

# 国家级生物大数据中心展望

马英克<sup>1,2</sup>, 鲍一明<sup>1,2,3</sup>

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, 中国科学院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:** 大数据时代下, 科学大数据已经成为科技创新和社会经济发展的新动力。我国是生物数据生产大国, 生命大数据是人口健康和国家安全的重要战略资源。面对我国生物数据因存储零散、缺乏系统监管而大量丢失和流失, 以及严重依赖国际生物组学大数据中心的局面, 亟需从国家层面建设我国自己的生命大数据保存和管理体系。本文以美国 NCBI 为例介绍了国际生物大数据中心的发展历程及现状, 阐明我国建立国家级生物大数据中心的重要性、迫切性、当前历史机遇和发展前景。中国科学院北京基因组研究所生命与健康大数据中心为此做了大量努力, 并在数据存储、汇交和转化应用上取得了阶段性成果, 以期推进我国生物大数据中心的建设, 提高生命科学研究的国际竞争力和影响力。

**关键词:** 生命与健康; 大数据; 国家级; 大数据中心

## Prospects for national biological big data centers

Yingke Ma<sup>1,2</sup>, Yiming Bao<sup>1,2,3</sup>

1. BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
2. CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
3. University of Chinese Academy of Sciences, Beijing 100049, China

**Abstract:** In the era of big data, scientific big data have become the new driving force for both science and technology innovation and social and economic development. China is a powerhouse in generating vast quantities of biological data, which are an essential strategic resource for population health and national security. The current situation of data loss due to the isolated data storage and the lack of systematic data monitoring and management, and the heavy dependency on international biological data centers urgently calls for China's own life big data storage and management system at the national level. Taking NCBI as an example, this article introduces the development history and present situation of the

收稿日期: 2018-07-02; 修回日期: 2018-09-19

基金项目: 国家重点研发计划项目(编号: 2016YFE0206600), 中国科学院“十三五”信息化建设专项(编号: XXH13505-05)和中国科学院率先行动“百人计划”项目资助[Supported by National Key Research and Development Program of China (No. 2016YFE0206600), the 13th Five-year

Informatization Plan of Chinese Academy of Sciences (No. XXH13505-05) and the 100-Talent Program of Chinese Academy of Sciences]

作者简介: 马英克, 博士, 助理研究员, 研究方向: 生物信息学。E-mail: mayk@big.ac.cn

通讯作者: 鲍一明, 博士, 研究员, 研究方向: 生物信息学。E-mail: baoyim@big.ac.cn

DOI: 10.16288/j.yczs.18-180

网络出版时间: 2018/11/13 11:32:50

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20181113.1132.002.html>

international biological big data centers. In addition, the importance, urgency, current historical opportunity and prospect of establishing a national biological big data center in China are also expounded in detail. In order to promote the development of the national center and improve China's international competitiveness and influence in life science research, the BIG Data Center at Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, has taken many efforts on big data deposition, integration and translation and achieved initial progress.

**Keywords:** life and health; big data; national level; big data center

20 世纪 90 年代发起和实施的人类基因组计划极大地推动了高通量测序技术的进步和应用<sup>[1]</sup>。人类基因组草图绘制完成后, 欧美发达国家又纷纷启动了基于测序技术的生命科学大数据研究计划, 如国际千人基因组计划<sup>[2]</sup>、DNA 元件百科全书计划<sup>[3]</sup>、英国万人和 10 万人基因组计划<sup>[4,5]</sup>、美国精准医学计划(<https://allofus.nih.gov/>)、癌症基因组图谱计划<sup>[6]</sup>和微生物组计划<sup>[7,8]</sup>, 以及日本<sup>[9]</sup>、冰岛<sup>[10]</sup>、加拿大<sup>[11]</sup>和荷兰<sup>[12]</sup>等国家的基因组人群队列研究, 这些计划的实施带动了生物信息学技术、蛋白质组学技术、代谢组学技术、图像处理技术以及其他高通量组学技术的发展, 使得人体成为大数据重要产出源, 以目前多种组学数据、医学影像和临床资料在内统计的生物信息数据产出达到了 10TB/人的水平(基于美国 NetApp.com 公司数据), 全球每年产生的生物数据总量已达 EB 级<sup>[13]</sup>, 标志着生命科学已经从实验数据积累阶段进入大数据科学时代。对生物大数据开展有效的管理和应用, 将信息技术与生物技术有效融合, 正在给生命科学及相关产业领域带来一次新的革命, 尤其在人口健康领域, 大数据贯穿从基础研究、药物开发、临床诊疗到健康管理的所有环节。能否拥有这些生命大数据资源及对其高效存储、管理和应用, 已经成为一个国家综合国力的重要体现。本文以美国 NCBI 为例介绍了国际生物大数据中心的发展历程及现状, 阐明我国建立国家级生物大数据中心的重要性、迫切性、当前历史机遇和发展前景。

## 1 国际生物大数据中心的发展及现状

国际核酸序列数据库联盟(International Nucleotide Sequence Database Collaboration, INSDC)由美

国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)、欧洲生物信息学研究所(European Bioinformatics Institute, EBI)和日本 DNA 数据库(DNA Database of Japan, DDBJ)组成, 掌握和管理着全世界绝大部分的组学生物信息数据。欧、美、日这几大国际生物信息中心建设起步早, 多年来一直引领着全球生物大数据及生物信息领域的发展。以 NCBI(<http://www.ncbi.nlm.nih.gov/>)为例, 早在 1988 年, 美国国会就关注到生物技术领域的重要性, 意识到利用由 DNA 测序带来的大数据的迫切性, 专门成立了 NCBI。30 多年来, 美国政府一直提供持续稳定的支持。NCBI 初建时仅几个人, 发展到今天 700 多人的规模, 它所开发和维护的 PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>)、BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>)和 GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>)等上百个数据库和软件, 已经成为生命科学研究开发领域必不可少的资源。1997 年, 当时的美国副总统戈尔亲自启动了 PubMed 的在线搜索系统, 足见政府对 NCBI 的重视程度。NCBI 还被一些科学家戏称为美国政府做的唯一有用的事情<sup>[14]</sup>。在政府的全额拨款支持下(预算额度最高的 2014 财政年度达到 9480 万美元), NCBI 现今已经形成了具有数十 Petabytes 存储、千万亿次计算资源及 110 Gbps 网络带宽资源的全球领先的国家生物信息中心。NCBI 拥有一支强大的研究开发团队, 为美国乃至全球科学家提供基础设施及大数据研究与应用服务, 有力地支持了美国生命科学研究领域的领跑式发展。

由于国际几大数据中心在生物大数据领域的领导地位, 国际主流期刊杂志要求论文递交者把发表的数据递交到 NCBI 等国际知名数据库, 供全世界科研人员免费使用。另外, 作为美国最大的生物医

学基金资助机构, 美国国立卫生研究院(National Institutes of Health, NIH)资助的科研项目明确要求所产出的基因组信息必须及时在 NCBI 的 GenBank 等数据库公开, 这在很大程度上保证了 NCBI 有稳定的数据来源。这些政策使得全球生命科学研究产生的生物医学大数据, 源源不断地进入国际上极少数的核心数据中心, 数据量不断地暴涨, 截止到 2018 年 8 月, 仅 NCBI 的 Sequence Read Archive (SRA, <http://www.ncbi.nlm.nih.gov/sra/>) 数据已接近 20PB (<https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/>)。

在数据量剧增的同时, 国际大数据中心的经费支持和人员总数却趋于平稳, 给这些中心的运行和维护带来了巨大的挑战。为了应对这一问题, 国际大数据中心一方面在积极寻求新的运维模式, 比如将数据存储到商业云; 另一方面, 不得不削减一些服务, 例如从 2017 年开始, NCBI 的 dbSNP (<http://www.ncbi.nlm.nih.gov/snp/>) 和 dbVar (<http://www.ncbi.nlm.nih.gov/pubmed/dbvar/>) 数据库不再接收、支持除人以外物种的变异数据(<https://ncbiinsights.ncbi.nlm.nih.gov/2017/05/09/phasing-out-support-for-non-human-genome-organism-data-in-dbsnp-and-dbvar/>)。

## 2 我国建立国家级生物大数据中心的重要性和迫切性

我国幅员辽阔, 生物样本资源丰富, 人口数量居世界第一, 仅重大慢病患者就超过 3.4 亿人, 是生物数据产出大国。随着国家在人口健康领域的战略性部署, 越来越多的大型人群队列研究, 如“国家大型健康队列”、“重大疾病专病队列”、“罕见病的临床队列”等项目, 以及人类遗传资源库、主要入侵生物的动态分布与资源库建设等等(基于国家重点研发计划 2018 年度项目申报指南 [http://service.most.gov.cn/2015tztg\\_all/20170526/2179.html](http://service.most.gov.cn/2015tztg_all/20170526/2179.html)), 将要收集数十万甚至数百万人群和生物的百 PB 级数据。这些海量数据, 将会成为我国未来医学模式转变和相关产业创新的重要支撑。对这些数据的存储、管理和计算分析是有效利用生物大数据的前提基础和重要保障。

但是在生物大数据领域, 我国目前尚缺乏国家

层面上对生物大数据进行有效存储、管理和利用的体制、机制和环境, 造成了以下 3 大问题。

### 2.1 我国产出的生物数据得不到永久保存以及共享利用

一方面, 由于缺乏强制性的数据递交和共享政策, 很多数据散落在各个研究人员和单位的电脑里。随着研究人员的流动或电脑设备的淘汰, 有些数据就被丢失了。另一方面, 即使是有些科研项目要求数据汇交共享, 但是由于没有一个国家级的生物大数据中心, 这些数据往往只存放在本项目支持的数据库, 仅支持项目内部有限的共享, 而且也存在着项目结束后缺乏维护更新, 不可持续的问题。更为严重的是, 依托项目的数据库通常是由全新组合的团队构建, 在数据标准的建立、数据管理和共享等方面往往达不到国家级大数据中心的专业水平, 而且有些数据库是简单重复建设。近年来, 我国生物领域数据科学家依托国家项目经费扶持建立了大量的数据库资源, 据最新统计(基于 Database Commons 数据库, <http://databasecommons.org/>), 我国生物数据库资源总位数居世界第二, 但是利用率极低。由于这类数据库大多分散保存、不成系统、水平低下, 不利于数据共享, 造成了国家宝贵生物数据资源和投入资金的巨大浪费。

### 2.2 目前我国严重依赖国际主要生物数据库

一方面, 绝大部分的生物数据以及主要的生物信息分析工具都存放在国际主要生物数据库中, 出于科研的需要, 研究人员必须使用这些数据库搜索、下载和分析生物数据。另一方面, 随着国家科研支持力度的加大和国内整体科研水平的提升, 越来越多的科研单位正在产出更多的生物组学原始数据, 而且用这些数据在国际期刊上发表了大量的论文。按照期刊要求, 在论文发表前, 需要将这些数据递交到期刊认可的公开数据库中。由于我国一直以来缺乏这类国际期刊认可的公开数据库, 几乎所有的数据都必须提交到国际主要数据库(如 NCBI 的 SRA)。由于递交系统的复杂, 绝大多数的科研机构需要雇用专门的生物信息人员, 或者是外包给私营公司来递交这类数据。在这方面的花费甚至超过了

数据产生本身,从国家整体层面上看,这是一笔很大的支出和浪费。除此之外,受国际网络带宽的限制,数据传输缓慢,加上语言障碍等方面的因素,经常造成数据递交的延误,影响论文的及时发表。在当今世界上研究成果发表分秒必争的时代,这些问题让我国在国际科研竞争舞台上失去先机。另外还有断网的风险,一旦国际数据库出现问题(如美国财政预算导致 NCBI 停摆),国内相关搜索和分析都将中断,会对国内的科研造成巨大影响。

### 2.3 我国目前对生物数据的使用尚缺乏有效的监管

出于对个人隐私和知识产权等方面保护的需求,国际主要生物数据库对一些敏感数据(比如人类遗传资源数据)会设置不同层级的数据共享权限和管理原则,按共享层级分成公开级、学术共享级和授权共享级等。我国虽然起草了《人类遗传资源管理条例》并在积极推动其立法实施,而且成立了专门机构对人类遗传资源采集、收集、买卖、出口和出境的申报进行审批,但是由于研究项目和团队众多,数据存放分散,缺乏国家统一的数据汇交和管理平台,对人类遗传资源不能起到全面追踪和系统监管,造成了一些敏感数据的不当流失,严重损害了国家利益。

为了充分保存和有效利用国家宝贵生物数据资源,维护国家在生物资源方面的合法利益,避免在世界科技赛跑中受制于他人,我国亟需建立国家级生物大数据中心,把我国的生物大数据存好、管好、用好。

## 3 建设国家级生物大数据中心的机遇

最近,国家就大数据、科学大数据以及生物大数据的发展布局出台了一系列战略方针和政策。中共中央政治局 2017 年 12 月 8 日就实施国家大数据战略进行学习,强调加快推动实施国家大数据战略,加快建设大数据基础设施,推进数据资源整合和开放共享,加快建设数字中国,更好服务我国经济社会发展和人民生活改善。2018 年 3 月 17 日,国务院办公厅发布《科学数据管理办法》,标志着我国开始从国家层面实施科学数据管理。2018 年 4 月 10

日,科技部发布“十三五”生物技术创新专项规划,明确提出:建设国家生物信息中心、人类遗传资源库和生物和医学大数据等战略资源平台,构建一批资源共享库及共享服务体系。这些大数据战略方针、政策为建立我国国家级生物大数据中心带来了前所未有的历史机遇。

我国具有庞大的生物数据资源优势及世界领先的数据产出能力,为国家级生物大数据中心提供了充足的数据储备。多年来,我国相继建立了包括北京大学生物信息学中心、国家人口与健康科学数据共享服务平台、中国科学院生物物理研究所健康大数据中心、凤凰中心、中国科学院微生物研究所大数据中心、国家基因库和上海生物医学大数据中心等在内的各种类型的大数据中心,已逐步具备形成国家级生物大数据中心的研究基础、设施架构、技术支撑体系。特别值得一提的是,2016 年成立的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)<sup>[15,16]</sup>,被国际同行列入与美国的 NCBI 和英国的 EBI 齐名的全球主要数据中心<sup>[17]</sup>,标志着我国生物大数据中心开始同国际接轨。

1999 年 6 月,已故郝柏林院士在院士建议书中就提出了建立“国家生物医学信息中心”的建议。2013 年 1 月,中国科学院专门组织了由 30 多名生物信息学领域相关院士和专家组成的调查组,召开了多次调研会,专程访问了 NCBI 等国际生物信息中心,并于 2015 年底向国家有关部门提交了调研报告:《我国亟待建设“国家生物信息中心”的建议》。由此可见,建立国家级的生物大数据中心是我国广大科研人员的共识。多年来,我国培养并积累了一大批从事生物大数据研究的学者,更有超过 10 位曾经在 NCBI 和 EBI 等国际知名生物大数据中心工作过的专家回到国内工作,他们具有丰富的实践经验,可以为我国国家级生物大数据中心的建设发挥十分重要的指导作用。

## 4 结语与展望

虽然我国大数据产业起步晚,但是发展迅速且势头强劲,已经在各个领域成为推动经济发展和提升政府治理能力的重要引擎。中国信息通信研究院

《中国大数据发展调查报告(2017)》指出 2016 年中国大数据市场规模达 168 亿元, 预计 2017~2020 年仍将保持 30% 以上的增长速度。

为了加强对我国生物大数据的管理, 解决我国生物大数据流失的问题, 中国科学院北京基因组研究所生命与健康大数据中心于 2016 年初成立。成立两年多来, 中心已建成了中国首个具有自主知识产权的组学原始数据归档系统(Genome Sequence Archive, GSA)<sup>[18,19]</sup>, 目前已有来自近 100 家科研单位的 300 多用户向 GSA 提交过数据, GSA 存储的总数据量将近 600TB。中心还建成了 6 大基础数据库和多个特色资源库, 提供跨库检索功能, 形成了多组学数据资源体系<sup>[20]</sup>。另外, 在精准医学方面, 中心基于已有的中国人群基因组数据建成了中国人群动态基因组数据库和中国人群全基因组序列的基因组变异数据库<sup>[21]</sup>。

尽管如此, 建设我国国家级生物大数据中心还面临着诸多挑战: (1) 获取尽可能多的我国生物大数据资源的机制还不完善。这一方面需要我们创建更好的数据获取模式来维系生物大数据的生态系统, 例如和国内外有影响力的期刊合作, 建立数据引用机制, 激励更多的用户提交更多更有价值的数据; 另一方面也需要国家的政策支持, 例如国家财政课题产生的数据必须强制递交到国家级生物大数据中心等。(2) 缺乏全方位支撑生物大数据深入解析的平台。这需要我们进一步完善现有的数据库资源体系和计算分析工具, 例如: 持续改进现有的生物组学大数据汇交、存储与管理系统, 形成国内生物信息数据汇聚基地; 整合计算机硬件与生物信息软件、工具和流程等资源, 形成面向生物信息大数据分析 with 挖掘的“生物云”平台; 分类整合与挖掘汇聚数据, 完善多组学数据库系统, 形成面向不同研究方向的数据服务体系等。(3) 国内生物数据中心分散, 缺乏统一的数据标准和规范, 给建立国家级生物数据库体系带来困难。这就需要科研人员研究基于元数据的多个数据库的模式整合技术, 研究基于多模态数据索引的高速搜索排序算法, 开发生物大数据智能搜索引擎等。

我们希望并且坚信, 在生命健康领域, 将会很快通过统一生物大数据存储和共享的标准, 同时建

成国家级生物大数据中心, 将信息科学、生命科学、计算科学和临床医学有效交叉, 开展多维数据的深度挖掘, 揭示海量数据中蕴含的深刻科学规律, 获取新知识和新发现, 形成有效的管理能力和技术体系, 以承接我国生物资源、人口健康、环境与农业等大数据和国家人类遗传资源, 利用数据开展实时分析、预测分析、个性化分析和解析复杂相关性等等, 为数据使用者提供更方便、迅捷、准确的技术服务, 形成立足我国和具有国际影响力的生物信息研究和应用中心。

## 致谢

感谢北京大学罗静初教授和国家蛋白质科学中心朱伟民教授对文章的宝贵意见, 感谢北京基因组研究所吴双秀对文章的编辑和整理工作。

## 参考文献(References):

- [1] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 2004, 431(7011): 931–945. [\[DOI\]](#)
- [2] 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. *Nature*, 2010, 467(7319): 1061–1073. [\[DOI\]](#)
- [3] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 2004, 306(5696): 636–640. [\[DOI\]](#)
- [4] Parry V. Commit to talks on patient data and public health. *Nature*, 2017, 548(7666): 137. [\[DOI\]](#)
- [5] Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereau A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, 100 000 Genomes Project. The 100 000 genomes project: bringing whole genome sequencing to the NHS. *BMJ*, 2018, 361: k1687. [\[DOI\]](#)
- [6] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 2008, 455(7216): 1061–1068. [\[DOI\]](#)



- [7] Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*, 2007, 449(7164): 804–810. [DOI]
- [8] NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. The NIH human microbiome project. *Genome Res*, 2009, 19(12): 2317–2323. [DOI]
- [9] Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, Sato Y, Mimori T, Tsuda K, Saito R, Pan X, Nishikawa S, Ito S, Kuroki Y, Tanabe O, Fuse N, Kuriyama S, Kiyomoto H, Hozawa A, Minegishi N, Douglas Engel J, Kinoshita K, Kure S, Yaegashi N, ToMMo Japanese Reference Panel Project, Yamamoto M. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun*, 2015, 6: 8018. [DOI]
- [10] Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadóttir HT, Johannsdóttir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdóttir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdóttir H, Steingrimsdóttir T, Gudmundsdóttir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdóttir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardóttir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdóttir U, Helgason A, Sulem P, Stefansson K. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet*, 2015, 47(5): 435–U420. [DOI]
- [11] Reuter MS, Walker S, Thiruvahindrapuram B, Whitney J, Cohn I, Sondheim N, Yuen RKC, Trost B, Paton TA, Pereira SL, Herbrick JA, Wintle RF, Merico D, Howe J, MacDonald JR, Lu C, Nalpathamkalam T, Sung WWL, Wang Z, Patel RV, Pellecchia G, Wei J, Strug LJ, Bell S, Kellam B, Mahtani MM, Bassett AS, Bombard Y, Weksberg R, Shuman C, Cohn RD, Stavropoulos DJ, Bowdin S, Hildebrandt MR, Wei W, Romm A, Pasceri P, Ellis J, Ray P, Meyn MS, Monfared N, Hosseini SM, Joseph-George AM, Keeley FW, Cook RA, Fiume M, Lee HC, Marshall CR, Davies J, Hazell A, Buchanan JA, Szego MJ, Scherer SW. The personal genome project Canada: findings from whole genome sequences of the inaugural 56 participants. *CMAJ*, 2018, 190(5): E126–E136. [DOI]
- [12] Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*, 2014, 46(8): 818–825. [DOI]
- [13] May M. Life science technologies: Big biological impacts from big data. *Science*, 2014, 344(6189): 1298–1300. [DOI]
- [14] Somerville C, Flanders D, Cherry JM. Plant biology in the post-Gutenberg era - everything you wanted to know and more on the world wide web. *Plant Physiol*, 1997, 113(4): 1015–1022. [DOI]
- [15] BIG Data Center Members. The BIG data center: from deposition to integration to translation. *Nucleic Acids Res*, 2017, 45(d1): D18–D24. [DOI]
- [16] BIG Data Center Members. Database resources of the BIG data center in 2018. *Nucleic Acids Res*, 2018, 46(d1): D14–D20. [DOI]
- [17] Rigden DJ, Fernandez XM. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res*, 2018, 46(d1): D1–D7. [DOI]
- [18] 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]
- [19] 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]
- [20] Zhang YS, Xia L, Sang J, Li M, Liu L, Li MW, Niu GY, Cao JB, Teng XF, Zhou Q, Zhang Z. The BIG data center's database resources. *Hereditas (Beijing)*, 2018, 40(11): 1039–1043. 张源笙, 夏琳, 桑健, 李漫, 刘琳, 李萌伟, 牛广艺, 曹佳宝, 滕徐菲, 周晴, 章张. 生命与健康大数据中心资源体系. *遗传*, 2018, 40(11): 1039–1043. [DOI]
- [21] Song SH, Teng XF, Xiao JF. Database resources of the reference genome and genetic variation maps for the Chinese population. *Hereditas (Beijing)*, 2018, 40(11): 1048–1054. 宋述慧, 滕徐菲, 肖景发. 中国人群参考基因组及基因组变异图谱资源库. *遗传*, 2018, 40(11): 1048–1054. [DOI]