



**National Genomics Data Center**

## **Database Resources of CNCB-NGDC**

---

**Yiming Bao**

**Director**

**National Genomics Data Center**

**Beijing, China**

**The 9<sup>th</sup> Big Data Forum**  
**Oct. 16, 2024 • Beijing**



**中国科学院北京基因组研究所（国家生物信息中心）**

BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

# AI needs data support

- **GPT-3:** 175 billion parameters
  - Cost (2020): \$4.6 million
- **GPT-4 (Human Brain):** 100 trillion parameters
  - Cost (2020): \$2.6 billion
  - Cost (2024): \$325 million
  - Cost (2028): \$40 million
  - Cost (2032): \$5 million





# AI needs data support

nature

<https://doi.org/10.1038/s41586-024-07487-w>

Accelerated Article Preview

## Accurate structure prediction of biomolecular interactions with AlphaFold 3

7R6R



Ground truth shown in gray

### Inputs and data sources

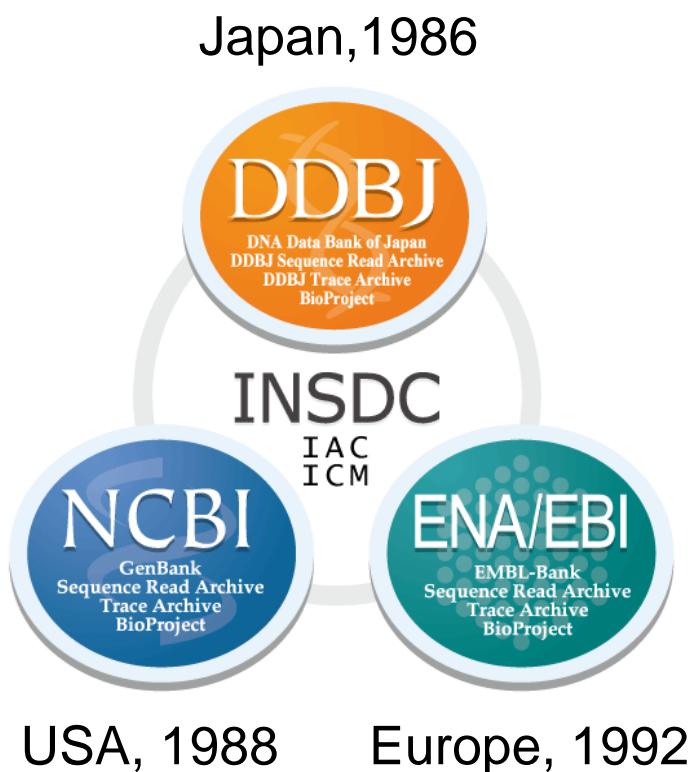
Inputs to the network are the primary sequence, sequences from evolutionarily related proteins in the form of a MSA created by standard tools including jackhmmer<sup>60</sup> and HHBlits<sup>61</sup>, and 3D atom coordinates of a small number of homologous structures (templates) where available. For both the MSA and templates, the search processes are tuned for high recall; spurious matches will probably appear in the raw MSA but this matches the training condition of the network.

One of the sequence databases used, Big Fantastic Database (BFD), was custom-made and released publicly (see 'Data availability') and was used by several CASP teams. BFD is one of the largest publicly available collections of protein families. It consists of 65,983,866 families represented as MSAs and hidden Markov models (HMMs) covering 2,204,359,010 protein sequences from reference databases, metagenomes and metatranscriptomes.

BFD was built in three steps. First, 2,423,213,294 protein sequences were collected from UniProt (Swiss-Prot&TrEMBL, 2017-11)<sup>62</sup>, a soil reference protein catalogue and the marine eukaryotic reference catalogue<sup>7</sup>, and clustered to 30% sequence identity, while enforcing a 90% alignment coverage of the shorter sequences using MMseqs2/Linclust<sup>63</sup>. This resulted in 345,159,030 clusters. For computational efficiency, we removed all clusters with less than three members, resulting in 61,083,719 clusters. Second, we added 166,510,624 representative protein sequences from Metaclust NR (2017-05; discarding all sequences shorter than 150 residues)<sup>63</sup> by aligning them against the cluster rep-

Nature (2021, 2024)

# International Nucleotide Sequence Database Collaboration (INSDC)



- NCBI: 1988, by US congress
- EBI: 1992, by EMBL
- DDBJ: 1986, by NIG of Japan
- NCBI, EBI and DDBJ form INSDC
- Establish international standard, exchange data daily, hold annual meeting
- Before papers are published, data need to be deposited into an international recognized database

# BIG Data Center

## Beijing Institute of Genomics (BIG), CAS

The BIG Data Center, officially founded in 2016, advances life & health sciences by providing freely open access to a variety of data resources, with the aim to translate big data into big knowledge and support worldwide research activities in both academia and industry.

*Translating big data into big discoveries*



**Deposition**



**Integration**



**Translation**

# Measures for the Management of Scientific Data

## 国务院办公厅印发《科学数据管理办法》

国务院办公厅印发《科学数据管理办法》（以下简称《办法》）

进一步加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好地为国家科技创新、经济社会发展和国家安全提供支撑

科学数据是国家科技创新发展和经济社会发展的重要基础性战略资源

《办法》明确了我国科学数据管理的

总体原则、主要职责、数据采集汇交与保存、共享利用、保密与安全等方面内容，着重从五个方面提出了具体管理措施

一  
明确各方职责分工，强化法人单位主体责任，明确主管部门职责，体现“谁拥有、谁负责”，“谁开放、谁受益”

二  
按照“分级分类管理，确保安全可控”的原则，主管部门和法人单位依法确定科学数据的密级及开放条件，加强科学数据共享和利用的监管

三  
加强知识产权保护，对科学数据使用者和生产者的行为进行规范，体现对科学数据知识产权的尊重

四  
要求科技计划项目产生的科学数据进行强制性汇交，并通过科学数据中心进行规范管理和长期保存，加强数据积累和开放共享

五  
提出法人单位要在岗位设置、绩效收入、职称评定等方面建立激励机制，加强科学数据管理能力建设

新华社发（朱禹制图）

2018/03

- Establishment of National Scientific Data Centers (NSDCs)
- Mandatory deposition in NSDCs for data from government-funded projects



# Establishment of 20 National Scientific Data Centers

## 科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知

国科发基〔2019〕194号

教育部、自然资源部、农业农村部、卫生健康委、市场监管总局、林草局、中科院、地震局、气象局、药监局科技、财务主管部门，广东省科技厅、财政厅：

为落实《科学数据管理办法》和《国家科技资源共享服务平台管理办法》的要求，规范管理国家科技资源共享服务平台（简称国家平台），完善科技资源共享服务体系，推动科技资源向社会开放共享，科技部、财政部对原有国家平台开展了优化调整工作，通过部门推荐和专家咨询，经研究共形成“国家高能物理科学数据中心”等20个国家科学数据中心、“国家重要野生植物种质资源库”等30个国家生物种质与实验材料资源库。

请你们组织依托单位进一步加强对各国家平台的管理，根据相关管理办法要求，制定国家平台五年建设运行实施方案，进一步明确国家平台功能定位和目标任务，梳理本领域科技资源体系架构，推进相关领域科技资源向国家平台汇聚与整合，强化科技资源开发应用与分析挖掘利用，提升科技资源使用效率和科技创新支撑能力，完善科技资源存储、管理和安全所需基础设施，健全网络安全保障体系，创新运行管理机制，加强评价考核组织管理，开展国际交流与合作，充分发挥法人单位主体责任，为科学研究、技术进步和社会发展提供高质量的科技资源共享服务。

特此通知。

附件：国家科技资源共享服务平台名单

科 技 部

财 政 部

2019年6月5日

- Undertaking the integration and exchange of scientific data in relevant fields
- Taking responsibility for the grading and categorizing, processing, and analysis of scientific data
- Ensuring the safety of scientific data and promoting the open sharing of scientific data in accordance with laws and regulations
- Strengthening scientific data exchanges and cooperation both domestically and internationally

# National Genomics Data Center (NGDC)

国家科技资源共享服务平台名单

序号	国家平台名称	依托单位	主管部门
1	国家高能物理科学数据中心	中国科学院高能物理研究所	中科院
2	国家基因组科学数据中心	中国科学院北京基因组研究所	中科院
3	国家微生物科学数据中心	中国科学院微生物研究所	中科院
4	国家空间科学数据中心	中国科学院国家空间科学中心	中科院
5	国家天文科学数据中心	中国科学院国家天文台	中科院
6	国家对地观测科学数据中心	中国科学院遥感与数字地球研究所	中科院
7	国家极地科学数据中心	中国极地研究中心	自然资源部
8	国家青藏高原科学数据中心	中国科学院青藏高原研究所	中科院
9	国家生态科学数据中心	中国科学院地理科学与资源研究所	中科院
10	国家材料腐蚀与防护科学数据中心	北京科技大学	教育部

11	国家冰川冻土沙漠科学数据中心	中国科学院寒区旱区环境与工程研究所	中科院
12	国家计量科学数据中心	中国计量科学研究院	市场监管总局
13	国家地球系统科学数据中心	中国科学院地理科学与资源研究所	中科院
14	国家人口健康科学数据中心	中国医学科学院	卫生健康委
15	国家基础学科公共科学数据中心	中国科学院计算机网络信息中心	中科院
16	国家农业科学数据中心	中国农业科学院农业信息研究所	农业农村部
17	国家林业和草原科学数据中心	中国林业科学研究院资源信息研究所	林草局
18	国家气象科学数据中心	国家气象信息中心	气象局
19	国家地震科学数据中心	中国地震台网中心	地震局
20	国家海洋科学数据中心	国家海洋信息中心	自然资源部



# China National Center for Bioinformation



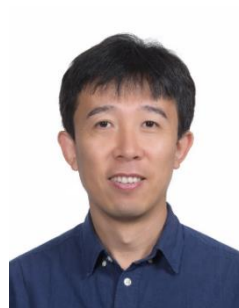
- China National Center for Bioinformation (CNCB) is affiliated with Beijing Institute of Genomics
- Bioinformation data archiving, storage, management and sharing
- Perform frontier research
- Achieve translation and application

# The NGDC Team

## ❑ Steering Advisors



## ❑ Professors



74  
students

56  
Staff

# NGDC national network

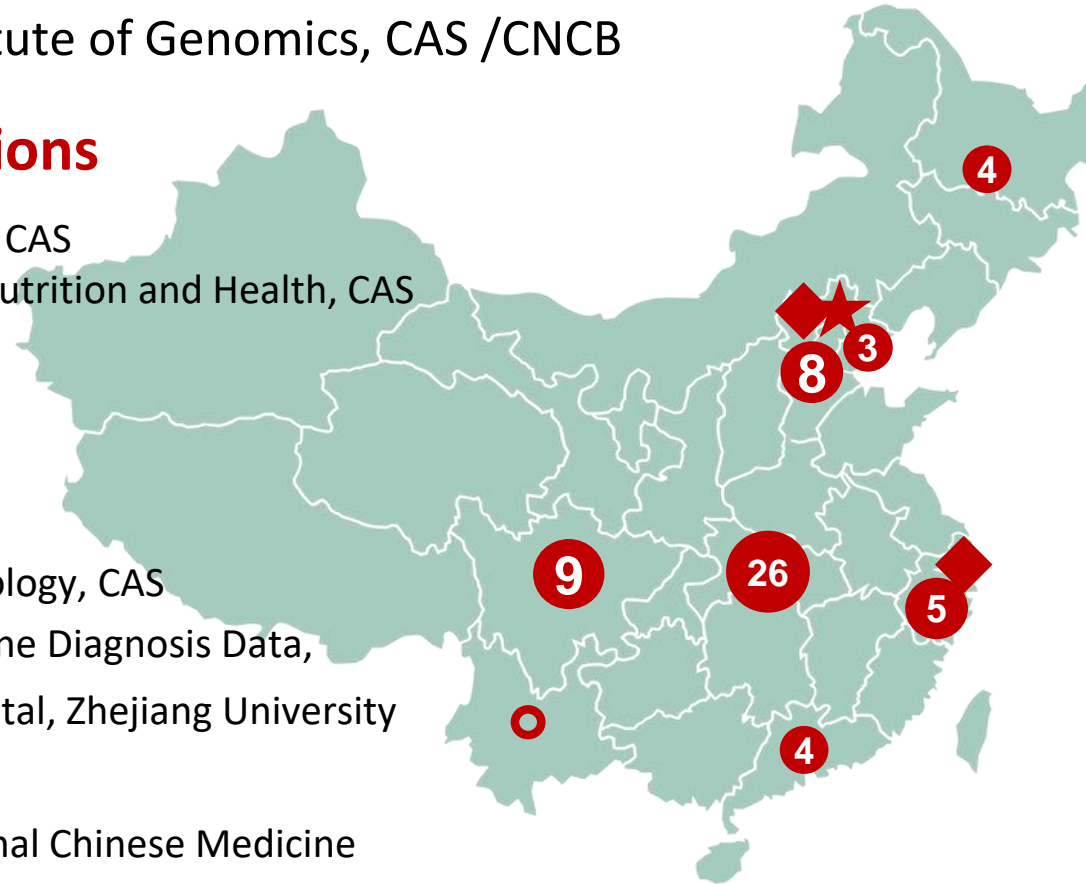
★ **Host:** Beijing Institute of Genomics, CAS /CNCB

## ◆ **Joint Organizations**

- ✓ Institute of Biophysics, CAS
- ✓ Shanghai Institute of Nutrition and Health, CAS

## ○ **Subcenters**

- ✓ Biodiversity Subcenter,  
Kunming Institute of Zoology, CAS
- ✓ Subcenter of Tumor Gene Diagnosis Data,  
The First Affiliated Hospital, Zhejiang University  
School of Medicine
- ✓ Subcenter of Traditional Chinese Medicine  
Experimental Research Center, CACMS

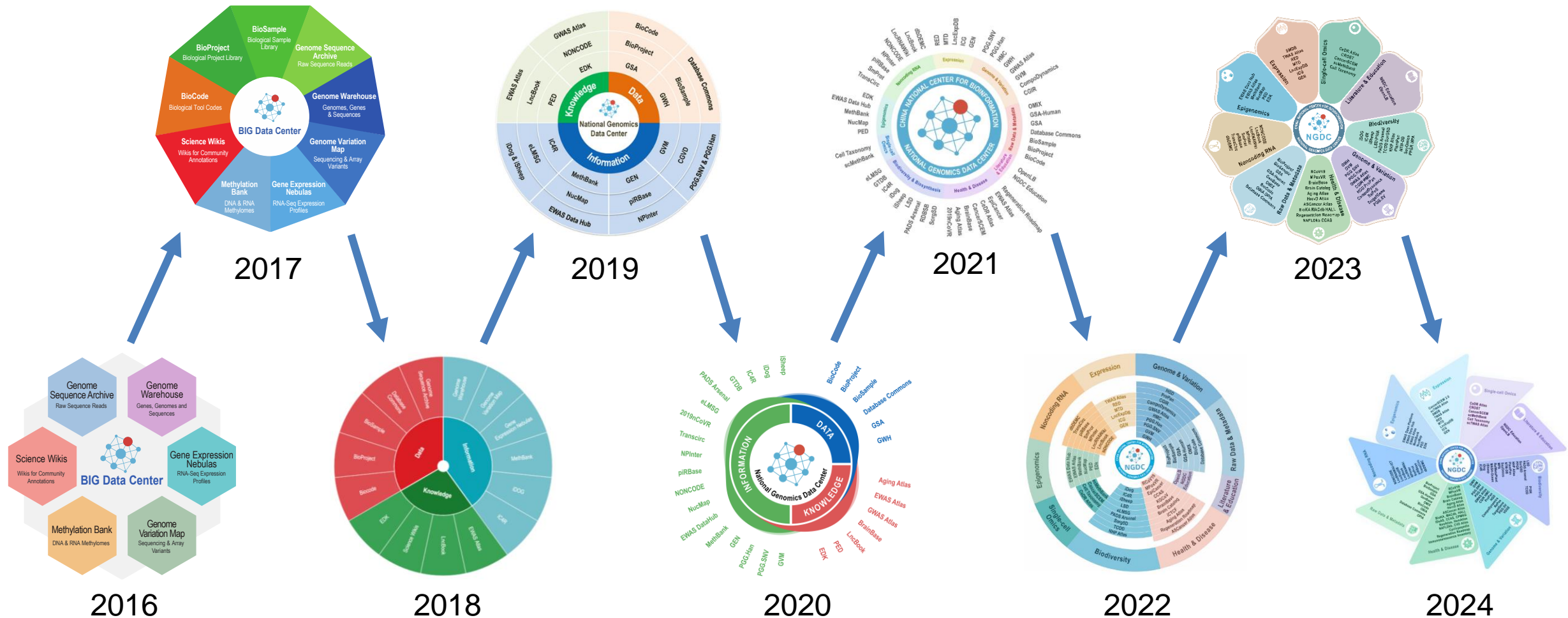


## ● **Partners**

- Peking University
- Capital Medical University
- Beijing Institutes of Life Science, CAS
- Institute of Zoology, CAS
- Tianjin University
- Huazhong University of Science and Technology
- Huazhong Agricultural University
- Harbin Medical University
- Institute of Medical Biology, CAMS
- Sun Yat-sen University
- West China Hospital, Sichuan University
- Zhejiang University



# The growing of capability



*Nucleic Acids Research 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024*

# Comprehensive Resources at CNCB-NGDC

The National Genomics Data Center (NGDC) advances life & health sciences by providing open access to a suite of resources, with the aim to translate big data into big discoveries and support worldwide activities in both academia and industry.

Find a bioproject, biosample, gene, protein, tool, database...

e.g., PRJCA000126; SAMC000385; tp53; EGFR; human; KaKs\_Calculator; GenBank

Scientific Data Archive System

Submit SDAS HGRIP BLAST RCoV19 OpenLB

**Resources**

- Raw Data & Metadata
- Genome & Variation
- Expression
- Noncoding RNA
- Epigenomics
- Single-cell Omics
- Biodiversity & Biosynthesis
- Health & Disease
- Literature & Education
- Tools

[See a full list of resources >>](#)

**Popular Resources**

- BioCode: Biological Tool Codes
- BioProject: Biological Project Library
- BioSample: Biological Sample Library
- GSA: Genome Sequence Archive
- GSA-Human: GSA for Human
- OMIX: Miscellaneous data
- GWH: Genome Warehouse
- GVM: Genome Variation Map
- Database Commons: Biological Database Catalog
- GEN: Gene Expression Nebulas
- MethBank: Methylation Bank
- BIT: Bioinformatics Toolkit

## ➤ Core omics databases

- BioProject
- BioSample
- Genome Sequence Archive (GSA)
- GenBase
- Genome Warehouse (GWH)
- Gene Expression Nebulas (GEN)
- Genome Variation Map (GVM)
- Methylation Bank (MethBank)

## ➤ Specialized databases

- RCoV19
- IC4R
- DogSD
- LncRNAWiki
- Database Commons

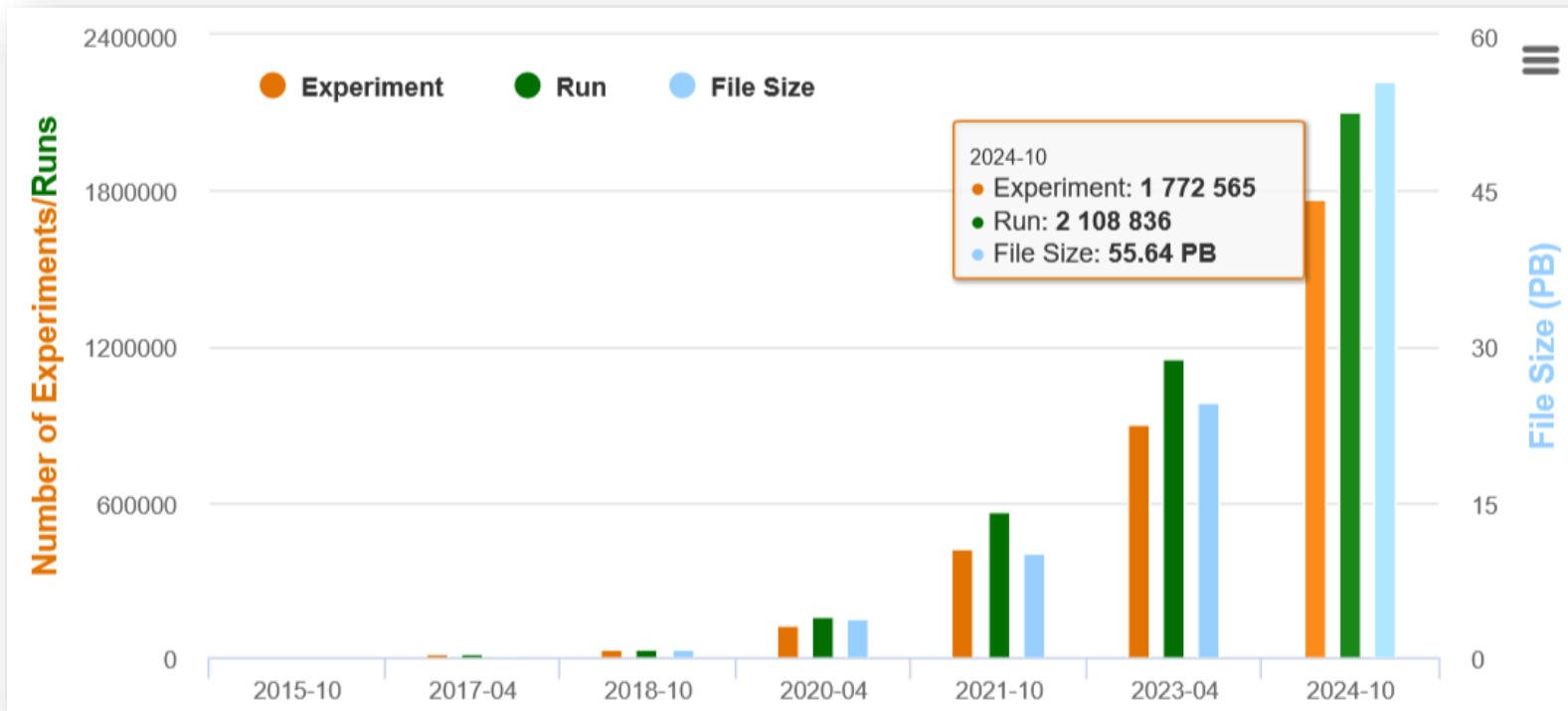
## ➤ Literatures

- OpenLB

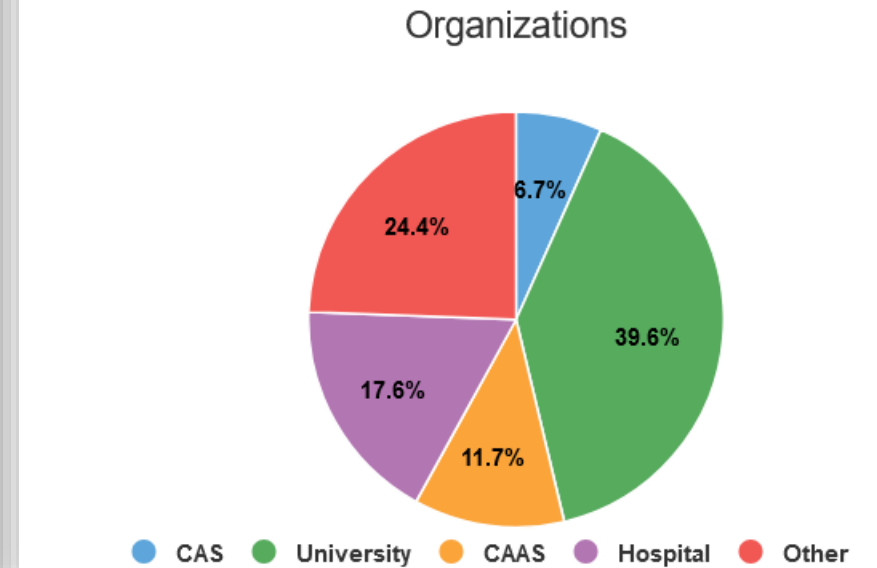
## ➤ Tools

- BLAST
- BIT

# Rapid Data Growth



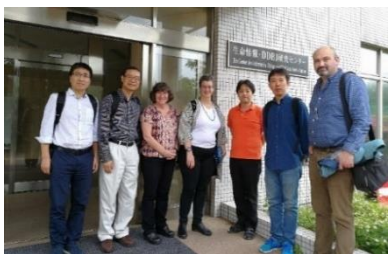
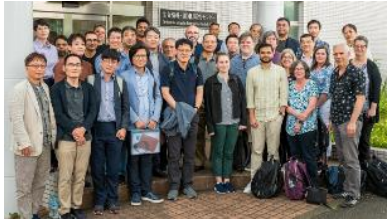
Submitters: 7322 / Organizations: 955 (890 public)



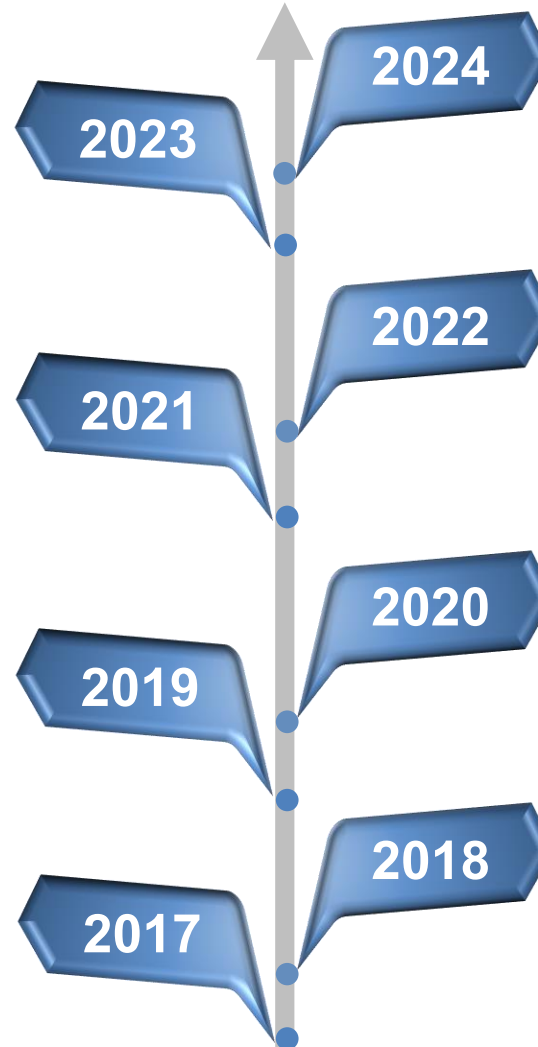
>55 PBase as of 2024-10-15



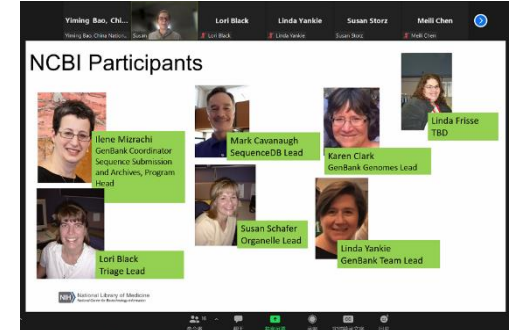
# Collaborations with INSDC



- **INSDC annual meeting (Japan)**
- **Visited NCBI**
- **INSDC annual meeting (online)**
- **INSDC drafted criteria for new members**
- **INSDC annual meeting (UK)**
- **INSDC annual meeting (Japan)**
- **Visited NCBI**



- **INSDC annual meeting (US)**
- **NGDC applied to join INSDC**
- **INSDC annual meeting (online)**
- **NCBI provided technical trainings for NGDC**
- **INSDC annual meeting (online)**
- **Exchanged data with NCBI**
- **Two-weeks' training in NCBI**



# Data Sharing with NCBI

## Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/Gilgit1/human/2020/PAK, complete genome

GenBank: MT240479.1

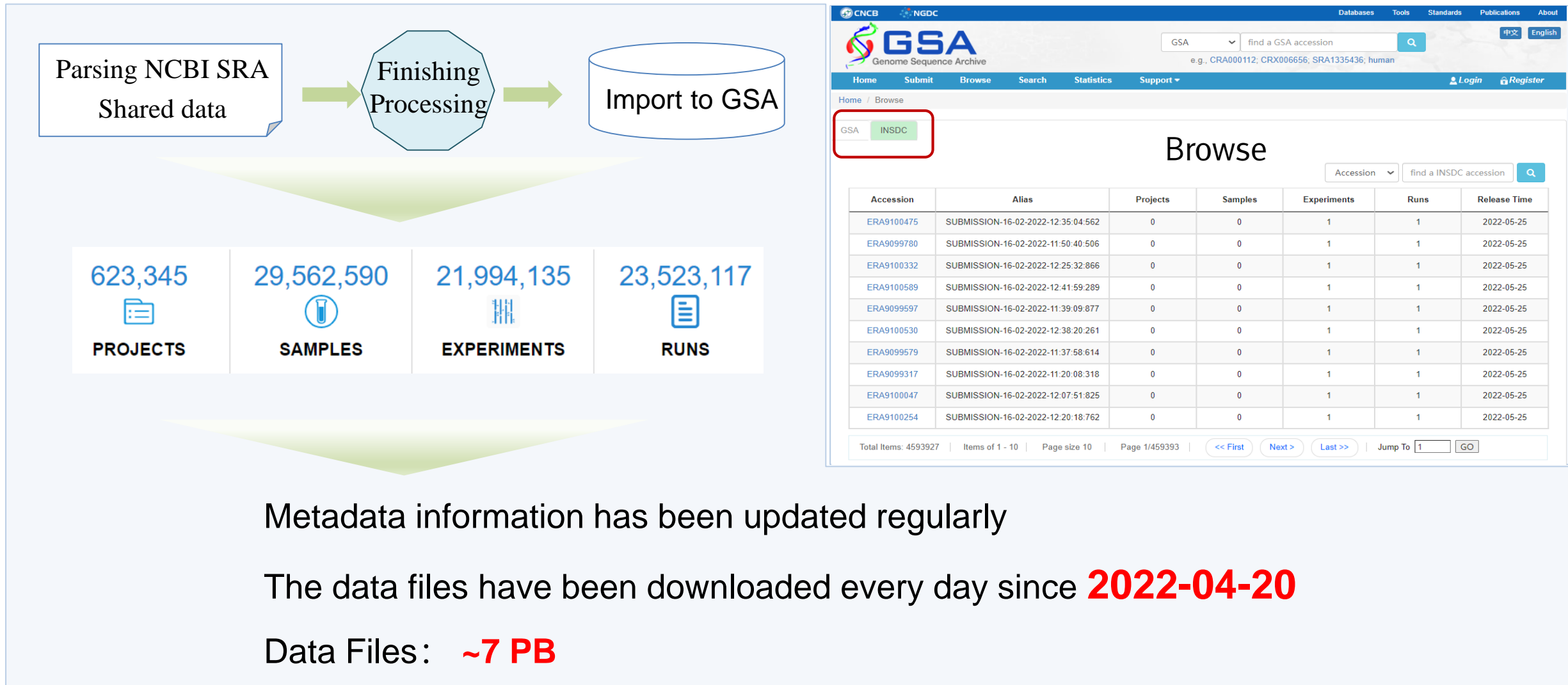
[FASTA](#) [Graphics](#)

[Go to:](#) ☐

LOCUS	MT240479	29836 bp	RNA	linear	VRL 25-MAR-2020
DEFINITION	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/Gilgit1/human/2020/PAK, complete genome.				
ACCESSION	MT240479	GWHACDD01000001			
VERSION	MT240479.1				
KEYWORDS	.				
SOURCE	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)				
ORGANISM	<u>Severe acute respiratory syndrome coronavirus 2</u> Viruses; Riboviria; Nidovirales; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.				
REFERENCE	1 (bases 1 to 29836)				
AUTHORS	Javed,A., Niazi,S.K., Ghani,E., Saqib,M., Janjua,H.A., Corman,V.M. and Zohaib,A.				
TITLE	Direct Submission				
JOURNAL	Submitted (25-MAR-2020) Department of Healthcare Biotechnology, National University of Sciences and Technology (NUST), Islamabad, Islamabad 46000, Pakistan				
COMMENT	This record was submitted to GenBank on behalf of the original submitter through Genome Warehouse (GWH, <a href="https://bigd.big.ac.cn/gwh/">https://bigd.big.ac.cn/gwh/</a> ) of the China National Center for Bioinformation (CNCB)/National Genomics Data Center (NGDC, <a href="https://bigd.big.ac.cn">https://bigd.big.ac.cn</a> ).				

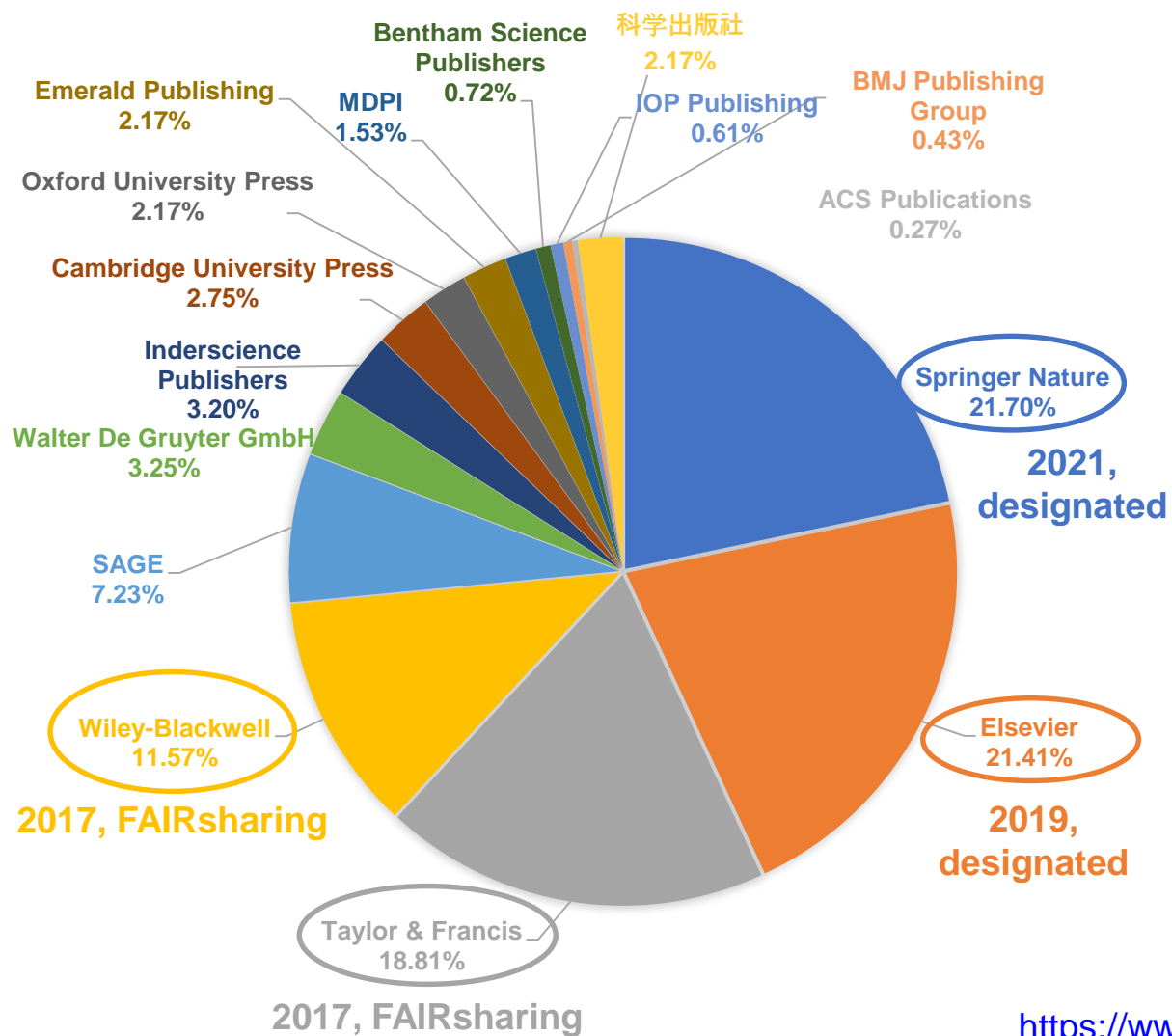
- Released the first genome sequence of a SARS-CoV-2 isolate from Pakistan
- Shared the sequence with INSDC through a data exchange mechanism established with NCBI
- Accession numbers of both NCBI and GWH of CNCB-NGDC are displayed and searchable
- This sets a good model for data sharing between databases

# Integration of International Data - GSA





# GSA Endorsed by Springer Nature and Major Publishers



## SPRINGER NATURE



### ✓ Nucleic acid sequence & Omics

Nucleic acid sequence data and metadata should follow the Genome Standards Consortium (GSC) guidance, which can be browsed at [FAIRsharing GSC collection](#).

#### Data types

DNA sequence data\*  
RNA sequence data\*  
Genome assembly data\*  
  
Genetic variation data

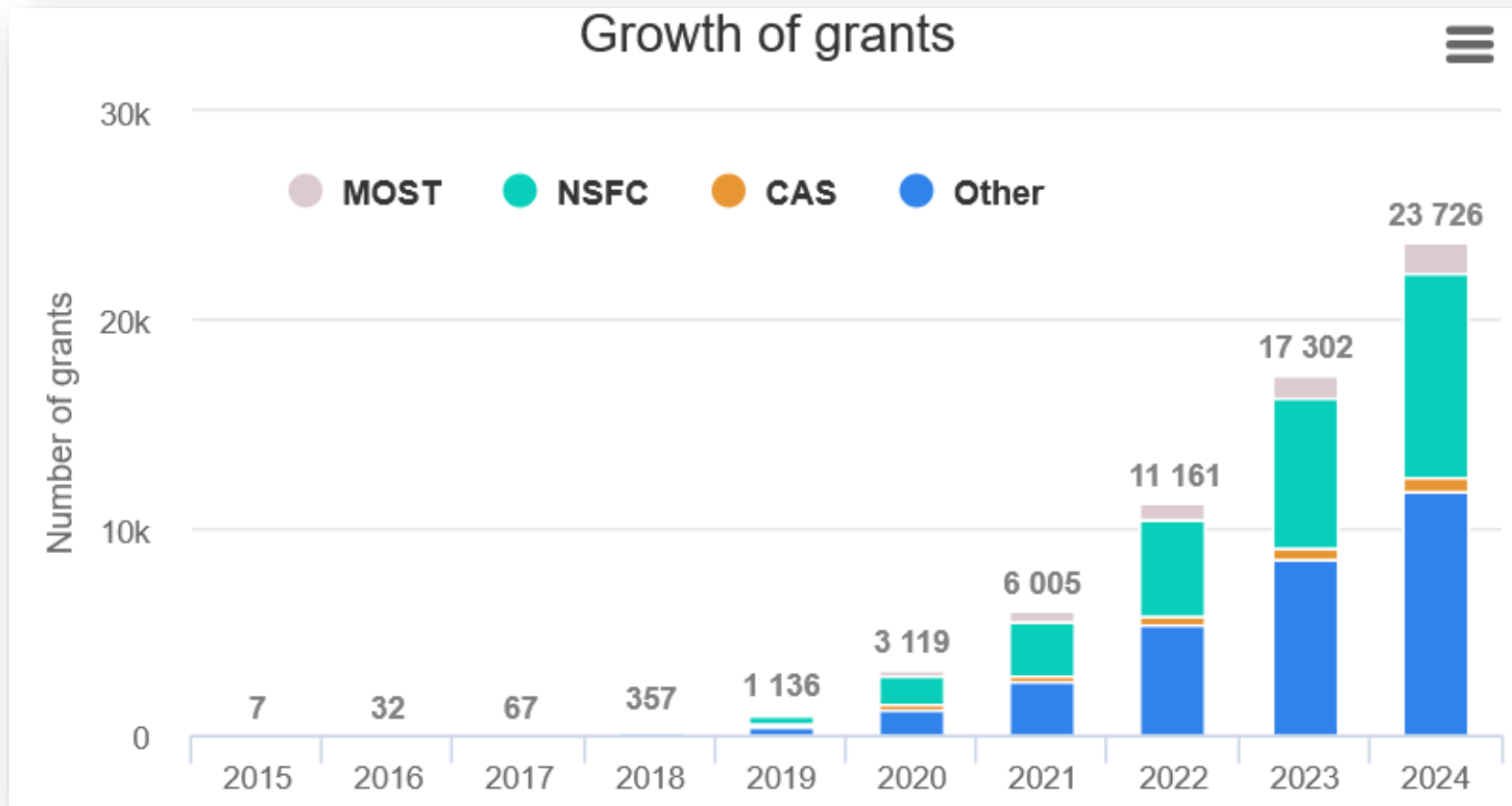
#### Repositories

Any [INSDC member repository](#)  
**Genome Sequence Archive (GSA)**  
  
dbSNP (human variations less than 50bp)  
dbVar (human variations greater than 50bp)  
European Variation Archive (EVA) (all species)  
**Genome Sequence Archive for Human (human variation)**

\* Novel DNA sequence, novel RNA sequence, and novel genome assembly data must be deposited to repositories that are part of the [International Nucleotide Sequence Collaboration](#) (INSDC), or those which are working towards INSDC inclusion (included in the table), unless there are privacy or ethics restrictions that prevent open sharing of such data. Novel DNA sequence, novel RNA sequence, and novel genome assembly data may in addition be deposited to any other repository (including regional or national repositories) as required.

<https://www.springernature.com/gp/authors/research-data-policy/repositories-bio/>

# Supporting >23k Research Grants




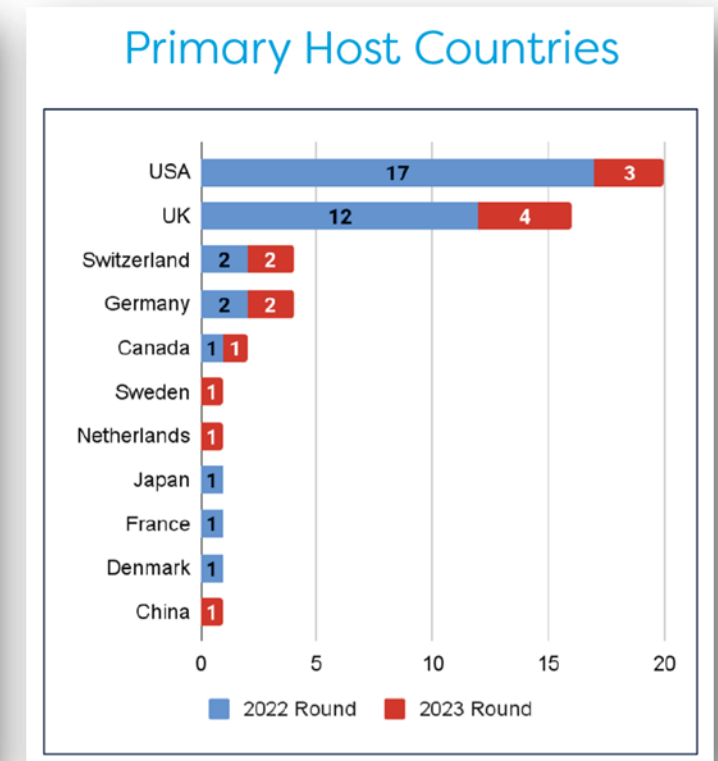
# International Submitters from 22 countries

所属单位					
Duke University	University of Southampton	Medizinische Hochschule Hannover	Germany	2	24
Temple University	Walailak University	IPK	Germany	1	12
University of Chicago	Chulalongkorn University	University Hospital Cologne	Germany	1	6
Yale University	Umeå University	Universität Witten/Herdecke	Germany	1	3
Iowa State University	National University of Singapore	Inserm	France	6	1347
Cornell University	Nanyang Technological University	Université de Lyon	France	3	106
University of California, San Francisco	Temasek Life Sciences Laboratory	Centre National de la Recherche Scientifique	France	1	77
Washington University (St. Louis)	Yonsei university	Université de Caen Normandie	France	1	44
University of Texas MD Anderson Cancer Center	Hamad Bin Khalifa University	MIVEGEC	France	1	3
Stanford School of Medicine	Quaid-i-Azam University	Cairo University	Egypt	1	3
	The University of Agriculture, Peshawar	Faculty of Health and Medical Sciences	Denmark	1	96
	Monash University Malaysia	Samplix	Denmark	1	8
	Kobe University	The University of Adelaide	Australia	1	8
	National University of Ireland, Galway	Macquarie University	Australia	1	6
		Innsbruck Medical University	Austria	2	11
		University Antwerpen	Belgium	1	191
		National Laboratory of Scientific Computation LNCC/MCTIC	Brazil	20	6608
		University of British Columbia	Canada	1	1



# GSA became a Global Core BioData Resource (GCBR)

 GLOBAL BIODATA COALITION		
WHAT WE DO ▾ MEMBERSHIP ▾ NEWS & RESOURCES ▾ CONTACT X in		
NAME	OVERVIEW	FUNDERS, 2017–2023
Host Countries: Primary, Additional		
<u><a href="#">GSA: Genome Sequence Archive</a></u>  China	The Genome Sequence Archive (GSA) is a data repository for collecting, archiving, managing and sharing raw genome sequence data.	Ministry of Science and Technology of China, National Key R&D Program of China, Chinese Academy of Sciences, International Union of Biological Sciences, The Professional Association of the Alliance of International Science Organisations, National Natural Science Foundation of China



<https://globalbiodata.org/what-we-do/global-core-biodata-resources/list-of-current-global-core-biodata-resources/>

# Integration of International Data - GenBase



GenBase

Home

Submit

Search

Downloads

Statistics

Standards

Help

Login

Register

语言/Language

**GenBase** is a genetic sequence database that accepts user submissions (mRNA, genomic DNAs, ncRNA, or small genomes such as organelles, viruses, plasmids, phages from any organism) and integrates data from INSDC.

Nucleotide

Type your keywords

Search

Advanced

e.g. C\_AA001108.1; MH011443.1; GB0003962

## Archived Data



Species  
592,438



Nucleotides  
374,114,285



Proteins  
618,500,336

## Recent Updates

2024-07-25

The sequence update function is available. [link](#).

2024-06-25

GenBase paper published online (PMID:38913867).  
Welcome to the citation.

2024-05-13

GenBase is one of the registered repositories in  
FAIRsharing.

2024-05-08

The statistics of data exchange is available.

## INSDC(GenBank) Integration

Update Date: 2024-10-15  
#Nucleotides: 6,405  
#Proteins: 29,294

## GenBase Release

Release Date: 2024-10-13  
#Nucleotides: 588  
#Proteins: 6,920

## New

SARS-CoV-2 Fast Submission

## GenBase-supported Deposition

GenBase is one of the registered  
repositories in **FAIRsharing** and  
**r3data**.

## Related Links

GSA

## Problems or Questions?

### Guide for submission

If you have any question or would like to give us any  
suggestion/comment or report a bug, please feel free to  
contact us.

Email: [genbase@big.ac.cn](mailto:genbase@big.ac.cn)

QQ group: 629388189

Tel: 86-10-84097298

Work hours: 9:00 - 17:00

## How to cite

### Recommended citation style:

The data reported in this paper have been deposited in the GenBase[1] in National Genomics Data Center [2], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number C\_AA000000 that is publicly accessible at <https://ngdc.cncb.ac.cn/genbase>.

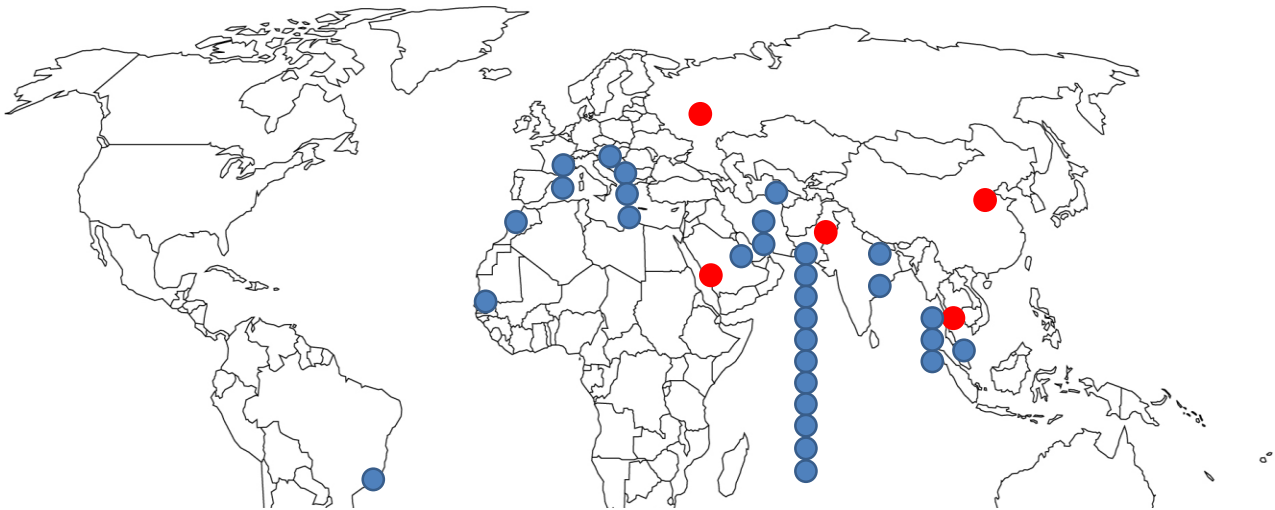
### References:

[1] GenBase: A Nucleotide Sequence Database. Genomics Proteomics Bioinformatics 2024, qzae047 [PMID=38913867].

[2] Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. Nucleic Acids Res 2024, 52(D1):D18-D32 [PMID=38018256].

# Global Biodiversity and Health Big Data Alliance (BHBD)

❑ The BHBD alliance is established based on the International Union of Biological Sciences (IUBS), aiming to promote the open sharing of global biodiversity and health big data



● Council Member 5      ● Regular Member 30



BIG, CAS  
China



QAU  
Pakistan



VIGG  
Russia



KAUST  
Saudi Arabia



CU  
Thailand

**35** **17**  
**Members Countries**  
(As of Jun. 2024)

**Regular Members:**

Brazil	1	Kazakhstan	1	Qatar	1
France	2	Malaysia	1	Senegal	1
Hungary	1	Morocco	1	Serbia	3
India	1	Nepal	1	Thailand	3
Iran	2	Pakistan	11		

The BHBD Alliance was officially established on Oct 14, 2018

BHBD Members and Their Distribution

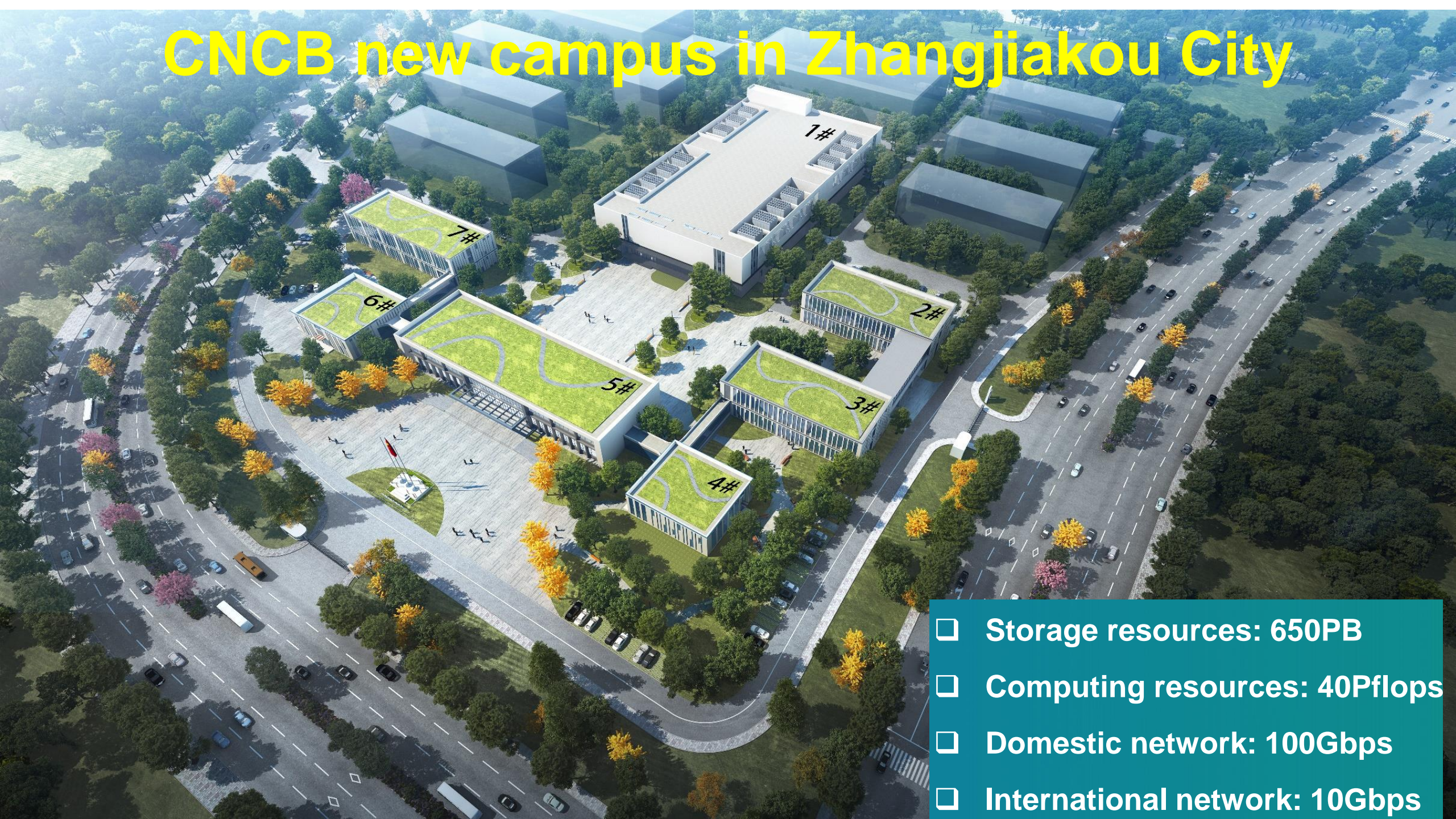


# Grants Awarded for International Collaboration

Funding Agency	Project Title	Duration	Collaborators	Amount
IUBS	Open Biodiversity and Health Big Data Initiative	2019-2022, 2023-2026	Multiple countries	Euro 30,200 Euro 22,500
ANSO	Global Biodiversity and Health Big Data Alliance	2020-2022, 2023-2026	Multiple countries	RMB 750,000 RMB 300,000
ANSO	Precision warning method for high-risk variants of emerging infectious diseases	2023-2025	Brazil, France, Pakistan	RMB 1,300,000
ANSO	Whole genome sequencing and miRNA biomarkers for an enhanced understanding of mechanism of tuberculosis infection in cynomolgus macaques ( <i>Macaca fascicularis</i> ): A translational knowledge to clinical study	2023-2025	Thailand, USA	US\$ 150,000
NSFC	SARS-CoV-2 Network for Genomic Surveillance in Brazil, Russia, India, China and South Africa (NGS BRICS)	2021-2022	Brazil, Russia, India, South Africa	RMB 2,000,000
CAS	Global Genomics Data Sharing	2023-2025	USA	RMB 800,000



# CNCB new campus in Zhangjiakou City



- ❑ Storage resources: 650PB
- ❑ Computing resources: 40Pflops
- ❑ Domestic network: 100Gbps
- ❑ International network: 10Gbps



# Acknowledgements

## Steering Advisors:

- Runsheng Chen
- Guoping Zhao

## Scientific Advisors:

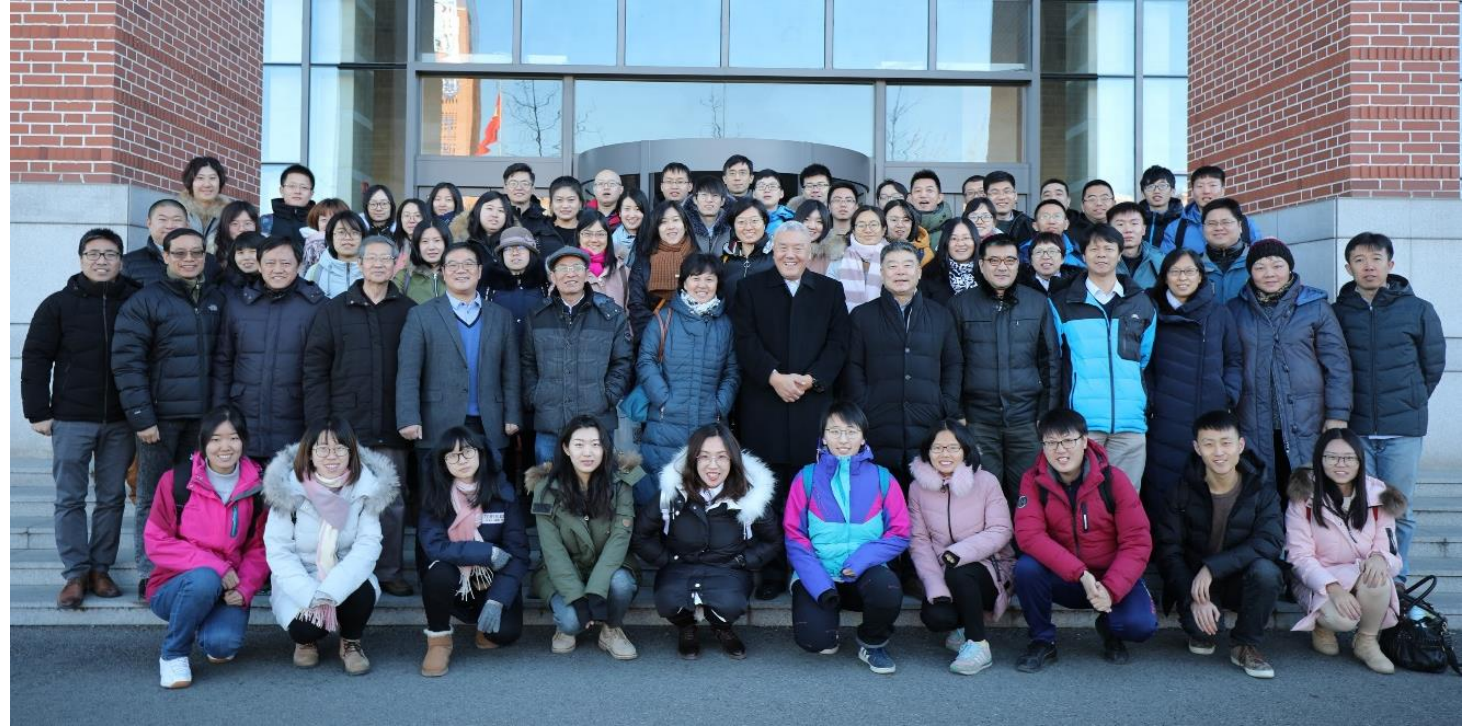
- Amos Bairoch (SIB)
- Guy Cochrane (EBI)
- Frank Eisenhaber (BI)
- Takashi Gojobori (KAUST)
- Yixue Li (CAS)
- Jingchu Luo (PKU)
- Ilene Mizrachi (NCBI)
- Yasukazu Nakamura (DDBJ)
- Weimin Zhu (CAS)

## Center Collaborators:

- SINH: Guoqing Zhang
- IBP: Shunmin He

## Strategic Partners:

- Ming Chen
- Qinghua Cui
- Feng Gao
- Ge Gao
- Xin Gao
- An-Yuan Guo
- Tao Jiang
- Cheng Li
- Chuan-Yun Li
- Xia Li
- Jian Ren
- Yun Xiao
- Yu Xue
- Yong Zhang
- Fangqing Zhao



中华人民共和国科学技术部  
Ministry of Science and Technology of the People's Republic of China



中国科学院  
CHINESE ACADEMY OF SCIENCES



中华人民共和国国家卫生健康委员会  
National Health Commission of the People's Republic of China



# Thank You!



**NGDC**



**BHBD Alliance**



**baoym@big.ac.cn**