

中国人群参考基因组及基因组变异图谱资源库

宋述慧^{1,2,3}, 滕徐菲^{1,3}, 肖景发^{1,2,3}

1. 中国科学院北京基因组研究所, 中国科学院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: 随着人类基因组计划和国际千人基因组计划的实施, 已公开数百个中国人个体的全基因组数据。建立高精度的中国人群参考基因组序列, 发现并解析中国人群特有的序列变异, 是我国未来精准医学研究的基础。为满足未来精准医学研究中国人基因组数据持续增长的科学管理和深入研究的需求, 中国科学院北京基因组研究所发展并建立了基于中国人群全基因组测序数据的虚拟中国人基因组数据库(Virtual Chinese Genome Database, VCGDB)和中国人群基因组变异数据库(Genome Variation Map, GVM), 面向国内外用户提供数据检索、共享、下载和在线分析服务。本文重点介绍了这两个数据库的特点和功能, 以及未来发展与应用前景, 以期为中国人群参考基因组及基因组变异图谱资源库的推广使用、发展完善提供有益信息。

关键词: 中国人群; 参考基因组; 变异图谱

Database resources of the reference genome and genetic variation maps for the Chinese population

Shuhui Song^{1,2,3}, Xufei Teng^{1,3}, Jingfa Xiao^{1,2,3}

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

Abstract: With the implementation of the international human genome project and 1000 genome project, hundreds of Chinese individual genome sequences have been published. Establishing a high-precision Chinese population reference genome and identifying the unique genome variations are fundamental for future precision medicine research in China. To

收稿日期: 2018-05-24; 修回日期: 2018-09-10

基金项目: 国家自然科学基金项目(编号: 31771465), 中国科学院“十三五”信息化建设专项: 大数据驱动的生物信息领域创新示范平台项目(编号: XXH13505-05)和中国科学院青年创新促进会项目(编号: 2017141)资助[Supported by the National Natural Science Foundation of China (No.31771465), the 13th Five-year Informatization Plan of Chinese Academy of Sciences (No. XXH13505-05) and Youth Innovation Promotion Association (No.2017141)]

作者简介: 宋述慧, 博士, 副研究员, 研究方向: 生物信息学。E-mail: songshh@big.ac.cn

滕徐菲, 在读硕士研究生, 专业方向: 生物信息学。E-mail: tengxufei@big.ac.cn

宋述慧和滕徐菲并列第一作者。

通讯作者: 肖景发, 博士, 研究员, 研究方向: 生物信息学。E-mail: xiaojingfa@big.ac.cn

DOI: 10.16288/j.ycz.18-177

网络出版时间: 2018/9/2 19:52:03

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20180902.1951.002.html>

further meet the needs of scientific management and deep mining on the rapidly growing Chinese genomic data, Beijing Institute of Genomics, Chinese Academy of Sciences, has developed a Virtual Chinese Genome Database (VCGDB, <http://bigd.big.ac.cn/vcg/>) and Genome Variation Map (GVM, <http://bigd.big.ac.cn/gvm/>) based on the public whole genome sequencing data, which provides the worldwide services of data retrieval, sharing, downloading and online analysis. This paper presents the brief introduction of characteristics and functions of the two databases, as well as their future development and application prospects, aiming to provide useful information for the promotion and development of the reference genome and genome variation map database in China.

Keywords: Chinese population; reference genome; variation map

自 1990 年 10 月 1 日启动“人类基因组计划”, 到 2003 年 4 月 15 日, 国际人类基因组组织正式宣布全部完成, 历时 10 多年的国际人类基因组计划绘制了物理、遗传、序列和基因 4 张图谱^[1,2], 开启了人类对自身(包括癌症在内的人类疾病的发生)的深入认识和研究, 推动了测序技术、基因组学和生物信息学等的发展, 并相继启动了国际单倍体型计划(HapMap 计划)^[3,4]、“国际千人基因组计划”^[5]、“肿瘤基因组解剖计划”^[6-8]和“环境基因组学计划”^[9]等一系列与健康相关的研究计划。其中, 2008 年 1 月 22 日启动的“国际千人基因组计划”^[5]是举世闻名的人类基因组计划的延续和发展, 该计划于 2012 年 3 月 29 日完成, 是基因组科学研究向临床医学迈进的重要转折点, 不仅绘制了迄今为止最详尽的、最有医学应用价值的人类基因组遗传多态性图谱, 还贡献了海量的源于不同国家和不同人群的、包含着大量遗传变异信息的个人基因组数据^[10,11]。该计划产生的 392 个中国人(283 个汉族和 109 个少数民族)样本的全基因组测序数据, 为中国人特异的遗传特征和相关医学分析研究提供了宝贵的数据资源。科学家们通过分析, 发现不同人种之间的基因组单核苷酸多态性位点及频率存在明显的差异, 因此许多国家纷纷启动了面向本国或本地区的基因组测序计划, 目标是建立更加精细的参考基因组及变异组。例如, 英国于 2010 年和 2012 年分别启动了英国万人基因组计划和 10 万人基因组计划^[12,13], 旨在通过大规模的基因组测序寻找英国人群特有的基因组变异, 挖掘与健康 and 疾病相关联的遗传风险因素。2016 年, 日本人参考基因组计划(Japanese Reference Genome)通过新一代 DNA 测序技术构建了日本人参

考基因组序列^[14]。此外, 澳大利亚、冰岛、加拿大、新加坡、韩国、荷兰、丹麦、沙特阿拉伯等国家和地区都纷纷启动了相应的基因组计划。中国人要有自己的基因组数据和参考基因组序列, 才能解决中国人特有的疾病遗传问题。2007 年 10 月, 第一个黄种人个人基因序列“炎黄一号”完成^[15], 是首例基于二代测序技术完成的参考基因组序列。2016 年 6 月, 中国人个体基因组“华夏一号”公布, 该个体基因组采用三代单分子测序和二代测序技术相结合, 大幅度提高了基因组组装的完整性和准确性^[16]。

基因组数据的测定为鉴定和研究遗传变异及多态特点提供了基础。国际人类基因组单体型图谱计划(HapMap 计划)测定了全球 11 个人群, 获得约 500 万单核苷酸多态性位点(single nucleotide polymorphisms, SNPs)。国际千人基因组计划对全球不同人类种群的 2500 人进行了全基因组测序, 获得了 8470 万 SNPs、360 万序列插入删除(insertion or deletion)和 6 万结构变异。20 世纪 90 年代以来, 我国也先后启动和实施了“中华民族基因组 SNP 研究”、“中华民族基因组中若干位点基因结构的研究”和“中国人若干群体的基因组多态性研究”等重大项目。这些项目的开展都为创建我国人群遗传资源库打下了重要的基础, 如: 通过对 20 635 例中国人样本的主要组织相容性复合体(major histocompatibility complex, MHC)目标区域进行高深度测序和分析, 建立了世界上最大样本量的中国人 MHC 全区域完整遗传变异数据库^[17], 展示了中国人 MHC 区域突变位点和 HLA 基因的多态性图谱, 为开展中国人复杂疾病与 MHC 区域的相关性研究奠定了坚实的基础。

尽管在人类(尤其是中国人)基因组的解析和发

展中取得了长足的进步,但在基因组研究中广泛用于序列比对分析的人类基因组参考序列,仅是基于有限的人类个体全基因组测序后的结果,这个不包含任何遗传变异信息的静态基因组显然不足以支持高度复杂的基因组学、转录组学、表观基因组学以及全基因组关联分析等研究;此外,目前国际上公开的人类基因组变异数据也主要来源于西方白种人,利用这些变异数据作为参比数据,常造成我国基因组研究和临床应用结果的不准确。面向未来中国精准医学研究的新需求,中科院北京基因组研究所发展并建立了基于中国人群全基因组测序数据的虚拟中国人基因组数据库(Virtual Chinese Genome Data Base, VCGDB)^[18, 19]和基因组变异数据库(Genome Variation Map, GVM)^[20]资源(图 1),有效并全面展示了中国人群的遗传变异特征,更好服务于中国的人类遗传学、基因组学和生物医学的研究和应用。

1 虚拟中国人基因组数据库

国际千人基因组计划提供了丰富的全基因组测序数据资源,其中包含中国南方汉族人群数据(Sou-

thern Han Chinese, CHS)、北方汉族人群数据(Han Chinese in Beijing, CHB)以及中国西双版纳傣族的中国人基因组数据(Dai Chinese in Xishuangbanna, CDX)。为了充分利用这些信息,选取该计划中包含中国南方人群和北方人群数据共计 194 个高覆盖度个体的全基因组序列数据,通过标准化数据分析和处理流程^[18],构建了虚拟中国人基因组数据库(VCGDB, <http://bigd.big.ac.cn/vcg/>) (图 2)。VCGDB 提供了中国人群基因组多态性信息,共包括 3500 万个单核苷酸变异位点信息(SNPs)、50 万个基因组插入删除片段信息、2900 万个罕见变异位点信息,及其对应的基因组注释信息^[18]。同时 VCGDB 还分别提供了中国人群、南方人群体和北方人群体的一致性基因组参考序列。此外,通过真实的基因组测序数据序列比对分析,将其与已有的人类基因组参考序列以及“炎黄一号”进行比较,表明基于中国人群体高频遗传变异位点构建的中国人基因组一致性参考序列更能体现中国人群体的基因组特征。

虚拟中国人基因组数据库具有以下特点:

(1) VCGDB 是一个“动态”的数据库,通过信息熵等方法来计算中国人群体之间各个位点遗传变异的动态变化水平和发生率,能够展示基因组中不同位

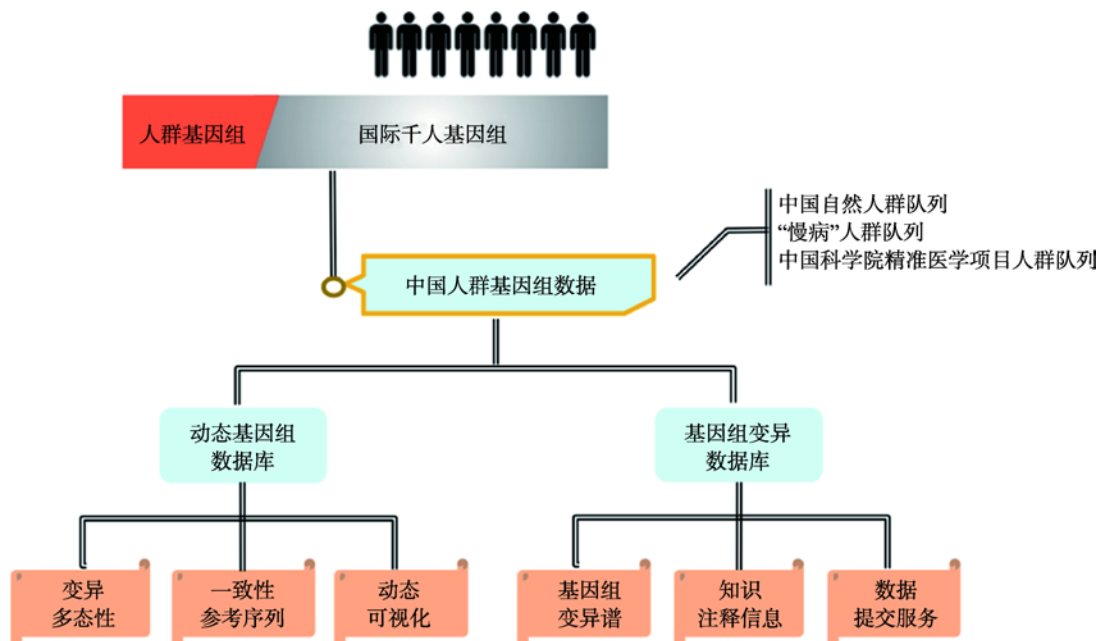


图 1 中国人群参考基因组和变异组数据库建立示意及主要特征

Fig. 1 Schematic for Chinese reference genome and variome databases and their main characteristics

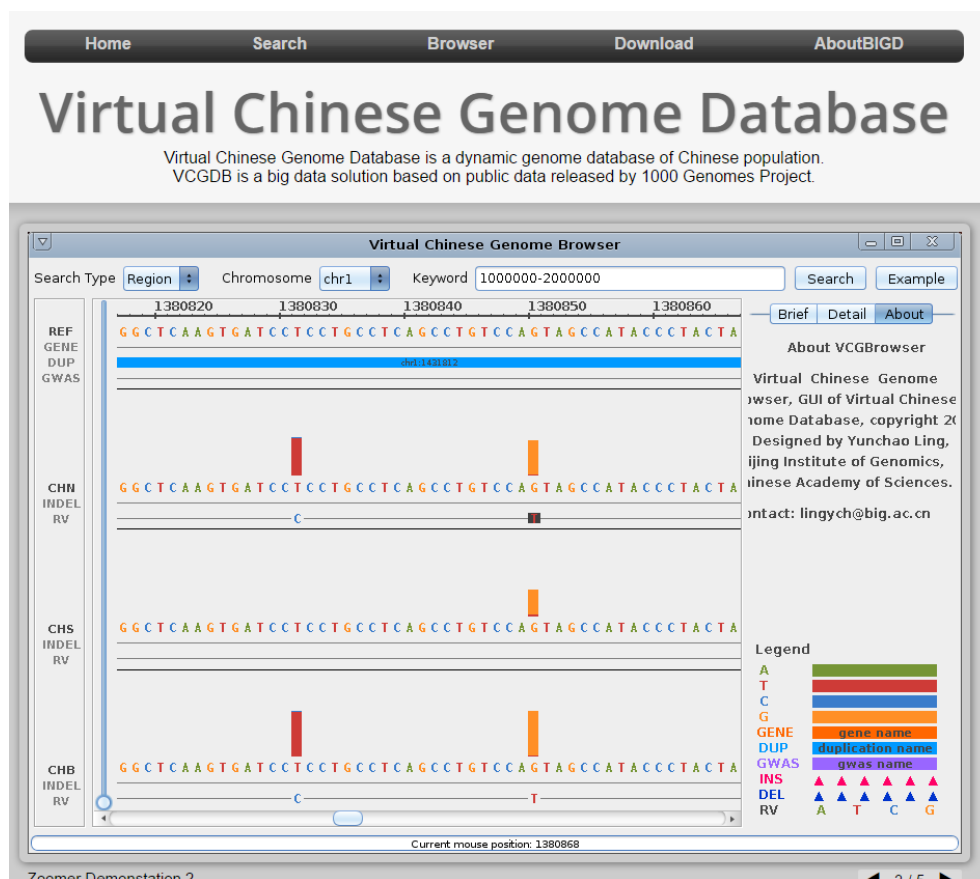


图 2 虚拟中国人基因组数据库主页

Fig. 2 A screen shot of the home page of VCGDB

点的遗传变异多态性信息和各位点不同基因型的发生频率信息 ;(2) VCGDB 是一个“虚拟”的数据库,通过整合中国人群体高频遗传变异位点信息,以标准参考基因组为参照,分别构建了中国人群体、南方人群体和北方人群体的一致性基因组参考序列。构建的一致性基因组参考序列并不属于和代表任何一个真实存在的个体,而是源于对 200 多个个体 TB 级大规模数据进行综合分析的结果,也因此可以更好地描述中国人群体的遗传变异特征 ;(3) VCGDB 提供高度交互的、友善的、融合多种全新功能的中国人动态基因组浏览器(VCGBrowser),相较于传统的基因组浏览器如 UCSC Genome Browser 和 JBrowse Genome Browser 并不能显示群体的基因组动态信息,VCGBrowser 可根据用户的不同需求,从染色体、固定片段、指定基因和指定位点等多层次展示所有位点在不同群体的位点动态信息以及相关的基因组注释信息。总体上,虚拟中国人基因组数据库实现了

对国际千人基因组计划中中国人群基因组测序数据的精细整合分析,并提供了中国人群体基因组变异的动态信息,为今后开展大规模人群基因组测序数据的分析和展示提供了参照^[18]。

2 中国人群基因组变异数据库

虚拟中国人基因组数据库的建设为中国人基因组数据比对分析提供了较为精准的参考基因组,为进一步满足基于变异位点基因型和表型的关联分析及知识发现的研究需求,在 VCGDB 的基础上又发展和构建了中国人群基因组变异数据库(Genome Variation Map, GVM) (<http://bigd.big.ac.cn/gvm>, 图 3)^[20]。利用国际千人基因组计划中的 215 个(测序覆盖度>5)全基因组序列数据,采用统一的变异位点鉴定和注释分析流程^[20],提供了截至日前最全的中国人群变异位点、人群频率和位点知识的注释信息,

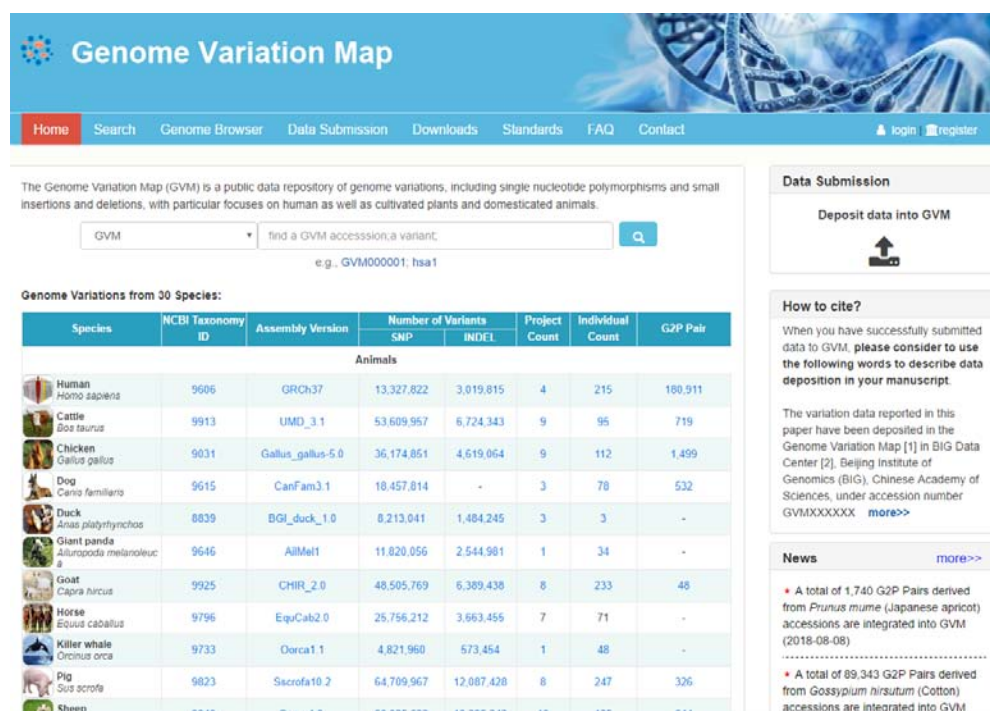


图 3 基因组变异数据库主页

Fig. 3 A screen shot of the home page of GVM

共包括 13327822 个单核苷酸变异位点信息(SNPs)、3 019 815 个基因组插入删除片段信息、16 739 583 个低发生概率(minor allele frequency, MAF <0.05)的变异位点信息, 5 343 882 个罕见发生概率(MAF <0.005)的变异位点信息, 以及与这些位点和序列片段相关的基因组注释信息, 包括位点突变效应、临床表型效应、人类孟德尔遗传疾病效应等。

GVM 数据库中的中国人群变异模块具有以下特点: (1)提供了强大实用的变异数据检索服务, 支持用户根据变异位点编号、变异类型、变异效应、基因名称、基因功能、已知临床效应、遗传病和其他表型等条件进行组合检索, 初步检索结果还可以进一步根据上述条件组合进行二次过滤。检索结果以表格形式返回, 并提供了变异位点的详细注释信息、各个体的基因型在线浏览、全部检索结果条目的下载服务等; (2)提供了高度交互的、友善的、融合多种功能和信息的基于 GBrowser 技术的在线浏览功能, 支持用户自主选择感兴趣的个体和区间, 提供统一的坐标系直接展示和比较全基因组水平的所有动态变异信息, 可以根据用户需求缩放至单碱基水平, 从基因组的水平展示某个变异位点的频率

信息和其他详细细节; (3)提供了数据递交服务, 支持用户选择在线或离线两种不同的方式递交数据, 用户通过创建 BioProject 填写数据元信息, 系统对所提交变异数据自动分配唯一编号(GVMXXXXXX), 该编号可以直接应用于科技论文的数据获取。总体上, 中国人群基因组变异数据库实现了对国际千人基因组计划海量中国人数据的完整、全面的整合和高效展示, 体现了中国人群的变异特征。是未来精准医学研究表型与基因型关联分析的重要基础, 为未来我国精准医学队列人群计划大数据的处理和分析、数据管理等提供了示范指导, 也为基于基因组序列变异的遗传检测、药物研发等提供数据支持。

3 未来展望

国际大型基因组研究计划如 HapMap 和千人基因组计划等虽已将汉族样本作为主要亚洲人群进行了研究。然而, 这些重要参考数据却存在着很大的局限: 一方面样本量较少, 因而对于低频基因组多态性的代表性差; 另一方面对应样本没有表型信息, 无法将遗传多态性与表型进行有效关联。中国科学

院于 2015 年率先启动了“中国人群精准医学研究计划(中科院)”重点部署项目,已产出上千人的高覆盖度的全基因组测序数据,并已经提交到中国生命与健康大数据中心^[21]的组学原始数据库 GSA 中管理^[22]。2016 年,我国“十三五”重点研发计划中已设置了为实施精准医学研究而构建百万人以上的自然人群国家大型健康队列和重大疾病专病队列的项目,随着国家“慢病”、“精准医学”大型人群队列项目的启动,未来将产生百万人群的多达 EB 级组学数据。这些人群队列的基因组数据将进一步丰富和完善中国人群的参考基因组和基因组变异数据。此外,VCGDB 和 GVM 数据库将为未来大规模人群队列基因组数据的汇交、存储、管理、共享与分析提供支持和指导。

国家“慢病”、“精准医学”大型人群队列项目,未来不仅产出多层次的组学分子数据,还将产出海量多维的生物医学数据,包括基线数据、随访和临床表型组数据等。未来将建立包含多层次多维度信息数据的 VCGDB 和 GVM 数据库,开展大数据驱动的创新应用研究和疾病防治方案的精准化研究,助力于发现疾病或表型、用药等遗传关联位点,为生物医学数据转换为能够支撑临床决策的辅助诊疗信息,服务大众实现疾病预测和预警提供重要基础性保障,真正实现数据到信息和知识的转化。

发展可用于精准医学研究组学数据分析的大数据系统的应用体系和解析体系,开发及完善基因组分析算法及软硬件,开发基于通路分析和网络模块的基因型-表型深度解析方法,形成可用于海量基因型-表型数据解析流程、汇交管理、数据挖掘整合的算法体系,并在 VCGDB 和 GVM 中为临床和其他科研人员提供完善的在线分析和研究系统,助力我国的精准医学研究与应用。

参考文献(References):

- [1] Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. *Science*, 2003, 300(5617): 286–290. [DOI]
- [2] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 2004, 431(7011): 931–945. [DOI]
- [3] International HapMap Consortium. The international HapMap project. *Nature*, 2003, 426(6968): 789–796. [DOI]
- [4] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 2005, 437(7063): 1299–1320. [DOI]
- [5] 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]
- [6] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 2008, 455(7216): 1061–1068. [DOI]
- [7] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 2011, 474(7353): 609–615. [DOI]
- [8] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 2013, 45(10): 1113–1120. [DOI]
- [9] Wilson SH, Olden K. The environmental genome project: phase I and beyond. *Mol Interv*, 2004, 4(3): 147–156. [DOI]
- [10] Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, Eichler EE. Diversity of human copy number variation and multicopy genes. *Science*, 2010, 330(6004): 641–646. [DOI]
- [11] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*, 2015, 526(7571): 68–74. [DOI]
- [12] Parry V. Commit to talks on patient data and public health. *Nature*, 2017, 548(7666): 137. [DOI]
- [13] 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]
- [14] 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. [\[DOI\]](#)
- [15] Wang J, Wang W, Li RQ, Li YR, Tian G, Goodman L, Fan W, Zhang JQ, Li J, Zhang JB, Guo YR, Feng BX, Li H, Lu Y, Fang XD, Liang HQ, Du ZL, Li D, Zhao YQ, Hu YJ, Yang ZZ, Zheng HC, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan JJ, Zhou Y, Qin JJ, Ma LJ, Li GQ, Yang ZT, Zhang GJ, Yang B, Yu C, Liang F, Li WJ, Li SC, Li DW, Ni PX, Ruan J, Li QB, Zhu HM, Liu DY, Lu ZK, Li N, Guo GW, Zhang JG, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su YY, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng HK, Ren YY, Yang L, Gao Y, Yang GH, Li Z, Feng XL, Kristiansen K, Wong GKS, Nielsen R, Durbin R, Bolund L, Zhang XQ, Li SG, Yang HM, Wang J. The diploid genome sequence of an asian individual. *Nature*, 2008, 456(7218): 60–65. [\[DOI\]](#)
- [16] Shi LL, Guo YF, Dong CL, Huddleston J, Yang H, Han XL, Fu AS, Li Q, Li N, Gong SY, Lintner KE, Ding Q, Wang Z, Hu J, Wang DP, Wang F, Wang L, Lyon GJ, Guan YT, Shen YF, Evgrafov OV, Knowles JA, Thibaud-Nissen F, Schneider V, Yu CY, Zhou LB, Eichler EE, So KF, Wang K. Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun*, 2016, 7: 12065. [\[DOI\]](#)
- [17] Zhou FS, Cao HZ, Zuo XB, Zhang T, Zhang XG, Liu XM, Xu RC, Chen G, Zhang YW, Zheng XD, Jin X, Gao JP, Mei JP, Sheng YJ, Li QB, Liang B, Shen J, Shen CB, Jiang H, Zhu CH, Fan X, Xu FP, Yue M, Yin XY, Ye C, Zhang CC, Liu X, Yu L, Wu JH, Chen MY, Zhuang XH, Tang LL, Shao HJ, Wu LM, Li J, Xu Y, Zhang YJ, Zhao SL, Wang Y, Li G, Xu HS, Zeng L, Wang JN, Bai MZ, Chen YL, Chen W, Kang T, Wu YY, Xu X, Zhu ZW, Cui Y, Wang ZX, Yang CJ, Wang PG, Xiang LH, Chen X, Zhang AP, Gao XH, Zhang FR, Xu JH, Zheng M, Zheng J, Zhang JZ, Yu XQ, Li YR, Yang S, Yang HM, Wang J, Liu JJ, Hammarstrom L, Sun LD, Wang J, Zhang XJ. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat Genet*, 2016, 48(7): 740–746. [\[DOI\]](#)
- [18] Ling YC, Jin Z, Su MM, Zhong J, Zhao YB, Yu J, Wu JY, Xiao JF. VCGDB: a dynamic genome database of the Chinese population. *BMC Genomics*, 2014, 15: 265. [\[DOI\]](#)
- [19] 凌望超. 虚拟中国人动态基因组数据库[学位论文]. 中国科学院大学, 2014. [\[DOI\]](#)
- [20] Song SH, Tian DM, Li CP, Tang BX, Dong LL, Xiao JF, Bao YM, Zhao WM, He H, Zhang Z. Genome variation map: a data repository of genome variations in BIG Data Center. *Nucleic Acids Res*, 2018, 46(d1): D944–D949. [\[DOI\]](#)
- [21] BIG Data Center Members. Database Resources of the BIG Data Center in 2018. *Nucleic Acids Res*, 2018, 46(d1): D14–D20. [\[DOI\]](#)
- [22] Wang YQ, Song FH, Zhu JW, Zhang SS, Yang YD, Chen TT, Tang BX, Dong LL, Ding N, Zhang Q, Bai ZX, Dong XN, Chen HX, Sun MY, Zhai S, Sun YB, Yu L, Lan L, Xiao JF, Fang XD, Lei HX, Zhang Z, Zhao WM. GSA: genome sequence archive. *Genomics Proteomics Bioinformatics*, 2017, 15(1): 14–18. [\[DOI\]](#)

(责任编辑: 赖江华)