Introduction to the Korea National Genome Project

Ji-Hwan Park

Korea Bioinformation Center (KOBIC)

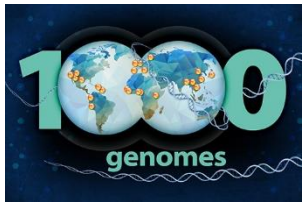
Korea Research Institute of Bioscience & Biotechnology (KRIBB)

KRIBB
Korea Research Institute of Bioscience & Biotechnology

kobic

Why do we need the National Genome Project?

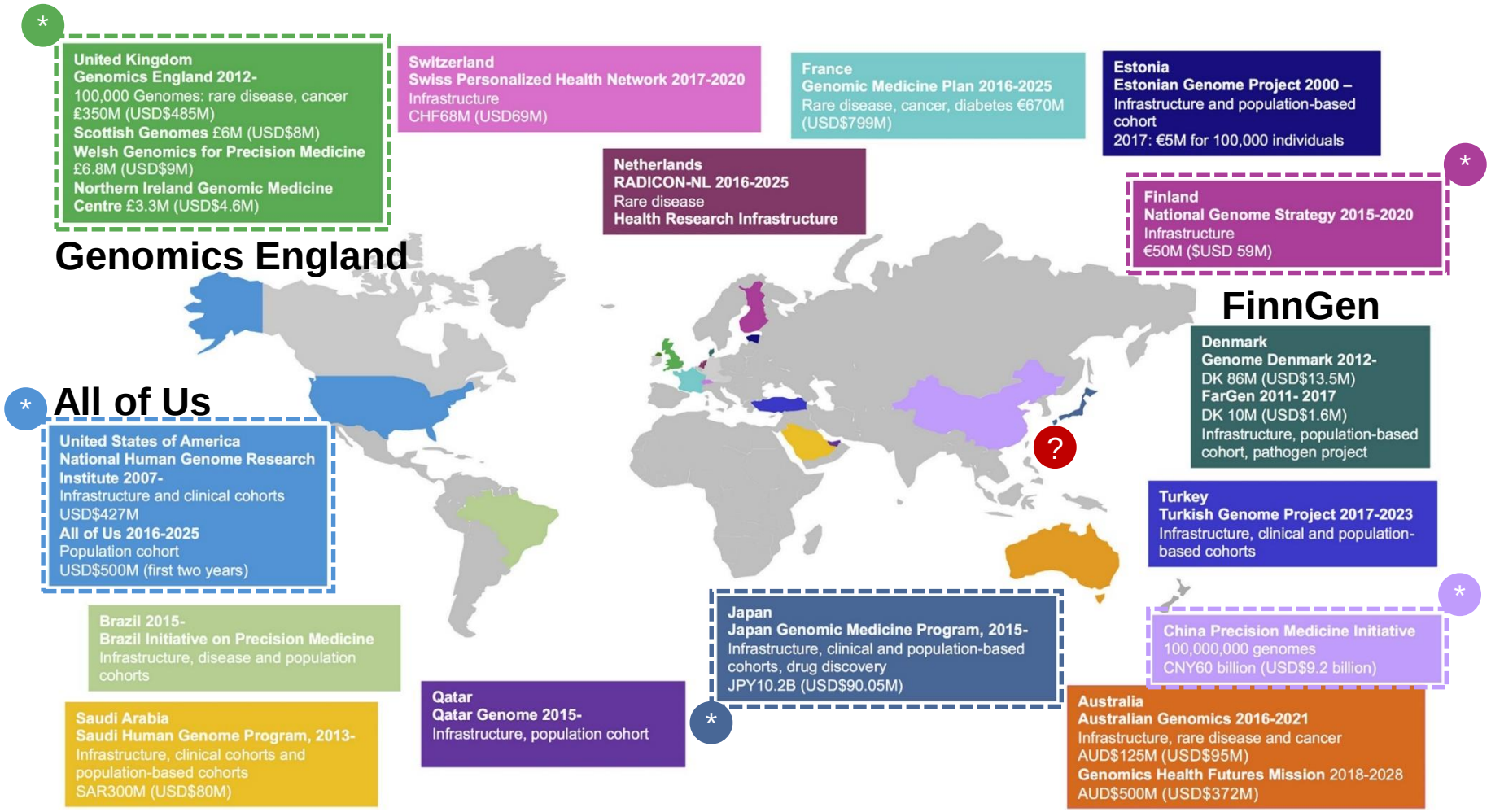
Drastic advances in high-throughput sequencing and
data analysis and sharing technologies



A new wave of population genomics & precision medicine

Why do we need the National Genome Project?

National genomic projects have been launched in > 41 countries



Stark *et al*., *AJHG* (2019)
Kovanda et al., *BMC Human Genomics* (2021)

The National Genome Project

Providing large-scale genomic & clinical data
to scientific & industrial community
in the purpose of studying precision medicine,

while protecting the participants' privacy

Workflow of the NGP

Collection of Samples & Data → **Data Processing & Quality Control** → **Data Analysis & Sharing**

Clinical report & assessment

- Rare Disease (RD)
- Normal Participants (KoGES)
- Normal Participants (KGP)
- Alzheimer's Disease (AD)
- Autism Spectrum Disorder (ASD)
- Colorectal Cancer (CRC)
- Non-Smoking Lung Adenocarcinoma (NSLA)

Whole-genome sequencing (WGS)

Clinical data

WGS & clinical data collected from previous projects

NGP consortium

kobic

KiSTi www.kisti.re.kr

질병관리청 KDCA

Research Environment Platform

- Computing resources
- Cloud services
- Data sharing

Integrative data analysis in the secure platform

→ Workflow

Cohorts in the NGP, pilot phase (2020 – 2022)

The pilot phase of the NGP provides genome and clinical information from up to **25,000 samples**:

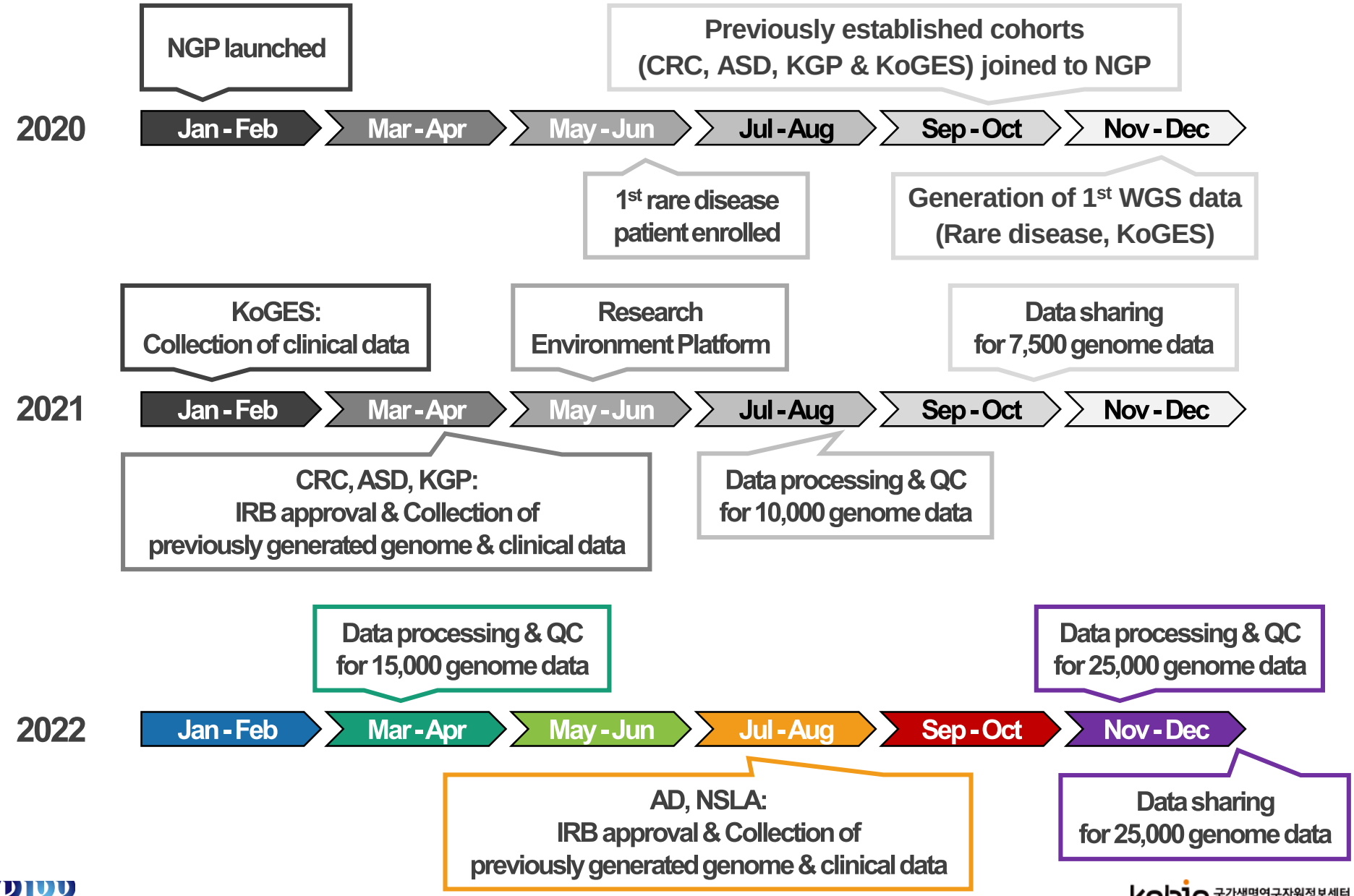
Rare	**Rare Disease (RD)**	Total 15,000 participants (predominantly trio)
Normal /Healthy	**Normal Participants (KoGES)**	Total 5,000 participants
	Normal Participants (KGP)	Total 2,383 participants
Currently incurable	**Alzheimer's Disease (AD)**	Total 500 Alzheimer's disease patients and 500 normal participants
	Autism Spectrum Disorder (ASD)	Total 849 participants (>200 patients & their parents; >149 families)
Cancer	**Colorectal Cancer (CRC)**	Total 300 patients (600 tumor and paired normal adjacent tissues)
	Non-Smoking Lung Adenocarcinoma (NSLA)	Total 84 never-smoker lung adenocarcinoma patients (168 tumor and paired normal adjacent tissues)

Progress of the NGP pilot phase, as of 1st Nov. 2022

- Clinical data from **23,833 participants** (= 25,000 samples) have been collected (from 15,000 RD, 300 CRC, 84 NSLA, 849 ASD, 1,000 AD, 1,600 KGP, & 5,000 KoGES cohorts)

- **20,000 whole-genome sequencing (WGS) data** have been **newly generated** and **5,000 WGS data** were **collected from NGP cohort partnership**

Progress of the NGP pilot phase, as of 1st Nov. 2022




Collection of
Samples & Data

Collection of samples & clinical data based on patient consent

***Bioethics and BioSafety Act*, Republic of Korea**

Participants' consent

■ 생명윤리 및 안전에 관한 법률 시행규칙 [별지 제34호서식]

인체유래물 연구 동의서

동의서 관리번호		(앞쪽)
인체유래물 기증자	성 명	생년월일
	주 소	
	전화번호	성별
법정대리인	성 명	관계
	전화번호	
연구책임자	성 명	
	전화번호	

이 동의서는 귀하로부터 수집된 인체유래물등(인체유래물과 그로부터 얻은 유전정보를 말합니다)을 질병의 진단 및 치료법 개발 등의 연구에 활용하기 위한 것입니다. 동의는 자발적으로 이루어지므로 아래의 내용을 읽고 궁금한 사항은 상담자에게 묻고 질문할 기회를 가지고 충분히 생각한 후 결정하시기 바라며, 이 동의서에 대한 동의 여부는 귀하의 향후 검사 및 치료 등에 어떤 영향도 미치지 않습니다.

Approval of research plans

+ IRB

Institutional Review Board

- The research participation agreements and plans are approved by multiple Institutional Review Boards (IRBs).
- Blood or saliva samples and clinical information are collected, according to the strictly managed quality control (QC) processes.

Rare disease recruitment

The rare disease patients take part in the NGP via the following hospitals:

- Seoul National University Hospital (SNUH)
- Samsung Medical Center (SMC)
- Asan Medical Center (AMC)
- Pusan National University Yangsan Hospital (PNUYH)
- The Catholic Univ. of Korea Seoul St.Mary's Hospital (SSMH)
- Kyungpook National University Chilgok Hospital (KNUH)
- Seoul National University Bundang Hospital (SNUBH)
- Severance Hospital (YUHS)
- Ajou University Hospital (AJOUMC)
- Inha University Hospital (INHAUH)
- Cheju Halla General Hospital (CHH)
- Inje University Busan Paik Hospital (INJE)
- Jeonbuk National University Hospital (JBUH)
- Chonnam National University Hwasun Hospital (CNUHH)
- Chungnam University Hospital (CNUH)
- Chungbuk National University Hospital (CBNUH)
- Wonju Severance Christian Hospital (WSCH)

NGP cohort partnership

In cooperation with previous projects (NGP cohort partnership),
NGP collected diverse cohort datasets,

Normal Participants (KoGES)

Normal Participants (KGP)

Autism Spectrum Disorder (ASD)

Colorectal Cancer (CRC)

Alzheimer's Disease (AD)

Non-Smoking Lung Adenocarcinoma (NSLA)

Current research & medical staffs in **NGP cohort partnership**
may participate in the **next phase of NGP**,
which aims to **collect the data from > 1M participants**

NGP cohort partnership

The NGP cohort partnership encompasses **diverse types of cohorts**
(**cancer** and **currently incurable disease** patients & **normal/healthy** participants)

Following implications from the cohort partnership:

Establishment of **the standard operating procedure (SOP)**
for diverse types of diseases and healthy participants

- **Ethical, Legal and Social Implications (ELSI)**
  (*e.g.*, Consent forms and research plans
  regarding the prospective / retrospective studies in the NGP)

- **Strategies for sample and data collection**
  (*e.g.*, sampling and banking solid tumors rather than blood)

- **Data standardization or strategies for update of clinical information**
  (*e.g.*, trajectory analysis of clinical information & outcomes)

Standardization & QC of clinical information

The Research Environment Platform provides **structured clinical data**:

	Rare Disease (RD)	Human phenotype ontology (HPO)
Normal/Healthy	**Normal Participants (KoGES)**	General health checkup including interview medical examination, radiography examination, blood test, and urinaylsis
	Normal Participants (KGP)	
	Alzheimer's Disease (AD)	Blood test, Urinalysis, Education, and Mini-Mental State Examination (MMSE)
	Autism Spectrum Disorder (ASD)	ASD-specific information, including Social communication questionnaire, Autism diagnostic observation schedule and interview
Cancer	**Colorectal Cancer (CRC)**	Pathological diagnosis, metastasis, survival, medical treatment, and radio- and chemotherapy
	Non-Smoking Lung Adenocarcinoma (NSLA)	


Genome sequencing

Whole-genome sequencing (WGS) process

Sample QC | **Library prep & QC** | **Sequencing**

Input Sample (gDNA) QC

- gDNA conc. & amount
- A260/A280 ratio
- A260/A230 ratio
- DNA Integratiy Number

Library QC

- Library conc. & amount
- Library size distribution

Sequencing

< **141,000** samples are able to be sequenced by using Illumina NovaSeq 6000

Runtime	Num. of Runs / Year	Maximum data size / Run	Maximum data size / Year	WGS samples / year
48 hr / Run	156 Runs / yr	6 Tbase / Run	12,168 Tbase / yr	110,000 samples / yr

Newly generated whole-genome sequencing (WGS)

As of 1st Nov. 2022, a total of **20,000 samples were sequenced**

Rare Disease (RD) 15,000 participants

Normal Participants (KoGES) 5,000 participants

For each sample, **138 GBase (>30X) on average** was generated

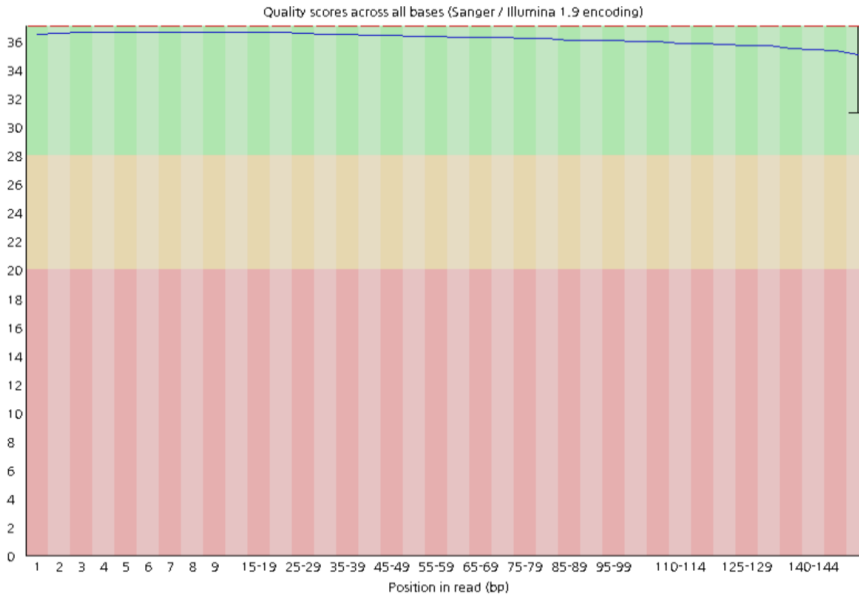
WGS data were generated by Macrogen, DNALink, TheragenBio, and LabGenomics

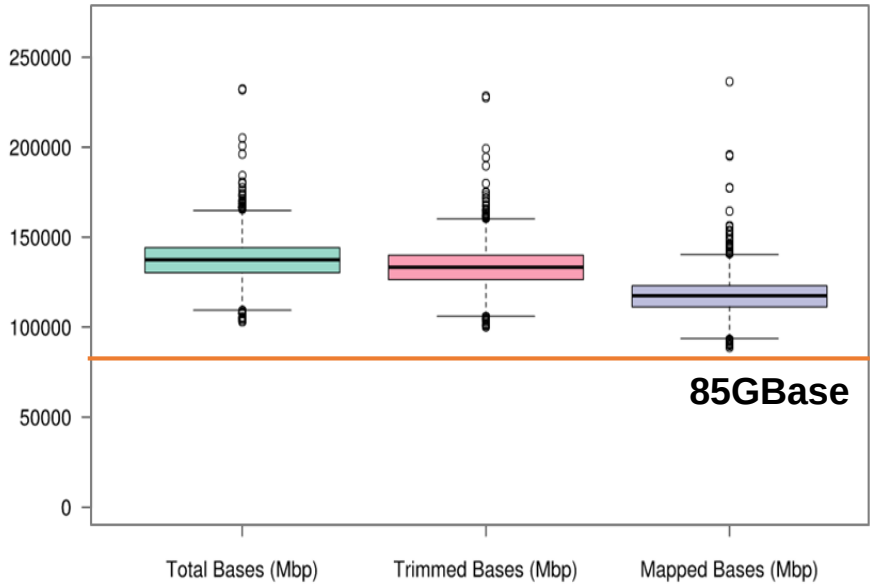
Sequence data QC

For each sample, **138 GBase (>30X) on average** was generated

Phred Q30	91.6%
De-duplicate read	87.8%
Mappable read	99.8%

Genome Coverage	≥ 1X	94.9%
Genome Coverage	≥ 10X	94.3%
Genome Coverage	≥ 30X	86.7%

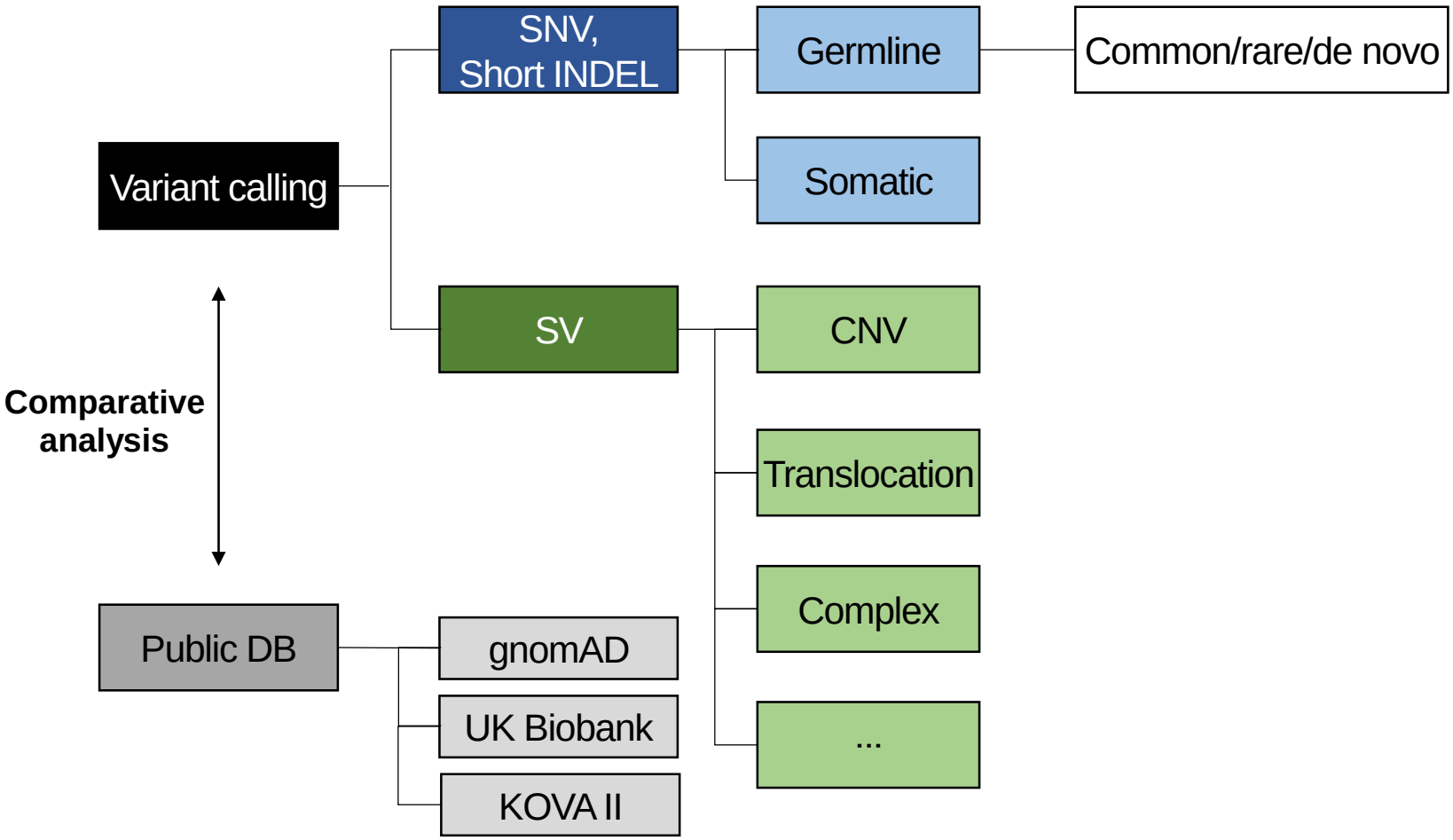





Data processing
& Quality control

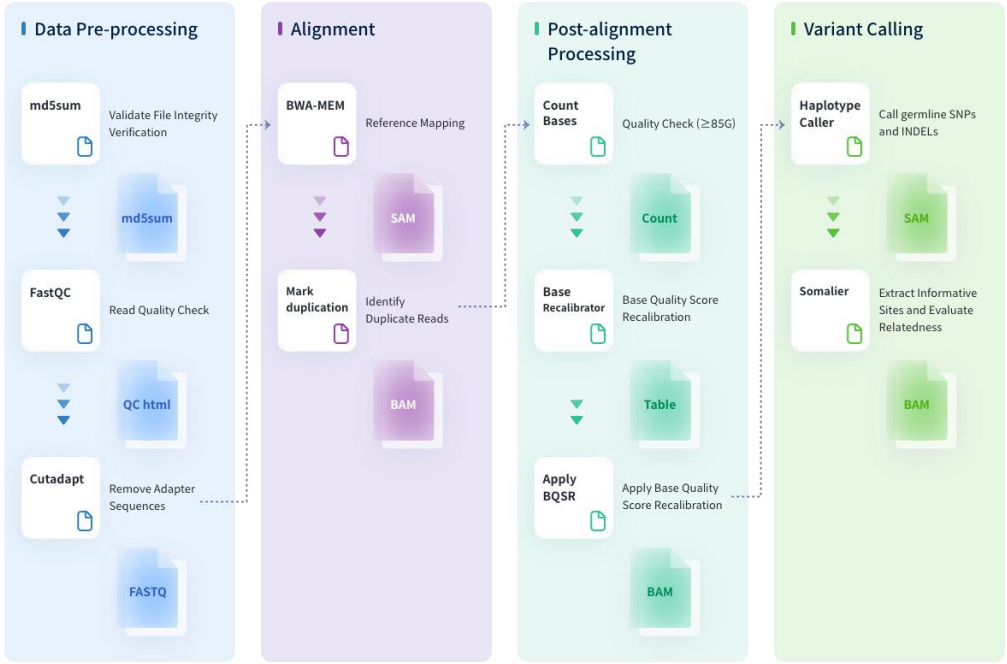
WGS data analysis & QC

Data processing and QC pipelines for the WGS data have been established to call germline or somatic variants with high consistency and confidence.



Data processing and QC pipelines for the WGS

Data processing and QC pipelines for the WGS data have been established to call germline or somatic variants with high consistency and confidence.



- **<u>Germline mutation</u>**

  GATK-Spark: MapReduce-based distributed parallel computing

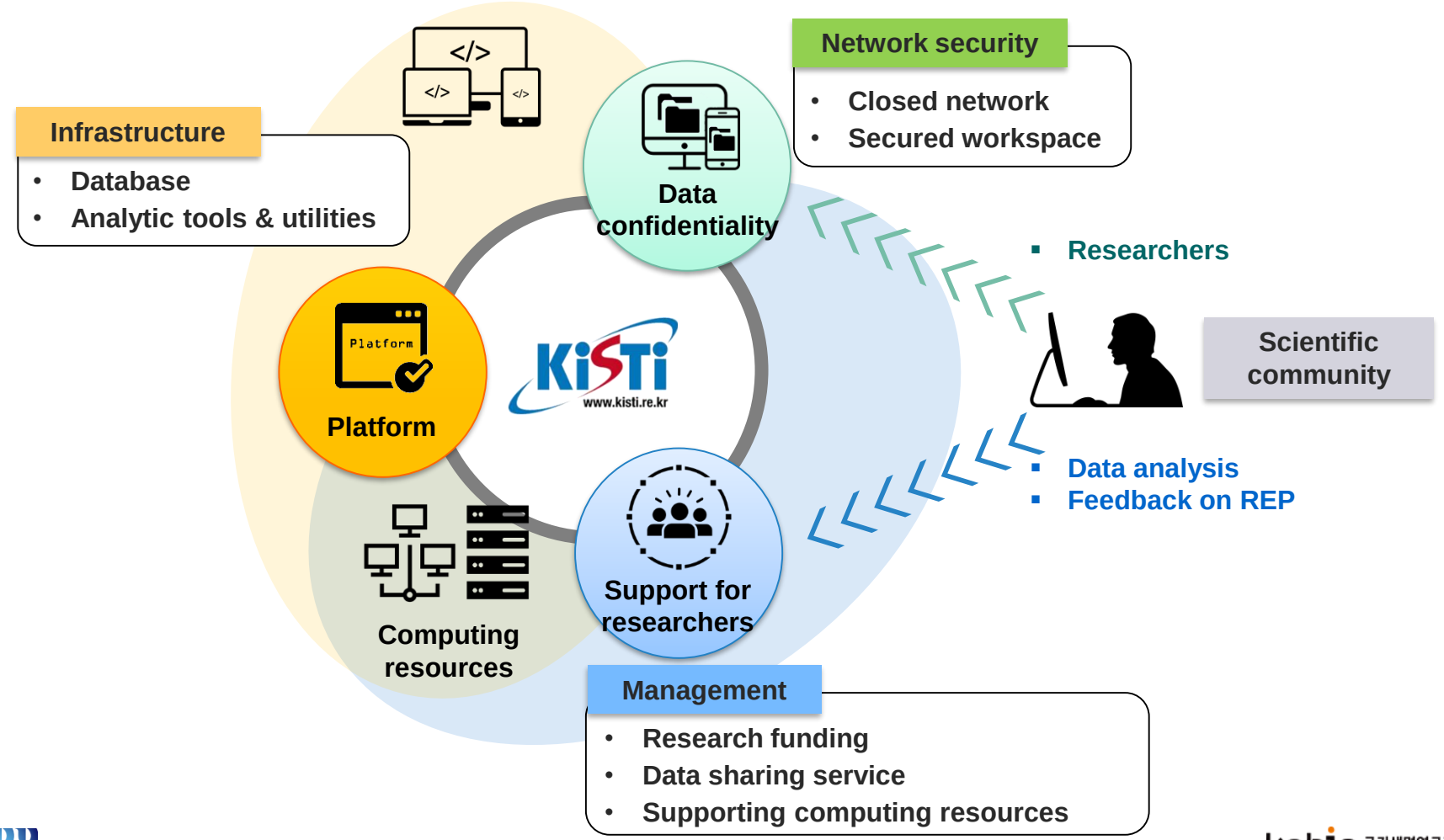
  GATK-Spark pipeline performs **<u>3~4 times faster</u>** than Java-based GATK

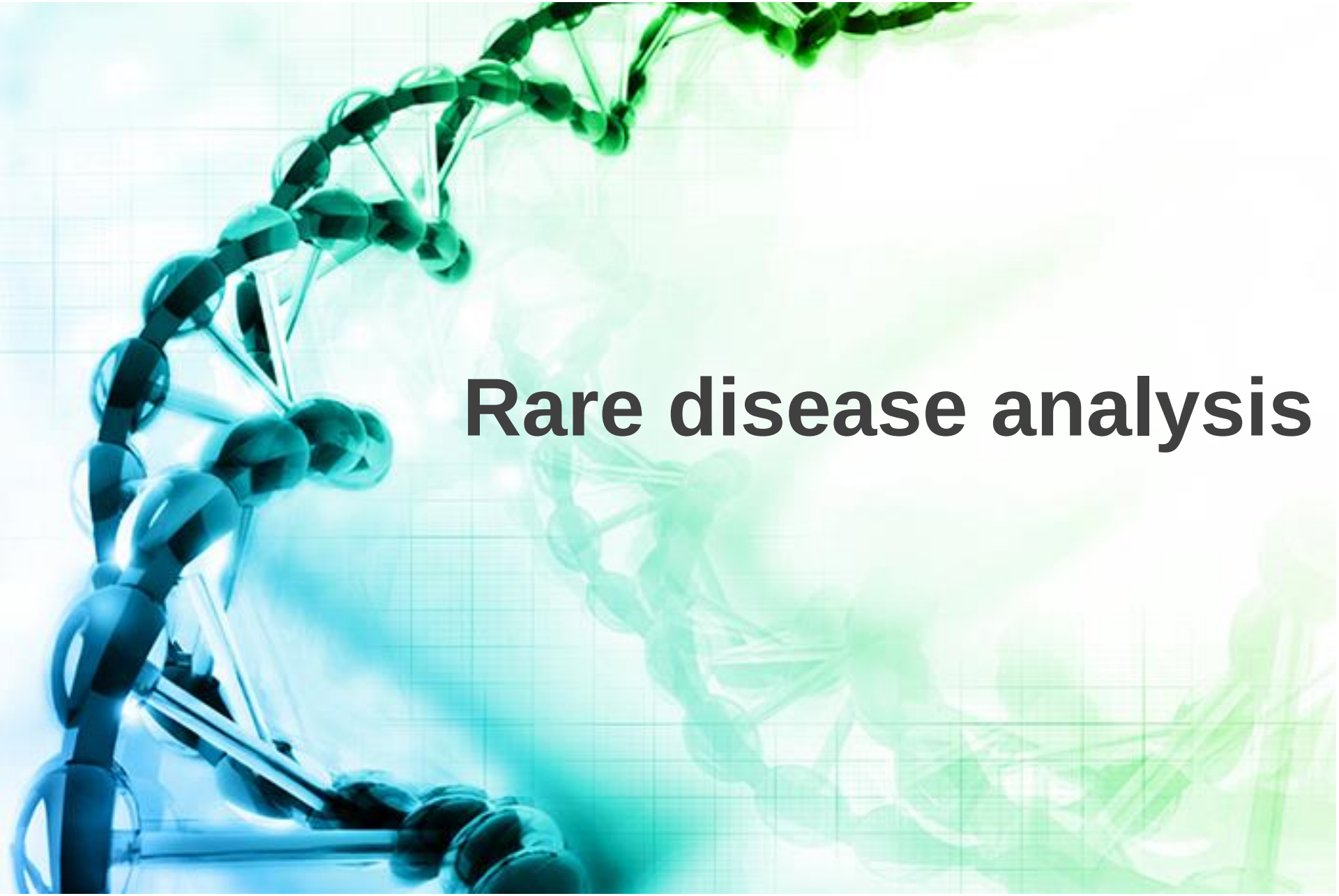
- **<u>Somatic mutation</u>**

  FPGA-based DRAGEN platform, the Hardware accelerated variant analysis, enables somatic variant analysis **<u>20 times faster</u>** than GATK-Spark (60 min per sample)

Research Environment Platform (REP)

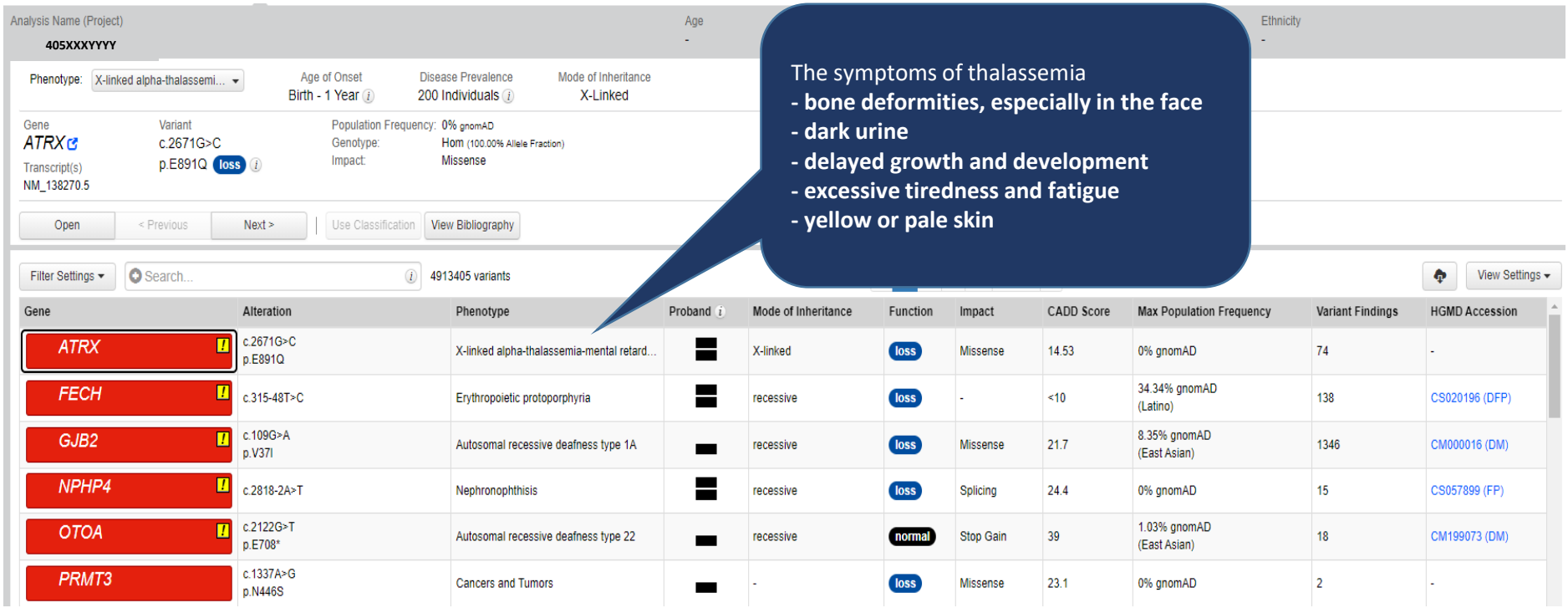
The platform provides processed WGS data and **de-identified clinical information**, enabling the researchers to conduct **a multitude of integrative data analyses** in the **secured workspace**.




Rare disease analysis

Rare disease data analysis (example)

Sample_ID: 405XXXYYYY
Sex: Male
Age: <7
Diagnosis : Neurodevelopmental disorders (Delayed speech and language development)



Gene	Alteration	Phenotype	Proband	Mode of Inheritance	Function	Impact	CADD Score	Max Population Frequency	Variant Findings	HGMD Accession
ATRX	c.2671G>C p.E891Q	X-linked alpha-thalassemia-mental retard...		X-linked	loss	Missense	14.53	0% gnomAD	74	-
FECH	c.315-48T>C	Erythropoietic protoporphyria		recessive	loss	-	<10	34.34% gnomAD (Latino)	138	CS020196 (DFP)
GJB2	c.109G>A p.V37I	Autosomal recessive deafness type 1A		recessive	loss	Missense	21.7	8.35% gnomAD (East Asian)	1346	CM000016 (DM)
NPHP4	c.2818-2A>T	Nephronophthisis		recessive	loss	Splicing	24.4	0% gnomAD	15	CS057899 (FP)
OTOA	c.2122G>T p.E708*	Autosomal recessive deafness type 22		recessive	normal	Stop Gain	39	1.03% gnomAD (East Asian)	18	CM199073 (DM)
PRMT3	c.1337A>G p.N446S	Cancers and Tumors		-	loss	Missense	23.1	0% gnomAD	2	-

Rare Disease (RD) 8,000 participants

Genome Sequencing for Undiagnosed Genetic Diseases

Annual Review of Medicine

What Has the Undiagnosed Diseases Network Taught Us About the Clinical Applications of Genomic Testing?

David R. Murdock,[1] Jill A. Rosenfeld,[1] and Brendan Lee[1,2]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA; email: blee@bcm.tmc.edu

[2]Texas Children's Hospital, Houston, Texas 77030, USA

Genetic testing has undergone a revolution in the last decade, particularly with the advent of next-generation sequencing and its associated reductions in costs and increases in efficiencies. The Undiagnosed Diseases Network (UDN) has been a leader in the application of such genomic testing for rare disease diagnosis. This review discusses the current state of genomic testing performed within the UDN, with a focus on the strengths and limitations of whole-exome and whole-genome sequencing in clinical diagnostics and the importance of ongoing data reanalysis. The role of emerging technologies such as RNA and long-read sequencing to further improve diagnostic rates in the UDN is also described. This review concludes with a discussion of the challenges faced in insurance coverage of comprehensive genomic testing as well as the opportunities for a larger role of testing in clinical medicine.

Table 1 Comparison of reportable variant types detected by different genetic testing assays

Genetic test	SNVs/ Indels	CNVs	Mosaic variants	Repeat expansions	Balanced and complex SVs	mtDNA	Genes tested	VUS potential[a]	Cost[a]
Sanger	Yes	No	Limited	No	No	If included	Single–few	+	$$
CMA	No	Yes	Yes (CNV only)	No	No	No	Up to ~20,000	++	$$
NGS panel	Yes	Yes	Yes	No	No	If included	Few–hundreds	++	$
WES	Yes	Limited	Limited (depends on coverage)	No	No	If included	~20,000 (coding regions only)	+++	$$$
WGS	Yes	Yes	Limited (depends on coverage)	Yes	Yes	Yes	~20,000 (coding and noncoding regions)	++++	$$$$

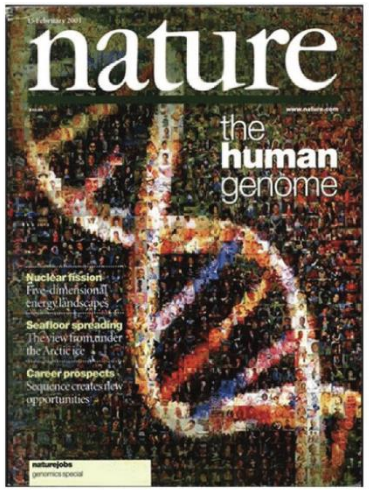
Table 2 Comparison of published RNA-seq studies and their contrasting approaches and results

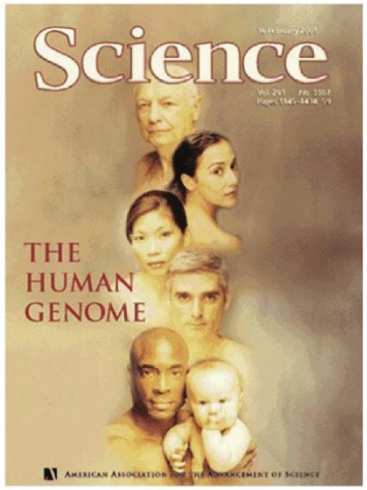
Reference	UDN study	Phenotypes	Tissue	Events detected	Analysis method[a]	RNA-seq diagnostic rate
35	No	Neuromuscular	Muscle	Splicing	Candidate + outlier	35%
34	No	Neuromuscular	Muscle, fibroblasts, fibroblast-derived myotubes	Expression, splicing, ASE	Outlier	36%
29	Yes	Multiple (neurologic most common)	Whole blood	Expression, splicing, ASE	Outlier	7.5%
30	Yes	Multiple (neurologic most common)	Whole blood, fibroblasts, muscle, bone marrow	Expression, splicing, ASE	Candidate	18%
23	Yes	Multiple (neurologic most common)	Whole blood, fibroblasts	Expression, splicing, ASE	Outlier	17%

Long-read sequencing is necessary for the undiagnosed rare diseases

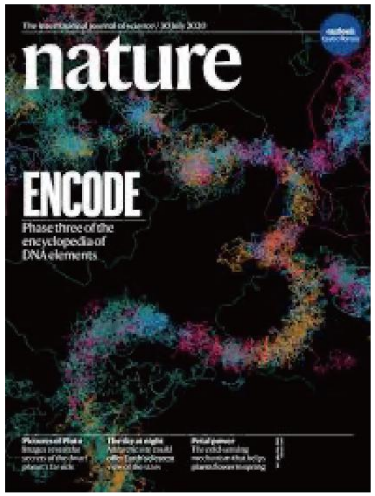

The Korean
Pan-genome Project

The complete sequence of a human genome



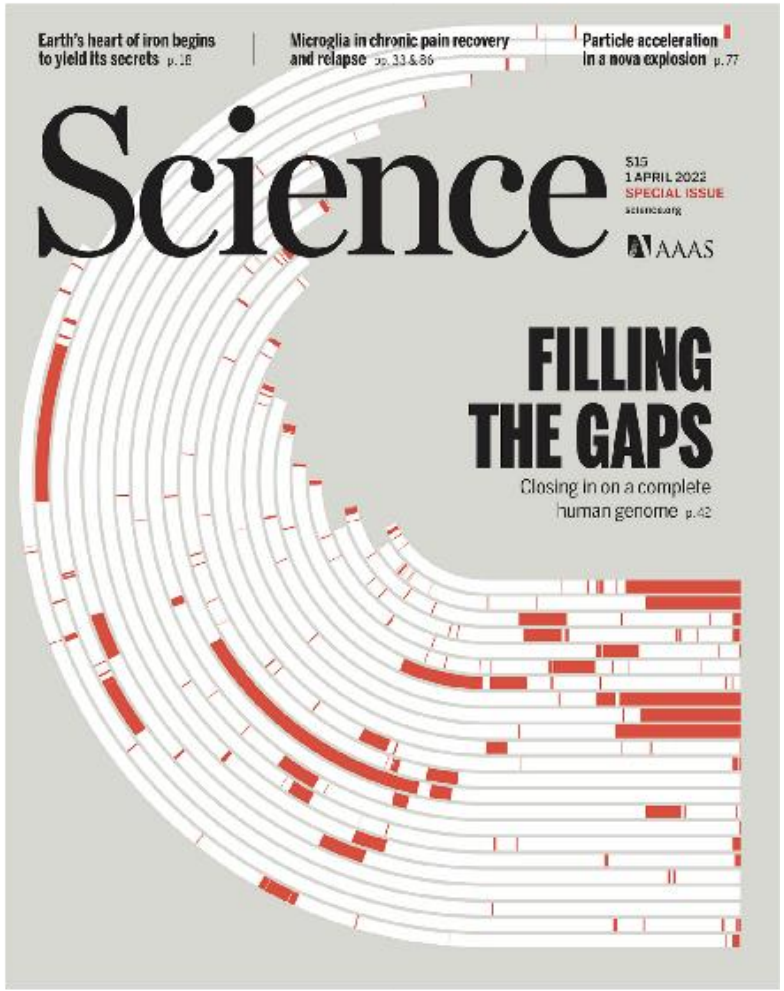


1990~2003



2003~2017(4th)

2008~2015



2014~2022

T2T consortium to the Human Pangenome Project

Science

Current Issue First release papers Archive About

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

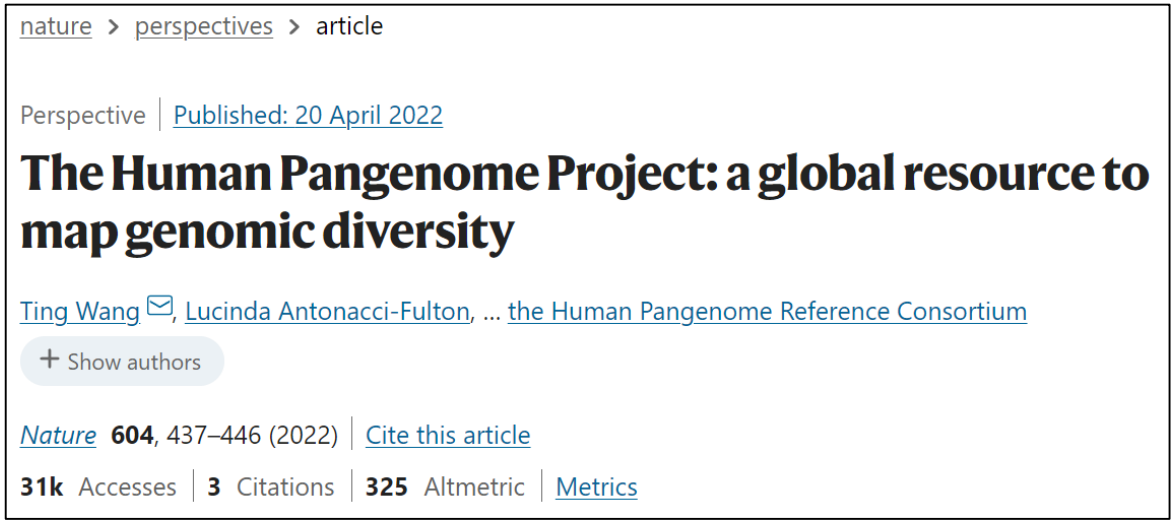
SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

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, [...] ADAM M. PHILLIPPY +91 authors Authors Info & Affiliations

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

CHM13 reference genome



nature > perspectives > article

Perspective | Published: 20 April 2022

The Human Pangenome Project: a global resource to map genomic diversity

Ting Wang, Lucinda Antonacci-Fulton, ... the Human Pangenome Reference Consortium

+ Show authors

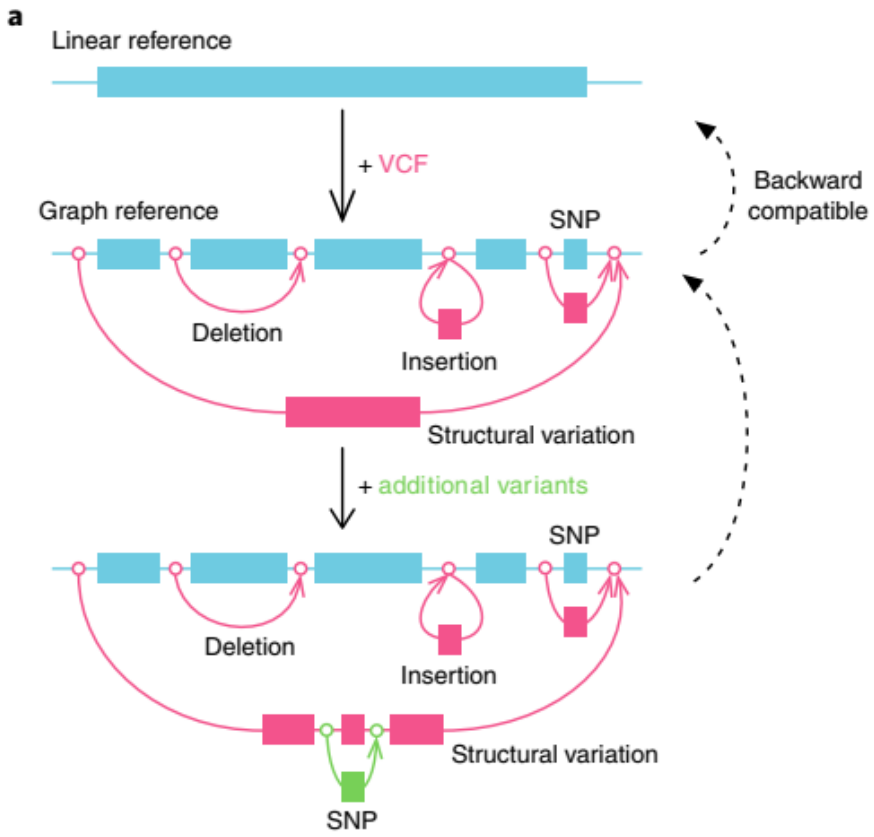
Nature **604**, 437–446 (2022) | Cite this article

31k Accesses | **3** Citations | **325** Altmetric | Metrics

Needs for a pan-genome (genome graph)

TECHNICAL REPORT
https://doi.org/10.1038/s41588-018-0316-4
nature genetics

Fast and accurate genomic analyses using genome graphs



Rakocevic (2020) Nat Genet

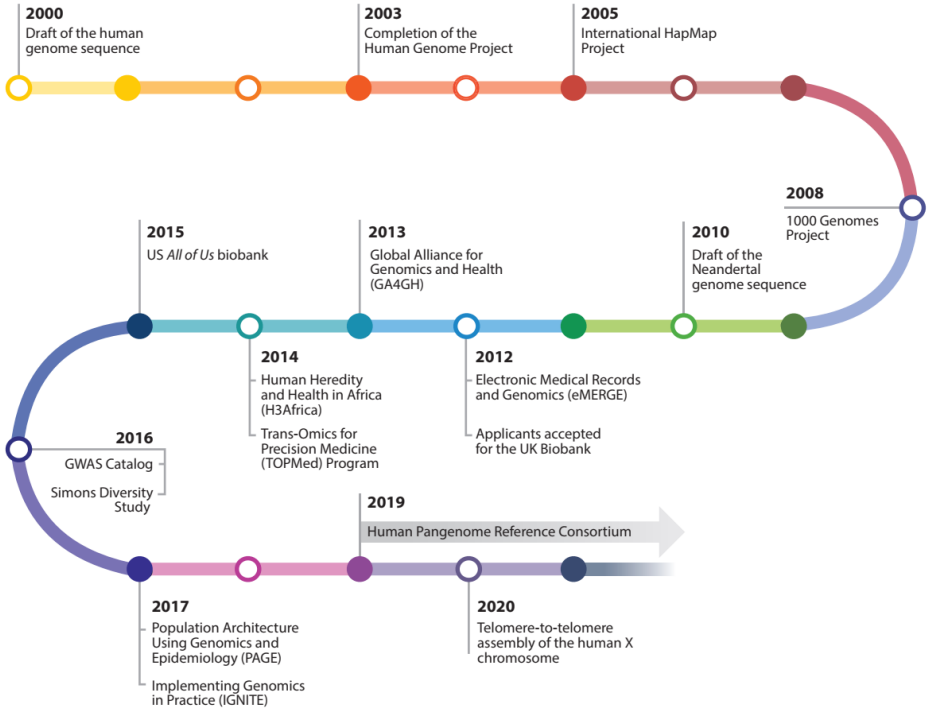
Annual Review of Genomics and Human Genetics

The Need for a Human Pangenome Reference Sequence

Karen H. Miga[1] and Ting Wang[2]

[1]UC Santa Cruz Genomics Institute and Department of Biomedical Engineering, University of California, Santa Cruz, California 95064, USA; email: khmiga@ucsc.edu

[2]Department of Genetics, Edison Family Center for Genome Sciences and Systems Biology, and McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri 63110, USA; email: twang@wustl.edu

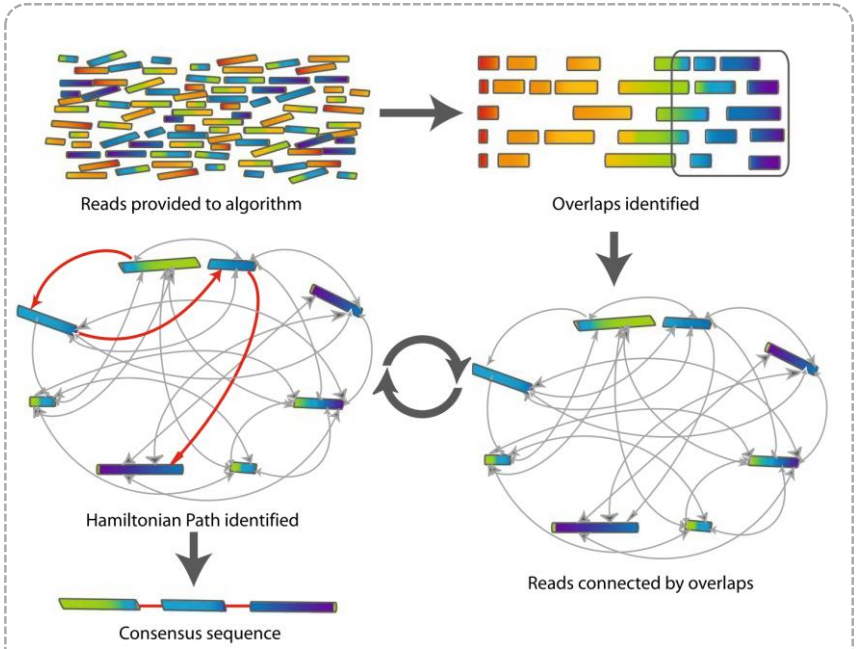


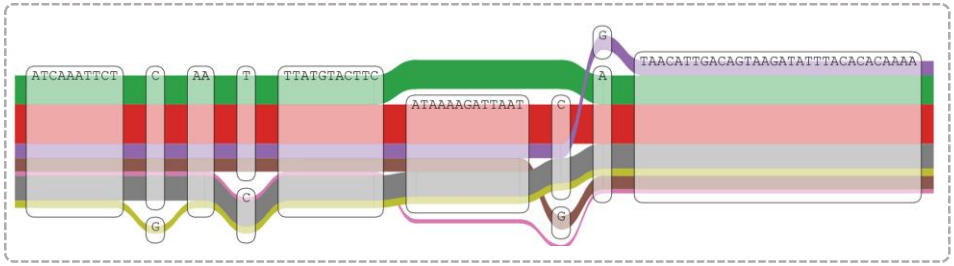
Miga & Wang (2021) ARGHG

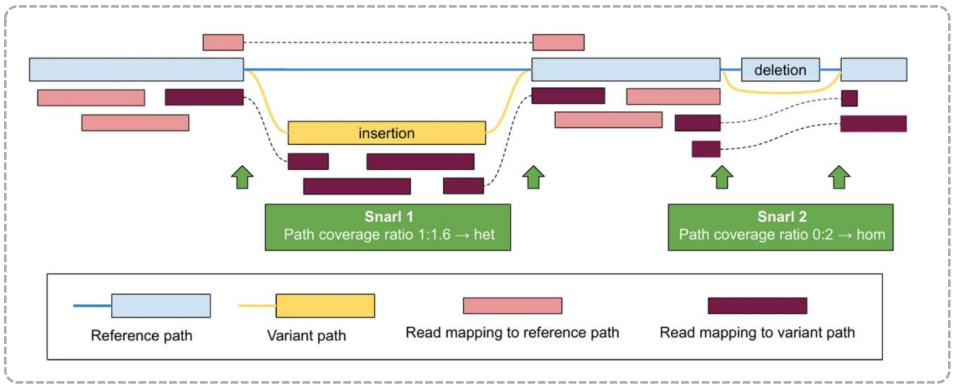
Construction of the human pan-genome

1. *De novo* assembly of long-read sequencing data

3. Identification of variants (SNVs, small INDELs, & SVs)
4. Construction of the pan-genome





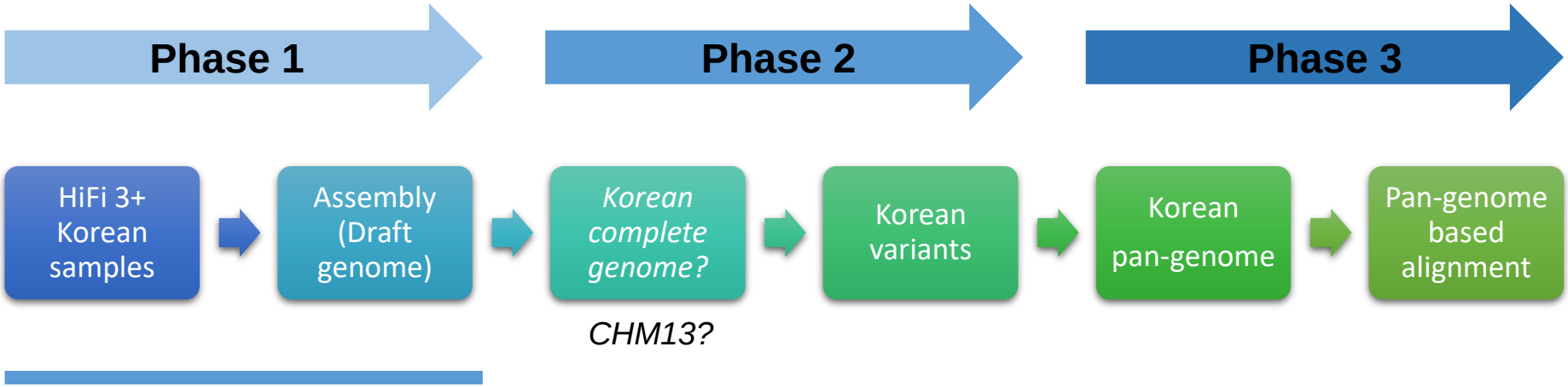


2. Construction of complete reference genome (template)

5. Pan-genome based alignment

Science (2021),
Genome biology (2020)

Roadmap of the Korea pan-genome project



During the phase 1, we performed *de novo* assembly of the human WGS data, generated from the long-read sequencing technique (PacBio HIFI)

Summary

- **The Korea National Genome Project (NGP)** enables the researchers to explore **clinical and genomic data from 23,833 participants (25,000 samples)** in the secure cloud service (Research Environment Platform)

- KOBIC launched **the Korean Pan-genome Project** for the better performance to **identify the uncharted structural variations** and **integrate numerous genetic variants from Korean population**

Acknowledgement

Dr. Seon-Young Kim, **Jin Ok Yang**, Ik Su Byeon, Dr. Jong-Hwan Kim, Dr. Kiwon Jang, Jae Ho Lee, Gun Woo Park, Dr. Byungwook Lee, **Dr. Sang-Ok Kim**, Bang-Hyuck Lee, Dr. Jongbum Jeon, Dr. Wonyong Jeong, Young Mi Sim, Dr. Seon-Kyu Kim, Dr. Pan-Gyu Kim, GunHwan Ko, Jong-Cheol Yoon, Dr. Jaeeun Jung, Taeyeon Hwang, Dongmin Jang, **Dr. Soobok Joe**

Dr. Jun-Hak Lee, Dr. Hyojin Kang, Dr. Yukyung Jun

Dr. Kwan Ik Lee, Dr. Misook Kwak, Dr. Jong-suk Park

Dr. Kang-Woo Lee, Dae Sung Kim

Dr. Hyun-Young Park, Dr. Myungguen Chung, Namhee Kim, Dr. Hee Youl Chai, Dr. Min Jin Go, Dr. Eugene Kim, Dr. Jung-Eun Kim

Dr. Murim Choi, **Dr. Jun Kim**, Jeongeun Lee

Dr. Joon-Yong An, Lizzy Choi, Ganghee Lee

NGP cohort Partnership

SNUH Seoul National University Hospital, UNIST, Yonsei University, Chosun University, 국립암센터 National Cancer Center

Supported & Funded by:

Ministry of Health and Welfare

Ministry of Science and ICT

Ministry of Trade, Industry and Energy

NRF

Thank you!