

GSA: 组学原始数据归档库

张思思^{1,2}, 陈婷婷^{1,2}, 朱军伟^{1,2}, 周晴^{1,3}, 陈旭^{1,2}, 王彦青^{1,2}, 赵文明^{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: 生命科学的发展已进入组学大数据时代, 然而我国至今尚未形成公共数据库存储体系。为弥补国内空白, 组学原始数据归档库(Genome Sequence Archive, GSA, <http://bigd.big.ac.cn/gsa>)系统遵循国际核苷酸序列数据联盟(International Nucleotide Sequence Database Collaboration, INSDC)相关数据库建设标准, 广泛收集各类生命组学原始数据。自 2015 年底上线运行以来, 已获得了包括 *Cell*、*Nature*、*PNAS*、*GPB* 等 30 余个国内外期刊的认可, 收录的数据量呈显著增长趋势, 提供的数据服务受到国内外广大科研人员的认可。GSA 有效缓解了当前我国生命组学数据汇交、存储与共享困难的问题, 为我国国家生物信息中心的建设奠定了坚实基础。本文对目前 GSA 数据汇交、审核、发布与管理等机制进行了深入阐述, 以方便用户了解 GSA 的各项功能, 提供更高效的数据服务。

关键词: 组学原始数据归档库 (GSA); 组学大数据; 数据汇交; 数据共享

GSA: Genome Sequence Archive

Sisi Zhang^{1,2}, Tingting Chen^{1,2}, Junwei Zhu^{1,2}, Qing Zhou^{1,3}, Xu Chen^{1,2},
Yanqing Wang^{1,2}, Wenming Zhao^{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: The Genome Sequence Archive (GSA), a new data repository for raw sequence reads in China, has been

收稿日期: 2018-06-29; 修回日期: 2018-09-06

基金项目: 国家重点研发计划“国家生物信息平台支撑技术项目”和“精准医学项目”(编号: 2017YFC1201200, 2016YFC0901603), 中国科学院战略性先导科技专项基金项目(编号: XDB13040500, XDA08020102), 国家自然科学基金项目(编号: 91731304), 中国科学院关键技术人才基金项目和中国科学院“十三五”信息化建设专项(编号: XXH13505-05)资助[Supported by the National Key R&D Program of China (Nos. 2017YFC1201200, 2016YFC0901603), the Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDB13040500, XDA08020102), the National Natural Science Foundation of China (No. 91731304), Key Technology Talent Program of the Chinese Academy of Sciences and the 13th Five-year Informatization Plan of Chinese Academy of Sciences (No. XXH13505-05)]

作者简介: 张思思, 博士, 工程师, 研究方向: 生物信息学。E-mail: zhangss@big.ac.cn

陈婷婷, 硕士, 工程师, 研究方向: 生物信息学。E-mail: chentt@big.ac.cn

张思思和陈婷婷并列第一作者。

通讯作者: 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.yczs.18-178

网络出版时间: 2018/9/28 11:32:26

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20180928.1132.003.html>

developed in compliance with the International Nucleotide Sequence Database Collaboration (INSDC) standards. It supports data generated from a variety of sequencing platforms ranging from Sanger sequencing to single-cell sequencing and provides data storing and sharing services freely for worldwide scientific communities. Since it went online in late 2015, GSA has archived more than 500 TB data and been acknowledged by many high-profile journals, including *Cell*, *Nature*, *PNAS*, *GPB*, etc. Focusing on omics data submission, storing and sharing typically for Chinese users, GSA promotes the initiative of the National Bioinformatics Center of China. This paper introduces the specifics of GSA as data collection, curation, management and exchange to facilitate users to understand and use GSA database.

Keywords: Genome Sequence Archive (GSA); omics data; data submission; data sharing

基因组测序技术的迅速发展,使生命科学研究产生的组学数据呈现爆炸性增长的态势,生命科学亦进入大数据时代。“数据堪比石油”,是国家重要战略资源,生物数据尤为如此。未来,生物数据的存储、管理、共享和开发利用将在一定程度上反应国家科技发展实力和水平。

国际上,美国、欧洲和日本于 20 世纪 80 年代相继建立核酸序列数据库^[1~3],并于 2002 年成立了国际核酸序列共享联盟(International Nucleotide Sequence Database Collaboration, INSDC)^[4],制定了生命科学研究领域数据管理和共享标准,收集并存储来自全世界科学家提交的组学数据,提供共享服务。随着数据量的持续增加和学术论文发表的数据共享要求,大量组学数据通过国际互联网递交到 INSDC 变得十分困难。我国国际网络出口带宽的瓶颈问题,更使得数据传输效率低下。以中国科学院北京基因组研究所 150 Mbs 出口带宽为例,向美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)数据库递交 1 TB 的数据需要花费两周甚至更长的时间。数据下载亦是如此,国内科研人员饱受从国际数据库下载数据效率低下的困扰。这种低下的数据传输效率以及对国际数据库使用的不便,使得我国生物学家在一定程度上失去了生命科学研究领域的竞争力。

随着国家一系列重大生命科学研究计划的部署和实施,预期我国每年将产生超过 100 PB 的组学数据。为解决这些海量数据存储、管理、共享与发布,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)^[5,6]研发并构建了组学原始数据归档系统(Genome Sequence Archive, GSA)^[7],

专注于生命组学原始数据收集与管理,并提供免费的数据存储、共享与访问服务。GSA 遵循国际 INSDC 的数据标准及数据库建设标准,可收集来自不同测序平台产出的数据,并存储序列数据及其对应的元数据信息,确保数据的完整性。本文主要从数据类别与使用、运行与效果等方面,全面解析 GSA 数据库在数据汇交、存储及共享各项功能,以便更高效为用户提供数据服务。

1 GSA 的数据类别与使用

1.1 数据库建设

GSA 系统建设采用数据分类存储、集中管理的指导思想及高内聚、低耦合的程序设计原则,既确保各类数据的独立性及关联性,又保证整体系统的完整性和有效性,避免信息孤岛的存在,可实现任一数据的向上回溯及向下追踪。所谓数据分类是根据表现形式将数据分为元数据信息和测序数据文件,集中管理则指通过内在的关联信息,对数据进行关联显示、调用与管理。

GSA 数据类别主要包括项目信息(BioProject)、样本信息(BioSample)、实验信息(Experiment)、测序反应(Run)信息以及测序数据文件。项目信息是用来描述所开展研究的目的、涉及物种、数据类型、研究思路等信息;样本信息是指本研究涉及的生物样本描述,如样本类型、样本属性等;实验信息包括实验目的、文库构建方式、测序类型等信息;测序反应信息包括测序文件和对应的校验信息。各类数据之间采用线性、一对多的模式进行关联,从而形成“金字塔”式的信息组织与管理模式。

为确保与国际同类数据库系统的兼容性, GSA 遵循 INSDC 联盟的数据标准和命名规范。如 Bio-Project 序列号(accession number)以 PRJC 开头, 前 3 位字母 PRJ 为 Project 的缩写, 第 4 位字母 C 代表中国(China), 第 5 位是英文字母 A~Z, 其余为 6 位自然数, 例如 PRJCA000001; BioSample 序列号以 SAMC 开头, 前 3 位字母 SAM 为 Sample 的缩写, 第 4 位字母 C 代表中国(China), 第 5 位是英文字母 A~Z, 其余为 6 位自然数, 例如 SAMCA000001。其他数据类型的编码遵循相同标准, 既确保数据编码的全局性与唯一性, 又便于后续数据使用者的信息检索与访问。

1.2 数据汇交与审核

2017 年 6 月, 数据递交系统(BIG Sub)正式上线, 作为生命与健康大数据中心数据统一汇交入口, BIG Sub 承载 GSA 系统的数据汇交功能, 并为用户提供一站式数据递交服务。在元数据信息汇交方面, BIG Sub 提供两种数据递交方式: 在线递交和离线递交。在线递交即通过 WEB 页面实现信息输入, GSA 系统提供可视化及向导化的操作模式, 最大限度地规范信息录入并实现各类数据的质量控制; 离线递交即采用离线模板的形式, 由科研人员事先整理文件, 然后通过 GSA 系统或 GSA 审编人员进行数据批量导入。在线递交较适合小量样本的数据递交(如样本数小于 10 个), 而离线递交适合大量样本的数据递交, 这两种互为补充的提交方式为科研人员提供简单、便利、高效的数据递交服务。在测序序列文件汇交方面, 提供在线 FTP 上传服务和邮寄硬盘两种服务模式, 如超过 500 GB 的序列文件, 数据递交者可以选择采用邮递硬盘的模式, 由 GSA 系统审编人员协助上传数据。针对 FTP 数据上传服务, BIG Sub 为每位数据递交者提供独立的数据存储空间, 以防止不同递交者之间的数据干扰及信息泄露, 充分确保数据的隐私性和安全性。

GSA 系统具有元数据审核和序列数据质量控制功能。针对元数据信息, 采用自动校验和人工校验相结合的模式进行审核, 以保证信息的有效性。而对于测序序列数据, GSA 内置数据质量审核的标准化流程, 以防止序列文件在处理、传输、压缩、拷

贝等过程中出现损坏。审核内容包括: (1) 文件压缩的正确性; (2) 文件格式的正确性; (3) 序列的测序质量。针对某一数据递交, 只有当元数据和序列数据均审核通过, GSA 系统方可为该数据分配正式的访问序列号(accession number)。

1.3 数据发布与管理

GSA 系统提供两种数据状态控制权限: 公开访问(public)或受控访问(confidential), 公开即意味着数据可被任何人访问或下载使用, 受控即在数据公开发布前, 他人的访问将被限制, 且无法通过系统检索获取相关的信息, 更无法下载相关数据文件。同时, GSA 系统提供个性化的数据状态及权限管理方案, 即由数据递交者自行设置数据受控保护期限, 最大限度的满足论文发表前的数据保密需求, 亦可方便论文审稿人对数据在线访问与审核(peer reviewer link), 还可快速实现文章发表后的数据发布与共享。

2 GSA 的运行效果

GSA 系统自 2015 年上线运行以来, 获得包括 *Cell*、*Nature*、*PNAS*、*AJHG*、*GPB*、*Cell Research* 等在内的 30 余个国内外期刊的认可, 支持 40 余篇科研论文的数据归档与发布任务。截止 2018 年 7 月, GSA 接收的数据来自国内外 93 个机构的 300 余名科研用户, 累计递交项目信息达 535 个, 涵盖的生物物种数量超过 178 个, 涉及的生物学样本 21 843 个, 生物学实验 28 050 个, 测序反应 29 624 个, 测序序列数据总量超过 TB, 且各类数据呈现显著增长的趋势(图 1)。同时, GSA 系统收录的数据受到国内外科研人员的广泛关注, 经统计发现, GSA 系统访问用户来自于 70 余个国家/地区, 累计访问量超过 13 305 人次。数据下载用户来自 39 个国家/地区, 日平均上传下载量达到 1 TB。

GSA 系统不仅为国内科研人员提供了良好的数据存储和共享服务, 同时也为国家的重大科研计划提供了良好的数据管理和支撑平台。如为国家科技部重点研发计划“精准医学”专项提供数据集中存储和管理保障, 为中国科学院战略先导专项“分子模块设计育种创新体系”和“动物复杂性状的进化

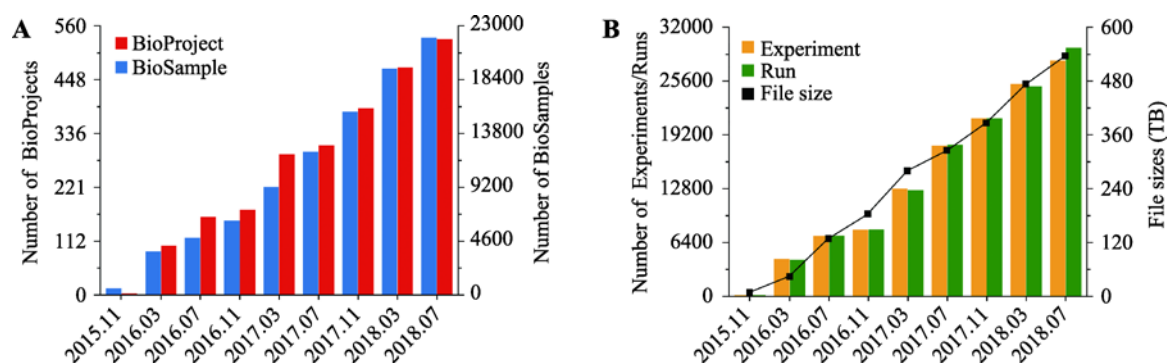


图 1 GSA 数据量统计图

Fig. 1 Statistics of data submissions of GSA

A: GSA 数据库中项目和样本数量统计图; B: GSA 数据库中实验、测序反应和文件数量统计图。数据统计截至 2018 年 7 月。

解析与调控”等项目提供数据集中存储、统一管理和共享服务,并取得了较好的应用效果。

3 结语与展望

GSA 是一个公共的、免费的组学原始数据存储库,在建设标准上遵循国际 INSDC 联盟的数据标准和数据库结构标准,在内容上收集生命科学研究中产生的组学测序数据及其元数据信息,并且接受来自全世界科研人员的数据递交与获取请求。在组学大数据时代,GSA 不仅作为当前 INSDC 联盟体系的补充以缓解组学大数据远距离传输与储存的压力,而且承担推动国际组学大数据共享的责任。

GSA 致力于我国组学数据汇交、管理、共享与应用体系的建设,促进我国在生命组学大数据领域的发展,力争我国在国际组学数据共享领域的地位,为国内外生命科学研究与产业创新应用提供服务。2017 年,中国科学院北京基因组研究所发起“中国基因组数据共享倡议”(http://bigd.big.ac.cn/gdsd),呼吁中国产出的组学数据递交 GSA 进行统一存储、管理与共享。截止 2018 年 6 月,得到国内 570 多个机构 1400 余人支持。

立足现在、着眼未来,GSA 将不断完善系统的功能,更加重视数据资源使用者的需求,开发类似 fastq-dump (如 SRA toolkit)的辅助数据下载,实现数据便捷共享。此外,还将开发基于数据分析的云计算平台,提供免费数据在线分析服务,届时,用户可以不用下载数据便可利用云计算资源进行数据分

析。顺应国家大数据发展战略及科技创新和产业发展需求,存好、管好国家生物数据资源,推动“国家生物信息中心”的建立。

参考文献(References):

- [1] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2018, 46(Database issue): D8–D13. [DOI]
- [2] Silvester N, Alako B, Amid C, Cerdeño-Tarraga A, Clarke L, Cleland I, Harrison PW, Jayatilaka S, Kay S, Keane T, Leinonen R, Liu X, Martínez-Villacorta J, Menchi M, Reddy K, Pakseresht N, Rajan J, Rossello M, Smirnov D, Toribio AL, Vaughan D, Zalunin V, Cochrane G. The European Nucleotide Archive in 2017. *Nucleic Acids Res*, 2018, 46(Database issue): D36–D40. [DOI]
- [3] Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T. DNA Data Bank of Japan (DDBJ) progress report. *Nucleic Acids Res*, 2016, 44(Database issue): D51–57. [DOI]
- [4] Karsch-Mizrachi I, Takagi T, Cochrane G, International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 2018, 46(Database issue): D48–D51. [DOI]
- [5] BIG Data Center Members. The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res*, 2017, 45(Database issue): D18–D24. [DOI]
- [6] BIG Data Center Members. Database resources of the BIG Data Center in 2018. *Nucleic Acids Res*, 2018, 46(Database issue): D14–D20. [DOI]
- [7] 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 Bioinform*, 2017, 15(1): 14–18. [DOI]