A report on recent activities at Korea Bioinformation Center

Seungwoo Hwang

Korea Bioinformation Center

kobic 국가생명연구자원정보센터 Korea Bioinformation Center

Korea Research Institute of Bioscience and Biotechnology

KRIBB Korea Research Institute of Bioscience & Biotechnology

The history of ABC and KOBIC

Symposium

Training course & Symposium

Training course

ABC

CJK

JK

CNCB-NGDC

2021

2019

2018

2017

2016

2015

2014

2012

2011

2010

2009

2008

2007

Scbit

2006

2005

2004

2003

2002

kobic Korea Bioinformation Center

NGIC

6. Seon-Young Kim

5. Yong-Kyung Choi

4. Ryan W. Kim

3. Kiejung Park

2. Sanghyuk Lee

1. Jong Bhak

Sangsoo Kim

KOBIC has been skyrocketing since 2019

New building!!!

New server room!!!

More people!!!

(2019. 2.)
27

(Now)
57
* Excluding interns

More funding!!!

(2019. 2.)
~4,800 million KRW
(~40 M USD)

(Now)
~27,700 M KRW
(~320 M USD)

* Including WGS costs

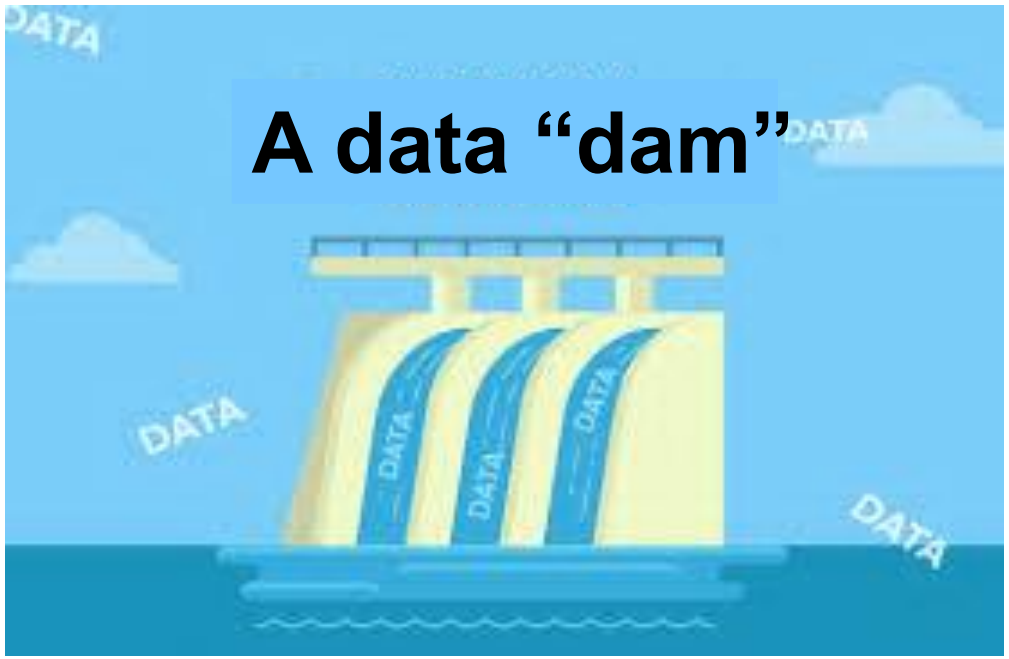
New big projects!!!

1. BioData Station Project
2. National Genome Project

BioData Station Project (2020~)

Goal of BioData Station Project

Long-term goal:
To collect **all types** of biological data (i.e., not just omics) that were produced from **all grants** from government funding

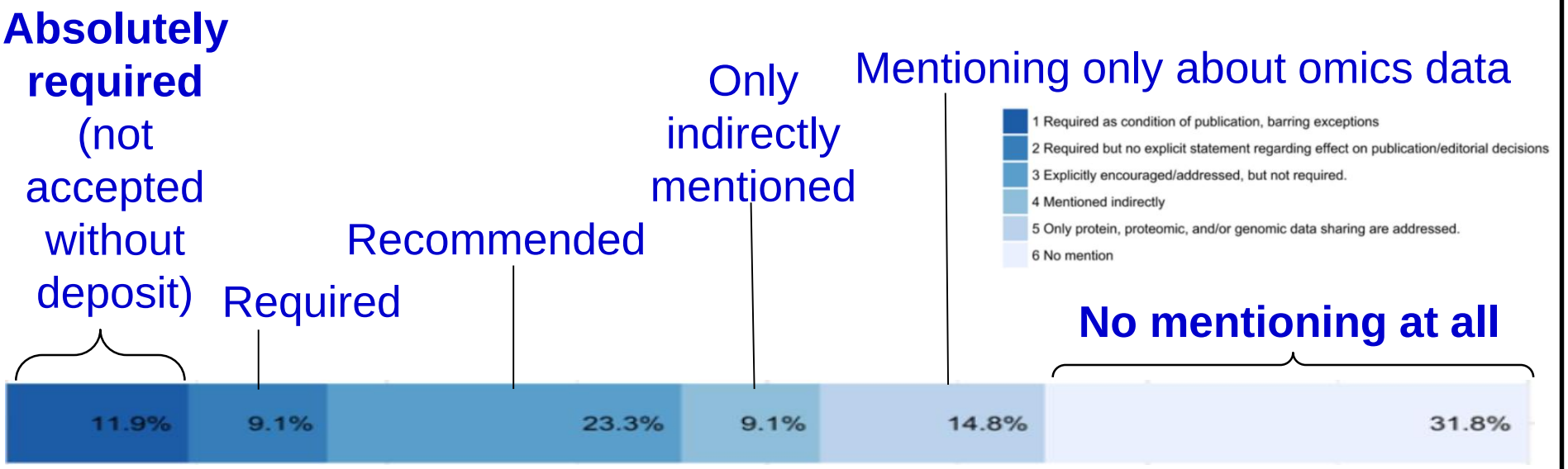


Why can't just we leave data sharing to journal publishers?

FAQ:

To publish a paper, I already deposit data somewhere (e.g., established international data repositories) after all. So, why do you bother to make another repository in Korea?

Data deposit requirements from journals are surprisingly low

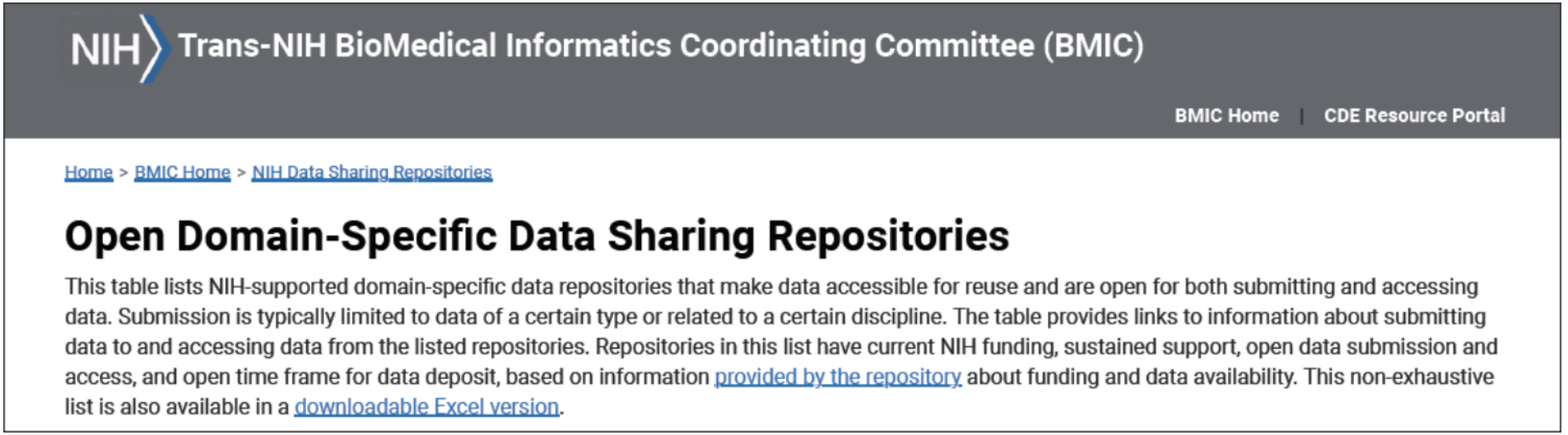


Vasilevsky et al (2017) PeerJ PMID:28462024
Reproducible and reusable research: are journal data sharing policies meeting the mark?

To make data sharing more active, funding agencies need to act too.

For more data sharing, funding agencies need to act too

While journals are international, funding agencies are national.



NIH does not list non-US repositories.

Need to establish a policy on data management

Final NIH Policy for Data Management and Sharing

Notice Number:

NOT-OD-21-013

Key Dates

Release Date:	October 29, 2020
Effective Date:	January 25, 2023

To get an approval for NIH-funded grants (of any amount), researchers need to file a Data Management Plan (DMP), and prove that they deposited the data according to the plan.

→ Korea is preparing such a policy.

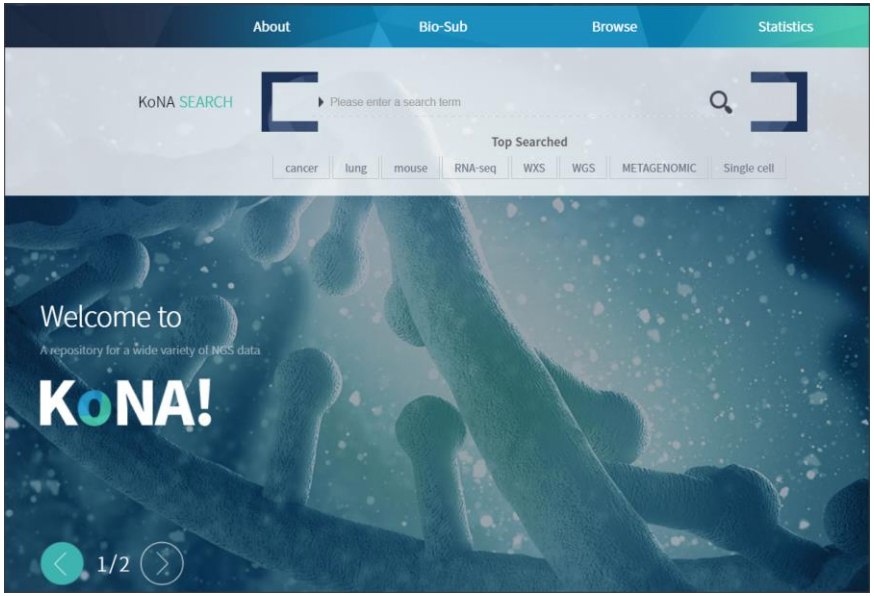
Need a proper format for data and metadata deposit

- **To maximize information content of deposited data**
- **Generally following existing formats**
  - **NGS:** INSDC
  - **Microarray:** NCBI GEO
  - **Nucleotide sequence:** NCBI GenBank
  - **Proteomics:** EBI PRIDE
  - **Metabolomics:** EBI MetaboLights
  - **Chemical activity:** NCBI PubChem Bioassay
  - **Microscopy image:** IDR, EBI EMPIAR, EBI BioImage
  - **Medical imaging:** DICOM, BIDS, ADNI
  - **Medical instrument R&D:** ClinicalTrials
  - **Public health:** NCD-RisC, WHO STEPS
  - **Pre-clinical:** SEND
  - **Plant metadata:** UPOV
  - **Other unstructured data:** EBI BioStudies, Dryad, FigShare, Zenodo
  - **…**

A pilot system for data deposit is now open

KBDS: Korea BioData Station

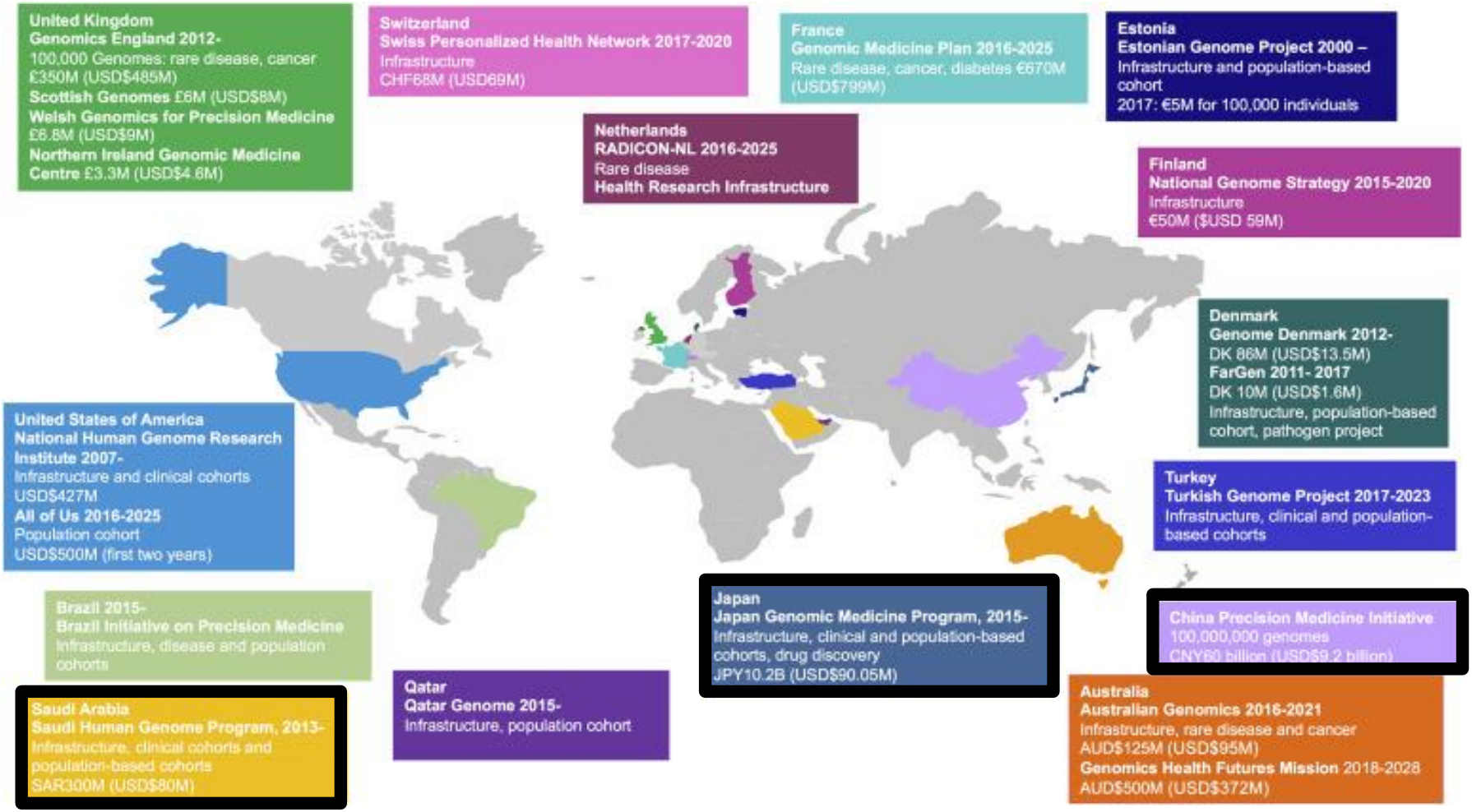
KoNA: Korean Nucleotide Archive



National Genome Project (2020~)

Nation-wide genome projects across the world

14 countries with nation-wide genome projects (as of 2019)



Stark *et al*., *AJHG* (2019)

National Genome Project (NGP) in Korea

Phase 1. Pilot project (2020~2022) 25,000 genomes

Phase 2. Main project (2023~2028) 1 million genomes

Mission Statement:
To provide large-scale genomic and clinical data
to scientific and industrial community
for the purpose of studying precision medicine,
while protecting the participants′ privacy

Overall scheme of the pilot phase of NGP

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

Clinical report & assessment

2 cohorts (newly sequenced by NGP)

Rare Disease (RD)
15,000 (mostly trio)

Normal Participants
2,500 (Korean Genome Epidemiological Study)

Sequencing data

Clinical info

3 cohorts (sequenced prior to NGP)

Normal Participants
1,600 (KGP)(Jong Bhak)

Autism Spectrum Disorder (500, familial)

Colorectal Cancer
(200, tumor and adjacent normal)

Sequencing data

Clinical info

NGP consortium

kobic

KiSTi
www.kisti.re.kr

질병관리청
KDCA

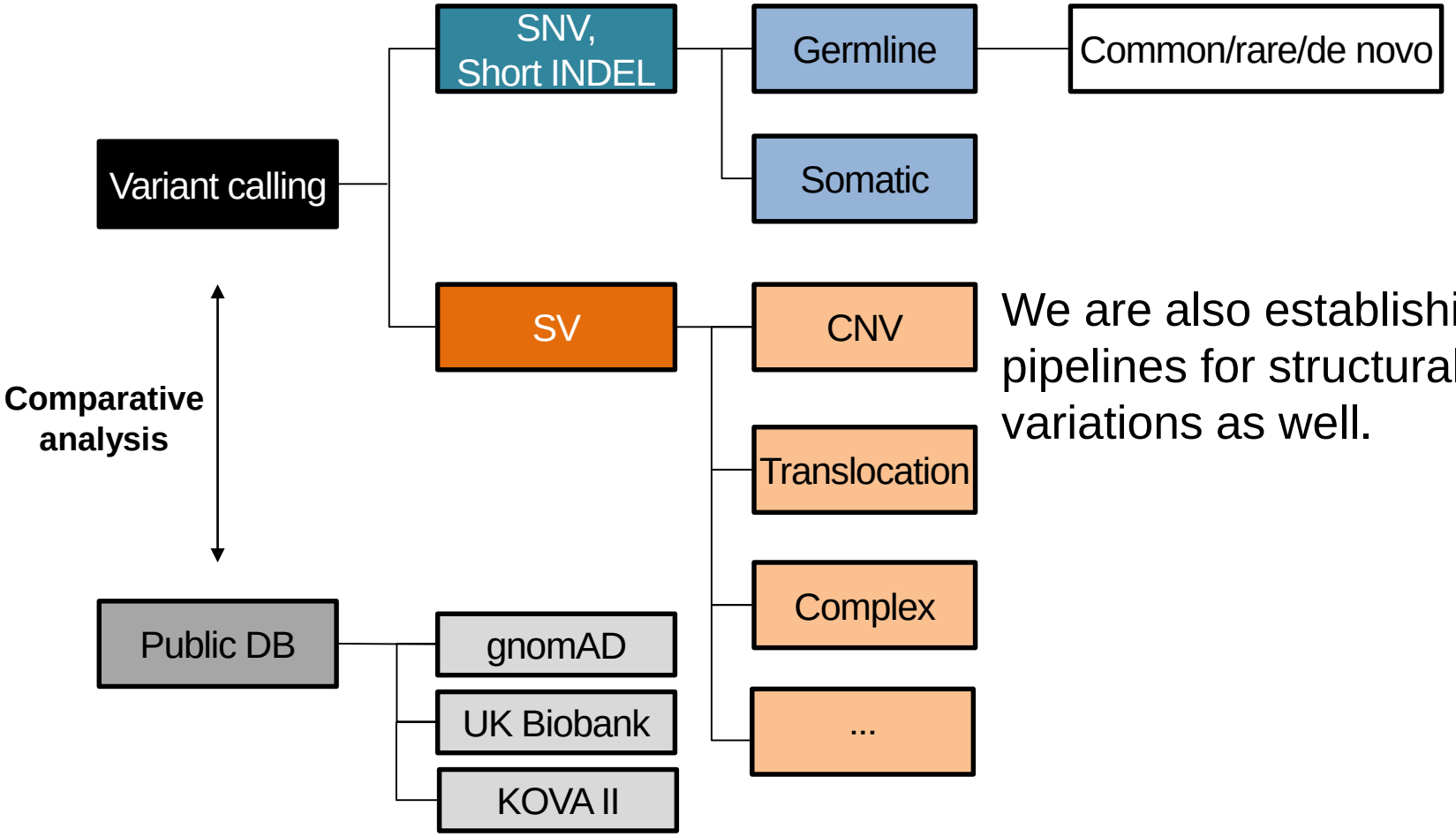
Research Environment Platform

- **Computing resources**
- **Cloud services**
- **Data sharing**

Integrative data analysis in the secure platform

Data processing and QC

We established data processing and QC pipelines to identify germline or somatic variants with high consistency and confidence.



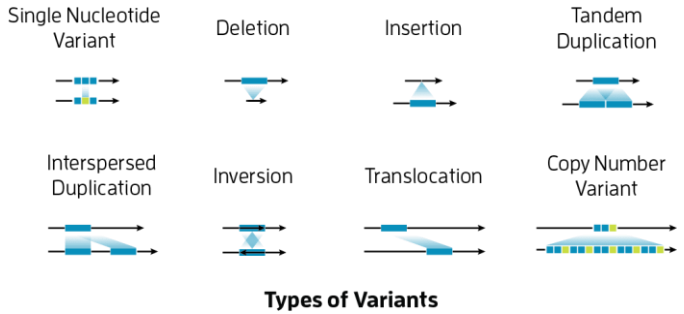
We are also establishing pipelines for structural variations as well.

We are also plan on performing a comparative analysis between our results and other public databases.

Data processing and QC report

The result from data processing and QC goes into the Research Environment Platform and provided to researchers.

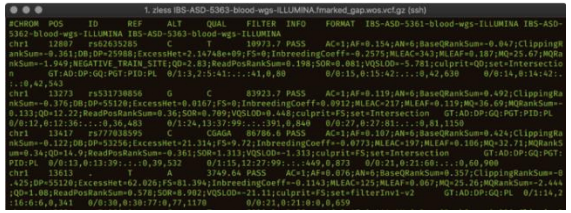
Variant calling & Data QC



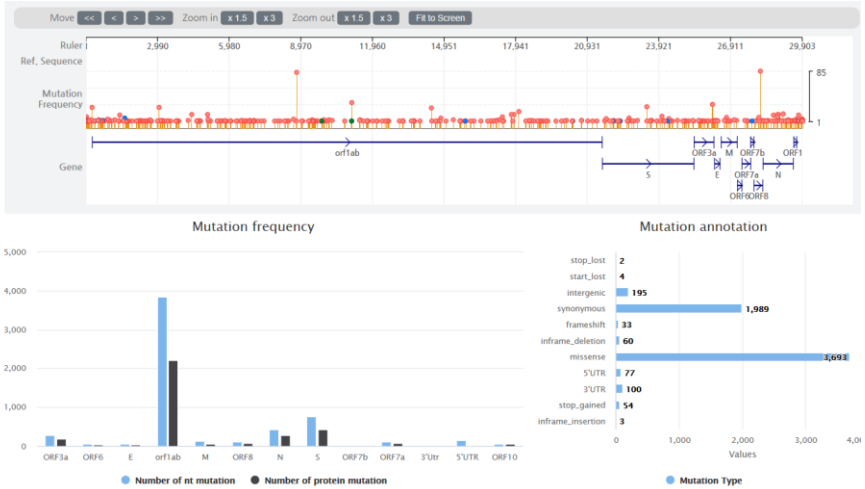
Types of Variants

https://www.pacb.com/

Variant data (VCF, TSV, *etc*.)



Data processing and QC report



Show 10 entries Search:

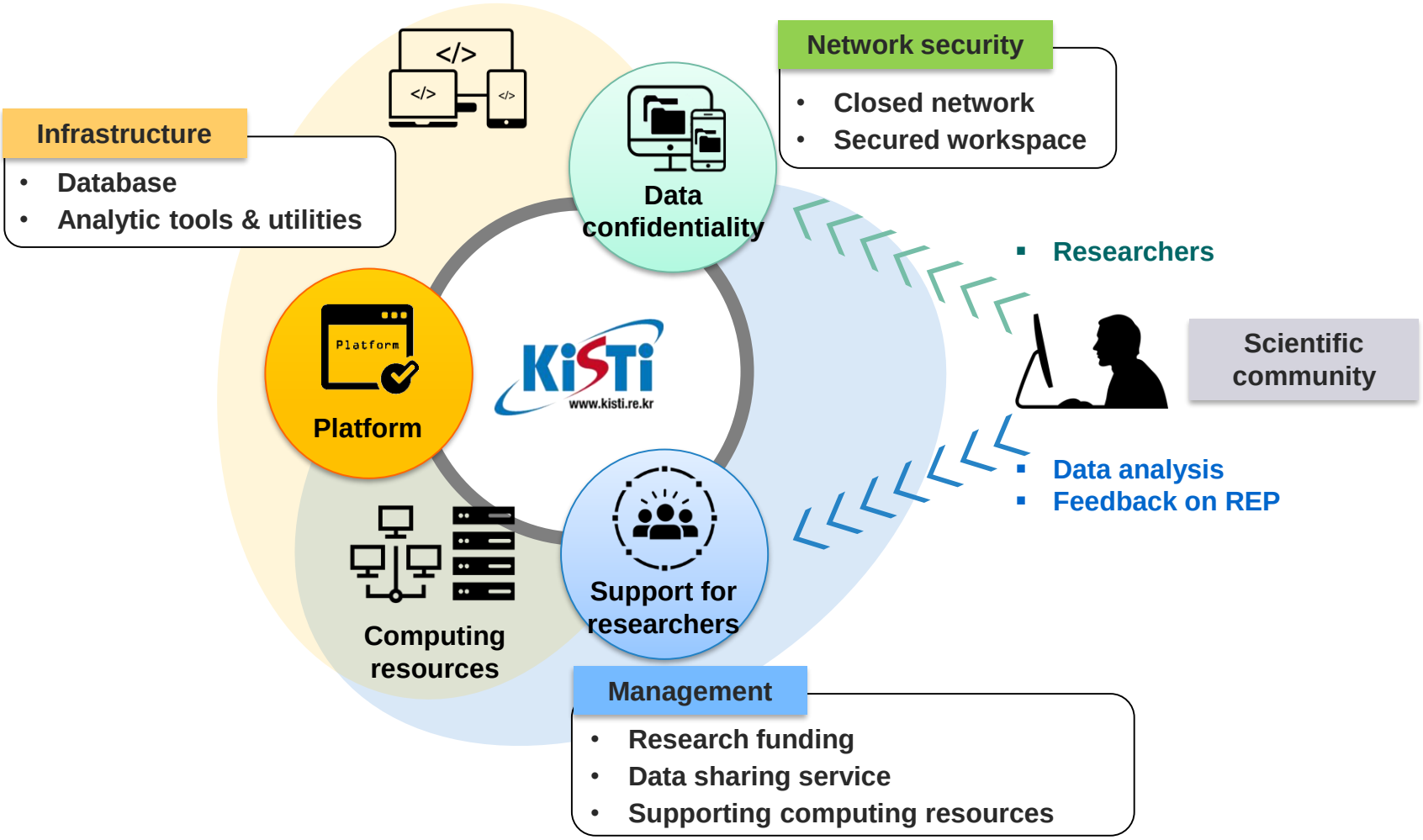
Effect	novel	novel%	Total
***TOTAL	17,252,867	100%	17,252,867
5_prime_UTR_truncation+exon_loss_variant	17	0%	17
5_prime_UTR_variant	400	0%	400
chromosome_number_variation	73	0%	73
disruptive_inframe_deletion	983	0.01%	983
disruptive_inframe_insertion	590	0%	590
downstream_gene_variant	7,973,598	46.22%	7,973,598
exon_loss_variant	17	0%	17
frameshift_variant	4,349	0.03%	4,349
inframe_deletion	473	0%	473

Showing 1 to 10 of 23 entries Previous 1 2 3 Next

The report contains important genetic features of each sample.

Research Environment Platform

Researchers can access and analyze the NGP data (processed WGS data and de-identified clinical information) at the Research Environment Platform, a secure workspace.



Next phase (2023~2028) of NGP

- The next phase of NGP aims to collect samples and data from **1 million participants**:
  - 0.3 million from existing cohorts
  - 0.7 million from newly recruited participants

- The 1 million participants will consist of
  - Rare disease patients
  - Cancer patients
  - Severe incurable disease patients
  - Normal/Healthy participants

- The data collection will include not only WGS, but also multi-omics data as well as longitudinal clinical information and personal health records.

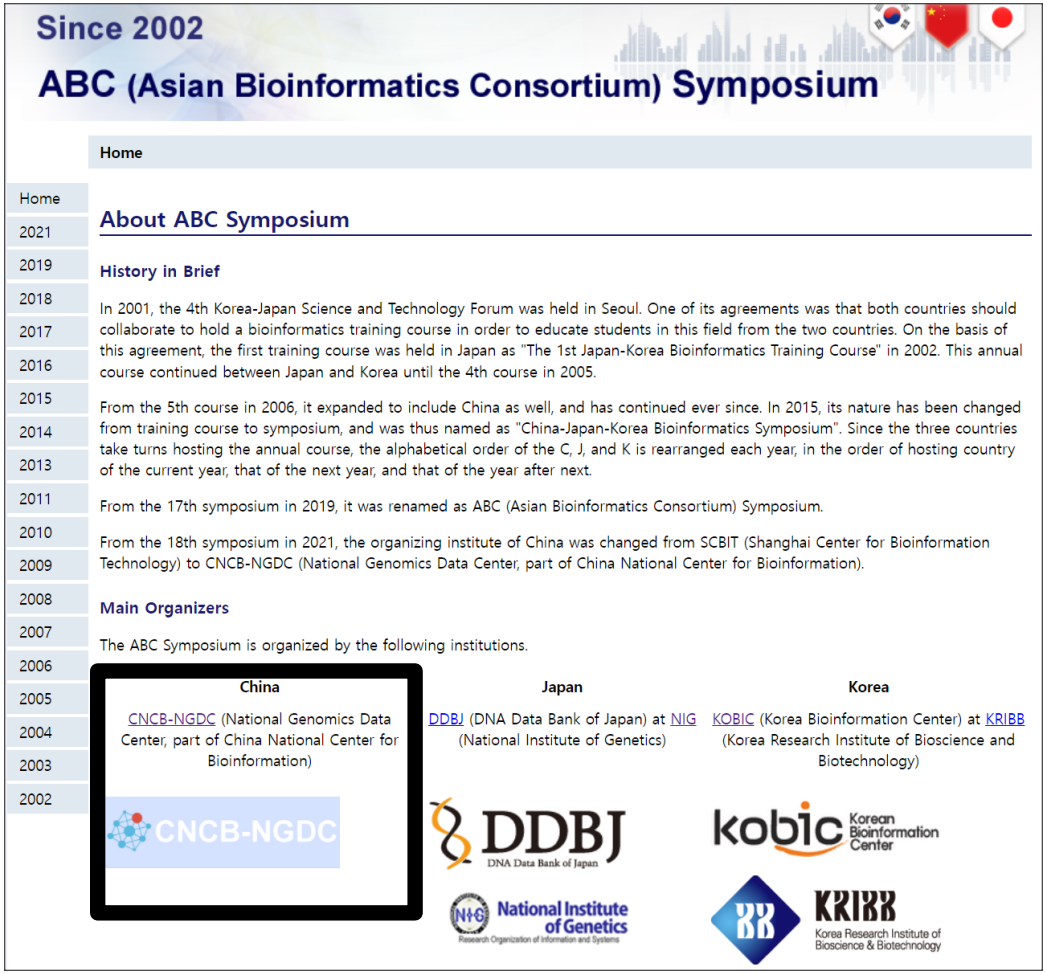
- All these datasets will be available in the expanded Research Environment Platform with dynamic consent system. So the researchers can use the data for a wide range of purposes without any ethical and legal issues.

Summary

National Genome Project

Korea BioData Station

May our collaboration continues!



http://cjk-bioinfo.org