基因组科学数据的安全管理与应用

李茹姣[1]，张欣[1]，宋述慧[1]，王彦青[1]，邹东[1]，肖景发[1,2]，赵文明[1,2]，章张[1,2]，鲍一明[1,2]
1. 中国科学院北京基因组研究所（国家生物信息中心）国家基因组科学数据中心，北京 100101；
2. 中国科学院大学，北京 100049

摘要
基因组科学数据是人口健康和国家安全的重要战略资源，存好、管好和用好基因组科学数据具有重要意义。面对我国生物数据大量产出但因存储零散、缺乏系统监管而丢失和流失，以及严重依赖国际生物组学数据库的局面，亟须从国家层面建设我国自己的生物大数据管理体系。以国家基因组科学数据中心为例，阐述了基因组科学数据汇交共享体系和标准规范、数据安全管理机制，给出了数据挖掘与应用的典型案例，并从政策机制、基础设施、软件研发、学科建设、人才培养和国际合作等方面提出对策建议。

关键词
科学数据；基因组学；汇交共享；数据安全管理；数据应用

中图分类号：Q34　　文献标志码：A　　doi: 10.11959/j.issn.2096-0271.2022004

Safety management and application of genomics data

LI Rujiao[1], ZHANG Xin[1], SONG Shuhui[1], WANG Yanqing[1], ZOU Dong[1], XIAO Jingfa[1,2], ZHAO Wenming[1,2], ZHANG Zhang[1,2], BAO Yiming[1,2]
1. National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, Beijing 100101, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract
Genomics data is essential strategic resources for population health and national security. It is of great significance to deposit, manage and utilize genomics data in a scientific manner. China is a powerhouse in generating vast quantities of biological data, while facing the situation of data loss due to the isolated data storage and the lack of systematic data monitoring and management, also with the heavy dependency on international biological data centers. Therefore, it urgently calls for China's own life big data storage and management system at the national level. Taking the National Genomics Data Center for example, the deposition, sharing and safety management system and standards of genomics data were summarized, with data mining and application cases. Suggestions were also given on the aspects of policy-making, infrastructure, software research and development, principle building and talent development, as well as international cooperation.

Key words
scientific data, genomics, deposition and sharing, data safety management, data application

0 引言

科学数据是国家科技创新和经济社会发展的重要基础性战略资源，做好科学数据资源的汇交共享、安全管理与挖掘利用具有重要的科学意义和价值。2019年6月10日，科学技术部和财政部联合发布了《关于国家科技资源共享服务平台优化调整名单的通知》，公布了多个学科领域的20个国家科学数据中心。其中，国家基因组科学数据中心（National Genomics Data Center，NGDC）（以下简称中心）依托中国科学院北京基因组研究所（国家生物信息中心）建设。中心面向我国人口健康和社会可持续发展的重大战略需求，建立基因组科学数据汇交存储、安全管理、开放共享与整合挖掘的研究体系，研发基因组科学大数据前沿交叉与转化应用的新方法和新技术，其目标是成为国际领先的基因组科学数据中心，支撑我国生命与健康科学创新发展。

中心自成立以来，面向人口健康和重要战略生物资源，以“存好”“管好”和“用好”基因组科学数据的实际需求为前提，已初步建成具有自主知识产权、安全可控、涵盖国家人类遗传资源和重要战略生物资源的基因组科学数据资源体系[1]。中心汇聚全球数据，提供公共服务，形成了组学“数据—信息—知识”一体化资源系统，主要分为：①原始数据仓储，包括生物项目数据库（BioProject）、生物样本数据库（BioSample）、组学原始数据归档库（genome sequence archive，GSA）[2-3]、人类遗传资源组学原始数据归档库（genome sequence archive for human，GSA-Human）[4]等；②组学信息库，包括基因组数据库（genome warehouse，GWH）[5]、基因组序列变异库（genome variation map，GVM）[6-7]、基因表达数据库（gene expression nebulas，GEN）[8]、甲基化数据库（methylation bank，MethBank）[9-10]等；③组学知识库，包括水稻多组学数据资源（IC4R）[11]、犬类组学资源库（iDog）[12]、绵羊组学资源库（iSheep）[13]、2019新型冠状病毒信息库（RCoV19）[14-15]、动植物基因组变异-表型关联知识库（GWAS Atlas）[16]、表观组关联分析知识库（EWAS Atlas）[17]等；④在线工具和文献情报信息平台，包括生物大数据跨库搜索引擎BIG Search、基因组科学数据在线分析平台等。中心已获得国际同行的高度认可，被国际生物数据领域权威期刊*Nucleic Acids Research*（《核酸研究》）称为与美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）、欧洲生物信息学研究所（European Bioinformatics Institute，EBI）并列的“全球主要数据中心”[18]。中心解决了长期以来我国基因组科学数据汇交共享严重依赖国际数据库的问题，为国家基因组科学数据的汇交共享、安全管理和挖掘利用提供了重要支撑。

1 基因组科学数据安全管理

数据的安全管理指在数据的收集、存储、使用、加工、传输、提供、公开等过程中采取必要的措施，确保数据处于有效保护和合法利用的状态。中心严格遵循《科学数据管理办法》和《中华人民共和国数据安全法》等相关法规，目前已建立较完整的基因组科学数据汇交共享机制和安全管理规范，研发形成具有自主知识产权的数据库管理系统和共享平台，为我国基因组科学数据安全可控的汇交存储、共享管

理与有效利用提供重要保障。

对于所有用户递交的数据，如生物研究项目和生物样本元数据、组学原始数据、基因组序列数据、基因组变异数据等，中心借鉴国际核酸序列数据库联盟（International Nucleotide Sequence Database Collaboration，INSDC）[19]的数据汇交标准规范，分别建立相应数据管理系统对其进行收集和管理。在数据管理系统中，内置多套受控词表，提供在线向导化信息提交功能，规范化、结构化管理各类信息，并通过在线校验和人工审编实现信息的质控和审核，以此确保用户递交数据的完整性和可靠性。审核通过后，系统分别为每个递交到中心的项目、样本、数据分配唯一可识别的编号，作为检索和访问的标识。数据管理系统根据递交用户设定的数据公开时间进行可控管理，并依托中心高性能存储和异地容灾的备份机制，定期进行数据更新与异地备份，以全面保证数据的完整性与安全性。对于大型项目及数据，中心则提供高效、安全、专业化的项目分级管理。

特别强调的是，中心遵循《中华人民共和国人类遗传资源管理条例》的指导原则，对人类遗传组学数据资源采取如下六方面的安全管理机制和策略[2,4]。①在数据访问方式方面，面向人类遗传资源，提供公开访问和受控访问两种方式。受控访问的数据采用“申请—审核”的共享方式，即数据使用者需要先向数据管理委员会（data access committee，DAC）提交申请，审核通过后才有权限访问、下载并使用数据。②在身份安全认证方面，采取双重认证方式，用户除了需要通过单点登录（single sign-on，SSO）系统的密码认证，还需要在数据提交和申请下载的人工审核阶段进行项目负责人（principal investigator，PI）身份信息核实，以确保数据的可溯源性。③在数据上传权限方面，系统规定数据上传必须使用PI账号，且需对元数据信息进行脱敏，即不能包含受试者的隐私信息。④在数据存储空间方面，系统为每个用户提供独立的数据存储空间，有效避免不同用户之间相互干扰，降低信息泄露的可能性，充分确保数据的安全性和私密性。⑤在数据申请访问方面，为了保证数据访问安全，系统规定只有注册为PI的用户才能申请下载数据。⑥在存储策略和备份机制方面，针对不同访问级别的数据采用分级存储策略，并建立完善的多点备份和异地灾备机制，以确保数据的安全存储。

2 基因组科学数据挖掘应用

中心在做好数据资源存储和管理的同时，十分注重数据的整合及应用系统的建设，研发了一站式跨库检索系统和在线分析平台，并支撑国内外用户开展组学大数据挖掘应用研究，为科学技术部、国家自然科学基金委员会、中国科学院等资助的4 000多个项目提供数据汇交存储和共享管理服务。

2.1 生物大数据跨库搜索引擎

生物大数据跨库搜索引擎BIG Search是目前整合全球生物数据库数量最多的生物大数据跨库检索平台，为全球科研人员提供秒级响应、一站式的跨库检索服务，支撑生物大数据的快速发现与利用。BIG Search整合了中心28个重要的生物数据库资源[1]以及国内众多合作伙伴的39个生物数据库资源，包括北京市神经外科研究所江涛教授团队的中国脑胶质瘤基因组图谱数据库（CGGA）[20]、北

京大学崔庆华教授团队的长非编码RNA疾病数据库（LncRNADisease）[21]、北京大学高歌研究员团队的植物转录因子数据库（PlantTFDB）[22]、华中科技大学郭安源教授团队的动物转录因子数据库（AnimalTFDB）[23]，以及哈尔滨医科大学肖云教授团队的细胞标记物知识库（CellMarker）[24]等。此外，还整合了国际知名生物信息数据中心的数据资源，包括NCBI的35个数据资源库[25]和EBI的115个数据集[26]，累计数据索引量达到1 TB，记录数超过11.5亿条。

2.2 基因组科学数据在线分析平台

为了促进基因组科学数据的有效挖掘利用，中心已初步建立了基因组科学数据在线分析平台，目前主要包括：①序列比对在线分析工具，集成了生命科学领域最常用的序列比对软件BLAST（basic local alignment search tool）[27-28]，不仅整合了nt、nr、Swiss-Prot等常用的核酸和蛋白数据库，还发挥了中心的特色数据资源优势，提供多种特有的核酸、蛋白序列比对数据库，包括GWH转录本和蛋白序列库、GEN转录本和蛋白质序列库、新型冠状病毒基因组代表序列库、人类长非编码RNA数据库LncBook[29]、万种原生生物核酸和蛋白质序列库、水稻/高粱/胡蜂等特色物种基因库[30]；②冠状病毒在线分析平台[31]，由基因组拼接、序列比对、基因组注释、变异鉴定和注释、谱系和进化分析等11个模块组成，满足快速增长的新型冠状病毒基因组数据的分析需求，已为国际生物多样性与健康大数据联盟（Global Biodiversity and Health Big Data Alliance，BHBD）成员以及来自全国10多个重要口岸的海关检疫人员提供了线上或现场的使用培训，为国内外用户完成了11 628个病毒数据的分析任务。

2.3 基于多维组学数据的典型应用

中心建立的基因组科学数据多维资源体系为新型冠状病毒的分子溯源与传播演化、动植物分子育种与遗传改良、精准医学与人口健康等多个研究领域提供了强有力的数据和信息支撑。新型冠状病毒信息库RCoV19有效支撑了世界卫生组织的SARS-CoV-2全球溯源研究—中国部分[32]、北京新发地疫情分子溯源[33]和巴基斯坦境内早期新型冠状病毒传播演化规律[34]等研究工作，在全球抗疫过程中发挥了科技支撑作用。武汉大学研究团队对从新型冠状病毒肺炎（COVID-19）患者的支气管肺泡灌洗液（bronchoalveolar lavage fluid，BALF）和外周血单个核细胞（peripheral blood mononuclear cell，PBMC）样本中提取的RNA进行了转录组测序，揭示了新型冠状病毒肺炎患者支气管肺泡灌洗液与外周血单个核细胞的转录组学特征[35]，并将数据递交至GSA（CRA002390），该成果发表后得到了广泛的关注。华中农业大学的研究人员利用GVM中猪、马、牛、山羊、水牛、鸡、野马和熊猫等物种的高密度基因型数据，经过数据再分析与处理，构建了经基因型填补后的13个动物的高质量参考变异组，同时开发了专业数据库Animal-ImputeDB[36]，用于在线基因型估算、基因变异搜索和免费下载，为动物遗传育种和遗传改良提供了丰富的数据资源，促进了基因型填补在动物遗传研究中的应用。

3 结束语

在科学技术部及有关部门的大力支持

和资助下，中心在数据汇交共享、安全管理和挖掘应用等方面都取得了突破性进展，已建成涵盖国家人类遗传资源和重要战略生物资源的多组学数据资源体系，研发一站式跨库检索系统和在线分析平台，数据资源总量已超过10 PB，为公益性科学研究和产业创新发展，尤其是全球抗疫，提供了重要数据资源和科技支撑。然而，在生物数据统一汇交政策机制、基础设施和数据智能管理能力、生物信息专业人才队伍以及生物数据的国际互通共享等方面仍需极大的提升。为此，笔者提出如下建议。

● 加快完善生物信息资源共享的政策保障措施：加快推动建立科技信息公开制度，确保各类科技项目产生的科学数据能够全面、及时开放共享，健全科学数据共享管理过程中的保障机制。

● 加强生物信息基础设施建设和核心软件系统研发：以生命科学研究的实际需求为导向，建立面向生物信息大数据的基础设施环境，研发多维数据资源的生物数据库、信息库和知识库系统及其关键核心软件和工具，加大对生物信息算法、模型、软件、工具、数据库等方面的资助支持力度，切实形成综合性、权威性的生物信息数据库以及具有自主知识产权的核心软件。

● 加大我国生物信息学学科建设及人才队伍的培养：建议尽快推进生物信息学的学科布局和整体规划，提升生物信息学的学科级别，成立生物信息学一级学会，并在有较好基础的大学设立生物信息学院，以此加强基础人才培养，为未来我国生命科学领域的可持续发展提供充足的人才储备。

● 加强生物信息数据与资源的国际合作：一方面，根据国家“一带一路”倡议，加强与相关国家的科技合作和技术探讨，在生命科学领域开展联合研究，扩大我国生物信息数据体系的影响力；另一方面，加强国内外科学共同体的交流合作，探索与国际社会的数据交换和合作交流，保障资源的全球化利用，最大限度发挥数据的价值。

致谢

感谢国家基因组科学数据中心的陈梅丽、陈婷婷、杜政霖、郝丽丽、马利娜、唐碧霞、张思思等在本文撰写过程中给予的支持和帮助。

参考文献:

[1] CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2022[J]. Nucleic Acids Research, 2021.

[2] CHEN T T, CHEN X, ZHANG S S, et al. The genome sequence archive family: toward explosive data growth and diverse data types[J]. Genomics, Proteomics & Bioinformatics, 2021.

[3] WANG Y Q, SONG F H, ZHU J W, et al. GSA: genome sequence archive[J]. Genomics, Proteomics & Bioinformatics, 2017, 15(1): 14-18.

[4] 张思思, 陈旭, 陈婷婷, 等. GSA-Human: 人类遗传资源数据管理的公共系统[J]. 遗传, 2021(10): 988-993.
ZHANG S S, CHEN X, CHEN T T, et al. GSA-Human: genome sequence archive for human[J]. Hereditas (Beijing), 2021(10): 988-993.

[5] CHEN M L, MA Y K, WU S, et al. Genome warehouse: a public repository housing genome-scale data[J]. Genomics, Proteomics & Bioinformatics, 2021.

[6] LI C P, TIAN D M, TANG B X, et al.

Genome variation map: a worldwide collection of genome variations across multiple species[J]. Nucleic Acids Research, 2020, 49(D1): 1186–1191.

[7] SONG S H, TIAN D M, LI C P, et al. Genome variation map: a data repository of genome variations in BIG Data Center[J]. Nucleic Acids Research, 2018, 46(D1): 944–949.

[8] ZHANG Y S, ZOU D, ZHU T T, et al. Gene expression nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels[J]. Nucleic Acids Research, 2021.

[9] ZOU D, SUN S X, LI R J, et al. MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data[J]. Nucleic Acids Research, 2014, 43(D1): 54–58.

[10] LI R J, LIANG F, LI M W, et al. MethBank 3.0: a database of DNA methylomes across a variety of species[J]. Nucleic Acids Research, 2017, 46(D1): 288–295.

[11] SANG J, ZOU D, WANG Z N, et al. IC4R-2.0: rice genome reannotation using massive RNA-seq data[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(2): 161–172.

[12] TANG B X, ZHOU Q, DONG L L, et al. iDog: an integrated resource for domestic dogs and wild canids[J]. Nucleic Acids Research, 2018, 47(D1): 793–800.

[13] WANG Z H, ZHU Q H, LI X, et al. iSheep: an integrated resource for sheep genome, variant and phenotype[J]. Frontiers in Genetics, 2021, 12: 714852.

[14] SONG S H, MA L N, ZOU D, et al. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR[J]. Genomics, Proteomics & Bioinformatics, 2020, 18(6): 749–759.

[15] 赵文明, 宋述慧, 陈梅丽, 等. 2019新型冠状病毒信息库[J]. 遗传, 2020, 42(2): 212–221. ZHAO W M, SONG S H, CHEN M L, et al. The 2019 novel coronavirus resource[J]. Hereditas(Beijing), 2020, 42(2): 212–221.

[16] TIAN D M, WANG P, TANG B X, et al. GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals[J]. Nucleic Acids Research, 2019, 48(D1): 927–932.

[17] LI M W, ZOU D, LI Z H, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies[J]. Nucleic Acids Research, 2018, 47(D1): 983–988.

[18] RIGDEN D J, FERNÁNDEZ X M. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection[J]. Nucleic Acids Research, 2020, 49(D1): 1–9.

[19] KARSCH-MIZRACHI I, NAKAMURA Y, COCHRANE G, et al. The International Nucleotide Sequence Database Collaboration[J]. Nucleic Acids Research, 2012, 40(D1): 33–37.

[20] ZHAO Z, ZHANG K N, WANG Q W, et al. Chinese glioma genome atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients[J]. Genomics, Proteomics & Bioinformatics, 2021, 19(1): 1–12.

[21] BAO Z Y, YANG Z, HUANG Z, et al. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases[J]. Nucleic Acids Research, 2018, 47(D1): 1034–1037.

[22] JIN J P, TIAN F, YANG D C, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants[J]. Nucleic Acids Research, 2016, 45(D1): 1040–1045.

[23] HU H, MIAO Y R, JIA L H, et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors[J]. Nucleic Acids Research, 2018, 47(D1): 33–38.

[24] ZHANG X X, LAN Y J, XU J Y, et al. CellMarker: a manually curated resource of cell markers in human and mouse[J]. Nucleic Acids Research, 2019, 47(D1): 721–728.

[25] SAYERS E W, BOLTON E E, BRISTER J R, et al. Database resources of the national center for biotechnology information[J]. Nucleic Acids Research, 2021.
[26] CANTELLI G, BATEMAN A, BROOKSBANK C, et al. The European Bioinformatics Institute (EMBL-EBI) in 2021[J]. Nucleic Acids Research, 2021.
[27] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.
[28] ALTSCHUL S F, MADDEN T L, SCHÄFFER A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.
[29] MA L N, CAO J, LIU L, et al. LncBook: a curated knowledgebase of human long non-coding RNAs[J]. Nucleic Acids Research, 2019, 47(5): 2699.
[30] LIU Y M, WANG Z H, WU X Y, et al. SorGSD: updating and expanding the sorghum genome science database with new contents and tools[J]. Biotechnology for Biofuels, 2021, 14(1): 165.
[31] GONG Z, ZHU J W, LI C P, et al. An online coronavirus analysis platform from the National Genomics Data Center[J]. Zoological Research, 2020, 41(6): 705-708.
[32] Joint WHO-China Study Team. WHO-convened global study of origins of SARS-CoV-2: China part (text extract)[J]. Infectious Diseases & Immunity, 2021, 1(3): 125-132.
[33] PANG X H, REN L L, WU S S, et al. Cold-chain food contamination as the possible origin of COVID-19 resurgence in Beijing[J]. National Science Review, 2020, 7(12): 1861-1864.
[34] SONG S H, LI C P, KANG L, et al. Genomic epidemiology of SARS-CoV-2 in Pakistan[J]. Genomics, Proteomics & Bioinformatics, 2021.
[35] XIONG Y, LIU Y, CAO L, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients[J]. Emerging Microbes & Infections, 2020, 9(1): 761-770.
[36] YANG W Q, YANG Y B, ZHAO C C, et al. Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation[J]. Nucleic Acids Research, 2019, 48(D1): 659-667.

作者简介

李茹姣(1976-),女,博士,中国科学院北京基因组研究所(国家生物信息中心)高级工程师,主要研究方向为组学大数据整合和挖掘。

张欣(1981-),男,中国科学院北京基因组研究所(国家生物信息中心)工程师,主要研究方向为组学大数据合作共享。

宋述慧（1981– ），女，博士，中国科学院北京基因组研究所（国家生物信息中心）副研究员，主要研究方向为基因组变异大数据整合分析与挖掘应用。

王彦青（1982– ），女，中国科学院北京基因组研究所（国家生物信息中心）高级工程师，主要研究方向为基因组学原始数据汇聚、管理与共享体系构建。

邹东（1986– ），男，中国科学院北京基因组研究所（国家生物信息中心）高级工程师，主要研究方向为生物数据库系统研发、多维组学大数据跨库检索平台建设。

肖景发（1973– ），男，博士，中国科学院北京基因组研究所（国家生物信息中心）研究员，主要研究方向为多维组学数据整合挖掘和微生物泛基因组学算法软件开发等。

赵文明（1977– ），男，中国科学院北京基因组研究所（国家生物信息中心）高级工程师，国家基因组科学数据中心副主任，主要研究方向为生物信息大数据整合挖掘、生物信息工具与平台研发。

章张（1980– ），男，博士，中国科学院北京基因组研究所（国家生物信息中心）研究员，国家基因组科学数据中心副主任，主要研究方向为生物大数据整合与信息挖掘。

鲍一明(1965-),男,博士,中国科学院北京基因组研究所(国家生物信息中心)研究员,国家基因组科学数据中心主任,主要研究方向为生物数据库、病毒基因组注释、病毒进化与分类等。

收稿日期: 2021-12-10

通信作者: 鲍一明, baoym@big.ac.cn

基金项目: 国家重点研发计划资助项目(No.2018YFD1000505, No.2021YFC0863300);中国科学院战略性先导专项(No.XDB38030200);中国科学院基因组科学数据中心能力建设项目(No.XXH-13514-0202)

Foundation Items: The National Key Research and Development Program of China (No.2018YFD1000505, No.2021YFC0863300), Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDB38030200), Genomics Data Center Construction of Chinese Academy of Sciences (No.XXH-13514-0202)