China National Center for Bioinformation
Promotes Open Science

Yiming Bao
Director
National Genomics Data Center
Beijing, China

APANAC 2023
Sep. 28, 2023 • Panama
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

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 jackhammer60 and HHBlits61, 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 catalogue7, and clustered to 30% sequence identity, while enforcing a 90% alignment coverage of the shorter sequences using MMseqs2/Linclus63. 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-

UNESCO Recommendation on Open Science
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 (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

10K patients, ~2.3 PB data

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

Translating big data into big discoveries

Deposition → Integration → Translation
The Team

- **Steering Advisors**
- **Professors**

- 67 students
- 53 Staff

https://ngdc.cnccb.ac.cn/people
The growing of capability

Establishment of National Scientific Data Centers (NSDCs)

Mandatory deposition in NSDCs for data from government-funded projects
Establishment of 20 National Scientific Data Centers

➢ 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)

<table>
<thead>
<tr>
<th>序号</th>
<th>国家称号</th>
<th>国家名称</th>
<th>依托单位</th>
<th>主管部门</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>国家高能物理科学数据中心</td>
<td>中国科学院高能物理研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>国家基因组科学数据中心</td>
<td>中国科学院北京基因组研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>国家微生物科学数据中心</td>
<td>中国科学院微生物研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>国家空间科学数据中心</td>
<td>中国科学院国家空间科学中心</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>国家天文科学数据中心</td>
<td>中国科学院国家天文台</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>国家对地观测科学数据中心</td>
<td>中国科学院遥感与数字地球研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>国家极地科学数据中心</td>
<td>中国极地研究中心</td>
<td>自然资源部</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>国家青藏高原科学数据中心</td>
<td>中国科学院青藏高原研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>国家生态科学数据中心</td>
<td>中国科学院地理科学与资源研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>国家材料腐蚀与防护科学数据中心</td>
<td>北京科技大学</td>
<td>教育部</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>序号</th>
<th>国家称号</th>
<th>国家名称</th>
<th>依托单位</th>
<th>主管部门</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>国家冰川冻土科学数据中心</td>
<td>中国科学院寒区旱区环境与工程研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>国家计量科学数据中心</td>
<td>中国计量科学研究院</td>
<td>市场监管总局</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>国家地球系统科学数据中心</td>
<td>中国科学院地理科学与资源研究所</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>国家人口健康科学数据中心</td>
<td>中国医学科学院</td>
<td>卫生健康委</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>国家基础学科公共科学数据中心</td>
<td>中国科学院计算机网络信息中心</td>
<td>中科院</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>国家农业科学数据中心</td>
<td>中国农业科学院农业信息研究所</td>
<td>农业农村部</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>国家林业和草原科学数据中心</td>
<td>中国林业科学研究院资源信息研究所</td>
<td>林草局</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>国家气象科学数据中心</td>
<td>国家气象信息中心</td>
<td>气象局</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>国家地震科学数据中心</td>
<td>中国地震台网中心</td>
<td>地震局</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>国家海洋科学数据中心</td>
<td>国家海洋信息中心</td>
<td>自然资源部</td>
<td></td>
</tr>
</tbody>
</table>
China National Center for Bioinformation (CNCB) is affiliated with Beijing Institute of Genomics

- Bioinformation data archiving, storage, management and sharing
- Perform frontier research
- Achieve translation and application
Comprehensive Resources at CNCB-NGDC

➢ Omics databases
- BioProject
- BioSample
- Genome Sequence Archive (GSA)
- GenBase
- Genome Warehouse (GWH)
- Gene Expression Nebulas (GEN)
- Genome Variation Map (GVM)
- Methylation Bank (MethBank)

➢ Specialized databases
- RCoV19
- IC4R
- DogSD
- LncRNAWiki
- Database Commons

➢ Literatures
- OpenLB

➢ Tools
- BLAST
- BIT
Collaborations with INSDC

DDBJ

2017, 2020, 2023
INSDC Annual Meetings

NCBI

2017, 2018, 2021
Visit and training

EBI

2016, 2019, 2022
Visit and INSDC meeting
Following INSDC Data Structure & Standard

- **BioProject**
  - PRJNA - PRJNZ: NCBI
  - PRJEA - PRJEZ: EBI
  - PRJDA - PRJDZ: DDBJ
  - PRJCA - PRJCY: NGDC

- **BioSample**
  - SAMN: NCBI
  - SAME: EBI/ENA
  - SAMD: DDBJ
  - SAMC: NGDC

- **Sequence Read Archive**
  - DRA: DDBJ, Submission object
  - DRP: DDBJ, Study object
  - DRR: DDBJ, Run object
  - DRS: DDBJ, Sample object
  - DX: DDBJ, Experiment object
  - DRZ: DDBJ, Analysis object
  - ERA: ENA/EBI, Submission object
  - ERP: ENA/EBI, Study object
  - ERR: ENA/EBI, Run object
  - ERS: ENA/EBI, Sample object
  - EX: ENA/EBI, Experiment object
  - ERZ: ENA/EBI, Analysis object
  - SRA: NCBI, Submission object
  - SRP: NCBI, Study object
  - SRR: NCBI, Run object
  - SRS: NCBI, Sample object
  - SRX: NCBI, Experiment object
  - SRZ: NCBI, Analysis object

- **GSA**
  - CRA: Submission object
  - CRP: Study object
  - CRR: Run object
  - CRS: Sample object
  - CRX: Experiment object
  - CRZ: Analysis object

Image From DDBJ
Rapid Data Growth

>33 PB as of 2023-09-27
Integration of International Data - GSA

Metadata information has been updated regularly
The data files have been downloaded every day since **2022-04-20**

Data Files: ~5 PB
GenBase in sync with GenBank

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

https://ngdc.cncb.ac.cn/genbase/
Supporting >15k Research Grants

## Agencies

<table>
<thead>
<tr>
<th>Agencies</th>
<th>Grants</th>
<th>GSAs</th>
<th>Experiments</th>
<th>Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ministry of Science and Technology of the People's Republic of China (MOST)</td>
<td>967</td>
<td>686</td>
<td>46360</td>
<td>48686</td>
</tr>
<tr>
<td>National Natural Science Foundation of China (NSFC)</td>
<td>6327</td>
<td>2562</td>
<td>122430</td>
<td>130721</td>
</tr>
<tr>
<td>Chinese Academy of Sciences (CAS)</td>
<td>508</td>
<td>563</td>
<td>31116</td>
<td>35526</td>
</tr>
<tr>
<td>Others</td>
<td>7343</td>
<td>703</td>
<td>55082</td>
<td>61504</td>
</tr>
</tbody>
</table>

https://ngdc.cncb.ac.cn/gsa/statistics
GSA Endorsed by Springer Nature and Major Publishers

**Springer Nature**

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**
- DNA sequence data*
- RNA sequence data*
- 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/](https://www.springernature.com/gp/authors/research-data-policy/repositories-bio/)
International Submitters from 22 countries

Crop Science, The Queensland Alliance for Agriculture and Food Innovation, University of Queensland
Upuli Nakandala, u.nakandala@uq.edu.au
Upuli Nakandala, Robert Henry, Agnelo Furtado, Ardy Kharabian Masouleh

Reference citrus genomes

6 (7) Australian species
4 (5) domesticates
High quality haplotype resolved de novo


The Queensland Alliance for Agriculture and Food Innovation (QAAFI) is a research institute of The University of Queensland (UQ), supported by the Queensland Government.
# GSA for Human Database – Controlled Access

The Genome Sequence Archive for Human (GSA4Human), 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
- Providing controlled access data management services
- Providing secure online data request services.

## Available Data

### HRA006158
- **Title**: Single-cell immunological landscape of peripheral blood mononuclear cells of patients with COVID-19 disease
- **Organization**: National Clinical Research Center for Infectious Diseases
- **DAC**: HDAC000080
- **Access**: Controlled
- **Requests**: 163
- **Effective Requests**: 103
- **Approved**: 94
- **Sharing Rank**: ★★★★★
- **Last Processed**: 2023-08-26

### HRA006051
- **Title**: scRNA-seq of gastric cancer
- **Organization**: Institute of Military Cognition and Brain Sciences
- **DAC**: HDAC00002S
- **Access**: Controlled
- **Requests**: 122
- **Effective Requests**: 76
- **Approved**: 51
- **Sharing Rank**: ★★★★★
- **Last Processed**: 2023-08-31

### HRA006155
- **Title**: Global Characterization of CD4+ Immune Cells States in Peripheral Blood and Synovial Tissues of ACPA-negative and ACPA-positive Rheumatoid Arthritis Patients by Single-cell Sequencing
- **Organization**: Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College
- **DAC**: HDAC000096
- **Access**: Controlled
- **Requests**: 101
- **Effective Requests**: 64
- **Approved**: 12
- **Sharing Rank**: ★★★★★
- **Last Processed**: 2023-08-14

### HRA001748
- **Title**: scRNA-seq of liver cancer
- **Organization**: Peking University First Hospital
- **DAC**: HDAC001033
- **Access**: Controlled
- **Requests**: 100
- **Effective Requests**: 62
- **Approved**: 60
- **Sharing Rank**: ★★★★★
- **Last Processed**: 2023-09-04

## Request Access

For more information, please visit the GSA for Human website:

[https://ngdc.cncb.ac.cn/gsa-human/](https://ngdc.cnbc.ac.cn/gsa-human/)
Human Data Backup & Registration Protocol

Backup Center

Registration ID

Backup ID

Researchers

Submission

Accession #s

Backup Center

Registration Center

Backup ID

CNCB-NGDC GSA-Human
Cross-database search engine: BIG Search

BIG Search

NCBI DBs

NGDC DBs

API

JSON

Index files

EBI DBs

Partner DBs

AlphaFold Protein Structure Database

Partner databases

https://ngdc.cncb.ac.cn/search/
“Google” for biology data

BIG Search

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

- **All Databases**: human
  - e.g., PRJCA000126;SAMC000385;tp53;EGFR; human; KaKs_Calculator

<table>
<thead>
<tr>
<th>Database</th>
<th>Records Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlphaFold DB</td>
<td>307623</td>
<td>AlphaFold Protein Structure Database</td>
</tr>
</tbody>
</table>

Powered by EBI AlphaFold DB
Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution.

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

The OpenLB’s literature texts are sourced from NCBI PubMed, bioRxiv and medRxiv, including title, abstract, author, journal, reference, etc.
Bioinformatics Tolls - BIT

Highly used

<table>
<thead>
<tr>
<th>Tool</th>
<th>#Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>10689</td>
</tr>
<tr>
<td>plot_heatmap</td>
<td>857</td>
</tr>
<tr>
<td>plot_venn</td>
<td>756</td>
</tr>
<tr>
<td>plot_bar</td>
<td>530</td>
</tr>
<tr>
<td>plot_box</td>
<td>366</td>
</tr>
</tbody>
</table>

News & updates

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

Contact

If you have any questions or
BLAST

Customized databases

- Gene Expression Nebulas (GEN)转录本序列
- Genome Warehouse (GWH)转录本序列
- LncBook人类长非编码RNA序列
- IC4R水稻转录本序列
- NCBI核酸序列集 (nt)
- 冠状病毒基因组数据库
- SARS-CoV-2基因组数据库
- SARS-CoV-2 PANGO谱系基因组
- 高粱核酸序列
- 原生生物P10K基因组
- 大黄蜂基因组序列
- 大黄蜂转录本序列
- Gene Expression Nebulas (GEN)转录本序列
- 热带作物基因组
RCoV19


https://ngdc.cncb.ac.cn/ncov/
Machine learning detection of high-risk SARS-CoV-2 variants

Briefings in Bioinformatics, 2023

https://ngdc.cncb.ac.cn/ncov/monitoring/risk
Data Sharing with NCBI

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

The 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 new database account for 48 fields (1) while 92 (not previously apparent) in the history of new and existing resources. The remaining 52 papers discuss various aspects of NGS technologies, with the EC platform and the Genomic and Sequencing technologies platform offering different approaches. The database of National centers for biotechnology and biotechnology-related information provides a useful resource for readers interested in learning more about NGS technologies.

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 are these groups: (i) nucleic acid sequence databases, (ii) protein sequence databases, (iii) genome browsers, and (iv) other databases, such as the PubChem database, which includes chemical structures, biological activities, and human disease associations. The NCBI database, which was one of the first public databases, now includes over 150 million sequences and is updated daily. The EBI database, which was one of the first databases to be developed, now includes over 100 million sequences and is updated daily. The BIG Data Center, which was one of the first databases to be developed, now includes over 50 million sequences and is updated daily.
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

National Genomics Data Center (NGDC)  
China National Center for Bioinformatics (CNCB)

BHBD:

Established in: 2022

Membership:

- Council Member: 5
- Regular Member: 23

Regular Members:

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

Members: 28
Countries: 12

(As of Dec 2022)
International Meetings/Trainings

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

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+

Genomic Epidemiology of SARS-CoV-2 in Pakistan

Shuhui Song, Cuiping Li, Lu Kang, Dongmei Tian, Nazih Badar, Wentai Ma, Shilei Zhao, Xuan Jiang, Chun Wang, Yongqiao Sun, Wenjie Li, Meng Lei, Shuangli Li, Qinhui Qi, Aamer Braham, Muhammad Salman, Massab Umair, Huma Shireen, Fatima Batool, Bing Zhang, Hua Chen, Yan-Gui Yang, Qihui Wang, Ana Tereza Ribeiro de Vasconcelos, Georgi A. Basykin, Yiming Bao, Mingkun Li.
Association with ANSO

- BHBD became one of the international associations of ANSO (Alliance of International Science Organizations) since 2020
- BHBD contributed for ANSO’s activities to fight against COVID-19
- CNCB-NGDC data resources were introduced in *ANSO Highlight for Open Data* and recommended by ANSO President Prof. BAI Chunli
## Grants Awarded for International Collaboration

<table>
<thead>
<tr>
<th>Funding Agency</th>
<th>Project Title</th>
<th>Duration</th>
<th>Collaborators</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>IUBS</td>
<td>Open Biodiversity and Health Big Data Initiative</td>
<td>2019-2022</td>
<td>Multiple countries</td>
<td>Euro 30,200</td>
</tr>
<tr>
<td>ANSO</td>
<td>Global Biodiversity and Health Big Data Alliance</td>
<td>2020-2022</td>
<td>Multiple countries</td>
<td>RMB 750,000</td>
</tr>
<tr>
<td>ANSO</td>
<td>Precision warning method for high-risk variants of emerging infectious diseases</td>
<td>2023-2025</td>
<td>Brazil, France, Pakistan</td>
<td>RMB 1,300,000</td>
</tr>
<tr>
<td>ANSO</td>
<td>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</td>
<td>2023-2025</td>
<td>Thailand, USA</td>
<td>US$ 150,000</td>
</tr>
<tr>
<td>NSFC</td>
<td>SARS-CoV-2 Network for Genomic Surveillance in Brazil, Russia, India, China and South Africa (NGS BRICS)</td>
<td>2021-2022</td>
<td>Brazil, Russia, India, South Africa</td>
<td>RMB 2,000,000</td>
</tr>
<tr>
<td>CAS</td>
<td>Global Genomics Data Sharing</td>
<td>2023-2025</td>
<td>USA</td>
<td>RMB 800,000</td>
</tr>
</tbody>
</table>
Summary

- A comprehensive bioinformatics resource
  - Multi-omics DBs (GSA、GenBase、GWH、GVM、MethBank)
  - Knowledgebases (TCOD、RCoV19)
  - Tools and literatures (BLAST、OpenLB)

- The establishment of 3 national centers/platform
  - CNCB
  - NGDC
  - HGRIP

- International recognitions
  - Publishers (Springer Nature、Elsevier)
  - Peers (NAR Database Issue)
  - Major global centers (INSDC)
Take home messages

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

Steering Advisors:
• Runsheng Chen
• Guoping Zhao

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

Center Collaborators:
• SINH: Guoqing Zhang
• IBP: Shunmin He

Strategic Partners:
• Ming Chen
• Qinghua Cui
• Feng Gao
• Ge Gao
• Xin Gao
• An-Yuan Guo
• Tao Jiang
• Cheng Li
• Chuan-Yun Li
• Xia Li
• Jian Ren
• Yun Xiao
• Yu Xue
• Yong Zhang
• Fangqing Zhao
Thank You!
We are open for collaborations

NGDC
BHBD Alliance
baoym@big.ac.cn