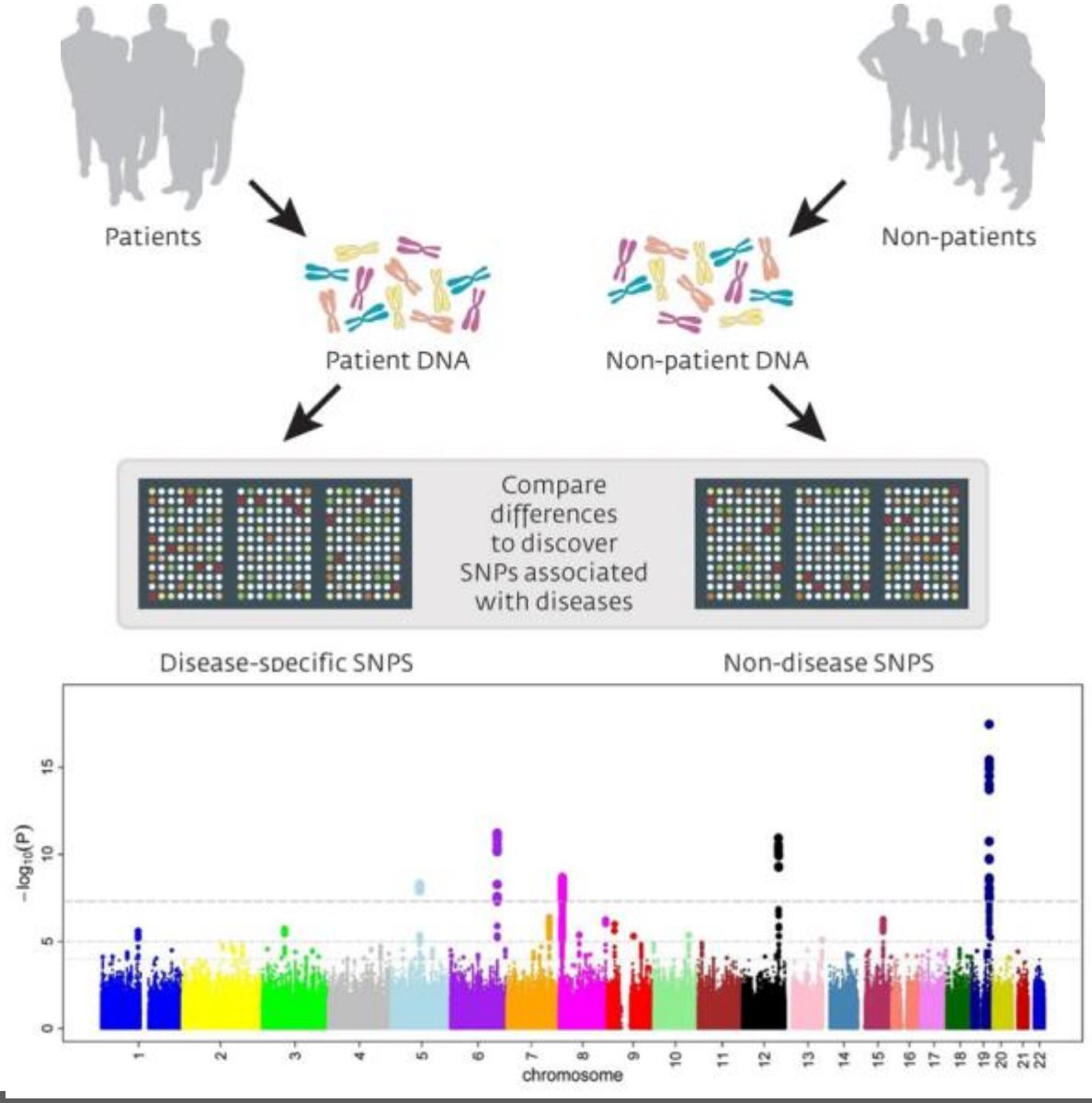


부모 양쪽이 모두 E4변이 일때 무려 10배의 확률이 올라감



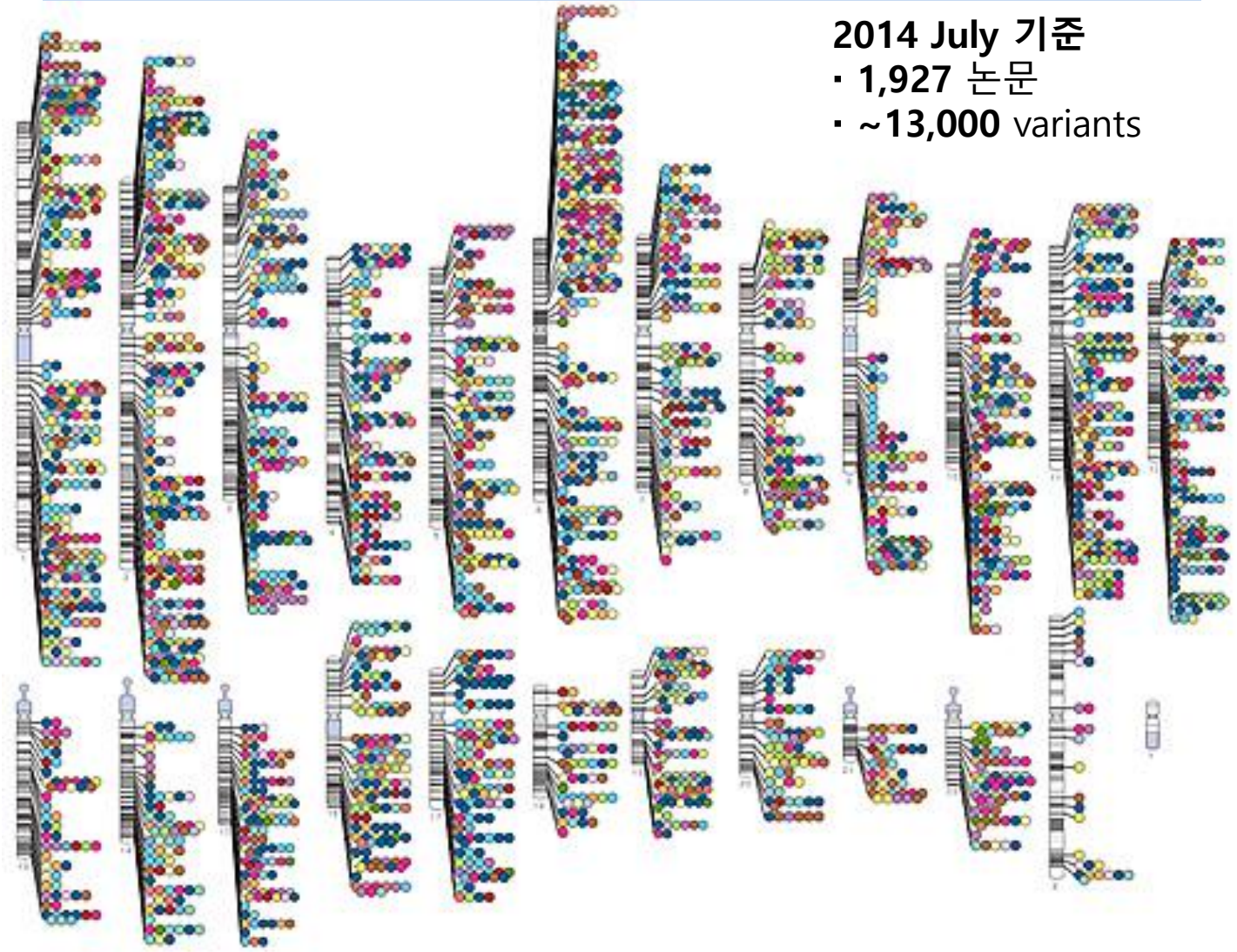
(Cure Alzheimer's Fund)

전장유전체 연관분석 (GWAS: Genome-wide association studies)



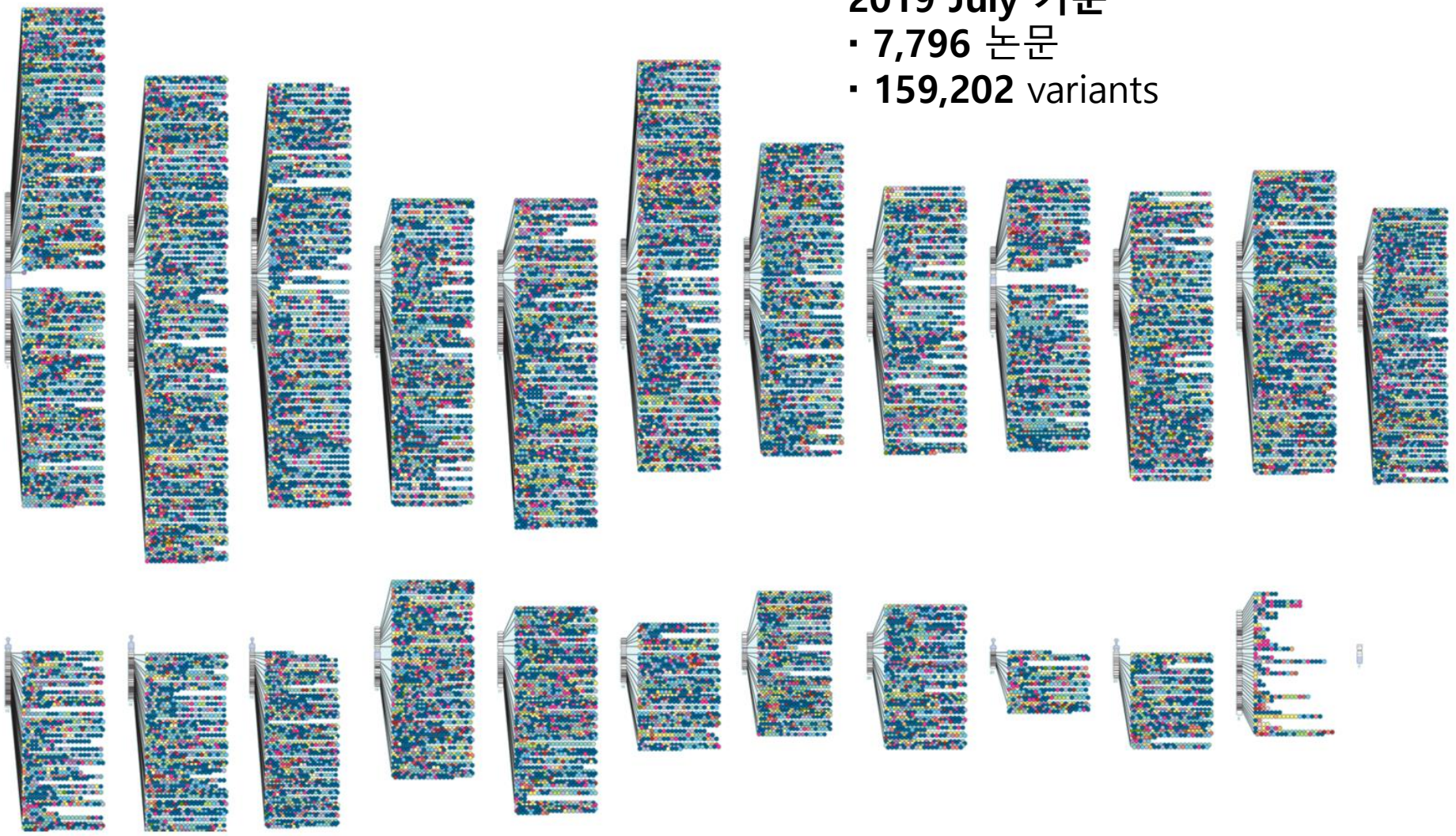
전장유전체 연관분석

GWAS (Genome-wide association studies) Catalog



2014 July 기준

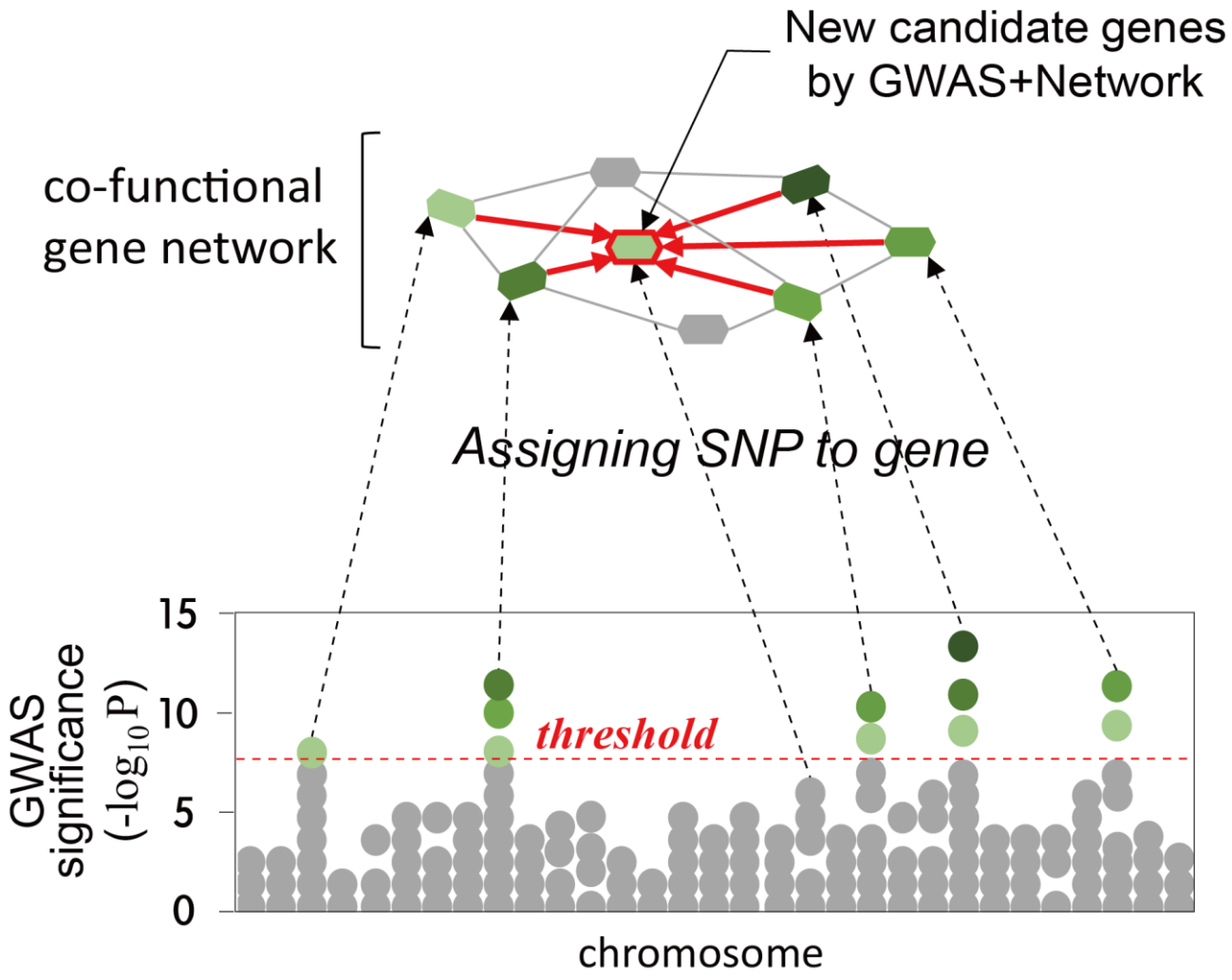
- **1,927** 논문
- **~13,000** variants



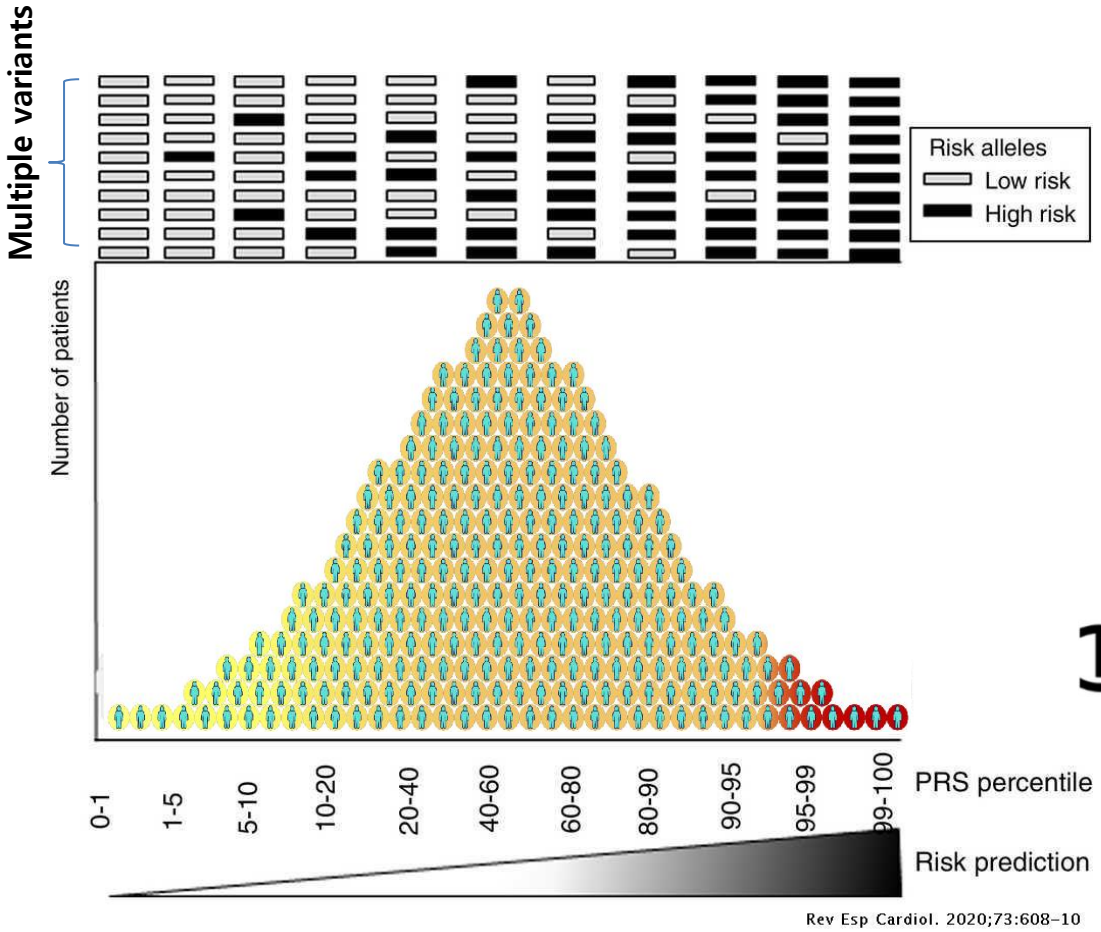
2019 July 기준

- **7,796** 논문
- **159,202** variants

Polygenic causes



질병 위험도 예측



Rev Esp Cardiol. 2020;73:608-10

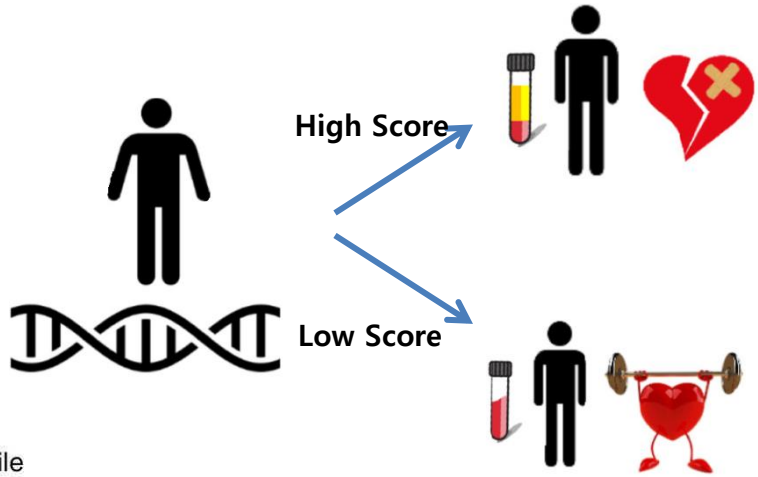
위험도 예측모델
Polygenic risk model

$$\hat{\phi}_j = \sum_{i=1}^{M} \hat{\beta}_i x_{ij}$$

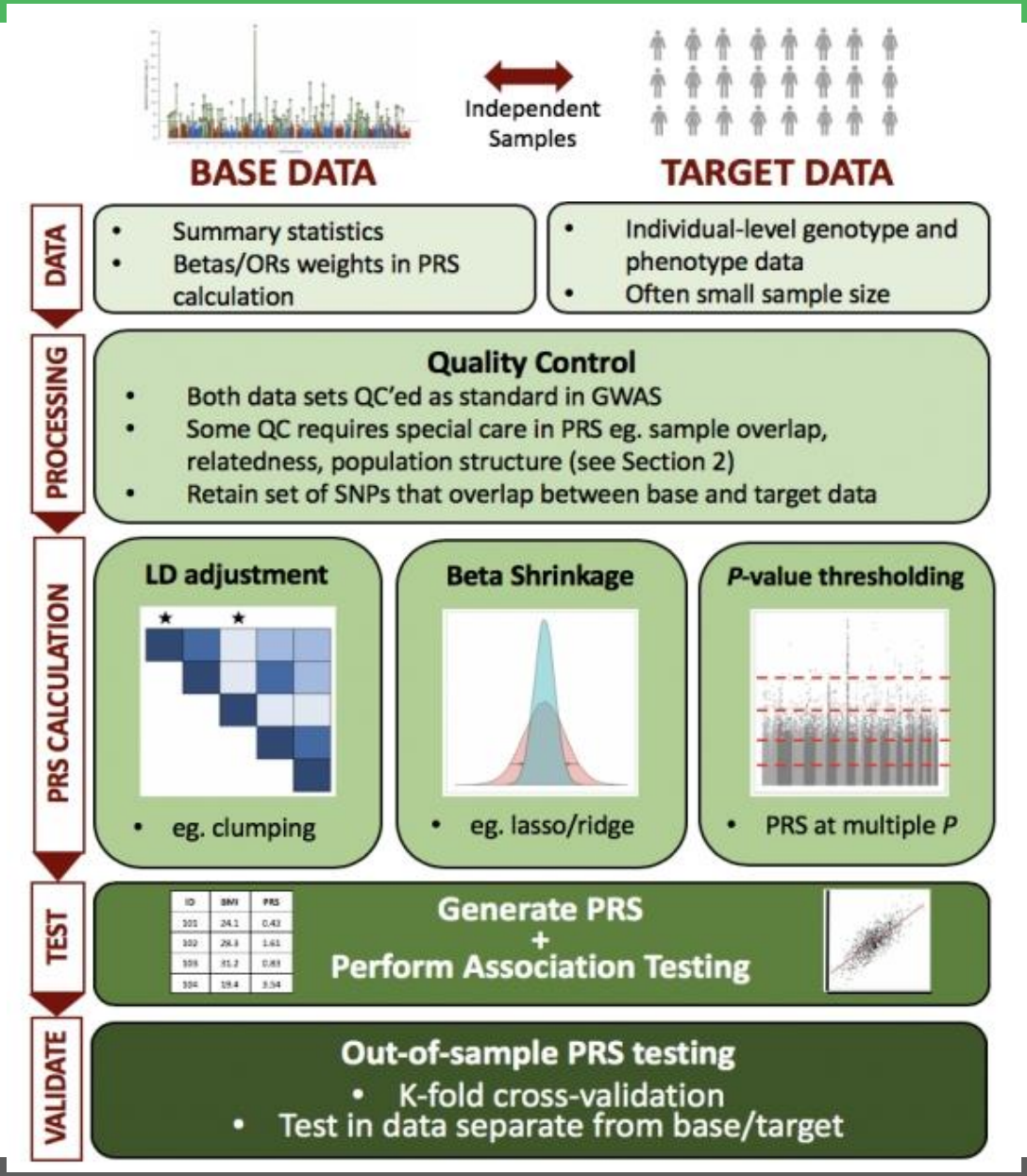
PRS

Estimated per-allele log odds

Genotype (# of variant alleles)

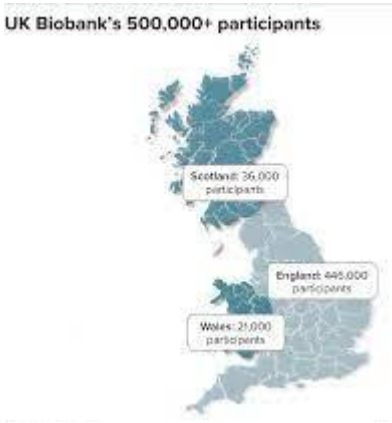


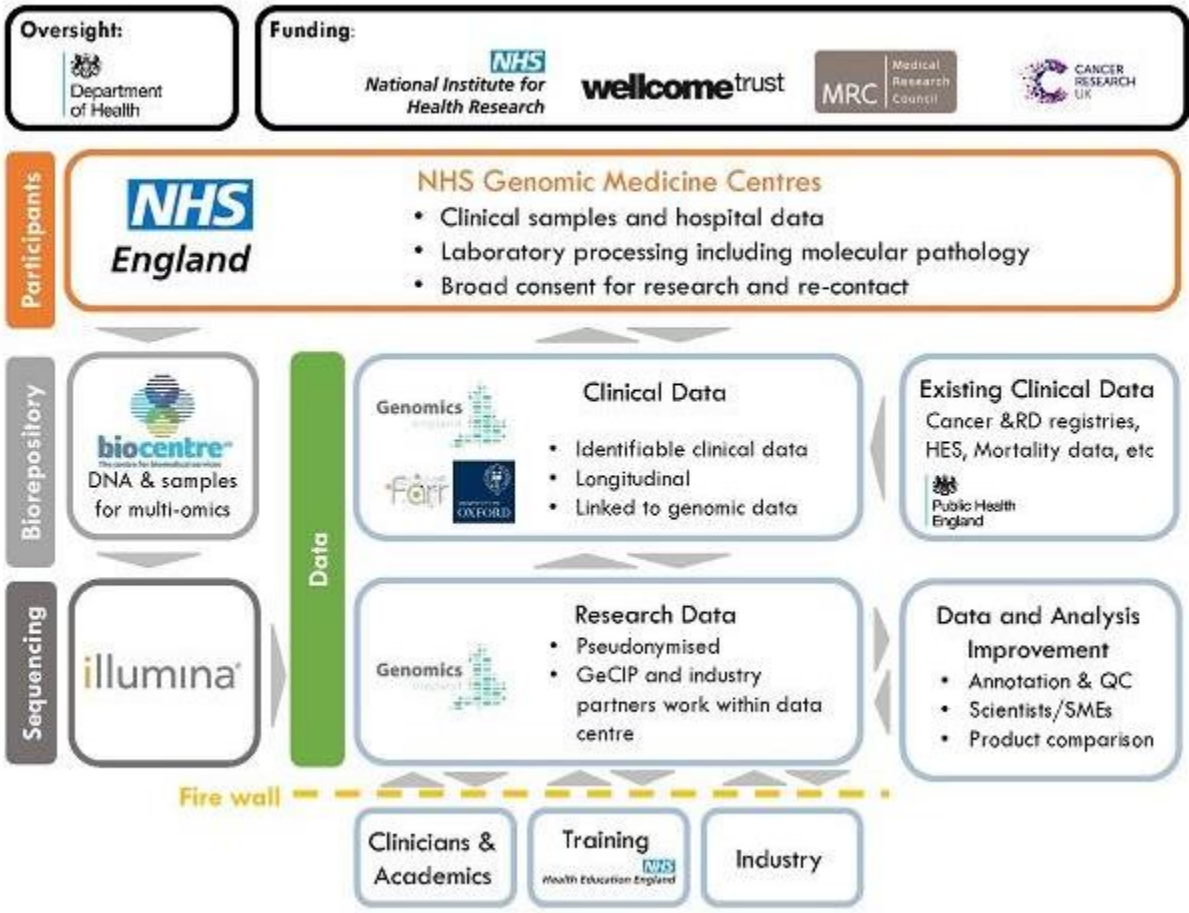
질병 위험도 예측


Independent Samples
BASE DATA
TARGET DATA
DATA
Summary statistics
Betas/ORs weights in PRS calculation
Individual-level genotype and phenotype data
Often small sample size
PROCESSING
Quality Control
Both data sets QC'ed as standard in GWAS
Some QC requires special care in PRS eg. sample overlap, relatedness, population structure (see Section 2)
Retain set of SNPs that overlap between base and target data
PRS CALCULATION
LD adjustment
eg. clumping
Beta Shrinkage
eg. lasso/ridge
P-value thresholding
PRS at multiple P
TEST
Generate PRS
+
Perform Association Testing
VALIDATE
Out-of-sample PRS testing
K-fold cross-validation
Test in data separate from base/target

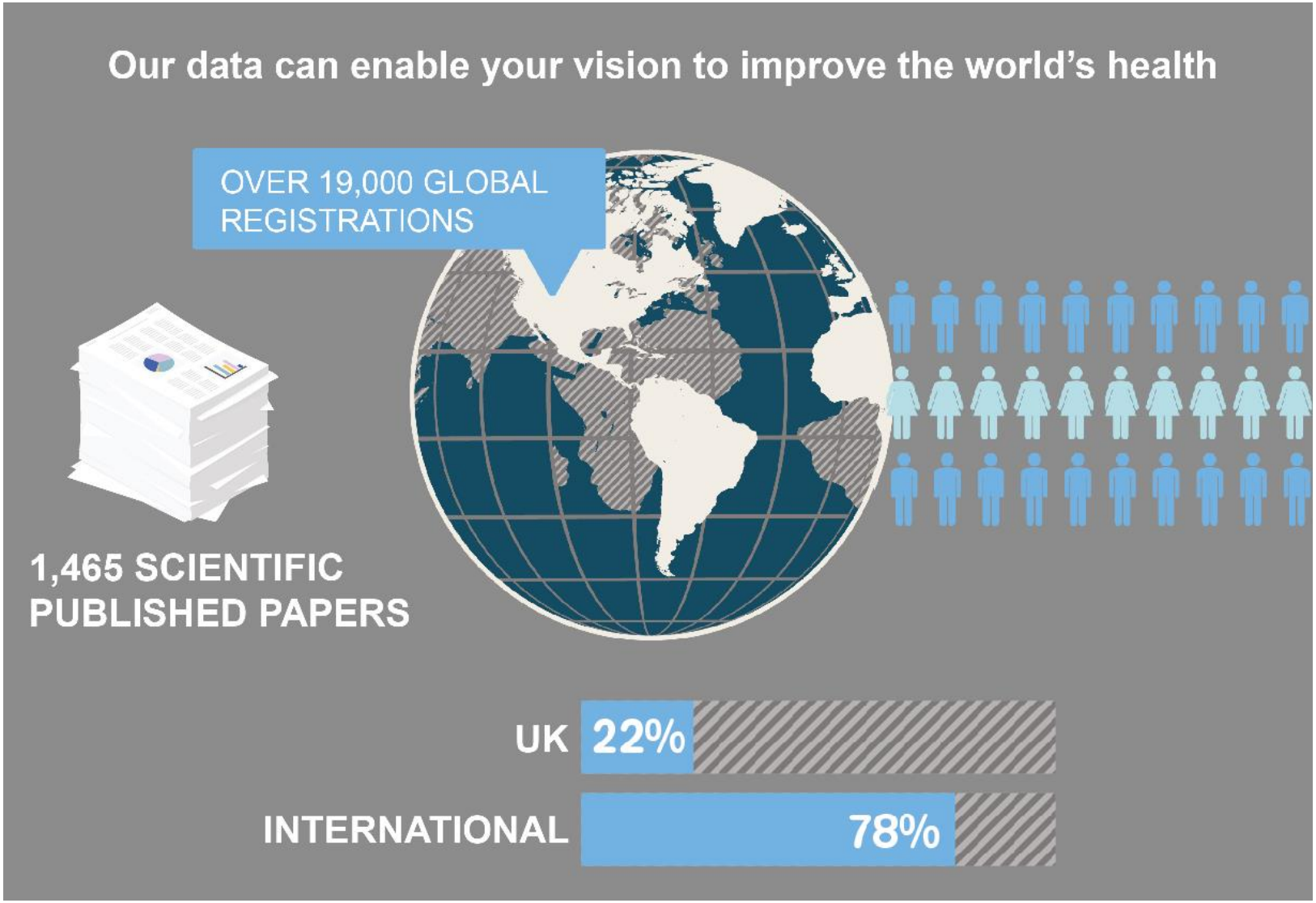
UK Biobank Project

500,000 whole human genomes



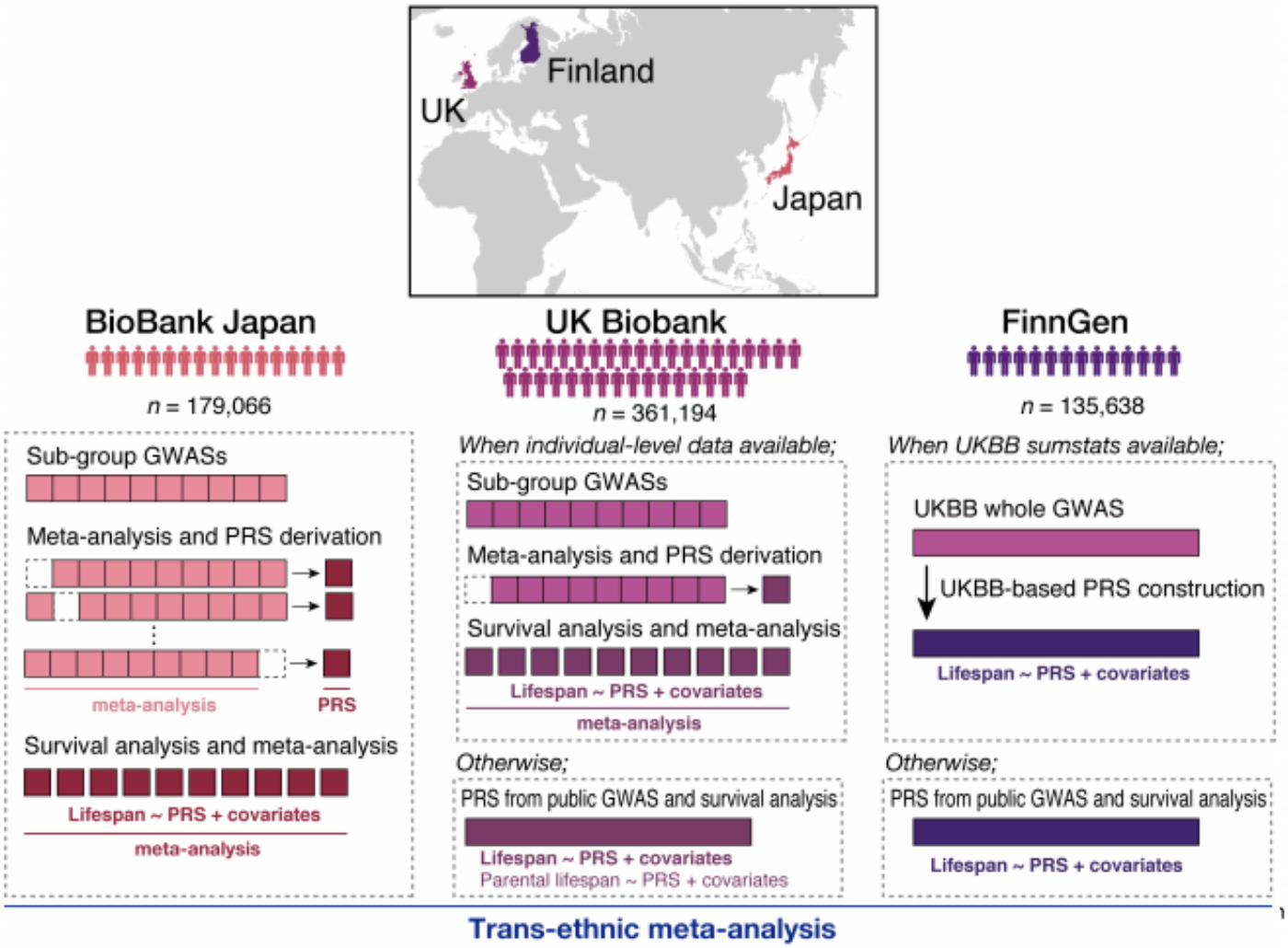


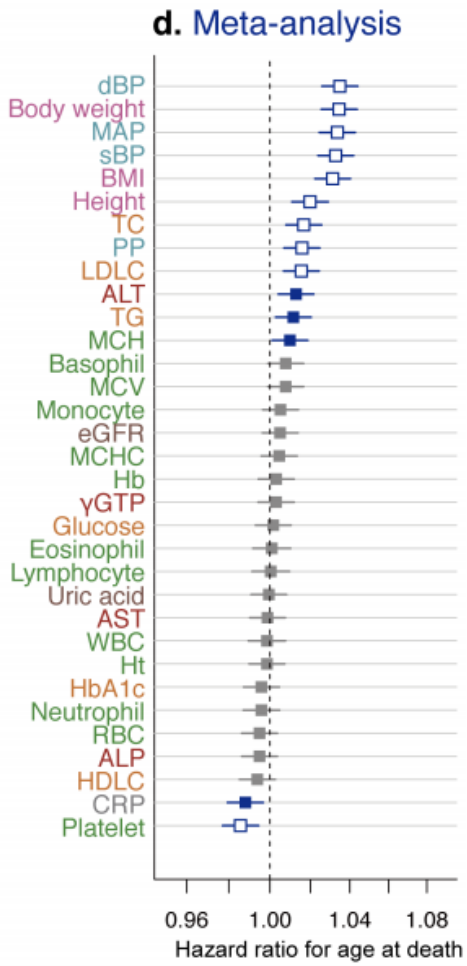
UK Biobank Project



Trans-biobank analysis with 676,000 individuals

The association of polygenic risk scores of complex traits with human lifespan

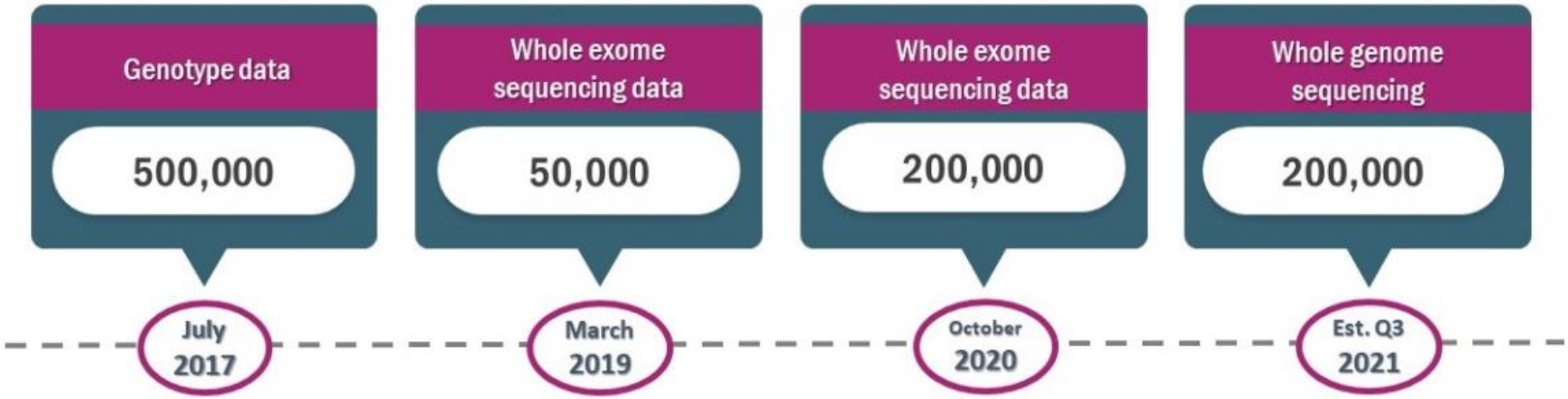




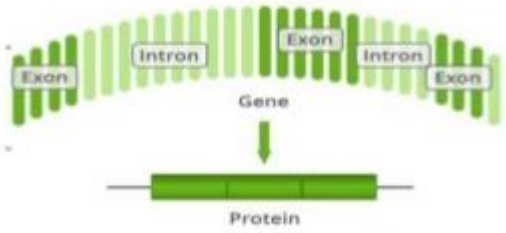
(Nature Medicine, 2020)

UK Biobank Project

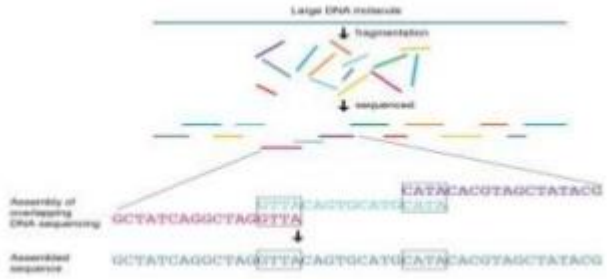
Genetic data release timeline



850,000 positions

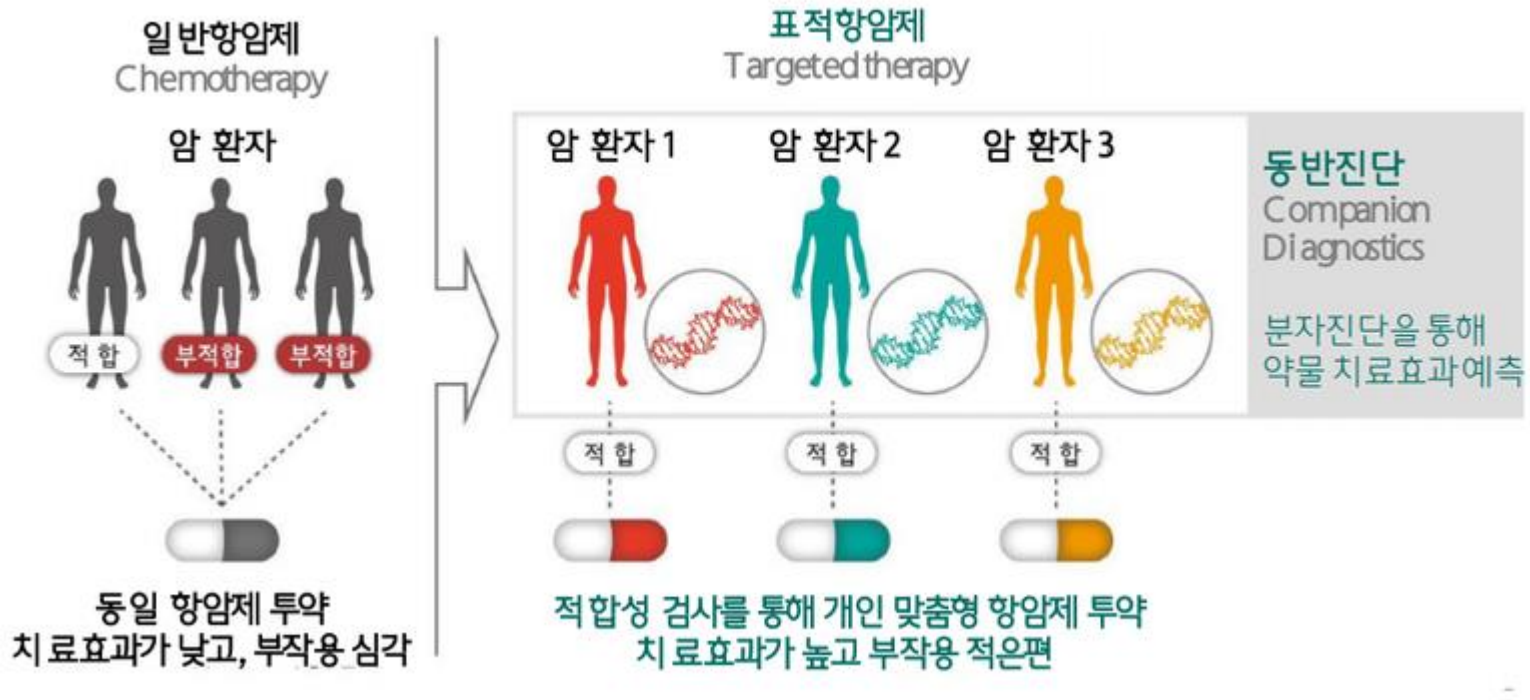


~2% of genome



3 billion in genome

정밀의학

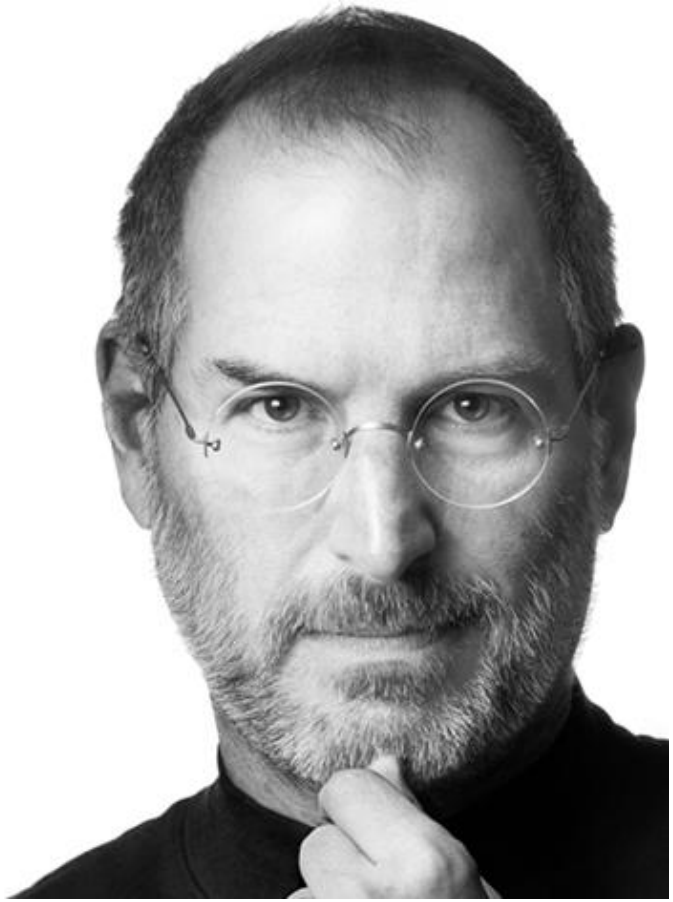


출처: 메디파나뉴스

- 유전정보, 임상정보, 환경정보
- 효과적인 치료방법(표적항암제 등)을 선택
- 각 개인에 최적화된 진단 및 치료를 적용

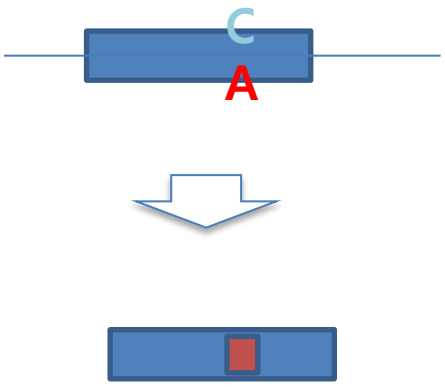
Steve Jobs

1955-2011



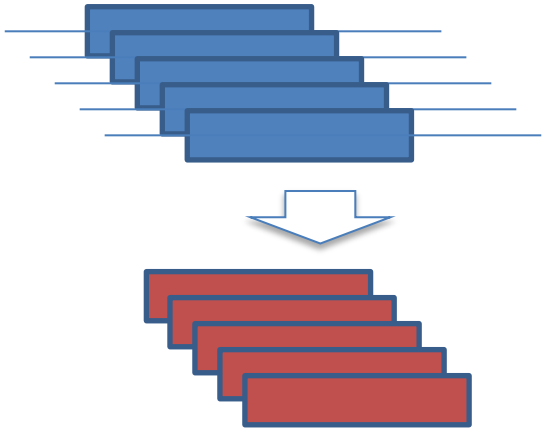
종양 세포들에서 유전자 변이들

1 포인트 변이



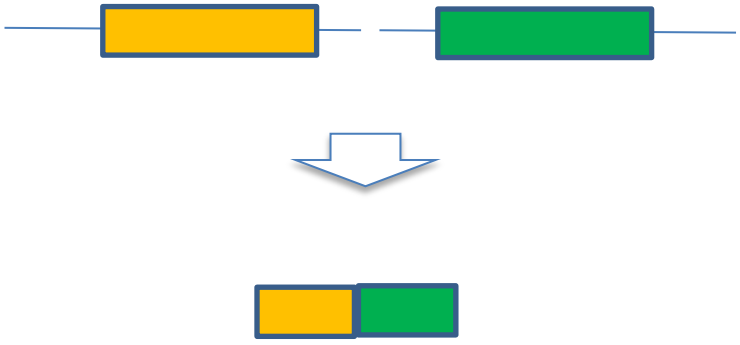
비정상 시그널

2 복제수 변이



증가/삭제 시그널

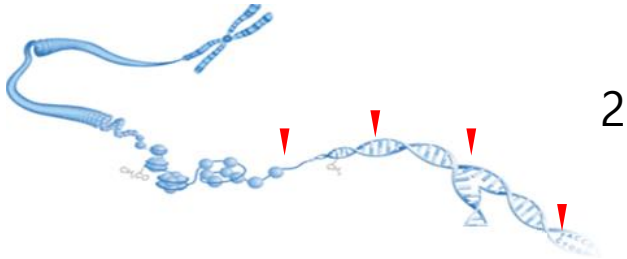
3 유전자 융합



비정상 시그널

유전자 패널 검사

요양기관명	유전성	고형암	혈액암
서울대병원	o	o	o
인천성모병원	o	o	o
부산대병원	o	o	o
분당서울대병원	o	o	o
아주대병원	o	o	o
서울의과학연구소 용인의원	o	o	o
서울아산병원	o	o	o
이원의료재단 이원의원	o	o	o
서울성모병원	o	o	o
연세대 세브란스병원	o	o	o
랩지노믹스진단검사의학과의원	o	o	o
고려대 안암병원	o	o	
고려대 구로병원	o	o	
삼성서울병원	o	o	
연세대 강남세브란스병원	o	o	
인하대병원			
길병원			
순천향대 서울병원			
녹십자의료재단 녹			
국립암센터			
삼광의료재단삼광의			
분당차병원			



2017년 3월 부터

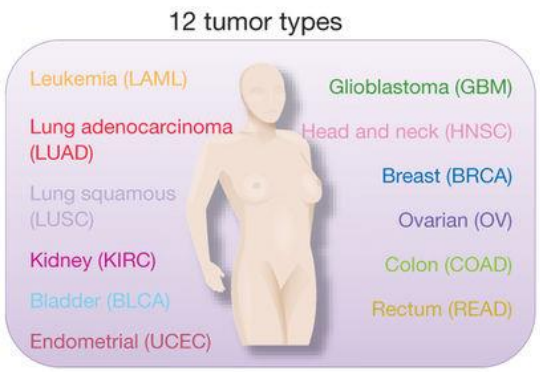
NGS 검사

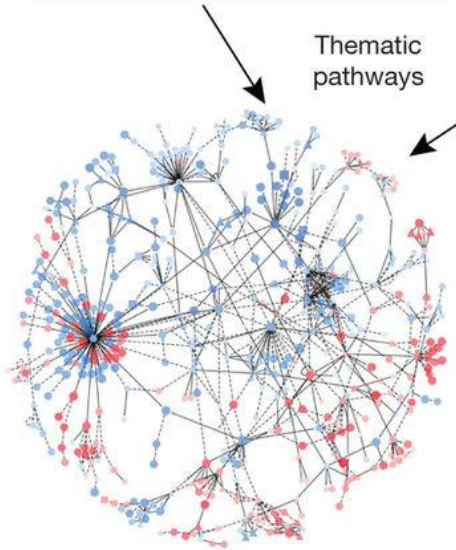
건강보험 적용 NGS 검사 대상 (단위:개)

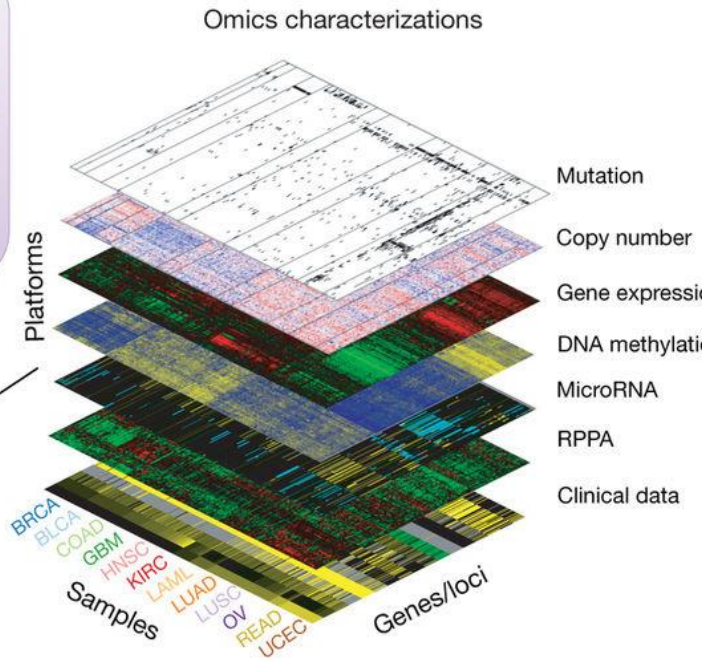
고형암(10종)	위암, 폐암, 대장암, 유방암, 난소암, 흑색종, 악성뇌종양, 위장관 기저종양, 소아신경모세포종, 원발성 불명암(총 14)
혈액암(5개군)	급성 골수성 백혈병(9), 급성 림프구성 백혈병(5), 악성 림프종(3), 형질세포종(3), 골수형성이상골수증식종양(11)
유전질환(4개군)	유전성 망막색소변성(7), 유전성 난청(4), 샤르코마리투스병(4), 기타(0)

기존 유전자 검사

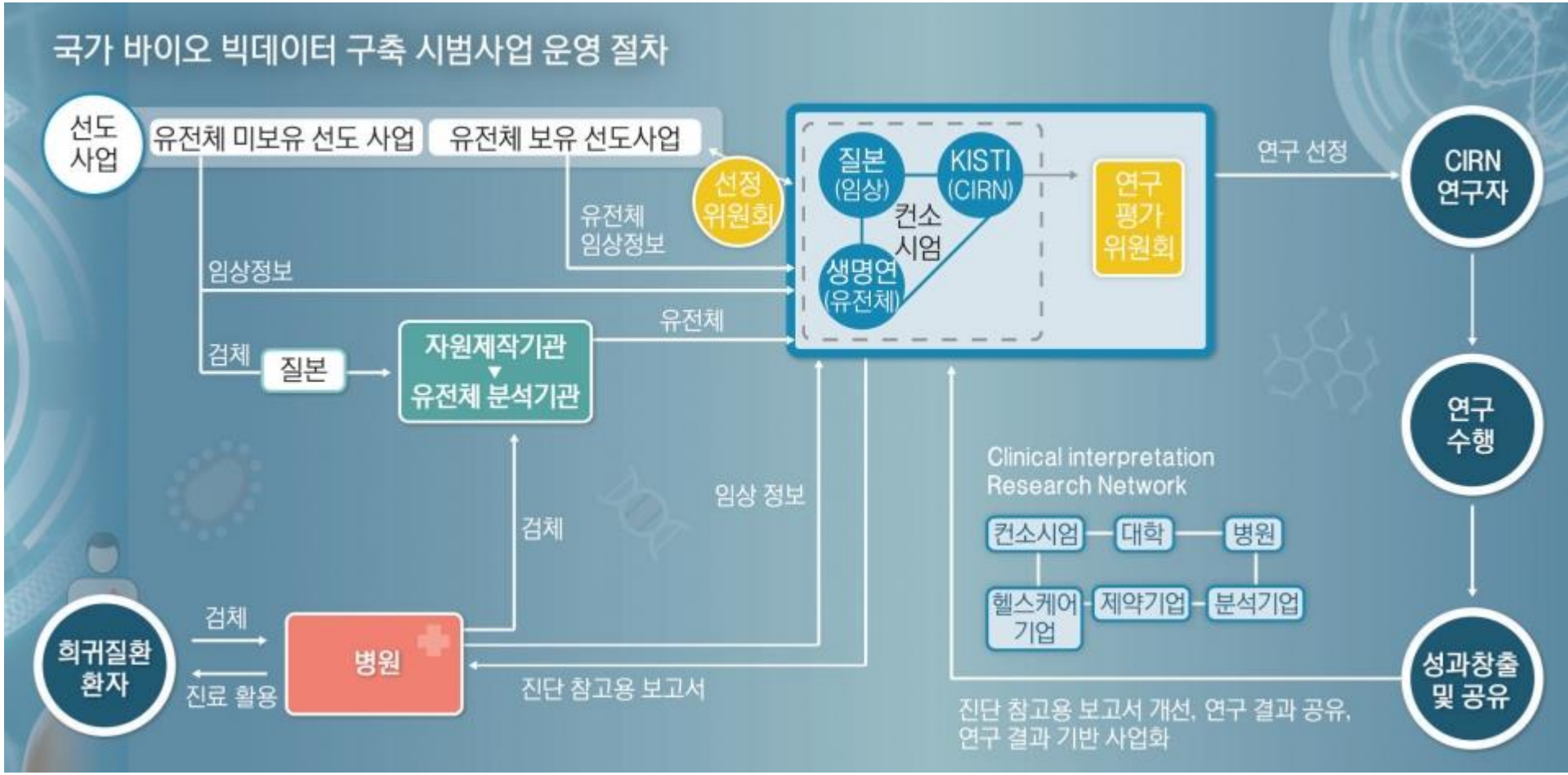
The Cancer Genome Atlas







국가 바이오 빅데이터 구축 시범사업



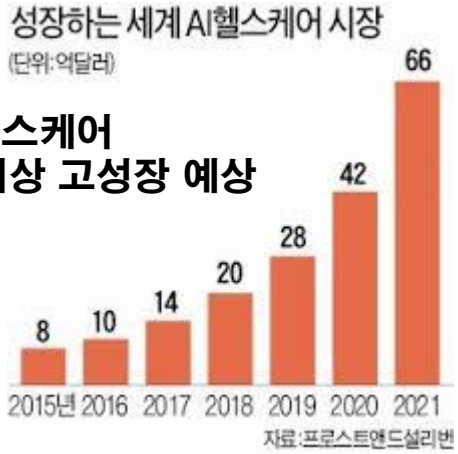
국가 바이오 빅데이터 구축 사업

시범사업 기간 동안 총 25,000명 ('20~'21)

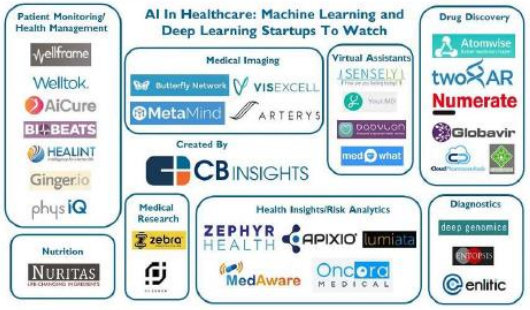
구분	주요 항목	규모
희귀질환	환자 및 부모의 임상 및 유전체데이터 희귀질환 분류, (추정)진단명, HPO 등	15,000명
대장암	환자의 임상 및 유전체 데이터 대장암 진단, 생존 및 치료 정보	400명
자폐증	환자 및 부모의 임상 및 유전체 데이터 행동, 지능지수, 적응능력, 주의력, 인지전환능력, 이상행동, 강박증상, 질병력, 약물투여력 등	500명
일반인	역학 및 유전체 데이터 질병력, 가족력, 생활습관, 신체 및 혈액 검사 및 소변검사 등 국가 일반검진 정보	울산1만명 프로젝트 1,600명 KoGES 5,000명

인공지능헬스케어의 급성장

The next major advancein medicine will be the use of AI



AI 헬스케어
연평균 40%이상 고성장 예상



자료: CB insight

900이상의 헬스케어+AI 스타트업에
$15억(1조6천억투자)

의료 빅데이터

기존 데이터	신규형태 데이터
병원 진료 기록	웨어러블 기기 생체 데이터
보험 청구 정보	유전체 데이터
논문/서적 정보	소셜 데이터

→ **AI 기반 통합 정보 분석** →

- ✓ 의사결정에 도움을 줌
- ✓ 의료 업무 프로세스 개선
- ✓ 새로운 제품/서비스

유전체 분야 AI 기술 적용 패러다임

Deep learning
Cognitive computing
…

classifying
prioritizing
interpreting
linking

AI 알고리즘 적용

IBM 왓슨
cleveland clinic
resistant to treatments

Geisinger
MyCode clinical genomics

Pathway Genomics
smartphon app of personal genetic info

Deep Genomics
genetic variant analysis

Mendel.ai
clinical trial matching platform

Sophia Genomics
genome interpretation

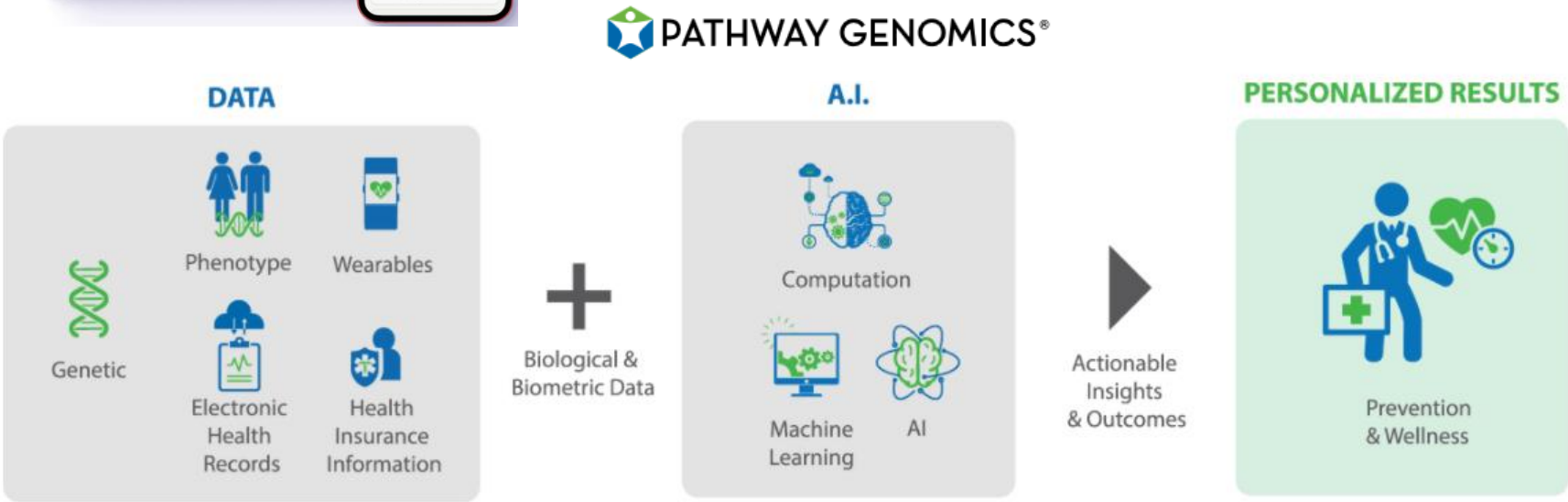
iCarbonX
genomics & health factors

글로벌 기관 /기업

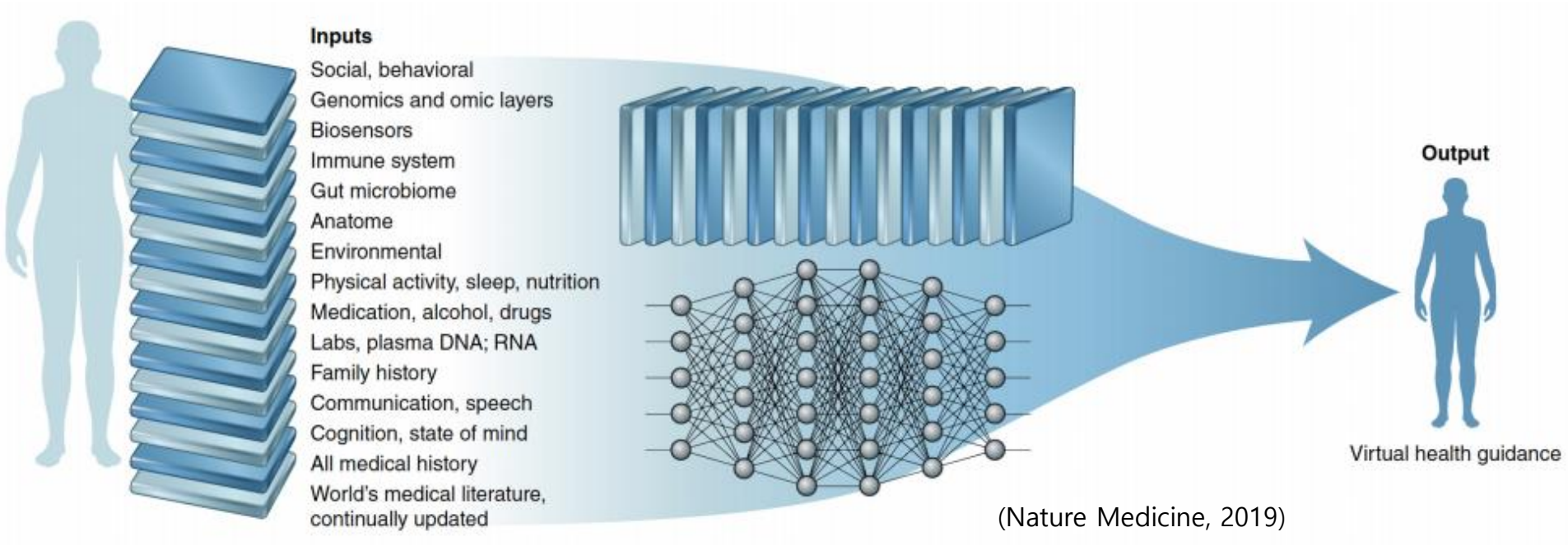
- ✓ **Pathogenic variants**
- ✓ **Drug development**
- ✓ **Drug/treatment response**
- ✓ **Cancer risk**
- ✓ **Clinical trials**

의료 분야 적용

- 개인유전체 정보를 이용한 AI 및 딥러닝
- 암, 유전질환, 심혈관질환, 약물반응성 등 예측
- 건강 히스토리를 근거로 개인 건강 지식 제공



The Virtual Medical Assistant



(Nature Medicine, 2019)