

# 国家基因库：共有、共为、共享

王博<sup>1</sup>, 刘芳<sup>2</sup>, 张二春<sup>1</sup>, 沃晨亮<sup>1</sup>, 陈振家<sup>1</sup>, 钱璞毅<sup>1</sup>, 卢浩荣<sup>1</sup>,  
曾文君<sup>1</sup>, 陈泰<sup>1</sup>, 危金普<sup>1</sup>, 万仟<sup>1</sup>, 王韧<sup>1</sup>, 徐讯<sup>1,2</sup>

1. 深圳国家基因库, 深圳 518120
2. 深圳华大生命科学研究院, 深圳 518083

**摘要:** 基因资源是国家的重要战略资源, 保存、保护和合理利用基因资源将成为未来维护国家安全、打造核心竞争力的坚实基础和有效保障, 然而我国在基因数据存储、样本存储等方面均起步较晚, 无法满足国内日益增长的生命科学相关领域的研究发展需求。针对上述问题, 2011 年中国政府批复依托深圳华大生命科学研究院(原深圳华大基因研究院)建设我国首个读、写、存一体化的综合性生物遗传资源基因库——深圳国家基因库(亦称国家基因库)。本文总结了国内外较有影响力的基因资源大平台的发展概况, 着重阐述了国家基因库的定位与使命, 以及“三库两平台”的结构与功能——生物遗传资源的存储、读取、合成运用和开放共享。自 2016 年 9 月正式运行以来, 国家基因库作为公益性、开放性、支撑性、引领性的战略性科技平台, 已具备千万级可溯源样本存储能力, 十万级基因组/年的存储和计算能力, 建成首个国产化 Pb (Petabases)级基因组数据产出平台以及千万碱基/年高效合成平台, 同时基于自身平台能力, 国家基因库建立了全面的开放共享机制, 开展资源数据共享和公共平台服务, 对生命科学研究和生物产业创新发展的支撑和助力初见成效。

**关键词:** 基因库; 样本库; 数据库; 基因测序; 合成生物学平台

## The China National GeneBank—owned by all, completed by all and shared by all

Bo Wang<sup>1</sup>, Fang Liu<sup>2</sup>, Erchun Zhang<sup>1</sup>, Chenliang Wo<sup>1</sup>, Jason Chen<sup>1</sup>, Puyi Qian<sup>1</sup>,  
Haorong Lu<sup>1</sup>, Wenjun Zeng<sup>1</sup>, Tai Chen<sup>1</sup>, Jinpu Wei<sup>1</sup>, Qian Wan<sup>1</sup>, Ren Wang<sup>1</sup>, Xun Xu<sup>1,2</sup>

1. China National GeneBank, Shenzhen 518120, China
2. BGI-Shenzhen, Shenzhen 518083, China

**Abstract:** Genetic resources are important national strategic resources. Their preservation, protection and rational utilization form a solid foundation to guarantee national security and to build national competitiveness for the future. Due to a relatively late starting point, China is actively catching up with global peers in storing genetic samples and data. In view of this, in 2011 China approved a plan to build its first nation-level comprehensive gene bank, the China

收稿日期: 2019-05-23; 修回日期: 2019-07-29

作者简介: 王博, 硕士, 研究方向: 基因组学。E-mail: wangbo@cngb.org

通讯作者: 徐讯, 博士, 研究员, 研究方向: 基因组学、生物信息学等。E-mail: xuxun@genomics.cn

王韧, 博士, 研究员, 研究方向: 农学。E-mail: wangren@cngb.org

DOI: 10.16288/j.ycz.19-148

网络出版时间: 2019/8/5 20:59:59

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20190805.2059.007.html>

National GeneBank (CNGB), and entrusted BGI-Research to implement its construction and operation. It is China's first gene bank for "reading, writing and storing" bioresources. In this paper, we summarize the development of influential platforms at home and abroad, and focus on CNGB's position, mission, and its structure of "Three Banks and Two Platforms". CNGB launched its official operation in September 2016 and aims to develop a world-class, non-profit and strategic platform that supports science and technology development. It has built capacities to store tens of millions of traceable samples and to analyze hundreds of thousands of WGS each year. It has also set up China's first Pb-level digitalization platform and a high-efficient synthesis platform with a production rate of ten million bases per year. Based on such capacities, CNGB has established its open sharing mechanism for biological samples and data, provided public platform services for life science research, and achieved initial results in supporting innovation and development of the bio-industry.

**Keywords:** gene bank; biorepository; database; genome sequencing; synthetic biology platform

基因是地球生物繁衍生息的写证, 基因的存续代表着一个物种的延续。随着人类基因组计划的启动和实施, 基因与人的关系正逐渐被解析, 基因组学研究及相关技术的跨越式发展, 全球重要物种的基因图谱破译标志着基因组学研究进入了黄金时代。基因资源是国家的重要战略资源, 保存、保护和合理利用基因资源将成为未来维护国家安全、打造核心竞争力的坚实基础和有效保障。国际上部分发达国家在国家级基因资源的部署早已遥遥领先, 美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)、欧洲生物信息研究所(European Bioinformatics Institute, EMBL-EBI)、日本 DNA 数据库(DNA Data Bank of Japan, DDBJ)等三大库早已确立了在全球科学发展中的地位。

中国在 2011 年, 由国家发展改革委员会、财政部、工业和信息化部、卫生健康委员会(原卫生部)联合批复依托深圳华大生命科学研究院(原深圳华大基因研究院)筹建中国首个读、写、存一体化的综合性生物遗传资源基因库——深圳国家基因库(后简称国家基因库)。历时 5 年, 国家基因库于 2016 年 9 月 22 日建成并开始运营。这是继世界三大库之后, 全球第四个建成的国家级基因库。与三大库相比, 国家基因库不仅仅以保存数据为主要功能, 更是一个以“三库(生物样本资源库、生物信息数据库和动植物资源活体库)两平台(数字化平台和合成与编辑平台)”为结构部署的综合性基因库, 读、写、存全生命周期, 样本、活体、数据全贯穿。国家基因库

作为我国生命科学研究和生物产业发展提供基础性和公益性服务平台, 对生物遗传资源进行存储、读取、合成运用和开放共享, 为我国生物技术和生命经济发展提供有力支撑。本文总结了国内外较有影响力的基因资源大平台的发展概况, 着重阐述了国家基因库“三库两平台”的功能建设和开放共享机制, 并展望了国家基因库对中国生命科学领域发展的助力与支撑作用。国家基因库的发展与壮大, 将为中国在全球生命科学领域的引领地位提供有力的基石。

## 1 国内外较有影响力的大平台概况

### 1.1 生物样本库

国际上不乏知名的样本库, 首先最具有国际影响力的组织是国际生物和环境样本库协会(International Society for Biological and Environmental Repositories, ISBER)。该组织 1999 年成立, 是美国研究病理学会下辖的一个分支机构。该机构通过建立规范和标准, 利用培训等方式影响发展中国家的样本库建设, 使其达到一定的质量和标准。目前 ISBER 下辖有 6 个不同类型的生物样本库, 分别为动物样本库、环境样本库、人体样本库、微生物样本库、博物馆样本库、植物/种子样本库。此外, ISBER 设置了若干个专门性的工作组, 通过白皮书或其他出版物, 及时解决生物样本库建设和管理过程中遇到的问题, 逐步推进 ISBER 在生物样本库建设过程

中各个领域内的专业性和权威性。

在美洲与 ISBER 同样具有影响力的还有美国国家癌症中心生物样本库和生物样本研究办公室(Office of Biorepository and Biospecimen Research, OBBR)。OBBR 在 2005 年由美国国立癌症研究所(National Cancer Institute, NCI)成立, 致力于制定一个共同的生物样本库标准, 用于指导、协调和发展机构搜集生物样本资源的能力和提高了所搜集生物样本的质量, 确保其满足研究需要。

英国在生物样本库的建设部署也同样始于 1999 年, 而 1999 年提议设立的英国生物样本库(UK Biobank)历经 7 年于 2006 年开始试运营。UK Biobank 是目前世界上已建成的规模最大的人类遗传队列生物样本库, 该样本库搜集了 50 万来自英国各地 40~69 岁志愿者捐赠的样本, 涵盖血样、尿样、遗传数据和生活方式等个人医疗详细信息, 分析来自他们生活习惯、环境和遗传因素对健康的影响<sup>[1]</sup>。

欧盟国家起步于 2008 年计划筹建泛欧洲生物体样本库与生物分子资源研究平台(Better Biology Makes Reality Interesting, BBMRI)。BBMRI 旨在整合和升级欧洲现有生物样品收集、技术和专家资源平台, 目前来自欧洲 24 个国家超过 200 个机构加入了这一平台, 有超过 1000 万例样本的资源<sup>[2]</sup>。

值得一提的是丹麦国家生物样本库(Danish National Biobank), 丹麦国家生物样本库建于 2012 年, 目前储存有 798 万余份生物样本。丹麦国家生物样本库作为未来丹麦国家样本的中心不仅从物理上整合新收集样本, 并从网络数据管理层面统筹全国各个卫星库的样本信息。在完备的人口登记制度和医疗体制下, 每名丹麦人的生物样本都可以被保留下来, 且具有完善的相关信息。

国际上这些代表性的样本库呈现大型化、专业化、信息化、产业化趋势, 而科学的管理计划、综合的经营模式是其成功的因素<sup>[3,4]</sup>。与欧美相比, 我国的生物样本资源库起步时间并不算晚, 并且我国人口众多, 疾病和生物样本资源极其丰富, 是任何国家无可比拟的, 但是在规模化、信息化、管理操作规范化方面仍然落后于世界发达国家, 主要表现在缺乏统一的建设标准、封闭式发展导致资源分散、整体生物样本资源质量堪忧、存在不少“垃圾库”、

缺乏产业化为应用导向等方面。如何建设高质量、高标准的生物样本库, 建立共享机制促进样本资源共享, 打破行业“孤岛”, 保证其可持续性发展, 是当前亟待解决的问题<sup>[5]</sup>。

## 1.2 数据库

国际上数据库主要为起步较早的三大库, 均是在 20 世纪 80~90 年代已开始搭建, 分别为美国国家生物技术信息中心(NCBI)、欧洲生物信息研究所(EBI)和日本 DNA 数据库(DDBJ), 结合生物信息的免费共享政策, 已经成为国际生物信息数据存储、交换、获取方面的核心机构。它们对常用的生物数据类型已经积累了 PB 级的数据量, 其中涵盖了多种数据类型, 如核酸与蛋白质序列数据、蛋白质二、三级结构数据、化学小分子数据与代谢通路数据等。

国内在基因相关的数据库搭建上近年来也有不同维度的代表, 例如生命与健康大数据中心(BIG Data Center, BIGD, <http://bigd.big.ac.cn/>), 建立多组学数据的生命与大数据资源系统, 建立生物大数据汇交存储、整合挖掘、共享管理与转化应用体系, 研发生物大数据汇交管理平台和多组学数据资源体系, 主要包括组学原始数据归档库(Genome Sequence Archive, GSA)、基因组数据库(Genome Warehouse, GWH)、基因组变异数据库(Genome Variation Map, GVM)、基因表达数据库(Gene Expression Nebulas, GEN)、甲基化数据库(Methylation Bank, MethBank)、生物信息工具库(Biological Tool Codes, BioCode)和生命科学维基知识库(Science Wikis)等<sup>[6]</sup>。再有国家组学数据百科全书(The National Omics Data Encyclopedia, NODE)隶属于中国科学院上海生命科学研究院, 主要归档高通量测序数据, 包括近 10 年来市面可见到的测序公司生产的测序仪产出的数据, 如 454、Ion Torrent、Illumina、SOLiD、Helicos 和 Complete Genomics 等。另蛋白质组学整合资源库(Integrated Proteome Resources, iProX, <https://www.iprox.org/>), 是一家在中国建立的蛋白质组学数据与知识中心, 旨在促进蛋白质组学资源在世界范围内的共享。iProX 目前由一个蛋白质组数据提交系统和一个蛋白质组数据库组成。这些数据库多数具有各自独特的专长, 但均以数据单独形式存在, 缺少与样本的

关联,未能发挥出更大的价值。

### 1.3 活体库

活体库是一个全新的提法,国内外暂无统一的概念和完全一致的机构,我们认为凡是能用来衍生出下一代的生物可繁殖体均称之为“活体”,对生物的延续有着重要的意义。列举几个国内外著名的在物种保护及生物多样性保护相关的机构。

挪威的斯瓦尔巴德全球种子库被称为世界末日种子库(The Svalbard “Doomsday” Seed Vault, <https://www.nordgen.org/sgsv/>)也被称为“植物诺亚方舟”。位于距离北极点约 1000 公里的挪威斯瓦尔巴群岛的一处山洞中,旨在为国际植物种质资源提供备份,保护作物的多样性。该种子库堪称全球最安全的基因储存库,其安全性堪比美国国家黄金储藏库,甚至可以抵御地震和核武器。

史密森尼热带研究所(The Smithsonian Tropical Research Institution, STRI, <https://stri.si.edu/>),其主要致力于热带地区物种多样性的研究和保护。它主导巴拿马以及南美地区多个保护区的规划与建设,如通过森林动态监测和物种保育繁殖相结合,为巴拿马热带雨林的重建提供了新的解决方案,再造巴拿马本土物种森林,保护巴拿马本土物种和生物多样性。

圣地亚哥动物园的“冷冻动物园”,是目前世界上最大的冷冻动物园,其冷库中保存着来自于 800 多个物种的 8400 多只动物身上的超过 10 000 种活细胞、精卵细胞和胚胎组织等。圣地亚哥冷冻动物园期望能够在全球范围内建立一个冷冻动物园网络,利用干细胞和试管受精等技术,从而为地球上最珍稀物种的未来提供终极保障。

中国科学院昆明野生动物细胞库利用我国西南地区动物种类繁多、资源丰富的特点,从动物遗传种质资源保护和利用角度,建立具有我国动物资源特色的野生动物细胞库。

阿拉伯濒危野生动物繁育中心开展了多个当地最濒危物种繁育的项目,目前该繁育中心总共繁育了接近 200 种阿拉伯濒危野生动物,涉及范围非常广泛,极大程度上保存了当地珍稀动物的遗传资源。

我国是生物多样性最丰富的国家之一,同时也

是生物多样性受威胁最严重的国家之一。我国拥有复杂多样的生态系统类型,高等植物 35 000 多种,居世界第 3 位,脊椎动物 6400 多种,占世界总种数的 13.7%。尽管我国生物多样性保护取得了积极进展,但生物多样性下降的总体趋势尚未得到有效遏制<sup>[7]</sup>。生物多样性保护的最佳方式之一即是活体库的建设,预期未来 10 年,中国将会有更多、更新型的活体库。

### 1.4 测序平台

随着基因产业的发展,基因数据的产出已逐步被商业化为各类不同的服务,而作为研究型、开放公益的测序平台在国内外却寥寥无几。

英国维康桑格研究所(Wellcome Sanger Institute, <https://www.sanger.ac.uk/>)是世界最著名的基因组测序研究中心之一,建于 1992 年,已具备 12.7 Pb/年的基因组数据产出能力,主要开展癌症与基因变异、细胞遗传学、人类基因学和寄生虫与微生物等方面的研究。凭借其强大的数据产出能力,桑格研究所使用 UK Biobank 的样品开展科研,不断地将数据存入 EBI,持续丰富和更新 EBI 原有的数据库。

近十几年来,各国出现了大量的测序技术服务公司,部分知名基因组研究中心也纷纷建立了自己的测序中心,具备强大的基因组数据产出能力。除维康桑格研究所外,美国的博罗德研究所(Broad Institute, <https://www.broadinstitute.org/>)也已经具备 6 Pb/年的基因组数据产出能力。而法国也投资 7.6 亿欧元开展“法国基因组医疗 2025 计划”并在全法国范围内建立 12 个基因测序平台中心。

我国基因组学在过去十几年中也有较大发展,部分科研机构 and 高校具备一定的数据产出能力。同时,我国涌现出大量以营利为目的的测序技术服务公司,这些公司大多依赖国外测序技术设备、试剂和耗材等。

### 1.5 合成生物学平台

DNA 合成是合成生物学研究关键核心支撑技术,对多个生物技术领域都具有重要价值。随着合成生物学的快速发展,全球 DNA 合成的需求以及 DNA 合成产业的规模呈现出逐年递增的趋势<sup>[8]</sup>。建

设高通量自动化 DNA 合成中心, 抢占发展先机, 已成为世界各国新的生物经济竞争点。

美国伊利诺伊高级生物制造工厂 (Illinois Biological Foundry for Advanced Biomanufacturing, iBioFAB, <https://experts.illinois.edu/>), 是世界上第一个用于合成生物学应用的自动化中心。该平台在功能建设上主要强调以算法驱动、以全自动方式完成项目研究。具体功能模块由两部分组成: (1) 是自动化循环运行“设计-构建-测试-分析”的计算机设计架构; (2) 是自动化集成系统, 可实现自动化 DNA 组装、克隆、表达、表达产物检测等实际功能。

爱丁堡基因组平台 (The Edinburgh Genome Foundry, EGF, <https://www.genomefoundry.org/>), 于 2015 年在英国爱丁堡大学建设, 致力于使用高度自动化设备组装大型 DNA 片段的研究机构。功能上包括: (1) 可用于复杂 DNA 构建的软件 Genetic Constructor, 用于设计大型 DNA 模块或组合文库; (2) 多功能自动化设备, 适用于多种模式 DNA 组装, 如 Golden Gate 组装、Gibson 组装, 以及酵母基因重组、质粒 DNA 制备等实验; (3) 具备细胞表型分析能力, 包括定量 PCR、质谱、高通量生物反应器及发酵动力学分析等。同时, 可实现对所有的设计、构建和测试过程的进行全程监控, 为深度机器学习和流程提供数据基础。

帝国理工学院 DNA 合成平台, 2016 年成立, 目标是建立一个先进的 DNA 合成和构建平台, 支持英国的合成生物学工业发展, 并为学术界和工业界的研究人员提供培训和支持。功能上包括: (1) 一系列商业和专有的 BioCAD/CAM 工具, 用于快速和可扩展的生物学实验设计; (2) 常规和超声液体处理工作站, 利用软件编程与调度, 实现多个 DNA 片段的并行构建和组装; (3) 装配用于数据采集的高通量分析/测试设备, 实现对生物产品的表征和定量分析; (4) 建设基于网络的实验室信息管理系统 (LIMS), 用以追踪实验的工作流程及数据参数, 并通过数据分析来完成建造-设计循环, 利用对数据的统计解释和模型创建来优化原来的设计循环。

三大平台爱丁堡基因组平台 (EGF)、伊利诺伊高级生物制造工厂 (iBioFAB)、帝国理工学院 DNA 合成平台的功能定位在合成生物学自动化高通量“设

计-合成-测试-学习”工程化闭环的基础建设与应用。从顶层设计层面, 各平台均存在长期运用模式上的缺陷, 其共性问题因技术源端控制力不足, 导致下游技术对接领域如医疗、环保等产业应用示范的支撑作用受到很大限制。

## 2 国家基因库的战略定位与重要使命

国家基因库 (China National GeneBank, CNGB, <https://www.cnbg.org>) 是服务于国家战略的国家级公益性创新科研及产业基础设施建设项目, 也是我国首个获批筹建的国家级综合性基因库, 将建设成为引领我国生命科学和生物经济发展的战略性科技力量。它坐落于深圳大鹏区金沙湾, 以海量生物资源的“读、写、存”能力为基础, 实现样本、数据、活体全贯穿, 搭建起基因资源挖掘的公益性、开放性、支撑性、引领性服务平台, 促进基因组学在精准医学、精准健康、未来农业、海洋开发、微生物应用等方面的前沿探索与产业转化, 真正实现基因资源的共有、共为、共享。

农业时代, 一个国家拥有的耕地越多优势越大; 工业时代, 拥有的石油、矿产等能源越多优势越大; 而在生命时代, 拥有更多的基因资源同时能对基因资源进行认知和利用, 则意味着更大的优势。国家基因库的建成对于中国具有重大意义。

(一) 有利于保护并充分利用我国特有的遗传资源。建设国家基因库有利于维护我国生物信息安全, 提高我国基因组数据的储存、分析和管理能力, 促进基因及数据资源共享利用, 增强我国在生命科学大数据时代的国际话语权, 对于推动我国生命科学和生物产业发展, 抢占未来生物经济的战略制高点、掌控基因战略资源, 具有极其重要的战略意义。国家基因库将充分整合各方资源, 形成特色突出、平台开放、资源共享的合作交流机制, 提高样本资源使用的合理性、合法性和资源效益最大转化性, 支撑民生、医疗健康及科研探索, 推动生物创新技术成果落地。

(二) 有利于提升我国生命科学的创新能力和信息共享程度。积极贯彻落实党的十九大关于“加快建设创新型国家”的战略部署, 国家基因库将搭建

基因信息资源公共服务平台, 打造可持续的新型基因库系统, 支持科研机构深入挖掘基因信息数据, 在健康管理、医药卫生、环境改造、农业生产等领域探索新前沿, 拓展新学科, 实现新突破, 不断提升我国生命科学创新能力, 并通过大项目、大科学、大数据的整合, 加快从资源存储、基础科研到技术转化应用的周期, 促进基因组学向临床医学、分子育种、生殖健康等领域的转化, 在大健康领域不断催生新技术、新产品和新模式, 推动相关产业快速发展, 对我国生物产业发展和科技创新能力的提升必起到巨大的推动作用。

### 3 致力于全球最大的综合性国家级基因库

#### 3.1 “三库两平台”贯穿生命科学领域的“读、写、存”能力

国家基因库已初步建成“三库两平台”的业务结构和功能。“三库”, 由生物样本资源库、生物信息数据库和动植物资源活体库组成, 建立了样本、数据、生命体“存”的能力; “两平台”为数字化平台、合成与编辑平台, 建立“读”与“写”的能力(图 1)。

##### 3.1.1 生物样本资源库

国家基因库生物样本资源库致力于建设理念超前、大规模高通量、低成本、全自动的生物样本库(图 2), 形成特色精品库、资源库, 保存本国特有的遗传资源, 建立行业新标准, 提升国内样本库的整体水平, 促进科研成果转化和生物产业的发展。

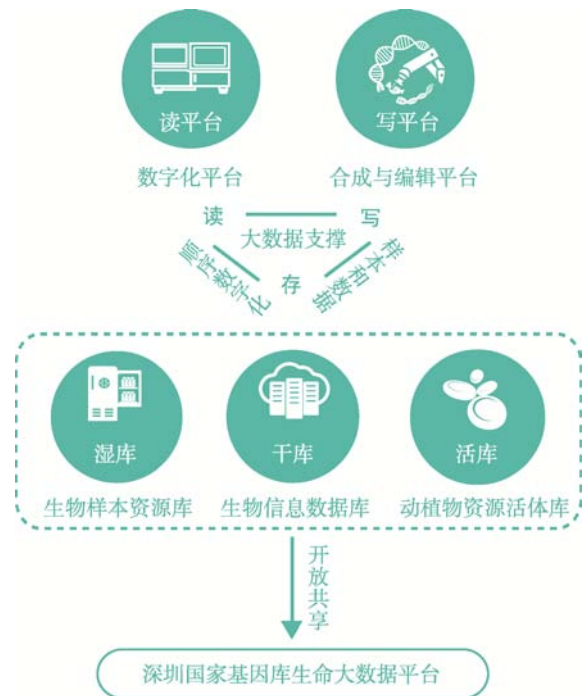


图 1 国家基因库功能关系图

Fig. 1 The functional diagram of CNGB

在硬件建设上, 已具备多温层存储系统、自动化样本处理流水线、实验室及配套设施, 并联合国内制冷企业研发大型高效深低温-80 自动化存储设备, 实现关键技术自主可控, 打破行业壁垒, 驱动深低温存储产业的技术革新, 引领国内深低温存储产业向安全、低温、低能耗新方向发展。

在存储能力上, 已实现 1000 万管可溯源生物样本存储能力, 收集的样本涵盖血液、尿液、唾液、粪便、组织、细胞、蛋白、核酸等多种类型, 形成支撑大型人群特色资源库、植物资源库、动物资源



图 2 生物样本资源库组图

Fig. 2 Group diagram of biorepository

A: -80 自动化冷库; B: 自动化液氮罐; C: 入库操作系统。

库和微生物资源库的全球最大综合性生物资源样本库之一。

在信息化管理上,通过自主研发针对生物样本全周期管理的多个软件,实现支撑样本采集、预处理、运输、库存管理、数据产出、开放共享的全人群覆盖、全流程贯穿、全过程质控的信息化系统,形成了整体统一、模块独立、灵活组合的分布式信息化管理平台,保障数据标准化、行为记录质控及信息安全可控,全面提升资源整合与应用能力。

在支撑服务水平方面,已形成面向科研和产业应用的可溯源、高质量、特色鲜明的人群资源库、植物资源库、动物资源库和微生物资源库等,完善生物样本库行业标准体系,引领生物样本库行业向国际规范化、流程标准化及管理科学化的发展方向。

在标准化建设方面,致力于研究和制定生物样本采集、存储和管理相关的标准和技术规范,促进样本库行业管理体系标准化和规范化的发展。

### 3.1.2 生物信息数据库

相比于国内外数据库,国家基因库基于数据资源的“共有、共为、共享”原则,搭建形成国家基因库生命大数据平台(China National GeneBank Database, CNGBdb, <https://db.cngb.org/>),面向全球科研用户提供样本信息及组学数据归档、计算分析、知识搜索、管理授权和关联可视化等数据服务,实现了样本与数据资源的整合贯穿共享。同时,为了更好地管理海量、多维度、多源头的组学数据,国家基因库自主研发了《全样本管理系统》、《国家基因库数据处理平台》等多个管理软件,实现了数据采集、数据存储、自动化生物信息计算分析流程、数据质量标准化和数据服务灵活化的全贯穿信息化系统,形成了统一、自动化的数据处理和数据管理平台,保障数据标准化、统一化,全面提升了数据整合与应用能力。

以 CNGBdb 为基础,国家基因库对外开放了数据、平台和应用的 3 个能力层,且在平台开放层中,构建了数据信息加密和管理的模块和流程。通过不可逆的加密和对敏感数据的存储隔离,实现数据的受控与权限访问,保障数据的存储和共享安全。基于 3 个能力开放层, CNGBdb 实现了内外部样本

信息与组学数据的整合,并依据《国家科技资源共享服务平台管理办法》规定,逐步推进和实现样本与数据资源面向社会开放共享。2017 年 10 月, CNGBdb 上线试用,2018 年 10 月正式发布。截至 2019 年 4 月已累积注册用户 2979 个,访客数 102 783 人次,接收并处理了样本与数据使用相关业务咨询 970 余次。

#### (1) 数据资源归档与共享

国家基因库核酸序列归档系统(CNGB Nucleotide Sequence Archive, CNSA, <https://db.cngb.org/cnsa/>)作为 CNGBdb 的数据归档模块,为科研用户提供了快捷的在线提交生物研究项目、样本、实验、组装等信息和数据的功能,同时也提供公开数据下载、数据受控管理等服务。截止 2019 年 4 月,该系统已归档 565.75 TB 的测序数据,其中 221.26 TB 数据已开放共享。

#### (2) 样本资源整合与共享

CNGBdb 中的样本资源信息共享服务平台(E-BioBank, EBB, <https://db.cngb.org/ebb/>),致力于建立全球生物样本库目录、制定统一规范的资源信息整合体系,创造公平、可靠的资源样本信息共享环境,促进样本资源科学合理地利用。截止 2019 年 4 月,E-BioBank 平台已整合共享的样本资源信息累计 48.9 万条份,涉及我国农业、林业、微生物、海洋等领域的生物遗传资源等。

#### (3) 专业科学的数据库

国家基因库在原有 40 个数据库的基础上进行优化整合,形成覆盖动植物生物多样性、微生物和人类疾病健康等领域的 10 个专业生物信息数据库和数据分析平台(表 1)。

国家基因库生命大数据平台(CNGBdb)自上线以来,已支撑 29 篇文章发表,如蒙古族人全基因组学研究“*Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia*”<sup>[9]</sup>等。CNGBdb 充分发挥了科研支撑平台的公共服务作用,为这些项目的样本信息和组学数据的梳理、归档、脱敏、分析与受控访问提供了全面支撑,得到国际高水平学术期刊的认可,支持在其期刊上发表的基因组学文章所使用的数据仅保存于 CNGBdb,

表 1 专业生物信息数据库及数据分析平台列表  
Table 1 List of professional biological information databases and data analysis platforms

数据库	数据库简介
千种植物数据库(1000 Plants, OneKP)	OneKP 是一个国际多学科联盟项目,对超过 1000 种植物进行了大规模测序研究。数据库基于千种植物数据构建了在线 BLAST 平台,提供在线 BLAST 服务。
谷子数据库(Millet Database, MilletDB)	MilletDB 创新地将谷子的表型和基因型贯穿起来,通过谷子的表型信息可以查询和检索谷子的基因型信息,通过基因型可以查到对应的表型信息。
万种鸟基因组数据库 (Bird 10 000 Genomes, B10K)	B10K 将完成整个鸟类生物的基因组水平生命树,解读遗传变异与表型变异之间的联系。
千种鱼转录组计划 (Transcriptomes of 1000 fishes, FishT1K)	FishT1K 数据库将建立专门针对鱼类组学研究的数据存储,应用,深化对鱼类的比较生理学、生物地理学认识,促进鱼类资源医用价值的挖掘、经济和生态价值的开发,以及生物多样性保护等问题的解决。
炎黄中国人基因频率数据库 (Chinese Millionome Database, CMDB)	CMDB 是由 BGI 构建的迄今为止最大规模的中国基因组数据库,数据存放于国家基因库。CMDB 通过分析数百万中国人的测序数据,提供定期、有用的变异信息和科学见解。
癌症数据集成与整合分析平台 (Data Integration Solution for Systematic Exploration of Cancer Traits, DISSECT)	DISSECT 已在中国内地首次建立 ICGC Data Portal(目前最大规模的癌症基因组数据库)镜像站点,为国内研究者提供重要的资源渠道。
免疫数据库 (Pan Immune Repertoire Database, PIRD)	PIRD 主要关注人体相关的免疫数据,收集了多种疾病的 BCR 和 TCR 测序数据,与及对应个体的实验信息,表型信息等。
罕见病数据库 (Genetic Disease and Rare Disease, GDRD)	GDRD 是一个综合的遗传病和罕见病研究与应用平台。整理了共计约 7000 余篇文献,1 万多个致病变异,近 300 个遗传病家系的信息。
微生物组数据库(Microbiome Database, MDB)	MDB 是一个关注人体共生微生物研究的数据库,提供该领域相关的样本和微生物数据,以及一个迄今为止最完整的人肠道微生物基因组。
病原数据库(Pathogen Variation Database, PVD)	PVD 整合了各种病原微生物的基因数据及相关的注释信息,提供全面的基因测序数据的病原鉴定功能,通过数据分析和可视化手段,一目了然地展示鉴定结果。

无需备份至境外数据库,符合我国的人类遗传资源条例管控。这代表着国家基因库完全有能力存储、管理和保护我国重要遗传资源,对我国遗传资源信息和基因数据安全保护具有重要意义。

3.1.3 动植物资源活体库

国家基因库动植物资源活体库的建设,立足于基因库“存”和“读”的能力(即生物资源样本库的保存和数字化平台的测序能力),将活体库建设成为数字化的生物多样性基地和生物资源库。

其优势在于生物资源的转化技术能力,能够将生物物种转化为信息准确和质量可控的生物遗传资源,促进生物物种遗传资源的保存和保护及合理利用。活体库是数字化生态型植物园建设者,其联合开展的中国云南瑞丽珍稀植物园数字化项目(图 3),

全面采集植物园植物物种进行基因组学测序,完成了世界首个生态型植物园基因组数字化研究,从基因组学的角度认知瑞丽地区生物多样性和物种的保护工作。

3.1.4 数字化平台

国家基因库数字化平台,又称“读”平台,读出生物的碱基,即为测序平台。其拥有一系列具有自主知识产权的国产化测序仪,是世界领先的基因组数据产出中心。支撑引领数字化农业、数字化地球、数字化人生、百万传感染病研究及防治计划、百万出生大人研究计划等重大科研项目。其综合优势目前在国内外任何机构中仍无可比拟。

(1)规模大。目前已经具备 15 Pb/年的基因组数据产出能力,即每年可完成 15 万人全基因组测序,



图 3 云南瑞丽珍稀植物园

Fig. 3 The Ruili botanical garden



图 4 国产化超高通量测序中心

Fig. 4 Self-developed high-throughput sequencing center

并且将逐步拓展，成为世界最大的基因组数据产出中心。(2)效率高。在样品准备、测序、测序数据质控和交付处理等实验操作过程全部实现了自动化操

作和信息化管理，测序仪实现全年 365 天、24 小时不间断运行。(3)能力全面。具备全面的应用技术能力，包括全基因组测序、全外显子组测序、全基因组甲

基化测序、RNA-Seq、Meta 基因组测序、ChIP-Seq 文库、单细胞测序、ATAC-Seq、人线粒体基因组测序、免疫组库测序、大片段测序、10× genomics 测序和 stLFR 测序等。全面的技术能力可全面支撑精准医疗、农业育种、海洋开发、微生物应用、生物多样性等领域的各类民生、科研和产业项目。(4)成本低和自主可控。数字化平台使用国产化新型超高通量测序仪,试剂、耗材均不依赖进口。该系列测序仪具有单位产出高、成本低和成本自主可控等优点。

### 3.1.5 合成与编辑平台

国家基因库合成与编辑平台,又称“写”平台。目标围绕形成可支撑每年千万至亿级碱基合成产能的自主知识产权合成软硬件体系,持续保持基因组合成与编辑在能力、效率、成本领先地位;支撑工业微生物、动植物及疾病相关基因组合成,推动基因组合成与编辑的科研及产业化应用。国家基因库建设的合成与编辑平台,在设立和发展上突破了三大平台技术源端控制力不足的壁垒,定位于通用性及专业性、规模化的基因组编写基础设施,将支撑合成生物学科学与产业发展,聚集资源培养并引进创新型人才,有利于增强我国在新一轮科技革命中抢占先机。

在功能上与国家基因库“存”、“读”平台相辅相成,打通基因科技“读、写、存”技术联动,通过技术迭代进行核心技术的成本指数级降低,促进基因科技产业应用的广泛拓展,形成的国家级资源库将成为全球范围首个打通基因科技上下游应用的创新性体系。数字化平台快速解读基因信息的能力以及生物信息数据库存储将提供海量基因组学数据与资源,通过基因组泛写技术切入能够在功能快速验证及产业价值挖掘方面有重大潜在价值。

例如在环境治理方面,通过对高耐受性微藻进行全基因组和转录组等数据收集与分析,定向性在耐污高降解力方向进行基因组改写以进一步提高极端环境的耐受性与降解能力,产出的具有自主知识产权生物资源库将为环保产业应用提升提供新的思路。

再如在疾病检测标准建立方面,在多维度基因

组解读数据的积累基础之上,利用基因组泛写技术可定向构建含高频疾病突变位点的细胞系,为疾病的临床检测提供相应的检测标准品,也可为相应位点的地贫疾病研究构建细胞模型。

## 3.2 开放共享

国家基因库作为服务于国家战略的国家级公益性、开放性、支撑性基础科研平台,旨在建立全面的开放共享机制支撑公共科研需求。一方面积极贯彻落实国家关于生物遗传资源与科学数据管理的政策法规,推动行业内资源数据的整合,加强我国生物遗传数据与生命科学数据的规范管理和利用。另一方面,在满足国家基因库公益类服务、科研合作类服务开展的前提下,利用国家基因库“三库两平台”的剩余能力向公众提供技术与服务以支撑科研,支撑我国生物产业提升创新能力,推动我国生命经济快速发展。

### 3.2.1 资源数据开放共享

国家基因库积极响应国家相关政策法规,开放共享自有的样本、数据资源,为积极推动行业内资源数据的整合提供平台,努力打破资源数据“孤岛”现状,合理与国内外国家级资源数据共享平台实现互通,促进全国乃至全球资源保护与战略性开发。

目前,已通过国家基因库生命大数据平台,面向全球科研用户提供样本信息及组学数据归档、计算分析、知识搜索、管理授权和关联可视化等数据服务,实现了样本与数据资源的整合贯穿共享,进一步提升了国家基因库的开放共享程度。截止 2019 年 4 月,CNGBdb 已标准化、规范化管理和整合生物样本资源信息达百余万份及组学数据 565.75 TB。其中 48.9 万条样本信息和 221.26 TB 组学数据已向全国乃至全球的科研用户开放共享。参与共享体系的医院、高校及科研机构达 30 余家。此外,国家基因库还积极与国内外资源数据开放共享平台交流、合作、互通。国际上,国家基因库已与国际癌症基因组联盟(International Cancer Genome Consortium, ICGC)、全球基因组生物多样性联盟(Global Genome Biodiversity Network, GGBN)实现资源互通;在中国,

积极参与国家人类遗传资源共享服务平台华南中心的建设, 实现与中国战略生物资源网络服务平台等国家级资源平台的互通。

### 3.2.2 仪器设施开放共享

国家基因库作为生物遗传资源的综合性公共平台, 响应并落实国家对大型基础设施开放共享的号召, 充分发挥国家基因库国家级科研支撑平台的公共服务作用, 积极探索并初步形成大型仪器设施开放共享及可持续性发展新模式。以提供公共平台服务的方式来实现大型仪器设施开放共享, 促进生命科学研究及生物产业的发展。

## 4 结语与展望

国家基因库将持续秉承“共有、共为、共享”的理念与宗旨, 充分发挥自身优势, 为粤港澳大湾区建设、“一带一路”、健康中国2030等国家战略部署提供技术储备和实施平台, 并持续保持基因领域的国际竞争力与影响力, 支撑和助力中国生命科学领域发展。

在生物样本库建设上, 进一步提升自动化及信息化存储能力、加强对外服务水平, 形成特色鲜明、开放共享的资源平台, 充分发挥平台的支撑和对外服务功能, 在统一标准规范下, 推动信息、技术、资源的共享与交流, 促进基因组学在精准健康、精准农业、海洋开发、微生物应用等方面的前沿探索与产业转化, 催生新技术、新产品和新模式。在生物信息数据库建设上, 将积极发挥建设与运营的经验和能力, 通过国内外的联盟组织, 参与标准制定、安全认证、技术输出以及开放共享等相关工作, 以此推动技术改革、促进应用普及、引领行业发展, 达到增强我国在国际生命科学研究和应用领域的话语权、实现资源数据共有、共为、共享的效果。在动植物资源活体库建设上, 努力成为中国乃至世界著名的生物资源中心, 促进资源数字化和多组学研究, 支撑基础科研开展与成果突破, 带动产业的发展、实践、创新性应用, 保障我国生物科技和产业发展所需生物多样性资源的安全。在数字化平台建

设上, 终将建成世界规模最大、成本最低的基因组数据产出中心, 提高我国在生命科学领域的核心竞争力、提高我国生物遗传资源的安全性做出贡献。在合成与编辑平台建设上, 将不断推动软件控制、硬件设备和合成生物学应用整合, 弥补国际上现有合成设施的不足。在增强自身对基因组编写设计能力的同时, 促进存储的生命大数据向与国计民生息息相关的精准医疗、环境修复等方向应用转化, 最终建立基因组“读、写、存”一体化的创新路径, 扩大国家基因库公共服务平台的支撑广度, 推动全球样本和信息数据资源产业链发展。

基因组学的发展能有力推动精准医学、传染病防治、农业育种、海洋开发、微生物应用、生物多样性等领域的发展。随着各国都竞相推出精准医学计划和开展大人群基因组项目, 数据产出能力将成为各国重要的竞争力, 国家基因库的发展与壮大, 将为引领中国在全球生命科学领域的地位提供了有力的基石, 为我国生物技术和生命经济发展提供有力支撑。

## 参考文献(References):

- [1] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 2018, 562(7726): 203–209. [DOI]
- [2] Kuhn K, Bild R, Spengler H. BBMRI Catalogue. *J Clin Bioinforma*, 2015, 5(S1): S19–S19. [DOI]
- [3] Liu KX, Zheng L, Wang Y, Zhao XM, Ji X, Jiang JJ, Guo YC. The current situation and research progress of biobank. *Chin Med Record*, 2014, 15(9): 32–34.  
刘克新, 郑琳, 王莹, 赵秀梅, 吉栩, 江婧婧, 郭渝成. 生物样本库的现状 & 研究进展. *中国病案*, 2014, 15(09): 32–34. [DOI]
- [4] Zhang XJ, Li HY, Gong SS. Status analysis and countermeasures of china biobanks. *Chin Hosp Manag*, 2013, 33(07): 76–77.  
张雪娇, 李海燕, 龚树生. 国内生物样本库建设现状分析与对策探讨. *中国医院管理*, 2013, 33(07): 76–77. [DOI]
- [5] Man QH, Yu N, Yan F, Xu Q, Wang WY. Current status

- and future development strategy of China biobanks construction. *Chin Med Biotechnol*, 2018, 13(4): 289–293.
- 满秋红, 于农, 闫飞, 许庆, 王伟业. 我国生物样本库建设现状与未来发展的思考. *中国医药生物技术*, 2018, 13(04): 289–293. [DOI]
- [6] Zhang YS, Xia L, Sang J, Li M, Liu L, Li MW, Niu GY, Cao JB, Teng XF, Zhou Q, Zhang Z. The big data center's database resources. *Hereditas (Beijing)*, 2018, 40(11): 1039–1043
- 张源笙, 夏琳, 桑健, 李漫, 刘琳, 李萌伟, 牛广艺, 曹佳宝, 滕徐菲, 周晴, 章张. 生命与健康大数据中心资源. *遗传*, 2018, 40(11): 1039–1043. [DOI]
- [7] Bai CS, Cui P. The current situation and direction of biodiversity conservation in China. *Environ Prot*, 2015, 43(5): 16–20.
- 柏成寿, 崔鹏. 我国生物多样性保护现状与发展方向. *环境保护*, 2015, 43(5): 16–20. [DOI]
- [8] World Synthetic Biology Market Opportunities and Forecasts, 2014–2020. [DOI]
- [9] Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, Zhang Y, Bond SR, Pei Z, Zhang Y, Zhang D, Jirimutu J, Zhang D, Yang X, Morigenbatu M, Zhang L, Ding B, Guan B, Cao J, Lu H, Liu Y, Li W, Dang N, Jiang M, Wang S, Xu H, Wang D, Liu C, Luo X, Gao Y, Li X, Wu Z, Yang L, Meng F, Ning X, Hashenqimuge H, Wu K, Wang B, Suyalatu S, Liu Y, Ye C, Wu H, Leppälä K, Li L, Fang L, Chen Y, Xu W, Li T, Liu X, Xu X, Gignoux CR, Yang H, Brody LC, Wang J, Kristiansen K, Burenbatu B, Zhou H, Yin Y. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. *Nat Genet*, 2018, 50(12): 1696–1704. [DOI]

(责任编辑: 张勇)