微生物学报
Acta Microbiologica Sinica
2021, 61(12): 3761–3773
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20210691

国家微生物科学数据中心数据资源服务与应用

范国梅，孙清岚，史文聿，亓合媛，孙定中，李芳慧，庞慧芳，马俊才*，吴林寰*

中国科学院微生物研究所，北京 100101

摘要：作为解决生命领域复杂科学问题的关键要素以及驱动科学发现与决策的基础资源，微生物科学数据资源已成为国家的重要战略资源。国家微生物科学数据中心(https://nmdc.cn/)的建设使得海量微生物数据资源可以得到有效的整理整合和开放共享，这对于微生物资源的研究、利用和可持续发展都起着至关重要的作用。本文从核心资源、服务内容、功能特色等多方面总结了国家微生物科学数据中心平台的建设进展，并提出了面向微生物领域科研及产业用户的应用实践。

关键词：微生物，数据中心，数据共享，数据库

近年来，随着基因组学和新一代测序技术的迅猛发展，生命科学领域的研究逐渐呈现出“大样本、大数据、大平台、大合作”的特点，引领数据出现爆发式增长，同时也对生物数据的整合、管理、共享及安全等问题提出了新的需求[1–4]。而满足这些需求的第一步是建立起能够大范围收集数据并支撑其应用的数据中心。早在 20 世纪 80–90 年代，美国、欧洲和日本等国就相继成立了国家级生物信息中心—美国国家生物技术信息中心(National Center for Biotechnology Information，NCBI)、欧洲生物信息研究所(European Bioinformatics Institute，EBI)和日本 DNA 数据库(DNA Data Bank of Japan，DDBJ)，现已成为国际生物数据获取、存储等方面的权威机构[5]。为了加强我国科学数据资源的合规管理和开放共享，2019 年 6 月，我国科技部、财政部也启动了一批包括国家微生物科学数据中心(以下简称“中心”)在内的国家科学数据中心的建设。国家微生物科学数据中心围绕海量的微生物数据资源，进行数据的有效整合、利用、存储和共享，建立了完善的数据管理和服务体系，从而推动我国微生物领域科学数据的共享与应用，并力争推动中国在微生物资源的开发应用和数据共享方面占领国际微生物研究前沿和主导地位。

基金项目：国家微生物科学数据中心项目；中国科学院微生物科学数据中心能力建设(XXH-13514-0203)；国家科技基础条件平台中心(2020WT11)；模式微生物基因组测序多组学数据整合的功能挖掘研究(153211KYSB20190021)；世界微生物数据中心秘书处建设方案项目

*通信作者。E-mail：马俊才，ma@im.ac.cn；吴林寰，wulh@im.ac.cn

收稿日期：2021-11-10；**修回日期：**2021-11-25；**网络出版日期：**2021-11-26

1 系统资源建设

国家微生物科学数据中心(http://nmdc.cn/)建立了完善的微生物领域数据体系(图 1)。数据内容覆盖微生物资源、研究过程及工程、微生物组学、微生物技术、合成生物学等学科方向以及微生物文献、专利、专家、成果等知识库，旨在重点推进微生物领域科技资源向国家平台汇聚与整合，并将海量的微生物科学数据进行多层次数据划分和专业化类别归属，同时给科研用户提供便利的数据提交渠道和高速的下载通道，为科学研究、技术进步和社会发展提供高质量的科学数据资源共享服务。目前，平台已汇聚资源总量超过 3PB，数据记录数超过 40 亿条，重点数据库 232 个，年访问量逾 2000 万人次，其中，数据来源覆盖中国科学院及国内其他科研院所、高校、企业等百余家单位，此外，中心还接收了全球 50 个国家 146 个单位的数据汇交和全球共享。与微生物领域国际著名数据库如美国的微生物基因组数据库 IMG/M (单菌序列数据为 364 万条[6])，以流感病毒数据为核心的全球共享流感数据倡议组织 GISAID (数据量超过 65 万条[7])相比，无论是数据量还是数据体系的全面性已经处于同

图 1. 国家微生物科学数据中心主页
Figure 1. The homepage of NMDC.

一水平。中心已经逐步成为全球微生物领域最重要的数据中心。海量的数据资源和多元化的数据类型是中心平台重要特色之一。

中心平台网站主要涵盖数据资源、元数据、数据下载、数据汇交、分析工具、服务案例、标准规范等多样化内容栏目，支持高级检索、智能排序，并且突出了数据资源、数据汇交和数据分析云平台等核心资源(图 2)。

目前，中心采用“自建+提交+外采”的数据整合方式，对不同来源、不同类型的数据资源进行了统一规划、整合加工、深度挖掘分析，针对顶层应用进行专业化数据分类整理，构建了一套完善的数据整合体系，为用户在庞大的数据资源体系内提供了一页即可浏览以及可进行全量数据资源搜索的展示页面，用户可通过一键搜索快速精准定位资源。

图 2. 国家微生物科学数据中心数据资源页面

Figure 2. Overview of NMDC’s data resources.

2 科学数据的汇交和发布

中心面向各领域广大科研用户提供一站式在线数据汇交服务，支持科研用户全程在线填报信息、提交数据，有效助力科学数据的快速发布及全球共享使用。数据汇交主要收集两大类科学数据，一类是用于文章发表的数据，一类是科研专项的数据。

依据国际通用的元数据标准与流程，中心建立了一套数据汇交体系(图 3)。用于文章发表的数据提交可以支持生物项目数据、生物样本数据、核酸序列数据、原始组学数据、基因组数据、宏基因组数据、晶体结构数据和期刊附件数据的汇交。中心会在数据提交后 3 个工作日内，完成审核并发放数据编号，该数据编号可在国际及国内期刊中直接使用，可支撑国内外科学家发表的文章数据的存储、共享。

按照国家《科学数据管理办法》的要求，公共财政经费支持的科研项目所形成的科学数据，在项目绩效评估前，需要将该数据汇交到国家科学数据中心。因此，目前中心配合国家重点研发计划、科技基础资源调查专项等项目开展数据汇交的工作，已经支持了生物安全关键技术研发、食品安全关键技术研发、合成生物学、绿色先进制造、公共安全风险防控与应急技术装备等多个专项的科学数据汇交。

图 3. 国家微生物科学数据中心数据汇交流程

Figure 3. Overview of data submission process in NMDC.

按照数据共享的要求，中心对汇交的科学数据进行系统性、科学性的分级分类、加工整理及开放共享，对多源异构的各类型专项数据进行深度精细化加工及元数据抽取，并根据数据共享方式提供在线的下载、浏览或者检索等服务(图 4)。

3 特色专题数据库

通过在微生物领域的完整的数据资源体系建设和数据服务体系的布局，中心开展了大量的技术创新和服务整合，配合重大项目、重要国际合作计划、重点研究方向等有针对性地建立了一系列特色专题数据库。同时，中心作为世界微生物数据中心(The World Data Centre for Microorganisms，WDCM)的承担单位，是国际微生物领域数据注册和管理的权威机构，与国内外各微生物资源保藏机构和主要数据中心建立了良好的合作关系，形成了稳定的国内外数据汇交渠道，因此，目前中心所形成的特色数据资源在微生物领域具有较大的国际、国内影响力。

3.1 全球微生物保藏机构数据库

由中心建立的全球微生物保藏机构数据库(Culture Collections Information Worldwide，CCINFO)[8] (http://ccinfo.wdcm.org/)(图 5)是全球所有保藏机构的注册中心，也是下属机构对自身资源进行统一管理和国际合作的平台。所有的保藏中心通过注册并且填报详细的元数据信息，通过 WDCM 的审核及在线发布数据后，保藏中心才可提供符合国际标准的对外共享资源和服务。

图 4. 国家微生物科学数据中心专项数据展示页面

Figure 4. Overview of project-specific data in NMDC.

图 5. CCINFO 数据库

Figure 5. The homepage of Culture Collection Information database.

该数据库是世界微生物保藏联合会的官方数据平台，截至 2021 年 10 月已经有 78 个国家的 801 个微生物资源保藏机构在此注册，其中，中国的微生物、藻类、细胞资源库馆共 48 个。

3.2 全球微生物资源目录数据

中心建立的全球微生物菌种保藏目录(Global Catalogue of Microorganisms，GCM)[9](http://gcm.wdcm.org/)(图 6)旨在为分散于全球各个保藏中心和科学家手中的宝贵微生物资源提供一个全球统一的数据仓库，并以统一数据门户的形式，对全世界科技界和产业界提供微生物菌种资源的信息服务。截至 2021 年 10 月已经有来自美国、法国、德国、荷兰等 50 个国家和地区的 146 个国际微生物资源保藏机构正式参加这一计划。同时中心也与亚洲微生物资源保藏联盟、亚洲生物资源网络、欧洲微生物资源中心联盟等区域性网络和俄罗斯、泰国、葡萄牙等国家网络建立了实质性合作。到目前为止，GCM 已经整合了超过 47 万的微生物实物资源的详细信息，其中不乏来自特殊生态环境、具有重要的科研和工业应用价值的微生物。作为一个微生物数字资源整合的大数据平台，GCM 还利用先进的数据挖掘手段，从全球超过 600 万已发表的微生物相关文献及专利中，进一步提取了可供后续研究和利用的微生物资源信息。因此，该信息平台对微生物实物资源的采集、保藏、跨国转移、学术和商业应用以及利益分享的各个环节都能提供有效的数据支持，为《生物多样性公约》在微生物领域的实施和执行提供了重要支撑(https://www.cbd.int/abs/doc/protocol/icnp-1/wfcc-en.pdf)。

图 6. GCM 数据库

Figure 6. The homepage of the Global Catalogue of Microorganisms.

3.3 模式微生物基因组数据库

由中心建立的模式微生物基因组数据库(Global Catalogue of Type Strain，gcType)[10] (http://gctype.wdcm.org/)(图 7)是目前国际上在模式微生物基因组方面数据资源最为全面、功能最为完善的数据平台，截至 2021 年 9 月，平台整合了全部有效发表的模式微生物菌株资源 67000 余株，物种 18000 种，基因组 14000 余个。平台不仅集成了目前所有公共来源的模式微生物物种和基因组数据，还发布了大量自测模式微生物基因组数据，并且集合了数据搜索下载，新种鉴定，基因组拼接与注释等在线分析工具，是目前国内外模式微生物基因组数据最丰富的平台，为分类学家进行基因组研究、新种鉴定提供了一个重要途径。平台在物种鉴定模块还可以计算平均核苷酸一致性(Average Nucleotide Identity，ANI)等用于微生物分类鉴定的重要参数，可直接用于 *International Journal of Systematic and Evolutionary Microbiology* 上微生物新种的文章发表。与国际上其它的类似数据库相比[11−12]，gcType 的优势在于紧密依托 WDCM 的菌种资源和 10K 测序计划，其中超过 1000 种模式菌的基因组为项目成果，比之前的数据更加准确。

3.4 全球微生物组数据库

中心建立的全球微生物组数据平台(Global Catalogue of Metagenomics，gcMeta)[13] (https://gcmeta.wdcm.org/) (图 8)，集成了目前国际主要微生物组(人体微生物组、地球微生物组)项目以及中国科学院微生物组计划和国内重点研发等项目产生的微生物组数据。截至 2021 年 10 月，总集成数据量超过 150TB，总数据量超过 50 万条，其中，国内自有数据超过 2000 个样本。目前已经覆盖本领域国内外超过 80%的公开数据。平台充分采用了国际通用的微生物基因组及宏

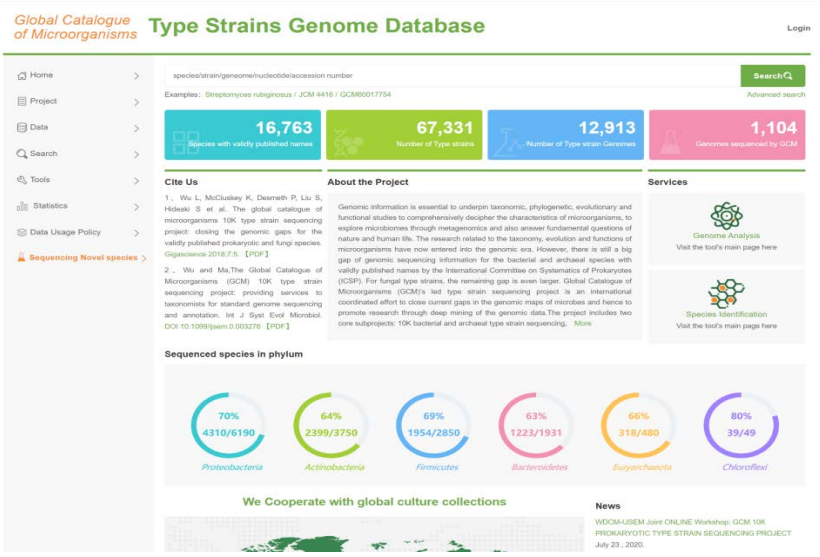

图 7. gcType 数据库

Figure 7. The homepage of gcType database.

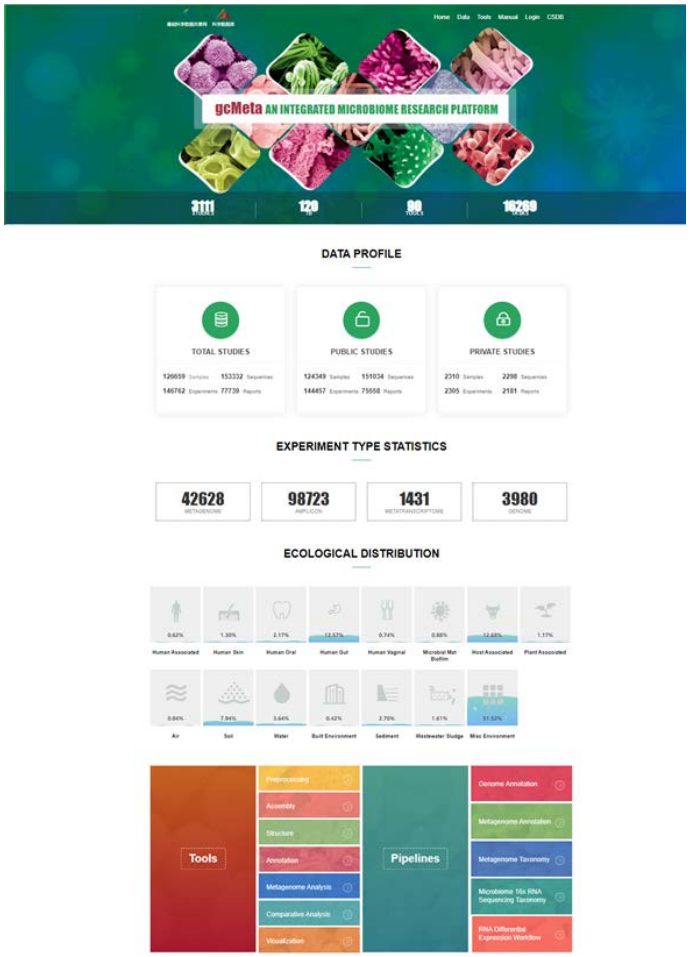

图 8. gcMeta 数据库

Figure 8. The homepage of gcMeta database.

基因组数据标准(MIGS/MIMS)[14]，并且使用了国际通用的环境信息概念系(ENVO)对元数据进行了标准化[15]，从而确保本数据库能与国内外相关数据进行无缝对接。

该数据库不仅提供了数据存储、数据管理和数据发布的一条龙服务，还能利用平台提供的资源对数据进行在线分析、可视化并生成分析报告。目前，该数据库已经逐步发展成国内外微生物领域最具权威和影响力的数据库，将持续接收国内外相关数据并提供长期服务。

3.5 新冠病毒虚拟突变与预警数据库

随着全球新冠疫控的持续进行，新型冠状病毒基因组也在流行过程中持续发生变异[16–17]。迄今，在全球科学技术人员的共同努力下，已经对超过 400 万例病毒基因组进行了测序，并构建了多个病毒基因组数据库[18]。然而，随着对变异研究的深入，变异造成的功能影响日渐成为科学家关注的焦点。目前，在全球多个国家和地区均发现了包括 Alpha、Beta、Delta 和 Omicron 在内的多种感染力增强的变异毒株，尤其是关键位点积累的氨基酸变异，极大地改变了病毒的免疫学特征，增加了病毒免疫逃逸的风险，这可能会降低现有疫苗、抗体、药物等疫情控制方法的保护性，影响核酸诊断试剂的适用性，因而对疫情防控构成了严峻挑战[19–21]。因此，现有的以收集、展示数据为主的基本数据库已经难以满足未来疫情防控的需求，亟需一个基于大数据的病毒变异风险评估及预警系统，对现有及未来可能出现的各种变异造成的影响进行系统性评估和解读，从而实施更加精准有效的疫情防控策略。

为此，中心建立了新冠病毒虚拟突变与预警数据库[22] (https://nmdc.cn/ncovn/)，该数据库从基因组学和结构生物学角度入手，在基于变异位点频率评估的基础上，从核苷酸变异发生难易程度、氨基酸替换难度、变异对蛋白质二级结构的影响、单个氨基酸突变引起的 ACE2 及中和抗体结合自由能变化等参数对变异进行多维度的评估，全面综合分析已知变异和潜在的虚拟变异对病毒功能造成的影响。在此基础上，该系统采用了人工智能分类器算法，从传播性和中和抗体亲和力两方面对变异株进行有效分组，实现了基于病毒序列的风险评估和预警。基于虚拟变异和风险评估模型，该系统可以作为全球病毒变异监测和追踪的工具，为针对新型变异毒株的精准防控和抗体疫苗设计提供有效的参考信息。目前基于该系统的分析结果为精准高效应对 SARS-CoV-2 突发疫情提供了重要的决策依据，同时也为应对其他突发传染性公共卫生事件提供了技术储备。

3.6 数据分析云平台

通过整合超级计算和分布式系统等基础设施，中心利用 Docker 技术构建了微生物交互式应用分析云平台，可提供生物信息在线分析工具、计算资源、公共参考数据的整合服务(图 9、图 10)。用户无需自己编写代码，无需配置 Linux 操作系统，无需安装复杂的生物信息分析软件和下载庞大的生物数据库，即可利用线上分析工具进行微生物数据的分析。目前，中心的分析工具模块集成了微生物领域常用的七大类共 88 种线上分析工具，包括宏基因组分析流程、基因组拼接工具、基因组结构分析、基因组注释分析、元基因组分析、比较基因组分析、便捷分析工具。

4 总结、建议和展望

目前，国家微生物科学数据中心已经初步形

成了海量微生物数据资源的整合与安全共享的技术体系，为我国的科学数据管理与共享实践打下了扎实的基础。但是，在数据管理与共享过程中的数据采集、加工、保存、质量控制等各个环节的标准规范还有待进一步完善，科研数据的学术贡献和评价体系还相对缺乏[23]。随着科学数据汇交的数据量逐渐增多，用户的个性化需求不断增加，平台需要及时改善和优化系统的功能和性能，加强与交叉学科的科学数据平台的融汇贯通，加强与行业科学数据平台的共享合作。在进

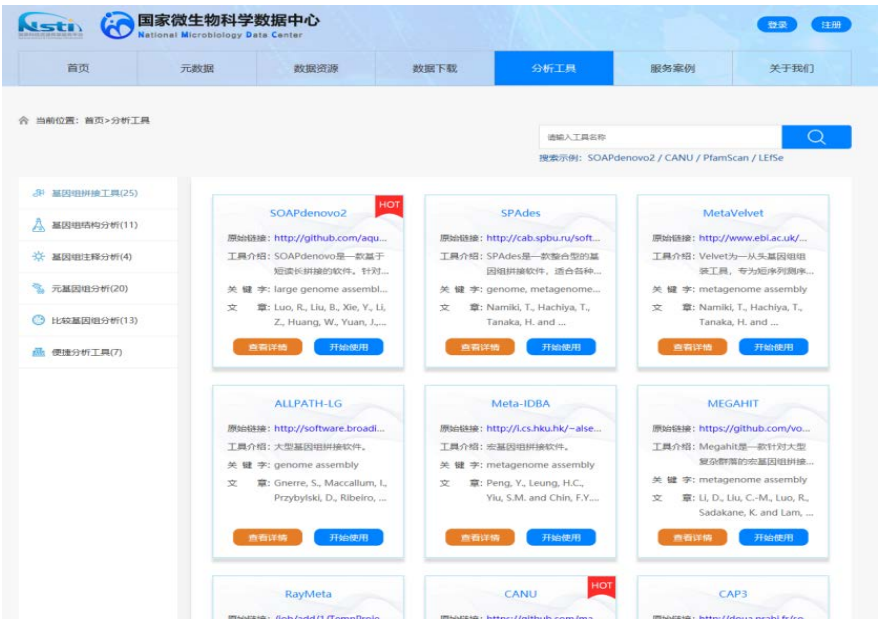

图 9. 国家科学数据中心分析工具页面

Figure 9. Analytic tools in NMDC.

图 10. 微生物云平台软件工具集

Figure 10. Tools and software available on the cloud.

一步细化科学数据分级分类的同时，还需要提升数据资源深度挖掘和深层次加工的能力，促进多类型资源融合，利用机器学习、人工智能、大数据分析多种手段，增强数据延展性，在保障数据和系统的安全的同时，更好地为用户服务。

另一方面，2021 年，《国家生物安全法》和《数据安全法》都已相继出台，《网络数据安全管理条例(征求意见稿)》也对信息保护法、数据安全法等上位法的相关规定在网络数据安全应用中进行了细化和补充。以微生物数据为核心的生物安全数据，是国家生物安全体系建立的重要组织部分，因此，中心一方面将依据国家相关法律法规，建立系统、完善的微生物数据安全分级分类的框架体系，另一方面将加强数据安全技术研究和应用，在保障数据安全的情况下，持续推动科学数据的共享。

参 考 文 献

[1] Stein LD. Integrating biological databases. *Nature Reviews Genetics*, 2003, 4(5): 337–345.

[2] Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. *Journal of Biological Research: Thessalonike, Greece*, 2015, 22(1): 9.

[3] Green ED, Rubin EM, Olson MV. The future of DNA sequencing. *Nature*, 2017, 550(7675): 179–181.

[4] Navarro FCP, Mohsen H, Yan CF, Li ST, Gu MT, Meyerson W, Gerstein M. Genomics and data science: an application within an umbrella. *Genome Biology*, 2019, 20(1): 109.

[5] Stevens H. Globalizing genomics: the origins of the international nucleotide sequence database collaboration. *Journal of the History of Biology*, 2018, 51(4): 657–691.

[6] Chen IMA, Chu K, Palaniappan K, Ratner A, Huang JH, Huntemann M, Hajek P, Ritter S, Varghese N, Seshadri R, Roux S, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Research*, 2021, 49(D1): D751–D763.

[7] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 2017, 1(1): 33–46.

[8] Wu LH, Sun QL, Desmeth P, Sugawara H, Xu ZH, McCluskey K, Smith D, Alexander V, Lima N, Ohkuma M, Robert V, Zhou YG, Li JH, Fan GM, Ingsriswang S, Ozerskaya S, Ma JC. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. *Nucleic Acids Research*, 2017, 45(D1): D611–D618.

[9] Wu LH, Sun QL, Sugawara H, Yang S, Zhou YG, McCluskey K, Vasilenko A, Suzuki KI, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Philippe D, Ma JC. Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. *BMC Genomics*, 2013, 14: 933.

[10] Shi W, Sun QL, Fan GM, Hideaki S, Moriya O, Itoh T, Zhou YG, Cai M, Kim SG, Lee JS, Sedlacek I, Arahal DR, Lucena T, Kawasaki H, Evtushenko L, Weir BS, Alexander S, Dénes D, Tanasupawat S, Eurwilaichitr L, Ingsriswang S, Gomez-Gil B, Hazbón MH, Riojas MA, Suwannachart C, Yao S, Vandamme P, Peng F, Chen ZH, Liu DM, Sun XQ, Zhang XJ, Zhou YC, Meng Z, Wu LH, Ma JC. gcType: a high-quality type strain genome database for microbial phylogenetic and functional research. *Nucleic Acids Research*, 2021, 49(D1): D694–D705.

[11] Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, Kandimalla M, Chen IMA, Kyrpides NC, Reddy TBK. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research*, 2021, 49(D1): D723–D733.

[12] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao YM, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li WJ, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun HZ, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi

A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 2016, 44(D1): D733–D745.

[13] Shi W, Qi HY, Sun QL, Fan GM, Liu SJ, Wang J, Zhu BL, Liu HW, Zhao FQ, Wang XC, Hu XX, Li W, Liu J, Tian Y, Wu LH, Ma JC. gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Research*, 2019, 47(D1): D637–D648.

[14] Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 2008, 26(5): 541–547.

[15] Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 2016, 7(1): 57.

[16] Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, Mlisana K, von Gottberg A, Walaza S, Allam M, Ismail A, Mohale T, Glass AJ, Engelbrecht S, Van Zyl G, Preiser W, Petruccione F, Sigal A, Hardie D, Marais G, Hsiao NY, Korsman S, Davies MA, Tyers L, Mudau I, York D, Maslo C, Goedhals D, Abrahams S, Laguda-Akingba O, Alisoltani-Dehkordi A, Godzik A, Wibmer CK, Sewell BT, Lourenço J, Alcantara LCJ, Kosakovsky Pond SL, Weaver S, Martin D, Lessells RJ, Bhiman JN, Williamson C, de Oliveira T. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 2021, 592(7854): 438–443.

[17] Kupferschmidt K. Evolving threat. *Science*, 2021, 373(6557): 844–849.

[18] Rigden DJ, Fernández XM. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 2021, 49(D1): D1–D9.

[19] Rella SA, Kulikova YA, Dermitzakis ET, Kondrashov FA. Rates of SARS-CoV-2 transmission and vaccination impact the fate of vaccine-resistant strains. *Scientific Reports*, 2021, 11: 15729.

[20] Hoffmann M, Arora P, Groß R, Seidel A, Hörnich BF, Hahn AS, Krüger N, Graichen L, Hofmann-Winkler H, Kempf A, Winkler MS, Schulz S, Jäck HM, Jahrsdörfer B, Schrezenmeier H, Müller M, Kleger A, Münch J, Pöhlmann S. SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell*, 2021, 184(9): 2384–2393.e12.

[21] McCarthy KR, Rennick LJ, Nambulli S, Robinson-Mccarthy LR, Bain WG, Haidar G, Duprex WP. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science*, 2021, 371(6534): 1139–1142.

[22] Sun QL, Shu C, Shi W, Luo YF, Fan GM, Nie JY, Bi YH, Wang QH, Qi JX, Lu J, Zhou YC, Shen ZH, Meng Z, Zhang XJ, Yu ZF, Gao SH, Wu LH, Ma JC, Hu SN. VarEPS: an evaluation and prewarning system of known and virtual variations of SARS-CoV-2 genomes. *Nucleic Acids Research*, doi(10.1093): nar.

[23] Tu ZF, Yang ZP. Practice of scientific data management and sharing in China: focusing on two models. *Documentation, Information & Knowledge*, 2021(1): 103–112. (in Chinese)
涂志芳, 杨志萍. 我国科学数据管理与共享实践进展: 聚焦两种主要模式. 图书情报知识, 2021(1): 103–112.

The services and applications of national microbiology data center

Guomei Fan, Qinglan Sun, Wenyu Shi, Heyuan Qi, Dingzhong Sun, Fanghui Li, Huifang Pang, Juncai Ma*, Linhuan Wu*

Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

Abstract: Microbiology data centers has become an essential part of strategic resources of a country. The National Microbiology Data Center (NMDC, https://nmdc.cn/) allows a huge amount of microbiological data to be organized and integrated in an effective way, and shared in an open manner, which is crucial for the research, utilization, and sustainable development of microbiological resources. This paper summarizes the progress of the recently founded NMDC platform, with respect to core resources, services, and functions, in order to foster the applications and practices of microbiological knowledge in academic and industrial users.

Keywords: microbiology, data center, data sharing, database

（本文责编：李磊）

Supported by the National Microbiology Data Center, by the Capacity Building of the Microbiology Data Center of Chinese Academy of Sciences (XXH-13514-0203), by the National Science and Technology Infrastructure Platform Center (2020WT11), by the Global Cooperation Project of Type Microorganisms Genome Sequencing, Data Mining and Functional Analysis (153211KYSB20190021) and by the World Data Centre for Microorganisms Secretariat Construction Plan

*Corresponding authors. E-mail: Juncai Ma, ma@im.ac.cn; Linhuan Wu, wulh@im.ac.cn

Received: 10 November 2021; Revised: 25 November 2021; Published online: 26 November 2021

吴林寰，中国科学院微生物研究所正高级工程师，国家微生物科学数据中心副主任。长期从事微生物大数据集成和挖掘的方法研究和系统构建。在《核酸研究》、*Nature Communication*、*GigaScience*、*BMC gemonics* 等杂志发表文章 30 余篇，申请软件著作权 40 余项，承担了国家“863 计划”、国家重点研发计划、中国科学院先导专项应急计划等多个项目。2017 年，获得 World Data System (WDS) Data Stewardship Award。近年来，带领国家(世界)微生物数据中心团队建设了国际引领的微生物大数据平台体系。其中全球微生物菌种目录(GlobalCatalogueof Microorganism，gcm)集成了来自全球 50 个国家 133 个微生物资源中心 46 万微生物菌种资源数据(Nucleic AcidsRes. 2016)，是目前最大的微生物实物资源数据平台。全球模式微生物基因组数据库(Global Catalogue of TypeStrain，gcType)，是目前模式微生物基因组数据最为全面，功能最为完善的数据平台(*Nucleic Acids Research* 2020)。全球微生物组数据库(The GlobalCatalogueof Metagenomics，gcMeta)整合了超过 200TB 宏基因组相关数据和超过 100 个在线数据分析工具及整合的工作流(*Nucleic Acids Research* 2019)。