Information Retrieval Methods for DNA Data Storage

Lee Wohlen Organick

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Luis Ceze, Chair
Karin Strauss
Jeff Nivala

Program Authorized to Offer Degree:
Paul G. Allen School for Computer Science and Engineering

University of Washington

Abstract

Information Retrieval Methods for DNA Data Storage

Lee Wohlen Organick

Chair of the Supervisory Committee:
Luis Ceze
Paul G. Allen School of Computer Science and Engineering

This thesis presents methods of data retrieval in DNA data storage systems. While DNA has been considered a promising data storage medium for decades for its data density, durability, programmability, and eternal relevance, the cost has, until recently, been much too high for implementation. Now that the cost of DNA synthesis and sequencing has fallen significantly, the field has begun to examine DNA data storage at meaningful scale. This thesis includes scaling up PCR-based random access, evaluating PCR-based random access performance under increasingly sparse and complex conditions, exploring the feasibility of implementing a CRISPR-Cas9-based similarity-search data retrieval method, and finally, comparing DNA archival methods to successfully retrieve data after long periods of time. The work presented here is part of a small but quickly growing body of work showing the scalability and viability of DNA data storage, with broader application to the world of synthetic biology and the many other fields where DNA is present.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

DSDNA: Double-stranded DNA.

LIBRARY: A collection of DNA. In this document, synonymous with *pool*.

MULTIPLEXABILITY: The ability to retrieve multiple files in a single reaction.

POLYMERASE: An enzyme that replicates DNA.

POOL: A collection of DNA. In this document, synonymous with *library*.

PRIMER: A short ssDNA strand, typically 20nt, used during PCR to signify the start or end of a sequence to be amplified.

SEQUENCE: A sequence of nucleotides that form a particular species of DNA. Many strands of DNA may have the same sequence.

SSDNA: Single-stranded DNA.

STRAND: An individual, whole, piece of DNA. This strand may share the same sequence as other strands.

ACKNOWLEDGMENTS

This thesis would not have been possible without the collaboration of many scientists and engineers across many countries and disciplines.

Special thanks to my advisors Karin Strauss and Luis Ceze. I cannot thank you enough for letting me join MISL before it was MISL back in 2015, and for your unwavering support, enthusiasm, and encouragement since then.

Special thanks also to Jeff Nivala who was my advisor for the similarity search project. Thank you for your optimism and for being so quick to make a joke.

For the work presented in Chapter 3, this work would not have been possible without co-authors Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Chris Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Callie Bee, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Karin Strauss, and Luis Ceze. It was a pleasure getting to work with you all!

For the work presented in Chapter 4, this work would not have been possible without co-authors Yuan-Jyue Chen, Siena Dumas Ang, Randolph Lopez, Xiaomeng Liu, Karin Strauss, and Luis Ceze. Thank you all for your mentorship, your advice, and your patience as I filled up many a white-board with plans and calculations. Thanks also to Leila Zelnick for your help in reviewing statistics, Sergey Yekhanin for your help in troubleshooting the small file decoding behavior, Melissa Queen for your helpful discussions regarding logical and physical density. Thank you also to Jennifer Brennan and Nick Nuechterlein for your insights on modeling data, and Chris Takahashi for your help finding bugs in the code.

For the work presented in Chapter 5, this was a true team effort that would not have

been possible without Callie Bee, Jeff McBride, Jessica Dunstan, Luis Ceze, Karin Strauss, and Jeff Nivala. Thank you all for all your patience and help. I have learned so much from all of you and I cannot thank you all enough!

For the work presented in Chapter 6, this work would not have been possible without co-authors Bichlien Nguyen, Rachel McAmis, Weida Chen, A. Xavier Kohll, Siena Dumas Ang, Robert Grass, Luis Ceze, and Karin Strauss. Thank you all for making such an involved experiment so much fun!

Chapter 1

INTRODUCTION

1.1 The Current State of Data Storage

Of the 64 zettabytes (ZB) of data created in 2020, it is estimated that only two percent persisted into 2021 [1]. By 2025, 175 ZB of data creation is predicted [1]. Even if the two percent trend persists, that is still multiple zettabytes of data stored per year. For context, in 2013 Facebook built an entire warehouse to hold one exabyte (0.1% of one ZB) [2].

Of the small percentage of data that is stored, some data are *hot*, that is, frequently accessed, while the majority of data are *cold*, hardly if ever accessed [3]. In current commercial practices, cold data intended for archival storage are written to magnetic tape and stored in warehouses kept at carefully controlled temperature and humidity levels.

Magnetic tape is used for two main reasons: its high data density, and its durability. However, with the ever increasing demand for archival data storage, there is great interest in creating a storage medium even denser and more durable. Ideally, this storage medium would also not require the costly maintenance like climate controlled warehouses that magnetic tape requires.

Yet there is a third category of data that is *warm*, only sometimes accessed. As more and more data are generated each year, there's pressure to not only store this data, but to search through it efficiently. As databases grow ever larger, searching through data is becoming much more time and energy intensive, leading to increased interest in finding a new ways to efficiently search through data.

The solution to both archival data storage and very large data search may be to employ DNA as a storage and search medium.

1.2 Why Use DNA?

DNA is an excellent candidate for data storage and subsequent data retrieval for four main reasons.

First, DNA is able to be used as a data storage and searchable medium because of its programmability and predictable kinetics. Since the 1970's we have been able to sequence (*read*) DNA, and since the 1980's, we have been able to synthesize (*write*) DNA sequences of our choosing - though both have various limitations that persist today [4]. While the basics of DNA kinetics have been known since the early 1900s, increasingly precise models of various DNA interactions are being developed and refined and enable the design and implementation of DNA based computations, some of which will be discussed in Chapters 2 and 5.

Second, DNA is incredibly durable. In contrast to tape's roughly decade-long durability, DNA has been recovered from multiple mammoth specimens over one million (M) years old [5]. More recent work has examined various ways to store synthetic DNA and will be covered in Chapters 2 and 6, but in essence, in the right conditions it is trivial to store and recover DNA after many hundreds of years with no climate control. This means that in addition to not rewriting data every decade or so as is common with current magnetic tape storage, data stored in DNA could easily be kept at a wide range of conditions, effectively eliminating the need for costly climate control.

Third, DNA is orders of magnitude more data dense than current data storage methods, with a measured density of approximately 17 exabytes/gram [6, 7]. At this density, an entire super-sized warehouse of magnetic tape could instead be stored in DNA with a footprint roughly the size of a sugar cube. This work will be covered in more detail in Chapter 4.

Fourth, because of the biochemical properties of DNA, parallelism is an innate feature of DNA. When searching through large amounts of data, this parallelism is immensely valuable, and will be touched on throughout this thesis.

1.3 Summary of Contributions

My PhD work closely examines DNA retrieval methods in the context of data storage and search, and broadly explores the following two hypotheses:

1. DNA retrieval using polymerase chain reaction (PCR) enables random access information retrieval and supports a DNA data storage architecture that is information dense, efficient and durable data storage.

2. DNA retrieval using CRISPR-Cas9 enables similarity search that is energy efficient and rapid, but has less accurate recall than previous work implementing similarity search in DNA.

To summarize the contributions of each chapter very briefly:

- Chapter 2 provides the high-level background information needed to understand the work presented in the rest of the dissertation.

- Chapter 3 presents work exploring PCR random access with a dataset roughly three orders of magnitude larger than previous DNA random access work. The encoding method, sequence architecture method, error profiles, and considerations for implementing future large-scale DNA random access systems are described in detail.

- Chapter 4 presents work that closely examines how many copies of each DNA sequence are required for successful data retrieval per PCR random access reaction. PCR random access in more complex conditions is also examined.

- Chapter 5 presents simulation data using CRISPR-Cas9 for similarity search, comparing its recall performance, speed, and energy efficiency to prior work.

- Chapter 6 presents work closely examining the durability of DNA data storage systems over time with nine different preservation methods. The error profile for each preservation is presented in detail.

- Chapter 7 concludes the dissertation, examining the implications and practical constraints of the work presented here.

Chapter 2

BACKGROUND

2.1 A Brief History and Primer of PCR

To understand the nuances of using PCR for random access, and in certain contexts why CRISPR-Cas9 is an attractive alternative to PCR-based DNA retrieval, one must first understand what PCR is and some of the properties of DNA that allow it to be amplified[1].

Broadly, polymerase chain reaction (PCR) is a technique to copy specific sequences of DNA. As noted in Chapter 1.2, we could sequence DNA by the 1970's, but it was as destructive of a process then as it largely remains today. Huge biological samples had to be collected to have sufficient DNA to prepare for sequencing, and once the DNA was isolated, there was usually a genome's worth of material to sift through when often only a small proportion was of interest. Using enzymes called *polymerases*, whose purpose is to duplicate DNA for cell division, initially discovered in hyperthermophilic bacteria and archaea living in heated sea floors, hot springs, and other extreme conditions [8], the story goes that in the spring of 1983, Kary Mullis came up with the idea of the modern PCR technique while driving [9].

Using DNA polymerases, a DNA sample, and two short fragments DNA known as *primers* (along with other reagents that are unimportant for the discussion of this project), a specific region of DNA can be easily amplified to have on the order of 10^{10} copies in roughly an hour. The primers are known as *forward* (or, *front*) and *reverse* (or, *back*) primers, where the forward primer is identical to the first 20 bases of the DNA region you'd like to amplify, and the reverse primer is the reverse complement[2] of the last 20 bases you'd like to amplify.

---

[1] If you are already an expert in DNA and PCR, you may wish to skip to Section 2.1.2

[2] DNA has four bases: A, T, G and C. As and Ts bind to each other and Gs and Cs bind to each other. DNA also has directionality, one terminus is noted as 3' and the other 5'. A sequence of DNA reading 5' ATTGC 3' would bind to its reverse complement, here, the sequence 5' GCAAT 3'. Standard notation

To implement PCR, a series of temperature cycles must be performed. In each cycle there are three steps, each at a different temperature and slightly different duration, though generally on the order of 20-30 seconds for this work. In the first step, DNA is heated to at least 94°C and denatured so that all DNA is single-stranded. In the second step, the mixture is cooled to approximately 50-60°C where it is now energetically favorable for the primers to anneal to their complementary DNA sequences. In the last step, the mixture is heated slightly (typically near 70°C) and the polymerase attaches to the double stranded part of the sequence where the primer has bound and extends the DNA so the rest of the sequence is now double-stranded. In each cycle, the number of amplified strands is doubled. With this exponential growth, if you started with 100 copies of your sequence of interest, after just 15 cycles you'd have 3.3 M copies ($100 * 2^{15}$).

While the first use of PCR was for diagnosing sickle cell anemia [10], the uses for PCR go far beyond diagnostics. Now, PCR is virtually indispensable to the fields of synthetic biology, molecular programming, and arguably any field that utilizes DNA.

2.2 A Brief History of DNA Data Storage and PCR Random Access

Why use PCR for random access? Say you have 10^{10} DNA sequences[3] but you only want to sequence a small fraction of those sequences. DNA sequencing remains relatively expensive [11], so rather than needlessly waste resources sequencing the entire DNA library, PCR can be used to amplify only the sequence(s) of interest. After PCR, instead of a library with each sequence being present roughly equally, your library is now heavily biased towards your sequence(s) of interest, which means your sequencing results are now almost exclusively your sequences of interest and thus your sequencing resources are utilized much more efficiently.

---

dictates that all written sequences should be read 5' to 3' unless otherwise specified.

[3]The terms *DNA sequences* and *DNA strands* are not interchangeable in this work. DNA strands refer to the individual, physical strands of DNA. DNA sequences refer to the sequence of bases that make up a strand. For example, "I have 100 strands of DNA. In that mixture, I have two DNA sequences and each has 50 copies."

In a typical PCR protocol, an aliquot of the original pool[4] is taken so that the original pool does not become biased.

Though DNA data storage had been proposed as early as the 1960's [12, 13], it wasn't until the 1990's that data was experimentally encoded in DNA [14, 15], and the early 2010's when sequencing and synthesis became cheap enough that experiments were able to scale to hundreds of kilobytes (KB) of data.

However, early work did not include random access [16–19] and was thus not suitable for large scale archival data storage.

In 2015, a DNA data storage architecture with random access was published [20], encoding 3 KB of data. In this work, the traditional strand architecture of [primer - payload - primer] was employed successfully. Then, in 2016 the same general strand architecture was employed on a 150 KB dataset [21], though with a different data encoding scheme. The work successfully demonstrated the potential for archival, large scale DNA data storage. However promising on this small scale, the question of scalability became more pressing. Namely, would this architecture work on a database several orders of magnitude larger, and with more files? This is explored in detail in Chapter 3.

In short, results from this large scale experiment (to date the largest implementation of DNA data storage) were encouraging. However, it still left many important questions unanswered. Could this architecture efficiently retrieve files from a highly complex library where the file of interest only comprised a minuscule percentage of DNA strands before PCR? Would this architecture require such a large physical redundancy (number of copies of each DNA sequence) that it wouldn't be practical to implement at a large scale? And if this architecture was used, how would one store the DNA for optimal durability? I have since explored all of these questions and present answers to them in Chapter 4. It is important to note that all my work utilizes the general sequence architecture of [forward primer - payload - reverse primer], but there are several variations of this architecture that are worth noting

[4]The words *pool* and *library* are used interchangeably in this document, both referring to a collection of DNA.

and they are discussed below.

2.3 A Brief History of Alternative DNA Random Access Methods

While careful primer selection, as utilized in the work presented in this dissertation, may yield higher fidelity sequence retrieval, it drastically limits the number of primers available and thus the number of files that can be encoded in one pool. To increase the number of files with unique file IDs while preserving stringent primer requirements, there are currently two main approaches to sequence architecture: hierarchical and combinatorial primer regions.

While traditional primer architecture is [primer A - payload - primer B], in a hierarchical (also known as a *nested*) primer method strand architecture resembles the following: [primer A - primer B - payload - primer C], though any number of forward or reverse primer regions could theoretically be used. With the same 28,000 primers designed in the work presented in Chapter 3 that can identify a maximum of 14,000 files, using two forward primers and one universal reverse primer can identify a maximum of 783,916,002 files [22]. Other work has explored using only six forward primers [23]. However, while these methods can encode many more files, they lack the simple ease of random access that a more traditional primer design with PCR offers. Because of the increased number of primer regions (or no back primer, which then only allows linear, not exponential, amplification), one simple round of PCR would amplify multiple files, as any given primer region would be present in at least two files depending on the number of primer regions used. In these hierarchical methods, magnetic beads are used to physically isolate the sequences of interest (essentially, complementary primer regions are bound to magnetic beads so when they bind to the primer(s) of interest, a magnet can be used to pull all retrieved DNA to the magnet and the rest of the material can be removed). Unfortunately, magnetic bead extraction has three main downsides: the reagents are more expensive, they require more sequence space (which causes each DNA strand to be more expensive to synthesize, and also more errors in synthesis as the rate of errors increases with strand length), and retrieval has significantly less fidelity, often resulting in approximately 15% of sequencing reads being irrelevant. However, the ability to physically

isolate strands of interest in a less destructive manner than PCR is attractive (recall that PCR heavily biases a sample, so typically an aliquot of the original pool is used to perform PCR). Furthermore, at a great enough scale, it is possible that the cost of many rounds of PCR to bring the proportion of your file of interest sufficiently high might exceed the cost of performing one bead extraction with follow up PCR to amplify the bead-retrieved strands.

To avoid the use of magnetic bead extraction, we have recently explored a combinatorial primer approach (not presented in this dissertation) in which the traditional [primer - payload - primer] architecture is used, but in this work we allowed for any unique combination of forward and reverse primers to be used [24]. Thus with the same 28,000 primers design in previous work [25], we can encode up to 196,000,000 files ($14,000^2$). Encouragingly, in a small scale implementation of 81 files we found no evidence of a practical drop in retrieval fidelity, and with multiple rounds of PCR this technique could scale well to be implemented with very large databases.

Other groups have used magnetic bead extraction not to perform random access of a specific file, but to perform a selection based on metadata tags [26]. In this work, each file's DNA is encapsulated in nanoparticles, such as silica beads, and the bead surface is tagged with DNA sequences corresponding to metadata such as 'cat' and 'orange'. It is then possible for one to easily extract all the DNA associated with orange cats. While this technique would be difficult to scale due to the difficulty of manually tagging each item, it is an intriguing data storage method due to its selectivity and accurate recall.

Another recent work exploits usually unwanted variations in PCR thermodynamics to enable a file preview in a DNA database [27]. Here, the authors showed that by tuning concentration and PCR temperature, one could retrieve either the entire encoded file, or retrieve only a subset of sequences to allow for a low resolution file preview. While again difficult to scale to a large dataset due to the precise requirements of the PCR primers, this is another data retrieval method the field would do well to keep in mind as DNA data storage matures and becomes more viable with the dropping price of synthesis and sequencing.

Recent work still under review has also implemented random access with CRISPR-Ca9

rather than PCR [28] and will be discussed further in Section 2.6.

The creative use of molecular processes is a recurring theme in this quickly evolving space of DNA data storage and retrieval.

2.4 A Brief History of Examining DNA Durability

A concern recognized early in the field of DNA data storage was how to preserve DNA so that a DNA database could be preserved and be retrieved after preservation without catastrophic error [29, 30]. With PCR, the concern is that if the DNA backbone breaks, when the solution is heated the DNA will separate into fragments. With the template fragmented, a primer will only (1) be able to linearly amplify (2) a fragment of the strand. These fragments will likely not be able to be used in the decoding process because they were not amplified enough times to be read by the sequencer compared to the exponentially amplified strands, though a painstaking retrieval process could likely be implemented if necessary, and if the strand breaks largely only occur once per strand.

While the field of DNA preservation dates back to at least the 1960s when biological specimens were being shipped to various labs via filter paper [31], the focus has largely been in biological DNA. Synthetic DNA differs in that (1) it is typically much shorter in length, (2) it has different sequence profiles with fewer, if any, repeats or homopolymers, and (3) it has different contaminants that drastically impact DNA durability. However, all DNA has the same basic mechanisms of damage. Best practice has long dictated that DNA be kept away from all UV radiation, oxidative reagents like water, radiation, and mechanical sheer (though normally mechanical sheer is a greater concern for DNA strands on the order of thousands of bases long) [32].

However, when various groups began exploring methods to store synthetic DNA for long term data storage, there were two main problems with these methods in the DNA data storage context.

First, some of the storage methods explored are not information dense. Though the tolerance for information density will vary, one method encapsulates DNA in small, hermet-

ically sealed tubes, thus drastically lowering the physical information density [33, 34], while another encapsulates DNA in silica beads (as alluded to in Section 2.1.3) [18] which lowers information density to a much lesser degree. Notably, both methods are also more difficult to automate than, for example, more information-dense methods requiring only the mixture of sugars. And importantly, the de-encapsulation protocol for the beads used in this work requires hydrofluoric acid (HF), a particularly dangerous de-encapsulation agent[5], to dissolve the bead and retrieve DNA. While at small scales (i.e., in standard de-encapsulation protocols in standard labs today) the process is relatively safe, as only low concentrations and quantities of HF are used per de-encapsulation process. However, to perform each protocol, a dangerous amount of HF is required to be present and thus requires a non-trivial amount of caution. Notably, if this method were to be scaled up significantly, the handling of so much HF would need to be taken into consideration.

Second, storage methods are typically presented individually, and rarely directly compared to one another [29, 35, 36]. Without direct comparisons, subtle differences in lab protocols can drastically influence the reported half-life of preservation methods and make comparing methods' durability virtually impossible. In our work (presented in Chapter 6), we stored two files in DNA and directly compared aging samples preserved with a variety of methods, and importantly, unlike previous work we examine sequencing data from these aged samples. This was important to understanding important details about file recovery and sequence design as well as providing the most comprehensive comparison of preservation methods to date.

2.5 A Brief History of Similarity Search With and Without DNA

Sometimes a file will be recovered based not on its ID as with simple random access, but its content. In silicon-based computers, a classic implementation of this might be with approximate nearest neighbor algorithms, known as k-NN algorithms, or just *similarity search*. In

[5]HF easily passes through organic material including skin and can lead to terrible injury or death if not treated immediately.

these algorithms, k is the number of items to be returned that most closely resemble the query item. While previously mentioned work in the molecular programming space utilized metadata-tagged encapsulated DNA [26], this was not the same as implementing similarity search.

Similarity search is more scalable than the comparatively simple metadata tagging problem presented in that previous work. In similarity search, machine learning can used for a more scalable approach, no longer requiring precise metadata labels, but instead organizing documents added to a database by their relative distances from one another. Various similarity search algorithms have been developed for use in silicon-based computers [37], but one commonality is that at a very large scale the algorithms become much slower as they sift through the data.

Document similarity is typically framed as a geometric problem [38], where the document's contents are summarized by a vector (known as a *feature vector*) with neighboring feature vectors representing similar documents. With images, for example, neural networks trained on a large library of real-world images can be generalized to other real-world images, making it a relatively simple task to simple feed these pre-trained networks previously unseen images and obtain useful output feature vectors [39, 40].

Using the pre-trained image classification network VGG16 [39], recent work has, for the first time, demonstrated similarity search in DNA [41]. Briefly, in this work, Bee et al. took 1.6 M images[6] and ran them through VGG16. The resulting 4096-dimensional feature vectors from the FC2 layer were trained to now encode a feature vector of length 80 not in bits, but in nucleotides. This work leveraged NUPACK, a previously built model that predicts DNA hybridization kinetics [42], so that a query image's DNA sequence is complementary to similar images in the encoded database. Essentially, each image has a feature vector space and a lookup ID so that when the feature vector sequence is read, the data corresponding to that file can be accessed in another, separate database. In a

[6]The number 1.6 M is a recurring one because currently the largest commercial synthesis method can synthesize 1.6 M sequences on one chip.

database of 1.6 M encoded images, one would then pass a query image through the encoder and take the reverse complement. Because DNA will often hybridize to sequences that are similar (Hamming/Euclidean distance is a good proxy, but there are enough nuances that a hybridization model is necessary to predict actual kinetics [42]), we can introduce magnetic beads with the query sequence bound to the surface and slowly let the reaction cool from 95°C to 21°C at a rate of 1°C per 20 min. At the end of this reaction, a magnet physically separates all the beads, now bound to strands from the database, and those strands are sequenced to reveal the file IDs we can now go look up.

Remarkably, this approach worked and is comparable to state-of-the-art in silico similarity search algorithms (such as hnsw, annoy, or rpforest). Encouragingly, because DNA strands in solution diffuse and constantly interact with any strand they come into contact with, the parallelism of this DNA-based method and density of DNA means that this 1.6 M image database could be scaled up and have recall times competitive with state-of-the-art similarity search algorithms, assuming the DNA was already synthesized. However, because the protocol requires the solution be heated to 95°C, at a very large scale (on the order of milliliters or liters, unlike the microliters used in this work) the energy costs of this method, while still much better than a traditional silicon-based computer, would be very high and potentially prohibitive. This led us to wonder if we could implement similarity search not based on DNA hybridization coupled with bead extraction, but instead with CRISPR-Cas9, as Cas9 reactions can happen at nearly room temperature. The answer to that is presented in Chapter 5.

2.6 A Brief History and Primer of CRISPR

In 2020, Jennifer Doudna and Emmanuelle Charpentier won the Nobel Prize in Chemistry for developing the CRISPR-Cas9 gene editing tool, which, as we'll discuss, is used for much more than just gene editing. Clustered regularly interspaced short palindromic repeat (CRISPR) systems are commonly found in nature as acquired immunity systems in archaea and bacteria [43]. In a striking resemblance to the development of PCR, hyperthermophilic organisms were

again the key to developing a world-changing technology. Several genes thought to encode DNA repair proteins in hyperthermophilic bacteria and archaea were found to be strictly associated with CRISPR and were designated *cas* (CRISPR-associated) genes [44]. This allowed for the discovery that Cas proteins (products of *cas* genes) work together to form an acquired immune system in prokaryotes.

Essentially, for our purposes, this system works by having a protein (Cas9) with a strand of RNA approximately 18-20 nt long (known as the *guide RNA*) attached to it act as a kind of molecular scissors. This complex will interact with double-stranded DNA strands it comes in contact with, and if there is a PAM [7] site (a short sequence of *NGG* [8] in the case of Cas9) 3' adjacent to a sequence of DNA complementary with the guide sequence, the double stranded DNA will be cleaved at a site a few bases upstream of the PAM site [45] (the exact distance is Cas-dependent).

CRISPR systems can be far more complicated than simple 'molecular scissors' though. In organisms, cleaved DNA may be repaired with either non-homologous end joining (often resulting in random insertion or deletion of DNA), or homology directed repair (in which a replacement sequence is provided by Cas9 and used as a repair template). Homology directed repair is exceptionally useful for gene editing. With gene editing, one can heal certain genetic diseases by replacing the offending DNA sequence, induce *Saccharomyces cerevisiae* (a common yeast) to synthesize desired compounds, or any number of other useful functions [46].

In addition, Cas9 is not the only Cas protein, and there are different versions of Cas9 as well [43]. For the purposes of this work we will ignore these other Cas proteins as they either require *in vivo* conditions to function, induce non-specific cleavage (i.e., Cas12), or, as in the case of virtually all variants of Cas9, are more precise in their cleavage and thus have fewer off-target effects. When using CRISPR-Cas9 as a tool for similarity search, we do

[7] PAM stands for *protospacer adjacent motif*, functioning as a kind of flag for CRISPR complexes to come investigate an area for complementary sequences and allowing cleavage to occur if there is a complementary sequence.

[8] N refers to a 'wildcard', a symbol to represent any base.

not want an *in vivo* system for a plethora of reasons (i.e., much more complex interactions that we still cannot fully predict, and low information density when compared to synthetic systems). Non-specific cleavage, while useful when acting as a reporter [9] [47], is not useful for our purposes of selective data retrieval. Conversely, a Cas protein with too specific of cleavage would not be useful for similarity search, and would instead be better suited for precise random access of one exact sequence. While the vast majority of the world seeks to discover or engineer increasingly precise Cas proteins for delicate tasks such as gene editing, this work seeks to utilize the less precise wild-type Cas9 to take advantage of the slight cleavage imperfections and leverage this variation for similarity search.

Conversely, as previously mentioned, other work from members of our lab has leveraged the precision of Cas9 to perform random access of individual files of interest, much like our previous work using PCR for random access [28]. Unlike PCR, Cas9 has two major advantages. First, recall that PCR protocols require the solution to be heated to nearly boiling and proceed through several heat cycles, while Cas9 is activated near room temperature. For this reason, energy savings for large scale retrieval could be enormous using Cas9 rather than PCR. Another advantage of Cas9 over PCR is its ability to perform multiplex reactions, meaning that more than one file could be accessed at a time with Cas9. Attempting to retrieve multiple files at once with PCR leads to exceptionally uneven sequencing coverage, with some files present nearly an order of magnitude more than others, which means sequencing resources are wasted in an attempt to read all files from the same reaction [48]. Nevertheless, the simplicity of PCR is still an attractive alternative to Cas9, and Cas9's use of RNA guides makes handling protocols much more strict, as RNA is much more subject to degradation and must be kept in much more sterile conditions.

[9] Typically, Cas12 is used in diagnostic or other reporter settings. Once the presence of a guide's complement is detected by Cas12, it begins cleaving DNA indiscriminately. Diagnostic settings take advantage of this by flooding the system with fluorescent-bound DNA such that when it is cleaved it begins to fluoresce, leading to a convenient, light-based detection system.

2.7 A Brief History of Cas9 Predictor Models

If we are to use Cas9 in our work, we must first be able to predict Cas9 activity. While we know as a general rule Cas9 will cleave a perfectly complementary sequence with a PAM site, we know that Cas9 is not perfect and must understand the nuances of its activity to design a system we hope can take advantage of this nuance.

Prior to 2021, most Cas activity was measured in genomic contexts. Comparing the off-target effects of Cas proteins was virtually impossible due to the use of different DNA and conditions (i.e., different organisms, different DNA libraries if synthetic DNA was even used, different reaction conditions, etc.) and often the use of readout methods that only indirectly measure activity [49]. As the number of Cas proteins available has grown and scientists attempt to compare Cas activity, it became clear that a benchmark tool was necessary. Much like the ANN-benchmark tool mentioned before for benchmarking similarity search [37], a CRISPR-Cas benchmark tool called NucleaSeq was published in 2021 [49].

Like ANN, NucleaSeq provides a standard for measuring performance allowing new developments to be easily and accurately compared. In NucleaSeq, over 10,000 DNA targets containing mismatches, insertions and deletions relative to the guide RNA are examined, with the cleavage specificity of various Cas proteins then able to be compared exactly to one another because sequences are identified and categorized by flanking barcodes that provide a direct measurement of Cas activity (unlike previous work which used read counts as an indicator). Due to their construction of an interpretable biophysical model, they showed that mismatches further away from the PAM region (PAM-distal) have less effect on cleavage than PAM-proximal mismatches, and that wtCas9 tolerates rG-dT mismatches (when an RNA guide's G is paired with a DNA target's T, rather than the canonical dC we normally expect) more than other mismatches, among other insights that are not relevant to our work with similarity search.

Nucleaseq also helpfully provides a function where one can specify the Cas protein, PAM sequence, gRNA sequence and DNA sequence, and it outputs a cleavage score. In this

output, zero is the maximum cleavage rate and negative infinity is the minimum cleavage rate (though in practice the model's minimum score is clipped to approximately -3.7). The higher the score, the more the gRNA-Cas complex will cleave the DNA sequence. In general, the most dissimilar the gRNA and DNA sequence are, the less cleavage will occur and the lower the score will be, however, this rule of thumb should not be relied on. Changing a single base near the PAM sequence, for example, will have much more profound impact on cleavage rate than changing multiple bases far from the PAM sequence.

While the use of a synthetic library of approximately 10,000 sequences and only a few different guide sequences is a strong start to modeling Cas behavior, we are conscious that there may be more nuances to discover with a larger library but are encouraged by the model's agreement with previous work and use of a synthetic environment that most closely resembles the environment we plan to use [49].

2.8 A Brief Primer of Metric Learning

Though not directly related to DNA search methods, our Cas9-based similarity search method described in Chapter 5 employs metric learning and will be briefly discussed here for context.

Metric learning strives to learn a representation function that maps objects into an embedded space such that the more distance there is between objects, the more dissimilar they are. Conversely, the more similar two object are, the less distance there is between them. This is exactly the task at hand when encoding a dataset to be used for similarity search in DNA - in our context, similar images will have similar Cas9 activity, while dissimilar images will have dissimilar Cas9 activity.

Mapping objects based on their similarity is inherently difficult, even for humans. When comparing images, one may wish to focus on subject similarity for one context, and, say, background similarity for another context. To take a more holistic approach to comparing objects, it is a much easier task to collect pairs of objects and label them 'similar' or 'dissimilar', or even take triplets of objects and form conclusions like, 'x is more similar to y

than z' [50]. In fact, it appears that the method in which metric learning models are trained matters a great deal in facilitating an encoding scheme that embeds objects in space by their similarity, with triplet loss (the x vs y vs z method mentioned) having been shown to be a useful, at times crucial, tool for training models [51].

Chapter 3

RANDOM ACCESS IN LARGE-SCALE DNA DATA STORAGE

A version of this work was originally presented in Nature Biotechnology, Organick et al. 2018 [25]. In addition to myself, the co-authors for this work are Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss.

3.1 Chapter Abstract

Here, we encode and store 35 distinct files (over 200 MB of data), in more than 13 million DNA oligonucleotides, and show that we can recover each file individually and with no errors, using a random access approach. We design and validate a large library of primers that enable individual recovery of all files stored within the DNA. We also develop an algorithm that greatly reduces the sequencing read coverage required for error-free decoding by maximizing information from all sequence reads. These advances demonstrate a viable, large-scale system for DNA data storage and retrieval.

3.2 Overview

Storing digital data using synthetic DNA requires information to be encoded into nucleotide sequences and the corresponding molecules to be synthesized and stored in an appropriate environment. To extract the stored information, one has to sequence the DNA and decode it back into digital data. Here, we provide an end-to-end DNA storage workflow (Fig. 3.1a). We focus on scaling up data volumes and solving the associated challenges. Specifically, we

address the need to access data selectively, rather than in bulk, to minimize the amount of sequencing required to recover the desired stored data.

For many years, high cost and low throughput have limited the applications of DNA data as a storage medium [13, 52]. Recently, various groups have observed that the biotechnology industry has made substantial progress and DNA data storage is nearing practical use [16–20, 53]. However, most prior DNA data storage efforts sequenced and decoded the entire amount of stored information, with no random access [16–19, 53]. However, this type of redundant sequencing becomes impractical as the amount of data increases (Fig. 3.1b,c). Being able to selectively access only part of the written information (e.g., retrieving only one image from a collection) is therefore necessary to make DNA data storage viable, but so far accessing part of stored information has only been demonstrated on a small scale [20, 21]. Our work demonstrates that PCR-based random access can be scaled up to reliably extract files of widely varying size and complexity from a DNA pool three orders of magnitude larger than those used in prior random access experiments.

Both DNA synthesis and sequencing are highly error-prone [54]. It is not unusual to observe aggregate insertion, deletion, and substitution rates at approximately 0.01 errors/base. Even complete loss of specific data strands can occur during library synthesis or amplification. Prior work has shown that it is possible to recover data even from such noisy conditions if proper encoding schemes are used. Although efforts have been made to minimize the amount of logical redundancy (i.e., the amount of additional information encoded) required for complete data recovery at a given error rate, existing approaches rely on a high degree of sequencing redundancy (i.e., having many copies of each sequence and deep sequencing coverage).

Here, we present a coding algorithm that explicitly reduces sequencing redundancy, hence requiring fewer sequencing resources and, in turn, fewer physical copies of any given molecule to fully recover the stored data. Our scheme tolerates aggressive settings of uneven low coverage and high coordinate error rates of insertions, deletions, and substitutions (Appendix A.1), while maintaining a logical density (bits per nucleotide) competitive with previously proposed

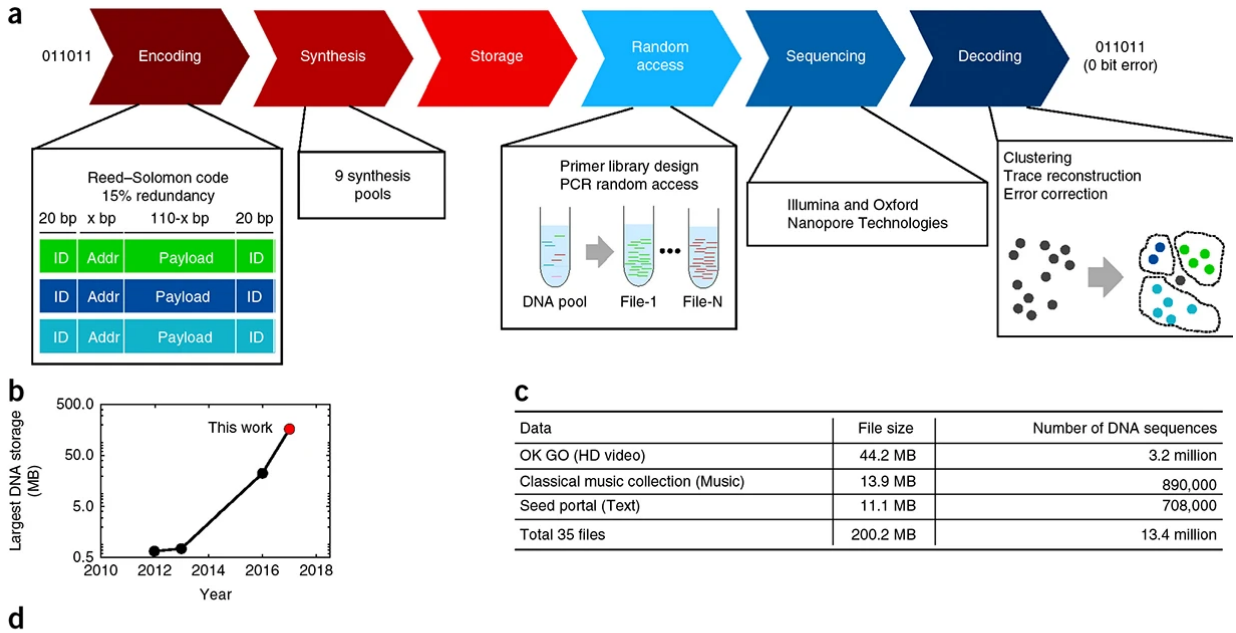


Figure 3.1: **(a)** The encoding process maps digital files into a large set of 150-nucleotide DNA sequences, including Reed–Solomon code redundancy to overcome errors in synthesis and sequencing. The resulting collection of sequences is synthesized by Twist Bioscience. The random access process starts with amplifying a subset of the sequences corresponding to one of the files using PCR. The amplified pools are sequenced using either sequencing by synthesis (Illumina NextSeq) or nanopore sequencing (Oxford Nanopore Technologies). Finally, sequencing reads are decoded using our clustering, consensus and error correction algorithms. **(b)** We encoded a total of 200 MB of data, about an order of magnitude more than prior work. **(c)** Example files encoded within these 200 MB of data. **(d)** A comparison to prior work shows that our coding scheme has similar logical redundancy, but requires lower sequencing coverage to recover files.

schemes (Fig. 1d, Appendix A.2). As DNA data storage technology matures, the goals of increasing throughput and lowering costs will likely drive coordinate error rates in the DNA data storage channel even higher than the current value.

To investigate challenges associated with increasing DNA data storage size, we created a large DNA library of modern data types, such as high-definition video, images, audio, and text. These included the "Universal Declaration of Human Rights" in over 100 languages [55], a high-definition music video of the band "OK Go" [56], and a CropTrust database of the seeds stored in the Svalbard Global Seed Vault [57].

3.3 Results

3.3.1 Coding method and random-access primer design

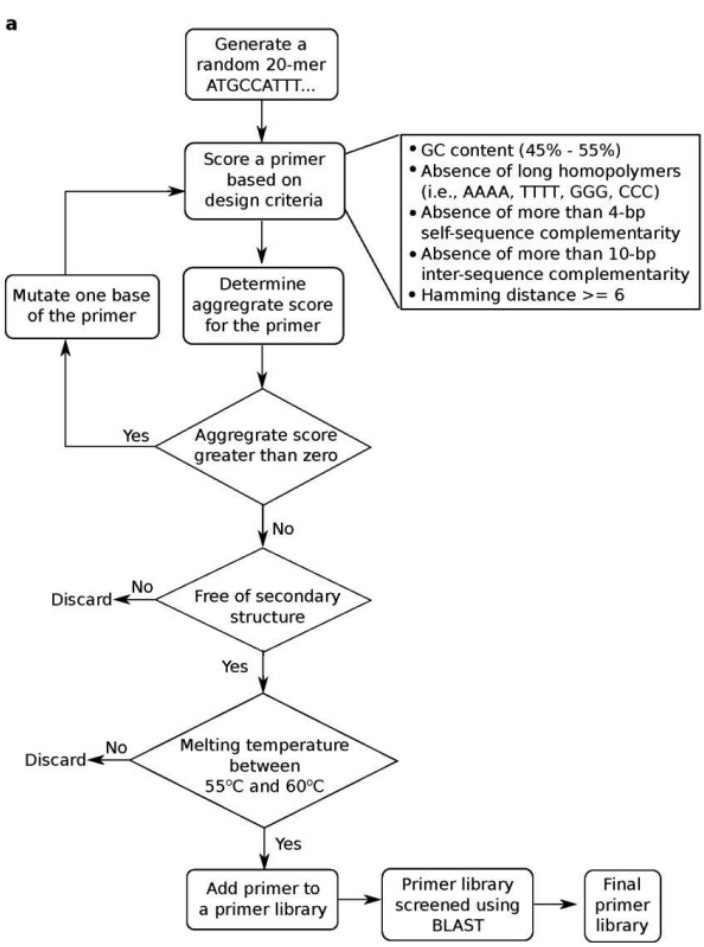
We encoded 35 files ranging from 29 KB to over 44 MB, totaling over 200 MB of unique (compressed) data (Fig. 1c lists a few examples; Supplementary Note 3 and Supplementary Table 3 found in the original work [25] provide the full list). We added 15% logical redundancy for robust error correction to 33 of our files and 25% to the other two, resulting in an additional 32.2 MB of data encoded in DNA. For DNA synthesis, we segmented each input file into a large number of oligonucleotides, each containing the same PCR primer target sequences that form a unique file ID. Moreover, each strand also includes a unique, strand-specific address to order strands within a file. The resulting synthetic DNA library contains 13,448,372 unique DNA sequences of lengths ranging from 150 to 154 bases, synthesized using Twist Bioscience's oligo pool services in a total of nine synthesis pools. Our resulting combined pool of about 2 billion bases represents an increase of about an order of magnitude in the amount of information stored in and retrieved from DNA, relative to prior work [53].

Achieving robust random access in a large DNA data storage system requires effective PCR primers to reliably amplify a specific file without crosstalk. We thus devised a framework for designing a primer library with thousands of pairs of orthogonal primers (file IDs). Our design method (Fig. 3.3a.i. shows a broad overview, Fig. 3.2 shows the specifics)

optimizes primers for several properties: avoidance of secondary structure formation and primer–dimer formation, absence of long stretches of homopolymers, melting temperature constrained to a narrow range (55–60°C), and a minimum of 30% of their sequence unique compared to other primers. To increase the stringency of sequence orthogonality, we used the basic sequence alignment program BLAST to screen out primers with long stretches of similar sequences [58].

Primer sequence design starts with a random 20-mer, and the sequence is "scored" based on several heuristic design criteria: absence of long homopolymer regions (maximum of 3 consecutive As or Ts, 2 consecutive Gs or Cs), absence of more than 4 bases of self-complementarity and absence of more than 10 bases of inter-sequence complementarity, 45%-55% GC content, and a minimum Hamming distance of 6 to other primers. If the sequence cannot satisfy a design criterion, all bases related to violating that criterion receive a +1 score. After evaluating the primer with all the design criteria, the score of every base is aggregated to form a final score of that primer. In short, the primer scores roughly correlate with the likelihood that the sequence will not act as an ideal primer. If the primer scores are greater than zero, the program continues to mutate the sequence until all design criteria are satisfied. After that, the primer sequence is filtered by secondary structure and melting temperature. This process is used to generate a large library of primers which are further screened against each other to avoid stretches with greater than 12 high similarity pairings (HSP) using the basic alignment program BLAST.

Before incorporating selected primer sequences into the library files, we tested primer performance on a pool of 3,240 synthetic "mini-files" ranging from 1 to 200 103-mers. We successfully accessed and sequenced up to 48 mini-files from this pool in a one-pot multiplex PCR experiment (Fig. 3.3a.ii), validating our primer library design approach (Appendix A.3). Here, we synthesized a pool of approximately 100,000 strands containing sets of size 1 to 200 DNA sequences each, surrounded by one of the 3,240 candidate primer pairs, and then randomly selected 48 of those pairs for amplification. The product was sequenced, and all 48 desired files were recovered, but with uneven coverage.



b Primer: AGGGGAGTA...C

	First position (A)	Second position (G)	Nth position (G)
GC content	0	1	1
Long homopolymers	0	1	1
Primer-dimer formation	0	0	1
Hamming distance	0	0	0

Figure 3.2: **(a)** Method for primer design. A random 20-mer continues to mutate until it satisfies the design criteria explained above. After satisfying these criteria, the primer is filtered by secondary structure and melting temperature. After generating a library of primers, the library is screened using BLAST to further improve sequence orthogonality. **(b)** Example of scoring for a primer. If the primer violates a design criterion, all bases related to the violation receive a +1 score.

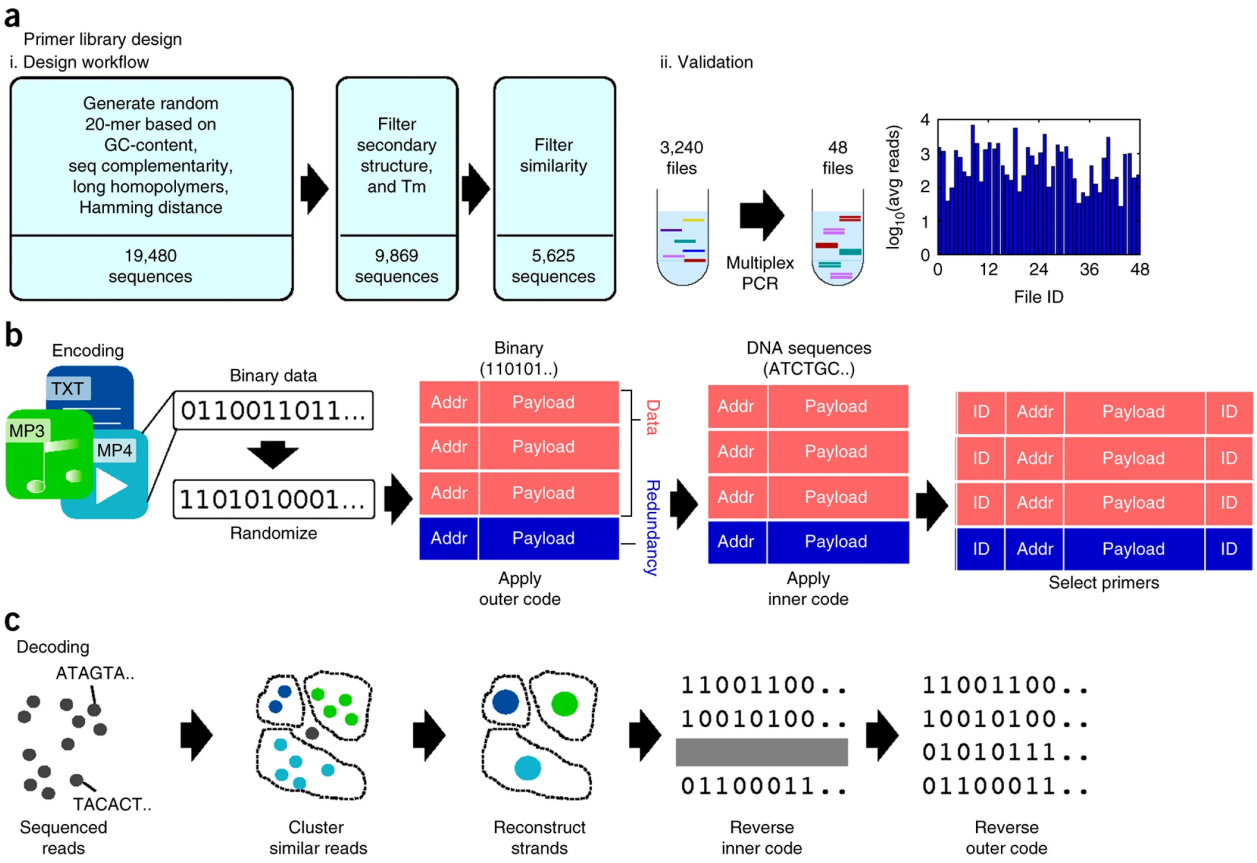


Figure 3.3: **(a, i)** We designed a primer library for our PCR-based random access method using an in silico process. **(a, ii)** The resulting set of candidate primers is then validated experimentally by accessing 48 arbitrary files. Each of the 48 primer pairs appear among sequencing reads, albeit at different relative proportions when normalized to the number of sequences in each set. **(b)** Our encoding process starts by randomizing data to reduce chances of secondary structures, primer–payload non-specific binding, and improved properties during decoding. It then breaks the data into fixed-size payloads, adds addressing information (Addr), and applies outer coding, which adds redundant sequences using a Reed–Solomon code to increase robustness to missing sequences and errors. The level of redundancy is determined by expected errors in sequencing and synthesis, as well as DNA degradation. Next, it applies inner coding, which ultimately converts the bits to DNA sequences. The resulting set of sequences is surrounded by a primer pair chosen from the library based on (low) level of overlap with payloads. **(c)** The decoding process starts by clustering reads based on similarity, and finding a consensus between the sequences in each cluster to reconstruct the original sequences, which are then decoded back to digital data.

Next, we created a coding scheme to convert digital information to DNA sequences and back to digital information. Similar to prior work [18, 53], our approach employs concatenated codes with Reed–Solomon (RS) as the outer code (Fig. 3.3b). (However, unlike most earlier work, we used very long codes (length up to 65,536) to handle large variations in the number of errors between code words.) Input data are then randomized by XOR with a pseudo-random sequence. Randomization facilitates coping with errors by breaking multi-bit repeats (e.g., 00000000) and ensures that the DNA sequences we produce are dissimilar, which makes decoding less computationally costly.

The encoder first partitions the randomized digital file into multiple blocks, up to a megabyte in size. We represent each block by a matrix M with up to ten rows and up to 55,000 columns, where every matrix cell carries a 16-bit value. Next, we encode each row of M with a Reed–Solomon code to obtain a larger matrix M′ that extends M by appending redundant columns. Every column of M′ is later converted into a DNA sequence of length 110 (114 for File 33; details about each file can be found in the published work's Supplementary Note 3 and Supplementary Table 3 [25]). When Reed–Solomon redundancy is set to 15%, 87% of the DNA sequences carry raw input data (systematic RS coordinates), while 13% carry redundant data used for error correction (redundant RS coordinates).

The conversion of columns of M′ to DNA sequences involves representing a column in base 4, appending a prefix with address information (block index and column index), breaking the column into consecutive fragments of size three each, treating the content of each fragment as a number between 0 and 63 written in base four, representing this number in base three to obtain a fragment of size four, putting the new fragments together, and applying a rotating code [17] to turn a base-three representation into a base-four representation that eliminates homopolymers.

Finally, all DNA sequences are appended with 20-base PCR primer targets selected from the primer library on both ends to allow random access to the file. Resulting DNA sequences are synthesized into DNA strands, which can then be preserved using a variety of methods, and later selected via random access.

To avoid collisions between primers and payloads, we perform a "collision check" by aligning the randomized data against the primers with BLAST (see Fig. 3.4). Payloads have a "collision" with primers if BLAST finds pairing segments longer than 12 base pairs. The process repeats to reduce the number of collisions. We found it challenging to recover files from complex pools without proper primer design and collision avoidance. We compared file amplification with and without proper primer design and collision avoidance in both simple and complex pools. for pools containing a single file, we successfully amplified the file with or without the design and collision avoidance. However, using a complex pool where the file of interest was 18.0% of its pool, we were unable to amplify the same file without the design and collision avoidance methods. In contrast, the use of these methods allows robust amplification of a file that was 17.4% of its complex pool. A gradient PCR was performed to ensure the annealing temperature did not impact the comparison, and qPCR was also performed to determine the number of subsequent PCR cycles.

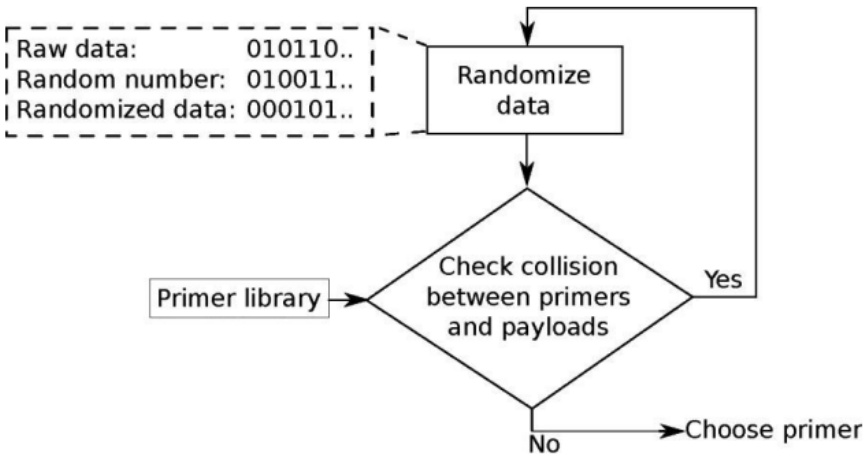


Figure 3.4: Digital data are iteratively randomized to reduce collisions between primers and payloads.

Our proposed random-access approach and associated primer design and conflict detection methodology scales to physically isolated pools of several terabytes each (Appendix A.3). In dehydrated spots, these would measure on the order of one millimeter, which in turn can

be organized in dense arrays. Such a system would be orders of magnitude denser than tape.

When servicing a read request, we retrieve and rehydrate the DNA. Sequencing these DNA strands produces a collection of noisy reads, which do not necessarily include all original DNA sequences; sequences may be lost by sampling, storing, retrieving and preparing the DNA for sequencing. Sequences belonging to a specific file are obtained by aligning and filtering based on the primer sequence and length. Frugality with respect to coverage was a key consideration when designing our decoding approach. Therefore, we do not require reads to be the correct length5, and we do not filter out reads with errors in their primer region [19]. Instead, noisy reads whose length is within 20 nucleotides of the original length are selected and passed to the decoder.

The decoder operates in four stages (Fig. 3.3c). First, it clusters noisy reads by similarity, based on their entire content, not just the addresses [20], to collect all available reads that likely correspond to one of the unique DNA sequences originally stored. To do so, we employ an algorithm that leverages the input randomization done during encoding. At a high level, we initially consider each noisy read a separate cluster and iteratively merge clusters based on random representatives, leveraging the fact that noisy reads of any specific DNA sequence are similar and noisy reads of different DNA sequences are dissimilar. Our algorithm runs in time that is close to linear in the input size and utilizes a series of filters to avoid unnecessary and slow edit distance computations. Using a locality-sensitive hashing scheme for edit distance, we compare only a small subset of representatives. We also use a lightweight check based on a binary embedding to further filter pairs. If a pair of representatives passes these two tests, edit distance determines whether the clusters are merged. A less computationally efficient, but functionally equivalent alternative, approach to clustering that uses off-the-shelf software is discussed in Appendix A.4.

The second stage of the decoder then processes each cluster to recover the original sequence. This stage, which we call trace reconstruction, uses a variant of the Bitwise Majority Alignment algorithm (BMA) [59], adapted to support insertions, deletions, and substitutions. The algorithm follows BMA in that pointers for noisy reads are maintained and moved from

left to right, and at every step of the process the next symbol of the original sequence is estimated via a plurality vote. For the noisy reads that agree with plurality, the pointer is moved to the right by 1 (hypothesizing that the read had the correct symbol at the respective position), just like in BMA. But for the samples that do not agree with plurality, the algorithm tries to decide what the reason for the disagreement is: is it due to a deletion, an insertion, or a substitution? The classification of mismatches is done by looking at the context around the symbol under consideration in the noisy read. Once this is estimated, the pointers are then moved to the right accordingly.

In the third stage, the decoder unwinds the no-homopolymer representation to obtain matrices M corresponding to different blocks. In each recovered matrix some columns may be missing (erasures), and others may contain errors. In stage four, we decode the outer Reed–Solomon (RS) code to correct errors and erasures in rows of matrices M′ and invert randomization. Successful decoding is possible if for each row of each matrix M′ the 'used error resilience' ratio

$$\frac{2 * (\text{ number of errors }) + (\text{ number of erasures })}{\text{number of redundant RS coordinates}} \tag{3.1}$$

is at most 1.

3.3.2 Error analysis and decoding from Illumina sequencing

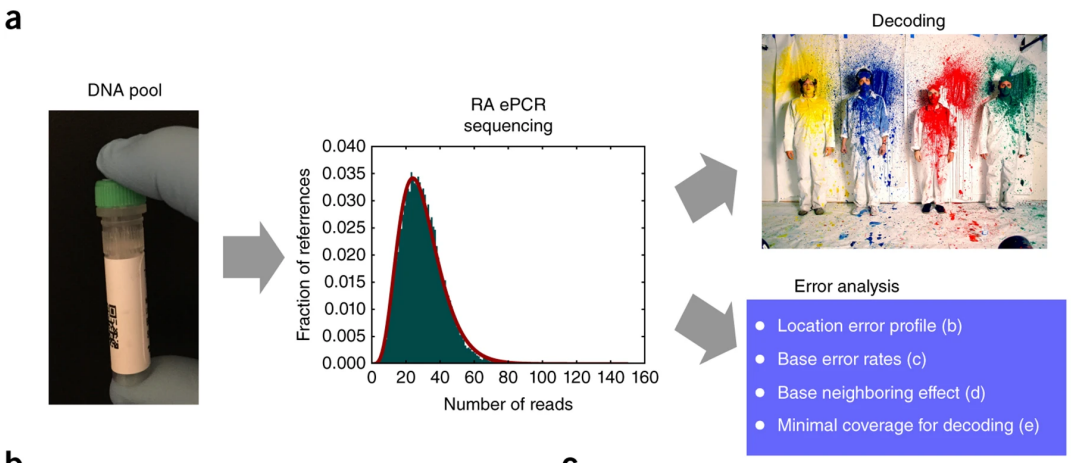
We received nine synthesized DNA pools periodically over several months. In each case, we immediately individually amplified every file in the pool using random-access emulsion PCR, attached Illumina sequencing primers and adapters, and then sequenced the files for a total of ten sequencing runs. We have aggregated about 723 million reads of more than 13 million distinct synthetic DNA sequences. The mean coverage (i.e., number of reads for a given DNA sequence) across the data set was 53.8 reads with a s.d. of 48.7 reads. We observed considerable variance across files, ranging from a mean of 6.7 reads with s.d. of 3.4 reads, to a mean of 298.6 reads with s.d. of 139.6 reads. For most files, the empirical coverage distribution was reasonably well-approximated by a gamma distribution with matching mean

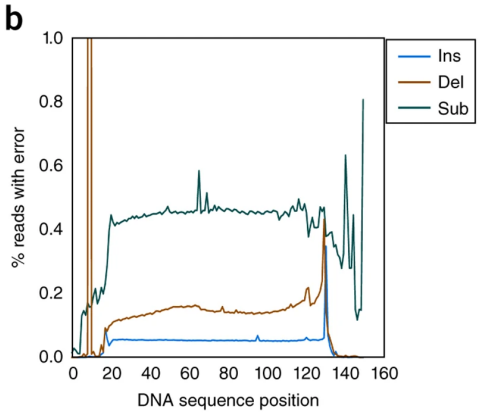
and variance (Fig. 3.5a, center).

The sequencing information serves two purposes: (1) error analysis of processes related to DNA manipulation, including synthesis, random access, and sequencing, when used in conjunction with knowledge of the encoded DNA sequences; (2) and decoding of data stored in DNA and analysis of code resilience, that is, its ability to recover the information under the observed error regime. The decoding process uses only information that would be available at read time in a real storage scenario, that is, no knowledge of the encoded DNA sequences, other than the information received from sequencing, is used.

The error analysis in Figure 3.5b (Appendix A.5) reveals an average error rate per position of 0.6%. Substitutions were the most prominent type (0.4%), twice as likely as deletions (0.2%) and ten times as common as insertions (0.04%). In some files, specific positions showed higher error rates owing to systematic errors in the reading or writing processes. Also, primer target regions (first and last 20 positions) suffered from fewer errors because of the nature of PCR, which favors amplification of perfect primers and primer target regions. A clear exception is the spike at position 9, which was caused by a single primer sequence with an error at that position. However, in all cases, error rates in the primer region were low enough to associate most reads coming from sequencing to the sequenced files via sequence alignment. Analysis of the non-primer region is shown in Figure 3.5c,d. Figure 3.5c shows the percent breakdown of errors per base type, where almost half of the insertions are associated with type G and about a third of the substitutions are associated with type T. Deletions are evenly distributed. This highlights the fact that insertion and substitution errors are biased toward certain base types. Figure 3.5d also shows variation of error rates across differing neighboring base types. Again, types G and T are associated with higher error rates.

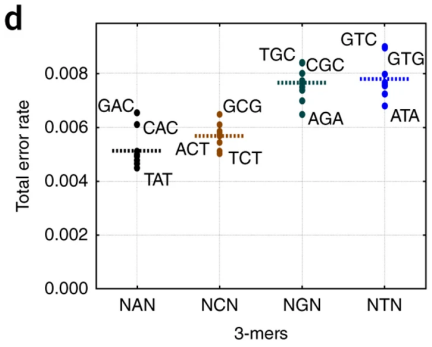
Next, we proceeded to decode the data. In practice, current preparation and sequencing technologies yield some unusable reads owing to their length being outside the acceptable range or too low-confidence in base calling (6.5% on average in our experiments). We expect this number to improve as sequencing technology and wet lab protocols mature. We randomly subsampled the usable sequencing reads for each individual file, gradually increasing the

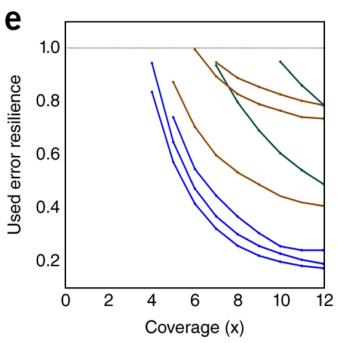




c

	Insertion rate	Deletion rate	Substitution rate	% Total reads
A	1.1×10^{-4}	4.1×10^{-4}	7.5×10^{-4}	24.6
C	9.1×10^{-5}	3.6×10^{-4}	9.8×10^{-4}	25.1
G	2.5×10^{-4}	3.8×10^{-4}	1.3×10^{-3}	25.1
T	8.4×10^{-5}	3.7×10^{-4}	1.5×10^{-3}	25.2
Total	5.4×10^{-4}	1.5×10^{-3}	4.5×10^{-3}	100.0





	Ins	Del	Sub
Run 2	3.4×10^{-4}	9.4×10^{-4}	3.7×10^{-3}
Run 3	3.8×10^{-4}	1.2×10^{-3}	3.2×10^{-3}
Run 4	2.8×10^{-4}	1.2×10^{-3}	4.5×10^{-3}

Figure 3.5: **(a)** A file of interest is randomly accessed via ePCR amplification from a stored pool and sequenced. A file-specific error profile is generated with the recovered file stored in DNA. **(b)** Per-position average read error profile averaged over the first 150 positions of all 13 million sequences' corresponding reads. **(c)** Error rates and number of reads for different nucleotide types in the payload region. **(d)** Error rates depend not only on nucleotide type but also on the type of neighboring nucleotides. Each dot corresponds to a 3-mer in the payload region colored according to the central base: black for A, red for C, green for G, and blue for T. The horizontal bars represent the weighted mean of the dots in that particular column, as not all 3-mer appear the same number of times. **(e)** Estimating the minimal coverage required for decoding. Each curve corresponds to a different file, each color corresponds to a different sequencing run, and numbers in the legend correspond to the average insertion, deletion, and substitution errors for the corresponding sequencing run.

number of reads supplied to the decoder. We were able to recover all 200 MB of data (zero-byte difference when compared to the original digital data) stored in the DNA with median coverage of only 5 reads per DNA sequence, with different files ranging from 4 to 14 reads per DNA sequence. If we include unusable reads in the calculation, the median goes up to 6.2 reads per DNA sequence. This is half as much as the minimum coverage ever reported in decoding digital data from DNA (Fig. 3.1d). The impact is lower cost because decoding from lower coverages allows for a larger number of different DNA sequences to be read with the same sequencing kit. To understand the effect of coverage on our ability to decode files with no bit errors, we supplied the decoder with increasing coverage of reads, and measured the 'used error resilience' for several of our files (Fig. 3.5e). As expected, the ratio decreased with coverage because the total number of errors and erasures decreases with extra read information. Redundant information is more scarce at lower coverages, resulting in higher 'used error resilience'.

3.3.3 DNA assembly for nanopore sequencing

To further stress-test our decoding algorithm, we sequenced two files (32 KB and 1.3 KB) using the Oxford Nanopore Technologies (ONT) MinION sequencer (Fig. 3.6). The compactness and potential for scalability makes nanopore-based sequencing an intriguing option for integration in future stand-alone DNA data storage systems. A key advantage is a very long read length of potentially thousands of nucleotides; however, with current technology, only a limited number of reads can be obtained from a single sequencing flow cell. To best utilize these characteristics, we developed a protocol to concatenate multiple oligonucleotides into longer reads (Fig. 3.6a,b). Using this approach, we successfully recovered a 32-KB file sequenced with nanopore technology at a coverage of 36× and a 1.3-KB file at a coverage of 80× despite a high coordinate error rate of approximately 12%, computed using exhaustive minimum edit distance. We observed that reads of incorrect length constituted over 88% of all reads, and ignoring those reads makes recovering the files impossible even at maximal available coverages of 74× and 147×, respectively, indicating the importance of using as

many reads as possible (Fig. 3.6d), a unique feature of our proposed decoder. Results above bode well for building an integrated, scalable DNA data storage system that is tolerant of the high error rates that could accumulate over millennia.

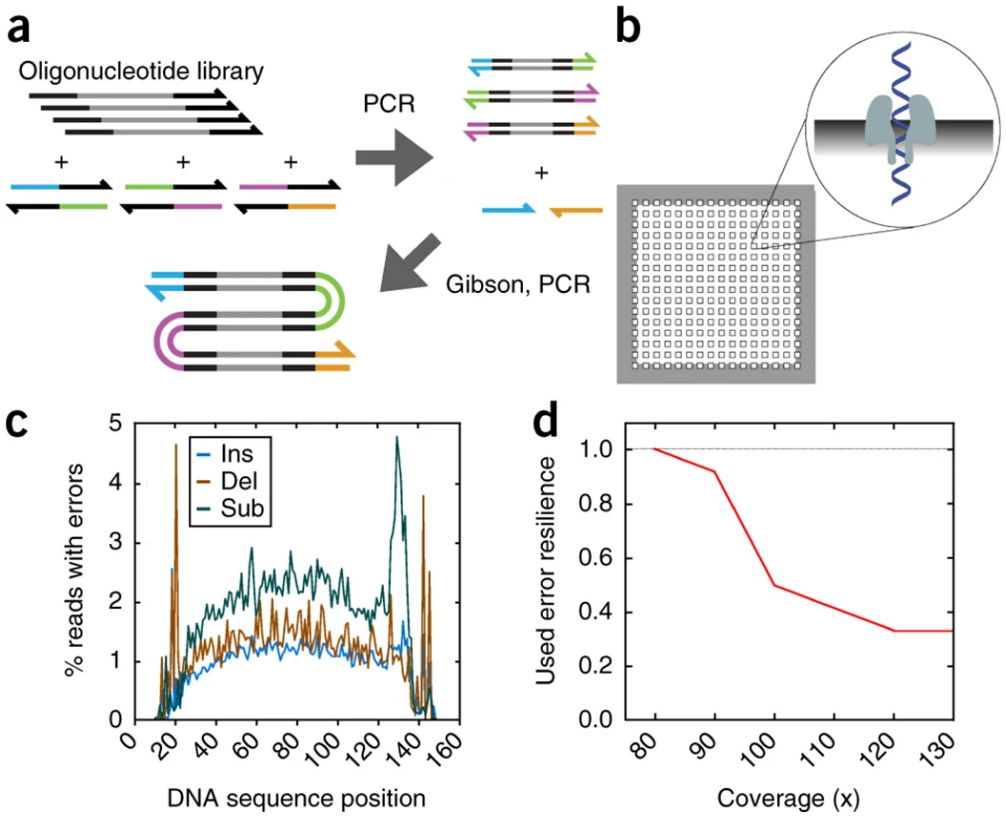


Figure 3.6: **(a)** A file of interest is amplified using PCR with primers containing complementary overhang sequences. Subsequently, amplification products are mixed in a Gibson assembly reaction and amplified using primers corresponding to the unique overhangs present at the 5' and 3' of the Gibson assembly product. **(b)** Amplicon consisting of concatenated oligonucleotides is sequenced using the MinION and thousands of reads are generated. **(c)** Per-position average read error profile averaged over all 88 strands of a 1.3-KB file and their corresponding 2D reads. Error rates are higher than in Illumina-sequenced reads. **(d)** Estimating the minimal coverage required for decoding. Higher error rates are offset by higher coverage, making decoding the original stored data possible.

3.4 Methods

3.4.1 Workflow of DNA pool preparation and sequencing

The general workflow from receiving a synthesized pool to sequencing a file from the pool is divided into three steps (Fig. 3.7): (1) ePCR amplification that selects only the desired file while simultaneously adding a random 25-nucleotide region, (2) ligation of Illumina sequencing adapters to the amplified DNA, and finally (3) sequencing the product of ligation in an Illumina NextSeq instrument.

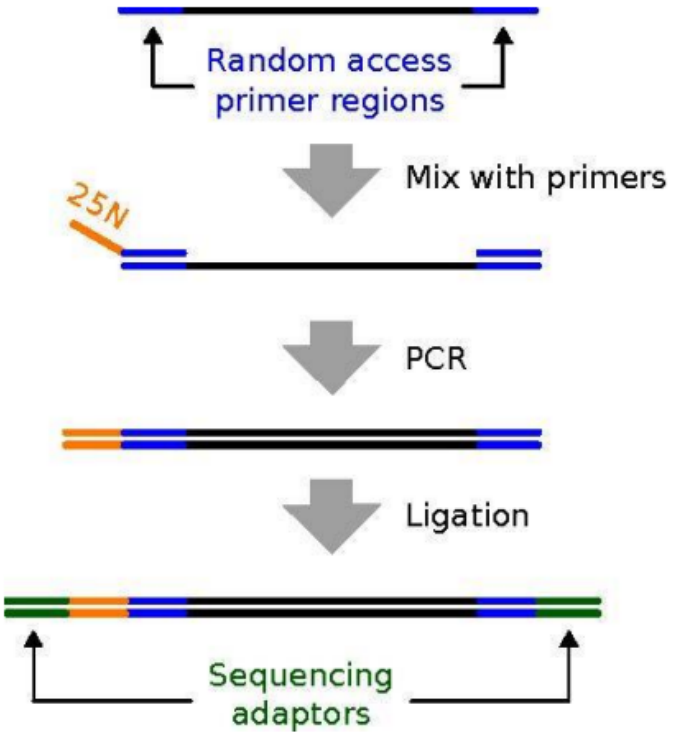


Figure 3.7: First, random access regions are used to select files for sequencing. Through ePCR, a 25N region is added to the oligos to improve nucleotide diversity. Then, samples are ligated to Illumina sequencing adaptors with modified Illumina TruSeq Nano kit protocol. Finally, prepared samples are sequenced on an Illumina NextSeq instrument with a 10%-20% PhiX spike-in.

We refer to the unique primer regions in these DNA strands as "random access regions". Due to the conserved random access regions across DNA strands in each sequencing run, we found it necessary to introduce a random 25-nucleotide overhang (25N region) on one of the random access primers. This is because the NextSeq requires highly diverse sequences for the first 4-7 cycles to determine cluster positions, and for the first 25 cycles to determine the percent of clusters passing chastity filter. Roughly 30-80% of a sequencing sample must be highly diverse for the first 25 bases for good quality results, which was often not the case for our samples. Adding the 25N overhang allowed us to ensure the first 25 nucleotides were in adequately equal proportions of A, C, G, and T across all strands of DNA so samples could be sequenced efficiently.

After ePCR, samples were prepared for sequencing with the Illumina TruSeq Nano kit with the modified protocol described in Methods. After quantification, samples were mixed proportionally, diluted to 1.3pM and loaded into the NextSeq, including an additional 10% to 20% of Illumina's PhiX control genome spiked in. Interestingly, only 2 to 11% of all sequencing data aligned to PhiX. This is most likely due to the significantly slower clustering of most PhiX fragments, which are longer than our sequencingready strands.

3.4.2 Selectively amplifying DNA

Whenever we received a pool of synthetic DNA, we rehydrated the pool in 1× TE buffer and used the following protocol to amplify each file individually (see Fig. 3.7).

Mix 10 ng of ssDNA pool (1 uL) with 1 μL of 100 uM of the forward primer with a 25 nucleotide random overhang and 1 μL of 100 uM of the reverse primer (with no overhang), 25 μL of 2× Kapa HiFi enzyme mix, 20 μL of molecular grade water, and 2 μL of 1.25 mg/mL acetylated bovine serum albumin. All primers were ordered from Integrated DNA Technologies. This 50 μL mix was then mixed with the 300 μL oil surfactant mixture detailed in the EURx ePCR kit. The resulting mixture was then attached to a benchtop vortexer in a refrigerator and vortexed for 5 min at the highest setting.

After vortexing, the now milky-appearing product was split evenly into three PCR tubes

and placed in a thermocycler with the following protocol: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 62°C for 20 s, (4) 72°C for 15 s, (5) go to step 2 a varying number of times depending on the proportion of the pool being amplified, (6) 72°C for 30 s. The reaction was then purified according to the instructions in the EURx ePCR kit. Total yield typically ranged between 30 ng and 1 ug because ePCR yield is directly proportional to the size of the file. The reverse micelles that make up the emulsion should all theoretically have the same amount of primer and one strand of DNA, so the larger the file, the greater the proportion of micelles have targeted strands. Recall that regardless of file size, 10 ng of the pool was used (see Supplementary Table 3 of the published work [25] for the percent of the amplified pool each file comprised). This resulted in approximately 80k copies of each strand present at the start of each ePCR reaction.

When necessary, qPCR was performed to determine the ideal number of cycles to amplify a file according to the following recipe: mix 1 ng of ssDNA pool (1 uL) with 0.5 μL of 10 uM of the forward primer (with no overhang) and 0.5 μL of 10 uM of the reverse primer (with no overhang), 10 μL of 2$\times$ Kapa HiFi enzyme mix, 7 μL of molecular grade water, and 1 μL of 20x Eva Green. The thermocycling protocol was: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 62°C for 20 s, (4) 72°C for 15 s, then repeat steps 2–4 as needed.

After amplification with ePCR, the length of the dsDNA products was confirmed with a Qiaxcel fragment analyzer, with sample concentration measured by Qubit 3.0 fluorometer.

3.4.3 Ligation of amplified DNA files for sequencing

After ePCR, amplified products were ligated to the Illumina sequencing adapters with a modified version of Illumina TruSeq Nano ligation protocol and TruSeq ChIP Sample Preparation protocol. Briefly, samples were first converted to blunt ends with the ERP2 reagent and directions provided in the Illumina TruSeq Nano kit, then purified with AMPure XP beads according to the TruSeq ChIP protocol. An 'A' nucleotide was added to the 3' ends of the blunt DNA fragments with the TruSeq Nano's A-tailing ligase and protocol, followed by ligation to the Illumina sequencing adapters with the TruSeq Nano reagents and proto-

col. We then cleaned the samples with Illumina sample purification beads and enriched the sample using PCR to yield enough product for sequencing. The length of enriched products was confirmed using a Qiaxcel bioanalyzer.

3.4.4 Sample preparation for sequencing

When multiple separate samples were prepared for sequencing, these samples were mixed proportionally (e.g., a 10,000 oligonucleotide file to be sequenced with a 500,000 file would comprise 1.96% of the DNA material in this mix). The mixed sample was then prepared for sequencing by following the NextSeq System Denature and Dilute Libraries Guide. The sequencing sample was loaded into the sequencer at 1.3 pM, with a 10 to >20% PhiX spike-in as a control (PhiX is a reliable, adaptor-ligated, well-characterized genomic DNA sample provided by Illumina).

3.4.5 Sequencing with Oxford Nanopore Technologies' MinION

First, we used PCR to amplify an oligonucleotide library with primers containing orthogonal overhang sequences. Then, we combined the amplified products into one Gibson assembly reaction where each overhang allowed for multiple library members to be concatenated. Finally, we used PCR to amplify the resulting concatenated product with primers that hybridize to each respective end of the assembly product. Using this approach, we generated a 3-fragment and a 6-fragment assembly for a 1.3-KB and a 32-KB file, respectively.

To amplify the original file and add the overhangs, a 100-fold diluted sample of ssDNA library was amplified using a KAPA SYBR FAST qPCR kit with the following thermal profile: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 69°C for 20 s, (4) 72°C for 20 s. The total number of cycles of steps 2–4 was determined by monitoring the fluorescence of the qPCR instrument as the amplification reached the plateau phase. Each amplification reaction was performed separately with primers containing distinct overhang regions necessary for a subsequent Gibson assembly reaction. Overhang sequences were designed using the NUPACK15 design module to avoid secondary structure formation. After amplification, each reaction

was purified using Agencourt AMPure XP. Subsequently, amplification products mixed at equal molar ratio were added to NEB Gibson assembly master mix (1:1 volume ratio) and incubated at 50°C for 30 min.

Upon AMPure XP clean-up, the ligated product was amplified using the same qPCR protocol described above. Amplification was performed with primers corresponding to unique overhang sequences present at the 5' and 3' ends of the DNA. After amplification, a DNA band corresponding to the expected size was gel-extracted from a 2% agarose gel and quantified by a Qubit 3.0 fluorometer. The final product had the expected size corresponding to the number of fragments and overhangs used in the assembly.

Sequencing sample preparation of the 1.3-KB file was performed according to the Oxford Nanopore Technologies (ONT) Amplicon (R9) protocol for 2D sequencing. Metrichor sequencing metrics indicated 37,478 reads with workflow successful exit status out of 130,573 total reads. Sequencing sample preparation of the 32-KB file was performed according to the Oxford Nanopore Technologies (ONT) Amplicon (R9.4) protocol for 1D2 sequencing. ONT Albacore basecalling software yielded 57,012 1D2 reads. In both cases, these reads were then successfully decoded into the original digital file.

3.5 Discussion

Given the current trends in data production and the rapid pace of progress of DNA manipulation technologies, DNA data storage has the potential to complement or eventually replace tape, the densest commercially available storage medium for archival storage.

The global demand for synthetic DNA in 2015 was estimated to be 4.8 billion bases of single-stranded oligos and approximately 1 billion bases of longer double-stranded oligos, or just under 6 Gigabases in total [11]. To provide a sense of scale, the size of the largest known eukaryotic genome is about 149 Gigabases [60]. The first practical 'DNA drive' should have a throughput of at least a few kilobytes per second. At the coding density demonstrated here, this is a few kilobases per second, or the equivalent of the entire synthetic DNA industry annual production in just 2 weeks. Clearly, synthetic DNA production will have to increase

to meet this goal. We contend this is attainable because the synthetic DNA needed for data storage can be significantly more error prone than DNA used by life sciences, and very few copies per sequence are required. This is due to error-correcting algorithms such as the one described in this paper.

Even at kilobyte-per-second throughput, a DNA drive can be interesting because of the long-term durability and relevance DNA can offer to the preservation of high value-per-bit data. However, large-scale, deployed storage technologies today offer throughputs of hundreds of megabits per second, which will be more challenging to match. At this point, even DNA sequencing technologies, which are currently capable of reading megabases per second, will require improvement. The cost per bit offered by current storage devices is also much lower than what is possible with DNA today. Luckily, both DNA synthesis and sequencing technologies use array-based designs, which are readily replicable and amenable to scaling. This scaling increases the number of DNA sequences that can be read or written at a time, simultaneously increasing throughput and decreasing costs. Additional cost reductions can be obtained by optimizing fluidic delivery and exploiting large-scale efficiencies in the chemistry of reagents. We expect to see substantial activity in these areas in the upcoming years.

This chapter describes large-scale random access, low redundancy, and robust encoding and decoding of information stored in DNA, as well as a notable increase in the volume of data stored (200 MB). To encourage more work in this area, we have made samples available to groups interested in DNA data storage.

Chapter 4

PROBING THE PHYSICAL LIMITS OF RELIABLE DNA DATA RETRIEVAL

A version of this work was originally presented in Nature Communications, Organick et al. 2020 [7]. In addition to myself, the co-authors for this work are Yuan-Jyue Chen, Siena Dumas Ang, Randolph Lopez, Xiaomeng Liu, Karin Strauss, and Luis Ceze.

4.1 Chapter Abstract

Here, we demonstrate reliable file recovery with PCR-based random access when as few as 10 copies per sequence are stored, on average. This results in density of about 17 exabytes/gram, nearly two orders of magnitude greater than prior work has shown. We successfully retrieve the same data in a complex pool of over 10^{10} unique sequences per microliter with no evidence that we have begun to approach complexity limits. Finally, we also investigate the effects of file size and sequencing coverage on successful file retrieval and look for systematic DNA strand drop out. These findings substantiate the robustness and high data density of the process examined here.

4.2 Overview

For synthetic DNA data storage to become a viable alternative to electronic archiving, many unique DNA sequences must be physically storable in a single pool and then randomly and reliably accessed. Random access requires far fewer resources to recover data since only relevant files are sequenced and analyzed. Theoretically, for maximum density, only one copy of each sequence would be necessary to perform the Polymerase Chain Reaction (PCR) random access reaction. In practice, however, this is not the case for two reasons: stochastic

variations in copy numbers that arise from sub-sampling the pool during random access, and copy number variations that arise from synthesis. Knowing the minimum copy number for each PCR reaction is crucial for storing DNA data; without it, one might store too few copies to access the data or too many, wasting orders of magnitude of density.

Previous work in this space recognized the importance of storage density for DNA to become a practical archival storage [16, 17, 19, 25], but the greatest complexity surrounding random access in those works reached just over 10^7 unique sequences [25], and was presented in Chapter 3. In more recent work, a different method of random access using bead extraction of desired strands prior to PCR random access utilized 10^{18} unique sequences [61]. However, in addition to different random access techniques and strand architecture, those methods differ substantially from the methods presented here and make it difficult to compare to this work. Notably, while in this paper random strands of DNA are encoded by 150Nmers where each base is randomly attached during synthesis, their methods employ mutagenetic PCR on template strands which introduces an approximate 5% error rate to the final PCR product [62] while maintaining a conserved PCR primer region. Both factors likely minimize interactions between the desired strands and the strands added for complexity. Additionally, the concentration is substantially changed in those methods depending on the complexity examined, while this work only examines random access at approximately 1ng/μL concentration. Nevertheless, both works are consistent in concluding that the limit to random access complexity is not 10^{10} sequences per random access reaction.

It is also important to note that PCR random access is not unique to synthetic biology, as biological research often involves using PCR in complex conditions to amplify only desired parts of a genome or gene expression data. Though the commonly analyzed human and mouse genome are both nearly 3 gigabases [63], approximately 2 orders of magnitude fewer bases than examined in this work's most complex condition, genomic libraries are often exceedingly complex because of the genome fragmentation process. While genomic analysis is also not immune to missing expected sequences (also known as dropout) and the field has developed techniques to cope with this problem, the mechanisms behind the absent

sequences and the steps to mitigate it are different from more synthetic applications such as this work [64–67]. This further motivates the work presented here.

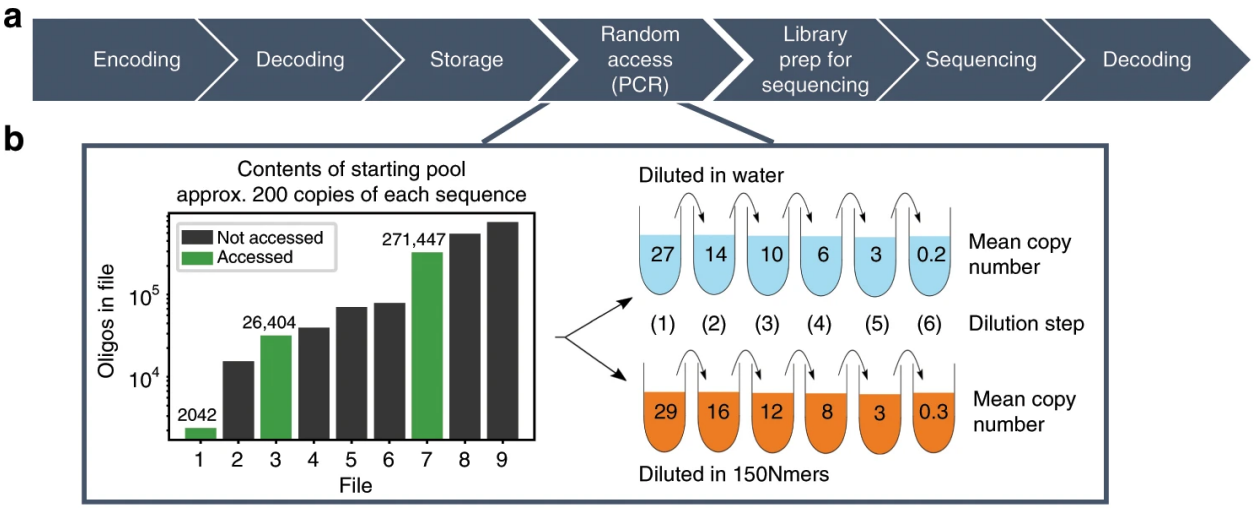
This chapter examines the ability of PCR to recover three files, each an order of magnitude larger than the other, from truly random pools ranging from over 10^6 to over 10^{10} unique sequences per microliter (Fig. 4.1). An average of 10 copies of each sequence is necessary for successful data retrieval for the three files, regardless of how many sequences per microliter are used. In addition, we also examine individual sequence behavior and find no systematic sequence loss, thus showing sequence loss is stochastic and not due to sequence design. We also look at the effects of increasing sequencing coverage to ten times greater than the coverage used in this work's analysis and find that it is minimally effective at retrieving more sequences. This work further supports the robustness and high density storage potential of DNA, for we demonstrate we have not yet reached the limit of permissible pool complexity, and with a minimum copy number of 10 we show this process yields the densest DNA storage system to date at 17 exabytes/gram (EB g^{-1}).

4.3 Results

4.3.1 Reducing Copy Number and Increasing Pool Complexity

In this work, we randomly accessed three files from a large pool of DNA at varying copy numbers (Fig. 4.1b). The small file, comprised of 2,042 sequences, represents approximately 0.1KB of digital data. The medium file consists of 26,404 sequences and 1.7KB of digital data, and the large file has 271,447 sequences and 18KB.

We then sequenced all three files at all stages of dilution to measure the rate of sequences lost (Fig. 4.2a, 4.2b). A copy number of 10, for example, means that, on average, each sequence was present 10 times in solution. The original, undiluted pool of DNA encoded nine files (1.6M total oligonucleotides (oligos)) and was subsequently serially diluted with water to result in copy numbers ranging three orders of magnitude (Fig. 4.1b). To dramatically increase pool complexity, we repeated this process, this time diluting the samples with 1ng

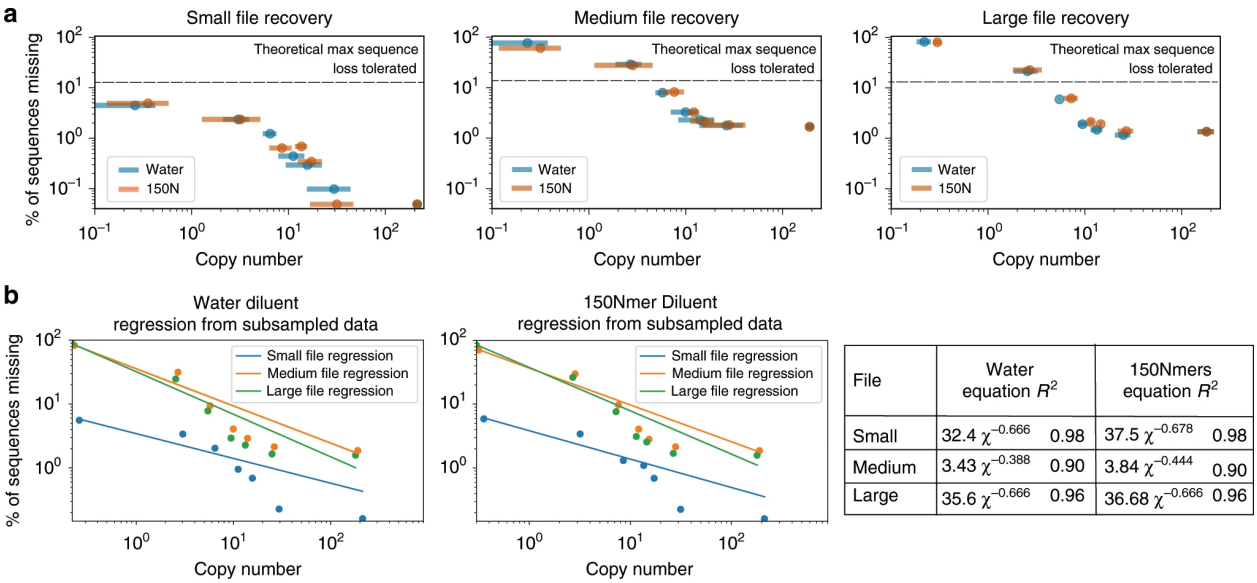


c

Dilution step	Initial pool	1	2	3	4	5	6
µL Prior sample used	n/a	11.0	57.9	63.8	40.0	19.0	5.0
µL Diluent used	n/a	99.0	52.1	21.3	20.0	31.0	45.0
Water diluent: unique sequences/µL	1.60E+6	1.60E+6	1.60E+6	1.60E+6	1.60E+6	1.60E+6	3.20E+5
150N Diluent: unique sequences/µL	1.60E+6	9.87E+9	1.04E+10	1.06E+10	1.07E+10	1.08E+10	1.09E+10
Smallest percent of pool accessed in a single PCR reaction	0.13	2.07E–5	1.96E–5	1.93E–5	1.91E–5	1.88E–5	1.87E–5

Figure 4.1: **(a)** A high-level representation of the DNA data storage pipeline. **(b)** (Left) The bar chart depicts contents of the initial, undiluted pool. (Right) The illustration shows the serial nature of subsequent dilutions. Mean copy number refers to the mean number of copies of each file's unique sequences as determined by qPCR (Appendix B.1). One serial dilution used water as the diluent in each step; the other used a solution of 150Nmers to dilute the pool to much greater complexity. **(c)** Details of how the samples were diluted. Note that the dilution steps were identical regardless of diluent. The smallest percent of pool accessed is calculated by dividing the size of the smallest file by the number of unique sequences in the 1 uL of solution used for PCR random access. This percentage refers to the 150Nmer diluent pool since the small file in the water diluent pool is a constant 0.13%.

μL^{-1} 150Nmers, random sequences of DNA the same length as our original pool (Fig. 4.1c). Theoretically, we would expect to encounter a duplicate sequence only after 4^{150} $(2 * 10^{90})$ sequences had been synthesized; thus, we would expect each strand to be unique. By diluting samples with 1ng μL^{-1} 150Nmers, pool complexity increased from $1.6{*}10^6$ unique sequences in the initial pool (Fig. 4.1b) to over 10^{10} unique sequences per microliter (see Appendix B.2).



File	Water equation R^2		150Nmers equation R^2	
Small	$32.4\,\chi^{-0.666}$	0.98	$37.5\,\chi^{-0.678}$	0.98
Medium	$3.43\,\chi^{-0.388}$	0.90	$3.84\,\chi^{-0.444}$	0.90
Large	$35.6\,\chi^{-0.666}$	0.96	$36.68\,\chi^{-0.666}$	0.96

Figure 4.2: **a**) Each plot illustrates a file's loss of sequences recovered at 20x coverage, directly comparing the samples diluted in water to those diluted in 150Nmers (150nt sequences comprised of random nucleotides). The threshold of the maximum number of sequences that can be lost while still permitting file recovery is plotted for reference, as determined by previous work [25]. Error bars represent 95% confidence intervals. X-axis errors are taken from triplicate qPCR data (see Methods), and y-axis errors are the result of 100 simulations of the original sequencing data sub-sampled to 20x sequencing coverage (see Methods). (**b**) Each plot illustrates behavioral similarities for each file in each diluent condition, with a power regression overlayed (see Appendix B.4). The data used here are also sub-sampled to a sequencing coverage of 20x.

At approximately 15.6 bytes of data encoded per strand [25], this encoding scheme and experimental protocol emulate approximately 150 GB of digital data per microliter, and 7 TB in the final 50μL solution.

If PCR file retrieval fails in complex settings, we would observe a marked difference between sequences recovered in complex versus less complex settings. In addition, we might observe an inability to recover small files in complex conditions. Encouragingly, we observed neither of these symptoms. We found no distinguishable difference when comparing sequence loss between water and complex 150Nmer dilution conditions (Fig. 4.2a). For the large file, only three samples distinguishably differed for the two complexity conditions; this difference

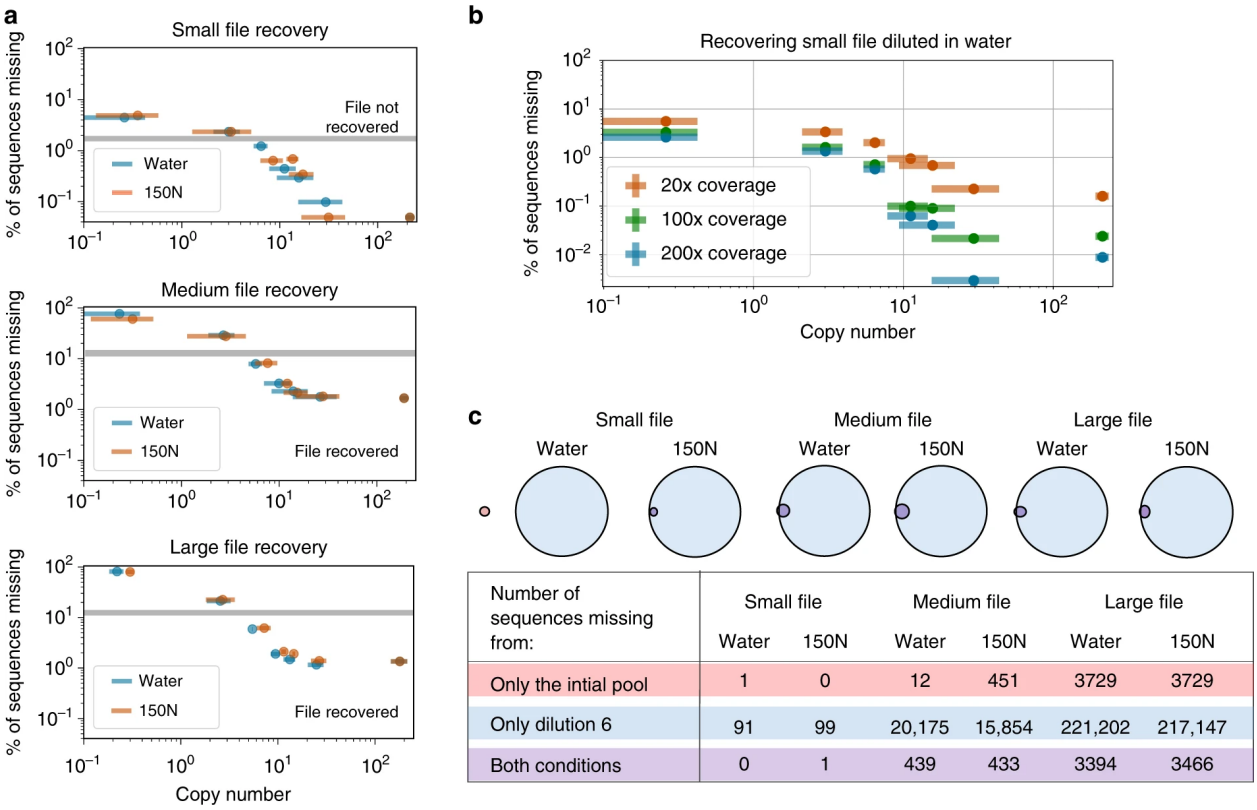
was negligible, with a mean difference of sequences missing of 0.98% and standard deviation of 1.82%.

Regardless of file accessed, decreasing the copy number yielded similar behavior (Fig. 4.2). The loss of sequences for all three files was modeled with a power regression with R^2 values ranging from 0.90 to 0.98 (Fig. 4.2b). However, though the medium and large files behaved almost indistinguishably, the small file did not lose sequences at a similar rate after the copy number fell below approximately 10 (see Fig. 4.2a). Instead, it lost fewer sequences than the larger two files. This is likely due to a combination of copy number being slightly higher than calculated, fewer sequences initially missing due to variation in synthesis, and the distribution of sequences being slightly more uniform (see Appendix B.3 for detailed analysis).

Encouragingly, the size of the file being recovered also does not impact the copy number required in complex pool conditions (Fig. 4.2, Fig. 4.3a). Regardless of pool complexity or size of file accessed, only 10 copies of each strand on average, with a standard deviation of 3, are required for successful recovery with no bit errors. A pool complexity emulating nearly 150 GB of digital data per PCR reaction did not hinder file recovery, and the fact that recovering data from this complex pool was indistinguishable from the water-diluted pool with orders of magnitude fewer strands suggests that we have not approached the limit of pool complexity. Storing many unique sequences in one pool reduces the need for physical isolation, one of the largest density overheads facing this technology.

4.3.2 Data Density

Determining the need for a minimum of approximately 10 copies of each DNA sequence in the PCR reaction to successfully recover a file enables us to calculate a density of 17 EB g^{-1}, nearly two orders of magnitude denser than prior work [19] and closer to the maximum theoretical density predicted for DNA data storage [16] (see Appendix B.5). This is due both to the wet lab techniques used (such as synthesis, sequencing, or library preparation), as well as the encoding and decoding scheme's error tolerance (logical redundancy). In this context,



Number of sequences missing from:	Small file		Medium file		Large file	
	Water	150N	Water	150N	Water	150N
Only the intial pool	1	0	12	451	3729	3729
Only dilution 6	91	99	20,175	15,854	221,202	217,147
Both conditions	0	1	439	433	3394	3466

Figure 4.3: **(a)** The limit of successful, no bit error decoding is shown with the grey bar in each graph. Data points below the grey bar are samples where the file was successfully decoded and recovered with no bit errors. Done post-sequencing, decoding involves clustering sequences, finding consensus, then correcting errors [25]. A more detailed view of this data including exact copy number and sequencing coverage is in Appendix B.7. **(b)** For the small file, when each sample diluted in water was sequenced at greater depth, minimal improvement on the proportion of missing sequences occurred. The 100x coverage data was found by sub-sampling the data used to create the 200x coverage data. **(c)** Missing sequences are compared between the initial pool prior to any dilutions where the mean copy number was 194 (in red) and the last dilution where the mean copy number was less than 1 (in blue). Note the different total number of sequences in the small (2,042) medium (26,404) and large (271,447) files. The fact that some sequences are missing only from the initial pool but "reappear" in the final dilution suggests that the lost sequences are a result of stochastic variation that occurs during sub-sampling for file recovery, rather than irretrievably lost due to some property of the sequence. This pattern of sequences reappearing in subsequent dilutions is shown for every dilution step in Appendix B.6. Note that for **(a)** and **(b)**, x-axis and y-axis error bars represent 95% confidence intervals. X-axis errors are taken from triplicate qPCR data (see Methods), and y-axis errors are the result of 100 simulations of the original sequencing data sub-sampled to 20x sequencing coverage (see Methods).

physical redundancy is the number of copies of each sequence of DNA; logical redundancy in this context is the amount of extra digital information added to aid error correction and mitigate the effect of erasures (missing sequences) and it therefore increases the total number of sequences. Physical and logical redundancy are closely related regarding data density. With more logical redundancy, the system tolerates lower physical redundancy as more sequences can be lost while still allowing perfect recovery of digital data. The logical redundancy used in this work was 15%, which tolerates a maximum of 13% of sequences missing (if there are 100 sequences and we apply 15% logical redundancy, there are now 100+15 sequences and we can tolerate 15/115 missing). If the logical redundancy had been 50%, a maximum of 33% of the strands could be lost ($\frac{50}{100+50}$) [25]. Based on the rate at which sequences are lost (Fig. 4.2b) and ignoring all other errors (such as insertions, deletions, substitutions), the maximum data density is 38 EB g^{-1} for the scheme with 15% logical redundancy and 124 EB g^{-1} for 50% logical redundancy. A relatively small increase in logical redundancy allows for a disproportionately lower physical redundancy with no data loss. This results in higher overall data density. It is important to note that determining the appropriate logical and physical redundancy depends on the specific DNA data storage workflow, for errors incurred by DNA synthesis, sequencing, and wet protocols used can affect sequence recovery.

4.3.3 Sequence Behavior

To further investigate the role of complex pools on sequence recovery, the behavior of each individual sequence was compared for both dilution conditions. We measured sequence behavior by examining the proportion of each sequence present in each dilution. Having a proportion that consistently changes for a subset of sequences would indicate that some strands are being systematically disproportionately accessed and amplified. Yet we observed no difference in sequence behavior between the two different dilution conditions. Thus, sequences may be changing proportions or are observed to be missing due to stochastic variation in sub-sampling. Further supporting this hypothesis is the fact that, although most

sequences absent from the initial pool are also absent from subsequent dilutions, many of the strands "reappear" in subsequent dilutions as shown briefly in Fig. 4.3c; this indicates that these sequences disappeared and reappeared due to stochastic effects rather than systematic interactions that made certain sequences irretrievable. See Appendix B.6 for more detail and analysis.

This significant finding demonstrates the robustness of PCR itself and primer design methodology presented in Chapter 3 and other prior work [25, 68]. It is important to note that the encoding scheme presented in Chapter 3 [25] and used here encodes the payload between the primer regions by using a randomized seed to generate a random payload and exclude homopolymers. While the effect of homopolymers in the payload has recently found to be minimal in practice [69], it is encouraging to note that the care taken to randomize payloads yields no systematic sequence loss. The finding that sequences are lost stochastically coupled with the finding that there is no recovery threshold difference between samples diluted with water and those in more complex settings diluted with 150Nmers thus assures users of this encoding scheme that we have not yet reached the limit of how complex a pool can be before it inhibits the ability to recover desired information.

4.3.4 Sequencing Depth

While storage density is a crucial component of DNA data storage, sequencing efficiency is also critical due to its significant time and cost. We found that the samples successfully recovered at the lowest possible copy number had a mean sequencing coverage of 35x with a standard deviation 11x, with a mean copy number of 10 and a standard deviation of 3 when accounting for all files and all diluent conditions (Appendix B.7, Fig. 4.3a). Thus, we show that it is possible to successfully recover files with no bit errors using the encoding scheme presented in Chapter 3 [25] with a physical density that approaches the maximum theoretical density of one copy per sequence, without compromising sequencing efficiency.

We next examined the degree to which simply sequencing more of the prepared sample aids file recovery. To do so, all water-diluted samples from the small file were sequenced a

second time from the exact same post-library-preparation material, this time with a much higher sequencing coverage. Previously, all samples had a mean sequencing coverage of 24.5x with a standard deviation of 1.6x and were sub-sampled randomly with replacement to 20x coverage. Here, samples had a mean sequencing coverage of 549x with a standard deviation of 148x and were sub-sampled randomly to 200x coverage. This resulted in a mean of 1.8% fewer missing sequences for 200x coverage, with a standard deviation of 1.1% (Fig. 4.3b). This method of increasing sequencing coverage has the benefit of not requiring additional material from the original pool because there is extra material post library preparation to sequence many times over. However, to significantly improve levels of recovery, over an order of magnitude more sequencing resources must be used. Since increasing sequencing coverage significantly increases the cost of recovery, this process is useful only as a last resort.

4.4 Methods

4.4.1 Dilution

The process of diluting the starting pool, once in water and once in 150Nmers (strands of DNA 150nt in length, with 'N' as the input when ordering from IDT to result in random sequences, see Appendix B.2 for a gel electrophoresis analysis) was repeated twice to confirm consistency in qPCR behavior. However, only one of the dilutions was sequenced, and those are the copy numbers reported throughout this paper. Fig. 4.1c details the volume of diluent and sample used for each dilution step. To minimize variation in copy number due to pipetting error, the pipette used to perform the dilutions for the two samples was not adjusted between uses.

4.4.2 qPCR Protocol

From all dilution samples, the same three files were amplified in triplicates via qPCR. To find the most accurate standard curve for each file, an arbitrary ultramer from the relevant file was used (also in triplicate). We ordered all ultramers from IDT. See Appendix B.2 for

primer and ultramer sequences as well as amplification efficiencies.

Each file was amplified using the following qPCR recipe: 1 μL of diluted pool, 0.5 μL of the appropriate forward primer at 10μM, 0.5 μL of the appropriate reverse primer at 10μM, 10 μL of 2x Kapa HiFi enzyme mix, 7 μL of molecular grade water, and 1 μL 20x Eva Green. The following qPCR protocol was used: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 62°C for 20 s, (4) 72°C for 15 s, (5) repeat steps 2-4 as needed.

4.4.3 Calculating Copy Numbers

For the large file, qPCR measurements were taken in triplicate, with an arbitrary ultramer from the file used as a custom qPCR standard (also measured in triplicate) to measure the number of copies of each oligo for every sample. The file's mean amplification efficiency and difference from its standard curve are detailed in Appendix B.2. This identical method of qPCR with a custom standard was used to calculate the first (undiluted) copy number for the medium and small files. However, only the first, undiluted sample could be quantified in this way; this was due to the difference in amplification efficiency between the small and medium files and their respective standards, likely the result of the low number of target strands leading to non-specific amplification (see Appendix B.2). A different method was thus used to determine copy numbers for the remaining samples associated with the small and medium files.

For the small and medium files, the diluted samples' copy numbers were calculated using the large file's dilution factors. Here, because diluting the pool dilutes all three files simultaneously, the dilution factor between subsequent large file samples was the same as it was between subsequent samples for the small and medium files. Thus, the dilution factor (DF) between each subsequent dilution was found for the large file's samples. The initial, undiluted sample's copy number for the small or medium file (CN_0) was then multiplied by the first (DF_1) to yield the copy number of the first dilution:

$$CN_1 = CN_0 * DF \tag{4.1}$$

This has the general formula:

$$CN_n = CN_{n-1} * DF_n \tag{4.2}$$

4.4.4 Calculating Margin of Error for Copy Numbers

For the large file, error is presented as the 95% confidence interval found from the variation of the triplicate qPCR reactions. This was calculated from the standard error for the triplicate qPCR, translated from cycle standard error to copy number standard error using the standard curve. Next, to find the 95% CI for each copy number, the standard error was multiplied by the appropriate t-table value with 2 degrees of freedom (4.303).

However, for the small and medium files, calculating the 95% CI entails incorporating both the error caused by initial qPCR measurement and that from the large file's observed dilution factors used to calculate subsequent copy numbers. This is because the equation used to calculate copy numbers is shown, where CN_{n-1} and DF_n both have a previously calculated 95% confidence interval:

$$CN_n = CN_{n-1} * DF_n \tag{4.3}$$

To calculate the qPCR measurement error for each sample (δCN_n), the first undiluted sample's 95% CI was calculated in the same way as it was for the large file (detailed above). For each remaining sample, because qPCR data was not used and copy number was calculated with the large file's dilution factor, the prior dilution sample's copy number variation (δCN_{n-1}) was multiplied by the observed dilution factor (DF_n) as measured from the large file. Thus:

$$\delta CN_n = \delta CN_{n-1} * DF_n \tag{4.4}$$

To calculate the error that results from the large file's dilution factor measurements (δDF_n), we first found the greatest dilution factor that could have been calculated within the 95% confidence interval for calculated copy numbers by the following:

$$DF_{max} = \frac{CN_n + \delta CN_n}{CN_{n-1} - \delta CN_{n-1}} \tag{4.5}$$

Thus, the variation in copy number due to the dilution factor (δDF) was:

$$\delta DF_n = |CN_{n-1} * DF_{max} - CN_{n-1} * DF_n| \tag{4.6}$$

Finally, to determine the final 95% CI for small and medium files' diluted samples' copy numbers (δCN_n), the propagation equation incorporated both δCN_{n-1} and δDF_n and was represented by the following, easily solvable, standard error propagation equation:

$$\frac{\delta CN_n}{|CN_n|} = \sqrt{\left(\frac{\delta CN_{n-1}}{CN_{n-1}}\right)^2 + \left(\frac{\delta DF_{n-1}}{DF_{n-1}}\right)^2} \tag{4.7}$$

4.4.5 Library Preparation and Enrichment

All files were amplified using the following recipe (primer sequences can be found in Appendix B.2): 1 μL of sample, 0.5 μL of the appropriate forward primer at 10 μM, 0.5 μL of the appropriate reverse primer at 10 μM, 10 μL of 2x Kapa HiFi enzyme mix, and 8 μL molecular grade water. The following PCR protocol was used: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 62°C for 20 s, (4) 72°C for 15 s, (5) repeat steps 2-4 as needed according to prior qPCR.

Subsequent sequencing preparation via ligation was done with a modified version of Illumina TruSeq Nano ligation protocol and TruSeq ChIP Sample Preparation protocol. Step by step instructions are in Appendix B.8 for convenience, but briefly, samples were first converted to blunt ends with the ERP2 reagent and directions provided in the Illumina TruSeq Nano kit, then purified with AMPure XP beads according to the TruSeq ChIP protocol. An 'A' nucleotide was added to the 3' ends of the blunt DNA fragments with the TruSeq Nano's A-tailing ligase and protocol, followed by ligation to the Illumina sequencing adapters with the TruSeq Nano reagents and protocol. We then cleaned the samples with Illumina sample purification beads and enriched the sample using an 8 cycle PCR protocol given in the TruSeq Nano protocol.

For the enrichment, all samples were enriched using the following recipe: 3 μL of a ligation sample, 3 μL of the PCR Primer Cocktail provided in the TruSeq Nano kit, 12 μL of Enhanced PCR Mix provided in the TruSeq Nano kit, and 12 μL molecular grade water. The following PCR protocol was used: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 60°C for 15 s, (4) 72°C for 30 s, (5) repeat steps 2-4 for a total of 8 times. The length of enriched products was confirmed using a Qiaxcel bioanalyzer. Notably, these are the same ligation and sequencing preparation methods presented in Chapter 3 [25]. Reformatted instructions are given in Appendix B.8 for convenience.

4.4.6 Next Generation Sequencing

When multiple separate samples were prepared for sequencing, these samples were mixed proportionally (e.g., a 10,000 oligonucleotide file to be sequenced with a 500,000 file would comprise 1.96% of the DNA material in this mix). The mixed sample was then prepared for sequencing by following the NextSeq System Denature and Dilute Libraries Guide. The sequencing sample was loaded into the sequencer at 1.3 pM, with a 10 to 20% PhiX spike-in as a control (PhiX is a reliable, adapter-ligated, well-characterized genomic DNA sample provided by Illumina).

4.4.7 Sub-sampling Data to Calculate Percent of Sequences Missing

To remove the effect of varying sequencing coverage, aligned sequences were randomly sub-sampled with replacement down to 20x coverage (*20x coverage = number of oligos in file * 20*), and the resulting number of missing strands was recorded. This was performed 100 times per file per sample to yield mean percent sequences missing and the 95% confidence interval.

4.5 Discussion

In summary, as DNA data storage becomes an increasingly viable alternative to mainstream data storage methods, the ability to perform random access on small subsets of densely

stored DNA is increasingly important. With the maximum theoretical information density of 2 bits per nucleotide and and ideal copy number of 1, one could achieve a maximum theoretical density of 455 exabytes g^{-1} [16]. However, there are many practical design trade offs including the ability for random access and error correction that the DNA data storage community has made. By demonstrating files with a copy number of approximately 10 can be successfully recovered, we present the most practically dense system to date at 17 EB g^{-1}, nearly two orders of magnitude greater than the densest prior work [19]. Furthermore, by showing that we can reliably access files that encode 0.1 KB, 1.7KB, and 18KB from a complex pool emulating nearly 150 GB of digital data per PCR random access reaction, we demonstrate the ability of the storage method used in this work to enable efficient, reliable data retrieval in complex settings. We have no reason to think that we have experimentally reached the limit of how complex the pool could be for this DNA data pipeline, further supporting the robustness and unprecedented storage potential of DNA.

Chapter 5

CRISPR-CAS9 FOR SIMILARITY SEARCH IN DNA DATA STORAGE SYSTEMS

5.1 Chapter Abstract

While previous chapters have explored PCR-based random access, this chapter examines the efficacy of using CRISPR-Cas9 to perform similarity search. Using Cas9 to retrieve data is much more energy efficient than PCR or prior work implementing similarity search with DNA hybridization, and has the additional benefit of being a much simpler and faster lab protocol. Here, we demonstrate the use of CRISPR-Cas9 in similarity search by encoding a 1.7 million image database and simulating multiple queries to successfully return a selection of the most similar images. While similarity search performance is less than traditional silicon-based computer implementations and prior DNA-based similarity search work using hybridization, the speed and energy-efficiency of DNA-based similarity search using Cas9 may be an attractive alternative at very large scale when speed and energy-efficiency are crucial.

5.2 Overview

In early DNA data storage architectures, accessing specific files required sequencing the entire DNA library [6, 16, 17, 70]. To reduce costs and increase resource efficiency, scaling up DNA data storage requires methods to selectively retrieve and read only particular pieces of data, i.e., random access. Therefore, recent emphasis has been placed on developing increasingly efficient DNA retrieval methods in the context of DNA data storage, with virtually all of these methods employing PCR-based random access [21, 22, 24, 25]. This PCR-based approach, however, is time-consuming, requires careful design and validation of primers to

avoid crosstalk [25, 71], and has limited multiplexability [25, 70] (the ability to retrieve multiple files in a single reaction). Accessing specific portions of a DNA data storage pool in multiplex has been previously demonstrated, but only on a small scale [70]. To efficiently recover data at scale, large-scale, multiplex random access will likely be necessary.

However, retrieving multiple items simultaneously can be achieved with methods other than random access, and may have distinct advantages. Similarity search is one such method that returns multiple items, where the items returned from a database are similar to the query item. This is entirely distinct from random access because while a well curated database might assign metadata or even a unique key for precise lookup for each item, modern search platforms do not assume users know the exact metadata tags or keys of items they wish to access. Recently, similarity search was successfully performed in DNA by leveraging DNA hybridization [41]. The advantage of using DNA rather than traditional silicon-based computers is that the query molecules are able to diffuse in solution and interact with all items in the database in parallel, which could significantly improve search times at large scale. However, the previous hybridization-based similarity search work requires a significant and lengthy temperature gradient to perform search, which at scale could require a prohibitive amount of energy.

Here, we present a CRISPR-Cas9-based similarity search (C9SS) method paired with nanopore sequencing (Fig. 5.1). CRISPR-Cas systems have varying rates of off-target cleavage activity [49], and in this work we exploit off-target cleavage activity as a feature using the Cas variant with the greatest off-target cleavage (wtCas9) to perform similarity search. While wtCas9 similarity search trades data retrieval accuracy for energy efficiency and speed, retrieval performance is still much better than random and may be useful for applications that value speed and/or energy efficiency over retrieval accuracy.

5.3 Results

5.3.1 Similarity Search Sequence Encoding and Simulation

To perform similarity search using CRISPR-Cas9, each image is represented by one DNA sequence as part of the encoding process. Each sequence corresponds to one item from the database and has the same front and back primer sequences as all other items in the database (see Appendix C.1).

To retrieve sequences similar to a query image sequence, the query image is itself run through the encoder to generate the complementary Cas9-RNPs that cut the similar sequences in the database. Adapters for nanopore sequencing are ligated only at the cut sites where there is an exposed phosphate group, thus only these sequences with the ligated adapter are sequenced and retrieved. The retrieved File IDs can then be used to look up the encoded files.

The first step in this process is to encode the image database into DNA sequences. A detailed description of this process is in the Methods section and a summary is shown in Figure 5.1C, but in brief, to train the encoder model, we introduce batches of image triplets from the training data set. Each triplet consists of an 'anchor' image, one image similar to the anchor, and one image dissimilar to the anchor. Each image is transformed to a feature vector using the FC2 layer of the VGG16 image classification model, and the feature vector acts as a summary of the image. The shorter the Euclidean distance between a pair of feature vectors, the more similar the corresponding images are. All three feature vectors from a triplet are passed through two fully connected layers (the 'Encoder') resulting in one-hot vectors for each image, which represent the encoded DNA sequence. A Cas9 cleavage model (modified from previous work [49] to be differentiable and thus efficiently used in this machine learning model) then operates on the sequences to predict the Cas9 activation rate for the anchor-similar and anchor-dissimilar sequence pairs. Then, the activation rates are used to compute a loss, which is back-propogated to train the encoder to maximize the activation on similar pairs while minimizing it on dissimilar pairs.

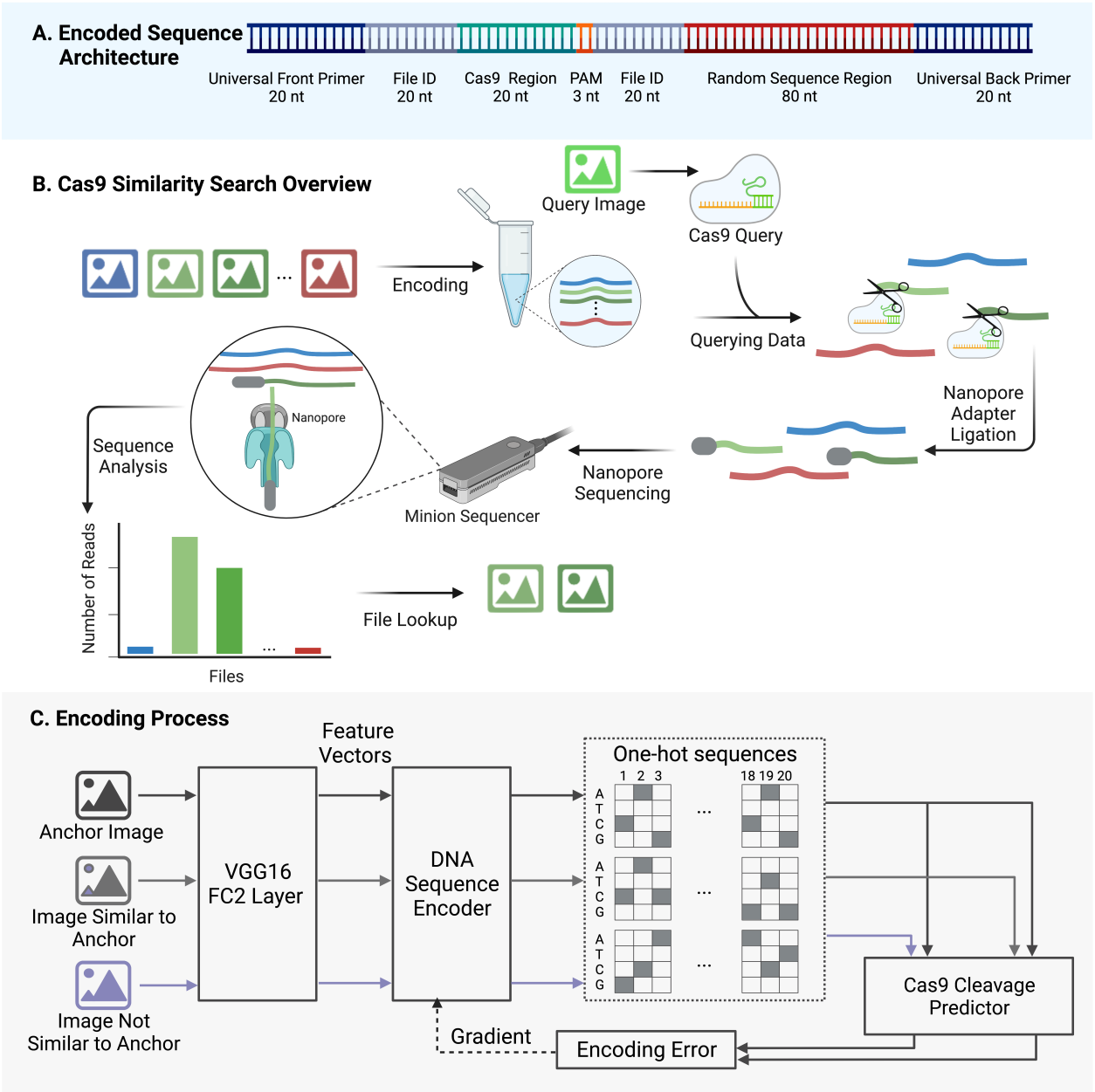


Figure 5.1: (**A.**) The sequence design after encoding. (**B.**) Similarity search workflow. (**C.**) Encoding model overview.

Sequence encoder performance was evaluated by simulating similarity search via two different methods (Fig. 5.2).

In one method of model performance evaluation, we simulated three different image queries. One query is an image of a tuxedo cat, one is of Lego sushi, and one is the Taipei 101 building (Fig. 5.2A). These three image queries are encoded using the same encoding process as the rest of the database, the only difference is that the output sequence is then ordered as an RNA guide to be used with Cas9. Each query sequence is paired with every target sequence and run through the Cas9 cleavage predictor. If the score is greater than 0.1 (which represents a relatively small rate of cleavage), then the target is considered retrieved. As shown in Fig. 5.2A, our model performs significantly better than an untrained model. However, the query performance varies significantly. Ideally, similar images (Euclidean distances $<= 75$) should be returned with a score above the threshold, and all similar images should be retrieved while no dissimilar images (Euclidean distances > 75) should be retrieved. Each query returns a large proportion of similar images, but also returns a large portion of dissimilar images.

However, model performance changes drastically depending on the threshold used. To examine our model more comprehensively than using three image queries and one threshold, our second method of model performance evaluation draws 50,000 random image pairs where half the pairs are two images similar to one another, and half the pairs are dissimilar (Fig. 5.2B). The proportion of the 50,000 image pairs mistakenly retrieved is shown on the x-axis (in other words, the proportion of dissimilar images retrieved), while the proportion of correctly retrieved image pairs is on the y-axis (the proportion of similar images retrieved). Each data point represents the performance of a different threshold. An ideal model with an ideal threshold would be represented as one data point in the top left corner, where all similar images are retrieved and no dissimilar images are retrieved. Random performance is represented by the grey dashed line. To compare model performance, the area under the curve above random performance was calculated. We trained three models, the model that had the greatest area under the curve was considered the best, and that model's results are

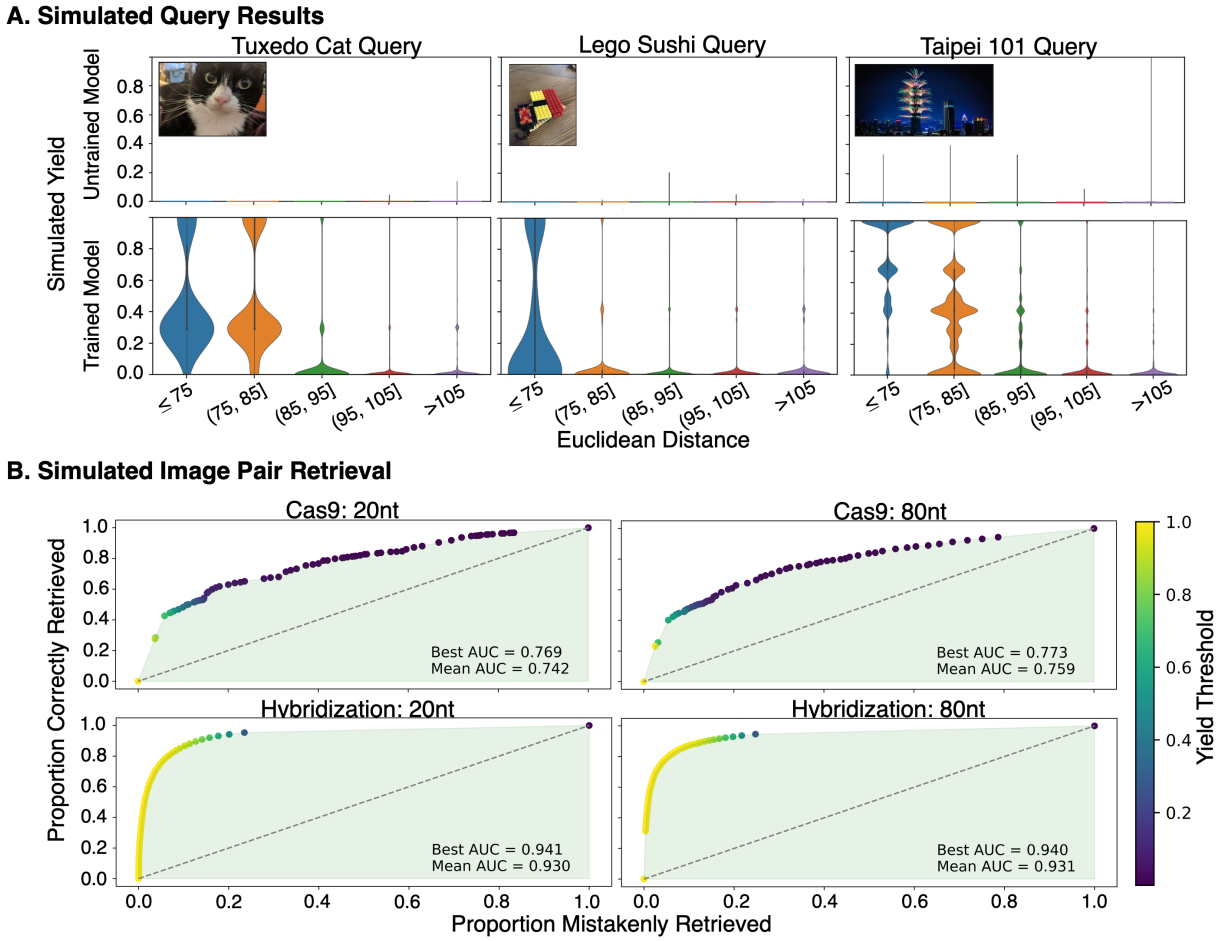


Figure 5.2: (**A.**) From a dataset of 1.7 M images, all images were encoded into 20nt DNA sequences. Three queries were encoded and the reaction between the Cas9 guide sequence (the query) and all targets was simulated. The results show the difference in recall between an untrained encoder model and the trained encoder model. Images with Euclidean Distance $<= 75$) are considered similar and should have a higher yield than dissimilar images, which should have little to no yield. (**B.**) Here, 50,000 image pairs were generated. The pairs are balanced, meaning half the image pairs are similar to each other (the two images have feature vectors with a Euclidean distance $<= 75$), and half are not similar to each other (Euclidean distance > 75). Each data point represents the proportion of the 50,000 image pairs that was correctly retrieved or mistakenly retrieved (sometimes called a false positive) using a given yield threshold, as denoted by the color bar. A high yield threshold indicates very high rates of the molecular mechanism in question- Cas9 cleavage of DNA or DNA hybridization, respectively- are required to be classified as a recall. The dashed line represents random recall. The area under the curve (AUC) is shaded green and is quantified for ease of comparison between methods. 1.0 is the maximum possible AUC value.

shown in Fig. 5.2.

When evaluating model performance, it is clear from these results that using one 20nt wtCas9 site to perform similarity search does not perform as well as prior work using an 80nt site for DNA hybridization [41]. We hypothesized there are two possible reasons the Cas9 similarity search does not perform as well as DNA hybridization similarity search. First, a 20nt search site may provide insufficient sequence space to encode a large dataset. Second, the fundamental behavior of Cas9 is too selective for performing similarity search.

To examine these hypotheses we trained three Cas9 encoding models to use four 20nt Cas9 sites and thus a total sequence space of $4 \cdot 4^{20}$nt, where a Cas9 cleavage at any of the four sites would retrieve the target. While we see improved similarity search performance, we do not see meaningfully significant improvement. We then trained three DNA hybridization models using the methods from prior work [41], and three models identical to prior work except that the search site was reduced to only 20nt. These 80nt and 20nt hybridization models yield similar results with the longer sequence model performing only slightly better than the shorter sequence model. Both DNA hybridization models perform significantly better than the Cas9 models, which suggests that the molecular mechanisms behind Cas9 and hybridization are what drive the difference in similarity search performance.

This makes intuitive sense because DNA hybridization is less sensitive to mutations between two sequences. If we consider each possible sequence as a node in a graph, with edges between any pair of nodes with 'high' activation rates, the Cas9 graph is far less connected than the hybridization graph for a given activation threshold. The rate of hybridization is hardly affected after 16 mutations, but with Cas9 it is often reduced by as much as a factor of 10^{-3} after just 2 mutations. This makes Cas9 ill-suited to representing complex distance relationships between images (e.g., A is similar to B, B is similar to C, but C is not similar to A) for a large database compared to hybridization.

Cas9 encoder model behavior is explored in more detail by examining the encoded sequences and activation behavior in Appendix C.2. The lack of sequence diversity in the PAM proximal positions (where any sequence substitutions result in drastically lowered activation)

and the increased diversity in PAM distal positions (where sequence substitutions result in minor changes in activation) further suggest that Cas9 sensitivity to sequence perturbations limits its ability to perform similarity search. Furthermore, when examining activation behavior as a function of sequence edit distance (Fig. C.3 in Appendix C.2), we observe that Cas9 cleavage behavior is far more stringent than DNA hybridization, regardless of sequence size, as evidenced by Cas9's small range of edit distances that can be acted on by Cas9. Essentially, by substituting only 20% of bases with one Cas9 site, or 30% of bases with four Cas9 sites, the mean activation score is 0, and the drop off in activation score per edit distance is very steep. In contrast, hybridization, regardless of sequence space size, still has activation values above 0 until 70% of sequence positions are mutated, with a much less steep drop off in activation score per edit distance. Taken together, it is clear that while the wtCas9 variant used here performs similarity search much better than random, it struggles to perform similarity search well due to its sensitivity to sequence mutations not allowing for the preservation of complex distance relationships. However, it is not clear that using a less precise Cas9 variant would have the same limitations, and this is discussed in the Discussion section below.

5.4 Methods

5.4.1 Similarity Search Datasets

With the exception of the query images, all images were collected from Open Images V4, a dataset of over 9 million URLs for images with Creative Commons licenses. Of these, approximately 1.7 million are hosted by the CVDF and available for download; the rest are raw Flickr URLs and may or may not be available. For the image database used in our experiments, we took all images from the hosted set. For training, we took images from the full set of 9 million that were not used for training, testing, or experiments.

5.4.2 Similarity Search Feature Extraction

To extract image features, we processed each image with VGG16, a convolutional neural network designed for image classification. The weights were loaded from the publicly available trained model and left unchanged during our processing. We used the activations of FC2 (the second fully-connected layer) as 4096-dimensional feature vectors. Using the same metric as prior work [41], pairs of images with Euclidean distance of 75 or less tend to be consistently similar, so during training we label these pairs as "similar" and all other pairs as "not similar".

5.4.3 Similarity Search Sequence Encoding

The sequence encoder is a fully-connected neural network. The 4096-dimensional FC2 vectors are fed into a 2048-dimensional hidden layer with a rectified linear activation, followed by an output layer with a "one-hot" sequence representation. The output layer has dimensions N by 4, where N denotes the number of nucleotides in the sequence, typically 20 nucleotides (though this was modified to 80 to compare the effects of having multiple Cas sites). In this representation, each sequence position has four channels, one for each base. A straight through estimator is used to convert each base to a one-hot vector with a hardmax function, while a softmax activation function is applied during backpropagation to estimate the gradients despite the non-differentiable hardmax function used in the forward pass. A DNA sequence can be read off by picking the channel with the maximum activity at each position.

The yield predictor (a modified version of the previously developed NucleaSeq [49]) takes a pair of one-hot sequence representations and produces an estimate of the yield of the cas9 cleavage reaction between the first and second sequence.

5.4.4 Modifying NucleaSeq

NucleaSeq, the original Cas cleavage prediction model developed by Jones et al. [49], was modified for this work. Most notably, NucleaSeq takes into account substitutions, deletions and insertions between the gRNA sequence and the DNA sequence. However, because our encoding process uses a uniform length of 20 (or multiples of 20 when investigating the affect of multiple Cas sites), the differences between our strands are entirely described by substitutions. By only taking substitutions into account, we could then make the model differentiable and thus use it in a machine learning context.

To make the model differentiable, the model weights from NucleaSeq were used and the original NucleaSeq functions were rewritten in a tensorflow to be efficiently used in our training process. Our implementation can be found at `https://github.com/uwmisl/cas9-similarity-search/`.

5.4.5 Similarity Search Encoder Training

During each round of encoder training, we draw a batch of pairs of feature vectors from the training set. This batch of pairs is formed by randomly choosing "anchor" images - please note these are not images of anchors. Each anchor image is then paired with a dissimilar image, (defined as having a Euclidean distance between the image feature vectors >75), and a similar image (Euclidean distance between the image feature vectors <=75). This process of using anchor images in this way is known as triplet loss. Due to memory constraints, the training dataset is broken up into 16 batches, one batch is loaded at a time, and training samples are drawn from the currently loaded batch. Throughout training, the batch is periodically changed by randomly selecting a new batch file to load into memory. After selecting random anchors, it is sometimes not possible to find a vector similar to the anchor. If, after a fixed amount of searching for a similar vector one is not found, the anchor vector is duplicated as the similar vector.

The batch of triplets is processed by the encoder, which produces triplets of one-hot

sequences. The positive and negative pairings from each triplet of sequences are processed by the yield predictor, which outputs the estimated yield of Cas9 cleavage for the positive and negative pairs in each triplet. The estimated yield is calculated by adapting the wtCas9 cleavage prediction model from prior work [49] to be a differentiable function and thus able to be used by our TensorFlow workflow. Loss for each triplet is calculated using the log of the yield predictor. In order to achieve a large (near to zero) log-activation for the positive, and maintain a small (much less than zero) log-activation for the negative sample, the loss is calculated as the difference: the log-activation of the negative sample minus the log-activation of the positive sample. In order to focus the backpropagation training on harder, not yet learned samples, the contribution of both the negative sample and positive sample are clipped at acceptable levels. Positive log-activation scores greater than -0.5 and negative log-activation scores less than -3.0, are clipped at these values preventing them from contributing to training gradients. The encoder weights are trained using the Adaptive Gradient algorithm (Adagrad).

5.5 Discussion

In conclusion, we have evaluated data retrieval using CRISPR-Cas9 for similarity search. We have presented a novel DNA data storage similarity search architecture that enables file IDs to be accessed with CRISPR-Cas9 and decoded using a nanopore sensor array. The advantages of this approach over previous architecture are 1) the isothermal search protocol, which is more energy efficient than the previous DNA hybridization method [41] requiring high temperatures, 2) the quick reaction time of Cas9 that performs similarity search far faster than DNA hybridization, and 3) a much simpler and quicker protocol to implement than similarity search with hybridization.

Taken together, using Cas9 for similarity search may be much more energy efficient and faster at large scale than current silicon-based methods. However, while Cas9 is able to perform similarity search with better than random performance, it may not meet the needs of a similarity search user demanding high precision and low false-negative recall. A Cas9

variant with greater off-target effects may improve performance, but there is no such variant currently commercially available. Therefore, we propose using Cas9 as a hierarchical tool when performing similarity search. For example, rather than extract the 4,096 feature vector from the VGG model used to classify the image dataset in this work, one could use the original VGG output that categorizes images with 1,000 unique labels. With these 1,000 labels, one could take a query image and label it with the VGG model, then essentially perform random access to return all targets with the same label. Then one could either perform a search on this subset of targets with further molecular processes (i.e., strands could have DNA hybridization architecture on the remaining part of the sequence to perform a second round of search using DNA hybridization), or simply use the retrieved target IDs to perform search on a subset of the database in a silicon-based system.

So while using the specificity of Cas9 can be leveraged to perform random access, the off-target effects from less stringent Cas9 variants can be leveraged to perform similarity search. In fact, Cas9 similarity search may even be improved with a different, less precise variant than the wtCas9 used in this work. It is possible that engineering a much "sloppier" Cas9 would allow for a less steep cleavage activity drop off, and thus be a better tool for similarity search. However, this goes against all current trends of Cas9 engineering, as virtually all other applications using Cas9 are in a context where precision is paramount (i.e., therapeutics or gene editing) and increasingly precise Cas9 variants are commercialized each year.

Ultimately, this method of using wtCas9 for similarity search offers greater energy efficiency and retrieval speed than current molecular methods, and has the potential to be used in a broader framework of molecular data retrieval despite its lack of recall precision.

Chapter 6

AN EMPIRICAL COMPARISON OF PRESERVATION METHODS FOR DNA DATA STORAGE

A version of this work was originally presented in Small Methods, Organick et al. 2021 [72]. In addition to myself, the co-authors for this work are Bichlien H. Nguyen, Rachel McAmis, Weida D. Chen, A. Xavier Kohll, Siena Dumas Ang, Robert N. Grass, Luis Ceze, and Karin Strauss.

6.1 Chapter Abstract

To ensure that information encoded in DNA is safely recovered after storage, it is essential to appropriately preserve the physical DNA molecules encoding the data. While preservation of biological DNA has been studied previously, synthetic DNA differs in that it is typically much shorter in length, it has different sequence profiles with fewer, if any, repeats (or homopolymers), and it has different contaminants. In this paper, nine different methods used to preserve data files encoded in synthetic DNA are evaluated by accelerated aging of nearly 29,000 DNA sequences. In addition to a molecular count comparison, the DNA is also sequenced and analyzed after aging. These findings show that errors and erasures are stochastic and show no practical distribution difference between preservation methods. Finally, the physical density of these methods is compared and a stability versus density trade-offs discussion provided.

6.2 Overview

Synthetic DNA has been growing in popularity as a promising new technology for long-term digital data storage [19, 20, 22, 25]. DNA is very dense, with expected practical densities

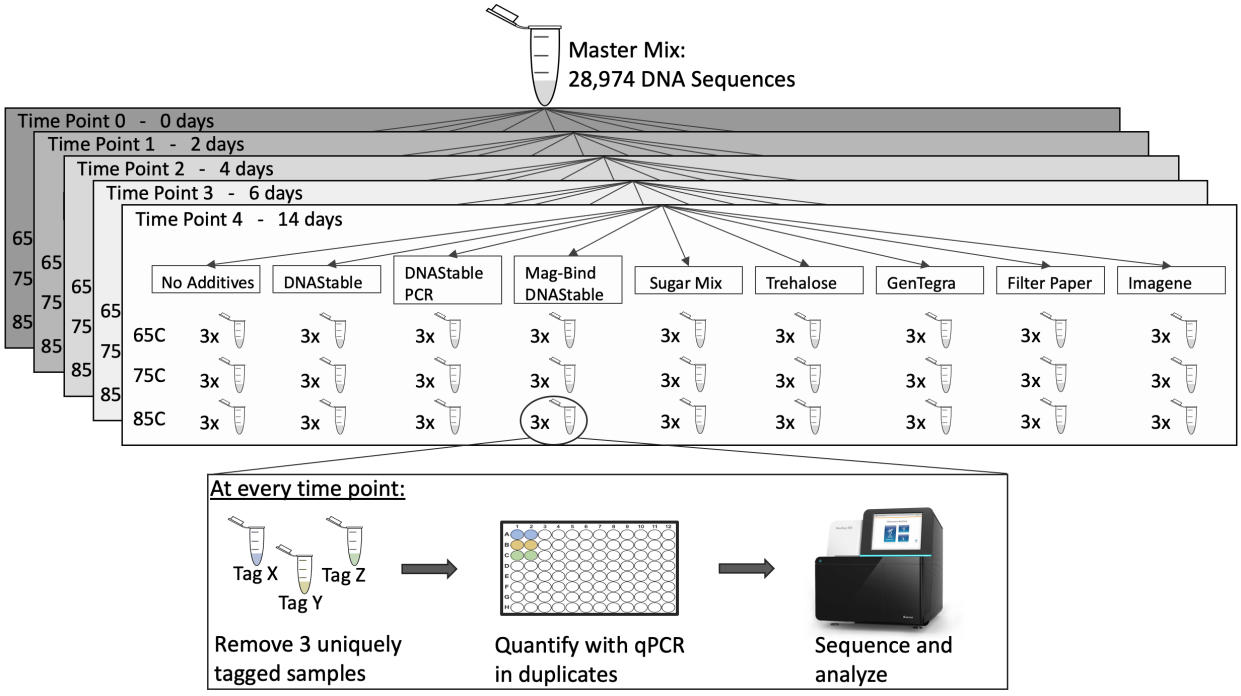


Figure 6.1: DNA sequences were amplified with PCR and mixed together into one master mix tube. Shown in this figure, 81 aliquots were taken from that one tube and prepared for Next-Generation Sequencing (NGS) via ligation, which also added a unique sequencing index (tag) to each aliquot so all 81 samples could be sequenced in the same sequencing run. (The number 81 comes from the three triplicates at each of the three experiment temperatures for each of the nine preservation techniques. Each time point therefore had 81 samples.) Each of these 81 samples was then split into five equal parts, one for each time point. At each time point, the triplicates were removed from each condition and quantified with qPCR in duplicate, then sequenced with NGS if there was sufficient material. Note that the time points were different for Imagene and samples aged at ETH-Z as shown in Fig. 6.2 and Fig. 6.3, respectively.

higher than one exabyte (10^{18} bytes) per cubic inch. This is orders of magnitude higher than current storage media. Unlike other media, whose reading technology quickly becomes obsolete as the technology evolves, DNA is expected to always be readable and compatible with existing and future DNA sequencing platforms due to its prominence in life sciences and clinical applications. Long-term digital data storage relies on the integrity of the physical medium for data endurance over decades up to possibly thousands of years. This is no

exception for digital data storage in synthetic DNA. After storage, anywhere from 10 to over 1000 intact copies of each sequence are required for data retrieval [7, 19].

Biological DNA samples are commonly frozen or chilled significantly below room temperature for preservation. The simplicity of this method makes it attractive in a laboratory setting for storing small quantities of DNA. However, this is significantly less appealing for DNA data storage due to a different set of constraints to reach wide deployment: access to the DNA must be fully automated, density must be kept as high as possible, and the cost and energy to store and maintain the samples as low as possible.

For these reasons, there has been growing interest at storing DNA at room temperature. Most notably, individual studies examine synthetic DNA preservation at room temperature with nanoparticles, encapsulation in metal capsules, trehalose, and various other sugar matrices [29, 73]. While there are a number of commercial products designed to preserve DNA, there have been no large, comprehensive comparison studies of the stability of digital data storage in synthetic DNA under the various methods until now. We selected DNA preservation methods for evaluation by factoring in protocol simplicity and reported ability to store DNA at room temperature. The most "primitive" methods chosen were to store DNA dehydrated in a standard polypropylene eppendorf tube (referred to as "No Additives") and to store DNA dehydrated on filter paper, a relatively old method of preserving and transporting biological specimens dating back to the 1960s [74]. The next set of methods involved mixing DNA with various additives prior to dehydration in an eppendorf tube. These methods included using trehalose (referred to as "Trehalose"), a disaccharide that enables several organisms to survive desiccation [75], and also a mixture of sugars comprised of trehalose, raffinose, maninitol, and uric acid ("Sugar Mix"). Similarly, we included commercially available proprietary sugar mixtures advertised for room temperature DNA storage ("DNAStable" and "GenTegra"). We also investigated more sophisticated encapsulation techniques to preserve DNA. These methods included storing DNA in Imagene DNAshells ("Imagene"), in which DNA is stored in a borosilicate glass insert inside a stainless steel shell and cap, which is hermetically sealed and filled with non-reactive gas, and on magnetic

nanoparticles further coated with DNAStable ("Mag-Bind DNAStable"). While all of these preservation methods were sequenced immediately after rehydration or de-encapsulation with no further preparation or manipulation, we also examined our ability to manipulate a pool of DNA after preservation with a preparation method that included polymerase chain reaction, PCR, ("DNAStable + PCR").

We present an analysis of these different DNA storage methods used to preserve synthetic DNA encoding two different data files. We performed accelerated aging of the samples at 65°C, 75°C, and 85°C to determine the first order decay kinetics of each storage method and analyzed the sequencing data to determine the percentage of sequences recovered. We also tested DNA preservation in four methods at a different lab (ETH-Z) in addition to those performed at our UW lab. Three preservation methods were tested in both labs "No Additives", "Trehalose", "DNAStable") and one was performed solely at ETH-Z ("Magnetic NP").

6.3 Results

6.3.1 qPCR of DNA Material

To evaluate the effectiveness of each storage method, we selected a small data file encoded in 7,373 unique DNA sequences and a larger data file encoded in 21,601 unique sequences. The overall architecture of these sequences features a 110bp payload encoding the digital data flanked by a 20bp forward primer and a 20bp reverse primer, resulting in a sequence 150bp long in total. Each digital file has a unique pair of forward and reverse primer sequences (the details of this architecture are reported in previous work [25]).

The two files were amplified individually via PCR and combined into one solution, totalling 28,974 unique sequences. Aliquots of this solution were used for the aging experiment. Triplicates of each aliquot were subjected to accelerated aging at three storage temperatures (65°C, 75°C, and 85°C) for five different time points. All samples were kept in ovens maintaining 50% relative humidity. More details about each storage method can be found in

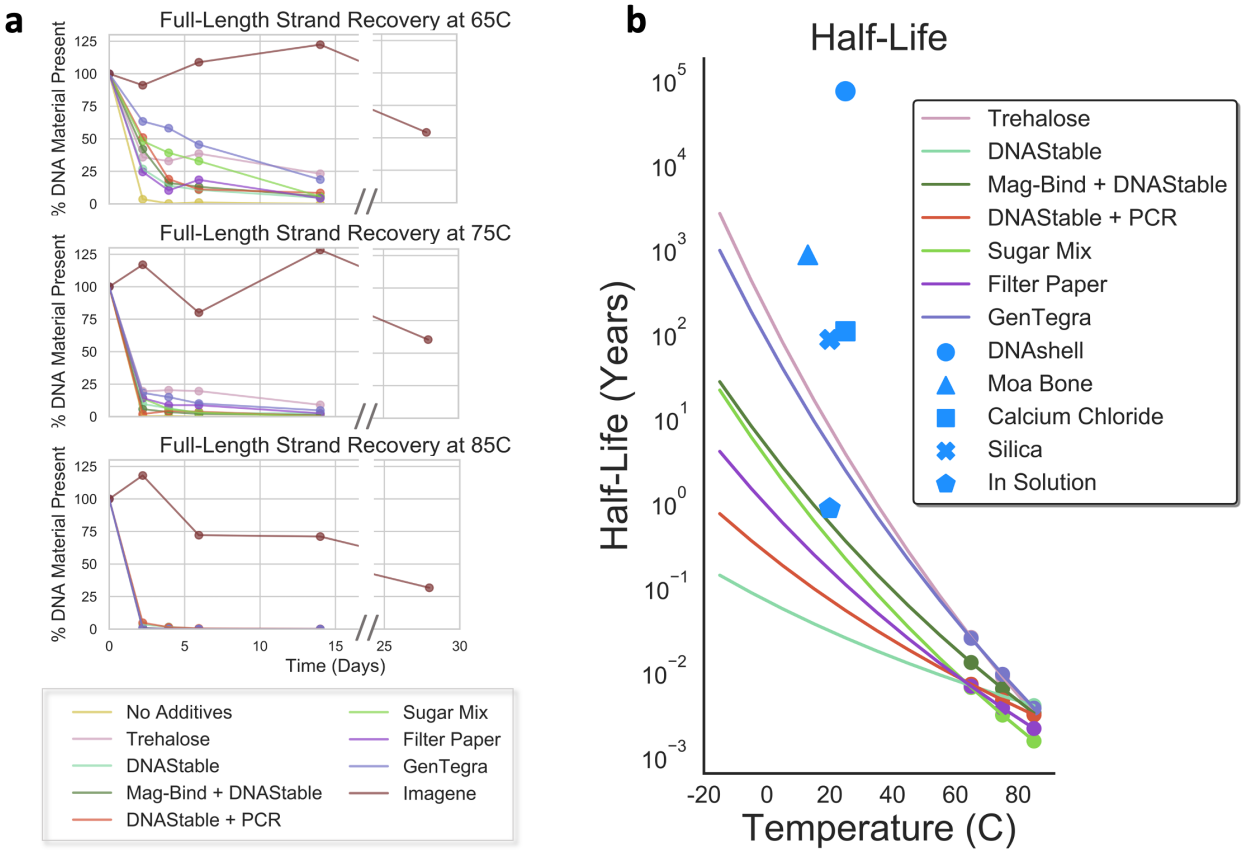


Figure 6.2: (a) The percent of full-length DNA material present after each time point, as measured with a quantitative polymerase chain reaction (qPCR). (b) Extrapolated DNA half life of 150bp DNA strands at various temperatures for all preservation methods that had measurable qPCR data for at least two time points for each temperature. Our Imagene data is not shown because it did not degrade enough to reliably measure a rate of decay, however it has been extrapolated in prior work [76] and is shown here in blue ("DNAshell"). The five data points in blue are shown as a comparison to other reported DNA preservation results in literature (DNAshell technology from Imagene [76], Moa bone [77], calcium chloride [29], silica [73], in solution [78]). All data are scaled to reflect the half life of a 150bp DNA strand following previously established scaling methods [73] of ${t_{1/2}}^{150bp} = {t_{1/2}}^{1bp}/150$. For more calculation information, see Appendix D.6.

their respective Methods sections. To reduce errors associated with sequencing preparation, each sample was ligated with Illumina adapters and tagged with an unique index prior to the accelerated aging experiment and were ready for sequencing directly after aging (note

that the "DNAStable + PCR" method was the exception to this, details in Methods). An overview of this experiment is shown in Fig. 6.1.

At each time point, the concentration of each of the samples was measured by quantitative polymerase chain reaction (qPCR), as shown in Fig. 6.2. Only strands that didn't break are detectable with qPCR, for exponential amplification is only possible when both the forward and reverse primer are present on the same DNA strand. While all DNA preservation methods degraded at a slower rate than only dehydrating the sample (No Additives), there was no clear ranking of preservation methods only looking at one experimental temperature. For example, GenTegra consistently degraded slower than Trehalose at 65°C, but not at 75°C. This could be a result of the various proprietary ingredients included in the GenTegra mix having different temperature-dependent protective properties. The GenTegra user guide specifies that the GenTegra DNA preservation material is "designed to tolerate temperatures of -80°C to 76°C during transport" [79].

To compare the various preservation methods more comprehensively, we measured the decay kinetics by incorporating data from all three temperatures and time points. Assuming first-order kinetics, consistent with prior work [73], we calculated the temperature dependence of the per-nucleotide fragmentation rate (k) using the Arrenhius equation and solved the half-life of the DNA samples at 65°C, 75°C, and 85°C (Fig. 6.2). Details can be found in the Methods section. We did not solve for the half-life of DNA preservation methods that degraded completely by the first time point (No Additives), or did not degrade enough for analysis (Imagene). Our finding that Imagene's DNAshells degraded exceptionally slowly is consistent with previous work performed with biological samples [33, 76].

Each method preserved DNA better than no preservation material at all (No Additives). Based on the extrapolated half-lives, we found that samples preserved with trehalose and GenTegra had nearly indistinguishable longest half-lives, while the two methods utilizing only DNAStable had the shortest half-lives. However, the method utilizing DNAStable in conjunction with magnetic beads performed nearly identically to samples preserved in the mix of sugars and these two methods had the next shortest half-lives. The filter paper

sample degraded only slightly faster than the DNAStable and magnetic beads method and the sugar mix. However, we caution that these extrapolations are sensitive to the amount of DNA material stored and the amount of preservation material, therefore extrapolations shown here are more likely a relative ranking than an exact half-life (see the Discussion section for more details).

6.3.2 qPCR of ETH-Z Material

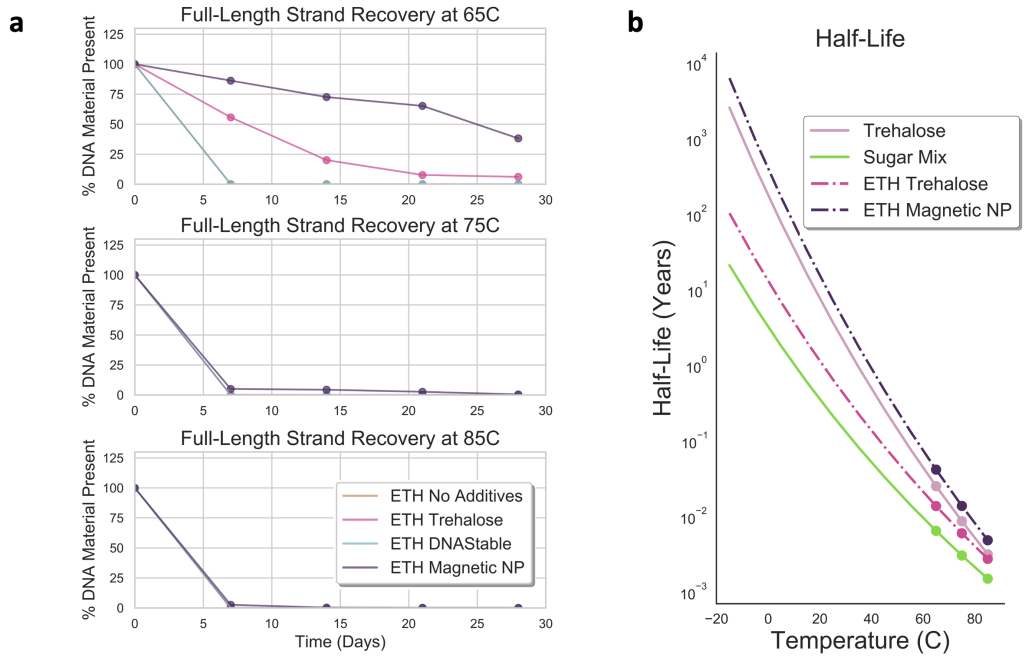


Figure 6.3: (a) The percent of full-length DNA material present after each time point, as measured with qPCR. (b) Extrapolated DNA half-life at various temperatures for all preservation methods that had measurable qPCR data for at least two time points for each temperature. All data are scaled to reflect the half life of a 150bp DNA strand following previously established scaling methods [73] of ${t_{1/2}}^{150bp} = {t_{1/2}}^{1bp}/150$. For more calculation information, see Appendix D.6.

Since the samples aged at ETH-Z contained the DNA sequences prior to library preparation, ETH-Z samples were less than half the length of the remaining samples (150 bp and

310 bp, respectively). Therefore, we derived the per-nucleotide fragmentation rate (k) of all samples and scaled all data to the half-life of a 150bp sequence to provide a direct comparison between the two studies, as shown in Fig. 6.2 and Fig. 6.3. Note that the preservation methods "No Additives" and "DNAStable" both degraded almost entirely before the first time point at all three temperatures tested and were therefore excluded from the half-life extrapolation.

We observed a significant difference in decay kinetics between the UW and ETH-Z samples stored with trehalose. We attribute the differences in preservation to two main factors: difference in the ratio of amount of preservative material to DNA material (0.1M vs 0.02M trehalose), and the different amount of DNA material stored (350 ng vs 24 ng) [36, 80]. We hypothesize the former also explains the difference observed between the DNAStable samples aged at UW and ETH-Z. Nonetheless, the relative rankings of the methods performed at ETH-Z are the same as the rankings found at UW.

It is interesting to note that all preservation methods including trehalose (Trehalose, Sugar Mix) performed well, and exhibit concentration-dependent behavior as illustrated by Fig. 6.3b: the greater the concentration of trehalose, the slower the DNA degrades. Note that DNAStable and GenTegra are comprised of proprietary mixtures and the presence or absence of trehalose is unknown.

6.3.3 Sequence Analysis of DNA Material

The samples aged at UW were the only samples to be sequenced. All samples were prepared for sequencing prior to aging so that no perturbation of the library (e.g., PCR, ligation steps) were necessary after aging. Each sample had a unique index (i.e., tag) that allowed the identical sequences to be sequenced in the same Illumina NextSeq sequencing run to minimize quality score variation.

First, we explored the difference in observed error rates. We compared the rates of insertions, deletions, and substitutions between all preservation methods at different temperature and time points and between the two files used. As shown in Fig. 6.4, we found minimal

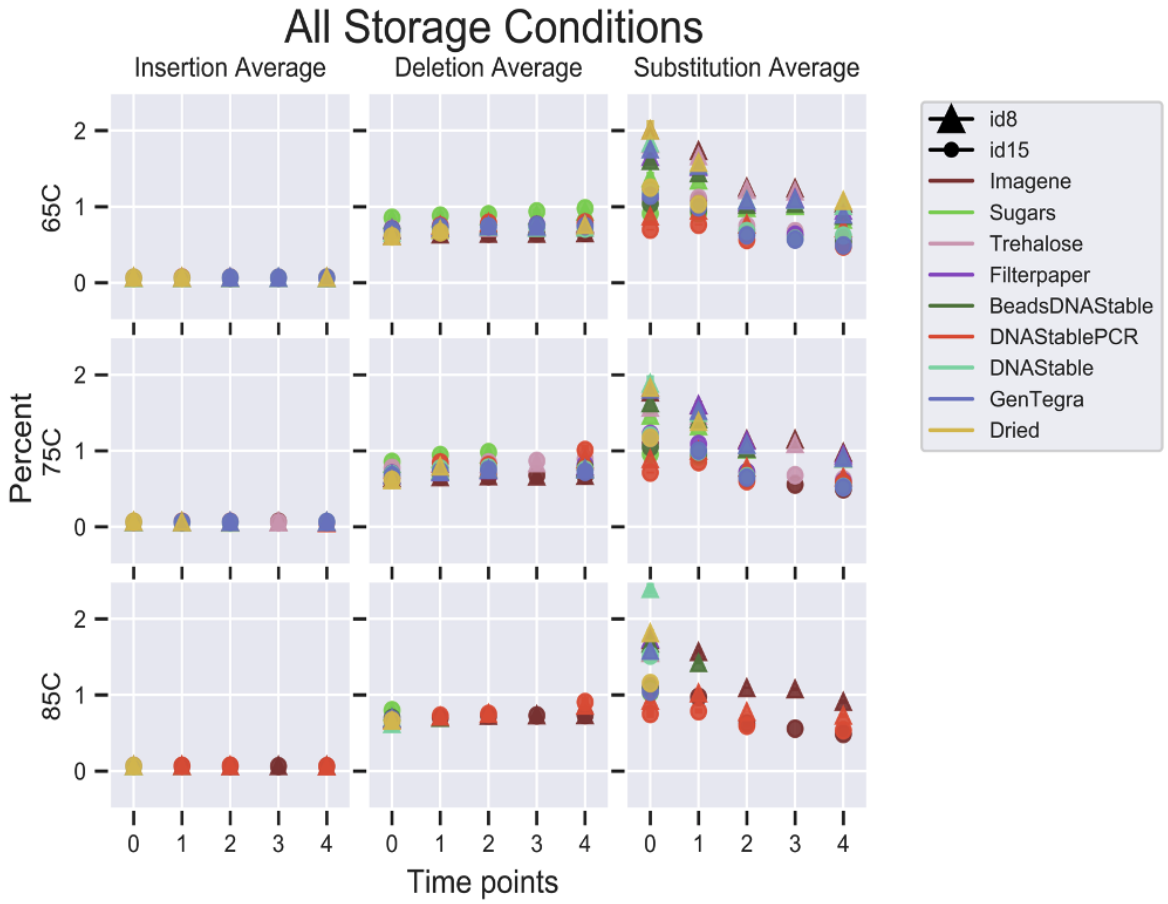


Figure 6.4: The error bars are the standard deviation of the method's triplicates for each time point sequenced. Note that ETH-Z samples were not sequenced. When a certain time point is not plotted, there was either no data for that time point or the file's average coverage is under the lowest tolerated threshold of 14x (see Appendix D.2 for more information). It is important to note that the errors reported here are the combined errors of both DNA synthesis and sequencing, however all samples underwent the same sequencing process in the same sequencing machine so all samples are expected to have similar sequencing error rates.

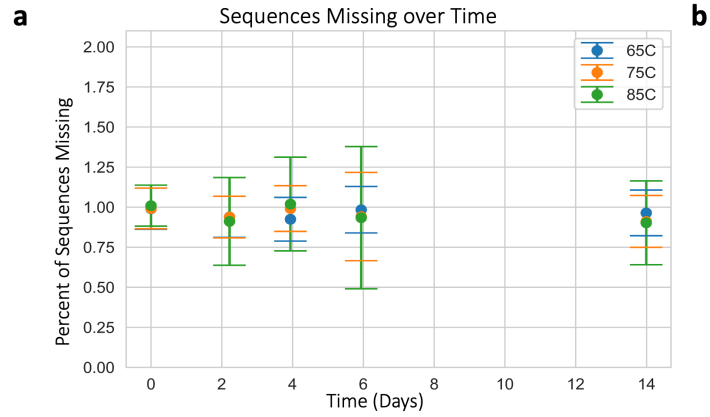
variation between the storage methods, with no practical difference between methods and/or files for insertion and deletion errors. Even when looking intently at the substitution rate, which has the most variation, we still observed a maximum difference of just over one percent, which is not practically significant.

More importantly, there is no preservation method that consistently had more or less errors than the others except for the method that employed PCR prior to sequencing (we hypothesize that PCR is, in effect, selecting the most intact and thus less error-prone strands for replication). In fact, as shown at time point 0, there is already a significant amount of variation prior to any aging, and furthermore those initial orderings of methods' errors do not predict the subsequent method rankings. There was no particular storage method(s) that showed more or fewer errors than other methods across the different temperatures and time points, which suggests that insertion, deletion, and substitution errors are independent from the storage method (see Appendix D.3 for more analysis).

We next explored the relationship between missing sequences in the files between the aged triplicates at different time points to determine if there was sequence-dependent degradation. We examined both the total number of sequences missing from sequencing at each time point as well as which individual sequences were missing. If the total number of sequences missing increased after the pre-aging time point 0, we could hypothesize that there was some sequence-dependent degradation as more-vulnerable sequences degraded. However, we observed no difference in the number of sequences missing across all time points (Fig. 6.5a). This suggests that sequence loss is stochastic across all storage methods.

This is further supported by our analysis of each individual sequence missing. If a sequence was missing from one time point due to the nature of the sequence being prone to degradation over time, we would not expect to see that sequence again in all subsequent sequencing runs, or in any of the triplicates. However, when we observe a missing sequence, it is often sequenced at a later time point, shown in Fig. 6.5b (as detailed in Appendix D.4.2). We hypothesize this is due to the stochastic nature of subsampling the sample for sequencing, rather than systematic bias against particular sequences.

That sequence loss is stochastic is yet further supported by results from a trimer and GC content analysis, in which we looked at the prevalence of each trimer (ACA, AGA, ATA, etc.) and the proportion of the sequence that is comprised of guanine (G) and cytosine (C). We found no significant difference between trimer or GC compositions of sequences



Temperature and Time Point	Total Percent Missing	Percent Missing that Reappear	Percent that Stay Missing
65C: TP0	0.616	0.0	0.616
TP1	0.625	99.259	0.005
TP2	0.694	99.333	0.005
TP3	0.588	99.213	0.005
TP4	0.593	96.875	0.019
75C: TP0	0.579	0.0	0.579
TP1	0.593	96.094	0.023
TP2	0.63	94.853	0.032
TP3	0.579	94.4	0.032
TP4	0.602	90.769	0.056
85C: TP0	0.588	0.0	0.588
TP1	nan	nan	nan
TP2	nan	nan	nan
TP3	nan	nan	nan
TP4	nan	nan	nan

Figure 6.5: (a) The mean number of DNA sequences missing from sequencing data over time with error bars representing the standard error of the mean. (b) An example of sequence loss behavior. The "Total Percent Missing" is the mean percentage of the sequences not recovered from each time point and temperature for File 8. Of the sequences missing from each time point and temperature, a non-zero number of sequences typically reappear in subsequent time points, and that percentage is given in "Percent Missing that Reappear". "Percent that Stay Missing" reports the percentage of total sequences in the file that stay missing. Data for all conditions and files can be found in the Supplemental Files in the published work [72] ending with "missing sequences analysis.csv". (c) A portion of the trimer and GC content analysis giving the percent of samples that had a statistically significant difference in trimer composition between the top 5% most present sequences and the bottom 5% sequences, including missing sequences. For all trimers data for both files, see Appendix D.3.

missing/lowly present and sequences highly present, as shown in Fig. 6.5c. More details can be found in Appendix D.3.

6.3.4 Density

Each preservation method in our study improved the half-life of DNA when compared to dehydrated DNA stored with no additives, but at a cost to the physical DNA density of the sample and, at times, simplicity of DNA recovery. Though the physical overhead (i.e., molecules added to the sample to preserve DNA) of each technique and the time required to extract the DNA may not be significant for many DNA applications, high physical overhead is less appealing for data storage where DNA sequences per volume translates to bytes per volume, and extraction complexity of the DNA limits the rate at which files can be processed. That is to say, the more DNA that can be stored in a given storage container and the faster the samples can be processed, the higher the DNA data density of that container and the higher the throughput.

The highest DNA density was ETH-Z Magnetic NP preservation technique at 3.4 wt% of DNA to encapsulant, though it required the most complex protocol with roughly a 10-20 minute DNA release time [35]. For methods that involved mixing or simply drying the DNA files with preservatives (Trehalose/Sugars/GenTegra/DNAStable, and Filter Paper), release times are under a minute; however, the DNA loading in the Trehalose and Sugar mix drastically reduced to 0.13 wt% and 0.095 wt%, respectively. Due to the proprietary formulations of GenTegra and DNAStable, we could not accurately calculate the DNA loading but estimate the mix to be similar to the former. DNA stored by mixing or drying the DNA with preservatives degrade significantly faster than the DNA stored in Imagene DNAshells. DNAshells presently have the highest physical overhead due to the borosilicate glass insert and stainless steel shell and cap, though the physical overhead could be reduced as the company develops smaller tubes.

6.4 Methods

6.4.1 DNA Material Stored

It is important to note there are slight differences between the storage protocols used by UW and ETH. The protocol used by UW is shown in Fig. 6.1. 28,974 unique sequences of dsDNA 150nt in length were prepared for NextGen sequencing for a final length of 310 base pairs. Storage experiments performed by ETH contained only 7,373 of the DNA sequences (File 15) and did not prepare them for NextGen sequencing, so the length of DNA stored in their ovens was 150nt.

6.4.2 Sample preparation

From 150mer DNA pools synthesized by Twist Biosciences, two files were amplified. File 8 was comprised of 21,601 unique DNA sequences while File 15 was comprised of 7,373 unique DNA sequences. Within each file, the primer sequences (first and last twenty bases of the 150mer) are conserved and the middle 110 bases are distributed such that there are no homopolymers and the GC content is approximately 50%.

Each of the files was PCR amplified in multiple rounds to minimize PCR bias (when there is an uneven distribution of each DNA sequence due to the stochastic nature of PCR) using the following recipe: 5 μL of 1 ng/μL of DNA template, 2.5 μL of each primer at 10μM, 50 μL of Kappa HiFi 2x, and 40 μL of molecular grade water. The PCR protocol was then: (1) 95°C for 3 min, (2) 98°C for 20 sec, (3) 62°C for 20 sec, (4) 72°C for 15 sec, (5) repeat steps 2-4 a total of 13 or 20 times for files 8 and 15 respectively,(6) 72°C for 30 sec.

The samples were then PCR amplified following the same protocol, but with forward primes now having a 25N overhang to allow for sequencing (this is an artifact of the Illumina NextSeq, as largely uniform sequences at the beginning of sequencing do not allow for accurate cluster calling and so the first 25 bases being completely random solves this problem).

The samples were then quantified with qPCR, and then pooled together so that File 8

had approximately twice as many copies of each sequence as File 15.

This pool of File 8 and File 15 sequences was then split into 96 identical alequots and prepared for aging and sequencing with ligation. (Note that 9 of these alequots that were to become the basis for DNAStable + PCR also had 323,875 extra sequences added to its sample to make the PCR conditions slightly more complex, as would be more realistic in a typical setting, though the number of copies of each sequence remained the same as the other samples.) Ligation was done with a modified version of Illumina TruSeq Nano ligation protocol and TruSeq ChIP Sample Preparation protocol. Step by step instructions are in Appendix D.5 for convenience, but briefly, samples were first converted to blunt ends with the ERP2 reagent and directions provided in the Illumina TruSeq Nano kit, then purified with AMPure XP beads according to the TruSeq ChIP protocol. An 'A' nucleotide was added to the 3' ends of the blunt DNA fragments with the TruSeq Nano's A-tailing ligase and protocol, followed by ligation to the Illumina sequencing adapters with the TruSeq Nano reagents and protocol. We then cleaned the samples with Illumina sample purification beads and enriched the sample using an 8-cycle PCR protocol given in the TruSeq Nano protocol.

For the enrichment, all samples (each now with a unique ligation index) were enriched using the following recipe: 3μL of a ligation sample, 3μL of the PCR Primer Cocktail provided in the TruSeq Nano kit, 12μL of Enhanced PCR Mix provided in the TruSeq Nano kit, and 12μL of molecular grade water. The following PCR protocol was used: (1) 95°C for 3 min, (2) 98°C for 20 sec, (3) 60°C for 15 sec, (4) 72°C for 30 sec, (5) repeat steps 2–4 for a total of eight times. The length of enriched products was confirmed using a Qiaxcel bioanalyzer. Reformatted instructions are given in Appendix D.5 for convenience.

After enrichment, the PCR products with the same ligation index were pooled together and PCR purified with a QIAquick PCR purification kit with molecular grade water at the elution step.

Each sample was then quantified with a Qubit 2.0 fluorometer and each preservation method took 350 ng of material per sample. The details of each preservation method are below.

6.4.3 No Additives

From samples described in the "Sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five alequots of 350ng were made and placed in 0.6mL eppindorf tubes (one alequot for each time point). Each tube was then dehydrated in a SPD 1030 speedvac on high.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer. The samples then sat at room temperature for 20 minutes prior to quantification with qPCR (see Methods section, "Quantifying Degradation with qPCR").

6.4.4 DNAStable

From samples described in the "Sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five alequots of 350ng, with a volume ranging from 3.4-5μL, were made and placed in 0.6mL eppindorf tubes (one alequot for each time point). Each tube then had 20μL of DNAStable LD (product number 53001-066) added, and the mixture was pipetted up and down three times. Each tube was then dehydrated in a SPD 1030 speedvac on high.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer. The samples then sat at room temperature for 20 minutes prior to quantification with qPCR (see Methods section, "Quantifying Degradation with qPCR").

6.4.5 DNAStable PCR

The preservation and rehydration protocol is identical to DNAStable.

For DNAStable PCR, the samples were analyzed with qPCR in an identical manner to

the other methods, then File 8 and File 15 were accessed with PCR (see Appendix D.1.1 for primer sequences), re-ligated with the same protocol described above (see Appendix D.1.2), and sequenced using the same protocol as all other methods. The amplification protocol for File 8 and File 15 was as follows:

First, amplify each sample using the primers listed in Appendix D.1.1. For time point 0, 0.2μL of sample was used and 10 cycles were performed. For time point 1, 8μL of sample was used and 12 cycles were performed for the 65°C samples, while 16 cycles were performed for the 75°C and 85°C samples. In addition to the sample DNA, there were 10μL of 2x Kapa HiFi, 1μL of each primer at 10μM, and 7μL of molecular grade water. The thermocycle protocol was: (1) 95°C for 3 min, (2) 98°C for 20 sec, (3) 62°C for 20 sec, (4)72°C for 15 sec, (5) repeat steps2-4 for a total of the number of times stated at the beginning of this paragraph.

Then, amplify the resulting sample again except that now there is a 25Nmer (25 random nucleotides) on the 5' end of the forward primer. Follow the same protocol as above, but now there are always 12 cycles and 1μL of DNA product.

Then, rather than performing ligation in preparation for sequencing, perform one last PCR with appropriate overhanging sequences. Follow the same PCR protocol as the previous paragraph, but this time using primers containing each files' primer sequence, as well as the relevant index sequence and Illumina adapter and sequencing primer, as listed in Appendix D.1.

6.4.6 Mag-Bind DNAStable

From samples described in the "sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five aliquots of 350 ng, with volumes ranging 20-30μL, were made and placed in 0.6mL eppendorf tubes (one aliquot for each time point). Each tube than had 1.2X ratio of Mag-Bind beads to DNA volume added, and the mixture was vortexed and allowed to sit at room temperature for 5 minutes. After magnetic bead separation, the supernatant was discarded.

Then the samples were washed with 200 uL of 70% ethanol, allowed to sit for 1 minute before separating the beads and discarding the supernatant (performed 2x). Any residual ethanol was air dried for 10 minutes. Each tube then had 20μL of DNAStable LD (product number 53001-066) added, and the mixture was vortexed to mix. Each tube was then dehydrated in a SPD 1030 speedvac on high.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer. The samples then sat at room temperature for 20 minutes before magnetic bead separation. The supernatant was then used for qPCR quantification (see Methods section, "Quantifying Degradation with qPCR").

6.4.7 Sugar Mix

From samples described in the "sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five aliquots of 350 ng, with volumes ranging 7-9μL, were made and placed in 0.6mL eppendorf tubes (one aliquot for each time point). Each tube then had equal volumeμL of 0.2M Sugar Mix (0.1 M trehalose, 0.05 M raffinose, 0.05 M mannitol, 0.125 mM uric acid) added and the mixture was pipetted up and down three times. Each tube was then dehydrated in a SPD 1030 speedvac on high.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer. The samples then sat at room temperature for 20 minutes prior to quantification with qPCR (see Methods section, "Quantifying Degradation with qPCR").

6.4.8 Trehalose

From samples described in the "sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five aliquots of 350 ng, with volumes ranging 7-9μL, were made and placed in 0.6mL eppendorf tubes (one aliquot for each time point). Each tube then had equal volumeμL of 0.2M Trehalose added, and the mixture was pipetted up and down three times. Each tube was then dehydrated in a SPD 1030 speedvac on high.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer. The samples then sat at room temperature for 20 minutes prior to quantification with qPCR (see Methods section, "Quantifying Degradation with qPCR").

6.4.9 GenTegra

From samples described in the "Sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five alequots of 350ng, with volumes ranging 6-15μL, were made and placed in GenTegra tubes (one alequot for each time point). This mixture was then pipetted up and down 10 times as per GenTegra's instruction. The GenTegra tubes were 0.5mL with GenTegra's proprietary mixture at the bottom of each tube (product number GTD2100-S). Each tube was then dehydrated in a SPD 1030 speedvac on high.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer. The samples then sat at room temperature for 20 minutes prior to quantification with qPCR (see Methods section, "Quantifying Degradation with qPCR").

6.4.10 Filter Paper

From samples described in the "Sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. From each ligation index, five alequots of 350ng, with volumes ranging 6-10μL, were pipetted 2μL at a time onto 2.5mm diameter VWR Grade 415 filter paper (product number 28320-121) and dried at 30°C between rounds of pipetting. The final, dry 2.5 diameter circles of filter paper were then individually placed in 0.6mL eppindorf tubes.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature, then rehydrated with 50μL of molecular grade water, capped, and vortexed on a benchtop vortexer with the filter paper still in the tube. The samples then sat at room temperature for 20 minutes prior to quantification with qPCR (see Methods section, "Quantifying Degradation with qPCR").

6.4.11 Imagene

From samples described in the "Sample preparation" section, nine samples, each with a unique ligation index, were used for this preservation method. Each sample was then shipped overnight on dry ice to Imagene with two extra samples prepared the same as the others, but only used as shipping controls. Imagene then deposited 350 ng of the shipped DNA material into each DNAshell. The samples were then shipped back to the UW with the shipping controls. The samples were stored at -20°C until the aging process started.

After aging, nine samples (one alequot from each index, in other words, three indices per temperature) were removed from their respective heat sinks and allowed to come to room temperature. The metal DNAshells were then pierced with the included shell piercer, and then rehydrated with 50μL of molecular grade water and pipetted up and down 10 times. The samples then sat at room temperature for 20 minutes and were transferred out of DNAshells to PCR tubes prior to quantification with qPCR (see Methods section,

"Quantifying Degradation with qPCR").

6.4.12 No Additives - Performed by ETH-Z

A sample of F15 and primers were shipped to ETH-Z. F15 was PCR amplified and purified using QIAquick PCR purification Kit (from Qiagen). Following elution with MiliQ water, the final concentration of DNA was 12 ng/uL.

Five aliquots of 24 ng (2μL), were made and placed in 2 mL eppendorf tubes (one aliquot for each time point). Each tube was then dehydrated in a vacuum centrifuge for 2 hours at 60°C and aged in a desiccator containing a saturated NaBr solution within an oven. After aging, samples were removed from the oven and allowed to come to room temperature. They were then rehydrated and quantified with qPCR.

6.4.13 Trehalose - Performed by ETH-Z

Five aliquots of 24 ng (2μL), were made and placed in 2 mL eppendorf tubes (one aliquot for each time point). Each tube then had 4 uL of 0.02M Trehalose solution added and was mixed. Each tube was then dehydrated in a vacuum centrifuge for 2 hours at 60°C and aged in a desiccator containing a saturated NaBr solution within an oven. After aging, samples were removed from the oven and allowed to come to room temperature. They were then rehydrated and quantified with qPCR.

6.4.14 DNAStable - Performed by ETH-Z

Five aliquots of 24 ng (2μL), were made and placed in 1.5 mL eppendorf tubes pre-coated with DNAStable (one aliquot for each time point). Each tube was then dehydrated in a vacuum centrifuge for 2 hours at 60°C and aged in a desiccator containing a saturated NaBr solution within an oven. After aging, samples were removed from the oven and allowed to come to room temperature. They were then rehydrated and quantified with qPCR.

6.4.15 Magnetic Nanoparticles - Performed by ETH-Z

Samples were encapsulated and deencapsulated in magnetic nanoparticles following the protocol reported in Chen et al.

Five aliquots of 40μL final particle solution were made and placed in 1.5 mL eppendorf tubes (one aliquot for each time point). Each tube was then dehydrated in a vacuum centrifuge overnight at 30°C and aged in a desiccator containing a saturated NaBr solution within an oven. After aging, samples were removed from the oven and allowed to come to room temperature.

6.4.16 Aging DNA

After preservation, all samples except time point 0 samples were placed in their respective ovens concurrently and kept uncapped at 50% relative humidity (except for the sealed Imagene DNAshell capsules).Relative humidity was controlled through a well-established technique using saturated salt solutions [81]. Sodium bromide was dissolved in water until it was no longer dissolvable with percipitate to create a saturated solution of sodium bromide. A deep 1L petri dish was filled with the saturated salt solution and placed on the bottom oven rack to maintain 50% RH atmosphere at the three different temperatures.

The ovens used were Quincy Lab, Inc. Model 10E-LT Lab Oven, with form-fitted insulation over the body of the oven above the control system made of polyisocyanurate rigid foam insulation board that was one inch thick to prevent condensation.

An aluminum heat sink (a one inch thick block of aluminum) with holes the width and depth of each tube was milled and placed in each of the three ovens.

In each oven, a thermometer was placed through the vent and rested on the heat sink and monitored daily for temperature fluctuations greater than 2 degrees, which were not observed during the experiment except immediately after removing the heat sinks at each time point.

6.4.17 Quantifying Degradation with qPCR

After each aging time point, samples were diluted 1:100 with molecular water and the amount of full-length product present was quantified with qPCR. The standard used was an ultramer ordered from IDT, whose sequence is given in Appendix D.1. The qPCR recipe is as follows: 1μL DNA sample, 0.5μL of each post-ligation primer (given in Appendix D.1) at 10μM, 10μL Kapa HiFi, 7μL molecular water, 1μL 20x Eva Green. The qPCR thermocycling protocol was: (1) 95°C for 30 sec, (2) 98°C for 20 sec, (3) 60°C for 20 sec, (4) 72°C for 30 sec, (5) repeat steps 2-4 39 times. A negative control was included on each plate that had no DNA sample included with the mixture of all other reagents.

6.4.18 Calculating Half-Life

We use the standard Arrhenius equation for first-order decay, as detailed in Appendix D.6, and this is implemented in the script entitled *Halflife_Calc.py* available on the GitHub repository `https://github.com/uwmisl/aging_data.git`.

6.5 Discussion

DNA samples aged in preservation methods without exposure to any water or air (Imagene) were found to decay much slower than samples exposed to water and air, which supports previous findings [33, 73, 76, 82]. Samples preserved with trehalose exhibited concentration dependence, with the higher concentration of trehalose conferring more protection against degradation (Fig. 6.3b), and the amount of DNA stored per sample and the ratio of DNA to storage molecules may also play an important role on the rate of degradation, depending on the preservation method used [36, 80].

Depending on the storage method, there are sometimes significant changes in the DNA encapsulation and retrieval process. Imagene DNAShells (Imagene) and magnetic nanoparticles (ETH-Z Magnetic NP) were the most complex storage methods to retrieve DNA from, as they either required specialized equipment or involved several steps. In contrast, DNA

combined with preservatives and simply dehydrated have the simplest recovery procedures for they only require a simple re-hydration protocol. This leads to an inverse relationship between stability and process complexity. Surprisingly, DNA stored in the presence of trehalose or GenTegra did not fit this storage complexity trend and instead degraded much slower than anticipated.

There was a real concern that half-life values would not be the only difference between storage methods, and that the sequences would accumulate insertion, deletion, or substitution errors at varying rates. However, we observed that error rates did not differ with any practical significance between the preservation methods. Furthermore, within the tight constraints of the sequences examined here (i.e., no homopolymers in the payload region, balanced GC content), sequence degradation was found to be stochastic. This is encouraging to the field of DNA data storage, where stochastic errors and sequence erasures are already dealt with easily with various means of error correction such as Reed-Solomon codes, as presented in Chapter 3 and other work [25, 73, 83].

Automation will very likely be the key in obtaining a scalable DNA data storage system. It is foreseeable that a solution containing DNA files and a preservative could be mixed together and dried on a thin film for storage. When the file needs to be accessed, it would be re-hydrated for a short time (likely on the order of a few seconds) and then processed for sequencing. Several early examples of this already exist on a glass top plate using a digital microfluidic system [80, 84] but could extend to other liquid dispensing systems such as acoustic or ink-jet dispensing. Development of an automated DNA storage and retrieval system would require thorough investigation of the effects of sampling and preserving the same sample multiple times, as studies in genomic DNA have shown this causes a non-negligible effect on recovery and may be preservation material dependent [85].

For a more comprehensive guide to DNA degradation, more long-term studies should be conducted in which DNA is allowed to degrade for years at a time, rather than days or a few weeks as most aging experiments have done, and at concentrations and temperatures more reflective of storage conditions users anticipate. This is because DNA decay depends

on storage material, temperature, and other factors [77, 86].

Yet it is important to note that two different labs using different experimental aging setups and different length sequences produced the same relative ranking of those methods, though the half-life values did differ. This underscores the fact that exact half-life values are difficult to extrapolate and very sensitive to the conditions used to achieve them. We encourage a more robust examination of all these variables depending on a user's end goals and likely storage conditions.

Users of future DNA data storage systems will require a broad range of stability, from hundreds to thousands of years, but the error analysis results presented here demonstrate the interchangeable nature of all the methods examined regarding sequencing data quality. As this work shows, users must choose their preferred storage method based on their desired half-life, data density, and process complexity.

Chapter 7

CONCLUSION

In this dissertation, we have addressed some of the practical concerns of scaling and implementing DNA data storage, including concerns of efficient information retrieval at scale, physical redundancy, durability, and the implementation of two different information retrieval algorithms for broader DNA data storage use cases. Based on these findings, we are encouraged that DNA data storage may one day be a viable alternative to current tape-based storage systems, and we see a number of future explorations that could further elucidate DNA data storage implementations as well as a number of things to consider as this technology continues to develop.

7.1 Potential Future Directions

In Chapter 3 we examined PCR-based random access and showed that we can efficiently retrieve a file of interest with negligible background noise and a relatively simple workflow. However, there are many instances when one might wish to access more than one file at a time (this is known as retrieving files in multiplex). As mentioned in this chapter, PCR is a poor tool for multiplex retrieval because it accesses files with wildly different efficiency, leading to inefficient sequencing. Recent work from the Molecular Information Systems Lab has explored multiplex retrieval with much more success using CRISPR-Cas9 [87], which also solves another potential problem that PCR-based access presents, and was briefly touched on in Chapter 5. The problem with PCR at large scale is that the energy required to take solutions up to near boiling temperatures and then rapidly cool and heat the solutions over several cycles is untenable. This may be an acceptable cost if we assume that information in DNA will only very rarely, if ever, be accessed, but we can safely assume many users would

not tolerate this. As discussed in Chapter 5, CRISPR-Cas9 is an attractive alternative to PCR not only for its multiplex ability, but for its energy efficiency. With comparable lab protocol complexity, we can now eliminate energy intensive thermocycling that lasts on the order of an hour and instead incubate a solution at near room temperature on the order of seconds to minutes.

While this examination of Chapter 3 may leave readers dubious of the future of PCR-based random access, it is important to note that this work provides an excellent scaffold on which to analyze potential future DNA-based information retrieval strategies, and still provides us with a crucial understanding of PCR-based random access while we continue to examine other aspects of DNA data storage at a scale amenable to PCR. In addition, the work presented in Chapter 3 also functions as a benchmark to which potential future information retrieval methods can be compared.

If we think of using our work on PCR random access as a benchmark, then the work presented in Chapter 4 is absolutely crucial to understanding the information density of DNA data storage. While Chapter 4 determined the minimum physical redundancy needed to perform PCR random access, it also examined much more complex conditions than those explored in Chapter 3. However, the limit of those complex conditions (how many sequences can be present in solution before PCR random access falters or fails) was not found. The most obvious extension of this work is to reach and define these limits. Whether the limit of PCR random access is 10 or 10,000 nanograms of material per reaction would drastically change how DNA data storage is implemented at commercial scale, and thoroughly investigating these limits would serve as another useful benchmark for any future information retrieval method. A perfectly energy efficient retrieval method would be completely unusable if it required high physical redundancy and/or only performed well in conditions with few other DNA sequences, and having PCR as a benchmark will likely be a useful comparison for any technology to come.

Moving past random access, in Chapter 5 we examined using Cas9 for similarity search. While in theory Cas9 was an intriguing molecular tool for similarity search because of the

potential to harness off-target cleavage activity for similarity search, we were not able to use Cas9 as a highly effective similarity search tool due to its high false-positive recall rate. However, the energy efficiency and speed at which this molecular computation occurs (on the order of seconds) is an important consideration, and future work could very well expand on this preliminary experiment and attempt more sophisticated Cas9 modeling and encoding to improve performance. And in a broader context, Cas systems other than just Cas9 should be more readily considered in the field of molecular programming. Cas9 random access and similarity search are just two molecular programming applications that utilize Cas9, but we encourage future molecular programmers to consider the full range of highly programmable Cas systems as the field continues to leverage biological tools in creative ways.

And finally, in Chapter 6 we examined various methods of preserving synthetic DNA and evaluated their durability. To formulate an end-to-end pipeline for DNA storage, choosing a preservation method is one of the most critical steps. From this work, DNA data storage researchers and commercial developers can now begin to design potential DNA data storage workflows in detail. By knowing storage conditions and the half-life of DNA in those conditions, we can back-calculate from our expected storage duration exactly how much DNA must be stored initially, which in turn informs users how to organize the information stored in DNA. Everything from the size of pools to the amount of space DNA data requires can now utilize the work presented in this chapter, regardless of the information retrieval method used.

7.2 Practical and Ethical Implications of DNA Data Storage

If DNA data storage is indeed poised to replace tape or other traditional data storage media, we must think critically about the implications of this technology. Broadly, these considerations can be categorized into biological considerations and societal considerations.

7.2.1 Biological Considerations

Magnetic tape and other traditional storage media have an impact on our environment at every stage of their workflow, from mining silicon to taking up vast amounts of land storing the data to the e-waste after the medium has degraded to the point where it is no longer usable. Yet this is the same consideration we must make for any information storage medium. If we are to replace silicon with DNA, however, we must do so conscientiously, for DNA has biological significance. Theoretically, DNA used for data storage could be inadvertently incorporated into the genome of an organism, where it could then be translated into some sort of harmful protein. Thankfully, the odds of this happening are astronomically low by my calculation, and I will briefly explain why here.

First, an organism would have to come into contact with the pool of DNA. This will undoubtedly happen unless all samples are kept in a clean room, however, it is unlikely that DNA will be directly exposed to the environment. It is far more likely that DNA will be coated in a substrate such as a silicon bead or a mixture of sugars. It is likely the preservation method will act as a barrier to any biological uptake.

Second, even if uptake were to occur, the DNA must then mean something *and* get inserted to a portion of the genome that is translated into some kind of protein. The odds of this are very, very slim, as not all DNA gets translated into protein, and furthermore not all DNA is kept after it has been incorporated.

Third, even if DNA were translated into protein, proteins are usually formed by 50-1,000 amino acid residues, and each amino acid residue is encoded by three DNA bases [88]. All DNA data storage sequences presented in this work are less than 200 bases, which would allow for a maximum of only 66 amino acid residues.

Fourth, again assuming the unlikely event of DNA being translated into protein, DNA is translated into proteins using START and STOP sequences (called codons) that signal the start and stop of protein translation. Each of these sequences is three bases long, and in general there are three STOP codon sequences: TAG, TAA and TGA. The odds of any one

of those sequences appearing in a "random" 200nt sequence ("random" from the perspective of biology, not random at all to data scientists), as in our work, is nearly guaranteed (see Appendix E for the calculation). With every strand likely containing multiple STOP codon(s) to stop translation, the risk of a biologically meaningful protein being expressed becomes reduced even further.

Fifth, again assuming the unlikely event of DNA being translated into protein and assuming a STOP codon does not truncate the resulting protein sufficiently or is not present, this would very likely not have a significant impact on anything but the organism's personal health because DNA sequences used for DNA data storage are random with respect to biology, and therefore largely nonsense, from the perspective of biology. The odds of a terrible biological accident scenario happening are vanishingly slim, though should be continually considered as the technology evolves.

7.2.2 Societal Considerations

A much more relevant consideration is the way we preserve information and how it impacts our society. While similar to the data preservation considerations we give to traditional storage media, using DNA as a storage medium lasts so much longer that I believe there are two main considerations to make.

The first is how to preserve our privacy; we must know how to selectively destroy data once someone has asked for it to be destroyed. This could be done as simply as removing the DNA and dousing it in bleach or by using some other destructive method, but may get more complicated depending on how data is co-localized.

The second, more complex, question we must grapple with is what the societal implications of storing data for so long are. It is entirely possible that one day DNA data storage will be used for its superior information density, not its remarkable durability. However, the field of DNA data storage currently values information longevity, and by developing DNA data storage technology we may fall into the futile trap of trying to preserve all information created. By solving the "problem" of running out of space to store our data, we will likely

create a myriad of unintended dilemmas society must grapple with. I suspect that these dilemmas will not only be technical (i.e., how to organize so much data), but also philosophical (i.e., why do we feel the need to keep all this data? Are there types of data we do not want preserved for so long? Could preserving data for so long be harmful in some circumstances? Could it be good?).

So while it is highly unlikely DNA data storage poses any immediate threat to the biological world or society at large, these are implications best considered sooner rather than later.

BIBLIOGRAPHY

[1] Alex Woodie. "Big Growth Forecasted for Big Data". In: *Datanami* (Jan. 12, 2022). https://www.datanami.com/2022/01/11/big-growth-forecasted-for-big-data/ (Cited on page 1).

[2] Rich Miller. "Facebook Builds Exabyte Data Centers for Cold Storage". In: *Data Center Knowledge* (Jan. 18, 2013). https://www.datacenterknowledge.com/archives/2013/01/18/facebook-builds-new-data-centers-for-cold-storage (Cited on page 1).

[3] Osamu Shimizu and Hitoshi Noguchi. "Tape storage for cold data archive and future technologies". In: *SNAI Data Storage Innovation Conference* (2015). https://www.snia.org/sites/default/orig/DSI2015/presentations/ColdStorage/OasamuShimizu_Tape_storage_for_cold_data_archive.pdf (Cited on page 1).

[4] Emily Leproust. "DNA Synthesis – An Integral Force in the Founding and Future of Precision Medicine". In: *The Journal of Precision Medicine* (2021). https://www.thejournalofprecisionmedicine.com/wp-content/uploads/dna-synthesis.pdf (Cited on page 2).

[5] Tom van der Valk, Patrícia Pečnerová, David Díez-del-Molino, Anders Bergström, Jonas Oppenheimer, Stefanie Hartmann, Georgios Xenikoudakis, Jessica A. Thomas, Marianne Dehasque, Ekin Sağlıcan, Fatma Rabia Fidan, Ian Barnes, Shanlin Liu, Mehmet Somel, Peter D. Heintzman, Pavel Nikolskiy, Beth Shapiro, Pontus Skoglund, Michael Hofreiter, Adrian M. Lister, Anders Götherström, and Love Dalén. "Million-year-old DNA sheds light on the genomic history of mammoths". In: *Nature* (Mar. 2021). ISSN: 1476-4687. DOI: 10.1038/s41586-021-03224-9 (Cited on page 2).

[6] Victor Zhirnov, Reza M. Zadegan, Gurtej S. Sandhu, George M. Church, and William L. Hughes. "Nucleic acid memory". In: *Nature Materials* (Mar. 23, 2016). ISSN: 1476-4660. DOI: 10.1038/nmat4594 (Cited on pages 2, 55).

[7] Lee Organick, Yuan-Jyue Chen, Siena Dumas Ang, Randolph Lopez, Xiaomeng Liu, Karin Strauss, and Luis Ceze. "Probing the physical limits of reliable DNA data retrieval". In: *Nature Communications* (Jan. 30, 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-14319-8 (Cited on pages 2, 40, 69).

[8] Claire Vieille and Gregory J. Zeikus. "Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability". In: *Microbiology and Molecular Biology Reviews* (Mar. 2001). ISSN: 1092-2172. DOI: 10.1128/MMBR.65.1.1-43.2001 (Cited on page 5).

[9] John M. S. Bartlett and David Stirling. "A Short History of the Polymerase Chain Reaction". In: *PCR Protocols*. Ed. by John M. S. Bartlett and David Stirling. Methods in Molecular Biology™. Totowa, NJ: Humana Press, 2003. ISBN: 978-1-59259-384-2. DOI: 10.1385/1-59259-384-4:3 (Cited on page 5).

[10] Randall K. Saiki, Stephen Scharf, Fred Faloona, Kary B. Mullis, Glenn T. Horn, Henry A. Erlich, and Norman Arnheim. "Enzymatic Amplification of -Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia". In: *Science* (Dec. 20, 1985). DOI: 10.1126/science.2999980 (Cited on page 6).

[11] Rob Carlson. *On DNA and Transistors*. http://www.synthesis.cc/synthesis/2016/03/on_dna_and_transistors. Mar. 29, 2016 (Cited on pages 6, 38).

[12] Norbert Wiener. "Interview: Machines Smarter than Men?" In: *News World Rep.* (1964) (Cited on page 7).

[13] M. S. Neiman. “On the molecular memory systems and the directed mutations.” In: *Radiotekhnika* (1965) (Cited on pages 7, 20).

[14] Joe Davis. “Microvenus”. In: *Art Journal* (Mar. 1, 1996). ISSN: 0004-3249. DOI: 10.1080/00043249.1996.10791743 (Cited on page 7).

[15] Catherine Taylor Clelland, Viviana Risca, and Carter Bancroft. “Hiding messages in DNA microdots”. In: *Nature* (June 1999). ISSN: 1476-4687. DOI: 10.1038/21092 (Cited on page 7).

[16] George M Church, Yuan Gao, and Sriram Kosuri. “Next-Generation Digital Information Storage in DNA”. In: *Science* (2012). ISSN: 0036-8075. DOI: 10.1126/science.1226355 (Cited on pages 7, 20, 41, 45, 54, 55, 113).

[17] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M Leproust, Botond Sipos, and Ewan Birney. “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA”. In: *Nature* (2013). ISSN: 0028-0836. DOI: 10.1038/nature11875 (Cited on pages 7, 20, 26, 41, 55, 113).

[18] Robert N. Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J. Stark. “Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes”. In: *Angewandte Chemie International Edition* (2015). ISSN: 1521-3773. DOI: 10.1002/anie.201411378 (Cited on pages 7, 11, 20, 26, 113).

[19] Yaniv Erlich and Dina Zielinski. “DNA Fountain enables a robust and efficient storage architecture”. In: *Science* (2017). ISSN: 0036-8075. DOI: 10.1126/science.aaj2038 (Cited on pages 7, 20, 28, 41, 45, 54, 67, 69, 113).

[20] S. M. Hossein Tabatabaei Yazdi, Yongbo Yuan, Jian Ma, Huimin Zhao, and Olgica Milenkovic. “A Rewritable, Random-Access DNA-Based Storage System”. In: *Scientific Reports* (Sept. 18, 2015). ISSN: 2045-2322. DOI: 10.1038/srep14138 (Cited on pages 7, 20, 28, 67, 113).

[21] James Bornholt, Randolph Lopez, Karin Strauss, Luis Ceze, Douglas M Carmean, and Georg Seelig. "Toward a Dna-Based Archival Storage System Storing Data in Dna Molecules Offers Extreme Density and Durability". In: *ASPLOS* (2016) (Cited on pages 7, 20, 55, 113).

[22] Kyle J. Tomek, Kevin Volkel, Alexander Simpson, Austin G. Hass, Elaine W. Indermaur, James M. Tuck, and Albert J. Keung. "Driving the Scalability of DNA-Based Information Storage Systems". In: *ACS Synthetic Biology* (June 21, 2019). DOI: 10.1021/acssynbio.9b00100 (Cited on pages 8, 55, 67).

[23] Billy Lau, Shubham Chandak, Sharmili Roy, Kedar Tatwawadi, Mary Wootters, Tsachy Weissman, and Hanlee P. Ji. "Magnetic DNA random access memory with nanopore readouts and exponentially-scaled combinatorial addressing". In: *bioRxiv* (Sept. 16, 2021). DOI: 10.1101/2021.09.15.460571 (Cited on page 8).

[24] Claris Winston, Lee Organick, David Ward, Luis Ceze, Karin Strauss, and Yuan-Jyue Chen. "Combinatorial PCR Method for Efficient, Selective Oligo Retrieval from Complex Oligo Pools". In: *ACS Synthetic Biology* (2022). PMID: 35191684. DOI: 10.1021/acssynbio.1c00482 (Cited on pages 9, 55).

[25] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z. Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N. Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss. "Random access in large-scale DNA data storage". In: *Nature Biotechnology* (Mar. 2018). ISSN: 1546-1696. DOI: 10.1038/nbt.4079 (Cited on pages 9, 19, 22, 26, 36, 41, 44, 46–48, 53, 55, 56, 67, 70, 89, 120, 134).

[26] James L. Banal, Tyson R. Shepherd, Joseph Berleant, Hellen Huang, Miguel Reyes, Cheri M. Ackerman, Paul C. Blainey, and Mark Bathe. "Random access DNA memory using Boolean search in an archival file storage system". In: *Nature*

Materials (June 10, 2021). ISSN: 1476-4660. DOI: 10.1038/s41563-021-01021-3 (Cited on pages 9, 12).

[27] Kyle J. Tomek, Kevin Volkel, Elaine W. Indermaur, James M. Tuck, and Albert J. Keung. "Promiscuous molecules for smarter file operations in DNA-based data storage". In: *Nature Communications* (June 10, 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-23669-w (Cited on page 9).

[28] N. Cardozo, K. Zhang, J. Nivala, L. Ceze, K. Strauss, D. Wilde, and C. Anderson. "Cas9-Mediated Random Access in DNA Data Storage". In: *2021 Synthetic Biology: Engineering, Evolution & Design (SEED)* (June 15, 2021) (Cited on pages 10, 15).

[29] Jacques Bonnet, Marthe Colotte, Delphine Coudy, Vincent Couallier, Joseph Portier, Bénédicte Morin, and Sophie Tuffet. "Chain and conformation stability of solid-state DNA: Implications for room temperature storage". In: *Nucleic Acids Research* (Dec. 7, 2009). ISSN: 03051048. DOI: 10.1093/nar/gkp1060 (Cited on pages 10, 11, 69, 71).

[30] B. Roder, K. Fruhwirth, C. Vogl, M. Wagner, and P. Rossmanith. "Impact of Long-Term Storage on Stability of Standard DNA for Nucleic Acid-Based Methods". In: *Journal of Clinical Microbiology* (Nov. 1, 2010). ISSN: 0095-1137. DOI: 10.1128/JCM.01230-10 (Cited on page 10).

[31] Robert Guthrie. "Blood Screening for Phenylketonuria". In: *JAMA* (Nov. 1961). ISSN: 0098-7484. DOI: 10.1001/jama.1961.03040470079019 (Cited on page 10).

[32] Karishma Matange, James M. Tuck, and Albert J. Keung. "DNA stability: a central design consideration for DNA data storage systems". In: *Nature Communications* (Mar. 1, 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-21587-5 (Cited on page 10).

[33] Dominique Clermont, Sylvain Santoni, Safa Saker, Maite Gomard, Eliane Gardais, and Chantal Bizet. "Assessment of DNA Encapsulation, a New Room-Temperature

DNA Storage Method". In: *Biopreservation and Biobanking* (June 1, 2014). ISSN: 1947-5535. DOI: 10.1089/bio.2013.0082 (Cited on pages 11, 72, 88).

[34] Kevin Washetine, Simon Heeke, Camille Ribeyre, Camille Bourreau, Corinne Normand, Hélène Blons, Pierre Laurent-Puig, Claire Mulot, Dominique Clermont, Maha David, Bruno Clément, Georges Dagher, and Paul Hofman. "DNAshell Protects DNA Stored at Room Temperature for Downstream Next-Generation Sequencing Studies". In: *Biopreservation and Biobanking* (Mar. 26, 2019). ISSN: 1947-5535. DOI: 10.1089/bio.2018.0129 (Cited on page 11).

[35] Weida D. Chen, A. Xavier Kohll, Bichlien H. Nguyen, Julian Koch, Reinhard Heckel, Wendelin J. Stark, Luis Ceze, Karin Strauss, and Robert N. Grass. "Combining Data Longevity with High Storage Capacity—Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles". In: *Advanced Functional Materials* (2019). ISSN: 1616-3028. DOI: 10.1002/adfm.201901672 (Cited on pages 11, 78).

[36] Susanne E. Howlett, Hilda S. Castillo, Lora J. Gioeni, James M. Robertson, and Joseph Donfack. "Evaluation of DNAstable™ for DNA storage at ambient temperature". In: *Forensic Science International: Genetics* (Jan. 1, 2014). ISSN: 1872-4973. DOI: 10.1016/j.fsigen.2013.09.003 (Cited on pages 11, 74, 88).

[37] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. "ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms". In: *Similarity Search and Applications*. Ed. by Christian Beecks, Felix Borutta, Peer Kröger, and Thomas Seidl. Springer International Publishing, 2017. ISBN: 978-3-319-68474-1 (Cited on pages 12, 16).

[38] Piotr Indyk and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality". In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. STOC '98. New York, NY, USA: Association for Computing

Machinery, May 23, 1998. ISBN: 978-0-89791-962-3. DOI: 10.1145/276698.276876 (Cited on page 12).

[39] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* (2015). DOI: 10.48550/arXiv.1409.1556 (Cited on page 12).

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012 (Cited on page 12).

[41] Callista Bee, Yuan-Jyue Chen, Melissa Queen, David Ward, Xiaomeng Liu, Lee Organick, Georg Seelig, Karin Strauss, and Luis Ceze. "Molecular-level similarity search brings computing to DNA data storage". In: *Nature Communications* (Aug. 6, 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-24991-z (Cited on pages 12, 56, 61, 63, 65).

[42] Joseph N. Zadeh, Conrad D. Steenberg, Justin S. Bois, Brian R. Wolfe, Marshall B. Pierce, Asif R. Khan, Robert M. Dirks, and Niles A. Pierce. "NUPACK: Analysis and design of nucleic acid systems". In: *Journal of Computational Chemistry* (Jan. 15, 2011). ISSN: 01928651. DOI: 10.1002/jcc.21596 (Cited on pages 12, 13).

[43] Yoshizumi Ishino, Mart Krupovic, and Patrick Forterre. "History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology". In: *Journal of Bacteriology* (Jan. 22, 2018). DOI: 10.1128/JB.00580-17 (Cited on pages 13, 14).

[44] Ruud Jansen, Jan D. A. van Embden, Wim Gaastra, and Leo M. Schouls. "Identification of genes that are associated with DNA repeats in prokaryotes". In: *Molecular Microbiology* (Mar. 2002). ISSN: 0950-382X. DOI: 10.1046/j.1365-2958.2002.02839.x (Cited on page 14).

[45] Melody Redman, Andrew King, Caroline Watson, and David King. "What is CRISPR/Cas9?" In: *Archives of Disease in Childhood - Education and Practice* (Aug. 1, 2016). ISSN: 1743-0585, 1743-0593. DOI: 10.1136/archdischild-2016-310459 (Cited on page 14).

[46] Vanja Tadić, Goran Josipović, Vlatka Zoldoš, and Aleksandar Vojta. "CRISPR/Cas9-based epigenome editing: An overview of dCas9-based tools with special emphasis on off-target activity". In: *Methods in Enzymology*. New Methods for Extracting Function from the Mammalian Genome (July 15, 2019). ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2019.05.003 (Cited on page 14).

[47] James P. Broughton, Xianding Deng, Guixia Yu, Clare L. Fasching, Venice Servellita, Jasmeet Singh, Xin Miao, Jessica A. Streithorst, Andrea Granados, Alicia Sotomayor-Gonzalez, Kelsey Zorn, Allan Gopez, Elaine Hsu, Wei Gu, Steve Miller, Chao-Yang Pan, Hugo Guevara, Debra A. Wadford, Janice S. Chen, and Charles Y. Chiu. "CRISPR–Cas12-based detection of SARS-CoV-2". In: *Nature Biotechnology* (July 2020). ISSN: 1546-1696. DOI: 10.1038/s41587-020-0513-4 (Cited on page 15).

[48] Peter Ney, Lee Organick, Jeff Nivala, Luis Ceze, and Tadayoshi Kohno. "DNA Sequencing Flow Cells and the Security of the Molecular-Digital Interface". In: *Proceedings on Privacy Enhancing Technologies* (2021). DOI: 10.2478/popets-2021-0054 (Cited on page 15).

[49] Stephen K. Jones, John A. Hawkins, Nicole V. Johnson, Cheulhee Jung, Kuang Hu, James R. Rybarski, Janice S. Chen, Jennifer A. Doudna, William H. Press, and Ilya J. Finkelstein. "Massively parallel kinetic profiling of natural and engineered CRISPR nucleases". In: *Nature Biotechnology* (Jan. 2021). ISSN: 1546-1696. DOI: 10.1038/s41587-020-0646-5 (Cited on pages 16, 17, 56, 57, 63–65).

[50] Aurelien Bellet, Amaury Habrard, and Marc Sebban. *Metric Learning*. Morgan Claypool, 2015. ISBN: 9781627053662. DOI: 10.2200/S00626ED1V01Y201501AIM030 (Cited on page 18).

[51] Alexander Hermans, Lucas Beyer, and Bastian Leibe. "In Defense of the Triplet Loss for Person Re-Identification". In: *arXiv* (Nov. 21, 2017) (Cited on page 18).

[52] Jonathan P. L. Cox. "Long-term data storage in DNA". In: *Trends in Biotechnology* (July 1, 2001). Publisher: Elsevier. ISSN: 0167-7799, 1879-3096. DOI: 10.1016/S0167-7799(01)01671-7 (Cited on page 20).

[53] Meinolf Blawat, Klaus Gaedke, Ingo Hütter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, and George M. Church. "Forward Error Correction for DNA Data Storage". In: *Procedia Computer Science*. International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA (Jan. 1, 2016). ISSN: 1877-0509. DOI: 10.1016/j.procs.2016.05.398 (Cited on pages 20, 22, 26, 114).

[54] Sriram Kosuri and George M Church. "Large-scale de novo DNA synthesis: technologies and applications". In: *Nature Methods* (May 1, 2014). ISSN: 1548-7105. DOI: 10.1038/nmeth.2918 (Cited on page 20).

[55] "Universal declaration of human rights". In: *The International Journal of Human Rights* (1998). DOI: 10.1080/13642989808406748 (Cited on page 22).

[56] OK Go. *OK Go - This Too Shall Pass - Rube Goldberg Machine - Official Video*. https://www.youtube.com/watch?v=qybUFnY7Y8w. Mar. 1, 2010 (Cited on page 22).

[57] Svalbard Global Seed Vault. *Svalbard Global Seed Vault*. https://seedvault.nordgen.org/ (Cited on page 22).

[58] Qikai Xu, Michael R. Schlabach, Gregory J. Hannon, and Stephen J. Elledge. "Design of 240,000 orthogonal 25mer DNA barcode probes". In: *Proceedings of the National Academy of Sciences* (2009). DOI: 10.1073/pnas.0812506106 (Cited on page 23).

[59] Tugkan Batu, Sampath Kannan, S. Khanna, and A. Mcgregor. "Reconstructing strings from random traces". In: *SODA '04, University of Pennsylvania Penn Libraries* (2004) (Cited on page 28).

[60] Jaume Pellicer, Michael F. Fay, and Ilia J. Leitch. "The largest eukaryotic genome of them all?" In: *Botanical Journal of the Linnean Society* (2010). ISSN: 1095-8339. DOI: 10.1111/j.1095-8339.2010.01072.x (Cited on page 38).

[61] Kyle J. Tomek, Kevin Volkel, Alexander Simpson, Austin G. Hass, Elaine W. Indermaur, James M. Tuck, and Albert J. Keung. "Driving the scalability of DNA-based information storage systems". In: *ACS Synthetic Biology* (May 24, 2019). ISSN: 2161-5063, 2161-5063. DOI: 10.1021/acssynbio.9b00100 (Cited on page 41).

[62] Manuela Zaccolo and Ermanno Gherardi. "The Effect of High-frequency Random Mutagenesis on *in Vitro* Protein Evolution: a Study on TEM-1 β-Lactamase". In: *Journal of Molecular Biology* (1999). ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1998.2262 (Cited on page 41).

[63] Lewis Y. Geer, Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H. Bryant. "The NCBI BioSystems database". In: *Nucleic Acids Research* (suppl_1 Jan. 1, 2010). ISSN: 0305-1048. DOI: 10.1093/nar/gkp858 (Cited on page 41).

[64] Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J. Garry. "DrImpute: imputing dropout events in single cell RNA sequencing data". In: *BMC Bioinformatics* (June 8, 2018). ISSN: 1471-2105. DOI: 10.1186/s12859-018-2226-y (Cited on page 42).

[65] Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. "Bayesian approach to single-cell differential expression analysis". In: *Nature Methods* (July 2014). ISSN: 1548-7105. DOI: 10.1038/nmeth.2967 (Cited on page 42).

[66] Sabine Verboven, Karlien Vanden Branden, and Peter Goos. "Sequential imputation for missing values". In: *Computational Biology and Chemistry* (Oct. 1, 2007). ISSN: 1476-9271. DOI: 10.1016/j.compbiolchem.2007.07.001 (Cited on page 42).

[67] Hyunsoo Kim, Gene H. Golub, and Haesun Park. "Missing value estimation for DNA microarray gene expression data: local least squares imputation". In: *Bioinformatics* (Jan. 15, 2005). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth499 (Cited on page 42).

[68] Qikai Xu, Michael R. Schlabach, Gregory J. Hannon, and Stephen J. Elledge. "Design of 240,000 orthogonal 25mer DNA barcode probes". In: *Proceedings of the National Academy of Sciences* (2009). ISSN: 0027-8424. DOI: 10.1073/pnas.0812506106 (Cited on page 48).

[69] Yuan-Jyue Chen, Christopher N. Takahashi, Lee Organick, Kendall Stewart, Siena Dumas Ang, Patrick Weiss, Bill Peck, Georg Seelig, Luis Ceze, and Karin Strauss. "Quantifying Molecular Bias in DNA Data Storage". In: (Mar. 4, 2019). DOI: 10.1101/566554 (Cited on page 48).

[70] S. M. Hossein Tabatabaei Yazdi, Ryan Gabrys, and Olgica Milenkovic. "Portable and Error-Free DNA-Based Data Storage". In: *Scientific Reports* (July 10, 2017). ISSN: 2045-2322. DOI: 10.1038/s41598-017-05188-1 (Cited on pages 55, 56, 114).

[71] Qikai Xu, Michael R. Schlabach, Gregory J. Hannon, and Stephen J. Elledge. "Design of 240,000 orthogonal 25mer DNA barcode probes". In: *Proceedings of the National Academy of Sciences of the United States of America* (Feb. 17, 2009). ISSN: 1091-6490. DOI: 10.1073/pnas.0812506106 (Cited on page 56).

[72] Lee Organick, Bichlien H. Nguyen, Rachel McAmis, Weida D. Chen, A. Xavier Kohll, Siena Dumas Ang, Robert N. Grass, Luis Ceze, and Karin Strauss. "An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage". In: *Small Methods* (2021). ISSN: 2366-9608. DOI: 10.1002/smtd.202001094 (Cited on pages 67, 77, 157).

[73] Robert N. Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J. Stark. "Robust chemical preservation of digital information on DNA in silica with error-correcting codes". In: *Angewandte Chemie - International Edition* (2015). ISBN: 2012290248. ISSN: 15213773. DOI: 10.1002/anie.201411378 (Cited on pages 69, 71–73, 88, 89).

[74] Robert Guthrie. "Blood Screening for Phenylketonuria". In: *JAMA* (Nov. 25, 1961). Publisher: American Medical Association. ISSN: 0098-7484. DOI: 10.1001/jama.1961.03040470079019 (Cited on page 69).

[75] Thomas C. Boothby, Hugo Tapia, Alexandra H. Brozena, Samantha Piszkiewicz, Austin E. Smith, Ilaria Giovannini, Lorena Rebecchi, Gary J. Pielak, Doug Koshland, and Bob Goldstein. "Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation". In: *Molecular Cell* (Mar. 16, 2017). ISSN: 1097-2765. DOI: 10.1016/j.molcel.2017.02.018 (Cited on page 69).

[76] Anne-Lise Fabre, Aurélie Luis, Marthe Colotte, Sophie Tuffet, and Jacques Bonnet. "High DNA stability in white blood cells and buffy coat lysates stored at ambient temperature under anoxic and anhydrous atmosphere". In: *PLOS ONE* (Nov. 30, 2017). Publisher: Public Library of Science. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0188547 (Cited on pages 71, 72, 88).

[77] Morten E. Allentoft, Matthew Collins, David Harker, James Haile, Charlotte L. Oskam, Marie L. Hale, Paula F. Campos, Jose A. Samaniego, M. Thomas P. Gilbert, Eske Willerslev, Guojie Zhang, R. Paul Scofield, Richard N. Holdaway, and Michael Bunce. "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils". In: *Proceedings of the Royal Society B: Biological Sciences* (Dec. 7, 2012). Publisher: Royal Society. DOI: 10.1098/rspb.2012.1745 (Cited on pages 71, 90).

[78] Tomas Lindahl and Barbro Nyberg. "Rate of depurination of native deoxyribonucleic acid". In: *Biochemistry* (Sept. 12, 1972). Publisher: American Chemical Society. ISSN: 0006-2960. DOI: 10.1021/bi00769a018 (Cited on page 71).

[79] GenTegra. *GenTegra DNA User Guide - Ambient temperature storage and transport of purified DNA, Version C*. https://nbsscientific.co.uk/wp-content/uploads/sites/8/2020/02/GenTegraDNA-User-Guide.pdf. Sept. 11, 2020 (Cited on page 72).

[80] A. Xavier Kohll, Philipp L. Antkowiak, Weida D. Chen, Bichlien H. Nguyen, Wendelin J. Stark, Luis Ceze, Karin Strauss, and Robert N. Grass. "Stabilizing synthetic DNA for long-term data storage with earth alkaline salts". In: *Chemical Communications* (2020). DOI: 10.1039/D0CC00222D (Cited on pages 74, 88, 89).

[81] Lewis Greenspan. "Humidity fixed points of binary saturated aqueous solutions". In: *Journal of Research of the National Bureau of Standards Section A: Physics and Chemistry* (Jan. 1977). ISSN: 0022-4332. DOI: 10.6028/jres.081A.011 (Cited on page 87).

[82] Marthe Colotte, Delphine Coudy, Sophie Tuffet, and Jacques Bonnet. "Adverse Effect of Air Exposure on the Stability of DNA Stored at Room Temperature". In: *Biopreservation and Biobanking* (Mar. 1, 2011). Publisher: Mary Ann Liebert, Inc., publishers. ISSN: 1947-5535. DOI: 10.1089/bio.2010.0028 (Cited on page 88).

[83] I. S. Reed and G. Solomon. "Polynomial Codes Over Certain Finite Fields". In: *Journal of the Society for Industrial and Applied Mathematics* (June 1, 1960). Publisher: Society for Industrial and Applied Mathematics. ISSN: 0368-4245. DOI: 10.1137/0108018 (Cited on page 89).

[84] Sharon Newman, Ashley P. Stephenson, Max Willsey, Bichlien H. Nguyen, Christopher N. Takahashi, Karin Strauss, and Luis Ceze. "High density DNA data storage library via dehydration with digital microfluidic retrieval". In: *Nature*

Communications (Apr. 12, 2019). Number: 1 Publisher: Nature Publishing Group. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09517-y (Cited on page 89).

[85] Natalia V Ivanova and Masha L Kuzmina. "Protocols for dry DNA storage and shipment at room temperature". In: *Molecular Ecology Resources* (Sept. 2013). ISSN: 1755-098X. DOI: 10.1111/1755-0998.12134 (Cited on page 89).

[86] Reimer C. Dobberstein, Jan Huppertz, Nicole von Wurmb-Schwark, and Stefanie Ritz-Timme. "Degradation of biomolecules in artificially and naturally aged teeth: Implications for age estimation based on aspartic acid racemization and DNA analysis". In: *Forensic Science International* (Aug. 6, 2008). ISSN: 0379-0738. DOI: 10.1016/j.forsciint.2008.05.017 (Cited on page 90).

[87] Nicolas Cardozo. "Developing Multiplexed Molecular Assays for Synthetic Biology and DNA Data Storage with Nanopore Sensing Technology". Accepted: 2022-04-19T23:41:48Z. Thesis. 2022 (Cited on page 91).

[88] Mingming Su, Yunchao Ling, Jun Yu, Jiayan Wu, and Jingfa Xiao. "Small proteins: untapped area of potential biological importance". In: *Frontiers in Genetics* (Dec. 16, 2013). ISSN: 1664-8021. DOI: 10.3389/fgene.2013.00286 (Cited on page 94).

[89] Jinny X. Zhang, John Z. Fang, Wei Duan, Lucia R. Wu, Angela W. Zhang, Neil Dalchau, Boyan Yordanov, Rasmus Petersen, Andrew Phillips, and David Yu Zhang. "Predicting DNA hybridization kinetics from sequence". In: *Nature Chemistry* (Jan. 2018). ISSN: 1755-4349. DOI: 10.1038/nchem.2877 (Cited on page 116).

[90] Cairong Yan, Xue Zhao, Qinglong Zhang, and Yongfeng Huang. "Efficient string similarity join in multi-core and distributed systems". In: *PloS one* (2017). DOI: 10.1371/journal.pone.0172526 (Cited on page 118).

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Stress-testing the encoder/decoder

We summarize our experience decoding two files that were sequenced with nanopore technology in Section 3.3.3, along with results of experiments we carried out in which we synthetically generated noise to assess the performance of our encoder/decoder (codec) against various levels of high noise, modeling a nanopore sequencing scenario. The two nanopore-sequenced files were 1.3KB and 32KB, respectively. When sequenced with a nanopore device, they experienced a coordinate error rate of 11.4% and 13.2%, respectively. Errors were distributed at around 30% insertions, 30% deletions, and 40% substitutions for both files. For the synthetic noise experiments, we encoded a 200 KB file with 12,799 strands of length 110 using a Reed-Solomon code with 15% redundancy, applied synthetic noise, and decoded the file reporting minimal coverages required for successful decoding. In these experiments, we varied the total coordinate error rate from 6% to 14% and designated 40% of errors to be substitutions, 30% to be deletions, and 30% to be insertions, respectively, as observed in practice.

The results, summarized in Section 3.3.3, indicate that the codec is capable of handling very aggressive noise scenarios. Note that at 10% error rate, when 30% of errors are deletions, 30% are substitutions, and strands are of length 110, a typical noisy read suffers at least three insertions and three deletions, which may be challenging to correct for other codecs. In synthetic experiments, we assumed that errors at different coordinates happen independently. In reality, errors are correlated. That is the likely explanation for existing discrepancies in critical coverages between real and synthetic experiments.

The last column in Table A.1 indicates how much coverage is needed for successful

Real reads	Size	Sequences	Coordinate error rate	Error distribution	Critical coverage	Critical coverage when dropping reads of incorrect length
File A	1.3KB	88	11.4%	Ins: 31%, Del: 29%, Sub: 40%	80x	>147x
File B	32KB	2,042	13.2%	Ins: 30%, Del: 31%, Sub: 39%	36	>74x
Synthetic Noisy Reads	200KB	12,799	6% 8% 10% 12% 14%	Ins: 30%, Del: 30%, Sub: 40%	8x 12x 17x 28x 60x	45x 65x 109x 195x 416x

Table A.1: Minimum required coverages to decode a file in different settings. The top two results came from decoding on a real nanopore sequencer. The remaining results were created by applying synthetic noise to an encoded file.

decoding if reads of incorrect length are ignored by our decoder, and shows that utilizing those reads significantly reduces critical coverage. This is in contrast with the case of (less noisy) Illumina sequenced data, where utilizing reads of incorrect length typically yields relatively small benefit (e.g., less than 5% reduction in critical coverage).

A.2 Calculating coverage and net density for DNA storage systems

Church et al. [16] report encoding 5.27 million bits of data in 54,898 DNA sequences of length 115 and using primers of length 22, yielding a net density of 0.83 bits per base (0.60 bits per base including primers). The paper reports using 3,000x coverage for retrieving the data.

Goldman et al. [17] report encoding 5.2 million bits of data in 153,335 DNA sequences of length 117 and using primers of length 33, yielding a net density of 0.29 bits per base (0.19 bits per base including primers). Ten percent of 79.6*106 noisy reads were used to recover the data, implying a mean coverage of 51x.

Grass et al. [18] report encoding 679 thousand bits in 4,991 DNA sequences of length 117 and use primers of length 21, yielding a net density of 1.16 bits per base (0.86 bits per base including primers). Sequencing produced 1,858,027 noisy reads that were used for decoding, implying a mean coverage of 372x.

Yazdi et al. [20] used 27 DNA sequences of length 1,000 to store two text files of total size 17Kb using a DNA coding scheme that allows for random access. However, the coding scheme is specifically designed to only store text data in English language. Therefore, we are unable to report its bits per base density. As such, we do not cover this scheme in Figure 3.1.

Bornholt et al. [21] modified the coding scheme developed in Goldman et al. [17] to improve the density to 0.85 bits per base (0.57 bits per base including primers). They have stored 150KB of data using DNA sequences of length 80 with primers of length 20. Decoding experiments (see Fig. 11 in Bornholt et al. [21]) included successfully recovering 16,994 DNA strands from 3.3% of 20.8 million noisy reads, implying a coverage of 40x.

Erlich and Zielinski [19] report encoding 2.11MB of data in 72,000 DNA sequences of length 152 and use primers of length 24, yielding a density of 1.55 bits per base (1.18 bits per base including primers). In one experiment, data was successfully retrieved from 750,000 noisy reads, i.e., at a mean coverage of 10.4x. In another experiment, DNA strands

were subject to nine successive PCR amplification reactions. The data was then retrieved from five million reads, i.e., at a coverage of 69x.

Blawat et al. [53] report encoding 22MB of data in 900,000 DNA strands of length 190 and use primers of length 20, yielding a density of 1.08 bits per base (0.89 bits per base including primers). Sequencing produced 144,475,005 noisy reads that were used for decoding, implying a mean coverage of 160x.

Yazdi et al. [70] used 17 DNA sequences of length 1000 to store 3,633 bytes of information, yielding a net density of 1.74 bits per base (1.71 bits per base including primers). Strands were sequenced using the Oxford Nanopore Technologies' MinIon. Decoding used approximately 200x coverage [O. Milenkovic, personal communication].

In this work, we have used 13,448,372 DNA sequences of length 110 and 114 with primers of length 20 to store 35 different files of total size 200MB yielding a net density of approximately 1.1 bits per base (0.8 bits per base including primers). Our system provides random access to individual files, and requires different amounts of coverage for decoding, as Table A.2 shows. Required decoding coverage ranges from 4x to 14x with a median of 5x, while required preparation coverage ranges from 4.2x to 20.5x with a median of 6.2x.

File Number	Minimum average coverage	Standard deviation
1	11.0	10.6
2	7.0	4.8
3	7.0	3.9
4	6.0	3.9
5	6.0	3.5
6	5.0	3.1
7	4.0	2.5
8	4.0	2.5
9	3.8	2.4
10	5.0	3.3
11	4.0	2.6
12	4.0	2.5
13	4.0	2.4
14	4.0	2.6
15	4.0	2.5
16	5.0	3.0
17	5.0	3.2
18	4.0	2.5
19	10.0	9.3
20	14.0	13.7
21	9.0	8.2
22	7.0	6.1
23	7.0	5.9
24	6.0	5.0
25	7.0	6.4
26	14.0	11.9
27	8.0	7.3
28	5.0	3.3
29	5.0	3.4
30	6.0	3.7
31	6.0	3.3
32	11.0	7.3
33	4.0	2.4
34	5.0	3.5
metadata	4.0	2.6
median across all files	**5.0**	

Table A.2: Files stored and the minimum average coverage needed to decode these files, along with the standard deviation of coverage across different DNA sequences in each file.

A.3 Primer library scalability

To estimate the scalability of our primer design, we generated six pools of random 20-mers, with total number of sequences ranging from 10,000 to 300,000, and applied the selection criteria explained in Fig. 3.2. The total number of primers that pass the design criteria is proportional to the log of the number of random 20-mers (Fig. A.1a). Extrapolating this trend gives us an estimated maximum library size of 14,000 20-mer primer pairs when using this method.

We also tested primer passing ratios of the collision detection algorithm explained in Supplementary Fig. 3. We used three arbitrarily chosen files (ID6, ID7, and ID21) with file sizes ranging from 370KB to 13.9MB. The primer passing ratio decreases linearly with the log of the file size (Fig. A.1b). Extrapolating this trend shows that it is possible to reach TBs of data in a single pool. These results indicate that we can scale the above methods to pools of a few TBs of data, organized into roughly 1,000s files (or, more generally, random access data blocks), each containing a few GBs of data.

When testing our primer library, we used multiplex PCR and observed large variations of sequencing coverage (Fig. 3.3a). Note that this variation does not speak to the validity of our method: It is an artifact of multiplex PCR, which typically generates uneven amplification stemming from multiple sources:

1. Primer sequence differences can cause different hybridization kinetics and affect PCR yield [89].

2. Payload sequence differences (e.g., GC content, secondary structures) can also affect PCR yield.

3. The synthesis process may create an initial uneven distribution of the DNA pool.

4. Errors in mixing equal volume of different primers.

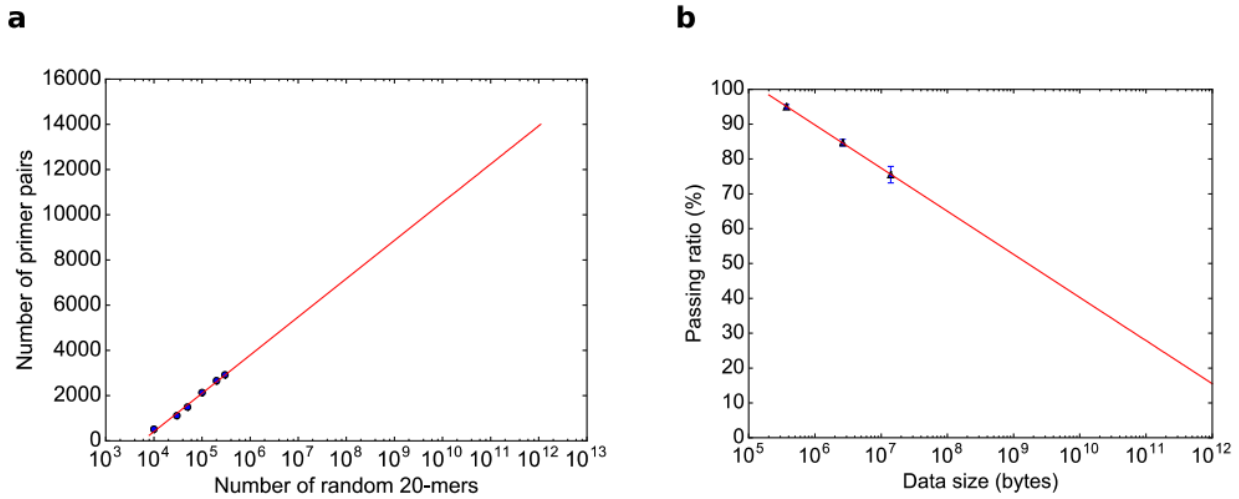


Figure A.1: **(a)** The total number of primer pairs that pass the selection criteria described in Supplementary Fig. 1 (y-axis) increases with the log of the number of starting random 20-mers (x-axis, logarithmic scale). The six blue dots are primer libraries generated from different numbers of starting random 20-mers. **(b)** The ratio of primers passing the primer-payload collision detection algorithm described in Supplementary Fig. 4 (y-axis) decreases as the log of the amount of information to be stored increases (x-axis, logarithmic scale). Blue points represent the average passing ratio of the six different primer libraries generated in a. Error bars indicate standard deviation calculated from these six primer libraries. The measures of the centre indicate mean calculated from these six primer libraries.

The read coverage is a result of PCR over the uneven sequences resulting from multiplex PCR, further amplifying small variations and exacerbating uneven coverage. Uneven sequencing coverage is inconvenient because it requires higher overall sequencing coverage to generate at least a few reads of each encoded sequence in a file. However, error correction mitigates this need because it allows missing sequences, trading off higher logical redundancy for lower sequencing coverage. Note that this is the only experiment in this study to use multiplex PCR – single-primer PCR coverage is much more even.

A.4 Clustering using off-the-shelf software

The clustering step aims to group noisy reads into clusters, where each cluster corresponds to a single unknown reference strand. For an example open-source clustering method, we recommend using StarCode31. The user-friendly code and executable may be obtained from the following repository: https://github.com/gui11aume/starcode. We decoded a file successfully using a distance threshold of four (i.e., the "-d 4" parameter), with the sphere clustering option (i.e., the "-s" flag). Larger distance thresholds will lead to fewer clusters, containing more reads on average.

Computing a string similarity join also suffices to cluster the reads. A similarity join is the set of all pairs of input strings that have distance shorter than or equal to a specified threshold. After computing the similarity join of all noisy reads, the desired clustering corresponds to the transitive closure of the similar pairs. Examples of similarity join algorithms appear in Yan et. al. [90] and the references therein. Again, setting a distance threshold of at least four will produce a sufficiently accurate clustering. Note that longer or noisier reads would require using a larger distance threshold.

We make available (via http://misl.cs.washington.edu/data) two files that enable the reproduction of key parts of our decoding pipeline: (1) A FASTQ file containing reads associated with a single file; and (2) A list of references corresponding to these reads. The file corresponds to File 16; Front Primer: AGTGCAACAAGTCAATCCGT; Reverse Primer: AATTGAATGCTTGCTTGCCG; File Type: Portable Document Format (BZip2 Compressed); Size (Bytes): 9503066; of Sequences: 607,150.

The FASTQ file contains Illumina-sequenced reads that correspond to this file. There are roughly 16 million reads. To extract the relevant portion of the read, the front/reverse primers must be aligned (or directly matched) to portions of the read. Then, the payload lies between the primers, and the ends may be truncated. Note that the read can be either in the correct, template format, or it can appear as a reverse complement. Therefore, the primers must be searched for in both configurations. An additional file contains a list of

reference IDs that generated each read in the FASTQ file. This file was constructed by aligning the reads to the actual reference strands.

A.5 Detailed Error Analysis

Dataset: The dataset consists of 35 files, including one metadata file (see the published work's Supplemental Section 3 for details about the files [25]). All but two files have 15% redundancy (file number 33 and the metadata file have 25% redundancy). The total size of the data is approximately 200MB. The data was synthesized in nine pools, and sequenced in ten sequencing runs.

Errors

- **Error rates.** The overall error rate per position across the whole dataset and the entire sequences (including primer regions) is 0.637%. Substitutions are most prominent, with an error rate per position of 0.407%, deletions have an error rate per position of 0.188%, and insertions are least common, with an error rate per position of 0.042%.

- **Error rates across positions.** Different files show different error rate profiles. In general, the primers at the ends of the read have lower error rates than the central part of the read containing the address and the payload; across the whole dataset the error rate for the payload is 0.662%, while the error rate for the primers is 0.570% (caused by a deletion error in the primer region of one file). The substitution, insertion, and deletion rates per position are reasonably uniform throughout positions 21 through 130, although several files contain deviations (typically spikes) in a few positions.

- **Error rates depending on the base.** Deletion errors in the payload are roughly uniform across bases, i.e., A, C, G, T are deleted roughly 25% each. Substitution errors show more variance, with most errors involving Ts. For example, G to T and T to C each represent over 15% of all substitutions, while A to G represents only 2% of them. For insertion errors, Gs are the most common: about 46% of all insertions.

Error rates depending on nearby bases. We examined the effect of neighboring bases on the error rate (i.e., the 3-mer centered on a position), and found that there is an effect, although secondary, when compared to the effect of the center base itself. Figure 3d shows that average payload error rates are affected by the base at a given position and the two neighboring positions.

Appendix B

APPENDIX FOR CHAPTER 4

B.1 Primer sequences, amplification efficiency, and fragment analysis

Each file has a unique primer pair, facilitating random access in DNA data storage.

File	Forward Primer	Reverse Primer
Small	5' ATAATTGGCTCCTGCTTGCA 3'	5' TTGCACTTTCCGCCTACATT 3'
Medium	5' AATCATGGCCTTCAAACCGT 3'	5' AACAAGACTTTCGGAGCGTT 3'
Large	5' AACATCGTGTCCAAGCAAGT 3'	5' TTGTTTGTCCACGCTTTCGA 3'

Table B.1: Forward and reverse primer sequences needed to amplify each file

Each file had its own ultramers to act as standards. The sequences and their percent efficiency as determined by qPCR is below. Each reaction was performed in triplicate, and to determine the standard efficiency each standard was diluted serially by an order of magnitude six times.

Ultramer	Sequence	%Amplification Efficiency
Small File Ultramer 1	ATAATTGGCTCCTGCTTGCAAGCTAGCTGCTATA CACACTCACACACGTCTCGTGTCTGCTGCTGCGA TCTGTGATAGCGTGTGAGAGACGCTAGATGCGCT GACATCTGTGCACACAGACACGAGCTAGAATGTA GGCGGAAAGTGCAA	91

Small File Ultramer 2	ATAATTGGCTCCTGCTTGCAGCTAGCTAGTGCGT GCAGCGTCTCTCAGCTCTACACTCTCTATCTACG CTAGTACGTATGTGTGACATGCGTCTGTGCAGAG CATCTACTCGCGCGAGCATACATAGCTAAATGTA GGCGGAAAGTGCAA	100
Small File Ultramer 3	ATAATTGGCTCCTGCTTGCACTAGCTAGTAGCTA CGTATGATATATACACATCAGATGCGCAGCGCGT AGCTACTATGTCACGATACTACAGCTATCGCGAT ACATATAGACGTGCACTCAGCGAGCTAGAATGTA GGCGGAAAGTGCAA	88
Small File Ultramer 4	ATAATTGGCTCCTGCTTGCATAGCTAGCGTGTCT ATCAGATGCGTGACAGTCTGTATGCGCATACATA CTGCTACTACGTACTCTACGCTATACACTAGTCT GCGACTACGAGTATGAGCAGCTCTAGCTAATGTA GGCGGAAAGTGCAA	98
Med. File Ultramer 1	AATCATGGCCTTCAAACCGTAGCTAGCTAGCGTC TACATATACAGTACTATGCAGTATATGTCATCTC AGTCAGTGTGCATCAGCTGTACAGTGCGCTACGC TACTCTATAGATAGACAGAGAGTGCATAAACGCT CCGAAAGTCTTGTT	92
Large File Ultramer 1	AACATCGTGTCCAAGCAAGTAGCTAGCTAGCTAG TCACTCGAGCGTGCACGTGCTACAGTGCGATGAC GTCTCTCTACAGATACAGTATGTATACACTATGT GCAGACGAGTCAGATATCTGCACACGAGTCGAAA GCGTGGACAAACAA	90

Table B.2: Each ultramer's sequence used for qPCR standards, as well as the resulting percent amplification efficiency of the qPCR reaction.

There are four ultramers for the small file in order to explore how much variation in amplification efficiency might be observed. A wide range of amplification efficiency results would mean it's not informative to use ultramers as quantification standards in the qPCR reactions calculating copy number. However, there was an observed 12% range, which satisfied us that picking an ultramer at random to serve as a quantification standard was sufficient. An ultramer with 110 random bases in the payload region of the strand was not ordered because it does not exclude homopolymers as our encoding scheme does, which was thought to be a potentially confounding factor.
When comparing the qPCR standard (the ultramer(s) listed above), it becomes clear that the smaller the file, the greater the difference in amplification efficiency from the standard.

Mean Percent Amplification Efficiency			
File	Standard	Water Dilution	150Nmer Dilution
Small	94%	145%	149%
Medium	92%	113%	114%
Large	90%	98%	103%

Table B.3: Comparing the percent amplification efficiency of each file's standard to its samples diluted in water and diluted in 150Nmers.

We hypothesize that due to the low number of target strands in solution, the small and medium file begin spuriously amplifying primers or other fragments of DNA in solution to

create an inaccurate amplification curve. This is further supported by the fact that the small file's smallest dilution has a C_t value that completely overlaps the negative control, and also supported by the fragment analyzer results (see Fig. B.1) which clearly shows the small file has more side products due to spurious amplification than the medium and large file for the same dilution.

As detailed in the Methods section of the main text, we only used qPCR data to determine the first (undiluted) copy number for the small and medium file. This is because at first, the authors tried to determine copy number based off all the qPCR results for each file. However, the amplification efficiency wasn't accurate because of the spurious amplification happening in each PCR cycle (as discussed in detail above). The resulting copy number for the last last dilutions were found to be 0.3/0.4 (diluted in water or 150Nmer respectively) for the large file, 0.7/0.6 for the medium file, and 28/29 for the small file. This would be reasonable if the small file had a much larger starting copy number than the large and medium file. However, all three have a very similar starting copy number (180, 190, and 213 for the large, medium and small file respectively). A final dilution copy number nearly two orders of magnitude higher than the other two files was clearly not accurate. In addition- the Cq value (essentially, the qPCR cycle where the Ct value was crossed) was indistinguishable from the negative control at the lowest dilution for the small file, further rendering the qPCR data unhelpful.

We therefore decided that since the large file had the best amplification efficiency when compared to its qPCR standard than the medium and small file, the most accurate way to determine copy number would be to rely on the results of the large file's qPCR. Specifically, utilizing its dilution factor. While we were able to accurately determine the medium and small file's starting copy number (pre-dilution) using custom, accurate dilution standards for each of the files, we calculated subsequent copy numbers by leaning on the qPCR data for the large file as described in the Methods.

We performed a gradient qPCR in duplicate on all dilutions and conditions for the small file. The annealing temperatures were 57, 60, 62, 65 and 67 degrees Celcius while the rest

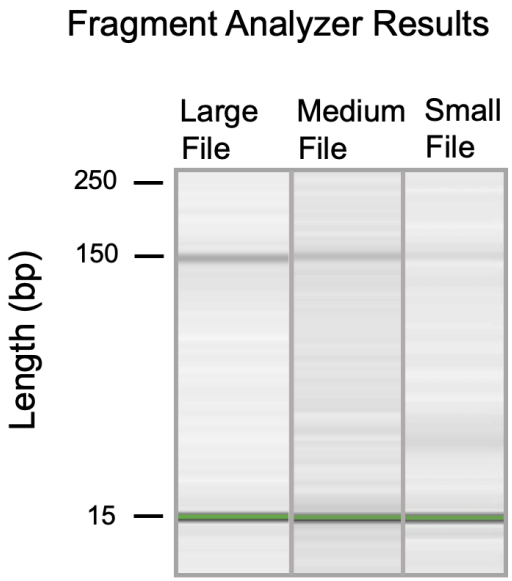


Figure B.1: Results of a fragment analyzer performed on a Qiagen QIAxcel. The desired band is at 150 bp. All samples shown here are the PCR results of the final dilution in water. The large file has the most desired product and the least spurious amplification, while the small file has the least desired product and the spurious amplification appears in such great quantity a wide band can be seen at a length of approximately 20-30bp. The medium file likely also has spurious amplification, but in such low quantities the fragment analyzer cannot detect it. All images have the same contrast and other image settings.

of the protocol followed those given in the Methods section of the paper. The annealing temperature used in the paper was 62°C.

When compared to the paper's annealing temperature to other the other four temperatures, we find the C_t values to be either worse or indistinguishable from the 62°C C_t value. A C_t value is considered worse if it took longer to amplify than the sample with an annealing temperature of 62°C.

Of all the conditions that did not have a worse C_t value, the mean positive difference in C_t value was found to be 0.09 and the standard deviation was 0.05 cycles.

B.2 Calculating Pool Complexity and Emulated Data

To ensure the 150Nmer ordered from IDT was largely the full 150nt product, gel electrophoresis was done, as shown in Fig. B.2. Using FIJI analysis software, the band at the desired 150nt length was found to comprise 80% of the total product, and this was incorporated in our calculations.

The 150Nmer ordered from IDT was run through a 10% PAGE-urea gel with 1x TBE buffer and run under 200V for 25 minutes.

For each sample, the Sample Buffer was the BioRad TBE-Urea Sample Buffer (Catalog 161-0768). The final volume added to each gel well was 15 μL.

The 10bp ladder was the ss10 DNA Ladder from Simplex Sciences. The final solution was 1 μL ladder, 9 μL molecular grade water, and 10 μL Sample Buffer. The 50bp ladder was from New England BioLabs, product NEB B7025. The final solution was 0.5 μL ladder, 9.5μL molecular grade water, and 10 μL Sample Buffer. To denature the ladder, this sample was denatured by incubating at 95°C for 5 minutes, then immediately transferred into the gel well. The 150Nmer sample contained 5 μL of 1 ng/μL 150Nmer sample as measured by a ssDNA standards with a Qubit 3.0 Fluorometer. 5 μL 1x TE buffer was added, along with 10 μL Sample Buffer.

To determine how complex the pool diluted in 150Nmers actually is, we must find out how many strands there are per microliter. We can then do simple arithmetic to calculate the number of sequences per microliter (the starting sample of each PCR reaction).

Using the DNA Copy Number Calculator provided by ThermoFisher, we can input that the 150Nmers are 325 (g/mol)/bp because it is single stranded, and that it is a custom DNA fragment 150 nucleotides in length. With this information, we can now see that there are ideally $1.2 * 10^{10}$ strands of DNA per nanogram, but in reality due to imperfect synthesis we multiply this by 80% to determine a more accurate number of full length strands per nanogram. The result is $0.96 * 10^{9}$ full length strands of DNA per nanogram.

Due to the fact that each step of the dilution protocol adds a specified amount of 1ng/μL

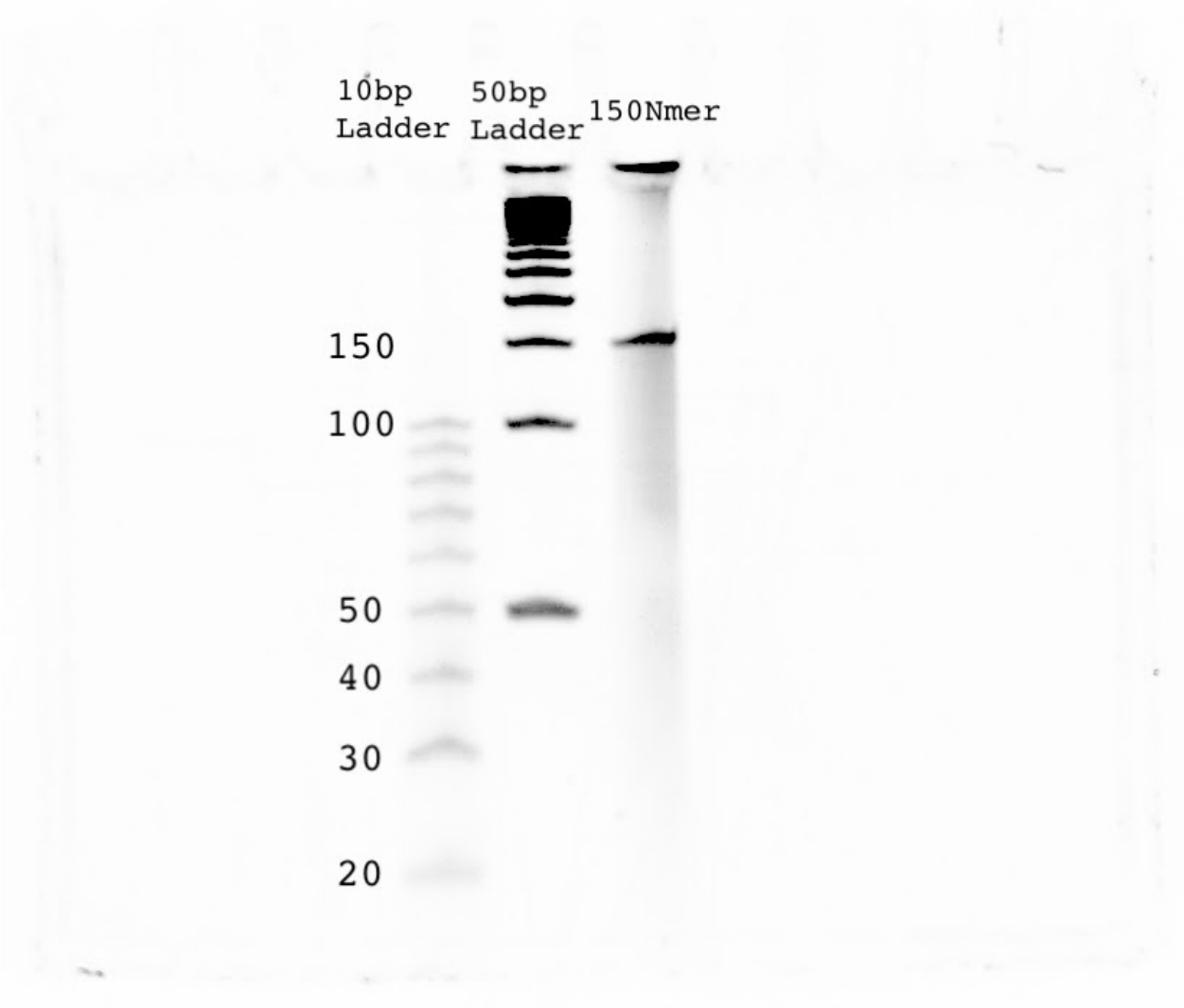


Figure B.2: Gel electrophoresis of the 150Nmer. Please note that the residue left behind in the well for both the 50bp ladder and the 150Nmer is not thought to be DNA product, as it did not migrate.

150Nmer solution, it is simple to multiply that quantity by $0.96 * 10^9$ and divide by the total resulting volume to get the number of unique strands per μL being added to the previous unique strands per μL.

The amount of digital data these strands emulate is found by knowing that 200.2MB are encoded in 13,448,372 unique DNA sequences (from Organick et al. 2018), so each unique

DNA sequence contains about 15.6 bytes. To calculate the total amount of data, we multiply this number by the number of different DNA sequences in solution.

The exact numbers are shown below:

*g/mol/bp**: 325

Fragment length (nt): 150

Strands/ng: 9,882,256,410

Bytes/MB: 1024^2

GB/μL: 143.6

TB/45μL: 6.310

TB in Final 50μL Solution: 7.002

**Note that these strands were ssDNA.*

B.3 Investigation of Small File Missing Sequence Behavior

As seen in Fig. 4.2 and Fig. 4.3, it appears that the small file does not lose sequences at the same rate as the other two files. Upon further investigation, this is likely due to a combination of factors. First, the copy number is likely slightly higher than what was calculated, thus explaining the lack of sequences lost. Second, there are fewer sequences initially missing from the file. Third, as a byproduct of synthesis, the distribution of sequences is slightly more uniform (Fig. B.3), further increasing the likelihood of greater sequence recovery.

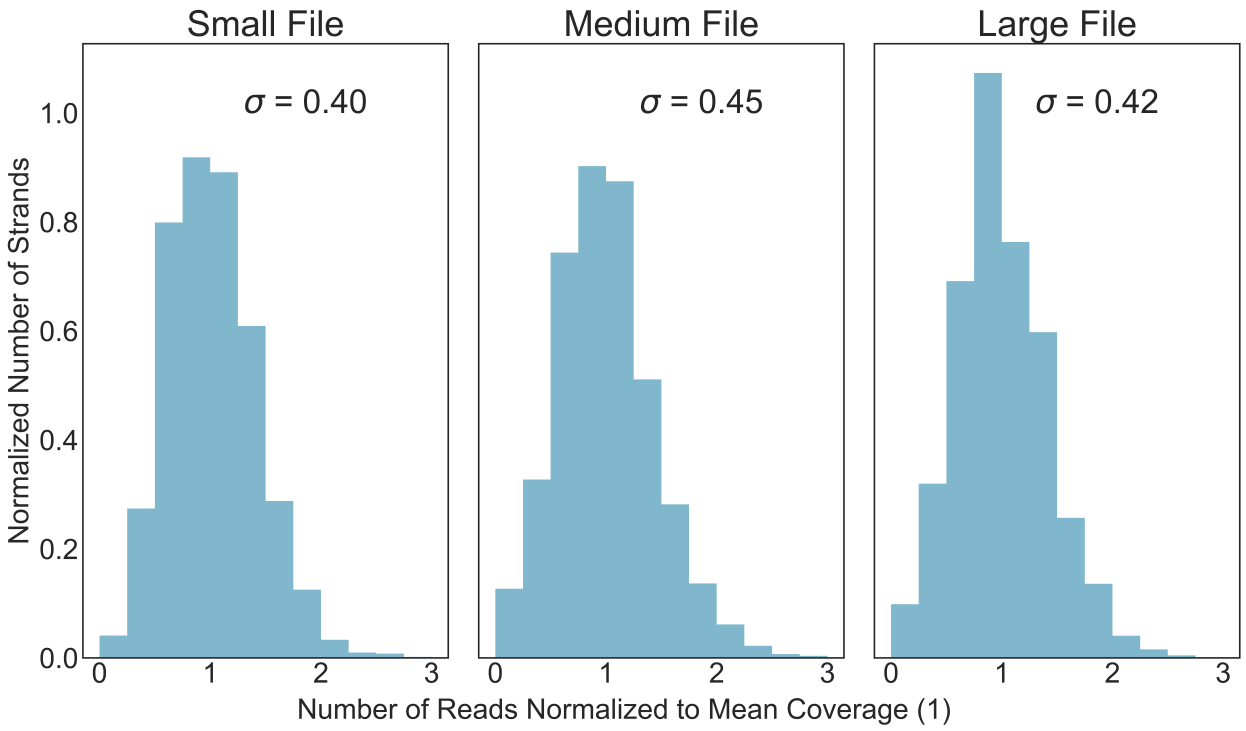


Figure B.3: Comparisons of normalized initial sequence distributions prior to dilution (with a mean copy number of 194) with standard deviation from the mean overlayed.

One hypothesis was that poor alignment scores were incorrectly identifying sequences for the small file, therefore giving the appearance of fewer sequences missing. However, as

shown in Table B.4, this is not supported.

File	Water	150Nmers
Small	130 ± 38	68 ± 71
Medium	29 ± 56	31 ± 58
Large	140 ± 26	140 ± 24

Table B.4: Alignment scores (mean $\pm$ standard deviation) for each file and diluent at the last dilution step.

Another hypothesis was that many sequencing reads were in fact aligning to more than one reference sequence, resulting in what's known as a chimera alignment. A chimera alignment is when one read is incorrectly identified as matching two or more reference sequences. However, as shown in Table B.5, we did not find the small file to have significantly fewer reads with only one alignment.

File	Water	150Nmers
Small	97	99
Medium	99	99
Large	97	98

Table B.5: Percent of reads with one alignment for each file and diluent at the last dilution step.

To ensure that chimera alignments weren't causing false sequence recovery rates, we then examined the number of alignments per chimera read. If the small file had many more alignments per chimera read, that could explain the observed sequence recovery rates. However, as shown in Table B.6, there was virtually no difference between files.

For discussion of the decoding behavior of the small file, please see Appendix B.7.

File	Water	150Nmers
Small	2.1 ± 0.25	2.0 ± 0.21
Medium	2.0 ± 0.17	2.0 ± 0.16
Large	2.1 ± 0.30	2.1 ± 0.23

Table B.6: Number of alignments (mean ± standard deviation) per chimera read for each file and diluent at the last dilution step.

B.4 Calculating Power Regressions

Power regressions for fitting curves in Fig. 4.3 were found in Python. The following is the code used to generate both the line of best fit and the associated R^2 values:

```
# Python 2.7 was used
from scipy.optimize import curve_fit

def power(x, a, b):
    y = a * x ** -b
    return y

def get_regressions(xdata, ydata):
    """
    Given xdata (a list of copy numbers) and ydata (a list of the
    percentage of file missing), prints the resulting power line
    of best fit and associated R^2 value
    """
    popt, pcov = curve_fit(power, xdata, ydata)
    # popt[0] is the coefficient of x given by function "power" (a)
    # popt[1] is the exponent of x given by function "power" (b)
    x_linspace = np.linspace(min(xdata), max(xdata), 50)
    power_y = popt[0]*x_linspace**-popt[1]

    # Manually Calculating R^2 value
    residuals = ydata - power(xdata, popt[0], popt[1])
    ss_res = np.sum(residuals**2)
    ss_tot = np.sum((ydata - np.mean(ydata))**2)
    rsq = 1 - (ss_res / ss_tot)

    power_equation = str( y = popt[0]x**(-popt[1]) )
    print "The line of best fit is", power_equation
    print "The R^2 value is", rsq
```

B.5 Calculating Information Density per Gram

Using the DNA Copy Number Calculator provided by ThermoFisher, we can input that the 150Nmers are 325 (g/mol)/bp because it is single stranded, and that it is a custom DNA fragment 150 nucleotides in length. With this information, we can now see that there are $1.2 * 10^{10}$ strands of DNA per nanogram.
By knowing 15.6 bytes are encoded in each strand, as discussed in Chapter 3 [25], we can then do simple arithmetic to determine the maximum amount of data that can be stored and retrieved using the encoding scheme also used in Chapter 3 [25].

The exact numbers are shown below:

*g/mol/bp**: 325

Fragment length (nt): 150

Strands/ng: 12,352,820,513

Bytes/strand: 15.6

Minimum copy number: 10

$$\frac{Bytes}{ng} = \frac{\frac{strands}{ng}}{minimum\ copy\ number} * 15.6\frac{Bytes}{strand} = 19,270,400,000$$

$$\frac{EB}{g} = \frac{\frac{Bytes}{ng}}{1024^6 \frac{Bytes}{EB}} * 10^9 \frac{ng}{g} = 16.7$$

* Note DNA here is single stranded.

B.6 Effect of Pool Complexity on Sequence Recovery

To investigate the role of complex pools on sequence recovery, the behavior of each individual sequence was compared between dilution conditions to check for systematic trends in recovery.

For example, each sequence in the third water dilution was compared to its same sequence in the third 150Nmer dilution. First, each file was normalized by the population fraction equation:

$$\tilde{C} = \frac{C}{T} \tag{B.1}$$

Here, $\tilde{C}$ is the normalized coverage for a given strand, C is the raw number of times the strand was seen by the sequencer and T is the total number of reads the sequencer read for that sample.

The difference in recovery behavior between the two identical sequences was then found by taking the difference between $\tilde{C}$ values for each sequence (Fig. B.4, Fig. B.5, Fig. B.6). A skewed recovery might show sequences behaving in a systematic way, either failing to appear in the water diluted sample (negative values) or appearing much more in the water diluted sample (positive values). However, this was not found to be the case, as all samples were found to have a mode of 0 and clustered heavily around 0 largely symmetrically, showing that most sequences do not have much change in frequency regarding pool complexity.

Some of the variation in normalized coverage may have come from stochastic variation as a product of subsampling the pool at the dilution step, and at the PCR retrieval step. This is supported by Fig. B.7, in which Venn diagrams compare the lost sequences from one dilution to the next. In Fig. B.7, it is interesting to note that the lost sequences in subsequent dilutions are not merely supersets of those from the prior dilution, despite them being serially diluted. However, when the initial undiluted sample's missing strands are compared to the final dilution's missing strands (with copy numbers 194 vs 0.3

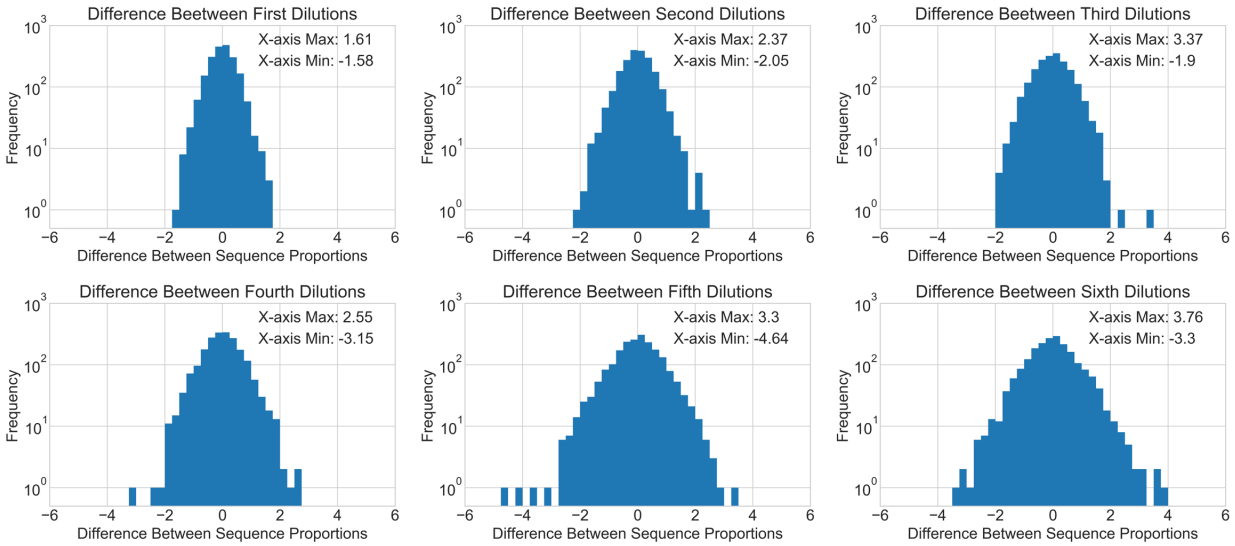


Figure B.4: All comparisons between dilution conditions for the small file.

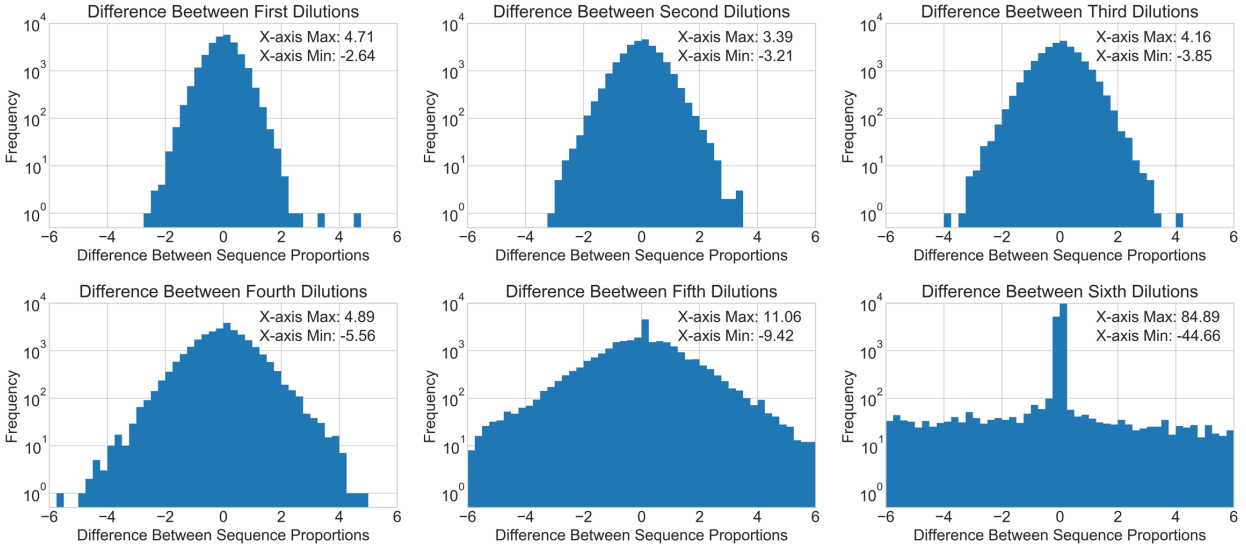


Figure B.5: All comparisons between dilution conditions for the medium file.

respectively), we see that the missing strands from the final dilution are still not a perfect superset. This illustrates the stochastic nature of sampling sequences, and does not indicate a strong, underlying property of strands that makes them unrecoverable.

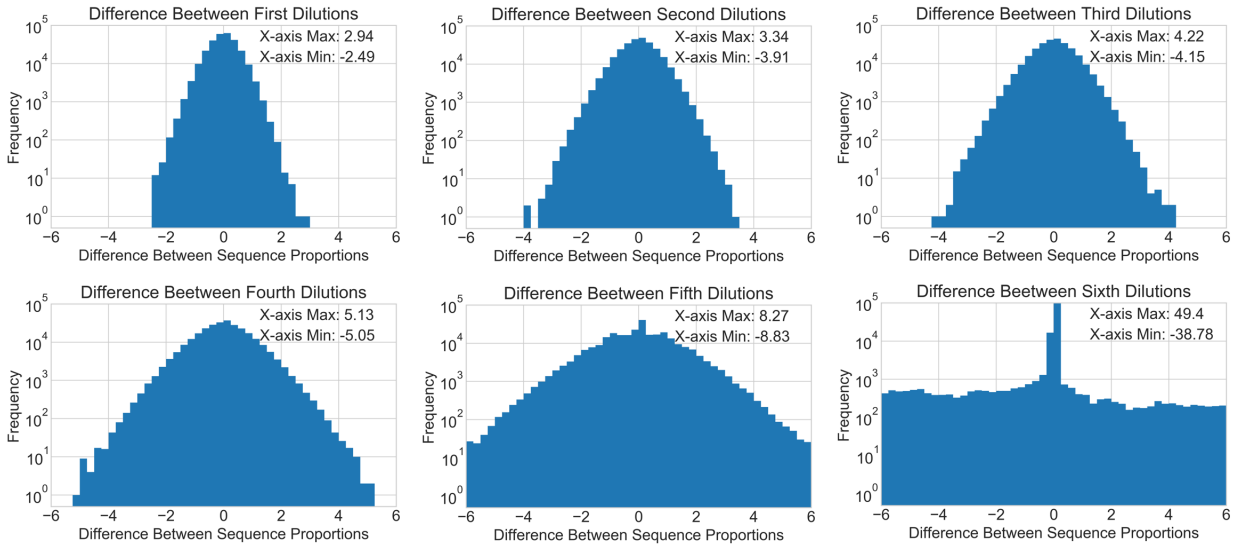


Figure B.6: All comparisons between dilution conditions for the large file.

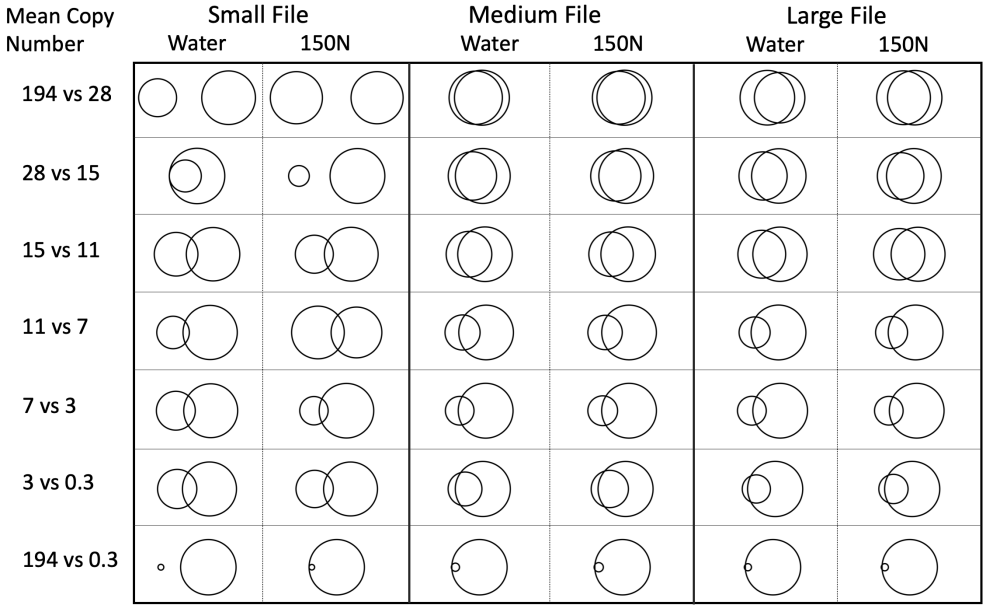


Figure B.7: For each file and diluent condition, a Venn diagram compares the set of sequences missing from the two copy numbers compared (given in the leftmost column). Each circle represents the set of sequences that were not recovered, and the overlap between the circles represents the sequences missing from both dilutions.

B.7 Decoding thresholds

As shown in Fig. 4.3a, there is a threshold at which the data can no longer be decoded with no bit errors. The following provides a more granular presentation of that data for the small file (Table B.7), medium file (Table B.8), and large file (Table B.9). Sequencing coverage is determined by comparing alignment of actual sequences recovered to sequences expected by the reference template; it is the result of total alignments divided by sequences in the file. Decoding is blind to any references and is thus more vulnerable to losing sequences.

Diluent	Copy Number	Sequencing Coverage	Decoded?
Water	11.2	22.5	Yes
Water	6.5	25.5	Yes
Water	3.0	25.0	No
Water	0.3	21.9	No
150Nmers	13.6	20.5	Yes
150Nmers	8.5	27.9	Yes
150Nmers	3.2	20.5	No
150Nmers	0.4	20.8	No

Table B.7: Sequencing and recovery data for the small file

Note that as shown in Fig. 4.3a the small file fails to decode with a lower percent of sequences missing than the medium or the large file. However, the ultimate difference is not as great as it might seem as the copy numbers are different for each data point, and in fact the small file successfully decodes with the average lowest copy number of 7.5 (when averaging both water and 150Nmer diluent conditions), while the medium and large file are 12.7 and 10.5 respectively. This data is presented in tables B.7, B.8, and B.9.
Furthermore, we have performed an in-depth analysis of the decoding process for all three

Diluent	Copy Number	Sequencing Coverage	Decoded?
Water	14.0	22.0	Yes
Water	10.0	30.7	Yes
Water	5.8	23.6	No
Water	2.7	23.3	No
150Nmers	28.1	21.8	Yes
150Nmers	15.4	29.0	Yes
150Nmers	12.1	22.6	No
150Nmers	7.6	23.7	No

Table B.8: Sequencing and recovery data for the medium file

files and did not see significant deviations in other error metrics such as global rate of errors or rate of insertion, deletion, or substitution errors per coordinate. The difference in the observed behavior likely comes from the small file size of 2,042 sequences, for at small sizes random effects of noise become more pronounced and larger deviations become more likely. Philosophically, this is similar to the law of large numbers. Behavior is more predictable when a larger number a random variables are summed together.

Diluent	Copy Number	Sequencing Coverage	Decoded?
Water	180.2	22.2	Yes
Water	9.5	54.5	Yes
Water	5.5	33.5	No
Water	2.6	29.0	No
150Nmers	14.6	23.1	Yes
150Nmers	11.5	42.5	Yes
150Nmers	7.2	25.6	No
150Nmers	2.7	37.9	No

Table B.9: Sequencing and recovery data for the large file

B.8 Ligation Protocol

B.8.1 Ligation

The ligation protocol used is a combination of the Illumina TruSeq ChIP Sample Preparation protocol and the Illumina TruSeq Nano Library Prep protocol, using the Illumina TruSeq Nano reagents. The following are the step-by-step directions used in this work.

1. Add 40 μL ERP2 to each well with 60 μL of PCR product
2. Pipette up and down to mix
3. Place on the thermal cycler and run the ERP program (30 min at 30°C) then place on ice. Each well contains 100 μL.
4. Add 160 μL well-mixed AMPure XP Beads to each well of the PCR plate containing 100 μL End Repair Mix.
5. Gently pipette the entire volume up and down 10 times to mix thoroughly.
6. Incubate the PCR plate at room temperature for 15 minutes.
7. Place the PCR plate on a magnetic stand at room temperature for 15 minutes or until the liquid is clear.
8. Using a 200 μL single channel or multichannel pipette set to 127.5 μL, remove and discard 127.5 μL of the supernatant from each well of the PCR plate.
9. Repeat step 8 one time.

   NOTE- Leave the PCR plate on the magnetic stand while performing the following 80% EtOH wash steps (10-12).

10. With the PCR plate on the magnetic stand, add 200 μL freshly prepared 80% EtOH to each well without disturbing the beads.

11. Incubate the PCR plate at room temperature for 30 seconds, and then remove and discard all of the supernatant from each well. Take care not to disturb the beads.

12. Repeat steps 10 and 11 one time for a total of two 80% EtOH washes.

13. Let the PCR plate stand at room temperature for 15 minutes to dry, and then remove the plate from the magnetic stand.

14. Resuspend the dried pellet in each well with 17.5 μL Resuspension Buffer (RSB). Gently pipette the entire volume up and down 10 times to mix thoroughly.

15. Incubate the PCR plate at room temperature for 2 minutes.

16. Place the PCR plate on the magnetic stand at room temperature for 5 minutes or until the liquid is clear.

17. Transfer 15 μL of the clear supernatant from each well of the PCR plate to the corresponding well of a new 96-well 0.3 ml PCR plate.

18. Add 15 μL of A Tailing Ligase (ATL) to the 15 μL of supernatant from previous step.

19. Briefly spin down on a standard, small benchtop centrifuge.

20. Place on the thermal cycler and run the ATAIL70 program listed below. Each well contains 30 μL.

    (a) Choose the preheat lid option and set to 100°C

        i. 37°C for 30 minutes

        ii. 70°C for 5 minutes

iii. 4°C for 5 minutes

21. Add the following reagents from the Illumina TruSeq Nano ligation kit IN ORDER:

    (a) RSB (2.5 μL)

    (b) LIG2 (2.5 μL)

    (c) DNA adapter (2.5 μL)

22. Pipette up and down, centrifuge briefly

23. Run lig program listed here in step a.

    (a) Choose the preheat lid option and set to 100°C

        i. 30°C for 10 minutes

        ii. Hold at 4°C

    (b) Add 5 μL STL to each well, and then mix w/pipette. NOTE- You can hold this mixture at -20°C overnight with no trouble.

24. Vortex Sample Purification Beads (SPB) until well dispersed.

25. Perform steps a through l using the Round 1 volumes.

    (a) Add SPB to each well, and then mix thoroughly as follows.

    i.

Round 1	42.5 μL
Round 2	50 μL

    (b) Pipette up and down.

    (c) Incubate at room temperature for 5 minutes.

    (d) Place on a magnetic stand and wait until the liquid is clear (2–5 minutes).

    (e) Remove and discard all supernatant from each well.

(f) Wash 2 times as follows.

i. Add 200 μL freshly prepared 80% EtOH to each well.

ii. Incubate on the magnetic stand for 30 seconds.

iii. Remove and discard all supernatant from each well.

(g) Use a 20 μL pipette to remove residual EtOH from each well.

(h) Air-dry on the magnetic stand for 5 minutes.

(i) Add RSB to each well.

i.

Round 1	52.5 μL
Round 2	27.5 μL

(j) Remove from the magnetic stand, and then mix thoroughly as follows. Pipette up and down.

(k) Incubate at room temperature for 2 minutes.

(l) Place on a magnetic stand and wait until the liquid is clear (2–5 minutes).

(m) CONTINUE TO STEP 26

26. ROUND1- Transfer 50 μL supernatant to the corresponding well of the CAP plate.

27. Repeat steps a through l with the new plate using the Round 2 volumes.

28. ROUND2- Transfer 25 μL supernatant to the corresponding well of the PCR plate.

SAFE STOPPING POINT If you are stopping, seal the plate and store at -25°C to -15°C for up to 7 days.

At this point, ligation and purification is done.

B.8.2 Enrichment of Post-Ligation Sample

The following is the recipe for each ligated sample:

Sample	Concentration	Volume (μL)
DNA mix		3
PPC (PCR Primer Cocktail)	10x	3
EPM (Enhanced PCR Mix)	2.5x	12
Molecular Grade Water		12
Total		30

Follow the following thermocycling protocol with the above 30 μL mixture:

[noitemsep]Choose the preheat lid option and set to 100°C 95°C for 3 minutes 8 total cycles of:[noitemsep]

- – 98°C for 20 seconds
  - – 60°C for 15 seconds
  - – 72°C for 30 seconds
- 72°C for 3 min
- Hold at 4°C

Appendix C

APPENDIX FOR CHAPTER 5

C.1 Similarity Search Universal Primer Sequences

All encoded images are encoded, then flanked with a universal primer pair to allow PCR amplification of the entire pool. Table C.1 contains the primer sequences.

Forward Primer Sequence	Reverse Primer Sequence
TACTCGCTGCGTGCAATTTA	TTGACACGTTCGGACACTTT

Table C.1: Universal primer sequences for similarity search work.

C.2 Examining Sequences After Similarity Search Encoding

After encoding with the best performing similarity search model, a 1.74 million image database is represented by 457 unique sequences. In Figure C.1, we show how many images are represented by each unique DNA sequence.

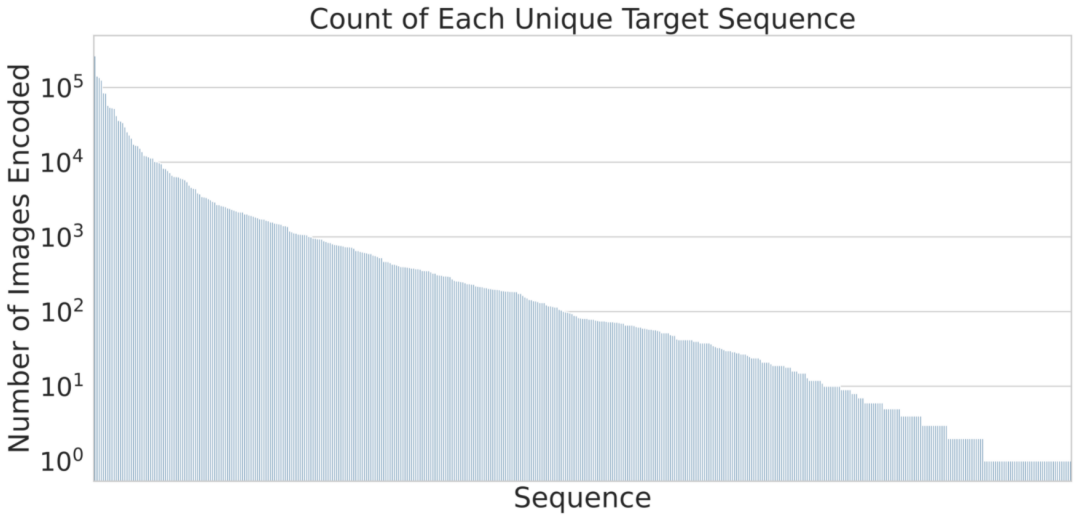


Figure C.1: Examining the Encoded Sequence Relationship to Number of Represented Images. Each column in the x-axis is a unique sequence encoded by our best performing encoder. They y-axis shows the number of images that one sequence represents.

To understand the composition of each sequence, we examine all unique sequences and report the rate each base is present at each position as shown in Figure C.2. We see the sequences are highly conserved, with only positions 2, 3, 4, 6 and 11 having high diversity.

To understand the difference in activation behavior between Cas9 models with one site or four sites (20nt or 80nt respectively), and DNA hybridization models with hybridization sites of length 20nt or 80nt, we examine activation behavior as the distance between two sequences grows as shown in Figure C.3.

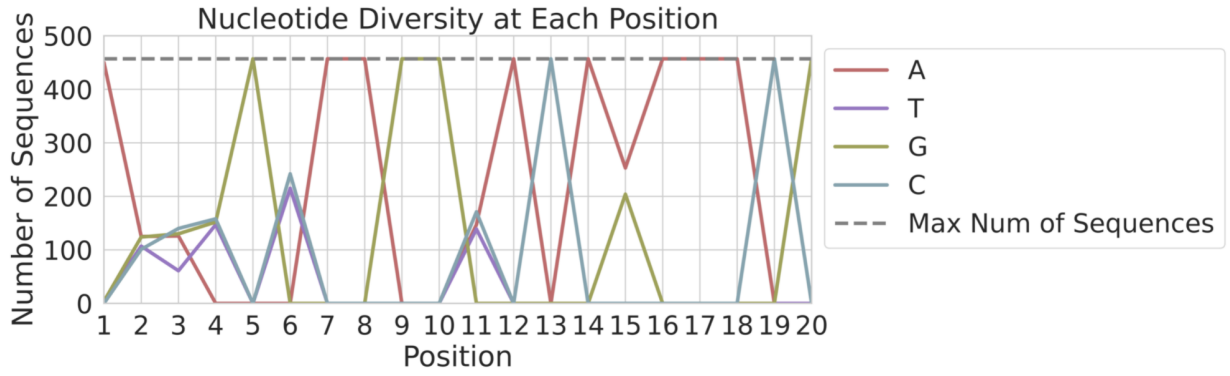


Figure C.2: Examining the Encoded Sequence Diversity. The PAM distal region is to the left and the PAM proximal region is to the right.

We generated 100,000 random sequence pairs and recorded both the the edit distance between each sequence in the pair, and the activation score the pair yielded when run through the appropriate predictor model.

For both the Cas9 and hybridization method, more sequence space yields more possible activation scores, which enables the preservation of similarity relationships (this is discussed in detail in the Results section in the main text). Furthermore, the hybridization method, regardless of the sequence space, spans a greater edit distance range than Cas9. This also supports the hypothesis presented in the main text that Cas9 behavior is too stringent, with any increase in edit distance drastically impacting Cas9 activation behavior which makes similarity search encoding difficult if not impossible at large scale.

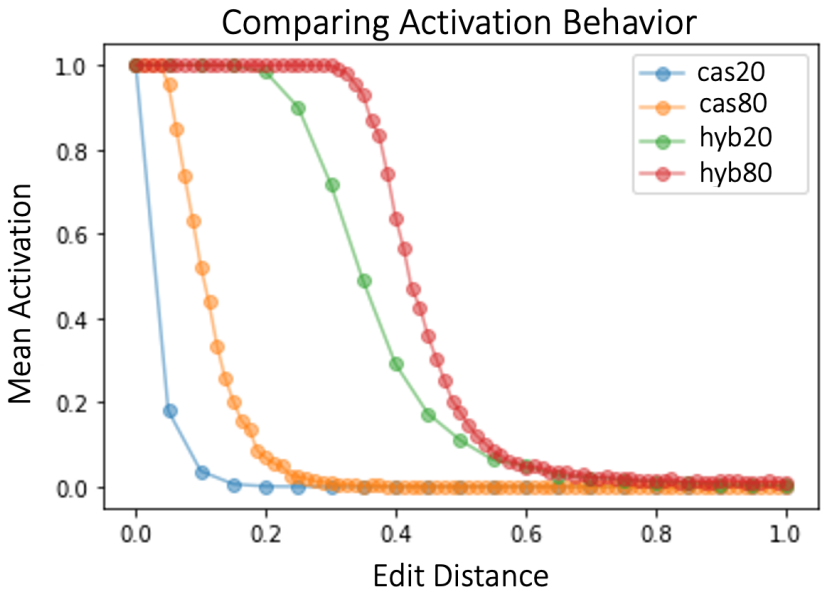


Figure C.3: Examining Activation Behavior. An edit distance of 0 indicates there are no differences between two sequences, an edit distance of 1 indicates all positions in a sequence are different between two sequences. An activation of 0 indicates no activity (Cas9 cleavage of a sequence, or the hybridization of two sequences, respectively) while a score of 1 indicates maximum activity.

Appendix D

APPENDIX FOR CHAPTER 6

D.1 Performing and Analyzing qPCR

D.1.1 Primer Sequences

Each file has a unique primer pair, facilitating random access in DNA data storage. The primer sequences are in Table D.1.1.

File	Forward Primer	Reverse Primer
ID8	TTCGTTCGTCGTTGATTGGT	ACAAACTCATGGCTCCGTTT
ID15	ACATTCCGTGCCATTGGATT	TTTGTGGAACGATTTGCCGA
Post-Ligation	AATGATACGGCGACCACCGA	CAAGCAGAAGACGGCATACG

Table D.1: The forward and reverse primers needed to amplify each file, as well as the primer pair used to enrich the sequences after ligation, but prior to aging.

The ultramer used to quantify the UW (non-ETH) samples after degradation utilized the post-ligation primers. The ultramer sequence is:

5' AATGATACGGCGACCACCGAGATCTTGACCAAACACTCTTTCCCTACAC
GACGCTCTTCCGATCTACTGACTGAGTACGACATGCATGCAGCCGTCCATA
GCCTTGTTCGTTGTCATAGTATGTAGCTACTGCACGTATGCATACTGAGTC
TGTACATGAGTAGTGCACAGTAGTAGACGCGTCATCAGAGTATGCATATCA
GCATCGCGTAGCTAGCTAGGCGGAAACGTAGTGAAGGTAGATCGGAAGAGC
ACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG 3'

D.1.2 Primer Sequences for the DNAStable PCR Samples

Here, Table D.1.2 and Table D.1.2 show the overhanging PCR primers used to prepare samples in the *DNAStable PCR* aging condition after random access, prior to sequencing. Thus, the samples did not have to be ligated again.
The spaces within the sequences listed in the tables are to help the reader separate the different domains of the primer sequence, which are: Illumina universal flow cell adapter, Illumina index, Illumina sequencing primer, file 8/15 primer.

Primer Description	Sequence
Forward, ID8, index H	CAATGATACGGCGACCACCGAGATCTACACGTACTGACACACTCTTTCCCTACA CGACGCTCTTCCGATCTTTCGTTCGTCGTTGATTGGT
Reverse, ID8, index H4	CAAGCAGAAGACGGCATACGAGGGAATCTCGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H5	CAAGCAGAAGACGGCATACGAGTTCTGAATGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H6	CAAGCAGAAGACGGCATACGAGACGAATTCGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H7	CAAGCAGAAGACGGCATACGAGAGCTTCAGGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H8	CAAGCAGAAGACGGCATACGAGGCGCATTAGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H9	CAAGCAGAAGACGGCATACGAGCATAGCCGGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H10	CAAGCAGAAGACGGCATACGAGTTCGCGGAGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H11	CAAGCAGAAGACGGCATACGAGGCGCGAGAGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT
Reverse, ID8, index H12	CAAGCAGAAGACGGCATACGAGCTATCGCTGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCACAAACTCATGGCTCCGTTT

Table D.2: The forward and reverse primers needed to attach the Illumina indices, sequencing primer, and adapters to file 8 via PCR, only utilized for the DNAStable + PCR aging method since these samples were aged, then PCR'ed with the short primers, then prepped for sequencing using these overhang PCR sequences.

Primer Description	Sequence
Forward, ID15, index H	AATGATACGGCGACCACCGAGATCTACACGTACTGACACACTCTTTCCCTACA CGACGCTCTTCCGATCTACATTCCGTGCCATTGGATT
Reverse, ID15, index H4	CAAGCAGAAGACGGCATACGAGGGAATCTCGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H5	CAAGCAGAAGACGGCATACGAGTTCTGAATGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H6	CAAGCAGAAGACGGCATACGAGACGAATTCGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H7	CAAGCAGAAGACGGCATACGAGAGCTTCAGGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H8	CAAGCAGAAGACGGCATACGAGGCGCATTAGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H9	CAAGCAGAAGACGGCATACGAGCATAGCCGGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H10	CAAGCAGAAGACGGCATACGAGTTCGCGGAGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H11	CAAGCAGAAGACGGCATACGAGGCGCGAGAGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTGTGGAACGATTTGCCGA
Reverse, ID15, index H12	CAAGCAGAAGACGGCATACGAGCTATCGCTGTGACTGGAGTTCAGACGTGT GCTCTTCCGATCTTTGTGGAACGATTTGCCGA

Table D.3: The forward and reverse primers needed to attach the Illumina indices, sequencing primer, and adapters to file 15 via PCR, only utilized for the DNAStable + PCR aging method since these samples were aged, then PCR'ed with the short primers, then prepped for sequencing using these overhang PCR sequences.

D.2 Determining a Coverage Threshold

The purpose of determining a coverage threshold is to exclude files that have so few average reads that the data could inaccurately reflect the number of strands that have been degraded. In other words, a strand that is missing is more likely to be due to random chance of not being read rather than actually having been degraded. To determine a threshold of when to assume that files might more likely be "missing" from random chance rather than being degraded, a plot of average file coverage versus percent missing was created. Only time point zero files (from all storage conditions) were included to determine the coverage threshold. The coverage threshold is the same for both id8 and id15 files. Figure D.1 displays how an average coverage of 14 and beyond had a relatively low percent of strands missing from reads:

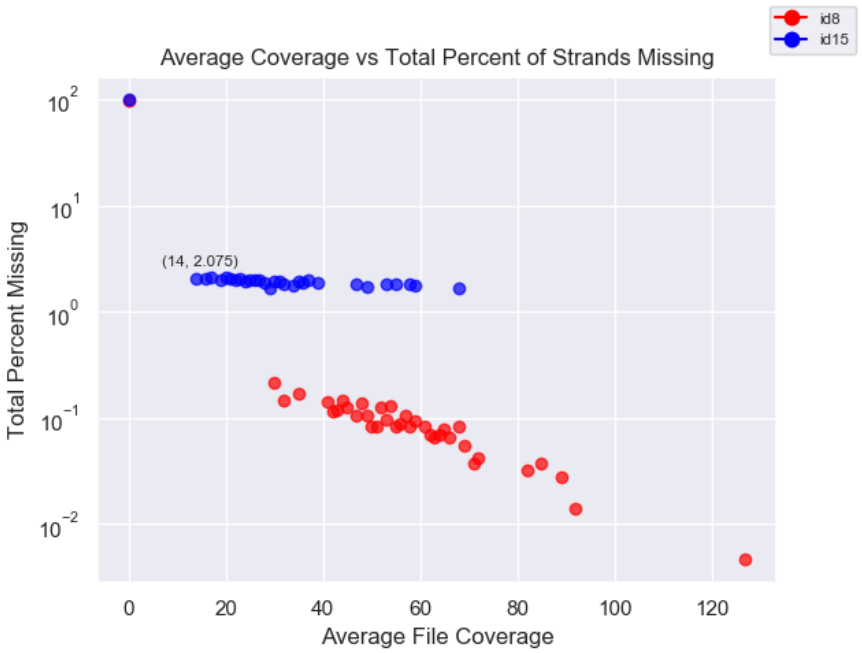


Figure D.1: Relationship between sequencing coverage and sequences missing.

All files were subsampled to the determined coverage threshold, which is an average of 14 reads per sequence. Every file with greater than or equal to an average coverage of 14 was

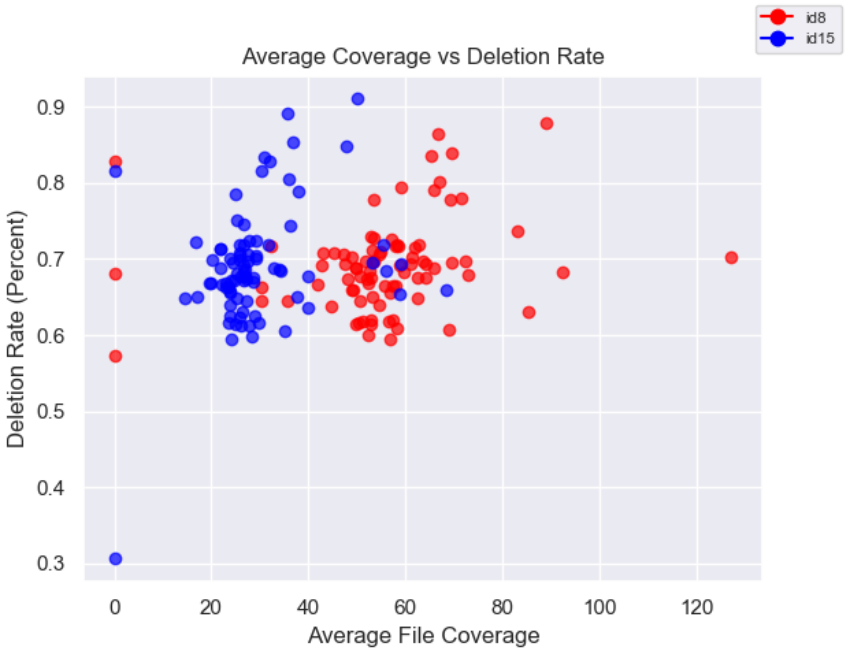


Figure D.2: Relationship between sequencing coverage and deletion error rate.

subsampled 100 times, and the average coverage for each sequence was calculated and used in all subsequent analysis.

D.2.1 Insertion/Substitution/Deletion Rates as a Function of Coverage

Average coverage versus insertion/deletion/substitution rate were plotted to ensure no correlation between average coverage and error rate. Like with determining coverage threshold, only time point zero data are plotted. We find no practically significant relationship between coverage and error rate (Figures D.2, D.3, and D.4).

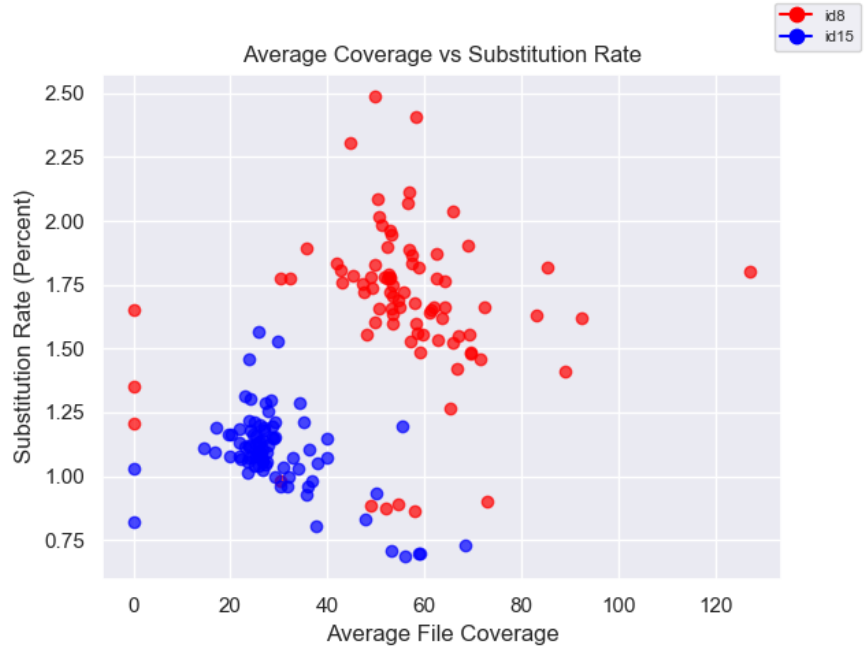


Figure D.3: Relationship between sequencing coverage and substitution error rate.

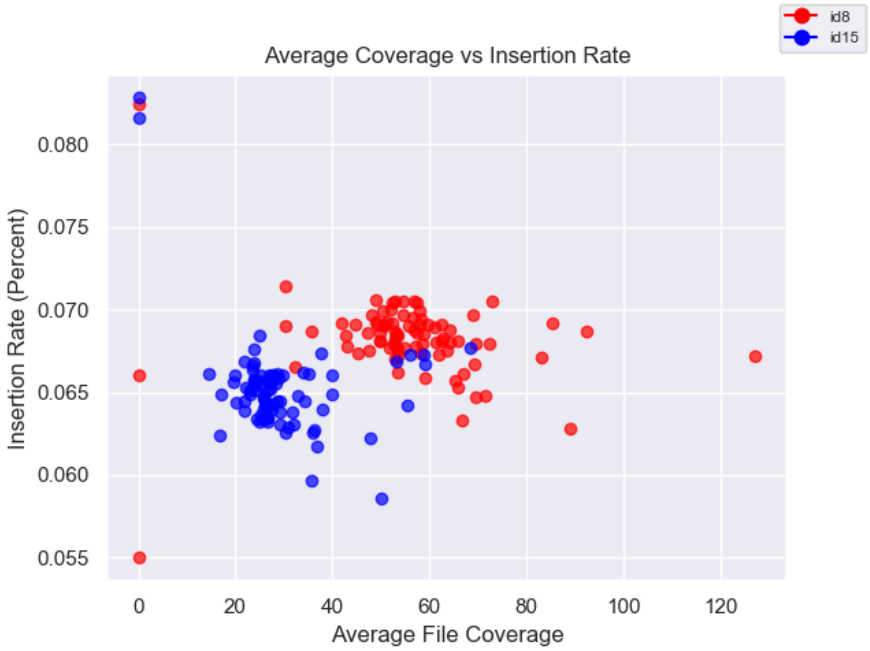


Figure D.4: Relationship between sequencing coverage and insertion error rate.

D.3 Sequence behavior analysis

Here, we present the sequence behavior observed after next-generation sequencing, as it is important to know if any particular types of sequences degrade faster than others, or if sequence degradation is stochastic in nature.
We analyze both the absence of sequences in varying levels of detail, as well as individual sequence composition in the form of a trimer analysis.

D.3.1 Percent of sequences missing over time

Here, we show that the percent of sequences missing is not dependent on degradation. All samples that had enough material (as measured by the qPCR) to directly sequence were sequenced.
As shown in the main text, the percent of sequences missing is constant. Here, we present each preservation method individually in Figures D.5 to D.13. Note that several times the data are so similar not all conditions are visible because they are stacked so closely on top of one another.
That the percent of sequences missing constant shows degradation likely does not impact certain sequences more than others, as those hypothetical "degradation-prone" sequences would have degraded faster than the surrounding material and would not have been sequenced.

D.3.2 A sequence once lost is lost forever?

One might think a DNA sequence not seen by the sequencer in one time point is never recovered in all subsequent time points, but this is not the case. Typically, a sequence will be missing in one sample only to reappear later (as shown in the supplemental CSV files ending in "missing sequences analysis" available with the published work [72]). This is likely due to the stochastic nature of sampling material to sequence. The fact that sequences tend not to stay missing supports the hypothesis that DNA degradation for these

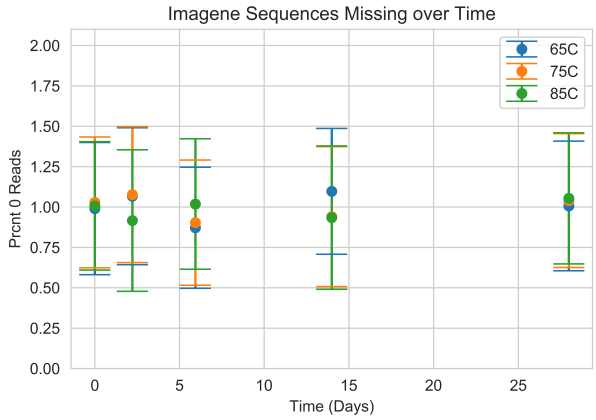


Figure D.5: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

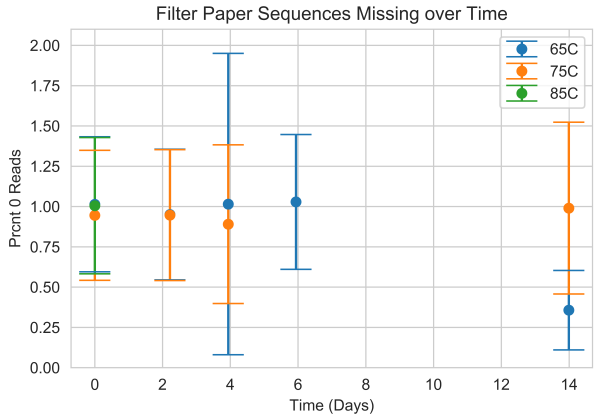


Figure D.6: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

DNA sequences is stochastic.

For each temperature and time point, we present the data in three columns.

"Total percent missing" is calculated by finding the set of sequences that have zero coverage in ANY one of the triplicates and dividing that number by the number of total sequences (21,601 or 7,373, depending on the file being analyzed).

The "percent missing that reappear" examines all the sequences that are in the "total

Figure D.7: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

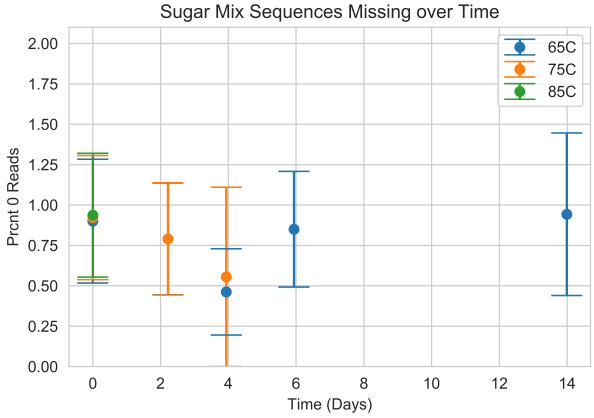


Figure D.8: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

percent missing" set and determines what percent of those sequences are found either in another triplicate in the same time point, or a subsequent time point.

The "percent that stay missing" is calculated by finding the set of sequences that have zero coverage in all replicates in the examined time point as well as all subsequent time points, then dividing that number by the number of total sequences (21,601 or 7,373).

NaN ("not a number") in the table means the file was not sequenced to sufficient

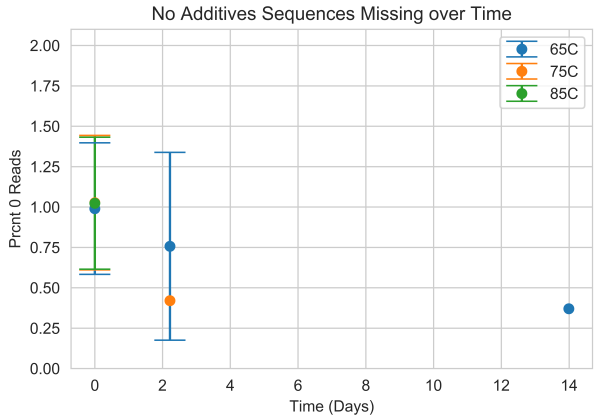


Figure D.9: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

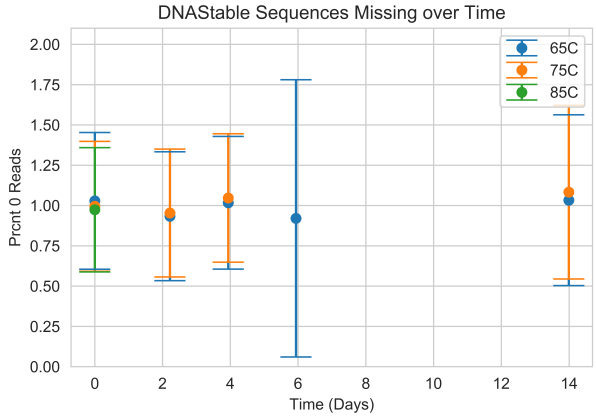


Figure D.10: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

sequencing depth at the given condition, had only one replicate, or the subsequent time point was not sequenced to sufficient sequencing depth and therefore we could not determine if the sequences stayed missing or not.

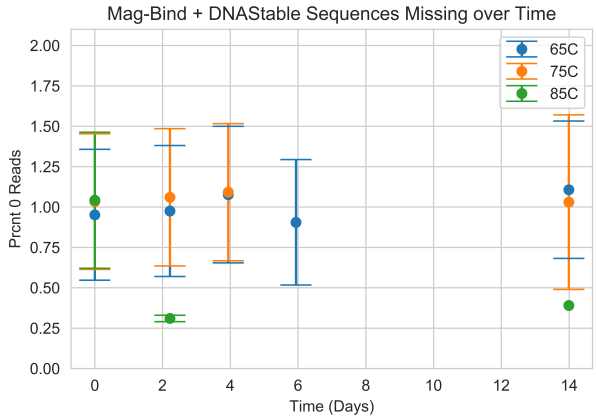


Figure D.11: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

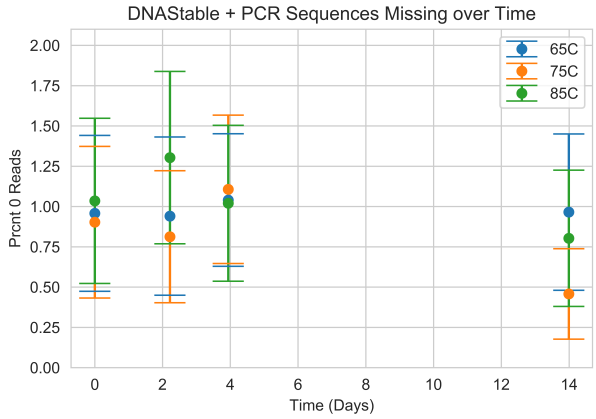


Figure D.12: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

D.3.3 Trimer analysis

In this analysis shown below in Tables D.4 and D.5, we look at each trimer in each DNA sequence. For example, the sequence "ATCG" has the trimers "ATC" and "TCG". In the last column, we also examine the sequence's overall GC ratio.

We then look at each sample of DNA that was degraded and examine two groups of sequences: First, the sequences that were most sequenced (the top 5th percentile) and

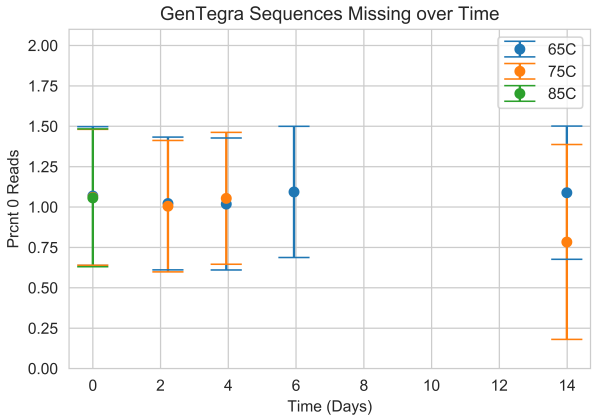


Figure D.13: Sequences missing from each time point and temperature for all triplicates, where error bars are the standard error from the mean.

second, the sequences that were least sequenced (the bottom 5th percentile, including missing sequences).

We examined whether or not those two groups of sequences were statistically different with a t-test. By definition, this yielded many hundreds of p-values as there is one p-value for each trimer in each file in each method in each time point, temperature, and replicate. Thus, we created a summary table in which we can clearly see that no trimer is statistically significantly different in all conditions, and the rates can vary quite a lot between the two files files aged in this experiment. The GC content metric is also not statistically significant in all conditions. Note that the primer sequences were not included in analysis here because they are constant for each file.

Even in instances where there is a statistically significant difference in the trimer presence after aging, for example, 92% for the trimer "AGA" when looking at file 15 data, the number is quite different for file 8 where the trimer "AGA" yields 74%. Or take the trimer "AGT" with percentages of 62 for File 15 and 18 for File 8. This supports the hypothesis that there is no clear relationship between a particular trimer and increased degradation.

ACA	ACG	ACT	AGA	AGC	AGT	ATA	ATC	ATG
92	18	14	92	47	18	91	8	2

CAC	CAG	CAT	CGA	CGC	CGT	CTA	CTC	CTG
35	70	50	14	39	42	4	88	92

GAC	GAG	GAT	GCA	GCG	GCT	GTA	GTC	GTG
21	80	33	9	21	85	32	57	64

TAC	TAG	TAT	TCA	TCG	TCT	TGA	TGC	TGT
40	73	16	1	78	78	0	87	70

GC:AT
85

Table D.4: The percentage of significant t-test results comparing the most prevalent sequences to the least prevalent sequences across all conditions for file 8.

ACA	ACG	ACT	AGA	AGC	AGT	ATA	ATC	ATG
81	65	2	74	13	62	83	4	55

CAC	CAG	CAT	CGA	CGC	CGT	CTA	CTC	CTG
67	67	27	6	6	60	4	74	72

GAC	GAG	GAT	GCA	GCG	GCT	GTA	GTC	GTG
11	18	65	72	13	79	60	72	79

TAC	TAG	TAT	TCA	TCG	TCT	TGA	TGC	TGT
30	69	11	0	81	74	0	74	81

GC:AT
81

Table D.5: The percentage of significant t-test results comparing the most prevalent sequences to the least prevalent sequences across all conditions for file 15.

D.4 Overview Table

An comprehensive comparison of each preservation method. The times listed to encapsulate/dehydrate and de-encapsulate/re-hydrate serve to show the order of magnitude in which these processes currently take.
Note that Imagene does not list a half-life due to the fact that its samples degraded so little over time that no accurate half-life could be derived.
The DNAStable sample aged by ETH and the samples with no preservation other than dehydration degraded too quickly to measure adequately and so no accurate half-life could be derived.

Method	Time to dehydrate and encapsulate	Half life per nt at 20°C (years)	Order of time to rehydrate/de-encapsulate sample
Imagene	minutes	N/A	seconds
Trehalose	minutes	4,900	seconds
GenTegra	minutes	3,000	seconds
ETH Magnetic NP	minutes	2,300	minutes
Mag-Bind + DNAStable	minutes	360	minutes
Sugar Mix	minutes	230	seconds
ETH Trehalose	minutes	170	seconds
Filter Paper	minutes	100	seconds
DNAStable + PCR	minutes	45	seconds
DNAStable	minutes	19	seconds
ETH DNAStable	minutes	N/A	seconds
No Additives	N/A	N/A	seconds

Table D.6: An overview of the storage methods explored.

D.5 Ligation Protocol

The ligation protocol used is a combination of the Illumina TruSeq ChIP Sample Preparation protocol and the Illumina TruSeq Nano Library Prep protocol, using the Illumina TruSeq Nano reagents. The following are the step-by-step directions used in this work.

1. Add 40 μL ERP2 to each well with 60 μL of PCR product

2. Pipette up and down to mix

3. Place on the thermal cycler and run the ERP program (30 min at 30°C) then place on ice. Each well contains 100 μL.

4. Add 160 μL well-mixed AMPure XP Beads to each well of the PCR plate containing 100 μL End Repair Mix.

5. Gently pipette the entire volume up and down 10 times to mix thoroughly.

6. Incubate the PCR plate at room temperature for 15 minutes.

7. Place the PCR plate on a magnetic stand at room temperature for 15 minutes or until the liquid is clear.

8. Using a 200 μL single channel or multichannel pipette set to 127.5 μL, remove and discard 127.5 μL of the supernatant from each well of the PCR plate.

9. Repeat step 8 one time.

   NOTE- Leave the PCR plate on the magnetic stand while performing the following 80% EtOH wash steps (10-12).

10. With the PCR plate on the magnetic stand, add 200 μL freshly prepared 80% EtOH to each well without disturbing the beads.

11. Incubate the PCR plate at room temperature for 30 seconds, and then remove and discard all of the supernatant from each well. Take care not to disturb the beads.

12. Repeat steps 10 and 11 one time for a total of two 80% EtOH washes.

13. Let the PCR plate stand at room temperature for 15 minutes to dry, and then remove the plate from the magnetic stand.

14. Resuspend the dried pellet in each well with 17.5 μL Resuspension Buffer (RSB). Gently pipette the entire volume up and down 10 times to mix thoroughly.

15. Incubate the PCR plate at room temperature for 2 minutes.

16. Place the PCR plate on the magnetic stand at room temperature for 5 minutes or until the liquid is clear.

17. Transfer 15 μL of the clear supernatant from each well of the PCR plate to the corresponding well of a new 96-well 0.3 ml PCR plate.

18. Add 15 μL of A Tailing Ligase (ATL) to the 15 μL of supernatant from previous step.

19. Briefly spin down on a standard, small benchtop centrifuge.

20. Place on the thermal cycler and run the ATAIL70 program listed below. Each well contains 30 μL.

    (a) Choose the preheat lid option and set to 100°C

        i. 37°C for 30 minutes

        ii. 70°C for 5 minutes

iii. 4°C for 5 minutes

21. Add the following reagents from the Illumina TruSeq Nano ligation kit IN ORDER:

    (a) RSB (2.5 μL)

    (b) LIG2 (2.5 μL)

    (c) DNA adapter (2.5 μL)

22. Pipette up and down, centrifuge briefly

23. Run lig program listed here in step a.

    (a) Choose the preheat lid option and set to 100°C

        i. 30°C for 10 minutes

        ii. Hold at 4°C

    (b) Add 5 μL STL to each well, and then mix w/pipette. NOTE- You can hold this mixture at -20°C overnight with no trouble.

24. Vortex Sample Purification Beads (SPB) until well dispersed.

25. Perform steps a through l using the Round 1 volumes.

    (a) Add SPB to each well, and then mix thoroughly as follows.

    i.

Round 1	42.5 μL
Round 2	50 μL

    (b) Pipette up and down.

    (c) Incubate at room temperature for 5 minutes.

    (d) Place on a magnetic stand and wait until the liquid is clear (2–5 minutes).

    (e) Remove and discard all supernatant from each well.

(f) Wash 2 times as follows.

i. Add 200 μL freshly prepared 80% EtOH to each well.

ii. Incubate on the magnetic stand for 30 seconds.

iii. Remove and discard all supernatant from each well.

(g) Use a 20 μL pipette to remove residual EtOH from each well.

(h) Air-dry on the magnetic stand for 5 minutes.

(i) Add RSB to each well.

i.

Round 1	52.5 μL
Round 2	27.5 μL

(j) Remove from the magnetic stand, and then mix thoroughly as follows. Pipette up and down.

(k) Incubate at room temperature for 2 minutes.

(l) Place on a magnetic stand and wait until the liquid is clear (2–5 minutes).

(m) CONTINUE TO STEP 26

26. ROUND1- Transfer 50 μL supernatant to the corresponding well of the CAP plate.

27. Repeat steps a through l with the new plate using the Round 2 volumes.

28. ROUND2- Transfer 25 μL supernatant to the corresponding well of the PCR plate.

SAFE STOPPING POINT If you are stopping, seal the plate and store at -25°C to -15°C for up to 7 days.

At this point, ligation and purification is done.

The following is the post-ligation recipe for each ligated sample:

Sample	Concentration	Volume (μL)
DNA mix		3
Forward post-ligation primer (see Appendix D.1)	10μM	1.5
Reverse post-ligation primer (see Appendix D.1)	10μM	1.5
Kapa HiFi 2x	2.5x	12
Molecular Grade Water		12
Total		30

Follow the following thermocycling protocol with the above 30 μL mixture:

- Choose the preheat lid option and set to 100°C
- 95°C for 3 minutes
- 8 total cycles of:
  - 98°C for 20 seconds
  - 60°C for 15 seconds
  - 72°C for 30 seconds
- 72°C for 3 min
- Hold at 4°C

D.6 Arrhenius Equation

Assuming DNA degradation is a first-order decay process, then the rate of the reaction is dependent on the rate constant, k, and the concentration of the DNA, $[DNA]$.

$$r = k[DNA] \tag{D.1}$$

From equation (1), we can derive the relationship between k and the initial and end concentrations of DNA for a given time, t.

$$ln[DNA] = -kt + ln[DNA]_0 \tag{D.2}$$

The temperature dependence of k can be modeled by the Arrhenius equation, where A is the pre-exponential factor, Ea is the activation energy, R is the gas constant, and T is temperature.

$$k = Ae^{\frac{-E_a}{RT}} \tag{D.3}$$

Taking the natural logarithm of equation (3), we can calculate Ea.

$$ln[k] = \frac{-E_a}{RT} + ln[A] \tag{D.4}$$

The half-life of the DNA ($t_{1/2}$) can be calculated by solving equation (2), where at $t_{1/2}$ the concentration of DNA is half of the original concentration.

Appendix E

APPENDIX FOR CHAPTER 7

E.1 Calculating the Odds of a Meaningful STOP Codon

Here, we will calculate the odds of finding a meaningful STOP codon in a random sequence of 200nt. We will only consider the standard STOP codons TAG, TAA, and TGA.

- Given any random sequence of three nucleotides, there are 4^3 (64) possible combinations, thus the odds of seeing any one of our STOP codons is 3/64.

- For a sequence of length 150 (the typical length of sequences presented in this dissertation), we have $\frac{148}{3}$ (49.33) blocks of 3 nucleotides. Note that there are only 148 blocks of 3 nucleotides, as the last two positions would be stunted blocks of two and one nucleotide. The number of blocks we'd expect to see be a STOP codon is thus $\frac{148}{3} \cdot \frac{3}{64}$ (2.3).

Note that this actually means we can expect to see approximately two STOP codons per reading frame. When DNA is read, it is read in one of three reading frames. For example, the sequence AAGTAATCCAA could be read as:

1. AGG—TAA—TCC—AA: in which a STOP codon is seen after one amino acid

2. A—AGT—AAT—CCA—A: in which no STOP codons are seen

3. AA—GTA—ATC—CAA: in which no STOP codons are seen

So a typical DNA data storage sequence of 150nt would actually contain $2.3 \cdot 3$ (6.9) total STOP codons per sequence, though likely only approximately two would be relevant for any given reading frame.