China National Center for Bioinformation

Resources of China National Center for Bioinformation (CNCB)

Yiming Bao

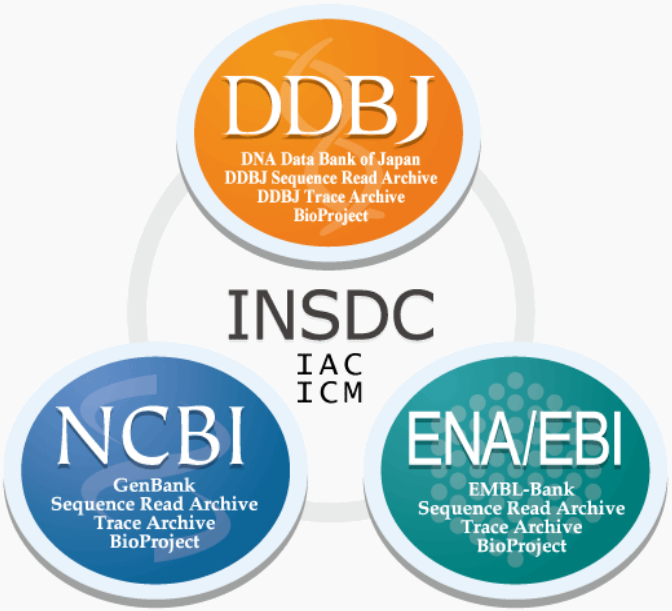
Director
National Genomics Data Center
Beijing, China

The 18th ABC Symposium
Dec. 08, 2021 • Zoom

中国科学院北京基因组研究所
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES

International Nucleotide Sequence Database Collaboration (INSDC)

Japan,1986



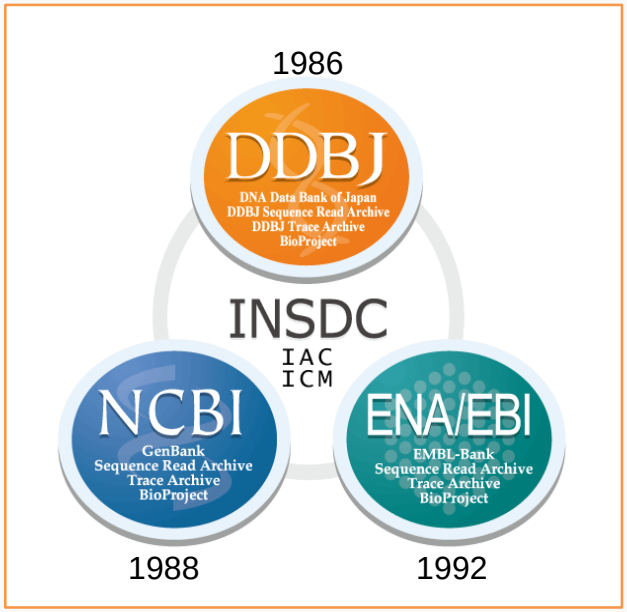
USA, 1988 Europe, 1992

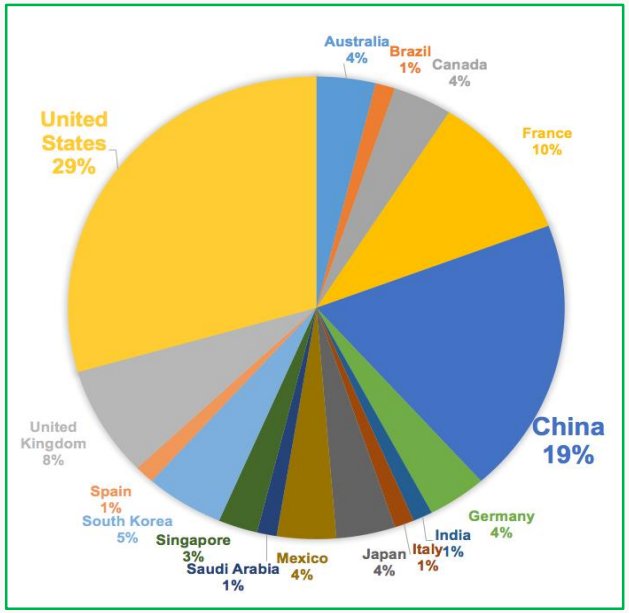
- NCBI: 1988, by US congress
- EBI: 1992, by EMBL
- DDBJ: 1986, by NIG of Japan
- NCBI, EBI and DDBJ form INSDC
- Establish international standard, exchange data daily, hold annual meeting
- Before papers are published, data need to be deposited into an international recognized database

Background in China

- **Big Data generated from Large-scale National Research Projects based on genome sequencing**

Big Data, Big Challenges





It is estimated that at least 20% of data in INSDC is produced by China

Background in China

- Big Data generated from Large-scale National Research Projects based on genome sequencing
- **Lack of data sharing in China**

Background in China

- **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

Background in China

- **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

Background in China

- **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

Background in China

350TB of data were shipped to China National Center for Bioinformation in servers through express service

Li et al. A genomic and epigenomic atlas of prostate cancer in Asian populations. Nature 2020

Large Data Submission to GSA-Human

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

Background in China

- **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

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

NGDC SAB Members

Prof. Amos Bairoch

Vice-president of the Basic Medical Science section, University of Geneva and Swiss Institute of Bioinformatics

Prof. Yasukazu Nakamura

Genome Informatics Laboratory
DNA Data Bank of Japan (**DDBJ**)
National Institute of Genetics, Japan

Prof. Jingchu Luo

China Node Manager of the European Molecular Biology Network (EMBnet)
Peking University, China

Dr. Guy Cochrane

Head of European Nucleotide Archive
European Bioinformatics Institute (**EBI**)
United Kingdom

Dr. Frank Eisenhaber

Head of Biomolecular Function Discovery Division, Executive Director of Bioinformatics Institute, Singapore

Dr. Ilene Mizrachi

GenBank Coordinator
National Center for Biotechnology Information (**NCBI**), United States

Prof. Takashi Gojobori (Chair)

Distinguished Professor of King Abdullah University of Science and Technology, Saudi Arabia

Prof. Yixue Li

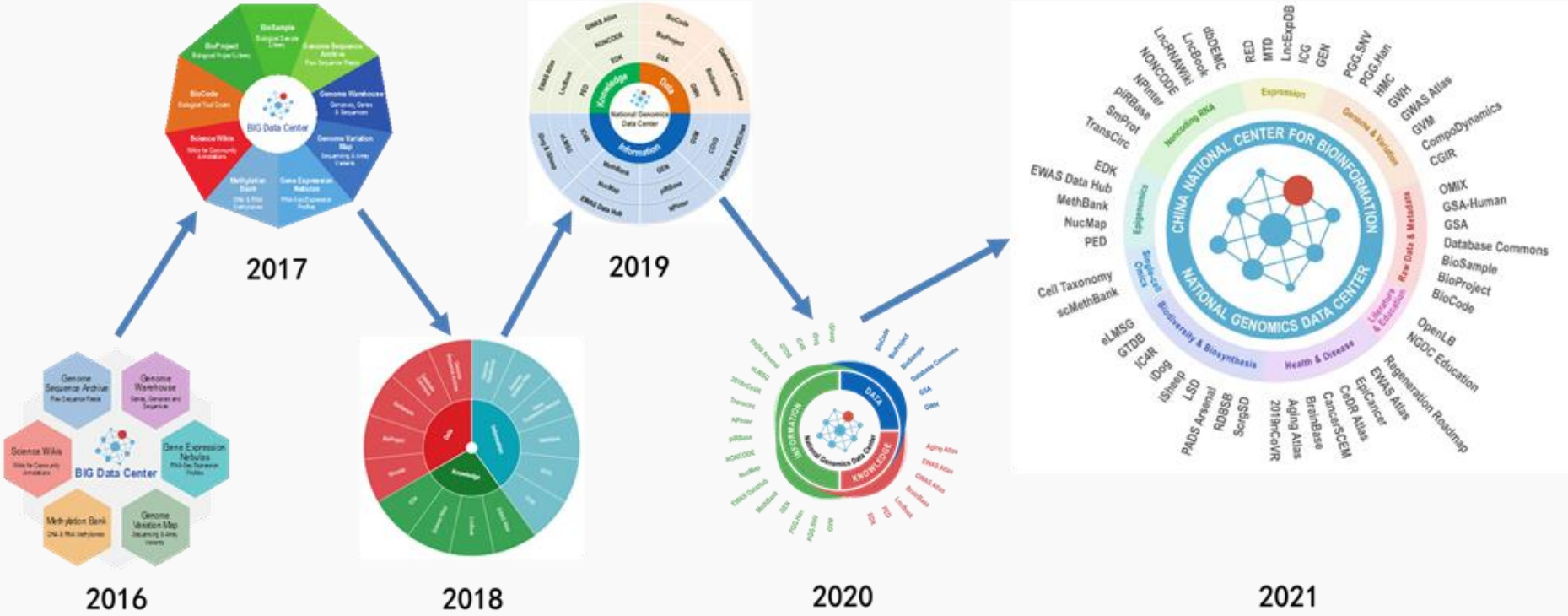
CAS-MPG Partner Institute for Computational Biology
Chinese Academy of Sciences, China

Prof. Weimin Zhu

Head of Data Science
National Center for Protein Sciences
China

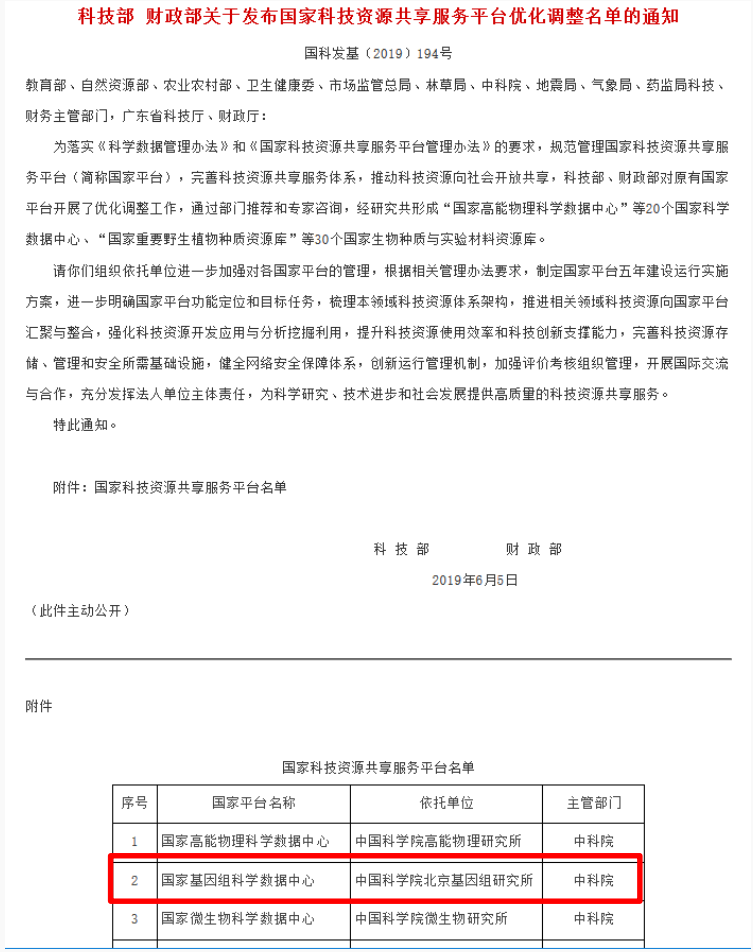
https://ngdc.cncb.ac.cn/board

The Growing of Capability



Nucleic Acids Research (2017, 2018, 2019, 2020, 2021, 2022)

National Genomics Data Center



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

（此件主动公开）

附件

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

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

National Genomics Data Center announced, 2019/06/05

China National Center for Bioinformation - CNCB

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

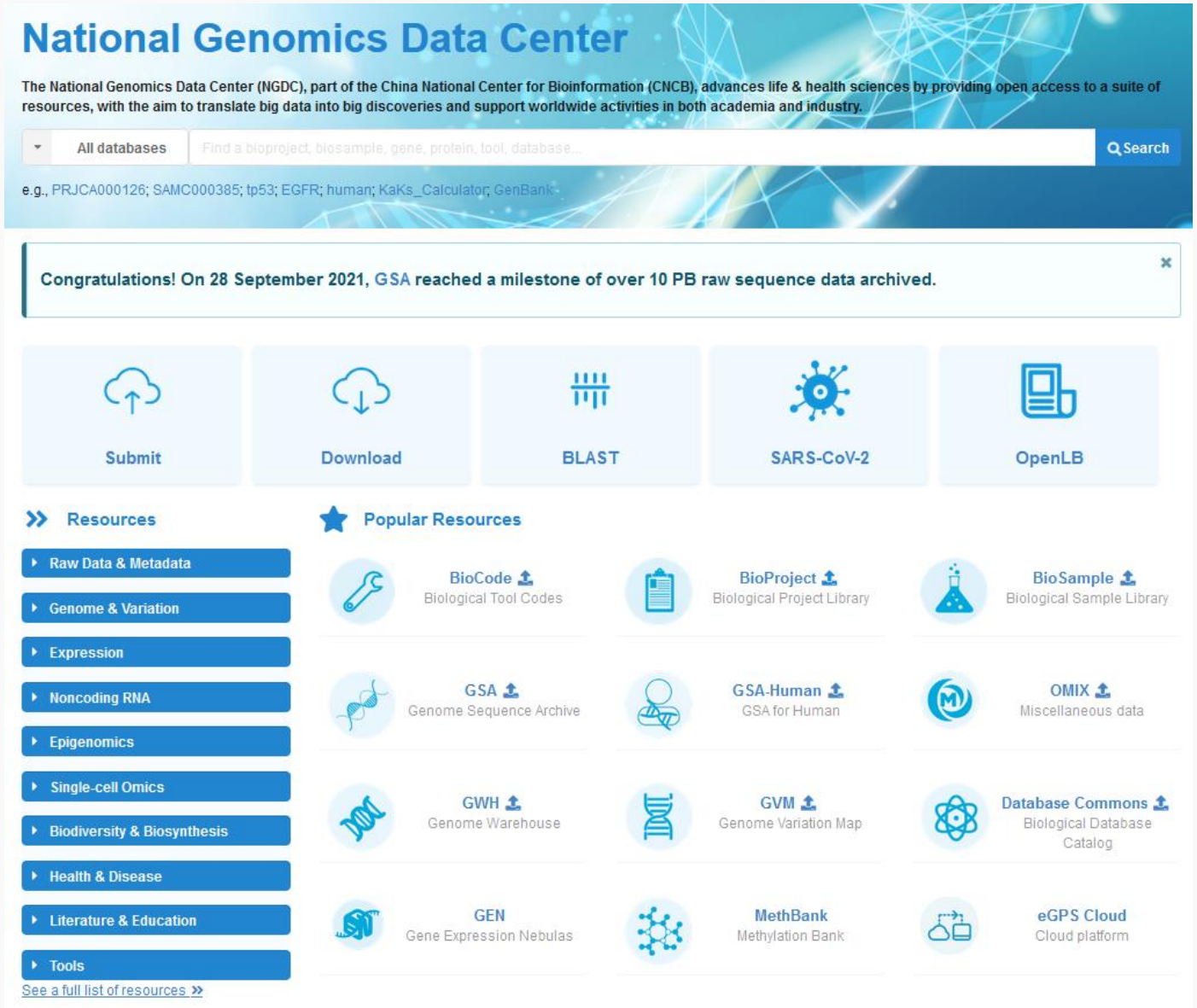
院属各单位、院机关各部门：

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。

2019年12月16日

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

NGDC Resources



GSA Genome Sequence Archive

GSA | find a GSA accession

e.g., CRA000112; CRX006656; human

中文 English

Home | Submit | Browse | Search | Statistics | Standards | Documentation | Login | Register

Genome Sequence Archive

The Genome Sequence Archive (GSA) is a data repository for archiving raw sequence reads. It accepts data submissions from all over the world and provides free access to all publicly available data for global scientific communities.

Submit
Submit data to GSA

Download
Download data to your computer

Browse
Browse publicly available records

Document
Find help information and documents

Data Statistics

New Genome Sequence Archive for Human

New 2019-nCov Raw Sequences

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 improvements.

GSA-supported Deposition

Data submissions to GSA have been reported by multiple high-profile journals. GSA has been designated as supported data repository by **Springer Nature** and **Elsevier**. GSA is one of the registered repositories in **FAIRsharing**, and is supported by **Wiley** and **Taylor & Francis**.

NSR GPB Cell Research Science Cell PNAS

more >>

Latest Released GSA

Accession	Description
CRA002283 (2021-10-08)	Genome sequencing in trachypithecus p oliocephalus

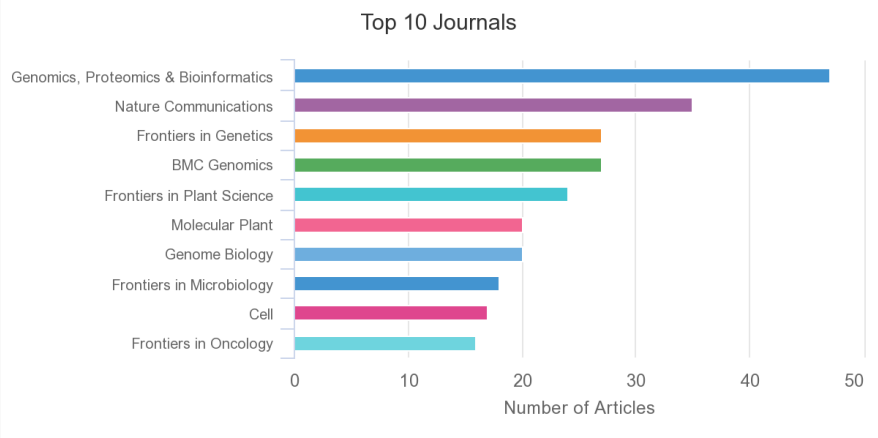
GSA Supported Major Grants

Agencies	Grants	GSAs	Experiments	Runs
Ministry of Science and Technology of the People's Republic of China (MOST)	772	268	16796	17435
National Natural Science Foundation of China (NSFC)	2756	974	51408	54658
Chinese Academy of Sciences (CAS)	525	202	12062	14394

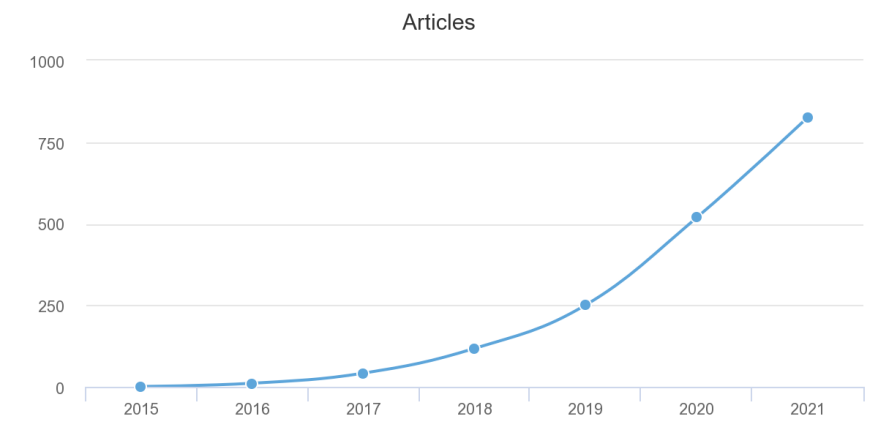
Total Grants: 4053

GSA Supported Publications

Total Journals: 275



Total Articles: 826



GSA Endorsed by Elsevier

ELSEVIER About Elsevier Products & Solutions Services Shop & Discover Search

Genes & Gene Expression

Data Repository	How articles and data are linked	More information
Allele Frequency Net Database (AFND)	Authors should specify AFND accession numbers, e.g. AFND: AFND001243	AFND homepage Submitting data
ArrayExpress	Authors should specify ArrayExpress accession numbers, e.g. ArrayExpress: E-MEXP-3783.	ArrayExpress homepage Submitting data Example article
GenBank	Authors should specify GenBank accession numbers, e.g. *GenBank: BA123456.* ScienceDirect displays and visualizes supporting information using information from and linking to the repository.	GenBank homepage Submitting data
Gene Expression Omnibus (GEO)	Authors should specify GEO accession numbers, e.g. *GEO: GSE27196; GEO: GPL5366; GEO: GSM9853.* ScienceDirect displays supporting information using information from and linking to the repository.	GEO homepage Submitting data Example article
Genome Sequence Archive	Authors should specify GSA identifiers, e.g. GSA: CRA000134	GSA homepage

GSA Endorsed by Springer Nature

SPRINGER NATURE

Search EN

Authors

Research Data Policies

- Data policy types
- Data availability statements
- Data policy FAQs
- Journal policies & services
- Recommended repositories list
- Research Data Helpdesk

Biological sciences

Imaging

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)

Genome Warehouse - GWH

Genome Warehouse

The Genome Warehouse (GWH) is a public repository housing genome-scale data for a wide range of species and delivering a series of web services for genome data submission, storage, release and sharing.

China Genomic Data Sharing Initiative

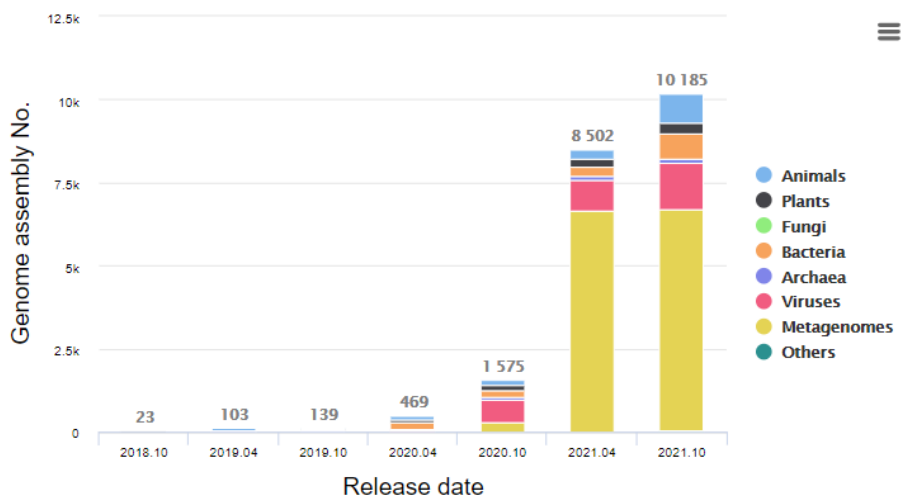
Submit
Deposit meta-information into GWH databases

Download
Transfer GWH data to your computer

Browse
View genome information about the released data

Documentation
Find help documents to learn more about GWH

Data Growth



Statistics

- **Integrated Animals**: 61 genomes, 62 assemblies
- **Integrated Plants**: 77 genomes, 88 assemblies
- **Release of Direct Submissions** (Total: 10185): 884 Animals; 316 Plants; 24 Fungi; 765 Bacteria; 104 Archaea; 1410 Viruses; 6653 Metagenomes; 29 Others
- Direct Submissions (Total: 21388): 8566 Animals; 2285 Plants; 126 Fungi; 842 Bacteria; 122 Archaea; 2287 Viruses; 6663 Metagenomes; 497 Others

GWH-supported Deposition

Data submissions to GWH have been reported by multiple journals, including:

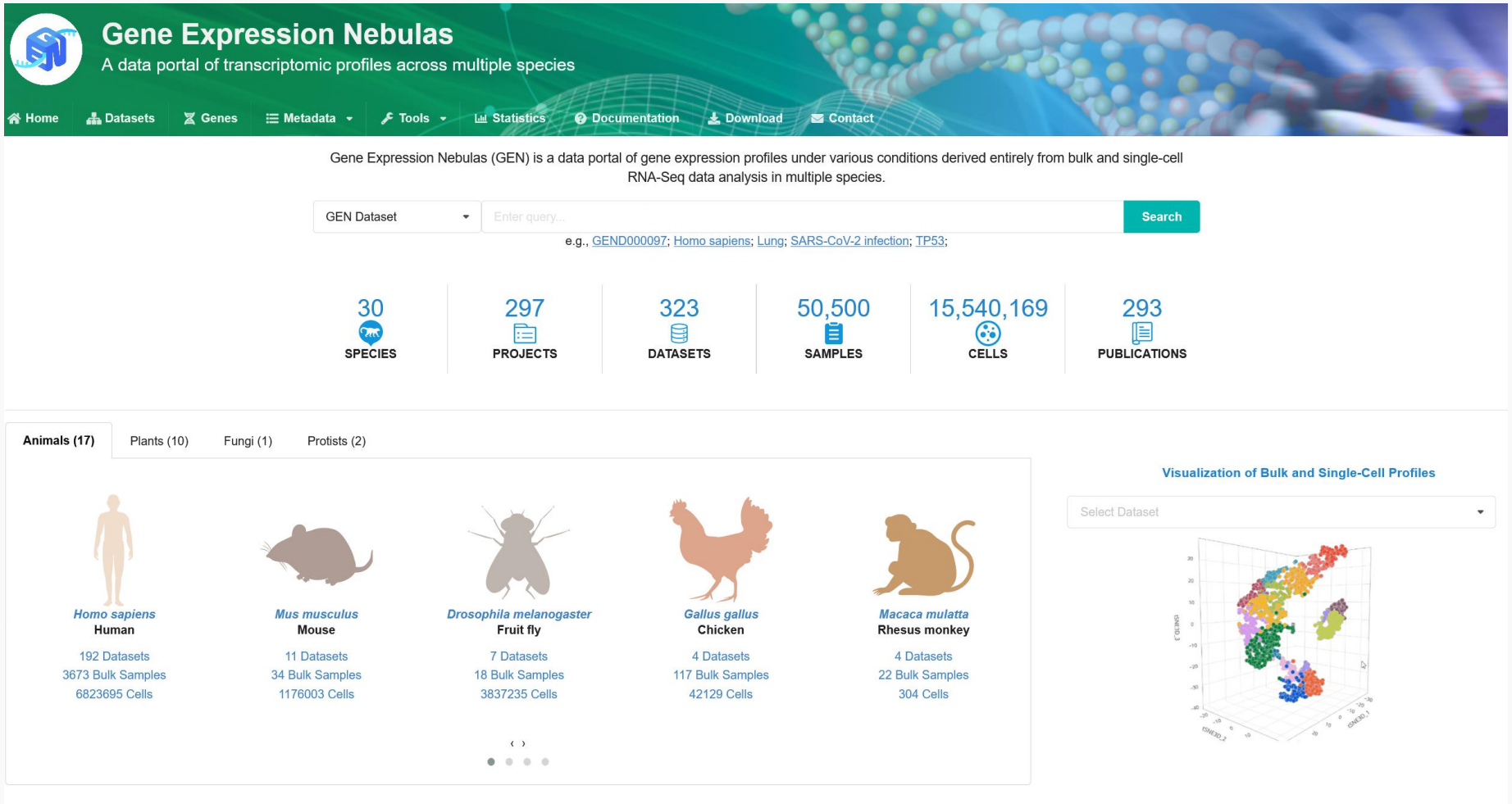
More Journals >>

Latest Released GWH

Organism	Accession
	GWHBFHH00000000

Genomics Proteomics Bioinformatics (2021)

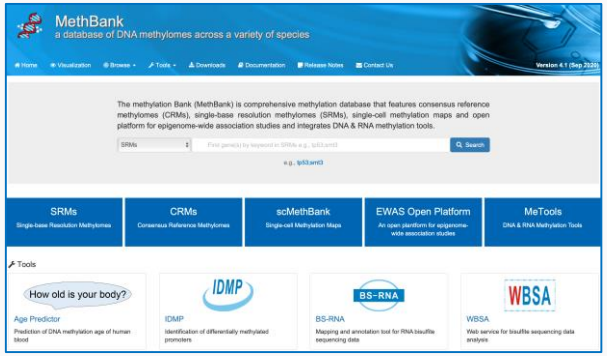
Gene Expression Nebulas - GEN



Nucleic Acids Research (2022)

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

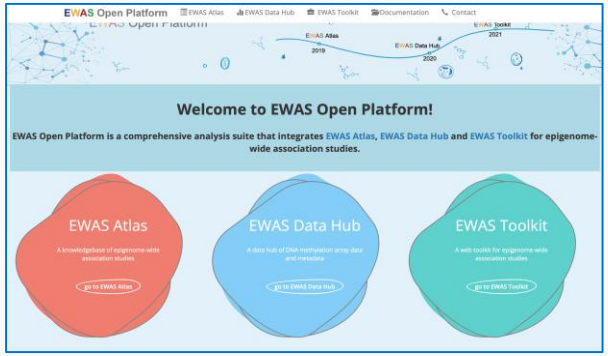
Epigenomic Resource



MethBank

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

***Nucleic Acids Research* (2018)**



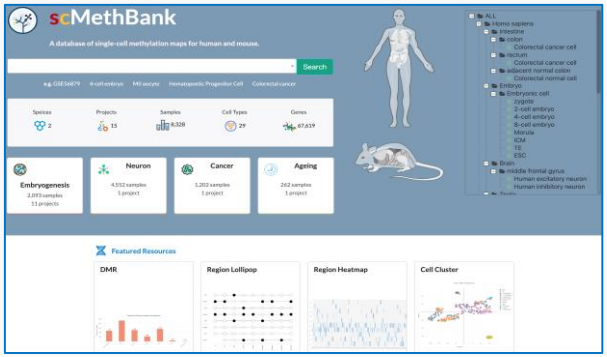
EWAS Open Plantform

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

***Nucleic Acids Research* (2019, 2020, 2021)**

Epigenomic Data
+
Knowledge

17 Species
125,260 Samples



scMethBank

https://ngdc.cncb.ac.cn/methbank/scm/

***Nucleic Acids Research* (2021)**

NucMap

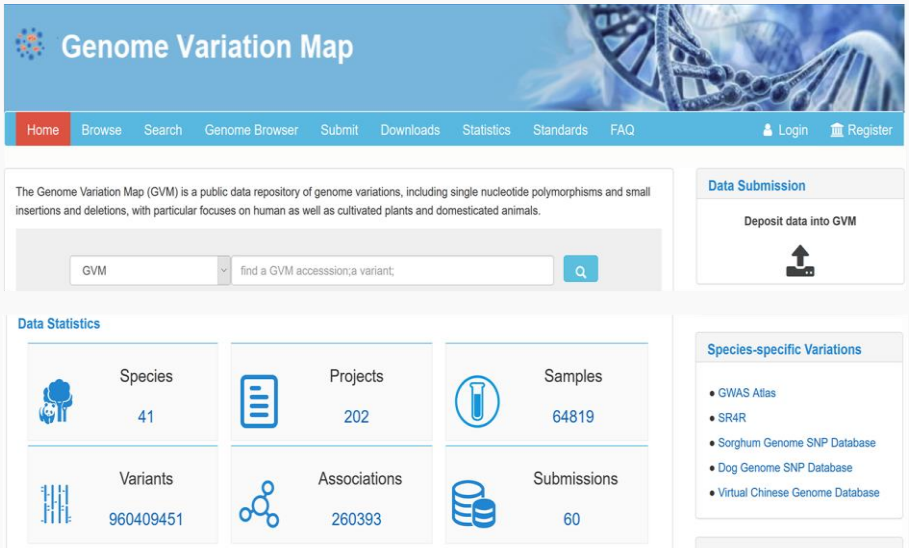
https://ngdc.cncb.ac.cn/nucmap/

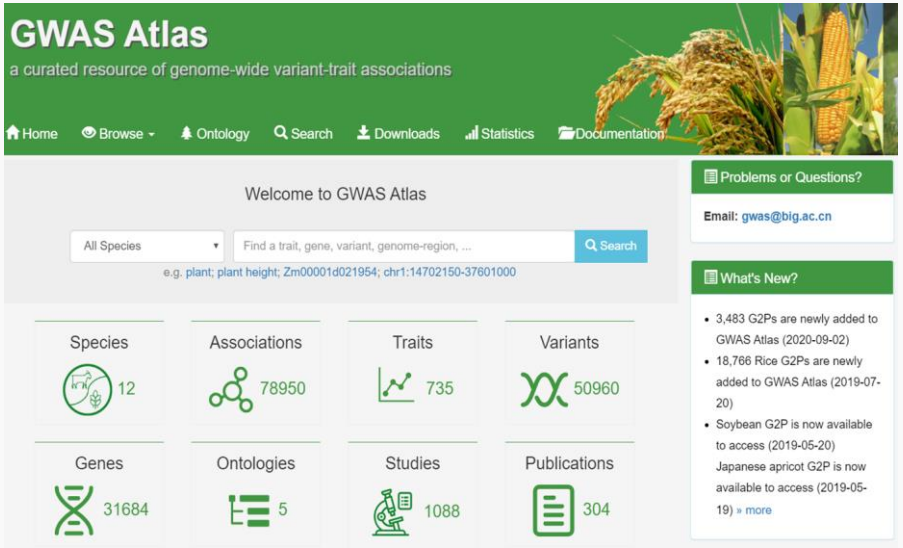
***Nucleic Acids Research* (2019)**

618 Traits
617,018 Associations

~ 260,000 PV
~ 70,000 UV

Genomic Variation Resources





Genomic Variant Map

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

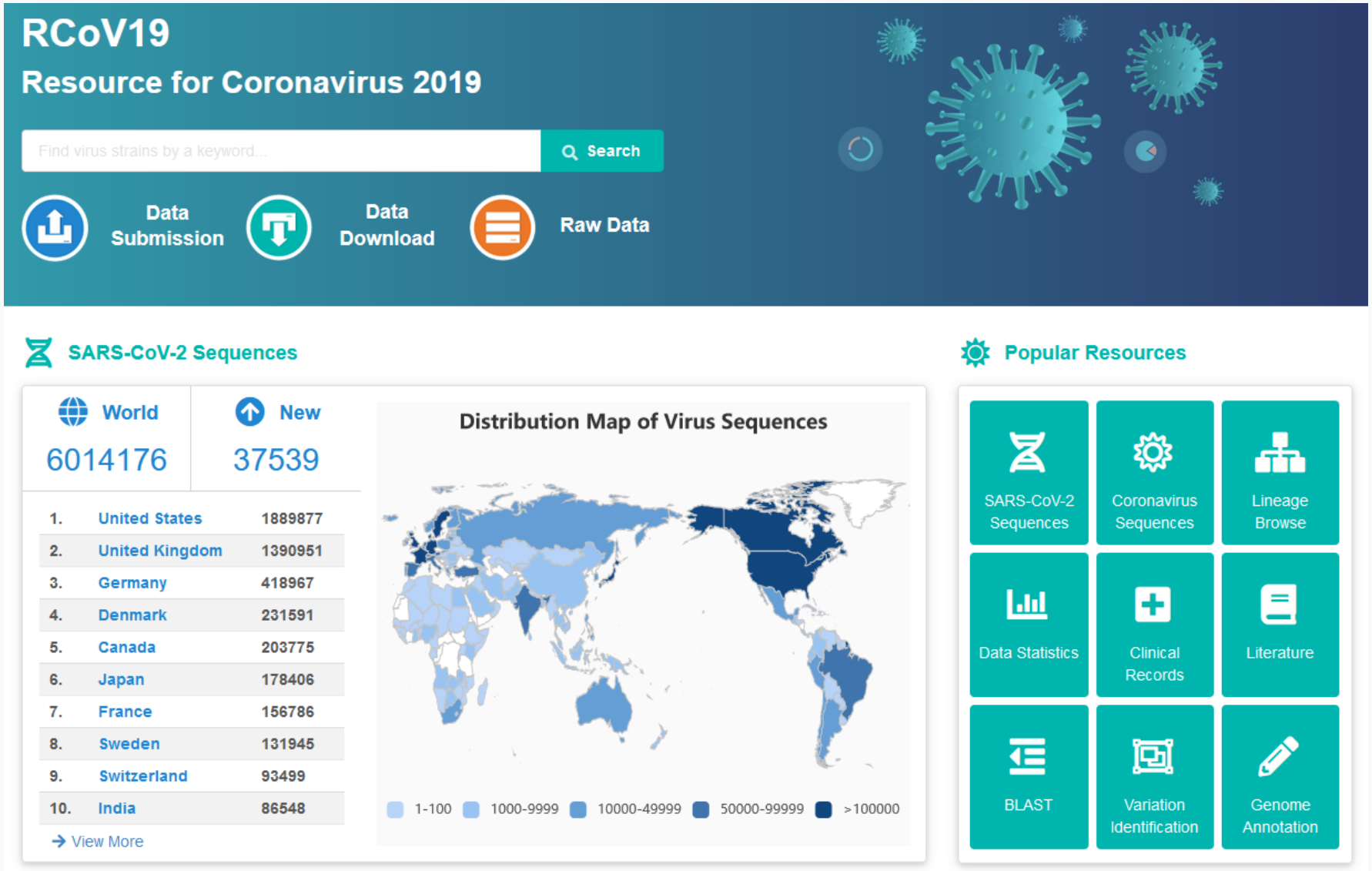
Genomic Variant Knowledge

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

41 Species, **64819** Individuals, **960** million Variants, **260,393** Genotype to Phenotype, **22,736** visits & **182** submission

Nucleic Acids Research (2018, 2020, 2021)

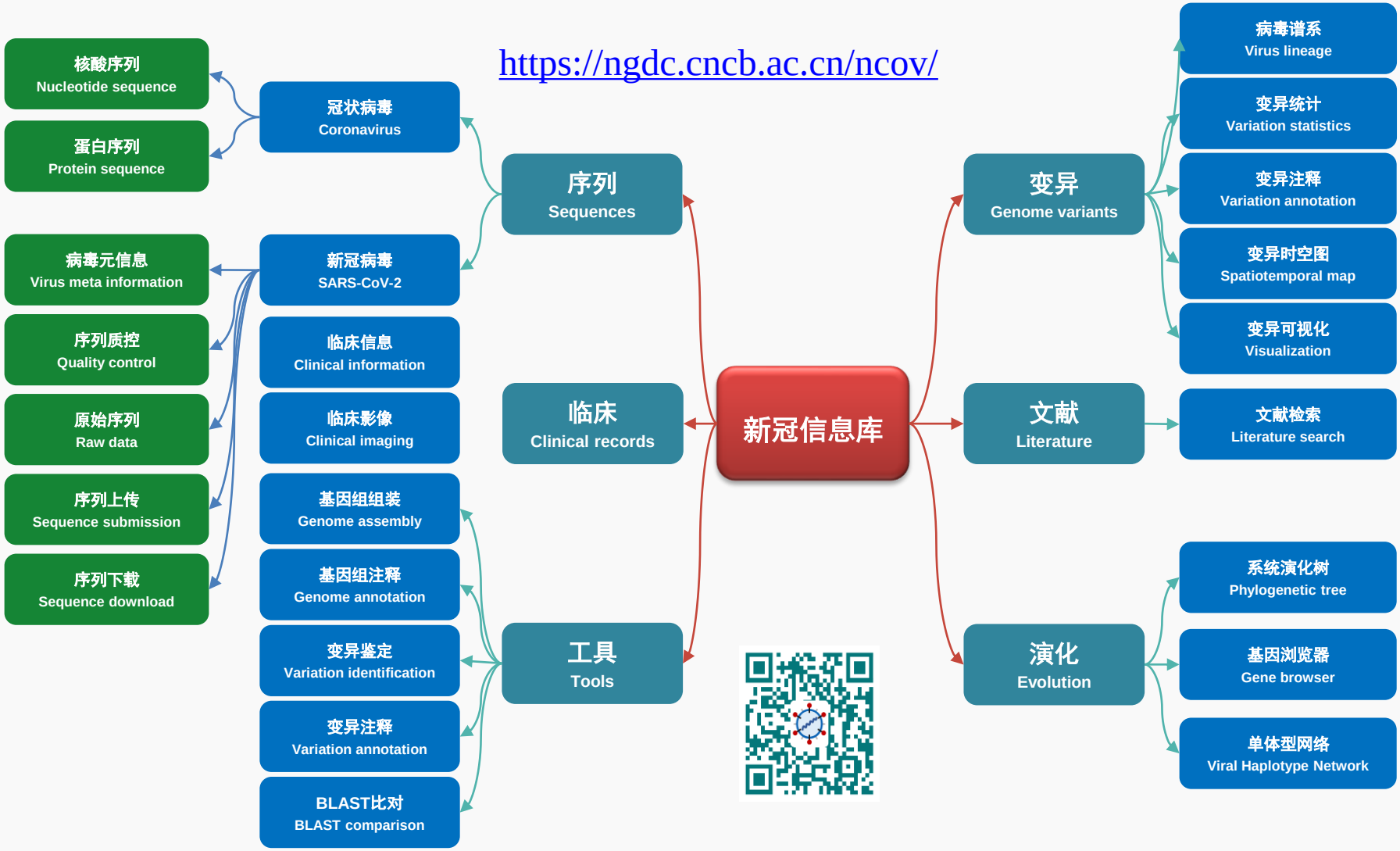
RCoV19


RCoV19
Resource for Coronavirus 2019
Find virus strains by a keyword...
Search
Data Submission
Data Download
Raw Data
SARS-CoV-2 Sequences
World
6014176
New
37539
1. United States 1889877
2. United Kingdom 1390951
3. Germany 418967
4. Denmark 231591
5. Canada 203775
6. Japan 178406
7. France 156786
8. Sweden 131945
9. Switzerland 93499
10. India 86548
View More
Distribution Map of Virus Sequences
1-100
1000-9999
10000-49999
50000-99999
>100000
Popular Resources
SARS-CoV-2 Sequences
Coronavirus Sequences
Lineage Browse
Data Statistics
Clinical Records
Literature
BLAST
Variation Identification
Genome Annotation

Sequence Growth

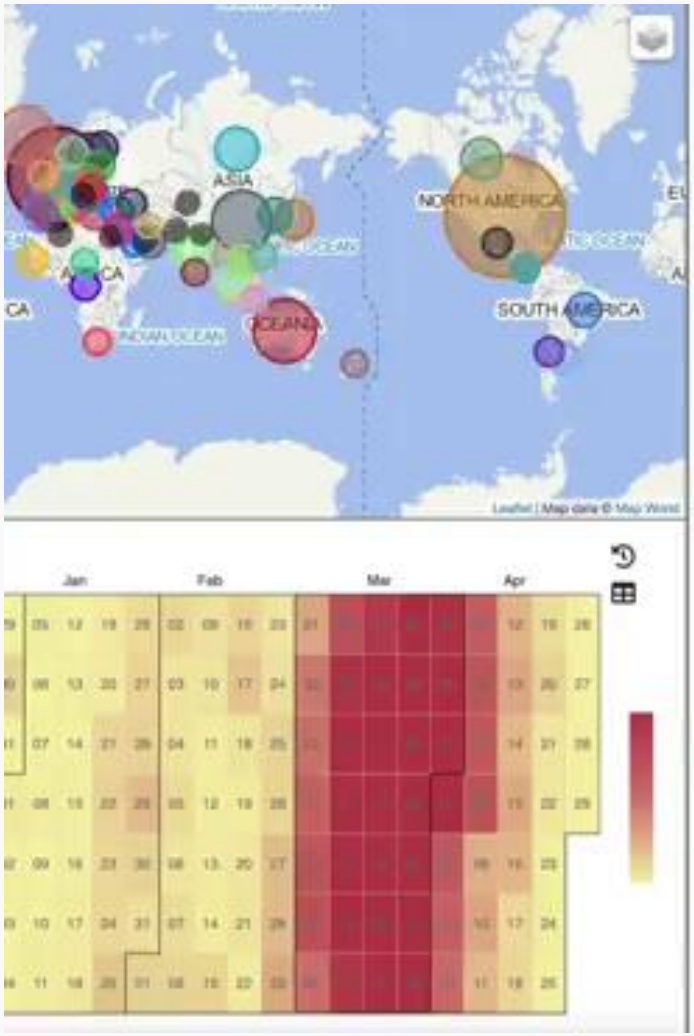
RCoV19 Modules

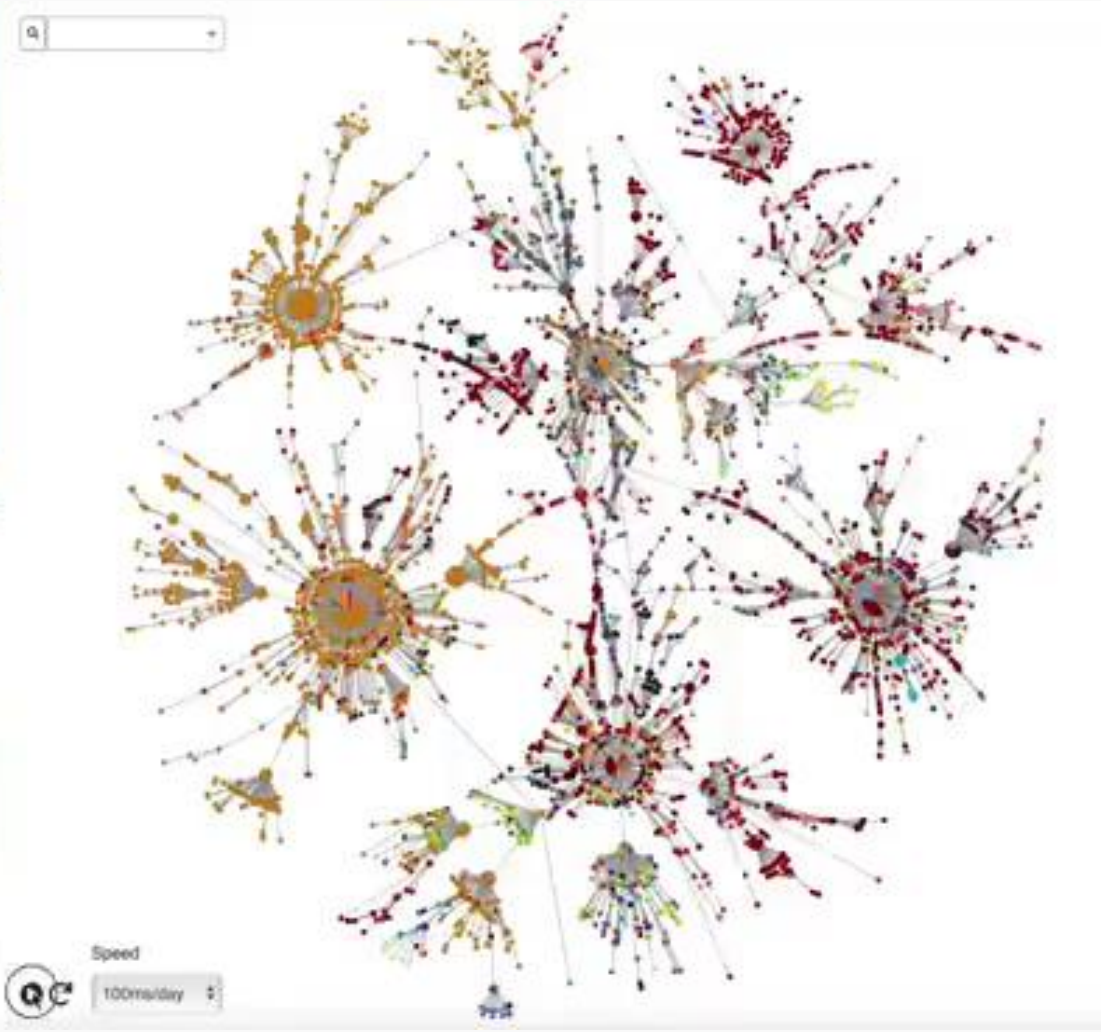
https://ngdc.cncb.ac.cn/ncov/



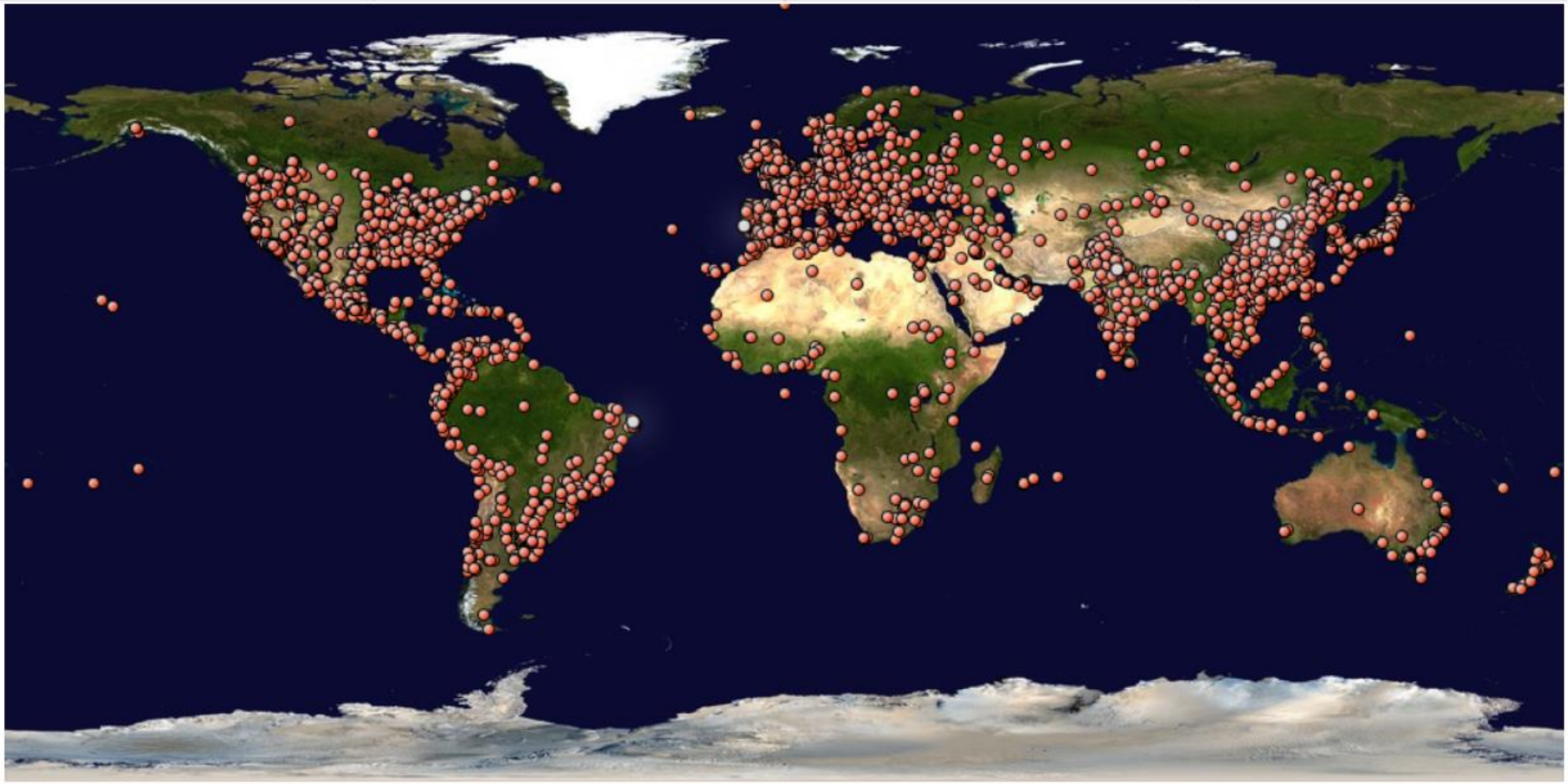
YiChuan 2020, *GPB* 2020, *ZR* (2020)

Haplotype Network Map



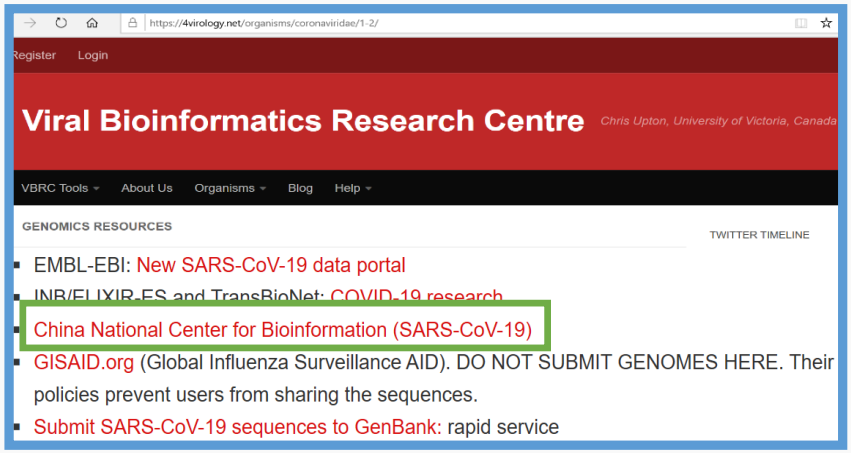


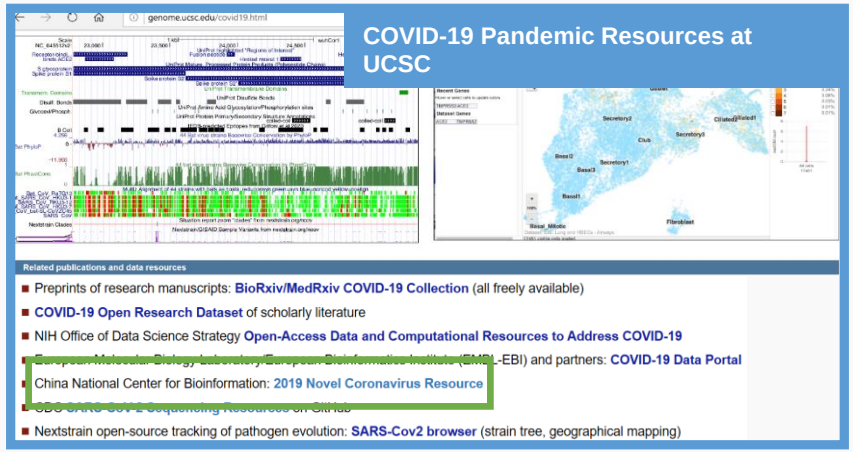
Users across the world

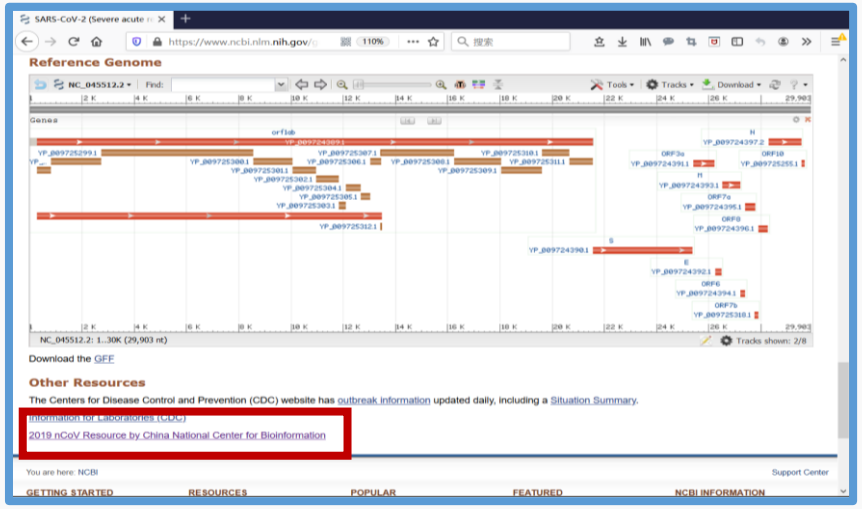


Countries/regions: 179; Users: >1.3 million; Data downloads: 1.9 billion

Referred by other resources







Open Archive for Miscellaneous Data (OMIX)

OMix

Quickly upload your data via OMix.

- ✓ transcriptome, epigenome, microarray
- ✓ lipidome, metabolome, proteome
- ✓ clinical information, demographic data, questionnaire *etc*.

Serving as a critical complement to GSA/GSA-Human, OMIX is an open archive for miscellaneous data.

Released Data : 142 records

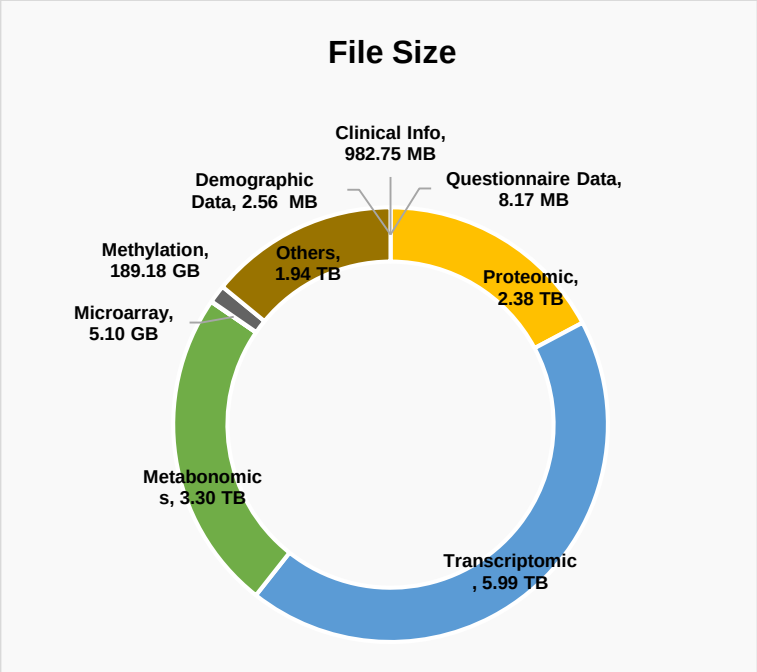
Show 10 entries

OMIX ID	Project	Title	Organism	Access Type	Release Time	Operation
OMIX691	PRJCA006797	scRNAseq data of anatomically distinct fibroblast subsets deter mine skin autoimmune disease patterns	Homo sapiens	Controlled	2021-10-07	
OMIX567	PRJCA006458	mNGS in the diagnosis of HHV 1 reactivation in a critically ill C OVID-19 patient	Severe acute respiratory syndrome coronavirus 2	Open	2021-10-06	
OMIX665	PRJCA006637	Database of O-GlcNAC modifie d proteins in human placental tr ophoblast cell line	Homo sapiens	Open	2021-09-25	
OMIX663	PRJCA005262	Interference PTPRZ1 expressio n profiling chip in glioma stem cells	Mus	Open	2021-09-24	
OMIX661	PRJCA006391	The mechanism of embryonic d iseases	Mus musculus	Open	2021-09-23	
OMIX561	PRJCA006378	Copy number variation in Chine se children with complete atriov entricular canal and single vent ricle	Homo sapiens	Open	2021-09-22	
OMIX655	PRJCA006631	Decipher unwinding mechanis ms of RNA helicases MOV10 an d MOV10L1	Mus	Open	2021-09-19	
OMIX656	PRJCA006631	Proximal single-stranded RNA d estabilizes human telomerase R NA G-quadruplex and induces it s distinct conformers	Homo sapiens	Open	2021-09-19	
OMIX657	PRJCA006631	Discover a new hRPA-activated BLM unwinding mode.	Gallus gallus	Open	2021-09-19	
OMIX658	PRJCA006631	Identify critical interaction sites between SpCas9 and DNA.	Streptococcus pyogenes	Open	2021-09-19	

Showing 1 to 10 of 142 entries

Previous 1 2 3 4 5 ... 15 Next

Date File: 869; Total File Size: 13.79 TB



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

Open Library of Bioscience(OpenLB)

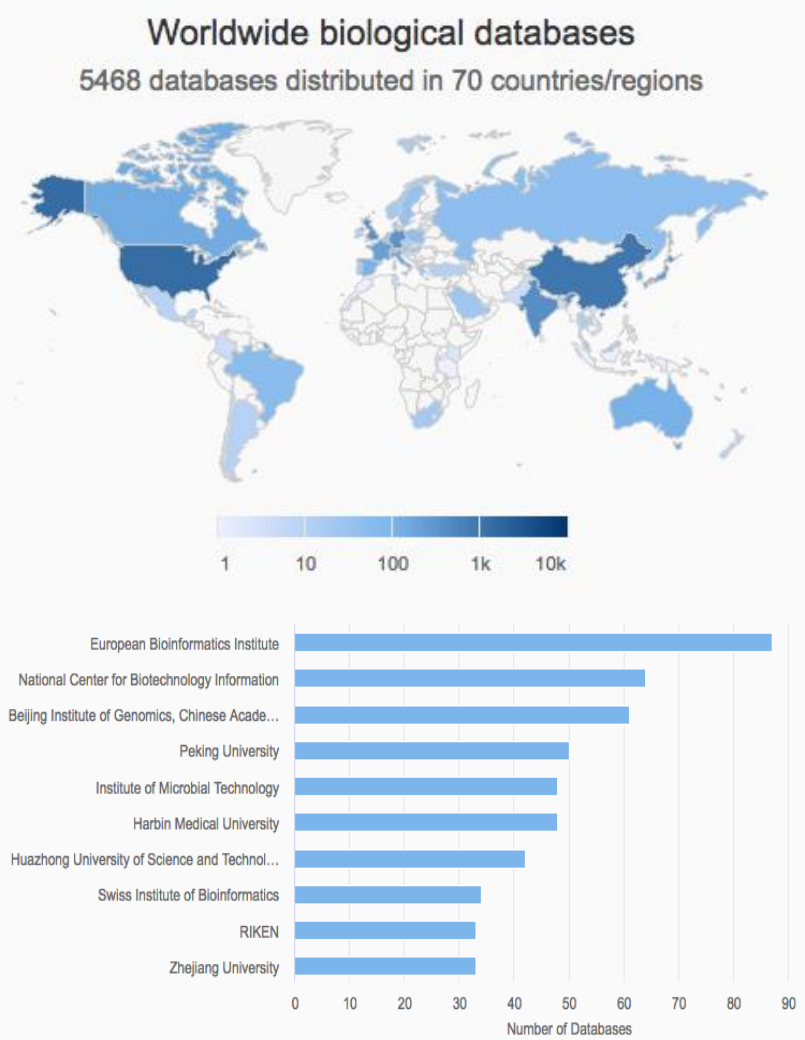
A biological literature library links to relevant resources in CNCB-NGDC based on open access and open data principles.

Abstract text: ~33 millions

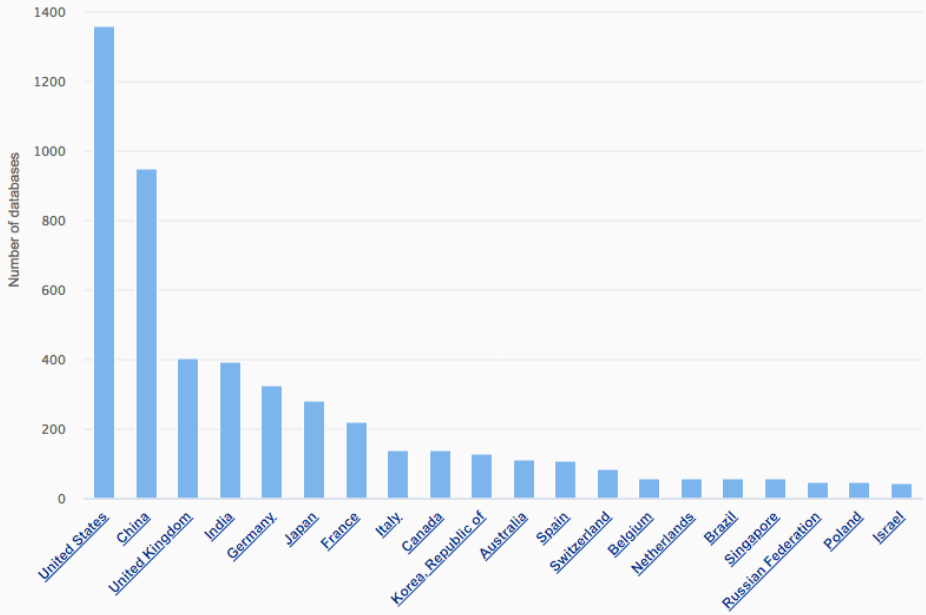
Source: [NCBI PubMed](), [bioRxiv](), [medRxiv]()

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

Database Commons



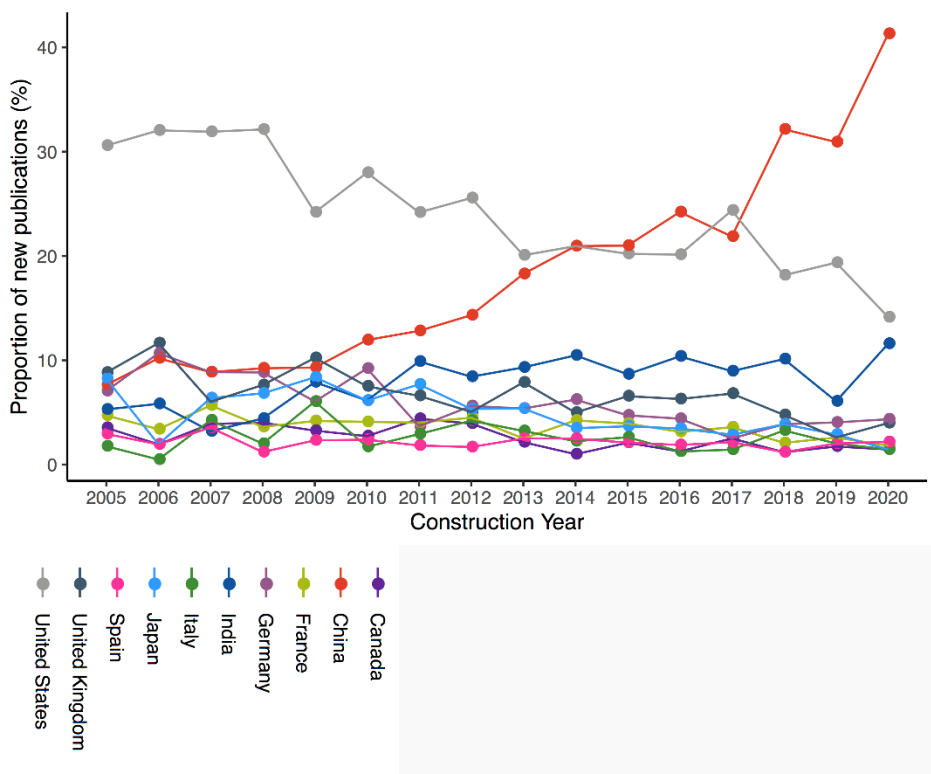
5468	70	1291
DATABASES	COUNTRIES / REGIONS	ORGANISMS
8138	2098	13
PUBLICATIONS	INSTITUTIONS	CATEGORIES



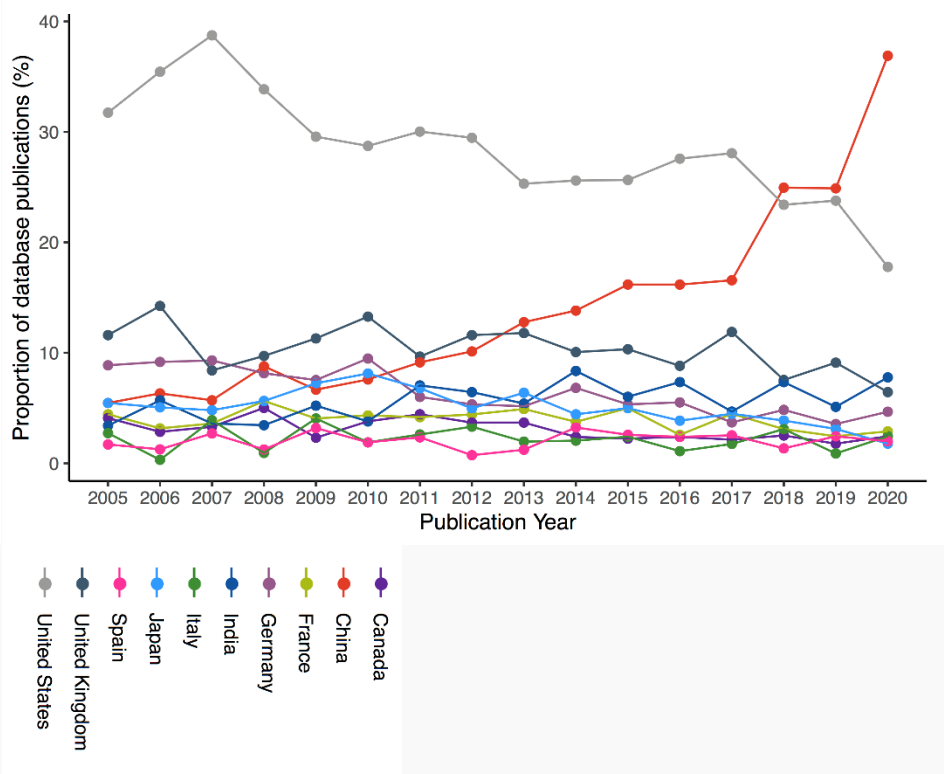
https://ngdc.cncb.ac.cn/databasecommons/

Database Commons

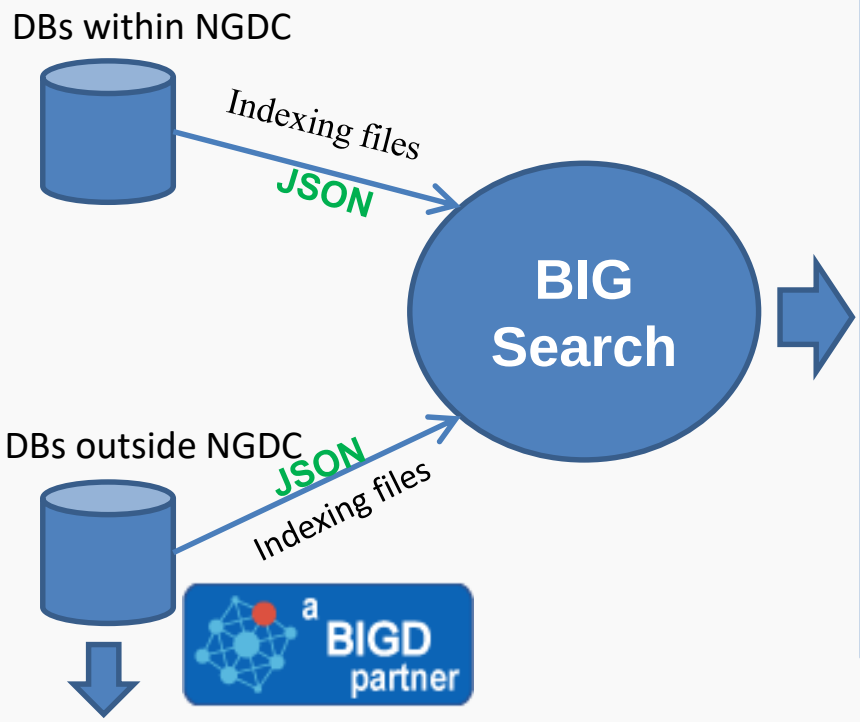
Trends in new database construction

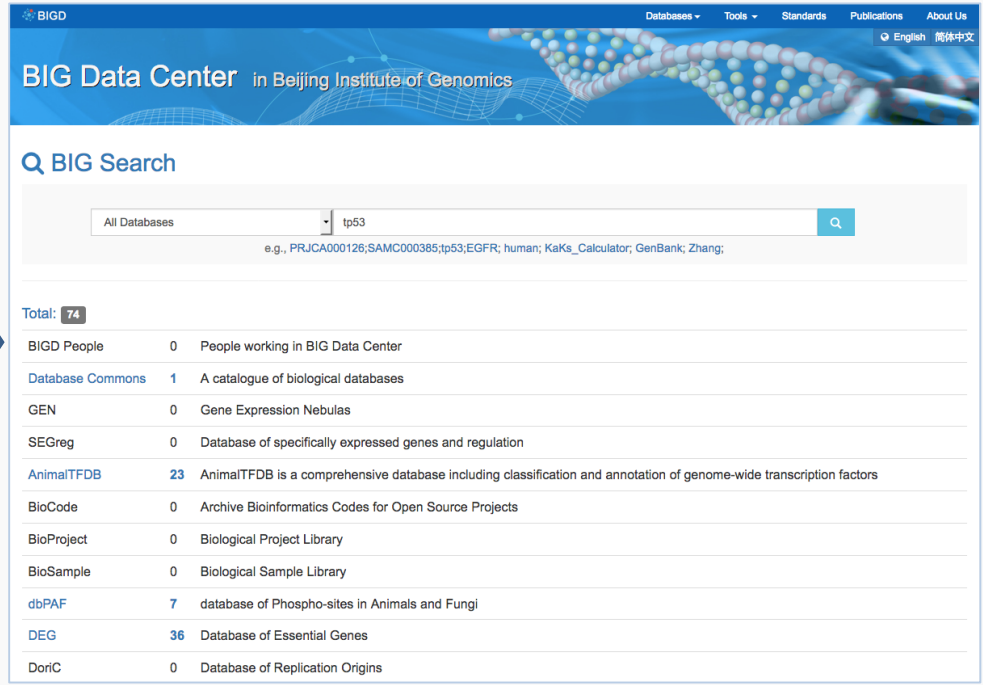


Trends in database publication



BIG Search：Cross-database search engine





Partner Databases

AnimalTFDB2.0 · dbPSP · dbPPT · EKPD · CGDB · WERAM · DEG · DoriC · MiCroKiTS · hTFtarget · dbPAF · THANATOS · PceRBase · PlantTFDB · RhesusBase · FLMD · iDUCD · SEGreg · The LncRNA and Disease Database · lncRNASNP2 · PceRBase

BIG Search：NGDC & partners

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.

All Databases | human | Search

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

NGDC & Partners Databases | EBI Databases | NCBI Databases

25,626,940 records from 45 NGDC & Partner databases.

Show 10 entries

Filter:

Database	Records Number	Description
lnCAR	28,420	lnCAR | A comprehensive resource for lncRNAs from Cancer Arrays
LncBook	409,204	A curated knowledgebase of human long non-coding RNAs.
LncExpDB	101,293	Expression Database of Human Long non-coding RNAs
lncRNASNP2	4,443,771	
Methbank CRMs	60,415	Methbank, Consensus Reference Methylomes (CRMs)
MethBank SRMs	60,479	Methbank, Single-base Resolution Methylomes (SRMs)
NODE	31	The National Omics Data Encyclopedia
OMix	1	OMix

BIG Search：EBI databases

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.

All Databases | human | Search

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

NGDC & Partners Databases | **EBI Databases** | NCBI Databases

82,248,602 records from 112 EBI databases.

Show 10 entries Filter:

ArrayExpress	38,693	ArrayExpress Archive is a MIAME compliant public database for microarray data.
Assembly	42,925	Genome Assembly
Assembly contig set	28,927	European Nucleotide Archive(Whole Genome Shotgun Set)
Baseline Expression Atlas Genes	776	Large scale meta-analysis of public transcriptomics data
bio.tools	799	Bioinformatics Tools and Services Discovery Portal
BioModels	756	Database of Mathematical models of biological interest
BioModels	21	Biomodels Autogenerated

BIG Search：NCBI databases

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.

All Databases | human | Search

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

NGDC & Partners Databases | EBI Databases | **NCBI Databases**

935,941,769 records from 35 NCBI databases.

Show 10 entries

Filter:

Assembly	526	Functional categorization of proteins by domain architecture
BioAssays	541,056	Bioactivity screening studies
BioCollections	6	Museum, herbaria, and other biorepository collections
BioProject	82,938	Biological projects providing data to NCBI
BioSample	6,579,858	Descriptions of biological source materials
Bookshelf	162,404	Books and reports
ClinVar	843,722	Human variations of clinical significance

BIG Search: NGDC+EBI+NCBI+AlphaFold

CNCB NGDC

Databases Tools Standards Publications About

国家基因组科学数据中心
National Genomics Data Center

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

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.

All Databases | human | Search

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	20309	AlphaFold Protein Structure Database

Powered by EBI AlphaFold DB

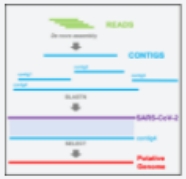
Online Tools

Genome Tracing

'Genome tracing' can be used to screen for the closest relatives of the query sequences in the SARS-CoV-2 database and display their spatiotemporal distributions.

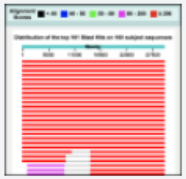
Denovo Assembly

'Denovo Assembly' can be used to assemble NGS sequencing reads,and identify SARS-CoV-2 genome sequences from the assembly contigs.



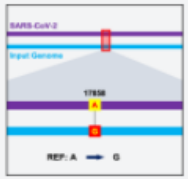
BLAST

Using 'BLAST' tool, users can perform sequence alignments with coronavirus genome database, SARS-CoV-2 reference and genome database.



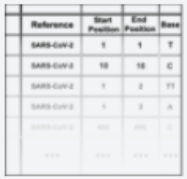
Genome-to-Variants

'Genome-to-Variants' can align submitted genome sequences to the SARS-CoV-2 genome and identify SNPs and indels.



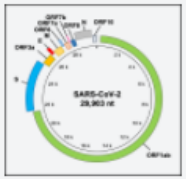
Variant Annotation

'Variant Annotation' can perform functional annotation for the variants, and show the information of genes, code and amino acid changes.



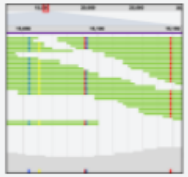
Genome Annotation

'Genome Annotation' supports accurate gene annotations for submitted SARS-CoV-2 genome sequences.



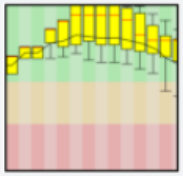
Fastq-to-Variants

'Fastq-to-Variants' can align NGS sequencing data to the SARS-CoV-2 genome, then identify and annotate SNPs and indels.



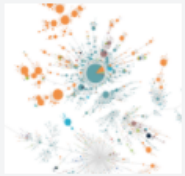
SeqQC

SeqQC can evaluate the sequencing quality of the uploaded FASTQ or BAM files using FastQC and MultiQC.



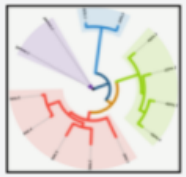
Haplotype Network

The haplotype network will be constructed via Haplotype network construction algorithm based on minimum-cost arborescence.



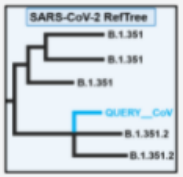
Phylogenetic Tree

Build Phylogenetic Tree using maximum likelihood method by IQ-TREE or RAxML for the submitted sequences.



Lineage & Phylogenetic

The Pango Lineage Assigner supports Pango lineage assigned for submitted SARS-CoV-2 genome sequences based on the software Pangolin.



Online Tools — BLAST

BLAST
Basic Local Alignment Search Tool

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

References:
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.

Submit a job
My BLAST jobs
BLAST results

blastn | blastp | blastx | tblastn

Query

Input **FASTA** sequence(s): Example Clear

Maximum of 20 sequences (type in plain text or FASTA format)

Upload a file Choose query file

Target

Align sequences in database Align your input sequences

Database:

SARS-CoV-2 genome database

Task selection

Optimize for: Highly similar sequences (megablast)

Submit your job

Set job title Notify by email

Submit or Clear

Set your parameters

Database:

Genome Warehouse (GWH) transcript sequences
LncBook human lncRNA sequences
IC4R rice transcript sequences
NCBI Nucleotide Collection (nt)
Coronaviridae genome database
SARS-CoV-2 genome database
SARS-CoV-2 PANGO lineage genomes
Sorghum nucleotides
Protists P10K genomes

nucleotide

Genome Warehouse (GWH) protein sequences
IC4R rice protein sequences
NCBI non-redundant protein sequences (nr)
Sorghum proteins

protein

International Collaboration

- SARS-CoV-2 genomes sequenced and analyzed for 150 samples from Pakistan, with 350 more samples newly received
- BRICS grant awarded for SARS-CoV-2 genome sequencing and analyses (with Brazil, Russia, India & South Africa)
- CAS-NSTDA (Thailand) research grant applied

Genomic Epidemiology of SARS-CoV-2 in Pakistan

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 Shreen[7], Fatima Batool[7], Bing Zhang[8], Hua Chen[1,4,5,8], Yungui Yang[1,4,5], Amir Ali Abbasi[7,*], Mingkun Li[1,4,5,9,*], Yongbiao Xue[1,4,9,*], Yiming Bao[1,2,3,4,*]

[1] China National Center for Bioinformation, Beijing 100101, China
[2] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[3] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[4] University of Chinese Academy of Sciences, Beijing 100049, China
[5] CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[6] Department of Virology and Immunology, National Institute of Health, Islamabad 45500, Pakistan
[7] National Center for Bioinformatics, Programme of Comparative and Evolutionary Genomics, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan
[8] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China
[9] State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, The Innovation Academy of Seed Design, Chinese Academy of Sciences, Beijing 100101, China

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

CAS - NSTDA
Joint Research Project (2021)

Research Area

☑ Life Sciences
☐ Material Sciences

Note: You can only choose ONE (1) research area in which the proposal will be reviewed.

Title of Cooperative Research Project

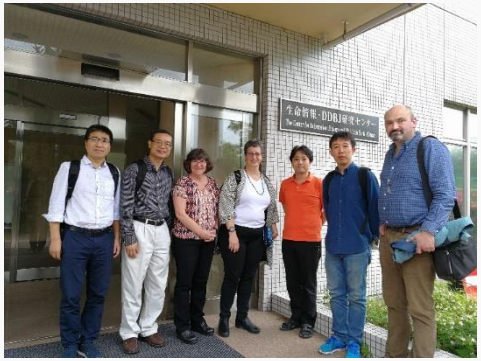
Understanding the miRNA regulation mechanism in naturally TB infected cynomolgus macaques (*Macaca fascicularis*)

Chinese Research Leader

Name (First) Yiming (Family) Bao
Organization Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation
Division /Department National Genomics Data Center Title Director & Prof.

Collaboration with INSDC

DDBJ

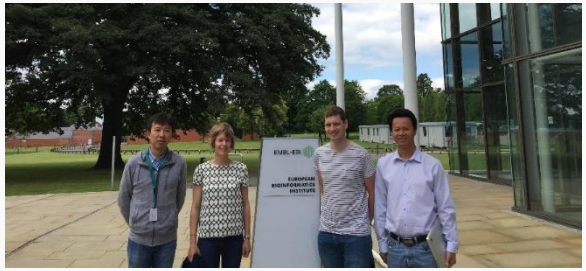


May, 2017, INSDC Meeting

NCBI

Sept, 2017 & 2018, visit & training

EBI



May, 2019 INSDC meeting

Data Sharing with INSDC

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).
```

Challenges

- Stable funding
- Rapid growth of data - storage
- Data security and backup
- CNCB development

Acknowledgements

Funding

- National Programs for High Technology Research and Development
- National Key Research Program of China
- Strategic Priority Research Program of the Chinese Academy of Sciences
- Key Program of the Chinese Academy of Sciences
- IUBS
- ANSO

NGDC Members

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