让云端数据应用更轻松

科学家呼吁更便捷地获取及分析云端大型生物数据集

以前，人类基因组学研究人员总被获取巨大数据集的挑战阻碍；今年年初，当研究人员看到原来的主要路障消失后，该研究领域为此集体欢呼。今年 3 月，美国国立卫生研究院（NIH）取缔了自 2007 年开始对其库存的数百万人类基因组和其他遗传信息的使用云计算储存和分析的限制，该储藏库包括基因型和表型数据。

在量入为出的基础上，云服务可以为客户提供大规模储存和计算能力。因为这些服务在互联网上可以获得，大量用户可共享硬件，许多资助机构担心，客户使用基因组学信息会威胁到提供样本者的私人信息。NIH 态度的回转部分上是因为旨在解决人类基因组研究挑战的呼声越来越高，获取大数据集的挑战正在阻碍科学家的科研能力，尤其是那些在原来工作基础上复制和建立的科研工作。

为了充分发挥云计算提供的潜力，加拿大多伦多安大略癌症研究所信息学和生物运算部主任 Lincoln D. Stein 和同事近日在《自然》杂志发文，敦促 NIH 和其他机构为储存的最受欢迎的主要基因数据集买单。通过这种方式，才不会让数千万研究工作者因为要从一个储存库向他们选择的云端独立传输数据而浪费时间和金钱，被授权的科学家才可以在需要时便捷、经济地进入全球云共享。

> 通过正确的途径进行云计算，人类基因组学界将为在许多领域中与大数据战斗的研究者铺平道路。

谷歌云服务是分析大型基因数据集的研究人员使用的工具之一。

图片来源：KeystoneUSA-ZUM

海量数据

多亏测序技术的迅速发展，呈交给公共档案库的基因组数据量现在已经达到数千万亿字节（PB）范围。例如，在国际癌症基因组学会（ICGC），来自 17 个国家的团队在仅仅 5 年内已经积累了超过 2PB 的数据集——约相当于 50 万个光盘的容量。

利用一个普通的大学互联网连接，要花费超过 15 个月才能把如此大规模的数据集从储存库中传输到一名研究人员的本地连接计算机网络中。不说处理数据，单是需要用于储存的硬件就要花费 100 万美元左右。

云服务则提供了“弹性”，它意味着研究人员可以根据需要，用尽可能多的计算机迅速完成一项分析，而且只为使用的计算时间付费。通过从研究人员笔记本终端控制的基于云的虚拟计算机进行分析，若干名研究者可以轻松地实现平行工作，共享数据和方法。因此以前花费数月才能完成的大型基因组数据分析现在数日或数周内就可以解决。

近来，云服务也已经和大多数学术数据中心一样安全，而且往往比后者更加安全。现在，相关服务由包括亚马逊、谷歌、微软等在内的大型商业公司提供，而规模小一些的公司则聚焦于基因组研究，如加州的 Annai 系统，还有若干家学术机构，如英国辛克斯顿的欧洲生物信息研究所，这些服务商使用强加密——如防火墙和秘钥链——管理数据和系统，这些可以控制谁可以获取数据，并给数据拥有者提供密切监管相关使用情况的工具。

但一些人类基因组研究的主要资助机构对此却非常审慎，例如一些欧盟资助机构建议研究人员遵照欧盟隐私权法案，将基因组数据放在这些机构的司法权监管之下。但是由于云计算的经济性、灵活性、可靠性和安全性已经发展到今天的程度，Stein 等人期望，在未来数月内可以看到相关交易大规模转向云服务，他们对 NIH 加速这一转变的决策也表示拥护。

现在，在降低研究成本的同时，已经是时候建立机制和实践，让云计算的效率和利用最大化了，Stein 等人指出。

通道控制

为了获取储存在中心数据库如dbGaP或欧洲基因组档案（EGA）中的人类基因组和其他数据，研究人员必须获得数据获取委员会（DAC）的批准。目前，如果两家独立研究团队想利用一个私人云或商业云的同一组数据，它们需要分别获取相关 DAC 的批准，才能在互联网上复制数据并把其储存在它们选择的云端。

两个团队都需要等待数据的复制，而且当数据复制后，只要它们需要这些数据，每个团队就需要为相应的储存付费，由于数以千万计的研究组开始做同样的事情，这一过程会浪费研究人员数年时间和纳税人数亿美元。即便是可以不受约束地获取云服务，当前对大多数团队来说，利用大规模公共基因组数据集仍然有些不切实际，因为在把数据从储存库传输到云端需要花费大量时间和成本。

好的解决方式是向有关资助机构要求，被上传到最受欢迎的学术云和商业云中的每个主要基因数据集都可以获得，并且为这些数据在云端长期储存付款。通过这种方式，数据就需要仅被复制一次，研究人员也只需要在进行分析时，只对暂时储存付费。

目前，若干家云服务供应商正在提供免费储存研究数据集的服务，或是在大量补贴率的基础上促使更多研究人员使用它们的服务。例如，亚马逊网络服务并未对千人基因组计划——一项统计人类基因变异的国际项目，目前数据总量已超过 200 兆字节（TB）——发布的测序结果征收任何费用。而 Annai 系统则储存了日益增长的 ICGC 数据集的一个子集。

Stein 等人设想，诸如 dbGaP 或 EGA 等实体将会继续作为主要数据保管机构，它们的 DACs 将仍然会审核以及授权云端的数据使用。如此，基因组云计算甚至可以产生微观经济现象。例如，一名向云端贡献了有价值数据集的遗传生物学家会在处理过程中接收到信誉积分。同理，一名计算机科学家如果贡献了可以让其他遗传学家更有效地找到癌症变异的软件包，那么，每次有人在使用这个软件包时，他本人就会收到信誉积分。

基因标准

“人类基因组学界也为战斗在数据超负荷战役中的其他领域的研究人员铺平道路。”随着时间的发展，将会出现良性循环。可以融合大数据集的能力将会让研究人员把罕见的基因变异和疾病产生联系，而类似的成功会鼓励其他人储存更多数据集，并促进更强大软件的发展。这样的机制也可以和资助机构把一些数据集储存在特定云端的要求相结合。

当云服务上升至主导地位后，一种可能的风险是，单独一家云服务供应商可能会控制价格，因此会对科学的执行产生微妙影响。为了阻止这种可能性发生，资助机构应该在多个云端储存同样重要的数据集。这样做还有助于解决管辖权症结问题，例如基因组数据起源于欧洲，所以就被限制储存在欧洲的云端。

实现这一设想需要工作、技术和法律，Stein 等人指出。例如，目前对于囊性纤维化研究员来说，没办法写出用于搜索 dbGAP 数据库的软件，从而从相关疾病人群中找到获得的基因序列。而系统地对这些数据进行标注，例如特别是对样本组织的来源作标注，就有助于解决这一问题。自 2001 年起，期刊出版商已同意接收核糖核酸微型阵列研究结果，研究需要用一种微型阵列实验标准的“最小信息量”描述其数据。对于基因组学数据来说，同样如此。

在法律层面，必须建立相应的规则以阐明资助机构、数据保管机构、云服务供应商和利用基于云的基因组数据的研究人员的角色和责任。例如，如果有人把一个 ICGC 的基因传输到脸谱网上，在以上这些参与者中，应该由谁对其负责？幸运的是，在过去两年中，全球基因组学和健康联盟已经准备了一个规范——《共享基因组和有关健康数据责任人框架》。

同时，美国国家癌症研究所也设立了若干试点项目，探索共享和分析云端基因组数据的实践活动。而 NIH 和其他资助机构也已经开始讨论各种“生物医药共享”概念，其中一些概念包括：通过正确的途径进行云计算，人类基因组学界将为在许多领域中与大数据战斗的研究者铺平道路。（红枫）

与老祖宗抢地盘

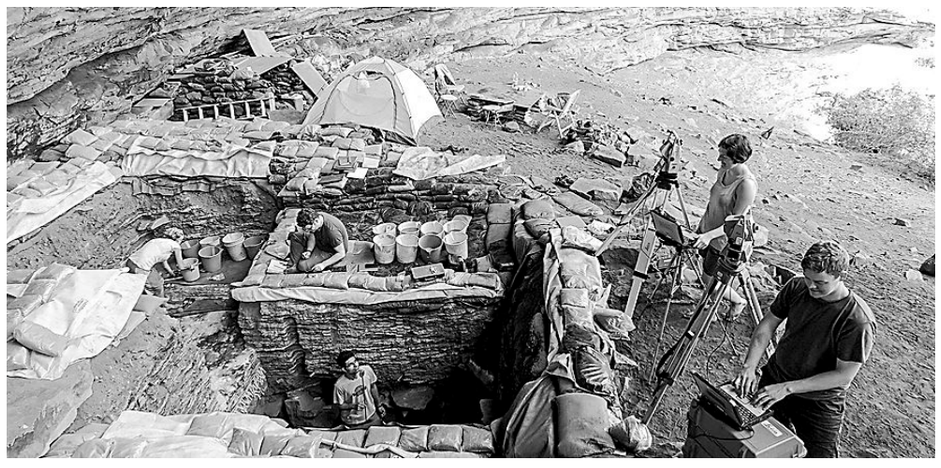
南非社会发展威胁早期人类家园

早期人类对南非 Sibudu 洞穴——在河畔悬崖上凿出的一处岩石居所——如此着迷，他们几乎连续在那里生活了好几万年。在那里，早期人类留下了许多重要技术的首批有记录的证据：包括寝具、弓箭、捕捉小猎物的套圈，他们的手工制品也对追溯早期人类文化如何发展提供了关键线索。但是现在，21 世纪的人类开始在原来隐蔽的洞穴旁比邻而居，考古学家表示，人类祖先的无价记录正处于危险之中。

一场声势浩大的工业和住宅开发计划（KDC）已决定在 Sibudu 洞穴——位于夸祖鲁－纳塔尔省、距离德班北部约 40 里——附近上马，最近几周，考古学家和支持者已经发出了两封请愿信，反对省政府批准的 621 公顷土地的发展计划。该计划将建立重工业和轻工业、两所学校以及约 2700 个别墅与公寓混合的住宅单元。该计划预计仅在这些建筑结构和 Sibudu 洞穴之间预留 200~300 米的缓冲区，考古学家认为这样近的距离将会带来灾难性的后果；而开发者的土地可延伸至距离洞穴 20 米附近。

“事实上，根本就没什么缓冲区。”南非约翰尼斯堡金山大学领导 1998 年 Sibudu 洞穴首次主要出土工作的Lyn Wadley说。她担心，在附近河流收集柴火的居民或四处闲逛的青少年会踩坏洞穴脆弱的地层。

考古学家称，Sibudu 是世界上最重要的理解早期智人文化和技术如何发展的考古遗址；该遗址正在考虑被列为联合国教科文组织世界文化遗产项目，并且已经被提名为南非国家级遗产。洞穴厚厚的沉积层是世界上时间最久、年代序列最全的早期人类住址。那里的地层涵盖了距今 7.7 万年前至 3.8 万年前的历史，甚至有可能延伸至距今 10 万年以前。到目前为止，出土文物包括从精美的石器、骨制工具到具有象征意义的手工制品，如用海贝制成的穗子，还表明早期人类已具有诸如事先作好计划等认知能力。



考古学家已经在南非 Sibudu 洞穴出土了古工具、珠宝，甚至还有早期人类留下来的寝具。

图片来源：NICHOLAS CONARD

Sibudu 洞穴发现于上世纪 60 年代，但是直到上世纪 90 年代才开始全面发掘出土工作，研究人员目前已发表了 100 多篇相关同行评议文章，包括目前的发掘工作指导人、德国图宾根大学考古学家 Nicholas Conard 近日在美国《公共科学图书馆·综合》上发表的一篇论文。这篇论文报告了约 5.8 万年以前，Sibudu 洞穴工具制造策略发生的 3 次转变，证明了当时那里居住的早期人类的文化适应性。

这个遗址“给我们提供了一幅在人种发展过程关键时期文化演化的独一无二的图景。”法国波尔多大学考古学家 Francesco d'Errico 说，“如果可以保存下来，它仍然具有可以让我们洞察历史的巨大潜力。”

因此，当 Sibudu 洞穴附近辽阔的土地项目——KDC 项目和巴利托开发计划在 2010 年宣布开发其所属区域后，Wadley 和其他人奋力阻止其实施。目前，621 公顷的开发用地主要种植甘蔗，但是未来住宅单元和工厂会导致大批人涌入该地区，导致当前遗址附近稀少的交通量迅速膨胀。

但今年年初，该项目已经通过了环境影响评估，并在 4 月份被夸祖鲁－纳塔尔省经济开发、旅游和环境事务部快速批准。“我很难让那些人相信，Sibudu 非常重要，区域开发会对它造成很大威胁。”Wadley 沉痛地说。夸祖鲁－纳塔尔大学教育专家、电影制作人 Charlotte Mbali 说，省政府“很明显没有考虑考古学家。”Mbali、Wadley 和其他人正在呼吁在洞穴附近留出更大的保护区。

一名不愿具名的夸祖鲁－纳塔尔省官员则表示：“在南非，确实很难阻止开发工作。”KDC 一名负责人 Patrick Conway 说：“我们对洞穴和它的重要性非常尊重。”他表示计划的缓冲带对于保护该遗址安全完全足够。他还强调，Sibudu 的支持者应该为遗址安全负责。“那不是我们的土地。他们应该在周围建栅栏。他们对保护遗迹没做什么事情。”他说。

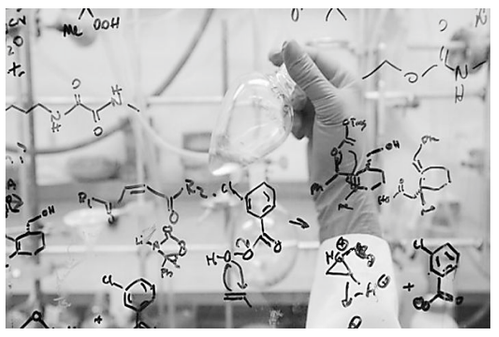
Wadley 还击称，实际上缓冲区最多只有 200 米。而 Conard 强调，至今为止 Sibudu“距离被踩踏出来的路仍很远，而且不需要保护”。

“当地居民和考古学家已经建立了长期的信任和互相尊重的关系。”Conard 说，“如果搬迁过来的新居民‘不在乎 Sibudu 是祖先之地’，继续让他们维持这一传统可能会有难处，但并非不可能。”（鲁捷）

科学线人

全球科技政策新闻与解析

生化学家启动众包网站抵制错误数据



生化学家希望新因特网入口可以提高类似药物的小分子的信息，相关小分子可用于研究对人类健康和疾病至关重要的蛋白。

图片来源：LEN RUBENSTEIN

根据 7 月 21 日一个国际专家小组发布的报告，寻找设计和测试新药的生化学家经常被海量不合格数据困扰。这个由非营利机构、大学以及生物技术和制药公司等 46 家机构组成的专家组说，他们正在筹建一个类似旅行顾问网站的众包窗口，以解决他们认为处于问题核心的最新化学探针信息。

近年来，错误化学探针发展势头迅猛。这些类似药物的小分子主要用于阻止生物化学中决定特定蛋白角色的活动。在理念上，它可以帮助研究人员设计药物复方，该复方可执行类似功能，但却保留了成功制药的一些特征，比如无毒性和人体内可传输性。今天，已经存在着成千上万的类似化学探针。但是他们中大多数会和非目标蛋白结合，或是具有其他不想要的“非目标”作用。

“这已经成为一个真正意义上日益严重的问题。”生化学家、英国伦敦癌症研究所（ICR）执行主任 Paul Workman 说，他也是上述专家组的成员之一。Workman 和其他专家表示，要强调的是，许多化学探针会产生虚假的结果，把研究人员误导向他们研究的蛋白和药物分子的错误结论。

生化学家表示，他们希望网络众包窗口可以解决这个问题。在 ICR、博德研究院、结构基因组学联盟和惠康基金的帮助下，他们已经设立了一个名为化学探针窗口的网站。研究人员可以在该网站给不同的化学探针加注释，以确保同行获得所需要的最新对比信息。

然而，这一工作也会存在挑战，因为即便是利用同样探针的研究也经常会存在不同的使用条件和剂量，美国加州大学生化学家 Kevan Shokat 说。然而，他补充说，只要研究人员重复利用探针，稳定提高对不同探针的理解，就会对生化学家有所帮助。“我认为，这确实是一个好的服务系统。”Shokat 说。（鲁捷）

气候学家宣称全球变暖控制目标“高度危险”

一项新研究称，气候变化导致的负面反馈回路可能会致使海洋层化和冰层融化更加迅速。

图片来源：MARIUSZ KLUZNIAK

气候学家 James Hansen 发起了新一轮气候战役。在一份报纸中，Hansen 和同事警告称，当前限制全球变暖的国际计划远不足以避免诸如冰盖退化和海平面上升等气候灾难，Hansen 在一场新闻发布会上告诉一家媒体称，他希望相关报道可以影响今年 12 月在巴黎举行的全球气候谈判，鼓励谈判者重新思考各国把全球气候控制在不超过工业革命前 2 摄氏度的目标，一些科学家认为这个目标虽值得称赞，但却仍然不够。然而由于该报道出现的错误，Hansen 的言论究竟会产生哪些影响目前仍不清楚。

这项新研究有 66 页之多，包含了近 300 个参考书目，研究争论称，2 摄氏度的目标尽管在政治上来之不易，但该目标却仍然不够，事实上还“相当危险”。研究称，如果按照这一目标，会让大量冰盖融化，并形成负面反馈回路——导致更多冰盖融化和海平面上升。Hansen 和同事表示，更好的目标应该是向大气中排放 350ppm（百万分之一）二氧化碳。而目前相应的标准是 400ppm。

对此，一些科学家赞同类似的讨论是必要的。“通常，在气候变化风险讨论中，起始点都从这个假设开始，即当全球变暖超过 2 摄氏度就会给人类带来灾难。”美国宾夕法尼亚州立大学帕克分校气候学家 Michael Mann 说，“新报告值得表扬的一点是，提出即便是温度升高 2 摄氏度也极度危险。”

然而，对于该文章的发现结果，Mann 表示：“我对其中一些具体细节却持怀疑态度。”例如，他表示，该研究包含了一种可能性，即来自冰盖融化的淡水会随着时间推移呈指数增长，“这可能不太现实”。该研究还利用一个低分辨率海洋模型，该模型并不包括向高纬地带传送热量的关键湍流，如墨西哥湾流。（红枫）