# 用于在DNA上存储渐进式图像的自适应采样

Xavier Pic*、Nimesh Pinnamaneni**、Raja Appuswamy*

* EURECOM，数据科学系，法国索菲亚-安提波利斯** Helixworks Technologies 有限公司，爱尔兰科克 Bishopstown，MTU校区

arXiv:2603.05220v1 [eess.IV] 5 Mar 2026

**摘要**

传统数据存储介质的使用寿命较短，加之存储需求呈指数级增长，使长期归档成为数据存储行业乃至更广泛领域的一项根本性难题。因此，研究人员正在寻找创新的数据存储介质解决方案，以极低的成本长期保存数据。DNA分子具有高密度、长寿命和低能耗等特点，已成为数字数据归档的一种可行替代方案。然而，当前的DNA数据存储技术在成本和可靠性方面仍面临挑战。因此，编码率和错误鲁棒性对于扩大DNA存储规模，并使其在技术和经济层面切实可行至关重要。此外，不同文件对应的DNA分子通常位于同一个寡核苷酸池中。如果没有寡核苷酸层级的随机访问方案，从这些混合池中解码特定文件将非常不切实际，因为必须先对所有寡核苷酸进行测序和解码，才能检索目标文件，这会大幅增加读取成本。

本文介绍了一种将图像高效编码并存储到DNA分子中的方案，旨在降低读取图像低分辨率版本所需的读取成本。该图像存储系统基于JPEG2000编解码器的渐进式解码功能，但也可适用于任何传统的渐进式编解码器。每个分辨率层均使用JPEG DNA VM编解码器编码为一组寡核苷酸；这是一种基于DNA的编码器，旨在以高可靠性还原文件。根据所需读取的分辨率，可相应调整寡核苷酸集合，以及需要进行测序和解码的寡核苷酸部分。测序时，借助纳米孔测序仪提供的自适应采样方法选择这些寡核苷酸，从而实现无PCR随机访问。

索引词— DNA数据存储、JPEG 2000、渐进式、随机访问、JPEGDNAVM

## 1. 引言

数据存储行业正面临前所未有的挑战，因为存储需求仍在以指数速度持续增长。传统存储设备存在耐久性和存储密度方面的根本性限制，使得长期数据存储难以实现。合成DNA分子已成为一种应对这些挑战的替代方案，主要得益于其超长寿命、高密度和低能耗。 [1],

将数据存储到DNA分子中的经典方案由生化过程和计算过程组成。两种主要的生化过程是合成，即创建具有所需核苷酸序列的DNA分子；以及测序，即获取其中所含的核苷酸序列。主要的计算过程是适配DNA的编解码器，它会将数据转换为四元序列以供合成，并在测序过程中将其转换回来。编解码器还会压缩数据，以降低合成和测序成本。以往的DNA存储研究着重于优化生化过程和计算过程，旨在使DNA成为可靠的数字存储介质，尽管读写流程中的多个环节都会出现错误。然而，现有的DNA存储方案并未设计为支持一种重要的访问模式：将图像存储在DNA中时触发的解码图像分辨率自适应选择。

自适应分辨率选择适用于多种场景。例如，根据用于查看图像的客户端，社交媒体应用可能会选择在资源受限的移动设备上显示低分辨率版本的图像，而在台式机上显示高分辨率版本。要有效支持自适应分辨率选择，需要采用高效的图像编码技术来压缩图像，以实现渐进式解码；还需要底层存储介质支持随机访问，从而能够根据分辨率要求检索图像的特定部分。以往的DNA数据存储研究已展示了使用聚合酶链式反应（PCR）实现此类随机访问的方法。另一种基于纳米孔测序仪Read Until功能的替代方案，则有助于检索具有特定参考序列的寡核苷酸，而无需进行特定引物的PCR扩增。目前，此类随机访问仅应用于任意二进制文件，尚未用于图像存储场景中实现解码分辨率自适应选择等功能。

本文提出了一种新颖的方法，将渐进式图像编码与纳米孔测序仪的 Read Until功能相结合，以支持对存储在DNA中的大型图像集合自适应选择分辨率，且读取成本远低于现有方案。

本研究获得欧盟“地平线”研究与创新计划项目 GLACIATION（资助编号：101070141）以及 ANR PEPR 计划 MOLECU-LARXIV 的资助

## 2. 背景

### 2.1. DNA适配编码

过去几年，DNA数据存储领域的研究活动迅速发展。本节概述几种开创性方法，既包括DNA存储的一般情况，也特别介绍将图像存储在DNA中的方法。近期综述 [2] 对使用DNA存储通用二进制数据的各种方法进行了详细比较。2012年，Church等人 [1], 提出了一种方法，可将数据大规模编码至合成DNA分子中，并从中解码。该论文不仅给出了将数据编码至DNA分子并进行解码的方法，还指出了将数据编码至DNA分子时必须遵守的主要编码约束。2013年，Goldman等人[3] 提出了首批能够遵守其中部分约束的编码器之一。该熵编码器能够将任意文件编码为类似DNA的数据。2015年，Grass等人[4] 首次在DNA数据存储方案中引入纠错码。该纠错机制旨在提高整个存储过程的鲁棒性，以抵御生化操作所导致的错误（替换、插入、删除）。此后，其他研究通过为二进制数据添加冗余来检测和纠正错误，并提出了其他纠错方案 [5, 6, 7, 8, 9, 10, 11], 。最后，人们还研究了随机访问方案，旨在实现对存储于DNA中的二进制数据进行选择性访问并降低读取成本 [12, 13]。

除了用于编码通用二进制数据的方案外，DNA适配图像编解码器也应运而生。早期方案之一是，Di mopoulou等人[14]开发了一种适用于DNA数据存储的基于JPEG的图像编码器。同样，在 [15], 中，作者介绍了一种基于JPEG的DNA适配编码方案。Pic等人采用改进的DNA适配熵编码器[16] ，提升了这种基于JPEG的DNA适配图像编码器的性能。Lazzarotto等人[17] 开发了JPEG DNA VM软件，该软件采用基于Raptor码[18]的系统对数据进行编码。该软件由JPEG DNA特别工作组发布，作为后续开发的验证模型。此外，人们也研究了基于学习的图像压缩方法 [19, 20], ，其中作者使用变分自编码器将图像压缩到潜在空间中，再将其编码到DNA中。在 [21], 中，作者利用像素域表示，以更好的压缩性能重建图像。

### 2.2. DNADataStorage 中的随机访问

大多数随机访问方法都需要加入唯一的地址序列或物理把手，以便选择性识别和分离目标DNA链。这些寻址方案必须满足若干关键要求：具备正交性以防止交叉反应、在操作条件下保持热力学稳定性，并与后续测序和解码过程兼容 [12, 22, 23, 24, 25, 26]。

### 2.3. 基于PCR的随机访问

基于PCR的随机访问是DNA存储系统中首个可扩展的选择性数据检索方案。该方法通过在寡核苷酸上添加引物序列，利用PCR选择性扩增目标文件。华盛顿大学与微软合作展示了首个大规模实现，成功从含有超过1,300万条DNA寡核苷酸的池中检索出单个文件；该池编码了200兆字节的数据 [12]。

该系统架构要求精心设计引物，以确保正交性并防止非特异性扩增[12, 25, 27, 28]。引物库必须满足严格的热力学约束，包括将GC含量控制在40–60%、退火温度高于58°C，并尽量减少二级结构的形成 [25, 27]。

基于PCR的系统展现出出色的可扩展性，理论上可支持 $8.98 \times 10^{21}$ 个可寻址目标，系统容量可达65.8 ZB [22]。然而，若干限制因素制约了其实际应用。扩增过程本质上具有破坏性，每次访问操作都需要完全消耗原始DNA池 [12]。由于需要进行热循环，访问时间长达数小时；此外，为克服可能导致不同序列表示失衡的扩增偏倚，还需采用乳液PCR [12, 22] 和微胶囊化技术等复杂流程 [29]。此外，在寡核苷酸中使用引物序列，尤其是嵌套引物或分层引物 [12, 22] ，需要占用20–60 nt，减少了200 nt寡核苷酸中可分配给数据载荷的核苷酸数量，从而增加了合成成本。

### 2.4. 带表面标记的二氧化硅封装

二氧化硅封装方法通过将DNA物理保护在不透性二氧化硅胶囊内，解决了基于PCR的系统的若干局限；这些胶囊表面标记有与地址或文件元数据相对应的单链DNA条形码 [30]。与PCR检索不同，基于FACS的二氧化硅封装DNA分选可在无需扩增的情况下对目标文件进行非破坏性选择，避免偏差，并将地址标签与载荷链解耦，从而提高净数据密度。

封装过程利用溶胶-凝胶化学在DNA分子周围形成稳定的二氧化硅基质 [30]。实验研究表明，封装在二氧化硅中的DNA在相当于中欧2000年的加速老化后仍保持结构完整性并可测序 [4]。二氧化硅基质可防止紫外线辐射、氧化和水解造成的降解 [31]。

荧光激活细胞分选（FACS）能够从复杂混合物中精确筛选目标胶囊 [30]。该系统在每个光学通道中可实现每 $10^6$ 个文件中筛选出一个文件的灵敏度，并且能够扩展至 $10^6 N$ 个文件；常见的商用 FAS 系统最多可提供 $N = 17$个光学通道 [30]。FACS 通过检测

与二氧化硅胶囊表面条形码序列杂交的带荧光标记的寡核苷酸探针，实现快速识别并物理分离目标胶囊，同时使其余样本池保持完整，以便日后访问 [32]。

封装可在重复查询中保持DNA池的完整性，而二氧化硅基质则可提供长期环境保护，抵御水解、氧化和紫外线损伤 [30, 4]。然而，这种方法需要复杂的溶胶-凝胶封装和表面标记流程；在测序前，还必须经过化学脱保护步骤才能从胶囊中释放DNA，因此会引入延迟；此外，它还依赖FACS等复杂且昂贵的仪器。除此之外，通量可能受到荧光探针杂交动力学和分选速度的限制，而设计能够避免大型文库中交叉反应性的探针，则会进一步增加复杂性和成本 [30, 31, 32]。

### 2.5. 使用纳米孔的自适应采样

自适应采样利用纳米孔测序的实时控制能力，实现无需PCR的随机访问。当每条DNA链开始易位时，经过缓冲的电流信号会进行碱基识别，并与用户提供的地址或元数据序列进行比对；如果检测到匹配，则继续测序；否则，电压反转会将该DNA链排出并送回DNA池。Roman Sokolovskii等人证明，该方法可以从异质DNA池中选择性富集目标子集，在运行过程中动态切换目标，并采用文库回收方案，使非目标DNA链能够留待未来检索 [33]。其核心机制要求对大约 $\sim$ 400nt的信号进行缓冲和处理，并加上碱基识别与比对延迟（总计800nt），之后才能作出保留或排出的决定 [34], ；这意味着需要足够长的DNA构建体（$\geq$ 1000–1500nt），才能针对较短的地址基序实现可靠的早期比对。

与基于PCR的随机访问相比，自适应采样可从根本上避免扩增偏倚和池耗竭；而自适应采样会完整地弹出不匹配的链，从而保留文库多样性并实现重复利用。运行过程中动态切换目标，无需额外的制备步骤，即可从单份等分样品中检索多个文件 [33]。与结合基于FACS的分选的二氧化硅封装相比，自适应采样无需采用溶胶-凝胶封装流程、表面标记、化学脱保护及专用分选仪器；它通过测序过程中的读段内比对进行在线筛选，简化操作并减少制备开销。尽管在未预先富集的情况下，对于大型样本池中的极低丰度目标，自适应采样的效率会有所下降，而且其对长构建体的需求需要通过组装或类似方法解决，但其非破坏性、动态且无需扩增的特点，使其成为基于PCR和二氧化硅的随机访问方法的有力替代方案或补充手段。

## 3. 提出的方法

### 3.1. 寡核苷酸结构与载荷

随机访问的自适应采样方法依赖于参考序列，这些序列位于寡核苷酸的起始位置。这些序列用作后续长载荷的索引，长载荷连接在参考序列末端（图 1）。该长载荷通过合成并利用接头连接较短的寡核苷酸构建而成。

我们的图像编码方案基于工作于渐进编码模式的JPEG2000编解码器。该编解码器输出一个比特流，按 $N_{levels}$ 大小递增的分辨率层组织。此比特流被切分为一组 $N_{levels}$ 二进制文件，分别对应不同的层。每个分辨率二进制文件都使用JPEG DNA VM编解码器编码为一池适配DNA的短载荷寡核苷酸。随后，每一池寡核苷酸将分别进行合成和制备，以供自适应采样。在此设计中，前文提到的参考序列用于索引特定的图像分辨率层。

### 3.2. 参考序列与数据集准备

每种分辨率所用的参考序列是在制备寡核苷酸时设计的。我们构建了一个参考序列字典，将图像的每个分辨率级别与制备用于该级别寡核苷酸池时所用的参考序列关联起来。数据集中所有图像的所有分辨率级别制备完成后，即可将它们合并为一个寡核苷酸总池，也就是该数据集。对于每幅图像的每个分辨率层，字典中都会存储一个参考序列，并在图像解码时提供该序列。需要注意的是，如果将 [35]中介绍的基于缩略图的访问系统改造用于自适应采样，则可以在每个缩略图寡核苷酸末端附加一段较短的DNA序列，并将其用作设计参考序列的种子，从而无需在外部存储这些序列。

### 3.3. 无需 PCR 的随机访问与图像检索

过去，人们会采用渐进式解码来缓解网络连接不佳给图像交换带来的限制：在数据传输过程中，可以逐步解码最初的几层，从而获得目标图像的低质量预览。同样，在DNA数据存储中，也可以利用这种渐进式解码技术来降低测序成本。通过在参考序列字典中查找特定图像不同分辨率层对应的序列，可以从通用寡核苷酸池中逐步提取图像。提取图像的不同层时，从较低分辨率开始。在测序过程中，图像编解码器会依次将不同的参考序列输入纳米孔测序仪，以更新正在运行的Read Until过程的参数。当前输入Read Until过程的参考序列将使测序仪仅对以该输入序列开头的寡核苷酸进行测序，并将其他寡核苷酸拒绝送回通用寡核苷酸




**图 1**。自适应采样测序模式的寡核苷酸结构

池中。通用图像编解码器（图2）用于逐步选择、测序和解码这些不同的编码层，从而逐渐提升最终目标图像的质量。当编解码器输出的分辨率质量达到令人满意的程度时，可以向测序仪发送提前停止指令，以避免测序不必要的数据。

## 4. 实验结果

所提出的编解码器主要旨在降低从寡核苷酸池中检索图像所需的成本。衡量这一成本的第一步，是确定图像检索的实际使用场景。对于在寡核苷酸层级不实现任何随机访问的 DNA 编码系统，读取图像的成本等于读取整个数据集所需的寡核苷酸读取次数。对于这类编码器，公式 (1) 可表示检索图像所需的读取成本。另一方面，将随机访问或渐进式解码集成到核心编码系统中的图像检索过程，可以降低公式 (1) 中的读取成本。首先，如果启用了渐进式解码，读取成本可限制为公式 (2)，因为只需解码前几层。

$$R_c(I,K) = \frac{\sum_{i=0}^{N_{images}} \sum_{k=0}^{N_{levels}} nucs(i,k)}{input_image_pixels} \quad (1)$$

$$R_{c_pd}(I,K) = \frac{\sum_{i=0}^{N_{images}} \sum_{k=0}^{K} nucs(i,k)}{input_image_pixels} \quad (2)$$

数值 $nucs(i,\ k)$ 表示为解码一层而需要测序的核苷酸数量：

$$nucs(i,k) = coverage(i,k) \times number_oligos(i,k) \quad (3)$$

由此，我们可以定义读取成本增益，用于衡量渐进式解码方法带来的改进：

$$G_{pd}(I,K) = \frac{R_c(I,K)}{R_{c_pd}(I,K)} \quad (4)$$

### 4.1. 性能评估

任何DNA编码方法的性能都可以通过一系列指标进行评估，例如率失真曲线、读取成本和写入成本。由于我们的工作主要关注渐进式解码（PD），因此研究中重点关注的主要指标是读取成本。我们尤其研究给定编码图像在逐层读取每个分辨率层时读取成本的变化。图像被编码为一系列分辨率层。每个分辨率层都将图像在每个维度上的大小缩小为原来的 1/2（总计缩小为 1/4）。此处呈现的结果来自对 Kodak[1] 数据集中的 5 幅图像进行编码，每幅图像包含 3 个分辨率

[1]https://r0k.us/graphics/kodak/

| 层 | $L_0$ | $L_1$ | $L_2$ |
|---|---|---|---|
| # 寡核苷酸 | 802 | 2229 | 6208 |
| 理论 $G_{pd}$ 读取成本增益 | 7.74 | 5.71 | 3.30 |

**表 1**。$G_{pd}$ 各目标分辨率级别的平均读取成本增益 $L_k$，为所选图像的平均值。

级别。寡核苷酸的数据块长度为 148。在合成过程中，这些数据块将连接在一起，并向每个连接后的分子中添加特定于图像和分辨率级别的参考序列，以便通过自适应采样实现寡核苷酸随机访问。我们观察到了理论上的读取成本增益（表 1），该增益体现了渐进式解码相较于类似编码方法所带来的改进；后者不提供任何自适应读取成本降低机制，例如分辨率选择。

表 1 所示的结果中，我们测量了解码每幅图像直至某一分辨率级别所需的寡核苷酸数量，并对所有图像的结果取平均。在这些条件下，仅针对初始层时，渐进式解码器带来的增益最高可达 7.5×。如果针对更多层，这一增益会迅速降低；如果针对所有层，则无法从渐进式解码中获得任何增益，因为所有寡核苷酸都需要进行测序。不过，这些增益在很大程度上取决于所选分辨率层的尺寸：较小的层能带来更大的增益，但图像失真也会更严重。

获得的增益与通过调整 JPEG DNA VM 和 JPEG2000 的编码选项所能实现的编码性能和读取成本改进相互独立。JPEG DNA VM 软件参数可以针对每个分辨率层分别调整，尤其是冗余参数，从而更好地保护较低分辨率层，使其免受错误影响。

## 5. 结论

本文介绍了一种适用于DNA数据存储的渐进式图像编解码器，通过渐进式解码和基于缩略图的随机访问，降低了检索图像所需的读取成本。这种新型编解码器使我们能够利用读取成本优势：与不采用渐进式解码或基于缩略图的随机访问相比，读取成本可降低数倍。这些理论结果促使我们订购了这些寡核苷酸的合成，并计划在未来几个月开展实际湿实验。

## 6. 参考文献

[1] G. M. Church等人，“DNA中的下一代数字信息存储”，《科学》，第337卷，第6102期，2012年。[2] T. Heinis等人，“DNA信息编码技术综述”，《ACM计算机评论》，第56卷，2023年11月。[3] N. Goldman等人，“迈向合成DNA中实用、高容量、低维护的信息存储”，《自然》，2013年。[4] R. N. Grass等人，“利用纠错码在二氧化硅中的DNA上稳健地化学保存数字信息”，《应用化学国际版》，2015年。

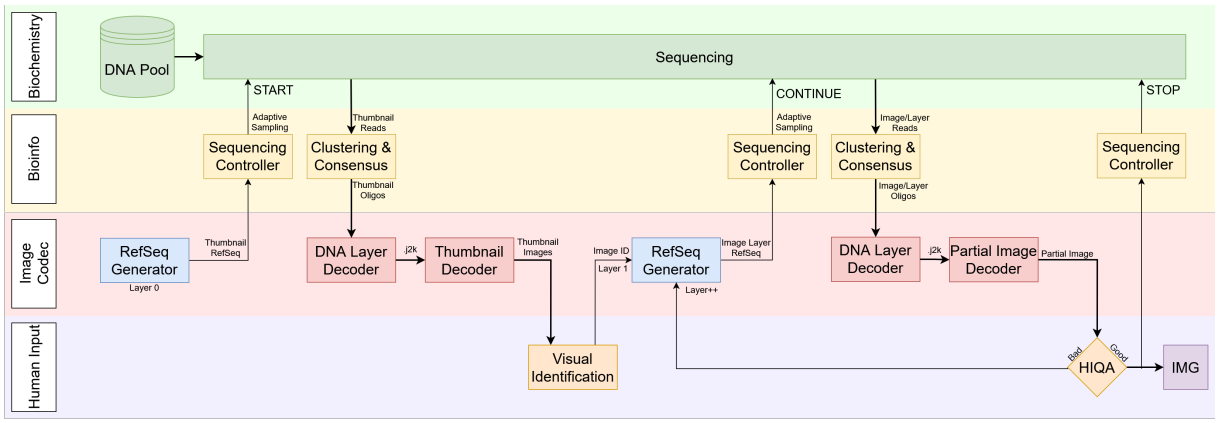


**图2**。基于缩略图的自适应采样DNA图像编解码器解码流程


[5] Y. Erlich和D. Zielinski，“DNA喷泉码实现稳健高效的存储架构”，《科学》，2017年。[6] S. M. H.T. Yazdi等人，“便携式、无错误的基于DNA的数据存储”，《科学报告》，第7卷，第1期，第1–6页，2017年。[7] M. Welzel等人，“DNA-Aeon为DNA存储提供灵活的算术编码，以满足约束并实现纠错”，《自然·通讯》，2023年。[8] M. Blawat等人，“DNA数据存储的前向纠错”，《计算机科学 Procedia》，2016年。[9] Ben C.等人，“GCNSA：采用图卷积网络和自注意力的DNA存储编码”，《iScience》，第26卷，第3期，第106231页，2023年。[10] Ben Cao等人，“高存储密度、低覆盖率DNA存储的自适应编码”，《NPJ系统生物学与应用》，第8卷，2022年。[11] O. Sabary等人，“DNA存储系统的重建算法”，《科学报告》，第14卷，2024年。[12] L. Organick等人，“大规模DNA数据存储中的随机访问”，《自然·生物技术》，2018年。[13] E. Marinelli等人，“CMOSS：可靠的基于基序的柱状分子存储系统”，载于《第17届ACM国际系统与存储会议论文集》，2024年。[14] M. Dimopoulou等人，“用于DNA数据存储的基于JPEG的图像编码方案”，欧洲信号处理会议，2021年。[15] B. Li等人，“IMG-DNA：用于图像的近似DNA存储”，载于《第14届ACM系统与存储国际会议论文集》，美国纽约州纽约，2021年，SYSTOR '21，计算机协会。[16] X. Pic和M. Antonini，“用于合成DNA图像存储的受约束的香农-范诺熵编码器”，欧洲信号处理会议（EUSIPCO），2022年。[17] D. Lazzarotto等人，“洛桑联邦理工学院参与JPEG DNA征稿启事的技术说明”，arXiv预印本，2023年。[18] A. Shokrollahi，“Raptor码”，《IEEE信息论汇刊》，第52卷，第2551–2567页，2006年6月。[19] Y. Zheng等人，“DNA-QLC：一种高效可靠的DNA存储图像编码方案”，《BMC基因组学》，2024年。[20] G. Franzese等人，“生成式DNA：基于DNA的近似图像存储的表征学习”，载于《2021年国际视觉通信与图像处理会议（VCIP）》，2021年，第01–05页。

[21] S. Seo 等人，“利用有损压缩提高DNA数据存储的信息密度”，2024年。[22] K. J. Tomek 等人，“推动基于DNA的信息存储系统实现可扩展性”，《ACS合成生物学》，2019年。[23] A. El-Shaikh 等人，“DNA存储系统中的大规模随机访问”，《NAR基因组学与生物信息学》，第4卷，2022年。[24] K. N. Lin 等人，“动态且可扩展的基于DNA的信息存储”，《自然·通讯》，第11卷，2020年。[25] S. K. Subramanian 等人，“一组经实验验证、可用于组合指定遗传组分的相互正交引物”，《合成生物学》（牛津），2018年。[26] G. Gowri 等人，“正交DNA条形码库的可扩展设计”，《自然·计算科学》，第4卷，2024年。[27] Qiang Z. 等人，“一种用于热力学设计DNA序列的新约束条件”，《公共科学图书馆·综合》，第8卷，2013年。[28] A. S. Booeshaghi 等人，“量化用于序列普查分析的正交条形码”，《生物信息学进展》，第4卷，第1期，第vbad181页，2023年。[29] B. B. B¨ogels 等人，“热响应微胶囊中的DNA存储：实现反复随机多重数据访问”，《自然·纳米技术》，第18卷，第8期，第912–921页，2023年。[30] J. L. Banal 等人，“利用布尔搜索在档案文件存储系统中实现DNA存储器随机访问”，《自然·材料》，第20卷，第1272–1280页，2021年。[31] D. Kapusuz 和 C. Durucan，“探究DNA和单核苷酸在溶胶-凝胶衍生二氧化硅中的封装机制”，《生物材料应用杂志》，第32卷，2017年。[32]Bio-Rad，“荧光激活细胞分选（FACS）”，https://www.bio-rad.com/en-ie/feature/fluorescence-activated-cell-sorting.html，访问日期：2025-06-20。[33] R. Sokolovskii 等人，“用于DNA数据存储无PCR随机访问的纳米孔测序自适应采样”，bioRxiv，2024年。[34]牛津纳米孔科技，“自适应采样”，https://nanoporetech.com/document/ adaptive-sampling#introduction-advanced，访问日期：2025-06-20。[35] X. Pic 和 R. Appuswamy，“结合渐进式图像压缩与DNA数据存储随机访问”，第25届国际数字信号处理会议，2025年。