

人类基因数据资源管理研究

渠鸿竹^{1,2,3} 方向东^{1,2,3}

(1. 中国科学院北京基因组研究所(国家生物信息中心), 中国科学院基因组科学与信息重点实验室, 北京 100101; 2. 中国科学院大学, 北京 100049; 3. 基因组与精准医学检测技术北京市重点实验室, 北京 100101)

摘要: 人类基因数据是破译人类遗传信息的重要科学数据。随着DNA测序技术的发展, 人类基因数据量呈爆发式增长。研究和探讨人类基因数据资源的管理对于更好地共享利用数据资源具有重要的意义。本文阐述人类遗传资源的管理历程, 分析人类基因数据现状以及国内外基因数据资源管理现状, 对人类基因数据共享利用提出建议。

关键词: 基因数据; 人类遗传资源; 资源共享; 资源利用; 资源保护

中图分类号: Q1

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2021.01.003

Research on Management of Human Genetic Data Resource

QU Hongzhu^{1,2,3}, FANG Xiangdong^{1,2,3}

(1. CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences/ China National Center for Bioinformation, Beijing 100101; 2. University of Chinese Academy of Sciences, Beijing 100049; 3. Beijing Key Laboratory of Genome and Precision Medicine Technologies, Beijing 100101)

Abstract: Human genetic data is important scientific data for deciphering human genetic information. With the development of DNA sequencing technology, the amount of human genetic data has exploded. Investigating the management of human genetic data resources is of great significance for better sharing and utilization of data resources. Here, we describe the management process of human genetic resources, analyze the current status of human genetic data and the management status of domestic and foreign genetic data resources, and make suggestions for the sharing and utilization of human genetic data.

Keywords: Genetic data, human genetic resource, resource sharing, resource utilization, resource protection

0 引言

中国人口约占世界人口总数的22%, 除了占绝大部分的汉族以外, 还有55个少数民族生活在全国各个相对固定的地域内。中国庞大的人口基数、丰富的民族多样性、独特的地理隔离人

群及疾病核心家系资源构成了丰富典型的中国人遗传资源, 是研究中华民族起源、基本生命现象、生理和病理机能以及行为的物质基础, 也是防治重大疾病、推动医药科技创新、保障人口健康以及维护人口安全的重要物质基础^[1]。一直以来, 我国高度重视人类遗传资源的保护。《中华

作者简介: 渠鸿竹(1982—), 女, 中国科学院北京基因组研究所(国家生物信息中心)副研究员, 研究方向: 生物大数据整合分析与精准医学; 方向东(1969—), 男, 中国科学院北京基因组研究所(国家生物信息中心)研究员, 研究方向: 生物大数据整合分析与精准医学(通信作者)。

基金项目: 国家“精准医学研究”重点研发计划项目(2016YFC0901700; 2018YFC0910700)。

收稿时间: 2019年9月2日。

《中华人民共和国人类遗传资源管理条例》的实施,更加明确了人类遗传资源的概念,更加强调了要保护遗传材料产生的数据形式的信息材料。加强对人类遗传基因数据等信息资源的保护,是为了更好地共享利用。近年来,随着DNA测序技术的发展,基因数据量激增,如何能够更好地共享利用这些资源是我们面临的挑战。因此,研究和探讨人类基因数据资源的管理对于更好地共享利用数据资源具有重要的意义。本文将阐述人类遗传资源的管理历程,分析国内外基因数据资源管理的现状,并对基因数据共享利用提出建议,以供相关研究人员参考。

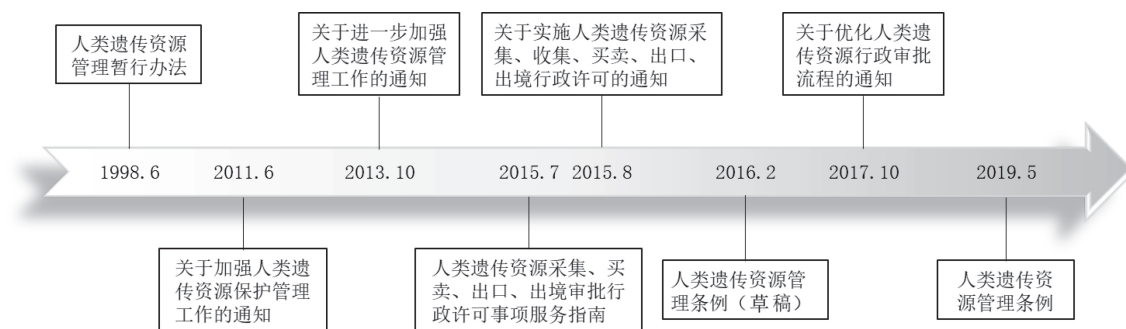
1 我国人类遗传资源的管理历程

中国丰富的人类遗传资源是我国重点保护的對象,也是国外一些利益追逐者觊觎和窃取的对象^[2]。在这个大背景下,我国先后出台了一系列的管理办法,对有效保护和合理利用我国人类遗传资源发挥了积极的作用。1998年,国务院办公厅颁布了《人类遗传资源管理暂行办法》^[3]。2011年、2013年科技部先后发布了《关于加强人类遗传资源保护管理工作的通知》《关于进一步加强人类遗传资源管理工作的通知》。2015年,科技部发布了《人类遗传资源采集、买卖、出口、出境审批行政许可事项服务指南》。同年,又发布了《关于实施人类遗传资源采集、收集、买卖、出口、出境行政许可的通知》^[4]。该通知的发布进一步明确了“分级管理、统一审批”的监管体制。但是,随着形势的发展,人类遗传资源

源非法外流时有发生,我国人类遗传资源的利用还不够规范、缺乏统筹,有关制度还不够完善。因此,我国政府对人类遗传资源加强了管理,科技部于2016年发布了《中华人民共和国人类遗传资源管理条例(草案)》,并于2017年发布了《关于优化人类遗传资源行政审批流程的通知》^[5],进一步简化了审批程序,提高了审批效率。当“基因编辑婴儿”事件发生后,司法部会同科技部对《中华人民共和国人类遗传资源管理条例(送审稿)》作了进一步的修改完善,并于2019年7月1日开始正式实施了《中华人民共和国人类遗传资源管理条例》^[6]。该条例的颁布与实施,进一步促进了人类遗传资源的有效保护与合理利用。

2 人类基因数据现状及其保护利用

人类基因数据是破译人类遗传信息的重要科学数据。随着DNA测序技术的发展,基因数据量呈爆发式增长。DNA测序技术是测定DNA序列的技术。从1977年第一代DNA测序技术(桑格法)开始发展至今经历了三次技术革命,测序技术快速发展。测序技术的每一次变革,都对基因组研究、疾病研究、药物研发等领域产生巨大的推动作用。测序技术的迅速发展降低了测序成本,致使DNA测序数据在2007年以前呈指数增长,近10年来更是达到了超指数增长。据不完全统计,全球每年产生超过200PB组学数据,生命科学成为继天文、地理后第三大数据科学。之前完成的人类基因组计划、DNA元件百科全书



注:数据来源参考文献[7]。

图1 中国人类遗传资源管理的法规历程图

(ENCODE)计划、癌症和肿瘤基因组图谱计划(TCGA)、千人基因组计划、人类表观基因组计划等都已经积累了大量的人类基因数据,并促进了人类对疾病发生发展的认知。随着精准医学时代的到来,各国政府更是进行大人群队列的基因组分析,如英国十万基因组计划、美国精准医疗、韩国万人基因组计划、澳大利亚十万人基因组计划、法国基因组医疗2025、中国十万人基因组计划等(表1)。这些研究计划的实施必将在未来几年加剧数据的积累,也进一步表明各国政府和科学家都已经非常重视人类遗传资源的开发与利用。

基因数据不仅关乎个人的健康,还与国家安全相关。基因是生命体进化过程中遗传信息的载体,是生物体的遗传密码,存储人类种族、分化发育、生长凋亡、疾病发生进展等过程的全部信息,是决定生命健康的内在因素。据OMIM网站统计,到目前为止已经有8 000多种人类遗传疾病被发现和研究^[8],GWAS Catalog数据库中收录的与3 676种疾病或性状相关的变异位点有近10万个(截至2019年8月29日)^[9]。随着大数据技术的发展,基因组数据与医疗数据、健康数据相互融合挖掘,能够帮助人类更好地破解遗传疾病的难题、服务于人类的健康、延续人类的寿命。利用基因数据进行有效的个体识别已经在国家安全防范领域得到广泛的应用。在群体范围以及多学科交叉研究的背景下,对人类遗传资源数

据进行有效管控,已经是国家安全领域的重要课题。尽管某单个个体或少量个体的基因组测序数据对国家安全不可能构成危害,但是面对大量的基因组测序数据,在结合特殊地域特征、特殊民族特征、特殊种群特征、特殊疾病特征、特殊代际关系特征的情况下,就有可能对国家安全、民族安全、种族安全甚至特殊对象安全造成严重危害^[10],因此应加强对基因大数据的保护和管理。

3 国内外基因数据资源的管理现状

鉴于基因数据的重要性以及基因数据指数级的积累,不少发达国家已经建立了基因数据存储平台,实现了国家层面对基因数据的统一管理。20世纪80年代,以美国、英国和日本为代表的国际社会陆续起步建设国家级的生物医学大数据中心,并逐渐形成了垄断全球生物医学大数据的国际三大生物信息数据中心,即美国国立生物技术信息中心(National Center for Biotechnology Information, NCBI)、欧洲生物信息学研究所(European Bioinformatics Institute, EBI)和日本国立遗传研究所DNA数据中心(DNA Data Bank of Japan, DDBJ)。以美国NCBI为例,由政府全额拨款,建立了全球领先的生物医学大数据基础设施,具备数十PB存储、千万亿次计算资源及110Gbps网络带宽资源。同时拥有700余人的科研与专业技术团队,为美国乃至全球的科学家提供基础设施及大数据研究与应用服务,构建

表1 部分国家的基因组计划

国家	起始时间/年	研究时长/年	研究人数/万人	主要内容
英国	2012	4	10	推进基因组医疗整合至英国国家医疗服务体系,加速对癌症和罕见病的了解,提升患者的诊断和精准治疗
美国	2015	5	100	鉴定新的癌症亚型,与药厂等部门合作测试精准疗法的临床效果,拓展对癌症疗法的认识
澳大利亚	2015	4	10	绘制罕见病和癌症患者的基因组图谱,创建国民基因数据库,推动相关药物的研究,构建基于基因组学的新医疗卫生服务系统
韩国	2015	4	1	绘制韩国人基因组图谱,建立韩国标准化的基因数据库,发现罕见遗传疾病的突变位点
法国	2016	10	—	将基因组医疗整合至患者常规监测流程中,建立起国家基因组医疗产业
中国	2017	4	10	绘制中国人基因图谱和中国人健康地图,研究疾病健康与基因遗传的关系

了PubMed、OMIM、GeneBank、SRA、dbGaP、GEO等一批侧重不同类型组学数据存储管理的专业数据库,有力地支持了美国及他国生命科学研究领域的领跑式发展^[1]。除了国家级生物大数据中心之外,基于不同生命组学项目的研究,各国还建立了项目内产生的数据的存储管理数据库,如ENCODE数据库、TCGA数据库等,并建立了完善的数据管理机制,如在TCGA中将数据分为不同的等级,其分析的三级数据可以免费下载,而原始数据以及医疗数据都需要经过严格的审批机制才能获取。

我国开始对生物医学研究和人类基因组研究的时间较长,积累了丰富的经验和资源,并且相关领域的研究资金的支持力度逐年增加。在国家重点研发计划的支持下,人类基因组研究取得了长足的进展。早在1998年,经国家批准,分别在上海和北京成立了中国南方人类基因组研究中心和中国北方人类基因组研究中心。依托这两个平台,产生了许多组学技术,构建了少量的专业数据库。在科研过程中,部分科研院所与公司也形成了自己的数据存储平台,但这些数据都零散存放在不同机构,没有形成统一的国家级基因组数据存储管理平台。但近年来,我国在构建生物医学大数据中心方面呈现快速发展的局面。如2016年1月,中国科学院北京基因组研究所成立生命与健康大数据中心,面向人口健康和社会可持续发展的重大战略需求,建立生物大数据储存、整合与挖掘分析研究体系。生命与健康大数据中心的GSA(Genome Sequence Archive)数据库已开展原始组学数据存储与共享服务^[12],并与国际接轨,成为与NCBI、EBI、DDBJ并列的数据中心,打破了国际垄断,保护了我国组学数据的主权。中国科学院生物物理研究所的非编码RNA数据库、中国科学院微生物研究所的微生物菌株保藏数据库等一批特色数据库都产生了一定的国际影响力。为进一步规范化基因数据的资源共享管理,2019年6月5日,科技部、财政部联合发布了《关于国家科技资源共享服务平台优化调整名单的通知》,公布了不同学科领域的20

个国家科学数据中心,“国家基因组科学数据中心”“国家微生物科学数据中心”名列其中。“国家基因组科学数据中心”依托中国科学院北京基因组研究所建设,其共建单位包括中国科学院上海生命科学研究院和中国科学院生物物理研究所;“国家微生物科学数据中心”依托中国科学院微生物研究所建设^[13]。国家对基因数据的资源共享管理进一步规范化。

由于基因数据自身具有大数据特性,并且其与人类健康密切关联,在国家层面和科学研究方面,政府人员和科学家们都对严格管理基因数据资源达成共识。然而,目前基因数据资源存储分散、集成困难、缺乏有效的挖掘技术和安全保护技术等,有效管理和共享利用基因数据资源还面临诸多挑战。

4 人类基因数据资源共享利用的建议

中国虽然是资源大国,但不是资源强国,只有将保护和开发利用有效结合起来,才能成为国家真正的战略资源。对人类基因数据的安全保护就是为了更好地有效利用。随着人类基因数据量的增长,数据的有效利用则成为技术瓶颈与研究热点。实现人类基因数据的共享利用可能需要在以下几方面加快建设。

(1) 加快数据标准化建设,为数据的统一汇交提供标准。目前,生物数据来源广泛、测序仪器种类众多,数据类型和格式各异,存储设备和存储结构不完善,很难保证数据的延续性、完整性^[14]。在数据共享方面,各资源收集保存单位的数据标准不一致,缺乏可比性,影响着人类遗传资源的原始质量,使得数据资源难以实现整合,阻碍了数据资源的共享利用。因此,需要制定统一的国家人类遗传资源标准平台标准体系,为我国人类遗传资源的整理整合以及信息共享系统的建立提供必要的基础条件。“十三五”规划中设立的“精准医学大数据处理利用的标准化技术体系建设”项目,致力于基因数据共享利用的标准建设,相关标准正在制定中。

(2) 加快资源共享平台与机制建设,为共

享提供合法途径与保障。中国的资源和数据具有存储碎片化、管理分散、无安全保障和标准化质控的特征，难以形成规模化资源和有效共享转化，因此需要统一布局，建立类似NCBI、EBI、DDBJ的具有一定规模、由国家统一管理、与国际接轨和国际认可的国家级资源战略保藏平台。目前国家基因组科学数据中心一直致力于该方向的一定研究，建立了组学数据存储平台，摆脱了中国科学家只能到外国数据库中上传数据的局面。科技部通过“十三五”规划发布“精准医学研究”重点研发专项，从顶层设计上设立了国家数据平台的建设项目，但仍需国家制定完善的共享机制，尽快出台与国际接轨的数据管理和共享使用的规范和指南，在机制指引与平台汇聚的共同作用下实现数据共享。

(3) 加快共享利用的相关技术研发，为基因数据有效保护和共享利用提供技术保障。在大数据时代，数据带来了巨大价值的同时，也带来了用户隐私保护方面的难题，因此应在技术层面通过匿名化、差分隐私、数据脱敏、加密等方法保护用户隐私，防止敏感信息泄露的问题，使数据提供者安心地上交数据。同时还要研发追踪溯源与预警技术，即使发生数据流失泄露，也可以在第一时间发现泄露的人员与地址，及时截留制止，防止危害进一步发生，并可根据相关法律追究责任人责任。另外，多模态数据的整合分析技术是基因数据有效利用的基础。除了多组学数据的整合分析，与医疗健康数据的深度融合，建立疾病预警预测与诊断、个性化干预推荐以及预后评估等工具，最大程度地有效利用数据、提升疾病诊疗水平。

(4) 加快资源保护立法，运用法律手段解决资源共享过程中产生的问题。虽然国家已经颁布《中华人民共和国人类遗传资源管理条例》，对知情同意、伦理审查、审批备案、合作权益等制定了详细的规则，但是对违反规定应承担的惩罚没

有做出详细规定。近年来，生物技术的发展，很大程度上得益于对遗传信息的获取、解读，但非法组织和机构对中国人类遗传资源的攫取和非法利用的行为屡禁不止，因此应尽快出台适应当前发展需求的《管理法》，通过法律手段解决基因数据应用与共享过程中产生的利益均衡问题、生物安全问题，加大对违法行为的处罚和约束，提升监管机构的监管效力，为中国人类遗传资源提供切实有效的保护，降低生物安全风险。

参考文献

- [1] 曹宗富, 曹彦荣, 马立广, 等. 中国人类遗传资源共享利用的标准化研究[J]. 遗传, 2008, 30(1): 51-58.
- [2] 徐慈根. 人类遗传资源的开发与争夺[J]. 中学生物学, 2005, 21(6): 2-5.
- [3] 国务院办公厅. 人类遗传资源管理暂行办法[Z]. 北京: 国务院, 1998.
- [4] 科技部办公厅. 关于实施人类遗传资源采集、收集、买卖、出口、出境行政许可的通知[Z]. 北京: 科技部办公厅, 2015.
- [5] 科技部办公厅. 关于优化人类遗传资源行政审批流程的通知[Z]. 北京: 科技部办公厅, 2017.
- [6] 国务院. 中华人民共和国人类遗传资源管理条例[Z]. 北京: 国务院, 2019.
- [7] 曹诗雨, 武志昂. 我国临床试验中人类遗传资源政府规制研究[J]. 中国药事, 2019, 33(2): 131-136.
- [8] OMIM [DB/OL]. [2019-08-15]. <https://omim.org/statistics/entry>.
- [9] GWAS Catalog [DB/OL]. [2019-08-15]. <https://www.ebi.ac.uk/gwas/>.
- [10] 李伟国. 大数据对国家秘密范围的挑战[J]. 保密科学技术, 2019 (6): 16-22.
- [11] 马英克, 鲍一明. 国家级生物大数据中心展望[J]. 遗传, 2018, 40(11): 938-943.
- [12] WANG Y, SONG F, ZHU J, et al. GSA: Genome sequence archive[J]. Genomics Proteomics Bioinformatics, 2017, 15(1): 14-18.
- [13] 科技部, 财政部. 关于发布国家科技资源共享服务平台优化调整名单的通知[Z]. 北京: 科技部, 2019.
- [14] 操利超, 陈凤珍, 严志祥. 生物数据标准化研究进展[J]. 生物信息学, 2015, 13(1): 31-34.