ABSTRACT

MATANGE, KARISHMA. Synthesis and Storage Stability of DNA in DNA Data Storage Systems. (Under the direction of Dr. Albert Keung).

Data storage in DNA is a rapidly evolving technology that could be a transformative solution for the rising energy, materials, and space needs of modern information storage. Given that the information medium is DNA itself, its efficient synthesis and stability under different storage and processing conditions will fundamentally impact and constrain design considerations and data system capabilities. Here we introduce an enzymatic method for DNA synthesis of short oligos and discuss its use in production of long strands. We also analyze the storage conditions, molecular mechanisms, and stabilization strategies influencing DNA stability and pose specific design configurations and scenarios for future systems that best leverage the considerable advantages of DNA storage.

Synthesis and Storage Stability of DNA in DNA data storage systems

by
Karishma Rajeev Matange

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Chemical Engineering

Raleigh, North Carolina
2021

APPROVED BY:

Dr Albert Keung. Chair of Advisory Committee	Dr Carol Hall
Dr Nathan Crook	

BIOGRAPHY

Karishma Rajeev Matange was born in October 1997 to Rajeev and Bhagyashree Matange in Nashik, India. Growing up in Pune with sister, Kavita, she attended City International school. She completed her undergraduate education from University of Pune before moving to the United State of America for a master’s degree in the Chemical and Biomolecular Engineering Department at North Carolina State University. Here, she found an excellent fit in the research group of Dr. Albert Keung working on a project in collaboration with Dr. James Tuck from the Electrical and Computer Engineering Department at NC State.

ACKNOWLEDGMENTS

I am a deeply grateful for everyone who has been a part of my master's journey. First, I would like to thank my adviser Dr. Albert Keung for his mentorship and guidance. I am also thankful to the DNA data storage team and all Keung Lab members for their collaboration, help and support. A special thanks to my committee members Dr. Crook and Dr. Hall for agreeing to serve on my committee. Last but not the least, I want to thank my friends and family for their constant love and unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction to DNA data storage systems

In 1964, just 7 years after the discovery of the structure of DNA[1], Wiener and Neiman discussed the potential density advantage of using nucleic acids as a form of memory storage[2]. Over half a century later, advancements in our understanding of the properties of DNA have confirmed its high theoretical information density of nearly 455 billion GB of data per gram[3], ~6 orders of magnitude greater than even the most advanced magnetic tape storage systems[4]. DNA also provides a host of other unique potential advantages including highly parallelized computation within the storage system itself [5,6], low energy requirements[7-9], rapid high-capacity transportation of data[10], potentially longer lifetimes, and stabilities of decades or centuries compared to conventional media which are replaced every 3–5 years, as well as ease of replication[11] by molecular biology approaches to ward off degradation. While it has taken more than just the decade Wiener predicted, a large body of knowledge in molecular biology[12,13] and computer and information systems[14,15] has been developed in the intervening period that has set the foundation for the recent interest and investment in DNA information storage technologies.

DNA data storage systems work by encoding digital information into DNA sequences. Many physical copies of each strand of approximately 150-200bp length are synthesized. The strands are physically conditioned and stored in environments on the basis of the access requirements. In order to access or retrieve the data, the strands are selectively accessed, the sequences are read and decoded back into binary data as illustrated in figure **1**. This work focusses on synthesis and storage stability of DNA molecules suitable for DNA data storage systems.

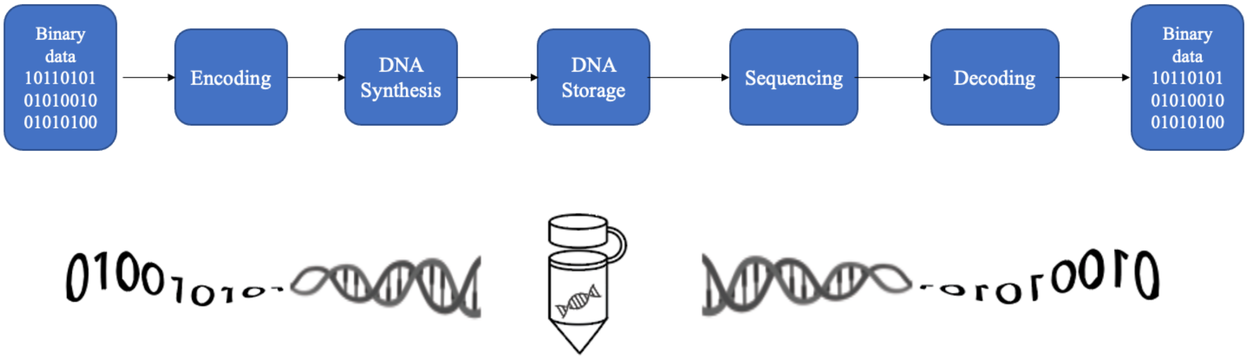

Binary data 10110101 01010010 01010100
Encoding
DNA Synthesis
DNA Storage
Sequencing
Decoding
Binary data 10110101 01010010 01010100
0100 1010
0010 1010

Figure 1: Working of DNA information storage systems. Binary data is encoded into a string of DNA bases (A, C, G, T) to form a sequence. Several encoding strategies may be used on the basis of error probabilities and types of damage. The strands are then stored in specific environments depending upon the desired timeframe of storage required. For retrieval, strands are selectively accessed and decoded based on decoding schemes required for various modes of damage to obtain the original binary data.

Chapter 1

Continuous Enzymatic synthesis of DNA oligos

1.1 Introduction:

DNA as a storage medium shows promise for future data storage systems by providing high data density, energy efficiency and durability [4,16]. An important part of the system is the synthesis of DNA strands that contain the data. However, current methods to synthesize DNA are expensive, slow, and generate byproducts toxic to the environment. Here we will briefly discuss current methods that are commonly used for DNA synthesis along with their advantages and limitations. We will then describe the goal of this thesis chapter to address these limitations.

Chemical phosphoramidite synthesis has been the most widely used technique for almost 35 years. It uses a solid support to immobilize the first phosphoramidite and facilitates the cyclic addition of new bases to build sequences of up to 200 base pairs with a cost ranging between \$0.005 to \$0.017 per base[17]. Error rates of ~1 in 200nt are common, mainly due to the depurination of adenosine caused by the use of strong acids (trichloroacetic acid and dichloroacetic acid) and prevent the production of long error-free oligos. Moreover, yields as high as 99% are required for each turn of the cycle in order to obtain a final yield of 13% for a 200-nt synthesis[18]. Such limitations of chemical methods motivate the study of enzymatic approaches to DNA synthesis in aqueous conditions.

An enzyme belonging to the DNA polymerase X family- Terminal deoxynucleotidyl transferase (TdT), has been studied in the context of template-independent DNA synthesis. This method adds dNTPs to the 3' termini of DNA strand and reduces synthesis costs to $4.38/MB due to the reusability of enzymes. However, raw synthesis only produces 19% of perfect strands due to errors caused by mismatches, insertion errors and missing nucleotides. Although yield can be increased to 89% on further processing, high error-probability in DNA data storage systems means low efficiency of storage[19].

Enzymatic methods have clear advantages over chemical synthesis, but sequential addition of bases can be slow. Hence, block-based assembly methods like Ligase cycling [20], Gibson assembly[21] and CPEC[22] have an obvious advantage. Assembly reactions typically make use of overhangs coupled with specific enzymes that bind short DNA strands to create longer sequences at costs less than $5 per reaction. Since these methods simply combine already created strands, they have error rates as low as 1 per 50 DNA molecules joined[21].

A feasible, environmentally friendly alternative for base-by-base synthesis that specifically caters to the requirements of DNA information storage systems is required. In this study, we present a template-based enzymatic method that allows continuous synthesis of short DNA strands that have low error probability and cost as little as $1.1/PB. We showcase a system that produces 60nt long strands we are calling 'codewords' and their subsequent selective assembly into two separate 280nt length sequences.

1.2 Materials and Methods:

1.2.1 Oligonucleotides, enzymes and other reagents:

All oligonucleotides were purchased from Genewiz Inc. Nicking enzymes and polymerases including Nt.AlwI (NEB, #R0627S), Klenow fragment (3' -> 5' exo) (NEB, #M0212S), Large (Klenow) Fragment (NEB, #M0210S), phi29 DNA polymerase (NEB, #M0269S), Deep Vent DNA polymerase (NEB, #M0258S) and Q5 High-fidelity DNA Polymerase (NEB, #M0515) were acquired from New England Biolabs (NEB).

1.2.2 Hairpin design: The Hairpin was designed in a way that it could self-prime with a binding interaction of 15nt. The double-stranded region contains a recognition site for the Nickase (NEB, #R0627S) (GGATCNNNN) and is Biotinylated at the loop. Details of the sequence can be found in table **1**. To create the hairpin from a single strand, 2ug of the oligo was taken in a PCR tube and reconstitute with 50µl of 1X TE buffer. A slow annealing protocol was run on a thermocycler. First temperature at 95C for 3 minutes followed by a slow ramp down of 2C per minute until temperature reached 25C.

1.2.3 Template: Hairpin structure specific for synthesis of a particular sequence ('codeword') was ordered with an internal Biotin modification. 22µl solution of streptavidin magnetic beads (NEB #S1420S) was pipetted into a low retention tube and washed twice with 100µl Bind-wash buffer (containing 20nM Tris-HCl pH7.4, 2M NaCl, 2mM EDTA pH8), each time placing the tube on a magnetic stand to separate out the beads. The beads were then added to 2.67µl (Concentration: 1E12 strands/µl) of Hairpin and 50µl bind-wash buffer before being incubated at room temperature on a rotisserie for 30 minutes. After further washing twice with bind-wash buffer, the bead-hairpin ('template') was eluted in 25.5µl of water.

1.2.4 Codeword synthesis: 0.5µl DNA polymerase, 0.5µl Nickase (NEB, #R0627S), 2.5µl polymerase buffer and dATP (NEB, #N0440S), dCTP (NEB, #N0441S), dGTP (NEB, #N0442S), dTTP (NEB, #N0443S) each at 10mM were added to 25.5µl of template, bringing the total reaction volume up to 30µl, and incubated at 25C for 1 hour. Following this, the template was magnetically separated using a magnetic stand, and the supernatant was collected for analysis.

1.2.5 Assembly reactions:

1.2.5.1 MOEPCR assembly: 7E11 strands each of six single-stranded codewords of 60 bp length with 20bp complementary overlapping regions were mixed together along with 10ul 5X Q5 reaction buffer, 1µl 10mM dNTPs and 0.5µl Q5 enzyme (NEB #M0515). The reaction mixture was eluted to 50µl with water and cycled through temperatures as indicated in table **2a**. This step was followed by a PCR amplification. 1ul each of forward and reverse primers (P3, P4, Concentration: 1E13 strands/ul) were added to the 50ul assembly reaction mixture along with 1ul 5X Q5 reaction buffer, 1µl 10mM dNTPs and 0.5ul Q5 enzyme. The sample was eluted with water up to 60ul and incubated according to table **2b**.

1.2.5.2 Amplification using codeword as primer: 0.848µl of DNA template S2 (Concentration: 1E12 strands/µl) was added to 5µl 5X Q5 PCR Buffer, 0.25µl Q5 polymerase, 0.5µl 10mM dNTPs, 1.06µl reverse primer (P4, Concentration: 1E13 strands/µl) and 12.5 µl of produced codeword as forward primer (codeword1, Concentration: 8.47E11 strands/ul). The sample was eluted with water to 25µl and run at normal PCR temperatures as mentioned in figure **8**.

Products from these two reactions were then purified using gel electrophoresis and sent for Genewiz Sanger sequencing.

1.2.6 Analysis: Samples were diluted to fit the recommended concentration range of the fragment analyzer of 5pg/µl – 500 pg/µl. 2µl of each diluted sample was pipetted into an unused well along with 22µl of the High Sensitivity NGS Diluent Marker (1bp-6000bp, part #DNF-373) and run on the 12 capillary Fragment Analyzer (Advanced Analytical) using the High Sensitivity NGS Fragment Analysis Kit (Advanced Analytical, DNF-474).

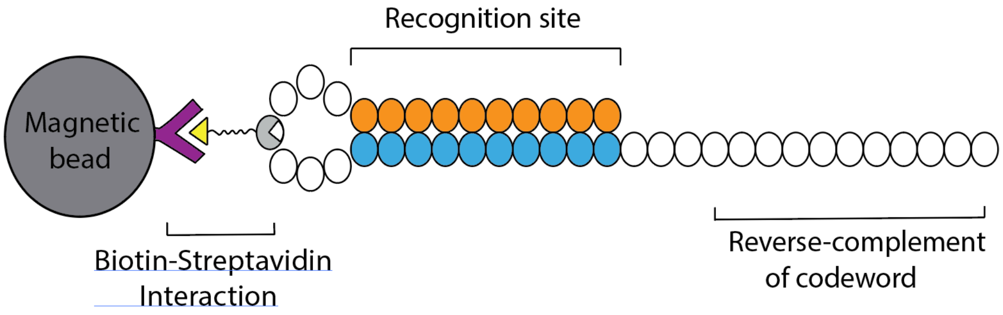

Recognition site
Magnetic bead
Biotin-Streptavidin Interaction
Reverse-complement of codeword

Figure 2: Template. 90 nt sequence with a BIOdT on the 75th nucleotide. Top part of the strand forming the double-stranded region is a recognition site (GGATCNNNN) for the nickase Nt.alw1. The BIOdT in the loop interacts with streptavidin on the magnetic beads.

Table 1: Sequences of strands, codewords and primers

	Sequence	Length (nt)	Modification
Hairpin	AGATGCGAGCTGTCATCTCGAGTCTACGTACGTACGTACGGAGTGCTAACTGCGTACCTGTCTCGATCCCTCGCTTGCGAGGGATCGAGA	90	BIOdT on 75th nucleotide
Codeword 1	CAGGTACGCAGTTAGCACTCCGTACGTACGTACGTAGACTCGAGATGACAGCTCGCATCT	60	
P3	CAGGTACGCAGTTAGCACTC	20	
P4	CGTGGCAATATGACTACGGA	20	
S2	CAGGTACGCAGTTAGCACTCCGTACGTACGTACGTAGACTCGAGATGACAGCTCGCATCTACGAGCTCGAGATGACACAGAGTATCGCATCTACGACACAGTCTCTCGCGAGCTAGAGATGAGTGATCGAGCTCTGCTCGCTCTCGTGACAGTCGACTCGAGAGTGCAGAGCAGACTCATTCCGTAGTCATATTGCCACG	200	
Codeword 2	TGTGTCGTAGATGCGATACTCTGTGTCATCTCGAGCTCGTAGATGCGAGCTGTCATCTCG	60	
Codeword 3	AGTATCGCATCTACGACACAGTCTCTCGCGAGCTAGAGATGAGTGATCGAGCTCTGCTCG	60	
Codeword 4	ATGAGTCTGCTCTGCACTCTCGAGTCGACTGTCACGAGAGCGAGCAGAGCTCGATCACTC	60	
Codeword 5	AGAGTGCAGAGCAGACTCATGAGTATCGCATCTGACTCTCATCGCACTCGCATCGATGAG	60	
Codeword 6	CGTGGCAATATGACTACGGACGATGAGAGTAGAGAGTCGACTCATCGATGCGAGTGCGAT	60	
Full length assembled strand	CAGGTACGCAGTTAGCACTCCGTACGTACGTACGTAGACTCGAGATGACAGCTCGCATCTACGAGCTCGAGATGACACAGAGTATCGCATCTACGACACAGTCTCTCGCGAGCTAGAGATGAGTGATCGAGCTCTGCTCGCTCTCGTGACAGTCGACTCGAGAGTGCAGAGCAGACTCATGAGTATCGCATCTGACTCTCATCGCACTCGCATCGATGAGTCGACTCTCTACTCTCATCGTCCGTAGTCATATTGCCACG	260	

1.3 Results and discussion:

As depicted in Figure **3**, our method works by the consecutive and repeated action of a polymerase and a nicking endonuclease. First, the polymerase binds to the end of the double strand and adds DNA bases to the single strand one-by-one in the 5’ to 3’ direction. The nickase then nicks the strand after the recognition site and in the next cycle of polymerization, the polymerase pushes the 60 nt long strand out, producing a single-stranded codeword. This cycle repeats.

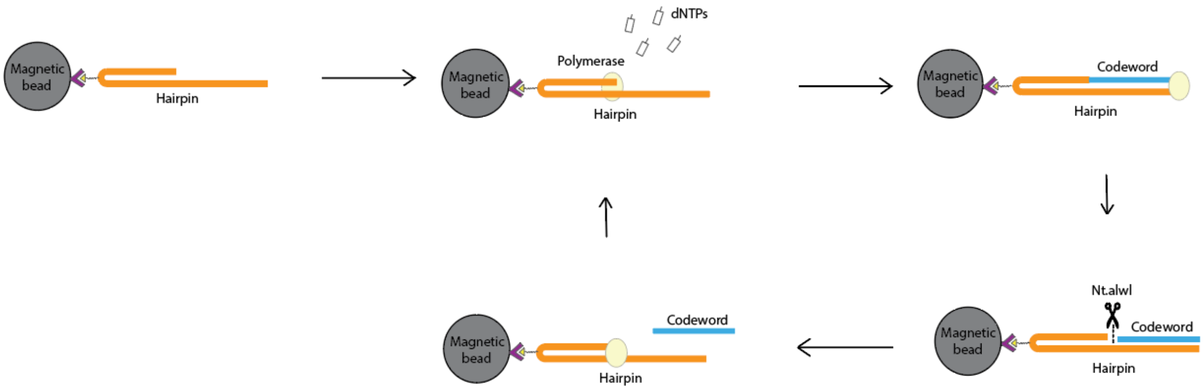
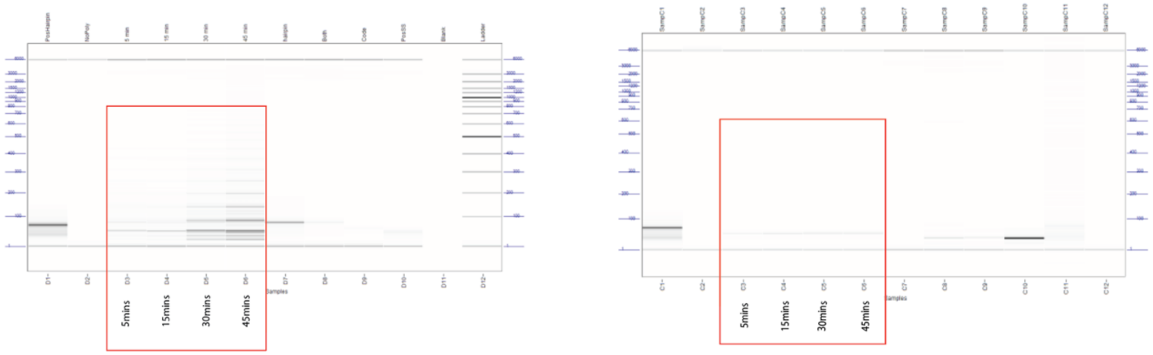


Figure 3: **Continuous synthesis of codeword**

Initially, in order to demonstrate proof of concept, hairpin was added to the reaction mixture and incubated for 5, 15, 30, and 45 minutes. Different times were chosen so as to provide enough time for the system to produce codeword and to determine the point at which codeword concentration plateaus i.e. no more codeword is produced. The results, highlighted in figure **4a**, showed strong bands at 60bp length confirming the production of codeword. However, a striated banding pattern suggested some non-specific interaction. We hypothesized that this might be due to the hairpin unfolding and binding to other hairpin structures to form longer strands. Therefore, we attached the hairpin to a magnetic bead so that we could avoid non-specific interactions while also being able to control the hairpin in the reaction magnetically and repeated the experiment. As shown in figure **4b**, these samples exhibit a single band at 60bp length indicating a higher concentration of pure codeword being produced with no non-specific binding. Although initial experiments were successful, optimization of the components of this system was essential in the improvement of efficiency.



(a) **(b)**

Figure 4: **Proof of concept experiments**; **(a) Hairpin without attached magnetic beads**; D1: Hairpin in water, D2: Polymerase + nickase + dNTPs, D3: Reaction mixture incubated for 5 minutes, D4: Reaction mixture incubated for 15 minutes D5: Reaction mixture incubated for 30 minutes D6: Reaction mixture incubated for 45 minutes, D7: Polymerase + hairpin + dNTPs, D8: Polymerase + hairpin + dNTPs+ codeword, D9: Polymerase + dNTPs+ codeword, D10: Codeword in water, D11: Blank, D12: DNA ladder. Samples D1 and D10 serve as positive controls for hairpin and codeword respectively. D2 is a negative control for hairpin in the reaction mixture. D3, D4, D5 and D6 represent reaction mixtures incubated for 5, 15, 30 and 45 minutes respectively while samples D7, D8 and D9 test for any non-specific interactions that may occur in reactions. **(b) Hairpin attached to magnetic beads;** C1: Hairpin in water, C2: Polymerase + nickase + dNTPs, C3: Reaction mixture incubated for 5 minutes, C4: Reaction mixture incubated for 15 minutes C5: Reaction mixture incubated for 30 minutes C6: Reaction mixture incubated for 45 minutes, C7:

Figure 4 (continued)

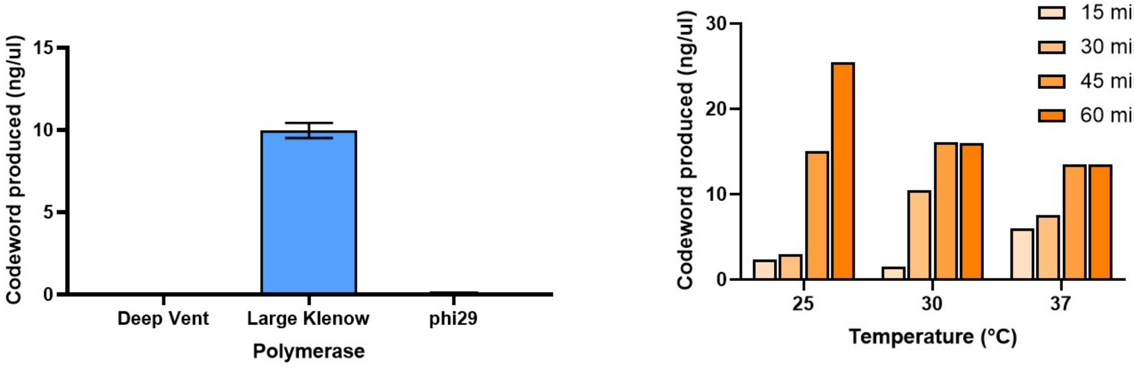
Polymerase + dNTPs, C8: Polymerase + dNTPs+ codeword, D9: Polymerase + dNTPs+ codeword, D10: Codeword in water, D11: Blank, D12: DNA ladder. Samples D1 and D10 serve as positive controls for hairpin and codeword respectively. D2 is a negative control for hairpin in the reaction mixture. D3, D4, D5 and D6 represent reaction mixtures incubated for 5, 15, 30 and 45 minutes respectively while samples D7, D8 and D9 test for any non-specific interactions that may occur in reactions. Template was magnetically separated out from samples C3, C4, C5, C6, C7 and C8 before running on the fragment analyzer.

1.3.1 Optimization:

1.3.1.1 Polymerase:

Synthesis methods used in DNA data storage systems require production of specific and accurate sequences. Therefore, any tailing in strands would be undesirable. Tailing is an enzymatic process in which a non-templated nucleotide is added to the 3' end of a blunt, double-stranded DNA molecule. Often, the use of some polymerases produces an A-tail which is undesirable. We selected three polymerases that produce blunt ends and compared them on the basis of their ability to produce codeword when incubated for 30 minutes. The polymerases were the Deep Vent DNA polymerase, Large Klenow fragment and the phi29 polymerase. Samples were cycled between the incubation temperature of the polymerase and that of Nickase (37C). As shown in figure 3a, Large (Klenow) Fragment produced maximum yield. Low yield in samples containing Deep Vent DNA polymerase could be attributed to a high incubation temperature (75 C) which would disrupt biotin-streptavidin interactions and cleave hairpin off the beads. On the other hand, low yield in samples containing phi29 DNA polymerase could be due to its low strand displacement capability as shown in figure **5a**.

Naturally, in the next step, we tested the optimum temperature and incubation time for Large (Klenow) Fragment for this system. We tested the codeword production at constant temperatures of 25C, 30C and 37C for 15, 30, 45 and 60 minutes. Samples incubated at 25C for 60 minutes produced maximum codeword yield, while samples kept at 30C and 37C seemed to plateau after 45 minutes as shown in figure **5b**. Therefore, incubation temperature and time of 25C and 60 minutes was used for further experimentation.



(a) **(b)**

Figure 5: (a) Relationship between codeword produced and the type of polymerase used; Samples were cycled 30 minutes between 75C and 37C, 25C and 37C and 30C and 37C for Deep vent polymerase, Large Klenow fragment and phi29 polymerase respectively for. Samples with Large Klenow fragment produced an average of 9.9ng/ul of codeword while samples containing Deep vent and phi29 polymerases produced 0ng/ul and 0.09ng/ul codeword respectively. **(b) Relationship between codeword produced and incubation times and temperatures.** Samples were subjected to incubation temperatures of 25C, 30C and 37C for 15, 30, 45 and 60 minutes. Samples incubated at 25C for 1 hour produced maximum codeword of concentration of 25ng/ul.

1.3.1.2 Magnetic beads:

In order to optimize the binding interaction between hairpin and the beads, we tested varying ratios of bead binding sites to hairpins. Initial experiments used a 10:1 ratio and the calculations for amount of bead added are given below. As shown in figure **6**, higher ratios appear to give better codeword yield. A possible explanation for low concentration for codeword in lower ratios could be crowding. Crowding of hairpin at bead binding sites would result in ineffective Biotin-streptavidin interactions and lead to more unbound template.

Bead binding site calculation

1 mg of bead = 500 pmol binding sites = $500 pmol * \frac{6.022E23}{1E12} \frac{bindingsites}{pmol}$ = 3.01E14 binding sites

Amount of hairpin added to each reaction = 2.67E12 strands

Amount of bead required for each reaction = $2.67E12\ binding\ sites * \frac{10}{3.01E14} \frac{mg}{binding\ sites}$

(maintaining a 10:1 ratio) = 0.088 mg

Density of beads = 4 mg/ml

Amount of bead required for each reaction = $\frac{0.088\ mg}{4\ mg/ul}$ = 22 ul

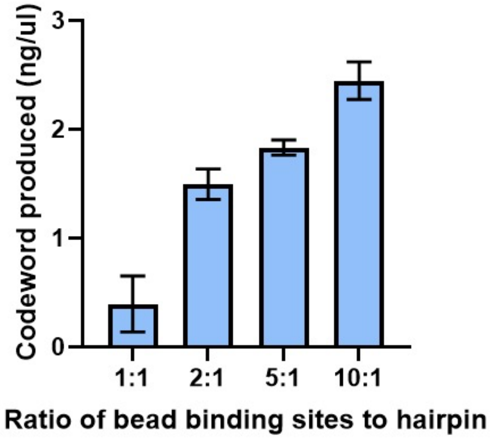


Figure 6: Bead optimization. Beads were linked to hairpin according to bead binding sites to hairpin ratios of 1:1, 2:1, 5:1 and 10:1. Samples were run for 1 hour at 25C using Large Klenow fragment. As ratio increases, the concentration of codeword produced increases. A considerable rise is observed between samples with ratios 1:1 and 2:1 with the latter producing 5 times as much codeword as the former. This jump may be utilized for optimization.

1.3.2 Assembly

Once the production of 60nt length sequences was confirmed, it was important to test whether the codeword sequence was accurate and could be used in assembly reactions to build longer strands. This was done in 2 ways. (1) Codeword was used in a Multiple overlap extension PCR or MOEPCR assembly reaction using 6 strands of 60nt length. A MOEPCR assembly[23] works by having strands containing complementary overlaps run on a simple PCR protocol. The strands bind through the complementary overlaps and the PCR fills in the gaps to create a full-length strand. In this case, we had six 60nt strands with 20 nt complementary overlaps that assembled to form a product of 260bp length as illustrated in figure **7**. (2) Codeword was used as a primer to amplify a larger strand with a simple PCR. In this case, as shown in figure **8**, we amplified a strand

S2 of 200 bp length using codeword as the forward primer and a reverse primer P4. Products from both the reactions were Sanger sequenced and the results verified the accuracy of the codeword sequence and its ability to participate in assembly reactions.

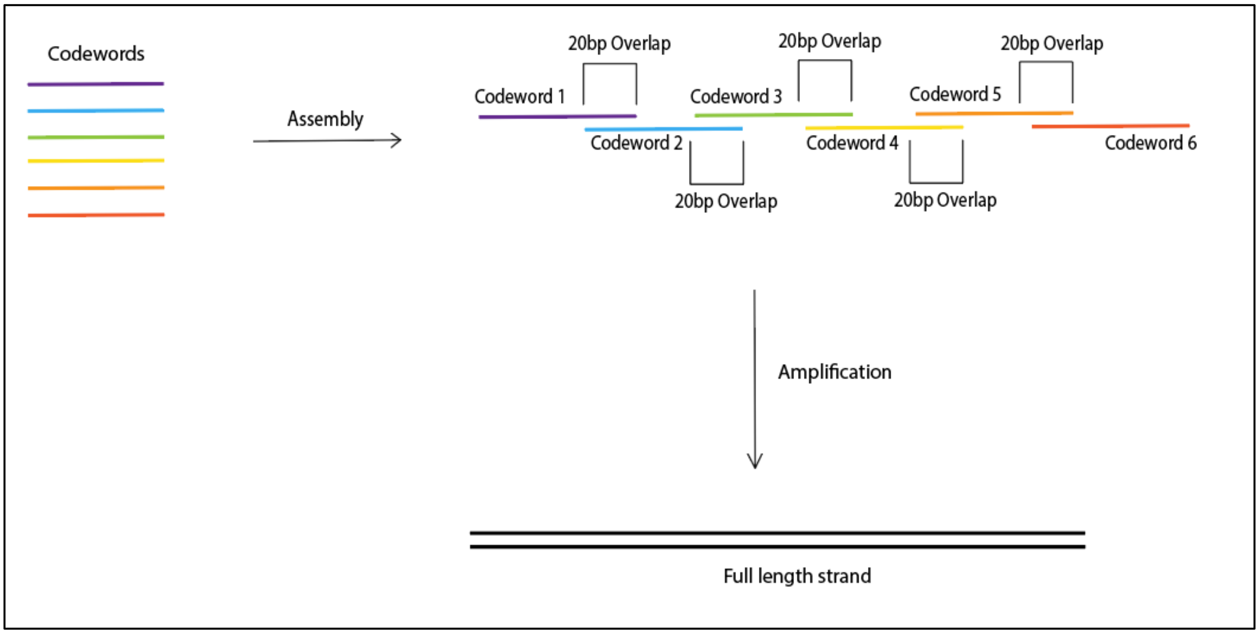
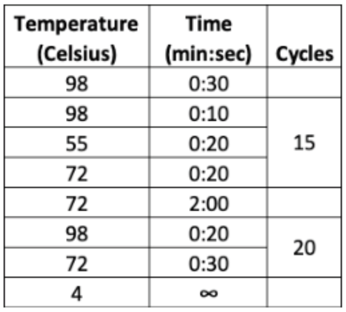
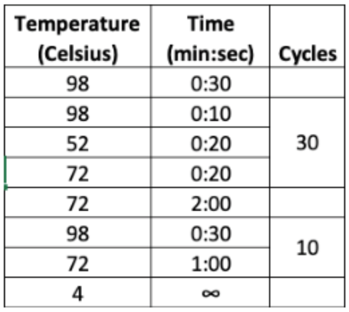


Figure 7: Illustration of a MOEPCR assembly reaction. Codeword 1 represents the 60 nt sequence we generated using the enzymatic method. The assembly and amplification step involved incubations as recorded in table **2a** and **2b** respectively. Sequences of the codewords can be found in table **1**.



Temperature (Celsius)	Time (min:sec)	Cycles
98	0:30	
98	0:10	15
55	0:20	
72	0:20	
72	2:00	
98	0:20	20
72	0:30	
4	∞	

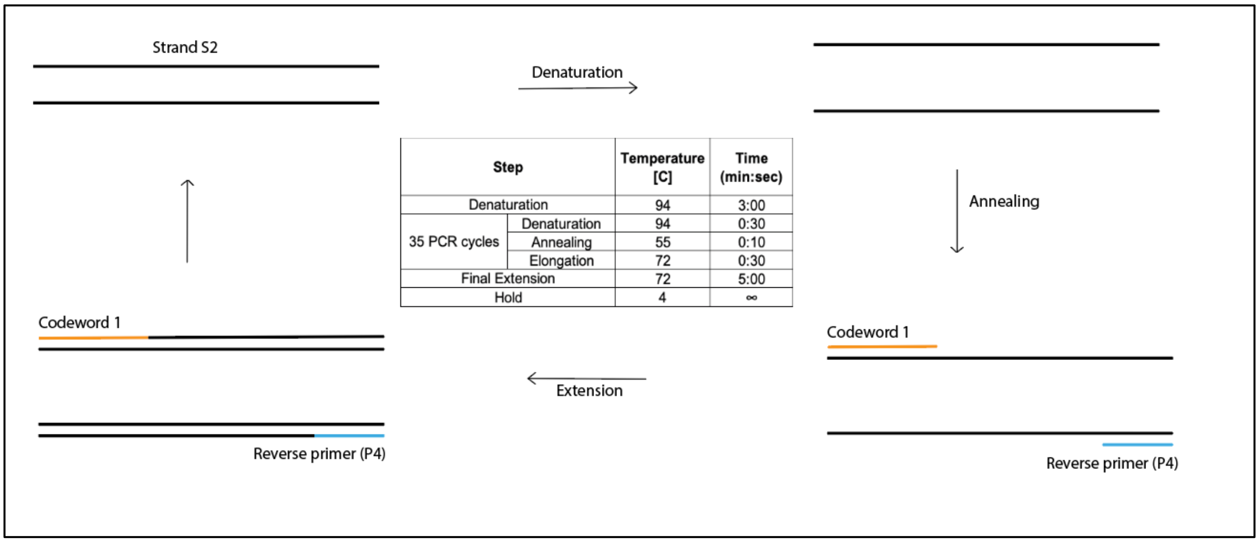
(a)



Temperature (Celsius)	Time (min:sec)	Cycles
98	0:30	
98	0:10	30
52	0:20	
72	0:20	
72	2:00	
98	0:30	10
72	1:00	
4	∞	

(b)

Table 2: (a) Assembly protocol for MOEPCR assembly reaction. The protocol design was such that it allowed for the annealing of the overlaps followed by the extension of the strand to form a full-length strand[23]. **(b) Amplification protocol for MOEPCR assembly reaction.** The amplification protocol was similar to that of a simple PCR reaction protocol. The samples were cycled through the denaturation, annealing and extension steps to amplify a strand of length 200 bp.



Step		Temperature [C]	Time (min:sec)
Denaturation		94	3:00
35 PCR cycles	Denaturation	94	0:30
	Annealing	55	0:10
	Elongation	72	0:30
Final Extension		72	5:00
Hold		4	∞

Figure 8: Amplification of strand S2 using codeword as forward primer. S2 was amplified using codeword and a short 20bp sequence P4 as the forward and reverse primer respectively.

1.4 Future Work:

The advancement of DNA information storage systems is heavily dependent on the development of commercially viable and environmentally friendly methods to write DNA. For our synthesis method to be feasible, much more can be done to improve the efficiency of this system. a 50-fold improvement must be made in the current system's ability to produce codeword for the method to be cost-comparable to the current industry standard; phosphoramidite chemistry (table **3**). We can already optimize bead concentration on the basis of figure **6**. Furthermore, an increase in amount of template and addition of optimal quantities of enzymes and dNTPs would further improve the efficiency of the system as indicated by table **4**.

However, the first step in optimization should be the study of reusability. Reusability of template and enzymes would substantially reduce cost and facilitate the continuous synthesis of oligos without a need for replenishment. As a step in this direction, we tested the sustainability of template and enzymes by running the standard reaction for a long duration while periodically replenishing dNTPs and recorded the concentration of codeword produced every hour. As shown in figure **9**, the codeword concentration kept rising which demonstrates the sustainability of the system for duration up to 6 hours. However, the system must be studied further to determine the limiting component and the rate of degradation of hairpin and enzymes to establish reusability. When optimized, this method could potentially be the transformative solution the field of DNA synthesis requires in terms of cost-efficiency, environmental friendliness and low error probabilities.

From a much broader perspective, the future lies in automation. There is a need for DNA writing and assembly to be standardized to eliminate human error and improve speed of production. From automation through liquid handling robots by Catalog Technologies Inc. to direct printing of DNA through printers in 2D[24] and 3D [25] ideas are rapidly evolving to fit the use. A system of continuous synthesis and assembly of DNA fragments, perhaps through a microfluidic device that is capable of precise control and manipulation of small volumes would greatly improve the speed and feasibility of DNA information systems for large scale data storage.

Table 3: Cost comparison of our synthesis method to synthesis by Phosphoramidite chemistry. The calculations for this analysis are detailed in Appendix A

Method	**Cost per Petabyte**
Phosphoramidite method	$0.024
Our Synthesis method	$1.1

Table 4: Cost calculated for a single run of a 30μl reaction.

Sr. No.	**Components**	**Cost**		**Standard run contains (ul)**	**Cost per run**
		Amount	**Quantity**		
1	Hairpin	$156.92	50nmol	2.67	**$0.01**
2	Bead (10:1)	$334	5000 ul	22.2	**$1.48**
3	Polymerase	$258.30	200 ul	0.5	**$0.65**
4	Nickase	$72	50 ul	0.5	**$0.72**
5	dNTPs	$53.89	500 ul	1	**$0.11**
6	Buffer			2.5	
7	Water			0.63	
	Total			**30**	**$2.97**

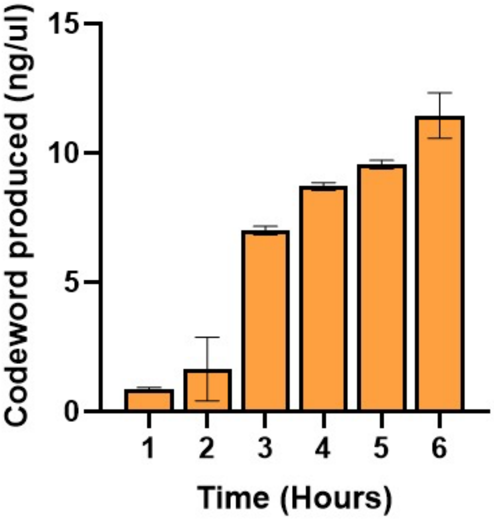


Figure 9: Reusability. Six identical sets of samples were incubated at 25C. Every hour, one set of samples were removed from incubation for analysis and 0.5µl dNTPs were added to each of the rest of the samples. This was repeated, the longest sample being incubated for 6 hours.

Chapter 2

DNA stability: a central design consideration for DNA data storage systems

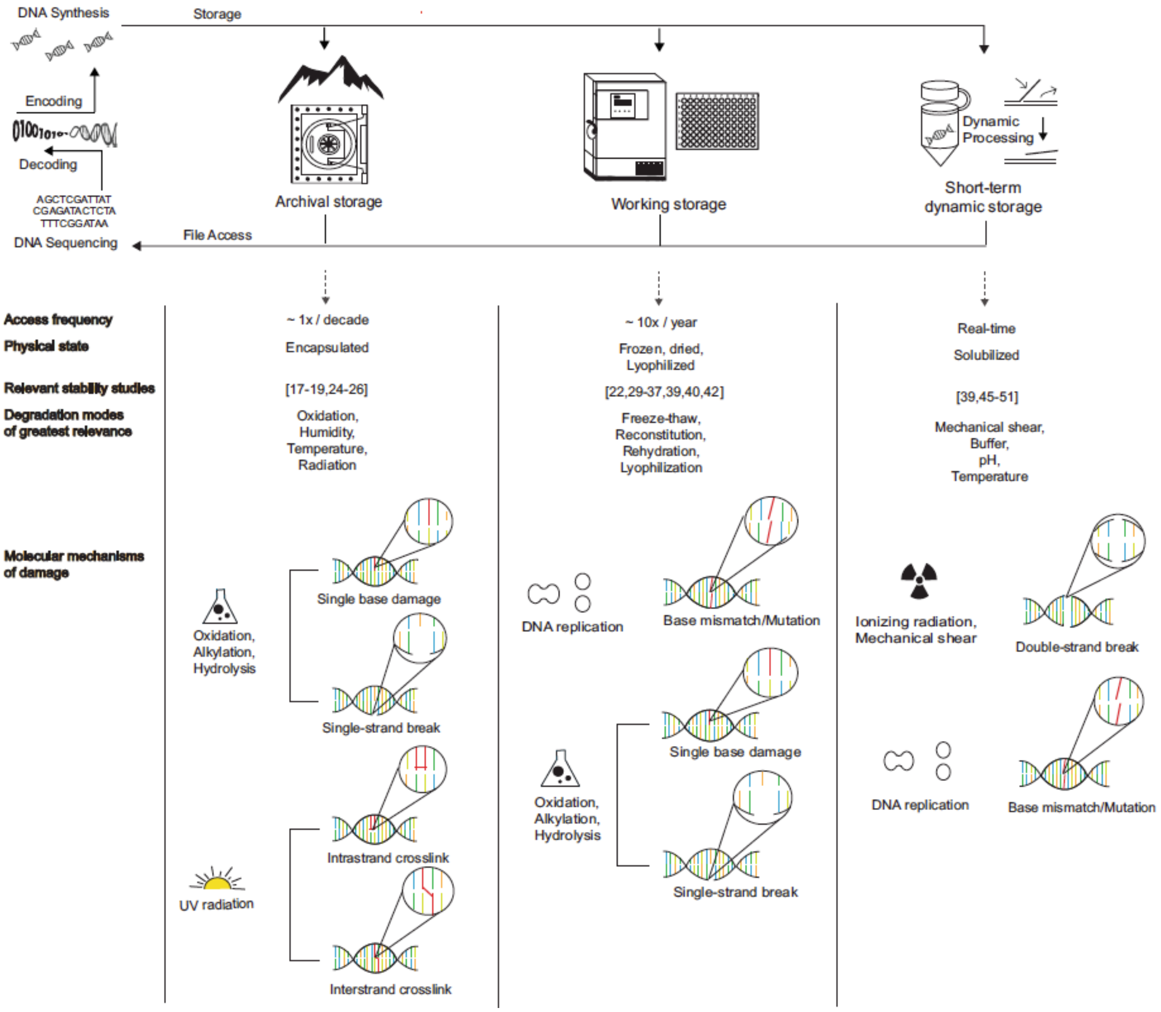
2.1 Introduction:

DNA information storage systems remain in constant flux and development. Indeed, multiple types of systems may arise to address different applications, from long-term ‘write-once-read-never’ archival storage to highly dynamic and frequently accessed data storage, potentially with in-storage computational capabilities. It is important to imagine these possible types of DNA information storage systems and the different unit processes that will comprise these systems, and identify how the chemical, physical, and encoding properties of DNA will influence their design. As DNA or an analog will be the substrate of this class of polymeric systems, its stability under different environmental and process conditions will be a central design consideration, informing the nature of both physical unit processes and encoding algorithms.

An end-to-end DNA storage system is depicted in Fig. **10** with generic unit processes. As applications range from cold archival storage to frequently accessed or even dynamically manipulated data, the DNA is exposed to more manipulation such as phase changes or physical shearing through liquid handling, and to more distinct types of environmental conditions such as buffers with different salt concentrations and pHs. These present more opportunities for degradation as well as specific degradation mechanisms that may influence encoding strategies and sources for data and decoding errors (Fig. **10**). Here we review what is known about the stability of DNA under each of these conditions, organized by their relevance to systems with different operating timescales. We then provide a quantitative analysis of the relative tradeoffs in density, physical redundancy, and encoding strategies that must be sacrificed to achieve increasingly sophisticated system capabilities such as increased access frequency and in-storage computation.

2.2 Write-once-read-never archival storage (access frequency once every ~10+ years):

One of the most likely first applications for DNA-based information storage will be long-term ‘cold’ storage intended for the preservation of historical or other records across decades or centuries. This is because the current high cost of DNA synthesis and sequencing can be justified when they are infrequent and amortized over many years[26]. What is crucial for these applications is a strong understanding of the long-term stability of DNA. The successful recovery of DNA from fossils or microbes preserved in permafrost, some millions of years old, is commonly used as motivating evidence for the long-term stability of DNA[27,28]. However, there are several caveats to consider. First, there are substantial concerns that some of these results were misinterpreted due to potential contamination from contemporary bacteria or human DNA[29]. Therefore, the current best estimate for the recovery of DNA from natural samples based upon a rough agreement between theoretical calculations[29] and physical measurements[30] is roughly 400,000 years, with the major degradation mechanism in these samples thought to be crosslinking between strands that inhibit PCR-based amplification and detection of the DNA.



	Archival storage	Working storage	Short-term dynamic storage
Access frequency	~ 1x / decade	~ 10x / year	Real-time
Physical state	Encapsulated	Frozen, dried, Lyophilized	Solubilized
Relevant stability studies	[17-19,24-26]	[22,29-37,39,40,42]	[39,45-51]
Degradation modes of greatest relevance	Oxidation, Humidity, Temperature, Radiation	Freeze-thaw, Reconstitution, Rehydration, Lyophilization	Mechanical shear, Buffer, pH, Temperature
Molecular mechanisms of damage			

Figure 10: Categories of DNA storage with distinct longevities, functional characteristics, and degradation modes. (Top) A generic DNA-based system showing distinct types of storage modes. (Middle) Functional and physical characteristics of each storage mode. (Bottom) Molecular mechanisms of damage most relevant to each storage mode.

Second, estimations of DNA stability from natural samples may be overly optimistic for DNA storage applications. In the context of recovering DNA from fossils, ribosomal and mitochondrial DNA are often used to identify different species of organisms but can be present at hundreds of copies per cell. Yet, these are the sequences used when claiming successful isolation of DNA when in fact genomic DNA sequences are much more difficult to recover than multi-copy mitochondrial DNA. Thus, while some DNA can be recovered from fossilized or permafrost samples, often only the most abundant sequences are detectable due to substantial degradation. This is further reflected by the estimated half-life of fossilized DNA being only ~500 years, corresponding to a per nucleotide fragmentation rate of 5.50E−6 per year assuming an average 242-nt long mitochondrial DNA sequence in the relatively cold temperatures of permafrost (5.5E−6/nt/yr ∗ 242 nt ∗ 500 yrs ~ 50%)[27]. This is substantially lower than the estimated 400,000 years estimated to recover any DNA signal. Thus, in the context of DNA storage applications where a substantial fraction of the data should be recoverable depending on the encoding strategy and amount of physical redundancy, the ‘useful’ stability of storage systems, based upon data from fossilized DNA, suggests stabilities of a few hundred years or less.

While there are likely substantial stability limitations of natural permafrost or fossilized samples, and frozen aqueous samples in general, fortunately DNA storage systems can be engineered with highly controlled materials and environmental conditions that could substantially augment its stability. Several approaches have been tested, most involving dehydrated forms of DNA to reduce hydrolysis of its phosphate backbone. For example, DNA has been adsorbed onto Flinders Technology Associate (FTA) filter cards, stored in biopolymeric storage matrices such as the commercial product DNA Stable, embedded into silk matrices, or simply stored as lyophilized powder[30,31]. In many cases, the adsorption of the DNA onto a matrix stabilized the DNA. For example, over 40 days at 25, 37, and 45 °C, 80% of the DNA embedded in silk was recoverable compared to 20% when unprotected[32]. Silk also offered protection against UV radiation. Salts have also been shown to offer stabilizing effects for dried DNA particularly against high ambient humidity and can maintain high loading of DNA (>20 wt%) while keeping the DNA relatively accessible[33].

Of the many different methods tested, the current leading approach that consistently exhibits the best stability has been encapsulation of DNA within an inorganic matrix comprised of silica, iron oxide, or a combination of both. Multiple groups have shown that encapsulation can substantially enhance DNA stability[25,34-36]. For example, Puddu and colleagues directly compared the stability of encapsulated DNA with unprotected DNA at 100 °C for 30 min. Eighty percent of the encapsulated DNA was recovered while only 0.05% of the unprotected dried DNA survived. Grass and colleagues estimated that encapsulation in silica particles could maintain DNA for 20–90 years at room temperature[35], 2000 years at 9.4 °C[34], to over 2 million years at −18 °C. These estimates were derived from accelerated aging models applied to data obtained from DNA exposed to elevated temperatures of 60–70 °C. In a few studies, diverse populations of DNA strands encoding actual data were decoded after accelerated aging, providing estimates not only of half-lives, but also evidence that complete files could be retrieved[34,35]. In several cases, data could be retrieved and re-embedded multiple times, although degradation was observed. ‘Break points’ in these systems would depend on the physical redundancy as well as density overhead sacrificed to enable degradation-tolerant encodings.

This work provides substantial evidence for the long-term stability of DNA, especially encapsulated in silica. However, there are several potential limitations to consider. First, the physical processes of encapsulation and retrieval take some time, suggesting this approach is best suited for cold storage applications. Second, the encapsulation of the DNA inherently reduces the information density of the storage system. A layer-by-layer design with alternating DNA and cationic polyethylenimine with a silica final encapsulation has achieved the best storage density to date in such systems, ~3.4 wt% DNA[35]. This is a sacrifice of 1–2 orders of magnitude in information density; yet, while not as dense as pure DNA, given the 5–6 orders of magnitude advantage of DNA storage over conventional storage media, sacrificing 1–2 orders of magnitude of density would still yield a very space-efficient system.

The size of DNA strands is also an important design consideration. One hope in the DNA storage field is to develop synthesis technologies that are able to create longer DNA strands. The benefit of longer strands would be to reduce the percentage of overhead per strand devoted to file addresses and indices, thereby increasing the information density of the system. However, prior work has already suggested diminishing returns with regards to density with increasing strand length[3]. Further supporting the use of shorter DNA strands, empirical evidence suggests that shorter strands are more resilient to environmental insults. For example, exposure to 830 W/m2 of sunlight irradiation led to nearly 2 orders of magnitude greater degradation for a 113 nt compared to a 53 nt DNA strand. While constant exposure to UV is unlikely for a DNA storage system, it may reflect a general principle of length-dependent degradation sensitivities. Indeed, other environmental insults exhibit similar length dependencies, with exposure to 90 °C thermal treatment leading to nearly 3 orders of magnitude greater degradation for longer strands[37], while shorter strands were more resilient to freeze-thaw cycles[38]. As these are accelerated degradation studies, it is likely that longer strands will be sufficiently stable for practical use; however, shorter strand lengths would likely provide improved stability profiles.

Finally, it is important to note that while seemingly simple, accurately determining long-term DNA stability is not trivial nor a solved problem. There are two primary challenges. First, the accelerated aging models used in many studies that can lead to stability predictions of hundreds or even millions of years are inherently extrapolations and could therefore deviate from actual stabilities over long periods of time. This is especially true if there are unknown degradation mechanisms that are important over long timescales but that do not exhibit the exponential dependency on temperature specifically as commonly assumed by many models. Given that the accuracy of data retrieval is paramount for cold storage systems, additional studies assessing accelerated degradation due to other parameters other than temperature would provide increased confidence in extrapolated stabilities. For example, a broader variety of storage conditions should be tested to develop a more comprehensive mechanistic model of DNA stability, including models that not only artificially accelerate aging through elevated temperature but through elevated exposure to electromagnetic radiation and/or subatomic particles, high or low pH, hydrolysis and humidity, and mechanical rearrangements (e.g., freeze thaws, liquid handling) at the molecular level. In addition, DNA concentration and encapsulation conditions could have inherent effects on stability as there could be degradation mechanisms arising from chemical interactions between DNA molecules or between DNA and the encapsulating materials.

A second challenge of accelerated aging studies is that methods to measure degradation often are not sensitive enough, so rely on large degradation effects or require amplification steps (e.g., through PCR) that could skew or bias results. These issues motivate new future gold standards to assess long-term stability. New approaches could include a combination of deep next generation sequencing of short-term samples to detect very rare degradation events. In addition, it may be advisable for the field to collaboratively initiate real time, non-accelerated, long-term studies of DNA stability over the course of a human generation or more, and compare these results to deeply sequenced short-term studies as well as those using different modes of accelerated degradation. These could be performed on actual commercially active storage systems as both a way to monitor stability and to provide insights into long-term stability. This work could be paired with studies of different environmental insults to identify dominant degradation modes and their relative contributions to aging. Overall, this data could be used to update and benchmark the stabilities of any DNA storage systems that were created in the past, and to continually inform and adjust models predicting their stability.

2.3 Working storage (~ accessed multiple times per year):

While there is promise for the ability to store DNA stably over centuries or even millennia, methods capable of this type of stability typically require fully encapsulating and sealing DNA within a matrix, like silica. This is because encapsulation protects the DNA from exposure to humidity, radiation, fluctuations in temperature, and other potential reactants. Furthermore, it may be the most robust and economical storage method, requiring the least amount of specialized storage equipment such as tightly controlled refrigeration and humidity. However, due to substantial DNA loss associated with retrieval from encapsulated media, and the relatively involved encapsulation and retrieval processes, this form of storage is best matched with applications where access is infrequent or never occurs. In a generic DNA storage system, encapsulated DNA could be used to store the 'master' or preservation copy of data, while 'working' copies are maintained using less stable but more accessible methods.

There are three semi-accessible forms of storage that might be compatible with working storage, all of which are similar to how research labs in the biological sciences currently store DNA samples: refrigerated in aqueous solution, frozen in aqueous solution (typically at −20 or −80 °C), or as a dry solid. While DNA recoveries near 100% have been reported from all of these storage forms after ~2 years[39-42], accelerated aging studies at elevated temperatures suggest storing DNA in these forms would not be sufficient for multi-decade stability[31,43] especially when compared to encapsulated storage strategies. For example, at 70 °C, substantial degradation of dried, adsorbed, and aqueous DNA was observed at 15–70% after only 5 days[32,33,44,45], while encapsulated DNA showed no appreciable degradation after 7 days[32]. This may be surprising to many working in the biological sciences where plasmid DNA preparations are stored for entire scientific careers; however, it is important to consider that many such research samples are usually single plasmid constructs stored at very high copy number (1010 copies/µL − 1 µg/µL) so that recovery of the plasmid through bacterial retransformation can occur even with 99% DNA degradation. Regardless, there is strong evidence that storage at 4 °C or below in aqueous or dried form would provide at least a couple of years of stability, appropriate for storing working copies of DNA-based information[40-42], with lyophilized DNA showing better stability than aqueous DNA solutions[46]. One additional important caveat for DNA solutions appears to be the starting concentration of DNA, with dilute solutions of ~0.02 ng/mL exhibiting substantial degradation within weeks when stored at −20 °C[39]. Inclusion of non-specific carrier RNA or DNA was able to abrogate this degradation.

If stored for only a few years, the key concern for working copies of data then becomes the amount of degradation that occurs each time information is accessed, e.g., freeze-thaw of solubilized DNA or rehydration of dried DNA. Likely due to the ability to directly measure the effect of multiple access events within a reasonable experimental time frame rather than needing to simulate or accelerate degradation over time, many more studies have investigated these mechanical effects on DNA stability as opposed to the basal stability of unperturbed samples.

During the freezing and thawing process, ice crystals are formed that generate forces that could lead to breakage of DNA polymers. There is also evidence that freezing makes DNA more susceptible to breakage than non-frozen DNA at similar tensional forces[47]. Several studies have repeatedly frozen and thawed DNA samples, generally observing exponential degradation. An approximately 10% degradation of lambda DNA in Tris-EDTA buffer was observed after 1 freeze-thaw, and 75% degradation was observed after 20 freeze thaws[48]. An exponential fit modeled this process relatively closely (%intact DNA=0.9484*e−0.068*freeze-thaws). A study of DNA stored in water or 50% glycerol found similar degradation rates with more than 75% degradation after 16 freeze thaws, and samples in 50% glycerol performing worse than in water[49].

As with archival storage, one important consideration in assessing degradation from freeze thaws is the size of the DNA strands. Empirical evidence suggests smaller DNA strands would be preferable for improved stability against freeze thaws. This effect was observed comparing genomic DNA above and under 100 kb, but also at smaller length scales of 5 kb[38]. In addition, increased DNA concentrations exhibited a self-protective effect. As freeze thaws cause mechanical stresses on DNA, there is some intuition for why longer DNA strands would be more susceptible to breakage. The effect of concentration of DNA stability may be due to less intuitive mechanisms that should be investigated in more detail, such as altering the phase transition boundaries of the aqueous solution[50].

Dehydration is another means of storing DNA at intermediate (~2–10 years) timescales. Vacuum dried DNA, for example, was shown to be stable at 5 °C for at least 22 months[45]. While detailed studies of repeated rehydrations are lacking and need to be performed, a few studies have assessed the recovery efficiency from dried DNA and the impact of prior exposure to elevated temperatures on this recovery. For example, a library of 2042 unique sequences comprising ~20 kb of data was dehydrated on a microfluidic "PurpleDrop" device[51]. Rehydration with water was performed at various dwell times ranging from 1 to 120 s. While increasing the dwell time generally increased the amount of DNA recovered, just 1 s was able to recover ~77% of the DNA by mass. The degradation rate was not directly reported but with sufficient physical redundancy (copy number) and sequencing depth, the entire file was recovered and successfully decoded. Of note, the consistency of the retrieval process exhibited substantial variability and suggests further work investigating mechanisms of the dehydration process, potential irreversible denaturation of the DNA into different helical forms (i.e., forms A or B), and molecular degradation chemistries will be important. For example, exposure of dried DNA to elevated temperatures could irreversibly denature the secondary helical structure of DNA[44], resulting in samples that are difficult to process or sequence.

Other forms of DNA instability should also be considered for both solution and dried sample storage methods. These include depurination (removal of adenine or guanine bases from the DNA backbone) and oxidation (8-oxo-dG). Both in solution and in dried form, the rate of depurination and oxidation are both higher by 1–2 orders of magnitude compared to strand breakage. Roughly 6% of strands that are 200 nt would develop a depurinated or 8-oxo-dG nucleotide per year at room temperature[44], with rapid increases in these rates with temperature, following an Arrehenius dependence. These degradation mechanisms have functional impacts. Depurination would result in missing or incorrect nucleotides during sequencing, while 8-oxo-dG can lead to cross-linking between DNA strands which inhibits both rehydration of dried DNA as well as amplification and sequencing of the DNA. Depending on the encoding method and physical redundancy of the storage system, these could further reduce the estimated longevity of solution, frozen, and dried storage methods.

2.4 Short-term storage for dynamic handling of data:

The active development of DNA-based computation preceded the development of DNA storage systems by over a decade[5,6,52,53]. The idea of now merging these two fields to perform in-storage computation has intriguing implications including potentially providing the ability to directly search[54] and edit[55] DNA databases. In addition to computation, there is also the hope that DNA storage systems may be capable of dramatically lower latencies of operation than the current multi hour to multi-day process of reading and writing information. For both of these applications, physical manipulation of DNA will be necessary. An important factor to consider for dynamically accessible storage is the complexity of storage methods. Most likely, DNA for these applications will need to be stored in a soluble aqueous form with buffers compatible with molecular processes including dynamic DNA–DNA hybridizations, transcription or polymerization, and other enzyme-driven reactions. Alternatives may include the use of magnetic particles as long as the adsorption/desorption of DNA is relatively rapid. Understanding and potentially improving the stability of DNA in these contexts will inform the operating lifetime of systems as well as dictate the types of error tolerant encodings that will be necessary.

Almost all manipulations performed on DNA will involve liquid handling, often through small apertures such as pipette tips, microfluidic channels, or tubing. These manipulations impart shear forces on DNA that can lead to strand breakage. Several studies have investigated the effect of vortexing or the use of other shearing devices on DNA stability and showed substantial degradation, for example nearly 70% fragmentation after 3 min[48,56-58] on a standard table-top vortexer at maximum speed. While these types of shearing devices are unlikely to be used in DNA storage systems, they may be useful in providing relationships between degradation rates and shear forces or energy inputs if used in controlled settings, like rheometers. More directly relevant to storage systems is the measurement of DNA fragmentation due to actual pipetting. With a relatively rapid single pipetting action through a 1 mL tip, 70% of long lambda phage DNA was fragmented. Slower and gentler pipetting still led to more than 50% of lambda DNA being fragmented[48]. This is somewhat concerning given liquid handling often uses even smaller 200, 20, and 2 μL tips that would result in higher shear forces. With automation, the pipetting actions could be slowed significantly to be less turbulent; however, this might present a significant trade off in increasing the time required to perform each physical operation or manipulation. What is clear is whatever form of liquid handling is used, careful assessment of DNA stability will be important and processing parameters tuned to ensure DNA stability. New forms of liquid handling should also be explored but also quantitatively assessed for their impacts on DNA stability. For example, acoustic liquid handling of nanoliter to microliter droplets, may exhibit distinct impacts on DNA stability through additional types of forces such as exposure to surface tensions through high surface area to volume ratios. In addition, as in the case of freeze-thawing where smaller DNA strands were qualitatively shown to be more resistant to fragmentation, comprehensive determinations of thresholds and quantitative relationships between DNA length and resiliency to shear forces are currently lacking but would be useful for designing storage systems with low latency or in-storage computation.

While physical manipulations will likely impart some level of unavoidable damage to DNA, the biochemical environment of the DNA and the properties of the DNA molecules themselves could improve stability and resilience to these processes as well as enhance stability in aqueous and unfrozen states. For example, as was already discussed, DNA stored at 4 °C can remain relatively stable over a few years and maintaining a high concentration of DNA or including carrier DNA or RNA helps slow down degradation rates. DNA degradation can occur through several mechanisms including oxidation and hydrolysis of the phosphate backbone or the base from the sugar (depurination). In addition to temperature which generally accelerates rates of most degradation reactions, controlling pH is perhaps the next most obvious candidate for improving DNA stability against these mechanisms. Both acidic and basic conditions can enhance the hydrolysis rate of DNA by either increasing the electrophilicity of the DNA or the nucleophilicity of water. For example, it was estimated that a change from pH 6 to 5 could increase the degradation rate of DNA by an order of magnitude[59], and even just a 90-min exposure to pH 4.0 at 65 °C led to almost complete degradation of DNA samples[60]. While these experiments were performed at elevated temperatures that might be argued could be avoided, there are many molecular biology manipulations that require transient exposures to elevated temperatures above at least 37 °C and often above 70 °C where many polymerases function optimally. These include DNA in-storage computation, in-storage editing, or even simply copying information by PCR. Furthermore, for search-type functions that rely on DNA–DNA hybridizations, stable hybridizations for ~20 nt sequences occur at between ~50 and 60 °C. Room temperature operation would not provide adequate energy to first melt any secondary structures that might have formed that would block hybridization and would also result in non-specific sequences hybridizing with each other. Shorter DNA sequence lengths that hybridize near room temperature would be too short to confer adequate sequence specificity amongst large, highly diverse libraries of DNA strands that will comprise most DNA storage systems. Therefore, tightly controlling pH as well as the time exposure to elevated temperatures will be important design considerations for storage systems. The use of many standard buffers such as PBS and TE could help maintain pH levels at appropriate targets. Other additives may also have protective effects including DNAstable™ [61], salts[33], and trehalose[31,39].

2.5 Tradeoffs between DNA stability and information density

This work has focused on empirical measurements of DNA stability under a range of different conditions. Together with theoretical analyses[4,62,63], there is strong evidence for the utility of DNA as a storage medium. Nevertheless, DNA has finite stabilities, and is especially labile in conditions relevant for frequent access and dynamic processing. This does not preclude the use of DNA for storage applications but does affect the information density of systems by requiring higher redundancy in the number of copies of each distinct strand as well as in the encoding methods used to achieve certain reliability. Understanding the effect of DNA stability on these tradeoffs will be important in designing unit processes and systems for specific types of storage applications. Here we discuss these tradeoffs through a series of analyses shown in Fig. **11**, and in particular demonstrate how relationships between copy number, strand loss, strand length, information density can be modeled and used to inform system design. The goal here is to show how models can reveal both intuitive and unintuitive relationships between parameters; it is important to note that ranges of parameter values and observed trends shown here may change depending on the specific details of each system.

Most DNA data storage systems use short sequences or ‘addresses’ written into strands of each file for retrieval. These addresses typically are complementary to short DNA oligomers used to amplify the files through PCR[3] or to extract them through affinity-based methods[64]. To function at reasonable temperature ranges (<100 °C) the addresses are generally limited to <25 nt as longer sequences would require higher temperatures to ‘melt’ during PCR cycling. However, addresses cannot be much shorter without sacrificing diversity in addresses[55]. The addresses take up space on each strand and do not encode data (e.g., overhead). Index sequences are also overhead and are necessary to include to know how the different strands comprising a file should be ordered.

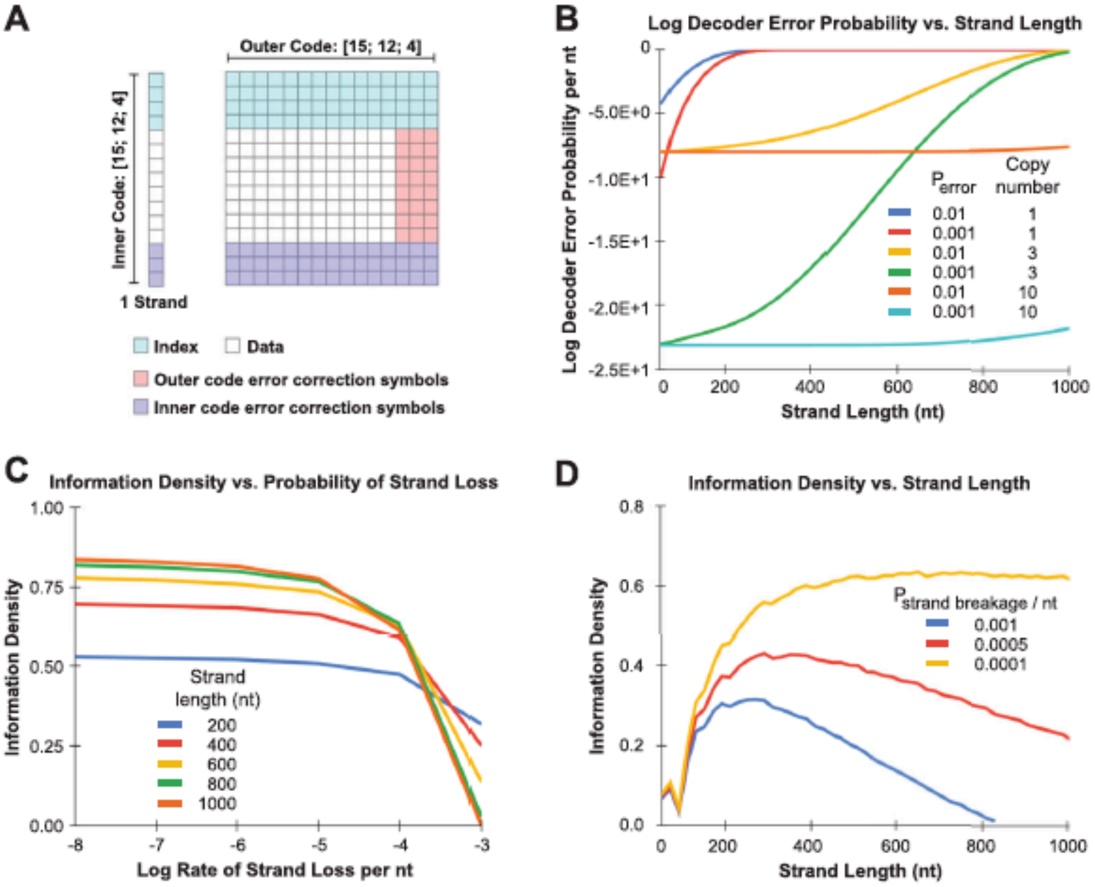


Figure 11: Analysis of tradeoffs between system reliability, information density, strand length, and error rates. A Reed-Solomon inner-outer encoding scheme. **B** Relationship between log decoder error probability during RS decoding and DNA strand length, including the effects of symbol error rate (mutations, insertions, and deletion, Perror) and copy number. **C** Relationship between information density of a DNA storage system and the probability of symbol erasure (strand loss due to breakage) as a function of strand length. D Relationship between information density and strand length as a function of the probability of strand breakage. C and D assume a copy number of 1.

In addition to the address and index, error correction codes may use additional overhead. Electronic storage systems adopt and employ many error correction mechanisms to ensure the reliability of stored data. Similarly, to cope with frequent errors, DNA storage systems can leverage error correction codes that are capable of detecting and correcting errors that occur as a result of strand breakage or loss and due to substitutions, insertions, and deletions within strands. Error correction codes work by adding enough redundancy to recompute the original data even in the presence of errors or missing strands. Hence, to maintain the same reliability of information transfer or recovery, error correction forces a trade-off between density of information and tolerance to errors. The higher the likelihood of strand error or loss due to reduced DNA stability, the more overhead must be spent on error correction. While a variety of codes have been proposed for DNA storage, Reed-Solomon (RS) codes are particularly popular given their configurability and tolerance to errors[34,65-67]. The error correction properties of RS codes are well known, and we use them here to explore the combined effects of strand errors, breakage, and length on the information density and reliability of DNA storage systems.

We will not focus on the details of different codes but rather use RS codes to illustrate some general trends to consider when designing DNA storage systems. Additional details of popular error correction approaches can be found in a few recent references[14,15], while details about our implementation of RS codes and parameters used (shown in brackets) can be found in the Supplemental Information section. In brief, the key features of the implementation used here are (1) that the RS code is capable of detecting and differentiating two kinds of errors, symbol errors such as insertions, deletions, and mutations, collectively having a probability of perror, and symbol erasures such as the breakage or loss of a DNA strand with probability pstrand erasure; (2) inner and outer RS codes are interleaved where the inner code protects against errors within a strand and the outer code corrects for missing or erroneous strands (Fig. **11A**); (3) an error probability can be calculated to estimate the probability the code will not be able to account for and correct errors given certain error and strand loss rates; (4) the amount of redundancy in the code is tunable in order to achieve a certain error probability or system reliability; (5) the code accounts for multiple copies of each distinct DNA strand, with strand loss or ‘erasure’ only occurring when all copies of a distinct DNA strand are lost; and (6) the length of the DNA strands can be tuned with effects on density due to the ‘overhead’ of addresses and indices.

2.5.1 Decoding error analysis. To provide a non-exhaustive example of the potential importance of trade-off analyses, here we focus on a major mode of DNA degradation, strand loss through hydrolysis or mechanical breakage. First, to better understand the impact of strand loss on the probability of a decoding error, we analytically model its effect on the decoding error probability of an outer RS[255,223,33] code. Based upon experimental observations, we conservatively model the probability of strand breakage as linearly dependent on strand length, and we assume that strand loss due to breakage is equivalent to an erasure in the outer code. We further assume that multiple copies of a strand exponentially reduce the likelihood of loss because all copies of the strand must be lost to cause an erasure. An exact analytical formula for this probability is unknown so we estimate the relationship as pstrand erasure(length L, copies c) = (L ∗ 5E−3)c. This equation is chosen so that strands of length 1000 nt have a probability of 50% or less of erasure, a value that can be tuned depending on the stability measured for any particular DNA storage system. We fix the perror to 1e−2 and 1e−3 per nt to cover the range of typical synthesis and sequencing errors[68,69].

Figure **11B** shows the impact of longer strands on the decoder error probability for several system configurations with different copy numbers of each strand and different error rates. The y-axis shows the log decoding error probability and the x-axis is the strand length. We have chosen for comparison a system with similar read error rates to hard drives (on the order of 10−14 per bit, log error probability of −1.5E + 01)[70]. For this system that has a linearly dependent probability of strand breakage on strand length, the trend is that longer strands increase the likelihood of strand loss, and substantially increase the decoder's probability of error. Other parameters matter, too. Lower perror significantly reduces the residual likelihood of error. Also, higher numbers of copies per strand also have a large effect since it becomes exponentially less likely that all copies of a strand are lost. Current experimental systems tend to operate in a regime with large numbers of copies; however, even if future systems may wish to be more lean, this analysis suggests copy number requirements may not need to be as high as intuition may suggest particularly if strand length designs remain in the 200–500 nt range.

2.5.2 Information density. To demonstrate how the relationship between strand loss and information density can be studied, we vary the probability of strand breakage per nt from 10−3 to 10−8 and analyze the information density for many different strand lengths. For a given strand length, there are many different possible designs for a RS inner-outer code. We use an algorithm to find a design with a residual error probability <10−14 and maximum information density while keeping the outer block size at 255, and the index to 4. In Fig. **11C**, we report the density on the y-axis vs. the probability of strand loss on the x-axis for each strand length (L) considered. We assume a single copy of each strand to maximize information density and fully exploit the error-correcting capability of RS codes.

In this system, shorter strands achieve superior density at high strand breakage rates. This is a result of shorter strands being overall less likely to break, and therefore the outer code needs fewer error correction symbols. Longer strands have a higher overall likelihood to break and need more error correction in the outer code to compensate. However, as the overall probability of breakage per nt decreases (to the left), all strands are less likely to break, and this gives longer strands an advantage since they can hold a larger fraction of information per strand and the outer code can work effectively even with a relatively small number of error correction symbols. A deeper analysis further shows that the magnitude of the strand breakage rate can fundamentally alter the relationship between information density and optimal strand length (Fig. **11D**). In fact, it is not simply that shorter strands are better for high strand loss rates but that there can be optimal lengths that balance the overhead needed for file addresses and indices with the encoding overhead required to account for strand loss.

This analysis is just one example of many that can be performed interrogating the effects of diverse parameters of DNA storage systems. It underscores the need to tailor error correction and system parameters like strand length and copy number to different settings including error rates that vary according to environmental conditions and data storage applications. For example, long-term archival storage will likely encapsulate the data in silica helping keep the probability of strand loss or breakage low, allowing longer DNA strands to be used and achieving higher information densities. In contrast, working storage or short-term dynamic storage would benefit from shorter strand lengths to compensate for higher strand loss rates.

2.6 Future prospects

The robustness and failure rates of information storage systems are of utmost importance as the reliability of data retrieval must be concretely reported, verifiable, and trustworthy[71-74]. While we have some rough estimates and measurements of DNA stability in a variety of conditions, often measurements exhibit considerable noise and variability between experimentalists and research groups in addition to substantial noise between samples in an individual experiment. There are likely experimental details affecting the accurate interpretation of measurements including the manipulation of DNA itself in setting up experiments, or confounding parameters like DNA solubility. Experiments exploring a more comprehensive set of parameters and that assess sources of variability in results could provide more confidence in the design and utility of DNA storage technologies. Fine-tuning exact buffer conditions, assessing changes in its composition, and maintaining strict control and provenance records over the environmental exposures of the DNA throughout its complete lifetime starting with DNA synthesis will be important for commercial DNA storage products.

In addition, while many studies have quantified DNA degradation through some form of quantitative PCR, mass measurements, or even next-generation sequencing, the definition of DNA degradation remains unprecise and too limited despite its considerable impact on the physical design and encoding of reliable storage systems. For example, there are many ways a DNA storage system can be degraded. (1) The loss of DNA strands could be biased toward strands with certain properties such as length, base content, or presence of specific sequences. (2) DNA strands may be fragmented in different patterns and frequencies depending on base content, length, and physical processing conditions. (3) Depurination or chemical alterations of bases may occur and not be directly assessed by sequencing or QPCR based approaches. How each of these types of degradation mechanisms are affected by environmental conditions is important for system design and should be carefully assessed.

There is also the intriguing possibility that the physical stability of storage systems could be enhanced by the use of another polymer that is chemically more stable than DNA, although substantial work would likely be needed to replicate the synthesis, processing, and sequencing technologies and infrastructure available for DNA[75]. For example, 'locked' nucleic acid monomers possess a methylene bridge between a 2′ oxygen and the 4′ carbon of the pentose ring and offer substantial resistance to nuclease enzymes present abundantly in the ambient environment. There are many other potential chemistries of nucleic acid polymer backbones that may offer differing stabilities tuned for specific environmental conditions or applications, including bicyclo-DNA or glycerol-DNA that have altered sugar backbone chemistries[76] or nuclease resistant nucleic acids[75]. In addition to altering the biopolymer substrate itself, protective additives or even active repair systems similar to those in natural biological systems may improve storage system reliability.

There are clearly many opportunities and an important need to better understand, characterize, and improve DNA stability. While seemingly a straightforward concept, assessing the stability of DNA and making appropriate choices of storage methods is not trivial. DNA stability has been experimentally measured and reported in many diverse ways, including mutational rate, breakage rate per base, and loss of intact DNA strands. Degradation rates have also been reported in a mix of many different environmental, temperature, buffer, and temporal conditions. Furthermore, the functional impact of different types of degradation will depend on the nature of the storage system.

For example, mechanical degradation may affect systems that use longer DNA strands compared to shorter strands, degradation rate may be complicated to predict depending on the density of storage systems due to its potential nonlinear dependency on DNA concentration, and some encoding algorithms may sacrifice information density but be more resistant to the loss of strands. However, it is clear that even with our current nascent knowledge of DNA stability, reliable DNA storage systems can already be created with existing technologies. In addition to developing a better understanding of DNA stability, what will be important is the recognition that the appropriate tradeoffs and limitations in system properties and capabilities should be made and can be supported through models. With improving measurements, a better understanding of degradation mechanisms, new technologies, models, and enhanced encoding algorithms, the efficiency of these systems will continue to improve the commercial viability of DNA-based information storage.

REFERENCES

(1) Watson, J., Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171, 737–738 (1953). https://doi.org/10.1038/171737a0

(2) Wiener, N. Interview: Machines Smarter Than Men? US. News World Rep. 1964, 84-86 pp.

(3) Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G., Strauss, K., 2016. A DNA-Based Archival Storage System, ASPLOS '16, https://doi.org/10.1145/2872362.2872397

(4) Ceze, L.; Nivala, J.; Strauss, K. Molecular Digital Data Storage Using DNA. Nat. Rev. Genet. 2019, 20 (8), 456–466. https://doi.org/10.1038/s41576-019-0125-3

(5) Adleman LM: Molecular computation of solutions to combinatorial problems. Science 1994, 266:1021-1024.

(6) Cox JC, Cohen DS, Ellington AD: The complexities of DNA computation. Trends Biotechnol 1999, 171:151-154.

(7) Shehabi, A., Smith, S.J., Horner, N., Azevedo, I., Brown, R., Koomey, J., Masanet, E., Sartor, D., Herrlin, M., Lintner, W. 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775

(8) Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith and Jonathan Koomey, Recalibrating global data center energy-use estimates, Science 367 (6481), 984-986 DOI: 10.1126/science.aba3758

(9) Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. Nature, 561(7722), 163–166. https://doi.org/10.1038/d41586-018-06610-y

(10) Ann Bednarz, Amazon, Google, IBM, Microsoft: how their data-migration appliances stack up, https://www.networkworld.com/article/3295937/migrate-data-to-the-cloud-how-appliances-from-amazon-google-microsoft-and-ibm-stack-up.html

(11) Mullis, Kary B. “The Unusual Origin of the Polymerase Chain Reaction.” Scientific American, vol. 262, no. 4, 1990, pp. 56–65., www.jstor.org/stable/24996713. Accessed 31 Dec. 2020.

(12) Hughes, R.A., Ellington, A.D., 2017. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harbor Perspectives in Biology 9, a023812.. doi:10.1101/cshperspect.a023812

(13) Burel, A., Carapito, C., Lutz, J-F. and Charles, L. Macromolecules 2017 50 (20), 8290-8296 DOI: 10.1021/acs.macromol.7b01737

(14) De Silva, P.Y., Ganegoda, G.U. "New Trends of Digital Data Storage in DNA", BioMed Research International, vol. 2016, Article ID 8072463, https://doi.org/10.1155/2016/8072463

(15) Erlich, Y., Zielinski, D., 2017. DNA Fountain enables a robust and efficient storage architecture. Science 355, 950–954.. doi:10.1126/science.aaj2038

(16) Lim, C.K., Nirantar, S., Yew, W.S., Poh, C.L., 2021. Novel Modalities in DNA Data Storage. Trends in Biotechnology.. doi:10.1016/j.tibtech.2020.12.008

(17) Hughes, R.A., Ellington, A.D., 2017. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harbor Perspectives in Biology 9, a023812.. doi:10.1101/cshperspect.a023812

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

(19) Lee, H.H., Kalhor, R., Goela, N., Bolot, J., Church, G.M., 2019. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nature Communications 10.. doi:10.1038/s41467-019-10258-1

(20) Kok, S.D., Stanton, L.H., Slaby, T., Durot, M., Holmes, V.F., Patel, K.G., Platt, D., Shapland, E.B., Serber, Z., Dean, J., Newman, J.D., Chandran, S.S., 2014. Rapid and Reliable DNA Assembly via Ligase Cycling Reaction. ACS Synthetic Biology 3, 97–106.. doi:10.1021/sb4001992

(21) Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods 6, 343–345.. doi:10.1038/nmeth.1318

(22) Quan, J., Tian, J., 2009. Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. PLOS ONE 4, e6441.. doi:10.1371/journal.pone.0006441

(23) Kadkhodaei, S., Memari, H.R., Abbasiliasi, S., Rezaei, M.A., Movahedi, A., Shun, T.J., Ariff, A.B., 2016. Multiple overlap extension PCR (MOE-PCR): an effective technical shortcut to high throughput synthetic biology. RSC Advances 6, 66682–66694.. doi:10.1039/c6ra13172g

(24) Roquet, Nathaniel (Cambridge, MA, US), Park, Hyunjun (Cambridge, MA, US), Bhatia, Swapnil P. (Boston, MA, US), Hazel, Mike (Boston, MA, US), Day, Richard (Boston, MA, US), Hammond, Richard (Boston, MA, US), Brown, James (Boston, MA, US), Richardson, Rodney (Boston, MA, US), Redman, Thomas (Boston, MA, US), Leake, Devin (Boston, MA, US) 2019, PRINTER-FINISHER SYSTEM FOR DATA STORAGE IN DNA, United States Catalog Technologies, Inc. (Boston, MA, US) 20190351673 https://www.freepatentsonline.com/y2019/0351673.html

(25) Koch, J.; Gantenbein, S.; Masania, K.; Stark, W. J.; Erlich, Y.; Grass, R. N. A DNA-of-Things Storage Architecture to Create Materials with Embedded Memory. Nat. Biotechnol. 2020, 38 (1), 39–43. https://doi.org/10.1038/s41587-019-0356-z.

(26) Byron, J., Long, D. D. E., & Miller, E. L. (n.d.). Measuring the Cost of Reliability in Archival Systems.

(27) Allentoft, M. E.; Collins, M.; Harker, D.; Haile, J.; Oskam, C. L.; Hale, M. L.; Campos, P. F.; Samaniego, J. A.; Gilbert, T. P. M.; Willerslev, E.; et al. The Half-Life of DNA in Bone: Measuring Decay Kinetics in 158 Dated Fossils. Proc. R. Soc. B Biol. Sci. 2012, 279 (1748), 4724–4733. https://doi.org/10.1098/rspb.2012.1745.

(28) Bada, J. L.; Wang, X. S.; Poinar, H. N.; Pääbo, S.; Poinar, G. O. Amino Acid Racemization in Amber-Entombed Insects: Implications for DNA Preservation. Geochim. Cosmochim. Acta 1994, 58 (14), 3131–3135. https://doi.org/10.1016/0016-7037(94)90185-6.

(29) Hofreiter, M.; Serre, D.; Poinar, H. N.; Kuch, M.; Pääbo, S. Hofreiter_Ancient DNA_NatRevGen_2001_1556040. Nature.Com 2001, 2 (May), 3–9.

(30) Willerslev, E.; Hansen, A. J.; Rønn, R.; Brand, T. B.; Barnes, I.; Wiuf, C.; Gilichinsky, D.; Mitchell, D.; Cooper, A. Long-Term Persistence of Bacterial DNA. Curr. Biol. 2004, 14 (1), 13–14. https://doi.org/10.1016/j.cub.2003.12.012.

(31) Organick, L. W.; Nguyen, B. H.; McAmis, R.; Chen, W. D.; Kohll, A. X.; Ang, S. D.; Grass, R. N.; Ceze, L. H.; Strauss, K. An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage. bioRxiv 2020, 2020.09.19.304014.

(32) Liu, Y., Zheng, Z., Gong, H., Liu, M., Guo, S., Li, G., Wang, X., & Kaplan, D. L. (2017). DNA preservation in silk. Biomaterials Science. 1279–1292. https://doi.org/10.1039/c6bm00741d

(33) Kohll, A. X.; Antkowiak, P. L.; Chen, W. D.; Nguyen, B. H.; Stark, W. J.; Ceze, L.; Strauss, K.; Grass, R. N. ChemComm Storage with Earth Alkaline Salts. 2020, 3613–3616. https://doi.org/10.1039/

(34) Grass, R. N.; Heckel, R.; Puddu, M.; Paunescu, D.; Stark, W. J. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angew. Chemie - Int. Ed. 2015, 54 (8), 2552–2555. https://doi.org/10.1002/anie.201411378.

(35) Chen, W. D.; Kohll, A. X.; Nguyen, B. H.; Koch, J.; Heckel, R.; Stark, W. J.; Ceze, L.; Strauss, K.; Grass, R. N. Combining Data Longevity with High Storage Capacity—Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles. Adv. Funct. Mater. 2019, 29 (28), 1–8. https://doi.org/10.1002/adfm.201901672.

(36) Clermont, D.; Santoni, S.; Saker, S.; Gomard, M.; Gardais, E.; Bizet, C. Assessment of DNA Encapsulation, a New Room-Temperature DNA Storage Method. Biopreserv. Biobank. 2014, 12 (3), 176–183. https://doi.org/10.1089/bio.2013.0082.

(37) Mikutis, G.; Schmid, L.; Stark, W. J.; Grass, R. N. Length-Dependent DNA Degradation Kinetic Model : Decay Compensation in DNA Tracer Concentration Measurements. 2019, 65 (1). https://doi.org/10.1002/aic.16433.

(38) Shao, W.; Khin, S.; Kopp, W. C. Characterization of Effect of Repeated Freeze and Thaw Cycles on Stability of Genomic DNA Using Pulsed Field Gel Electrophoresis. Biopreserv. Biobank. 2012, 10 (1), 4–11.

(39) Baoutina, A.; Bhat, S.; Partis, L.; Emslie, K. R. Storage Stability of Solutions of DNA Standards. Anal. Chem. 2019, 91 (19), 12268–12274. https://doi.org/10.1021/acs.analchem.9b02334.

(40) Ivanova, N. V. & Kuzmina, M. L. Protocols for dry DNA storage and shipment at room temperature. Molecular Ecology Resources 13, 890–898 (2013).

(41) Madisen, L. The effects of storage of blood and isolated DNA on the integrity of DNA. American Journal of Medical Genetics 27, 379–390 (1987).

(42) Smith, S. & Morin, P. A. Optimal Storage Conditions for Highly Dilute DNA Samples: A Role for Trehalose as a Preserving Agent. Journal of Forensic Sciences 50, 1–8 (2005).

(43) Dhanasekaran, S.; Doherty, T. M.; Kenneth, J. Comparison of Different Standards for Real-Time PCR-Based Absolute Quantification. J. Immunol. Methods 2010, 354 (1–2), 34–39.

(44) Bonnet, J.; Colotte, M.; Coudy, D.; Couallier, V.; Portier, J.; Morin, B.; Tuffet, S. Chain and Conformation Stability of Solid-State DNA: Implications for Room Temperature Storage. Nucleic Acids Res. 2009, 38 (5), 1531–1546. https://doi.org/10.1093/nar/gkp1060.

(45) Trapmann, S.; Catalani, P.; Hoorfar, J.; Prokisch, J.; Van Iwaarden, P.; Schimmel, H. Development of a Novel Approach for the Production of Dried Genomic DNA for Use as Standards for Qualitative PCR Testing of Food-Borne Pathogens. Accredit. Qual. Assur. 2004, 9 (11–12), 695–699. https://doi.org/10.1007/s00769-004-0872-4.

(46) Podivinsky, E.; Love, J. L.; Colff, L. van der; Samuel, L. Effect of Storage Regime on the Stability of DNA Used as a Calibration Standard for Real-Time Polymerase Chain Reaction. Anal. Biochem. 2009, 394 (1), 132–134.

(47) Chung, W. J., Cui, Y., Chen, C. S., Wei, W. H., Chang, R. S., Shu, W. Y., & Hsu, I. C. (2017). Freezing shortens the lifetime of DNA molecules under tension. Journal of Biological Physics, 43(4), 511–524. https://doi.org/10.1007/s10867-017-9466-3

(48) Yoo, H.-B. Flow Cytometric Investigation on Degradation of Macro-DNA by Common Laboratory Manipulations. J. Biophys. Chem. 2011, 02 (02), 102–111.

(49) Schaudien, D.; Baumgärtner, W.; Herden, C. High Preservation of DNA Standards Diluted in 50% Glycerol. Diagnostic Mol. Pathol. 2007, 16

(50) Woutersen, S.; Ensing, B.; Hilbers, M.; Zhao, Z.; Austen Angell, C. A Liquid-Liquid Transition in Supercooled Aqueous Solution Related to the HDA-LDA Transition. Science (80-.). 2018, 359 (6380), 1127–1131. https://doi.org/10.1126/science.aao7049.

(51) Newman, S.; Stephenson, A. P.; Willsey, M.; Nguyen, B. H.; Takahashi, C. N.; Strauss, K.; Ceze, L. Dehydration with Digital Micro Fluidic Retrieval. Nat. Commun. No. 2019, 1–6. https://doi.org/10.1038/s41467-019-09517-y.

(52) Boneh D, Dunworth C, Lipton RJ. Breaking DES using a molecular computer. In DNA Based Computers. American Mathematical Society; 1996.

(53) Winfree, E. Simulations of Computing by Self-Assembly. Technical Report CaltechCSTR:19998.22.

(54) Bee, C.; Chen, Y.-J.; Ward, D.; Liu, X.; Seelig, G.; Strauss, K.; Ceze, L. H. Content-Based Similarity Search in Large-Scale DNA Data Storage Systems. bioRxiv 2020, 2020.05.25.115477. https://doi.org/10.1101/2020.05.25.115477.

(55) Lin, K. N., Volkel, K., Tuck, J. M., & Keung, A. J. (2020). Dynamic and scalable DNA-based information storage. Nature Communications, 11(1), 1–12. https://doi.org/10.1038/s41467-020-16797-2

(56) Lengsfeld, C. S.; Anchordoquy, T. J. Shear-Induced Degradation of Plasmid DNA. J. Pharm. Sci. 2002, 91 (7), 1581–1589. https://doi.org/10.1002/jps.10140.

(57) Freitas, S.; Monteiro, G. A.; Prazeres, D. M. F.; Wu, M. L.; Santos, A. L. Stabilization of Naked and Condensed Plasmid DNA against Degradation Induced by Ultrasounds and High-Shear Vortices. 2009, 246, 237–246. https://doi.org/10.1042/BA20080215.

(58) Levy, M. S.; Collins, I. J.; Yim, S. S.; Ward, J. M.; Titchener-Hooker, N.; Ayazi Shamlou, P.; Dunnill, P. Effect of Shear on Plasmid DNA in Solution. Bioprocess Eng. 1999, 20 (1), 7–13. https://doi.org/10.1007/s004490050552.

(59) Lindahl, T.; Nyberg, B. Rate of Depurination of Native Deoxyribonucleic Acid. 1972, 11 (362), 3610–3618. https://doi.org/10.1021/bi00769a018.

(60) Bauer, T.; Weller, P.; Hammes, W. P.; Hertel, C. The Effect of Processing Parameters on DNA Degradation in Food. Eur. Food Res. Technol. 2003, 217 (4), 338–343. https://doi.org/10.1007/s00217-003-0743-y.

(61) Howlett, S. E.; Castillo, H. S.; Gioeni, L. J.; Robertson, J. M.; Donfack, J. Evaluation of DNAstable^{TM} for DNA Storage at Ambient Temperature. Forensic Sci. Int. Genet. 2014, 8 (1), 170–178. https://doi.org/10.1016/j.fsigen.2013.09.003.

(62) Zhirnov, V.; Zadegan, R. M.; Sandhu, G. S.; Church, G. M.; Hughes, W. L. Nucleic Acid Memory. Nat. Mater. 2016, 15 (4), 366–370. https://doi.org/10.1038/nmat4594.

(63) Thomas J. Anchordoquy and Marion C. Molina. Preservation of DNA. Cell Preservation Technology, Volume 5, Number 4, 2007, DOI: 10.1089/cpt.2007.0511

(64) Tomek, K. J.; Volkel, K.; Simpson, A.; Hass, A. G.; Indermaur, E. W.; Tuck, J.; Keung, A. J. Driving the Scalability of DNA-Based Information Storage Systems. ACS Synthetic Biology, 2019, 8 (6), 1241-1248. https://doi.org/10.1021/acssynbio.9b00100

(65) Shu Lin and Daniel Costello, Error Control Coding, Prentice Hall, 2004.

(66) Meinolf Blawat, Klaus Gaedke, Ingo Hütter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, George M. Church, Forward Error Correction for DNA Data Storage, Procedia Computer Science, Volume 80,2016, Pages 1011-1022, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.05.398

(67) Organick, L.; Ang, S. D.; Chen, Y. J.; Lopez, R.; Yekhanin, S.; Makarychev, K.; Racz, M. Z.; Kamath, G.; Gopalan, P.; Nguyen, B.; et al. Random Access in Large-Scale DNA Data Storage. Nat. Biotechnol. 2018, 36 (3), 242–248. https://doi.org/10.1038/nbt.4079.

(68) Ma, S., Saaem, I., & Tian, J. (2012). Error correction in gene synthesis technology. Trends in Biotechnology, 30(3), 147–154.https://doi.org/10.1016/j.tibtech.2011.10.002

(69) Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports, 8(1), 1–14. https://doi.org/10.1038/s41598-018-29325-6

(70) Jim Gray, Catharine van Ingen. Empirical Measurements of Disk Failure Rates and Error Rates Microsoft Research Technical Report MSR-TR-2005-166 December 2005

(71) Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In Proceedings of the nineteenth ACM symposium on Operating systems principles (pp. 29-43).

(72) David A. Patterson, Garth A. Gibson, and Randy H. Katz. A case for redundant arrays of inexpensive disks (RAID). In Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data, pages 109–116, Chicago, Illinois, September 1988.

(73) R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse. “Fundamental limits of DNA storage systems”. In: IEEE International Symposium on Information Theory (ISIT). 2017, pp. 3130– 3134.

(74) I. Shomorony and R. Heckel. “Capacity Results for the Noisy Shuffling Channel”. In: IEEE International Symposium on Information Theory (ISIT). 2019.

(75) Yang, K.; McCloskey, C. M.; Chaput, J. C. Reading and Writing Digital Information in TNA. ACS Synth. Biol. 2020, acssynbio.0c00361. https://doi.org/10.1021/acssynbio.0c00361.

(76) Epple, C., Leumann, C., 1998. Bicyclo(3.2.1]-DNA, a new DNA analog with a rigid backbone and flexibly linked bases: pairing properties with complementary DNA. Chemistry & Biology 5, 209–216.. doi:10.1016/s1074-5521(98)90634-2

APPENDIX

APPENDIX A

Cost calculation:

1 base pair (bp) = 4 bits

8 bits = 1 byte

Therefore,

2 bp = 1 byte

In our case,

1 strand = 60 bp = $\frac{60\ bp}{2\ bp/byte}$ = 30 bytes

Phosphoramidite synthesis:

Cost of 1.5E16 strands = $10.80

1 byte = 1E-15 Petabytes (PB)

1.5E16 strands = $1.5E16\ strands * 30\,\frac{bytes}{strand}$ = $4.5E17\ bytes * 1E - 15\,\frac{PB}{byte}$ = 450 PB

Therefore,

Cost of synthesis of 1 PB DNA = $(\frac{\$10.80}{1.5E16\ strands}) * (\frac{1.5E16\ strands}{450\ PB})$ = **$0.024 /PB**

Our synthesis method:

Cost of 9E13 strands = $2.97

1 byte = 1E-15 PB

9E13 strands = $9E13\ strands * 30\,\frac{bytes}{strand} = 2.7E15\ bytes * 1E - 15\,\frac{PB}{byte}$ = 2.7 PB

Therefore,

Cost of synthesis of 1 PB DNA = $\left(\frac{\$2.97}{9E13\ strands}\right) * \left(\frac{9E13\ strands}{2.7\ PB}\right)$ = **$1.1/PB**

SUPPLEMENTARY INFORMATION

Additional details of the encoding method and analysis. RS codes are linear codes and specified in terms of three parameters, RS[n, k, d]. n is the length of the coded message, k is the original message length, and d is the minimum distance of any two blocks and equals n-k+1. The distance, d, indicates the error correcting capability of the code. It is worth noting that n is restricted to be of the form q^m-1, where q is a prime number and m is an integer greater than or equal to 1, but k or d can be chosen arbitrarily. Genomic DNA is fundamentally a base-4 code, and hence n may be chosen to be one less than an even power of 2 in DNA storage systems.

Error correction techniques usually differentiate two kinds of errors, symbol errors and symbol erasures. An erasure occurs when the decoder has extra knowledge that an input symbol is missing or erroneous. For example, an erasure occurs if a strand is never sequenced and therefore is never provided to the decoder. The decoder algorithm knows the strand is absent and can leverage that to increase the likelihood of correct decoding. Erasure-only decoding is common in electronic file systems where hard disks are often modeled to either succeed or fail as an entire unit.

A symbol error is what occurs in any other case when no such erasure knowledge is present. The input symbol to the decoder differs from what was originally encoded, and the decoder is unaware of the error and must identify that the symbol is incorrect and fix it through the decoding process.

The RS decoder is less capable of correcting symbol errors than erasures. A maximum of d erasures can be corrected, whereas, a maximum of only (d-1)/2 symbol errors can be corrected. Any combination of erasures and errors can be corrected such that d > 2*e_{sym} + $e_{erasure}$, were $e_{erasure}$ is the number of erasures and e_{sym} are all other errors.

An RS code can be designed to tolerate some number of errors by setting the distance parameter, d, large enough. The probability that it fails to correct can be estimated using the decoder error probability of an RS code[65,66,70], given by P_E([n,k,d], p_{error}, $p_{erasure}$), which calculates the probability that the number of errors and erasures exceeds the bound required for correct decoding, namely d ≥ 2*e_{sym} + e_{erase} - 1 given the block length, number of error correction symbols, and the probability of errors, p_{error}, and erasures, $p_{erasure}$.

P_E can be estimated by treating the number of errors and erasures as a random variable following a binomial distribution[65,66]. As a reminder, the binomial probability mass function for a random variable X is given by:

$$f(t;n;p) = \Pr(X = t) = \binom{n}{t} p^t (1-p)^{n-t}$$

The cumulative binomial distribution, denoted $F(t;n;p) = \Pr\,(X \leq t)$, is the summation of $f(t;n;p)$ over the range where $X \leq t$. Then, for an RS[n, k, d] code, and using the substitution $t = \frac{d-1}{2}$:

$$P_{UE}([n,d,k], p_{error}, p_{erasure})$$
$$= (1 - F(t, n, p_{error})) + \sum_{i=0}^{t} f(i;t;p_{error}) \times (1 - F(2(t-i);n;p_{erasure}))$$

Simply stated, P_{UE} is sum of the probabilities of either more than t errors or a combination of errors and erasures such that the number of erasures plus twice the errors is greater or equal to the distance, d, of the code.

The decoder error probability is dependent on the likelihood of symbol errors and erasures in the message sent to a decoder. We can estimate p_{error} and $p_{erasure}$ based on previous studies[64,67,71,72]. The probability of a single base error in a DNA strand is a function of the combined effects of synthesis and sequencing errors and has been measured empirically to be in the range of 10^{-3} to 10^{-2}. For the case of probability of strand loss or breakage, we sweep the likely error ranges predicted from empirical measurements of DNA stability to estimate $p_{erasure}$.

DNA storage systems often use a combination of both inner and outer RS codes. The inner RS code protects against errors within a strand and the outer code corrects for missing or erroneous strands. The block length of an inner code, n_{inner}, and its error correcting ability, d_{inner}, would be limited to the number of symbols that fit in a strand of a given length and the requirement that an index be present within each strand[67,73,76]. The index must be large enough to uniquely identify each strand of a file, and this is on the order of O(log M), where M is the desired number of strands per file.

The outer RS code is formed as a set of n_{outer} strands, wherein k_{outer} strands hold the information and n_{outer}-k_{outer} strands hold the additional error correction symbols required for the RS code. If a strand is never sequenced or discarded due to too many errors by the RS inner decoder, such a missing strand is treated as an erasure by the outer code. If fewer than d_{outer} strands are lost in a block, the RS outer decoder may recover the block provided that no strands contain erroneous data. If some strands are missing and some strands are erroneous, the previously stated relationship, $d_{outer} > 2*e_{sym} + e_{erase}$, also applies to the outer code as well.

Also, DNA storage systems incorporate in each strand an index that indicates which part of a file a strand corresponds to[58,64,68]. The index must be large enough to uniquely partition all of the data for a large file across strands. The index may be considered part of the data with respect to error correction and protected using the inner RS code. However, with respect to information density, it is usually considered overhead. Hence, the information density, or Rate, of the code is estimated as: $(k_{inner} - i)*k_{outer} / (n_{inner} * n_{outer})$, where i is the number of symbols devoted to the index. Figure **11A** illustrates this ratio as white data units divided by all of the other symbols in the code. Figure **11C** and 10D use this formula to calculate information density assuming a copy number per strand of 1.