健康管理与深度测序数据解析的挑战*

白晋伟 沈百荣

（苏州大学图书馆数字化部 苏州 215006） （苏州大学系统生物学研究中心 苏州 215006）

〔摘要〕 介绍深度测序技术对生物医学研究和社会的影响，讨论深度测序数据解析及其在健康管理应用中所面临的挑战和机遇，包括数据存取、计算技术、数据应用、人才缺失与跨学科人才教育等方面。

〔关键词〕 健康管理；深度测序；数据分析；研究范式

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.01.002

Challenges from Health Management and Deep Sequencing Data Parsing *BAI Jin－wei, Digital Department of Library, Soochow University, Suzhou 215006, China; SHEN Bai－rong, Center for Systems Biology, Soochow University, Suzhou 215006, China*

〔**Abstract**〕 The paper introduces impact of deep sequencing technology on biomedical research and the society, discusses challenges confronting and opportunities enjoyed by deep sequencing data parsing and its application in health management, including aspects like data access, computing technology, data application, lack of talent and interdisciplinary talent education, etc.

〔**Keywords**〕 Health management; Deep sequencing; Data analysis; Research paradigm

1 引言

21 世纪初测定基因组图谱是基于毛细管电泳芯片的 Sanger 测序法，测定一个全基因组图谱的花费要高于 1 000 万美元，随着下一代测序技术的进步，测序速度的提高和测序成本的下降[1]，全基因组测序只需要 1 000 美元就可以实现[2-3]。深度测序技术的发展将对科学和社会产生深远的影响。

2 深度测序技术对生物医学研究和社会的影响

2.1 生物医学大数据与生物医学研究范式的改变

2.1.1 *生物医学大数据的产生和数据驱动的生物学研究* 与传统的测序技术相比，深度测序技术的应用更为广泛，不仅可以应用于定性的 DNA 序列的测定和基因组的结构分析，同时可以用于定量的表达分析。在表达定量分析方面，不仅可以测定已知基因的表达情况，也可以用于新基因的发现和测定。深度测序技术的应用除在研究具体生物问题方面的广度得以拓展之外，同样在物种演化比较、基因与环境相互作用方面得到广泛的应用。由于深度测序的应用范围拓展，基于一个生物标本可进行的测序式样有很多种，如可以对一个生物组织进行基因组测序、分析其基因组结构的变化，也可以测定其表达组信息、表观遗传组信息、变异信息等，因

〔修回日期〕 2017－09－21

〔作者简介〕 白晋伟，硕士，馆员，发表论文 5 篇。

〔基金项目〕 国家自然科学基金资助项目“前列腺癌演变过程中的关键基因和模块及其作用机制”（项目编号：31670851）。

此少量的样本就可以产生大量的数据，这些数据相对于商业领域的个人商品和书籍等购买信息而言，有不同的特征，前者被称为小的大数据，后者为大的小数据，随着数据的累积如人群基因数据的测定，也逐渐形成大的大数据，如 23andMe 公司收集大量的个性化数据可用于大数据分析[4-5]。

2.1.2 个性化医学、P4 医学和精准医学研究范式的兴起 个性化医学、P4 医学和精准医学等研究范式都是近几年在深度测序技术迅猛发展的基础上提出来的，个人基因组测定的可行性使得大众有可能测定和分析自己的基因组、寻找个人健康相关的基因风险因素，从而可以在生活习惯、饮食等方面提前进行个性化预测和预防。由于互联网的发展和即时检验技术（Point - of - care - testing，POCT）的应用，人们可以通过网络进行交流和参与到整个诊疗过程，这便是 P4 医学的概念：预测性（Predictive）、预防性（Preventive）、个性化（Personalized）、参与性（Participatory）。与个性化医学的范式相比 P4 医学更强调早期预测和预防，强调对患者了解的系统性和参与性。精准医学的概念则是在基因测序普及的基础上，将个体的各种信息如生理信息（利用可穿戴设备可以即时监控和收集）和肠道菌群变化、各种组学信息（深度测序测定）整合进行精准的疾病分型、治疗和预防。随着老年化时代的到来以及临床资源的限制，基于这 3 种范式的健康管理和精准诊疗将成为生物医学研究和应用的基本范式，走向大众生活。

2.2 深度测序技术对社会和经济的影响

俄国经济学家康狄夫的长波理论认为商品经济中存在着为期 50 ~ 60 年的周期性波动，根据这一理论，商品经济的第 6 波创新驱动将由信息技术向心理社会健康方面转移[6]。可以相信全球老年化社会到来后的经济主战场将是健康行业。而以基因测序预测健康和临床精准分型的市场将会越来越大。深度测序相关的经济市场有两方面：一是测序仪器和技术相关的市场的竞争达到狂热的程度，中国也有几台国产的测序仪，如华大 BGISEQ500、中科紫鑫 BIGIS 和华因康 PSTAR - II 等，随着测序技术的进一步普及，期待国产测序仪在今后的市场上有一席之地。另一方面是测序的应用市场的竞争，测序的应用产业除科研服务外，最主要的是围绕疾病个性化诊疗的精准分型和健康管理产业、基于基因测试的基因与疾病的风险预测和干预。其他的基因测试市场是疾病诊疗和健康管理业的衍生行业，如保险业务的相关疾病风险预测、遗传咨询相关的基因检测，食品、营养和农业相关的基因检测等。“23andMe”是美国的一家著名 DNA 测定公司[4,7]。可以选择提交 DNA 数据，进行遗传学分析，参与 230 多项研究，寻找疾病治疗和治愈方式。在健康领域的分析还有很多细节，尤其是复杂疾病，中国人的基因突变谱和疾病风险有其个性化的特征，另外还需要收集大量的数据、建立准确的风险分析模型，才能在市场上得到准确应用和认可。

3 深度测序数据处理的挑战

3.1 概述

早期的测序技术是“测定速度没有计算速度快”，下一代测序技术发展以来，变为“计算速度跟不上测定速度”。测序速度给计算带来巨大的挑战。深度测序数据分析针对各种生物问题处理的复杂性有很多，如测序对比、测序组装、碱基识别、蛋白质与 DNA 结合位点分析、诊断应用、分析流程工具软件、转录组测序、变异检测、可视化等[8]，除此之外深度测序技术测定数据的处理和计算可以分为 10 个挑战，根据数据存取、数据运算分析、数据应用以及人才培养分为 4 个方面。

3.2 数据存取方面的挑战

3.2.1 挑战 1：数据存储与提取 深度测序测定的数据同样具有大数据的 4 个 V 的特点，即：大量（Volume）、高速（Velocity）、多样（Variety）、真实性（Veracity）；目前 1 台 HiSeq X Ten 每天可以产生 600GB 的数据，3 天可测定 1.8TB 的数据，通常个人电脑是难以处理这样的巨大数据的。深度测序数据的多样性在于不同的组学数据，如基因组学、转录组学、表观组学以及变异组学等，其真实

性与实验室带进去的污染、仪器误差和计算的假阳性等有关。大数据存取需要很大的硬件和计算代价，开发高效快速的存取格式和模型是深度测序大数据的首要挑战。

3.2.2 挑战 2：数据格式与标准化 大数据的存储与提取与数据的格式和结构密切相关，同时也与数据的重复使用性相关，由于深度测序数据的多样性以及基因之间的相关性，数据的标准性还涉及到数据之间的关联，将高效易解析的数据格式、数据标准化与各种不同的本体论结合起来，将会为数据库和知识库的构建提供基础。

2.2.3 挑战 3：伦理、隐私与数据共享 如果深度测序数据是个人的各种组学数据，将涉及到个人数据的隐私性，保护个人的隐私（包括能力、疾病、习惯等）可以避免在就业、婚恋、保险和社会生活中被歧视等伦理问题。数据的共享必须以隐私保护为前提，然而隐私保护并不是将姓名等个人信息去掉那么简单，据文献报道根据 20 多个基因可以推断一个人的长相，去隐私性需要对基因数据进行复杂的预处理，才能保障数据提供者的信息安全[9-11]。

2.2.4 挑战 4：数据库、知识库 数据库的构建应该考虑从原始数据到知识发展的过程，为知识发现提供好的数据库构架，没有好的数据、数据关联和建模相关的元素，就没有好的知识发现。传统构建的独立元素的数据库，不能用于发现复杂的元素关联网络结构。好的知识库可以用于对数据库的深度解释和机制发现，如 GO，KEGG 等是广为使用的知识库[12-13]。但在临床表型数据和知识方面的缺乏，使得这些知识库在精准基因型和表型关联方面的应用受到制约。将深度测序数据建立成一个大的数据库便于搜查，建立基因与疾病风险之间的知识库对精准的健康监控尤为关键。

3.3 计算技术方面的挑战

3.3.1 挑战 5：运算速度与算法 大数据的处理速度是关键，前面讨论到的数据库系统的整体效率即存取速度。这里讨论的速度是指对数据进行各种操作和分析的速度，需要考虑数据结构、分布和运算资源的分配，最常用的解决办法是高性能计算和并行计算，在计算方案设计过程中需要对生物医学问题有机制性的理解和分析：如数据的重复度、数据的分布结构等，针对深度测序数据不同特征的新的算法是计算的一大挑战，甚至需要考虑软件、硬件整合的计算构架。

3.3.2 挑战 6：数据降维与可操作数据的选择 由于基因数据关联度大、演变性强，在生物医学大数据中寻找驱动基因和子网络，对大数据进行合理的降维，寻找可操作的变量是促进生物医学大数据分析走向应用的关键。

3.4 数据应用方面的挑战

3.4.1 挑战 7：多组学数据的整合与应用 多组学的整合一直是近几年来的研究热点，如将盲人摸象中多个盲人的部分信息进行整合。近几年用于数据整合的系统有很多，如：BTRIS[14]，dbGaP[15]，Enterprise Data Trust[16]，i2b2[17]，IMMPORT[18]，NDAR[19]，STRID[20]，TCGA[21]，TCIA[22]，TRAM[23]等，这里多组学组合不只是在分子组学层次，还有基因与图像组学、生理组学和临床表型组学等方面的整合和应用。

3.4.2 挑战 8：系统建模分析 对复杂系统的理解，系统建模是关键。如何将深度测序得到的数据重建一个基因调控网络的图像，进行动态模拟和静态分析，绝非一件简单的事情。因为这个系统中有多种元素，如编码基因的表达、Non - coding RNA 的表达、表观修饰、变异、肠道菌群还有环境影响因素等，研究这些复杂系统的变化和干扰，构建一个多元素的网络是第 1 步。另外，对这个系统的动力学研究还需要必要的动力学参数和初始条件信息。在系统层次上的方法还需要新的数学模型，对复杂系统的分子机制进行深入了解。

3.4.3 挑战 9：数据的个性化应用和个性化模型 疾病体系是由基因、环境和生活习惯 3 者复杂的相互作用导致的，生物系统是一个异质性和鲁棒性很强的系统，很难开发出一个实用性很强的模型和软件，针对不同的数据结构和复杂系统开发出不同的个性化的模型和软件是精准医学和精准健康管理

的必然挑战。

3.5 人才缺失与跨学科人才教育的挑战

挑战 10：跨学科教育的挑战。深度测序和大数据处理都是新生事物，将深度测序数据应用到临床更需要数学统计、计算机和生物、医学临床领域的多学科交叉高级人才，这些人才的培养往往不是短时间内能够实现的，加上深度测序、大数据处理、生物医学的迅猛发展和学科内容的不断更新，需要有良好的社会机制和教学模式才可能培养出真正有用的人才。

4 结语

本文从深度测序的技术发展出发，讨论这一技术对社会、经济和科学研究范式演变的影响，总结深度测序数据处理在健康领域应用的 10 大挑战，期待对今后的健康大数据科学研究有启示作用。由于世界人口老年化和医疗、人力资源缺乏的矛盾日益加重，各国政府都在推进大数据、大健康的科学和产业发展，促进疾病防控关口前移，医疗资源下沉。健康管理相关的大数据包括基因数据、生活习惯数据、临床表观组学数据和暴露组学数据，这些复杂、动态的数据用来描述生命现象和了解疾病与健康。然而目前最容易得到的是基因测序的数据（包括静态的遗传学和动态的表达数据等），同时也是最重要的表型驱动因素。人们的生活习惯和饮食的偏好也常常与遗传和基因动态表达相关，所以基于深度测序技术的基因数据分析是精准医学和大健康管理的关键因素，解决好深度测序技术产生的大数据存储、分析和应用等方面的挑战无疑是精准医学和健康管理科学的前提和保障。

参考文献

1 Stevens H, Dr Sanger, Meet Mr. Moore: next - generation sequencing is driving new questions and new modes of research [J]. Bioessays, 2012, 34 (2): 103 - 105.

2 Erlich Y. A Vision for Ubiquitous Sequencing [J]. Genome Res, 2015, 25 (10): 1411 - 1416.

3 Mardis E R. Anticipating the 1 000 Dollar Genome [J]. Genome Biol, 2006, 7 (7): 112.

4 Abbasi J. 23andMe, Big Data, and the Genetics of Depression [J]. JAMA, 2017, 317 (1): 14 - 16.

5 Hyde C L, Nagle MW, Tian C, et al. Identification of 15 Genetic Loci Associated with Risk of Major Depression in Individuals of European Descent [J]. 2016, 48 (9): 1031 - 1036.

6 Akaev A A, Sadovnichiy V A. A Closed Dynamic Model to Describe and Calculate the Kondratiev Long Wave of Economic Development [J]. Herald of the Russian Academy of Sciences, 2016, 86 (5): 371 - 383.

7 Wynn J, Chung W K, 23andMe Paves the Way for Direct - to - Consumer Genetic Health Risk Tests of Limited Clinical Utility [J]. Ann Intern Med, 2017, 167 (2): 125 - 126.

8 沈百荣主编. 深度测序数据的生物信息学分析及实例 [M]. 北京：科学出版社, 2017.

9 Vayena E, U Gasser. Between Openness and Privacy in Genomics [J]. PLoS Med, 2016, 13 (1): e1001937.

10 Wang S, Jiang X, Singh S, et al. Genome Privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States [J]. Ann N Y Acad Sci, 2017, 1387 (1): 73 - 83.

11 Huang Z, Anday E, Lin H, et al. A Privacy - preserving Solution for Compressed Storage and Selective Retrieval of Genomic data, 2016, 26 (12): 1687 - 1696.

12 Harris M A, Clark J, Ireland A, et al. The Gene Ontology (GO) Database and Informatics Resource [J]. Nucleic Acids Res, 2004, 32 (Database issue): D258 - 61.

13 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes [J]. Nucleic Acids Res, 2000, 28 (1): 27 - 30.

14 Cimino J J, Ayres E J, Remennik L, et al. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date [J]. J Biomed Inform, 2014, (52): 11 - 27.

15 Tryka K A, Hao L, Sturcke A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP [J]. Nucleic Acids Res, 2014, 42 (Database issue): 975 - 979.

16 Chute C G, Beck S A, Fisk T B, et al. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data [J]. J Am Med Inform Assoc, 2010, 17 (2): 131 - 135.

（下转第 32 页）

认证，在设备初次注册时记录其硬件信息，保障用户身份的合法性。一旦完成身份验证，将通过对通信隧道的加密实现网络之间的加密传输，满足可信要求。隔离访问机制针对不同的异构存储通过强制读写分离进行管控，同时将强制关闭移动设备热点及 WIFI 功能直至客户端安全断开连接。隔离远程客户端在安装过程中将在本地生成一个安全数据拷贝盘，该数据拷贝盘是对私有云存储资源池中客户端历史访问文件的同步拷贝。当客户端和业务服务器进行通信时，数据访问和操作将作用于临时数据拷贝盘，从而减少网络通信压力。离线状态下，同样能够利用客户端访问本地数据拷贝盘读取文件数据。

3 结语

基于虚拟隔离机制的医院私有云存储架构为医院的内外网隔离和互通提供一个便捷、安全可靠的实施方案，改变以往需要配备两套网络的格局，节省了开支。系统立足于私有云存储系统架构，对异构存储资源虚拟化、文件分层组织管理、私有云客户端可信隔离访问、本地安全虚拟桌面构建及分布式目录同步等技术内容进行系统应用。一方面为医疗企业的虚拟化存储云架构实施应用提供一定的实践经验和理论架构，另一方面也是本地安全虚拟桌面构建远程桌面、私有云访问等虚拟化技术的综合应用，对创新型虚拟隔离数据保护环境构建技术虚拟化技术在医疗企业的应用具有一定的理论和现实意义。

参考文献

1 魏智，黄昊. 虚拟存储技术在医院信息化建设中的作用［J］. 中国数字医学，2013，(11)：86－88.

2 李先锋，王凯芸，吕强，等. 三甲医院虚拟化技术的研究与实践［J］. 中国医院，2012，16（2）：12－14.

3 孟群，屈晓晖. 虚拟化技术在医院信息平台服务器整合中的应用［J］. 中国数字医学，2011，6（7）：8－12.

4 孟庆伟，刘婷. 基于 Fstor Phantosys 云桌面虚拟化平台的构建［J］. 计算机安全，2014，(5)：24－27.

5 肖玮炜. 医院信息系统平台建设与存储虚拟化技术研究［J］. 电脑编程技巧与维护，2016，(12)：67－68.

（上接第 11 页）

17 Murphy S N, Weber G, Mendis M, et al. Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2) [J]. J Am Med Inform Assoc, 2010, 17 (2): 124－130.

18 Bhattacharya S, Andorf S, Gomes L, et al. ImmPort: disseminating data to the public for the future of immunology [J]. Immunol Res, 2014, 58 (2－3): 234－239.

19 Payakachat N, Tilford J M, Ungar W J. National Database for Autism Research (NDAR): Big Data Opportunities for Health Services Research and Health Technology Assessment [J]. Pharmacoeconomics, 2016, 34 (2): 127－138.

20 Lowe H J, Ferris T A, Hernandez P M, et al. STRIDE—An integrated standards－based translational research informatics platform [J]. AMIA Annu Symp Proc, 2009, (2009): 391－395.

21 Tomczak K, P Czerwinska, M Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge [J]. Contemp Oncol (Pozn), 2015, 19 (1a): A68－77.

22 Clark K. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository [J]. J Digit Imaging, 2013, 26 (6): 1045－1057.

23 Wang X. Translational Integrity and Continuity: personalized biomedical data integration [J]. J Biomed Inform, 2009, 42 (1): 100－112.