

今年是人类基因组草图公布20周年,中国参与并完成了其中1%的任务

破译“生命天书”,生命科学走向“大数据时代”

■晋楠

今年是人类基因组草图公布 20 周年。2001 年,被称为破译“生命天书”的人类基因组序列草图正式发表。20 年来,这一划时代成就为人类对疾病和物种演化的认知带来了革命性变化。

我国科学家参与并完成了国际“人类基因组计划”(HGP)1%的任务,成为当时世界

上少数几个能独立完成大型基因组分析的国家。正是从这 1%起步,中国的基因测序技术逐渐跻身全球第一梯队。

随着基因测序成本以“超摩尔速度”直线下降,被解密的致病基因越来越多,新的基因药物不断问世。展望未来,一个由遗传学、生物化学、分子生物学和信息科学等学科共同构建的生命科学“大数据”时代已然来临。

DNA 测序价格“断崖式”下跌。

“人类基因组计划”在 1985 年被提出时,其预算可谓是一次“世纪拍板”。当时,并没有人知道测序究竟要花多少钱,于是科学家就按 1 美元一对碱基,“拍脑袋”做出了 30 亿美元的总预算,计划周期为 15 年。

自上世纪 70 年代桑格等人发明第一代测序技术,到上世纪 90 年代科学家把多种渐进性技术创新整合在一起,最终成为绘制人类首个基因组图谱的基础技术。

然而,这种测序技术耗时长、费用高,无法满足现代科学发展对生物基因序列获取的迫切需求。当第一个人类基因组被完全“破译”后,以更低成本、更短时间、更高效率,准确测定每个人的基因组,成为科学界的迫切需求和普遍共识。

第二代高通量测序是对传统桑格测序的革命性变革,一次运行即可同时得到几十万到几百万条核酸分子的序列,大大降低了获取核酸序列所需的成本。这一发展速度甚至超越了芯片发展的“摩尔定律”,使科学家获得基因序列所需的时间和资金成本直线下降。

2007 年 5 月,“人类基因组计划”完成四年后,“DNA 之父”詹姆斯·沃森获得了世界上首份个人基因组图谱。而绘制沃森基因组图谱的“吉姆工程”前后只用了不到两年时间,花费仅 200 万美元。尝到“甜头”之后,科学家的创新热情持续高涨,又发明了第三代单分子测序技术、第四代核糖核酸测序技术,可进行单条序列长度更长、更准确的测序。

到 2010 年,全基因组测序费用已经可以控制在 1 万美元以内。而今年 6 月初,华大基因全资子公司武汉华大医学检验所中标的“四川大学华西医院十万例罕见病患者全基因组测序计划采购项目”,其标的金额更是低至每例 1348 元——已下降到 300 美元以下。这说明我国在人类全基因组测序的成本控制方面已达全球领先地位。

基于四代测序技术的接力,人类基因组序列终于有了更完整的版本。近日,由美国加州大学圣克鲁斯分校的卡伦·米加和美国国家人类基因组研究所的亚当·菲利普领导的“端到端粒”(T2T)联盟绘制的新版本基因组问世。

相较于 2001 年的版本,这一版增加了近 2 亿个碱基对、2226 个新基因,使人类所知的碱基对信息上升至约 30.55 亿对。这一结果填补了之前剩余的大部分缺口,是自人类参考基因组首次发布以来的最大改进,也使人类第一次获得了自身 DNA 的“完整版”信息。

据悉,最新的序列由两种互补的全新 DNA 测序技术推导而来——英国牛津纳米孔公司的技术允许更长的 DNA 片段以非常高的精确度进行测序;美国太平洋生物科学公司的技术则可产生超长的连续 DNA

序列。这些新技术使得拼图片段可以长达数千甚至数百万个碱基对,组装也变得更加容易。

正是得益于测序成本的大幅下降,与基因组相关的科学研究取得了突飞猛进的发展。20 年来,许多规模化人类基因组计划先后涌现。例如,国际人类基因组单体型图(HapMap)计划,旨在发现人类基因组中普遍不同的位点;DNA 元素百科全书(ENCODE)计划,目标是找出人类基因组中的功能元件;蛋白质组计划是第一个国际人类组织/器官的蛋白质组计划,也是中国科学家首次领导执行的重大国际科技协作计划。

不止于此,基因测序技术还在其他众多领域得到广泛应用,包括生物的基因组图谱绘制、环境基因组学和微生物多样性、转录水平动态响应及其调控机制,疾病相关基因的确定和诊断、考古学(古代 DNA)、物种演化过程等等。

正如 2018 年 9 月 30 日詹姆斯·沃森等人在《自然》杂志共同撰文所说:“人们常常以为,先有科学发现再有技术发明,科学发现和技术发明存在单向关系,其实不然。”“很多科学发现都是技术发明推动的。比如,由于玻璃磨制技术的改进,发明出了望远镜,天文学才得以空前进步;再如,莱特兄弟发明了飞机,其后才有空气动力学。有了 DNA 测序技术的进步和工具的发展,基因组科学才有如此惊人的突破性发展。”

从生物学到生命组学,生命科学迎来“大科学”融合

20 年来,“人类基因组计划”所取得的划时代成就,为人类对疾病和物种演化的认知带来了革命性变化。

美国国家人类基因组研究院主任埃里克·格林曾想象,有一天,基因组学可能会成为临床治疗的一部分,“但我真的没想到,这会在有生之年发生”。

2010 年,科学家怀揣了数十年的梦想成为现实。当年,罹患罕见致命性肠道感染病的 6 岁美国男孩尼古拉斯·沃尔克成为世界上第一个被基因测序技术拯救的儿童。基因测序发现,这个男孩有一个出人意料的基因突变,并指出了一种治疗方法——医生从脐带血中取出细胞进行骨髓移植,取得了良好疗效。

这一成功带给了科学家巨大信心。近年来,又有多位地中海贫血、白血病患者受益于基因技术而被成功治愈。这些成果向人们展示了基因科技造福人类的美好前景。

如果说,技术的更新换代和价格的大幅下降,为基因测序“飞入寻常百姓家”提供了保



图/视觉中国

障,那么政策支持就为基因技术行业的发展提供了红利。

2015 年 1 月,美国宣布启动“精准医学计划”,目的是让所有人获得健康个性化信息。同年 3 月,我国首次召开精准医学战略专家会议,计划在 2030 年前投入 600 亿元加速中国精准医疗的发展。

在此背景下,医疗应用已成为基因测序最大的增长点。在临床上,它被应用于生育健康、肿瘤个体化诊断和治疗、遗传病及传染病检测等方面。在肿瘤检测、个性化用药领域,随着应用技术和数据解读技术的不断深入,基因检测市场发展空间也越来越大。

我国庞大的人口基数也成为发展医学基因检测产业的巨大优势。近年来,我国启动了聋病基因组计划、万人双胞胎基因组计划、百万人群基因组大数据、十万新生儿出生缺陷筛查项目……这些项目将为实现疾病精准诊断提供重要支持。

杨焕明说:“人类基因组序列就像化学元素周期表之于化学一样重要,因为基因组所提供的遗传信息可以绘制出基因组图,而基因组图与人的生老病死密切相关。”杨焕明说。

今年 2 月 11 日,《自然》杂志发表了美国西北大学网络科学研究所亚历山大·盖茨等人的一项研究,他们通过对 1900 年至 2017 年间的 70 余万篇相关科研论文分析后发现,得到“注释”的基因数量迅猛增加。自 2001 年开始,每年关于蛋白编码基因的学术论文数量在一至两万篇,很多研究集中于诸如 TP53、TNF 和 EGFR 等“超级明星基因”上,这对深入认识重要基因十分关键。

“人类基因组计划”还进一步确定了基因组中非编码序列的重要性。这些序列的改变不会影响蛋白质的序列,但会干扰蛋白表达和调控网络,进而影响生物学功能。

在 2001 年之前,明确某种药物全部蛋白靶点的概率不到 50%。“人类基因组计划”完成后,美国每年通过审批的药物几乎都有清楚的作用靶点说明。

尽管如此,从整体来看,生命科学的组学时代才刚刚露出曙光,我们对生命的认识还远远不够。一方面,经过 20 年研究,生物学家目前对定义生命的网络组织和动态有了初步认知,但尚不足以充分理解任一系统。另一方面,“人类基因组计划”不仅促进了生物学和生物医学的发展,而且正在积极深化遗传学、生物化学、分子生物学和信息科学等多学科合作的“大科学”融合,共同构建生命科学的“大数据”时代。

延伸阅读

人类基因组计划

人类基因组计划(HGP)是一项规模宏大的、跨学科的科学探索工程,旨在测定组成人类染色体(指单倍体)中所包含的 30 亿个碱基对组成的核苷酸序列,从而绘制人类基因组图谱,并辨识其载有的基因及其序列,达到破译人类遗传信息的最终目的。

HGP 在研究过程中建立起来的策略、思想与技术,构成了生命科学领域的新学科——基因组学,可用于研究微生物、植物及其他动物。HGP 与“曼哈顿工程”和“阿波罗计划”并称为三大科学计划,是人类科学史上的又一个伟大工程,被誉为生命科学的“登月计划”。

1%项目

人类基因组包含近 2 万个编码蛋白质的基因,由约 30 亿个碱基对组成,分布在细胞核的 23 对染色体之中。

中国在“人类基因组计划”中负责测定和分析 3 号染色体短臂上,从端粒到标记 D3S3610 间大约 30 厘摩尔(相当于 3 千万个核苷酸)的区域,因此被称为“1%项目”。

国际人类基因组单体型图计划

国际人类基因组单体型图(HapMap)计划是人类基因组中常见遗传多态位点的目录,它描述了这些变异的形式、在 DNA 上存在的位置、在同一群体内部和不同人群间的分布状况。

HapMap 为研究者提供了相关信息,使之能够将遗传多态位点和特定疾病风险联系起来,从而为预防、诊断和治疗疾病提供新的方法。

DNA 元素百科全书计划

DNA 元素百科全书(ENCODE)计划于 2003 年由国际科学界发起创建,旨在研究出人类基因组上各个基因的功能信息。该计划获得并分析了超过 15 万亿字节的原始数据,以确定哪些能打开和关闭特定基因,以及不同类型细胞之间的“开关”存在什么差异。

国际人类蛋白质组计划

国际人类蛋白质组计划(HPP)是继国际人类基因组计划之后的又一项大规模的国际性科技工程。首批行动计划包括由中国科学家牵头的“人类肝脏蛋白质组计划”和美国科学家牵头的“人类血浆蛋白质组计划”,这也是中国科学家第一次领导执行重大国际科技协作计划。(杨霞整理)

从“1%”到第一梯队,在基因组学烙下“中国印记”

1953 年,英国科学家弗朗西斯·克里克和詹姆斯·沃森发现了 DNA(脱氧核糖核酸)双螺旋结构——由含有四种碱基(A、T、C、G)的脱氧核苷酸连接而成的长链。这四个“字母”的排列组合构成了所有物种基因组的“天书”。

如何破译这部“天书”,从而了解生命的奥秘?1977 年,英国生物化学家弗雷德里克·桑格等发明了末端终止测序法,第一代 DNA 测序技术由此诞生。同年,他们测定了第一个基因组序列——全长 5375 个碱基的噬菌体 X174。自此,人类获得了窥探生命遗传密码的能力。此后,解码 DNA 序列成为生命科学发展的重大任务,至今还在不断深入、再深入。

1990 年 10 月,“人类基因组计划”在美国率先启动,英、日、法、德等国相继参与,组成了国际“人类基因组计划”协作组,其核心内容是测定人类基因组的全部 DNA 序列,获得人类全面认识自我最重要的生物学信息。中国要参与这项生命科学领域的“阿波罗登月计划”吗?答案是肯定的。1994 年,中国“人类基因组计划”(CHGP)在谈家桢、吴旻、强伯勤、陈竺,以及沈岩、杨焕明等科学家的倡导下启动。

1997 年 11 月,在中国遗传学组青年委员会第一次会议上,杨焕明、于军、汪建、贺福初、贺林、余龙、夏家辉等青年科学家提议加入国际“人类基因组计划”,推动中国基因组科学的发展。

1998 年 8 月,在中国科学院和国家南方人类基因组研究中心同行的支持下,中国科学院遗传所(现遗传发育所的部分前身)人类基因组中心成立。按遗传所原所长陈受宜的话说,该中心“将有志于此的人才汇聚一堂”,解决了当时我国大规模基因组研究人才匮乏的窘境。

1999 年 7 月 7 日,国际“人类基因组计划”协作组公布了中国加入“人类基因组计划”的申请,并于同年 9 月 1 日正式向全球宣布。这标志着我国成为美、英、日、法、德之外,第六个“人类基因组计划”的参与国,也是参与该计划唯一的发展中国家。

2001 年 8 月 26 日,遗传所人类基因组中心、北京华大基因研究中心和国家南方人类基因组中心共同完成了“1%项目”的基因组图谱——相比国际同行,提前两年高质量完成任务。

对整个项目而言,“1%”也许有些微不足道,但它的实施对于我国基因组学发展意义重大。通过“1%项目”,我国公众也接受了一次声势浩大的基因科学普及教育,为中国生命科学研究和生物产业发展开拓了无限空间。

中国科学院院士、华大基因联合创始人杨焕明认为,加入国际“人类基因组计划”,可使中国平等分享该计划所建立的所有技术、资源和数据,并使我国成为世界上少数几个能独立完成大型基因组分析的国家。

2003 年 4 月 25 日,距离发现 DNA 双螺旋整整半个世纪后,耗资近 30 亿美元、历时 13 年的“人类基因组计划”宣告完成。这项人类科学史上的伟大工程开启了生命科学的新时代——在计划实施过程中建立起来的基因组学、生物信息学技术,对全球生物相关学科和产业的发展,起到了巨大推动作用,各种生命科学新兴技术和生物产业如雨后春笋般涌现。

中国科学家在其中“抢”到的 1%份额,将“中国”二字镌刻在了这一人类科技史的重要里程碑上。同时,它还推动中国基因测序技术从“跟跑”到“并跑”,并逐渐走向全球第一梯队。

通过参与这一计划,中国科学家得以在短时间内学习并追赶发达国家的先进生物技术。我国基因组研究团队先后完成了水稻基因组、小麦 A 基因组、SARS 冠状病毒的基因组研究,并完成了对熊猫、家猪、家鸡、家蚕等动物的基因组测序工作,使中国基因组研究跻身世界前列。

同时,我国基因组测序产业规模与创新研究“比翼齐飞”。在全球基因测序市场,华大基因等公司已占有一席之地;中科院遗传发育所、北京基因组研究所、农科院基因组研究所等科研机构,成为我国基因组学原始创新研究、创新人才培养的重要基地。

从 30 亿到 300 美元,测序降价带来基因组学大飞跃

20 年来,随着基因组测序技术飞速发展,

