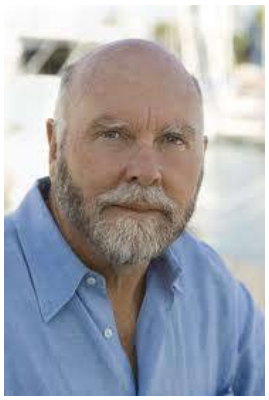
2. 个人多个基因组测序——拭目以待



原文作者：J. Craig Venter现在美国加利福尼亚州的J. Craig Venter研究所工作。Venter于1972年获得生物化学本科学位，1975年获得加州大学圣迭戈分校的生理和药学的博士学位。在NIH期间，Venter利用一种快速判定细胞中的mRNA的技术研究人类的脑基因，并利用该技术发现了被称作表达序列标签（EST）的短cDNA序列片段。Venter是Celera Genomics的创始人和前总裁。该公司因与人类基因组计划竞争并努力完成一个以商业为目的的人类基因组而著名。Celera Genomics计划创建一个付费的基因数据库。这在遗传学家社群内被证明是非常不受欢迎的，并激发了其他研究组加倍努力完成测序工作并将这些序列发布为开放获取。Celera Genomics计划使用了5个人作为人类基因组测序的对象，Venter则是这5个人中的一个。而人类基因组计划则使用了来自全世界的很多人群作为样本，这本身使得将基因组数据私有化的努力变得徒劳。尽管动机不同，Venter还是和他的对手Francis Collins以及美国总统Bill Clinton于2000年联合宣布人类基因组草图已完成，并一起分享了当年的生物年度奖。

联系方式：jcventer@jcvi.org

J. Craig Venter说，基因组数据将很快成为一种商品；下一个挑战——联接人类遗传变异、生理学与疾病——将与十年前任何一位基因组学家面临的挑战同样重大。

近十年前，Francis Collins和我站在白宫，与Bill Clinton总统一起，宣布关于人类基因组的两份草图。自此至今，DNA测序技术引人注目的进展超越了我们任何人的预言。人类基因组计划吸取了世界范围的努力，数十亿美元被投入来实现这个过去被某些人认为不可能完成的目标。今天，多亏改革创新，使得单个人类基因组测序可以用一个简单的仪器在一天之内完成，并且只花费数千美金。这场革新有一部分是由以下单位发起的：我的Celera Genomics公司以及Francis Collins领导的团队的努力。

然而，要使这项成就对医疗和健康产生显著影响，还有一段路要走。随着测序花费持续下降，数据质量需要不断提高。如果没有个体可见特征的相应表型信息，以及将两者联系起来的计算工具，基因组数据的增加将价值微小。研究人员今天面临的挑战，至少与我和我的同事十年前面对的挑战一样令人畏怯。

后见之明

人类基因组计划从一开始就因为一些原因而充满争议，特别是研究基金很可能从其它生物学研究计划转而向它倾斜。一些早期的决定对研究策略有着长远效应。1989年，测序预算计划在数年内降到每个碱基对花费一美元，因此一些相关人员决定向美国国会提议，拨出30亿美金

来完成一个含有30亿碱基对的单倍体基因组测序，而不是用60亿美金来完成一个含有60亿碱基对的双倍体基因组测序，因为这个代表着两套染色体的双倍体被认为太昂贵了。

人们还一致认为，基于人类基因组计划的雄心壮志，需要庞大的科学家队伍来对基因组片段测序。一旦做出这些决定，就没有多少空间进行实质性的革新了。我相信，历史已经证明这些决定是一个错误。比如，尽管Leroy Hood争辩说，他最开始的测序机器就跟最老的福特车一样，需要很大的努力来发展测序技术，但是，伟大计划还是不顾一切地前进着。

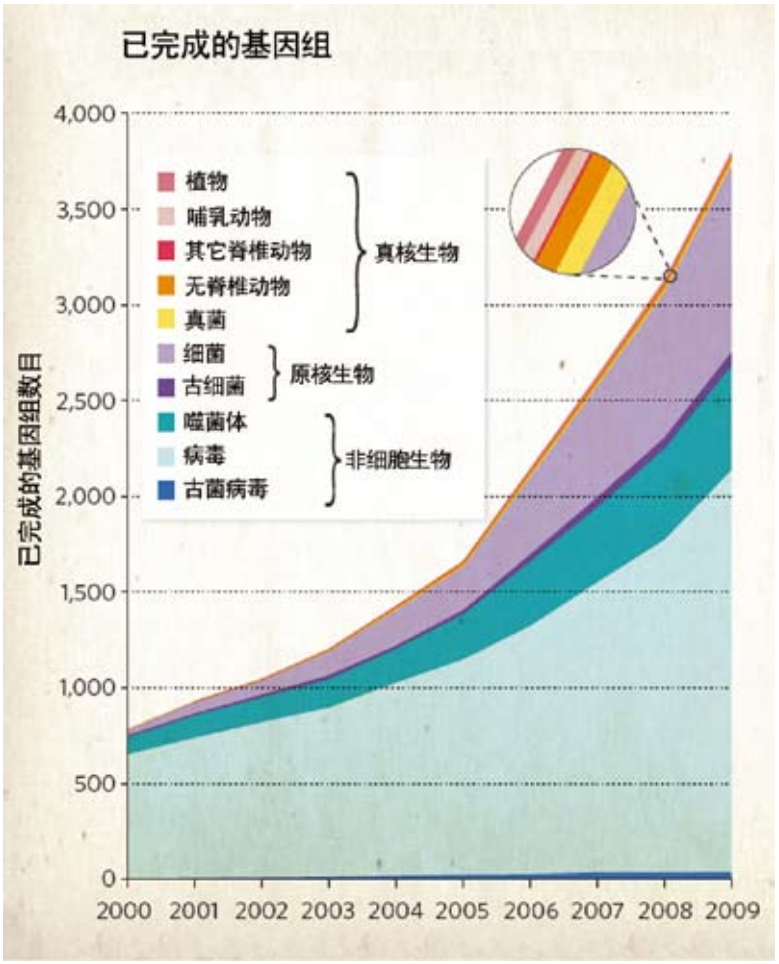


图1 已完成的基因组。经过测序的非人类基因组数目增至今天的3,800多个。这些基因组序列储存在国际核酸序列数据库协作体（International Nucleotide Sequence Database Collaboration）。

1994年，鉴于人类基因组计划进展缓慢、效率较低，所以我在基因组研究机构（Institute for Genome Research）的团队开发了“全基因组鸟枪测序”（whole genome shotgun sequencing）方法。我们用它在三个月内完成了一个细菌全部基因组的测序。五年后，我们在Celera应用这个方法对果蝇和人类基因组进行了测序。自2001年全基因组鸟枪测序引发众议起，它几乎被用于每一个基因的测序。

Celera的工作是在一个独立的大实验室完成的。这个实验室拥有300台DNA自动测序仪和一台高效计算机。而人类基因组计划则使用了分布在世界各地实验室的大约600台DNA测序仪。这两项计划相结合，对人类进行了不平常的早期洞察。其中最有意义的发现是人类基因数目只有26,000，比早先估计的300,000要小，而且个体之间的变异只有0.1%这么少。

随着人类基因组草案的发布，许多分析家预言DNA测序技术市场将走向终结。正如我们现在所知，一个非比寻常的故事展开了。测序中心转向了动物学，经过测序的非人类基因组数目增至今天的3,800多个（图1）。同时，全世界的实验室继续为人类基因组草图增添数据，生成了2004年的升级版本。我的团队集中精力于完成我个人的基因组测序，并在2007年发布了来自一个个体标本的人类基因组二倍体首次测序（称作HuRef基因组）。

更大的差异

这个人类基因组二倍体测序展示了人类多样化的新图景。测序显示，在DNA序列中核苷酸插入和缺失被包含进单核苷酸多态性——遗传变异的另一个普遍形式时，我的两个亲代基因组产生了0.5%的差异。这是一个引人注目的增长，它超过了2001年只关注单核苷酸多态性时估计的0.1%。随后发现不同个体之间基因组的差异介于1%~3%。

为什么人类基因组两份草图的数据并没有显示出个体之间基因组的差异如此显著呢？在人类基因组计划中，人们设计了一个单倍体基因组计划。单倍体测序涉及克隆片段测序，因此没有办法来直接检测多态性、插入和缺失。相反，Celera计划的问题是遗传变异太多了。DNA来自不同种族的两个男性和三个女性，分别是非裔美籍人、中国人、西班牙人和高加索人。Celera公司的基因组集装软件的早期版本使用一个叫做“多数规则”的方法来生成单个共有序列。这使得每个基因组都遗漏了大量的插入和缺失计数。如果我们只使用一个人的DNA，那么就可能可以更加完整地了解个体之间遗传学差别的程度。

尽管两项计划都有限制，关于人类基因组测序的竞赛仍然激发了许多基础研究实验室和公司去发展测序技术，并在过去数年里崭露头角。2003年，为了激励商业和政府投资，我的机构设立了500,000美元的奖金，奖励取得最重大进展的团队——令每个基因组的测序费用等于或少于1,000美元。这个奖项自此发展成为千万美元的阿康艾克斯大奖（Archon X-Prize），以奖励建立某设备的第一团队。此设备的目标是在10天内以最少花费进行高度精确的100个人类基因组测序。

这些激励和其它的激励措施很快加速了革新的步伐（图2）。今年初，两家公司——Illumina和Life Technologies宣布出售新的测序仪器。这些新仪器每天分别可以生成250亿和1,000亿个碱基对。“Life Technologies”还宣布了一个未来版本，它可以每天生成3,000亿碱基对。两家公司都宣称它们目前的仪器可以在一天之内进行一个人类基因组测序，并且花费少于6,000美元。另外一家公司——Complete Genomics宣称能在5,000~8,000美元之内完成一个人类基因组测序，但是它并不出售生产仪器。这让我回想起我在NIH的实验室里的第一台应用生物系统测序仪——Applied Biosystems sequencer。1987年时，它每天仅能进行4,800碱基对测序。如今这令人难以置信的进展能够匹配甚至超越同时期的任何高性能计算研究。23年过去了，最新的生命技术测序仪已将测序提高了约8个量级。

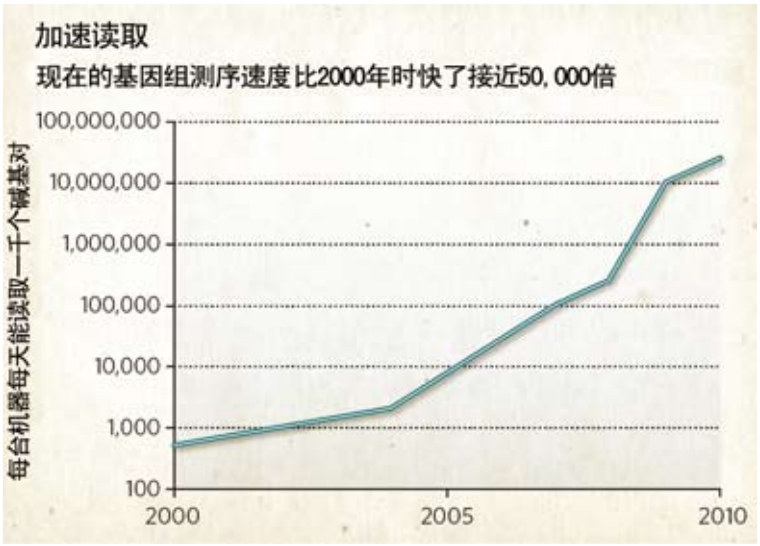


图2 加速读取。

然而，这些大幅提升也伴生了大量的负效应。大多数高速测序仪一次只能对少于100碱基对的片段测序（或称“读取”）。这比第一代仪器“Sanger”、第二代仪器“Roche 454”都要少得多，前者和后者分别可以一次性读取800~900和500碱基对。短读取显著阻碍了将序列集成长链的过程，而长链代表着染色体。测序组已经努力试图克服这些限制。他们的方法是将成果在已经公开的人类基因组序列上分层，而不是试图从混乱的数据中装配整个序列。这对任何单个基因组都造成了曲解，而尽管进程中已有如此的进步，得到的数据质量仍然在特定标准以下。

提高数据质量非常重要，因为如果一个人类基因组不能被独立装配，那么测序数据便不能被分装入两套亲代染色体，或者单倍体。而单倍体期将是基因药物中最有用的部分之一。确立我们从每一个亲代所得到的整套基因信息，对于理解遗传率、基因功能、调节序列以及我们对疾病的易感性都十分重要。幸运的是，这条路上很多令人振奋的进展对我们有益，比如Pacific Biosciences公司和Life Technologies公司的新方法，可以从单条DNA链生成序列信息。这个方法允许了在数以千计碱基对范围内的序列读取，从而将大大提高基因组序列数据的质量。

下一场挑战

以目前技术进步的速率，DNA测序可能很快成为一种商品，而生产价廉、高质量的测序数据将不再是一个问题。因为人类生物学和临床信息的复杂，下一个障碍——表型——提出了比基因型更为严峻的挑战。

即将改变医学的实验揭示了人类基因变异和生物成果的关系，比如生理学和疾病的关系。这些实验将一起获得成千上万完整的人类基因组，以及广泛的数字化表型数据。一个简化的问题可以被简易刻录，比如，“你有糖尿病吗：是或否”。一个更加全面的认识应该囊括发病年龄和疾病临床表现范围的评分，包括神经损伤的程度、血管问题，以及药物治疗的剂量和家族史。评分系统将会包括亚分类特征，比如疾病类型、进展和严重程度。

尽管我们拥有当下的所有信息，我们可能很难利用它们，因为我们没有计算基础来进行哪怕数千个基因型和表型的相互比较。而对这样的分析的需求可以成为建立“路由器”巨型计算机的佐证，它将比今天最快的计算机还快1,000倍。科学家需要在全球范围内合作来建立关于表型数据的标准，而推动力可以来自学术界、政府或者产业。

数十年后基因组学将何去何从？随着全世界测序能力及数据质量的提高，我们将定下比现在一人一个基因组更高的目标，即一人多个基因组测序，而来源包括精子、卵细胞、胚胎细胞、干细胞、肿瘤前细胞和癌细胞。这将使得我们能选择健康的细胞来进行生殖和组织移植，或者更好地理解老化和肿瘤进展。对于医学发展同等重要的，是我们体内寄居的数以百万计微生物的基因组测序。基因组革命才刚刚开始。

原文检索：J. Craig Venter. (2010) Multiple personal genomes await. *Nature*, 464(1):676-677.

姚宇亮/编译

注：本文观点仅代表作者个人意见，不代表本刊立场！