特约评述

DOI: 10.12211/2096-8280.2021-001

DNA信息存储：生命系统与信息系统的桥梁

韩明哲[1,2]，陈为刚[1,3]，宋理富[1,2]，李炳志[1,2]，元英进[1,2]
（[1]天津大学，合成生物学前沿科学中心，系统生物工程教育部重点实验室，天津 300072；[2]天津大学化工学院，天津 300072；[3]天津大学微电子学院，天津 300072）

摘要：DNA信息存储通过编解码、合成、编辑和测序等过程，实现数字信息写入、存储与读出。其在密度、寿命、能耗和抗电磁干扰等方面较磁、光、电等常规的信息存储介质有较大优势。随着全球数据总量的快速增长，DNA信息存储的优势特性和发展潜力受到了研究者的广泛关注。本文阐述了DNA信息存储的基本原理和技术流程，分析了DNA信息存储与生命系统和信息系统的关联，并依据读写技术特点归纳近年来涌现的“DNA硬盘”“DNA光盘”“DNA磁带”等几种主要模式、发展现状及技术路线。在此基础上，探讨DNA信息存储商业化、大规模应用面临的主要挑战，讨论更低成本的数据写入和更快速的数据读出，并指出可行的发展路线。最后，展望了DNA作为新型存储介质在现代存储系统中的发展演化趋势。

关键词：合成生物学；数字信息存储；DNA合成；DNA测序；信息编码

中图分类号：Q819　　**文献标志码：**A

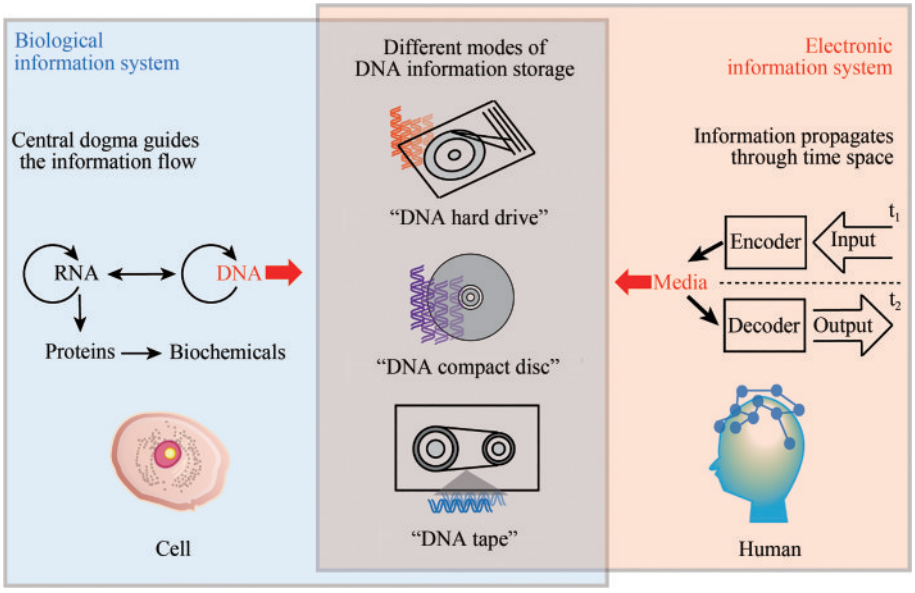
DNA information storage: bridging biological and digital world

HAN Mingzhe[1,2], CHEN Weigang[1,3], SONG Lifu[1,2], LI Bingzhi[1,2], YUAN Yingjin[1,2]
([1]*Frontier Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China*; [2]*School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China*; [3]*School of Microelectronics, Tianjin University, Tianjin 300072, China*)

Abstract: The external preservation of information enables reliable inheritance of human thoughts, playing important roles in the progress of human civilization. Starting from tying knots in ropes to storing data in magnetic and optical media, these technologies have documented and will continue to record the splendid civilization. However, driven by the global digitalization, the global data volume is growing rapidly and challenging the storage capability of existing storage technologies. DNA, as the natural carrier of genetic information, is believed to be a potential candidate to deal with the data storage challenge due to the revealed high density, long-term duration and low maintaining cost features. In this review, we first describe the fundamental principles and technical processes of DNA information storage. The pivotal position of DNA information storage bridging the biological and digital world is also pointed out. Then, according to the different characteristics of data writing and

收稿日期：2021-01-04 修回日期：2021-03-02
引用本文：韩明哲，陈为刚，宋理富，李炳志，元英进. DNA信息存储：生命系统与信息系统的桥梁[J]. 合成生物学，2021，2(3)：309-322
Citation: HAN Mingzhe, CHEN Weigang, SONG Lifu, LI Bingzhi, YUAN Yingjin. DNA information storage: bridging biological and digital world[J]. Synthetic Biology Journal, 2021, 2(3): 309-322

reading, we categorize these technologies into three storage modes, termed as "DNA hard drive", "DNA compact disc" and "DNA tape", by analogy with the popular storage media correspondingly. "DNA hard drive" mode shows the potential in the volume enlargement of the existing information storage system using oligonucleotide pools. "DNA compact disc" mode provides direct *in vivo* processing on DNA data storage enabling massive data distribution at low cost. "DNA tape" mode provides intracellular information recoding solutions, which may promote the future developments of cellular computing and communication. The up-to-date progress of these three modes is also summarized. We then discuss the main obstacles and potential technical routes towards practical applications of DNA information storage. We envision a cheaper, faster DNA information storage technology, and its appropriate integration with information storage systems in the future. Finally, we conclude that DNA information storage is a cutting-edge interdisciplinary technology and hope this review can bring more focus and research efforts from various fields to DNA information storage.



Keywords: synthetic biology; digital information storage; DNA synthesis; DNA sequencing; information encoding

信息存储是文明传承的基础。人类是地球上最具智慧的生命体，从结绳记事开始，生命体外的数据存储就成为了人类思想的延续，记录了灿烂文明。造纸与印刷术的发明，使得人类能够存储的数据量在几百年内获得了大约5个数量级的提升[1]；在计算机时代，尤其是近年来随着信息技术的快速发展，人类生活的方方面面都逐渐实现数字化转变，人类产生的数据爆发式增长。基于磁、光及集成电路的现代数据存储介质历经发展，存储体积密度已经可达到10^{10}～10^{12} bit/cm^3[2]。与之相比，DNA存储具有更高密度存储潜力，如大肠杆菌染色体DNA的存储体积密度据估算达约10^{19} bit/cm^3[3]。近年来，随着合成生物学的快速发展，以高通量DNA合成技术[4]和人工合成染色体的工作为代表[5-6]，标志着人类对DNA的设计[7]、合成[8]、编辑[9]和读取[10]能力已经进入到一个崭新的时代。在此背景下，利用合成DNA进行高密度信息存储成为一个非常有前景的研究方向[11]，得到了相关领域研究者、信息技术企业与生物科技企业的广泛关注。2020年11月，微软、西部数据等传统信息技术企业与Twist Bioscience、Illumina等新兴生物技术公司一道，共同宣布成立了第一个DNA数据存储联盟，将制定全面的行业路线图，为经济高效的商业档案存储奠定基础[12]。

1 DNA存储数字信息

利用人工合成的脱氧核糖核酸（DNA）存储数字信息，简称DNA信息存储[13]。DNA用作信息存储载体，具有存储高密度、不受电磁干扰、长期高可靠和维护低成本等优势[13-16]。DNA作为天然的信息载体，以“A/T/C/G”数字信号的表示形式，存储了亿万年来无数生物的遗传信息，依托中心法则造就生命繁衍、进化演化及生物多样性。人类产生的海量信息，记录在各类数字存储介质，保存并得以延续，支撑了文明的传承与繁荣。利用DNA存储数字信息连通了生物系统与信息系统，发展了多种应用模式，成为近年重要的研究热点。

利用DNA存储数字信息的原理和技术流程如图1所示。其原理是：数字化信息在二进制码流、四进制碱基序列和实际DNA片段之间的转化与流动[3, 13-14]。目前，基于此原理的技术流程主要包含两个方面：①信息写入，首先对文本、图片或视频等信息的二进制码流进行编码，得到A/T/C/G组成的碱基序列，随后利用DNA合成技术将信息写入对应的DNA片段，并对其进行多模式保存[17-18]；②信息读取，首先对制造的数据DNA片段进行测序，随后进行识别、组装、纠错与解码等，将存储在DNA介质中的数据还原为原始数字化信息，得到原始文本、图片、声音和视频等。

2 DNA信息存储的若干模式

依据DNA片段读写技术的特点，类似传统数据存储，也可划分为“硬盘”“光盘”“磁带”等应用模式。“DNA硬盘”具有高通量读写特征，面向海量数据的高密度存储；“DNA光盘”具有低成本快速复制特征，支持单写多读，面向数据的海量分发；“DNA磁带”具有体内串行刻写特征，面向数据或状态的顺时间记录。以下将对各个存储模式的特点和相关研究进展进行详细介绍。

2.1 “DNA硬盘”模式

2012年哈佛大学George Church等在《科学》

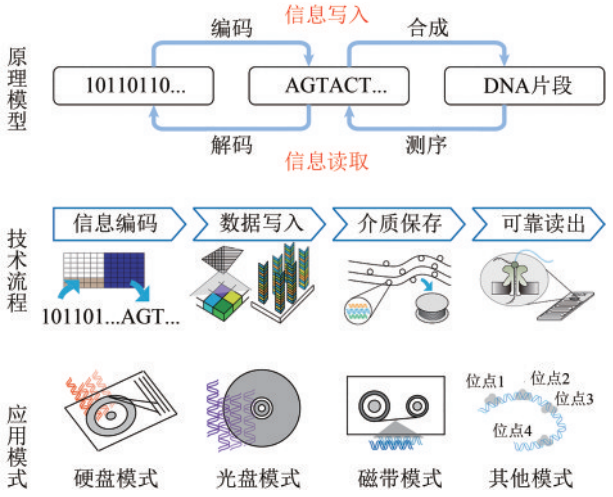


图1 DNA存储的原理模型、技术流程和应用模式

Fig. 1 The basic principle, technical work flow and storage modes of DNA information storage

(The basic principle is the conversion of digital information between binary code stream, quaternary base sequence and actual DNA fragment. The technical work flow includes Information encoding, data writing, media storage and reliable reading. Storage modes include "DNA hard drive", "DNA CD", "DNA tape" and others)

杂志发表研究成果[19]，成功存储和读取了5.27 Mb包含文字、图像和JavaScript程序的数字化信息，出错率仅为百万分之二。随后在*Johns Hopkins Magazine*上首次提出“DNA硬盘”（DNA hard drive）[20]。该模式依托高通量DNA芯片合成技术和高通量二代测序技术来写入和读出数据。与传统的硬盘类似，具有面向海量数据的高密度存储潜质。由此衍生的类似研究，可归纳为“DNA硬盘”。

“DNA硬盘”的数据端到端可靠性远不及传统硬盘，需要解决DNA作为载体的数据可靠性问题[21]。目前商业硬盘的读写错误率低至10^{-15}以下，而高通量合成寡核苷酸的错误率一般在1/2000到1/200[22-23]，二代测序的错误率在1/1000到1/100[24]。为了解决这些错误对信息可靠性的影响，多个信息领域的信息编码方法被引入到了“DNA硬盘”框架。欧洲分子生物学实验室的Goldman教授[25]通过添加四倍冗余和简单的校验机制实现了数据的可靠恢复，但是由于四倍冗余的设计，该方法实现的逻辑密度（bit/nt）和成本控制都不理想。苏黎世联邦理工大学Grass团队[26]引入了里德-所罗门（RS）纠删码，解决了寡核苷

酸链池中部分片段丢失以及片段内碱基替代错误，在保证数据可靠恢复的同时使数据部分的逻辑密度超过了1 bit/nt。Erlich等[27]引入了喷泉码，更好地适配海量片段化的存储模式，将数据部分的逻辑密度进一步提升到1.57 bit/nt。另一思路，Anavy等[28]和Choi等[29]分别使用了简并碱基来拓展DNA的多进制表示方法，将“硬盘”模式下的逻辑密度推升到了2 bit/nt以上，但是此方法也面临需要更高测序覆盖度（覆盖度 > 150×）的问题。除此之外，在未来引入非天然碱基拓展存储单元，可进一步提升逻辑密度[30]。总而言之，在确保数据可靠性的前提下，逼近数据承载能力的极限是DNA信息存储发展的趋势[31]。

值得关注的是，“DNA硬盘”中合成与测序会引入碱基的插入和缺失错误（insertion/deletion，简称Indel），这有别于传统存储介质，处理较为困难[3]。针对该问题，Press等[32]提出了基于哈希编码和贪婪穷举解码的编码方案，该方案能够在单分子拷贝的情况下纠正插入和缺失错误，但是需要较高的冗余度来实现纠错，且解码复杂度较高。Sabary等[33]提出了几种动态的DNA重构算法，可直接用于较高错误率下的DNA序列重建。天津大学Song等[34]设计了一个基于德布莱英图（de Bruijn Graph）的DNA序列高鲁棒重建算法，如图2所示，可以从包含大量插入缺失和替代错误的多序列快速重建无错误的DNA片段序列。该方法可以从低质量的PCR产物（序列长度完全错误）中可靠地读取数据，实现高鲁棒读取。

为降低“DNA硬盘”写入成本，提高写入速度，2019年，Lee等[35]采用非阻断型的末端脱氧核酸转移酶（TdT）合成DNA，实现了一种专用于信息存储的DNA酶法合成技术。2020年，Lee等[36]进一步利用图案化紫外光快速解离Co^{2+}激活TdT，成功编码了110位的数据信息，初步验证了在阵列表面实现大规模DNA并行合成的可行性。

为解决“DNA硬盘”多轮PCR造成的偏好性累积和部分DNA片段丢失的问题，Lin等[37]通过对原始文库修饰并引入RNA逆转录过程，构建了

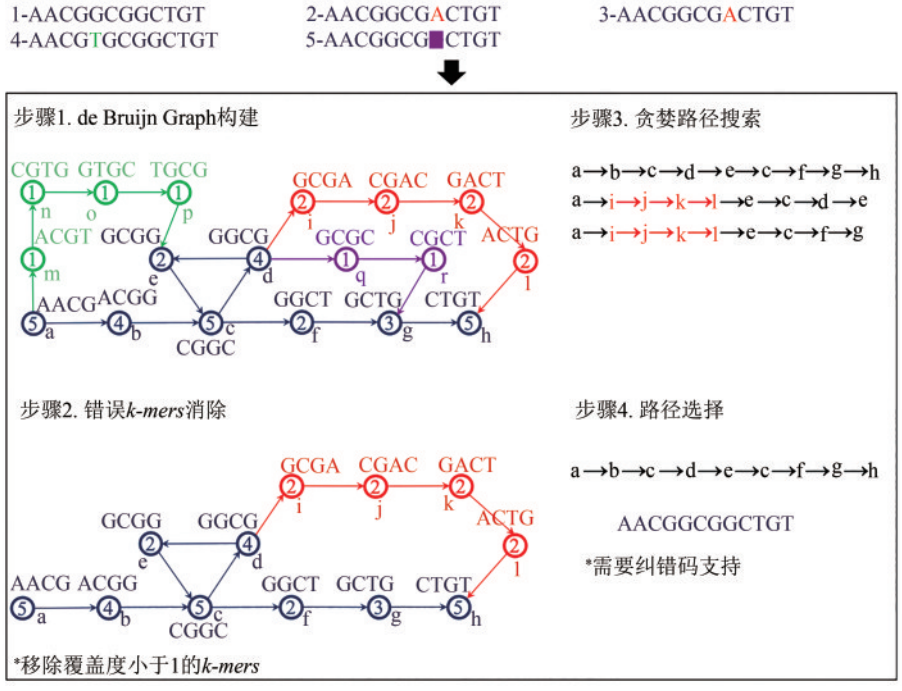


图2 基于de Bruijn图论的DNA序列重建算法[34]

Fig. 2 Algorithm of de Bruijn graph-based reconstruction of DNA strands[34]

(There are four major steps in this algorithm: Step 1—Construct the de Bruijn graph using the sequencing results; Step 2—Eliminate the noisy *k-mers* by removing the low coverage *k-mers*; Step 3—Greedy path search with the simplified de Bruijn graph to reveal all possible paths associated with specific indexes; Step 4—Select the correct paths, i.e. the correct strands, based on the embedded EC codes)

始终以原始文库为模板的扩增方法，在一定程度上降低了多次访问对原始文库的影响。Choi等[38]将原始文库固定在具有二维码编号的微盘上，实现了对文库的原位（*in situ*）扩增，经过20轮扩增未发现产物片段分布的明显变化，显著降低了扩增带来的偏好性，同时还通过二维码实现了数据库管理。天津大学Gao等[39]将原始文库固定在磁珠上，通过等温链置换扩增技术，实现了对文库低偏好性、稳定重复的扩增。

“DNA硬盘”的应用模式已实现了一定规模的存储验证[40-47]。2018年，华盛顿大学和微软公司的研究团队实现了200 MB的数据存储和部分数据文件的随机访问[40]，并于2019年开发了原型设备，实现了“HELLO”的自动读写[41]，同时还设计了DNA保存和访问的微流控平台[42]；2019年，美国Catalog公司[43]利用独创的DNA写入技术，存储了16 GB的维基百科数据，是目前最大规模的“DNA硬盘”。在国内，天津大学陈为刚等[44]采用LDPC码与RS码的乘积码保证可靠性，采用27万条的寡核苷酸池存储超过3 MB数据，存储了两段有历史价值的音视频片段以及13 000多汉字，实现了低样本浓度、低测序覆盖度的可靠读出（图3）。深圳华大生命科学研究院Ping等[45]设计的“阴-阳”编码策略可调整均聚物长度或GC含量等以满足不同用户需求，实现了2.02 MB数据的存储。

2.2 “DNA光盘”模式

与“DNA硬盘”的体外存储方式不同，一种生命体内的DNA信息存储模式也被提出，其特征类似光盘，本文归纳为“DNA光盘”[48]。该模式的主要特征是采用较长DNA片段，通过细胞体内组装完成写入、借助细胞自身的快速低成本的DNA复制能力，快速且均一拷贝数据。虽然“CD母版”的制作成本较高，即合成与组装成本较高，但是其类似CD的低成本大量拷贝，使得“母版”成本得以分摊。受益于常用模式生物较低的突变率[49-50]，“DNA光盘”亦可高保真拷贝，支持数据长期传代复制[51]。利用小型纳米孔测序器件，有望实现数据快速读出，便携式“DNA光驱”呼之欲出。值得注意的是，纳米孔测序错误率高达10%，并且包含难以处理的插入与缺失错误[52]。因此如何保证数据在纳米孔测序下的可靠读出，是一个值得研究的方向。

“DNA光盘”开始于早期细胞体内存储数字信息的概念验证，探索单个细胞内存储的数据量是个有价值的问题。概念验证多使用质粒在大肠杆菌内存储数据，编码的DNA长度通常不超过1 kbp[53-59]。2010年，Venter等[60]在化学合成蕈状支原体时，第一次在原核生物基因组中嵌入了超过4 kbp的编码DNA存储外部信息。本文作者[48]从头设计合成了一条254 886 bp的存储专用染色体，其中数据编码部分占95.27%，将单菌内数据存储DNA数量提升到了百kbp级，存储了37.8 KB图片、视频以及文字，利用叠加编码方案，有效克服三代测序的高错误率问题，实现了数据的可靠恢复。这项工作突破性地将单菌内数据存储DNA数量提升到百kbp级，初步打通了单细胞数据存储容量这个限制“DNA光盘”模式存储通量提升的关键因素（图4）。

“DNA光盘”模式除了提高单细胞数据容量外，增加并行通量也是提升数据存储容量的关键。Shipman等[61]通过CRISPR/CAS1-CAS2系统捕捉DNA小片段整合进大肠杆菌群体的CRISPR序列中，分别编码了494字节的21色图片和2.6 KB的动画短片。天津大学Hao等[62]构建了携带不同短信息片段质粒的大肠杆菌分布式混菌存储系统，在维持低成本的同时实现较大的体内存储通量，将445 KB的数字文件存储在11 520个115 bp的合成DNA中。

2.3 “DNA磁带”及其他模式

运用动态基因组工程（dynamic genome engineering）[63]在生命体内“书写”DNA来记录信息的新模式，一定程度上类似磁带，本文称之为“DNA磁带”。“书写”包括对特定DNA靶向插入、删除、倒位和单碱基突变等操作，类似于在磁带上磁化刻录以记录信息[64]。目前已经验证的模型中，“书写”过程的开启信号可以是对抗生素或病毒的暴露、营养底物的改变和对光及特定诱

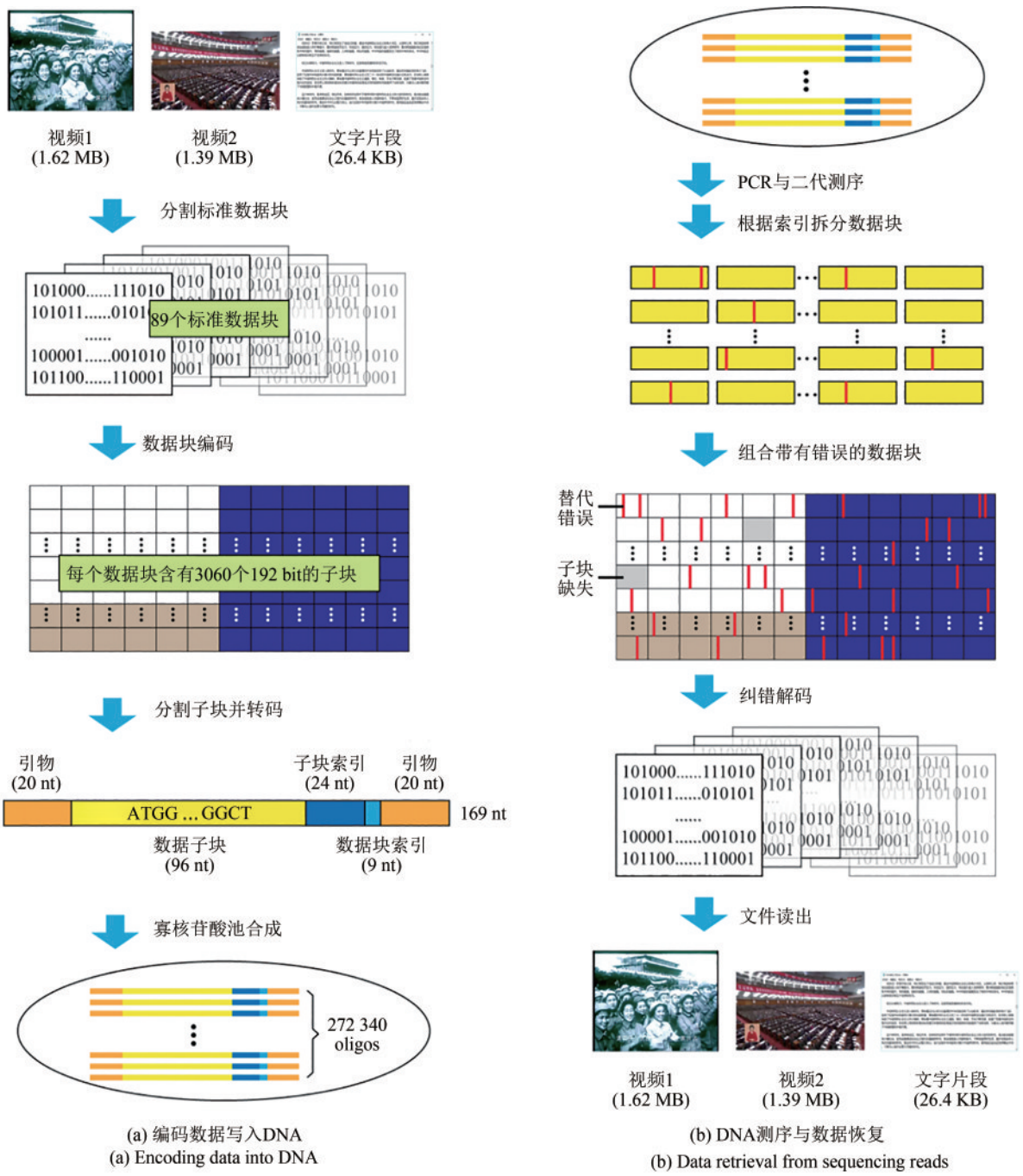


图3 “DNA硬盘”模式示意图[44]

Fig. 3 Schematic diagram of "DNA hard drive"[44]

[(a) Encoding data into DNA. The original files were converted into 89 standard data blocks and then divided into sub-blocks after error correction coding. Each 192-nt sub-block was transcoded into a 96-nt DNA sequence, and then the sub-block index, block index, and primers are added to form an oligonucleotide structure. (b) Data retrieval from sequencing reads. After PCR and 2nd-generation sequencing of synthesized oligo pool, data blocks with substitutions and erasures were obtained from sequencing reads according to indexes. With the help of error correction coding, original files were finally recovered]

导剂的响应等[65-69]。起初“DNA磁带”主要记录细胞内的特定事件或状态，Harries Wang团队[70]首次构建了基于电刺激的“人-胞”输入接口，利用电压控制胞内的氧化还原对状态，从而诱导CRISPR/Cas1-Cas2系统在特定位点插入不同的DNA序列，实现信息写入。这使得未来半导体-生

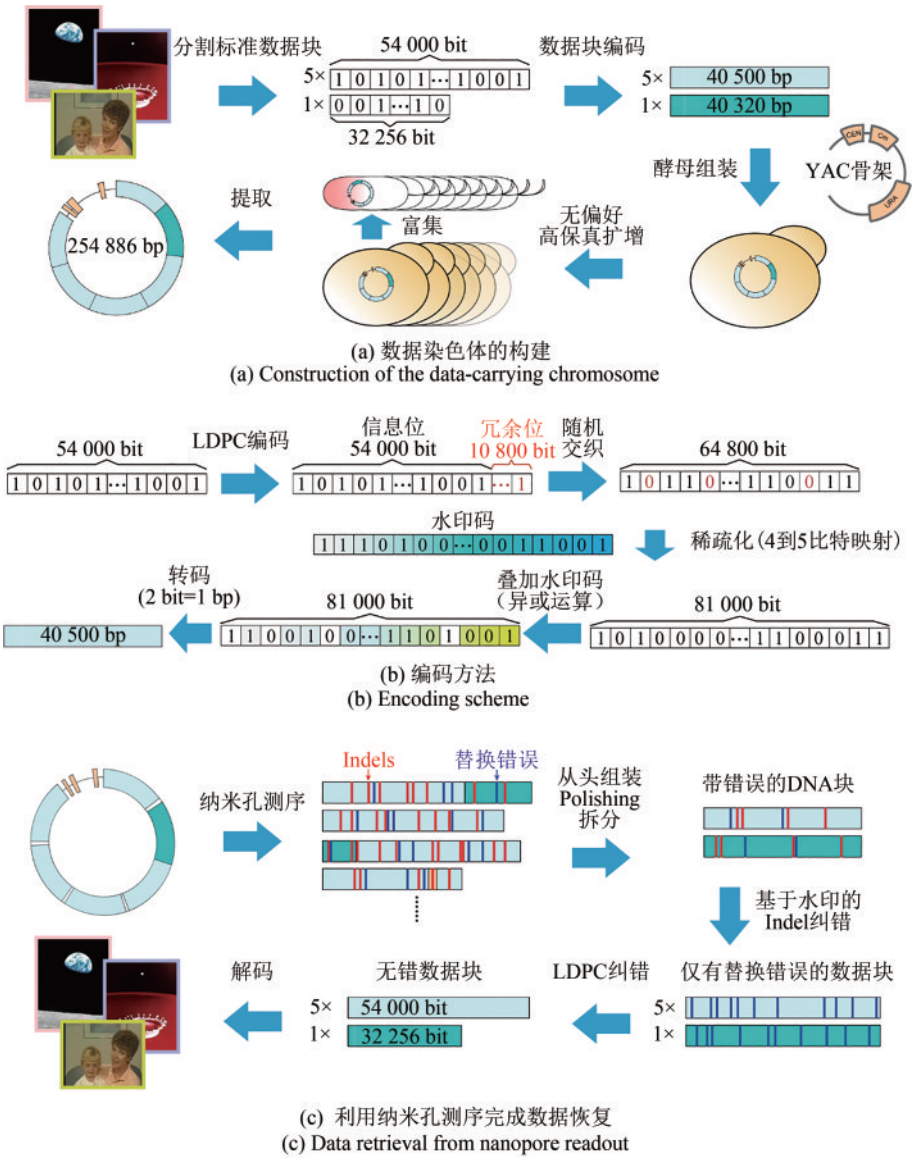


图4 “DNA光盘”模式示意图[48]

Fig. 4 Schematic diagram of "DNA CD"[48]

[(a) Construction of the data-carrying chromosome. The DNA sequences encoded from original files were synthesized and *in vivo* assembled into a 254 886 bp chromosome. (b) Encoding scheme. The encoding process included LDPC coding, random interleaver, sparsification, superposition with watermark sequences, and transcoding into DNA sequences. (c) Data retrieval from nanopore readout. The data retrieval process included *de novo* assembly, polishing, locating, indel correction with watermark sequences, and substitution error correction with the LDPC code]

物接口的发展成为了可能。进一步，得益于基因线路设计的发展，生物“逻辑门”可与“DNA磁带”相结合，为生物细胞计算提供记录。然而，“DNA磁带”依然存在逻辑密度低、数据响应延迟和精准性较低等问题。此外，目前通常是基于菌群进行记录，通过加标签（barcode）对不同菌群进行区分[70]，随机访问的难度较大。

与“DNA磁带”模式类似，为避免人工合成DNA产生的高昂成本，美国UIUC的Tabatabaei等[71]模仿古老的打孔卡存储方式，以天然的DNA分子链（例如基因组DNA、克隆或PCR扩增产物）为“卡纸”，以特定的酶为“打孔机”，建立了一种“打孔卡”DNA存储方法。该方法通过在DNA磷酸骨架上预设位置“打孔”来表示二进制数据中的“0”和“1”，从而避免了昂贵的DNA合成。与之相似，以天然M13噬菌体单链DNA为骨

架，Chen等[72-73]在骨架上间隔插入带有生物素标记的支链DNA用以记录信息，并通过纳米孔测序检测是否带有标记物来读取数据的“0”和“1”。然而，这种基于天然DNA分子链的存储技术没有发挥DNA存储密度大的优势。

除此之外，华盛顿大学和微软公司的研究团队[74]也尝试了对组装后的寡核苷酸池进行纳米孔测序。上海交通大学Zhang等[75]利用DNA折纸技术实现信息的加解密，这种基于结构的信息表示和加密方法，为保证重要信息的安全性提供了新的方案。

3 挑战与展望

当前DNA信息存储的主要挑战为单位信息存储成本高，信息读写速度慢，无法高效对接现有信息系统。因此，DNA信息存储当前发展的重点是进一步降低成本，提高读写速度，实现与现有信息系统的融合。

3.1 更低成本的信息写入

目前，寡核苷酸池的商业合成价格大约为0.002美元/base，折合0.001美元/bit（约8.6×10^{6}美元/GB）[23, 76]，写入成本较高，是硬盘的10^{8}倍[77]，如图5所示。美国情报高级研究计划局（IARPA）分子信息存储技术（MIST）项目的目标是到2023年DNA信息写入成本将降低至10^{-10}美元/bit（约0.86美元/GB）[78]。

DNA信息存储成本在未来有很大下降的潜力。首先，Twist Bioscience的首席技术官在2016年声称其合成成本已经低于10^{-12}美元/base[79]。但是，运行维护、合成芯片、试剂耗材、质量控制以及人工等其他成本造成了现有DNA信息写入成本较高的现状。可以从优化合成反应、改良芯片结构、替换廉价耗材、优化试剂分配量等多方面着手，有望大幅降低合成成本。其次，传统上DNA合成主要用于生命科学研究，其技术指标与DNA信息存储的需求不匹配。面向DNA信息存储的合成，可容忍合成步骤产生的更多错误，降低精度与纯度要求，减少质量控制成本，在保证数据准确性而不是序列准确性的基础上提升合成的长度和通量，从而有望大幅降低合成成本[80]。再者，由于信息存储领域市场规模巨大，随着半导体器件、微纳加工在DNA信息存储领域的应用，该领域的巨大投入将对DNA合成技术产生重大影响，DNA合成技术与装备快速迭代升级，合成通量快速提升，成本有望快速下降。

3.2 更快速的数据读取

DNA信息存储的读取依赖测序技术，与磁、光、电等存储相比，读取速度较慢，如图6所示。进一步提升读取速度，是DNA信息存储发展的一个需求。DNA的测序技术与现有电、磁存储技术的串行读取不同，具有高并行读取特点，以Illumina为代表的二代测序技术可以同时读取0.04亿～11亿个位点[81]。然而，每轮测序反应和

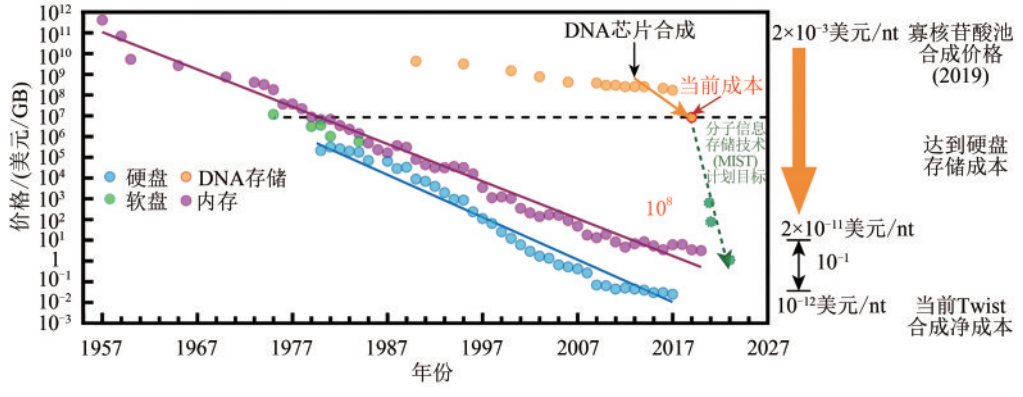


图5 DNA信息存储成本比较与预测

Fig. 5 Comparison and forecast of cost by DNA information storage

(Blue spots, HDD; green spots, floppy disk; purple spots, RAM; orange spots, DNA information storage)

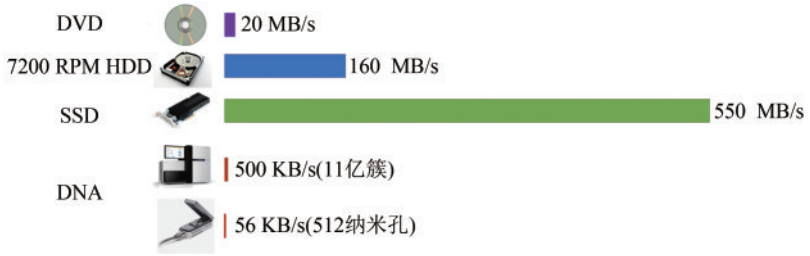


图6 DNA信息存储读取速度对比

Fig. 6 Comparison of reading rate for DNA information storage

信号采集时间长达2.2～19 min[82]，所有反应所耗时间约占运行时间的90%。通过高通量（也即空间并行度）弥补反应时间较慢的缺陷，读取速度可达5～500 KB/s[81]（最大数据产出/最长运行时间），但是需测序完全结束后才能获取原始数据。三代纳米孔测序已经做到便携化和低延迟数据生成，单通道测序速度约为450 bp/s（约112 B/s）[83]，基于MinION测序芯片（最多支持512通道同时读取）的最高读取速度约为56 KB/s（不包含电信号到碱基转换时间）。而现有电、磁存储技术通常每秒可读取几十到几百兆字节数据。基于二代测序的数据读取受化学反应限制，较难突破性地降低反应时间，可以通过进一步增大通量满足未来大规模冷数据读取需求；基于三代纳米孔测序的数据读取，依然有较大潜力提升单孔读取速度，如固相纳米孔的发展有望在保证分辨率的前提下继续提升读取速度1～3个数量级[84]，甚至在未来超越现有存储的读取速度。此外，提高并行化读取的集成程度，构建一体化、自动化的读取专用设备也面临很大挑战，需要机械、生化、信息、控制等的多学科协同解决。

3.3 DNA信息存储与现代存储系统的融合

依据DNA合成与读取的技术发展现状和特点，DNA信息存储有望率先在冷数据存储方面获得应用[85]。图7为DNA信息存储在开放系统互联（OSI）、模型中的映射关系以及存储系统分等级架构。DNA作为新介质，融入现代存储系统的过程，也是信息存储系统不断演化完善的过程。

在物理层，造成DNA数据存储不可靠的因素主要包括：合成、扩增以及测序处理过程的非理想，体现在碱基的插入、缺失、替代（IDS）错误以及DNA分子或片段丢失等[86]；按照信息理论研

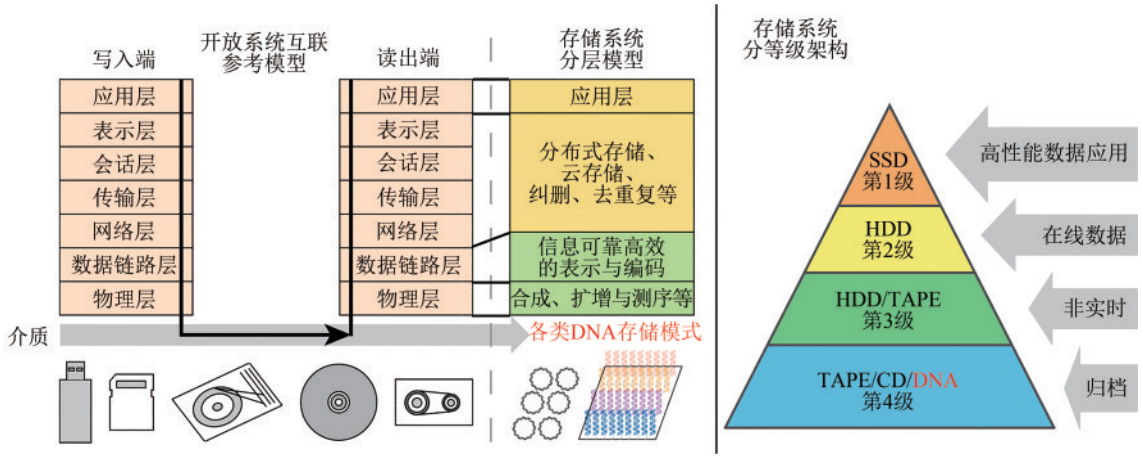


图7 DNA信息存储与现代存储系统的融合

Fig. 7 Fusion of DNA information storage and information storage system

(Synthetic DNA enriches the modern data storage media including SSD, HDD, CD, tape. The introduced new property by DNA requires not only suited low-level techniques including synthesis, amplification and sequencing, but also reliability guarantee schemes such as error correction codes, erasure correction, redundancy eliminating schemes, etc. New application paradigms also should be explored to dig out this new medium. The new medium changes all the elements in the 7 levels of OSI reference model and reforms the classic storage hierarchical architecture)

究范式，一旦建立了准确的碱基错误模型，就可以设计匹配的信息编码方法与数据恢复方法[31]，设计有效的数据链路层。但是，由于DNA信息存储信道的一些新特点，例如包含Indel错误、信道容量尚无法准确计算[87]，值得深入研究[13, 32, 88]。中间各层是DNA信息存储融入现代存储系统的桥梁。传统数据存储领域的关键技术，需要结合DNA介质与DNA存储的新特点进行优化设计。例如，目前纠删码已经在基于寡核苷酸池的信息存储模式得到了很好的应用[27, 40]。同时，纠删码也广泛应用于存储系统的中间各层，如何协调设计是一个非常有价值的问题。在应用层，提供的用户服务需要与DNA存储特点相适配[89]。例如，数据检索、聚类分析、数据挖掘、特征识别等，需要方便地读取数据，而现阶段DNA信息存储将大块数据封装于无法实时读取的DNA介质。因此，探索结合DNA信息存储特点的“存算一体化”的处理引擎，设计跨层的直达DNA介质的机制就显得极为重要。

存储系统的分等级架构是存储系统充分发挥作用的基础，DNA作为新的存储介质，短期内其技术特性与大容量冷数据归档存储最为匹配。据预测，归档的冷数据比例高达60%[90]，冷数据的DNA存储展现出了巨大的发展潜力，有望平稳融入现代数据存储体系。

值得一提的是，DNA信息存储也可能给传统信息系统带来安全方面的隐患。研究者可将计算机病毒信息存储于DNA，通过DNA测序以及处理过程，访问并进入非合作方的计算机系统，造成信息安全风险[91-92]。而DNA分子极小的物理尺度、特定条件下稳定的物理性质和无金属特征的非电/磁存储，为隐蔽数据传递提供了新途径。将携带信息的DNA封装为可打印材料，存储到常见的生活物品中并隐蔽传递[26, 93]，可能造成敏感数据泄露。

3.4 总 结

近年来，DNA信息存储的基本原理、技术流程和应用模式引起了研究者的广泛关注。DNA信息存储连接了生命系统与信息系统，推动相关研究与应用的发展。以“DNA硬盘”为主的体外存储与电子信息系统耦合更多，拓展了现有基于磁、光、电的电子信息存储系统；以“DNA光盘”和“DNA磁带”为主的体内存储与生命信息系统耦合度比较大，提供了细胞内的信息存储器或记录器，为未来细胞计算或细胞通信的发展提供了更广阔的空间。DNA信息存储是一个新兴的、多学科深度交叉融合的研究方向。进一步推动其走向实用化，仍面临很多挑战。为应对挑战，美欧的相关企业、大学与研究机构已经组成了DNA数据存储联盟，通过广泛合作共同制定全面的行业路线图，以推动DNA信息存储的产业化发展。据高德纳咨询公司预测，到2024年，将有30%的数字业务进行DNA存储试验[94]，以应对指数级增长的数据存储需求。面对未来的存储需求，国内也亟需布局和发展DNA信息存储研究与应用。本文从合成生物学与信息科学交叉融合的视角，对近年来DNA信息存储的研究进行了综述与展望，希望能吸引更多研究者在该交叉框架下提出有价值的研究问题，推动DNA信息存储的发展与应用。

参 考 文 献

[1] Semiconductor Industry Association. Decadal plan for semiconductors [EB/OL]. [2020-12-02]. https://www.semiconductors.org/wp-content/uploads/2020/10/SRC-SIA-Decadal-Plan-Webinar-Dec-2-2020.pdf.

[2] FONTANA R E, DECAD G M, HETZLER S R. Volumetric density trends (TB/in.3) for storage components: TAPE, hard disk drives, NAND, and Blu-ray[J]. Journal of Applied Physics, 2015, 117(17): 17E301.

[3] ZHIRNOV V, ZADEGAN R M, SANDHU G S, et al. Nucleic acid memory[J]. Nature Materials, 2016, 15(4): 366-370.

[4] KOSURI S, CHURCH G M. Large-scale *de novo* DNA synthesis: technologies and applications[J]. Nature Methods, 2014, 11(5): 499-507.

[5] WU Y, LI B Z, ZHAO M, et al. Bug mapping and fitness testing of chemically synthesized chromosome X[J]. Science, 2017, 355(6329): eaaf4706.

[6] XIE Z X, LI B Z, MITCHELL L A, et al. "Perfect" designer chromosome V and behavior of a ring derivative[J]. Science, 2017, 355(6329): eaaf4704.

[7] 陈欣懋, 欧阳颀. 生物逆向工程设计在合成生物学中的应用[J]. 合成生物学, 2020, 1(1): 29-43.
CHEN X M, OUYANG Q. The application of biological reverse engineering in synthetic biology[J]. Synthetic Biology Journal, 2020, 1(1): 29-43.

[8] 彭凯, 逯晓云, 程健, 等. DNA合成、组装与纠错技术研究进展[J]. 合成生物学, 2020, 1(6): 697-708.
PENG K, LU X Y, CHENG J, et al. Advances in technologies for *de novo* DNA synthesis, assembly and error correction[J]. Synthetic Biology Journal, 2020, 1(6): 697-708.

[9] 曹中正, 张心怡, 徐艺源, 等. 基因组编辑技术及其在合成生物学中的应用[J]. 合成生物学, 2020, 1(4): 413-426.
CAO Z Z, ZHANG X Y, XU Y Y, et al. Genome editing technology and its applications in synthetic biology[J]. Synthetic Biology Journal, 2020, 1(4): 413-426.

[10] 王会, 戴俊彪, 罗周卿. 基因组的"读-改-写"技术[J]. 合成生物学, 2020, 1(5): 503-515.
WANG H, DAI J B, LUO Z Q. Reading, editing, and writing techniques for genome research[J]. Synthetic Biology Journal, 2020, 1(5): 503-515.

[11] 丁明珠, 李炳志, 王颖, 等. 合成生物学重要研究方向进展[J]. 合成生物学, 2020, 1(1): 7-28.
DING M Z, LI B Z, WANG Y, et al. Significant research progress in synthetic biology[J]. Synthetic Biology Journal, 2020, 1(1): 7-28.

[12] Business Wire. Twist Bioscience, Illumina and Western Digital form alliance with Microsoft to advance data storage in DNA [EB/OL]. [2020-11-12]. https://www. businesswire. com/news/home/20201112005759/en/.

[13] CEZE L, NIVALA J, STRAUSS K. Molecular digital data storage using DNA[J]. Nature Reviews Genetics, 2019, 20(8): 456-66.

[14] MEISER L C, ANTKOWIAK P L, KOCH J, et al. Reading and writing digital data in DNA[J]. Nature Protocols, 2020, 15(1): 86-101.

[15] DONG Y, SUN F, PING Z, et al. DNA storage: research landscape and future prospects[J]. National Science Review, 2020, 7(6): 1092-1107.

[16] PING Z, MA D, HUANG X, et al. Carbon-based archiving: current progress and future prospects of DNA-based data storage[J]. GigaScience, 2019, 8(6): giz075.

[17] LIM C K, NIRANTAR S, YEW W S, et al. Novel modalities in DNA data storage[J]. Trends in Biotechnology, 2021. DOI: https://doi.org/10.1016/j.tibtech.2020.12.008.

[18] 周廷尧, 罗源, 蒋兴宇. DNA数据存储:保存策略与数据加密[J]. 合成生物学, 2021, 2(3): 371-383.
ZHOU T Y, LUO Y, JIANG X Y. DNA data storage: preservation approach and data encryption[J]. Synthetic Biology Journal, 2021, 2(3): 371-383.

[19] CHURCH G M, GAO Y, KOSURI S. Next-generation digital information storage in DNA[J]. Science, 2012, 337(6102): 1628.

[20] KEIGER D. DNA hard drive [EB/OL]. [2012-11-30]. https://hub.jhu.edu/magazine/2012/winter/dna-hard-drive/.

[21] 毕昆, 顾万君, 陆祖宏. DNA存储中的编码技术[J]. 生物信息学, 2020, 18(2): 76-85.
BI K, GU W J, LU Z H. Coding algorithms in DNA storage[J]. Chinese Journal of Bioinformatics, 2020, 18(2): 76-85.

[22] Twist Bioscience. Product sheet of Twist oligo pools [EB/OL]. [2019-08-29]. https://www.twistbioscience.com/sites/default/files/resources/2019-09/ProductSheet_OligoPools_29Aug19_Rev5.1.pdf

[23] GenScript. Precise synthetic oligo pools [EB/OL]. [2021-02-01]. https://www. genscript. com/precise-synthetic-oligo-pools. html.

[24] LIU L, LI Y, LI S, et al. Comparison of next-generation sequencing systems [J]. Journal of Biomedicine and Biotechnology, 2012, 2012: 251364.

[25] GOLDMAN N, BERTONE P, CHEN S Y, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA [J]. Nature, 2013, 494(7435): 77-80.

[26] GRASS R N, HECKEL R, PUDDU M, et al. Robust chemical preservation of digital information on DNA *in silica* with error-correcting codes[J]. Angewandte Chemie International Edition, 2015, 54(8): 2552-2555.

[27] ERLICH Y, ZIELINSKI D. DNA Fountain enables a robust and efficient storage architecture[J]. Science, 2017, 355(6328): 950-953.

[28] ANAVY L, VAKNIN I, ATAR O, et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters[J]. Nature Biotechnology, 2019, 37(10): 1229-1236.

[29] CHOI Y, RYU T, LEE A C, et al. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases[J]. Scientific Reports, 2019, 9(1): 6582.

[30] HOSHIKA S, LEAL N A, KIM M J, et al. Hachimoji DNA and RNA: a genetic system with eight building blocks[J]. Science, 2019, 363(6429): 884-887.

[31] HECKEL R, SHOMORONY I, RAMCHANDRAN K, et al. Fundamental limits of DNA storage systems[C]//2017 IEEE International Symposium on Information Theory (ISIT). Aachen, Germany: IEEE, 2017: 3130-3134.

[32] PRESS W H, HAWKINS J A, JONES S K, et al. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints[J]. Proceedings of the National Academy of Sciences of the United States of America, 2020, 117(31): 18489-18496.

[33] SABARY O, YUCOVICH A, SHAPIRA G, et al. Reconstruction Algorithms for DNA-Storage Systems[EB/OL]. [2021-05-25]. https://doi.org/10.1101/2020.09.16.300186.

[34] SONG L, GENG F, GONG Z, et al. Super-robust data storage in DNA by de Bruijn graph-based decoding[EB/OL]. [2021-05-25]. https://doi.org/10.1101/2020.12.20.423642.

[35] LEE H H, KALHOR R, GOELA N, et al. Terminator-free template-independent enzymatic DNA synthesis for digital information storage[J]. Nature Communications, 2019, 10(1): 2383.

[36] LEE H, WIEGAND D J, GRISWOLD K, et al. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage[J]. Nature Communications, 2020, 11(1): 5246.

[37] LIN K N, VOLKEL K, TUCK J M, et al. Dynamic and scalable DNA-based information storage[J]. Nature Communications, 2020, 11(1), 2981.

[38] CHOI Y, BAE H J, LEE A C, et al. DNA micro-disks for the management of DNA-based data storage with Index and Write-Once-Read - Many (WORM) memory features[J]. Advanced Materials, 2020, 32(37): 2001249.

[39] GAO Y, CHEN X, QIAO H, et al. Low-bias manipulation of DNA oligo pool for robust data storage[J]. ACS Synthetic Biology, 2020, 9(12): 3344-3352.

[40] ORGANICK L, ANG S D, CHEN Y J, et al. Random access in large-scale DNA data storage[J]. Nature Biotechnology, 2018, 36(7): 242-248.

[41] TAKAHASHI C N, NGUYEN B H, STRAUSS K, et al. Demonstration of end-to-end automation of DNA data storage[J]. Scientific Reports, 2019, 9(1): 4998.

[42] NEWMAN S, STEPHENSON A P, WILLSEY M, et al. High density DNA data storage library *via* dehydration with digital microfluidic retrieval[J]. Nature Communications, 2019, 10(1): 1706.

[43] SHANKLAND S. Startup packs all 16GB of Wikipedia onto DNA strands to demonstrate new storage tech [EB/OL]. [2019-07-02]. https://www.cnet.com/news/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/.

[44] 陈为刚, 黄刚, 李炳志, 等. 音视频文件的DNA信息存储[J]. 中国科学:生命科学, 2019, 50(1): 81-85.
CHEN W G, HUANG G, LI B Z, et al. DNA information storage for audio and video files[J]. SCIENTIA SINICA Vitae, 2019, 50(1): 81-85.

[45] PING Z, CHEN S, ZHOU G, et al. Towards practical and robust DNA-based data archiving by codec system named 'Yin-Yang'[EB/OL]. [2021-05-25]. https://doi.org/10.1101/829721.

[46] BORNHOLT J, LOPEZ R, CARMEAN D M, et al. Toward a DNA-based archival storage system[J]. IEEE Micro, 2017, 37(3): 98-104.

[47] YAZDI S M H T, YUAN Y, MA J, et al. A rewritable, random-access DNA-based storage system[J]. Scientific Reports, 2015, 5(1): 14138.

[48] CHEN W G, HAN M Z, ZHOU J T, et al. An artificial chromosome for data storage[J]. National Science Review, 2021, 8(5): nwab028.

[49] ZHU Y O, SIEGAl M L, HALL D W, et al. Precise estimates of mutation rate and spectrum in yeast[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(22): E2310-E2318.

[50] 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.

[51] SONG L F, ZENG A P. Orthogonal information encoding in living cells with high error- tolerance, safety, and fidelity [J]. ACS Synthetic Biology, 2018, 7(3): 866-874.

[52] DEAMER D, AKESON M, BRANTON D. Three decades of nanopore sequencing[J]. Nature Biotechnology, 2016, 34(5): 518-524.

[53] DAVIS J. Microvenus[J]. Art Journal, 1996, 55(1): 70-74.

[54] BANCROFT C, BOWLER T, BLOOM B, et al. Long-term storage of information in DNA[J]. Science, 2001, 293(5536): 1763-1765.

[55] WONG P C, WONG K K, FOOTE H. Organic data memory using the DNA approach[J]. Communications of the ACM, 2003, 46(1): 95-98.

[56] GUSTAFSSON C. For anyone who ever said there's no such thing as a poetic gene[J]. Nature, 2009, 458(7239): 703.

[57] YACHIE N, SEKIYAMA K, SUGAHARA J, et al. Alignment-based approach for durable data storage into living organisms[J]. Biotechnology Progress, 2007, 23(2): 501-505.

[58] AILENBERG M, ROTSTEIN O D. An improved Huffman coding method for archiving text, images, and music characters in DNA[J]. Biotechniques, 2009, 47(3): 747-751.

[59] NGUYEN H H, PARK J, PARK S J, et al. Long-term stability and integrity of plasmid-based DNA data storage[J]. Polymers,

2018, 10(1): 28.

[60] GIBSON D G, GLASS J I, LARTIGUE C, et al. Creation of a bacterial cell controlled by a chemically synthesized genome[J]. Science, 2010, 329(5987): 52-56.

[61] SHIPMAN S L, NIVALA J, MACKLIS J D, et al. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria[J]. Nature, 2017, 547(7663): 345-349.

[62] HAO M, QIAO H, GAO Y, et al. A mixed culture of bacterial cells enables an economic DNA storage on a large scale[J]. Communications Biology, 2020, 3(1): 416.

[63] AUSLÄNDER S, FUSSENEGGER M. Dynamic genome engineering in living cells[J]. Science, 2014, 346(6211): 813-814.

[64] FARZADFARD F, LU T K. Emerging applications for DNA writers and molecular recorders[J]. Science, 2018, 361(6405): 870-875.

[65] FARZADFARD F, LU T K. Genomically encoded analog memory with precise *in vivo* DNA writing in living cell populations[J]. Science, 2014, 346(6211): .

[66] SHETH R U, YIM S S, WU F L, et al. Multiplex recording of cellular events over time on CRISPR biological tape[J]. Science, 2017, 358(6369): 1457-1461.

[67] TANG W, LIU D R. Rewritable multi-event analog recording in bacterial and mammalian cells[J]. Science, 2018, 360(6385): eaap8992.

[68] PERLI S D, CUI C H, LU T K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells[J]. Science, 2016, 353(6304): aag0511.

[69] FARZADFARD F, GHARAEI N, HIGASHIKUNI Y, et al. Single-nucleotide-resolution computing and memory in living cells[J]. Molecular Cell, 2019, 75(4): 769-780.

[70] YIM S S, MCBEE R M, SONG A M, et al. Robust direct digital-to-biological data storage in living cells[J]. Nature Chemical Biology, 2021, 17(3):246-253.

[71] TABATABAEI S K, WANG B, ATHREYA N B M, et al. DNA punch cards for storing data on native DNA sequences *via* enzymatic nicking[J]. Nature communications, 2020, 11(1): 1742.

[72] CHEN K, KONG J, ZHU J, et al. Digital data storage using DNA nanostructures and solid-state nanopores[J]. Nano Letters, 2018, 19(2): 1210-1215.

[73] CHEN K, ZHU J, BOSKOVIC F, et al. Nanopore-based DNA hard drives for rewritable and secure data storage[J]. Nano Letters, 2020, 20(5): 3754-3760.

[74] LOPEZ R, CHEN Y J, ANG S D, et al. DNA assembly for nanopore data storage readout[J]. Nature Communications, 2019, 10(1): 2933.

[75] ZHANG Y, WANG F, CHAO J, et al. DNA origami cryptography for secure communication[J]. Nature Communications, 2019, 10(1): 5469.

[76] Semiconductor Research Corporation. 2018 semiconductor synthetic biology roadmap [R]. Durham, NC, USA: SRC, 2018.

[77] MCCALLUM J C. Disk drive prices 1955+ [EB/OL]. [2021-05-25]. https://jcmit.net/diskprice.htm.

[78] IARPA. IARPA BAA on molecular information storage [EB/OL]. [2021-05-25]. https://www. iarpa. gov/index. php/research-programs/mist/mist-baa.

[79] MARKOWITZ D. SRC/IARPA Workshop on DNA-based massive information storage [EB/OL]. [2021-05-25]. https://www.src. org/program/grc/semisynbio/semisynbio-consortium-roadmap/6043-full-report-dna-based-storage-final-twg1.pdf.

[80] ANTKOWIAK P L, LIETARD J, DARESTANI M Z, et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction[J]. Nature Communications, 2020, 11(1): 5345.

[81] Illumina. Illumina sequencing platforms [EB/OL]. [2021-05-25]. https://www.illumina.com/systems/sequencing-platforms.html

[82] Illumina. Run time estimates for each sequencing step on the Illumina sequencing platforms. [EB/OL]. [2021-05-25]. https://emea. support. illumina. com/bulletins/2017/02/run-time-estimates-for-each-sequencing-step-on-illumina-sequenci.html.

[83] KONO N, ARAKAWA K. Nanopore sequencing: review of potential applications in functional genomics[J]. Development, Growth & Differentiation, 2019, 61(5): 316-326.

[84] YUAN Z, LIU Y, DAI M, et al. Controlling DNA translocation through solid-state nanopores[J]. Nanoscale Research Letters, 2020, 15: 80.

[85] STANLEY P M, STRITTMATTER L M, VICKERS A M, et al. Decoding DNA data storage for investment[J]. Biotechnology Advances, 2020, 45: 107639.

[86] HECKEL R, MIKUTIS G, GRASS R N. A characterization of the DNA data storage channel[J]. Scientific Reports, 2019, 9(1): 9663.

[87] MAO W, DIGGAVI S N, KANNAN S. Models and information-theoretic bounds for nanopore sequencing[J]. IEEE Transactions on Information Theory, 2018, 64(4): 3216-3236.

[88] MERCIER H, BHARGAVA V K, TAROKH V. A survey of error-correcting codes for channels with symbol synchronization errors[J]. IEEE Communications Surveys & Tutorials, 2010, 12(1): 87-96.

[89] 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 H L, CHEN S H, et al. Chamaeleo: an integrated evaluation platform for DNA storage[J]. Synthetic Biology Journal, 2021, 2(3): 412-427.

[90] Semiconductor Research Corporation. Decadal plan for semi-

conductors full report [R]. Durham: SRC, 2021.

[91] PUZIS R, FARBIASH D, BRODT O, et al. Increased cyber-biosecurity for DNA synthesis [J]. Nature Biotechnology, 2020, 38(12): 1379-1381.

[92] NEY P, KOSCHER K, ORGANICK L, et al. Computer security, privacy, and DNA sequencing: compromising computers with synthesized DNA, privacy leaks, and more[C]//26th USENIX Security Symposium (USENIX Security 17). Vancouver, BC, Canada: USENIX Association, 2017: 765-779.

[93] 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.

[94] Gartner. Gartner unveils top predictions for IT organizations and users in 2021 and beyond [EB/OL]. Gartner [2020-10-21]. https://www. gartner. com/en/newsroom/press-releases/2020-10-21-gartner-unveils-top-predictions-for-it-organizations-and-users-in-2021-and-beyond.

通讯作者：元英进（1963—），男，教授，博士生导师。研究方向为合成生物学及人工基因组化学合成。

E-mail：yjyuan@tju.edu.cn

第一作者：韩明哲（1996—），男，博士研究生。研究方向为合成生物学及DNA信息存储。

E-mail：mickeyhan@tju.edu.cn