农业大数据学报 2024, 6(3): 325-332
Journal of Agricultural Big Data
DOI: 10.19788/j.issn.2096-6369.000053

开放科学（资源服务）标识码（OSID）：

生命组学大数据安全管理实践

王彦青[1,2]，陈婷婷[1,2]，张思思[1,2]，朱军伟[1,2]，陈焕新[1,2]，肖景发[1,2,3]，宋述慧[1,2,3]，章张[1,2,3]，赵文明[1,2,3*]，鲍一明[1,2,3*]

1.国家生物信息中心，国家基因组科学数据中心, 北京 100101；2.中国科学院北京基因组研究所, 北京 100101；3.中国科学院大学, 北京 100049

摘要：生命组学大数据是国家重要基础性、战略性资源，对支撑生命科学基础研究和应用创新、推动生物经济创新发展、维护国家安全具有重要意义。随着数据规模的不断增长，生命组学大数据的安全管理问题逐渐凸显。国家基因组科学数据中心（National Genomics Data Center, NGDC）面向我国人口健康和社会可持续发展的重大战略需求，建立了生命与健康大数据汇交存储、安全管理、开放共享与整合挖掘研究体系，形成了一系列数据安全管理的制度和措施。本文聚焦于生命组学大数据全生命周期的安全管理问题，探讨生命组学大数据安全管理框架，全面分析在数据汇交、存储、管理、共享全生命周期中涉及的安全管理内容，并总结了 NGDC 在生命组学大数据安全管理方面的成效。最后，本文展望了生命组学大数据安全管理的发展方向，包括完善数据分级分类制度、提升数据分级安全管理技术和加强数据异地灾备建设，以期实现生命组学大数据的安全管理与可持续发展。

关键词：生命组学大数据；数据汇交；数据共享；安全管理

1 引言

生命组学大数据是生命体通过高通量测序技术所衍生出的基因组、变异组、转录组、表观组等多维生物数据，是生命科学研究范式转变和产业创新发展的核心驱动力。生命组学大数据推动生物安全、人口健康、社会可持续发展等国家重大战略和世界科学前沿的基础科学研究，对支撑我国抢占未来生命科学和健康医学发展制高点具有重大战略意义、科学价值和社会经济效益。

随着数据量的急剧增长和应用领域的不断拓展，在全球数字化进程加速、网络威胁增加以及跨境生物信息流通的背景下，生命组学大数据安全问题逐渐凸显。全球范围内，各个国家在生物数据安全方面不断加强立法保护。美国于 2021 年提出的《基因组学支出和国家安全增强法案》（Genomics Expenditures and National Security Enhancement Act, 简称基因法案, the GENE Act）[1]和《基因组数据安全法案》（Genomics Data Security Act）[2]，聚焦基因数据安全保护，对外资企业，特别是受关注国家的企业在基因数据采集、使用、共享等方面制定了严格的监管措施；2024 年提出的《生物安全法》（BIOSECURE Act）[3]草案，计划禁止美国行政机构与受关注的生物科技公司进行特定交易，防止美国人的基因数据流入受关注国家（Countries Of Concern）；2024 年初，美国总统拜登签发的《关于防止受关注国家获取美国人大量敏感个人数据和美国政府相关数据的行政命令》（Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern）[4]，提出建立对受关注国家在包括个人生物特征数据、健康数据、组学数据等敏感个人数据方面交易监管制度。欧盟也发布了一系列数据安全相关的法规，其中最为重要的是 2018 年发布的《通用数据保护条例》（General

收稿日期：2024-06-13；**录用日期：**2024-09-13
基金项目：国家重点研发计划（2023YFC2605700, 2023YFC2604400, 2021YFF0703704），中国科学院基因组科学数据中心运行维护（CAS-WX2022SDC-XK05）。
联系方式：王彦青，E-mail：wangyanqing@big.ac.cn；陈婷婷，E-mail：chentt@big.ac.cn；张思思，E-mail：zhangss@big.ac.cn。王彦青、陈婷婷、张思思为同等贡献作者。通信作者赵文明，E-mail：zhaowm@big.ac.cn；通信作者鲍一明，E-mail：baoym@big.ac.cn。

Data Protection Regulation, 简称 GDPR）[5]，该条例强调个人数据保护的重要性，要求研究者在收集、存储、处理和传输个人数据时必须获得明确同意，并对将个人数据转移到欧盟以外的国家或地区进行了严格的规定，确保这些国家或地区的保护水平与欧盟一致。

在中国，2021 年《中华人民共和国生物安全法》[6]、《中华人民共和国数据安全法》[7]和《中华人民共和国个人信息保护法》[8]的陆续实施标志着生物信息数据安全在政策层面得到了高度重视。这些法律的出台促使国家逐步建立了一个以“法律/法案-法规-指南”为框架的数据安全管理体系。在人类遗传资源管理方面，分别于 2019 年和 2023 年发布了《中华人民共和国人类遗传资源管理条例》（国务院令第 717 号）[9]和《人类遗传资源管理条例实施细则》（科学技术部令第 21 号）[10]，涵盖人类遗传资源的收集、保存和利用，明确监管职责和法律责任。通过这一管理体系，中国致力于在保障数据安全和个人权益的同时，实现生物科技的创新发展和社会责任的平衡，在保护个体和社会利益同时，也为生物信息管理提供了新的发展方向。

在生物数据安全问题备受关注的态势下，构建统一、安全的生命组学大数据汇交、管理和共享平台，促进数据的合规、有序、安全共享，不仅具有现实的科学意义，而且具有重要的社会意义。国家基因组科学数据中心（National Genomics Data Center, NGDC）[11]正是在这样的背景下，在中国科学院北京基因组研究所生命与健康大数据中心[12]的基础上，由科技部、财政部于 2019 年 6 月 5 日发文成立。NGDC 也是 20 个国家科学数据中心之一，其目标是面向我国人口健康和社会可持续发展的重大战略需求，建立生命与健康大数据汇交存储、安全管理、开放共享与整合挖掘研究体系，建设支撑我国生命科学发展、国际领先的基因组科学数据中心。经过几年的发展，NGDC 已初步建成具有自主知识产权、安全可控的多维组学数据汇交、存储、管理和共享体系，包括组学原始数据归档库（Genome Sequence Archive, GSA）[13-14]、人类遗传资源组学原始数据归档库（Genome Sequence Archive for Human, GSA-Human）[14-15]、基因组数据库（Genome Warehouse，GWH）[16]、基因序列数据库（GenBase）[17]、基因组变异数据库（Genome Variation Map，GVM）[18]和多元数据归档库（Open Archive for Miscellaneous Data, OMIX）[14]等，承载着我国生物数据安全管理的使命，为科研用户提供不同组学数据的汇交、存储、管理和共享，以及国家重大科技项目数据管理服务。

2 生命组学大数据安全管理框架

面向生命组学大数据的汇交存储和共享应用，以促进生命组学大数据安全共享为目标，针对数据汇交、审核、存储、共享全生命周期安全管理的需求，构建生命组学大数据安全管理框架。从管理制度、网络安全、核心技术、系统服务等 4 个层面出发，制定标准和规范，搭建安全网络防护环境，研发核心技术和系统，建立数据服务平台，全方位保障生命组学大数据的安全管理。整体框架如图 1 所示。

2.1 安全管理制度

依照国家法律法规及行业规范，制定生命组学大数据安全管理相关规范和制度，用于指导数据管理系统设计、建设、运行和维护的全流程操作。遵照中华人民共和国《生物安全法》《数据安全法》《人类遗传资源管理条例》等相关规定，结合生命组学大数据的多维、多模态的特点，制定适用于生物信息领域的数据分级分类标准。在数据分级分类标准的基础上，建立数据分级操作规范，针对不同的数据级别，对数据管理者和用户进行权限分级，明确不同级别数据存储、处理和访问的安全边界。数据安全管理制度主要面向数据操作人员（包括数据提交者、数据管理者、数据使用者等），建立数据安全相关的管理制度和指南，实现数据全流程操作的安全、合规。系统开发管理规范主要面向系统开发和运维人员，对系统设计、开发、测试、运维的每个环节制定安全操作准则，最大化降低系统层面的安全漏洞，确保数据安全管理。

2.2 网络和系统环境安全防护

网络安全建设是生命组学大数据安全管理的重要部分，为数据汇交和共享提供安全可靠的网络环境。网络安全环境建设需要充分结合管理数据的重要性分级情况，既要保证数据得到充分的安全防护，也要为用户的数据共享提供便利的获取途径。除部署基础的防火墙、堡垒机等必要的网络安全设备和入侵防御、态势感知等安全系统外，还要据生命组学大数据不同程度的安全需求，建设不同安全等级的网络保护系统，并设置不同的访问控制策略。如对重要数据，建立网络安全等级保护三级系统，在网络层面规划数据独立

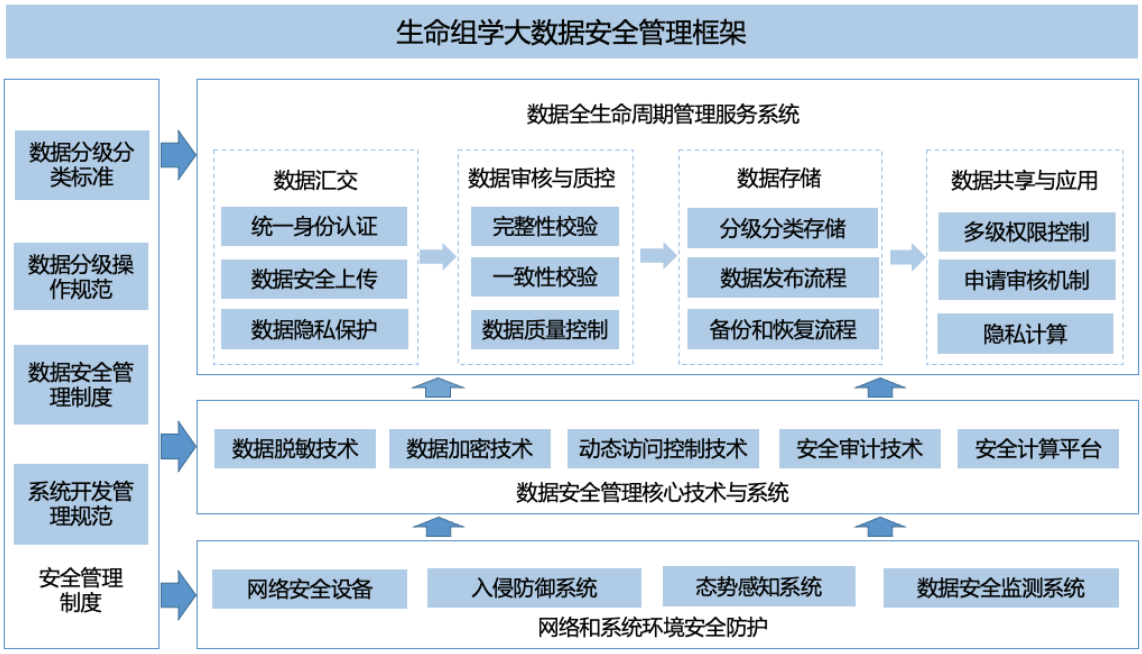

图 1　生命组学大数据安全管理框架

Fig. 1　Big data security management framework of life omics

管理区域，建立更加严格的边界控制、访问确权、运维审计等，保证数据可控访问的实现。最后，需要在网络层面建立数据安全监测系统，对数据的流入、流出进行监测和管控，保证数据的入口和出口安全。

2.3　数据安全管理核心技术与系统

针对生命组学大数据管理的全生命周期，开发通用的数据安全管理算法与工具，形成组学数据安全管理核心技术体系，为数据汇交、管理、共享等服务系统提供技术支撑。具体包括数据脱敏、数据加密等数据安全保护技术，数据动态访问控制技术和数据安全审计技术。数据脱敏技术包括元数据脱敏及组学序列文件脱敏，需要采用特定的算法，对可能包含的隐私信息如姓名、年龄、身份证号等信息进行删除或修改；另外，针对图像、测序数据等文件中的敏感信息，研发相应的方法和技术，实现多组学数据文件脱敏。数据加密是指对于安全级别较高的数据，采用加密算法进行加密，保证数据在流转过程中的安全。数据动态访问控制技术针对多种类、多级别的数据，研发细粒度的数据权限和访问控制技术，在保证数据访问安全的情况下，促进数据的开放共享。安全审计技术针对数据汇交、存储、共享全流程操作过程，建立数据审计方法，实现数据操作可追踪、可溯源。此外，利用云计算技术，整合中心数据资源和算力资源，建立安全计算平台，为用户提供在线数据计算服务，实现数据不出域情况下的安全计算，方便用户使用数据，促进数据的共享利用。

2.4　全生命周期数据服务系统

围绕数据汇交存储、共享应用全生命周期，建立全链条式数据服务系统，为用户提供包括数据汇交、数据审核与质控、数据共享、数据分析等服务。数据汇交系统主要包括用户身份认证、元数据递交和数据文件上传等过程，需要实现用户登录、权限分配以及数据安全上传和数据隐私保护；数据审核与质控系统主要实现数据的完整性和一致性校验和数据的质量控制，为用户提供高质量的数据资源，并保证数据在递交、归档、发布和共享全流程的一致性；数据存储系统根据数据分级分类标准，建立数据分级分类存储机制，为数据汇交、归档、发布的全过程提供数据存储空间和转移通道，确保数据安全流转；此外，为保障数据存储安全，还需要建立备份及恢复策略，实现数据的实时备份和灾难恢复。数据共享与应用系统以数据开放共享为目标，根据数据的不同级别，建立生命组学数据多级共享

和访问控制机制；此外，针对组学大数据安全、便捷利用的需求，结合隐私计算等技术，打通数据系统与安全计算平台的通道，研发数据隐私计算服务模块，实现数据的“可用不可得”。

3 生命组学大数据安全管理实践

国家基因组科学数据中心在生命组学大数据汇交、管理、共享实践中，贯穿数据的全生命周期管理，在数据汇交、审核、存储、管理、发布和共享等阶段实施了一系列的安全措施，保障数据的全流程安全。

3.1 数据递交

数据递交包括用户登录、元数据信息递交和数据文件上传等过程。用户必须在系统中注册账号，经过身份认证后才能进行后续的操作。

在用户账号管理和身份认证方面，NGDC 引入 Single Sign On (SSO)单点登录系统[19]，实现了各汇交子系统的用户统一注册、管理、登录和鉴权。从安全角度考虑，统一单点登录减少了用户密码管理的负担，降低了弱密码和密码重复使用的风险；通过提供统一登录入口，减少多个登录界面，降低了网页漏洞和网页攻击的风险；SSO 可以更容易地实施和执行复杂的安全策略，如多因素认证(MFA)。此外，集中的身份验证系统能更有效地监控和管理用户的访问，可及时发现并响应异常活动。因此，统一单点登录系统提高了数据汇交体系整体认证系统的安全性。

鉴于人类遗传资源数据的重要性，NGDC 对于人类遗传资源组学原始测序数据的递交制定了更加严格的规范。在 SSO 统一账号安全管理的基础上，增加了对人遗相关系统数据提交者账号的要求。例如，GSA-Human 系统只允许以课题研究组长的身份进行数据提交，在提交数据的人工审核阶段，系统会对数据提交者进行项目负责人身份认证，以确保数据提交者是人类遗传资源数据的责任人，保证数据全生命周期的安全可控管理[15]。

在数据文件上传方面，结合 SSO 统一认证技术，为用户提供增强的数据访问控制权限。例如，系统的 FTP 上传服务通过 SSO 账号进行统一身份认证，并为每个注册用户分配独立的数据上传空间，用户只能访问和读写自己的上传目录。用户上传文件时，必须使用 SSO 账号进行登录，并将文件上传至自己的数据空间。这在一定程度上保护了用户数据上传的安全性。特别地，针对人类遗传资源数据，也开辟了专用上传通道，以加强数据安全保障。

3.2 数据审核与质控

数据的完整性和一致性直接影响数据共享平台的可靠性和可用性。为确保数据不被非法修改和破坏，中心制定了严格的数据审核与质控策略。在数据递交、质控、归档、发布和共享的每一步骤，都会进行数据一致性校验。在方法上，利用文件的 MD5 码来验证数据的一致性。系统规定用户在提交数据时，必须同时提供文件的 MD5 码，并在数据关联、归档、发布等各个环节，进行数据 MD5 码复验，以保证数据在流转过程中的一致性。该策略保证了用户本地存储的数据、系统归档和发布的数据以及用户下载的数据之间的一致性。此外，针对不同的数据类型，分别建立数据质量控制流程，对用户汇交的数据进行严格的审核和质量控制，保证数据的完整性、高质量。例如，Fastq 格式的原始测序文件的质控流程，除了审核文件格式，还会对文件内容进行审核，该流程目前共能检测出 30 余种常见的文件错误。此外，还对错误类型进行细分编码并实时反馈用户质控结果，方便用户了解错误情况。数据一致性、完整性校验，是数据管理和数据安全的基础，能够为用户提供可信、高质量的数据，促进数据资源的高效流转和利用。

3.3 数据存储

中心初步制定了分级分类存储策略，以保障数据存储安全。将数据按照两个维度进行分类：数据类型和访问权限。首先，根据不同的类型，如原始测序数据、基因组组装数据、基因数据、变异数据等，为数据划分独立存储空间；其次，按照访问权限，将数据划分为私有数据、受控数据和公共数据，并对每类数据采取物理隔离的方式进行存储。将私有数据和受控数据分别存储在私有存储空间和受控存储空间，不提供对外访问接口；公共数据存放于公共存储空间，并建立数据访问接口，用户可通过 FTP 或 HTTPS 等访问接口进行访问。系统包含完善的数据发布流程，当数据发布后，会自动从私有空间转存到公共空间。此外，还建立了完善的数据备份和恢复策略，所有数据都以磁带库的方式进行备份，确保数据存储安全。

3.4 数据共享

数据的安全共享是生命组学大数据安全管理的重

要环节。NGDC 以数据开放共享为目标，遵循我国生命组学数据相关法律法规，参考国际生命组学数据共享政策，建立生命组学数据开放共享机制。在数据共享方式方面，将数据分为一般数据和重要数据，并根据数据的不同类别，制定不同的数据访问方式。一般数据主要是指病原体、动植物、微生物等普通物种的原始组学测序数据、分析过的数据如基因组组装数据、变异数据、转录组数据、表观组数据等；重要数据是指人类遗传资源相关的原始组学测序数据、基因组变异数据等。一般数据采用公开访问方式，全球用户都可通过 NGDC 的 BIG Search 平台在线检索、浏览、下载已发布的公开访问数据。对于重要数据，采用“申请–审核制”的受控访问方式，数据使用者须通过数据平台向数据所有者提出数据使用申请，获得授权后才有权限下载使用。在受控数据安全访问和下载方面，NGDC 提供了基于 Apache Mina FtpServerp[20]开发的文件传输服务系统。系统引入单点登录账号信息和 NGDC 账号内部 ID 信息双重认证机制，对登录鉴权进行多因子校验，保证用户登录安全；此外，为通过数据授权的用户开辟专有的数据目录，并对数据设置只读权限，通过安全隔离和最小操作权限的方式，进一步保证数据下载的安全性。

为了更好地落实执行国家在人类遗传资源数据管理方面的制度，NGDC 遵照《中华人民共和国人类遗传资源管理条例》《人类遗传资源管理条例实施细则》等法律法规，制订了《人类遗传资源数据共享政策》[21]，数据使用者需要在遵循该政策的基础上，申请、下载和使用数据。该政策符合相关国际规范，包括禁止对下载的数据进行再分发、再传播等行为。NGDC 的人类遗传资源数据共享机制有效保护了人类遗传资源数据的合理、合法和合规使用，降低了安全风险和隐患。

3.5　网络和系统环境安全防护

在系统和网络安全防护方面，NGDC 也部署了一系列的网络安全设备，为数据汇交访问提供安全可靠的网络环境。目前 NGDC 已建立网络安全等级保护二级、三级系统各一个，建成了较为完整的网络安全防御体系，在互联网接入、办公、安全运维、业务生产区域间设置不同的访问控制策略，部署 IPS、WAF、防火墙、堡垒机、防病毒、日志审计、数据库审计等必要的网络安全设备，并完善数据全生命周期安全管理制度。

实践中，NGDC 已开展数据分级工作，将一般数据和相关应用系统纳入网络安全等级保护二级系统管理，重点关注数据的完整性和可获得性，为科研用户提供公开的数据汇交、共享服务。针对具有一定规模的人类遗传资源等重要数据，建立网络安全等级保护三级系统，建立严格的数据授权访问机制，在网络层面规划数据独立管理区域，建立更加严格的边界控制、访问确权、运维审计、加密策略等，保证数据可控访问的实现。

4　成效

面向国家生物安全、人口健康、生物多样性等重大战略需求，国家基因组科学数据中心建立了自主安全可控的综合性数据汇交管理体系，保障国家数据安全、主权和发展需求。截至 2024 年 6 月，数据汇交体系服务用户 2,735 万人，累计服务各级各类科技计划项目（含课题和子课题）2 万余个，支撑发表文章 3,584 篇，汇交的数据总量超过 50.7 PB，数据日均下载量超 6 TB。

在服务国家人类遗传资源信息管理方面，受中华人民共和国科学技术部委托，中心自 2022 年 7 月 18 日起开始承担我国人类遗传资源信息统一汇交管理任务，整合已有数据汇交管理系统（包括 GSA-Human、GenBase、GWH、GVM 和 OMIX），建成人类遗传资源信息管理、备份、发布与共享一体化管理体系。截至 2024 年 6 月，备份平台已分配备份编号 3,564 个，关联归档数据的总量超过 2.6 PB。在此期间，受控数据累计申请 1,658 次，授权数据下载总量 756.70 TB。

此外，中心的网络安全防御体系已形成了网络边界和内部安全计算区域多层防护、重要系统持续审计和周期性网络安全检测等安全机制，高峰时网络边界每周记录攻击告警总数近 100 万次，年主动开展安全检测网站页面超 10 万页面/次。

5　问题与建议

近年来，国际社会越来越重视生物安全，涉及生物数据安全、隐私保护、数据共享等方面的法律法规不断出台和完善。国家基因组科学数据中心虽然在生物组学大数据汇交、管理、共享等方面取得了一定的成果，也实施了一系列的生物数据安全管理措施，但是，在生物数据隐私保护、数据分级分

类管理、数据高效共享和利用等方面，还有待进一步加强。

5.1 推进生命组学大数据分级分类标准制定与实施

基因组科学与生物安全、临床医学等多个领域的交叉发展，使得科研领域内通用的数据开放政策与数据保护法规之间的矛盾逐渐显现[22]。为了在国家政策和开放科学之间达到平衡，应尽快在国家层面推进生命组学大数据分级分类标准的制定和实施，并依据标准研发相应的数据分级保护技术，促进生命科学数据安全、合规、高效共享。

对于生命组学大数据，应根据数据的重要性、敏感性及数据遭到破坏后的危害性等对其进行细致的分级分类，以便采取相应的安全措施进行精准保护。2021年11月14日，国家互联网信息办公室发布了《网络数据安全管理条例（征求意见稿）》[23]，根据对国家安全、公共利益及个人、组织合法权益的影响，将数据分为一般数据、重要数据和核心数据三级。袁康等人在此基础上构建了5级数据安全分级体系,并给出了重要数据特征的相关分析[24]。对于生命组学大数据，可先基于数据的组学类型，初步分类为基因组、变异组、转录组、翻译组、表观组、蛋白质组、代谢组等。其次，可根据数据挖掘程度，将每个类型的组学数据细分为个体原始输出数据、个体分析数据、群体统计数据等不同层次。最后，对不同样本类型、数据类型的数据进行风险评估，并根据风险级别和重要程度进行分级。在数据安全管理和共享系统建设中，研发相应的数据分类分级安全保护技术，保护数据存储管理和共享利用的安全性，保障生命科学研究和健康医学安全、快速发展。

5.2 融合先进信息和计算技术，进一步强化生命组学大数据安全管理技术体系建设

生物组学大数据安全与国家生物安全息息相关，目前国家级生物安全大数据分析核心算法匮乏，也是我国生物安全面临的重要问题[25]。作为生物组学大数据管理平台，应该从数据收集、存储、共享、利用等各个环节进行安全技术提升，以加强数据的安全性保障。目前，在基因组学数据隐私处理领域，已产生相关技术体系，如利用转换、聚合、混淆、合成等数据扰乱技术实现数据隐私保护，使用同态加密、安全多方计算、可信执行环境、区块链等加密技术实现基因数据的安全使用[26]等。在数据共享与应用方面，除了常规的数据受控管理模式外，采用云计算、联邦计算等可提供灵活便捷的计算资源和数据处理流程的新型计算模式[27-28]在一定程度上可以保护人类遗传数据的安全可控。未来，我们将针对生物组学数据及其共享应用的需求，利用数据加密解密、云计算、联邦计算等先进技术，从数据要素安全、数据共享和应用安全等多角度出发，研发数据安全管理技术体系，保障生物组学大数据的安全、高效利用。

第一，数据要素安全。针对人类遗传资源数据等重要生物数据安全保护的需求，结合基因组多组学数据的特点，开发加密解密、数据脱敏等技术，保护数据的私密性和安全性。

第二，数据共享安全。利用零信任技术，根据数据的分级分类情况，建立细粒度的访问控制及风险监测机制，保障数据访问和流转安全。

第三，数据应用安全。原始测序数据通常体量较大，数据传输不够便捷，并且数据分级策略通常将重要原始测序数据划分在受控级别较高的等级，不便公开获取。因此，结合云计算、隐私计算等技术，构建数据在线分析处理平台，使用户可以在不接触原始数据的前提下获得分析结果，真正实现原始数据的可用不可得，提升数据利用效率。

5.3 加快数据异地灾备设施建设，提升数据安全保障能力

灾难备份是开展大数据安全工作的基础，必须能够确保出现极端危害事件的情况下可以获得有效的数据副本。目前通行的做法是建立空间独立，在同一时刻无相同危害风险的备份中心。

NGDC正在规划相关的基础设施的建设工作，可选方案包括具有设施基础的分中心，此方案具有地域距离远、同发灾难概率低的优势，但需进行独立的设施投入。利用未来国家生物信息中心基础设施资源，也是较为可行的方案之一，具有无需重复投入、设施标准高的优势，但国家生物信息中心与NGDC现有设施的距离较近，存在同发灾难事件的可能性。

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。

参考文献

[1] Genomics Expenditures and National Security Enhancement Act [EB/OL]. https://www.congress.gov/bill/117th-congress/senate-bill/1745/text.

[2] Genomics Data Security Act [EB/OL]. https://www.congress.gov/bill/117th-congress/senate-bill/1744/text.

[3] BIOSECURE Act [EB/OL]. https://www.congress.gov/bill/118th-congress/house-bill/7085/text

[4] Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern [EB/OL]. https://www.federalregister.gov/documents/2024/03/01/2024-04573/preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related.

[5] General Data Protection Regulation [EB/OL]. https://gdpr-info.eu/.

[6] 中华人民共和国生物安全法 [EB/OL]. https://www.gov.cn/xinwen/2020-10/18/content_5552108.htm?eqid=ee76ba160000091a000000036465eef7.

[7] 中华人民共和国数据安全法 [EB/OL]. https://www.gov.cn/xinwen/2021-06/11/content_5616919.htm.

[8] 中华人民共和国个人信息保护法 [EB/OL]. https://www.gov.cn/xinwen/2021-08/20/content_5632486.htm.

[9] 中华人民共和国人类遗传资源管理条例 [EB/OL]. https://www.safea.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/flfg/201906/t20190612_147044.html.

[10] 人类遗传资源管理条例实施细则 [EB/OL]. https://www.gov.cn/zhengce/202306/content_6887562.htm.

[11] CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024[J]. Nucleic Acids Research, 2024, 52(D1): D18-D32.

[12] BIG Data Center Members. The BIG Data Center: from deposition to integration to translation[J]. Nucleic Acids Research, 2017,45(D1): D18-D24.

[13] WANG Y, SONG F, ZHU J, et al. GSA: Genome Sequence Archive[J]. Genomics Proteomics Bioinformatics, 2017, 15(1):14-18.

[14] CHEN T, CHEN X, ZHANG S, et al. The Genome Sequence Archive Family: Toward explosive data growth and diverse data types[J]. Genomics Proteomics Bioinformatics, 2021,19(4):578-583.

[15] 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, 43(10):988-993.

[16] CHEN M, MA Y, WU S, et al. Genome Warehouse: A public repository housing genome-scale data[J]. Genomics Proteomics Bioinformatics, 2021,19(4):584-589.

[17] BU C, ZHENG X, ZHAO X, et al. GenBase: A nucleotide sequence database[J]. Genomics Proteomics Bioinformatics, 2024, qzae047.

[18] LI C, TIAN D, TANG B, et al. Genome Variation Map: A worldwide collection of genome variations across multiple species[J]. Nucleic Acids Research, 2021, 49(D1):D1186-D1191.

[19] Single Sign-On [EB/OL]. https://www.apereo.org/projects/cas.

[20] Apache MINA FtpServer [EB/OL]. https://cwiki.apache.org/confluence/display/FTPSERVER/Index.

[21] 国家基因组科学数据中心人类遗传资源数据共享政策 [EB/OL]. https://ngdc.cncb.ac.cn/gsa-human/document/Principle_of_Accessing_Human_Genetic_Resource_Data_in_NGDC_V1.pdf.

[22] Mark Phillips. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR)[J]. Human genetics, 2018.

[23] 网络数据安全管理条例（征求意见稿）[EB/OL]. https://www.cac.gov.cn/2021-11/14/c_1638501991577898.htm.

[24] 袁康,鄢浩宇. 数据分类分级保护的逻辑厘定与制度构建——以重要数据识别和管控为中心[J]. 中国科技论坛, 2022(7):167-177.

[25] 王秉,朱媛媛.大数据环境下国家生物安全情报工作体系构建[J/OL]. 情报杂志, 2021, 40(6):82-88. https://kns.cnki.net/kcms/detail/61.1167.G3.20210511.1427.028.html.

[26] WAN Z, HAZEL J W, CLAYTON E W, et al. Sociotechnical safeguards for genomic data privacy[J]. Nature Reviews Genetics, 2022, 23:429-445.

[27] Genomic Data Science Community Network. Diversifying the genomic data science research community[J]. Genome Research 2022, 32: 1231-1241. doi:10.1101/gr.276496.121.

[28] LANGMEAD B, NELLORE A. Cloud computing for genomic data analysis and collaboration[J]. Nature Reviews Genetics, 2018, 19(4): 208-219. DOI: 10.1038/nrg.2017.113.

引用格式： 王彦青,陈婷婷,张思思,朱军伟,陈焕新,肖景发,宋述慧,章张,赵文明,鲍一明.生命组学大数据安全管理实践[J].农业大数据学报,2024,6(3): 325-332. DOI: 10.19788/j.issn.2096-6369.000053.

CITATION: WANG YanQing, CHEN TingTing, ZHANG SiSi, ZHU JunWei, CHEN HuanXin, XIAO JingFa, SONG ShuHui, ZHANG Zhang, ZHAO WenMing, BAO YiMing. Practice of Security Management of Omics Big Data in Life Sciences[J]. Journal of Agricultural Big Data,2024,6(3): 325-332. DOI: 10.19788/j.issn.2096-6369.000053.

Practice of Security Management of Omics Big Data in Life Sciences

WANG YanQing[1,2], CHEN TingTing[1,2], ZHANG SiSi[1,2], ZHU JunWei[1,2], CHEN HuanXin[1,2], XIAO JingFa[1,2,3], SONG ShuHui[1,2,3], ZHANG Zhang[1,2,3], ZHAO WenMing[1,2,3*], BAO YiMing[1,2,3*]

1. National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China; 2. Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China; 3. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Omics big data is a significant foundational and strategic resource for the country, which plays an important role in supporting the basic research and application innovation of life sciences, promoting the innovative development of bioeconomy, and maintaining national security. With the rapid accumulation of omics data, the security of data management has become increasingly prominent. Facing the major strategic needs of China's population health and sustainable social development, the National Genomics Data Center (NGDC) has established a comprehensive research architecture for collecting, storing, managing, sharing, and mining of big data in omics, forming a series of practices and measures for the security management of the data. This paper delves into the issues of security management of omics big data throughout its lifecycle, elaborating on NGDC's security management measures implemented in the collecting, storing, managing and sharing of the data. Furthermore, it summarizes NGDC’s achievements in the security management of omics big data. Finally, this paper envisions the future directions for the security management of omics big data, including enhancing the data classification and categorization system, enhancing data hierarchical security management technologies and strengthening the construction of off-site disaster recovery, in order to achieve the security management and sustainable development of omics big data in life sciences.

Keywords: omics big data; data archive; data sharing; security management