中国科学: 生命科学 2024 年 第 54 卷 第 6 期: 1088 ~ 1100
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述 生物样本资源专辑

国家基因组科学数据中心: 现状及展望

陈梅丽[1†], 王彦青[1†], 李茹姣[1,2†], 马英克[1], 张思思[1], 张欣[1], 宋述慧[1,2], 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
† 同等贡献
* 联系人, E-mail: baoym@big.ac.cn

收稿日期: 2023-09-20; 接受日期: 2024-01-30; 网络版发表日期: 2024-05-13
中国科学院战略性先导科技专项(B类)(批准号: XDB38030200)和国家重点研发计划(批准号: 2021YFF0703704)资助

摘要 面向我国人口健康和社会可持续发展的重大战略需求, 国家基因组科学数据中心(National Genomics Data Center, NGDC)自2019年成立以来, 已初步建成具有自主知识产权、安全可控、涵盖领域广的多维组学数据汇交、存储、管理和共享体系, 涵盖基础组学数据资源、国家人类遗传资源、重要战略生物资源、生物安全资源以及生物信息分析工具和平台等, 为人口健康、公共安全、育种改良、生物多样性等相关研究提供重要资源和参考信息. 截至目前, NGDC已存储和管理27.6 PB的数据量, 数据编号被Springer Nature, Elsevier, Wiley, Taylor & Francis等全球主要出版集团推荐或认可. 尽管NGDC已连续六年被本领域国际权威期刊《核酸研究》称为与美国NCBI、欧洲EBI齐名的国际主要生物数据中心, 但与国际一流数据中心仍存在一定差距. 展望未来, NGDC将重点聚焦于数据智能审编、数据融合检索、生物大数据云平台、前沿算法工具等, 同时在经费争取、人才培养和国际合作方面加大工作力度, 建成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领域的科技创新发展和自立自强.

关键词 基因组, 生物信息, 大数据, 多组学, 人类遗传资源, 汇交管理, 国家基因组科学数据中心, 国家生物信息中心

科学数据被视为推动国家科技创新和促进经济社会发展的关键基础性战略资源[1]. 20世纪八九十年代, 国际上已成立了具有世界权威性的三个国家级生物信息中心[2], 引领全球生物信息学领域发展. 我国在生物信息数据的存储和管理方面起步晚, 为应对生物信息数据大量缺失、丢失和流失, 以及严重依赖国际生物信息数据库的问题, 2019年, 我国正式建立了国家基因组科学数据中心(National Genomics Data Center, NGDC). 虽然NGDC的建设取得了一定的进展, 但是与国际一流生物数据中心相比, 在人员及团队规模、经费体量及获取渠道和国际影响力等方面还具有不小的差距. 本文介绍了NGDC的成立背景, 总结了NGDC在多组学资源体系建设及算法工具等方面的进展、服务成效和影响力, 以及与国际主要生物数据中心之间

引用格式: 陈梅丽, 王彦青, 李茹姣, 等. 国家基因组科学数据中心: 现状及展望. 中国科学: 生命科学, 2024, 54: 1088–1100
Chen M L, Wang Y Q, Li R J, et al. Current status and prospect of the National Genomics Data Center (in Chinese). Sci Sin Vitae, 2024, 54: 1088–1100, doi: 10.1360/SSV-2023-0150

© 2024 《中国科学》杂志社

的对比和思考，并进一步展望了NGDC未来的重点发展布局.

1　国家基因组科学数据中心成立背景

随着生命科学的不断进步，基因组科学数据在国民经济和社会发展中的作用日益突出，提升基因组科学数据的保存、管理和利用水平对保障国家生物数据安全、推动科技和产业创新发展、支撑全民健康、粮食安全等国家重大战略具有重要科学价值和现实意义[1]. 20世纪80年代以来, 以美国、欧洲和日本为代表的发达国家和地区竞相在生物信息资源领域投入大量资金，陆续成立了具有国际权威性的国家级生物信息中心，包括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)和日本DNA数据库(DNA Data Bank of Japan, DDBJ). 2005年, NCBI, EBI和DDBJ达成协议, 建立了国际核酸序列数据库联盟(International Nucleotide Sequence Database Collaboration, INSDC)[3]，每天交换更新数据和信息，涵盖测序原始数据、组装序列和注释信息、以及相关的元信息，每年召开内部会议，讨论并制定了一系列统一的标准和政策，在国际生物信息大数据汇交共享方面具有重要影响力. 作为惯例，主流生命科学期刊发表论文前都会要求将相关数据公开，大部分要求或推荐上传到INSDC成员数据库中. 我国已是生物信息数据产出大国，但由于体系分散，形成一个个“数据孤岛”，导致大量数据都要上传到国外生物信息数据库, 对国家生物数据安全造成严重隐患. 因此, 迫切需要建立国家级生物数据管理平台[1].

2019年6月, 科技部和财政部联合发布《关于国家科技资源共享服务平台优化调整名单的通知》，其中，依托中国科学院北京基因组研究所建设国家基因组科学数据中心，合作共建单位为中国科学院生物物理研究所和上海营养与健康研究所. NGDC的前身是成立于2016年的中国科学院北京基因组研究所生命与健康大数据中心(BIG Data Center, BIGD)[4]. 在NGDC的基础上, 2019年11月, 经中央编制委员会批准, 中央机构编制委员会办公室发文，在中国科学院北京基因组研究所加挂“国家生物信息中心”牌子，整合中国科学院和全国优势力量，建设国家生物信息中心(China National Center for Bioinformation, CNCB)[5]. NGDC立足国内, 面向全球, 积极建设组学生物大数据存储、整合与分析研究体系，致力于成为国际一流的基因组科学数据中心，为支撑我国生命科学创新发展贡献力量[1].

2　国家基因组科学数据中心建设进展

NGDC面向国家大数据和健康中国战略，已初步建成具有自主知识产权、安全可控、涵盖领域广的多维组学数据汇交、存储、管理和共享体系. NGDC建立了基因组数据分析软件工具和平台体系，为数据挖掘与利用提供便利. 同时提供跨库检索服务，可实现对所有开放共享的数据资源的一站式检索. 受科技部委托，2022年7月18日，NGDC开通运行了人类遗传资源信息管理备份平台(Human Genetic Resources Information Platform, HGRIP), 实现了人类遗传资源信息管理、备份、发布与共享的一站式管理.

2.1　组学数据基础库的建设进展

面向国家人口健康和社会可持续发展的重大战略需求, 解决数据孤岛和数据外流问题, 提高组学数据管理、挖掘与再利用的效率, NGDC成立几年来, 已经逐步建设完成了一系列组学数据基础库，支撑多维组学数据的汇交、存储、管理和共享. NGDC的组学数据基础库主要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家族(Genome Sequence Archive Family, GSA Family)[6~8]、基因组数据库(Genome Warehouse, GWH)[9]、基因序列数据库(GenBase)和基因组变异库(Genome Variation Map, GVM)[10,11]等(表1), 其中组学原始数据归档库家族包括组学原始数据归档库(Genome Sequence Archive, GSA)[6,7]、人类遗传资源组学原始数据归档库(Genome Sequence Archive for Human, GSA-Human)[6,8]、多元数据归档库(Open Archive for Miscellaneous Data, OMIX)[6]. 这些数据库在不同组学数据信息的汇交、管理和共享中扮演着相应的角色，同时也承担着国家科学项目数据汇交、管理和发布的重要任务.

为更好地开展多维组学数据的系统化、规范化汇交工作, NGDC开发了多维组学数据汇交服务系统. 该

表 1 NGDC多维组学基础数据库基本信息概览(截至2023年6月30日)

Table 1 An overview of basic information on NGDC's fundamental multi-omics databases (as of June 30, 2023)

NGDC数据库简称	中文名	数据库功能	数据汇交共享情况		NCBI数据镜像整合情况				
			汇交数据量	共享数据量	NCBI对标数据库简称	元信息镜像记录数(万条)	FTP镜像数据量(完成比例)	可浏览和检索	可下载
BioProject	生物项目数据库	生物项目信息的汇交、管理和共享	12662个项目	9022个项目	BioProject	68	NA	是	NA
BioSample	生物样本数据库	生物样本信息的汇交、管理和共享	1129114个样本	646148个样本	BioSample	3346	NA	是	NA
GSA	组学原始数据归档库	原始测序数据的汇交、管理和共享	26.90 PB	13.42 PB	SRA	5087	4.76 PB(26%)	是	是
GWH	基因组数据库	多物种基因组组装序列和注释信息汇交、管理和共享	62532个基因组组装	18418个基因组组装	GenBank	166	125万(75%)	是	是
						36	26万个(72%)	是	是
GenBase	基因序列数据库	病毒、细胞器或基因片段的核酸序列和注释信息的汇交、管理和共享	28392条核苷酸序列	25087条核苷酸序列		26192	2.6亿条(100%)	是	是
			261353条注释蛋白序列	237047条注释蛋白序列		25051	2.5亿条(100%)	是	是
GVM	基因组变异数据库	基因组变异数据的汇交、管理和共享	323441个样本	156761个样本	dbSNP	–	305 GB(100%)	否	是
OMIX	多元数据归档库	基因组统计分析、蛋白、代谢、临床等数据的汇交、管理与共享	11520个文件	5261个文件	–	–	–	–	–

系统的建设遵循INSDC数据汇交标准规范，主要包括以下5大类: (ⅰ) 通用元信息: BioProject收录和共享生物学研究项目信息，提供生物项目结构化描述信息的递交和发布服务，项目类型涵盖常规组学研究的基因组、转录组、表观组和元基因组等，并针对大型项目提供高效、安全、专业化的项目分级管理. BioProject收集的内容包括: 数据递交者、项目基本描述、项目类型、项目资助和相应的项目成果等信息. BioSample收集与共享生物样本信息，提供生物样本结构化描述信息的递交和发布服务，样本类型涵盖人、动物、植物、微生物、病毒和元基因组等. BioSample主要收集数据递交者、样本基本描述以及样本属性等信息. (ⅱ) 测序原始数据: GSA对标NCBI的序列读取存档数据库(Sequence Read Archive, SRA), 接收不同高通量测序平台产生的组学原始下机数据，如Illumina, PacBio SMART, Ion Torrent等, 并支持通用的数据文件格式, 如FASTQ, BAM等; GSA-Human作为GSA的子库, 主要接收人类遗传资源组学原始下机数据. (ⅲ) 基因组和基因序列: GWH和GenBase对标NCBI的GenBank数据库，负责接收基因组、宏基因组、线粒体、叶绿体、细胞器、质粒和噬菌体等序列和注释数据. (ⅳ) 变异数据: GVM作为全基因组变异数据的存储、管理和分析平台，接收包括人、主要动植物和病原微生物等多物种的变异数据信息，是NCBI的变异库dbSNP (Single Nucleotide Polymorphism Database)/dbVar的有效补充. (ⅴ) 其他数据: OMIX作为生命科学领域通用型数据存储库, 主要存储测序分析数据、芯片数据, 涵盖代谢组、脂质组、蛋白组等多元组学数据，以及医学影像、临床信息等数据，它作为上述数据库的重要补充，有效地解决了用户提交多元类型数据的需求. 截至2023年6月，NGDC已经归档了来自28个国家的27.6 PB数据，其中，GSA和GSA-Human共收录来自25个国家，共计6016个用户提交的11611个数据集，共包含976097个实验、1148611个测序反应，数据量达到26.9 PB(图1).

在多维组学数据质控方面，为了提供更好的科学数据服务，NGDC采取了多种数据质量控制方法和流程来确保数据的准确性和完整性. 不同数据库采用的方法略有不同，但大致流程均采用自动校验与人工审核相结合，来审核数据及其格式的一致性和完整性，

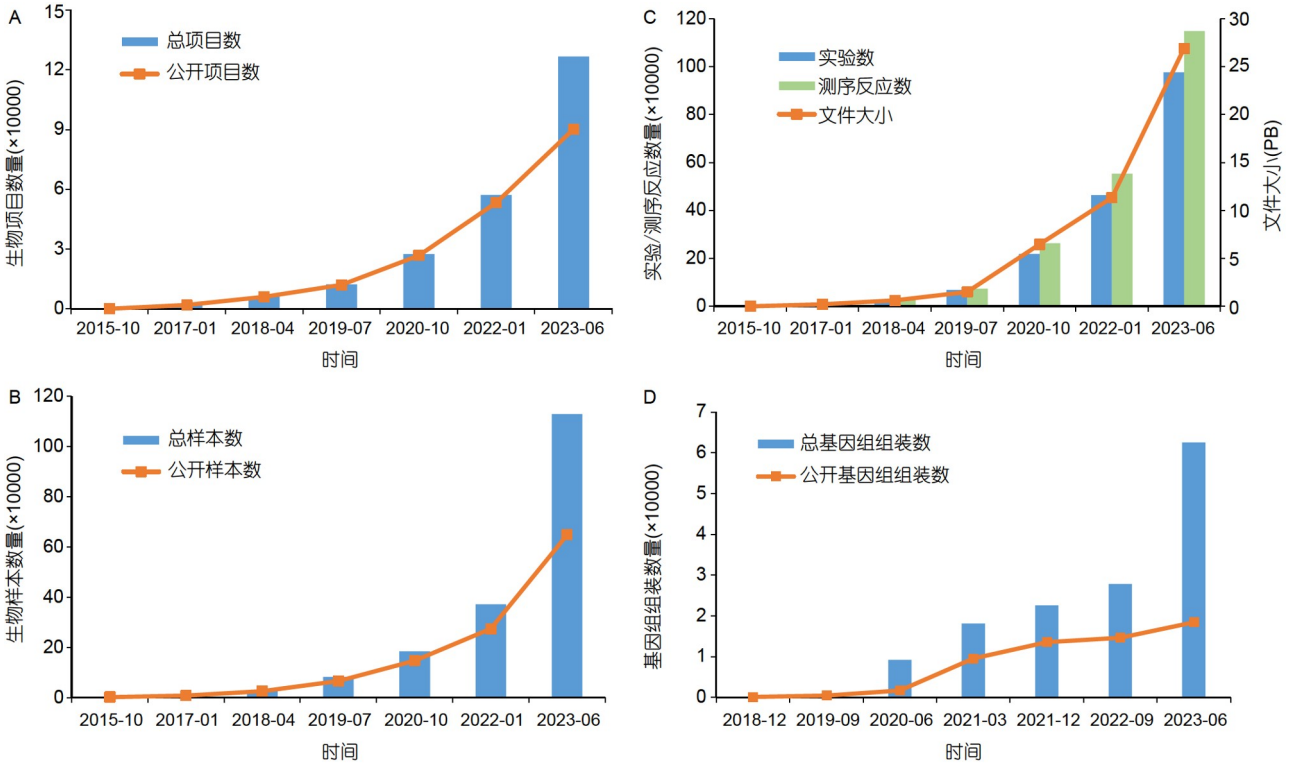

图 1　NGDC多维组学基础数据库汇交和发布的数据量(截至2023年6月30日). A: BioProject汇交和发布的数据量; B: BioSample汇交和发布的数据量; C: GSA Family汇交和发布的数据量; D: GWH汇交和发布的数据量

Figure 1　Volume of data deposited and released in the NGDC fundamental multi-omics databases (as of June 30, 2023). A: Volume of data deposited and released by BioProject; B: volume of data deposited and released by BioSample; C: volume of data deposited and released by GSA Family; D: volume of data deposited and released by GWH

检查数据内容与描述间的一致性，以及核实文件名和MD5码等信息. 例如，GenBase遵循INSDC的相关标准，采用基于规则的自动验证和人工审编为一体的质控系统，分别从序列、物种、元信息、注释信息、审核等角度校验，严格把控数据质量，保障基因序列数据的准确性、完整性和可用性. 经过上述系列数据质量控制流程，增强了数据的可信度和实用性，提升了数据的有效性和应用价值，为更广泛的科学研究提供了可靠的支撑.

在多维组学数据共享方面，数据基础库采用标准化的数据格式和命名规则组织数据开放共享, 以“开放为常态，不开放为例外”为指导原则，提供高效、规范化的科学数据发布与共享服务，并确保用户隐私和个人信息安全. 建立快速便捷的数据访问和使用方式, 提供实时更新的数据信息，以保证数据的时效性和质量. 截至目前, 组学基础数据库发布共享数据13.75 PB (图1). 同时，NGDC与华为云开展合作，目前已建立GSA与华为云的数据传输通道，实现GSA公开数据到华为云的同步. 同时，在GSA的数据访问页面中增加华为云的下载链接，实现用户从华为云下载数据的功能. 截至2023年6月30日，已完成5360个数据集, 约1.5 PB数据的同步.

在数据的国际化同步交换方面，得到INSDC的支持和NCBI的技术协助，以新冠病毒基因组序列为起点，NGDC与NCBI初步建立了数据交换共享机制[9]，并且已逐步推广到其他物种的基因序列. 目前NGDC的BioProject, BioSample, GSA, GWH, GenBase与NCBI对接标准和流程，持续推进和探索测序原始数据、基因组、基因序列等全面的数据交换共享. 此外, NGDC也镜像了NCBI多个核心基础库的数据资源(表1), 包括: (ⅰ) BioProject和BioSample镜像了全部元信息记录，实现了一体化整合，并进行每日更新; (ⅱ) GSA每日更新镜像SRA[12]的全部元信息(5087万条记录), 通过一体化整合实现了数据可浏览和检索，并同

步了超4.6 PB测序原始数据供下载; (iii) GWH库镜像了GenBank的基因组数据库(Assembly[13], 166万条记录)和参考序列数据库(Reference Sequence, RefSeq[14], 36万条记录)的全部元信息, 并持续更新和同步了超150万个基因组组装数据, 与自己收录的数据实现了一体化整合, 并且支持数据的检索、展示和下载; (iv) GenBase库镜像了传统GenBank库[15]最新发布的2.6亿条核苷酸和2.5亿条蛋白序列及其元信息数据, 与自己收录的数据实现了一体化整合, 并且支持数据的检索、展示和下载; (v) GVM库镜像了dbSNP[16]的全部数据供下载, 共计305 GB. 当前已基本实现了上述数据在NGDC网站的一站式检索、访问和下载, 通过与NGDC自主开发数据库进行一体化整合, 并持续更新镜像序列等文件, 极大提升了国内科研人员查询和获取数据的效率.

2.2 人类遗传资源数据管理的建设进展

人类遗传资源是国家重要的战略性资源, 是人类“生命说明书”, 对于生命科学研究至关重要. 研究人类遗传资源信息有助于探索生命奥秘, 理解生老病死的发生和发展规律. NGDC在人类遗传资源数据的管理上, 严格遵循《中华人民共和国人类遗传资源管理条例》《科学数据管理办法》等国家法律法规, 建立了符合国家法律法规的数据管理制度和安全策略, 制定一系列人遗数据分级分类和数据共享管理规范, 用于指导NGDC各数据资源库数据共享方面技术规范的制定和系统功能开发, 主要包括: 《国家基因组科学数据中心数据分级使用管理规定》《国家基因组科学数据中心人类遗传资源数据管理办法》和《国家基因组科学数据中心人类遗传资源数据申请审核办法》.

NGDC逐步形成了人类遗传资源数据汇交与共享管理的核心规范和要求. 在数据汇交管理方面, 核心内容包括: (i) 伦理合规性要求: 在提交数据之前, 数据提供者必须获得相关研究对象的知情同意, 以确保数据的获取和使用符合伦理准则和法律规定; (ii) 隐私保护性原则要求: 在提交数据时, 数据提供者必须对研究对象的个人身份信息进行脱敏处理, 以维护其隐私权; (iii) 政策合法性规定: 数据提供者在发布其数据集之前, 必须在科学技术部人类遗传资源服务系统完成事先报告[8]. 在数据使用管理方面, 核心内容包括: (i) 使用原则和限制: 数据使用者在获得数据后, 应保障数据集的合规和隐私保护; (ii) 使用者责任与义务: 在数据管理和处理中出现问题时, 其职责将延伸至其所属组织或机构的责任人; (iii) 安全责任与应急措施: 数据使用者必须保障数据安全、及时报告和处理安全事件; (iv) 知识产权及引用要求: 数据使用者拥有使用数据所获得的研究结果的知识产权, 但在成果发表时应正确引用被使用数据集.

NGDC推进人遗数据分类分级的有效实施. 在人类遗传资源分级管理方面, 遵循国家互联网信息办公室印发的《网络数据安全管理条例(征求意见稿)》, 将数据分为一般数据、重要数据、核心数据三级. 在人类遗传资源分类管理方面, 中心构建了一个人类遗传资源数据综合的管理体系, 包括数据收集、层次化存储、安全管理和受控共享等多个功能[8]. 基于数据分析层次分类, 将数据分为原始输出数据、组学个体级分析数据、群体特征分析数据以及其他类数据, 分别提供了组学研究的起点、个体层面、群体层面及其补充信息. 数据管理体系覆盖原始数据、基因组、变异组、转录组和表观组等多维组学数据, 包括GSA-Human, OMIX, GWH, GenBase和GVM等, 为用户提供人类遗传资源数据汇交、存储、发布和共享服务. 基于数据共享方式分类, 将数据共享方式分为两种: 公开访问和申请审核. 公开访问意味着数据释放后, 所有的用户都能直接查看和下载该数据; 申请审核则是在数据释放后, 用户需要向数据所有者提出申请, 获得批准后才能访问和下载数据.

2.3 特色资源库和知识库的建设进展

随着多维组学数据的爆炸性增长, 科学研究向数据密集型研究范式转型, 带动了信息挖掘和知识转化的快速发展. NGDC通过整合分析和生物审编相结合的方式促进数据增值, 基于新数据资源的不断挖掘和海量文献的知识审编, 构建专业化的特色资源库和多维组学知识库. 更好地组织和利用科学研究产出的多维组学相关知识, 有助于理解生命的本质, 助力人口健康、公共卫生安全、植物育种、动物改良和生物多样性保护等相关研究, 为特定领域的国家决策提供科学数据支撑.

面向国家战略需求, NGDC采用智能审编和人工审编相结合的模式, 整合并挖掘生物信息知识, 建立了涵盖不同领域的特色资源库和知识库, 服务科学研

究和产业应用. 包括: (ⅰ) 人口健康: 聚焦疾病、生理过程等相关研究, 如人类癌症可变剪接知识库(ASCancer Atlas)[17]、脑疾病知识库(BrainBase)[18]、衰老生物学多组学数据库(Aging Atlas)[19]; (ⅱ) 公共卫生安全: 服务于疫情防控并支撑相关研究, 如新冠病毒信息库(RCoV19)[20]和猴痘病毒信息库(MPoXVR)[21]; (ⅲ) 植物育种: 围绕经济作物和林业植物中基因功能和性状调控等, 如大豆多维组学数据库(SoyOmics)[22]、高粱基因组数据库(SorGSD)[23]、植物叶片衰老数据库(LSD)[24]等; (ⅳ) 动物改良: 助力于全球范围内动物疾病和品种改良等研究, 如犬科动物数据库(iDog)[25]、绵羊组学数据库(iSheep)[26]等; (ⅴ) 生物多样性保护: 涵盖动物、植物、微生物等不同物种类型, 支撑生物多样性保护, 如同源基因数据库(HGD)[27]、叶绿体基因组综合信息库(CGIR)[28]、原核生物泛基因组数据库(ProPan)[29]等.

基于同一组学数据类型统一标准化处理的指导原则, 围绕多维组学研究不同领域的需求, NGDC建立了一系列特色专题的多维组学信息库和知识库, 在全球范围内为推动生命和健康科学的发展提供了开放的资源家族, 包括: (ⅰ) 基因组: 如分子序列组分动态图谱数据库(CompoDynamics)[30]、原核生物防御系统基因数据库(PADS Arsenal)[31]等; (ⅱ) 转录组: 如基因表达数据库(GEN)[32]、全转录组关联研究知识库(TWAS Atlas)[33]、人类长非编码RNA表达数据库(LncExpDB)[34]等; (ⅲ) 变异组: 如动植物基因组变异与表型关联知识库(GWAS Atlas)[35]等; (ⅳ) 表观组: 如表观组关联研究开放平台(EWAS Open Platform)[36]、甲基化数据库MethBank[37]、核小体定位数据库(NucMap)[38]和lncRNA人类长非编码RNA数据库(LncBook)[39]等; (ⅴ) 代谢组: 如人类癌症代谢物关联知识库(MACdb)[40]; (ⅵ) 多组学: 如热带作物多组学数据平台(TCOD)、再生生物学多组学数据库(Regeneration Roadmap)[41]等. 近年来, 单细胞研究呈现出迅猛增长的态势, 为支撑和服务单细胞领域研究, NGDC建立了癌症单细胞表达图谱数据库(CancerSCEM)[42]、单细胞全基因组甲基化数据库(scMethBank)[43]、细胞分类库(Cell Taxonomy)[44]和细胞药物反应知识库(CeDR Atlas)[45]等.

为解决目前生命科学研究领域对国外数据文献库严重依赖的问题, 通过整合PubMed, bioRxiv和medRxiv公共文献库, 初步建成了生命科学文献库(OpenLB). 为更好地了解数据库发展态势及其对生命健康科学的影响提供全球视角, 进一步建立了生物医学领域已发表数据库的信息整合系统Database Commons[46]. 总之, NGDC持续贯彻落实生物大数据国家科技发展战略和重大任务需求, 紧跟生命科学领域前沿, 持续推进生物信息挖掘和知识转化进程, 为国家基因组科学数据的挖掘利用提供重要支撑.

2.4　软件工具的建设进展

在服务于基因组科学数据存储和管理的同时, 为了促进生物数据的分析挖掘和有效利用、将数据转化为知识, NGDC研制和整合了包括多组学整合与分析工具, 进化、肿瘤和病毒等专题的软件工具, 并分别根据基础库和特色库服务用户的不同需求, 在组学数据存管用的实际服务场景下、自主开发和部署了10套数据分析和可视化算法工具, 用数据和工具紧密结合的方式给用户提供更有价值的信息.

NGDC研制和整合的工具包括: (ⅰ) 基础工具: 为满足用户最基本和最通用的生物信息分析需求, NGDC建立了生物信息基础工具平台BIT. 目前平台整合了23个数据可视化和序列比对工具, 提供在线计算服务. 其中, NGDC的BLAST在线工具整合NCBI BLAST+软件包, 提供在线的核酸和蛋白质序列局部两两比对服务. 目前用户可以选择的比对目标数据库共有353个, 包括整合的几乎所有的NCBI BLAST[47~50]目标数据库(如Nt, Nr, RefSeq, Swiss-Prot[51], Betacoronavirus等)和NGDC自主构建的目标数据库(包括GWH, GEN, LncBook, SoyOmics, TCOD等), 其中从NCBI整合的数据库和部分NGDC构建的数据库保持每日更新. (ⅱ) 专题工具: 包括服务单倍型参考面板重建的基于卷积神经网络的基因组插值方法Refrgim[52], 服务基因表达定量、单细胞转录组降维、聚类、轨迹推断、细胞类型注释等转录组分析的在线工具集[32], 服务表观组关联分析的工具集EWAS Toolkit[36], 支持编码和非编码序列选择压力计算的工具包KaKs_Calculator 3.0[53], 基于随机森林的体细胞变异鉴定方法TSomVar[54], 多组学水平上对个体癌症基因组的一站式综合生物标志物鉴定平台CCAS[55], 冠状病毒在线分析平台[56], 冠状病毒基因组浏览器(Coronavirus GenBrowser)[57], 服务原始读段分类的

GSA的原始测序数据物种分类工具，新型冠状病毒资源库单体型网络可视化模块网络高效构建工具McAN[58].

2.5 服务成效和中心影响力

NGDC的多维组学数据汇交服务系统为国内外用户提供数据汇交、存储、管理与共享服务，数据编号被Springer Nature, Elsevier, Wiley, Taylor & Francis等全球主要出版集团推荐或认可. 系统现已服务于国家重点研发计划、国家自然科学基金、中国科学院战略性先导科技专项等科研项目的数据汇交与管理，已成为国内科研用户递交与共享基因组数据的首选，数据递交量呈显著增长趋势. 截至2023年6月，共为国内外1700多家单位7400余用户提供数据汇交服务，支持13000多个科技项目数据汇交管理，同时，支撑了2700多篇研究文章在近600种学术期刊上发表. 除国内用户外，还服务了包括巴基斯坦、巴西、英国、美国、新加坡等在内的22个境外国家.

NGDC的多维组学数据资源体系支撑了国内多个大型基因组计划、古基因组研究等领域的数据汇交和共享工作，如中国玉米界首篇*Nature*论文数据的共享[59]，以及中国境内最大规模古基因组研究数据的汇交[60]. 该体系也为新型冠状病毒分子溯源和传播演化、动植物分子育种与遗传改良、疾病与人口健康等研究提供了广泛的数据支持. 此外, NGDC的数据资源还被多个领域的科学家下载和复用，完成多项多组学数据深度挖掘整合和知识库构建工作[1]. 例如, 在新冠病毒疫情暴发期间, 新冠病毒信息库RCoV19有力地支撑了北京新发地疫情分子溯源[61]，世界卫生组织开展的SARS-CoV-2全球溯源研究——中国部分[62]，以及巴基斯坦境内早期新冠病毒传播演化规律研究[63]等工作; 来自荷兰的研究团队使用GSA-Human中的数据揭示了唾液酸化聚糖在免疫抑制中的关键作用[64]; 中国科学院遗传与发育生物学研究所的团队整合了GSA中的水稻全基因组测序数据，开发了水稻分子育种整合组学知识库MBKbase[65].

NGDC的多维组学数据资源体系还为用户提供了一站式在线检索、数据浏览和下载服务. 自2019年至今，用户访问总量达351万人次，独立IP地址320万个, 服务范围覆盖全球190多个国家和地区，FTP数据下载超过6000万次, 总下载量超过7 PB.

依托NGDC的多维组学数据资源体系，NGDC在科技部的授权下，于2022年7月18日成功完成了人类遗传资源信息管理备份平台系统的迁移、升级和上线运行. 新平台有效实现了数据汇交、管理、发布和共享的一体化，大大减少了对人类遗传资源信息备份与管理的工作量，确保了备份数据与发布共享数据的一致性. 自系统迁移至2023年6月底，平台新增注册用户695名，新生成备份号2015个. 产生有效文件记录7554条, 去冗余后备份数据总量926.12 TB.

3 与国际主要生物数据中心的比较

3.1 优势

我国在生物数据管理系统及其技术体系研究方面起步较晚, 近年来NGDC逐渐发展起来, 逐步在生物信息领域占有一席之地，尤其是在特色资源库建设、数据整合等方面形成了一些优势. 与国际一流生物信息中心(如美国NCBI和欧洲EBI)相比，NGDC的数据资源建设围绕汇交存储、整合挖掘、转化应用三个维度进行, 除了传统的基因、基因组数据库, 在特色资源库和整合类知识库方面更具优势，如新冠病毒信息库(RCoV19)在数据的来源、数据类型、数据量等方面, 都处于领先地位. 此外，在数据的一体化整合获取方面, NGDC不仅整合了NCBI和EBI等国际数据资源, 也整合了国内多个合作单位的44个数据资源库，所有的数据资源都可以通过BIG Search进行一站式检索，从而实现高性能、灵活的全文索引和检索，帮助用户更有效、快速、准确地获取知识.

3.2 不足

与国际知名数据中心对比，我国的人员投入少、经费紧缺、配套设施薄弱，国际影响力有待进一步提升，应尝试扩大数据中心团队规模，加强专业人才培养，持续增加科研经费投入和政府管理力度，全方位支持国家级生物数据中心的基础建设. 这些不足主要体现在以下方面:

(1) 人员及团队规模差距明显. 我国在生物数据中心方面的建设起步晚，专业人才队伍建设有待扩大和提升，才能有望依靠团队攻关解决生物医学大数据的“卡脖子”关键技术问题. 美国NCBI和欧洲EBI均有大约700人的专业团队，而我国NGDC目前专职人员约

100人(含学生), 并且人员的差距不仅仅体现在数量的差别, 更重要的是人才引进、人才发展以及人才评价的政策和机制.

(2) 经费体量和获取渠道差距悬殊. 生命组学数据的爆发式增长与数据存储等资源紧张的矛盾问题日益突出. 公开数据显示, 2017年, 美国NCBI的年运行经费高达1.54亿美元, 且全部来自美国国会拨付. 2022年, EBI年运行经费1.09亿美元, 其中41%来自欧盟各国特别是英国政府拨款, 17%来自资本投资. 而当前NGDC年运行经费约0.04亿美元, 除~0.01亿美元为财政部经费直接支持外, 所需运转费用大部分由科研人员从国家不同渠道的科技资助项目中竞争获得, 这使得科研人员不得不花费大量的时间和精力争取和完成各类项目来维持中心的日常运转, 一方面其公益属性不可持续, 另一方面迫使科研团队成员难以完全聚焦主责主业, 专注于中心的专业化建设和发展(表2).

(3) 国际合作力度和国际影响力有待进一步加强. 虽然NGDC连续六年被本领域国际权威期刊《核酸研究》称为与美国NCBI、欧洲EBI齐名的国际主要生物数据中心, 但是, 在国际合作网络方面, INSDC联盟的成员单位, 即美国的NCBI、欧洲的EBI和日本的DDBJ, 仍然处于国际生物信息的垄断地位, 在数据标准制定、数据交换和共享方面具有重要影响力. NGDC目前正在积极与INSDC联盟进行对接, 努力加入该联盟, 以期改变当前现状. 此外, NGDC还需专注于建设一批有国际影响力的特色资源和软件工具, 可与RefSeq和BLAST等相媲美.

4　国家基因组科学数据中心成就及展望

4.1　成就

2018~2023年, NGDC连续6年被本领域国际权威期刊《核酸研究》称为与美国NCBI、欧洲EBI齐名的国际主要生物数据中心[66~69]. NGDC整体建设成果入选国家“十三五”建设成就展、“奋进新时代”主题成就展, 多项研究成果入选“中国生物信息学年度十大进展”. NGDC开发的新冠病毒信息库RCoV19被列入《抗击新冠肺炎疫情的中国行动》白皮书, 入选2021年度中科院科技创新亮点成果、2022年世界互联网领先科技成果, 作为亮点工作入选《“一带一路”创新发展报告(2021)》, NGDC团队荣获“全国科技系统抗击新冠肺炎疫情先进集体”、中国科学院第三届“科苑名匠”等荣誉称号.

4.2　展望

NGDC建设虽已取得诸多实质性进展, 被誉为全球主要数据中心之一, 但仍需对标国际主要生物数据中心, 争取加大稳定经费支持、加大人员队伍建设、加强国际交流与合作, 在充分发挥自身优势的同时, 努力缩小差距. 另外, 结合大数据发展的新形势和新要求, 在标准化、智能化、自动化等方面继续提升. 展望未来, NGDC将重点着眼以下7个方面.

(1) 加大争取稳定经费支持力度. NGDC建设对我国具有举足轻重的作用和意义, 它不仅是生命科学研究的需要, 更是国家生物数据战略储备和国家安全的需要. NGDC建设不仅是硬件设施建设, 更需要大量从事生物信息学、计算机科学、数学等交叉学科背景的人才团队持续攻关. 在硬件设施能够投入使用之前, 对硬件的需求和核心算法软件的攻关都亟需大量的经费投入. 目前, NGDC的经费主要来自于国家纵向项目, 这难以支撑NGDC的公益性发展和需求, 国际大型数据中心多是采用政府稳定支持的方式运行, 因此, 未来NGDC的发展必须加大争取对NGDC给予稳定的运行经费支持的力度.

表 2　NGDC与国际主要生物数据中心运行情况对比(2022年数据)[a)]

Table 2　Comparison of NGDC with major international biological data centers (data in 2022)[a)]

机构名称(简称)	成立年份	人员规模	数据量(PB)	数据库个数	年度经费(亿美元)
美国国家生物技术信息中心(NCBI)	1988	~700	>60	67	~1.54* (美国国会拨付)
欧洲生物信息学研究所(EBI)	1992	672	~45	69	~1.09
中国国家基因组科学数据中心(NGDC)	2019	117 (含学生)	~20	54	0.04 (几十个项目资助)

a) 表格中欧元和人民币对美元使用2022年1月1日汇率. *表示该数据来自于2017年公开数据

(2) 加强人才培养和人才队伍建设. 目前, NGDC的人员队伍在体量和质量上都与国际大的数据中心之间存在差距, 所以, 下一步, NGDC在获得稳定经费支持的前提下, 制定完善的人才发展以及人才评价的政策和机制, 一方面要加强基础人才培养, 优化人员结构, 另一方面加强人才引进. 在为NGDC进行人才培养和人才队伍建设的同时, 为未来我国生命科学领域的可持续发展提供人才储备.

(3) 扩大国内和国际合作. 当前NGDC已初步建立国内外合作网络, 在生命科学领域高质量共建“一带一路”, 促进人类命运共同体的构建, 未来将继续积极与国内外相关研究机构拓展合作关系, 通过务实合作形成合力共同构建生命科学发展新格局. 国内合作方面, 将重点建设国内生物信息数据的互通互联机制, 促进良好的生物信息数据生态建设, 主要包括: (ⅰ) 建立国内数据物理整合机制, 开展跨领域合作实现联合攻关、互利互赢; (ⅱ) 通过BIG Search等工具和平台实现国内数据逻辑整合, 实现生物信息大数据的一体化检索和快速获取; (iii) 建立打通不同类型数据间壁垒的机制和方案, 促进组学数据与农业口、临床医学等领域的数据融合应用. 国际合作方面, NGDC在确保数据安全的基础上, 将深入开展数据共享服务和国际合作交流, 主要包括: (ⅰ) 与INSDC开展合作交流, 探索全面的数据镜像和数据交换, 争取尽早成为其合作伙伴; (ⅱ) 在“一带一路”国际科学组织联盟(Alliance of International Science Organizations for the Belt and Road Regions, ANSO)和国际生物科学联合会(International Union of Biological Sciences, IUBS)框架下推进大数据全球合作, 巩固和发展国际生物多样性与健康大数据联盟(Global Biodiversity and Health Big Data Alliance, BHBD), 推进建设国际合作网络; (iii) 与外方联合开展科研合作, 形成共同的学术成果, 增进互利共赢. NGDC将持续创新服务模式、实现合作双赢, 支撑我国生命科技创新发展, 形成显著的国内外影响力.

(4) 建立智能化、自动化的数据审编系统. 数据审编是生物信息大数据管理的核心内容. 未来, NGDC将根据各种组学数据的特点, 研发个性化的智能审编应用系统, 涵盖生物项目、生物样本、原始序列、基因组组装、基因序列、变异信息等, 实现从数据递交、接收、质控、审编、发布、保存、更新等过程的流程自动化和智能化, 完成用户汇交数据的一站式审编, 以期达到序列号快速分配和日处理PB级数据的能力, 提升国家生物信息大数据管理效率, 完善国家生物信息大数据管理体系.

(5) 开发先进和精准的搜索引擎. 面向NGDC的多组学数据资源体系, 建立多数据源融合检索引擎, 实现海量、多维、异构数据的跨库关联检索. 抽取和整合多源数据信息, 形成多源数据信息分层级、分布式存储管理系统, 实现数据库及知识库的多源融合; 利用关系型及非关系型数据存储模式、分片索引算法、分布式索引存储等, 实现海量数据毫秒级查询响应; 利用跨模态、多粒度语义关联搜索方法, 实现多源数据库的跨库检索; 利用数据接口、高速数据传输等技术, 实现检索结果的高效获取.

(6) 开发数据处理工具和分析流程云平台. 在自主工具研发方面, NGDC将紧跟基因组学技术的发展前沿, 针对新兴技术产生的基因组学数据在分析中遇到的突出问题和数据处理瓶颈, 以及数据库、知识库构建和使用过程中产生的实际问题, 开发新工具和算法, 推动基因组学研究进展, 更好地服务用户. 在工具整合方面, NGDC将继续整合生物信息学基础工具, 搭建组学数据处理常用流程, 提高整合工具和流程的数量和质量, 并把它们部署到虚拟化的云计算和云存储平台上, 形成服务研究者的“生物云”在线计算平台. 同时, NGDC将聚焦基因组学研究的前沿领域, 面向热点应用, 整合最新的算法技术, 建立标准流程, 形成一站式主题在线服务(Web Server), 帮助研究者系统地分析和挖掘研究过程中产生的基因组学数据, 从数据中获得知识.

(7) 利用和发展AI技术. NGDC目前已在新冠病毒高风险株系的预警、基于基因组序列的病毒分类、多模态医学影像检索、基于序列的疾病预测、动物的图像检索、基于图像识别的植物性状提取、脑疾病的预测和分型、体细胞变异鉴定、基因型填补等工作中使用了AI技术, 取得了良好的效果. 未来, 在数据存管用、数据库建设、数据分析挖掘工作的更多方面, NGDC将拓展AI技术的应用范围. 例如, 运用大语言模型进行OpenLB文献库的知识提取, 帮助用户从海量文献中快速提取最相关的知识, 提高文献审编的效率; 为应对RNA三维结构预测这一挑战性的问题, 利用中心积累的海量RNA相关数据的优势, 运用图卷积神经网

络等前沿AI技术，构建深度学习模型对RNA序列和RNA结构之间的关系建模，准确快速地预测RNA三维结构；在使用AI技术对病毒高风险变体预警方面，未来将把相关研究推广到其他的病毒和病原体上，在高风险变体出现的早期及时预警，为疫情防控提供科学数据支撑.

致谢　感谢NGDC全体成员的敬业工作; 感谢北京大学的罗静初教授、中国科学院生物物理研究所的陈润生院士、中国科学院分子植物科学卓越创新中心的赵国屏院士的指导; 感谢科学技术部、国家卫生健康委员会、中国科学院以及NGDC的两个共建单位——中国科学院生物物理研究所和中国科学院上海营养与健康研究所的支持. 感谢NGDC的陈婷婷、赵学彤、田东梅、杨飞、徐添翼在材料整理方面给予的协助与支持.

参考文献

1 Li R J, Zhang X, Song S H, et al. Safety management and application of genomics data (in Chinese). Big Data Research, 2022, 8: 37–45 [李茹姣, 张欣, 宋述慧, 等. 基因组科学数据的安全管理与应用. 大数据, 2022, 8: 37–45]
2 Luo J C. Life sciences and bioinformation technology in the Internet era (in Chinese). 2016, Available from URL: http://abc.gao-lab.org/jcl/Luo-JC_2003_100Puzzles.pdf [罗静初. 网络时代的生命科学和生物信息技术. 2016, URL: http://abc.gao-lab.org/jcl/Luo-JC_2003_100Puzzles.pdf]
3 Cochrane G, Karsch-Mizrachi I, Takagi T, et al. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res, 2016, 44: D48–D50
4 Luo J. GSA and BIGD: filling the gap of bioinformatics resource and service in China. Genomics Proteomics Bioinformatics, 2017, 15: 11–13
5 Luo J C. EMBOSS and EMBnet (in Chinese). Chin J Bioinform, 2021, 19: 223–231 [罗静初. EMBOSS和EMBnet. 生物信息学, 2021, 19: 223–231]
6 Chen T, Chen X, Zhang S, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics, 2021, 19: 578–583
7 Wang Y, Song F, Zhu J, et al. GSA: genome sequence archive. Genomics Proteomics Bioinformatics, 2017, 15: 14–18
8 Zhang S S, Chen X, Chen T T, et al. GSA-Human: genome sequence archive for human (in Chinese). Hereditas, 2021, 43: 988–993 [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. 遗传, 2021, 43: 988–993]
9 Chen M, Ma Y, Wu S, et al. Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics, 2021, 19: 584–589
10 Li C, Tian D, Tang B, et al. Genome Variation Map: a worldwide collection of genome variations across multiple species. Nucleic Acids Res, 2021, 49: D1186–D1191
11 Song S, Tian D, Li C, et al. Genome Variation Map: a data repository of genome variations in BIG Data Center. Nucleic Acids Res, 2018, 46: D944–D949
12 Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res, 2012, 40: D54–D56
13 Kitts P A, Church D M, Thibaud-Nissen F, et al. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res, 2016, 44: D73–D80
14 Haft D H, DiCuccio M, Badretdin A, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res, 2018, 46: D851–D860
15 Sayers E W, Cavanaugh M, Clark K, et al. GenBank 2023 update. Nucleic Acids Res, 2023, 51: D141–D144
16 Sherry S T, Ward M H, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res, 2001, 29: 308–311
17 Wu S, Huang Y, Zhang M, et al. ASCancer Atlas: a comprehensive knowledgebase of alternative splicing in human cancers. Nucleic Acids Res, 2023, 51: D1196–D1204
18 Liu L, Zhang Y, Niu G, et al. BrainBase: a curated knowledgebase for brain diseases. Nucleic Acids Res, 2022, 50: D1131–D1138
19 Aging Atlas Consortium . Aging Atlas: a multi-omics database for aging biology. Nucleic Acids Res, 2021, 49: D825–D830
20 Song S, Ma L, Zou D, et al. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics Proteomics

Bioinformatics, 2020, 18: 749–759
21 Ma Y, Chen M, Bao Y, et al. MPoxVR: a comprehensive genomic resource for monkeypox virus variant surveillance. Innovation, 2022, 3: 100296
22 Liu Y, Zhang Y, Liu X, et al. SoyOmics: a deeply integrated database on soybean multi-omics. Mol Plant, 2023, 16: 794–797
23 Liu Y, Wang Z, Wu X, et al. SorGSD: updating and expanding the sorghum genome science database with new contents and tools. Biotechnol Biofuels, 2021, 14: 165
24 Li Z, Zhang Y, Zou D, et al. LSD 3.0: a comprehensive resource for the leaf senescence research community. Nucleic Acids Res, 2020, 48: D1069–D1075
25 Tang B, Zhou Q, Dong L, et al. iDog: an integrated resource for domestic dogs and wild canids. Nucleic Acids Res, 2019, 47: D793–D800
26 Wang Z H, Zhu Q H, Li X, et al. iSheep: an integrated resource for sheep genome, variant and phenotype. Front Genet, 2021, 12: 714852
27 Duan G, Wu G, Chen X, et al. HGD: an integrated homologous gene database across multiple species. Nucleic Acids Res, 2023, 51: D994–D1002
28 Hua Z, Tian D, Jiang C, et al. Towards comprehensive integration and curation of chloroplast genomes. Plant Biotechnol J, 2022, 20: 2239–2241
29 Zhang Y, Zhang H, Zhang Z, et al. ProPan: a comprehensive database for profiling prokaryotic pan-genome dynamics. Nucleic Acids Res, 2023, 51: D767–D776
30 Jiang S, Du Q, Feng C, et al. CompoDynamics: a comprehensive database for characterizing sequence composition dynamics. Nucleic Acids Res, 2022, 50: D962–D969
31 Zhang Y, Zhang Z, Zhang H, et al. PADS Arsenal: a database of prokaryotic defense systems related genes. Nucleic Acids Res, 2020, 48: D590–D598
32 Zhang Y, Zou D, Zhu T, et al. Gene Expression Nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels. Nucleic Acids Res, 2022, 50: D1016–D1024
33 Lu M, Zhang Y, Yang F, et al. TWAS Atlas: a curated knowledgebase of transcriptome-wide association studies. Nucleic Acids Res, 2023, 51: D1179–D1187
34 Li Z, Liu L, Jiang S, et al. LncExpDB: an expression database of human long non-coding RNAs. Nucleic Acids Res, 2021, 49: D962–D968
35 Liu X, Tian D, Li C, et al. GWAS Atlas: an updated knowledgebase integrating more curated associations in plants and animals. Nucleic Acids Res, 2023, 51: D969–D976
36 Xiong Z, Yang F, Li M, et al. EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. Nucleic Acids Res, 2022, 50: D1004–D1009
37 Zhang M, Zong W, Zou D, et al. MethBank 4.0: an updated database of DNA methylation across a variety of species. Nucleic Acids Res, 2023, 51: D208–D216
38 Zhao Y, Wang J, Liang F, et al. NucMap: a database of genome-wide nucleosome positioning map across species. Nucleic Acids Res, 2019, 47: D163–D169
39 Li Z, Liu L, Feng C, et al. LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations. Nucleic Acids Res, 2023, 51: D186–D191
40 Sun Y, Zheng X, Wang G, et al. MACdb: a curated knowledgebase for metabolic associations across human cancers. Mol Cancer Res, 2023, 21: 691–697
41 Kang W, Jin T, Zhang T, et al. Regeneration Roadmap: database resources for regenerative biology. Nucleic Acids Res, 2022, 50: D1085–D1090
42 Zeng J, Zhang Y, Shang Y, et al. CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic Acids Res, 2022, 50: D1147–D1155
43 Zong W, Kang H, Xiong Z, et al. scMethBank: a database for single-cell whole genome DNA methylation maps. Nucleic Acids Res, 2022, 50: D380–D386
44 Jiang S, Qian Q, Zhu T, et al. Cell Taxonomy: a curated repository of cell types with multifaceted characterization. Nucleic Acids Res, 2023, 51: D853–D860
45 Wang Y Y, Kang H, Xu T, et al. CeDR Atlas: a knowledgebase of cellular drug response. Nucleic Acids Res, 2022, 50: D1164–D1171
46 Ma L, Zou D, Liu L, et al. Database Commons: a catalog of worldwide biological databases. Genomics Proteomics Bioinformatics, 2023, 21: 1054–1058
47 Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215: 403–410

48 Altschul S, Madden T L, Schaffer A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, 25: 3389–3402

49 Boratyn G M, Thierry-Mieg J, Thierry-Mieg D, et al. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics, 2019, 20: 405

50 Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics, 2009, 10: 421

51 Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol, 2001, 3: 47–55

52 Shi S, Qian Q, Yu S, et al. RefRGim: an intelligent reference panel reconstruction method for genotype imputation with convolutional neural networks. Brief Bioinform, 2021, 22: bbab326

53 Zhang Z. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. Genomics Proteomics Bioinformatics, 2022, 20: 536–540

54 Shi S, Wang Q, Shang Y, et al. TSomVar: a tumor-only somatic and germline variant identification method with random forest. Brief Bioinform, 2022, 23: bbac381

55 Zheng X, Zong W, Li Z, et al. CCAS: One-stop and comprehensive annotation system for individual cancer genome at multi-omics level. Front Genet, 2022, 13: 956781

56 Gong Z, Zhu J W, Li C P, et al. An online coronavirus analysis platform from the National Genomics Data Center. Zool Res, 2020, 41: 705–708

57 Yu D, Yang X, Tang B, et al. Coronavirus GenBrowser for monitoring the transmission and evolution of SARS-CoV-2. Brief Bioinform, 2022, 23: bbab583

58 Li L, Xu B, Tian D, et al. McAN: a novel computational algorithm and platform for constructing and visualizing haplotype networks. Brief Bioinform, 2023, 24: bbad174

59 Huang Y, Wang H, Zhu Y, et al. THP9 enhances seed protein content and nitrogen-use efficiency in maize. Nature, 2022, 612: 292–300

60 Kumar V, Wang W, Zhang J, et al. Bronze and Iron Age population movements underlie Xinjiang population history. Science, 2022, 376: 62–69

61 Pang X, Ren L, Wu S, et al. Cold-chain food contamination as the possible origin of COVID-19 resurgence in Beijing. Natl Sci Rev, 2020, 7: 1861–1864

62 Joint WHO-China Study Team. WHO-convened global study of origins of SARS-CoV-2: China part (text extract). Infect Dis Immun, 2021, 1: 125–132

63 Song S, Li C, Kang L, et al. Genomic epidemiology of SARS-CoV-2 in Pakistan. Genomics Proteomics Bioinformatics, 2021, 19: 727–740

64 Rodriguez E, Boelaars K, Brown K, et al. Sialic acids in pancreatic cancer cells drive tumour-associated macrophage differentiation via the Siglec receptors Siglec-7 and Siglec-9. Nat Commun, 2021, 12: 1270

65 Peng H, Wang K, Chen Z, et al. MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. Nucleic Acids Res, 2020, 48: D1085–D1092

66 Rigden D J, Fernández X M. The 27th annual *Nucleic Acids Research* database issue and molecular biology database collection. Nucleic Acids Res, 2020, 48: D1–D8

67 Rigden D J, Fernández X M. The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection. Nucleic Acids Res, 2021, 49: D1–D9

68 Rigden D J, Fernández X M. The 2022 *Nucleic Acids Research* database issue and the online molecular biology database collection. Nucleic Acids Res, 2022, 50: D1–D10

69 Rigden D J, Fernández X M. The 2023 *Nucleic Acids Research* database issue and the online molecular biology database collection. Nucleic Acids Res, 2023, 51: D1–D8

Current status and prospect of the National Genomics Data Center

CHEN MeiLi[1], WANG YanQing[1], LI RuJiao[1,2], MA YingKe[1], ZHANG SiSi[1], ZHANG Xin[1], SONG ShuHui[1,2], 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

Since its foundation in 2019, the National Genomics Data Center (NGDC) has played a pivotal role in addressing China's strategic goals related to population health and sustainable social development. Over time, NGDC has preliminarily established a suite of systems designed for the submission, deposition, management and sharing of multi-omics data. These systems encompass fundamental omics data repositories, national human genetics repositories, strategically significant organisms and biosecurity databases resources. Additionally, NGDC has incorporated a range of bioinformatics analysis tools and platforms into its infrastructure. These resources collectively serve as valuable references for research areas such as population health, public health security, breeding improvement, biodiversity conservation and related fields. Up to date, NGDC has successfully archived an impressive 27.6 PB of omics data. The Data made accessible through NGDC garnered recognition and endorsement from prominent international publishers, including Springer Nature, Elsevier, Wiley, and Taylor & Francis. Moreover, NGDC has consistently been recognized as one of the major global data centers by *Nucleic Acids Research* for six consecutive years. However, NGDC acknowledges the need for further improvement to reach the level of world-class data centers. In the future, NGDC will prioritize the development of intelligent and automated data curation systems, data fusion and retrieval capabilities, and the establishment of cloud platform tailored for biological big data. Advanced algorithms will also be a key focus of NGDC's technological advancements. Simultaneously, NGDC is committed to enhancing its efforts in grant applications, talent development and international collaborations. These endeavors collectively contribute to NGDC's aspiration to become a world-leading genomics data center, supporting the innovative advancements in life and health sciences in China, driven by self-reliance and continuous self-improvement.

genome, bioinformatics, big data, multi-omics, human genetic resources, repository management, National Genomics Data Center, China National Center for Bioinformation

doi: 10.1360/SSV-2023-0150