ABSTRACT

TOMEK, KYLE JOHN. Addressing Practical Barriers to Extreme-Scale DNA-based Data Storage Systems. (Under the direction of Dr. Albert J. Keung).

The clear need to increase data storage capacities and mitigate the exponential rise in materials, space, and energy demands of information storage have stimulated interest in the development of DNA as a data storage medium. DNA holds significant promise due to its density, durability, and resource and energy conservation. While gigabyte-scale DNA-based systems have been demonstrated, there remain challenges in scaling systems to the capacities necessary for a transformative data storage solution. Fundamental obstacles to data organization, file retrieval, and DNA synthesis arise from the fact that as systems continue to scale, DNA databases will become increasingly complex, crowded, and physically disordered. Here we develop scalable methods to organize and access files stored in DNA, harness off-target molecular interactions to increase system functionality, experimentally investigate file address interactions, and explore enzymatic DNA assembly methods for constructing strands for data storage.

Existing DNA data storage systems have few enough strands to be completely read by modern DNA sequencing technologies. Eventually, high-capacity systems will no longer be able to be sequenced entirely, nor will lower-latency systems with smaller capacities (e.g., semiconductor-based systems) be able to process entire DNA databases. **Chapter 1** starts by addressing how to specifically access individual files from complex databases. We use chemical handles to extract unique files from a 5 TB background database. Additionally, we implement this technology in a microfluidic device capable of automation. These advancements enable the development and future scaling of DNA-based data storage systems with modern capacities through augmented file access capabilities.

High-capacity DNA storage systems will require many available file addresses for data organization. However, as systems scale-up, the probability for off-target biomolecular interactions increases. Consequently, addresses must be sufficiently different from each other in sequence and are, therefore, finite in number and a limiting factor of system capacities. **Chapter 1** also discusses the design and application of a file address scheme that uses file addresses multiple times in hierarchical combination to increase the maximum capacity of DNA storage systems by five orders of magnitude. In **Chapter 2** we exploit underutilized file addresses and leverage thermodynamic tuning of biomolecular interactions to create useful data access and organizational

features. Specific reaction conditions including temperatures, reagent compositions, and DNA concentrations were screened for their ability to controllably access DNA strands encoding complete image files or subsets of those strands encoding low-resolution portions. We demonstrate this using four JPEG images in a GB-sized background database and provide an argument for the economic benefit of this generalizable data organization strategy.

Chapter 3 seeks to further understand DNA interactions through the development of a high-throughput experimental strategy to screen many combinations of variable sequences. Specifically, we uncover biased sequence interactions during DNA ligation, test a polymerase-based reaction for screening interactions, and describe plans to explicitly investigate DNA hybridization. These platforms will not only identify useful sequences for DNA storage systems, but also inform computational primer design models.

Synthetic DNA used for data storage is predominantly created using phosphoramidite chemistry which is limited to base-by-base synthesis of ~300mer oligonucleotides and is only scalable by the reaction surface. In **Chapter 4** we design and implement multiple enzymatic DNA assembly reactions which use short oligonucleotide ‘codewords’ as data building blocks. These methods will allow for the synthesis and storage of codewords at massive scale as feedstocks for economical, enzyme-based DNA strand assembly for data storage.

These key innovations unlock the potential for DNA storage systems to scale to extreme capacities with improved functionalities and set the stage for the broader incorporation of molecular and synthetic biology techniques in engineering DNA databases.

Addressing Practical Barriers to Extreme-Scale DNA-based Data Storage Systems

by
Kyle John Tomek

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Chemical Engineering

Raleigh, North Carolina

2021

APPROVED BY:

Dr. Albert Keung Committee Chair	Dr. Robert Kelly
Dr. Cranos Williams	Dr. Joshua Pierce

DEDICATION

To Adriel, my wife, best friend, better half, my everything.

You and me together, we can do anything.

BIOGRAPHY

Kyle Tomek was born on May 31, 1990, and grew up in Flushing, MI. He attended high school at Flint Powers Catholic where he met his future wife, Adriel, while competing on the alpine ski team. He went on to study Human Biology and Chemical Engineering and joined the water ski team at Michigan State University. During college, Kyle discovered an interest in all things science and research in the lab of Dr. Timothy Whitehead. Here, he was excited to cultivate many fundamental lab skills and went on to develop methods to increase biomass fermentation efficiency through protein and reaction engineering. He then interned at Zoetis Inc. under the guidance of Dr. Lisa Bergeron where he engineered monoclonal antibodies in support of veterinary immunogenic and pharmacokinetic assays. Upon graduation from MSU, Kyle worked as an Engineer at Wolverine Fire Protection where he designed fire suppression sprinkler systems for data server rooms, cooling towers, and coal and nuclear power plants. His scientific curiosity was sparked once again when he learned to homebrew and about the biological processes that make craft beer so good. He knew a career in research was the right path for him, and upon discovering the strong biomolecular focus NC State had to offer, he decided to pursue a PhD in Chemical Engineering at NC State in Raleigh, NC. He quickly felt at home while discussing projects with Dr. Albert Keung and was excited to join his lab to study the scalability of DNA-based data storage systems in collaboration with Dr. James Tuck. In the mountains of North Carolina Kyle was introduced to rock climbing, a sport he and his wife learned and enjoy together. In addition to the beauty of the mountains, he loves the challenge of the sport, both mentally and physically. It has helped him overcome mental barriers and has reinforced his love and respect for the Earth and his community. During his graduate studies, Kyle particularly enjoyed teaching, lecturing, and mentoring students, especially the younger student researchers he worked with in lab. He is excited to follow his passion in the next phase of life, the establishment of DNAli Data Technologies. Dr. Albert Keung, Dr. James Tuck, and Kyle have co-founded DNAli to bring DNA storage to the world as a practical data storage solution.

ACKNOWLEDGMENTS

First, I'd like to thank my advisor Dr. Albert Keung. Our meetings where we would brainstorm new experiments and projects are some of my favorite moments from the past 5 years. I always left with a sense of clarity and a better feeling about where my projects were headed. Your work ethic, efficiency, and devotion to conducting only the purest science are remarkable and each make you an impressive scientist and mentor. Most of all, thank you for your unwavering support and confidence in me as I've grown into the scientist I am today.

To Dr. Joshua Emrick, my oldest friend and greatest, most consistent mentor. Your continuous drive for excellence and work ethic are inspiring. Thank you for being my toughest competition and biggest rival in all our games, hobbies, and sports – we've helped each other become Jacks of all trades, much to the chagrin of Grandma G. Thank you for being a sounding board for my scientific curiosities and for helping me through struggles with experiments and projects. I love that we can be beer and career consultants for each other and look forward to continuing to grow together as scientists and brothers.

Thank you to Dr. James Tuck, your group is the 1 to our 0. Your perspective is refreshing and has been instrumental in further guiding my thought process in DNA storage experiments. Your demand for concise clarity has made me a better scientist and presenter. I'm excited to continue our collaboration and for all the great things ahead for DNAli.

To the DNA data storage subgroup, our biweekly meetings planted the seeds for so many new, exciting ideas. Thank you for your curiosities, feedback, brainstorms, and help with optimizations and troubleshooting. To the Tuck group members, thank you for constantly fact checking my knowledge of computer systems, designing countless oligos, encoding and decoding all our files, and for all the computation and analysis – this work could not have been so impactful without you.

To the members of the Keung lab, you each had a part in making this project into what it is here. Thank you for providing critical feedback during countless practice talks, memos, roundtable discussions and for the judgement free zone for all my random ideas, questions, and curiosities. Your own groundbreaking projects, hard work, and determination have inspired me to strive for greatness in my own work and your demand for scientific rigor and difficult questions have made journal reviewers seem easy.

I am incredibly grateful to have mentored four outstanding undergraduate and high school students throughout my project. Lainey, Austin, Zach, and Antonio, I'm fortunate to have worked with such talented, smart, and hard-working scientists. I hope you learned as much from me as I have from each of you.

I would like to thank my committee members: Dr. Robert Kelly, Dr. Cranos Williams, Dr. Joshua Pierce, and Dr. Gregory Reeves for their guidance. Your pragmatic questions and demand for clarity helped me nail down the important details of my experiments while also focusing on the bigger picture of my entire project and its place in the field.

To the NC State Office of Research Commercialization, especially Kultaran Chohan, thank you for the support throughout this project—from patent applications, through the I-Corps program, to helping us start DNAli and your guidance towards funding opportunities and business mentorship. To the administrative support staff in the CBE offices, thank you for making all of my learning and research possible. I couldn't have navigated program requirements, course registrations, or completed any orders for reagents without you.

Thank you to the Graduate Assistance in Areas of National Need (GAANN) Fellowship Program in Molecular Biotechnology for the financial support as well as the scientific training and professional development opportunities throughout my time at NC State.

To my undergraduate advisor, Dr. Timothy Whitehead, thank you for helping build a strong foundation in research and for the helpful letters of recommendation defining me as a well-rounded candidate—laser tag skills and all.

To Dr. Lisa Bergeron, thank you for the industrial research experience and for helping plant the seed of a PhD by bluntly asking "why not?" when I said I wasn't going to go to grad school.

Mom and Dad, thank you for making me finish my homework every day before going to play and for making me choke down all of my tomato soup before going to watch TV; my work ethic seemingly originated during events like those. More importantly, thank you for instilling the manners, morals, and values that have made me the man that I am today.

To my sisters, Anna and Ellen, when is your beatification for dealing with me as a kid? For real though, you both were and remain to be amazing role models and inspirations to me daily.

To Grandma Aggie, Pops, Grandma G, and Papa Frank, thank you for your unconditional love and for imparting your genuine values as I've grown up. Your hard work, perseverance, and determination serve as a gold standard for me to strive towards.

To the Egner Family, MomE, PappE, Alyse, Dane and now Nick, I'm so fortunate to have 2 sets of parents and siblings. Thank you for loving me like a son and a brother and for putting up with, and even encouraging, all of my shenanigans.

To friends who have provided support and much needed distractions along the way. To the NC climbing community and my climbing crew, thank you for all of the beta, soft catches, encouragement, and endless psych; your hard work, attention to detail, and adventurous spirits are a daily motivation. To the CBE Basketball players, thanks for being the first to welcome me to the department - Tuesday night 3on3 grad/faculty/staff division champs for life!! To my kickball team, for teaching me that I should bring a competitive ferocity to everything I do, even a children's game.

And to Adriel, thank you for indulging me when I'm analyzing anything and everything to try and figure out how and why it works. Your presence during this project has made celebrating the victories so much sweeter and enduring the difficult times that much easier. Your love, confidence in me, extraordinarily difficult questions, and unwavering encouragement make me the best version of myself. I cannot imagine where I would be or what my life would be like without your love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Driving the scalability of DNA-based information storage systems

Kyle J. Tomek[1,4], Kevin Volkel[2,4], Alexander Simpson[2], Austin G. Hass[1,3], Elaine W. Indermaur[1], James M. Tuck[2,*], and Albert J. Keung[1,*]

1. Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27606

2. Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606

3. Department of Structural and Molecular Biochemistry, North Carolina State University, Raleigh, NC 27606

4. These authors contributed equally to this work

*Correspondence (jtuck@ncsu.edu, ajkeung@ncsu.edu)

Abstract

The extreme density of DNA presents a compelling advantage over current storage media; however, to reach practical capacities, new systems for organizing and accessing information are needed. Here, we use chemical handles to selectively extract unique files from a complex database of DNA mimicking 5 TB of data and design and implement a nested file address system that increases the theoretical maximum capacity of DNA storage systems by five orders of magnitude. These advancements enable the development and future scaling of DNA-based data storage systems with modern capacities and file access capabilities.

Introduction

DNA is an excellent candidate for archival data storage as it offers high raw information density as well as durability and energy efficiency[1–4]. Motivated by these compelling properties, pioneering work has tackled many important features needed for a DNA storage system. For example, encoding and decoding algorithms have been developed to be tolerant to errors while also being highly efficient in terms of density and computational intensity[5–13]. Strategies such as nested polymerase chain reaction (PCR) architectures have also been proposed to increase the number of file addresses in storage systems[14]. In addition, molecular manipulations have been developed to access files through PCR amplification[8,9,11,15,16], encrypt and rewrite information using PCR and Sanger sequencing[8,17], and implement DNA-based computations or search functionalities through extraction of specific DNA strands using biotin-functionalized DNA oligomers[18,19]. Further accelerating the field, a recent implementation of a 200 MB DNA storage system demonstrated that current DNA synthesis technologies are already capable of reasonable modern storage capacities[15].

Given these rapid advancements in DNA storage, it is timely to anticipate the challenges that will arise as systems continue to scale in capacity and density. Broadly encompassing these challenges is the fact that as systems continue to scale, DNA databases will become ever more diverse, crowded, and physically disordered, thus posing inherent barriers to data organization and retrieval. This analysis can be broken down further into specific issues. For example, existing systems have few enough strands to be completely read by modern DNA sequencing technologies; in contrast, future high-capacity systems will not be able to be sequenced in their entirety (**Fig. 1.1, Supplementary Fig. A.1**), nor will entire databases be able to be decoded and stored using low latency systems with much smaller capacities that are higher in storage hierarchies (i.e. semiconductor-based systems). In addition, high-capacity DNA storage systems will also require a large number of available file addresses (i.e. PCR primer sequences[5,8–12,15,16]) to organize the data. However, due to increasing probabilities for potential off-target molecular interactions as systems scale in capacity, addresses must be sufficiently different from each other in sequence and are, therefore, finite in number and limit total system capacities (**Fig. 1.1, Supplementary Fig. A.1**).

Our goal was to develop a robust platform with an easy to adopt implementation that could address these capacity limitations. Here we leverage, innovate, and integrate prior[14,18,19] and new robust biomolecular tools and encoding strategies to implement a platform capable of scaling storage system capacities. In particular, we present a system for non-destructively accessing specific data from high-capacity DNA-based databases in conjunction with a nested file address system that can handle the organization of exascale databases. We will refer to this overall storage system, which uses DNA Enrichment and Nested SEparation, as DENSE data storage. This system, through the integrated use of magnetic bead purifications and nested PCR primers, directly addresses the challenges arising from the molecularly crowded nature of high-capacity DNA storage systems while functioning within a single physical pool of DNA. Therefore, it not only harnesses the raw capacity and density advantages of DNA but also drives the practical scalability of high-capacity data storage systems.

Results

The current state-of-the-art file access method uses many cycles of PCR to amplify a desired file's corresponding DNA strands (referred to as random access[8,9,11,15,16]). However, random access is theoretically predicted to exhibit decreasing sequencing efficiencies with increasing database size as eventually PCR will not be able to overwhelm large quantities of non-target database strands. To experimentally measure this transition point, we generated a library of five files, each with unique PCR primer sequences (**Fig. 1.2a, Supplementary Fig. A.2a**), and mixed it with increasing quantities of background database strands. As it is currently cost prohibitive to order large databases of completely unique strands of DNA, large DNA databases can be mimicked in mass proportions by mixing copies of an individual file (i.e. 1.94 'non-unique' GB of File 3 strands = 1.14E9 total strands) with many more background database strands (i.e. 6.22 GB to 19.4 TB of a single non-specific DNA strand, 3.66E9 to 1.14E13 total strands, respectively). After 30 cycles of random-access PCR to amplify File 3 from this series of databases, the relative abundance of File 3 strands to background DNA was compared by quantitative PCR. As predicted, the percentage of the sample that was File 3 monotonically decreased as a function of increasing background DNA (**Fig. 1.2b**). File 3 fell below 50% of the total sample once the database size reached 31.1 GB and higher. Thus, in high-capacity systems random access becomes ineffective for specific file retrieval.

To address this database capacity limitation, we sought to physically separate newly created copies of specific files from the database, while preserving the original library, allowing for the non-destructive and efficient sequencing and analysis of only desired data. Inspired by prior examples of biotin-mediated separations of DNA[18,19], we modified this approach to create moiety-labeled copies of target file strands while leaving original unmodified file strands in the database. We did this by using moiety-modified primers in 1 cycle of PCR to create chemically labeled copies of a desired file's DNA strands (**Fig. 1.2c**). These labeled copies of individual files were then separated from the database of five files using magnetic beads and fully recovered, as confirmed by next generation sequencing (NGS) (**Fig. 1.2d, e, Supplementary Fig. A.2b-d**). We also expanded this approach to three other distinct modification systems and showed they were all capable of efficient and complete file access (biotin-streptavidin, fluorescein-antibody, digoxigenin-antibody, polyA-polyT oligomers). NGS results indicated sequencing efficiencies above 86%, representing a reduction in wasted sequencing throughput. Of note, to access files in this manner, we found that only a single emulsion PCR cycle was needed to chemically label files prior to their separation from the database. Importantly, we observed no destruction of the original database in the remaining solution following separation (**Fig. 1.2d, e, Supplementary Fig. A.2d**). Furthermore, we determined that the same or a different file could be repeatedly accessed from this previously ‘used’ database solution (**Fig. 1.2d**). Taken together, this approach to physically separate files is non-destructive and represents a reusable DNA-based storage system.

To directly compare the performance of DENSE storage with random access in high-capacity systems, we compared the relative enrichments of File 3 from a 5.53-TB database (**Figure 1.2f, g**). In this experiment, to better mimic a true high-capacity and high-diversity database, File 1 was mutagenized by two rounds of error prone PCR[20] to an estimated 5.53 TB of unique data. Whereas random access was not able to significantly enrich File 3 strands from this high-capacity database, all four DENSE separation methods enriched File 3 to above 99% of the total sample after 30 cycles of emulsion PCR using the corresponding chemically modified primers.

High-capacity systems require many unique addresses in order to store and access information, yet there are roughly 28,000 usable primer addresses that will not cross interact[15]. Thus, in a storage system comprised of 3 GB file sizes, 28,000 primers limit total system capacity to ~84 TB (**Fig. 1.1, ‘Single primer encoding’**), given our strand organization and encoding strategy. To address this database capacity limitation, we were inspired by nested PCR architectures that were previously posed as a way to expand the number of possible addresses[14]. We integrated this strategy into DENSE by using a hierarchical encoding scheme where primer sequences are nested and used in sequential combination (**Fig. 1.3a**). Theoretically, this architecture can more than exponentially increase the number of unique addresses for files without increasing the total number of unique primers needed: nesting 2 primers would increase theoretical system capacity by five orders of magnitude to enable exascale capacities (**Fig. 1.1, ‘Two primer hierarchy encoding’**), while nesting more than 2 primers would result in exponentially larger numbers of total addresses (i.e. 28,000 unique primers$^{\text{N number of nests}}$). Using this hierarchical PCR architecture with nested primers used in sequential combination (but with no physical extractions), both File 4 and File 5 were separately and selectively accessed using opposite temporal amplification sequences, albeit there were substantial amounts of contaminating off-target strands (**Fig. 1.3b, Supplementary Fig. A.3**). This contamination arises because the original strands of both File 4 and File 5 are still present in the 2nd PCR step and both files would therefore be amplified in both PCR steps. This problem would be further exacerbated in higher capacity systems because of the high background and larger file sizes. Therefore, we combined this hierarchical strategy with biotin separations after each PCR step to remove the background strands (including the undesired contaminating file) and saw a reduction of these contaminating off-target strands. Specifically, the desired file in each case comprised either 96.9% or 86.8% of the sample, as measured by quantitative PCR, showing the specificity of nested addresses when used in the correct hierarchical temporal sequence and in conjunction with file separation (Fig. 1.3b, Supplementary Fig. A.3).

Building off the nested separation platform we focused on augmenting the storage of DNA files through the development of a microfluidic device capable of automating fluid handling and file access. The design creates polydimethylsiloxane (PDMS) microfluidic reaction chambers for storage and extraction of specific files (Fig. 1.4a). The rectangular PDMS molds underwent O2 plasma treatment, were fixed to a glass microscope slides, and created variable width reaction chambers. Fluid flow and magnetic manipulations of the beads in the channel were optimized; specifically, the PDMS channels were treated with a blocking buffer (0.1% SDS, 5 mg/mL BSA, and 750 mM NaCl) prior to the additions of DNA and magnetic bead to the reaction chamber. All solutions contained 5% BSA in addition to the DNA or beads and the components used in previous manipulations (see "Biotin-Streptavidin file extractions" Methods Section). To visually demonstrate DNA retrieval, a mixture of two distinct, fluorescently labeled oligomer sequences was introduced to the channel (**Fig. 1.4b**). One sequence was specifically extracted using a 20-nt complementary oligomer bound to functionalized magnetic beads. Following annealing, a stationary wash step (see washing protocol in "Biotin-Streptavidin file extractions" Methods Section) was used to remove any unbound or non-specific DNA. The fluorescent tags (ATTO550 and FITC are red and green, respectively) were monitored via fluorescence microscopy throughout the reaction (**Fig. 1.4c**) and only the perfectly complimentary oligomer was separated by the magnetic beads while the nonspecific oligomer was washed away.

We then expanded the microfluidic system to store and access Files 1 and 3. An equimolar mixture of the two files was introduced to separate microfluidic devices. Notably, the device was able to withstand thermocycling; therefore, one cycle of amplification (PDMS channel mounted on glass slide was placed directly on thermal mixture surface) to create chemically labeled files and subsequent file extractions were performed. The starting mixture, reaction supernatant, and file elution were analyzed via qPCR to determine the percentage of file present in each solution (**Fig. 1.4d**). Each file was able to be selectively enriched above 75% while maintaining the roughly equimolar ratio of the two files in the supernatant. This PDMS, microfluidic device paves the way for future designs of a reusable, automation capable DNA-based data hard drives.

Discussion

DENSE is practical in that it reduces the number of PCR cycles needed compared to random-access methods: only 1 cycle was necessary to access data from the 5-file database. This not only reduces the amount of dNTPs and other reagents needed, but also reduces the chances of mutational errors and alterations in strand distributions that may arise from PCR (**Supplementary Figs. A.4-6**). Consequently, in conjunction with DENSE, encoding algorithms may not need to sacrifice as much information density towards error correction. We do note that when the capacity of databases increase, more PCR cycles are required to access target files using DENSE (**Fig. 1.2g**). At higher capacities DENSE outperforms PCR alone (random access was not able to access target files at all), but this requirement for increased PCR cycles suggests additional biochemical engineering should be pursued to improve the specificities and affinities of the many complex molecular interactions that can occur during file separation.

Although initially designed to address barriers to scaling to the extreme capacities anticipated in the future (~PB and higher), DENSE storage will already be useful and needed for smaller and imminently achievable capacities. For example, while the largest system created to date is 200 MB[15], GB or TB amounts of DNA are routinely achieved by mainstream DNA synthesis companies in their aggregate purchase orders. Even for such modest systems, if common file sizes of ~25 MB are desired, there will be challenges in providing enough unique addresses without harnessing nested address architectures (**Supplementary Fig. A.1, '25 MB files'**). These nested architectures will also need to be integrated with physical file separations to avoid obtaining undesired contaminating strands, as each sequential PCR would otherwise have all database files available as templates, defeating the purpose of a nested architecture. Furthermore, without physical file separations, reading data from GB to TB level systems will be wasteful and perhaps infeasible even using state-of-the-art sequencing capabilities. For instance, Illumina's NovaSeq6000 can read only 20-30 GB of data when conservatively accounting for 10 redundant copies per strand (i.e. read depth of 10) (**Fig. 1.1, Supplementary Fig. A.1**). Critically, this work demonstrates the enrichment and physical separation of 9.15 unique kBs of targeted DNA strands from 5.53 unique TBs of undesired database strands (**Fig. 1.2g**). When considering the file's raw capacity instead of unique data, DENSE was able to enrich 1.94 GB of non-unique DNA strands from 5.53 TB of background strands. In other words, target strands starting at only 0.025% of the original database were enriched to over 99% purity in the separated sample. Therefore, as systems continue to scale, DENSE could be used to store and access individual files containing at least GBs of data. Thus, this file access approach can be combined with a hierarchical, nested-address system to increase the theoretical total capacity of DNA storage systems by over five orders of magnitude (see **Fig. 1.1, Supplementary Fig. A.1, and Equations 1.1 & 1.2** in the Methods section for calculations).

While there are many challenges, and likely many still unanticipated, there are recent promising breakthroughs in all necessary aspects of DNA storage: advances continue to be made in DNA synthesis and sequencing, in encoding and error correction, and in physical file access and system architecture. This work provides a conceptual and quantitative framework to think about DNA storage systems and their challenges, proposes practical strategies to address key barriers to scaling system capacities, and suggests that DNA-based data storage systems with reasonable modern capacities and file access capabilities are not only immediately achievable but also scalable to extreme capacities in the future.

Figures

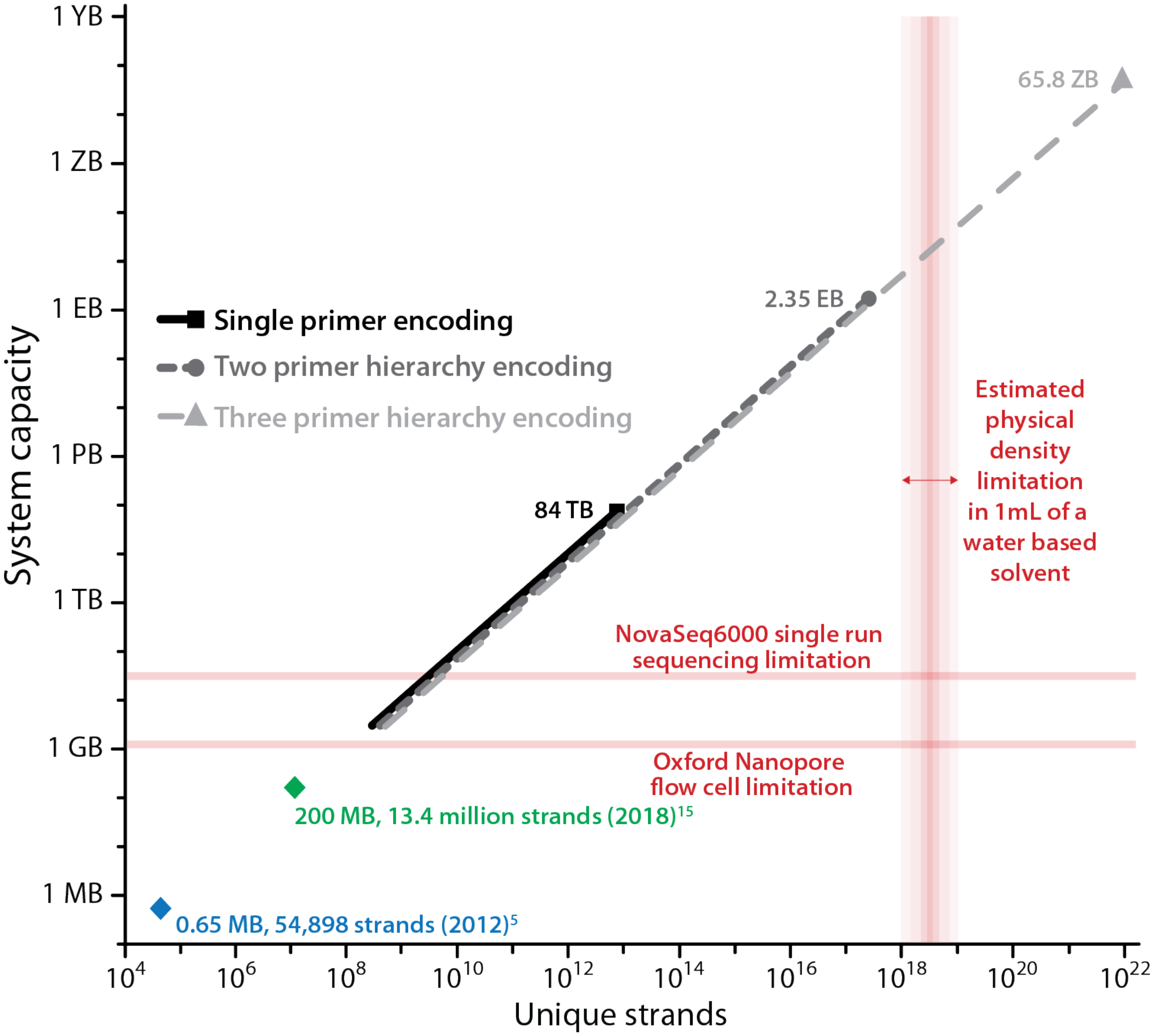


Figure 1.1 Theoretical analysis of readable files sizes, total system capacity limits, and improvements through physical data extraction and nested encoding.

Limited readable files sizes: Current sequencing platforms can only sequence a fraction (~20-30 GB) of the theoretical maximum capacity of current systems (84 TB) assuming a sequencing depth of 10. *Capacity limits*: The linear plots of system capacities are based on current best estimates of 28,000 usable primers[15] and an average file size stored per unique address of 3 GB. As the total number of unique strands within a database increases, so does the total system capacity, limited ultimately by the number of primers available. Thus, the availability of non-interacting primers limits the theoretical maximum capacity of storage systems. The system capacity limit for current one-primer encodings using 28,000 primers (all 27,999 files sharing 1 antisense primer) storing 3-GB files is 84 TB (Total capacity = total file addresses * file size); this corresponds to 7.88×10^{12} unique 200 bp long strands. In contrast, using the same distinct primers in double or triple nested architectures increases the number of possible addresses exponentially (Total file addresses = $27{,}999^{\text{N number of nests}}$). As a result, the total capacities also increase to 2.35 EB (2.52×10^{17} unique strands) and 65.8 ZB (8.98×10^{21} unique strands), respectively. The limits of commonly used next generation sequencing platforms are included for reference: Oxford nanopore flow cells can sequence 1.5×10^{11} bases, or roughly 1.27 GB per flow cell using our encoding scheme and average sequencing depth of 10. Illumina's Novaseq6000 platform can sequence 2×10^{10} of our 200 bp strands per run, or roughly 28.9 GB. The aqueous solubility of DNA is roughly between 10^{18} and 10^{19} per milliliter, depending on ionic concentrations.

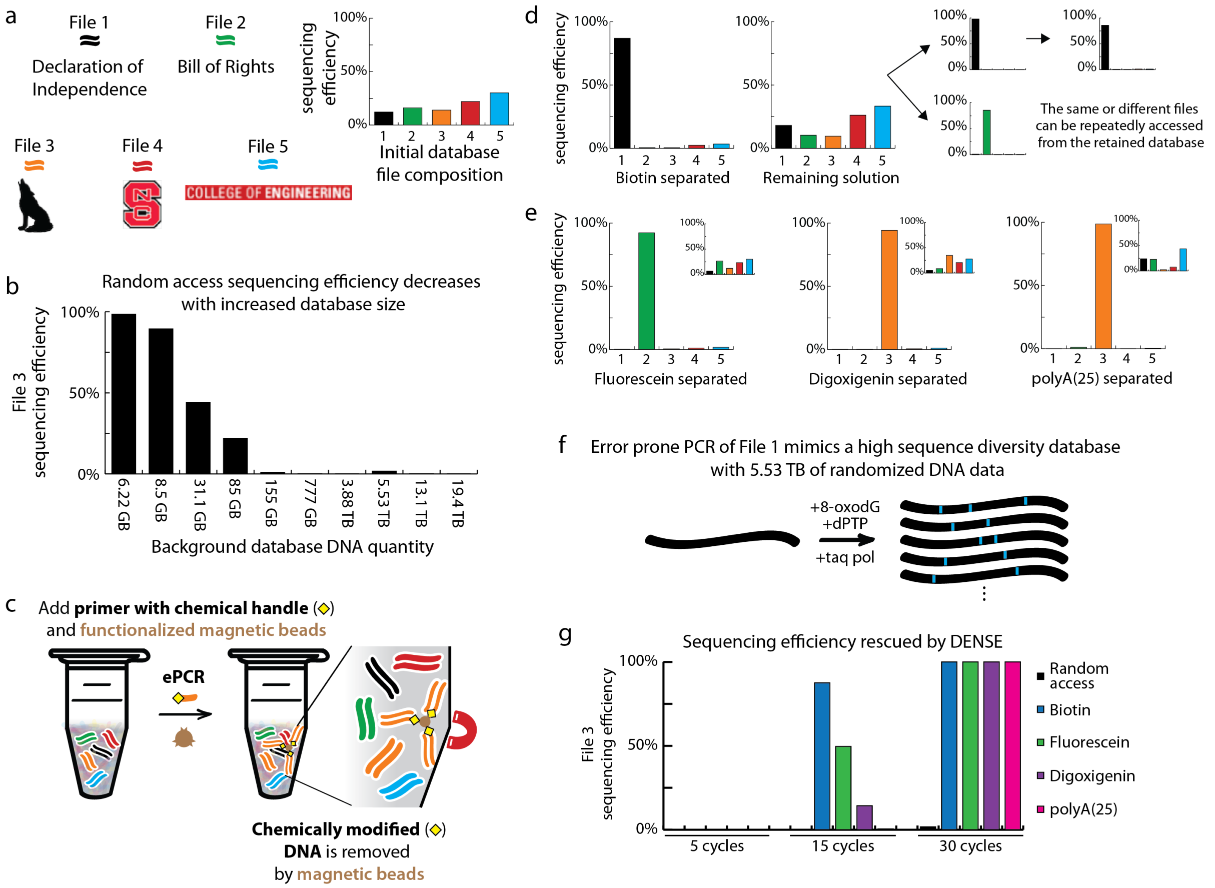


Figure 1.2 Physical file separations in DENSE storage rescue the decreased sequencing efficiency experienced by high-capacity databases.

(a) A library of five files was ordered and analyzed using NGS to confirm an even file distribution. (b) File 3 strands were enriched over increasingly higher capacity backgrounds of non-specific DNA strands using 30 cycles of random-access PCR. Random access failed to enrich File 3 to above 50% of the total sample once the background capacity reached 31.1 GB, as measured by quantitative PCR. (c) DENSE physically extracts a file (orange) from the database so only its strands are sequenced. A primer functionalized with a chemical handle (yellow diamond) is used to execute one emulsion PCR cycle to create chemically labeled copies of the desired file's strands. Functionalized magnetic beads (brown) that bind to the chemical handle are added to the sample. The desired file is bound to the bead, and the unbound solution containing the original database is removed and saved for future reuse. The bound file is then eluted from the bead. (d) After biotin-streptavidin file extractions, the remaining solution still contained all files while the target files were enriched and physically separated, as measured by NGS. By mapping sequencing reads to the original file sequences, all targeted data were confirmed recovered. The target file was retained in the supernatant containing the database and was able to be copied and extracted again. File 1 was extracted three sequential times, and File 2 was extracted from the solution remaining after an initial extraction of File 1. (e) File extractions using fluorescein, digoxigenin, and polyA(25) as chemical handles also successfully separated target files from the database. (f) A large-scale background mimicking diverse data was created using error prone PCR[20] to mutagenize and amplify File 1. (g) Random access was compared directly to chemical handle extractions. File 3 strands, with a starting fraction of 0.025% of the total number of strands, were enriched over a high-capacity background equivalent to 5.53 TB of undesired, non-specific strands using either random access (black) or PCR followed by chemical handle primer extractions (blue, green, purple or pink). After 5, 15, and 30 cycles of PCR (random access), enrichment of File 3 was 0.0%, 0.0%, and 1.69% of the total sample, respectively. After biotin-modified PCR followed by extraction, the enrichment of File 3 was 0.2%, 87.5%, and 100% of the total sample, respectively. After fluorescein-modified PCR followed by extraction, the enrichment of File 3 was 0.1%, 49.6%, and 100% of the total sample, respectively. After digoxigenin-modified PCR followed by extraction, the enrichment of File 3 was 0.2%, 14.2%, and 100% of the total sample, respectively. After poly(A)-25-modified PCR followed by extraction, the enrichment of File 3 was 0.09%, 0.47%, and 100% of the total sample, respectively.

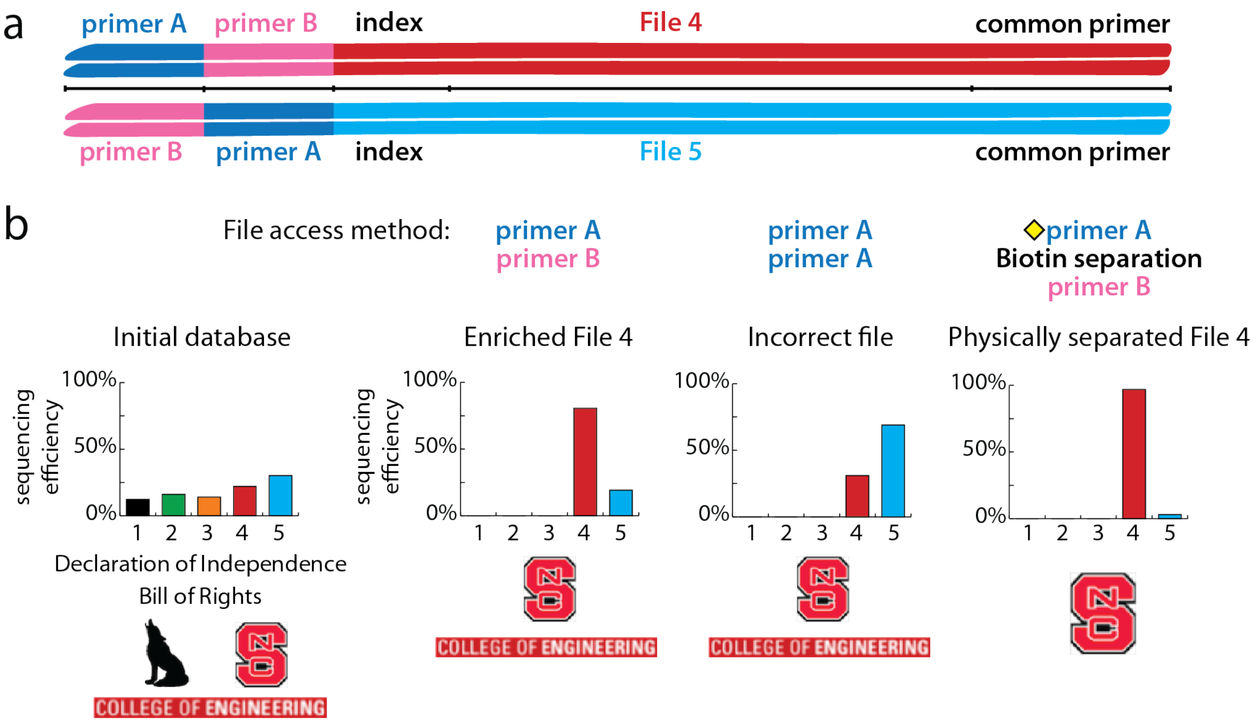


Figure 1.3 Combining a nested, hierarchical address strategy with physical separations results in purified enrichment of the desired file

(a) Strand architectures of Files 4 and 5 exhibit nested primer addresses. Binding sites for primers A and B are shared by both files but in opposite orders. Both files share a common antisense primer. (b) Experimental demonstration that PCR using primer A followed by primer B enriches for File 4. PCR amplifications using two rounds of the same primer enriches for the incorrect file. In conjunction with physical extractions, File 4 is specifically accessed using hierarchical PCRs. The extraction after the first PCR amplification increases File 4 enrichment from 81% to 97% over no extraction, as measured by qPCR.

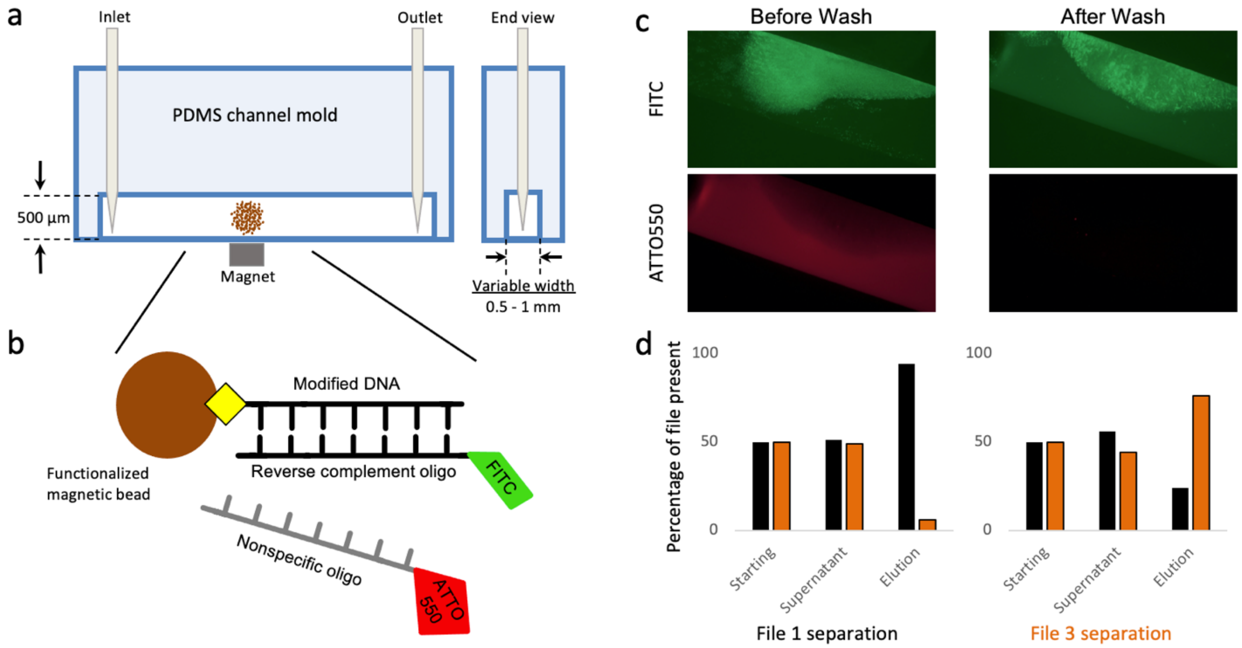


Figure 1.4 DENSE performed in a microfluidic device.

(a, b) Schematic of the PDMS channel and magnetic bead-based reaction used for a microfluidic channel. A solution containing two different fluorescently labeled DNA sequences is loaded into the channel. Functionalized magnetic beads containing reverse complementarity to one of the DNA sequences are introduced to the solution, allowed to interact, and held in place with a magnet while the supernatant is removed. (c) Fluorescent imaging of the channel before and after removing the supernatant and washing the magnetic beads; Green = FITC; Red = ATTO550. (d) The channel is scaled up to separate one of two files from an equimolar mixture. After biotin-streptavidin file extractions, the supernatant still contains both files in comparable proportions (measure via qPCR) while the target file is enriched and physically separated.

Materials and Methods

Data Representation, Encoding, and Decoding.

We adopted a similar approach for representing and encoding data as reported in recent work[6,9,15]. We partitioned a digital file into blocks of data that fit in DNA strands that are 200 bp long. Each strand consists of multiple fields. A primer binding site occupies each end and enables DNA polymerase chain reactions. Between the primers, we placed three fields that represent the index of the strand within the file, the data payload, and a checksum to detect errors within the strand. We used a fixed length index that is 16 bp-long. This leaves the remaining 136 bp-long sequence to represent the data payload of each strand. We designed 8 bp-long codewords to represent one byte of data. The codewords have no repetition of bases both individually and when appended, and they are GC balanced. Each byte of file is converted one byte at a time into a corresponding codeword and appended together to form the payload of a strand. The checksum is a single-byte XOR-accumulation of all the data in the payload that is encoded and appended to the end of the data payload. The checksum allows each strand to self-check its own data. The only notable difference for hierarchical encoding is that it requires an additional primer binding site in each strand, thereby reducing the size of the data payload.

We also adopted a redundant XOR-style encoding proposed by Bornholt et al.[9] to enhance the reliability of our system. In our design, indices with even values hold data, and odd indices store the XOR-ed content of their adjacent strands. This redundancy enables recovery of data even if some strands are lost or discarded due to an invalid checksum. The decoder algorithm for our encoding is similar to previous work[9], with the modification that we can disregard any read with an invalid checksum. It is important to note that for clarity of analysis and ease of comparison across systems, the file and database sizes estimated in the figures do not take into account the overhead required to implement XOR or other encodings that may be used. Thus, we present best case scenarios, whereas true capacity challenges and limitations are likely even more severe than described in this work.

Primer design.

Primers used in this work were designed to achieve multiple goals. First, they must facilitate effective PCRs. The primers were designed such that GC content is between 40% and 60%, and their melting temperature is between 50°C and 60°C. We required that the last base is G but the GC content in the last 5 bases could not exceed 60%. Second, primers were designed to reduce the likelihood of non-specific binding with other primer binding sites. We required a Hamming distance of >10 between all primers to minimize the likelihood of such binding. We also performed NUPACK simulations of homodimer, hairpin, and heterodimer bindings[21]. We required a Gibbs free energy greater than -10 kcal/mol at 50°C on all likely complexes to select the primer. Note, we compared each candidate primer to all other primers to ensure no heterodimer bindings are likely, and we included the Illumina NEXTERA primers in this process. Third, to reduce the likelihood of non-specific binding between a primer and the data payload, we required that primers must contain a repeating nt every 5 bases. This guaranteed that primers would differ from all length 20 sub-sequences of the data payload.

We used a computer program written in Python to automate the generation of candidate primer sequences and screened them against the requirements stated above. The python program invoked the relevant analysis in NUPACK as needed.

Emulsion PCR.

The emulsion PCR (ePCR) protocol from Schutze et al.[22] was modified slightly and used for all PCR steps. Emulsions were created by mixing 150 µL of emulsion oils (73% v/v Tegosoft DEC (Evonik, 99068594), 20% v/v mineral oils (Sigma Aldrich, 330779), 7% v/v ABIL WE (Evonik, 99068358)) with 25 µL aqueous PCR samples. Samples were then vortexed for 5 minutes until a persistent emulsion was formed. Samples were aliquoted into four PCR tubes and a standard Q5 polymerase PCR protocol was used. Twenty cycles were sufficient to reach the maximum yield of DNA product. After amplification, aliquots were pooled in an Eppendorf tube and emulsions were broken with the addition of 1 mL of isobutanol followed by a 5 second vortex. Five volumes of (125 µL for 25 µL PCR reaction volume) binding buffer (Biobasic Canada Inc. BS664) was added to samples, gently mixed and centrifuged at 2,400 g for 30 seconds. The organic phase was removed and discarded while the remaining aqueous phase was purified using AMPure XP beads (Beckman Coulter, A63881). DNA was eluted in 50 µL of water.

Biotin-Streptavidin file extractions.

File-specific sense ('coding') primers were ordered with a biotin modification on the 5' end. PCR amplified samples were purified (AMPure XP beads) and added to prewashed streptavidin magnetic beads (NEB #S1420S) (wash and bind buffer: 20 mM Tris-HCl pH 7.4, 2M NaCl, 2 mM EDTA pH 8) and incubated at room temperature on a rotisserie for 30 minutes. The database files were retained by collecting the supernatant. The beads were then washed once with

100 μL of the binding buffer and once with 100 μL of a low-salt wash buffer (20 mM Tris-HCl pH 7.4, 150mM NaCl, 2 mM EDTA pH 8). Amplified DNA was subsequently eluted (elution buffer: 95% formamide (Sigma, F9037) in water). DNA sizes and concentrations of the purified (AMPure XP beads) supernatants and elutions were measured on a Fragment Analyzer (Advanced Analytical, DNF-474) before the addition of Illumina sequencing adapters. Representative DNA gel images of biotin separations are shown in Supplementary Figure 2b.

Fluorescein and digoxigenin file extractions.

File-specific sense ('coding') primers were ordered with either fluorescein or digoxigenin on the 5' end (Eurofins Genomics). Antibodies (anti-fluorescein: Novus Biologicals, NB600-493, Lot 19458; anti-Digoxigenin (21H8): Novus Biologicals, NBP2-31191, 17E16) were bound to magnetic protein A or G beads (BioRad Cat. #s 161-4013 & 161-4023) through a 30-minute room temperature incubation (bind and wash buffer: 20 mM Tris-HCl pH 8, 300 mM NaCl, 2 mM EDTA). PCR amplified samples were purified (AMPure XP beads) and added to the antibody-linked beads and incubated at room temperature on a rotisserie for 2 hours. The database files were retained by collecting the supernatant. The beads were washed once with 100 μL of the binding buffer and once with 100 μL of a low salt wash buffer (20 mM Tris-HCl pH 7.4, 150mM NaCl, 2 mM EDTA pH 8). DNA sizes and concentrations of the purified (AMPure XP beads) supernatants and elutions were measured on a Fragment Analyzer (Advanced Analytical, DNF-474) before the addition of Illumina sequencing adapters. Representative DNA gel images of a fluorescein separation are shown in Supplementary Figure 2c.

Oligo-d(T) magnetic bead separation.

File-specific sense ('coding') primers were ordered with a poly(A)-25 tail on the 5' end (Eurofins Genomics). Oligo-d(T)$_{25}$ beads (NEB #S1419S) were washed twice with 100 μL wash and bind buffer (20 mM Tris-HCl pH 7.4, 2M NaCl, 2 mM EDTA pH 8). PCR amplified samples were purified (AMPure XP beads) and added to the desired amount of bead based on the amount of DNA present and theoretical binding capacity. The mixture was heated in a thermal mixer at 90°C and 500 rpm for 2 minutes, allowed to cool to room temperature, and the database files were retained by removing the supernatant. The beads were washed twice with 100 μL of a low salt wash buffer (20 mM Tris-HCl pH 7.4, 150mM NaCl, 2 mM EDTA pH 8). Beads were then resuspended in 1x TE buffer, heated in the thermal mixer at 50°C and 500 rpm for 2 minutes. The desired file was extracted while the mixture was still hot by removing the eluted sample from the beads. DNA sizes and concentrations of the purified (AMPure XP beads) supernatants and elutions were measured on a Fragment Analyzer (Advanced Analytical, DNF-474) before the addition of Illumina sequencing adapters.

Calculation of data quantity from total number of DNA strands.

In Figures 1.1, 1.2, and Supplementary Figs. A.1 & A.2 we refer to file and database sizes (MB, GB, etc.). For clarity and ease of comparison all values were calculated based on the total number of DNA strands. Each strand is comprised of 200 nts, 20 of which are used for each primer sequence, 16 for the index, and 8 for the checksum. 8 nts comprise each 1-byte codeword. Thus, each strand addressed with a single primer pair contains 17 bytes of data. Specifically, in Figure 2, we assumed a 10-copy physical redundancy per unique strand to provide a conservative estimate for a realistic system where multiple copies of each strand would likely be needed to avoid strand losses and inhomogeneous strand distributions. Thus, in Figure 1.2 total file and database sizes are divided by 10.

Calculation of System Capacity.

In Figure 1.1, and Supplementary Fig. A.1, we calculate the system capacity by following Equation 1.1 and Equation 1.2.

$$System\ Capacity\ (B) = P * U * D \qquad (1.1)$$

$$Strand\ Density\ (B/Strand) = \frac{Strand\ length\ -\ Strand\ Overhead}{Encoding\ Density} \qquad (1.2)$$

Where P is the number of primers available to the system, U is the number of unique strands that can be supported for each file, and D is strand density in units B/Strand. The density, D, in B/Strand can be calculated by dividing the number of bases available for data encoding by the encoding density in units of B/Base. For Figure 1.1 and Supplementary Fig. A.1, we start with a strand length of 200 and subtract off the overhead associated with both flanking primers, which will be a total of 40 bases in the case of a single primer system, and will be 60 bases in the case of a hierarchical primer system. The leftover bases can then be either allocated to the index region of the strand, or to the payload region. With the number of bases selected for the index region, the number of unique strands supported for each file, U, can then be determined by applying the encoding method utilized by the system for the index. In our examples we conservatively choose a base-3 encoding, thus U will be equal to 3^N, where N is the number of bases allocated to the index region. With the remaining bases, strand Density can be calculated by dividing the number of remaining bases by the encoding density in units of B/Base, where in our examples we conservatively choose an encoding density of 0.125 B/Base (8 bases for each byte).

Error prone PCR.

Template DNA was amplified using 0.5 µL of Taq DNA polymerase (5 units/µL, Invitrogen, 100021276) in a 50 µL reaction containing 1X Taq polymerase Rxn Buffer (Invitrogen, Y02028), 2 mM MgCl2 (Invitrogen, Y02016), the sense and antisense primers at 1E13 strands each, and dATP (NEB, N0440S), dCTP (NEB, N0441S), dGTP (NEB, N0442S), dTTP (NEB, N0443S), dPTP (TriLink, N-2037), 8-oxo-dGT (TriLink, N-2034), each at 400 mM. PCR conditions were 95°C for 30 seconds, 50°C for 30 seconds and 72°C for 30 seconds for 35 cycles with a final 72°C extension step for 30 seconds.

qPCR.

qPCR was performed using SsoAdvanced Universal SYBR Green Supermix (BioRad). qPCRs were performed in 5 µL format using SYBR Green (95°C for 2 min, and then 50 cycles of: 95°C for 10 s, 50°C for 20 s, and 60°C for 20 s). qPCR results were compared to next generation sequencing results for samples that were analyzed using both methods. File compositions measured using both methods showed strong agreement (Supplementary Table A.1).

Illumina library preparation.

Illumina TruSeq Nano DNA Library Preps (Illumina, 20015965) were performed according to manufacturer instructions beginning from the 'Repair Ends and Select Library Size' step, as DNA fragmentation was unnecessary. The quality and band sizes of libraries were assessed using the High Sensitivity NGS Fragment Analysis Kit (Advanced Analytical, DNF-474) on the 12 capillary Fragment Analyzer (Advanced Analytical) at multiple steps during each protocol, typically after size selection and after PCR amplification. Unless otherwise stated, libraries were normalized to balance estimated sequencing depth across similar samples (e.g. all elutions had estimated sequencing depth of ~100 reads) using the molar concentrations measured on the Fragment Analyzer. The pooled sample had a concentration of 8 nM and was sequenced using the MiSeq v2 chemistry 150 PE kit that was operated as a 300 SR run. PhiX DNA was added at 20% of total DNA to increase sequence diversity.

Error Analysis.

Before proceeding with an error analysis of sequenced strands, the error-free reference strand for each sequenced strand needed to be determined. To find the error-free reference strands, a mapping operation was performed to match each sequenced strand with its original database strand. Due to the large number of sequenced strands in samples (up to 571k reads), the mapping operation was carried out in two steps: the first step partitioned the large read space using the primer sequences of the different files, and the second step further analyzed each partition to match each strand in a partition with its corresponding database strand.

The first step of mapping divided the initial sequencing read space into partitions, one for each file in the database, with the exception of Files 4 and 5 (hierarchical encodings) where each of these files had 2 partitions. These 2 partitions were used to separate nested address strands that were truncated from the first PCR step and reads where the nested address strands were not truncated. Other partitions were also created for special strands like the background strands used to simulate high-capacity data storage and for unknown strands that could not be categorized into a file's partition. A strand from sequencing was placed into a partition by looking for a subsequence that matched a file's sense primer, or the reverse complement of the anti-sense primer. The reverse complement of the anti-sense primer was used because all NGS sequencing reads are in the 5' to 3' direction. A subsequence was deemed acceptable if it matched a sense primer or anti-sense primer's reverse complement within a Levenshtein distance of 4. A Levenshtein distance of 4 was chosen as the cut-off point to ensure that the matched subsequence was not data within a DNA strand, but one of the primers of interest. When a primer of interest is found in a sequenced strand, the sequenced strand is placed in the primer's respective partition.

After categorizing each strand in a sample's sequence pool, each partition was analyzed further to determine the original database strand for each sequenced strand in the partition. To find out the correct original strand, each original strand from a file was compared to each sequenced strand placed in the file's partition by calculating the Levenshtein distance between the sequenced strand and the original strand. If the distance was less than or equal to 12, the original strand was considered as a candidate for a match. Because some of the original strands in the database have small edit distances between them, file strands that are close to the candidate were also checked against the sequenced strand to make sure the correct original strand was chosen. Once a candidate was concluded to correspond to a specific original strand, the location of the matching strand in the file along with the sequenced strand's location in the read space was recorded. A distance of 12 was chosen as the threshold in order to reduce the amount of checking that was required once a candidate was found, while ensuring that error rates would not be artificially low due to choosing candidates that were within a small number of edit operations.

With a completed mapping of sequenced strands to their corresponding database strands, analyses such as error rates per base, strand error rates, and read distributions were performed. To calculate the error rate for a nt position, Equation 1.3 was used. Where L is the number of unique edit operations considered (insertions, deletions, substitutions), M is the number of unique strands in the database, s_j is the jth strand in the database, N_j is the number of sequenced strands that map to strand s_j, s_k is the kth strand that maps to database strand s_j, T is the total number of strands from the sample that has been mapped to some database strand, and $EO_l(s_j , s_k)_i$ is the number of edit operations of type l at the ith nt position to transform s_j to s_k. This equation calculates the total error rate for base position i by summing up all of the edit operations of each type at the ith position needed to transform each original database strand to the sequenced strands that map to it, and then dividing by the total number of mapped strands in the sample.

$$Total\ Error\ Rate_i = \frac{\sum_{l=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{N_j} EO_l(s_j, s_k)_i}{T} \qquad \mathbf{(1.3)}$$

Similarly, the error rate for each strand in the original database was calculated using Equation 1.4. Where L is the number of unique edit operations, s_j is a strand from the original database, N_j is the number of sequenced strands that map to strand s_j, s_k is the kth strand that maps to s_j, T_{sj} is the total number of mappings in the sample for s_j, and $EO_l(s_j , s_k)$ is the number of edit operations of type l to transform s_j to s_k.

$$Total\ Error\ Rate_{s_j} = \frac{\sum_{l=1}^{L} \sum_{k=1}^{N_j} EO_l(s_j, s_k)}{T_{s_j}} \qquad \mathbf{(1.4)}$$

Acknowledgements

We acknowledge Kevin N. Lin for helpful discussions. This work was supported by the National Science Foundation (#CNS-1650148), a North Carolina State University Research and Innovation Seed Funding Award (#2018-2509), and the North Carolina Biotechnology Center Flash Grant to AJK and JT. KJT was supported by a Department of Education Graduate Assistance in Areas of Need Fellowship. EWI and AH were supported by funds from the North Carolina State University Research Experience for Undergraduates, Provost's Professional Experience Programs, and NCSU startup funds.

Author Contributions

KJT, JT, and AJK conceived the study. KJT, EWI, and AJK developed the wet experimental system. KV, AS, and JT developed the software and simulations. KJT, EWI, AGH, and AJK planned and performed the wetlab experiments with guidance from all. KV, AS, and JT planned and performed simulations and next generation sequencing analysis with guidance from all. KJT, AJK, KV, and JT wrote the paper with input from all.

References

(1) Jonathan P.L. Cox. (2001) Long-Term Data Storage in DNA. TRENDS Biotechnol. 19, 247–250.

(2) Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013) Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse. Nature 499, 74–78.

(3) Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., Stark, W. J. (2015) Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angew. Chemie - Int. Ed. 54, 2552–2555.

(4) Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M., Hughes, W. L. (2016) Nucleic Acid Memory. Nat. Mater. 15, 366–370.

(5) Church, G. M., Gao, Y., Kosuri, S. (2012) Next-Generation Digital Information Storage in DNA. Science 337, 1628.

(6) Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., Birney, E. (2013) Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized DNA. Nature 494, 77–80.

(7) Shah, S., Limbachiya, D., Gupta, M. K. (2014) DNACloud: A Potential Tool for Storing Big Data on DNA. arXiv: 1310.6992.

(8) Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H., Milenkovic, O. (2015) A Rewritable, Random-Access DNA-Based Storage System. Sci. Rep. 5, 14138.

(9) Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., Strauss, K. (2016) A DNA-Based Archival Storage System. ASPLOS ’16 637–649.

(10) Blawat, M., Gaedke, K., Huetter, I., Chen, X.-M., Turczyk, B., Inverso, S., Pruitt, B., Church, G. (2016) Forward Error Correction for DNA Data Storage. Procedia Comput. Sci. 80, 1011–1022.

(11) Yazdi, S. M. H. T., Gabrys, R., Milenkovic, O. (2017) Portable and Error-Free DNA-Based Data Storage. Sci. Rep. 7, 5011.

(12) Erlich, Y., Zielinski, D. (2017) DNA Fountain Enables a Robust and Efficient Storage

Architecture. Science 355, 950–954.

(13) Agrawal, A., Limbachiya, D., M., R., Saiyed, T., Gupta, M. K. (2019) BacSoft: A Tool to Archive Data on Bacteria. arXiv: 1903.01902.

(14) Kashiwamura, S., Yamamoto, M., Kameda, A., Shiba, T., Ohuchi, A. (2003) Hierarchical DNA Memory Based on Nested PCR. DNA8, LNCS 2568, 112–123.

(15) Organick, L., Ang, S. D., Chen, Y.-J., Lopez, R., Yekhanin, S., Makarychev, K., Racz, M. Z., Kamath, G., Gopalan, P., Nguyen, B., et al. (2018) Random Access in Large-Scale DNA Data Storage. Nat. Biotechnol. 36, 242–249.

(16) Organick, L., Chen, Y., Ang, S. D., Lopez, R., Strauss, K., Ceze, L. (2019) Experimental Assessment of PCR Specificity and Copy Number for Reliable Data Retrieval in DNA Storage. bioRxiv. DOI: 10.1101/565150.

(17) Zakeri, B., Carr, P. A., Lu, T. K. (2016) Multiplexed Sequence Encoding: A Framework for DNA Communication. PLoS One 11.

(18) Adleman, L. M. (1994) Molecular Computation of Solutions to Combinatorial Problems. Science 266, 1021–1024.

(19) Stewart, K., Chen, Y.-J., Ward, D., Liu, X., Seelig, G., Strauss, K., Ceze, L. (2018) A Content-Addressable DNA Database with Learned Sequence Encodings. DNA Comput. Mol. Program. 55–70.

(20) Zaccolo, M., Gherardi, E. (1999) The Effect of High-Frequency Random Mutagenesis on in Vitro Protein Evolution : A Study on TEM-1 b -Lactamase. J. Mol. Biol. 285, 775–783.

(21) Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., Pierce, N. A. (2011) Software News and Updates NUPACK: Analysis and Design of Nucleic Acid Systems. J Comput. Chem 32, 170–173.

(22) Schütze, T., Rubelt, F., Repkow, J., Greiner, N., Erdmann, V. A., Lehrach, H., Konthur, Z., Glökler, J. (2011) A Streamlined Protocol for Emulsion Polymerase Chain Reaction and Subsequent Purification. Anal. Biochem. 410, 155–157.

CHAPTER 2: Promiscuous molecules for smarter file operations in DNA-based data storage

Kyle J. Tomek[1], Kevin Volkel[2], Elaine W. Indermaur[1], James M. Tuck[2]*, Albert J. Keung[1]*

1. Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27606

2. Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606

*Correspondence (jtuck@ncsu.edu, ajkeung@ncsu.edu)

Abstract

DNA holds significant promise as a data storage medium due to its density, longevity, and resource and energy conservation. These advantages arise from the inherent biomolecular structure of DNA which differentiates it from conventional storage media. The unique molecular architecture of DNA storage also prompts important discussions on how data should be organized, accessed, and manipulated and what practical functionalities may be possible. Here we leverage thermodynamic tuning of biomolecular interactions to implement useful data access and organizational features. Specific sets of environmental conditions including distinct DNA concentrations and temperatures were screened for their ability to switchably access either all DNA strands encoding full image files from a GB-sized background database or subsets of those strands encoding low resolution, File Preview, versions. We demonstrate File Preview with four JPEG images and provide an argument for the substantial and practical economic benefit of this generalizable strategy to organize data.

Introduction

Information is being generated at an accelerating pace while our means to store it are facing fundamental material, energy, environment, and space limits[1]. DNA has clear potential as a data storage medium due to its extreme density, durability, and efficient resource conservation[2–6]. Accordingly, DNA-based data storage systems up to 1 GB have been developed by harnessing the advances in DNA synthesis and sequencing, and support the plausibility of commercially viable systems in the not too distant future[7–11]. However, in addition to continuing to drive down the costs of DNA synthesis and sequencing, there are many important questions that must be addressed. Foremost among them are how data should be organized, accessed, and searched.

Organizing, accessing, and finding information constitutes a complex class of challenges. This complexity arises from how information is commonly stored in DNA-based systems: as many distinct and disordered DNA molecules free floating in dense mutual proximity[8,9,12–16]. This has two major implications. First, an addressing system is needed that can function in a complex and information dense molecular mixture. While the use of a physical scaffold to array the DNA would ostensibly solve this challenge, analogous to how data are addressed on conventional tape drives, this would abrogate the density advantage of DNA as the scaffold itself would occupy a disproportionate amount of space. Second, while the inclusion of metadata in the strands of DNA could facilitate search, ultimately there will be many situations in which multiple candidate files contain very similar information. For example, one might wish to retrieve a specific image of the Wright brothers and their first flight, but it would be difficult to include enough metadata to distinguish the multiple images of the Wright brothers as they all fit very similar search criteria. Additionally, data stored using DNA could be maintained for generations[6] with future users only having access to a limited amount of metadata and cultural memory or knowledge. Given the costs associated with DNA retrieval and sequencing, a method to preview low resolution versions of multiple files without needing to fully access or download all of them would be advantageous.

In previously reported DNA systems, files were organized, recognized and accessed through specific DNA base-pair interactions with ~20 nucleotide (nt) address sequences in both PCR-based file amplifications[8,13–16] and hybridization-based separations[9–11]. However, these address sequences participate in thermodynamically driven interactions that are not cleanly all-or-none as they are for conventional electronic storage addresses[17,18]. To bypass this limitation, current DNA system architectures and encoding strategies avoid any untoward cross-interactions between addresses by setting a threshold for sequence similarity (e.g., Hamming distance, HD)[16,19,20] (Fig. 1a). These limits on the address sequence space result in a reduction in the storage capacity of systems[14,21], as well as in the amount of metadata that could be included for use in search functions. Both limitations pose significant practical barriers for this technology and restrict the engineering of more advanced and useful functions[10,11].

We hypothesize that so called non-specific interactions in DNA-based data storage systems, conventionally viewed as a thermodynamic hinderance, can actually be leveraged to expand file address space, increase data storage capacity, and implement in-storage functions in DNA storage systems. This hypothesis is inspired by intentional non-specific interactions that have been leveraged for DNA editing in molecular biology (e.g., site-directed mutagenesis) and more recently in DNA storage for in-storage search[10,11,22]. Here we develop a theoretical and experimental understanding of factors that impact DNA-DNA interactions and show that we can predictably tune molecular interactions between imperfectly matched sequences in both isolated and competitive systems. To further demonstrate this concept and its potential utility in a true data storage system, individual files are encoded into three or four distinct subsets of strands (i.e., fractions of the file) that can be differentially accessed using the same accessing primer by tuning only the PCR conditions. In this approach, a small portion of a file can be previewed as a low-resolution version of the original higher resolution image, an operation with closest analogy to Quick Look, a function found on modern Mac operating systems or Progressive JPEG but with fundamentally different implementation. Importantly, this function uses address sequences (i.e., primer binding sites) that would have previously been discarded due to their mutual sequence similarities, and therefore does not impact total database capacities. We successfully implement File Preview for four different image files in the presence of a randomized, non-specific 1.5 GB background to more accurately represent a realistic DNA storage system. This approach to encoding and accessing strands harnesses the intrinsic properties of DNA and implements practical functionality while remaining cost competitive with traditional DNA storage systems. We also anticipate that this general principle of leveraging uniquely biochemical aspects of DNA molecules could be extended to implement diverse and useful functions including encoding metadata, increasing or decreasing the stringency of a search function, and prioritizing information by differentially encoding files to increase the read efficiency of frequently versus infrequently accessed data.

Results

PCR stringency is thermodynamically tunable.

In PCR-based DNA storage systems, data payloads, file addresses, and PCR primers that bind those addresses have typically been designed to avoid non-specific interactions by requiring that all primers are at least six to ten or more mismatches from all other sequences in a database (6-10+ HD) (Fig. 1a)[14,16]. To test this design criterion, we incorporated the widely used NuPACK thermodynamic model into a Monte Carlo simulation and found that a HD of greater than 10 was likely required to minimize unwanted hybridizations (Fig. 1b, black line)[14,23]. We confirmed this experimentally by measuring the percentage of successful PCRs using a primer with 10 strand addresses of each successively greater even-numbered HD (Fig. 1b, dashed line). Non-specific amplifications were minimized beyond mismatches of ~6 HD and greater. Indeed, the likelihood of amplification was expected to be lower than the likelihood of hybridization since in wet experimental conditions a primer samples the reaction volume with the potential to interact with other strand regions, and also must interact with a DNA polymerase molecule to carry out the amplification.

Systems based upon such stringent criterion have had success with small-scale systems, but this criterion constrains the set of potential non-interacting addresses to a few thousand from the theoretical maximum of 4^{20} (for 20-nt addresses)[12–14,16](Fig. 1b, inset). This severely limits the functionality of DNA storage systems. We hypothesized that rather than viewing non-specific interactions as a hindrance, they could instead be potentially useful if controllable. In particular, it could be possible to tune the access of different subsets of DNA strands by simply changing environmental conditions while using the same file-access primer.

Towards this goal, we considered how biomolecular interactions are governed by thermodynamics (Fig. 1c), with more negative Gibbs free energy (ΔG), lower temperature, or higher primer concentration leading to more template binding sites being bound. Sequences with a higher HD have a less negative ΔG (they are less favorable to bind thermodynamically) but this can be compensated by changes in temperature or primer concentration. Embedded in this equilibrium equation also is that the equilibrium constant itself can be dependent on other environmental factors such as ionic strength and the presence of detergents. Based on thermodynamics[24–26] and significant practical work in molecular biology and biochemistry[27–29] we tested how a range of temperatures and primer concentrations would shift the likelihood of PCR amplification. As expected, lower annealing temperatures and higher primer concentrations increased non-specific amplifications, while higher annealing temperatures and lower primer concentrations decreased non-specific amplifications (Fig. 1d,e and Supplementary Table 1).

Thermodynamics tune amplification within competitive PCRs.

DNA strands in a storage system do not function in isolation, so we designed a competitive system with two unique template strands that had closely related addresses. In this reaction we added a single 20-nt PCR primer pair used to amplify both strands: a 200 bp template with perfectly complementary primer binding sites (0 HD) and a 60 bp template with primer binding sites containing 2 mismatches (2 HD). In this competitive context using only one pair of PCR primers, only the 0 HD strands were amplified using stringent conditions (e.g., high annealing temperature and/or low primer concentration). Both 0 HD and 2 HD strands were amplified using promiscuous conditions (e.g., low annealing temperature and/or high primer concentration) (Fig. 2a).

To further tune the relative yield of 0 and 2 HD strands, strands with six distinct 2 HD forward primer-binding addresses and five distinct 2 HD reverse primer binding addresses, paired in all combinations, were amplified with the same primer set in PCRs at both stringent and promiscuous conditions. This yielded a range of ratios of promiscuous to stringent amplifications (Fig. 2b). Interestingly, the 2 HD addresses that exhibited tunability when varying annealing temperature also tended to be more likely to exhibit tunability when varying primer concentration.

Implementing File Preview of jpeg images through thermodynamic swings.

We hypothesized that this tunable promiscuity could provide a useful framework for organizing data and implementing functionalities. We focused on engineering a practical data access function, File Preview (Fig. 3a). For an image, this could be implemented where stringent PCR conditions would amplify and access only a subset of strands encoding a low-resolution pixelated Preview version (or thumbnail) of an image. In contrast, promiscuous PCR conditions would amplify both the Preview strands and the rest of the strands comprising the full image. The same exact primers would be used in both stringent and promiscuous conditions. We asked if this tunability could be applied to entire files (NCSU Wuflab logo – 25.6kB, two Wright glider images – 27.9 and 30.9kB – Figure 3f left and right respectively, Earth – 27.2kB) rather than just toy DNA strands. Furthermore, we expanded our screen for primers and addresses and asked if this principle of tunable promiscuity could be extended to more distant HDs to create multiple Preview layers. We screened four 20-nt primer pairs and up to 30 distinct 0, 2, 3, 4, and 6 HD addresses per pair. We screened them individually and in competitive reactions using a diverse range of PCR conditions incorporating salt concentrations, detergents, temperature, primer concentration, number of unique mismatch strands present, and size and ratios of template strands (Supplementary Tables 1 and 2, Supplementary Figs. 1 and 2).

Based upon these results, we designed files containing 0, 4, and 6 HD strands (Supplementary Fig. 1). Selecting the most consistent primer and its variable addresses, the Preview data strands were encoded with the fully complementary primer binding addresses (0 HD) while the rest of the file was encoded with the 4 HD (Intermediate Preview) and 6 HD (Full Access) addresses. The most stringent condition successfully accessed only the Preview image (Fig. 3b). Furthermore, the distribution of sequencing reads showed this Preview condition cleanly accessed only the Preview strands (Fig. 3c, top). When conditions were made less stringent, both 0 and 4 HD strands were accessed as expected, and the intermediate preview image with higher resolution was obtained. However, when we attempted to access the full file, we did not obtain any 6 HD strands. Instead, we discovered that there were problematic sequences in the data payload region that had been inadvertently encoded to be only 5 HD from the primer sequence. While the full file was therefore not accessed, this accident serendipitously revealed that a relatively sharp transition of just 1 HD (between 4 and 5 HD) could be cleanly achieved between the intermediate and full file access conditions (Fig. 3c). We also found in this experiment that because the 0 HD strands amplified efficiently in all conditions, it often dominated the distribution of sequencing reads. We therefore found that increasing the physical copy number of mismatched strands (alternatively one could encode more data in the higher HD partition of the file) resulted in a more even sequencing read distribution between 0 and 4 HD strands. Furthermore, by using more promiscuous access conditions, the balance of 0 and 4 HD strands that were accessed could be tuned and evened out (Fig. 3c, middle vs. bottom).

To explore these transitions and develop more informed control over them, diverse factors were individually varied to determine their impact on reaction specificity/promiscuity (Fig. 3d,e and Supplementary Figs. 2 and 3). 0, 2, 4, and 6 HD strands were used, each having unique restriction sites that allowed for digestion and facile quantification of each strand type by capillary electrophoresis. The accidental 5HD strands were still present so their contributions were quantified as well. The most important factor in Preview tunability was PCR annealing temperature, with a low temperature (40-45°C) resulting in an increased proportion of mismatched strands when compared to high annealing temperatures (55-60°C). Other parameters and reagents were nonetheless important for fine tuning the system. Primer and magnesium chloride ($MgCl_2$) concentrations had inverse relationships with specificity, while potassium chloride (KCl) concentration exhibited a direct relationship to specificity up to 150mM (beyond which PCR amplification was completely inhibited, Supplementary Fig. 3a,b). In aggregate, a gradient of distinct conditions were identified that were able to specifically access 0, 0-2, 0-2-4, and 0-2-4-5 HD strands as well as successfully decode low, intermediate, and higher resolution images (Fig. 3f,g and Supplementary Fig. 3c).

In a true data storage system, each file will be a small fraction of the total data. Biochemical interactions may be affected by the high background of other DNA strands and potential non-specific interactions; we therefore asked if File Preview could still function in a high background system. A text file encoding the United States Declaration of Independence was amplified via error prone PCR[30] to create strands equivalent to 1.5 GB of data (Supplementary Fig. 4a), and each image file (NCSU Wolf, two Wright glider images, Earth) was amplified in the presence of this non-specific, noisy background (Fig. 3h). In this setting, the Preview strands (0 HD) were merely ~0.036% of the total number of strands present in the reaction. Encouragingly, we were still able to reliably amplify and decode the Preview strands for each of the four files using stringent PCR access conditions. When promiscuous PCR conditions were used all four files were able to be accessed, decoded, and displayed without background contamination (Fig. 3i and Supplementary Fig. 4b).

File Preview can reduce next generation sequencing costs.

This system provides an innovative functionality for DNA data storage systems; however, it is important to consider what the potential benefits and tradeoffs of this system may be from a practical and quantitative perspective. When implementing Preview there are two main trade-offs to consider: physical storage density and sequencing cost (Fig. 4). In our current balance of Preview versus full access strands, we are previewing ~5% of a file's strands (5% file preview). This requires 100x more copies of each unique 4 HD strand than each unique 0 HD strand (1:100 ratio) to account for differences in PCR efficiency. With this current configuration, a file in which 5% of the strands are used for Preview requires 95x the physical space to be stored (Fig. 4a, black line) compared to normal encoding. Further reducing the copy number of full file strands by a factor of ten (1:10 ratio) or twenty (1:5 ratio) allows a file to be stored in 9.5x or 4.8x of the theoretical minimum physical space, respectively (Fig. 4a, grey dashed and light-grey dashed lines). This loss of physical efficiency is tunable based on the percent of the file to be Previewed and, subsequently, the number of excess copies of each unique full file strand to be stored. For example, when the Preview strands account for a smaller fraction of a file (~0.1-1%), the total number of full file strands will already be in a sufficient ratio to Preview strands to account for PCR efficiency differences; therefore, excess copies will not need to be stored. This removes the negative tradeoff in storage density. In the future, for any desired percentage of a file that one wishes to encode with Preview strands, one may be able to match access conditions, polymerase type, or primer selection so that all unique strands are present at equivalent copy numbers.

With regards to cost, when searching for a file in a database or recovering only key portions of data in a series of files, costs may be lowered by requiring the sequencing of fewer strands when quickly Previewing a file (or multiple files) rather than needing to sequence entire files. To understand this trade-off, envision a small database with 15 very similar files where: the full contents of the files are unknown, all 15 pairs of access primers are known, and a user is trying to find and sequence a target file of interest from amongst these 15 files based upon information that is not included in any metadata system. Without File Preview, one would potentially sequence 15 full files before finding the correct one. Using File Preview, one would sequence only the Preview strands of each of the 15 files until the correct file was found. Then that full file would be sequenced. Assuming all 15 files were searched, it would cost 85.3% less to find and fully sequence a file using a 5% Preview system (5% of all unique strands are Preview strands) compared to a normal encoding system (Fig. 4b). This cost advantage only increases as the percentage of strands encoding the Preview strands decreases, and as the number of files needed to be searched increases. Encouragingly, even without further engineering the access conditions, by reducing the percent of the file being Previewed from 5% to 1% it would cost 91.7% less to find and fully sequence a file from the 15-file library using the Preview system compared to a normal encoding system.

Discussion

The File Preview function is practical in that it reduces the number of strands that need to be sequenced when searching for a desired file. This will reduce the latency and cost of DNA sequencing and decoding. Consequently, one will be able to search a database of files much more rapidly and cost effectively using Preview than if each file needed to be fully sequenced. Beyond the Preview function, this inducible promiscuity technology could be used for many other data or computing applications. It may have broad application to how data is managed or organized in a file system. For example, files could be differentially encoded to make it cheaper and easier to access frequently versus infrequently used data. Another interesting use case is support for deduplication of data, a ubiquitous need in large and small data sets in which replicated blocks of data are detected and optimized[31]. Rather than storing many copies of duplicated data, a single copy could be shared amongst files by taking advantage of the promiscuous binding.

Although we initially designed our File Preview system to include 0, 2, 4 and 6 HD file addresses for each file, there were problematic sequences that arose within the data payload region. Specifically, when two particular codewords were adjacent to each other their sequences combined to create a binding site 5 HD from one of the accessing primers. While this was unintended, similar sequences can be avoided in the encoding process using thorough quality control measures that screen through all possible codeword combinations. Primer sequences are typically designed to be more than 8 HD from data payloads[16]; accordingly, we expect data encoding densities can remain unchanged when implementing File Preview since only a single primer pair is used per file.

However, it is important to note and consider that using more promiscuous conditions could increase off-target interactions more generally in the data payload regions even if all <10 HD sequences are avoided. This possibility should be investigated in the future as part of expanding our overall understanding of off-target interactions, particularly in extreme-scale systems. However, our work (Figures 2 & 3) suggests that the presence of <10 HD addresses in File Preview systems will outcompete interactions with higher HD off-target sequences that may be present in data payload regions. For example, while 4, 5, and 6 HD binding sites were very similar in sequence, stepwise decreases in accessing each HD set could be cleanly achieved by tuning PCR conditions. Thus, the chances of off-target interactions are most likely to occur within strands of the same file that have higher HD addresses rather than in strands of an undesired file. In addition, we did not observe off-target access from the randomized 1.5 GB data background in Figure 3h-i. Despite this, it would be prudent in the future to carefully assess within extreme-scale systems how increasing promiscuity of access conditions statistically increases the chances of inadvertently accessing strands from off-target files.

While previous DNA-based storage systems draw inspiration from conventional storage media and have had success, shifting the design paradigms to naturally leverage the intrinsic structural and biophysical properties of DNA holds significant promise that could transform the functionality, practicality, and economics of DNA storage. This work provides an archetype for a biochemically driven and enhanced data storage system.

Figures

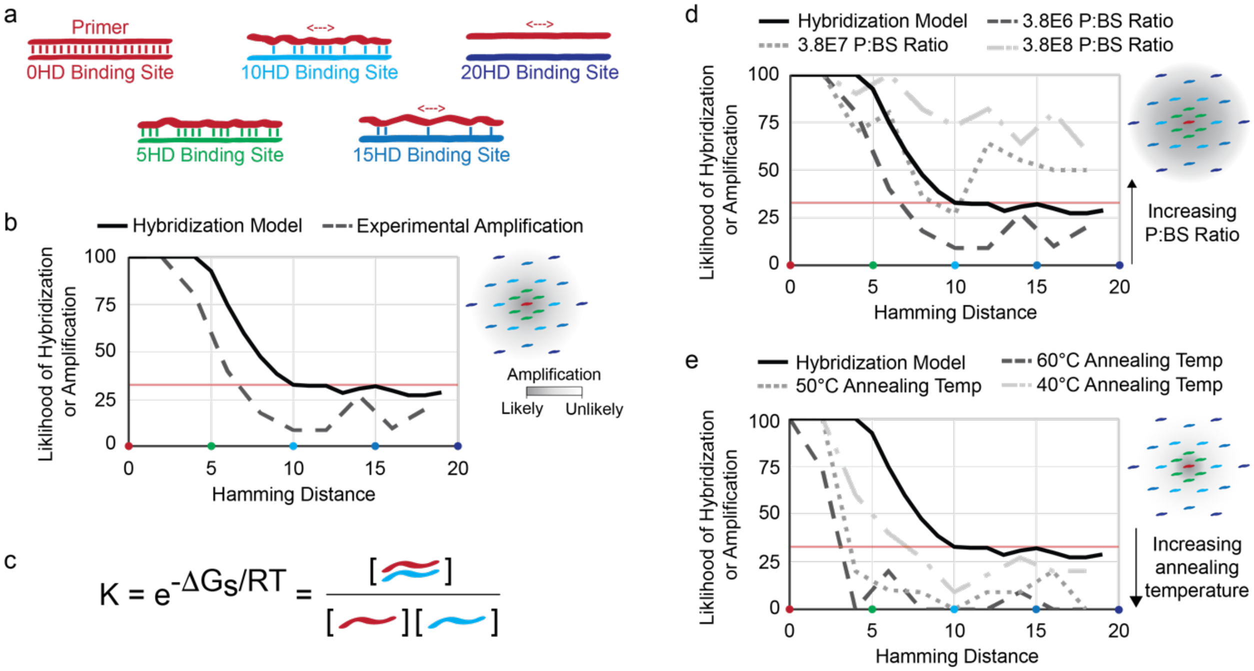
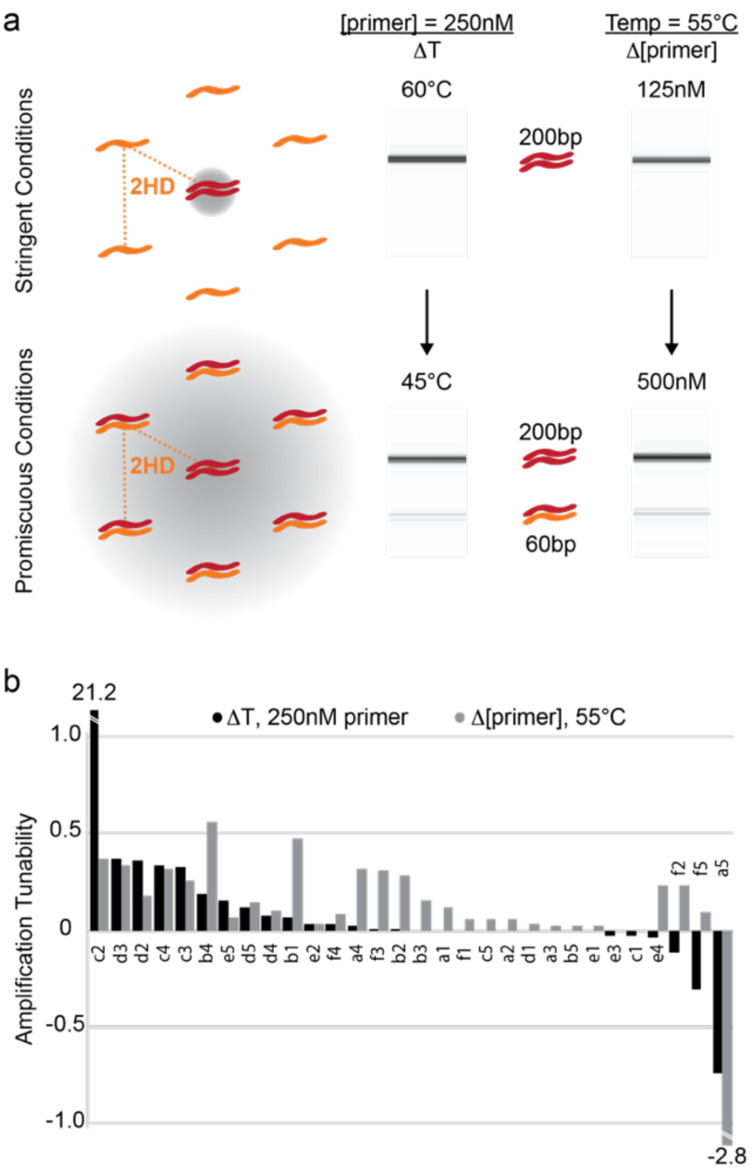


Figure 2.1 Stringency of PCR reactions is tunable via annealing temperature and primer concentration. (a) File address sequence similarity is inversely proportional to Hamming distance (HD – total number of nucleotide differences in a given sequence). While perfectly matching (0 HD) primers tightly bind their complementary binding sequence, primers with increasing HDs can still bind with gradually diminishing effect. (b) A thermodynamic model shows that the likelihood of hybridization (black trace) reaches a plateau (red line) around 10 HD and remains level out to 20 HD. Likelihood of amplification (grey dashed line) is represented as a percent of the 10 sequences experimentally tested at each HD (0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20) that created PCR product. Source data are provided as a Source Data file. (b, d, e, spherical insets) Visualization of address space with a perfect primer-address match at the center in red. DNA storage systems currently implement addresses that are at least 10 HD (light blue strands) apart. This disregards and wastes much of the potential address space (i.e. green strands). Grey shading indicates likelihood of hybridization/amplification with a red primer. (c) The equilibrium constant of thermodynamically controlled interactions can be computed based on the sequences' Gibbs Free energy (ΔG), the gas constant (R), and the PCR annealing temperature (T). Annealing temperature and primer concentration impact the amount of template amplified via PCR. (d) Primer concentration, represented by primer to binding site (P:BS) ratio, is experimentally varied at a constant 50°C annealing temperature. The likelihood of non-specific amplification increases with increasing P:BS ratio. Source data are provided as a Source Data file. (e) The annealing temperature is experimentally varied at a constant 1.9E7 P:BS ratio. The likelihood of non-specific amplification decreases with increasing annealing temperature. Source data are provided as a Source Data file.

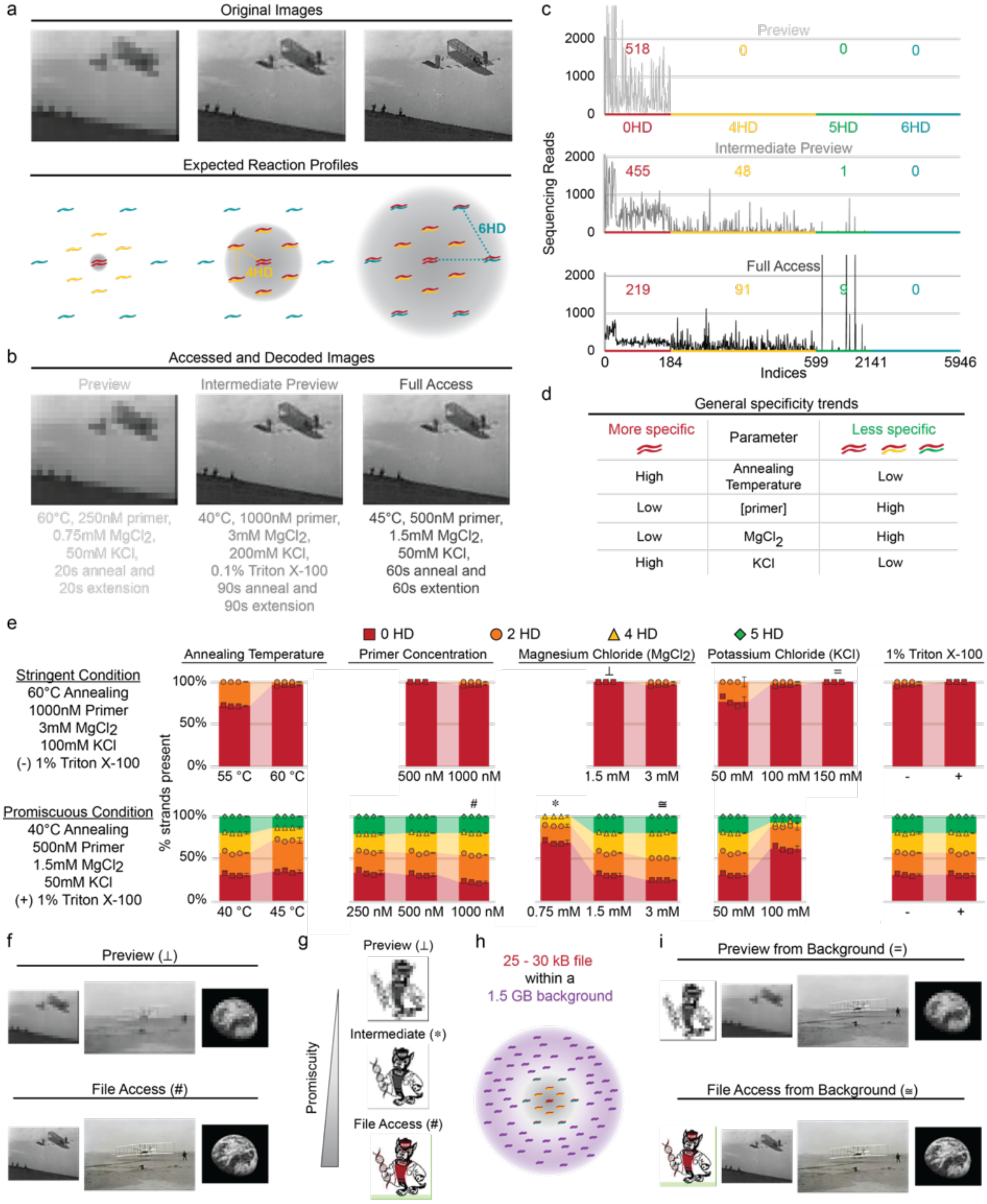


2HD forward binding sequence	2HD reverse binding sequence
a - AAGGTACGCAGTAAGCACTC	1 - ACCGTAGTCTTATTGCCACG
b - CAGGGACACAGTTAGCACTC	2 - TCCGTAGTCAGATTGCCATG
c - CAGGTACACAGTGAGCACTC	3 - TGCGTAGTCATATAGCCACG
d - CAGGTACGCAGTTTGCAATC	4 - TTCGTAGTCATATGGCCACG
e - CAGGTCCGCAGTTAGAACTC	5 - TTCGTAGTCGTATTGCCACG
f - CATGTTCGCAGTTAGCACTC	

Figure 2.2 Thermodynamic tuning of amplification within competitive PCRs.
(a) Strands with 2 HD binding sites (60 bp, orange) are screened for non-specific amplification in a competitive reaction against 0 HD strands (200 bp, red). Two stringent conditions are individually tested (top row): (left) 250nM primer at 60°C and (right) 125 nM at 55°C. The promiscuous conditions individually tested (bottom row) are (left) 250 nM primer at 45°C and (right) 500 nM at 55°C. Grey spheres encompass strands that are expected to be amplified, and the gel electrophoresis lanes show experimental results. Source data are provided as a Source Data file. (b) A screen of a library of sequences is conducted to find sequences to be used in scaling to full files and databases. Each forward binding sequence (letters a-f) is paired with every reverse binding sequence (numbers 1-5) in the reactions described in 2a. Amplification Tunability is defined as the difference in the ratio of mismatch (60bp) strands to perfect match (200 bp) strands from promiscuous to stringent conditions. Positive values represent tunability in the expected direction. Tunability using annealing temperature (black) and primer concentration (grey) are shown. Source data are provided as a Source Data file.

Figure 2.3 Implementing File Preview of jpeg images through thermodynamic swings. (a) Subsets of strands encoding increasingly more data to create a higher resolution image. The expected reaction profiles show that higher HD strands (0+4 and 0+4+6HD, middle and right, respectively) must be amplified to obtain the desired resolution image. (b) Experimental results showing high stringency to low stringency conditions are used to access and decode images. Intermediate Preview and Full Access result in identical images due to non-specific 5HD binding sites interfering with amplification of 6HD strands. (c) Next generation sequencing read counts versus strand index number for preview, intermediate preview, and full access conditions. Average read depth per strand index is listed above the corresponding HD regions. Most 5HD indices appear within 6HD strands, but their truncated amplification products are uniquely distinguished by NGS. Including 5HD products over-represents the number of unique file sequences: 5946 indices represent the number of amplification products; only 4405 unique strands actually encode the file. Source data are provided as a Source Data file. (d) A screen of environmental parameters reveals trends controlling the specificity of PCR amplification (data shown in Supplementary Figure 3a,b). (e) Environmental parameters are independently varied from one stringent and one promiscuous base condition. The percentage of 0 (red), 2 (orange), 4 (yellow), and 5HD (green) strands are measured by capillary electrophoresis. Wuflab logo file data shown here. Data points represent triplicate reactions. Error bars represent standard deviation. Center of error bars represents the mean of the triplicate reactions. Source data are provided as a Source Data file. (f) Preview (⊥) and File Access (#) conditions from (e) are selected to access three jpeg files, followed by NGS analysis. All files are successfully decoded. The image resolutions all increase from the Preview to the File Access conditions. (g) A gradient of Preview conditions is also achieved. Preview (⊥), Intermediate (*) and File Access (#) conditions from (e) successfully accessed the Wuflab logo as measured by NGS analysis. (h) A text file containing the Declaration of Independence is amplified using error prone PCR to create a noisy, non-specific background equivalent to 1.5GB of data. (i) Preview works in context of the GB-scale background, and all files are successfully decoded after amplification from the 1.5GB noisy background and NGS.

Figure 2.3 (continued)



General specificity trends

More specific	Parameter	Less specific
High	Annealing Temperature	Low
Low	[primer]	High
Low	$MgCl_2$	High
High	KCl	Low

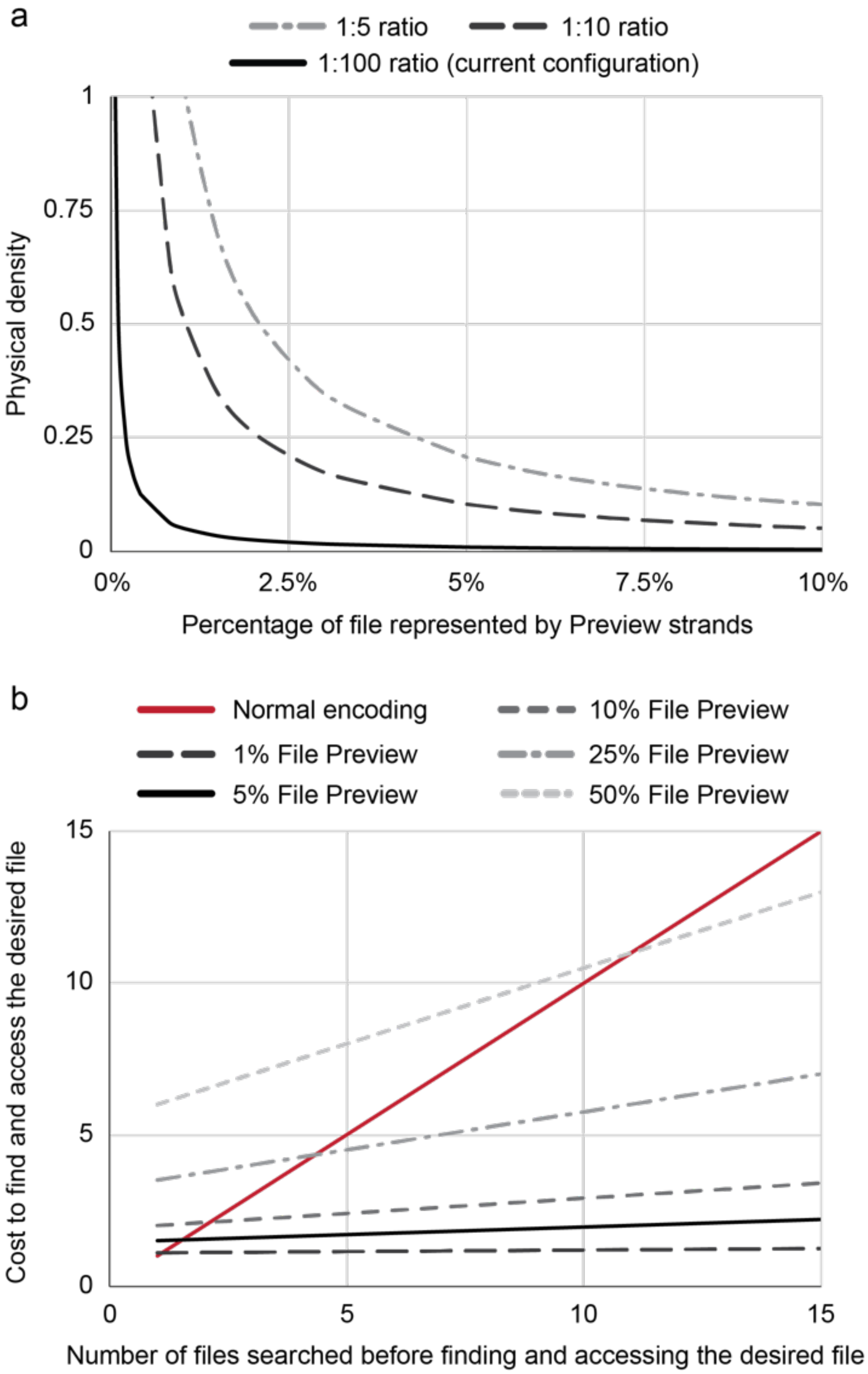


Figure 2.4 File Preview can be implemented at similar densities while reducing sequencing costs. (a) Physical density to store a given file using Preview encoding normalized to the physical density of normal encoding. When holding the ratio of total preview strands to total full file strands constant at 100 (black line, current configuration), 10 (grey dashed line), or 5 copies (light grey dashed-dot line), the physical density exponentially decreases as more of the file is stored in the Preview strands. Source data are provided as a Source Data file. (b) File Preview can cost effectively find a file in a library. Cost to find a file is defined as the normalized cost to fully sequence an entire file within a library, e.g., sequencing a 15-file database costs 15 on the y-axis. File preview can be used to quickly and cheaply find a file by sequencing a fewer total number of strands than is needed in a normally encoded library. Decreasing the percentage of each file stored in Preview strands further decreases the cost of finding a file. Source data are provided as a Source Data file.

Materials and Methods

Hybridization Model.

Hamming distance is frequently used as a metric in the design of primers to support random access of files in DNA-based storage systems because a high Hamming distance is an effective proxy for low hybridization likelihood. Hamming distance is a measure of how many symbols differ between two symbolic strings and in our case the strings of interest are two DNA primer sequences.

When analyzing two primers, p_1 and p_2, we compared their Hamming distance directly by lining up the sequences and counting the positions in which they are different. Hence, a Hamming distance of 0 means that the two primers were in fact the same sequence. If the Hamming distance was equal to the length of the primer, then every position was different. However, in terms of hybridization, we were interested in whether p_1 will bind to the reverse complement of p_2, as that binding site was present on the data strand. For convenience, we describe the Hamming distance of the two coding primers, but for hybridization, we analyzed the hybridization likelihood for p_1 against the reverse complement of p_2. Hence, a Hamming distance of 0 implies hybridization was guaranteed, however a high Hamming distance implies that hybridization was unlikely, although caveats existed. For example, if p_1 was the same as p_2 but merely shifted over one position, it had a high Hamming distance but a low edit distance. Such a high Hamming distant primer almost certainly bound due to low edit distance. To ensure that low edit distances do not skew the findings, primers with a much lower edit distance than Hamming distance were screened.

While high Hamming distances of 10 or more were common in past literature, low Hamming distances and their relationship to hybridization were of particular interest to our design. To better understand the potential of exploiting primer binding among similar but non-complementary primers, an in-silico analysis was used to predict the likelihood of primer hybridization as a function of Hamming distance. Our approach was based on a Monte Carlo simulation that considered the likelihood of hybridization among many primer pairs. One primer was designed specifically for data storage using procedures common in the field, namely it must have had GC-balance between 40-60%, melting temperature between 50°C and 55°C, and avoided long homopolymers. Then, it was randomly mutated into a new primer with varying Hamming distances, from 1 to N, where N was the length of the string. The mutated primer was produced by generating a random index from 1 to N with equal likelihood and randomly picking a new base for that position from the other three bases with equal probability. The mutation process was repeated until a primer with a suitable distance was achieved. Primers with a much lower edit distance were screened in this step, and it is worth noting that such primers had a very low probability due to the probabilistic nature of the mutate step; only a handful were observed over all trials. Using NUPACK's complex tool, the ΔG for the complex arising from the original primer binding to the reverse complement of the mutated primer was estimated[23]. Negative values beyond a threshold of -10 kcal/mol were interpreted as binding in our analysis. The Monte Carlo simulation included at least 10000 trials for each given Hamming distance to estimate the likelihood of hybridization. The percentage of mutated primers with a high chance of hybridizing for each Hamming distance is reported as the Hybridization % in Figure 1b.

The python program that performed this analysis is included in our code repository as part of the supplementary material[32].

Hamming Distance Primer Design.

Primers were selected for use in File Preview using a similar screening process as that for the Hybridization Model. However, instead of generating many trials, only a handful of primers were produced at each desired Hamming distance. These primers were then subjected to additional experimental screening.

Experimental model verification – qPCR amplification.

Using one primer sequence as the 0 Hamming distance amplifying primer, 10 variable strand addresses at each even numbered Hamming distance were used as templates strands for qPCR amplification (Supplementary Table 1). All strands were amplified using the same primer pair since they contained the same forward primer binding sequence while varying the reverse primer binding sequence. Reactions were performed in 6μL format using SsoAdvanced Universal SYBR Green Supermix (BioRad). A range of primer concentrations (125nM-500nM), template strand concentrations (2E3-2E6 strands/μL) and annealing temperatures (40-60°C) were tested. Thermocycler protocols were as follows: 95°C for 2 min and then 50 cycles of: 95°C for 10s, 40-60°C for 20s, and 60°C for 20s followed by a melt curve increasing from 60°C to 95°C in 0.5°C increments held for 5s each. Data were analyzed using Bio-Rad CFX Maestro. Cq value (i.e., cycle number at which a sample reached a defined fluorescence) and melt curve data (i.e., temperature a sample was denatured while being heated) were used for analysis. Successful amplifications were defined as crossing the Cq threshold before the negative control while also creating an individual peak (i.e., single product) on the melt curve.

Competitive PCR primer reactions.

Using four distinct primer pairs as the 0 Hamming distance amplifying primers, 5-30 unique strands (60bp) containing variable address pairs at 2, 3, 4, or 6 Hamming distance were tested as template strands alongside 0 HD strands (200bp) in competitive qPCR amplifications (Supplementary Table 2). All strands designed using the same original primers were amplified using the 0 HD primer pair. Reactions were performed in 6µL format using SsoAdvanced Universal SYBR Green Supermix (BioRad). Template strand concentrations were in equal copy number concentration for the 0 HD and variable HD strands (1.67E5 strands/µL). A range of primer concentrations (125nM-500nM) and annealing temperatures (40-60°C) were tested. Thermocycler protocols were as follows: 95°C for 2 min and then 50 cycles of: 95°C for 10s, 40-60°C for 20s, and 60°C for 20s. Final products were diluted 1:60 in 1xTE before analysis using high-sensitivity DNA fragment electrophoresis (Agilent DNF-474; Advanced Analytical 5200 Fragment Analyzer System; Data analysis using Prosize 3.0). The ability for a primer to variably amplify a strand with a non-specific primer binding site at different PCR conditions, or Amplification Tunability, was calculated using the following equation (concentrations in nmole/L):

$$(2.1)\ \text{Amplification Tunability} = \Delta\left(\frac{[\text{nonspecific strand}]}{[\text{specific strand}]}\right) = \left(\frac{[\text{ns strand}]}{[\text{s strand}]}\right)_{\text{Promiscuous}} - \left(\frac{[\text{ns strand}]}{[\text{s strand}]}\right)_{\text{Stringent}}$$

JPEG Encoding for File Preview Operations.

File Preview was performed on JPEG images due to their widespread popularity, their small storage footprint, and their support for organizing data within a file that works synergistically with the goals of File Preview in this work. In particular, JPEG's progressive encoding[33] allowed for image data to be partitioned into scans by color band and by frequency. Through careful organization of the file, a low-resolution grayscale image was constructed from a small percentage of the file's data or an increasingly higher resolution image was obtained from reading a greater percentage of the file[32]. For the File Preview operations, the JPEG information was arranged in such a way that a 0-HD access pulled out a small amount of data and produced a low-resolution image. By tuning the access conditions as described, more of the file was accessed and greater resolution image was produced.

Important details of the JPEG file format. The most important aspects of the JPEG format are described for the sake of explaining how Preview works. JPEG holds information in three color bands known as Y, Cb, Cr that together encode information for all pixels in an image. Y represents luminosity, Cb is a blue color band, and Cr is a red color band. Together, these components may represent any conventional RGB color. Each pixel of an image can be thought of as having a tuple of Y, Cb, and Cr components although they are not actually stored that way.

JPEG does not store images in a naïve matrix of (Y,Cb,Cr) pixel values. This would waste storage since many pixels have the same color. Instead, each 8x8 block of pixels from each color band are converted into a frequency domain representation using the 2-D Discrete Cosine Transform (DCT). The 2-D DCT has the interesting effect of partitioning the data into low frequency and high frequency components. Each 8x8 block becomes a linearized vector of 64 values ordered from low frequency to high frequency. The first value in the vector is known as the DC value because it represents an average value across the original 8x8 pixel block. For example, if the original 8x8 block were entirely white, the Y band would have a DC value of 255, indicating the average value over the block was white. The remaining 63 entries represent the higher frequency components known as the AC band. For an all-white block, the rest of the vector would be 0, indicating no other content.

In a progressive encoding, each color band is encoded in scans. A scan is the aggregation of all values from a given position in the linearized vector across all 8x8 blocks. For example, the first scan of a file would include all of the DC values from the Y band across all 8x8 blocks. The scan of DC values for a given band is given as Y[0], Cr[0], or Cb[0]. The Y[0] scan by itself is essentially a low resolution grayscale image. Cr[0] and Cb[0] would add low resolution color information.

The DC scans precede the AC scans. The AC scans group the following AC components together, and these scans could include a single value from the linearized vector or multiple values. For example, Y[1:5] would include indices 1 through 5 of the linearized vectors taken from all 8x8 blocks in the Y band. All indices from 1 through 63 must be included in at least one scan. This is repeated for all bands. The JPEG standard additionally compresses each scan to save storage space, but the details of that mechanism are not pertinent to Preview and are omitted. Furthermore, the scans follow the standard and are stored in compressed form.

Partitioning the JPEG file for Preview. The JPEG files were first encoded into 42 scans: Y[0], Cr[0], Cb[0], Y[1:5], Cb[0] ,Cr[0], Y[6:10], Y[11:15], Y[16:20], Y[21:25], Y[26:30], Y[31:35], Y[36:40], Y[41:45], Y[46:50], Y[51:55], Y[56:60], Y[61:63], Cb[1:5], Cr[1:5], Cb[6:10], Cr[6:10], Cb[11:15], Cr[11:15], Cb[16:20], Cr[16:20], Cb[21:25], Cr[21:25], Cb[26:30], Cr[26:30], Cb[31:35], Cr[31:35], Cb[36:40], Cr[36:40] ,Cb[41:45], Cr[41:45], Cb[46:50], Cr[46:50], Cb[51:55], Cr[51:55], Cb[56:60], Cr[56:60], Cb[61:63], Cr[61:63].

The scans were grouped into partitions. Wuflab logo and Wright Glider 2 had 4 partitions, and Wright Glider 1 and Earth had 3 partitions. In all cases, the first and second partitions, if accessed alone, provided low resolution images that are recognizable as the image. For the Wuflab logo and Wright Glider 2 files, the third partition contained all remaining scans. For the others, the third partition added DC color information and some higher frequencies of the Y band to improve image quality, and the fourth partition contained all remaining scans.

Each partition was treated as a block of data and encoded into DNA as a unit. Each partition was tagged with primers. Higher numbered partitions were given primers with a greater Hamming distance.

Encoding for Error Correction. The encoding process is described in Supplementary Figure 7. Each partition was encoded into DNA using a multi-level approach. First, the JPEG file was partitioned into scans. Then, each partition was divided into blocks of 1665 bytes, which were interpreted as a matrix with 185 rows and 9 columns with one byte per position. Blocks smaller than 1665 bytes at the end of a file or partition were padded out with zeros. An RS outer code with parameters of [n=255,k=185,d=71] added additional rows to each block to compensate for possible loss of strands within a block. Each row was given a unique index that was two bytes long. Then, each row was appended with error correction symbols using an RS inner code given as [n=14,k=11,d=4] that protected both the data and index bytes.

Each row of byte was converted into a DNA sequence using a comma-free code that maped each byte to a unique codeword sequence. The codewords were designed using a greedy algorithm to be GC-balanced and have an edit distance of at least 2 to all other codewords. Each codeword had a length of 8 nts. The last step was the appending of primers to each end of the sequence and insertion of a restriction enzyme cut site in the middle of the strand. Each partition of the JPEG file used different primer binding sites, so these primer sequences were given as inputs for each partition as it was encoded.

An additional set of flanking primers were added to each strand to enable the entire library to be amplified at once using a single common primer. The final set of strands for each file were synthesized into a DNA library.

PCR condition screening and File Preview.

The four-file synthetic DNA library was ordered from Twist Biosciences. Flanking primer amplifications unique to each subset of strands (Supplementary Table 3) were optimized and the resulting products were used in screening and preview reactions. Each subset of strands within a file encodes an increasing percentage of the stored image and contains a unique restriction enzyme cut site to allow for rapid sample analysis. It was determined that each block of data encoded in strands with increasing Hamming distance binding sites (2, 4, and 6HD), needed to be physically stored with extra copies of the non-specific strands: 10x, 100x, and 1000x, respectively. A screen of variable PCR conditions was conducted on files from the library prior to preview and full access reactions. Reactions were performed in 6-50µL format using SsoAdvanced Universal SYBR Green Supermix (BioRad) or Taq polymerase (Invitrogen). Conditions varied during testing include: Annealing Temperature (40-60°C), annealing and extension timing (20-90s), number of cycles (25-40), primer concentration (62.5-1000nM), polymerase concentration (0.5-2x recommended units), dNTP concentration (200-800µM), $MgCl_2$ concentration (0.75-3mM), KCl concentration (50-200mM), and absence or presence of 0.1-1% Triton X-100, 0.1-1% BSA, 0.1-1% Tween-20, 2-8% DMSO, 0.1-3.5mM Betaine, or 2% DMSO plus 0.1mM Betaine. Reaction products (1µL) were added to restriction enzyme reactions to cut 0, 2, 4, or 6 HD sections of the products. Digestion products were diluted 1:3 in 1xTE for analysis using high-sensitivity DNA fragment electrophoresis (Agilent DNF-474; Advanced Analytical 5200 Fragment Analyzer System; Data Analysis using Prosize 3.0). Quantification data was taken directly from Fragment Analyzer. Undigested preview, full access and intermediate samples were then analyzed via Illumina Next-Generation Sequencing (Genewiz and AmpliconEZ).

Error Prone PCR.

Template DNA was amplified using 0.5µL of Taq DNA polymerase (5 units/µL, Invitrogen) in a 50µL reaction containing 1× Taq polymerase Rxn Buffer (Invitrogen), 2mM

MgCl2 (Invitrogen), the sense and antisense primers at 1E13 strands each, and dATP, dCTP, dGTP, dTTP (NEB), dPTP (TriLink), and 8-oxo-dGTP (TriLink), each at 400mM. PCR conditions were 95°C for 30s, 50°C for 30s, and 72°C for 30s for 35 cycles with a final 72°C extension step for 30s.

Calculation of Data Quantity of Error Prone Background.

In Figure 3h, i and Supplementary Figure 4, we refer to background size (GB). For clarity and ease of comparison, this value was calculated based on the total number of DNA strands. Each strand is comprised of 200 nts, 20 of which are used for each primer sequence, 16 for the index, and 8 for the checksum. Eight nts comprise each 1-byte codeword. Thus, each strand addressed with a single primer pair contains 17 bytes of data. We assumed a 10-copy physical redundancy per unique strand to provide a conservative estimate for a realistic system where multiple copies of each strand would likely be needed to avoid strand losses and inhomogeneous strand distributions. Thus, the total background size is divided by 10.

Next Generation Sequencing and File Preview Decoding.

FASTQ files obtained from sequencing were all decoded successfully into images. Decoding occurred in the reverse order shown in Supplementary Figure 7. Files were reconstructed by placing all data blocks and JPEG file partitions into the correct order based on their index. Since error correction was applied separately to each partition, each partition succeeded or failed at partition boundaries. If a partition was incomplete, it was omitted from the JPEG image. As long as omitted partitions were the latter partitions taken from AC scans, their absence only reduced the quality of the JPEG image and made it appear lower resolution or grayscale, depending on the scans that were lost in the partition. However, if the first partition in the file was missing or too erroneous to decode, the image would be unreadable. No experiment yielded an undecodable or unreadable image. The successfully decoded images are shown in Figure 3b, f, and h.

To gain deeper insight into which strands were sequenced and their relative abundance, a clustering analysis was performed on all sequenced reads[32]. The Starcode algorithm is an open source and efficient algorithm for performing an all-pairs search on a set of sequencing data to find all sequences that are within a given Levenshtein distance to each other[34]. To derive the number of reads for each encoded strand in the library, the algorithm was seeded with 20 copies of each strand from the library. The Starcode algorithm was additionally given the following parameters: Levenshtein distance set to 8 edits, the clustering algorithm set to message passing, and the cluster ratio set to 5. The Levenshtein distance parameter defines the maximum edit distance allowed when determining whether a strand belongs to a cluster. The clustering algorithm attributed all reads for a given strand *S* to another strand *V* provided that the ratio of *V's* reads to *S'*s reads were at least the cluster ratio. Hence, providing 20 copies of each expected strand ensured that each was well represented during clustering such that it was considered a centroid. With the clusters formed, each centroid was interrogated to make sure that it was a strand from the library and not an unexpected strand present during sequencing. If the centroid matched an expected strand defined by the encoded file(s), the number of reads for that strand was adjusted to match the size of the cluster less 20. These results are reported in Figure 3c.

Acknowledgments

We thank Kevin N. Lin, Karishma Matange, Magdelene Lee, and Doug Townsend for fruitful discussions during group meetings. We thank the NC State Biotechnology Program for providing the Wuflab logo. We also thank Joshua Emrick for his helpful comments on the manuscript. This work was supported by the National Science Foundation (Grants CNS-1650148, CNS-1901324, ECCS 2027655) and a North Carolina State University Research and Innovation Seed Funding Award (Grant 1402-2018-2509). K.J.T. was supported by a Department of Education Graduate Assistance in Areas of National Need Fellowship. E.W.I. was supported by funds from the North Carolina State University Research Experience for Undergraduates, Provost's Professional Experience Programs, and NCSU startup funds.

Author contributions

K.J.T., J.M.T., and A.J.K. conceived the study. K.J.T., E.W.I., and A.J.K. developed, performed, and analyzed the wet lab experiments. K.V. and J.M.T. developed and performed the software, simulations, file design, and next generation sequencing analysis with guidance from all. K.J.T. and A.J.K. wrote the paper with input from all.

References

1. Rydning, J. & Reinsel, D. Worldwide Global StorageSphere Forecast, 2021–2025: To Save or Not to Save Data, That Is the Question. IDC White Pap. (2021).

2. Cox, J. P. Long-term data storage in DNA. Trends Biotechnol 19, 247–250 (2001).

3. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed Engl 54, 2552–2555 (2015).

4. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. Nat. Mater. 15, 366–370 (2016).

5. Valk, T. Van Der et al. Million-year-old DNA sheds light on the genomic history of mammoths. Nature 591, 265–269 (2021).

6. Matange, K., Tuck, J. M. & Keung, A. J. DNA stability: a central design consideration for DNA data storage systems. Nat. Commun. 12, (2021).

7. Illumina , Microsoft , Twist Lead New DNA Data Storage Alliance. Genetic Engineering & Biotechnology News (2020). Available at: https://www.genengnews.com/news/illumina-microsoft-twist-lead-new-dna-data-storage-alliance/.

8. Yazdi, S. M., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. Sci Rep 5, 14138 (2015).

9. Lin, K. N., Volkel, K., Tuck, J. M. & Keung, A. J. Dynamic and scalable DNA-based information storage. Nat. Commun. 11, (2020).

10. Stewart, K. et al. A Content-Addressable DNA Database with Learned Sequence Encodings. in 24th International Conference on DNA Computing and Molecular Programming 11145 LNCS, 55–70 (2018).

11. Bee, C. et al. Content-Based Similarity Search in Large-Scale DNA Data Storage Systems. bioRxiv 2020.05.25.115477 (2020). doi:10.1101/2020.05.25.115477

12. Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. Science 337, 1628 (2012).

13. Bornholt, J. et al. A DNA-based archival storage system. in Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS '16 337, 637–649 (2016).

14. Tomek, K. J. et al. Driving the Scalability of DNA-Based Information Storage Systems. ACS Synth. Biol. 8, (2019).

15. Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. Nat. Commun. 11, (2020).

16. Organick, L. et al. Random access in large-scale DNA data storage. Nat. Biotechnol. 36, 242–249 (2018).

17. Mathews, D. H., Burkard, M. E., Freier, S. M., Wyatt, J. R. & Turner, D. H. Predicting oligonucleotide affinity to nucleic acid targets. Rna 5, 1458–1469 (1999).

18. Zhang, J. X. et al. Predicting DNA hybridization kinetics from sequence. Nat. Chem. 10, 91–98 (2018).

19. Tanaka, F., Kameda, A., Yamamoto, M. & Ohuchi, A. Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. Nucleic Acids Res 33, 903–911 (2005).

20. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. Science 355, 950–954 (2017).

21. Yamamoto, M., Kashiwamura, S., Ohuchi, A. & Furukawa, M. Large-scale DNA memory based on the nested PCR. Nat. Comput. 7, 335–346 (2008).

22. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. BMC Biotechnol. 8, 91 (2008).

23. Zadeh, J. N. et al. Software News and Updates NUPACK: Analysis and Design of Nucleic Acid Systems. J Comput. Chem 32, 170–173 (2011).

24. Yamamoto, M., Kameda, A., Matsuura, N., Shiba, T. & Ohuchi, A. Simulation analysis of hybridization process for DNA computing with concentration control. 1, 85–90 (2002).

25. Tanaka, F., Kameda, A., Yamamoto, M. & Ohuchi, A. Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop. Biochemistry

43, 7143–7150 (2004).

26. SantaLucia, J. & Hicks, D. The thermodynamics of DNA structural motifs. Annu. Rev. Biophys. Biomol. Struct. 33, 415–440 (2004).

27. Abu Al-Soud, W. & Rådström, P. Capacity of nine thermostable DNA polymerases to mediate DNA amplification in the presence of PCR-inhibiting samples. Appl. Environ. Microbiol. 64, 3748–3753 (1998).

28. Kramer, M. F. & Coen, D. M. Enzymatic Amplification of DNA by PCR: Standard Procedures and Optimization. Curr. Protoc. Cytom. 37, A.3K.1-A.3K.15 (2006).

29. Feng, B. et al. Hydrophobic catalysis and a potential biological role of DNA unstacking induced by environment effects. Proc. Natl. Acad. Sci. U. S. A. 116, 17169–17174 (2019).

30. Zaccolo, M. & Gherardi, E. The Effect of High-frequency Random Mutagenesis on in Vitro Protein Evolution : A Study on TEM-1 b -Lactamase. J. Mol. Biol. 285, 775–783 (1999).

31. Manber, U. Finding Similar Files in a Large File System. in USENIX Winter 1994 (1994).

32. Tomek, K. J., Volkel, K., Indermaur, E. W., Tuck, J. M. & Keung, A. J. Promiscuous molecules for smarter file operations in DNA-based data storage. dna-storage/ncomm-file-preview. doi:10.5281/zenodo.4747693 (2021).

33. Wallace, G. K. The JPEG still picture compression standard. IEEE Trans. Consum. Electron. 38, xviii–xxxiv (1992).

34. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: Sequence clustering based on all-pairs search. Bioinformatics 31, 1913–1919 (2015).

CHAPTER 3: High-throughput primer binding analysis

Kyle J. Tomek[1], Kevin Volkel[2], James M. Tuck[2], Albert J. Keung[1]

1. Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27606

2. Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606

Introduction

High-capacity DNA storage systems will require a large number of available file addresses to organize and access data. This is true for both PCR-based access systems and non-PCR-based systems[1–4]. However, due to increasing probabilities for potential off-target molecular interactions as systems scale in capacity, addresses must be sufficiently different from each other in sequence and are, therefore, finite in number and limit total system capacities[5]. Sequence limitations are a product of database design: file addressing, retrieving, and data encoding currently depend on the same pool of nucleotide sequences[2,4–6]. In other words, complex systems – those with larger datasets – have data and addresses that are too similar in sequence to file addresses used in polymerase chain reaction (PCR) based random access and DNA hybridization based systems[2–4]. Consequently, the file addresses used for access (i.e., primers) begin to bind 'off-target' sequences in other file addresses and data payload regions resulting in a diminished ability to selectively retrieve desired information. Yet, there has not been a concerted effort to fully understand these primer interactions in DNA storage systems[7]. Using open-source software (NUPACK), suggestions from prior publications, and software models, we developed tools that allowed us to design primers and pick ones that are better to use together and with our data within the same library. Here we develop an experimental strategy to screen many combinations of primers and mismatched binding sites. We find a directional bias in ligation-based screening, work to optimize a polymerase-based test, and plan future hybridization reactions for enzyme free systems. This study will not only identify real primer sequences that can be used in practical storage systems, but also inform a computational primer design model.

Experimental Design and Results

Although there have been studies using numerous files within complex background libraries[5,6,8] and comprehensive analysis of 4-nt overhangs during DNA ligation[9], there has yet to be a high-throughput approach experimentally screening the ~20-nt address sequence space. Our strategy aims to screen 3,798 known sequences against a library of totally randomized, degenerate sequences capturing the entire address space of ~1-trillion sequences (4 bases^20 nt > 1 trillion unique sequences). The 3,798 sequences were designed to be a variable region of a hairpin DNA structure[9]. Each hairpin oligo has this variable overhang primer binding site, a unique barcode region, consensus primer binding site, restriction enzyme recognition site, spacer, and hairpin loop sequence (**Fig. 3.1a and Supplementary Fig. C.1**). The reverse complement sequences are located 3' to their complementary regions and located on the same strand to create self-complementarity. Once the DNA forms the hairpin structure, only the ssDNA variable overhang region can interact with external sequences unless denatured through thermocycling. A degenerate 20-nt primer (i.e., totally randomized 20-nt sequence), also containing a 20-nt poly-A tail for later processing, was added to the hairpin library. Variant sequences that bind to the ssDNA overhang sequences are ligated using T4 ligase which lock the primer binding event into the sequence of the DNA. The degenerate oligo library was ordered with a 5' phosphate group to facilitate efficient ligations. Since the barcode sequence associated with each unique overhang is located on the same strand as the newly bound and ligated primer sequence, we can determine which overhang sequence bound to which degenerate oligo sequence directly using NGS. The ligation product was digested with a restriction enzyme to remove the hairpin loop and amplified using the consensus primer sequence and an oligo-dT20 primer to ensure there was enough DNA for NGS.

The system was initially tested and optimized using a single hairpin sequence ordered from Genewiz to determine if this design is suitable for testing primer binding interactions in a high throughput setting. The hairpin structure was formed by heating the oligo to above the hairpin Tm (78C) and cooled slowly to room temperature (1C/30s). A perfectly complimentary oligo (with the additional 3' poly-A tail and 5' phosphate) was also ordered to bind the overhang on the hairpin structure. These hairpin and primer oligos were combined in three distinct molar ratios (primer:hairpin = 1:1, 5:1, and 1:5) to be annealed, ligated, digested, and amplified. The original strands and a portion of the product from each step were analyzed via capillary electrophoresis (**Supplementary Fig. C.2**). The highest primer to hairpin ratio was the best for binding the most hairpin structures (**Supplementary Fig. C.2e**). Importantly, the amplification using the consensus and oligo-dT primers resulted in the same size DNA for all three reaction conditions (**Supplementary Fig. C.2f-h**). A panel of 20-nt sequences designed to not bind (hamming distance above 10) the overhang region of this test hairpin was tested for annealing. Analysis showed no indication of non-specific binding (**Supplemental Fig. 3**). One of these sequences was arbitrarily chosen for subsequent experiments as the negative control.

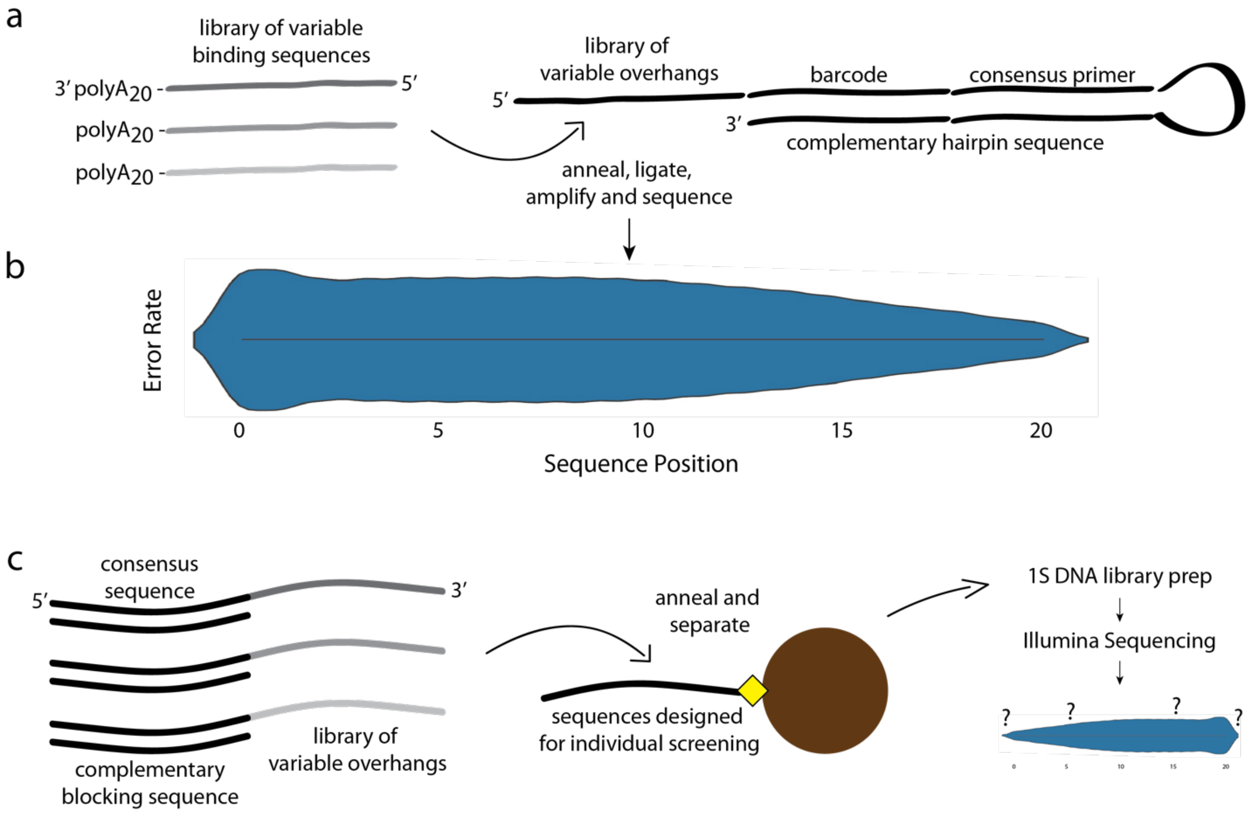

a
library of variable binding sequences
3′ polyA20
5′
polyA20
polyA20
library of variable overhangs
5′
barcode
consensus primer
3′
complementary hairpin sequence
anneal, ligate, amplify and sequence
b
Error Rate
0
5
10
15
20
Sequence Position
c
consensus sequence
5′
3′
library of variable overhangs
complementary blocking sequence
anneal and separate
sequences designed for individual screening
1S DNA library prep
Illumina Sequencing
?
?
?
?
0
5
10
15
20

Figure 3.1. High-throughput platforms used to explore DNA interactions.
(a) Experimental design and simplified strand architecture of a ligation-based screening approach. Each of the 3798 hairpin oligomers has a variable overhang primer binding site, unique barcode region, consensus primer binding site, and consensus hairpin loop sequence. A totally randomized library of variable binding sequences with polyA tails bind the overhanging hairpin strand, are ligated using T4 DNA ligase, amplified, via PCR, and sequenced using Illumina based sequencing. (b) The sequence position of mismatched base pairs on the library of variable binding sequences is biased towards fewer errors at the 3' end of the variable overhang. (c) Experimental design and strand architecture of a hybridization-based screening approach. Individual sequences are designed as potential primers to be used in a DNA storage system. The totally randomized library of variable overhangs also contains a consensus sequence for binding a complementary blocking sequence. The variable library will be added to the reaction containing the individual strands bound to magnetic beads, allowed to hybridize, separated, and prepped for Illumina based sequencing and subsequent analysis.

To properly capture the sequence space during the high throughput experiment, there will need to be adequate copies of each hairpin oligo present in each step of the protocol. Therefore, starting template copy numbers were serially diluted in 10-fold increments to test the lower limits of the system. Results show that 1E9 and 1E8 starting hairpins allow for amplification of the ligated and digested product while 1E7 and 1E6 show a reduction in the final PCR product yield

(**Supplementary Fig. C.3**). Skeptical if the digestion step is necessary for proper PCR amplification, we tested the same starting copy numbers while skipping restriction digestion. Results again show proper amplification starting with 1E9 and 1E8 copies and limited amplification with 1E7 and 1E6 copies (**Supplementary Fig. C.4**). Additionally, when comparing results with and without the restriction digest, samples that were digested have much cleaner peaks (i.e., purer PCR product).

After proving the system would work experimentally, the library of 3,798 variable overhang hairpin sequences was designed and ordered alongside the randomized binding sequences. Successfully screening sequences and analyzing data requires knowledge of how abundant, on average, each random primer should be in the screening process. We look at four different coverages (1, 5, 10, 50), where coverage is the expected number of copies for each random primer. For example, a coverage of 50 means that, on average, 50 copies of each variable primer should be present at the initialization of the experiment. Since the primers that we are investigating are 20 bases long, the total number of random primers for any given coverage, *C*, is $C*4^{20}$. Experimentally, too high of a coverage can lead to a prohibitively high variable primer concentration (i.e., synthesis at such a scale would be cost prohibitive or too many random sequences would cause uncontrollable interactions), but too low of a coverage will not allow for screening the entire sequence space. After performing the previously discussed experimental protocol for each of the four different coverages we sequence and analyze the hybridization events using next generation sequencing. For each coverage, 1, 5, 10, and 50, there were 41,416, 35,961, 33,690, 14,518 recorded hybridization events, respectively, out of a total of 45,580, 44,531, 41,796, and 38,027, raw sequencing reads, respectively. There are fewer recorded hybridizations than raw reads since our sequencing data processing throws out strands in which there is not high confidence that a strand conveys a hybridization event. This can occur if the sequenced strand is too short, the suspected barcode region is unrecognizable, and if the consensus primer cannot be found. The number of unique events for coverages 1 and 5 tracks closely to the number of considered reads (**Supplementary Fig. C.5, red line**). Once the coverage gets to 10 and 50, there is an increasing departure from the considered reads indicating that there are events that are measured more than once. These results indicates that coverage may indeed influence what is perceived as a strong hybridization binding. The significant decrease in total hybridization events for a coverage of 50 conceivably stems from random primer probes that have not hybridized to any given target library primer remaining in solution throughout the reaction and into sequencing.

Supplementary Figure C.6 compares the distribution of hybridization events for a given edit distance (similar to hamming distance). Both coverages observe reads with edit distances of 0 (perfect matches between the random and library primer), but only a coverage of 50 is able to separate these events from those that have very few reads due to poorer matches with high edit distance. Interestingly, the distribution for edit distance (right side bar graph) experiences an increase of event counts at lower edit distances when going from a coverage of 5 to 50. This is likely because the space of mismatches at high edit distances is much larger than that of exact matches and low edit distances, and a higher coverage increases the probability that lower edit distance matches can interact.

Lastly, as we are interested in determining design rules for primers, we look at the locations in which mismatches occur between the library and random primer (**Figure 3.1b and Supplementary Figure C.7**). Clearly from these graphs, we can see that total mismatches decrease significantly towards the higher order indexes of the primers (3' end of the 3,798 variable overhang library). This indicates that matches at this end are more important for observing hybridization events, which is consistent with the experiment setup since ligation is done at the 3' end relative to the library primer in the experiment schematic.

Future Directions

Since 3' sequence bias is found to impact primer ligation we have designed a PCR-based experimental workflow to investigate this bias in polymerase-based reactions. As most DNA data storage systems utilize PCR for making copies and accessing files a PCR-based system will be important to study. We expect a similar 3' bias trend is likely to be found based on the enzymes' mechanisms of DNA binding and amplification. Preliminary trials of a system containing the 3,798 addresses interacting with the >1 trillion random 20-nt primers have been unsuccessful to date. We believe that the unsuccessful reactions are been due to the ever-increasing complexity of the reactions captured by the inherent properties of PCR. When strands are denatured during thermocycling, new binding sites for the randomized primer library are presented. When primers bind to unpredictable sequences, an almost infinite number of unexpected products can be formed. When these new copies are then amplified, they will participate in subsequent amplifications leading to an ever-increasing complex set of interactions throughout the reaction. Future work will continue to optimize and troubleshoot PCR based reactions.

While most DNA data storage systems are reliant on PCR, there are recent examples of systems which utilize DNA-DNA interactions for file access, database searches, and system functionality. We have designed a third high throughput workflow to invest an enzyme independent protocol for screening DNA-DNA hybridizations (Figure 3.1c). Several 20 nt sequences were designed as potential primers for a DNA storage system to be screened against another totally random variable library of >1 trillion, 20 nt sequences. For this experiment the randomized sequence will also contain a consensus sequence which will help facilitate NGS analysis. A sequence complementary to the consensus sequence will be annealed to the library as a blocking sequence before screening against the known primers. After the potential primers are bound to a magnetic bead, the randomized library will be added to the known primers, allowed to anneal, physically separated using the magnetic beads, and prepped for Illumina based sequencing and analysis. This study will potentially identify real primer sequences that can be used in practical storage systems and help build and inform a more sophisticated computational primer design model.

References

1. Bornholt, J. et al. A DNA-based archival storage system. in Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS ’16 337, 637–649 (2016).

2. Organick, L. et al. Random access in large-scale DNA data storage. Nat. Biotechnol. 36, 242–249 (2018).

3. Stewart, K. et al. A Content-Addressable DNA Database with Learned Sequence Encodings. in 24th International Conference on DNA Computing and Molecular Programming 11145 LNCS, 55-70 (2018).

4. Lin, K. N., Volkel, K., Tuck, J. M. & Keung, A. J. Dynamic and scalable DNA-based information storage. Nat. Commun. 11, (2020).

5. Tomek, K. J. et al. Driving the Scalability of DNA-Based Information Storage Systems. ACS Synth. Biol. 8, 1241–1248 (2019).

6. Tomek, K. J., Volkel, K., Indermaur, E. W., Tuck, J. M. & Keung, A. J. Promiscuous molecules for smarter file operations in DNA-based data storage. Nat. Commun. 12, (2021).

7. DNA Data Storage Alliance. PRESERVING OUR DIGITAL LEGACY : AN INTRODUCTION TO DNA DATA STORAGE. (2021).

8. Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. Nat. Commun. 11, 1–7 (2020).

9. Potapov, V. et al. Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. ACS Synth. Biol. 7, 2665–2674 (2018).

CHAPTER 4: DNA-based data synthesis via codeword assembly methods

Kyle J. Tomek[1], Kevin Volkel[2], James M. Tuck[2], Albert J. Keung[1]

1. Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27606

2. Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27606

Introduction

Current DNA data library production relies on parallelizing phosphoramidite synthesis reactions[1,2]. This method of synthesis can create any arbitrary DNA sequence; therefore, it is important for synthesizing PCR primers, genetic engineering, and now, DNA data storage. Although optimization of the base-by-base reactions has been ongoing for decades, processes today can only create libraries of DNA strands up to 200-300 bases in length while maintaining tolerable error rates[3–5]. The cost and environmental toxicity of the reagents has already been scaled significantly by a mature industry yet remain prohibitive for the immense scale needed for DNA data storage compared to current biomedical applications. Additionally, the process is slow as it relies on microfluidic devices for liquid handling and requires as many cycles as the length of desired DNA. Consequently, there has been a push from academic and industrial research groups to develop higher-throughput and cheaper DNA synthesis technologies[6,7]. Recently, enzymatic approaches to DNA synthesis have promised to decrease latency and cost while improving oligonucleotide throughput and sequence accuracy[8,9]. Fortunately, enzymatic synthesis methods use sustainable aqueous reagents, which produce fewer waste byproducts.

A key insight is that for DNA storage it is not necessary to be able to create every arbitrary DNA sequence. In binary data storage systems, a byte is made up of 8 bits (0 or 1), therefore there are 2^8 (256) potential bytes which can be arranged to represent any and all data. In a DNA based system, each of these 256 bytes can be represented by a unique block of nucleotides (i.e., ‘codeword’) which can be assembled to produce long strands with tolerable error rates. The advantage of this is threefold. First, short codewords can be synthesized in bulk at considerable cost-savings since it is relatively inexpensive to synthesize many copies of a few oligos while it is currently cost prohibitive to synthesize a few copies of many oligos. Second, data stored using longer strands needs fewer total strands to store the same amount of data, lowing the overhead devoted to file addresses, file indexes, and error correction. Third, the fields of molecular biology and synthetic biology have developed many high throughput one-pot DNA assembly methods for synthesizing large genes or operons. These include Golden Gate Assembly and Multiple Overlap

Extension PCR where pieces of DNA have complementary overhangs that ligate to each other without the need for multiple slow reaction cycles (**Fig. 4.1A & 4.1C**)[10–12]. The sequences of the overhangs are designed in a way that dictates the sequence in which codewords are assembled. Here we design and test a system to assemble small sequences of DNA into data encoding strands to be stored in a DNA data storage system.

Experimental Design and Results

Ligation-based assembly

An example of creating long DNA strands from short oligonucleotides using ligation is shown below (**Figure 4.1A**). We start with a set of double-stranded short oligonucleotides that each of these 'codewords' represents a unit of data and have single-stranded, 'sticky' overhangs on each side. Long strands are assembled by annealing the sticky overhangs to their homologous binding partners and locking in the event using a ligase enzyme to construct a full-length strand.

Three sets of 16 unique oligos, each 19, 24, or 29 nt long, were designed to contain a 15, 20, or 25 nt codeword, respectively, and a 4 nt overhang to allow for annealing and ligation to its neighboring block[12]. The single stranded oligos were first treated with polynucleotide kinase (PNK) to add a phosphate group to the 5' end of DNA to allow ligation to proceed. Complimentary strands were then annealed together by heating the mixture to 10C above the pair's melting temperature then reducing to room temperature 1C every 30 seconds. Various ligation reactions were tested to determine optimal conditions (1h at 25C or 37C; 18h at 25C or 37C; 30 cycles of 1min at 37C and 1min at 16C; 30 cycles of 5min at 37C and 5 min at 16C). To determine if the appropriate full-length strands were assembled a PCR amplification using the end sequences of the strands as primers was conducted. We have demonstrated the ability to assemble monomer blocks that are each 15, 20, and 25 nt long into 300, 380, and 460bp DNA strands, respectively, as well as create incomplete mixtures of DNA strands that are multiples of 15, 20, and 25 nt (**Fig. 4.1B**). Incomplete assemblies are hypothesized to be due to inefficiencies during PNK treatment of the oligos; future work is planned to conduct similar assembly reactions using oligos with confirmed 5' phosphate groups for more efficient ligations.

Multiple Overlap Extension PCR-based Assembly

Another example of assembling long DNA strands from short oligonucleotides, this time using a polymerase chain reaction (PCR) method, is shown in **Figure 4.1C**[11]. We start with a set of single-stranded short oligonucleotides that have homologous overlaps with adjacent oligonucleotides. Each ‘fragment’ also has a middle, data encoding region containing one or more codewords while the first and last fragments have primer binding regions for full-strand amplification. A one-pot PCR based protocol first allows the strands to act as primers for each other to create longer strands containing an increasing number of the fragments. A final PCR step with the full-strand primers only exponentially amplifies the fully assembled strands and finalizes the DNA strand assembly.

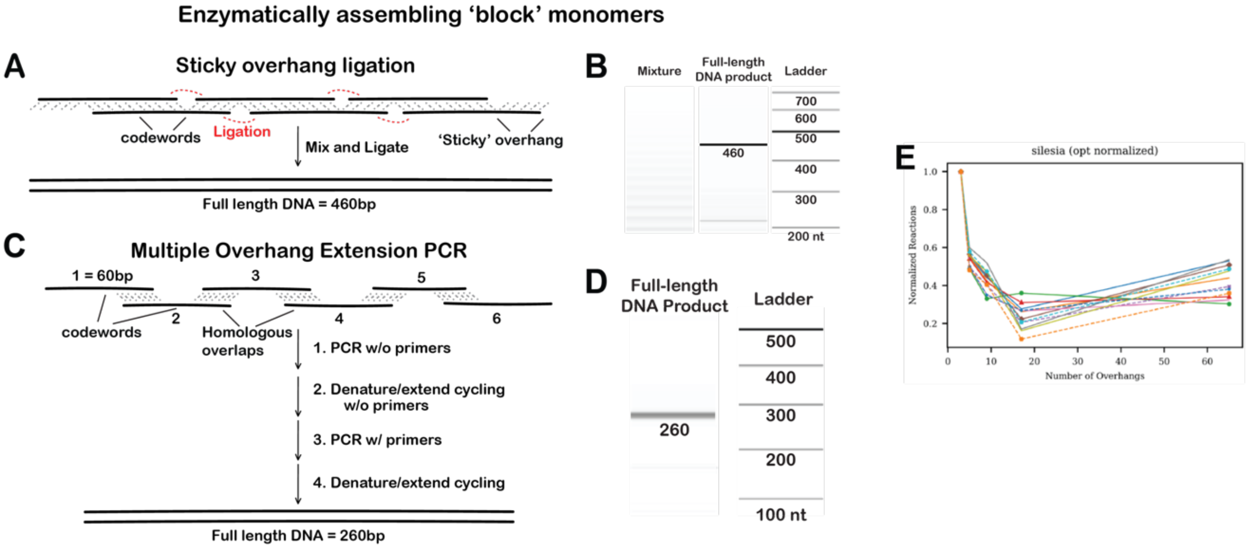


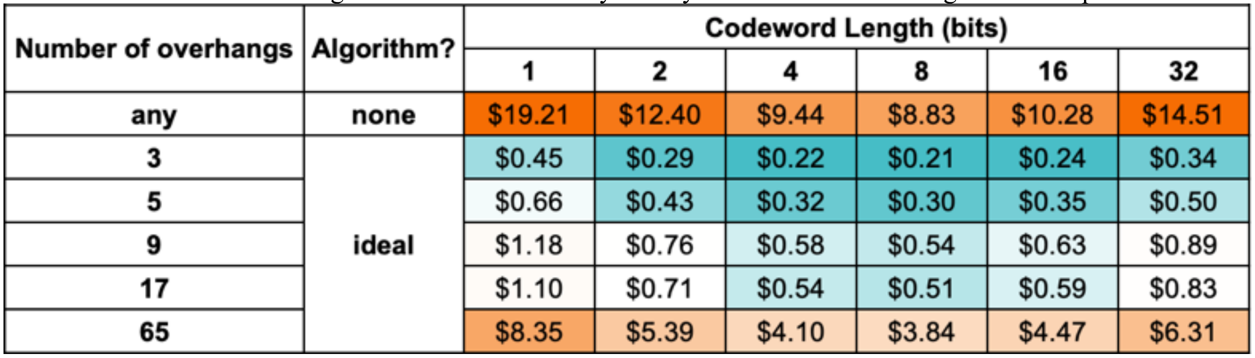
Figure 4.1 Enzymatic DNA assembly methods.
We used one-pot enzyme-based (A) Golden-Gate and (C) Overlap Extension PCR DNA assembly methods that use complimentary overhang sequences on the ends of ‘codeword’ monomers to (B, D) assemble mixtures of DNA strands of different sizes as well as specific strands. (E) An optimal number of distinct overhangs sequences minimizes the number of reactions needed to assemble DNA strands.

Six oligos, each 60 nt long, were designed to have a middle region of 20 nt to represent a codeword while adjacent oligos contain 20 nt homologous overlaps (i.e., complimentary sequences) with each other. Primer binding sequences flank fragment 1 and fragment 6 to allow for PCR amplification of the completed strand. All six fragments were combined in the same reaction with 20ng of the four middle fragments (2-5) per reaction and 10ng of the two end fragments (1, 6) per reaction. A 15 cycle PCR in the absence of primers followed by 20 cycles of denaturation and extension allows the fragments to act as each other’s primers and fill in single stranded gaps to build strands containing 2 or more combined fragments. A second PCR in the presence of end fragment primers only exponentially amplifies the strands with all fragments in the proper order while a final denaturation and extension step fills in any gaps and completes the unfinished strands. We were able to demonstrate the one-pot assembly of a full length 260bp strand of DNA (**Figure 4.1D**) assembled from 6 smaller, 60nt single-stranded DNA fragments.

Cost analysis

Accounting for all costs of reagents including the codeword monomers, enzymes, and buffers, the cost of synthesizing 1 TB of information using this assembly method would average ~$10 (rather than ~$1 billion estimated by phosphoramidite chemistry). The immense cost savings comes from the economies of scale of chemically synthesizing large amounts of each distinct codeword monomer and then using enzymes to assemble them into many distinct long DNA strands, rather than synthesizing every arbitrary DNA strand chemically using traditional synthesis. We have also computationally optimized the assembly method to reduce the number of total reactions needed to build files and tested it on all the files in the Silesia compression corpus. Interestingly, there is an optimal number of overhangs to reduce the number of single-pot reactions needed (**Fig. 4.1E**) to make all the strands of a file; this has major implications for decreasing synthesis costs and increasing synthesis speed. We have also performed simple calculations showing that the putative $10/TB cost can be further reduced simply by optimizing the length of each codeword and the number of distinct overhangs used, which in turn affects the number of distinct codeword monomer blocks that need to be synthesized. This simple optimization is able to further reduce the putative cost to 21 cents per TB (**Table 4.1**). We will build upon this experimental and computational work to increase the speed of assembly reactions, reduce the masses and volumes of reagents needed per assembly reaction, and assess error rates and improve the accuracy of assemblies.

Table 4.1 Codeword assembly economics.
Comparing the cost to synthesize 1TB of data based on number of overhangs used in the system, length of each codeword in bits, and with or without optimizing the redundancy of reactions. Optimization of codeword length and the number of distinct overhangs reduces DNA assembly cost by another 2 orders of magnitude on a per TB basis.



Number of overhangs	Algorithm?	Codeword Length (bits)					
		1	2	4	8	16	32
any	none	$19.21	$12.40	$9.44	$8.83	$10.28	$14.51
3	ideal	$0.45	$0.29	$0.22	$0.21	$0.24	$0.34
5		$0.66	$0.43	$0.32	$0.30	$0.35	$0.50
9		$1.18	$0.76	$0.58	$0.54	$0.63	$0.89
17		$1.10	$0.71	$0.54	$0.51	$0.59	$0.83
65		$8.35	$5.39	$4.10	$3.84	$4.47	$6.31

Several assumptions were used to determine the cost of de-novo synthesis and DNA ligation reactions. A laboratory preferred vendor was used to price the necessary reagents. Since costs were determined based on typical laboratory scales, it can be assumed both reactions will benefit from the economies of scale. The calculations used the cheapest currently available scale of oligo synthesis ($1.00/nt at a scale of 1umole)[13]. Assuming 10,000 copies of each codeword needed per ligation reaction, each codeword will have the potential to be reused for $6.02x10^{13}$ total reactions. The total amount of ligase needed per reaction is dependent upon the minimum amount of enzyme needed to ligate 100ng of DNA, which can vary between enzymes[14–16]. The amount of DNA (ng) to be ligated per reaction was determined by equation 4.1.

$$0.00000298\,\frac{ng}{reaction} = \frac{10{,}000\ copy}{codeword} * \frac{10\ codewords}{reaction} * \frac{29\ basepairs\ (bp)}{copy} * \frac{618.5\ ng}{nmole\ of\ bp} * \frac{1x10^{9} nmole\ of\ bp}{1\ mole\ of\ bp} * \frac{1\ mole\ of\ bp}{6.022x10^{23}\ bp} \qquad (4.1)$$

References

1. Takahashi, C. N., Nguyen, B. H., Strauss, K. & Ceze, L. Demonstration of End-to-End Automation of DNA Data Storage. *Sci. Rep.* **9,** 1–5 (2019).

2. TwistBioscience. *Oligo Pools product information. www.twistbioscience.com DSOPUS_031218*. (2018).

3. Caruthers, M. H. & McBride, L. J. An investigation of several deoxynucleoside phosphoramidites useful for synthesizing deoxyoligonucleotides. *Tetrahedron Lett.* (1983). doi:10.1017/CBO9781107415324.004

4. TwistBioscience. A Simple Guide to Phosphoramidite Chemistry and How it Fits in Twist Bioscience's Commercial Engine. https://www.twistbioscience.com/blog/science/simple-guide-phosphoramidite-chemistry-and-how-it-fits-twist-biosciences-commercial. (2018).

5. Sigma Aldrich. DNA Oligonucleotide Synthesis. https://www.sigmaaldrich.com/technical-documents/articles/biology/dna-oligonucleotide-synthesis.html.

6. Roquet, N., Park, H. & Bhatia, S. P. Patent Application Number PCT/US2017/062098: Nucleic Acid-Based Data Storage. 1–100 (2017).

7. DNA Data Storage Alliance. PRESERVING OUR DIGITAL LEGACY : AN INTRODUCTION TO DNA DATA STORAGE. (2021).

8. Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.* **10,** 1–12 (2019).

9. Palluk, S. *et al.* De novo DNA synthesis using polymerasenucleotide conjugates. *Nat. Biotechnol.* **36,** 645–650 (2018).

10. Halleran, A. D., Swaminathan, A. & Murray, R. M. Single Day Construction of Multigene Circuits with 3G Assembly. *ACS Synth. Biol.* **7,** 1477–1480 (2018).

11. An, Y. *et al.* A rapid and efficient method for multiple-site mutagenesis with a modified overlap extension PCR. *Appl. Microbiol. Biotechnol.* **68,** 774–778 (2005).

12. Potapov, V. *et al.* Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. *ACS Synth. Biol.* **7,** 2665–2674 (2018).

13. Eurofins. Oligo Price List. Available at: https://eurofinsgenomics.com/en/products/dnarna-synthesis/oligo-price-list/.

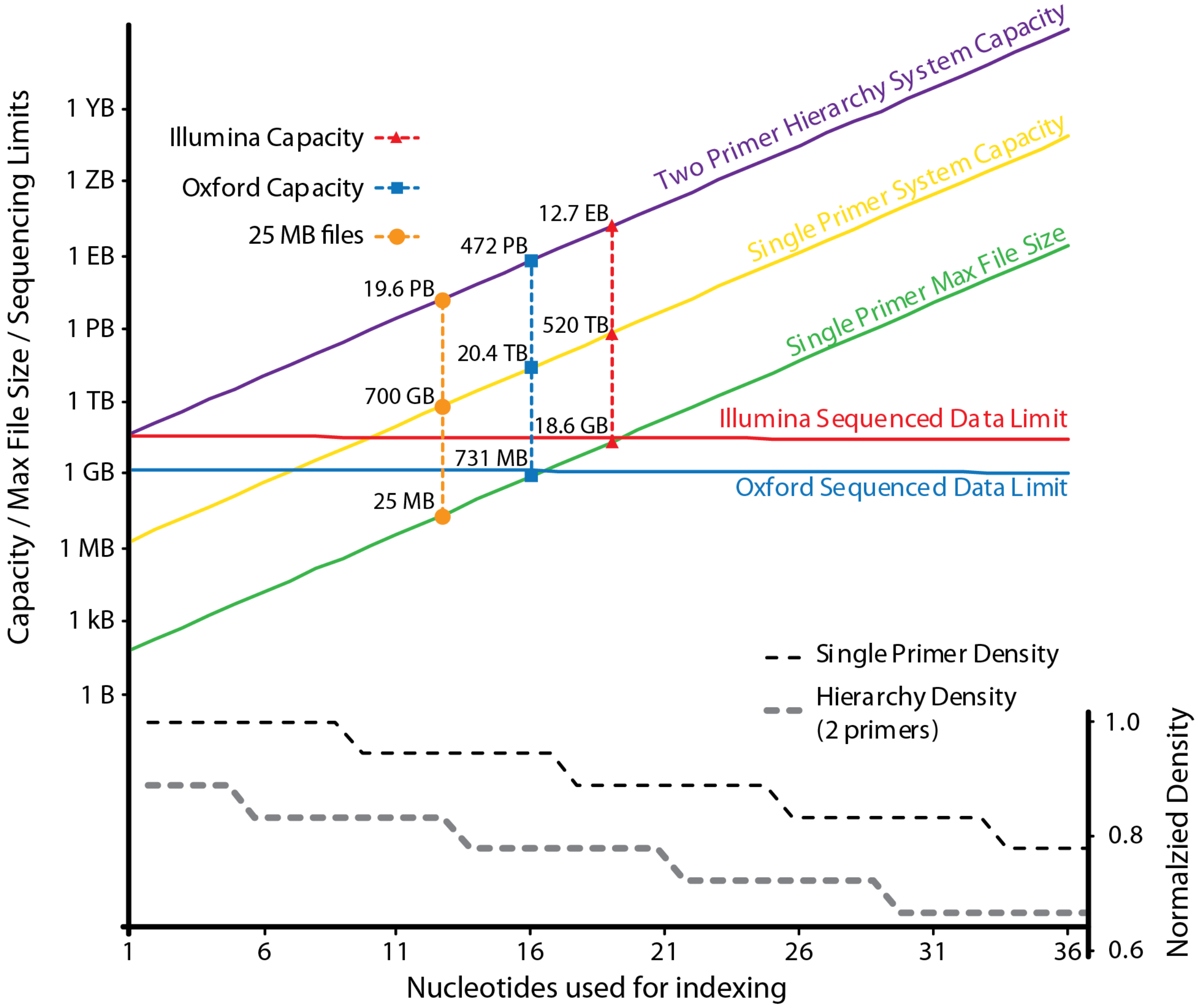
14. Enzymatics. T4 DNA Ligase. Available at: http://www.enzymatics.com/wp-content/uploads/2014/12/L6030-HC-L-REV-F-T4-DNA-Ligase-PSF-EFF04SEP2014.pdf.

15. NEB. Protocol with T4 DNA Ligase. Available at: https://www.neb.com/protocols/0001/01/01/dna-ligation-with-t4-dna-ligase-m0202.

16. NEB. Protocol for Taq DNA Ligase. Available at: https://www.neb.com/protocols/2014/04/04/protocol-for-taq-dna-ligase-m0208.

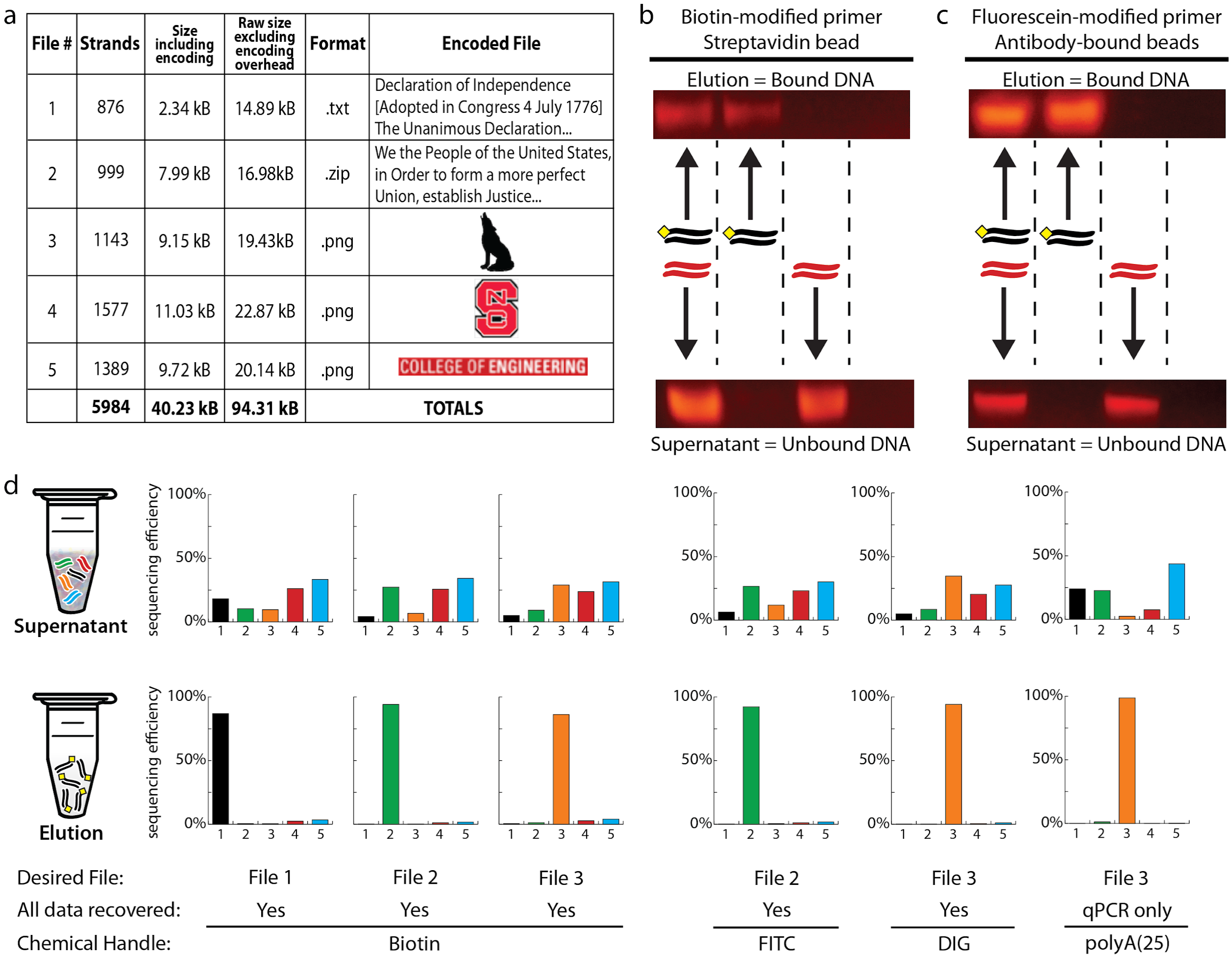
APPENDICIES

APPENDIX A: Supporting Information for Chapter 1



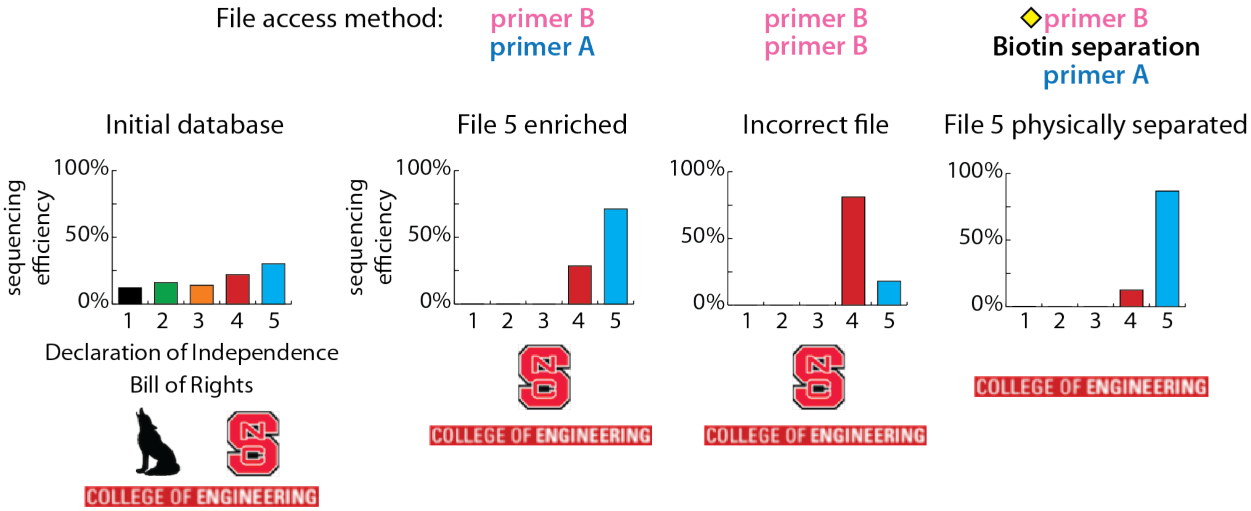
Supplementary Figure A.1 Implementation of high-capacity DNA-based data storage presents physical and architectural challenges.

This plot presents a quantitative analysis of tradeoffs in selecting system parameters. The maximum file size as well as total storage capacities for both hierarchy (nested addresses) (purple line) and single primer (yellow line) systems increase with increasing number of nucleotides (nt) allocated to the index region of a strand (x-axis). Also plotted are the densities of single primer (black dash) and hierarchical (grey dash) strand architectures (normalized to the density of a single primer with 1 nt for indexing), the amount of data that can be sequenced with two different sequencing methods (Illumina – solid red line and Oxford Nanopore – solid blue line), and the maximum file size that can be attained as a function of index length (solid green line). The maximum amount of data that can be sequenced assumes a 10x sequencing depth. The maximum file size is plotted for only the single primer configuration. For clarity, the amount of data that can be sequenced and the maximum file sizes for hierarchy systems are not plotted but can be calculated using the provided densities and would be only minimally offset from the single primer system.



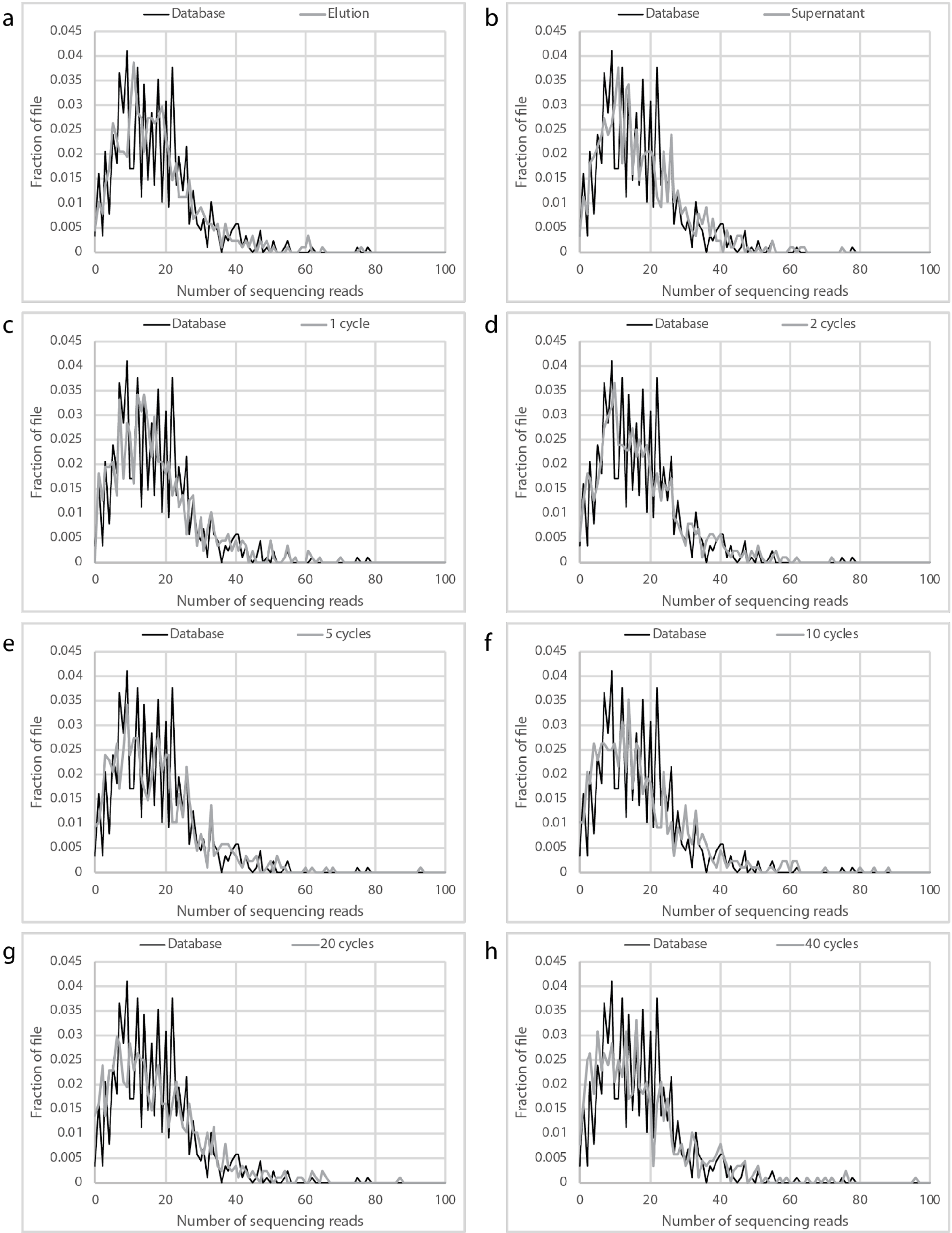
File #	Strands	Size including encoding	Raw size excluding encoding overhead	Format	Encoded File
1	876	2.34 kB	14.89 kB	.txt	Declaration of Independence [Adopted in Congress 4 July 1776] The Unanimous Declaration...
2	999	7.99 kB	16.98kB	.zip	We the People of the United States, in Order to form a more perfect Union, establish Justice...
3	1143	9.15 kB	19.43kB	.png	
4	1577	11.03 kB	22.87 kB	.png	
5	1389	9.72 kB	20.14 kB	.png	COLLEGE OF ENGINEERING
	5984	40.23 kB	94.31 kB	TOTALS	

Supplementary Figure A.2 Library description, preliminary data, and complete analysis of file separations. (a) Description of the experimental database. Five files totaling 40.23 KB (5,984 unique strands), were encoded, synthesized, pooled and stored as one database. (b,c) Proof of concept DNA strand extraction. Biotin (b) or fluorescein (c) primers were used to amplify a single 200 bp-long strand of DNA. Extraction reactions were performed with either: a mixture of modified (black with yellow diamond) and unmodified (red) DNA (lane 1), modified DNA (lane 2), unmodified DNA (lane 3) or water (lane 4). The resulting elutions and supernatants were visualized on an agarose gel. (d) After biotin-streptavidin, fluorescein-antibody, digoxigenin-antibody, and polyA(25)-Oligo-dT file extractions, the supernatants still contain all files (top) while the eluents were enriched for the target files (bottom), as measured by next generation sequencing. By mapping sequencing reads to the original file sequences, all targeted data were confirmed recovered.



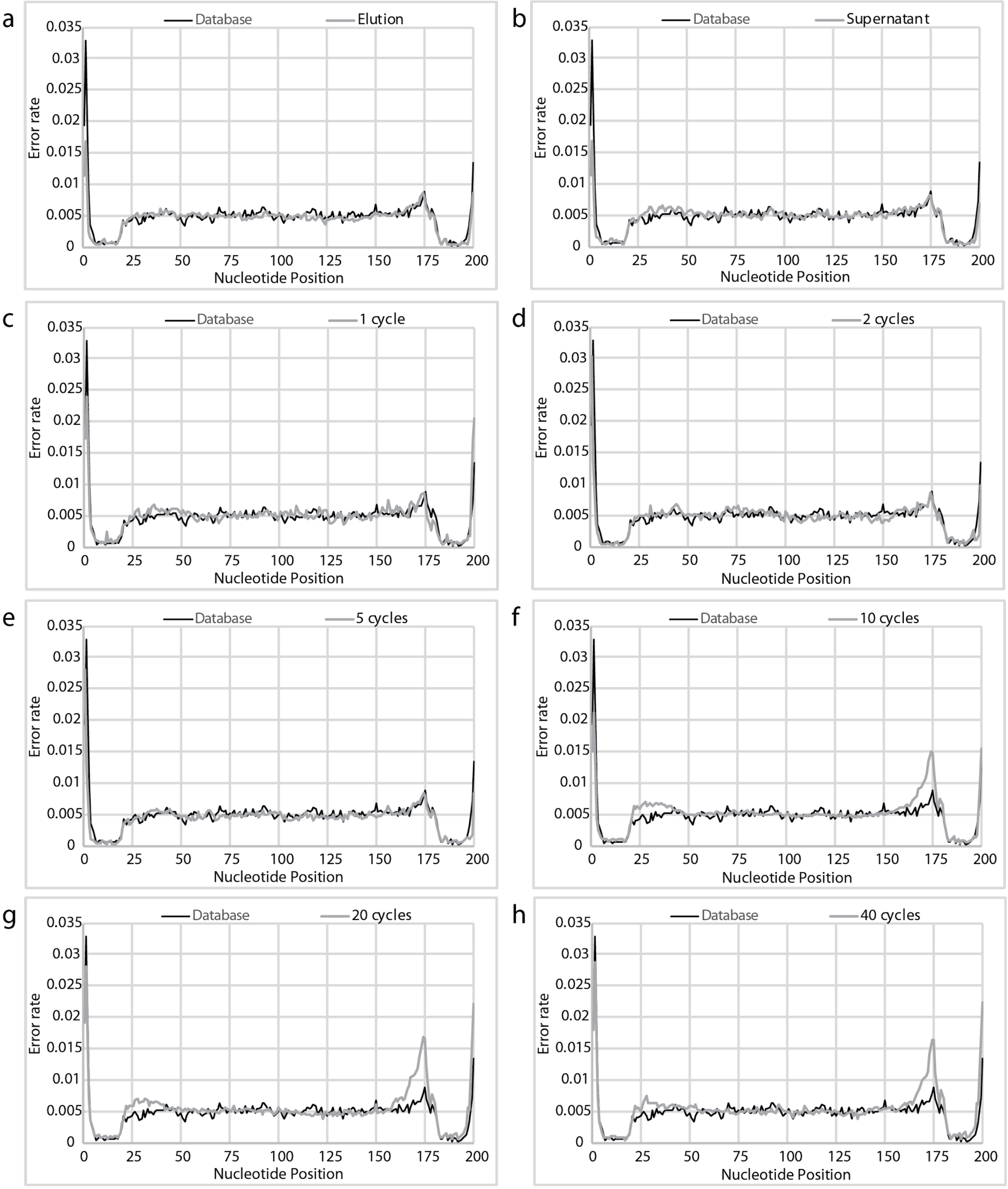
Supplementary Figure A.3 Combining a nested, hierarchical address strategy with physical separations resulted in purified enrichment of the desired file.

Experimental demonstration that PCR using primer B followed by primer A enriched for File 5. PCR amplifications using two rounds of primer B enriched for the incorrect file. In conjunction with physical extractions, File 5 was specifically accessed using hierarchical PCRs. The extraction after the first PCR amplification increased File 5 enrichment from 71% to 86% over no extraction, as measured by qPCR.



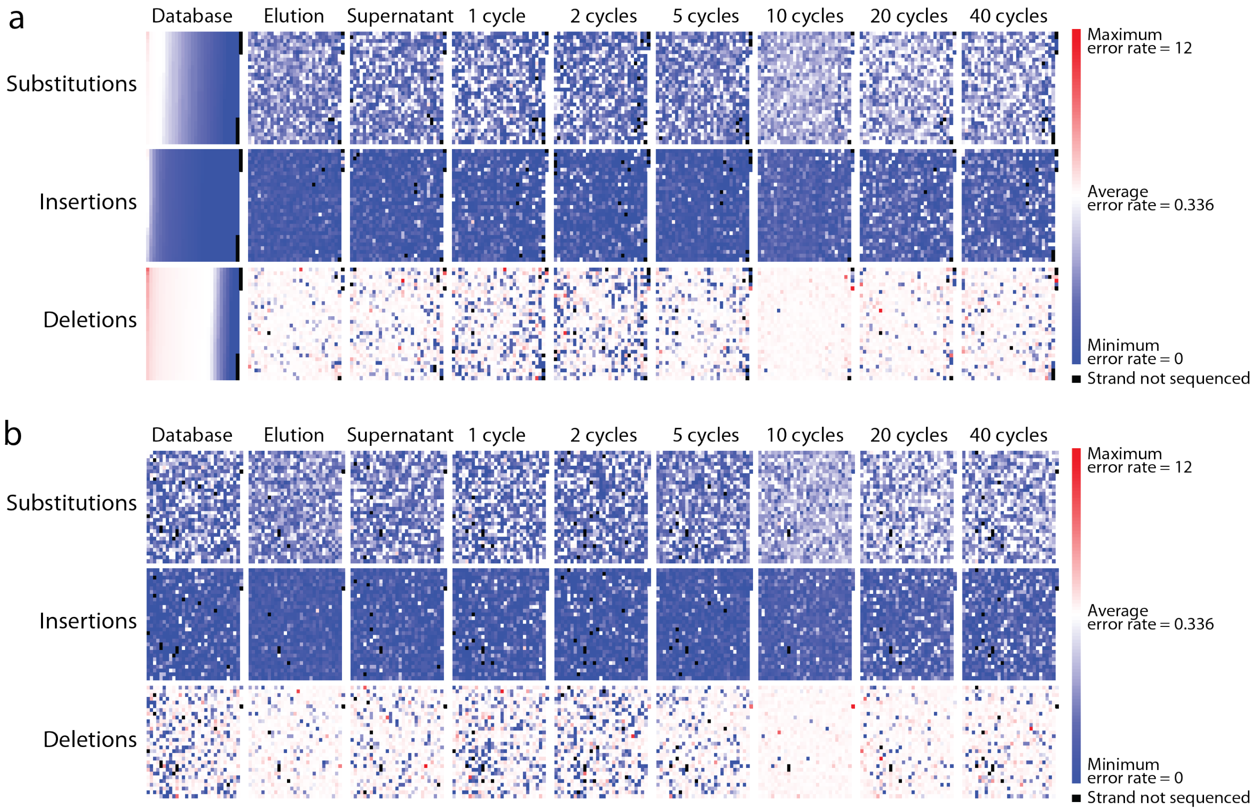
Supplementary Figure A.4 The strand distribution (frequencies of sequencing depths per each unique strand) was not noticeably affected by 40 PCR cycles nor by DENSE.

(a) Elution and (b) supernatant samples from a biotin separation of File 1. Samples after (c) 1, (d) 2, (e) 5, (f) 10, (g) 20, and (h) 40 PCR cycles amplifying File 1. For equal comparison, all sequencing data were randomly downsampled to included only 10,000 File 1 reads.



Supplementary Figure A.5 The error rate at a given nt position (1-200) is plotted as an average across all File 1 strands.

File 1 biotin separation (a) elution and (b) supernatant. (c) 1, (d) 2, (e) 5, (f) 10, (g) 20, and (h) 40 PCR cycles amplifying File 1. For equal comparison, all sequencing data were randomly downsampled to included only a random sample of 10,000 File 1 reads. Enriching File 1 using 10 or more cycles of PCR increased the error rate between nts 28-32 and 167-176. Error rates for File 1 after DENSE, both in the elution and supernatant, remain largely unchanged.



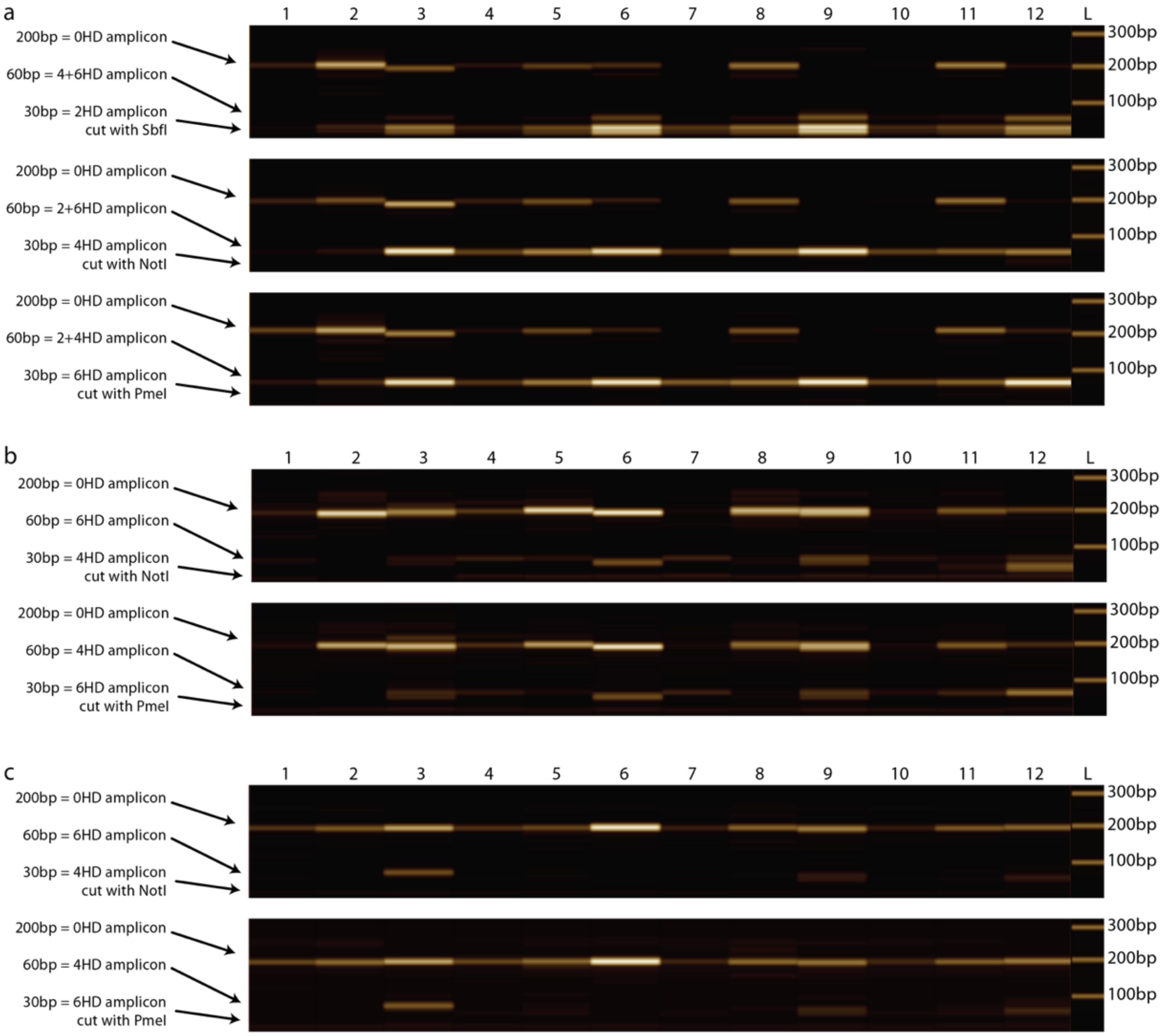
Supplementary Figure A.6 Error rate heatmaps of all File 1 strands after file access by different methods. (left to right) Original database, biotin separation elution, biotin separation supernatant, 1 PCR cycle (without separation), 2 PCR cycles, 5 PCR cycles, 10 PCR cycles, 20 PCR cycles, and 40 PCR cycles. (a) In each row, strands are placed in order from highest to lowest rate of substitutions, insertions, or deletions, and the same order was maintained within each row. (b) To determine if particular strands have correlations in similar error types, all heat maps across all rows display the File 1 strands in the same order: Each unique strand is sorted by its index.

Supplementary Table A.1 Comparison of next generation sequencing and qPCR measurements of file ratios within samples indicate strong agreement between measurement methods.

All initial sample measurements were made using qPCR. Once NGS was conducted, results were compared to validate the accuracy of qPCR file ratio quantifications. Unknown sequences are those that did not fit the mapping criteria discussed in the Methods section.

	File 1		File 2		File 3		File 4		File 5		unknown sequence
Sample	**NGS**	**qPCR**	**NGS**	**qPCR**	**NGS**	**qPCR**	**NGS**	**qPCR**	**NGS**	**qPCR**	**NGS**
Database	12.23%	11.00%	16.12%	19.00%	13.96%	17.00%	26.59%	33.00%	24.99%	26.00%	6.10%
File 1 biotin elution (round 1)	86.87%	94.00%	0.48%	1.00%	0.44%	1.00%	2.46%	3.00%	3.24%	2.00%	6.40%
File 1 biotin supernatant (round 1)	18.10%	32.00%	10.30%	16.00%	9.55%	19.00%	26.99%	18.00%	32.34%	15.00%	2.70%
File 2 biotin elution	0.25%	1.00%	94.17%	95.00%	0.12%	0.00%	1.20%	2.00%	1.60%	2.00%	2.60%
File 2 biotin supernatant	4.14%	6.00%	27.27%	29.00%	6.67%	11.00%	26.59%	24.00%	33.03%	29.00%	2.30%
File 3 biotin elution	0.60%	1.00%	1.20%	2.00%	86.15%	92.00%	2.87%	3.00%	3.81%	3.00%	5.30%
File 3 biotin supernatant	4.99%	5.00%	9.18%	13.00%	28.91%	41.00%	24.61%	22.00%	30.53%	20.00%	1.80%
File 2 fluorescein eluiton	0.25%	0.00%	92.29%	85.00%	0.49%	0.00%	1.24%	8.00%	1.64%	7.00%	4.00%
File 2 fluorescein supernatant	6.38%	6.00%	26.54%	10.00%	11.83%	8.00%	24.64%	34.00%	28.11%	43.00%	2.50%
File 3 digoxigenin elution	0.16%	0.00%	0.18%	0.00%	94.20%	88.00%	0.64%	6.00%	1.03%	6.00%	3.70%
File 3 digoxigenin supernatant	4.92%	5.00%	8.48%	5.00%	34.71%	9.00%	21.95%	38.00%	25.58%	43.00%	4.30%
Repeated File 1 biotin elution (round 2)	97.90%	100.00%	0.06%	0.00%	0.01%	0.00%	0.06%	0.00%	0.07%	0.00%	1.80%
Repeated File 1 biotin supernatant (round 2)	48.30%	56.00%	1.70%	2.00%	1.70%	4.00%	20.20%	18.00%	26.30%	20.00%	1.80%
Repeated File 1 biotin elution (round 3)	85.80%	96.00%	0.50%	1.00%	0.40%	1.00%	1.50%	1.00%	1.20%	1.00%	10.40%
Repeated File 1 biotin supernatant (round 3)	59.80%	64.00%	0.40%	1.00%	0.40%	1.00%	16.30%	19.00%	19.10%	15.00%	4.10%
Repeated File 2 biotin elution (round 2)	0.90%	3.00%	85.70%	97.00%	0.03%	0.00%	0.36%	0.00%	0.32%	0.00%	12.60%
Repeated File 2 biotin supernatant (round 2)	8.70%	12.00%	25.50%	31.00%	2.20%	4.00%	27.10%	27.00%	33.60%	26.00%	2.80%
File 1 - 1 PCR cycle	35.30%	18.00%	5.60%	8.00%	6.10%	7.00%	24.40%	59.00%	20.60%	8.00%	7.90%
File 1 - 2 PCR cycles	50.20%	39.00%	4.10%	8.00%	4.80%	11.00%	18.60%	36.00%	16.30%	6.00%	6.00%
File 1 - 5 PCR cycles	85.00%	59.00%	1.10%	3.00%	1.00%	3.00%	4.10%	30.00%	3.80%	6.00%	5.00%
File 1 - 10 PCR cycles	95.50%	29.00%	0.50%	6.00%	0.80%	7.00%	0.70%	22.00%	0.90%	36.00%	1.70%
File 1 - 20 PCR cycles	96.10%	67.00%	0.30%	2.00%	0.40%	2.00%	0.40%	15.00%	0.40%	13.00%	2.40%
File 1 - 40 PCR cycles	94.70%	95.00%	0.40%	0.00%	0.30%	0.00%	0.40%	0.00%	0.40%	4.00%	3.90%

APPENDIX B: Supporting Information for Chapter 2

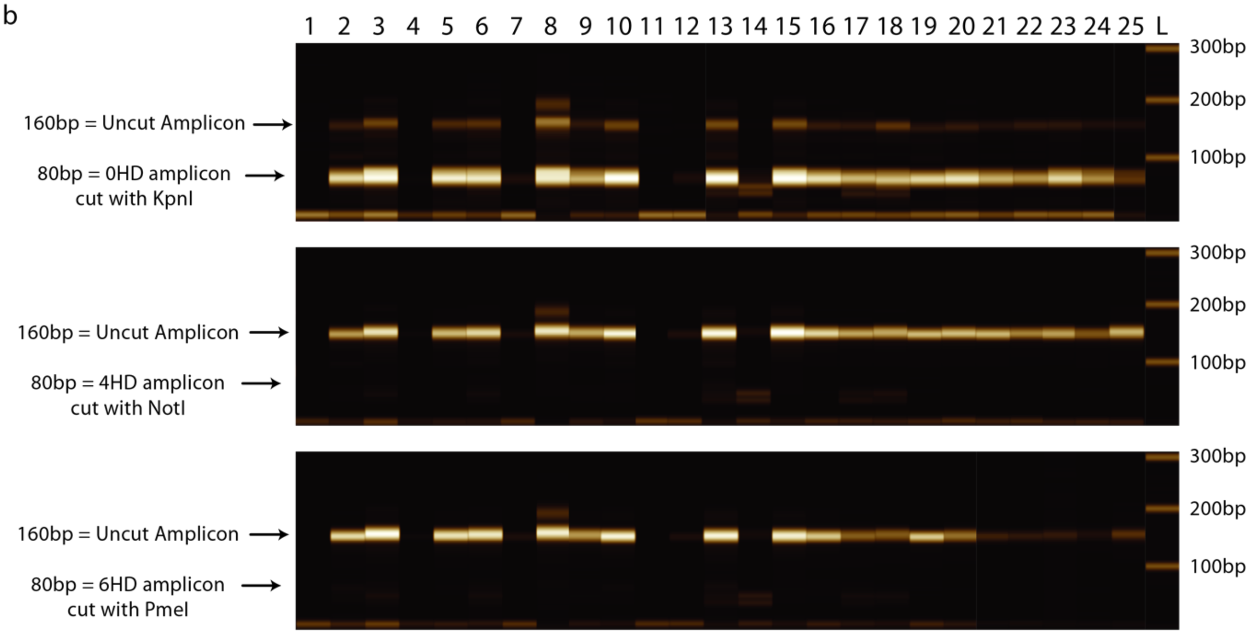


Supplementary Figure B.1 Capillary DNA gel electrophoresis of different Hamming distance PCR products derived from a screen of temperature and primer conditions.

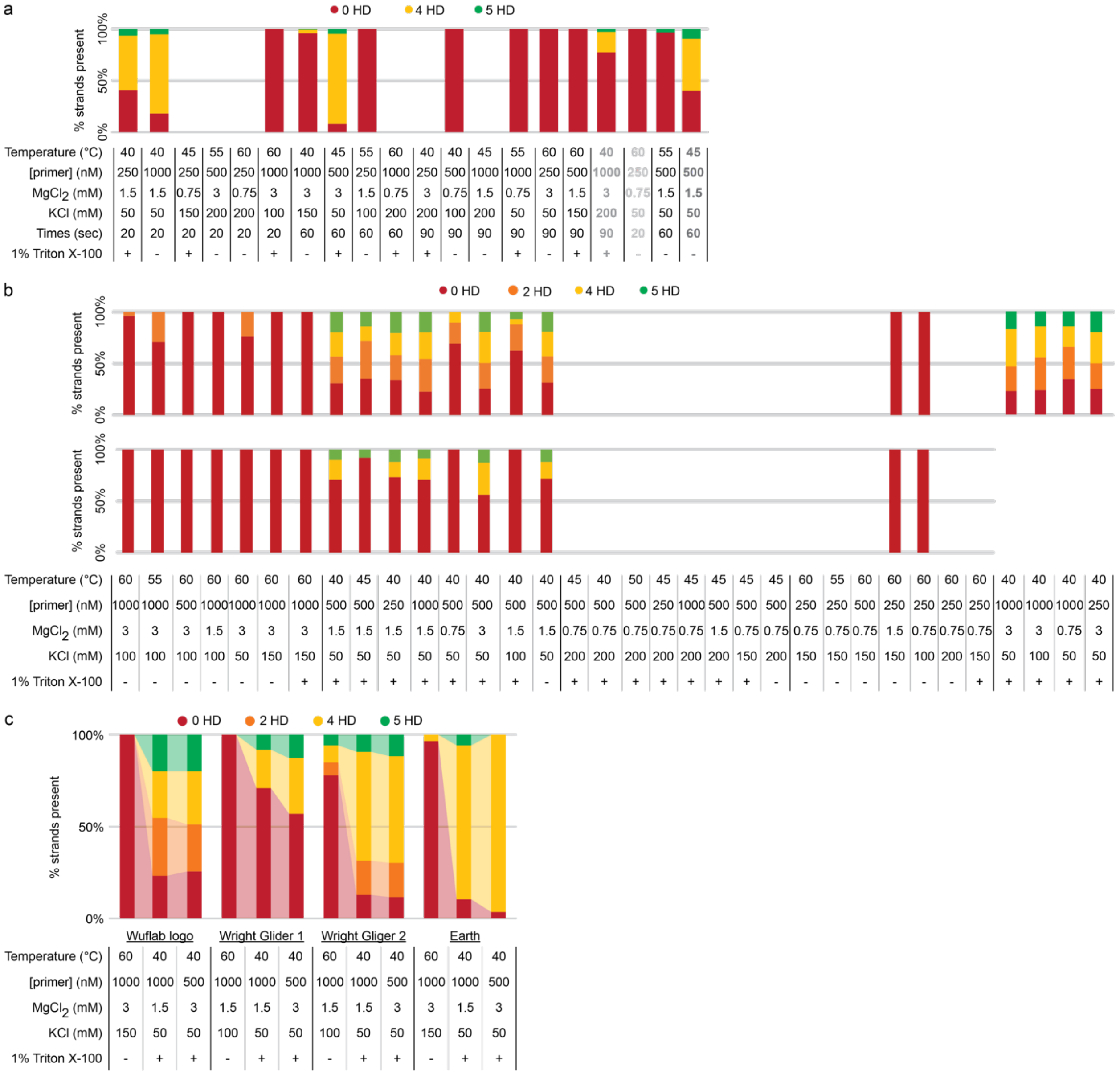
(a) 0, 2, 4, and 6 HD or (b, c) 0, 4, and 6 HD strands were combined and amplified at the various conditions described below. Full length PCR products were 200 bp for 0HD strands and 60 bp for 2, 4, and 6 HD strands. 0 HD amplicons were left uncut while 2 HD strands were cut only by SbfI, 4 HD amplicons were cut only by NotI, and 6HD amplicons were cut only by PmeI. Gel lane numbers correspond to the following qPCR primer and annealing temperature descriptions: 1 – 125 nM primer, 60°C; 2 – 250 nM primer, 60°C; 3 – 500 nM primer, 60°C; 4 – 125 nM primer, 55°C; 5 – 250 nM primer, 55°C; 6 – 500 nM primer, 55°C; 7 – 125 nM primer, 50°C; 8 – 250 nM primer, 50°C; 9 – 500 nM primer, 50°C; 10 – 125 nM primer, 40°C; 11 – 250 nM primer, 40°C; 12 – 500 nM primer, 40°C. L equals ladder. Each experiment was run a single time. Source data are provided as a Source Data file.

a

Rxn #	Annealing Temp	[Primer]	Anneal/Ext Time	Cycles	Taq Polymerase	[dNTPs]	MgCl2	KCl (in buffer)	BSA	Triton X-100	Tween 20	DMSO	Betaine
1	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
2	40C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
3	45C	1000nM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
4	45C	250nM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
5	45C	500mM	20s/20s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
6	45C	500mM	60s/60s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
7	45C	500mM	30/30s	25 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
8	45C	500mM	30/30s	40 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	-
9	45C	500mM	30/30s	30 cycles	0.625U	0.2 mM	1.5mM	1x	-	-	-	-	-
10	45C	500mM	30/30s	30 cycles	2.5U	0.2 mM	1.5mM	1x	-	-	-	-	-
11	45C	500mM	30/30s	30 cycles	1.25U	0.8 mM	1.5mM	1x	-	-	-	-	-
12	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	0.75mM	1x	-	-	-	-	-
13	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	3mM	1x	-	-	-	-	-
14	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	2x buffer	-	-	-	-	-
15	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	0.10%	-	-	-	-
16	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	1%	-	-	-	-
17	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	0.10%	-	-	-
18	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	1%	-	-	-
19	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	0.10%	-	-
20	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	1%	-	-
21	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	2%	-
22	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	8%	-
23	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	0.1 mM
24	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	-	3.5mM
25	45C	500mM	30/30s	30 cycles	1.25U	0.2 mM	1.5mM	1x	-	-	-	2%	0.1mM



Supplementary Figure B.2 Description of a preliminary PCR condition screen for full file access.
(a) This table shows the reaction number to reference in panel b, annealing temperatures, primer concentrations, annealing and extension times, cycles counts, polymerase concentration, dNTP concentration, $MgCl_2$ concentration, KCl concentration, % BSA, % Triton X-100, % Tween20, % DMSO, and % Betaine. (b) Capillary DNA gel electrophoresis. Wright Glider 2 (0, 4, and 6 HD) strands were combined and amplified at the various reaction conditions (#1-25) described in panel a. Full length PCR products were 160 bp for all strands. 0 HD amplicons were cut only by KpnI, 4 HD amplicons were cut only by NotI, and 6HD amplicons were cut only by PmeI. Each experiment was run a single time. Source data are provided as a Source Data file.



a

Temperature (°C)	40	40	45	55	60	60	40	45	55	60	40	40	45	55	60	60	40	60	55	45
[primer] (nM)	250	1000	250	500	250	1000	1000	500	250	1000	250	500	1000	1000	250	500	1000	250	500	500
$MgCl_2$ (mM)	1.5	1.5	0.75	3	0.75	3	3	3	1.5	0.75	3	0.75	1.5	0.75	3	1.5	3	0.75	1.5	1.5
KCl (mM)	50	50	150	200	200	100	150	50	100	200	200	100	200	50	50	150	200	50	50	50
Times (sec)	20	20	20	20	20	20	60	60	60	60	90	90	90	90	90	90	90	20	60	60
1% Triton X-100	+	-	+	-	-	+	-	+	-	+	+	-	-	+	-	+	+	-	-	-

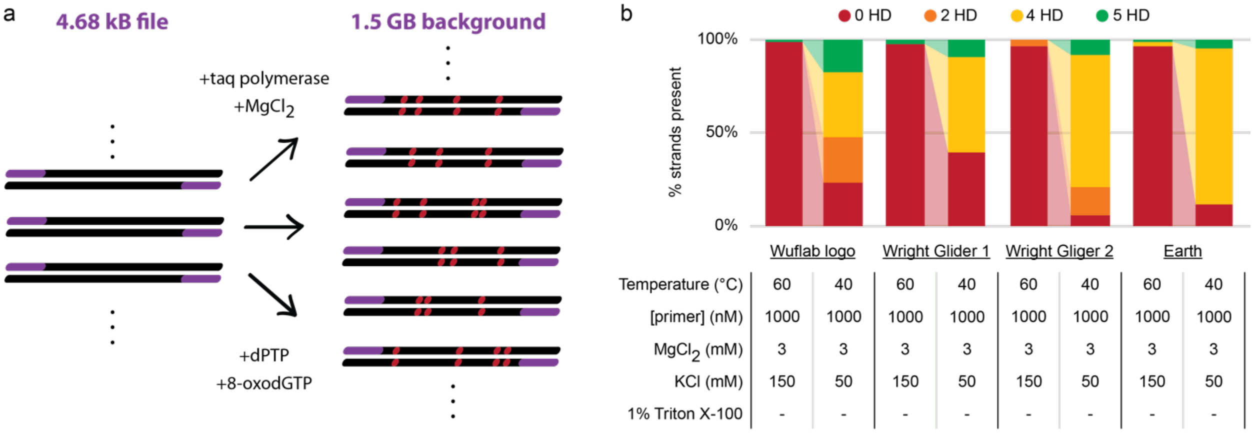
b

Temperature (°C)	60	55	60	60	60	60	60	40	45	40	40	40	40	40	40	45	40	50	45	45	45	45	45	60	55	60	60	60	60	60	40	40	40	40
[primer] (nM)	1000	1000	500	1000	1000	1000	1000	500	500	250	1000	500	500	500	500	500	500	500	250	1000	500	500	500	250	250	500	250	250	250	250	1000	1000	1000	250
$MgCl_2$ (mM)	3	3	3	1.5	3	3	3	1.5	1.5	1.5	1.5	0.75	3	1.5	1.5	0.75	0.75	0.75	0.75	0.75	1.5	0.75	0.75	0.75	0.75	0.75	1.5	0.75	0.75	0.75	3	3	0.75	3
KCl (mM)	100	100	100	100	50	150	150	50	50	50	50	50	50	100	50	200	200	200	200	200	200	150	200	150	150	150	150	100	200	150	50	100	50	50
1% Triton X-100	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	-	-	-	-	-	-	-	+	+	+	+	+

c

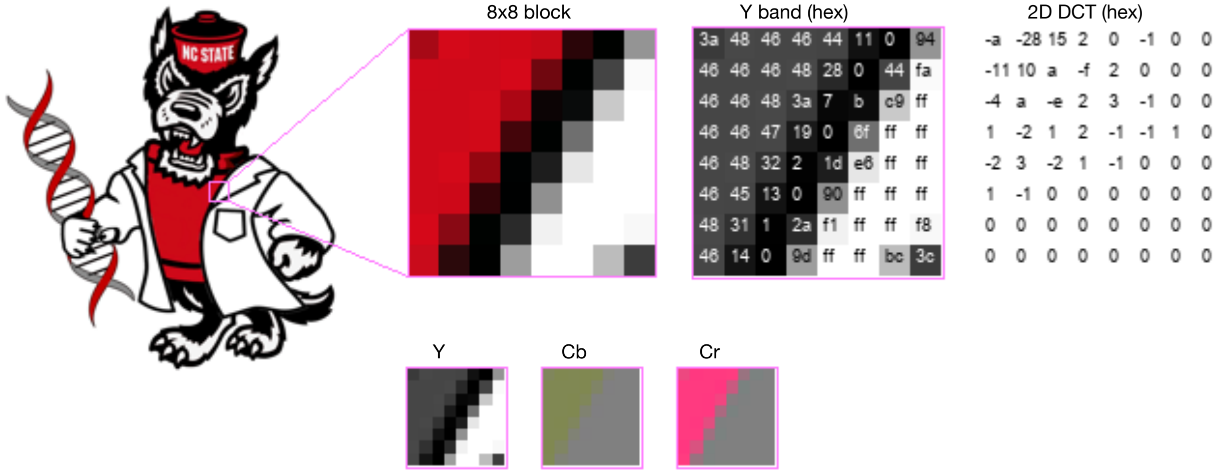
Temperature (°C)	60	40	40	60	40	40	60	40	40	60	40	40
[primer] (nM)	1000	1000	500	1000	1000	500	1000	1000	500	1000	1000	500
$MgCl_2$ (mM)	3	1.5	3	1.5	1.5	3	1.5	1.5	3	3	1.5	3
KCl (mM)	150	50	50	100	50	50	100	50	50	150	50	50
1% Triton X-100	-	+	+	-	+	+	-	+	+	-	+	+

Supplementary Figure B.3 Assessing the effects of specific environmental conditions on File Preview.
(a) PCR condition screen amplifying Wright Glider 1 using relevant parameters determined in Supplementary Figure 2. (b) Fine-tuned (single variable changed per reaction) screen of PCR conditions amplifying Wuflab logo (top row) and Wright Glider 1 (bottom row). The four rightmost conditions of the top row include an attempt to find synergistic variable conditions. (c) Capillary electrophoresis analysis of Wuflab logo, Wright Glider 1, Wright Glider 2, and Earth samples which were accessed, then sent for NGS and decoded. Source data are provided as a Source Data file.



Temperature (°C)	60	40	60	40	60	40	60	40
[primer] (nM)	1000	1000	1000	1000	1000	1000	1000	1000
$MgCl_2$ (mM)	3	3	3	3	3	3	3	3
KCl (mM)	150	50	150	50	150	50	150	50
1% Triton X-100	-	-	-	-	-	-	-	-

Supplementary Figure B.4 File Preview amongst a high background of data.
(a) Schematic of Error Prone PCR used to generate a 1.5 GB background by mutagenizing a file that encodes the Declaration of Independence. (b) Capillary electrophoresis analysis of Wuflab logo, Wright Glider 1, Wright Glider 2, and Earth samples which were accessed in the presence of the error prone background, then sent for NGS and decoded. Source data are provided as a Source Data file.



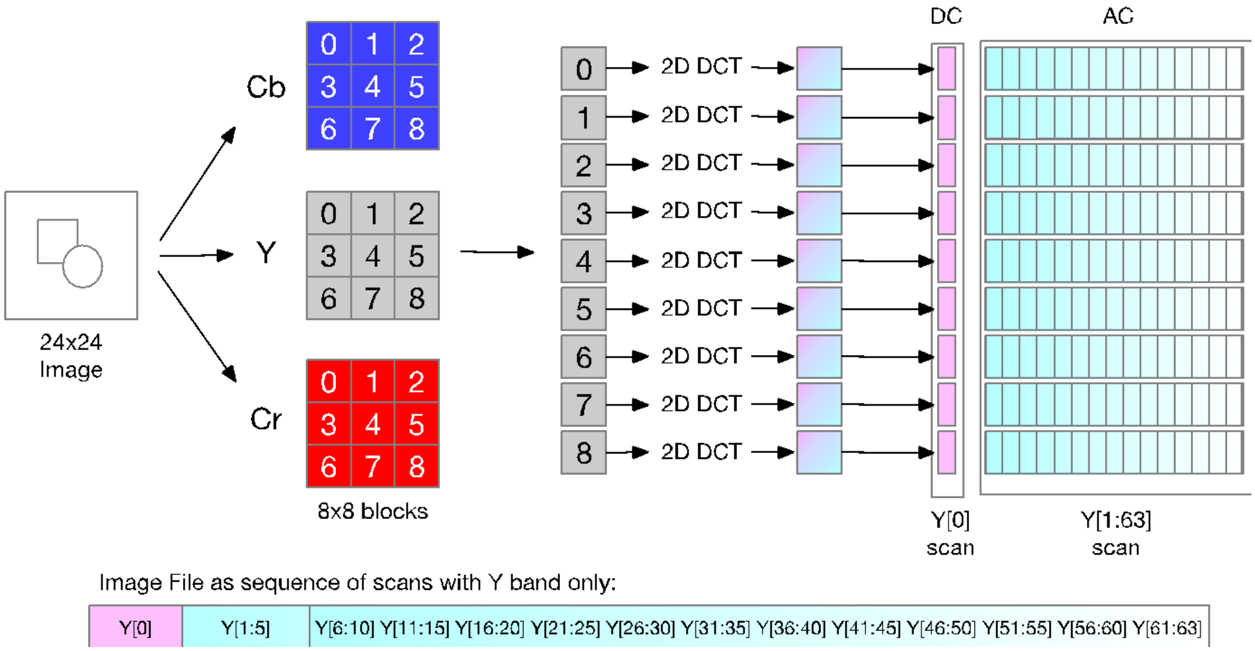
Y band (hex)

3a	48	46	46	44	11	0	94
46	46	46	48	28	0	44	fa
46	46	48	3a	7	b	c9	ff
46	46	47	19	0	6f	ff	ff
46	48	32	2	1d	e6	ff	ff
46	45	13	0	90	ff	ff	ff
48	31	1	2a	f1	ff	ff	f8
46	14	0	9d	ff	ff	bc	3c

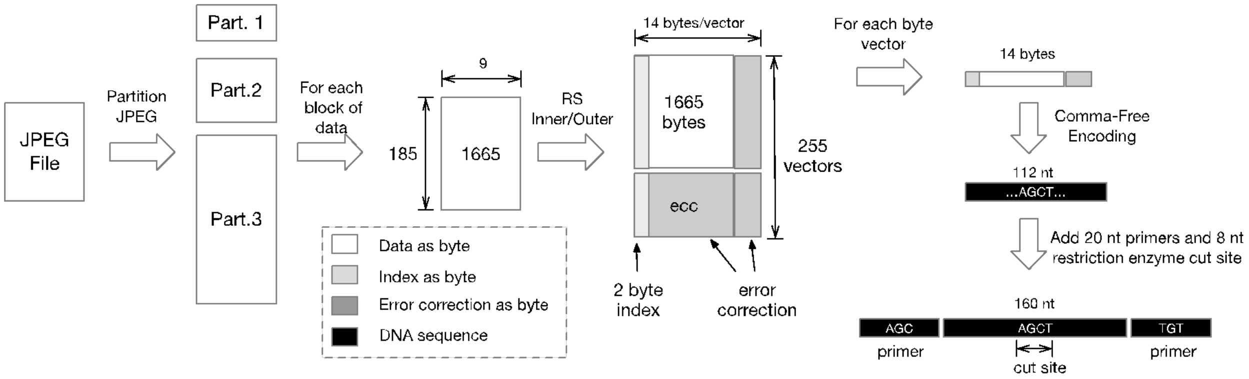
2D DCT (hex)

-a	-28	15	2	0	-1	0	0
-11	10	a	-f	2	0	0	0
-4	a	-e	2	3	-1	0	0
1	-2	1	2	-1	-1	1	0
-2	3	-2	1	-1	0	0	0
1	-1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Supplementary Figure B.5 WufLab logo with an 8x8 block extracted and magnified.
The Y, Cb, and Cr bands are shown individually across the bottom row. The Y band values are shown in hexadecimal in the top middle. The top right shows the 2D DCT transformation in hexadecimal on the 8x8 block. Note the prevalence of 0 values in the bottom right corner enable significant compression.


24x24
Image
Cb
Y
Cr
0 1 2
3 4 5
6 7 8
8x8 blocks
0
1
2
3
4
5
6
7
8
2D DCT
DC
AC
Y[0]
scan
Y[1:63]
scan
Image File as sequence of scans with Y band only:
Y[0]
Y[1:5]
Y[6:10] Y[11:15] Y[16:20] Y[21:25] Y[26:30] Y[31:35] Y[36:40] Y[41:45] Y[46:50] Y[51:55] Y[56:60] Y[61:63]

Supplementary Figure B.6 An illustration of progressive encoding for JPEG images.
A 24x24 is split into the Y, Cb, and Cr bands and divided into nine 8x8 blocks. The Y band is further shown with each 8x8 block converted into the frequency domain using the 2D DCT. The DC component across all blocks are collected into the Y[0] scan and the remaining AC components follow it in subsequent scans.



Supplementary Figure B.7 Encoding process from JPEG file into DNA.

Supplementary Table B.1 Additional screening of 30 variable binding sites.
This table shows primer sequence information, original primer homology of the variable binding site, Hamming distance of the binding site to the original primer, percentage of the 30 strands amplified at constant annealing temperature while varying primer concentration (same calculation as Figure 1b, d, and e), percentage of the 30 strands amplified at constant primer concentration and varying annealing temperature, and percentage of the 30 strands that were amplified at all 4 of these conditions.

Original primer 1 sequence: CAGGTACGCAGTTAGCACTC				Original primer 1' sequence: CGTGGCAATATGACTACGGA			
Original Primer Homology	Hamming Distance	Variable binding sites tested	55°C		250nM		Amplified at all conditions
			125nM	500nM	40°C	60°C	
1 / 1'	2	30	90%	100%	53%	77%	43%
1 / 1'	4	30	7%	77%	10%	10%	0%
1 / 1'	6	30	0%	67%	10%	7%	0%

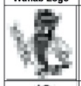
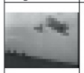
Supplementary Table B.2 Competitive screening of three additional primer sequences.
This table shows original primer sequences, homology, Hamming distance of the binding site to the original primer, ratio of mismatch to perfect match strands at constant annealing temperature while varying primer concentration and each strand's tunability (same calculation as Figure 2b), and ratio of mismatch to perfect match strands at constant primer concentration while varying annealing temperature and each strand's tunability.

Original primer 2 sequence: CAGGAGAATGCCTTCCTAGG	Original primer 2' sequence: CCTCGGTTCTTCTTGACCAG
Original primer 3 sequence: AGGCTGGAGGTCCAATCTTG	Original primer 3' sequence: ATTCTGGCCACTTCCTGAAG
Original primer 4 sequence: AACTAAACGGAGGCCAACAG	Original primer 4' sequence: TTTGTCCAGGAGCCTTTGAG

Original Primer Homology	Hamming Distance	55°C			250nM		
		125nM	500nM	Tunability	40°C	60°C	Tunability
2 / 2'	2	0.03	0.22	0.20	10.30	0.24	10.06
2 / 2'	2	1.79	0.06	-1.73	6.69	0.38	6.31
2 / 2'	2	0.06	0.16	0.09	6.05	0.79	5.26
2 / 2'	2	0.06	0.15	0.09	4.22	0.48	3.74
2 / 2'	2	0.01	0.02	0.02	10.97	0.25	10.72
2 / 2'	3	0.21	0.00	-0.21	0.49	0.00	0.49
2 / 2'	3	0.39	0.00	-0.39	0.46	0.56	-0.10
2 / 2'	3	0.04	0.00	-0.04	3.58	0.09	3.48
2 / 2'	3	0.05	0.00	-0.05	0.00	0.14	-0.14
2 / 2'	3	0.01	0.00	-0.01	0.00	0.00	0.00
3 / 3'	2	0.12	0.83	0.71	4.79	0.04	4.75
3 / 3'	2	0.00	0.10	0.10	1.64	0.00	1.63
3 / 3'	2	0.01	0.40	0.39	6.68	0.47	6.20
3 / 3'	2	10.88	0.30	-10.59	7.08	0.49	6.59
3 / 3'	2	0.07	0.54	0.48	4.45	0.04	4.41
3 / 3'	3	0.00	0.00	0.00	0.01	0.00	0.01
3 / 3'	3	0.00	0.00	0.00	0.14	0.00	0.14
3 / 3'	3	0.00	0.00	0.00	0.02	0.00	0.02
3 / 3'	3	0.00	0.01	0.01	0.38	0.00	0.38
3 / 3'	3	0.00	0.00	0.00	0.51	0.00	0.51
4 / 4'	2	0.06	0.13	0.07	6.64	0.05	6.59
4 / 4'	2	0.00	0.00	0.00	0.07	0.00	0.07
4 / 4'	2	0.00	0.00	0.00	0.00	0.00	0.00
4 / 4'	2	0.00	0.00	0.00	0.14	0.01	0.12
4 / 4'	2	0.19	0.50	0.31	1.85	0.20	1.65
4 / 4'	3	0.00	0.00	0.00	0.22	0.02	0.20
4 / 4'	3	0.00	0.00	0.00	0.00	0.00	0.00
4 / 4'	3	0.00	0.16	0.16	5.39	0.00	5.39
4 / 4'	3	0.00	0.00	0.00	2.80	0.01	2.79
4 / 4'	3	0.00	0.00	0.00	0.12	0.00	0.12

Supplementary Table B.3 Details of the File Preview library.

Four files are encoded and used in the Preview experiments. This table shows a description of the image, the size of the file in bytes, the number of strands per partition, and the fraction contained in each partition (% bytes and % strands), the forward and reverse primer binding sequence, their Hamming distance with respect to the 0 HD primers, the percent of file accessed including that partition (0 HD strands are accessed with 2 HD which are accessed with 4 HD which are accessed with 6 HD), the recognition and cut site for the restriction enzymes used for analysis, and the flanking primer sequences for the initial strand amplification when the library was received.

Wuflab Logo	Bytes	Percent of bytes stored	Strands	Percent of strands stored	Forward primer sequence	Reverse primer sequence	HD from primers	Percent of file accessed	Restriction site	Forward flanking primer	Reverse flanking primer
	707	2.77%	183	4.42%	**CAGGTACGCAGTTAGCACTC**	**CGTGGCAATATGACTACGGA**	0	4.42%	GGTACC	CAGGAGAATGCCTTCCTAGG	CCTCGGTTCTTCTTGACCAG
	1763	6.91%	338	8.16%	CATGTTCGCAGTTAGCACTC	CGTGGCTATATGACTACGCA	2	12.57%	CCTGCAGG	AGGCTGGAGGTCCAATCTTG	ATTCTGGCCACTTCCTGAAG
	3462	13.58%	597	14.41%	CCGATACGTAGTTAGCGCTC	CGGAGAAATATGACGACGGA	4	26.98%	GCGGCCGC	AACTAAACGGAGGCCAACAG	TTTGTCCAGGAGCCTTTGAG
	19565	76.73%	3026	73.02%	TTATTACGCGGTGAGCACTC	CATAGCAATAAGGCTCCGGT	6	100.00%	GTTTAAAC	CGTGGATTCAATTCGGAACG	TTGTTCGCCGAACTGGTTAG
Wright Glider 1	**Bytes**	**Percent of bytes stored**	**Strands**	**Percent of strands stored**	**Forward primer sequence**	**Reverse primer sequence**	**HD from primers**	**Percent of file accessed**	**Restriction site**	**Forward flanking primer**	**Reverse flanking primer**
	726	2.60%	184	4.18%	**CAGGTACGCAGTTAGCACTC**	**CGTGGCAATATGACTACGGA**	0	4.18%	GGTACC	CTGCCAACCTCGGATAACCG	GAACCGAACGGCCACAATAG
	2455	8.78%	415	9.42%	CCGATACGTAGTTAGCGCTC	CGGAGAAATATGACGACGGA	4	13.60%	GCGGCCGC	AGGTCGGACAACGCCTTAAG	TCCACCAACGAACATTTACG
	24785	88.63%	3806	86.40%	TTATTACGCGGTGAGCACTC	CATAGCAATAAGGCTCCGGT	6	100.00%	GTTTAAAC	CGGCACCAACGAAAGAATCG	CAACGAAGGTCCGTCCTTAG
Wright Glider 2	**Bytes**	**Percent of bytes stored**	**Strands**	**Percent of strands stored**	**Forward primer sequence**	**Reverse primer sequence**	**HD from primers**	**Percent of file accessed**	**Restriction site**	**Forward flanking primer**	**Reverse flanking primer**
	1187	3.84%	236	4.84%	**CAGGTACGCAGTTAGCACTC**	**CGTGGCAATATGACTACGGA**	0	4.84%	GGTACC	AGCCTGAACCGTTCTTCCTG	CGCGGAAGGAGGATTAACAG
	4433	14.36%	705	14.47%	CATGTTCGCAGTTAGCACTC	CGTGGCTATATGACTACGCA	2	19.31%	CCTGCAGG	TACCAACATTGCCGCAACTG	CGACCAACAAGGTTCTTACG
	4437	14.37%	705	14.47%	CAGGTAAGTAGCCAGCACTC	CAGGGCAATATGAGAACGGA	4	33.78%	GCGGCCGC	CAACTTGTCCTCCATAAGCG	ACTTAAGCCAGGTTGATTCG
	20822	67.43%	3226	66.22%	CTGGTATGCCCTTAACACCC	CGTGGGGCTATGACTATGTC	6	100.00%	GTTTAAAC	AATGTTCCTGTTGGCGGTTG	ACAATCCTAAGTCCGGTAGG
Earth	**Bytes**	**Percent of bytes stored**	**Strands**	**Percent of strands stored**	**Forward primer sequence**	**Reverse primer sequence**	**HD from primers**	**Percent of file accessed**	**Restriction site**	**Forward flanking primer**	**Reverse flanking primer**
	696	2.56%	181	4.19%	**CAGGTACGCAGTTAGCACTC**	**CGTGGCAATATGACTACGGA**	0	4.19%	GGTACC	ACCAACTAACGGCTTCGTTG	GTAACCATCCGGAGGAAGAG
	1950	7.17%	359	8.31%	CAGGTAAGTAGCCAGCACTC	CAGGGCAATATGAGAACGGA	4	12.50%	GCGGCCGC	GTTCTTGGCTCCAGGTAAGG	ACCGGTCAATTACAACGAAG
	24555	90.27%	3781	87.50%	CTGGTATGCCCTTAACACCC	CGTGGGGCTATGACTATGTC	6	100.00%	GTTTAAAC	ACGGCGAAGGACAATTACGG	GTTAACACCGTGGCAACCAG

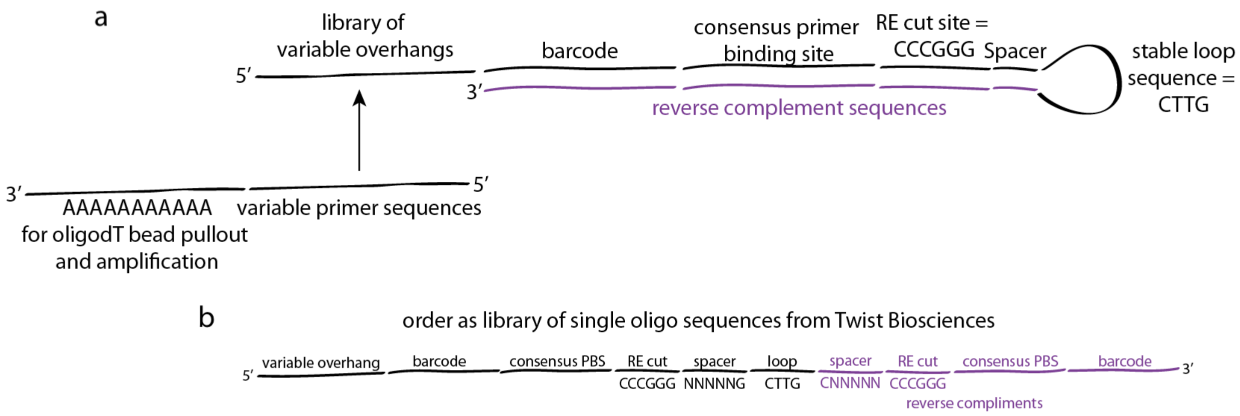
Supplementary Table B.4 Wuflab logo and Wright Glider 2 files are partitioned into four parts. The scans in each partition are described in each row.

Partition	JPEG Progressive Encoding Components per Partition	Preview Description
1	JFIF header, Q0, SOF, Y[0]	**Preview.** JFIF is the JPEG header. Q0 is a quantization matrix. SOF marks beginning of scans. Y[0] provides a grayscale Preview.
2	Y[1:5]	**Intermediate Preview**. Improves resolution of grayscale.
3	Cb[0] ,Cr[0], Y[6:10], Y[11:15]	**Intermediate Preview**. Adds color and improved grayscale resolution.
4	Y[16:20], Y[21:25], Y[26:30], Y[31:35], Y[36:40],Y[41:45], Y[46:50], Y[51:55], Y[56:60], Y[61:63], Cc[1:5], Cr[1:5],CB[6:10], Cr[6:10], Cb[11:15], Cr[11:15], Cb[16:20], Cr[16:20], Cb[21:25], Cr[21:25], Cb[26:30], Cr[26:30], Cb[31:35], Cr[31:35], Cb[36:40], Cr[36:40] ,Cb[41:45], Cr[41:45], Cb[46:50], Cr[46:50], Cb[51:55], Cr[51:55], Cb[56:60], Cr[56:60], Cb[61:63], Cr[61:63], EOI	**Full Image.** Adds remaining components for full image reconstruction. EOI marks the end of the image.

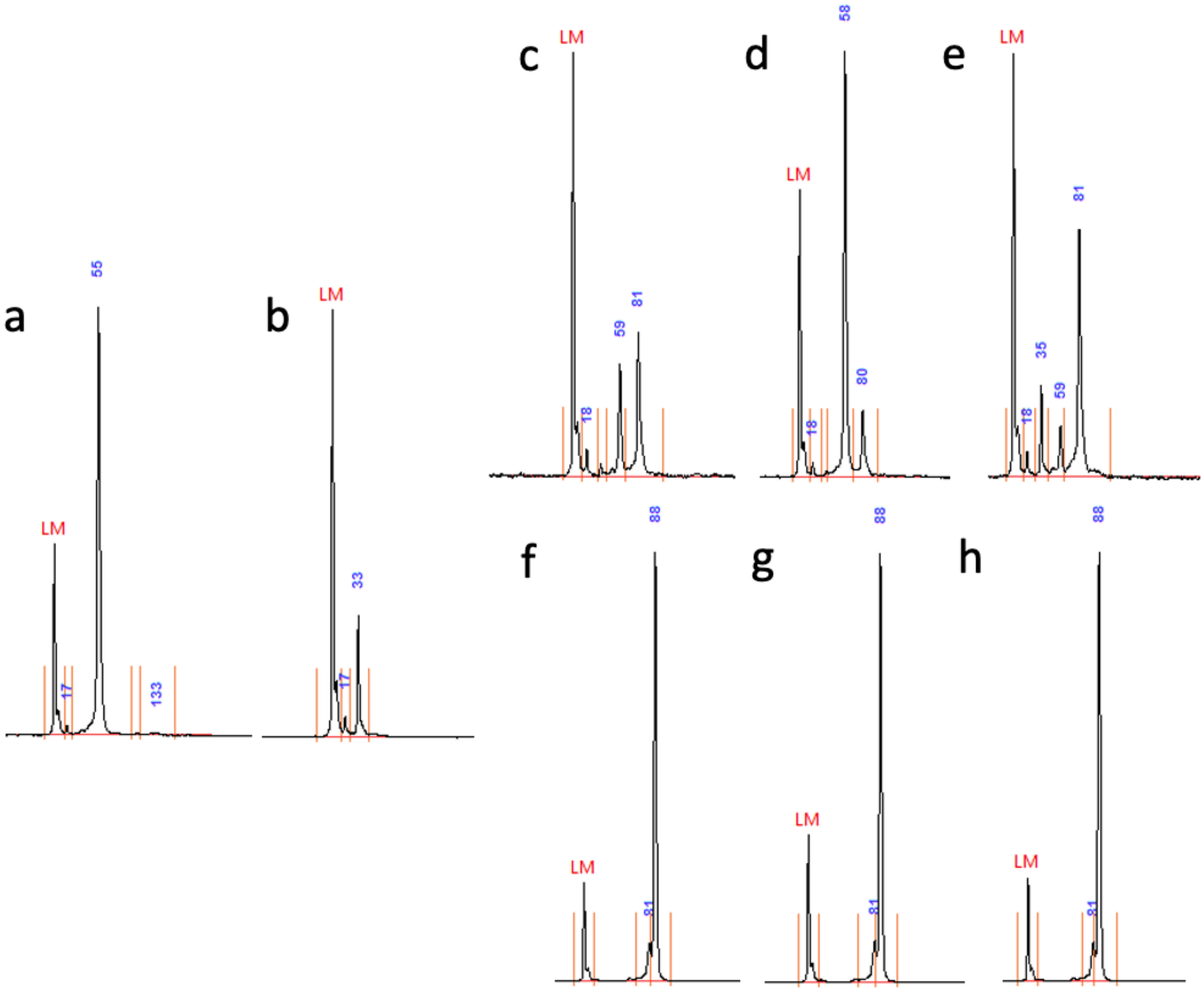
Supplementary Table B.5 Wright Glider 1 and Earth files are partitioned into three parts. The scans in each partition are described in each row.

Partition	Progressive Encoding Components per Partition	Preview Description
1	JFIF header, Q0, SOF, Y[0]	**Preview.** JFIF is the JPEG header. Q0 is a quantization matrix. SOF marks beginning of scans.
2	Y[1:5]	**Intermediate Preview**. Improves resolution of grayscale.
3	Cb[0], Cr[0], Y[6:10], Y[11:15], Y[16:20], Y[21:25], Y[26:30], Y[31:35], Y[36:40],Y[41:45], Y[46:50], Y[51:55], Y[56:60], Y[61:63], Cb[1:5], Cr[1:5],Cb[6:10], Cr[6:10], Cb[11:15], Cr[11:15], Cb[16:20], Cr[16:20], Cb[21:25], Cr[21:25], Cb[26:30], Cr[26:30], Cb[31:35], Cr[31:35], Cb[36:40], Cr[36:40] ,Cb[41:45], Cr[41:45], Cb[46:50], Cr[46:50], Cb[51:55], Cr[51:55], Cb[56:60], Cr[56:60], Cb[61:63], Cr[61:63], EOI	**Full Image.** Adds remaining components for full image reconstruction. EOI marks end of the image.

APPENDIX C: Supporting Information for Chapter 3

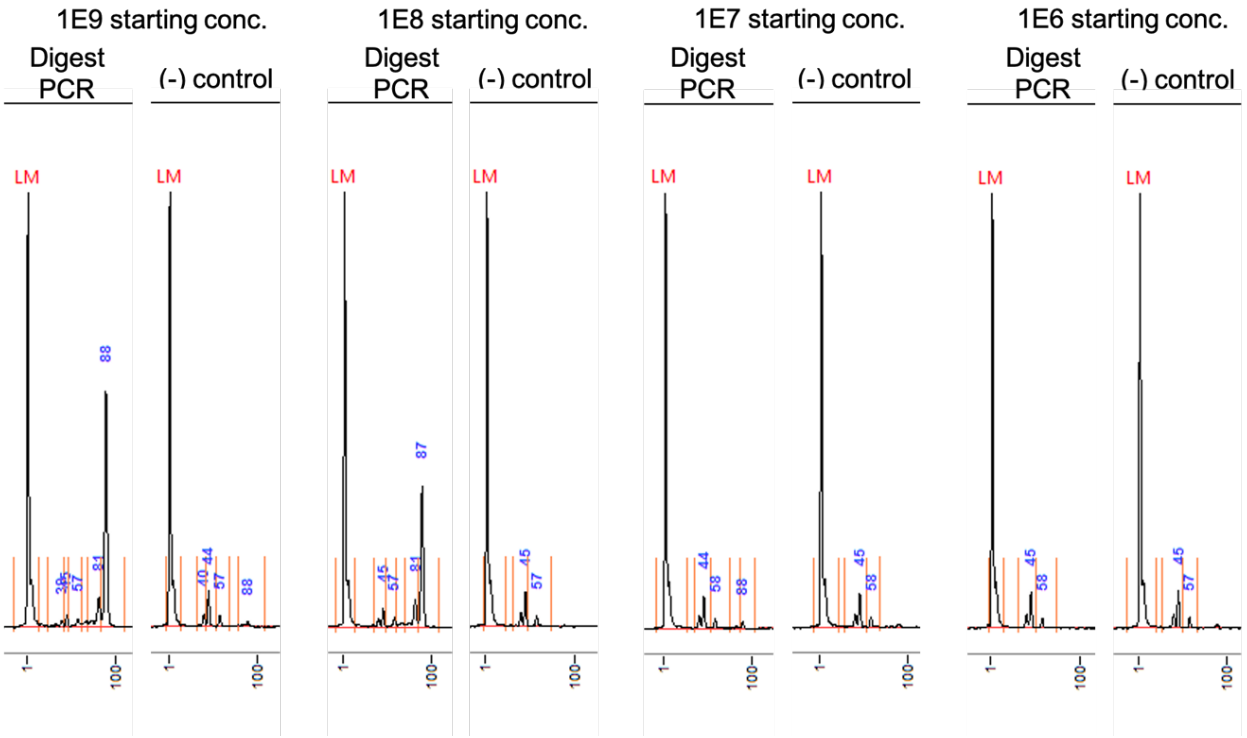


Supplementary Figure C.1. Detailed ligation-based experimental design and strand architecture. (a) Each hairpin oligo has a variable overhang primer binding site, unique barcode region, consensus primer binding site, restriction enzyme recognition site, spacer, and hairpin loop sequence. A primer sequence with a polyA tail will be used to bind the overhanging hairpin strand. (b) The reverse complement sequences are located 3’ to their complementary regions and ordered on the same strand to create self-complementarity.



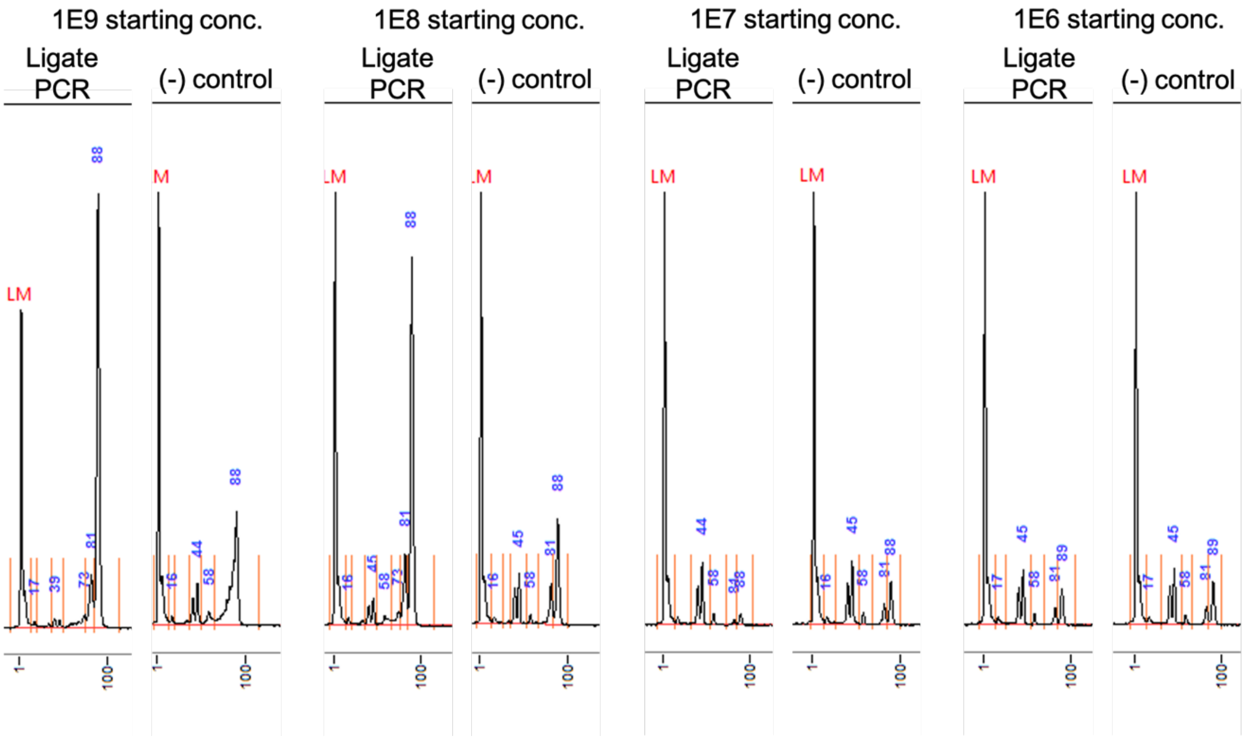
Supplementary Figure C.2. Products from each step of ligaion-based sequence screening match expected size and concentration.

(a) Hairpin structure alone, (b) Primer oligo alone, (c) Primer to hairpin ratio of 1:1 after annealing, (d) Primer to hairpin ratio of 1:5 after annealing, (e) Primer to hairpin ratio of 5:1 after annealing, (f) Primer to hairpin ratio of 1:1 after PCR amplification, (g) Primer to hairpin ratio of 1:5 after PCR amplification, (h) Primer to hairpin ratio of 5:1 after PCR amplification. Data from capillary electrophoresis.



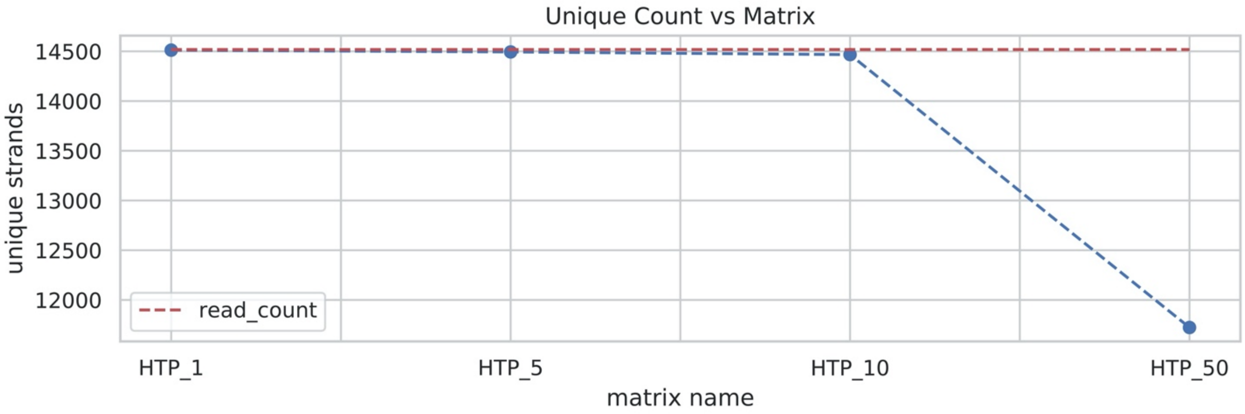
Supplementary Figure C.3. Minimum starting DNA concentration for ligation-based sequence screening determined via capillary electrophoresis.

Decreasing the starting copy number present in the reaction allows for proper amplification using 1E9 and 1E8 copies but not 1E7 and 1E6 copies.

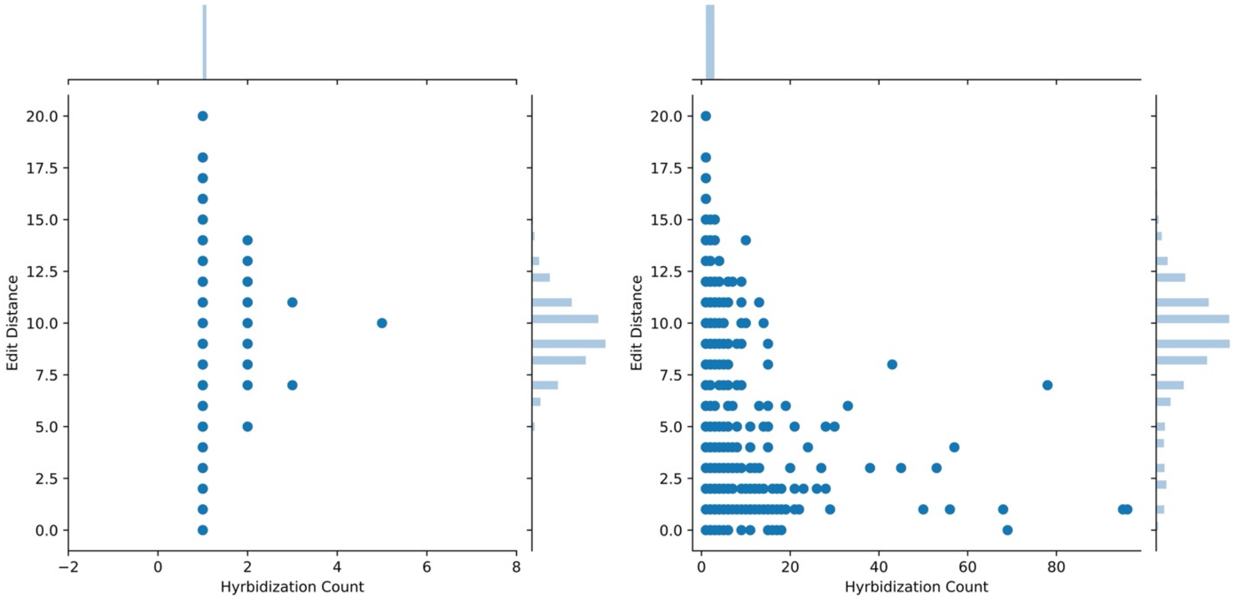


Supplementary Figure C.4. Restriction digestion helps purify products of ligation-based sequence screening.

Skipping the restriction digestion does not impact the ability to amplify the hairpin at 1E9 and 1E8 starting copies but does not help increase yield using 1E7 and 1E6 starting copies. Additionally, negative control samples exhibit amplification when the hairpin is not digested.

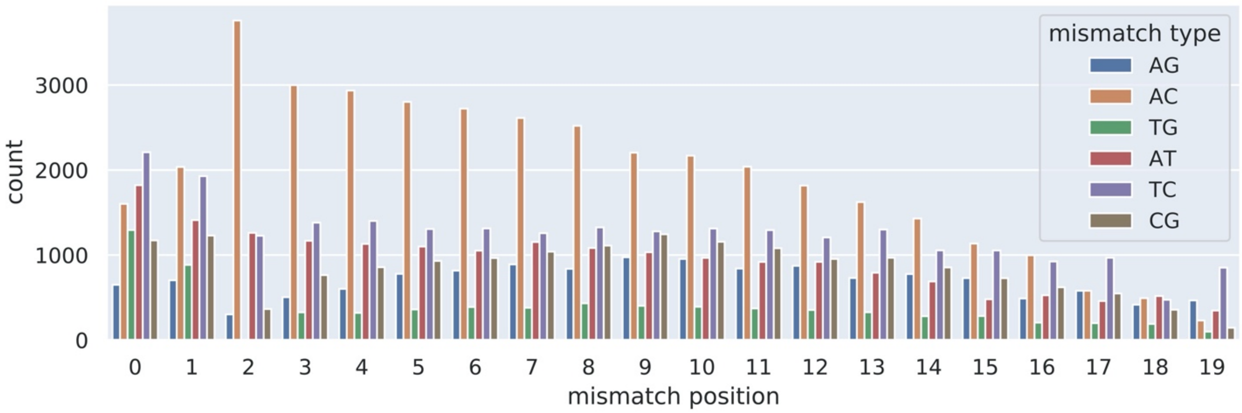


Supplementary Figure C.5. Unique hybridization events decrease as variable primer coverage increases. The number of unique hybridization events measured for each of the different coverages, e.g., HTP_1 indicates a coverage of 1. The number of events is normalized by randomly down sampling the sequencing reads for the coverages of 1, 5, and 10 to the number of reads for the coverage of 50 since it has the fewest recorded events.



Supplementary Figure C.6. Higher variable primer coverage increases the probability that lower edit distance matches will hybridize.

A comparison of the distributions for the number of observed hybridization events and edit distance for each recorded event for the coverages of 5 and 50 (left and right, respectively). The edit distance is the number of edits needed to transform between the random primer and the according library primer to which it annealed.



Supplementary Figure C.7. Total mismatches increase significantly towards the 5' end of primers. Each location on the x-axis indicates the position within the 20 base primers, and each bar shows the raw count of mismatches of each given type. A mismatch type, e.g., 'A-G', indicates that the random primer and library primer disagreed at this location, and one had base 'A' while the other had 'G'. No distinction is made on which primer had which base in this scenario. Data for a coverage of 50 is shown.