“唐尧”基因组已取得历史性突破——

建立中国人自己的基因组技术体系路有多远

深瞳工作室出品

采 写：本报记者 操秀英
策 划：刘 恕 李 坤

何忠（化名）没有想到，自己身上不到20毫升的血液样本，竟成就了一项被中国工程院院士、哈尔滨医科大学党委书记张学评价为“我国乃至世界范围内里程碑式的事件”的成果。

利用何忠的血液样本，北京大学人民医院教授高占成团队和中国科学院北京基因组研究所（国家生物信息中心）研究员康禹团队首次在世界范围内成功完成从端粒到端粒的中国人全基因组，获得包括Y染色体在内的高质量真实人类二倍体以及完整无间隙的全基因组参考序列（44+XY）。

因为这个采样点位于山西省临汾市——几千年前尧帝建立的古唐国遗址附近，研究团队将该参考基因组命名为“唐尧”。

在人们印象中，人类基因组图谱早已公布，如今普通人的基因组也很容易被测出来。为何“唐尧”基因组会被评价为“里程碑式的事件”，这一基础研究领域的突破意味着什么？科技日报记者对此进行了采访。

现有人类参考基因组用于中国人有偏差

这是一项由临床应用需求催生的基础研究。

过去几十年，北京大学人民医院呼吸与危重症医学科主任高占成的主要工作是接诊来自全国各地的呼吸科疑难杂症患者。他带领团队首次诊断出多例孤儿肺病，如弥漫性肺淋巴管瘤病、肺泡蛋白沉积症等。

诸多案例丰富了他的医学实践，但也给他带来了诊疗困惑。不少疾病综合征在不同种族人群中的临床表现存在不小的差异。

“目前所有的肿瘤、遗传病等测序诊断报告，均根据美国主导的GRCh37/38为人类参考基因组序列来判定正常或变异。”高占成说，GRCh37/38是来自多个人类个体基因组序列嵌合而成的一套基因组，主要来源是非洲和欧洲人。它不但不完整、错误多，而且难以代表中国乃至亚裔族群。

以遗传性肺囊性纤维化为例，这种病在欧美白人中表现为跨膜氯离子转录因子突变导致的功能缺失。但在中国患者中，该转录因子突变的发生率要小得多。

“预测疾病风险和诊疗时，对于亚洲人种而言，仅对照现有参考组，可能会产生较大的偏差。”高占成说，这种偏差还会影响靶向药物的研发。

2003年，国际知名药厂阿斯利康在全球率先研发成功表皮生长因子受体酪氨酸激酶抑制剂（EGFR-TKI）——吉非替尼，适用于存在表皮生长因子受体（EGFR）基因突变的非小细胞肺癌患者。

随后的研究发现，EGFR基因突变存在明显的种族特异性。中国和东亚种族不吸烟肺腺癌患者的突变率明显高于欧美白人患者。

“目前的主流观点认为，不同人种基因组之间的差别只有千分之一。但从临床实践来看，实际差别可能远大于这个数字。”高占成说，“所以，我们有必要构建中国人自己的参考基因组。”

但对于一个临床医生来说，这是个全新且较难攻克的课题。

2020年，一个合适的契机到来。

这一年，设在山西省临汾市中心医院的高占成呼吸病学山西工作室开始筹建。

“这个工作室绝不能仅仅挂个牌子，要有具体的课题，能解决实实在在的问题。”高占成说，绘制中国人自己的参考基因图谱被提上日程。

他立即联系他带的第一个博士生，也是多年的合作伙伴——中国科学院北京基因组研究所研究员康禹。

“我当然很高兴能参与这项工作。”康禹说，“我们判断，现在的技术发展是构建中国人参考基因组的最佳时机，可以让我们以较少花费、较短时间完成这件事情。”

为中国人基因组研究提供更准确的坐标系

何忠何许人？为什么何忠的基因组就可以称为参考基因组？

康禹说，选择合适的样本是第一步。悠久的历史、多样的地理气候环境，塑造了中华民族独特的遗传多样性。“‘唐尧’基因组是研究的起点，我们决定从人数最多的汉族开始。”康禹说。

“构建中国人自己的参考基因图谱，目的是为了更好地服务现代医学应用，所以样本需要更好地代表现代中国人的基因组特征。”康禹说，最终他们确定的样本来自一名现在生活在山西省洪洞县一个古老村庄的健康男青年——何忠。

这个地区是明代洪洞移民，即历史上有名的“大槐树”移民的起点。600多年前的这场迁徙持续了近半个世纪，大量移民遍布中国各地，有些进入东南亚。“我们认为何忠的基因组有望成为现代汉族人群的代表。”高占成说。

根据祖源分析，“唐尧”基因组的绝大部分为东亚人群特征。“这个样本的Y染色体的分型在中国除了新疆、西藏等地外都有广泛分布，极具代表性。”康禹说。

“唐尧”基因组提示了中国人和欧洲人基因组水平的显著差异。对照国际科学团队“端粒到端粒（T2T）”联盟（以下简称“T2T”联盟）于2022年发布的新版本人类参考基因组T2T-CHM13，“唐尧”显示出11%差异序列和5%差异基因。

中国科学院院士陈润生说，“唐尧”弥补了汉族高质量基因组的空白，完整的中国人基因组序列的发布，也将改变以往认为不同人种基因组之间只有千分之一区别的认知。

张学认为，“唐尧”基因组将为汉族中国人基因组研究提供更准确的定位基因和变异的坐标系，同时解决欧洲血源参考基因组不适于中国人基因组研究的技术障碍。这将为我国医学基因组研究，包括遗传病诊断、常见病风险预测、肿瘤基因组变异、药物基因组学等领域，建立技术体系和质量基准。

中国工程院院士程京认为，“唐尧”基因组测序分析工作不仅具有非常重要的跨学科、跨领域的基础研究意义和应用价值，而且从DNA水平上回答了“何以中国人”这个重要的社会科学问题，将帮助我们回答中国人起源、迁徙、历史沿革和交流等问题。

用两年时间完成国际领先的质量标准

配置最先进的测序仪器和最精干的研发人员，“唐尧”项目以最快速度启动。仅用了不到两年时间，2023年8月，项目组获得何忠的完整无间隙高质量基因组序列。

结果超出课题组的预期。

经国际通用的评估基因组质量的重要工具Merqury评估，“唐尧”的质量值达到了参考基因组的质量标准，质量值为Q74.69，而T2T-CHM13的质量值为Q73.94。

“这个数字说明我们的参考基因组的错误更少，拼接质量高于T2T-CHM13。”康禹说。

将时间指针拨回到30多年前。1990年，在生命科学领域被誉为“登月计划”的人类基因组计划启动。11年后，该计划发布了人类基因组工作草图。又过了两年，研究人员公布了当时被称为人类基因组“完成图”。

此后数年，研究团队不断完善人类基因组空白区，但仍有约8%的序列缺失。

直到2022年，“T2T”联盟填补了缺失的“拼图”碎片，发布了T2T-CHM13新版本参考基因组。在这项成果中，科学家们成功地在人类基因组中增加了大约2亿个碱基，解码了从1号到22号染色体上的大部分空缺。而唯一被遗漏的，是人类所有染色体中最小的一条——Y染色体。

2023年，随着两篇研究论文发表在顶尖学术期刊《自然》上，人类Y染色体的完整序列终于展现在世人面前。

也就是说，国际基因组计划用了30多年的时间才获得包括Y染色体在内的人类完整单倍体基因组序列。

“唐尧”课题组同样拿到了这一结果。他们在世界上首次获得包括46条染色体的真实人类二倍体基因组序列（44+XY），能99.99%准确地区分来自父本和母本的两套单倍体基因组序列。

2022年，“T2T”联盟测的是一个单倍体，即所采用的DNA序列不是来自自然人的组织样本，而是来自女性子宫中的水泡状胎块（葡萄胎）细胞株——CHM13。

当时，“T2T”联盟联合主席、美国华盛顿大学霍华德·休斯医学研究所研究员艾文·艾克勒对媒体表示：“我们现在已经补全了一个人类基因组，下一个重点任务是补全二倍体基因组的父系和母系。”

“唐尧”课题组做到了。

“和‘T2T’联盟能补上最后的‘拼图’一样，我们之所以能快速获得这一成果，也得益于DNA测序和拼接技术的快速进步，以及包括国际基因组计划在内的大量技术和理论积累。”康禹说，“我们取得成果是因为站在了前人的肩上。”

这并不是一项只要有仪器、有资金就能完成的工作。“两年里，我们的团队夜以继日，创新了大量算法和拼接方式。这才能够实现高准确度地区分相似度极高的基因片段，实现高于NIH参考基因组的准确度。”高占成说。

避免“西方人比中国人更了解中国人”的尴尬

“这是中华民族群体遗传学研究的一个新起点。”中国科学院北京基因组研究所原副所长于军说，“接下来，我们将推进其他有代表性的个体参考基因组测序，并开展不同民族等群体的测序，最终我们希望能启动全民基因组测序工程。”

回顾过去，中国在基因组学技术领域的发展，可以说是从参与到同步。

陈润生回忆说，1994年，国家自然科学基金资助开展中华民族基因组若干位点基因结构研究项目，标志着我国人类基因组研究正式启动。

1999年，中国拿到了国际人类基因组计划1%任务。以华大基因和中国科学院基因组所研究人员为主力的科学家团队，高质量完成了这一测序任务，带动我国基因组学快速发展。在过去的20多年里，我国的基因组技术和研究取得了飞跃式的进步。

在构建中华民族自己的参考基因组方面，我国科学家也一直在努力。

“炎黄一号”是全球第一例中国人标准基因组序列图谱，也是全球20亿黄种人的首个个人基因序列图。该项目完成于2007年10月11日，是我国科学家继承担国际人类基因组计划1%任务、国际人类单体型图谱10%任务后，用新一代测序技术100%独立完成的中国人基因组图谱。

随后暨南大学、中国科学院北京基因组研究所等单位陆续开展了类似研究。但受限于当时的技术手段，这些基因组并未成为我国实际应用中的参考基因组，未发挥应有价值。

2023年，复旦大学、西安交通大学、中国医学科学院等26家单位联合发布了中国人群泛基因组联盟一期研究进展。该研究初步构建了首个中国人群专属的泛基因组参考图谱，且该成果全部由中国科学家独立完成。

在此基础上，专家们认为，我国要加快构建中国人自己的基因组研究“坐标系”的步伐。

20多年前，在人类基因组计划基础上，美国正式提出全新的大科学计划——精准医学计划。该计划最终目标是测定每一个人的基因组，也称为“全民基因组计划（All of Us研究计划）”。2022年，该计划研究项目公布了第一批近10万人的全基因组测序数据供研究人员使用。数据包括身高、体重和血压等基础数据和调查数据，例如关于参与者的人口统计数据、生活方式和总体健康状况的数据。

高占成说，一旦美国的全民基因组计划完成包括500万美籍华人在内基因组测序，完全有可能形成“别人比我们自己更了解中国人基因组”的局面。

近年来，国际科学家联合成立了人类泛基因组联盟（HPRC），试图建立更精准完整的世界主要人群的参考基因组，了解世界人口的多样性。去年5月，HPRC制作的首个人类泛基因组参考草图在《自然》发布，纳入了全球47个样本，其中包括3例中国南方汉族样本。

张学关注到一个现象：基因组领域最主要的两个国际联盟——国际人类泛基因组联盟、国际T2T基因组联盟，其中的重要成员都是来自欧美的大学和研究所，我国研究机构和实体并不在内。

“这种形势下，建立中国人自有的高质量参考基因组是防止被‘卡脖子’的关键一步。”张学说。

“接下来我们将对‘唐尧’进行进一步的解析和注释，让它能更好地应用于临床。”康禹说，我们希望基于自己的参考基因组发展出服务华人的靶向测序、基因组分析和诊疗技术，并推动未来的新药研发。

亟待构建中国人自己的基因组技术体系

受访专家预计，T2T-CHM13以其完整性和高质量，有望逐渐取代目前正在使用的GRCh38参考基因组。

陈润生和中国检验检疫科学院体外诊断试剂所副所长黄杰均建议，在新旧参考基因组交接之际，我国应建立国家标准，推广使用“唐尧”作为中国人群基因组研究和临床应用中测序和分析的标准物质和参考基因组，不再使用欧洲人的参考基因组来定义中国人的遗传变异。同时，在此基础上建立中国人基因组学知识框架和应用技术体系。

于军等科学家认为，要实现上述目标，我国人类基因组研究亟待进一步加强顶层设计和规划。“由谁来测，给谁用，数据安全如何保障，这些问题都需要系统研究。”

1993年，于军参与人类基因组计划这一里程碑式的科学计划。他在导师梅纳德·奥尔森的全力支持下，促成了中国科学家参与人类基因组计划。

多年来，中国的基因组研究计划是什么，如何建立自主的基因测序技术和数据体系，这些问题在于军的脑海中挥之不去。

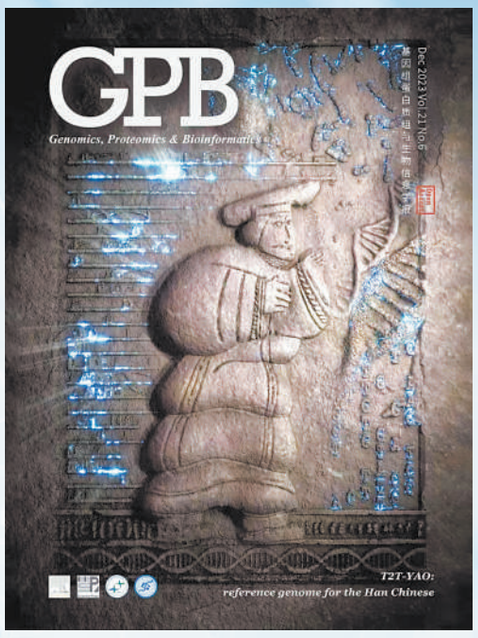
于军认为，我们目前的相关研究仍然是相对零散的，开展的群体研究规模较小，且数据所有权分散在不同研究者手中，无法共享数据集成创新，造成了资源浪费。

研究与应用的分离，也是目前存在的突出问题。于军说，我国基因组领域的基础科学研究、临床准入、应用规范由不同部门管理，信息沟通效率不高，造成应用需求难以对基础科研起到有效牵引作用，基础科研和临床应用之间无法形成有效反馈和良性循环。为了促进基因组领域基础研究和临床医学的合作与交流，北京大学人民医院于今年1月成立了人类基因组研究中心，以深入拓展“唐尧”基因组的相关研究和医学应用。

于军认为，在陆续构建中国人自己的参考基因组的基础上，未来如何推动更大规模的人群测序，最终实现全民测序，真正推动精准医学的发展，都是当前必须面对的课题。“你测几百人，我测几千人，这些数据除了发表一些看起来还不错的论文，大部分并没有推动临床诊断、新药研发等实际应用。”

针对这种现状，专家认为，目前亟待整合有限资源，包括资金、人才、样本资源、基础设施等条件，集中管理样本和数据，有效协调资源。

“我们可以探索成立一个类似国家人类基因组研究与管理中心这样的机构。”于军建议，该机构采用中央决策、专家委员会监督指导、中心执行的管理模式，统筹科技资金，协调社会资源，规范技术标准，促进科技转化，防范安全风险。“以此实现自主建立我国具有国际竞争力的人类基因组技术体系和知识框架的目标。”



“唐尧”基因组相关研究成果发表在《基因组蛋白质组与生物信息学报》上，图为当期杂志封面。 受访者供图

本版图片除标注外由视觉中国提供