遗传 Hereditas (Beijing) 2021 年 10 月, 43(10): 988－993
www.chinagene.cn

GSA-Human：人类遗传资源数据管理的公共系统

张思思[1,2], 陈旭[1,2], 陈婷婷[1,2], 朱军伟[1,2], 唐碧霞[1,2], 王安可[1,2], 董丽莉[1,2], 张哲文[1,2], 孙艳玲[1,2], 俞彩霞[1,2], 翟爽[1,2], 孙玉彬[1,2], 陈焕新[1,2], 杜政霖[1,2,3], 肖景发[1,2,3], 章张[1,2,3], 鲍一明[1,2,3], 王彦青[1,2], 赵文明[1,2,3]

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

摘要: GSA-Human 是人类遗传资源数据汇交、存储、管理与共享的数据库系统，可提供人类遗传资源数据的上传、下载、浏览、检索等公共服务，并有效支撑了国家重点研发计划科技项目数据的汇交与管理工作。系统具有符合《中华人民共和国人类遗传资源管理条例》数据安全管理策略，提供公开访问和受控访问相结合的数据使用模式。公开访问数据允许用户自由下载与获取；受控访问数据采用申请-审核的模式，即需要通过数据管理委员会(Data Access Committee, DAC)的授权方可获得下载和使用权限。系统自上线以来，截至 2021 年 7 月，汇集数据总量已超 5.27 PB。

关键词: 人类遗传资源数据管理系统；组学数据；数据汇交；数据共享

收稿日期: 2021−07−13; **修回日期:** 2021−09−16

基金项目: 国家重点研发计划资助项目(编号：2016YFC0901603，2017YFC0907502，2020YFC0847000)，中国科学院战略性先导科技专项基金资助项目(编号：XDB38050300，XDB38050200)，中国科学院关键技术人才基金资助项目(王彦青)，中国科学院“十四五”网络安全和信息化项目(编号：WX145XQ07-04)资助[Supported by the National Key R&D Program of China (Nos. 2016YFC0901603, 2017YFC0907502, 2020YFC0847000), the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDB38050300, XDB38050200), the Key Technology Talent Program of the Chinese Academy of Sciences (to Yanqing Wang) and the 14th Five-year Network Security and Informatization Plan of Chinese Academy of Sciences (No. WX145XQ07-04)]

作者简介: 张思思，博士，工程师，研究方向：基因组学、生物信息学。E-mail: zhangss@big.ac.cn
陈旭，硕士，工程师，研究方向：生物信息学、计算机科学。E-mail: chenx@big.ac.cn
陈婷婷，硕士，工程师，研究方向：基因组学、生物信息学。E-mail: chentt@big.ac.cn
张思思、陈旭和陈婷婷并列第一作者。

通讯作者: 王彦青，硕士，高级工程师，研究方向：生物信息学、计算机科学。E-mail: wangyanqing@big.ac.cn
赵文明，硕士，正高级工程师，研究方向：生物信息学。E-mail: zhaowm@big.ac.cn

DOI: 10.16288/j.yczz.21-248
网络出版时间: 2021/9/28 11:11:50
URI: https://kns.cnki.net/kcms/detail/11.1913.R.20210927.1137.001.html

GSA-Human: Genome Sequence Archive for Human

Sisi Zhang[1,2], Xu Chen[1,2], Tingting Chen[1,2], Junwei Zhu[1,2], Bixia Tang[1,2], Anke Wang[1,2], Lili Dong[1,2], Zhewen Zhang[1,2], Yanling Sun[1,2], Caixia Yu[1,2], Shuang Zhai[1,2], Yubin Sun[1,2], Huanxin Chen[1,2], Zhenglin Du[1,2,3], Jingfa Xiao[1,2,3], Zhang Zhang[1,2,3], Yiming Bao[1,2,3], Yanqing Wang[1,2], Wenming Zhao[1,2,3]

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

Abstract: The Genome Sequence Archive for Human (GSA-Human) is a data repository specialized for human genetic related data derived from biomedical researches, and also supports the data collection and management of National Key Research and Development Projects. GSA-Human has a data security management strategy according to the national regulations of human genetic resources. It provides two different models of data access: Open-access and Controlled-access. Open-access data are universally and freely accessible for global researchers, while Controlled-access ensures that data are accessed only by authorized users with the permission of the Data Access Committee (DAC). Till July 2021, GSA-Human has housed more than 5.27 PB of data from 750 datasets.

Keywords: GSA-Human; omics data; data submission; data sharing

数据是 21 世纪的珍贵财产。人类遗传资源数据关系到人口健康和人类社会的可持续发展，是国家重要战略资源。2019 年 7 月 1 日开始实施的《中华人民共和国人类遗传资源管理条例》(简称“条例”)明确了人类遗传资源范围，即人类遗传资源数据是指利用含有人体基因组、基因等遗传物质的器官、组织、细胞等材料产生的数据。该条例规定了人类遗传资源数据在采集、保藏、利用和对外开放方面的审批事项，为我国人类遗传资源数据的管理提供了指导思想。国家生物信息中心–国家基因组科学数据中心(National Genomics Data Center, China National Center for Bioinformation, CNCB-NGDC)于 2015 年建立的组学原始数据归档库(Genome Sequence Archive, GSA)(https://ngdc.cncb.ac.cn/gsa/)[1~3]，汇交、存储、管理和共享全球生命组学测序数据，为我国的科学数据管理发挥了重要作用。依托 GSA 系统，以人类遗传资源管理条例为指导原则，以《科学数据管理办法》和《数据安全法》为参照，CNCB-NGDC 建立了集数据汇交、分级存储、安全管理、受控共享等多个功能为一体的人类遗传资源数据管理系统(Genome Sequence Archive for Human, GSA-Human)(https://ngdc.cncb.ac.cn/gsa-human/)，以此推动我国人类遗传资源数据的安全管理与开放共享，促进数据开发与利用，服务于国家面向人口健康和生命安全相关的科学研究。

1　GSA-Human 系统建设

1.1　支持数据类型

通常，生命组学数据可分为三级：一级为原始测序数据，大多为通过基因测序仪或相关仪器设备产生的并经过简单整理和质量控制后的数据，这类数据含有最全面的信息；二级为原始测序数据经过一些处理，如序列比对、基因组拼接等操作后所产生的数据；三级为在二级数据的基础上进一步深加工产生的数据，如基因组的变异、基因注释、转录组表达量、表观组调控位点等分析结果数据。GSA-Human 主要面向一级测序序列数据，支持当前主流测序平台，如二代测序平台 Illumina、BGISEQ 等，三代测序平台 PacBio SMART、Bionano Genomics、Oxford Nanopore 等。针对二、三级数据，CNCB-NGDC 已建立了多个数据库系统收录并整合数据，如基因组数据库(Genome Warehouse, GWH)[4]，基因

组变异数据库(Genome Variation Map, GVM)[5]，基因组表达库(Gene Expression Nebulas, GEN)[6]，甲基化数据库(Methylation Bank, MethBank)[7]，多元数据归档库(Open Archive for Miscellaneous Data, OMIX)等。各数据库通过项目编号(BioProject accession)进行相互关联，相辅相成，形成了我国人类遗传资源数据安全存储和统一管理的公共平台。

1.2 数据组织模式

GSA-Human中的数据包括元数据信息和测序序列数据。元数据信息主要为测序序列数据的描述信息，鉴于人类遗传资源承载的基本对象是人，GSA-Human使用“个体”(individual)来描述研究对象，并组织与此研究对象相关联的信息，主要包括“样本信息”(sample)、“实验信息”(experiment)、“测序反应”(run)信息以及对应的测序序列数据(sequence)。其中，“个体信息”是用于收集取样对象的基本信息，主要收集包括性别、身体形态指标、生活习惯、疾病、治疗情况以及其他属性信息。当取样对象为细胞系时，主要收集原代培养物或细胞系的取样组织、生理性别和种族来源等信息。“样本信息”是主要收集研究涉及的生物样本描述，如样本类型、样本属性等。为更加灵活的实现个体和样本的元数据信息的收集管理，GSA-Human采样用固定词条与自定义属性相结合的方式组织数据，即系统设置个性化的数据描述字段以满足不同的数据管理需求。“实验信息”包括实验目的、文库构建方式、测序类型等信息。“测序反应”信息为测序文件所对应的校验信息，测序文件则为各种测序平台的测序原始数据，主要测序格式包括Fastq、BAM等。GSA-Human系统中，一个或多个个体组成的数据组由“研究信息”(study)数据模型进行统一管理，包括研究类型、数据访问机制、数据备份号与备案号[①]等信息。因此，“研究信息”被定义为GSA-Human中的一个独立数据集(dataset)，并以“HRA+6 位数字”(如“HRA000001”)编码进行唯一标识。各类数据元素之间采用层级及关联的模式进行组织，从而形成包括“研究(study)-个体(individual)-样本(sample)-实验(experiment)-测序反应(run)-序列数据(sequence)”的“金字塔”式的数据组织与管理模式。

1.3 数据质控与审核

GSA-Human 系统建立了元数据实时审核、人工审编和数据文件审编三个层次的数据质控与审核功能。元数据实时审核发生在数据录入过程中，审核内容包括数据合规性、一致性、控制词汇、专有术语和数据结构等。人工校验发生在数据录入之后，由 GSA-Human 的系统审编员执行，人工校验可以防止一些内容不当或垃圾信息进入系统并被公布，从而确保元数据信息的准确性，并使得系统中的数据干净整洁。数据文件审编由后台监控程序自动检测并触发运行，该过程主要检查用户递交序列数据的完整性和可靠性，防止数据文件在处理、压缩、拷贝、传输和存档过程中出现异常，自动化程序审核过程和内容包括：(1)文件压缩的正确性；(2)文件格式的合规性，目前主要的文件格式包括 Fastq 和 Bam 格式；(3)序列信息的统计，包括 reads 数量、碱基数量、reads 长度、碱基数量分布和 reads 长度分布等。针对用户递交的数据集，只有当元数据和序列数据均通过审核，GSA-Human 才为该数据集分配正式的访问序列号(accession number)。

1.4 数据管理委员会

GSA-Human 设置数据管理委员会(Data Access Committee, DAC)对数据的访问权限进行管理和控制。DAC 由数据递交者提供并在递交数据时创建，每一个需要受控管理的数据集均需设置 DAC，DAC 中可包含一个或多个成员，一般由资深专家组成，且需要设定一名 DAC 联系人(DAC contact)。DAC 是 GSA-Human 中审批数据使用请求的最终决策方，DAC 成员负责审核用户请求，DAC 联系人负责接收数据申请、组织 DAC 成员对数据申请进行审核、处理相关的决策决议。GSA-Human 为每个 DAC 分配一个编号，并实现与其管理数据的关联与访问。

1.5 数据安全保障措施

为保证人类遗传资源数据的存储安全，GSA-Human 从系统架构整体设计了多重安全防护措施。在用户身份认证方面，采取双重认证方式，用户既

[①]数据备份号与备案号为中华人民共和国科学技术部为人类遗传资源信息对境外机构提供或开放使用提供的审批编号。

需要通过 CNCB-NGDC 的单点登录系统(single sign-on, SSO)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，进行项目负责人身份信息核实，以确保数据的可溯源性。针对数据上传服务，GSA-Human 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。在数据存储方面，采用磁盘和磁带库相结合的数据备份方式，防止因意外事故造成数据丢失。在用户下载数据方面，实现了用户身份认证和数据访问目录权限控制的系统开发，并通过数据文件软连接(soft link)、授权账户关联以及自动权限控制的模式实现数据的受控共享，既保证了数据的安全性，也保障了多用户同时访问同一数据时的效率。

2　人类遗传资源数据汇交与共享

2.1　数据汇交原则与方法

为了有效管理和保护我国人类遗传资源数据，促进数据有序共享与合理利用，GSA-Human 建立了人类遗传资源数据汇交的基本规范，核心内容包括：(1)数据递交者身份认证，只允许以课题研究组长的身份进行数据提交，从而确保数据的可溯源性；(2)伦理合规性，即数据递交者应已经从数据集对应的研究对象处获得知情同意书，并符合伦理原则，通过相应的伦理审查；(3)隐私保护性，数据递交者提供的信息必须对其研究对象的个人信息进行脱敏处理；(4)政策合法性，数据递交者在对外发布其数据集前，遵循科技部人类遗传资源信息备案流程获得数据集备份号及备案号；(5)遵守科研诚信与道德，数据递交者对其提交的数据质量负责。

按照数据的组织模式，GSA-Human 的数据递交包含两部分内容：元数据递交和序列文件递交。元数据递交主要为在线递交(https://ngdc.cncb.ac.cn/gsa-human/submit/hra/submit)，即通过 WEB 页面实现信息输入、勾选、导入或确认；GSA-Human 提供可视化及向导化的操作模式，内置多种控制词汇表，最大限度地规范信息录入；此外，系统还提供批量表格在线导入与校验功能，实现元数据信息实时在线质控和信息反馈，为科研人员提供简单、便利、高效的元数据信息递交服务。在测序序列文件汇交方面，可支持 Aspera 和 FTP 两种在线数据上传方式。对于一次性上传数据量超过 1 TB 的数据递交，可以选择采用邮递硬盘的模式，由 GSA-Human 系统审编人员协助上传数据。

2.2　数据共享模式

GSA-Human 提供公开访问和受控访问两种共享访问模式。公开访问即已经发布的数据可被任何人浏览和下载，用户对数据的使用无须向数据递交者申请；受控访问即对数据使用在一定限制下进行，用户在下载数据之前需要先获得该数据的使用授权。共享模式的选择由数据递交者自行设定，但需要遵守相关的规则：尚未获得人类遗传资源数据备案编号的数据集(商用细胞系和古人类数据除外，依照相关规定此两类数据无须备案备份)不能设置为公开访问，已获得备案编号的数据集，可以设置为公开访问或受控访问。GSA-Human 支持的受控访问被称为“申请–审核制”(图 1)，即用户检索到所需数据

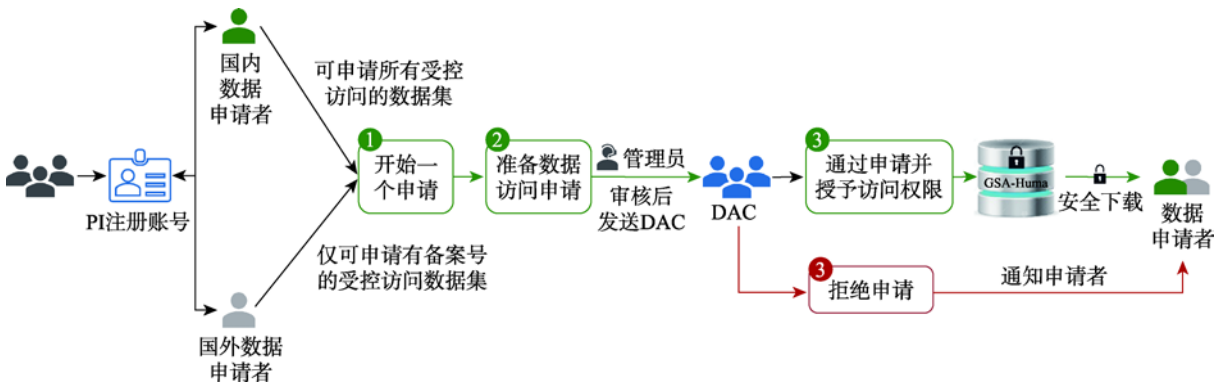

图 1　受控访问数据集申请下载流程

Fig. 1　The Data Access Request wizard

集后(https://ngdc.cncb.ac.cn/gsa-human/browse/)，通过数据申请下载模块(request)在线提交数据应用“申请单”，该数据的管理委员会(DAC)对“申请单”进行审核，如果审核通过，则赋予数据使用权利，用户需使用授权账号的用户名和密码登录系统，并通过 FTP 工具或命令行完成授权数据下载。GSA-Human 的数据共享政策遵循了相关的国际规范，比如获得数据后不能再分发、再传播，数据申请者要经过审核与认证等[8,9]。GSA-Human 的数据共享模式有效地保护了人类遗传资源数据的合理、合法、合规使用，降低了安全风险和隐患。

3 GSA-Human 运行状况

GSA-Human 自正式上线运行以来，已接收了来自全球用户递交的人类遗传数据集 750 个，已发布数据集 395 个，其中受控访问数据集 313 个，公开访问数据集 82 个，而受控访问数据集中已获得备案备份号的仅 43 个)；共合计收录个体数(individual) 71,283 个，生物学样本数(sample)159,747 个，实验数(experiment) 180,231 个，测序反应数(run)216,546 个，总数据量超过 5.27 PB，数据日增量统计如图 2 所示。GSA-Human 已接收来自 550 个用户的数据下载申请共 808 份，总数据下载量超过 300 TB。GSA-Human 已支撑数据递交用户在 *Cell*、*Science*、*Nature*、*Cancer Cell*、*Nature Immunology* 等 66 种国内外期刊发表论文 117 篇。此外，GSA-Human 承担国家重点研发计划与人类遗传资源相关的多组学数据汇聚与统一管理工作，截至 2021 年 7 月，已接收来自国家重点研发计划项目的原始测序下机数据共计 1.57 PB。

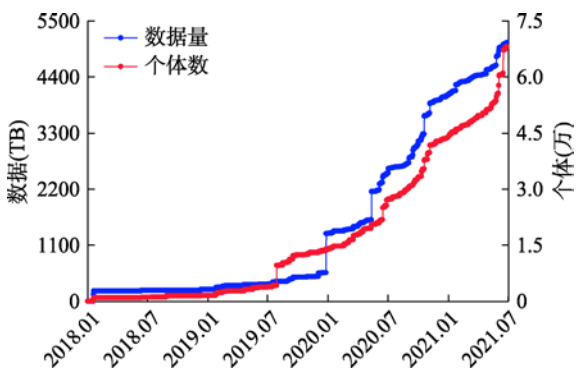

图 2 GSA-Human 数据增长情况统计图

Fig. 2 Statistics of data submissions of GSA-Human

4 结语与展望

GSA-Human 作为人类遗传资源组学数据汇交、存储和受控访问管理系统，接受来自全球的科研工作者的数据提交和共享请求，为人类遗传资源数据共享与利用提供了良好的平台。同时，GSA-Human 系统承担国家科技项目数据汇聚与管理任务，有力支撑了我国重大科研任务的科学数据管理。

GSA-Human 推行数据“申请–审核”制共享模式，采用数据管理委员会审批数据使用权限的机制，提升数据递交者对数据管理的自主权，在充分保障数据权益的同时激发了数据汇交的积极性，促进了我国人类遗传资源数据的共享与再利用。但随之而来的问题是大量的数据汇交与存储 GSA-Human 需求，这对当前系统的性能和数据存储能力，尤其是数据长期保存能力提出严峻的考验。因此，未来，GSA-Human 将从软件和硬件两方面出发，加强自身能力的建设。针对软件系统层次，在数据汇交和共享方面，将进一步优化数据提交、审核和申请流程，以及管理和共享机制；在数据信息检索方面，完善检索机制，逐步实现数据特性化检索；在数据自动化处理方面，不断完善流程和算法，实现智能化数据处理。此外，在遵守国内外法律法规和道德规范的前提下，实现更加安全、快捷、高效的人类遗传资源数据管理和共享。在硬件系统层次，将加强计算机存储系统和网络带宽资源的建设，优化硬件设施以提升大数据传输与存储效率，同时，借鉴区块链、云计算、流计算等数据安全管理的特性和理念，建立人类遗传资源数据共享和使用的新模式。

参考文献(References):

[1] Chen TT, Chen X, Zhang SS, Zhu JW, Tang BX, Wang AK, Dong LL, Zhang ZW, Yu CX, Sun YL, Chi LJ, Chen HX, Zhai S, Sun YB, Lan L, Zhang X, Xiao JF, Bao YM, Wang YQ, Zhang Z, Zhao WM. The Genome Sequence Archive family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics*, 2021, doi:

10.1016/j.gpb.2021.08.001. [DOI]

[2] 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]

[3] 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]

[4] Chen ML, Ma YK, Wu S, Zheng XC, Kang HE, Sang J, Xu XJ, Hao LL, Li ZH, Gong Z, Xiao JF, Zhang Z, Zhao WM, Bao YM. Genome Warehouse: A Public Repository Housing Genome-scale Data. *Genomics Proteomics Bioinformatics*, 2021, doi: 10.1016/j.gpb.2021.04.001. [DOI]

[5] Li CP, Tian DM, Tang BX, Liu XN, Teng XF, Zhao WM, Zhang Z, Song SH. Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res*, 2021, 49(D1): D1186–D1191. [DOI]

[6] CNCB-NGDC Members and Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res*, 2021, 49(D1): D18–D28. [DOI]

[7] Zou D, Sun SX, Li RJ, 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(D1): D54–D58. [DOI]

[8] Tryka KA, Hao LN, Sturcke A, Jin YM, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res*, 2014, 42(D1): D975–D979. [DOI]

[9] Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, Paschall J, Flicek P. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet*, 2015, 47(7): 692–695. [DOI]

(责任编委: 朱波峰)