生命奥秘

总136期 2021/09
LIFEOMICS

人类基因组学20年回顾

无奇不有

生命世界

阅读生命

走进科学

目录 CONTENTS

专题:人类基因组学20年回顾

专题

人类基因组学20年回顾

前言

通过重新提出数据共享，人类基因组计划的研究人员可以实现长久以来一直被遗忘的承诺。

20年前发表在《自然》（*Nature*）和《科学》（*Science*）杂志上的人类基因组初稿，为所谓的“生物学世纪”打开了大门。仅仅在本世纪初的前二十年时间里，信息库已经从两个粗糙的、充满错误的基因组序列发展到对世界各地成千上万名个体的遗传突变的全面描述，并且出现了越来越多的研究工具。《自然》杂志的本期特刊回顾了人类基因组学方面取得的成就，以及未来还需要继续深入挖掘的方面。但是围绕人类基因组的研究生态系统的某些方面几乎没有改变，仍然是一个令人担忧的问题。

基因组研究的许多伦理、法律和社会影响——包括隐私、知情同意以及研究人员和参与者的公平代表——仍未解决。此外，免费开放获取基因组数据在落实上也存在问题。近日，研究人员指出了在大流行期间无法获取冠状病毒基因组所带来的问题。研究人员、资助者和期刊若想实现人类基因组计划的承诺——更好地理解疾病并改进诊断和治疗——就需要解决这些问题。

基因组序列草图在《自然》上一经发表，就马上可以免费获取——事实上，最初的序列组装在草图正式发布之前的7个月前就已经在网上发布了。这是符合《百慕大原则》（Bermuda Principles）（1996年2月，人类基因组计划领导者聚集在百慕大，同意将超过一定规模的基因组测序数据在获取后的24小时内提交到公共数据库）的，这是一项由国际联盟成员签署的数据共享协议，它使人类基因组计划成为可能。

《自然》杂志早在1996年就致力于基因组学研究的数据开放原则。通过发表人类基因组计划的第一篇论文，我们与一个致力于数据共享、公共资助的项目合作。但《自然》承认，保持信息的自由、开放流动将面临挑战，研究团体可能需要在这些原则上做出妥协，例如当数据来自私营公司时。事实上，在2001年，《科学》杂志的同事们协商出版了马里兰州罗克维尔的Celera公司的基因组草图。该研究论文立即可以免费访问，但对完整数据的访问有一些限制。

20年后，妥协和拖延在基因组研究的三个领域逐渐成为常态：从参与者那里收集数据；在批准的、可公开查阅的数据库中存储；基于研究和医疗保健的目的获取数据。完全开放数据共享环境这一承诺尚未实现。

若让基因组学真正改变医学，则需要与表型数据——物理特征、医疗历史和其他可以与基因组突变相联系的可识别特征——相结合。然而收集这些数据增加了研究参与者的隐私风险。因此目前人类基因组数据研究受到了更多必要的限制，比如如何使用参与者的数据。此外，相关科学家需要接受审查，以确保参与者的知情同意，并确保参与者的利益得到保护。

下一步是将收集到的基因组序列及其相关数据存入得到批准的国际数据库，以便继续保护参与者的利益。但研究人员经常指出，他们无法迅速储存数据，理由是隐私和同意方面的担忧，或者是与提供数据的公司达成了协议。技术上的限制意味着储存数据的过程也可能非常耗时。同时，科学家们正在生成越来越多、越来越复杂的数据——这是资源不足的存储库所无法承受的。

为此，研究人员努力寻找那些一发表就能立刻获取的数据，然而即使在定位数据之后，他们也会发现很难访问。

多样性缺乏

在人类基因组计划（Human Genome Project）公布第一份序列草案后的几年里，研究人员认识到，基因组数据库过多地代表了生活在高收入国家的欧洲后裔的DNA。

真正的全球数据库和存储库需要能够正确代表人类巨大基因多样性的数据。但是二十年来这一点一直没能实现，这是科学界忽略和不平等对待低收入国家人群的证明，尤其是对非洲人和土著居民。可以理解的是，来自这些人群的许多人认为，参与研究可能也不会造福于自己，甚至有可能造成伤害，因此他们本身对参与研究也持谨慎态度。例如，当疾病与某一特定人群有关时，就可能导致污名化和歧视。

由非洲科学院召集的一个研究人员委员会正在敦促国际资助者更多地考虑那些为基因组学贡献数据的参与者的需求和愿望。这包括更适合特定研究目的的知情同意协议，而不是通常要求的广泛同意。最好的方法是由来自不同社区的人组成的团队共同进行这项研究，所有人在研究过程中享有平等的份额，在研究结果中享有平等的利益。

在这个具有里程碑意义的周年纪念日里，基因组学界——包括来自世界各地的资助者、期刊、研究人员和参与者——需要重新承诺开放数据共享。与此同时，研究人员必须与参与者建立更紧密的伙伴关系——投入更多时间参与、建立信任、倾听和采取行动来解决问题。这必须被视为基因组学研究的必要部分，并将是其未来的关键。

研究人员还需要承诺完善数据存储库的标准。存储库必须变得更易于访问，并且上传数据也更方便。此外，管理需要更好地反映不同的视角，不仅是全球基因组学研究界的视角，也应包括基因信息提供者的视角。

正如在大流行期间反复看到的那样，迅速的数据共享可以为科学带来巨大利益，并通过科学造福社会所有人。现在是时候巩固百慕大原则，改进数据贡献实践了，但前提始终是保持公平和尊重。

一、数据共享受限将阻碍人类基因组研究

数据共享是20年前人类基因组计划成功的核心原则。现在科学家们正在努力保持信息的自由共享。

David Haussler还记得，2000年7月，当他看到第一个完全组装好的人类基因组呈现在他的电脑屏幕上时，他情不自禁流下了眼泪。他和当时的一名研究生Jim Kent创建了第一个基于网络的工具，用于探索人类基因组的30亿个碱基对。在人类基因组计划（Human Genome Project, HGP）开始10年后，HGP的成员Haussler等人在完成将测序信息拼接起来的11天，他们在网上发布了基因组的草图。几个月后，该团队在《自然》（*Nature*）杂志上发表了基因组分析报告，并且数据已经准备好与大家分享了。

据加州大学圣克鲁兹基因组研究所（University of California Santa Cruz Genomics Institute）的科学主任Haussler回忆，当时大家想的是，数据就在那里，我们将其分享给全世界。很快，世界上的每个人都可以通过网络来探索一个个染色体、一个个基因和一个个碱基。Haussler指出，这是一个历史性的时刻。在HGP于20世纪90年代早期启动之前，生物医学研究中并没有关于数据共享的严肃讨论。通常的做法是，成功的研究人员会尽可能长时间地保存自己的数据。

这一标准显然不适用于如此大规模的协同工作。如果一些国家或科学家不分享自己获取的数据，这将使该项目脱轨。因此，1996年，人类基因组计划的研究人员聚集在一起，提出了所谓的百慕大原则，各方同意在公共数据库中共享人类基因组序列，理想情况下在24小时内——没有延误，没有例外。

如今20年过去了，得益于对整个基因组测序技术以及通过对数百万个选择点进行测序以快速捕捉其中变异的基因分型技术的改进，基因组数据的规模不断暴增。这些努力已经为数以千万计的个体生成了基因读出，它们保存在全球各地的数据仓库中。人类基因组计划提出的原则后来被期刊和资助机构采用，这意味着任何人都应该能够访问为已发表的基因组研究创建的数据，并利用它们为新发现提供动力。

事情没有想象的那么顺利。

数据的爆炸式增长导致政府、资助机构、研究机构和私人研究财团开发了自己的定制数据库，以处理复杂且有些敏感的数据集。Haussler表示，各种各样的存储库，有各种各样的访问规则，没有标准的数据格式，导致了一种“巴别塔”（巴别塔指的是混乱和语言不通）的局面。

尽管一些研究人员不愿分享基因组数据，但与其他学科相比，该领域通常被认为是数据共享最充分的。即便如此，旨在促进共享的存储库经常给那些上传和下载数据的人带来障碍。研究人员讲述了他们花费数月或数年追踪数据集的故事，结果却发现了无法获取或无法使用的文件。期刊编辑和资助机构也难以监督科学家们是否遵守了协议。

许多科学家都在推动变革，但变革的速率始终跟不上基因组研究的发展速度。

临床基因组学家Heidi Rehm表示，该领

域已经认识到，重大的科学进步需要大量与疾病和健康特征相关的基因组数据。但目前基因组数据并不是兼容和共享的。Rehm就职于Hospital）和剑桥的布罗德研究所（Broad Institute）。该如何才能让世界上每一个人——患者、临床医生和研究人员——都共享数据?

2000年，加州大学圣克鲁斯分校（University of California, Santa Cruz）的研究生Jim Kent帮助收集，并分享了长达十年的人类基因组计划的成果。

障碍重重

人类基因组测序使研究与单个基因突变相关的疾病——如非综合征性听力丧失之类的孟德尔疾病变得更加容易。但是，要确定更常见的复杂疾病的遗传根源，包括心血管疾病、癌症和其它主要死亡原因，就需要在整个基因组中确定多种遗传风险因素。为了做到这一点，研究人员在2000年代中期开始使用一种被称为全基因组关联研究（genome-wide association studies，简称GWAS）的方法，对成千上万患有或没有某种特定疾病或状况的人的基因型进行比较。

事实证明，这种方法很受欢迎——自2005年以来，已经进行了超过10700个GWAS试验。新加坡基因组研究所（Genome Institute of Singapore）研究青光眼遗传基础的小组负责人Chia Chuen Khor指出，这产生了海量的数据。Khor举例说，一项有1万人参与的研究，对每个人的100万个基因标记进行研究，将生成一个包含100亿个条目的电子表格。

大多数个体层面的基因组数据现在存储于“控制访问”数据库中。它们的设立是为了应对与个人信息相关的基因组数据带来的棘手的法律和道德问题——“表型数据”包括医疗记录、疾病状况及生活方式选择。即使在匿名数据集中，从技术上讲，个人也可以被重新识别。因此，控制访问数据库会审查申请者的相关材料，并确保数据只用于参与者同意的目的。

美国国立卫生研究院（US National Institutes of Health, NIH）要求其资助接受者将GWAS数据放入其官方知识库——基因型和表型数据库（Database for Genotypes and Phenotypes,dbGaP）。欧洲研究人员可以将数据存入位于英国Hinxton的欧洲生物信息研究所（European Bioinformatics Institute, EMBL-EBI）的欧洲基因组-表型档案库（European Genome-phenome Archive, EGA）。同样，其他大型基因组数据生成机构，如加利福尼亚州森尼维尔市的营利性公司23andMe，以及伦敦的非营利性机构英格兰基因组学（Genomics England），都经营着自己的受控存取数据库。

但是将数据上传到这些存储库中通常需要很长时间。因此，Khor表示，数据库里的数据通常是“最小的和稀疏的”，因为研究人员为了避免麻烦，往往只存储了符合要求的数据。

有时数据存储在多个地方，产生了其它挑战。Rasika Mathias是马里兰州巴尔的摩约翰霍普金斯大学（Johns Hopkins University）的遗传流行病学家，其课题是研究非洲血统的人的哮喘基因。她指出，权力下放是一个巨大的问题。她是美国国立卫生研究院国家心肺血液研究所（National Heart, Lung, and Blood Institute）运行的精准医学项目TOPMed的一员。TOPMed由来自80多项研究的超过15.5万名研究参与者组成，并在多个资料库中共享数据，包括dbGaP和一些大学门户网站。

Mathias认为这是一个非凡的资源。但对于一个局外人来说，找到所有可用的数据并请求访问是很麻烦的。申请访问需要提交一系列说明和证明材料。这是不必要的困难。

许多人会寻找变通办法。例如，纽约市西奈山伊坎医学院（Icahn School of Medicine）的遗传流行病学家Ruth Loos不去下载dbGaP

数据，而是直接去找那些数据的贡献者，问他们是否愿意合作。几年前，她试图访问dbGaP数据集，提交了多轮电子文件，但都被拒绝了。甚至登录dbGaP也很麻烦。这对研究人员并不友好。

马里兰州贝塞斯达的美国国立卫生研究院生物技术信息中心（National Center for Biotechnology Information）负责dbGaP的Stephen Sherry承认，提交和获取数据的过程是“不完美和痛苦的”；而且，复杂的、异类的数据需要逐案审查，不能简单地通过“让更多的人参与来加快速度”以加速审查。

但是，Sherry指出，NIH正在投资使这个系统现代化，以使其更加精简和灵活。NIH负责科学政策的副主任Carrie Wolinetz表示，目前还没有确定究竟是会升级成dbGaP 2.0，还是会找一个替代资源。是在原有基础上进行升级，还是全部从头翻新？

许多研究人员表示，尽管在共享基因组数据时受控访问会带来种种问题，但dbGaP和英国生物银行（UK BioBank）等数据库仍是无价的。这些数据库保存着50万人的基因组数据。Mathias极力保护TOPMed的参与者，并认为控制访问提供的保护是有价值的。像许多人一样，她希望看到存储库能提供更好的资源获取方式。但是，她提倡权力制衡。而另一些人则乐于申请访问权限，尽管很难获得。在澳大利亚墨尔本的沃尔特和伊莱扎霍尔医学研究所（Walter and Eliza Hall Institute of Medical Research）经营统计遗传学实验室的Melanie Bahlo认为，产生那么多的数据超出了我们的能力范围。她的实验室非常愿意通过数字文件来使用dbGaP，并且已经为十多个项目这样做了。最近，她还花了六个月来寻找一组本应通过研究机构的数据门户公开的数据，但却一无所获。

Khor指出，没有什么比从dbGaP和EGA中获取数据更困难的了，但怎么都比从一个不愿意分享的研究人员那里获取数据要简单得多。

数据分享管理

人类基因组计划出台20年后，还没有具体的普遍政策去规定研究团体必须共享他们的人类基因组数据，或者以特定的格式或数据库共享。尽管如此，许多期刊仍继续遵守百慕大原则，要求在发表文章时基因组数据必须在批准的数据库中共享。但这些政策的执行可能并不“走心”。

伦敦《自然》杂志遗传学和基因组学的高级编辑Michelle Trenkmann表示，作者往往不愿分享，理由是担心参与者的隐私、同意，或国家/公司对数据所有权有相关规定。值得注意的是，遗传学家希望数据被共享，但有时他们不想分享自己的数据。在这种情况下，Trenkmann回应指出，如果挑战无法克服，作者必须在论文中直接阐明他们要求数据保护的理由。（《自然》杂志的新闻团队在编辑上独立于期刊团队。）

《基因组研究》（*Genome Research*）杂志有一个“不例外（no exceptions）”的政策。该杂志的执行主编Hillary Sussman解

释，编辑们将在个案的基础上克服与作者共享数据的障碍，找到解决方案，甚至可以要求作者重新向他们的机构审查委员会申请批准，回到参与者那里重新获得他们的同意，或者在删除不可共享的数据后重新运行分析。《基因组研究》杂志会拒绝那些预先声明他们不能共享数据的作者。社区和资助者要求这种透明度和再现性。

但是，即使作者同意共享数据，编辑和审稿人也无法确定数据共享是否真的执行。他们可能没有时间或访问控制访问数据库来检查数据质量、格式或完整性。

Trenkmann表示，资助者应该要求研究人员从项目一开始就有一个具体的数据共享计划。这可能有助于转变人们的态度，让研究人员将分享视为一种责任。

一项将于2023年1月实施的NIH范围内的数据共享政策就做到了这一点。它要求所有NIH的拨款申请人在他们的拨款提案中加入一个数据管理和共享（Data Management and Sharing, DMS）计划，并允许研究人员将他们的部分预算分配给这项任务。

这应该确保统一的数据共享与伦理和隐私原则，以及公平原则——这意味着数据必须是可发现的、可获取的、可互操作和可重用的。国家人类基因组研究中心（National Human Genome Research Institute）基因组科学部主任Carolyn Hutter如此评价：这并不意味着，我把我的数据随意放在那里，然后希望有人发现它。

Hutter补充指出，它的执行部分是棘手的，因为数据共享往往是在项目的最后一部分。与期刊编辑一样，拨款管理人员只能对出现在年度进度报告中的任何数据共享链接进行抽查。

寻找解决方案

有一些方法可以更简单地分享数据，而不涉及专利或隐私问题。许多基因组的利益相关者同意，GWAS数据的聚合形式，称为GWAS摘要统计，可以而且应该广泛而自由地共享。这些总结包括与多个基因组中发现的疾病或状况相关的每个遗传变异的汇总得分。它们对研究人员来说更容易合作，也保护了参与者的隐私。

许多研究财团在其网站或门户网站上分享这些。EMBL-EBI和NHGRI之间的一项开放获取的合作，称为GWAS目录，正在朝着一个集中的、标准化的解决方案努力。

从2020年开始，GWAS目录为研究人员提供了一种提交摘要统计数据以及描述研究和参与者元数据的方式。研究人员将获得一个出版前的登录ID，用于访问预印本和提交的手稿。

但许多研究人员表示，汇总统计数据不足以推动基因组科学的发展。Chris Amos认为，这是GWAS的一个主要威胁。Amos是德克萨斯州休斯顿贝勒医学院（Baylor College of Medicine）研究肺癌的遗传流行病学家。研究人员需要个体层面的基因组数据和相关的表型性状数据来准确揭示遗传变异在疾病中是如何发挥作用的。他们也需要完整的数据来检验分

析的可靠性。Amos 认为，如果你没有原始数据，你就无法考察数据质量。这还不足以得出可重复的发现。

而且像23andMe和Genomics England这样的大型群组的数据所有者，不会无限制地提供他们的摘要统计数据。他们表示担心参与者的数据隐私，并希望保留数据的所有权。实际上，他们运行自己的控制访问数据库，使用自定义的过程来访问和重新分析他们的数据。研究者如果需要获得大部分数据，并利用这些数据发表文章，那么一个先决条件是需要将这些公司添加为共同作者。Bahlo指出，这种要求对她和其他希望分析英国基因组学10万个基因组计划数据的生物信息学家来说，可能有些不能接受。

Hutter承认，并不是目前所有的基因组数据共享带来的麻烦都能简单地通过改进dbGaP或在GWAS目录中共享摘要统计信息来解决。dbGaP的定位并不是发展和处理每一种新型数据。例如，存储全基因组数据的成本与存储GWAS数据的成本非常不同。因此，NHGRI创建了一个基于云的基础设施，名为分析、可视化和信息学实验室空间（Analysis, Visualization, and Informatics Lab-space, AnVIL），研究人员可以在这里共享和分析大型基因组数据集，包括整个基因组和外显子组序列。

NIH的另一个项目是研究人员认证服务（Researcher Auth Service, RAS），它将授权研究人员访问AnVIL、dbGaP和其它一些数据资源。Sherry表示，他们希望像推出签证一样，给科研人员颁发数据库图书证。这样，研究人员最终就可以在基于云的系统中随意合并和分析数据了。Sherry等人正在为研究人员建立第一批图书证系统。

Haussler和其他一些大数据专家也有自己的想法。2013年，由于数据共享方面的挫折不断出现，Haussler、David Altshuler、Eric Lander等人成立了全球基因组与健康联盟（Global Alliance for Genomics and Health, GA4GH）（go.nature.com/3app3xr）。它与HGP有着同样的理念。Haussler指出，他们的初衷是要让全世界在一个大数据库上共享数据，并就如何使用这些数据达成一致。但很快，他们就发现，那是完全不可能的。

相反，GA4GH现在专注于为世界各地的众多基因组数据库建立标准。这项工作的假设是，从技术上来说，协调数据（就像规模更大的GWAS目录）和联合（或松散地链接）不同的数据仓库是可能的。

GA4GH首席执行长Peter Goodhand用全球移动电话通信进行了类比。手机制造商和服务提供商之间存在着巨大的竞争，但最终，他们都必须在同一个网络上工作。Goodhand表示，为了实现真正的互操作性，提供者之间必须有工作关系，你可以建立允许共享的系统，让共享变得更容易。

例如，科学家们使用GA4GH标准来创造媒介交换。这项服务可以让临床医生和研究罕见疾病的研究人员搜索由8个国际数据库组成的单一联合网络，以找到与他们正在研究的病例具有相似基因型或表型的个人。如果返回匹配项，则双方以既保护患者私密性又保护研究所有权和作者的方式进行连接。NIH的RAS也将使用GA4GH标准，称为数据存储库服务，这是一种帮助不同存储库进行通信的软件界面。

Bahlo等人指出，随着深入挖掘显型数据成为该领域的核心，数据联合的努力变得更加重要，显型数据的范围和复杂性都在增长。这些吸烟状况、医学成像数据。

她和其他人认为数据联盟是将全球公平理念贯彻到基因组数据共享中的大好机会。来自发展中国家的研究人员可以访问和使用数据集

，而不需要生成他们自己的数据或拥有他们自己的超级计算资源。更好的数据共享也应该改善非白人、非欧洲人的全球祖先的代表性。在GWAS参与者中，非洲大陆血统的人所占比例不足0.5%，这一点尤其明显。Haussler认为，积极的同伴压力应该说服科学家以更好的方式分享。这种需求只会越来越大。在互联网上发布第一个人类基因组20年后，他的团队为任何人探索SARS-CoV-2病毒基因组开拓了一条途径。

Haussler表示，数据应该是有生命的东西。点击一下数据库中他人共享的数据之后就能获取和处理。这是我们数据共享的动机。如果你不分享，你就无法实现这一点。

二、人类基因组计划20年总结与反思

一项新的分析追溯了自2001年以来基因组草图对基因组学的影响，总结了人类基因组计划对文章发表、药物批准和理解疾病的影响。

人类基因组初稿发表20周年之际，我们对以下事项进行了回顾与总结：该项目推动了人类疾病遗传根源的研究，改变了药物发现，并帮助革新了我们对基因本身的看法。

本文提炼了这些影响和趋势。我们结合了几个数据集来量化已经被发现的、不同类型的基因元件、相关的出版物，以及科学发现和文章发表模式发生的变化。我们的分析涵盖了包括38546个RNA转录本、大约100万个单核苷酸多态性（single nucleotide polymorphism, SNP）、1660种有记载的遗传根源的人类疾病、7712种获批和测试中的药物，以及在1900年至2017年期间发表的704515篇科学论文。

我们的分析结果显示，人类基因组计划（Human Genome Project, HGP）通过其全面的蛋白质编码基因列表，开辟了一个揭示基因组非编码部分功能的新时代，并为药物开发铺平了道路。至关重要的是，研究人员在揭示细胞构建模块之间的相互作用时，从传统单基因视角的基础上，开始转变为从系统视图来看待生物学问题（图：大规模合作成为趋势）。

当然，我们承认本分析具有一定局限性。例如，对于一个基因的起始和终止位置，甚至某些基因的具体序列，人们还没有达成共识。由于一些基因组元素有多个名字，所以有时我们的方法没有将它们联系起来。并且，一些作者可能没有将出版物和基因元件之间的关联添加到数据库中。最后，考虑到文章的发表时间和数据库的收录时间可能存在时间差，我们的图表只包含2017年及之前的数据。

然而，我们不认为这些问题会影响基因组研究随着时间的推移而改变的整体趋势。当我们归一化同期生物出版物的增长时，这个趋势仍然存在。虽然我们没有对基因发现后的时间进行归一化处理，但我们认为，这并不会影响结论。

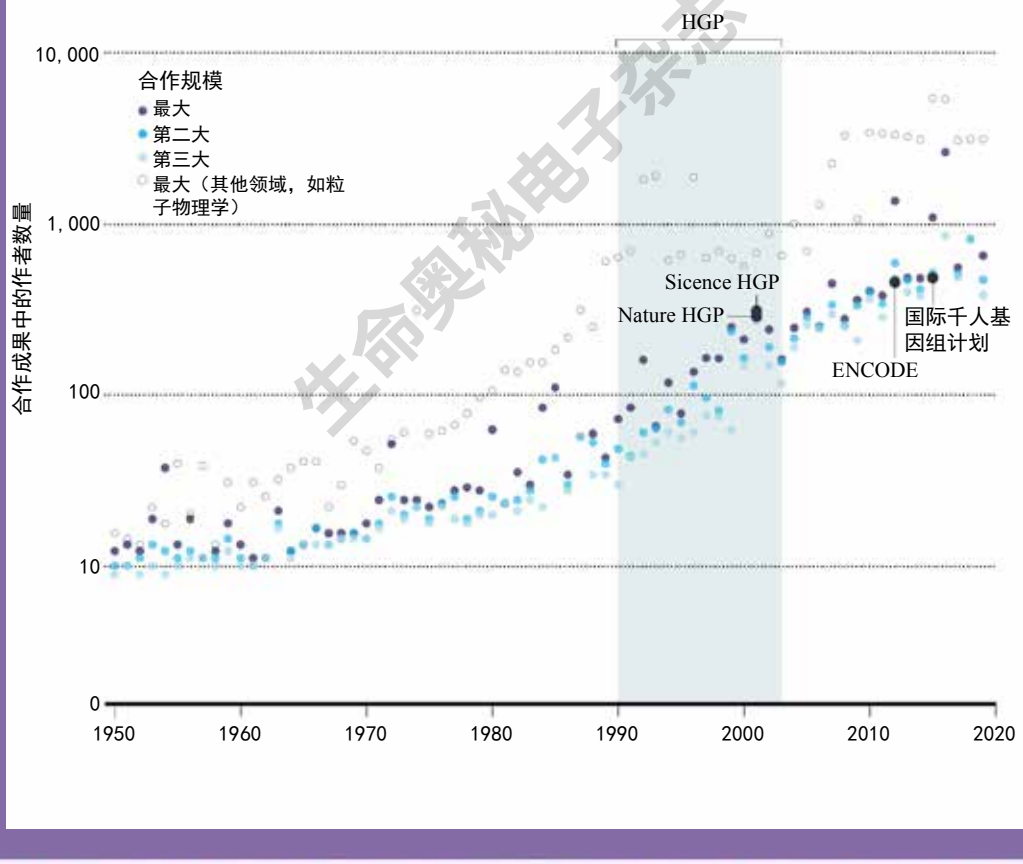
这些趋势提供了HGP前后研究格局演变的快照。数据显示，研究人员的大部分注意力聚焦于少数“顶流”蛋白质编码基因上，这可能会相应地削弱对其它潜在基因的研究。基因组的非蛋白质编码部分，以及理解遗传物质和蛋白质之间的相互作用，已经成为一个重点领域。药物的发现往往集中在少数蛋白质靶点上。

其中一些趋势对生物学家来说是熟悉的，但是量化和图表化能让人有更深刻、更直观的了解。

我们难以构想没有HGP的世界，因此也不能下定论说，没有HGP，这些趋势就不会出现。其它因素，从增加的计算能力到复杂的测序方法，也推动了这些趋势的发生。然而不容置疑的是，HGP的目录催化了持续的基因革命。

大规模合作成为趋势

人们普遍认为，HGP 论文合作作者的数量标志着一个重大变化。事实上，自 20 世纪 50 年代以来，生物学的团队规模一直在增长。



明星基因

普遍的看法是，HGP标志着对蛋白质编码基因密集搜索的开始。事实上，2001年人类基因组草图标志着长达数十年搜寻工作的结束。1902年随着激素分泌素（*SCT*基因）的发现，第一个蛋白质编码基因的证据就出现了，这比DNA结构的发现早了50年，比基因组测序开始普及早了75年。我们的分析显示，从1990年HGP开始到2003年完成（基因组草图于2001年发表），被发现（或“注释”——指标明基因功能）的人类基因数量急剧增加。到2005年左右，被发现的蛋白质编码基因数量突然趋于平稳，约为2万个（图：二十年成果统计），这一数字远低于科学界此前所估计的10万以上。

虽然蛋白质编码基因的发现进入了平稳期，但在HGP之后，对单个基因的兴趣迅速增长。自2001年以来，每年都有1万到2万篇提到蛋白质编码基因的论文发表。

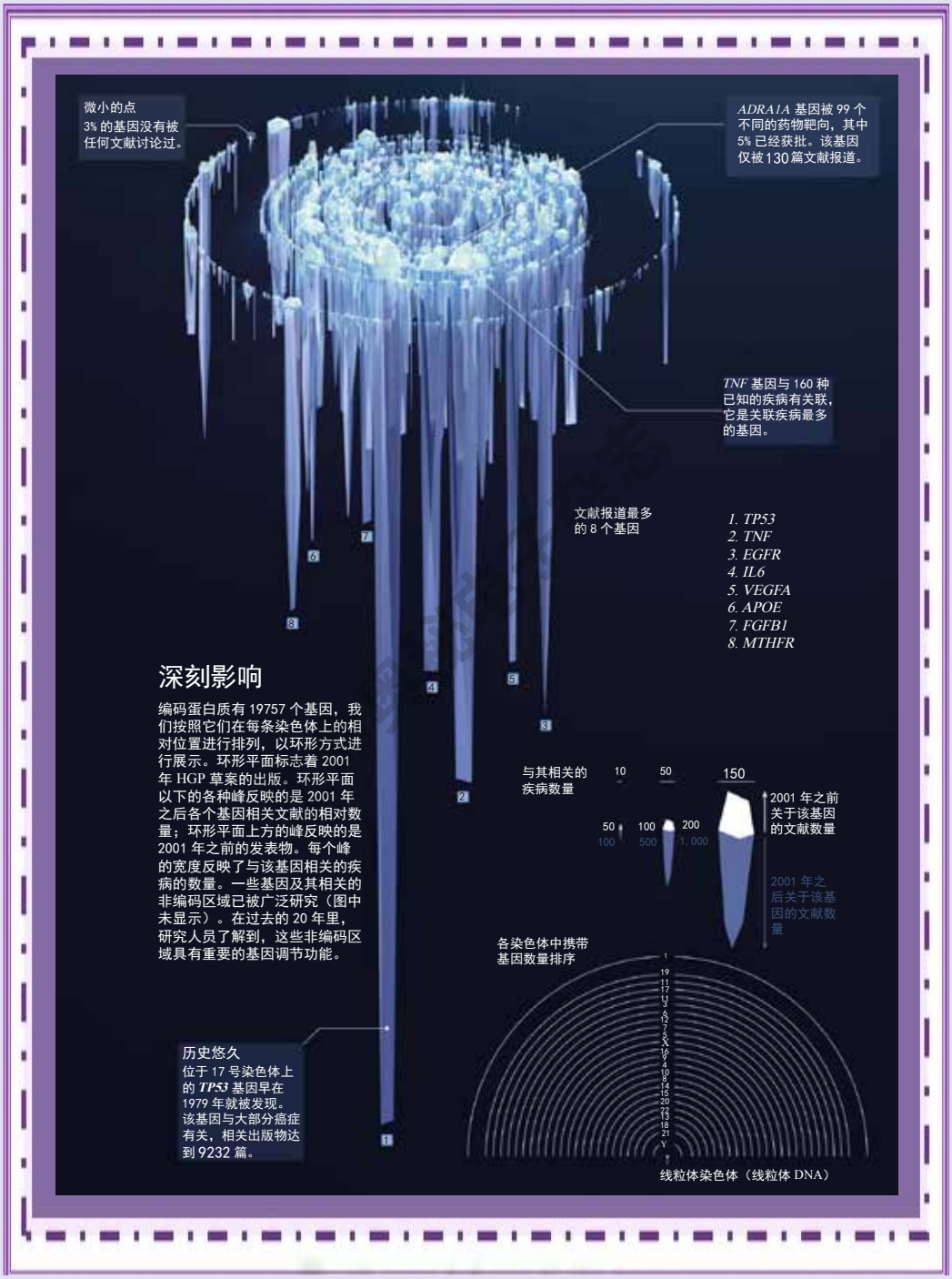
然而，这种兴趣主要集中在少数基因上。在1990年之前，*HBA1*是研究最多的基因，因为它编码了成人血红蛋白中的一种蛋白质。从1990年开始，由于CD4蛋白参与T细胞免疫和作为艾滋病毒的细胞受体，人们的注意力转向了*CD4*（根据累计发表的论文数量）。然而，与2001年人类基因组草图之后人们对少数几个明星基因的关注相比，这两个基因就算是“十八线”了。一些超级明星基因，包括*TP53*、*TNF*和*EGFR*，每年都有数百篇研究论文，而其它大多数基因很少受到关注（图：深刻影响；图：二十年成果统计）。我们发现，到2017年，22%的基因相关出版物只引用了1%的基因。

当然，对具有深远生物学意义的基因进行深入研究是合理的。*TP53*就是一个很好的例子，它对细胞的生长和死亡至关重要，失活或改变后会导致癌症。在超过50%的肿瘤序列中发现了这种基因的变异。1976年至2017年，有9232篇论文提及该基因。

一些人可能会认为，对特定基因了解得越多，就越有动力去探索基因组的其它部分。然而过去20年却与之相反：更多的注意力被花在了少数基因身上。尽管在该基因组草图发表十周年之际，这个问题就被反复提及，但并没有进行实质性的纠正。

之前对其它非常不同的系统（从人类社会网络到万维网）的研究表明，这种巨大的不平衡可以用一种植根于社会因素的马太效应来解释。或许随着关注*TP53*的论文数量不断增加，*TP53*的后续工作会更容易获得资金、指导、工具和引用，因为这很容易出成果。在社交网络中，这种现象被称为“偏好连接”（preferential attachment）。事实上，我们发现，关注特定基因的年出版物往往与既往的、与该基因相关的出版物数量成正比。

现在生物学面临的一个挑战是理清下一步研究的方向。研究人员究竟是该把金钱、时间和精力投入到最重要或最紧急的事情上，还是投入到更有可能获得资助和回报的重复工作上？

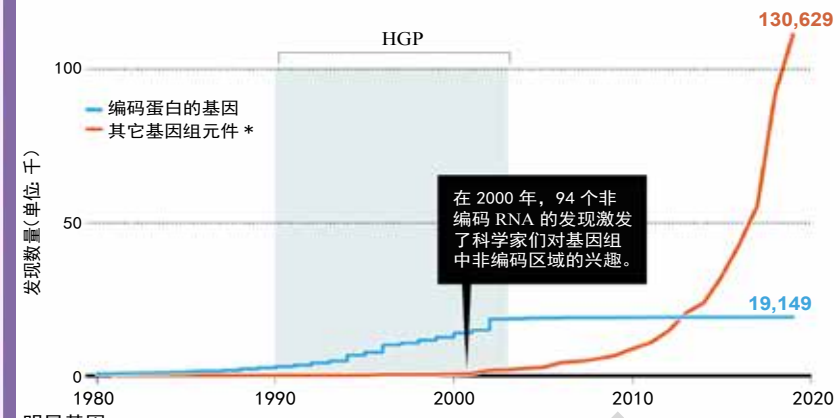

微小的点
3% 的基因没有被任何文献讨论过。
ADRA1A 基因被 99 个不同的药物靶向，其中 5% 已经获批。该基因仅被 130 篇文献报道。
TNF 基因与 160 种已知的疾病有关联，它是关联疾病最多的基因。
文献报道最多的 8 个基因
1. TP53
2. TNF
3. EGFR
4. IL6
5. VEGFA
6. APOE
7. FGFB1
8. MTHFR
深刻影响
编码蛋白质有 19757 个基因，我们按照它们在每条染色体上的相对位置进行排列，以环形方式进行展示。环形平面标志着 2001 年 HGP 草案的出版。环形平面以下的各种峰反映的是 2001 年之后各个基因相关文献的相对数量；环形平面上方的峰反映的是 2001 年之前的发表物。每个峰的宽度反映了与该基因相关的疾病的数量。一些基因及其相关的非编码区域已被广泛研究（图中未显示）。在过去的 20 年里，研究人员了解到，这些非编码区域具有重要的基因调节功能。
与其相关的疾病数量
10
50
150
50
100
100
500
200
1, 000
2001 年之前关于该基因的文献数量
2001 年之后关于该基因的文献数量
各染色体中携带基因数量排序
历史悠久
位于 17 号染色体上的 TP53 基因早在 1979 年就被发现。该基因与大部分癌症有关，相关出版物达到 9232 篇。
线粒体染色体（线粒体 DNA）

二十年成果统计

通过文献计量分析追踪，我们梳理了基因组学研究人员研究了哪些内容、对应时间，以及为什么。

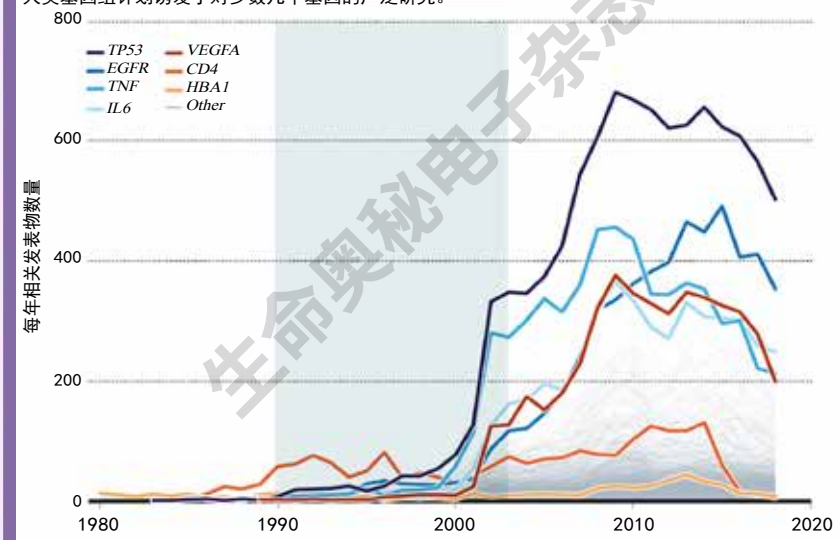
非编码元件

大多数蛋白质编码基因是在 HGP 初稿发布（2001 年）之前发现的。在那之后，许多以前被称为“垃圾 DNA”的其它基因组元件也开始受到关注。



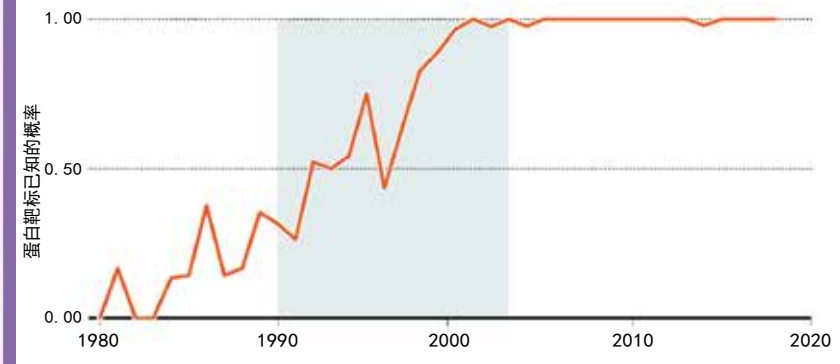
明星基因：

人类基因组计划诱发了对少数几个基因的广泛研究。



药物靶标

在 2001 年之后，几乎所有美国获批的药物都明确了潜在靶点。



* 包括单核苷酸多态性、假基因、非编码 RNA、启动子等

“垃圾”DNA不垃圾

早在HGP开始之前，就有一场大争论：花费大量人力物力去绘制被称为垃圾DNA（junk DNA）或基因组暗物质（dark matter of the genome）的大量基因组非编码区是否值得？在很大程度上多亏了HGP，现在人们认识到人类基因组中的大多数功能序列并不编码蛋白质。相反，长链非编码RNA、启动子、增强子和无数的基因调控基序共同工作，使基因组变得有生命。这些区域的变异不会改变蛋白质，但会扰乱控制蛋白质表达的网络。随着HGP草案的发布，非蛋白质编码元件的发现呈爆炸式增长。到目前为止，这种增长速度比发现的蛋白质编码基因快了5倍，而且没有放缓的迹象。同样，在我们的数据集所涵盖的这段时间（1900年至2017年），关于调控基因表达的非编码RNA的论文就有几千篇。

HGP还提供了一种对人类基因突变（包括SNP）进行分类的方法。其他一些大的举措降低了对数千人进行共同差异分析的成本，这些项目包括人类基因组国际单体型图计划（International HapMap Project，第三阶段是最后阶段，于2010年完成）和国际千人基因组计划（1000 Genome，2015年完成）。这些数据集，结合统计分析的进展，开创了无数性状的全基因组关联研究（genome-wide association study, GWAS），包括身高、肥胖和对复杂疾病（如精神分裂症）的易感性。

现在每年有超过3万篇论文将SNP与性状联系起来。这些SNP中有很多位于曾经被忽略的非编码区域。

细胞功能依赖于遗传物质和蛋白质之间的弱联系和强联系。绘制出这个网络是对孟德尔遗传定律的补充。现在，超过30万个调控网络的相互作用（蛋白质与非编码区域或与其它蛋白质结合）已被绘制出来。

助力药物发现

大约在20世纪80年代之前，药物发现主要靠偶然。药物分子和蛋白质目标通常是未知的。在2001年之前，知道一种药物所有蛋白质靶点的概率小于50%。HGP改变了这一情况。现在，几乎所有每年在美国获批的药物都有明确的蛋白质靶点（图：二十年成果统计）。

在HGP揭示的约20000种潜在药物靶点蛋白中，我们发现到目前为止只有约10%（2149种）是获批药物的靶点。这意味着，90%的蛋白质不会受到现有药物的影响。在我们的数据集中，实验药物使这个数字增加到3119。同样，对这些问题的关注是极不均衡的。目前批准的所有药物中有5%（99种不

同分子）靶向ADRA1A蛋白，该蛋白参与细胞生长和增殖。

如前所述，这种不平衡或许有更好的解释。例如，有些蛋白质可能对人类健康更重要，或者更有可能作为药物靶点。有些可能无法成药。如果研究人员、资方和出版商不那么厌恶风险，未来可能会有更多的蛋白质被探索，成为新的药物靶点。

事实上，大多数成功的药物并非直接针对单个疾病基因。相反，它们靶向的是蛋白质的一两种相互作用，控制该作用缺失的结果。例如，对现有的、可用于抗COVID-19的药物进行大规模筛选发现，只有1%有希望的候选药物以病毒蛋白为靶点——大多数药物调节的是人类蛋白。这种作用于相互作用网络的药物具有巨大潜力。

结论

总之，我们认为HGP最重要的贡献是，它开创了基因组学的新时代，而不是揭示了蛋白质目录本身。正如复杂系统理论所表明的那样，要理解任何系统，对组件的准确调查是必要的，但不是充分的。复杂性来自于组件之间相互作用的多样性。在HGP的基础上进行了20年的研究之后，生物学家现在对定义生命的网络结构和动态有了初步了解。

三、对非洲三百万人进行基因测序

尼日利亚埃德救世主大学（Redeemer’s University）的 Christian Happi 计划对人类基因组进行测序。

捕捉全方位的基因变化，以改善全球的医疗保健、公平和医学研究。

在人类基因组计划（Human Genome Project, HGP）完成二十年后，要确保基因组学为全球公共利益服务，还有很多工作要做。对高收入国家人口的关注，实际上妨碍了对人类健康和疾病的了解。迄今为止，被分析的人类基因组中只有不到2%是非洲人的基因组，尽管事实上非洲作为人类起源地，比其他大陆都包含更多的遗传多样性。由于医疗保健系统的不平等、当地研究人员人数少和资金缺乏，基因组学的知识和应用很少使非洲受益。

我目前领导的非洲人类遗传学学会(African Society of Human Genetics, AfSHG)成立于 2003 年，旨在帮助解决非洲在遗传学研究领域的滞后、改善教育、加强网络和建设研究能力。尽管最近取得了进展和投资，但在非洲进行的基因组研究大多是由欧洲和美国研究人员推动的。为什么这是个问题？因为欧美研究者的优先事项可能与非洲大陆人民的需要脱节。测试新疗法更有可能产生高影响力的论文，但为现有疗法测试更有效的改进方法通常更有可能挽救生命并减轻痛苦。

HGP 构建的参考基因组序列缺少来自非洲祖先基因组的许多变体。2019 年的一项研究估计，代表非洲人口 DNA 的基因组中的DNA比现有的参考基因组多10%。去年，仅对非洲 50 个民族语言群体中426 人的全基因组序列分析揭示了超过 300 万个以前未知的突变。

这些突变是一项耗资 1.8 亿美元、为期 10 年的倡议——非洲人类遗传与健康 (H3Africa) 联盟——的部分成果。该倡议支持 30 个非洲国家的研究机构，并由 AfSHG 与美国国立卫生研究院（NIH）和英国生物医学资助机构维康信托（Wellcome Trust）合作推动。H3Africa 现在即将结束，所以是时候考虑下一步的计划。

粗略估计，所有非洲遗传变异需要对大约 300 万个体进行测序，这些个体需要在整个非洲范围内被精心挑选以涵盖各个民族语言、区域和其他群体。因此，我们的目标是启动一个名为三百万非洲基因组（Three Million African Genomes, 3MAG）的项目，该项目将在非洲大陆进行基因组研究及其应用和治理方面的能力建设。这些发现将给全世界带来好处，包括一些难以预料的好处，正如应对埃博拉疫情为如今抗击COVID-19疫情奠定了基础：在 COVID-19 大流行期间应用的许多知识——从公共交流到共享生物样本和数据——其实得益于前几年埃博拉疫情中受到的锤炼。

3MAG 的开发大概需要十年左右。我们估计核心资金每年大约需要 4.5 亿美元（每位参与者总计约1,500美元）。这将用于建立和运行生物储存库以及开发数据基础设施和技术。我们计划在第一年对大约三十万个非洲基因组进行测序和表型分析。

那些认为这太大胆的人忘记了启动 HGP 所需的雄心。HGP用了超过13年，并将遗传科学注入到医疗保健的所有领域。现在，基因组测序的成本不到1,000 美元——构建第一份参考基因组草案的成本约为 3 亿美元（参见 go.nature.com/3pfy2kh）。30年前，在 HGP 成立之初，美国国家人类基因组研究所

（US National Human Genome Research Institute）将 NIH 总资金约 95% 分配给了人类遗传学研究项目。去年，这个数字是 10%，因为所有其他 NIH 研究所也支持以遗传学为基础的特定学科研究。

3MAG 的目标是在整个非洲对足够多的基因组进行测序，以构建具有代表性的人类参考基因组，并建立一个泛非的临床信息和样本生物库。从这个角度来看，英国生物银行正在进行一项为期27个月的项目，该项目对500,000个基因组进行测序（参见 go.nature.com/3ciohcj），而英国的人口约占非洲人口的 5%。（根据英国的人口测序比例，非洲需要对1000万人进行测序。）

研究收益

非洲基因组可以揭示导致健康和疾病的基因和变异，这在以前以欧洲为中心的研究中无法被发现。（来自一小群祖先的某些欧洲人口，例如冰岛，其基因组具有重要意义，因为从少数祖先繁衍而来的后代具有高度的遗传同质性，这有助于揭示具有强大影响的环境因素和单基因变异。）同理，非洲血统的人口是世界上基因最多样化的，他们共同拥有更多的遗传变异，与其他非非洲人群的混合更少，这更容易找到可能导致特定疾病的变异。

例如，*PCSK9* 基因的变异在欧洲人中极为罕见（不到 0.1%），但在非裔美国人中相对常见（约 2%）。该基因与某些血脂水平较低相关，这一发现导致至少一种新的药物（例如 evolucomab）用于治疗与心脏病发作和中风相关的血脂异常，造福了全人类。一项对大约 900 名患有精神分裂症的科萨血统非洲人的研究与 900 名没有精神分裂症的非洲人相匹配，发现了许多导致这种疾病的罕见突变，并对其机制有所理解。2016 年在瑞典人群中进行的一项研究发现了许多相同的突变——但需要的样本量是非洲研究的4倍多。

南非库努的科萨妇女。科萨人血统中存在罕见的突变。

对非洲基因组的研究也将有助于纠正不公正现象。使用在欧洲人中运行良好的工具对非洲人后裔的遗传风险评分进行预测，例如预测心肌病或精神分裂症的可能性或许并不可靠，甚至具有误导性。为了促进发现和产生可靠的临床工具，必须使用来自更多人群的基因组重新优化基因分型和分析。

关于听力障碍和镰状细胞病（sickle-cell disease, SCD）（本文原作者的研究重点）的研究说明了研究由单个基因引起的疾病可能带来的许多好处。有两个镰刀突变拷贝的人会产生畸形的香蕉形红细胞。这些畸形红细胞在血管中聚集在一起，阻止了氧气向组织分配。在高收入国家，患有 SCD 的人通常能活到50 多岁。在许多较贫穷的国家，患者容易在童年时死于细菌感染、贫血、肺部疾病、中风或其它并发症。

如果患有 SCD的人携带其它基因的变异，例如延长胎儿血红蛋白的产生，或者发生α-地中海贫血症（在该疾病中，一种或多种珠蛋白肽链合成受阻或完全抑制，导致血红蛋白成分组成异常，引起慢性溶血性贫血），那么他们会更健康。相比之下，基因 *APOL1* 的变异增加了对肾脏疾病的易感性。

大多数 SCD 的遗传修饰物是从欧洲和美国的研究中确定的，通常使用开发的基因芯片来发现欧洲人群中常见的变异。每年约有 300,000 名婴儿出生时携带这种突变，其中约 75% 在非洲，其医保并没有涵盖基因诊断相关的费用。跨大陆的多中心、协调良好的纵向研究可以发现许多变体。这将有助于预测病程，提出新的治疗途径（可能包括基因编辑），在可以进行产前基因检测时为父母提供更好的建议，并有助于控制疾病。这样的研究也可以成为理解基因组变异如何影响其它单基因疾病的模型。

三个优先事项

对 300 万个非洲基因组进行测序需要非洲政府、学术界和国际组织的支持。

合作。最直接的起点是在全球和非洲国家之间建立更多合作，包括学术和企业研究。我们可以利用 H3Africa和其它项目的技术、工作流程和最佳实践（表：铺平道路）。例如，位于尼日利亚拉各斯的 54gene 公司正在建立设施来对100,000 名尼日利亚人的基因组进行测序。据报道，硅谷投资者已投入450万美元建立了一个生物银行（https://54gene.com）。

人力。另一个需求是培训医学和技术研究人员，重视人类遗传学、信息学和计算机科学。理想情况下，这可通过在非洲大学设立研究生课程以及在非洲卫生保健设施中建立卓越的基因医学中心来实现。

治理。最具挑战性的将是开展临床研究、与政府打交道，以及建立社会基础设施，以拥抱非洲各地的不同文化和国家。来自 3MAG 和相关生物库的知识将具有深远的伦理意义。目前，缺乏关于非洲人民对某些问题的看法的研究，包括知情同意、社区参与、隐私和保密以及遗传信息的使用，也不了解他们对生物

储存库管理、利益共享和研究成果回报的看法，或者对研究合作和商业化中所存在的剥削的恐惧。当英国的维康信托桑格研究所计划将基于非洲基因组见解的基因分析套件商业化时，这些缺陷引起了争议（参见 go.nature.com/3r7elep）。南非立法机构多年来一直在讨论《个人信息保护法》。但是，迫切需要涵盖这些问题的正式伦理、法律和社会影响框架。应努力协调一致，以建立以公平为导向的基因组学研究。这些应该借鉴全球正义理论以及基于非洲的概念，例如 Ubuntu，可粗略地翻译为社区精神。

第一步

为了制定全面的目标和计划，来自非洲人类遗传学学会、非洲科学院和 H3Africa 联盟的成员必须与学者、科学家、专业协会、政府代表、卫生工作者和患者倡导者等合作。与 H3Africa 一样，这些目标可以通过工作组的定期会议（例如医学遗传学、培训、生物库、公众参与和可持续性）来完善，并穿插更广泛的聚会。

政府则需要致力于建设数据中心、开发基因医学服务和创建学术项目。他们还需要拓展公私合作伙伴关系，以研究、开发并转化到诊所，同时围绕个人数据和知情同意建立法律和道德规则，等等。在世界卫生组织或非洲联盟内设立一个有权援助和协调这一基础设施以简化跨国研究的委员会至关重要。税收和市场激励措施应鼓励私营部门将基因组学研究引向被忽视的疾病。

该项目提出了一个显而易见的问题：当人们仍然死于营养不良、疟疾和艾滋病毒时，如何证明大规模的基因组测序是合理的？

本文原作者认为 3MAG 将提高一系列生物医学学科的能力，使非洲能够更公平地应对公共卫生挑战，并产出可以使弱势群体受益的知识。事实上，有一些证据表明严重营养不良的影响与治疗受遗传变异的影响具有相关性。HGP 加速了医学发展，包括在非洲广泛使用的用于诊断 HIV 和结核病的分子技术。它还为艾滋病毒的预防和治疗提供了方法。

3MAG 可以扩大和扩展这些好处。已经在高收入人群中发现了影响 HIV 药物代谢的基因变异。多达 47% 的非洲和非裔美国人携带一种称为 *CYP2B6*6* 的变体，该变体与 HIV 药物依法韦仑的严重副作用有关，这增加了人们逃避服药和出现病毒耐药性的可能性。还应进一步开展药物遗传学研究，以将人们与非洲最有效的疗法相匹配。

3MAG 必须为将要研究的人群提供服务，并带来更好的基本医疗保健。试想一下，在无法进行乳房 X 光检查或高血压检查以及可能无法获得医疗服务的人群中，识别乳腺癌或心血管疾病的变异有什么好处？最终，3MAG 与 H3Africa 一样，必须服务于全球基因医学和非洲人口，而不是仅局限于遗传学。

铺平道路 300 万非洲基因组可以从现有项目中提取				
项目	启动时间 / 持续时间	基金	目的	成果
非洲人类遗传与健康（Human Heredity and Health in Africa, H3Africa）	2011 年；10 年	NIH，维康信托（非洲人类遗传学学会（African Society of Human Genetics）的合作伙伴），基金共计 1.8 亿美金	为非洲人建立由非洲科学家领导的合作和遗传学研究	来自非洲 30 个国家的 79,254 个人的全基因组和测序数据。建立三个生物库（尼日利亚、乌干达和南非）
疟疾基因流行病学计划（Malaria Genomic Epidemiology Network，MalariaGEN)	2005 年；仍在继续	维康信托，英国医学研究委员会（UK Medical Research Council），盖茨基金会（Bill & Melinda Gates Foundation），NIH	将基因组学研究人员与疟疾流行国家的临床医生联系起来	来自 39 个国家（其中 12 个来自非洲）的 17,000 人的全基因组数据
利用数据科学促进非洲的健康发现和创新（Harnessing Data Science for Health Discovery and Innovation in Africa，DS-I Africa）	2020 年；5 年	NIH 提供的 5800 万美金	在非洲推进数据学，以惠及临床护理、公共卫生及相关研究	尚未确定
精准医学跨组学（Trans-Omics for Precision Medicine，TOPMED）	2014 年；仍在继续	NIH	整合测序、分子和临床数据以实现个性化治疗	来自超过 80 项不同研的 155,000 名个体的、超过 90,000 个全基因组测序和全基因组数据；47,020 名非洲血统的参与者
基因组聚合数据（Genome Aggregation Database, gnomAD）	2017 年；仍在继续	博德研究所（Broad Institute）	在公共网站中聚合和协调来自大型测序项目的数据	来自各种测序项目的、超过 140,000 个外显子组（蛋白质编码区）和基因组；20,744 名非洲裔 / 非裔美国人参加
英国生物银行（UK Biobank）	2006 年；仍在继续	由英国多个机构提供的 3.323 亿美金	为医学研究提供来自多人的生物医学和遗传数据	来自 500,000 名参与者、8,066 名非洲黑人的临床和全基因组数据；200,000 个参与者外显子组

四、破解人类参考基因组中的未解之谜

自人类基因组于2001年发表以来，原始序列中的许多空白已被填补，让人们得以更详细地了解基因组调控、结构和功能。

2001 年人类基因组草图的发布是一个里程碑式的成就。 科学家们第一次可以逐个碱基地研究每条人类染色体的长链。基于此，研究人员可以开始了解单个基因的排列方式，以及周围的非蛋白质编码 DNA 的结构和组织方式。尽管取得了惊人的进展，但基因组草图仍然不完整，缺失了超过 1.5 亿个碱基。在此期间的技术进步使研究人员能够不断对草案进行补充，最终在 2020 年实现了染色体的完整测序。因此，人类基因组中新的和未表征的部分开始浮出水面，迎来另一个生物发现激动人心的新时期。

基因组草案中究竟包含了什么？原始草案包含许多以前未探索过的基因间区域，还包含绝大多数基因。国际人类基因组测序联盟（International Human Genome Sequencing Consortium）最初估计基因组包含 30,000-40,000 个蛋白质编码基因，尽管 2004 年发布的更新基因组以及改进的基因预测方法导致该数字修正为约 20,000。2004 年的基因组提供了来自常染色质（DNA更松散的包装区域，这些区域富含基因，约占人类基因组的 92%）的 28.5 亿个核苷酸的高分辨率图谱。

参考基因组将科学界带入了基因组探索时代，将重点从单一基因转移到更完整的全基因组研究。然而，23 对人类染色体中的每一对仍然存在缺口，估计包含超过150 兆碱基的未知序列（图 1）。最大的缺口位于富含高度重复的 DNA 或序列的位置，这些 DNA 或序列有许多几乎相同的副本。这些部分最初难以克隆、测序和正确组装。因此，人类基因组计划故意低估了这些重复序列。尽管研究人员对这些区域的序列性质已有一个非常基本的了解，但这些区域的高分辨率基因组组织仍然难以研究清楚。

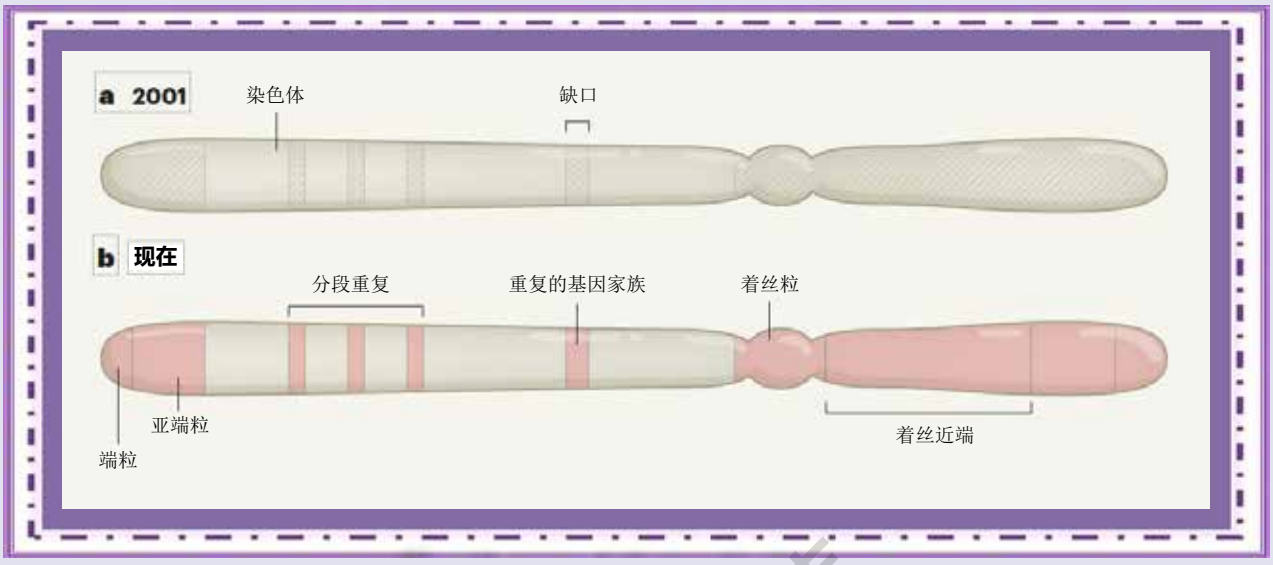


图1 填补人类基因组中缺失的序列。a，2001 年的人类基因组草案涵盖了大部分富含基因的 DNA，这些 DNA 松散地包装在细胞核中。但是，许多缺口仍然存在于富含重复 DNA 序列的紧密包装区域中，这些区域通常是未转录的（为了便于看清楚，此处夸大了缺口的整体范围）。b，随着测序和生物信息学的进步，研究人员现在可以研究所有这些缺失的序列。具体包括覆盖染色体的端粒和亚端粒区域； 细胞分裂所必需的着丝粒结构；尤其是短而高度重复的染色体臂，即(染色体)近端着丝部分。现在也可以分析一个区域内重复或分段重复的序列。

早期尝试包括使用长序列读取来跨越重复序列——但这种读取最初非常容易出错。在 2010 年代，由于读取更长序列能力的进步以及可扩展生物信息学工具的开发，这个难题被部分解决。数十到成百上千碱基的序列读数让科学家们能够研究许多中等长度缺口的基因组组织。这让科学家对一些亚端粒区域（与染色体末端端粒结构相邻的富含重复序列的 DNA）有了新的见解。它还使研究第一个着丝粒卫星阵列成为可能，其中短序列串联重复约 300 kb。部分重复的子集（共享 90-100% 碱基并在多个位置发现的序列片段）也得到解决，其中许多包含以前在参考基因组中缺失的基因。然而，许多最大的、多兆碱基大小的、重复序列丰富的区域仍然难以处理。

在过去的几年中，超长读取与高度准确的长读取数据的结合彻底改变了这一情形，首次揭示了极长的串联重复序列和丰富的分段重复序列。通过打破这些技术障碍，科学家们现在发现了广泛的富含重复序列的区域，这些区域可以跨越数百万个碱基，并构成染色体的整个短臂。

研究人员尚未完全理解为什么人类基因组的某些部分以这种方式组织。但获得这样的理解无疑是有价值的，因为这些富含重复序列的序列通常位于对生命至关重要的位点。例如，

核糖体 DNA (rDNA) 的长片段重复编码细胞蛋白质合成机制的 RNA 成分，并在核组织中发挥重要作用。着丝粒区域的重复 DNA 对于细胞分裂过程中正确的染色体分离至关重要。

这些大量重复的 DNA 从基因组织方式和进度来看，符合不同的规则集。它们还受到不同的表观遗传调控（DNA 和相关蛋白质的分子修饰，不会改变潜在的 DNA 序列），这导致重复 DNA 在其组织、复制时间和转录活性方面与常染色质不同。许多全基因组工具和数据集还无法从重复的 DNA 区域中完全捕获所有这些信息，因此科学家们还没有完整地了解哪些转录因子与它们结合、这些区域在细胞核中如何分布，或者我们基因组中这些部分的调控如何在发育和疾病状态下发生变化。现在，就像几十年前基因组的首次发布一样，研究人员面临着人类基因组中一个崭新的、未探索的功能景观。获取这些信息将推动技术创新，再次拓宽我们对基因组生物学的理解。

在过去的一年里，科学家们使用极长且高度准确的序列读取来重建从端粒到端粒的整个人类染色体。去年还看到从有效的“单倍体”人类细胞系中释放出近乎完整的人类参考基因组，仅剩下五个标记 rDNA 阵列位点的缺口（go.nature.com/3rgz93y）。在该品系中，细胞具有两对相同的染色体，与典型的人类细胞（二倍体，具有从母亲和父亲遗传的不同染色体）相比，简化了重复组装的挑战。这些信息结合在一起，首次提供了对着丝粒区域、节段重复、亚端粒重复和五个着丝近端染色体的高清照片。其中着丝近端的臂非常短，在一端几乎完全由高度重复的 DNA 组成。

人们很容易认为科学家们终于接近终点线了。然而，单个基因组组装，即使以近乎完美的序列准确完成，也不足以作为研究存在于人群中的序列变异的参考。现有的绘制基因组常染色体部分多样性的图谱必须扩展到完全捕获重复区域，而且这些重复序列的拷贝数和重复组织因个体而异。这样做将需要制定常规生成和分析完整人类二倍体基因组的策略。实现更完整和更全面的人类参考的理想目标无疑将提高我们对基因组结构及其在人类疾病中作用的理解，并与人类基因组计划的承诺和意旨保持一致。

五、罕见病医学的基因革命

孟德尔病由单个基因的突变引起，2001年发表的人类基因组草图对如何诊断、管理和预防这些疾病产生了广泛的影响。

当人类基因组的第一份草图公布时，预计它将对医学产生变革性影响。人们大胆预测，医学将发生范式转变，变得具有个性化、可预测性以及可预防性。对许多人来说，这样的转变并没有实现，可能是由于公众关注的是糖尿病和冠状动脉疾病等常见疾病。但对于孟德尔疾病（Mendelian diseases，由单基因突变引起的疾病），如遗传性癌症和多种形式的儿童发育迟缓，这些预测则是正确的。

在基因组草案发布之前，关于单个突变基因的序列和基因组位置的基本信息必须通过一个名为“克隆”（cloning）的过程来确定。在这个过程中，研究人员借助酶将人类的DNA切割出短小的染色体片段，并使其在细菌中复制，产出够的数量以供分析。克隆是一项极其繁重的工作，往往需要数年，而且只有少数几个实验室能够完成。因此，大多数孟德尔疾病的遗传基础是未知的，这使得诊断工作非常困难。即使对于少数有已知潜在遗传基础的疾病（如脆性X染色体综合征，fragile X syndrome），由于疾病在临床表现出的显著差异性及其罕见性，临床专家仍然可能无法做出诊断。

在20世纪90年代，“定位图谱”（positional mapping）方法的发展使确定与孟德尔疾病相关的基因异常变得更加容易。早期的定位图谱绘制工作涉及到比较几个患有相同疾病个体的DNA，使用包含少数已知序列的原始基因组图谱，这些序列在个体之间存在变异，这些序列作为位置标记，帮助研究人员瞄准了一个候选的致病区域。这张原始图谱可以追溯到1987年，对早期的基因发现工作至关重要。然而，分辨率低是基因探索工作的一个主要障碍。

因此，很难夸大人类基因组草图对孟德尔疾病患者及其家庭的影响。该草图并没有将单个基因与疾病直接联系起来，但它确实为诊断技术的革新提供了必要元素。起初，

它提供了丰富的标记图，使定位图谱的分辨率大大提高。然而，当基因组草图与“下一代”测序技术（next-generation sequencing technologies）结合使用时，真正的“游戏规则”改变了，这种技术可以读取整个基因组，而不是单个基因。这使研究人员有能力比之前更快地在整个基因组中识别潜在的致病变异基因。

得益于这一技术的进步，由已知遗传原因引起的孟德尔疾病的数量从2001年的1257种增加到编写本报告时的4377种（来源于OMIM数据库，这是一个有关人类基因和疾病的在线目录；go.nature.com/ omimdb）。现在，越来越多的患者从长期存在的诊断瓶颈中解放出来。许多人在紧急状态下可以在几个小时内获取诊断，其精确性在医学水平上是无与伦比的。这为真正的个性化疾病管理打开了大门。例如，对于一些特定的致病基因变异，如那些在*CFTR*基因处变异而导致的囊性纤维化疾病，也已经有了治疗方法。我们还可以避免无效的干预措施，如生长激素疗法，因为这种疗法对患有孟德尔病症——塞克尔综合征（Seckel syndrome，一种侏儒症）的儿童是无效的。

一旦孟德尔疾病和基因之间的关联被建立起来，这种疾病就是高度可预测的，这意味着预防疾病变成可能。例如，美国医学遗传学与基因组学学会（American College of Medical Genetics and Genomics）建议，出于任何诊断目的而对其基因组进行测序的人，如果他们在59个基因中的任何一个基因中携带致病变异（这些变异与可能危及生命的孟德尔病有关），都应被告知，因为这些疾病是可以进行预防性治疗的。最近在英国进行了一项测序研究，该研究包括约5万名年龄在40至69岁之间的志愿者。结果显示，2 %的志愿者携带这种可以进行改造的变异基因。早期的数据显示，基于人群的这些变异基因筛查可以提高风险管理程序的接受率。检测这些变异基因和更多后续变异基因所带来的影响（例如变异基因影响机体对药物的反应），让我们看到了未来基因组测序普及后的潜在医疗效益。

大规模基因组测序的另一个好处是提高了生殖能力。携带者筛查（Carrier screening）可以确定一个人是否携带了一个“隐性”的变异基因（如果相关基因存在两个变异，就会导致疾病），通常是父母双方都携带变异并将变异传给孩子（图1）。通过这些筛查，携带者可以做出明智的生殖选择。Tay-Sachs病（Tay–Sachs disease）和地中海贫血症（thalassaemia），这两种由常染色体隐性遗传变异引起的会危及生命的疾病，通过携带者筛查，已分别在纽约和塞浦路斯的高风险区域基本消除。未来，我们可以将这种模式扩展应用到所有严重或致命的隐性孟德尔疾病基因，并会受到私营部门和公共资助计划的拥护。然而，重要的是，我们需要注意，使用基因筛查进行生殖选择，存在许多伦理上的争论，人们担心会“筛出”某些群体，并带来其它社会风险。此外，对与健康无关的性状进行基因筛查被认为是不道德的。

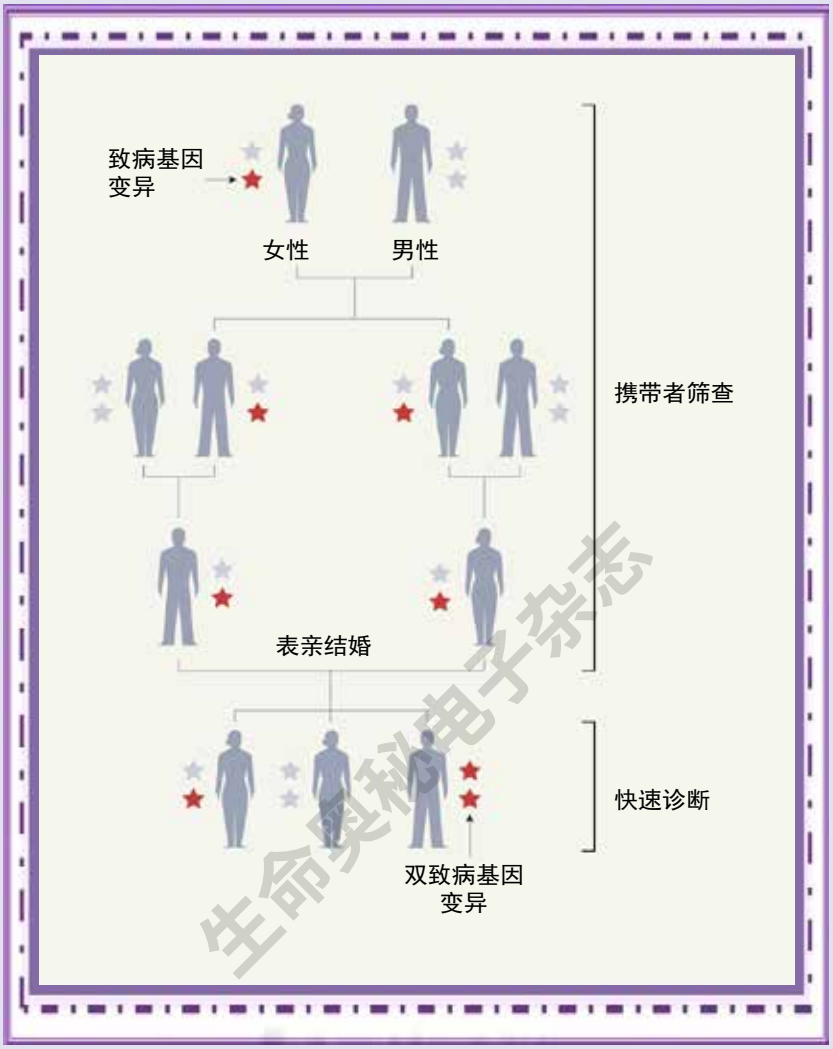


图1 孟德尔病的筛查。孟德尔病是指当一个人携带的两个致病基因变异时产生的疾病。在这个假设的家谱中，两个孩子从他们的母亲那里遗传了一个致病的变异基因，从他们的父亲那里继承了一个非损伤性的变异基因。反过来，他们又各自将一个致病的变异基因遗传给他们的孩子。如果这些堂兄弟姐妹（或任何两个携带该变异基因的人）生下孩子，每个后代都有可能遗传到两个致病的变异基因，从而患上这种疾病。人类基因组测序提升了我们识别致病变异基因的能力。今天，人们可以进行筛查，以确定他们是否携带这种变异基因，并且可以通过基因组测序快速诊断出患有这种疾病的人。

携带者筛查影响最大的地方莫过于表亲之间结合普遍的国家。由于表亲比无血缘关系的人共享更多的变异基因，他们更有可能共有以及传递有害的隐性遗传变异，从而导致隐性遗传疾病。沙特阿拉伯就是一个典型的例子。当人类基因组草图公布时，沙特阿拉伯是世界上有记载的隐性疾病发生率最高的国家。20年后，该国几乎所有的主要隐性疾病都在基因水平上得到了确认。无数夫妇在变异鉴定的基础上寻求生殖选择，该国即将推出一项扩大性的筛查方案。

我们对孟德尔疾病认识的提高，也开始使那些具有更复杂遗传基础的常见病患者受益。例如，在2020年开展的一项测序研究显示，对于一小部分患有常见病的患者来说，一个单一的基因变异就是病因。也就是说，他们患有孟德尔形式的疾病。除了因果关系外，与孟德尔病相关的基因也被发现是许多常见疾病的风险因素。常见疾病的新疗法完全是以人类基因组学为依据的，且孟德尔基因在其中起到了不小的作用。

医学遗传学界经常被指责做出空洞的承诺，但现在基因组学正在真正地改善人们的健康。这不仅是一种辩护，也激励着我们继续用DNA改写医学。

六、扩大人类基因组学多样性

在人类基因组初稿公开后的20年里，基因组测序的迅猛发展揭示了如何通过分析人类基因组的多样性来理解人类进化史和健康状况。

2001年人类基因组测序的成功被许多人认为是生物学上最伟大的成就之一。所发表的序列是由少数几个不同种族背景的匿名志愿者的DNA生成的。然而，一个单一的基因组（即使是一个基于许多人DNA生成的基因组）提供的信息非常有限。显而易见的是，如果我们要利用基因组编码的信息来更好地了解人类的健康和遗传，就需要获得更多个体的基因组信息，并进行比较。到目前为止，我们已经生成了成千上万名个体的基因组——比20年前想象的还要多。即便如此，我们才刚刚开始对不同的种群进行测序，以达到实现基因组学的多样性。

尽管在人类基因组中99.9%的部分是相似的，但0.1%的不同意味着数百万个单核苷酸多态性（single nucleotide polymorphism, SNP）——个体之间存在遗传变异的单碱基。一份包含约142万个SNP的图谱与基因组草图一起发布，这些SNP源于为该基因组草图贡献DNA的个体之间的差异。因此，人类基因组计划为分析人类突变的大规模计划提供了一个框架。

2003年，一个由研究人员组成的联盟开始从不同的个体中生成SNP的基因图谱，即国际单核苷酸多态性图谱计划（International HapMap Project）。2007年公布的第一次迭代的图谱是一个重要的里程碑，它记录了在日本、中国、美国和尼日利亚270名个体中发现的300多万个SNP。这项工作揭示了基因组是如何组织的，揭示了我们的DNA片段是如何作为一个个模块遗传在一起的，并强调了这些模块在种群内部和种群之间是如何变化的。HapMap最终扩展到包括11个群体，强调了常见人类遗传变异（human genetic variants, HGV）在全球分布方式上的差异。

HapMap项目还有助于生物技术和计算方法的发展，如全基因组关联研究（Genome-wide association studies, GWAS），它允许科学家搜索数千个个体基因组，发现与特定性

状相关的遗传变异。GWAS已经成功地识别出增加常见疾病风险的基因组区域，如糖尿病、冠状动脉疾病和克罗恩病。但GWAS主要是在欧洲血统的人群中进行的，截至2020年12月，78%的GWAS患者具有欧洲血统（go.nature.com/3ocyhql）。有几个因素导致了这种不平衡，包括对现有队列的依赖、对同质人口群体的偏好、对招募代表性不足群体的资金有限以及早期认为欧洲人的研究结果应可推广到其他群体的看法等。GWAS缺乏多样性一直被强调为科学和公平地实现基因组学前景的主要障碍之一。

1000个基因组计划（1000 Genomes Project）创建于2008年，目的是通过对来自不同地理位置的数千名个体的基因组进行系统测序，从而生成一个更全面的HGV目录，以识别常见和罕见的遗传变异。由于测序的成本不断降低，到项目完成时，该项目已经收集了来自5大洲26个族群的2504名个体（包括几个祖先混合的族群）的基因组信息，以原来无法想象的规模提供了一个详细的基因变异目录。

产出的数据导致了一系列关于HGV全球分布的前所未有的发现。例如，研究发现，大多数常见变异是全球共有，但更罕见的变异是由密切相关的人群所共有，86%的罕见变异仅限于一个大陆群体。该项目还证实，非洲人的遗传多样性高于其他群体。

大约10万年前离开非洲居住到世界其他地方的一小群人只携带了当时存在的变异的一个子集。这意味着遗留下来的HGV子集只能在非洲人身上进行研究。由于非洲各国政府的资金和投资不足，非洲在基因组研究中的代表性历来不足。几年前，拥有基因组学专业知识的非洲科学家数量有限，生物医学研究和计算基础设施欠缺。去年，非洲人类遗传与健康联盟（Human Heredity and Health in Africa，H3Africa，两位作者均是该联盟的成员），报告了来自非洲50个种族语言群体的426名个体的全基因组序列。H3Africa发现了超过300万的变异——主要来自于以前未被研究的种族群体。该研究还观察到祖先混合的复杂模式，并确定了62个基因组区域。这些基因区域在进化中保持了高频率，或许是因其在病毒免疫、DNA修复和代谢方面具有保护作用。

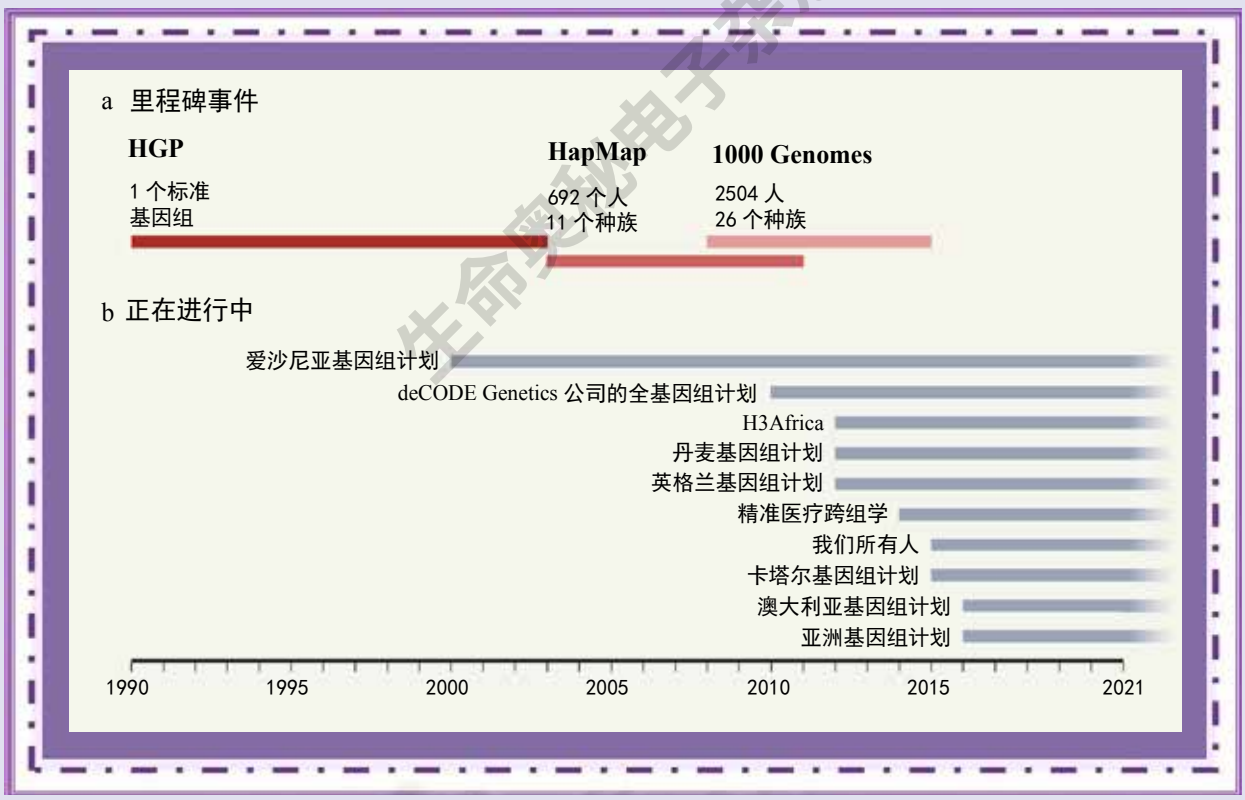
这些发现强调了——正如我们和其他人多年来一直争论的那样——增加基因组科学多样性的必要性（图“增加基因组学的多样性”）。显然，以欧洲为中心的研究不会广泛适用于所有人群。有些疾病风险变异是特定人群独有的，而多基因风险评分（根据个体携带变异的总和来量化个体发生某种特定特征或疾病的风险）可能不能很好地在多个人群中推广。2型糖尿病中的研究结果就反映了这一点。尽管2型糖尿病有一组众所周知的跨人群共有的风险变异，但在东亚、墨西哥和非洲人群中似乎已经发现了人群特有的变异。

了解人类基因组如何根据个体和群体的祖先背景聚类，无疑是有价值的。然而，我们推断出的集群可能与“黑人”、“拉美人”、“亚洲人”和“欧洲人”等社会描述词并不重叠。这一假设曾被一些人用来为种族分类辩护。迄今为止最好的证据表明，社会类别和基因群是不一致的。事实上，一项鉴定了21个全球祖先的研究报告称，6000名个体平均拥有4个祖先的DNA。这表明在基因组科学中使用非洲/黑人、西班牙/拉丁裔、亚洲或欧洲/白人等标签时需要谨慎。事实上，在基因组科学中使用这些术语应该被劝阻，除非作为自我报告的描述词或用于提供社会人口背景。使用这些术语可能会歪曲我们对历史和健康中HGV分布的理解。

未来，我们将越来越多地利用基因组学来了解人类的进化史，预测个体的疾病风险，开发疫苗等疗法，并使用DNA编辑技术治愈

包括镰状细胞贫血在内的疾病。为了能够实现这些期望，我们必须应对若干持续的挑战，包括但不限于与多样性有关的三个因素。首先是增加来自不同祖先背景的个体对基因组研究的参与。大型财团（如H3Africa财团）和国家基因组项目对基因组研究和能力建设的支持，正在应对这一挑战。第二种是发展全球合作，建立至关重要的国家间生物医学基础设施、伦理框架和公平的数据共享——国际合作的共同障碍。第三种是公平地利用基因组进步，以避免加剧健康差距，特别是在世界各地面临资源挑战的环境中。

实现这些目标将极大地提高我们对人类遗传多样性的认识，鼓励发现致病基因的努力，并促进我们对人类生物学的理解。从拥有一个参考基因组到数十万个基因组信息的过程，为研究人类遗传变异和复杂的种群历史提供了前所未有的洞见，带来了许多实际的科学和医学益处。让全人类都能享受到这些好处是下一个前沿。



原文检索：

The next 20 years of human genomics must be more equitable and more open. *Nature*, 590: 183-184.

Kendall Powell. (2021) The broken promise that undermines human genome research. *Nature*, 590: 198-201.

Alexander J. Gates, Deisy Morselli Gysi , Manolis Kellis & Albert-László Barabási. (2021) A wealth of discovery built on the Human Genome Project — by the numbers. *Nature*, 590: 212-215.

Ambroise Wonkam. (2021) Sequence three million genomes across Africa. *Nature*, 590: 209-211.

Karen H. Miga. (2021) Breaking through the unknowns of the human reference genome. *Nature*, 590: 217-218.

Fowzan S. Alkuraya. (2021) A genetic revolution in rare-disease medicine. *Nature*, 590: 218-219.

Charles N. Rotimi & Adebowale A. Adeyemo. (2021) From one human genome to a complex tapestry of ancestry. *Nature*, 592: 220-221.

张洁、郭庭玥/编译

特约编辑招聘启事

为了及时收集生命科学最新资讯、提高《生命奥秘》办刊质量，现面向从事生命科学或对这学科有浓厚兴趣的科研人员、学生诚聘特约编辑（兼职）。

职位职责：

独立完成《生命奥秘》专题的策划：对基因组学、蛋白组学、生物信息学和细胞生物学等学科的发展以及生物医学领域相关技术（例如基因诊断技术、干细胞和克隆技术、生物芯片技术等）的应用进行翻译及深入评述。

选题要求内容新颖、评述精辟、注重时效和深入浅出。尤其欢迎以自身系统研究为基础的高水平译述与评论，结合所从事的科研工作提出自己的见解、今后设想或前瞻性展望。

要求：

1.具备基因组学、蛋白组学、生物信息学、细胞生物学等生命科学学科背景；
2.具备良好的生命科学前沿触觉；
3.具备较高的外文文献翻译、编译水平；
4.具备较强的选题策划、资料搜集、组织能力，以及专业稿件撰写能力；
5.具有高级职称；或者拥有（正在攻读）该领域的最高学位。

有意者请将个人简历发送至 editor@lifeomics.com

GeneCopoeia™ Expressway to Discovery 复能基因 FulenGen 易锦生物 iGeneBio

黄金时节 惠享PLUS

促销时间：2021年9月1日~11月30日

一、克隆

1、ORF克隆 …… 699元起

2、Next-Day克隆 …… 489元起

二、qPCR系列产品

1、CytoCt™ 细胞裂解直接RT-qPCR 系列 …… 6折起

2、BlazeTaq™ SYBR® mRNA qPCR系列 …… 买一赠一

3、BlazeTaq ™ Probe qPCR Mix …… 买一赠一

4、All-in-One™ miRNA qPCR系列 …… 6折起

5、mRNA引物 …… 6折

6、miRNA引物 …… 8折

三、荧光素酶检测试剂

1、非分泌型 Luc-Pair ™ Luciferase Assay Kits …… 8折加送载体

2、分泌型 Secrete-Pair ™ Luciferase Assay Kits …… 8折

3、同一订单加购 miTarget™ miRNA 靶标（3′ UTR）克隆或 Gluc-ON™ 启动子报告克隆 …… 再享折上9折

四、外泌体解决方案

1、外泌体相关标记慢病毒 …… 7.5折

2、外泌体相关miRNA表达检测阵列 …… 7折

地址：广州科学城揽月路3号F区B楼(510663) 电话：4006-020-200 020-32068595 邮箱：sales@igenebio.com 网址：www.fulengen.com www.igenebio.com

五、病毒相关产品和服务

1、慢病毒包装小规格 ---- 2000元

2、慢病毒包装大规格 ---- 7折起

3、热门预制慢病毒 ---- 7.5折

- CRISPR 相关预制慢病毒
- 诱导 iPSC 重编程预制慢病毒
- 细胞永生化预制慢病毒
- 外泌体相关标记预制慢病毒

4、AAV包装 ---- 8.5折

- 定制纯化AAV病毒颗粒（12种血清型可选高安全性，高滴度）
- AAV血清型筛选试剂盒

5、病毒包装相关试剂 ---- 任意2件，9.5折；任意3件，8.5折；任意4件或以上7.5折

- 慢病毒包装试剂
- 慢病毒滴度检测试剂
- 慢病毒浓缩试剂
- AAV滴度检测试剂盒
- 支原体检测试剂盒（生物化学发光法/PCR法）

六、细胞株

1、预制稳转株 ---- 8.5折；加送STR鉴定

- Cas9 预制稳转株
- CRISPRa/i预制稳转株
- 荧光素酶+GFP双标签肿瘤细胞预制稳转株
- GFP标记肿瘤细胞预制稳转株
- 细胞结构相关预制稳转株

2、哺乳动物细胞系 ---- 送STR鉴定

七、转染试剂

1、EndoFectin ™ Max ---- 7折

2、EndoFectin ™ Lenti ---- 7折

3、EndoFectin ™ Expi293 ---- 7折

4、CRISPR-Fectin ™ ---- 7折

让科研更高效 让实验更顺利


合办专题专刊
网站广告合作
邮件群发推广
请致电(020)32051255

www.LifeOmics.com