引文格式
史越，贾李佳，刘翟. DNA 数据存储——机遇与挑战 [J].集成技术, 2024,?(?):??

Citing format
Shi Y, Jia LJ, Liu D. DNA-Based Data Storage—Opportunities and Challenges[J]. Journal of Integration Technology,2024,?(?):??

DNA 数据存储——机遇与挑战

史越[1]，贾李佳[1]，刘翟[1]

[1]（中国科学院武汉病毒研究所，武汉 430207）

摘要：自人类进入信息时代以来，全球信息总量飞速增长，对数据存储行业带来极大挑战。目前的信息存储工具存在如信息密度低、使用年限短、环境污染等许多缺陷，而脱氧核糖核酸（DNA）作为天然的遗传信息载体，其信息密度高、稳定性高、保存时间长、维护成本低等特点，或可成为信息存储领域的一项优秀选择。尽管 DNA 存储目前面临着读写成本高、速度慢、错误率高等挑战，但其在许多领域也有着独特的优势，如“冷”数据存储、军事加密存储等。目前 DNA 存储的潜在发展方向主要包括在军事、空天等特殊场景下的应用，高容错的编解码方案，生物活体存储体系，脱离测序的信息读取方法，集成化的存储系统及统一行业标准等。希望在不久的将来，DNA 存储可实现规模化的应用落地，开启数据存储新纪元。

关键词：DNA 存储；合成生物学；生物信息交叉；数字信息存储
中图分类号：Q819 doi: 10.12146/j.issn.2095-3135.20231128002

DNA-Based Data Storage—Opportunities and Challenges

SHI Yue[1], JIA Lijia[1], LIU Di[1]

[1] (Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, 430207, China)
Corresponding Author: Di LIU. **Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, 430207, China**. Email: liud@wh.iov.cn

Abstract: Since the human civilization entered the information age, an exponential growth of digital information globally posed great challenges to data storage. Current data storage devices have many defects, such as limited data density, short lifespan, environment pollution and so on. Deoxyribonucleic acid (DNA), the natural carrier of genetic information, was proposed to be a reasonable alternative due to its high information density, robustness, long half-life and low maintenance cost. Although DNA storage currently faces the challenges of high reading and writhing costs, slow speed and high error rate, it has unique advantages in many fields, such as long-time archival storage, military data encryption and so on. The potential future directions of DNA storage mainly include applications under special scenarios such as space data center and military, encoding-decoding algorithms robust to base errors, *in vivo* DNA storage, information retrieval without sequencing, and integrated DNA storage system as well as a unified evaluation standard. We hope that in the future, DNA storage can achieve large-scale application, and open a new era of data storage.

Key words: DNA storage; synthetic biology; BT-IT; digital information storage

来稿日期：2023-11-28 修回日期：2023-11-29
基金项目：国家自然基金项目(2020YFA0907000)

Funding: This project is supported by the National Natural Science Foundation of China (No. 2020YFA0907000).

引言

自 20 世纪末全球迈入信息时代以来，信息化、数字化浪潮的冲击，带动数字经济、5G、人工智能、物联网等新兴技术的高速发展，随之而来的是全球数据量产出爆发式增长。据国际权威机构 Statista 预测，至 2035 年，全球数据产出量预计将达到 2143ZB，较 2020 年的 47ZB 增长近 45 倍[1]。2016 年 3 月，我国国家发展和改革委员会（发改委）在《十三五规划纲要》中正式将大数据列为国家战略，确立了大数据在经济社会发展中的重要地位。然而，传统数据存储手段受限于信息密度低、能耗高、使用年限短等问题，恐难以在可预见的未来实现突破。飞速增长的数据存储需求与逐渐落后的数据存储能力之间的矛盾，使得脱氧核糖核酸（Deoxyribonucleic Acid, DNA）存储的概念应运而生。

DNA 存储概念最早由 Norbert Wiener 和 Mikhail Neiman 于 1960 代中期提出[2]，并在 1988 年得到初步验证[3]。但受限于 DNA 高通量合成及测序技术的发展，直到 2010 年后 DNA 存储领域才开始有了实质性的进步。DNA 分子作为承载生物遗传信息的天然信息存储介质，其相比基于硅基介质的传统存储方式有着独特优势。一方面，DNA 分子可承载的信息密度极高，与目前存储密度最高的闪存相比，DNA 存储可达到每立方厘米 10^{19} 比特信息，为闪存存储密度的 1000 倍[4]；另一方面，DNA 存储所需的维护成本大大低于传统存储，同时在使用年限、抗干扰能力和稳定性上亦存在显著优势。因此，在数据爆炸的大背景下，DNA 存储这一新兴技术有望成为未来新一代大数据存储的解决方案。

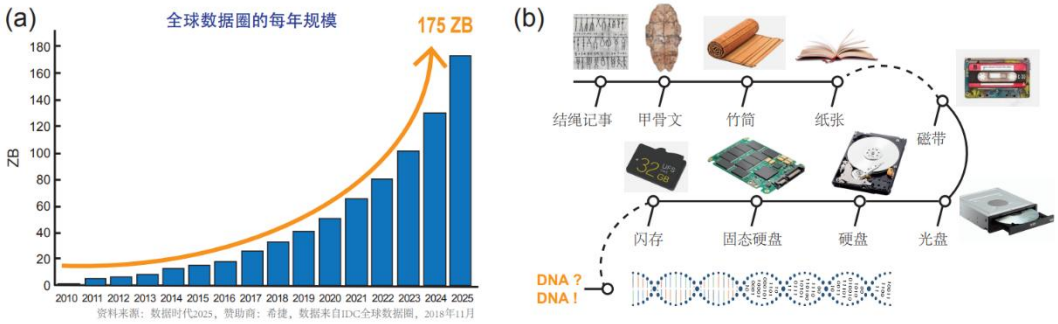


图 **1**　(a)全球数据产出增长预测[1]；(b)信息存储介质发展简史

Fig. 1　A. Global data output forecast [1]; B. History and current progress of information storage devices

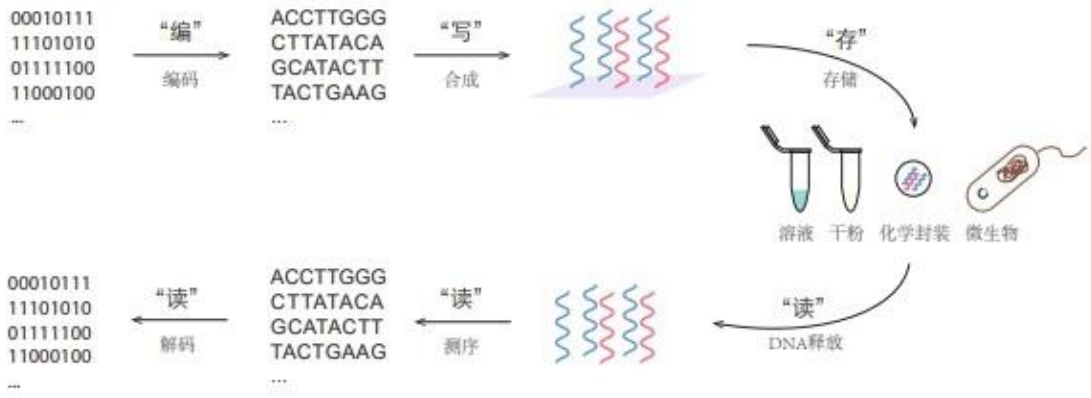
1 DNA 存储研究进展

1.1 DNA 存储基本流程

DNA 存储的全流程可大致分为“编”“写”“存”“读”四个步骤（图 2(a)）。“编”即将待存储二进制信息通过编码算法转换为碱基信息。目前常用的 DNA 存储编码算法包括碱基与二进制信息直接映射的 Church 编码[5]、基于霍夫曼编码和三进制轮换编码的 Goldman 编码[6]、DNA 喷泉码[7]等。另外，与传统信息存储不同，受 DNA 合成技术限制，单条 DNA 序列的合成长度有限，无法将所有信息存入单个分子中，因此，目前所有编码方法均需要将信息片段化，并对每个信息片段添加相应的位置索引。“写”即将编码好的碱基信息通过 DNA 合成写入到 DNA 分子中。主流的 DNA 合成方式包括化学合成与生物合成两大方向[8]，化学合成法为目前较为成熟的 DNA 合成方式，包括传统的柱式合成法和以 Twist

（https://www.twistbioscience.com/）、迪赢生物（https://www.dynegene.com/）为代表的簇式高通量合成法等。生物合成法则是 2013 年左右逐渐开始发展的新技术，通过酶促合成提高长链 DNA 分子的合成效率，降低合成及组装成本，并减少有害废液的产出[9]。“存”步骤将合成好的 DNA 分子在适宜条件下长期稳定储存，干粉和溶液是实验室保存 DNA 分子的传统方法，可在低温环境下稳定保存数年至数十年；DNA 物理封装也是一种可有效保护 DNA 分子的方式，苏黎世联邦理工学院的 Grass 团队于 2013 至 2019 年的一系列研究中提出了一种适用于 DNA 存储的封装技术，将 DNA 与聚乙烯亚胺交替包裹在纳米磁珠上，最外层添加二氧化硅层，可使 DNA 在室温下保存 20～90 年，10℃下则可保存 527 年之久[10-12]。其团队与哥伦比亚大学合作，运用该技术与 3D 打印技术培育了五代包含 DNA 信息的“斯坦福兔子”，并完成了信息的完美恢复，证实了 DNA 封装拥有优良的稳定性和保真性[13]。除体外存储之外，DNA 信息体内存储也有着不容忽视的优点，将包含数据信息的质粒载体转化进大肠杆菌中，可实现信息低成本复制和长期保存[14]。

“读”步骤即信息还原，依赖于 DNA 测序技术。首先对承载待读取信息的 DNA 分子进行测序，然后对测序结果进行解码纠错后还原到原始内容。如前文所述，DNA 存储有着片段化、离散化的特性，因此通常选择可以同时对大量 DNA 分子进行测序的 NGS 高通量测序[15]，而三代 Nanopore 测序由于其快速便捷的特性，在 DNA 存储领域也有部分应用[16-17]。DNA 信息的解码过程为编码算法的逆运算，考虑到 DNA 分子在合成、储存、PCR 扩增、测序过程中可能发生的碱基突变、插入及丢失，需要设计相应的纠错算法，通过添加冗余的方式提高信息读取的容错率，确保在错误低于一定量时能完美还原到原始内容。里德-所罗门(Reed-Solomon, RS)纠错码、低密度奇偶校验（Low-Density Parity-Check, LDPC）码、异或（exclusive OR, XOR）算法均为经典的纠容错算法。



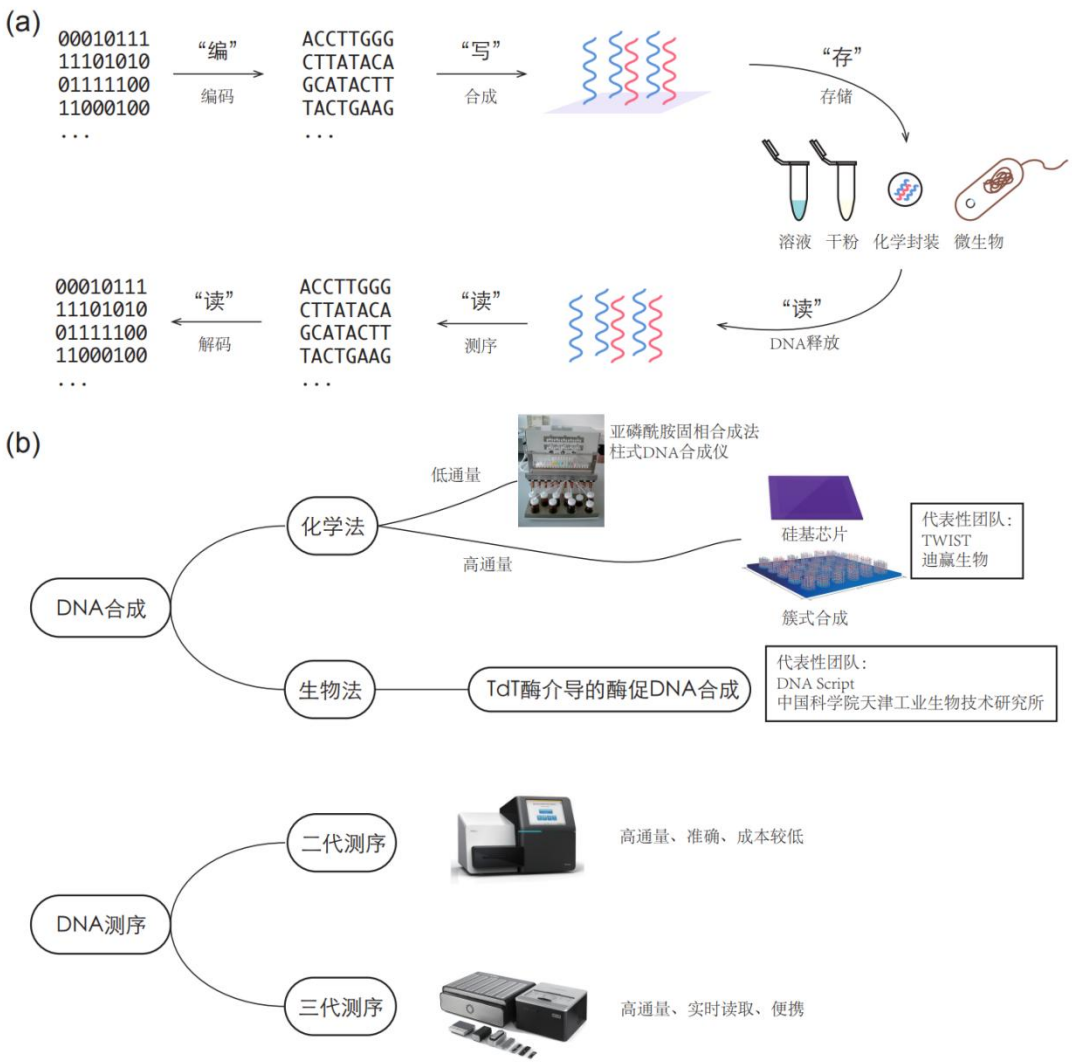


图 **2**　(a) DNA 信息存储流程示意图；(b)DNA 合成、测序技术路线

Fig. 2　A. Schematic diagram of DNA information storage; B. DNA synthesis and sequencing techniques

1.2 DNA 存储领域国内外研究现状

DNA 存储的首个里程碑式成果由哈佛医学院的 Church 研究组在 2012 年实现，其提出了一种使用两种碱基映射一种比特的灵活编码方式，可以有效规避高 GC 含量、长均聚物和 DNA 二级结构等影响合成和测序效率的问题，并成功将一本 650 KB 的图书编码存储到亿万分之一克 DNA 中[5]。2013 年，欧洲生物信息研究所的 Goldman 团队在 Nature 发表论文，使用三进制霍夫曼编码实现了 ASCII 文本、PDF、JPG 和 MP3 格式合计 739 KB 文件的 DNA 存储，文件还原准确率超过 99.99%，也是纠错码在 DNA 存储领域的首次应用[18]。2015 和 2016 年，Grass 和 Blawat 团队分别将 RS 码和 LDPC 码这两种计算机领域常用的纠错码运用在 DNA 信息编码中[19-20]。2017 年，哥伦比亚大学和纽约基因组中心的 Erlich 团队参考通信领域的喷泉码，设计了一种适合 DNA 存储的 DNA 喷泉码，该编码方式在保持高容错率的前提下大大降低了冗余度，在 1.57 比特/碱基的净信息密度下可达成完美还原[7]。

我国在 DNA 存储领域的发展虽然较西方略为滞后，但近年来也在强势追赶，达成了多项技术突破，取得一系列国际领先水平的研究成果。2017 年，深圳华大生命科学研究院

实现了多片段连接存储的原理验证，并于 2021 年获得专利授权[21]。2021 年，天津大学元英进团队从头设计合成了一条 254,886 bp 的专用于 DNA 信息存储的人工染色体，编码了合计 37.8 KB 的文本、图片和视频文件，将单菌内数据存储 DNA 数量提升到了百 kbp 级，并实现了数据的可靠恢复[22]。同年，深圳华大生命科学研究院平质团队开发出了一套“阴阳”双编码算法，其结合中国古代阴阳哲学思想，并综合参考了 Goldman 编码和 DNA 喷泉码的设计思路，在低分子拷贝数下数据恢复率可达到 88%[23]。2022 年 3 月，国务院发布《“十四五”数字经济发展规划》，提出“抢先布局前沿技术融合创新”“推进前沿学科与交叉研究平台建设，将 DNA 存储列为需要重点布局的新兴战略技术，显示出国家层面对 DNA 存储技术的高度重视。

除科研领域，DNA 存储在商业化应用方面的潜能也受到多方认可。IT 研究与顾问咨询公司 Gartner 预测，至 2024 年，约有 30%的信息存储企业将会使用 DNA 存储以对应爆炸式增长的信息总量[24]。目前，西方国家已在 DNA 存储上有多项商业化应用尝试。Helixworks 公司于 2016 年在 Amazon 上提供 512 KB 容量的 DNA 存储硬盘，可以支持在适宜环境下存储 1 世纪之久，是首个提供商业化 DNA 存储服务的公司[25-26]。2019 年，CATALOG 公司推出了其自主研制的高通量 DNA 存储设备，并成功将 16 GB 的英文维基百科页面存储到了 DNA 中，证实了商业化大规模 DNA 存储的可能性[27]。2023 年，Biomemory 公司推出了 DNA 数据存储卡，该款存储卡仅有信用卡大小，支持 1 KB 容量的文本数据存储，最短寿命为 150 年[28]。我国在 DNA 存储商业化上的摸索主要集中在 DNA 高通量合成技术上的突破，目前还未见成熟的 DNA 存储商业化产品，但包括中科碳原（http://carbon-atom.com/#/products）、苏州泓迅生物（http://www.synbio-tech.com.cn/synthetic-biology/dna-ubar/）等企业已对外提供 DNA 存储服务。

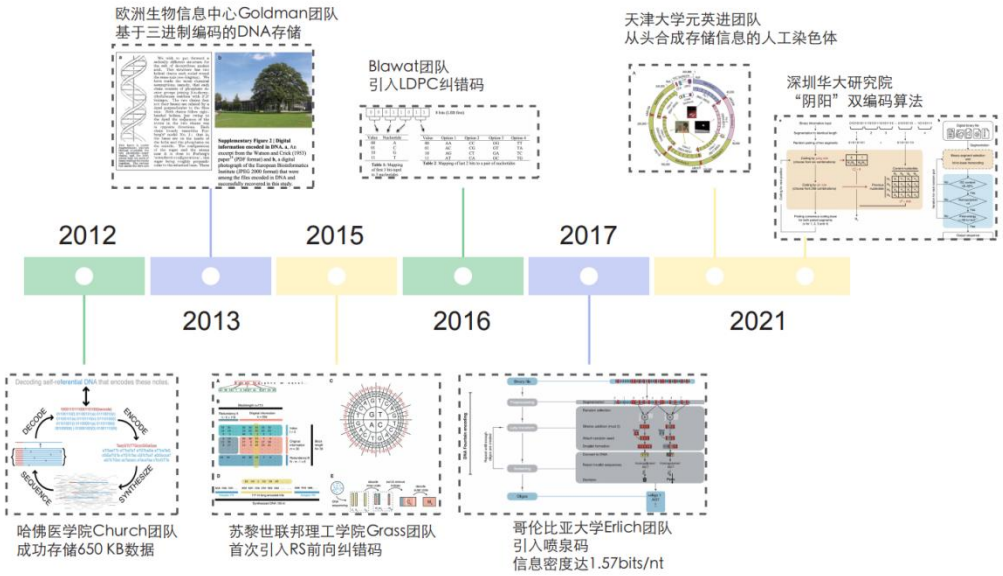


图 3　国内外 DNA 存储发展里程碑事件

Fig. 3　Milestones in the development of DNA storage

2 DNA 存储目前的挑战

DNA 存储是生物学与计算机科学、密码学、信息学等学科的跨领域有机结合，其在解决传统信息存储的痛点问题时，也不可避免地带来了一些生物分子本身的特性，这使得

DNA 存储在实现上有着与传统信息存储领域截然不同的难点。尽管从 2012 年至今 DNA 存储已经有了蓬勃的发展，但多数已发表的成果都局限在 KB 级、MB 级存储，更大规模的 DNA 存储应用目前仍是空白。DNA 存储从实验室走向实际应用的最大障碍主要集中在相对较高的错误率以及随之而来的高纠容错算法需求、高昂的合成和读取成本、繁琐的实验流程这几大方面。

2.1 DNA 存储相对较高的错误率与低容错性

2.1.1 信息写入 DNA 分子时引入的碱基错误

在将编码好的信息写入 DNA 分子中时，目前主流的化学合成 DNA 技术基于碱基之间的偶联反应，但其每个步骤均可能存在反应不完全和副反应等问题，这是化学反应本身的局限性，该限制使得合成产物中往往会存在一定比例的错误[29]。影响 DNA 序列准确性的错误类型主要分为碱基替换、插入和缺失三种。目前主流的 DNA 化学合成手段在每个位置上发生碱基替换的概率约为 0.2-0.5%[30-32]。由于 DNA 合成是逐位添加碱基，随着合成 DNA 链的长度增加，碱基替换概率也将相应累加。一条 100 bp 长的 DNA 链，其不发生碱基替换的概率为$(1\sim0.2\%)^{200}$即 81.8%，而当长度上升到 200 bp 时，序列不发生碱基替换的概率仅为 67%。除碱基替换以外，碱基缺失也是 DNA 合成中较为常见的错误，其概率大约在 0.1%左右，而碱基插入的概率则相对较低，仅在 0.01-0.1%[30-32]，这些也大大限制了人工合成 DNA 分子的序列准确率。

2.1.2 从 DNA 中读取信息时引入的碱基错误

在常规的 DNA 存储实验流程中，存储信息的 DNA 分子在合成完成后以干粉、封装等形式长期保存，在需要读取信息时，则利用预先设计的引物区域，将待读取信息通过 PCR 扩增的方式提取出来，进行后续的测序步骤。然而，PCR 扩增步骤也会引入碱基错误，这些错误一方面来源于 DNA 聚合酶本身的保真性有限[33-34]，另一方面则在于序列本身的扩增偏好性[35-36]。亦有研究发现，过低的模板浓度可导致 PCR 产物的突变率大幅提升，而这些错误无法通过使用高保真聚合酶降低[37]。一项针对 DNA 存储全流程错误率的研究显示，最终测序产物的错误率约为每千碱基 6.7±6.9 个缺失错误，7.9±2.0 个替换错误，以及低于 0.3±0.2 个插入错误。方差分析（Analysis of Variance, ANOVA）显示产物中的碱基缺失错误绝大部分由合成环节贡献，而替换错误则主要在 PCR 过程中产生[38]。相较之下，传统硬盘的读取错误率约为每比特 10^{14} 个错误[39]，比 DNA 存储的错误率低约十个数量级。

2.1.3 DNA 存储对纠容错算法的高需求

在传统硬盘存储领域，常见的信息错误为误码和信道损坏，分别对应 DNA 存储中的碱基替换错误和整条 DNA 序列丢失。因此，这两类错误均可通过如 RS 码、LDPC 码等主流纠删码算法和添加冗余的方式纠正。然而，DNA 分子存在独特的碱基插入和丢失错误，这些错误如果无法被纠正，可能导致分子生物学中常见的移码突变问题[40]，使得后续所有信息均出错，而这种错误是一般的纠删码算法难以处理的[41]，使得整条序列均无法被还原。同时由于 DNA 存储流程中涉及繁琐的分子生物学实验步骤，也容易引入大量系统误差，无法依靠处理测序数据的方式来规避[41]。DNA 存储独特的错误类型，使得其往往需要更加复杂的纠容错算法以满足信息还原的需求。

目前主流的 DNA 存储纠容错策略包括添加冗余及内外校验码，如 Goldman 团队采取的 RS 校验码配合四倍冗余的编码策略[18]。然而，添加冗余降低了信息密度并增加了需要合成的 DNA 分子数量，使得读写成本进一步升高。DNA 喷泉码可以达到极高的净信息密度，但其恢复能力波动较大，部分情况下甚至低于不添加纠错码的数据恢复率[7, 41]。探究低冗余、高纠错能力的纠容错算法也是 DNA 存储面临的一项重大挑战。天津大学合成生物学团队于 2022 年提出了一种基于德布莱英图的序列重建算法，存储了 6.8 MB 的敦煌壁画，并在长达十周的加速老化实验后仍能达到完美还原，为 DNA 存储中可能出现的独特

错误类型提供了一种解决方案[42]。

表 **1**　DNA 存储各环节碱基错误发生率[30-32]

Table 1　Base error rates in different steps of DNA storage[30-32]

	DNA 合成	DNA 储存	DNA 测序
碱基替换	0.2%～0.5%	0.0164%/半	0.1%～0.26%
碱基缺失	0.1%	0.0083%/半	<0.01%
碱基插入	0.01%～0.1%	—	<0.01%

2.2 DNA 存储的成本居高不下

2.2.1 高昂的 DNA 合成及测序成本

DNA 存储最大的经济成本来源于 DNA 分子的合成。如前文所述，由于 DNA 信息存储的低容错性，使得其需要更多的数据冗余以确保在出现错误时能够纠正并还原内容。然而，与传统信息存储不同的是，DNA 存储的所有信息均以 DNA 分子为载体，而大量的冗余意味着大量额外的合成成本。目前，商业合成寡核苷酸池的价格约为每碱基 0.002 美元，以信息密度较高的喷泉码编码（1.57 比特/碱基）为例计算，折合每比特信息 0.0013 美元，即约 1.04×10^{7} 美元/GB[43]。在信息读取方面，根据测序服务提供商不同，二代高通量测序（next-generation sequencing, NGS）每 GB（Gigabase，即十亿碱基）数据的价格约在 5～20 美元不等[44]。尽管 NGS 测序的单价较低，但为了保证原始数据能够可靠还原，最少需要约 35×的测序覆盖度，即测到目标碱基数的 35 倍[45]。同样以喷泉码为例计算，折合每比特信息的读取成本至少为 1.11×10^{-7} 美元，即约 892 美元/GB。高昂的读写成本，令 DNA 存储的经济成本高于硬盘存储约 8 个数量级（硬盘存储的成本仅为约 0.013 美元/GB）[46]，极大地阻碍了 DNA 存储走向规模化应用。

2.2.2 实验流程耗费巨大时间成本

除了经济成本外，现阶段 DNA 存储的时间成本也居高不下。在信息写入阶段，合成所需时间由合成仪的最高通量和合成周期决定。目前市面上的 DNA 柱式合成最多可达 1536 通量，合成寡核苷酸长度上限在 150～200 bp 之间，每添加一个碱基耗时约为 30 分钟[47-48]。以一个 5 MB 的小文件为例，其包含的信息在不添加任何冗余和寻址标签的情况下至少需要 260 万碱基来存储，如果使用柱式合成法，这些信息需要进行至少 8 轮合成，耗时超过三周。目前二代高通量合成技术理论上可在单张芯片上同时合成数万条乃至数百万条寡核苷酸，但产物量极低，仅能达到 fmol 水平，在后续实验中需要通过 PCR 增加产物量，变相延长了时间周期[49]。除 DNA 合成耗时之外，DNA 信息的读取速度也与硬盘存储存在较大差距。Illumina 测序平台的主流测序仪产出数据速度在 50 KB～1 MB/秒[50]，与传统的机械硬盘（160 MB/秒）和固态硬盘（550 MB/秒）相比差了 2～4 个数量级。

表 **2**　DNA 存储与传统存储工具的比较[51-53]

Table 2　Comparison of DNA storage vs. traditional information storage[51-53]

	读写成本	能耗	存储密度	读写速度	保存年
光盘	<0.005 美元	～0.1W/GB	$\sim10^{10}$bit/cm^3	20MB/s	5～20 年
硬盘	0.013 美元	～0.4W/GB	$\sim10^{13}$bit/cm^3	160～	5～20 年
闪存	<0.2 美元/GB	0.01～	$\sim10^{16}$bit/cm^3	1.0～1.5MB/s	～10 年
DNA 存储	1.04×10^{7} 美元/GB	<10^{10}W/GB	$\sim10^{19}$bit/cm^3	读：0.05～1MB/s 写	>1000 年

3 DNA 存储未来发展方向

现阶段，DNA 存储的优缺点十分鲜明，一方面有着信息密度高、保存年限长、维护成本低等优势，另一方面却又有着读写成本高、周期漫长繁琐等缺陷，这些问题受限于

DNA 合成和测序技术的发展，难以在可预见的未来取得颠覆性进展。不过，DNA 存储仍然在许多领域有着独特的优势，如“冷”数据存储、军事加密存储等。随着科技的不断发展，生物科技（Biological Technology, BT）和信息科技（Information Technology, IT）领域的深度融合，未来有望在 DNA 存储的许多关键问题上取得重要突破，为推动 DNA 存储从实验阶段走向应用阶段迈出坚实的一步。

3.1 DNA 存储的高稳定性与生物加密特性开辟的特殊应用场景

DNA 分子可以在适宜条件下稳定储存数百年之久，有研究者曾在距今 56～78 万年的化石中提取出完整的野马基因组序列[54]。理论上，DNA 分子在面对脱嘌呤和脱氨基这两种最常见的降解方式时，在零下 4 摄氏度的水中可分别保存 1,108,965 年和 283,001 年，而在同样温度的空气中保存年限则能上升约 3 个数量级[55]。

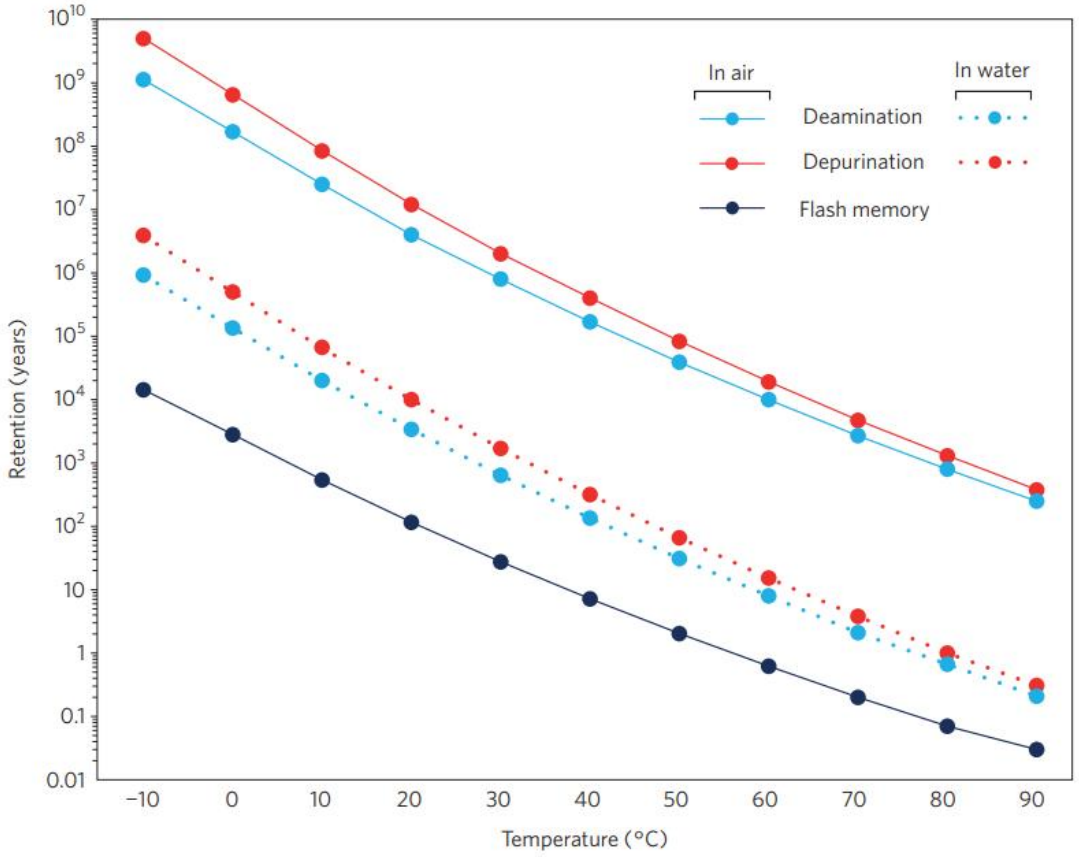


图 4　不同温度下 DNA 与闪存在空气及水中的保存年限对比[55]

Fig. 4　Comparing retention years of DNA vs. flash memory under different temperature in air and water [55]

这种极高的稳定性使得 DNA 存储在一些特殊场景、极端环境下拥有比传统存储方式更高的应用潜力。譬如，我国目前正在设计部署空天数据中心，有望在太空中实现数据的大规模存储和备份，而 DNA 存储的高稳定性、抗电磁辐射与低能耗等特征，使得其在空天数据存储上有着天然优势。相较于电子存储技术，面对太空环境中可能发生的电磁辐射干扰、温度波动和存储空间有限等种种问题，DNA 存储展现出巨大的潜力与前景。

DNA 存储的另一个独特性是其作为生物化学分子，天然拥有区别于计算加密的生物加密特性，可以通过细菌抗性、荧光探针杂交、DNA 二级结构等生物反应对存储信息进行加密，大幅提高信息截获难度，降低信息泄露风险，这使得 DNA 存储在军事领域亦有出色的应用价值。上海交通大学樊春海团队开发了一种基于 DNA 折纸的信息多重加密方式，通过指定的密钥将携带信息的 DNA 骨架链以正确方式折叠，才可以对信息进行解码[56]。湖南科技大学张翼飞团队开发了一系列基于探针杂交的 DNA 存储加密技术，读取前必须

选择正确的处理步骤，否则将造成信息错误甚至自毁，且截获方处理后必定留下可检出的痕迹[57-59]。

3.2 以信息密度换取高容错、低成本的编码思路

目前，西方主流的 DNA 存储编码方式，无论是 Goldman 编码、Church 编码还是喷泉码，均用碱基与二进制信息直接对应，本质上仍然是计算机信息的思路，在对编码方式的优化上也以尽量降低完美还原所需的冗余量、使信息密度逼近理论上的香农极限（2 比特/碱基）为基本方向。然而这种编码方式每次都需要单独合成 DNA，其成本难以控制，这也是 DNA 存储规模化、商业化应用的最大阻碍。对此，近年来包括中国科学院[60]、天津大学[61]、美国 CATALOG[62]等在内的多个研究团队采取了新的编码思路，借鉴中国古代活字印刷的思想，用一段 DNA 短序列表示一段信息（如一个汉字），存储时通过酶促连接将所需短序列组合成长链 DNA。这种 DNA 活字存储牺牲了编码密度，但由于 DNA 短序列“活字”可以单次合成反复使用，在进行大规模存储时反而比逐位合成更加节省成本，且两者之间的差距随存储规模的上升而增加。同时，由于 DNA 活字存储使用一段序列对应一段信息，对于 DNA 存储中常见的单碱基替换、缺失等错误有着极高的鲁棒性，变相节省了传统 DNA 存储中用于纠删码和冗余的空间，进一步降低使用成本。

3.3 探究生物活体 DNA 存储体系

在 DNA 存储中，信息的复制和传递需要通过对储存信息的原始 DNA 分子进行 PCR 扩增。然而，对双链 DNA 进行 PCR 扩增时首先需要对原始 DNA 模板进行加热变性，将双链解离成单链以作为后续扩增模板，这一步会导致原始模板被消耗[63]。随着信息被多次复制，PCR 过程引入的碱基错误会同时在原始信息和拷贝信息中积累，变相增加信息错误率，这是 DNA 存储与传统存储的又一大显著区别。针对这一特点，部分研究者尝试探究新的存储介质，使用质粒、人工染色体等方式保存 DNA 信息，并转化进生物体内（大肠杆菌等）长期保存。该技术路线的优势是可以利用菌体自身的复制制造信息拷贝，而大肠杆菌、酵母菌等常用模式生物复制过程中突变率低，可以达成信息的高保真复制[64]。韩国危害监测生物纳米研究中心的 Moonil Kim 团队于 2018 年完成了通过利用质粒长期存储 DNA 信息的验证，将《国际人权宣言》编码进 22 种包含 400 bp 信息的质粒并转化进大肠杆菌中培养，并完成了信息的完美还原[65]。该研究预计质粒 DNA 可在-20 摄氏度环境下稳定保存约 20 年，进一步证明了生物活体存储体系的发展潜力。另一方面，由于菌体自身可以进行快速、低成本的复制，使得生物活体存储或可作为一种信息冗余备份的手段，进一步提升 DNA 信息存储的可靠性。中国科学院微生物研究所及合作团队在 2022 年提出了一种基于微生物体系的高容纠错 DNA 存储阵列技术(Bio-RAID)，通过携带存储信息质粒的微生物的低成本、无限量复制，达到传统存储中必须用多个硬盘阵列才能达到的效果，是该技术路线下的一次有力尝试[66]。

3.4 探索高速便捷的信息读写方法

如前文所述，现阶段 DNA 存储的信息读写依赖 DNA 合成和测序，而不论是 DNA 化学合成和二代测序，其时间周期均远远长于传统硬盘存储，无法满足日常热数据的高频读写需求。因此，探究脱离合成——测序的快速、高通量、操作简便、成本低廉的数据读写方式，使 DNA 存储在冷数据存储之外亦可应对热数据存储，也是 DNA 存储领域一个重要的发展方向。湖南科技大学生命科学学院张翼飞团队与中科院武汉病毒所合作，于 2022 年提出了一种基于寡核苷酸芯片杂交的信息存储方法，通过读取荧光探针杂交产生的荧光信号即可快速准确地还原信息，成本比基于合成和测序的读写方法降低了 2 个数量级[67]。

除此之外，一些来自其他碳基信息存储领域的研究也为 DNA 存储提供了另一种全新的技术思路，即用不同分子量的生物分子编码不同信息，再通过质谱等目前十分成熟的大分子检测分离技术进行读取。香港理工大学的姚钟平/刘重明研究团队于 2022 年提出一种

基于多肽序列的信息存储方案，使用不同氨基酸作为信息存储单元，并通过串联质谱达到信息的快速精确读取[68]。而在传统的分子生物学实验中，亦会使用聚丙烯酰胺凝胶电泳（PAGE）对 DNA 分子进行精度高达 1 bp 的分离[69]；高效液相色谱（HPLC）、电喷雾质谱（ESI-MS）等分析手段也可对 DNA 进行基于序列长度或分子量的高精度分辨，甚至可分辨 DNA 二级结构[70-72]，或可使囊括碱基信息、序列长度、分子量、二级结构的多维度编码成为可能，也是一种有潜力的研究方向。

3.5 集成化的存储体系与行业标准

自 2012 年 DNA 存储首次实现应用以来，DNA 存储领域的研究百花齐放，在“编”“写”“存”“读”各步骤上均有突破性进展。然而，与传统电子信息存储可以在计算机上一次性完成写、存、读的所有步骤相比，DNA 存储的各流程仍处于割裂状态，DNA 的合成、保存、测序往往需要在不同平台之间完成，过程中涉及到大量实验人员和仪器设备参与，缺乏一套方便快捷的集成化、自动化的信息存取流程，也使得 DNA 存储现阶段难以脱离实验室场景走向实际应用。因此，如何整合 DNA 存储中的各个流程，建立 DNA 存储“编”“写”“存”“读”一体化系统，也是 DNA 存储走向实际应用的重要课题。东南大学生物科学与医学工程学院刘宏团队设计开发了一套在单电极上同时完成 DNA 合成与测序的实验方案，为未来实现 DNA 存储存读一体化提供了硬件基础[73]。

另一方面，DNA 存储领域目前存在多条主流技术路线，而不同技术路线之间的编解码算法和实验流程难以兼容，也阻碍着 DNA 存储的应用化。而随着 DNA 合成、封装、测序技术的发展，新的 DNA 存储技术路线也将不断涌现，一套具有良好拓展性、可兼容不同文件类型、编解码算法、存储介质的软硬件结合系统，建立行业统一的存储性能综合评判机制，对于未来实现 DNA 存储全流程一体化也是至关重要。深圳华大生命科学研究院研究团队发表了 DNA 存储不同编解码算法的集成与评估平台 Chamaeleo，集成了 6 种 DNA 存储主流编解码算法，可以针对文件分析不同编解码算法的表现评分，并提供拓展接口以应对新的算法[41]。

结语

随着大数据时代的到来，DNA 存储这一新兴数据存储技术也备受关注。作为保存生物遗传信息的天然信息介质，与传统的硅基存储介质相比，DNA 分子在信息密度、稳定性、使用年限、维护成本上均有着显著优势。目前 DNA 存储领域的发展仍处于初级阶段，虽然在全球已有不少实验室级别的小规模应用实现，但其错误率偏高、读写成本昂贵、流程繁杂冗长等痛点问题，仍阻碍着 DNA 存储走向大规模、商业化应用。本文讨论了 DNA 存储发展目前面临的主要挑战，并提出了未来该领域可能的几大重点研究方向。随着 DNA 合成和测序技术的不断进步，分子生物学、信息学、计算机科学等多学科交叉日趋深入，希望在不久的将来能实现 DNA 存储在民用、军事、空天等领域的应用落地，开启数据存储的新纪元。

致谢

本工作由国家重点研发计划“合成生物学”重点专项（2020YFA0907000）支持，感谢项目团队在执行过程中作出的贡献和共同努力，感谢《集成技术》编委会提供的资源平台。

参考文献

请参照国家标准 GB/T 7714-2005 执行。电子文献代号：数据库 DB；计算机程序 CP；电子公告 EB；磁带 MT；磁盘 DK；光盘 CD；联机网络 OL。

[1] 中国信息通信研究院. 大数据白皮书（2020 年）[R/OL]. (2020-12) [2023-10-31]. http://www.caict.ac.cn/kxyj/qwfb/bps/202012/t20201228_367162.htm

China Academy of Information and Communications Technology. Big Data White Paper (2020) [R/OL]. (2020-12) [2023-10-3]. http://www.caict.ac.cn/kxyj/qwfb/bps/202012/t20201228_367162.htm

[2] Joshua L. Machines Smarter Than Men? Interview with Dr. Norbert Wiener, Noted Scientist [EB/OL]. U.S. News & World Report, Inc., 1964(1964-02-24) [2023-10-31]. https://profiles.nlm.nih.gov/spotlight/bb/catalog/nlm:nlmuid-101584906X7699-doc

[3] Davis J. Microvenus [J]. Art Journal, 1996, 55(70).

[4] Extance A. How DNA could store all the world's data [J]. Nature, 2016, 537(7618): 22–24.

[5] Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA [J]. Science, 2012, 337(6102): 1628.

[6] Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA [J]. Nature, 2013, 494(7435): 77-80.

[7] Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture [J]. Science, 2017, 355(6328): 950-954.

[8] Hoose A, Vellacott R, Storch M, et al. DNA synthesis technologies to close the gene writing gap [J]. Nature reviews. Chemistry, 2023, 7(3): 144–161.

[9] Eisenstein M. Enzymatic DNA synthesis enters new phase [J]. Nature biotechnology, 2020, 38(10): 1113–1115.

[10] Paunescu D, Puddu M, Soellner JO, et al. Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils' [J]. Nature Protocols. 2013, 8(12): 2440-2448.

[11] Paunescu D, Fuhrer R, Grass RN. Protection and deprotection of DNA--high-temperature stability of nucleic acid barcodes for polymer labeling [J]. Angewandte Chemie (International ed. in English). 2013, 52(15): 4269-4272.

[12] Chen WD, Kohll AX, Nguyen BH, et al. Combining Data Longevity with High Storage Capacity—Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles [J/OL]. Advanced Functional Materials. 2019, 29. https://doi.org/10.1002/adfm.201901672.

[13] Koch J, Gantenbein S, Masania K, et al. A DNA-of-things storage architecture to create materials with embedded memory [J]. Nature Biotechnology. 2020, 38(1): 39-43.

[14] 郜艳敏, 唐梦童, 刘倩, 等. DNA 信息存储中关键生化方法的研究 [J]. 合成生物学, 2021, 2(3): 384-398.

Gao YM, Tang MT, Liu Q, et al. The pivotal biochemical methods in DNA data storage [J]. Synthetic Biology Journal, 2021, 2(3): 384-398.

[15] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies [J]. Nature reviews. Genetics, 2016, 17(6): 333–351.

[16] Organick L, Dumas Ang S, Chen Y, et al. Random access in large-scale DNA data storage [J]. Nature Biotechnology, 2018, 36: 242-248.

[17] Lopez R, Chen YJ, Dumas Ang S, et al. DNA assembly for nanopore data storage readout [J]. Nature communications, 2019, 10(1): 2933.

[18] Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA [J]. Nature, 2013, 494(7435): 77-80.

[19] Grass RN, Heckel R, Puddu M, et al. (2015). Robust chemical preservation of digital information on DNA in silica with error-correcting codes [J]. Angewandte Chemie (International ed. in English), 2015, 54(8): 2552–2555.

[20] Blawat M, Gaedke K, Hütter I, et al. Forward Error Correction for DNA Data Storage [J]. Procedia Computer Science, 2016, 80: 1011-1022.

[21] Shen Y, Chen T, Liu LY, et al. Method for using DNA to store text information, decoding method therefor and application thereof, CN2016/081037 [P]. 2017-09-11.

[22] Ping Z, Chen SH, Zhou GY, et al. Towards practical and robust DNA-based data archiving using the yin-yang codec system [J]. Nature computational science, 2022, 2: 234-242.

[23] Chen WG, Han MZ, Zhou JT, et al. An artificial chromosome for data storage [J]. National Science Review, 2021, 8(5): nwab028.

[24] Gartner. Gartner Top 10 Strategic Predictions for 2021 and Beyond [EB/OL]. [2023-12-29]. https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-predictions-for-2021-and-beyond.

[25] Amazon. DNADrive [EB/OL]. [2023-12-29]. https://www.amazon.in/Helixworks-dsDNA300-750hw-DNADrive/dp/B01IVSH4OM.

[26] McKenna C. Helixworks is the first start-up to offer DNA data storage on Amazon [EB/OL]. (2016-12-31) [2023-12-29]. https://www.siliconrepublic.com/video/helixworks-dna-data-storage.

[27] Shankland S. Startup packs all 16GB of Wikipedia onto DNA strands to demonstrate new storage tech [EB/OL]. (2019-06-29) [2023-12-29]. https://www.cnet.com/tech/computing/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/.

[28] Trueman C. Biomemory launches first commercially available DNA storage solution [EB/OL]. (2023-12-06) [2023-12-29]. https://www.datacenterdynamics.com/en/news/biomemory-launches-first-commercially-available-dna-storage-solution/.

[29] 赛百盛. DNA 合成原理概述 [EB/OL]. [2023-12-29]. http://www.sbsbio.com/cuxiao/DNAziliao1.pdf.

SBS Genetech. An introduction of DNA synthesis [EB/OL]. [2023-12-29]. http://www.sbsbio.com/cuxiao/DNAziliao1.pdf.

[30] Antkowiak PL, Lietard J, Darestani MZ, et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction [J]. Nature communications, 2020, 11(1): 5345.

[31] Heckel R, Mikutis G, Grass RN. A Characterization of the DNA Data Storage Channel [J]. Scientific reports, 2019, 9(1): 9663.

[32] Masaki Y, Onishi Y, Seio K. Quantification of synthetic errors during chemical synthesis of DNA and its suppression by non-canonical nucleosides [J]. Scientific reports, 2022, 12(1): 12095.

[33] Cline J, Braman JC, Hogrefe HH. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases [J]. Nucleic Acids Research. 1996, 24(18): 3546-3551.

[34] Lubock NB, Zhang D, Sidore AM, et al. A systematic comparison of error correction enzymes by next-generation sequencing [J]. Nucleic Acids Research. 2017, 45(15): 9206-9217.

[35] Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing[J]. Nature Reviews Genetics. 2014, 15(1): 56-62.

[36] Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR) [J]. Journal of Bioscience and Bioengineering. 2003, 96(4): 317-323.

[37] Akbari M, Hansen MD, Halgunset J, et al. Low copy number DNA template can render polymerase chain reaction error prone in a sequence-dependent manner [J]. The Journal of Molecular Diagnostics. 2005, 7(1): 36-39.

[38] Gimpel AL, Stark WJ, Heckel R, et al. A digital twin for DNA data storage based on comprehensive quantification of errors and biases [J]. Nature communications, 2023, 14(1): 6026.

[39] Gray J, Van Ingen C. Empirical measurements of disk failure rates and error rates [J/OL]. ArXiv,2007.abs/cs/0701166.

[40] National Human Genome Research Institute. Frameshift Mutation [EB/OL]. (2023-12-29) [2023-12-30]. https://www.genome.gov/genetics-glossary/Frameshift-Mutation.

[41] 平质, 张颢龄, 陈世宏, 等. Chamaeleo：DNA 存储碱基编解码算法的可拓展集成与系统评估平台 [J]. 合成生物学, 2021, 2(3): 412-427.

Ping Z, Zhang HL, Chen SH, et al. Chamaeleo: an integrated evaluation platform for DNA storage [J]. Synthetic Biology Journal, 2021, 2(3): 412-427.

[42] Song L, Geng F, Gong ZY, et al. Robust data storage in DNA by de Bruijn graph-based de novo strand assembly [J]. Nature Communications. 2022, 13(1): 5361.

[43] GenScript. Precise synthetic oligo pools [EB/OL]. [2021-02-01].

[44] Yuzuki D. A 2022 NGS Cost and Throughput Comparison [EB/OL]. (2022-03-20) [2023-10-31]. https://www.yuzuki.org/a-2022-ngs-cost-and-throughput-comparison/

[45] Organick L, Chen YJ, Dumas Ang S, et al. Probing the physical limits of reliable DNA data retrieval [J]. Nature communications, 2020, 11(1): 616.

[46] McCallum JC. Disk drive prices 1955+ [EB/OL]. (2023-05-07) [2023-10-31]. https://jcmit.net/diskprice.htm

[47] Cheng JY, Chen HH, Kao YS, et al. High throughput parallel synthesis of oligonucleotides with 1536 channel synthesizer [J]. Nucleic acids research, 2002, 30(18): e93.

[48] LeProust EM, Peck BJ, Spirin K, et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process [J]. Nucleic acids research, 2010, 38(8): 2522–2540.

[49] 江湘儿, 王勇, 沈玥. DNA 合成技术与仪器研发进展概述 [J]. 集成技术, 2021, 10(5): 80-95.

Jiang XE, Wang Y, Shen Y. The Review of DNA Synthesis Technologies and Instruments Development [J]. Journal of Integration Technology, 2021, 10(5): 80-95.

[50] Illumina, Inc. Illumina sequencing platforms [EB/OL]. [2023-10-31]. https://www.illumina.com/systems/sequencing-platforms.html

[51] Folio Photonics. Futureproof Data Storage [OL]. [2023-10-31]. https://foliophotonics.com/product

[52] FUJIFILM Corporation. Optical Disk Storage vs. Tape Storage [EB/OL]. [2023-10-31]. https://www.tape-storage.net/en/storage_comparison/article_02/

[53] BestBuy. USB Flash Drives [EB/OL]. [2023-10-31]. https://www.bestbuy.com/site/hard-drives/usb-flash-drives/abcat0504010.c?id=abcat0504010&i

ntl=nosplash

[54] Orlando L, Ginolhac A, Zhang GJ, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse [J]. Nature, 2013, 499(7456): 74–78.

[55] Zhirnov V, Zadegan RM, Sandhu GS, et al. Nucleic acid memory [J]. Nature materials, 2016, 15(4): 366–370.

[56] Zhang Y, Wang F, Chao J, et al. DNA origami cryptography for secure communication [J]. Nature Communications. 2019, 10(1): 5469.

[57] 张翼飞, 王海华, 赵椰, 等. 一种基于探针封闭与解封的 DNA 杂交信息存储加密方法: 中国, CN113539363A [P]. 2022-06-17.Zhang YF, Wang HH, Zhao Y, et al. DNA hybridization information storage encryption method based on probe blocking and unblocking: China, CN113539363A [P]. 2022-06-17.

[58] 张翼飞, 王海华, 肖祖颖, 等. 一种基于双探针特异性分离的 DNA 杂交信息存储加密方法: 中国, CN113345517A [P]. 2022-07-29.

Zhang YF, Wang HH, Xiao ZY, et al. DNA hybridization information storage encryption method based on double-probe specific separation: China, CN113345517A [P]. 2022-07-29.

[59] 王海华, 张翼飞, 肖祖颖, 等. 一种基于编码链发卡结构添加与移除的 DNA 杂交信息存储加密方法: 中国, CN113539379A [P]. 2022-07-08

Wang HH, Zhang YF, Xiao ZY, et al. DNA hybridization information storage encryption method based on addition and removal of coding chain hairpin structure: China, CN113539379A [P]. 2022-07-08.

[60] 陈非, 卜东波, 马灌楠, 等. DNA 活字存储系统和方法: 中国, CN111858510A [P]. 2020-10-30.

Chen F, Bu DB, Ma GN, et al. DNA movable type storage system and method: China, CN111858510A [P]. 2020-10-30.

[61] Gong ZY, Song LF, Pei GS, et al. Engineering DNA Materials for Sustainable Data Storage Using a DNA Movable-Type System [J]. Engineering, 2023, 29(10): 130-136.

[62] Roquet N, Bhatia SP, Flickinger SA, et al. DNA-based data storage via combinatorial assembly [EB/OL]. bioRxiv, 2021 (2021-04-20) [2023-12-30]. https://doi.org/10.1101/2021.04.20.440194.

[63] National Human Genome Research Institute. Polymerase Chain Reaction (PCR) Fact Sheet [EB/OL]. (2020-08-17) [2023-10-31]. https://www.genome.gov/about-genomics/fact-sheets/Polymerase-Chain-Reaction-Fact-Sheet

[64] Lee H, Popodi E, Tang H, et al. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing [J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(41): E2774-E2783.

[65] Nguyen HH, Park J, Park SJ, et al. Long-Term Stability and Integrity of Plasmid-Based DNA Data Storage [J]. Polymers, 2018, 10(1): 28.

[66] 杨怀义, 范婷文, 吴迪, 等. 一种分布式阵列存储及基于微生物的高容量纠错 DNA 存储技术（Bio-RAID）: 中国, CN114927169A [P]. 2022-08-19.

Yang HY, Fan TW, Wu D, et al. Distributed array storage and microorganism-based high-capacity error correction DNA storage technology (Bio-RAID): China, CN114927169A [P]. 2022-08-19.

[67] 刘翟, 张翼飞, 贾李佳, 等. 一种基于寡核苷酸芯片杂交的信息存储方法: 中国,

CN114093401A [P]. 2022-02-25.

Liu D, Zhang YF, Jia LJ, et al. Information storage method based on oligonucleotide chip hybridization: China, CN114093401A [P]. 2022-02-25.

[68] Ng CCA, Tam WM, Yin H, et al. Data storage using peptide sequences [J]. Nature Communications. 2021, 12(1): 4242.

[69] Green MR, Sambrook J. Polyacrylamide Gel Electrophoresis [J]. Cold Spring Harbor protocols. 2020, 2020(12).

[70] Beverly M, Hagen C, Slack O. Poly A tail length analysis of in vitro transcribed mRNA by LC-MS [J]. Analytical and bioanalytical chemistry, 2018, 410(6): 1667–1677.

[71] Hofstadler SA, Sannes-Lowery KA. Applications of ESI-MS in drug discovery: interrogation of noncovalent complexes [J]. Nature reviews. Drug discovery, 2006, 5(7): 585–595.

[72] Guo XH, Bruist MF, Davis DL, et al. Secondary structural characterization of oligonucleotide strands using electrospray ionization mass spectrometry [J]. Nucleic Acids Research, 2005, 33(11): 3659–3666.

[73] Xu CT, Ma B, Gao ZL, et al. Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage [J]. Science advances, 2021, 7(46): eabk0100.