National Genomics Data Center

Resources of China National Center for Bioinformation

Yiming Bao

Director
National Genomics Data Center
Beijing, China

ABC 2023
Oct. 11, 2023 • Seoul

中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

AI needs data support

- GPT-3: 175 billion parameters
  - Cost (2020): $4.6 million

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

AI needs data support

DeepMind

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

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

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

Inputs and data sources

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

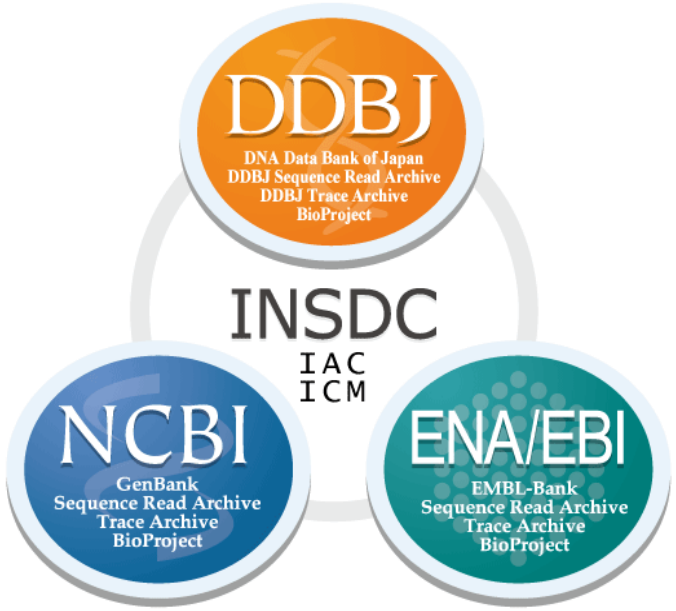
One of the sequence databases used, Big Fantastic Database (BFD), was custom-made and released publicly (see 'Data availability') and was used by several CASP teams. BFD is one of the largest publicly available collections of protein families. It consists of 65,983,866 families represented as MSAs and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

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

Jumper, J et al. Nature (2021).

International Nucleotide Sequence Database Collaboration (INSDC)

Japan,1986



USA, 1988 Europe, 1992

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

Background in China (probably your country too)

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

Large Data Submission to NGDC

Open access Protocol

SVN Stroke and Vascular Neurology

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

10K patients, ~2.3 PB data

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

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

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

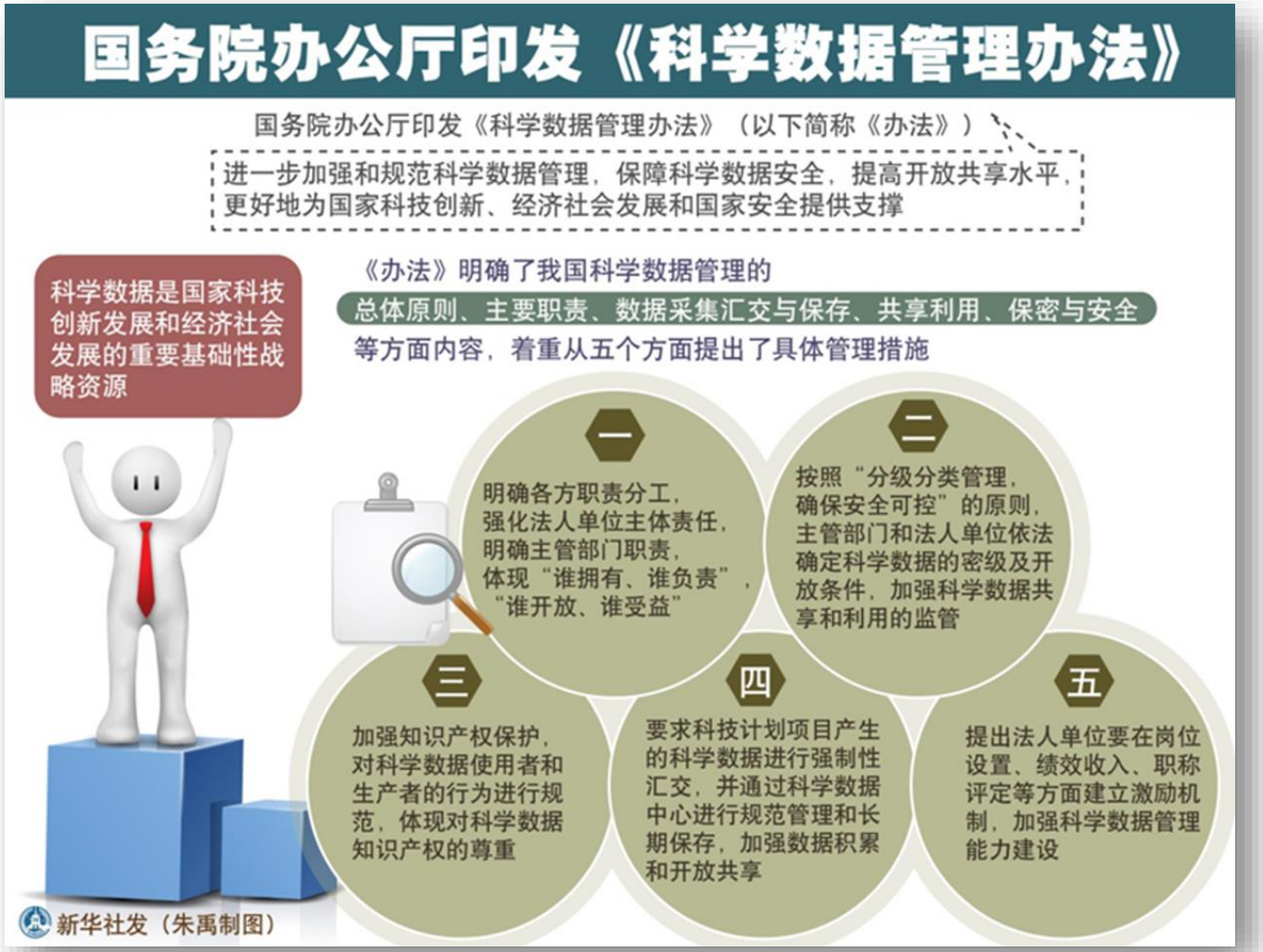
Translating big data into big discoveries

Deposition

Integration

Translation

Measures for the Management of Scientific Data



2018/03

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

Establishment of 20 National Scientific Data Centers

科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知

国科发基（2019）194号

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

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

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

特此通知。

附件：国家科技资源共享服务平台名单

科 技 部　　财 政 部

2019年6月5日

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

National Genomics Data Center (NGDC)

国家科技资源共享服务平台名单

序号	国家平台名称	依托单位	主管部门
1	国家高能物理科学数据中心	中国科学院高能物理研究所	中科院
2	国家基因组科学数据中心	中国科学院北京基因组研究所	中科院
3	国家微生物科学数据中心	中国科学院微生物研究所	中科院
4	国家空间科学数据中心	中国科学院国家空间科学中心	中科院
5	国家天文科学数据中心	中国科学院国家天文台	中科院
6	国家对地观测科学数据中心	中国科学院遥感与数字地球研究所	中科院
7	国家极地科学数据中心	中国极地研究中心	自然资源部
8	国家青藏高原科学数据中心	中国科学院青藏高原研究所	中科院
9	国家生态科学数据中心	中国科学院地理科学与资源研究所	中科院
10	国家材料腐蚀与防护科学数据中心	北京科技大学	教育部
11	国家冰川冻土沙漠科学数据中心	中国科学院寒区旱区环境与工程研究所	中科院
12	国家计量科学数据中心	中国计量科学研究院	市场监管总局
13	国家地球系统科学数据中心	中国科学院地理科学与资源研究所	中科院
14	国家人口健康科学数据中心	中国医学科学院	卫生健康委
15	国家基础学科公共科学数据中心	中国科学院计算机网络信息中心	中科院
16	国家农业科学数据中心	中国农业科学院农业信息研究所	农业农村部
17	国家林业和草原科学数据中心	中国林业科学研究院资源信息研究所	林草局
18	国家气象科学数据中心	国家气象信息中心	气象局
19	国家地震科学数据中心	中国地震台网中心	地震局
20	国家海洋科学数据中心	国家海洋信息中心	自然资源部

China National Center for Bioinformation

中国科学院文件

科发人字〔2019〕105号

中国科学院关于中国科学院北京基因组研究所加挂国家生物信息中心牌子的通知

院属各单位、院机关各部门：

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

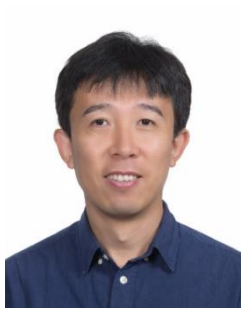
2019年12月16日

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

The Team

❑ Steering Advisors


❑ Professors

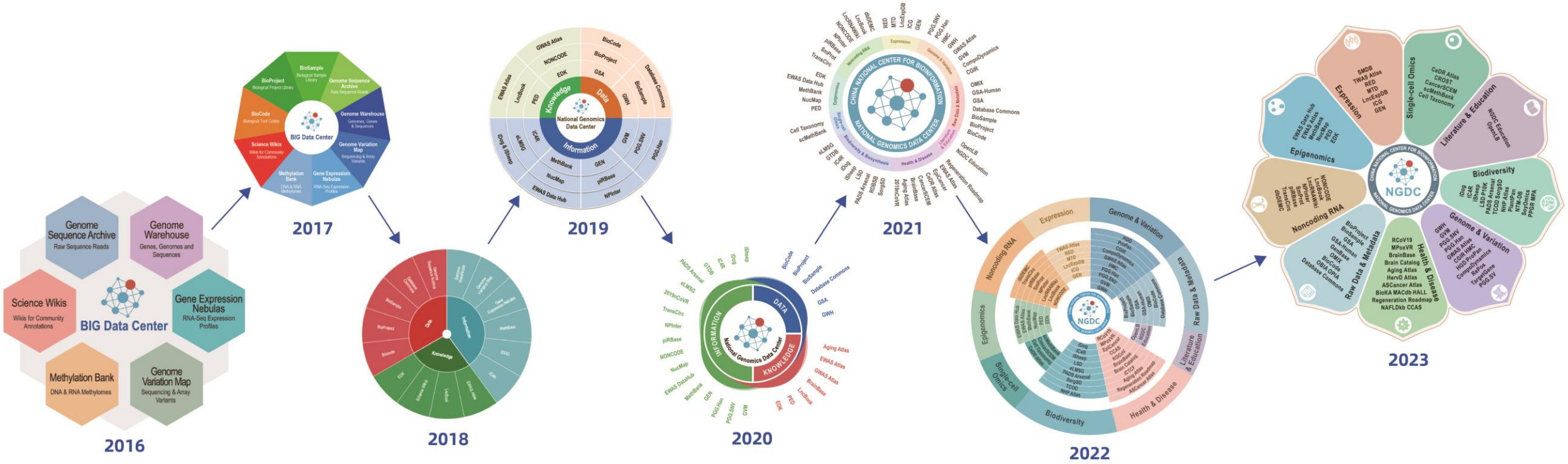






67 students

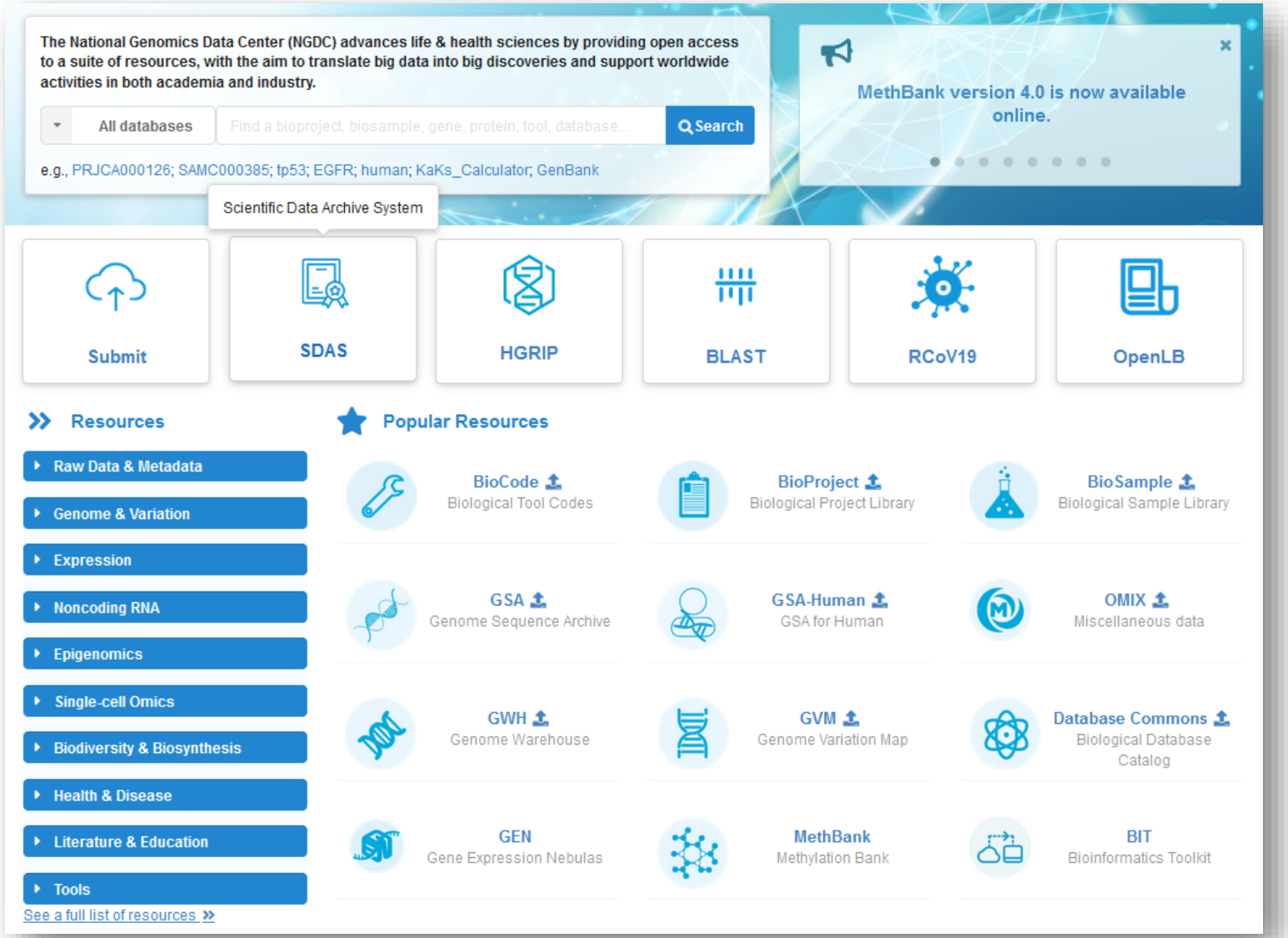
53 Staff

The growing of capability



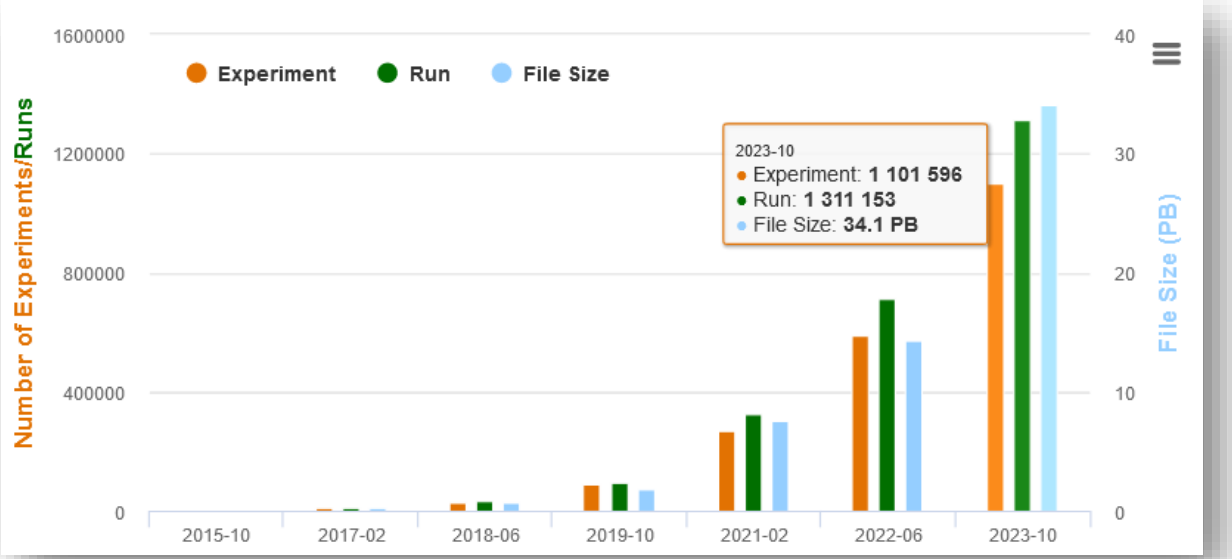
***Nucleic Acids Research*: 2017, 2018, 2019, 2020, 2021, 2022, 2023**

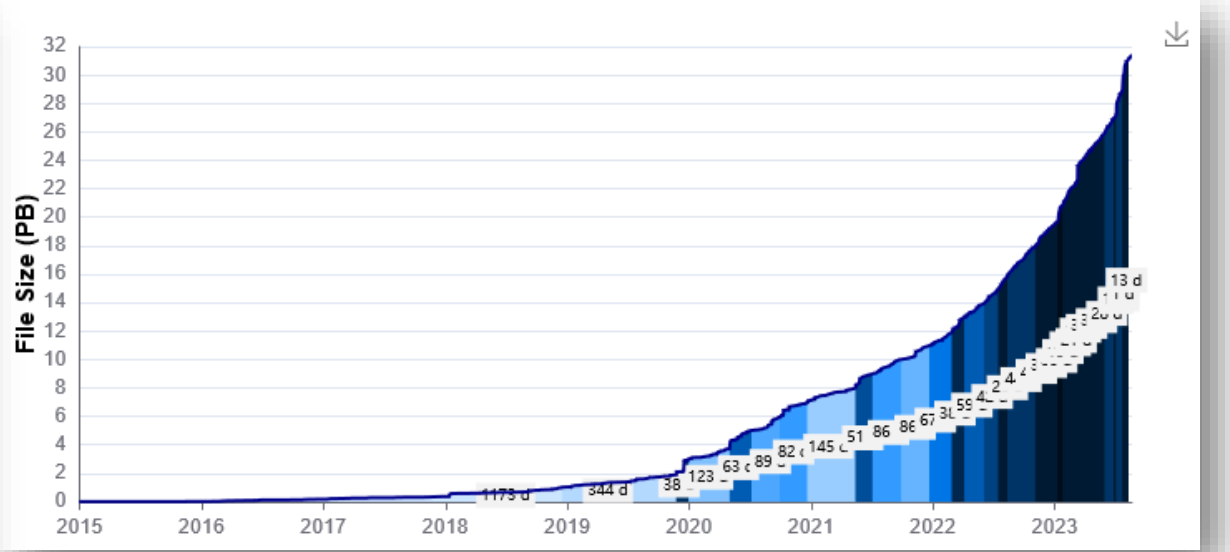
Comprehensive Resources at CNCB-NGDC



- Omics databases
  - BioProject
  - BioSample
  - Genome Sequence Archive (GSA)
  - GenBase
  - Genome Warehouse (GWH)
  - Gene Expression Nebulas (GEN)
  - Genome Variation Map (GVM)
  - Methylation Bank (MethBank)
- Specialized databases
  - RCoV19
  - IC4R
  - DogSD
  - LncRNAWiki
  - Database Commons
- Literatures
  - OpenLB
- Tools
  - BLAST
  - BIT

Rapid Data Growth

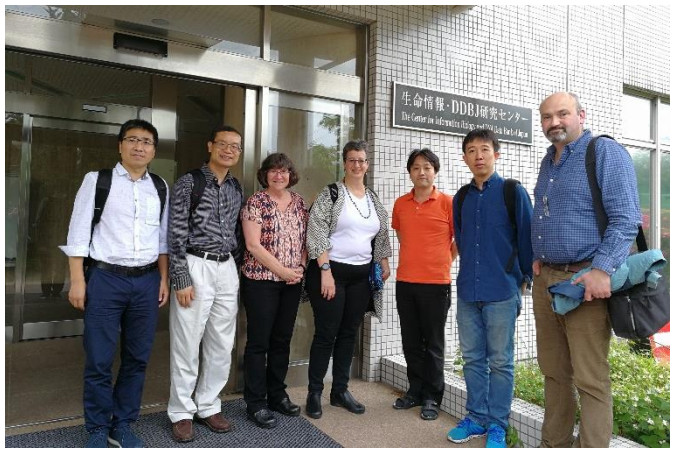




>34 PB as of 2023-10-11

Collaborations with INSDC

DDBJ

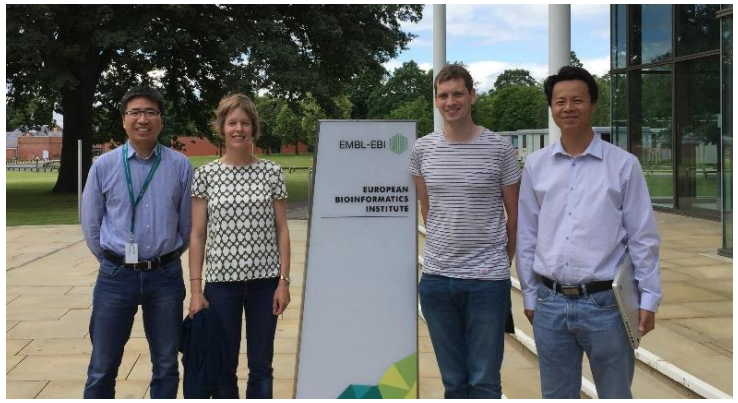


2017, 2020, 2023
INSDC Annual Meetings

NCBI

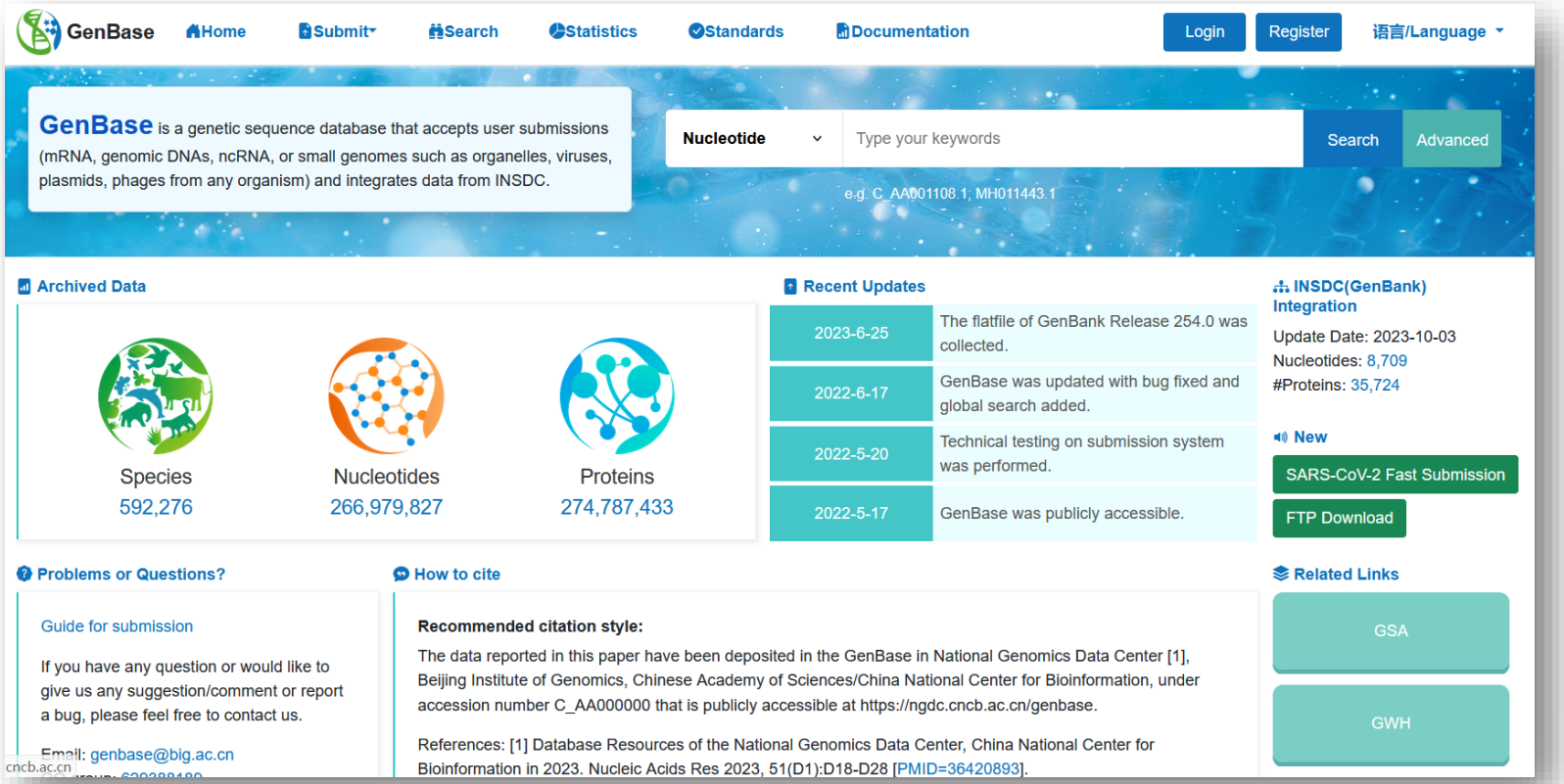
2017, 2018, 2021
Visit and training

EBI

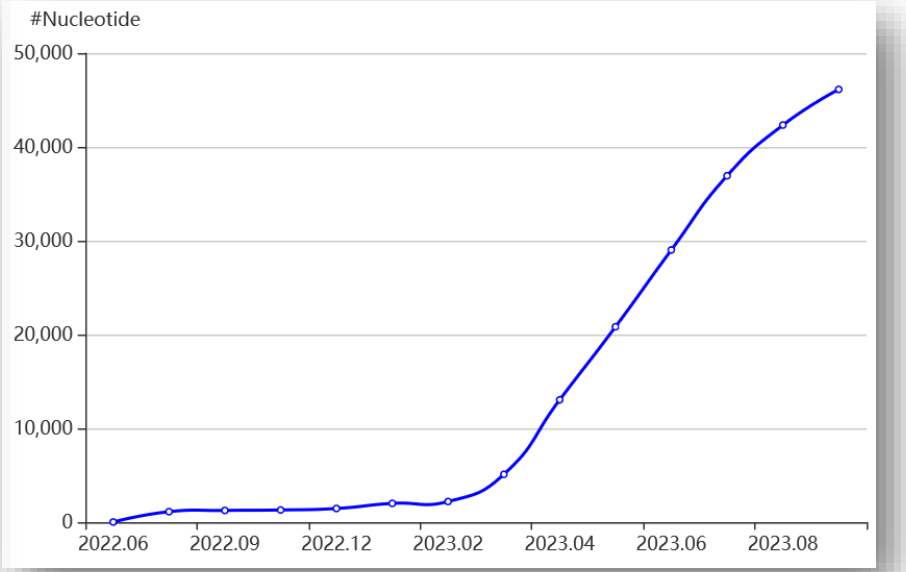


2016, 2019, 2022
Visit and INSDC meeting

GenBase in sync with GenBank



Direct submissions



- **GenBank Release 254.0** has been integrated, with daily updates
- **In total**: **592,276** Species, ~**267 mil.** Nucleotides, ~**274 mil.** Proteins
- **Direct submissions**: **46 k** Nucleotides, **484 k** Proteins

Data Sharing with NCBI

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

GenBank: MT240479.1

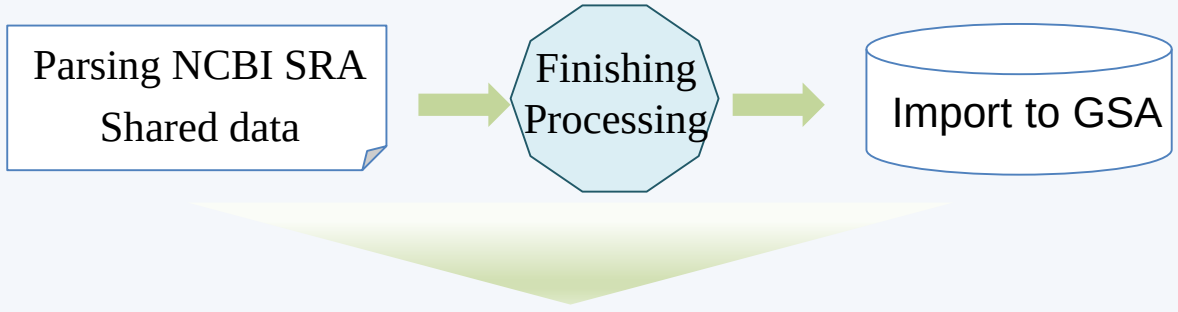
FASTA Graphics

Go to:

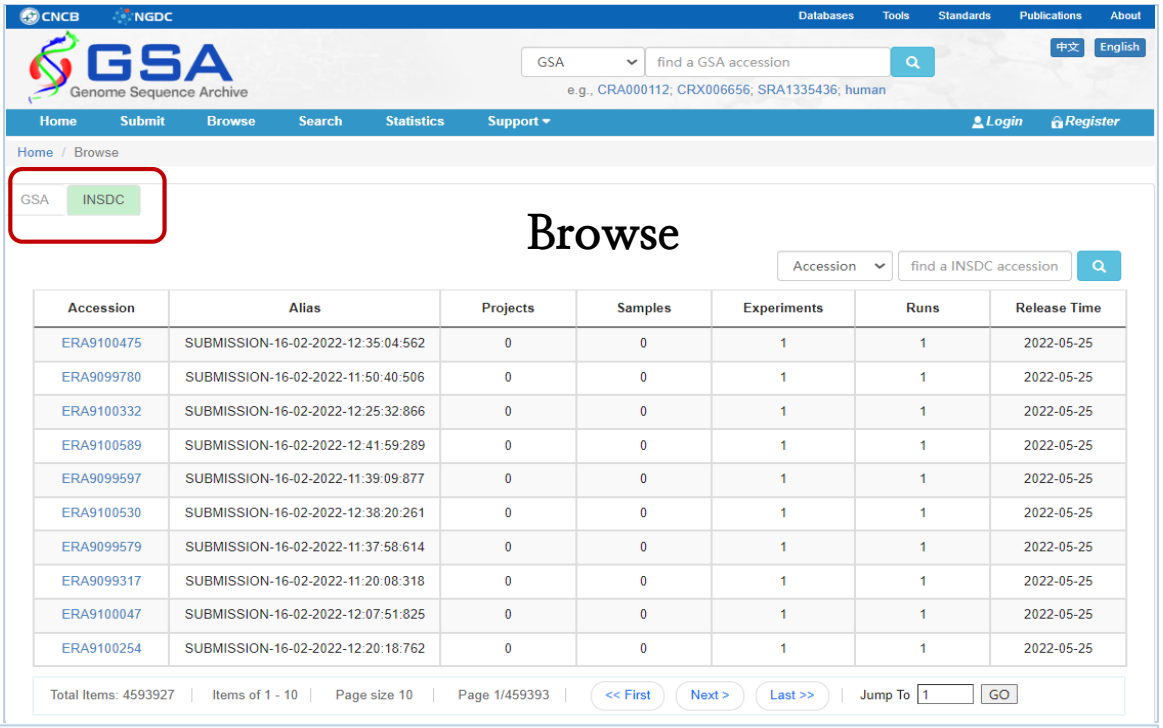
```
LOCUS       MT240479                29836 bp    RNA     linear   VRL 25-MAR-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate
            SARS-CoV-2/Gilgit1/human/2020/PAK, complete genome.
ACCESSION   MT240479 GWHACDD01000001
VERSION     MT240479.1
KEYWORDS    .
SOURCE   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

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

Integration of International Data - GSA



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



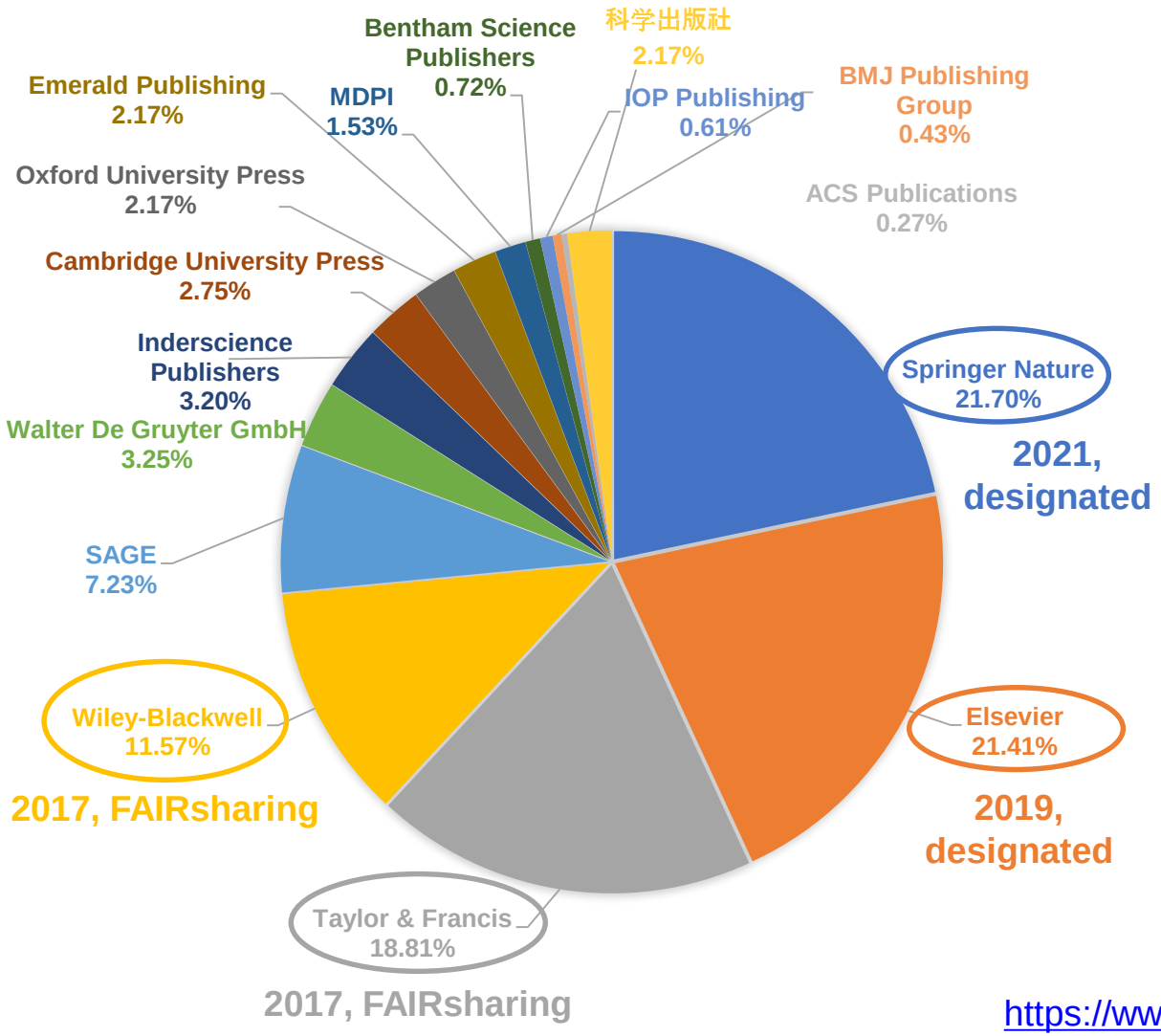
Accession	Alias	Projects	Samples	Experiments	Runs	Release Time
ERA9100475	SUBMISSION-16-02-2022-12:35:04:562	0	0	1	1	2022-05-25
ERA9099780	SUBMISSION-16-02-2022-11:50:40:506	0	0	1	1	2022-05-25
ERA9100332	SUBMISSION-16-02-2022-12:25:32:866	0	0	1	1	2022-05-25
ERA9100589	SUBMISSION-16-02-2022-12:41:59:289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

Metadata information has been updated regularly

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

Data Files： **~5 PB**

GSA Endorsed by Springer Nature and Major Publishers



SPRINGER NATURE

< Research data

Research data policies

Nucleic acid sequence & Omics

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

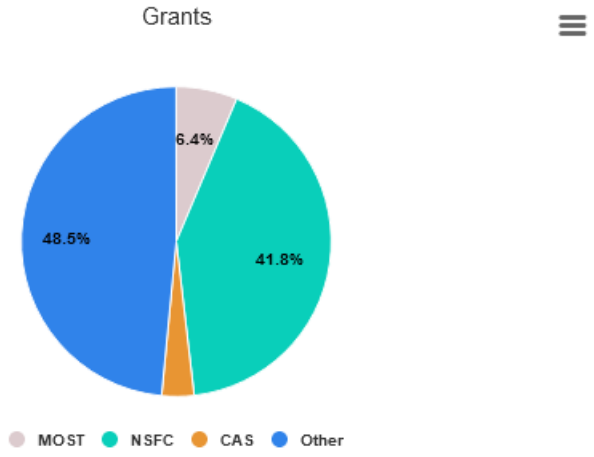
Data types	Repositories
DNA sequence data*	Any INSDC member repository
RNA sequence data*	Genome Sequence Archive (GSA)
Genome assembly data*	
Genetic variation data	dbSNP (human variations less than 50bp)
	dbVar (human variations greater than 50bp)
	European Variation Archive (EVA) (all species)
	Genome Sequence Archive for Human (human variation)

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

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

Supporting >15k Research Grants

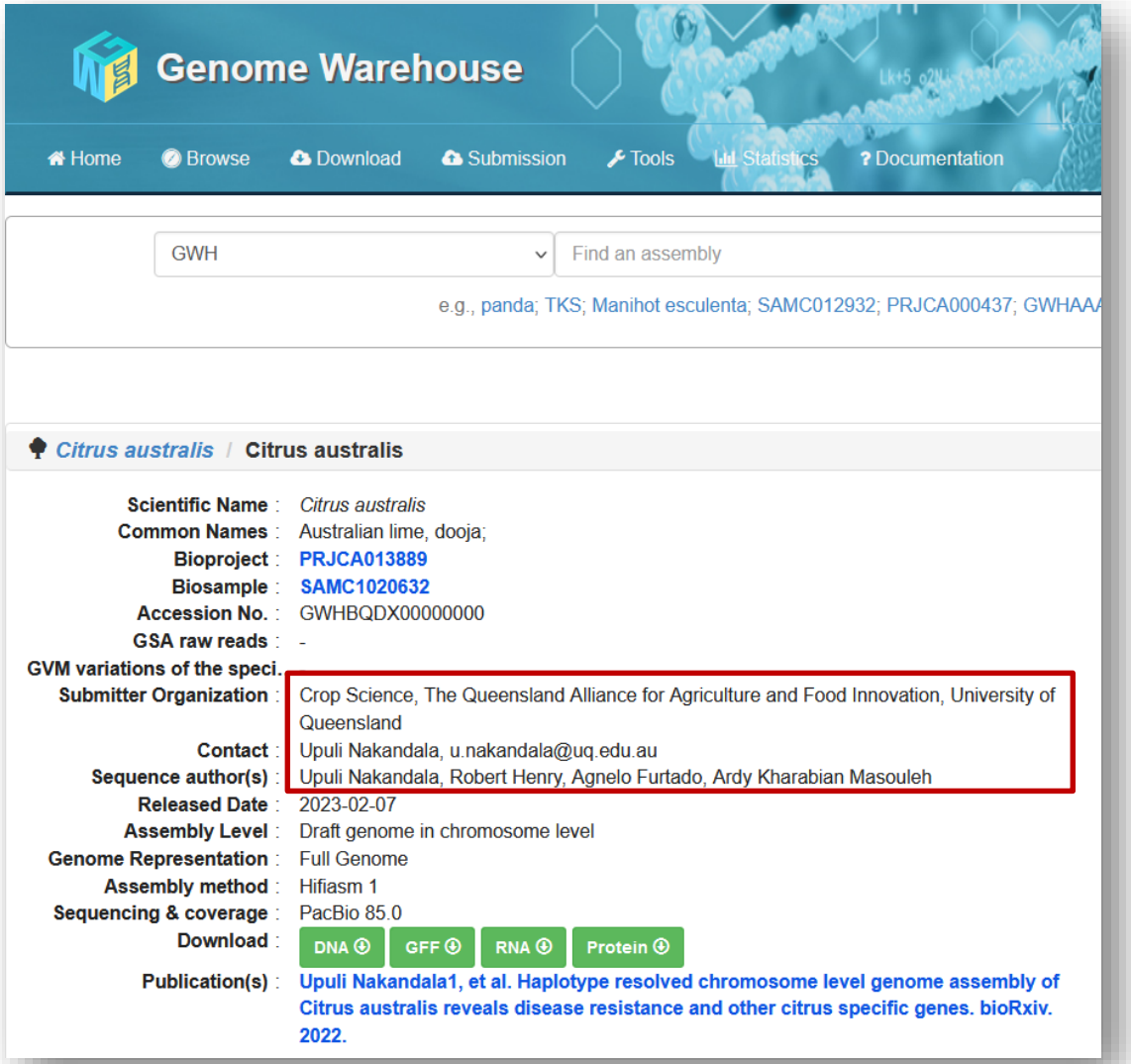
Total Holdings | Sample Types | Organisms | Platforms | Organizations | Downloads | Journals | Articles | Agencies | Global Visits

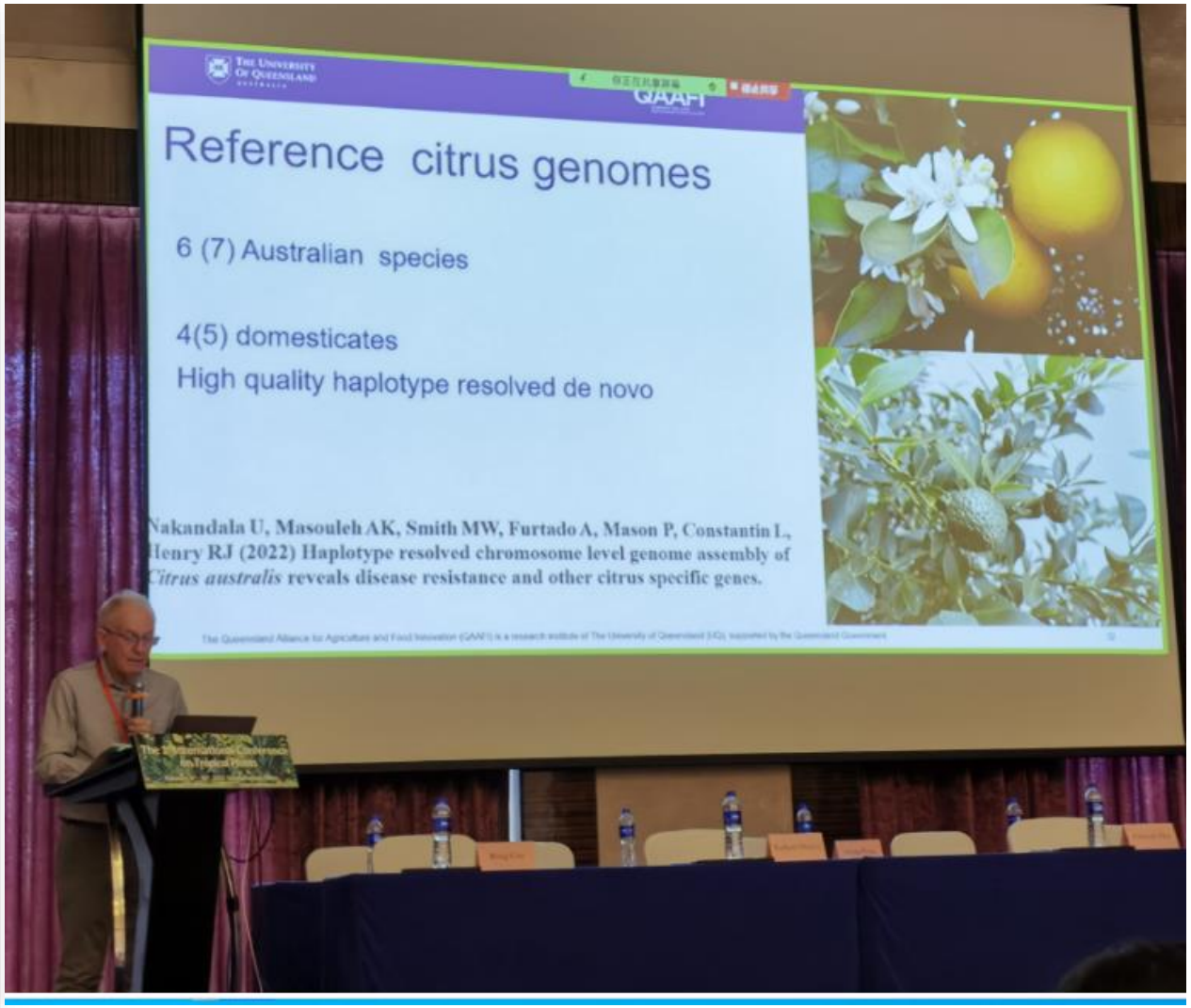


Search:

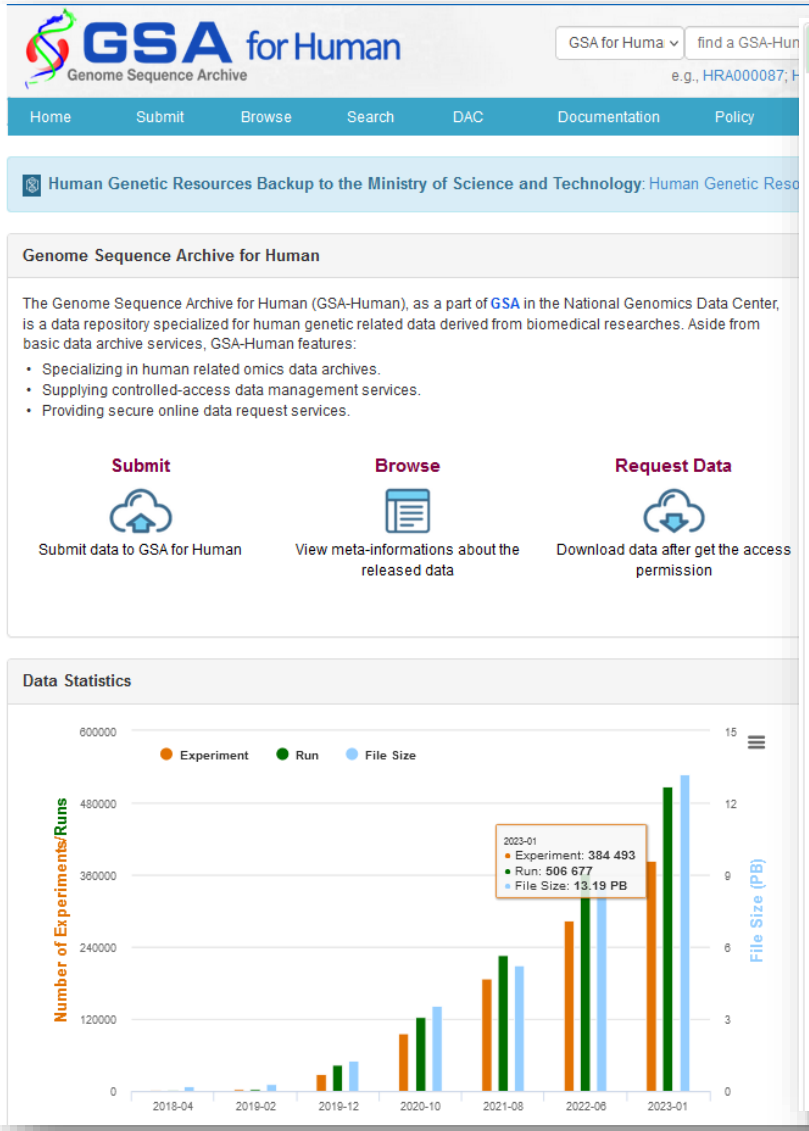
Agencies	Grants	GSAs	Experiments	Runs
Ministry of Science and Technology of the People's Republic of China (MOST)	967	685	45350	48585
National Natural Science Foundation of China (NSFC)	6327	2562	122430	130721
Chinese Academy of Sciences (CAS)	508	563	31115	35526
Others	7343	703	55082	61504

International Submitters from 22 countries





GSA for Human Database – Controlled Access



Available | Unavailable

Filter:

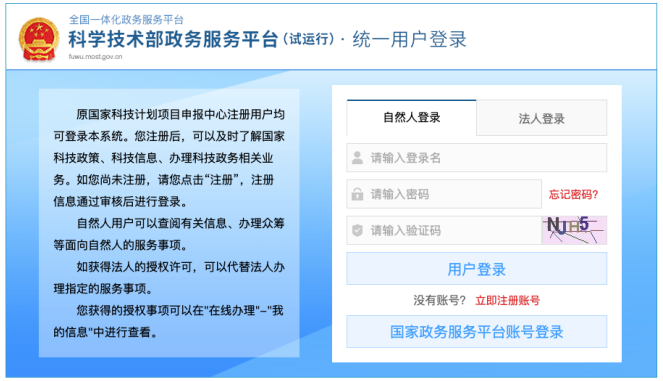
Study	Title	Organization	DAC	Access	Requests	Effective Requests	Approved	Sharing Rank	Last Processed	Request
HRA000150	Single-cell immunological landscape of peripheral blood mononuclear cells of patients with COVID-19 disease	National Clinical Research Center for Infectious Diseases	HDAC000089	Controlled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

Human Data Backup & Registration Protocol

Backup Center

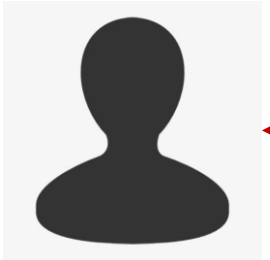
Backup

Backup ID



Registration ID

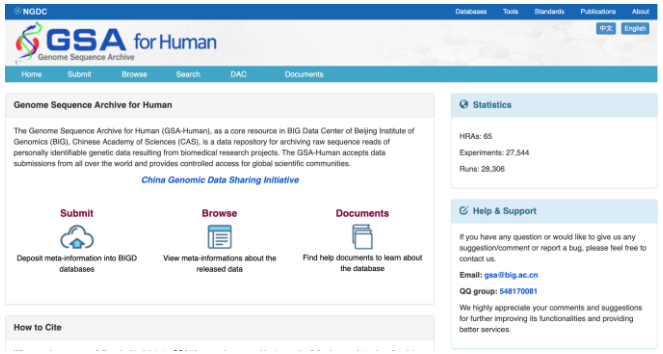
Registration Center



Researchers

Submission

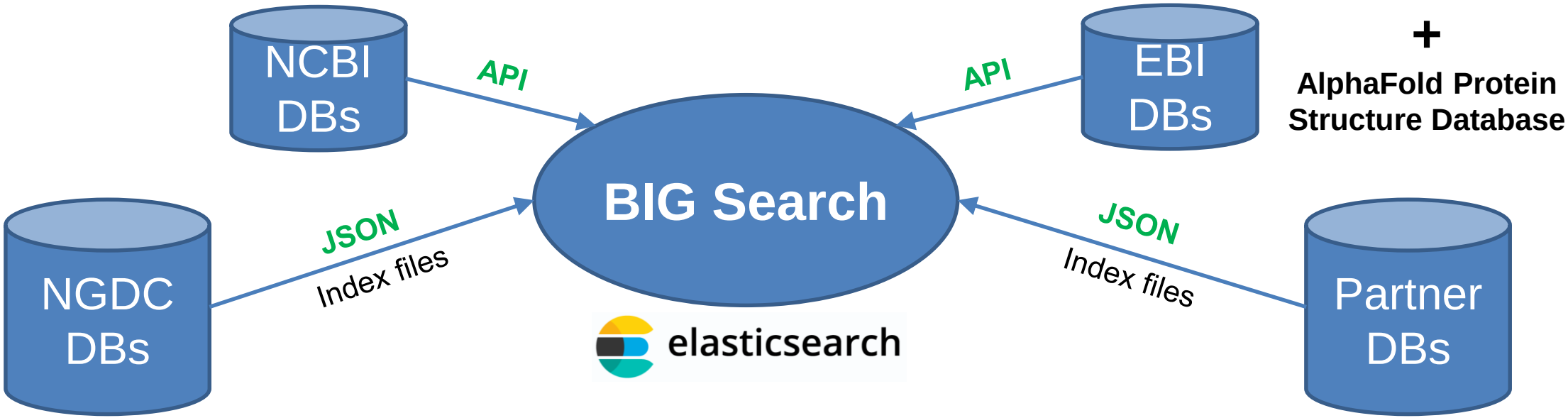
Accession #s



Backup ID

CNCB-NGDC GSA-Human

Cross-database search engine： BIG Search



Partner databases

“Google” for biology data

BIG Search

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

All Databases | human | Search

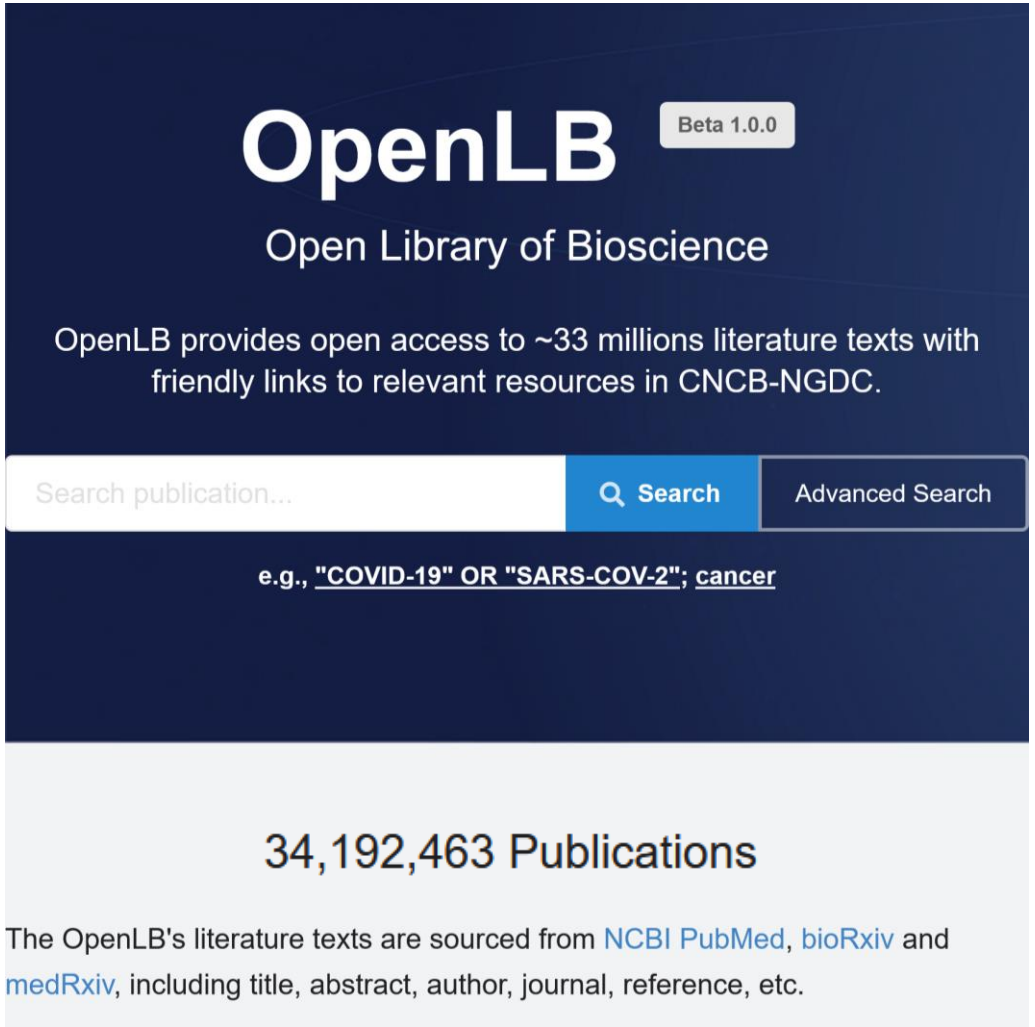
e.g., PRJCA000126;SAMC000385;tp53;EGFR; human; KaKs_Calculator

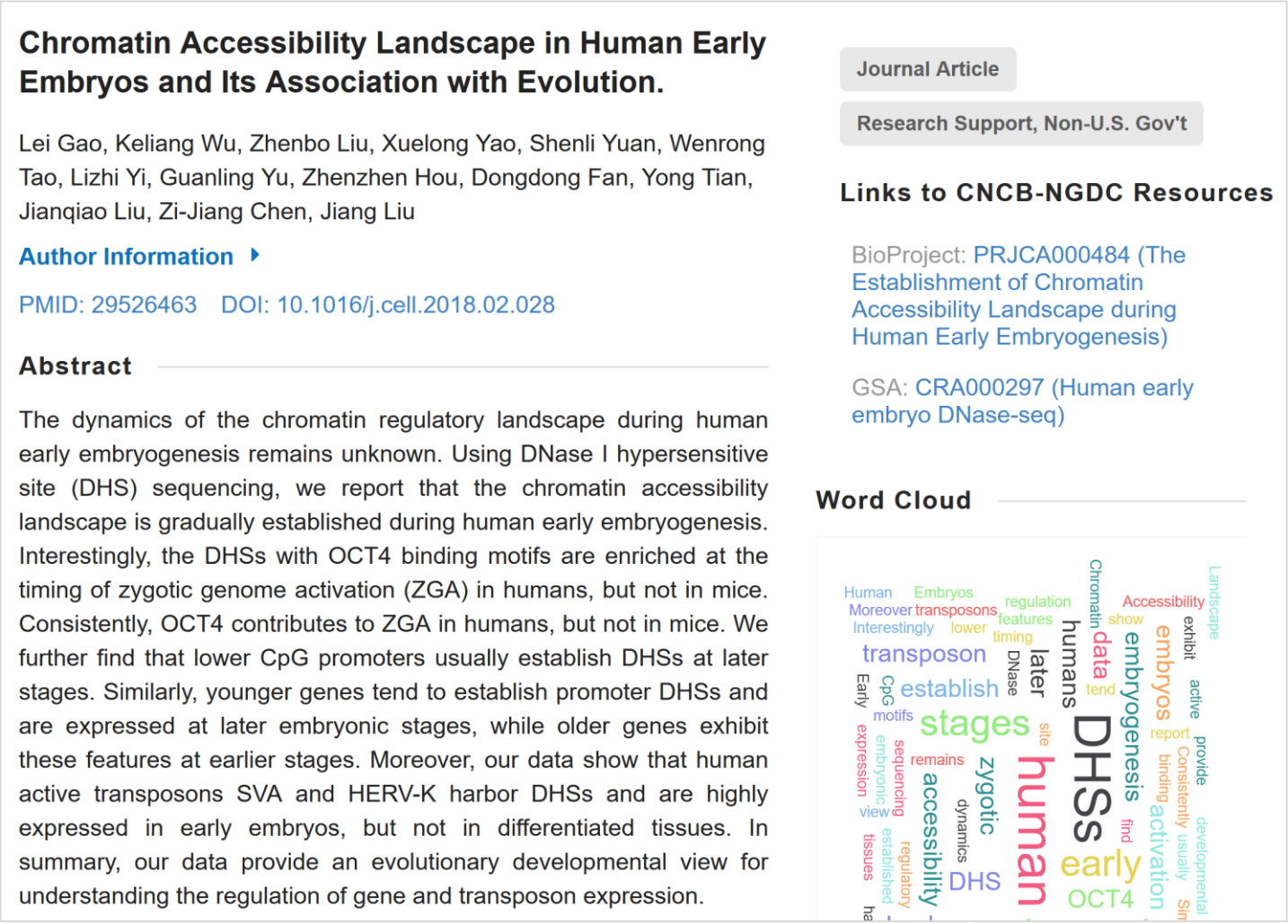
NGDC & Partners Databases | EBI Databases | NCBI Databases | **AlphaFold Protein Structure Database**

Database	Records Number	Description
AlphaFold DB	307623	AlphaFold Protein Structure Database

Powered by EBI AlphaFold DB

Literatures： Open Library of Bioscience





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

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

Author Information

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

Abstract

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

Journal Article

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

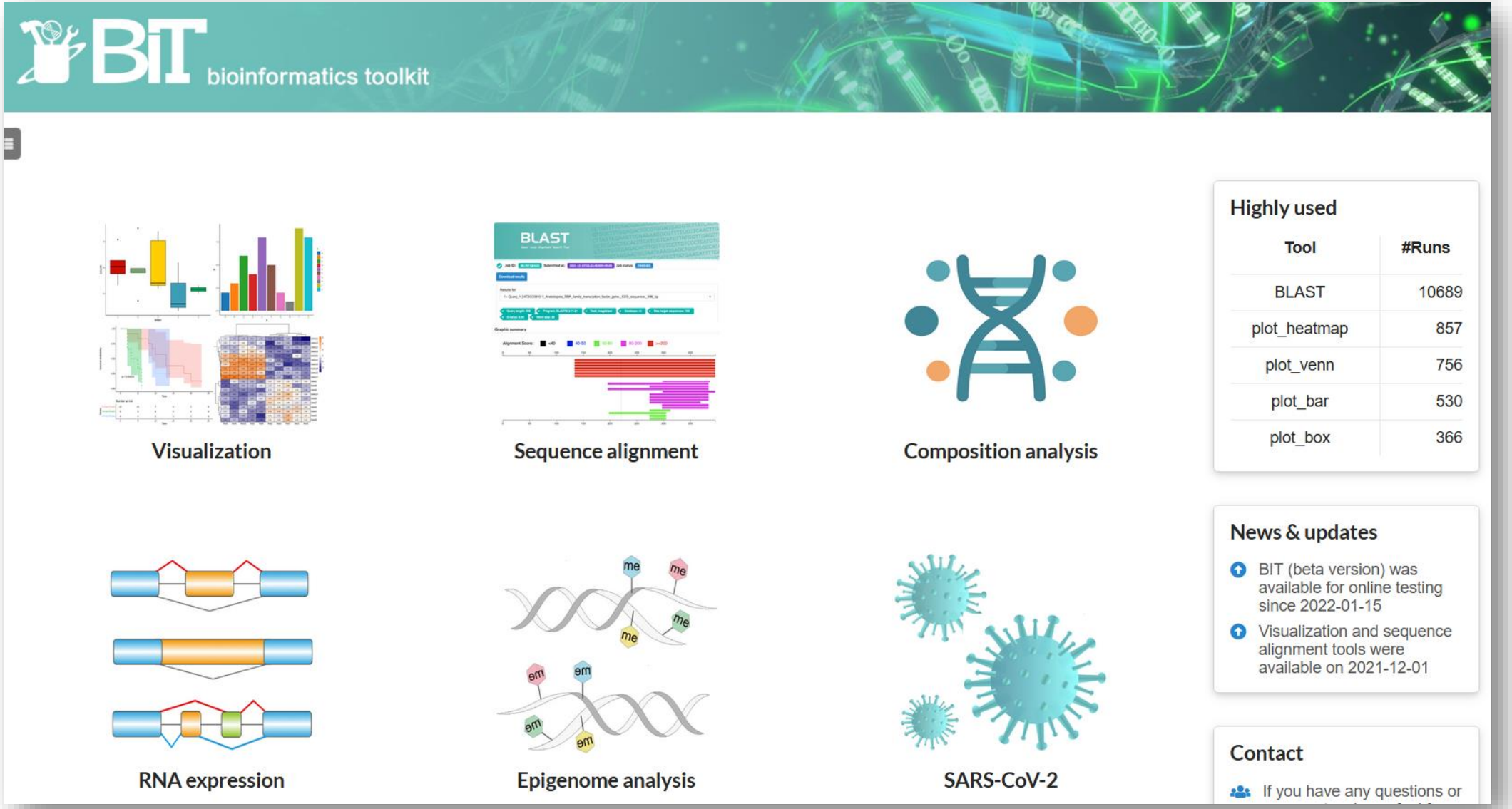
Links to CNCB-NGDC Resources

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

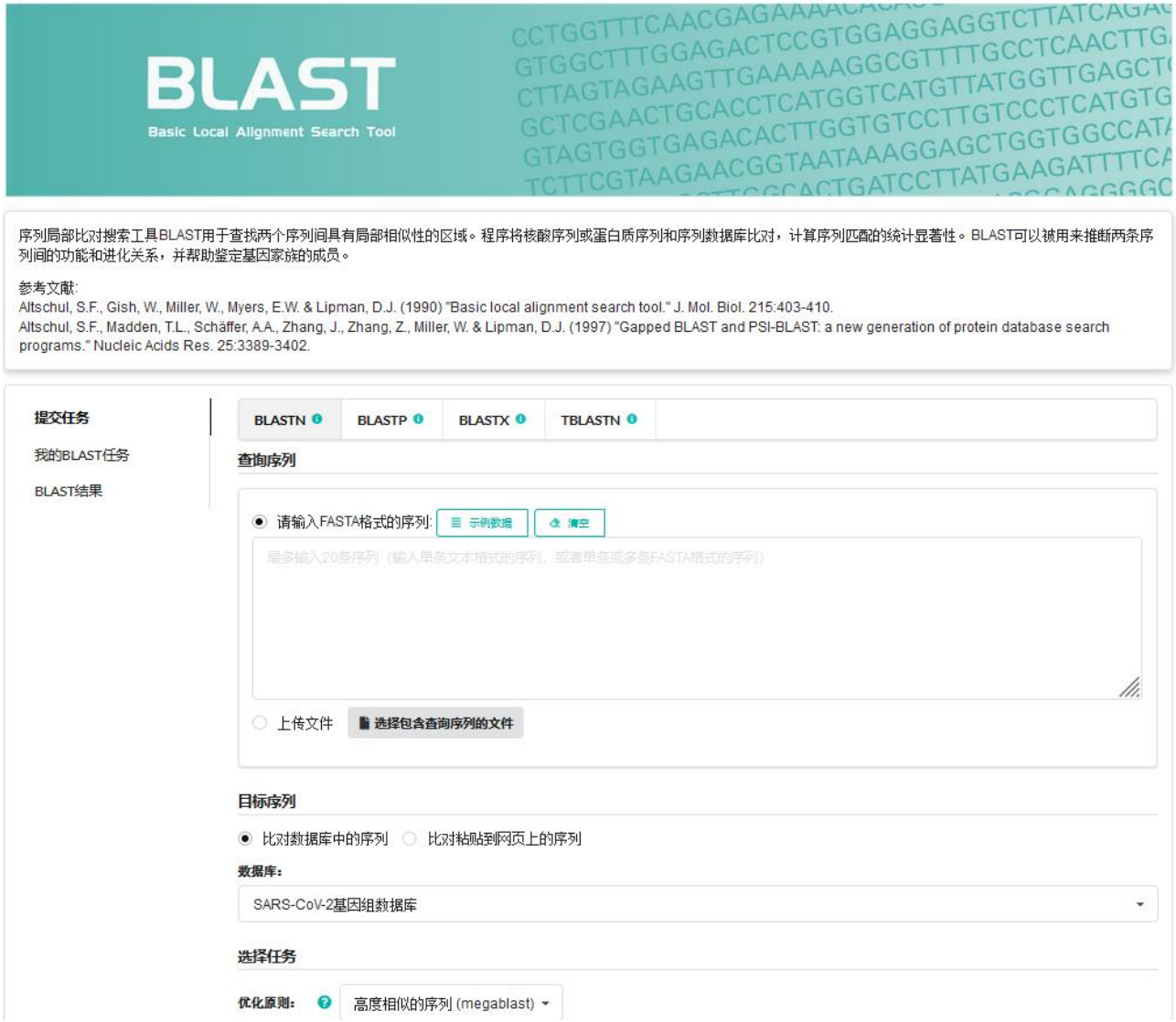
GSA: CRA000297 (Human early embryo DNase-seq)

Word Cloud

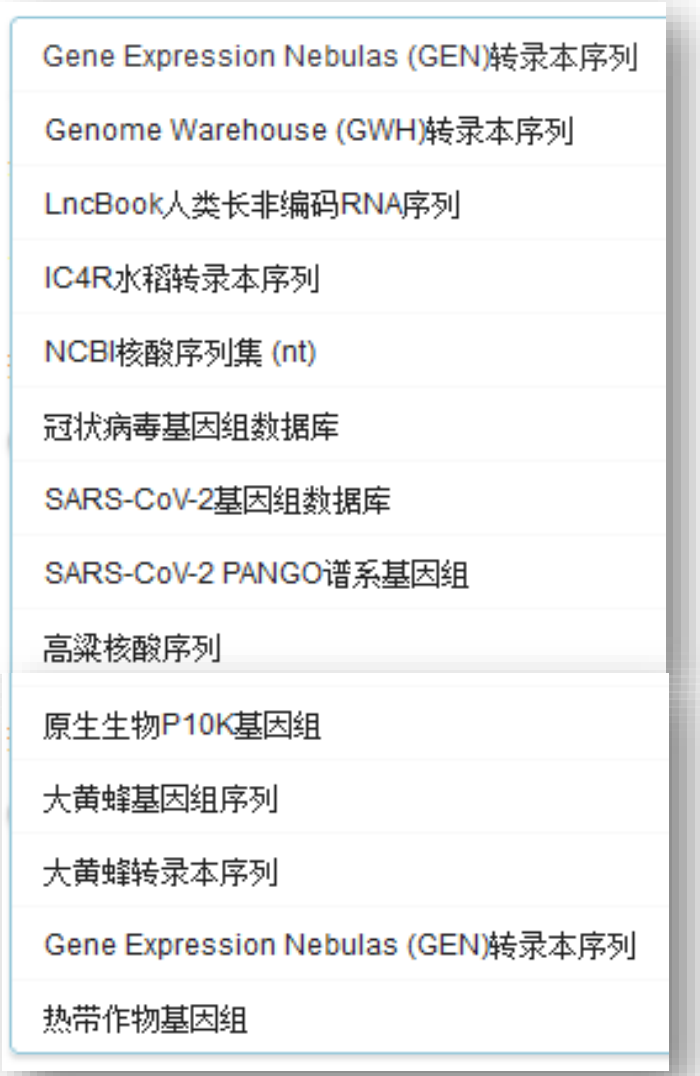
Bioinformatics Tolls - BIT



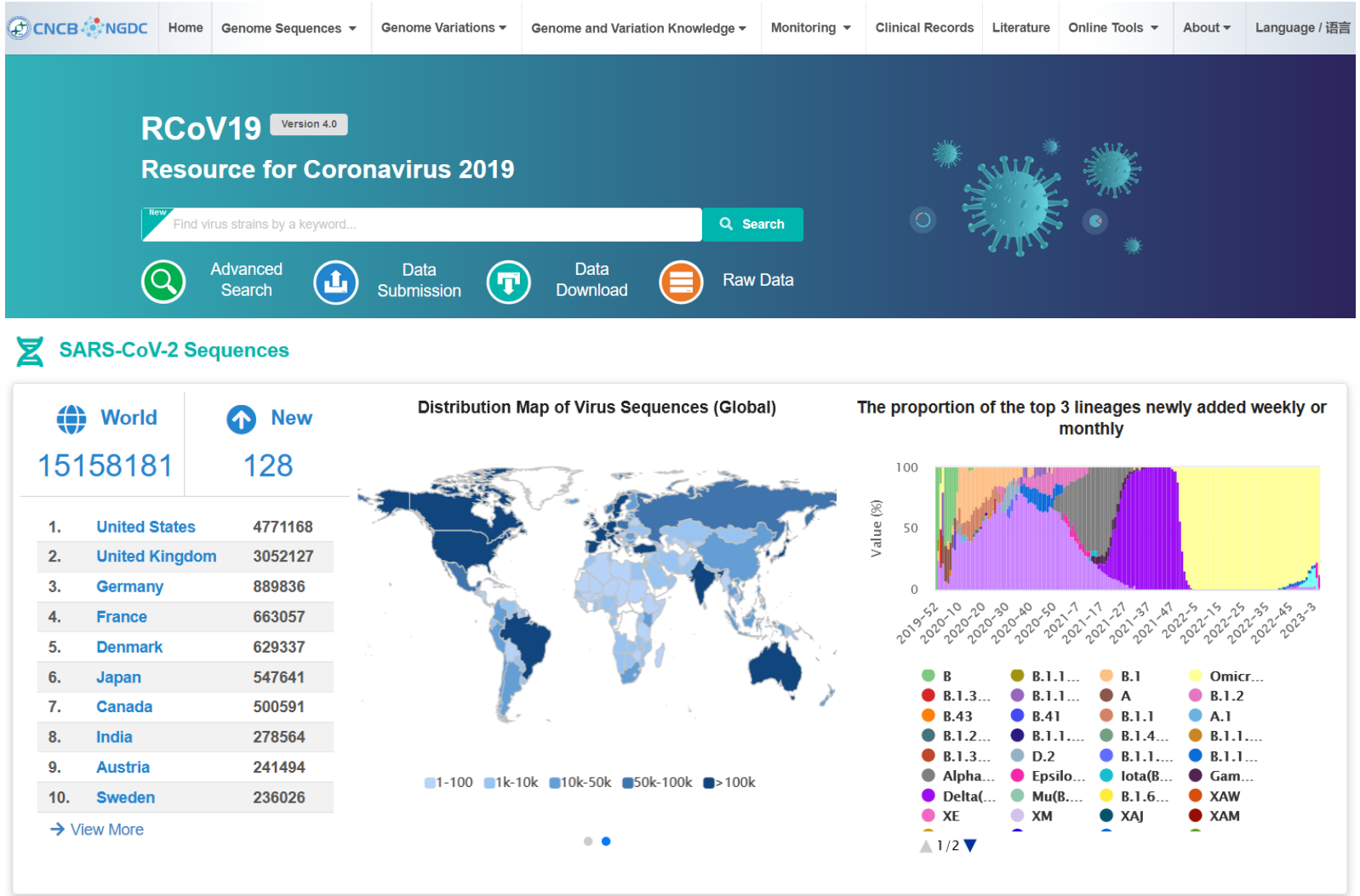
BLAST



Customized databases

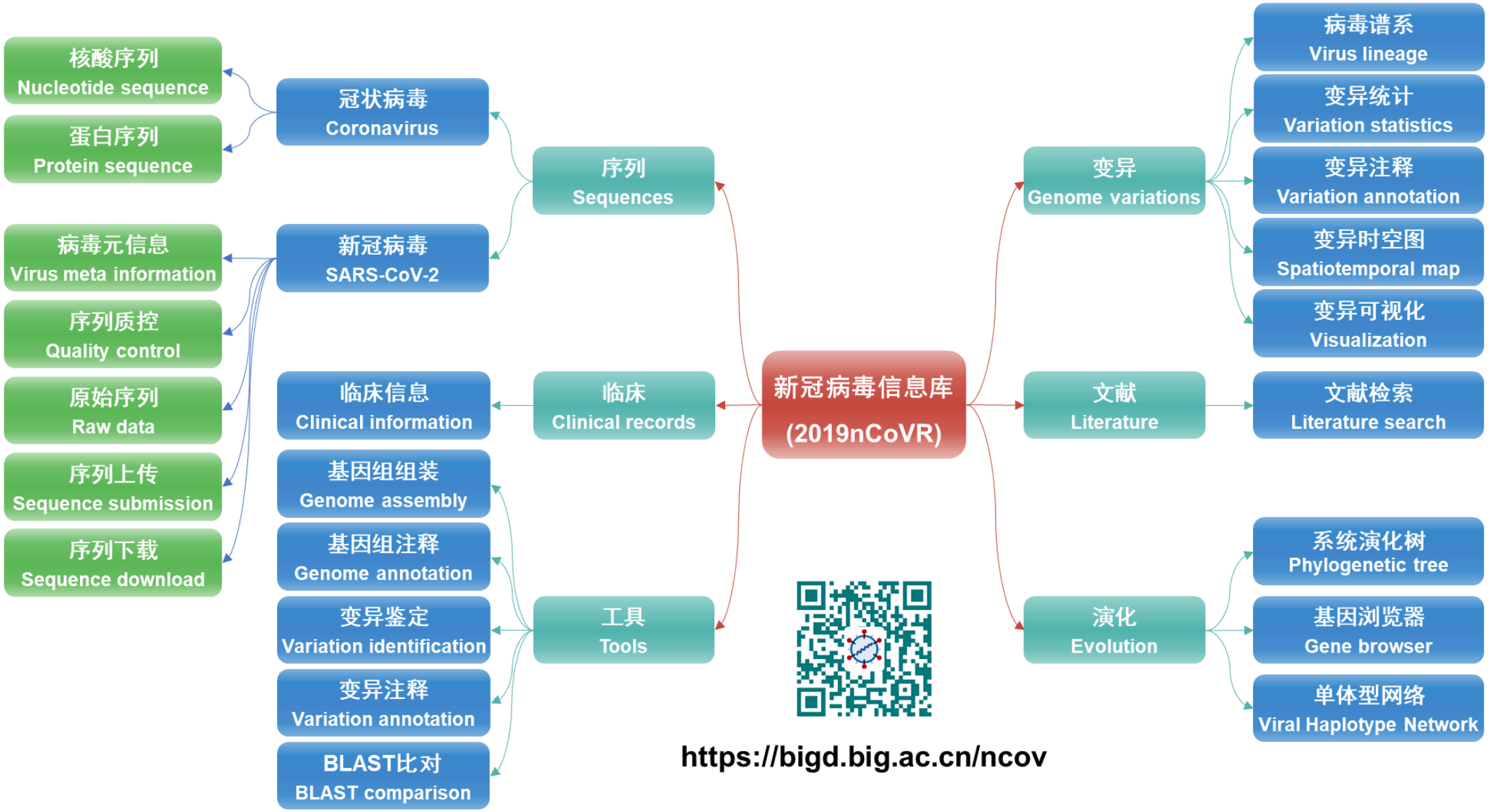


RCoV19

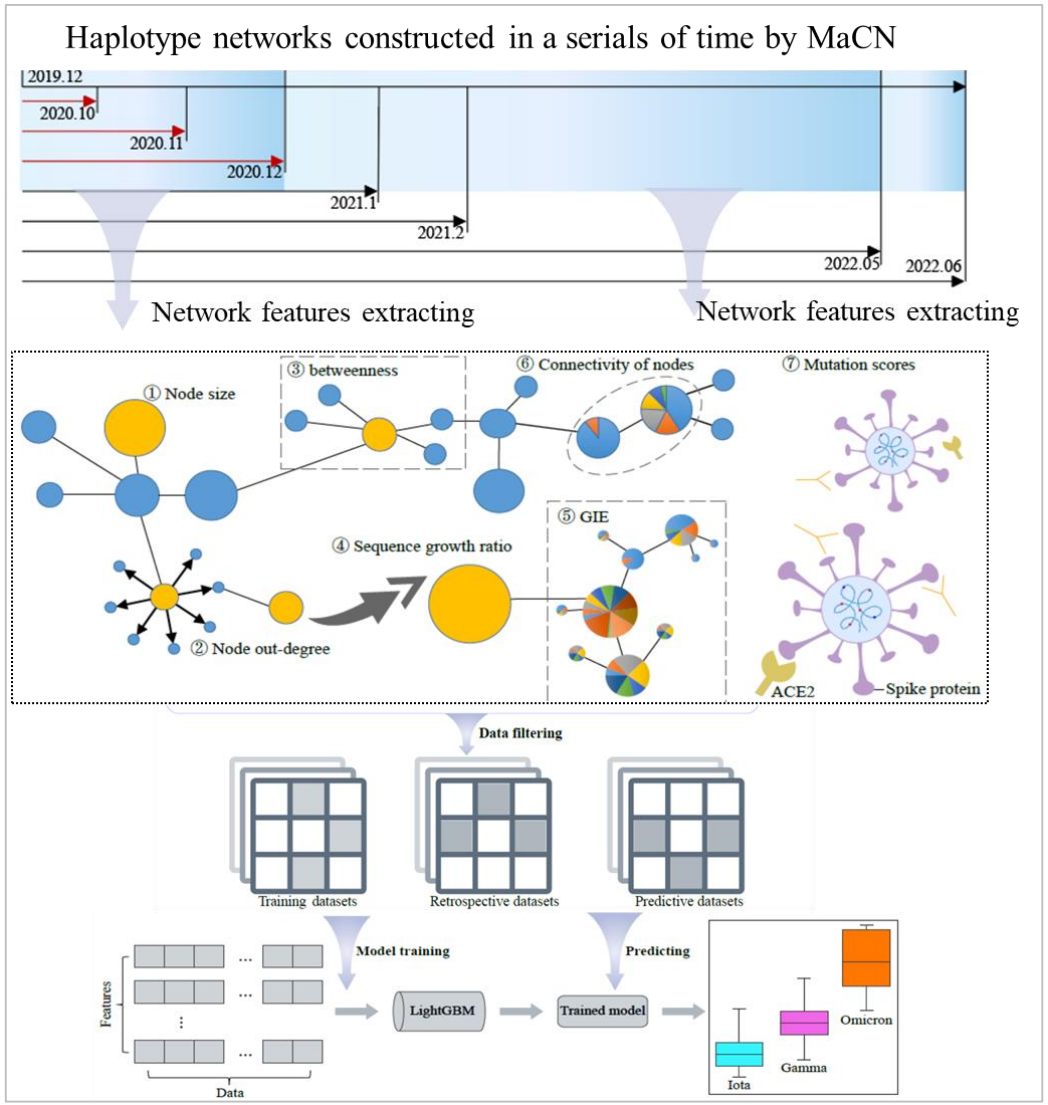


Yi Chuan, 2020; *Zoological Research*, 2020; *Genomics Proteomics Bioinformatics*, 2020; *Nucleic Acids Research*, 2021

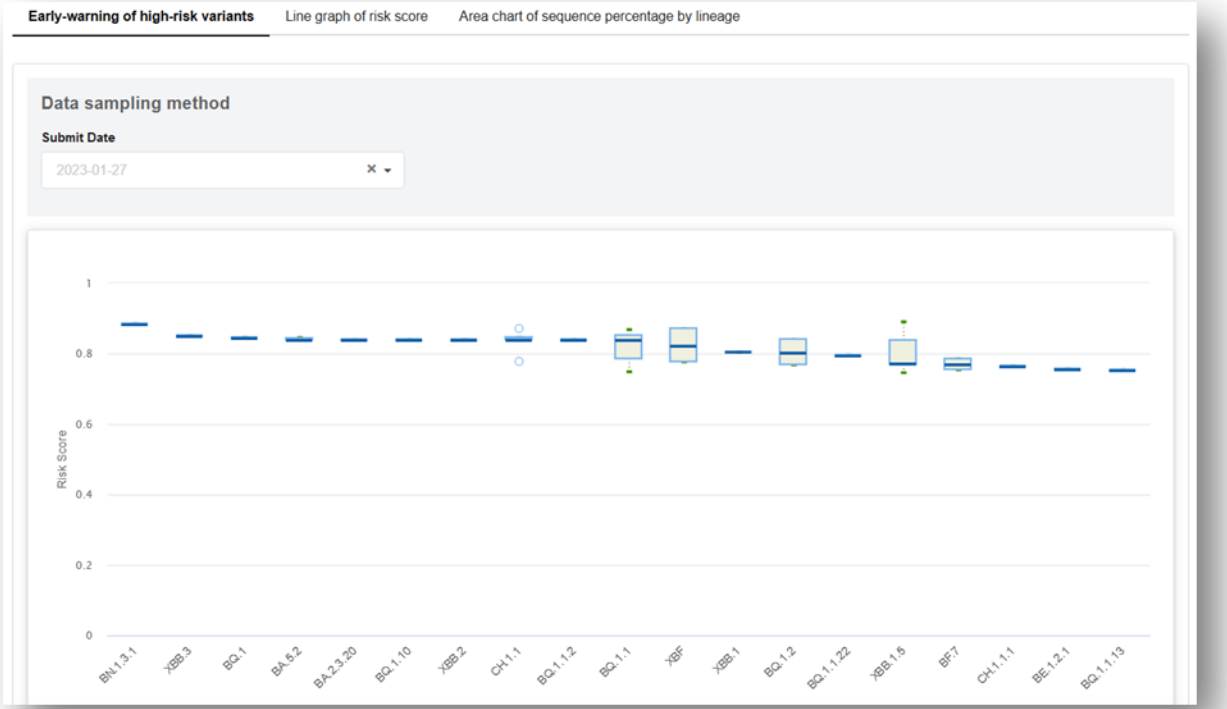
RCoV19


新冠病毒信息库
(2019nCoVR)
序列
Sequences
临床
Clinical records
工具
Tools
变异
Genome variations
文献
Literature
演化
Evolution
冠状病毒
Coronavirus
新冠病毒
SARS-CoV-2
核酸序列
Nucleotide sequence
蛋白序列
Protein sequence
病毒元信息
Virus meta information
序列质控
Quality control
原始序列
Raw data
序列上传
Sequence submission
序列下载
Sequence download
临床信息
Clinical information
基因组组装
Genome assembly
基因组注释
Genome annotation
变异鉴定
Variation identification
变异注释
Variation annotation
BLAST比对
BLAST comparison
病毒谱系
Virus lineage
变异统计
Variation statistics
变异注释
Variation annotation
变异时空图
Spatiotemporal map
变异可视化
Visualization
文献检索
Literature search
系统演化树
Phylogenetic tree
基因浏览器
Gene browser
单体型网络
Viral Haplotype Network
https://bigd.big.ac.cn/ncov

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



Briefings in Bioinformatics, 2023



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

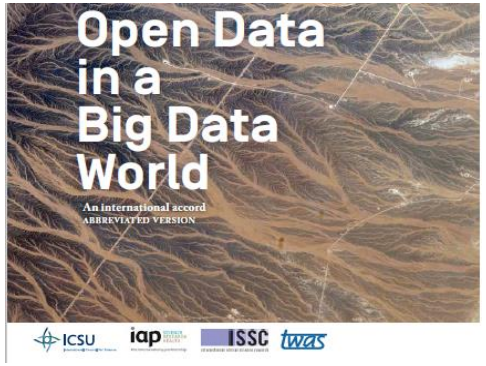
BHBD Alliance

About BHBD

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

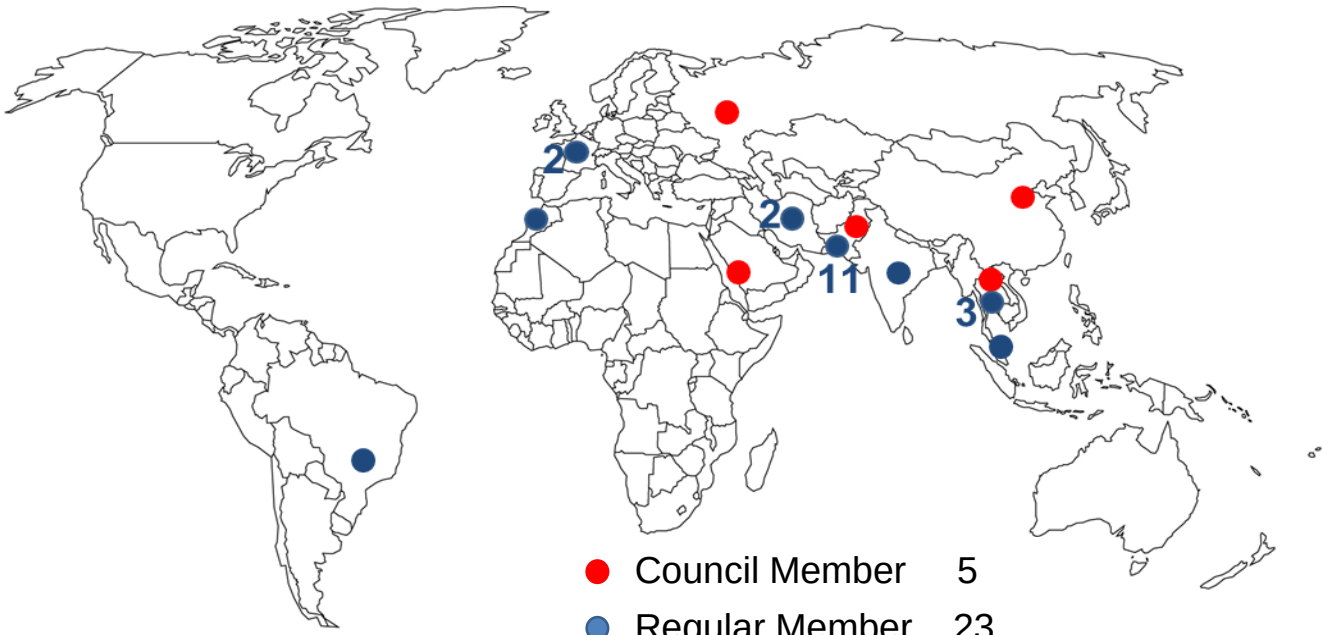
Vision of BHBD

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



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

BHBD Establishment and Membership Expanding



- Council Member 5
- Regular Member 23

BIG, CAS/CNCB China

QAU Pakistan

VIGG Russia

KAUST Saudi Arabia

CU Thailand

28 Members

12 Countries

(As of Dec 2022)

Regular Members:

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

International Meetings/Trainings

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

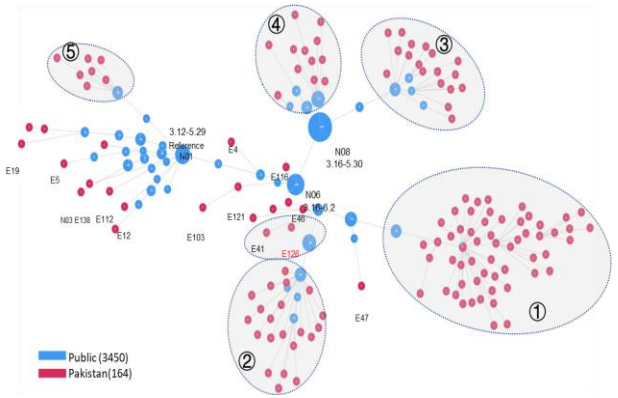
Visiting scholars

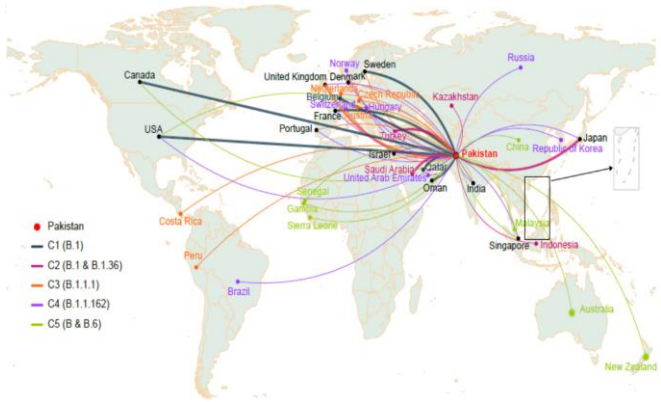
BHBD Int'l Symposium
Jul., 2019, Pakistan

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

International Joint Research

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





Genomics Proteomics Bioinformatics 19 (2021) 727–740

Genomics Proteomics Bioinformatics

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

ELSEVIER

ORIGINAL RESEARCH

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

NSFC · SAMRC · CNPq · science & innovation · Department of Science & Technology, Ministry of Science & Technology, Government of India · RFBR Russian Foundation for Basic Research · FASIE · 中华人民共和国科学技术部 Ministry of Science and Technology of the People's Republic of China · 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

Genomics Proteomics Bioinformatics 20 (2022) 60–69

Genomics Proteomics Bioinformatics

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

ELSEVIER

ORIGINAL RESEARCH

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

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

Grants Awarded for International Collaboration

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

Take home messages

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

Acknowledgements

Steering Advisors:

- Runsheng Chen
- Guoping Zhao

Scientific Advisors:

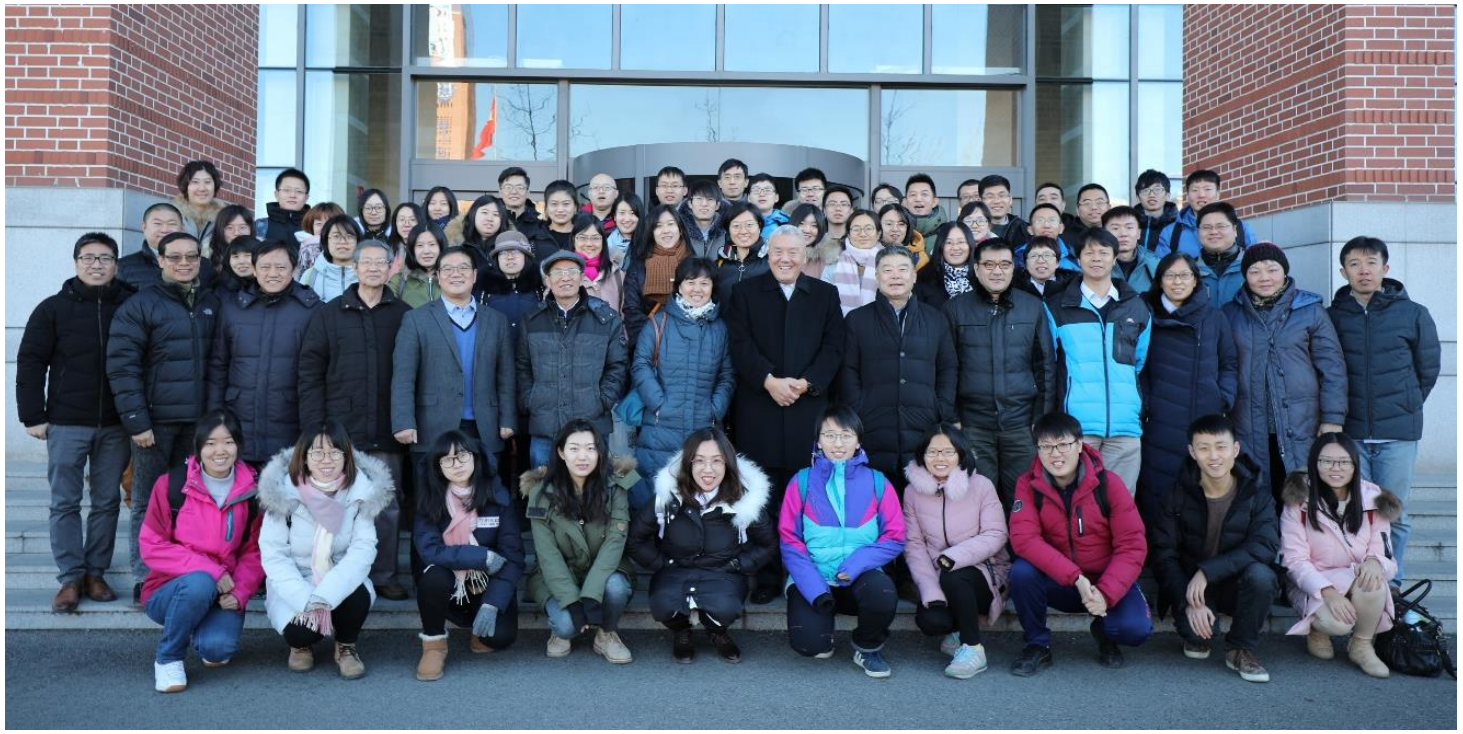
- Amos Bairoch (SIB)
- Guy Cochrane (EBI)
- Frank Eisenhaber (BI)
- Takashi Gojobori (KAUST)
- Yixue Li (CAS)
- Jingchu Luo (PKU)
- Ilene Mizrachi (NCBI)
- Yasukazu Nakamura (DDBJ)
- Weimin Zhu (CAS)

Center Collaborators:

- SINH: Guoqing Zhang
- IBP: Shunmin He

Strategic Partners:

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



中国科学院-马普学会计算生物学伙伴研究所 (2005-2020)

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

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

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

中国科学院 CHINESE ACADEMY OF SCIENCES

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

国家自然科学基金委员会 NSFC

ANSO Alliance of International Science Organizations

IUBS

Thank You!

NGDC

BHBD Alliance

baoym@big.ac.cn