

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)



- 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

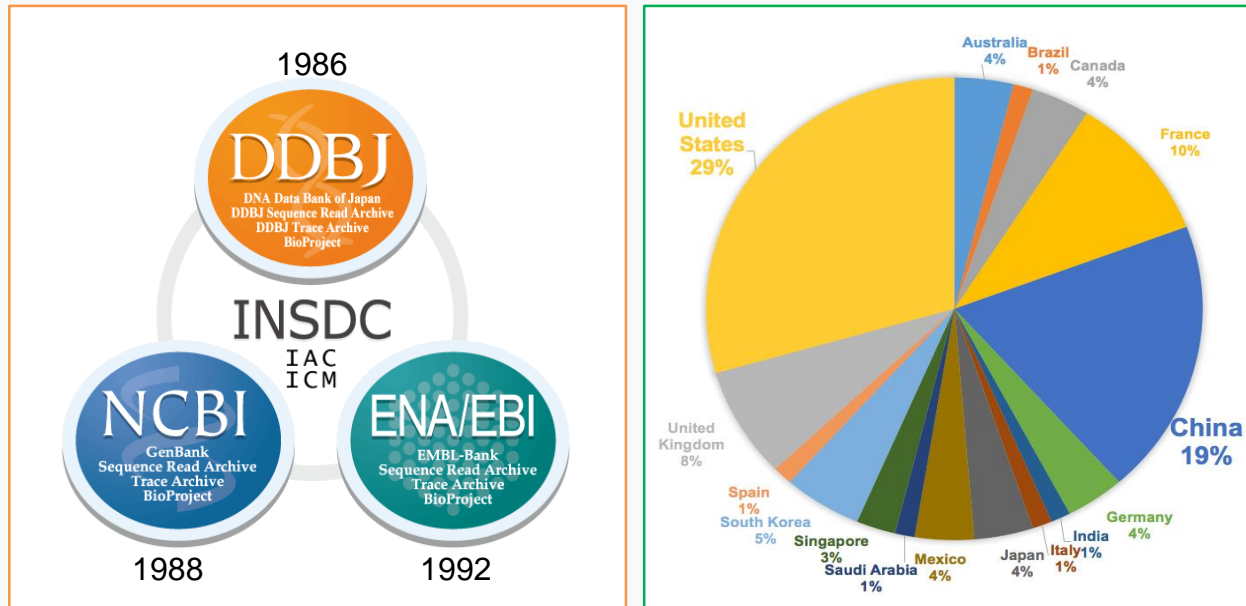


Background in China

- **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



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 Mizrahi

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



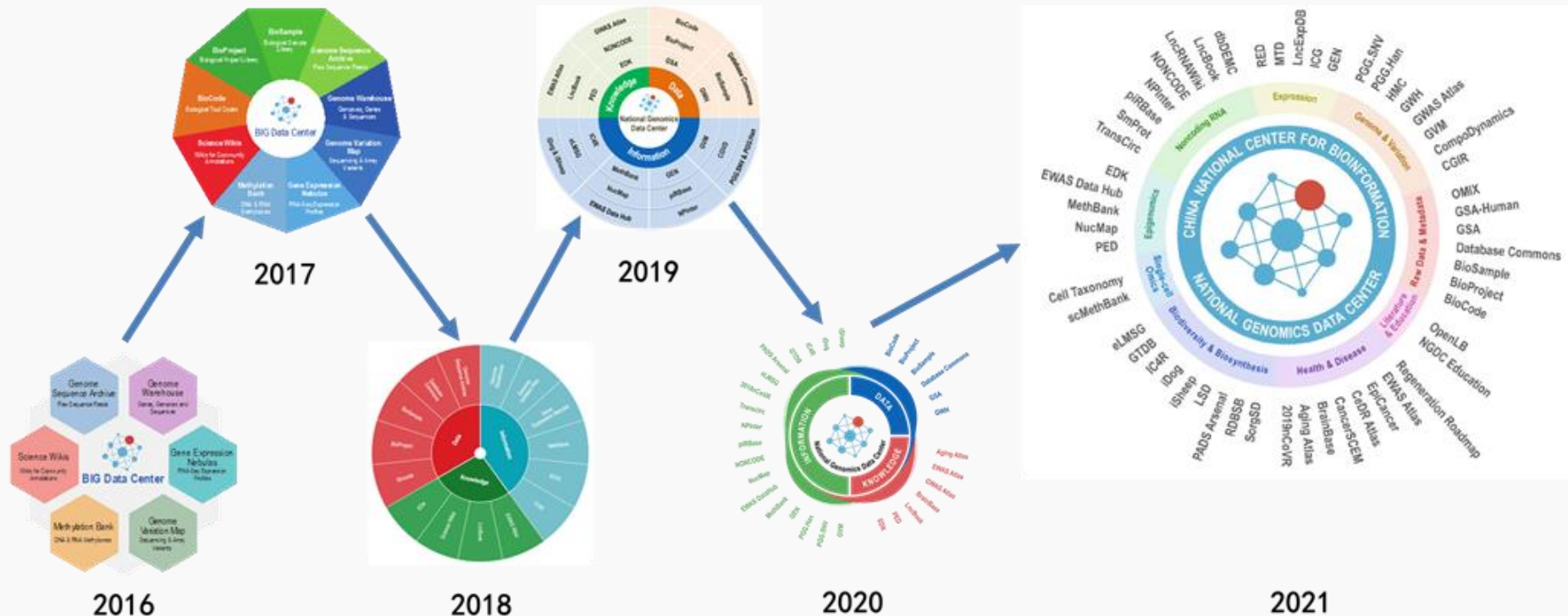
Prof. Weimin Zhu

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

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



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

National Genomics Data Center
announced, 2019/06/05



China National Center for Bioinformation

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

China National Center for Bioinformation - CNCB

中国科学院文件

科发入字〔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。



- 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

National Genomics Data Center

The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB), 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... Q Search

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

Congratulations! On 28 September 2021, GSA reached a milestone of over 10 PB raw sequence data archived.



Submit



Download



BLAST



SARS-CoV-2



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



eGPS Cloud
Cloud platform



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

New [Genome Sequence Archive for Human](#)

New [2019-nCov Raw Sequences](#)

Help & Support

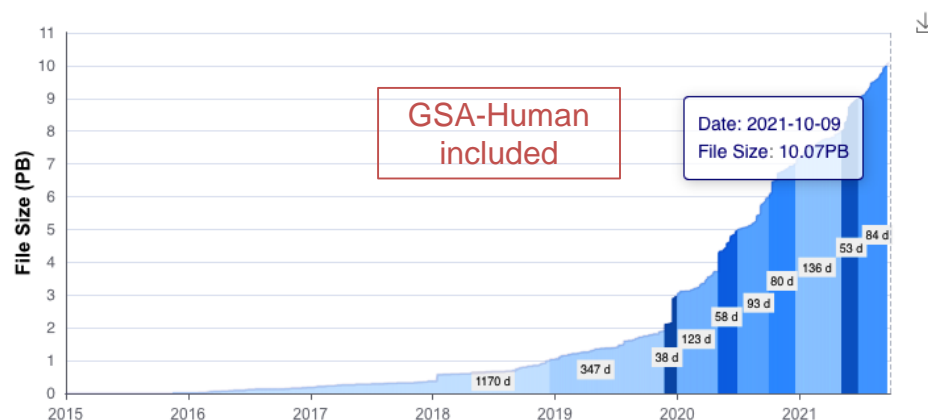
If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

Email: gsa@big.ac.cn

QQ group: [548170081](#)

We highly appreciate your comments and suggestions for further improvements.

Data Statistics



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
National Science Review

GPB
Cell

Cell Research
PNAS

[more >>](#)

Latest Released GSA

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



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

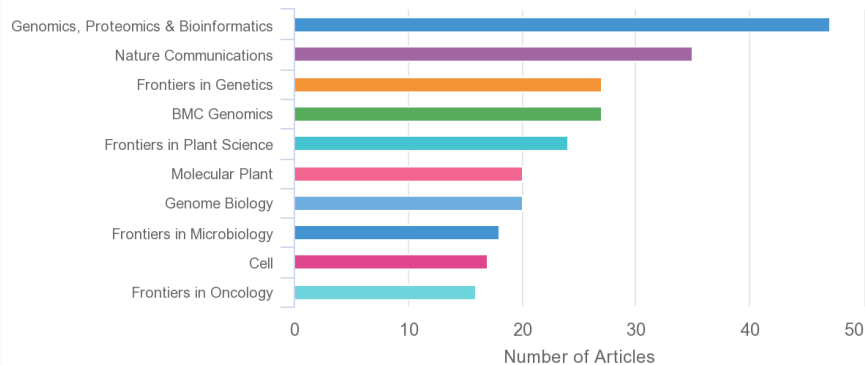


GSA Supported Publications

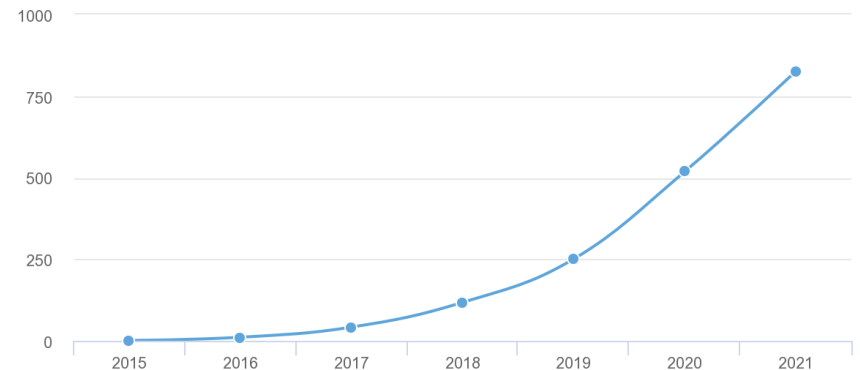
Total Journals: **275**

Total Articles: **826**




Top 10 Journals



Articles

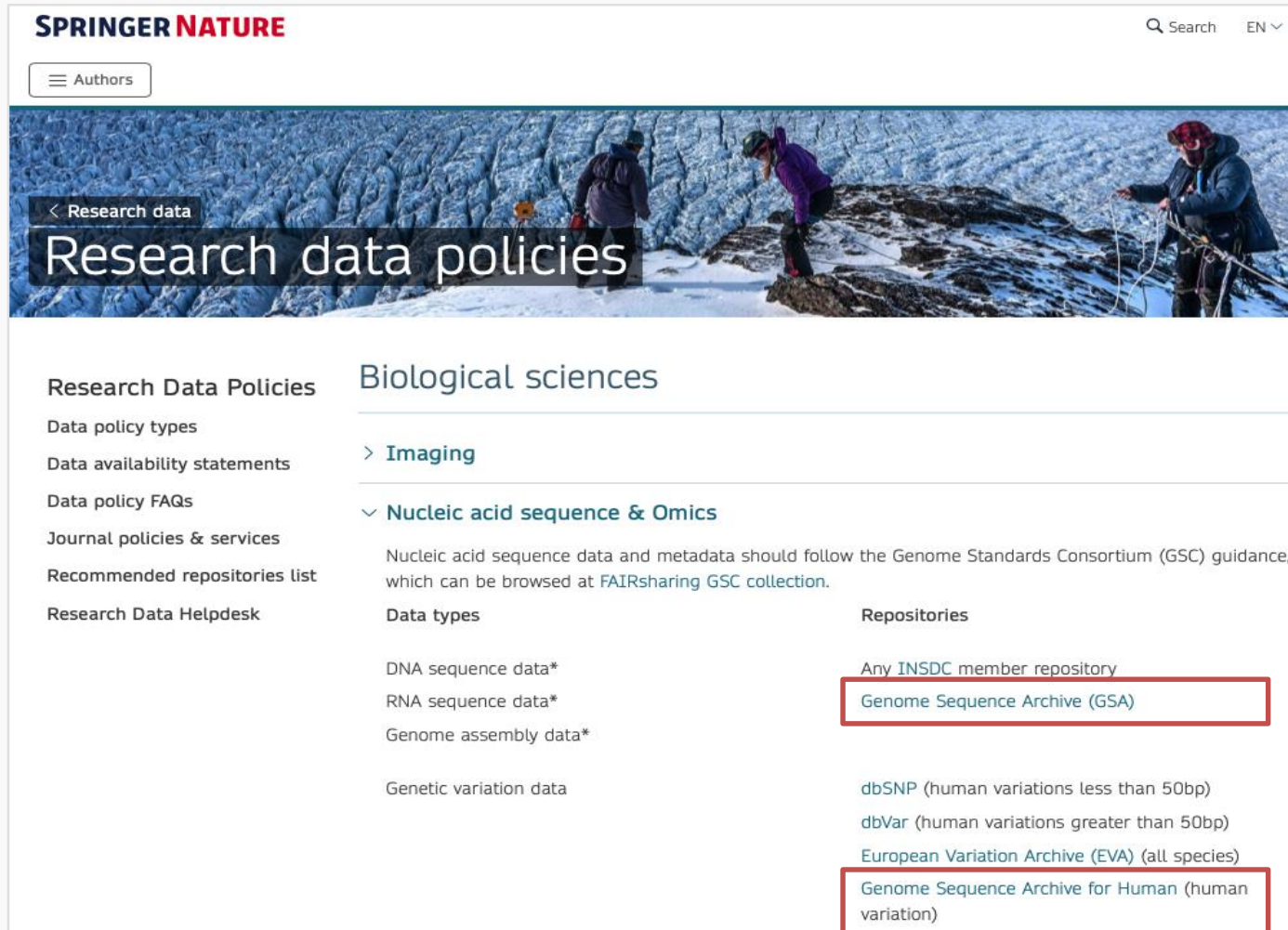


GSA Endorsed by Elsevier

 ELSEVIER	About Elsevier	Products & Solutions	Services	Shop & Discover	<input type="text" value="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



The screenshot shows the Springer Nature website's 'Research data policies' page for Biological sciences. The page features a navigation menu on the left with links to 'Data policy types', 'Data availability statements', 'Data policy FAQs', 'Journal policies & services', 'Recommended repositories list', and 'Research Data Helpdesk'. The main content area is titled 'Biological sciences' and includes a sub-section 'Nucleic acid sequence & Omics'. Under this section, it states that nucleic acid sequence data and metadata should follow the Genome Standards Consortium (GSC) guidance, which can be browsed at FAIRsharing GSC collection. Below this, there are two columns: 'Data types' and 'Repositories'. The 'Data types' column lists 'DNA sequence data*', 'RNA sequence data*', 'Genome assembly data*', and 'Genetic variation data'. The 'Repositories' column lists 'Any INSDC member repository', 'Genome Sequence Archive (GSA)', 'dbSNP (human variations less than 50bp)', 'dbVar (human variations greater than 50bp)', 'European Variation Archive (EVA) (all species)', and 'Genome Sequence Archive for Human (human variation)'. The 'Genome Sequence Archive (GSA)' and 'Genome Sequence Archive for Human (human variation)' are highlighted with red boxes.

SPRINGER NATURE Search EN

Authors

< Research data

Research data policies

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

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

Data Growth

Release date	Genome assembly No.
2018.10	23
2019.04	103
2019.10	139
2020.04	469
2020.10	1,575
2021.04	8,502
2021.10	10,185

GWH-supported Deposition

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

[More Journals >>](#)


Latest Released GWH

Organism	Accession
<i>Frankia</i>	GWHBFHH000000000

Genomics Proteomics Bioinformatics (2021)



Gene Expression Nebulas - GEN



Gene Expression Nebulas

A data portal of transcriptomic profiles across multiple species

- Home
- Datasets
- Genes
- Metadata
- Tools
- Statistics
- Documentation
- Download
- Contact

Gene Expression Nebulas (GEN) is a data portal of gene expression profiles under various conditions derived entirely from bulk and single-cell RNA-Seq data analysis in multiple species.

GEN Dataset

Enter query...

Search

e.g., [GEND000097](#); [Homo sapiens](#); [Lung](#); [SARS-CoV-2 infection](#); [TP53](#);

30
SPECIES

297
PROJECTS

323
DATASETS

50,500
SAMPLES

15,540,169
CELLS


293
PUBLICATIONS

Animals (17)

Plants (10)


Fungi (1)

Protists (2)




Homo sapiens
Human

192 Datasets
3673 Bulk Samples
6823695 Cells




Mus musculus
Mouse

11 Datasets
34 Bulk Samples
1176003 Cells




Drosophila melanogaster
Fruit fly

7 Datasets
18 Bulk Samples
3837235 Cells



Gallus gallus
Chicken

4 Datasets
117 Bulk Samples
42129 Cells

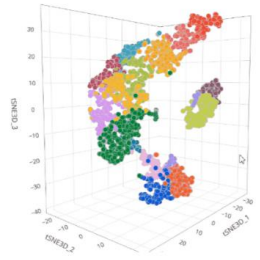


Macaca mulatta
Rhesus monkey

4 Datasets
22 Bulk Samples
304 Cells

Visualization of Bulk and Single-Cell Profiles

Select Dataset



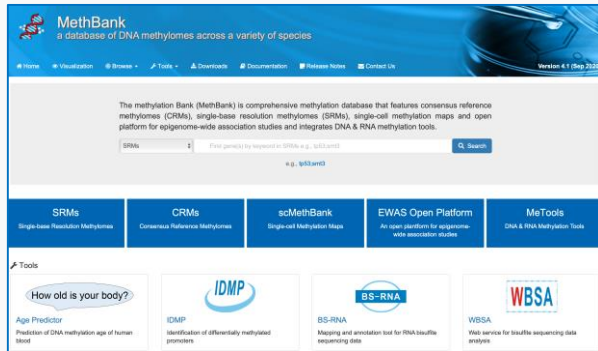
Nucleic Acids Research (2022)

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

China National Center for Bioinformation

24

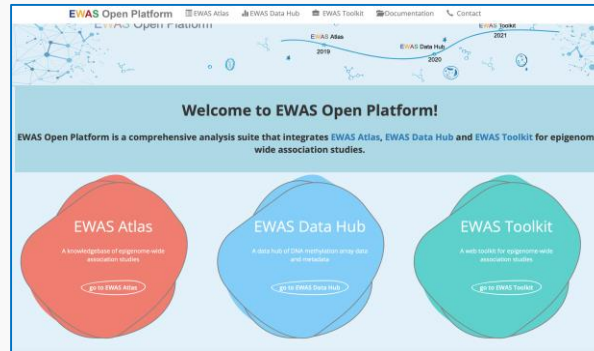
Epigenomic Resource



MethBank

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

Nucleic Acids Research (2018)



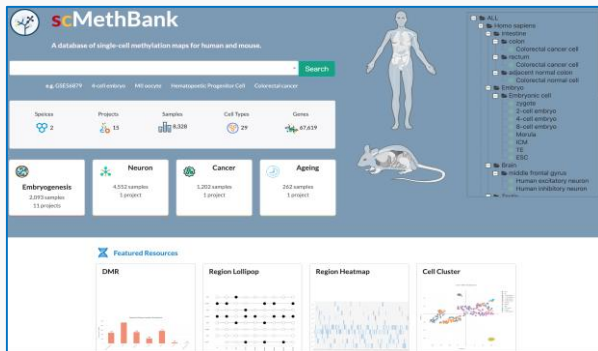
EWAS Open Platform

<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/>

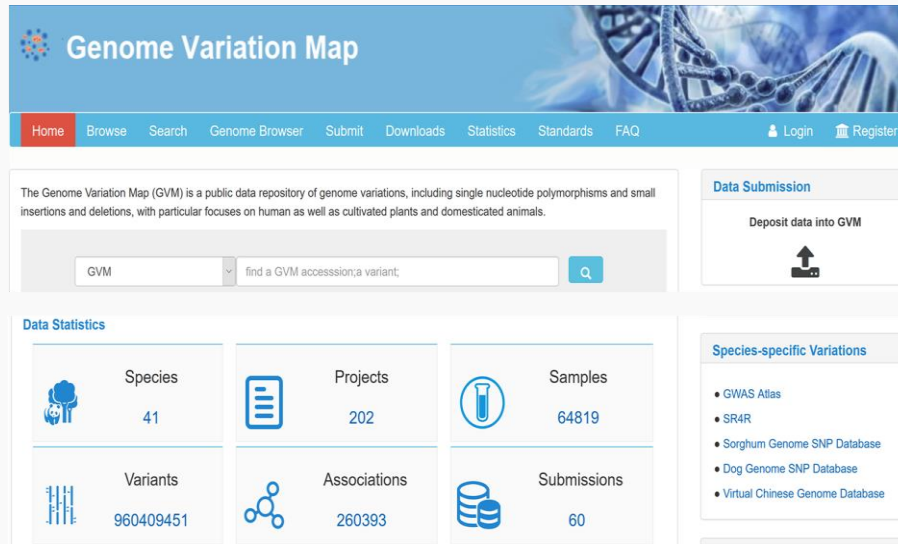
Nucleic Acids Research (2019)

**618 Traits
617,018 Associations**

**~ 260,000 PV
~ 70,000 UV**



Genomic Variation Resources



Genome Variation Map

The Genome Variation Map (GVM) is a public data repository of genome variations, including single nucleotide polymorphisms and small insertions and deletions, with particular focuses on human as well as cultivated plants and domesticated animals.

Data Submission

Deposit data into GVM

Data Statistics

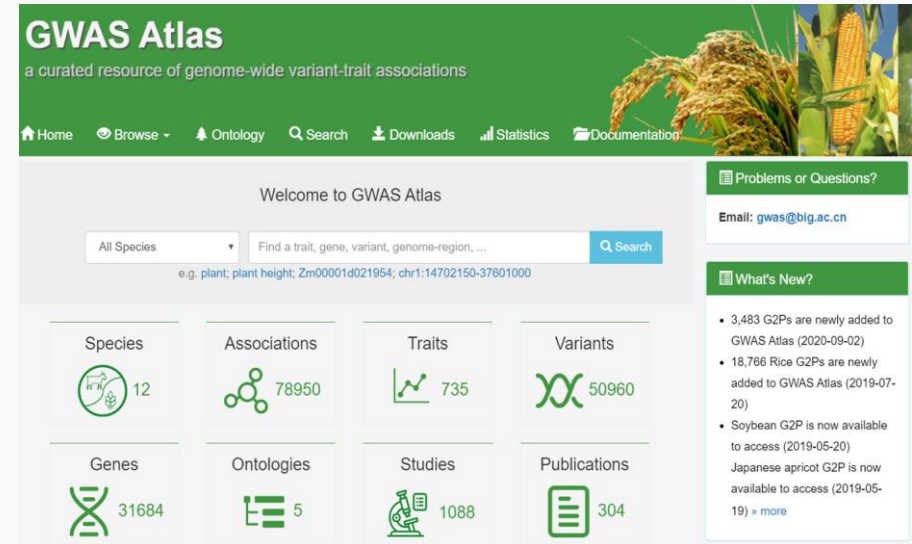
Species	Projects	Samples
41	202	64819
Variants	Associations	Submissions
960409451	260393	60

Species-specific Variations

- GWAS Atlas
- SR4R
- Sorghum Genome SNP Database
- Dog Genome SNP Database
- Virtual Chinese Genome Database

Genomic Variant Map

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



GWAS Atlas

a curated resource of genome-wide variant-trait associations

Welcome to GWAS Atlas

Search

All Species Find a trait, gene, variant, genome-region, ... Search

e.g. plant; plant height; Zm00001d021954; chr1:14702150-37601000

Species	Associations	Traits	Variants
12	78950	735	50960
Genes	Ontologies	Studies	Publications
31684	5	1088	304

Problems or Questions?

Email: gwas@big.ac.cn

What's New?

- 3,483 G2Ps are newly added to GWAS Atlas (2020-09-02)
- 18,766 Rice G2Ps are newly added to GWAS Atlas (2019-07-20)
- Soybean G2P is now available to access (2019-05-20)
- Japanese apricot G2P is now available to access (2019-05-19) » more

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)



China National Center for Bioinformation

RCoV19

RCoV19

Resource for Coronavirus 2019

Find virus strains by a keyword...

Q Search



Data
Submission



Data
Download



Raw Data

SARS-CoV-2 Sequences



World

6014176



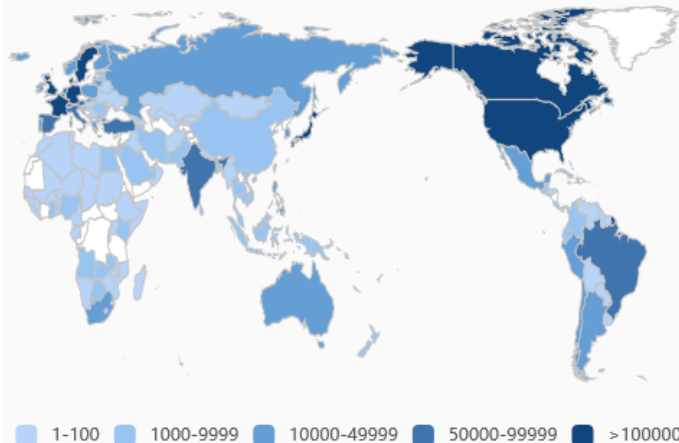
New

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



Popular Resources



SARS-CoV-2
Sequences



Coronavirus
Sequences



Lineage
Browse



Data Statistics



Clinical
Records



Literature



BLAST



Variation
Identification



Genome
Annotation



Sequence Growth

Navigation bar for the NCDC (China National Center for Bioinformation) website, showing the URL ngdc.cncb.ac.cn/ncov/monitoring and various search and navigation links.

Navigation links: 应用, 视频, Google 翻译, co-methylation, database commons, <https://www.ncbi.nlm.nih.gov/>, Sci-Hub: removing, SCI-Hub论文下载...

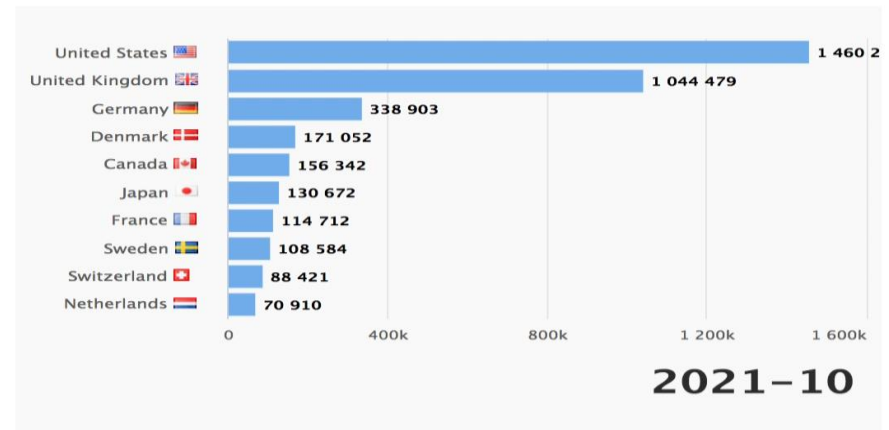
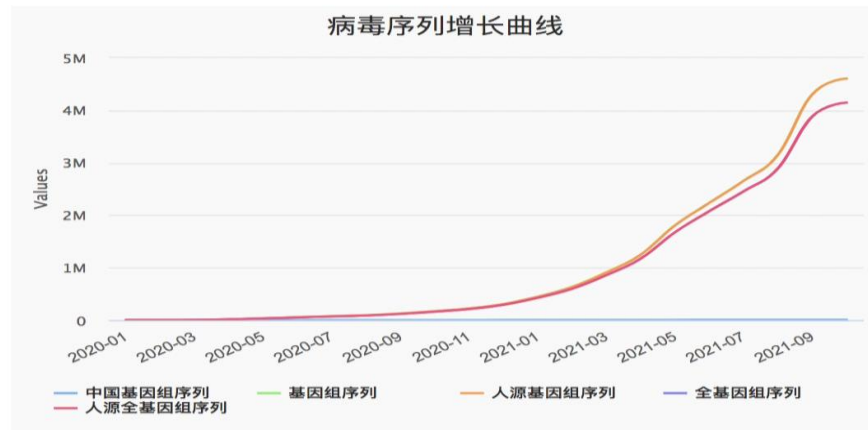
Site navigation: 首页, 基因组序列, 监测预警, 基因组变异, 临床信息, 在线工具, 文献情报, 关于

Language: 语言 / Language

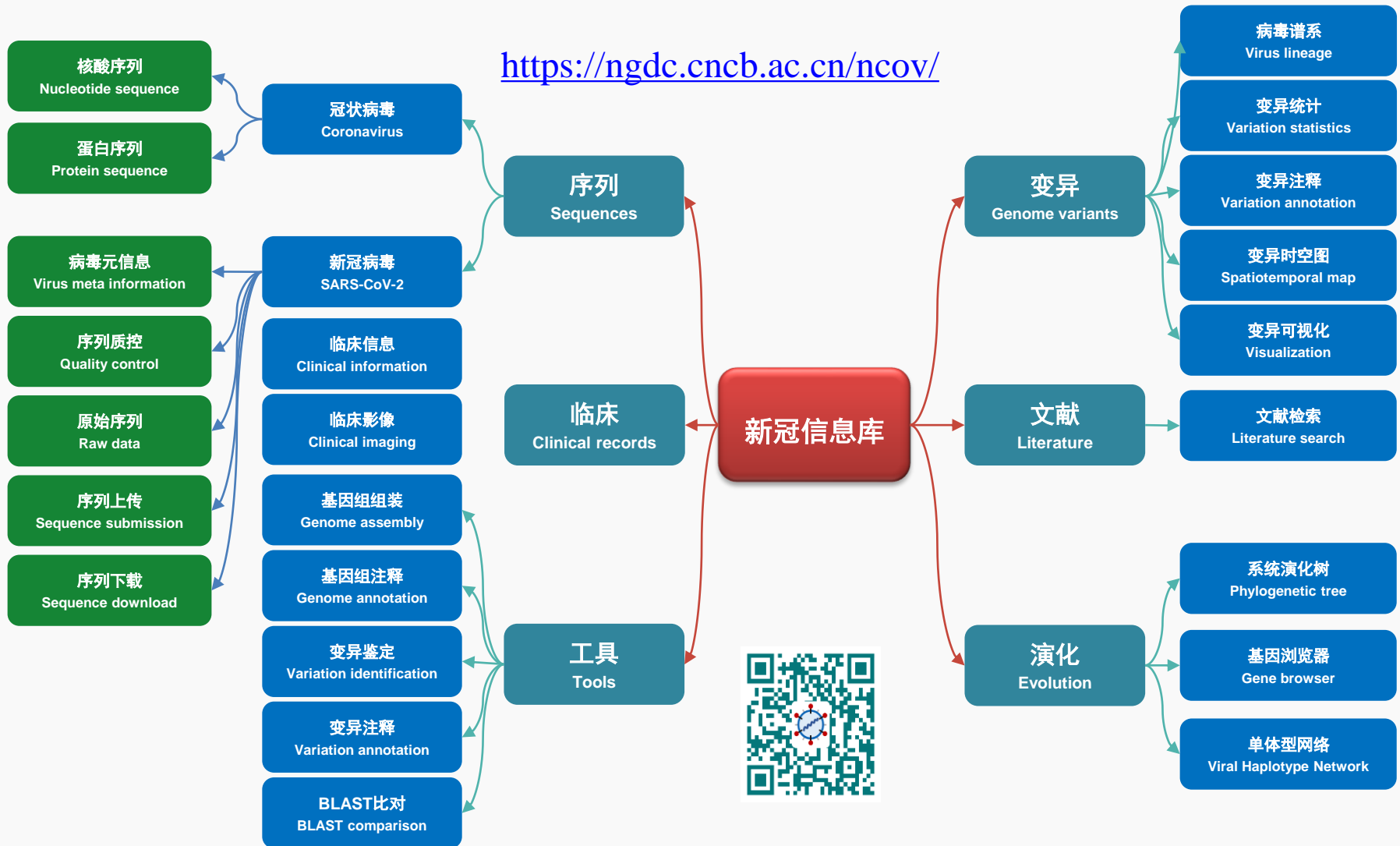
Breadcrumb: 首页 > 监测预警

监测预警

全球基因组序列增长趋势



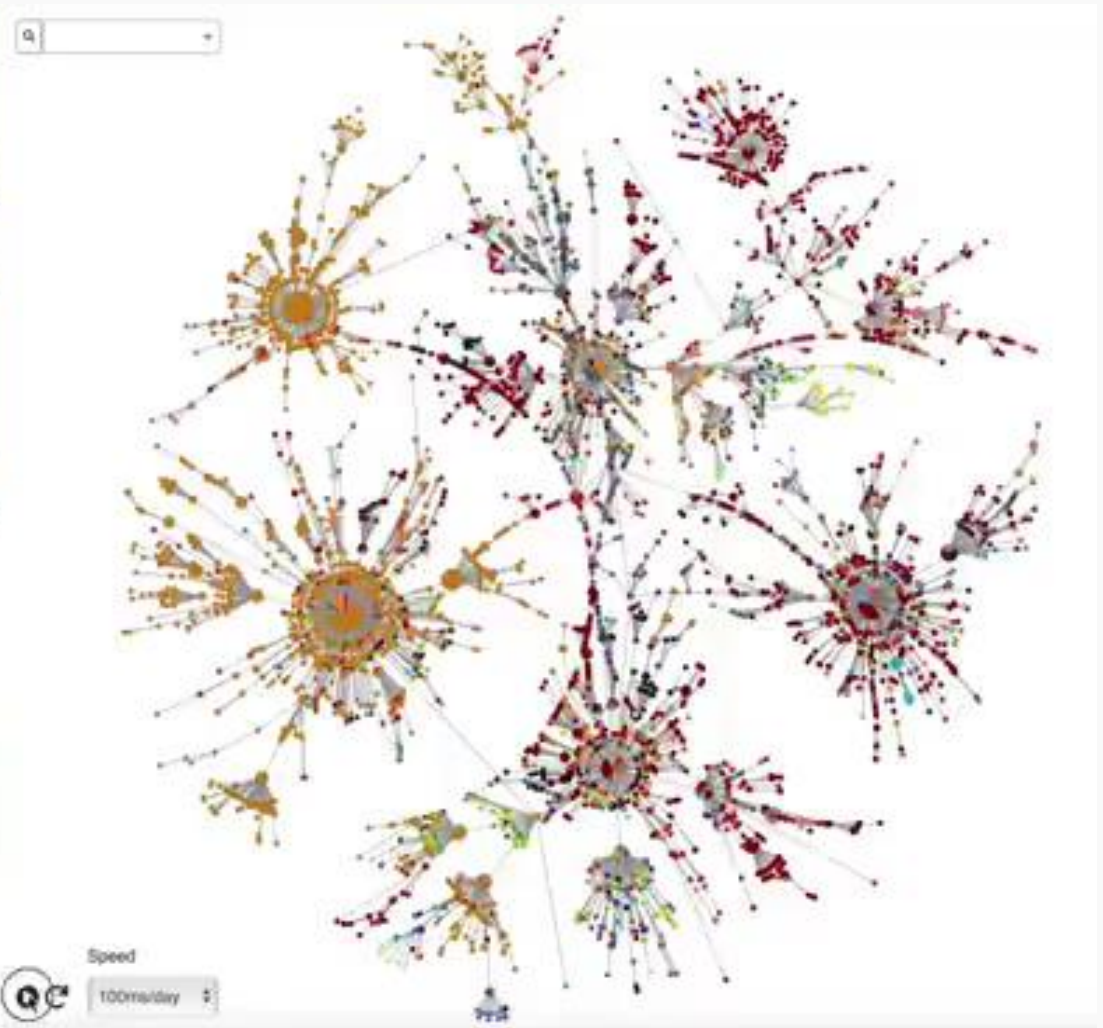
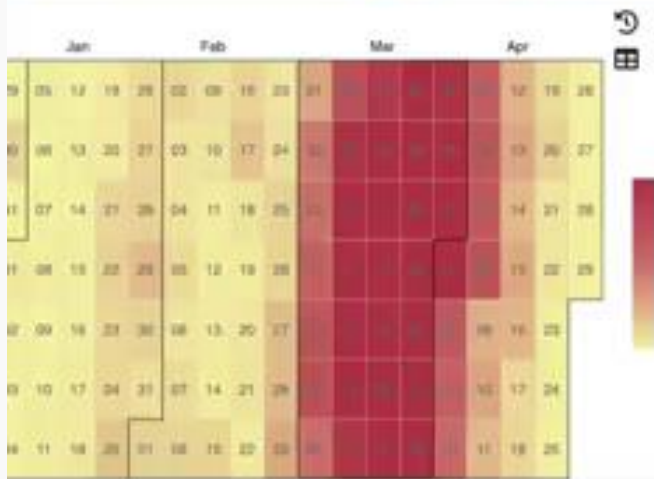
RCoV19 Modules



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

https://virology.net/organisms/coronaviridae/1-2/

Register Login

Viral Bioinformatics Research Centre

Chris Upton, University of Victoria, Canada

VBRC Tools About Us Organisms Blog Help

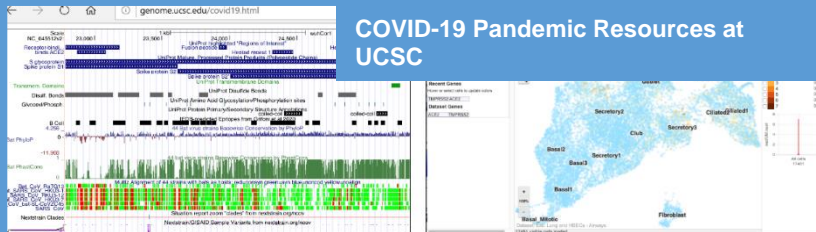
GENOMICS RESOURCES

- EMBL-EBI: [New SARS-CoV-19 data portal](#)
- INR/ELIXIR-ES and TransBioNet: [COVID-19 research](#)
- [China National Center for Bioinformation \(SARS-CoV-19\)](#)
- [GISAID.org](#) (Global Influenza Surveillance AID). DO NOT SUBMIT GENOMES HERE. Their policies prevent users from sharing the sequences.
- Submit SARS-CoV-19 sequences to [GenBank](#): rapid service

TWITTER TIMELINE

genome.ucsc.edu/covid19.html

COVID-19 Pandemic Resources at UCSC



Related publications and data resources

- Preprints of research manuscripts: [BioRxiv/MedRxiv COVID-19 Collection](#) (all freely available)
- [COVID-19 Open Research Dataset](#) of scholarly literature
- NIH Office of Data Science Strategy [Open-Access Data and Computational Resources to Address COVID-19](#)
- [European Molecular Biology Laboratory-European Bioinformatics Institute \(EMBL-EBI\)](#) and partners: [COVID-19 Data Portal](#)
- [China National Center for Bioinformation: 2019 Novel Coronavirus Resource](#)
- [SARS-CoV-2 Sequencing Resources on GISAID](#)
- Nextstrain open-source tracking of pathogen evolution: [SARS-Cov2 browser](#) (strain tree, geographical mapping)

ViralZone

Virus DB:

[China National Center for Bioinformation](#)

VIPR

CVR

Nextstrain

CORONAVIRUS TYPING TOOL

DB LINKS

Nucleotide DB: NCBI

Protein DB: UniProtKB

Virus DB:

CVR

Nextstrain

CORONAVIRUS TYPING TOOL

TAXONOMY

Group IV: ssRNA positive-strand viruses

Order: Nidovirales

Family: Coronaviridae

Subfamily: Coronavirinae

Genus: Betacoronavirus

ETYMOLOGY

Coronavirus: from latin corona (crown), referring to the shape of proteins around the virion

SPECIES

SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2)

https://www.ncbi.nlm.nih.gov/

Reference Genome

NC_045512.2: 1.30K (29,903 nt)

Download the GFF

Other Resources

The Centers for Disease Control and Prevention (CDC) website has outbreak information updated daily, including a [Situation Summary](#).

[Information for Laboratories \(CDC\)](#)

[2019 nCoV Resource by China National Center for Bioinformation](#)

You are here: NCBI

GETTING STARTED RESOURCES POPULAR FEATURED NCBI INFORMATION Support Center



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

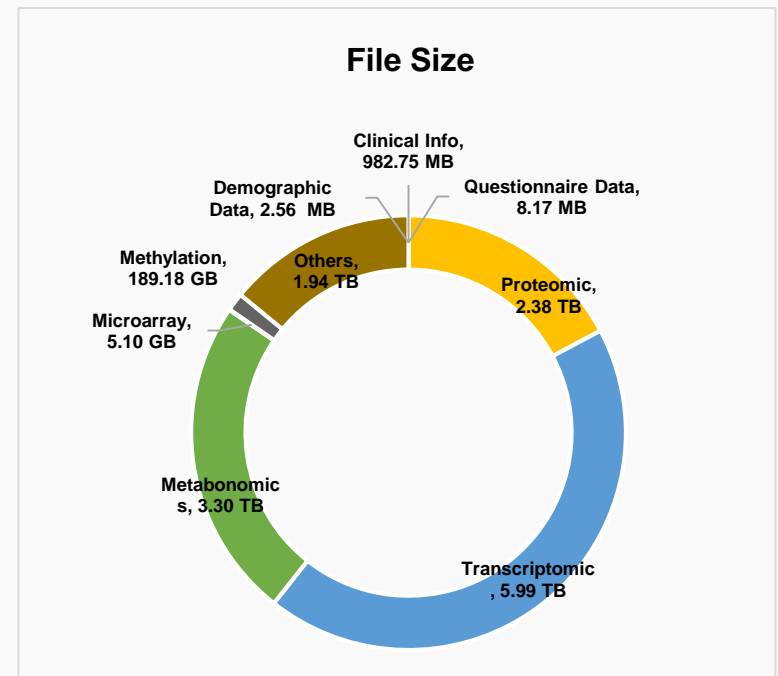
Show 10 entries

OMIX ID	Project	Title	Organism	Access Type	Release Time	Operation
OMIX691	PRJCA006797	scRNAseq data of anatomically distinct fibroblast subsets determine skin autoimmune disease patterns	Homo sapiens	Controlled	2021-10-07	
OMIX687	PRJCA006458	mNGS in the diagnosis of HHV-1 reactivation in a critically ill COVID-19 patient	Severe acute respiratory syndrome coronavirus 2	Open	2021-10-06	 
OMIX685	PRJCA006637	Database of O-GlcNAc modified proteins in human placental trophoblast cell line	Homo sapiens	Open	2021-09-25	 
OMIX683	PRJCA005262	Interference PTPRZ1 expression profiling chip in glioma stem cells	Mus	Open	2021-09-24	 
OMIX681	PRJCA006391	The mechanism of embryonic diseases	Mus musculus	Open	2021-09-23	 
OMIX661	PRJCA006378	Copy number variation in Chinese children with complete atrioventricular canal and single ventricle	Homo sapiens	Open	2021-09-22	 
OMIX655	PRJCA006631	Decipher unwinding mechanisms of RNA helicases MOV10 and MOV10L1	Mus	Open	2021-09-19	 
OMIX656	PRJCA006631	Proximal single-stranded RNA stabilizes human telomerase RNA G-quadruplex and induces its distinct conformers	Homo sapiens	Open	2021-09-19	 
OMIX657	PRJCA006631	Discover a new HRP-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>



China National Center for Bioinformation

Open Library of Bioscience(OpenLB)

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

OpenLB

Open Library of Bioscience

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

Beta 1.0.0

Q Search
Advanced Search

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

Rearing system causes changes of behavior, microbiome and gene expression of chickens.

Siyu Chen, Hai Xiang, Hui Zhang, Xu Zhu, Dan Wang, Jikun Wang, Tao Yin, Langqing Liu, Minghua Kong, Hua Li, Xingbo Zhao

Author Information ▶

PMID: 30916350 DOI: 10.3382/ps/pez140

Abstract

It has been long demonstrated that cage rearing (CR) deprives the animal of the possibility to express natural behaviors and results in stress. However, the effect of the rearing system on gene expression and the molecular levels of the gut microbiome are unknown. 10-wk-old Beijing You chickens were studied in parallel CR and free-range (FR) systems for 30 wk, to investigate the effect of rearing systems on behavior, microbiota composition, and gene expression. From week 40, a match-mismatch design was conducted for 5 wk. The results indicated that CR deprives the animals of natural behaviors, evidenced by sham dust-bathing behavior. A decreased alpha diversity of gut microbiome composition of CR chickens was seen in FR compared to CR-FR chickens ($P < 0.001$), and the alpha diversity of gut microbiome composition of FR-CR was decreased as compared to FR chickens ($P = 0.045$). The heat map and beta-diversity analysis showed that the cluster of gut microbial compositions were similar between the mismatch groups (FR-CR and CR-FR), while those of CR showed the lowest diversity from the 4 groups. The relative abundance of gut microbes at genera and species levels was different between comparisons ($P < 0.05$). Moreover, the CR (both CR and FR-CR) triggered the downregulation of most Kyoto encyclopedia of genes and genomes pathways, while it was upregulated in 2 genetic information processing pathways, compared to FR hens regardless of long or short term. In conclusion, CR deprived chickens of their normal behavior and resulted in changes in the microbiome diversity and pathways and gene expression of chickens.

Keywords

behavior chicken gene expression gut microbiome rearing system

MeSH Term

- Animal Husbandry
- Animals
- Bacteria
- Behavior, Animal
- Chickens
- Female
- Gastrointestinal Microbiome

Links to CNCB-NGDC Resources

BioProject: PRJCA007797 (Free-ranged vs caged chickens)

GSA: CRA000833 (LSY_caecum_microbiome)

GSA: CRA000834 (nLQ_caecum_microbiome)

GSA: CRA000836 (nSY_caecum_microbiome)

GSA: CRA000837 (SYLQ_caecum_microbiome)

GSA: CRA000845 (LQ-S_transcriptome)

GSA: CRA000848 (LSY-S_transcriptome)

GSA: CRA000850 (SY-S_transcriptome)

GSA: CRA000844 (SYLQ-S_transcriptome)

Word Cloud

Abstract text: ~33 millions

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

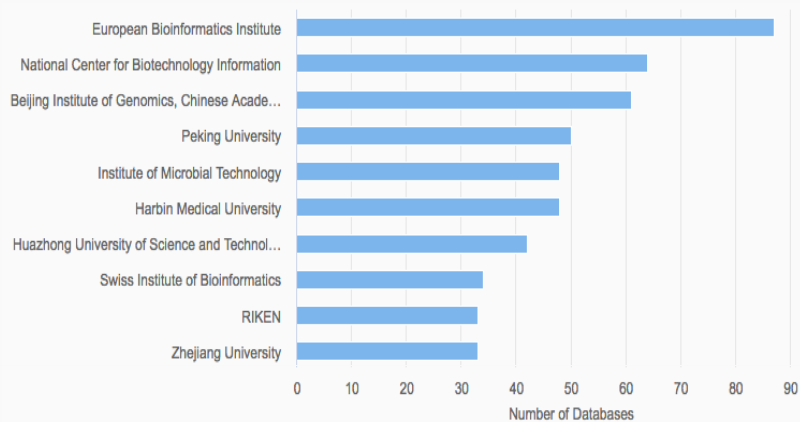
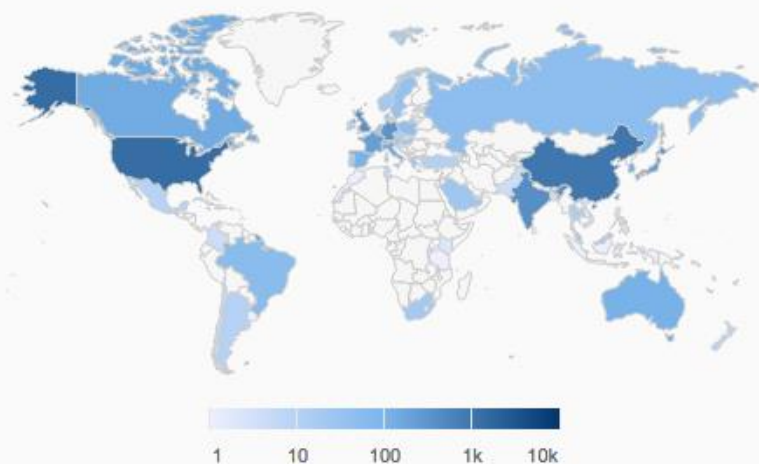
<https://ngdc.cncb.ac.cn/openlb/home>



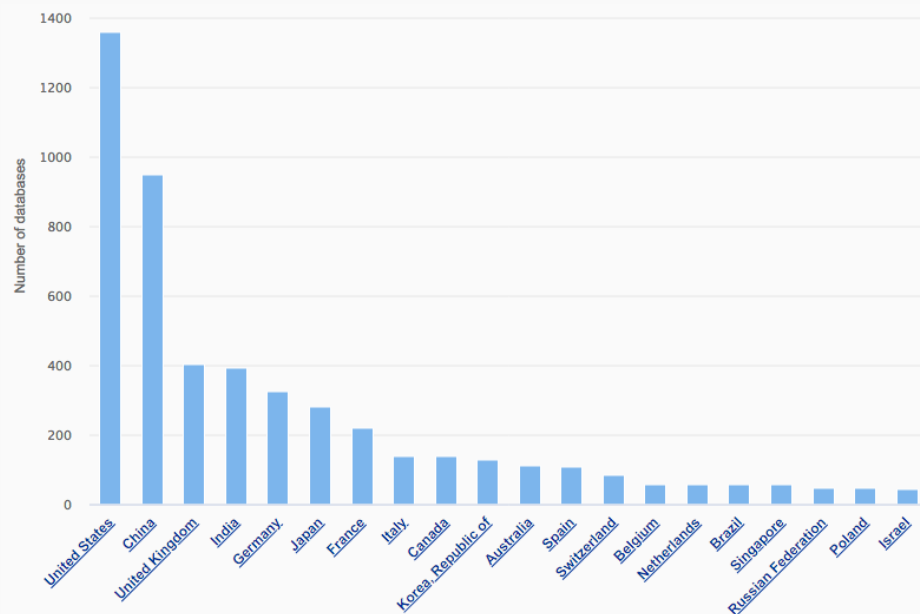
Database Commons

Worldwide biological databases

5468 databases distributed in 70 countries/regions



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

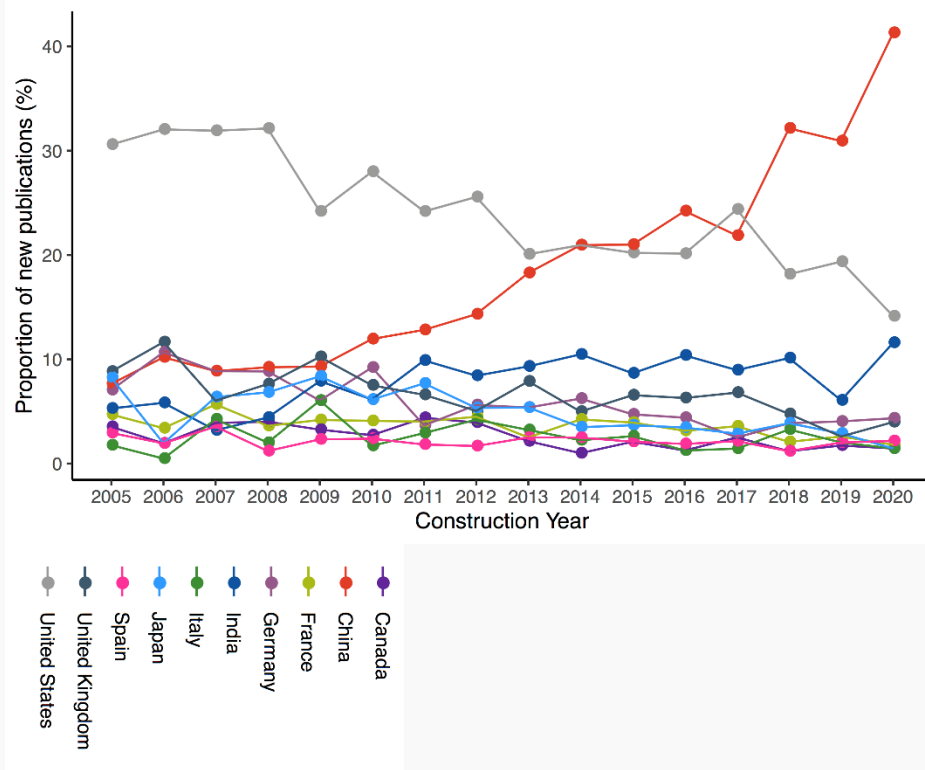


<https://ngdc.cncb.ac.cn/databasecommons/>

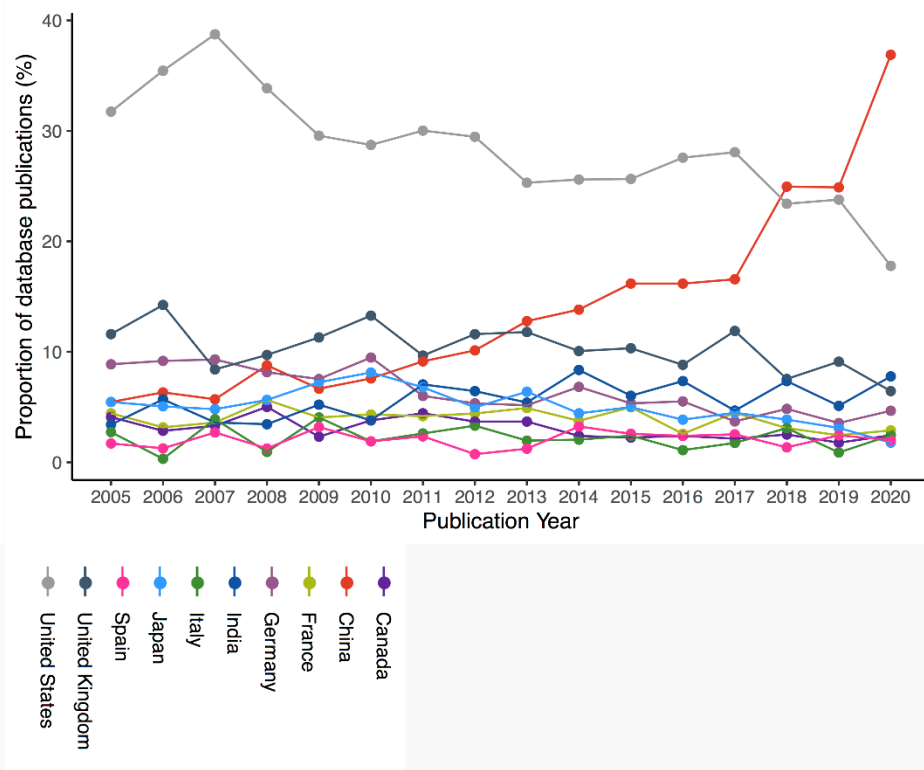


Database Commons

Trends in new database construction

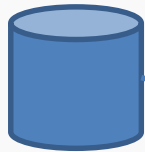


Trends in database publication



BIG Search: Cross-database search engine

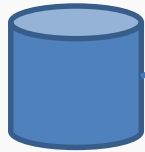
DBs within NGDC



Indexing files
JSON



DBs outside NGDC



Indexing files
JSON



BIG Data Center in Beijing Institute of Genomics

BIG Search

All Databases

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

Total: **74**

BIGD People	0	People working in BIG Data Center
Database Commons	1	A catalogue of biological databases
GEN	0	Gene Expression Nebulas
SEGreg	0	Database of specifically expressed genes and regulation
AnimalTFDB	23	AnimalTFDB is a comprehensive database including classification and annotation of genome-wide transcription factors
BioCode	0	Archive Bioinformatics Codes for Open Source Projects
BioProject	0	Biological Project Library
BioSample	0	Biological Sample Library
dbPAF	7	database of Phospho-sites in Animals and Fungi
DEG	36	Database of Essential Genes
DoriC	0	Database of Replication Origins

Partner Databases

AnimalTFDB^{2.0}

dbPSP

dbPFT

EKPD

CCDB

WERAM

DEG

DoriC

MiCroKiTS

hTfTarget

dbPAF

TMANATOS

PceRBase

PlantTFDB

RhesusBase

PLMD

iUUCD

SEGreg

The LncRNA and Disease Database

lncRNASNP2

PceRBase



National Genomics Data Center

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

Q 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
InCAR	28,420	InCAR 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

Q 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

Q 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



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

Q 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.



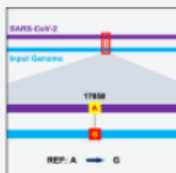
BLAST

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.



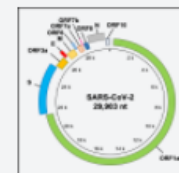
Variant Annotation

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

Reference	Start Position	End Position	Base
SARS-CoV-2	1	1	T
SARS-CoV-2	10	10	G
SARS-CoV-2	1	2	TT
SARS-CoV-2	1	2	AA
SARS-CoV-2	100	100	C
SARS-CoV-2	100	100	T

Genome Annotation

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



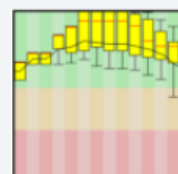
Fastq-to-Variants

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



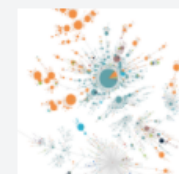
SeqQC

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



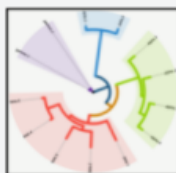
Haplotype Network

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



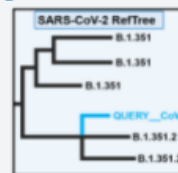
Phylogenetic Tree

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

CCTGGTTTCAACGAGAT
GTGGCTTTGGAGACTCCGTGGAGGAGGTCCTCAACTTG
CTTAGTAGAAGTTGAAAAAGGCGTTTGCCTCAACTTG
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCT
GTAGTGGTGAGACACTTGGTGTCTTGTCCCTCATGTG
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATA
TCTTCGCACTGATCCTTATGAAGATTTTCA

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

Genome Warehouse (GWH) protein sequences

IC4R rice protein sequences

NCBI non-redundant protein sequences (nr)

Sorghum proteins

nucleotide

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^{1,2,3,4}, Cuiqing Li^{1,2,3,4}, Lu Kang^{1,4,5,6}, Dongmei Tian^{1,2,3,4}, Nazish Badar^{4,6}, Wentai Ma^{1,4,5}, Shalei Zhao^{1,4,5}, Xuan Jiang^{1,5}, Chun Wang^{1,4,5}, Yongqiao Sun¹, Wenjie Li¹, Meng Lei¹, Shuangli Li¹, Qitun Qi¹, Amer Ikram¹, Muhammad Salman¹, Maysah Usman¹, Huma Shireen¹, Fatima Batool¹, Bing Zhang¹, Hua Chen^{1,4,5,6}, Yimeng Yang^{1,4,5}, Amir Ali Abbasi^{1,4}, Mingkun Li^{1,4,5,6}, Yongbiao Xue^{1,4,5,6}, Yimeng Bao^{1,2,3,4,7}

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) Yimeng (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

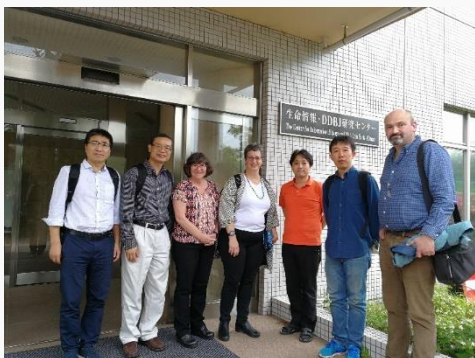
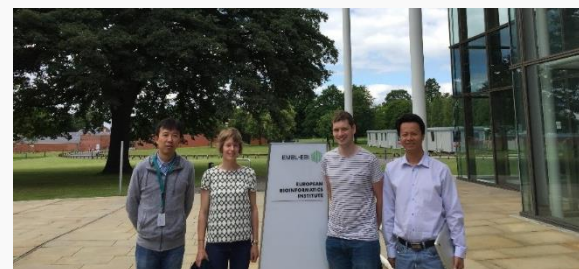
DDBJ



NCBI



EBI



May, 2017, INSDC Meeting



Sept, 2017 & 2018, visit & training



May, 2019 INSDC meeting



China National Center for Bioinformation

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; Coronidovirineae; 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**





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>



China National Center for Bioinformation