“阴阳”编解码系统实现信息高效存储

华大智造DNBSEQ测序平台赋能新型DNA数据存储

“阴阳”编解码系统是基于华大智造DNBSEQ测序技术开创的一套独具优势的编解码系统，用以解决当前DNA信息存储领域的技术难题[1]。该研究是由深圳华大生命科学研究院主导，深圳国家基因库、首都师范大学、美国哈佛大学等多个研究团队共同参与。

具体研究成果于2022年4月25日发表于《自然-计算科学》（*Nature Computational Science*），题为“Towards practical and robust DNA-based data archiving using the yin–yang codec system”。

推荐应用：前沿技术 -DNA 存储技术
推荐机型：DNBSEQ-T7RS

- **测序数据质量高**
  DNBSEQ™ 测序技术具有高准确性、低重复序列率、低标签跳跃等的重要特性。

- **信息恢复稳定性高**
  “阴阳”编解码在保证高信息密度、高信息转换效率和高技术兼容性的同时，大幅提高信息恢复的稳定性。

背景介绍

DNA 存储技术是指利用 DNA 的分子结构来进行数据存储，与传统信息存储的“信息写入 - 保存 - 读取”步骤类似，DNA 存储流程图如图 1 所示。我们正处在前所未有的信息大爆炸时代，传统的标准存储介质已不能满足指数级增长的数据存储需求，DNA 作为生物体内古老而高效的信息载体凭借其独家的超高信息密度、超长待机时间以及超强生物兼容性的天然优势，在信息密度、复制与维护成本、使用寿命等方面都具有颠覆现有技术的巨大潜能 [2-4]。

DNA 存储的编解码是 DNA 存储中最重要的环节之一，不仅决定信息转换的效率（信息密度），还直接影响存储信息的稳定性及可靠恢复性。

从 2012 年起，DNA 存储技术的发展主要聚焦于提升信息密度，而技术兼容性和原始信息的稳定恢复方面的考虑尚不全面。2017 年以前，编解码技术都未能实现完全的技术兼容。2017 年，美国哥伦比亚大学研究团队开发的 DNA 喷泉码几乎解决了此前的技术瓶颈 [5]，但实际应用中也出现灵活性与适用性的问题。因此，如何在保证信息转换效率和技术兼容的同时大幅提高信息恢复的稳定性成为实现 DNA 存储的一大难题。

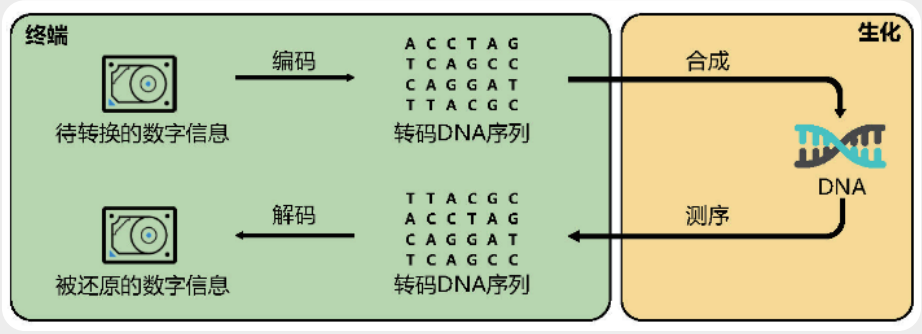

图1.DNA存储流程

研究描述

基于华大智造 DNBSEQ 测序平台开创的“阴阳”编解码系统，巧妙地将起源于几千年前的中华“阴阳”思想应用到 DNA 编解码系统当中，以两套不同的规则，分别对两条二进制信息进行“一对一”编译转换，再取两者统一交集的部分为最终解，实现将两条独立的信息组合统一为一串 DNA 序列，其“阴阳”编解码规则示意图如图 2 所示。经验证，该系统在信息密度、技术兼容性、数据恢复稳定性等多方面的具有很大的优势。

实验方法

使用 YYC 编解码和 DNA 喷泉码对文件进行编码

使用YYC编解码中第888种编码规则将文件的二进制信息转换为DNA序列，用“YYC-screener”选择可行的DNA序列，最后得到一个包含10,103个单链200 nt DNA序列的寡核苷酸池。

对于DNA喷泉码，除了冗余之外，使用原始报告中的默认参数设置生成DNA寡核苷酸文库，确定文件解码的最小冗余。最后得到一个编码.tar归档压缩文件（9,185个序列）的寡核苷酸库和一个编码混合的三个单独文件（10,976个序列）的寡核苷酸库。

文库制备与测序

三个寡核苷酸池由Twist Biosciences外包合成，并以DNA粉末的形式交付用于测序。

对于体内存储时，首先将DNA片段分割成具有重叠区域的亚片段，然后进一步分割成构建块。

使用Q5高保真DNA聚合酶将 80 nt 寡核苷酸组装到块上并克隆到载体中进行桑格测序。通过酶切载体释放测序验证的块，PCR扩增，凝胶纯化，最后利用酵母的天然同源重组，将完整组装的片段整合到酵母II染色体中的YBR150C基因上。通过带有的LEU2标记筛选阳性酵母菌落，用于基因组DNA提取和测序。

寡核苷酸池的文库制备：将DNA粉末溶解在双蒸水（ddH_2O）中得到标准溶液，将标准溶液连续稀释10倍，得到七种工作溶液（WS6-WS0），平均浓度依次为每微升10^6至10^0个DNA分子。接下来，每个工作溶液通过PCR扩增获得P2的扩增产物以及P1和P3的三个不同文件中的每一个扩增产物。使用凝胶电泳和Qubit荧光计测量产物的浓度。使用DNBSEQ-T7测序仪对所有扩增的DNA文库进行测序。

酵母基因组DNA制备及文库制备：将酵母细胞在YPD培养基中培养两天，离心收集颗粒。添加破菌缓冲液重悬沉淀。加入玻璃珠和PCI，涡旋混匀，离心。然后将水层转移至试管中，加入异丙醇沉淀基因组DNA，静置，离心，最后将基因组DNA重悬于TE缓冲液中。利用Covaris仪器片段化基因组DNA，随后进行末端修复，利用磁珠法进行片段筛选，然后在3‘端加入A尾，连接带有标签的双端测序衔头，最后利用PCR富集连接后的产物。使用MGISEQ-2000和DNBSEQ-E5测序平台对样品进行测序。

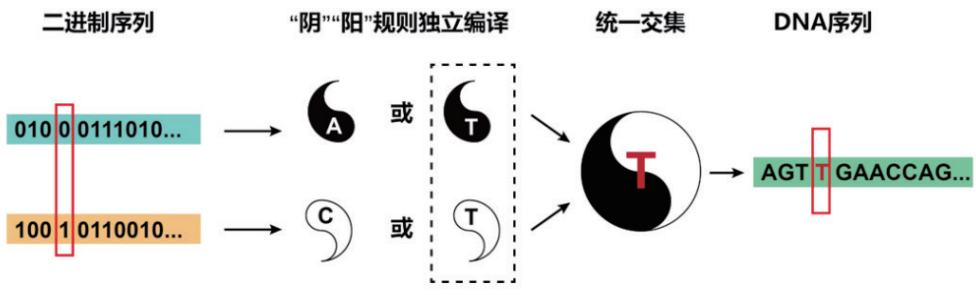

图2.“阴阳”编解码规则的原理

数据分析

总共为体外储存实验验证生成了 >3 G PE150 读数。对平均深度为 100 倍的测序数据进行随机二次抽样以进行信息检索。首先对读数进行聚类和组装，以完成每种寡核苷酸的序列。去除侧翼引物区域，使用编码的反向操作将 DNA 序列解码为二进制片段，并使用 RS 代码纠正替换错误。二进制段根据地址区域重新排序。在此过程中，根据地址删除了“伪二进制”段。然后将完整的二进制信息转换为数字文件。数据恢复率计算使用。对于误差分析，使用不同的随机种子对平均深度为 100×、300×、500×、700× 和 900× 的测序数据进行六次随机二次抽样。

总共为体内存储生成了 >50 M PE-100 读数，其中 SOAPnuke 过滤了 10% 的低质量读数（Phred 分数 <20）。在 BWA 作图后，使用 samtools 去除宿主基因组的读数，然后通过 SOAPdenovo 将短读取组装成重叠群。Blastn 用于发现 contigs 之间的连接。编写了一个 Python 脚本来合并重叠群并获得每个菌株的组装序列。进行多序列比对以通过比对组装的序列，以进行多数表决过程，以识别结构变异、插入和缺失。预先添加的 RS 代码用于替换的纠错。通过反转编码操作来解码完整的 DNA 序列以恢复二进制信息。

结果分析

YYC 系统的应用范围广

该研究从 DNA 双链模型中受到启发，与中华文化中“阴阳”对立统一的思想相结合，巧妙地应用于 DNA 编解码系统，以两套不同的规则，分别对两条二进制信息进行“一对一”编译转换，再取两者统一交集的部分为最终解，实现将两条独立的信息组合统一为一串 DNA 序列，如图 3 所示。另一方面，通过引入筛选机制，与现有合成测序技术兼容性不佳的序列通过预先设置的筛选条件进行过滤。根据不同的组合方法，该系统共能提供 1536 种不同的编码规则组合，大大扩展了其应用场景范围。

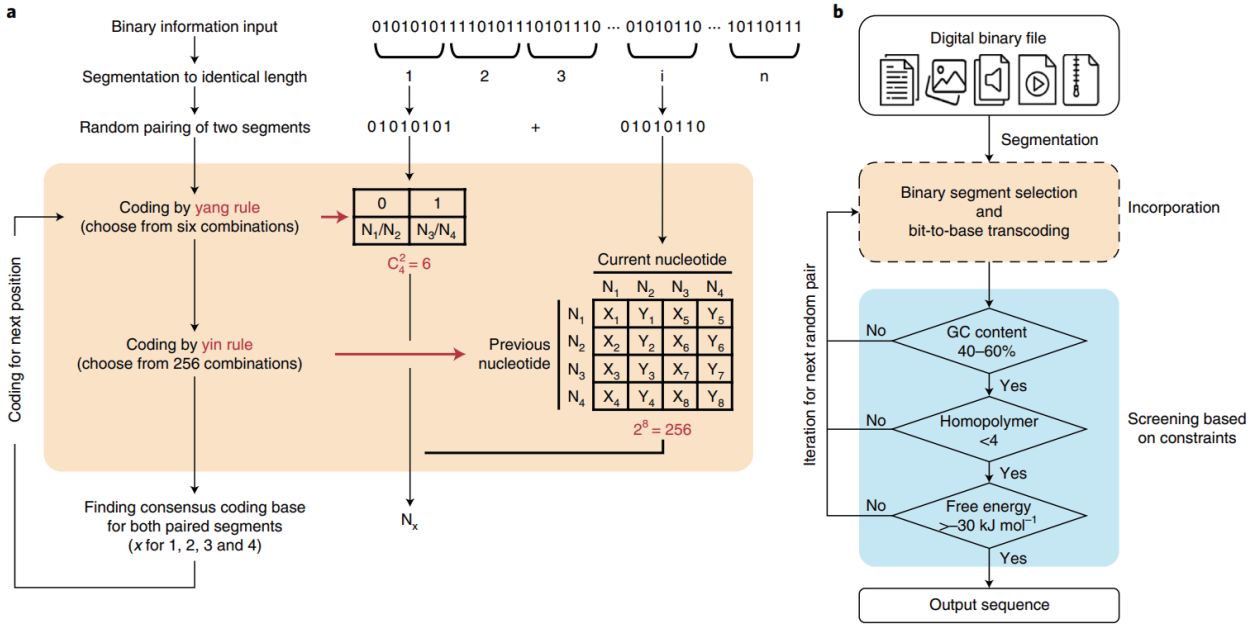

图3.YYC的原理图

YYC 系统存储数据恢复能力强

该研究通过在 DNA 序列中以 0.01%-1% 的平均速率随机引入三种最常见的错误来测试 YYC 对系统误差的鲁棒性。并且，在不引入误差修正机制的情况下，与 DNA Fountain 编码方案进行对比分析相应的数据恢复率。结果如图 4 所示，可以看到无论 Indels 还是 SNV 存在时，YYC 的数据恢复性能都优于 DNA Fountain，数据恢复率保持在相当稳定的 98% 以上的水平。

YYC 体外存储数据可恢复率高

该研究评估成功恢复文件所需的寡核苷酸的最小拷贝数以及对 DNA 分子损失的鲁棒性分析如图 5b 所示。其测序结果显示，在 10^3 以上的 AMC 编号处，P1 对应的数据可恢复到 99.9%；当 AMC 个数为 10^2 时，平均数据恢复率下降到 71.2%，每个存储的文件从 65.69% 到 87.53% 不等。当 AMC 的数字小于 10^1 时，它进一步下降到 10% 以下。总的来说，YYC 呈线性恢复趋势，与数据编码的 DNA 分子保留量呈正相关，如图 5c 所示。对于 DNA Fountain 算法，当 AMC 数字在 10^4 以上时，数据恢复率与 YYC 相当，但当 AMC 数字在 10^3 以下时，数据恢复率明显下降到单拷贝水平。

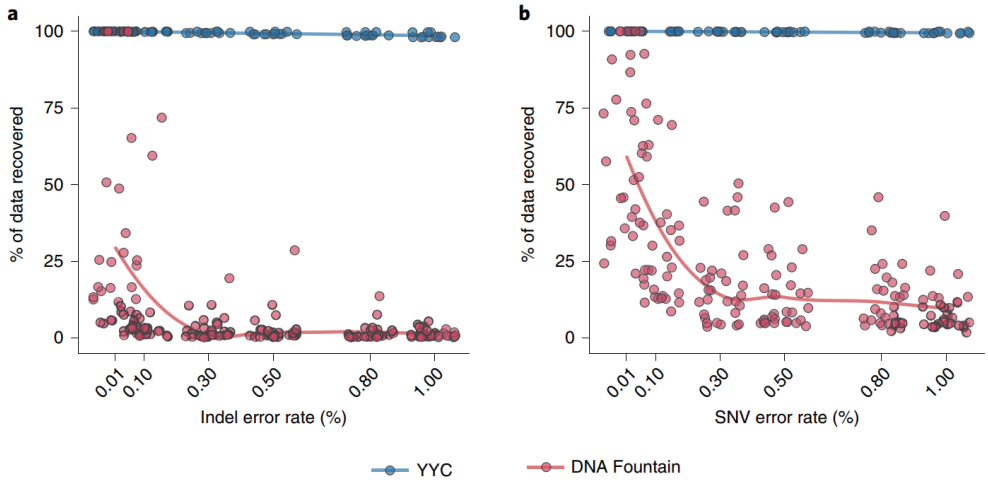

图4.YYC和DNA喷泉码的数据恢复能力分析

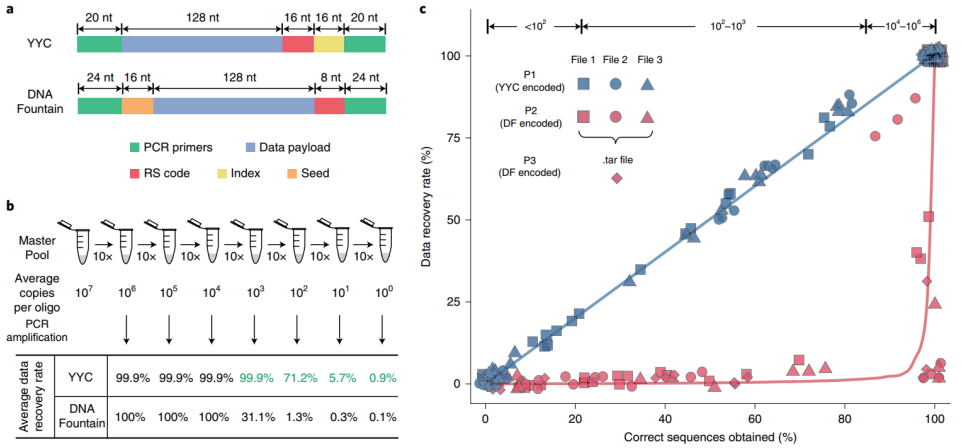

图5.YYC体外存储实验验证结果

YYC 体内储存数据稳健性好

该研究使用 YYC 将文本文件的一部分编码到一个 54240 bp 的 DNA 片段中，如图 6a 所示。并且利用酵母较高的同源重组效率，将这些片段与线性化的低拷贝的着丝粒载体 pRS416 直接转化到酵母菌株 BY4741 中，实现体内一步组装全长 DNA。通过批量转移细胞培养物约 1000 代后，通过对 15 个单菌落进行全基因组测序来评估 YYC 方案的稳定期，如图 6b 所示。此外，观察到所有 15 个单菌落中存在不同程度的部分片段丢失，丢失范围在 ~21.1 kbps 到 ~51.4 kbps 之间，数据恢复水平从 38.9% 到 95.0% 不等。通过采用单个简单多数投票策略，对多个菌落的重建和比对产生一致序列。该研究重建了一个包含 66 个 SNV 的完整序列，这些 SNV 不能被 RS 码引入到数据块中进行校正，并完全恢复了存储的数据。

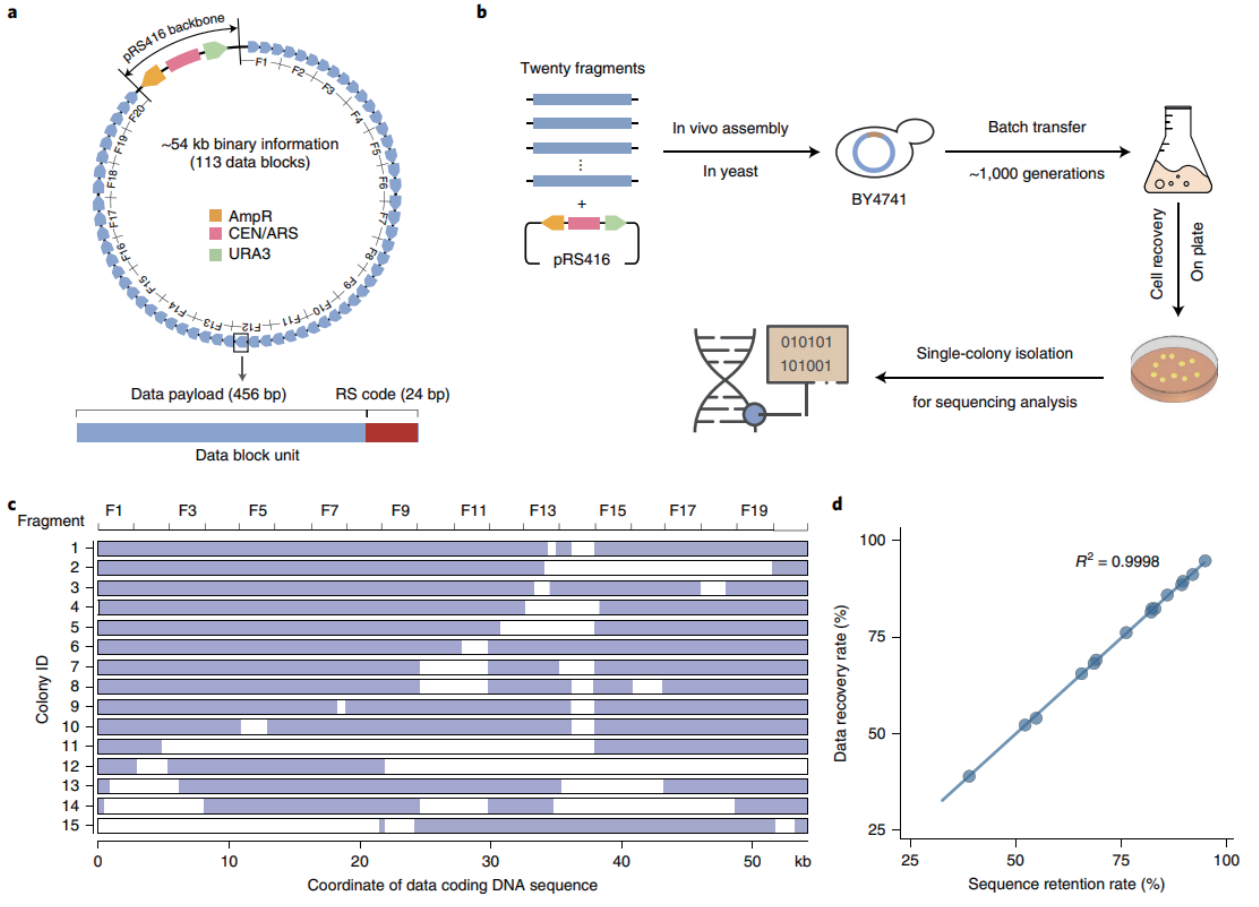

图6.YYC体内存储实验验证结果

总结

基于华大智造 DNBSEQ 测序平台开创的一套独具优势的“阴阳”编解码系统，为 DNA 信息存储的应用提供了一种高密度、高稳定性的比特 - 碱基编解码方法，并完成了体内外两种模式的信息存储实验验证；研究开发了一种全新的 DNA 存储编码方法，为 DNA 存储的多类型应用提供了重要工具。

DNBSEQ 测序平台与华大智造全流程测序产品搭配，性价比高，工作稳定性强，为研究开发了一种全新的 DNA 存储编码方法提供完全国产化且高效可靠的高通量技术支持。研究团队表示，他们对该系统在未来的普及和发展很有信心，有望在海量数据长期存储的新型介质研究中起到积极的推动作用。

基因测序仪 DNBSEQ-T7RS

参考文献

1. Ping Z., Chen S., Zhou G., et al. Towards practical and robust DNA-based data archiving using the yin–yang codec system[J]. *Nature Computational Science*, 2022, 2(4): 234-242.
2. Church G. M., Gao Y., Kosuri S. Next-generation digital information storage in DNA[J]. *Science*, 2012, 337, 1628.
3. Allentoft M. E., Collins M., Harker D., et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils[J]. *Proceedings of the Royal Society B: Biological Sciences*, 2012, 279,4724-4733.
4. Bhat W. A. Bridging data-capacity gap in big data storage[J]. *Future Generation Computer Systems*, 2018, 87, 538-548.
5. Erlich Y., Zielinski D. DNA Fountain enables a robust and efficient storage architecture[J]. *Science*, 2017, 355, 950-954.

推荐订购信息

产品类型	产品名称	产品货号
仪器	基因测序仪 DNBSEQ-T7RS	900-000236-00
	MGIDL-T7RS 全自动样本加载系统	900-000237-00
	基因测序仪 MGISEQ-2000RS	900-000035-00
	基因测序仪 DNBSEQ-E25RS	900-000490-00
软件	数据存储一体机	900-000593-00
	MegaBOLT 生信分析加速器（工作站式服务器）	970-000085-00
建库试剂	MGIEasy 通用 DNA 文库制备试剂套装（16 RXN）	1000006985
测序试剂	DNBSEQ-T7RS 高通量测序试剂套装（FCL PE150）V3.0	940-000268-00
	MGISEQ-2000RS 高通量测序试剂套装（FCL PE150）	1000012555
	DNBSEQ-E25RS 高通量测序试剂套装 (FCL PE150)	940-000567-00

深圳华大智造科技股份有限公司

深圳市盐田区北山工业区综合楼11栋

4000-688-114

www.mgi-tech.com

MGI-service@mgi-tech.com

股票简称：华大智造

股票代码：688114

版本：2023年11月版

撰稿：陈丽琴 黎汉平

责任编辑：王其伟

审稿：江遥