China National Center for Bioinformation

Genomics Data Resources in China

Yiming Bao

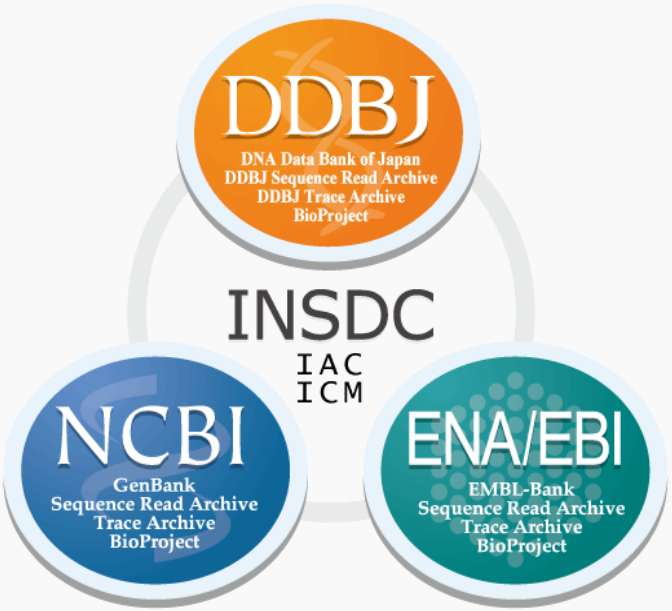
Director
National Genomics Data Center
Beijing, China

1st Sino-EU Permed Science & Technology Webinar
Sep. 13, 2022 • 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



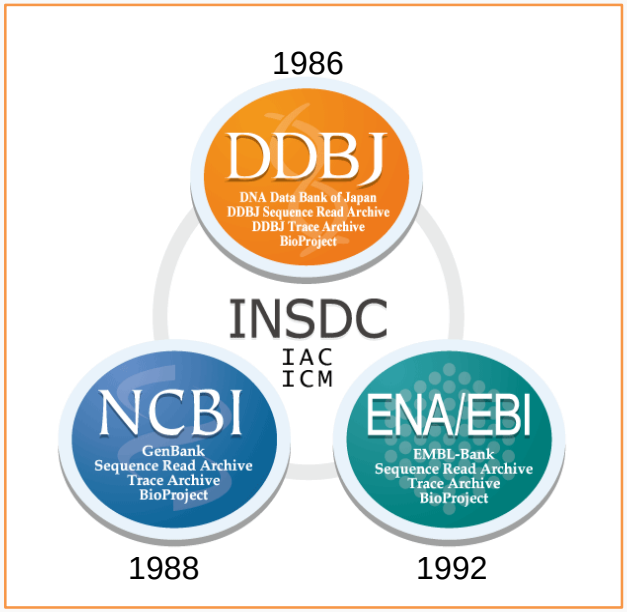
USA, 1988 Europe, 1992

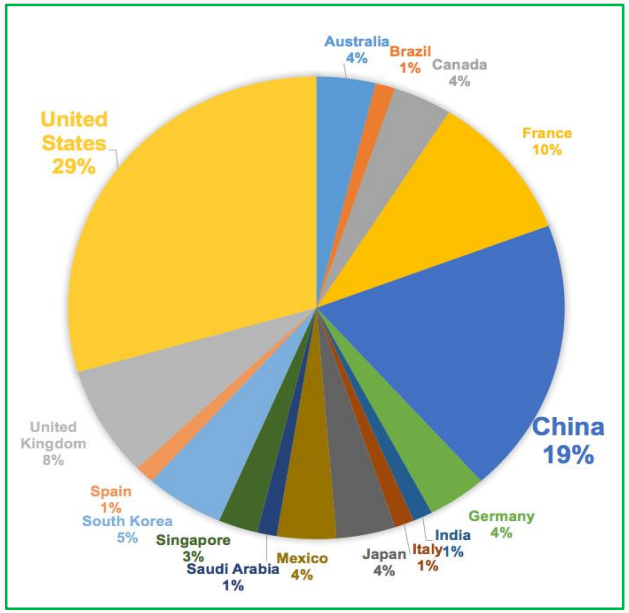
- NCBI: 1988, by US congress
- EBI: 1992, by EMBL
- DDBJ: 1986, by NIG of Japan
- NCBI, EBI and DDBJ form INSDC
- Establish international standard, exchange data daily, hold annual meeting
- Before papers are published, data need to be deposited into an international recognized database

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

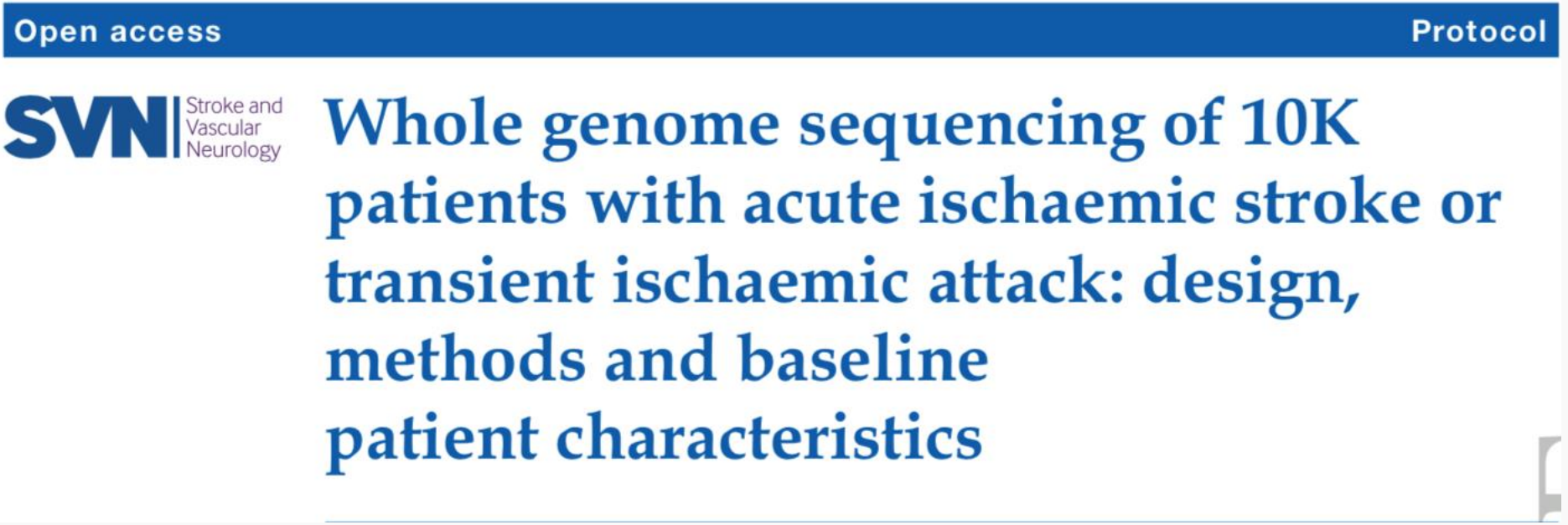
Background in China

- **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

Large Data Submission to GSA-Human



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.

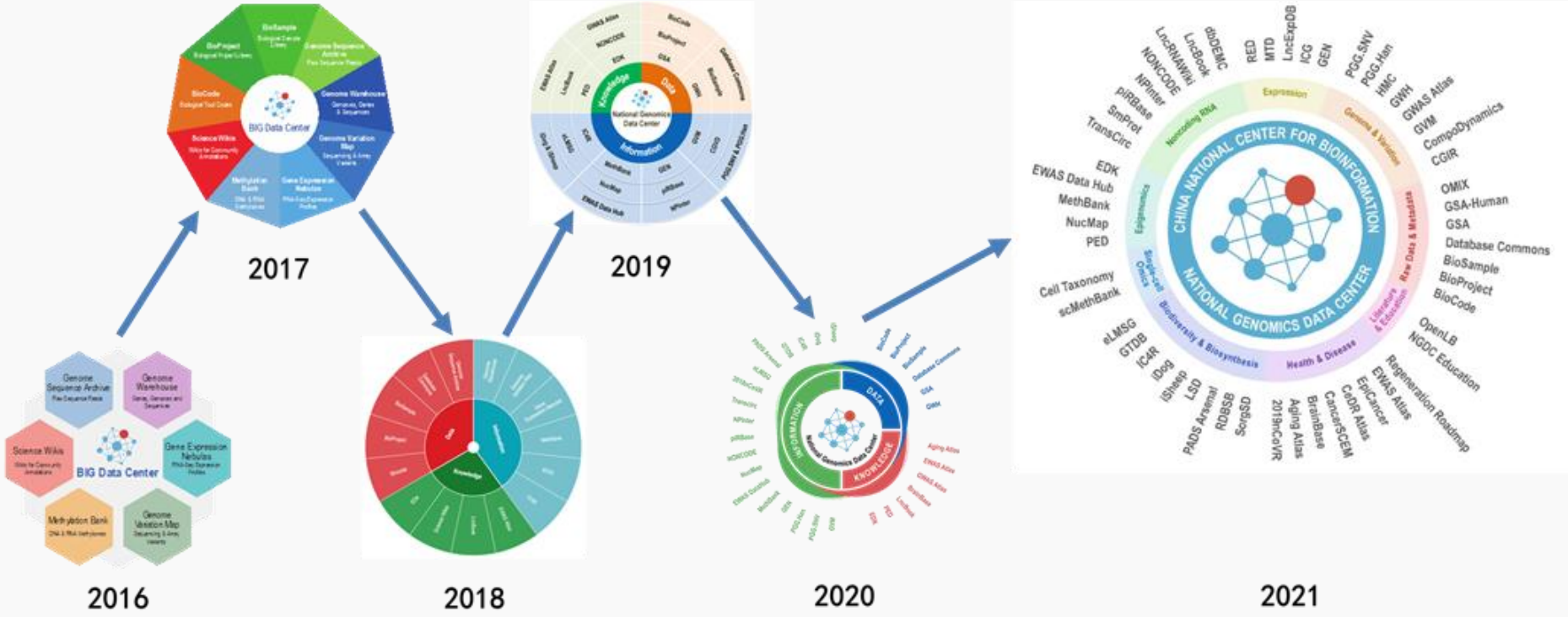
Translating big data into big discoveries

Deposition

Integration

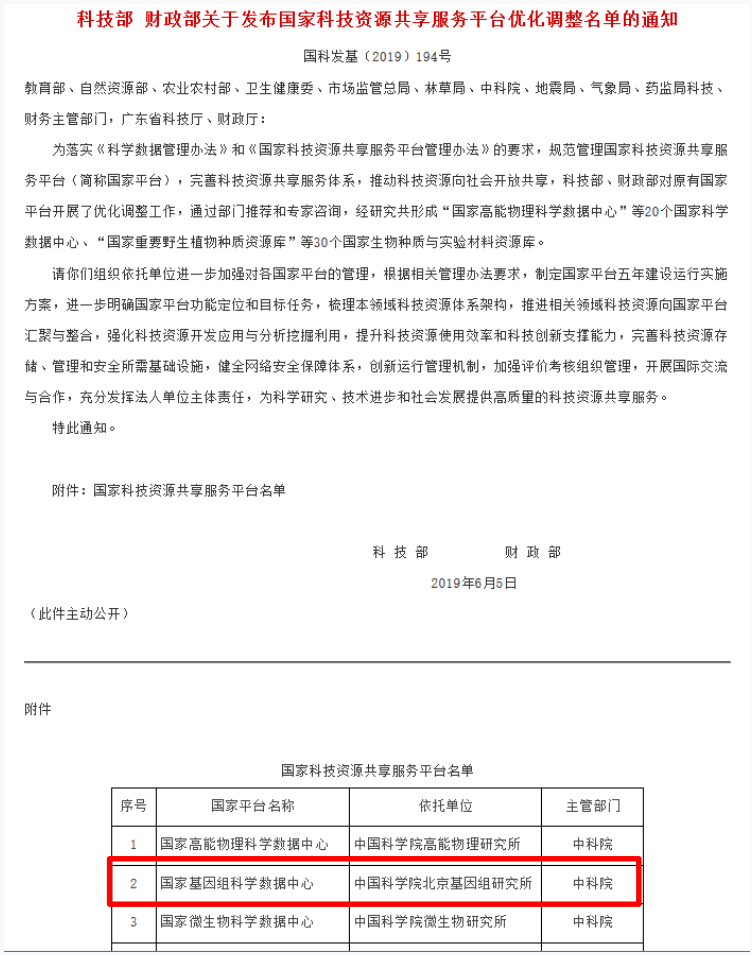
Translation

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

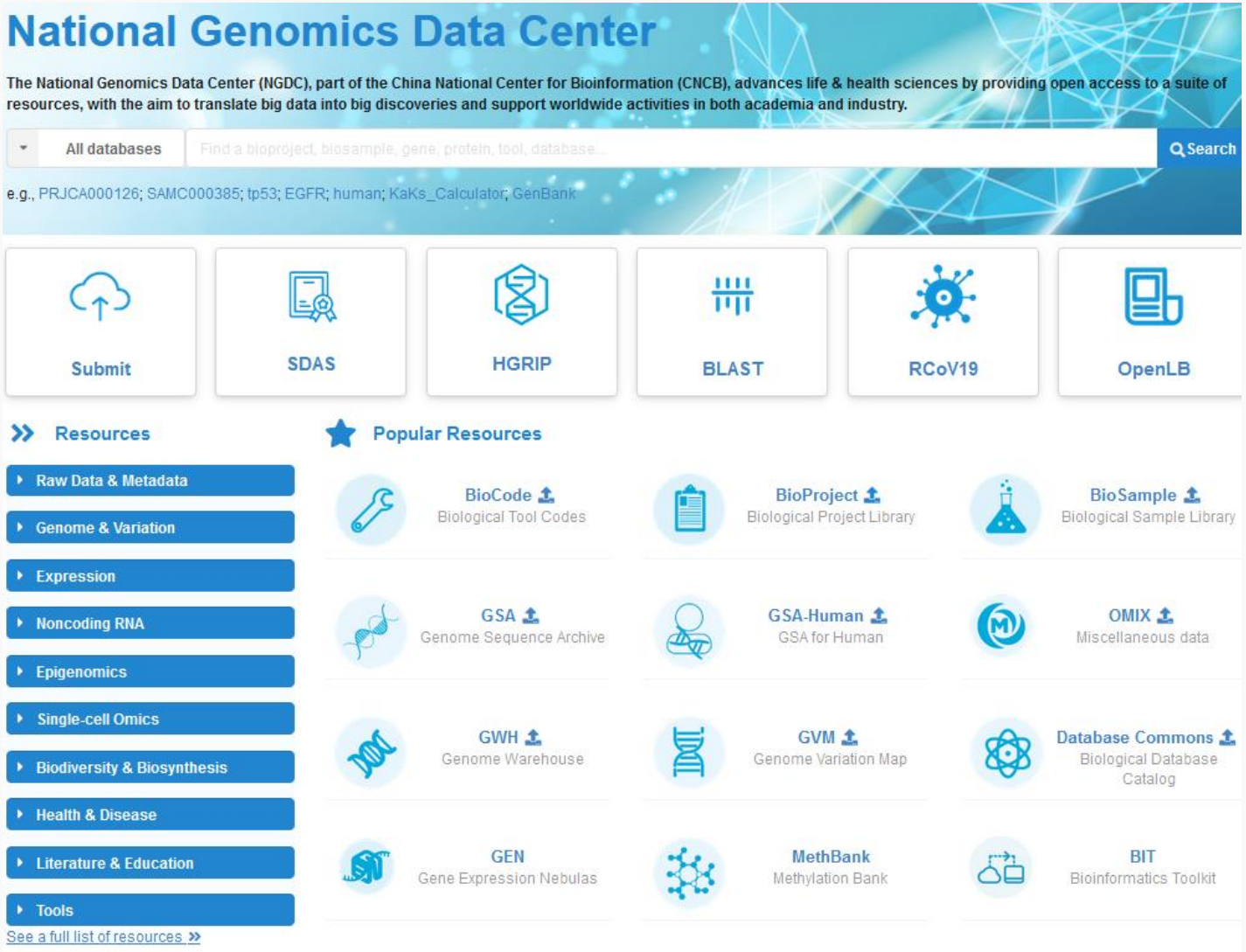
院属各单位、院机关各部门：

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



- Omics databases
- Specialized DBs
- Literatures
- Tools

Genome Sequence Archive

The Genome Sequence Archive (GSA) is a data repository for collecting, archiving, managing and sharing raw sequence data, which is the first repository of the genome sequence data with international journal recognition in China.

Submit
Submit data to GSA

Download
Download data to your computer

Browse
Browse publicly available records

Document
Find help information and documents

Collected and Archived Data

PROJECTS	SAMPLES	EXPERIMENTS	RUNS
599,867	28,160,347	20,717,799	22,211,560

Data of Concern

- Monkeypox Virus Sequence Data
- Omicron Sequence Data in South Africa
- SARS-CoV-2 Sequence Data
- Human Genetic Resources Data
- Open Archive for Miscellaneous Data

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

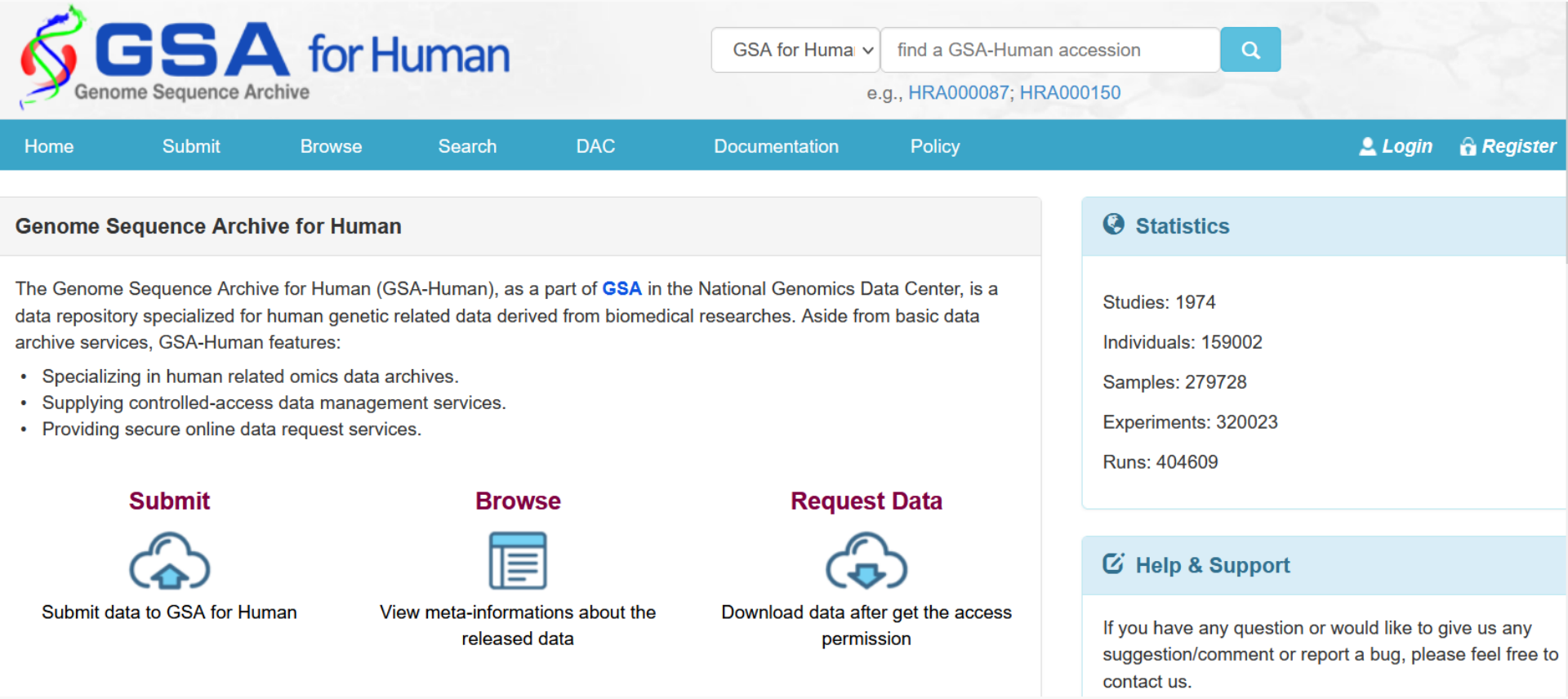
QQ group: 548170081

We highly appreciate your comments and suggestions for further improvements.

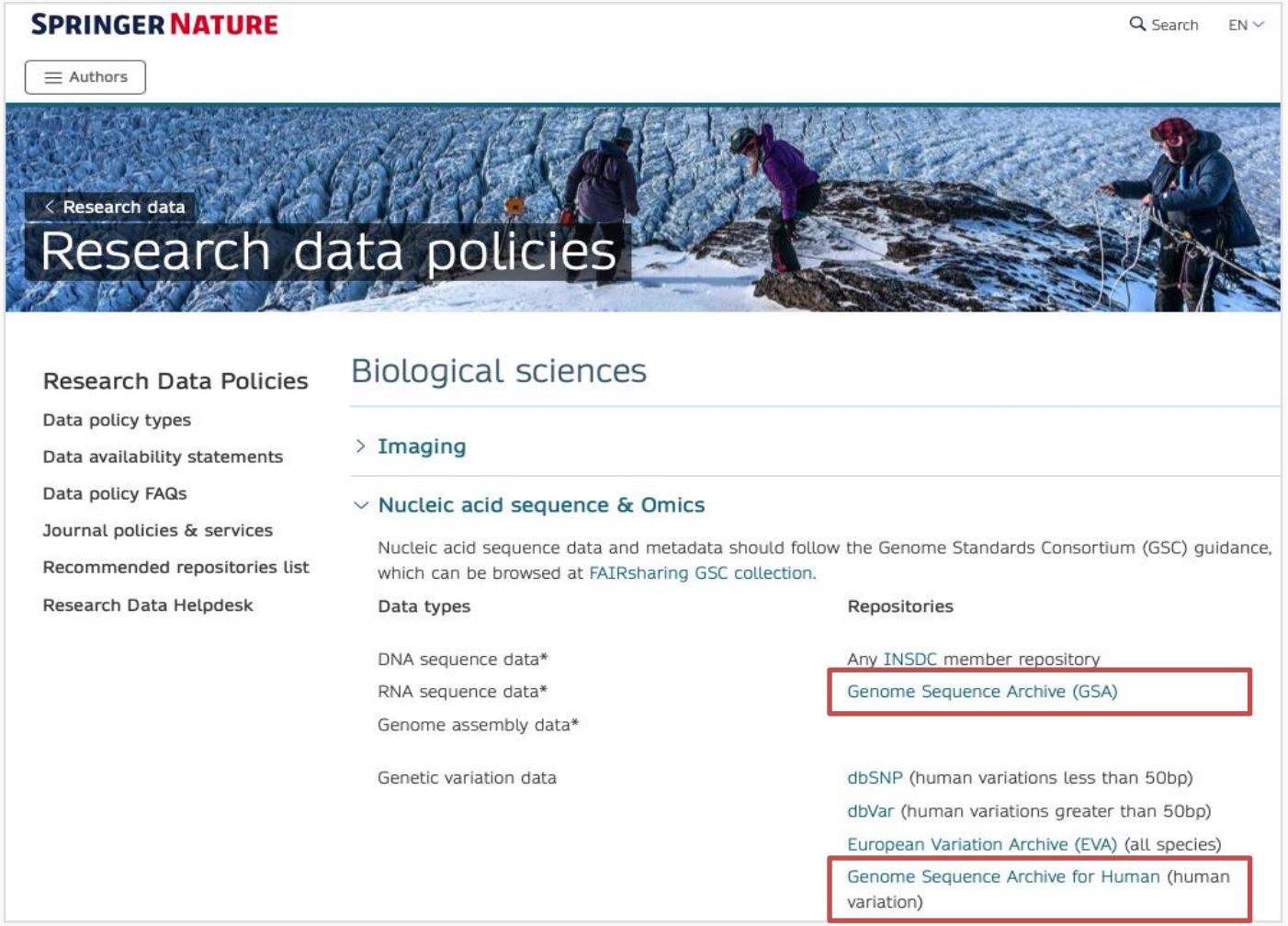
GSA-supported Deposition

Data submissions to GSA have been reported by

GSA-Human


GSA for Human
Genome Sequence Archive
GSA for Huma
find a GSA-Human accession
e.g., HRA000087; HRA000150
Home
Submit
Browse
Search
DAC
Documentation
Policy
Login
Register
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
Statistics
Studies: 1974
Individuals: 159002
Samples: 279728
Experiments: 320023
Runs: 404609
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.

GSA Endorsed by Springer Nature



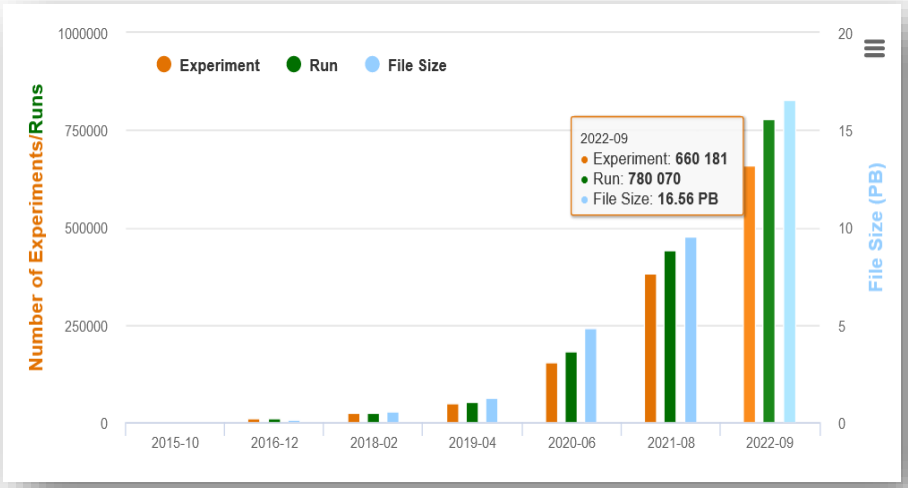
GSA Supported Major Grants

Agencies	Grants	GSAs	Experiments	Runs
Ministry of Science and Technology of the People's Republic of China (MOST)	1175	432	26064	28887
National Natural Science Foundation of China (NSFC)	4525	1565	80087	86207
Chinese Academy of Sciences (CAS)	874	383	23187	26691

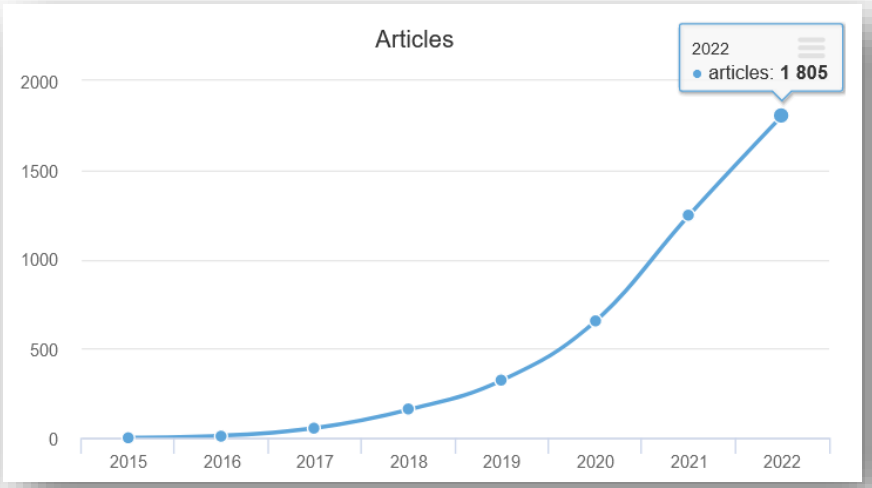
Total Grants: 6,574

GSA Statistics

Total file size: >16 PB



Data reported in papers: 1805



Human Data Backup & Registration Protocol

Backup Center

Backup

Backup ID

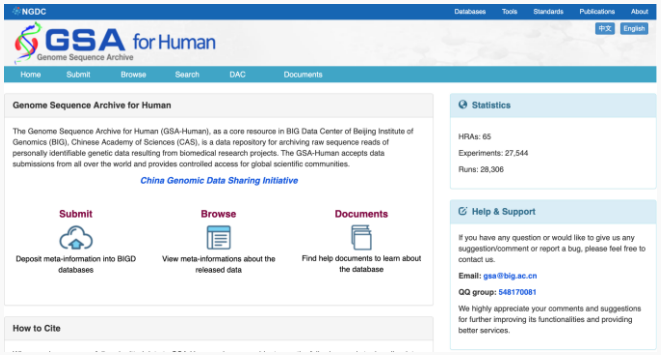
Registration Center

Registration ID

Researchers

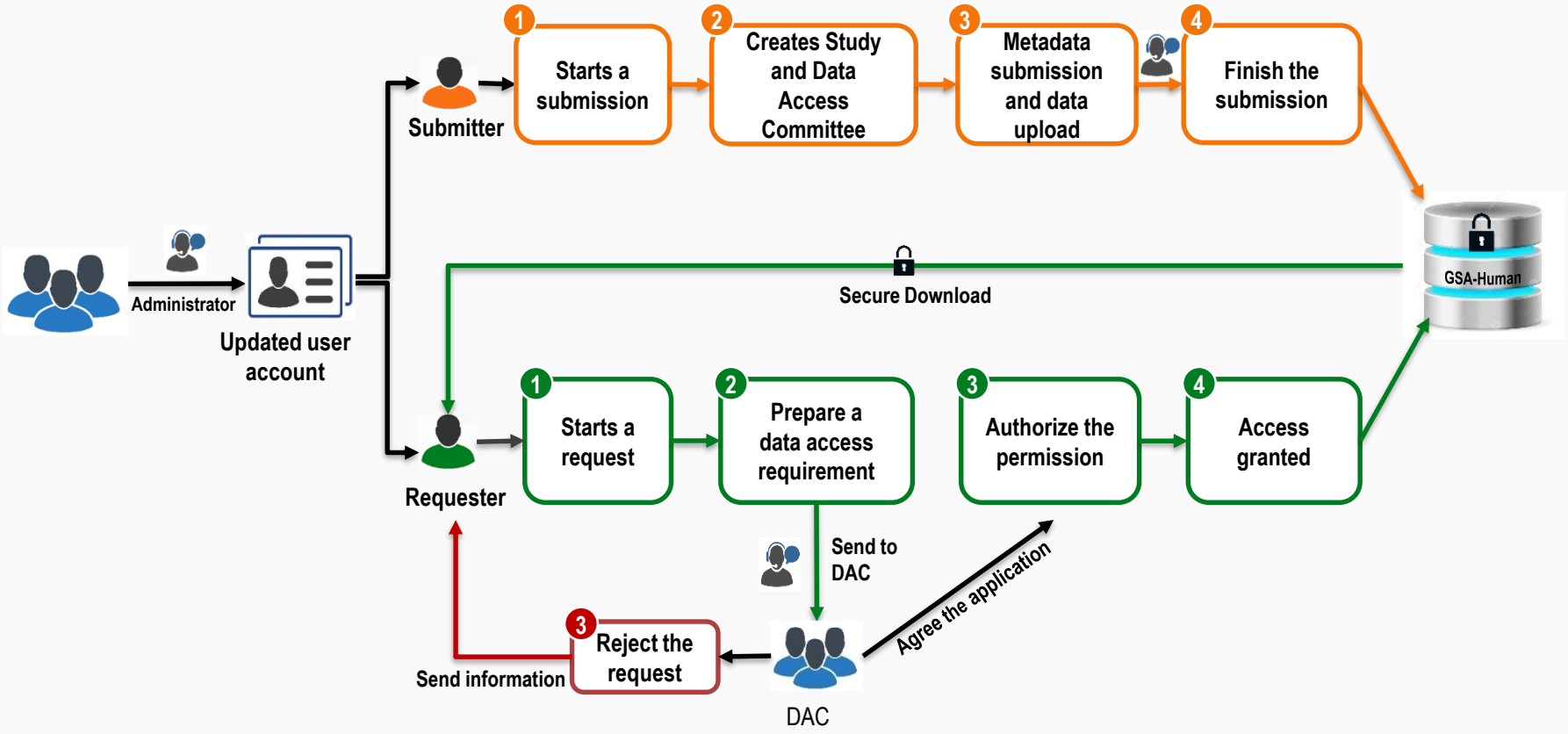
Submission

Accession #s



CNCB-NGDC GSA-Human

Workflow for data upload/download



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

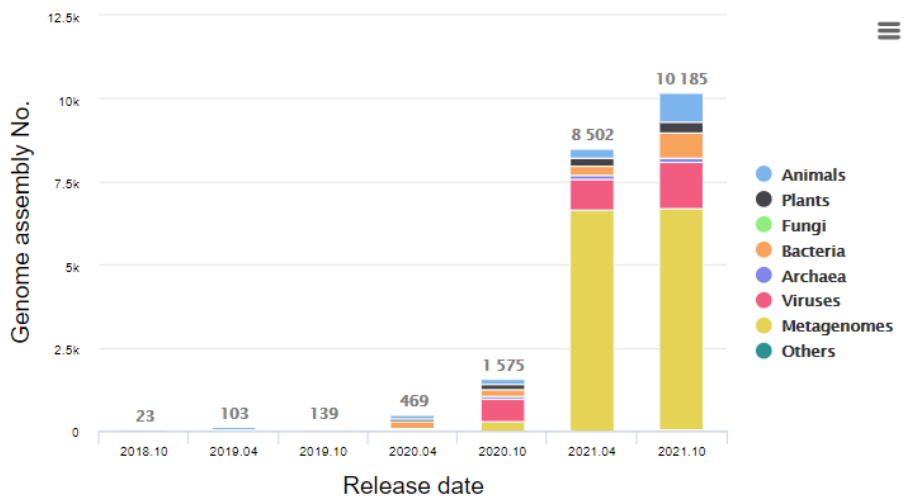
Submit
Deposit meta-information into GWH databases

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:

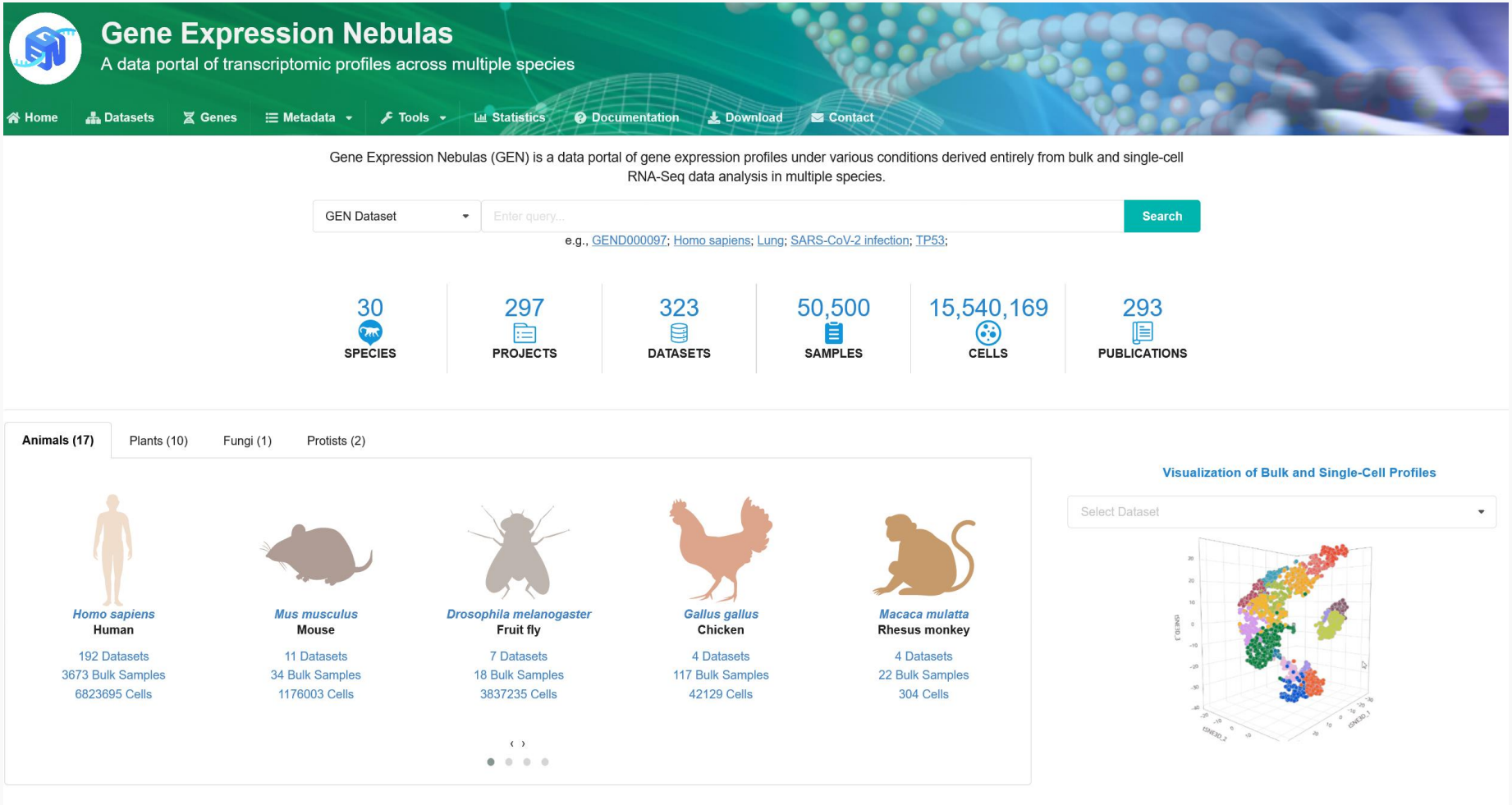
More Journals >>

Latest Released GWH

Organism	Accession
	GWHBFHH00000000

Genomics Proteomics Bioinformatics (2021)

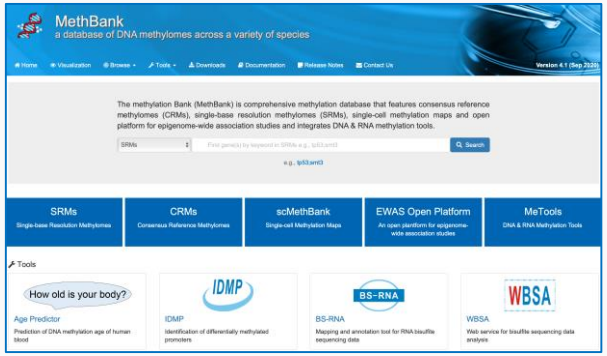
Gene Expression Nebulas - GEN



Nucleic Acids Research (2022)

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

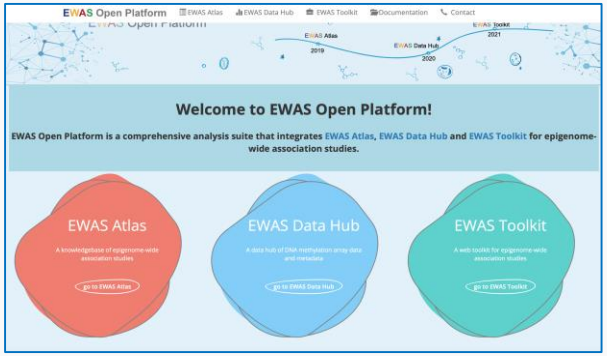
Epigenomic Resource



MethBank

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

Nucleic Acids Research (2018)



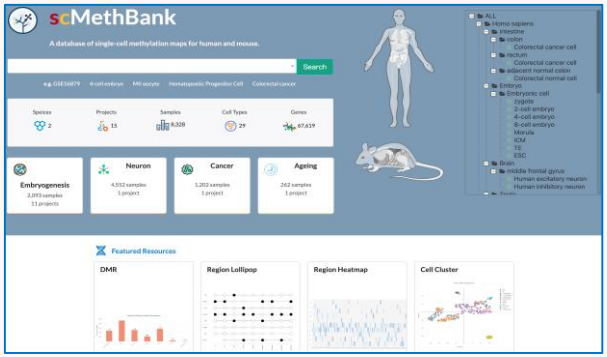
EWAS Open Plantform

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

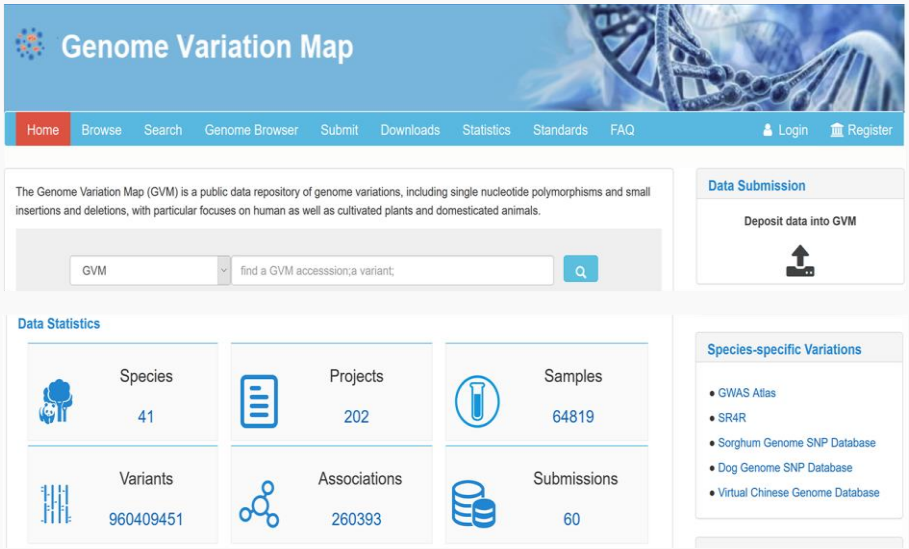
https://ngdc.cncb.ac.cn/nucmap/

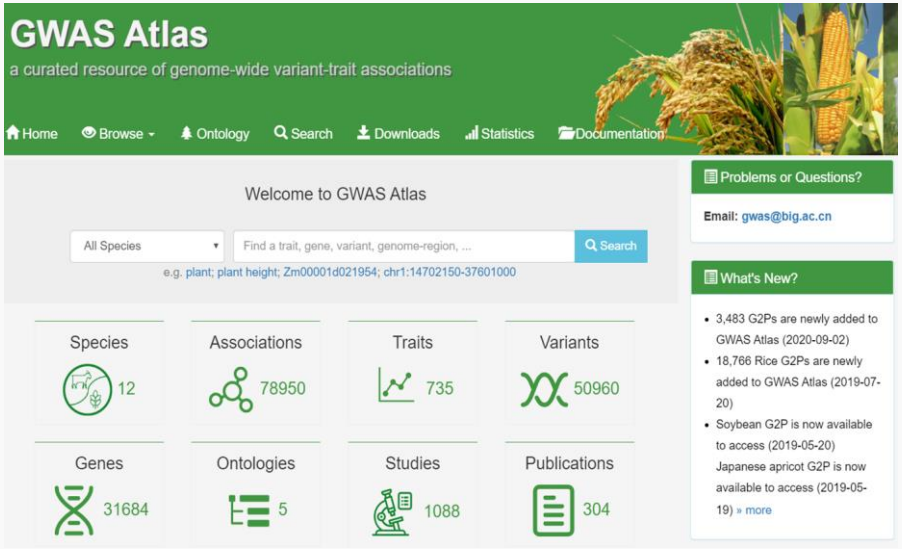
Nucleic Acids Research (2019)

618 Traits
617,018 Associations

~ 260,000 PV
~ 70,000 UV

Genomic Variation Resources





Genomic Variant Map

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

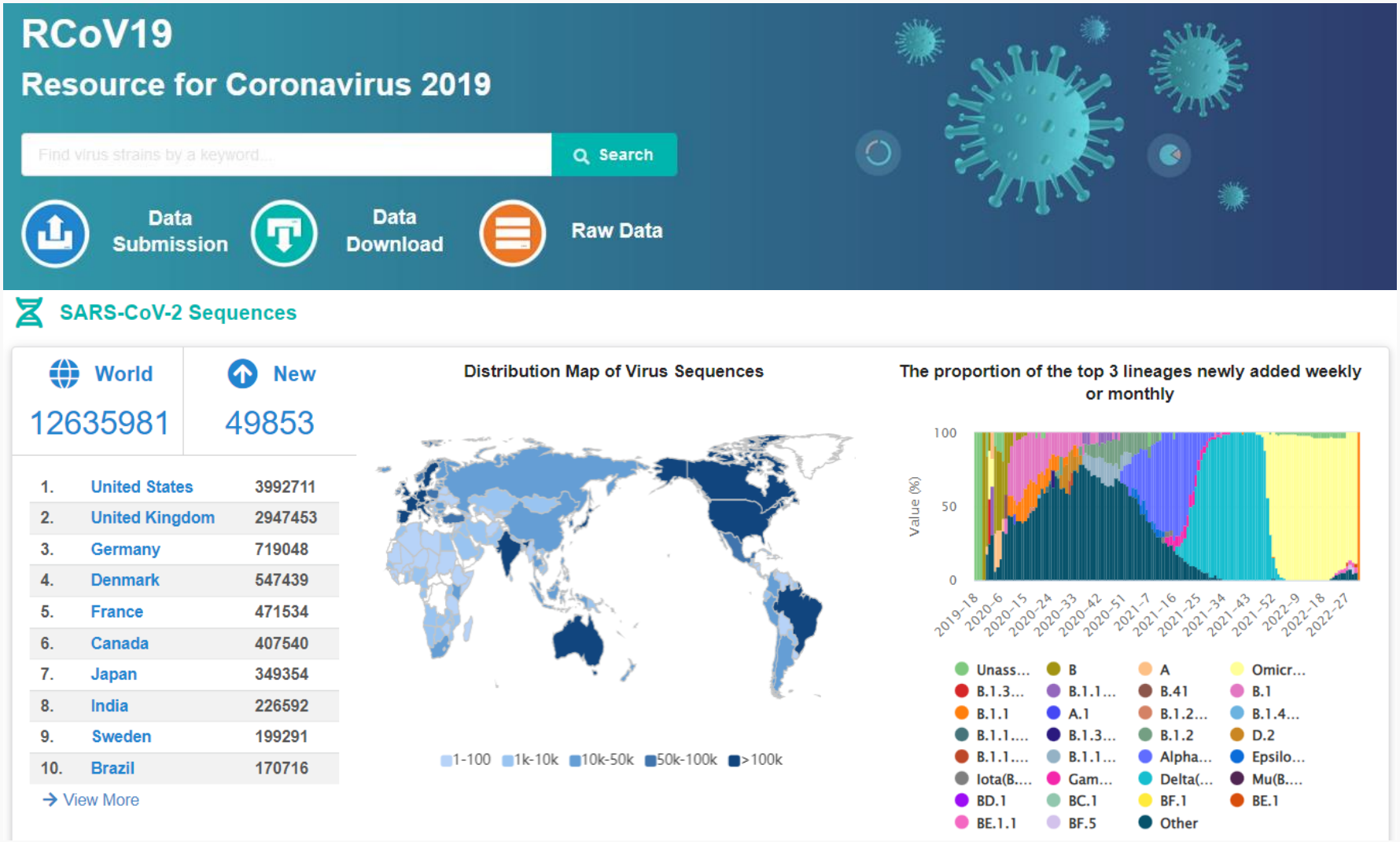
Genomic Variant Knowledge

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

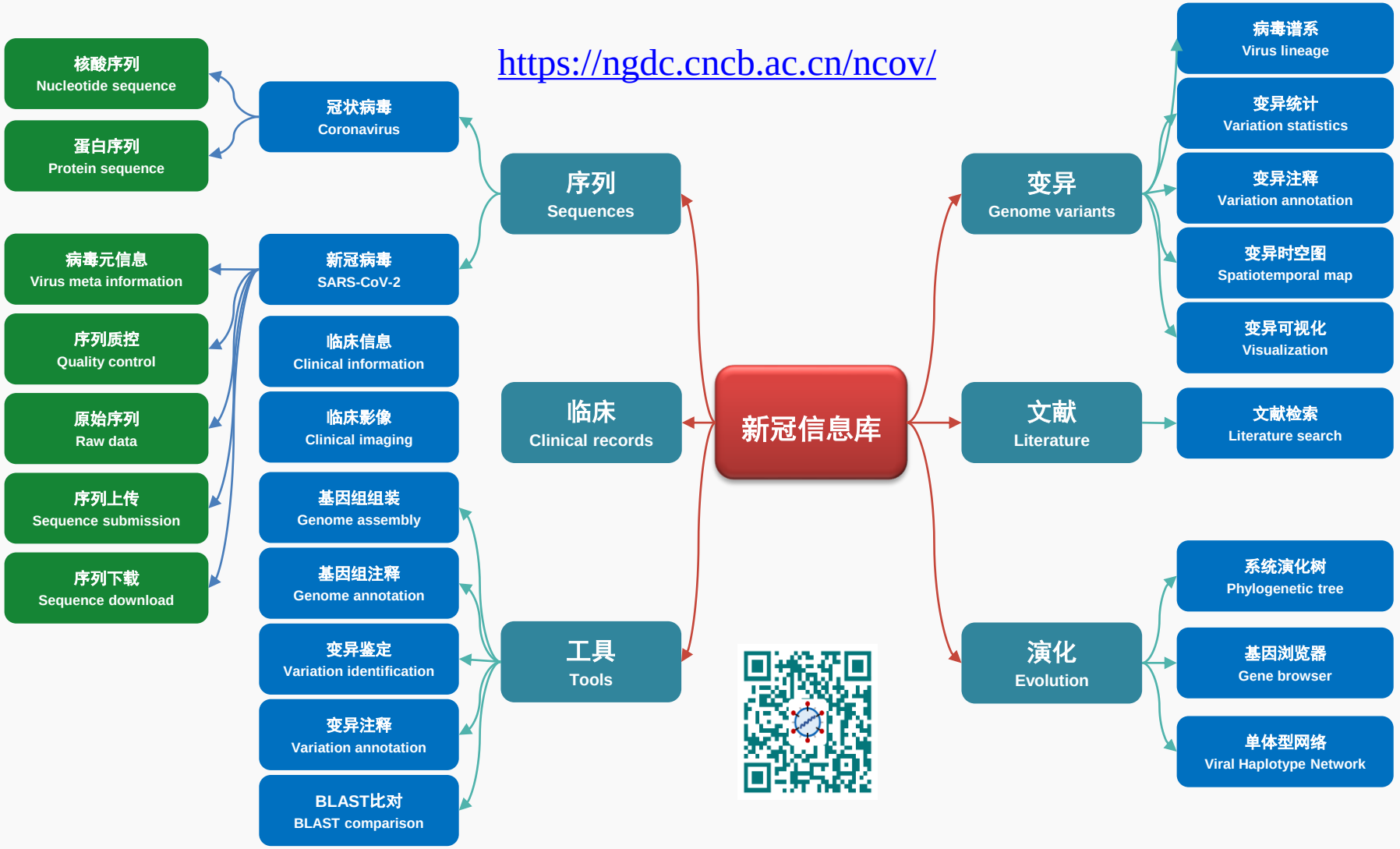


YiChuan 2020; *GPB* 2020, 2021; *ZR* 2020

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

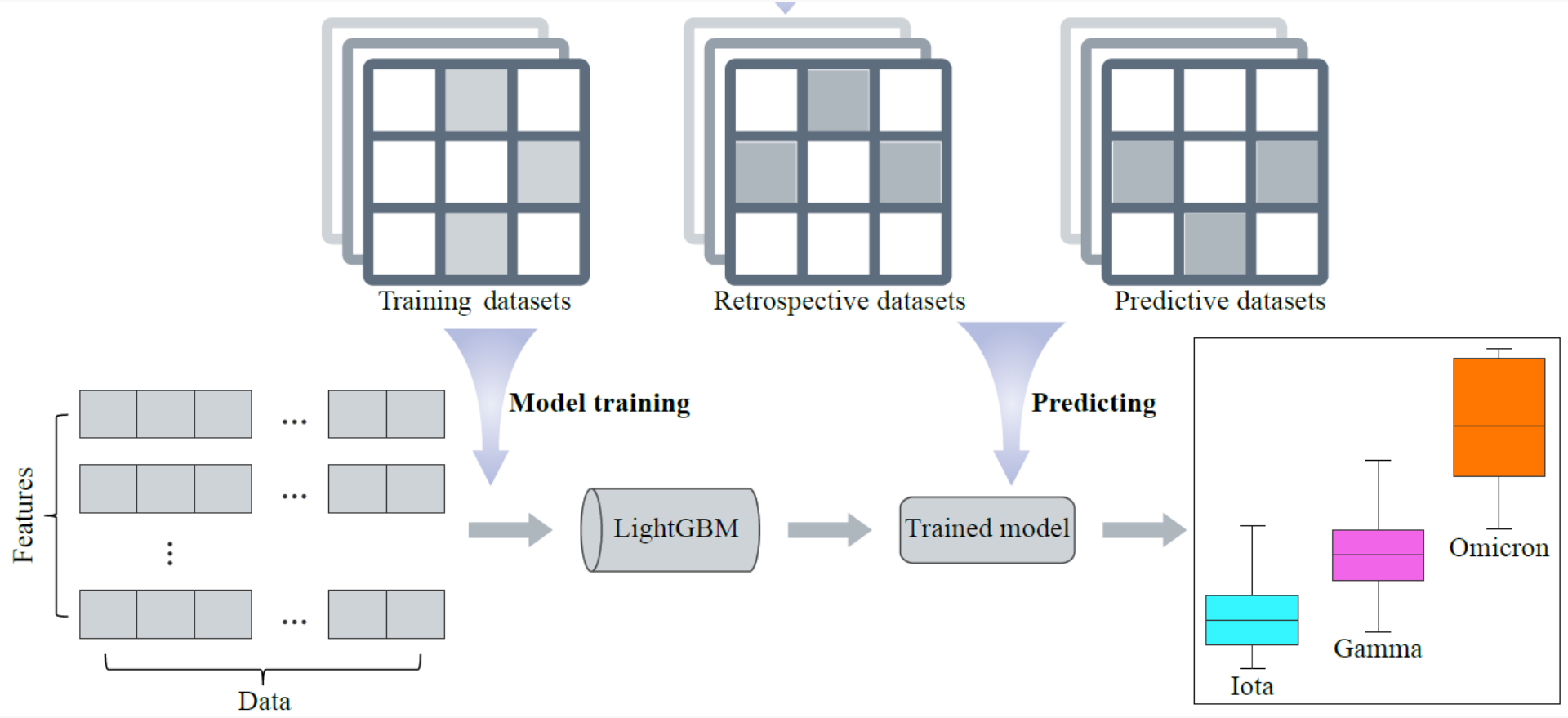
RCoV19 Modules

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

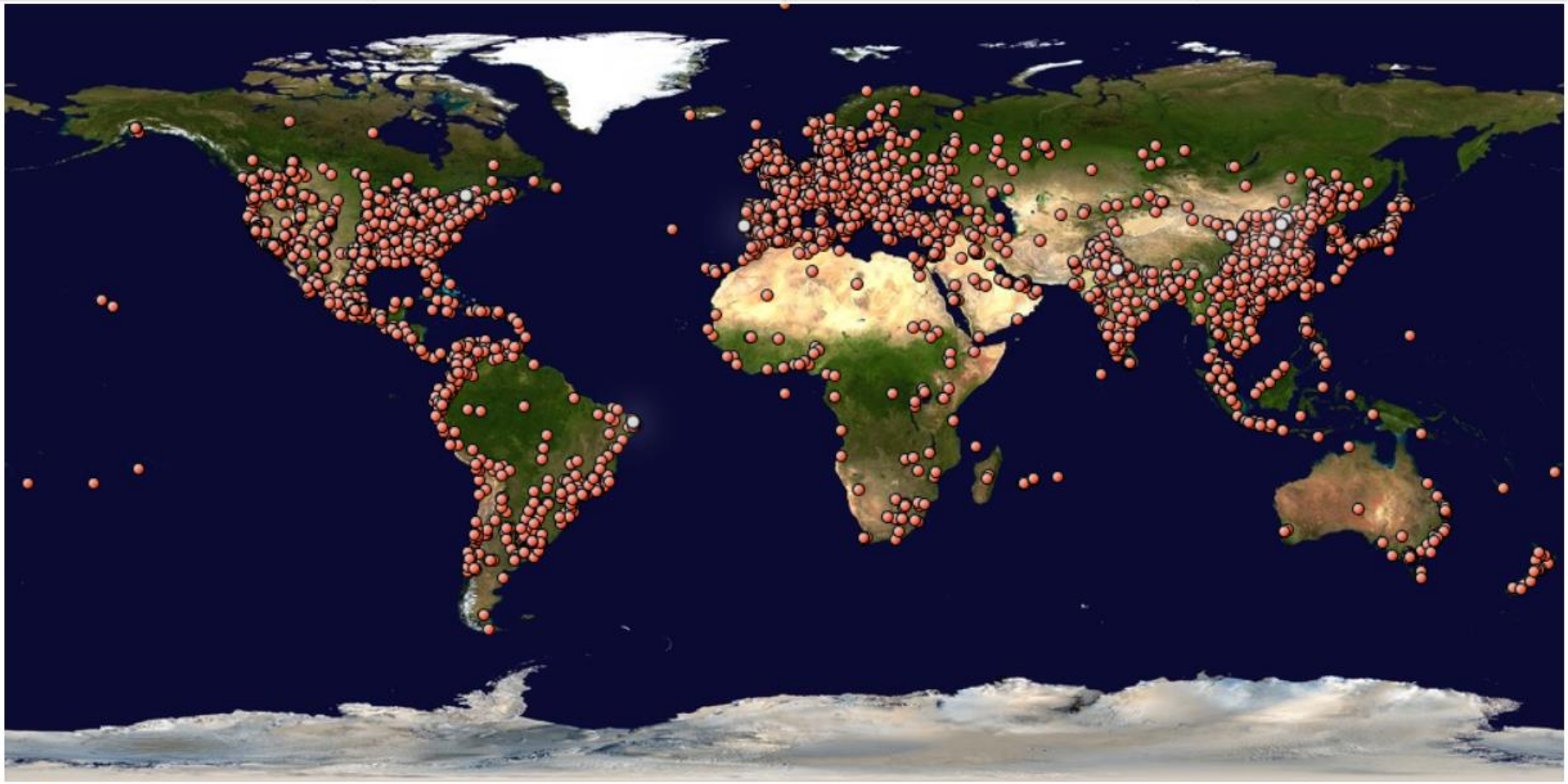


YiChuan 2020, *GPB* 2020, *ZR* (2020)

Early warning for high-risk COVID

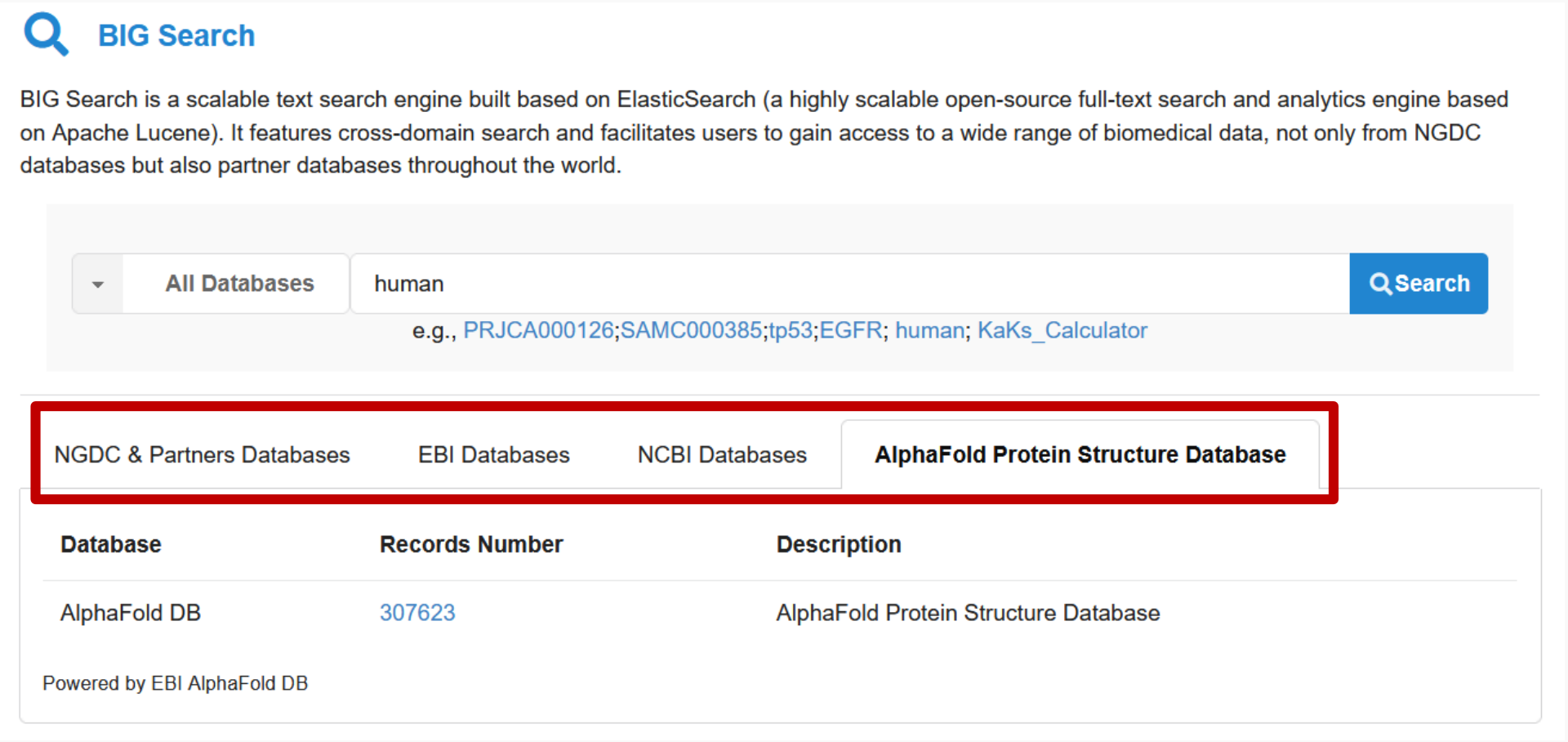


Users across the world



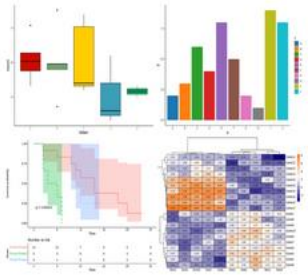
Countries/regions: 181; Users: >2 million; Data downloads: >7 billion

BIG Search: NGDC+EBI+NCBI+AlphaFold



Online Tools

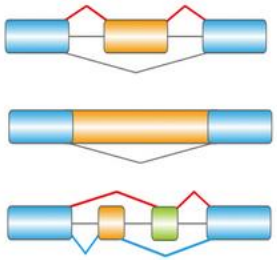
BIT bioinformatics toolkit



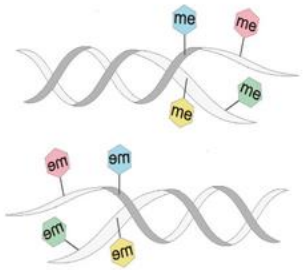
Visualization

Sequence alignment

Composition analysis



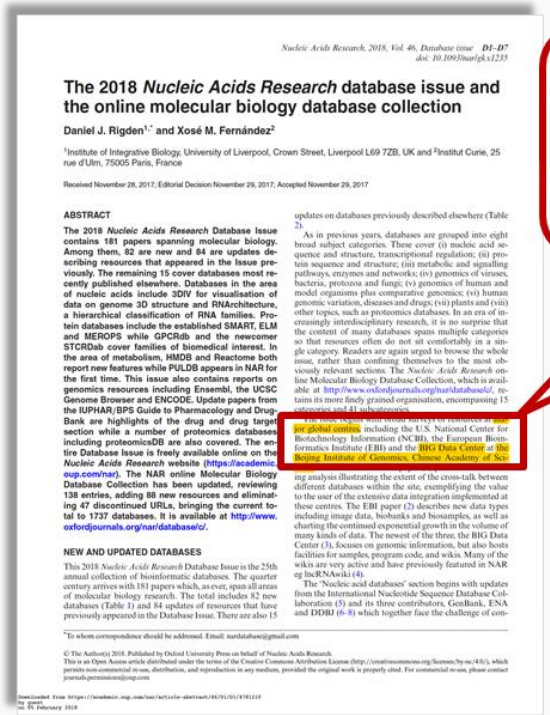
RNA expression



Epigenome analysis

SARS-CoV-2

NGDC became a major global center



Nucleic Acids Research, 2018, Vol. 46, Database issue D1–D7
doi: 10.1093/nar/gkx1235

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/cf, 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-

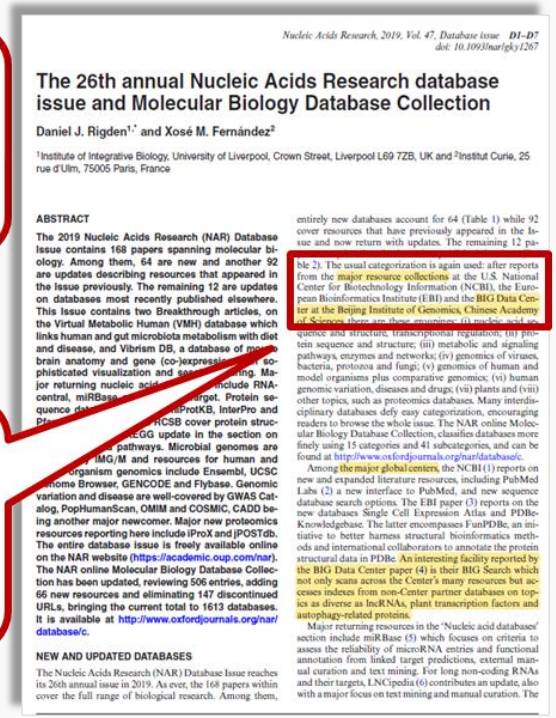
*To whom correspondence should be addressed. Email: nardatabase@gmail.com

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article-abstract/46/D1/D1/4781210
by guest
on 05 February 2018

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-



Nucleic Acids Research, 2019, Vol. 47, Database issue D1–D7
doi: 10.1093/nar/gky1267

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 mouse brain anatomy and gene (co-)expression [illegible] sophisticated visualization and [illegible]ing. Major returning nucleic acid [illegible] include RNAcentral, miRBase [illegible]rget. Protein sequence [illegible]ProtKB, InterPro and [illegible]RCSB cover protein structure [illegible]EGG update in the section on [illegible] pathways. Microbial genomes are [illegible] IMG/M and resources for human and [illegible]organism genomics include Ensembl, UCSC Genome 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 papers (Table 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 2018, 46:D1–D7

Nucleic Acids Research 2019, 47:D1–D7

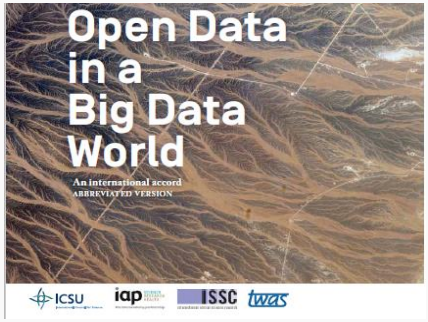
Global Biodiversity and Health Big Data Alliance (BHBD)

About BHBD

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

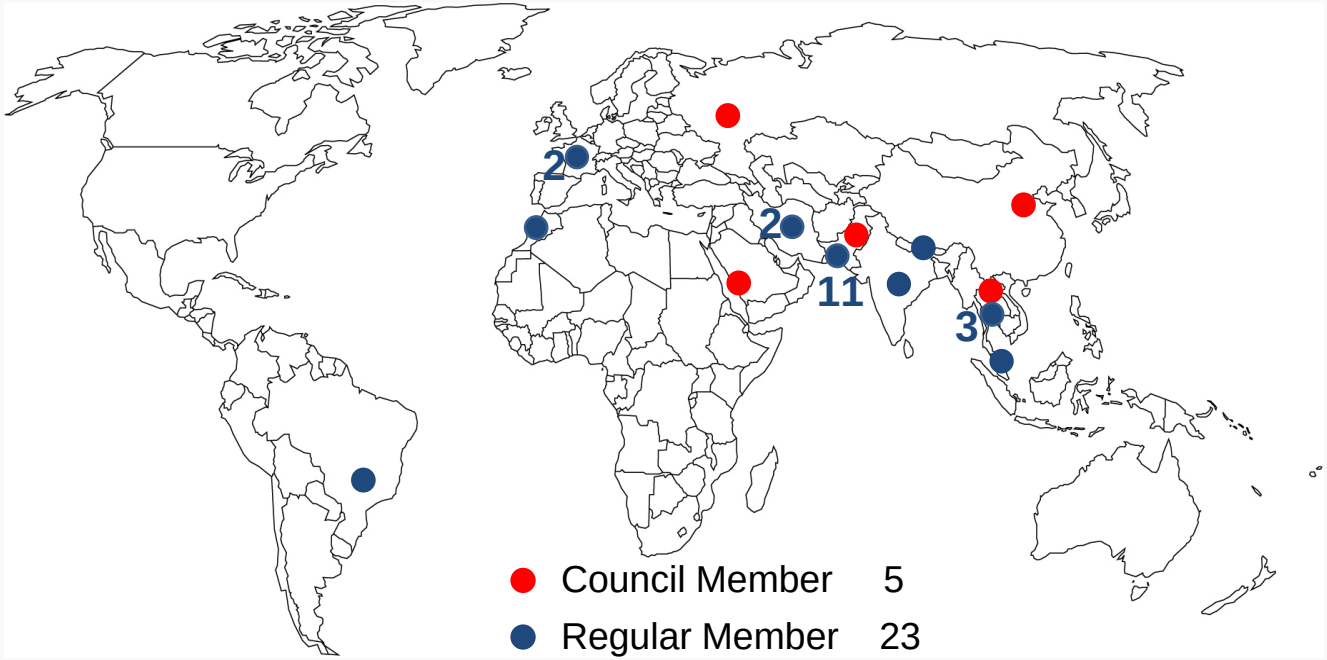
Vision of BHBD

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



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

BHBD Members



28 Members **12** Countries

(As of Jun 2022)

Regular Members:

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

BHBD Achievements

- MoUs signed: 8
- Int'l meetings: 8
- Int'l trainings: 200+ person-time
- Visiting scholars to China: 13 person-time
- Joint research programs: 6 grants obtained
- Data sharing: 300+ datasets
- Joint publications: 10+
- SARS-CoV-2 sample sequencing & analyses: PAK & BRICS

Sequence analysis

Characterization and identification of long non-coding RNAs based on feature relationship

Guangyu Wang[1,2,3,☒], Hongyan Yin[1,2,3,#], Boyang Li[4], Chunlei Yu[1,2,3], Fan Wang[1,2], Xingjian Xu[1,2,3, §], Jiabao Cao[1,2,3], Yiming Bao[1,2], Liguo Wang[5], Amir A. Abbasi[6], Vladimir B. Bajic[7], Lina Ma[1,2,*], Zhang Zhang[1,2,3,*]

of Genomics, Chinese Academy
Genomics, Chinese Academy of
ciences, Beijing 100049, China,
onnecticut 06520, United States,
Medicine, Rochester, Minnesota
f Comparative and Evolutionary
amabad 45320, Pakistan, [7]King
al Bioscience Research Center
ivision (CEMSE), Thuwal 23955-

RESEARCH ARTICLE Open Access

Evolutionary history of the human multigene families reveals widespread gene duplications throughout the history of animals

Nashaiman Pervaiz[1], Nazia Shakeel[1], Ayesha Qasim[1], Rabail Zehra[1], Saneela Anwar[1], Neenish Rana[1], Yongbiao Xue[2], Zhang Zhang[2], Yiming Bao[2*] and Amir Ali Abbasi[1*]

NSFC · 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

RFBR RUSSIAN FOUNDATION FOR BASIC RESEARCH · F A S I E

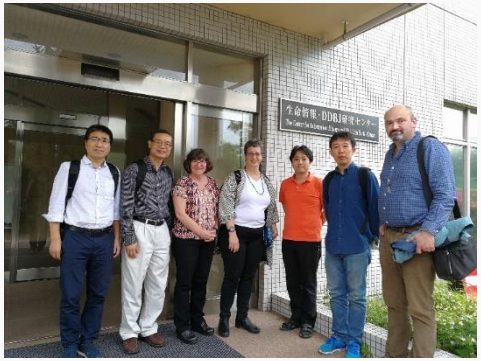
中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

Department of Biotechnology
Ministry of Science & Technology
Government of India

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

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

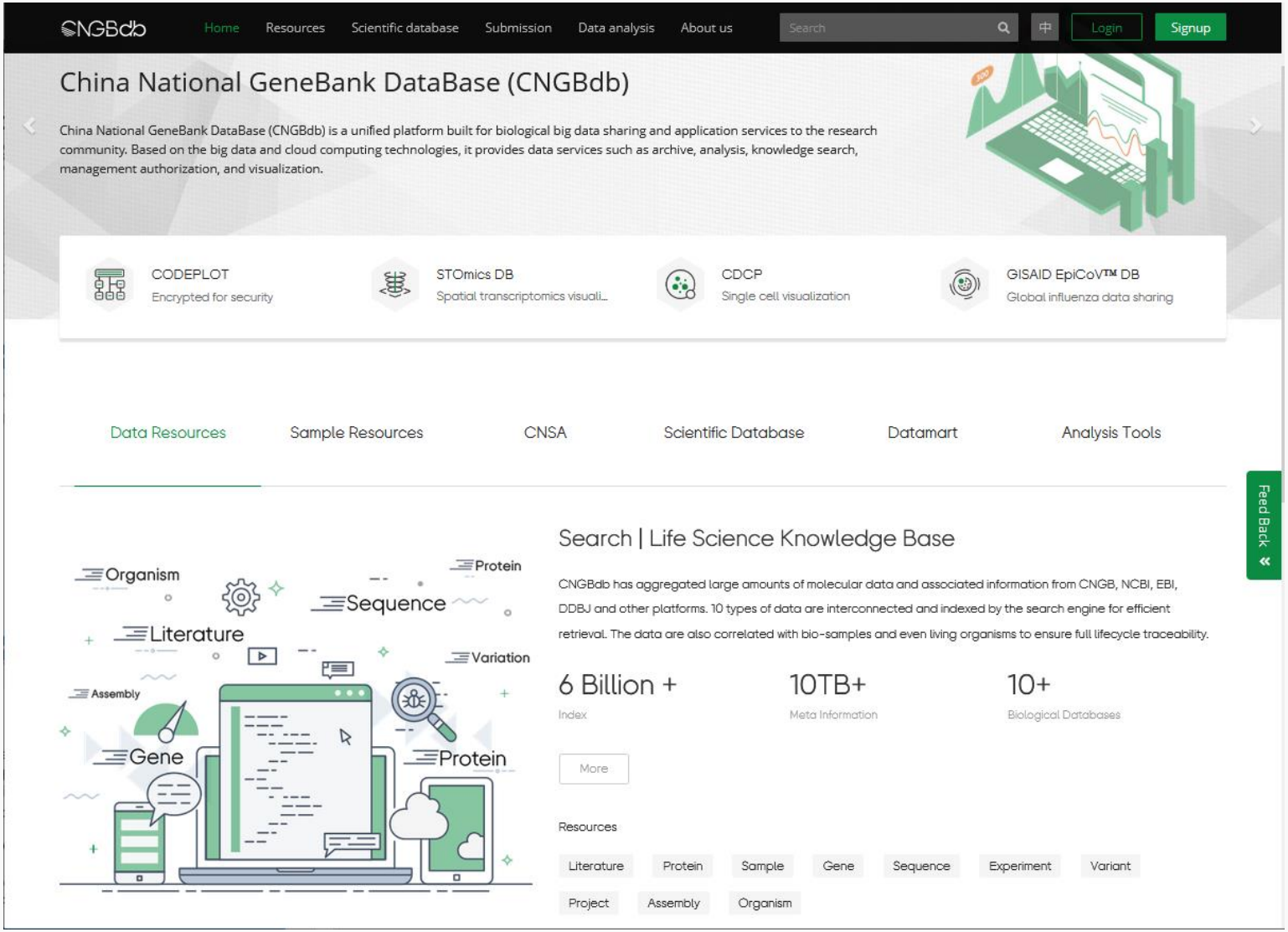
GenBank: MT240479.1

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).
```

Other Genomics Data Resources



China National GeneBank DataBase (CNGBdb)

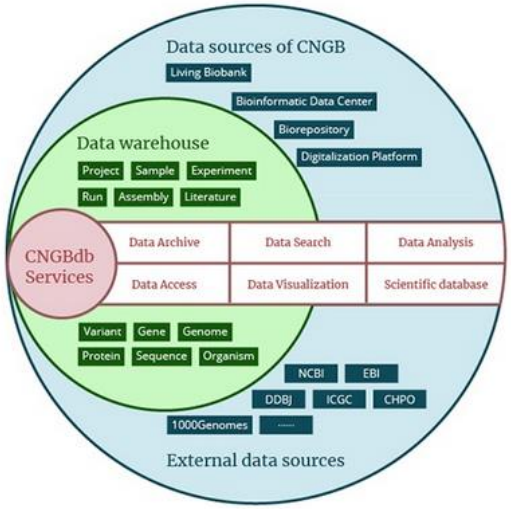
About us | Contribution wall | News | Updates

What is CNGBdb ?

China National GeneBank DataBase (CNGBdb) is a unified platform built for biological big data sharing and application services to the research community. Based on the big data and cloud computing technologies, it provides data services such as archive, analysis, knowledge search, management authorization, and visualization. At present, CNGBdb has integrated large amounts of internal and external molecular data and other information from CNGB, NCBI, EBI, DDBJ, etc., indexed by search, covering 10 data structures. Moreover, CNGBdb correlates living sources, biological samples and bioinformatic data to realize the traceability of comprehensive data.

CNGBdb Data Resources

Based on the internal data source from CNGB and external data source from NCBI, EBI, DDBJ, etc., and following the data standards of international standard alliance such as INSDC, DataCite, GA4GH, GGBN, ACMG, CNGBdb builds data structures covering literature, gene, variation, protein, etc., and provides data sharing and application services such as data archive, query and retrieval, and analysis.



Overview of CNGB

China National GeneBank (CNGB) is one of the key science infrastructures in Shenzhen, approved and funded by the Chinese government. CNGB is committed to supporting public welfare, life science research, innovation and industry development, through effective bioresource conservation, digitalization and utilization.

China National GeneBank (CNGB) is owned by the government and under the guidance and supervision of State High-Tech Industrial Innovation Center, Shenzhen (HIIC).

CNGB is a public, non-profit, open, and enabling platform. It facilitates advanced genomics R&D and technology transfer to industrial application, including precision medicine, agriculture, marine sciences and microbial application, towards its mission of making genetic resources "Owned by All, Completed by All and Shared by All".

Quick Links

1.CNGB Nucleotide Sequence Archive (CNSA)

2.BLAST service of CNGB

3.1, 000 Plants project(OneKP or 1 KP)

4.Millet database (MilletDB)

5.The Microbiome Database (MDB)

6.Data Integration Solution for Systematic Exploration of Cancer Traits(DISSECT)

7.Pan Immune Repertoire Database (PIRD)

8.The Chinese Millionome Database(CMDB)

9.Pathogen Variation Database (PVD)

Popular news

1.China National GeneBank DataBase (CNGBdb) Officially Released *[October...*

2.The latest upgrades to CNSA: large capacity, batch submission, easy to hold your sequ...

3.CNGB Nucleotide Sequence Archive (CNSA) released *[October 25, 2017]*

4.China National GeneBank (CNGB) joins Data Center Alliance and Open Data Center...

5.China National GeneBank (CNGB) series database adding new members *[May 8 , 2018]*

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**

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

Thank you!