Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

An outlook on the current challenges and opportunities in DNA data storage

Muhammad Hassan Raza[a], Salil Desai[b,c], Shyam Aravamudhan[a,c,*], Reza Zadegan[a,c,*]

[a] Department of Nanoengineering, Joint School of Nanoscience & Nanoengineering, Greensboro, NC 27401, USA
[b] Department of Industrial & Systems Engineering, North Carolina Agricultural & Technical State University, Greensboro, NC 27411, USA
[c] Center of Excellence in Product Design and Advanced Manufacturing (CEPDAM), North Carolina Agricultural & Technical State University, Greensboro, NC 27411, USA

ARTICLE INFO

Keywords:
Information storage
DNA
Microfluidics
Automation

ABSTRACT

Silicon is the gold standard for information storage systems. The exponential generation of digital information will exhaust the global supply of refined silicon. Therefore, investing in alternative information storage materials such as DNA has gained momentum. DNA as a memory material possesses several advantages over silicon-based data storage, including higher storage capacity, data retention, and lower operational energy. Routine DNA data storage approaches encode data into chemically synthesized nucleotide sequences. The scalability of DNA data storage depends on factors such as the cost and the generation of hazardous waste during DNA synthesis, latency of writing and reading, and limited rewriting capacity. Here, we review the current status of DNA data storage encoding, writing, storing, retrieving and reading, and discuss the technology's challenges and opportunities.

1. Introduction

Deoxyribonucleic acid (DNA) is the blueprint of living organisms that carries information vital for their growth, development, reproduction, and survival. The unique biochemical and structural characteristics of DNA, such as the Wattson-Crick base pairing, programmable molecular self-assembly, and ability to form nanostructures, have enabled researchers to use it across a diversity of research areas such as drug delivery, biosensing, molecular computation, and nanorobotics (Arter et al., 2020; Dey et al., 2021; Kim et al., 2020; Zhang et al., 2020; Zhou et al., 2020). Besides pharmaceutical sciences, DNA has been used for developing various applications in agriculture and forensics as well (Alarcon et al., 2019; McCord et al., 2019). Over the past few years, DNA is also being explored for unconventional applications such as a medium for storing digital information(G M Church et al., 2012; Lee et al., 2020; Seth L Shipman et al., 2017). Silicon-based information storage systems are the gold standard for storing digital information. Due to the drastic surge in the production of digital information, the global supply of silicon will exhaust soon; therefore, researchers are developing alternative means for storing digital information.

Retention of information, stability and longevity, data storage capacity and density, the energy of operation, and data transfer cost are key factors determining the efficiency of the information storage systems. DNA has an estimated half-life of 521 years under appropriate storage conditions, and it can last over 2 million years if stored in silica (Allentoft et al., 2012; Grass et al., 2015). In contrast, silicon-based information storage systems have limited lifespans and data retention time (Zhirnov et al., 2016). The predicted theoretical digital information storage capacity of DNA is 455 Exabytes of data per gram of single-stranded DNA (ssDNA) (G M Church et al., 2012). High-volume and high-density silicon-based memory fabrication relies on costly manufacturing technologies such as lithography, which are not required

Abbreviations: 2D, two dimensional; A, adenine; AA, amino acid; ASCII, American Standard Code for Information Interchange; BCH, Bose–Chaudhuri–Hocquenghem; C, cytosine; CRC, cyclic redundancy; CRISPR, clustered regularly interspaced short palindromic sequences; CRISPR-MAGE, CRISPR-multiplex automated genome engineering; DENSE, DNA enrichment and nested separation; DMF, digital microfluidics; DNA, deoxyribonucleic acid; EC, error correction; ePCR, emulsion PCR; Gb, gigabase; HTML, HyperText Markup Language; iDNA, information DNA; IPTG, Isopropyl ß-D-1-thiogalactopyranoside; LT, Luby transform; Mb, megabase; MB, magnetic beads; MIC-Drop, multiplexed intermixed CRISPR droplets; mLSI, microfluidic large-scale integration; mVLSI, microfluidic very-large-scale integration; nL, nanoliter; NPMM, nested primer molecular memory; ONT, Oxford Nanopore Technologies; OTP, one-time pad; PAM, protospacer adjacent motif; PEI, polyethyleneimine; PRD, partial restriction digestion; qPCR, quantitative polymerase chain reaction; RA, repeat accumulate; RAD, recombinase addressable data; RS, Reed Solomon; SCRIBE, cellular recorders integrating biological events; ssDNA, single-stranded DNA; T, thymine; TdT, terminal deoxynucleotidyl transferasel; uL, microliter; VLC, variable-length constrained sequence.

* Corresponding authors.

E-mail address: rzadegan@ncat.edu (R. Zadegan).

https://doi.org/10.1016/j.biotechadv.2023.108155
Received 5 December 2022; Received in revised form 23 March 2023; Accepted 12 April 2023
Available online 15 April 2023
0734-9750/© 2023 Published by Elsevier Inc.

by DNA-based memory. Encoding data in nucleic acids also offers lower energy of operation, with efficiency reaching several magnitudes better than flash memory devices (Zhirnov et al., 2016). These properties of DNA make it an excellent choice for information storage.

To storge data on DNA in digital format comprises several steps; comprises several steps (Fig. 1). First, the process involves converting digital information into binary data and converting (encoding) binary data into a nucleotide sequence. DNA is afterward synthesized and stored in vitro or in vivo. Finally, the stored information can be accessed, sequenced, and decoded to the original file. Additionally, some works have focused on storing pixel or vector data based on DNA (Dimopoulou and Antonini, 2021; Seth L. Shipman et al., 2017). Although various approaches for encoding and reading digital information for DNA-based memory have been developed, methods for writing digital data into DNA are limited. Furthermore, information encoding, writing, storage, retrieval, and reading strategies have pros and cons. In this paper, we review the foundations of DNA data storage and discuss the critical challenges and opportunities that correlate with the scalability and sustainability of DNA data storage.

2. Encoding of information in DNA

Various methods for encoding information into DNA are being developed and used (Ailenberg and Rotstein, 2009; G M Church et al., 2012; N Goldman et al., 2013; R N Grass et al., 2015) (Table 1). The feasibility of an information-encoding approach depends on multiple factors, such as the quantitative measurement of information density,

Table 1
Encoding strategies used for storing digital information in DNA.

Work	Information density (nt/bit)	Error correction	Storage	Decoding approach
Portney et al. (2008)	~9.1	No	In vivo	Electrophoresis
G M Church et al. (2012)	0.83	No	In vitro	Illumina sequencing
Nick Goldman et al. (2013)	0.33	Yes	In vitro	Illumina sequencing
Robert N. Grass et al. (2015)	1.14	Yes	In vitro	Illumina sequencing
Bornholt et al. (2016)	0.88	No	In vitro	Illumina sequencing
Shipman et al. (2016)	~2	No	In vivo	Illumina sequencing
Blawat et al. (2016)	0.92	Yes	In vitro	Illumina sequencing
Erlich and Zielinski (2017)	1.57	Yes	In vitro	Illumina sequencing
Organick et al. (2018)	1.1	Yes	In vitro	Illumina sequencing
Wang et al., 2019	1.67	Yes	In vitro	Illumina sequencing
Lee et al. (2019)	1.58	Yes	In vitro	Nanopore sequencing
Dimopoulou et al. (2021)	1.31	No	In vitro	Illumina sequencing

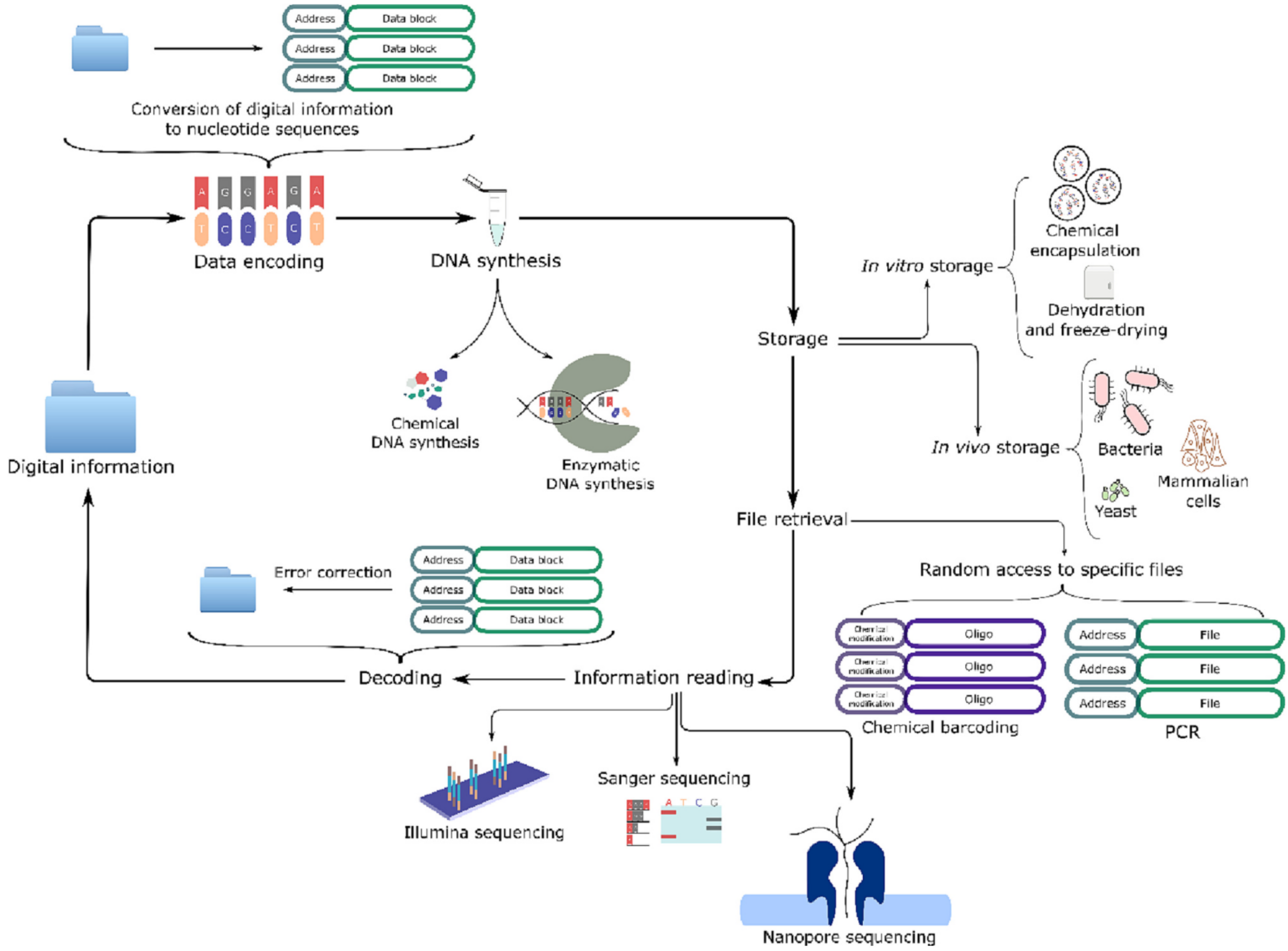


Fig. 1. Foundational principles of conventional DNA data storage: Data is first encoded into nucleotide sequences using a predefined data encoding strategy, followed by synthesizing the DNA using chemical or enzymatic DNA synthesis. DNA then can be stored in vitro or in vivo. Random access in DNA memory has primarily focused on PCR (polymerase chain reaction) to retrieve information read by DNA sequencing and decoding sequencing data into the original digital data.

the nature of the storage medium, i.e., in vitro or in vivo, and the information decoding and error correction methods (Heinis, 2019). Geoffrey Bate performed one of the pioneering works in DNA data storage and reported the comparative analysis of storing digital data in magnetic recordings and DNA (Bate, 1978). Microvenus - a 5 × 7 bit-mapped database - is another earlier study on encoding data in DNA (Davis, 1996). The study demonstrated encoding an icon file into DNA and transforming it into *Escherichia coli* (Bate, 1978; Davis, 1996). The study used molecular weights of the nucleotide bases in an incremental order for encoding digital information. Clelland et al. used substitution cipher for encoding digital information in DNA triplets(Clelland et al., 1999). In another work, a text encoding method used three nucleotide bases (adenine, cytosine, and thymine) to develop information DNA units (Bancroft et al., 2001). The encoding process constructed 27 DNA codons corresponding to the English alphabet using the ternary code. DNA codon "AAA" produced the letter A, followed by the gradual addition of the C and T nucleotide bases at defined positions to generate the 27 codons. The estimated capacity of the information DNA units was ~200 texts in a microchip comprising 10,000 microwells, each containing ~100 unique units. Wong et al. stored digital information in DNA using a predefined coding scheme denoting each character in English text with a triplet of nucleotide (Wong et al., 2003). Smith et al. studied three codes, including the Huffman code, comma code for punctuation of the encoded information and generation of an automatic reading frame, and alternative code that produced a substitute for the existing DNA sequence (Smith et al., 2003). The study concluded Huffman code is practical for short-term digital information storage, and the comma code and the alternatively developed code.

Most of these methods use DNA sequencing technology to decipher the encoded data. Alternatively, Portney et al., developed a DNA sequencing-independent data encoding technique based on partial restriction digestion (Portney et al., 2008). The authors presented a pattern of four and eight base pairs DNA fragments separated by *Alu*I recognition sites in the target DNA to generate either 0 or 1 bits. They partially digested the DNA and performed gel electrophoresis analysis to retrieve the data. Gustafsson et al. generated DNA sequences from protein sequence based on the *Rangifer tarandus* codon bias (Gustafsson, 2009). In summary, most of these approaches focused on encoding text and limited digital information with constrained data capacity and density.

Ailenberg and Rotstein introduced an improved version of the Huffman coding approach to store digital information in DNA (Ailenberg and Rotstein, 2009). An enhanced version of the Huffman coding involved the synthesis of a plasmid library, with each plasmid containing data that can be encoded in 10,000 bp alongside an index plasmid consisting of generic information about the plasmid library. Codons based on the conventional Huffman code were GC-rich; therefore, they replaced Gs with Ts and Cs with As, to improve the DNA synthesis. For encoding the text, three low-base codons, including G, TT, and TA, were utilized as group headers, with the remaining codons listed according to increase in base numbers. To store text, the authors used a single-column version of the customized Huffman code to store music and image files.

Church et al. improved existing information encoding methodologies by introducing one bit per base data mapping (G M Church et al., 2012). Digital information was converted into HyperText Markup Language (HTML), then converted into bits and, eventually, into nucleotides. Goldman et al. utilized Huffman code to convert information to base-3 digits and achieved a Shannon information of 5.2 × 10^6 bits (N Goldman et al., 2013).

The information encoded in DNA is prone to errors during DNA synthesis, DNA storage, and retrieval of information, particularly by sequencing by synthesis (SBS). Heckel et al. reported a qualitative and quantitative analysis of the distribution of such errors in DNA data storage system (Heckel et al., 2019). Numerous studies have focused on integration of EC schemes to tackle such errors. For instance, Grass et al. reported the utilization of Reed-Solomon error correction to encode digital information in DNA (R N Grass et al., 2015). Digital information was mapped to a Galois field of size 47. The two layers of error correction consisted of the inner layer to remove the single base errors and the outer layer to remove the complete DNA string errors. Similarly, Yazdi et al. reported a system for random access to the encoded information in DNA based on a constrained coding technique (Yazdi et al., 2015). The coding strategy implemented several constraints such as GC content of ~50% in the addressing blocks and prefixes, with larger Hamming distances, and no secondary structures. Additionally, the authors introduced increased mutual Hamming distances for proper address selection. Furthermore, the study attempted to remove any correlation between the prefix of an address and the suffix of the same or different address. Lastly, the code tries to reduce the secondary structures in the address sequences to limit the possible interference of secondary structures in the write and read processes.

Blawat et al. introduced a forward error-correction method for encoding information in DNA channels (Blawat et al., 2016). Eight information bits were mapped to a DNA symbol (a DNA fragment comprising five nucleotides). The first nucleotide in the DNA symbol denotes first two binary bits including bit 0 and 1, with bits 2 and 3 represented by second nucleotide, bits 4 and 5 represented by fourth nucleotide, bits 6 and 7 represented by third and fifth nucleotide. For instance, 00 was mapped to A, 01 was mapped to C, and 10 was mapped to G. Similarly, 11 was mapped to nucleotide T. Bose–Chaudhuri–Hocquenghem (BCH) error-correction codes were used to ensure the accuracy in oligo addresses. Furthermore, Reed Solomon protected the continuous oligo blocks. Lastly, 16 bits variant of the Cyclic Redundancy Check (CRC) was used to detect the errors in individual oligos. Based on Goldman's coding method, Bornholt et al. reported the development of exclusive-or operation (XOR) encoding to reduce the overhead generated in the encoding (Bornholt et al., 2016). The authors segmented the nucleotides that were encoded with digital information into blocks which were synthesized as separated DNA fragments. The theoretical information encoding density of the reported method was estimated to be higher than Goldman's encoding, while consisting of 1.5 times the average repetition of each nucleotide.

Kiah et al. developed an alternative approach based on DNA storage channels and utilization of profile vectors for computational modeling of the data reading process (Kiah et al., 2016). The study reported error-correction codes for tackling errors due to DNA synthesis and sequencing, followed by their categorization into asymmetric errors. Erlich & Zielinski et al. developed a data encoding technique referred to as DNA fountain (Erlich and Zielinski, 2017). The DNA fountain data encoding was initiated with binary information preprocessing, followed by subjection to Luby transform (LT), which generated data packets called droplets. Consequently, a DNA sequence was created after converting droplets to nucleotide sequence and screening the DNA sequence for defined parameters such as desired GC percentage and restrains for homopolymers formation. The data encoding approach allowed for digital information storage of up to 500 Megabytes. In another work, Shipman et al. developed an approach for encoding digital images based on a defined assignment of pixels to specific nucleotides (Seth L Shipman et al., 2017).

The decompression of digital information often results in loss of data redundancy and proliferates the errors caused during DNA synthesis and sequencing (Hossein et al., 2017). Yazdi et al. tackled such errors and encoded data into DNA codewords with a defined length of 1000 bp (Hossein et al., 2017). Using Base64 conversion, digital information was compressed and converted to binary information (Josefsson, 2006). The data was converted into nucleotide sequences followed by balancing the GC content of the substring consisting of 8 nucleotides using a constrained coding approach for tackling the formation of the secondary structures and errors arising during DNA synthesis and sequencing. The residual nucleotides were used as addresses for code blocks to be later accessed via polymerase chain reaction.

Organick et al. introduced a coding scheme for reducing sequence redundancy (L Organick et al., 2018). A predefined percentage of logical redundancy to the files led to increased data storage capacity. The data encoding technique utilized Reed-Solomon as the outer layer followed by XOR-based randomization of the information using a pseudo-random sequence. The segmentation of the randomized information initiated the data encoding into several blocks, with each block denoted by a matrix consisting of a defined number of rows and columns. Subsequently, each row of the matrix was encoded using Reed-Solomon, resulting in a modified version of the matrix, later converted into DNA sequences. Zhong et al. encoded digital information using a comparatively shorter data block consisting of a 44-nt (Zhong et al., 2018). The proposed BitDNA encoding scheme was developed based on the representation of every bit by one base-4 number. The storage architecture comprised a data block, indexing sequence, and flanking addresses for accessing the information and paired indexing codes for error-correction. Similarly, the conversion of text files to quadruplets of nucleotides based on the base-4 numeral system has also been reported (Nguyen et al., 2018). Similar storage architecture, i.e., division of binary information into data blocks encoded into DNA strands of a predefined length, has been reported by Tomek et al. (Kyle J. Tomek et al., 2019). In this work, encoding and decoding of information were based on the methods that were reported by Bornholt et al. (Bornholt et al., 2016).

Wang et al. used packet-level repeat accumulate (RA) codes to encode data (Wang et al., 2019). Given the biological constraints faced during encoding information into DNA, two mapping methods, including interleaved mapping defined as the core alongside the variable-length constrained sequence (VLC) mapping as a substitute, were used to convert binary information into a DNA sequence. The storage density of the primary mapping scheme was 1.995 bits per nucleotide, while the VLC mapping scheme comprised comparatively less complex encoding and decoding parameters and reached the information storage density of 1.976 bits per nucleotide. However, combining both mapping schemes increased the efficiency to approximately 1.98 bits per nucleotide.

Takahashi et al. used a one-time pad (OTP) and a data encoding scheme consisting of two layers (Takahashi et al., 2019). Lopez et al. developed a two-layer information encoding scheme (Lopez et al., 2019). Following the randomization of data, it was divided into a specific storage architecture consisting of payloads of predefined size and addresses. As part of the outer layer, Reed-Solomon introduced redundancy into the data, and inner coding generated the DNA sequences from binary information.

The utilization of degenerate bases and four standard nucleotides resulted in increased information (Choi et al., 2019). Eleven degenerate bases were utilized for encoding the data, resulting in an information storage capacity of 3.37 bits/base pair.

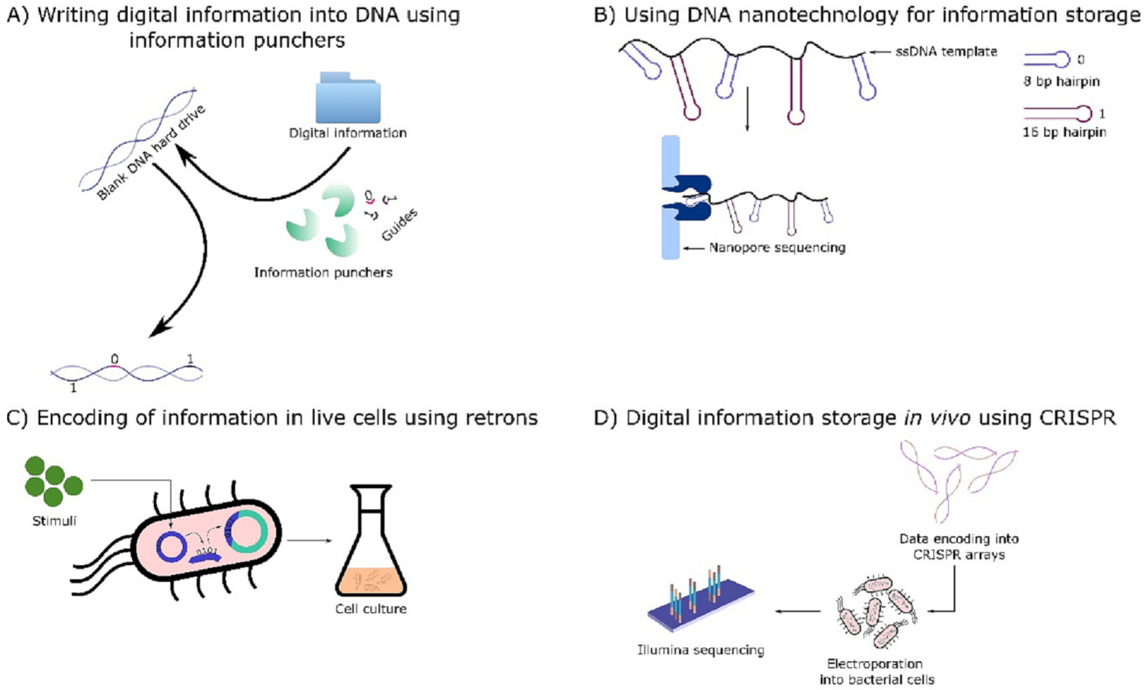
Similarly, Anavy et al. used six-letter composite DNA alphabets for reducing the synthesis cycles (Anavy et al., 2019). The authors used a dedicated approach for encoding information in composite letters alongside error-correction based on a combination of Reed-Solomon and Fountain code. For large composite alphabets, information stored in DNA was converted to binary sequence using standard ASCII encoding, followed by utilization of Huffman encoding for generation of the DNA sequence. Dickinson et al. used multilayer error-correction approach for encoding digital information (G D Dickinson et al., 2021). Data was divided into equally sized substrings, then combined into different combinations using an XOR strategy to form data blocks referred to as droplets. Each droplet was encoded into matrixes of predefined size, followed by the addition of indexing and orientation information alongside the checksum and parity bits. Conclusively, various information encoding approaches have been with each one having a variety of pros and cons in different areas, such as information storage capacity, data compression, data recovery, and error-correction.

Emerging technologies such as artificial intelligence and machine learning can be utilized for improving data encoding approaches (Pan et al., 2022). Furthermore, structuring data in an efficient manner has been little explored in DNA data storage community. Data structures are commonly used for organizing and storing data in computer programs and allow the data to undergo different types of operations such as accessing, searching, querying, modifying, deleting, or sorting (Odeh and Knuth, 1969). Data structures serve as the fundamental building blocks of intricate computational algorithms, owing to their ability to achieve efficient information organization. Lopiccolo, Annunziata, et al. modeled a stack data structure for storing and retrieving digital information in a last-in first-out approach (Lopiccolo et al., 2021). The data structuring strategy was based on polymerizing DNA chemistry. As described earlier, the DNA data storage process is prone to several technical challenges such as DNA breaks and rearrangements. Song et al. used de Bruijn graph and greedy search algorithms to tackle these challenges (Song et al., 2022). As a proof-of-concept, the authors were successfully able to recover 6.8 MB of information from a sample incubated at 70 °C for 70 days and were able to reach a physical density of 295 petabytes/g which indicates the improvement in the robustness of the DNA data storage using their approach. Other approaches such as Bloom filters and suffix trees have been used for various applications related to DNA assembly and can be explored for potential applications in DNA data storage applications (Huo et al., 2007; Nayak and Access, 2019). Likewise, data compression techniques focus on optimization of the ratio between raw data and the headers such as metadata, EC codes, addressing information. There are several techniques used for compressing data to be stored in DNA. Lossless compression techniques such as Huffman coding, Run Length Encoding, and Lempel-Ziv-Welch (LZW) algorithm preserve the information in a file while maximizing the compression of the data. For instance, Huffman coding is one of the most widely used coding methods with a compression ratio of 20%-90% (Cao et al., 2022). In contrast, lossy compression algorithms such as Transform Coding, Discrete Cosine Transform, and Fractal Compression tend to keep the core information whilst discarding some of the data. As discussed earlier, most of the DNA data storage approaches rely on synthesizing DNA using phosphoramidite chemistry associated with generation of vast amounts of hazardous waste. Developing efficient data structuring approaches alongside data compression techniques can significantly reduce the amount of DNA required to store a given amount of digital information which will help to reduce the carbon footprint associated with synthesis of DNA.

3. Writing digital information on DNA

Different approaches can be used for writing digital information on DNA (Fig. 2). Here, we refer to the writing of digital information as denoting specific parts of a sequence of DNA to encode information. Conventional DNA data storage strategies have primarily focused on directly encoding digital information into DNA and synthesizing DNA using chemical or enzymatic DNA synthesis. However, one of the most significant disadvantages of this approach is the generation of massive amounts of toxic waste (Palluk et al., 2018). Other factors, such as the cost associated with synthesizing large DNA fragments, nonexistent or limited approaches for rewriting information, and the latency associated with the encode-to-synthesis process, drastically impact the scalability of DNA memory. Alternatively, a comparatively environment-friendly DNA data storage system would focus on utilizing a standardized blank DNA hard drive that can be used for storage, random access, and rewriting of digital information, for instance, using different enzymes. Different alternative approaches focus on utilizing DNA-modifying enzymes or *information punchers,* such as clustered regularly interspaced palindromic sequences (CRISPR) system, homologous recombination, and site-specific nucleases have been used for writing digital information onto DNA (Bonnet et al., 2012; George M. Church et al., 2012; Farzadfard and Lu, 2014; Li et al., 2018; Seth L Shipman et al., 2017; Tabatabaei et al., 2020; Yim et al., 2021).

Bonnet et al. developed a recombinase addressable data module to store rewritable digital information storage in living cells (Bonnet et al.,


A) Writing digital information into DNA using information punchers

B) Using DNA nanotechnology for information storage

C) Encoding of information in live cells using retrons

D) Digital information storage *in vivo* using CRISPR

Fig. 2. Unconventional approaches for writing information into DNA. A) A blank DNA hard drive is synthesized, and information is written using information punchers or programmable enzymes such as CRISPR, recombinases, or artificial restriction enzymes. B) M13mp18 ssDNA has been used to encode binary information; 8 bp DNA hairpins as 0 and 16 bp hairpins as 1. C) On-demand encoding of information using expression of ssDNA and their utilization for inducing mutations into the target loci as a genomic memory. D) Digital information can be converted into pixels encoded into CRISPR arrays. CRISPR arrays are then transformed into bacterial cells, and Illumina sequencing can be utilized for reading the information.

2012). It comprised a two-state latch system to change its configuration in response to external stimuli. Serine integrase and excisionase were utilized for site-specific editing of the templates to store information. A similar study reported the usage of recombinase to develop Synthetic Cellular Recorders Integrating Biological Events (Farzadfard and Lu, 2014). It contained stimuli-responsive production of ssDNA using retrons in *E. coli.* These ssDNAs, when expressed together with recombinase, could induce mutagenesis in the target loci. The stimuli-encoded memory in target loci was programmable with dependency on the modularity of the sequences of ssDNAs. Another study reported the concept of enzymatic nicking using *Pyrococcus furiosus* Argonaute (*Pf*ago), an artificial restriction enzyme, to store information in DNA (Tabatabaei et al., 2020). *Pf*ago uses short 16 nt guide sequences for controlled cleavage of DNA (Enghiad and Zhao, 2017). Digital information was converted to binary information followed by division into defined blocks of *m* bits, where *m* referred to the number of cleavage locations on DNA. Subsequently, cleavage was introduced using *Pf*ago at the determined locations using the DNA guides.

Since its development for genome engineering applications, CRISPR systems have progressively grown to write digital information into DNA, attributed mainly to their reprogrammable and controllable nature (Li et al., 2018; Shipman et al., 2016; Seth L Shipman et al., 2017; Yim et al., 2021). Shipmen et al. reported one of the earliest works for using CRISPR system in DNA memory, where specific DNA sequences were integrated into a genomic CRISPR array in the form of spacer sequences (Shipman et al., 2016). DNA fragments were electroporated and integrated into the genome of *E. coli.* The diverse class of mutants of Cas1 and Cas2 generated using directed evolution allowed the acquisition of defined DNA sequences. High throughput DNA sequencing was used for sequencing inserted DNA fragments. Similarly, the work was expanded to store digital information using the type 1-E CRISPR-Cas system (Shipman et al., 2017), where pixel values from the digital information were converted into synthetic oligonucleotides integrated into living cells containing unique CRISPR arrays. Yim et al. engineered a redox-responsive CRISPR-directed information writing system and encoded CRISPR arrays with digital information in live cells that were subsequently barcoded to address scalability (Yim et al., 2021). CRISPR-Cas system has also been utilized in the DNA steganography concept, one of the critical pillars of DNA data storage. Li et al. used the CRISPR-Cas genome engineering approach to access the key to the information stored using DNA memory (Li et al., 2018). The study developed a pre-key via mixing of real key with the fake key or alternatively using a real key modified with additional sequence at the 3' end. Subsequently, the authors used CRISPR/Cas12a to cleave the fake key or the additional sequences added at the 3' end to produce the real key for accessing the concealed information. In the security aspect of DNA data storage, little work has been done and it requires exploration for further studies. On the other side, majority of these works enable information writing methods that do not require de-novo DNA synthesis, two significant concerns, including the programmability and controllability of the enzymes employed in the information writing process, are critical factors that determine the efficiency of the involved processes and impact the scalability of DNA memory. Development of approaches such as CRISPR-directed base editing that allow precise and modular single-base resolution editing of DNA will affect the expandability of the DNA data storage (Kim, 2018).

4. DNA synthesis for information storage

A variety of DNA syntheses approaches, such as chemical and enzymatic DNA synthesis, have been utilized in DNA data storage with a significant emphasis on chemical DNA synthesis; mainly attributed to the well-established methodology (Anavy et al., 2019; Antkowiak et al., 2020; Nick Goldman et al., 2013; Robert N. Grass et al., 2015; Lee et al., 2020, 2019; Xu et al., 2021; Yoo et al., 2021). The chemistry underlying the chemical DNA synthesis process is beyond this article's scope and can be explored elsewhere (Hughes and Ellington, 2017; Kosuri and Church, 2014).

Chemical DNA synthesis is the gold standard in the life sciences industry; and, therefore, has been used for the synthesis of DNA fragments of variable lengths, from a few bases to large DNA fragments (G M Church et al., 2012; Yazdi et al., 2015). For instance, Church et al. synthesized 54,898 oligonucleotides using a microarray-based approach (Church et al., 2012). Similarly, another study reported the utilization of microarray-based DNA synthesis platform for synthesizing 153,335 'strings' of DNA, each consisting of 117 nt (N Goldman et al., 2013). Depending on the information storage approach, DNA sequences encoded with digital information have been synthesized differently, such as oligonucleotide pools and standardized ssDNA templates. A significant disadvantage of chemical DNA synthesis is the cost incurred and the length of the oligos that can be produced with high accuracy. The size limitation of chemical DNA synthesis affects storage of digital information in DNA in several ways. For instance, for storing large amounts of digital information, long DNA sequences are required. However, the current limitation of chemical DNA synthesis is that it can reliably produce oligonucleotides that are up to a few hundred nucleotides in length with high accuracy. This means that long DNA sequences, such as those required for large-scale data storage, cannot be synthesized with high accuracy using chemical synthesis alone. Furthermore, the accuracy of the synthesized oligonucleotides decreases as their length increases. This can lead to errors in the encoded data, which can result in loss of information. Many studies have utilized versatile information encoding approaches with different DNA synthesis approaches to tackle these challenges. E.g. A 12 K chip was used to synthesize 4991 nucleotide sequences, focusing on reducing the cost (~ $2500/pool) (R N Grass et al., 2015). Takahashi et al. automated information storage in DNA and used phosphoramidite chemistry for DNA synthesis (Takahashi et al., 2019). The addition of composite DNA letters to address the higher cost associated with DNA synthesis has also been demonstrated (Anavy et al., 2019). In this work, using phosphoramidite DNA synthesis, foundational bases, including A, T, G, and C, along with the composite letters were synthesized, eventually reducing the number of DNA synthesis cycles. Phosphoramidite DNA synthesis has also been used to synthesize l-DNA for data storage (Fan et al., 2021). However, phosphoramidite chemistry is attributed with generation of chemical waste and the large-scale DNA chemical DNA synthesis is not cost-effective. These disadvantages make the chemical DNA synthesis considerably a less efficient approach for DNA data storage (Palluk et al., 2018).

Several studies have focused on enzymatic DNA synthesis, specifically for DNA data storage applications (Lee et al., 2020, 2019; Palluk et al., 2018; Yoo et al., 2021). Enzymatic DNA synthesis uses engineered DNA polymerases to synthesize user-defined DNA sequences (Eisenstein, 2020). A template-independent DNA polymerase called terminal deoxynucleotidyl transferase (TdT) was used for enzymatic DNA synthesis to store digital information (Lee et al., 2019). Information reading using Illumina sequencing revealed the presence of missing nucleotides, which was reduced when MinION was utilized for the sequencing, potentially attributed to the sequencing efficiency of nanopore sequencing (Lee et al., 2019). For encoding information into DNA, often many oligonucleotide strands are required. Therefore, a parallel DNA synthesis strategy can drastically improve the utilization of enzymatic DNA synthesis for DNA data storage-related applications. Furthermore, due to the inherent tendency of TdT to be non-specific in its enzymatic activity, its sequence-specific control of is difficult. H. Lee et al. reported parallel enzymatic synthesis by photolithographic modulation of TdT in a multiplexed array and stored 110 bits of information using base transition (Lee et al., 2020). The photolithographic control of TdT relied on the spatiotemporal concentration of the Co^{2+} cofactor required for TdT. During the information decoding, errors with single-base deletions accounted for 25.8% of the errors, followed by single-base insertions (13.4%) and mismatches (8.9%). Such errors can impact the recovery of the information. However, integrating several factors into the information storage architecture, such as the physical redundancies utilized in the study, as mentioned earlier, can dramatically improve information recovery. Enzymatic DNA synthesis poses a viable alternative to traditional approaches such as phosphoramidite synthesis. However, there are several disadvantages, including the error-prone nature of the enzymatic DNA synthesis process and limitation on the size of the synthesizable DNA. Synthesizing longer DNA sequences is of key importance for DNA data storage because it enables the storage of larger amounts of data per DNA pool. Longer DNA sequences can also increase the robustness and reliability of the stored data. Addressing these challenges would make the environment-friendly synthesis of DNA using enzymes a feasible process for information storage applications. Recently, a study reported TdT from *Zonotrichia albicollis* with catalytic activity surpassing that of the commonly used mammalian TdT by three orders of magnitude when using 3′-ONH2-dNTPs (Lu et al., 2022). Similarly, Padhy et al. used microfluidics to develop a dielectrophoretic bead-droplet reactor for carrying out high-fidelity enzymatic DNA synthesis (Padhy et al., 2022). The authors carried out solid-phase enzymatic DNA synthesis to produce oligonucleotides on beads using dielectrophoretic force. Such studies can pave the way for automated and miniaturized enzymatic DNA synthesis approaches. Life sciences companies are also focusing on reducing the overall cost associated with enzymatic DNA synthesis; particularly by reducing the required amount of costly nucleoside triphosphate (NTP) during the enzymatic synthesis process (Blois, 2022).

In contrast to phosphoramidite and enzymatic DNA synthesis, alternative methods for DNA synthesis have also been reported. For instance, Antkowiak et al. focused on utilizing a light-modulated maskless array approach to synthesize 6383 sequences for information storage (Antkowiak et al., 2020). Xu et al. reported the synthesis of DNA using an electrochemical approach on Au electrodes (Xu et al., 2021). DNA synthesis approaches for data storage applications with greener and environmentally friendly approaches that enable synthesis of longer DNA strands with minimal mutations are yet to be established.

5. Random access in DNA memory

Selective retrieval of data from a pool of files stored is referred to as random access. Various methods have been developed for random access in DNA based-memory, focusing significantly on PCR utilization (Fig. 3). PCR allows to amplify specific regions of the DNA that comprise of the encoded digital information. Using primers that bind to the flanking regions of a target sequence, PCR can only amplify the desired target region, enabling more accurate retrieval of encoded data which reduces the errors during the decoding of digital information. Kashiwamura et al. introduced Nested Primer Molecular Memory (NPMM), where data was divided into specific sequences of DNA, referred to as data blocks addressed using primer address sites (Kashiwamura et al., 2003). Specific data blocks were extracted using nested PCR. Primers were designed with specific constraints; including GC content to increase the specificity, hamming distance to prevent hybridization between oligos, and a 3' end complementary evaluation parameter to avoid mispriming. As an extension of the work by Kashiwamura et al. nested PCR was used to develop DNA memory with 16.8 million addresses (Yamamoto et al., 2008a, 2008b). Nested PCR was utilized to reterive specific data blocks in a given pool of files. PCR-based methods have been widely used in the past for accessing desired files. For instance, S. M. Yazdi et al. used 1000 bps data blocks flanked on both sides by an address block allowing selective retrieval of information using PCR (Yazdi et al., 2015). Random extraction of 1000 bps sequences was performed using PCR from a pool of DNA fragments.

Similarly, Organick et al. used a primer-based approach to extract desired files from a pool of 13,448,372 nucleotide sequences (L Organick et al., 2018). Each strand comprised a unique address for identifying the strand and a file ID. Random access to specific files using PCR has been reported elsewhere (Organick et al., 2020).

On the other hand, PCR-based random access has several associated disadvantages. For example, the selective retrieval of files using PCR

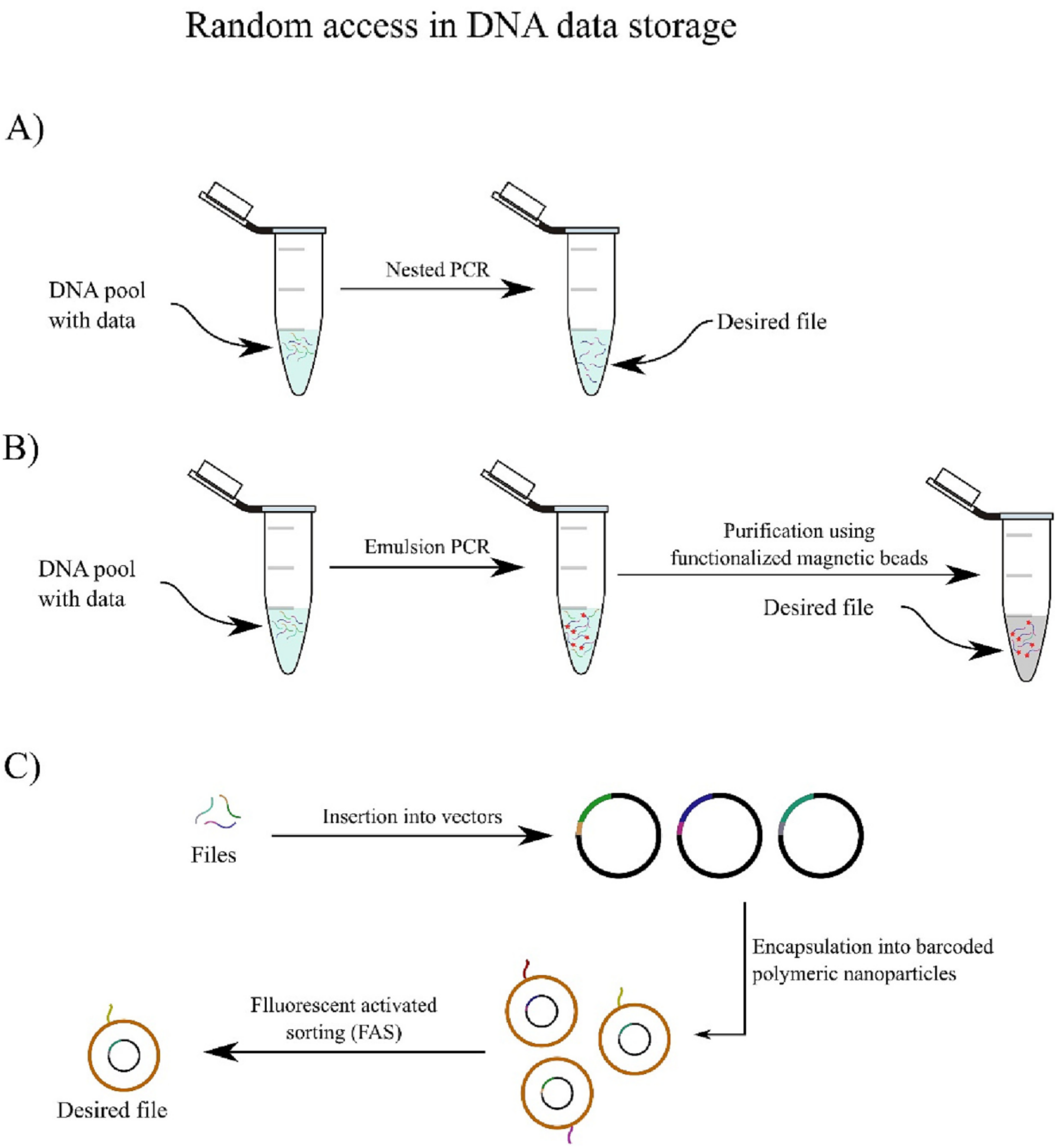


Fig. 3. Generic approaches for random access. A) Files are generally divided into two sections: an addressing section and a data block. Nested PCR can be used for amplification of the desired file. B) Files are amplified using emulsion PCR to generate a pool of the amplified file containing a chemical tag that can be used together with functionalized magnetic beads to extract the file. C) Data is encoded into DNA and inserted into plasmid vectors. The vectors are eventually encapsulated into polymeric nanoparticles that are labeled with barcodes. Subsequently, FAS can be used to separate the desired file physically.

may decrease the relative abundance of the extracted file in the pool after a certain number of PCR cycles (Kyle J. Tomek et al., 2019). Tomek et al. developed DNA Enrichment and Nested Separation to tackle the depletion of the original file (Kyle J. Tomek et al., 2019). The method was based on the extraction of the target file using chemically functionalized primers and production of labeled copies of the required file using emulsion PCR (ePCR). Subsequently, functionalized magnetic beads were added to the solution containing the chemically labeled file. Enrichment of a pre-specified file referred to as File 3 was conducted after random access showing the presence of File 3 as 0.2, 87.5, and 100% of the pool following biotin-modified PCR. Similarly, enrichment of File 3 following the fluorescein-modified PCR and random access was 0.1, 49.6, and 100%; 0.2, 14.2, and 100% in the case of digoxigenin-modified PCR; and 0.09, 0.47, and 100% in case of poly(A)-25 modification which conclusively revealed the efficiency of file retrieval approach Recently, silica capsules encapsulated DNA sequences encoding the digital information using a sol-gel chemical approach and were selectively labeled with individual barcodes using 25 nt long ssDNA strands (James L. Banal et al., 2021). The study focused on using fluorescently tagged 15-nt ssDNA fragments complementary to the barcodes to retrieve the file using fluorescence-activated sorting (FAS). Files were accessed using a Boolean search based on AND, OR, and NOT gates.

To conclude, nested PCR was used as a foundation for developing random access in DNA memory, and it has been widely used for the selective retrieval of files. Over the past few years, a combination of approaches, including PCR, chemical modification of oligos, and encapsulation of oligos comprising individual files, have been reported, and the developments in these areas will drive the feasibility of DNA memory (Kashiwamura et al., 2003; Yamamoto et al., 2008a, 2008b; Yazdi et al., 2015; L Organick et al., 2018; Organick et al., 2020; Kyle J. Tomek et al., 2019; James L. Banal et al., 2021).

6. Reading digital information encoded in DNA

Information encoded in DNA can be read by sequencing of the DNA using well-established approaches such as Illumina sequencing and nanopore sequencing, with each having its advantages and disadvantages in terms of reading length, minimal generation of errors in the sequencing data, and portability (J L Banal et al., 2021; G M Church et al., 2012; G D Dickinson et al., 2021; Erlich and Zielinski, 2017; N Goldman et al., 2013; R N Grass et al., 2015; Meiser et al., 2020; Newman et al., 2019; L Organick et al., 2018; Organick et al., 2020; Seth L Shipman et al., 2017; Takahashi et al., 2019; Kyle J Tomek et al., 2019; Hossein et al., 2017). Illumina sequencing provides higher accuracy (>99.9%) as compared with that of ONT (87-98%) (Lin et al., 2021). The estimated cost per gigabase (Gb) for ONT's PromethION is $21-$42 whilst the same for Illumina's NovaSeq 550 is between $50-$63 (Lin et al., 2021). Illumina has a wide range of products for both short and long-read sequencing. However, nanopore sequencing is known to provide the longest reads as compared with other NGS technologies with highest read record of 2.3 megabase (Mb) (Amarasinghe et al., 2020). Additionally, nanopore sequencing is comparatively more portable and provides the sequencing data in real-time (McNaughton et al., 2019). Conclusively, these factors collectively impact the choice of DNA sequencing technology to be opted for. In DNA data storage, both Illumina and nanopore sequencing have been widely used to read digital information encoded in DNA. Church et al. used HiSeq 2000 to sequence the encoded information in one of the earliest published landmark studies in DNA memory. Authors located 22 discrepancies among the designed and read sequences, with 20 positioned in the last 15 bases of the oligonucleotide sequence (Church et al., 2012). Illumina's HiSeq 2000 has also been used elsewhere (N Goldman et al., 2013). Illumina's MiSeq platform has also been used for reading the digital information stored in DNA (Erlich and Zielinski, 2017; R N Grass et al., 2015; Seth L Shipman et al., 2017).

One of the biggest challenges driving DNA memory's scalability is the portability of reported approaches. The information reading is usually accomplished using DNA sequencing, while the cost, physical space requirements, and time associated with most DNA sequencing approaches drastically impact the portability aspect of DNA memory. Yazdi et al. used MinION, introduced by Oxford Nanopore Technologies (ONT), read information from a portable DNA memory device (Hossein et al., 2017). Using the R7 version of nanopore, the authors sequenced the DNA blocks with a ~ 75 bp/s speed. As DNA sequencing using MinION is prone to significantly higher error rates than the traditional DNA sequencing techniques, an error rate of 0.2% was reported after the consensus determination without the implementation of error-correction. Likewise, nanopore sequencer was used to recover two 32-KB and 1.3-KB files, respectively, with an approximate coordinate error rate of 12% (L Organick et al., 2018). Another study focused on using nanopore sequencer for reading the information encoded into DNA reported the association of decreased ligation efficiency with low decoding rate and payload generation (Takahashi et al., 2019) On the other hand, Dickinson et al. employed super-resolution microscopy to read the digital information encoded into DNA origami (G D Dickinson et al., 2021). Dickinson et al. used DNA-Points Accumulation for Imaging in Nanoscale Topography (DNA-PAINT), for information reading via imaging of the DNA origami. A mean of 7.3 ± 1.2 false errors per DNA origami was detected, while the information decoding algorithm successfully recovered information. False-positive errors were also reported with a mean of 1.7 ± 0.5.

Various approaches have been introduced for reading the digital information encoded into DNA. Emerging DNA sequencing technologies such as Oxford Nanopore will pave the way for developing portable and rapid information reading approaches.

7. Information storage in vivo vs in vitro

DNA is a principal carrier of information in living organisms. Inspired by the natural ability of DNA to store digital information, microorganisms such as bacteria have been engineered for recording molecular events and for storage of digital information (Bonnet et al., 2012; F R et al., 2013; Farzadfard and Lu, 2014; Wong et al., 2003; Yachie et al., 2007, 2008). One of the earliest studies focusing on information storage in DNA utilized *E. coli and D. radiodurans* to store digital information. *D. radiodurans* was selected primarily attributed to its ability to survive in harsh environments (Wong et al., 2003). Encoding information into the genome of other bacteria, such as *Bacillus subtilis,* has also been demonstrated (Yachie et al., 2007). Digital information was first encoded into plasmids which were then utilized to integrate the cassettes containing relevant information to *B. subtilis.* Bonnet et al. showed rewritable information storage in an *E. coli* chromosome (Bonnet et al., 2012) Similarly, Farzadfard & Lu used ssDNA molecules for genomically encoded memory (Farzadfard and Lu, 2014) CRISPR-based genome engineering has also been used for writing information into genomes of live cells (Shipman et al., 2016; Seth L Shipman et al., 2017). One of the most important criteria for selecting a suitable organism for information storage applications is its ability to maintain the stability of DNA sequences for extended periods of time, particularly in extreme environments. Recently, Liu et al., used a bacterial artificial chromosome in a *Bacillus* chassis for information storage (Liu et al., 2022). The authors used inducer molecules including nisin, xylose, and Isopropyl β-D-1-thiogalactopyranoside (IPTG) to induce respective promoters for random access of desired files from the bacterial artificial chromosome. The *Bacillus* spores were stored in acacia gum and activated charcoal and were exposed to harsh conditions including exposure to high temperatures of 60 °C for a period of 14 days, exposure to oxidizing agents (10% H_2O_2) for 2 days and UV irradiation (105 W/m2 at UV 254 nm) for 120 h. The data retrieval process indicated error rates of <1% from the spores exposed to high-temperature and oxidant stress treatment, and <4% from those subjected to UV irradiation treatment. Conclusively, the diversity of approaches used for storing digital information in living organisms is vast, yet the in vivo information storage is viable for archival storage at this stage.

Digital information has also been stored in DNA in other forms, such as encapsulated silica particles and DNA origami (G D Dickinson et al., 2021; R N Grass et al., 2015). Grass et al. synthesized silica particles to encapsulate DNA for long-term storage and showed that DNA preserved in silica could be stored for >2 million years if appropriate storage conditions were maintained. In another work, DNA encoded with digital information was utilized together with polyethyleneimine (PEI) in a layer-by-layer approach, followed by the addition of silica to protect DNA from harsh environmental conditions (W. D. Chen et al., 2019). Similarly, Antkowiak et al. encapsulated DNA in silica nanoparticles and utilized digital microfluidics for the on-demand retrieval of information (Antkowiak et al., 2022). Several studies have also focused on the storage of information in DNA nanostructures. DNA nanotechnology takes advantage of the fundamental biochemical characteristics of DNA to form nanostructures that can be modulated using different mechanisms, such as toehold-mediated strand displacement (Ijäs et al., 2018; Rothemund, 2006; Zhang and Winfree, 2009). Chandrasekaran et al. stored digital information in conformational states of DNA nano-switches possessing binary switching properties (Chandrasekaran et al., 2017). The storage architecture was based on a loop formed by partially hybridizing parts of the DNA nano-switches with external strands called data strands. Toehold-mediated strand displacement was used to displace the data strands to enable the rewriting capability of the DNA nano-switches with latency ranging from minutes to hours, depending on the concentration of the data strands. A 5-bit DNA data storage system was demonstrated, while an 8-bit system with potentially expandable storage capacity was also developed. Chen et al. used M13mp18 ssDNA as a DNA carrier to assemble hairpins to store digital information readable via a solid-state nanopore (K. Chen et al., 2019). Two types of DNA hairpins, including eight bp hairpin encoding 0 while 16 bp hairpin encoding 1 with a distance of 114 bp, were placed on the DNA carrier and resulted in 56 hairpins per 7228 bp DNA carrier. 112-bit of digital information was encoded into 2 DNA carriers, and the Bayesian inference approach was employed to tackle the errors arising during the information reading process. Dickinson et al. used DNA origami to store digital information (George D. Dickinson et al., 2021). Oligonucleotide strands, also called staple strands, stored digital information as 1 or 0 in DNA origami and were replaced with the desired strands during the information rewriting process. To summarize, different methods have been used to store information using DNA origami. DNA nanostructures have several disadvantages associated with them. For instance, information stored in DNA origami may be impacted by the changes in the thermodynamics. On the other hand, DNA origami poses a potential choice for information storage approaches focuses on data security.

8. Automation of the DNA memory writing and reading processes

Automating the digital information writing and reading processes may improve portability and scalability of DNA memory. The conceptualization and progress in DNA memory have primarily focused on the encoding and decoding of digital information, with minimal focus on the automation, miniaturization, and portability aspects. These aspects contribute to the commercialization of DNA memory, and the reported work in the DNA memory field is currently far from the concept of commercialization. Here, we review the studies focused on the automation and miniaturization of generic molecular memory concepts and discuss the potential integration of microscale technologies, such as microfluidics with DNA memory, for a scalable DNA memory system that meets industry standards.

Takahashi et al. did one of the pioneering works on the automation of the DNA memory (Takahashi et al., 2019). The authors showed the automation concept with a focus on three individual components. Each had unique functions for writing and reading the digital information encoded into DNA: i) software for encoding and decoding the information, ii) DNA synthesis column, and iii) DNA sequencing using nanopore sequencer. The information was first translated into DNA sequences, followed by synthesizing and storing the synthesized DNA until a retrieval for reading the encoded information was requested. Eventually, a specified DNA amount was eluted and sent for DNA sequencing using Oxford Nanopore's MinION, and information was decoded using the decoding software. In standard semiconductor-based information storage systems, the latency of writing and reading digital information is critical. The earlier study reported latency of 5 bytes in ~21 h, with a significant portion of the time, i.e., ~8.4 h, devoted to synthesizing DNA. The study's limitations are the physical space consumption with a footprint of a benchtop and a cost of ~10,000$. Microfluidic lab-on-a-chip platforms (Table 2) pose an excellent alternative to tackle the cost and space consumption requirements associated with the automation process (Gach et al., 2016; Iwai et al., 2018, 2022). Newman et al. used a digital microfluidic platform to retrieve information stored in dehydrated DNA spots (Newman et al., 2019). Water droplets were utilized to elute the dehydrated DNA droplets, and information was read using DNA sequencing. DNA spots with physical mass ranging from 5 ng to 60 ng comprising 2042 distinct DNA strands and ~ 20 KB of digital information dwelled for 1 min. Following dwelling, >99.8% of the sequences were subjected to the information reading process at least once, enabling data recovery. Similarly, 398,000 copies of 276,000 DNA sequences were stored in dehydrated DNA spots of ~30 ng, and 1.2% of sequences were reported missing while sequencing the DNA. While writing the information, various DNA-modifying enzymes; or information writers such as recombinases and the CRISPR-Cas9 system are employed, a scalable DNA storage system may require various information writers and many DNA templates, which may be challenging to achieve, attributed to the consumption of many reagents in a substantial quantity contributing to an exponential increase in the cost and space consumption, and the amount of physical labor. Automation may be achieved using laboratory automation robots introduced by different companies. However, physical space and consumption of hefty amounts of laboratory reagents that contribute to the process's cost still exist. Digital microfluidics systems can be utilized for automation with reagent requirements of as few microliters.

Table 2
Different approaches used for fabrication of microfluidic devices and their characteristics.

Characteristics	Soft lithography	3D printing	Xurography
Opticaltransparency	Excellent	Variable. Limited materials are transparent.	Good
Feature resolution	Very high. Limited by the approach used in photolithography.	Very high. Limited by the 3D printing technique used to 3D print features.	Good
3D geometries	Achieved using multi-layer soft lithography by stacking multiple layers	Variable	Limited
Biocompatibility	Excellent	Variable. Limited materials are biocompatible.	Good
Thermostability	Limited	Variable	Limited
Flexibility	Flexible and stretchable	Limited flexible materials	Flexible
Throughput	High	High	Limited

Gach et al. demonstrated a hybrid droplet microfluidics platform to automate the fundamental processes associated with genetic engineering, such as transformation and cell culture in microorganisms (Gach et al., 2016). Miniaturization and automation using the hybrid droplet microfluidics device decreased the cost of the reagents by 100-fold compared to the standard laboratory protocols. Similarly, Iwai et al. developed a droplet microfluidic device combined with a 10 × 10 element array to manipulate droplets using electric fields for CRISPR-multiplex automated genome engineering (CRISPR-MAGE) of *E. coli* (Iwai et al., 2022). The device had 100 individually addressable chambers with the reagent requirement of <2 uL. Likewise, Parvez et al. introduced Multiplexed Intermixed CRISPR droplets (MIC-Drop), a droplet microfluidics platform for generating large-scale reverse genetic screens for zebrafish (Parvez et al., 2021). 188 genes were screened using nanoliter-sized droplets. Similar to hybrid microfluidic platforms that focus on manipulation of droplets of defined size and volume using sigital microfluidics, approaches such as microfluidic large-scale integration (mLSI) and microfluidic very-large-scale integration (mVLSI) are potentially suitable alternatives for scalable DNA memory where many information writers and DNA templates are involved (Fig. 4) (Araci and Quake, 2012; Liu et al., 2003; Thorsen et al., 2002).

mLSI and mVLSI utilize microvalves to route the reagents, and the placement of these microvalves can address hundreds of individual reaction chambers (Gómez-Sjöberg et al., 2007; Melin and Quake, 2007; Thorsen et al., 2002; Unger et al., 2000). These concepts have been implemented mainly in biomedical sciences for applications such as cell culture (Briones et al., 2021; Gómez-Sjöberg et al., 2007; Hu et al., 2018; Kellogg et al., 2014). Similarly, emerging hybrid microfluidic platforms such as complementary metal-oxide semiconductor (CMOS)-integrated microfluidic devices will pave the way for developing platforms for writing, retrieving, and reading digital information in DNA. To summarize, microfluidics devices can address the challenges, such as automation and miniaturization in DNA memory, that will eventually drive the scalability of DNA-based information systems to a commercial level.

9. Concluding remarks and future perspective

At the current pace of the digital information generation, existing silicon-based information storage systems infrastructure will exhaust soon (Zhirnov et al., 2016). DNA data storage offers a potential alternative to traditional data storage platforms due to theoretically higher information storage density, and data retention coupled with low energy

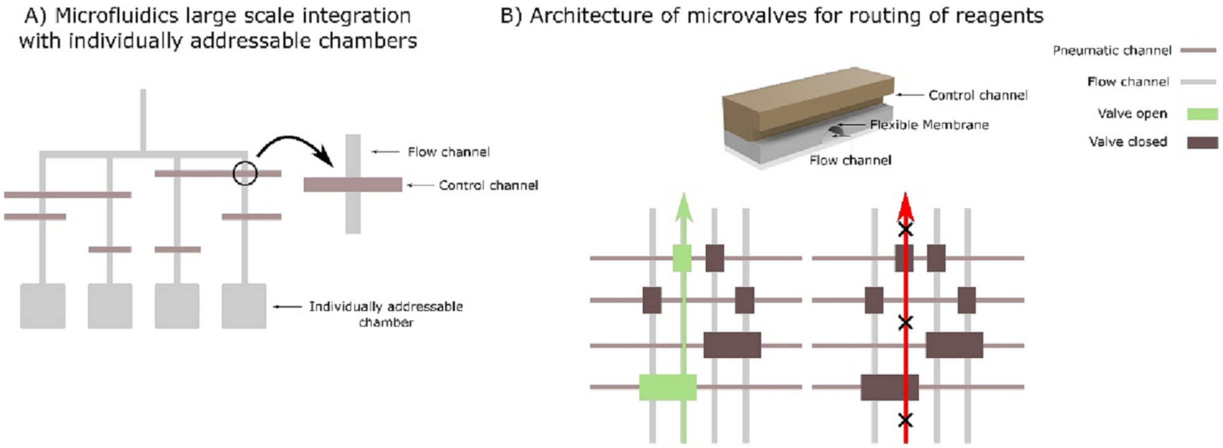

A) Microfluidics large scale integration with individually addressable chambers
Flow channel
Control channel
Individually addressable chamber
B) Architecture of microvalves for routing of reagents
Control channel
Flexible Membrane
Flow channel
Pneumatic channel
Flow channel
Valve open
Valve closed

Fig. 4. Microfluidic large-scale integration. A) A set of reagents can be combined and dispensed into desired locations based on the placement of microvalves in the microfluidic device. B) Typical membrane microvalves comprise a control layer consisting of pneumatic channels that form a perpendicular junction with the flow channels. Pressurizing the pneumatic channels deflects the thin, flexible membrane into the flow channel to block the fluid flow.

of operation requirements. Significant work has been done on different aspects of DNA data storage, including data encoding algorithms, error-correction schemes, DNA synthesis approaches, and information writing processes, yet it is still in its infancy stage with a number of challenges (Bonnet et al., 2012; George M. Church et al., 2012; Kosuri and Church, 2014; Lee et al., 2019, 2020; Lopez et al., 2019; Seth L Shipman et al., 2017; Tabatabaei et al., 2020).

Firstly, conventional DNA data storage relies on the synthesis of DNA. Although the cost of DNA sequencing has drastically reduced over the past few years, the cost associated with DNA synthesis is still comparatively high. Storage of a large amount of digital information in DNA will lead to the synthesis of a large amount of DNA which will not only incur high costs, time and labor, but will also lead to the generation of mostly toxic chemical wastes (Zhirnov et al., 2016). Rewritable hard drives made up of DNA may tackle the challenges posed by conventional DNA data storage. For instance, the M13mp18 genome was used as a rewritable hard drive to store digital information in pre-annotated domains of the DNA (Chen et al., 2020). Inspired by this work, a rewritable information storage system based on the one-time synthesis of rewritable DNA hard drive can be developed and utilized as needed. Similarly, as chemical DNA synthesis is associated with the production of toxic waste, enzymatic DNA synthesis is a feasible alternative to chemical DNA synthesis and has been utilized in the DNA data storage community (Lee et al., 2019, 2020; Yoo et al., 2021). H. H. Lee utilized TdT to synthesize DNA encoding 144 bits of data (H. H. Lee et al., 2019). Another study demonstrated the parallel synthesis of enzymatic DNA using maskless photolithography and encoded 110 bits of digital information (Lee et al., 2020). Taken together, enzymatic DNA synthesis of rewritable DNA hard drives that can eventually be translated into synthesis-free storage of digital information by encoding information into predefined regions of DNA can aid in tackling the challenges such as the cost and produced waste associated with chemical DNA synthesis.

Secondly, the scalability of DNA data storage, particularly the development of rewritable DNA templates or hard drives that are not quantitatively impacted by selected information retrieval, is still a significant hindrance. An ePCR-based method referred to as DENSE was developed to overcome the physical depletion of selectively retrieved data from the DNA pool (Kyle J. Tomek et al., 2019). DENSE utilized chemically modified oligos to produce copies of the desired file containing chemical tags, which were then selectively retrieved using magnetic beads functionalized using a variety of chemical handles. Such approaches that can enable the selective extraction of desired files from a pool of oligos without its quantitative depletion in the original DNA pool may allow random access required for a scalable DNA data storage system.

Thirdly, the majority of the demonstrated DNA data storage systems involve in vitro experimentation which requires handling expensive reagents at a microliter scale that may multiply as the storage system is expanded, leading to an exponential increase in the cost of the experimental procedures. The reagent consumption can be reduced using a variety of methods. For instance, laboratory automation robots have been demonstrated to handle reagents as low as a few nanoliters (Formulatrix, n.d.). Similarly, miniaturization technologies such as microfluidics are well established since the early 2000s to reduce the sample required to perform experiments involving expensive reagents (Table 1) (Hong et al., 2006; Iwai et al., 2022; Marcus et al., 2006). For instance, Liu et al. developed a microfluidic platform for performing 400 unique PCR reactions with only 41 pipetting steps as compared to 1200 pipetting steps using conventional approaches (Liu et al., 2003). Similarly, Marcus et al. developed a microfluidic device for performing 72 RT-PCR reactions in 450 pL reaction chambers with sensitivity similar to conventional RT-PCRs (Marcus et al., 2006). Likewise, Iwai et al. developed a digital microfluidics-based approach for strain engineering with a sample requirement of <2 uL (Iwai et al., 2022). A significant reduction of the sample size requirement for performing experiments focusing on the writing and selective retrieval of desired digital information to and from the DNA pool can be established using microfluidic devices. For instance, microfluidic-based miniaturization has shown to reduce cost by over a million times with a 1000 fold increase in speed, and a million fold reduction in the reaction volume (Agresti et al., 2010; Leman et al., 2015; Li et al., 2014).

Lastly, parallel encoding, writing, and reading of digital information into DNA is also a critical issue that remains to be tackled. To truly replace silicon-based data storage devices or at least to become a part of a hybrid information system, the latency of information writing and retrieval of DNA memory should suffice the industrial demands. The latency of writing and retrieving information in DNA is still not practically scalable. For instance, write-to-read latency for 5 bytes of information was ~21 h (Takahashi et al., 2019). One of the several factors that impact the write-to-read latency is the labor-intensive experimental approach. Microfluidics, particularly mLSI and mVLSI, have already been demonstrated to have hundreds of individually addressable chambers with different reaction conditions, which may be adapted for DNA memory for encoding data blocks into different blank DNA templates (Li et al., 2005; Liu et al., 2003; Marcus et al., 2006; Thorsen et al., 2002; Vollertsen et al., 2020; Vyawahare et al., 2010). Taken together, emerging technologies in synthetic biology, such as enzymatic DNA synthesis, high-throughput sequencing, and integration of artificial intelligence with DNA sequencing, will play a vital role in transforming DNA data storage into an industrially scalable technology.

DNA data storage has emerged as a potential alternative or complement to conventional silicon-based information storage devices. DNA data storage technologies have primarily been focused on a synthesis-based approach for encoding digital information into DNA (Choi et al., 2019; G M Church et al., 2012; Lee et al., 2019, 2020; Seth L Shipman et al., 2017). Synthesis-based encoding of information typically involves the encoding of digital information into DNA sequences followed by the synthesis of DNA using chemical or enzymatic DNA synthesis. However, chemical DNA synthesis is associated with the generation of significant amounts of hazardous waste (Palluk et al., 2018). On the other hand, enzymatic DNA synthesis offers a comparatively environmentally friendly approach and has been explored for the synthesis of DNA for storing digital information (Lee et al., 2019, 2020; Yoo et al., 2021). At current stage, enzymatic DNA synthesis is slow, costly, and poorly developed. Following the synthesis of DNA, the DNA pools comprising oligos are stored using a variety of methods such as in vitro in dehydrated form or encapsulated in polymeric materials such as silica microparticles (Antkowiak et al., 2022; Newman et al., 2019). Once stored, information can be selectively retrieved using different methods such as nested PCR, which precisely extracts the file containing a unique barcode from the pool of DNA (Hossein et al., 2017; Lee Organick et al., 2018; Organick et al., 2020; Kyle J. Tomek et al., 2019; Yamamoto et al., 2008a, 2008b; Yazdi et al., 2015). Other approaches focusing on using silica to encapsulate the DNA into particles barcoded with unique ssDNA strands followed by retrieval of information have also been developed (James L. Banal et al., 2021). Once retrieved, the digital information is typically read using DNA sequencing. Advances in the development of portable and rapid DNA sequencing approaches have enabled the community to use DNA sequencing alternatives such as nanopore sequencing for reading the desired file extracted from the DNA pool (Chen et al., 2020; K. Chen et al., 2019; Lopez et al., 2019).

As a solution to challenges associated with synthesis-based DNA data storage, researchers have focused the development of rewritable DNA hard drives to reduce the number of DNA synthesis cycles (Chen et al., 2020; K. Chen et al., 2019). In a different approach, once synthesized, the DNA hard drive can be utilized as a template with pre-specified domains to encode the digital information. The information can be stored and rewritten using various approaches, such as strand displacement reactions (Chen et al., 2020). Such approaches offer a feasible alternative to the current DNA data storage, where DNA is repeatedly synthesized, which can impact the scalability of large-scale information storage systems due to the cost of DNA synthesis and the generation of massive amounts of chemical waste. Furthermore, the conventional approaches also require the whole DNA pool to be resynthesized if a file is to be modified; attributed largely to the currently available limited data encoding approaches.

For writing digital information into DNA, information can be encoded using a variety of molecular writers that are classified as pseudo-random writers or precise writers, depending on the nature of modifications/mutations (Farzadfard and Lu, 2018). Genome engineering technologies such as recombineering and the CRISPR-Cas system have been used for molecular recordings (Bonnet et al., 2012; Farzadfard et al., 2019; Farzadfard and Lu, 2014; Seth L Shipman et al., 2017). Similarly, artificial restriction enzymes have been used to encode digital information into predefined DNA templates (Tabatabaei et al., 2020). CRISPR-Cas-based DNA editing approaches, such as base editing and integration of zinc finger nucleases, are some of the examples of precise DNA writers that can introduce defined mutations into DNA (Kim, 2018; Komor et al., 2016; Mok et al., 2020). Adapting such technologies for storing digital information may dramatically improve the processes associated with the writing of digital information into DNA.

Likewise, the DNA data storage cycle has been automated in vitro (Takahashi et al., 2019). Digital microfluidics has also been utilized for the automated storage and retrieval of digital information in DNA (Antkowiak et al., 2022; Newman et al., 2019). The processes associated with writing the digital information involving approaches that focus on reducing the number of DNA synthesis cycles by encoding the digital information into DNA using molecular writers are performed in vitro manually. These processes are expensive due to the cost of reagents and labor. Miniaturization of the experimental procedures involved in DNA memory and their automation can be achieved using microfluidic devices. To summarize, modulation of precise DNA writers for programmable single-base resolution mutations can improve the efficiency of encoding information into DNA. The development of alternative methods for writing information on DNA that improve the information density and parallel encoding can significantly impact the information encoding process. Furthermore, microfluidic automation of DNA memory can pave the way for a scalable and sustainable DNA data storage system.

Acknowledgments

The authors acknowledge the funding from the National Science Foundation (NSF) SemiSynBio-II Award #2027738, CMMI Awards #2100739, and #2100850, and NIH Award #1R16GM145671. This work was performed at the Joint School of Nanoscience and Nanoengineering, a member of the Southeastern Nanotechnology Infrastructure Corridor (SENIC), which is supported by NSF Award #2025462.

References

Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.C., Marquez, M., Klibanov, A.M., Griffiths, A.D., Weitz, D.A., 2010. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proc. Natl. Acad. Sci. U. S. A. 107, 4004–4009. https://doi.org/10.1073/PNAS.0910781107/SUPPL_FILE/SM3.AVI.

Ailenberg, M., Rotstein, O.D., 2009. An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques 47, 747–754. https://doi.org/10.2144/000113218.

Alarcon, C.M., Shan, G., Layton, D.T., Bell, T.A., Whipkey, S., Shillito, R.D., 2019. Application of DNA- and protein-based detection methods in agricultural biotechnology. J. Agric. Food Chem. 67, 1019–1028. https://doi.org/10.1021/ACS.JAFC.8B05157/ASSET/IMAGES/LARGE/JF-2018-05157K_0005.JPEG.

Allentoft, M.E., Collins, M., Harker, D., Haile, J., Oskam, C.L., Hale, M.L., Campos, P.F., Samaniego, J.A., Gilbert, M.T.P., Willerslev, E., Zhang, G., Scofield, R.P., Holdaway, R.N., Bunce, M., 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proceedings of the Royal Society B: Biological Sciences 279, 4724–4733. https://doi.org/10.1098/rspb.2012.1745.

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 1 (21), 1–16. https://doi.org/10.1186/S13059-020-1935-5.

Anavy, L., Vaknin, I., Atar, O., Amit, R., Yakhini, Z., 2019. Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nat. Biotechnol. 10 (37), 1229–1236. https://doi.org/10.1038/s41587-019-0240-x.

Antkowiak, P.L., Lietard, J., Darestani, M.Z., Somoza, M.M., Stark, W.J., Heckel, R., Grass, R.N., 2020. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. Nat. Commun. 11, 1–10. https://doi.org/10.1038/s41467-020-19148-3.

Antkowiak, P.L., Koch, J., Nguyen, B.H., Stark, W.J., Strauss, K., Ceze, L., Grass, R.N., 2022. Integrating DNA encapsulates and digital microfluidics for automated data storage in DNA. Small 18, 2107381. https://doi.org/10.1002/smll.202107381.

Araci, I.E., Quake, S.R., 2012. Microfluidic very large scale integration (mVLSI) with integrated micromechanical valves. Lab Chip 12, 2803–2806. https://doi.org/10.1039/c2lc40258k.

Arter, W.E., Yusim, Y., Peter, Q., Taylor, C.G., Klenerman, D., Keyser, U.F., Knowles, T.P.J., 2020. Digital sensing and molecular computation by an enzyme-free DNA circuit. ACS Nano 14, 5763–5771.

Banal, James L., Shepherd, T.R., Berleant, J., Huang, H., Reyes, M., Ackerman, C.M., Blainey, P.C., Bathe, M., 2021. Random access DNA memory using Boolean search in an archival file storage system. Nat. Mater. 20, 1272–1280. https://doi.org/10.1038/s41563-021-01021-3.

Bancroft, C., Bowler, T., Bloom, B., Clelland, C., 2001. Long-term storage of information in DNA. Science 1979. https://doi.org/10.1126/SCIENCE.293.5536.1763C.

Bate, G., 1978. Bits and genes: A comparison of the natural storage of information in DNA and digital magnetic recording. IEEE Trans. Magn. 14, 964–965.

Blawat, M., Gaedke, K., Huetter, I., Chen, X., Turczyk, B., Inverso, S.A., Pruitt, B., Church, G., 2016. Forward Error Correction for DNA Data Storage. https://doi.org/10.1016/j.procs.2016.05.398.

Blois, M., 2022. Enzymatic DNA synthesis gets a significant new player. Chem. Eng. News. https://doi.org/10.1021/CEN-10003-BUSCON3.

Bonnet, J., Subsoontorn, P., Endy, D., 2012. Rewritable digital data storage in live cells via engineered control of recombination directionality. Proc. Natl. Acad. Sci. U. S. A. 109, 8884–8889. https://doi.org/10.1073/pnas.1202344109.

Bornholt, J., Lopez, R., Carmean, D., Ceze, L., Seelig, G., Strauss, K., 2016. A DNA-based archival storage system. In: Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems.. https://doi.org/10.1145/2872362.2872397.

Briones, J., Espulgar, W., Koyama, S., Takamatsu, H., Tamiya, E., Saito, M., 2021. A design and optimization of a high throughput valve based microfluidic device for single cell compartmentalization and analysis. Sci. Rep. 11, 1–12. https://doi.org/10.1038/s41598-021-92472-w.

Cao, B., Zhang, X., Cui, S., Zhang, Q., 2022. Adaptive coding for DNA storage with high storage density and low coverage. npj Syst. Biol. Appl. 8, 1–12. https://doi.org/10.1038/s41540-022-00233-w.

Chandrasekaran, A.R., Levchenko, O., Patel, D.S., Macisaac, M., Halvorsen, K., 2017. Addressable configurations of DNA nanostructures for rewritable memory. Nucleic Acids Res. 45, 11459–11465. https://doi.org/10.1093/nar/gkx777.

Chen, K., Zhu, J., Bošković, F., Keyser, U.F., 2020. Nanopore-based dna hard drives for rewritable and secure data storage. Nano Lett. 20, 3754–3760. https://doi.org/10.1021/acs.nanolett.0c00755.

Chen, W.D., Kohll, A.X., Nguyen, B.H., Koch, J., Heckel, R., Stark, W.J., Ceze, L., Strauss, K., Grass, R.N., 2019. Combining data longevity with high storage capacity—layer-by-layer DNA encapsulated in magnetic nanoparticles. Adv. Funct. Mater. 29, 1901672. https://doi.org/10.1002/adfm.201901672.

Choi, Y., Ryu, T., Lee, A.C., Choi, H., Lee, H., Park, J., Song, S.H., Kim, S., Kim, H., Park, W., Kwon, S., 2019. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. Sci. Rep. 9, 6582. https://doi.org/10.1038/s41598-019-43105-w.

Church, G.M., Gao, Y., Kosuri, S., 2012. Next-generation digital information storage in DNA. Science (1979) 337, 1628. https://doi.org/10.1126/science.1226355.

Clelland, C., Risca, V.I., Bancroft, C., 1999. Hiding messages in DNA microdots. Nature. https://doi.org/10.1038/21092.

Davis, J., 1996. Microvenus. Art J. 55, 70–74.

Dey, S., Fan, C., Gothelf, K.V., Li, J., Lin, C., Liu, L., Liu, N., Nijenhuis, M.A.D., Saccà, B., Simmel, F.C., Yan, H., Zhan, P., 2021. DNA origami. Nat. Rev. Methods Prim. 1 https://doi.org/10.1038/s43586-020-00009-8.

Dickinson, G.D., Mortuza, G.M., Clay, W., Piantanida, L., Green, C.M., Watson, C., Hayden, E.J., Andersen, T., Kuang, W., Graugnard, E., Zadegan, R., Hughes, W.L., 2021. An alternative approach to nucleic acid memory. Nat. Commun. 12, 2371. https://doi.org/10.1038/s41467-021-22277-y.

Dimopoulou, M., Antonini, M., 2021. Image storage in DNA using vector quantization. European Signal Processing Conference 2021-January, 516–520. Doi:10.23919/EUSIPCO47968.2020.9287470.

Dimopoulou, M., Antonini, M., Barbry, P., Appuswamy, R., 2021. Image storage onto synthetic DNA. Signal Process. Image Commun. 97, 116331 https://doi.org/10.1016/J.IMAGE.2021.116331.

Eisenstein, M., 2020. Enzymatic DNA synthesis enters new phase. Nat. Biotechnol. 38, 1113–1115. https://doi.org/10.1038/S41587-020-0695-9.

Enghiad, B., Zhao, H., 2017. Programmable DNA-guided artificial restriction enzymes. ACS Synth. Biol. 6, 752–757. https://doi.org/10.1021/ACSSYNBIO.6B00324/SUPPL_FILE/SB6B00324_SI_001.PDF.

Erlich, Y., Zielinski, D., 2017. DNA fountain enables a robust and efficient storage architecture. Science 1979 (355), 950–954. https://doi.org/10.1126/SCIENCE.AAJ2038/SUPPL_FILE/ERLICH.SM.PDF.

F R, J, S, V, S, M, 2013. Preventing data loss by storing information in bacterial DNA. Int. J. Comput. Appl. 69, 53–57. https://doi.org/10.5120/12083-8322.

Fan, C., Deng, Q., Zhu, T.F., 2021. Bioorthogonal information storage in L-DNA with a high-fidelity mirror-image Pfu DNA polymerase. Nat. Biotechnol. https://doi.org/10.1038/s41587-021-00969-6.

Farzadfard, F., Lu, T.K., 2014. Genomically encoded analog memory with precise in vivo dna writing in living cell populations. Science 1979, 346. https://doi.org/10.1126/science.1256272.

Farzadfard, F., Lu, T.K., 2018. Emerging applications for DNA writers and molecular recorders. Science 1979 (361), 870–875.

Farzadfard, F., Gharaei, N., Higashikuni, Y., Jung, G.Y., Cao, J.C., Lu, T.K., 2019. Single-nucleotide-resolution computing and memory in living cells. Mol. Cell 75, 769–+. https://doi.org/10.1016/j.molcel.2019.07.011.

Formulatrix, n.d. MANTIS® - Microfluidic Liquid Handler - FORMULATRIX® [WWW Document]. URL https://formulatrix.com/liquid-handling-systems/mantis-liquid-handler/ (accessed 6.19.22).

Gach, P.C., Shih, S.C.C., Sustarich, J., Keasling, J.D., Hillson, N.J., Adams, P.D., Singh, A. K., 2016. A droplet microfluidic platform for automating genetic engineering. ACS Synth. Biol. 5, 426–433. https://doi.org/10.1021/acssynbio.6b00011.

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B., Birney, E., 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77–80. https://doi.org/10.1038/nature11875.

Gómez-Sjöberg, R., Leyrat, A.A., Pirone, D.M., Chen, C.S., Quake, S.R., 2007. Versatile, fully automated, microfluidic cell culture system. Anal. Chem. 79, 8557–8563. https://doi.org/10.1021/ac071311w.

Grass, R.N., Heckel, R., Puddu, M., Paunescu, D., Stark, W.J., 2015. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew. Chem. Int. Ed. Eng. 54, 2552–2555. https://doi.org/10.1002/anie.201411378.

Gustafsson, C., 2009. For anyone who ever said there's no such thing as a poetic gene. Nature 458, 703. https://doi.org/10.1038/458703a.

Heckel, R., Mikutis, G., Grass, R.N., 2019. A characterization of the DNA data storage channel. Sci. Rep. 9 https://doi.org/10.1038/s41598-019-45832-6.

Heinis, T., 2019. Survey of information encoding techniques for DNA. arXiv. arXiv:1906.11062.

Hong, J.W., Chen, Y., Anderson, W.F., Quake, S.R., 2006. Molecular biology on a microfluidic chip. J. Phys. Condens. Matter 18, S691. https://doi.org/10.1088/0953-8984/18/18/S14.

Hossein TabatabaeiYazdi, S.M., Gabrys, R., Milenkovic, O., Yazdi, S.M.H.T., Gabrys, R., Milenkovic, O., 2017. Portable and error-free DNA-based data storage. Sci. Rep. 7, 1–6. https://doi.org/10.1038/s41598-017-05188-1.

Hu, B., Liu, Y., Deng, J., Mou, L., Jiang, X., 2018. An on-chip valve-assisted microfluidic chip for quantitative and multiplexed detection of biomarkers. Anal. Methods 10, 2470–2480. https://doi.org/10.1039/C8AY00682B.

Hughes, R.A., Ellington, A.D., 2017. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. Cold Spring Harb. Perspect. Biol. 9, a023812 https://doi.org/10.1101/CSHPERSPECT.A023812.

Huo, H., V.S.-2007 I, 7th I.S., 2007, undefined, 2007. A Suffix Tree Construction Algorithm for DNA Sequences. ieeexplore.ieee.org. https://doi.org/10.1109/BIBE.2007.4375711.

Ijäs, H., Nummelin, S., Shen, B., Kostiainen, M.A., Linko, V., 2018. Dynamic DNA origami devices: from strand-displacement reactions to external-stimuli responsive systems. Int. J. Mol. Sci. https://doi.org/10.3390/ijms19072114.

Iwai, K., Ando, D., Kim, P.W., Gach, P.C., Raje, M., Duncomb, T.A., Heinemann, J.V., Northen, T.R., Martin, H.G., Hillson, N.J., Adams, P.D., Singh, A.K., 2018. Automated flow-based/digital microfluidic platform integrated with onsite electroporation process for multiplex genetic engineering applications. In: Proceedings of the IEEE International Conference on Micro Electro Mechanical Systems (MEMS). Institute of Electrical and Electronics Engineers Inc, pp. 1229–1232. https://doi.org/10.1109/MEMSYS.2018.8346785.

Iwai, K., Wehrs, M., Garber, M., Sustarich, J., Washburn, L., Costello, Z., Kim, P.W., Ando, D., Gaillard, W.R., Hillson, N.J., Adams, P.D., Mukhopadhyay, A., Garcia Martin, H., Singh, A.K., 2022. Scalable and automated CRISPR-based strain engineering using droplet microfluidics. Microsyst. Nanoeng. 8, 1–10. https://doi.org/10.1038/s41378-022-00357-3.

Josefsson, S., 2006. The base16, base32, and base64 Data Encodings.

Kashiwamura, S., Yamamoto, M., Kameda, A., Shiba, T., Ohuchi, A., 2003. Hierarchical DNA Memory Based on Nested PCR. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2568, pp. 112–123. https://doi.org/10.1007/3-540-36440-4_10.

Kellogg, R.A., Gómez-Sjöberg, R., Leyrat, A.A., Tay, S., 2014. High-throughput microfluidic single-cell analysis pipeline for studies of signaling dynamics. Nat. Protoc. 9, 1713–1726. https://doi.org/10.1038/nprot.2014.120.

Kiah, H.M., Puleo, G.J., Milenkovic, O., 2016. Codes for DNA sequence profiles. IEEE Trans. Inf. Theory. https://doi.org/10.1109/ISIT.2015.7282568.

Kim, J., Bae, J.H., Baym, M., Zhang, D.Y., 2020. Metastable hybridization-based DNA information storage to allow rapid and permanent erasure. Nat. Commun. 11 https://doi.org/10.1038/s41467-020-18842-6.

Kim, J.S., 2018. Precision genome engineering through adenine and cytosine base editing. Nat. Plants 4, 148–151. https://doi.org/10.1038/s41477-018-0115-z.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., Liu, D.R., 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 7603, 420–424. https://doi.org/10.1038/nature17946.

Kosuri, S., Church, G.M., 2014. Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods 11, 499–507. https://doi.org/10.1038/nmeth.2918.

Lee, H., Wiegand, D.J., Griswold, K., Punthambaker, S., Chun, H., Kohman, R.E., Church, G.M., 2020. Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. Nat. Commun. 11, 1–9. https://doi.org/10.1038/s41467-020-18681-5.

Lee, H.H., Kalhor, R., Goela, N., Bolot, J., Church, G.M., 2019. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nat. Commun. 10, 1–12. https://doi.org/10.1038/s41467-019-10258-1.

Leman, M., Abouakil, F., Griffiths, A.D., Tabeling, P., 2015. Droplet-based microfluidics at the femtolitre scale. Lab Chip 15, 753–765. https://doi.org/10.1039/C4LC01122H.

Li, N., Hsu, C.H., Folch, A., 2005. Parallel mixing of photolithographically defined nanoliter volumes using elastomeric microvalve arrays. Electrophoresis 26, 3758–3764. https://doi.org/10.1002/elps.200500171.

Li, S.Y., Liu, J.K., Zhao, G.P., Wang, J., 2018. CADS: CRISPR/Cas12a-assisted DNA steganography for securing the storage and transfer of DNA-encoded information. ACS Synth. Biol. 7, 1174–1178. https://doi.org/10.1021/acssynbio.8b00074.

Li, Y., Yan, X., Feng, X., Wang, J., Du, W., Wang, Y., Chen, P., Xiong, L., Liu, B.F., 2014. Agarose-based microfluidic device for point-of-care concentration and detection of pathogen. Anal. Chem. 86, 10653–10659. https://doi.org/10.1021/AC5026623/SUPPL_FILE/AC5026623_SI_004.AVI.

Lin, B., Hui, J., Mao, H., 2021. Nanopore technology and its applications in gene sequencing. Biosensors 11, 214. https://doi.org/10.3390/BIOS11070214.

Liu, F., Li, J., Zhang, T., Chen, J., Ho, C.L., 2022. Engineered spore-forming Bacillus as a microbial vessel for long-term DNA data storage. ACS Synth. Biol. 11, 3583–3591. https://doi.org/10.1021/ACSSYNBIO.2C00291/SUPPL_FILE/SB2C00291_SI_001.PDF.

Liu, J., Hansen, C., Quake, S.R., 2003. Solving the "world-to-chip" interface problem with a microfluidic matrix. Anal. Chem. 75, 4718–4723. https://doi.org/10.1021/ac0346407.

Lopez, R., Chen, Y.-J., Ang, S.D., Yekhanin, S., Makarychev, K., Racz, M.Z., Seelig, G., Strauss, K., Ceze, L., 2019. DNA assembly for nanopore data storage readout. Nat. Commun. 10, 1–9.

Lopiccolo, A., Shirt-Ediss, B., Torelli, E., Olulana, A.F.A., Castronovo, M., Fellermann, H., Krasnogor, N., 2021. A last-in first-out stack data structure implemented in DNA. Nat. Commun. 12, 1–10. https://doi.org/10.1038/s41467-021-25023-6.

Lu, X., Li, J., Li, C., Lou, Q., Peng, K., Cai, B., Liu, Y., Yao, Y., Lu, L., Tian, Z., Ma, H., Wang, W., Cheng, J., Guo, X., Jiang, H., Ma, Y., 2022. Enzymatic DNA synthesis by engineering terminal deoxynucleotidyl transferase. ACS Catal. 12, 2988–2997. https://doi.org/10.1021/ACSCATAL.1C04879/ASSET/IMAGES/LARGE/CS1C04879_0005.JPEG.

Marcus, J.S., Anderson, W.F., Quake, S.R., 2006. Parallel picoliter RT-PCR assays using microfluidics. Anal. Chem. 78, 956–958. https://doi.org/10.1021/AC0513865/SUPPL_FILE/AC0513865SI20051111_090452.PDF.

McCord, B.R., Gauthier, Q., Cho, S., Roig, M.N., Gibson-Daw, G.C., Young, B., Taglia, F., Zapico, S.C., Mariot, R.F., Lee, S.B., Duncan, G., 2019. Forensic DNA analysis. Anal. Chem. 91, 673–688. https://doi.org/10.1021/ACS.ANALCHEM.8B05318/ASSET/IMAGES/ACS.ANALCHEM.8B05318.SOCIAL.JPEG_V03.

McNaughton, A.L., Roberts, H.E., Bonsall, D., de Cesare, M., Mokaya, J., Lumley, S.F., Golubchik, T., Piazza, P., Martin, J.B., de Lara, C., Brown, A., Ansari, M.A., Bowden, R., Barnes, E., Matthews, P.C., 2019. Illumina and Nanopore methods for whole genome sequencing of hepatitis B virus (HBV). Sci. Rep. 1 (9), 1–14. https://doi.org/10.1038/s41598-019-43524-9.

Meiser, L.C., Antkowiak, P.L., Koch, J., Chen, W.D., Kohll, A.X., Stark, W.J., Heckel, R., Grass, R.N., 2020. Reading and writing digital data in DNA. Nat. Protoc. 15, 86–101. https://doi.org/10.1038/s41596-019-0244-5.

Melin, J., Quake, S.R., 2007. Microfluidic large-scale integration: the evolution of design rules for biological automation. Annu. Rev. Biophys. Biomol. Struct. https://doi.org/10.1146/annurev.biophys.36.040306.132646.

Mok, B.Y., de Moraes, M.H., Zeng, J., Bosch, D.E., Kotrys, A.V., Raguram, A., Hsu, F.S., Radey, M.C., Peterson, S.B., Mootha, V.K., Mougous, J.D., Liu, D.R., 2020. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature 583, 631–637. https://doi.org/10.1038/s41586-020-2477-4.

Nayak, S., Access, R.P., 2019. A Review on Role of Bloom Filter on DNA Assembly. ieeexplore.ieee.org.

Newman, S., Stephenson, A.P., Willsey, M., Nguyen, B.H., Takahashi, C.N., Strauss, K., Ceze, L., 2019. High density DNA data storage library via dehydration with digital microfluidic retrieval. Nat. Commun. 10, 1706. https://doi.org/10.1038/s41467-019-09517-y.

Nguyen, H.H., Park, J., Park, S.J., Lee, C.S., Hwang, S., Shin, Y.B., Ha, T.H., Kim, M., 2018. Long-term stability and integrity of plasmid-based DNA data storage. Polymers (Basel) 10, 28. https://doi.org/10.3390/polym10010028.

Odeh, R.E., Knuth, D.E., 1969. The art of computer programming. Volume 1: fundamental algorithms. J. Am. Stat. Assoc. 64 https://doi.org/10.2307/2283757.

Organick, L., Ang, S.D., Chen, Y.J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M.Z., Kamath, G., Gopalan, P., Nguyen, B., Takahashi, C.N., Newman, S., Parker, H.Y., Rashtchian, C., Stewart, K., Gupta, G., Carlson, R., Mulligan, J., Carmean, D., Seelig, G., Ceze, L., Strauss, K., 2018. Random access in large-scale DNA data storage. Nat. Biotechnol. 36, 242–248. https://doi.org/10.1038/nbt.4079.

Organick, L., Chen, Y.-J., Ang, S.D., Lopez, R., Liu, X., Strauss, K., Ceze, L., 2020. Probing the physical limits of reliable DNA data retrieval. Nat. Commun. https://doi.org/10.1038/s41467-020-14319-8.

Padhy, P., Zaman, M.A., Jensen, M.A., Cheng, Y.-T., Huang, Y., Galambos, L., Davis, R. W., Hesselink, L., 2022. Bead-Droplet Reactor for High-Fidelity Solid-Phase Enzymatic DNA Synthesis.

Palluk, S., Arlow, D.H., de Rond, T., Barthel, S., Kang, J.S., Bector, R., Baghdassarian, H. M., Truong, A.N., Kim, P.W., Singh, A.K., Hillson, N.J., Keasling, J.D., 2018. De novo DNA synthesis using polymerase-nucleotide conjugates. Nat. Biotechnol. 7 (36), 645–650. https://doi.org/10.1038/nbt.4173.

Pan, C., Tabatabaei, S.K., Tabatabaei Yazdi, S.M.H., Hernandez, A.G., Schroeder, C.M., Milenkovic, O., 2022. Rewritable two-dimensional DNA-based data storage with machine learning reconstruction. Nat. Commun. 1 (13), 1–12. https://doi.org/10.1038/s41467-022-30140-x.

Parvez, S., Herdman, C., Beerens, M., Chakraborti, K., Harmer, Z.P., Yeh, J.R.J., MacRae, C.A., Joseph Yost, H., Peterson, R.T., 2021. MIC-drop: A platform for large-scale in vivo CRISPR screens. Science 1979 (373), 1146–1151. https://doi.org/10.1126/science.abi8870.

Portney, N.G., Wu, Y., Quezada, L.K., Lonardi, S., Ozkan, M., 2008. Length-based encoding of binary data in DNA. Langmuir 24, 1613–1616. https://doi.org/10.1021/la703235y.

Rothemund, P.W.K., 2006. Folding DNA to create nanoscale shapes and patterns. Nature 7082 (440), 297–302. https://doi.org/10.1038/nature04586.

Shipman, S.L., Nivala, J., Macklis, J.D., Church, G.M., 2016. Molecular recordings by directed CRISPR spacer acquisition. Science 1979, 353. https://doi.org/10.1126/science.aaf1175.

Shipman, Seth L., Nivala, J., Macklis, J.D., Church, G.M., 2017. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature 547, 345–349.

Smith, G.C., Fiddes, C.C., Hawkins, J.P., Cox, J.P.L., 2003. Some possible codes for encrypting data in DNA. Biotechnol. Lett. 25, 1125–1130.

Song, L., Geng, F., Gong, Z.Y., Chen, X., Tang, J., Gong, C., Zhou, L., Xia, R., Han, M.Z., Xu, J.Y., Li, B.Z., Yuan, Y.J., 2022. Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. Nat. Commun. 1 (13), 1–9. https://doi.org/10.1038/s41467-022-33046-w.

Tabatabaei, S.K., Wang, B., Athreya, N.B.M., Enghiad, B., Hernandez, A.G., Fields, C.J., Leburton, J.-P., Soloveichik, D., Zhao, H., Milenkovic, O., 2020. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. Nat. Commun. 11, 1–10.

Takahashi, C.N., Nguyen, B.H., Strauss, K., Ceze, L., 2019. Demonstration of end-to-end automation of DNA data storage. Sci. Rep. 9, 1–5. https://doi.org/10.1038/s41598-019-41228-8.

Thorsen, T., Maerkl, S.J., Quake, S.R., 2002. Microfluidic large-scale integration. Science (1979) 298, 580–584. https://doi.org/10.1126/science.1076996.

Tomek, Kyle J., Volkel, K., Simpson, A., Hass, A.G., Indermaur, E.W., Tuck, J.M., Keung, A.J., 2019. Driving the scalability of DNA-based information storage systems. bioRxiv. https://doi.org/10.1101/591594.

Unger, M.A., Chou, H.P., Thorsen, T., Scherer, A., Quake, S.R., 2000. Monolithic microfabricated valves and pumps by multilayer soft lithography. Science 1979 (288), 113–116. https://doi.org/10.1126/science.288.5463.113.

Vollertsen, A.R., de Boer, D., Dekker, S., Wesselink, B.A.M., Haverkate, R., Rho, H.S., Boom, R.J., Skolimowski, M., Blom, M., Passier, R., van den Berg, A., van der Meer, A.D., Odijk, M., 2020. Modular operation of microfluidic chips for highly parallelized cell culture and liquid dosing via a fluidic circuit board. Microsyst. Nanoeng. 1 (6), 1–16. https://doi.org/10.1038/s41378-020-00216-z.

Vyawahare, S., Griffiths, A.D., Merten, C.A., 2010. Miniaturization and parallelization of biological and chemical assays in microfluidic devices. Chem. Biol. https://doi.org/10.1016/j.chembiol.2010.09.007.

Wang, Y., Noor-A-Rahim, M., Zhang, J., Gunawan, E., Guan, Y.L., Poh, C.L., 2019. High capacity DNA data storage with variable-length oligonucleotides using repeat accumulate code and hybrid mapping. J. Biol. Eng. 13, 1–11. https://doi.org/10.1186/s13036-019-0211-2.

Wong, P.C., Wong, K.-K., Foote, H., 2003. Organic data memory using the DNA approach. Commun. ACM 46, 95–98. https://doi.org/10.1145/602421.602426.

Xu, C., Ma, B., Gao, Z., Dong, X., Zhao, C., Liu, H., 2021. Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage. Sci. Adv. 7, 100. https://doi.org/10.1126/sciadv.abk0100.

Yachie, N., Sekiyama, K., Sugahara, J., Ohashi, Y., Tomita, M., 2007. Alignment-based approach for durable data storage into living organisms. Biotechnol. Prog. 23, 501–505. https://doi.org/10.1021/bp060261y.

Yachie, N., Ohashi, Y., Tomita, M., 2008. Stabilizing synthetic data in the DNA of living organisms. Syst. Synth. Biol. 2, 19–25. https://doi.org/10.1007/s11693-008-9020-5.

Yamamoto, M., Kashiwamura, S., Ohuchi, A., 2008a. DNA memory with 16.8M addresses, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 99–108. https://doi.org/10.1007/978-3-540-77962-9_10.

Yamamoto, M., Kashiwamura, S., Ohuchi, A., Furukawa, M., 2008b. Large-scale DNA memory based on the nested PCR. In: Natural Computing. Springer, pp. 335–346. https://doi.org/10.1007/s11047-008-9076-x.

Yazdi, S.M., Yuan, Y., Ma, J., Zhao, H., Milenkovic, O., 2015. A rewritable, Random-Access DNA-Based storage system. Sci. Rep. 5, 14138. https://doi.org/10.1038/srep14138.

Yim, S.S., McBee, R.M., Song, A.M., Huang, Y.M., Sheth, R.U., Wang, H.H., 2021. Robust direct digital-to-biological data storage in living cells. Nat. Chem. Biol. 23 https://doi.org/10.1038/s41589-020-00711-4.

Yoo, E., Choe, D., Shin, J., Cho, S., Cho, B.K., 2021. Mini review: enzyme-based DNA synthesis and selective retrieval for data storage. Comput. Struct. Biotechnol. J. 19, 2468–2476. https://doi.org/10.1016/J.CSBJ.2021.04.057.

Zhang, D.Y., Winfree, E., 2009. Control of DNA strand displacement kinetics using toehold exchange. J. Am. Chem. Soc. 131, 17303–17314. https://doi.org/10.1021/ja906987s.

Zhang, T., Tian, T., Zhou, R., Li, S., Ma, W., Zhang, Y., Liu, N., Shi, S., Li, Q., Xie, X., 2020. Design, fabrication and applications of tetrahedral DNA nanostructure-based multifunctional complexes in drug delivery and biomedical treatment. Nat. Protoc. 15, 2728–2757.

Zhirnov, V., Zadegan, R.M., Sandhu, G., Church, G.M., Hughes, W., 2016. Nucleic acid memory. Nat. Mater. https://doi.org/10.1038/nmat4594.

Zhong, Y., Qi, S., Sheng, F., Tian, J., Zhu, P., Yang, P., Cai, X., 2018. A new digital information storing and reading system based on synthetic DNA. Sci. China Life Sci. 61, 733–735. https://doi.org/10.1007/s11427-017-9131-7.

Zhou, Z., Brennan, J.D., Li, Y., 2020. A multi-component all-DNA biosensing system controlled by a DNAzyme. Angew. Chem. Int. Ed. 59, 10401–10405.