

千人基因组视角下的人类起源演化

程耀霆, 李玉春, 孔庆鹏*

中国科学院昆明动物研究所, 遗传资源与进化国家重点实验室, 昆明 650223

* 联系人, E-mail: kongqp@mail.kiz.ac.cn

2014-11-20 收稿, 2015-01-27 接受, 2015-03-11 网络版发表

国家自然科学基金(81272309, 31322029)资助

摘要 千人基因组工程作为大规模基因组测序时代的一个里程碑计划, 与此前的人类基因组计划和人类基因组单体型图谱计划一起成为推动人类遗传学研究发展的重大公共科学事件. 利用该计划所产生的人类群体全基因组测序数据, 研究人员可以从一个前所未有的深度探寻人类遗传多样性与人类起源、演化、疾病之间的关系. 本文从千人基因组计划发展的历史轨迹以及利用这一大规模数据集所开展的人类遗传学研究成果等方面入手, 对大数据时代背景下人类遗传演化研究的特点进行了梳理和总结, 并对未来基于如千人基因组这类大规模人类遗传数据工程的工作做了展望.

关键词

千人基因组计划
人类遗传
起源演化
基因组测序
大数据

人类的起源、演化和疾病一直是遗传学研究关注的焦点. 早期, 受技术水平的限制, 针对人类遗传多样性的研究往往十分粗糙. 此后, 随着遗传学技术的不断发展, 越来越多先进的分子技术被应用于遗传学检测, 从单个位点检测, 到Y染色体、线粒体DNA分型检测, 再到后来的基因组芯片技术, 研究人员所掌握的遗传数据信息量越来越大. 这些数据的不断积累给人类进化遗传与疾病遗传研究提供了前所未有的重大机遇. 然而, 这些进展同时也让人们意识到, 对人类基因组全序列的测定以及人类群体遗传多态性分布图谱的绘制是所有后续深入研究的前提基础. 但是考虑到实际条件, 无论是样本收集还是数据产生与分析都是某一实验室甚至某个国家无法单独完成的, 而全球化进程给这种大规模遗传工程提供了一个千载难逢的国际合作机遇. 在该背景下, 不同国家的研究人员根据不同时期遗传学检测手段的发展状况以及自身需求联合开展了一系列大型人类遗传检测计划, 以期能够全面、细致、权威地对人类遗传基础以及群体遗传多态性进行揭示.

2003年, 历时13年耗资数十亿美元的人类基因

组计划(human genome project, HGP)宣布完成了人类基因组全图的绘制, 这项20世纪80年代开始实施的庞大计划, 使人们首次能够系统完整地研究人类基因组的结构和特点, 同时也为以后的遗传学研究提供了一个核心蓝图, 而该计划的成功完成也从一定程度上推动了后续一系列大型遗传学研究工程的开展. 随着基因芯片技术的发展以及全基因组关联研究(genome-wide association study, GWAS)的不断深入, 大量遗传数据得以积累. 2002年, 由美、加、中、日、英、尼日利亚等国科学家共同提出并发起了“国际人类基因组单体型图谱计划”(International Hap-Map Project)^[1], 该项计划与更先一步启动的人类基因组多样性计划(human genome diversity project, HGDP)^[2]一起为研究人员提供了覆盖全球范围人类群体的基因组群体遗传数据, 从而将HGP工作推向深入化, 这些工作通过对涵盖欧、亚、非、北美、南美等地区人群的样本进行全基因组遗传分析, 绘制了全球人群遗传多态性以及单倍型差异的图谱. 这些数据发布之后, 利用这些数据作为平台对人类进化遗传学以及疾病遗传学展开了大量人类遗传学研

引用格式: 程耀霆, 李玉春, 孔庆鹏. 千人基因组视角下的人类起源演化. 科学通报, 2015, 60: 1057–1065

Cheng Y T, Li Y C, Kong Q P. Application of the 1000 genomes project in human evolutionary genetics (in Chinese). Chin Sci Bull, 2015, 60: 1057–1065, doi: 10.1360/N972014-01245

究研究,取得了一批重要的成果.

HGP, HapMap和HGDP等计划在此前为研究人员提供了丰富的人类遗传数据,让研究人员可以有机会详尽地研究人类基因组的一些演化特点和功能作用,但是这些工作仍然存在一些缺憾.(i) 尽管作为同一物种的遗传差异相较于种间差异变化很小,但是人类不同群体之间的表型和遗传差异仍然不可忽视,不同人群面对不同的生活环境,其遗传特征也可能不尽相同,这就使得种内遗传多态性分布模式成为了一个极为重要的影响因素.而HGP作为单个人类基因组测序计划,显然无法对人类多态分布规律提供更多的群体信息;(ii) HapMap和HGDP计划的基础数据来自于基因芯片,无法对稀有突变进行检测,而且芯片的设计没有考虑不同人群之间的遗传差异,这可能会导致大量信息丢失;同时,基因芯片也存在分辨率较低等问题.由于这些问题的存在,使得对于建立高覆盖度人群基因组测序数据库的需求愈发迫切,加之新一代基因组测序技术的快速发展以及测序成本的不断下降等技术背景,促使大规模组织人类基因组测序成为可能.

1 千人基因组计划的发展历程与基本数据产生

2008年,由中国华大基因、英国Sanger研究所及美国国立人类基因组研究所3家机构联合启动了“千人基因组计划”(1000 genomes project, 1KG),并在随后几年中合作完成了该计划的一部分,目前分为几期公布数据.千人基因组计划可以被看作是人类基因组计划的后期工程.该计划主要目的是通过新一代高通量测序技术对全球范围内的人类群体进行全基因组测序,以期从更深层次上绘制人类基因组多态性分布水平以及稀有突变的情况,这个工作对医学遗传学、进化遗传学、群体遗传学等人类遗传学分支学科有着巨大的指导意义.

2010年,千人基因组计划发布了首阶段数据,数据的产生主要利用3种高通量测序技术完成(Illumina genome analyzer, AB SOLiD system和454 GS FLX)^[3].该数据主要包括3部分:(i) 家系全基因组深度测序数据(Trio project).利用全基因组测序法对2个家系的6个个体进行了高深度测序,这2个家系分别来自尼日利亚的约鲁巴人家庭(Yoruba in Ibadan, YRI)以及来自犹他州的欧洲族源家庭(Utah residents with

northern and western European ancestry, CEU),平均测序深度42x;(ii) 人群低覆盖度全基因组测序(low-coverage project).主要针对4个不同族源人群无关个体的全基因组测序数据,包括59个非洲个体(YRI),60个欧洲个体(CEU),30个北京汉族个体(Han Chinese in Beijing, CHB),30个日本东京个体(Japanese in Tokyo, JPT),平均测序深度仅为2x~6x;(iii) 高深度外显子组测序数据(exon project).测定了7个人群的697个个体906个随机选择基因的外显子序列,平均深度达到了50x.依照这些测序数据,研究人员最终得到了1500万左右的单核苷酸多态性位点(single nucleotide polymorphisms, SNPs),100万的短片段的插入缺失(short insertions and deletions)及2万左右的大规模的结构变异(structural variants),而且这些基因组多态性多数未在以往报道中被发现.

继而在2012年,千人基因组计划在*Nature*杂志上发表了第1期的数据分析结果^[4].相较于2010年发布的首阶段数据,2012年发布并开始释放的第1期数据规模更为庞大,一共释放了代表全世界14个族群的1092个个体的全基因组测序数据以及外显子组测序结果.这些样本能够较好地覆盖世界范围内的主要几个代表人群:欧洲人群(CEU, TSI(Toscani in Italy), GBR(British in England and Scotland), FIN(Finnish in Finland)和IBS(Iberian populations in Spain)),非洲人群(YRI, LWK(Luhya in Webuye)和ASW(African Ancestry in Southwest US)),东亚人群(CHB, JPT和CHS(southern Han Chinese))以及美洲人群(MXL(Mexican Ancestry in Los Angeles), CLM(Colombian in Medellin)和PUR(Puerto Rican in Puerto Rico)).利用大规模测序结果,研究人员最终鉴定出3800万个单核苷酸多态性位点、140万个短片段插入缺失,以及超过1.4万个大片段缺失.未来千人基因组计划最终的数据量将会达到26个群体,2500多个个体的测序数据.届时,千人基因组数据基本覆盖了全世界范围内的所有主要人群,为人类遗传学研究提供前所未有的大规模数据集.

2 千人基因组数据的应用

以国际化、大样本量、高通量、开放性等特点,千人基因组计划提供了一个大规模标准化的人类基因组多态性基础数据库,它有效地弥补了前千人基因组时代大型公共遗传学数据的很多短板和缺陷,

为人类进化遗传学、医学遗传学、生物信息学等学科提供了一个开放的大规模人类基因组遗传信息数据集。借助这一公共数据集,从事遗传学相关分支领域研究的不同研究人员可以灵活地根据自身工作目的以及实验设计而利用千人基因组数据集进行数据分析。正是由于其具有目前为止对人类群体基因组数据最为广泛的代表性,自释放之日起,不同学科利用这些数据开展了大量研究,取得了令人瞩目的研究成果。这些研究使人们能够更加全面深入了解人类现今群体遗传多态性分布模式、人群进化历史、自然选择压力等信息^[3-17],同时也给其他物种及古人类学研究提供了一个极为良好的参照系^[18-20]。而在利用这一数据集开展研究的过程中,对研究策略、遗传参数计算、人类基因组多态性水平衡量等方面的革新产生了巨大推动作用。

2.1 千人基因组数据应用与分析方法的改进

对千人基因组数据的有效利用离不开数据分析方法的革新。在前千人基因组时代,数据通量较小,所需处理的遗传多态性信号数量也并不大,所以更多的进化遗传学、医学遗传学分析方法和手段主要针对较低通量的遗传数据而设计。但这些数据分析手段是否仍然适用于大规模基因组测序数据集,成为能否有效利用千人基因组数据展开研究工作首要问题。

千人基因组数据集提供了迄今最全面的人类群体遗传信息数据库^[3,4],然而由于其低覆盖度、较多稀有突变等因素导致传统分析方法出现难以定位低频突变、检测效力降低、均一化处理困难等问题^[21-24]。结合针对低覆盖度数据特点所设计的一些算法,同时利用位点之间的连锁关系,可以采用低覆盖度数据进行较为准确的推算,从而保证多态性位点读取的准确性,而这一思路被广泛应用于千人基因组数据的分析之中^[3,23]。千人基因组数据分析过程中另一个问题是如何利用低覆盖度测序数据准确分析群体遗传结构、自然选择压力等参数。许多不同的统计方法和模型被用于提高基因组测序数据分析结果的精确性。例如,利用最大似然法(maximum likelihood, ML)和贝叶斯法对群体水平等位基因频率进行估算^[25];利用贝叶斯法对千人基因组低覆盖度测序数据进行分析,估算祖先人群数量、分化时间、迁移率等参数^[9];利用单倍型共享水平对千人基因组数据进行估算,推测人类群体历史^[26]。另一方面,较为传

统的如主成分分析(principal component analysis, PCA)、固定系数(fixation index, Fst)等方法也被用于高通量测序数据的群体遗传学分析。对不同频率多态性位点进行分析之后,研究人员发现,PCA对于较高频率的多态性位点有较好的分析能力^[13],而针对低覆盖度高通量测序数据进行相应的调整之后,这些传统方法可以较好地地区分人类群体遗传关系^[27]。另外一个受数据和方法制约的遗传学研究领域是自然选择,在人类群体基因组测序数据为代表的大数据时代,如何有效、准确、精确定位受自然选择压力的区段成为很多研究人员需要面对的新课题^[28]。利用千人基因组数据集,研究人员对传统自然选择检测方法进行了更新和整合,使得人们可以更加准确地检测基因组上不同区域所受到的不同选择压力水平^[6,16,28,29]。

2.2 千人基因组数据的应用与人类的起源、分化研究

有效利用千人基因组数据,使得研究人员可以更加全面深入地开展针对人类起源、分化、适应、疾病等问题的研究,并取得了大量的新成果。

基因组重组对人类及其他物种的健康以及物种差异起重要作用,由于这种变化属于大尺度基因组突变,往往会导致基因组片段的插入、缺失甚至是染色体级别的变化^[30-32]。人类基因组与其他灵长类近亲在重组率、重组热点等方面存在显著的差异^[33-35],然而,由于人类基因组数据积累较多,使得相关方面的研究主要集中于人类,其他灵长类的研究相对滞后^[36-39]。2012年,Abecasis等人^[3]以千人基因组数据中的欧洲(CEU)与非洲(YRI)个体为参照对10个西部黑猩猩(*Pan troglodytes*)的深度高通量测序数据进行了研究^[40],结果显示,人类与黑猩猩的重组热点区域无重叠。利用千人基因组家系数据可以估算,人类基因组的位点突变频率在 1×10^{-8} 左右,而这一参数对灵长类的进化时间同样有重要的指导和借鉴意义。结合化石记录,利用千人基因组数据所估算的突变率与灵长类世代时间进行计算,可以有效计算出不同物种的分化时间^[11,41,42]。利用这一方式,人们确定了人类-黑猩猩以及人类-黑猩猩-大猩猩(*Gorilla gorilla gorilla*)的分化时间为600~800万年及1000万年左右,并利用相同的思路确定了黑猩猩物种内部的分歧时间以及遗传交流的模式^[41,42]。

2010年,发布了2个隶属于完全不同种类的古人类全基因组测序数据:尼安德特人和丹尼索瓦人^[18,43,44],为人类遗传学研究提供了一个极为重要的外群参照.这些研究发现,非非洲人群与尼安德特人有1%~4%的遗传共享^[45].古人类与现代人类之间的分化关系此前存在2个模型:早期非洲子群体模型与后期基因交流模型^[43~45].前者认为,尼安德特人群与现代人祖先一同在非洲发生分化后分别迁出非洲大陆,因此由于同源关系导致遗传组分相近;而后一种模型则认为,古人类群体与现代人类群体在非洲之外产生了基因交流而导致存在遗传共享组分.利用千人基因组数据对欧洲人群的古人类单倍型组分进行分析,研究人员估算出,距今4.7~6.5万年2群体发生了基因交流,支持后期基因交流模型^[20,46].相似地,其他研究也利用D statistic方法推测证实,现代人类祖先与古人类间存在一定的基因交流,进而导致了遗传组分的共享^[45].与尼安德特人相似,丹尼索瓦人也与非洲之外的人群存在一定的遗传共享水平,特别在东南亚和澳洲人群中尤为显著,与千人基因组数据的比较分析表明,这种遗传上的共享有可能是后期基因交流的结果^[19,20,47].

千人基因组数据更多的是被用于人类自身群体遗传学以及群体历史研究中.在重建人群历史的过程中,分子钟对人群分化及扩张时间的估算起着至关重要的作用.而由于受到数据和方法的影响,人类遗传突变速率的估算还存在着较大争议,这导致估算人类群体历史及分化时间产生了较大差异^[48].千人基因组数据对突变速率的重新估算提供了强大的数据支撑,对解决这些争议提供了良好契机.例如,早期科伊桑人(Khoi-San)的分化时间依据不同分子标记所估算的时间存在较大差异,利用线粒体DNA和Y染色体的估算结果为12~25万年,而核基因的估算结果(25~30万年)则早于该时间^[48~50],但此后利用千人基因组数据重新进行估算得到南部非洲科伊桑与其他人类群体的分化时间为10.8~15.7万年^[9].另一个存在争议的是非洲人群与非非洲人群的分化时间,以往的研究所估算分化时间在6万年左右,该结果受到来自中东地区以及阿拉伯半岛的古DNA证据(约10万年前)的质疑.尽管在Scully等人^[41,48]的研究中,利用较慢的突变速率所计算出的分歧时间与后者吻合,但是这一争议似乎仍然没有停止.在随后基于千人基因组数据的研究中,非洲群体与非非洲群体的分

歧时间为3.8~6.4万年^[9,10,51],在此后一段时期内,各群体之间仍然存在基因交流^[9,10,26].

2.3 千人基因组数据的应用与人类的适应性演化研究

千人基因组数据在人类进化研究方面另一个重要应用方向是人类对生存环境的遗传适应机制研究.人类对环境的适应性进化一直是人类进化遗传学一个重要的研究分支并取得了许多重要研究成果^[52~56],然而由于早期研究数据不足导致研究结果存在可重复性较低、假阳性率较高、精度较差等问题^[8,16,57].千人基因组数据释放后,许多研究组利用这一大数据集的优势对人类受自然选择压力的信号进行了筛查,同时对以往研究结果可能遇到的偏差进行了检验^[6,8,16,28,58~62].人群历史事件、背景选择、基因重组、数据深度、检测方法等因素对于自然选择压力检测的准确性存在一定影响,利用千人基因组数据可以在一定程度上估算这些遗传因素对结果的影响^[8,28,63],同时有效排除这些因素的干扰^[6,8,58].另一方面,借助于千人基因组的大数据集,研究人员可以更加深入地研究自然选择与基因功能、表达、调控之间的关系^[8,60,64,65].2011年,Lappalainen等人^[60]利用千人基因组多态性数据以及相应样本的RNA测序数据开展了一系列的分析,结果发现,调控区与编码区受到较强的净化选择作用以保证较少有害突变的积累.此后的研究发现,人类基因组上大量受到正选择、净化选择的区域位于调控区^[8,64,65],针对microRNA的研究也发现其结合区往往在群体间存在显著差异,而这些差异也与一定的生存条件存在直接关系^[65],这些结果恰恰验证了King和Wilson^[66]提出的自然选择作用主要影响调控区域而非基因编码区的观点.

2.4 千人基因组数据的应用与稀有多态性位点

前千人基因组时代由于受检测方法的限制,其更多地围绕一些具有较高发生频率的多态性位点展开研究,而千人基因组数据第一次能够让研究人员以群体为单位审视稀有变异、低频率多态位点对于人类进化遗传的影响.

相较于高频率多态位点,稀有多态位点的发生更加分散,而且受较低净化选择压力的影响,导致一些有害的稀有突变出现在某些基因上,从而影响这些基因的功能.正是由于这个特点,使得稀有多态位

点逐渐取代传统的高频常见突变成为近年来疾病遗传学研究的焦点^[23,67,68]。然而,检验稀有多态位点的成本远高于检验高频位点:(i)检验需要利用高通量测序才能有效覆盖基因组范围内的所有位点;(ii)对稀有变异的检验需要大样本量筛查后才能确定一个群体内这些位点的水平。而千人基因组计划的提出和开展恰恰满足了以上2个严苛的条件,成为研究人类群体水平稀有变异的第一手也是最重要的资料^[3,4,21]。利用高通量测序技术对人类群体进行大规模基因组测序分析后发现,基因组上绝大部分变异都处于低频状态,而这一结果也符合群体遗传学对于突变分布的预期。研究人员利用千人基因组群体数据分析稀有突变在人类群体基因库中的分布特点时发现,稀有变异在不同人群中呈现出极为显著的差异分布状态。人群依照亲缘距离的远近,稀有位点的共享水平逐渐增高,但是,即便是亲缘较近的2个群体仍然存在较为显著的差异,甚至这种稀有变异大部分为单一群体所特有^[22,24]。而且这种稀有变异的积累与基因组上某些区域受到的自然选择压力存在一定正相关关系,其中进化保守区、调控区、基因区等区域积累稀有变异的水平最低^[24]。而稀有位点占据了人类基因组中多态位点的绝大多数(86%)^[67],其中具有功能影响的占另一个相当大的比例,这对于人类健康水平以及人类群体进化都有较为深远的影响^[67,68]。稀有位点的大规模积累被认为与人类群体数量有着直接关系,因此,人类爆炸式的人口增长对基因库中稀有多态性位点的迅速积累有直接的推动作用^[25]。

3 未来对千人基因组数据的应用与展望

前文对千人基因组的发展历程及其在应用过程中对遗传学研究带来的一些巨大变化进行了粗浅地梳理和总结。千人基因组计划的顺利开展及大规模数据的公开对许多人类遗传学领域的深入研究起巨大的推动作用。有了这样一批高质量的遗传数据,使得不同研究团队根据其实际需求基于同一批数据开展各自不同方向的研究,这样大大节约了研究成本,加快了研究进度。就目前的数据挖掘程度及千人基因组数据的释放程度来看,仍然有很多有趣的工作有待发掘。

尽管千人基因组数据集具有前所未有的人类群体代表性,但是仍然存在一些重要的缺失环节。具体

来说,主要是样本群体代表性以及测序覆盖水平2个问题,这给人类群体历史研究、群体遗传结构研究、医学遗传学研究都造成一定制约。(i)为了弥补千人基因组计划人类群体样本覆盖的缺失环节,一些研究机构针对各自国家的人群情况开展了各种人群高通量测序计划,最为熟知的是新加坡研究机构开展的对于马来西亚人、印度人的测序计划。这些计划利用高通量测序手段对所遴选出的代表人群样本个体进行深度测序(30×),利用千人基因组数据作为参照系对这些群体进行分析,最终发现印度群体拥有十分高的遗传多态性而且在基因组的遗传关系上与欧洲群体相距较远,然而在线粒体DNA水平上与欧洲群体存在一定的关联性^[14,47]。由此可见,在千人基因组基础之上对更多未被覆盖人群进行研究成为了不可或缺的后继工作;(ii)由于千人基因组计划的主体数据来自低覆盖度的高通量测序检测,这必然导致多态位点读取结果存在较高的错误率。前文提到很多研究团队针对这一缺失环节发展了许多研究方法以提高研究结果的准确性,然而这些努力依然无法与直接提高测序质量带来的数据提升相提并论。因此,未来对于千人基因组数据集一个重要的补足方向是对于数据质量的提高以及位点的验证,这需要增加后续投入,进一步提高每个样本的测序深度以及后续的重复性验证工作。

千人基因组提供了前所未有的详实的人类遗传多样性数据,然而该信息如何与基因表达、表观调控、人群表型等联系起来仍需大量工作。而之前所提到的关于人类基因组自然选择压力区段与遗传调控区域相关性的研究则在一定程度上揭示了遗传多样性与表达调控、表型适应之间的密切联系^[60]。利用千人基因组样本对mRNA及microRNA的表达水平及遗传多样性进行相关性分析发现,大量变异作用于调控区域并对人类基因表达水平以及转录因子结构产生重要影响,该研究从整体上对人类多态性与基因表达进行了分析,这对人类基因表达预测以及疾病易感性位点的预测有重要意义^[67]。虽然这些研究在基因组多态性-基因表达调控-表型变化的复杂关系面前仍显不足,但是千人基因组计划也提供了一个对该网络关系进行深入检验的机会。

千人基因组数据对于医学遗传学的意义同样十分重要。(i)千人基因组数据作为最具代表性的人群基因组测序数据可以为研究人员提供一个基础参

照人群,避免人群结构差异对于疾病关联研究的影响;(ii)每个研究人员在开展相关研究前可以利用千人基因组数据对自己所关注的基因进行初步筛查,这对后续工作的开展有着极为关键的指导意义;(iii)如前所述,千人基因组给予研究人员详细检验人类基因组多态性与基因功能、表型之间关系的机会,而这一工作的顺利开展将对人类疾病遗传学基础研究提供一个最为关键的证据;(iv)在关联分析与基因功能研究取得一定成果之后,千人基因组数据可以以全球为单位,对一些疾病表型进行易感性的预测。

综上所述,大数据时代背景下的千人基因组计划作为一个极为庞大的人类群体高通量基因组测序计划,弥补了前千人基因组时代的遗传数据缺陷,提供了极为丰富的人类遗传信息.利用这一平台,可以针对不同研究领域的具体情况制定研究方案,最终解决相应的科学问题.而实践也证明了这一计划已经对人类起源、演化、疾病等方面的研究提供了强有力的数据支撑,并取得了一系列有影响的研究成果.与此同时,这一计划的成功实施及相应的后续研究也拓展了对人类遗传学的认识,特别对于人类基因组中稀有多态性位点的分布规律、基因组多态性对基

因表达与功能的影响、疾病易感性的遗传基础等方面提供了重要的研究机遇.然而,千人基因组计划所提供的信息量如此巨大以至远未完全挖掘出这一大数据集所有的潜力,这也给未来的研究留下了无限可能.随着数据的不断完善,千人基因组计划也会不断更新.而从发展的角度来判断,该计划也很可能会被其他更为大规模的生物学数据计划所取代,这一不断交替上升的过程终将使人们越来越接近人类生命本质的答案.毋庸置疑,从人类基因组计划到千人基因组计划,这些大规模公共数据集对于研究人类遗传学规律研究起到了极为重要的作用,而从经济学的角度讲,这些计划所带来的收益更是难以衡量.重要的是,从这些宏大的计划之中人们认识到,只有公开透明的交流才能更好地整合全球资源,推动科学研究的快速发展.从自身利益来看,数据的公开化有助于寻找更好的协作者,取长补短,进一步完善自身的工作;从长远利益来看,数据的广泛公开化符合现今全球协同的趋势,可以进一步调动全球科研资源实现效益最大化.因此,希望数据的公开化不要仅停留在一纸书面承诺,而数据的全球无差别共享势必成为生物学领域的下一个里程碑事件.

参考文献

- 1 The International HapMap Consortium. The international HapMap project. *Nature*, 2003, 426: 789–796
- 2 Cann H M, de Toma C, Cazes L, et al. A human genome diversity cell line panel. *Science*, 2002, 296: 261–262
- 3 Abecasis G R, Altshuler D, Auton A, et al. A map of human genome variation from population-scale sequencing. *Nature*, 2010, 467: 1061–1073
- 4 Abecasis G R, Auton A, Brooks L D, et al. An integrated map of genetic variation from 1092 human genomes. *Nature*, 2012, 491: 56–65
- 5 Kidd J M, Gravel S, Byrnes J, et al. Population genetic inference from personal genome data: Impact of ancestry and admixture on human genomic variation. *Am J Hum Genet*, 2012, 91: 660–671
- 6 Fagny M, Patin E, Enard D, et al. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Mol Biol Evol*, 2014, 31: 1850–1868
- 7 Andersen K G, Shylakhter I, Tabrizi S, et al. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philos Trans R Soc Lond B-Biol Sci*, 2012, 367: 868–877
- 8 Enard D, Messer P W, Petrov D A. Genome-wide signals of positive selection in human evolution. *Genome Res*, 2014, 24: 885–895
- 9 Gronau I, Hubisz M J, Gulko B, et al. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet*, 2011, 43: 1031–1034
- 10 Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*, 2011, 475: 493–496
- 11 Veeramah K R, Hammer M F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat Rev Genet*, 2014, 15: 149–162
- 12 Wei W, Ayub Q, Xue Y, et al. A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. *Forensic Sci Int Genet*, 2013, 7: 568–572
- 13 Zhang Y, Shen X, Pan W. Adjusting for population stratification in a fine scale with principal components and sequencing data. *Genet Epidemiol*, 2013, 37: 787–801

- 14 Wong L P, Ong R T, Poh W T, et al. Deep whole-genome sequencing of 100 Southeast Asian Malays. *Am J Hum Genet*, 2013, 92: 52–66
- 15 Ayub Q, Yngvadottir B, Chen Y, et al. FOXP2 targets show evidence of positive selection in European populations. *Am J Hum Genet*, 2013, 92: 696–706
- 16 Grossman S R, Andersen K G, Shlyakhter I, et al. Identifying recent adaptations in large-scale genomic data. *Cell*, 2013, 152: 703–713
- 17 Mills R E, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 2011, 470: 59–65
- 18 Prufer K, Racimo F, Patterson N, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 2014, 505: 43–49
- 19 Skoglund P, Jakobsson M. Archaic human ancestry in East Asia. *Proc Natl Acad Sci USA*, 2011, 108: 18301–18306
- 20 Wall J D, Yang M A, Jay F, et al. Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics*, 2013, 194: 199–209
- 21 Harismendy O, Ng P C, Strausberg R L, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 2009, 10: R32
- 22 Suzuki S, Ono N, Furusawa C, et al. Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS One*, 2011, 6: e19534
- 23 Wang Y, Lu J, Yu J, et al. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res*, 2013, 23: 833–842
- 24 Nielsen R, Paul J S, Albrechtsen A, et al. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, 2011, 12: 443–451
- 25 Keightley P D, Halligan D L. Inference of site frequency spectra from high-throughput sequence data: Quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics*, 2011, 188: 931–940
- 26 Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 2013, 9: e1003521
- 27 Fumagalli M, Vieira F G, Korneliussen T S, et al. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, 2013, 195: 979–992
- 28 Korneliussen T S, Moltke I, Albrechtsen A, et al. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 2013, 14: 289
- 29 Grossman S R, Shlyakhter I, Karlsson E K, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 2010, 327: 883–886
- 30 Coop G, Przeworski M. An evolutionary view of human recombination. *Nat Rev Genet*, 2007, 8: 23–34
- 31 Hassold T, Hunt P. To err (meiotically) is human: The genesis of human aneuploidy. *Nat Rev Genet*, 2001, 2: 280–291
- 32 Inoue K, Lupski J R. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet*, 2002, 3: 199–242
- 33 Lynn A, Ashley T, Hassold T. Variation in human meiotic recombination. *Annu Rev Genomics Hum Genet*, 2004, 5: 317–349
- 34 Myers S, Bowden R, Tumian A, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 2010, 327: 876–879
- 35 Oliver P L, Goodstadt L, Bayes J J, et al. Accelerated evolution of the *Prdm9* speciation gene across diverse metazoan taxa. *PLoS Genet*, 2009, 5: e1000753
- 36 Hinch A G, Tandon A, Patterson N, et al. The landscape of recombination in African Americans. *Nature*, 2011, 476: 170–175
- 37 Myers S, Bottolo L, Freeman C, et al. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 2005, 310: 321–324
- 38 Frazer K A, Ballinger D G, Cox D R, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 2007, 449: 851–861
- 39 Kong A, Thorleifsson G, Gudbjartsson D F, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 2010, 467: 1099–1103
- 40 Auton A, Fledel-Alon A, Pfeifer S, et al. A fine-scale chimpanzee genetic map from population sequencing. *Science*, 2012, 336: 193–198
- 41 Scally A, Dutheil J Y, Hillier L W, et al. Insights into hominid evolution from the gorilla genome sequence. *Nature*, 2012, 483: 169–175
- 42 Langergraber K E, Prufer K, Rowney C, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci USA*, 2012, 109: 15716–15721
- 43 Green R E, Krause J, Briggs A W, et al. A draft sequence of the Neandertal genome. *Science*, 2010, 328: 710–722
- 44 Reich D, Green R E, Kircher M, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 2010, 468: 1053–1060
- 45 Durand E Y, Patterson N, Reich D, et al. Testing for ancient admixture between closely related populations. *Mol Biol Evol*, 2011, 28: 2239–2252

- 46 Sankararaman S, Patterson N, Li H, et al. The date of interbreeding between Neandertals and modern humans. *PLoS Genet*, 2012, 8: e1002947
- 47 Wong L P, Lai J K, Saw W Y, et al. Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. *PLoS Genet*, 2014, 10: e1004377
- 48 Scally A, Durbin R. Revising the human mutation rate: Implications for understanding human evolution. *Nat Rev Genet*, 2012, 13: 745–753
- 49 Behar D M, van Oven M, Rosset S, et al. A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am J Hum Genet*, 2012, 90: 675–684
- 50 Behar D M, Vilems R, Soodyall H, et al. The dawn of human matrilineal diversity. *Am J Hum Genet*, 2008, 82: 1130–1140
- 51 Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 2014, 46: 919–925
- 52 Sabeti P C, Varilly P, Fry B, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 2007, 449: 913–918
- 53 Voight B F, Kudaravalli S, Wen X, et al. A map of recent positive selection in the human genome. *PLoS Biol*, 2006, 4: e72
- 54 Williamson S H, Hubisz M J, Clark A G, et al. Localizing recent adaptive evolution in the human genome. *PLoS Genet*, 2007, 3: e90
- 55 Akey J M, Zhang G, Zhang K, et al. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res*, 2002, 12: 1805–1814
- 56 Hancock A M, Witonsky D B, Alkorta-Aranburu G, et al. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet*, 2011, 7: e1001375
- 57 Akey J M. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res*, 2009, 19: 711–722
- 58 Hernandez R D, Kelley J L, Elyashiv E, et al. Classic selective sweeps were rare in recent human evolution. *Science*, 2011, 331: 920–924
- 59 Fumagalli M, Sironi M, Pozzoli U, et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*, 2011, 7: e1002355
- 60 Lappalainen T, Montgomery S B, Nica A C, et al. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet*, 2011, 89: 459–463
- 61 Pybus M, Dall’Olio G M, Luisi P, et al. 1000 Genomes selection browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res*, 2014, 42: D903–D909
- 62 Li M J, Wang L Y, Xia Z, et al. dbPSHP: A database of recent positive selection across human populations. *Nucleic Acids Res*, 2014, 42: D910–D916
- 63 Hodgkinson A, Casals F, Idaghdour Y, et al. Selective constraint, background selection, and mutation accumulation variability within and between human populations. *BMC Genomics*, 2013, 14: 495
- 64 Ward L D, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 2012, 337: 1675–1678
- 65 Li J, Liu Y, Xin X, et al. Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution. *PLoS Genet*, 2012, 8: e1002578
- 66 King M C, Wilson A C. Evolution at two levels in humans and chimpanzees. *Science*, 1975, 188: 107–116
- 67 Lappalainen T, Sammeth M, Friedlander M R, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013, 501: 506–511
- 68 Gravel S, Henn B M, Gutenkunst R N, et al. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA*, 2011, 108: 11983–11988

Application of the 1000 genomes project in human evolutionary genetics

CHENG YaoTing, LI YuChun & KONG QingPeng

State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

The 1000 Genomes Project (1KG), a milestone plan of the large-scale genome sequencing era, greatly promotes the advancement of human genetics studies together with the Human Genome Project and the International HapMap Project. 1KG whole-genome sequencing data have been generated from different human populations, with the aim of elucidating deeper insights into the relationships between genetic diversity and human origins, evolution, and disease. Here, we briefly review the developmental history of 1KG, paying particular attention to its application in human evolutionary genetics.

1000 genomes project, human genetics, origin and evolution, whole-genome sequencing, big data

doi: 10.1360/N972014-01245