

生命与健康大数据中心资源

张源笙^{1,2,3}, 夏琳^{1,2,3}, 桑健^{1,2,3}, 李漫^{1,2,3}, 刘琳^{1,2,3}, 李萌伟^{1,2,3},
牛广艺^{1,2,3}, 曹佳宝^{1,2,3}, 滕徐菲^{1,2,3}, 周晴^{1,2,3}, 章张^{1,2,3}

1. 中国科学院北京基因组研究所, 生命与健康大数据中心, 北京 100101
2. 中国科学院北京基因组研究所, 中国科学院基因组科学与信息重点实验室, 北京 100101
3. 中国科学院大学, 北京 100049

摘要: 生命与健康多组学数据是生命科学研究和生物医学技术发展的重要基础。然而, 我国缺乏生物数据管理和共享平台, 不但无法满足国内日益增长的生物医学及相关学科领域的研究发展需求, 而且严重制约我国生物大数据整合共享与转化利用。鉴于此, 中国科学院北京基因组研究所于 2016 年初成立生命与健康大数据中心 (BIG Data Center, BIGD), 围绕国家人口健康和重要战略生物资源, 建立生物大数据管理平台和多组学数据资源体系。本文重点介绍 BIGD 的生命与健康大数据资源系统, 主要包括组学原始数据归档库、基因组数据库、基因组变异数据库、基因表达数据库、甲基化数据库、生物信息工具库和生命科学维基知识库, 提供生物大数据汇交、整合与共享服务, 为促进我国生命科学数据管理、推动国家生物信息中心建设奠定重要基础。

关键词: 大数据; 组学; 数据共享; 数据资源; 生物信息学

The BIG Data Center's database resources

Yuansheng Zhang^{1,2,3}, Lin Xia^{1,2,3}, Jian Sang^{1,2,3}, Man Li^{1,2,3}, Lin Liu^{1,2,3}, Mengwei Li^{1,2,3},
Guangyi Niu^{1,2,3}, Jiabao Cao^{1,2,3}, Xufei Teng^{1,2,3}, Qing Zhou^{1,2,3}, Zhang Zhang^{1,2,3}

1. BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
2. CAS Key Laboratory of Genomics and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
3. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Omics data in life and health sciences are of fundamental significance for scientific research and biomedical

收稿日期: 2018-07-05; 修回日期: 2018-09-12

基金项目: 中国科学院战略性先导科技专项(编号: XDA19050302, XDB13040500, XDA08020102), 国家重点研发计划(编号: 2016YFC0901603)和中国科学院“十三五”信息化建设专项(编号: XXH13505-05)资助[Supported by Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA19050302, XDB13040500, XDA08020102), the National Key Research & Development Program of China (No. 2016YFC0901603) and the 13th Five-year Informatization Plan of Chinese Academy of Sciences (No. XXH13505-05)]

作者简介: 张源笙, 硕士研究生, 专业方向: 生物信息学。E-mail: zhangyuansheng@big.ac.cn

夏琳, 博士研究生, 专业方向: 生物信息学。E-mail: xialin@big.ac.cn

桑健, 博士研究生, 专业方向: 生物信息学。E-mail: sangj@big.ac.cn

张源笙、夏琳和桑健并列第一作者。

通讯作者: 章张, 博士, 研究员, 研究方向: 生物信息学。E-mail: zhangzhang@big.ac.cn

DOI: 10.16288/j.yczs.18-190

网络出版时间: 2018/9/18 10:01:08

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20180918.1000.002.html>

technology development. However, there is yet to be a platform for biological data management and sharing in China, making it difficult to meet the development needs of biomedical and related fields and consequently leading to severe issues in big data management, sharing and translation. To address these issues, Beijing Institute of Genomics (BIG) of Chinese Academy of Sciences founded the BIG Data Center (BIGD) in 2016, which is dedicated to establish a biological big data management platform and multi-omics databases, with a particular focus on national population healthcare and important strategic biological resources. In this paper, we describe core database resources in BIGD, including GSA (Genome Sequence Archive), GWH (Genome Warehouse), GVM (Genome Variation Map), GEN (Gene Expression Nebulas), MethBank (Methylation Bank), BioCode and Science Wikis. Taken together, all these resources provide a series of services for data deposition, integration and sharing, laying solid foundations for enhancing national biological science data management and further promoting the construction of national bioinformatics center.

Keywords: big data; omics; data sharing; data resource; bioinformatics

随着高通量测序技术的迅猛发展及测序成本的不断下降,生命与健康数据迎来爆发式增长,我国已经成为世界上基因组数据产出大国。预计到 2023 年,我国将年产超过 200 PB 的基因组数据。然而,当前我国生命与健康数据存在两大问题:一是数据外流。虽然我国生物组学数据产量约占全球 40%,但是长期缺乏国际认可的国家级生物数据库系统,科学家们不得不把宝贵的数据资源提交至美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)、欧洲生物信息学研究所(European Bioinformatics Institute, EBI)及日本 DNA 数据库(DNA Data Bank of Japan, DDBJ),造成我国宝贵的生物遗传数据外流。二是数据孤岛。由于缺乏数据共享管理机制,宝贵的组学数据分散在国内不同实验室和机构内部,未能有效共享整合,从而形成数据孤岛,严重制约我国生物大数据的整合共享与转化利用。因此,从国家战略层面和行业发展出发,亟需建立面向我国科学数据管理的生物大数据汇交共享平台,研发面向我国人口健康和重要战略生物资源的多组学数据资源体系,彻底改变我国在世界上处于“测序大国、数据弱国”的尴尬地位。针对上述问题,中国科学院北京基因组研究所的生命与健康大数据中心,围绕国家人口健康和重要战略生物资源的组学数据,建设生物大数据汇交、共享与管理平台,研发生命与健康多组学数据资源体系,为促进我国生命科学数据管理、推动国家生物信息中心建设奠定重要基础。

1 生命与健康大数据中心的任务与使命

生命与健康大数据中心(BIG Data Center, BIGD)于 2016 年 2 月在中科院北京基因组研究所正式成立^[1,2]。BIGD 主要任务是围绕我国人口健康和重要战略生物资源,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,支撑服务我国公益性科学研究与产业创新发展。成立至今, BIGD 已建立形成涵盖多组学数据的生命与健康大数据资源系统,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,面向全球提供生物大数据资源共享与访问服务,形成支撑我国科学发现和产业发展的重要数据资源和基础条件。

2 面向全球服务的数据资源

2.1 组学原始数据归档库(GSA)

GSA(<http://bigd.big.ac.cn/gsa>)^[3,4]是一个面向全

球的组学原始数据汇交、存储、管理与共享的公共数据管理平台。作为国内首个被国际期刊认可的组学数据发布平台, GSA 面向全球接收不同测序平台产生的组学原始数据, 免费提供长期的存储、管理与共享服务。除了收录原始测序数据, GSA 还收录 BAM (Binary Alignment Map)、VCF (Variant Call Format) 等格式的二级分析文件。GSA 中的数据元素包括元数据和原始序列两个部分, 其中元数据可以按照由小到大的逻辑顺序建立多对一的关联关系, 即测序信息(Run)、实验信息(Experiment)、生物样本信息(BioSample)和生物项目信息(BioProject)。为确保与国际核酸序列数据库联盟(International Nucleotide Sequence Database Collaboration, 简称 INSDC)系统的兼容性, GSA 为用户递交的每一组数据分配唯一的数据获取号(Accession Number), 项目数据以“PRJCA”为前缀, 样本数据以“SAMC”为前缀, 测序数据以“CRR”为前缀。截至 2018 年 7 月, GSA 已接收来自全球 93 个研究机构的 308 名用户递交的数据, 包括 535 个 BioProjects, 21 843 个 BioSamples, 28 050 个 Experiments, 29 624 个 Runs, 覆盖 178 个物种, 累积存储组学原始数据超过 536TB。目前 GSA 已获得包括 *Cell*、*Nature*、*PNAS*、*AJHG*、*GPB*、*Cell Research* 等在内的 30 多个国际权威期刊认可, 支撑服务国家重大科研任务。

2.2 基因组数据库(GWH)

GWH (<http://bigd.big.ac.cn/gwh>) 提供多物种基因组序列和基因注释信息的汇交、存储、发布和共享等数据服务, 涵盖动物、植物、真菌、细菌等国家重要战略资源物种。GWH 遵循国际 INSDC 数据标准, 将数据组织成 3 个对象, 即生物样本信息(BioSample)、生物项目信息(BioProject)和基因组组装信息(Assembly)。GWH 存储的数据有两个来源: (1) 用户直接递交; (2) 整合已发布的重要物种基因组信息。针对用户递交数据, GWH 建立了严格的质量控制标准, 检查基因组序列 ID、序列内容、基因结构的完整性和一致性等。截至 2018 年 7 月, GWH 已收录 254 个物种基因组数据, 包括用户直接递交的 116 个物种的基因组数据(7 个动物、10 个植物、1 个真菌、74 个细菌、23 个古细菌、1 个病毒)和整

合了已发布的 138 个新发布的物种基因组(61 个动物和 77 个植物)。

2.3 基因组变异数据库(GVM)

GVM (<http://bigd.big.ac.cn/gvm>)^[5,6] 是集基因组变异数据汇交、管理、整合与检索的重要数据库, 提供多物种的遗传多样性信息及遗传变异与表型关联信息。GVM 以物种为单位, 收录其基因组中变异位点及其注释信息, 涉及的数据类型主要包括单核苷酸多态性(single nucleotide polymorphism, SNP)、小片段插入与缺失(insertion and deletion, InDel)等。区别于其他变异数据库(如 dbSNP), GVM 收录了我重要战略资源与生物多样性物种。截至 2018 年 7 月, GVM 已涵盖包括人, 畜牧如猪(*Sus scrofa*)、牛(*Bos taurus*)、羊(*Capra hircus*)、鸡(*Gallus gallus*)、鸭(*Anas platyrhynchos*)等, 农作物如水稻(*Oryza sativa*)、玉米(*Zea mays*)、高粱(*Sorghum bicolor*)、小麦(*Triticum aestivum*)、番茄(*Solanum lycopersicum*)、大豆(*Glycine max*)等在内的 25 个物种, 囊括了约 50 亿条变异信息。此外, 基于科研文献的人工审编, GVM 整合了基因型与表型(genotype-phenotype, G2P)关联信息, 包括 180 911 条人类的 G2P 信息和 13 262 条非人物种的 G2P 信息。同时, GVM 提供多条件关联组合检索及变异信息的可视化功能, 支持不同类型的基因组变异数据提交和下载。同时, 针对重要特色物种, 已建立虚拟中国人基因组数据库(Virtual Chinese Genome Database)^[7]、高粱基因组变异数据库(*Sorghum* Genome SNP Database)^[8]以及家狗基因组变异数据库(Dog Genome SNP Database)^[9]。

2.4 基因表达数据库(GEN)

GEN (<http://bigd.big.ac.cn/gen>) 是基因表达数据的汇集和共享平台, 系统整合多物种、多组织、多样本的基因表达数据, 为精准医学、分子育种、生物安全、生物多样性等研究提供基因表达数据信息。目前, GEN 已经涵盖了基于二代测序的哺乳动物如人(*Homo sapiens*)、猪(*S. scrofa*)和小鼠(*Mus musculus*)等、植物如水稻(*O. sativa*)等的表达数据以及基于人工审编的多物种内参基因(internal control gene)相关信息, 分别建立哺乳动物转录组数据库(Mammalian

Transcriptomic Database, MTD, <http://mtd.cbi.ac.cn>)^[10], 水稻表达数据库(Rice Expression Database, RED, <http://expression.ic4r.org>)^[11]和内参基因知识库(Internal Control Genes, ICG, <http://icg.big.ac.cn>)^[12]。随着高通量测序数据的积累, GEN 将支持更多物种的表达数据汇集与管理, 开发实现同源基因在多物种尺度下的表达谱可视化比较分析, 整合单细胞水平的基因表达数据, 为研究人员提供多维全面的数据信息与展示功能。

2.5 甲基化数据库(MethBank)

MethBank (<http://bigd.big.ac.cn/methbank>)^[13,14]整合人类和重要动植物的全基因组高精度 DNA 甲基化图谱, 提供界面友好的数据浏览、检索、分析和下载功能。MethBank 提供基因甲基化、差异甲基化、特异性甲基化、年龄相关甲基化等信息, 为精准医学、公共安全、动植物育种等研究提供重要基础数据资源和分析平台: (1)在人类衰老方面, 收录了 4577 个健康人外周血样本的 450K 芯片数据, 整合审编成 34 个不同年龄组的参比甲基化组(consensus reference methylomes, CRMs); (2)在动物方面, 整合两个模式动物(斑马鱼 *Danio rerio* 和小鼠 *M. musculus*)的配子与早期胚胎发育的 18 个单碱基精度甲基化组; (3)在植物方面, 整合 5 个重要经济作物, 包括水稻 (*O. sativa*)、大豆(*G. max*)、木薯(*Manihot esculenta*)、番茄(*S. lycopersicum*)和菜豆(*Phaseolus vulgaris*)不同发育阶段多个组织的 336 个单碱基精度甲基化组(single-base resolution methylomes)。此外, MethBank 提供在线分析工具 Age Predictor 和 IDMP (Identification of Differentially Methylated Promoter), 分别用于预测人的甲基化年龄和识别差异甲基化启动子。

2.6 生物信息工具库(BioCode)

BioCode (<http://bigd.big.ac.cn/biocode>)是整合开源生物信息学相关软件及工具的数据库, 其主要功能是分类搜集整理生物信息软件工具源代码、软件包以及重要元数据信息, 包括软件名称、功能描述、分类、发表文章、文章引用情况、开发人员信息及单位信息等。截止 2018 年 7 月, 已经收录软件包 6980 个, 主要来源于生物信息学领域期刊杂志, 包

括 *Bioinformatics*、*Genome Biology*、*Nucleic Acids Research* 等。BioCode 允许用户提交自己开发的工具及软件包, 实现了生物信息学软件工具的集中归档与管理, 从而支持生物信息学工具的一站式检索和公开访问, 不仅方便开发者托管、存档及发布生物信息学工具, 同时也可以让用户有效浏览、搜索和下载任何感兴趣的内容。

2.7 生命科学维基知识库(Science Wikis)

Science Wikis (<http://bigd.big.ac.cn/sciencewikis>)是基于维基(Wiki)技术或维基思想, 对生命科学知识和数据进行整合、集成和共享的数据库系统。Science Wikis 允许用户对生命科学知识和数据进行添加和编辑, 旨在利用集体智慧实现生物知识和数据的整合与审编。目前, Science Wikis 主要包括以下子库: (1) ICG: 内参基因知识库, 整合收录 209 个物种(73 种动物, 115 种植物, 12 种真菌及 9 种细菌)中的超过 700 个内参基因信息; (2) LncRNAWiki^[15]: 人类长非编码 RNA 知识库, 整合了 106 063 个人类的长非编码 RNA, 并对超过 1000 个有文献支持的 lncRNA 进行审编与注释; (3) RiceWiki^[16]现收录了粳稻与籼稻共 86 216 个基因, 并对超过 1000 个基因实现了高质量审编, 注释信息包含表达信息、功能信息、进化信息以及参考文献信息; (4) Database Commons: 全球生物数据库目录, 整合近 5000 个已发表的生物学数据库信息。

3 结语与展望

BIGD 在成立至今的两年建设阶段已取得了突破性进展, 2018 年被国际 *Nucleic Acids Research* 数据库专刊评价为全球主要生物信息数据中心之一。展望未来, 在数据资源建设方面, BIGD 将面向人口健康和重要战略生物资源, 不仅注重数据的汇交存储(deposition), 同时也充分考虑数据的整合挖掘(integration)与转化应用(translation), 建立以数据汇交为基础、以数据整合为途径、以挖掘分析为导向、以前沿领域为牵引的生物信息大数据中心, 形成具有中国特色的生物大数据汇交共享平台和多组学数据资源体系; 在方法技术方面, BIGD 将充分采

用多项前沿交叉技术(包括云计算、人工智能、深度学习、Wiki、生物审编等),应用于大数据中心共享访问平台建设。尽管前路漫漫,BIGD 将积极开展与 NCBI 和 EBI 等国际主要生物信息中心合作,联合国内其他生物信息资源管理和服务单位,共同推动国家生物信息中心建设。

参考文献(References):

- [1] BIG Data Center Members. The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res*, 2017, 45(Database Issue): D18–D24. [DOI]
- [2] BIG Data Center Members. Database resources of the BIG Data Center in 2018. *Nucleic Acids Res*, 2018, 46(Database Issue): D14–D20. [DOI]
- [3] Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, Tang B, Dong L, Ding N, Zhang Q, Bai Z, Dong X, Chen H, Sun M, Zhai S, Sun Y, Yu L, Lan L, Xiao J, Fang X, Lei H, Zhang Z, Zhao W. GSA: Genome sequence archive. *Genom Proteom Bioinf*, 2017, 15(1): 14–18. [DOI]
- [4] Zhang S, Chen T, Zhu J, Zhou Q, Chen X, Wang Y, Zhao W. 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]
- [5] Song S, Tian D, Li C, Tang B, Dong L, Xiao J, Bao Y, Zhao W, He H, Zhang Z. Genome Variation Map: a data repository of genome variations in BIG Data Center. *Nucleic Acids Res*, 2018, 46(Database Issue): D944–D949. [DOI]
- [6] Song S, Teng X, Xiao J. Database resources of the reference genome and genetic variation maps for the Chinese population. *Hereditas (Beijing)*, 2018, 40(11): 1048–1054.
宋述慧, 滕徐菲, 肖景发. 中国人群参考基因组及基因组变异图谱资源库. *遗传*, 2018, 40(11): 1048–1054. [DOI]
- [7] Ling Y, Jin Z, Su M, Zhong J, Zhao Y, Yu J, Wu J, Xiao J. VCGDB: a dynamic genome database of the Chinese population. *BMC Genomics*, 2014, 15: 265. [DOI]
- [8] Bai B, Zhao W, Tang B, Wang Y, Wang L, Zhang Z, Yang H, Liu Y, Zhu J, Irwin D, Wang G, Zhang Y. DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res*, 2015, 43(Database Issue): D777–D783. [DOI]
- [9] Luo H, Zhao W, Wang Y, Xia Y, Wu X, Zhang L, Tang B, Zhu J, Fang L, Du Z, Bekele W, Tai S, Jordan D, Godwin I, Snowdon R, Mace E, Jing H, Luo J. SorGSD: a sorghum genome SNP database. *Biotechnol Biofuel*, 2016, 9(1): 6. [DOI]
- [10] Sheng X, Wu J, Sun Q, Xian F, Li X, Sun M, Fang W, Chen M, Yu J, and Xiao J. MTD: a mammalian transcriptomic database to explore gene expression regulation. *Brief Bioinform*, 2017, 18(1): 28–36. [DOI]
- [11] Xia L, Zou D, Sang J, Xu X, Yin H, Li M, Wu S, Hu S, Hao L, Zhang Z. Rice Expression Database (RED): an integrated RNA-Seq-derived gene expression database for rice. *J Genet Genomics*, 2017, 44(5): 235–241. [DOI]
- [12] Sang J, Wang Z, Li M, Cao J, Niu G, Xia L, Zou D, Wang F, Xu X, Han X, Fan J, Yang Y, Zuo W, Zhang Y, Zhao W, Bao Y, Xiao J, Hu S, Hao L, Zhang Z. ICG: a wiki-driven knowledgebase of internal control genes for RT-qPCR normalization. *Nucleic Acids Res*, 2017, 46(Database Issue): D121–D126. [DOI]
- [13] Zou D, Sun S, Li R, Liu J, Zhang J, Zhang Z. MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. *Nucleic Acids Res*, 2015, 43(Database Issue): D54–D58. [DOI]
- [14] Li RJ, Liang F, Li MW, Zou D, Sun SX, Zhao YB, Zhao WM, Bao YM, Xiao JF, Zhang Z. MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic Acids Res*, 2018, 46(Database Issue): D288–D295. [DOI]
- [15] Ma L, Li A, Zou D, Xu X, Xia L, Yu J, Bajic VB, Zhang Z. LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res*, 2015, 43(Database Issue): D187–192. [DOI]
- [16] Zhang Z, Sang J, Ma L, Wu G, Wu H, Huang D, Zou D, Liu S, Li A, Hao L, Tian M, Xu C, Wang X, Wu J, Xiao J, Dai L, Chen LL, Hu S, Yu J. RiceWiki: a wiki-based database for community curation of rice genes. *Nucleic Acids Res*, 2014, 42(Database Issue): D1222–D1228. [DOI]