National Genomics Data Center

China National Center for Bioinformation Promotes Open Science

Yiming Bao

Director

National Genomics Data Center

Beijing, China

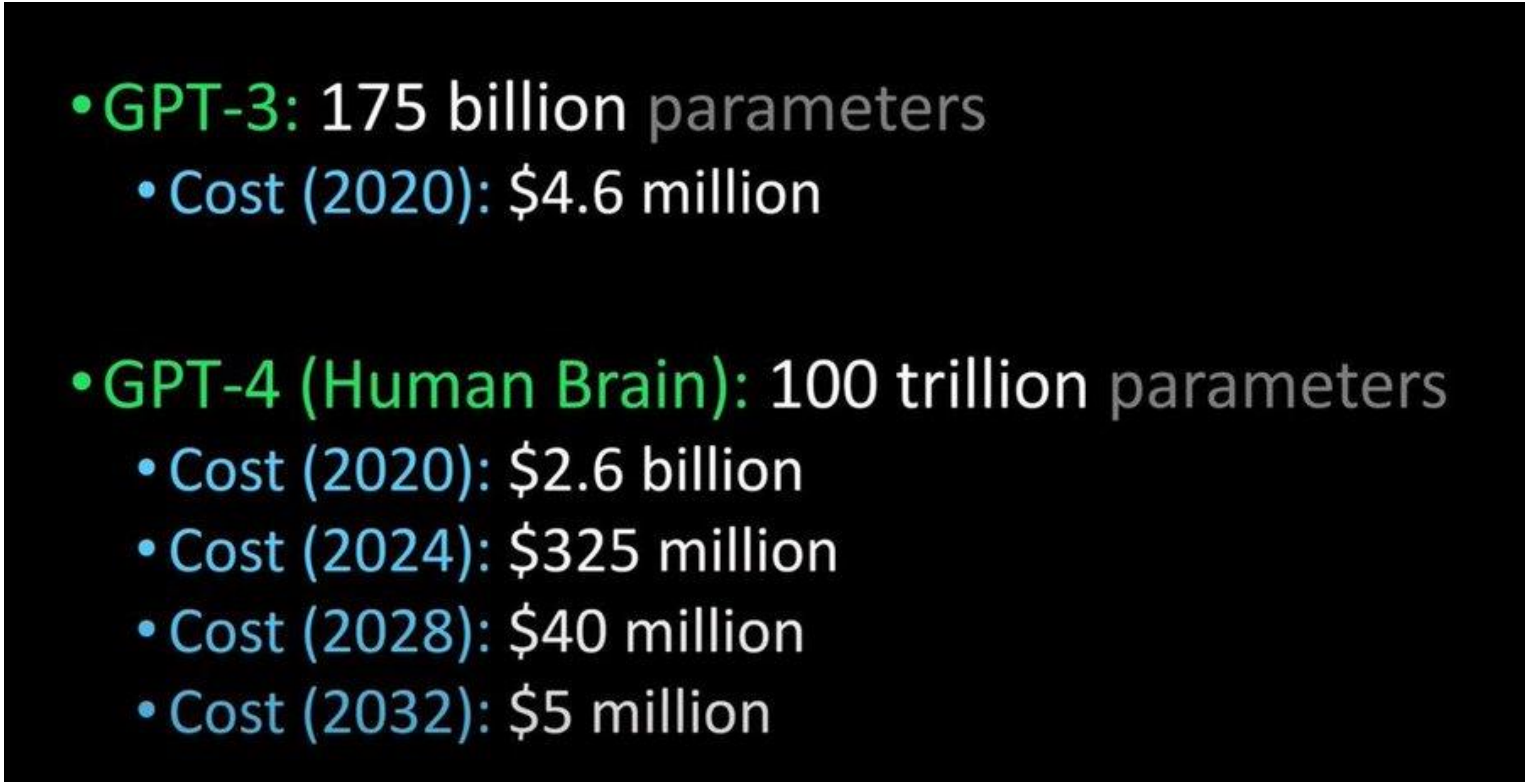
APANAC 2023

Sep. 28, 2023 • Panama

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



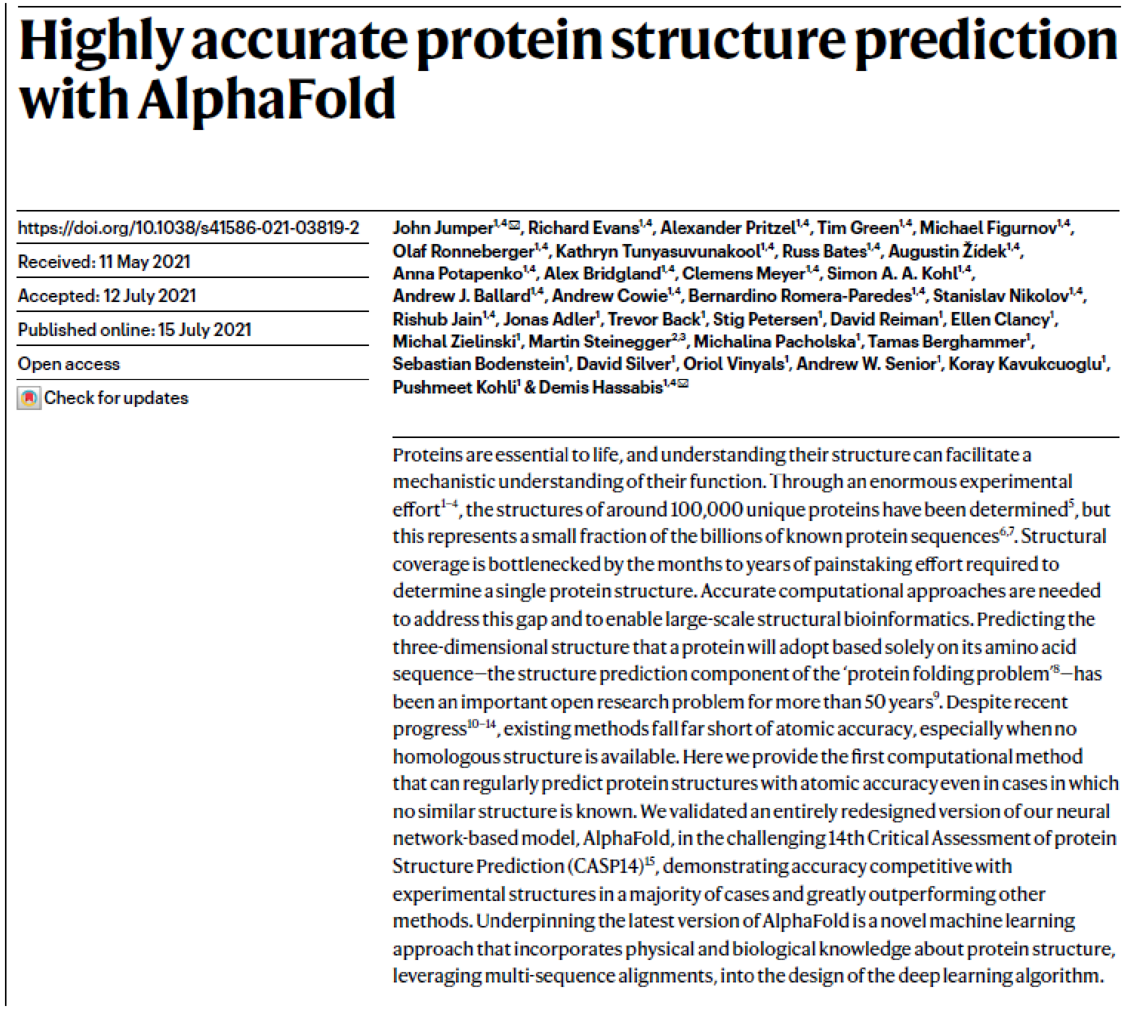
- GPT-3: 175 billion parameters
  - Cost (2020): $4.6 million
- GPT-4 (Human Brain): 100 trillion parameters
  - Cost (2020): $2.6 billion
  - Cost (2024): $325 million
  - Cost (2028): $40 million
  - Cost (2032): $5 million

OpenAI

ChatGPT: Optimizing Language Models for Dialogue

AI needs data support

DeepMind



Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

Accepted: 12 July 2021

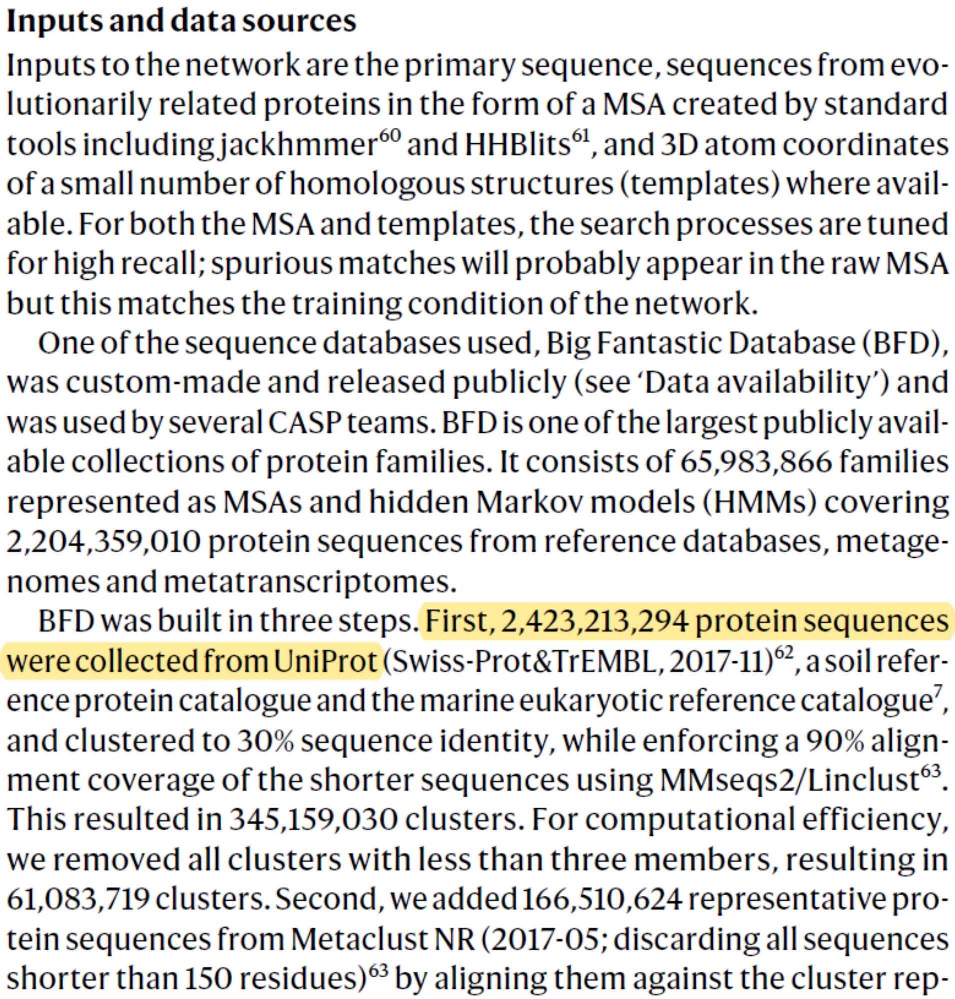
Published online: 15 July 2021

Open access

Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.



Inputs and data sources

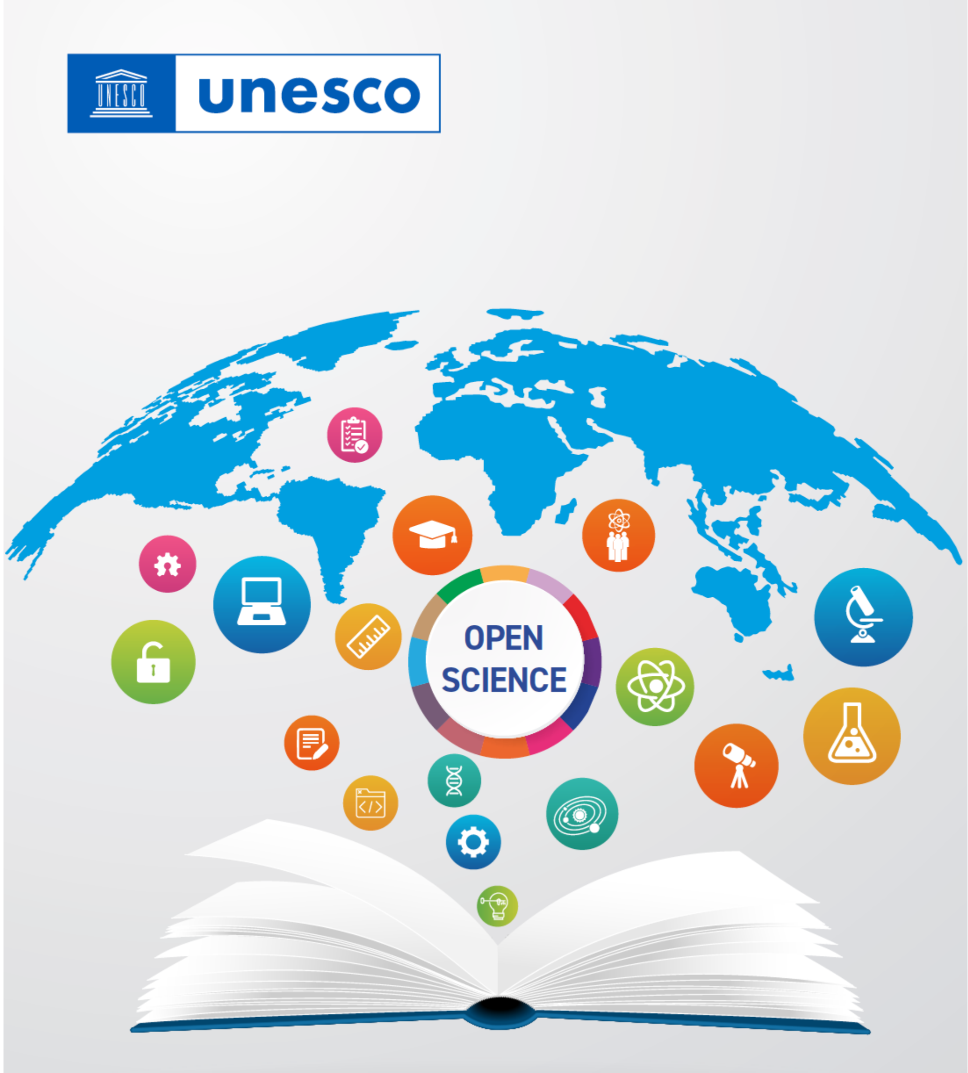
Inputs to the network are the primary sequence, sequences from evolutionarily related proteins in the form of a MSA created by standard tools including jackhmmer[60] and HHBlits[61], and 3D atom coordinates of a small number of homologous structures (templates) where available. For both the MSA and templates, 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.

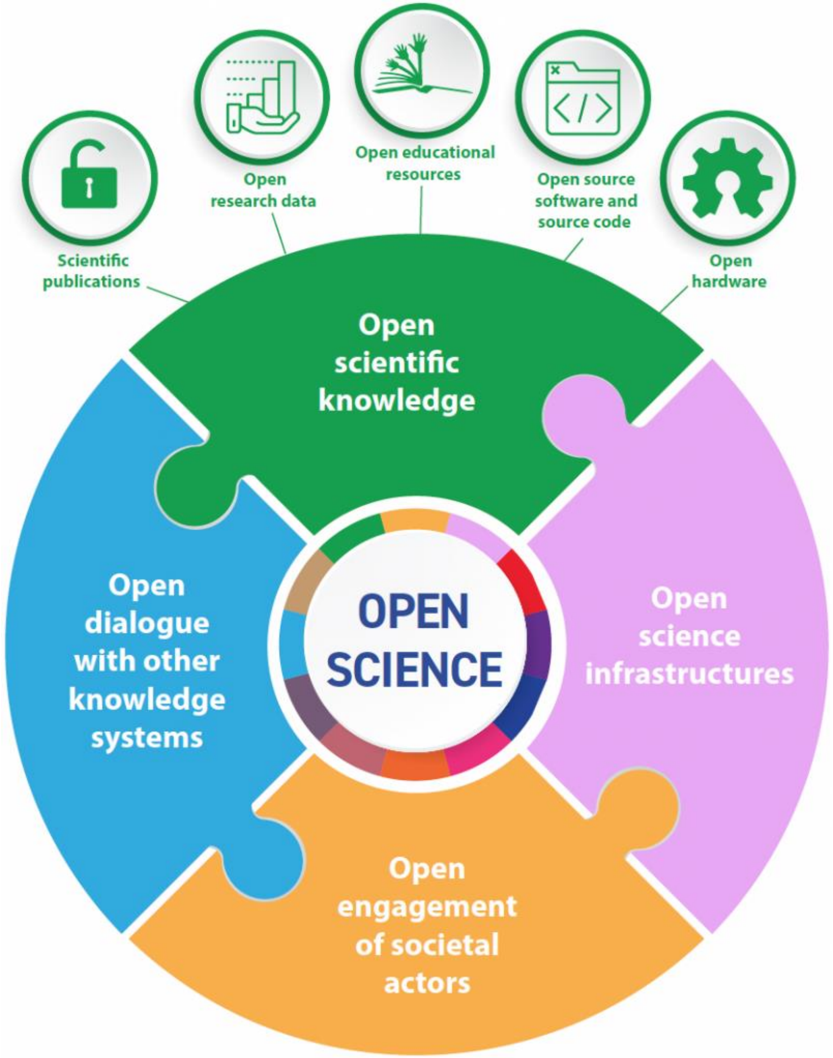
One of the sequence databases used, Big Fantastic Database (BFD), was custom-made and released publicly (see 'Data availability') and was used by several CASP teams. BFD is one of the largest publicly available collections of protein families. It consists of 65,983,866 families represented as MSAs and 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)[62], a soil reference protein catalogue and the marine eukaryotic reference catalogue[7], and clustered to 30% sequence identity, while enforcing a 90% alignment coverage of the shorter sequences using MMseqs2/Linclust[63]. 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)[63] by aligning them against the cluster rep-

Jumper, J et al. Nature (2021).

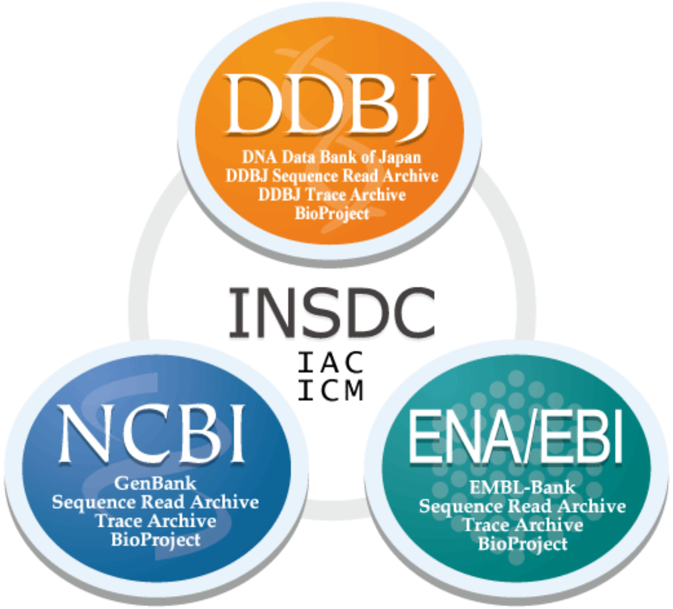
UNESCO Recommendation on Open Science




Scientific publications
Open research data
Open educational resources
Open source software and source code
Open hardware
Open scientific knowledge
Open dialogue with other knowledge systems
OPEN SCIENCE
Open science infrastructures
Open engagement of societal actors

International Nucleotide Sequence Database Collaboration (INSDC)

Japan,1986



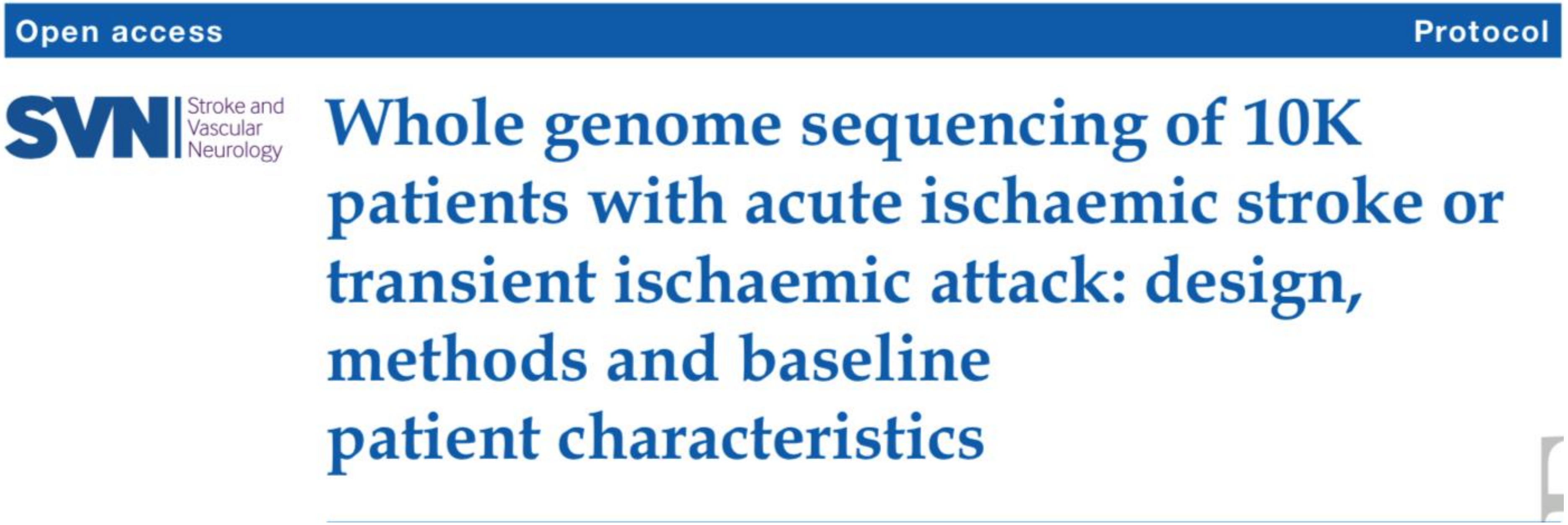
USA, 1988 Europe, 1992

- 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

Background in China (probably your country too)

- Big Data generated from Large-scale National Research Projects based on genome sequencing
- Lack of data sharing in China
  - No policy to enforce data sharing
  - Data sharing at INSDC mostly publication-driven
  - Technical issues (international network bandwidth, language barrier) make such sharing very difficult
  - No incentive to share data

Large Data Submission to NGDC



Open access Protocol

SVN Stroke and Vascular Neurology

Whole genome sequencing of 10K patients with acute ischaemic stroke or transient ischaemic attack: design, methods and baseline patient characteristics

10K patients, ~2.3 PB data

Cheng S, Xu Z, Liu Y, et al. Whole genome sequencing of 10K patients with acute ischaemic stroke or transient ischaemic attack: design, methods and baseline patient characteristics. Stroke & Vascular Neurology 2020;**0**. doi:10.1136/svn-2020-000664

BIG Data Center
Beijing Institute of Genomics (BIG), CAS

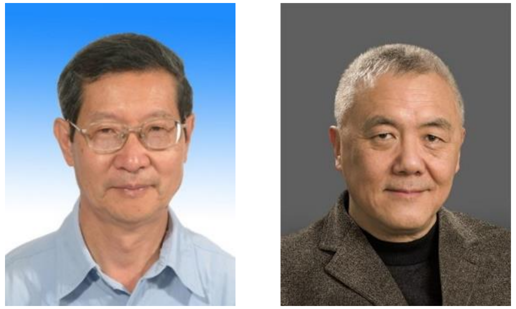
The BIG Data Center, officially founded in 2016, advances life & health sciences by providing freely open access to a variety of data resources, with the aim to translate big data into big knowledge and support worldwide research activities in both academia and industry.

Translating big data into big discoveries

Deposition **Integration** **Translation**

The Team

❑ Steering Advisors

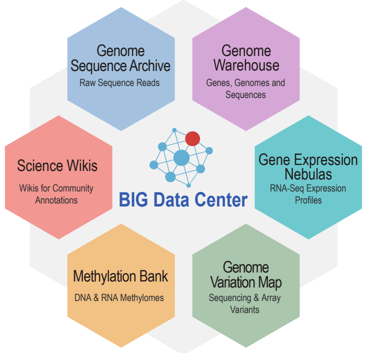


❑ Professors

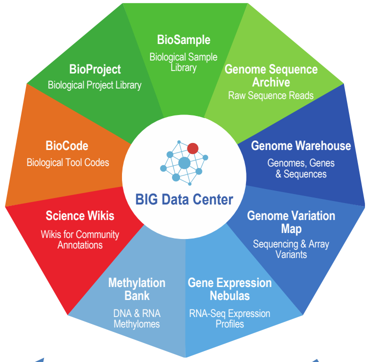
67
students

53
Staff

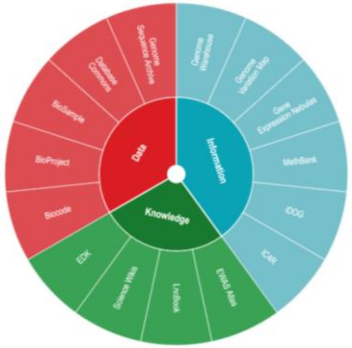
The growing of capability



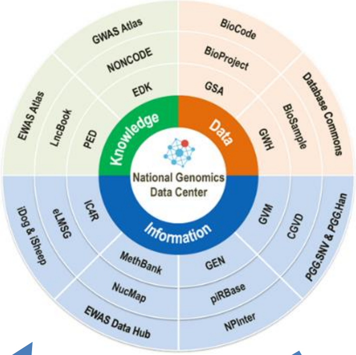
2016



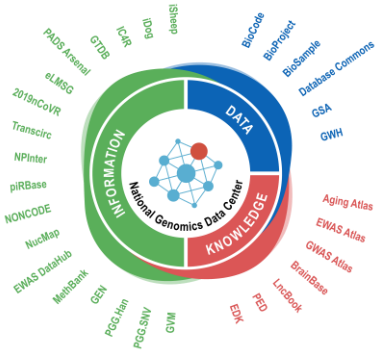
2017



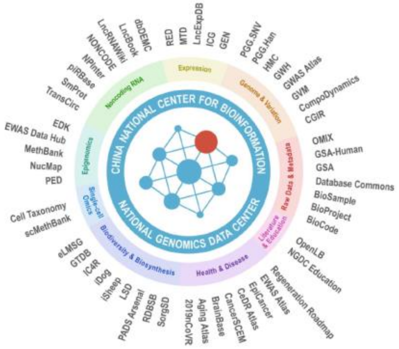
2018



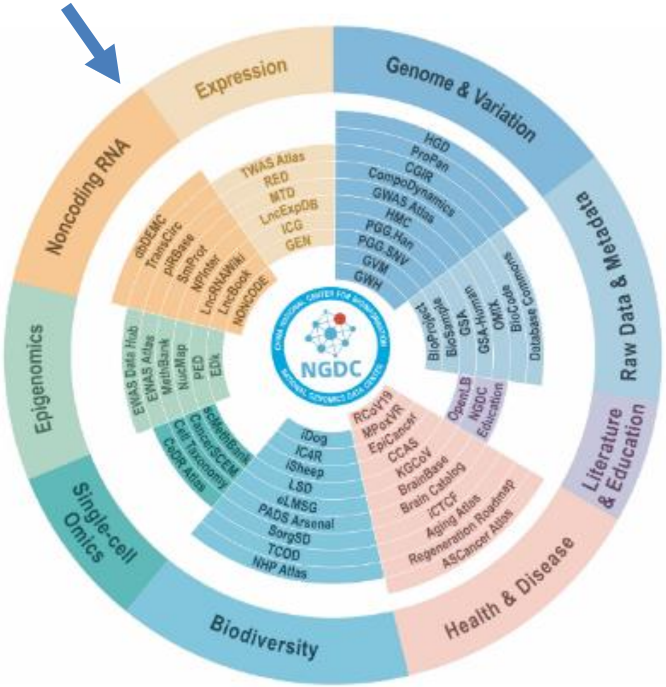
2019



2020



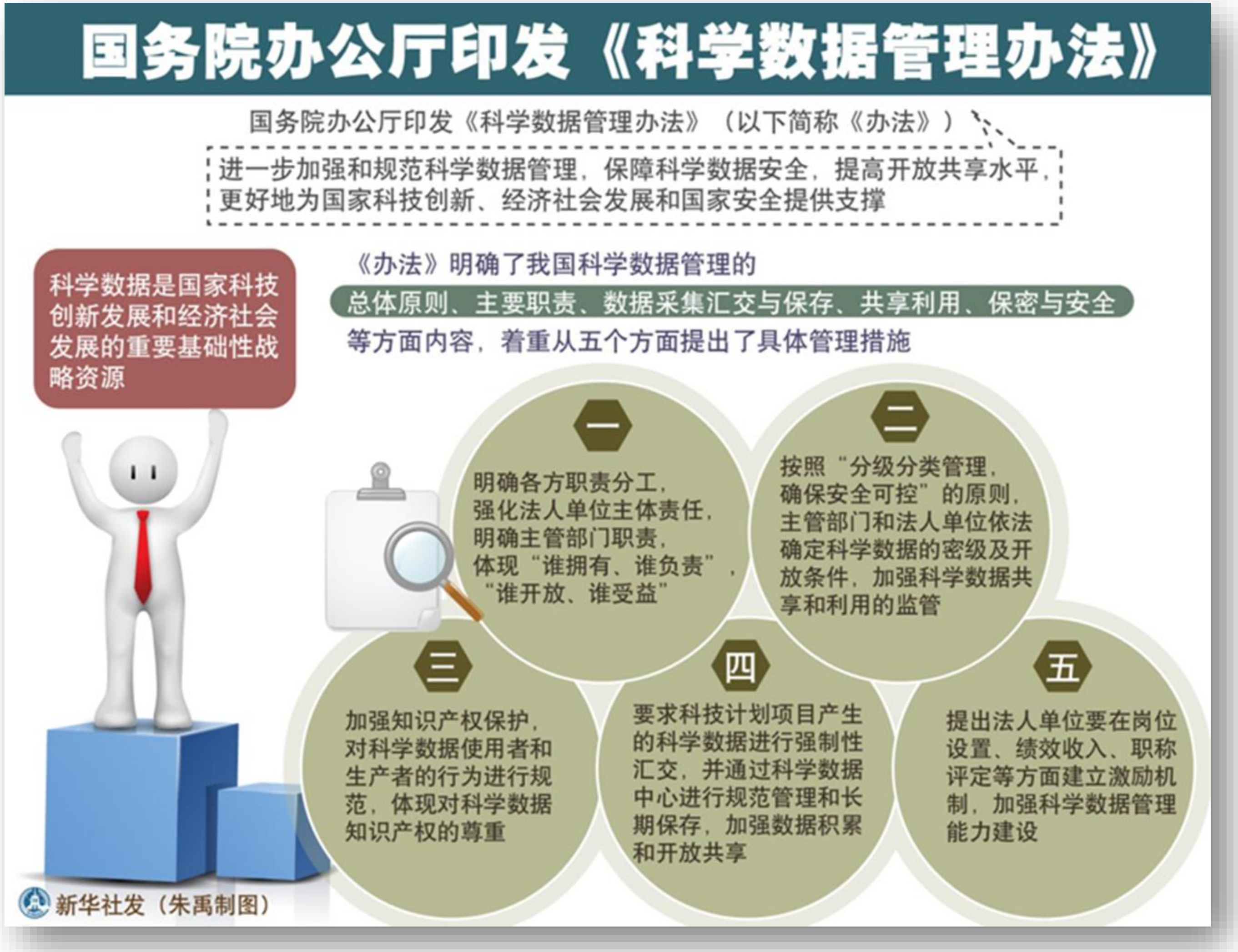
2021



2022

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

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

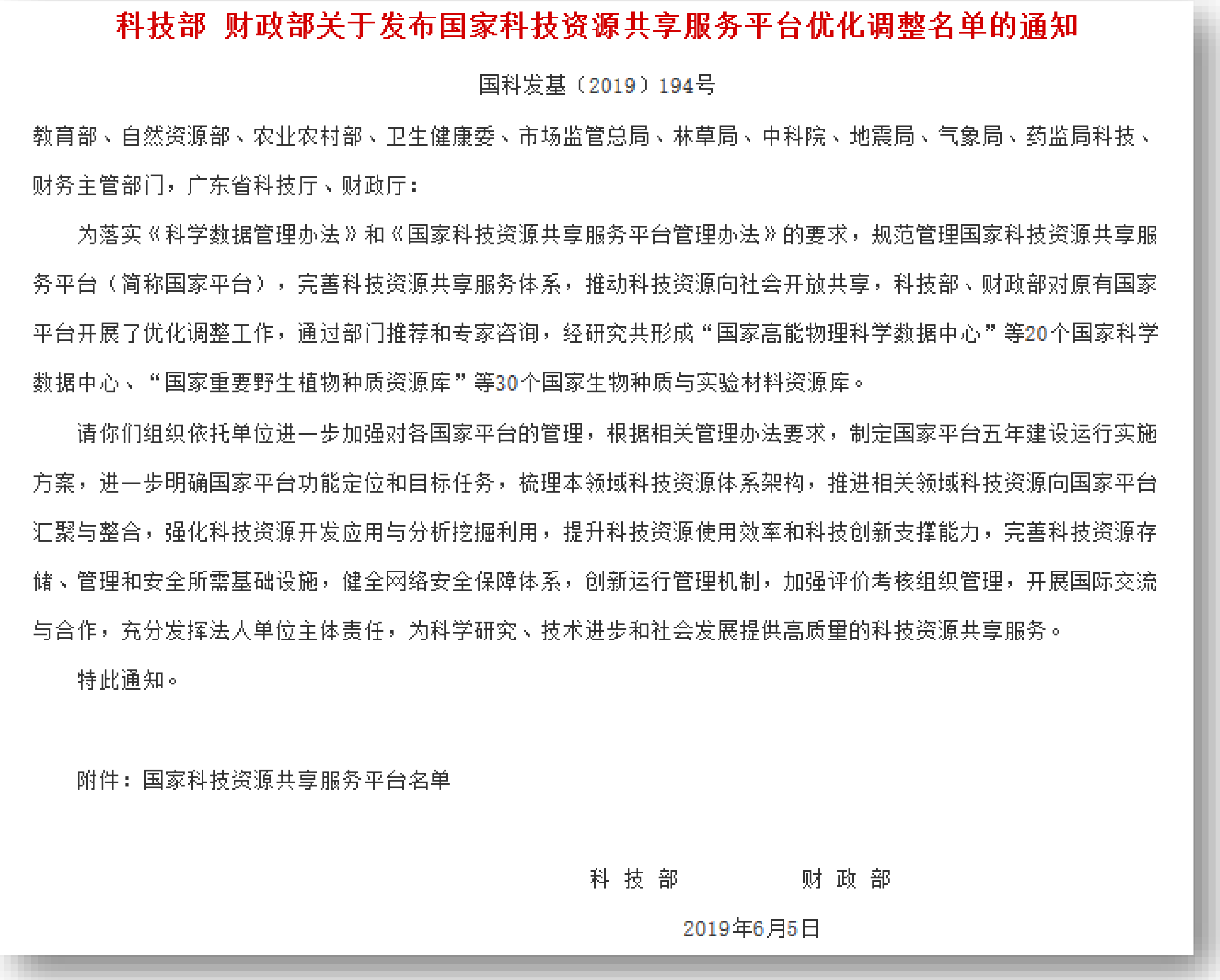
新华社发（朱禹制图）

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。

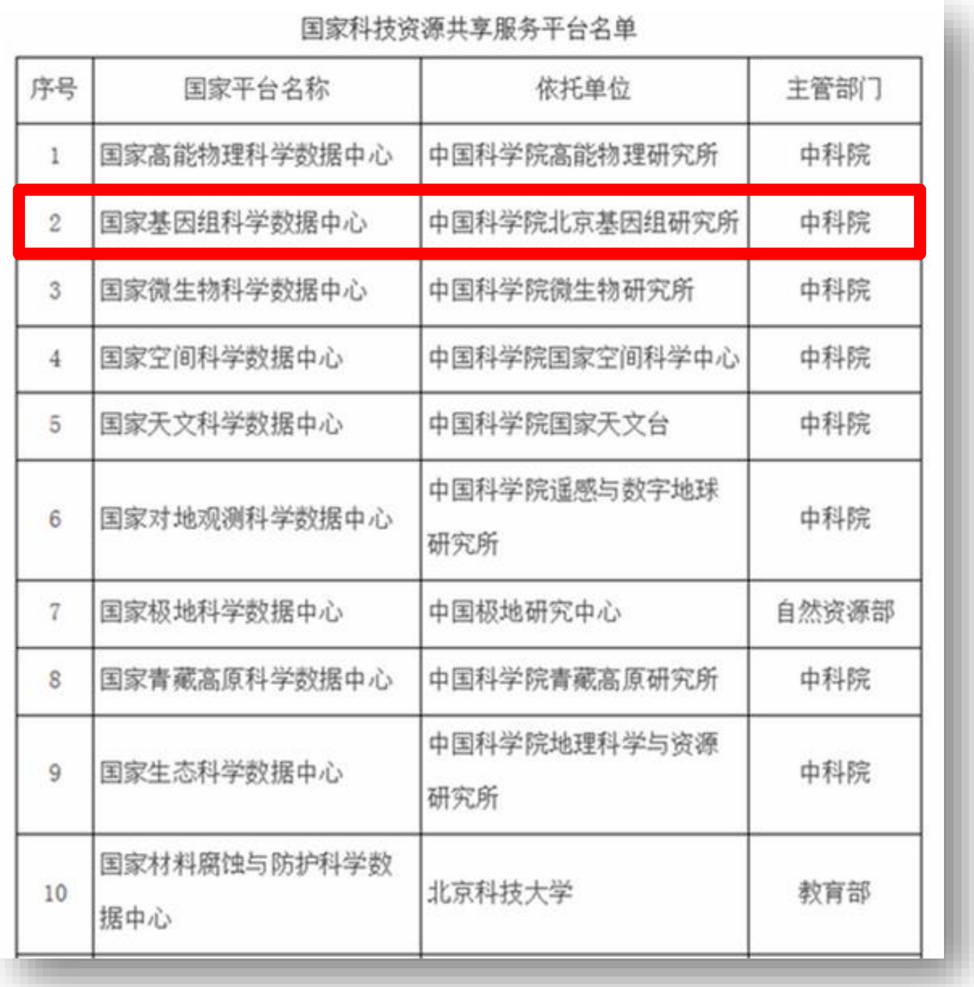
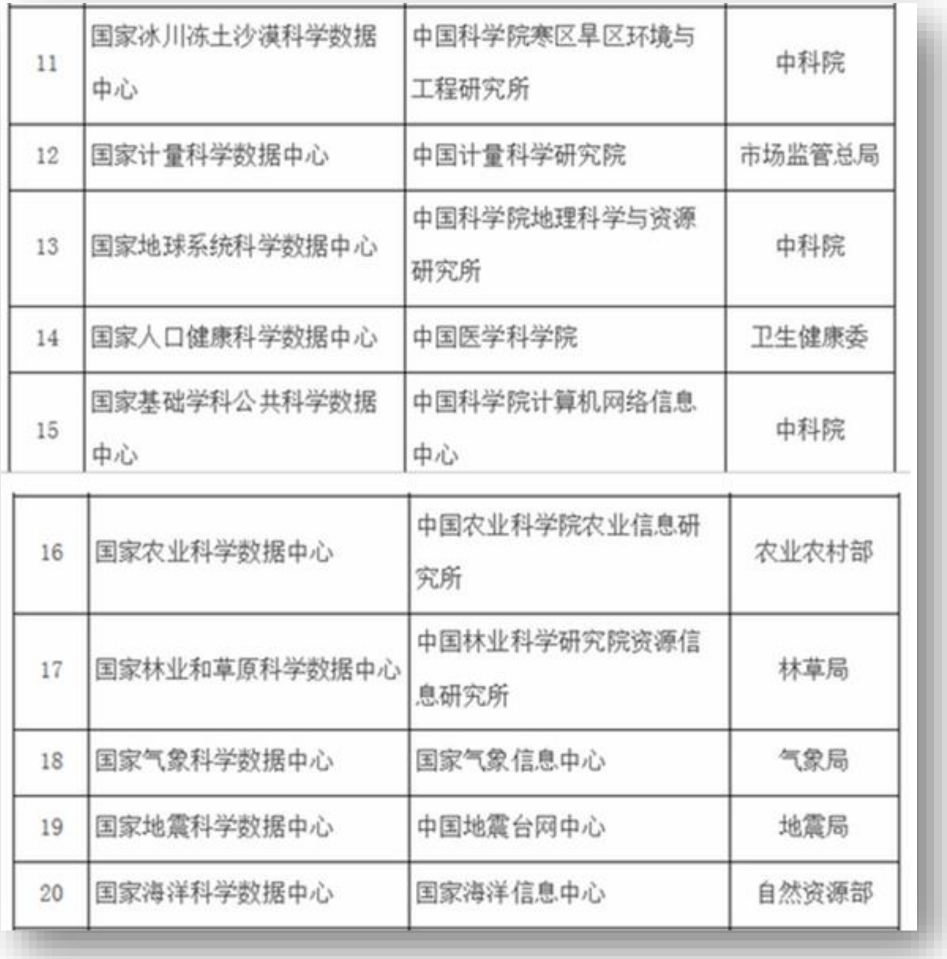
附件：国家科技资源共享服务平台名单

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

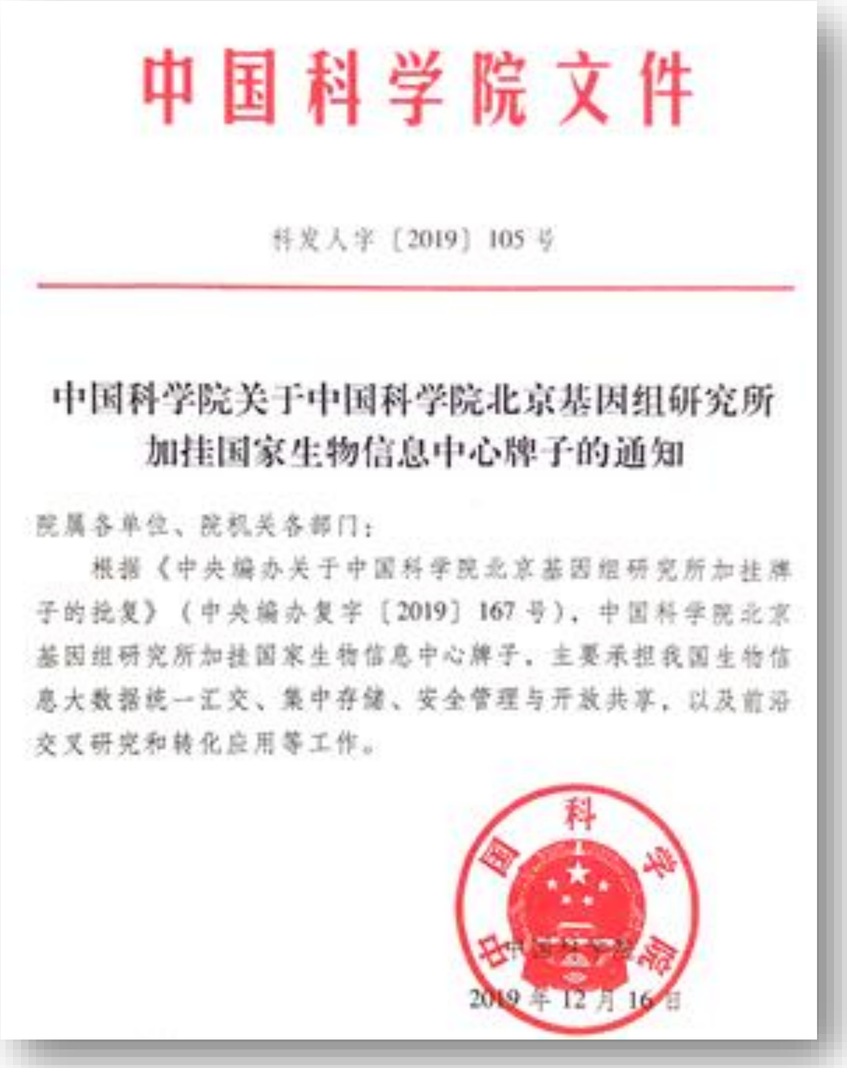
National Genomics Data Center (NGDC)

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



中国科学院文件

科发人字〔2019〕105号

中国科学院关于中国科学院北京基因组研究所加挂国家生物信息中心牌子的通知

院属各单位、院机关各部门：

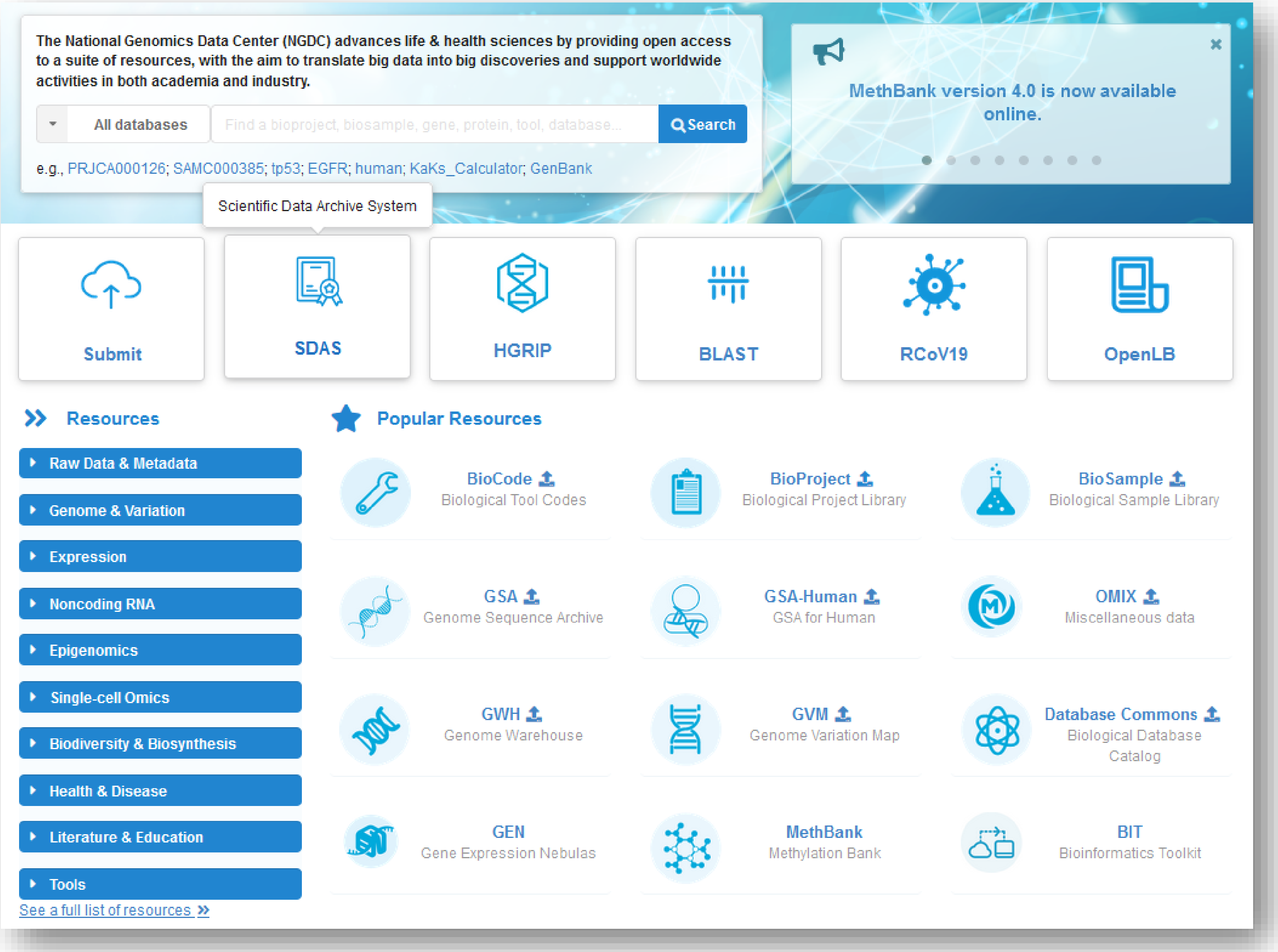
根据《中央编办关于中国科学院北京基因组研究所加挂牌子的批复》（中央编办复字〔2019〕167号），中国科学院北京基因组研究所加挂国家生物信息中心牌子，主要承担我国生物信息大数据统一汇交、集中存储、安全管理与开放共享，以及前沿交叉研究和转化应用等工作。

中国科学院

2019年12月16日

- ➢ 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

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.

All databases | Find a bioproject, biosample, gene, protein, tool, database... | Search

e.g., PRJCA000126; SAMC000385; tp53; EGFR; human; KaKs_Calculator; GenBank

Scientific Data Archive System

MethBank version 4.0 is now available online.

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 »

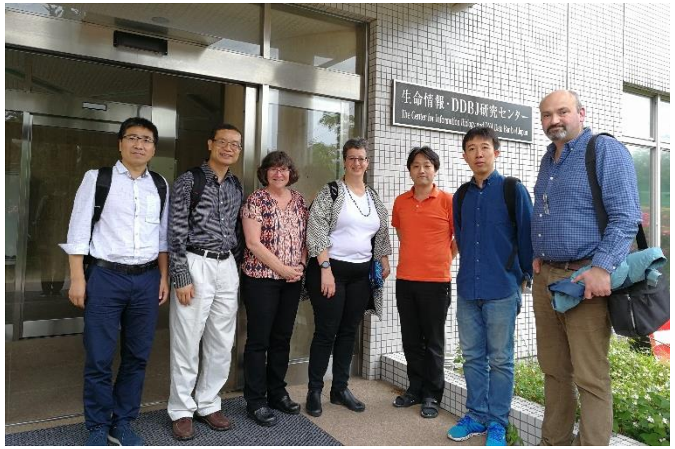
Popular Resources

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

- 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

Collaborations with INSDC

DDBJ



China National Center for Bioinformation

CNCB/NGDC Repo

Yiming Ba
Direct
National Genomics Data Cent
Beijing, Chi

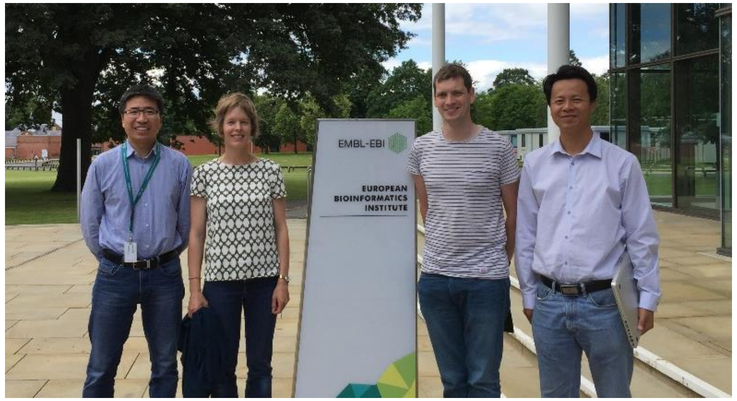
INSDC Annual Meeting
June 5, 2020 • Zoom

2017, 2020, 2023
INSDC Annual Meetings

NCBI

2017, 2018, 2021
Visit and training

EBI



2016, 2019, 2022
Visit and INSDC meeting

Following INSDC Data Structure & Standard

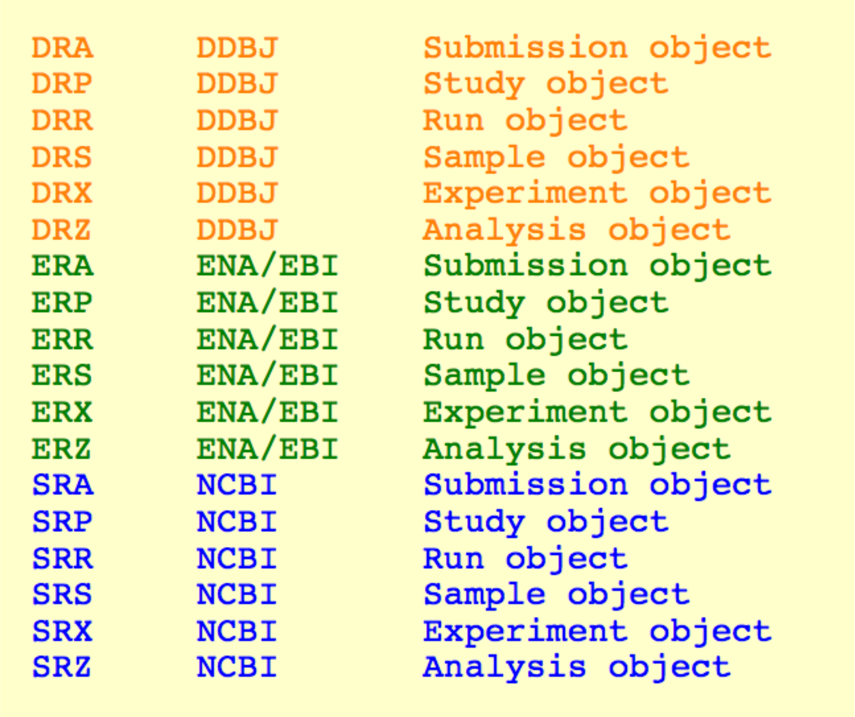
BioProject

PRJNA – PRJNZ	NCBI
PRJEA – PRJEZ	EBI
PRJDA – PRJDZ	DDBJ
PRJCA – PRJCZ	**NGDC**

BioSample

SAMN	NCBI
SAME	EBI/ENA
SAMD	DDBJ
SAMC	**NGDC**

Sequence Read Archive



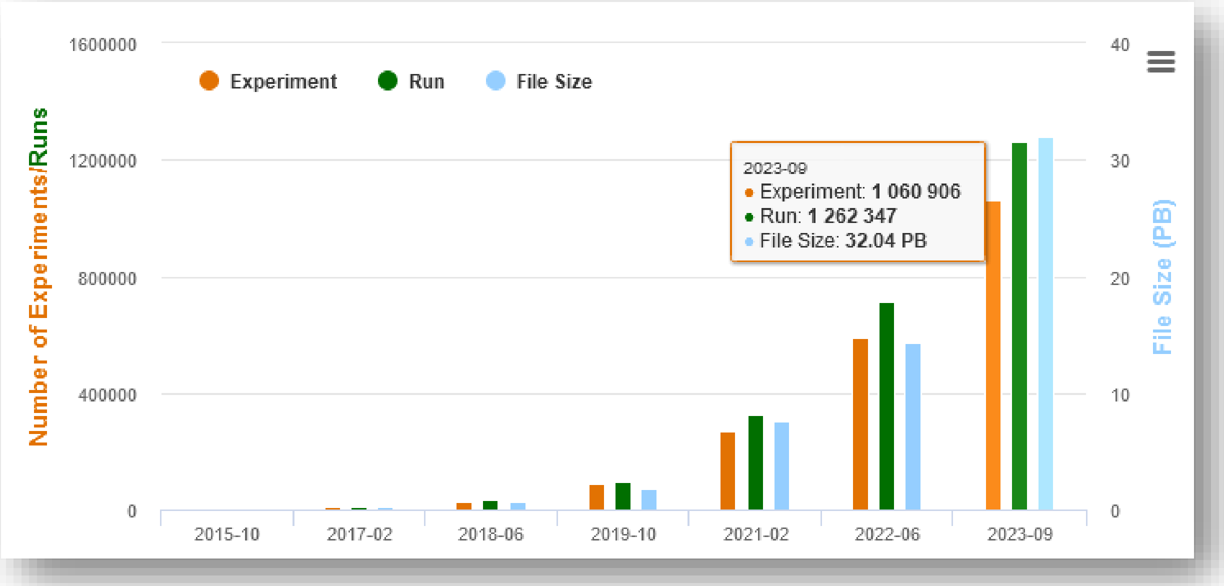
DRA	DDBJ	Submission object
DRP	DDBJ	Study object
DRR	DDBJ	Run object
DRS	DDBJ	Sample object
DRX	DDBJ	Experiment object
DRZ	DDBJ	Analysis object
ERA	ENA/EBI	Submission object
ERP	ENA/EBI	Study object
ERR	ENA/EBI	Run object
ERS	ENA/EBI	Sample object
ERX	ENA/EBI	Experiment object
ERZ	ENA/EBI	Analysis object
SRA	NCBI	Submission object
SRP	NCBI	Study object
SRR	NCBI	Run object
SRS	NCBI	Sample object
SRX	NCBI	Experiment object
SRZ	NCBI	Analysis object

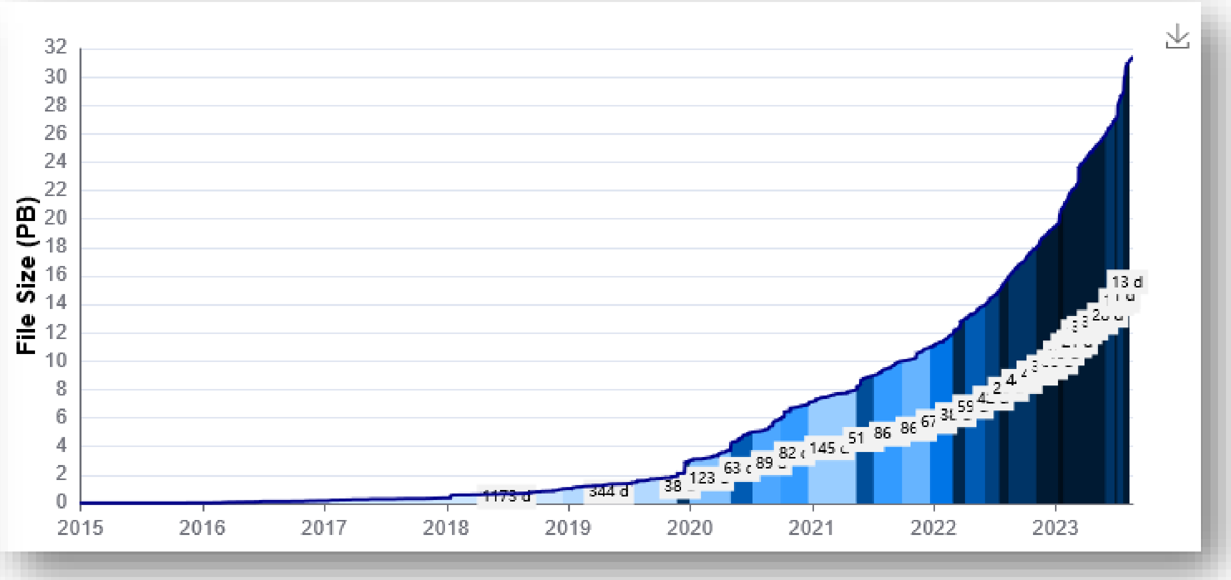
Image From DDBJ

GSA

CRA	Submission object
CRP	Study object
CRR	Run object
CRS	Sample object
CRX	Experiment object
CRZ	Analysis object

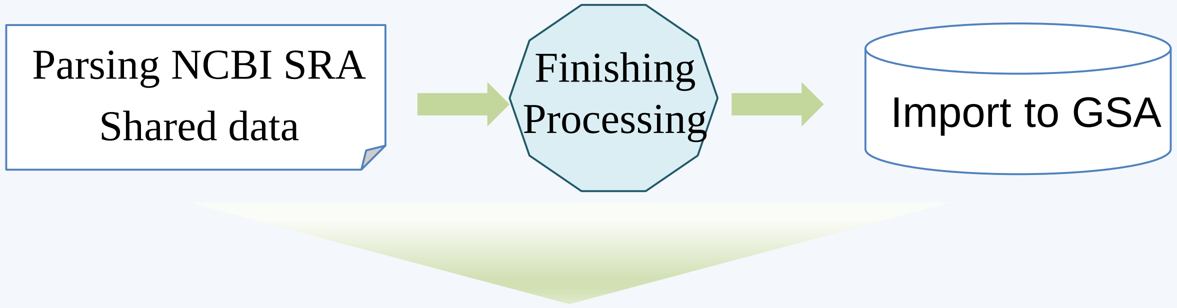
Rapid Data Growth



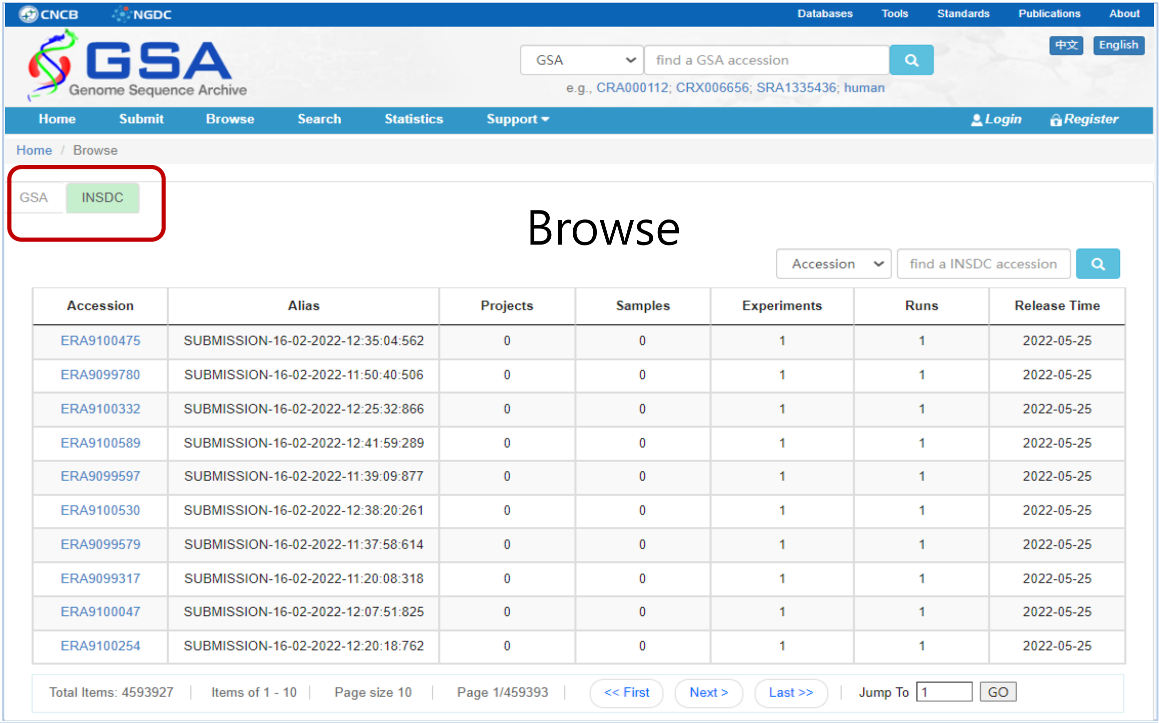


>33 PB as of 2023-09-27

Integration of International Data - GSA



623,345	29,562,590	21,994,135	23,523,117
PROJECTS	SAMPLES	EXPERIMENTS	RUNS



Browse

Accession	Alias	Projects	Samples	Experiments	Runs	Release Time
ERA9100475	SUBMISSION-16-02-2022-12:35:04:562	0	0	1	1	2022-05-25
ERA9099780	SUBMISSION-16-02-2022-11:50:40:506	0	0	1	1	2022-05-25
ERA9100332	SUBMISSION-16-02-2022-12:25:32:866	0	0	1	1	2022-05-25
ERA9100589	SUBMISSION-16-02-2022-12:41:59:289	0	0	1	1	2022-05-25
ERA9099597	SUBMISSION-16-02-2022-11:39:09:877	0	0	1	1	2022-05-25
ERA9100530	SUBMISSION-16-02-2022-12:38:20:261	0	0	1	1	2022-05-25
ERA9099579	SUBMISSION-16-02-2022-11:37:58:614	0	0	1	1	2022-05-25
ERA9099317	SUBMISSION-16-02-2022-11:20:08:318	0	0	1	1	2022-05-25
ERA9100047	SUBMISSION-16-02-2022-12:07:51:825	0	0	1	1	2022-05-25
ERA9100254	SUBMISSION-16-02-2022-12:20:18:762	0	0	1	1	2022-05-25

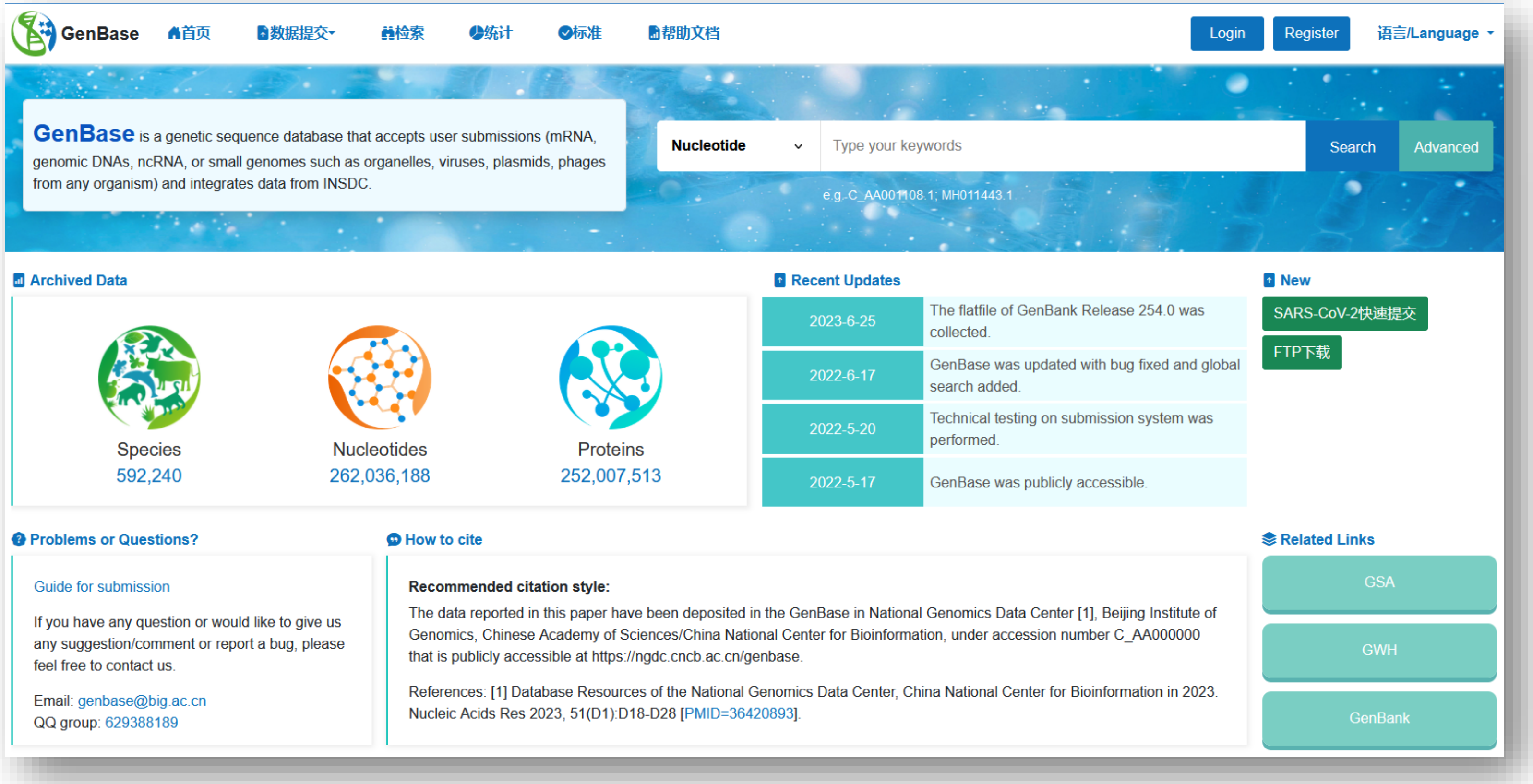
Total Items: 4593927 | Items of 1 - 10 | Page size 10 | Page 1/459393 | << First Next > Last >> | Jump To 1 GO

Metadata information has been updated regularly

The data files have been downloaded every day since **2022-04-20**

Data Files： **~5 PB**

GenBase in sync with GenBank



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.

e.g. C_AA001108.1; MH011443.1

Archived Data

Species	Nucleotides	Proteins
592,240	262,036,188	252,007,513

Recent Updates

2023-6-25	The flatfile of GenBank Release 254.0 was collected.
2022-6-17	GenBase was updated with bug fixed and global search added.
2022-5-20	Technical testing on submission system was performed.
2022-5-17	GenBase was publicly accessible.

New

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
QQ group: 629388189

How to cite

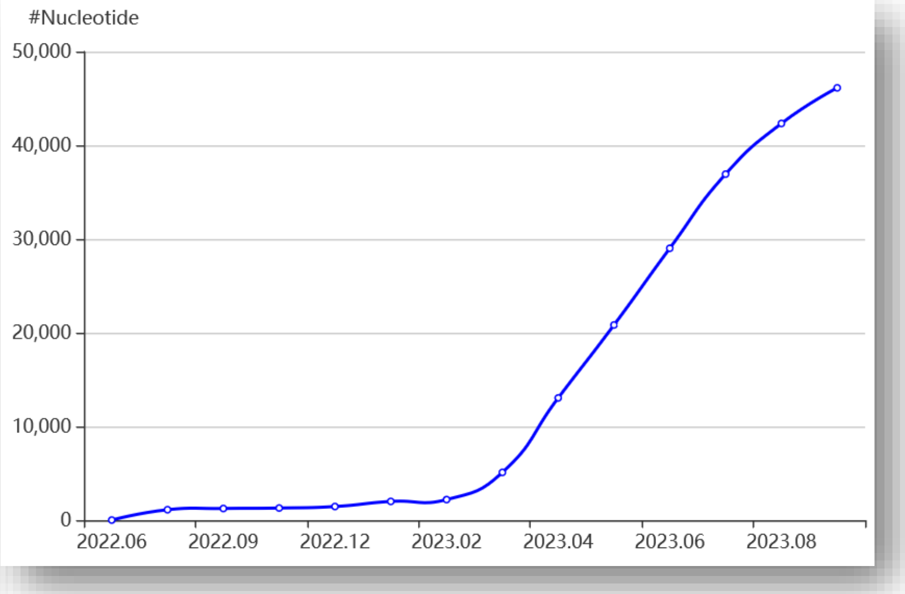
Recommended citation style:

The data reported in this paper have been deposited in the GenBase in National Genomics Data Center [1], 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] Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. Nucleic Acids Res 2023, 51(D1):D18-D28 [PMID=36420893].

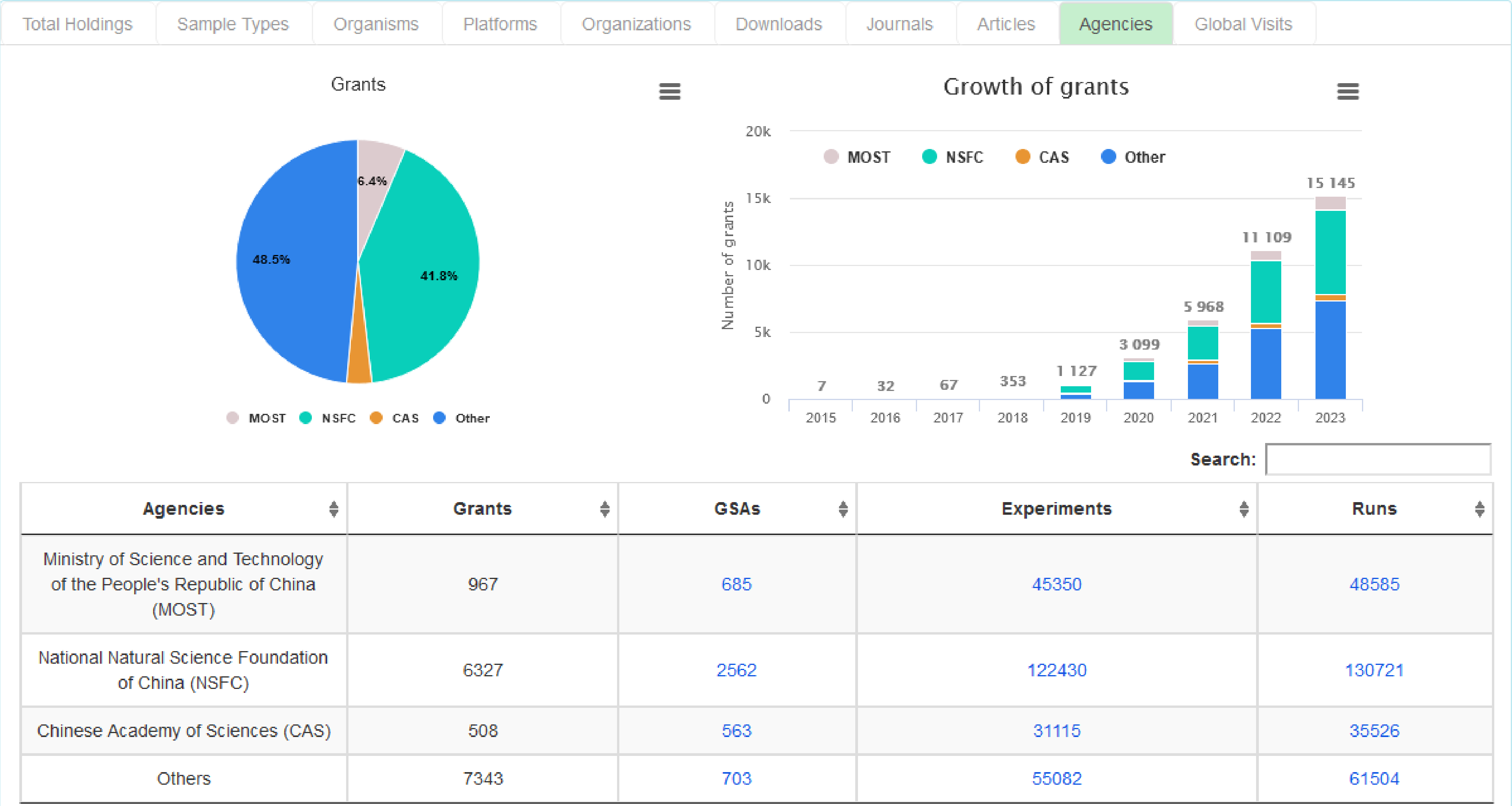
Related Links

Direct submissions



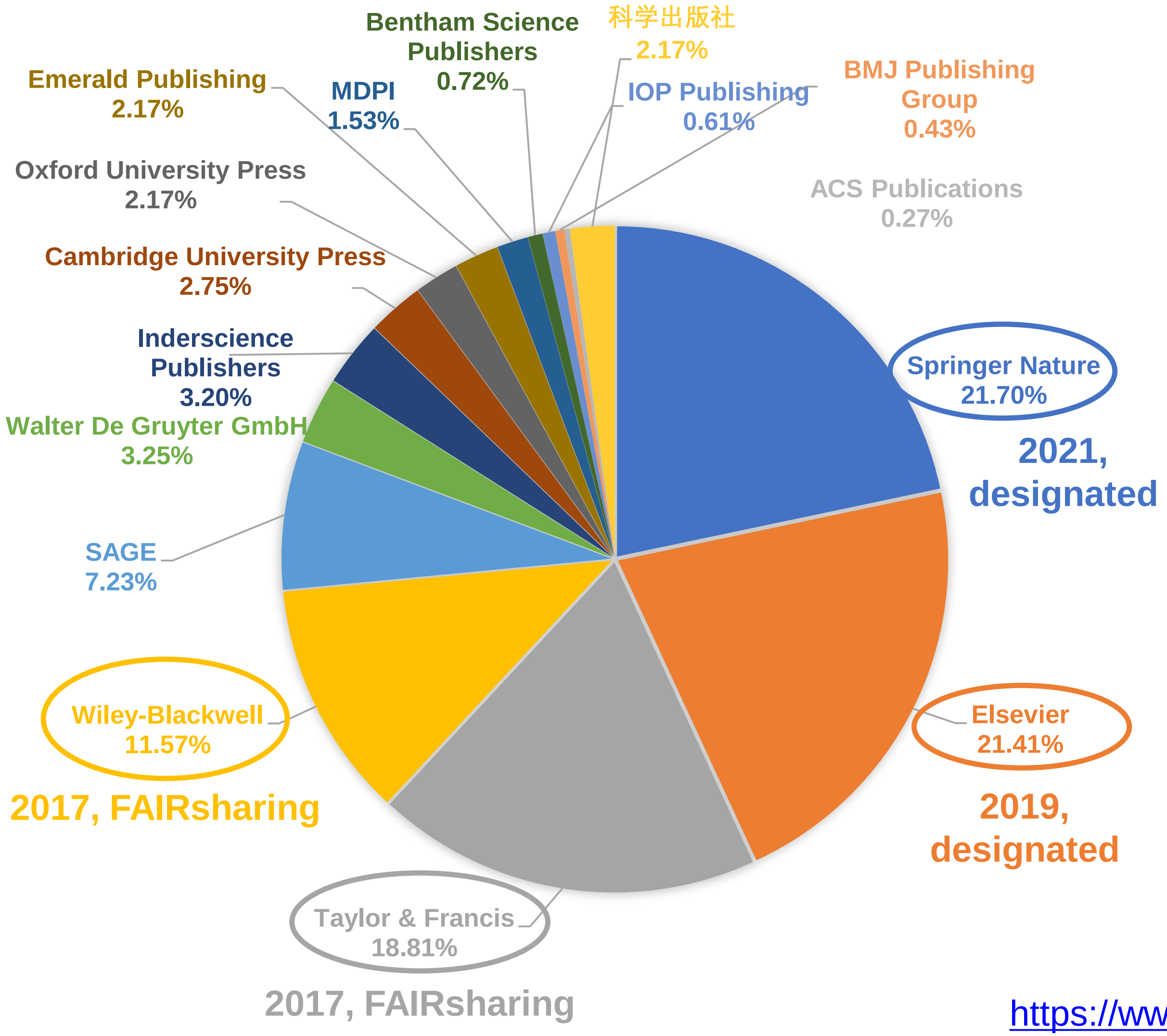
- **GenBank Release 254.0** has been integrated, with daily updates
- **In total**: **592,276** Species, ~**267 mil.** Nucleotides, **274 mil.** Proteins
- **Direct submissions**: **46k** Nucleotides, **466k** Proteins

Supporting >15k Research Grants



Agencies	Grants	GSAs	Experiments	Runs
Ministry of Science and Technology of the People's Republic of China (MOST)	967	685	45350	48585
National Natural Science Foundation of China (NSFC)	6327	2562	122430	130721
Chinese Academy of Sciences (CAS)	508	563	31115	35526
Others	7343	703	55082	61504

GSA Endorsed by Springer Nature and Major Publishers

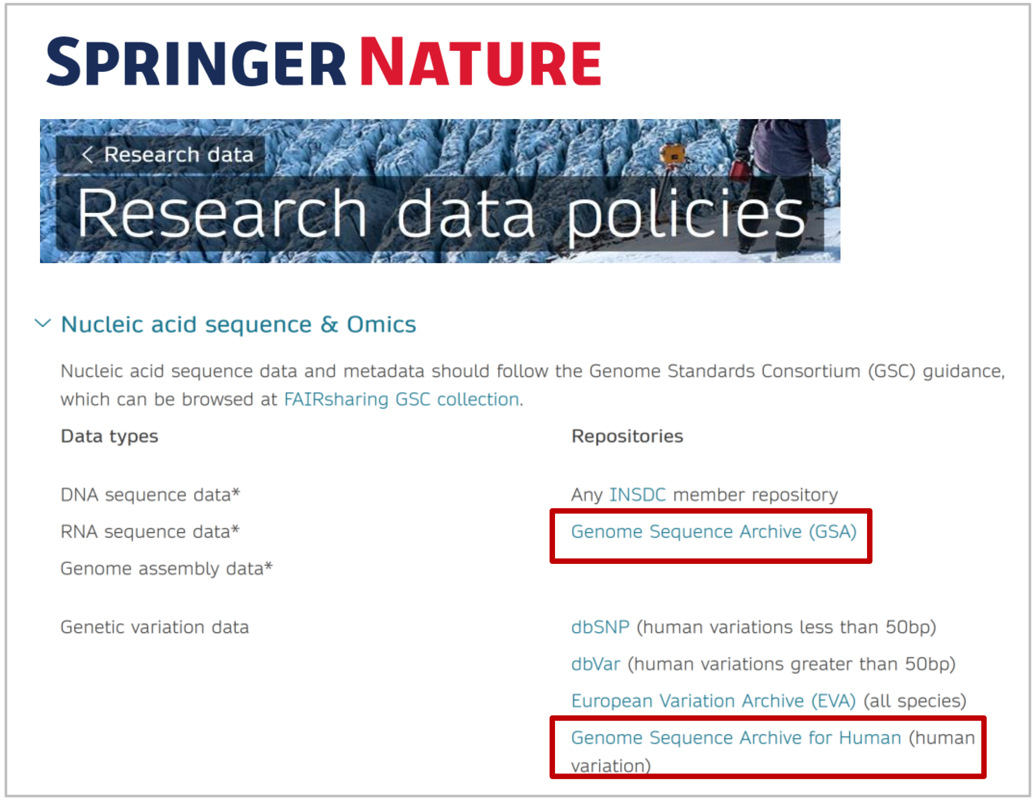


2021, designated

2019, designated

2017, FAIRsharing

2017, FAIRsharing



SPRINGER NATURE

< Research data

Research data policies

Nucleic acid sequence & Omics

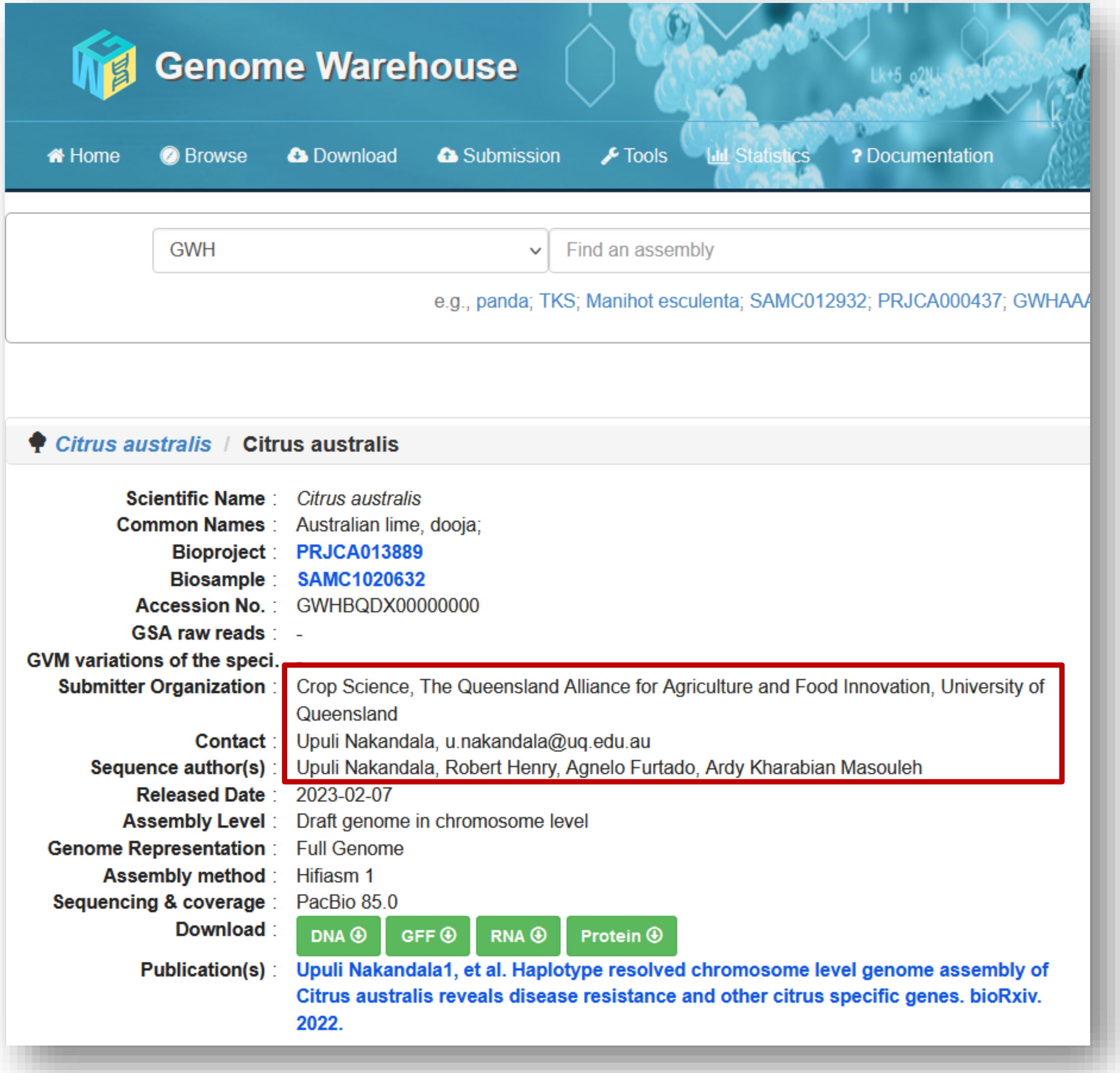
Nucleic acid sequence data and metadata should follow the Genome Standards Consortium (GSC) guidance, which can be browsed at FAIRsharing GSC collection.

Data types	Repositories
DNA sequence data*	Any INSDC member repository
RNA sequence data*	Genome Sequence Archive (GSA)
Genome assembly data*	
Genetic variation data	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/

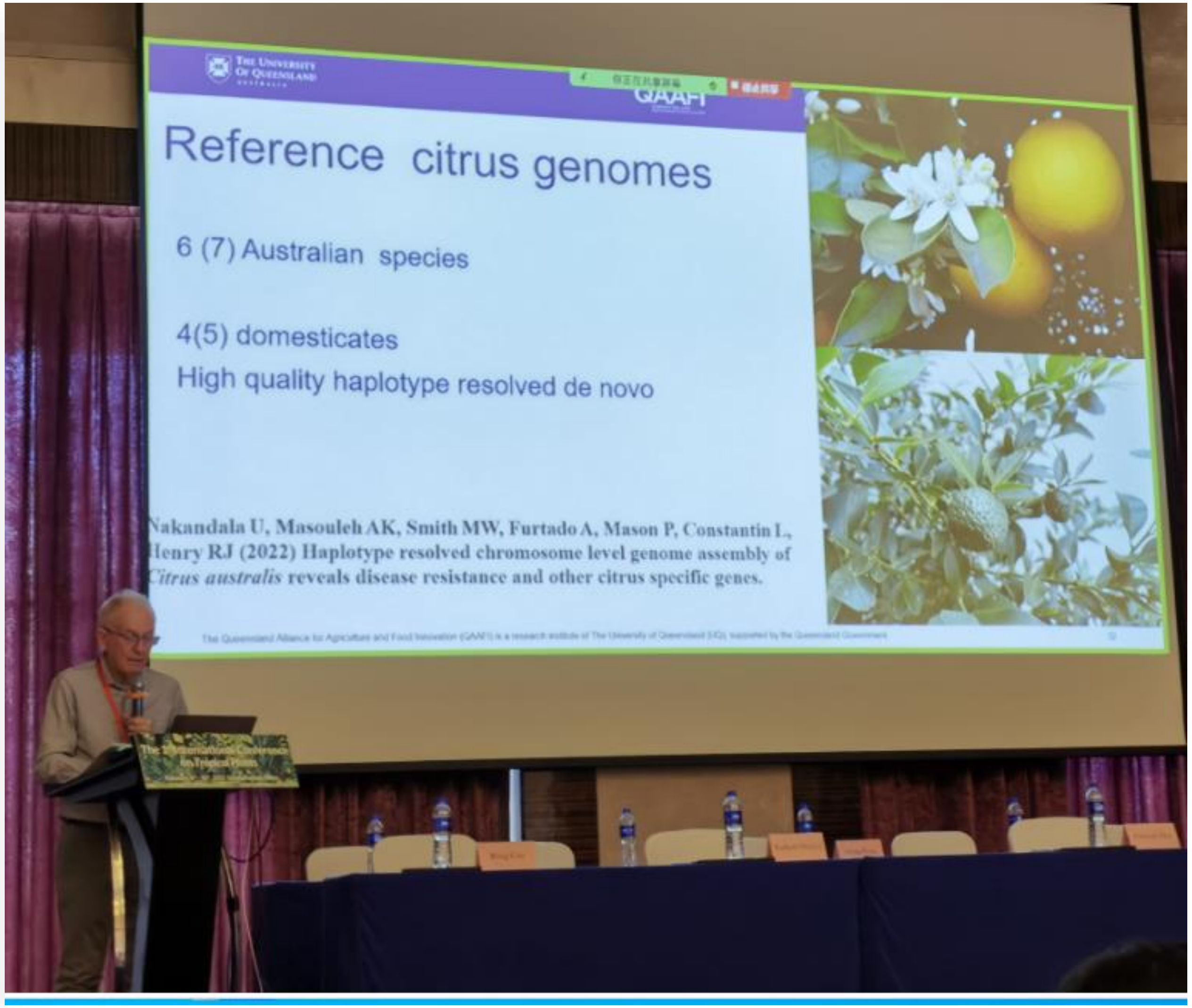
International Submitters from 22 countries



e.g., panda; TKS; Manihot esculenta; SAMC012932; PRJCA000437; GWHAA

Citrus australis / Citrus australis

Scientific Name :	*Citrus australis*
Common Names :	Australian lime, dooja;
Bioproject :	PRJCA013889
Biosample :	SAMC1020632
Accession No. :	GWHBQDX00000000
GSA raw reads :	-
GVM variations of the speci. :	-
Submitter Organization :	Crop Science, The Queensland Alliance for Agriculture and Food Innovation, University of Queensland
Contact :	Upuli Nakandala, u.nakandala@uq.edu.au
Sequence author(s) :	Upuli Nakandala, Robert Henry, Agnelo Furtado, Ardy Kharabian Masouleh
Released Date :	2023-02-07
Assembly Level :	Draft genome in chromosome level
Genome Representation :	Full Genome
Assembly method :	Hifiasm 1
Sequencing & coverage :	PacBio 85.0
Download :	DNA GFF RNA Protein
Publication(s) :	Upuli Nakandala1, et al. Haplotype resolved chromosome level genome assembly of Citrus australis reveals disease resistance and other citrus specific genes. bioRxiv. 2022.



Reference citrus genomes

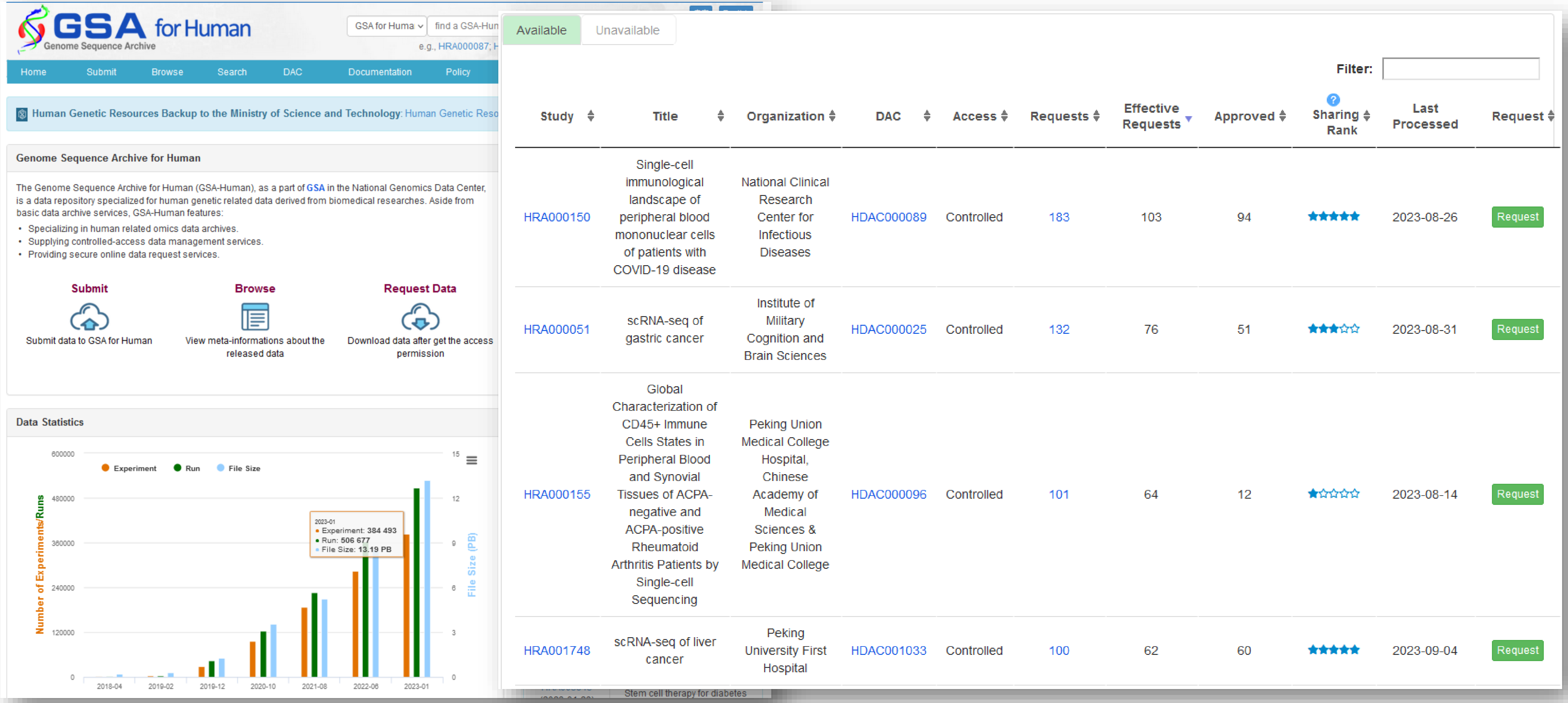
6 (7) Australian species

4(5) domesticates

High quality haplotype resolved de novo

Nakandala U, Masouleh AK, Smith MW, Furtado A, Mason P, Constantin L, Henry RJ (2022) Haplotype resolved chromosome level genome assembly of *Citrus australis* reveals disease resistance and other citrus specific genes.

GSA for Human Database – Controlled Access



Human Genetic Resources Backup to the Ministry of Science and Technology: Human Genetic Reso

Genome Sequence Archive for Human

The Genome Sequence Archive for Human (GSA-Human), as a part of GSA in the National Genomics Data Center, is a data repository specialized for human genetic related data derived from biomedical researches. Aside from basic data archive services, GSA-Human features:

- Specializing in human related omics data archives.
- Supplying controlled-access data management services.
- Providing secure online data request services.

Submit
Submit data to GSA for Human

Browse
View meta-informations about the released data

Request Data
Download data after get the access permission

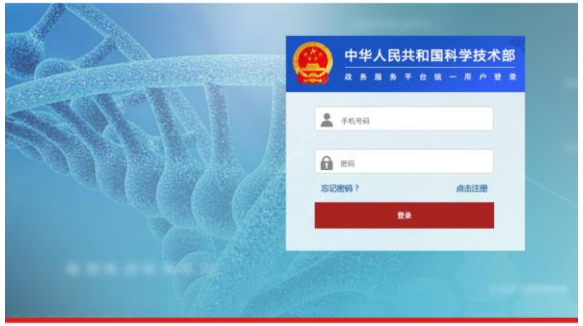
Data Statistics

Study	Title	Organization	DAC	Access	Requests	Effective Requests	Approved	Sharing Rank	Last Processed	Request
HRA000150	Single-cell immunological landscape of peripheral blood mononuclear cells of patients with COVID-19 disease	National Clinical Research Center for Infectious Diseases	HDAC000089	Controlled	183	103	94	★★★★★	2023-08-26	Request
HRA000051	scRNA-seq of gastric cancer	Institute of Military Cognition and Brain Sciences	HDAC000025	Controlled	132	76	51	★★★☆☆	2023-08-31	Request
HRA000155	Global Characterization of CD45+ Immune Cells States in Peripheral Blood and Synovial Tissues of ACPA-negative and ACPA-positive Rheumatoid Arthritis Patients by Single-cell Sequencing	Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College	HDAC000096	Controlled	101	64	12	★☆☆☆☆	2023-08-14	Request
HRA001748	scRNA-seq of liver cancer	Peking University First Hospital	HDAC001033	Controlled	100	62	60	★★★★★	2023-09-04	Request

Human Data Backup & Registration Protocol

Backup Center

人类遗传资源信息备份平台



中华人民共和国科学技术部

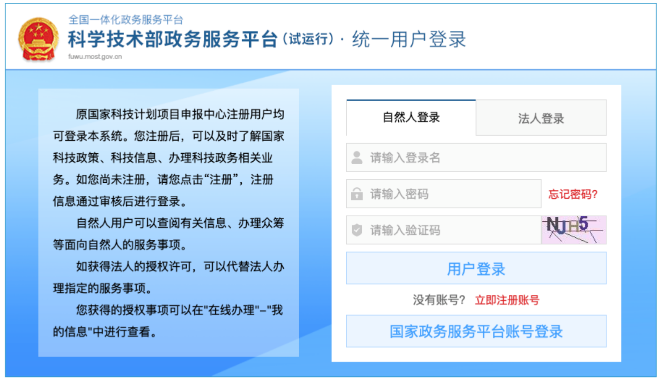
政务服务平台统一用户登录

Backup

Backup ID

Researchers

Registration ID



全国一体化政务服务平台

科学技术部政务服务平台（试运行）· 统一用户登录

fuwu.most.gov.cn

原国家科技计划项目申报中心注册用户均可登录本系统。您注册后，可以及时了解国家科技政策、科技信息、办理科技政务相关业务。如您尚未注册，请您点击“注册”，注册信息通过审核后进行登录。

自然人用户可以查阅有关信息、办理众筹等面向自然人的服务事项。

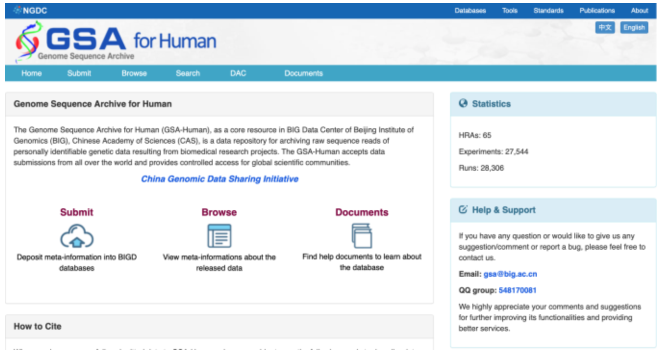
如获得法人的授权许可，可以代替法人办理指定的服务事项。

您获得的授权事项可以在"在线办理"–"我的信息"中进行查看。

Registration Center

Submission

Accession #s



Genome Sequence Archive for Human

The Genome Sequence Archive for Human (GSA-Human), as a core resource in BIG Data Center of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), is a data repository for archiving raw sequence reads of personally identifiable genetic data resulting from biomedical research projects. The GSA-Human accepts data submissions from all over the world and provides controlled access for global scientific communities.

China Genomic Data Sharing Initiative

Submit
Deposit meta-information into BIGD databases

Browse
View meta-informations about the released data

Documents
Find help documents to learn about the database

How to Cite

Statistics

HRAs: 65

Experiments: 27,544

Runs: 28,306

Help & Support

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: gsa@big.ac.cn

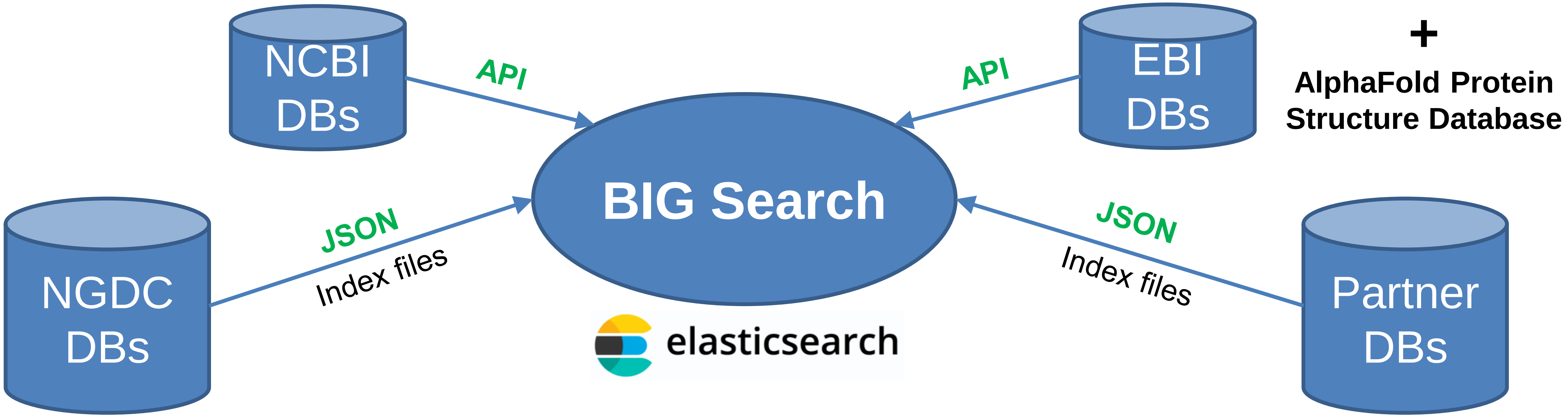
QQ group: 548170081

We highly appreciate your comments and suggestions for further improving its functionalities and providing better services.

CNCB-NGDC GSA-Human

Backup ID

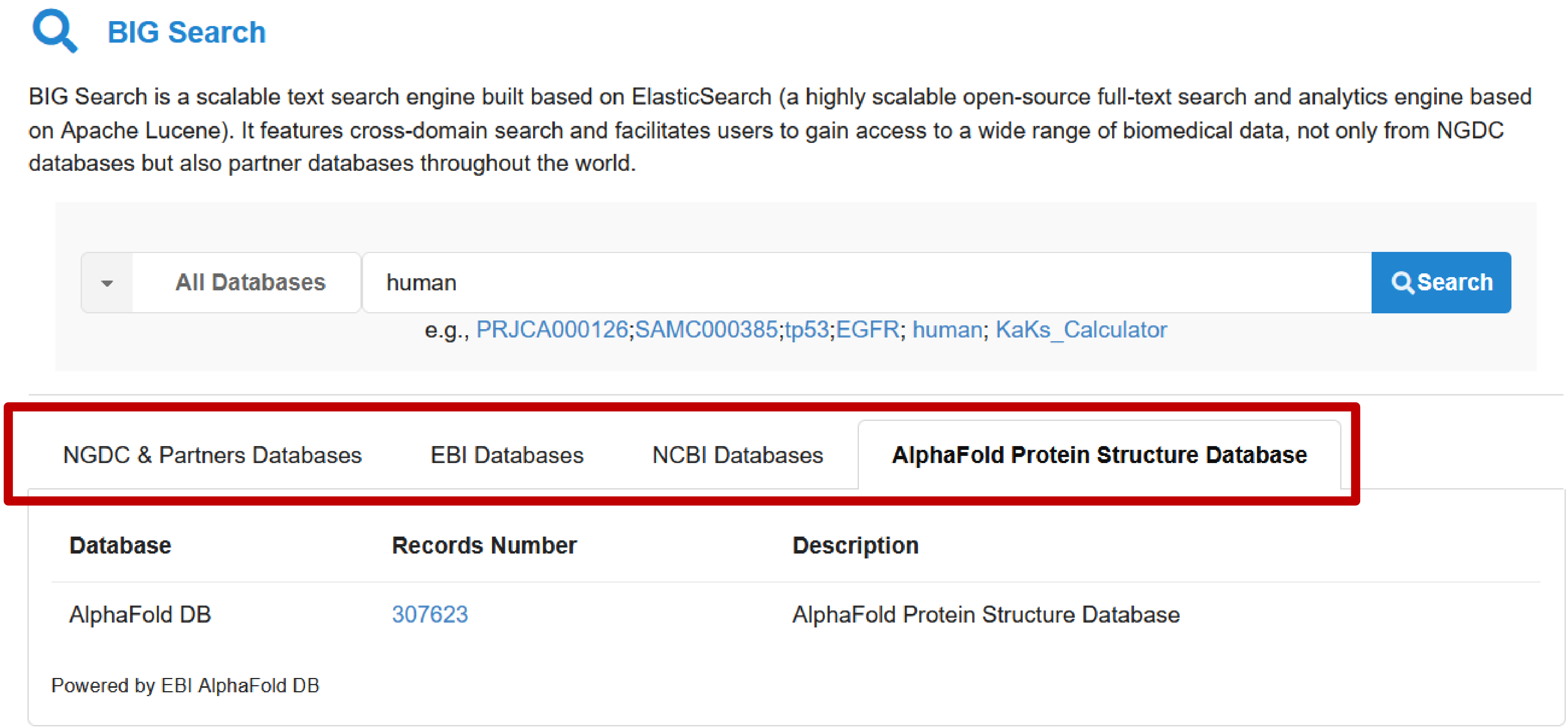
Cross-database search engine： BIG Search



Partner databases

https://ngdc.cncb.ac.cn/search/

“Google” for biology data



BIG Search

BIG Search is a scalable text search engine built based on ElasticSearch (a highly scalable open-source full-text search and analytics engine based on Apache Lucene). It features cross-domain search and facilitates users to gain access to a wide range of biomedical data, not only from NGDC databases but also partner databases throughout the world.

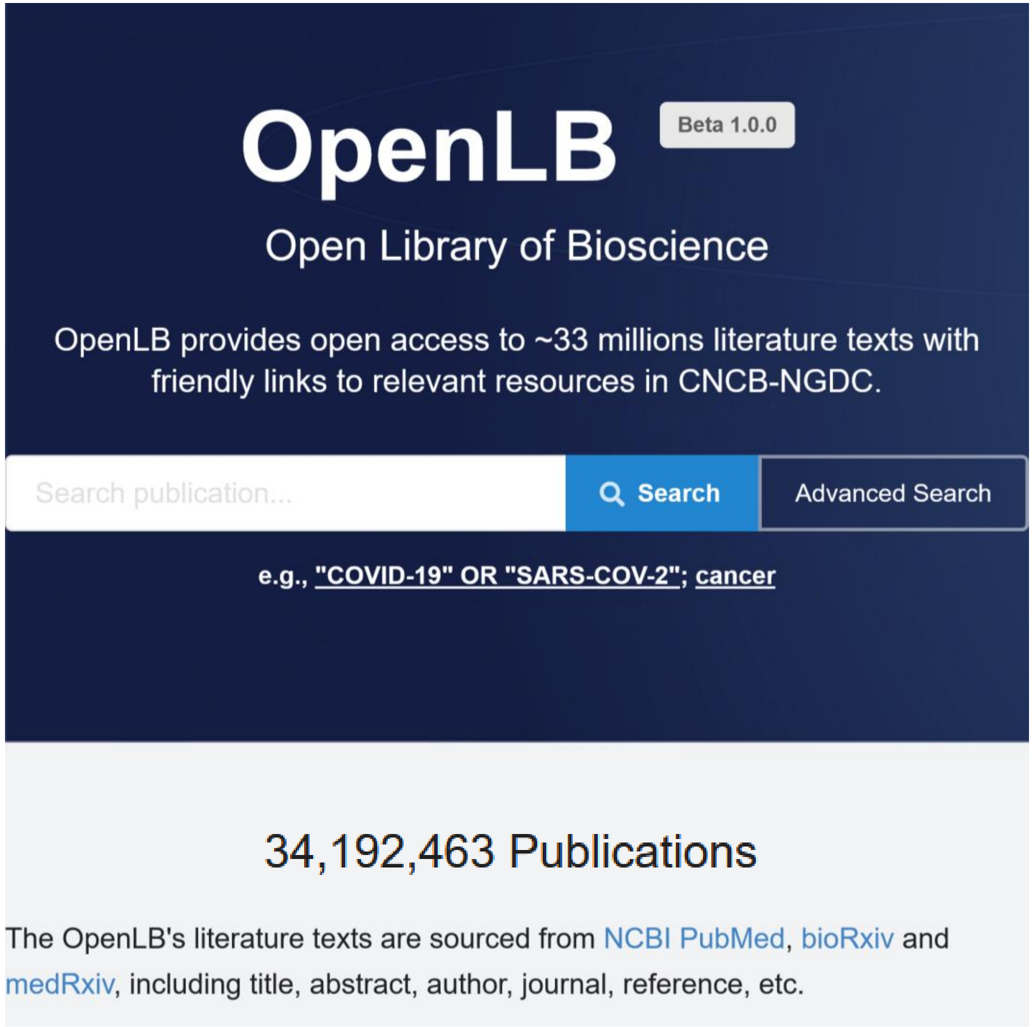
e.g., PRJCA000126;SAMC000385;tp53;EGFR; human; KaKs_Calculator

NGDC & Partners Databases
EBI Databases
NCBI Databases
AlphaFold Protein Structure Database

Database	Records Number	Description
AlphaFold DB	307623	AlphaFold Protein Structure Database

Powered by EBI AlphaFold DB

Literatures：Open Library of Bioscience



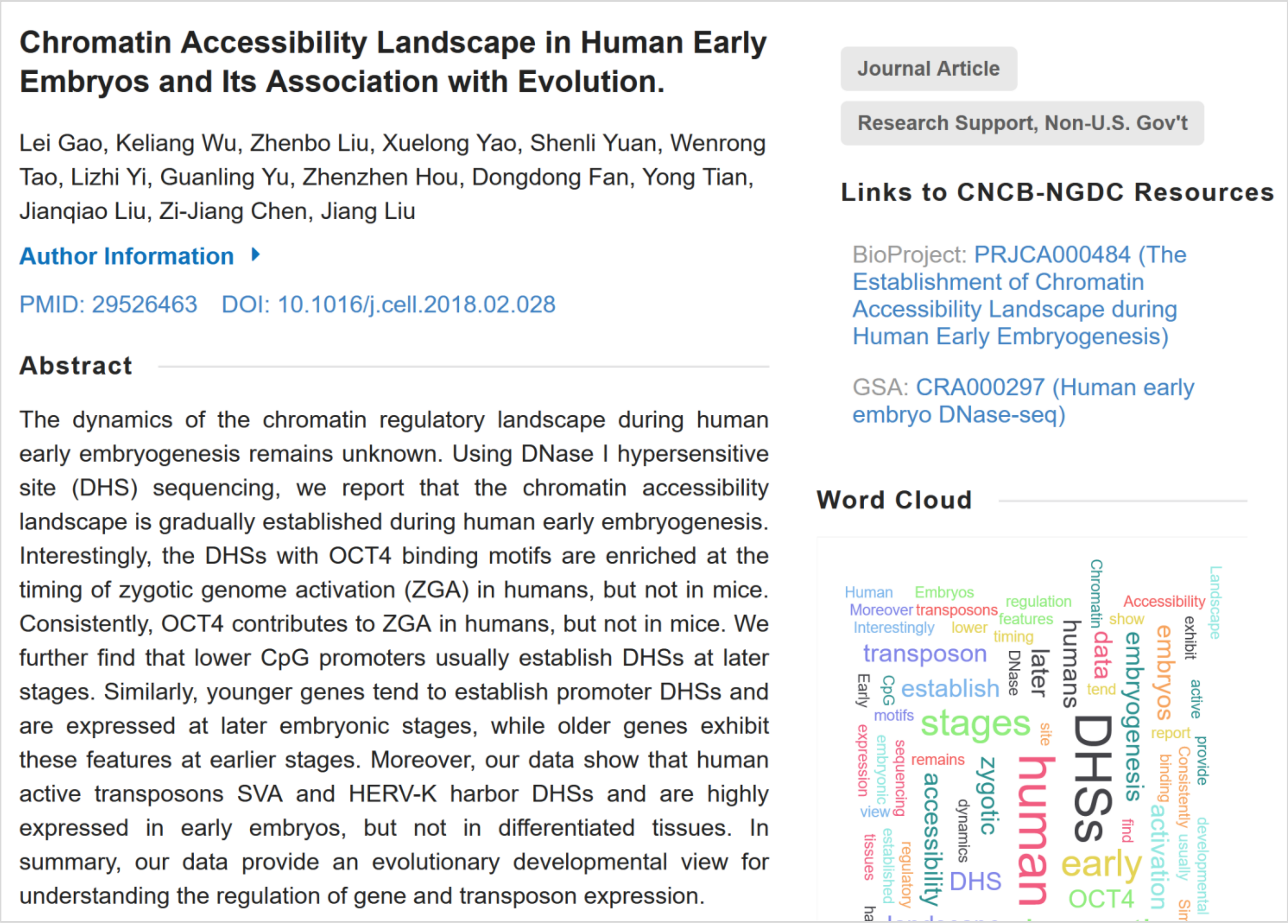
Open Library of Bioscience

OpenLB provides open access to ~33 millions literature texts with friendly links to relevant resources in CNCB-NGDC.

e.g., "COVID-19" OR "SARS-COV-2"; cancer

34,192,463 Publications

The OpenLB's literature texts are sourced from NCBI PubMed, bioRxiv and medRxiv, including title, abstract, author, journal, reference, etc.



Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution.

Lei Gao, Keliang Wu, Zhenbo Liu, Xuelong Yao, Shenli Yuan, Wenrong Tao, Lizhi Yi, Guanling Yu, Zhenzhen Hou, Dongdong Fan, Yong Tian, Jianqiao Liu, Zi-Jiang Chen, Jiang Liu

Author Information ▸

PMID: 29526463 DOI: 10.1016/j.cell.2018.02.028

Abstract

The dynamics of the chromatin regulatory landscape during human early embryogenesis remains unknown. Using DNase I hypersensitive site (DHS) sequencing, we report that the chromatin accessibility landscape is gradually established during human early embryogenesis. Interestingly, the DHSs with OCT4 binding motifs are enriched at the timing of zygotic genome activation (ZGA) in humans, but not in mice. Consistently, OCT4 contributes to ZGA in humans, but not in mice. We further find that lower CpG promoters usually establish DHSs at later stages. Similarly, younger genes tend to establish promoter DHSs and are expressed at later embryonic stages, while older genes exhibit these features at earlier stages. Moreover, our data show that human active transposons SVA and HERV-K harbor DHSs and are highly expressed in early embryos, but not in differentiated tissues. In summary, our data provide an evolutionary developmental view for understanding the regulation of gene and transposon expression.

Journal Article

Research Support, Non-U.S. Gov't

Links to CNCB-NGDC Resources

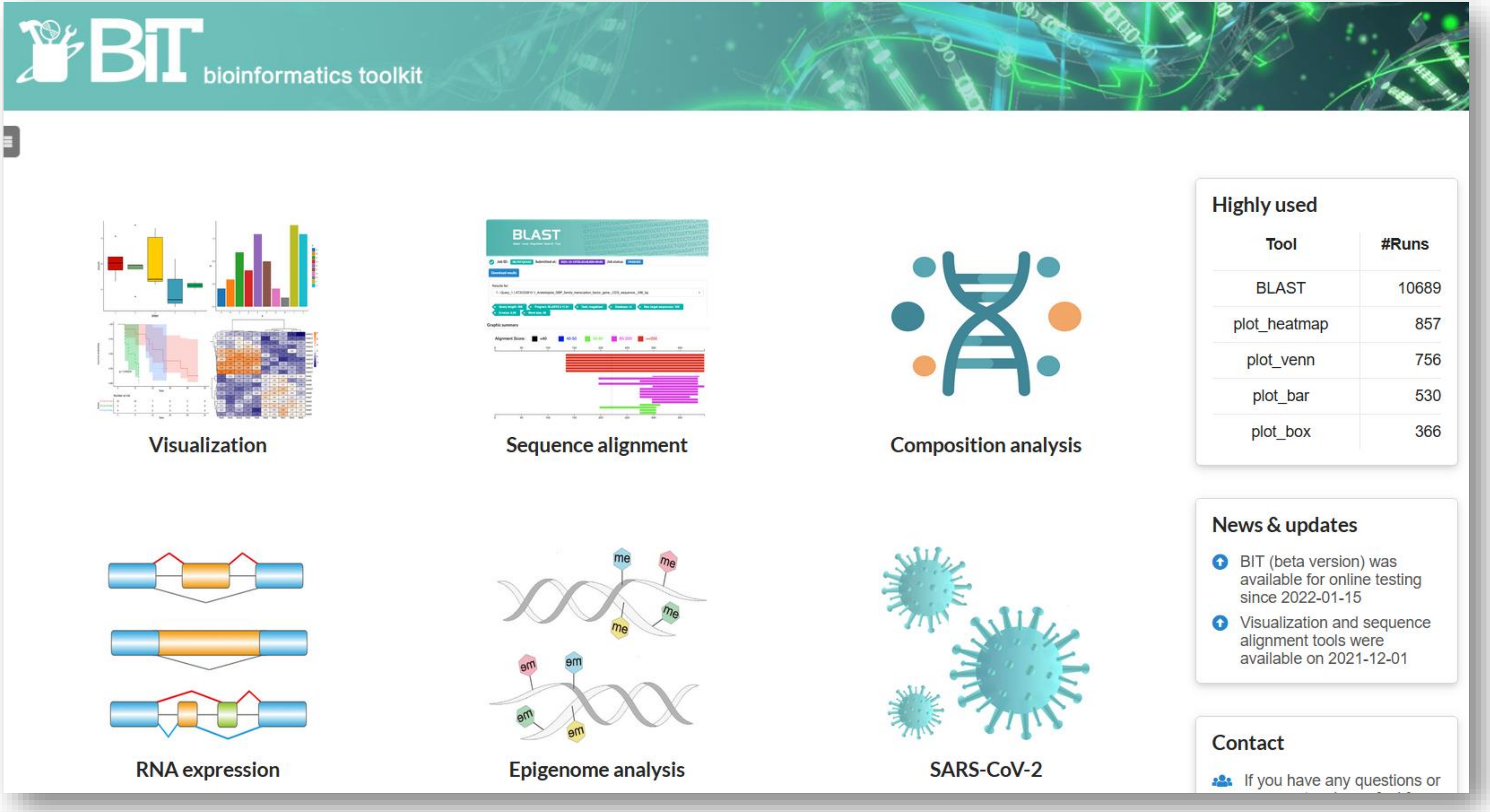
BioProject: PRJCA000484 (The Establishment of Chromatin Accessibility Landscape during Human Early Embryogenesis)

GSA: CRA000297 (Human early embryo DNase-seq)

Word Cloud

https://ngdc.cncb.ac.cn/openlb

Bioinformatics Tolls - BIT



Highly used

Tool	#Runs
BLAST	10689
plot_heatmap	857
plot_venn	756
plot_bar	530
plot_box	366

News & updates

- BIT (beta version) was available for online testing since 2022-01-15
- Visualization and sequence alignment tools were available on 2021-12-01

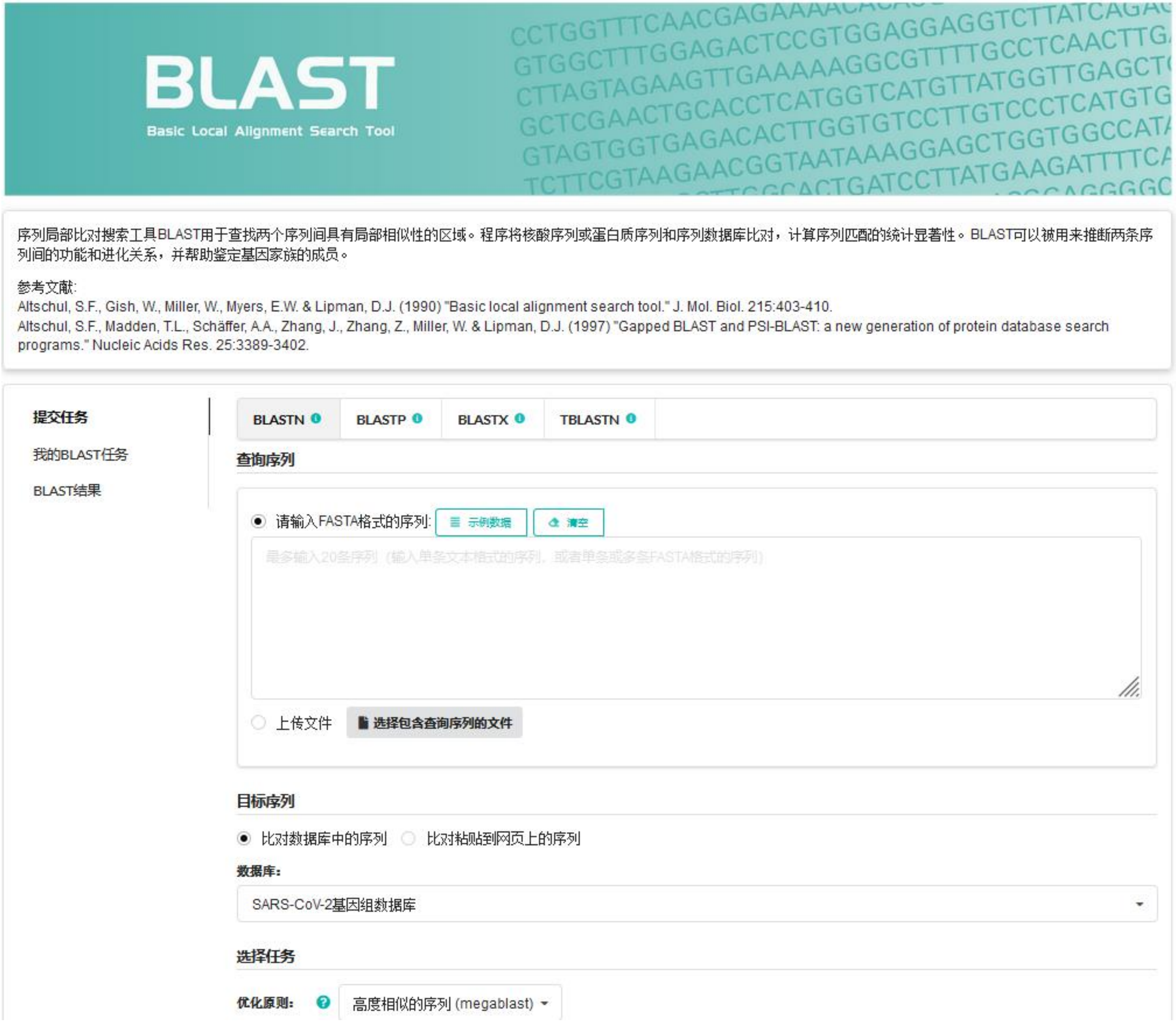
Contact

If you have any questions or

CNCB-NGDC

https://ngdc.cncb.ac.cn/bit/

BLAST



序列局部比对搜索工具BLAST用于查找两个序列间具有局部相似性的区域。程序将核酸序列或蛋白质序列和序列数据库比对，计算序列匹配的统计显著性。BLAST可以被用来推断两条序列间的功能和进化关系，并帮助鉴定基因家族的成员。

参考文献:
Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402.

查询序列

● 请输入FASTA格式的序列: 示例数据 清空

最多输入20条序列（输入单条文本格式的序列，或者单条或多条FASTA格式的序列）

○ 上传文件 选择包含查询序列的文件

目标序列

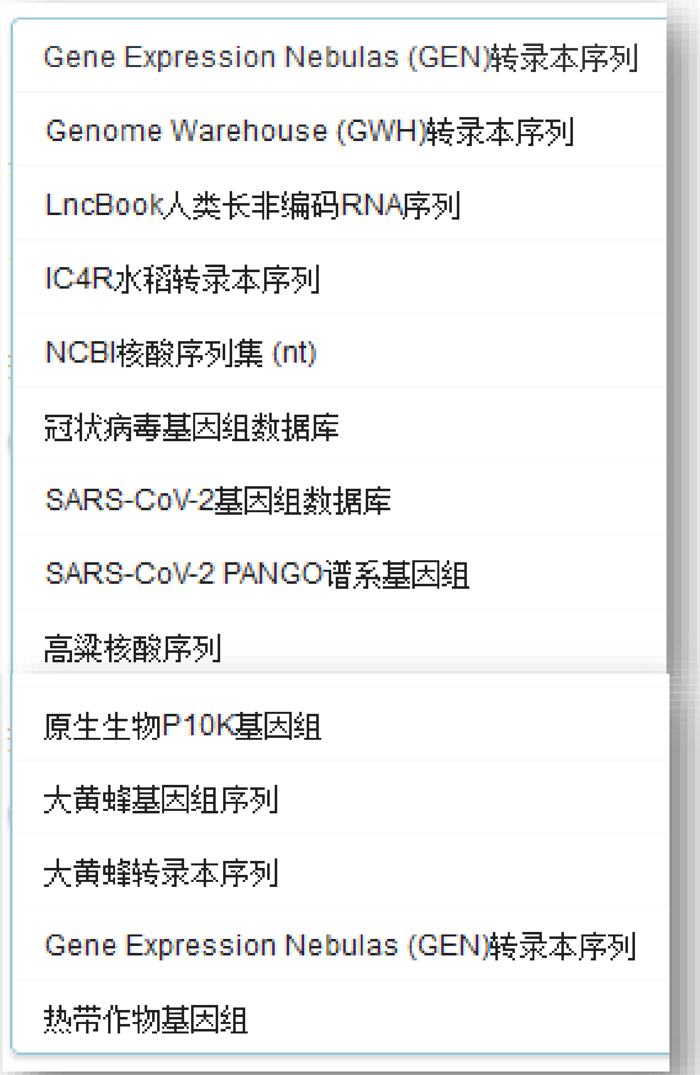
● 比对数据库中的序列 ○ 比对粘贴到网页上的序列

数据库: SARS-CoV-2基因组数据库

选择任务

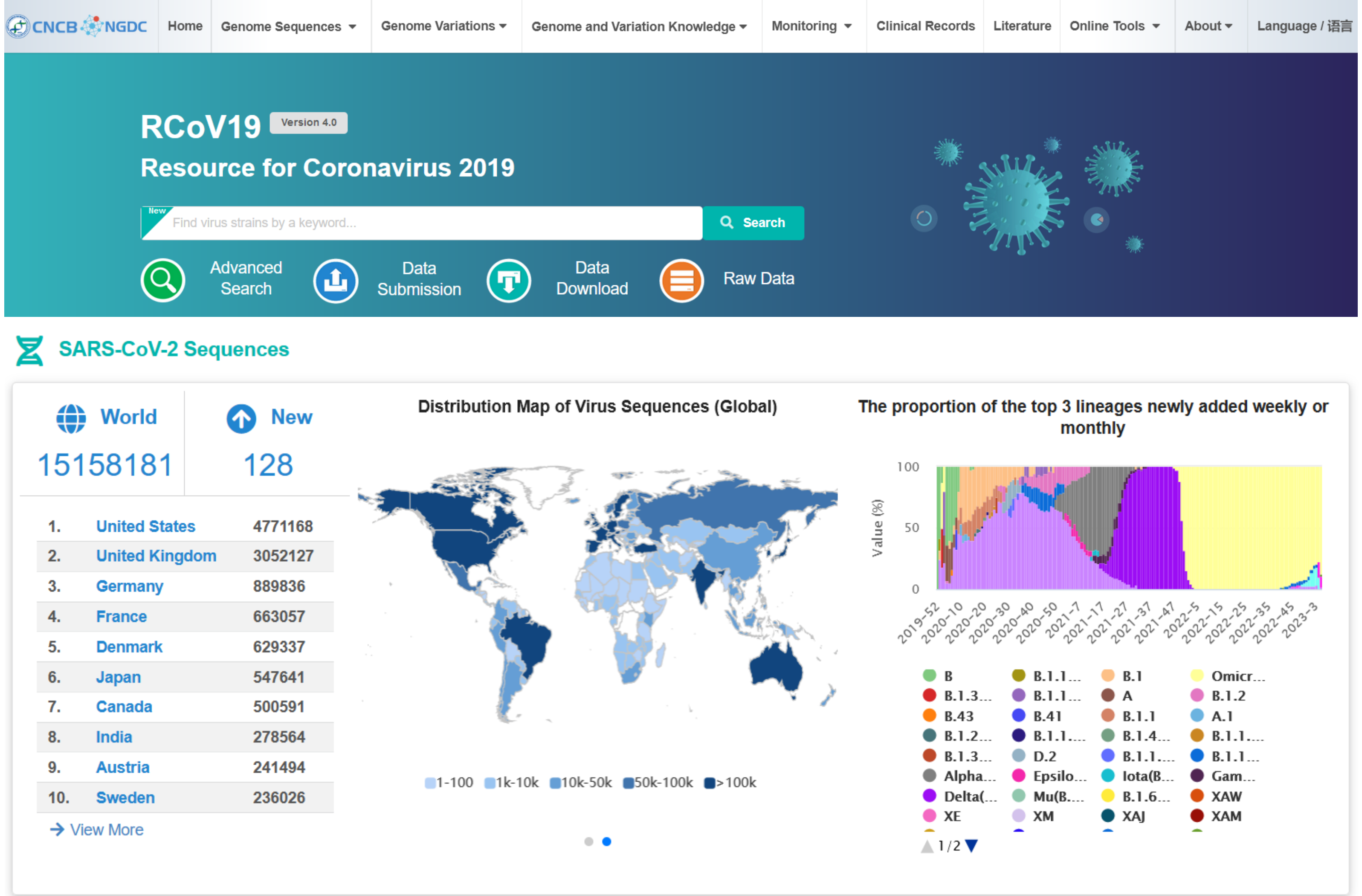
优化原则: 高度相似的序列 (megablast)

Customized databases



- Gene Expression Nebulas (GEN)转录本序列
- Genome Warehouse (GWH)转录本序列
- LncBook人类长非编码RNA序列
- IC4R水稻转录本序列
- NCBI核酸序列集 (nt)
- 冠状病毒基因组数据库
- SARS-CoV-2基因组数据库
- SARS-CoV-2 PANGO谱系基因组
- 高粱核酸序列
- 原生生物P10K基因组
- 大黄蜂基因组序列
- 大黄蜂转录本序列
- Gene Expression Nebulas (GEN)转录本序列
- 热带作物基因组

RCoV19



RCoV19 Version 4.0

Resource for Coronavirus 2019

SARS-CoV-2 Sequences

World	New
15158181	128

1.	United States	4771168
2.	United Kingdom	3052127
3.	Germany	889836
4.	France	663057
5.	Denmark	629337
6.	Japan	547641
7.	Canada	500591
8.	India	278564
9.	Austria	241494
10.	Sweden	236026

→ View More

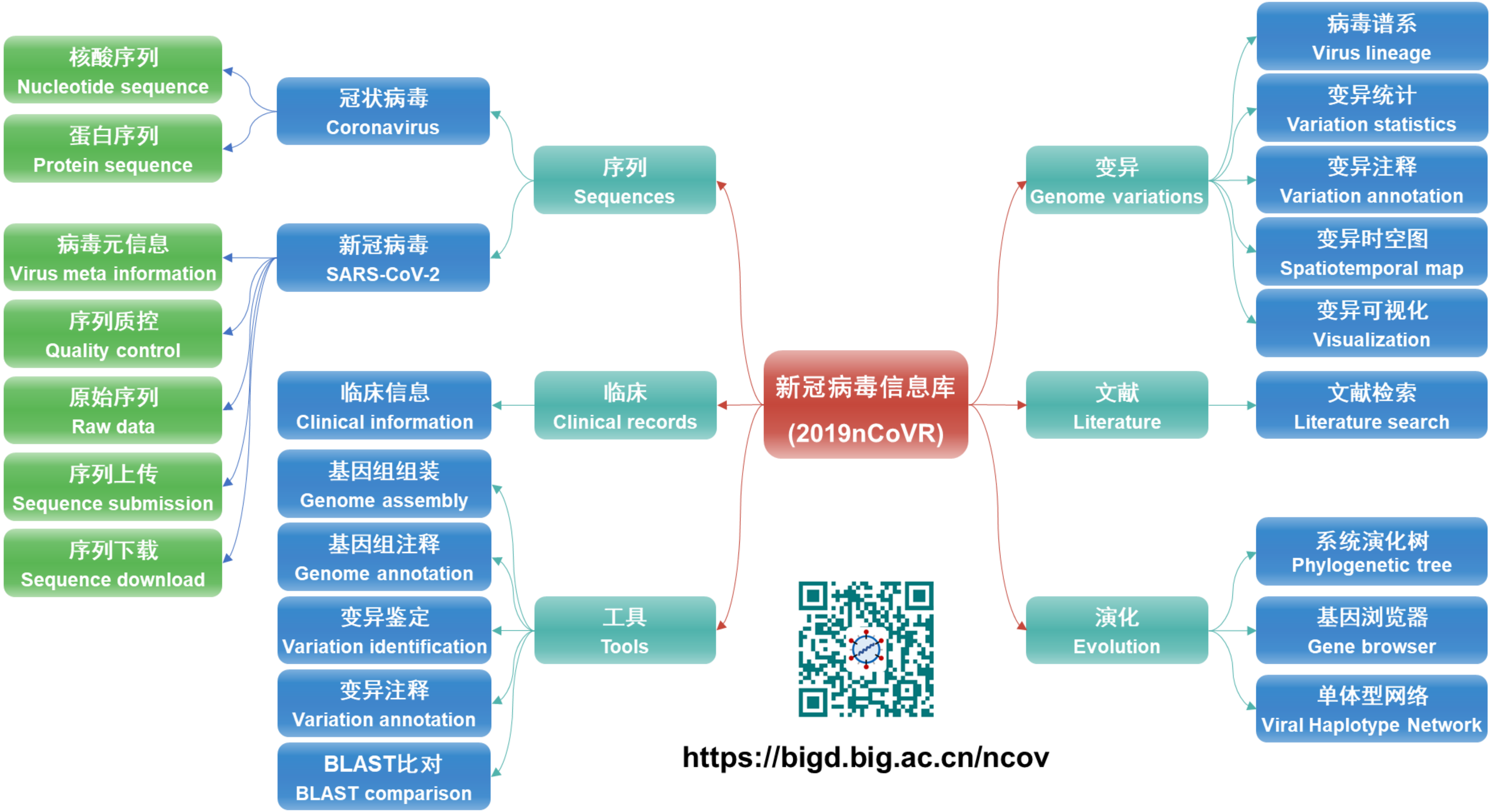
Distribution Map of Virus Sequences (Global)

The proportion of the top 3 lineages newly added weekly or monthly

Yi Chuan, 2020; *Zoological Research*, 2020; *Genomics Proteomics Bioinformatics*, 2020; *Nucleic Acids Research*, 2021

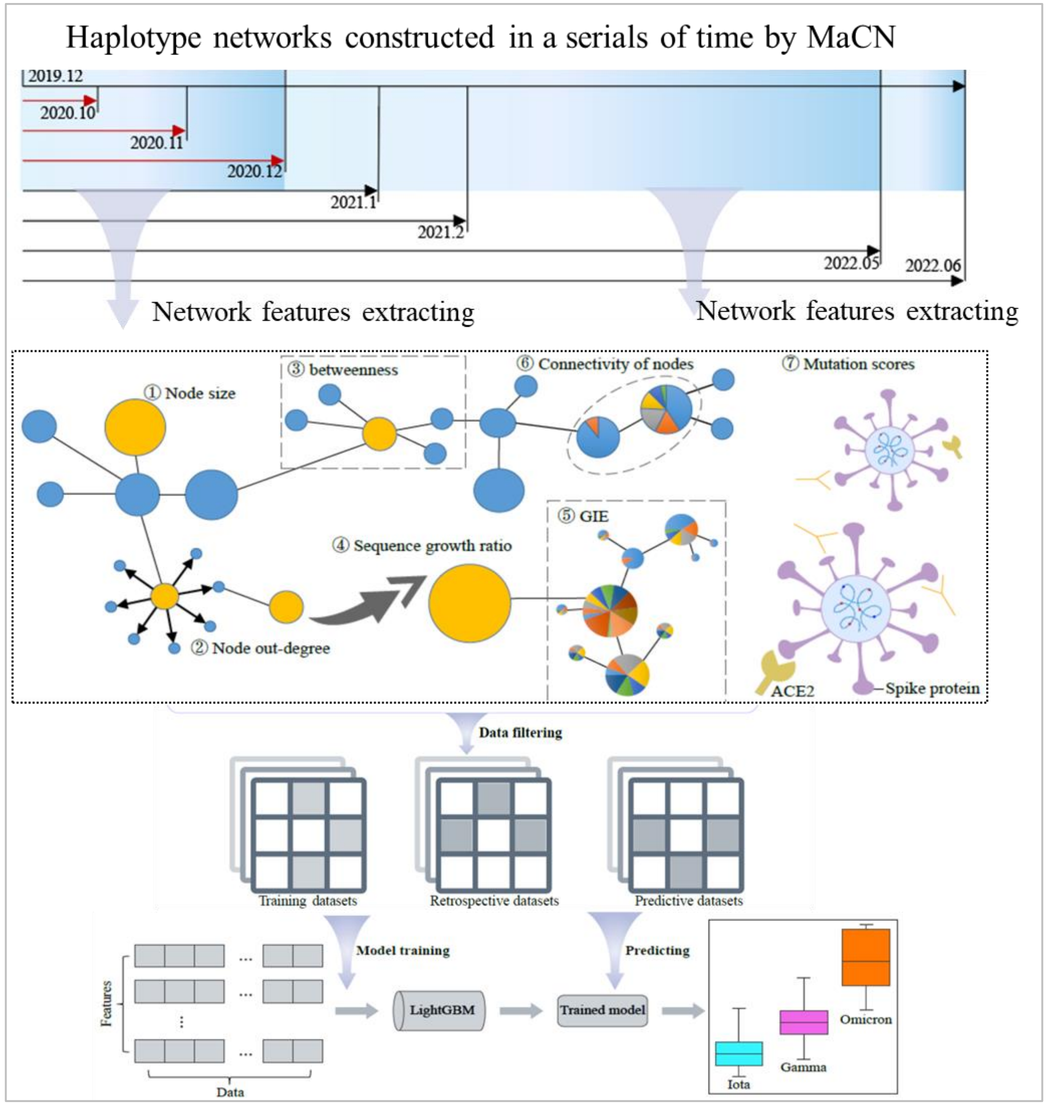
https://ngdc.cncb.ac.cn/ncov/

RCoV19

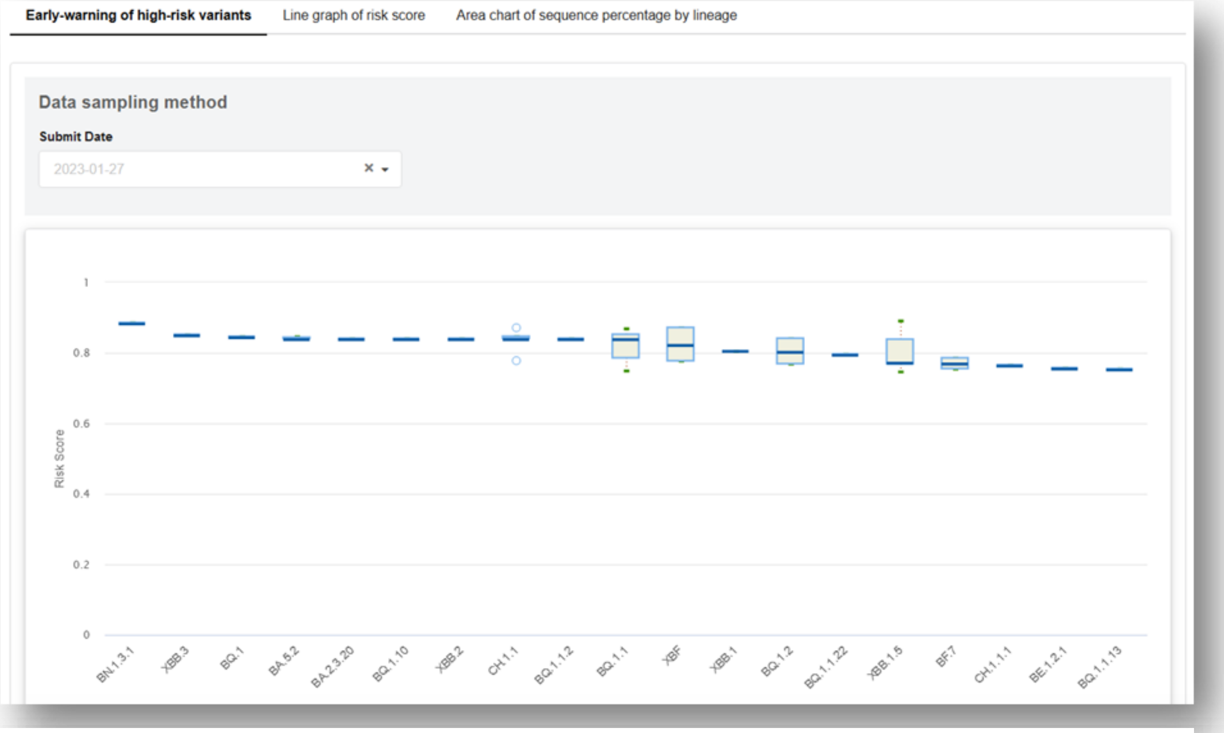


https://bigd.big.ac.cn/ncov

Machine learning detection of high-risk SARS-CoV-2 variants



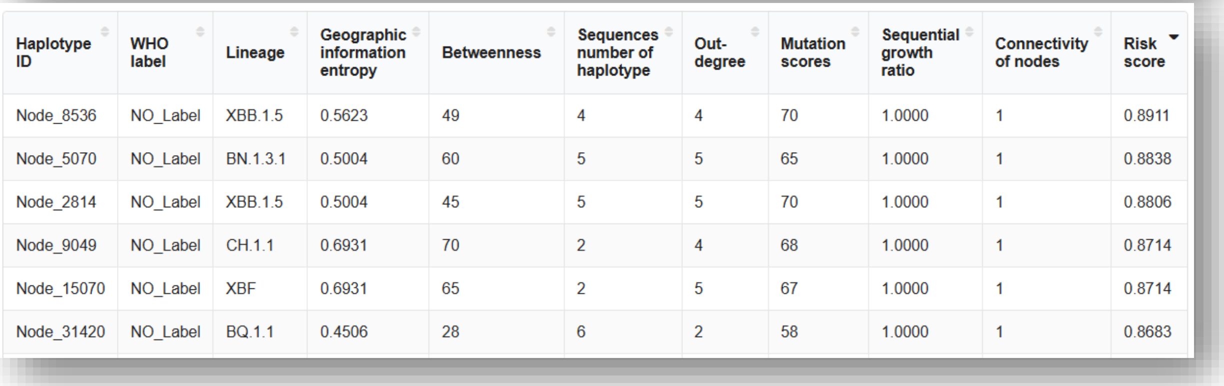
***Briefings in Bioinformatics*, 2023**



Early-warning of high-risk variants
Line graph of risk score
Area chart of sequence percentage by lineage

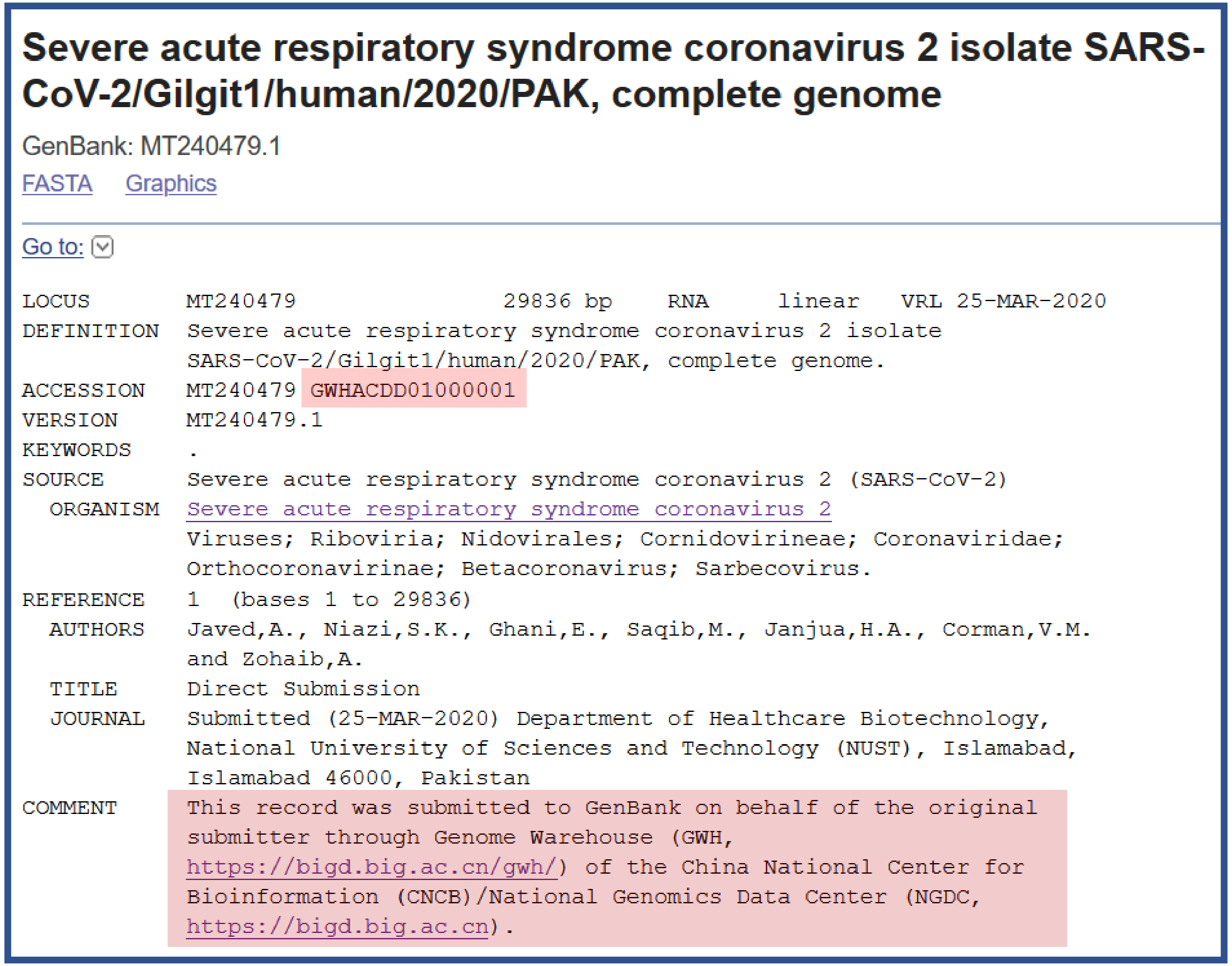
Data sampling method

Submit Date: 2023-01-27



Haplotype ID	WHO label	Lineage	Geographic information entropy	Betweenness	Sequences number of haplotype	Out-degree	Mutation scores	Sequential growth ratio	Connectivity of nodes	Risk score
Node_8536	NO_Label	XBB.1.5	0.5623	49	4	4	70	1.0000	1	0.8911
Node_5070	NO_Label	BN.1.3.1	0.5004	60	5	5	65	1.0000	1	0.8838
Node_2814	NO_Label	XBB.1.5	0.5004	45	5	5	70	1.0000	1	0.8806
Node_9049	NO_Label	CH.1.1	0.6931	70	2	4	68	1.0000	1	0.8714
Node_15070	NO_Label	XBF	0.6931	65	2	5	67	1.0000	1	0.8714
Node_31420	NO_Label	BQ.1.1	0.4506	28	6	2	58	1.0000	1	0.8683

Data Sharing with NCBI



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

GenBank: MT240479.1

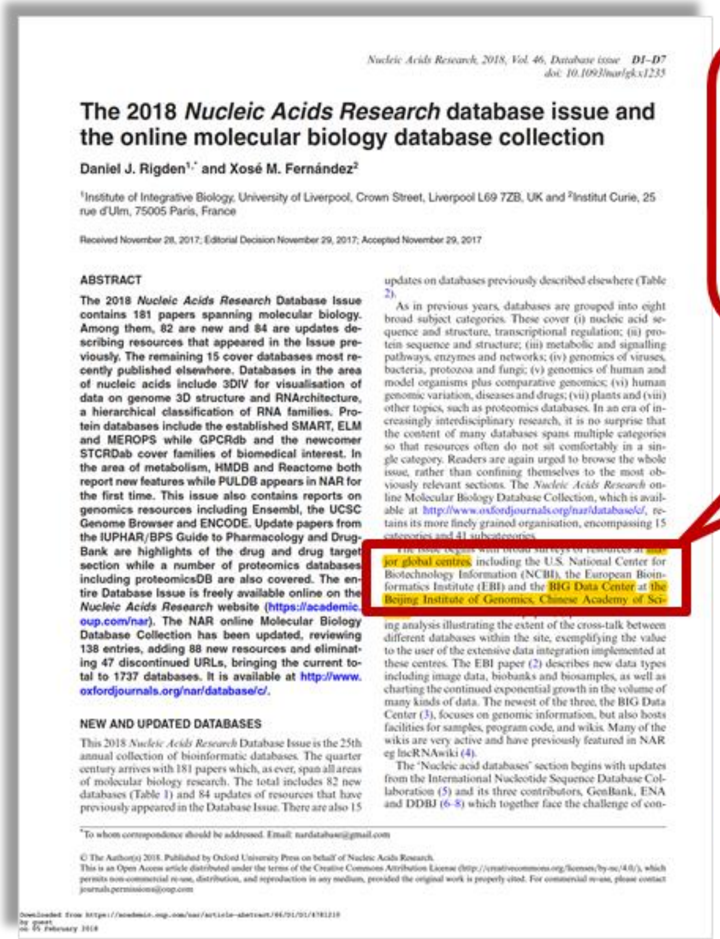
FASTA Graphics

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  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Nidovirales; Cornidovirineae; 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,
            https://bigd.big.ac.cn/gwh/) of the China National Center for
            Bioinformation (CNCB)/National Genomics Data Center (NGDC,
            https://bigd.big.ac.cn).
```

- 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

NGDC became a major global center



The 2018 *Nucleic Acids Research* database issue and the online molecular biology database collection

Daniel J. Rigden[1,*] and Xosé M. Fernández[2]

[1]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and [2]Institut Curie, 25 rue d'Ulm, 75005 Paris, France

Received November 28, 2017; Editorial Decision November 29, 2017; Accepted November 29, 2017

ABSTRACT

The 2018 *Nucleic Acids Research* Database Issue contains 181 papers spanning molecular biology. Among them, 82 are new and 84 are updates describing resources that appeared in the Issue previously. The remaining 15 cover databases most recently published elsewhere. Databases in the area of nucleic acids include 3DIV for visualisation of data on genome 3D structure and RNArchitecture, a hierarchical classification of RNA families. Protein databases include the established SMART, ELM and MEROPS while GPCRdb and the newcomer STCRDab cover families of biomedical interest. In the area of metabolism, HMDB and Reactome both report new features while PULDB appears in NAR for the first time. This issue also contains reports on genomics resources including Ensembl, the UCSC Genome Browser and ENCODE. Update papers from the IUPHAR/BPS Guide to Pharmacology and DrugBank are highlights of the drug and drug target section while a number of proteomics databases including proteomicsDB are also covered. The entire Database Issue is freely available online on the *Nucleic Acids Research* website (https://academic.oup.com/nar). The NAR online Molecular Biology Database Collection has been updated, reviewing 138 entries, adding 88 new resources and eliminating 47 discontinued URLs, bringing the current total to 1737 databases. It is available at http://www.oxfordjournals.org/nar/database/c/.

NEW AND UPDATED DATABASES

This 2018 *Nucleic Acids Research* Database Issue is the 25th annual collection of bioinformatic databases. The quarter century arrives with 181 papers which, as ever, span all areas of molecular biology research. The total includes 82 new databases (Table 1) and 84 updates of resources that have previously appeared in the Database Issue. There are also 15 updates on databases previously described elsewhere (Table 2).

As in previous years, databases are grouped into eight broad subject categories. These cover (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure; (iii) metabolic and signalling pathways, enzymes and networks; (iv) genomics of viruses, bacteria, protozoa and fungi; (v) genomics of human and model organisms plus comparative genomics; (vi) human genomic variation, diseases and drugs; (vii) plants and (viii) other topics, such as proteomics databases. In an era of increasingly interdisciplinary research, it is no surprise that the content of many databases spans multiple categories so that resources often do not sit comfortably in a single category. Readers are again urged to browse the whole issue, rather than confining themselves to the most obviously relevant sections. The *Nucleic Acids Research* online Molecular Biology Database Collection, which is available at http://www.oxfordjournals.org/nar/database/c/, retains its more finely grained organisation, encompassing 15 categories and 41 subcategories.

The issue begins with broad surveys of resources at major global centres, including the U.S. National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the BIG Data Center at the Beijing Institute of Genomics, Chinese Academy of Sciences. The NCBI Resources paper (1) presents an interesting analysis illustrating the extent of the cross-talk between different databases within the site, exemplifying the value to the user of the extensive data integration implemented at these centres. The EBI paper (2) describes new data types including image data, biobanks and biosamples, as well as charting the continued exponential growth in the volume of many kinds of data. The newest of the three, the BIG Data Center (3), focuses on genomic information, but also hosts facilities for samples, program code, and wikis. Many of the wikis are very active and have previously featured in NAR eg lncRNAwiki (4).

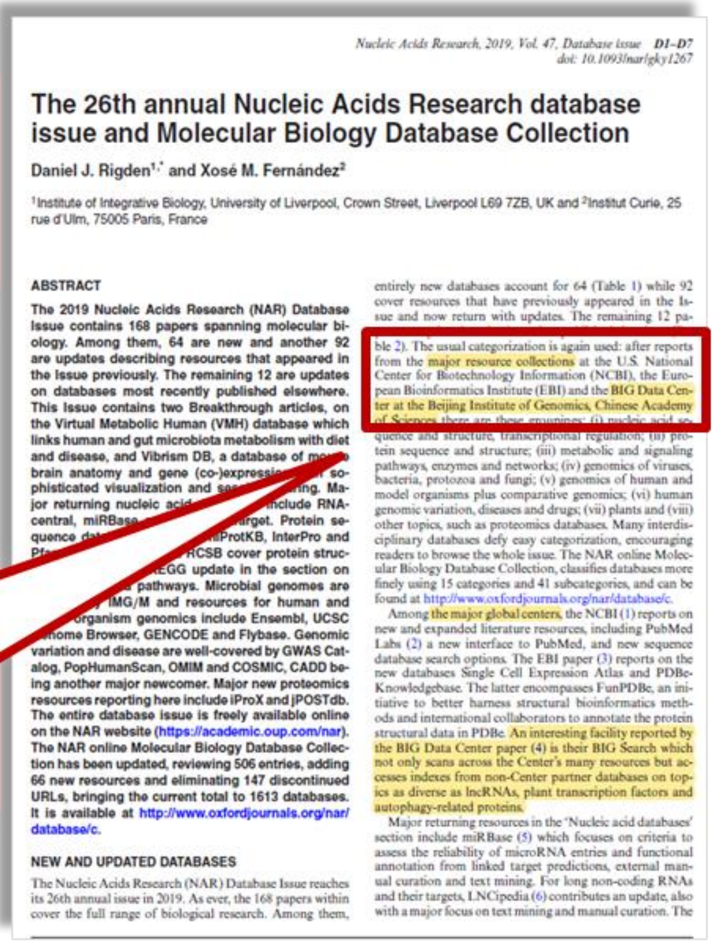
The 'Nucleic acid databases' section begins with updates from the International Nucleotide Sequence Database Collaboration (5) and its three contributors, GenBank, ENA and DDBJ (6–8) which together face the challenge of con-

[*]To whom correspondence should be addressed. Email: nardatabase@gmail.com

Nucleic Acids Research 2018, 46:D1–D7

The issue begins with broad surveys of resources at major global centres, including the U.S. National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the BIG Data Center at the Beijing Institute of Genomics, Chinese Academy of Sciences. The NCBI Resources paper (1) presents an interest-

ble 2). The usual categorization is again used: after reports from the major resource collections at the U.S. National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the BIG Data Center at the Beijing Institute of Genomics, Chinese Academy of Sciences there are these groupings: (i) nucleic acid se-



The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection

Daniel J. Rigden[1,*] and Xosé M. Fernández[2]

[1]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and [2]Institut Curie, 25 rue d'Ulm, 75005 Paris, France

ABSTRACT

The 2019 Nucleic Acids Research (NAR) Database Issue contains 168 papers spanning molecular biology. Among them, 64 are new and another 92 are updates describing resources that appeared in the Issue previously. The remaining 12 are updates on databases most recently published elsewhere. This Issue contains two Breakthrough articles, on the Virtual Metabolic Human (VMH) database which links human and gut microbiota metabolism with diet and disease, and Vibrism DB, a database of mo brain anatomy and gene (co-)expressio so-phisticated visualization and se ing. Major returning nucleic acid include RNA-central, miRBase rget. Protein sequence da niProtKB, InterPro and Pf RCSB cover protein struc- EGG update in the section on pathways. Microbial genomes are IMG/M and resources for human and organism genomics include Ensembl, UCSC nome Browser, GENCODE and Flybase. Genomic variation and disease are well-covered by GWAS Catalog, PopHumanScan, OMIM and COSMIC, CADD being another major newcomer. Major new proteomics resources reporting here include iProX and jPOSTdb. The entire database issue is freely available online on the NAR website (https://academic.oup.com/nar). The NAR online Molecular Biology Database Collection has been updated, reviewing 506 entries, adding 66 new resources and eliminating 147 discontinued URLs, bringing the current total to 1613 databases. It is available at http://www.oxfordjournals.org/nar/database/c.

NEW AND UPDATED DATABASES

The Nucleic Acids Research (NAR) Database Issue reaches its 26th annual issue in 2019. As ever, the 168 papers within cover the full range of biological research. Among them, entirely new databases account for 64 (Table 1) while 92 cover resources that have previously appeared in the Issue and now return with updates. The remaining 12 pa- ble 2). The usual categorization is again used: after reports from the major resource collections at the U.S. National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the BIG Data Center at the Beijing Institute of Genomics, Chinese Academy of Sciences there are these groupings: (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure; (iii) metabolic and signaling pathways, enzymes and networks; (iv) genomics of viruses, bacteria, protozoa and fungi; (v) genomics of human and model organisms plus comparative genomics; (vi) human genomic variation, diseases and drugs; (vii) plants and (viii) other topics, such as proteomics databases. Many interdisciplinary databases defy easy categorization, encouraging readers to browse the whole issue. The NAR online Molecular Biology Database Collection, classifies databases more finely using 15 categories and 41 subcategories, and can be found at http://www.oxfordjournals.org/nar/database/c.

Among the major global centers, the NCBI (1) reports on new and expanded literature resources, including PubMed Labs (2) a new interface to PubMed, and new sequence database search options. The EBI paper (3) reports on the new databases Single Cell Expression Atlas and PDBe-Knowledgebase. The latter encompasses FunPDBe, an initiative to better harness structural bioinformatics methods and international collaborators to annotate the protein structural data in PDBe. An interesting facility reported by the BIG Data Center paper (4) is their BIG Search which not only scans across the Center's many resources but accesses indexes from non-Center partner databases on topics as diverse as lncRNAs, plant transcription factors and autophagy-related proteins.

Major returning resources in the 'Nucleic acid databases' section include miRBase (5) which focuses on criteria to assess the reliability of microRNA entries and functional annotation from linked target predictions, external manual curation and text mining. For long non-coding RNAs and their targets, LNCipedia (6) contributes an update, also with a major focus on text mining and manual curation. The

Nucleic Acids Research 2019, 47:D1–D7

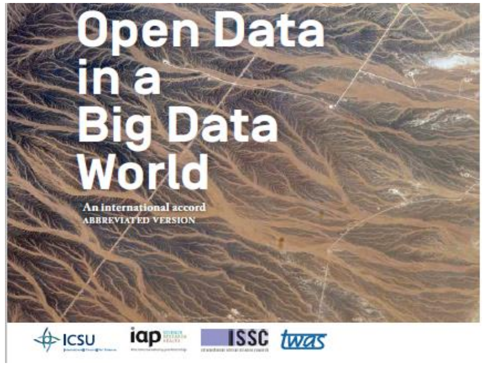
BHBD Alliance

About BHBD

BHBD Alliance is a non-profit, non-governmental organization founded in October 2018 for promoting biodiversity and health big data sharing in the world, under the framework of “Open Biodiversity and Health Big Data Initiative” by IUBS.

Vision of BHBD

BHBD is committed to developing a world-wide open platform for biodiversity and health big data integration, translation and sharing, under the FAIR principles.



Open Data in a Big Data World

An international accord
ABBREVIATED VERSION

ICSU iap ISSC twas

https://ngdc.cncb.ac.cn/bhbd-alliance

BHBD Establishment and Membership Expanding

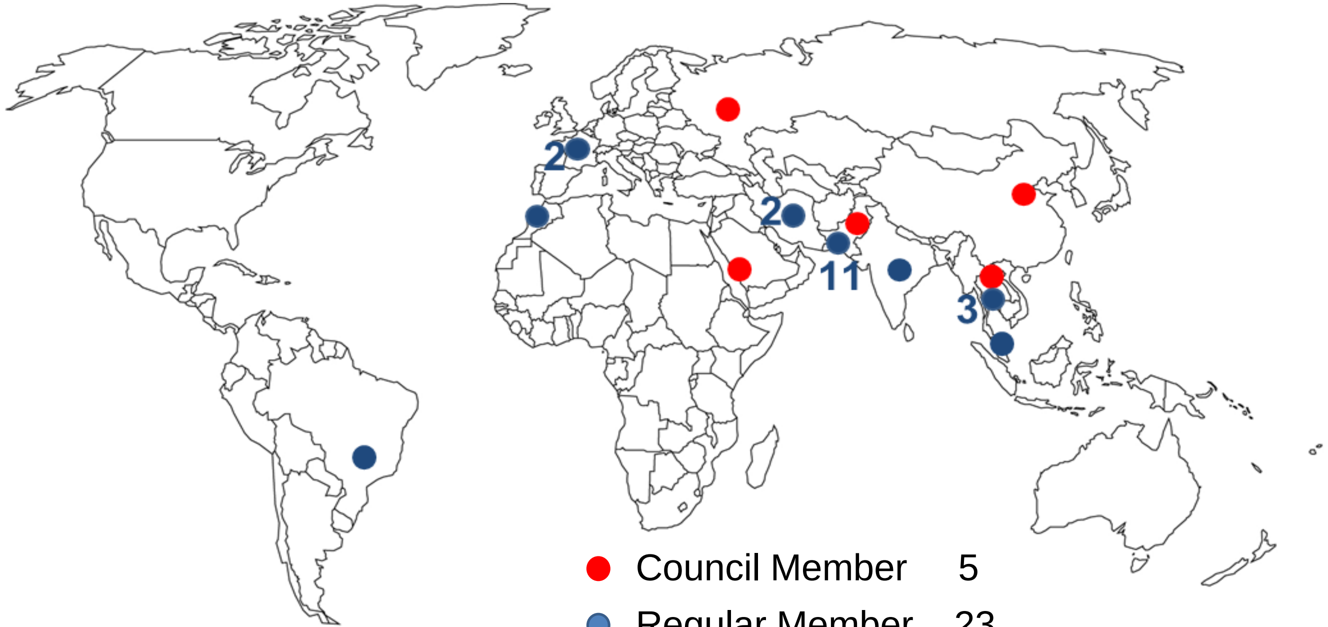
BIG, CAS/CNCB China

QAU Pakistan

VIGG Russia

KAUST Saudi Arabia

CU Thailand



● Council Member 5

● Regular Member 23

28 Members **12** Countries

(As of Dec 2022)

Regular Members:

Brazil	1	Malaysia	1	Thailand	3
France	2	Morocco	1		
India	1	Nepal	1		
Iran	2	Pakistan	11		

International Meetings/Trainings

- ❑ **Organization of Int’l meetings: 10**
- ❑ **International trainings: 200+ persons**
- ❑ **Visiting scholars to China: 13 persons**

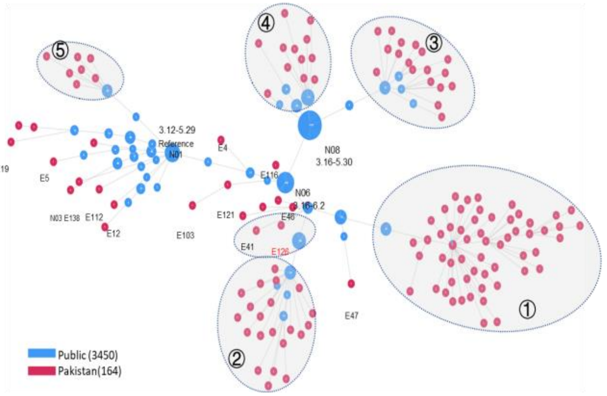
Visiting scholars

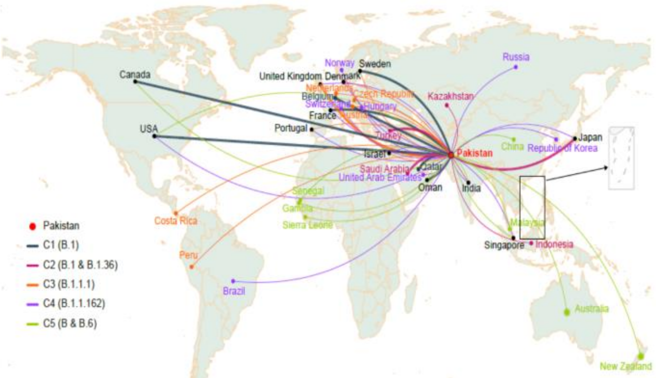
BHBD Int’l Symposium
Jul., 2019, Pakistan

Big Data Forum on Life and Health
Oct., 2019, Beijing

International Joint Research

- SARS-CoV-2 sample sequencing & analyses: Pakistan & BRICS
- Data sharing: 300+ datasets
- Joint publications: 10+





N S F C

saMRC advancinglife

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

Department of Science & Technology
Ministry of Science & Technology
Government of India
सत्यमेव जयते

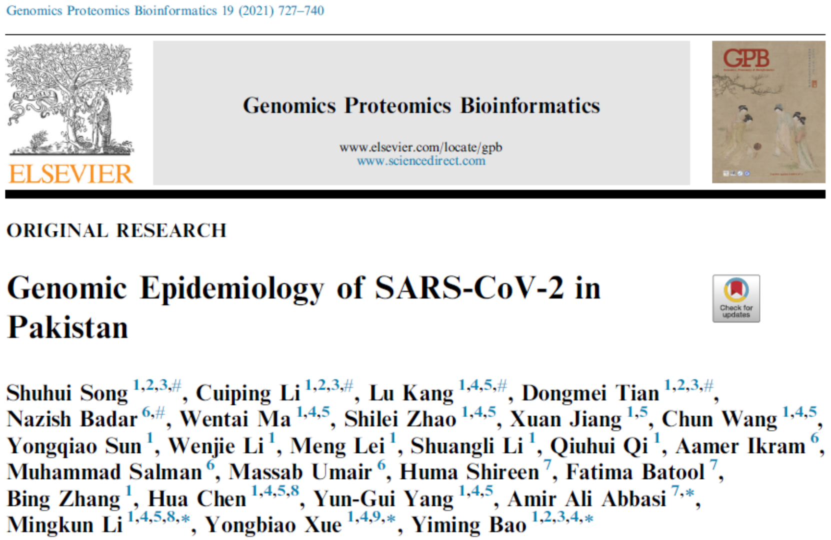
RFBR
RUSSIAN FOUNDATION FOR BASIC RESEARCH

FASIE

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

जैव प्रौद्योगिकी विभाग
Department of Biotechnology
Ministry of Science & Technology
Government of India
सत्यमेव जयते

BRICS STI Framework Programme
Response to COVID-19 pandemic coordinated call for BRICS multilateral projects 2020


Genomics Proteomics Bioinformatics 19 (2021) 727–740

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

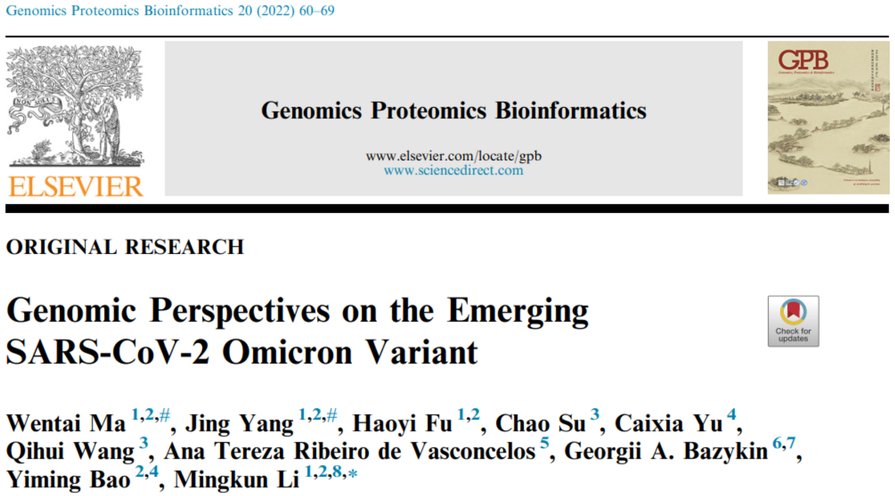
GPB

ORIGINAL RESEARCH

Genomic Epidemiology of SARS-CoV-2 in Pakistan

Check for updates

Shuhui Song [1,2,3,#], Cuiping Li [1,2,3,#], Lu Kang [1,4,5,#], Dongmei Tian [1,2,3,#], Nazish Badar [6,#], Wentai Ma [1,4,5], Shilei Zhao [1,4,5], Xuan Jiang [1,5], Chun Wang [1,4,5], Yongqiao Sun [1], Wenjie Li [1], Meng Lei [1], Shuangli Li [1], Qiuhui Qi [1], Aamer Ikram [6], Muhammad Salman [6], Massab Umair [6], Huma Shireen [7], Fatima Batool [7], Bing Zhang [1], Hua Chen [1,4,5,8], Yun-Gui Yang [1,4,5], Amir Ali Abbasi [7,*], Mingkun Li [1,4,5,8,*], Yongbiao Xue [1,4,9,*], Yiming Bao [1,2,3,4,*]


Genomics Proteomics Bioinformatics 20 (2022) 60–69

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

GPB

ORIGINAL RESEARCH

Genomic Perspectives on the Emerging SARS-CoV-2 Omicron Variant

Check for updates

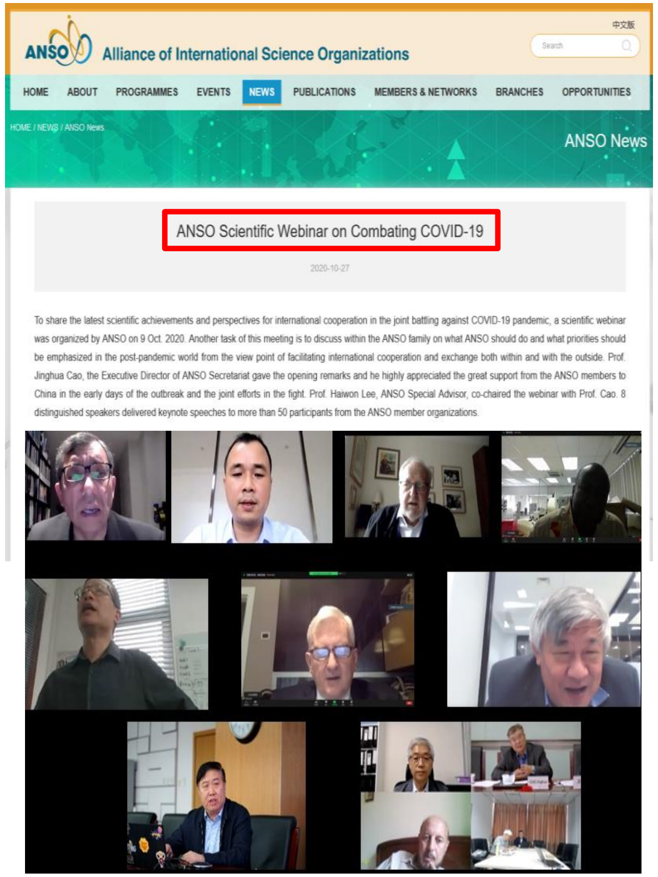
Wentai Ma [1,2,#], Jing Yang [1,2,#], Haoyi Fu [1,2], Chao Su [3], Caixia Yu [4], Qihui Wang [3], Ana Tereza Ribeiro de Vasconcelos [5], Georgii A. Bazykin [6,7], Yiming Bao [2,4], Mingkun Li [1,2,8,*]

Association with ANSO

- BHBD became one of the international associations of ANSO (Alliance of International Science Organizations) since 2020

- BHBD contributed for ANSO's activities to fight against COVID-19

- CNCB-NGDC data resources were introduced in *ANSO Highlight for Open Data* and recommended by ANSO President Prof. BAI Chunli

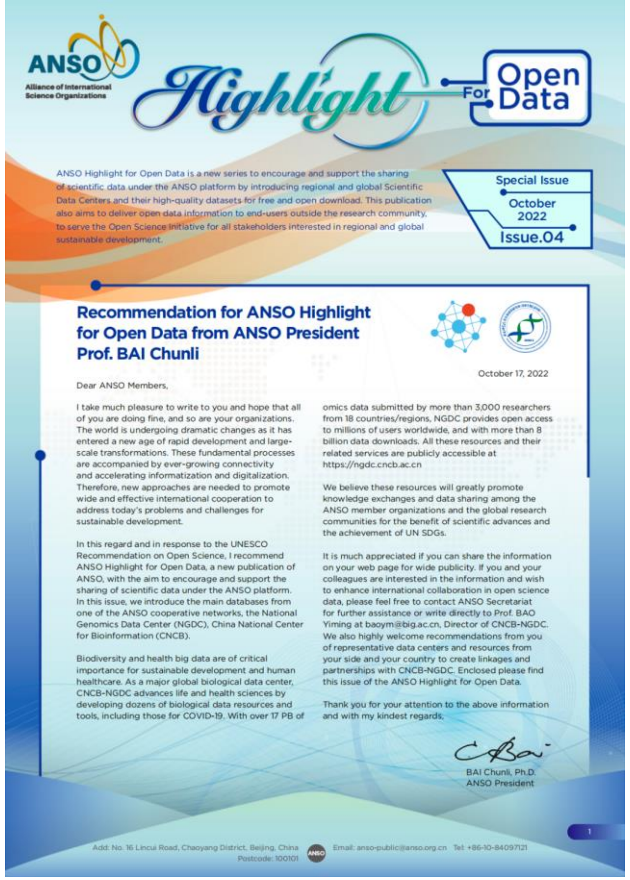


ANSO Scientific Webinar on Combating COVID-19

2020-10-27

To share the latest scientific achievements and perspectives for international cooperation in the joint battling against COVID-19 pandemic, a scientific webinar was organized by ANSO on 9 Oct. 2020. Another task of this meeting is to discuss within the ANSO family on what ANSO should do and what priorities should be emphasized in the post-pandemic world from the view point of facilitating international cooperation and exchange both within and with the outside. Prof. Jinghua Cao, the Executive Director of ANSO Secretariat gave the opening remarks and he highly appreciated the great support from the ANSO members to China in the early days of the outbreak and the joint efforts in the fight. Prof. Haiwon Lee, ANSO Special Advisor, co-chaired the webinar with Prof. Cao. 8 distinguished speakers delivered keynote speeches to more than 50 participants from the ANSO member organizations.

ANSO Webinar on COVID-19
Oct 2020


ANSO
Alliance of International Science Organizations
Highlight
For
Open
Data

ANSO Highlight for Open Data is a new series to encourage and support the sharing of scientific data under the ANSO platform by introducing regional and global Scientific Data Centers and their high-quality datasets for free and open download. This publication also aims to deliver open data information to end-users outside the research community, to serve the Open Science Initiative for all stakeholders interested in regional and global sustainable development.

Special Issue
October 2022
Issue.04

Recommendation for ANSO Highlight for Open Data from ANSO President Prof. BAI Chunli

October 17, 2022

Dear ANSO Members,

I take much pleasure to write to you and hope that all of you are doing fine, and so are your organizations. The world is undergoing dramatic changes as it has entered a new age of rapid development and large-scale transformations. These fundamental processes are accompanied by ever-growing connectivity and accelerating informatization and digitalization. Therefore, new approaches are needed to promote wide and effective international cooperation to address today's problems and challenges for sustainable development.

In this regard and in response to the UNESCO Recommendation on Open Science, I recommend ANSO Highlight for Open Data, a new publication of ANSO, with the aim to encourage and support the sharing of scientific data under the ANSO platform. In this issue, we introduce the main databases from one of the ANSO cooperative networks, the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB).

Biodiversity and health big data are of critical importance for sustainable development and human healthcare. As a major global biological data center, CNCB-NGDC advances life and health sciences by developing dozens of biological data resources and tools, including those for COVID-19. With over 17 PB of omics data submitted by more than 3,000 researchers from 18 countries/regions, NGDC provides open access to millions of users worldwide, and with more than 8 billion data downloads. All these resources and their related services are publicly accessible at https://ngdc.cncb.ac.cn

We believe these resources will greatly promote knowledge exchanges and data sharing among the ANSO member organizations and the global research communities for the benefit of scientific advances and the achievement of UN SDGs.

It is much appreciated if you can share the information on your web page for wide publicity. If you and your colleagues are interested in the information and wish to enhance international collaboration in open science data, please feel free to contact ANSO Secretariat for further assistance or write directly to Prof. BAO Yiming at baoym@big.ac.cn, Director of CNCB-NGDC. We also highly welcome recommendations from you of representative data centers and resources from your side and your country to create linkages and partnerships with CNCB-NGDC. Enclosed please find this issue of the ANSO Highlight for Open Data.

Thank you for your attention to the above information and with my kindest regards,

C Bai

BAI Chunli, Ph.D.
ANSO President

1

Add: No. 16 Lincui Road, Chaoyang District, Beijing, China
Postcode: 100101
Email: anso-public@anso.org.cn Tel: +86-10-84097121

ANSO Highlight for Open Data
Oct 2022

Grants Awarded for International Collaboration

Funding Agency	Project Title	Duration	Collaborators	Amount
IUBS	Open Biodiversity and Health Big Data Initiative	2019-2022	Multiple countries	Euro 30,200
ANSO	Global Biodiversity and Health Big Data Alliance	2020-2022	Multiple countries	RMB 750,000
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 (Macaca fascicularis): 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

Summary

- ❑ A comprehensive bioinformatics resource
  - ➢ Multi-omics DBs（GSA、GenBase、GWH、GVM、MethBank）
  - ➢ Knowledgebases（TCOD、RCoV19）
  - ➢ Tools and literatures（BLAST、OpenLB）
- ❑ The establishment of 3 national centers/platform
  - ➢ CNCB
  - ➢ NGDC
  - ➢ HGRIP
- ❑ International recognitions
  - ➢ Publishers（Springer Nature、Elsevier）
  - ➢ Peers（NAR Database Issue）
  - ➢ Major global centers（INSDC）

Take home messages

- Genome data archiving at INSDC is the consensus for the community
- It should not be taken for granted, considering technical difficulties
- Regional/national data centers can play big roles in promoting data sharing and archiving, thus are complementing INSDC
- Data exchange mechanism can be established between local centers and INSDC to facilitate data sharing and preservation
- Compared to OA of literature, OA of genomic data is still challenging, and needs new mechanisms/business models

Acknowledgements

Steering Advisors:

- Runsheng Chen
- Guoping Zhao

Scientific Advisors:

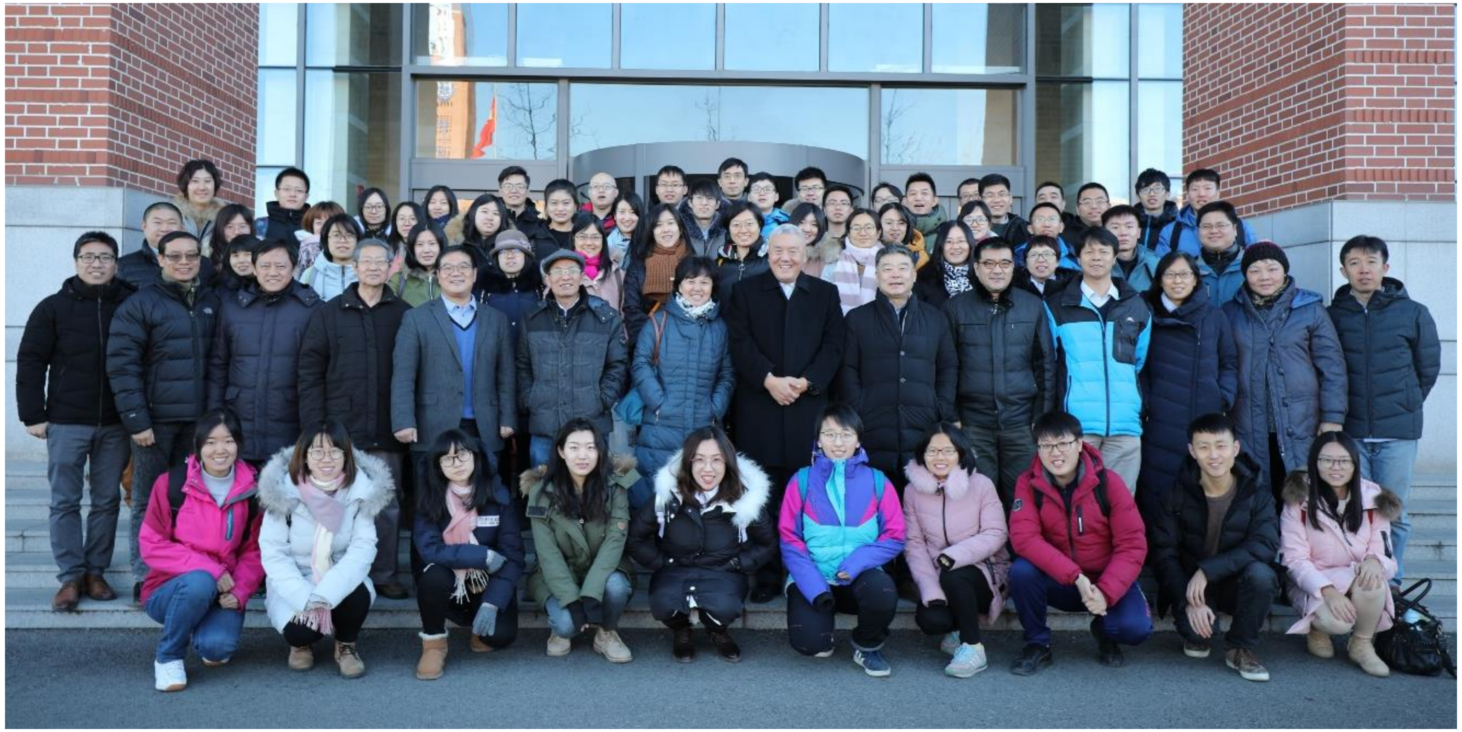
- Amos Bairoch (SIB)
- Guy Cochrane (EBI)
- Frank Eisenhaber (BI)
- Takashi Gojobori (KAUST)
- Yixue Li (CAS)
- Jingchu Luo (PKU)
- Ilene Mizrachi (NCBI)
- Yasukazu Nakamura (DDBJ)
- Weimin Zhu (CAS)

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



中国科学院–马普学会
计算生物学伙伴研究所
(2005–2020)

中国科学院生物物理研究所
Institute of Biophysics, Chinese Academy of Sciences

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

Nsti
国家科技资源共享服务平台
National Science & Technology Infrastructure

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

国家自然科学基金委员会
N S F C

ANSO
Alliance of International Science Organizations

IUBS

Thank You!
We are open for collaborations

NGDC

BHBD Alliance

baoym@big.ac.cn