六、更多阅读

1. 基因组测序问题终结者

科学家正在努力开展基因组测序方面的工作，加紧减少片段之间的空隙，纠正错误，希望最终真正完成人类基因组的测序工作。《自然》（*Nature*）医学助理新闻编辑Elie Dolgin将带领我们去了解一下这些科学家们的最新工作进展。

在Deanna Church位于NIH五楼的没有窗户的办公室里，她几乎很难有机会分散精力从事工作以外的事情。在Deanna Church的电脑里装满了有关人类基因组序列的各种难题。虽然这些难题的数量可能有时会有所改变，但它们还是有一个潜在的压力在时刻提醒着Deanna Church带领的美国国家生物技术信息中心（NCBI）的研究小组，他们距离完成课题还有很长的一段路要走。这项课题正是早在20年前就已经启动的大名鼎鼎的人类基因组计划。

来自全球的科学家们已经为这个项目花费了30亿美元。2000年，在美国白宫举行了一场意义重大的新闻发布会，科学家们在会上宣布他们已经获得了人类基因组的序列草图。在次年公布这一草图时，科学家们用各种充满着诗意的文字对这本具有里程碑式意义的实验笔记大加赞誉。当2003年第一次有官方表态人类基因组测序工作已经完成时，科学家们又一次打开了庆祝的香槟。从那以后，媒体在报道基因组测序工作所取得的新进展的同时也刊登一些“负面”消息。比如全球顶极科技期刊《新科学家》（*New Scientist*）就曾经做过这样的报道：“科学家们保证，这一次是真的。”一年后，分析结果被公布，两年后（2006年），将最终的论文公众于世，文中对染色质进行了“全景描绘”。三年过去了，Church已经因为长期在电脑前辛勤的工作而略显驼背。她轻点鼠标，静静地浏览电脑屏幕上的那些基因组序列，思考着问题。在由她的合作者和世界各地的序列使用者们提交的问题当中有好几条问题都是有关序列缺失的。还有一些人反映有些序列有误。当然还有意想不到的，更为复杂的问题，比如复杂的DNA重排问题等，这些问题都是需要花费好几年的时间才能弄清楚的。

美国贝勒大学医学院（Baylor College of Medicine）人类基因组测序中心（Human Genome Sequencing Center）的主任Richard Gibbs说道：“这非常令人沮丧，这是一个质量非常高的基因组测序结果，是目前为止最好的结果，但是问题在于哪怕是非常小的一点差错也会造成非常严重的后果。”

Church的团队正在努力构建一个精确可靠的参考序列，但是他们的努力又再一次证实了这是根本行不通的。比如序列本身并不能代表任何一个人的基因组，序列只是男人、女人等不同

人DNA的集合。这是因为我们在进行人类基因组测序时本来就是采集了来自世界各地不同人种个体的DNA样品。正如人类基因组项目现任主席Francis Collins说的那样：“我们拥有同样的遗传背景。”

但是这种全人类共有的遗传特质却很难捕捉。任何两个个人的基因组看起来都不如我们原本设想的那么相像。我们对人体基因组进行测序时并不是按照30亿个碱基对中一个碱基一个碱基地进行测序，而是根据个人将这么一个巨大的DNA序列分解成数百条各不相同的片段，其中就可能会导致数百万个碱基对组成的序列发生丢失、插入、重复或者反转等现象。

如果我们真的能够得到一份完整的基因组参考序列，那么它将与我们最初得到的那份很不一样。而这就是Church等人现在正在进行的工作。他们正在努力消减这些基因组序列之间的差异，同时他们也希望能够打造出一个更富有弹性的平台，得以从中发现所有人类基因组的共性和独特性。尽管有人认为这纯属浪费时间和精力，因为现在可以只用花费十年前测序费用的很少一部分就能对个人的基因组进行测序了，不过绝大部分人还是认为这种参考序列是非常重要的，因为它可以对我们将来将要开展的人类基因组测序工作起到重要的参考作用。

即使Church解决了这些问题，她也不可能得到太多的赞誉。她既不可能像以前那些从事基因组测序工作的人那样到白宫与总统会面，也不可能在什么具有高影响因子的杂志上发表文章。不过如果Church留下什么问题，还是会有其他人来解决这些问题的，因为“这些问题虽然不是那么的让人感兴趣，但是它们的确很重要”，Church这样说道。

通力协作

到2003年4月为止，人类基因组计划的测序工作成果已经超出了当时的预定目标，即平均每10,000个碱基的出错率小于1个碱基，同时对基因组当中编码基因部分的测序覆盖率超过95%。但是问题仍然存在，比如在序列中还有大约350个缺口，还有很多结构变异区域没有得到测序等。

2004年，Church和其他一些科研人员聚集到英国维尔康姆基金会桑格研究所（Wellcome Trust Sanger Institute），一起讨论基因组和结构变异方面的问题。其中有一个问题得到大家一致的响应，那就是没有一个简单的办法能够让我们用新的测序数据来修补现有的基因组测序结果或者对该结果进行升级。在上世纪90年代，基因组测序还是一件非常不确定的事情，世界各地的科研人员们都可以与分布在世界各地的参与人类基因组计划的那几大测序中心（这几个中心分别负责一些染色体的测序工作）进行联系，报告任何测序方面的错误。但是到了2004年，几乎没有哪家测序中心还会主动检查测序错误了，他们也不再有热情去重新检查一遍以往的测序数据，于是问题出现了。美国国立人类基因组研究所（National Human Genome Research Institute，NHGRI）的项目主任Adam Felsenfeld这样说道：“必须要有一些人对基因组数据负责，只有这样才能在发现错误时进行及时的纠正。”

Church与英国茵格司顿欧洲生物信息学研究所（European Bioinformatics Institute）

的Ewan Birney一起向美国国立人类基因组研究所和维尔康姆基金会提交申请，希望他们能够提供资金支持。为此他们争取了两年多时间，但是最后美国国立人类基因组研究所只同意从每年3,000多万的测序资金当中拨出100万美元，供美国密苏里州圣路易斯的华盛顿大学（Washington University）使用。桑格研究所和维尔康姆基金会也提供了数额相当的资金支持，欧洲生物信息学研究所和美国国家生物技术信息中心负责他们擅长的生物信息工作。上述这四家单位共同合作组成的基因组参考序列合作体（Genome Reference Consortium, GRC）是目前世界上进行基因组序列改进工作的主要力量。

GRC为了进一步改进人类基因组参考序列的质量从而设立了三个主要目标：纠正序列组装错误；填补基因组当中的现有缺口；以及找出基因组当中高度可变区域的可能序列。

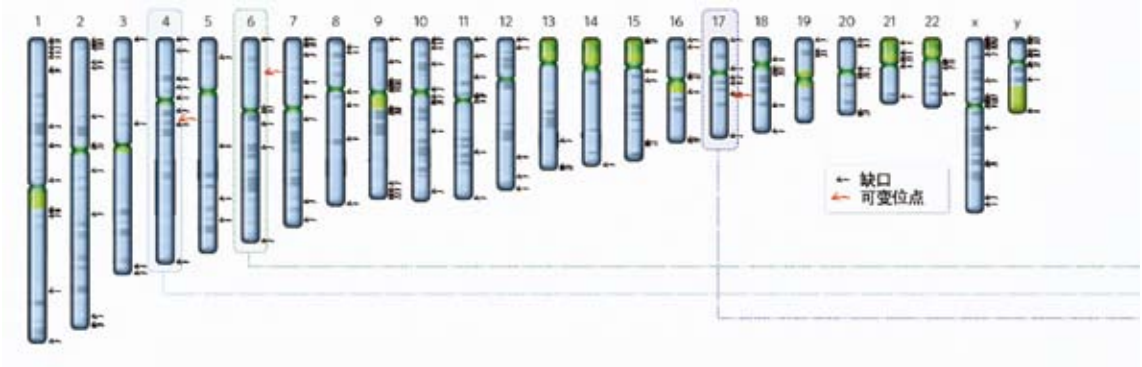

图1 人类基因组图谱当中现存的缺口。2009年3月，GRC公布了他们的第一份人类基因组图谱。该图谱相比人类基因组计划项目公布的图谱填补了25个缺口，增加了3个可变区域（即图中红色箭头所示）。不过这幅图谱当中仍然存在近300个缺口（即图中黑色箭头所示）。这些缺口的范围从700个碱基对到3,000万个碱基对不等，而且还不包括目前尚无法进行测序的染色体端粒部分和着丝粒部分（即图中绿色箭头所示）。

自从人类基因组计划宣告结束的那一刻起，全世界的科学家们就已经开始为上述前两个目标努力工作了，不过科学家们屡屡受挫，基因组当中有些区域总是无法解决。比如对于一些重复序列，科研人员们希望能够在细菌当中得到这些序列的多拷贝片段（这也是测序工作当中必需的一个环节），但是一直都没能成功，幸好最新的技术已经能够解决这个问题了。在今年年初，美国Broad研究所基因组测序和分析中心的共同主任Chad Nusbaum领导的一个科研小组使用了一种不需要利用细菌来扩增待测DNA片段的新一代的测序技术解决了上述问题。Nusbaum小组已经将测序结果递交给GRC，这些最新成果将被收录到人类基因组参考序列当中。

第三个目标则是最近才开始逐渐形成的。这是因为，最开始研究人员假设人们个体之间的遗传差异主要是由单碱基突变造成的。但是现在，随着对基因组结构变异，比如片段缺失、插入、反转和扩增等情况的了解越来越深入，我们发现实际情况并非如此。虽然这些变异当中有一些与遗传性疾病有关，但是与发现单碱基突变不同，我们很难发现这些结构变异情况，因为很难在参考序列当中将这些结构变异情况表现出来。因此，GRC除了要向我们提交一份准确的30亿个碱基对组成的DNA序列图谱之外，还必须用各种方式将人体基因组的多样性展现出来。

变异问题

人类主要组织相容性复合体（major histocompatibility complex, MHC）编码区就是这样一种高度可变的区域。该区域位于6号染色体上，由大约400万个碱基对组成，含有很多与人体免疫机制相关的基因，该区域被认为是人体基因组当中变异程度最高的区域。最初的参考序列中掺杂有大量的源自不同个体的被称为单倍体型（haplotype）的重复DNA序列，因此这种参考序列实际上是一个不属于任何个体的基因组序列。为了得到一份源自某个人体的参考序列，英国伦敦大学癌症研究院（University College London Cancer Institute）的Stephan Beck研究小组对一个单一的MHC单倍体型进行了测序。然后他们将测序结果与另外七名普通欧洲人的单倍体型序列进行了比较，结果发现了37,000多个单碱基突变和7,000多处结构变异。这种程度的遗传多样性要比全基因组水平的遗传多样性高出一个数量级。目前Beck小组的测序结果已经被收入GRC的默认参考序列当中，而上述那7名欧洲人的单倍体型序列则被当成了参考序列的可选序列（alternative pathway）。

在人类基因组当中还有两个区域也和MHC编码区差不多，具有高度的可变性，即具有多种单倍体型序列。其中一个区域位于4号染色体中编码UGT2B17蛋白（该蛋白参与了多种类固醇激素和药物的代谢过程）的基因附近。在目前已经完成的参考序列当中错误地组装了两个单倍体型序列，还出现了一个假缺口。后来经过更正之后发现，这个假缺口实际上只是在某些个体当中会出现的一段DNA序列缺失而已，缺失部位的两端各自有一大段DNA重复序列。目前这段区域也被GRC收录，作为了一段参考序列的可选序列。

另一个类似的区域位于第17号染色体的MAPT基因周围。该区域可供我们进行参考序列起源方面的个案研究，因为该区域只存在一种单倍体型。这种单倍体型是原始序列的复杂反转。在2005年进行的一项大规模家族样本研究中发现，该单倍体型只见于大约20%的欧洲人群当中，说明该区域承受着某种正向选择（positive selection）。但是到了2006年，美国西雅图华盛顿大学的遗传学家Evan Eichler和另外两个科研小组发现该反转区域非常容易出现自发性缺失，从而导致智力低下。这种既具有正向适应作用同时又会因自发性缺失而导致疾病发生的情况很像中国文化当中描述的一个物质所具有的阴阳两面性质。但是问题在于这到底是为什么呢？

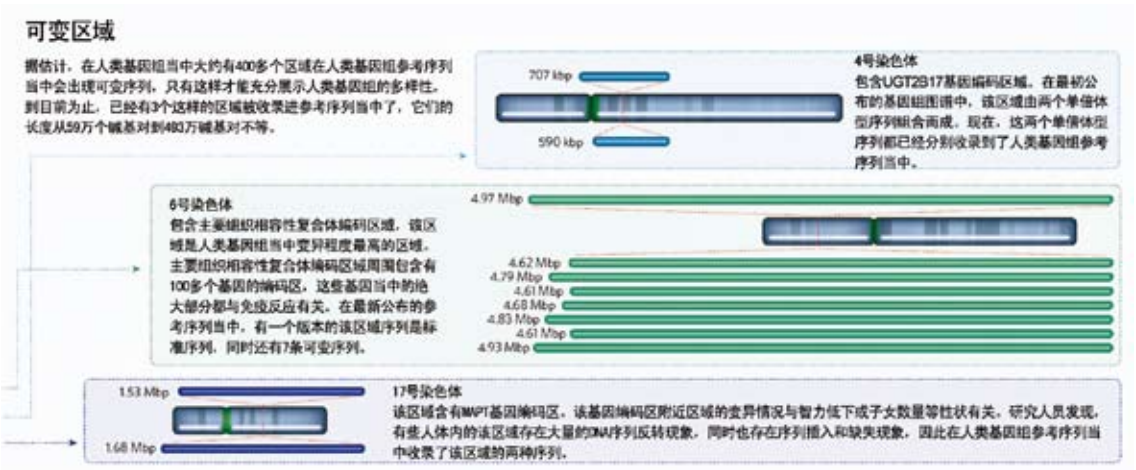

图2可变区域。据估计，在人类基因组当中大约有400多个区域在人类基因组参考序列当中会出现可变序列，只有这样才能充分展示出人类基因组的多样性。到目前为止，已经有3个这样的区域被收录进参考序列当中了，它们的长度从59万个碱基对到493万碱基对不等。

为了解答这个问题，Eichler急需人体基因组序列。他与Board研究所基因组生物学项目的首席技师Michael Zody合作，组成了研究小组，对人类全基因组进行了重新测序。结果他们发现反转的单倍体型区域非常容易发生与智力发育低下有关的序列缺失。2008年，Eichler和Zody发表了他们的这一研究成果。当时GRC正在全力以赴准备公布最新版的人类基因组参考序列。Eichler和Zody将他们的发现提交给了GRC。他们发现的这两种单倍体型序列都被收录到最新的参考序列当中。Zody说道：“GRC为我们提供了一个中枢结算所。”

Eichler介绍说，由于这些区域都具有非常重要的临床意义，因此对这类区域进行深入研究，尽可能地发现更多的参考序列对于临床工作当中发现这类区域中的突变具有非常重要的意义。“一旦弄清楚了这些区域的几种可变结构，相信我们就能够发现以前不可能知道的疾病间的联系。”他继续说道。Eichler估计在人类基因组当中大约有5%的区域（约有400个位点）会存在多种参考序列，弄清楚了这些区域就足以发现人类基因的多样性问题了。这些区域当中涵盖了1,000多个基因，这些基因能够广泛影响各种生理学过程，比如免疫反应过程、药物解毒过程和繁殖过程等。

共同的任务

GRC的第一项成果，即更为精确的人类基因组序列于2009年3月在网上发布了。该序列收录了几种参考序列，对以往公布的序列图谱中的3个部分进行了补充和更正，同时纠正了150多个比对问题，填补了25个缺口。但是仍然遗留了300多个缺口。2009年9月，来自GRC的20名核心成员齐聚英国茵格司顿，参加了GRC组织每两年举行一次的例会，讨论了GRC组织下一步的发展规划。此时，实验室里的工作人员们还在忙碌地敲击着电脑键盘，生物信息学家们也正在为GRC组织里讨论最为激烈的一项问题（即如何改变参考序列，让它只显示“普通的”基因变异）而努力着。GRC组织的9人科学顾问组（Eichler和Gibbs都是该顾问组成员）建议，只要可能，在基因组序列当中就应该包括DNA序列的“普通”版本。但是他们并没有详细解释普通的标准。到底应该是高频率的变异呢，还是在人群中出现较为广泛的变异呢？究竟应该是以全球六十多亿人口这个整体来进行衡量呢，还是应该就某一个地区的人口或者某一个种族的人口来进行衡量呢？结果还得等待千人基因组项目（1000 Genomes Project）的工作成果，因为它将影响GRC组织的最终判断。

有一些GRC成员不同意对人类基因组参考序列进行这样的改动。比如在EBI中领导脊椎动物基因组研究小组的Paul Flicek就说道：“我认为我们不应该对整个基因组的碱基序列进行逐一审查，从中一个一个找出发生突变的碱基。因为从信息学角度来看，这种突变根本就不碍事，只要基因组能够发挥作用就足够了。”

还有一些非GRC组织成员的人也质疑整个项目的合理性。加拿大安大略省癌症研究所（Ontario Institute for Cancer Research）的生物信息学家Lincoln Stein就奇怪为什么一定要纠结于十年前公布的数据。他将目前GRC的工作称作“抽象意义大于实际意义”。不过至少Church并没有受到这些质疑的影响，她作为一个非常注重细节的人很清楚哪怕是很小的一点问

题也会造成很大的影响。要知道每一个科研工作者都很热衷于他们研究的那几个基因，这也正是为什么Church那里总有一堆问题的原因。随着基因组学技术正逐渐成为个性化医疗工作中的一个重要组成部分，这些问题也必须得到解决。Church说道：“对于对基因组当中某个与某种疾病相关区域非常感兴趣的科研人员来说，他们并不在乎人类基因组图谱是不是达到了99%的完整度，他们只在乎他们感兴趣的那个区域是不是得到了完整并且准确的测序。”

因此，尽管面临着种种质疑，GRC仍旧在平静地继续着他们的工作，将一些T改正为A或C或G。直到有一天，当我们在用目前使用的鸟枪测序法时不再需要参考序列来帮助我们拼接序列片段时，我们就可以深入研究人类基因组参考序列当中的基因组多样性问题了。2010年，GRC还将继续对小鼠基因组序列和斑马鱼基因组序列开展类似的工作。虽然这已经不可能成为头条新闻了，但是学术界的人还都非常清楚这项工作的价值。在美国纽约冷泉港实验室从事基因组结构变异问题研究的遗传学家Jonathan Sebat这样说道：“成立GRC组织本身就是一项非常明智的决定。这是非常显而易见的一件事情，很明显，一定得有一些人去解决人类基因组图谱当中存在的那些问题。”

原文检索: Elie Dolgin. (2009) The genome finishers. *Nature*, 462:843-845.

YORK/编译

2. 隐藏在基因组当中的宝藏