生命奥秘 LifeOmics

2016年 6月刊 总第86期

生物医学大数据

精神健康APP

铁甲水上漂？

无奇不有
生命世界

做读生命
走进科学

目录 CONTENTS

专　题

生物医学大数据

下一期（2016年7月刊）预告：精准医学

下一期《生命奥秘》将介绍精准医学。精准医学将研究学科与临床实践相结合从而指导个性化病人护理工作。这篇专题介绍了将研究人员、临床实验室、临床医生和病人整合成的一个精准的医学‘生态系统’；在实现个性化治疗的道路上，人们在药物基因组学和基因疗法方面取得的进步；以及重新设计临床试验以给予病人最优治疗方面的需求。

热　点

百　态

专题

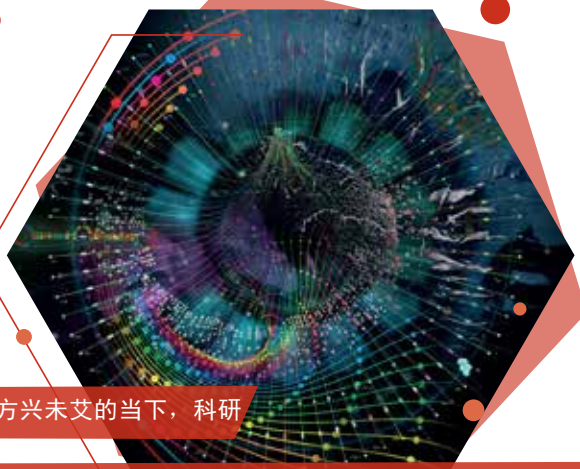
Worthy Issues

生物医学大数据

前 言

目前，对30亿DNA碱基对进行测序的费用比做一次脑部扫描还低。低廉的成本让我们获得了大规模的基因组数据，可是如何才能将这些庞大的基因组数据转化成有效的治疗方法呢？为此，生命科学家们将基因组测序、实验室研究和病人档案中获取的大规模信息整合了起来。对这些大数据的分析让我们逐渐步入精准医学时代——但是我们必须清楚地认识到，来自科学研究、工程技术和机构方面的挑战仍然存在。

一、大数据的威力
——寻找数据的意义



在个体化医疗方兴未艾的当下，科研人员们正在努力分析数据规模不断扩张的各种“组学”数据。

15年前，基因组测序技术（genome sequencing）只是一项标志性的成果；10年前，它还是一种非常有吸引力，但同时也是一项价格非常昂贵的研究手段。但到了今天，随着成本的快速下降、准确率的不断提升以及多年来的积累，它已经开始走进日常的临床诊疗工作当中。

越来越多的研究机构开始开展全基因组范围内的科研工作，希望能够发现罕见疾病的致病突变。美国斯坦福医学院（Stanford School of Medicine in California）的生物信息学家Russ Altman表示，他们在病人的基因组内发现致病变异的速度越来越快了。在某些机构里甚至能够在一半以上的患者基因组内有所发现。基因组变异（Genomic variants）还有助于我们发现“司机突变（driver mutations）”，这些都能帮助我们找到治疗肿瘤的新靶点，甚至帮助我们预测每一个个体对某种药物的反应，即了解所谓的“药物遗传学（pharmacogenetic）”信息。

只需要花费1000美元就能够完成一个人类个体的全基因组测序，这在当初被大家认为是测序技术能够被应用于个体化医疗工作的必要条件，到今天已经成为了现实。据英国剑桥欧洲分子生物学实验室下属欧洲生物信息学研究所（European Molecular Biology Laboratory's European Bioinformatics Institute in Cambridge, UK）的计算基因组学（computational genomics）专家Paul Flicek介绍，他们每单位成本获得数据的能力仍在不断提升，而且是以一种科学发展史上前所未有的速度在飞速地提升，在近七八年以来，已经增加了将近6个数量级之多。比如由美国加利福尼亚州圣地亚哥市的Illumina公司推出的HiSeq X Ten系统每年就能够完成1.8万个以上的人类基因组测序工作。

生物医学研究界正在不遗余力地用人群层面的方式去探寻基因组学信息对临床工作的价值和意义。2014年，英国启动了“10万人基因组项目（100,000 Genomes Project）”，

美国和中国也均启动了旨在分析100万个个体基因组数据的项目。

很多其它类似的工作也都在进行之中，虽然他们关注的面可能不是全基因组这么大的范围，但也都可以算是大数据级别的工作了。比如美国宾夕法尼亚州丹维尔市的Geisinger Health System公司（Geisinger Health System, based in Danville, Pennsylvania）与美国纽约的再生元制药公司（Regeneron Pharmaceuticals of Tarrytown, New York）开展的合作项目就是一项计划对25万多人进行基因测序的工作。与此同时，全世界越来越多的医院和医疗服务供应商也正在对大量的肿瘤患者和遗传病患者进行基因组测序。

与此同时，很多科研人员都担心海量的数据可能会“冲垮”我们的运算分析系统，而且对数据的存储也会提出更高的要求。有一篇文章曾经对此进行过预测，他们估计，基因组学研究产生的数据量很快就会大大超过YouTube的数据规模。然而仍旧还有很多人依旧担心如今的大数据还不够丰富，无法为临床工作提供有价值的帮助。Geisinger基因组医学研究所（Geisinger Genomic Medicine Institute）的所长Marc Williams就不太确定100万人的基因组数据是否足够满足我们的需要，但他肯定的是，现在的数据量是远远不够的。”

■ 突变的意义

当今的临床基因组学（Clinical genomics）主要关注单核苷酸变异（single-nucleotide variants），即那些在个体基因组当中有可能破坏基因功能的小错误。很多科研机构不是在整个基因组范围内寻找这些小变异，而将重点锁定在外显子组（exome）这个范围内，即对所有蛋白质编码区域进行搜索。这样一来就大大缩减了数据分析的工作量——可以减少99%的工作。但是在这个外显子组里仍然含有1.3万多个单核苷酸变异，而其中将近2%的变异可能会影响蛋白终产物的组成，因此要从中找出致病突变，也还是一项大海捞针般的工作。

数十年来，生物医学研究工作者们一直都在往“人类基因突变数据库（Human Gene Mutation Database）”或dbSNP这样的数据库里上传各种单核苷酸变异信息。而在研究工作中发现的这些突变的作用，往往都取决于使用的实验材料（比如使用的细胞系或者动物模型），甚至是理论预测模型。这就极大地削弱了这些遗传学信息在临床诊断工作中的作用和价值。据Williams介绍，在很多情况下，这些变异与疾病之间的关系其实是非常牵强的，证据往往都不够。

如果面对的是大片段复制或缺失这样的结构变异（structural variants），情况会变得更加复杂。与单核苷酸变异这样的突变相比，使用现有的测序技术将更难发现结构变异。在全基因组的层面上，每一个人都含有数百万个变异，其中有很多变异都位于非蛋白质编码序列里，但它们同样能够调节基因的活性，因此也可以成为致病突变。可惜的是，我们现在对这些调控区域的功能和范围还知之甚少。虽然科研人员们也希望能够发现这些变异，但至少在近期内，它们对临床测序工作还起不到太大的帮助作用。据Altman介绍，如果你找到了一堆你看不懂的数据，那就是自找苦吃。

不过也已经有人开始着手解决这个问题了。美国国立人类基因组研究院（US National Human Genome Research Institute）设立的临床基因组资源数据库（Clinical Genome Resource）就在专门搜集与疾病相关的变异信息，以及有可能对临床工作具有指导和参考意义的信息，同时也在搜集那些能够支持变异与疾病之间具备关联关系的各种证据。启动10万人基因组项目的Genomics England公司则正致力于构建“临床释义合作伙伴系统（clinical interpretation partnerships）”，即将临床医生与科研人员联合起来，大家共同构建各种疾病模型，以此来发现潜在的特异性疾病相关变异。

可是，数据的规模和数量与数据的质量同

样重要。能够起到决定性作用的突变会给生物进化带来不利的影响，因此这种突变肯定是极少的，所以只有通过对大规模人群进行筛查才能够发现它们的踪迹。同样，想要在疾病与效应量较小的变异之间寻找具有统计学意义的关联，也需要非常大的样本量。

冰岛的deCODE Genetics公司已经向我们展示了人群基因组学（population-scale genomics）研究的威力。他们将大量的家系与病历档案和15万人的基因组数据（其中包括1.5万人的外显子组序列数据）相结合，开展了相关科学研究。结果发现在人群中分布的遗传风险因子就包括了与乳腺癌、糖尿病和阿尔茨海默病相关的基因变异。

他们还在人体上开展过去只能够在经过人工遗传学改造的动物模型上开展的研究工作。据该公司的CEO Kári Stefánsson介绍，他们已经发现大约有1万名冰岛人的基因组里含有纯合型功能缺失突变，这些突变涉及1500个不同的基因。他们正在努力找出这些突变基因对个人的影响。

该公司之所以能够取得成功，得益于冰岛人的同源属性，而其它的一些研究反而需要更广泛的代表性。比如国际千人基因组计划（international 1000 Genomes Project）就是反映全世界基因多样性的一个代表。但其中的大部分数据仍然来自高加索人种，所以在临床上的应用价值并不大。美国哈佛医学院（Harvard Medical School in Boston, Massachusetts）的生物信息学家Isaac Kohane表示，因为他们全都来自同一个祖先，而源自非洲祖先的人要比源自非非洲祖先的人携带更多的遗传变异。在高加索人中比较罕见的变异在非洲人却比较常见，而且这些变异似乎对人体是无害的。

一部分原因还可能源自参考基因组数据，即由多国基因组参考数据组织（multinational Genome Reference Consortium）提供的、供所有科研人员进行序列比对用的“标准”序列。第一版参考序列的数据来源就是几个种族信息不详的志愿者提供的基因组样品，不过在最新的GRCh38版数据里包含了更多有关人类基因组多样性的信息。

■ 进入云端数据时代

对人群进行基因组或外显子组数据采集会产生大量的数据：每年可能会生产40pb（相当于4000万gb）的数据。但是数据存储还不是最让人担心的问题。Flicek表示，需要大量硬盘进行数据存储的基因组学家们毕竟人数还不太多，所以他们不认为数据存储会是一个大问题。

大家更加关注的问题其实是每一个人体内的、需要进行分析的变异的数量，此数量也非常惊人。据美国宾夕法尼亚州立大学（Pennsylvania State University in State College）的基因组研究人员 Marylyn Ritchie介绍，计算规模与被研究的个体数量呈线性关系。但是一旦你加入了更多的变量，比如寻找不同变量组合的关系，那计算规模就会立刻变成呈指数增长了。如果还存在其它的数据，比如与临床症状或基因表达等信息相关的数据，那这种情况就会变得尤为明显。如果对数千人进行这样的数据分析，就会让一个供小型实验室进行统计学分析的工具全都瘫痪。

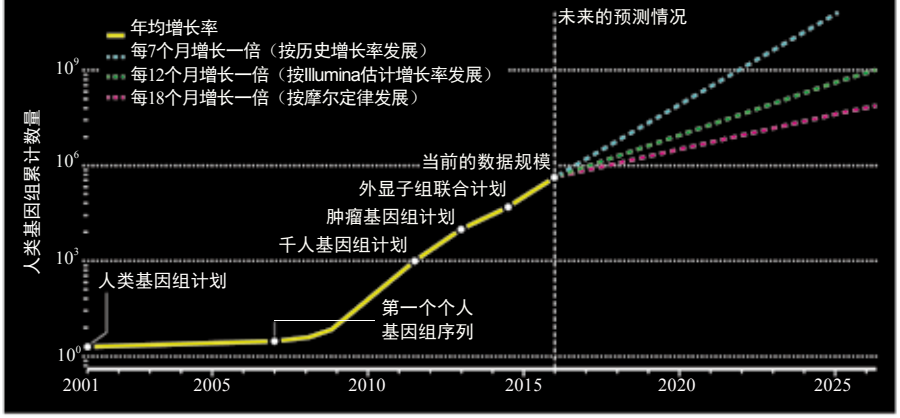
数据规模的扩大需要我们运用更多的创造力和想象力，但也不用完全从头开始。据Ritchie介绍，像气象学、金融学和天文学这种专业其实早就在处理这么大规模的数据了。Ritchie曾经参加过几个科学研讨会，他在会议上给大家介绍过他们的工作，比如对Google和Facebook的数据，或者对他们所谓的大数据进行的分析工作，但Ritchie等人的工作与那些人的大数据工作毫无可比性。所以我们应该向这些专业的人士好好请教，学学他们是怎么干做的，然后更好地开展自己的工作。

可惜的是，大量能够处理大数据的天才程序员还是更愿意去硅谷找工作。美国国立卫生研究院（National Institutes of Health, NIH）

数据科学部门的副主任Philip Bourne相信，其中有一部分原因是因为在当今这个以论文论英雄的科研大环境下，软件开发人员和数据分析经理还不是那么被人认可，发展机会也不够好。Bourne还指出，这其中有一些人也非常愿意成为学者，只不过他们在如今的环境下却不能在大学或科研机构里获得教员的身份，这是不对的。

DNA测序腾飞史

人类基因组测序正义前所未有的速度在飞速发展着。千人基因组计划已经获得了数百人的基因组信息，肿瘤基因组计划已经获得了数千个基因组信息，外显子组联合计划更是获得了6万多人的外显子组序列。图中的虚线表示未来的三种发展趋势。



科学技术的飞速发展正在改变着基因组学研究的面貌。

处理能力则是另外一个限制因素。Kohane表示，进行基因组大数据分析可不是打电脑游戏，真正的从业人员全都应该是能够同时用好几百台，甚至数千台CPU（每一台都拥有大容量存储空间）进行大规模平行运算的专业人士。很多从事序列分析的团队都已经开始使用云架构（'cloud'-based architectures）模式，这样他们的数据就能够在大量运算资源下进行处理和分析。

据基因组学英格兰公司的首席生物信息学家Tim Hubbard介绍，将你的算法带进数据里，这是一个渐进式的发展过程。对于他们公司而言，这个架构位于一个受到严格安全保护的机构里，外界很难进入该系统。但是其他的很多研究团队则会运用商业化的云服务，比如亚马逊或谷歌提供的商业云服务。

■ 个人隐私保护问题

从原则上来说，云架构可以鼓励大家进行数据分享，以及彼此之间的共同协作。但是对于患者个人隐私，尤其是涉及敏感临床信息的保护却是一大问题，因为这涉及伦理及法律的方方面面。

在欧盟内部，这种协作受到了种种限制，比如对数据处理遵循不同规则的团队之间就很难开展合作。而在与非欧盟成员进行数据共享的时候，则需要一整套繁琐的流程，或者是各合作方都必须签署限制性双边同意备忘录，以确保数据安全。为了解决这个问题，多国合作组织全球基因组学及健康联盟（Global Alliance for Genomics and Health）专门出台了一套基因组学及健康相关数据共享框架协议（Framework for Responsible Sharing of Genomic and Health-Related Data）。其中就包括关于个人隐私、知情同意、责任，以及违约之后需要承担的法律后果等相关问题的指南。

加拿大麦吉尔大学（McGill University in Montreal, Canada）的生物伦理学家Bartha Knoppers是全球基因组学及健康联盟管理及伦理委员会的主席，他表示，有了这套框架协议，我们就再也不需要签署一大摞的文件了，只需要所有参加合作的机构和个人都在这份协议书上签字就够了。该协议还号召大家共同打造一个“安全的天堂”，让整个科研界都能够使用这个个基因组数据的“央行”，能够使用里面的各种资源，不过这些资源都必须是隐去个人信息，但又没有彻底剥去个人信息（以保证其科学价值）的资源。Knoppers表示，他们希望将这些信息与临床数据和医疗档案相结合，因为他们还从未打算在其他地方获得医疗信息，所以他们只能够使用这些经过编码的数据。

将基因组数据与电子健康档案相结合正逐渐成为一种趋势，因为这对很多欧洲国家来说非常重要。据Hubbard介绍，他们的目标是将这些信息纳入标准的国家卫生服务（National Health Service）当中。英国的10万人基因组计划应该是目前在这方面走在最前面的一项工作，但是其他国家也不甘落后。比如比利时最近就宣布，他们将启动一项医学基因组项目。

所有这些国家全都得益于他们拥有一个由政府负责管理的中央医疗系统。但是在美国，情况较为复杂，因为美国有不同的服务提供商，他们也各自采用了不同的医疗档案记录系统，这些记录系统背后还有不同的服务供应商。他们在设计医疗档案记录系统时没有将复杂的基因组数据考虑在内。NIH于2007年启动了电子医疗记录及基因组网络（Electronic Medical Records and Genomics, eMERGE）系统，就是想在这方面有所改进。

■ 从数据到临床诊断

建立这些补充了基因组学信息的医疗档案系统，最直接的目的就是为了能够向临床医生们解释基因变异与疾病的关联，而其中首当其冲的就是药物遗传学（pharmacogenetics）信息。临床药物遗传学执行组织（The Clinical Pharmacogenetics Implementation Consortium）已经对PharmGKB数据库（这

是由Altman等人管理的一个数据库）中已知的药物与基因间相互作用全都做了一番临床解读。比如，某种抗凝剂可能对携带有某种变异的人的效果不太好，这会增加这类人群用药之后发生心梗的风险。据Altman介绍，问题在于，一个医生可能只有12分钟看一位病人，并且只有45秒给病人开处方，我们应该如何用更好的方式向他们传递这种信息，并对他们的临床实践工作带来影响和改变呢？

只要还有人继续从事将遗传学发现应用于临床这样一份工作，就需要花费大量的时间和精力。将基因型信息与表型信息相结合是一个非常有用的策略，也已经获得了丰厚的科研回报。大部分临床相关基因变异都是通过全基因组关联研究（genome-wide association studies, GWAS）发现的，其中有大量患有某种特定疾病的患者参加了研究，从中也发现了与该疾病最为相关的遗传信号。科研人员们现在已经可以重新寻求医疗记录档案的帮助，从中找出线索和证据，来证明某种遗传变异与特定疾病之间的关系了。

基因组学只是整个研究工作的一个组成部分，还有其它很多“组学”研究，也都是对人类健康非常有意义的。今年7月，中国华大基因的前CEO王俊从CEO的位置上退下来，成立了一个专门分析华大基因百万基因组，以及蛋白质组、转录组和代谢组数据的组织。王俊表示，他将成立一个新的研究所，专门致力于用人工智能处理这种大数据的工作。

■ 患者的力量

就在科研人员们努力将临床数据、基因组数据和其它生物数据结合到一起的同时，患者们也已准备好贡献自己的力量。 Ritchie表示，当我们关注行为、营养、体育锻炼、是否抽烟喝酒这些问题的时候，我们是不可能比患者自己更加了解他们的自身情况的。

智能手机以及FitBits这些可穿戴的设备正在尽力收集各种个人相关数据，比如体育锻炼的情况、心率等，由于每个人都很容易收集这些数据，所以这些数据的规模也在不断地攀升。

因此，我们每一个人都是一个大数据的生产者。Kohane认为，人们在家里或在户外生产的数据量都将远远超过医疗记录里的数据量。他们正在努力将所有这些数据，比如基因组数据、环境数据、医疗数据与每一个人一一对应。如果这一天真的能够到来，那将会迎来一个计算史上的大转折，今天所谓的大数据难题到那时也就是一个计算器就能够解决的小问题了。而患者将是最终的赢家。

二、高蛋白研究
——实用遗传学的挑战

斯坦福大学（Standord University）学者Michael Snyder分析了自己的基因组、RNA表达和蛋白合成。

当斯坦福大学（Standord University）的分子遗传学家Michael Snyder在自己身上使用组学工具时，他没有意识到他会迎来一些惊喜。例如，他发现他有2型糖尿病的遗传倾向，即使他没有任何常规危险因素，如肥胖或家族病史，他还是属于2型糖尿病高危人群。在接下来的14个月里，Snyder将会反复测试自己的RNA活性和蛋白质合成。

当他在研究中感染了一种呼吸道疾病时，他发现他的蛋白表达发生了变化，生物通路也被激活了。然后，他被诊断出患有糖尿病——整个事情看起来就像是感染引发了糖尿病。Snyder也监测了感染上莱姆病时，体内蛋白质发生的变化。

Snyder表示，他不知道，这个研究会变得这么有意思。他自己的组学数据已经达到了半拍字节（500000GB）。Snyder认为，这只是理论的证明罢了。

Snyder把研究规模扩展至100人，研究范围也扩展到采集测量蛋白组学等13个组学数据，包括居住在他们体内的微生物的蛋白质组和转录子组。他希望，他的团队可以收集100万名患者的深度数据，并应用大数据的工具，找出能预测疾病的差异，以加深学界对各类疾病的理解。他还希望他们能通过蛋白质组学来描述各种疾病亚型。Snyder指出，可能有100种不同类型的糖尿病。

美国东北大学（Northeastern University）的蛋白化学家William Hancock表示，Snyder的经验表明，“组学”对生物学研究具有重要意义。

■ 实用遗传学

基因为生物过程提供了指导手册，但基因合成的蛋白质则负责实现这些指令。全世界的科学家们都在努力发现和识别蛋白质，以确定它们在组织和细胞中的分布，同时也在计算特定情况下特定蛋白质的表达量，以及特定蛋白

质的亚型。来自这些研究的海量数据有助于揭示疾病的生物标志物，提供药物靶点。通过将蛋白质组学、基因组学、转录组学、代谢组学和其它组学结合起来，科学家们就能从分子层面上进一步加深对生物过程的理解。

美国密歇根大学（University of Michigan）前常务副校长，国际人类蛋白质组计划（HPP）主席Gilbert S. Omenn指出，蛋白质组学让遗传学变得更实用。Omenn表示，HPP的目的是创造人类身体的“完整的零件清单”。现在我们知道一个基因与某个疾病相关，我们想做的就是发现它真正的作用机制。

这是一份非常庞大的清单。人体包含大约20000个编码蛋白质的基因。每一个基因可以产生一种蛋白的多种亚型，这些蛋白亚型又可能发生多种不同的翻译后修饰：附带有磷酸或甲基基团，或是与脂肪或碳水化合物结合——所有这些都会影响蛋白功能。德国慕尼黑工业大学（Technical University of Munich ）的蛋白组学研究者Bernhard Küster指出，一个基因可能产生的分子非常多，很难估计，但是如果一个基因能产生100000种甚至更多的不同蛋白质，我完全不惊讶。

■ 国际合作

蛋白质组学研究是一个国际性的研究领域。人类蛋白质组组织创建了两个互补的HPP项目，两者都采用质谱法。一个HPP是基于染色体，把24个染色体的研究分配给19个国家。例如，日本负责研究3号和X染色体，伊朗则研究Y染色体。另一个HPP是基于生物/疾病，寻找特定组织和器官的蛋白质，专注于那些与糖尿病、结肠癌等疾病有关的蛋白。另一个独立的全球项目，人类蛋白质图集（Human Protein Atlas），则使用带有荧光分子或其它标签的抗体来结合和识别特定的蛋白质。

国内也有一些大型项目。中国在蛋白质组学研究上也投入了大量资金。例如国家建立了一个新的国家实验室凤凰工程，每年资金支持达到1000万美金。

无论使用什么技术方法，绘制人类蛋白质组图谱都不简单。相比而言，基因组比较简单，基因由四个核酸组装而成，除了癌症等特殊情况外，一个人的基因一生的变化不大。而蛋白质则会随时间变化，例如，运动、疾病和月经周期等都会引起蛋白质的变化。更复杂的是，人体内的蛋白质种类并不完全相同，一个人体内可能含有的最多种类的蛋白质可能是最少种类的100亿倍。基于染色体的HPP项目联合主席Hancock表示，你有一个基因组，可能会有一堆蛋白组，这取决于环境。

Küster表示，一个人体内都不存在一个固定的蛋白组，更何况是很多人体内。去年，他的团队发表了基于16857个人体组织、细胞和体液的质谱分析得到的人类蛋白质组图谱。他们还创建了ProteomicsDB数据库，并对这些数据进行分析。

■ 数据太多？

如何处理海量的蛋白组学数据是一大挑战。例如，人类蛋白质图集（Human Protein Atlas）收集使用荧光抗体标记的组织图片时，每个图像就有几十MB，上传到网络的压缩JPEG文件大约为10 MB。

与此同时，欧洲生物信息研究所（EBI）目前正在搭建一个分布式计算基础设施设计研究所ELIXIR，旨在分享欧洲各研究机构的蛋白组学和其它生物数据。瑞典皇家理工学院（Royal Institute of Technology）的微生物学家Mathias Uhlén表示，ELIXIR并不致力于打造一个庞大的数据库，更多的是连接不同国家的不同研究群体。EBI已经有蛋白识别（Protein Identifications, PRIDE）数据库，这个数据库收集了多个研究团体的质谱数据。

但科学家们就是否应该保留原始数据这个问题存在很大争议。伦敦大学玛丽女王学院（Queen Mary University of London）的生物信息学家Conrad Bessant表示，从原始数

据中识别蛋白质的技术在不断提高，所以研究者们会尽可能保留原始数据——但这极占空间。另一方面，他补充指出，领域发展非常快，你可能不会去看5年前的旧数据；仪器先进多了，你可能会选择使用新仪器再重复一次实验。

■ 补充图谱

然而，蛋白质组数据远未至完美。去年5月，《自然》（*Nature*）发表了Küster等人的基因组学结果，同时来自美国和印度的科学家们发表了据说涵盖人类基因组84%蛋白编码基因的草图。两者都基于质谱：使用酶消化蛋白，产生长约7到30个氨基酸的肽序列，这些肽的质谱用于推断蛋白质的组成。当学界质疑他们的解读方式时，两个团队最终都减少了声称发现的蛋白数量。Omenn表示，质谱是一种概率统计方法，无法排除两个不同的蛋白质含有相同的肽序列的可能性。

另一方面，人类蛋白质图集基于抗体，属于非概率方法，因为它标记单个蛋白质。人类蛋白质图集的创始人之一Uhlén表示， 这种方法的优点是它能精确地指出蛋白在器官、组织和细胞中的分别。Uhlén强调，他们提供的是蛋白分布的图谱，蛋白分布其实提示了蛋白功能。

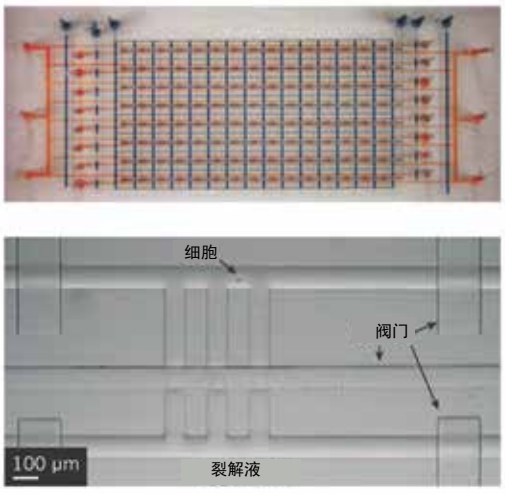
近几年来，基于抗体的单细胞蛋白组学分析微流控芯片取得了较大进展。这种方法在研究罕见细胞时非常重要，如循环肿瘤细胞。我们还能用这种芯片研究相同细胞类型的不同细胞群体之间的差异。例如，如果一个肿瘤细胞的特定蛋白质比周边细胞要多，或在一个细胞中的蛋白质发生甲基化修饰，而临近细胞没有，这就能解释肿瘤细胞发生耐药性的原因，从而提示新的药物靶点。

然而，即使抗体方法也有一定的局限性，因为一些抗体可以与多个蛋白结合，产生误导性的结果。Uhlén指出，更难的问题是如何分辨哪些数据是有效的；大数据的问题是，产生数据容易，解读数据难。

另外，还有蛋白缺失的问题。大约15%的人类基因编码的蛋白完全未知，这意味着有近3000个缺失蛋白。在某些情况下，这可能是因为这些蛋白的量很少，或只存在于组织的微小区域里。没有完整的蛋白目录，人类蛋白组学图谱总是失真。

Hancock担心，使用不完整或不准确的数据进行计算，可能会将研究者引入歧途。他认为，把生物学和数学结合起来会带来很多信息失真的问题。

但Hancock表示，随着测量技术的改进，以及科学家越来越多的发现，图谱会越来越清晰。而且由于测量技术的提高，可供筛选的数据量将继续飙升。Bessant 指出，从各种不同的实验里，我们会得到各种数据，得到数百GB或TB数据可能花不了多少时间。所以，计算带来的失真也许会越来越小。



蛋白组学芯片（上）解读了微室（下）里分别标记的细胞的蛋白。

三、发掘大数据的宝藏
——国际挖掘数据协作项目

著名演员Angelina Jolie在2013年向公众宣布，她切除了两侧的乳腺组织，以免患上乳腺癌。这是因为她的*BRCA*基因检测结果为阳性，而我们都知道，这是一个乳腺癌高危因素。在携带*BRCA1*基因有害突变的女性人群当中，有55~65%的人到70岁时会患上乳腺癌，而在携带*BRCA2*突变基因的女性人群当中，有45%的人到70岁时会患上乳腺癌。

Jolie只检出了*BRCA*有害基因，但这让她患上乳腺癌的几率大大增加，可是患上一种遗传疾病可远远不是这么简单的。最近这几年开展的一些大型研究项目正在逐渐向我们揭示这些遗传疾病的复杂特性。

数百名科研人员和计算生物学家们正在携手共同破解基因组学、蛋白质组学以及其它各种组学的奥秘。每天，都有各种蛋白质，或者各种修饰物被公布出来，科研人员们也在密切跟踪这些最新的科研进展。他们正在寻找与各种疾病相关的分子信号通路，同时也在检测其它因素的影响作用，比如细菌对人体的影响作用，即所谓的微生物组学（microbiomics）研究。科研人员们还在开发新的软件和算法，用以预测所有这些组学研究成果与人体健康的关系。与此同时，他们彼此之间也开展了广泛的交流，大家互通有无，确保每个人都能够更好地往前发展。

这些大型的研究项目有助于我们发现与某种疾病相关的风险因子，更有利于个性化医疗的研究工作，也产生了大量的数据。从基因组中发现罕见的变异，同时还必须确保没有任何遗漏，这意味着需要对数万人的基因组，进行一个碱基一个碱基的仔细筛查，而每一个人的基因组就包含30亿个碱基对。为了完成这样一份工作，全世界的临床医生正在与生物信息学家和计算科学家一起，进行着一项规模空前的研究工作。

在这个大数据的时代，他们也在不断开发出新的合作模式和策略。

■ 被掩盖的珍宝

以疾病为中心的基因组学研究方法往往都会用到GWAS研究手段，这一点已经在肿瘤研究工作当中得到了充分的证明。比如在乳腺癌的研究工作当中，使用GWAS研究方法就发现了90个乳腺癌相关变异体（即基因组序列上的小错误）。不过据美国马萨诸塞州哈佛大学（Harvard University in Boston, Massachusetts）的遗传流行病学家Sara Lindström介绍，这其中只有5个变异点位于蛋白质编码区域内。

其它85个乳腺癌相关变异的具体作用现在还是一个迷。据Lindström介绍，当你看到这种信号时，你并不清楚这是不是真的能够增加你罹患乳腺癌的风险，也许这只是与乳腺癌有关而已。要发现重要的变异体，首先需要先了解整个基因组的功能才行。

计算生物学家破解这个难题时手上能够利用的最大的一个资源就是于2003年启动的ENCODE项目，即DNA元件百科全书项目（Encyclopedia of DNA Elements）。ENCODE项目是由美国国立人类基因组研究院（US National Human Genome Research Institute）资助的合作项目，同时还运行着一个向公众开放的、可检索的基因组数据库。

2012年，来自32个实验室的442名科研人员联合发表了一篇论文，他们将人类基因组中80%以上的序列与特定的生物学功能联系到了一起，同时还发现了400多万个蛋白质DNA相互结合区域（J. R. Ecker et al. *Nature* 489, 52–55; 2012）。

美国斯坦福大学的遗传学家，同时也是ENCODE项目负责人之一的 Michael Snyder表示，如果你有一个感兴趣的基因，你就可以到ENCODE里查查，了解一下基因组里都有哪些区域可能对你感兴趣的这个基因起到调控作用。比如一位乳腺癌研究人员通过ENCODE项目就可以发现，其他人发现的遗传变异可能刚好就是某个转录因子的调控靶点。而这个转录因子很有可能成为一个新的治疗靶点。

被28家科研机构采用的补充研究策略（Complementary approaches）也被纳入了这项基因组百科全书当中。很多科研人员都在从事RNA相关研究，但也有很多人还在专注于转录因子，或者转录因子与DNA相互作用这类的研究。当然，也有另外一部分人在从事数据分析和基因作图这方面的工作。

不过据美国马萨诸塞州博大研究院（Broad Institute in Cambridge, Massachusetts）的计算生物学家Manolis Kellis介绍，有些时候，ENCODE项目的庞大体系相反会延缓一些项目的进展。比如一个博士后的新点子首先得让一大帮子专家点评一番，然后才能实施；有些时候，某些科研人员们又必须等其他实验室的工作结束之后，才能发表他们自己的研究结果。

但是Kellis也承认，总体来说，ENCODE项目带来的好处还是远远超过了这些小瑕疵的。他表示，如果你一个人单打独斗，会有很多问题，而且可能需要花费你好几年的时间才能发现这些问题。但是在ENCODE项目里就不存在这样的问题，我们有这么多的同事，其他人很快就能替你指出问题所在。这种合作的工作模式也有利于科研工作的标准化，比如在称呼某个基因或者某个调控元件时，所有的科研人员都得遵循同样的命名法，这样大家彼

此之间才方便交流，而且也便于在数据库里检索，数据库的人机交互方式也更友好。

■ 肿瘤基因组

在在处理复杂程度更高的数据时，这种标准化系统的意义就更加突出了。国际肿瘤基因组协会（International Cancer Genome Consortium, ICGC）于2008年成立，他们在成立之初就在建立标准化系统。

该项目的最初目标是对2.5万名健康人和肿瘤患者进行基因组测序。最开始的计划是对外显子组进行测序。该组织的负责人，加拿大多伦多大学安大略肿瘤研究所（Ontario Institute for Cancer Research in Toronto, Canada）的首席科学家Tom Hudson现在向我们介绍，他们已经收集了2pb（大约200万gb）的数据，他们还将继续更加深入地收集更多的数据。

ICGC现在计划进军ENCODE项目更擅长的非编码区域，同时他们也会融入更多的临床信息。这种对全基因组进行泛肿瘤分析（Pan Cancer Analysis of Whole Genomes）的工作会涉及更多的人，现在的计划是纳入25万人（包括正常人与肿瘤患者）参与该研究项目。

研究规模的扩大，参与人数的增多并不会带来供应保障上的问题。到目前为止，ICGC已经汇集了来自全球16个国家，78个科研项目的负责人。据德国海德堡欧洲分子生物实验室（European Molecular Biology Laboratory in Heidelberg, Germany）的计算生物学家Jan Korbel介绍，在一个大型全基因组比对项目的前期实验工作中，科研人员们对2600名参与试验志愿者（包括正常人和肿瘤患者）的基因组进行了配对比较，该工作的数据量已经达到了0.7pb。这已经是很大的数据规模了，但这我们只是在科研机构自己的计算机中心里就完成了所有的工作，所以是完全有能力应付今后的工作的。

不过Korbel也表示，他们的项目现在到了一个十字路口。他们面临两种选择，一是需要很大一笔资金，添置运算设备，一是使用云计算服务。Korbel表示，你可以使用好几个云，每一个国家用一个，只要这些云彼此之间能够互通就可以了。即能够对不同云之间的数据进行比对。

■ 其他观点

经过这样的努力工作，各个研究小组各自得出的结果才有了一个统一的标准，彼此之间才可以互相进行比较，而且也可以用同一套标准的系统进行检索，这就让信息的价值得到了最大化的利用。这一点非常重要，尤其在发现罕见变异时显得更为突出，因为只有对数万或数十万的样品进行分析，才有可能发现这些变异，如果没有一套标准化的体系，单靠一个团队是很难完成这样的任务的。IBM下属Thomas J.Watson研究中心（IBM's Thomas J.Watson Research Center in Yorktown Heights, New York）的计算生物学家Gustavo Stolovitzky还认为，协同工作模式还有助于加深我们对数据的分析。

虽然大数据分析能够发现一些之前尚未了解的模式和关联信息，但同时也能够给科研人员们提出的各种预设提供证据，所以有时候可能会掩盖真相。

一个比较常见的错误就是过度解读。Stolovitzky将这种情况比喻为通过背诵一大堆生僻字来准备高考。你可以非常辛苦地记住所有这些生僻字，但是高考试卷上却完全有可能一个生僻字都不考。即便考了，也可能是用一种你完全想象不到的方式在考你。

同样的，基于自己手上掌握的数据开发预测软件的科研人员也往往倾向于用他们自己手头的数据对软件进行验证，但是在对其他数据进行分析时就常常不那么管用了。

另外一个问题是人体自然的本质。据Stolovitzky介绍，当他们分析自己的工作时，他们都是不那么严格的。所以让更多的人加入进来会更有帮助，他们可能会提出很多你自己

根本就想象不到的意见和建议。

Stolovitzky认为，如果你没有一套值得信赖的分析，那么所谓的大数据其实毫无价值可言。将几套好的软件分析结果综合起来看，肯定要比单独看其中的每一个结果更有收获。这也再一次证明了科研人员该如何利用生物医学大数据——协作才是关键。

“外行”也优秀

俗话说当局者迷，旁观者清，“局外人”往往能够解决难题。所以美国国立癌症研究院（US National Cancer Institute, NCI）在今年6月举办了一场名为“快来挑战（Up for a Challenge）”的竞赛活动，希望能够找到分析乳腺癌相关大数据的新方法。NCI从多家科研机构采集了数据，各家参赛队都会拿到一份数据，他们需要做的就是给出一个可行的解决方案。本次大赛的奖金高达3万美元，而且还有机会在*PLoS Genetics*杂志上发表一篇论文。

大赛的评委会根据各个参赛队如何运用创新思维，发现新的乳腺癌相关变异的表现来评分，同时还会考虑这些新发现是否可被重复验证，是否符合现有的肿瘤生物学相关基础。如果参赛队还能够与其他人或团体达成合作，还会额外加分。据NCI的遗传流行病学家Elizabeth Gillanders介绍，他们希望获得超出预期的结果，也希望有来自各个领域的人加入进来，大家一起解决这个问题。

这次大赛只是众多同类竞赛中的一个，比如还有由美国IBM的计算生物学家Gustavo Stolovitzky举办的，旨在改善系统生物学中算法开发工作的DREAM挑战赛等。还有比赛要求参赛者预测肌萎缩性侧索硬化症（amyotrophic lateral sclerosis）疾病的进展情况，以及各种药物联用方案的疗效。

很多时候，最棒的选手往往都不是拥有生物学背景的人。“可只要给他们一个新的数据，他们就能够崭露头角。”Stolovitzky介绍道。

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

四、重塑癌症临床治疗的希望
——疾病个性化治疗

目前，关于癌症的大数据之战仍然处于初级阶段，但“前线”已经在推进。

癌症基因组图谱（The Cancer Genome Atlas）对癌症的变异进行编录，该数据库包含了250万GB数据。这个庞大的项目由美国国立卫生研究院（US National Institutes of Health）运营，它极大地增进了人们对各类癌症的了解。然而，这个项目却对临床信息的了解——医生如何治疗提供样本的患者——相对较少。

那么若是求助癌症治疗链的另一端——电子健康记录呢？它含有大量能用于改善癌症治疗的病例详细信息。但是，这类记录往往分散在私人医院和各种医疗体系中。结果，正如纽约纪念斯隆-凯特琳癌症中心（Memorial Sloan Kettering Cancer Center）的一位专攻乳腺癌的肿瘤学家Clifford Hudis所言，人们并没有很好地研究大多数患者的治疗经验。

为了努力改进癌症治疗，Hudis及许多研究者正在开展合作。他们努力整合源于研究、患者治疗和临床试验的大数据，并试图发现其中的意义。虽然，利用大数据的良机遍及医学的大部分领域，只不过如马里兰州切维蔡斯郡（Chevy Chase）的一位健康医疗顾问Lynn Etheredge所言——“癌症走在前头”。但是，总体来说，癌症的多样性和致命性意味着大量壁垒与突破口并存。

即便如此，Etheredge还是在2007年撰写了一篇具有影响力的文章，并发表于《保健事务》（*Health Affairs*）。文章呼吁建立一个“快速学习系统”来处理大数据。对此，他解释说，现在已知癌症是一种基因性疾病，同时

人们又拥有数据库和数据分析必需的计算机能力，因而可以相信，目前已经进入了癌症研究和治疗的历史性时期。

另一方面，肿瘤学家和计算机专家希望利用个体化抗癌药物来打下成功的基础，于是共同合作，以获取数字化信息，并将之应用于临床。于是，一些风险项目不断出现，它们一边为生意而相互竞争，一边试图解决隐私、数据所有权和可持续商业模型的难题。对此，Etheredge表示，大数据既是研究工具，也是专有商品。但它在行业中尚处于早期阶段，还有许多需要大家研究解决的地方。

在美国，癌症治疗的某些方面在世界处于领先地位，不少团体和方案都在将大数据引入癌症临床。下面，本文列出四个“一”阐述要点：一家快速发展的初创企业、一个具有主动性的专业组织、一台拥有强大认知运算能力和能够辅助健康医疗的计算机，以及一个学术性癌症中心网络。

■ 初创企业

由马萨诸塞州剑桥伯德研究所（Broad Institute in Cambridge）的科学家于2009年着手筹备的Foundation Medicine公司为其分析服务购买保险，由来自学术机构和社团的肿瘤学家提交患者的组织样本，Foundation Medicine公司负责测序，然后将测序结果与公司拥有的不断扩大的分子图谱数据库（截至目前为止，该数据库包含50000多名癌症患者的数据）和其它公共数据库的数据进行对比，筛选驱动癌症发生的基因。

Foundation Medicine公司的首席执行官Michael Pellini表示，公共数据库与Google不相同，肿瘤学家无法像在Google简单搜索那样，轻易地在公共数据库中搜出与其病人的肿瘤相关的驱动基因。因此，他们必须分析患者的组织，并汇报有效的治疗干预方法，例如建议采用FDA批准的某款药物，或开展某个临床试验。

另外，肿瘤学家还能向Foundation Medicine公司的客户端网络咨询针对不同病例的建议。据Pellini介绍，72小时内，他们就能将这些建议汇总并反馈给医生。这样，肿瘤学家就能评估某种特定药品或者方法是否有效了。而公司的目标就是使其客户端数据的适用范围更广，以帮助医生做出临床治疗的决定。

2015年1月，瑞士药品巨头Roche公司斥资10亿美元购入Foundation Medicine公司56%的股份，并期望公司今年的收益能超过8500万美元。

■ 专业组织

2015年末，美国临床肿瘤学会（American Society of Clinical Oncology, ASCO）打算筹建CancerLinQ平台，其设计思路是从成千上万例肿瘤临床治疗实践中采集电子健康记录，并进行分析汇总，然后传递有益的临床治疗方法。

此后，肿瘤学家将能通过咨询CancerLinQ来获知特定干预疗法的效果，也能对自己的治疗方法进行回顾分析，看是否能与已有的医疗标准相媲美，并发展出各种假说以进一步开展研究。

Hudis任职于CancerLinQ董事会。他表示，大家对癌症治疗的了解大部分来自临床试验，但这种临床试验每年仅能招募3%的癌症患者。而若是使用CancerLinQ，却能设法了解其它未参加此类研究的97%的患者的情况。

起初只有15个不同规模的“先锋实践”小组加入了这个系统。ASCO希望，到2016年，该系统能囊括500000条患者记录；这样，研究者和临床医生就能通过咨询这些记录来比较患者的治疗结果了。同时采集汇总如此大量的数据将有助于揭示特定药物或方法的效果。

ASCO质量研究所的医学主任、肿瘤学家Robert Miller表示，CancerLinQ能做的最重要的事情就是通报结果。比如说，患者在接受某种特殊的治疗之后，存活时间延长，或者减缓了疾病的进展。这些观点都将有利于患者的治疗。并且，总有一天，美国头号癌症治疗资

助机构——联邦医疗保险（Medicare）会将服务费补偿的方式转变为奖励更好结果的选择性付费模式。

2013年，CancerLinQ初版已在一项囊括170000名乳腺癌患者的研究中进行了测试。据Miller称，已有未公开数据表明，该系统能突出源自不同医疗实践的数据趋势——比如，它们如何激发红细胞的产生，以此对抗化疗后的贫血。

该平台从电子健康记录中提取患者数据，隐去私人资料并归集数据，然后与其它各种信息整合（包括医生的笔记和生物标记贮藏库）。总之，它的目标就是最终为医生的决定增加治疗的支持论点，从而有助于诊断和治疗出现问题的患者。

目前，CancerLinQ依靠外界捐赠运行，但Miller表示，它迟早会售出有效的报告和数据搜索工具，从而使其更为自给自足。目前，他们正在关注一系列与CancerLinQ相关的产品和服务，以补偿系统的运营成本。

认知计算

大数据需要庞大的运算。2013年，IBM成立了一个独立的业务部门——IBM Watson Health，藉此为公司的Watson认知计算系统（可将自然语言和学习能力结合起来）关注癌症相关的商业机会。Watson的生物医学知识储存库包括：Pubmed数据库的每一条摘要（目前该数据库大约计有2500万条摘要，而且还在不断增长中）、美国国家癌症研究所药物辞典信息（其数据包括获批药物信息以及仍处于临床阶段的药物信息）、英国剑桥桑格研究院（Wellcome Trust Sanger Institute）COSMIC（Catalogue of Somatic Mutations in Cancer）数据库中的癌症相关体细胞突变的完整目录，以及许多其它来源的数据。

2011年，Watson在美国电视问答节目《危险边缘》（*Jeopardy!*）上击败了人类冠军，从而一举成名。同时，它还能访问匿名的病人数据。对此，纽约约克城（Yorktown Heights）IBM 研究院的计算生物学中心主任Ajay Royyuru告诉我们，IBM Watson Health已与超过一打的医疗体系、癌症中心和研究组织建立了联系。

而在一项关于胶质细胞瘤（通常为致命性的脑肿瘤）的研究中，纽约基因组中心（The New York Genome Center）依靠Watson来筛选纳入招募的患者的DNA变异情况。

同时，纪念斯隆-凯特琳中心和德克萨斯州休斯顿MD Anderson癌症中心的医生正在将Watson培养为临床的支持工具，这需要将匿名和假设的病例提交给计算机。Royyuru举例说，某位肿瘤患者由于缺失*STK11*基因，导致检测结果可能呈阳性，而该基因可能会对降糖药二甲双胍产生应答。但是，Watson很可能不会推荐二甲双胍，因为这是一个超说明书用药指示（off-label indication）。所以，这只是它可能学到的“广撒网”中的一个案例而已。（注：超说明书用药，又称“药品未注册用法”，是指药品使用的适应证、适应人群、给药方法或剂量不在药品监督管理部门批准的说明书之内的用法）。

纪念斯隆-凯特琳中心的乳腺癌专家Andrew Seidman补充说，对Watson的使用必须公开透明，这样做的目的是能够轻易地评判它下的论断。同时，他提醒说，Watson还没为“黄金时间”做好准备。Seidman之所以持审慎态度，是因为他正参与Watson的开发工作，了解Watson的情况。而且，Watson的自然语言处理能力还有待进一步提高。而今，临床医生仍无法与计算机直接对话，只能通过手工录入数据。

学术性癌症中心网络

癌症研究面临的一个主要挑战就是如何将作用于罕见变异的目标药物与患者匹配起来，这是因为想要在临床试验中招募足够多的此类患者并非易事。但一个医院小组却发现了解决这个问题的途径。

2014年，佛罗里达州坦帕的墨菲特癌症中

心（Moffitt Cancer Center）建立了肿瘤研究信息交换网络（ORIEN），这个网络包含九个学术性癌症中心。患者提供分析所需的临床数据和组织样本，最为重要的是同意接受终生随访，这样才能被招募进入针对其独有的遗传构成而开展的新型临床试验。对此，ORIEN的创建负责人Bill Dalton表示，这是一种更积极主动的研究方式。

2003年，墨菲特癌症中心开发了名为“肿瘤终生呵护项目”（total cancer care）的科学计划，并于2006年创建了M2Gen公司来处理数据分析和组织储存。ORIEN的发展促使该计划达到了国家级别，目前为止，大约有130000人成为受试者。会员中心则分享临床和分子数据，这样就能合作解决研究问题。

■ 高昂的标价

然而，从大数据中提取临床问题的解决思路并指导治疗，成本并不低廉。例如，Foundation Medicine公司就耗费了近6000美元对源于单个实体瘤的数据进行测序和解析。而对一份血癌样本的分析则耗费了7000多美元。

所幸，与新开发的抗肿瘤药物相比，这不过是小巫见大巫。后者的标价是每种疗法或者每年超过100000美元。今年7月，美国联邦医疗保险同意为Amge公司旗下一种治疗白血病的药物支付费用，如果没有保险，那么每位患者将需要为此药物支付178000美元。

当然，其它国家也可以更为积极地与制药公司讨价还价，从而降低价格，或者通过其它机构，譬如英国国家卫生与保健研究所（UK National Institute for Health and Care Excellence），以成本价为基准，抛弃那些贵价药物。

在理想的情况下，这笔巨款将会买下高价值的个体化治疗回报。这当然也是美国联邦医疗保险和公共医疗援助官员所希望的，他们在未来十年内将面临着把超过13万亿美元投入健康护理事业的问题，其中大头就在癌症治疗领域。因此这些机构将投入极大的力量，把大数据引入临床实践。当然，与大数据商业模型和成本相关的问题将会在整个医学领域都得到应用，但Etheredge表示，正是癌症促使它们上了台面。

五、定制智能体检
——健康研究步入新时代

可以穿戴在身上的传感器和智能手机在为人们提供海量信息的同时，使得全民范围的研究成为可能。

几十年来，全世界的医生都在用一种简单的测试来检测病人的心血管健康情况。他们让患者在坚实平坦的地面上行走，看他们在六分钟内步行的路程。该测试可用于预测接受过肺移植的受试者的存活几率，以检测肌营养不良的发展状况以及评估心血管健康的整体状况。

因此，步行测试在多项试验中都得到应用。不过，此前即便是最大规模的研究，受试者人数也极少超过1000名。而在2015年3月由Euan Ashley开展的一项心血管研究中，竟然在起始的头两周内就获得了6000人的受试结果。用他自己的话说，那简直就是一个卓越的数字。Ashley是何许人也？他是一名名遗传学家，执掌斯坦福大学（Stanford University）遗传性心血管疾病中心。他毫不讳言地告诉我们，在过去，他们只能研究几百名患者，而且是在幸运的情况下。

他希望，通过目前这种大范围的数字，能获知更多生理活动与心脏健康之间关系的相关信息。而他们之所以能够完成这件事，原因就在于当今有上百万的人拥有智能手机与健康追踪器，从而能够记录各种生理活动。目前，健康研究者正在研究这类装置，以查看他们在分析几十万人日复一日的各种活动测量数据时，能够采集的数据种类，并确定数据的可靠性以及可能获知的信息，从而将庞大的数据算法应用到各种解读中。

7月，美国40000多人签署知情同意书参加Ashley的研究，期间使用一个名为MyHeart Counts的iPhone手机软件。Ashley希望使用者蜂拥而至，能让这个手机软件在全世界更为广泛地使用。这项由科学家设计的研究经伦理审查委员会（institutional review boards）批准，并要求知情同意——让患者回答关于自身

健康与风险因素的问题，同时使用电话带有的移动传感器收集7天内的活动数据。当然，受试者也要完成一个6分钟的步行测试，由电话来测量其步行距离。若他们的医生提出验血，使用者也能输入诸如胆固醇或血糖监测等信息。每隔3个月，手机软件都会与受试者进行联系，更新数据。

Ashley表示，医生明白：人的生理活动是其长期心脏健康的强烈预测因子。但到底哪一种是最能反映结果的活动，或者不同组别的人是否在不同类型的活动中有更好的反应，目前还不甚清楚。而MyHeart Counts可能可以打开解决这类问题的窗口。Ashley表示，他们能够开始通过观测亚组人群来寻找其中的差异。

大容量的数据可能可以促使上述类别的研究得以进行。在传统的研究中，缺乏足够的数据来获取具有统计意义的结果以进行这种亚组研究。同时，罕见事件可能不会在较小的样本量中出现，或者产生的信号太弱，以致在统计噪音中无法显现。而庞大的数据则能解决上述问题。并且，如果数据设置足够大，那么小的错误也能摆平。对此，Ashley认为，虽然可能会采集到干扰非常多的数据，但如果数据量足够，那么总能找到说明问题的信号。

每天一“苹果”

正是因为有一个叫做ResearchKit的苹果软件，使得采集如此大量的数据成为可能。该软件能用于基于iPhone开发的手机软件，以进行上述类别的研究。MyHeart Counts就是ResearchKit发布那天同时上线使用的5个手机软件之一，其余的则试图管控帕森氏综合征、乳腺癌、糖尿病以及哮喘研究的庞大数据。

今年6月，帕金森氏综合征的研究招募了约16000名受试者，均使用了步行测试。这是因为帕金森患者的临床表现中有运动失调。受试者沿一条直线行走20步，他们手机上的加速表和陀螺仪则会测量患者的步态，以评估其运动控制能力。此外，测试还要求患者对着电话说“啊”约10秒钟，以测试他们的声音颤抖度，从而协助医生了解其肌紧张情况。对此，Sage Bionetworks（一家非盈利生物医学研究咨询公司，总部位于华盛顿州西雅图）的一位公开数据支持者John Wilbanks表示，手机装置非常适合感官测试。该公司与纽约罗切斯特大学（University of Rochester）的博士们共同开发了一款适用于帕金森综合征的手机软件，该软件能使用调查问卷，并与健康追踪系统相连，从而可以采集更多的数据。

当然，人们还写出了较小的手机软件，用于其它智能手机操作系统（如Windows和安卓）以及其它相关的智能手表。此外，许多公司（包括Basis、Fitbit和Jawbone）还开发了衍生产品——各种可穿戴的健康装置。此外，研究人员正在开发其它可穿戴的传感器，以用于长时间采集数据，其中包括临时纹身以及可检测眼泪中的葡萄糖水平的隐形眼镜。同时，现有的装置，比如可持续监测糖尿病患者血糖的仪器，也正在迅猛发展中，并可将其数据录进智能手机的混合数据。

当前，研究者正尝试使用智能手机来监测更高级的生理健康状况。比如说，一些装置能够通过监听某人的声音来确认其受到的压力，或者通过追踪动作来确认受试者的社交状况，以弄清他们是否情绪低落，从而追踪一个人的精神状态和情感健康。

总之，随着便携式装置愈来愈多地用于测量人类的全部活动，同时计算机也具备了足够的能力对庞大的数据进行筛选，研究人员希望能够获得史无前例的关于人类健康的真知灼见。

关于测量

此前，相关检测的系列装置类型一直在不断增长，这为人们提供了大量检测方法，却也使得研究人员不得不考虑怎么解决由此带来的所有问题。对此，美国旧金山加利福尼亚大学（University of California）生物医学信息

部的负责人之一Ida Sim有独特的看法，她觉得，那不过是个很刺激的乱子。

Sim说这话是有底气的，因为她同时还是Open mHealth的合作创始人。这是一家非盈利公司，专门开发软件，能将采集自各种装置和手机软件的数据进行标准校验、储存和处理，以帮助归整乱七八糟的数据。对此，Sim自豪地表示，虽然每个人都讨厌标准化，但如果没有它，就很难将数据精确地归整到一块儿。比如说，一名医生想要正确地解读一份血糖数据，那么要弄清患者在一段时间内是否有在控制饮食就很重要了。

然而，想要建立标准，就必须解决两个关键的问题。一是源自各种装置的读数的精确度有多高，二是测量的准确度如何。当今设计的健康追踪器一般都能告知用户简单的结果，比如：本周步行的距离是否比上周多，但无法采集达到实验室水平的测量数据。因此，在马萨诸塞州波士顿东北大学（Northeastern University）研究个人健康信息的Stephen Intille表示，从这些装置获知的只是一般的运动结果，而将之转化为步数，这就稍微要高端些了。

美国东北大学（Northeastern University）的研究人员在实验室通过规定范围的活动校验传感器。

为了更好地感知到底哪一种装置具有精确的测量度，Intille把志愿者们带入他的实验室，然后将各种传感器缚在他们的双臂和双腿上——这不仅包括商业性装置，还有经实验室校验过的、可记录运动、心率、呼吸及其它数据指标的传感器。大约2-3小时之后，他就会提取志愿者步行、做杂务、骑自行车以及完成类似活动的读数。然后移去部分传感器，送受试者回家。接下来，他持续两日用保留在受试者身上的装置来采集真实世界中获取的数据。在随后的三个月内，他再将装置的数量缩减到自己所研究的一到两种。

就是用这种办法，Intille精准地观察到了商业性装置在人们进行特定活动中的记录情况。比如说，当用户在熨衣服时，一个Fitbit监测仪可能只会产生一定设置的读数，而实验室装置却能记录心率和呼吸。如果计算机能够在实验室条件下识别出不同的活动产生的不同Fitbit读数，那么或许也能够识别出真实世界中产生的活动，并分析出它们对生理健康的影响。

Intille表示，他个人并不相信这些装置能在与终端用户不进行某些交互的情况下仍能够运行得很好。因此，他试图运用电话来利用健康追踪器上获知的数据，并且了解个人的某些习惯，以及询问诸如“现在是否在遛狗？”等

问题。他还表示，在更完整的蓝图中，人们可能还需要穿戴不止一个装置，也许需要一个在手腕上，一个在脚踝上。

此外，斯坦福长寿研究中心（Stanford Center on Longevity）的荣誉退休医学教授、生理学家William Haskell表示，对于研究人员而言，需要这种细节信息来获知更多的了解。比如说，想要知道人们跑多远，以及跑步频率如何，或者他们在体育场上的锻炼强度有多大，这些都不是难事，但日复一日的活动对人的健康效果如何，却很少被了解。

Haskell一直与Intille合作，从事商业性追踪器的测量度验证批准工作，他认为，对于低强度范围的运动（站立或只是稍微走走），人们其实并不是很了解。比如说，有一个站台，人每天在那儿坚持站立3小时，这与绕着办公室来一场不错的步行，到底哪个更有用一些？大家都不知道。

早在40年前，Haskell就已开始用加速表来追踪生理活动。对于可能从可穿戴装置获取相关信息，他感到很激动，并且表示，科技就在这儿，他们只需对其进行验证批准，就能用这东西来观察人24小时的活动周期了。

■ 下一站穿戴？

获取庞大的数据能够加强健康研究的力量，而可穿戴技术还能开启另一种可能：采集其它此前无法获得的数据，如长期、昼夜监控处理公务的人们。

Sim表示，目前对于庞大数据的期望是这样的：不但能够观测到大量数据，而且还能观测到多种不同来源的大量数据。

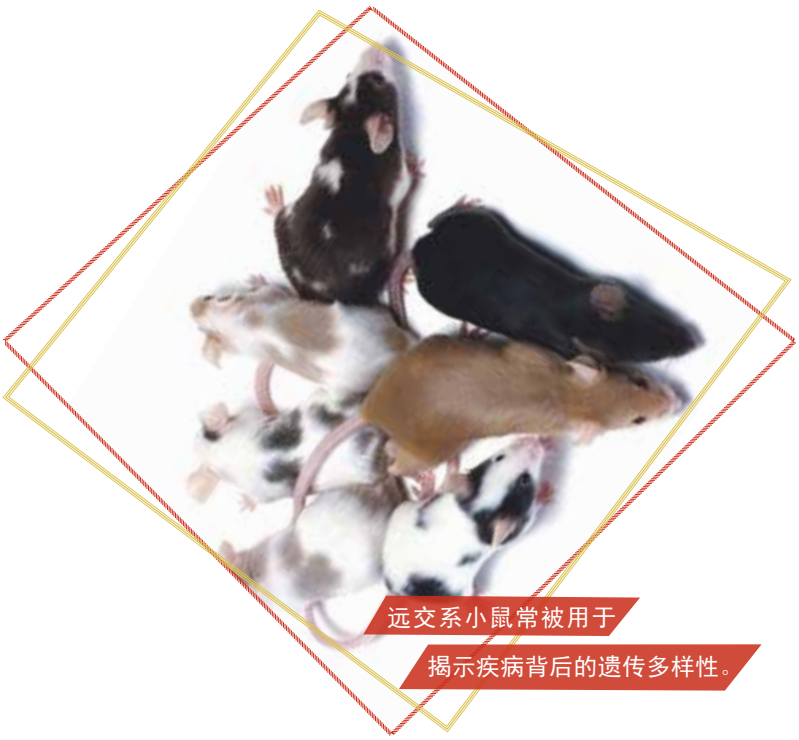
当Sim为患者诊治时，会与他们交流大约20分钟。她表示，当他们不在自己诊室时，她完全是一抹黑，根本不知道他们的肝脏在搞什么。而不间断的数据采集即能最终改变这种状况，协助医生为他们每一位患者进行量体裁衣式的照料。不过，现在他们还有一个关键的环节缺失：目前为止，尚未有一种方法能将商业装置上的有用数据传递给医生。所以，适合医生的工作流程还没有完全建好。

不过，这种无处不在的健康信息采集也可能会带来范围更广的社会利益。成千上万人的健康信息数据通过日渐普及的技术进行自愿采集，使关于影响健康的因子的研究得以在全民范围内开展。Ashley设想用一种移动-健康版本的装置开展数十年之久的Framingham心脏研究，该研究能协助识别心脏病的风险因素。他还开始将采集自iPhone的数据与遗传数据相关联（遗传数据采集自斯坦福医学中心患者用户）。

Intille相信，随着更大的数据装置被开发出来，健康研究者将能够回答范围更全面的问题。他表示，在个体水平，此前人们尚未获得过任何类似数据，因为在移动装置出现前，这样的检测根本是不可能的。因此，这与过去解决健康和医疗问题的方式完全不同。

想知道更多关于移动技术促进健康研究的信息，请点击链接：http://go.nature.com/lyvvkp

六、疾病细节
——通过深度表型分析形成精准医学



远交系小鼠常被用于揭示疾病背后的遗传多样性。

精准医学计划若想获得成功，就必须将基因组学深度分析与表型模型精确匹配起来。两者的联系越紧密，我们就越能获得更多的信息。

20年前，人们对仅进行人类基因组测序工作就可以进入精准医学时代这个想法非常乐观，但是Isaac Kohane则持谨慎的态度。没错，基因测序的确能让我们往前迈出非常重要的一步。但Kohane指出，如果真想从洪流般的基因组信息中捞取有临床价值的信息，那么靠的便是更为枯燥的表型研究了。表型是指能够表示健康或疾病，例如发热、皮疹、四肢无力或不规则心跳等的性状特征。

如果想要测量糖尿病的不同表型，我们就需要安排某个人专门梳理涵盖体重、血压和血糖水平等度量信息的医学记录。这可是一项单调、乏味，并且昂贵的工作。而且，新的测量方式，例如可提供关于疾病的有价值信息的动态血糖监测等可能并不包含在这些医学记录里面。

精准医学需要我们了解基因和表型之间的精确关联，而且想要将疾病细分成各种亚型则需要了解疾病背后的生物机制。但研究人员并未真正了解所有基因的功能，并且他们知道的也仅限于几种细胞类型、组织或生理状态。另外，对这些疾病表型的描述通常无法准确地抓住普通疾病的不同表现形式，或无法定义那些能够预测治疗反应的疾病亚型。据2012年一

篇评论指出，表型描述通常都是“马虎或不精确的”。

若想解决这些难题，就必须对表型性状的不同成分进行详尽的分析，而不是单单了解常规的病历表中的记录信息。众所周知，这种“深度表型分析”方法可以更为个性化和更为精准地获取疾病表现的细节，从而借助精妙的算法将获得的大量数据与其它种类的信息联系起来。

在这之前，表型分析并没能产生大数据。它只能产生一小部分的、无序的数据，并且要获取这些数据并不容易，是一件非常费时的事情。个体表型的信息一直没能与个体间遗传突变情况很好地匹配起来。幸运的是，深度表型分析可以为我们提供更为特异性的、新型的大数据，并且可能能够将疾病亚型和遗传突变情况相互关联起来。

这种方法将能帮助研究人员解决新问题。疾病细胞中的蛋白表达或基因调控的特异性模式是什么呢？细胞的代谢产物和其它生物化学特性又是什么呢？是否存在异常的肠道细菌？病人是否有其它看似无关的症状，例如自身免疫反应或精神障碍，而这些症状又是否疾病本身拥有相同的生物学通路？这种全面的深度表型分析，同时联合其它大数据（例如基因组数据）将可以精准地揭示每个个体疾病背后的机制。正如Kohane所说的，深度表型分析“可以为我们展示疾病的不同方面”。

■ 糖尿病分型

糖尿病可是充分体现了表型含糊不清的问题。美国缅印州巴港的杰克逊实验室（Jackson Laboratory in Bar Harbor）的小鼠遗传学家Gary Churchill指出，引发糖尿病的疾病通路有100种，涵盖了发生在胰腺、肝脏、肌肉、脑部和脂肪中的通路。如果不加以区分疾病的不同机制，而将它们看成是一个大类，那么遗传学研究就会失去统计检验的效能了。不同的基因会导致不同疾病亚型的发生，如果将它们混合在一起，就会掩盖携带同一遗传突变体的人，对同一治疗响应却存在差异的原因了。

与Churchill合作的威斯康星大学麦迪逊分校（University of Wisconsin-Madison）的生化学家Alan Attie指出，在体重和血糖水平上，致病基因引发疾病表型的过程分为多个步骤。每一个步骤都会受到遗传突变的影响，从而削弱基因和表型之间的联系。

Attie正在寻找个体基因组的差异是如何影响糖尿病的某个特定表型的。Attie研究的这个特定表型就是胰岛细胞的胰岛素分泌过程。他从遗传多样性小鼠身体中分离出胰岛细胞，然后测试它们对葡萄糖、脂肪酸、氨基酸和其它影响胰岛素分泌的分子的反应。初步数据显示，胰岛细胞对上述物质的反应差异非常大。

Churchill指出，之所以没有研究基因型相同的近交系小鼠，而是选择研究远交系小鼠，是因为后者能够更好地反映人类疾病，例如拥有多个致病因素的糖尿病的多样性。比如，研究人员经常使用的一种近交系小鼠——B6小鼠，当因为同一个原因而肥胖时，它们都会患上糖尿病。如果我们只研究这种近交系小鼠，结果只能反映某些病人的情况，而无法了解到其它的致病原因。

■ 头脑风暴

将深度表型分析与大“组学”数据联系起来并不是一件简单的事情。而且在某些神经精神疾病，例如自闭症方面，基因和表型之间的联系还有待进一步证实。

美国马萨诸塞州博德研究院（Broad Institute in Cambridge, Massachusetts）的神经科学家Steven Hyman指出，精准医学？那可与我们无关。我们几乎都不知道如何制药呢。在精神病学范畴，我们只有说明性的表型信息，而不是那些能够解释脑部障碍的机制性表型信息。对神经精神疾病进行深度表型分析可能有助于打破目前寻找更好的疗法的僵局。

大部分的脑部疾病都是多基因遗传性的，多种基因突变体联合作用导致了疾病的发生，

所以即便我们鉴定出致病基因，也无法解释大多数的疾病机制。拿自闭症来说，只有不到10%的病例与那些可以解释疾病机制的基因相关。麻省理工学院（剑桥）（Massa-chusetts Institute of Technology in Cambridge）的神经科学家Guoping Feng指出，一个自闭症基因可能也与精神分裂症、强迫性精神障碍以及双相情感障碍有关。每一种疾病都有一些特定的症状，但同时也有会一些与其它疾病重叠的症状。

另外，尽管大部分有自闭症的患者都拥有相同的核心症状（例如重复性的行为和社交障碍），但是某些患者另外还有肠易激综合征、感染、癫痫发作、精神分裂症或注意缺陷多动障碍等。Kohane表示，他们不但要考虑神经病学和行为学症状，还要考虑病人的其它症状，例如炎症和心脏功能障碍。仔细评估这些疾病亚类是精准医学的必要前提。

英国哈维尔医学研究委员会中心（Medical Research Council center）的哺乳动物遗传学家Steve Brown希望他这项与国际小鼠表型分析联盟（International Mouse Phenotyping Consortium, IMPC）共同完成的项目能够揭开这种并发症的神秘面纱。这个联盟正在对基因敲除小鼠基因组中的每一个基因进行系统的表型分析。

Brown表示，他们不能只查看一个或两个表型，因为他们根本不知道大部分基因的功能。“我们没办法假设该关注什么”。研究人员测试了每只老鼠的感官知觉、心血管、肺部功能、代谢情况、形态学和病理学，并且记录它们的环境条件和饮食。他们还记录了小鼠的活动、社会交往、梳洗、睡眠和喂养的行为学数据。

这个联盟里的基因敲除小鼠都是近交系。虽然这限制了研究人员对自然多样性的了解，但是却让他们得以开展相关的比较研究以及获得重复性结果。Brown指出，他们从来没想过要构建一个自闭症或精神分裂症模型。相反，他们的目标是为每个基因能做什么以及这个基因如何影响行为建立一个基准线。

■ 模型的局限性

然而，那些正在对动物模型进行深度表型分析的研究人员也承认在非人类物种的建模过程中存在局限性。对此，Hyman指出，人类的神经精神疾病涉及前额叶皮质区，而这个区域是最近才进化形成的。小鼠大脑中压根就没有人类大脑中的许多重要的细胞和神经通路。Hyman强调，科研人员应该关注在小鼠和人类体内共有的细胞和分子通路。很多研究人员都把转基因小鼠当作研究对象，试图用它分析精神分裂症或自闭症，并从中揭示疾病的分子机理。反过来进行研究倒有可能行得通。

德国神经退行性疾病中心（German Center for Neurodegenerative Diseases）的朊病毒病研究者Walker Jackson研究的是人类朊蛋白的一个氨基酸突变如何导致Creutzfeldt-Jakob病变的发生。同时他还以小鼠为研究对象，研究致死性家族性失眠症。Jackson通过评估小鼠的行为，逐渐了解了疾病的自然发生史，不过他并没有寻找与疾病相关的基因。Jackson表示，他没有试图发现基因突变如何影响行为，因为实在难以弄清究竟是什么因素改变了行为。

Jackson发现相同的突变总是固定地影响某些神经元。他想知道正常的神经元如何弥补受突变影响的神经元，以揭示治疗靶标。这些病变发生在海马区、小脑和丘脑处，这三处区域都与疾病的异常行为相关。Jackson指出，结果显示，疾病比我们想象的要复杂得多。受影响的神经元间的功能障碍各不相同，因此，适用于一种神经元的疗法却并不适用于另一种神经元。

另一边厢，斯坦福大学医学院（Stanford University Medical School）的研究人员正在从事对*NL3*基因的一个突变的研究。该突变可直接导致某些人类自闭症——一种罕见的精神疾病。他们把这个突变体导入小鼠体内，并追踪这个突变体对运动行为的影响，以及这种影

响对在一个意想不到的脑部区域里的某个神经元的受损多巴胺的抑制的作用。

Feng采用一种相似的方法在小鼠体内识别由另一个自闭症基因（*Shank3*）导致的神经回路失衡。但是这种方法无法推广，广泛使用，因为大部分的疾病都是由无数基因联合导致的，每一个基因仅仅起到一个细微的作用。Feng不认为仅仅对突变小鼠的行为进行深度表型分析就能让我们深入了解疾病机制。但研究那些来源于人类的细胞，就有可能达到这个目标。因为这些细胞已经携带了非常完美的、足以诱发人类疾病的基因组合。

■ 人类实验

因为动物研究有局限性，并且直接在人体细胞中研究疾病具有优势，所以现在人们将深度表型分析应用于复杂疾病的新型人类细胞模型的研究工作。例如，神经研究可以将皮肤细胞诱导成为干细胞，并且还可以让皮肤细胞分化成神经元或自组装细胞簇，即类器官，这样研究人员便可以研究表型之间的联系、基因组信息以及相关的生物学数据了。

Kohane正在领导这样一个研究项目，该项目名为N-GRID，它从神经精神疾病病人身上收集细胞，以研究个体基因组与转录组、蛋白质组、DNA甲基化模式和其它影响基因表达、对小分子的响应和临床特征的表观遗传标记之间的联系。Kohane表示，这个项目的深度表型分析方法包括“我们能够测量到什么东西，就看看这个东西的独特的疾病分类是否存在”。该项目的目标就是建立一个“更强大的精神疾病分类方案，也即疾病预后更可靠、对每个类别的生物畸变的了解更透彻，从而更有效地治疗患者。”

Hyman提议研究人员应该考虑保留动物模型，从而进行安全性和药代动力学研究。一种新疗法的效用可以通过基因工程人类细胞培养物或类器官来测试。Hyman问道，如果我们没法建立一个精神分裂症小鼠模型，那怎么办？这可不能阻挡我们对安全、有效的疗法的追求，万一动物模型无法提供良好的疗法的数据，那么对人类细胞进行深度表型分析可能可以弥补这个空白。

七、维系大数据生态系统
——构建更好的信息获取模型

Philip E. Bourne、Jon R. Lorsch和Eric D. Green指出，组织和访问生物医学大数据的工作需要完全不同的商业模式的参与。

虽然生物医学大数据为生物发现提供了无限可能，但是却很少有人关心维系这些数字资产所需的花费是多少，更不用说会有人关注让它们成为有价值的物品背后所涉及的资源了。许多国家，包括美国在通货膨胀调整后，研究预算都没有增长，甚至还下降了。但是，即便在这种情况下，数据量还是在以前所未有的速度增长，因此，研究团体必须找到一款更高效的模型，以便更好地存储、组织和访问生物医药数据。须知道，简单地在现有系统的基础上投入越来越多的钱，并不是长远有效的办法。

为了让大家更好地了解目前的数据存储状况，Philip等人认真查看了美国国立卫生研究院目前和预计在生物医药数据运营方面投入的费用。他们的初步分析表明，即便忽略国家生物技术信息中心（National Center for Biotechnology Information, NCBI）这个特殊的情况，NIH资助的50家最大的数据资源中心的年度预算共为1.1亿美元。而且，这个数字仅仅是未来所需费用的冰山一角。

■ 理解现有数据的使用情况

目前的生物医学数据资源毫无区别地对待它们收集的每一条记录。但是这种处理方法并不总是有意义，因为数据的使用模式各不相同。但是我们如何判断哪些数据需要获得更多关注呢？目前大数据的产生越来越容易，维持和注释这些数据的费用也越来越高，如何对待每一条数据记录的问题将变得日益重要。

要回答这个问题，就需要我们更好地理解这些研究数据的使用情况。在这之前，很少有人认真地探究过这些问题。以往，投资人感兴趣的主要是他们所资助的数据资源是如何使用的，以及被谁使用。他们往往不会仔细查看每一条记录或数据类型是如何和为什么被采用。

仔细分析这些数据可以给研究人员带来启发。初步研究表明，通常情况下，都是某一小部分数据子集被频繁使用，其余大部分数据都是很少被访问到的。然而，被大量采用的这一部分数据子集可能会随时间发生变化，而大部分的数据访问工作都是在数据被下载后开展的，所以这些变化并没有被记录下来。所有这些都意味着想要诠释绝对数（absolute numbers）（一定时空条件下，数据增减变化）并不是一件容易的事。

我们需要将上述存在的问题，以及关于数据使用的更多细节知会投资人，以便他们制定投资对策。逐渐地，这种数据使用模式将会告诉我们如何最好地开展注释和管理数据的工作：确认哪些数据最值得关注，从而投入最多人力物力；同时决定哪些数据可以长期持有。数据更新的成本也可能会影响保存数据的决定。

资助者应该支持开发一些新的度量标准，用以确定数据用法和评定数据价值，并说服数据资源中心应用这些新标准来度量他们维护的所有数据。我们也可以从私营企业中吸取经验：他们通过数据分析工作来理解数据使用模式的相关细节，从而将公司经营得非常成功。这些公司包括Amazon和Netflix。

■ 公平和效率

当我们能够更好地理解数据使用情况时，我们就可以开发出供需平衡的商业模式，并使其可以持续发展。另外，确定经济规模和驾驭市场力量也是必不可少的环节。

对于典型的生物数据资源来说，单纯维护数据的费用仅仅占管理数据的总费用的一小部分。其余大部分的费用都花在了数据发现、访问、互操作和重新使用（finding, accessing, interoperating and reusing, the FAIR principle; go.nature.com/ axkjiv）上面。这部分的费用却是经常被低估的。

FAIR部分的花费合理吗？来自不同数据资源的服务是否冗余？数据资源是否受到功能蔓延（feature creep）——一种附加的、昂贵的、作用有限的花哨功能的影响？我们的资助机制是否导致了这些问题的出现？并且，最重要的是，我们目前维持生物医学数据的方式对现在和未来需要进行的科学研究工作来说，是最佳的选择吗？

目前科研人员主要采用不同来源的数据开展相关研究。而这些数据以不同的访问模式被存放在不同的数据库中。正是这种缺乏集中性和通用性的数据管理模式会阻碍数据的使用和生产。我们需要很好地理解多种数据资源的运用情况，然后以此为基础改进现有模式，做好有价值数据的保存、管理、发现、访问、整合和再使用等工作。

另一方面，生物医药数据的综合管理和质量保证情况也必须有所改变。从文献中完整地和精确地自动或半自动提取数据时，必须提供相关的元数据和注释信息。我们应该考虑有合适的验证模式和激励机制的众包管理。此外，专业注释人员的作用必须得到数据使用者、他所服务机构以及资助者的认可。

长远来看，我们需要一个与研究寿命周期一致的数据管理模式。有一个不必要的成本就是研究人员分析数据后，将这些分析信息放入研究论文中，这样做的目的仅仅是为了让生物

注释人员从论文中提取那些信息，然后再将它们与原始数据重新联系起来。我们确实需要一些工具和奖励政策，从而激励研究人员将他们的数据放入数据库，这样才能最大限度地提高数据的质量和降低访问数据的难度。

■ 商业模式

一个非常值得探究的商业模式是“免费增值模式”（freemium model）。也就是说，使用基本数据是免费的，但是增殖的附加服务则是收费的，以此获得维护基本数据的费用。其它学科，尤其是化学，也采用这种模式。但这种模式有两个棘手的问题。营利机构是否应该与非盈利机构的收费一致？谁该拥有增殖内容的知识产权？

另一种有潜力的商业模式是“订阅模式”。也就是说，用户通过订阅，即可访问例如由拟南芥信息资源网站（The Arabidopsis Information Resource, TAIR）提供的基因和分子数据库。这种模式通过活跃用户来为数据资源提供支持，但是它的访问有一定的限制，这可能会对数据公共访问策略带来负面影响。

如果商业模式理念更前卫一点，例如数据资源网站被合并、收购或倒闭，那么将会发生什么呢？如果现有的资源网站在某种程度上被合并，那么是否会变得更加实用，并且性价比更高？某些缺乏需求的服务是否应该被取消，从而为新服务腾出资源和空间？逐渐减少对某些数据资源网站的投资，是否有助于提高运营效率？我们可以从私营企业和其他科学团体那里获得建议和帮助，从而解答这些问题。

■ 求同存异

云计算创建了数据虚拟化元素，它将计算引入数据处理领域，并且可能还有助于解决生物医学大数据所面临的一些问题。Philip等人在NIH内部提议构建一个可持续发展的模式“commons”，以便抓住机遇，好好地借助云计算处理数据。

从物理角度来说，commons就是用于存储数据和计算数据的公共和私人资源（包括云资源）的集合。为了让commons具有更好的依从性，这些资源必须遵从两条简单的原则。首先，commons的每一个研究对象，例如数据、软件、叙述或论文都必须有唯一的标识符、可共享（兼顾隐私问题），以及可以通过通用的标识符来辨析其来源；其次，每一个研究对象都必须按这种模式界定的最小的元数据量来定义。

NIH的大数据倡议——大数据到知识（the Big Data to Knowledge initiative, BD2K）项目（bd2k. Nih.gov）的目标是实现commons的构想。Commons鼓励12个新型BD2K中心在commons内部分享研究热点。其中一个BD2K团体正在为commons的内容建立索引，以方便用户查找。

Philip等人也在研究计算积分的概念。他们试图给予接受赠予者积分而不是资金，以此来支付计算时间所需的费用。首席调查员可以通过积分来购买任何符合commons规定的资源。另外，从对小部分数据进行大量计算的研究人员，到对大量数据进行小规模计算的研究人员都可以用他们的积分来购买不同的符合commons规定的资源。

这种模式与现有模式之间存在显著差异。前者将科学技术获奖人员和他们所在机构开展的硬件、数据和软件的维护工作的初始压力转介给第三方，尤其是云服务供应方。这种资助模式只为使用的服务埋单，这样做的目的是创造市场竞争，所以这种方式可以让投入的每一美元都获得更多的数据回报。

如果在NIH开展的这种模式的初步试验获得成功，那么接下来非常重要的事情就是考虑这种commons模式的长期效益。其中一个效益可能是专利审查期间和失效后，所涉及的数据和软件使用情况都会被跟踪记录下来。这种记录方式可以让我们获知重要的统计信息使用情况，从而指导后续的资助决定。

研究机构，例如博德研究所（Broad Institute）正快速改进他们储存和访问生物医药大数据的技术。

■ 团结资助者

医学研究团体手头的资金太少了，以致于无法启动新的数据资源项目或构建更为成熟的数据库和服务体系。而且，目前的资助计划并没有包括寻找最佳的工作方式，例如，每个数据资源通常都是被孤立地审查的。

我们需要同时在相关机构和国际间改变现有的资助模式。数据的产生和维护通常都是由国家资助的，但这些数据都是国际间共用的。因此，我们需要开发一个更合理的资助模式。资助机构第一步要做的事情就是更高效地开展数据科学问题的沟通工作，并寻找协同解决方案。科学家们已经为此努力了很长一段时间了。

维护生物医学大数据生态系统是所有利益相关者的责任，并且需要数据生产者、数据维护人员、数据使用人员、资助人员、出版人员和私营企业人员的共同努力。NIH BD2K项目与所有相关人员合作，正着手解决这些问题。

【小词典】功能蔓延，有时也被称为需求蔓延（requirements creep）或范围蔓延（scope creep），它是指在发展过程中产品或设计的需求增加大大超过他们原来预期的趋势，导致其功能不是原定计划的，并且要承担产品质量或生产进度的风险。功能蔓延可能是由于客户期望功能的增加或由于开发者自身发现了改善产品的机会而导致的。为了控制功能蔓延，设计管理工具有时被提倡使用，比如需求稳定性索引（RSI）。（http://blog.csdn.net/itcareerist/article/details/6008077）

八、人物访谈

1. Mark Caulfield访谈：英国获取大数据的方法

Mark Caulfield是Genomics England公司的首席科学家。Genomics England公司成立于2013年，旨在运行英国10万人基因组项目。该项目拟通过基因组测序来确定癌症、罕见疾病和感染相关的基因。作为一名心血管临床医生和研究人员，Caulfield在接受采访的过程中谈到了英国在获取生物医药大数据的工作的现状以及Genomics England公司在当中发挥的作用，包括该公司是如何很好地将基因组医学应用于英国国家医疗服务体系（Britain's National Health Service, NHS）的。

■ 将基因组医学与临床相结合的主要挑战是什么？

第一个挑战就是要建立一个有能力运行这个项目的平台。为了达到这个目的，我们在全英格兰构建了11个基因组医学中心。这些医学中心全都配备了由临床医生、科学家和学术人士组成的专业小组，他们可以帮助我们招募合适的病人、获得病人的知情同意以及获取临床数据和样品，从而开展研究分析工作。

另一个非常重要的问题就是如何大幅度提升解读那些基因组信息的准确性。我们与英国的创新机构——Innovate UK公司合作，将1000万英镑（1550万美金）的政府资金用于激励相关公司提升数据分析的质量。2014年12月，我们创建了 Genomics England Clinical Interpretation Partnership（GeCIP）项目，该项目旨在将NHS和学术界的研究人

员、临床医生以及实习人员聚集到一起，以提高基因组数据的分析水平。GeCIP项目主要涉及多个具体的研究课题，如血液肿瘤学，目前这个课题已经将英国所有从事白血病和淋巴瘤工作的人员全都聚集起来了。

■ 你们如何将解读的基因组数据更好地应用于健康保险体系？

如果我们手头上有立即可用的研究结果，例如病人基因组中一个致病突变体的信息，那么我们就会直接向相应的NHS基因组医学中心发一份临床报告。医学中心的临床医生随后便会仔细查看这些数据，并在向病人反馈结果前，根据他们的验证步骤验证这些数据，以判断它们的准确性。

类似的判断工作通常都是由NHS完成的。NHS的专业的基因组分析人员需要认真评核这些数据、熟悉这些数据、判断数据的准确性，因为他们将每天与那些病人待在一起，对这些病人开展临床护理工作。将自主权交给NHS，可以让我们在英国基因组项目结束后建立一个长期可用的资源库。

如果我们找不到任何与致病相关的信息，那么这些基因组将会被放入与病人疾病相关的GeCIP数据库中。这样可以帮助我们提升与疾病相关的基因组数据的解读准确性。同时，我们还会发报告给病人，以促进他们对项目进展的了解。

■ 基因组学产业的作用是什么？

一个充满活力的基因组产业对病人和我们整个社会来说都是好事，当然，对整个国家的总体财富来说也是百利而无一害的。

我们已经建立了一个由10个公司构成的联盟。这些公司既包括小公司（业务涉及疾病诊断或分析），又包括非常大的公司。我们还邀请这些公司参加合作前的选拔赛，与我们一起查看最先获得的5000个基因组信息。

所谓“合作前的选拔赛”，就是他们一起工作、一起分析数据，但却不能将任何成果，例如知识产权收入囊中。Genomics England公司代表英国纳税人拥有这些研究成果。因此，如果这些成果具有商业潜力，我们将会作为英国纳税人的代表为该研究成果申请专利，并为英国就该研究项目吸引外资。这同时也为基因组产业构建了一个有针对性的商业模式。

■ 你们有多依赖NHS？

英国的医院和大学都是NHS的一部分，这使得他们可以紧密地合作，并免费共享信息，这一点是在高度竞争，并各自独立的环境下不可能出现的情景。NHS是在整个国家的水平上运作的体系，并且是一种免费的运行模式，这使得它与众不同。

NHS可以让我们将学术界的研究人员与卫生保健系统联系起来，所以能够对各自的需求做出快速反应，例如，卫生保健系统可以实时接收收集数据和样品的请求，并且快速获得结果。

■ 你期待这种方法能显著地加快研究进度？

从基础研究到真正产生卫生保健作用通常需时17年。我们正试图将其缩减至3年。如果将健康系统与研究人员结合起来，那么就可以最大限度地提高你成功的机会。对于为研究人员提供资金的人来说，这可是一个非常有效，并且高效的方式。

因此，我不但将这种模式看作是英国健康系统转型的平台，还将其看作是全球健康系统的一个典范。

2. Perry Nisen访谈：药物研发

Perry Nisen以前是在学界和业界都有丰富经验的儿童肿瘤学家和分子生物学家，2014年他还担任了GlaxoSmithKline公司科学和创新部门的高级副总裁，之后成为了桑福德-伯纳姆-普利贝斯医学研究所（Sanford Burnham Prebys Medical Discovery Institute）的首席执行官。本次访谈中，他介绍了在大数据时代药物研发面临的挑战。

■ 把大数据引进药物研发领域，最大障碍是什么？

我们都知道，整合大数据非常有益，而且我们也要确保生物研发不要与大数据脱节。在药品研发外包和合作越来越多的今天，制药领域的障碍之一就是如何把生物医学知识和大数据联系在一起。这二者必须结合起来，否则就会误导药品研发。

在把实验室成果推进临床的过程中，研究大的、纵向的医疗数据都非常有意义。但是，在脱离临床知识的前提下，去挖掘这些数据意义不大。只有那些能把数据集和大量基础医学数据结合起来的研究人员，才能从大数据中受益。但我不认为，目前有谁能做到这点。

在学术界，两种以蛋白质PCSK9为靶标的高胆固醇药物是非常成功的例子。德克萨斯大学西南医学中心（University of Texas Southwestern Medical Center）的Helen Hobbs等人把海量的基因数据和生物知识以及化学知识结合起来，才成功研发了这两种药物。

在制药领域，Genetech公司似乎已经在生物学上作了长期投资，把临床数据和个性化治疗联系在一起，例如用于治疗乳腺癌的药物赫赛汀。

■ 制药公司怎么寻找最有可能突破的领域？

鉴于制药领域注意力缺失和研发外包的趋势，以及药物研发对风投和其他金融市场的日益需求，有时候行业会出现风险规避和从众效应。例如现有有20多家公司都在研发下一个PD-1（癌症免疫治疗目标）新药。

但我们也看到，有制药公司选择了一些非常大胆的目标，例如研发一些免疫疗法，而

目前没有多少可靠的证据显示这一疗法可能有效。

我认为对于公司研发外包和合作研究会引起很多争论，有了外包和合作他们才能发现最新的东西。但在这种情况下，是否进入和何时进入一个领域就非常关键了。

■ 举个例子，说说这种艰难决定？

以微生物研发为例吧。每个人都觉得肠道菌群很重要。数十亿的微生物生活在我们的肠道、皮肤等各处，影响各种疾病的发生和发展，甚至能影响我们对药物的响应。研究者们也已经发现了一些疾病状态和特定细菌爆发的关联，但问题是，药企什么时候才能下定决心在这个领域大量投资？我也不知道。

目前，很多用于鉴别微生物类别的数据都来自少部分人。分析大量个体的菌群数据确实非常棘手，这些受试者的长期医疗数据也非常有限。

最后，目前没有可靠的、可持续的调整菌群的方法，也没办法研究菌群的调整是否能对疾病发生、发展有所改善。

■ 为什么桑福德-伯纳姆-普利贝斯医学研究所（Sanford Burnham Prebys Medical Discovery Institute）会把基础研究和药物研发结合起来？这样做的好处是什么？

制药公司经常只看重应用科学，不关注深入的生物知识和最新的研究进展。而桑福德研究所有80个一生都致力于药物研发的工作人员。我们的首席研究员可以与药物研发者密切合作，这样能确保我们能研究一些其他人都没有做过的东西，因为这类东西太新、太不确定了。药企一般都是不见兔子不撒鹰。

建立大数据为这种合作带来了全新的局面。例如，当我们开始考虑自身免疫时，大数据能为很多关键问题提供指导：什么时候会发生什么变化，不同情况激活的是哪种T细胞或B细胞，要从哪里开始研究等。

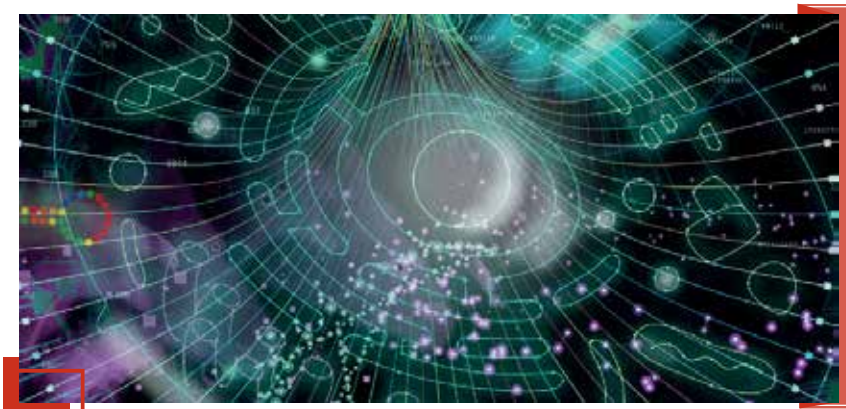
大数据生物信息学在这种情况下非常有用。我前所未有地意识到分析大数据的必要性，以及与理解生物网络的专家合作的重要性。

■ 寻找善于收集和分析大数据的研究人员是不是很难？

是的，我们也一直在努力挖掘这方面的人才。其中一个挑战就是训练和资助专人来学习这方面的技能。但这些掌握了技能的人又有可能跳槽去生物以外收入更高的行业。

我们一直在努力吸引和招募下一批系统生物学家和会分析大数据的研究者。你想找到两方面都擅长的人，太难了。有人能产生数据，有人能分析数据，但几乎没人两者都会。我相信，两种技能都能掌握的人肯定会很吃香。

九、生物医药大数据四问



尽管收集和理解大量生物医药研究数据和健康数据的工作仍然面临诸多挑战，发展缓慢，但是这项工作已然在快速改变着药学研究的全貌。

问题	它为何重要	接下来的工作	引述
怎样才能改善对相关研究至关重要的生物医药数据的长期获取工作？	数据储存，尤其是云计算环境下的数据存储的成本正逐渐降低。但是维护生物医药数据的总体花费仍然十分高昂，并且费用还在快速上涨中。而目前处理这些任务的应对模型仅仅是一种临时的替代品。	研究人员、投资者以及其他相关人员需要分析数据使用情况，并寻找可替代的模型（例如“数据共享空间”）来保证能够长期获取精准数据。另外，投资者应当在建立数据模型的过程中整合资源。	美国国立卫生研究院的Philip Bourne表示，他们的任务是利用数据科学的方法建立一个开放的电子生态系统，这将促使高效、性价比高的生物医学研究提高人们的健康水平、延长人们的寿命，并减少疾病和残疾的发生。
怎样才能更好地将临床病例结果和病人的健康信息运用于科学研究？	临床试验中未被标识（de-identified）的数据和病人的医疗信息为科学研究提供了新的契机，然而这之间存在着法律和技术的巨大障碍。临床研究数据很少能够共享使用，同时医疗记录也由于隐私、安全法规和法律层面的考虑而无法很好地应用于科学研究。	病人中的倡议者正在争取获得查看他们自身健康数据，包括遗传信息的权利。欧洲药品管理局正在发布临床报告，这份报告是新药申请所需资料的一部分。另外，如CancerLinQ项目等正在采集未被标识的病人数据。	费伦麦克德米德（Phelan-McDermid）综合征基金会的Megan O'Boyl表示，还有许多遗传信息至今无人能够解读，那么将这些信息交给病人是否适、安全和正确？不过，话又说回来，病人可能会提出，这是他们自己的遗传信息，如果他们想要这些信息，就应该得到这些信息。

（续下表）

（接上表）

数据研究中得到的知识融入护理点的医疗保健服务中？	开展精准医学研究将会很大程度上拓宽电子健康档案项目的范畴。完成这项卫生保健领域的巨大改变需要引入新的治疗手段，同时还需要对那些通过获取详细信息以做出治疗决策的临床工作者进行继续教育。	卫生系统正在试图将最新的治疗手段带入临床治疗，并建立整合了电子健康档案的“卫生保健学习系统”。例如，在难以确定癌症患者的最佳治疗方法的情况下，CancerLinQ项目可以为患者提供合理的建议。	哈佛大学医学院的Kenneth Mandl表示，为医疗改革者建立一个能够获得电子健康档案信息的标准界面，会将护理点和大数据以及网络的力量完美结合，打造出一个为健康而设的‘应用商店’（app store）。
学术界能否为生物信息学家创造更好的职业前景？	由于没有吸引人的生物信息学职业生涯规划，所以造成既掌握良好的统计学技术，又有生物学背景的科学家人才紧缺。如果这些数据分析学家流失到其他领域，那么将会阻碍医学研究前进的步伐。	科研机构将会采取措施来奖励从事多学科交叉研究的科学家，包括建立正式的职业规划途径。投资者也将采取更合理的方法来评估生物信息学家的贡献。	德克萨斯大学的Jeffrey Chang表示，如果我们缺乏生物信息学家，那么那些大数据最有前景的产物，即那些使实验室能够探索数不胜数的、难以想象的假设的产物的出现将会受到阻滞。

原文检索：

Michael Eisenstein. (2015) The power of petabytes. *Nature*, 527:s2-s4.
Neil Savage. (2015) High protein research. *Nature*, 527(1038): S6-S7.
Katherine Bourzac. (2015) Mining the motherlodes. *Nature*, 527:s8-s9.
http://www.chinadmd.com/file/pooootxpsprwtxpipzxpueoi_1.html
Charlie Schmidt. (2015) Reshaping the cancer clinic. *Nature*, 527, 10-11.
Neil Savage. (2015) Made to measure. *Nature*, 527: 12-13.
CATHRYN M. DELUDE. (2015) The details of disease. *Nature*, 527: s14-s15.
http://www.nature.com/nature/journal/v527/n7576_supp/full/527S16a.html
CLAIRE AINSWORTH. (2015) National genomics. *Nature*, 527: s5.
Eric Bender. (2015). Q&A: Perry Nisen. *Nature*, 527(1038): S18.
ERIC BENDER. (2015) 4 Big Questions. *Nature*, 527: S19.

Eason、张洁、筱悠、文佳、徐咏宁/编译

GeneCopoeia™
Expressway to Discovery

基因功能研究专家

Luc-Pair™ Duo-Luciferase HS Assay Kit

高灵敏性

双荧光素酶检测试剂盒

兼容性强 与 Promega 萤火虫和海肾荧光素表达载体相兼容！

灵敏性高 对萤火虫和海肾荧光素酶产生的光信号极度敏感！

适用性好 适用于多种不同的真核生物细胞系样品的检测！

操作简便 只需 2 步实验！

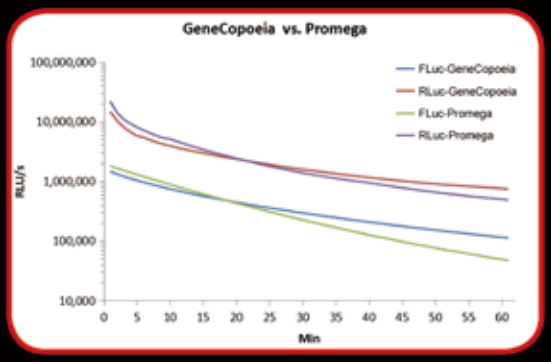


图 1. 用 GeneCopoeia 的 Luc-Pair™ 高灵敏性双荧光素酶检测试剂盒检测萤火虫和海参荧光素酶的活性。用 Promega 的 pGL4.13 和 pGL4.75 报告载体转染 HEK293 细胞，转染 48 小时后分别检测荧光信号，效果如上图所示。Fluc-GeneCopoeia 和 Rluc-GeneCopoeia 代表 GeneCopoeia 的检测结果，Fluc-Promega 和 RLuc-Promega 代表 Promega 品牌试剂盒的检测结果。

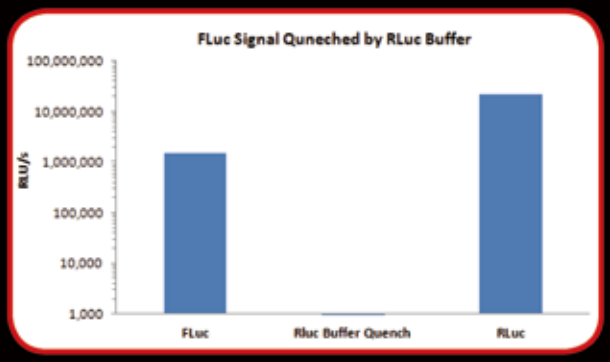


图 2. 萤火虫荧光素酶的活性被 Luo-H Buffer II 淬灭。用 Promega 的 pGL4.13 和 pGL4.75 报告载体转染 HEK293 细胞，转染 48 小时后检测萤火虫荧光信号（左柱），然后把 1 x Luc-H Buffer II【不包含（中柱）或包含（右柱）底物】加入到微孔板后，用光度计再测荧光值，实验结果表明约 99.9% 的萤火虫荧光素酶活性被淬灭（中柱），而海肾荧光素酶反应不受影响（右柱）。

产品名称	描述	货号	规格	目录价
Luc-Pair™ Duo-Luciferase HS Assay Kit	高灵敏性荧光素酶检测试剂盒，可高效、快速检测萤火虫（FLuc）和海肾（Rluc）荧光素酶的活性。	LF004	100 次反应	¥1190
		LF005	300 次反应	¥3000

Tel：4006-020-200 / 020-32068595　　Email：sales@igenebio.com　　Web：www.igenebio.com

荧光素酶检测试剂盒系列产品

产品名称	描述	货号	规格	目录价
Secrete-Pair™ Dual Luminescence Assay Kit	分泌型双荧光素酶 检测分泌型 Gaussia 荧光素酶（Gluc）和碱性磷酸酶（SEAP）的活性。	SPDA-D010	100 次反应	¥1110
		SPDA-D030	300 次反应	¥3110
Secrete-Pair Gaussia Luciferase Assay Kit	分泌型单荧光素酶 检测分泌型 Gaussia 荧光素酶（Gluc）活性。	SPGA-G010	100 次反应	¥620
Luc-Pair™ Duo-Luciferase Assay Kits 2.0	高稳定性双荧光素酶 方便、高效检测萤火虫（FLuc）和海肾（RLuc）荧光素酶活性。	LPFR-P010	100 次反应	¥1190
		LPFR-P030	300 次反应	¥3000
Luc-Pair™ Firefly Luciferase Assay Kits	高灵敏性萤火虫荧光素酶 方便、高效检测萤火虫（FLuc）荧光素酶活性。	LF007	100 次反应	¥600
		LF008	300 次反应	¥1000
Luc-Pair™ Renilla Luciferase Assay Kits	高灵敏性海肾荧光素酶 方便、高效检测海肾（RLuc）荧光素酶活性。	LF010	100 次反应	¥600
		LF011	300 次反应	¥1000

	Vector Type	Luc-Pair™ Luciferase Assay Kits		Secrete-Pair™ Luciferase Assay Kits	
		Duo-Luciferase Assay Kit 2.0	Duo-Luciferase HS Assay Kit	Dual Luminescence Assay Kit	Gaussia Luciferase Assay Kit
Reporter Genes		Firefly and Renilla Luciferase	Firefly and Renilla Luciferase	Gaussia Luciferase/ Secreted Alkaline Phosphatase	Gaussia Luciferase
Key Feature		For Enhanced Stability	For High Sensitivity	For Secreted Reporters	For Secreted Reporters
GLuc-ON™ Promoter Reporter Clones	pEZX-PG04			√	
	pEZX-PG02				√
	pEZX-LvPG04			√	
	pEZX-LvPG02				√
miTarget™ 3' UTR miRNA Target Clones	pEZX-MT05			√	
	pEZX-MT06	√	√		
GLuc-ON™ Transcriptional Response Element Reporter	pEZX-PG02				√
GeneCopoeia's Vector Options	pEZX-FR01*	√	√		
	pEZX-FR02*	√	√		
	pEZX-FR03*	√	√		
	pEZX-GN01*				√
	pEZX-GN03*				√
	pEZX-GA01*			√	
	pEZX-GA02*			√	
	pEZX-GA03*			√	
	pEZX-LvGN01*				√
	pEZX-LvGA01*			√	
Promega's Vector Options	pGL4 Luciferase Reporter Vectors	√	√		
	pmirGLO Dual-Luciferase miRNA Target Expression Vectors	√	√		

GeneCopoeia, Inc.

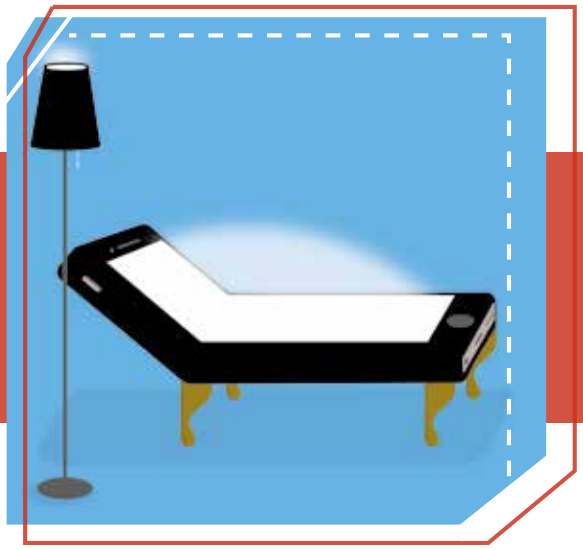
Tel： 4006-020-200 020-32068595
Email： sales@igenebio.com
Web： www.genecopoeia.com
www.igenebio.com

扫描二维码关注
官方微信账号
易锦生物

热点

Hot Topics

精神健康APP



智能手机应用号称能帮助缓解上瘾、精神分裂等精神疾病，但这些APP的有效性都还需验证。

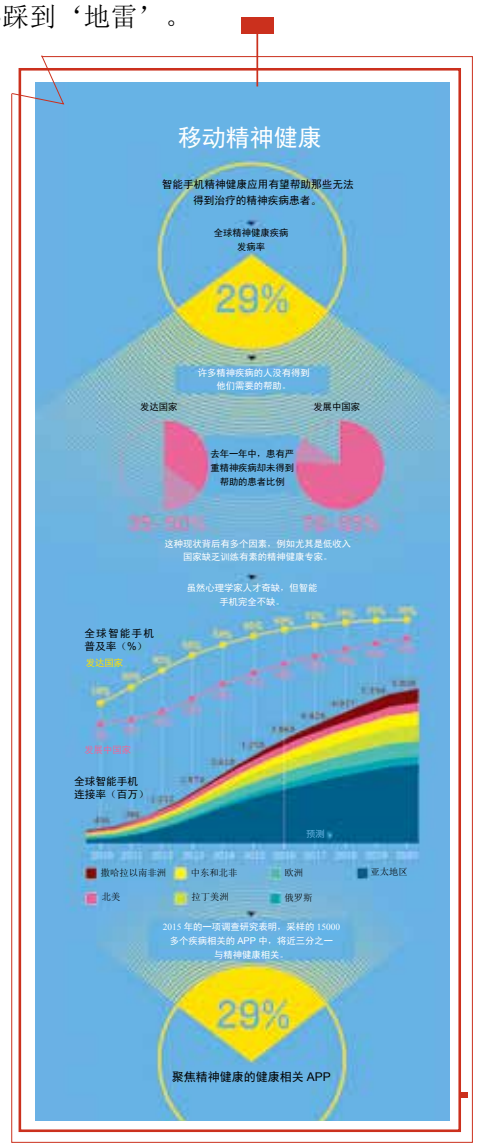
在苹果应用程序商店键入关键词“抑郁”开始搜索，屏幕上至少会弹出一百多个相关程序。这些程序五花八门，有诊断抑郁症的、有跟踪情绪的，还有帮助人们保持积极心态的。此外，还有抑郁症治疗催眠应用、感恩日记（号称是“最简单、最有效的心态调整方式，每天仅需五分钟”），诸如此类。这仅仅是冰山一角。焦虑、精神分裂症、创伤后应激障碍（post-traumatic stress disorder）、饮食失调和成瘾等精神问题的相关应用也层出不穷。

这个新兴产业从侧面反映了日益增长的精神卫生市场。据估计，大约29%的人一生中会至少经历一次精神障碍。来自世界卫生组织（WHO）的数据显示，发达国家里55%的精神障碍患者没有得到有效治疗；而在发展中国家，这个数字达到85%。移动健康应用程序可能有助于填补这个空白。鉴于智能手机的普遍性，应用程序可作为数字化的治疗工具，这相当于在每个人的口袋里都放了一个便携式的治疗师，这一点在农村和低收入地区非常重要。“这有助于我们治疗研究那些不会到医院就诊的病人。” 达特茅斯精神病学研究中心（Dartmouth Psychiatric Research Center）精神卫生项目的负责人Dror Ben-Zeev这样说道。

公共卫生组织已经接受精神健康APP。WHO在其2013-2020年精神卫生综合行动计划（Mental Health Action Plan 2013–2020）中提出，“通过使用电子和移动健康技术，促进精神疾病患者的自我护理”。英国国民保健服务（UK National Health Service, NHS）网站列出了其正式背书的在线心理健康资源，其中包括一些应用程序。

但APP技术的更新远快于精神健康科学的发展速度。尽管一些证据表明，基于经验、精心设计的心理健康APP可以改善患者的预后，但绝大多数APP的效果并不明确。有些可能有效，有些可能无效，有些甚至可能有害。科学家和卫生官员现在已经开始对精神健康APP的潜在好处和缺点进行深入的研究，但消费者能得到的信息和指导仍然非常少。

哈佛医学院（Harvard Medical School）的精神病学家、美国精神病学会（American Psychiatric Association）的智能手机APP评估计划（Smartphone App Evaluation Task Force）的主席John Torous表示，如果你键入“抑郁症”，很难知道弹出的程序的质量高不高，是否有效，以及是否安全。现在，精神健康APP质量参差不齐，消费者很容易不小心踩到‘地雷’。



■ 快乐APP

电子干预是心理学常用的手段。多篇文献显示了基于互联网的认知行为疗法（cognitive behavioural therapy, CBT）（一种旨在纠正有问题的思想和行为，有效治疗抑郁症、焦虑和饮食失调等疾病的方法）等。但是，很多这类在线治疗方案都需要患者长时间对着电脑完成。

而智能手机APP可随身使用，非常方便。“这种治疗方式更灵活，能配合患者的生活方式，并且解决了羞耻感的问题——很多患者认为精神疾病患者有羞耻感，不安于看医生。”英国国家健康研究所（National Institute for Health Research）MindTech研究中心的项目经理Jen Martin这样说道。Martin一直致力于研发和测试新的精神健康技术。

一个最有效的开放应用程序能够满足对灵活性的需求。2010年，美国政府心理学家对有创伤后应激障碍的退伍军人的研究发现，这些患者需要一个一旦发病就能使用的治疗工具。美国退伍军人事务部（US Department of Veterans Affairs）国家创伤后应激障碍中心的临床心理学家和移动APP项目负责人Eric Kuhn指出，这些军人需要一些可以随时使用的工具，比如说在超市排队发病时需要的工具。

该部们与美国国防部（US Department of Defense）启动了创伤后应激障碍教练（PTSD Coach）APP，一款于2011年初发布的免费智能手机APP。任何一个经历过创伤的人都可以使用这个APP来了解PTSD，跟踪症状，并建立朋友和家庭成员组成的支持网络。该APP还提供卓有成效的情绪应对策略；它会提醒用户观看YouTube上的搞笑视频来分心，或者引导用户参与可视化练习。

PTSD Coach在苹果应用商店上架三年多，在86个国家中的下载次数超过150000次。几个小型研究表明其有效；2014年一个受试者为45个老兵的研究中，超过80%的患者认为，该APP可以帮助他们跟踪和管理症状，并针对症状提出有效的解决方案。很快，更多的研究成果将出炉。Kuhn等人最近完成了一项包含120人的随机试验，荷兰的一个团队目前正在分析1300名患者使用SUPPORT Coah APP的试验数据。

智能手机APP可以与用户积极互动，询问用户的情绪、想法和总体幸福感。Ben-Zeev开发了一个名为FOCUS、针对精神分裂症患者的APP。该APP每天会几次提示用户回答问题，如“你昨晚睡得有多好”或者“你的心情怎样”。如果用户回答说，他们睡得不好，或是感到焦虑，该APP将提出解决策略，如限制咖啡因的摄入量或做一些深呼吸练习。

一些应用程序可以帮助人们与卫生保健人员相联系。ClinTouch，一个由英国曼彻斯特大学（University of Manchester）的研究人员设计的精神症状评估程序，能够分析用户的情绪状态，评估其复发的迹象；它甚至可以通知临床护理团队。

小的可行性研究：一般都用于评估一项干预是否可行，而非是否有效，即表明，患者喜欢并使用这两个APP。2014年的一项研究发现，FOCUS使用一个月以上，用户的精神症状和抑郁都有所改善。研究人员正在针对FOCUS和ClinTouch进行随机对照试验。

一些研究人员认为，智能手机收集了用户的移动模式或通信活动，这也能侧面反映用户的精神健康。“手机是一本有趣的日记，记录着你的生活。”旧金山Ginger. io公司的共同创始人和首席执行官Anmol Madan这样说道。研究表明，智能手机的使用可以预测心理健康症状的变化，例如，发送文字的减少可能意味着患者的焦虑加深。

Ginger.io仍处于测试阶段，该APP可监控用户的行动模式和通信模式，如果检测到恶化迹象，会通知用户的医生。

■ 证据缺乏

支持这些应用程序的证据在不断积累中。但这只是一门科学的起步阶段。许多研究还限

于试点研究和随机对照试验，规模小，且未经重复。许多研究都是由应用程序开发人员进行的，而不是由独立的第三方研究人员进行的。Torous表示，很多研究缺乏安慰剂对照组，因此“数字化安慰剂效应”有可能是部分研究有效的原因，一些研究人员也证实过这种可能性。Torous解释说，人们和智能手机联系密切，从熟悉的个人设备上接收信息和建议足以让很多人感觉好很多。

但事实上，大多数APP完全未经测试。2013年一次对囊括了商业化应用商店的1500多个抑郁症相关APP的研究表明，只有32个APP有发表的研究型文献支持。同年发布的另一项研究中，澳大利亚研究人员采用更严格的标准，寻找评估上架的APP对精神健康症状的影响的文献。他们只找到了8篇针对5个不同APP的论文。

同年，NHS推出了“安全可信”的健康APP名录，其中包括14个专门治疗抑郁症或焦虑的APP。但当两个团队的研究人员仔细研究这些应用程序时，他们发现只有4个APP提供了理论支持。英国Lifecode Solution中心的卫生经济学家Simon Leigh负责分析这些证据，他表示，由于有效性研究成本很高，APP开发商又不乐意削减营销费用，因此出现这种结果他毫不意外。

一项单独的分析显示，NHS最初推荐的35个移动健康APP使用网络传输信息——例如邮箱地址、姓名、出生日期等，三分之二对这些数据不加密。

去年，NHS撤消了这个APP库，精简了可靠的在线心理健康服务名单。NHS没有回应各方提出的质疑，也没有接受官方采访，但他作出了声明：“我们正在努力升级健康APP库，该数据库是根据2013年一项试点研究的评估结果选出的。评估是根据一系列明确的标准展开的，包括数据保护。”

心理健康APP的监管并不透明。一些需要在特定医疗背景下使用的APP被认为属于医疗设备，因此可能受到英国药品和保健品管理局（UK Medicines and Healthcare Products Regulatory Agency）、美国食药监局（US Food and Drug Administration, FDA）或类似的机构监管。但这条界线比较模糊。一般情况下，一个声称能预防、诊断或治疗某个疾病的APP会被归类为医疗设备，从而受到严格监管。但声称可能改善心情，提供指导的APP就不会面对这类问题。FDA表示，它仅监管那些会对患者带来最大风险的APP，即使是归类为医疗设备的精神健康APP，如果被认定为低风险，可能也不会受到监管。

但目前潜在的风险并不明确。Martin指出，从小的方面来说，风险是浪费钱和时间；从大的层面来说，尤其是精神健康方面来说，这些APP可能有害，会提出危险的建议，或阻碍病人接受合适的治疗。

澳大利亚一组研究人员对82个商业智能手机躁郁症相关APP进行了审查。他们发现有一些人提出了“非常错误”的信息。一个名为iBipolar的APP，建议人们在躁狂发作时，喝烈性酒助眠；而另一个名为What is Biopolar Disorder的APP则提出，躁郁症可能传染。这两个APP都已被下架。

Martin说，至少在欧洲，APP可分为两类，一类是商业APP，由公司开发，以营利为目的，少有证据支持，没有评估计划；另一类是政府或学界支持的，评估更严格。问题是，前者通常更具吸引力，而后者上架更耗时，而且由于美观、人性化等问题，后者看起来非常过时。Martin指出，这是普遍存在的一种现象。

■ 意外后果

即使是善意的APP也可能产生不可预测的结果。以Promillekoll为例，Promillekoll是由瑞典的国有白酒类零售商开发的一款智能手机APP，旨在帮助遏制酗酒。用户在酒吧或者派对时，该APP能记录下他们喝的酒量，显示用户血液中乙醇的浓度。

当瑞典研究人员以大学生为对象测试该

APP时，他们发现被随机分配到使用该APP的男性饮酒量没有增加，但饮酒次数增加了。2014年这篇论文发表，研究人员写道：“我们只能推测，该APP的用户可能会感到更加自信，他们可以依赖APP来减轻饮酒造成的负面影响，因此认为可以更经常地喝酒。”

科学家们说，还有一种可能——该APP让男性将饮酒视为一场游戏。“我认为，这些APP有点像玩具，”斯德哥尔摩卡罗林斯卡医学院（Karolinska Institute）的临床心理学家、该研究报告的作者之一Anne Berman这样说道。此外，还有其他风险。曾帮助过ClinTouch的开发、曼彻斯特大学（University of Manchester）的研究者John Ainsworth指出，ClinTouch早期试验中，研究人员发现，这个症状监测APP实际上会加重少数精神障碍患者的症状。他指出，他们需要小心监控用户使用该APP的初始阶段，并确保患者得到了良好的监控。

今年早些时候发表的一个试验性试验中，二十名患有PTSD的美国退伍军人被随机分配到两组，A组自行使用PTSD Coach，B组在初级护理人员的指导下使用该APP。试验时间为8周。在试验结束时，B组10名患者中有7名PTSD症状明显减轻，A组中仅有3名有明显症状减轻。

但是，如果APP需要医疗支持，那么把它作为一个简单、低成本、适用于大众的医疗手段就不可行。“人们认为，APP能彻底解决问题。”澳大利亚新南威尔士大学（University of New South Wales）黑犬研究所（Black Dog Institute）的主任Helen Christensen这样说道。Christensen在研发和研究精神健康APP方面非常有经验。他指出，事实上，只有围绕着APP建立各种系统，人们才能获取各种医疗资源。这才是解决问题的关键。

在发展中国家推广精神卫生APP的使用，是个巨大挑战。虽然移动技术正在迅速发展，但很多人没有或者买不起智能手机或移动流量。APP的内容还需要被翻译成本国语言，并体现本地文化。Ben Zeev指出，即使精神健康APP在发达国家有效，它也不一定在发展中国家有效。不同国家对于同一个疾病的专有名词，可能都有很大差别。“在美国，我们说，精神健康患者要‘听见声音’，意思是说要多与长辈沟通，但不同地区对这句话的理解会有很大差别。”

在这一点上，APP可以在低收入地区提供优质医疗服务的概念仅仅是理论。正如南非医学研究委员会（South African Medical Research Council）科学家Natalie Leon所说，精神卫生领域现在正处于这样的尴尬阶段，这种理论提供了“可能有效”的希望。

良好实践

为了实现这个愿景，必须测试各个APP。有报告称，2013年到2015年间，ClinicalTrials.gov网站上注册的移动医疗试验数目翻了一倍，从2013年的135增加到2015年的300。其中聚焦心理和行为健康的试验数量增加了32%。

赢得称赞的数字健康公司是Big Health，该公司由英国牛津大学（University of Oxford）睡眠科学家Colin Espie和企业家Peter Hames共同创立。该公司总部位于伦敦，第一个产品是Sleepio，一个可以通过智能手机或在线获取的失眠治疗APP。该APP可以提供一系列以证据为基础的失眠解决策略，包括管理焦虑和侵入性的思想，促进放松，建立良好睡眠环境和生活习惯。

在Sleepio进入测试之前，Espie坚持创建该APP的安慰剂版本，该版本和原版本具有相同的感官，从而可全面排除干扰因素。2012年发表的研究中，Espie等人发现，在随机试验中，相对于安慰剂组，使用Sleepio的失眠症患者睡眠效率——实际入睡时间和分配给睡眠的时间的比例——有明显提高，白天的工作状态也有轻微提升。在后续的2014年随访研究中，这些使用者也减少了侵入性的思考

（这类思考往往会干扰睡眠）。

Sleepio现在正招募志愿者，以展开更大规模、更国际化的试验。有多个研究团队会参与到这项研究中。参与这项研究的受试者可免费试用Sleepio。

Espie认为，这种方式能推进数字健康的发展。基于移动电话的治疗手段也应当得到测试和评估。不能因为治疗手段是通过APP实现的，就相对疏忽，这是对患者健康的不负责。

原文检索：

Emily Anthes. (2016) Mental health: There's an app for that. *Nature*, 532(1038): 20-23.

张洁/编译

百态

Amazing Lives

铁甲水上漂？

慢速拍摄睡莲小萤叶甲虫（*Galerucella nymphaeae*）在水面飞行的图像。

别眨眼，不然你可要错过一场精彩的演出了！前一秒，睡莲小萤叶甲虫（waterlily beetle，*Galerucella nymphaeae*）还停在池塘表面，可就在一瞬间，它竟然消失了！对此，美国斯坦福大学（Stanford University）的Manu Prakash指出，那速度快得根本来不及看，简直不可思议，唯有湖面上泛起的阵阵涟漪，才能证明它们曾经出现过。正是因为观察到这种不同寻常的现象，Prakash才决心找出它们顺利完成神技的原因所在。忆及此事，他不由得再次提到当初的情形：用餐盘盛水充当甲虫的“池塘”，大小正好，然后在厨房里让它们自由活动，并进行拍摄。这么做的原因有点滑稽：因为如果让它们在实验室里（自由活动），恐怕就难觅芳踪了。而当他看到拍到的第一部“影片”时，立即意识到自己正在做一件很特别的事情：甲虫看上去在滑水，但其实，它们是在以高达0.5 m s^{-1}的速度前行——相当于一个人以大约500 km h^{-1}的速度行进——且仅用其翅膀作为推动力。这样，它们看上去就像是在飞行，而实际上只不过是在贴着水面滑翔。看到这儿，Prakash觉得太有意思了，于是想进一步了解甲虫这种神秘的水面飞行到底是怎么回事。

Prakash与暑期实习生Thibaut Bardon、Dong Hyun Kim以及研究生Haripriya Mukundarajan合作，用高速相机拍下了甲虫的奇异动作。结果用Prakash的话说，简直难以想象这些小家伙们是怎么做到的。Mukundarajan描述了它们复杂的飞翔前准备运动：先是提起位于中部的对足——以防在飞行中阻碍翅膀运动，然后甩干每条腿上的水珠，最后轻轻地把肢末端的爪垫浸入水中再抬起来，准备起飞。当四对对足的末端平衡之后，甲虫就会张开背部的翅膀，连续拍上几次，扬帆飞翔，瞬间转变为飞翔模式。它们以115Hz左右的频率拍打着翅膀，呈8字形特征推动自身前行。可是，当它们飞过自己划出的波痕链时，却无法平稳地划过光滑的水面，而是几乎像在到处坑洼不平的路上前进，看上去就像是沿着过山车猛冲一样，其实这些“坑洼不平”不过是它们自己制造出来的波痕。

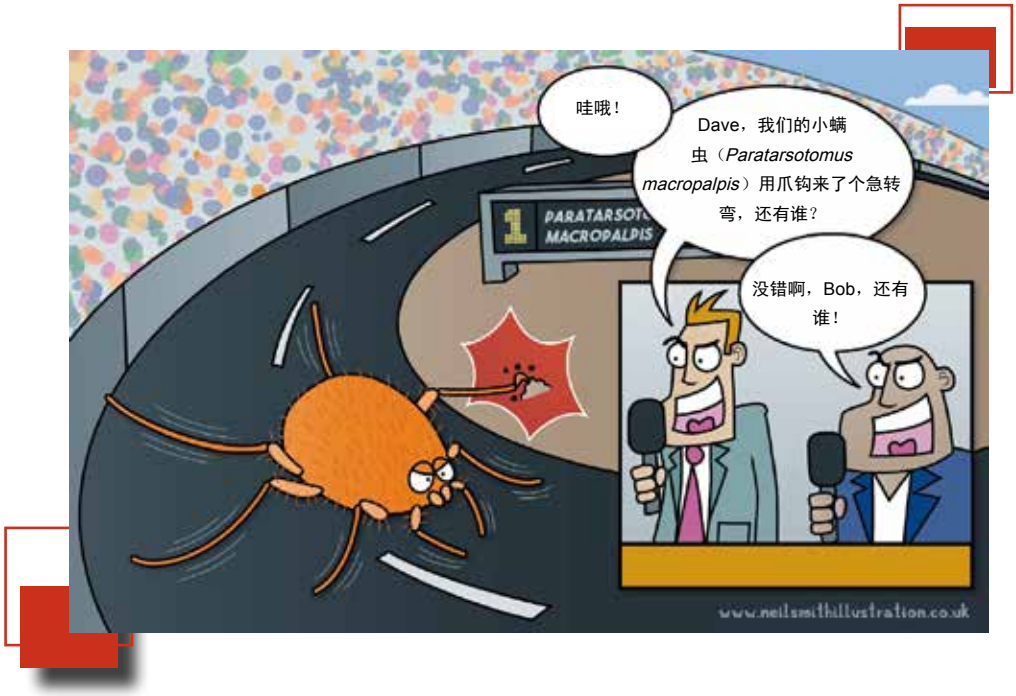
对这种令人意外的颠簸飞行，Mukundarajan和Prakash感到很是迷惑。他们对甲虫划过水面时反作用于它们的力进行分析，结果意识到：它们施展了一手极为漂亮的绝技，使其脚爪附有的表面张力（能使它们牢牢地固定于水面）和翅膀产生的升力间产生极为精细的平衡。然后，Mukundarajan用一套方程式描述甲虫的运动，解释了那泄露行踪的波痕——甲虫高速飞翔行为的唯一可见指标——是如何产生的。据Prakash解释，甲虫翅膀的每一次拍击，都能产生瞬间推动它们向下的力，使之在水的表面上下弹动。而当它们达到某个特定速度时，还会自发产生另一种波——重力表面张力波（capillary gravity waves）。据两位研究者所言，一种昆虫能够沿着池塘表面飞行，同时保持与水体接触而不会飞走，这种情况是比较少见的。那么，我们知道这些知识有什么用呢？Prakash解释说，他们做出的这个新数学模型或许能够解释其它外来物种（包括海蝇，一种特别享受冲浪运动的水生动物）的滑翔运动。对此，他们还是蛮有信心的。

原文检索：

Mukundarajan, H., Bardon, T. C., Kim, D. H. and Prakash, M. (2016). Surface tension dominates insect flight on fluid interfaces. *J. Exp. Biol.* 219, 752-766.

文佳/编译

螨虫是如何打破世界纪录的？



在华纳公司制作的动画片《群星总动员》（Looney Tunes）中，你能想到的加州第一速度的担当者是谁？当然是老被歪心狼（Wile E. Coyote）紧追不舍的BB鸟（Roadrunner）——名副其实的路跑者。但其实你不知道，还有一种更小的生物更值得你关注，那就是来自美国波莫纳学院（Pomona College）的Jonathan Wright想要介绍的一种螨虫：在某种程度上说，它拥有动物世界中所记载的最快的相对速度。为此，研究小组专门去拍摄了本土螨虫（*Paratarsotomus macropalpis*）。这种虫子在学院园区旁边的马路上驰行，横跨炙热的混凝土，并引以为乐。你别看它们个子小，却能以高达0.26 m s^{-1}的极速奔跑。研究小组测量了螨虫的跑速，并与其虫体大小相比较，结果其中一只幼虫的速度竟然达到了每秒行进323个体长距离（bodylengths，(BL) s^{-1}）。这个华丽丽的成绩使大家完全惊呆了，因为原世界纪录保持者——澳洲虎甲虫（tiger beetle）只能跑到171 BL s^{-1}，与它相比真是弱爆了。

螨虫轻盈敏捷的速跑固然令人惊讶，但它们到底是如何做到的呢？研究小组继续研究它们的步法。他们记录下螨虫的步频，结果产生了一个令人惊喜的数据——135Hz，据称是已有报道中肌肉所能承受的最高值。然后，研究小组对螨虫的转弯行为进行了分析，结果发现它们在跑步中使用了两种策略：一是高速急转，速度约为795 deg s^{-1}，以第三对内侧足为中心快转，其脚爪如同钩子一样紧紧地抓住地

面；二是较慢的转弯，速度约为567 deg s^{-1}，它们在转弯处用外侧腿迈大步，内侧腿迈小步。Wright等人还从螨虫起跑开始研究，结果也惊讶地发现，这种小动物在短短15-20 ms内就能达到0.15 m s^{-1}的速度，当起跑时，加速度可达7.2 m s^{-2}；当要停下来的时候，减速度又可达到10.1 m s^{-2}。

太神奇了，不是吗？

原文检索：

Rubin, S., Young, M. H.-Y., Wright, J. C., Whitaker, D. L. and Ahn, A. N. (2016). Exceptional running and turning performance in a mite. *J. Exp. Biol.* 219, 676-685.

文佳/编译


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