生物研究大数据的开放与共享探讨

侯 勇
（湖北华大基因研究院，湖北武汉 430074）

摘要：首先介绍生物研究大数据概念及其开放与共享现状，认为生物研究大数据开放与共享对科学研究具有促进的作用。然后针对生物研究大数据开放与共享的困难，以GigaScience为例探讨了生物研究大数据开放共享具体研究与做法。最后针对生物研究大数据开放与共享面临的技术挑战，提出了进一步的发展建议，以推动生物研究大数据开放与共享。

关键词：生物研究；大数据基因组；生物数据开放共享；GigaScience；开放存取
中图分类号：G203 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2017.04.002

Study of Openness and Sharing for Bioresearch Big Data

HOU Yong
(BGI-Wuhan, Wuhan 430074)

Abstract: This paper presents the concept of bioresearch big data and introduces its openness and sharing statue, and it is believed that biological data openness and sharing can promote scientific research. Due to difficulty of openness and sharing for bioresearch big data in practice exploit, so taking GigaScience for example, its specific research and method are studied in it. Finally, confronting its technological challenge, further development proposes and preliminary forecast are proposed to it.

Keywords: bioresearch, big date genome, biological data openness and sharing, GigaScience, open access

1 引言

近年来，生物研究越来越离不开计算和数据。以基因组为代表的生物大数据是近年来兴起的一个生命科学与数据科学交叉融合的名词。基因组学、转录组学、蛋白质组学、代谢组学、环境组学、宏基因组学、影像组学等一系列生物大数据的发表为生物研究大数据的开放与共享提供了巨大的机遇与挑战。近期调查和研究证明，数据开放与共享是促进科学研究，消弭科学研究重复性陷阱的必经之路。在促进生物研究大数据开放与共享方面已经进行了一些有益的尝试并取得了不错的效果。但是，未来生命科学正朝着定量化与数字科学方向发展，生物研究大数据的开放与共享也将面临着更大的技术挑战。本文将重点讨论生物研究大数据开放共享的问题，以期为促进数据开放与共享、打造一个更为开放和透明的生物大数据研究领域提供一定的参考建议。

2 生物研究大数据概述

生物大数据是近年来兴起的一个生命科学与数据科学交叉融合的名词。究其本源，离不开以基因组学技术为代表的组学技术的发展。人类基因组计划完成以来，绘制一个人完整基因组的

作者简介：侯勇（1989—），男，博士，华大基因研究院副院长，研究员，主要研究方向：肿瘤药物基因组研究、单细胞分析转化医学研究。

收稿日期：2017 年 6 月 24 日。

成本已经从数十亿美元降低至1000元美元。众所周知，人类基因组有30亿个碱基对，一个人的基因组学数据可以达到100G之多。与此同时，由于生命的中心法则和生命的复杂性，转录组学、蛋白质组学、代谢组学、可穿戴组学、环境组学、宏基因组学、影像组学等相关技术也都蓬勃发展起来。诸如，生化实验检测结果、MRI/CT、超声、皮肤影像检测、三维晶体结构等目前也在朝着数据开放和共享方向发展。这样，每个人实时动态的完整生命组学数据就接近1Tb，而这些数据正以超过指数增长的方式在持续增长。如今，基因组学的兴起已经渗入到进化研究、临床医学、法医学、考古学、健康管理等研究和应用领域。近年来，越来越多的生物医学研究使用了生物大数据，而生物大数据也与这些学科交叉融合。这些大数据对传统研究成果的发表方式、对没有数据科学背景的生物学家甚至是科研基金机构等都带来了巨大的挑战，特别是对以往的重结果轻过程、重论文轻数据、重竞争轻共享、重保密轻开放的惯性科研思维产生重大的影响。同时，海量的生物大数据需要海量的存储资源和计算资源，也决定了生物大数据具有共享和开放的特征，因为任何一家研究机构或者科学家都不能独立地从所有大数据中挖掘出所有有价值的信息。然而在未来，生物大数据也将结合计算机领域的深度学习、人工智能等技术，创造出全新的生物科学与医学知识。

3　生物研究大数据开放与共享的现状

在近几个世纪里，研究论文都是学术交流的主要载体，不管是网上出版还是开放获取，都没有从根本上改变研究论文出版的过程和结构。随着生物和生物医学研究越来越朝着数据驱动方向发展，像基因组、影像等方面出版物的信息量、计算工具、编码等已经呈现指数级增长，然而，如果缺少支撑这些出版物的原始资源特别是数据，就会造成所谓的“重复性陷阱”[1]。

重复性陷阱的出现正是由于期刊或者文章的作者仅仅提供有限的信息。在对不同研究领域的研究可重复性进行测试时，Ioannidis和同事们发现不少发表的结果是错误或者过分强调的，并且估计约有85%的研究资源因此而被浪费[2]。例如，一项研究表明，在微阵列研究中，9项工作中仅有一项能够被重复出来[3]。类似的情况也出现在肿瘤研究领域，仅仅有11%的研究可以被重复出来[4]。从过去很多发表的图片以及临床前研究中发现的不可重复的结构累积已经超过了50%。换句话说，每年有高达28亿美元的临床前研究无法得到重复[5]。随着撤稿率的不断升高，特别是高水平期刊的撤稿率的提升[6]，如何降低甚至消除重复性陷阱确实已经成为避免浪费研究资源和得到错误的研究结论以及增强公众对科研工作者信心的头等大事。

虽然大家都意识到“重复性陷阱”对科学界的信誉以及科学的发展是一个巨大的问题，但是在现实世界中数据或者源码的开放并自由获取却遇到了巨大的挑战。目前，很多作者仅仅依赖于在自己的网页上提供的数据和材料，这样的做法已经被证明对数据的开放不会产生太大的作用。最近，由癌症生物学组织发起了一项针对研究重复性的研究，这项研究旨在对研究重复性进行定量分析。研究人员选取了在2010年至2012年的50篇高引用率论文，他们获取每项研究相关的数据平均耗时2个月，而在50项研究中甚至有4项的作者在一年后才配合提供相关的研究数据[7]。进一步地，基于ACM会议和期刊论文的评估，唯一能证明可重复性和重复利用性的源代码平均需要2个月才能获取，而且只是其中44%的文章有回应。又有一项针对200篇经济学论文的调查发现，只有64%的作者有回应，而其中有56%的作者表示不愿意提供或者共享补充材料。最近出现的长期未发现的数据造假丑闻（包括最近一位作者通过数据作假发表了172篇文章），进一步凸显了能够十分轻松地访问数据的重要性，因为这样可以确保其他人进行独立的重复验证，并且增强科学界的相互信任。

对于一些基金机构来说，比如美国国家自然科学基金已经开始计划对所有资助项目进行数

据共享和开放，而美国国立卫生研究院已经走在了前面。美国国立卫生研究院投资了“大数据到知识”项目，计划的名字叫做“bioCADDIE（http://biocaddie.ucsd.edu/）”，希望以此来推动生物医学和健康管理方面的数据发现和索引生态系统的建设工作。数据发现索引使得“bioCADDIE”项目瞄准了医学和生命科学领域数据库PubMed中已经存储的数据，比如Pubmed以及PubMed Central，来提供存档数据的结构建设和管理。在欧洲，研究和创新基金以及OpenAIRE计划都要求参与者在开放存取的期刊上发表他们的结果。他们也有一个数据开放早期研究项目已经被“地平线2020”项目选中，进行数据管理计划以及数据存储计划的研究。英国研究委员会已经草拟了一个数据开放相关的草案，敦促各方开放数据并使得数据可以重复使用。然而，在国际间并没有一个专门的机构进行协调和统一，因此也给国际化研究的数据开放造成了一定的困难。

生物研究大数据开放的另一重要参与方就是期刊。他们如果能够加强管理政策制定与执行就能够起到一定的效果，比如在生态学期刊中的数据联合存档政策。尽管有不少积极的信号已经在鼓励开放存取的出版商开始解决生物研究大数据开放共享的问题，但是生物研究大数据开放共享还有很长的路要走[8]。最近的一项调查发现，在影响力最高的50种期刊中，44种已经制定了数据共享的政策，但是能够访问数据的仅仅是其中的很小一部分，而在一些情况下能够访问到原始数据则更少，不到10%[9−10]。随着生物医学研究越来越依赖计算机，计算方法和代码共享要比数据共享更为困难[11]。

4　生物研究大数据开放与共享实例分析

虽然在生物研究中，大数据的开放与共享存在着上述诸多的困难和挑战，但是也有一些很好的例子在推进生物大数据的开放与共享，促进科学研究。比如由华大基因主办的开放获取期刊《GigaScience》[1,12−13]。GigaScience是一个崭新的开放型在线期刊，于2012年7月12日创刊。GigaScience采用标准全文文献、数据库信息以及信息分析工具相结合的模式，为科研工作者提供免费公开的有效数据以及生物学发现等资源。在同行评议的过程中，GigaScience可以为审稿人提供所有支撑性的信息和数据。这些通过ftp访问的数据在有些研究中已经达到了100G。比如，在SOAPdenovo2[14]的发表过程中，编辑与审稿专家完整地测试了不同工具的表现情况，以确保测试结果与作者在文章中的陈述相一致。而参与这个过程的8名审稿人均在一个含有他们名字的报告上签名，作为文章发表前的历史参考。同时，GigaScience的审稿过程是公开化与透明化的。GigaScience在选择审稿人时通常会选择愿意公开自己的审稿人，这就使得审稿人在接受审稿后变得异常的谨慎，既确保了审稿的质量，也确保了所有的相关作者得到相对公正的对待。同时，GigaScience创造性地给数据集和工具集以DOI，使其可以被单独检索和引用，从而保障了所有利益相关方的利益，同时也促使更多的作者愿意以这种方式公开自己的数据和相关的分析工具。

目前，GigaScience的数据库中已经有超过200个关于数据的DOI，是世界上最大的组学数据库，包括测序相关的基因组、转录组、表观基因组、宏基因组，同时还有质谱技术的蛋白质组和代谢组。最近，也增加了许多MRI/CT等影像学数据以及电生理等其他系统生物学数据。目前，已经有30T的数据可供自由下载，其中最大的是农业相关数据库，包括379只牛、3000株水稻以及人类肿瘤数据。

GigaScience还解决了一个长期困扰科学界的问题，就是在文章发表前，数据的传播速度很慢，而这些数据如果能够尽早公布，就能够获得更多的科学发现，促进科学的发展，甚至很有可能挽回一些人的性命。比如，帝企鹅、北极熊等数据在论文发表前3年就已经通过GigaScience发表，而在这3年间，已经在一些群体遗传学和进化生物学研究中引用了这些数据[15]，然而这些引用并没有影响论文在2014年《细胞》杂志上以封面文章的形式发表[16]。再比如2015年出

版的《科学》杂志鸟类专刊。该专刊的出版意味着科学家首次能够在分子层面上解析鸟类的进化之谜，阐述了控制声音学习的分子通路在一些鸟类和人类的大脑语言控制区域中的独立演化过程、鸟类性染色体复杂的演化历程、鸟类在早期演化过程中是如何丢失牙齿、鸟类近亲鳄鱼的基因组是怎样演化的、鸟类歌唱行为在大脑内的基因调控机制以及一种利用大规模基因组数据构建演化树的新方法。在《科学》鸟类专刊发表前，GigaScience已经陆续公布了多达4TB的43种鸟类的基因组数据供全球的科学家研究和解析，包括了鸟类的基因组以及光学遗传图谱的数据。

另外再举一个例子。致死性大肠杆菌在欧洲造成了50人死亡、1000余人感染的恶果。这是一个非常实用的证明数据，提前公开可以增进人类健康福祉的例子。2011年由华大基因和德国科学家共同完成了致死性大肠杆菌的基因序列[17]。GigaScience和华大基因第一时间公开了基因测序数据，而后相关文章在新英格兰医学杂志上发表。基因序列公布后，很多科学家参考这种方式公布了他们的测序结果。很快地，来自北美、欧洲、澳大利亚、埃及等国家和中国香港等地区的科学家以及热心的科研爱好者都加入了分析，这使得科学家快速定位出了致病基因与致死基因以及传染源，有效地制定了公共卫生政策来应对疫情。除此之外，大数据公开的时效性还体现在另外一个及时公布的数据上。在2014年9月GigaScience公布了全球首个用OXFORD NANOPORE技术获得的微生物全基因组序列[18]。这些数据高达125G，结果是如何让全球的科学家快速获取这些数据却成了难题。GigaScience又与EBI的科学家合作，通过镜像转移数据，使得许多欧洲科学家也能够在第一时间获取数据。这些数据的公开，渐渐平息了科学界关于新一代单分子测序数据质量问题的争论，因为任何一个实验室都可以自由访问这些数据，可以自由且自主地评价这些数据的质量。

除了上述内容外，GigaScience还特别重视大数据分析工具的提供。GigaScience在GITHUB上开通了专门的网页，提供发表文章的源代码，以供感兴趣的科学家对大数据进行同步分析。

5　生物研究大数据开放与共享的建议与展望

综上所述，生物研究大数据开放与共享离不开期刊、基金机构，特别是研究人员的推动。因此，提出几点相关的建议。

（1）期刊编辑和审稿人要提出严格要求。如果高水平期刊的编辑应该要求作者必须执行数据公开和共享的政策，那么上述出现的很多重复性陷阱或许能够避免。另外，通过改革审稿方式，如果能够使审稿人的信息公开透明，那么可以营造一个更加公平的同行评议环境。

（2）训练科学家的数据科学思维。对于高标准的数据分析来说，必要的基础知识和技能是不可或缺的。因此，如果能够对青年科学家进行数据科学的训练，让他们学会使用诸如Github之类的工具去共享自己的代码，使用Github上的代码进行研究和引用。对于没有计算机或者相关背景的科学家来说，可以在实验室里保留一个Github来管理和共享所有相关的代码，并且通过规范化管理，加强对数据科学的重视。除此之外，也可以利用网络学习等自学方式，寻找适合自己的工具或者习题来提高自己的知识和技能。

（3）在公开的数据库中提供或者引用计算工具的源代码。软件是数据驱动的生命科学研究的重中之重。如果缺乏计算方法或者工具，研究人员无法完全理解研究者提供的大数据真正的意义。为其他研究人员提供计算工具的源代码，比如放在Github上能够方便其他研究者进行相关研究。在论文中引用源代码具体的位置，尽量避免引用不提供源代码的文章。而对于作者来说，尽量使用规范化和格式化的语言进行编程，这样可以使自己的工作通俗易懂，也有可能更加广泛地为人所使用。对于期刊来说，除了提供数据的存储外，也提供一些与数据相对应的分析流程，并且可以将这些分析流程部署在云上，使没有计算生物学或者计算机背景的科学家或者没有计算资源的科学家也能对开放共享的生物研究大数据进

行深度研究。

目前，生物医学研究数据呈现指数增长的趋势。很多研究的数据体量已经从G级别提升到了T级别。要共享如此大的数据量，在技术上将面临非常大的挑战。目前，已经有一些方法解决了这个问题。比如，使用工业化的大数据云计算技术，这也就使期刊从发表数据变为发表计算资源。再如，利用虚拟机、Bioboxes等方式使海量数据共享实现标准化、规模化和规范化。未来生命科学将向着定量化与数字科学方向发展，生物研究大数据的开放与共享将面临着更大的技术挑战。但是，我们坚信，只要科学家达成促进数据开放与共享的共识，未来我们一定能够打造一个更为开放和透明的生物大数据研究领域。

参考文献

[1] EDMUNDS S C, LIP, HUNTER C I, et al. Experiences in integrated data and research object publishing using GigaDB[J]. International Journal on Digital Libraries, 2016, 18(2): 99−111.

[2] IOANNIDIS J P. How to make more published research true[J]. PLoS Med, 2014, 11(10): e1001747.

[3] IOANNIDIS J P, ALLISON D B, BALL C A, et al. Repeatability of published microarray gene expression analyses[J].Nat Genet, 2009, 41(2): 149−155.

[4] BEGLEY C G, ELLIS L M. Drug development: raise standards for preclinical cancer research[J]. Nature, 2012, 483(7391): 531−533.

[5] FREEDMAN L P, COCKBURN I M, SIMCOE T S. The economics of reproducibility in preclinical research [J]. PLoS Biol, 2015, 13(6): e1002165.

[6] FANG F C CASADEVALlA. Retracted science and the retraction index[J]. Infect Immun, 2011, 79(10): 3855−3859.

[7] VANNoorden R. Sluggish data sharing hampers reproducibility effort[J]. Nature, 2015, 10: 1038/nature.2015.17694.

[8] BLOOM T, GANLEY E, WINKER M. Data access for the open access literature: PLOS’s data policy[J]. PLoS Medicine, 2014, 11(2): e1001607.

[9] MAVERGAMES C, SAVAGE C J, VICKERS A J. Empirical study of data sharing by authors publishing in PLoSJournals[J]. PLoS ONE, 2009, 4(9): e7078.

[10] BOUTRON I, ALSHEIKH−ALI A A, QURESHIW, et al. Public availability of published research data in High−Impact Journals[J]. PLoS ONE, 2011, 6(9): e24357.

[11] ZAYKIN D, STODDEN V, GuoP, et al. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals[J]. PLoS ONE, 2013, 8(6): e67111.

[12] EDMUNDS S C. Peering into peer−review at GigaScience[J]. GigaScience, 2013, 2(1): 10.1186/2047−217X−2−1.

[13] KENALL A, EDMUNDSS, GOODMANL, et al. Better reporting for better research: a checklist for reproducibility[J]. GigaScience, 2015, 4(1): 10.1186/s13059−015−0710−5.

[14] LUO R, LIUB, XIEY, et al. SOAPdenovo2: an empirically improved memory−efficient short−read de novo assembler[J]. GigaScience, 2012(1): 10.1186/2047−217X−1−18.

[15] NACHMAN M W, CAHILLJ A, GREENR E, et al. Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution[J]. PLoS Genetics, 2013, 9(3): e1003345.

[16] LIU S, LORENZEN Eline D, FUMAGALLIM, et al. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears[J]. Cell, 2014, 157(4): 785−794.

[17] ROHDE H, QINJ, CUIY, et al. Open−source genomic analysis of Shiga−Toxin–Producing E. coliO104: H4[J]. New England Journal of Medicine, 2011, 365(8): 718−724.

[18] QUICK J, QUINLANA R, LOMAN N J. A reference bacterial genome dataset generated on the MinION™ portable single−molecule nanoporesequencer[J]. GigaScience, 2014, 3(1): 10.1186/2047−217X−3.