一个样本都没测，博士生从二手数据中“ 榨 ”出新发现

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/39320.html

一个样本都没测，博士生从二手数据中“ 榨 ”出新发现。编译 张晴丹

年轻的科研人员容易陷入一个思维定式：想要做科研，必须先自己产生数据，数据等同于经费，经费等同于门槛。无数青年学者被困在这个闭环里寸步难行。

资金紧张、资源有限、人脉不足，这是许多人面临的共同困境。然而，有一座被忽视的“ 金矿 ”就藏在公共数据库中，等待着拥有新眼光的人去挖掘。

一位年轻的病毒学家Rhys Parry，无意间走上了一条非传统的科研之路。他没有花费巨额经费去测序，没有耗时数年去采样，而是靠着一台笔记本电脑和一双善于发现的眼睛，从全世界公开可的“ 二手数据 ”中“ 榨 ”出了令人惊喜的科学发现。

最终，Parry不仅发了论文，还拿下了国家级项目，走出了一条属于普通科研人的突围之路。近日，他在《自然》杂志的“ 职业专栏 ”发文，分享自己的经历。

对于所有渴望证明自己却苦于资源不足的科研人员来说，或许答案就藏在那些已经被储存却鲜被重访的数据之中。

3000份数据，一次意外的发现

2018年，在澳大利亚昆士兰大学的实验室里，博士生Parry正在与埃及伊蚊细胞系打交道。这是一种常见的蚊子细胞，用于研究蚊媒传播的疾病。然而，在一次常规的实验观察中，他发现了一个异常——细胞系中似乎存在着一种以前从未记录过的病毒。

这个发现本身并不惊人。事实上，昆虫细胞经常携带持续性的、未被注意的病毒感染，就像老房子里总会有一些不请自来的“住客”。但真正让这位年轻研究者感到好奇的是，这种新病毒无法感染哺乳动物细胞，而且出人意料的是，它竟然能适度降低登革热病毒的复制水平。

这个结果瞬间点亮了整个研究的意义。

登革热，一种由蚊子传播的急性传染病，每年威胁着全球数亿人的健康。如果有一种昆虫特异性病毒能够干扰登革热的传播，那将意味着什么？这极有可能成为人类理解甚至阻断蚊媒传播疾病的关键钥匙。

博导、分子病毒学家Sassan Asgari敏锐地捕捉到了这个发现的潜力，并鼓励Parry：“去查查我们实验室的其他数据集，把搜索范围扩大。”Asgari想知道，这种病毒在自家实验室及其他实验室的埃及伊蚊细胞中究竟有多普遍。

于是，一场跨越全球数据海洋的探险就此开始。

分析埃及伊蚊的现有数据集，帮助Rhys Parry鉴定出了一种新病毒。图源：James Gathany

幸运的是，全球各地从事蚊子研究的科研人员，早已将大量转录组数据共享。这些数据散落在不同的数据库中，像一颗颗无人捡拾的石头。Parry了大约3000个数据集，并日复一日地整理、对比

、分析，在海量信息中抽丝剥茧，一点点还原出这种病毒在全球范围内的分布与进化历史。

他没有飞往任何一个国家，没有采集一个样本，凭借3000个数据集、一台笔记本电脑，就完成了一次全球范围内的病毒流行病学调查。而这，仅仅是他在二手数据挖掘之路上的第一步。

从“ 旧数据 ”中发现新现象

博士生涯接近尾声时，一次偶然的机会，他在网上点开了昆士兰大学病毒学家Alexander Khromykh实验室已发表的数据集。Khromykh的研究方向是病毒感染期间细胞外囊泡中非编码RNA的作用，这是一个听起来相当小众的领域。

然而，就在这些已被分析过、发表过的数据中，Parry看到了一些“ 不对劲 ”的地方：病毒似乎在以一种前所未见的方式切割细胞RNA。

这不是原作者“ 漏掉 ”了什么。原作者与Parry的研究问题完全不同，他们的分析框架也完全不同。就像一个画家专注于画面的色彩，而一个建筑师却从中看到了结构。同样的数据，在不同的问题视角下，呈现出截然不同的面貌。

于是，Parry给Khromykh写了一封邮件介绍自己的新发现，并奏效了。一封邮件换来了一次交谈，然后促成了双方的一项合作。如今，他和Khromykh已经成为一个国家资助项目的共同研究者，而那个项目的基石正是从“ 旧数据 ”中发现的“ 新现象 ”。

Parry表示：“ 根据我的经验，大多数研究人员都很高兴看到他们的数据被这样使用。我发出的一些邮件促成了合作，另一些则让作者分享了原始发表论文中未包含的元数据。有时，原作者拥有你所不具备的样本或设备，能够以你无法实现的方式对结果进行验证；他们顺手做的一个小实验，或许就能证实某种关联，进而成为你下一份申请的初步数据。 ”

挖掘“ 二手数据 ”不是“ 次等科学 ”

这个故事的核心，指向一个让人惊讶的事实：海量的公共数据正在被闲置。

以美国国立卫生研究院下属的国家生物技术信息中心管理的序列读段档案库（SRA）为例，其拥有超过50PB的数据，而其中大部分在被存储之后，很少被再次使用。2022年，一个名为Serratus的项目将这些海量读数与病毒参考基因组进行比对，识别出数千个新的病毒序列，将已知RNA病毒的多样性扩展了一个数量级。

Parry强调，数千个新病毒从旧数据中发现。这些开创性的努力，展示了当人们真正重视并深耕二手数据分析时，它所撬动的可能性远超我们的想象。

这种模式并非基因组学领域独有。放眼整个科学界，临床试验数据集、生态学调查记录、医学影像档案……大量高质量资源都可在网上公开获取，正等待着被重新“ 拾取 ”。而绝大多数已发表的分析通常只是触及了数据所能揭示信息的表面，就像只读了小说的第一章，却以为知道了整个故事。

资助机构和出版商要求研究人员归档数据，初衷是为了确保结果的可重复性和可验证性。但归档数据的用途远不止于此。每个数据集都包含超出其生成者所发现之外范围的关联。Parry认为，

新方法会出现，新假设会涌现，研究领域的变迁可能让旧数据焕发新生。“ 我们有机会为现有数据带来新的视角，发现新的关联，并在理想情况下验证它们。 ”

最有趣的重新分析往往涉及整合不同类型的数据，比如蛋白质组学与转录组学，或者卫星图像与调查数据。Parry建议，从那些你理解其基础科学原理的数据集开始，但要能提出原作者未曾提出的问题。不过第一步永远是要检查元数据。如果需要费力挖掘才能理解相关系统、处理方式、时间点、重复实验以及实验平台，那么重新分析这些数据可能并不值得。

当然，并非所有数据集或所有分析都能产出新东西。Parry自己也承认：“ 我了数千个数据集，最终一无所获。 ”但搜索的成本很低，而阴性结果和阳性结果一样能提供信息。一次执行良好的二次分析可以发表、被引用，并作为初步数据使用，与任何其他科学产出地位相当。

为了表彰在严谨的二次数据分析领域作出杰出贡献的研究者，一个名为“ 研究寄生虫奖 ”的奖项诞生了——它由美国宾夕法尼亚大学支持并在2025年由GigaScience和GigaByte期刊赞助。但Parry认为“ 寄生虫 ”这个类比并不恰当。

关于“ 研究寄生虫奖 ”的介绍 图源：PSB

“ 当一名研究人员存入数据以支持可重复性，而另一个人利用这些数据发现了新东西，这不是剥削，而是科学在按预期方式运作。而且双方都受益：二次分析者发表了论文，而原始数据的生成者则获得了新的引用、潜在的合作者，以及其工作影响力的新证据。 ”Parry说。

当然，Parry也表示，他并不是建议用重新分析取代原始数据的生成。但对于那些无法获取人脉网络和大量资源的年轻科研人员来说，已经发表的数据是一座“ 金矿 ”，可以为他们提供一种以

极低甚至零成本来发表论文和申请资助生成数据的方法。“ 只需要一个问题、一台安装了R或Python编程语言的笔记本电脑，以及一双愿意重新审视旧数据的眼睛。 ”

参考链接：

https://www.nature.com/articles/d41586-026-00434-x

https://doi.org/10.1128/JVI.00224-18

作者：张晴丹 来源：科学网微信公众号