Molecular digital data storage using DNA

Luis Ceze[1]*, *Jeff Nivala*[1] *and Karin Strauss*[1,2]

Abstract | Molecular data storage is an attractive alternative for dense and durable information storage, which is sorely needed to deal with the growing gap between information production and the ability to store data. DNA is a clear example of effective archival data storage in molecular form. In this Review, we provide an overview of the process, the state of the art in this area and challenges for mainstream adoption. We also survey the field of in vivo molecular memory systems that record and store information within the DNA of living cells, which, together with in vitro DNA data storage, lie at the growing intersection of computer systems and biotechnology.

Archival storage
A method of retaining information outside of the internal memory of a computer.

Random access
The ability to select a portion of the data stored and thus avoid the need to read all the data in storage.

Digital data production has been growing exponentially[1], outpacing growth of the installed base (that is, all devices in active use) of mainstream storage, including magnetic (for example, tape or hard disk drives), optical (for example, Blu-ray) and solid state (for example, flash). Most data generated are discarded, but a portion still needs to be stored. Unfortunately, based on these trends, the portion that can be retained is declining. This is a strong motivator for research in new storage media. Important aspects of storage media are density (bits per unit of physical volume), retention (time that the data are still recoverable), access speed (latency and bandwidth of accessing data) and energy cost of data, both at rest and per access. Density, durability and energy cost at rest are primary factors for archival storage, which aims to store vast amounts of data for long-term, future use. Mainstream digital data storage typically works by changing the properties of materials: electrical properties in flash and phase-change memories, optical properties in Blu-ray disks or magnetic properties in hard disk drives and tape. Although these technologies have made fast-paced progress, they are all approaching their density limits. By contrast, using as few atoms as possible to store one bit of information leads to higher density, making molecular data storage attractive.

Although there are many forms of molecular-level or atomic-level data storage, DNA stands out as an especially attractive alternative. For example, using DNA for data storage offers density of up to 10^{18} bytes per mm^3, approximately six orders of magnitude denser than the densest media available today[2–4]. The sheer density also facilitates preservation of the data in molecules for long periods of time at low energy costs. A particularly unique advantage is the ease of replication of DNA, for example, using PCR, which offers the ability to copy large amounts of data at very low time and resource cost. Once data are stored in DNA, we can also leverage the DNA hybridization process to perform operations over the data, for example, image similarity searches[5]. Finally, DNA is likely to be eternally relevant because there will always be DNA sequencers (readers), given their expanding use in life sciences and medicine. Data storage can also benefit from fast progress in DNA writing and reading by the biotechnology industry for life sciences purposes.

DNA also offers other important advantages over traditional media. DNA is time tested by nature, with DNA sequences having been read from fossils thousands of years old. When kept away from light and humidity and at reasonable temperatures, DNA can last for centuries to millenia[6] compared with decades, the typical lifetime for archival storage media such as commercial tape and optical disks. Dealing with obsolescence is a substantial component of maintenance costs for archival storage — for example, the Library of Congress spends a considerable proportion of its resources moving data over to new generations of tape, and tape drives are only backward-compatible with a limited number of past generations. In addition, copying times for traditional storage media are proportional to the amount of data to be copied and the number of replicas to be made. This is in contrast with DNA storage, through which a large number of copies can be made with a fixed time process such as PCR.

The basic process in DNA data storage involves encoding digital information into DNA sequences (encoding), writing the sequences into actual DNA molecules (synthesis), physically conditioning and organizing them into a library for long-term storage, retrieving and selectively accessing them (random access), reading the molecules (sequencing) and converting them back to digital data (decoding). In the remainder of this article, we describe the state of the art in encoding and decoding

[1]*Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA.*
[2]*Microsoft Research, Redmond, WA, USA.*
**e-mail: luisceze@cs.washington.edu*

https://doi.org/10.1038/s41576-019-0125-3

for DNA data storage, highlighting the key proposed mechanisms. We also discuss systems aspects such as random access and preservation. We then highlight the remaining challenges to mainstream adoption and contrast them with challenges to gene synthesis. We also provide an overview of DNA data storage in vivo, discuss its target use and contrast it with in vitro data storage. For readers interested in an in-depth review focused on such in vivo DNA-based memory systems that record cellular activity, we recommend a recent review by Sheth and Wang[7].

A brief history of DNA data storage

The basic concept of using DNA for data storage can be dated back to the mid-1960s, when Norbert Wiener and Mikhail Neiman discussed 'genetic memory' ideas[8,9]. However, DNA sequencing and synthesis technologies were still in their infancy then, and it was not until more than 20 years later (around the same time that Richard Dawkins discussed the idea in his 1986 book *The Blind Watchmaker*[10]) that the concept of DNA data storage was first demonstrated experimentally with Joe Davis' bioart piece 'Microvenus'. Davis encoded a 35 bit image of the ancient Germanic rune for 'female Earth'[11]. The concept was demonstrated again in 1999 as a means of hiding secret messages (steganography) in DNA microdots on paper[12]. The microdot work is unique in that it was not only the first but also remained until 2012, the only demonstration of DNA data storage that did not include an in vivo step in the storage or recovery process: beginning with Davis' project, all other subsequent works (until 2012) stored data within living cells[11–18]. Ostensibly, this was as much a practical as it was a strategic decision, as synthetic DNA was typically cloned into replicative vectors to facilitate sequencing and selection of correctly synthesized sequences.

A major breakthrough occurred in the early 2010s, when Church et al.[19] and Goldman et al.[20] independently revisited the idea of DNA data, storing on the order of hundreds of kilobytes of data and making the observation that progress in writing and reading could make DNA data storage viable in the foreseeable time frame. FIGURE 1 shows a timeline of experimental results with data volumes and the major techniques used. Notably, there is a clear exponential rate of progress in capacity, with improvements of approximately 3 orders of magnitude in a mere 6 years. Most of the studies use phosphoramidite-based DNA synthesis, a process that has been perfected over decades; enzymatic DNA synthesis is still an emerging area of research, yet it has already been successfully used for data storage[21]. For readouts, most of the studies use sequencing by synthesis, which is a commercially available sequencing method popularized by Illumina. Recently, multiple groups have had success decoding data with nanopore sequencing using the Oxford Nanopore Technology MinION platform, although at more modest data volumes. We expect these volumes to increase in the near future. In FIG. 1, note the important step in data volumes with array-based DNA synthesis, which we discuss below.

As mentioned above, most early work on DNA data storage involved in vivo cloning and storage components, for convenience, watermarking or steganography[11,13–18,22]. More recently, an emerging branch of in vivo DNA data storage harnesses advances in synthetic biology to record new information within regions

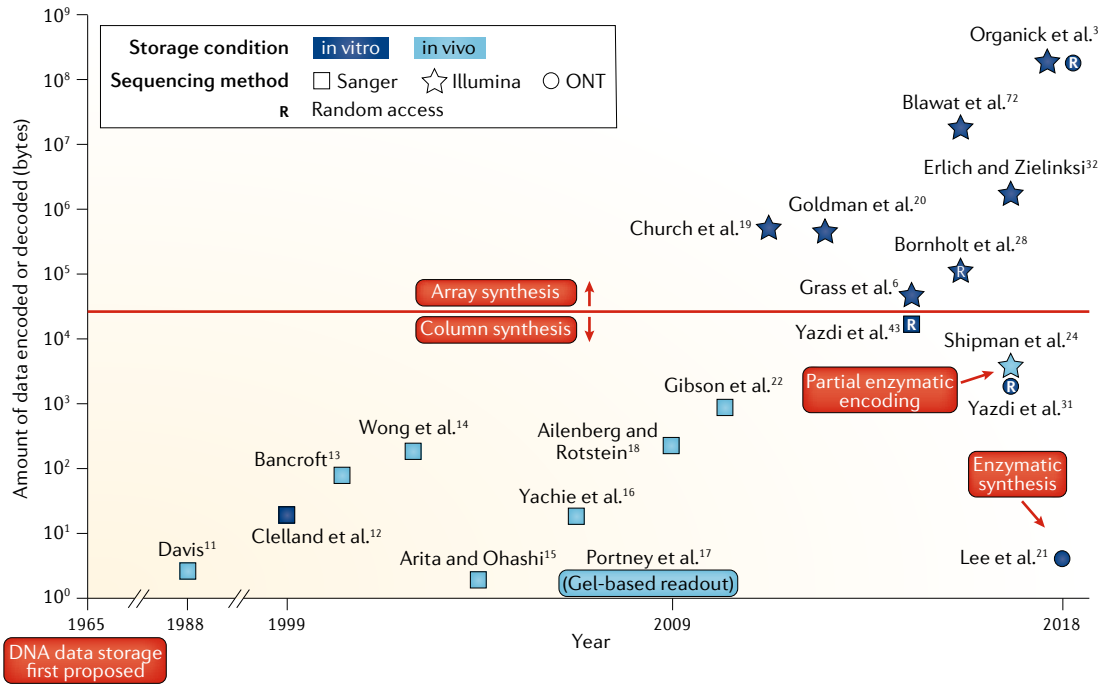

Fig. 1 | **Timeline of major published works on digital data storage with DNA.** The timeline comprises studies that included a wet-laboratory experimental demonstration. Details include how the DNA data were synthesized, stored and read and whether retrieval supported random access. Superscript numbers correspond to citations in the references section. ONT, Oxford Nanopore Technologies.

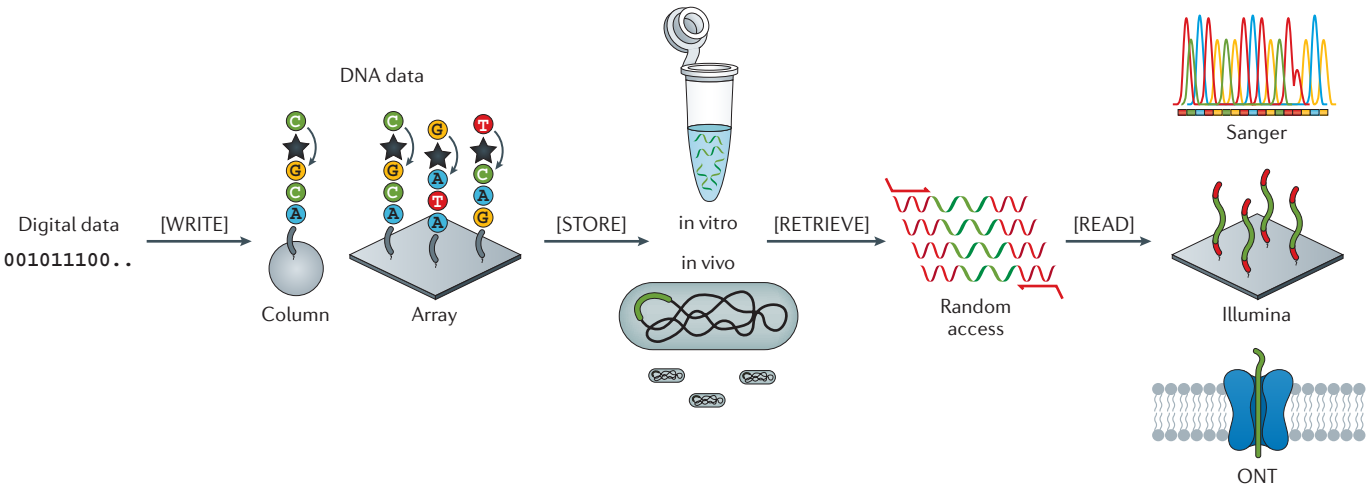

Fig. 2 | **Overview of the major steps of digital data storage in DNA.** First, a computer algorithm maps strings of bits into DNA sequences. The DNA sequences are then machine synthesized (write), thereby generating many physical copies of each sequence. Solid-phase synthesis via phosphoramidite-based chemical synthesis can be done on a column (low-throughput) or array (high-throughput) solid support. After synthesis, the resulting DNA material can be cloned and stored within a biological cell (in vivo) or, more commonly, stored in vitro, such as being frozen in solution or dried down for protection from the environment (store). DNA data requested to be read can be selectively retrieved from the DNA pool in a process called random access (retrieve). Random access within DNA data pools can be accomplished with PCR-based enrichment with primer pairs that map to specific data items generated during the encoding process. Finally, automated sequencing instruments are used to generate a set of reads that correspond to the molecules they can detect (read). The most common sequencing methods are Sanger (low-throughput) and sequencing-by-synthesis instruments (high-throughput, for example, by Illumina). More recently, nanopore sequencing (for example, from Oxford Nanopore Technologies (ONT)) has been used for real-time data reading.

of the genomes of living organisms, as described in greater detail below. Although it is unlikely that in vivo DNA data storage will be a viable alternative for general mainstream digital data storage because of its lower overall storage density owing to the relatively large size of cells, in addition to the extra complexity of modifying and/or adding to the natural DNA within living cells, in vivo DNA data recording and storage can enable new applications such as logging information about cellular history and environment[23–26]. Such systems can be thought of as molecular ticker tapes that record dynamic time-ordered molecular events within a cell and save the data for readouts at a later time[27].

Process overview

DNA data storage involves four major steps, as illustrated below and in FIG. 2.

Write. Writing data in DNA starts with encoding: a computer algorithm maps strings of bits into DNA sequences. The resulting DNA sequences are then synthesized, which generates many physical copies of each sequence. DNA sequences are of arbitrary but finite length such that bit strings are broken into smaller chunks that need to be later reassembled into the original data. To enable reassembly, it is necessary to either include an index in each chunk[6,19,28] or store overlapping chunks in different DNA sequences[20]. Heckel et al.[29] characterized the storage capacity under an index approach from a theoretical perspective and proved that a simple index-based coding scheme is optimal. For any interesting amount of information, a very large number of different DNA sequences needs to be synthesized, making it attractive to use array-based synthesis, which enables the synthesis of many unique sequences in parallel[30].

Store. Once synthesized, the resulting DNA needs to be stored. Organick et al.[3] estimate that a single physically isolated DNA pool can store on the order of 10^{12} bytes. A library of such pools is needed to scale out to large storage systems[28].

Retrieve. Once a data item is requested, the corresponding DNA pool needs to be physically retrieved and sampled. In order to avoid having to read all the data in a pool, we need what computer designers call random access, or the ability to choose a specific data item to be read out of a larger set. Whereas this feature is easy to support in mainstream digital storage media, it is more challenging in molecular storage because of the lack of physical organization across data items in the same molecular pool. Random access in DNA data storage can be supported via selective processes such as magnetic bead extraction with probes mapped to data items or PCR using primers associated to data items during the encoding process[28,31].

Read. After a sample of DNA is selected, the next step is to sequence it, producing a set of reads that correspond to the molecules detected by the sequencer. These reads are then decoded back into the original digital data with high probability. The success of this operation depends on the sequencing coverage and the error rate experienced throughout the process.

Erasures
The removal of writing, recorded material or data.

Error correcting codes
The results of mathematical manipulation of data to correct errors inserted in the data as bits are stored, transmitted and so on. The process typically involves computing a summary of the data and storing and/or transmitting it with the data and using the redundant information to correct those errors. An inner code refers to coding within a single strand to correct local errors. An outer code refers to whole new additional strands to deal with errors that are not covered by inner codes, for example, erasures.

Coping with errors during encoding and decoding

DNA synthesis and sequencing are error prone. Several papers on DNA data storage report errors of approximately 1% per base per position[28,31,32]. More precisely, for a given position in a DNA strand, when synthesized and sequenced back, approximately 1% of the reads will have an error in that position. This is for DNA synthesized in arrays using Caruthers' chemistry[33] and sequenced with sequencing by synthesis using Illumina instruments. It is interesting to note that Yazdi et al.[31] and Bornholt et al.[28] observed that most errors are due to sequencing. Organick et al.[3] also included nanopore-sequenced error information showing errors of approximately 10%. Heckel et al. have characterized the coding channel further[34] by analysing results from three previous studies[6,20,32], showing that errors mostly stem from synthesis and sequencing and that DNA manipulation, PCR and storage may cause erasures, that is, certain sequences becoming disproportionately under-represented in a mix.

It would be catastrophic to expose this level of error to end users in storage applications. Therefore, it is paramount to overlay error correcting codes on top of the raw storage media. Interestingly, modern magnetic media has a raw error rate of approximately 1% as well[35]. In a nutshell, reliable data storage and retrieval need error correction methods for all types of media (and communication channels such as radios). In fact, there is a whole field of computer science called information theory, or coding theory, that focuses on developing coding schemes that deliver digital data reliably over noisy media and communication channels. One especially interesting aspect about DNA is that, unlike other storage channels, which have only substitution errors, DNA channels can also manifest base insertions and deletions, which makes coding more challenging.

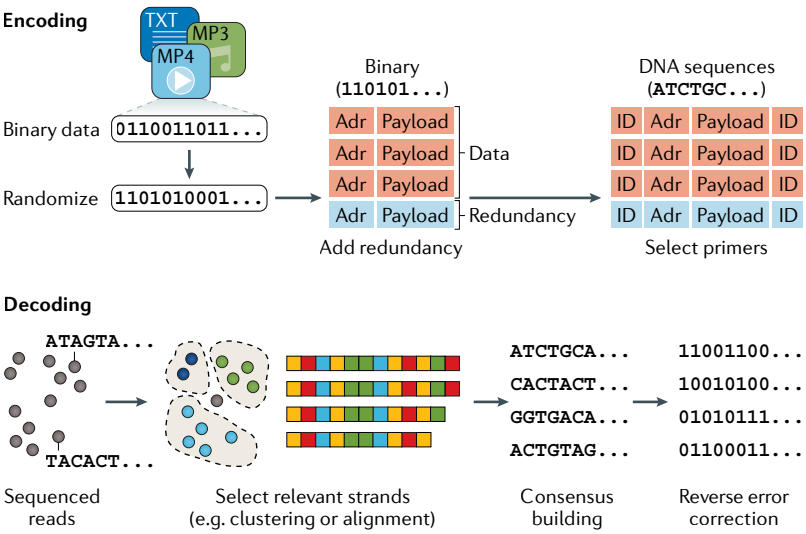

Fig. 3 | **Overview of the encoding–decoding process.** The process starts by partitioning the binary data into smaller bit sequences (payloads), each of which will fit within a DNA strand. Sequence numbers (addressing information (Adr)) are added to each bit sequence, and the data may be optionally randomized, for instance, by performing an exclusive-or operation with a pseudo-random key generated with a seed. Next, additional redundancy can be added with a code that summarizes the data to be stored. To conclude the encoding process, the bits are converted into bases, and primer target sites are added to the ends of each sequence (ID). For decoding, once the DNA is sequenced, the resulting reads are used to reconstruct the (putative) original strands, which are subsequently decoded back into bits. The originally added redundancy is used by the codes to recover any missing information and to correct remaining errors. This process is fairly similar for most of the work listed in TABLE 1, and the use of an address to recompose data scattered over multiple sequences was first used by Church et al.[19]. Organick et al.[3] also uniquely include a step that clusters reads into groups containing reads that are likely to correspond to the same original bit sequences, which enables tolerance of deletions, insertions and substitutions with potentially lower coverage. Each group is processed to reconstruct the likely original DNA strand, and finally, the codes are reverted, as in the other schemes described. Adapted from REF.[3], Springer Nature Limited.

The basics of error correction codes. Error correction codes boil down to adding redundant information, which increases the probability that the original information can be retrieved even in the presence of errors or missing data. The more redundancy, the more tolerant of errors (or losses) the resulting storage process will be. Redundancy can be of two basic types. Physical redundancy comes from having many physical copies of a given DNA sequence. Logical redundancy comes from embedding additional information when encoding data in the DNA sequences. DNA synthesis, using either columns or arrays, naturally produces a large number of physical copies of the same DNA sequence. Although physical redundancy helps with tolerating decay and allows some errors in synthesis and sequencing to be averaged out, it is not sufficient to guarantee error freedom with high probability. For example, despite a large degree of physical redundancy, Church et al.[19] did not have any form of logical redundancy and thus did not achieve zero-bit error. There are several methods for logical redundancy, which we cover below.

Multiple research projects have developed coding for DNA data storage to cope with read and write errors and with degradation of DNA over time[3,6,28,29,31,32]. FIGURE 3 provides an overview of the encoding and decoding process. When bits are encoded into DNA sequences, they go through a series of transformations and checks. As the number of bits is greater than can fit in a single DNA sequence that is synthesizable with current technologies, the bits are divided into smaller sequences. Most DNA data storage studies add indices to each smaller sequence to identify its relative location in the original file, as shown in the reassembly column of TABLE 1. These sequences may go through a logical 'exclusive-or' operation with a pseudo-random number generated from a known seed to ensure that DNA strands are dissimilar[3], and then redundancy is added via codes, such as Reed–Solomon codes, for error correction purposes. Different encoders may then use different rules to convert bits to sequences. A common conversion rule is to avoid repeated bases (homopolymers), and primer target sites are added on both ends.

The first modern error correcting codes appeared in the 1940s[36]. All error correcting codes add redundancy to the original data to be stored or transmitted over a channel. Receivers can use this extra redundant data to check whether the received message is consistent and, if it is not, to potentially reconstruct the original data. The amount of redundant data to be added can vary depending on the noise profile of a channel, on the code used and on the desired probability of successful decoding.

Table 1 | **Notable demonstrations of in vitro DNA data storage**

Study	Total data	Synthesis	Sequencing	Physical redundancy (coverage)	Reassembly	Strand length (nucleotides)	Logical density (bits per nucleotide)	Logical density (payload only)	Random access?
Church et al.[19]	650 kB	Phosphoramidite (deposition)	Sequencing by synthesis	3,000×	Index	115	0.60	0.83	No
Goldman et al.[20]	630 kB	Phosphoramidite (deposition)	Sequencing by synthesis	51×	Overlap	117	0.19	0.29	No
Grass et al.[6]	80 kB	Phosphoramidite (electrochemistry)	Sequencing by synthesis	372×	Index	158	0.86	1.16	No
Bornholt et al.[28]	150 kB	Phosphoramidite (electrochemistry)	Sequencing by synthesis	40×	Index	117	0.57	0.85	Yes
Erlich and Zielinsky[32]	2 MB	Phosphoramidite (deposition)	Sequencing by synthesis	10.5×	Luby seed	152	1.18	1.55	No
Blawat et al.[72]	22 MB	Phosphoramidite (deposition)	Sequencing by synthesis	160×	Index	230	0.89	1.08	No
Organick et al.[3]	200 MB	Phosphoramidite (deposition)	Sequencing by synthesis	5×	Index	150–200	0.81	1.10	Yes
Anavy et al.[40]	8.5 MB	Phosphoramidite (deposition)	Sequencing by synthesis	164×	Index	194	1.94	2.64	No
Choi et al.[39]	854 B	Phosphoramidite (column)	Sequencing by synthesis	250×	Index	85	1.78	3.37	No
Yadzi et al.[31,43]	3 kB	Phosphoramidite (column)	Nanopore sequencing	200×	Index	880–1,000	1.71	1.74	Yes
Organick et al.[3]	33 kB	Phosphoramidite (deposition)	Nanopore sequencing	36×	Index	150	0.81	1.10	Yes
Lee et al.[21]	18 B	Enzymatic (column)	Nanopore sequencing	175×	NA (single)	150–200	1.57[a]	1.57	No

The table includes the synthesis and sequencing methods used, encoding and decoding schemes and random-access properties for each approach. Studies differ in the availability of computer code underlying the encoding and decoding processes. The code for Church et al.[19] can be found in http://science.sciencemag.org/content/sci/suppl/2012/08/15/science.1226355.DC1/Church.SM.pdf. Data for Goldman et al.[20] are available from https://www.ebi.ac.uk/goldman-srv/DNA-storage/. Sequencing data for Organick et al.[3] can be found in https://github.com/uwmisl/data-nbt17. Computer code and sequencing data for Erlich & Zielinsky[32] can be found in http://dnafountain.teamerlich.org/. NA, not available. [a]Bits per transition in this case. TABLE 1 is adapted from REF.[3], Springer Nature Limited.

Reed–Solomon codes[37] date back to the 1960s. They are commonly used for DNA data storage purposes[3,6,38] but have also been used in a variety of other applications such as optical disks (such as compact, digital video and Blu-ray disks), 2D visual codes (such as quick response (QR) codes) and data transmission (such as WiMax). The basic idea in Reed–Solomon codes is to map the original data into a set of symbols, which is a fairly small basic unit of encoded data. Symbols map to coefficients in a system of linear equations, whose solutions are mapped back to the original data. These codes can correct two issues: a missing original symbol (called an erasure) and a corrupted original symbol (called an error). Correcting errors and erasures may require different amounts of redundancy and computational effort. The key to DNA data storage and other uses is that the same code can correct both types of issue.

Other error correction and mitigation mechanisms have been proposed for DNA data storage[20,28,32]. While not an error correcting code per se, a shifting pattern, as used by Goldman et al.[20], can be used such that a piece of data is repeated at different offsets in four different DNA sequences. For example, a piece of data may appear in the first quarter of the first DNA sequence, the second quarter of a different sequence, the third quarter of yet another sequence and the fourth quarter of a final strand. Effectively, the main goal is to position that piece of information differently in different sequences to avoid systematic errors related to the relative location on the strand, which could happen owing to synthesis and sequencing. The result is that sequences are reassembled on the basis of their overlap. This mechanism may require disproportionate synthesis efforts for the level of error protection it offers.

A slightly less intensive solution is proposed by Bornholt et al.[28]. Instead of repeating the same information in multiple locations, multiple pieces of information are summarized into one or more additional pieces by an exclusive-or operation. This mechanism reduces the amount of additional overhead effort required but is still not as efficient or robust as the Reed–Solomon-based codes. Grass et al.[6] were the first to propose using inner and outer codes as well as employing Reed–Solomon-based codes. Finally, fountain codes have been used by Erlich and Zielinski[32] in the context of DNA data storage. Fountain codes are codes that, given k original pieces of data, can generate a potentially limitless number of encoded symbols, k', such that the original data can be recovered from any k or slightly more than k pieces of data. Although fountain codes can be optimal to handle erasures, they require additional measures to detect and correct other errors (for example, an additional inner

Physical redundancy
The number of copies of each DNA species stored. Physical redundancy is not always available in the referenced work in TABLE 1, so we used the sequencing coverage as an upper bound for this number.

Logical 'exclusive-or' operation
A logic operation that outputs true only when inputs differ (that is, 0 xor 0 = 0; 0 xor 1 = 1; 1 xor 0 = 1; or 1 xor 1 = 0).

Logical density
The number of bits per nucleotide in the DNA sequences produced by the encoder.

code to detect errors). As these codes are geared towards erasure correction, their use in DNA data storage is most appropriate at low error rates (for example, sequencing by synthesis, as done by Illumina machines).

Recently, two studies have proposed the use of degenerate bases[39,40]. The basic principle is to enable the use of preset mixes of bases as symbols. For example, in addition to the basic symbols A, T, C and G, this method may also use 50% As and 50% Ts as additional symbols. This means that, at a certain position in a sequence, we may find either an A or a T with 50% probability. These methods increase logical redundancy because the composite symbols offer more than just four choices per position. These methods trade off improvements in logical density for additional physical redundancy, which is needed to resolve a symbol, because without extra physical copies, it is not possible to determine the ratio of bases in a given position. In the example of a 50:50 A and T mix as an additional symbol, the minimum theoretical physical redundancy to resolve this symbol is 2 to enable this mixed site to be represented by A in one copy and T in another copy; with mixes of two bases at 50% each, the maximum theoretical logical density per position is $\log_2(10)$ or less.

Comparison of practical DNA data storage efforts. In TABLE 1, we provide a comparison of notable practical coding efforts in DNA data storage to date. For each study, we report the total amount of data encoded in DNA, the synthesis and sequencing methods, coverage used in sequencing (equivalent to the physical redundancy available for decoding), the method used for reassembly, the length of DNA strands adopted for the work, the overall logical density, the logical density for the payload only (that is, ignoring overheads that are not intrinsic to the data, such as PCR primer target bases) and whether random access was available.

The maximum theoretical logical density achievable with the natural bases of DNA is 2, as the four bases that are possible at a single position can represent at most 2 bits. However, most strategies end up with lower density, typically around 1. This is due to overheads required for primer target sites that enable random access functionality (included only in the overall logical density value) and to redundancy added by the encoding process to facilitate the error correcting process (included in both the overall and payload-only logical density values). Error correcting codes require this redundancy to ensure that no single DNA sequence is essential to recover the data. The level of physical redundancy required for reliable data recovery is a function of the sequencing preparation protocol, raw sequencing error rates (which can be mitigated by averaging over multiple reads) and the level of logical redundancy provisioned, which enables tolerance to raw errors and loss of pieces of information. Lower physical redundancy leads to a greater likelihood of loss of sequences, thus requiring higher logical redundancy for recovery with high probability. It is worth noting that logical density is not a direct measure of physical information storage density, measured in bits per gram. This is because methods that achieve higher logical density may require higher physical redundancy, hence potentially leading to lower overall physical density. Higher logical redundancy, though, implies that more unique DNA sequences need to be synthesized and hence could lead to higher costs and lower bit-writing throughput.

DNA synthesis for data storage

Most experiments with DNA data storage so far have used the established method of phosphoramidite-based oligonucleotide synthesis[33] for writing DNA. TABLE 1 shows that the most recent demonstrations have stored the largest amounts of data, which attests to the maturity of the process. This method builds DNA strands one nucleotide at a time by using cyclic additions of reversibly 'blocked' mononucleotides to prevent the formation of unwanted homopolymers. Removing the blocking group is done with an acid solution or with a light-induced reaction using photolabile groups. Each synthesis cycle incorporates a chosen monomer into an existing polymer, strengthens the bond via oxidation, washes out excess monomers with a solvent, removes the blocking group in the last added monomer and invokes the next synthesis cycle or concludes the synthesis process. The most common synthesis errors stem from issues in removing the blocking group in the last added monomer, making insertions and deletions the prevalent error types in synthesis.

DNA synthesis can be made parallel via control mechanisms to select which bases to add to which strands. This enables the synthesis of different sequences in different spots in a solid substrate and is often called array-based synthesis[30]: sequences are seeded on a surface and reagents flow in succession to add bases in a cyclic fashion. Several technologies have been proposed, from which three are most commonly used: light-based arrays (light-activated pH change or photolabile chemistry), electronic arrays, which selectively deblock sequences and add the same base to all deblocked sequences simultaneously, and deposition-based arrays, which selectively deposit bases where they are to be added.

Achieving greater throughput for DNA synthesis depends on increased parallelism. This can be attained by one or a combination of two methods: increasing the area of the solid substrate on which the DNA is grown to fit an increased number of spots or making these spots smaller. To reduce spot dimensions, the processes described above have to be miniaturized further, which creates physical challenges: light wavelengths have to be scaled or light interference has to be used to target individual smaller spots; electronic devices have to be fabricated to manipulate smaller spots individually; and droplets have to be deposited in smaller areas. Any of these methods will inevitably result in a more erroneous synthesis process that produces fewer copies of each DNA sequence. This is problematic for biotechnology applications, which need large amounts of DNA with low rates of defects but is acceptable for DNA data storage purposes: error correcting codes allow for lower physical redundancy and higher error rates.

Enzymatic synthesis is based on an aqueous medium, as opposed to the hazardous chemicals used in phosphoramidite chemistry. In this process, template-independent

DNA-polymerizing enzymes, such as terminal deoxynucleotidyl transferases (TdTs), incorporate bases in a controllable fashion without a template. This method promises to be faster and cleaner than phosphoramidite-based synthesis. A major challenge to enzymatic synthesis is controlling single-base additions, as the TdT enzyme tends to catalyse the addition of multiple bases per cycle. Additionally, enzymatic synthesis can potentially create longer strands in faster cycle times with lower error rates[41,42]. Although enzymatic synthesis is still an emerging strategy, as can be seen by the amount of information recently stored (TABLE 1), it could be a promising method for inexpensive synthesis of longer strands as it matures. Lee at al.[21] recently demonstrated successful writing of a short message in DNA using enzymatic synthesis and reading it back using a nanopore sequencer. This method explored encoding information in the transition between different runs of the same base, cleverly sidestepping both the TdT single-base addition and the higher nanopore sequencing error challenges.

In reference to TABLE 1, a relevant comparison point for DNA data storage is strand length: although one study[43] uses long strand lengths, most use between approximately 150 and 230 nucleotide strands. The reason is longer lengths are challenging to achieve: acid-based deblocking methods suffer from depurination owing to repeated acid washes, and optical methods suffer from image drift. As phosphoramidite synthesis cannot easily achieve much longer lengths, the longer sequences are obtained via an assembly process (very similar to gene assembly), which adds to the cost and time of the writing process. Using longer strands amortizes indexing and PCR primer overheads. For most encoding schemes described earlier, the overhead reduction is substantial when increasing the length of the synthesized DNA sequences from 100 nucleotides to 400 nucleotides, but it quickly leads to diminishing returns (5% or less) for further increases beyond this length.

DNA sequencing for data retrieval

The most widespread DNA-sequencing platform today, popularized by Illumina, is based on image processing and a concept called sequencing by synthesis[44]. Single-stranded DNA sequences are attached to a substrate surface and amplified into small physical clusters of copies via PCR, and complementary bases with fluorescent markers are then attached one by one to individual sequences (in parallel for all sequences). The spatial fluorescence pattern created by the fluorescent markers is captured in an image, which is then processed, and the colour of the fluorescent spots reports the individual bases in the sequences. The fluorescent markers are then chemically removed, leaving complementary bases behind and setting up the next base in the sequence to be recognized. Scaling such technology to higher throughput will depend on more precise optical setups and improvements in image processing[45].

Another DNA-sequencing method that has been gaining momentum is nanopore technology[46]. The cornerstone of nanopore technology is the capture of DNA molecules and ratcheting them through a voltage-clamped nanoscale pore, which causes small fluctuations in electrical current through the pore that are dependent on the passing DNA strand sequence. In regard to DNA data storage, the major advantage of nanopore sequencing over competing methods is the real-time readout of the data. That is, sequence data can be streamed out of the device in essentially real time, potentially enabling data access applications that are not practical using other technologies. The main challenges in using nanopore devices for DNA storage are mitigating the higher error rates in addition to synthesizing or assembling relatively long DNA strands that can take advantage of the nanopore platform's extended read length to increase sequencing throughput[46]. A small but increasing number of studies have used nanopore technology for DNA data readout[3,21,43], specifically, the portable MinION device from Oxford Nanopore Technologies. The MinION currently achieves lower read throughput than Illumina machines, although a bench-top-sized nanopore sequencing machine is now recently available (PromethION) that may be more practical for large-scale data retrieval applications. Despite producing higher error rates than Illumina sequencers, MinION can accurately recover data by sequencing at higher coverage (more reads of the same sequence) and inferring a consensus sequence.

Achieving higher data read rates will come from higher parallelism and faster sequencing cycles. This means faster chemistry, denser sensing regions and larger flow cells. This may be an issue for optical readouts in sequencing by synthesis because the clusters need to have sufficient spatial separation to avoid overlap of fluorescence signals. A high-sequencing throughput for data storage would require a very large flow cell, potentially making this sequencing approach impractical. Nanopore-based sequencing, by contrast, is likely to offer substantially denser readout sensors because pore sizes can fundamentally be on the same order of magnitude as DNA strands. Hence, nanopore sequencing seems to offer a better scalability path and, again, error rates can be tolerated by appropriate error correction.

TABLE 1 shows that most of the demonstrations have so far used sequencing by synthesis (Illumina) instruments (column 4). Physical redundancy (column 5) — the number of copies per unique sequence — is difficult to assess accurately; therefore, we use sequencing coverage as a proxy, which is the number of physical molecules of a given sequencing observed by the sequencing instrument. This number varies from 5× to 3,000× and is a function of the ability of the decoding scheme to deal with the sequencing errors.

Random access

Scaling up DNA data storage requires a method for selectively reading pieces of data, referred to as random access in the computer science field. This is because having to sequence all the DNA in a pool to retrieve the desired data item is impractical owing to performance and cost reasons. Fortunately, selective extraction of DNA fragments is common practice in molecular biology work. Two popular methods are PCR amplification and magnetic bead extraction.

Using PCR for DNA data storage works as follows: during writing, the system assigns unique primer pairs to the different pieces of data and includes those primer sites when synthesizing the corresponding DNA sequences for the data. When a user requests a piece of data, the system finds the corresponding primers, amplifies the DNA sequences that contain the desired data and sequences a sample of the resulting pool. As indicated in TABLE 1, two independent studies proposed and demonstrated such a PCR-based system that maps data identifiers to primers when mapping digital data to DNA sequences[28,31]. A key challenge to this approach is designing primers that do not conflict with the payloads and that may enable multiplex PCR in case multiple data items are requested simultaneously.

Baum[47] proposed a theoretical system that uses magnetic bead extraction to build an associative search memory. The idea is to tag data items with an identifier that hybridizes to a molecular probe 'query'. The access is accomplished by synthesizing the desired query probe, annealing in the solution, extracting the probe, melting the molecular data attached and sequencing the results. Stewart et al.[5] recently experimentally demonstrated a method for a similarity search of images directly in DNA. The proposed method worked by encoding feature vectors of a collection of images in a DNA pool and subsequently searching for images similar to the query image (for example, find all pictures similar to an input image of binoculars). This is one example of the potential of doing more than a simple direct access of data in DNA.

Even with a method for random access within a pool, it is impractical to have to collect all data in a single pool of DNA. Very large complex mixtures will have long diffusion times and lead to less specific extractions. Organick et al.[3] provide an estimate that a DNA pool provisioned for PCR-based random access scales to the order of terabytes of data, which is sizeable but not enough to deliver on the promises of molecular data storage. To go beyond that limit, it may be necessary to create a library of physically isolated pools that are retrieved on demand. This needs to be done in a way that does not sacrifice much density and is currently an active area of research.

Challenges to mainstream adoption

It is likely that access latency (time to read) will continue to be high (minutes to hours) in the short and medium terms, but as long as bandwidth (throughput of data writing and reading) is high, in vitro DNA data storage can coexist with or potentially replace commercial media for archival data storage applications. This is because archival storage can tolerate higher latencies and would benefit considerably from smaller footprints and lower energy costs of data at rest.

The current overall writing throughput of DNA data storage is likely to be in the order of kilobytes per second. We estimate that a system competitive with mainstream cloud archival storage systems in 10 years will need to offer writing and reading throughput of gigabytes per second. This is a 6 orders-of-magnitude gap for synthesis and approximately 2–3 orders of magnitude for sequencing. On the cost gap, tape storage cost about US$16 per terabyte in 2016 (REF.[48]) and is going down approximately 10% per year. DNA synthesis costs are generally confidential, but leading industry analyst Robert Carlson estimates the array synthesis cost to be approximately US$0.0001 per base[49], which amounts to US$800 million per terabyte or 7–8 orders magnitude higher than tape. Although the throughput and cost gaps seem daunting, as has been mentioned throughout this Review, the requirements for DNA data storage are different from those of life sciences: accuracy can be sacrificed for speed, and physical redundancy can also be significantly lowered, both owing to the use of error correcting codes. This enables further scaling and performance improvements for both synthesis and sequencing methods. We expect that this will come with commensurate cost reductions because costs will be amortized over larger synthesis substrates and larger batches of DNA. Relatedly, as the number of copies per sequence required by data storage is orders of magnitude lower than it is for the life sciences, throughput improvements via more parallel synthesis and hence smaller spot size will also lead to proportional savings in reagent usage.

Finally, an important consideration is physical storage and preservation of the DNA molecules. Although there have been demonstrations of reading DNA that is thousands (or sometimes hundreds of thousands) of years old[50], DNA may degrade much faster than that, depending on the conditions to which it is exposed (for example, high temperatures, high humidity and exposure to ultraviolet light may contribute to its degradation)[4,51]. To address this issue, different groups have proposed a variety of methods to provide the appropriate conditions for DNA preservation[6,51–53]. Chemical solutions include dehydration and/or lyophilization, additives (for example, Biomatrica DNAStable or trehalose) or chemical encapsulation with protecting materials such as silicon dioxide[6]. Preparation is faster for chemical solutions and additives, whereas encapsulation provides longer shelf-life and better protection in higher humidity (50%) environments. Containers for DNA storage come in various materials and forms, such as filter paper (for example, from Whatman), airtight stainless steel minicapsules (for example, from Imagene) and plastic well plates (from Biomatrica). These items are tailored for biological samples and optimized for purity; hence, density and cost are compromised. Physical libraries for DNA data storage need to offer a path to full automation and scalability without significantly compromising density. This is still a largely open research topic. There are multiple challenges to automation of these systems to enable their use for large-scale archival storage. Such environments typically operate with minimal human interference. Most DNA manipulations outside of synthesis and sequencing are still being performed by humans in laboratory environments. Recently, Takahashi et al.[54] have made public a first demonstration of a fully automated DNA data storage system. Recent advances in microfluidics[55–58] are encouraging, and we expect them to be used for automation of DNA data storage.

In vivo data storage

Although we have discussed DNA data storage as a relatively nascent field, the concept could be regarded as ancient as life itself. Living organisms have used DNA to store and propagate their biological blueprints for

Access latency
The time needed to retrieve data.

billions of years. The human genome, for instance, contains slightly more than 3 billion bp of DNA. This equates to approximately 1.5 GB of information per diploid human cell (assuming the maximum of 2 bits per bp). Early work in DNA data storage focused on archiving digital data within genomic or plasmid DNA and explored in vivo storage as a means of heritable information storage and steganography[13,14]. More recently, with the emergence of the field of synthetic biology, in vivo genetic systems have been engineered as 'molecular recorders' that enable both the collection and permanent storage of new data within cellular DNA[7].

A number of researchers have demonstrated different approaches to in vivo molecular recording and data storage that can be categorized by both the mechanisms used for writing the data into the cell (the recording process) and the biopolymer used for storing the information (for example, DNA, RNA or protein)[59]. The most relevant to this Review are synthetic systems that store biological memory or digital data in DNA permanently within a cell (FIG. 4). The other major class of memory systems is based on transcriptional circuits that store information through regulatory feedback loops and toggle switches at the RNA or protein level[60]. These transcriptional circuit systems require the input of energy to maintain state, meaning that cell death results in information loss. By contrast, DNA-based data stored within a cell — for example, within a genome or plasmid — share many of the features that make in vitro DNA data storage attractive (for example, stability, scalability and eternal medium relevance), are naturally copied with high fidelity during cellular replication and are preserved after cell death[61].

One robust approach to in vivo DNA-based data storage is through the use of a class of enzymes known as recombinases[61,62]. Site-specific recombinases recognize specific flanking sequence motifs and enzymatically invert the intervening region of DNA. Information is stored digitally in the directionality of the DNA segment, which has two possible states and hence encodes a single bit of data. Readout is typically done by sequencing, PCR or reporter protein, whereas writing can be chemically controlled with input molecules (inducers) that activate the expression of the recombinase. These types of recombinase-based recording systems can be either irreversible (single write) or reversible with the addition of a second enzyme (excisionase) that participates in catalysing segment reversion[61,62]. Systems harnessing several unique site-specific recombinases have been used to build more complex memory systems such as genetic counters and finite-state machines with the potential to remember events (molecular inputs) and their order[63,64]. However, an important limitation with these types of memories is the need to have an orthogonal recombinase for every unique bit. To address this issue, bioinformatic mining has been used to uncover new recombinases with the potential to expand the number that could be used together in a single system, thereby increasing information capacity and scalability[65]. This recent work led to the discovery of a set of 13 orthogonal recombinases

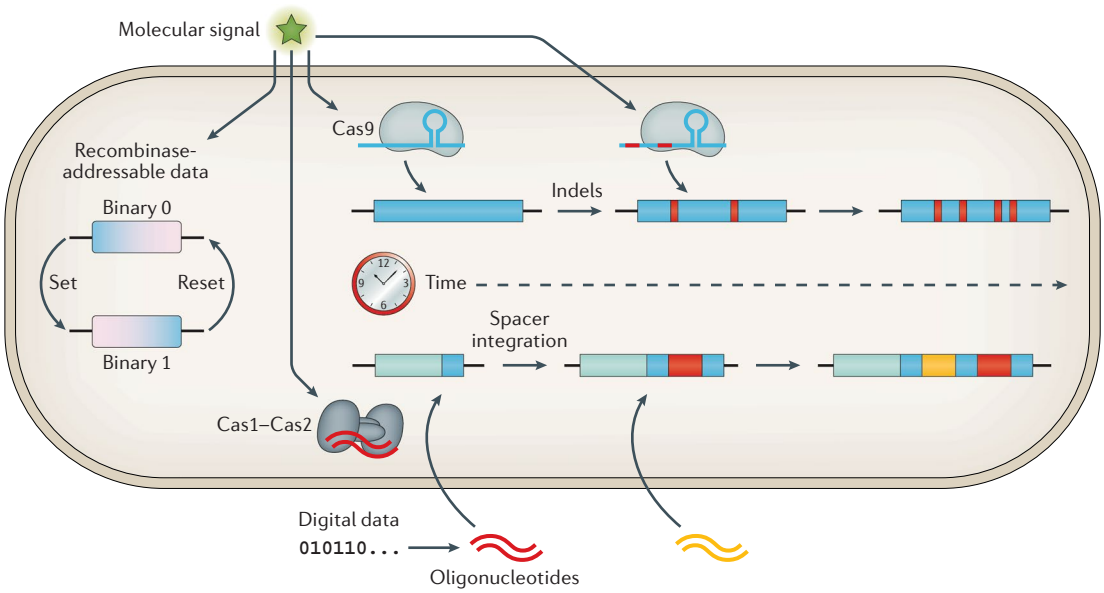

Fig. 4 | **Overview of in vivo strategies for molecular recording and storage of data in DNA.** Recording modules can be coupled to sensor components in the cell, enabling transduction of specific molecular signals (such as those from inducer molecules) into modifications of genomic or plasmid DNA at defined loci. In recombinase-addressable data systems, expression of a recombinase flips a defined segment of DNA, which corresponds to setting a bit from 0 to 1. Expression of an excisionase can reset the bit back to 0 by flipping the DNA segment into its original orientation. CRISPR–Cas9-based recording strategies use Cas9 and CRISPR guide RNAs to target specific sites within the genome for editing. Cleavage of these sites by Cas9 results in small insertions or deletions (indels) that can be used to log the magnitude and duration of molecular signals over time. In CRISPR–Cas1–Cas2 methods, the Cas1–Cas2 integration complex inserts short pieces of DNA (approximately 30 bp) called spacers into a specific locus known as the CRISPR array. New spacers are integrated upstream of any previously acquired spacers, forming a temporal memory bank of spacer sequences. Digital data can be encoded within synthetic oligonucleotides and integrated within the CRISPR array.

that could theoretically be used together to store up to 1.375 bytes of information within a single cell. It remains to be seen how much further the encoding capacity of these types of recombinase-based memory systems can be expanded, through either additional part-mining or protein-engineering efforts. A potentially more scalable recording strategy has been shown that uses a non-site-specific recombinase and single-stranded DNA as the input[66]. The main advantage of this system is the ability to introduce defined mutations at arbitrary loci, although the writing efficiency is probably too low for digital data storage applications.

Another class of emerging strategies for in vivo recording and data storage systems harnesses components of the recently discovered CRISPR–Cas system[67]. CRISPR–Cas is a microbial adaptive immune system that protects prokaryotic cells from invading viruses. This defence system 'remembers' DNA and RNA sequences derived from viral genomes and has evolved a repertoire of Cas proteins that have unique capabilities to process nucleic acids. The most popular is Cas9, a programmable DNA-cutting enzyme that has revolutionized the genome-editing field[68]. In one recording strategy, Cas9 expression is put under the control of a specific input and programmed to bind to and cut its own targeting sequence (encoding the guide RNA)[23]. Each round of cutting and subsequent repair of the target site results in unique changes, such as point mutations and insertions or deletions (indels), that serve as 'evolving barcodes', which report on the magnitude and duration of the input. Similar systems have been used in other applications to track aspects of cellular histories such as lineage[69]. An advantage of these Cas9-based recording systems is their ability to record a nonbinary range of mutations (barcodes) into DNA over time. This feature enables analogue writing of data with greater recording capacity than the more digital recombinase-based systems covered above. By contrast, recording and reading out predefined digital information, as discussed in the previous sections, would be difficult to implement with a Cas9-based encoding–decoding scheme compared with the binary recombinase writers because the exact sequence of the resulting edits is semi-random. Recently, a solution to this problem was demonstrated with a catalytically inactive Cas9 (dCas9)–cytidine deaminase fusion protein, which enabled the introduction of additional defined mutations at specific locations[26].

Finally, in addition to Cas9, Cas1 and Cas2 are another pair of Cas proteins that have been used for molecular information storage in vivo. The Cas1–Cas2 complex is an integrase that is essential to the adaptation phase of CRISPR–Cas immunity and is responsible for integrating short pieces of viral DNA (called spacers) into a precise location within the cellular genome (the CRISPR array). A critical feature of the integration process is that new spacers are (almost) always put ahead of any older spacers previously acquired, making the CRISPR array a temporal memory bank of nucleic acid sequences. For use as a molecular recorder, digital data can be encoded into pools of short segments of synthetic DNA and introduced into a population of cells expressing Cas1–Cas2, in which the synthetic DNA will be integrated into the CRISPR array within the cellular genome[24,70]. This recording process can be repeated over time by introducing unique data during each round while the temporal ordering of the spacer sequences is preserved within the array. This approach was the first in vivo recording method to enable the encoding and decoding of meaningful amounts of digital data. This highlights the unique advantages of the Cas1–Cas2 recording system in that arbitrary information can be stored within defined sequences and uploaded into the genome on demand, as demonstrated with the encoding of a short movie (totalling 2.6 kilobytes of data) within a population of bacteria[70]. In addition to the near-unconstrained sequences of the spacers themselves (mode 1) and their ordering within the array (mode 2), expression of wild-type and mutant forms of Cas1–Cas2 can be modulated to control the direction of spacer integration (mode 3), enabling an additional layer of information to be stored (similar to the recombinase-based systems)[24]. The Cas1–Cas2 recording system has also been used to report on biological inputs using exogenous DNA and integration of orientation modulation[24] and with the use of endogenously generated DNA spacers under the control of an inducible copy number promoter[25].

Conclusions

In comparing in vitro and in vivo DNA data storage, it is clear that in vitro storage is currently the most practical form of storage with regard to cost, scalability and stability. However, the in vivo storage systems can be used as biological recording devices that are better suited to collecting new data than preserving digital data already in hand. That being said, additional advantages of in vivo data storage may not yet have been realized. For instance, it has recently been demonstrated that in vivo storage of data within *Escherichia coli* may be a practical means of micro-scale random data access[71]. As the field of synthetic biology continues to mature, in vivo data storage may yet provide answers to lingering drawbacks of in vitro storage methods.

The Digital Revolution has transformed humanity's relationship with data, ushering society into the Information Age. The ever-expanding types and sheer quantity of data that we are generating are overwhelming our current technological storage capacities. New forms of digital data storage are required to keep pace. DNA data storage is a promising alternative to contemporary mainstream formats such as tape and disk, which are now approaching their density limits. The time-tested durability and eternal relevance of DNA make it a natural choice for long-term data archival. At the same time, DNA synthesis, sequencing and retrieval technologies, originally developed for life sciences applications, can be repurposed in data storage systems. As research into DNA data storage continues to progress, we anticipate technological innovations that are tailored for DNA data storage, which promise to gradually decrease barriers to its mainstream adoption.

Published online: 08 May 2019

1. Reisel, D., Gantz, J. & Rydning, J. Data age 2025: the digitization of the world from edge to core. *Seagate* https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (2018).
2. Rutten, M. G. T. A., Vaandrager, F. W., Elemans, J. A. A. W. & Nolte, R. J. M. Encoding information into polymers. *Nat. Rev. Chem.* **2**, 365–381 (2018).
3. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018). **This study presents an end-to-end discussion of DNA data storage, demonstrating the ability to perform random access at a large scale, the first error correction that tolerates insertions and deletions, and the largest amount of digital data in DNA as of 2019**.
4. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366–370 (2016). **This paper presents a detailed analysis of properties of DNA as a data storage medium and compares it with other media**.
5. Stewart, K. et al. in *DNA Computing and Molecular Programming* (eds Doty, D. & Dietz, H.) 55–70 (Springer International Publishing, Cham, 2018).
6. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem.* **54**, 2552–2555 (2015). **This study introduces the first robust system based on error correcting codes using inner codes and outer codes for DNA data storage, and it demonstrates silica encapsulation for greater durability**.
7. Sheth, R. U. & Wang, H. H. DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.* **19**, 718–732 (2018).
8. Wiener, N. Interview: machines smarter than men? *US News World Rep.* **56**, 84–86 (1964).
9. Neiman, M. S. On the molecular memory systems and the directed mutations. *Radiotekhnika* **6**, 1–8 (1965).
10. Dawkins, R. *The Blind Watchmaker* (Longman Scientific & Technical, 1986).
11. Davis, J. Microvenus. *Art J.* **55**, 70–74 (1996).
12. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399**, 533–534 (1999).
13. Bancroft, C. Long-term storage of information in DNA. *Science* **293**, 1763–1765 (2001).
14. Wong, P. C., Wong, K.-k. & Foote, H. Organic data memory using the DNA approach. *Commun. ACM* **46**, 95–98 (2003).
15. Arita, M. & Ohashi, Y. Secret signatures inside genomic DNA. *Biotechnol. Prog.* **20**, 1605–1607 (2004).
16. Yachie, N., Sekiyama, K., Sugahara, J., Ohashi, Y. & Tomita, M. Alignment-based approach for durable data storage into living organisms. *Biotechnol. Prog.* **23**, 501–505 (2007).
17. Portney, N. G., Wu, Y., Quezada, L. K., Lonardi, S. & Ozkan, M. Length-based encoding of binary data in DNA. *Langmuir* **24**, 1613–1616 (2008).
18. Ailenberg, M. & Rotstein, O. D. An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47**, 747–754 (2009).
19. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628–1628 (2012).
20. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013). **Church et al. (2012) and Goldman et al. (2013) feature key work on the modern reincarnation and demonstration of DNA data storage ideas**.
21. Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Enzymatic DNA synthesis for digital information storage. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/348987v1 (2018).
22. Gibson, D. G. et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–57 (2010).
23. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* **353**, aag0511 (2016).
24. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016). **This paper describes the first demonstration that the CRISPR–Cas adaptation system can be used to store DNA oligonucleotides of arbitrary sequence within the genome**.
25. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).
26. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018).
27. Glaser, J. I. et al. Statistical analysis of molecular signal recording. *PLOS Comput. Biol.* **9**, e1003145 (2013).
28. Bornholt, J. et al. A DNA-based archival storage system. Presented at the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '16) (2016).
29. Heckel, R., Shomorony, I., Ramchandran, K. & Tse, D. N. Fundamental limits of DNA storage systems. Presented at the 2017 IEEE International Symposium on Information Theory (ISIT) (2017).
30. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
31. Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A. Rewritable, random-access DNA-based storage system. *Sci. Rep.* **5**, 14138 (2015). **This paper proposes PCR-based random access**.
32. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
33. Caruthers, M. H. The chemical synthesis of DNA/RNA: our gift to science. *J. Biol. Chem.* **288**, 1420–1427 (2013).
34. Heckel, R., Mikutis, G. & Grass, R. N. A characterization of the DNA data storage channel. Preprint at *arXiv* https://arxiv.org/abs/1803.03322 (2018).
35. Albrecht, T. R. et al. Bit-patterned magnetic recording: theory, media fabrication, and recording performance. *IEEE Trans. Magn.* **51**, 0800342 (2015).
36. Shannon, C. The mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
37. Reed, I. S. & Solomon, G. Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.* **8**, 300–304 (1960).
38. Rashtchian, C. et al. Clustering billions of reads for DNA data storage. *NIPS* https://papers.nips.cc/paper/6928-clustering-billions-of-reads-for-dna-data-storage.pdf (2017).
39. Choi, Y. et al. Addition of degenerate bases to DNA-based data storage for increased information capacity. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/367052v1 (2018).
40. Anavy, L., Vaknin, I., Atar, O., Amit, R. & Yakhini, Z. Improved DNA based storage capacity and fidelity using composite DNA letters. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/433524v1 (2018).
41. Jensen, M. A. & Davis, R. W. Template-independent enzymatic oligonucleotide synthesis (TiEOS): its history, prospects, and challenges. *Biochemistry* **57**, 1821–1832 (2018).
42. Palluk, S. et al. De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.* **36**, 645–650 (2018).
43. Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Sci. Rep.* **7**, 5011 (2017).
44. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
45. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
46. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
47. Baum, E. B. Building an associative memory vastly larger than the brain. *Science* **268**, 583–585 (1995).
48. Fontana, R. E. & Decad, G. M. Moore's law realities for recording systems and memory storage components: HDD, tape, NAND, and optical. *AIP Adv.* **8**, 056506 (2018).
49. Carlson, R. Guesstimating the size of the global array synthesis market. *Synthesis* http://www.synthesis.cc/synthesis/2017/8/guesstimating-the-size-of-the-global-array-synthesis-market (2017).
50. Allentoft, M. E. et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.* **279**, 4724–4733 (2012).
51. Bonnet, J. et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage. *Nucleic Acids Res.* **38**, 1531–1546 (2009).
52. Ivanova, N. V. & Kuzmina, M. L. Protocols for dry DNA storage and shipment at room temperature. *Mol. Ecol. Resour.* **13**, 890–898 (2013).
53. Howlett, S. E., Castillo, H. S., Gioeni, L. J., Robertson, J. M. & Donfack, J. Evaluation of DNAstable™ for DNA storage at ambient temperature. *Forens. Sci. Int. Genet.* **8**, 170–178 (2014).
54. Takahashi, C. N., Nguyen, B. H., Strauss, K. & Ceze, L. H. Demonstration of end-to-end automation of DNA data storage. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/439521v1 (2018).
55. Choi, K., Ng, A. H., Fobel, R. & Wheeler, A. R. Digital microfluidics. *Annu. Rev. Anal. Chem.* **5**, 413–440 (2012).
56. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
57. Willsey, M. et al. in *Proc. 24th Int. Conf. on Architectural Support for Programming Languages and Operating Systems* 183–197 (ACM, 2019).
58. Newman, S. et al. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat. Commun.* **10**, 1706 (2019).
59. Inniss, M. C. & Silver, P. A. Building synthetic memory. *Curr. Biol.* **23**, R812–R816 (2013).
60. Burrill, D. R. & Silver, P. A. Making cellular memories. *Cell* **140**, 13–18 (2010).
61. Ham, T. S., Lee, S. K., Keasling, J. D. & Arkin, A. P. Design and construction of a double inversion recombination switch for heritable sequential genetic memory. *PLOS ONE* **3**, e2815 (2008).
62. Bonnet, J., Subsoontorn, P. & Endy, D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc. Natl Acad. Sci. USA* **109**, 8884–8889 (2012).
63. Friedland, A. E. et al. Synthetic gene networks that count. *Science* **324**, 1199–1202 (2009).
64. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based state machines in living cells. *Science* **353**, aad8559 (2016).
65. Yang, L. et al. Permanent genetic memory with >1-byte capacity. *Nat. Methods* **11**, 1261–1266 (2014).
66. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).
67. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526**, 55–61 (2015).
68. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
69. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).
70. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).
71. Tavella, F. et al. DNA molecular storage system: transferring digitally encoded information through bacterial nanonetworks. Preprint at *arXiv* https://arxiv.org/abs/1801.04774 (2018).
72. Blawat, M. et al. Forward error correction for DNA data storage. *Procedia Comput. Sci.* **80**, 1011–1022 (2016).

Acknowledgements

The authors thank S. Yekhanin for input on coding methods and R. Carlson, D. Carmean, G. Seelig, B. Nguyen, L. Organick, Y.-J. Chen, K. Stewart, S. D. Ang, M. Willsey, C. Takahashi and R. Lopez for helpful general discussions on DNA data storage. This work was supported, in part, by sponsored research agreements with Microsoft and Oxford Nanopore Technologies and gifts from Microsoft and DARPA under the Molecular Informatics Program.

Author contributions

All authors contributed to all aspects of the manuscript.

Competing interests

L.C. is a consultant to Microsoft and a Venture Partner at Madrona Venture Group. K.S. is employed by Microsoft. J.N. is a consultant to Oxford Nanopore Technologies.

Publisher's note

Reviewer information

Nature Reviews Genetics thanks R. Heckel and the other anonymous reviewer(s) for their contribution to the peer review of this work.