Published online 9 April 2021 *Nucleic Acids Research, 2021, Vol. 49, No. 10* ***5451–5469***
https://doi.org/10.1093/nar/gkab230

SURVEY AND SUMMARY

Uncertainties in synthetic DNA-based data storage

Chengtao Xu[†], Chao Zhao[†], Biao Ma and Hong Liu[*]

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu 210096, China

Received November 11, 2020; Revised February 16, 2021; Editorial Decision March 18, 2021; Accepted March 22, 2021

ABSTRACT

Deoxyribonucleic acid (DNA) has evolved to be a naturally selected, robust biomacromolecule for gene information storage, and biological evolution and various diseases can find their origin in uncertainties in DNA-related processes (e.g. replication and expression). Recently, synthetic DNA has emerged as a compelling molecular media for digital data storage, and it is superior to the conventional electronic memory devices in theoretical retention time, power consumption, storage density, and so forth. However, uncertainties in the *in vitro* DNA synthesis and sequencing, along with its conjugation chemistry and preservation conditions can lead to severe errors and data loss, which limit its practical application. To maintain data integrity, complicated error correction algorithms and substantial data redundancy are usually required, which can significantly limit the efficiency and scale-up of the technology. Herein, we summarize the general procedures of the state-of-the-art DNA-based digital data storage methods (e.g. write, read, and preservation), highlighting the uncertainties involved in each step as well as potential approaches to correct them. We also discuss challenges yet to overcome and research trends in the promising field of DNA-based data storage.

INTRODUCTION

Modern society is characterized by advanced information technologies, generating increasingly vast amounts of digital data (1,2). Therefore, a variety of data storage techniques have been developed, including those based on magnetic media (e.g. tape, hard disk drive), electronic media (e.g. flash memory), and optical media (e.g. digital video disc, blue-ray disc). However, the amount of data that can be stored on the conventional media will very soon be exceeded by one that humans will produce, not to mention for most data media the retention time is only about 10 years. It is estimated that the data generated worldwide in 2025 will reach 175 zettabytes ($\approx 10^{21}$ bytes), and then 3 yottabytes ($\approx 10^{24}$ bytes) in 2040, which will greatly challenge the storage capability of our digital memory if a new medium is still not available at that time (https://www.seagate.com/our-story/data-age-2025/, (3,4)). Therefore, novel data storage media with high performance in data density, retention time, energy cost and read-write speed are urgently needed.

Since billions of years ago, DNA has been naturally selected as the storage medium of gene information which serves as the blueprint to construct and maintain the most intricate biological system in the world. The naturally selected DNA is also an extraordinary candidate as a new data storage medium featuring high volumetric density, long retention time, and low energy cost (5). As a promising biological molecular data storage media, the most significant advantage would be its unprecedented data density (i.e. theoretical density up to 455 EB g^{-1} (6)), which is approximately three orders of magnitude greater than that of the state-of-the-art flash memories (5,7,8). Although reading this amount of data at a high speed is still challenging even with the well-established second generation high-throughput sequencing (9), it can still serve as the complementary to traditional storage media when high density and long retention time are more critical than short access time (e.g. for information archiving and backup), not to mention the promising third generation single-molecule DNA sequencing technique is on its way with long read lengths and real-time data acquisition (10).

DNA is also a remarkably stable biomacromolecule that can be preserved for thousands of years if kept away from high humidity, irradiation, and air. The molecule chain can accommodate any base sequence and form a double helix structure with specific Watson–Crick base pairing. The base pairs are formed based on the weak and reversible hydrogen bonding, which enables duplication and transcription of the gene information with small amounts of energy consump-

[*] To whom correspondence should be addressed. Tel: +86 25 83792245; Email: liuh@seu.edu.cn
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

tion. Taking advantage of these unique features, enormous replicas of DNA can be accurately prepared in a short time starting from a single copy (e.g. the polymerase chain reaction, PCR) (11,12). The advances in biotechnology also facilitate facile DNA synthesis and editing of the DNA sequence. Therefore, writing, copy, and paste of data into the DNA molecule can be rather efficient in the near future.

In the past few years, many efforts have been devoted to encoding non-biological information in the synthetic DNA sequences for digital data storage (13–16). Since current DNA synthesis and sequencing techniques inevitably lead to various uncertainties in data storage, many researchers have been trying to figure out and deal with the data errors, which can be one of the major challenges for large-scale DNA-based data storage. Therefore, in this review, we first give an overview of the mainstream procedures for synthetic DNA-based data storage and then discuss the uncertainties involved in each step, highlighting the mechanisms and potential methods for error correction. Other than data writing and reading, we also summarize the immobilization and preservation of DNA that are also highly relevant to the robustness of DNA-based data storage. At last, the challenges and future research trends toward the large-scale application of DNA-based data storage are also discussed.

OVERVIEW OF DNA-BASED DATA STORAGE

Akin to conventional electronic memory, the operation of the DNA-based data storage system generally involves five major steps: encoding, writing, preservation, retrieval and decoding, as illustrated in Figure 1.

Encoding

It is well known that the DNA sequence consists of four different bases (i.e. adenine (A), thymine (T), cytosine (C), guanine (G)) that are arranged in a specific order. The first step for DNA-based data storage is to transfer a binary data stream into quaternary DNA base sequences. Considering the sequence length of DNA that can be currently synthesized with relatively high fidelity, original binary data is usually broken into chunks and transferred into DNA sequences with lengths no more than two hundred nucleotides (nt). For accurate data recovery by the reassembly of these chunks, strategies based on data overlapping in adjacent chunks or adding addressable index in each chunk are often involved (13). Although the former is well established in gene assembly, the latter has proven to be more efficient for large databases (e.g. on the petabyte scale). Besides, error detection and correction algorithms, including those based on Fountain codes or Reed-Solomon codes (RS codes), are usually applied to deal with the errors in subsequent processes such as DNA synthesis and sequencing. After the encoding step, the DNA sequences are obtained for the next synthesis step.

Writing

Current nucleic acid synthesis methods are based on chemical or enzymatic methods. The chemical method usually relies on phosphoramidite chemistry developed by Caruthers and co-workers (17,18). To be specific, a nucleotide protected by a photolabile or acid-labile group at the 5′ end is deblocked under light or exposed to acid, then another nucleotide can react with the deblocked nucleotide, forming a phosphite internucleotide linkage. The array-based chemical synthesis, which can produce a large number of DNA strands in parallel, is usually preferred over column-based synthesis for DNA-based data storage. Enzymatic synthesis based on terminal deoxynucleotidyl transferase (TdT) has been of great interest owing to its advantages in speed, efficiency, and costs compared with the conventional methods (19,20), so it has been developed rapidly as a compelling candidate for DNA synthesis in data storage applications.

Preservation

Proper preservation condition is indispensable for the long-term integrity of DNA databases (21,22). It is well known that high humidity, light, and heat can induce irreversible damage to DNA molecules, leading to data errors or even complete loss of data. Meanwhile, in certain cases, DNA strands have to be immobilized on various carriers, so the linkage chemistry also directly determines the robustness of the system. Favourable preservation conditions in combination with proper immobilization promise the outstanding lifetime of DNA databases up to thousands of years.

Data retrieval

To selectively read certain data rather than sequentially read the whole database, specific chunks in the DNA pool need to be extracted and assembled, akin to random data access in the conventional electronic memories. However, unlike flash drivers or hard disks, locating the specific DNA strands with the desired data is difficult, especially for DNA that is freely dissolved in a solution. So base sequences for specific address information should be introduced in the encoding step. To retrieve the data, two approaches have been usually involved, which are based on DNA extraction (e.g. binding to a complementary probe immobilized on magnetic beads) and selective PCR amplification of required sequences using a specific primer, respectively. However, it is still very challenging to achieve random access for a very large database on the terabyte or even petabyte scale. The DNA sequence for address information will be prohibitively long, which can greatly affect the efficiency of encoding and accuracy of data locating (23).

Sequencing and decoding

Taking advantage of the high-throughput sequencing technique, the human genome sequence with a size of 3.2 gigabytes (GB) can be obtained within hours (24). The commercial DNA sequencers can be used to read enormous sequences in the DNA data storage system followed by using decoding algorithms to convert these sequences back to the original data. More compact third-generation single-molecule sequencers with greater reading lengths and ability of real-time sequencing (e.g. Oxford Nanopore Technology,

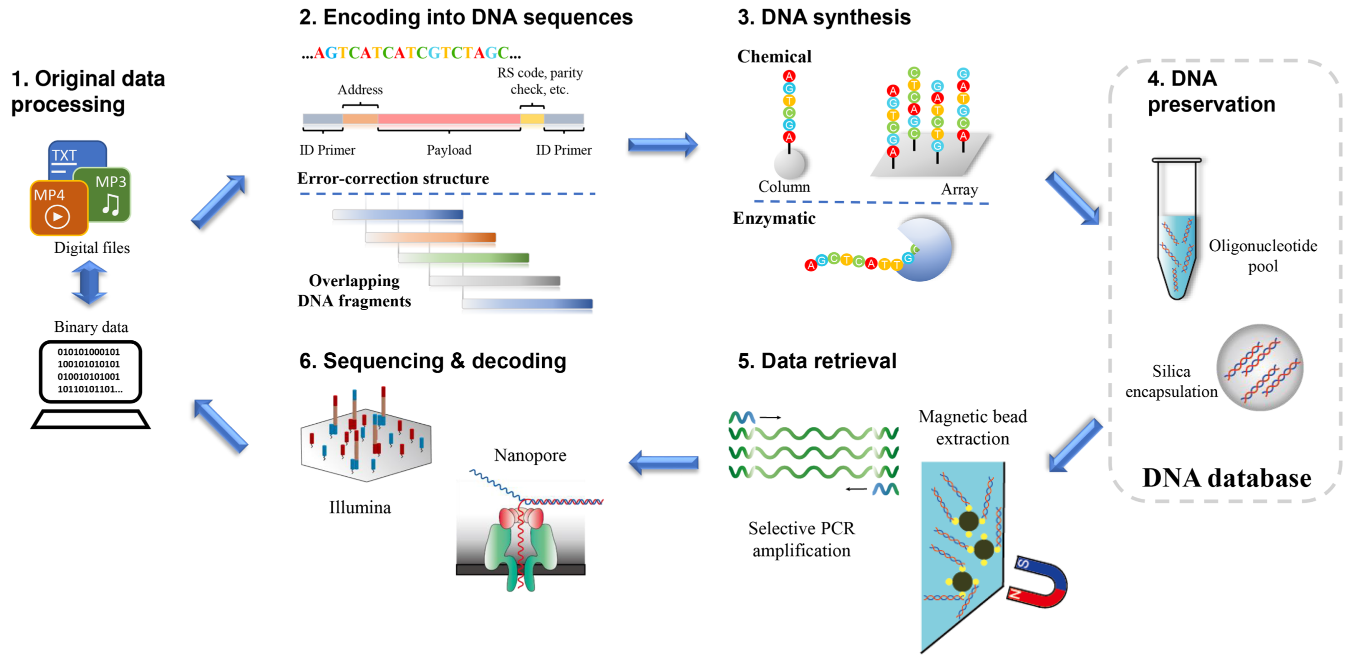

Figure 1. Schematic diagram of the DNA-based data storage and its operation. (**1**) The original data was transformed into binary data. (**2**) Binary data was encoded into corresponding DNA sequences. One usual strategy was to split the binary data into chunks, each of which was converted into a specific DNA base sequence. Additional base sequences including address information, error correction, and DNA amplification were also attached to the data sequence (16). Another approach was to split the data into chunks and transfer them to sequences having overlapping DNA fragments, which served as both address information and data redundancy for error correction (13). (**3**) All DNA strands with the specific sequences were synthesized, either chemically or enzymatically. The chemical synthesis included column-based and array-based phosphoramidite methods, while the enzymatical synthesis relied on polymerases such as template-independent terminal transferase. (**4**) The synthesized DNA samples were preserved in aqueous solutions or encapsulated in silica beads for preservation. (**5**) Data was retrieved by the selective extraction or amplification of relevant DNA strands in the pool. This could be achieved by using magnetic beads with specific ligands or polymerase chain reaction. (**6**) The relevant DNA sequences were obtained by a DNA sequencer, which was usually based on a sequencing-by-synthesis method or a nanopore. Then the DNA base sequences were decoded to obtain the original binary data.

ONT) can also be utilized in the future for the data reading (25), which can potentially reduce the reading time and the size of the DNA memory devices.

Note that the DNA-based database is essentially a chemical system governed by chemical thermodynamics and kinetics, so each step aforementioned, except the encoding and decoding steps, is error-prone like the conventional digital data storage system. Therefore, to establish a robust and efficient DNA database, uncertainties and their correcting methods should be concerned, as discussed in detail in the following sections.

UNCERTAINTIES IN DNA-BASED DATABASE

The origin of uncertainties in DNA-based data storage has been of great interest. It is believed that the main sources of errors are from DNA synthesis and sequencing. For example, Bornholt *et al.* (26) found that phosphoramidite-based synthesis and Illumina sequencing led to an error rate of about 1% in the final consequence (i.e. one error out of 100 bases), and the majority of the errors were from the sequencing process. A prior work from Organick *et al.* (14) reported a higher error rate of up to ∼10% from nanopore sequencing. To avoid data corruption, advanced error detection and correction algorithms are requisite for robust DNA-based data storage. In this section, uncertainties, their origins, and approaches for error correction are discussed. Key parameters in recent works are summarized in Table 1.

Uncertainties in DNA synthesis

One of the key steps in DNA storage is to synthesize the DNA strands having data-encoded sequences. Current DNA synthesis can be realized by chemical and enzymatic methods. As early as the 1970s, Caruthers *et al.* reported the phosphoramidite-based oligonucleotide synthesis, which has been widely used for decades (17,18,27,28). In a typical synthesis, nucleotide monomers with hydroxyl groups at the 5′ terminal were blocked with a protecting group (e.g. dimethoxytrityl (DMT) group). The protecting group was labile, which could be detached under certain conditions (e.g. acid, light). Then the deblocked nucleotide was exposed to the next monomer for coupling. To generate a specific sequence, only one type of nucleotide was introduced to the reaction pool at a time for the coupling reaction, which was repeated until a DNA molecule of hundreds of nucleotides was obtained. After the introduction of each nucleotide into the DNA strand, the phosphite linkage was further converted to a more stable phosphate bond using moderate oxidant.

The chemical synthesis was usually accomplished on a glass substrate with controlled pores (29,30). Initial nucleotides were seeded on the glass via linking to the surface hydroxyl groups. The chemical reactions were accom-

Table 1. Summary of current works on DNA-based data storage. The table is presented chronologically with the earliest on top. The encoding techniques are illustrated in detail in Figure 2. The error correction methods refer to the approaches shown in Figure 3. Synthesis technologies include the phosphoramidite-based chemical synthesis and the emerging enzyme-based synthesis. Random access indicates the ability to acquire specific pieces of data from the whole data pool of the DNA storage system. Sequencing technologies include Illumina's high-throughput technology based on the sequencing-by-synthesis principle, and ONT's nanopore technology with real-time data reading. The information density denotes the average number of binary information (bits) encoded in each nucleotide (nt). The binary information includes the data payload and the auxiliary sequences for error-correction, primers, and so forth. Table 1 is adapted from reference (14,16)

	Coding features	Data size	Error correction method	Synthesis technology	Random access	Sequencing technology	Coverage	Information density (bits/nt)
Church *et al.* (15)	Conventional	650KB	Repetition	Phosphoramidite (Agilent)	No	Illumina	3000×	0.60
Goldman *et al.* (13)	Conventional	750KB	Repetition (Overlapped segments)	Phosphoramidite (Agilent)	No	Illumina	51×	0.19
Grass *etal.* (108)	Conventional	80KB	RS codes	Phosphoramidite (CustomArray)	No	Illumina	372×	0.86
Bornholt *et al.* (26)	Conventional	150KB	Repetition (XOR operation)	Phosphoramidite	Yes	Illumina	40×	0.57
Blawat *et al.* (109)	Conventional	22MB	RS codes	Phosphoramidite (Agilent)	No	Illumina	160×	0.89
Yazdi *et al.* (107)	Conventional	3KB	Consensus estimation	Phosphoramidite (IDT)	Yes	Nanopore	200×	1.71
Erlich and Zirlinsky (16)	Conventional	2.15MB	Fountain codes	Phosphoramidite (Twist)	No	Illumina	10.5×	1.18
Organick *et al.* (14)	Conventional	200.2MB/32KB	Repetition (XOR operation) & RS codes	Phosphoramidite (Twist)	Yes	Illumina / Nanopore	5×/36×	0.81
Lopez *et al.* (77)	Conventional	1.67MB	RS codes & assembly	Phosphoramidite (Twist)	Yes	Nanopore	43×	0.84
Tomek *et al.* (63)	Conventional	94.3KB	Repetition (XOR operation)	Phosphoramidite	Yes	Illumina	10×	0.66
Organick *et al.* (6)	Conventional	19.8KB	Logical redundancy	Phosphoramidite	Yes	Illumina	35×	0.003
Anavy *et al.* (53)	Degenerate bases	8.5MB	RS codes	Phosphoramidite (Twist)	No	Illumina	164×	1.94
Choi *et al.* (54)	Degenerate bases	854 bytes	RS codes	Phosphoramidite (column)	No	Illumina	250×	1.78
Lee *et al.* (61)	Transition coding	18 bytes	Synchronization nucleotides	Enzymatic (column)	No	Nanopore	175×	1.57

plished by sequentially introducing the oxidizing agent, deblocking acid, and monomer solution, respectively, onto the substrate. These steps were repeated to synthesize DNA molecules with specific sequences. The synthesis could also be fully automated using a programmable microfluidic system to save time and reduce costs (31,32). After decades of development, various methods for the chemical synthesis of DNA have been reported, including those based on ink-jet printing (33,34), photolithography (35–37) and electrochemistry (38–40). The major difference between these methods is how the nucleotides are deblocked/activated. For example, Chow *et al.* reported the chemical DNA synthesis on a Ti array that was achieved on an undoped α-Si substrate. When the back of the substrate was exposed to light, the Ti array generated protons to trigger acidic deblocking of the nucleotide (41). Affymetrix reported another chemical DNA synthesis method based on the standard photolithography using phosphoramidite derivatives that were unstable to the light with a specific wavelength (42,43). In addition, the electrochemical microelectrode array has also been used for DNA synthesis, which is a competitive method because of its low cost and portability. The activation was usually induced by protons from an electrochemical reaction (44,45). For the construction of the DNA databases, a diverse group of DNA strands having different sequences needs to be synthesized in parallel. So, the high-throughput array-based synthesis is usually preferred over the conventional column-based synthesis.

Unfortunately, current chemical synthesis methods still have drawbacks such as low yield and synthetic errors (average error rate $\sim$0.7%) owing to substitution (0.5%), insertion (0.1%) and deletion (0.1%) (46,47). For instance, owing to incomplete coupling reaction, part of active sites remains on the substrate, which can subsequently react with the next nucleotide, leading to an erroneous insertion of a base. Likewise, incomplete deblocking of the protecting groups results in deletion errors. The residual acidic deprotection reagents may cause depurination or even cleavage of the backbone (48,49). Other problems can be associated with paralleled array-based synthesis. For example, protons produced at the reaction sites are likely to diffuse to the proximity, causing significant cross-talks between reaction spots, which is also called the edge effect (50). Therefore, array-based DNA synthesis possibly increases the error frequency and offsets the advantage of high density. Additionally, with the increasing length of the synthesized DNA, the uncertainty increased dramatically. For example, if the accuracy for adding one nucleotide is about 99.5%, the final accuracy for the synthesis of the whole DNA strand of 200 nt is only 36.7%. ($0.995^{200} = 0.367$).

Various methods have been proposed to solve these problems. For example, the remaining reaction sites can be de-activated by acetic anhydride to prevent insertion errors (30,51). LeProst *et al.* reported that unwanted depurination resulted from residual detritylation solution could be avoided by using relatively moderate reagents and by optimizing fluid introduction and washing steps (50). Error correction reaction (ECR) of the synthesized oligonucleotides via surveyor DNA endonuclease was also reported by Saaem *et al.* (52). In ECR, DNA with the mismatch was located and excised by the endonucleases, and then the fragments were reassembled by overlap extension PCR (OE-PCR) to obtain the correct strands.

To address the problem from errors in the chemical DNA synthesis and pursue a high DNA storage density, composite encoding relied on base combinations have been previously reported, as an improvement based on conventional encoding structure (53,54). For the traditional coding framework (Figure 2A), four kinds of nucleotides at one site were capable to store 2 bits of data ($\log_2 4 = 2$ bits). In the degenerated base encoding method (Figure 2B), data was encoded not by specific type of nucleotides (A, T, C, G), but by their combinations/ratios. For example, the letter R could be encoded as equal amounts of A and G at the same site on different DNA strands (R = 50% A + 50% G), and the letter D could be encoded as equal amounts of A, G and T at the same site (D = 33% A + 33% G + 33% T) on different DNA strands. So, part of random errors could be eliminated during the data reading step. Note that, to obtain the statistic base ratio, each sequence must be of sufficient physical redundancy. For example, at least reading of several DNA strands at one site could determine whether R (A: G = 1:1) or D (A: G: T = 1:1:1) was encoded. The demanded redundancy was usually acquired in the DNA synthesis step and the PCR step. Also, increased sequencing depth and corresponding statistical algorithms were required to ensure accurate data reading. Anavy *et al.* put logical redundancy into sequences via fountain codes and introduced extra composite bases in the encoding (53). Apart from four basic nucleotides, extra composite bases K (K = 50% A + 50% C) and M (M = 50% G + 50% T) were also used to encode a file of 2.12 megabytes (MB). In this way, the storage density could be improved by 24% compared with Erlich's previous work (16). Furthermore, Choi *et al.* (54) utilized an alphabet consisting of 11 additional composite bases to encode 854 kilobytes (KB) files, which greatly shortened the DNA length and led to a dramatical leap of theoretical data coding density up to 3.9 bits per character ($\log_2 15 \approx 3.9$ bits). The design of composite bases could store more data in a specific length but it required a greater number of DNA copies and sequencing depth for data reading. In other words, the application of composite bases was equivalent to reduce synthesis cost at the expense of increasing sequencing overhead. But under ideal conditions, the overall expected cost would be reduced greatly (∼52% as estimated, (53)).

The emerging enzyme-based DNA synthesis is accomplished in aqueous solutions, which is more biocompatible than the organic solutions involved in the chemical synthesis. The length of the DNA strand that can be enzymatically synthesized was estimated to be ∼8000 nt, which is greater than that of the chemical approach (55–57). For the enzymatic synthesis, deoxynucleoside triphosphates (dNTPs) were coupled consecutively with the help of template-independent DNA polymerase, such as TdT. Since TdT could lead to random coupling of any type of dNTP in the solution, the controlled addition of a specific dNTP was crucial for the enzymatic synthesis, similar to the chemical methods. Besides, the blocking strategy for chemical synthesis could also be utilized by the enzymatic synthesis. The 3′ terminal of the nucleotide monomers could be blocked while the 5′ terminal was associated with triphosphate for the chain extension. The blocking group was removed under certain conditions to expose the hydroxyl group for the downstream reaction (58,59).

Palluk *et al.* (60) reported the use of a TdT-dNTP complex linked by a covalent disulfide bond for the enzymatic DNA synthesis. When the complex was coupled with the terminal of the DNA strand, the dNTP could not be detached from the TdT, thus other dNTPs or complexes could not react with it and the reaction stopped. When the disulfide bond between TdT and dNTP was cleaved by dithiothreitol (DTT), the DNA strand was ready for the next TdT-dNTP complex addition to the terminal. The averaged stepwise yield was estimated to be 97.7% from a total of 4861 reads based on the comparison between the sequencing results and the target sequence. Alternatively, Lee *et al.* (61) combined enzymatic synthesis with nanopore sequencing for DNA-based data storage. Different from the traditional encoding method, the data was encoded using a specific transition of one nucleotide to another in the sequence, as illustrated in Figure 2C. For instance, the transition of bases from C to G represented '1' and that from A to T represented '2'. A competitive mechanism, in which the TdTs assisted the coupling of dNTPs with the strands, while apyrase served to inhibit this process, was used to controllably add several nucleotides to the DNA strand. Synchronous nucleotides were also included in the sequences to serve as fixed 'spots' between the information-encoding nucleotides, which facilitated the comparison between sequences, reconstruction as well as subsequent error correction. Despite the uncertainties in enzymatic synthesis and nanopore sequencing, this layout with the synchronous nucleotides enabled the reconstruction of template sequences. Although they contained all kinds of erroneous nucleotides, a statistical model could be used to evaluate the error probability and all the reconstructed sequences were scored. The final DNA sequences were obtained based on majority voting, in which the most probable nucleotide at each site was determined according to the statistics.

Current enzymatic synthesis is still less well-established than the chemical method, but it is a promising candidate for DNA-based data storage because of the use of less toxic reagents, mild reaction conditions, and a higher reaction rate. Moreover, longer DNA strands can be obtained using the enzymatic method, leading to a decrease of overheads for both address information and PCR primer in one strand, and consequently a remarkable increase in data density. Research in this area is now focused on optimizing the enzymatic DNA synthesis to further reduce errors.

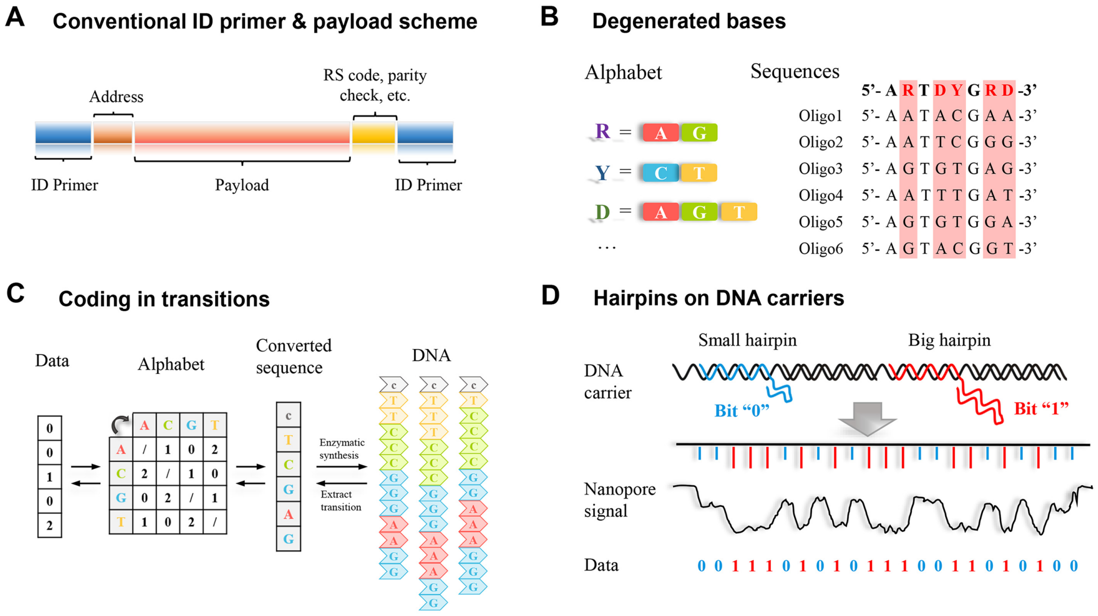

Figure 2. Encoding for DNA-based data storage. (**A**) Schematics of a DNA strand with the encoded address information, error-correction information (RS codes), data (payload), and adapters (ID primer). The address represented the relative location of the payload in the original data, which was required for data reassembly. The error-correction segment contained additional information that ensured error-free data recovery. Adapter sequences at both ends provided binding sites of complementary primers for PCR amplification and random access of certain data (16). (**B**) Schematics of the composite encoding strategy based on degenerated bases. Instead of directly transforming each binary data into four bases, the information was encoded by the ratio of A, T, C, G at the same sites on different DNA strands. For example, the letter R could be represented by A and G in a 1:1 ratio. This composite strategy improved the theoretical information density for DNA-based storage (54). (**C**) Schematics of an encoding method based on transitions of bases (e.g. C to G), which was applied in enzymatic synthesis having low synthesis accuracy but high speed. The data was represented by the transition of one nucleotide to another in the sequence. For instance, the transition of bases from C to G represented '1' and A to T represented '2' (61). (**D**) Schematics of an encoding method for nanopore-based DNA database. The data was encoded by oligonucleotide hairpins of two lengths so that two types of electronic signals could be recorded when they penetrated the nanopore. By monitoring the current signals during the DNA passed the nanopore, binary data could be obtained (80).

Uncertainties in data retrieval

Random access has been well established in the computer-based data system to selectively read data from the whole database. Random access enhances data reading efficiency by eliminating unnecessary reading tasks. However, selective retrieval of data from a large DNA database is still challenging, especially when the DNA molecules are dissolved in an aqueous solution or preserved as a lyophilized powder with no fixed physical location for direct access. Several approaches, including physical separation and PCR amplification, have been usually employed to specifically collect the DNA strands for random access of data. For the physical separation, magnetic beads carrying complementary probes or chemical labels were introduced into the pool, then the target DNA strands bound to the probes or labels and were selectively separated by a magnet. Another method was to store DNA as isolated partitions (solved in individual droplets or dehydrated as powder spots). For the PCR amplification, the target DNA strands were selectively hybridized with primers and then amplified to a predominant amount by PCR.

In a large-scale DNA database, physical partitioning of the whole database is an efficient way, considering both storage scale and random accessibility. In other words, instead of managing all data in one bulky mixed pool, dividing them into several sub-pools and retrieving data can avoid the use of a long sequence occupied by address information. Newman *et al.* (62) recently reported the use of digital microfluidics for DNA data retrieval. Dehydrated DNA was stored in sub-pools on a glass plate, which was interfaced with a digital microfluidic chip. When a file was requested, a droplet of an aqueous solution was moved to the target site via electrowetting enabled by the embedded electrode array on the chip. DNA was then dissolved in the aqueous droplet and transported for subsequent data reading. Taking advantage of the high-throughput digital microfluidics, the proposed system was promising to build automatic large-scale DNA databases. To conclude, the physical distribution of several sub-pools addressed the adversity of the crowded molecular composition in a highly dense single DNA database and improved the practical scalability of the DNA storage system. However, physical separation decreased the storage density since isolated partitions occupied more volume than a mixed large database. Besides, more duplication was needed to eliminate retrieval failure due to insufficient DNA redissolution or low separation efficiency (e.g. 9× extra redundancy applied in the work of Newman *et al.* (62)).

For traditional PCR-based random access, its ability to acquire a specific file is limited to terabyte-scale in a single pool (14). It can be interpreted that the retrieval efficiency monotonically decreases with the increasing amount of background DNA because the accurate random access of data, which highly depends on proper hybridization of the probe DNA or PCR primer, can be affected by the nontarget strands. Therefore, many efforts have been devoted to improving the orthogonality and new coding structure as data addresses (63–67). For example, Yamamoto *et al.* (65) theoretically designed a file retrieval system based on nested PCR. The concatenated nested PCR of three layers resulted in 10 million individual addresses for data storage. When a certain file was requested, the system searched for the target sequences layer by layer using cascaded PCR. In addition, the orthogonality of primers was an important concern for successful file extraction, which has been designed and optimized in previous works (14,26). Recently, Song *et al.* (67) proposed a hybrid nested and semi-nested PCR address structure based on the previous work. A virtual 4-dimensional address structure was created, in which the DNA encoded with data was assigned to an address including row, column, higher layer, and block-level information. For the random access of data, forward/reverse primer pairs from the same or different address layers were leveraged to virtually locate and retrieve data in the form of rows, columns, layers, and blocks. For example, the forward and reverse primers close to the data payload indicated the data location in the columns and rows, respectively. An additional pair of primers indicated the data location in the other two dimensions (i.e. layer and block). The virtual data structure enabled multiple random-access modes and reduced the number of primers requested from a very large address space. Theoretically, $k*n$ pairs of orthogonal primers were capable of covering n^k addresses, where k referred to the number of virtual dimensions and n represented the number of data sequences stored in each dimension. Yazdi *et al.* (64) proposed several constraints for designing PCR primers with high orthogonality. Large mutual Hamming distance and irrelevance of the addresses were beneficial to construct a robust address system. The former reduced the probability of disordered address selection, while the latter prevented the prefixes of one address from acting as suffixes of the same or another address to minimize the incorrect hybridization of primer pairs. Based on these principles, the primers designed were highly reliable for random access to the data. Tomek *et al.* (63) first reported the experimental combination of physical file separation based on magnetic extraction and nested PCR structure, which led to selective retrieval of 9.15KB of data from a database as large as 5 terabytes. The required sequences were modified with chemical labels via emulsion PCR prior to the magnetic extraction. The combined approach retrieved little unrelated DNA strands so that high sequencing efficiency and accuracy were achieved compared with that by PCR amplification alone.

However, there are also several limitations with PCR-based data retrieval. Heckel *et al.* (47) reported that the initial physical abundance was not consistent for different DNA strands due to the natural deviation in synthesis. The difference in the initial copy numbers of these DNA strands would be further magnified by PCR. Furthermore, the affinity of DNA polymerases to each primer is slightly different probably due to secondary structure formation (68,69). Commercial DNA polymerases exhibit a bias on amplification of sequences with high GC content, which significantly change the coverage and distribution of the DNA strands after cycles of amplification (70). Ideally, in each PCR cycle, sequences are amplified by a factor of two. But the actual amplification factor can be slightly below two, which also differs between sequences (PCR bias). These features lead to issues in the DNA-based database. For example, if the factor for PCR amplification of sequence A is 1.8, while the factor of sequence B is 1.9, after 60 PCR cycles, the ratio of copy numbers between these two sequences will be $(1.8/1.9)^{60} = 0.039$. Therefore, multiple PCR cycles can result in remarkably unbalanced physical redundancy for different DNA strands, leading to a biased data recovery or even complete data loss.

For these problems associated with PCR-based random access of data, there are two general solutions: (i) ensure that every sequence has sufficient and similar physical abundance in the synthesis step and reduce the number of PCR cycles appropriately to minimize PCR bias; (ii) choose orthogonal PCR primers with large affinity differences, and improve the efficiency of primers from the kinetic perspective. In other words, the more the primers are different, the less the non-specific and erroneous amplification will usually expect (71,72).

Uncertainties in DNA sequencing

Reading digital information stored in DNA strands is started from DNA sequencing. For this purpose, high-throughput second-generation sequencing (e.g. Illumina, Roche) or real-time third-generation single-molecule sequencing (e.g. Pacific Biosciences, ONT) is often involved. For instance, Illumina sequencing relies on the principle of sequencing by synthesis (73,74). DNA fragments to be sequenced were first linked with adapters at both ends and then hybridized with primers in the flow cell. Distal ends of the DNA interacted with other nearby primers to form a ‘bridge’ structure. These DNA strands subsequently underwent several rounds of ‘bridge’ amplification to generate multiple individual clusters of DNA copies. Then the sequencing step started, in which four types of fluorophore-labelled and 3′ blocked dideoxynucleotides consecutively coupled with the anchored strands, respectively. After incorporation of each dNTP to the DNA strands, the unbound dNTP was washed and the fluorescence in the flow cell was imaged to identify where this type of dNTP was coupled at the cluster. Different dNTPs were characterized by their corresponding fluorophore labels. After the fluorescent imaging, the fluorophore and blocking groups were cleaved to initiate the next cycle.

Nanopore-based technology enables real-time single-molecule DNA sequencing (10,75,76). In this technique, solid-state (e.g. graphene) or biological nanopores such as α-hemolysin, MspA porin from *Mycobacterium smegmatis*, were sandwiched by two separate chambers filled with ionic solution. When a constant voltage bias was applied, an ionic current was generated through the nanopore. The DNA was driven to pass through the nanopore towards the other chamber. An enzyme (e.g. a polymerase or a helicase) could combine with the DNA strand tightly and ratcheted the DNA through the pore step by step. The narrowest part of the nanopore (i.e. the sensing region) was sensitive to the type of nucleotide passing through, which led to different levels of blockades on the ionic current. So, the DNA sequence could be inferred by analysing the profile of the current signals.

Second-generation sequencing has shown its higher throughput and accuracy, while third-generation sequencing exhibits promising features of long read lengths and real-time readouts. Besides, the readout for nanopore-based sequencing relies on relatively simple instrumentation, such as an integrated circuit for small current measurement. So, it can be as small as a memory stick (USB). The apparatus for the second-generation sequence is relatively bulky, which usually involves sophisticated fluidic and optical systems. For the second-generation Illumina sequencing, continuous fluorescence images need to be taken and analysed for sequence information. On the other hand, for the nanopore sequencing, only the current signals produced by the DNA translocation are recorded so that the DNA sequences can be obtained in real-time. The real-time reading ability makes it superior for digital data storage. Besides, when only a few files are requested by assessing a small part of the database, nanopore sequencing is also more practical and economical than high-throughput second-generation sequencing.

For instance, Lopez *et al.* (77) reported a method for assembly of multiple DNA fragments into longer DNA strands using OE-PCR. The spliced DNA strands with overlapping primers were sequenced using a MinION nanopore sequencer from ONT. Random access for demanded data relied on multiple sequence alignments of file addresses. Data payloads with an identical address were collected and sorted for a consensus to reconstruct the original information. In their work, small fragments with the length of hundreds of nucleotides were integrated into long DNA strands, which could readily be sequenced based on nanopore technology. It is worth mentioning that the consensus algorithm was modified. The mismatched sequences were not immediately discarded, but labelled as out of sync and further searched for possible alignments in subsequent steps. Therefore, fewer coverages ($<30\times$ for three files) were required for proper decoding.

The reading length of nanopore-based DNA sequencing can be a hundred kbps (a thousand base pairs) without sophisticated sample pre-treatments (e.g. ONT MinION platform), which is much greater than that of the second-generation sequencing. Taking advantage of this, nanopore-based DNA sequencing has also promoted the development of new data encoding strategies for the DNA databases (78,79). Chen *et al.* (80) attached short DNA hairpins to double-strand DNA with different stem lengths (Figure 2D). When such a structured DNA translocated the nanopore, the hairpins blocked the nanopore and generated a secondary current decrease corresponding to the hairpin length. In their experiments, the 8 and 16 bp represented bit '0' and '1', respectively. After optimization, data of 56 bits was encoded in a DNA carrier of 7228 nt. Recently, they reported the data rewriting, deleting, and encrypting via strand displacement reaction (81). Although the efficiency of this encoding method was relatively low, the data writing through direct strand hybridization was simple, and the reading using highly portable nanopore devices avoided the use of complex instruments or vulnerable enzymes. The data storage was based on larger structures rather than single bases, which increased the signal-to-noise ratio but at the expense of decreased data density. Automated data storage devices by a combination of chemical synthesis and nanopore sequencing have been previously reported (82).

However, errors are also inevitable in DNA sequencing. For most sequencing technologies, high GC content and long homopolymers lead to sequencing errors. To be specific, the probability of insertion and deletion errors rise significantly for homopolymers consisting of more than six repeated nucleotides, especially for polyG because they tend to form guanine tetraplex structures (83). These structures are of high thermal stability, thus difficult to sequence. Additionally, sequencing coverage of DNA fragments with GC content <20% or >75% is much lower in conventional sequencing methods (84–86). At optimal conditions, Illumina sequencing accounts for <1% of errors in the retrieved data (26,87,88). Any DNA strand with a broken primer sequence, which can result from DNA damage or erroneous DNA synthesis, also cannot be read properly either for most second-generation sequencing.

For the nanopore sequencing, current fluctuation, noises, nanopore defects, and uncontrolled DNA translocation speed may result in interference and subsequent base-calling failure (89,90). For example, the sensing region of α-hemolysin is relatively large, thus the ion current can be affected by multiple nucleotides simultaneously. So, convoluted signal profiles are obtained, and difficult to correlate with a specific base (91). Besides, the signals from biological protein nanopores are highly sensitive to chemical conditions such as pH, temperature, and ionic strength, which complicate the interpretation of data and reduce reproducibility (92). As for the solid-state nanopores, despite they have a comparable dimension with the size of the DNA nucleotide, unwanted interactions with DNA nucleotides like adsorption on the substrate affect the translocation of DNA strands through the nanopore (93). Furthermore, compared with biological nanopores, the solid-state nanopores are less sensitive to different nucleotides, leading to low specificity for base recognition (94,95). Therefore, despite many advantages of nanopore sequencing, it still suffers from a relatively high error rate (∼10%) even under optimal conditions (14,96). Extensive error correction and signal analysis algorithms are highly required to improve the accuracy for practical applications.

Therefore, to eliminate the mistakes and recover loss during the DNA sequencing, the DNA strands in the database should include an appropriate amount of data redundancy. Besides, enhancing sequencing coverage and depth is an effective way widely used to ensure correct data retrieval, despite the expense of high costs. Signal processing is also of great importance for nanopore-based sequencing and accurate data recovery. Approaches based on hidden Markov model (HMM) (97–100), deep neural network (Clair (101)), likelihood alignment (Samtools/Bcftools package (102)) and so forth have been reported for the sequencing data processing (103–106). For example, in the HMM-based method, the signal of six nucleotides was treated as an event. A series of emission distribution functions were used so that the event could be correlated to the corresponding identities of the 6 nucleotides. After repeated training, the model could be used to obtain an unknown base sequence from an electric signal.

Error corrections by encoding and decoding

Error correction is one of the crucial functions of the data encoding and decoding steps. Error-correction codes (ECCs) are initially utilized in error-free information transmission and communication technology. A general idea for error correction is adding data redundancy to original data so that the errors can be corrected based on a statistical voting principle. Many efforts have been devoted to using substantial data redundancy to achieve acceptable error tolerance (26,107–109). Data redundancy can be categorized into two types: physical redundancy and logical redundancy. Physical redundancy means the presence of multiple copies of identical DNA segments, akin to backups of computer files. Note that, the physical redundancy generated either in the synthesis or in the PCR amplification process is not sufficient to eliminate all potential errors in the DNA database. Church *et al.* (15) reported that even with abundant physical redundancy added in their encoding process and high sequencing depth (3000×), manual intervention and error correction algorithm were still required owing to inevitable mistakes. Therefore, considering both efficiency and robustness, logical redundancy is usually preferred. For the logical redundancy, extra information was added into the DNA sequences apart from the original data. Usually, the encoding algorithm converted binary data into DNA sequences, which were then split into many pieces (payload). Then, a piece of sequence for error correction, an index sequence indicating the relative location of the data (address), and two terminal sequences that served as the PCR primers were linked with the payload to form a complete DNA strand in the data pool. Such an encoding framework has been proven to exhibit ideal storage capacity (110). Note that, high GC contents, homopolymers, or self-complementary sequences may lead to low synthesis yields and difficulty in sequencing, and they should be avoided in the DNA sequence.

For the data encoding with the logical redundancy, various approaches including those based on Huffman codes (111), RS codes (112,113), logical exclusive-or (XOR) operation (14,26) and fountain codes (114,115) have been reported. Huffman codes are a kind of variable length code to compress data. In Huffman encoding, each symbol in data was sorted by frequency in increasing order and regarded as several nodes. Then, two nodes (child) with the lowest frequency (e.g. 0.02 and 0.1) were combined as a new node (parent) with a summated frequency (0.02 + 0.1 = 0.12). Two steps above were iterated until all the nodes were assembled to generate a binary tree of nodes, where each node had two directions (left as '0' and right as '1') and was assigned with a unique codeword determined by the path from the root to the node. In this way, more common symbols were represented using fewer bits than less common symbols, so that the length of data was reduced.

RS codes can detect and correct multiple errors. In RS encoding, the data matrix was multiplied with a pre-calculated encoding matrix, and the redundant information would be added as new rows or columns of the result. The added rows or columns of the matrix aided to detect errors in the message and correct them if the number of errors did not exceed the correction capability (Figure 3A).

Fountain codes are a kind of codeless erasure code, in which the original data can be recovered from any subset of the encoding data which is equal to or slightly larger than the number of the original information. In the encoding process, binary data was divided into multiple non-overlapping segments of a certain length. Then, Luby transformation selected a random subset of segments according to a special distribution function, applied bitwise operation to add them up under a binary field, and wrapped them up into desired numbers of small fragments called droplets. For further recovery, each of them was assigned with seeds, which was used to generate pseudorandom numbers related to the previous transformation (Figure 3B).

As a popular ECC, RS codes are widely used for information encoding in optical disks, quick response (QR) codes, and other storage media. In previous works, RS codes have also been applied in synthetic DNA-based databases (14,77,108,109,116). For example, Grass *et al.* (108) used inner and outer RS encoding in the 2D data matrix to eliminate single-base errors or sequence loss. It was calculated that inner RS codes could correct ∼0.7 errors per 100 bases, and the outer codes rectified the loss of 0.7% further in the raw result. The cascaded encoding and in silica encapsulation strategy exhibited good robustness as demonstrated in an aging experiment, in which the DNA samples were preserved at a high temperature to accelerate the generation of errors and data loss. Based on the estimation using the Arrhenius Equation, the robust DNA data system based on the hybrid RS codes and silica encapsulation was effective for error-free preservation of DNA-based data for 1000 years in Zurich (∼ 9.4°C), or over 2 million years at the Global Seed Vault (-18°C). In another work by Organick and co-workers (14), the unprecedented 200 MB data of audio, video, and text files were stored in DNA using the RS codes. With selective PCR amplification and the Illumina sequencing, only a sequencing depth of 5× was sufficient for high-quality data retrieval without error. For nanopore sequencing with a high error rate, increased sequencing depth (e.g. 36× or 80× according to the size of files) and extracting information from all sequences were critical for promoting data recovery using corresponding statistical algorithms.

Blawat *et al.* (109) developed an encoding method based on the RS codes for storing 22MB files in DNA. Every eight bits in the original data were transformed into five successive bases in the encoded DNA sequences. Such an encoding approach reduced the possibility of self-complementation in DNA sequences. Furthermore, robust RS codes and the additional parity-check process were collaborated to encode data in synthesized oligonucleotides. The raw readout from sequencing underwent a series of decoding steps including majority voting for the most possible sequences, incomplete oligonucleotide reconstruction, and RS decoding. The possibility of residual errors in the decoded data was relatively low, which was comparable with modern hard disk drivers (HDDs), or even less if more data redundancy was included in the encoding step.

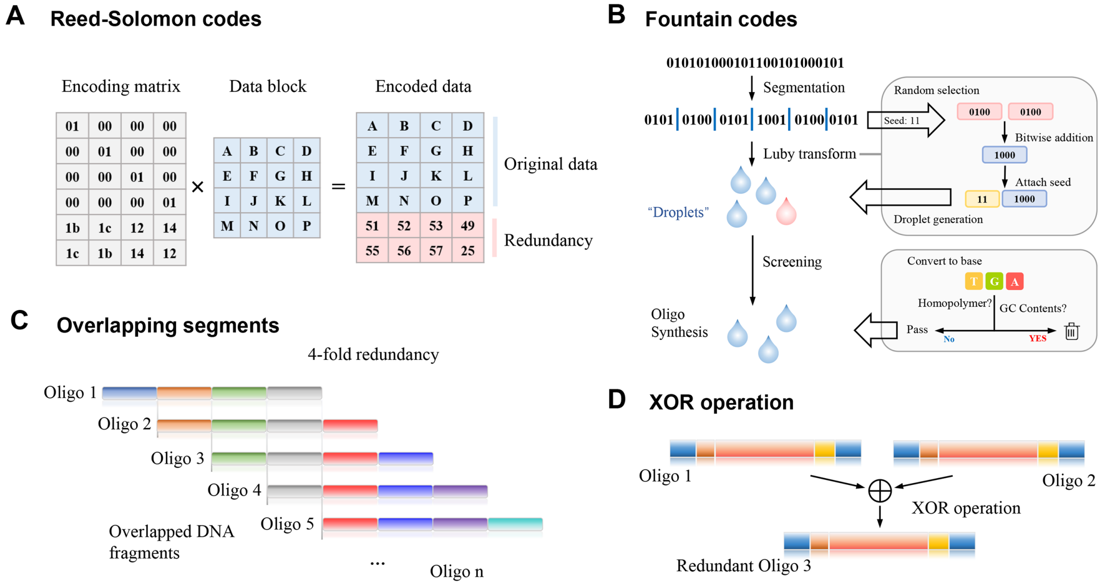

Figure 3. Error correction methods for the DNA database. (**A**) Reed-Solomon codes multiplied the data block with a pre-calculated encoding matrix, and the redundant information would be added at the end of the result (108). (**B**) Fountain codes separated data into many segments, selected them using a special distribution function and packaged them into many 'droplets'. Unqualified sequences were excluded from the screening procedure. Droplets of good quality were used for oligonucleotide synthesis (16). (**C**) Overlapping was a simple but effective method to avoid errors. Repeated pieces of the sequence under a shifting pattern were included in different oligonucleotide strands to obtain available copies (For example, four-fold redundancy generated in the figure) (13). (**D**) Exclusive-or operations between any two information strands generated the third strand. Any two of them could restore the remaining strand (26).

In addition to the RS codes, other encoding methods used in information technology have also been applied in the DNA-based database. Based on fountain codes, Erlich *et al.* (16) reported the theoretically highest physical density for the data storage. To eliminate unqualified droplets for synthesis and sequencing, an additional screening step was designed to rule out those sequences which consisted of homopolymers or high GC base content. Considering the restriction of synthesis and sequencing, unqualified droplets with high GC contents and homopolymers were abandoned in the screening step. Finally, a file with the size of 2.14MB was converted to 72 000 strands of oligonucleotides for storage. Then the data was recovered with no mistake, even 2000 of the 72 000 DNA strands were lost after the whole storage process. Furthermore, with PCR amplification and dilution, researchers managed to retrieve the data from only 10 pg of DNA. Therefore, the physical density was astonishingly as large as 215PB/g.

Depending on the type of errors during DNA synthesis, a specific coding method can be more appropriate. RS codes are more efficient in rectifying substitution errors, while the fountain codes are preferred to deal with deletion or addition errors (117,118). RS codes can effectively detect and correct substitution errors when the length of DNA is unaltered (no addition or deletion errors). On the other hand, when part of the DNA pool is lost or other sequences are added, fountain codes may be the better option. Erlich *et al.* reported that the combination of these two codes could provide better results (16).

Other than fountain codes, other concepts were also applied in recent studies for error correction. Goldman *et al.* (13) used Huffman codes to transform each byte in original files into several ternary digits to eliminate homopolymers. The resulting information was divided into overlapping segments under a shifting pattern, as demonstrated in Figure 3C. For a certain segment, in addition to itself, encoded information could be identified from the other three 'backup' sequences, resulting in a fourfold redundancy. Segments were designed as reverse complement alternatively, which reduced the probability of data loss and failed recovery. Bornholt *et al.* (26) also reported a new way to add data redundancy. To be specific, two original strands produced a new strand via an XOR operation so that any two out of the three DNA strands were sufficient to recover the original data (Figure 3D). This method reduced the overhead, and the data redundancy could be flexibly tuned by the number of replicated XOR operation. Multiple XOR operations between strands provided higher redundancy and *vice versa*.

Adding error correction information and data redundancy in the encoding step is indispensable for the accurate recovery of data in DNA. Various encoding methods such as RS codes, fountain codes, Huffman codes, and XOR operation are introduced in this section. The choice of the most appropriate coding and their combination for a specific project highly depends on the application scenario and the type of data. Furthermore, there is always a trade-off between data redundancy for error correction and coding density (6), so the reduction of error rate from the origin (e.g. DNA synthesis and sequencing) is much preferred from this respect.

DNA IMMOBILIZATION AND PRESERVATION

DNA is a naturally selected molecule storing our essential gene information which can inherit from generation to generation. It serves as the blueprint to construct all kinds of biological creatures, and maintain homeostasis despite the environmental changes. Although DNA can be degraded when exposed to air, high humidity, or radiation, the half-life of DNA can readily exceed thousands of years when preserved under suitable conditions (119–121). For example, the complete genome of a Neanderthal from about 50 000 years ago was successfully obtained and sequenced (122). The sequences of the 16S ribosomal DNA identified as a spore-forming bacterium belonging to *Bacillus* species were acquired from a 250 million-year-old salt crystal from the Permian Salado Formation (123). Taking advantage of the long retention time of DNA, long-term archival storage of molecular digital data can be accomplished. To achieve the long-term stability of DNA, several critical factors need to be considered, such as the DNA immobilization and the preservation conditions, which we will discuss in this section.

DNA immobilization

A vital step in DNA-based data storage is the attachment of DNA molecules to a substrate. The stability of the DNA immobilization and the DNA density on the substrate have a dramatic influence on the robustness of the storage system and the data density. DNA is a polymer built up with four nucleotides covalently bonded by phosphate linkage. Due to the phosphate backbone, DNA is negatively charged under physiological conditions (pH 7.4). Physical and chemical methods are often used for DNA immobilization on substrates (124).

The physical immobilization is based on relatively weak interactions between DNA and substrates, such as the van der Waals force, hydrophobic interaction, and electrostatic adsorption. Among them, electrostatic adsorption based on the negative charges on the DNA backbone is widely used for DNA immobilization. The layer-by-layer structure is formed by alternatively stacking two oppositely charged polymers (Figure 4A) (116,125,126). For example, positively charged polymers such as poly ethylenimine and poly (dimethyl diallyl ammonium chloride) have been used to form condensed layer-by-layer structure with the negatively charged DNA on a substrate. Chen *et al.* (116) reported a layer-by-layer assembly method to preserve the DNA database on magnetic nanoparticles with large surface areas. An additional encapsulation layer of silica was deposited to protect the DNA molecules in the accelerated aging experiments. By this method, 83 KB data was encoded in DNA with a storage density of 160 ng/cm^2 on the nanoparticles. Alternatively, Saurer *et al.* (127) assembled plasmid DNA with hydrolytically degradable cationic poly (β-amino ester) on microneedles via electrostatic adsorption. The microneedle could preserve bioactive DNA molecules, and controllably release the DNA molecules, which could be used for DNA data storage in the future.

For the chemical immobilization of DNA, the modified DNA reacts with the functional groups on the substrate surface. The most widely used reactions include Au-S interaction (128–131) and amine-based reactions (e.g. EDC (1-(3-dimethyl aminopropyl)-3-ethyl carbodiimide hydrochloride)-NHS (*N*-hydroxy succinimide) reaction and Schiff base reaction) (132–134). The immobilization of DNA on Au surfaces by forming Au-S bonds has been developed. For example, thiol-modified DNA covalently binds to Au after incubation at room temperature (Figure 4B). Cao *et al.* (128) reported the attachment of thiol-modified DNA to the surface of core/shell Ag/Au nanoparticles for colorimetric detecting. The DNA hybridization-induced nanoparticle aggregation led to visible colour change. Kim *et al.* (130) developed a programmable system based on DNA-linked colloidal gold nanoparticles building blocks termed nBLOCKs, in which gold nanoparticles were used as the nodes and thiol-modified DNA served as the scaffolds. By controlling the number, placement, and relative orientation of the DNA scaffolds on the surface of the gold nanoparticles, they could adjust the structures of the final self-assembled complexes, from simple linear dimers to complex 3D pyramidal and octahedral shapes. Although thiol-modified DNA may be detached from the Au substrate under high temperature or react with other competitive reagents like DTT (135,136), the biocompatible reaction conditions and relatively stable bonds make it appropriate for DNA-based data storage.

DNA immobilization using amino-based reactions is another widely used approach. The amino group is also highly reactive, which can be activated by EDC/NHS reaction and then react with a carboxylic group to form an amide bond (137–139). Another way to immobilize the amino-modified DNA is to react with an aldehyde group to form a Schiff base (Figure 4C). The resulting imine bond is unstable which can be further reduced by sodium borohydride to generate a stable product. For instance, Fuentes *et al.* (134) attached amino-modified DNA to the surface of superparamagnetic nanoparticles with a heterofunctional polymer (aldehyde-aspartic-dextran) to detect Hepatitis C Virus cDNA. Recently, Nguyen *et al.* (133) reported that the immobilization of DNA in microfluidic channels via the amino-based modification and the complementary fluorophore-labeled strands with controllable ON/OFF states helped to develop a new type of DNA storage system.

Recently, DNA immobilization by acrylate-based polymerization for data storage has been reported by Choi *et al.* (140). The acrylate moiety on the DNA could react and covalently link to poly (ethylene glycol) diacrylate (PEGDA) hydrogel. Auxiliary QR codes were printed on the DNA hydrogel based on photolithography. Details and related primer information for data retrieval could be obtained by scanning the QR codes. This covalent linking was also stable, which ensured reproducible reading with a long retention time (half-life >100 years under 10°C according to the decay kinetics estimation).

Except for the covalent immobilization above, another way that relies on specific recognition between molecules has also been used. For example, streptavidin and biotin exhibit strong specific interactions. Their binding strength is close to that of covalent bonding, as proven by its large affinity constant of $\sim 10^{15}$ mol/L (141,142). Therefore, the attachment of biotin-labeled DNA to the streptavidin-modified surface is an effective way for DNA immobilization (Figure 4D) (143). Song *et al.* (144) reported a hybridization system regulated by the electric field for DNA data storage, where the streptavidin-modified hydrogel ma-

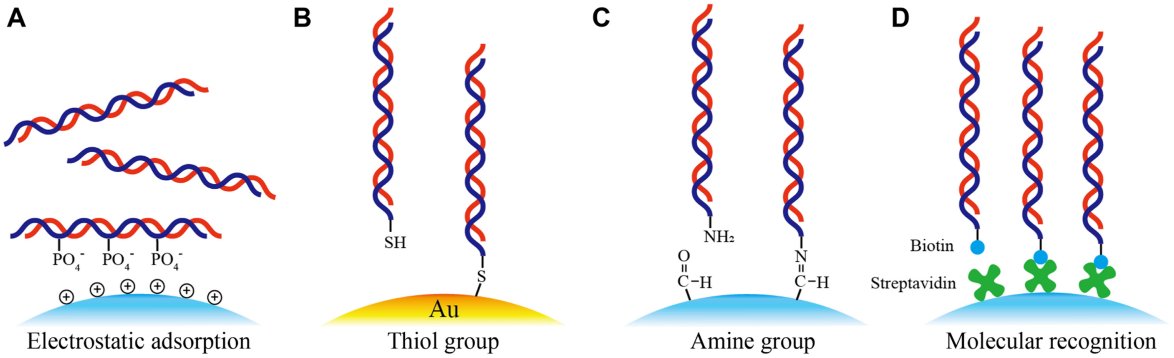

Figure 4. Schematics of various DNA immobilization methods. (**A**) DNA immobilized on the substrate using the electrostatic adsorption between polycations on the surface and negative phosphate groups of DNA molecules. (**B**) Immobilization of DNA with a thiol group on gold nanoparticles via Au–S bonding. (**C**) Immobilization of aminated DNA strands on the surface of the aldehyde group-modified substrate. The reaction between the amine group and the aldehyde group formed a Schiff base. (**D**) Immobilization of biotinylated DNA strands on the streptavidin-modified substrate.

trix was attached to an electrode array. When a positive potential was applied to the electrodes, the biotinylated polyanionic DNA was allowed to combine with the hydrogel. Three hybridization regions were designed on the fixed DNA molecules so that three kinds of complementary DNA probes with different fluorescence labels could selectively hybridize with the DNA molecule on the hydrogel. The hybridization of three DNA probes on one fixed DNA strand led to a total of eight (2^3) combinations or one byte for data storage. Moreover, editing, rewriting, and bit shifting were realized via strand displacement reaction, and random access was also achieved on the individually addressable electrodes.

For two immobilization methods, the chemical method based on covalent bonding provides stable and localized DNA immobilization, which is compatible with array-based DNA processes. For frequent data reading and writing, the chemically immobilized DNA exhibits far better reliability against solvation, diffusion, and electric migration than physical adsorption. However, other than the long-term stability, the data storage density is also an important aspect. For the chemical method, only a monolayer of DNA can be deposited on the surface of the substrate, so the density is limited by the steric hindrance and surface area. In contrast, the physical method (i.e. layer-by-layer assembly) can accommodate large amounts of DNA on the surface, despite that the relatively weak physical interaction may lead to poor stability and robustness. To conclude, from the perspective of storage density, the layer-by-layer assembly method is usually preferred, but possible data loss and damage to DNA should be further studied.

Critical conditions for DNA preservation

Paleontological genome DNA stored in fossils, ambers, and salt crystals can retain their biological functions, which implies their incredible lifespan. Under favourable preservation conditions, DNA molecules can be stable for thousands of years, making it an ideal medium for non-volatile memory. However, as a biomacromolecule, DNA is also susceptible to degradation. For example, DNA can be irreversibly damaged when exposed to acids, light, high humidity, high temperature, or reactive oxygen species (ROS). So, the protection of a DNA-based database requires the elimination of these factors from the storage conditions.

DNA has good solubility in water, and most of the DNA-related biological processes are accomplished in an aqueous solution. Therefore, directly storing DNA in water is straightforward for subsequent reading operations. Unfortunately, the dissolved DNA in an aqueous solution is prone to depurination, deamination, depyrimidination, and hydrolytic cleavage of the phosphate backbone (145–149). Under alkaline conditions, the DNA degradation catalysed by acids is prevented (150), and the depurination of DNA can also be suppressed in solutions of high ionic strength (151). Under favourable alkaline conditions with high ionic strength, DNA stability is mainly affected by oxidation. Apart from dissolved oxygen, trace amounts of metals (e.g. Fe^{3+}, Cu^{2+}, etc.) notably enhance oxidative damage to DNA by producing hydroxyl radicals through the Fenton reaction (152–154). Unfortunately, the presence of transition metal ions in a DNA sample is almost inevitable. DNA purified at the highest standard still contains ~30 ppb of Fe, and that common stabilizers and buffers lead to additional metal contamination (154). Lanthanide ions like Ce^{IV} ions have been proven to activate the phosphodiester linkages in DNA, which is susceptible to nucleophilic attack, leading to DNA hydrolysis (155). The addition of metal chelators such as ethylenediaminetetraacetic acid (EDTA) may mitigate the metal pollution, but it cannot avoid oxidation damage generated by the Fenton reaction (156,157). To prevent DNA from oxidation, antioxidants and radical scavengers can be added to the storing medium.

Another environmental factor affecting DNA preservation is temperature. Under high temperature, the probability of molecular collision increases so that a greater amount of molecules acquires the activation energy for unwanted reactions, as predicted by Arrhenius's Law (158). In a pyrolysis experiment, complete degradation of DNA was observed at 180°C in a few minutes (159). Although cryopreservation of DNA decreases the DNA degradation reactivity, cryolysis can also be induced by the formation of cracks within the ice at low temperatures (160). To address these problems, additives such as trehalose can improve the stability and integrity of DNA during prolonged preservation (161–

163). Strong hydrogen bonds between trehalose molecules and DNA strands offset the electrostatic repulsion between DNA backbones (161). The effect of limited molecular mobility of DNA can be enhanced by the vitreous matrix provided by trehalose (164,165). Other additives such as resveratrol and DNAStable have also been examined as effective preservatives (166,167).

On the other hand, despite low temperature is beneficial for long-term preservation, it increases the cost remarkably. So, natural permafrost is concerned as a favourable choice for low-frequency archival storage to reduce the cost. However, low temperature attributes less to the robustness of DNA storage compared with keeping at the dry state (168,169). Although DNA can maintain its integrity in an aqueous solution for tens of thousands of years at 10°C, it is still three magnitudes lower than that stored in the dry state (8). The dehydrated DNA is more chemically inert owing to decreased molecular mobility and absence of water or dissolved oxygen that usually participated in the DNA damage. Consequently, for DNA-based data storage, keeping the DNA samples dry is superior to keeping them at a low temperature.

Ionising radiation, including β, γ and X rays, can also induce the damage of DNA molecules, such as chain break and chemical changes of bases. The damage can be attributed to a combination of direct and indirect reactions. For the direct reactions, ionising radiations indiscriminately transfer energy to electrons in the irradiated molecules, which are excited to react with other reagents (170). For example, a C-H bond in DNA can be activated to generate a highly reactive carbon-centred radical, leading to oxidization, reduction, and even strand breakage. Sanche *et al*. reported that transient radical anions in DNA molecules formed under the radical attacks (171), which might further induce bond breakage.

For the indirect reactions, ionising radiation activate water molecules to produce aggressive ·OH radicals (172), which can react with sugar and base moieties of DNA and cause subsequent side reactions such as ring-opening, pyrimidine oxidation, and bond breakage (173,174). Due to DNA repair mechanisms (175) and radical scavengers (176), the problem of radiation damage can be alleviated *in vivo*. As previously estimated, about one single-strand break occurred per 10^7 bp/Gy of radiation (177). However, the molecular damage is much more significant for *in vitro* DNA systems in absence of repair pathways. Besides, for certain conditions, such as in an space station, the intensity of cosmic rays is much higher than that at sea level (178), which lead to much more DNA damages. So, the effect of ionising radiation on a DNA database can also be a significant concern.

Considering all conditions mentioned above, a currently feasible solution for DNA database preservation is to encapsulate the DNA in a dehydrated state (Figure 5A) (116,179,180). Paunescu *et al.* (180) reported the preservation of DNA by adsorption on the surface of submicron-sized silica particles modified with positive-charged ammonium groups and subsequently depositing a silica layer on it. The encapsulation of the silica layer prevented the DNA from damages under high temperature, ROS, and UV irradiation. The dense silica layer acted as a hermetic diffusion barrier to external oxygen, ROS, and metal ions. Although the outmost amorphous silica was a UV-permeable material (>60% transparent at 170 nm wavelength), the geometric shape was able to scatter light of low wavelengths so that the DNA inside could maintain its integrity even under the UV radiation. Recently, Koch *et al.* reported the dispersion of as-synthesized microparticles in polycaprolactone (PCL) for three-dimensional (3D) printing (179). The files encoded in the DNA molecules were required for 3D modelling and printing. They iterated as high as six times of these procedures including extracting DNA from printed objects via PCR amplification, decoding DNA to obtain the data file, and 3D printing objects using PCL fibres mixed with microparticles according to the files. The hard silica shell was capable of protecting the DNA from the violent stress in filament extrusion (Figure 5B). To conclude, the encapsulation of DNA is a currently feasible and effective method for long-term preservation, which endows the DNA storage system with relatively high stability and considerable storage density.

CHALLENGES BEFORE LARGE-SCALE APPLICATIONS

Although DNA-based data storage has theoretically higher data density, longer retention time, and lower power consumption than current electronic memory devices, the maximum storage capacity until now is less than 1GB, much lower than the state-of-the-art data storage devices. There remain several challenges to achieve the large-scale application of DNA-based data storage. Further improvements rely on the progress in all of the steps involved in the data storage, including data encoding, DNA synthesis, preservation, data retrieval, and DNA sequencing.

To ensure the correct retrieval of data from the DNA sequences, error-correction based on data redundancy is indispensable. Generally, more data redundancy leads to better correction of errors, either from DNA synthesis or sequencing. However, more data redundancy also results in decreased encoding efficiency, in other words, less information can be encoded in the DNA sequences. Thus, there is a trade-off between error-correction ability and logical data density. In addition, the performance of ECC differs when applied in different encoding frameworks shown in Figure 2. The choices of encoding framework and the ECC should be systematically considered for optimal error tolerance and coding efficiency.

The array-based DNA synthesis enables the paralleled synthesis of a large group of DNA molecules with different sequences, and the reagent consumption is less than the column-based synthesis method. However, it is still challenging to push the synthesis scale to the level that can be competitive against the scale of conventional electronic memory, in terms of encoded data. A larger synthesis scale relies on the fabrication of smaller features (e.g. electrodes) on the chip, leading to problems that affect the fidelity of DNA syntheses, such as the cross-talk between adjacent microelectrodes in the array and low reproducibility. The advance in this aspect probably relies on the advent of new enzymatic synthesis, which may considerably enhance the length, quality, and speed of the DNA synthesis (e.g. with

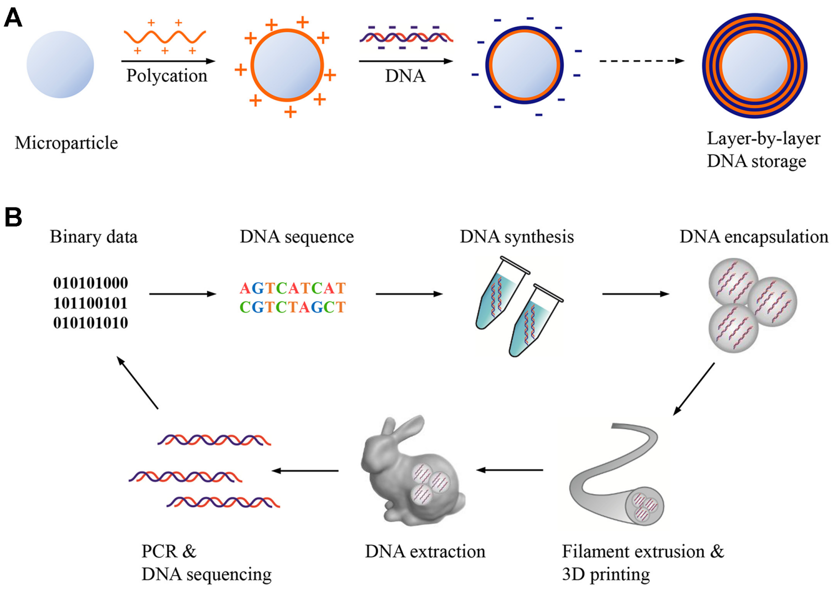

Figure 5. Schematics for long-term DNA preservation. (**A**) Preservation for layer-by-layer assembled DNA, in which cationic polymer and anionic DNA were organized into a layer-by-layer structure based on electrostatic interaction on the surface of microparticles (116). (**B**) DNA preservation by packaging in silica beads and then dispersed in polycaprolactone fibres. The DNA strands were encoded with data files that were used for 3D printing of a rabbit model. DNA molecules could be further extracted, amplified, and sequenced to acquire encoded files (179). Reproduced with permission (179). Copyright 2020, Springer Nature.

the use of template-independent TdT). The greater synthesis length, higher accuracy, and speed will also reduce the synthesis cost, which is a crucial factor that restricts the applicability of DNA-based data storage. Akin to the transition from column-based to array-based chemical synthesis, the development of paralleled high-throughput enzymatic DNA synthesis will be probably the focus of research in the future.

Preservation of DNA molecules away from adverse factors (e.g. by encapsulation in silica beads) can ensure long-term stability for data retention. Proper environmental conditions and additives should be investigated to pursue a long retention time to the theoretical limit. Additionally, the preservation by encapsulation avoids direct access to the data files. For example, DNA encapsulated in silica cannot be directly amplified and retrieved by PCR unless released from the beads. Therefore, an optimized preservation method should balance the long-term stability and the accessibility of data.

In the past few decades, high-throughput sequencing technology has achieved remarkable progress, which is characterized by a significant reduction in sequencing costs and time. Nevertheless, the data reading speed is still several orders of magnitude lower than that on conventional data media (e.g. flash memory). Therefore, the application of DNA storage at present is usually limited to archival storage and is impractical for low-latency data memory units used in mobile phones, computers, or other consumer electronic devices. So far, third-generation sequencing based on nanopore technology has been a promising technique to address this problem. But many problems hinder its utilization in DNA data storage. The nanopore sequencing relies on the detection of weak electric signals, which can be easily interfered with by background noises caused by vibrations and electric circuits. Furthermore, the translocation speed of DNA is difficult to control. High speed usually results in poor resolution of the electronic signals, which makes it hard to distinguish both the type and length of nucleotides. So, a technical breakthrough is still for fast and real-time data reading in large-scale application of DNA-based storage systems.

Another issue is the cost to store data in DNA. Despite the theoretical advantages compared with conventional digital memory, DNA-based data storage is still much more expensive, and the majority of the cost comes from DNA synthesis. Currently, synthesizing millions of different DNA sequences still requires cutting-edge technology. The cost for synthesis and sequencing has been estimated to be at least ∼\$3500/MB, as previously estimated by Erlich and Zielinski (16), which renders the large-scale DNA storage currently impractical. The technical development of DNA synthesis and sequencing is urgently demanded to reduce the expense of DNA data storage to an acceptable level.

Despite those challenges, we are still optimistic about the future of DNA-based data storage. It is always instructive to learn the wisdom of natural creatures, such as cells that have proofreading mechanisms for error correction of the DNA synthesized (181,182). This leads to *in vivo* storage of digital information or the use of proofreading enzymes for *in vitro* databases. In addition, genetic information can be exchanged between cells by plasmids, which are applied by bacteria to proliferate and overcome adversity (183–185). The plasmid can act as a carrier for information storage in living organisms. Emerging technologies, including the enzymatic synthesis of DNA, the development of computer and information technology, will lead to new robust encoding structures and error correction schemes. Other optimizations of DNA editing, encrypting and maintaining, immobilization, and storing approaches (e.g. clustered regularly interspaced short palindromic repeats (103,186,187), rewriting based on strand replacement reaction (66), and random number generation for encryption (188)) should help to reduce the uncertainties in large-scale DNA-based data storage and eventually makes the DNA a promising data storing medium for human beings.

CONCLUSION

With the rapid development of information technology, the worldwide explosion of data generation has increasingly challenged the production of conventional electronic memory based on silicon. Carbon-based DNA molecular data storage discussed in this review is probably one of the most promising technologies to address this issue considering its high data retention time, storage density, and relatively low energy consumption in data copying, pasting, editing, and reading. In this review, we introduced the general processes of data storage in DNA. Then we focused on the uncertainties involved in those steps and potential methods for error correction. Finally, we discussed the remaining challenges for the realization of large-scale DNA-based data storage systems and further research direction in this area. As a novel molecular media for data storage, DNA is still in its infancy. Indeed, substantial breakthroughs are still required to achieve the large-scale application. However, we believe it is a promising technology to solve the problem of data explosion, especially considering the rapid developments in DNA synthesis, sequencing, and other related technologies. As Grass *et al.* proposed (179), everything may act as a carrier of a DNA database in the future, which may have a profound influence on a range of fields, including data management, the internet of things (IoT), and blockchain.

FUNDING

National Natural Science Foundation of China [21635001 to H.L.]; State Key Project of Research and Development [2016YFF0100802 to H.L.]; Key Project and Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University. Funding for open access charge: National Natural Science Foundation of China.
Conflict of interest statement. None declared.

REFERENCES

1. Goda,K. and Kitsuregawa,M. (2012) The history of storage systems. *Proc. IEEE*, **100**, 1433–1440.
2. Hilbert,M. and Lopez,P. (2011) The world's technological capacity to store, communicate, and compute information. *Science*, **332**, 60–65.
3. Reisel,D., Gantz,J. and Rydning,J. (2018) Data age 2025: the digitization of the world from edge to core. *Seagate*, https://www.seagate.com/in/en/our-story/data-age-2025/.
4. Xu,Z.-W. (2014) Cloud-sea computing systems: Towards thousand-fold improvement in performance per watt for the coming zettabyte era. *J. Comput. Sci. Technol.*, **29**, 177–181.
5. Extance,A. (2016) Could the molecule known for storing genetic information also store the world's data? *Nature*, **537**, 22–24.
6. Organick,L., Chen,Y.J., Dumas Ang,S., Lopez,R., Liu,X., Strauss,K. and Ceze,L. (2020) Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.*, **11**, 616.
7. Gregory,T.R., Nicol,J.A., Tamm,H., Kullman,B., Kullman,K., Leitch,I.J., Murray,B.G., Kapraun,D.F., Greilhuber,J. and Bennett,M.D. (2007) Eukaryotic genome size databases. *Nucleic Acids Res.*, **35**, D332–D338.
8. Zhirnov,V., Zadegan,R.M., Sandhu,G.S., Church,G.M. and Hughes,W.L. (2016) Nucleic acid memory. *Nat. Mater.*, **15**, 366–370.
9. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
10. Branton,D., Deamer,D.W., Marziali,A., Bayley,H., Benner,S.A., Butler,T., Di Ventra,M., Garaj,S., Hibbs,A., Huang,X. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat. Biotechnol.*, **26**, 1146–1153.
11. Garibyan,L. and Avashia,N. (2013) Polymerase chain reaction. *J. Invest. Dermatol.*, **133**, 1–4.
12. Ochman,H., Gerber,A.S. and Hartl,D.L. (1988) Genetic applications of an inverse polymerase chain reaction. *Genetics*, **120**, 621–623.
13. Goldman,N., Bertone,P., Chen,S., Dessimoz,C., LeProust,E.M., Sipos,B. and Birney,E. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, **494**, 77–80.
14. Organick,L., Ang,S.D., Chen,Y.J., Lopez,R., Yekhanin,S., Makarychev,K., Racz,M.Z., Kamath,G., Gopalan,P., Nguyen,B. *et al.* (2018) Random access in large-scale DNA data storage. *Nat. Biotechnol.*, **36**, 242–248.
15. Church,G.M., Gao,Y. and Kosuri,S. (2012) Next-generation digital information storage in DNA. *Science*, **337**, 1628–1628.
16. Erlich,Y. and Zielinski,D. (2017) DNA Fountain enables a robust and efficient storage architecture. *Science*, **355**, 950–953.
17. Caruthers,M.H. (2013) The chemical synthesis of DNA/RNA: our gift to science. *J. Biol. Chem.*, **288**, 1420–1427.
18. Beaucage,S.L. and Caruthers,M.H. (1981) Deoxynucleoside phosphoramidites - a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.*, **22**, 1859–1862.
19. Mathews,A.S., Yang,H. and Montemagno,C. (2016) Photo-cleavable nucleotides for primer free enzyme mediated DNA synthesis. *Org. Biomol. Chem.*, **14**, 8278–8288.
20. Motea,E.A. and Berdis,A.J. (2010) Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochim. Biophys. Acta*, **1804**, 1151–1166.
21. Cano,R.J. and Borucki,M.K. (1995) Revival and identification of bacterial spores in 25-to 40-million-year-old Dominican amber. *Science*, **268**, 1060–1064.
22. Fish,S.A., Shepherd,T.J., McGenity,T.J. and Grant,W.D. (2002) Recovery of 16S ribosomal RNA gene fragments from ancient halite. *Nature*, **417**, 432–436.
23. Chen,Y.J., Takahashi,C.N., Organick,L., Bee,C., Ang,S.D., Weiss,P., Peck,B., Seelig,G., Ceze,L. and Strauss,K. (2020) Quantifying molecular bias in DNA data storage. *Nat. Commun.*, **11**, 3264.
24. Shendure,J., Balasubramanian,S., Church,G.M., Gilbert,W., Rogers,J., Schloss,J.A. and Waterston,R.H. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345–353.
25. Jain,M., Olsen,H.E., Paten,B. and Akeson,M. (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, **17**, 239.
26. Bornholt,J., Lopez,R., Carmean,D.M., Ceze,L., Seelig,G. and 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 - ASPLOS '16*. ACM, NY, pp. 637–649.
27. Caruthers,M.H. (2011) A brief review of DNA and RNA chemical synthesis. *Biochem. Soc. Trans.*, **39**, 575–580.
28. Hughes,R.A. and Ellington,A.D. (2017) Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harbor Perspect. Biol.*, **9**, a023812.
29. Damha,M.J., Giannaris,P.A. and Zabarylo,S.V. (1990) An improved procedure for derivatization of controlled-pore glass-beads for solid-phase oligonucleotide synthesis. *Nucleic Acids Res.*, **18**, 3813–3821.
30. Matteucci,M.D. and Caruthers,M.H. (1992) Synthesis of deoxyoligonucleotides on a polymer support. *J. Am. Chem. Soc.*, **24**, 92–98.

31. Moorcroft,M.J., Meuleman,W.R., Latham,S.G., Nicholls,T.J., Egeland,R.D. and Southern,E.M. (2005) In situ oligonucleotide synthesis on poly(dimethylsiloxane): a flexible substrate for microarray fabrication. *Nucleic Acids Res.*, **33**, e75.
32. Lee,C.C., Snyder,T.M. and Quake,S.R. (2010) A microfluidic oligonucleotide synthesizer. *Nucleic Acids Res.*, **38**, 2514–2521.
33. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
34. Butler,J.H., Cronin,M., Anderson,K.M., Biddison,G.M., Chatelain,F., Cummer,M., Davi,D.J., Fisher,L., Frauendorf,A.W., Frueh,F.W. *et al.* (2001) In situ synthesis of oligonucleotide arrays by using surface tension. *J. Am. Chem. Soc.*, **123**, 8887–8894.
35. Gao,X.L., LeProust,E., Zhang,H., Srivannavit,O., Gulari,E., Yu,P.L., Nishiguchi,C., Xiang,Q. and Zhou,X.C. (2001) A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res.*, **29**, 4744–4750.
36. Singh-Gasson,S., Green,R.D., Yue,Y.J., Nelson,C., Blattner,F., Sussman,M.R. and Cerrina,F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.*, **17**, 974–978.
37. Fodor,S.P.A., Read,J.L., Pirrung,M.C., Stryer,L., Lu,A.T. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science*, **251**, 767–773.
38. Egeland,R.D. and Southern,E.M. (2005) Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. *Nucleic Acids Res.*, **33**, e125.
39. Egeland,R.D., Marken,F. and Southern,E.M. (2002) An electrochemical redox couple activitated by microelectrodes for confined chemical patterning of surfaces. *Anal. Chem.*, **74**, 1590–1596.
40. Maurer,K., Cooper,J., Caraballo,M., Crye,J., Suciu,D., Ghindilis,A., Leonetti,J.A., Wang,W., Rossi,F.M., Stover,A.G. *et al.* (2006) Electrochemically generated acid and its containment to 100 micron reaction areas for the production of DNA microarrays. *PLoS One*, **1**, e34.
41. Chow,B.Y., Emig,C.J. and Jacobson,J.M. (2009) Photoelectrochemical synthesis of DNA microarrays. *Proc. Natl. Acad. Sci. USA*, **106**, 15219–15224.
42. Dalma-Weiszhausz,D.D., Warrington,J., Tanimoto,E.Y. and Miyada,C.G. (2006) The Affymetrix Genechip® platform: an overview. *Methods Enzymol.*, **410**, 3–28.
43. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
44. Liu,R., Munro,S., Nguyen,T., Siuda,T., Suciu,D., Bizak,M., Slota,M., Fuji,H., Danley,D. and McShea,A. (2006) Integrated microfluidic customarray device for bacterial genotyping and identification. *J. Am. Chem. Soc.*, **11**, 360–367.
45. Roth,K.M., Peyvan,K., Schwarzkopf,K.R. and Ghindilis,A. (2006) Electrochemical detection of short DNA oligomer hybridization using the combimatrix electrasense microarray reader. *Electroanalysis*, **18**, 1982–1988.
46. Antkowiak,P.L., Lietard,J., Darestani,M.Z., Somoza,M.M., Stark,W.J., Heckel,R. and Grass,R.N. (2020) Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nat. Commun.*, **11**, 5345.
47. Heckel,R., Mikutis,G. and Grass,R.N. (2019) A characterization of the DNA data storage channel. *Sci. Rep.*, **9**, 9663.
48. Efcavitch,J.W. and Heiner,C. (2006) Depurination as a yield decreasing mechanism in oligodeoxynucleotide synthesis. *Nucleosides. Nucleotides.*, **4**, 267–267.
49. Septak,M. (1996) Kinetic studies on depurination and detritylation of CPG-bound intermediates during oligonucleotide synthesis. *Nucleic Acids Res.*, **24**, 3053–3058.
50. LeProust,E.M., Peck,B.J., Spirin,K., McCuen,H.B., Moore,B., Namsaraev,E. and Caruthers,M.H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.*, **38**, 2522–2540.
51. Andrus,A., Efcavitch,J.W., McBride,L.J. and Giusti,B. (1988) Novel activating and capping reagents for improved hydrogen-phosphonate DNA-synthesis. *Tetrahedron Lett.*, **29**, 861–864.
52. Saaem,I., Ma,S., Quan,J. and Tian,J. (2012) Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Res.*, **40**, e23.
53. Anavy,L., Vaknin,I., Atar,O., Amit,R. and Yakhini,Z. (2019) Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat. Biotechnol.*, **37**, 1229–1236.
54. Choi,Y., Ryu,T., Lee,A.C., Choi,H., Lee,H., Park,J., Song,S.H., Kim,S., Kim,H., Park,W. *et al.* (2019) High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Sci. Rep.*, **9**, 6582.
55. Covo,S., Blanco,L. and Livneh,Z. (2004) Lesion bypass by human DNA polymerase μ reveals a template-dependent, sequence-independent nucleotidyl transferase activity. *J. Biol. Chem.*, **279**, 859–865.
56. Ramadan,K., Shevelev,I. and Hubscher,U. (2004) The DNA-polymerase-X family: controllers of DNA quality? *Nat. Rev. Mol. Cell Biol.*, **5**, 1038–1043.
57. Service,R.F. (2018) DNA printers poised to jump from paragraphs to pages. *Science*, **362**, 143.
58. Jensen,M.A. and Davis,R.W. (2018) Template-independent enzymatic oligonucleotide synthesis (TiEOS): its history, prospects, and challenges. *Biochemistry*, **57**, 1821–1832.
59. Metzker,M.L., Raghavachari,R., Richards,S., Jacutin,S.E., Civitello,A., Burgess,K. and Gibbs,R.A. (1994) Termination of DNA synthesis by novel 3′-modifieddeoxyribonucleoside 5′-triphosphates. *Nucleic Acids Res.*, **22**, 4259–4267.
60. 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. *et al.* (2018) De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.*, **36**, 645–650.
61. Lee,H.H., Kalhor,R., Goela,N., Bolot,J. and Church,G.M. (2019) Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.*, **10**, 2383.
62. Newman,S., Stephenson,A.P., Willsey,M., Nguyen,B.H., Takahashi,C.N., Strauss,K. and Ceze,L. (2019) High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat. Commun.*, **10**, 1706.
63. Tomek,K.J., Volkel,K., Simpson,A., Hass,A.G., Indermaur,E.W., Tuck,J.M. and Keung,A.J. (2019) Driving the scalability of DNA-based information storage systems. *ACS Synth. Biol.*, **8**, 1241–1248.
64. Yazdi,S.M., Yuan,Y., Ma,J., Zhao,H. and Milenkovic,O. (2015) A rewritable, random-access DNA-based storage system. *Sci. Rep.*, **5**, 14138.
65. Yamamoto,M., Kashiwamura,S., Ohuchi,A. and Furukawa,M. (2008) Large-scale DNA memory based on the nested PCR. *Nat. Comput.*, **7**, 335–346.
66. Lin,K.N., Volkel,K., Tuck,J.M. and Keung,A.J. (2020) Dynamic and scalable DNA-based information storage. *Nat. Commun.*, **11**, 2981.
67. Song,X., Shah,S. and Reif,J. (2019) Multidimensional data organization and random access in large-scale DNA storage systems. bioRxiv doi: http://dx.doi.org/10.1101/743369, 22 August 2019, preprint: not peer reviewed.
68. Ruijter,J.M., Ramakers,C., Hoogaars,W.M., Karlen,Y., Bakker,O., van den Hoff,M.J. and Moorman,A.F. (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.*, **37**, e45.
69. Warnecke,P.M., Stirzaker,C., Melki,J.R., Millar,D.S., Paul,C.L. and Clark,S.J. (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res.*, **25**, 4422–4426.
70. Pan,W., Byrne-Steele,M., Wang,C., Lu,S., Clemmons,S., Zahorchak,R.J. and Han,J. (2014) DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol.*, **14**, 10.
71. Henry,O.Y.F. and O'Sullivan,C.K. (2012) Rapid DNA hybridization in microfluidics. *TrAC, Trends Anal. Chem.*, **33**, 9–22.
72. Marimuthu,K. and Chakrabarti,R. (2014) Sequence-dependent theory of oligonucleotide hybridization kinetics. *J. Chem. Phys.*, **140**, 175104.
73. Guo,J., Xu,N., Li,Z., Zhang,S., Wu,J., Kim,D.H., Marma,M.S., Meng,Q., Cao,H., Li,X. *et al.* (2008) Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 9145–9150.

74. Ju,J., Kim,D.H., Bi,L., Meng,Q., Bai,X., Li,Z., Li,X., Marma,M.S., Shi,S., Wu,J. *et al.* (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 19635–19640.
75. Meller,A., Nivon,L., Brandin,E. and Golovchenko,J. Branton, D. (2000) Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 1079–1084.
76. Deamer,D., Akeson,M. and Branton,D. (2016) Three decades of nanopore sequencing. *Nat. Biotechnol.*, **34**, 518–524.
77. Lopez,R., Chen,Y.J., Dumas Ang,S., Yekhanin,S., Makarychev,K., Racz,M.Z., Seelig,G., Strauss,K. and Ceze,L. (2019) DNA assembly for nanopore data storage readout. *Nat. Commun.*, **10**, 2933.
78. Clarke,J., Wu,H.C., Jayasinghe,L., Patel,A., Reid,S. and Bayley,H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.
79. Ip,C.L.C., Loose,M., Tyson,J.R., de Cesare,M., Brown,B.L., Jain,M., Leggett,R.M, Eccles,D.A., Zalunin,V., Urban,J.M. *et al.* (2015) MinION analysis and reference consortium: phase 1 data release and analysis. *F1000Research*, **4**, 1075.
80. Chen,K., Kong,J., Zhu,J., Ermann,N., Predki,P. and Keyser,U.F. (2019) Digital data storage using DNA nanostructures and solid-state nanopores. *Nano Lett.*, **19**, 1210–1215.
81. Chen,K., Zhu,J., Boskovic,F. and Keyser,U.F. (2019) Secure data storage on DNA hard drives. bioRxiv doi: https://doi.org/10.1101/857748, 28 November 2019, preprint: not peer reviewed.
82. Takahashi,C.N., Nguyen,B.H., Strauss,K. and Ceze,L. (2019) Demonstration of end-to-end automation of DNA data storage. *Sci. Rep.*, **9**, 4998.
83. Poon,K. and Macgregor,R.B. Jr (1998) Unusual behavior exhibited by multistranded guanine-rich DNA complexes. *Biopolymers: Orig. Res. Biomol.*, **45**, 427–434.
84. Schwartz,J.J., Lee,C. and Shendure,J. (2012) Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods*, **9**, 913–915.
85. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
86. Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.
87. Schirmer,M., D'Amore,R., Ijaz,U.Z., Hall,N. and Quince,C. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinf.*, **17**, 125.
88. Ross,M.G., Russ,C., Costello,M., Hollinger,A., Lennon,N.J., Hegarty,R., Nusbaum,C. and Jaffe,D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, R51.
89. Yusko,E.C., Johnson,J.M., Majd,S., Prangkio,P., Rollings,R.C., Li,J., Yang,J. and Mayer,M. (2011) Controlling protein translocation through nanopores with bio-inspired fluid walls. *Nat. Nanotechnol.*, **6**, 253–260.
90. Fragasso,A., Schmid,S. and Dekker,C. (2020) Comparing current noise in biological and solid-state nanopores. *ACS Nano*, **14**, 1338–1349.
91. Wanunu,M. (2012) Nanopores: a journey towards DNA sequencing. *Phys. Life Rev.*, **9**, 125–158.
92. Dekker,C. (2007) Solid-state nanopores. *Nat. Biotechnol.*, **2**, 209–215.
93. Schneider,G.F., Xu,Q., Hage,S., Luik,S., Spoor,J.N., Malladi,S., Zandbergen,H. and Dekker,C. (2013) Tailoring the hydrophobicity of graphene for its use as nanopores for DNA translocation. *Nat. Commun.*, **4**, 2619.
94. Agah,S., Zheng,M., Pasquali,M. and Kolomeisky,A.B. (2016) DNA sequencing by nanopores: advances and challenges. *J. Phys. D: Appl. Phys.*, **49**, 413001.
95. Keyser,U.F. (2011) Controlling molecular transport through nanopores. *J. R. Soc., Interface*, **8**, 1369–1378.
96. Goodwin,S., Gurtowski,J., Ethe-Sayers,S., Deshpande,P., Schatz,M.C. and McCombie,W.R. (2015) Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.*, **25**, 1750–1756.
97. Timp,W., Comer,J. and Aksimentiev,A. (2012) DNA base-calling from a nanopore using a Viterbi algorithm. *Biophys. J.*, **102**, L37–L39.
98. Szalay,T. and Golovchenko,J.A. (2015) De novo sequencing and variant calling with nanopores using PoreSeq. *Nat. Biotechnol.*, **33**, 1087–1091.
99. Schreiber,J. and Karplus,K. (2015) Analysis of nanopore data using hidden Markov models. *Bioinformatics*, **31**, 1897–1903.
100. Simpson,J.T., Workman,R.E., Zuzarte,P.C., David,M., Dursi,L.J. and Timp,W. (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*, **14**, 407–410.
101. Luo,R., Wong,C.-L., Wong,Y.-S., Tang,C.-I., Liu,C.-M., Leung,C.-M. and Lam,T.-W. (2020) Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat. Mach. Intell.*, **2**, 220–227.
102. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.
103. Shipman,S.L., Nivala,J., Macklis,J.D. and Church,G.M. (2017) CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, **547**, 345–349.
104. Gilpatrick,T., Lee,I., Graham,J.E., Raimondeau,E., Bowen,R., Heron,A., Downs,B., Sukumar,S., Sedlazeck,F.J. and Timp,W. (2020) Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.*, **38**, 433–438.
105. Stranges,P.B., Palla,M., Kalachikov,S., Nivala,J., Dorwart,M., Trans,A., Kumar,S., Porel,M., Chien,M., Tao,C. *et al.* (2016) Design and characterization of a nanopore-coupled polymerase for single-molecule DNA sequencing by synthesis on an electrode array. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E6749–E6756.
106. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
107. Yazdi,S., Gabrys,R. and Milenkovic,O. (2017) Portable and error-free DNA-based data storage. *Sci. Rep.*, **7**, 5011.
108. Grass,R.N., Heckel,R., Puddu,M., Paunescu,D. and Stark,W.J. (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem., Int. Ed. Engl.*, **54**, 2552–2555.
109. Blawat,M., Gaedke,K., Hütter,I., Chen,X.-M., Turczyk,B., Inverso,S., Pruitt,B.W. and Church,G.M. (2016) Forward error correction for DNA data storage. *Procedia Comput. Sci.*, **80**, 1011–1022.
110. Heckel,R., Shomorony,I., Ramchandran,K. and Tse,D.N.C. (2017) Fundamental limits of DNA storage systems. In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, NY, pp. 3130–3134.
111. Ailenberg,M. and Rotstein,O. (2009) An improved Huffman coding method for archiving text, images, and music characters in DNA. *BioTechniques*, **47**, 747–754.
112. Guruswami,V. and Sudan,M. (1999) Improved decoding of Reed-Solomon and algebraic-geometry codes. *IEEE Trans. Inf. Theory*, **45**, 1757–1767.
113. Reed,I.S. and Solomon,G. (1960) Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.*, **8**, 300–304.
114. MacKay,D.J.C. (2005) Fountain codes. In: *IEEE Int. Sib. Conf. Control Commun., SIBCON-2005, Proc.* Vol. **152**, pp. 1062–1068.
115. Dimakis,A.G., Prabhakaran,V. and Ramchandran,K. Distributed fountain codes for networked storage. In: *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, NY, Vols **1–13**, pp. 6007–6010.
116. Chen,W.D., Kohll,A.X., Nguyen,B.H., Koch,J., Heckel,R., Stark,W.J., Ceze,L., Strauss,K. and Grass,R.N. (2019) Combining data longevity with high storage capacity—layer-by-layer DNA encapsulated in magnetic nanoparticles. *Adv. Funct. Mater.*, **29**, 1901672.
117. Luby,M. (2002) LT codes. In: *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings*. IEEE, NY, pp. 271–280.
118. Dong,Y., Sun,F., Ping,Z., Ouyang,Q. and Qian,L. (2020) DNA storage: research landscape and future prospects. *Natl. Sci. Rev.*, **7**, 1092–1107.

119. Paabo,S., Gifford,J.A. and Wilson,A.C. (1988) Mitochondrial-DNA sequences from a 7000-year old brain. *Nucleic Acids Res.*, **16**, 9775–9787.
120. Lindahl,T. (1993) Instability and decay of the primary structure of DNA. *Nature*, **362**, 709–715.
121. Lindahl,T. (1993) Recovery of antediluvian DNA. *Nature*, **365**, 700–700.
122. Prufer,K., Racimo,F., Patterson,N., Jay,F., Sankararaman,S., Sawyer,S., Heinze,A., Renaud,G., Sudmant,P.H., de Filippo,C. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.
123. Vreeland,R.H., Rosenzweig,W.D. and Powers,D.W. (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, **407**, 897–900.
124. Rashid,J.I.A. and Yusof,N.A. (2017) The strategies of DNA immobilization and hybridization detection mechanism in the construction of electrochemical DNA sensor: a review. *Sensing Bio-Sensing Res.*, **16**, 19–31.
125. Loan,P.T., Zhang,W., Lin,C.T., Wei,K.H., Li,L.J. and Chen,C.H. (2014) Graphene/MoS2 heterostructures for ultrasensitive detection of DNA hybridisation. *Adv. Mater.*, **26**, 4838–4844.
126. Pei,R.J., Cui,X.Q., Yang,X.R. and Wang,E.K. (2001) Assembly of alternating polycation and DNA multilayer films by electrostatic layer-by-layer adsorption. *Biomacromolecules*, **2**, 463–468.
127. Saurer,E.M., Flessner,R.M., Sullivan,S.P., Prausnitz,M.R. and Lynn,D.M. (2010) Layer-by-layer assembly of DNA- and protein-containing films on microneedles for drug delivery to the skin. *Biomacromolecules*, **11**, 3136–3143.
128. Cao,Y.W., Jin,R. and Mirkin,C.A. (2001) DNA-modified core-shell Ag/Au nanoparticles. *J. Am. Chem. Soc.*, **123**, 7961–7962.
129. Li,F., Zhang,H., Dever,B., Li,X.F. and Le,X.C. (2013) Thermal stability of DNA functionalized gold nanoparticles. *Bioconjugate Chem.*, **24**, 1790–1797.
130. Kim,J.W., Kim,J.H. and Deaton,R. (2011) DNA-linked nanoparticle building blocks for programmable matter. *Angew. Chem., Int. Ed. Engl.*, **50**, 9185–9190.
131. White,S.P., Dorfman,K.D. and Frisbie,C.D. (2015) Label-free DNA sensing platform with low-voltage electrolyte-gated transistors. *Anal. Chem.*, **87**, 1861–1866.
132. Moreno-Hagelsieb,L. (2004) Sensitive DNA electrical detection based on interdigitated Al/Al_2O_3 microelectrodes. *Sens. Actuators B*, **98**, 269–274.
133. Nguyen,H.H., Park,J., Hwang,S., Kwon,O.S., Lee,C.S., Shin,Y.B., Ha,T.H. and Kim,M. (2018) On-chip fluorescence switching system for constructing a rewritable random access data storage device. *Sci. Rep.*, **8**, 337.
134. Fuentes,M., Mateo,C., Rodriguez,A., Casqueiro,M., Tercero,J.C., Riese,H.H., Fernandez-Lafuente,R. and Guisan,J.M. (2006) Detecting minimal traces of DNA using DNA covalently attached to superparamagnetic nanoparticles and direct PCR-ELISA. *Biosens. Bioelectron.*, **21**, 1574–1580.
135. Li,Z., Jin,R.C., Mirkin,C.A. and Letsinger,R.L. (2002) Multiple thiol-anchor capped DNA-gold nanoparticle conjugates. *Nucleic Acids Res.*, **30**, 1558–1562.
136. Dougan,J.A., Karlsson,C., Smith,W.E. and Graham,D. (2007) Enhanced oligonucleotide-nanoparticle conjugate stability using thioctic acid modified oligonucleotides. *Nucleic Acids Res.*, **35**, 3668–3675.
137. Fixe,F., Dufva,M., Telleman,P. and Christensen,C.B. (2004) Functionalization of poly(methyl methacrylate) (PMMA) as a substrate for DNA microarrays. *Nucleic Acids Res.*, **32**, e9.
138. Ahangar,L.E. and Mehrgardi,M.A. (2012) Nanoporous gold electrode as a platform for the construction of an electrochemical DNA hybridization biosensor. *Biosens. Bioelectron.*, **38**, 252–257.
139. Guo,M., Chen,J., Liu,D., Nie,L. and Yao,S. (2004) Electrochemical characteristics of the immobilization of calf thymus DNA molecules on multi-walled carbon nanotubes. *Bioelectrochemistry*, **62**, 29–35.
140. Choi,Y., Bae,H.J., Lee,A.C., Choi,H., Lee,D., Ryu,T., Hyun,J., Kim,S., Kim,H., Song,S.H. *et al.* (2020) DNA micro-disks for the management of DNA-based data storage with index and write-once-read-many (WORM) memory features. *Adv. Mater.*, **32**, e2001249.
141. Green,N.M. (1963) Avidin .1. Use of 14c biotin for kinetic studies and for assay. *Biochem. J.*, **89**, 585–591.
142. Piran,U. and Riordan,W.J. (1990) Dissociation rate-constant of the biotin-streptavidin complex. *J. Immunol. Methods*, **133**, 141–143.
143. Jung,Y.K., Kim,T.W., Jung,C., Cho,D.Y. and Park,H.G. (2008) A polydiacetylene microchip based on a biotin-streptavidin interaction for the diagnosis of pathogen infections. *Small*, **4**, 1778–1784.
144. Song,Y., Kim,S., Heller,M.J. and Huang,X. (2018) DNA multi-bit non-volatile memory and bit-shifting operations using addressable electrode arrays and electric field-induced hybridization. *Nat. Commun.*, **9**, 281.
145. Zoltewicz,J.A., Clark,D.F., Sharpless,T.W. and Grahe,G. (1970) Kinetics and mechanism of acid-catalyzed hydrolysis of some purine nucleosides. *J. Am. Chem. Soc.*, **92**, 1741–1750.
146. Pruvost,M., Schwarz,R., Correia,V.B., Champlot,S., Braguier,S., Morel,N., Fernandez-Jalvo,Y., Grange,T. and Geigl,E.-M. (2007) Freshly excavated fossil bones are best for amplification of ancient DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 739–744.
147. Suzuki,T., Ohsumi,S. and Makino,K. (1994) Mechanistic studies on depurination and apurinic site chain breakage in oligodeoxyribonucleotides. *Nucleic Acids Res.*, **22**, 4997–5003.
148. Lindahl,T. and Nyberg,B. (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic-acid. *Biochemistry*, **13**, 3405–3410.
149. Lindahl,T. and Karlstro,O. (1973) Heat-induced depyrimidination of deoxyribonucleic acid in neutral solution. *Biochemistry*, **12**, 5151–5154.
150. Evans,R.K., Xu,Z., Bohannon,K.E., Wang,B., Bruner,M.W. and Volkin,D.B. (2000) Evaluation of degradation pathways for plasmid DNA in pharmaceutical formulations via accelerated stability studies. *J. Pharm. Sci.*, **89**, 76–87.
151. Lindahl,T. and Nyberg,B. (1972) Rate of depurination of native deoxyribonucleic acid. *Biochemistry*, **11**, 3610–3618.
152. Pogocki,D. and Schoneich,C. (2000) Chemical stability of nucleic acid-derived drugs. *J. Pharm. Sci.*, **89**, 443–456.
153. Hovorka,S.W. and Schoneich,C. (2001) Oxidative degradation of pharmaceuticals: theory, mechanisms and inhibition. *J. Pharm. Sci.*, **90**, 253–269.
154. Molina,M. and Anchordoquy,T.J. (2007) Metal contaminants promote degradation of lipid/DNA complexes during lyophilization. *Biochim. Biophys. Acta*, **1768**, 669–677.
155. Komiyama,M., Takeda,N. and Shigekawa,H. (1999) Hydrolysis of DNA and RNA by lanthanide ions: mechanistic studies leading to new applications. *Chem. Commun.*, **1999**, 1443–1451.
156. Graf,E., Mahoney,J.R., Bryant,R.G. and Eaton,J.W. (1984) Iron-catalyzed hydroxyl radical formation - stringent requirement for free iron coordination site. *J. Biol. Chem.*, **259**, 3620–3624.
157. Bucak,M.N., Tuncer,P.B., Sariozkan,S., Baspinar,N., Taspinar,M., Coyan,K., Bilgili,A., Akalin,P.P., Buyukleblebici,S., Aydos,S. *et al.* (2010) Effects of antioxidants on post-thawed bovine sperm and oxidative stress parameters: antioxidants protect DNA integrity against cryodamage. *Cryobiology*, **61**, 248–253.
158. Vilenchik,M.M. (1989) Studies of DNA damage and repair of thermal-induced and radiation-induced lesions in human-cells. *Int. J. Radiat. Biol.*, **56**, 685–689.
159. Bonnet,J., Colotte,M., Coudy,D., Couallier,V., Portier,J., Morin,B. and Tuffet,S. (2010) Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucleic Acids Res.*, **38**, 1531–1546.
160. Lyscov,V.N. and Moshkovsky,Y.S. (1969) DNA cryolysis. *Biochim. Biophys. Acta*, **190**, 101–110.
161. Zhang,M., Oldenhof,H., Sydykov,B., Bigalk,J., Sieme,H. and Wolkers,W.F. (2017) Freeze-drying of mammalian cells using trehalose: preservation of DNA integrity. *Sci. Rep.*, **7**, 6198.
162. Zhu,B., Furuki,T., Okuda,T. and Sakurai,M. (2007) Natural DNA mixed with trehalose persists in B-form double-stranding even in the dry state. *J. Phys. Chem. B*, **111**, 5542–5544.
163. Emanuele,E., Bertona,M., Sanchis-Gomar,F., Pareja-Galeano,H. and Lucia,A. (2014) Protective effect of trehalose-loaded liposomes against UVB-induced photodamage in human keratinocytes. *Biomed. Rep.*, **2**, 755–759.
164. Yoshioka,S. and Aso,Y. (2007) Correlations between molecular mobility and chemical stability during storage of amorphous pharmaceuticals. *J. Pharm. Sci.*, **96**, 960–981.

165. Hancock,B.C. and Zograf,G. (1997) Characteristics and significance of the amorphous state in pharmaceutical systems. *J. Pharm. Sci.*, **86**, 1–12.
166. Branco,C.S., Garcez,M.E., Pasqualotto,F.F., Erdtman,B. and Salvador,M. (2010) Resveratrol and ascorbic acid prevent DNA damage induced by cryopreservation in human semen. *Cryobiology*, **60**, 235–237.
167. Howlett,S.E., Castillo,H.S., Gioeni,L.J., Robertson,J.M. and Donfack,J. (2014) Evaluation of DNAstable for DNA storage at ambient temperature. *Forensic Sci. Int.: Genet.*, **8**, 170–178.
168. Ivanova,N.V. and Kuzmina,M.L. (2013) Protocols for dry DNA storage and shipment at room temperature. *Mol. Ecol. Resour.*, **13**, 890–898.
169. Wan,E., Akana,M., Pons,J., Chen,J., Musone,S., Kwok,P.-Y. and Liao,W. (2010) Green technologies for room temperature nucleic acid storage. *Curr. Issues Mol. Biol.*, **12**, 135–141.
170. Hutchinson,F. (1985) Chemical changes induced in DNA by ionizing radiation. *Prog. Nucleic Acid Res. Mol. Biol.*, **32**, 115–154.
171. Boudaıffa,B., Cloutier,P., Hunting,D., Huels,M.A. and Sanche,L. (2000) Resonant formation of DNA strand breaks by low-energy (3 to 20 eV) electrons. *Science*, **287**, 1658–1660.
172. Ward,J.F. (1988) DNA Damage Produced by Ionizing Radiation in Mammalian Cells: Identities, Mechanisms of Formation, and Reparability. *Prog. Nucleic Acid Res. Mol. Biol.*, **35**, 95–125.
173. Sutherland,B.M., Bennett,P.V., Sidorkina,O. and Laval,J. (2000) Clustered damages and total lesions induced in DNA by ionizing radiation: oxidized bases and strand breaks. *Biochemistry*, **39**, 8026–8031.
174. Teoule,R. (1987) Radiation-induced DNA damage and its repair. *Int. J. Radiat. Biol. Relat. Stud. Phys., Chem. Med.*, **51**, 573–589.
175. Wood,R.D., Mitchell,M. and Lindahl,T. (2005) Human DNA repair genes, 2005. *Mutat. Res.*, **577**, 275–283.
176. Ward,J. (1990) The yield of DNA double-strand breaks produced intracellularly by ionizing radiation: a review. *Int. J. Radiat. Biol.*, **57**, 1141–1150.
177. Elkind,M. and Redpath,J. (1977) Molecular and cellular biology of radiation lethality. In: Becker,F.F. (ed). *Radiotherapy, Surgery, and Immunotherapy*. Springer Press, Boston, Vol. **6**, pp. 51–99.
178. Reitz,G., Beaujean,R., Benton,E., Burmeister,S., Dachev,T., Deme,S., Luszik-Bhadra,M. and Olko,P. (2005) Space radiation measurements on-board ISS–the DOSMAP experiment. *Radiat. Prot. Dosimetry*, **116**, 374–379.
179. Koch,J., Gantenbein,S., Masania,K., Stark,W.J., Erlich,Y. and Grass,R.N. (2020) A DNA-of-things storage architecture to create materials with embedded memory. *Nat. Biotechnol.*, **38**, 39–43.
180. Paunescu,D., Fuhrer,R. and Grass,R.N. (2013) Protection and deprotection of DNA–high-temperature stability of nucleic acid barcodes for polymer labeling. *Angew. Chem., Int. Ed. Engl.*, **52**, 4269–4272.
181. Schaaper,R.M. (1993) Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. *J. Biol. Chem.*, **268**, 23762–23765.
182. Garmendia,C., Bernad,A., Esteban,J.A., Blanco,L. and Salas,M. (1992) The bacteriophage phi 29 DNA polymerase, a proofreading enzyme. *J. Biol. Chem.*, **267**, 2594–2599.
183. von Wintersdorff,C.J., Penders,J., van Niekerk,J.M., Mills,N.D., Majumder,S., van Alphen,L.B., Savelkoul,P.H. and Wolffs,P.F. (2016) Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.*, **7**, 173.
184. Acman,M., van Dorp,L., Santini,J.M. and Balloux,F. (2020) Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.*, **11**, 2452.
185. Davison,J. (1999) Genetic exchange between bacteria in the environment. *Plasmid*, **42**, 73–91.
186. Sheth,R.U. and Wang,H.H. (2018) DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.*, **19**, 718–732.
187. Ceze,L., Nivala,J. and Strauss,K. (2019) Molecular digital data storage using DNA. *Nat. Rev. Genet.*, **20**, 456–466.
188. Meiser,L.C., Koch,J., Antkowiak,P.L., Stark,W.J., Heckel,R. and Grass,R.N. (2020) DNA synthesis for true random number generation. *Nat. Commun.*, **11**, 5869.