DNA Data Storage Technology Review

Version 1.0

30-June-2025

Technical White Paper

ABSTRACT: This white paper provides an overview of the state of DNA data storage technology, the metrics important to measuring commercial readiness, and the challenges to reaching commercial readiness.

USAGE

DISCLAIMER

The information contained in this publication is subject to change without notice. SNIA makes no warranty of any kind with regard to this specification, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. SNIA shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this specification.

Suggestions for revisions should be directed to https://www.snia.org/feedback/.

Table of Contents

Contents

1 Introduction

Human society is creating vast amounts of digital data at ever increasing rates. This data has significant value when mined, stitched together, or otherwise searched and analyzed. Further, trends in AI/ML are accelerating the ability to do this search and analysis, increasing the potential value of saved data. This is leading to a “save/discard” dilemma as users wish to retain data for extended periods to maximize potential value, while the costs of retaining this data on existing storage technology is becoming prohibitive. The capital costs associated with traditional storage media are not scaling with the rate of data generation, and operational costs of refreshing data, or creating copies, using existing storage technologies is becoming prohibitive, with the refresh cycle of some large archives needing to start, or nearly so, by the time the previous refresh finishes. Underlying this trend is uncertainty regarding the future scaling potential of existing media. The rate of HDD and Tape media storage density growth is slowing, and media lifetime is not improving significantly. TCO seems challenged when considering the long term storage requirements. The market needs storage solutions that are very dense, durable for decades (or longer) at room temperature, use zero power at rest, and require no/minimal technology refresh.

Recent academic and industrial demonstrations establish DNA Data Storage as a viable potential solution for these requirements. DNA offers information density (bits/mm^3) orders of magnitude higher than traditional media. When stored away from oxygen, water, and UV light, DNA media is stable at ambient conditions for long periods, from several decades to centuries[1]. The ubiquity of DNA in biological systems and its centrality to human health ensures that the technologies to write and read it will never become unavailable, and its universal molecular format ensures that future reading technologies can be applied to DNA data archives (i.e., the reader need not be packaged away with the data). Provided DNA writing and reading technologies improve sufficiently, these attributes point to the potential for DNA as a sustainable, low-cost storage solution that does not require periodic technology refresh.

This document provides an overview of the progress toward commercialization of DNA data storage. We first review the state of the art for each step in the end-to-end DNA data storage workflow (Figure 1). These sections also introduce key performance attributes and associated metrics by which progress toward commercialization can be discussed and measured. Lastly, we summarize challenges to commercialization.

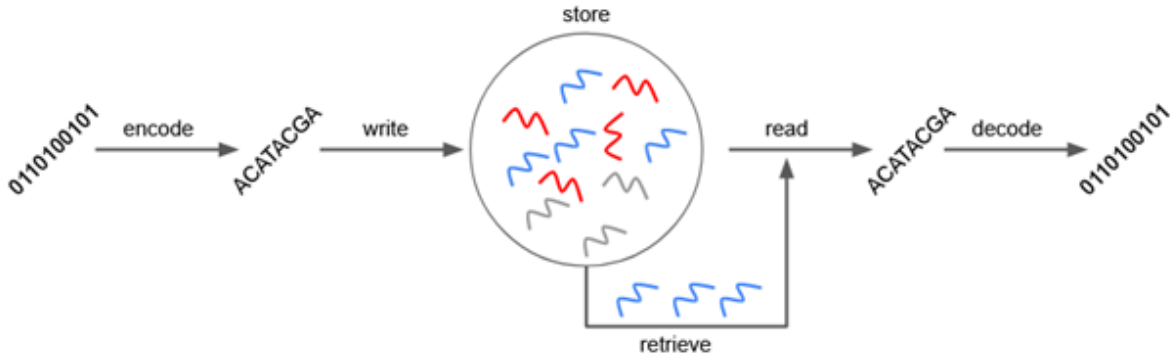

Figure 1: Key steps in end-to-end DNA data storage.

2 The DNA Codec

The DNA Codec is a software layer, often enhanced with hardware acceleration, that implements the DNA channel model for a DNA data storage system[2]. A DNA codec translates a source file of digital (1's and 0's) to DNA bases (A, C, G T), and back again. The codec operates in "symbol space," passing strings of symbols to the synthesizer, the "write" end of the DNA Physical Layer, and receiving strings of symbols from the sequencer, the "read" end of the Physical Layer. Figure 2 shows an example of a DNA codec flow. The order and the selection of the transformations and steps will vary, per codec, use case, and the properties of the DNA Physical Layer.

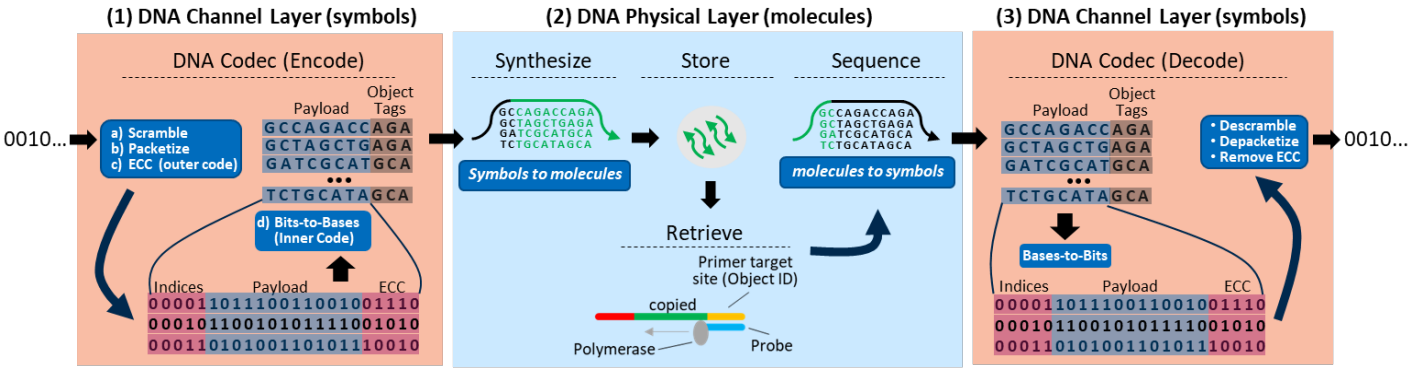

Figure 2 – Example of a DNA Channel (reprinted and edited from [2]):In step 1, the source bitstream is (1a) randomly scrambled, mitigating problematic sequences, (1b) packetized into large blocks which are then coded with ECC (outer code), and (1c) encoded from bits-to-bases (inner code), which divides the large blocks into small DNA sequences that are compatible with the properties of the Physical Layer chemistry. Also, in 1c, object tags (primers) may be added that can be used to retrieve all DNA segments in a pool associated with a particular digital object. Next (step 2), the now fully "line coded" DNA sequences are passed to the DNA Physical Layer for writing (synthesis), storing, retrieval, and reading (sequencing). Lastly (step 3), the recovered DNA sequences are passed back to the codec where they are converted back to bits and decoded, reversing all the transformations, error correction, packetization, etc., done on the encoding side.

While DNA codecs are, today, customized for the particular DNA data storage pipeline on which they are intended to be applied (because there are no standard pipelines), there are a number of common general areas of functionality which they all implement:

1. **Data representation:** The encode (bits-to-bases) and decode (bases-to-bits) functions.
2. **Error correction:** The addition of ECC (in digital or DNA symbol space) to enable correction of the various physical errors that may occur in the DNA Physical Layer.
3. **DNA sequence constraints:** Avoiding sequences of bases that increase the probability of causing synthesis or sequencing errors, or that violate a policy (e.g., biosecurity).
4. **Data storage protocols:** Various transformations and/or metadata that are used for managing, or managing the access to, the DNA archive, such as packetization, archive identity and structure discovery, object tagging/random access, etc.

The following sections go into further detail on the general functions of DNA Codecs. For a review of some popular DNA data storage codecs, please see the DNA Data Storage Alliance white paper, *DNA Data Storage Codecs: Examples, Requirements, and Metrics*.

2.1 Data Representation (Bits to Bases)

DNA molecules are composed of four naturally occurring bases (aka nucleotides, or nt): A, C, G, T. To encode digital information into this set of four bases, a variety of data representations are used; the most common are described below. This list is not comprehensive, as novel data representations are an area of active research, but all share the goal of balancing encoding density (bits/nucleotide) with computational complexity, the ability to avoid hard to write or read sequences, etc.

2.1.1 Binary Representation

The most basic representation of digital data in DNA is to simply encode the four DNA bases to binary values; that is, {A, C, G, T} is encoded as {00, 01, 10, 11}. In this representation, each base encodes 2 bits.

2.1.2 Ternary Representation

Another approach to encoding digital data in DNA is to represent the data in base-3 (ternary).[3] For example, converting the ternary string 10211_3 to bases using Table 1 below would yield the string “CGCTC”.

		Previous (or 1st) ternary number		
		0	1	2
Previous (or 1st) Base	T	A	C	G
	G	T	A	C
	C	G	T	A
	A	C	G	T

Table 1 - Ternary encoding example

This approach avoids generating strings of repeated bases (homopolymers), which are problematic for both synthesis and sequencing. In this representation, each base encodes $\log_2(3) = 1.58$ bits.

2.1.3 Representation by Transitions Between Homopolymers

This approach to encoding arbitrary data tolerates homopolymers.[4] Rather than assigning values to the bases directly, the data is represented by transitions between homopolymers. Each transition encodes $\log_2(3) = 1.58$ bits.

2.1.4 Representation by Combinatorial Assembly

This approach uses short DNA sequences (shortmers) as building blocks. By linking shortmers together through DNA assembly techniques, such as ligation enzymes, it enables the creation of significantly longer DNA molecules than traditional base-by-base synthesis methods allow[5]. The shortmers themselves, or assembled combinations of shortmers, can be used as an “alphabet” to encode data. These methods sacrifice some encoding density for potential benefits. (See section 3.1.2).

2.1.5 Representation by Topological Modification

In this technique, also known as “DNA punch card”, data is converted into a positional code that identifies locations where a template DNA sequence is topologically modified (i.e. nicked).[6] A nick is a cut in the sugar-phosphate backbone between two adjacent nucleotides in double-stranded DNA, and each nick encodes either $\log_2(2) = 1$ bit (only one strand nicked or not) or $\log_2(3) = 1.58$ bits (either or neither strand nicked).

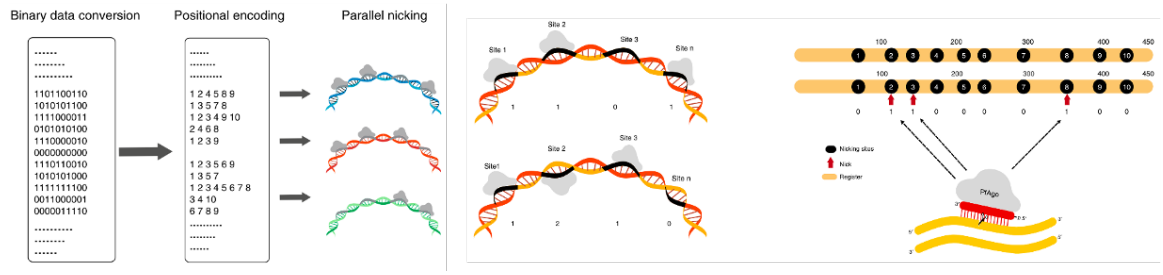

Figure 3: Demonstration of “DNA Punch Card” where information is encoded by nicking the sugar-phosphate backbone of double stranded DNA. Reprinted from [6].

2.1.6 Representation by DNA Nanostructures

Certain DNA nanostructures (“hairpins”, ***aka*** stem-loop structures) can be detected and resolved by solid-state nanopore sequencers.[7] (Figure 4) Given a library, H, of these structures, each structure h, where $h \in H$, can represent $\log2(|H|)$ bits.

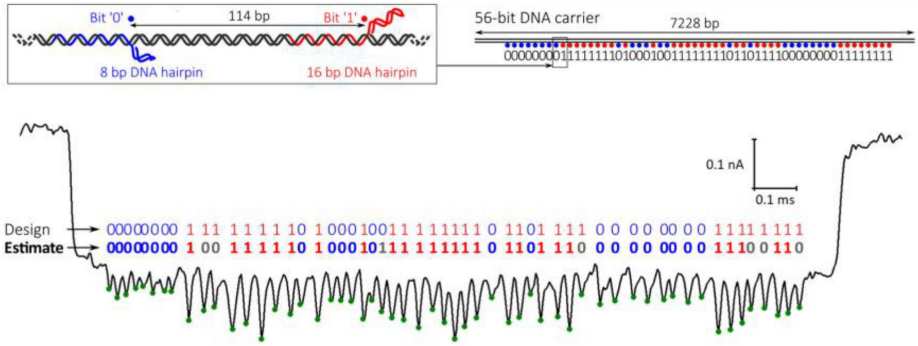

Figure 4: Digital data encoded by DNA hairpins (bit ‘0’ is 8 bp hairpin, bit ‘1’ is 16 bp hairpin) which hybridize to a double stranded carrier molecule. Bits are recovered by reading the position and magnitude of signal corresponding to hybridized hairpins. Reprinted from [7].

2.1.7 Composite DNA letters

In an effort to increase the information density (bit/nt) within an encoded DNA sequence, specific locations within the sequence are set aside for degenerate base addition, effectively increasing the size of the available alphabet.[8] During synthesis, when the degenerate positions are being synthesized, a predetermined mixture of two bases are included in the addition step. This approach relies on high-accuracy DNA sequencing to differentiate between strands whose sequences vary at degenerate locations.

2.2 Error Correction

All stages in the DNA pipeline introduce a certain percentage of errors that DNA Codecs must account for. Errors are characterized as insertions, deletions, substitutions, and erasures. Insertions, deletions, and substitutions occur at the base/bit level at specific locations within a given DNA strand (Table 2). Erasures occur at the whole molecule level, where whole molecules cannot be recovered. Error types that occur during DNA synthesis, storage and sequencing are discussed in more detail in their respective sections.

Target Sequence	GACTGGA
Insertion (red G)	GACTG**G**GA
Deletion (red X)	G**X**CTGGA
Substitution (red T)	G**T**CTGGA

Table 2 – Examples of insertion, deletion, and substitution error modes.

DNA synthesis and sequencing for genomic applications was developed to achieve extremely low error rates to ensure quality and accuracy in life science applications. The standard raw (i.e., uncorrected) error rates from pure DNA synthesis are typically in the 1% range[9], with downstream steps of precise error correction driving near perfect or perfect quality for all resulting synthesized products. For DNA sequencing, raw error rates range from ~0.1% for sequencing by synthesis (SBS)[46] to ~6% for nanopore sequencing[10], with a variety of techniques and parameters (consensus reads, read length, etc.) used to deliver an acceptably reliable final result. For data storage, there is more flexibility on tolerance of raw error rates than for life science applications both because of how a well-tuned DNA data storage channel (DNA codec and the DNA physical layer, see Figure 2) can compensate for channel errors, enabling the reliable recovery of the encoded digital data, and because of the inherently less stringent accuracy requirements for digital data storage vs. the requirements of life sciences use cases.

2.2.1 Common Error Correction Codes

Error correction codes, first introduced in the 1940's with Claude Shannon's mathematical theory of communication, are used and adapted by DNA codecs to avoid and recover from errors in the DNA channel. The following is a brief comparison of some of the most common ECC schemes:

- Parity codes work by adding an extra bit to a block of data so that the total number of 1's in the block is always even or odd, depending on the type of parity used.
    - Can detect but not correct single bit errors.
    - Does not work well with large blocks of data.
    - Linear complexity for generating and checking.

- CRC (Cyclic Redundancy Check) codes generate a checksum based on the data being transmitted and appends that checksum to the data.
    - Can detect single- and multi-bit errors but cannot correct errors.
    - Linear complexity for generating and checking.

- Hamming codes work by adding extra parity bits to the data being transmitted. The number of parity bits added depends on the number of data bits being transmitted.
    - Can detect 1-bit and 2-bit errors. Can correct 1-bit errors.
    - Linear complexity for generating quadratic complexity for checking.

- LDPC (Low Density Parity Check) works by solving a set of linear equations to correct errors in the data. LDPC is considered the state of the art when error rates are significantly high.
    - Can detect and correct single- and multi-bit errors.
    - Linear complexity for generating quadratic complexity for checking.

- Erasure codes work by adding redundant information to a datum that allows the receiver to reconstruct the datum even if parts of it are missing.
    - Can detect and correct errors, including the loss of entire data blocks.
    - No closed form for complexity (either for generating or checking).
    - Increases in datum size.

- Fountain codes work by using a randomized algorithm to generate an infinite stream of encoded packets from a single source packet. The receiver can reconstruct the original data by collecting enough of these encoded packets.
    - Can detect and correct errors, including the loss of entire data blocks.
    - No closed form for complexity (either for generating or checking).
    - Less efficient than other error correction codes requiring more storage to achieve the same level of error correction.

- Viterbi codes work by generating a series of encoded bits from the original data stream using a shift register and feedback. The receiver uses a technique called maximum likelihood decoding, which compares the received signal with all possible transmitted sequences to determine the most likely sequence.
    - Can detect and correct errors.

2.2.2 Concatenated Error Correction Codes

Multiple error correcting codes can be combined to utilize their respective strengths. In a process of concatenation (see Figure 5), concatenated codes[11] form a class of error-correcting codes that that are derived by combining an inner code and an outer code.

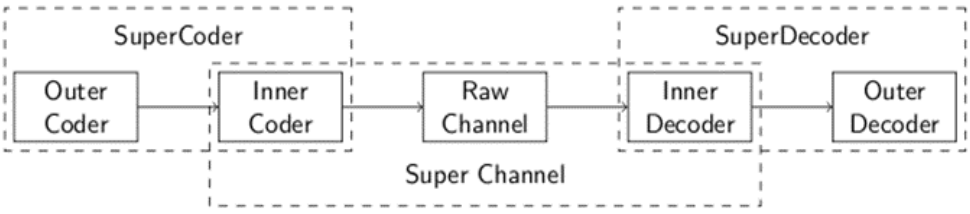

Figure 5: Schematic depiction of a concatenated code built upon an inner code and an outer code. Reprinted from [11].

Generally, the inner code operates on a smaller block of data, usually referred to as a "subblock" or "inner codeword". The primary responsibility of the inner code is to correct errors within this smaller block. The inner code generally offers stronger error correction capabilities and is capable of handling more significant error rates.

The outer code operates on a larger block comprising multiple subblocks generated by the inner code. This larger block is often referred to as a "superblock" or "outer codeword." The outer code is responsible for correcting residual errors that might not have been corrected by the inner code as well as any new errors that occurred during storage.

The concatenation approach has been applied to DNA storage where an inner Reed Solomon code corrects substitutions, and an outer Fountain code corrects all other error types[12]. The technique can be extended to 3 or more codes, at the cost of accommodating the computing overhead involved.

2.2.3 Error Model Considerations Specific to DNA-Based Data Storage Channels

2.2.3.1 Levenshtein Distance

In traditional data storage and transmission channels, codecs use Hamming distance for error correction and detection. For a DNA-based data storage channel, however, Levenshtein distance[13] (i.e., the minimum number of insertions, deletions and substitutions required to change one sequence to another) is a more typical metric. Unlike Hamming distance (an O(n) algorithm), calculating Levenshtein distance is a dynamic algorithm with time complexity O(n1xn2), where n1 and n2 are the lengths of the DNA molecules being compared. If both are the same length, n, then Levenshtein distance is an $O(n^2)$ algorithm.

2.2.3.2 Clustering-Correcting Codes

In a DNA-based data storage system, many copies of the same sequence are generated during the various pipeline phases; in this process, some of the molecules will have errors. In this case, when reading or selecting a group of these molecules, there is a need to establish a consensus sequence from this group of error-containing copies; *cluster analysis* or *clustering*, which aims to separate objects in set intro groups with similar members, is often employed to do this. Cluster analysis employs a *distance* metric as described in section 2.2.3.1. Two objects are considered similar if the distance between them is small (or zero).

Let’s take an example where the set of molecules in a DNA pool are identified by molecular tags for packetization or object identification (Section 2.4) and let’s assume that, due to PCR or other copying processes in the pipeline, there are many copies of each molecule in the pool. If none of the molecules in the pool contained errors, then we could expect that molecules with the same tag are in fact copies and carry the same payload (i.e., the portion(s) of the molecules bearing data). Errors in the index of a molecule, however, could result in molecules which are not in fact copies of each other (i.e., have a different payload) being identified as copies. This could lead to inaccuracies in recovering the payloads of molecules with a given tag.

In this example, clustering-correcting codes can ensure that if the distance between the tag fields of any two data-bearing DNA molecules is small, the distance between their payloads is large. This allows the clustering algorithm, during decoding, to accurately detect a miscategorized molecule (i.e., one associated with the wrong index due to a small index error, but identifiable as not correctly being associated to this index because the payload for this molecule is far distant from the other payloads in the group) and either discard it or place it in the correct group[14] Ensuring that all copies of molecules get into the same group, even if there are some tag errors, ensures that the correct consensus payload strand can be derived from each tag group, which in turn ensures that information can be successfully retrieved and decoded from an archive.

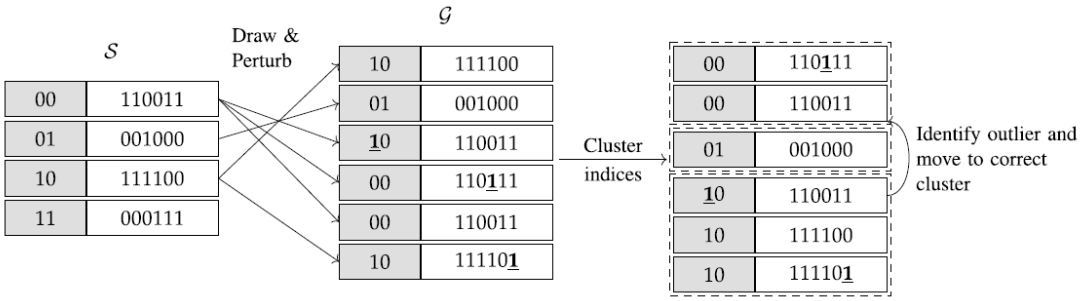

Fig. 2. Exemplary realization of the DNA channel model. A set $\mathcal{S}$ of $M = 4$ binary strands is stored and $\mathcal{R} = 6$ strands are drawn with errors (highlighted in bold). The strands are clustered according to their indices. The outlier can be identified as it has large distance w.r.t. all other strands in the cluster and be put into the correct cluster.

Figure 6: Example of Clustering-Correct codes example. Reprinted from [14].

There are many potential use cases for clustering-correcting codes. The central point is that, due to the nature of DNA data storage, which commonly creates many copies of the same molecule, clustering-correcting codes are a natural tool to enable various storage operations requiring consensus amongst those copies.

2.3 DNA Sequence-Constraints

Various DNA sequence patterns can elevate DNA synthesis and sequencing error rates; therefore, constraints are implemented in DNA codecs to avoid or minimize these patterns and thus reduce the error rate across all workflow steps. In addition, certain DNA sequences may be prohibited due to regulatory constraints, such as biosafety. There are two main categories of constraints: 1) Channel limitations; and 2) Biosafety and biosecurity considerations. We cover channel limitations here. Biosafety and biosecurity applications, which are common across all phases of the DNA data storage pipeline, are covered in section 6.4.

2.3.1 Channel Limitations

A DNA data channel comprises synthesis, storage, retrieval and sequencing. Codecs translate digital data into DNA sequences, which are used as input to DNA synthesis, and convert DNA sequences recovered during DNA sequencing back into digital data. The limitations of the components of the data channel are an important consideration of the codec, as follows:

- Synthesis introduces mainly deletion errors, and fewer insertions as well as substitutions.
- Long-term storage may lead to substitution errors, due to cytosine deamination, but in general the error rate during storage is very low, and not sequence dependent.

- Sequencing primarily introduces substitution errors, but very few insertions or deletions.

The Codec constraints deriving from channel limitations for DNA data storage are summarized in the following three sections.

2.3.1.1 Local constraints

Local constraints derive from concerns about the effects of a sequence of nucleotides in a localized area of a DNA molecule, such as homopolymers, secondary structures and catalytic sequences. In addition, some enzymes have difficulty synthesizing specific sequences. Of these local constraints, homopolymers and secondary structure are the most relevant. Homopolymers are discrete sequences where the same base is repeated. For example, a sequence 5'-AAAA-3' is an adenosine homopolymer of length 4. As discussed in their respective chapters, DNA write and read technologies have difficulties with these sequences.

2.3.1.2 Global Constraints

Global constraints concern sequence effects at the whole molecule level, where particular sequences can lead to the formation of secondary structures in DNA molecules, such as hairpin loops and bulges, interfering with both replication and sequencing. For example, while GC-rich regions are more stable than AT-rich regions (GC base pairs form three hydrogen bonds while the AT base pairs only form two), they are more difficult to synthesize because the G and C nucleotides are larger than the A and T nucleotides. Finally, as noted above, high GC content can lead to the formation of secondary structures.[15]

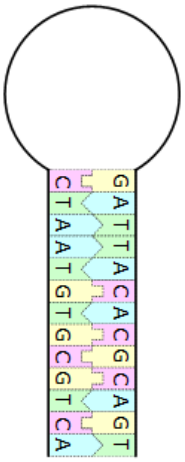

Figure 7 - A DNA hairpin structure.

2.3.1.3 Constraints on pools

Pool constraints are constraints on the total pool of molecules generated, such as data sizes allocated to single address spaces and the need for mutually uncorrelated encoding.

Within a large pool, it is important that there is limited sequence overlap between any two sequences, as this can lead to errors in random access recovery, read and decoding steps. Mutually uncorrelated encoding addresses these issues through sequence randomization. At any position within a sequence, there is a ~25% chance of any base occurring.[16,17,18,19]

2.3.2 Biosafety and biosecurity

Biosafety and biosecurity are broad topics which cross many aspects of the DNA data storage pipeline. See Section 6.4 for a discussion of this topic.

2.4 Data Storage Protocols

As shown in the example in Figure 2, DNA codecs must not only encode bits to bases, but also embed protocol to enable the DNA storage channel. This goes beyond pure "channel protocol", such as ECC and sequence constraints, to logical protocols for implementing storage operations. Some of these protocols are described here; others are touched on in Section 4.2 (e.g., similarity

search). This is a very active area of research and the coupling between the underlying chemistry and the DNA codec is, in the current state of the art, very tight.

2.4.1 Packetization

With even the highest projections for the ability to synthesize long DNA molecules, it is unlikely that a large digital object (GB, TB, or larger) can be encoded as a single DNA strand. Thus, as in network transmission, large digital objects need to be packetized into many small segments of DNA, which means these packets will require indices encoded within them to enable correct reassembly after storage. This is a common function of DNA codecs.

2.4.2 Archive identity and structure discovery

One of the strengths of DNA as an archival storage medium is that the underlying structure of DNA is universal; therefore, the raw sequence of bases in the molecules in a DNA archive will always be readable by the DNA reading technology available at the time an archive is being read. However, as with any encoded storage medium, the information encoded in a DNA archive can only be recovered if the codec originally used to encode the archive and, to a less vital but still important extent, the logical structure of the archive, can be discovered. And it is further desirable that any DNA codec can, in a standard way, do this discovery. The DNA Data Storage Alliance's Sector 0 and Sector 1 specifications[20] are an initial example of such a standard.

2.4.3 Object Tagging / Random Access

As shown in Figure 2, Object Tags are used to identify and differentiate digital objects, encoded as DNA molecules, within a molecular archive. Object tags are DNA sequences that are generated during encoding and written at predefined locations (typically at the distal ends) of molecules during synthesis. Object tags act as handles for random access (selective recovery), as discussed in the Storage and Retrieval chapter. These sequences are also used in decoding to assist in identifying payload sequences from the same object. Encoding algorithms attempt to maximize the distance between any two different Object Tags so that error generated during the various pipeline phases do not prohibit successful retrieval and decoding of desired data. This is complicated by sequence constraints discussed in section 2.3 above and mitigated by error correction approaches discussed in section 2.2.

2.5 Simulators

Codecs require an accurate representation of the error model of the information channel to recover data during decoding. The rate of synthesis and sequencing errors may vary an order of magnitude from one platform to another. Therefore, to ensure accurate reconstruction, Codecs must either be tuned for a particular platform, or they must account for the worst-case scenario. In either case, extensive write-read-rewrite experiments may be needed to estimate the error rates before encoding the error model in the molecular data stream. .[21] Simulators can be used to simulate various combinations of approaches to DNA storage, provide error models and test the ability of Codecs to recover data. A couple of popular simulators are described below.

2.5.1 DNA Storalator

The DNA-Storalator is a cross-platform software tool that simulates all the processes involved in DNA storage[22]. The tool consists of three main modules. The first is the error-simulator that receives encoded DNA sequences as input and outputs simulated noisy reads. (This module assumes that binary data was encoded into DNA sequences and simulates the synthesis, PCR and sequencing processes. The reads are generated by the simulator according to predefined error rates, that are based on analysis that was done on data from previous experiments, while the tool also supports user-defined error rates. The second module simulates the clustering, which is the process of partitioning the noisy reads into groups based on their original encoded strand. To perform this process, the DNA-Storalator has several algorithms, some of them are new suggested algorithms for this purpose, and some are implementations of previously published methods. The last module of the Storalator is reconstruction, which is the process of estimating the original designed strand from its noisy reads. The DNA-Storalator has several reconstruction algorithms that work on different complexities to solve this problem on various ranges of error rates. The main and the most direct use case of the DNA-Storalator is the design of new error-correcting codes for DNA storage: users can use the DNA-Storalator to simulate the process involved in DNA storage and test the performance and efficiency of the tool.

2.5.2 DNAssim (Avaneidi)

The simulation of the DNA storage pipeline involves different stages that should be modeled by a software framework capable of capturing the peculiarities of the information encoding and decoding (Figure 9).

Because of the number and complexity of the steps involved in the DNA storing process, the number of simulations is huge, and a “pure software” simulator can run very slowly. To overcome this limitation, DNAssim is built on a custom co-simulation (i.e. mix of hardware and software) platform.

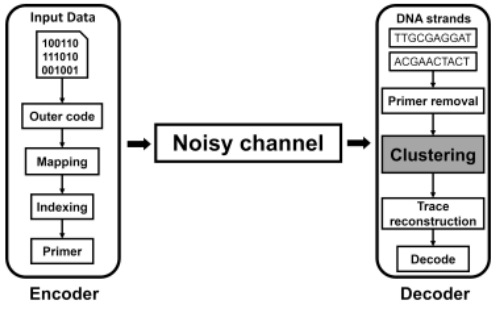

Figure 8: High-level description of the proposed DNAssim software simulation engine; courtesy Avaneidi.

PCIe-attached hardware is used to accelerate simulations. FPGAs (Field Programmable Gate Arrays) and GPUs (Graphics Processing Units) are great examples of high-performance cards using the PCIe interface to communicate to the host PC/workstation/server. FPGAs and GPUs can co-exist in the same platform, and they can both be instantiated multiple times. The choice of FPGA or GPU depends on the specific computing function that needs to be accelerated. [23,24,25,26]

2.6 CODEC Attributes and Metrics

2.6.1 Information Density (bits/nt)

The most essential metric for a DNA codec is information density (ID), the amount of information encoded at each position on the DNA strand, measured in bits/nt. The theoretical maximum of

data that can be stored per base is described by $ID = log_2(N)$; where N is the number of available characters and is measured in bits/base. With binary coding, the maximum is 2 bits / nt. With ternary codes, the maximum is ~1.58. These maximums assume that at any position within a sequence any available character can be used to encode data. The constraints described above, including constraining GC content and homopolymers to tailor the codec to synthesis and sequencing, can limit the characters used at any position, thus limiting ID. In addition, any positions within the sequence used for error correction (or random access, addressing, indexing, etc.) prevent encoding data and are considered overhead, thus lowering the ID.

2.6.2 Throughput & Latency

Throughput is defined as the bits encoded or decoded per unit time and is differentiated from latency, which is the time to the first encoded or decoded bit. These are important metrics in that they contribute to the overall flow of information through the information channel. For encoding, both throughput and latency are limited by constraints put on sequences and the approach and amount of error correction to be included in the sequences. For decoding, the throughput and latency are limited by the decoding algorithm and the overall error rate in the channel.

2.6.3 Power

Power consumption is defined as the number of watt hours consumed when performing encoding or decoding operations. This metric is important because it directly contributes to the total cost of ownership for operating a DNA Data Storage system. A significant component of the power requirements is determined by the computational hardware chosen to perform the operations. To minimize this cost, encoding and decoding software can be optimized to minimize computation cycles, memory usage, and communications with peripherals. For further discussion, please see the DNA Data Storage Alliance's white paper, *DNA Data Storage Codecs: Examples, Requirements, and Metrics*.

2.6.4 Biosecurity and biosafety

Biosafety and biosecurity are broad topics which cross many aspects of the DNA data storage pipeline. See Section 6.4 for a discussion of this topic.

2.6.5 Error recovery capability

Error recovery capability refers to the decoder's ability to correct errors incurred in the DNA information channel. This metric is important as it provides security that information stored and passed through the channel can be recovered, especially after very long storage periods. As discussed in the storage chapter, several studies have shown that the error correction techniques discussed in this chapter can successfully recover data over long storage periods in the presence of significant error rates. This capability will influence power consumption and total cost of ownership as error correction algorithms can be computationally expensive.

3 Synthesis

3.1 State of the Art

In general, today, DNA molecules are constructed either by (1) Base-by-Base (often called *de novo* synthesis) where a custom strand of DNA is built, one base at a time, or (2) DNA Assembly, where pre-existing DNA strands are assembled into longer DNA strands in a building block approach, using various molecular mechanisms such as ligation. DNA Assembly relies on some version of Base-by-Base in order to create the initial building blocks that are then assembled, but once the initial building blocks are created, the DNA Assembly method enables different efficiencies and tradeoffs than Base-by-Base.

3.1.1 Base-by-Base Synthesis

For base-by-base synthesis, two fundamental types of chemistries are used: 1) chemical and 2) enzymatic, both of which are employed by Twist Bioscience. Both of these employ a similar, cyclic process to add bases to an anchored strand.

3.1.1.1 Chemical DNA Synthesis

The chemical synthesis of single stranded DNA (ssDNA) was published as far back as 1965[27,28] and first commercialized in 1983, and is today the dominant source of synthetic DNA, driving a ~$7.7B/year market[29]. As shown in Figure 9, starting from a solid support, the process involves repetitive cycles of single base addition, oxidation, and 5'-deprotection, all occurring in the presence of organic solvents[30]. Given a desired sequence determined by the codec, the DNA synthesis process begins by flowing in a solution of the first desired base. The 5' position of the base is "protected" chemically, which prevents multiple bases from being added. Once added, the protecting group is removed, and the cycle can repeat again. The cycle is repeated for every desired base in the sequence until the entire sequence has been written. Once the desired sequence has been written, the strands are released from the support and purified. The stepwise cycle efficiency is often quoted as >99.5%, which means that errors accumulate as a function of strand length.

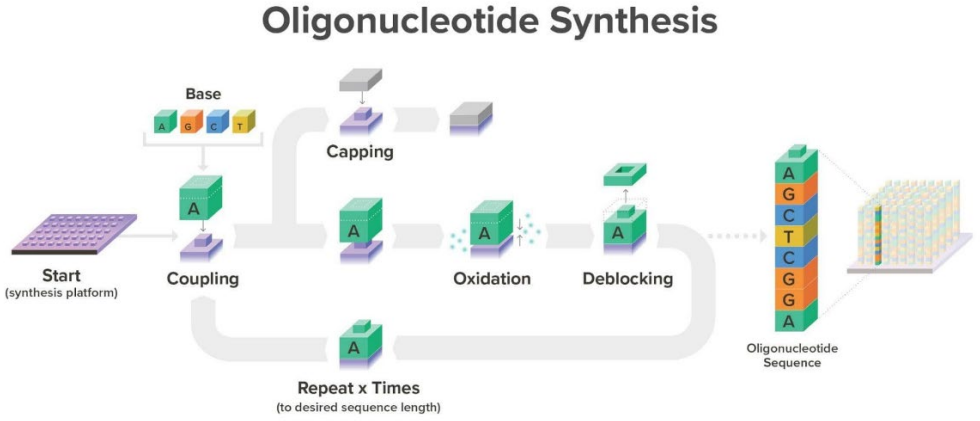

Figure 9 - Chemical DNA synthesis chemistry and workflow. (Image courtesy Twist Bioscience)

The stepwise efficiency typically limits the length of chemically synthesized oligonucleotides to fewer than 300 bases, though Twist Bioscience has reported extending this to 500 bases.

3.1.1.2 Enzymatic DNA Synthesis

Enzymatic synthesis of DNA is a method of producing single-stranded DNA (ssDNA) strands in a completely aqueous process (i.e., no oil-based solvents), using either a template-independent polymerase or a ligase (Figure 10). Analogous to the chemical synthesis method, the enzymatic synthesis process consists of repetitive cycles of enzymatic incorporation/ligation, followed by either chemical or enzymatic deprotection[31,32]. In this process, the base addition step is facilitated by an enzyme, rather than organic chemistry. The aqueous reaction conditions and the use of enzymatic addition improve the cycle efficiency, thus increasing the lengths of ssDNA that can be written error free.

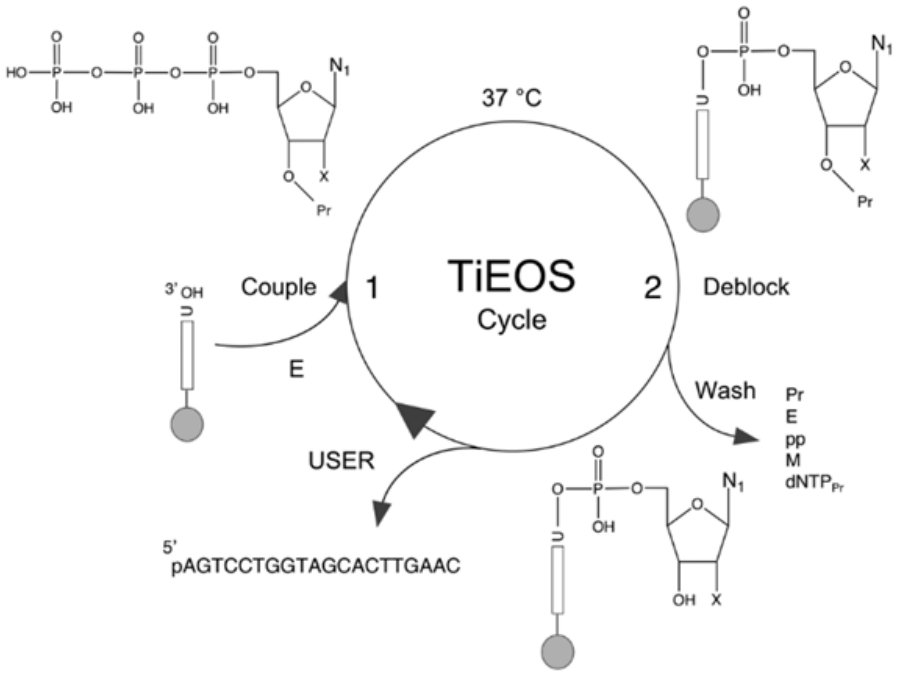

Figure 10 - Template Independent Enzymatic Oligonucleotide Synthesis (TiEOS). Reprinted from [31].

Another approach to enzymatic synthesis demonstrated by Ansa Biotechnologies uses the template independent polymerase molecule itself to act as the protecting group, by using a covalent complex of the base and the enzyme[33] (Figure 11). After each incorporation, the “protected” strand must be readied for the next cycle by removal of the tethered blocking enzyme molecule. Ansa claims economy in consumption of the enzyme-based reagent, but this could be off-set by the need to use an excess of reagent and its potentially higher cost.

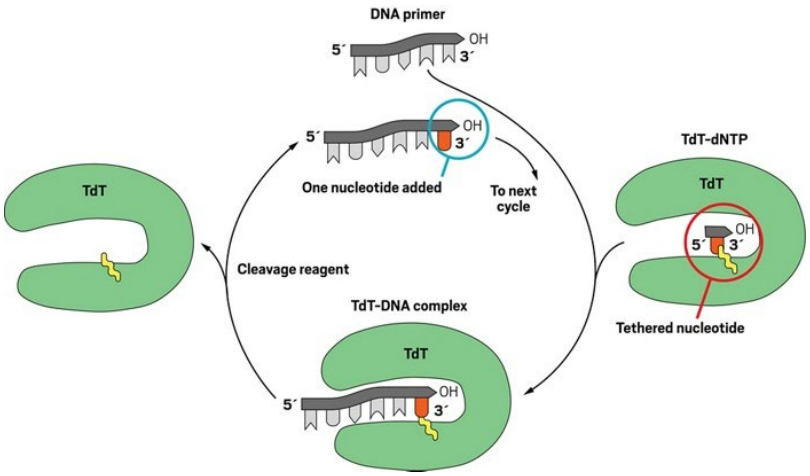

Figure 11 - Cleavable dNTP-TdT conjugates. Reprinted from [34].

A third approach to enzymatic synthesis, used by Kern Systems and Molecular Assemblies, uses unprotected bases, which neatly avoids the use of a deprotection step in each cycle[35,36] (Figure 12). The use of a free-running polymerase extension results in short homopolymers being synthesized, where the transition between homopolymers encodes bits, instead of using a single nucleotide to encode a bit. The length of the homopolymer is unimportant and need not be controlled too accurately, as long as the sequence detection method can accurately identify the presence of a homopolymer and the transition to the next homopolymer block. Since the bits are variable in length and accurate length determination is not required, any sequencing technology is suitable for readout, including nanopore readout. Homopolymer bit encoding is no more reagent consumptive than those using deprotection approaches, since template independent polymerases readily and rapidly form homopolymers in the presence of excess bases, which would be required in either enzymatic approach.

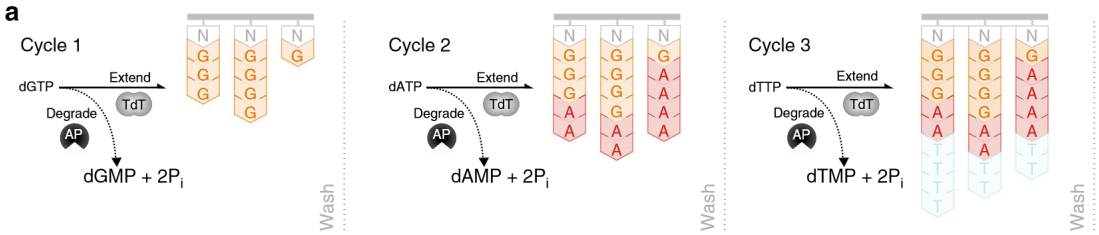

Figure 12 - Free running polymerase extension Reprinted from [35].

3.1.2 DNA Assembly

In contrast to *de novo* synthesis, DNA strands encoded with digital data can be synthesized from prefabricated oligonucleotides (oligos) or double-stranded DNA blocks. Catalog Technologies has described using prefabricated oligos to write digital information in double stranded nucleic acid molecules by encoding bit-value information with the presence or absence of unique nucleic acid sequences within a pool[37] (Figure 13). Biomemory utilizes bio-sourced double-stranded DNA blocks as their prefabricated building blocks to assemble long double stranded nucleic acid molecules. Unique bits in a bit stream are encoded by unique subsets of a nucleic acid sequence. In both these approaches, the prefabricated bits are generated either from base-by-base synthesis or isolation from natural DNA molecules and stored as pools that are then assembled

into a bitstream at the time of writing. The bitstream can be assembled using overlap - extension polymerase chain reaction, polymerase chain reaction (PCR), polymerase cycling assembly (PCA), sticky end ligation, BIOBRICKS® assembly, Golden Gate assembly, GIBSON ASSEMBLY®, recombinase assembly, ligase cycling reaction, or template directed ligation.

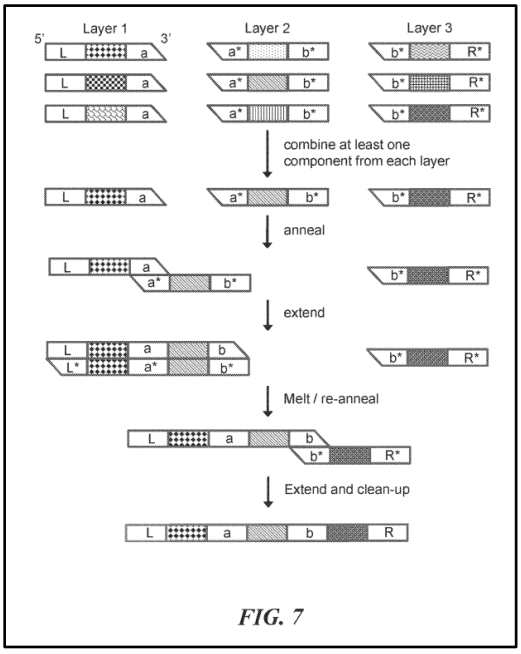

Figure 13 - Oligo Assembly Approach. Reprinted from [37].

3.1.3 Structure based encoding

As a contrast to writing bits in linear strands where bits are encoded by the sequence of bases in the strand, there are several structure-based approaches where bits are encoded in the physical structure of the molecules. Comprehensive reviews of these techniques can be found in the literature. Below are two specific examples of emerging approaches that highlight the utility of a DNA-based nanotechnology called “DNA origami”.

In the first (Figure 14), the above referenced oligo synthesis approaches are used to synthesize short strands that both encode digital data and are complementary to subsections of a large single strand of non-information encoding DNA. These short strands (staples) fold the larger strand (scaffold) into a unique three-dimensional shape with site-specific localization of digital information. Advanced microscopy techniques (i.e., DNA PAINT) can be used to read information back out of these structures.[38]

In the second (Figure 15), three dimensional structures are assembled onto a linear single stranded DNA backbone. In this case, the above referenced oligo synthesis approaches are used to synthesize short strands that fold into complex 3D shapes when hybridized to the backbone, which is also synthesized using the methods described above. This approach has many of the advantages of oligo assembly, because designs rely on a finite set of discrete sequences that greatly reduce synthesis overhead. This approach comes at the cost of information density as

several bases are required to encode a single bit and, currently, requires nanopore sequencing approaches to “read” encoded features.[39]

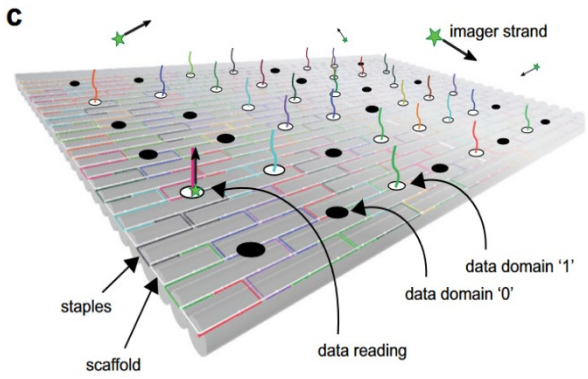

Figure 14 - Digital nucleic acid memory (dNAM). Reprinted from [38].

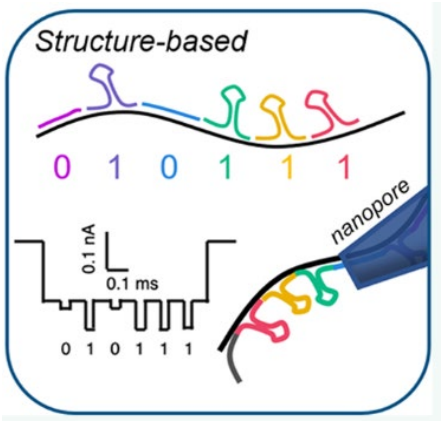

Figure 15 - Structure based encoding. Reprinted from [39].

3.1.4 Synthesis Modalities

Several implementation modalities have been used to control the synthesis chemistries discussed above. Each of these approaches increases the throughput by synthesizing many strands in parallel. In each instance, synthesis happens at the nanoscale with many copies of the same sequence synthesized at each location within an array. Although there may not be demonstrations in publicly accessible literature for each combination[40], it is probable that each of the modalities listed here could be applied to any of the chemistries. A demonstration is provided by the University of Washington and Microsoft Research[41] (Figure 16) who demonstrated chemical DNA synthesis within nanoscale wells on a semiconductor substrate. This modality has been extended to enzymatic synthesis in collaboration with Ansa Biotechnologies[42].

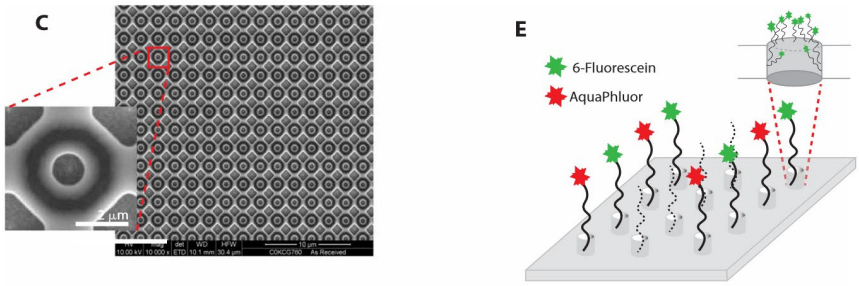

Figure 16 - Electrochemical DNA synthesis on a nanoscale array. (c) An overview of the nanoscale DNA synthesis array with scanning electron microscopy images of the 650-nm electrode array and enlarged view of one electrode. (e) Illustration of the wells patterned with ssDNA oligos with multiple copies of each oligo per synthesis location. Reprinted from [41].

Other modalities include photo-deprotection and inkjet printing[39] (Figure 17). Each of these approaches have unique capabilities and error profiles[43,44,45]. In regard to DNA data storage, the main takeaway is that these synthesis modalities enable high-density DNA writing and are being investigated to address challenges associated with higher synthesis capacity and throughput, while driving down costs.

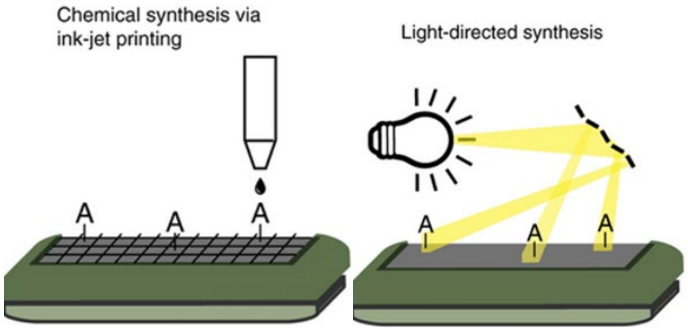

Figure 17 - Ink-jet printing and light directed DNA synthesis modalities. Reprinted from [39].

3.2 Attributes and Metrics

3.2.1 Synthesis errors

As seen in Table 2 (Section 2.2), an insertion occurs when an unintended base is inserted in the written sequence between two intended bases. A deletion occurs when an intended base is not added to the written sequence. A substitution occurs when an intended base is replaced by a different base in the written sequence. Depending on the location and frequency of errors, this could lead to a written sequence that cannot be read by a sequencing instrument, resulting in a fourth error type: erasures.

There are several factors that contribute to the creation of synthesis errors; these factors are cumulative, meaning that they can combine to increase the overall synthesis error rate. It is not unusual to observe aggregate error rates at ~0.01 errors / base[46].

While a comprehensive accounting of all factors is outside the scope of this roadmap, the most common synthesis errors stem from issues in removing the “protecting” group, making insertions and deletions the prevalent error types. Other contributors are specific to the type of synthesis (i.e., chemical or enzymatic) and the synthesis modality (e.g., pH, electrochemistry, photon/UV, etc.) and how they are integrated into a synthesis instrument and process. As methods to increase synthesis parallelism progress, reduced physical dimensions can lead to more error prone synthesis processes that produce fewer copies of each sequence[Error! Bookmark not defined.]. There will be a continual balancing of density and level of ECC that will be manifested in compute or other cost/performance tradeoffs.

3.2.2 Throughput & Latency

Synthesis throughput is measured as the amount of data that can be written per unit time (e.g., bytes/day). Since the low level operations involved in building DNA molecules are inherently slow (especially when compared to analogous operations in traditional storage media), the challenge of DNA synthesis throughput will be met through parallelism, as discussed in section 0.

Synthesis latency is defined as the time required to write the first byte. In a DNA Data Storage (DDS) context, write latency includes the time required to setup and initialize the instrument and to import sequence data from the encoding algorithm, in addition to the time required for the

chemistry to add the requisite number of bases to encode the first byte. This can be on the order of seconds to minutes. This compares well to incumbent media in the archival storage tier.

Table 3 summarizes the current state of DNA synthesis throughput and latency in comparison to today's storage media based on the synthesis modalities listed in 3.1.4.

Media	Typical Write Latency (time to first byte)	Typical Write Throughput
DNA Data Storage	Seconds to minutes	~100 MB/day = 0.001 MB/s
Tape	Seconds to minutes	~400 MB/s (uncompressed)
HDD	Tens of milliseconds	~300 MB/s
NVMe SSD	Hundreds of microseconds	~1000 MB/s

Table 3 - Latency and throughput associated with various storage media

As discussed above, array synthesis is the most common approach to parallelism, and various approaches are being used.[47,48,49] As seen in a technical analysis conducted by IARPA (Figure 18), in 2018 the projected state of the art was approximately 1 million (M) synthesis spots per substrate, with 100M expected in 2022. Twist Bioscience stated in 2022 that they had developed a chip which has a capacity of 1GB per run with on the order of (100M) synthesis spots[50], roughly tracking the projection. Systems with on the order of 100M synthesis spots should be capable of meeting minimum throughput requirements and by 2027, the IARPA projection anticipates there will be arrays approaching 100B unique spots, which could potentially achieve commercially relevant throughput. At present, these devices are approaching the limits of device area and feature size, and multichip systems may be required for continued scaling. Given limitations of photolithographic and on-demand deposition, the scaling and production roadmap of electronic devices is enticing and both academic and industry groups are pursuing this approach.

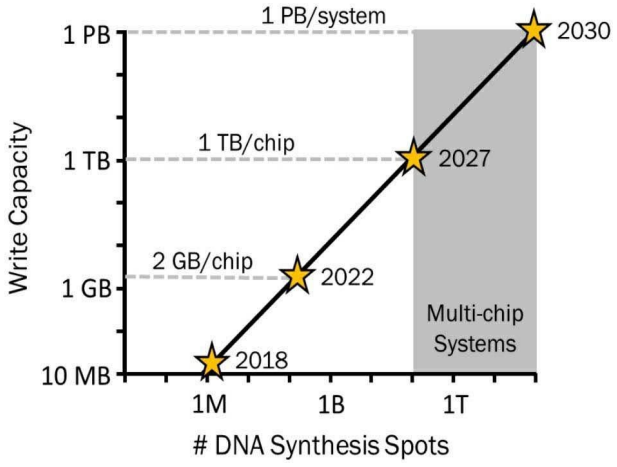

Figure 18 - 2022 IARPA Roadmap for DNA synthesis. Assumes ssDNA, 150 nt in length (20 nt flanking primers), encoded at 1 bit/nt. Courtesy David M. Markowitz, D. SRC/IARPA Workshop on DNA-based Massive Information Storage. (2016).

3.2.3 Environmental Impact & Sustainability

As described in the introduction to this chapter, there are two main categories of base-by-base DNA synthesis, chemical and enzymatic, each with their own unique sustainability profiles. The environmental impact and sustainability of synthesis systems are measured in the amount of greenhouse gas emissions, energy consumption and resource consumption (e.g., fresh water) per byte of data synthesized.

Chemical synthesis relies heavily on organic reagents, such as acetonitrile, which can be volatile, flammable and toxic. Phosphoramidites require complex chemical synthesis to produce and have waste streams similar to those of acetonitrile. Reagent reuse in chemical synthesis is difficult[51].

Enzymatic synthesis employs either native or lightly modified nucleotides that require less involved synthesis protocols with cleaner waste streams. The enzymes themselves are commonly manufactured in large-scale fermentation of engineered microorganisms and are not considered biohazardous waste. Enzymatic synthesis requires fewer steps and is performed predominantly in aqueous salt buffers at physiologic pH, which avoids many of the hazards associated with chemical synthesis. As such enzymatic synthesis creates waste that can be disposed of through drains rendering it compatible with data center infrastructure. No studies have explored the environmental impact and sustainability of DNA assembly methods using e.g. ligation of pre-synthesized DNA blocks. The assembly process being enzymatic, its impact should be comparable to enzymatic synthesis. The method used to produce the DNA blocks may further impact the environmental footprint of DNA assembly, depending on whether the DNA is synthesized chemically, enzymatically or produced biologically. In the latter case, used by Biomemory, the environmental impact is expected to be significantly lower than enzymatic synthesis.

A recent study from Microsoft[52] has shown that enzymatic DNA synthesis has the potential for significant reductions in greenhouse gas (GHG) emissions, energy and water consumption as compared to chemical DNA synthesis (Figure 18 & Figure 19).

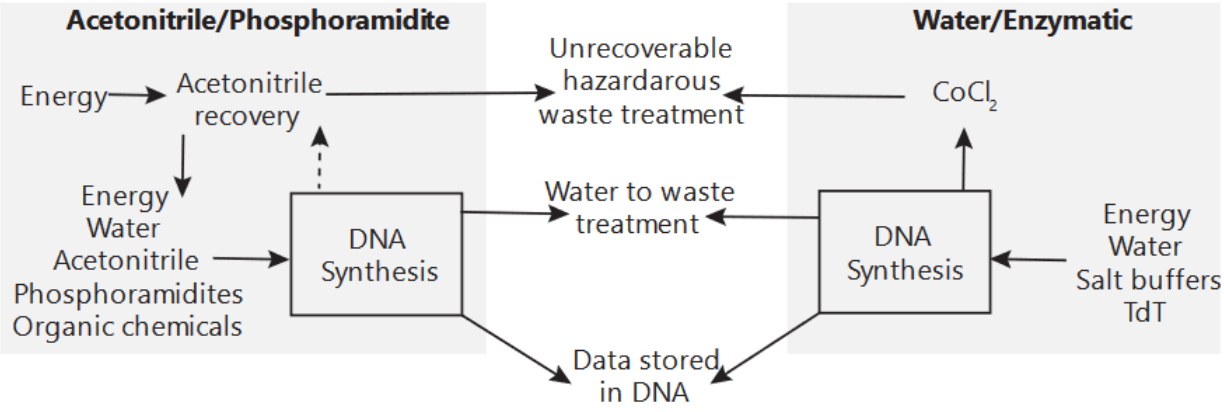

Figure 19 - Resource consumption in DNA synthesis [52]

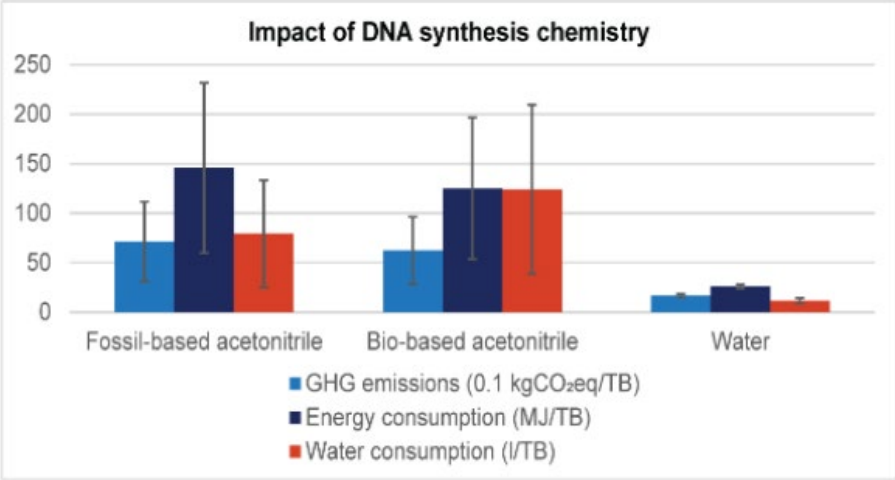

Figure 20 - Sustainability impact of DNA synthesis chemistries. Reprinted from [52].

3.2.4 Biosecurity/Biosafety

Biosafety and biosecurity are broad topics which cross many aspects of the DNA data storage pipeline. See Section 6.4 for a discussion of this topic.

4 Storage and Retrieval

4.1 State of the art - Storage

DNA as a molecule has been shown in nature to be very stable. The most recent record for decoding environmental DNA was set in 2022, where the genetic code of a new family of mammals was decoded from fossil DNA that was over 2M years old[53] .

However, to store DNA that contains encoded digital data, a manufactured DNA Containment System (DCS) is required. The characteristics of a DCS fall in two broad categories: 1) The physical form and arrangement of the containment vessels (sealed/unsealed, whole archive in a single vessel vs. archive sub-segments distributed in multiple vessels, etc.); and 2) the preservation steps (additives, drying, sealing, inert gasses, etc.) used to prepare and store the media in the vessels within the DCS.

The design choices made for a particular DCS, from physical configuration to complexity of the preservation method, are driven by user needs; longer term (e.g. century level) storage may be more expensive and higher latency, while shorter term (e.g., decades or even less, but needing density and low TCO over the shorter time) storage may offer lower costs and less usage complexity.

4.1.1 Physical manifestation of the containment vessels in a DCS

In the predominant forms of encoding digital information in DNA molecules today, individual encoded strands can store only on the order of tens of bytes, since longer strands are difficult to synthesize without unacceptable error rates (i.e., cost). It is nearly assured that no matter how robust synthesis techniques get, digital objects such as files or photos will have to be encoded into DNA sub-strands, which will thus have to encode indices for reconstruction at decode time, object tags for file selection, etc.

Physically, the most straightforward instantiation of a DNA archive is in a single pool, regardless of the object hierarchy within the pool. However, one can consider using multiple pools, based on some part of the object schema of the global pool. In this multiple pool case, the containment vessels which make up a DCS could be used to hold physical sub-segments of the whole archive. To say it another way, a DNA archive could be stored as one large pool in a single vessel or be subdivided into multiple pools (or groups of pools) stored in multiple vessels. Any grouping of pools to vessels is possible, depending on the characteristics of the DCS. In this example, the physical structure of the DCS is used to map some part of the object address space within the archive.

Physical sub-division in a DCS sacrifices space efficiency but it can avoid other complexity in coding/decoding and the chemistry of the pipeline. Total cost can be more or less, depending on the preservation processes and vessel handling needed for any particular DCS in relation to the use case (time of storage, frequency of access, etc.). Pool size is generally limited by the

characteristics of the DNA random access method (address space and its capacity to address files in parallel) combined with the user's needs for parallel random access.

4.1.2 Protecting DNA Media in a DCS

Regardless of the physical DCS instantiation considerations noted above, the primary purpose of a DCS is to protect the stored media from damage. DNA damage can occur at a variety of locations on a DNA molecule (Figure 21) and there is an extensive literature documenting this[54,55,56]. Most commonly, DNA degradation results in damage to a single base, where the base is modified or lost, leading to a strand break (i.e. a physical break in the sugar-phosphate backbone). The degradation mechanisms that ultimately result in strand breaks are from exposure of DNA to water, air (e.g., oxygen, ozone, atmospheric pollutants), mechanical shear, and/or ionizing radiation[57,58]; water is by far the most common degradation catalyst. Exposure to UV radiation can also cause new chemical bonds to form within a DNA molecule that can obfuscate the identity of a base or crosslink two strands.

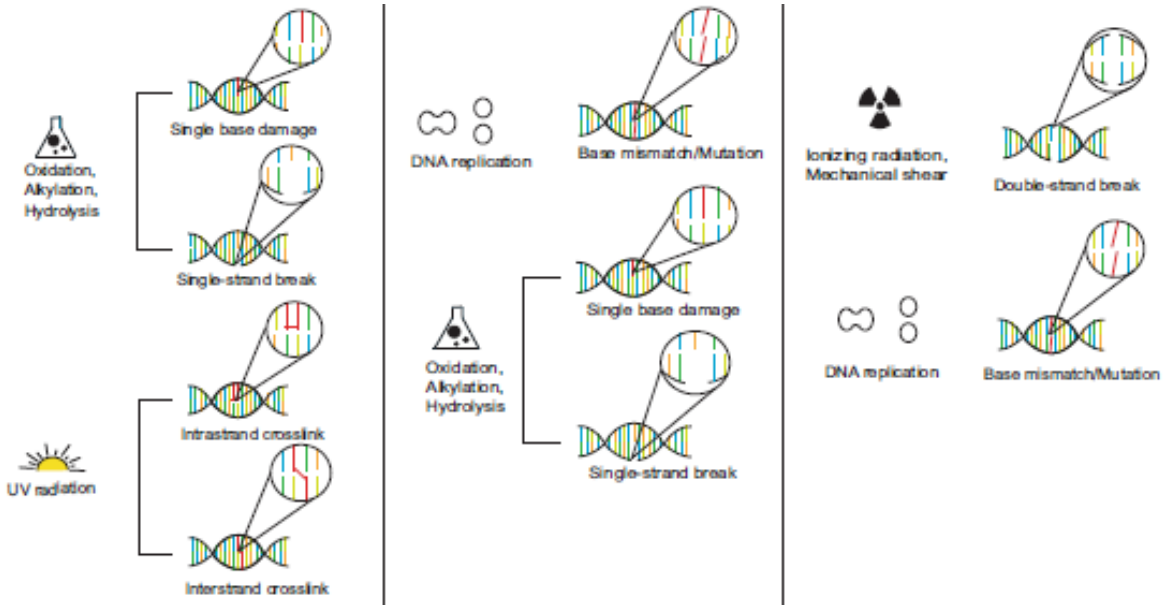

Figure 21 - Potential molecular mechanisms of DNA degradation during storage. Reprinted from [58].

Due to the centrality of water as the most dominant underlying causal factor for these failure mechanisms, the majority of the methods studied for preserving DNA for data storage (i.e., designing a DCS) depend on drying the media and/or completely isolating the media from the external atmosphere, as shown in Table 4.

Preservation Category	Preservation Substrate/Method	Drying	Protection from atmosphere	Stability estimation method
Chemical encapsulation	Encapsulation in salts[59,60,61]	✓		Accelerated aging
	Degradable Polymer[62,63]		✓	Accelerated aging
	Cationic Diblock Copolymer[64]		✓	
	Silica nanoparticles[65,66,67,68,69]	✓	✓	Arrhenius
	Magnetic silica nanoparticles[70]		✓	Accelerated aging
Lyophilization	3D-printed microfluidic chip[71]	✓		
	DNA moveable type blocks[72]			
	Living Memory Microspheroid[73]	✓		3 months at RT
	Storage Platform, Physical Data Partitioning[74]	✓		
	DNA Data Storage in Perl[75]	✓		Accelerated aging
Physical encapsulation	Stainless steel capsules[76,77]	✓	✓	Arrhenius
Inclusion in a matrix	DNAstable[99]	✓		Arrhenius
	Gentegra DNA[99]	✓		
	Pullulan[78]	✓		
	Silk[79]	✓		
	composite nucleic acid-polymer fibers[80]	✓		Accelerated aging
	300K matrix inclusion[81]	✓		
	Hierarchically structured polymeric microparticles[82]	✓		Accelerated aging
Absorption on paper	FTA paper[99]	✓		Arrhenius
	Chitosan treated paper[83]	✓		
Dehydration on solid supports	Glass[84,85]	✓		
	Silicon[86]	✓		
Dissolution in liquid salts	Imidazolium ammonium pyridinium cations[87]			
	Ammonium-Based Ionic Liquid[88]			>1 year at rt
Living organism	yeast genome[89,90]	✓		
	E. coli genome[90,91,92]	✓		
	yeast cells[93]	✓		
	Bacteria[88,91]	✓		
	Bacillus spores[94]	✓		
	Living Memory Microspheroid[73]	✓		3 months at RT
DNA beads	Magnetic Bead Spherical Nucleic Acid[89]	✓		Arrhenius
Storage in long DNA molecules	Experimental DNA storage platform [95]			
	Storage in an Extremophile Genomic DNA[96]			
	Construction, sequencing of long DNA sequences[97]			
	DNA as a universal chemical substrate[98]			

Table 4 - List of DNA preservation methods for storing DNA at rest [Courtesy J. Bonnet & M. Colotte]

Numerous accelerated wear studies have been done to test the molecular stability and data integrity of digital data stored in DNA. The results in general, with some examples shown in Figure 23, show that: 1) additives increase durability of DNA media; and 2) both dehydration and complete isolation can result in a DCS which provides a long half-life for molecules stored at room temperature[99].

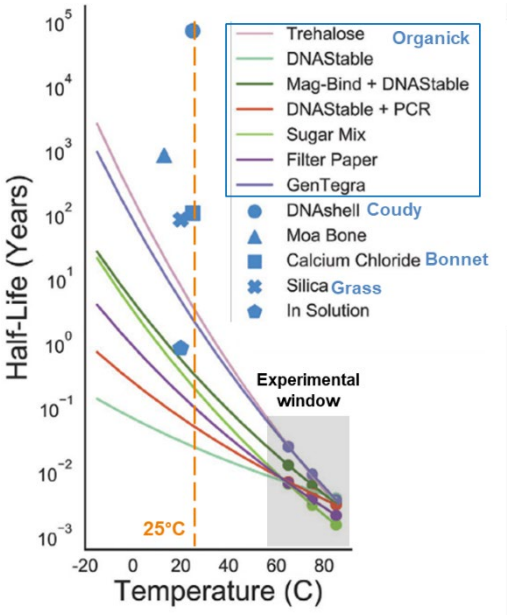

Figure 22 - Half-life plot for oligos of 150 nucleotide length, using various sample preservation methods. Reprinted with modifications from [99].

4.1.3 Interfacing Storage with Synthesis

In the DNA data storage workflow, molecules are transferred from a synthesizer to a stored state. There are several approaches to navigating this interface, most of which require some degree of liquid handling[69,100,101,102,73,103,86,104,105,106,107]. Traditionally these operations have been performed manually, with a human operating a pipette. Increasingly, there is a transition to automated systems designed specifically to manipulate a wide range of fluid volumes, from milliliters to sub-nanoliter. Liquid handling requirements are directly related to the storage container, with a number of emerging technologies employed. Increasingly, microfluidic approaches are used that include both pressure-induced flow and electrowetting approaches.[108,109,110]

4.2 State of the Art - Retrieval

The simplest way to access data stored in a DNA pool is to amplify and sequence the DNA sequences of interest. While this method of bulk retrieval is simple, it is impractical for sufficiently large databases (> ~10TB). Furthermore, the size of the library is dictated by the sequencing platform used for access. For example, suppose the Illumina NovaSeq X platform (currently Illumina's highest throughput sequencing machine) is used and data is encoded at the theoretical limit. In this case, a user could only store and access 312 GB of data at a time (assuming 52 billion total reads, a read depth of 10, 300 bp strands, 2 bits/base, two 20-base address primers, and a 20-base index), well below the level of modern external hard drives (~1-10 TB). Therefore, commercializing DNA data storage requires selective retrieval of DNA to be sequenced, termed "random access" in computer science.

Standard molecular biology techniques can be harnessed to allow selective retrieval[111,18,112,113,46,114,19,82,115,73,116,117,118,119,91]. By assigning unique sequences of DNA, called primers to the distal ends of each molecule comprising a file or block of data, the various retrieval methods described below can be used to access each file or block of data specifically.

4.2.1 PCR Based

PCR based random access[46] and retrieval is typically a destructive process that exploits the exponential copying (amplification) of DNA sequences flanked by *primer sequences*. When a file

is accessed using PCR a pair of short DNA oligos (primers) that match the 5' and 3' ends of all oligos associated with a particular file are used along with an enzyme to selectively amplify the file of interest (Figure 23). This reaction will add roughly 1-2 hours of latency to data retrieval. Once the abundance of the target file is several orders of magnitude higher than unaddressed files it can be sequenced such that the majority of the reads will belong to the addressed file. This method is considered destructive because the resulting sample is now heavily skewed toward one file and returning it to the DCS would make subsequent retrieval of other files difficult. This issue is mitigated by the fact that contemporary synthesis methods create very high physical redundancy, allowing for multiple samples of a pool to be retrieved for reading before the data must be rewritten. Other drawbacks of PCR based random access are that parallel access of multiple files is challenging (it is hard to amplify more than one file's strands at a time) and the address space of the primers is limited[19].

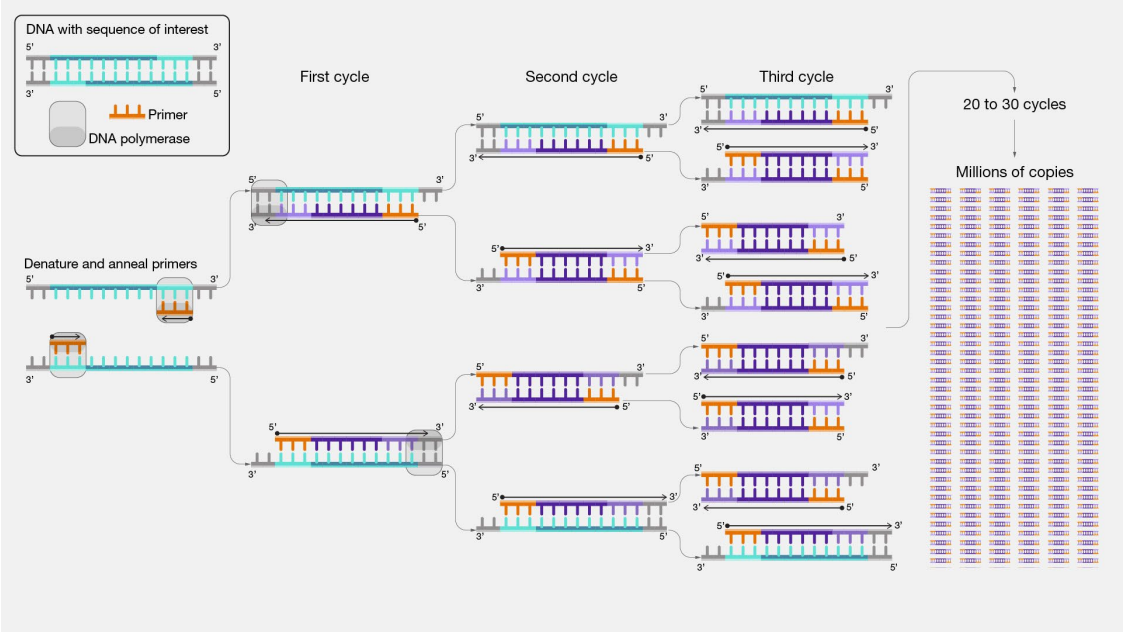

Figure 23 - Polymerase Chain Reaction (PCR): Original DNA templates are mixed in solution with complementary primers and nucleotides. The solution is heated and the original DNA denatures into two individual single strands. As the solution cools, primers anneal to the original DNA and are extended in an elongation phase. This process is repeated over many cycles resulting in exponential amplification of target sequences. Figure courtesy NHGRI (www.genome.gov).

4.2.2 Pull down approaches

The incorporation of molecular handles and magnetic bead separations have been used to increase the selectivity of file retrievals. One method attaches molecular handles on the PCR primers used for file amplification. These modifications, ranging from DNA barcodes, small molecules, and proteins, are recognized and bound by the magnetic beads used to physically separate the newly copied strands of a file. This PCR-based method, deployed in conjunction with a hierarchical file address encoding, exponentially increases the number of files stored with a minimally more complex workflow and successfully performs random access. This reaction will add roughly 1-2 hours of latency to data retrieval.

4.2.3 Hybridization based

Hybridization access methods rely on complementary DNA-DNA interactions, allowing for the binding and selective extraction of specific DNA strands. Combining DNA hybridizations with a pull-down technique allows for efficient and precise manipulation of DNA strands while reducing the reliance on PCR, which is known to introduce biases and errors. Numerous research groups have highlighted the potential for implementing DNA hybridizations to perform in-storage functions, including, renaming, locking, deleting, previewing, and searching data.

4.2.4 In vitro transcription - using RNA to access data

Drawing inspiration from natural biological systems, researchers have utilized In vitro transcription, a process that transcribes data stored in DNA into RNA for data retrieval. Despite introducing a latency of up to 8 hours, this approach offers promising advantages. This method enables reusable DNA libraries since DNA and RNA can be easily separated. Once separated, DNA remains in or is returned to the original library. Two pathways exist for decoding the data now stored in the RNA: reverse transcription into DNA for sequencing or direct sequencing using RNA sequencing methods. The feasibility of this approach has been successfully demonstrated[120], and research is ongoing that focuses on optimizing and refining it for practical implementations on a larger scale[82,82]. By incorporating hybridization and pull-down techniques, this approach has been used to enhance database capacity, reusability, and in-storage functionality.

4.2.5 Similarity search

In this approach[121,122], features of digital data are encoded as digital vectors and encoded into single stranded DNA sequences that are appended to the distal end of strands encoding the corresponding digital data. To query the archive, query statements are similarly encoded as digital vectors and encoded into single stranded DNA sequences. In this case, single stranded molecules created from an encoded target file and the reserve complement of an encoded query are likely to form stable hybridized structures when the query and the target vectors are similar, but not when they are distant. This hybridization results in double stranded DNA that can then employ any of the PCR, pull down or hybridization approaches described above.

4.3 Attributes and Metrics

4.3.1 Media Stability and Data Retention

There are two primary parameters that are commonly used to characterize how long DNA media can be stored at rest and maintain the ability for the data to be recoverable: (1) Media Stability and (2) Data Retention. Media Stability defines how long the media can be stored in the DCS such that, at the end of the storage period, there is enough of the media present to enable retrieval and reading of the data. Data Retention is defined as the period of time the media can be stored in a DCS such that, when the media is retrieved and read, the encoded data in the media can be successfully recovered. DNA Media Stability is typically defined as a function of the storage method, independent of the rest of the DNA data storage pipeline (and the codec), while Data Retention is dependent on the storage method, plus the synthesis, retrieval and sequencing methods used, the capabilities of each are visible to and accounted for by the codec.

The DNA Data Storage Alliance is defining standard metrics and methods for measuring and characterizing both Media Stability and Data Retention for DNA data storage use cases. The first such specification, the DNA Data Stability Evaluation Method for DNA Data Storage Containment Systems[123], defines Media Stability for a DCS (and the experimental method for characterizing it) as half-life; that is, the time it takes such that, at 25°C, only half of the molecules in the pool remain intact (i.e., can be successfully amplified or otherwise prepared for reading). In this way, competing claims of different DCS vendors can be objectively compared. Data Retention metrics and characterization are more complex, as they depend on the entire data storage pipeline. Work on these and other DNA data storage media reliability topics is ongoing.

4.3.2 Throughput and Latency

Throughput in a DCS is the rate at which bytes are either written (stored into) or read (retrieved from) the DCS vessel and is measured in bytes/unit time. While the throughput of Synthesis and Sequencing is a much larger factor in the overall performance of a DNA data storage end-to-end pipeline, different DCSs do have different rates at which data can be physically written into and retrieved from them.

Latency for a DCS is complex to characterize. For writing, this includes the time required to prepare the storage container and perform any required chemistry (additives, drying, vessel sealing, etc.) to preserve the DNA media. For reading, this includes the time to extract the desired molecules from the DCS vessel and move them to sequencing (i.e., time to first byte delivered to sequencer). Read latency can also include the time required to return molecules to the DCS vessel to maintain archive integrity and, also, the time required to identify the appropriate storage vessel from which to retrieve the desired media.

Selecting a DCS with the right Throughput and Latency will be use case dependent and will be affected by the TCO as required by the end-user and advertised by the product provider.

4.3.3 Archive reusability

Archive reusability is the number of times an archive can be sampled before a given file is no longer recoverable. As seen above, some retrieval approaches remove molecules from the archive, limiting the number of times data within the archive can be accessed.

4.3.4 Environmental Impact / Sustainability

The environmental impact and sustainability of storage and retrieval are measured in the amount of greenhouse gasses emissions, energy consumption and resource consumption (e.g., fresh water) per byte of data stored or retrieved. Because media degradation is temperature and humidity dependent, and various chemistries may be required for preservation, contributions to these measures include the manufacture and disposal of reagents used in preservation, environmental control systems (e.g., HVAC), and process automation (e.g., robotics).

4.3.5 Biosecurity/Biosafety

Biosafety and biosecurity are broad topics which cross many aspects of the DNA data storage pipeline. See Section 6.4 for a discussion of this topic.

4.3.6 Privacy/Media Security

Addressing privacy concerns is of utmost importance in the realm of any data storage system. A DDS must ensure the reliable retrieval and access of only the intended data, and the ability to delete/erase data. It must accommodate data with distinct privacy standards, originating from different users, or subject to varying security clearance levels necessitates the segregation of such data into separate DNA pools. To meet data privacy rules and regulations (GDPR, CCPA, HIPPA, etc.) and comply with governmental/judicial requests, it becomes imperative to implement mechanisms that allow for the specific deletion of data sets (e.g., personally identifiable information) when required.

5 Sequencing

5.1 State of the Art

Sequencing facilitates reading information stored in the molecular media. At the current state of the art, there are two main approaches to reading DNA sequences: 1) Sequencing by synthesis (SBS), where the sequence of the information contained in the original DNA template strand is read indirectly from a complementary strand that is synthesized from the template strand; and 2) Nanopore, where bases in the target strand are directly read as the molecule passes through a nanometer scale hole in an otherwise impermeable membrane.[124]

5.1.1 Sequencing by synthesis (SBS)

Sequencing by synthesis is the dominant approach to sequencing today. In this approach, the molecule to be sequenced is used as a template to synthesize a complementary DNA strand. The sequence of the complementary strand is determined during this synthesis process, as the identity of bases are determined as they are added to the complementary strand. In this section, the various approaches to implementing sequencing by synthesis are reviewed.

5.1.1.1 The Sanger method

While not considered as a practical option for DDS, the historical basis of today's SBS approaches is Sanger Sequencing[125]. Developed by Frederick Sanger in 1977 and commercialized by Applied Biosystems[126], the Sanger method remains the gold standard for sequencing approaches. Although throughput, costs and scaling limits of electrophoresis prohibit this approach for DNA Data Storage applications (Table 5), it is, due to its ability to read lengths >500 bases and accuracy around 99.99%, still used as validation for other sequencing approaches discussed in this chapter.

The method relies on chain termination chemistry with fluorescently labeled dideoxynucleotides. As seen in Figure 24, a purified single stranded template strand, a DNA primer, a DNA polymerase and both deoxy and dideoxynucleotides (bases) are added to a reaction. The primer, complementary to a portion of the target strand, acts as the initiator for the synthesis of a complementary DNA strand from the nucleotides. When a dideoxynucleotide is added to the complementary strand, the reaction terminates. Once the reaction is complete, the reaction products are separated by capillary electrophoresis, which separates the strands by length (smaller strands move faster) with a resolution of one base. As fluorescently terminated strands move past a detector, the color of the fluorescent tag is measured, identifying the terminal base. This produces a chromatogram identifying the sequence of the original target DNA strand.

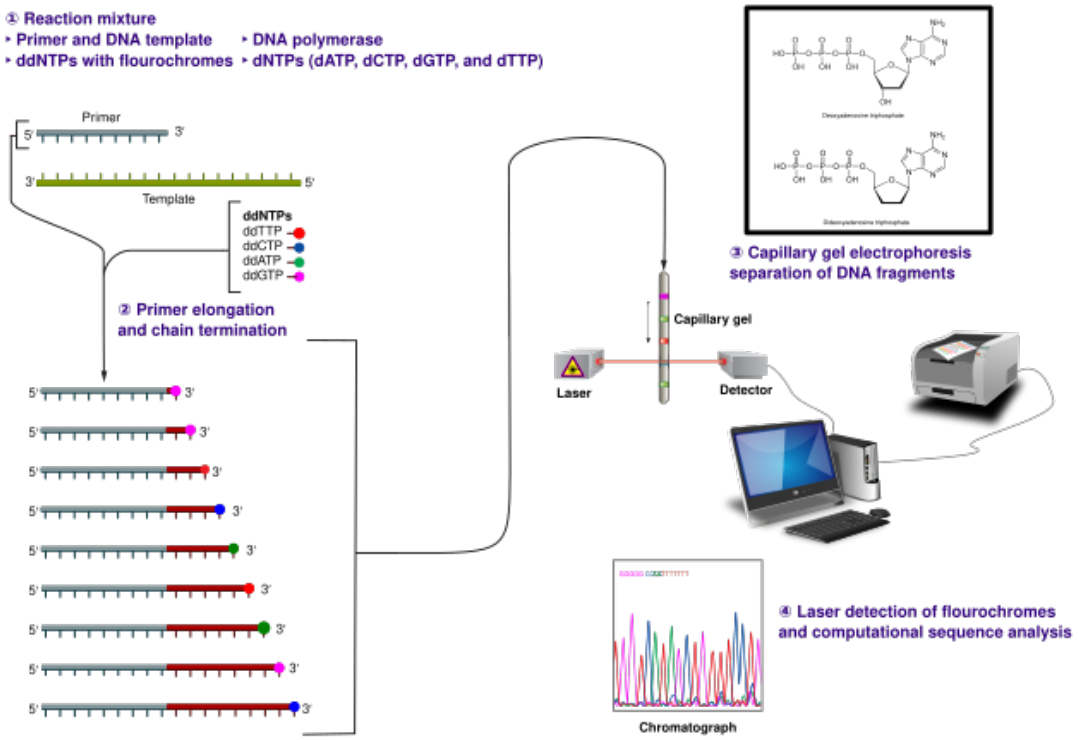

Figure 24 - Sanger Sequencing. Reprinted from https://en.wikipedia.org/wiki/Sanger_sequencing.

5.1.1.2 Cluster based SBS

The first example of cluster-based SBS is Illumina Sequencing (Figure 25). Cluster-based SBS is performed on a patterned surface within a flow cell consisting of nano wells on a 300-700 nm pitch. Each well in the flow cell contains multiple copies (“clones”) of a single stranded subsegment of the DNA sample under measurement, typically a few hundred bases long (100 - 300 nucleotides in length). During SBS, the single-stranded DNA in each well is used as a template to synthesize a complementary strand, resulting in double-stranded DNA using Watson-Crick pairing rules (A-G and C-T). At each step in the process, a population of A, C, G, and T nucleotides are flowed into the flow cell. These nucleotides are fluorescently labeled and chemically modified (“blocked”) so that only one nucleotide at a time can be added to the complementary strand. The complementary strands in each well are thus appended by a single complementary and marked nucleotide. After each binding cycle, the fluorophore on the newly incorporated nucleotides is excited with a light source and the wavelength (color) of fluorescence is used to determine which nucleotide (A, C, G, or T) was incorporated in any particular well. Lastly, the blocker on the newly incorporated nucleotide is then removed with a chemical reaction and the whole process is repeated to incorporate and interrogate the next nucleotide. This cycling process is repeated between 50 and 300 times depending on the length of the template strand. The process is performed in parallel on all wells. A high-performance flowcell contains between 1B and 25B wells, yielding between 300B and 7.5T base calls per flowcell.

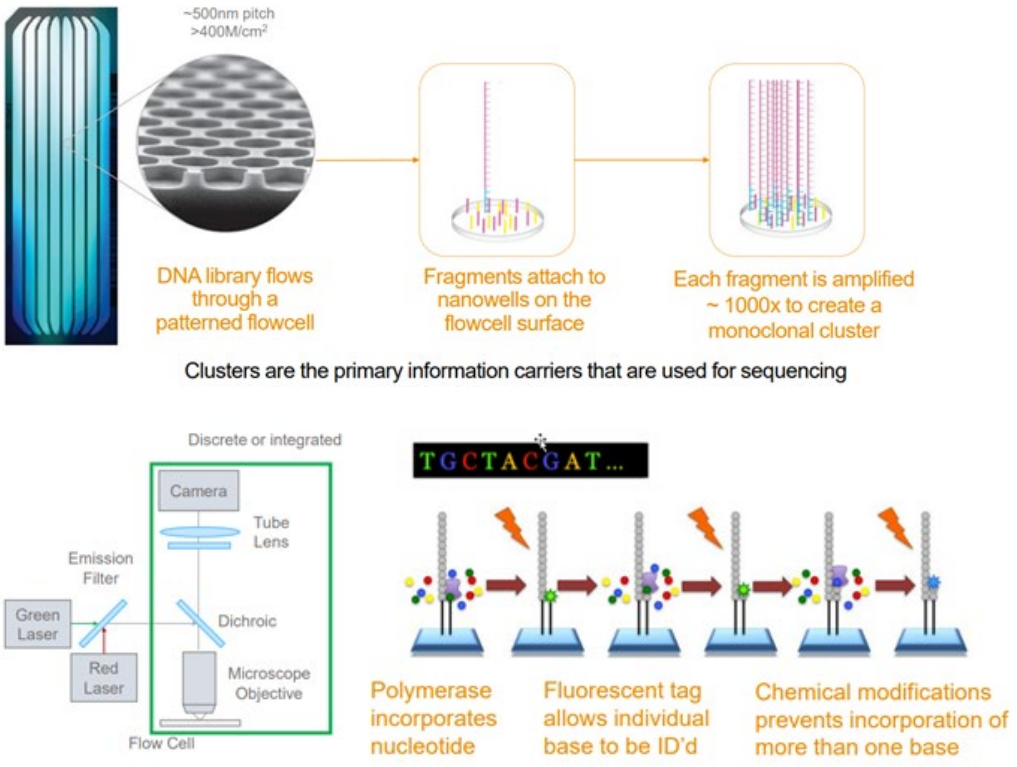

Figure 25 - Sequencing by Synthesis (Courtesy Illumina)

The broad success of cluster based SBS technology has led to a number of alternative approaches that employ clusters of clonal sequences as a sequencing template. Emerging approaches, including that commercialized by Ultima Genomics (Figure 26), employ alternative approaches to cluster generation and sequencing chemistry[127, 128]. In the case of Ultima Genomics[129], clusters are generated on beads prior to introduction to the instrument and immobilization in microfabricated locations on the flow cell. Similar to incumbent cluster-based approaches, fluorescence is used to monitor base addition to a growing daughter strand, and statistical approaches coupled with machine learning are used to identify base identity and quantity. As seen in Table 5 below, emerging approaches could provide scalable, cost efficient sequencing approaches suitable for DDS.

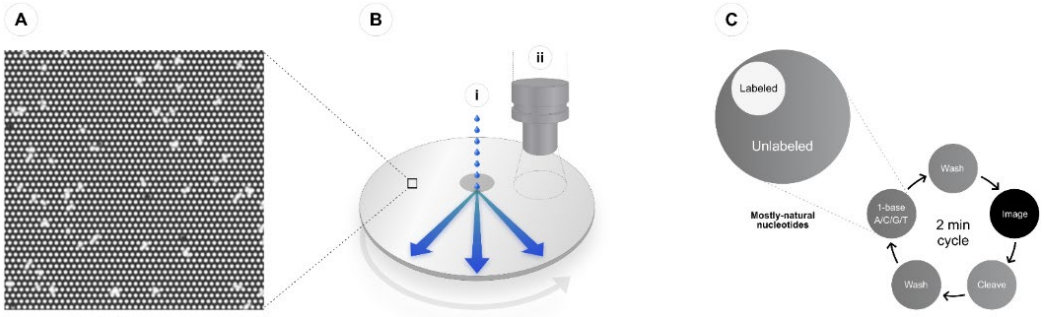

Figure 26 - Ultima Genomics Mostly Natural Sequencing by Synthesis (Reprinted from [129]): (A) Scanning electron micrograph of wafer surface patterned at micron resolution to allow binding and sequencing of billions of clonally amplified sequencing beads; (B) Open fluidics systems allows (i) dispensing of reagents from dedicated nozzles near the center of the rotating wafer to distribute reagents by centrifugal force and (ii) optical measurement of the entire wafer surface in one continuous step; (C) chemistry cycle includes addition of one type of mostly-natural nucleotide mix at a time (dA, dC, dG or dT) followed by imaging and cleavage of the sparse labels.

5.1.1.3 Single molecule optical SBS

Single molecule sequencing, also known as single-molecule real-time (SMRT) sequencing, originated from the efforts of researchers to develop innovative techniques for DNA sequencing[130,131,132]. The foundational work for single molecule sequencing was laid by Stephen Quake and Jingyue Ju in the late 1990s. They demonstrated the feasibility of sequencing single DNA molecules by monitoring the release of individual nucleotides during DNA polymerization. In 2003, the company Pacific Biosciences (later renamed PacBio) was founded with the goal of commercializing SMRT (Figure 27). SMRT sequencing involves the real-time monitoring of DNA polymerase activity as it incorporates fluorescently labeled nucleotides into the growing DNA strand. The fluorescence emitted during nucleotide incorporation is detected and recorded, enabling the determination of the DNA sequence. This technique eliminates the need for amplification, as sequencing is performed on individual DNA molecules.

The most recent PacBio sequencing technology is used in the Sequel II, Sequel IIe and Revio systems and offer several improvements over previous generations of PacBio sequencers. They feature an increased number of wells and improved chemistry, allowing for higher data output and enhanced read lengths. The Sequel IIe system specifically provides longer sequencing times, enabling relatively long reads (5k to 60k nucleotides). One notable feature of PacBio sequencing technology is its ability to generate long reads, often referred to as "HiFi reads." HiFi reads are achieved by using circular consensus sequencing (CCS) to generate multiple passes of the same DNA molecule. This approach greatly improves sequencing accuracy, making it comparable to or even surpassing short-read sequencing technologies in terms of base-level precision.

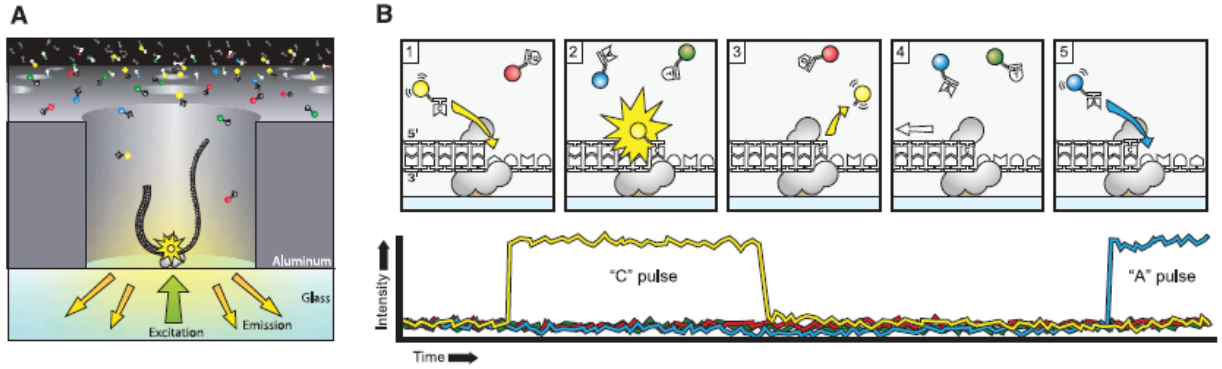

Figure 27 - PacBio SMRT sequencing (reprinted from [130]): Sequencing instruments contain several flow cells that run in parallel. Flow cells contain several zero-mode waveguide wells with a DNA polymerase immobilized at the bottom. During sample preparation, target DNA is assembled into a single stranded dumbbell template. Dumbbells are loaded into the flow cell and bind to immobilized polymerases. Next, fluorescently tagged nucleotides are flown into the flow cell. Signal is generated as nucleotides are added to the growing daughter strand. Circular consensus sequencing is achieved as the polymerase makes several copies of the dumbbell template.

5.1.2 Nanopore

5.1.2.1 Biological nanopore

Nanopore sequencing, a technology that has been gaining traction in recent years, represents a paradigm shift in the way DNA sequencing is approached[133]. This method, commercialized by Oxford Nanopore Technologies (ONT) and their portable, stapler-sized MinION device[134], is based on the principle of precision measurement of fluctuations in ionic current as DNA molecules move through a nanoscale protein pore (Figure 28). The distinctive electrical signals created by

different DNA bases translocating through the pore allow the sequence of bases in the DNA strand to be determined in near real-time. This real-time data streaming allows for immediate analysis, which could be a critical advantage over SBS-methods in time-sensitive data storage and retrieval use cases. Another attractive feature of nanopore sequencing technology, particularly exemplified by Oxford Nanopore Technologies' MinION device, is its portability and compatibility with everyday devices such as laptops, tablets, and even smartphones.

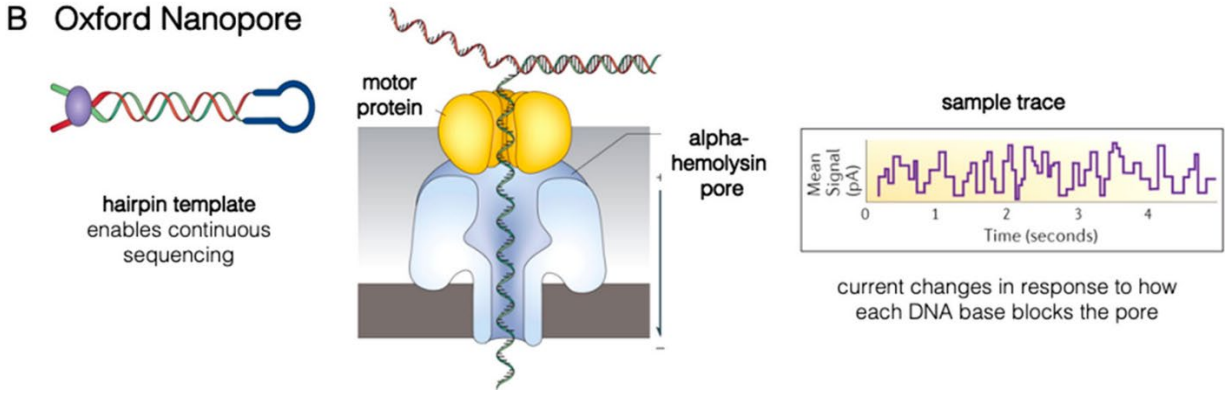

Figure 28 – Example of nanopore sequencing from Oxford Nanopore (reprinted from [39]): During the sample preparation step, leader and hairpin sequences are attached to sample DNA strands. The leader sequence primes the sample DNA in a motor protein that ratchets the DNA molecule through the nanopore. As the single stranded DNA molecule moves through the nanopore, ions are co-translocated under an applied electric field. Current (in pA) through the pore is modulated by the sequence of DNA present within the nanopore, and this measurement can be used to determine the target DNA sequence. The motor protein reduces the translocation speed and increases reading accuracy. The nanopore is embedded in a dielectric membrane, typically a lipid bilayer or synthetic polymer. Several types of nanopores have been employed, including alpha hemolysin, MspA, Refarc and CsgG. The electric field drives ion translocation through the nanopore, generating a transient ionic current, recorded by a current amplifier. Electro-osmotic forces across the membrane, hydrodynamic forces, entropic forces and electrostatics (attraction/repulsion), in addition to the applied electric field, are the main forces contributing the translocation process.

5.1.2.2 Solid state nanopore (SSN)

Solid state nanopores employ similar physical methods as biological nanopores to make single molecule measurements (i.e., ionic current fluctuations) [135]. SSNs are differentiated from biological pores in that the nanopore is a perforation in a relatively rigid membrane, with membrane examples including silicon, silicon nitride, silicon oxides, quartz, metals, polymeric films, nanowires and 2D materials such as graphene, boron nitride, and molybdenum disulfide. These materials have advantages over polymeric or lipid-based membranes used in biological nanopores, including increased resiliency to high-potentials required for translocation and solution conditions (e.g., non-neutral pH, chaotropes, detergents, etc.) that promote controlled single molecule translocation. These materials can also be functionalized to incorporate secondary, complementary, sensing approaches, such as photon-based, tunneling, field effect transistors, and plasmonics, to create systems with two simultaneous sensing modalities that can increase accuracy.

While various techniques for manufacturing membranes and nanopores have been demonstrated[135], some enabling array fabrication[136], controlling pore geometry remains a significant challenge. To date, this has prevented a demonstration of DNA sequencing with single nucleotide resolution. Block homopolymer sequencing has been demonstrated in an academic

setting[137] and could be compatible with block homopolymer encoding approaches, but an application has yet to be demonstrated. SSNs have shown significant utility in measuring the translocation of DNA strands decorated with secondary structures, including proteins and DNA nanostructures[Error! Bookmark not defined.]. These molecules can be used to encode information and DDS applications have been demonstrated[138].

SSNs have been commercialized, most notably by Northern Nanopore and Goeppert LLC, as well as tunable, elastomeric instrument system by Izon Science for single particle resolution analysis (e.g. size, concentration, surface charge).

Sequencing by Expansion[139] (SBX), by Roche, is a recently announced potential addition to the nanopore sequencing landscape. SBX uses chemical techniques to prepare a template molecule with structures that enable control of speed and other aspects of molecular translocation through a nanopore, with low error rates, high performance, and long reads.

5.2 Attributes and Metrics

5.2.1 Sequencing length

The sequencing length is defined as the number of continuous bases, or nucleotides (nt), within a given molecule that can be read during a single read event. This length varies for each sequencing modality, as does the definition of a read event.

Current synthesis capabilities favor the synthesis of relatively short oligonucleotide polymers, typically 100-300 nt in length. These factors align well with the current capabilities of cluster-based sequencing, which have sequencing lengths in the 100-300 nt range. Storage and retrieval approaches deliver double stranded DNA, and both strands are sequenced. Sequence information from each strand can be compared to increase the quality of the read (paired end read). This alignment in capability comes with tradeoffs, as the low error rates and limited length decrease sequencing throughput and increase its costs.

As synthesis capabilities evolve, with longer synthesis lengths at lower cost and lower error, nanopore sequencing and single molecule SBS approaches can confer distinct benefits to DNA data storage. For instance, shorter strands necessitate a larger proportion of their encoding space to be devoted to addressing as opposed to payload contents. Assembling these additional addresses can be computationally challenging and susceptible to errors, especially with larger datasets. Long reads can help alleviate these issues, simplifying the decoding process.

5.2.2 Throughput

Throughput is measured in the amount of data read per unit time and is unique from read latency, which is discussed below. Because the translation from bases to bytes requires a decoding step, throughput can be difficult to characterize. We use an approximation of 1 bit per base for discussion purposes but codecs commonly achieve higher bit density (Section 2).

As seen in Section 5.1, approaches to DNA sequencing are varied but parallelism has driven throughput performance of biotechnological sequencing systems within a few orders of magnitude of traditional storage media. As discussed in the Section 3 (Synthesis), uncompressed tape has throughput on the order of 400 MB/sec and HDD on the order of 300 MB/sec. Illumina's NovaSeq X platform delivers throughput on the order of 12 MB/sec and ONT's PromethION platform ~1MB/sec. Many sequencing technologies take advantage of microelectronics fabrication but they have not yet approached state of the art feature sizes and scaling will continue.

5.2.3 Error Profile

There are several sources of error in DNA sequencing and they are similar to those seen in DNA synthesis. Errors are measured in the rate of read errors per base read (errors/base). Per Table 2 (Section 2.2) there are three main sources of errors: insertions, deletions and substitutions. An insertion occurs when an unintended base is inserted in the read sequence between two intended bases. A deletion occurs when an intended base is not included in the read sequence. A substitution occurs when an intended base is replaced by a different base in the read sequence. The location and frequency of sequencing errors can lead to a read sequence that cannot be decoded, resulting in a fourth error type: erasures.

There are numerous underlying physical factors that contribute to sequencing errors; these factors are cumulative, meaning that they can combine to increase the overall read error rate. While a comprehensive accounting of all factors that contribute to errors are outside the scope of this document, the most common sequencing errors in SBS systems arise from issues in removing the blocking group, making insertions and deletions the prevalent error types in SBS. In nanopore sequencing, errors arise from noise in the measurement, making it difficult to discriminate between individual bases and leading to all three error types. There are higher-accuracy base callers that offer superior precision in DNA data reads. However, these require sophisticated algorithms that are more computationally demanding.

As synthesis typically results in many copies of each molecule (i.e., populations of molecules with the same sequence), errors generated in determining the sequence of a single molecule can be overcome by reading many copies of the same sequence and calculating a consensus sequence. This process is used by all sequencing modalities and is known as "coverage". While the error rate at high coverage varies by sequencing modality, it is not unusual to see raw (i.e., uncorrected) error rates for SBS at ~0.1% and nanopore at ~6%. Sequencing at high coverage leads to slower data processing times (decreased throughput), increased latency, and, generally, elevated cost due to higher computational resource needs.

Erasures can also arise from sample preparation steps that precede sequencing. There is some overlap with retrieval (Section 4.2), and we encourage a review of that chapter for details. In short, information containing molecules may need some processing to make them compatible with the sequencing chemistry. For example, cluster based sequencing requires the attachment of adapter sequences to the distal ends of each molecule, while nanopore based sequencing requires the attachment of a leader sequence to each molecule, which includes a motor protein that controls

translocation rate[140]. The success of this sample preparation step sets the quality ceiling for a sequencing approach and there is an opportunity to tune these for DDS applications.

5.2.4 Read Latency

Sequencing read latency is defined as the time from when molecules are retrieved from the archive to when the sequencer begins generating base calls from those molecules. Sequencing read latency can be measured in minutes to hours depending on the sequencing modality, degree of sample preparation automation, the need for consensus base calling and computational resources deployed.

Prior to generation of raw sequencing data, desired molecules from the archive are retrieved (Section 4.2) and prepared. The transition from the retrieval phase to sample preparation for sequencing is well defined in biotechnology applications, as molecules must be extracted from biological samples and purified in the retrieval step. As seen above in the discussion of erasures, sample preparation may require amplification (e.g., PCR) and attachment of adapters. In DDS applications this transition is less well defined. Encoding schemes may include sequences at the distal ends of molecules that increase sample retrieval and preparation efficiency. Storage approaches may minimize the complexity of molecule extraction and purification.

In the sequencing operation, there are two primary determining factors to latency. The first factor is the time required to generate raw read data. For SBS modalities, since raw data is generated by the addition and reading of each base in a cluster of complementary strands, the time required to generate raw read data for the entire sequencing length may take several hours. In nanopore, single molecule raw data for the entire sequence length is generated on the order of seconds. The second factor is determining a sequence from raw data. As discussed above, base calling and clustering algorithms are required to convert raw data to called sequences of bases. For current biotechnology applications, significant computational resources are required to deliver very low error rate sequences[141] and, in some cases, dedicated hardware is employed (e.g., Illumina's FPGA based DRAGEN platform).[142]

5.2.5 Environmental Impact / Sustainability

The environmental impact and sustainability of sequencing approaches are measured in the amount of greenhouse gasses emissions, energy consumption and resource consumption (e.g., fresh water) per byte of data read. Both SBS and nanopore sequencing approaches operate in aqueous solutions, therefore there is little use of hazardous (including biohazard) reagents and most liquid waste streams can be directed to municipal water treatment[52]. There is the potential for significant solid waste streams as current sequencing substrates are single use.

The largest contributor to resource consumption is the power consumed executing base calling algorithms that convert raw signals to called bases. Many of these approaches employ machine learning algorithms run on GPUs, TPUs or custom hardware, which can be energy intensive. For example, recent work has demonstrated that real-time base calling (processing 1.5M signals per second) can be achieved using a laptop-powered MinION nanopore sequencer and an optimized

TPU hardware accelerator while consuming ~60 microJoules/base called[141]. Assuming the same energy consumption in a PromethION system, this amounts to roughly 4 Gbases called/Watt.

6 Challenges to Commercialization

As the previous sections have documented, there is significant ongoing work, both academic and commercial, on building DNA data storage solutions. However, while there have been successful DDS demonstrations and the fundamentals have been shown to work on scalable technology platforms, challenges to commercial adoption remain. The following is our assessment of the most significant challenges to DDS commercialization.

6.1 Data throughput

The speed of data movement through the information channel, measured in bytes/sec, is an essential attribute of data storage systems, with use cases ranging from write many read never, to write once read many, and all combinations in between. The requirements for moving data into and out of traditional storage far exceed the current capabilities of writing and reading DNA in biotechnology use cases.

The most fundamental challenge for DDS systems is to increase the throughput of DNA write and read operations. The underlying write and read operations for DNA are relatively slow chemical reactions (high latency), so the emphasis for increasing throughput involves enabling parallelism. As use cases evolve into more flexible data center use cases, new levels of automated material handling and preparation steps will also be needed. To position DDS as a viable storage solution, DDS technology providers must increase the throughput of the underlying write and read operations, as well as reducing the time required to move molecules between operations, all while maintaining a competitive TCO for the use case at hand.

6.2 Total cost of ownership (TCO)

As the concept of DDS has been explored, the costs have tended to be assessed through the lens of biotechnology or traditional storage; that is, costs to write or read DNA in $/base. While the so-called “speeds and feeds” and costs of the components of DDS systems are critically important, DDS has many aspects which are new to data center environments, for example liquid waste streams from synthesis, fluid reagent movement between storage stages, rehydrating desiccated media, and specialized technical labor skill sets. Thus, developing a TCO model for DDS is a key to its commercial acceptance. Also important is the notion, often an attribute of new technologies, that the TCO requirements for early DNA data storage use cases, such as very long term archival storage of invaluable data libraries, where the write/read costs will be amortized over a very long period of time, will be different than the TCO for later use cases, such as more active archive in a working storage hierarchy. To position DDS as a viable storage solution for specific use cases, DDS technology providers must adopt total cost of ownership as a metric and develop standardized methodologies to assess TCO relative to traditional media over the lifetime of the data.

6.3 Media endurance and data retention metrics

Media endurance and obsolescence are issues for traditional media. Indeed, as we have discussed, one of the challenges of traditional storage media, especially for colder data, is that,

as the amount of digitized data explodes, the cost of keeping this data is becoming prohibitive in terms of needing to run fixity checks and to periodically migrate data to new generations of technology. Due to its endurance and immutability, it is expected that DDS will avoid these problems; however, DNA is a completely novel medium for data storage, with novel access methods, materials handling, etc., and "proving" that these problems are not problems with DNA is non-trivial.

The challenge for DDS technology developers is providing standard quantitative methods that provide confidence that DNA as a storage medium represents a true breakthrough in endurance and data retention (see Section 4.3.1).

6.4 Biosecurity and data security

Biosecurity (preventing the deliberate misuse of biological materials, protecting against bioterrorism, bio-crime, and other forms of biological aggression) and biosafety (protocols and practices designed to protect researchers, the public, and the environment from unintentional exposure to biological agents) have played a key role in DNA technologies across medical and scientific applications as synthetic biology technologies have enabled the manufacturing and manipulation of nucleic acids.

The DNA synthesis industry has, since the founding of the International Gene Synthesis Consortium[143] in 2009, worked to 'design and apply a common protocol to screen both the sequences of synthetic gene orders and the customers who place them.' This represents a substantial effort on the part of the synthesis industry to ensure that synthetic nucleic acid technologies and products are used only in responsible research.

The industry has worked alongside the United States government since 2009 resulting in publication of US government guidance on biosecurity for synthetic DNA in 2010[144] and a revised and expanded guidance document[145] on biosecurity for all synthetic nucleic acids in 2023. These documents identify a class of sequences subject to direct regulatory control or other misuse concerns described as 'sequences of concern' (SOC) and suggest specific practices for providers to detect and react to orders for SOCs. Also, organizations like the International Biosecurity and Biosafety Initiative for Science (https://ibbis.bio/) have evolved to free, distributed, open-source, automated software (i.e., the Common Mechanism[146]) for screening sequences of nucleic acids (including DNA and RNA) as well as resources to facilitate customer screening on sequences down to 50 base pairs.

In the context of using DNA as a data storage medium, the sheer number of DNA molecules being generated is vastly higher than in the current applications, which could create concern, but the fact that we are using the DNA for encoding digital data enables the means of significantly or completely mitigating the risks. To cite just two examples, within a DNA codec, encoding algorithms can be designed to avoid SOCs, and DNA synthesis hardware for DDS can be designed to never write a sequence without a corresponding certificate from the codec that the sequence is free of concern.

The DNA Data Storage Alliance is working internally, and in coordination with other parties in the synthetic nucleic acid ecosystem, to develop trusted device and trusted party models such that DDS technology providers, users, and regulatory stakeholders can create an ecosystem that allows DDS technologies to flourish while simultaneously addressing biosecurity and data security issues.

6.5 Standardization

There is a diverse ecosystem of DNA data storage solutions evolving today, in both academia and industry. While the field must stay open to innovation in this nascent phase, there is also value in the judicious use of standards to help a multi-vendor interoperable ecosystem emerge. A balance has to be made between innovation and interoperability.

The DNA Data Storage Alliance is attempting to find this balance and is working in the following areas:

1) **Interoperable Interfaces:** In a multi-vendor ecosystem, the discrete pieces of a complex system must be interoperable, both to enable flexibility of function for system designers, and to provide system component vendors a common interface around which they can innovate and compete. The Alliance is working on a generic DDS system interoperability model to enable plug and play amongst synthesis, storage, and sequencer vendors, as well as other mechanical aspects, such as fluid handling, reagent disposal, etc.
2) **Endurance and Data Retention:** Per the Media Endurance challenge (section 6.3), the Alliance is working on standard metrics and test methods to characterize endurance and data retention as it relates to DDS. The first such standard is the *DNA Data Stability Evaluation Method*[123], discussed in Section 4.3.1, enabling the comparison of media endurance claims for different vendors' DNA Containment Systems.
3) **DNA Archive Discovery and Identification:** We believe that the ability to recover the data within a DNA storage archive without prior knowledge of the contents of the archive, is critical to the mission of DNA archives in general, in particular for long term archival use cases. Standardizing such self-describing discovery methods enables users to read and decode a DNA archive far into the future, even if the provenance is not 100% clear. Section 2.4.2 references the Alliance work in this area.
4) **Biosecurity and Biosafety:** Biosecurity refers to measures taken to prevent the creation of or stop the misuse of or exposure to harmful biological agents. As the ability to write and read DNA molecules at scale proliferates, there is a need for alignment on the potential threats of using synthetic DNA in the context of data storage. Section 6.4 discusses work in this area.
5) **Codecs:** DNA Codecs are, today, still quite specific to the underlying chemistry and access schemas (e.g., random access probe address space) of the DNA data storage pipeline being implemented. Along with this document, the DNA Data Storage Alliance also published a white paper titled *DNA Data Storage Codecs: Examples, Requirements, and Metrics*, which surveys some DNA data storage codecs, and attempts to provide some guidelines on what attributes a good DNA codec should implement. Standard codecs or

codec certification (e.g., for discovery, biosafety) are probably premature for now, but may emerge in the future.

We anticipate that other areas of standardization will emerge as DNA data storage enters commercialization and future versions of this document will endeavor to tie technical advances to such standards. Please visit SNIA (https://www.snia.org) for more information on DNA data storage standards and white papers published or in development by the Alliance.

7 Conclusions

While DNA data storage is still quite nascent and there remain significant challenges to commercialization, the foundations of writing, storing, retrieving, and reading data using DNA have been shown to work on scalable technology platforms. Moreover, the ongoing investment in DNA technology, driven by biological and scientific applications, will continue to drive innovations that enhance DNA data storage capabilities.

It is important to see DNA data storage not as a replacement for any existing storage technology, but as a complementary capability that enables the data storage hierarchy to expand, resolving the “save/discard” dilemma with a viable TCO for zettabyte scale and data preservation.

We believe that the use of DNA for archival data storage use cases will emerge over the next 3-5 years, and that the continued investment in the segment will ultimately deliver more and more flexible capabilities enabling a wider variety of use cases.

8 Acknowledgements

We thank the following individuals for their contributions to this effort.

John M. Hoffman ()
David Landsman (Western Digital)
Omer Sabry (Technion)
Daniella Bar-Lev (Technion)
Manish Gupta (DA-IICT Gandhinagar)
Takashi Kobayashi (Fujitsu)
Don Doerner (Quantum Technologies)
Chris Takahashi (University of Washington)
Alessia Marelli (Avaneidi)
Bill Efcavitch (Molecular Assemblies)
Gerardo Bertero (Western Digital)
Gemma Mendonsa (Seagate)
Yvette Mimieux (Millipore Sigma)
Daniel Chadash (Twist Bioscience)
James Banal (CacheDNA)
Marthe Colotte (Imagene)
Jacques Bonnet (Imagene)
Julien Muzard (Entegris)
Kyle Tomek (DNAli)
Jeff Nivala (University of Washington)
Suayb S. Arslan (Boğaziçi University, MIT)
Mark Hahm (Illumina)
Stephane Lemaire (Sorbonne University, Biomemory)
Esther Singer (Twist Bioscience)

9 References

[1] Coudy D, Colotte M, Luis A, Tuffet S, Bonnet J (2021) Long term conservation of DNA at ambient temperature. Implications for DNA data storage. PLoS ONE 16(11): e0259868. https://doi.org/10.1371/journal.pone.0259868

[2] Landsman D, Strauss, K, "The DNA Data Storage Model" in Computer, vol. 56, no. 07, pp. 78-85, July 2023, doi: 10.1109/MC.2023.3272188

[3] Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77–80 (2013).

[4] Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nat. Commun. 10, (2019).

[5] Roquet, N. et al. DNA-based data storage via combinatorial assembly. 2021.04.20.440194 Preprint at https://doi.org/10.1101/2021.04.20.440194 (2021).

[6] Tabatabaei, S.K., Wang, B., Athreya, N.B.M. et al. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. Nat Commun 11, 1742 (2020). https://doi.org/10.1038/s41467-020-15588-z.

[7] Chen K, Kong J, Zhu J, Ermann N, Predki P, Keyser UF. Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. Nano Lett. 2019 Feb 13;19(2):1210-1215. doi: 10.1021/acs.nanolett.8b04715.

[8] Anavy, L., Vaknin, I., Atar, O. et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nat Biotechnology 37, 1229–1236 (2019); https://doi.org/10.1038/s41587-019-0240-x

[9] Masaki, Y., Onishi, Y. & Seio, K. Quantification of synthetic errors during chemical synthesis of DNA and its suppression by non-canonical nucleosides. Sci Rep 12, 12095 (2022). https://doi.org/10.1038/s41598-022-16222-2

[10] Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. Nat Biotechnol 34, 518–524 (2016). https://doi.org/10.1038/nbt.3423

[11] https://en.wikipedia.org/wiki/Concatenated_error_correction_code

[12] Welzel, M. *et al.* DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage. *Nat. Commun.* **14**, 628 (2023).

[13] https://en.wikipedia.org/wiki/Levenshtein_distance

[14] Shinkar, T., Yaakobi, E., Lenz, A. & Wachter-Zeh, A. Clustering-Correcting Codes. IEEE Trans. Inf. Theory 68, 1560–1580 (2022).

[15] Liu, Y., He, X. & Tang, X. Capacity-Achieving Constrained Codes with GC-Content and Runlength Limits for DNA Storage. in 2022 IEEE International Symposium on Information Theory (ISIT) 198–203 (2022). doi:10.1109/ISIT50566.2022.9834494.

[16] Levy, M. & Yaakobi, E. Mutually Uncorrelated Codes for DNA Storage. *IEEE Trans. Inf. Theory* **65**, 3671–3691 (2019).

[17] 1Yazdi, S. M. H. T., Kiah, H. M., Gabrys, R. & Milenkovic, O. Mutually Uncorrelated Primers for DNA-Based Data Storage. *IEEE Trans. Inf. Theory* **64**, 6283–6296 (2018).

[18] 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).

[19] Tomek, K. J. et al. Driving the Scalability of DNA-Based Information Storage Systems. ACS Synth. Biol. (2019) doi:10.1021/acssynbio.9b00100.

[20] DNA Data Storage Alliance, Sector 0 and Sector 1. https://www.snia.org/groups/snia-dna-technology-affiliate

[21] D. Bar-Lev, O. Sabary, R. Gabrys and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems," 2023 IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan, 2023, pp. 370-375, doi: 10.1109/ISIT54713.2023.10206882.

[22] G. Chaykin, N. Stein, O. Sabary, D. Ben-Shabat, and E. Yaakobi. "DNA-Storalator: End-to-End DNA Storage Simulator,", 13th Non-Volatile Memories Workshop, San Diego, California, 2022.

[23] Marelli, A. et al. Integrating FPGA Acceleration in the DNAssim Framework for Faster DNA-Based Data Storage Simulations. Electronics 12, 2621 (2023).

[24] Yuan, L., Xie, Z., Wang, Y. & Wang, X. DeSP: a systematic DNA storage error simulation pipeline. BMC Bioinformatics 23, 185 (2022).

[25] Alnasir, J. J., Heinis, T. & Carteron, L. DNA Storage Error Simulator: A Tool for Simulating Errors in Synthesis, Storage, PCR and Sequencing. Preprint at https://doi.org/10.48550/arXiv.2205.14437 (2022).

[26] Ou, L. Sonata165/DNA-Storage-Simulation. (2023).

[27] Oligonucleotides Synthesis on a Polymer Support. Letsinger, RL, Mahadevan V.J, Am Chem Soc. 1965 Aug 5;87:3526-7. doi: 10.1021/ja01093a058

[28] S.L. Beaucage, M.H. Caruthers, Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis, Tetrahedron Letters, Volume 22, Issue 20, 1981, Pages 1859-1862, ISSN 0040-4039, https://doi.org/10.1016/S0040-4039(01)90461-7.

[29] Oligonucleotide Synthesis Market Size, Share, Trends - [2022-2027]. *MarketsandMarkets* https://www.marketsandmarkets.com/Market-Reports/oligonucleotide-synthesis-market-200829350.html.
[30] Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
[31] Jensen, M. A. & Davis, R. W. Template-Independent Enzymatic Oligonucleotide Synthesis (TiEOS): Its History, Prospects, and Challenges. Biochemistry 57, 1821–1832 (2018).
[32] Tubbs, J. L. et al. Modified template-independent enzymes for polydeoxynucleotide synthesis. (2022).
[33] Palluk, S. et al. De novo DNA synthesis using polymerase-nucleotide conjugates. Nat. Biotechnol. (2018) doi:10.1038/nbt.4173.
[34] Henry, C., Enzyme conjugate synthesizes DNA, Method is possible replacement for phosphoramidite-based DNA synthesis. https://cen.acs.org/biological-chemistry/dna/Enzyme-conjugate-synthesizes-DNA/96/i26
[35]Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nat. Commun. 10, (2019).
[36] Efcavitch, J. W. & Holden, M. T. Homopolymer encoded nucleic acid memory. (2021).
[37] Roquet, N., Park, H. & Bhatia, S. P. Nucleic acid-based data storage. (2022).
[38] Dickinson, G. D. et al. An alternative approach to nucleic acid memory. Nat. Commun. 12, 2371 (2021).
[39] Doricchi, A. et al. Emerging Approaches to DNA Data Storage: Challenges and Prospects. ACS Nano (2022) doi:10.1021/acsnano.2c06748.
[40] Yu, M. et al. High-throughput DNA synthesis for data storage. Chemical Society Reviews, 9, (2024). doi.org/10.1039/D3CS00469D
[41] Nguyen, B. H. et al. Scaling DNA data storage with nanoscale electrode wells. Sci. Adv. 7, eabi6714.
[42] Smith J., Nguyen B., et al. Spatially Selective Electrochemical Cleavage of a Polymerase-Nucleotide Conjugate. ACS Synthetic Biology, 2023. DOI: 10.1021/acssynbio.3c00044
[43]Lietard, J. et al. Chemical and photochemical error rates in light-directed synthesis of complex DNA libraries. Nucleic Acids Res. 49, 6687–6701 (2021).
[44] Masaki, Y., Onishi, Y. & Seio, K. Quantification of synthetic errors during chemical synthesis of DNA and its suppression by non-canonical nucleosides. Sci. Rep. 12, 12095 (2022).
[45] Antkowiak, P. L. et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. Nat. Commun. 11, 5345 (2020).
[46] Organick, L. et al. Random access in large-scale DNA data storage. Nat. Biotechnol. 36, 242 (2018).
[47] Sack, M. et al. Express photolithographic DNA microarray synthesis with optimized chemistry and high-efficiency photolabile groups. J. Nanobiotechnology 14, 14 (2016).
[48]Li, X. et al. Inkjet Bioprinting of Biomaterials. Chem. Rev. 120, 10793–10833 (2020).
[49] Egeland, R. D. & Southern, E. M. Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. Nucleic Acids Res. 33, e125–e125 (2005).
[50] IEEE International Roadmap for Devices and Systems, “Mass Digital Storage.” Institute of Electrical and Electronics Engineers, 2023, doi: 10.60627/4HS9-2098.
[51] Andrews, B. I. et al. Sustainability Challenges and Opportunities in Oligonucleotide Manufacturing. J. Org. Chem. 86, 49–61 (2021).
[52] Nguyen, B. H. et al. Architecting Datacenters for Sustainability: Greener Data Storage using Synthetic DNA. 8.
[53] Kjær, K.H., Winther Pedersen, M., De Sanctis, B., et al, A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. Nature 612, 283–291 (2022). https://doi.org/10.1038/s41586-022-05453-y.
[54] Bonnet, J. et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage. Nucleic Acids Res. 38, 1531–1546 (2010).
[55] Lindahl, T. Instability and decay of the primary structure of DNA. Nature 362, 709–715 (1993).
[56] Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. Biochemistry 11, 3610–3618 (1972).
[57] Bruskov, V. I., Malakhova, L. V., Masalimov, Z. K. & Chernikov, A. V. Heat-induced formation of reactive oxygen species and 8-oxoguanine, a biomarker of damage to DNA. Nucleic Acids Res. 30, 1354–1363 (2002).
[58] Matange, K., Tuck, J. M. & Keung, A. J. DNA stability: a central design consideration for DNA data storage systems. Nat. Commun. 12, 1358 (2021).
[59] Kohll, A.X., et al., Stabilizing synthetic DNA for long-term data storage with earth alkaline salts. Chem Commun (Camb), 2020
[60] Antkowiak, P.L., et al., Anhydrous calcium phosphate crystals stabilize DNA for dry storage. Chem Commun (Camb), 2022
[61] Newman, S., Stephenson, A. P., Willsey, M., Nguyen, B. H., Takahashi, C. N., Strauss, K., & Ceze, L. (2019). High density DNA data storage library via dehydration with digital microfluidic retrieval. Nature communications, 10(1), 1706.
[62] Zelikin, A.N., et al., A general approach for DNA encapsulation in degradable polymer microcapsules. ACS Nano, 2007. **1**(1): p. 63-9.

[63] Prince, E., et al., Reversible Nucleic Acid Storage in Deconstructable Glassy Polymer Networks. Journal of the American Chemical Society, 2024.
[64] Korobko, A.V., C. Backendorf, and J.R. van der Maarel, *Plasmid DNA Encapsulation within Cationic Diblock Copolymer Vesicles for Gene Delivery.* J Phys Chem B Condens Matter Mater Surf Interfaces Biophys, 2006. **110**(30): p. 14550-6
[65] Antkowiak, P.L., et al., Integrating DNA Encapsulates and Digital Microfluidics for Automated Data Storage in DNA. Small, 2022. **18**(15): p. e2107381.
[66] Banal, J.L., et al., Random access DNA memory using Boolean search in an archival file storage system. Nature Materials, 2021.
[67] Chen, W.D., et al., Combining Data Longevity with High Storage Capacity—Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles. Advanced Functional Materials, 2019. **29**(28): p. 1901672.
[68] Koch, J., et al., *Preserving DNA in Biodegradable Organosilica Encapsulates.* Langmuir, 2022. **38**(37): p. 11191-11198
[69] Grass, R.N., et al., Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angew Chem Int Ed Engl, 2015. **54**(8): p. 2552-2555.
[70] Strauss, K.e.a.U.A., US2021291134 (A1) Silica Encapsulated DNA on Magnetic Nanoparticles. . 2021.
[71] Zhang, J., C. Hou, and C. Liu, CRISPR-powered quantitative keyword search engine in DNA data storage. Nat Commun, 2024. **15**(1): p. 2376.
[72] Xu, C., et al., Assembly of Reusable DNA Blocks for Data Storage Using the Principle of Movable Type Printing. ACS Appl Mater Interfaces, 2023. **15**(20): p. 24097-24108.
[73] Luo, H., et al., Engineered Living Memory Microspheroid-Based Archival File System for Random Accessible In Vivo DNA Storage. Adv Mater, 2025: p. e2415358.
[74] Luo, Y., et al., Integrated Microfluidic DNA Storage Platform with Automated Sample Handling and Physical Data Partitioning. Analytical Chemistry, 2022.
[75] Ui Jin, L., et al., *DNA Data Storage in Perl.* Biotechnology and Bioprocess Engineering, 2020. **25**(4): p. 607-615.
[76] Bonnet, J., et al., Chain and conformation stability of solid-state DNA: implications for room temperature storage. Nucleic Acids Res, 2010. **38**(5): p. 1531-46.
[77] Coudy, D., et al., Long term conservation of DNA at ambient temperature. Implications for DNA data storage. PLoS One, 2021. **16**(11).
[78] Jahanshahi-Anbuhi, S., et al., *Pullulan encapsulation of labile biomolecules to give stable bioassay tablets.* Angew Chem Int Ed Engl, 2014. **53**(24): p. 6155-8.
[79] Liu, Y., et al., *DNA preservation in silk.* Biomater Sci, 2017.
[80] Soukarie, D., et al., DNA data storage in electrospun and melt-electrowritten composite nucleic acid-polymer fibers. Mater Today Bio, 2024. **24**: p. 100900.
[81] Newman, S., et al., High density DNA data storage library via dehydration with digital microfluidic retrieval. Nature communications, 2019. **10**(1): p. 1706.
[82] Lin, K.N., Volkel, K., Cao, C. et al. A primordial DNA store and compute engine. Nat. Nanotechnol. 19, 1654–1664 (2024). https://doi.org/10.1038/s41565-024-01771-6
[83] Moon, W.C., US2007254294 (A1) Method for Storing Dna by Using Chitosan, and Products Using the Methods. 2007.
[84] Newman, S., et al., High density DNA data storage library via dehydration with digital microfluidic retrieval. Nature communications, 2019. 10(1): p. 1706.
[85] Horton, J.K., P.J. Tatnell, R. Stone, US20170151545A1 Oligonucleotide data storage on solid supports 2017.
[86] Bramlett, B.W. Peck, B.J., US20210142182A1 DNA-based digital information storage with sidewall electrodesTwist Bioscience. 2021.
[87] Ohno, H., US 20070196826 Al Solvent for dissolving nucleic acid, nucleic acid-containing solution and method of preserving nucleic acid. 2007.
[88] Singh, N., et al., Very High Concentration Solubility and Long-Term Stability of DNA in an Ammonium-Based Ionic Liquid: A Suitable Medium for Nucleic Acid Packaging and Preservation. ACS Sustainable Chemistry & Engineering, 2017. **5**(2): p. 1998-2005.
[89] Ping, Z., et al., Towards practical and robust DNA-based data archiving using the yin–yang codec system. Nature Computational Science, 2022.
[90] Sun, J., et al., *Digital information storage on DNA in living organisms.* Medical Research Archives, 2019. **7**(6).
[91] 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).
[92] Maes, A., et al., La révolution de l'ADN: biocompatible and biosafe DNA data storage. bioRxiv, 2022
[93] Hou, Z., et al., "Cell Disk" DNA Storage System Capable of Random Reading and Rewriting. Adv Sci (Weinh), 2024: p. e2305921.

[94] Liu, F., et al., Engineered Spore-Forming Bacillus as a Microbial Vessel for Long-Term DNA Data Storage. ACS Synthetic Biology, 2022.
[95] Boullé, O. and D. Lavenier, *Experimental DNA storage platform*2022
[96] Sun, F., et al., Mobile and Self-Sustained Data Storage in an Extremophile Genomic DNA. Adv Sci (Weinh), 2023: p. e2206201.
[97] Leblanc, J., et al., Fully in vitro iterative construction of a 24 kb-long artificial DNA sequence to store digital information. Biotechniques, 2024. **76**(5): p. 203-215
[98] Yang, S., et al., DNA as a universal chemical substrate for computing and data storage. Nat Rev Chem, 2024
[99] Organick L, Nguyen BH, McAmis R, Chen WD, Kohll AX, Ang SD, et al. An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage. Small Methods. 2021; 5(5): e2001094
[100] Banal, J. L. & Bathe, M. Scalable Nucleic Acid Storage and Retrieval Using Barcoded Microcapsules. ACS Appl. Mater. Interfaces 13, 49729–49736 (2021).
[101] Choi, Y. et al. DNA Micro-Disks for the Management of DNA-Based Data Storage with Index and Write-Once–Read-Many (WORM) Memory Features. Adv. Mater. 32, 2001249 (2020).
[102] Athreya, N., Khandelwal, A., Li, X. & Leburton, J.-P. Electrically Controlled Nanofluidic DNA Sluice for Data Storage Applications. ACS Appl. Nano Mater. 4, 11063–11069 (2021).
[103] Horton, J. K., TATNELL, P. J. & Stone, R. Oligonucleotide data storage on solid supports. (2017).
[104] Strauss, K., Nguyen, B. H., Grass, R. N., Kohll, A. X. C. & Chen, W. Dna data storage on two-dimensional support material. (2020).
[105] Newman, S. et al. High density DNA data storage library via dehydration with digital microfluidic retrieval. Nat. Commun. 10, 1706 (2019).
[106] Banal, J. L. et al. Random access DNA memory using Boolean search in an archival file storage system. Nat. Mater. 1–9 (2021) doi:10.1038/s41563-021-01021-3.
[107] Lemaire, S. Crozes, P., Xu, Z., Maes, A. & Peillet, J. L. Biocompatible nucleic acids for digital data storage. (2021).
[108] Peck, B. J. Nucleic acid based data storage. (2018).
[109] Lim, C. K., Nirantar, S., Yew, W. S. & Poh, C. L. Novel Modalities in DNA Data Storage. Trends Biotechnol. 39, 990–1003 (2021).
[110] Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. Nat. Mater. 15, 366–370 (2016).
[111] Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **337**, 1628–1628 (2012).
[112] Bornholt, J. *et al.* Toward a DNA-Based Archival Storage System. *IEEE Micro* **37**, 98–104 (2017).
[113] Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and Error-Free DNA-Based Data Storage. *Sci. Rep.* **7**, (2017).
[114] Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. Nat. Commun. 11, 616 (2020).
[115] Liu, (2023). "Sustainable DNA Data Storage on Cellulose Paper." Small Methods: e2201610
[116] Ma (2022). "Magnetic Microsphere/Silica Nanoparticle Composite Structures for Switchable DNA Storage." ACS Applied Nano Materials.
[117] Mao, (2023). "Metal-Organic Frameworks in Microfluidics Enable Fast Encapsulation/Extraction of DNA for Automated and Integrated Data Storage." ACS Nano 17(3): 2840-2850.
[118] Narvaez Villarrubia, (2022). "Long-term stabilization of DNA at room temperature using a one-step microwave assisted process." Emergent Mater 5(2): 307-314.
[119] Shen, (2023). "Magnetic Bead Spherical Nucleic Acid Microstructure for Reliable DNA Preservation and Repeated DNA Reading." ACS Synthetic Biology.
[120] 3) Lin, K.N., Volkel, K., Tuck, J.M. et al. Dynamic and scalable DNA-based information storage. Nat Commun 11, 2981 (2020). https://doi.org/10.1038/s41467-020-16797-2
[121] Bee, C. et al. Molecular-level similarity search brings computing to DNA data storage. Nat. Commun. 12, 4764 (2021).
[122] Tomek, K. J., Volkel, K., Indermaur, E. W., Tuck, J. M. & Keung, A. J. Promiscuous molecules for smarter file operations in DNA-based data storage. Nat. Commun. 12, 3518 (2021).
[123] SNIA. DNA Data Stability Evaluation Method for DNA Data Storage Containment Systems v1.0 r0 (https://www.snia.org/dnastability)
[124] 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).
[125] Sanger Sequencing | AAT Bioquest. https://www.aatbio.com/catalog/sanger-sequencing.
[126] ThermoFisher 3730xl data sheet. https://assets.thermofisher.com/TFS-Assets/GSD/Reference-Materials/3730xl-instrument-comparison-white-paper.pdf

[127] Lee, H. et al. Ultra high-throughput whole-genome methylation sequencing reveals trajectories in precancerous polyps to early colorectal adenocarcinoma. 2022.05.30.494076 Preprint at https://doi.org/10.1101/2022.05.30.494076 (2022).
[128] Simmons, S. K. et al. Mostly natural sequencing-by-synthesis for scRNA-seq using Ultima sequencing. Nat. Biotechnol. 41, 204–211 (2023).
[129] Almogy, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. 2022.05.29.493900 Preprint at https://doi.org/10.1101/2022.05.29.493900 (2022).
[130] Korlach, J. et al. Chapter 20 - Real-Time DNA Sequencing from Single Polymerase Molecules. in Methods in Enzymology (ed. Walter, N. G.) vol. 472 431–455 (Academic Press, 2010).
[131] Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13, 278–289 (2015).
[132] Eid, J. et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323, 133–138 (2009).
[133] Wang, Y., Yang, Q. & Wang, Z. The evolution of nanopore sequencing. Front. Genet. 5, (2015).
[134] Tyler, A.D., Mataseje, L., Urfano, C.J. et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci Rep 8, 10931 (2018). https://doi.org/10.1038/s41598-018-29334-5
[135] Xue, L. et al. Solid-state nanopore sensors. Nat. Rev. Mater. 1–21 (2020) doi:10.1038/s41578-020-0229-6.
[136] Ahmadi, A. G., Peng, Z., Hesketh, P. J. & Nair, S. Wafer-scale process for fabricating arrays of nanopore devices. J. MicroNanolithography MEMS MOEMS 9, 033011 (2010).
[137] Chien, C.-C., Shekar, S., Niedzwiecki, D. J., Shepard, K. L. & Drndić, M. Single-Stranded DNA Translocation Recordings through Solid-State Nanopores on Glass Chips at 10 MHz Measurement Bandwidth. ACS Nano 13, 10545–10554 (2019).
[138] Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. Kaikai Chen, Jinglin Kong, Jinbo Zhu, Niklas Ermann, Paul Predki, and Ulrich F. Keyser. Nano Letters 2019 19 (2), 1210-1215, DOI: 10.1021/acs.nanolett.8b04715
[139] Sequencing by Expansion (SBX) – a novel, high-throughput single-molecule sequencing technology
Mark Kokoris et al, bioRxiv 2025.02.19.639056; doi: https://doi.org/10.1101/2025.02.19.639056
[140] Shendure, J. et al. DNA sequencing at 40: past, present and future. Nature 550, 345–353 (2017).
[141] Perešíni, P., Boža, V., Brejová, B. & Vinař, T. Nanopore base calling on the edge. Bioinformatics 37, 4661–4667 (2021).
[142] DRAGEN Secondary Analysis | Variant calling and genomics software. https://www.illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html.
[143] Home | International Gene Synthesis Consortium. International Gene Synthesis Consortium | The Promotion of Biosecurity https://genesynthesisconsortium.org/
[144] https://www.federalregister.gov/documents/2010/10/13/2010-25728/screening-framework-guidance-for-providers-of-synthetic-double-stranded-dna
[145] https://aspr.hhs.gov/legal/synna/Documents/SynNA-Guidance-2023.pdf
[146] https://ibbis.bio/wp-content/uploads/2024/02/IBBIS-Common-Mechanism-FAQ.pdf