

BIG to CNCB: An Exploratory Journey from Genomics to Bioinformatics

YANG Yungui^{a,b,*}, XUE Yongbiao^{c,†}, WU Zhongyi^{d,‡}, YANG Huanming^{e,§}

^a China National Center for Bioinformatics, Beijing 100101, China

^b Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

^c Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

^d Sun Yat-sen University, Guangzhou 510275, China

^e Beijing Genomics Institute, Shenzhen 518080, China

Abstract: The Beijing Institute of Genomics (BIG) of the Chinese Academy of Sciences, as the leading Institute in Genomics, has walked through 20 year's journey since being founded in November 2003. From participating in the Human Genome Project (HGP) in completing the “1% task” to independently accomplishing the super-hybrid rice genome and other several national and international genome projects, BIG has made tremendous contributions in genomics research and development in China. In 2024, bearing great ambition and responsibility, BIG is transformed to the China National Center for Bioinformatics (CNCB), aiming to become a global hub in bioinformatics big data services, innovation, and entrepreneurship. With the completion of its new infrastructure in 2027, CNCB is looking into a brighter future.

Keywords: Human Genome Research Center (HGRC); Beijing Genomics Institute (BGI); Beijing Institute of Genomics (BIG); China National Center for Bioinformatics (CNCB); Genomics, Proteomics & Bioinformatics (GPB); Human Genome Project (HGP)

Cite this article as: YANG Yun-Gui, XUE Yongbiao, WU Chung-I, & YANG Huanming. (2024) BIG to CNCB: An Exploratory Journey from Genomics to Bioinformatics. *Bulletin of the Chinese Academy of Sciences*, 38(3), 164–169. DOI: <https://doi.org/10.1051/bcas/2024007>

History and Development

Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), officially founded on November 28, 2003, was born from the Human Genome Research Center (HGRC) established on August 11, 1998 at the Institute of Genetics (now the Institute of Genetics and Developmental Biology), CAS. In retrospect, therefore, the achievements of BIG can be divided into three major stages: (1) from 1998 to 2003, representing China officially as one of the six nations (other five: US, UK, France, Germany, and Japan) participating in the Human Genome Project (HGP); (2) from 2003 to 2019, leading role in advancing genomics research in China; and (3) from 2019 onwards, accelerat-

* Co-corresponding author as Director of BIG CAS and Director of CNCB since August 2023, to whom correspondence may be addressed at ygyang@big.ac.cn.

† Co-corresponding author as Director of BIG CAS between May 2014 to July 2023, to whom correspondence may be addressed at ybxue@genetics.ac.cn.

‡ Co-corresponding author as Director of BIG CAS between June 2008 to April 2014, to whom correspondence may be addressed at wzhongyi@mail.sysu.edu.cn.

§ Co-corresponding author as Director of BIG CAS between May 2003 to January 2008, to whom correspondence may be addressed at yanghm@genomics.org.cn.

ing phase in omics research and bioinformation infrastructure in China.

At the first stage, two Chinese institutions, including the Beijing Genomics Institute (BGI, a privately-funded non-profit research institution) and HGRC (IGD of CAS: the Predecessor of BIG), represented Chinese scientists to join the International Human Genome Project Sequencing Consortium on July 7, 1999 (International Human Genome Sequencing Consortium, 2001; 2004). The Chinese team took an assigned task of 1% human genome sequences (about ~ 30 Mb region on the short arm of human chromosome 3), also known as “the 1% Project” which was believed to be linked to esophageal cancer susceptibility. At that time, BIG and BGI were

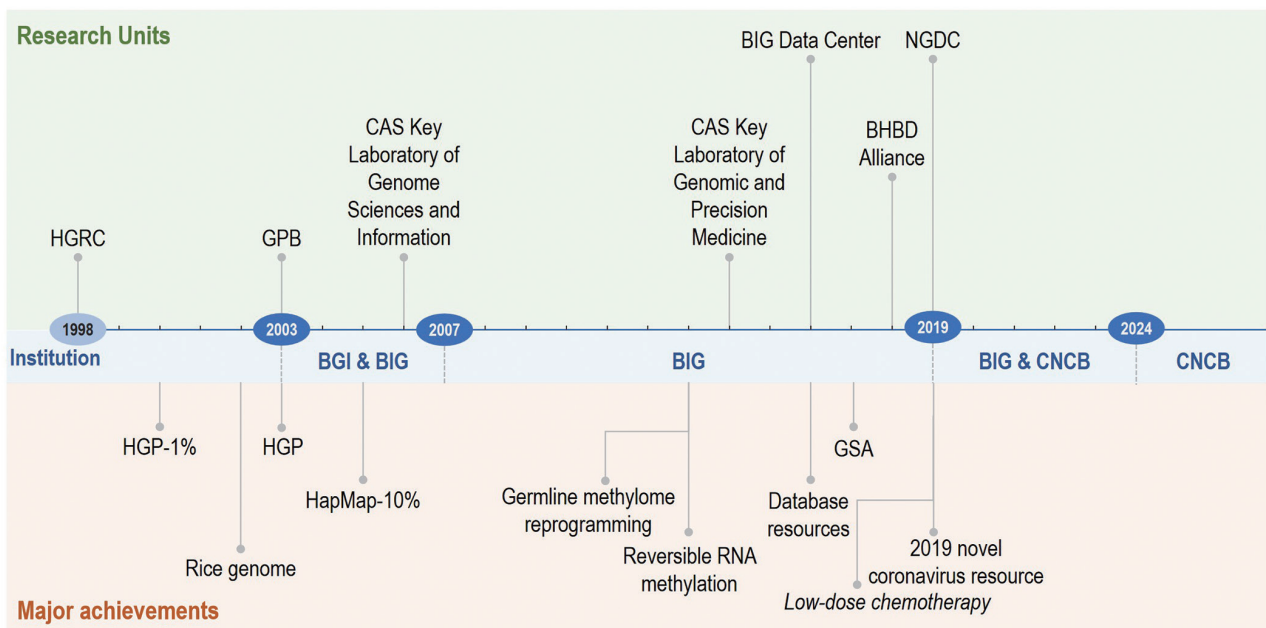
two institutions under the same roof, but organized differently. In addition, a journal named *Genomics, Proteomics & Bioinformatics* was launched at the time of BIG establishment in 2003 (Figure 1).

At the second stage, BIG has continued to actively engage in several national and international genome programs, including the projects of super-hybrid rice genome (YU *et al.*, 2002), the silkworm genome (XIA *et al.*, 2004), the International HapMap (The International HapMap Consortium, 2003), and the Saudi-Sino date palm genome (Al-Mssallem, *et al.*, 2013). In addition, BIG has made critical contributions on decoding the genomes of severe acute respiratory syndrome (SARS)-associated coronavirus (SARS-CoV) in 2003 (QIN *et al.*, 2003) and offering timely, valu-

able genomics support for forensic identification of tsunami victims in Thailand in 2004 (DENG *et al.*, 2005). In 2007, BIG moved to the CAS campus, while BGI moved to Shenzhen seeking further local governmental supports. Thereafter, two major research units have been successively established in BIG, including the CAS Key Laboratory of Genome Sciences & Information in 2006, the CAS Key Laboratory of Genomic and Precision Medicine in 2014. The two CAS key laboratories have made scientific breakthroughs in epigenetic heritability (JIANG *et al.*, 2013), aging associations (Ma *et al.*, 2020) and clinical therapeutic applications (HU *et al.*, 2019), and reversible RNA methylation (WANG *et al.*, 2021).

The BIG Data Center (BIGD) was established in 2016. This

Figure 1. Timeline of institution, its research units, and major achievements since 1998. According to the institution name, the history of our institution can be divided into three stages (light blue), namely, BGI & BIG (1999–2003), BIG (2003–2019), BIG & CNCB (from 2019 onwards), along with major achievements (light red) and involved parties (light green) labeled. Abbreviations used are: BGI for the Beijing Genomics Institute; BHBD for the Biodiversity and Health Big Data; BIG for the Beijing Institute of Genomics; CAS for the Chinese Academy of Sciences; CNCB for the China National Center for Bioinformation; GPB for the journal *Genomics, Proteomics & Bioinformatics*; GSA for the Genome Sequence Archive; HapMap for the International HapMap Project; HGP for the Human Genome Project; HGRC for the Human Genome Research Center; and NGDC for the National Genomics Data Center.



Graphic: CNCB

center developed the Genome Sequence Archive (GSA) (WANG *et al.*, 2017) as the first repository of raw sequence data in China, providing data archiving, retrieval, and sharing services for the worldwide community. Attributing to the excellent achievements made by BIGD and the growing importance and demand for big data in life, medicine, and health sciences, the National Genomics Data Center (NGDC) based on BIGD was founded on June 5, 2019 by the Ministry of Science and Technology and the Ministry of Finance of China.

At the third stage, on November 13, 2019, the China National Center for Bioinformation (CNCB) was officially entitled to be affiliated with the BIG. Since then, BIG has taken the great responsibilities to focus on fundamental bioinformatics data resources, including multi-omics big data deposition, data security management, and public sharing of multi-omics data, in support of interdisciplinary cutting-edge research and facilitating breakthroughs based on big data. One typical circumstance is the construction of the Resource for Coronavirus 2019 (RCoV19) during the onset of the COVID-19 pandemic, involving considerable efforts to integrate global SARS-CoV-2 genome sequences, decode their variants and haplotypes, conduct variant monitoring and high-risk variant warning, and offer online data analysis tools (ZHAO *et al.*, 2020). Currently, RCoV19 has served over 2.5 million visitors from 181 countries/regions, with more than 10 billion sequences downloaded, the value of which is embodied in the key supports not only for SARS-CoV-2 research, but also for the joint studies on tracing the origins of SARS-CoV-2 by the World Health

Organization (WHO) and China in 2021 (WHO-convened-global-study-of-origins-of-sars-cov-2-china-part, 2021).

To fulfil the national strategic needs, BIG has gradually switched to CNCB by fully focusing on big data deposition, integration, and translation, as demonstrated by a suite of database resources providing open access to the whole scientific communities as well as over 52 petabyte data deposit in GSA as of September 2024. One of the important developments of CNCB is the construction of bioinformatics infrastructure at a new Wanquan campus in the City of Zhangjiakou, Hebei Province, approximately 190 kilometers northwest of Beijing, which was formally approved by the National Development and Reform Commission, China in 2022. CNCB will have two campuses, Beijing and Wanquan, to coordinate talent, international exchanges, data resource systems, algorithms, and computing power. By taking advantage of this new campus, CNCB bears great potential to accelerate the development of life science, drive revolutionary innovations, and make substantial impacts on international scientific research in the coming years.

The major missions of CNCB are going to meet the key strategic needs of population health and social sustainable development in China and worldwide. By focusing on the omics data from national precision medicine and important strategic biological resources, CNCB aims to establish a big-data system with massive storage, integration, mining analysis and research, to develop new technologies and methods for big-data mining analysis and system construction, and to build a platform for big data exchange, application and sharing. Along

with the completion of these tasks, CNCB will become an internationally renowned big data center in life sciences.

Since its establishment, CNCB has continuously expanded, updated, and enriched. Its core data resources have systematically developed from the initial 6 to 109 databases containing comprehensive multi-omics data (<https://ngdc.cncb.ac.cn/databases>). Multiple data resources, information libraries and knowledgebases have been developed and presented great international influence. Meanwhile, bioinformatics analysis tools and platforms (Database Commons) have also been established, covering basic omics data resources (GSA), national human genetic resources (GSA human), important strategic biological resources (such as genome-scale database—GWH and genetic variation database—GVM), COVID-19 resources (RCoV19), *etc.* It covers various fields of life sciences, not only supporting multi-dimensional omics data exchange, storage, management, and sharing, but also providing important data resources and reference information for population health-related research, breeding improvement, biodiversity, *etc.* In terms of big data visualization, we have developed a visualization browser for monitoring and tracking the evolution of massive genomes, as well as a haplotype network-based evolution tracking platform, all of which significantly improve our big data service capabilities. CNCB has been continuously evaluated by the journal *Nucleic Acids Research* as one of the world's major biological data centers, alongside NCBI and EBI, greatly enhancing China's influence in the field of biological data (CNCB-NGDC Members & Partners, 2024). As of Septem-

ber 2024, we have served nearly 10,000 users and stored data for approximately 22,000 sci-tech projects. We have archived over 53 PB of data and supported over 3,800 research publications. CNCB has received a total of 598 million visits from over 190 countries/regions.

International Journey

HGP: The launch and completion of HGP in the turn of last century mark a new era of biomedical research, which facilitates the start-point of utilizing human genetic information for the clinical diagnostics and therapies. Through around 20 year's development, medicine has now truly entered its modern paradigm, and its fruitful achievements have contributed to the national healthcare programs and pharmaceutical industries worldwide. The timely participation in such a magnificent scientific project has profound influences on the advancement of research in fields of biomedical and agricultural sciences in China, including genomics, bioinformatics, stem cell research and genetic breeding. The younger generation of scientists have been well-trained to lead the research fields above, which meets the need of national research and development undertakings as well as that of higher education on genomics and other omics research frontiers in China. In addition, it has not only proven that Chinese scientists are always ready and capable of participating in and cooperating major international scientific and technological projects, but also encouraged Chinese scientists to build up confidence and skills in becoming global leaders in scientific discoveries and technological developments.

GSA: The Genome Sequence

Archive (GSA) has been broadly supported and endorsed by the scientific community, and has been designated as supported data repository by major global publishing groups, such as Springer Nature and Elsevier. GSA also serves as a data submission and management platform for a series of China's key funding sources, such as the key national research and development programs, the National Natural Science Foundation of China, and the Chinese Academy of Sciences' Strategic Priority Research Program.

In recent years, the data volume in GSA has grown exponentially. Until September 2024, it has collected a total of 1,712,736 experiments, 2,023,591 runs and over 52 PB files submitted by 8,900 users from over 1500 institutions, supporting more than 22,000 scientific programs. GSA has been visited by more than 980,000 users from 190 countries/regions, with an average of 6 TB of downloads per day. It has supported over 3,800 research articles in 640 scientific journals. Moreover, in Google Scholar, the articles introducing GSA have been cited 1,261 times. More importantly, GSA was selected as one of the Global Core Biodata Resources by the Global Biodata Coalition in 2023, demonstrating its pivotal roles in the global landscape of biological data resources.

BHBD: In response to the appeal "Open Data in a Big Data World" released by the International Council for Science (ICSU) in 2016, the Open Biodiversity and Health Big Data Initiative, initiated by BIG, CAS, was first approved by the International Union of Biological Sciences (IUBS) in 2017, and then turned into an IUBS scientific programme for the triennium from 2019 to 2022 and now the one

from 2023 to 2026. Under this framework, the Global Biodiversity and Health Big Data Alliance (BHBD) was launched in Beijing on October 14, 2018 with five founding members from China, Pakistan, Russia, Saudi Arabia and Thailand, with the aim to build an international platform for the open sharing of biodiversity and health big data. BHBD has also been an association of the Alliance of International Science Organizations (ANSO) since 2020.

Since the establishment of BHBD, great efforts have been made to expand international network and promote scientific cooperation. As of August 2024, 35 institutions from 17 countries have joined BHBD, covering Asia, Europe, South America and Africa. The alliance has conducted over 20 bilateral/multilateral activities, hosted 10 international meetings, implemented training programs for over 200 researchers, and published over 20 papers with international cooperation. A family of database resources and tools has been developed to support global academic and industrial communities (<https://ngdc.cncb.ac.cn>). Together with the BHBD members, several international research programs have been successfully completed or implemented. In particular, soon after the COVID-19 outbreak, BIG/CNCB developed RCoV19 on January 22, 2020 and shared it to international users, providing strong support for scientific researches on COVID-19 worldwide.

INSDC: We have also been cooperating with the International Nucleotide Sequence Database Collaboration (INSDC), including the National Center for Biotechnology Information (NCBI) in US, European Bioinformatics Institute (EBI) in UK, and DNA Data Bank of Japan (DDBJ) in Japan,



Graphic: CNCB

Beijing Campus, China National Center for Bioinformation.

and are establishing partnership with INSDC in the near future.

GPB: The journal *Genomics, Proteomics & Bioinformatics* (abbreviated as *GPB*) was launched at the year of Institution establishment in aim to track the research frontiers in the fields of life omics and bioinformatics across the globe. As the only official journal of the Institution, *GPB* reports leading-edge technologies and their applications in omics research to promote data sharing. Representative work published includes GSA, the NGDC's core database; Col-XJTU, the Arabidopsis genome; and T2T-YAO, the human reference genome.

In addition to the primary focus on publishing, *GPB*, with the support of the institution and the editorial board, organizes and releases the annual "Top Ten Advances in Bioinformatics in China", and also organizes the *GPB*

Omics and Bioinformatics Frontiers Symposium to help promote the development and cultivate youth talents of the field.

Taken together, these activities have promoted knowledge exchanges and sharing of the data on biodiversity and health among global researchers, and expanded the networks for international collaborations.

Perspectives

The future vision of CNCB is to become a global hub in the bioinformation big data services, innovation, and entrepreneurship. First, in the field of data integration and sharing, CNCB aims to be the global hub for multimodal data fusion. As standardization is key to achieving global data interoperability, CNCB will actively participate in the formulation of international standards, which will promote the establishment

of unified data coding, description, and exchange standards, as well as a common framework for data communication. In terms of global collaboration, to address the challenges of trust, verification, and traceability of biomedical data across geographical domains, CNCB will establish privacy-preserving data sharing and access control mechanisms, with aim of building a data marketplace being positioned as a major global hub for comprehensive bioinformation data.

On the path of technological innovation and development, CNCB will leverage cutting-edge technologies such as large artificial intelligence (AI) models, cloud computing, edge computing, and future non-Von Neumann computing architectures to create an open bioinformation algorithm and computing infrastructure to serve the whole world. Based on this infrastruc-

ture, CNCB will foster the growth of an open scientific community focusing on bioinformation innovation, establish itself as a key global center for biomedical conferences, and ultimately become a leading hub for talent cultivation and original innovation in bioinformation.

CNCB will integrate resources from research, education, industry, and capital to provide comprehensive support for startups. By consolidating advanced technology and extensive data resources, CNCB will establish dedicated funds, and collaborate with universities, research insti-

tutions, and enterprises to offer one-stop services, from office space and talent development to market analysis and capital matchmaking. CNCB will also promote the establishment of ethical standards and legal regulations in the field of bioinformation. Ultimately, CNCB will accelerate the transformation of scientific and technological achievements into market applications, drive the development of the bioeconomy, and become the “Global Silicon Valley of Bioinformation” to foster bioinformation innovation and entrepreneurship worldwide.

Acknowledgements

The authors thank all remarkable staff, coworkers, colleagues and friends who participated in and supported the establishment and the development of the Beijing Genomics Institute, CAS and the China National Center for Bioinformation for their hard work and great efforts. The authors thank Drs. YU Jun, ZHANG Zhang, ZHANG Zhihua, SONG Shuhui, JIAO Yuxia, MI Shuangli, ZHAO Wenming, and WANG Caiping, Messrs. ZHANG Xin and ZHANG Xiaoliang, and Ms PAN Liying for their help in writing the article.

References

- Ibrahim S. Al-Mssallem, Songnian Hu, Xiaowei Zhang, *et al.* (2013). Genome sequence of the date palm *Phoenix dactylifera* L. *Nature Communications*, 4: 2274, DOI: 10.1038/ncomms3274.
- BIOLOGY ANALYSIS GROUP, QINGYOU XIA, ZEYANG ZHOU, CHENG LU..... GENOME ANALYSIS GROUP, JUN YU, JUN WANG, RUIQIANG LI, *et al.* GENOME ANALYSIS GROUP, YU Jun, WANG Jun, LI Ruiqiang, *et al.* (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, 306: 1937–1940.
- CNCB-NGDC Members & Partners. (2024). Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Research*, 52: D18–D32.
- Ya-Jun Deng, Yuan-Zhe Li, Xiao-Guang Yu, *et al.* (2005). Preliminary DNA Identification for the Tsunami Victims in Thailand. *Genomics, Proteomics & Bioinformatics*, 3(3): 143–157.
- Yixin Hu, Aili Chen, Xinchang Zheng, *et al.* (2019). Ecological principle meets cancer treatment: treating children with acute myeloid leukemia with low-dose chemotherapy. *National Science Review*, 6(3): 469–479.
- International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409: 860–921.
- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431: 931–945.
- Lan Jiang, Jing Zhang, Jing-Jing Wang, *et al.* (2013). Sperm, but Not Oocyte, DNA Methylome Is Inherited by Zebrafish Early Embryos. *Cell*, 153: 773–784.
- Shuai Ma, Shuhui Sun, Lingling Geng, *et al.* (2020). Caloric Restriction Reprograms the Single-Cell Transcriptional Landscape of Rattus Norvegicus Aging. *Cell*, 180(5): 984–1001 e1022.
- QIN E'de, ZHU Qingyu, YU Man, *et al.* (2003). A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chin Sci Bull*, 48(10): 941–948.
- The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 426: 789–796.
- WANG Xiujie, LIU Feng, SUN Yingpu, YANG Yun-Gui, HE Chuan, WANG Hailin, ZHOU Qi, ZHOU Fangjian, CHONG Kang, YUAN Zengqiang, HUANG Min, MA Jinbiao, XIE Dan (authors are listed in the order of the number of strokes in their Chinese surnames); *Epitranscriptomics of RNA Methylation (China Basic Research Frontiers Series)*, Zhejiang University Press, February 2021.
- Yanqing Wang, Fuhai Song, Junwei Zhu, *et al.* (2017). GSA: Genome Sequence Archive. *Genomics, Proteomics & Bioinformatics*, 15(1): 14–18.
- JUN YU, SONGNIAN HU, JUN WANG, *et al.* (2002). A draft sequence of the rice (*Oryza sativa* L. ssp. *indica*) genome. *Science*, 296: 79–92.
- Wen-Ming Zhao, Shu-Hui Song, Mei-Li Chen, *et al.* (2020). The 2019 novel coronavirus resource. *Yi Chuan*, 42: 212–21. (In Chinese)