

CNGBdb: 国家基因库生命大数据平台

陈凤珍¹, 游丽金¹, 杨帆¹, 王丽娜¹, 郭学芹¹, 高飞¹, 华聪¹, 谈聪¹,
方林², 单日强³, 曾文君¹, 王博¹, 王韧¹, 徐讯^{1,2,4}, 魏晓锋¹

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
3. 深圳华大智造科技有限公司, 深圳 518083
4. 广东省高通量基因组测序与合成编辑应用重点实验室, 深圳 518120

摘要: 国家基因库生命大数据平台(China National GeneBank DataBase, CNGBdb)是一个致力于生命科学多组学数据归档和开放共享的数据库平台, 是深圳国家基因库的核心功能“三库两平台”中生物信息数据库的对外服务平台, 拥有深圳国家基因库丰富的样本资源、数据资源、合作项目资源和强大的数据计算和分析能力等优势。生命科学研究已经进入到了一个以高通量多组学数据为基础的大数据时代, 迫切需要加强国际合作和信息共享。随着中国经济的发展和在生命科学研究领域的研究项目投入力度的加大, 需要建立相关的生命大数据归档和共享的平台, 来促进我国生命科学研究项目中生成的基因组学数据的系统管理、开放共享与合理利用。目前, CNGBdb 主要提供生命科学研究相关的数据归档、知识搜索、数据管理、数据计算和数据服务等。其归档和共享的数据类型, 主要包括项目、样本、实验、测序、组装、变异、序列等。截止 2020 年 5 月 22 号, CNGBdb 已接受了全球生命科学科研工作者提交的研究项目达 2176 个, 归档的基因组学数据量超过 2221 TB。未来, CNGBdb 将继续推动生命科学研究多组学数据的开放共享和产业应用, 完善基因组学数据的归档和共享功能, 提升其服务生命科学数据开放共享的能力。CNGBdb 的网址是: <https://db.cngb.org/>。

关键词: 国家基因库生命大数据平台; 数据归档; 数据共享; 多组学数据

收稿日期: 2020-03-23; 修回日期: 2020-05-23

基金项目: 广东省高通量基因组测序与合成编辑应用重点实验室(编号: 2017B030301011)资助[Supported by Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011)]

作者简介: 陈凤珍, 本科, 研究方向: 生物大数据。E-mail: chenfengzhen@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

魏晓锋, 本科, 研究方向: 生物大数据。E-mail: weixiaofeng@cngb.org

DOI: 10.16288/j.ycz.20-080

网络出版时间: 2020/7/8 16:44:57

URI: <https://kns.cnki.net/kcms/detail/11.1913.R.20200707.1642.001.html>

CNGBdb: China National GeneBank DataBase

Fengzhen Chen¹, Lijin You¹, Fan Yang¹, Lina Wang¹, Xueqin Guo¹, Fei Gao¹, Cong Hua¹, Cong Tan¹, Lin Fang², Riqiang Shan³, Wenjun Zeng¹, Bo Wang¹, Ren Wang¹, Xun Xu^{1,2,4}, Xiaofeng Wei¹

1. China National GeneBank, Shenzhen 518120, China

2. BGI-Shenzhen, Shenzhen 518083, China

3. MGI-Shenzhen, Shenzhen 518083, China

4. Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen 518120, China

Abstract: China National GeneBank DataBase (CNGBdb) is a data platform aiming to systematically archiving and sharing of multi-omics data in life science. As the service portal of Bio-informatics Data Center of the core structure, namely, "Three Banks and Two Platforms" of China National GeneBank (CNGB), CNGBdb has the advantages of rich sample resources, data resources, cooperation projects, powerful data computation and analysis capabilities. With the advent of high throughput sequencing technologies, research in life science has entered the big data era, which is in the need of closer international cooperation and data sharing. With the development of China's economy and the increase of investment in life science research, we need to establish a national public platform for data archiving and sharing in life science to promote the systematic management, application and industrial utilization. Currently, CNGBdb can provide genomic data archiving, information search engines, data management and data analysis services. The data schema of CNGBdb has covered projects, samples, experiments, runs, assemblies, variations and sequences. Until May 22, 2020, CNGBdb has archived 2176 research projects and more than 2221 TB sequencing data submitted by researchers globally. In the future, CNGBdb will continue to be dedicated to promoting data sharing in life science research and improving the service capability. CNGBdb website is: <https://db.cn gb.org/>.

Keywords: China National GenBank Database; data sharing; data archiving; omics data

国家基因库生命大数据平台(China National Gene Bank DataBase, CNGBdb), 是深圳国家基因库(China National GeneBank, CNGB)^[1](以下简称“国家基因库”)核心功能“三库两平台”中生物信息数据库的对外服务平台。CNGB 是以公益性、开放性、支撑性、引领性为宗旨, 服务于国家战略的国家级创新科研及产业基础设施建设项目。其中, 生物信息数据库致力于存储人类健康及生物多样性相关的数字化遗传资源, 构建生物数据库及数据分析平台, 实现数据存储、分析的贯穿, 为后续科研及产业提供大数据源头保障。

随着基因组测序技术的飞速发展和测序成本的大幅下降, 生命科学研究已经进入到了以高通量多组学技术为基础的大数据时代。为了解决人类生存面临的诸多问题, 在过去的 20 多年里, 世界各国相继实施了一些大规模的包括人类、动植物和微生物

在内基因组测序项目, 如千人基因组项目^[2]、国际癌症基因组项目^[3]、水稻参考基因组项目^[4,5]、全球 3000 份水稻(*Oryza sativa* L.)种质资源测序项目^[6]、全球超过 2 万份大麦种质资源测序项目^[7]等。这些项目的实施促进了生命科学研究领域研究的快速发展, 特别是人类遗传疾病致病机制发现和动植物分子设计育种应用等领域。迄今, 世界范围有多达 11,508 种真核生物, 245,875 种原核生物和 35,746 种病毒样本经完成测序(依据 2020 年 4 月 17 日的 NCBI 已测序物种统计)。同时, 还有大量的正在进行或即将开始的大型基因组测序项目, 将导致基因组数据的爆炸式增长。

为了实现这些数据的安全保存和开放共享, 全球生命科学研究组织相继建立了 3 个国际生物数据库, 分别依托于美国国家生物信息中心(National Centre of Biotechnology, NCBI)的相关数据库^[8],

欧洲分子生物实验室(European Molecular Biology Laboratory)的欧洲生物信息研究所(European Bioinformatics Institute, EBI)系列数据库^[9]和日本国家遗传研究所的DNA数据库(the DNA Database of Japan, DDBJ)^[10]。这3个数据库的主要功能包括:(1)接收生物学领域研究人员提交在研究项目过程中生成的基因组测序数据,如测序仪下机数据,以及后续的生物信息分析结果数据,如组装的基因组序列和基因注释结果等;(2)维护覆盖人类、动植物及微生物的物种的参考基因组及基因注释信息,方便生物研究人员交流和使用。另外,还有大量由生物信息领域研究人员维护,同时由分子生物学领域研究人员逐一审核的高质量生物大分子知识数据库^[11],如依托于瑞士生物信息研究所(Swiss Institute of Bioinformatics, SIB)的系列生物数据库^[12]和由日本京都大学和东京大学联合开发的代谢途径/通路相关数据库(Kyoto Encyclopedia of Genes and Genomes, KEGG)^[13]。其中,NCBI、EMBL-EBI和DDBJ的核酸数据库组成了国际核酸序列数据库联盟(International Nucleotide Sequence Database Collaboration, INSDC)^[14],这3个核酸数据库之间,每日进行数据交换,在促进国际生物学数据的共享和利用方面发挥了重要作用。但是国外这3个核酸数据库的目的,主要还是促进其本国生物研究机构之间生命大数据的共享和合作。当其他国家人员使用这些数据库时,还是存在诸多不方便的地方,如网络基础设施、国家与国家之间合作态度的倾向,以及数据库维护人员与科研人员在沟通语言和方式等方面的限制。

随着中国经济的快速发展,中国政府正在加大科学研究的资助力度,特别是生物医学和现代农业领域。在过去的20年里,中国也相继实施了一些重大的基因组学研究项目,如炎黄基因组项目^[15]和大熊猫基因组项目^[16]等,生成了海量的基因组测序数据和大量珍贵的项目研究成果。目前,由中国不同研究机构分别承担的基因组学项目生成的生命科学相关数据和结果,面临着“数据孤岛”、“数据主权”等实际问题。为了更好地服务于中国的科研人员,管理好中国在基因组学领域重大项目实施过程中生成的数据,中国政府相关部门和生命科学研究共同体近几年已经开始布局并着手建设国家级的生命大数据平台或大数据中心,以解决中国生命科学大数

据产出面临的实际问题,促进基因组学数据的开放共享。建设属于中国自己的大型基因组数据库的基础设施,不仅可以更好地服务中国的科研人员,还可以在符合国家的利益和法律的前提下,促进与国际同行的信息数据合作与共享。目前,国内已经建成一定规模的生命科学数据中心主要有:依托于北京基因组研究所的国家基因组数据中心(National Genomics Data Center, NGDC)^[17,18]、依托于中科院微生物研究所的国家微生物科学数据中心(National Microbiology Data Center, NMDC)和依托于深圳国家基因库 CNGBdb 等。NGDC 平台(<https://bigd.big.ac.cn/>),除了支持组学原始数据归档,参考基因组及基因注释信息存储和查询,还建立了甲基化数据库,单核苷酸多态性数据库等多组学数据库系统以及以表观组关联分析为代表的综合数据系统^[19-21]。NMDC 平台(<http://nmcdc.cn/>),主要致力于微生物资源信息和微生物基因组数据的保存和共享,其整合的数据资源总量超过1PB,数据记录数超过40亿条。由NMDC平台维护的具有代表性的数据库资源主要有:微生物宏基因组数据库^[22],全球微生物菌种目录数据库^[23]和全球流感病毒数据库。

依托于国家基因库^[1]的生命大数据中心有以下优势:(1)国家基因库多年来开展的重大基因组项目,如万种鸟类基因组项目^[24]、万种鱼类基因组项目^[25]、千种植物转录组项目^[26]等,积累了海量珍贵数据资源;(2)国家基因库多年来已建成了世界级基因组高通量测序平台和高性能计算平台;(3)国家基因库与国内各省及其他国家相继合作开展的生物样本资源库及其数字化项目,如海洋生物样本资源库及数字化、云南药用植物资源样本资源库及数字化等项目;(4)国家基因库在长期大量基因组学项目中积累的生物信息分析能力和多组学数据深度整合的能力。国家基因库多年来积累的海量基因组学数据和强大的多组学数据计算分析和整合能力,将为CNGBdb提供丰富的生物数据资源和强有力的维护支撑能力。

本文将主要从数据归档、知识搜索、数据管理、数据计算和数据服务等方面介绍CNGBdb的相关功能模块和数据服务。目前,CNGBdb不仅归类存档了CNGB内部项目及与国内国际大量合作项目实施中产生的海量生物学数据,而且还支持研究人员在线提交包括项目、样本、实验、测序、组装和变异

数据信息。另外, CNGBdb 还积极与 NGDC、NMDC、SRA、ENA 和 DDBJ 等平台的依托单位开展合作交流, 促进与各大数据库平台之间数据交流与共享, 进而推动全球生命大数据资源的利用。

2 数据共享服务

2.1 数据归档

为提供便捷的测序数据归档和数据管理服务, CNGBdb 已构建了国家基因库序列归档系统(CNSA, <https://db.cngb.org/cnsa>)。CNSA 可以接受全球用户在线提交的生物研究项目、样本、实验、测序数据及后期项目研究结果等信息。CNSA 数据归档系统主要遵循了在全球生命科学领域广泛达成共识的 INSDC 和 DataCite 等数据库标准。CNSA 是一个测序数据归档和分享系统, 还提供早期数据的共享等服务, 方便科研文章在投稿过程中杂志编辑检查投稿文章中的数据是否已经全部成功上传。CNSA 系统采用了项目(project)、样本(sample)、实验(experiment)和测序(run) 4 个元数据结构进行原始测序数据的组织和归档。除原始数据归档外, CNSA 还支持组装数据、变异数据的在线批量归档。为了提高数据的通用性, CNSA 支持各种常用格式的数据文件的递交, 例如, 原始数据格式包括 FASTQ、BAM、SFF 和 PacBio_HDF5, 组装数据格式包含 FASTA, 变异数据格式包含 VCF 等。为了确保归档数据的完整性和提高其后续的可利用性, CNSA 对用户递交的数据进行校验和质控。在 CNSA 归档的数据, 递交者可以根据项目的保密级别以及研究进度, 自由决定归档数据的开放权限和开放时间等。

CNSA 自 2018 年 10 月上线以来, 其归档数据量快速增长。截至 2020 年 5 月 22 日, 在该平台归档的项目有 2176 个, 提交的数据量达到 2221 TB (图 1), 支撑文章发表 115 篇。为便于研究人员查找和利用数据, CNSA 为每个归档的项目分配 DOI, 索引项目。通过 DOI 为 CNSA 归档的数据能够在互联网环境下的访问建立便利的途径, 以增加人们对研究数据的认可, 将其作为对科学记录合法的、可引用的成果支持数据存档, 并允许这些数据在未来的研究中被验证以及被重新利用^[27]。

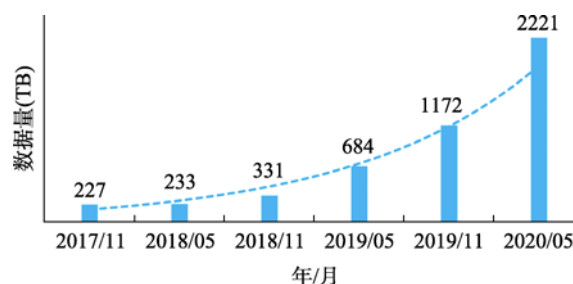


图 1 CNSA 归档数据量统计图

Fig. 1 Data statistics of CNSA

为实现活体资源、样本资源和数据资源的贯穿, 使得生命数据在全生命周期可追溯, 除归档核酸数据, CNGBdb 还构建了国家基因库样本信息共享平台(E-BioBank, EBB, <https://db.cngb.org/ebb/>), 支持活体资源和样本资源的递交和归档。EBB 制定了规范统一的样本信息整合标准, 建立了活体标本、样本、组学数据关联结构, 并创造公平、公开的生物样本共享环境, 促进生物样本的科学、合理利用, 是一个公益性、基础性、战略性的科技创新服务平台。E-BioBank 已归档 477,201 份样本, 1912 个物种, 23 个样本库。

2.2 知识搜索

除了国家基因库“三库两平台”的生命科学大数据资源, CNGBdb 还整合很多外部数据库的优秀数据资源, 如科研文献、基因、变异、蛋白质和序列等知识数据。为了使用户能够快速准确的检索到其需要的数据和信息, CNGBdb 平台中搭建了生命大数据搜索引擎。

CNGBdb 知识搜索的数据类型主要包括文献、项目、样本、实验、测序、组装、变异、基因、蛋白质、序列等。目前 CNGBdb 中可检索的知识条目数超过 30 亿条(表 1), 其中可被检索的文献数量超过 2947 万条记录, 基因序列超过 2274 万条记录, 蛋白质序列超过 22.7 亿条记录。CNGBdb 中的科研文献信息, 来源于对多个文献数据库系统的数据的整合, 包括 GigaScience、PubMed 和 Europe PMC 等。CNGBdb 知识检索服务, 可通过平台首页(<https://db.cngb.org/>)搜索入口, 选择不同的数据库, 进行跨多个数据库或者单个数据库高效快速检索。用户可在搜索输入框内输入任意的有意义的词或是编号来

表 1 知识搜索服务数据

Table 1 Data of knowledge search service

数据类型	索引量(万)	主要外源数据库	主要信息
文献	2947.19	GigaScience、PubMed 和 Europe PMC	文献标题、摘要、医学关键词、引用和参考文献和文献相关数据等
基因	2274.41	NCBI Gene	基因名称、染色体位置、基因产物和它的属性、基因所在的基因组、基因序列和基因变异等
变异	76323.01	dbSNP ^[28] 、dbVar 和 ClinVar ^[29]	变异名称(HGVS 名称)、基因组位置、相关物种、人群频率以及变异数据与疾病、表型和文献等
蛋白	13406.59	Uniprot ^[30]	蛋白名称、蛋白长度、物种和编码蛋白的基因等
序列	213665.12	NCBI Refseq ^[31] 和 GenBank ^[8]	序列名称、序列长度、物种和 fastq 序列文件等
项目	35.63	NCBI BioProject ^[32]	项目的名称、描述和数据类型等
样本	1007.36	NCBI BioSample ^[8]	样本的名称、物种、样本类型和描述等
实验	5515.46	NCBI SRA ^[33]	实验的题目、测序平台、文库构建策略、文库来源和文库选项等
组装	0.24	NCBI Assembly ^[34]	组装的名称、分子类型、测序技术和组装方法等

查找相关的信息。除此之外, CNGBdb 库与库之间的信息进行交叉互链, 形成数据信息的互联互通, 方便数据的关联查询和检索, 如搜索变异数据库, 除可检索到变异信息, 也可查看到变异关联物种、基因和文献等信息。这种数据互通互联的方式, 极大提升内容的检索效率, 便于用户进行相关知识的理解和深入研究。

CNGBdb 的知识搜索服务, 基于 Elasticsearch 搜索引擎, 支持全文检索功能^[2,35], 检索速度快。搜索引擎可对检索的结果进行综合评分排序, 将最匹配的最符合用户检索目的数据排在前列, 通过数据编号索引可以查看检索出的每一条数据的详细信息。CNGBdb 搜索引擎还实现了分布式的实时文件存储, 每个字段都被索引并可被搜索, 可以扩展到上百台服务器, 处理 PB 级结构化或非结构化数据, 提供更加深层次的数据、信息和知识的关联关系。

在 Elasticsearch 的基础上, CNGBdb 还拓展了基于生物数据特征的辅助搜索功能, 如文献推荐功能、同义词转换功能、高级检索功能和过滤检索功能。文献推荐功能, 根据文献的发表年份、杂志影响因子、作者、医学主题词等构建算法模型, 综合打分, 进行文献推荐, 帮助用户查找到与正在查阅的文献最相关的文献, 有助于其进行深入阅读和研究。为更深入地理解用户的检索意图, CNGBdb 搜索配置了物种同义词(同义词表主要来源于 NCBI 物种分类数据库^[3,36])及医学主题词(同义词表主要来源于 NCBI 医学主题词库^[37]), 在检索某个关键词的

时候, 该关键词的同义词也能检索到, 例如 *Oryza sativa*, 其学名为 *Oryza sativa* L., 常用名为 rice, Inherited blast name 为 monocots。您在检索 *Oryza sativa* 时, 它的同义词 *Oryza sativa* L.、rice 和 monocots 也能被检索到。高级检索功能, 可以帮助用户实现对指定字段进行检索, 如指定文献的标题、作者、期刊等字段进行检索。过滤检索功能, 可以根据用户设置的过滤条件实现对检索结果快速准确的过滤, 如根据文献是否免费进行免费全文检索, 根据文献发表年限, 对不同年限的数据进行过滤检索, 根据物种类型, 对不同类型物种数据进行过滤检索, 使得 CNGBdb 检索更加准确。

除此之外, CNGBdb 搜索引擎还结合了人工智能的智能语义识别和知识图谱技术, 使得搜索更加智能。在智能语义识别方面, CNGBdb 搜索系统可以实现自动补全功能和文本纠错功能。自动补全功能是能根据用户的输入的检索词自动识别用户的检索意图, 进行自动补全。文本纠错功能, 可对用户输入的错别词进行自动纠错, 使用正确的检索词进行检索, 如输入 “*Oryza sativa*”, 系统将识别为 “*Oryza sativa*” 后进行检索。在知识图谱技术方面, CNGBdb 构建了文献-作者-研究领域知识图谱, 通过文献引用与被引用的关系, 文献、作者和医学主题词(Medical Subject Headings, MeSH)关联关系, 构建文献-作者-研究领域知识图谱。文献知识图谱, 旨在帮助用户快速锁定某个研究方向的重要文献和同领域内具有重要影响力研究人员, 建立某个研究领域的发

用于打通 CNGBdb 的各数据库数据。UMS 系统还可以对用户在各个数据库的数据权限进行统一的授权和管理。为了最大化地提高平台的利用率, UMS 系统提供了各种丰富的 API 接口供各数据库使用, 主要有注册 API、登录验证 API、用户信息修改 API 和密码修改 API 等。

2.3.2 数据分类分级管理

CNGBdb 制定了数据资源分类和数据访问形式分类机制, 进行数据分类分级保护和统一管理。

在数据资源分类方面, CNGBdb 数据的资源类型分为去身份识别的人类遗传资源、生物多样性资源, 以及人源微生物资源, 定义如下: (1) 人类遗传资源数据: 是指利用人类遗传资源材料产生的数据等信息资料, 是未经过深层处理, 未过滤掉人体基因组信息的数据; (2) 生物多样性资源数据: 是指动物、植物以及微生物等物种资源的数据; (3) 人源微生物资源数据: 人源微生物是指微生物研究(包括培养以及宏基因组测序研究)的样本来源是人, 其本质是微生物。人源微生物资源数据又分为已过滤掉人体基因组的人源微生物资源数据和未过滤掉人体基因组的人源微生物资源数据。

在数据访问形式方面, CNGBdb 数据的访问形式包括公开、受控管理形式。(1) 公开: 数据公开是指元数据和数据文件都公开。数据递交者需要设置一个公开日期后, 元数据和数据文件都将在该公开日期公开, 公开数据将展示在 CNGBdb, 且面向全球开放, 用户可在 CNGBdb 自由访问或使用。(2) 受控: 即项目关联的元数据公开和数据文件受控。数据递交者需要设置一个元数据的公开日期, 元数据都将在该公开日期公开, 数据文件受控。受控数据仅在 CNGBdb 上展示元数据, 数据文件受控管理, 具有数据访问权限的用户可使用受控的数据。

2.4 数据计算

CNGBdb 数据计算服务是基于 CNGBdb 清洗和归档的数据部署的 BLAST 序列比对服务(<https://db.cngb.org/blast/>)。BLAST 功能基于 NCBI BLAST+ 2.8.1 standalone 版本开发, 支持大部分 NCBI BLAST 数据库的序列比对, 并逐步整合 CNGB 的公开特色

数据集, 如千种植物转录组数据集、万种鸟基因组项目数据集和千种鱼转录组数据集等。用户可根据研究需要, 自定义的设置比对数据集, 进行更加精准的比对分析, 为各领域的组学研究提供高效便捷的序列搜索服务。CNGBdb BLAST 比对数据资源列表见表 2。

2.5 数据应用

CNGBdb 基于底层数据结构和数据, 构建了包括动物、植物、微生物等不同专题数据库及分析数据库系统。目前 CNGBdb 已上线的上层应用数据库包括: 千种植物数据库(OneKP, <https://db.cngb.org/onekp/>)、万种鸟基因组数据库(B10K, <https://b10k.genomics.cn/>)、千种鱼转录组数据库(FishT1K, <https://db.cngb.org/fisht1k/>)、千种昆虫转录组进化研究数据库(1KITE, <https://1kite.cngb.org/>)、万种植物数据库(10KP, <https://db.cngb.org/10kp/>)、癌症数据集成与整合分析平台(DISSECT, <https://db.cngb.org/dissect/>)、微生物组数据库人类微生物数据库(Microbiome, <https://db.cngb.org/microbiome/>)、罕见病数据库(GDRD, <https://db.cngb.org/gdrd/>)、病原数据库(PVD, <https://db.cngb.org/pvd/>)和免疫数据库(PIRD, <https://db.cngb.org/pird/>)等。

为便捷和及时地共享科研数据, 在 CNGBdb 数据库平台, 除 CNGBdb 已经构建的不同研究领域的数据库, 还允许用户自定义创建数据集并共享发布。相比于传统的数据库共享, 用户不需要开发数据库、运营和维护数据库。在 CNGBdb 仅需上传数据、创建数据集和分享数据集 3 步, 即可将科研数据分享给科研领域的研究人员。CNGBdb 用户已创建的部分数据集见表 3。

3 结语与展望

CNGBdb 是一个自由开放的生命科学大数据共享平台, 致力于促进生命科学研究项目中生成的测序数据及研究项目所取得的成果的开发共享和合作利用。目前, CNGBdb 提供生物大数据归档、管理、搜索、计算、分析及应用一体化的生命大数据服务。

表 2 BLAST 工具数据资源

Table 2 Data resources of BLAST

数据源	数据库名称	数据库编号	数据库格式版本
CNGB	The 1000 Plants Project	onekp	v5
	Microbiome DataBase	microbiome	v5
	Pan Immune Repertoire Database	pird	v5
	The Transcriptomes of 1,000 Fishes Project	fisht1k	v5
	The Bird 10,000 Genomes Project	b10k	v5
NCBI	Nucleotide collection	nt	v5
	Reference proteins	refseq_protein	v4
	16S ribosomal RNA sequences	16smicrobial	v4
	Human genomic	human_genomic	v4
	RefSeq Representative genomes	refseq_representative_genomes	v4
	Non-human organisms genomic	other_genomic	v4
	Human RefSeqGene sequences	refseqgene	v4
	Genomic survey sequences	gss	v4
	Reference genomic sequences	refseq_genomic	v4
	High throughput genomic sequences	htgs	v4
	Transcriptome Shotgun Assembly	tsa	v4
	Expressed sequence tags	est	v4
	Patent sequences	pat	v4
	Sequence tagged sites	sts	v4
	Protein Data Bank	pdb	v4
	Metagenomic sequences	env	v4

表 3 CNGBdb 用户已创建的部分数据集

Table 3 Some datasets created by CNGBdb users

分类	数据集名称	简要介绍	网址
植物	10,000 Plant Genomes Project	万种植物基因组项目数据集	https://db.cngb.org/datamart/plant/DATApla1/
	The 3000 Rice Genomes Project	3000 水稻项目数据集	https://db.cngb.org/datamart/plant/DATApla2/
	1000 Plant Transcriptomes	千种植物转录组项目数据集	https://db.cngb.org/datamart/plant/DATApla4/
	Data of Ruili Botanical Garden	瑞丽珍稀植物园 689 种植物基因组测序数据	https://db.cngb.org/datamart/plant/DATApla5/
动物	The B10K Genomes Project	万种鸟基因组项目数据集	https://db.cngb.org/datamart/animal/DATAani1/
	1K Insect Transcriptome	千种昆虫转录组数据集	https://db.cngb.org/datamart/animal/DATAani3/
	Transcriptomes of 1000 Fishes	千种鱼转录组数据集	https://db.cngb.org/datamart/animal/DATAani2/
	Vertebrate Genomes 10K	万种脊椎动物数据集	https://db.cngb.org/datamart/animal/DATAani5/
	Life Periodic Plan (LPP)	生命周期表项目数据集	https://db.cngb.org/datamart/animal/DATAani6/
微生物	1520 reference genomes	覆盖人体肠道中所有主要细菌门和属的 1520 个基因组数据集	https://db.cngb.org/datamart/microbe/DATAmic1/
	Earth Microbiome Project	地球微生物组项目数据集	https://db.cngb.org/datamart/microbe/DATAmic4/
	1000 Fungal Genomes Project	千种真菌基因组项目数据集	https://db.cngb.org/datamart/microbe/DATAmic7/
	MetaHIT (metagenomics of humanintestinal tract)	欧盟的肠道微生物组计划数据集	https://db.cngb.org/datamart/microbe/DATAmic5/
	Human Microbiome Project	人类微生物组项目数据集	https://db.cngb.org/datamart/microbe/DATAmic3/
人群	WGS of 175 Mongolians	175 个蒙古人基因组数据集	https://db.cngb.org/datamart/other/DATAoth1/
	1000 Genomes Project (human)	千人基因组项目数据集	https://db.cngb.org/datamart/other/DATAoth2/

随着对生命科学大数据共享的需求的不断变化, CNGBdb 将在以下几个方面做出改进和提升。在数据归档上, 除归档项目、样本、实验/测序数据、组装、变异数据和样本实体信息, 在实现样本实体到组学数据的贯穿基础上, CNGBdb 还将扩展序列、蛋白、代谢、表达、临床和影像等多组学数据, 实现数据的组学贯穿。在知识搜索上, 除提供给用户主动搜索, CNGBdb 还将提供数据及知识推荐搜索, 实现主被动搜索联动, 提升搜索的准确度和搜索体验。在数据管理上, CNGBdb 将依据现有的伦理规范、现行法律、法规、条例、国际条约等, 制定更加完善的数据共享和应用政策。同时, CNGBdb 还将逐步建立数据可信计算环境和工具, 使得数据在可用而不可见的环境下进行安全计算, 并依托区块链技术, 对数据生命周期进行记账和监控, 实现生命科学数据的安全管理和应用。在数据应用上, CNGBdb 将在现有数据集的功能基础上, 提供更个性化、便捷化的多维度的统计分析工具, 数据比对工具, 数据可视化工具等, 实现数据的分享到数据应用的个性化、自动化。

CNGBdb 的建设和发展, 将促进我国生物遗传数据与生命科学数据的规范管理和利用, 为生物医药、生物农业和海洋生物等诸多生物产业的科学研究提供数据共享平台, 推动我国生命科学向更深入、更为广阔和更多创新的领域发展。CNGBdb 作为国家基因库的对外数据共享平台, 不仅促进扩大国内、国际交流与合作的范围, 还促进国内外生命科学数据的汇集、交流和互通。

参考文献(References):

- [1] Wang B, Liu F, Zhang EC, Wo CL, Chen J, Qian PY, Lu HR, Zeng WJ, Chen T, Wei JP, Wan Q, Wang R, Xu X. The China National GeneBank owned by all, completed by all and shared by all. *Hereditas(Beijing)*, 2019, 41(8): 761–772.
王博, 刘芳, 张二春, 沃晨亮, 陈振家, 钱璞毅, 卢浩荣, 曾文君, 陈泰, 危金普, 万仟, 王韧, 徐讯. 国家基因库: 共有、共为、共享. *遗传*, 2019, 41(8): 761–772. [DOI]
- [2] Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, Tassé AM, Flicek P. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Res*, 2017, 45(D1): D854–D859. [DOI]
- [3] Consortium ICG. International network of cancer genome projects. *Nature*, 2010, 464(7291): 993–938. [DOI]
- [4] Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, Cao ML, Liu J, Sun JD, Tang JB, Chen YJ, Huang XB, Lin W, Ye C, Tong W, Cong LJ, Geng JN, Han YJ, Li L, Li W, Hu GQ, Huang XG, Li WJ, Li J, Liu ZW, Li L, Liu JP, Qi QH, Liu JS, Li L, Li T, Wang XJ, Lu H, Wu TT, Zhu M, Ni PX, Han H, Dong W, Ren XY, Feng XL, Cui P, Li XR, Wang H, Xu X, Zhai WX, Xu Z, Zhang JS, He SJ, Zhang JG, Xu JC, Zhang KL, Zheng XW, Dong JH, Zeng WY, Tao L, Ye J, Tan J, Ren XD, Chen XW, He J, Liu DF, Tian W, Tian CG, Xia HG, Bao QY, Li G, Gao H, Cao T, Wang J, Zhao WM, Li P, Chen W, Wang XD, Zhang Y, Hu JF, Wang J, Liu S, Yang G, Zhang GY, Xiong YQ, Li ZJ, Mao L, Zhou CS, Zhu Z, Chen RS, Hao BL, Zheng WM, Chen SY, Guo W, Li GJ, Liu SQ, Tao M, Wang J, Zhu LH, Yuan LP, Yang HM. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 2002, 296(5565): 79–92. [DOI]
- [5] International RGSP. The map-based sequence of the rice genome. *Nature*, 2005, 436(7052): 793–800. [DOI]
- [6] RGP. The 3,000 rice genomes project. *GigaScience*, 2014, 3: 7. [DOI]
- [7] Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpfer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo GG, Xu DD, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao YS, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N. Genebank genomics highlights the diversity of a global barley collection. *Nat Genet*, 2019, 51(2): 319–326. [DOI]
- [8] Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. *Nucleic Acids Res*, 2019, 47(D1): D94–D99. [DOI]
- [9] Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*, 2019, 47(W1): W636–W641. [DOI]
- [10] Kodama Y, Mashima J, Kosuge T, Ogasawara O. DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acids Res*, 2019, 47(D1): D69–D73. [DOI]
- [11] Rigden DJ, Fernández XM. The 2018 Nucleic Acids

- Research database issue and the online molecular biology database collection. *Nucleic Acids Res*, 2018, 46(D1): D1–D7. [DOI]
- [12] Members SIB. The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res*, 2016, 44(D1): D27–D37. [DOI]
- [13] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, 2017, 45(D1): D353–D361. [DOI]
- [14] Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database Collaboration. The international nucleotide sequence database collaboration. *Nucleic Acids Res*, 2016, 46(D1): D48–D51. [DOI]
- [15] Wang J, Wang W, Li RQ, Li YR, Tian G, Goodman L, Fan W, Zhang JQ, Li J, Zhang JB, Guo TR, Feng BX, Li H, Lu Y, Fang XD, Liang HQ, Du ZL, Li D, Zhao YQ, Hu YJ, Yang ZZ, Zheng HC, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan JJ, Zhou Y, Qin JJ, Ma LJ, Li GQ, Yang ZT, Zhang GJ, Yang B, Yu C, Liang F, Li WJ, Li SC, Li DW, Ni PX, Ruan J, Li QB, Zhu HM, Liu DY, Lu ZK, Li N, Guo GW, Zhang JG, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su YY, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng HK, Ren YY, Yang L, Gao Y, Yang GH, Li Z, Feng XL, Kristiansen K, Wong GKS, Nielsen R, Durbin R, Bolund L, Zhang XQ, Li SG, Yang HM, Wang J. The diploid genome sequence of an Asian individual. *Nature*, 2008, 456(7218): 60–65. [DOI]
- [16] Li RQ, Fan W, Tian G, Zhu HM, He L, Cai J, Huang QF, Cai QL, Li B, Bai YQ, Zhang ZH, Zhang YP, Wang W, Li J, Wei FW, Li H, Jian M, Li JW, Zhang ZL, Nielsen R, Li DW, Gu WJ, Yang ZT, Xuan ZL, Ryder OA, Leung FCC, Zhou Y, Cao JJ, Sun X, Fu YG, Fang XD, Guo XS, Wang B, Hou R, Shen FJ, Mu B, Ni PX, Lin RM, Qian WB, Wang GD, Yu C, Nie WH, Wang JH, Wu ZG, Liang HQ, Min JM, Wu Q, Cheng SF, Ruan J, Wang MW, Shi ZB, Wen M, Liu BH, Ren XL, Zheng HS, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie XY, Lu ZH, Zheng HC, Li YR, Steiner CC, Tsan-Yuk Lam T, Lin SY, Zhang QH, Li GQ, Tian J, Gong TM, Liu HD, Zhang DJ, Fang L, Ye C, Zhang JB, Hu WB, Xu AL, Ren YY, Zhang GJ, Bruford MW, Li QB, Ma LJ, Guo YR, An N, Hu YJ, Zheng Y, Shi YY, Li ZQ, Liu Q, Chen YL, Zhao J, Qu N, Zhao SC, Tian F, Wang XL, Wang HY, Xu LZ, Liu X, Vinar T, Wang YJ, Lam TW, Yiu SM, Liu SP, Zhang HM, Li DS, Huang Y, Wang X, Yang GH, Jiang Z, Wang JY, Qin N, Li L, Li JX, Bolund L, Kristiansen K, Wong GKS, Olson M, Zhang XQ, Li SG, Yang HM, Wang J, Wang J. The sequence and de novo assembly of the giant panda genome. *Nature*, 2010, 463(7279): 311–317. [DOI]
- [17] Members NGDC. Database resources of the national genomics data center in 2020. *Nucleic Acids Res*, 2020, 48(D1): D24–D33. [DOI]
- [18] Ma YK, Bao YM. Prospects for national biological big data centers. *Hereditas(Beijing)*, 2018, 40(11): 938–943. 马英克, 鲍一明. 国家级生物大数据中心展望. *遗传*, 2018, 40(11): 938–943. [DOI]
- [19] Wang YQ, Song FH, Zhu JW, Zhang SS, Yang YD, Chen TT, Tang BX, Dong LL, Ding N, Zhang Q, Bai ZX, Dong XN, Chen HX, Sun MY, Zhai S, Sun YB, Yu L, Lan L, Xiao JF, Fang XD, Lei HX, Zhang Z, Zhao WM. GSA: genome sequence archive. *Genomics Proteomics Bioinformatics*, 2017, 15(1): 14–18. [DOI]
- [20] 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]
- [21] Zhang SS, Chen TT, Zhu JW, Zhou Q, Chen X, Wang YQ, Zhao WM. GSA: genome sequence archive. *Hereditas(Beijing)*, 2018, 40(11): 1044–1047. 张思思, 陈婷婷, 朱军伟, 周晴, 陈旭, 王彦青, 赵文明. GSA: 组学原始数据归档库. *遗传*, 2018, 40(11): 1044–1047. [DOI]
- [22] Shi WY, Qi HY, Sun QL, Fan GM, Liu SJ, Wang J, Zhu BL, Liu HW, Zhao FQ, Wang XC, Hu XX, Li W, Liu J, Tian Y, Wu LH, Ma JC. gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res*, 2019, 47(D1): D637–D648. [DOI]
- [23] Wu LH, Sun QL, Sugawara H, Yang S, Zhou YG, McCluskey K, Vasilenko A, Suzuki KI, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Philippe D, Ma JC. Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. *BMC Genomics*, 2013, 14: 933. [DOI]
- [24] Zhang GJ. Bird sequencing project takes off. *Nature*, 2015, 522(7554): 34. [DOI]
- [25] Fan GY, Song Y, Huang XY, Yang LD, Zhang SY, Zhang MQ, Yang XW, Chang Y, Zhang H, Li YX, Liu SS, Yu LL, Seim I, Feng CG, Wang W, Wang K, Wang J, Xu X, Yang

- HM, Chen NS, Liu X, He SP. Initial data release and announcement of the Fish10K: Fish 10,000 Genomes Project. *bioRxiv*, 2019, 787028. [DOI]
- [26] Initiative OTPT. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 2019, 574: 679–685. [DOI]
- [27] Paskin N. Digital object identifier (DOI®) system. *Encyclopedia of Library and Information Sciences*, 2010, 3: 1586–1592. [DOI]
- [28] Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res*, 2000, 28(1): 352–355. [DOI]
- [29] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu BS, Hart J, Hoffman D, Hoover J, Jang WH, Katz KK, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*, 2016, 44(D1): D862–D868. [DOI]
- [30] Consortium U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 2019, 47(D1): D506–D515. [DOI]
- [31] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 2007, 35: D61–D65. [DOI]
- [32] Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res*, 2012, 40(D1): D57–D63. [DOI]
- [33] Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, 2012, 40: D54–D56. [DOI]
- [34] Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res*, 2016, 44: D73–D80. [DOI]
- [35] Gormley C, Tong Z. Elasticsearch: the definitive guide: a distributed real-time search and analytics engine. “O’Reilly Media, Inc.”, 2015. [DOI]
- [36] Federhen S. The NCBI taxonomy database. *Nucleic Acids Res*, 2012, 40: D136–D143. [DOI]
- [37] Marc DT, Khairat SS. Medical Subject Headings (MeSH) for indexing and retrieving open-source healthcare data. *Stud Health Technol Inform*, 2014, 202: 157–160. [DOI]

(责任编辑:胡松年)