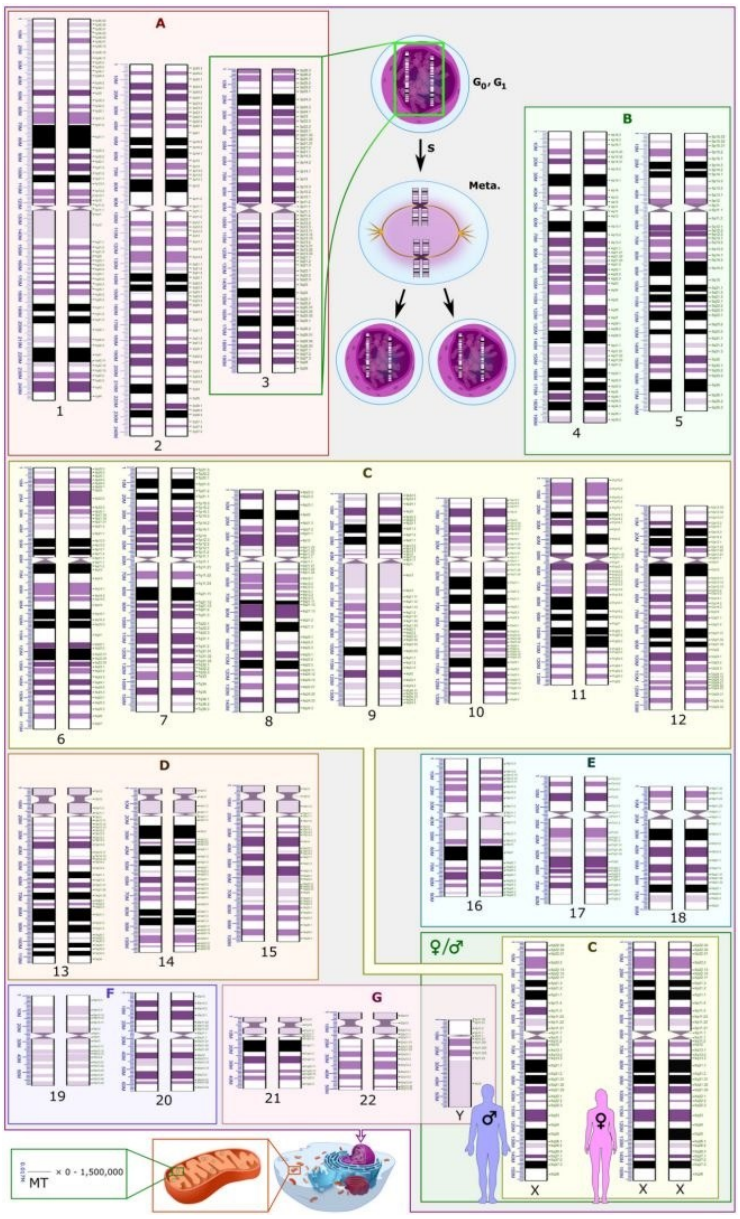
2025年9月19日 新世界时报
NEW WORLD TIMES

从“垃圾”到宝藏

——中国科学家揭开人类基因组中远古病毒的秘密



人类染色体的示意核型图，
展示了基于G显带技术的人类基因组概览

提到病毒，大多数人首先会想到疾病和感染。但你可能不知道，我们每个人的基因组中都携带着大量远古病毒的“化石”——它们占据了人类基因组的8%。这些被称为内源性逆转录病毒的序列，是数百万年前感染我们祖先的病毒留下的印记。

长期以来，科学家们认为这些序列只是基因组中的“垃圾”，没有任何功能。然而，最新的研究改变了这一看法。中国科学院上海药物研究所陈勋团队联合多个国际团队的一项突破性研究发现，这些远古病毒序列不仅不是“垃圾”，反而可能是调控人类基因表达的重要“开关”。

被误解的“垃圾”DNA——科学认知的转折

要理解本次发现的意义，我们需要先回到2000年，人类基因组计划初步完成，科学家们第一次看到了人类遗传密码的全貌。令人惊讶的是，编码蛋白质的基因只占基因组的不到2%，而剩下的98%被认为是没有功能的“垃圾DNA”。

在经典的G显带核型图（即通过吉姆萨染色技术呈现的染色体条带图谱）中，我们可以直观地看到人类基因组的组成奥秘。图中那些较暗的区域（DNA中鸟嘌呤和胞嘧啶碱基所占比例较少）和每条染色体中央狭窄的着丝粒（染色体中央的狭窄区域，在细胞分裂时起关键作用）区域，正是非编码DNA的主要聚集地。其中，内源性逆转录病毒序列就是这些“垃圾”的重要组成部分。

与普通病毒不同，逆转录病毒有一个独特的生命周期。它们携带的遗传物质是RNA，但在感染细胞后，会利用一种叫“逆转录酶”的特殊蛋白质，将自己的RNA“反向转录”成DNA。这就像是把一份手写的笔记（RNA）用打印机打印成标准文档（DNA）。

更关键的是，这些新合成的病毒DNA会整合到宿主细胞的染色体中，成为宿主基因组的一部分。这个过程就像是把一段外来的文字永久地粘贴到一本书中。HIV（人类免疫缺陷病毒）就是现代最著名的逆转录病毒例子。

人类基因组中这些病毒序列的来源可以追溯到几百万年前。当时，一些逆转录病毒感染了我们的祖先，并将自己的遗传物质整合到了宿主的基因组中。如果这种整合发生在生殖细胞中，病毒序列就会传递给后代。经过漫长的进化，这些病毒序列在突变和自然选择的作用下，大多数失去了原有的感染能力，成为了基因组中看似无用的“化石”。

科学家们根据序列差异，将这些内源性逆转录病毒分为500多个亚家族，包括HERV-E、HERV-K、HERV-H等。每个完整的病毒序列通常包含三个核心基因（gag、pol、env）和两端的长末端重复序列（LTRs）。但由于长期的进化，大多数序列已经支离破碎，看起来确实像是“垃圾”。

发现隐藏的宝藏——新方法带来新认识

然而，科学的魅力就在于不断挑战既有认知。近年来，越来越多的证据表明，这些所谓的“垃圾”DNA可能具有重要功能。特别是病毒序列两端的LTR区域，富含转录因子结合位点，可能作为基因调控元件影响邻近基因的表达。

但要验证这个假设面临一个巨大的挑战：由于这些病毒序列高度相似，就像同一本书的不同版本，传统的基于序列相似性的注释方法容易出错——可能把同一家族的序列误判为不同家族，或把不同家族的序列因局部相似而错误归类。

在基因组学中，所谓“注释”是指给DNA序列添加功能标签和分类信息的过程，就像给图书馆里的书籍贴标签分类一样。对于内源性逆转录病毒序列，注释就是标明每段序列属于哪个病毒家族或亚家族（如HERV-K、MER11-A等）。

而错误注释就像把一本经典文学小说放到了科技书架上，导致研究者在研究某个病毒家族功能时找到的是错误分类的序列，使实验结果混乱，无法得出正确结论。

为了解决这个问题，研究团队开发了一种全新的注释方法。这种方法不再单纯依赖序列相似性，而是结合了系统发育分析——通过追踪序列的进化历史来进行分类。这就像通过家谱来确定亲属关系，而不是仅凭长相相似。

研究人员首先聚焦于76个进化上较年轻的内源性逆转录病毒亚家族。令人震惊的是，他们发现其中26个亚家族存在近三分之一的注释错误。以MER11家族为例，原本被分为A、B、C三个亚家族，但新方法揭示了大量分类错误，并识别出了四个全新的亚家族：MER11_G1、G2、G3、G4，按照进化年龄从老到新排列。

从猴子到人：追踪病毒序列的进化轨迹

有了准确的分类，研究人员开始探索这些序列的功能。他们采用了一种名为“大规模平行报告系统”（lentiMPRA）的尖端技术。利用该方法，他们就像是同时进行了成千上万个实验，一次性测试了7000多条来自人类、大猩猩和猕猴的MER11序列，看它们是否能够调控基因表达。

实验在人类干细胞和早期神经细胞中进行，结果令人振奋。研究发现，最年轻的MER11_G4亚家族表现出强大的基因调控活性。更有趣的是，这种调控能力与一组特殊的DNA序列有关——SOX转录因子结合位点。

SOX转录因子是一类重要的基因调控蛋白，在胚胎发育、干细胞维持等关键生物学过程中发挥作用。研究发现，在灵长类进化过程中，MER11_G4序列通过单个碱基的缺失，意外地创造出了新的SOX结合位点。这个微小的变化，却带来了巨大的功能影响——显著增强了这些序列的调控活性。

更令人惊奇的是，这种进化在不同物种中呈现出不同的模式。人类和黑猩猩共有的一些MER11_G4序列获得了独特的突变，使它们在干细胞中具有更强的调控潜力。这意味着，这些远古病毒序列可能参与了人类特有的基因调控网络的形成。

单碱基的蝴蝶效应：微小变化的巨大影响

研究团队在单碱基分辨率水平上分析了这些变化。他们发现，仅仅一个碱基的插入或缺失，就能决定一个序列是否具有调控功能。这就像密码锁，只要一个数字错误，就无法打开。但在进化的长河中，偶然的“错误”反而可能创造出新的功能。

以SOX结合位点为例，原始的MER11序列并不包含这个位点。但在灵长类进化过程中，一个碱基的缺失意外地形成了SOX蛋白的识别序列。这个变化发生在人类和大猩猩的共同祖先中，距今约800万年。随后，在人类和黑猩猩的演化过程中，这些序列又积累了更多的变化，进一步优化了它们的调控功能。

这种现象展示了进化的精妙之处：看似随机的突变，在自然选择的作用下，可能被保留并赋予新的功能。远古病毒序列就这样从“入侵者”变成了“合作者”，成为人类基因组调控网络的一部分。

从基础研究到医学应用：打开新的大门

这项研究的意义远不止于满足科学好奇心。内源性逆转录病毒序列与多种人类疾病相关，包括癌症、自身免疫疾病和神经退行性疾病。准确了解这些序列的功能，对于理解疾病机制和开发新疗法至关重要。

例如，某些内源性逆转录病毒在肿瘤中异常激活，可能促进癌细胞的生长和转移。如果我们能够精确识别这些序列并理解它们的调控机制，就可能开发出新的癌症治疗策略。同样，在自身免疫疾病中，某些病毒序列的激活可能触发免疫反应，导致机体攻击自身组织。而若能精准干预这些序列的异常表达，或许能为此类疾病的治疗提供另一种思路。

此外，这项研究还为理解人类进化提供了新视角。人类与其他灵长类动物的基因组高度相似，但在认知能力、语言等方面存在巨大差异。这些差异的遗传基础一直是科学界的重大谜题。内源性逆转录病毒序列的物种特异性进化，可能是造成这些差异的重要因素之一。

研究团队的下一步计划是结合人工智能技术，全面解析内源性逆转录病毒的功能。通过机器学习算法，他们希望能够预测哪些病毒序列具有调控功能，以及它们在不同细胞类型和发育阶段的作用。这将为精准医疗和个性化治疗提供新的靶点。

同时，这种基于进化的研究方法也可以应用于其他领域。例如，研究流感病毒的突变模式，预测下一次流感大流行的可能性；或者分析肿瘤细胞的进化，开发更有效的抗癌策略。

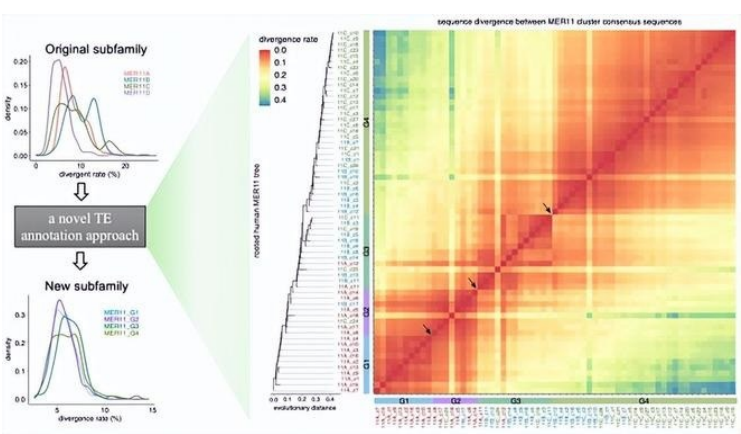
与病毒共舞的生命之歌

回顾生命演化的历史，病毒与宿主的关系远比我们想象的复杂。它们不仅是致病的“敌人”，也可能成为进化的“盟友”。内源性逆转录病毒序列就是这种复杂关系的见证——曾经的入侵者，如今成为我们基因组不可分割的一部分，甚至可能塑造了人类独特的生物学特征。

这个发现也让我们重新思考“自我”的定义。如果我们的基因组中8%来自远古病毒，那么什么才是真正的“人类基因”？也许，正是这种基因的“混血”，造就了生命的多样性和复杂性。

站在科学的前沿，我们看到的不是一个简单的黑白世界。今天的“垃圾”DNA，明天可能就是治愈疾病的关键；今天的有害病毒，明天可能成为基因治疗的工具。保持开放的心态，不断探索未知，这正是科学精神的真谛。

在基因组这部生命之书中，每一个序列都可能隐藏着进化的秘密。内源性逆转录病毒的故事告诉我们：生命的复杂性远超我们的想象，而探索这种复杂性的旅程，才刚刚开始。



基于进化的转座子序列注释新方法

不同物种的基因组规模