The Dissertation Committee for Siyuan Stella Wang
certifies that this is the approved version of the following dissertation:

# Scaling up DNA computation with next-generation sequencing and modified nucleic acids

Committee:

Andrew Ellington, Supervisor

David Soloveichik

Ilya Finkelstein

Edward Marcotte

Eric Anslyn

# Scaling up DNA computation with next-generation sequencing and modified nucleic acids

**by**

**Siyuan Stella Wang**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2022

# Acknowledgments

No accomplishment in life is the product of a singular effort. This dissertation is no exception.

First, I would like to thank the academic mentors who have made this journey possible. I would like to thank my advisor, Andy Ellington, for giving me the opportunity to be part of this program, pushing me to do my best, supporting me, and fostering an innovative spirit in all of us. The Ellington Lab is a truly unique lab environment and I am grateful for all that I have learned here that I would not have been exposed to elsewhere. I would like to acknowledge my committee: David Soloveichik, Ilya Finkelstein, Edward Marcotte, and Eric Anslyn. As I have had the unique privilege of working with nearly all of my committee members, they have provided me guidance that is above and beyond what is expected from a dissertation committee. I thank them for their support, feedback, and time through my candidacy.

I would like to thank, in approximate chronological order, the mentors and collaborators I have worked with who have made this research possible: my undergraduate advisor Lulu Qian, for introducing me to and inspiring me to pursue the field of DNA rational design; Edward, Jag, Angela, and Alex, for giving me the opportunity on the awesome peptide sequencing project for my very first rotation at UT, which I still remember fondly; Cheulhee, for his guidance in my first years in the Ellington Lab and especially for (very patiently) teaching me how to run CHAMP experiments; Erhu, for working with me on the phospho-

rothioate modeling project; Ilya, Steve, Jim, John, and Jami, for taking the time to be a regular source of guidance, feedback, and technical support on my work on the CHAMP platform and from whom I have learned a great deal about next-generation sequencing; and David, Boya, and Cameron, for being fantastic collaborators. I feel so fortunate to have worked on the SIMD||DNA project, and my experience here has helped me grow as a scientist. Last but not least, I want to thank UT GSAF, in particular Gabby and Heather, for working with me extensively and accommodating my numerous strange sequencing requests.

I would like to say a big thank you to the members of the Ellington Lab, past and present, for fostering a culture of creativity in research and casual scientific discussions. Thanks for being ready to help anyone in need and even making the time spent outside of the lab fun. I would like to thank Shaunak, Sanchita, and Jaydin for our many interesting discussions about nucleic acid rational design and beyond. Special thanks to the lab managers throughout the years - Michelle, Arti, Vidya, Cody, and Jose - who helped make the research in the lab possible.

Pursuing post-graduate education is a privilege that I sometimes took for granted all too easily. For this reason, I would like to say a sincere thank you to my parents for supporting my decision to go to grad school (all the way in Texas!) and for emotionally supporting me during this time, as well as my in-laws for being the kindest in-laws I could ever ask for - thank you, Anna, for being my cheerleader! I would like to thank the friends who have made this journey fun: Jay, for all our deep discussions about life, family, and career, and for making me feel heard especially during a time when I really needed meaningful female friendships and perspectives in STEM; Simon, Levi, and Jay, for making Austin feel like home; and Steve, John, and Henry, for our virtual Science Buddies meetings. Finally, I

would like to thank my therapist, the UT CMHC, and the UT dissertation support group, for not only helping me weather the pandemic but also for kickstarting my journey to becoming the best version of myself. I have grown so much as a person in the past two years thanks to what I have learned.

One lesson I will certainly take with me beyond grad school is the importance of maintaining a work-life balance. I am grateful for the hobbies that I have explored during this time and all the people who have helped me perpetuate them. I would like to thank my apartment's personal trainers Savannah and Michelle, and the ICMB running club, for encouraging me to stay active and keep running. I would also like to thank the Instagram makeup artistry community for being such an inviting and dramatically different space from the academic environment I am accustomed to. Thanks for being a supportive outlet for this nerd's creativity.

Most importantly, I would like to thank my partner, Patrick, for so many things that I would be here all day if I were to enumerate them all. But to highlight just a few: thanks for being my advocate, my biggest supporter, enabling my hobbies, enthusiastically answering my questions about linear algebra (even if it wasn't my first time asking!), being the voice of reason, and listening patiently and interestedly to my technical babble about niche experimental details despite being a theorist. Above all, thanks for believing in me.

# Scaling up DNA computation with next-generation sequencing and modified nucleic acids

Publication No. _____

Siyuan Stella Wang, Ph.D.
The University of Texas at Austin, 2022

Supervisor: Andrew Ellington

A central goal of biomolecular engineering is the construction of tools to manipulate nanoscale processes. DNA has proved to be a programmable material suited for this task. DNA strand displacement reactions can be designed to process chemical information in the form of concentrations and sequences. DNA nanotechnology has thus far produced devices for the detection of disease biomarkers, performed computation on chemical inputs, powered mechanical action at both the nanoscale and the macroscale, and assembled precise sub-micron structures from the bottom up.

This dissertation addresses three main topics. First, we develop predictive models for non-canonical nucleic acid hybridization that enable rational design. Second, we show how rationally designed DNA strand displacement reactions can be used to perform computations on information stored in DNA. Third, we present nucleic acid computation with both strand displacement and transcription and discuss strategies for facilitating the scale up of networks. Finally, we discuss data storage in nucleic acid variants in the appendix.

Rational design of DNA circuits and structures is possible because the thermodynamics of DNA and RNA hybridization can be approximated using a nearest-neighbor model. The parameters of this model are typically experimentally determined through the hyperchromism of denatured nucleic acids. This is measured through low-throughput UV-Vis spectrophotometry melting experiments that require a sizable amount of duplexes for a large set of sequences. For non-canonical nucleic acids or non-standard interactions, this characterization can be prohibitively costly and time consuming. Initially, we considered repurposing a next-generation sequencing (NGS) platform for high-throughput mapping of nucleic acid hybridization across a large sequence space; however, we found that the platform is suitable for mapping protein-nucleic acid interactions but not nucleic acid-nucleic acid hybridization due to its dynamic range. We then assessed whether high-resolution melting (HRM) can be used as a rapid method for determining approximate model parameters and found that HRM models can predict relative stabilities between duplexes of different sequences. Using this method, we developed a predictive model for phosphorothioate DNA which we then applied to the design of a phosphorothioate-modified catalytic hairpin assembly circuit.

DNA strand displacement reactions can be used not only to manipulate chemical information in the form of concentration, but also to read and write to more permanent forms of information, such as sequence and secondary structure. We developed and demonstrated a DNA data storage scheme that enables in-memory computation. DNA is a promising data storage medium for meeting today's rapidly growing data storage needs; however, because computation on the stored data is usually performed *in silico*, strands must be sequenced and re-synthesized at every read-write cycle. Our scheme circumvents the bottleneck of *de novo* oligonucleotide synthesis by updating information using strand displacement cascades

that result in sequence changes readable by NGS. We experimentally demonstrated two algorithms - binary counting and cellular automaton Rule 110 - and additionally showed that biologically-occurring DNA sequences without sequence design can be repurposed for storage and computation. Our scheme is capable of computation on multiple data in parallel, as well as random access and sequential computation, allowing for scaled up storage.

Programmable chemical computation is also possible with enzymatic reactions such as transcription. Catalytic activity from enzymes has the potential to simplify circuit design and produce biologically potent signals. Practical concerns to expanding chemical computation circuits such as transcription networks include limited readout of signals and time-consuming purification. We addressed these concerns by expanding on previous efforts to build scalable *in vitro* transcription networks. We updated a single-stranded inhibitory transcription switch design for compatibility with multiplexed NGS readout and developed an analogous single-stranded switch that is activated by nucleic acid signals.

# Table of Contents

**Acknowledgments** **iv**

**Abstract** **vii**

**List of Tables** **xiii**

**List of Figures** **xiv**

**Chapter 1. Introduction** **1**

1.1 DNA as a physical building block . . . . . 3
1.2 DNA as data storage . . . . . 6
1.3 DNA as software . . . . . 10
1.4 DNA as hardware . . . . . 20
1.5 Concluding remarks . . . . . 25

**Chapter 2. Repurposing next-generation sequencing platforms for high-throughput profiling of DNA-based interactions** **31**

2.1 Introduction . . . . . 32
2.2 Results . . . . . 34
2.2.1 Limitations of hybridization profiling on a repurposed sequencing platform 34
2.2.2 Profiling T7 RNA polymerase transcription activity for a library of synthetic promoters . . . . . 37
2.3 Discussion . . . . . 39
2.4 Materials and Methods . . . . . 41

**Chapter 3. Developing predictive hybridization models for phosphorothioated oligonucleotides using high-resolution melting** **52**

3.1 Introduction . . . . . 53
3.2 Results . . . . . 55

3.2.1 Derivation of thermodynamic parameters with high-resolution melting 55
3.2.2 Predictive models for duplexes with fully-PS strands . . . . . 57
3.2.3 Predictive models for duplexes with partially phosphorothioated strands 59
3.2.4 Predicting the impact of phosphorothiate modification on rationally designed nucleic acid circuits . . . . . 60
3.3 Discussion . . . . . 63
3.4 Materials and Methods . . . . . 65

**Chapter 4. Parallel and in-memory computation with data stored in DNA using strand displacement 87**
4.1 Introduction . . . . . 88
4.2 Results . . . . . 90
4.2.1 SIMD||DNA . . . . . 90
4.2.2 Binary Counting Program . . . . . 92
4.2.3 Rule 110 Program . . . . . 95
4.2.4 Random Access . . . . . 97
4.2.5 Sequential computation . . . . . 98
4.3 Discussion . . . . . 100
4.4 Materials and Methods . . . . . 104

**Chapter 5. Reading out in vitro transcription networks with high-throughput sequencing 128**
5.1 Introduction . . . . . 129
5.2 Results . . . . . 130
5.2.1 High-throughput readout of *in vitro* transcription networks . . . . . 130
5.2.2 Towards a single-stranded *in vitro* transcription activating promoter switch . . . . . 132
5.3 Discussion . . . . . 134
5.4 Materials and Methods . . . . . 136

**Appendices 144**

**Appendix A. Recovery of information stored in modified DNA with an evolved polymerase** **145**

A.1 Introduction . . . . . 146

A.2 Results . . . . . 148

A.2.1 Evolution and characterization of a polymerase that could read 2'-OMe DNA . . . . . 148

A.2.2 Encoding and recovery of DNA files . . . . . 151

A.2.3 Computational strategy for reading modified strands . . . . . 153

A.3 Discussion . . . . . 156

A.4 Methods . . . . . 157

**Bibliography** **170**

# List of Tables


3.1 Approximate thermodynamic parameters for PO-PO (phosphodiester-phosphodiester) duplexes derived from HRM data. . . . . 71

3.2 Approximate thermodynamic parameters for PS-PS (phosphorothioate-phosphorothioate) duplexes derived from HRM data. . . . . 73

3.3 Approximate thermodynamic parameters for PS-PO (phosphorothioate-phosphodiester) duplexes derived from HRM data . . . . . 74

3.4 Sequences used for nearest-neighbor parameter determination. . . . . 79

3.5 Sequences and domains used for high-temperature CHA. . . . . 80

3.6 Sequences and domains used for low-temperature CHA. . . . . 81

A.1 Selection conditions for the evolution of a 2'-O-methyl reverse transcriptase using RT-CSR. . . . . 163

A.2 NGS sequencing of the OMe RT-CSR Round 18 pool. . . . . 164

A.3 RTX-Ome variants constructed using NGS data and structure guided design. 169

# List of Figures


1.1 Determination of thermodynamic parameters for predictive models of DNA hybridization. . . . . 27
1.2 Forms of strand displacement. . . . . 28
1.3 Chemical computation: chemical reaction networks (CRNs) to strand-displacement implementation. . . . . 29
1.4 *In vitro* transcription gate designs. . . . . 30

2.1 Theoretical limits of profiling with the repurposed Illumina MiSeq flow cell. . 48
2.2 Experimental profiling of mismatched DNA-DNA hybridization interactions on a repurposed MiSeq platform. . . . . 49
2.3 Experimental profiling of T7 RNA polymerase transcription activity as a function of polymerase-promoter interactions using CHAMP. . . . . 50
2.4 Correlation of promoter variant relative activities to wildtype as measured using CHAMP to reported relative activity from the literature. . . . . 51

3.1 High-resolution melting (HRM) pipeline for determining duplex stability. . . 70
3.2 Comparison of $\Delta G$ predictions made by the HRM-derived model and reported UV-Vis models . . . . . 72
3.3 Predicting duplex stability of partially PS-modified duplexes as a function of sequence (a) or independently of sequence (b). . . . . 75
3.4 Reaction diagram of catalytic hairpin assembly. . . . . 76
3.5 HT-CHA with PO- and PS-H1. . . . . 77
3.6 Introducing PS backbones to select domains in LT-CHA. . . . . 78
3.7 Nearest-neighbor style models considered and the parameters included in each model. . . . . 82
3.8 Leave-one-out cross-validation on the PO-PO HRM dataset for $\Delta G_{50}$ and $T_m$ (concentration = 10 $\mu$M) . . . . . 83
3.9 Thermodynamic parameter determination at different EvaGreen dye concentrations. . . . . 84
3.10 Leave-one-out cross-validation on the PS-PS HRM dataset for $\Delta G_{50}$ and $T_m$ (concentration = 10 $\mu$M) . . . . . 85

3.11 Leave-one-out cross-validation on the PS-PO HRM dataset for $\Delta G_{50}$ and $T_m$ (concentration = 10 $\mu$M) . . . . . 86

4.1 Overview of SIMD||DNA. . . . . 110
4.2 Binary counting program on naturally-occurring sequences. . . . . 111
4.3 Sanger assessment of M13 Register addresses for 0010 and 0111. . . . . 112
4.4 SIMD||DNA single data binary counting program using M13 sub-registers 3 and 8. . . . . 113
4.5 Sanger assessment of different instruction temperatures for binary counting on M13 sub-registers 7, 8, and 9. . . . . 114
4.6 Multiple data readout of independently assembled initial values on M13 sub-register 8. . . . . 115
4.7 Multiple data binary counting on M13 sub-registers 7 and 9. . . . . 116
4.8 Assembly of initial values on M13 sub-registers 7 and 9 on the same M13 plasmids. . . . . 117
4.9 Rule 110 computation with chemically synthesized DNA. . . . . 118
4.10 Program implementation of one timestep of Rule 110 shown on an example register. . . . . 119
4.11 Readout of initial values assembled on chemically synthesized oligonucleotides designed register sequence prior to the computation done in Figure 4.9. . . . 120
4.12 Rule 110 computation on M13 sub-register 1. . . . . 121
4.13 Random access with chemically synthesized DNA. . . . . 122
4.14 Parallel random access for the binary counting algorithm. . . . . 123
4.15 Data erasure by random access. . . . . 124
4.16 Multiple rounds of sequential computation with chemically synthesized DNA. 125
4.17 Quantifying the loss of SIMD products following washing and computation steps. . . . . 126
4.18 Theoretical maximum rounds of computation possible for storing various numbers of unique registers. . . . . 127

5.1 Next-generation sequencing-compatible hairpin switch. . . . . 139
5.2 Transcribed products of sequencing-compatible hairpin switches. . . . . 140
5.3 Measuring transcription inhibition using different readout methods. . . . . 141
5.4 Multiplexed signal readout with NGS. . . . . 142
5.5 A single-stranded transcription activator switch. . . . . 143

A.1 Evolution of a xDNA/polymerase pair creates a platform to secure DNA information. . . . 149
A.2 Structural heat map of mutations that arose during RT-CSR using 2'-O-methyl challenge template. . . . 150
A.3 Primer extension and proofreading activity of RTX-Ome on DNA, RNA, and 2'-O-methyl templates. . . . 152
A.4 Encoding and decoding of information into oligonucleotides. . . . 155
A.5 Characterization of designed RTX-Ome polymerase variants. . . . 165
A.6 DNA Fountain scheme for encoding data files into unmodified and modified oligonucleotides. . . . 166
A.7 Distribution of NGS read sizes. . . . 167
A.8 NGS read reconstruction workflow. . . . 168
A.9 Potential cryptogenetic application for RTX-Ome and other xenopolymerases capable of reading information encoded in xNA oligonucleotides. . . . 169

# Chapter 1

# Introduction

*"...The cell should be considered as a logical and computational machine, processing and managing information. Our objective should be to identify what logical and computational modules operate in cells and how they are derived from the underlying molecular, biochemical and biophysical mechanisms."* - Paul Nurse, The Great Ideas of Biology

Today's concept of computation very often comes hand-in-hand with electronic, silicon-based computers and digital data. Computation, however, is nothing new to Nature and perhaps least of all to living things. Organisms survive by communicating and processing information in the form of chemical signals - in fact, one could even argue that life is simply information that perpetuates itself. Over the course of evolution, fine-tuned biological parts have emerged that allow organisms to effectively sense signals, interpret information, and execute tasks. Rather, it is man-made computing that parallels this architecture with components that perform data storage, software, and hardware.

The chemical computing observed in living systems, however, is markedly different from electronic computing. First, there is often no clear distinction between storage, computation, and execution. Epigenetics provide one example - histones, the protein hardware

that contains genomic DNA within eukaryotic cells, also store a layer of "meta" information in their post-translational modification states that dictates genetic expression. Conversely, although RNA was first understood as a biological implementation of "random access memory", various non-coding RNAs such as miRNA, lncRNA, and siRNA were later found to post-transcriptionally regulate genetic expression. Second, rather than operating with binary ON/OFF signals, chemical information usually exists in a continuous range, being less analogous to digital computers and more similar to analog computers. For instance, multicellular organisms uses concentration gradients of morphogens to determine a body plan during development. Third, biochemical computing tasks are usually distributed across the organism (or a population of organisms). Decentralized processes such as bacterial quorum sensing regulate various key processes for survival and adaptation - for example, microbial populations determine whether to initiate biofilm production, sporulate, or become competent as a function of secreted signaling molecule concentrations, which is often an indicator of cell density. When engineer biochemical computing systems, the challenge is to achieve modularity in a medium that is interconnected by nature.

Given the remarkable adaptability and complexity that life is capable of, it is no surprise that much research in recent decades has sought to understand biochemical processes and reprogram them with novel behaviors. Synthetic biology strives to manipulate organisms, cells, and even minimal *in vitro* systems into tools. Both the complexity and specialization of biology challenge this goal, however. If we want to better understand how to design synthetic life by distilling the design principles of biochemical networks, we need a "programmable" chemistry. DNA has proved itself suitable for this role as a building block of synthetic biochemical networks and applications beyond.

In this thesis, we look specifically at how chemical information can be encoded and processed with nucleic acid-based systems. We begin with what makes DNA suitable for rational design and later focus on the implementations of data storage, software, and hardware in DNA nanotechnology.

## 1.1 DNA as a physical building block

DNA exhibits well-studied and highly predictable chemical behaviors, and its production and usage have improved rapidly in recent decades, making it not only a useful handle for manipulating biological systems but also a customizable molecule for applications beyond molecular biology. DNA strands hybridize by Watson-Crick base pairing rules and accordingly fold into thermodynamically stable secondary structures. The stability of a double-stranded DNA duplex can be precisely calculated using nearest-neighbor models, which assume that the primary contribution to duplex thermodynamics comes from base stacking interactions between adjacent nucleobases. Therefore, thermodynamic properties for any duplex may be predicted using the combined contributions of all nearest-neighbors contained in the duplex [48, 41] (Fig. 1.1A). Duplex stability correlates positively with concentrations of monovalent cations (e.g. sodium, potassium) and is also a function of the concentration of other ions (magnesium) or reagents (dNTPs) present in the solution [141]. This allows the stability of a DNA duplex at any temperature to be estimated simply by its sequence and relevant buffer parameters.

The parameters of the nearest-neighbor model for DNA duplex stability were determined by various groups from the 1980s to the early 2000s using UV-Vis spectrophotometry

[76, 46, 50, 162, 183, 161], which directly measures the double-stranded to single-stranded transition during duplex melting by leveraging the innate hyperchromicity (i.e. increased UV absorbance in single-stranded form) of DNA. The melting curve that results may be analyzed to extract properties such as free energy, enthalpy, and entropy of formation (Fig 1.1B, [148, 132, 151]). The nearest-neighbor model can also be applied beyond perfectly complementary hybridization between DNA strands - parameters have been derived for DNA structural motifs [164], RNA [62, 195], DNA-RNA hybrids [184], and some non-canonical nucleic acids, such as variants with unnatural backbones (e.g. linked nucleic acids, 2'-o-methylation) [144, 107] or synthetic nucleobases [91]. Available softwares apply these parameters to predict the stability of two complementary strands or even the expected secondary structures of a set of strands [87, 229, 160]. These models also make it possible to analyze properties of nucleic acid duplexes that are crucial for molecular biology applications [142] or to design multi-stranded DNA-based systems that are programmed to switch between multiple conformations based on their input or environmental conditions [218].

While predictive parameters drive the rational design of nucleic acid-based systems, similar parameters have not yet been determined for most nucleic acid analogues, many of which contain chemical modifications (e.g. phosphorothioates, mesyl phosphoramidate, 2'-fluoro modifications) that are useful in diagnostics or therapeutics. This is in part due to the high material requirements and costs of UV-Vis melting experiments, which are low-throughput and necessitate materials on the order of nanomoles. In Chapters 2 and 3, we explore alternative experimental approaches to deriving sequence-based predictive models and consider their limitations and tradeoffs.

DNA is highly accessible as a commercial product. Oligonucleotides with custom

sequences can be rapidly and cheaply produced at scale - at time of writing, a 30 nucleotide (nt) oligonucleotide costs ∼10 USD for 25 nmoles and typically ships in 1 business day from the largest suppliers. Phosphoramidite chemistry on solid-phase synthesis is the dominant synthesis method currently in use, despite some limitations with yield and quality at longer lengths as a result of failed coupling [93]. Larger custom sequences on the order of hundreds to thousand bases can be pieced together from chemically synthesized fragments and biologically amplified as double-stranded duplexes. Alternatively, high-quality, kilobase-long single-stranded DNA, albeit with native sequences, can be extracted from sources such as bacteriophages; methods have been reported to produce long strands with custom sequences by cloning [53]. A plethora of modification chemistries are available for functionalizing DNA, including adding fluorescent signals (fluorophore conjugation), increasing resistance to degradation caused by nuclease enzymes (backbone modifications such as phosphorothioation or 2'-O-methylation), and attaching labels for later covalent (e.g. O6-benzylguanine aka SNAP-tag) or noncovalent (e.g. biotin) linking to other biomolecules such as proteins, antibodies, dyes, or even another oligonucleotide.

Rapid advances in sequencing in recent decades through the development of high-throughput next-generation sequencing platforms [126] have dramatically improved our capacity to work with DNA in many areas ranging from clinical contexts to biotechnology to fundamental research by combining chemical, electrochemical, optical, and enzymatic advances. Among the most popular technologies are Illumina sequencing-by-synthesis platforms, such as the Illumina MiSeq, which enable massively parallel sequence readout on the order of tens of millions of individual sequences for sequences up to several hundreds of base pairs in length [14]. While these platforms are predominantly used to analyze se-

quence information (as in genome science) or to detect the presence of sequences implicated in disease states (as in diagnostics), the DNA microarray-like parallel presentation of covalently-attached, heterogeneous sequences may be additionally useful in probing biomolecular interactions involving DNA. In Chapter 2, we explore the possibilities and limitations of repurposing parallel sequencing platforms for high-throughput analysis of DNA-based interactions. As the technology to read longer sequences matures, platforms for portable sequencing in low-resource conditions [128], low-cost rapid sequencing in real-time [158], and single-molecule sequencing [53] have also emerged. Together, these factors accelerate the use of DNA in applications beyond molecular biology.

## 1.2 DNA as data storage

Given that DNA evolved as a biological information storage medium, it is not surprising that it is currently being considered for the storage of man-made information. The volume of data consumption is increasing exponentially and has exceeded expectations in the last two years, possibly from higher demand due to the 2020 pandemic - 181 zettabytes of data are projected to be produced in 2025 alone [5]. With this explosive growth comes an urgent need for space-efficient, stable, and cost-effective means of archiving data. The capacity for electronic data storage, while also growing steadily in recent years, is dwarfed by data consumption - in 2020, the worldwide data storage capacity reached 6.7 zettabytes, little more than 10% of the total data produced that year, which is 64.2 zettabytes.

DNA's 4-letter nucleotide alphabet translates to a theoretical maximum information density of 455 exabytes/gram [39] in which each nucleotide position encodes 2 bits, with

recent experimental results demonstrating encoding and recovery of 215 petabytes per gram [59]. The successful extraction of sequencing-viable DNA from fossilized remains suggests that, under the right conditions, a specific DNA sample could be preserved for centuries to millennia [11, 130]. Moreover, genetic information is perpetuated in living cells through both replicative and error-correcting molecular mechanisms. Synthetic oligonucleotides are stable for years when lyophilized or dissolved in appropriate buffers and stored in a freezer [3], and double-stranded DNA can be kept at -20°C for years without significant loss. Stable room-temperature storage is achievable with encapsulation within a matrix to protect from heat, radiation, humidity, enzyme contaminants, and other factors [77]. Minuscule amounts of DNA down to just several copies are theoretically amplifiable with PCR. In contrast, the typical external hard drive has a lifespan of 5 years and costs 100 USD to replace in 2022. US data centers consumed nearly 70 billion kilowatt hours of electricity and over 600 billion liters of water in 2014. Flash memory has a theoretical maximum density of 70 GB/g; magnetic tape storage, while less consuming of resources and cheaper to produce, has a significantly lower data density capacity than DNA at a current maximum of 201 GB/in$^2$ as of 2017 [2]. Thus, over the past two decades DNA storage has become an increasingly appealing form of alternative data storage to meet the demands of the digital world.

Modern in vitro DNA storage schemes typically operate on a read-write cycle that involves encoding, synthesis, sequencing, and decoding. In this form of DNA storage, the DNA is more akin to hard drives than random access memory, because computation does not occur "in memory". Accessing any specific piece of information ("random access") can be achieved by PCR with primers that target the information of interest [138, 187, 210, 17]. Like other data storage media, DNA must be robust to data corruption. Assuming that en-

coding and decoding steps can be done perfectly in silico, these errors can accumulate in the synthesis (abasic sites, truncation, deletions), storage (temperature fluctuations, radiation damage, nuclease contamination), and sequencing (missing fluorophore, polymerase misincorporation) steps. Simple redundancy can resolve information loss by reconstructing lost data using a consensus; however, this strategy significantly reduces the desirable high data density promised by DNA storage and does not necessarily protect data from corruption. For this reason, error-resistant encoding and decoding schemes are critical to overcoming error accumulation. An early scheme demonstrated by Goldman et al. in 2013 encoded data in overlapping segments at four-fold redundancy to protect against missing oligos [74]. Bornholt et al. later built on this idea by applying a logical XOR to reduce redundancy to 1.5-fold [17]. Some schemes have borrowed from coding theory to apply existing error-correcting codes with great success. Works by Grass et al. and Blawatt et al. have incorporated Reed-Solomon codes to protect against dropout of oligos [77, 16]. Ehrlich and Zielinski's DNA Fountain scheme in 2017 adapted fountain codes for DNA data encoding and enabled both detection and correction of error that was tolerant of missing oligonucleotides, allowing full recovery at very low redundancy.

The cost and accessibility of sequencing and synthesis are the key bottlenecks to adapting DNA data storage for widespread use. Sequencing has seen many advances in recent years thanks to the development of novel high-throughput platforms. On the other hand, the cost of synthesis is decreasing much more slowly than for sequencing, and fundamental limitations prevent the large-scale production of long (>100 nts), high-quality oligonucleotides. One solution may involve enzymatic de novo oligonucleotide synthesis [146]. Although the incorporation of a determinate number of nucleotides (as opposed to homomer runs) remains

a challenge, Church and colleagues have demonstrated a scheme that can nevertheless utilize enzymatic synthesis by encoding information in transitions between homomer regions [117]. Another solution would be to repurpose biologically occurring DNA by storing information in the topology rather than the base sequence of duplexes. Tabatabaei et al. have developed a write system that involves the programmable restriction enzyme Pyrococcus furiosus Argonaute to store information by nicking specified positions in E. coli genomic DNA [186]. Although these schemes in theory have a lower data density by not encoding directly with the sequence (experimentally achieved 4 EB/g in Tabatabaei et al., theoretical maximum of 90 EB/g Lee et al.), they are no longer limited by strand length, less prone to data loss, and can in practice more efficiently use adaptor or metadata sequences.

Directly editing the encoded information could allow costly and time-consuming cycles to be skipped by reusing existing strands. Another work by Milenkovic and colleagues uses PCR-based methods to access and edit data in vitro [187]. This work used a prefix-synchronized code to store words in a lookup dictionary for efficient storage. While the schemes presented in this work circumvent de novo synthesis of an updated sequence, they require shorter fragments to be synthesized as primers containing new information. Recently, Wang et al. have proposed an in-memory computation scheme based on single instruction, multiple data operation using DNA strand displacement (SIMD||DNA) in which the encoded data is a direct function of the computational output [200]. This requires no intermediate synthesis steps, as information is encoded in the position of nicks in predetermined regions corresponding to bit values. In Chapter 4, we demonstrate the scalability of SIMD‖DNA by coupling computational outputs to sequencing readout by NGS. We additionally reduce the need for long oligonucleotide synthesis by repurposing naturally-occurring M13 DNA for

data storage.

Complementary to the maturation of DNA storage is the development of algorithms, chemistries, or proteins that make it possible to store and retrieve information in polymers other than DNA. At present, phosphodiester DNA is the primary medium in which data storage has been implemented. RNA is a significantly less stable medium considering its 2' hydroxyl group allows self-hydrolysis to occur spontaneously, and RNA nucleases are far more ubiquitous in the environment than DNA nucleases. Encoding information in chemically modified nucleic acids either as a means of protection against nuclease contamination or to build parallel channels of information is possible if the necessary protein tools are available. In Appendix A, we demonstrate how directed evolution methods can facilitate this goal by expanding the space of viable storage media to non-canonical nucleic acids through the development of novel polymerases. Additionally, we update the DNA Fountain coding scheme [59] to accommodate 2'-o-methyl-modified oligonucleotides, the chemical synthesis of which is more prone to deletion errors.

## 1.3 DNA as software

The programming language of biochemical networks is implemented with concentrations and chemical kinetics. Just as models of duplex stability allow the specification of precise binding through sequence, experimental data on the DNA hybridization kinetics enables the prescription of desirable hybridization behaviors by tuning reaction rates. DNA kinetics can be accurately and sensitively measured using fluorescence [135]. Association constants of $10^5$ to $10^7$ $M^{-1}$ $s^{-1}$ have been reported for two complementary strands from 10

to 30 nts long hybridizing at room temperature under high cation conditions [135, 65, 225]. A bimolecular association constant of $10^6$ $M^{-1}$ $s^{-1}$ approximately translates to a time-to-half completion of about 30 seconds [175]. Dissociation varies dramatically with length; for example, the half-life of dissociation can range from minutes to centuries for a 10-base pair duplex to a 20-base pair duplex [135]. The possibility of making G-C bonds or sequence repeats increase the chances of forming a stable initial bond that may then zipper or adjust into the hybridized form [139]. More recently, a weighted neighbor voting algorithm that predicts the rate constant of hybridization for a given sequence to within a factor of 3 with 91% accuracy has been developed [226].

Given this understanding, length and sequence composition can translate to tunable parameters with which to design hybridization reactions with custom kinetics. DNA strand displacement (DSD) is a reaction in which a hybridized (target) strand in a duplex changes its strand partner. This is energetically favorable when the exchange results in the overall maintenance or increase of the total number of bases paired (Figure 1.2). For such a reaction to occur, the new partner strand must make an initial contact with an unbound base pair on the target strand. Because the duplex is highly stable and fraying occurs slowly (Figure 1.2A), DSD may be accelerated by a toehold - a single-stranded region on the hybridized strand. Upon binding to the toehold, the new partner strand subsequently competes with the incumbent partner in a random walk process called branch migration. When the new partner fully exchanges all base pairing contacts with the incumbent partner, the incumbent partner is either completely unbound and dissociated (toehold-mediated strand displacement; Figure 1.2B) or remains bound to a toehold on the target that is not complementary to the new partner (toehold exchange; Figure 1.2C). In toehold exchange reactions, because

toeholds are usually only 2 to 6 nucleotides long, the incumbent strand may spontaneously dissociate after a short period of time, leaving behind a free toehold that may initiate further DSD reactions. The effective rate of toehold-mediated strand displacement is affected by the toehold binding strength (which is a function of sequence and length) and varies from 10 M-1s-1 to 107 M-1s-1 for toeholds 1 to 7 nucleotides in length [225]. Toehold exchange reaction rates depend on the lengths of the toeholds. When the first toehold is longer than the second, the first toehold length determines the rate (since part of the second toehold is indistinguishable from the branch migration); when the second toehold is longer, the reaction rate varies with the difference in length between the toeholds [225]. Additional control over reaction rates are possible through variations of toehold exchange such as remote toeholds, which can be used to initiate strand invasion, branch migration, and displacement at regions non-adjacent to the toehold (Figure 1.2D) [68]. Likewise, strand displacement rates can be controlled in RNA; computational simulation studies on RNA show that the rates of toehold-mediated strand displacement in RNA range from 1 $M^{-1}s^{-1}$ to $10^6$ $M^{-1}$ $s^{-1}$ (similar to DNA), with toeholds on the 5' end resulting in a much faster rate than the same toehold sequence on the 3' end [185].

This precise control over the rates of hybridization-based reactions makes it possible to engineer systems of chemical reactions that are driven by chemical equilibrium to exhibit defined behaviors. At an abstract level, any chemical reaction can be simplified to a set of reactant and product species and their corresponding stoichiometries. A system of such reactions can be represented as a chemical reaction network (CRN). To compute with CRNs, information is encoded as the concentration of a species, and reactions that involve the species act as the algorithm. In this manner, CRNs can be devised to carry out various functions

on chemical inputs [176, 31]. Examples of such functions include linear functions such as addition (Figure 1.3A) and nonlinear functions such as the maximum function (Figure 1.3B). DSD programming languages have been developed to translate CRNs to experimental DNA implementations [176, 24, 31]. In DSD implementations, a species is defined as the combination of a strand and its bound state - for instance, a target species that exists mostly in the bound state at the beginning of the computation may become increasingly unbound over the course of the computation, and the concentration of the unbound target strand is defined as the output signal. This convention is due to the fact that many DSD systems seek to be enzyme-free to maximize programmability, which comes at the cost of forgoing de novo synthesis or degradation of strands within the system. Such DSD systems are constructed at a far-from-equilibrium state so that equilibrium will drive forward the computation, like a compressed spring. A system will therefore have the same overall concentration of a particular DNA strand at the beginning and the end of the computation. To accelerate the forward reaction, many DNA-only systems use a high concentration of "fuel" molecules that are consumed (i.e. reached their energetically favored state) to amplify the signal, driving computation in the process. The output of a computation is generally coupled to fluorescence signaling through a molecular beacon - a double-stranded or hairpin structure with a single pair of covalently conjugated fluorophore and quencher molecules that become physically separated upon binding to the target strand. The fluorescence signal is monitored over the course of the computation as an indication of the output at any given time.

Various circuits have been experimentally implemented with DSD reaction motifs to perform computational tasks from digital logic to spatial pattern formation. The seesaw gate (Fig 1.3C) is a motif by Qian and Winfree that generalizes toehold exchange reactions

[152]. Input strands contain two longer regions, or domains, surrounding a central toehold. Seesaw gates are complexes with a top strand (similar in structure to an input strand) that is partially hybridized to a bottom strand. Seesaw gates each include two toeholds. Inputs are identified by the sequences of their domains - only if an input strand contains a complementary domain to a gate will toehold exchange occur and displace the top strand (output of the gate), which may act as input downstream. Boolean logic circuits that implement bitwise mathematical operations [152] and neural network pattern recognition circuits [153, 34] have been constructed using cascading layers of seesaw gates. Qian and colleagues trained neural networks in silico to determine appropriate weights to achieve memory for several 4-bit patterns that were then encoded in the concentrations of gate complexes. However, negative weights are required in neural networks to calculate the weighted summation of all inputs but cannot be represented as a negative concentration. As a workaround, Qian and colleagues have in one instance used dual-rail representations that separately encode "positive" and "negative" concentrations both as positive concentrations [153] and in another used winner-take-all motifs based on pairwise annihilation [34]. The output of DSD computations may also be dynamic and generate patterns over time or location, impressively without the use of enzymes. The "rock-paper-scissor" oscillator is a CRN containing three species that interconvert to produce waves with defined periods. Using only DNA hybridization reactions, Srinivas et al. compiled this CRN into a DSD system and demonstrated in vitro oscillatory behavior, with each target species completing more than 2 cycles over the course of 50 hours [178]. Chirieleison et al. constructed a DSD circuit that performs edge detection on the area of UV light exposure and to produce a fluorescent pattern [35]. The circuit uses photolabile linkers that become cleaved upon UV irradiation to release upstream signaling strands

while also inhibiting activation of downstream signals. This enacts an incoherent feedforward loop, which produces a pulse of signal over time. Released signals are amplified in a process called catalytic hairpin assembly [119], in which signal strands catalyze hybridization between kinetically trapped hairpin strands. The signal is amplified and transduced into a downstream signal in the form of a single-stranded domain, which in turn interacts with fluorescent reporter complexes. Because the system is spatially distributed across a heavily cross-linked media that slows diffusion, signals accumulate to form a signal gradient along the edge of the irradiated area. DSD systems that can generate lasting, complex patterns that are discrete [166] or continuous [165] have also been proposed. The assembly of large DNA structures can be conditionally initiated through the output of DSD circuits. Similar to catalytic hairpin assembly, the hybridization chain reaction uses a single-stranded DNA strand to catalyze enzyme-free hybridization between metastable hairpins [49]. This technique has applications in *in situ* hybridization, where the unbounded multi-stranded assemblies can serve as fluorescent detectors with signals orders of magnitude brighter than one-to-one in situ probes [36, 37]. Schulman and colleagues have adapted hybridization chain reaction for finite assembly to drive the controlled expansion of hydrogels to produce mechanical motion [23], as well as a "locked" design that becomes activated upon strand displacement by an upstream DNA signal [61]. This diversity of behaviors that may be achieved using DNA alone is a testament to the success of nucleic acid rational design strategies.

Despite its versatility, however, DNA is ultimately not as chemically reactive as enzymes. Enzyme components can not only more efficiently drive DNA circuits towards completion but can actuate biologically potent responses to molecular input. To balance the tradeoff between efficiency and programmability, recent works have included proteins from

routine molecular biology application as standard parts of *in vitro* chemical circuits, with strand displacement reactions as the customizable components. This has achieved the goal of recapitulating complex biochemical dynamics with minimal, rationally designed systems. For example, the Polymerase/Exonuclease/Nickase (PEN) system developed by Rondelez and colleagues consists of a DNA template, a DNA primer, Bst DNA polymerase, nickases, and exonuclease that operates by enzymatically synthesizing new strands and degrading existing strands [9]. Complex far-from-equilibrium behaviors have been shown using the PEN system, including oscillations [84], stable chemical spatial gradients [220], and traveling waves [145, 221, 219].

T7 RNA polymerase (T7 RNAP) is another commonly used enzyme tool for in vitro computation. T7 RNAP is readily used in vitro because it requires a single subunit, and is both highly specific to its promoter (reducible to a minimal double-stranded 17-nt sequence) and highly active [28, 180]. Using T7 RNAP transcription as a means of producing circuit parts, Schaffter and Strychalski developed an RNA version of the seesaw gate by Qian and Winfree for boolean logic operation [168]. These transcribed RNA gates have a practical advantage over DNA complex gates in that the RNA gate transcripts contain a self-cleaving ribozyme, ensuring one-to-one assembly in vitro without additional purification steps. Furthermore, the state of transcriptional activity itself can be the output of computation. Kim et al. introduced the T7 transcription gate in 2004, a circuit element that conditionally transcribes RNA strands based on the hybridization state of the T7 promoter (Figure 1.4A). As transcription is dependent on the presence of a double-stranded promoter region, gates are switched ON upon binding of a DNA signal strand complementary to the promoter sequence. RNA transcripts can act as inhibitors of input for downstream gates by toehold displace-

ment of the promoter top strand, or indirectly as activators by binding complements of DNA signal strands and thereby freeing the signal to interact downstream. RNase H is included in these transcriptional circuits, which cleaves RNA strands bound to DNA as in the case of the inhibited signal, allowing the system to produce dynamic outputs. Kim and Winfree have since proposed transcriptional circuit designs for logic circuits and neural networks [108] and experimentally demonstrated oscillatory, bistable, and pulse behavior using transcription [110, 111, 204, 109]. Schaffter and Schulman have expanded on this design of the transcription gate with additional motifs for state induction and signaling to scale up circuits, and have demonstrated circuits with feed-forward architecture and switchable states that contain up to four transcription gates and induction nodes [167]. Kar and Ellington have sought a more scalable conditional transcription gate design by developing a single-stranded hairpin transcription gate (Figure 1.4B) [104]. This hairpin circuit element can act as an inhibitor for one or two input strands, performing boolean NOT and NAND operations. Rationally designed transcription factors by Chou and Shih present another method to switch between active and inactive transcription that uses a DNA tether covalently attached to a T7 RNAP as a component of DNA-based transcription factors [38]. Upon strand displacement that results in the tether binding to a template strand - which contains complementary single-stranded domains to the tether, a double-stranded T7 promoter, and a double-stranded gene - transcription of the gene is activated.

Finally, DSD circuits can actuate phenotypic change by interfacing with enzymes in vivo. Several generations of riboswitches - RNA transcripts that are conditionally translated by the secondary structure of the ribosome binding site (RBS) - have been in development in the past two decades. Starting with Isaacs et al. in 2004, riboregulator designs initially

involved sequestering the ribosome binding site by direct hybridization with an upstream portion of the transcript. Later, Green et al. have presented designs with fewer sequence constraints that prohibit ribosome binding by secluding the RBS within a hairpin that becomes unfolded and available for translation upon binding of a linear trigger RNA strand [81]. This design has since been expanded to accommodate up to 4 inputs for logic computation [80], in addition to a repressor switch design that prohibits ribosome binding with a highly stable RNA three-way-junction [112]. These designs also yielded a fair number of mutually orthogonal sets (18 activators and 15 repressors) that may be used within the same system without significant crosstalk. In parallel, Chappell et al. have presented components using transcription attenuators to form upstream terminator and anti-terminator sequences to regulate transcription [29].

Nucleic acid circuits with and without enzymatic components have been successfully applied to molecular detection. Isothermal amplification techniques such as catalytic hairpin assembly, loop-mediated isothermal amplification [136], rolling circle amplification [47], and strand displacement amplification [198], are well-suited as a single-component circuit for point-of-care diagnostics because their isothermal operation does not require special equipment such as PCR machines [71]. These techniques have been used in a variety of assays, including detection of pathogens (e.g. HCV, chlamydia) [82, 18] and gene analysis (e.g. SNP detection) [47]. Detection of single disease analytes is usually computationally simple, involving activation or inhibition conditional on the presence of the target molecule. More complex circuits, like molecular classifiers, can be used to make more diagnoses based on multiple disease markers. For instance, Lopez et al. developed multi-gene classifiers to distinguish healthy plasma from cancer plasma or bacterial infections from viral infections by

detecting concentrations of seven RNA transcripts of interest, achieving correct classification on all 12 patient samples tested in vitro [127]. Since then, Zhang et al. have implemented a four-gene classifier that takes miRNA profiles from serum samples of healthy and lung cancer patients as input [223]. This circuit achieved an accuracy of 86.4% in distinguishing between healthy and diseased states in a total of 22 samples.

Several challenges to DNA computation bottleneck its progress. First, scaling up rationally designed circuits is often challenging. Most commonly used gate components are multi-stranded complexes and therefore must be stoichiometrically and correctly annealed prior to use to prevent free strands from interfering with downstream signals. For this reason, complexes generally require gel purification to remove unbound strands. As the size of the circuit increases, this time-consuming operation becomes less and less feasible. At present, the largest DNA-only circuit contains around a hundred gates [34] which, while impressive, is still far from the complexity of natural biochemical systems. Nucleic acid circuits involving transcription have presented some solutions to this issue; for instance, the single-stranded transcription switch design by Kar and Ellington [104] and the self-cleaving seesaw gates by Schaffter and Strychalski [168] ensure a one-to-one ratio between the top and bottom sequences. Second, because signals are generally represented as concentrations of specific molecules, it can be difficult to read multiple results in parallel within one sample. Signals are read out using fluorescence, meaning multiplexed signal readout requires distinct, conjugable fluorophores with non-overlapping emission spectra. It may therefore be helpful to couple signals to high-throughput quantification such as qPCR, NGS, or other methods that are equipped to analyze mixed populations. For instance, assays in which computational output is designed to produce a change in sequence (or possibly even use

sequence as an additional layer of instruction through mismatches, partial complementarity, etc.) could take advantage of high-throughput sequencing technologies. Third, while in theory even short sequence domains promise orthogonality because of the large space of possible sequences, similarities between domains could cause cross-talk and leak in reality and need to be addressed empirically. Strategies for reducing unwanted leak in DSD systems have been presented [202]. Addressing leak in a system is a slow troubleshooting process that could potentially also benefit from multiplexed sequencing readout which captures a snapshot of all output and intermediate signals at once for more transparent debugging.

In Chapter 4, we present a DSD scheme that results in a change of sequence, making it compatible with next-generation sequencing and therefore scalable. In Chapter 5, we show that next-generation sequencing and qPCR may also be used as a means of quantifying concentration at a scale.

## 1.4 DNA as hardware

The programmability of DNA not only makes it suitable as a medium for biochemical software, but also a choice substrate for prescribing exact structures at the nanometer scale. In doing so, DNA structural assemblies have filled a gap in the demand for customizable nanotechnology by taking a bottom-up approach. Although we will not cover any work by the author in this area in a later chapter, we will nevertheless address some of the notable works and key directions in this field given its rapid growth and relevance to the fields of DNA data storage and DNA computation.

DNA nanostructures are composed of repeating structural motifs. Holliday junctions

are a type of stable immobile multi-stranded DNA structure that may be used as a tile for 2D assemblies, or connected in 3D to produce wireframe structures [171]. Another common style of DNA structure is the origami tile, which folds a long, single-stranded DNA scaffold using shorter oligonucleotide staples to produce filled tiles [159]. Freeform structures can also be made using the scaffold and staple method [99]. The scaffold DNA, which is usually kilobases long and necessarily single-stranded, is sourced from bacteriophages that produce circular, ssDNA genomes [26]; because the native sequence is unaltered, the shorter oligonucleotide staples are designed to complement sequences in different regions and bring them together physically. If an artificial sequence is desired, it can be produced enzymatically and at the correct stoichiometry either by in vitro amplification techniques such as rolling-circle amplification or by replication in vivo with bacteriophage [53]. DNA nanostructures have been produced and assembled in vivo by genetically encoding components strands and reverse transcribing selected transcripts to produce the ssDNA components [56]. RNA has additionally been explored as a material for assembling nanostructures. Both standalone structures and repeating meshes (up to 100 nm without deformities) have been produced cotranscriptionally [67, 83, 123], and structures can even be assembled in vivo [120].

To produce more modular components that depend on local interactions for assembly rather than "seeding" interactions, designs using only short DNA oligonucleotides have emerged [212, 106]. These assemblies use fairly short (32 to 64 nts) oligonucleotides that hybridize across multiple strands to produce a web of connections between strands in a manner reminiscent of LEGO blocks. Recently, crisscross polymerization has been demonstrated as a rapid, highly multi-stranded assembly mechanism for building large 2D slat structures [134], with the latest designs reaching multiple microns in dimension [205]. Structures may

be periodically assembled with repeating strands to produce larger structures of unbounded dimensions, or aperiodically assembled using unique strands, the latter case trading off assembly size for the ability to uniquely address specific locations.

This bottom-up control over submicron geometry has broad applications across disciplines from therapeutics to optics. Targeted delivery of drugs in vivo has been widely pursued. Molecular payloads contained within DNA origami structures may be delivered to specific regions of the body by passive targeting (accumulation in target regions) [227], and the structures containing these drugs may be programmed for conditional release in the presence of protein biomarkers [6]. DNA and RNA nanostructures can also act as scaffolds for other biomolecules such as proteins and small molecules [66]. In certain disease states, the geometry between drug molecules can have a large impact on treatment efficiency; DNA structures can scaffold the precise positions and stoichiometries of drug molecules to more effectively deliver these drugs [222]. DNA origami scaffolds can serve as templates for organic synthesis for higher yield and improved chemoselectivity [196]. Functional structures such as synthetic lipid membrane channels have been constructed [116]. Nanofabrication applications have also utilized the precise control afforded by nucleic acid hybridization; the production of nanoscale metallic devices with programmable plasmonic properties can be achieved with the help of origami scaffolds [75, 170]. DNA nanostructures have also been used as tools to investigate biophysical [64, 63] or compositional [174] properties of proteins, and as tools for gene detection similar to microarrays [105, 181]. This space of applications can be further expanded with various conjugation techniques that enable the construction of functionalized, robust nanostructures [209]; for instance, Structures can be chemically coated to be made significantly more nuclease resistant for in vivo use [7].

DNA-based systems have also explored several methods for creating mechanical motion through the binding and unbinding of sequence domains. A well-studied example is the DNA walker, which is a nanoscale machine that makes directional progress (walk) on DNA duplexes (track) by alternatingly hybridizing its single-stranded overhangs (feet) with free overhangs from the tracks (footholds). Initial designs were inspired by the "walking" motion of kinesin and dynein along microtubule filaments. In 2004, Shin and Pierce developed a double-stranded DNA walker whose walking directional movement along a duplex track could be controlled by adding specific strands [172]. These strands can attach one foot of the walker to any foothold within reach, or release a bound foot by displacing a previous attachment strand. An autonomous design was developed around the same time by Yin et al. in which a DNA walker travels between posts by hybridization, followed by ligation and restriction digest to produce motion using a "scorched earth" strategy [213].

The main areas of improvement for dynamic DNA structures are processivity, speed, and directional control. Jung et al. showed that the addition of a simple 8-nt "cleat" - an extension of the toehold region - to a DNA walker resulted in up to 47 steps while remaining bound for 12 hours [101], significantly improving processivity (albeit trading off speed) compared to a previous cleat-less design at 36 steps in 40 minutes [100]. Going beyond the idea of a two-legged walker, Yehl et al. produced high-speed DNA-based rollers by coating spherical particles with DNA "feet" and sped up release from footholds using irreversible RNase H digestion, achieving speeds on the order of about 80 nm/s [211]; for comparison, this is 10% of the in vitro speed of conventional kinesins at 800 nm/s [92]. Impressively, electrical fields can be used to rotate a DNA robotic arm within milliseconds [114]. Other modes of top-down control have driven mechanical action at the nanoscale,

including light (hybridization/dissociation using azobenzene and its derivatives; [216]) and pH (structural change based in the i-motif; [55]).

As is the case in molecular biology where the lines between software and hardware are blurred, so it is in DNA nanotechnology where the nucleic acid components that process signals often also actuate responses. Beyond walking motion, dynamic DNA structures have accomplished other tasks at the molecular scale. Simmel and colleagues have used the aforementioned electric field-controlled DNA robot arm to modulate fluorescent signals in a computer-directed manner by moving a gold nanorod into and out of range of fluorophores immobilized on DNA origami [114]. Qian and colleagues demonstrated autonomous cargo sorting with a DNA walker that traveled via random walk along a DNA origami track, picking up oligonucleotide-labeled cargo and dropping them off at their corresponding goal locations [192]. DNA walkers have also found application in controllable plasmonic nanostructures. Zhou et al. have produced a large (35 x 10 nm) gold nanorod walker and "stator" pair that emits polarized light with distinct circular dichroic spectra as a result of coupling between the nanorods [228]. The relative position of the walker to the stator is mediated by strand displacement through the addition of strands. The precise, sequential gait of DNA walkers can provide a valuable method of control for multi-step organic synthesis. He and Liu have gone beyond DNA-templated organic synthesis strategies [121] and achieved a series of amine acylation reactions to form oligoamides with prescribed sequence through the motion of a DNA walker [86].

DNA origami seeks to bridge top-down fabrication techniques with bottom-up self-directed assembly by creating programmable and functional nanoscale tools. Among the challenges that DNA origami faces today include low yield, unstable larger-scale (i.e. beyond

micron) assemblies, lack of detailed analytical models of folding, high costs from synthesis of hundreds to thousands of custom oligonucleotides, and incompatibility with experimental conditions found in some applications. Attaining the promise of nanoscale control in a wide range of practical applications hinges on how well these hurdles can be addressed.

## 1.5 Concluding remarks

Given the exploration of DNA as a programmable material in the past few decades, we can now better identify in which applications nucleic acids excel. Even if Watson-Crick base pairing and predictive thermodynamic models afford researchers precise control over the structures and kinetics adopted by DNA or RNA in a given environment, nucleic acids inherently lack the chemical reactivity that is achievable with enzymes. In applications where higher reactivity is necessary (e.g. ligand-specific binding, catalysis), this can be partially remedied by modifications that expand the oligonucleotide alphabet to non-canonical or charged nucleobases [73, 43], but a more generalizable approach is to combine standard protein components (e.g. antibodies or well-characterized polymerases) as a part of rationally designed tools. For instance, oligonucleotide-tagged antibodies enable super-resolution microscopy on fixed samples by leveraging the kinetics of DNA probe binding in the technique known as DNA-PAINT [103]. Also, the inclusion of polymerases and nucleases as parts of synthetic biochemical networks can drive gene expression through the de novo production or degradation of RNA transcripts, as mentioned earlier. Another realization is that the design principles developed in DNA nanotechnology may pave the way for using novel, improved polymers to program chemistry into the future. Despite its biological relevance and the technologies that facilitate its use, DNA is not well-suited for all applications considered

in the field. It is possible that analogous molecules (e.g. synthetic polymers) may one day replace DNA in many of these technologies. Therefore, beyond its practical value in many applications, the broader value of DNA rational design comes from its ability to provide us with a first exploration of possible applications, where techniques for manipulating similar polymers may be matured.

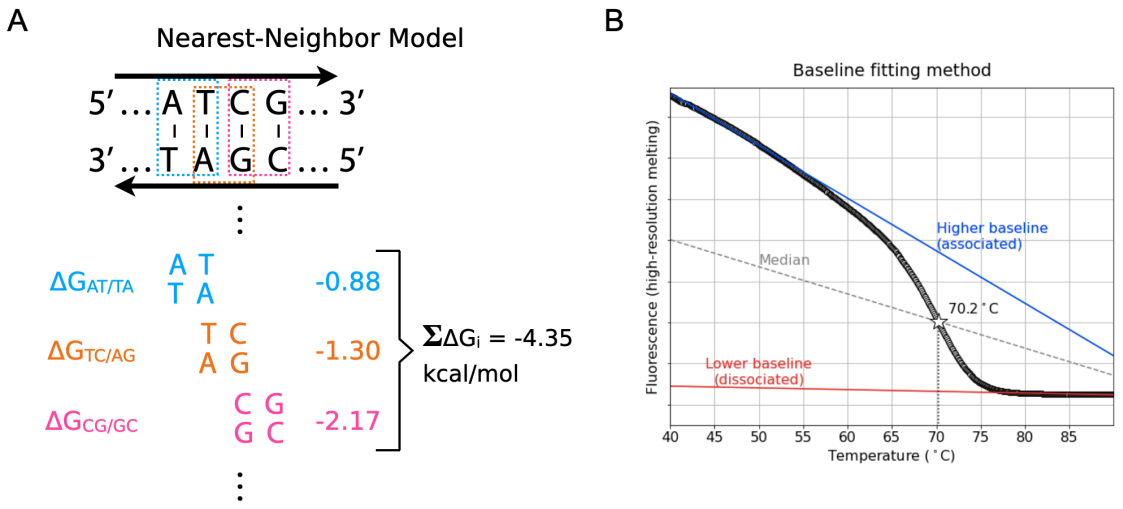


Figure 1.1: **Determination of thermodynamic parameters for predictive models of DNA hybridization.**

A. Using the nearest-neighbor model to predict duplex stability. The overall free energy of hybridization of the duplex is the sum of all nearest-neighbor parameters included in the duplex sequence. B. Baseline method for determination of overall duplex thermodynamics. Diagram is based on Figure 2B in Mergny and LaCroix 2003, data shown is collected using high-resolution melting.

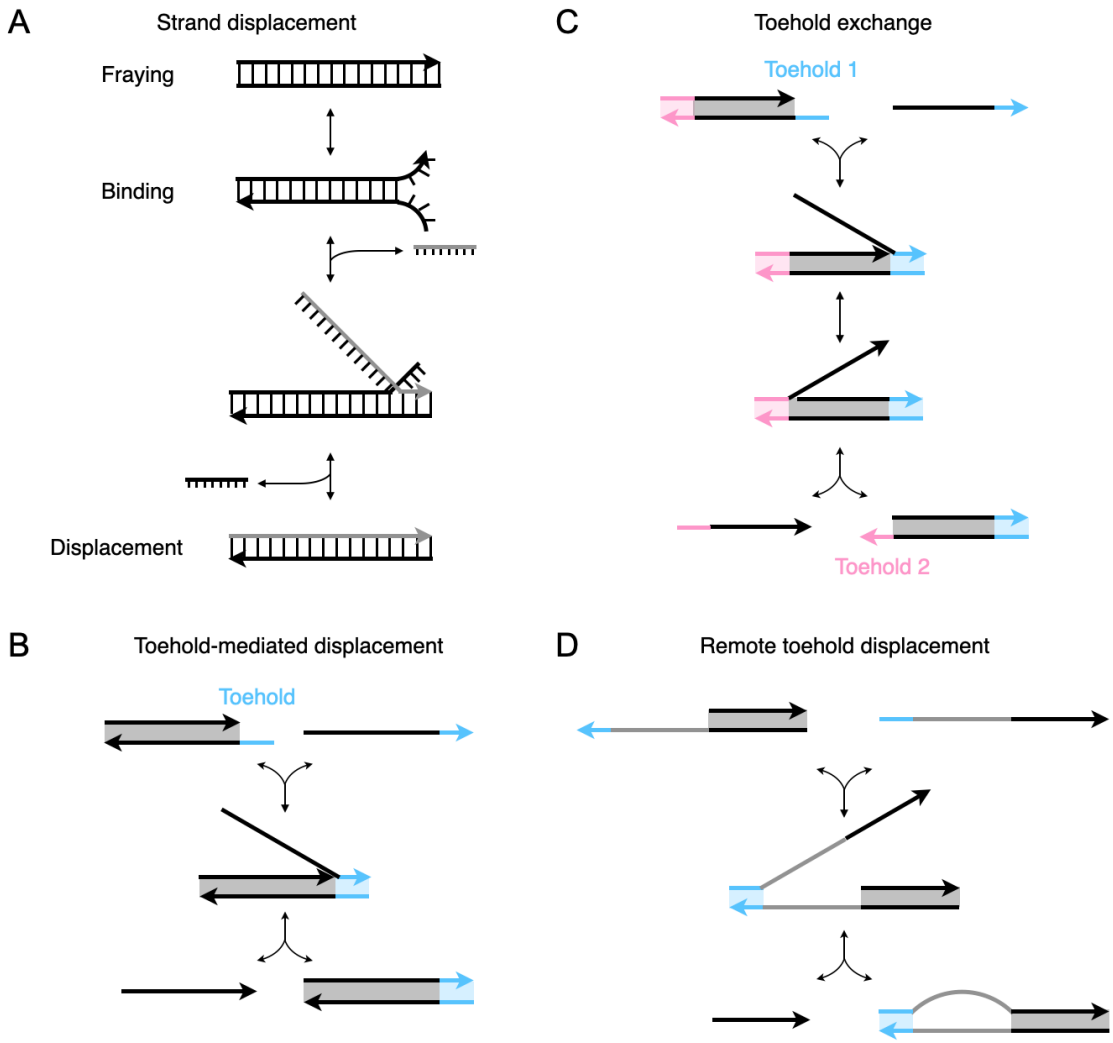


Figure 1.2: **Forms of strand displacement.**

A. Displacement due to end fraying. B. Toehold-mediated displacement. Domains with the same color have the same or complementary sequence; hybridized regions are represented by colored regions between strands. C. Toehold exchange. D. Remote toehold.

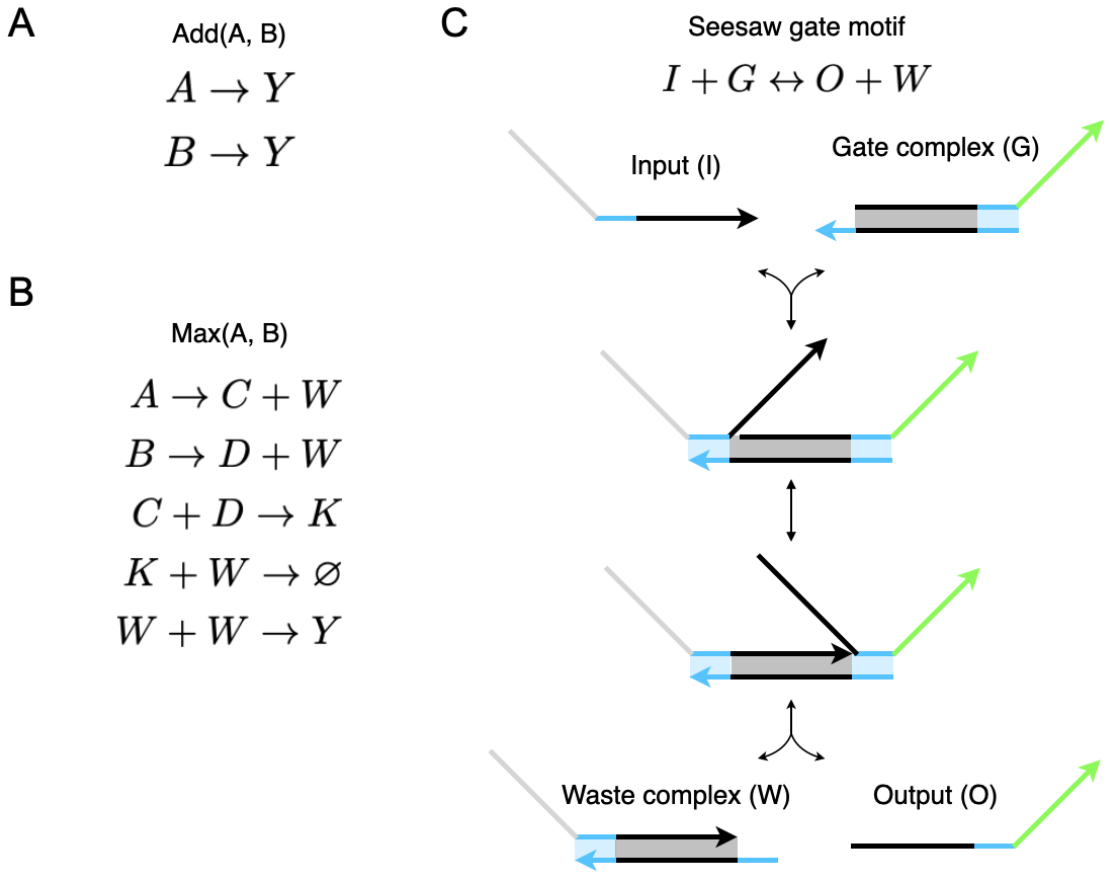


Figure 1.3: **Chemical computation: chemical reaction networks (CRNs) to strand-displacement implementation.**

A. CRN that encodes the addition function between species A and B, with species Y as output. B. A deterministic CRN that encodes the maximum function between species A and B, with species Y as output ([31]). C. Seesaw gate motif for implementing CRNs.

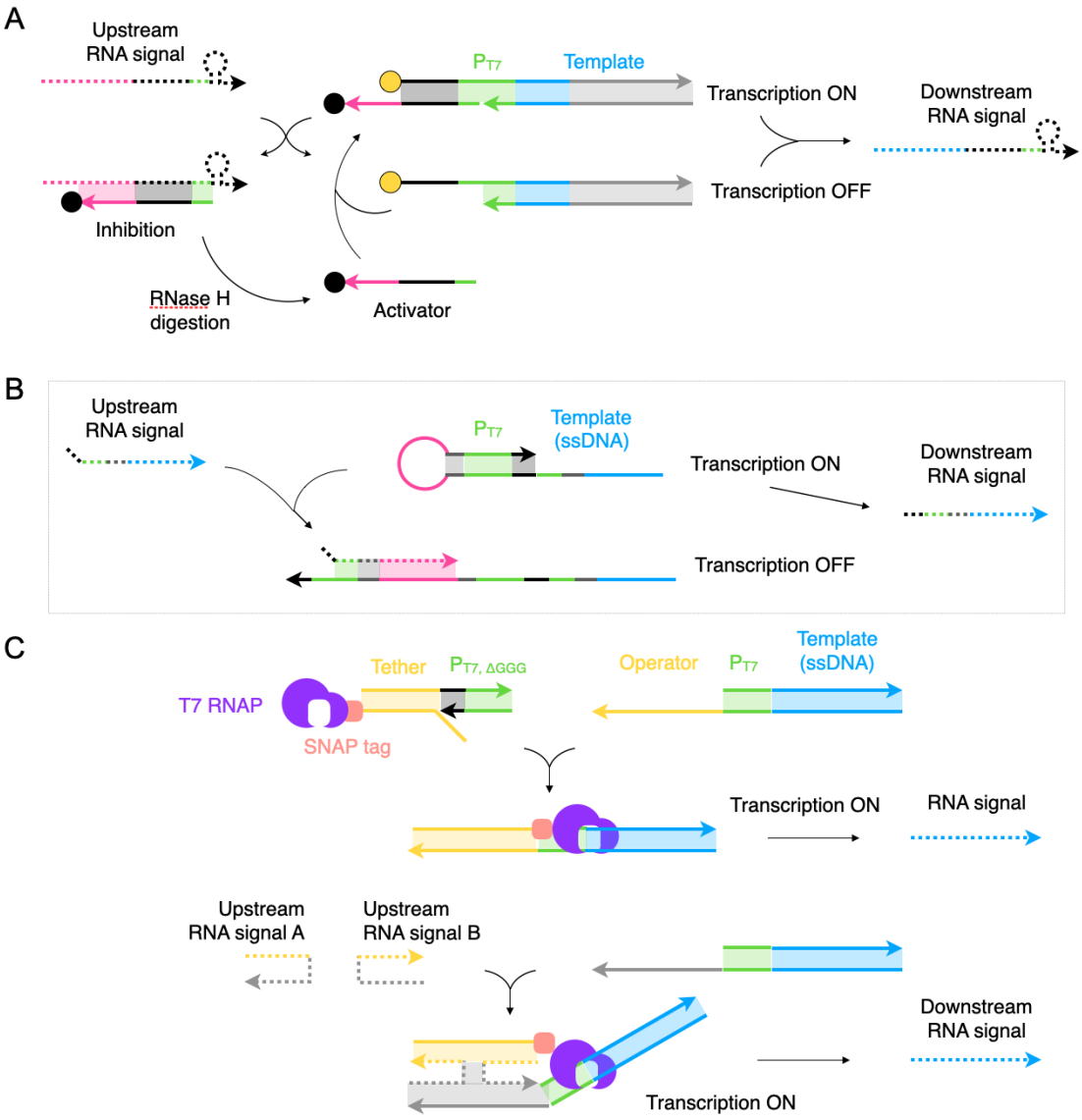


Figure 1.4: ***In vitro* transcription gate designs.**

A. Multi-stranded transcription gate by Kim et al. B. Single-stranded hairpin transcription gate by Kar and Ellington C. Tethered T7 RNAP design by Chou and Shih.

# Chapter 2

# Repurposing next-generation sequencing platforms for high-throughput profiling of DNA-based interactions[1]

**Abstract.** Next-generation sequencing (NGS) chips have been successfully repurposed as massively parallel platforms for mapping binding affinity of proteins of interest to libraries of DNA or RNA targets. We explored the possibility of using repurposed NGS platforms to map other biomolecular interactions - specifically, DNA-DNA hybridization and polymerase-promoter activity - at scale. We found that the size of Illumina MiSeq clusters limits the observable dynamic range to 4.24 kcal/mol in the best scenario, which is exceeded by many destabilizing nucleic acid hybridization motifs, including multiple mismatches, larger bulges, and loops. Using this platform, we were able to reproduce transcription activity rankings for T7 promoter variants as reported by previous solution-based assays.

[1]This chapter includes original work by SSW. SSW received guidance from Ilya Finkelstein and Andrew Ellington and funding for the project from Andrew Ellington. SSW would like to thank Stephen Jones, Jami Kuo, Jim Rybarski, and John Hawkins for helpful discussions and technical support.

## 2.1 Introduction

The thorough characterization of DNA and RNA hybridization has yielded sequence-dependent models for predicting secondary structure. This in turn has enabled the structural prediction of nucleic acid structures, including biologically potent RNAs (such as mRNAs, lncRNAs, miRNAs, crRNA) and rationally designed components engineered for DNA nanotechnology. Beyond the basic Watson-Crick base pairings, the thermodynamics of structural motifs such as bulges, loops, and dangling ends [164], as well as non-canonical interactions like wobble base pairing [70], inosine base pairing [208], mismatches [164], and even unnatural base pairing [91, 90] have been reported. Previous measurements of hybridization stability have either utilized the inherent hyperchromicity of DNA and RNA of the duplex to single-stranded transition (UV-Vis spectrophotometry) or the absorption of heat during melting (isothermal titration calorimetry and differential scanning calorimetry). These methods balance the advantage of producing absolute thermodynamic parameters through direct measurements with the disadvantages of being low-throughput and requiring large amounts of material. As a result, investigations of interactions that scale combinatorially with the features in question - for instance, multiple mismatches or mixed backbone modifications - would quickly become intractable at these smaller scales.

For this reason, alternative high-throughput methods would facilitate the collection of quantitative, predictive data, particularly for detailed models that include non-canonical nucleic acid-nucleic acid interactions. DNA microarrays, which contain ssDNA libraries on the order of thousands of variants, have been used to quantify the energetic impact and associated nearest-neighbor free energy parameters of mismatches in DNA-DNA hybridization [89]. Next-generation sequencing chips present a similar, larger-scale opportunity: each

sequencing chip flow cell can present up to tens of millions of unique DNA sequences which are spatially-addressable given positional data from its associated sequencing run. Previous works have indeed repurposed Illumina chip-based platforms to profile protein-DNA interactions, such as Gcn4-promoter affinity [137] or CRISPR-Cas target affinity [102], by assaying the affinity of a protein of interest to a DNA library of millions of sequences all at once. In fact, efforts have gone beyond mapping protein-DNA interactions to include RNA-protein interactions by using the DNA library as a template for *in vitro* transcription and producing a RNA library on the chip surface [20, 194].

Here, we investigated the feasibility of repurposing next-generation sequencing platforms for high-throughput profiling of DNA-DNA hybridizations and RNA polymerase-promoter activities. We expected that single-stranded probe-target binding would require a comparatively simple experimental setup and that enzyme catalytic activities may be likewise measurable if activity can be coupled to fluorescent signal intensity. We find that the specifications of the Illumina MiSeq platform in theory limit its range of detection to association reactions within a 4.24 kcal/mol range at optimal conditions, thus excluding significantly destabilizing DNA-DNA hybridization motifs such as multiple mismatches or larger loop (around length 6 or greater). Relevant RNA polymerase-promoter library interactions for the T7 RNA polymerase, however, can be successfully assayed by the platform, and transcriptional activities of promoter variants relative to the wildtype promoter sequence match reported values from *in vitro* studies in the literature. Our findings suggest that repurposed NGS flow cells can be valuable platforms for assaying protein-nucleic acid interactions at incredibly large scale, but are not suited for profiling equally large sequence spaces in nucleic acid-nucleic acid hybridization.

## 2.2 Results

### 2.2.1 Limitations of hybridization profiling on a repurposed sequencing platform

The Illumina MiSeq next-generation sequencing platform produces flow cells with tens of millions of spatially-addressable clonal clusters up to 600 base pairs in length and up to 1 micron in diameter [1]. Each cluster is produced by bridge amplification, which increases fluorescence signals up to 3 orders of magnitude above single-molecule techniques, by producing up to 1000 copies of the same sequence. After bridge amplification, clusters consist of double-stranded duplexes in which one strand is covalently attached to the slide surface and the other strand is its hybridized complement. Denaturation of the duplexes transforms the surface of the MiSeq chip into an microarray-like platform containing surface-immobilized single-stranded DNA strands ("targets") organized by sequence into localized individual clusters that bind a free-floating single-stranded probe.

Hooyberghs et al. have described the relationship between the fluorescence signal intensity observed from probe binding on a DNA microarray to the free energy of hybridization between the probe and target using the Langmuir model [89]:

$$I = A\theta_{\text{eq}} = \frac{Ace^{-\Delta G/RT}}{1 + ce^{-\Delta G/RT}}$$

where $I$, the observed intensity, is proportional to $\theta_{eq}$, the fraction of a cluster that is bound at equilibrium, by a constant scaling factor $A$; $c$ is the concentration of the probe; $\Delta G$ is the free energy of hybridization; $R$ is the gas constant; and $T$ is the temperature at equilibrium. Assuming that the system is far from chemical saturation at equilibrium (i.e.

only a small fraction of surface targets are hybridized, $ce^{-\Delta G/RT} \ll 1$ ), this approximates to

$$I = Ace^{-\Delta G/RT}$$

Solving for $\Delta G$ gives

$$\Delta G = -RT \ln\left(\frac{I}{Ac}\right)$$

The energetic penalty of a mismatch is defined as the difference in the probe's free energy of binding to a mismatched target ($\Delta G_{mm}$) compared to a perfectly matching target ($\Delta G_{pm}$) and can be experimentally observed as the ratio of their signal intensities (referred to from here as the "intensity ratio")

$$\Delta\Delta G = \Delta G_{mm} - \Delta G_{pm} = -RT \ln\left(\frac{I_{mm}}{Ac}\right) + RT \ln\left(\frac{I_{pm}}{Ac}\right) = RT \ln\left(\frac{I_{pm}}{I_{mm}}\right)$$

Illumina MiSeq clusters contain between one and several thousands of copies of a single sequence. This suggests that the largest signal intensity ratio possible is on the order of 1000 (i.e. in the mismatched target cluster, <10 targets are bound to the probe while in the perfect target cluster >1000 targets are bound). The detectable range of energetic penalties varies with the logarithm of this ratio; at a poor intensity ratio of 10, only interactions less than 1.41 kcal/mol from the perfect target-probe interaction can be measured, while at the best intensity ratio of 1000, the dynamic range is increased to a $\Delta\Delta G$ of 4.24 kcal/mol from perfect target-probe (Figure 2.1A). The $\Delta\Delta G$ of a single internal mismatch ranges from 1.86 to 5.97 kcal/mol in solution-based studies at 37°C [164]. Assuming it is possible to achieve the

highest intensity ratio of 1000, 140 out of 192 possible internal single mismatch triplets have energetic penalties that are within the theoretically detectable range at 37°C (Figure 2.1B). Beyond mismatches, some motifs involving imperfect hybridization have reported energetic penalties that fall within the detectable range; for example, loops (internal, bulge, or hairpin) can range from 2.9 to 6.6 kcal/mol at 37°C [164]. Increasing the incubation temperature can reduce the magnitude of energetic penalties, thereby fitting more interactions within the limit of detection, since the $\Delta G$ of duplex formation is negatively linear with respect to temperature; for instance, at an incubation temperature of 50°C, up to 160 single internal mismatches are within the best detection range (Figure 2.1B).

The value of repurposing NGS platforms lies in the ability to observe millions of interactions in parallel. This is particularly advantageous when the sequence space of interactions is very large, warranting such high-throughput methods. Given that the penalty of a single mismatch or other destabilizing motif already spans a large portion of the detectable range, however, most hybridization interactions involving more than one destabilizing motif are beyond the limit of detection on this platform. Further, contiguous mismatches are likely to form bulges and result in penalties greater than the additive impact from each single mismatch [164].

To experimentally confirm this limitation, we used a repurposed NGS platform adapted from the CHAMP platform [102] to measure the energetic penalties of mismatches. Similar to CHAMP, an Illumina MiSeq flow cell containing a library of target sequences (in our case, single and double mismatches as well as several negative controls; Figure 2.2A) was collected after a sequencing run and first denatured and washed to produce a single-stranded surface-bound library with minimal fluorescent background. Unlike CHAMP, however, we

did not regenerate the strand complements. Digoxigenin-labeled probe strand was added at 100 nM to the flow cell, and the cell was sealed and incubated at 50°C for 24 hours. Following incubation, bound probes were visualized using fluorophore-conjugated anti-digoxigenin antibodies and imaged by TIRF microscopy (Figure 2.2B). The measured intensities of the negative control targets were averaged and considered as the background signal. As expected, most single mismatches were distinguishable from the background and had observed intensities that correlated with the energetic penalty as predicted by parameters reported in the literature ($R^2 = 0.83$) [164]. The majority of double mismatches (excluding contiguous mismatches, for which thermodynamic parameters are unavailable) were compressed towards the lower end of the detection range, being only slightly above background.

### 2.2.2 Profiling T7 RNA polymerase transcription activity for a library of synthetic promoters

Despite their limitations in measuring nucleic acid-only interactions, repurposed NGS platforms have been used to map *in vitro* protein-nucleic acid affinities with great success, including protein-DNA [137, 102, 98] and protein-RNA interactions [20, 194]. We asked whether interactions beyond binding affinity, such as transcription activity, could be studied using this platform. We considered the T7 RNA polymerase, a single subunit DNA-dependent RNA polymerase that is highly specific in its recognition of its 17-nts promoter sequence [28, 180]. T7 RNAP is an invaluable tool in both molecular biology (e.g. for producing RNA transcripts *in vitro*) and synthetic biology (e.g. for driving gene expression for the production of proteins), where a library of promoter variants with known transcription activities could allow transcription to be quickly tuned across a range of expression levels,

which is particularly helpful for expressing toxic proteins or building multi-gene circuits. Because it contains only a single subunit, T7 RNAP is both easily used and easily studied. Highly related RNAPs (such as those from T3 or SP6) are similarly specific for their respective cognate promoter sequences, and orthogonal pairs of T7 RNAP-promoter interactions have been engineered [190, 133]; this suggests that the landscape of polymerase-promoter interactions is non-convex, and an exploration of the promoter sequence space may reveal variants not observed in enrichment-based selection assays.

We randomized the -12 to -7 region of the T7 promoter to produce a 6N promoter library (Figure 2.3A). The specificity loop of T7 RNAP recognizes the promoter through contacts between residues of its specificity loop and the -12 to -8 region promoter region [30]. Mutations to the promoter sequence at this region alters promoter recognition and consequently impacts transcription activity [157]. Our library fully covered all 4096 possible variants, with nearly every variant represented in 100 or more reads; in addition, as negative controls for transcription, a variant with a scrambled sequence in place of the T7 promoter was added. To assay protein activity as a function of RNA produced, a TerB sequence (Tus protein binding site) was inserted between the P7 and SP2a adapter sequences, and a template for the MS2 hairpin sequence was included. Once Tus protein is added to the flow cell and bound to the TerB site, transcribing T7 RNAPs stall at the Tus-TerB location, which in turn prevents the release of the elongating strand and ensures that transcripts from each promoter variant are location addressable (Figure 2.3B) [20, 194]. Fluorescently-labeled MS2 coat protein is added to visualize the amount of RNA corresponding to a cluster through association with transcribed tethered MS2 RNA hairpins, with signal intensity as a measure of overall transcriptional activity.

The highest measured intensity was observed in the wildtype promoter sequence (CGACTC in the randomized region) (Figure 2.3C). Weighing each variant by its average measured intensity and finding a "weighted consensus" reveals the wildtype sequence (Figure 2.3D). To assess the accuracy of the repurposed NGS chip platform for profiling in vitro transcription activity for various promoter sequences, we compared the relative transcription activities observed on our platform to those reported in three previous studies that measured transcription activity using either solution-based [154] or NGS-based methods [147, 113]. Between our study and each of the three literature studies, our measured transcription activities correlated well with the literature-reported activities for mutually-included variants (R2 of 0.84 or higher) (Figure 2.4). Two higher-activity promoter variants were common between our study and all three previous studies, and four higher-activity variants were common to our study and two previous studies (Figure 2.4A). In all datasets considered, these variants ranked among the top below wildtype. The repurposed NGS platform was able to measure differential activity between the mutually-included variants, whereas the distribution of measured activities for these variants were below the detection limit for methods used in other studies (Figure 2.4B and C). These results suggest that the repurposed NGS chip platform can be used to measure the relative *in vitro* activities of promoter variants for T7 RNA polymerase at a large scale.

## 2.3 Discussion

We observed that NGS-based affinity mapping platforms are suited to the study of protein-DNA interactions for a larger sequence space than for DNA-DNA interactions. This suggests that, given the small range of detection, variations in protein-DNA complex stability

that result from DNA sequence mutation are generally smaller than variations in DNA duplex stability arising from similar mutations. In other words, DNA-protein interactions may tolerate mutations better than DNA-DNA hybridization. This is not surprising considering that DNA duplexes are stabilized by protein binding, and energetic penalties due to slightly imperfect contacts between residues and minorly mutated dsDNA are relatively small in comparison to the stability of the protein-DNA complex. Specifically in the case of T7 RNAP initiation, while the thermodynamic impacts of specific promoter mutations have not been documented, the scale of interaction energies involved for transcription may shed some light on the impact of promoter mutations. Melting the initiation region of the duplex promoter requires unwinding of the helix and bending the single promoter strands and thus takes considerable energy; thus, it is estimated that extensive interactions between the polymerase and single-stranded portions of the promoter generate up to 68 kcal/mol that are used towards melting the promoter [214]. Part of this energy may come from polymerase binding, which is estimated at -13.3 kcal/mol for binding to the wildtype [189]. A mutated specificity region (-12 to -8) is unlikely to change the energy required to melt the initiation region (-4 to +2) to form the transcription bubble, suggesting that if T7 RNAP is capable of association with a mutated sequence, it will likely be capable of initiation.

The quantity measured with our platform, transcription activity, is a combination of multiple rates, including association, dissociation, initiation, promoter clearance, abortive cycling, elongation, and processivity [30, 214, 215, 54]. An advantage of surface-based approaches over solution-based methods is that with surface-based approaches it may be possible to individually study some of these factors. For example, association and dissociation may be measured as the amount of fluorescently-labeled polymerase bound, initiation by the

incorporation of fluorescently-labeled NTPs with one NTP depleted, and elongation by loading a single RNAP per strand for single turnover and measuring MS2 hairpins transcribed from a template of concatenated MS2 sequences.

Although the activities we measured matched reported activities from solution-based methods, it is still valuable to validate the findings using other experimental modalities, particularly for novel characterizations, since steric hindrance effects due to crowding at the flow cell surface may affect binding and activity. In vivo studies should be performed to confirm whether the interaction is robust in a cellular context. Chip-based high-throughput profiling techniques will best benefit protein-nucleic acid profiling tasks that require a large nucleic acid sequence space. Towards the future, it may be possible to use surface-addressable DNA (or RNA) clusters to map the affinities of a library of nucleoprotein variants to singular DNA or RNA probes, or even to express DNA sequences as peptide libraries for mapping protein-peptide or target-peptide affinity. This latter application could complement in vivo eukaryotic solution-based display platforms for high-throughput protein-target interaction assays such as yeast surface display [72] or mammalian cell display [95, 179] by providing biochemical insight on individual variants through large scale *in vitro* binding assays for titrations of target molecules.

## 2.4 Materials and Methods

**Oligonucleotides and reagents.** The target library for DNA-DNA mismatch hybridization was purchased as a custom library from CustomArray. All other oligonucleotides, including the T7 promoter library, fluorescent probes, digoxigenin probes, and primers, were

purchased as custom oligonucleotides from IDT. Next-generation sequencing was performed on the Illumina MiSeq platform using either a 2x75, 2x150, or 2x250 paired end reagent kit. Unless otherwise stated, all chemical reagents were purchased from Sigma Aldrich and all buffers and enzymes were purchased from NEB.

**Library preparation for next-generation sequencing.** The custom oligonucleotide T7 promoter library was PCR amplified using Q5 High-Fidelity DNA polymerase (NEB, M0491S) in 1x Q5 Reaction Buffer (NEB, B9027S) with a final concentration of 200 uM of each dNTP (ThermoFisher, R0181) and 400 nM of the forward and reverse primers on a PCR thermocycler with the following protocol: 3 min initial melting at 98C, followed by 10 cycles of 30 sec melting at 98C, 30 sec annealing at 67C, and 30 sec extension at 72C, followed by a 3 min final extension at 72C. After amplification, the PCR products were loaded onto a 1.2% agarose gel (NuSieve GTG, Lonza BioScience) and gel purified using a QIAquick Gel Extraction Kit (Qiagen) following manufacturer's instructions with the following exception: gel fragments were incubated for at least 20 minutes at 60C in Buffer QG and the DNA product was washed 3x with Buffer PE prior to elution in nuclease-free water. The finished library was then submitted to the UT Genome Sequencing Analysis Facility (GSAF) for next-generation sequencing. Because both the DNA-DNA mismatch and T7 promoter libraries both have low base diversity, to ensure that the final sequencing chip does not run into downstream analysis issues due to base diversity, an additional sample library prepared from HeLa genomic DNA (NEB, N4006) was added to represent approximately 50% of the final reads of all runs.

**DNA-DNA hybridization with CHAMP.** The target probe sequence was designed to have a GC-content close to 50%, a $T_m$ between 40C and 50C, and have a G:C

pairing on both 5' and 3' ends. Target sequences and subsets of target sequences included in the DNA library were assessed to find appropriate targets.

TIRF microscopy and microfluidics setup for CHAMP can be found in [102]. Unless otherwise specified, all washing and loading steps were performed at a flow rate of 100 $\mu$l/min. Following NGS, the physical chip was washed with an initial denaturing solution of 0.1 M NaOH (300 $\mu$l) and 1X TE (300 $\mu$l). Resetting the chip between experiments consisted of denaturation and removal of complementary strands using a 5 minute incubation in 60% DMSO [203] and a 300 $\mu$l wash in 1X NGS Wash Buffer (0.3X SSC, 0.1% Tween-20), followed by a proteinase K treatment consisting of incubating the channel for 45 minute at 42C in a 2 mg/ml RNA grade proteinase K solution (ThermoFisher, 25530049) in 1X TE and a 500 $\mu$l wash in 1X NGS Wash Buffer.

To perform hybridization experiments, the chip and its stage adapter were placed into a plate incubator with the ends of the tubing adapter sealed to prevent evaporation. Once the chip was equilibrated to the correct temperature, it was washed with 500 $\mu$l 1X NNE Buffer (0.5 M NaCl, 10 mM Na2HPO4, 1 mM EDTA, pH 7.0), after which probe mix (100 nM Target-Dig probe; 1X NNE buffer; 0.005% BSA, ThermoFisher; 1 $\mu$g/ml salmon sperm DNA, FisherScientific) heated to the appropriate temperature was loaded onto the chip and incubated. After the incubation period, the channel was washed with 1X NNE, the chip and attached stage adapter were removed from the incubator and onto the TIRF microscope. Primary (Anti-Dig rabbit, Invitrogen, 9H27L19) and secondary (Anti-Rabbit goat 647N, Sigma 40839) antibodies were each sequentially loaded onto the chip, incubated for 10 minutes at room temperature, and unbound antibodies were washed from the chip with a 5 minute NGS Wash Buffer flow step. The chip was illuminated with a 633 nm laser

(Ultralasers) at 100 mW and imaged at room temperature.

**Protein purification.** *Preparation of Flag-Tus.* The 6xHis-StrepTag-SUMO-Flag-Tus coding sequence was previously cloned into pET-19 plasmid by members of the Finkelstein Lab. The purified plasmid was sequence verified by Sanger and transformed into chemically competent BL21 (DE3) cells, and plated on LB agar with carbenicillin. An overnight culture was prepared in 1X LB media at 37C. The following day, subcultures were prepared using a 1:100 dilution of the overnight culture into 1X Superior Broth (SB) media containing antibiotics, shaken at 250 rpm at 37C in Erlenmeyer flasks, and grown until reaching OD600 of 0.6. The protein was then induced using IPTG (0.1 M IPTG final concentration in culture) and shaken at 37C for 4 hours. A 2 ml aliquot of the induced culture was taken and miniprepped (Qiagen Miniprep Kit, manufacturer's instructions) and Sanger sequenced to confirm the presence of the plasmid and to check that no deleterious mutations to the protein of interest had occurred. Cells were pelleted and frozen at -80C until use.

To purify the protein, frozen pellets were resuspended in Lysis Buffer, which consists of 50 mM sodium phosphate pH 7.5, 100 mM NaCl, 1 mM EDTA, 10% glycerol, 0.2 mg/ml lysozyme, 1 mM DTT, and 1 tablet cOmplete Protease Inhibitor Cocktail (Millipore Sigma) per 50 ml of buffer. Cells were lysed by sonication (Fisher Scientific Sonic Dismembrator, Amplitude = 75, total processing time = 1:30, pulse ON = 0:15, pulse OFF = 0:45) and ultracentrifuged at 35k rpm for 40 minutes at 4C. The resulting supernatant was purified by affinity chromatography in a Strep-Tactin (Iba Life Sciences) gravity column with 3 ml total resin volume. The column was equilibrated with 20 column volumes (CVs) of Lysis Buffer, after which the clarified supernatant was applied to the column. Bound protein was washed using 20x CV Wash Buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA,

1 mM DTT, 20% glycerol), eluted in 20 mM Elution Buffer (2.5 mM d-desthiobiotin, 50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA, 2 mM DTT, 20% glycerol), manually fractionated and concentrated with an Amicon Ultra-15 centrifugal unit (Millipore Sigma) to approximately 1 ml. Protein tags were cleaved using SUMO protease (purified by members of Finkelstein Lab) at approximately 3 uM final concentration in a rotator overnight at 4C. Cleaved Flag-Tus protein was isolated using a HiLoad 16/600 Superdex 200 pg size exclusion column (SEC) (Cytiva). The sample was quantified by SDS-PAGE and stored in SEC Buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM DTT, 10% glycerol) in 10 ul aliquots at -80C until use.

*Preparation of MCP-488.* The 6xHis-SUMO-MCP-SNAPf coding sequence was previously cloned into pET-19 plasmid by members of the Finkelstein Lab. Preparation of the pellet was identical to the preparation of Flag-Tus.

Procedures for pellet resuspension, sonication, ultracentrifugation, and gravity column purification were as previously described for Flag-Tus, except for the compositions of Lysis Buffer (50 mM HEPES pH 7.4, 500 mM NaCl, 1 mM EDTA, 0.1% Tween-20, 5% glycerol, 1 mg/ml lysozyme, 0.0025 U/$\mu$l DNase I, 1 mM DTT, and 1 tablet cOmplete Protease Inhibitor Cocktail per 50 ml buffer), Gravity Column Wash Buffer (50 mM HEPES pH 7.4, 500 mM NaCl, 1 mM EDTA, 5% glycerol), and Gravity Column Elution Buffer (10 mM d-desthiobiotin, 50 mM HEPES pH 7.4, 500 mM NaCl, 1 mM DTT, 10% glycerol). Following Strep-Tactin purification, the eluted sample was concentrated, and both SUMO protease and SNAP-Surface 488 (NEB, S9124S) was added to the concentrated sample which was gently agitated overnight at 4C covered by foil. The following day the cleaved and labeled sample was purified using the SEC and washed with SEC Wash Buffer (50 mM Tris-HCl pH

7.5, 500 mM NaCl, 2 mM DTT, 10% glycerol). Protein concentration was quantified and dye labeling was confirmed through SDS-PAGE. Samples were stored in 20 ul aliquots at -80C until use.

***In vitro* transcription assays on CHAMP.** TIRF microscopy and microfluidics setup for CHAMP can be found in [102]. Unless otherwise specified, all washing and loading steps were performed at a flow rate of 100 $\mu$l/min. Because a TerB sequence is inserted immediately following the SP2a adapter, the first 6 nts of the TerB sequence were used as the i7 index for NGS (AATTAG). Following NGS, the physical chip was washed with an initial denaturing solution of 0.1 M NaOH (300 $\mu$l) and 1X TE (300 $\mu$l). To regenerate complements to produce a double-stranded promoter library, a primer mix containing 500 $\mu$M each P7 primer and SP2b-complementary PhiX-digoxigenin primer in 1X Hybridization Buffer (5X SSC, 0.1% Tween) was first loaded into the chip and incubated on the heat block with the following protocol: 5 minutes at 85C, 30 minutes ramp down from 85C to 60C, 10 minutes ramp down from 60C to 40C, and 10 minutes at 40C with simultaneous 1X Wash Buffer flow (0.3X SSC, 0.1% Tween). After primer annealing PCR mix (0.1 U/$\mu$l Klenow Fragment exo- DNA polymerase, M0212L; 25 uM each dNTP, ThermoFisher; and 1X NEB Buffer 2) was loaded into the chip, incubated 30 minutes at 37C. To test alignment for the chip, primary antibody (Anti-Dig rabbit, Invitrogen, 9H27L19) and secondary antibody (Anti-rabbit Atto 488 Goat, Invitrogen, A-11008) was each sequentially added at 1 $\mu$g/ml, incubated in the chip for 10 minutes at room temperature, and washed with 1X NGS Wash Buffer for 5 minutes, after which the chip was imaged in the blue channel at 10 mM. The chip was then treated with Proteinase K as described earlier. The flow system was then switched to 1X Running Buffer (40 mM Tris-HCl pH 7.5, 150 mM NaCl, 6 mM MgCl2, 1

mM DTT, 0.1% Tween-20, 0.2 mg/ml BSA) and kept at 37C. The chip was washed for 5 minutes.

To assay *in vitro* transcription activity for the promoter library, Flag-Tus mix (500 nM Flag-Tus in 1X Running buffer) was loaded into the chip and incubated for 30 minutes. After washing with 1X RNAPol Reaction Buffer (B9012S), the T7 transcription mix (2 U/$\mu$l T7 RNA polymerase, M0251S; 200 uM each NTP; 5 mM DTT; 1X RNAPol Reaction Buffer) was loaded and incubated for 30 minutes, followed by a 5 minute wash with 1X Running Buffer. MCP mix was loaded (500 nM MCP-488 in 1X Running Buffer) and incubated for 30 minutes. The chip was then washed with 150 $\mu$l of 1X Running Buffer at 50 $\mu$l/min, illuminated at 10 mW with a 488 nm laser (Coherent), and imaged at room temperature.

**Data analysis.** TIRF images were aligned to positional and sequence data included in the fastq files associated with the sequencing run using the alignment code developed for the CHAMP platform [102] (github.com/hawkjo/champ). Literature reported dG values for mismatched DNA-DNA hybridization was calculated from nearest-neighbor parameters reported in [161] for matching base pairs and [164] for mismatched base pairs and adjusted to the experimental sodium concentration.

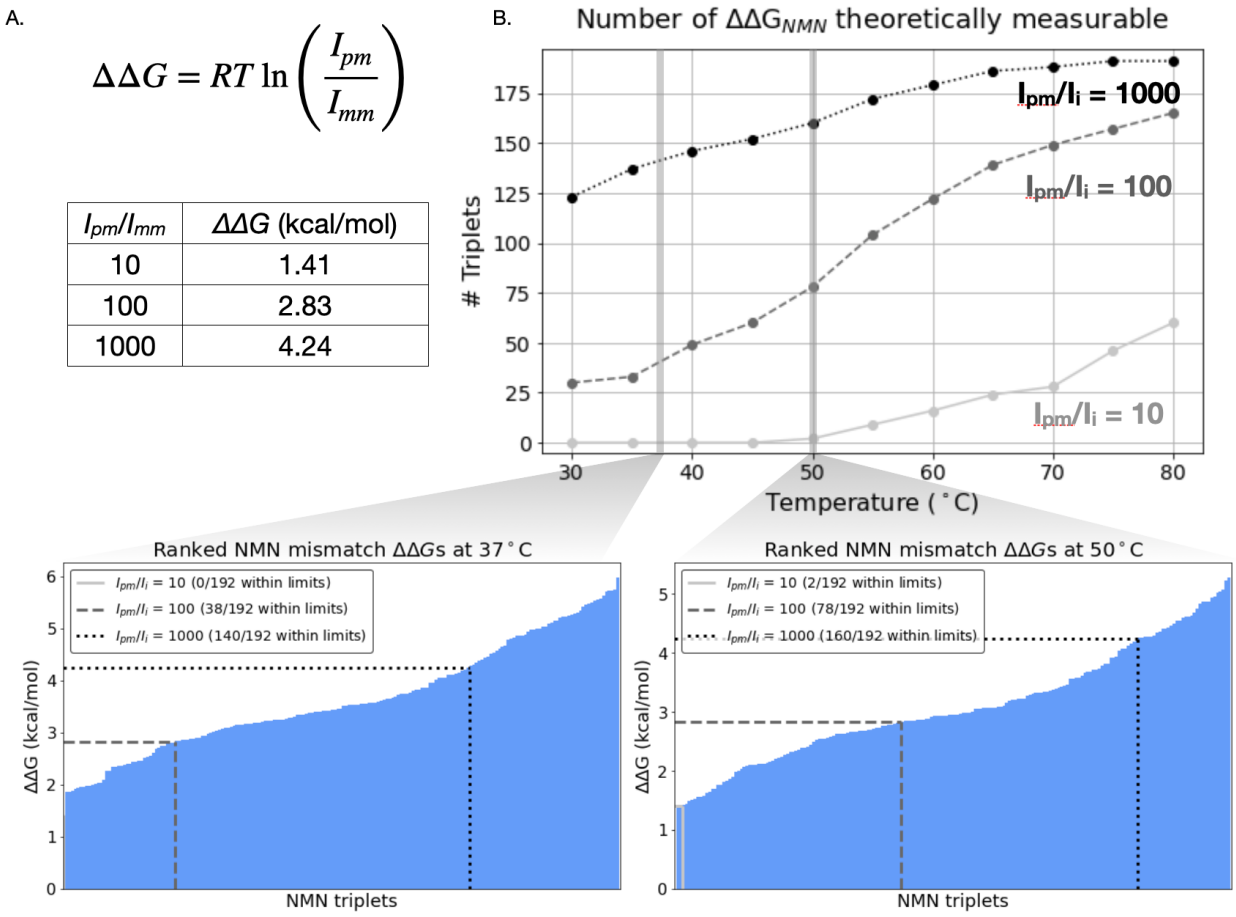


Figure 2.1: **Theoretical limits of profiling with the repurposed Illumina MiSeq flow cell.**

A. Dynamic range of energetic penalties relative to perfect match (pm) that are detectable for different ratios of perfect to mismatched (mm) observed intensities for an incubation temperature of 37°C. B. Number of single mismatch triplets with energetic penalties within the detectable range for an intensity ratio of 10, 100, and 1000. Lower plots show slices at 37°C and 50°C of all 192 single mismatch triplets arranged in increasing order of energetic penalty. Fractions show the number of triplets within the detectable range at a given intensity ratio; at 37°C the fraction at a ratio of 10 is omitted as no triplets are theoretically detectable. All values are calculated at [Na+] = 1 M; note that different sodium concentrations will not change the energy penalty [161].

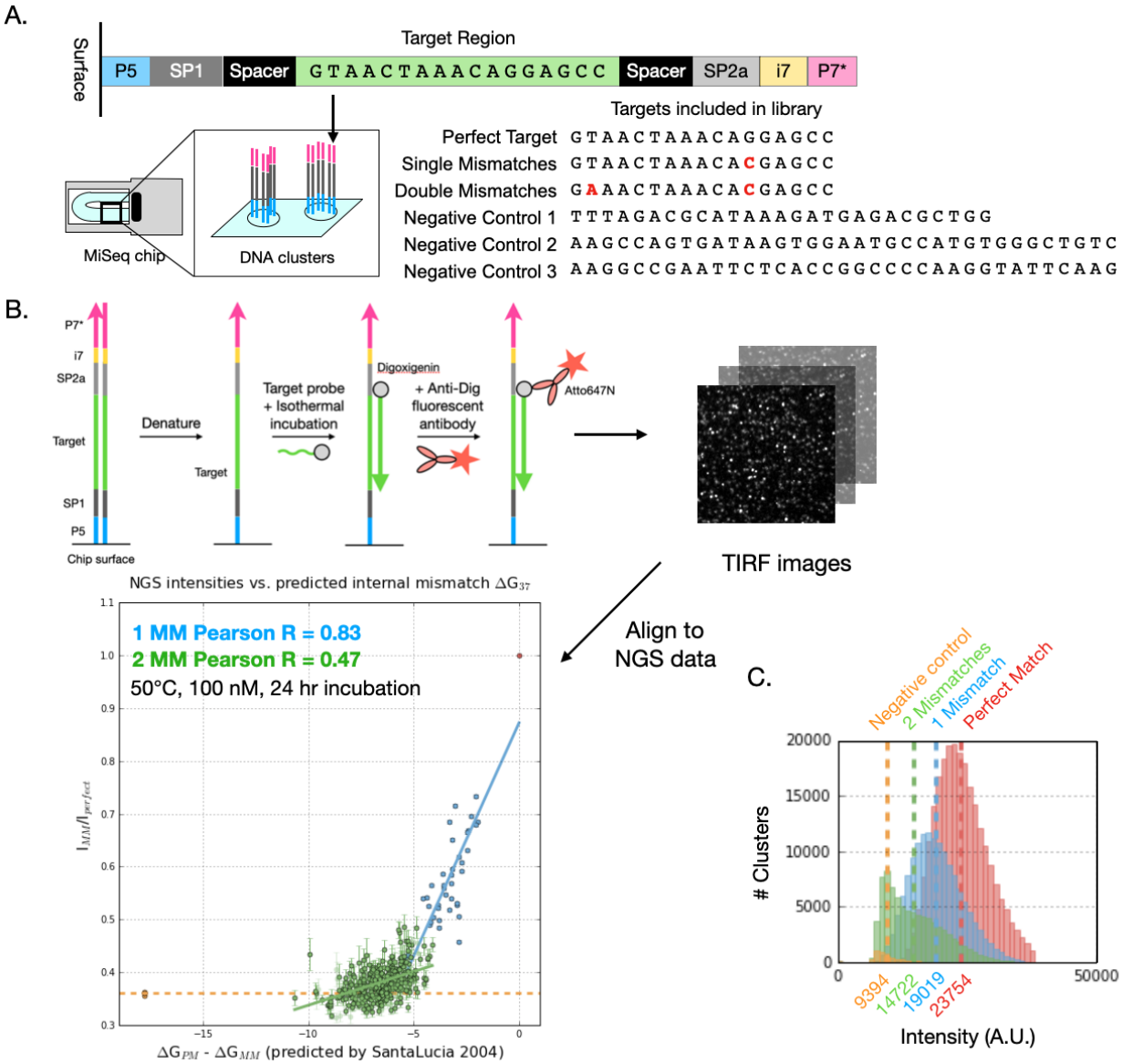


Figure 2.2: **Experimental profiling of mismatched DNA-DNA hybridization interactions on a repurposed MiSeq platform.**

A. NGS library design and types of mismatches considered. Contiguous mismatches were excluded from the library. B. Workflow with CHAMP for profiling mismatched hybridization and measured energetic penalties due to mismatches using the CHAMP platform. C. Experimentally observed typical mean signals for each type of target dynamic range.

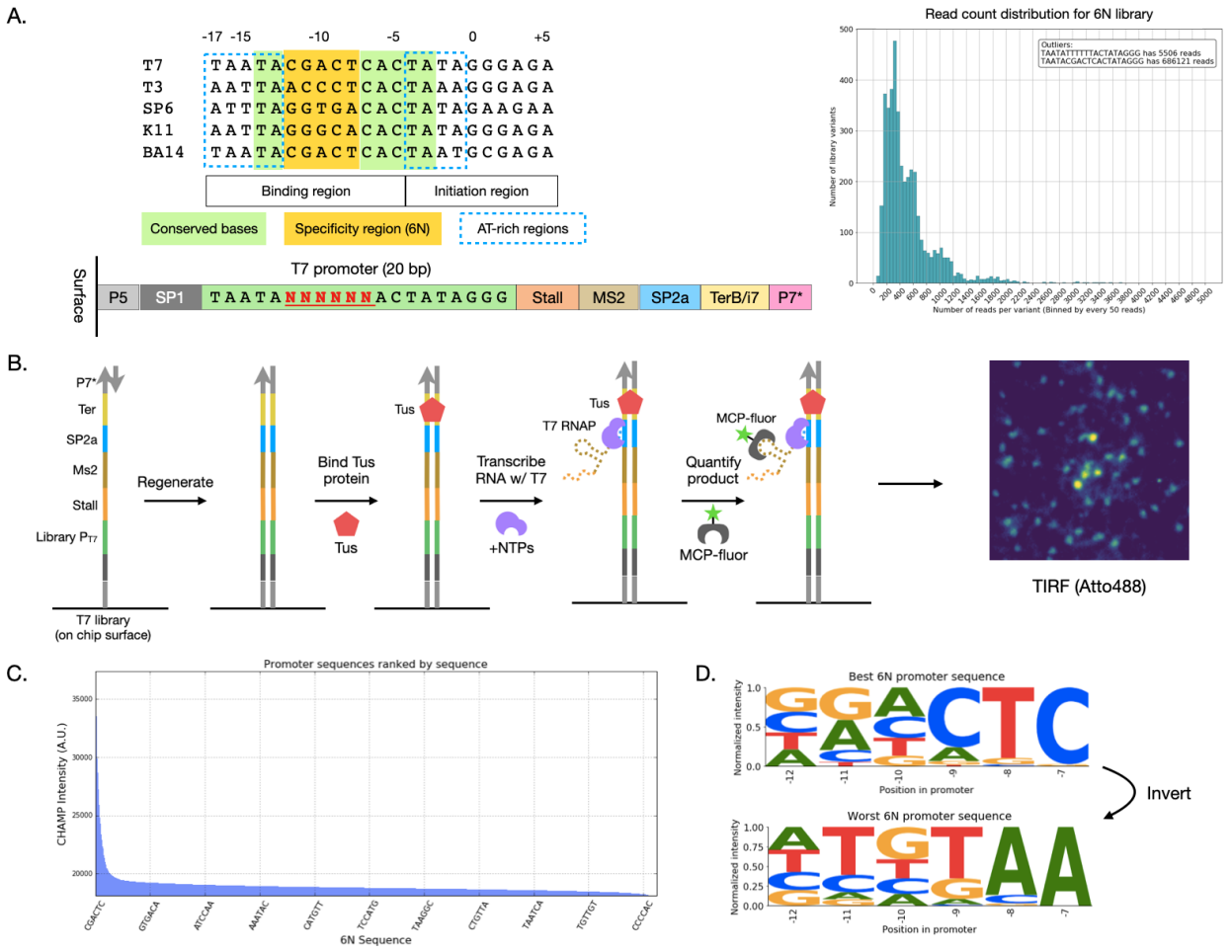


Figure 2.3: **Experimental profiling of T7 RNA polymerase transcription activity as a function of polymerase-promoter interactions using CHAMP.**

A. Breakdown of single-subunit polymerase family phage promoter sequences and NGS chip library design B. Experimental protocol and setup for T7 RNAP *in vitro* transcription on the CHAMP platform. C. Transcription activity as a function of promoter sequence measured by background-subtracted intensity. D. Specificity sequence weighted by transcription activity.

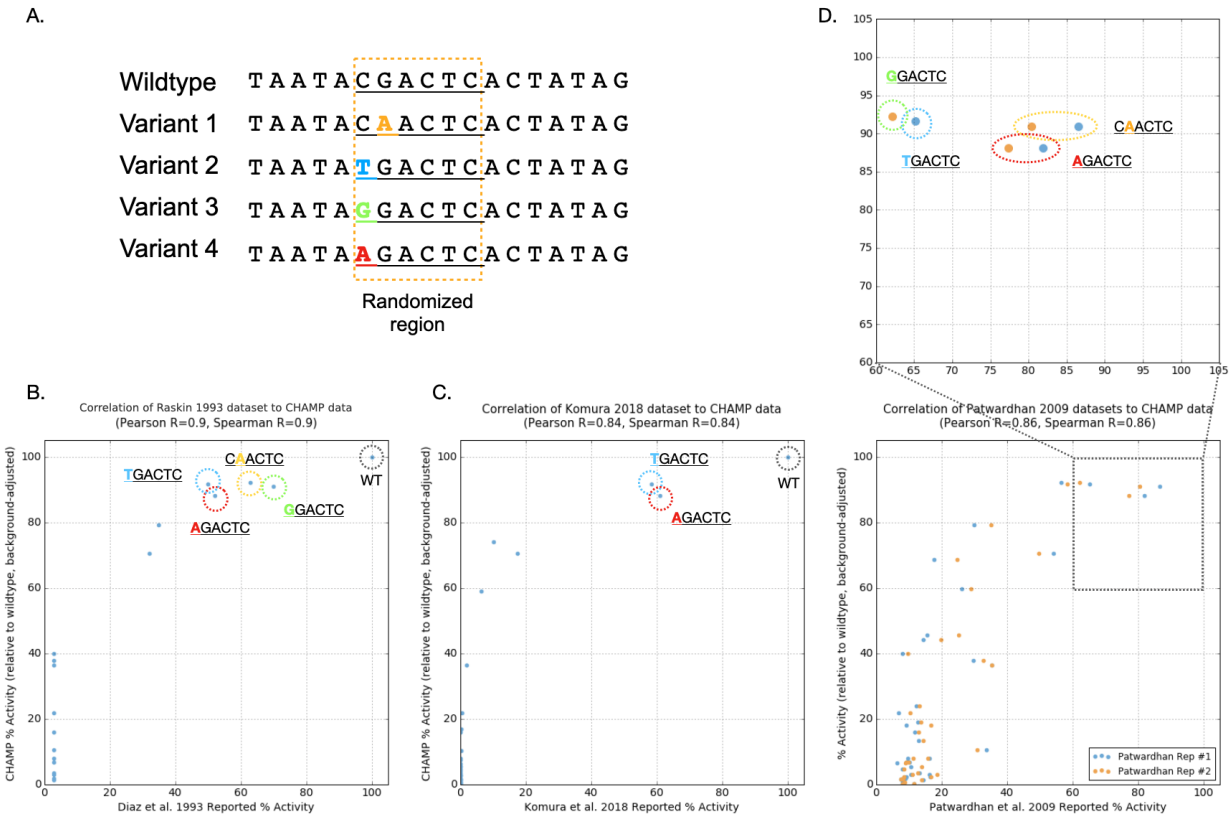


Figure 2.4: **Correlation of promoter variant relative activities to wildtype as measured using CHAMP to reported relative activity from the literature.**

A. Selected variants with the highest observed activity as determined by CHAMP. B. Correlation to Raskin et al. PNAS 1993 [154]. C. Correlation to Komura et al. PLOS ONE 2018 [113]. This dataset does not contain variants 1 (CAACTC) and 3 (GGACTC) D. Correlation to Patwardhan et al. Nature Biotechnology 2009 [147].

# Chapter 3

# Developing predictive hybridization models for phosphorothioated oligonucleotides using high-resolution melting[1]

**Abstract.** The ability to predict nucleic acid hybridization energies has been greatly enabling for many applications, but predictive models require painstaking experimentation, which may limit expansion to non-natural nucleic acid analogues and chemistries. We have assessed the utility of dye-based, high-resolution melting (HRM) as an alternative to UV-Vis determinations of hyperchromicity in order to more quickly acquire parameters for duplex stability prediction. The HRM-derived model for phosphodiester (PO) DNA can make comparable predictions to previously established models. Using HRM, it proved possible to develop predictive models for DNA duplexes containing phosphorothioate (PS) linkages, and we found

[1]This chapter is adapted from a manuscript by Wang SS, Xiong E, Bhadra S, and Ellington AD (2022). SSW and EX shared first authorship. SSW and EX conceived the project and performed the melting experiments. SSW designed the sequences and performed all analyses. EX developed the protocol and performed the CHA experiments. SB and ADE provided mentorship. ADE provided funding. SSW and ADE wrote the manuscript.

that hybridization stability could be predicted as a function of sequence and backbone composition for a variety of duplexes, including PS:PS, PS:PO, and partially modified backbones. Individual phosphorothioate modifications destabilize helices by around 0.12 kcal/mol on average. Finally, we applied these models to the design of a catalytic hairpin assembly circuit, an enzyme-free amplification method used for nucleic acid-based molecular detection. Changes in PS circuit behavior were consistent with model predictions, further supporting the addition of HRM modeling and parameters for PS oligonucleotides to the rational design of nucleic acid hybridization.

## 3.1 Introduction

The programmability of nucleic acids for biotechnology and nanotechnology applications is based on the highly predictive thermodynamic properties of DNA and RNA hybridization, which can be well-approximated by the nearest-neighbor model [41, 48, 193, 51]. In consequence, the stability of a given duplex can generally be accurately predicted from its sequence [161, 162, 62, 199, 183, 78, 182]. Typically, nearest-neighbor model parameters for nucleic acids are derived using UV-Vis spectrophotometry, relying on the hyperchromicity of single-stranded DNA and RNA to capture the transition from duplex to denatured strands, and consequently fit melting temperatures and other thermodynamic values pertaining to the duplex. While such hyperchromicity methods can produce thermodynamic parameters that are broadly applicable to various predictions because they result from direct measurements of duplex melting, the material cost and low throughput of UV-Vis spectrophotometry can be prohibitive, particularly in the case of expensive or precious non-canonical oligonucleotides.

As a result, while nearest-neighbor parameters have been found for some non-canonical bases [91] and unnatural backbones [107, 25], many other broadly employed chemical modifications to DNA and RNA have yet to be similarly adapted to predictive models. The rational design of nucleic acid hybridization for both structure and function is therefore generally limited to the use of unmodified oligonucleotides.

High-resolution melting (HRM) represents a higher throughput and more cost-efficient method for quantifying duplex stability and consequently deriving predictive parameters. In this method, sequence non-specific intercalating dyes such as EvaGreen or LC Green obviate the need for custom fluorescent probes or fluorophore-quencher modifications, and can be carried out in 96-well plate formats with volumes on the order of 10 $\mu$l and as little as pmoles of material. HRM has been widely employed in molecular diagnostics to rapidly discriminate between near-identical sequences through shifts in melting temperatures, and has enabled applications such as single-nucleotide polymorphism genotyping and quantification of mosaicism [197, 58].

In this study, we assessed the feasibility of HRM as a method for determining the sequence-dependent thermodynamic parameters for phosphorothioated (PS) oligonucleotides. We designed sets of phosphodiester (PO) DNA oligonucleotide duplexes with sequences that maximally spanned the space of nearest-neighbor nucleotide pair parameters and determined the $T_m$ of each duplex at various concentrations using HRM with EvaGreen intercalating dye. We fitted transition thermodynamic parameter enthalpy ($\Delta H$), entropy ($\Delta S$), and free energy ($\Delta G$) to the collected $T_m$ values using Van't Hoff analysis and then derived approximate nearest-neighbor parameters using singular value decomposition. While a potential drawback to using HRM to characterize nucleic acid duplex thermodynamics is the introduction of sys-

tematic errors due to binding interactions with intercalating dyes, we find that it is possible to apply a linear correction to HRM-derived model predictions (i.e. $\Delta G_{\text{Adjusted-37}} = \Delta G_{\text{HRM-37}}$ - 3.73 kcal/mol + 0.19 kcal/mol/base pair $\times$ sequence length) and thereby generate predictions comparable to those made by models derived from hyperchromicity data. Using HRM, predictive models for DNA duplexes containing PS modifications were fitted, PS modifications were incorporated into a DNA-based amplification circuit and changes to circuit behavior that corresponded to predictions were observed. HRM methods can therefore potentially accelerate the use of nucleic acid modifications in rationally designed oligonucleotides for a variety of applications, including for antisense oligonucleotide design.

## 3.2 Results

### 3.2.1 Derivation of thermodynamic parameters with high-resolution melting

To derive approximate thermodynamic parameters using HRM, we designed a set of sequences that achieved the maximum number of linearly independent sequences possible given constraints between parameters [78]. To evenly represent all parameters in sequence space, we designed 3 sets of sequences that each attained maximum rank in the stacking matrix (i.e., the combinations of nucleotide pairs needed to fully cover the parameter space) and combined these sets to produce a total of 66 sequences. The sequences ranged between 12 and 30 bases in length, with predicted $T_m$ values between 50°C and 80°C, as this suited the temperature range of the qPCR machine used for analysis (37°C to 98°C). Sequences were also designed to have secondary structure that were less stable than -1 kcal/mol at 37°C and 0.5 M NaCl, as calculated by NUPACK [218].

We performed HRM with an EvaGreen intercalating dye on thermally annealed duplexes that comprised each sequence in the designed set and its complement, at concentrations ranging from 1 $\mu$M to 20 $\mu$M. For each concentration, the $T_m$ was determined as the peak in the $-dF/dT$ of the melting curve. We applied linear regression to the $T_m$ series using the Van't Hoff equation and thereby determined $\Delta H$, $\Delta S$, and $\Delta G_{50}$ values (after adjusting to 1 M NaCl as reported by [161]) (Fig 3.1). In general, $R^2$ values were greater than 0.95. Experimentally derived, non-salt-adjusted $\Delta H$, $\Delta S$, and $\Delta G_{50}$ values are reported in the Supplemental Information.

Parameters for $\Delta H$, $\Delta S$, and $\Delta G_{50}$ for fitted PO-PO internal nucleotide pairs and terminal nucleotides are shown in Table 3.1. Since duplexes were predicted to have melting temperatures within a 50-80°C range, the reported $\Delta G$ was extrapolated to 50°C ($\Delta G_{50}$) to minimize the impact of heat capacity changes on unfolding. Although errors (standard deviation) for fitted $\Delta H$ and $\Delta S$ parameters were high, the fact that $\Delta H$ and $\Delta S$ are highly correlated led to much smaller errors for the derived $\Delta G$ parameters (which rely on entropy-enthalpy compensation) [163].

Parameter values from the HRM-derived model were on average higher (i.e. less stabilizing) than values reported in previous nearest-neighbor models. This is due to concentration-dependent interactions with the dye [173]; in general, while the dye stabilizes the duplex and therefore increases the measured $T_m$, the magnitude of the increase depends on the ratio of dye to duplex. We therefore measured $T_m$ at different dye concentrations for the same duplex concentrations and indeed observed that the shift in $T_m$ varies by dye/duplex concentration ratios (Fig 3.9), resulting in a larger upward shift to the calculated $\Delta G$ for lower dye/duplex ratios. To apply a correction for dye effects, we reasoned that the strength of the

effect should correlate with the number of intercalation sites on the duplex, which in turn is a function of overall duplex length. We selected a total of 16 duplexes whose $\Delta G$ values had been previously calculated from hyperchromicity measurements, ranging from 10 to 16 nucleotides in length [162, 183, 19, 143]. We avoided sequences containing homopolymer runs greater than 4 bases, as our own sequence designs originally excluded these. We used $\Delta H$ and $\Delta S$ parameters from our HRM model to predict $\Delta G_{37}$ of each sequence ($\Delta G_{37}$-HRM) and fitted a linear length-dependent correction that adjusted this value to match as closely as possible reported values extracted from melting as assessed via hyperchromicity. Uncorrected $\Delta G_{37}$-HRM predictions were consistently higher (i.e. less stable) than the reported value. An equation to correct for dye intercalation, $\Delta G_{\mathrm{HRM}} + A \times \mathrm{SequenceLength} + B$, was fitted to minimize the residual sum of squares (RSS) value between predicted and reported values, resulting in values 0.19 and -3.73 kcal/mol for A and B, respectively. The corrected model had a RSS of 12.44 compared to 7.25 for previously established hyperchromicity models [161], a great improvement over the uncorrected model, which had a RSS of 54.36 (Fig 3.2). These results show that, with some simple adjustments, HRM can be used to build predictive models for approximating duplex stability, and potentially provides a high-throughput and cost-efficient route to characterize novel nucleic acid duplexes that otherwise lack sequence-dependent models.

### 3.2.2 Predictive models for duplexes with fully-PS strands

Having proved the basic method's applicability, we attempted to establish approximate models for duplex stability with DNA strands containing entirely phosphorothioate (PS) linkages, for which sequence-dependent parameters had not been previously determined.

We anticipated that PS duplexes should be well-approximated by nearest-neighbor models since the thiol modification does not alter the structure of nucleobases and base-stacking has been shown to be the major energetic contributor to helix stability [41, 48]. While our study used non-stereospecific PS oligonucleotides, the different properties of the $R_p$ and $S_p$-stereoisomers have been shown to have relatively minor impacts on duplex stability, especially in comparison to the impact of sequence composition [23].

We studied two types of duplexes: a PS DNA strand paired with an opposing PS DNA strand (PS-PS), and a PS DNA strand paired with a PO DNA strand (PS-PO). Our sequence sets included the same 66 sequences described previously for PO-PO. Because PS-PO duplexes are hybrid duplexes that are not "symmetrical" about the base pairing axis (unlike PO-PO and PS-PS), a larger set of parameters was needed, since no nucleotide pair was redundant. This "asymmetrical" model contained a total of 16 internal nucleotide pair parameters and 8 terminal nucleotide parameters (Fig 3.7). We again performed leave-one-out cross-validation to compare the fit of the symmetrical model with and without terminal parameters for PS-PS duplexes (Fig 3.10), and the asymmetrical model with and without terminal parameters for PS-PO duplexes (Fig 3.11). As was previously observed for PO-PO duplexes, the inclusion of the terminal parameters significantly improved $T_m$ prediction accuracy (RMSE of 3.04°C to 1.60°C for PS-PS and 2.84°C to 1.73°C for PS-PO). Addition of terminal parameters largely improved $\Delta G$ prediction in PS-PS (RMSE of 0.69 kcal/mol to 0.38 kcal/mol), but not in PS-PO (0.63 kcal/mol to 0.62 kcal/mol). Fitted parameters for the symmetrical model for PS-PS and for the asymmetrical model for PS-PO duplexes are shown in Table 3.2 and Table 3.3, respectively. Terminal parameters are clearly important inclusions to HRM-derived duplex stability models for accurate prediction

of thermodynamic properties such as $\Delta G$ and $T_m$. Interestingly, we observed an increase in stability of the terminal parameters as the duplex included more fully PS strands: an average of 0.51, -0.52, and -1.25 kcal/mol for each terminal nucleotide for PO-PO, PS-PO, and PS-PS, respectively. This could indicate that although sequence composition is a key determinant of duplex stability in all 3 backbone conditions, it has a smaller impact on overall duplex stability in PS-PS duplexes than in PO-PO and PS-PO duplexes, possibly due to global helix destabilization by extensive phosphorothioation.

### 3.2.3 Predictive models for duplexes with partially phosphorothioated strands

Next, we investigated how to best model the thermodynamics of duplexes containing strands that contain a mix of PO and PS linkages. To increase the generality of our methods, we selected 2 new sequences unrelated to the previous 66 we had used and designed a set of partially PS-modified strands for each sequence ranging from 1 to 9 modifications. We combined these partial-PS strands with either fully-PO or fully-PS complement strands to produce 10 duplexes that varied in the number of total PS modifications: from 0 (i.e. two fully PO strands); to 1, 4, 9, and 19 (i.e., a fully PO top strand with a fully PS bottom strand and vice versa); and ultimately to 20, 23, 28, 38 (i.e., two fully PS strands). The $\Delta G$ values of each partially-modified duplex were once again experimentally determined using HRM at a range of concentrations. To predict the $\Delta G$ of duplexes with partially-modified strands, we used the parameters from models fitted without terminal parameters, since individual modifications likely have unique impacts on global stability, and our terminal parameter models were based on fits for the global stability of duplexes containing fully phosphorothioated strands.

Comparing the measured and predicted $\Delta G$ values, we found that the model predicted the stability of the partially-PS duplexes fairly well, resulting in an $R^2$ of 0.94 and 0.82 for the two sequences tested (Fig 3.3A). Predictions were more accurate for duplexes in which fewer than half of all linkages were PS. Across all sequences in our set, we found that PS linkages resulted in an energetic difference of 0.115 $\pm$ 0.04 kcal/mol per modification, on average (Fig 3.3B).

The stabilities of partially-modified duplexes can thus be approximated in a sequence-dependent manner by nearest-neighbor type models, with a few caveats. First, transitions from one phosphate backbone to the other may result in energetic penalties that depend on a sequence context beyond nearest neighbors, since more modifications will result in an overall change in structure. This was best seen by the departure in prediction accuracy with increasing phosphorothioate modifications. Second, the lack of terminal parameters means that predictions will only hold true for a limited range of sequences. In the absence of terminal parameters specifically determined for duplexes at various levels of modification, using internal parameters alone to make predictions will cause shorter partially-modified duplexes to proportionally depart greater from experimental values.

### 3.2.4 Predicting the impact of phosphorothiate modification on rationally designed nucleic acid circuits

Rationally designed nucleic acid systems have been used for a variety of applications, including enabling sensitive detection of analytes, precise assembly of nanoscale structures, and even chemical computation [175, 33, 88]. This programmability comes in part from the fact that experimental nucleic acid hybridization parameters often closely match theory,

allowing accurate designs.

As an example, catalytic hairpin assembly (CHA) is an in vitro DNA-based signal amplification reaction capable of achieving up to hundreds-fold amplification of nucleic acid inputs [212, 119], making it potentially useful for diagnostic applications [118]. CHA designs to date have derived in large measure from predictions by programs such as NUPACK [218] that in turn rely on experimentally determined nearest-neighbor parameters. By modifying the sequence of key regions, CHA circuits have been engineered to operate at various temperatures [96] and to have reduced background leakage [97, 15].

While changes to circuit stability can be achieved by introducing mismatches or shortening sequence domains, modification of the backbone with phosphorothioates could also serve to destabilize hybridization of a given duplex relative to a fully phosphodiester counterpart. For example, the use of PS modifications (combined with additives such as single-stranded DNA-binding proteins and urea) has already enabled enzyme-mediated isothermal amplifications to operate with high specificity at lower temperatures [22]. Moreover, PS modifications should also prove useful for imparting nuclease resistance to DNA circuits mixed with biological samples [177].

To further investigate whether and how PS modifications can impact circuit design, we generated a catalytic hairpin assembly (CHA) circuit that contained a hairpin (H1) that was fully phosphorothioated (PS-H1) (Fig 3.4; Table 3.5). This circuit was based on a previously published high-temperature CHA (HT-CHA) circuit with an operating temperature of 60°C [96]. We predicted that the circuit would now have a lower effective temperature range, and that its performance could be predicted via the models we have developed. In greater detail, at the maximum operating temperature of 60°C, the unmodified intermediate (i.e.

PO-H1:catalyst complex) and product (i.e. PO-H1:PO-H2 complex) species exhibit duplex stabilities of -23.3 kcal/mol and -37.8 kcal/mol in the hybridized region, respectively. Our model predicted that the modified versions of these complexes would have these same stabilities at 50.1°C (PS-H1:catalyst) and 46.0°C (PS-H1:PO-H2) (Fig 3.5A), suggesting that the circuit with PS-H1 would have a maximum operating temperature of around 50°C. In fact, when CHA was carried out with PS-H1 a decrease of activity beyond 50°C was observed (Fig 3.5B), in accord with modeling. A much lower overall signal was also observed with PS-H1 than with PO-H1 (e.g. peak activity of 25 a.u./min compared to 150 a.u./min). This was likely due to the reduced stability of the H1:Reporter complex as a result of phosphorothioation of the H1 strand.

We then tested how smaller-scale PS modifications, such as modification of individual domains, can impact circuit behavior. To this end, we started with a previously developed low-temperature CHA (LT-CHA) circuit designed for operation at 37°C [96] as a starting point and generated versions of LT-CHA circuits with strands that contained one or more PS-modified domains. We chose to modify LT-CHA since LT-CHA components are less stable and therefore more sensitive than their high-temperature counterparts to small energetic penalties (i.e., 0.12 kcal/mol per PS modification). These include a catalyst strand with a PS domain 1 (C*1), a catalyst strand with a PS domain 2 (C*2), a fully-modified catalyst strand (C*123), and a hairpin 1 with a PS toehold (PS-H1*1), as well as their PO counterparts (Table 3.6, Fig 3.6A). In the first step of CHA, the toehold of H1 binds to the single-stranded catalyst, and H1 is unfolded by the catalyst to form the H1:catalyst complex. Thus, the H1:catalyst complex must be energetically favored over the folded H1 structure to drive the reaction forward. To show the predictive power of our model, we estimated the difference in

duplex stability between the hairpin:PS catalyst complexes and the folded PO hairpin (i.e. $\Delta\Delta G$ = $\Delta G$-H1:C - $\Delta G$-Folded PO H1), which should correlate with circuit activity. In accord with PS destabilization and our model, a loss of activity was expected for CHA with PS-modified components.

In fact, the initial activity rates of chemically modified CHA circuits showed a good correlation with respect to $\Delta\Delta G$ (Fig 3.6B). For example, modifying domain 1 in only hairpin 1 of CHA with PS residues increased $\Delta\Delta G$ of the PS H1:C complex to to -0.65 kcal/mol and resulted in 75% of the original CHA activity, while modifying domain 1 in both hairpin 1 and the catalyst strand increased $\Delta\Delta G$ to -0.13 kcal/mol (i.e. only slightly favoring the forward reaction) and showed 30% of original activity.

## 3.3 Discussion

In this work, we carried out HRM experiments to develop approximate thermodynamic models for PO-PO, PS-PO, and PS-PS DNA duplexes, the latter two of which do not yet have published sequence-dependent models. Based on our analysis, $T_m$ determination by HRM with the EvaGreen intercalating dye resulted in models that slightly underestimated the stability of duplexes (ie. predicted higher $\Delta G$ values). While part of the skew may be due to dye intercalation [131], simply assuming a linear relationship between possible dye-binding positions (correlating with the total number of base-pairs) and the degree of destabilization allowed adjustments to be made, to the point where predictions were similar to those derived from UV-Vis hyperchromicity models. Overall, the biases accorded to dye binding were fairly minor, with an average correction of 0.19 kcal/mol per base.

More generally, there are notable differences between HRM and UV-Vis measurements that should be taken into account when fitting model parameters. The indirect nature of HRM allows high-throughput $T_m$ measurements (i.e., compatible with 96-well plates) and relatively low concentrations (down to 1 $\mu$M oligonucleotide), resulting in more rapid and scalable model development. However, $T_m$-HRM (the $T_m$ defined by HRM) must be derived from the -dF/dT plot rather than by regression curve fitting or baseline extrapolation methods [140, 148, 132, 151] typically used to determine $T_m$-UV-Vis (the $T_m$ defined by UV-Vis; the value at which half of all duplexes are bound), because curve fitting and baseline extrapolation are not sensitive enough to detect the duplex-to-single-stranded transition in HRM data at lower concentrations. Overall, this results in a $\sim$1-2°C difference between $T_m$-HRM and $T_m$-UV-Vis [140]. HRM-based models therefore trade off opportunities for rapid and high-throughput modeling with lower accuracy. Depending on ultimate applications, $T_m$-HRM should prove useful for quickly generating models for the increasing range of chemistries available to oligonucleotides, especially backbone or sugar ring modifications that introduce a new degree of freedom that, in conjunction with nucleobase sequence, might require a combinatorially large (and synthetically intractable) set of duplexes to fully characterize.

By demonstrating that phosphorothioate duplexes, like phosphodiester duplexes, can be represented by a nearest-neighbor type model, we set the stage for the development of predictive models that can inform the designs of modified sequences that contribute to practical applications, such as nucleic acid circuitry. Our results showed that duplex stability decreases with an increasing number of modifications, with each modification resulting in an average energetic penalty of 0.12 kcal/mol. Destabilization via phosphorothioation was shown to affect circuit dynamics in a predictable manner and therefore provides a design

strategy beyond merely editing sequence. In addition, considering that PS modifications have been regularly used in the design of therapeutic antisense oligonucleotides [45], our predictive models may narrow the range of designs, thereby reducing time and cost for testing candidates. For example, ATL1102 is a 20-nts antisense oligonucleotide designed for treatment of multiple sclerosis that is fully phosphorothioated and additionally includes 2'-O-(2-methoxyethyl) modifications and methylated cytosine and uracil bases [122, 8]. Based on the PS-PO HRM model and assuming a physiological sodium concentration of 141 mM [155] and an oligonucleotide concentration of 10 nM, for a PS-PO duplex of the same sequence with no additional modifications we predict a $T_m$ of 37.0°C, which is physiological temperature. In general, under physiological conditions, we predict that fully-PS DNA oligonucleotides with $T_m$ values within 0.25°C of 37°C can range in length from 13 to 26 nucleotides. Into the future, hybridization models rapidly determined by HRM for other commonly used (and currently unmodeled modifications) – such as 2'-O-methoxyethyl, morpholino, and peptide nucleic acids – may also impact the the efficient design of oligonucleotide therapeutics.

## 3.4 Materials and Methods

**Reagents and oligonucleotides.** All oligonucleotides were ordered from Integrated DNA Technology (IDT, Coralville, IA, USA). PS DNA oligonucleotides were produced through non-stereospecific chemical synthesis; as a result, PS oligonucleotides used in this study may contain either the $R_p$ or $S_p$ diastereomer at each modified position. All chemicals were purchased from Fisher Scientific (Waltham, MA, USA). Oligonucleotides used for model parameter determination are listed in Table 3.4, and those used for CHA are listed in Table 3.5 and 3.6. Oligonucleotides were stored at 100 $\mu M$ in nuclease-free water at

-20°C. Reactions were carried out in 1x NNE buffer (500 mM NaCl, 10 mM $Na_2HPO_4$, 1 mM EDTA, pH 7.0) for HRM experiments and 1x TNaK buffer (20 mM Tris-HCl, 140 mM NaCl, 5 mM KCl, pH 7.5) for CHA.

**Sequence design for parameter determination.** Each sequence can be represented as a linear combination of nearest-neighbor nucleotide pairs [79]; the linear combinations of pairs that make up a set of sequences can be represented together as a stacking matrix. The duplex thermodynamic value (i.e. $\Delta G$, $\Delta H$, $\Delta S$) of a given sequence is the sum of the contributions of each parameter in the duplex. Thus, in the example of $\Delta G$, given a set of sequences represented by stacking matrix $A$, we can represent the duplex $\Delta G$ of all sequences in the set as a vector $\vec{b}$, where $\vec{b}$ is the product of the stacking matrix and the vector of all parameter $\Delta G$ contributions $\vec{x}$.

$$\overbrace{\begin{bmatrix} n_{\text{seq1,AA/TT}} & n_{\text{seq1,AT/TA}} & \cdots \\ n_{\text{seq2,AA/TT}} & n_{\text{seq2,AT/TA}} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}}^{A} \cdot \overbrace{\begin{bmatrix} \Delta G_{\text{AA/TT}} \\ \Delta G_{\text{AA/TT}} \\ \vdots \end{bmatrix}}^{\vec{x}} = \overbrace{\begin{bmatrix} \Delta G_{\text{seq1}} \\ \Delta G_{\text{seq2}} \\ \vdots \end{bmatrix}}^{\vec{b}}$$

The sequence set was designed to have a rank of 20, which is the maximum rank for nearest-neighbor stacking matrices [78, 79]. The final set contains 66 total sequences and consists of three 20-sequence subsets that independently attain rank 20.

**$T_m$ measurement, determination of thermodynamic values, and model fitting.** Each duplex was annealed prior to melting experiments by adding equivalent amounts of top and bottom strands to obtain a final concentration of 25 $\mu$M and incubated for 5 minutes at 95°C followed by a 0.1°C/s ramp down to 20°C. The annealed sequences were

used to prepare 4 replicate samples at various final concentrations (1, 2.5, 5, 7.5, 10, 15, 20 $\mu$M), and each sample was adjusted to contain 1x NNE buffer and 1x EvaGreen dye (20x EvaGreen dye in water purchased from Biotium, Hayward, CA). HRM data was collected in the Roche LightCycler96 qPCR machine (Roche Molecular Systems, Inc., CA, USA) at excitation 470 nm and emission 514 nm. $dF/dT$ was calculated using the Roche LightCycler Software version 1.1.0 (Roche Diagnostics International) by selecting "Add Analysis" and "$T_m$ calling". $T_m$ is defined as the peak of the $dF/dT$ curve, and samples without distinct peaks were excluded from the analysis. We used linear regression to fit the melting data to the equation

$$\frac{1}{T_m} = \frac{R}{\Delta H} \ln\left(\frac{C_T}{4}\right) + \frac{\Delta S}{\Delta H}$$

to estimate duplex $\Delta H$, $\Delta S$, and by extension, $\Delta G$. $\Delta G$ was extrapolated to 50°C to minimize heat capacity changes of unfolding. Values of $\Delta S$ or $\Delta G$ were adjusted to 1 M NaCl during the fit using the salt correction reported in [161]. Unadjusted values are reported in the Supplemental Data. A total of 4 sequences in the PO-PO dataset, 1 in the PS-PO dataset, and 2 in the PS-PS showed high $\Delta H$ error (>30% of fitted $\Delta H$ value) were removed on the basis that high error during Van't Hoff analysis suggests either non-two-state behavior or incorrect concentration. In each dataset, the set of remaining sequences maintained the maximum rank of 20. All errors reported are standard deviations of the parameter fits. Sequence $\Delta S$ and $\Delta H$ variances for each sequence were determined by regression and used to calculate $\Delta G$ variances as described in [163].

For each model, the sequence variances were transformed into the parameter basis,

resulting in a covariance matrix (CNN). To allow us to drop covariances between parameters while not underestimating the error, we found the smallest diagonal covariance matrix $C'_{\mathrm{NN}}$ in the parameter space such that the matrix inequality $C_{\mathrm{NN}} \leq C'_{\mathrm{NN}}$ holds. Variances derived from $C_{\mathrm{NN}}$ are guaranteed to be equal to or overestimate the error on parameters; we report the standard deviations of these parameters. We performed all data analyses using Python, including linear regression to the Van't Hoff equation (scipy.optimize.curve_fit), singular value decomposition (numpy.linalg.svd), minimization of residual sum of squares (scipy.minimize), and convex optimization for finding $C'_{\mathrm{NN}}$ (cvxpy).

**CHA fluorescence kinetic reading.** A 2.5 $\mu$M stock of reporter complex was prepared by mixing 2.5 $\mu$L of RepF (100 $\mu$M stock in 1x TNaK buffer), 5 $\mu$L of RepQ (100 $\mu$M stock in 1x TNaK buffer), 10 $\mu$L of 10x TNaK buffer, and dH2O to reach a final volume of 100 $\mu$L, followed by annealing. A two-fold excess of RepQ was added to ensure efficient quenching of RepF, which is not expected to interfere with the readout of H1:H2. Prior to the experiments, folded solutions of H1 at 5 $\mu$M (5 $\mu$L of 100 $\mu$M stock solution, 10 $\mu$L of 10x TNaK buffer, and 85 $\mu$L of dH2O) and H2 at 10 $\mu$M (10 $\mu$L of 100 $\mu$M stock solution, 10 $\mu$L of 10x TNaK buffer, and 80 $\mu$L of dH2O) were individually prepared from their respective 100 $\mu$M stock solutions by a 5 minute incubation at 95°C followed by a 0.1°C/s ramp down to 20°C. Reaction mixtures (total volume of 25 $\mu$L) contained the following final concentrations in 1x TNaK buffer: 200 nM folded H1, 400 nM folded H2, 50 nM annealed reporter complex, 1 $\mu$M polyT (dT21), and various concentrations of the catalyst strand (500 nM, 250 nM, 125 nM, and 50 nM). Reaction mixtures were loaded to a 96-well plate and immediately transferred to the LightCycler96 plate reader (Roche Molecular Systems, Inc., CA, USA) for fluorescence measurements conducted at 37°C or higher (excitation: 470 nm, emission: 514

nm).

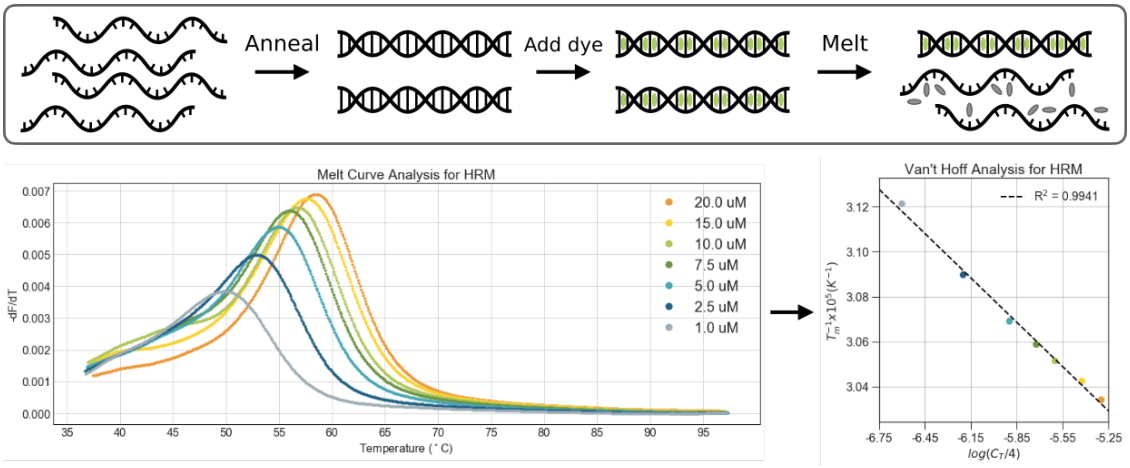


Figure 3.1: **High-resolution melting (HRM) pipeline for determining duplex stability.**

Peak change in fluorescence ($dF/dT$) indicates melting temperature. Thermodynamic parameters are derived from Van't Hoff analysis on HRM data. Since all sequences are non-self-complementary, $1/T_m$ is plotted against $\ln(C_T/4)$.

| Nucleotide Pairs (PO-PO) | $\Delta G_{50}$ (kcal/mol) | $\Delta H$ (kcal/mol) | $\Delta S$ (cal/K/mol) |
|---|---|---|---|
| AA/TT | -0.83±0.14 | -8.10±1.68 | -22.5±4.8 |
| AT/TA | -0.56±0.10 | -5.53±1.35 | -15.4±3.9 |
| TA/AT | -0.58±0.12 | -6.40±1.49 | -18.0±4.3 |
| CA/GT | -0.95±0.15 | -6.89±1.70 | -18.4±4.8 |
| GT/CA | -0.94±0.15 | -7.12±1.88 | -19.1±5.3 |
| CT/GA | -0.94±0.14 | -7.51±1.63 | -20.3±4.6 |
| GA/CT | -0.88±0.14 | -6.51±1.84 | -17.4±5.3 |
| CG/GC | -1.62±0.16 | -10.81±2.03 | -28.5±5.8 |
| GC/CG | -1.76±0.16 | -12.68±1.98 | -33.8±5.6 |
| GG/CC | -1.09±0.15 | -6.09±1.71 | -15.5±4.8 |
| EA/ET | 0.49±0.40 | 20.73±4.98 | 62.6±14.2 |
| AE/TE | 0.48±0.40 | 20.20±4.89 | 61.0±13.9 |
| EC/EG | 0.63±0.40 | 21.07±4.93 | 63.2±14.0 |
| CE/GE | 0.43±0.40 | 18.09±4.84 | 54.7±13.8 |

Table 3.1: **Approximate thermodynamic parameters for PO-PO (phosphodiester-phosphodiester) duplexes derived from HRM data.**

All reported values are adjusted to 1 M NaCl and 50°C. PO-PO = Phosphodiester-phosphodiester duplexes. Errors are defined as the standard deviations of the parameter fits.

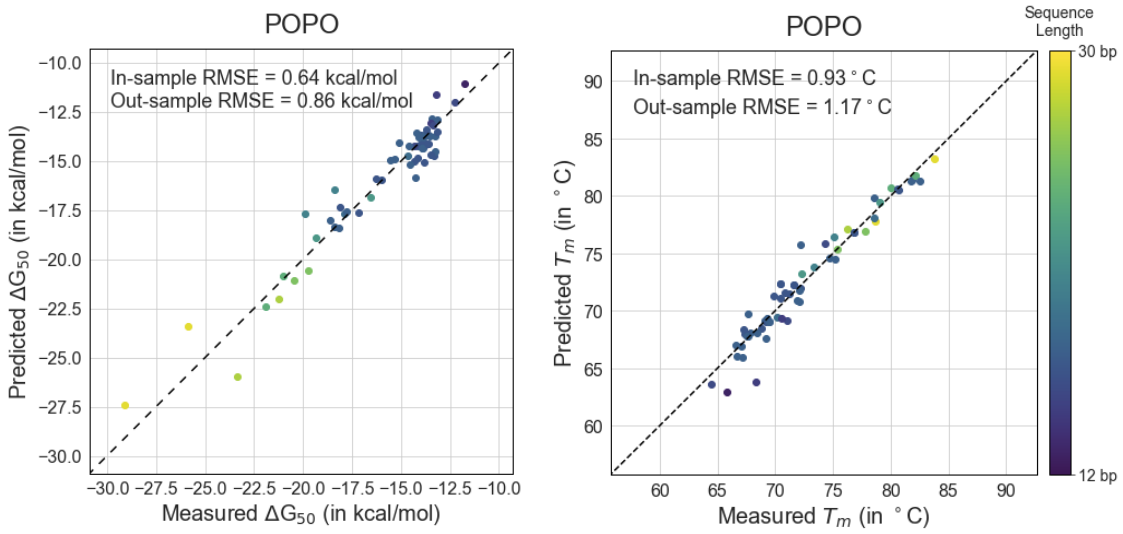


Figure 3.2: **Comparison of $\Delta G$ predictions made by the HRM-derived model and reported UV-Vis models**

14 literature reported sequences are included. RSS = residual sum of squares.

| Nucleotide Pairs (PS-PS) | $\Delta G_{50}$ (kcal/mol) | $\Delta H$ (kcal/mol) | $\Delta S$ (cal/K/mol) |
|---|---|---|---|
| AA/TT | -0.26±0.03 | -4.36±0.77 | -12.7±2.3 |
| AT/TA | -0.16±0.02 | -3.64±0.52 | -10.8±1.5 |
| TA/AT | -0.11±0.01 | -1.93±0.53 | -5.6±1.6 |
| CA/GT | -0.52±0.02 | -5.52±0.57 | -15.5±1.7 |
| GT/CA | -0.50±0.03 | -3.95±0.75 | -10.7±2.3 |
| CT/GA | -0.50±0.03 | -4.16±0.64 | -11.3±1.9 |
| GA/CT | -0.57±0.03 | -5.07±0.90 | -13.9±2.7 |
| CG/GC | -1.06±0.04 | -6.16±1.01 | -15.8±3.0 |
| GC/CG | -1.04±0.04 | -6.90±0.77 | -18.1±2.3 |
| GG/CC | -0.85±0.03 | -5.09±0.83 | -13.1±2.5 |
| EA/ET | -1.30±0.08 | 3.24±2.05 | 14.0±6.2 |
| AE/TE | -1.28±0.08 | 4.18±2.02 | 16.9±6.1 |
| EC/EG | -1.20±0.08 | 1.47±1.98 | 8.3±5.9 |
| CE/GE | -1.20±0.08 | 0.64±2.05 | 5.7±6.1 |

Table 3.2: **Approximate thermodynamic parameters for PS-PS (phosphorothioate-phosphorothioate) duplexes derived from HRM data.**

All reported values are adjusted to 1 M NaCl and 50°C. All internal nucleotide parameters have PS linkages both in the top nucleotide pair and in the bottom pair (e.g. 5'A*A/3'T*T). Errors are defined as the standard deviations of the parameter fits.

| Nucleotide Pairs (PS-PO) | $\Delta G_{50}$ (kcal/mol) | $\Delta H$ (kcal/mol) | $\Delta S$ (cal/K/mol) |
|---|---|---|---|
| AA/TT | -0.52±0.20 | -5.81±3.12 | -16.4±9.1 |
| AT/TA | -0.42±0.10 | -5.64±1.69 | -16.2±4.9 |
| AC/TG | -0.88±0.11 | -8.66±1.85 | -24.1±5.4 |
| AG/TC | -0.71±0.09 | -6.13±0.85 | -16.8±2.3 |
| TA/AT | -0.30±0.10 | -3.86±1.90 | -11.0±5.6 |
| TT/AA | -0.49±0.07 | -5.87±0.64 | -16.7±1.7 |
| TC/AG | -0.64±0.16 | -6.13±2.71 | -17.0±7.9 |
| TG/AC | -0.77±0.15 | -7.28±2.75 | -20.2±8.0 |
| CA/GT | -0.82±0.17 | -7.23±2.67 | -19.8±7.7 |
| CT/GA | -0.66±0.19 | -6.30±3.21 | -17.5±9.4 |
| CC/GG | -0.91±0.11 | -5.57±1.57 | -14.4±4.5 |
| CG/GC | -1.21±0.18 | -8.07±3.09 | -21.2±9.0 |
| GA/CT | -0.85±0.12 | -7.92±2.38 | -21.9±7.0 |
| GT/CA | -0.61±0.14 | -5.46±2.27 | -15.0±6.6 |
| GC/CG | -1.15±0.15 | -6.63±2.26 | -17.0±6.6 |
| GG/CC | -1.09±0.17 | -7.75±2.81 | -20.6±8.2 |
| EA/ET | -0.57±0.40 | 13.64±6.63 | 44.0±19.3 |
| AE/TE | -0.54±0.37 | 15.07±6.01 | 48.3±17.5 |
| ET/EA | -0.59±0.39 | 14.87±6.29 | 47.8±18.3 |
| TE/AE | -0.56±0.38 | 14.72±6.18 | 47.3±18.0 |
| EC/EG | -0.58±0.33 | 10.31±5.35 | 33.7±15.6 |
| CE/GE | -0.55±0.38 | 10.49±6.30 | 34.2±18.4 |

Table 3.3: **Approximate thermodynamic parameters for PS-PO (phosphorothioate-phosphodiester) duplexes derived from HRM data**



All reported values are adjusted to 1 M NaCl and 50°C. All internal nucleotide parameters have a PS linkage between the top nucleotide pair (e.g. 5'A*A/3'TT). Errors are defined as the standard deviations of the parameter fits.

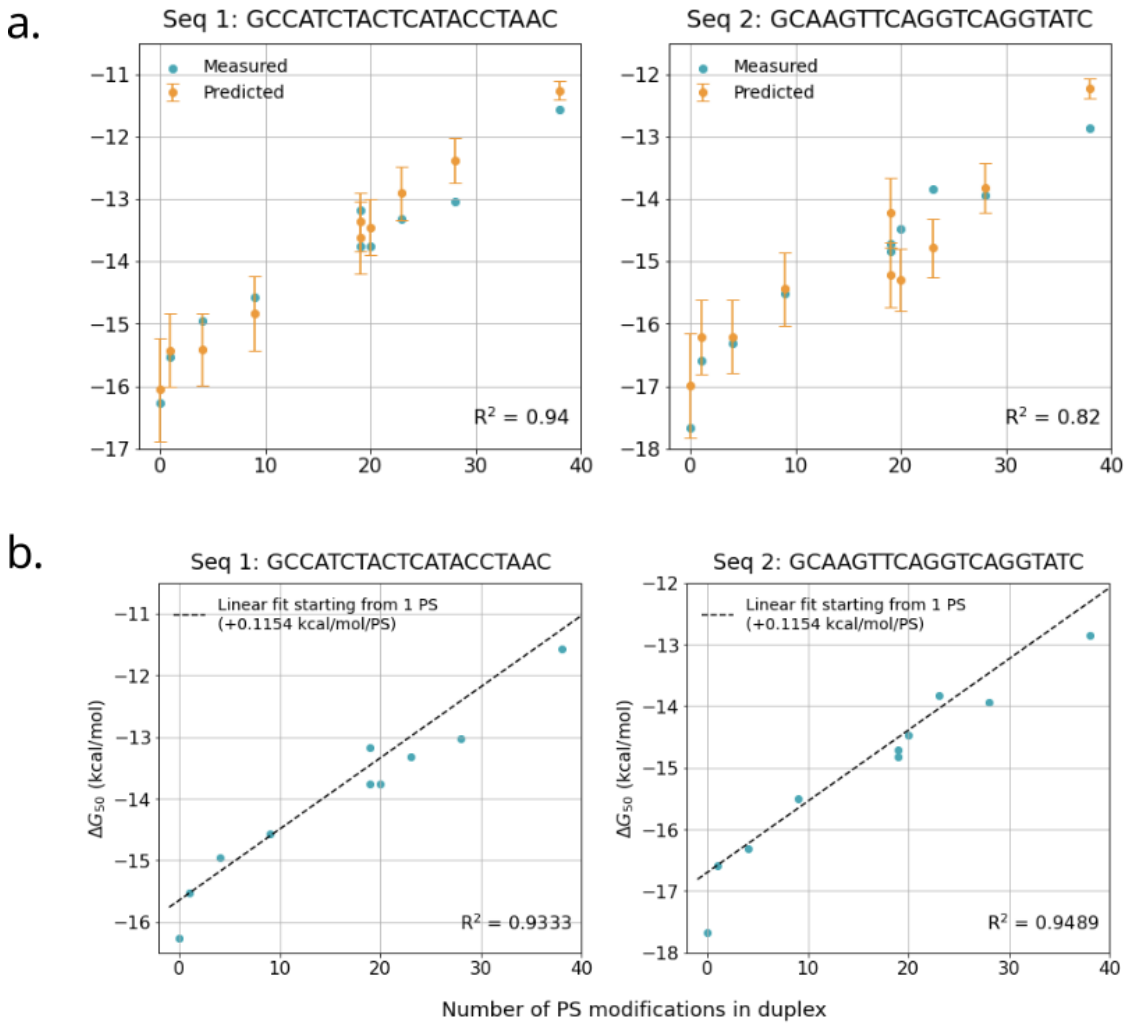


Figure 3.3: **Predicting duplex stability of partially PS-modified duplexes as a function of sequence (a) or independently of sequence (b).**

(a) Duplexes with partially-modified strands were considered as a linear combination of PO-PO (symmetrical), PS-PO (asymmetrical), and/or PS-PS (symmetrical) nucleotide pairs and $\Delta G$ parameters of these pairs across the three backbone conditions were used to predict the overall duplex stability. Errors are calculated using the variances of the parameter estimates. (b) Data from fully PS-PO or PS-PS duplexes was used to determine the average energetic contribution of a single PS backbone (0.1154 kcal/mol/PS). PO-PO duplexes (points at $x = 0$) are not included in the $R^2$ shown for sequence-independent predictions due to the large gap in $\Delta G_{50}$ between $x = 0$ and $x = 1$ seen in both sequences tested.

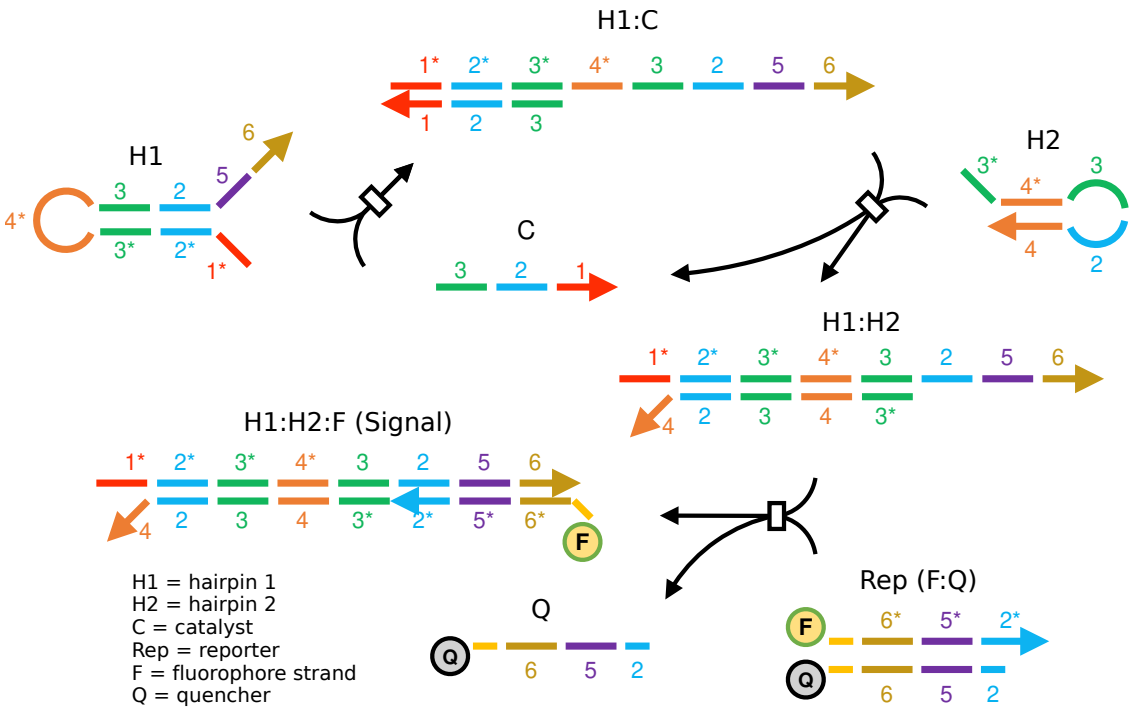


Figure 3.4: **Reaction diagram of catalytic hairpin assembly.**

Asterisks indicate sequence complement. Complexes of multiple strands are denoted with a colon.

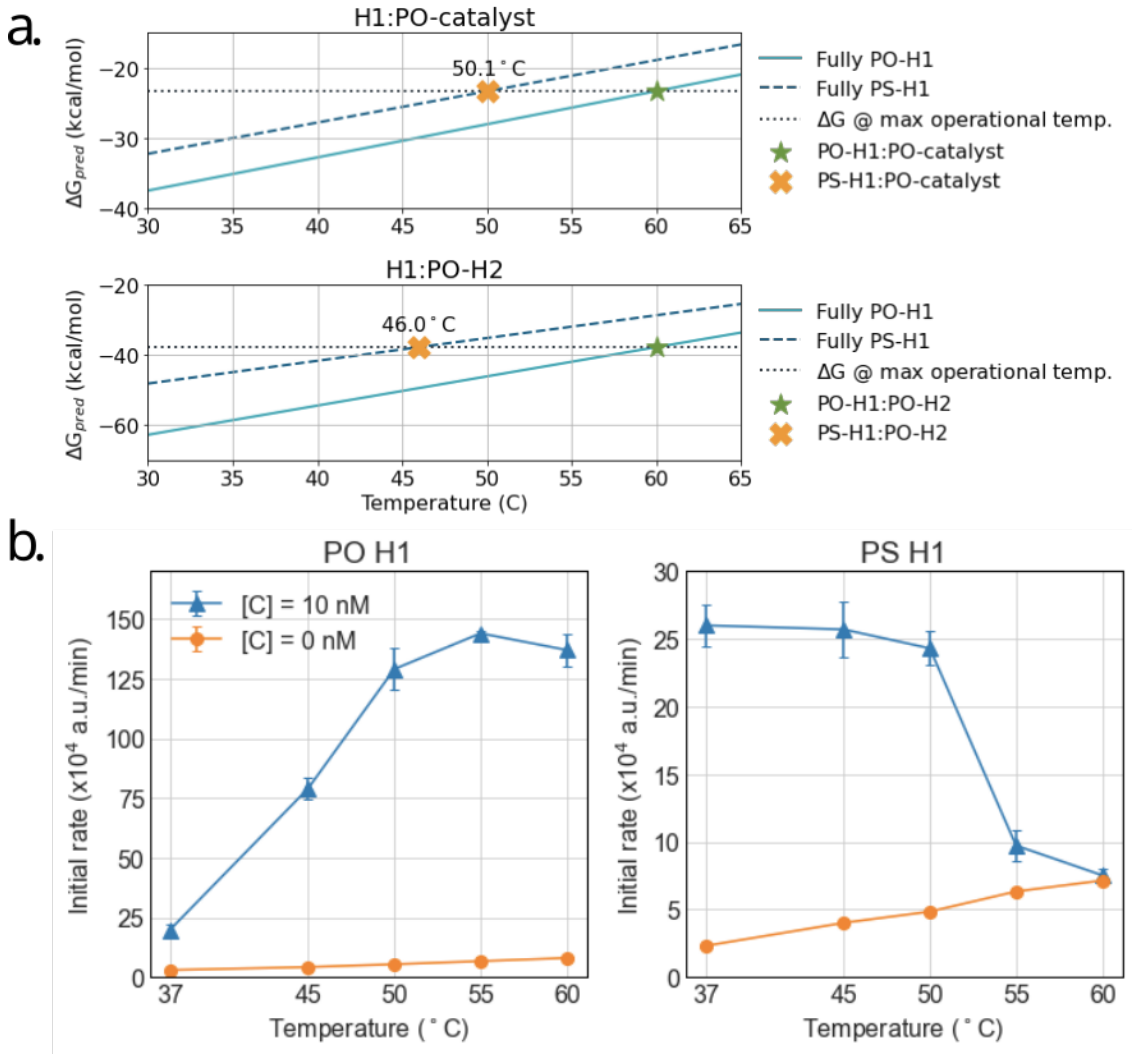


Figure 3.5: **HT-CHA with PO- and PS-H1.**

H1 strand backbones are either fully PO or PS. (a) Duplex stability predictions for interactions involving PO-H1 or PS-H1 (hybridized region only, symmetrical model). Gray dotted line indicates the target stability or the duplex stability of PO-H1:catalyst or PO-H1:H2 at 60°C, the temperature for which the HT hairpins were originally designed [96]. (b) Initial rates of HT-CHA with PO-H1 or PS-H1 at various incubation temperatures. Catalyst strand is fully PO.

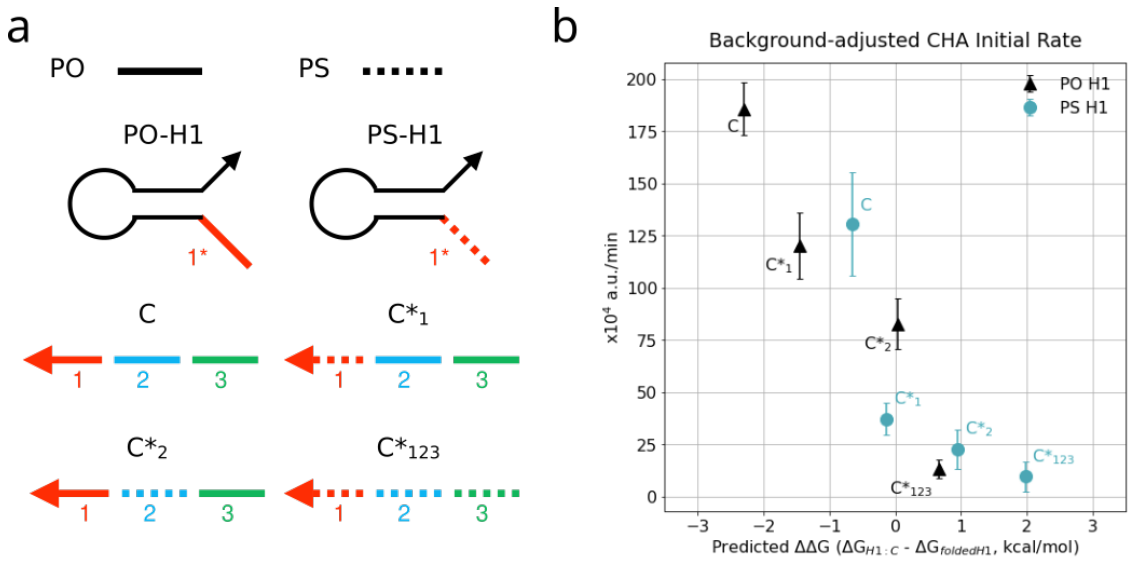


Figure 3.6: **Introducing PS backbones to select domains in LT-CHA.**

(a) Modified parts used in LT-CHA circuit. (b) Observed initial rate of LT-CHA using modified-domain hairpins with various modified versions of catalyst strand. Results show that H1 stability is crucial for maintaining activity. PO-H1 = hairpin 1 with phosphodiester backbones. PS-H1 = hairpin 1 with PS modifications on domain 1*. C*n = catalyst strand with PS on domain(s) n. (c) Predicted duplex $\Delta G$ of H1 using symmetrical NN model for different modified versions of catalyst. Predictions are made with unmodified H1 (left) or PS-domain 1*-H1 (right). LT-CHA

| Number | Sequence | Number | Sequence | Number | Sequence |
|---|---|---|---|---|---|
| 1 | CTGTAAGGCGATATGTT | 23 | TTAACCTGCAATGGATC | 45 | TACTTCCAACGTAGG |
| 2 | TGCCATGTTGAAAACC | 24 | GCGGTTGGTGGCCAAC | 46 | ACGGGTCGTTCCGTG |
| 3 | TATTCTGCCAATGGAAC | 25 | CACAGCGACATGTATGG | 47 | GTGGTACAAATGCGACC |
| 4 | GGTTGCGGTGGCCAAC | 26 | CCCAGGCCTCGGATTTA | 48 | AGCACGGTGGTACAACA |
| 5 | CGACATGTATGGCACAG | 27 | CTCAGAGTAGAGCCGA | 49 | GGTGGCGTTCTT |
| 6 | CTCGGAGGCCCCATTTA | 28 | AGGAGTGAATGAAATGG | 50 | TCTACACCGCGA |
| 7 | CCGCTCAGAGTAGAGA | 29 | GAACTGCCACTACCAA | 51 | ACTGTATCGCCCTA |
| 8 | AATGAGGAGTGAAATGG | 30 | CTACCAGCGGCGCCGTT | 52 | CCGTTGCTGCTAGG |
| 9 | GCTGAACTAACCACCA | 31 | GCAGGAGCCGCGACCTG | 53 | TAGACGCGGCCTCTTTCC |
| 10 | CCACTAGCGGCGCCGTT | 32 | AGCTGAGCCTGCCGCC | 54 | CTAAACTGTTATAGCCGG |
| 11 | GAGGCAGCCGCGACCTG | 33 | GCCCGAGCATGACTGCG | 55 | TGTAAGACTTCTGCCAGAAA |
| 12 | AGAGCGCCCTGCTGCC | 34 | AATGTCGAACAGCAATT | 56 | CGCGCGAGTATTTATAACCT |
| 13 | GATGCCGCAGCGACCTG | 35 | GTGAAACAATGCTGTAG | 57 | TTCTACATCCATCTTAATCCCA |
| 14 | AACGAATGTCAGCAATT | 36 | CAAGCCTCGATTTTGT | 58 | AGACATCCCATACGAGCATCCA |
| 15 | GCTGTGAAACAATGTAG | 37 | GTGAGCAGAAGGGGTT | 59 | ACATGACTCATCTTAGCCGGCGAG |
| 16 | CGATTTTGTCAAGCCT | 38 | ACTCGCTCTACCTTAAT | 60 | CGGGATTTCTGGCATCATTGTCCT |
| 17 | GAGCAGAAGGGGTTGT | 39 | CTTTTGTGCGGGTAGC | 61 | TAATTATACGAGTAGTTTCTGTCCTG |
| 18 | ATAACTTACTCTCGCCT | 40 | TTCGCGGTCTCCATTA | 62 | GATTGTATCATCGACATCACACTACC |
| 19 | CGGTGCTTTTGGTAGC | 41 | CGTGCAGCACTACTTG | 63 | CAAACTTAGTAATCACGCCCAGCAACCA |
| 20 | TCTCGCGGTTCCATTA | 42 | GTCATTGTGCTTTTGC | 64 | GATCTCTCTATCATCGTTTATTGGGTAT |
| 21 | CGTGTGGATAATTAGCT | 43 | GAACCGTTGATGATCTC | 65 | TTGTAGTTGACGTTTGTGATTTAGTGAATT |
| 22 | TTGAAAACCCATGTGC | 44 | TGTCGCACCCTACTA | 66 | TTTGGGTTAGTAAGAAGGCAGCAGTTGGGC |

Table 3.4: **Sequences used for nearest-neighbor parameter determination.**
Sequences were hybridized to their complements prior to HRM. Fully phosphodiester and fully phosphorothioate versions of each sequence listed were used in the study. Sequences maximally span the space of nearest neighbor pairs (i.e. sequences are maximally independent).

| HT-CHA | Sequence |
|---|---|
| HT-H1 | GTCACGTGA GCTAGCGTT AGCATCGTCG CCATGCTGCTAGCA CGACGATGCT AACGCTAGC CCTTGTCA TACGCAGCAC |
| HT-H1-PSall | G*T*C*A*C*G*T*G*A* G*C*T*A*G*C*G*T*T* A*G*C*A*T*C*G*T*C*G* C*C*A*T*G*C*T*G*C*T*A*G*C*A* C*G*A*C*G*A*T*G*C*T* A*A*C*G*C*T*A*G*C* C*C*T*T*G*T*C*A* T*A*C*G*C*A*G*C*A*C |
| HT-H2 | AGCATCGTCG TGCTAGCAGCATGG CGACGATGCT AACGCTAGC CCATGCTGCTAGCA |
| HT-H2-PSall | A*G*C*A*T*C*G*T*C*G* T*G*C*T*A*G*C*A*G*C*A*T*G*G* C*G*A*C*G*A*T*G*C*T* A*A*C*G*C*T*A*G*C* C*C*A*T*G*C*T*G*C*T*A*G*C*A |
| HT-Catalyst | CGACGATGCT AACGCTAGC TCACGTGAC |
| HT-RF | /56-FAM/CGA GTGCTGCGTA TGACAAGG GCTAGCGTT |
| HT-RQ | C CCTTGTCA TACGCAGCAC TCG /3IABkFQ/ |
| HT-Domain 1 | TCACGTGAC |
| HT-Domain 2 | AACGCTAGC |
| HT-Domain 3 | CGACGATGCT |
| HT-Domain 4 | CCATGCTGCTAGCA |
| HT-Domain 5 | CCTTGTCA |
| HT-Domain 6 | TACGCAGCAC |

Table 3.5: **Sequences and domains used for high-temperature CHA.**
Asterisks in sequence indicates positions with PS backbones. Different domains are indicated by different colors. /56-FAM/ = 5' Fluorescein; /3IABkFQ/ = 3' Iowa Black FQ.

| LT-CHA | Sequence |
|---|---|
| LT-H1 | GTCAGTGA GCTAGGTT AGATGTCG CCATGTGTAGA CGACATCT AACCTAGC CCTTGTCA TAGAGCAC |
| LT-H1-PS1 | G*T*C*A*G*T*G*A GCTAGGTT AGATGTCG CCATGTGTAGA CGACATCT AACCTAGC CCTTGTCA TAGAGCAC |
| LT-H2 | AGATGTCG TCTACACATGG CGACATCT AACCTAGC CCATGTGTAGA |
| LT-H2-PS3 | A*G*A*T*G*T*C*G TCTACACATGG CGACATCT AACCTAGC CCATGTGTAGA |
| LT-Catalyst | CGACATCT AACCTAGC TCACTGAC |
| LT-Catalyst-PS1 | CGACATCT AACCTAGC T*C*A*C*T*G*A*C |
| LT-Catalyst-PS2 | CGACATCT A*A*C*C*T*A*G*C* TCACTGAC |
| LT-Catalyst-PSall | C*G*A*C*A*T*C*T* A*A*C*C*T*A*G*C* T*C*A*C*T*G*A*C |
| LT-RF | /56-FAM/CGA GTGCTCTA TGACAAGG GCTAGGTT |
| LT-RQ | C CCTTGTCA TAGAGCAC TCG /3IABkFQ/ |
| LT-Domain 1 | TCACTGAC |
| LT-Domain 2 | AACCTAGC |
| LT-Domain 3 | CGACATCT |
| LT-Domain 4 | CCATGTGTAGA |
| LT-Domain 5 | CCTTGTCA |
| LT-Domain 6 | TAGAGCAC |

Table 3.6: **Sequences and domains used for low-temperature CHA.**
Asterisks in sequence indicates positions with PS backbones. Different domains are indicated by different colors. /56-FAM/ = 5' Fluorescein; /3IABkFQ/ = 3' Iowa Black FQ.

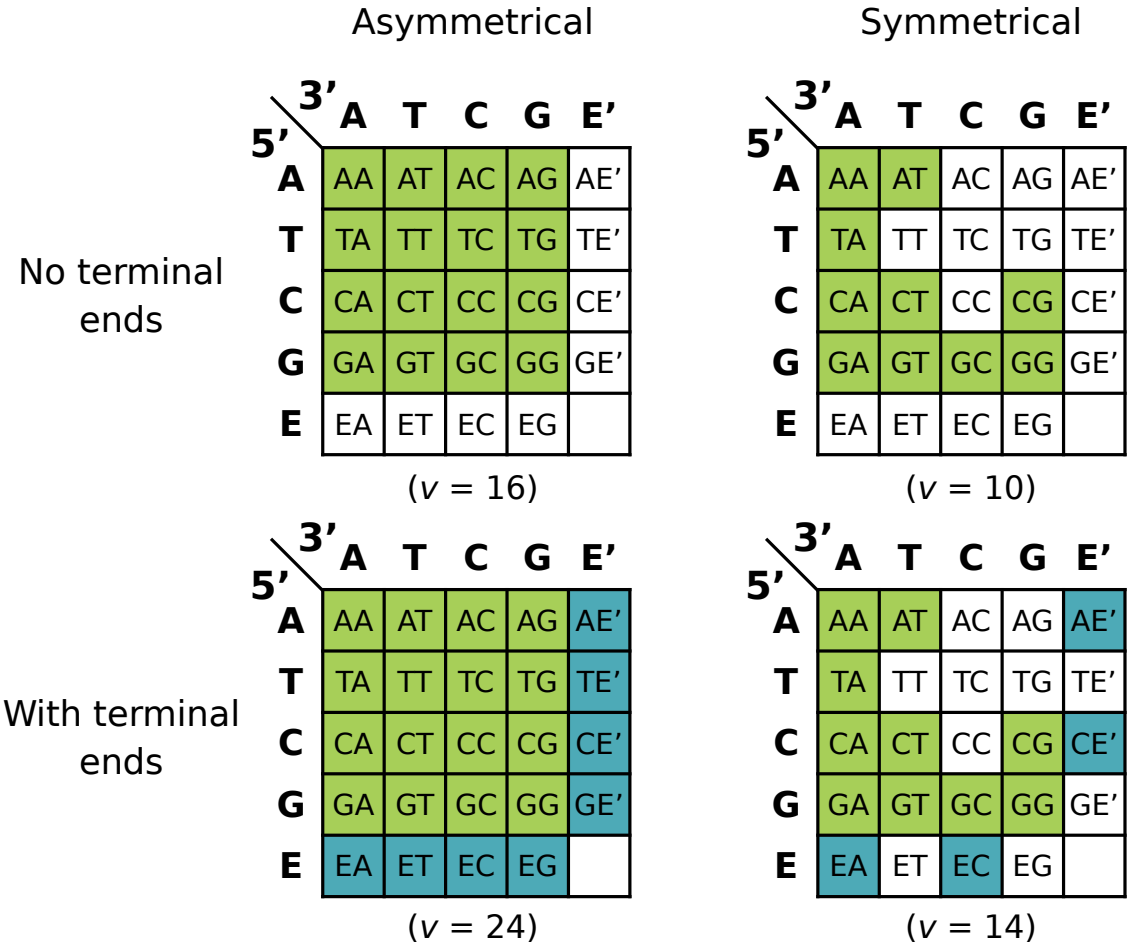


Figure 3.7: **Nearest-neighbor style models considered and the parameters included in each model.**

Filled cells indicate that the nearest-neighbor pair was added as a variable to the model. Green = nucleotide pair variable. Blue = terminal nucleotide variable. $v$ = total variables involved.

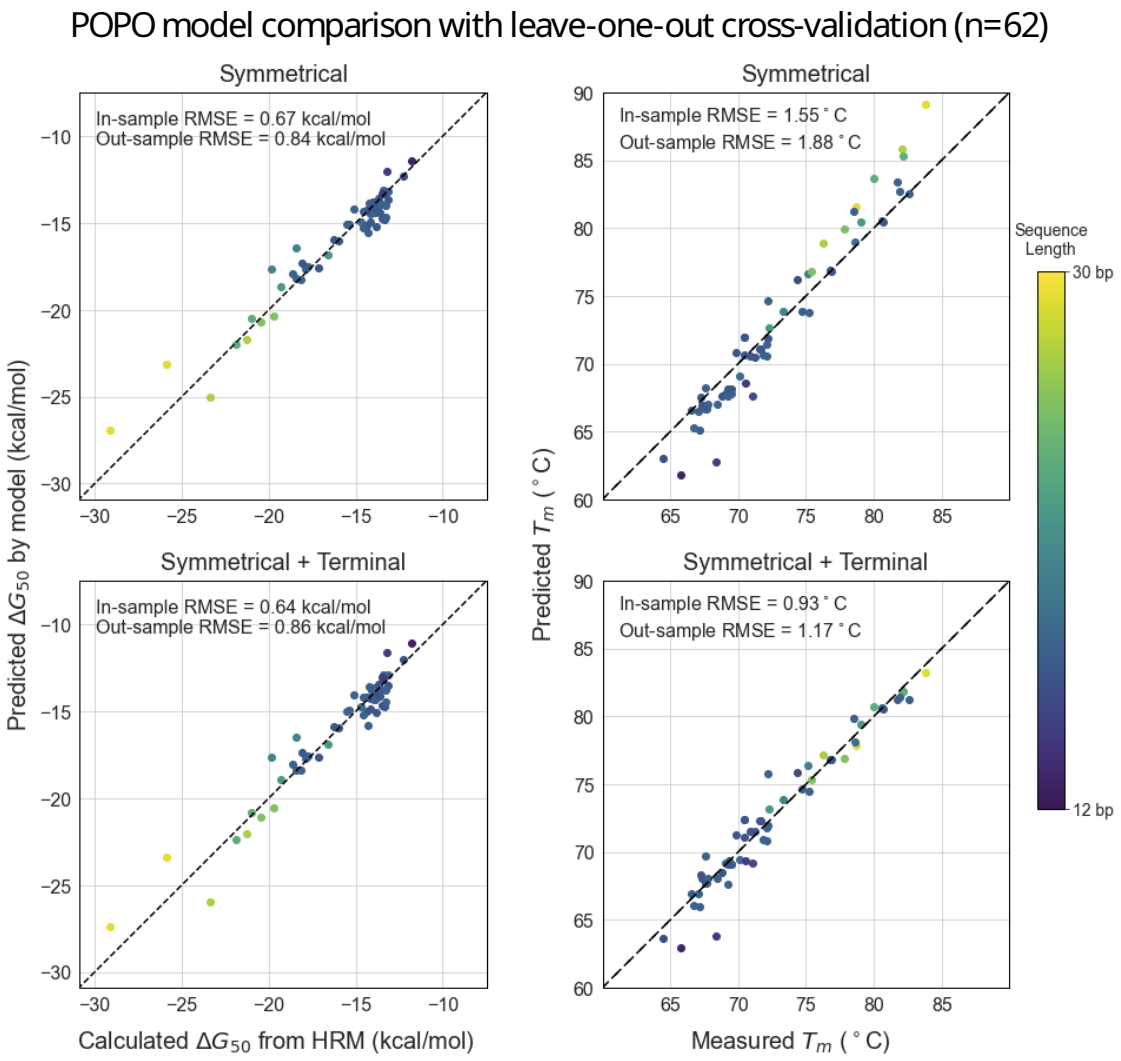


Figure 3.8: **Leave-one-out cross-validation on the PO-PO HRM dataset for $\Delta G_{50}$ and $T_m$ (concentration = 10 $\mu$M)**

Top row = without including terminal nucleotide variables, bottom row = including terminal nucleotide variables. Color of dots represents length of sequence. The dashed line $y = x$ is added to guide the eye. RMSE = root mean square error.

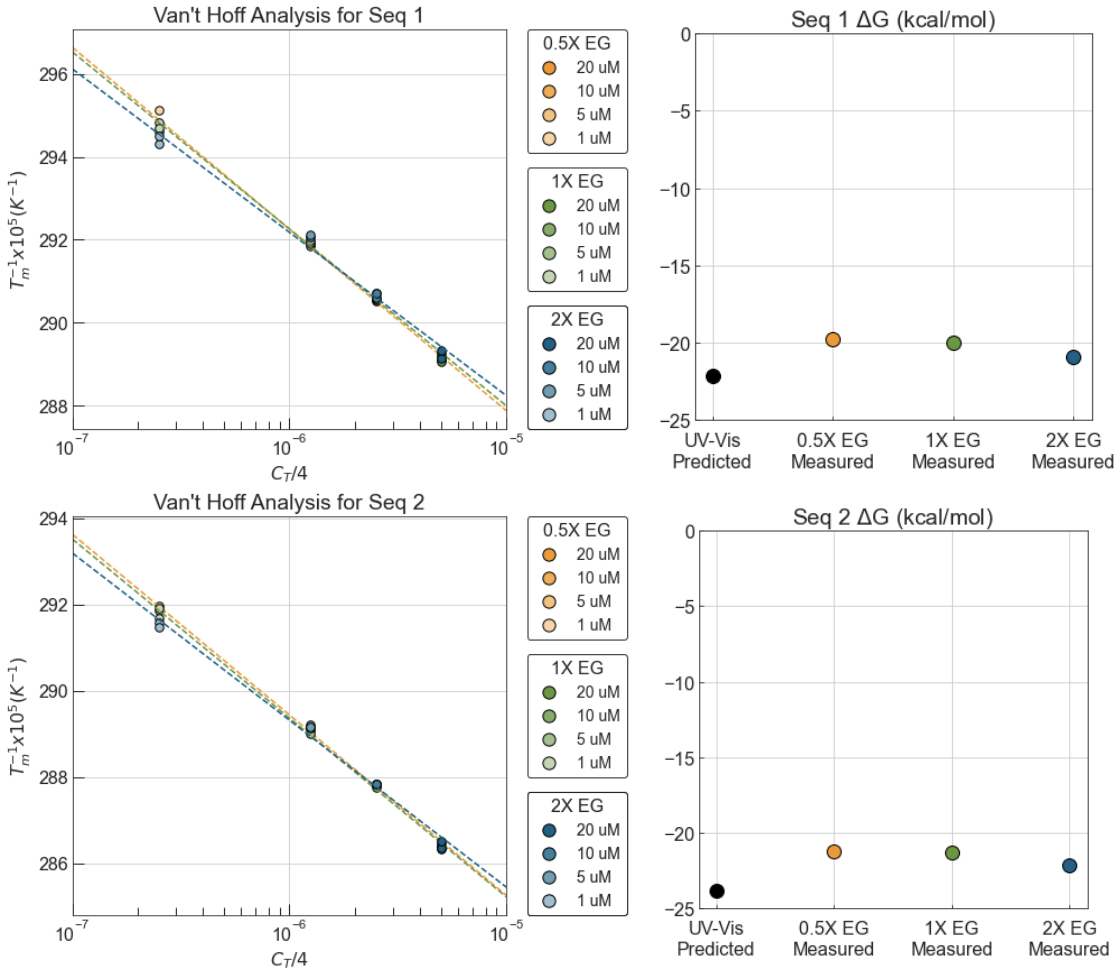


Figure 3.9: **Thermodynamic parameter determination at different EvaGreen dye concentrations.**

EG = EvaGreen dye. Predicted value is based on the model from [161].

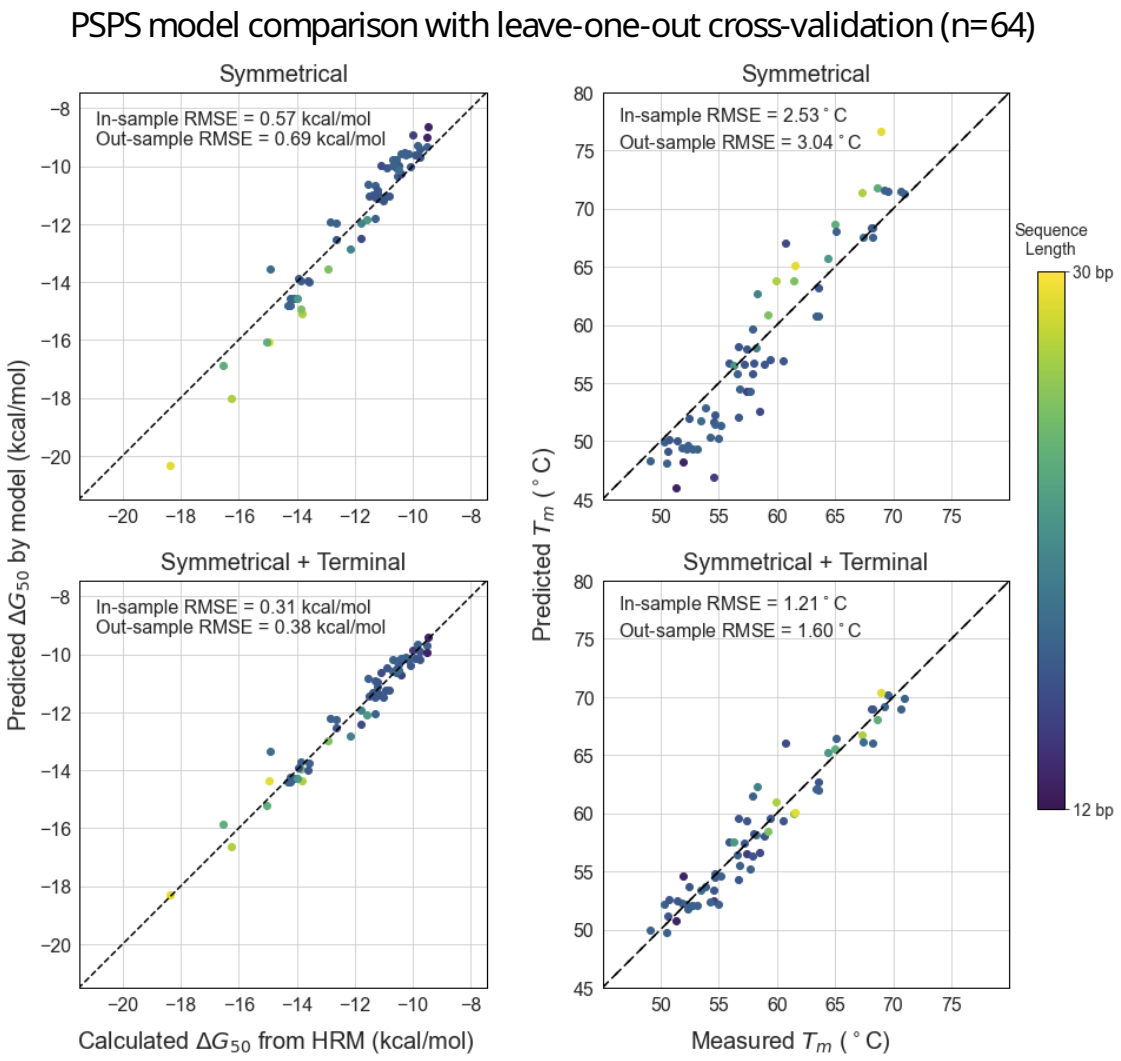


Figure 3.10: **Leave-one-out cross-validation on the PS-PS HRM dataset for $\Delta G_{50}$ and $T_m$ (concentration = 10 $\mu$M)**

Top row = without including terminal nucleotide variables, bottom row = including terminal nucleotide variables. Color of dots represents length of sequence. The dashed line $y = x$ is added to guide the eye. RMSE = root mean square error.

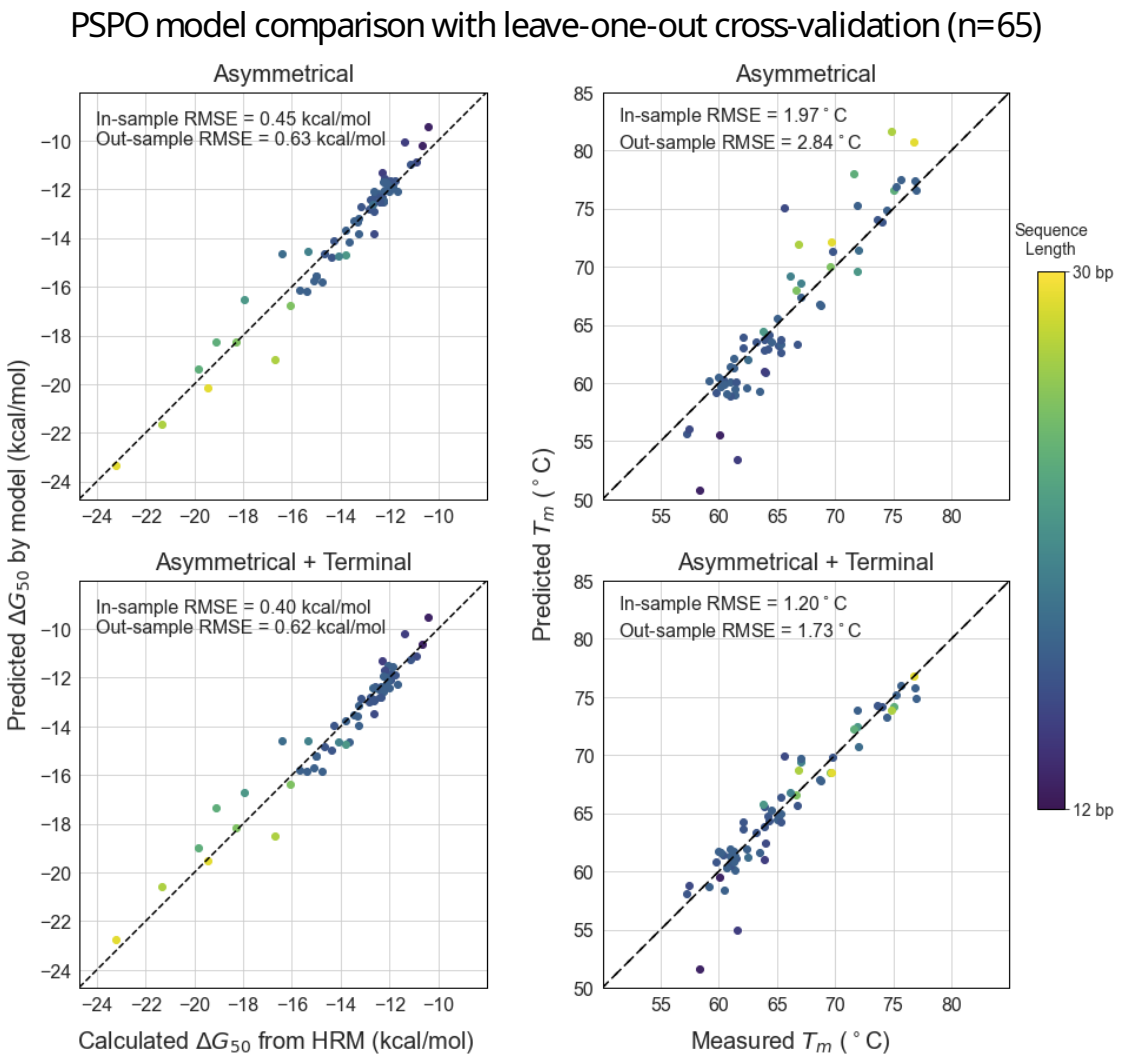


Figure 3.11: **Leave-one-out cross-validation on the PS-PO HRM dataset for $\Delta G_{50}$ and $T_m$ (concentration = 10 $\mu$M)**

Top row = without including terminal nucleotide variables, bottom row = including terminal nucleotide variables. Color of dots represents length of sequence. The dashed line $y = x$ is added to guide the eye. RMSE = root mean square error.

# Chapter 4

# Parallel and in-memory computation with data stored in DNA using strand displacement[1]

**Abstract.** DNA is an incredibly dense storage medium for digital data, but computing on the stored information is expensive and slow as it requires rounds of sequencing and *de novo* DNA strand synthesis. To augment DNA storage with "in-memory" molecular computation, we use strand displacement reactions to algorithmically modify data stored in the topological modification of DNA. A secondary sequence-level encoding allows high-throughput sequencing-based readout. We show that computation can occur in parallel across multiple data. We demonstrate multiple rounds of parallel binary counting and cellular automaton Rule 110

[1]This chapter is adapted from a draft manuscript by Wang B, Wang SS, Chalk C, Ellington AD, Soloveichik D. BW and SSW shared first authorship. BW, CC, and DS devised the project. CC and BW designed the SIMD DNA strand-displacement algorithms. BW designed the experimental protocol for SIMD DNA and performed all SIMD computations, post-computation ligation and displacement, and fluorescence experiments. SW performed all post-computational library preparation, sequence and data analysis, and qPCR assays. BW and SW prepared the figures with feedback from all authors. BW wrote the initial draft of the manuscript which was later edited by all authors. DS and AE obtained funding and provided guidance throughout the project.

computation on 4-bits data registers, as well as selective access and erasure. Avoiding stringent sequence design, we demonstrate large strand displacement cascades (122 distinct steps) on naturally-occurring DNA sequences. Our work merges DNA storage and DNA computing and sets the foundation of massively parallel algorithmic manipulation of digital information kept in DNA.

## 4.1 Introduction

DNA is an incredibly dense (up to 455 exabytes per gram, 6 orders of magnitude denser than magnetic or optical media) and stable (readable over millennia) digital storage medium [39, 27]. Storage and retrieval of up to gigabytes of digital information in the form of text, images, and movies have been successfully demonstrated [138]. Importantly, DNA's essential biological role ensures that the technology for manipulating DNA will never succumb to obsolescence. While these properties make DNA a promising storage medium, it is at present limited to the storage of rarely accessed data ("cold" storage) largely due to its inefficient read-write cycle. Performing computation on the stored data involves sequencing the DNA, electronically computing the desired transformation, and synthesizing new DNA, which is an expensive and slow loop.

Here we design a new paradigm called SIMD||DNA (Single Instruction Multiple Data DNA) which integrates DNA storage with massively parallel in-memory computation. As shown in Figure 4.1A, unlike traditional DNA data storage where information is encoded in the nucleotide sequence, SIMD||DNA encodes information in a register, a multi-stranded DNA complex with a unique pattern of nicks and exposed single-stranded regions. There are

as many independent registers as the number of molecules of the multi-stranded complexes, each capable of storing and manipulating a different value. To manipulate information, an instruction (a set of DNA strands) is applied to the registers. The strand composition of a register updates if the applied instruction strands trigger *strand displacement reactions* within that register. Strand displacement reaction has become a widely versatile building block in engineering nucleic acid based systems. Displacement occurs when an input strand invades a multi-stranded complex through binding to a *toehold* (single-stranded region with five to seven nucleotides) and then displaces the incumbent strand as an output. Through this mechanism, the strand composition, patterns of nicks, and exposed single-stranded regions in the registers are changed. Instruction strands are synthesized independently of the data stored in the registers, so that executing an instruction does not require reading the data. After the non-reacted instruction strands and reaction waste are washed away, subsequent instructions can be performed. Because all registers share the same sequence space, each set of instructions can perform multiple unique strand displacement reactions across multiple registers. This utilizes the parallelism granted by molecular computation (Figure 4.1B). Our DNA data processing scheme is capable of parallel, in-memory computation, eliminating the need for sequencing and *de novo* strand synthesis on each data update. Additionally, the doubly-parallel nature of SIMD||DNA programs allows instructions to act on all registers and multiple sites within a register in parallel.

We constructed the theoretical framework for SIMD||DNA and proved the correctness of two molecular programs: binary counting (a fundamental function in computer programming) and cellular automaton Rule 110 (a Turing universal computation) [200]. We then experimentally implemented these programs and demonstrated correct computation for in-

memory and parallel computation for a pool of 16 registers encoding all possible 4-bit binary values. To scale up the computational power, we show that the registers can be repeatedly processed prior to read out by conducting multiple rounds of computation. In addition, like a computer's memory, information stored in the SIMD||DNA paradigm can be specifically queried (random access) or erased. Registers can be constructed using both chemically synthesized DNA and naturally-occurring DNA (i.e. non-genetically modified sequences), further reducing the dependency on custom oligonucleotide synthesis. We show that unmodified kilobase-length M13 phage plasmid provides a large storage space that allows information size to be scaled up, by constructing multiple sub-registers for parallel computation. So far this is the largest strand displacement system using naturally-occurring DNA sequences: Using SIMD||DNA, we implemented 18 distinct strand displacement reactions in solution at the same time, and in total 122 distinct strand displacement reactions.

## 4.2 Results

### 4.2.1 SIMD||DNA

Figure 4.1 shows the overview of SIMD||DNA. Every register contains a long "bottom" strand and multiple short strands, called *top strands*, bound to the bottom strand. We use *domain* to represent consecutive nucleotides that act as a functional unit. Complementary domains are represented by a star ($*$). The length of the domains is chosen such that: (1) each domain can initiate strand displacement (i.e. can act as a toehold), (2) strands bound by a single domain readily dissociate, and (3) strands bound by two or more domains cannot dissociate. Each bottom strand is partitioned into sets of consecutive domains called *cells*

(Figure 4.1C). Each cell contains the same number of domains. Cells encode information with the binding configuration of their top strands (e.g. lengths, presence or absence of toeholds). For the programs we designed, we used a binary encoding with each cell representing one bit.

Each *instruction* of a program corresponds to the addition of a set of DNA strands at high concentration to a solution containing the registers. The registers are attached to magnetic beads, allowing washing away of beadless non-reacted instruction strands and reaction waste. Registers and instruction strands are allowed to react for a short amount of time before washing such that the high concentration instruction strands interact with the registers, but the low concentration waste products do not. The instruction strands can cause three different types of events (Figure 4.1D). **Attachment** reactions preserve all the strands originally bound to the register and attach new strands (as long as the new strand binds strongly enough—by two or more domains). The attachment of an instruction strand can lead to a partial displacement of a pre-existing strand on the register. **Displacement** reactions introduce new strands to the register and detach some pre-existing strands. Upon binding to a toehold on the register, the instruction strand displaces pre-existing strands through 3-way branch migration. Toehold exchange reactions are favored towards displacement by the instruction strand since they are added at high concentration. Two instruction strands can also cooperatively displace strands on the register. **Detachment** reactions detach pre-existing strands without introducing new strands to the registers. An instruction strand that is complementary to a pre-existing strand with an open overhang can use the overhang as a toehold and pull the strand off the register.

To experimentally implement SIMD||DNA, considerations for readout need to be

incorporated in the design (Figure 4.1E). To read out parallel computation results where registers share the same sequence space, registers with different initial values are given their own barcode sequences. Since information is encoded in the pattern of the top strands, direct readout requires obtaining the location of nicks. To read out the stored information through sequencing while preserving the desired computation logic, we modified the encoding by introducing mismatches between the top strands and the register in a manner that can coexist with the nick-based encoding. Since mismatches can affect strand displacement kinetics [129], the mismatch locations are carefully chosen to ensure that the desired strand displacement reaction is favorable. This secondary encoding allows us to read out the data stored in a heterogeneous pool of registers after ligating the nicks, PCR amplifying the products, and applying next generation sequencing (NGS). The resulting NGS reads, which correspond to proportionally amplified computation products, each encodes a 4-bit value and collectively represent the output of the computation. As in regular DNA storage, this readout method is destructive; however, a small sample can be taken, leaving most of the solution intact.

### 4.2.2 Binary Counting Program

We first start with the binary counting program: beginning from arbitrary initial counts stored in different registers, each computation step increments all the registers in parallel. Compared to counting in electrical circuits at the hardware level, where complicated modules are required (a full adder requires at least 2 *XOR* gates, 2 *AND* gates and an *OR* gate—18 transistors total), binary counting in SIMD||DNA requires only 7 instruction steps independent of the size of input. Binary counting requires changing all 1s to 0 starting from

the least significant (rightmost) bit to more significant bits until the first 0, and changing that 0 to 1. All bits to the left of the rightmost 0 remain the same. As shown in Figure 4.2, the SIMD||DNA program encodes states 0 and 1 by two different sets of top strands. One extra domain is included to the right of the rightmost cell which is used to initiate displacement. Starting from the rightmost domain, the program erases all 1's in between the rightmost cell and the rightmost state-0 cell (Instructions 1 and 2), and changes those cells to 0 at Instructions 4 and 5. The rightmost state-0 cell is first marked (Instruction 3), and then changed to state 1 (Instructions 6 and 7). We previously proved the correctness of the program in our previous work [200]. Note that the binary counting program requires a strand displacement cascade (Instructions 1) and the depth of the cascade is dependent on the number of consecutive 1's to the right of the rightmost 0.

To further reduce SIMD||DNA's dependence on artificially designed long oligonucleotides as bottom strands, we chose to assemble registers using the M13mp18 single-stranded DNA plasmid from the M13 bacteriophage without modifications to the original sequence. Phosphoramidite synthesis, currently the golden standard for *de novo* synthesis of single-stranded oligonucleotides, becomes increasingly error-prone as a function of strand length. On the other hand, naturally-derived DNA is ensured to have both high fidelity and high quality DNA as a result of biological error-correcting mechanisms. The single-stranded M13 bacteriophage plasmid is a staple of DNA nanotechnology that has been widely used as scaffolds for DNA origami [159]; similarly, it could potentially accommodate computation with SIMD||DNA on several hundreds of bits. Despite these advantages, naturally-occurring DNA is typically not used in strand displacement due to the potential for undesired sequence complementarity. While artificially designed sequences can be optimized to minimize sec-

ondary structure (e.g., using a 3-letter alphabet [224], computational tools like NUPACK sequence designer [206], or other tools [224, 60]), naturally-occurring DNA may contain thermodynamically stable secondary structures that trigger undesired spurious interactions or prevent desired displacement from completing and ultimately producing incorrect computation results.

Rather than designing the sequence, we pursued the use the naturally-occurring DNA without significant sequence optimization: We screened different regions on the M13mp18 plasmid for viability by first eliminating areas with undesirable secondary structures (specifically, G-quadruplexes and hairpins [A] [159]) from consideration and then selecting 9 random addresses as candidates at which we encoded sub-registers (Figure 4.2B). We tuned the domain strength and categorized the encoded registers according to the binding strengths of some domains: weak (sub-registers 1 through 3), medium (sub-registers 4 through 6) and strong (sub-registers 7 through 9). Each category is expected to react at different experimental conditions as a result of the domain strength; for example, registers with strong binding strength are expected to require higher temperature or longer reaction time. We tested initial values 0010 and 0011 with different reaction temperatures (Figure 4.3) on these 9 sub-registers, and then picked 5 for further experiments.

We first performed SISD (single instruction single data) computation on sub-register 8 for each of the 16 4-bit initial values. All registers within each test tube contained the same initial value of sub-register 8. After NGS sequencing, reads were organized according to the barcode sequences associated with their encoded initial values, and the percentage of reads representing the correct value was calculated. More than 90% of the registers can be successfully assembled, processed, and sequenced (Figure 4.4A). After a round of binary counting

computation, sub-registers affiliated with all 16 initial values show the correct output as the dominant output (Figure 4.2C), with the minimum correct percent at 68%. We observed similar results for sub-register 3 (Figure 4.4B). We then performed SIMD computation on sub-register 8 by pooling registers with all 16 initial values in the same test tube (Figure 4.6) for computation . Figure 4.2D shows that all the initial values were updated correctly, with the minimum correct ratio at 60%. After testing different incubation temperatures (Figure 4.5), we achieved similar computation results on sub-registers 7 and 9 (Figure 4.7) at a higher temperature.

We then investigated the ability to store and compute data on multiple registers simultaneously with SIMD||DNA. We tested parallel computation on multiple sub-registers assembled on M13. Each M13 molecule was assembled with both sub-registers 7 and 9 at the temperature compatible to both (Figure 4.8). For each step of the computation, instruction strands for both sub-registers were applied simultaneously. As shown in Figure 4.2E, most registers produced the highest readcount for the correct output, with the minimum correct ratio at 15%. This reduction of yield could be due to spurious cross-talk between the instruction strands for sub-registers 7 and 9. Additionally, as the success of computation is dependent on experimental conditions, this reduced accuracy may also stem from operating at a sub-optimal temperature for each register as a compromise for compatibility.

### 4.2.3 Rule 110 Program

In addition to binary counting, we also implemented a program that simulates elementary cellular automaton (CA) Rule 110. An elemental cellular automaton [207], one of simplest models of computation, consists of an infinite set of cells with two states, 0 or 1.

At each time step, updates to a cell depend on the states of its left and right neighbors. A simple two-rule characterization of Rule 110's transition rule is as follows: 0 updates to 1 if and only if the state to its right is a 1, and 1 updates to 0 if and only if both neighbors are 1. Critically, Rule 110 has been shown to be Turing universal [40].

The SIMD||DNA program for implementing one time step evolution is shown in Figure 4.10. Theoretically, SIMD||DNA's in-memory computation model is as powerful as any other space-bounded computing technique. In other words, our space-bounded simulation of Rule 110 immediately gives that any computable function can be computed by a SIMD||DNA program, if the required space is known beforehand. Note that the Rule 110 simulation invokes two sources of parallelism: instruction strands are applied to all registers in parallel, and every cell within a register can update concurrently. This contrasts with binary counting where instruction 1 requires a cascade of strand displacement reactions across multiple cells.

To experimentally implement the Rule 110 program, we used M13 sequences as well as artificially designed sequences. Since the encoding of information 1 contains an exposed region, to enable ligation and sequencing, a set of "seal" strands were applied to all the registers after performing parallel computation on all 16 initial values to fill in the gaps on the patterns of the top strands (Figure 4.9A). We confirmed that the Rule 110 program updated correctly for the 16 registers encoded with artificially designed sequences–the correct values are the dominant output (Figure 4.9B; the control for registers without computation are shown in Figure 4.11). We achieved similar results using the native M13 sequence as seen in Figure 4.12.

### 4.2.4 Random Access

A related desired functionality for DNA data storage is to be able to selectively address or read out a specific subset of data registers, a process commonly referred to as random access. Random access avoids reading out everything at once, thereby destroying all data. Traditional DNA storage uses PCR to selectively amplify data [138] or selectively pull out information by tuning the binding affinity between sequences [13]. However, designing sequences or multiplexed orthogonal PCR probes with high specificity can be challenging. Additionally, it is necessary to reconstruct the database for information update after if a single piece of data is read. On the other hand, strand displacement achieves specificity through kinetically and energetically favorable reactions that displace a pre-existing strand. In SIMD||DNA, every register is prepared with unique barcode sequences corresponding to different initial values; these sequences can serve as a point of access for specific registers. Another feature of random access is that it allows selective erasure. Accessing data can selectively destroy a subset of the database (data erasure) but leaves the remainder available for further computation. Instead of reconstructing the database, a new, edited register can simply be added to a previously-accessed database as an update. In principle, in SIMD||DNA programs, after computation on multiple registers, displacement strands with unique barcode sequences can be added to the solution to release registers with the matching barcodes from magnetic beads. Thus, every register can be queried separately for read out from the register mix.

We experimentally demonstrated parallel computation and random access of both the Rule 110 and the binary counting programs. We show that registers can be sequentially accessed by adding a series of different displacement strands with distinct barcodes (Fig-

ure 4.13A). We mixed all 16 registers to perform Rule 110 computation. After computation, we first added a displacement strand with a barcode corresponding to 0011 and processed the displaced registers (ligation, PCR amplification, sequencing). Next, we added another displacement strand with a barcode corresponding to 1001 to query the second register. Finally, we added all 16 different displacement strands (corresponding to all 16 barcodes to access all of the information. The sequencing results confirmed that, for the first and second queries, the desired register is the dominating register among the registers displaced from the mix.

Registers can be accessed in parallel by adding different displacement strands to different register mixes at the same time (Figure 4.14). All the queries were successful and at least 23% of registers show the correct value. Accessing a register also performs selective erasure of the data. Following displacement of one specific register, we added all 14 displacement strands to displace the remaining data from the register mix. We observed that reads corresponding to the displaced register were notably less abundant compared to reads corresponding to all other registers. (Figure 4.15)

### 4.2.5 Sequential computation

Finally, we scaled up the computational power of SIMD||DNA through sequential computation. We began with the Rule 110 program (Figure 4.16A) and prepared 4 sets of register mix containing 5 distinct registers, each encoding a unique initial value. Each set went through one of the following processes: no computation, one round of computation, two rounds of computation, and three rounds of computation. After these processes, all registers from the register mix were ligated, displaced from magnetic beads, PCR amplified,

and sequenced. In the first round of computation, we confirm that all initial values included in the register mix produced the correct value as the dominant output, with the correct value encompassing at least 83% of all reads. In the second round of computation, all initial values again achieved the correct value as the dominant output, with the correct value represented in at least 34% of all reads. In the final round, all but one initial value produced the correct value as the dominant output; for this initial value, the correct value was observed in approximately 10% of all reads.

For the binary counting program, we first prepared 7 sets of register mix containing all 16 registers (Figure 4.16B, left panel). One set did not go through any computation and served as a control. The other 6 sets initially went through one round of computation. As part of another experiment, a different register was random accessed (and therefore erased) from each set (results in Figure 4.13). For 3 of the 6 sets, all remaining registers were displaced and sequenced, and the analyzed results were pooled together to account for the missing registers (Figure 4.16B, middle panel). The other 3 sets were subjected to another round of computation, followed by access of all remaining registers, post-computation processing, and sequencing. Likewise, the two-round computational results of these 3 registers were pooled in our analysis (Figure 4.16B, right panel). After the first round of computation, the correct value was represented in at least 22% of all reads for each initial value; following the second round of computation, the correct value was present in at least 12% of all reads.

To investigate the limit of multi-round computation, we quantified the amount of product remaining after each round of computation using the $C_q$ value as determined by qPCR (i.e. the number of cycles needed to detect a signal above background) and quantitative electrophoretic techniques. In qPCR, the signal strength is dependent on the con-

centration of the sample and doubles at each cycle. Thus, for two samples with $C_q$ values $C_1$ and $C_2$, the ratio of their concentrations can be calculated as $2^{C_1-C_2}$. We calculated a yield of roughly 38% for each round of computation (Figure 4.17A). To corroborate these results, we additionally used an Agilent 2100 BioAnalyzer instrument to measure product concentrations for dilutions of the computation products. We observed a similar yield with multi-round Rule 110 computation, with an average of about 28% per round. This product loss can be attributed to magnetic beads lost due to washing ($\approx$ 59% yield) and imperfect ligation (possibly from gaps resulting from incorrect computation or incomplete ligation by T4 ligase). From our analysis, we determined that approximately 70% of the product loss results from bead loss during washing, and only about 30% is caused by imperfect ligation. This indicates that the yield can be significantly improved by a better washing technique. Theoretically, using the same protocol, we can perform up to 27 rounds when storing registers with 10,000 different values (Figure 4.18).

## 4.3 Discussion

We proposed and implemented the in-memory and parallel computation architecture SIMD||DNA as a new DNA data storage paradigm. In practice, we performed in-memory and parallel computation of two programs, binary counting and cellular automaton Rule 110, on 4-bit registers, which can be constructed using both naturally existing sequences and artificially designed sequences. To demonstrate that the computational power may be scaled up, we implemented random access memory and multiple rounds of sequential computation. We investigated the completion level of some of the instructions, which finish quickly since instruction strands are added at high concentration and no slow strand displacement

mechanisms (e.g. 4-way branch migration) are involved. However, strand displacement systems can be error prone. Undesired triggering reactions (i.e. leak) can come from fraying at the nicks in the registers and undesired opening of domains, both of which may lead to strands being mistakenly displaced or binding incorrectly. The SIMD||DNA programs presented here are not robust to leak. We mitigated leak by allowing registers and instructions strands to react for a short amount of time before washing. This favors the faster desired strand displacement events while slower leak reactions are unfavored. However, in situations where undesired reactions are fast, leak can be a major source of error; this raises the question of whether leakless design principles [191, 201, 202] can be imposed on SIMD||DNA constructions.

Our method of storing information in DNA is motivated by recent developments in DNA storage employing topological modifications of DNA to encode data [186]. Although we use chemically synthesized strands to assemble registers, it is possible to programmatically cut naturally existing DNA and form strand breaks at desired locations as a high-throughput method of writing information into registers. In contrast to storing data in the DNA sequence itself, encoding data in nicks sacrifices data density but could reduce the cost of large-scale *de novo* DNA synthesis by repurposing biologically-derived DNA. Other than the approach we have taken to adapt SIMD||DNA for sequencing (i.e. including a secondary sequence encoding with mismatches and performing ligation), recently developed Nanopore sequencing methods could potentially read information encoded in nicks and single-stranded gaps directly in double stranded DNA in a high-throughput manner [125]. Registers can also be affixed to the surface of a microfluidic chip to achieve autonomous control of reacting with instruction strands and elution, which could increase both the yield and scale of computation.

Information stored in the DNA sequence has been argued to be stable for thousands of years [39]. In contrast, SIMD||DNA stores information in the pattern of nicks, and as a result, stored data may be more prone to change since it is possible that the pattern of nicks is more readily disrupted than the DNA sequence itself (e.g. via undesired 4-way branch migration between different registers). In addition to the methods used in traditional DNA data storage to increase the longevity [77, 94], it is possible to seal the nicks reversibly through light-induced photochemical ligation [44]. Our current encodings in SIMD||DNA store data at a density of approximately 0.03 bit per nucleotide, a decrease from traditional storage schemes that encode information in the DNA sequence itself for a theoretical maximum data density of 2 bits per nucleotide. In principle, data density can be increased by using different encoding schemes, such as allowing overhangs on the top strands to encode information. In our current implementation of reading out SIMD||DNA products, we use mismatches to differentiate bit information, which is orthogonal to the logic encoding. It may be possible to increase data density by encoding logic information through mismatches so that the effect of an instruction depends on the difference in binding stability or kinetics between mismatched and perfectly matched sequences.

Designing DNA strand displacement systems that can readily utilize naturally-occurring sequences is still a challenge. There are several advantages to using naturally-occurring DNA over artificially designed and chemically-synthesized DNA. First, the length and fidelity of biologically-produced DNA far exceeds those attainable by chemical synthesis [93]. With phosphoramidite synthesis, currently the standard technique for *de novo* production of oligonucleotides, oligonucleotides longer than 100 nucleotides (such as those required for SIMD||DNA registers) are likely to be truncated and consequently trigger leak reactions.

Further, if the displacing strand is truncated it may not be able to fully displace the intended target, resulting in low completion. Thus, current chemical synthesis techniques have an upper bound of oligonucleotide length under reasonable yield requirements, which limits the design of DNA architectures. Most schemes therefore avoid using oligonucleotides of lengths longer than $\approx 70$ bases, because longer strands require higher levels of purification and a different, more expensive synthesis architecture (e.g., IDT Ultramers$^{\text{TM}}$ [4]). In contrast, bacteriophage DNA is typically on the order of kilobases in length and, importantly, single-stranded. Second, the cost to produce natural DNA biologically is far lower than that of producing custom DNA synthetically. M13mp18 plasmid can be easily cultured and harvested using minimal equipment, in contrast to custom oligonucleotide that require specialized synthesizers. Special synthesis architecture and additional purification steps are often needed to produce a similar yield compared to shorter oligonucleotides, adding to both the time and financial cost of production. At time of writing, M13 plasmid can be commercially purchased at less than \$5 for 1 $\mu$g at leading suppliers, whereas a typical oligonucleotide sequence of 200 nt costs around \$40 per 1 nmole. Further, recent technology developed for DNA origami can produce *both* short single stranded staples and the long M13 to achieve production costs of around \$0.025 per $\mu$g of folded DNA origami [150]. The same technology may be potentially applicable to SIMD||DNA, with instruction strands synthesized in the same manner as the staple strands for origami.

SIMD||DNA can potentially revolutionize DNA storage architecture for future applications. Given the current challenges in attaining high-quality, large-scale *de novo* long custom strand synthesis [146, 117] and the urgent, growing need for archival data storage worldwide, SIMD||DNA presents an intermediate solution that facilitates DNA storage

for practical settings. In a longer-term context, SIMD||DNA could remain relevant as an interface between DNA computation modules that process molecular inputs and a semi-permanent record of the output of those computations. This can both scale up strand displacement-based DNA nanotechnology while adding a "wet" sensor component to otherwise "cold" data storage. Towards this end, one could for instance envision a database of personal medical records that is collected through molecular detection programs taking daily samples from the patient as input and updating corresponding registers for later readout.

## 4.4 Materials and Methods

***DNA oligonucleotides** DNA oligonucleotides were synthesized by Integrated DNA Technologies (IDT). The bottom strands were ordered as PAGE purified Ultramer DNA Oligonucleotides. The unlabeled oligonucleotides for 4-bit registers were ordered PAGE purified. The fluorophore or phosphate labeled oligonucleotides were ordered HPLC purified. M13mp18 single-stranded DNA plasmid was purchased from NEB (# N4040S).

### Register preparation

**Anneal register** The bottom strand and all the top strands were mixed and then annealed with 5% excess of top strands. The buffer for the annealing process was TE/$Na^+$ (1 M) buffer (0.04 M Tris, 1 mM EDTA, 1 M $Na^+$). The annealing process was performed in a PCR thermocycler: DNA strands were incubated at 95 °C for 5 minutes and then slowly cooled down with rate 0.1 °C/s to 20 °C°C.

### Label register to magnetic beads

The "Dynabeads MyOne Streptavidin C1" magnetic beads were purchased from Invitrogen (# 65001). The SuperMag Multitube Separator was purchased from Ocean NanoTech (# MMS-1.5-8) To resuspend the beads, they were first vortexed for 30 sec. Then 5 $\mu$L beads were transferred to a tube and washed twice with the TE/$Na^+$ (1 M) buffer. The washed beads were incubated with the annealed register (25 $\mu$L at concentration 1 $\mu$M) on a rotator for 25 min. The beads were then washed twice by the washing buffer TE/$Na^+$ (0.5 M) buffer (0.04 M Tris, 1 mM EDTA, 0.5 M $Na^+$, 0.01% Tween 20) to remove the excess register. Finally we suspended the bead with 25 $\mu$L washing buffer. The register concentration was approximately 400 nM, estimated based on bead capacity.

**Computation experiments**

For each computation experiment, 5 $\mu$L labelled registers were transferred from the above stock and mixed with other instruction strands, diluting to 25 $\mu$L with approximate concentration 80 nM. The concentrations for the strands in each instruction are: 3 $\mu$M for instruction 1, 0.5 $\mu$M for instruction 2, 0.5 $\mu$M for instruction 3, 3 $\mu$M for instruction 4, 1 $\mu$M for instruction 5, 0.5 $\mu$M for instruction 6, 0.5 $\mu$M for instruction 7. The reaction temperature for instruction 1 varied from 25 °C to 40 °C. The reaction temperature for all other instructions was 25 °C. After incubating for 10 min, the magnetic beads were washed twice by the washing buffer. The 96-well super ring magnet separator plate (SKU:T480), purchased from Permagen was used for elution.

### Post-computation processing

#### Add adaptor strand

The adaptor strand (0.5 $\mu$M) located at the rightmost side for sequencing purposes was mixed with registers and incubated for 10 min. The beads were then washed twice by 1× T4 ligase buffer to remove excess adaptor strand. The 1× T4 ligase buffer was prepared by diluting the 10× T4 ligase buffer purchased from NEB (# B0202S) and mixing with Tween 20 to reach 0.01%.

#### Ligation

400 units of T4 ligase, purchased from NEB (# M0202S) were incubated with the register at 25 °C for 10 min. The product was washed twice with the above ligase buffer.

#### Displacing bead

The displacement strand (40 nM) was mixed with the ligated product at 25 °C for 10 min. The supernatant was transferred to a new tube and inactivated by heat for 10 min at 65 °C.

#### Library preparation

Prior to amplification, the displaced product was quantified by qPCR with the LightCycler96 instrument (Roche). Reaction mixtures contained 2.5 nM of the displaced product, 500 nM each of forward and reverse PCR primers containing NGS adaptors and unique barcodes, 400 $\mu$M dNTP, 1× EvaGreen intercalating dye (Biotium #31000), 0.4 U/$\mu$l Q5 DNA polymerase (NEB #M0491S), 1× Q5 Reaction Buffer (NEB). qPCR was performed on the samples using the following protocol: initial melting at 98 °C for 3 minutes, followed by 30 cycles of amplification with melting at 98 °C for 30 sec, annealing at 67 °C for 30 sec, and extension at 72 °C for 30 sec (measurement taken), followed by a final extension at

72 °C for 3 minutes (measurement taken). Once the $C_q$ of each sample was determined, PCR was repeated using the same thermocycling protocol in a thermocycler with the same concentrations, barcoded primers, and protocol as described for qPCR, except EvaGreen dye was replaced with nuclease-free water and the number of cycles was set to $C_q + 5$ for each sample to minimize the amplification of side products. After PCR, equivalent amounts of each sample were pooled together and gel purified for the expected size after running on a 1.8% NuSieve GTG agarose gel (Lonza #50081) using a QIAquick PCR & Gel Cleanup Kit (Qiagen #28506) as per manufacturer's instructions for gel purification with the following exceptions: gel fragments were incubated in Buffer QG for at least 20 minutes at 60 °C (instead of 10 minutes at 50 °C), and the column containing product was washed 3 times using Buffer PE (instead of once). The final samples were eluted in nuclease-free water and diluted to a concentration of 5 ng/ul as measured by Nanodrop.

**Next-generation sequencing**

Sample libraries were sequenced for 2x261 cycles using Illumina MiSeq 2x250 paired end reagent kits (v2). Because SIMD products exhibit very low base diversity (i.e. strands are very likely to have the same base composition at any given position within the target sequencing range), it is necessary to boost base diversity to avoid downstream analysis issues. We added a genomic DNA sample library (approximately 50% of all reads) on any runs in which SIMD products accounted for more than 30% of all reads.

**Sanger sequencing data analysis**

We included sequenced library prepped, single data computation products using both forward and reverse primers to gain confidence on the base call results and to maximize the

portion of the computation product with high quality base calls. Sanger sequencing traces were mapped to the expected SIMD product sequence using the "Map to Reference" feature in Geneious 2020.0.5. We determined the computation results using the composition of the base call at the nucleotide positions of interest. Although Sanger sequencing is generally used for discrete base calls (i.e. "A/T/C/G"), mixed populations can be detected when a position has more than one visible nucleotide. Because we expected single data SIMD products to have a mix of two possible nucleotides at each mismatch position, we interpreted the height of base call peaks in the raw trace to be representative of the relative proportions of each base in that population.

**Next-generation sequencing data analysis**

Next-generation sequencing was performed with the Illumina MiSeq V2 paired-end platform with 2x261 cycles. All data analysis was performed using Python. Each register sequence contains 4 cells, each of which contains a single nucleotide position the determines the bit value for that cell, or the "variable nucleotide position". In contrast, the sequences between consecutive variable nucleotide positions are expected to be constant regions, as no mutations are expected in these regions other than those arising from synthesis, PCR, or sequencing errors. An initial filter was applied to the raw reads such that reads with at least 3 consecutive constant regions, each with a maximum of 1 mutation (indel or substitution), were considered viable for analysis. If one read in a paired set of reads satisfied the criteria, its partner reads would also be included regardless of whether it passed the filter. Viable reads were then matched to its sample and initial value by identifying its barcodes. Reads with sample barcodes that contained no more than 2 mutations to the expected barcode

*and* that contained no more than 1 mutation in the register (i.e. initial value) barcode were included in the final analysis.

To read out the results of SIMD computation, each read in a qualified read pair was locally aligned to each expected cell sequences using pairwise2 from Biopython with match, mismatch, gap opening, and gap extending scores of 1, -0.5, -0.5, and -0.5, respectively. If the aligned nucleotide at a variable nucleotide position neither matched the original sequence nor was "G" in the forward read or "C" in the reverse read, as any SIMD-related sequences changes would result in an "N" $\rightarrow$ "G" mutation in the forward strand, the corresponding digit would be marked as undefined for that read. Finally, for each of the four digits, the read in the pair with the greater read quality score at the variable nucleotide was used to call the bit.

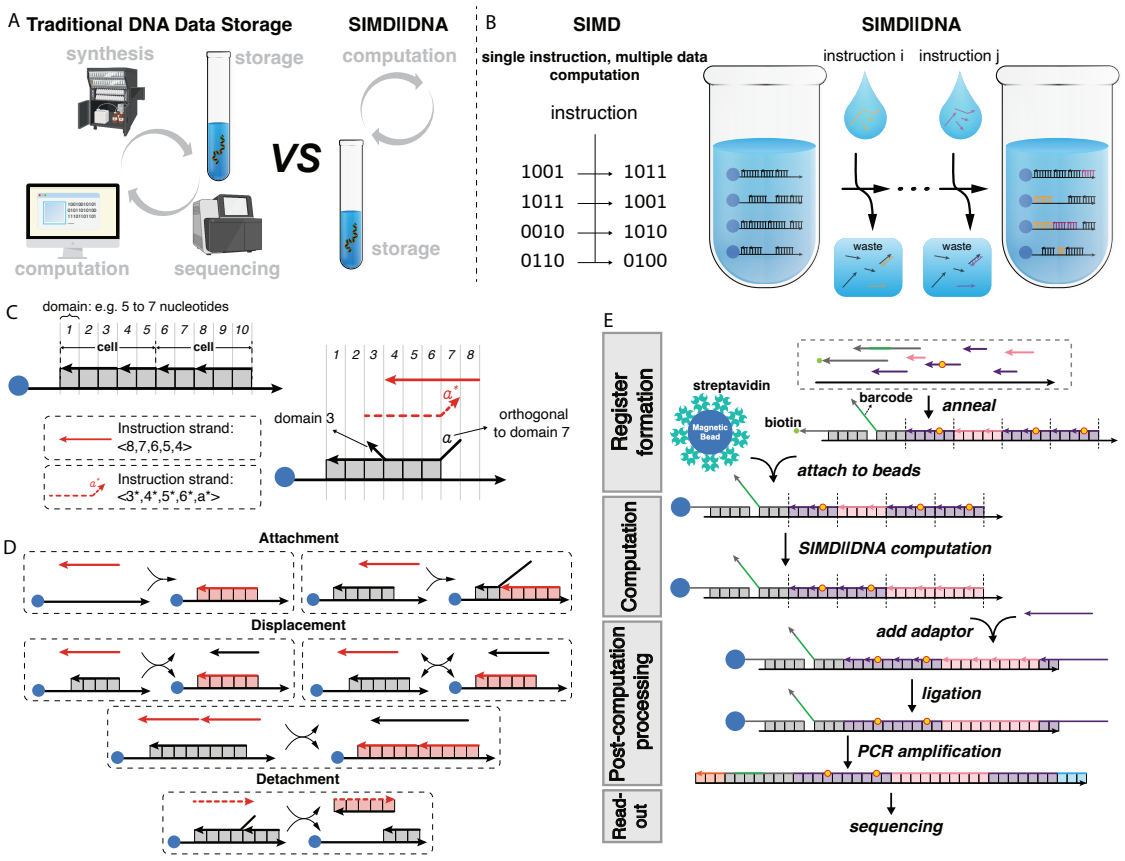


Figure 4.1: **Overview of SIMD||DNA.**

(A) Computation in traditional DNA storage paradigm relies on outsourcing computation processes to classical computer with additional required steps of sequencing and synthesis. The SIMD||DNA paradigm allows in-memory computation performed through DNA strand displacement reactions. (B) Analog to the single instruction multiple data (SIMD) computation in classical computer which enables processing multiple data by one single instruction, the SIMD||DNA paradigm can also perform parallel computation on multiple registers simultaneously. Each DNA register is a multi-stranded complex. Different information is encoded in the pattern of nicks and exposed single-stranded regions in the register. Registers are attached to magnetic beads (blue). At each instruction step, a set of instruction strands is added to the solution to react with all registers in parallel. Next, waste species (i.e. unreacted instruction strands and displaced reaction products) are washed away. After a series of sequential reaction and washing steps, the information stored on the registers is updated. (C) The notations for SIMD||DNA. Domains are represented by square boxes. We indicate complementarity of instruction strands to register domains by vertical alignment. If a domain label is given explicitly (e.g. $a$ and $a^*$), the domain is orthogonal to the other vertically aligned domains A strand can be described by listing the constituent domains in a bracket <> from 5'-end to 3'-end. Strands with solid lines are complementary to the corresponding domains in the bottom strand. Strands with dashed lines are complementary to the corresponding domains in the top strand. A dashed instruction strand indicates the domains in the instruction

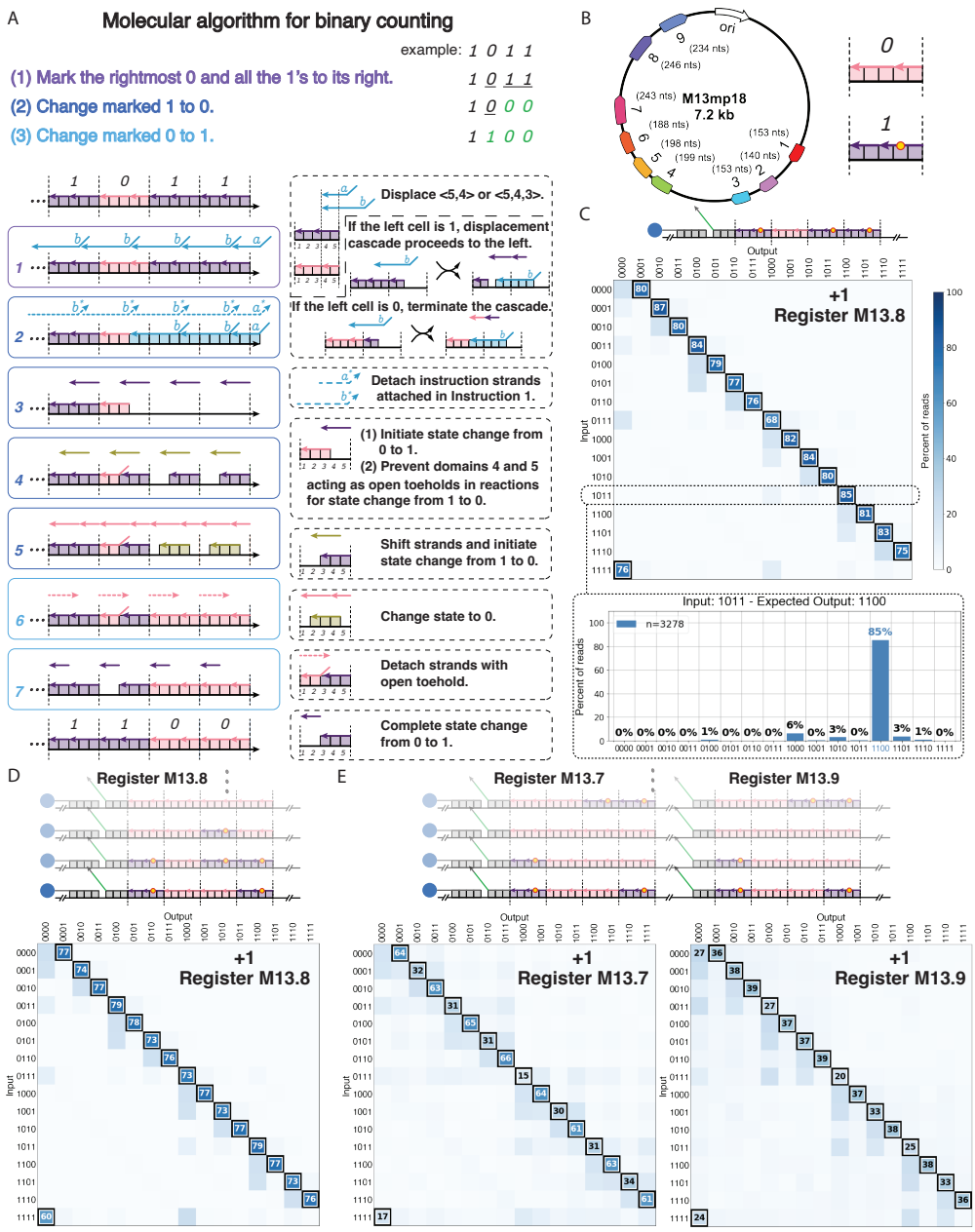


Figure 4.2: **Binary counting program on naturally-occurring sequences.**

(A) Molecular program implementing addition by 1 of a binary string on an example register. The top register shows the initial state of each cell. After 7 instructions, the register updates to the state shown at the bottom. Strand colors have three information categories: state 1 (purple), state 0 (pink), intermediates (other colors). Solid boxes show the instruction strands and the state of the register before the strands are applied. Dashed boxes explain the logical operation of the instructions. The overhang domains $a$ and $b$ are orthogonal to their vertically aligned domains. (B) (left) Locations of registers on the M13mp18 phagemid. (right) Mismatches (labeled as yellow dot) are introduced in top strands representing state 1. (C) Single data binary counting on register M13.8. For each initial value, the distribution of the output values are represented in the heat-map matrix. Lower bar plot shows an example of the data in one row of the heat map: the distribution of output values on reads associated with initial value "1011". (D) Multiple data Binary Counting


| | 1st Digit | | 2nd Digit | | 3rd Digit | | 4th Digit | | Expected: 0011 | 1st Digit | | 2nd Digit | | 3rd Digit | | 4th Digit | | Expected: 1000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fwd | Rev | Fwd | Rev | Fwd | Rev | Fwd | Rev | | Fwd | Rev | Fwd | Rev | Fwd | Rev | Fwd | Rev | |
| 1 | G | G | T | T | C | C | T | T | 0010 | N | N | T | T | A | A | T | T | ?000 |
| 2 | A | A | A | A | C | C | G | G | 0010 | A | A | C | C | C | C | C | C | 0111 |
| 3 | T | T | A | A | C | C | C | C | **0011** | C | C | A | A | A | A | T | N | **1000** |
| 4 | G | G | A | A | C | C | C | C | **0011** | G | G | C | C | A | A | A | A | 0100 |
| 5 | T | T | A | A | C | C | C | C | **0011** | N | T | N | N | T | T | A | A | 0?00 |
| 6 | T | T | A | A | C | C | A | A | 0010 | T | N | N | N | C | C | A | A | 0?10 |
| 7 | T | T | T | T | C | C | C | C | **0011** | N | N | T | T | T | T | G | G | ?000 |
| 8 | A | A | A | A | C | C | C | C | **0011** | C | C | A | A | A | A | A | A | **1000** |
| 9 | A | A | T | T | C | C | C | C | **0011** | N | A | N | N | N | N | A | A | 0??0 |

Figure 4.3: **Sanger assessment of M13 Register addresses for 0010 and 0111.** Mismatches to the native sequence as determined by Sanger are marked by a yellow circle. Computation products were sequenced both in forward and in reverse to maximize high-quality coverage of product and to improve confidence in base calls. Fwd = Forward read, Rev = Reverse read. Digits at which two bases show peaks of similar height are marked with a"?".

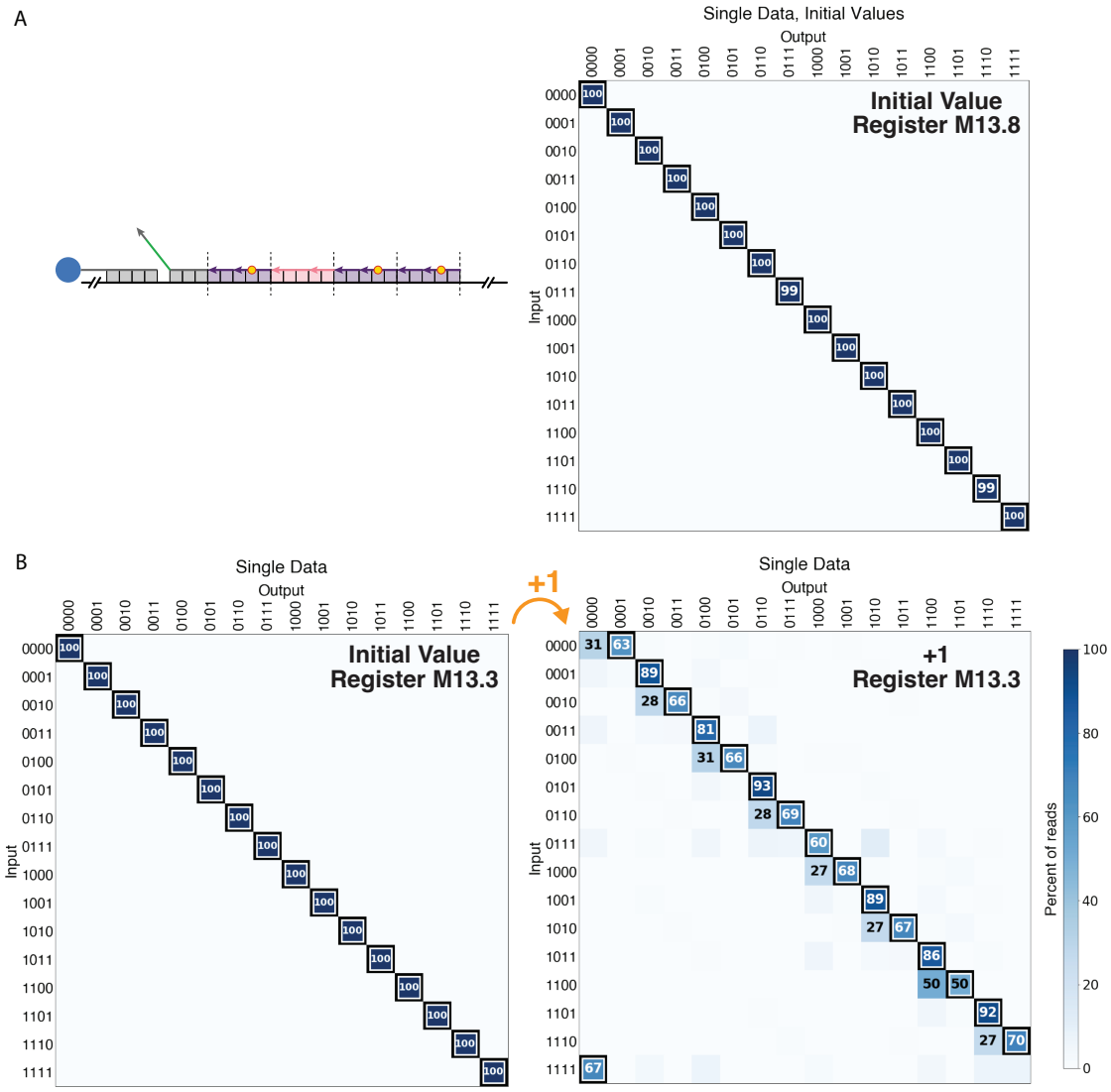


Figure 4.4: **SIMD||DNA single data binary counting program using M13 sub-registers 3 and 8.**

(A) Independent assembly of initial values on M13 sub-register 8. (B) Independent assembly of initial values on M13 sub-register 3 (left) and single data binary counting (right).

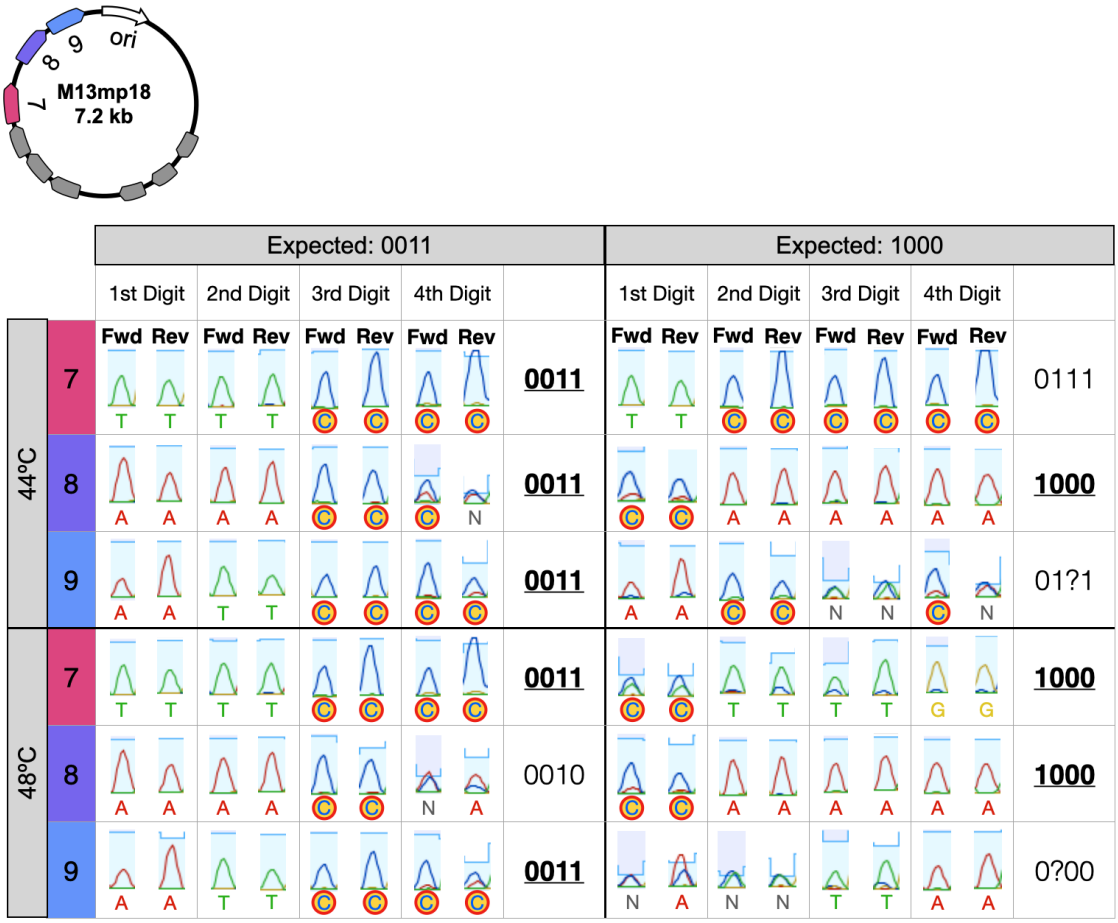


Figure 4.5: **Sanger assessment of different instruction temperatures for binary counting on M13 sub-registers 7, 8, and 9.**

Mismatches to the native sequence as determined by Sanger are marked by a yellow circle. Computation products were sequenced both in forward and in reverse to maximize high-quality coverage of product and to improve confidence in base calls. Fwd = Forward read, Rev = Reverse read. Digits at which two bases show peaks of similar height are marked with a"?".

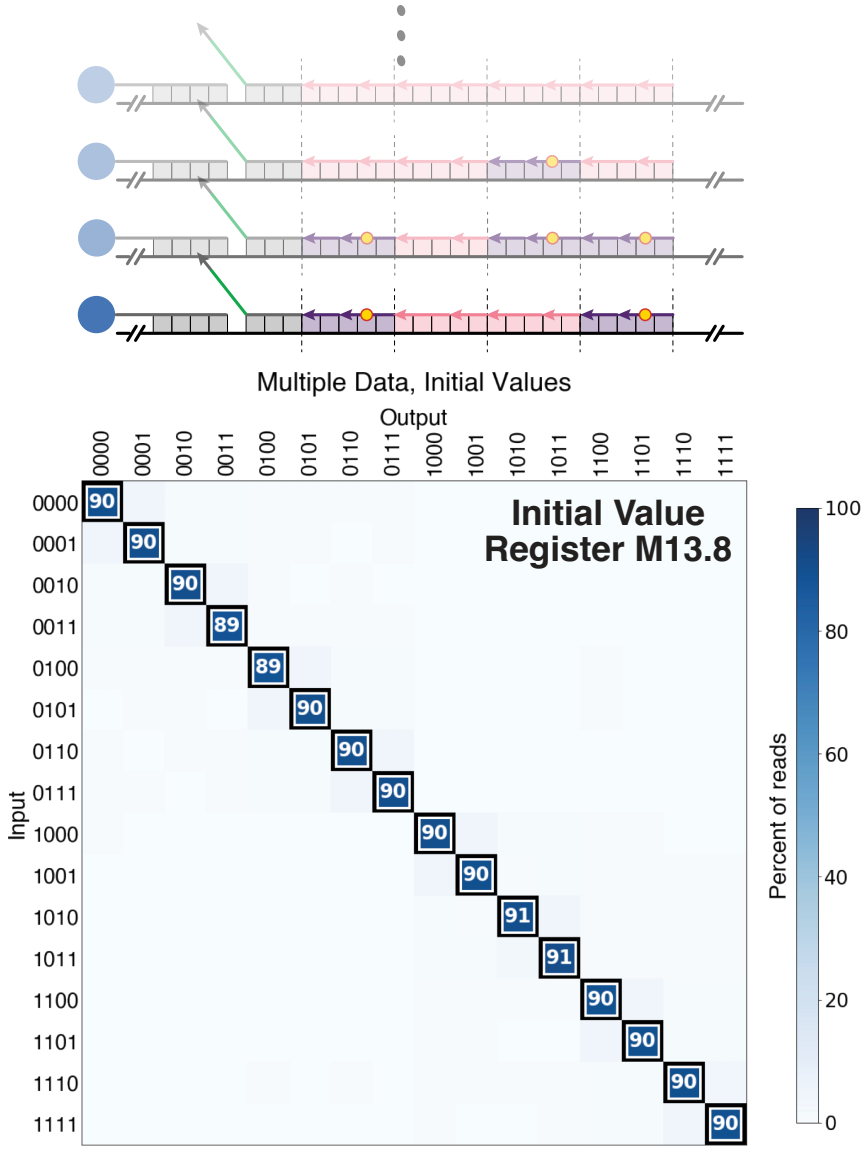


Figure 4.6: **Multiple data readout of independently assembled initial values on M13 sub-register 8.**

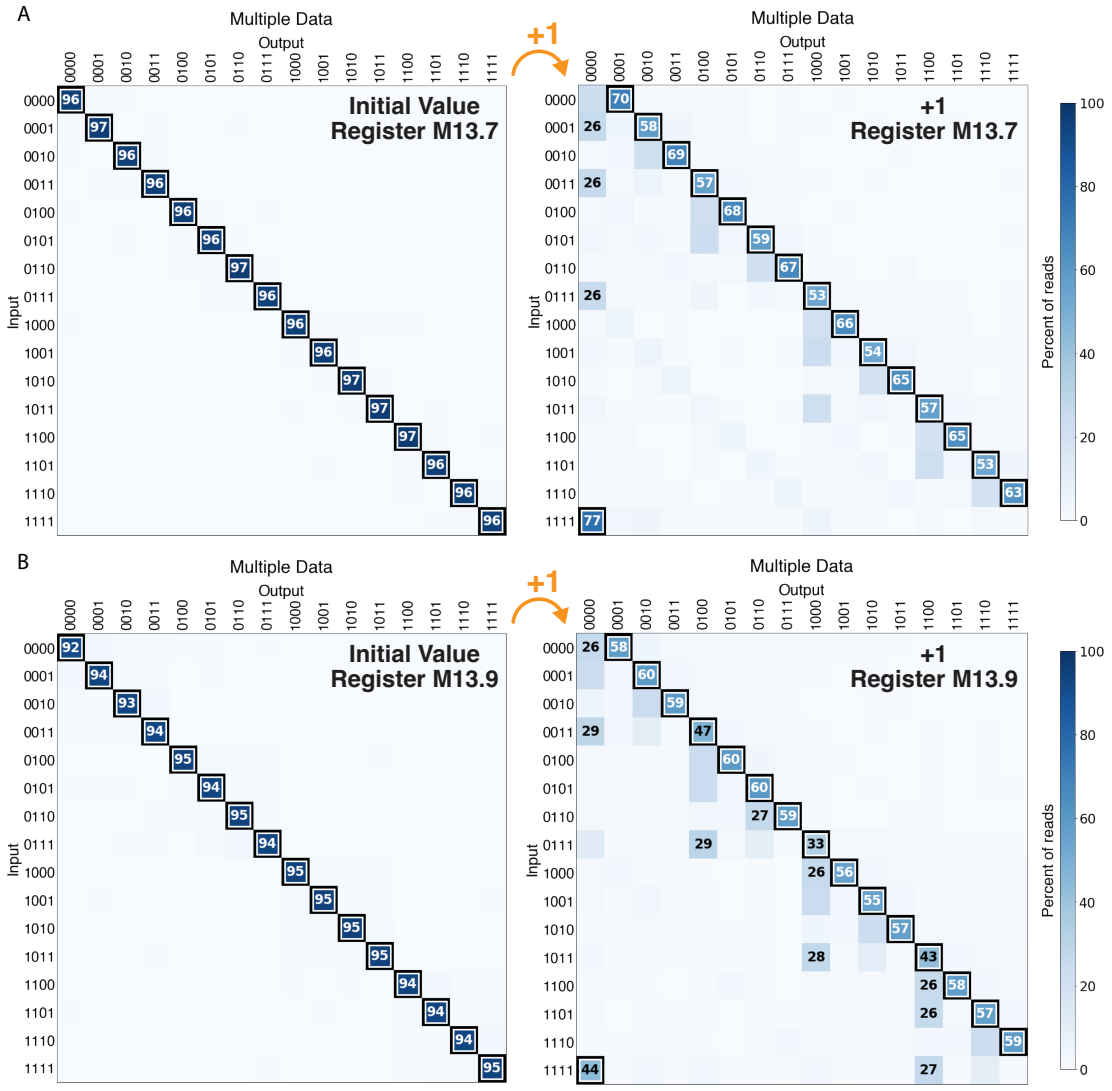


Figure 4.7: **Multiple data binary counting on M13 sub-registers 7 and 9.** (A) Initial values were independently assembled on M13 sub-register 7 and mixed together (left), then binary counting was performed on the register. (B) Initial values were independently assembled on M13 sub-register 9 and mixed together (left), then binary counting was performed on the register.

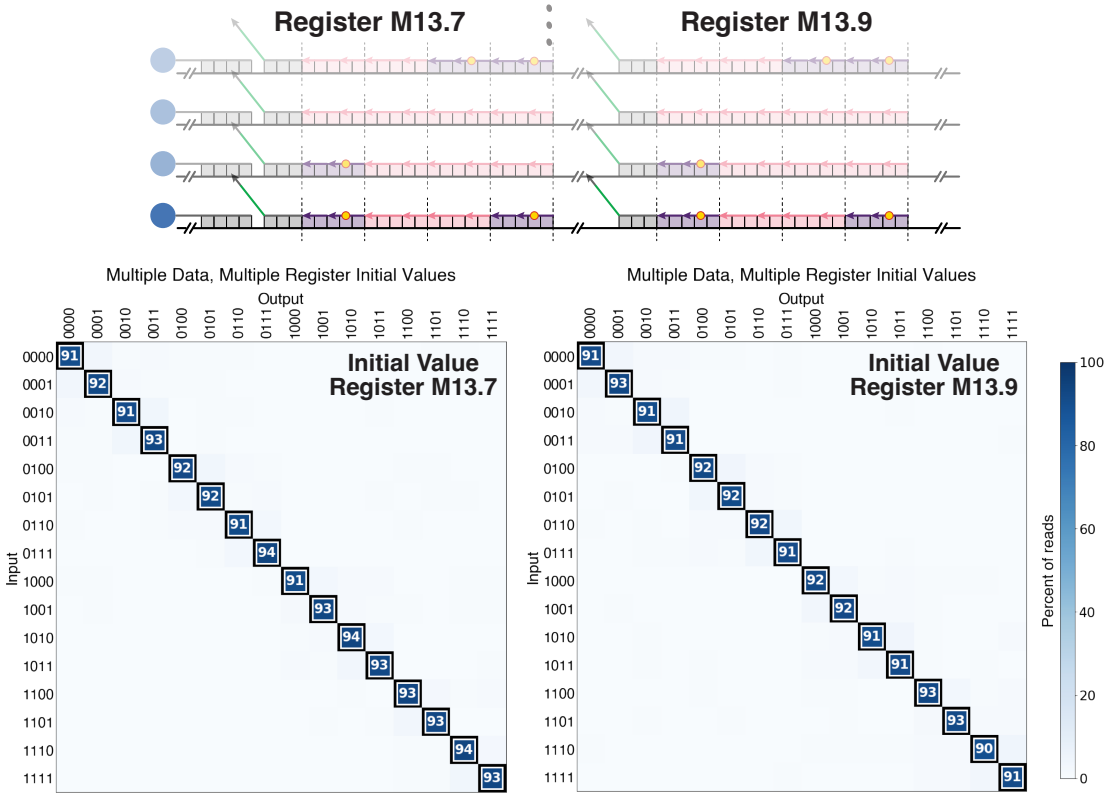


Figure 4.8: **Assembly of initial values on M13 sub-registers 7 and 9 on the same M13 plasmids.**

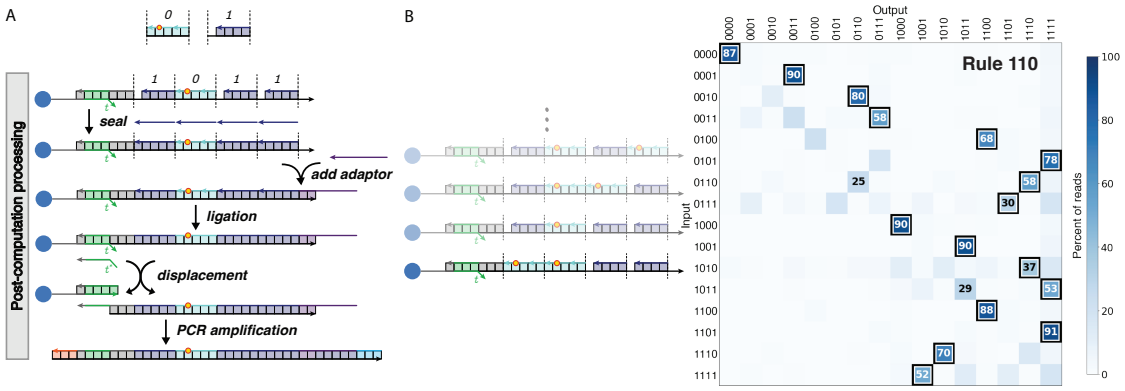


Figure 4.9: **Rule 110 computation with chemically synthesized DNA.**

(A) Post-computation process for the Rule 110 program with chemically synthesized DNA. After computation, a set of "seal" strands are added to the register to fill in the gap for cells representing bit 1 for the following ligation step. (B) Multiple data Rule 110 computation on 16 registers with unique initial values. The correct output value is indicated by a white and black border; values that appear in $> 25\%$ of all reads for a given sample are marked by text.

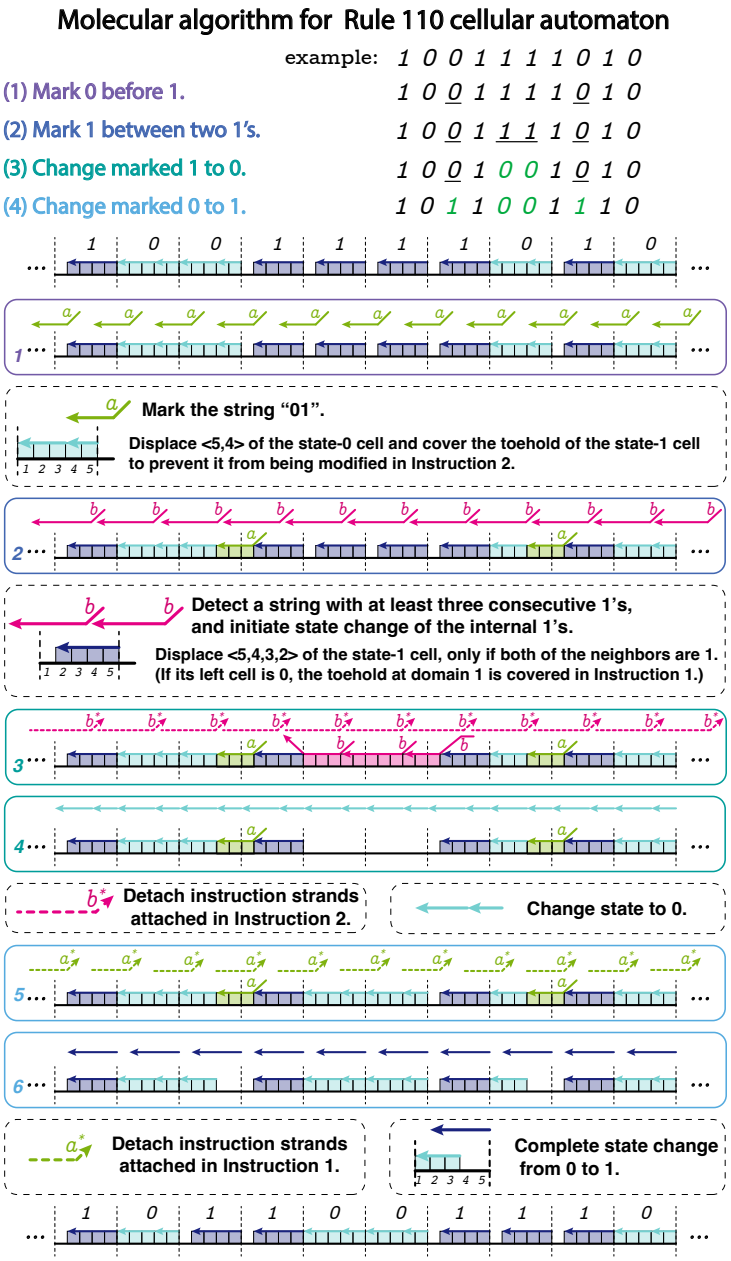


Figure 4.10: **Program implementation of one timestep of Rule 110 shown on an example register.**

The top register shows the initial state of each cell. After 6 instructions, the register updates to the state shown at the bottom. Strand colors have three information categories: state 1 (dark blue), state 0 (light blue), intermediates (other colors). Solid boxes show the instruction strands and the state of the register before the strands are applied. Dashed boxes explain the logical meaning of the instructions. The overhang domains $a$ and $b$ are orthogonal to their vertically aligned domains.

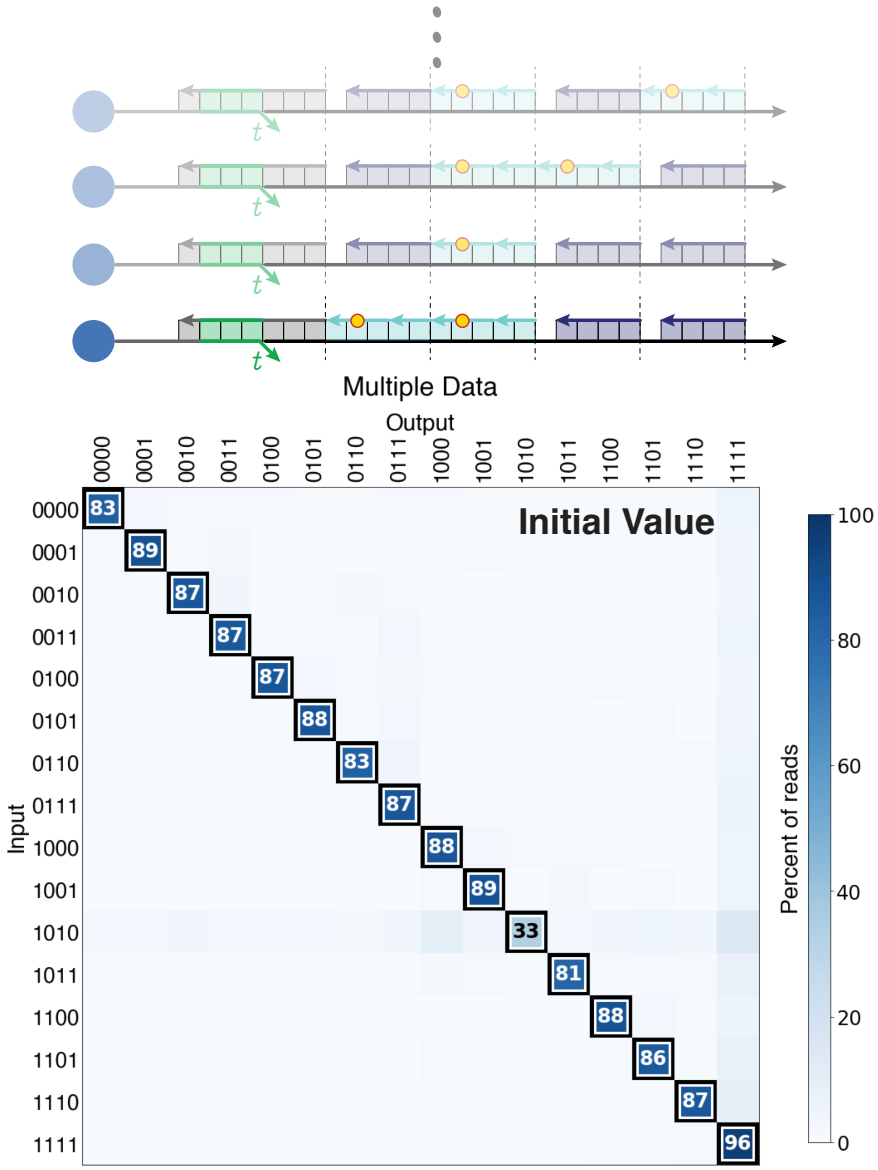


Figure 4.11: **Readout of initial values assembled on chemically synthesized oligonucleotides designed register sequence prior to the computation done in Figure 4.9.**

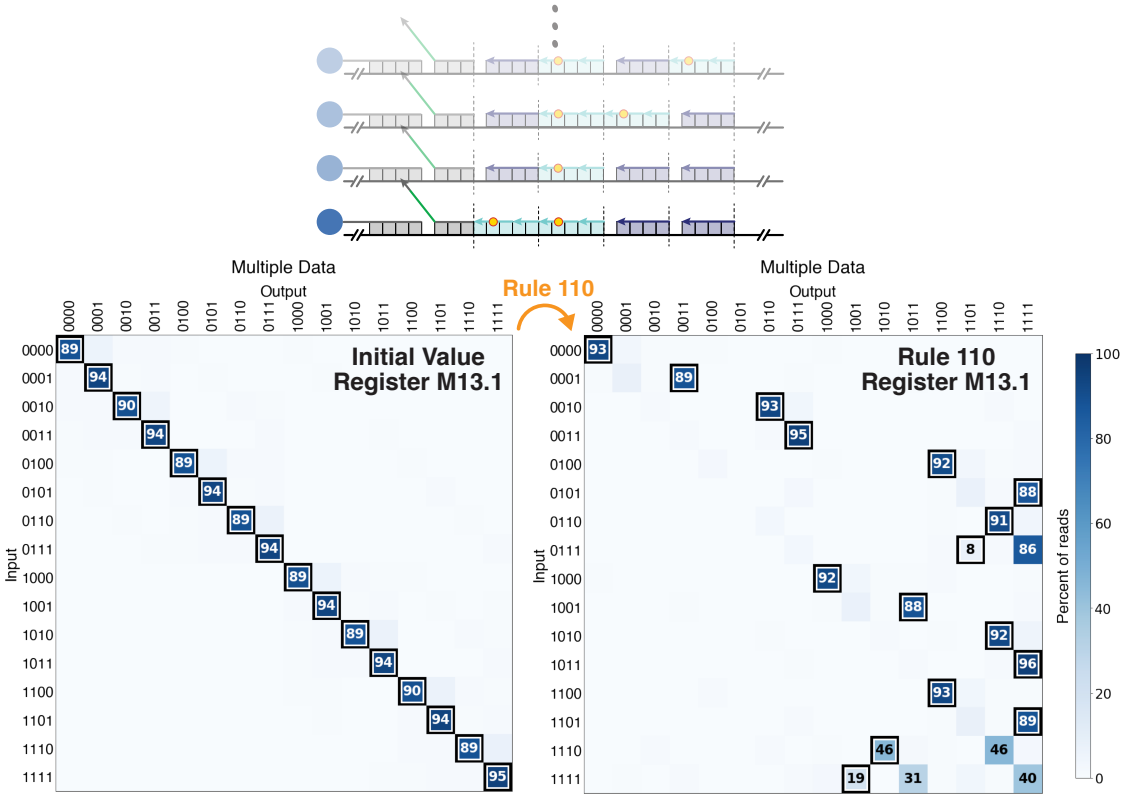


Figure 4.12: **Rule 110 computation on M13 sub-register 1.**

Readout of combined initial values (independently assembled) is on the left; multiple data computation is on the right

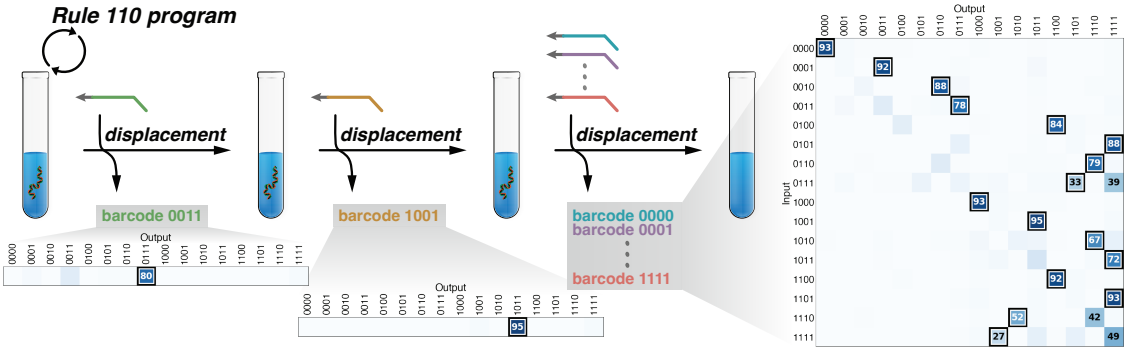


Figure 4.13: **Random access with chemically synthesized DNA.**

Sequential random access for the Rule 110 algorithm. Following Rule 110 computation, registers with initial value "0011" were accessed first (top), "1001" second (middle), and all remaining values last (bottom).

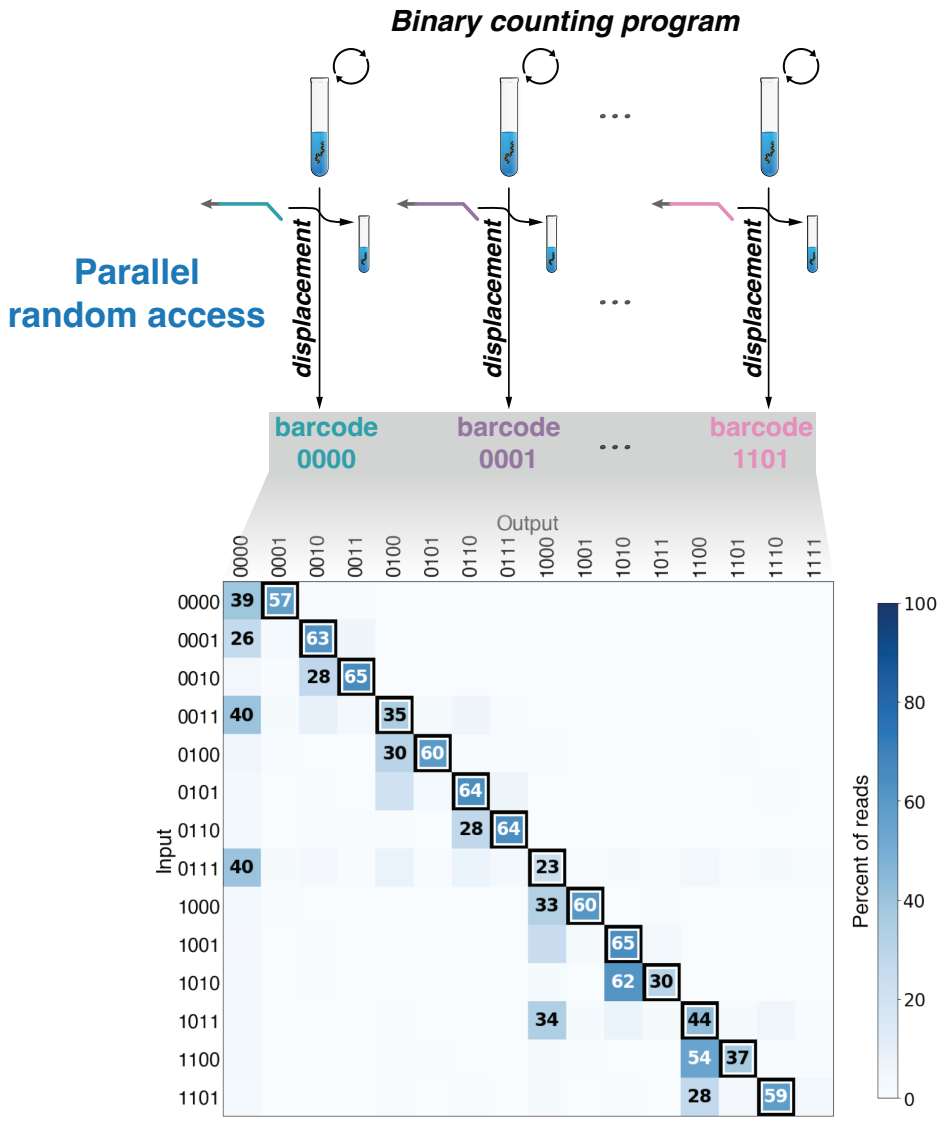


Figure 4.14: **Parallel random access for the binary counting algorithm.**
Computation is performed independently on multiple samples, after which a unique initial value is accessed from each sample. 14 initial values (0000 to 1101) were accessed in parallel following one round of binary counting. In parts B and D the correct output value is indicated by a white and black border; values that appear in >25% of all reads for a given sample are marked by text.

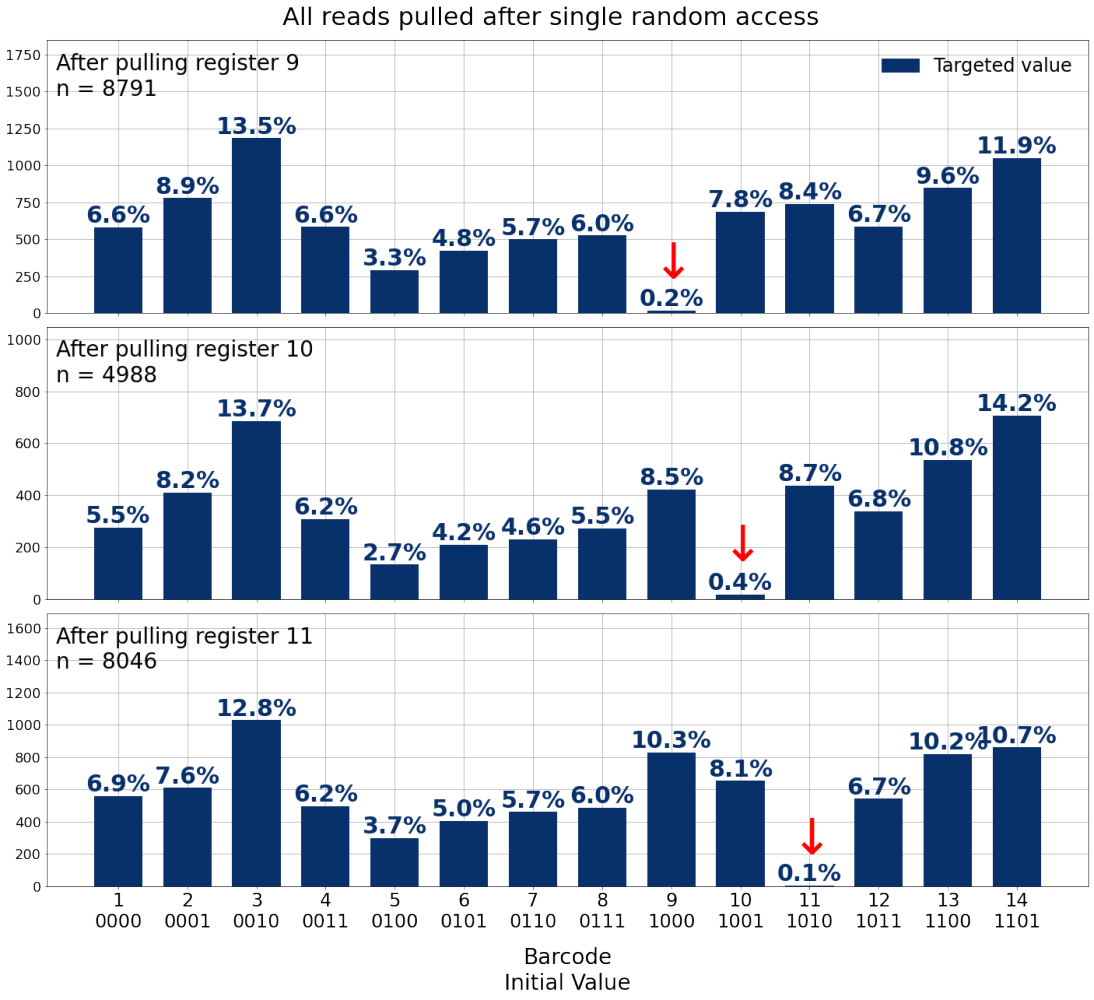


Figure 4.15: **Data erasure by random access.**

Red arrow indicates barcoded register that was previously accessed.

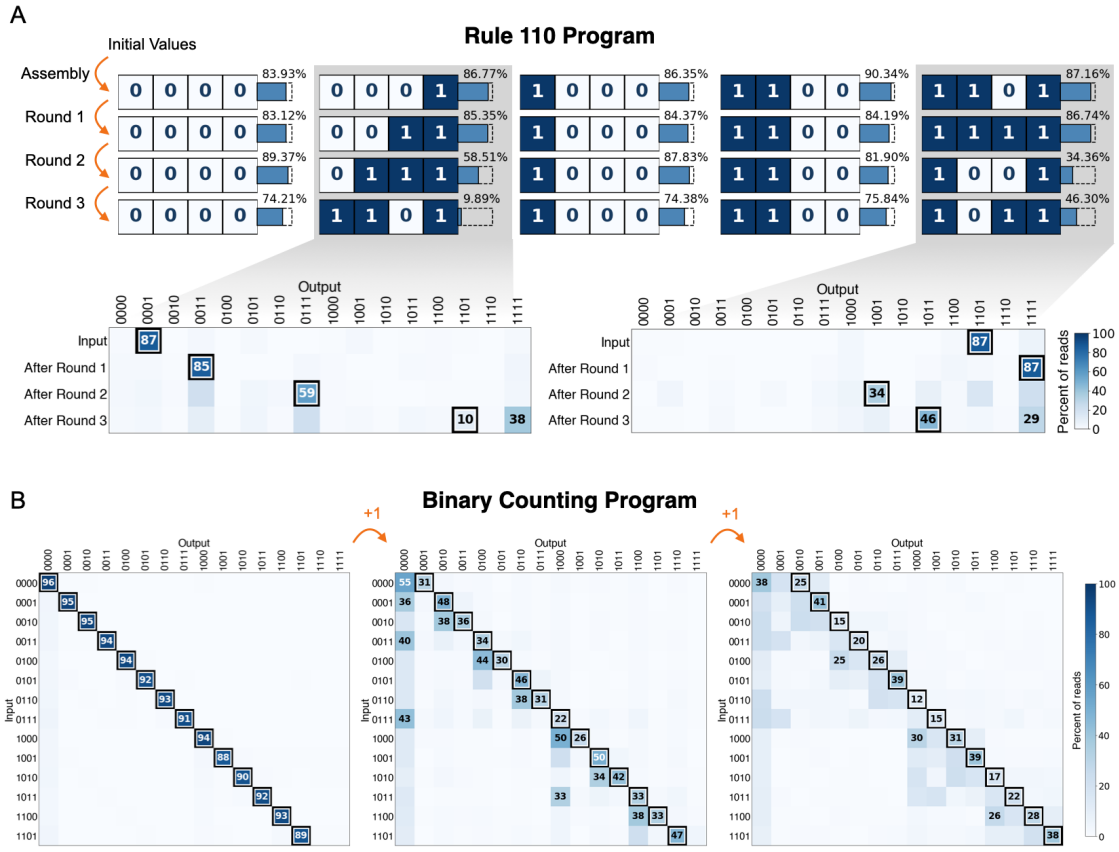


Figure 4.16: **Multiple rounds of sequential computation with chemically synthesized DNA.**

(A) Sequential computation of the Rule 110 program. Results are normalized to the total read count for each sample. Reads with one or more indeterminate digits were excluded. Lower panels show the distribution of outputs values for initial values "0001" and "1101". (B) Sequential computation of the binary counting program. In both the lower panels of (A) and (B), the correct value is indicated by a white and black border; values that were observed in $> 25\%$ of all reads are labeled.

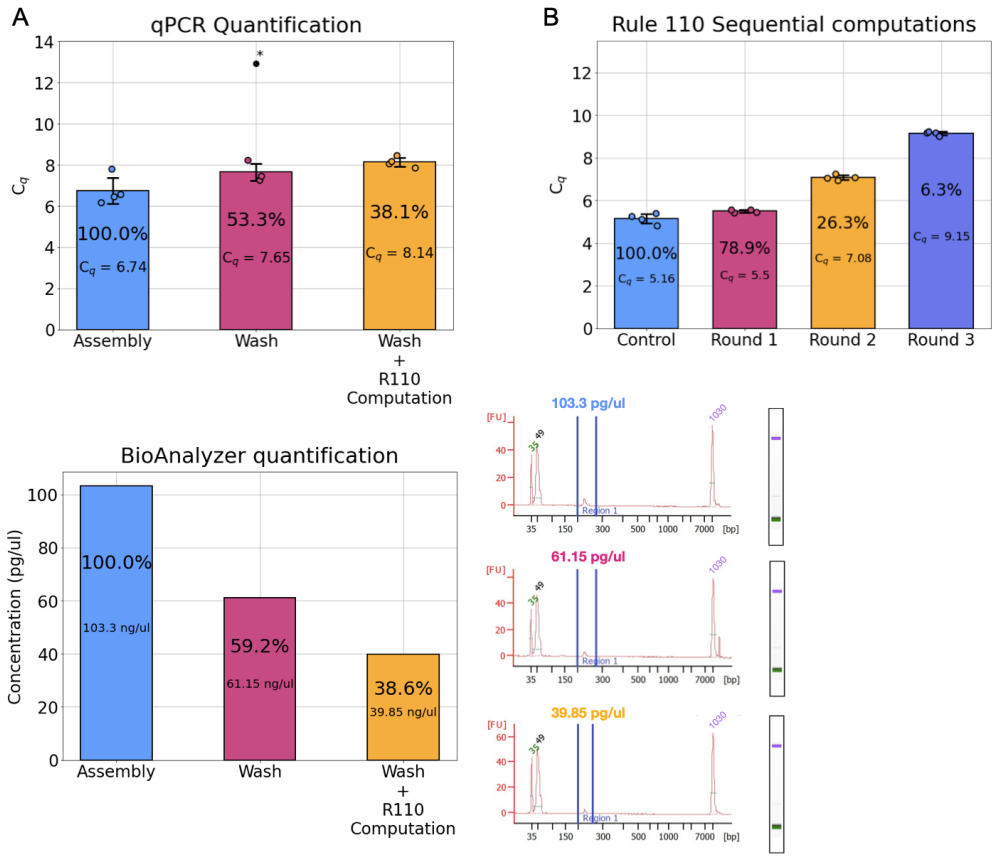


Figure 4.17: **Quantifying the loss of SIMD products following washing and computation steps.**

(A) SIMD products from the Rule 110 algorithm were quantified by both qPCR (top) and electrophoresis (bottom). The calculated percent yield is shown in text, as well as the $C_q$ and concentration as determined by qPCR and the BioAnalyzer, respectively. (B) Yield quantified by qPCR for Rule 110 sequential computation. In combination with the results from (A), each round of computation resulted in about 60% to 75% product loss. The value in the parentheses is the percent of product detected relative to the control (aka assembly) as described in (A). Note that the concentration of displacement strands and washing procedures are slightly different than in (A), which could account for the discrepancy in the yield from control/assembly to the first round of computation.

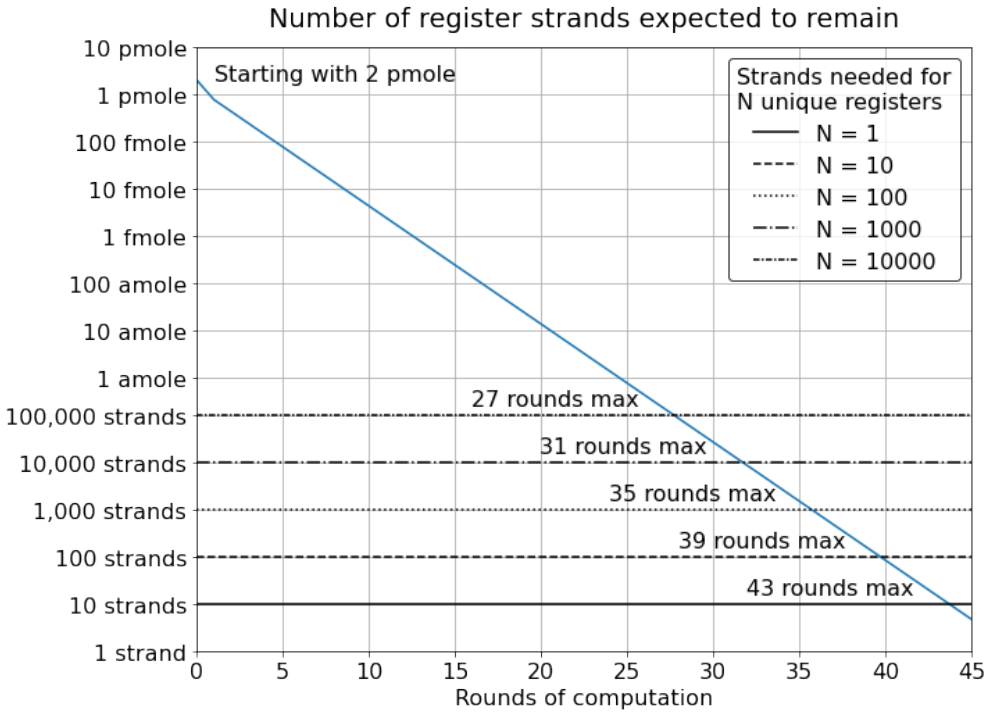


Figure 4.18: **Theoretical maximum rounds of computation possible for storing various numbers of unique registers.**

We made the following assumptions for the calculation: (1) Each round of computation except the final round has a yield of 56.25% (since products are only washed and not ligated). (2) The final round of computation has a yield of 38.25% due to ligation. (3) The reaction volume is 25 $\mu L$ and the starting total register concentration is 80 nM (i.e. starting with 1.2e+12 strands or 2 pmoles). (4) PCR is capable of amplifying as few as 10 copies of each register in the reaction volume. (5) There must be at least 1 copy of each unique register to determine data.

# Chapter 5

# Reading out in vitro transcription networks with high-throughput sequencing[1]

**Abstract.** Synthetic *in vitro* transcription networks have recapitulated complex and dynamic behaviors found in biological systems, such as oscillations and bistable switching, with minimal machinery. These networks have both elucidated principles for building artificial biochemical networks and demonstrated the computing capabilities of *in vitro* transcription regulatory elements. Here, we expand on the scalability and toolkit of transcription networks by modifying our previous single-stranded transcription switch and developing an associated protocol to read all signals, including intermediate signals, using next-generation sequencing. Additionally, we present a single-stranded transcription switch that activates upon binding by a cognate signal strand.

[1]This chapter includes original work by SSW. SSW received funding for the project from Andrew Ellington. SSW would like to thank Shaunak Kar for helpful discussions.

## 5.1 Introduction

Synthetic biology aims to create biochemical circuits to address a broad range of applications from disease diagnostics to biosynthesis of precious compounds. At their core, these circuits process chemical information, and engineered genetic circuits have been demonstrated to achieve computational tasks from simple boolean logic to molecular pattern recognition, as well as dynamic behaviors. Key challenges towards this goal are the modularity and predictability of circuit components. Nucleic acids adhere to predictable Watson-Crick base pairing rules [161, 62] and well-studied kinetics [225, 124], making DNA and RNA programmable substrates that can be used to implement minimal synthetic biochemical networks. This has enabled DNA nanotechnology for molecular detection [118], computation [152, 153, 34], and nanoscale actuation [172, 213] to be built *de novo*.

Despite their modularity and programmability, nucleic acid-based circuits suffer from issues of scalability and broad use. First, if the goal is to interface with biological systems, a protein, mRNA, or other biologically potent output must be produced or unlocked by the computation. However, conventional DNA computing solely involves DNA oligonucleotides and their hybridization reactions, making it difficult to actuate biological responses. To address this, previous works have explored the regulation of transcription [110, 167, 38] and translation [81, 80] through strand displacement reactions that result in the synthesis of new RNA strands *in vitro* or a phenotypic response *in vivo*. Second, most nucleic acid circuits involve multi-stranded complexes, which require time-consuming stoichiometric annealing and purification prior to use. This may become impractical as the number of total components scales. Some transcriptional circuits have addressed this issue by using single-stranded switch elements [104] or inserting self-cleaving ribozymes [168], both of which ensure a one-to-one

assembly of components. Third, fluorescence is typically used for immediate, real-time readout. However, this mode of readout has limited ability for multiplexed output. Stochastic photoswitching presents one solution to this challenge [103], while alternative readout methods such as sequencing (as presented in Chapter 3) could enable unlimited monitoring of all components involved.

Here we adapted our previously reported single-stranded transcription switch for quantitative, multiplexed readout using next-generation sequencing (NGS). We updated the transcribed RNA signals such that all products included in a network may be read and identified using RNA-Seq. To expand the toolkit of transcription switch elements, we additionally present a single-stranded transcription switch that is activated by hybridization with a cognate DNA signal. Our results present alternatives to conventions used in DNA computing that could improve the scalability of rationally designed nucleic acid circuits.

## 5.2 Results

### 5.2.1 High-throughput readout of *in vitro* transcription networks

The transcription hairpin is a hemi-duplex hairpin that acts as a transcription template for T7 RNA polymerase (T7 RNAP) conditional on the state of its promoter sequence. In the active state, it consists of a double-stranded T7 promoter sequence region, a single-stranded DNA loop, and a single-stranded templating region [104]. In the inactive state, an upstream signal strand binds to the loop and partially binds to the T7 promoter sequence, thereby separating the top and bottom strands of the promoter and prohibiting initiation by T7 RNAP. The hairpin switch is responsive to both single-stranded DNA and RNA signals

that include a complementary sequence to the hairpin loop region (i.e. cognate signals). It can implement NOT and NAND logic on inputs and transcribe an RNA signal as output. We modified our previously reported hairpin switches to be compatible with sequencing readout by extending the template region to include a signal-specific barcode [85] and a common reverse transcription priming site (Figure 5.1A). Following transcription, DNA templates and input signals are removed from the samples using DNase I, the treated samples are reverse transcribed, and signals are read out with qPCR (for singleplex signals) or PCR amplified and read out with NGS (for multiplex signals) (Figure 5.1B). This process can be scaled up through sample-specific barcodes that are introduced with PCR primers. To test these modifications, we designed several sequencing-compatible switches with different loop (input) and transcribed (output) signal sequences using NUPACK [218]. We confirmed transcription of signal RNAs at the expected size in the absence of the inhibiting signal by PAGE (Figure 5.2).

The transcription of the modified switch can be similarly observed using NGS, qPCR, or fluorescence. We designed a hairpin switch (H2) encoding the Spinach RNA aptamer [149] as well as an upstream hairpin (H1) that produces the cognate inhibiting RNA signal. The transcription activity of H2 in the absence or presence of upstream DNA signal (Inh2) or H1 was measured as the fluorescence signal. In the presence of Inh2 at 2X concentration, the transcription activity of H2 ws at background, and in the presence of H1 at 2X concentration, transcription was reduced by about 60% (Figure 5.3A). In parallel, we designed a sequencing-compatible version of the hairpin and performed qPCR quantification with the $C_q$ as a readout of relative concentrations. In the presence of the upstream inhibitor signal (OFF state), transcription activity was decreased to 37% of the ON state (Figure 5.3B). Finally, we

used NGS to measure the output, with the read count as the signal. We observed a graded response over a range of Inh2 concentrations. The maximally inhibited case (i.e. 400 nM of inhibitor) showed about 30% activity as compared to the uninhibited activity (Figure 5.3C). One reason for the diminished inhibition is that the concentration of templates in these assays exceeded the saturation concentration for T7 RNAP. Operating at a higher concentration of T7 RNAP should result in an improved response that is more responsive to changes in active template concentration.

NGS readout allows the signals of all switches in a network to be read out at once. We constructed a network of three inhibitors in series and observed their individual activities in response to increasing concentrations of Inh1 (Figure 5.4A). As expected, the activity of the first switch (H1) showed the largest dynamic range. The second switch (H2) showed the expected relative activities for up to 200 nM of Inh1. Beyond 200 nM, however, H2's inhibition response was no longer observed; this was explained by the saturation of H1 to input concentrations above 300 nM. We additionally used NGS to observe the transcription activities of two switches in series in response to varying switch concentrations (Figure 5.4B). As the signal for H1 (Inh2) increases with increasing concentrations of H1, inhibition of the downstream H2 switch saturates at around 50%. Multiplexed signal readout shows the individual response of components in a network for more transparent troubleshooting.

### 5.2.2 Towards a single-stranded *in vitro* transcription activating promoter switch

In an effort to expand the capabilities and scale of transcription-based networks, we sought to develop a single-stranded transcription activator analogous to the hairpin switch inhibitor. Multi-stranded conditionally-active transcription gates have been previously pre-

sented [110] that require stoichiometric annealing and PAGE purification prior to use. In our design, we aimed to satisfy the following requirements: switches must be single-stranded DNA, contain both the sense and antisense T7 promoter sequences, and respond only to cognate signals with a complementary sequence. From the first two requirements, it follows that in the absence of an activating signal, the activator is capable of forming an active double-stranded promoter and subsequently may leak signal. The inactive conformation (i.e. single-stranded promoter) must be more energetically favored in order to compete with the active conformation. To this end, we created a hemi-duplex activator that, in the inactive state, adopts multiple degenerate states that are at equilibrium with one another. The activator contains a single-stranded loop that disrupts the bottom strand of the promoter and "slides" between a range of positions. The range of this sliding loop is bounded by its sequence complementarity with the promoter. In the presence of an activating signal, hybridization between the signal strand and the activator stabilizes the active conformation, producing a stable double-stranded promoter region (Figure 5.5A)

To determine the downstream boundary of the sliding loop, we tested positions where the insertion of a single-stranded loop would disrupt transcription activity. We placed a 17 nt polyT loop at several locations between -17 and -7 (relative to initiation start site) in the templating strand of a double-stranded T7 promoter. This promoter was upstream of a malachite green RNA aptamer sequence and part of an otherwise linear transcription template. We avoided positions downstream of -5 because insertion of a loop here would not disrupt the specificity region [30] and could potentially still allow T7 RNAP to bind. This may lead to stalled enzymes in the OFF state and a delayed response to signal. On the other hand, we did not want to limit the sliding loop to positions further upstream (e.g.

upstream of -15) because a large fraction of switches in the inactive state would contain double-stranded promoters. Fluorescence transcription assay results showed that positions -13, -10, and -7 had similar decreases in activity, with the largest decrease at position -7 with a 3-fold reduction from a linear promoter (Figure 5.4B). We therefore proceeded with -7 as the downstream boundary.

We used NUPACK for the sequence design of the activators. Because the ON state involves a pseudoknot base pairing configuration, which is not supported by NUPACK, we separated the activator into two strands for the purposes of design. We then tested two versions of input strands: one with a complementary partial spacer sequence and another without (Figure 5.5C). We included a malachite green aptamer sequence in the templating region for fluorescence readout. In the presence of the signal strand, transcription activity was about 60% and 50% relative to a hemi-duplex hairpin template control for input versions 1 and 2, respectively (Figure 5.5D). In the presence of non-cognate or "scrambled" inputs (i.e. non-matching stem region), activity was reduced by 2.2-fold and 13.9-fold relative to the cognate input for designs 1 and 2, respectively.

## 5.3 Discussion

The observation of similar patterns of inhibition across different modes of readout shows that alternatives to fluorescence can be used to read transcriptional network output. Several remaining issues should be addressed to improve the consistency, strength of inhibition response, and scalability. First, the total concentration of switches in a network should be lower than the saturating concentration for T7 RNAP, and the concentration ratio be-

tween the total switch and enzyme should be kept constant across assays. This is because T7 RNAP transcription follows Michaelis-Menten enzyme kinetics, which dictates that rate of transcription is dependent on the concentration of available substrate (i.e. double-stranded promoters). This can be done by first determining the saturating concentration of T7 RNAP (by titrating a constant concentration of T7 RNAP with different concentrations of ON state switches) and later by adding an orthogonal "normalization" switch to networks to maintain the total switch concentration. Introducing an additional layer of barcoding at the reverse transcription step with barcoded primers could improve both consistency and scalability, thus ensuring that concentrations of signals are subject to the same fluctuations in the following processing steps.

Given the response of the hairpin activator switch to DNA signals, RNA signals should be tested to assess the utility of this switch design for layered transcription networks. Additionally, as with inhibitor switches, it is likely possible to adapt signals for multiplexed readout using NGS. Ultimately, the activation response should be improved prior to combining activator switches with inhibitor switches, since the current activator design has about 60% activity relative to ON state inhibitors. Increased activation would also enable deep networks to lose less signal over layers. Some adjustments to the activator switch design may potentially improve the signal-to-noise ratio. Reducing the length of the conserved region would destabilize binding overall and cause signal binding to become more dependent on sequence complementarity to the activator, thereby weakening partial stabilization by non-cognate signals. Altering the upstream "GC" clamp either to "AT" or an AT-rich sequence may improve transcription, as previous studies have reported increased transcription activity for an extended dsDNA region upstream of the promoter [10], particularly with an AT-rich

sequence [188].

## 5.4 Materials and Methods

**Oligonucleotides and reagents.** All oligonucleotides were purchased as custom oligonucleotides from IDT. Unless otherwise noted, all enzymes and reaction buffers were purchased from New England Biolabs and all chemical reagents were purchased from Sigma Aldrich.

**Hairpin switch in vitro transcription for sequencing readout.** Prior to in vitro transcription, templates (linear and hairpin) were individually annealed to a final solution of 4 uM in 1X T7 Annealing Buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl) with the following heating protocol: 5 minutes at 95°C, ramp down at 0.1°C/s to 25°C, 5 minutes at 25°C, hold at 4°C until use. Unless otherwise noted in figures, transcription reactions were 20 $\mu$l each and contained a 200 nM of each template (linear or hairpin), 10 U/$\mu$l (about 200 nM) of T7 RNAP, 5 mM of each NTP, 5 mM DTT, 1X T7 Buffer (NEB), and any concentration of DNA input signal specified in the figures. Transcription reactions were incubated at 37°C for 4 hours on a standard thermocycler.

**Sample preparation for sequencing.** Following transcription, reactions were treated with DNase I to remove templates in the following reaction mix: 12 $\mu$l of the transcription reaction, 4 units of DNase I (NEB), 1X DNase I Buffer (NEB), and nuclease-free water added to a final volume of 30 $\mu$l. Reactions were incubated at 37°C for 30 minutes, then EDTA solution was added to each reaction to a final concentration of 5 mM, and the reactions were incubated for 5 minutes at 75°C to inactivate enzymes. Reverse transcription was

then performed with the following reaction mix for each reaction: 5 $\mu$l of the DNase-treated RNA product, 500 nM of RT primer, 15 units AMV RT, 2 units murine RNase Inhibitor, 1 mM each dNTP (ThermoFisher), 2 mM $MgCl_2$, 1X AMV RT buffer, and nuclease-free water added to a final volume of 20 $\mu$l. Reactions were incubated for 1 hour at 42°C, followed by 5 minutes at 80°C for inactivation and stored at -20°C until use.

To prepare samples for NGS, cDNA samples were pooled by combining equal amounts of each sample across barcodes. The pooled samples were cleaned using a QIAquick PCR Purification Kit (Qiagen, 28104) according to manufacture's instructions with these exceptions: the sample-bound column was washed with PE buffer 3X and cDNA was eluted in 30 $\mu$l nuclease-free water. Pooled samples were then PCR-amplified with the following mix: 1.5 $\mu$l of cleaned pooled cDNA, 500 nM each of forward and reverse primers containing indexed NGS adaptors, 0.4 U Q5 DNA polymerase, 400 nM of each dNTP, 1X Q5 Buffer, and nuclease-free water added to a final volume of 20 $\mu$l. The sample was heated with the protocol: 3 minutes initial melting at 98°C, 16 cycles of amplification - 30 seconds of melting at 98°C, 30 seconds of annealing at 65°C, and 30 seconds of extension at 72°C - followed by a 3 minute final extension at 72°C and was stored at 4°C until use. Amplified samples were cleaned again according to manufacturer's instructions using the QIAquick kit, with the exceptions listed above, and stored at -20°C until sequencing.

**Sequencing and analysis.** Samples were sequenced with either on the Illumina MiSeq platform at the University of Texas Genome Sequencing and Analysis Facility using the MiSeq v2 500-cycle kit (MS-102-2003) or on the iSeq 100 platform using the iSeq 100 i1 Reagent v2 300-cycle kit (20031371). Because the samples contained large regions of identical sequences and therefore contain low base diversity, high base diversity samples (e.g. Illumina

PhiX, NEB HeLa genomic DNA) was added to form a high proportion (> 50%) of each run. After NGS, reads were sorted into their respective samples by the i5 and i7 indices of the read. Each read was identified as a transcribed signal using its signal-specific barcode. Up to 1 mismatch was tolerated in this barcode for identification; reads without a barcode match were removed from analysis. All analyses were performed in Python.

**Fluorescence-based in vitro transcription assays.** In fluorescence assays for hairpin inhibitors, a hairpin switch templating the Spinach aptamer [149] was used. For the spinach aptamer transcription assay, reaction mixes were as previously described for sequencing readout with the following exceptions: DFHBI solution in DMSO was added to a final concentration of 50$\mu$M, and each reaction contained 5 U/$\mu$l ($\sim$100 nM) of T7 RNAP instead of 10 U/$\mu$l ($\sim$200 nM). Following transcription, 18 $\mu$l of each reaction was transferred to a Nunc black flat bottom plate 384-well and the end point was measured with a Tecan Infinite M200 plate reader at 469 nm excitation and 501 nm emission.

For assays with hairpin activators, transcription templates included the malachite green aptamer [12]. The composition of the transcription mix was as previously described for sequencing readout with the following exceptions: Malachite Green dye solution in water was added to a final concentration of 25 $\mu$M, and each reaction contained 5 U/$\mu$l ( 100 nM) of T7 RNAP instead of 10 U/$\mu$l ($\sim$200 nM). Kinetic measurements at 630 nm excitation and 664 nm emission were collected every 3 minutes over the transcription period and the reported signal is the average of the final 5 measurements for each sample.

**qPCR analysis.** Samples were reverse transcribed but not pooled together. cDNA samples were similarly cleaned using a QIAquick kit. The qPCR heating protocol was the same as that for PCR with the exception that amplification was carried out in a LightCy-

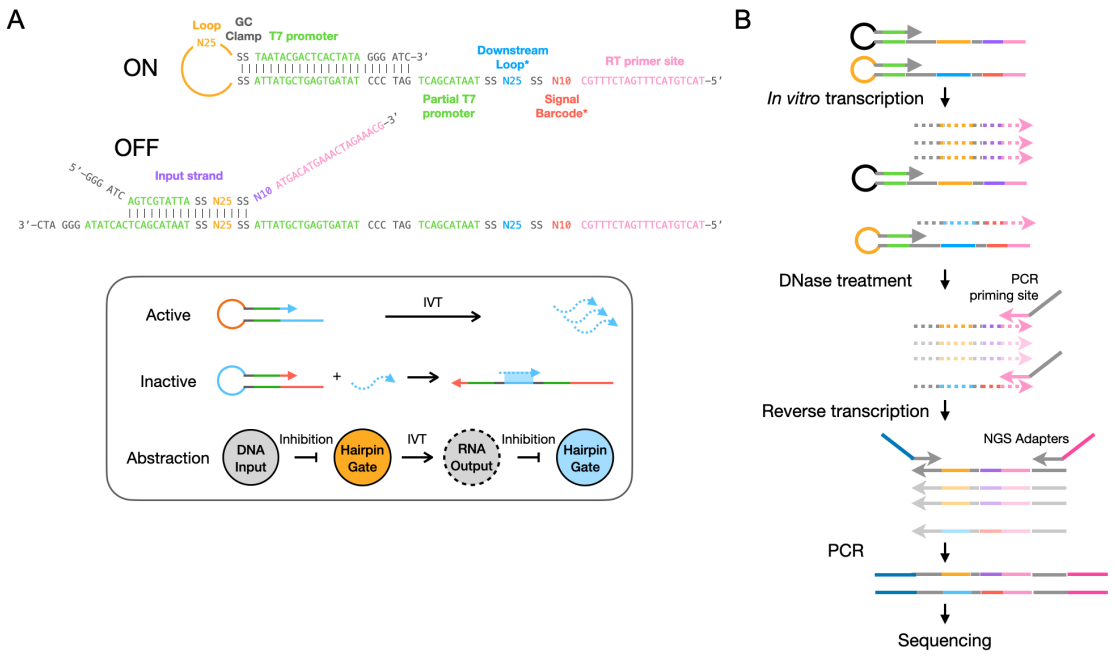


Figure 5.1: **Next-generation sequencing-compatible hairpin switch.**
A. Sequence design (top) and circuit diagram abstraction (bottom). Gray circles represent single-stranded signal and colored circles represent hairpin switches. Solid outlines represent DNA and dashed outlines represent RNA. B. Protocol for reading transcriptional output.

cler96 qPCR machine (Roche) and measurements were taken at the initial melting step, each extension step, and the final extension step. The qPCR mix was the same as the PCR mix with the addition of a final concentration of 1X EvaGreen dye (#31000). $C_q$ of samples was determined using the LightCycler96 software (Roche).

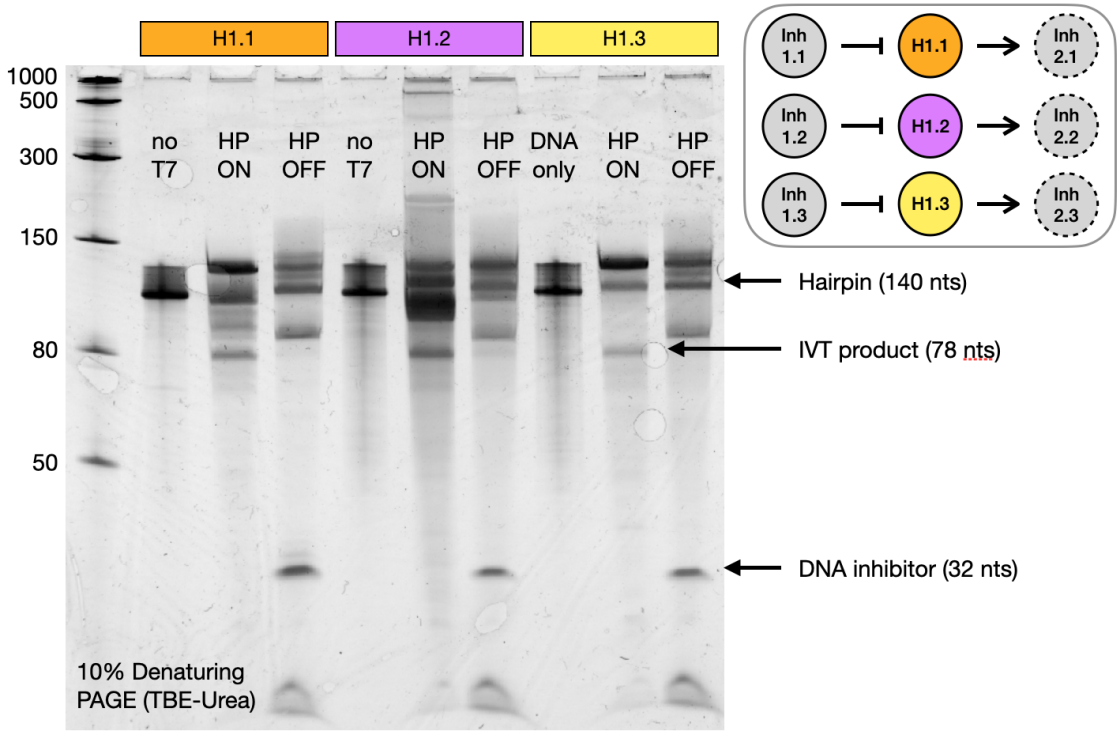


Figure 5.2: **Transcribed products of sequencing-compatible hairpin switches.**

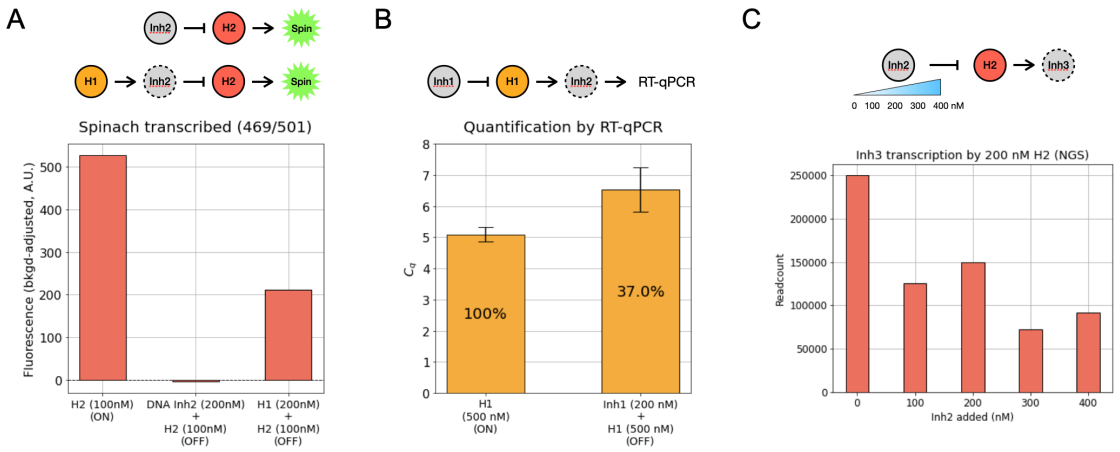


Figure 5.3: **Measuring transcription inhibition using different readout methods.** A. Fluorescence assay. B. qPCR quantification. Percent value is calculated from $C_q$ C. NGS read count.

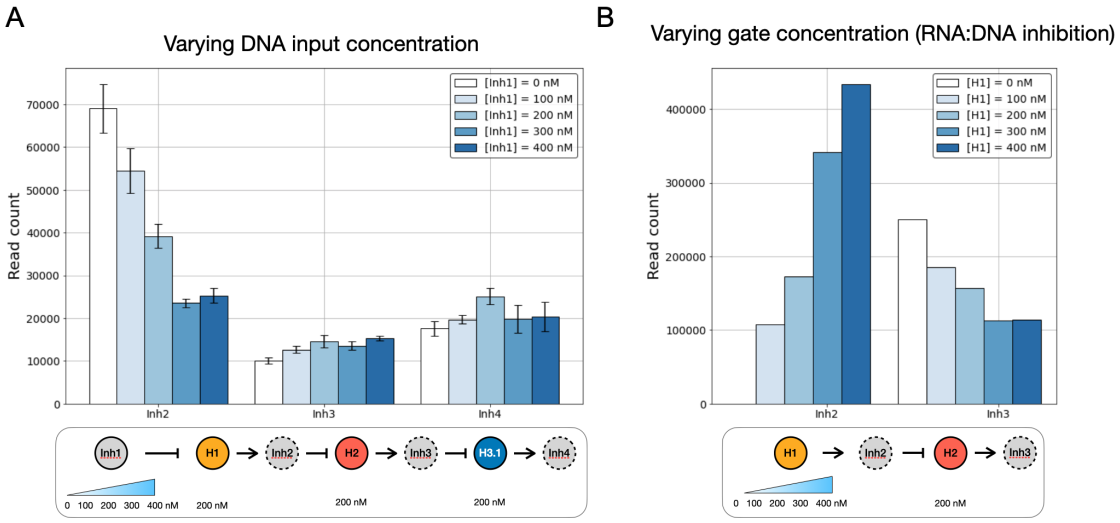


Figure 5.4: **Multiplexed signal readout with NGS.**

A. Network response as a function of DNA input concentration. DNA input inhibits the H1, which produces RNA signals that inhibit H2, etc. B. Network response as a function of switch concentration. Inhibition is caused by RNA transcribed by switch H1.

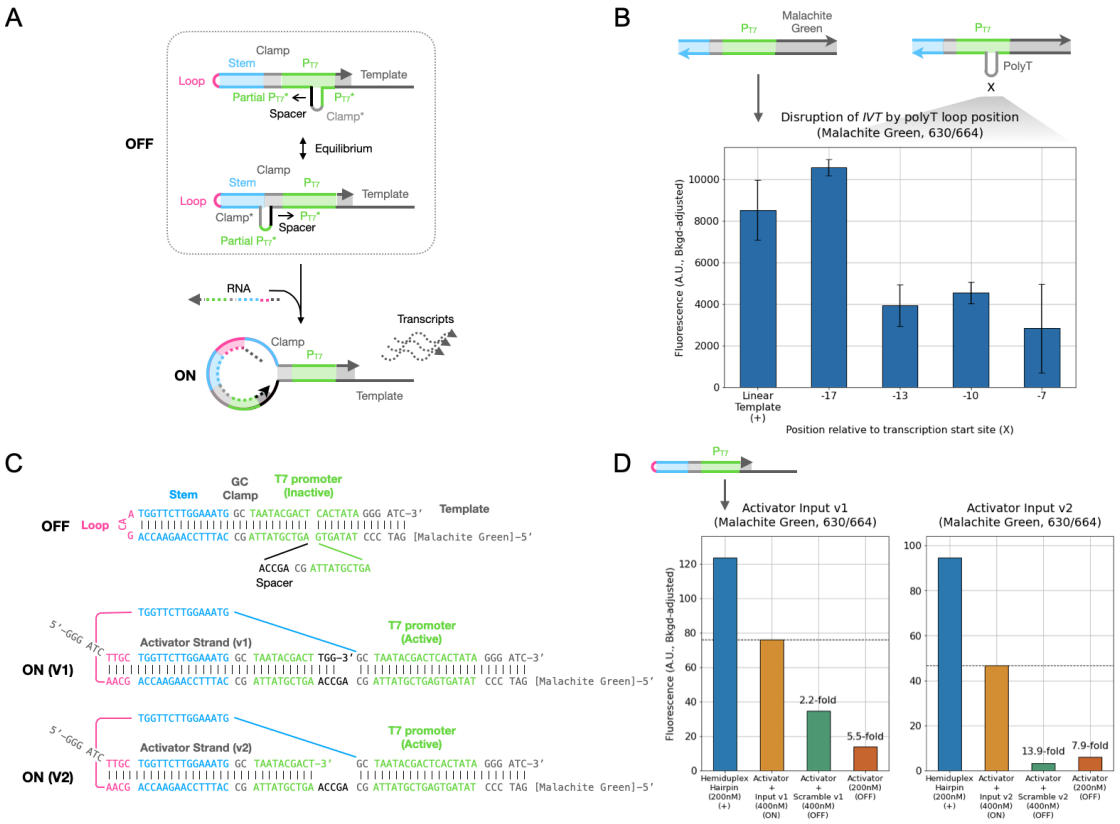


Figure 5.5: **A single-stranded transcription activator switch.**

A. Mechanism of transcription activation. In the OFF state, the activator exists in a mixture of degenerate states with the loop sliding between different positions. B. Disruption of transcription activity by insertion of a polyT loop in the templating strand. C. Sequence designs and activation complex for two versions of inputs. D. Activation fluorescence assay.

# Appendices

# Appendix A

# Recovery of information stored in modified DNA with an evolved polymerase[1]

**Abstract.** DNA is increasingly being explored as an alternative medium for digital information storage, but the potential information loss from degradation and associated issues with error during reading challenge its wide-scale implementation. To address this, we propose an atomic-scale encoding standard for DNA, where information is encoded in degradation-resistant analogues of natural nucleic acids (xNAs). To better enable this approach, we used directed evolution to create a polymerase capable of transforming 2'-O-methyl templates into double-stranded DNA. Starting from a thermophilic, error-correcting reverse transcriptase, RTX, we evolved an enzyme (RTX-Ome v6) that relies on a fully functional proofreading domain to correct mismatches on DNA, RNA, and 2'-O-methyl templates. In addition, we implemented a downstream analysis strategy that accommodates deletions

[1]This appendix is adapted from a manuscript by Shroff R, Ellefson JW, Wang SS, Boulgakov AA, Hughes RA, and Ellington A.D (2020). JE devised the project and carried out protein evolution experiments and assays. RS analyzed the NGS results and sequence decoding. AAB designed and performed the encoding scheme using DNA Fountain. RAH synthesized the modified and unmodified oligonucleotide pools. RS, JE, SSW, and AE wrote the manuscript with feedback from all authors.

that arise during phosphoramidite synthesis, the most common type of synthesis error. By coupling and integrating new chemistries, enzymes, and algorithms, we further enable the large-scale use of nucleic acids for information storage.

## A.1 Introduction

Global data aggregation is expected to outstrip society's storage capacity; by 2025, 163 zettabytes of data will be generated annually [187]. Accommodating this growth strains data centers with an unending battle of scalability. Where traditional electromechanical data storage technologies exhibit defined obsolescence, sensitivity to temperature and humidity, and significant energy maintenance cost, DNA data storage benefits from stability on the order of thousands of years, robustness to a broad range of environmental conditions [11], and a theoretical information density of 455 exabytes per gram of DNA [39] (215 petabytes per gram demonstrated experimentally [59]). Though the latency of information retrieval from DNA prohibits its use in real-time access, it does provide an attractive solution for large archival storage.

At present, molecular data storage is primarily developed using native DNA, in large part because a wide range of enzyme tools are available to reliably read, write, and, to a limited degree, edit information stored in DNA, while very few enzymes are known to be capable of utilizing nucleic acid analogues to a similar degree. Storing data in chemically modified oligonucleotides could not only expand the space of data archival technologies (for instance, multiple independently addressable channels of data) but might also help resolve roadblocks to the widespread use of DNA archives as a supplement to digital archives. These

include data loss arising from sequencing (high-throughput sequencing has error rates between 0.1% and 0.01% [85]), synthesis (oligonucleotide synthesis has error rates of around 0.5%, but significant errors accumulate at lengths beyond 100 nucleotides [93]), and degradation in the presence of contaminants. Efforts to combat such errors can be broadly categorized into chemical approaches (e.g. improved phosphoramidite chemistry or template-free enzymatic oligonucleotide synthesis), biochemical approaches (e.g. improved sequencing error rates, costs, and throughput) and algorithmic approaches (e.g. error-tolerant encoding schemes). Most efforts currently center on algorithmic approaches, with both public and private sectors exploring error-resistant or degradation-tolerant information encoding schemes [39, 59, 74, 77, 17, 138].

Biochemical innovations in the underlying nucleic acid storage "hardware" can complement algorithmic developments by providing more robust information-storage substrates and the corresponding enzymatic machinery for efficient reading and writing. We therefore propose a new paradigm for long-term nucleic acid data storage that uses naturally nuclease-resistant, chemically-modified nucleic acids. Critical to the goal of modified oligonucleotide data storage is the ability to easily read the encoded information at scale via sequencing. To this end, we have encoded information in highly stable 2'-O-methyl RNA, which is known to resist degradation by several ribonucleases as well as deoxyribonucleases [115, 217, 42]. In parallel evolved a polymerase with error-correcting capabilities that can read out the encoded information. Additionally, we present a bioinformatic strategy to retrieve encoded messages with multiple deletions arising from low synthesis fidelity. Overall, our results show the viable and valuable co-development of nucleic acid, protein, and computational components for improved DNA data storage.

## A.2 Results

### A.2.1 Evolution and characterization of a polymerase that could read 2'-OMe DNA

To develop an enzyme that was capable of reading 2'-O-methyl (2'-OMe) modified templates, we built off of previous efforts to evolve a thermostable, error-correcting reverse transcriptase enzyme (RTX), starting from an Archaeal family B DNA polymerase [57]. Previously, enzymes capable of reading 2'-O-methyl-modified templates have been engineered from the family A DNA polymerase Taq [32, 169] or derived from recombinant sources [21]; however, high-fidelity, proofreading polymerases more suited to long-term information storage have not previously been explored. RTX proved capable of accurately reverse transcribing RNA into DNA, but also showed minimal activity on 2'-OMe RNA (Figure A.5A). To further encourage the adoption of 2'-OMe templates by RTX, we made modifications to the emulsion-based selection scheme, reverse transcription compartmentalized self-replication (RT-CSR) [69]. In this scheme, polymerase variants are expressed in bacteria, which are subsequently ensconced within a water-in-oil emulsion mixture (Figure A.1). Upon thermal cycling, individual bacteria lyse, and individual polymerase variants gain access to primers that allow amplification of their own genes. In the current instance, these primers contain 2'-OMe residues, which the polymerase must be able to read through in order to complete the amplification of its own gene, which can then be carried into subsequent rounds of cloning, expression, selection, and amplification.

Starting from RTX, error-prone PCR was used to generate diversity, and bacteria expressing individual variants were emulsified with modified primers. The stringency of selection was tuned by gradually increasing the number of challenge positions in primers

over the course of evolution. For example, the challenge region initially contained a run of 5x 2'-OMe nucleotides in each primer; this number was progressively increased through the course of the selection until it reached 81 2'-OMe bases in the 18th and final round of selection (Table A.1).

After the final round of selection the library as a whole was examined via next-generation sequencing (NGS), revealing several predominant mutations (Figure A.2, Table A.2). Near the template entry site, the E251K and Q242R mutations increase the net positive charge and may provide tighter binding to the negatively charged backbone. The G498A and G350V mutations increase hydrophobicity near the 2' moiety at the +1/-1 positions. Mutations in the finger domain (I488L and K468N) may alter the overall fit of the helix (which is now more A-form than B-form) to the polymerase, an outcome that was also observed in the original selection for RTX.

Based on the distribution and predominance of the accumulated mutations, we rationally constructed a series of variants (Table A.3); mutations that were likely to inactivate the proofreading domain (a common outcome of selections for polymerase activity) were excluded from the rational design. Six designed variants (but not the ancestral enzyme) proved able to reverse transcribe a template with 44 sequential 2'-OMe bases. Of the six variants constructed, one (RTX-Ome v6) exhibited the most robust capability to use O-methyl substrates as templates, and was chosen for further analysis (Figure A.3). Since DNA:OMe duplexes are structurally closer to DNA:RNA than DNA:DNA duplexes, the similar extension profiles for the OMe and RNA templates matched our expectations and further support the hypothesis that the mutations in RTX-Ome v6 improve its ability to utilize A-form helices. This variant also surprisingly retained the ability to reverse transcribe RNA, perform

PCR amplification, and proofread on DNA templates, making it a generally useful enzyme for molecular biology applications.

### A.2.2 Encoding and recovery of DNA files

In order to demonstrate the utility of RTX-Ome v6 in information storage applications, we sought to store and recover information in 2'-OMe oligonucleotides with our engineered enzymatic tool. We transformed a series of short text files into unmodified DNA and a similar set of files into 2'-OMe RNA (Figure A.4A) using a previously reported "DNA Fountain" encoding scheme, which we chose based on its efficient encoding density, built-in substitution error correction, and erasure correction [59]. In total, the unmodified and modified files were encoded into 4000 and 2000 oligonucleotides, respectively, with only the modified oligonucleotides containing a modulo 2000 seed to ease downstream recovery. Oligonucleotide pools were individually synthesized on a 12k Customarray oligonucleotide chip, with a 16-nucleotide seed region for positional identification, 64 nucleotides of data containing payload, and 8 nucleotides containing a Reed-Solomon code. As redundancy is built into the encoding scheme, our simulations suggested that we would require an average of 2784 +/- 58 oligonucleotides to recover the unmodified files and 1245 +/- 46 oligonucleotides for the modified files (Figure A.6).

In order to show that we could selectively recover the information encoded in modified oligonucleotides, the unmodified DNA and modified DNA pools were mixed in a 1:20 ratio. RT-PCR was then used to amplify the oligonucleotides and append appropriate adapters for NGS. Beyond RTX-Ome v6, four additional polymerases (KOD, RTX, and a mix of MMLV/Taq) were assayed for their ability to recover either unmodified or modified DNA

(or both). RTX-Ome, along with the other three control polymerases, successfully amplified the unmodified oligonucleotide sequences, leading to their full recovery (Figure A.4B).

### A.2.3 Computational strategy for reading modified strands

Initially, following sequencing of our oligonucleotide libraries we discovered deletions in virtually every sequencing read. Because deletions were systematic across all libraries, we attribute these errors to oligonucleotide synthesis (Figure A.7). While DNA decoding schemes do have error-checking mechanisms (like the use of a Reed-Solomon code), these are primarily suitable for correcting substitutions and do not generally correct for indels, despite the fact that this may encompass the majority of oligonucleotide synthesis errors [85].

Because deletion errors during synthesis are common, especially when modified nucleotides are utilized, we developed a reconstruction method to account for deletions in NGS reads that expands on the DNA Fountain decoding scheme. Assuming that the deletions appear randomly, redundant coverage can be used to reconstruct the original sequence via sequence alignment. We created bins of similar oligonucleotides through sequence clustering and performed multiple sequence alignments on each bin to build consensus sequences (Figure A.8). If the length of the consensus sequence was less than the expected length for the read, gaps were filled by inserting positions at which a non-gapped base occurred most frequently and iterated to find a sequence that best matched the designed GC content, homopolymer stretch, and Reed-Solomon code. Reads that were more intact were given higher weight in the consensus search.

Our strategy ultimately generated a consensus sequence for each bin of strands that contains the data payload. We then used the DNA Fountain decoding mechanism to further

translate these consensus sequences and the random seed, and thus to reconstruct the original message. Fountain codes are inherently resilient to information loss, in that message reconstruction can still occur with missing packets; however, this type of data is especially sensitive to data corruption and information fidelity may be compromised. Thus, to fix possible "corruption" in our sequences, we further modified the decoding program to better accommodate potential, practical errors. Our solution was to give more weight to sets of sequences that were able to quickly reach a consensus (in 20 iterations or fewer) and then perform 1000 trials with different permutations of the other sequences as randomized sets. In addition, we used the md5 checksum to decode the message. In cases where trials produced different checksums (as was the case with KOD and RTX-Ome), the correct checksum was observed most frequently and no other checksum appeared more than once. To show the robustness of our decoding strategy, we repeated the method on a random sample subset containing 10% of our reads and observed fully correct recovery. Overall, RTX-OMe was the only one among the three tested Archaeal family B, error-correcting polymerases that was able to correctly recover the 2'-OMe-encoded files (Figure A.4B).

## A.3 Discussion

The storage of information in modified nucleic acid templates may eventually be a generally viable option, if several obstacles are overcome. First, more polymerases that can be readily adapted to a variety of nucleic acid analogues (xNAs) will likely prove key. In addition to the proofreading enzyme RTX-OMe, the viral reverse transcriptase MMLV RT is capable of reverse transcribing 2'-O-methyl RNA into DNA ([52]; Figure A.4B) and could therefore also potentially enable a 2'-O-methyl data reading scheme in conjunction with Taq

polymerase. Broadly speaking, however, for most xNAs there are no known RTs or polymerases capable of reverse transcription into sequencing-compatible DNA, especially at the low error rates amenable for data storage [156]. Therefore, the fact that RTX itself had initial broad specificities suggests that it might prove to be a useful starting point for engineering numerous high-fidelity, xNA-compatible polymerases. Indeed, the proofreading capabilities available via RTX may ultimately be compatible with further evolved xNA polymerases with more dramatically altered sugar backbones such as, HNA, LNA, or TNAs, enabling high fidelity reverse transcription of exotic substrates. Additionally, given that non-proofreading DNA polymerases have been evolved to utilize fully 2'-O-methyl modified templates or synthesize fully-modified products [32], it may be feasible to further engineer RTX-OMe to both utilize and transcribe fully-modified oligonucleotides for protected storage schemes. Second, even with improved proofreading, the errors and limitations inherent in both reading and writing require encoding schemes with error correction or tolerance for high error rates at both the software and hardware levels [59, 77, 17, 138, 187]. Fortunately, our computational method's primary dependence on universal alignment parameters (match, mismatch, gap opening, and gap extension scores) and sequence properties (GC content, homopolymer stretch) suggests that it should not only be applicable to various types of consensus searches, but also easily scale with the number of sequences.

Overall, modified nucleic acids offer an attractive, nuclease-resistant medium for long term data storage, especially when read out by evolved, low-error xenopolymerases and an associated, deletion-robust information decoding algorithm. Such xNA systems for information storage could also potentially provide a steganographic and cryptogenetic approach to hidden message storage among otherwise normal information, where privileged information

could only be discovered with privileged polymerases (Figure A.9). More broadly, with the development of multiple xNA-specific xenopolymerases, it may be possible to encode information in separate channels (i.e. sugar backbone variants) and to independently retrieve the information in each channel using a channel-specific polymerase.

## A.4 Methods

**Reverse Transcription CSR (RT-CSR).** RTX polymerase libraries were created through error prone PCR (unless otherwise indicated) to have a mutation rate of 1-2 amino acid mutations per gene. Libraries were cloned into tetracycline inducible vector and electroporated into DH10B E. coli. Library sizes were maintained with a transformation efficiency of at least 106, but more typically 107-108. Induced overnight library cultures were seeded at a 1:20 ratio into fresh 2xYT media supplemented with 100 $\mu$g / mL ampicillin and grown for 1 hour at 37°C. Cells were subsequently induced by the addition of anhydrotetracycline (typically at a final concentration of 200 ng/mL) and incubated at 37°C for 4 hours. Induced cells (200 $\mu$L total) were spun in a tabletop centrifuge at 3,000 x g for 8 minutes. The supernatant was discarded and the cell pellet was resuspended in 150 $\mu$L RTCSR mix: 1x Selection buffer (50 mM Tris-HCl (pH 8.4), 10 mM (NH4)2SO4, 10 mM KCl, 2 mM MgSO4), 260 $\mu$M dNTPs, 530 nM forward and reverse 2' O-methyl containing primers (detailed in Table A.1). The resuspended cells were placed into a 2 mL tube with a 1mL rubber syringe plunger and 600 $\mu$L of oil mix (73% Tegosoft DEC, 7% AbilWE09 (Evonik), and 20% mineral oil (Sigma-Aldrich)). The emulsion was created by placing the cell and oil mix on a TissueLyser LT (Qiagen) with a program of 42Hz for 4 minutes. The emulsified cells were thermal-cycled with the program: 95°C - 3min, 20x (95°C-30 sec, 62°C-30 sec, 68°C-2

min). Emulsions were broken by spinning the reaction (10,000x g - 5 min), removing the top oil phase, adding 150 $\mu$L of H2O and 750 $\mu$L chloroform, vortexing vigorously, and finally phase separating in a phase lock tube (5Prime). The aqueous phase was cleaned using a PCR purification column which results in purified DNA, including PCR products as well as plasmid DNA. Subamplification with corresponding outnested recovery primers ensures that only polymerases that reverse transcribed are PCR amplified. Typically this is achieved by addition of 1/10 the total purified emulsion using Accuprime Pfx (ThermoFisher) in a 20 cycle PCR, however challenging rounds of selection could require increasing the input DNA or cycle number to achieve desired amplification.

**Cloning and purification of polymerase variants.** Escherichia coli DH10B and BL21 (DE3) strains were used for cloning and expression, respectively. Strains were maintained on either Superior or 2X YT growth media. Polymerases were cloned into a modified pET21 vector using NdeI and BamHI sites. Overnight cultures of BL21 (DE3) harboring each of the variants were grown overnight in Superior broth at 37°C. Cells were then diluted 1:250, and protein production was induced with 1 mM IPTG during mid-log at 18°C for 20 hrs. Harvested cells were flash-frozen and lysed by sonication. Polymerases were purified using a gravity flow Ni-NTA column followed by HiTrap Heparin column (GE) using FPLC. Purified fractions were pooled and dialyzed into storage buffer (50 mM Tris-HCl, 50 mM KCl, 0.1 mM EDTA, 1 mM DTT, 0.1% nonidet p40, 0.1% Tween20, and 50% glycerol pH 8.0). Polymerase concentration was determined using a Bradford assay and diluted to a working stock of 0.2 mg / mL.

**Primer Extension Assay.** 10 pmol of 5' fluorescein labeled primer (RT.Probe or RT.Probe.3ddc) were annealed with 50 pmol of template DNA, RNA, or 2' O-methyl DNA

(DNA.TEMP, RNA.TEMP, or Ome.TEMP, respectively) and 0.2 $\mu$g of polymerase by heat denaturation at 80°C for 1 minute and allowing to cool to room temperature. Reactions were initiated by the addition of a "start" mix which contained: 1x Assay Buffer, 2 mM MgSO4 (total) and 200 $\mu$M dNTPs. Reactions were incubated for 5 minutes at 68°C. The labeled primer was removed from the template strand by heating sample at 75°C for 5 minutes in 1x dye (47.5% formamide, 0.01% SDS) and 1 nmol of unlabeled BigBlocker oligonucleotide (to competitively bind the template strand). Samples were run on a 20% (7 M urea) acrylamide gel.

**PCR Proofreading Assay.** 50 $\mu$L PCR reactions were set up with a final concentration of 1x Assay Buffer (60 mM Tris-HCl (pH8.4), 25 mM (NH4)2SO4, 10 mM KCl), 200 $\mu$M dNTPs, 2 mM MgSO4, 400 nM (DiDeTest.F/DiDeTest.R) forward and reverse primers, 20 ng of template plasmid and 0.2 $\mu$g polymerase. Reactions were thermal-cycled using the following program: 95°C - 1 min, 25x (95°C- 30 sec, 55°C- 30 sec, 68°C- 2 min 30 sec).

**RT-PCR Assay.** 50 $\mu$L reverse transcription PCR (RT-PCR) reactions were set up on ice with the following reaction conditions: 1x Assay Buffer, 1 mM MgSO4, 1 M Betaine (Sigma-Aldrich), 200 $\mu$M dNTPs, 400 nM reverse primer PolII.R, 400 nM forward primer PolII.F2, 40 units RNasin Plus (Promega), 0.2 $\mu$g polymerase and 1 $\mu$g of Total RNA from Jurkat cells (Ambion). Reactions were thermal-cycled according to the following parameters: 68°C - 30 min, 25x (95°C- 30 sec, 68°C - 30 sec, 68°C - 30 s/kb).

**Encoding of information into oligonucleotides.** We first combined each set of documents into a tar.xz file and padded the tail end with zeros such that the final file size was a multiple of 16 bytes. We then used DNA Fountain (Erlich) to generate 4000 oligonucleotides encoding the cover message. We confirmed that none of these 4000 nucleotides had

a DNA Fountain seed modulo 2000, a fact that will be used below to distinguish the hidden oligonucleotides from the cover set upon sequencing. For the hidden message, we first generated 2,000,000 DNA Fountain oligonucleotides, and kept only 2000 out of the 8933 whose DNA fountain seed was modulo 2000. We then computationally tested each oligonucleotide set, cover and hidden, to see how many sequences we required to recover each message. For each set, the oligonucleotides were shuffled into a random order and fed into DNA Fountain until the message was recovered (DNA Fountain terminates upon successfully recovering the message). This was repeated 1000 times. We recorded the number of oligonucleotides DNA Fountain required from each permutation before the message was decoded.

**Synthesis of DNA and O'Methyl DNA for DNA Data Storage.** The encoded oligonucleotide pools were each randomly arrayed on a 12,472 feature chip using the Customarray rearrayer software to give a $\sim$3 fold sequence coverage for the standard unencrypted DNA pool (4,000 unique oligonucleotides) and $\sim$6 fold sequence coverage for the encrypted 2'-O-Methyl-DNA oligonucleotide pool (2,000 unique oligonucleotides). The unencrypted DNA oligonucleotides were synthesized on the Customarray B3 oligonucleotide array synthesizer following standard phosphoramidite chemistry protocols. For the synthesis of the encrypted, 2'-O-Methyl oligonucleotides, 5 grams of each of the 2'-O-methyl phosphoramidites (2'-Ome Bz A, Cat. #27-1842; 2'-Ome Ac C, Cat. #27-1823; 2'-Ome U, 27-1825; 2'-Ome iBu G, Cat. #27-1846) were purchased from Thermo Scientific and resuspended in 100mL anhydrous acetonitrile and used for oligonucleotide synthesis on the chip following standard DNA synthesis protocols. Following the completion of the synthesis, the oligonucleotide pools were cleaved and deprotected directly from the chip surface using aqueous ammonia at 65°C for 4 hours. The cleaved and deprotected oligonucleotide pools

were resuspended in TE buffer and purified on a Micro Bio-spin column (Biorad) following the manufacturer's protocol. The column purified oligonucleotide pools were then used for further analysis.

**Preparation of DNA for NGS Sequencing.** Synthesized oligonucleotides were pooled in a ratio of 1 part DNA to 10 parts O-methyl DNA prior to amplification. To prepare oligonucleotides for NGS the pools were PCR amplified to add adaptor sequences. Reactions were indexed using Illumina small RNA primers (RPI1-KOD, RPI2-RTX, RPI3-RTX-Ome, RPI4-OneTaq One Step RTPCR (NEB)). For KOD, RTX, and RTX-Ome: 50$\mu$L PCR reactions were prepared with 1x Assay buffer, 200 $\mu$M dNTPS, 1 M Betaine, 400 nM RP1 primer, 400 nM RPI (1-3), 10 ng oligonucleotide pool, and 0.2 $\mu$g of KOD, RTX, or RTX-Ome polymerase (polymerase added after temperature reached 94°C). Reactions were cycled using a program: 94°C - 30s; 12x cycles (94°C - 15s, 65°C (-1°C/cycle) - 15s, 68°C - 10 minutes). For OneTaq One-step RT-PCR kit, the manufacturer's recommended protocol was used with the same concentration of pooled oligonucleotides. After thermal cycling, products were cleaned using Wizard SV PCR purification kit (Promega) and eluted in 15 $\mu$L H2O. A secondary PCR was used to further amplify products from the RT-PCR before submission to the UT GSAF facility. Accuprime PFX PCR (Thermo Scientific) was used to amplify 5 $\mu$L of the eluted primary amplification with universal outnested primers (Universal F/Universal R) for 25 additional cycles.

**Informatic recovery.** Starting with raw sequencing reads, we first trimmed adapters and filtered reads to be between 50bp and 90bp using flexbar. We clustered the resulting reads using cd-hit at a 70% sequence similarity. For each cluster, we performed multiple sequence alignment using mafft with a gap penalty of zero and weighted bases according

to the read's original length. A consensus sequence is built based on the most common base and gaps are filled until the sequence reaches our target length. Using knowledge of the Reed-Solomon code, GC content, and homopolymer constraints, we ensured that the constructed consensus sequence matched the initial design parameters and if not, iterated through the gaps until such a sequence was found. Sequences were inputted into a modified DNA fountain program, where sequences needing less than 20 iterations were fixed and the remaining shuffled. The aggressive flag in DNA fountain was utilized and run 1000 times, with the most commonly occurring md5 checksum used as the basis for decoding.

**Abbreviations.** RT = reverse transcriptase; OMe = 2'-O-methyl.

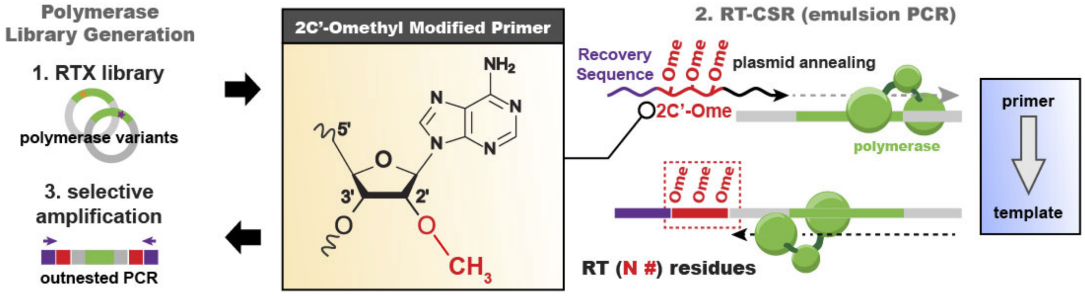


Figure A.1: **Evolution of a xDNA/polymerase pair creates a platform to secure DNA information.**

Evolution strategy to create RTX-Ome.

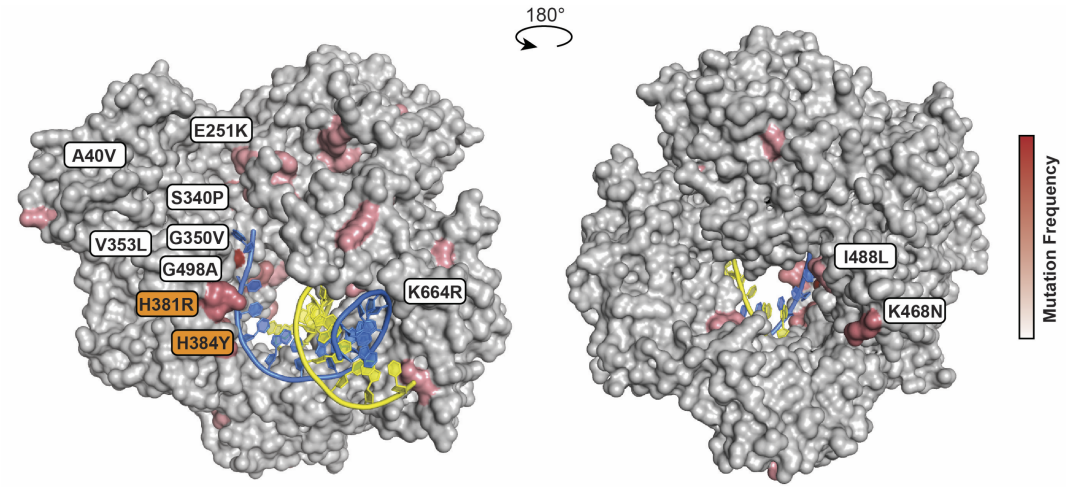


Figure A.2: **Structural heat map of mutations that arose during RT-CSR using 2'-O-methyl challenge template.**

Conserved mutations are colored incrementally darker red to indicate increased frequency. Mutations found in RTX-Ome polymerase are labeled with KOD polymerase reversions indicated in orange. The template strand is labeled in blue and primer strand in yellow.

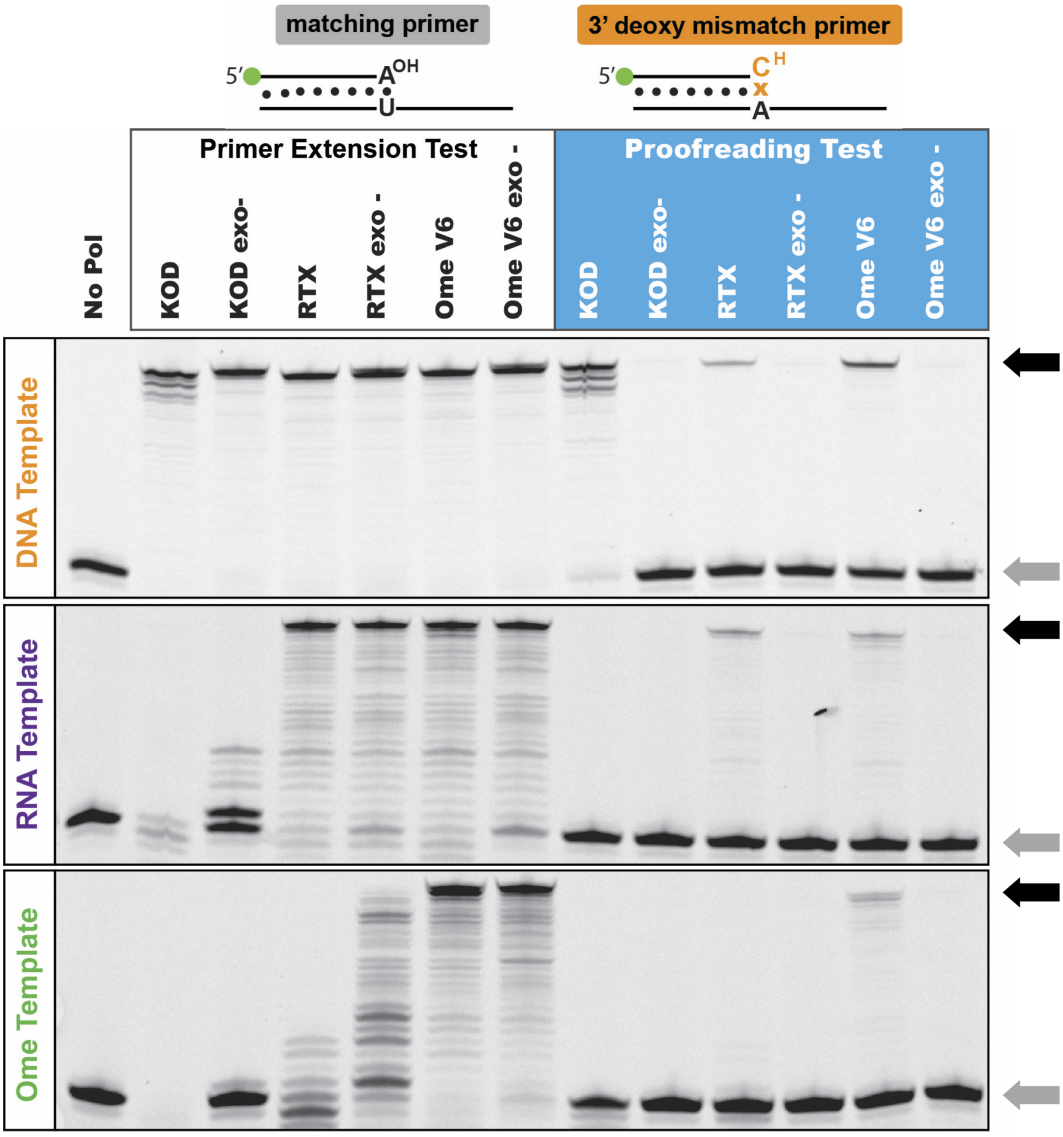


Figure A.3: **Primer extension and proofreading activity of RTX-Ome on DNA, RNA, and 2'-O-methyl templates.**

Primer extension reactions of KOD, RTX, and RTX-Ome polymerases and proofreading deficient counterparts (exo-) on DNA, RNA, and 2'-O-methyl templates. Extension reactions were performed with matched 3' primer:template (left) or a 3' deoxy mismatch primer (right), which must be cleaved prior to extension. Gray arrow indicates unextended fluorescently labeled primer and full length product is marked by a black arrow. In the DNA template uracil is replaced with thymine.

A

| Native DNA | | |
|---|---|---|
| File | Type | Size |
| AACS Encryption Flag | Image | 1 KB |
| Budget | Excel | 5 KB |
| Facebook Facial Recognition | HTML | 2 KB |
| GPS Location History | KML | 2 KB |
| Phone Contacts | Vcard | 1 KB |
| Edgar Allen Poe "Gold Bug" | Text | 30 KB |
| Tic Tac Toe | Game | 1 KB |
| Tortilla Recipe | Text | 1 KB |

| 2'-O-Methyl-Modified DNA | | |
|---|---|---|
| File | Type | Size |
| Cryptographie Indechiffrable | Text | 3 KB |
| Schroder Message | Enigma Text | 1 KB |
| Rasch Message | Enigma Text | 1 KB |
| Kryptos Panel 1 | Text | 1 KB |
| Kryptos Panel 2 | Text | 1 KB |
| Kryptos Panel 3 | Text | 1 KB |
| Kryptos Panel 4 | Text | 1 KB |
| Unbroken U Boat Message | Enigma Text | 1 KB |
| Von Looks Message | Enigma Text | 1 KB |
| Wikileaks Missing Statement | Text | 12 KB |
| Zimmerman Telegram | Text | 1 KB |

B

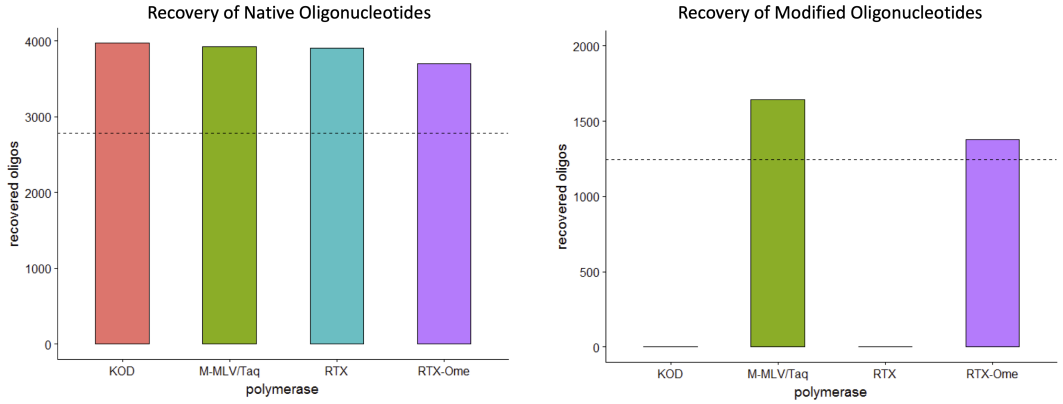


Figure A.4: **Encoding and decoding of information into oligonucleotides.**

A. The listed files were encoded into DNA. B. Recovery performance of each tested polymerase in native DNA oligonucleotides (left) and 2'-O-methyl (right) oligonucleotides. The dotted line indicates the average number of oligonucleotides needed for decoding based on our simulations.

| RTX-Ome Polymerase Evolution | | | |
|---|---|---|---|
| Round # | Mutagenesis | RTCSR Primers | Total Ome |
| 0 | Error Prone PCR | | |
| 1 | N/A | RTCSR.Ome5.F / RTCSR.Ome5.R | 10 |
| 2 | N/A | RTCSR.Ome5.F / RTCSR.Ome5.R | 10 |
| 3 | N/A | RTCSR.Ome10.F / RTCSR.Ome5.R | 15 |
| 4 | N/A | RTCSR.Ome10.F / RTCSR.Ome10.R | 20 |
| 5 | N/A | RTCSR.Ome10.F / RTCSR.Ome10.R | 20 |
| 6 | N/A | RTCSR.Ome20.F / RTCSR.Ome10.R | 30 |
| 7 | N/A | RTCSR.Ome20.F / RTCSR.Ome20.R | 40 |
| 8 | Gene Shuffling | RTCSR.Ome20.F / RTCSR.Ome20.R | 40 |
| 9 | N/A | RTCSR.Ome20.F / RTCSR.Ome20.R | 40 |
| 10 | N/A | RTCSR.Ome20.F / RTCSR.Ome20.R | 40 |
| 11 | N/A | RTCSR.Ome20.F / RTCSR.Ome51.R | 71 |
| 12 | N/A | RTCSR.Ome20.F / RTCSR.Ome51.R | 71 |
| 13 | N/A | RTCSR.Ome20.F / RTCSR.Ome51.R | 71 |
| 14 | N/A | RTCSR.Ome20.F / RTCSR.Ome51.R | 71 |
| 15 | N/A | RTCSR.Ome20.F / RTCSR.Ome51.R | 71 |
| 16 | N/A | RTCSR.Ome30.F / RTCSR.Ome51.R | 81 |
| 17 | N/A | RTCSR.Ome30.F / RTCSR.Ome51.R | 81 |
| 18 | N/A | RTCSR.Ome30.F / RTCSR.Ome51.R | 81 |

Table A.1: **Selection conditions for the evolution of a 2'-O-methyl reverse transcriptase using RT-CSR.**

| Amino Acid Position | RTX | Round 18 | Variant Frequency | Amino Acid Position | RTX | Round 18 | Variant Frequency |
|---|---|---|---|---|---|---|---|
| 498 | G | A | 45.00% | 559 | K | R | 15.20% |
| 251 | E | K | 41.80% | 276 | E | D | 15.20% |
| 350 | G | V | 41.00% | 741 | V | A | 14.60% |
| 159 | M | T | 33.60% | 484 | R | H | 14.30% |
| 381 | H | R | 24.20% | 755 | L | S | 13.80% |
| 488 | I | L | 22.20% | 168 | A | T | 13.70% |
| 340 | S | P | 22.10% | 353 | V | I | 13.50% |
| 384 | H | Y | 21.70% | 768 | W | R | 12.30% |
| 468 | K | N | 21.60% | 214 | F | L | 12.10% |
| 40 | A | V | 21.30% | 247 | R | L | 11.50% |
| 353 | V | L | 20.20% | 605 | T | A | 11.10% |
| 498 | G | S | 18.30% | 704 | L | I | 10.90% |
| 289 | K | R | 18.20% | 752 | K | E | 10.80% |
| 145 | L | P | 17.60% | 640 | V | I | 10.80% |
| 242 | Q | R | 17.50% | 684 | K | R | 10.70% |
| 664 | K | R | 17.20% | 703 | V | I | 10.60% |
| 44 | D | N | 16.50% | 523 | M | T | 10.50% |
| 244 | M | F | 16.10% | 248 | F | L | 10.50% |
| 152 | F | S | 15.60% | 298 | A | S | 10.10% |
| 418 | V | I | 15.30% | 309 | A | T | 10.00% |

Table A.2: **NGS sequencing of the OMe RT-CSR Round 18 pool.**

Mutations are mapped to the parental RTX polymerase. Only mutations with over 10% frequency are shown.

| Polymerase Variant | Mutations (RTX Reference Sequence) | Total Mutations |
|---|---|---|
| RTX-Ome V1 | A40V, D44N, Q242R, M244F, E251K, S340P, G350V, V353L, H381R, H384Y, V418I, K468N, R484H, I488L, G498A, K664R | 16 |
| RTX-Ome V2 | A40V, D44N, Q242R, M244F, E251K, S340P, G350V, V353L, V418I, K468N, R484H, I488L, G498A | 13 |
| RTX-Ome V3 | A40V, Q242R, M244F, E251K, S340P, G350V, V353L, K468N, I488L, G498A | 10 |
| RTX-Ome V4 | A40V, E251K, S340P, G350V, V353L, K468N, I488L, G498A | 8 |
| RTX-Ome V5 | A40V, E251K, S340P, G350V, V353L, H381R, H384Y, K468N, I488L, G498A | 10 |
| RTX-Ome | A40V, E251K, S340P, G350V, V353L, H381R, H384Y, K468N, I488L, G498A, K664R | 11 |

Table A.3: **RTX-Ome variants constructed using NGS data and structure guided design.**

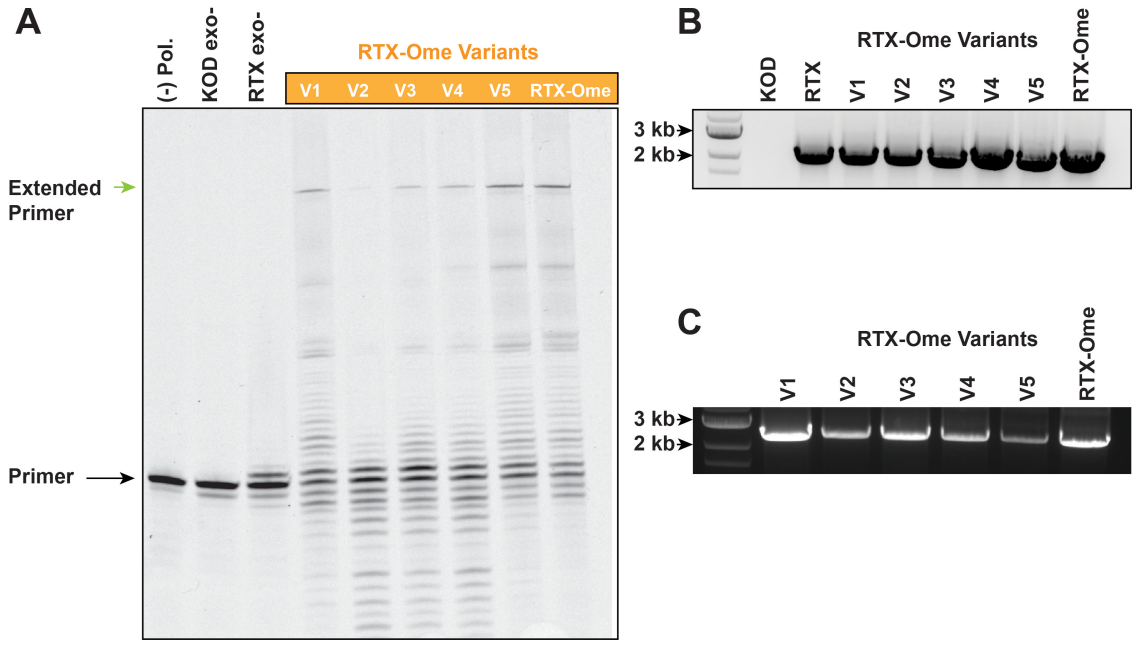


Figure A.5: **Characterization of designed RTX-Ome polymerase variants.**

A. Primer extension of designed polymerase variants (V1-V6) on 2'-O-methyl templates. Fluorescently labeled primers (OMe Probe F) were extended by 44 nucleotides before reaching the end of the template strand (OMe Long R). B. Polymerase variants were tested in a single-enzyme RT-PCR reaction to determine their efficacy for RNA reverse transcription. A 2 kb RNA fragment of PolR2A from human total RNA was amplified using primers PolII.F2 and PolII.R C. 3' Dideoxy mismatch primers (PCRTest.Dide.F /PCRTest.Dide.R) were used in a PCR reaction to determine proofreading capabilities on a DNA template.

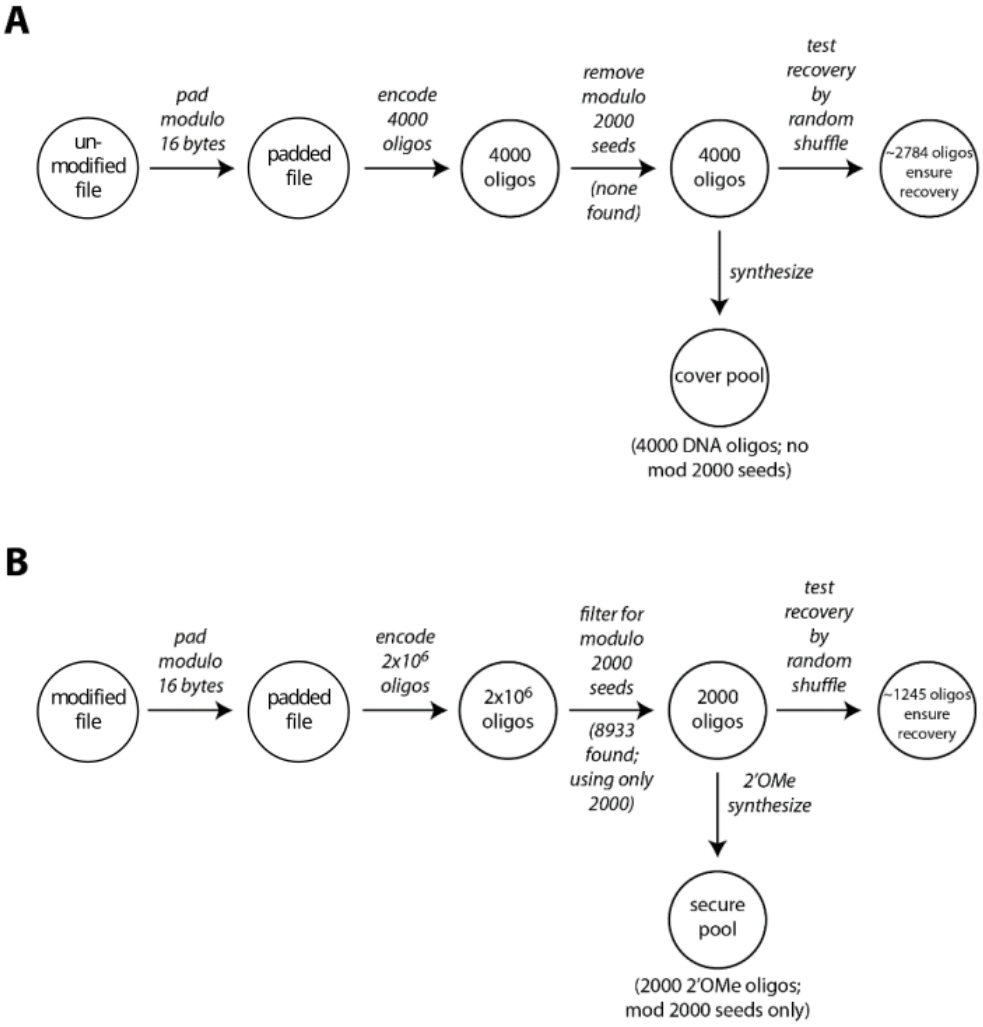


Figure A.6: **DNA Fountain scheme for encoding data files into unmodified and modified oligonucleotides.**

A. The unmodified file is encoded. First, it is padded to a multiple of 16 bytes for compatibility with DNA Fountain. We then let DNA Fountain generate 4000 oligonucleotides to encode it. We filter all oligos with a DNA Fountain seed modulo 2000 (by chance, none were found in our particular run). We then test how many oligos are sufficient to recover the original (padded) file by randomly shuffling the 4000 oligo file and feeding it into the DNA Fountain decoder. Since the decoder stops as soon as it recovers the file, we can tally how many oligos out of the 4000 are required. Repeating this test 1000 times gives statistics that indicate even in cases with loses larger than 1100 out of 4000 it is likely to recover the file. Finally, we perform next-generation sequencing on the 4000 oligos. B. We perform the analogous procedure for the modified file, except we want to encode only oligos with modulo 2000 DNA Fountain seeds, hence the large number of initial oligos generated. The remaining steps are identical, except synthesis uses 2'-OMe bases.

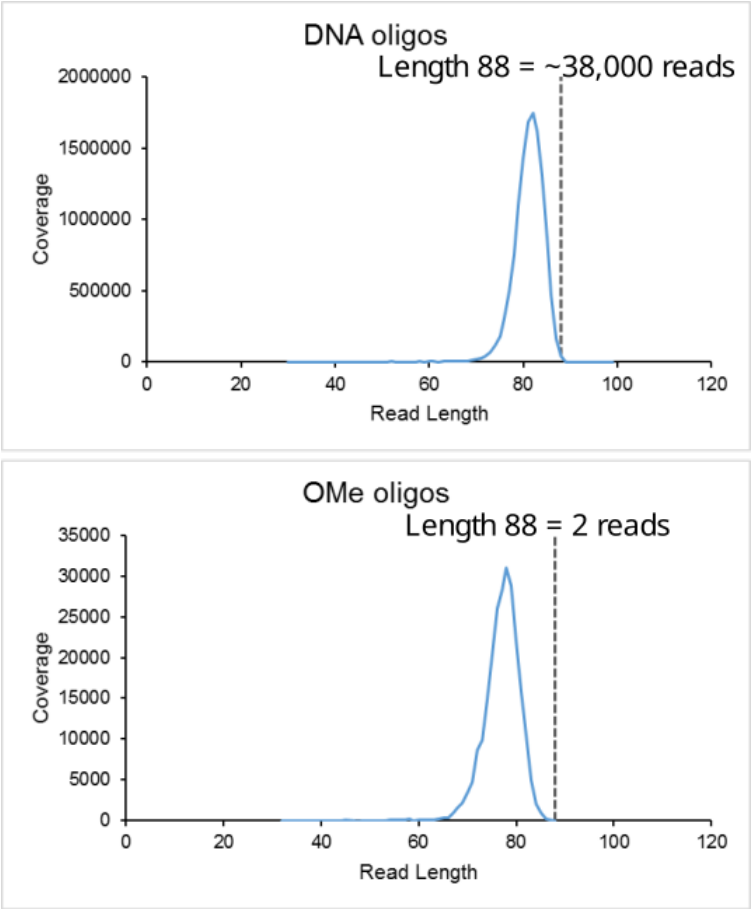


Figure A.7: **Distribution of NGS read sizes.**

The vast majority of sequences are less than the designed length of 88 bases.

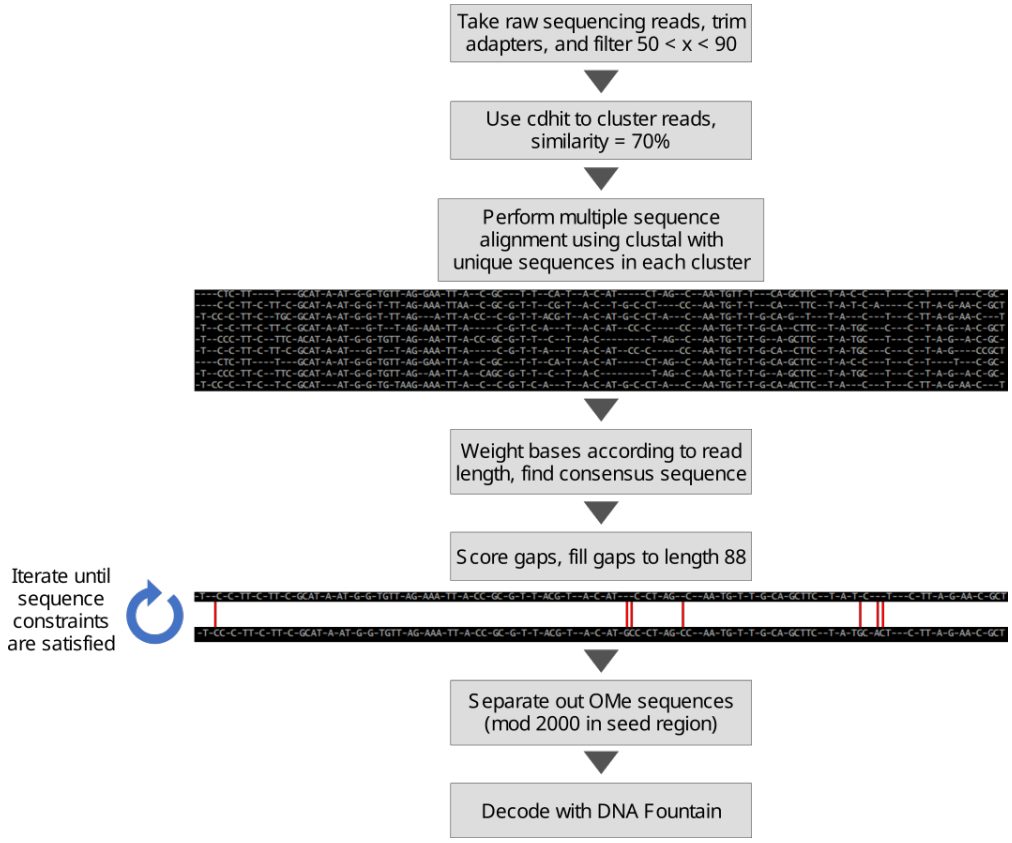


Figure A.8: **NGS read reconstruction workflow.**

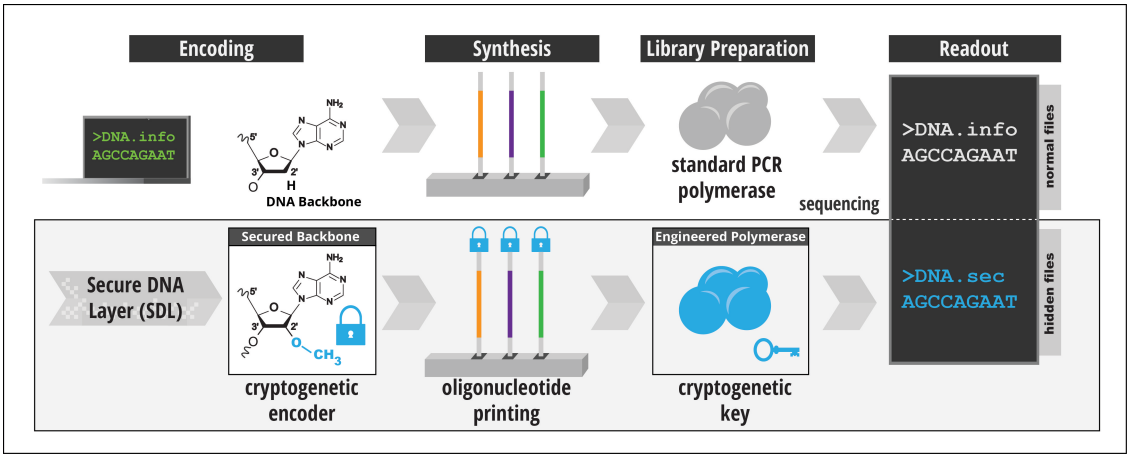


Figure A.9: **Potential cryptogenetic application for RTX-Ome and other xenopolymerases capable of reading information encoded in xNA oligonucleotides.**

# Bibliography


[1] Illumina sequencing technology: Highest data accuracy, simple workflow, and a broad range of applications (illumina). https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf. Accessed: 2022-03-09.

[2] Sony develops magnetic tape storage technology with the industry's highest*1 recording areal density of 201 gb/in2 (sony). https://www.sony.com/en/SonyInfo/News/Press/201708/17-070E/. Accessed: 2022-03-07.

[3] Storing oligos: 7 things you should know (idt). https://www.idtdna.com/pages/education/decoded/article/storing-oligos-7-things-you-should-know. Accessed: 2022-03-07.

[4] Ultramer$^{TM}$ DNA Oligonucleotides. https://www.idtdna.com/pages/products/custom-dna-rna/dna-oligos/ultramer-dna-oligos.

[5] Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (statista). https://www.statista.com/statistics/871513/worldwide-data-created/. Accessed: 2022-03-07.

[6] Yaniv Amir, Eldad Ben-Ishay, Daniel Levner, Shmulik Ittah, Almogit Abu-Horowitz, and Ido Bachelet. Universal computing by dna origami robots in a living animal. *Nature nanotechnology*, 9(5):353–357, 2014.

[7] Frances M Anastassacos, ZHAO Zhao, Yang Zeng, and William M Shih. Glutaraldehyde cross-linking of oligolysines coating dna origami greatly reduces susceptibility to nuclease degradation. *Journal of the American Chemical Society*, 142(7):3311–3315, 2020.

[8] May T Aung-Htut, Iain Comerford, Russell Johnsen, Kerrie Foyle, Sue Fletcher, and Steve D Wilton. Reduction of integrin alpha 4 activity through splice modulating antisense oligonucleotides. *Scientific reports*, 9(1):1–12, 2019.

[9] Alexandre Baccouche, Kevin Montagne, Adrien Padirac, Teruo Fujii, and Yannick Rondelez. Dynamic dna-toolbox reaction circuits: A walkthrough. *Methods*, 67(2):234–249, 2014.

[10] Michail M Baklanov, Larisa N Golikova, and Enrst G Malygin. Effect on dna transcription of nucleotide sequences upstream to t7 promoter. *Nucleic Acids Research*, 24(18):3659–3660, 1996.

[11] Carter Bancroft, Timothy Bowler, Brian Bloom, and Catherine Taylor Clelland. Long-term storage of information in dna. *Science*, 293(5536):1763–1765, 2001.

[12] Christopher Baugh, Dilârâ Grate, and Charles Wilson. 2.8 å crystal structure of the malachite green aptamer. *Journal of molecular biology*, 301(1):117–128, 2000.

[13] Callista Bee, Yuan-Jyue Chen, Melissa Queen, David Ward, Xiaomeng Liu, Lee Organick, Georg Seelig, Karin Strauss, and Luis Ceze. Molecular-level similarity search brings computing to DNA data storage. *Nature Communications*, 12(1):4764, December 2021.

[14] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53–59, 2008.

[15] Sanchita Bhadra and Andrew D Ellington. Design and application of cotranscriptional non-enzymatic rna circuits and signal transducers. *Nucleic acids research*, 42(7):e58–e58, 2014.

[16] Meinolf Blawat, Klaus Gaedke, Ingo Huetter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W Pruitt, and George M Church. Forward error correction for dna data storage. *Procedia Computer Science*, 80:1011–1022, 2016.

[17] James Bornholt, Randolph Lopez, Douglas M Carmean, Luis Ceze, Georg Seelig, and Karin Strauss. A dna-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 637–649, 2016.

[18] James Bremer, Marek Nowicki, Suzanne Beckner, Donald Brambilla, Mike Cronin, Steven Herman, Andrea Kovacs, and Patricia Reichelderfer. Comparison of two amplification technologies for detection and quantitation of human immunodeficiency virus type 1 rna in the female genital tract. *Journal of clinical microbiology*, 38(7):2665–2669, 2000.

[19] Kenneth J Breslauer, Ronald Frank, Helmut Blöcker, and Luis A Marky. Predicting dna duplex stability from the base sequence. *Proceedings of the National Academy of Sciences*, 83(11):3746–3750, 1986.

[20] Jason D Buenrostro, Carlos L Araya, Lauren M Chircus, Curtis J Layton, Howard Y Chang, Michael P Snyder, and William J Greenleaf. Quantitative analysis of rna-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature biotechnology*, 32(6):562–568, 2014.

[21] Paula E Burmeister, Scott D Lewis, Robert F Silva, Jeffrey R Preiss, Lillian R Horwitz, P Shannon Pendergrast, Thomas G McCauley, Jeffrey C Kurz, David M Epstein, Charles Wilson, et al. Direct in vitro selection of a 2-o-methyl aptamer to vegf. *Chemistry & biology*, 12(1):25–33, 2005.

[22] Sheng Cai, Cheulhee Jung, Sanchita Bhadra, and Andrew D Ellington. Phosphorothioated primers lead to loop-mediated isothermal amplification at low temperatures. *Analytical chemistry*, 90(14):8290–8294, 2018.

[23] Angelo Cangialosi, ChangKyu Yoon, Jiayu Liu, Qi Huang, Jingkai Guo, Thao D Nguyen, David H Gracias, and Rebecca Schulman. Dna sequence–directed shape change of photopatterned hydrogels via high-degree swelling. *Science*, 357(6356):1126–1130, 2017.

[24] Luca Cardelli. Two-domain dna strand displacement. *Mathematical Structures in Computer Science*, 23(2):247–271, 2013.

[25] Marta Carlucci, Elzbieta Kierzek, Anna Olejnik, Douglas H Turner, and Ryszard Kierzek. Chemical synthesis of lna-2-thiouridine and its influence on stability and selectivity of oligonucleotide binding to rna. *Biochemistry*, 48(46):10882–10893, 2009.

[26] Carlos Ernesto Castro, Fabian Kilchherr, Do-Nyun Kim, Enrique Lin Shiao, Tobias Wauer, Philipp Wortmann, Mark Bathe, and Hendrik Dietz. A primer to scaffolded dna origami. *Nature methods*, 8(3):221–229, 2011.

[27] Luis Ceze, Jeff Nivala, and Karin Strauss. Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20(8):456–466, August 2019.

[28] Michael Chamberlin, Janet McGrath, and Lucy Waskell. New rna polymerase from escherichia coli infected with bacteriophage t7. *Nature*, 228(5268):227–231, 1970.

[29] James Chappell, Melissa K Takahashi, and Julius B Lucks. Creating small transcription activating rnas. *Nature chemical biology*, 11(3):214–220, 2015.

[30] Graham MT Cheetham, David Jeruzalmi, and Thomas A Steitz. Structural basis for initiation of transcription from an rna polymerase–promoter complex. *Nature*, 399(6731):80–83, 1999.

[31] Ho-Lin Chen, David Doty, and David Soloveichik. Deterministic function computation with chemical reaction networks. *Natural computing*, 13(4):517–534, 2014.

[32] Tingjian Chen, Narupat Hongdilokkul, Zhixia Liu, Ramkrishna Adhikary, Shujian S Tsuen, and Floyd E Romesberg. Evolution of thermophilic dna polymerases for the recognition and amplification of c2-modified dna. *Nature chemistry*, 8(6):556–562, 2016.

[33] Yuan-Jyue Chen, Benjamin Groves, Richard A Muscat, and Georg Seelig. Dna nanotechnology from the test tube to the cell. *Nature nanotechnology*, 10(9):748–760, 2015.

[34] Kevin M Cherry and Lulu Qian. Scaling up molecular pattern recognition with dna-based winner-take-all neural networks. *Nature*, 559(7714):370–376, 2018.

[35] Steven M Chirieleison, Peter B Allen, Zack B Simpson, Andrew D Ellington, and Xi Chen. Pattern transformation with dna circuits. *Nature chemistry*, 5(12):1000–1005, 2013.

[36] Harry MT Choi, Victor A Beck, and Niles A Pierce. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. *ACS nano*, 8(5):4284–4294, 2014.

[37] Harry MT Choi, Maayan Schwarzkopf, Mark E Fornace, Aneesh Acharya, Georgios Artavanis, Johannes Stegmaier, Alexandre Cunha, and Niles A Pierce. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development*, 145(12):dev165753, 2018.

[38] Leo YT Chou and William M Shih. In vitro transcriptional regulation via nucleic-acid-based transcription factors. *ACS Synthetic Biology*, 8(11):2558–2565, 2019.

[39] George M Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012.

[40] Matthew Cook. Universality in Elementary Cellular Automata. *Complex Systems*, 15(1):1–40, 2004.

[41] Donald M Crothers and Bruno H Zimm. Theory of the melting transition of synthetic polynucleotides: evaluation of the stacking free energy. *Journal of molecular biology*, 9(1):1–9, 1964.

[42] Lendell L Cummins, Stephen R Owens, Lisa M Risen, Elena A Lesnik, Susan M Freier, Danny McGee, Charles J Guinosso, and P Dan Cook. Characterization of fully 2-modified oligoribonucleotide hetero-and homoduplex hybridization and nuclease sensitivity. *Nucleic acids research*, 23(11):2019–2024, 1995.

[43] Douglas R Davies, Amy D Gelinas, Chi Zhang, John C Rohloff, Jeffrey D Carter, Daniel O'Connell, Sheela M Waugh, Steven K Wolk, Wesley S Mayfield, Alex B Burgin, et al. Unique motifs and hydrophobic interactions shape the binding of modified dna ligands to protein targets. *Proceedings of the National Academy of Sciences*, 109(49):19971–19976, 2012.

[44] Angela F. De Fazio, Afaf H. El-Sagheer, Jason S. Kahn, Iris Nandhakumar, Matthew Richard Burton, Tom Brown, Otto L. Muskens, Oleg Gang, and Antonios G. Kanaras. Light-Induced Reversible DNA Ligation of Gold Nanoparticle Superlattices. *ACS Nano*, 13(5):5771–5777, May 2019.

[45] Alain De Mesmaeker, Robert Haener, Pierre Martin, and Heinz E Moser. Antisense oligonucleotides. *Accounts of Chemical Research*, 28(9):366–374, 1995.

[46] Scott G Delcourt and RD Blake. Stacking energies in dna. *Journal of Biological Chemistry*, 266(23):15160–15169, 1991.

[47] Vadim V Demidov. Rolling-circle amplification in dna diagnostics: the power of simplicity. *Expert review of molecular diagnostics*, 2(6):542–548, 2002.

[48] Howard DeVoe and Ignacio Tinoco Jr. The stability of helical polynucleotides: base contributions. *Journal of molecular biology*, 4(6):500–517, 1962.

[49] Robert M Dirks and Niles A Pierce. Triggered amplification by hybridization chain reaction. *Proceedings of the National Academy of Sciences*, 101(43):15275–15278, 2004.

[50] Mitchel J Doktycz, Robert F Goldstein, Teodoro M Paner, Frank J Gallo, and Albert S Benight. Studies of dna dumbbells. i. melting curves of 17 dna dumbbells with different duplex stem sequences linked by t4 endloops: Evaluation of the nearest-neighbor stacking interactions in dna. *Biopolymers: Original Research on Biomolecules*, 32(7):849–864, 1992.

[51] Mitchel J Doktycz, Max D Morris, Shelly J Dormady, Kenneth L Beattie, and K Bruce Jacobson. Optical melting of 128 octamer dna duplexes: effects of base pair location and nearest neighbors on thermal stability. *Journal of Biological Chemistry*, 270(15):8439–8445, 1995.

[52] Zhi-Wei Dong, Peng Shao, Li-Ting Diao, Hui Zhou, Chun-Hong Yu, and Liang-Hu Qu. Rtl-p: a sensitive approach for detecting sites of 2-o-methylation in rna molecules. *Nucleic acids research*, 40(20):e157–e157, 2012.

[53] Cosimo Ducani, Corinna Kaul, Martin Moche, William M Shih, and Björn Högberg. Enzymatic production of'monoclonal stoichiometric'single-stranded dna oligonucleotides. *Nature methods*, 10(7):647–652, 2013.

[54] Kimberly J Durniak, Scott Bailey, and Thomas A Steitz. The structure of a transcribing t7 rna polymerase in transition from initiation to elongation. *Science*, 322(5901):553–557, 2008.

[55] Johann Elbaz, Zhen-Gang Wang, Ron Orbach, and Itamar Willner. ph-stimulated concurrent mechanical activation of two dna "tweezers". a "set- reset" logic gate system. *Nano letters*, 9(12):4510–4514, 2009.

[56] Johann Elbaz, Peng Yin, and Christopher A Voigt. Genetic encoding of dna nanostructures and their self-assembly in living bacteria. *Nature communications*, 7(1):1–11, 2016.

[57] Jared W Ellefson, Jimmy Gollihar, Raghav Shroff, Haridha Shivram, Vishwanath R Iyer, and Andrew D Ellington. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science*, 352(6293):1590–1593, 2016.

[58] Maria Erali, Karl V Voelkerding, and Carl T Wittwer. High resolution melting applications for clinical laboratory medicine. *Experimental and molecular pathology*, 85(1):50–58, 2008.

[59] Yaniv Erlich and Dina Zielinski. Dna fountain enables a robust and efficient storage architecture. *science*, 355(6328):950–954, 2017.

[60] Constantine G. Evans and Erik Winfree. DNA Sticky End Design and Assignment for Robust Algorithmic Self-assembly. In David Soloveichik and Bernard Yurke, editors, *DNA Computing and Molecular Programming*, volume 8141, pages 61–75. Springer International Publishing, Cham, 2013.

[61] Joshua Fern and Rebecca Schulman. Modular dna strand-displacement controllers for directing material expansion. *Nature communications*, 9(1):1–8, 2018.

[62] Susan M Freier, Ryszard Kierzek, John A Jaeger, Naoki Sugimoto, Marvin H Caruthers, Thomas Neilson, and Douglas H Turner. Improved free-energy parameters for predictions of rna duplex stability. *Proceedings of the National Academy of Sciences*, 83(24):9373–9377, 1986.

[63] Jonas J Funke, Philip Ketterer, Corinna Lieleg, Philipp Korber, and Hendrik Dietz. Exploring nucleosome unwrapping using dna origami. *Nano letters*, 16(12):7891–7898, 2016.

[64] Jonas J Funke, Philip Ketterer, Corinna Lieleg, Sarah Schunter, Philipp Korber, and Hendrik Dietz. Uncovering the forces between nucleosomes using dna origami. *Science advances*, 2(11):e1600974, 2016.

[65] Yang Gao, Lauren K Wolf, and Rosina M Georgiadis. Secondary structure effects on dna hybridization kinetics: a solution versus surface comparison. *Nucleic acids research*, 34(11):3370–3377, 2006.

[66] Cody Geary, Guido Grossi, Ewan KS McRae, Paul WK Rothemund, and Ebbe S Andersen. Rna origami design tools enable cotranscriptional folding of kilobase-sized nanoscaffolds. *Nature chemistry*, 13(6):549–558, 2021.

[67] Cody Geary, Paul WK Rothemund, and Ebbe S Andersen. A single-stranded architecture for cotranscriptional folding of rna nanostructures. *Science*, 345(6198):799–804, 2014.

[68] Anthony J Genot, David Yu Zhang, Jonathan Bath, and Andrew J Turberfield. Remote toehold: a mechanism for flexible control of dna hybridization kinetics. *Journal*

*of the American Chemical Society*, 133(7):2177–2182, 2011.

[69] Farid J Ghadessy, Jennifer L Ong, and Philipp Holliger. Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences*, 98(8):4552–4557, 2001.

[70] Matthew R Giese, Kelly Betschart, Taraka Dale, Cheryl K Riley, Carrie Rowan, Kimberly J Sprouse, and Martin J Serra. Stability of rna hairpins closed by wobble base pairs. *Biochemistry*, 37(4):1094–1100, 1998.

[71] Pooria Gill and Amir Ghaemi. Nucleic acid isothermal amplification technologies—a review. *Nucleosides, Nucleotides and Nucleic Acids*, 27(3):224–243, 2008.

[72] Jule Goike, Ching-Lin Hsieh, Andrew Horton, Elizabeth C Gardner, Foteini Bartzoka, Nianshuang Wang, Kamyab Javanmardi, Andrew Herbert, Shawn Abbassi, Rebecca Renberg, et al. Synthetic repertoires derived from convalescent covid-19 patients enable discovery of sars-cov-2 neutralizing antibodies and a novel quaternary binding modality. *bioRxiv*, 2021.

[73] Larry Gold, Deborah Ayers, Jennifer Bertino, Christopher Bock, Ashley Bock, Edward Brody, Jeff Carter, Virginia Cunningham, Andrew Dalby, Bruce Eaton, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *Nature Precedings*, pages 1–1, 2010.

[74] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M LeProust, Botond Sipos, and Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *Nature*, 494(7435):77–80, 2013.

[75] Ashwin Gopinath, Evan Miyazono, Andrei Faraon, and Paul WK Rothemund. Engineering and mapping nanocavity emission via precision placement of dna origami. *Nature*, 535(7612):401–405, 2016.

[76] Osamu Gotoh and Yusaku Tagashira. Stabilities of nearest-neighbor doublets in double-helical dna determined by fitting calculated melting profiles to observed profiles. *Biopolymers: Original Research on Biomolecules*, 20(5):1033–1042, 1981.

[77] Robert N Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J Stark. Robust chemical preservation of digital information on dna in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8):2552–2555, 2015.

[78] Donald M Gray. Derivation of nearest-neighbor properties from data on nucleic acid oligomers. ii. thermodynamic parameters of dna· rna hybrids and dna duplexes. *Biopolymers: Original Research on Biomolecules*, 42(7):795–810, 1997.

[79] Donald M Gray and Ignacio Tinoco Jr. A new approach to the study of sequence-dependent properties of polynucleotides. *Biopolymers: Original Research on Biomolecules*, 9(2):223–244, 1970.

[80] Alexander A Green, Jongmin Kim, Duo Ma, Pamela A Silver, James J Collins, and Peng Yin. Complex cellular logic computation using ribocomputing devices. *Nature*, 548(7665):117–121, 2017.

[81] Alexander A Green, Pamela A Silver, James J Collins, and Peng Yin. Toehold switches: de-novo-designed regulators of gene expression. *Cell*, 159(4):925–939, 2014.

[82] Andrea Guichón, Héctor Chiparelli, Alfredo Martınez, Claudia Rodrıguez, Alfonsina Trento, José C Russi, and Guadalupe Carballal. Evaluation of a new nasba assay for the qualitative detection of hepatitis c virus based on the nuclisens® basic kit reagents. *Journal of clinical virology*, 29(2):84–91, 2004.

[83] Dongran Han, Xiaodong Qi, Cameron Myhrvold, Bei Wang, Mingjie Dai, Shuoxing Jiang, Maxwell Bates, Yan Liu, Byoungkwon An, Fei Zhang, et al. Single-stranded dna and rna origami. *Science*, 358(6369):eaao2648, 2017.

[84] Koshi Hasatani, Mathieu Leocmach, Anthony J Genot, André Estévez-Torres, Teruo Fujii, and Yannick Rondelez. High-throughput and long-term observation of compartmentalized biochemical oscillators. *Chemical Communications*, 49(73):8090–8092, 2013.

[85] John A Hawkins, Stephen K Jones, Ilya J Finkelstein, and William H Press. Indel-correcting dna barcodes for high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 115(27):E6217–E6226, 2018.

[86] Yu He and David R Liu. Autonomous multistep organic synthesis in a single isothermal solution mediated by a dna walker. *Nature nanotechnology*, 5(11):778–782, 2010.

[87] Ivo L Hofacker. Vienna rna secondary structure server. *Nucleic acids research*, 31(13):3429–3431, 2003.

[88] Fan Hong, Fei Zhang, Yan Liu, and Hao Yan. Dna origami: scaffolds for creating higher order structures. *Chemical reviews*, 117(20):12584–12640, 2017.

[89] Jeff Hooyberghs, Paul Van Hummelen, and Enrico Carlon. The effects of mismatches on hybridization in dna microarrays: determination of nearest neighbor parameters. *Nucleic acids research*, 37(7):e53–e53, 2009.

[90] Melissa C Hopfinger, Charles C Kirkpatrick, and Brent M Znosko. Predictions and analyses of rna nearest neighbor parameters for modified nucleotides. *Nucleic acids research*, 48(16):8901–8913, 2020.

[91] Shuichi Hoshika, Nicole A Leal, Myong-Jung Kim, Myong-Sang Kim, Nilesh B Karalkar, Hyo-Joong Kim, Alison M Bates, Norman E Watkins Jr, Holly A SantaLucia, Adam J Meyer, et al. Hachimoji dna and rna: A genetic system with eight building blocks. *Science*, 363(6429):884–887, 2019.

[92] J Howard. Mechanics of motor proteins and the cytoskeleton: Sinauer assoc. *Sunderland, MA*, 2001. `https://bionumbers.hms.harvard.edu/bionumber.aspx?&id=101506`.

[93] Randall A Hughes and Andrew D Ellington. Synthetic dna synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harbor perspectives in biology*, 9(1):a023812, 2017.

[94] Natalia V. Ivanova and Masha L. Kuzmina. Protocols for dry DNA storage and shipment at room temperature. *Molecular Ecology Resources*, 13(5):890–898, September 2013.

[95] Kamyab Javanmardi, Chia-Wei Chou, Cynthia I Terrace, Ankur Annapareddy, Tamer S Kaoud, Qingqing Guo, Josh Lutgens, Hayley Zorkic, Andrew P Horton,

Elizabeth C Gardner, et al. Rapid characterization of spike variants via mammalian cell surface display. *Molecular cell*, 81(24):5099–5111, 2021.

[96] Yu Jiang, Bingling Li, John N Milligan, Sanchita Bhadra, and Andrew D Ellington. Real-time detection of isothermal amplification reactions with thermostable catalytic hairpin assembly. *Journal of the American Chemical Society*, 135(20):7430–7433, 2013.

[97] Yu Sherry Jiang, Sanchita Bhadra, Bingling Li, and Andrew D Ellington. Mismatches improve the performance of strand-displacement nucleic acid circuits. *Angewandte Chemie*, 126(7):1876–1879, 2014.

[98] Stephen K Jones, John A Hawkins, Nicole V Johnson, Cheulhee Jung, Kuang Hu, James R Rybarski, Janice S Chen, Jennifer A Doudna, William H Press, and Ilya J Finkelstein. Massively parallel kinetic profiling of natural and engineered crispr nucleases. *Nature Biotechnology*, 39(1):84–93, 2021.

[99] Hyungmin Jun, Fei Zhang, Tyson Shepherd, Sakul Ratanalert, Xiaodong Qi, Hao Yan, and Mark Bathe. Autonomously designed free-form 2d dna origami. *Science advances*, 5(1):eaav0655, 2019.

[100] Cheulhee Jung, Peter B Allen, and Andrew D Ellington. A stochastic dna walker that traverses a microparticle surface. *Nature nanotechnology*, 11(2):157–163, 2016.

[101] Cheulhee Jung, Peter B Allen, and Andrew D Ellington. A simple, cleated dna walker that hangs on to surfaces. *ACS nano*, 11(8):8047–8054, 2017.

[102] Cheulhee Jung, John A Hawkins, Stephen K Jones Jr, Yibei Xiao, James R Rybarski, Kaylee E Dillard, Jeffrey Hussmann, Fatema A Saifuddin, Cagri A Savran, Andrew D

Ellington, et al. Massively parallel biophysical analysis of crispr-cas complexes on next generation sequencing chips. *Cell*, 170(1):35–47, 2017.

[103] Ralf Jungmann, Maier S Avendaño, Johannes B Woehrstein, Mingjie Dai, William M Shih, and Peng Yin. Multiplexed 3d cellular super-resolution imaging with dna-paint and exchange-paint. *Nature methods*, 11(3):313–318, 2014.

[104] Shaunak Kar and Andrew D Ellington. In vitro transcription networks based on hairpin promoter switches. *ACS Synthetic Biology*, 7(8):1937–1945, 2018.

[105] Yonggang Ke, Stuart Lindsay, Yung Chang, Yan Liu, and Hao Yan. Self-assembled water-soluble nucleic acid probe tiles for label-free rna hybridization assays. *Science*, 319(5860):180–183, 2008.

[106] Yonggang Ke, Luvena L Ong, William M Shih, and Peng Yin. Three-dimensional structures self-assembled from dna bricks. *science*, 338(6111):1177–1183, 2012.

[107] Elzbieta Kierzek, Anna Ciesielska, Karol Pasternak, David H Mathews, Douglas H Turner, and Ryszard Kierzek. The influence of locked nucleic acid residues on the thermodynamic properties of 2'-o-methyl rna/rna heteroduplexes. *Nucleic acids research*, 33(16):5082–5093, 2005.

[108] Jongmin Kim, John Hopfield, and Erik Winfree. Neural network computation by in vitro transcriptional circuits. *Advances in neural information processing systems*, 17, 2004.

[109] Jongmin Kim, Ishan Khetarpal, Shaunak Sen, and Richard M Murray. Synthetic circuit for exact adaptation and fold-change detection. *Nucleic acids research*, 42(9):6078–6089, 2014.

[110] Jongmin Kim, Kristin S White, and Erik Winfree. Construction of an in vitro bistable circuit from synthetic transcriptional switches. *Molecular systems biology*, 2(1):68, 2006.

[111] Jongmin Kim and Erik Winfree. Synthetic in vitro transcriptional oscillators. *Molecular systems biology*, 7(1):465, 2011.

[112] Jongmin Kim, Yu Zhou, Paul D Carlson, Mario Teichmann, Soma Chaudhary, Friedrich C Simmel, Pamela A Silver, James J Collins, Julius B Lucks, Peng Yin, et al. De novo-designed translation-repressing riboregulators for multi-input cellular logic. *Nature chemical biology*, 15(12):1173–1182, 2019.

[113] Ryo Komura, Wataru Aoki, Keisuke Motone, Atsushi Satomura, and Mitsuyoshi Ueda. High-throughput evaluation of t7 promoter variants using biased randomization and dna barcoding. *PloS one*, 13(5):e0196905, 2018.

[114] Enzo Kopperger, Jonathan List, Sushi Madhira, Florian Rothfischer, Don C Lamb, and Friedrich C Simmel. A self-assembled nanoscale robotic arm controlled by electric fields. *Science*, 359(6373):296–301, 2018.

[115] Maureen S Lalonde, Yuhong Zuo, Jianwei Zhang, Xin Gong, Shaohui Wu, Arun Malhotra, and Zhongwei Li. Exoribonuclease r in mycoplasma genitalium can carry out both

rna processing and degradative functions and is sensitive to rna ribose methylation. *RNA*, 13(11):1957–1968, 2007.

[116] Martin Langecker, Vera Arnaut, Thomas G Martin, Jonathan List, Stephan Renner, Michael Mayer, Hendrik Dietz, and Friedrich C Simmel. Synthetic lipid membrane channels formed by designed dna nanostructures. *Science*, 338(6109):932–936, 2012.

[117] Henry H Lee, Reza Kalhor, Naveen Goela, Jean Bolot, and George M Church. Terminator-free template-independent enzymatic dna synthesis for digital information storage. *Nature communications*, 10(1):1–12, 2019.

[118] Bingling Li, Xi Chen, and Andrew D Ellington. Adapting enzyme-free dna circuits to the detection of loop-mediated isothermal amplification reactions. *Analytical chemistry*, 84(19):8371–8377, 2012.

[119] Bingling Li, Andrew D Ellington, and Xi Chen. Rational, modular adaptation of enzyme-free dna circuits to multiple detection methods. *Nucleic acids research*, 39(16):e110–e110, 2011.

[120] Mo Li, Mengxi Zheng, Siyu Wu, Cheng Tian, Di Liu, Yossi Weizmann, Wen Jiang, Guansong Wang, and Chengde Mao. In vivo production of rna nanostructures via programmed folding of single-stranded rnas. *Nature communications*, 9(1):1–9, 2018.

[121] Xiaoyu Li and David R Liu. Dna-templated organic synthesis: nature's strategy for controlling chemical reactivity applied to synthetic molecules. *Angewandte Chemie International Edition*, 43(37):4848–4870, 2004.

[122] Volker Limmroth, Frederik Barkhof, Nuket Desem, Mark P Diamond, George Tachas, ATL1102 Study Group, et al. Cd49d antisense drug atl1102 reduces disease activity in patients with relapsing-remitting ms. *Neurology*, 83(20):1780–1788, 2014.

[123] Di Liu, Cody W Geary, Gang Chen, Yaming Shao, Mo Li, Chengde Mao, Ebbe S Andersen, Joseph A Piccirilli, Paul WK Rothemund, and Yossi Weizmann. Branched kissing loops for the construction of diverse rna homooligomeric nanostructures. *Nature Chemistry*, 12(3):249–259, 2020.

[124] Hao Liu, Fan Hong, Francesca Smith, John Goertz, Thomas Ouldridge, Molly M Stevens, Hao Yan, and Petr Šulc. Kinetics of rna and rna: Dna hybrid strand displacement. *ACS Synthetic Biology*, 10(11):3066–3073, 2021.

[125] Ke Liu, Chao Pan, Alexandre Kuhn, Adrian Pascal Nievergelt, Georg E. Fantner, Olgica Milenkovic, and Aleksandra Radenovic. Detecting topological variations of DNA at single-molecule level. *Nature Communications*, 10(1):3, December 2019.

[126] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012, 2012.

[127] Randolph Lopez, Ruofan Wang, and Georg Seelig. A molecular multi-gene classifier for disease diagnostics. *Nature chemistry*, 10(7):746–754, 2018.

[128] Hengyun Lu, Francesca Giordano, and Zemin Ning. Oxford nanopore minion sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, 14(5):265–279, 2016.

[129] Robert R. F. Machinek, Thomas E. Ouldridge, Natalie E. C. Haley, Jonathan Bath, and Andrew J. Turberfield. Programmable energy landscapes for kinetic control of DNA strand displacement. *Nature Communications*, 5(1):5324, December 2014.

[130] Karishma Matange, James M Tuck, and Albert J Keung. Dna stability: a central design consideration for dna data storage systems. *Nature communications*, 12(1):1–9, 2021.

[131] Chitrani Medhi, John BO Mitchell, Sarah L Price, and Alethea B Tabor. Electrostatic factors in dna intercalation. *Biopolymers: Original Research on Biomolecules*, 52(2):84–93, 1999.

[132] Jean-Louis Mergny and Laurent Lacroix. Analysis of thermal melting curves. *Oligonucleotides*, 13(6):515–537, 2003.

[133] Adam J Meyer, Jared W Ellefson, and Andrew D Ellington. Directed evolution of a panel of orthogonal t7 rna polymerase variants for in vivo or in vitro synthetic circuitry. *ACS synthetic biology*, 4(10):1070–1076, 2015.

[134] Dionis Minev, Christopher M Wintersinger, Anastasia Ershova, and William M Shih. Robust nucleation control via crisscross polymerization of highly coordinated dna slats. *Nature communications*, 12(1):1–9, 2021.

[135] Larry E Morrison and Lucy M Stols. Sensitive fluorescence-based thermodynamic and kinetic measurements of dna hybridization in solution. *Biochemistry*, 32(12):3095–3104, 1993.

[136] Tsugunori Notomi, Hiroto Okayama, Harumi Masubuchi, Toshihiro Yonekawa, Keiko Watanabe, Nobuyuki Amino, and Tetsu Hase. Loop-mediated isothermal amplification of dna. *Nucleic acids research*, 28(12):e63–e63, 2000.

[137] Razvan Nutiu, Robin C Friedman, Shujun Luo, Irina Khrebtukova, David Silva, Robin Li, Lu Zhang, Gary P Schroth, and Christopher B Burge. Direct measurement of dna affinity landscapes on a high-throughput sequencing instrument. *Nature biotechnology*, 29(7):659–664, 2011.

[138] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, et al. Random access in large-scale dna data storage. *Nature biotechnology*, 36(3):242–248, 2018.

[139] Thomas E Ouldridge, Petr Šulc, Flavio Romano, Jonathan PK Doye, and Ard A Louis. Dna hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic acids research*, 41(19):8886–8895, 2013.

[140] Richard Owczarzy. Melting temperatures of nucleic acids: discrepancies in analysis. *Biophysical chemistry*, 117(3):207–215, 2005.

[141] Richard Owczarzy, Bernardo G Moreira, Yong You, Mark A Behlke, and Joseph A Walder. Predicting stability of dna duplexes in solutions containing magnesium and monovalent cations. *Biochemistry*, 47(19):5336–5353, 2008.

[142] Richard Owczarzy, Andrey V Tataurov, Yihe Wu, Jeffrey A Manthey, Kyle A McQuisten, Hakeem G Almabrazi, Kent F Pedersen, Yuan Lin, Justin Garretson, Neil O

McEntaggart, et al. Idt scitools: a suite for analysis and design of nucleic acid oligomers. *Nucleic acids research*, 36(suppl_2):W163–W169, 2008.

[143] Richard Owczarzy, Peter M Vallone, Frank J Gallo, Teodoro M Paner, Michael J Lane, and Albert S Benight. Predicting sequence-dependent melting stability of short duplex dna oligomers. *Biopolymers: Original Research on Biomolecules*, 44(3):217–239, 1997.

[144] Richard Owczarzy, Yong You, Christopher L Groth, and Andrey V Tataurov. Stability and mismatch discrimination of locked nucleic acid–dna duplexes. *Biochemistry*, 50(43):9352–9367, 2011.

[145] Adrien Padirac, Teruo Fujii, André Estévez-Torres, and Yannick Rondelez. Spatial waves in synthetic biochemical networks. *Journal of the American Chemical Society*, 135(39):14586–14592, 2013.

[146] Sebastian Palluk, Daniel H Arlow, Tristan de Rond, Sebastian Barthel, Justine S Kang, Rathin Bector, Hratch M Baghdassarian, Alisa N Truong, Peter W Kim, Anup K Singh, Nathan J Hillson, and Jay D Keasling. De novo dna synthesis using polymerase-nucleotide conjugates. *Nature Biotechnology*, 36:645–650, 2018.

[147] Rupali P Patwardhan, Choli Lee, Oren Litvin, David L Young, Dana Pe'er, and Jay Shendure. High-resolution analysis of dna regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology*, 27(12):1173–1175, 2009.

[148] Matthew Petersheim and Douglas H Turner. Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with ccgg, ccggp, ccggap, accggp, ccggup, and accggup. *Biochemistry*, 22(2):256–263, 1983.

[149] Georgios Pothoulakis, Francesca Ceroni, Benjamin Reeve, and Tom Ellis. The spinach rna aptamer as a characterization tool for synthetic biology. *ACS synthetic biology*, 3(3):182–187, 2014.

[150] Florian Praetorius, Benjamin Kick, Karl L. Behler, Maximilian N. Honemann, Dirk Weuster-Botz, and Hendrik Dietz. Biotechnological mass production of DNA origami. *Nature*, 552(7683):84–87, December 2017.

[151] Joseph D Puglisi and Ignacio Tinoco Jr. Absorbance melting curves of rna. *Methods in enzymology*, 180:304–325, 1989.

[152] Lulu Qian and Erik Winfree. Scaling up digital circuit computation with dna strand displacement cascades. *science*, 332(6034):1196–1201, 2011.

[153] Lulu Qian, Erik Winfree, and Jehoshua Bruck. Neural network computation with dna strand displacement cascades. *Nature*, 475(7356):368–372, 2011.

[154] Curtis A Raskin, George A Diaz, and William T McAllister. T7 rna polymerase mutants with altered promoter specificities. *Proceedings of the National Academy of Sciences*, 90(8):3147–3151, 1993.

[155] Rebecca M Reynolds, Paul L Padfield, and Jonathan R Seckl. Disorders of sodium balance. *Bmj*, 332(7543):702–705, 2006.

[156] John D Roberts, Katarzyna Bebenek, and Thomas A Kunkel. The accuracy of reverse transcriptase from hiv-1. *Science*, 242(4882):1171–1173, 1988.

[157] Minqing Rong, Biao He, William T McAllister, and Russell K Durbin. Promoter specificity determinants of t7 rna polymerase. *Proceedings of the National Academy of Sciences*, 95(2):515–519, 1998.

[158] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011.

[159] Paul WK Rothemund. Folding dna to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.

[160] Steve Rozen and Helen Skaletsky. Primer3 on the www for general users and for biologist programmers. In *Bioinformatics methods and protocols*, pages 365–386. Springer, 2000.

[161] John SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465, 1998.

[162] John SantaLucia, Hatim T Allawi, and P Ananda Seneviratne. Improved nearest-neighbor parameters for predicting dna duplex stability. *Biochemistry*, 35(11):3555–3562, 1996.

[163] J SantaLucia Jr. The use of spectroscopic techniques in the study of dna stability. *Spectrophotometry and Spectrofluorimetry: A Practical Approach*, pages 329–56, 2000.

[164] John SantaLucia Jr and Donald Hicks. The thermodynamics of dna structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.

[165] Dominic Scalise and Rebecca Schulman. Designing modular reaction-diffusion programs for complex pattern formation. *Technology*, 2(01):55–66, 2014.

[166] Dominic Scalise and Rebecca Schulman. Emulating cellular automata in chemical reaction–diffusion networks. *Natural Computing*, 15(2):197–214, 2016.

[167] Samuel W Schaffter and Rebecca Schulman. Building in vitro transcriptional regulatory networks by successively integrating multiple functional circuit modules. *Nature chemistry*, 11(9):829–838, 2019.

[168] Samuel W Schaffter and Elizabeth Strychalski. Co-transcriptional rna strand displacement circuits. *bioRxiv*, 2021.

[169] Hayley J Schultz, Andrea M Gochi, Hannah E Chia, Alexie L Ogonowsky, Sharon Chiang, Nedim Filipovic, Aurora G Weiden, Emma E Hadley, Sara E Gabriel, and Aaron M Leconte. Taq dna polymerase mutants and 2-modified sugar recognition. *Biochemistry*, 54(38):5999–6008, 2015.

[170] Boxuan Shen, Veikko Linko, Kosti Tapio, Siim Pikker, Tibebe Lemma, Ashwin Gopinath, Kurt V Gothelf, Mauri A Kostiainen, and J Jussi Toppari. Plasmonic nanostructures through dna-assisted lithography. *Science advances*, 4(2):eaap8978, 2018.

[171] William M Shih, Joel D Quispe, and Gerald F Joyce. A 1.7-kilobase single-stranded dna that folds into a nanoscale octahedron. *Nature*, 427(6975):618–621, 2004.

[172] Jong-Shik Shin and Niles A Pierce. A synthetic dna walker for molecular transport. *Journal of the American Chemical Society*, 126(35):10834–10835, 2004.

[173] Lian CT Shoute and Glen R Loppnow. Characterization of the binding interactions between evagreen dye and dsdna. *Physical Chemistry Chemical Physics*, 20(7):4772–4780, 2018.

[174] Prakash Shrestha, Darren Yang, Toma E Tomov, James I MacDonald, Andrew Ward, Hans T Bergal, Elisha Krieg, Serkan Cabi, Yi Luo, Bhavik Nathwani, et al. Single-molecule mechanical fingerprinting with dna nanoswitch calipers. *Nature nanotechnology*, 16(12):1362–1370, 2021.

[175] Friedrich C Simmel, Bernard Yurke, and Hari R Singh. Principles and applications of nucleic acid strand displacement reactions. *Chemical reviews*, 119(10):6326–6369, 2019.

[176] David Soloveichik, Georg Seelig, and Erik Winfree. Dna as a universal substrate for chemical kinetics. *Proceedings of the National Academy of Sciences*, 107(12):5393–5398, 2010.

[177] Stephan Spitzer and Fritz Eckstein. Inhibition of deoxyribonucleases by phosphorothioate groups in oligodeoxyribonucleotides. *Nucleic acids research*, 16(24):11691–11704, 1988.

[178] Niranjan Srinivas, James Parkin, Georg Seelig, Erik Winfree, and David Soloveichik. Enzyme-free nucleic acid dynamical systems. *Science*, 358(6369):eaal2052, 2017.

[179] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310, 2020.

[180] F William Studier and Barbara A Moffatt. Use of bacteriophage t7 rna polymerase to direct selective high-level expression of cloned genes. *Journal of molecular biology*, 189(1):113–130, 1986.

[181] Hari KK Subramanian, Banani Chakraborty, Ruojie Sha, and Nadrian C Seeman. The label-free unambiguous detection and symbolic display of single nucleotide polymorphisms on dna origami. *Nano letters*, 11(2):910–913, 2011.

[182] Naoki Sugimoto, Kei-ichi Honda, and Muneo Sasaki. Application of the thermodynamic parameters of dna stability prediction to double-helix formation of deoxyribooligonucleotides. *Nucleosides, Nucleotides & Nucleic Acids*, 13(6-7):1311–1317, 1994.

[183] Naoki Sugimoto, Shu-ich Nakano, Mari Yoneyama, and Kei-ich Honda. Improved thermodynamic parameters and helix initiation factor to predict stability of dna duplexes. *Nucleic acids research*, 24(22):4501–4505, 1996.

[184] Naoki Sugimoto, Shu-ichi Nakano, Misa Katoh, Akiko Matsumura, Hiroyuki Nakamuta, Tatsuo Ohmichi, Mari Yoneyama, and Muneo Sasaki. Thermodynamic parameters to predict stability of rna/dna hybrid duplexes. *Biochemistry*, 34(35):11211–11216, 1995.

[185] Petr Šulc, Thomas E Ouldridge, Flavio Romano, Jonathan PK Doye, and Ard A Louis. Modelling toehold-mediated rna strand displacement. *Biophysical journal*, 108(5):1238–1247, 2015.

[186] S Kasra Tabatabaei, Boya Wang, Nagendra Bala Murali Athreya, Behnam Enghiad, Alvaro Gonzalo Hernandez, Christopher J Fields, Jean-Pierre Leburton, David Soloveichik, Huimin Zhao, and Olgica Milenkovic. Dna punch cards for storing data on native dna sequences via enzymatic nicking. *Nature communications*, 11(1):1–10, 2020.

[187] SM Tabatabaei Yazdi, Yongbo Yuan, Jian Ma, Huimin Zhao, and Olgica Milenkovic. A rewritable, random-access dna-based storage system. *Scientific reports*, 5(1):1–10, 2015.

[188] Guo-Qing Tang, Rajiv P Bandwar, and Smita S Patel. Extended upstream at sequence increases t7 promoter strength. *Journal of Biological Chemistry*, 280(49):40707–40713, 2005.

[189] Guo-Qing Tang and Smita S Patel. T7 rna polymerase-induced bending of promoter dna is coupled to dna opening. *Biochemistry*, 45(15):4936–4946, 2006.

[190] Karsten Temme, Rena Hill, Thomas H Segall-Shapiro, Felix Moser, and Christopher A Voigt. Modular control of multiple pathways using engineered orthogonal t7 polymerases. *Nucleic acids research*, 40(17):8773–8781, 2012.

[191] Chris Thachuk, Erik Winfree, and David Soloveichik. Leakless DNA strand displacement systems. In *Lecture Notes in Computer Science (Including Subseries Lecture*

*Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9211, pages 133–153. Springer Verlag, 2015.

[192] Anupama J Thubagere, Wei Li, Robert F Johnson, Zibo Chen, Shayan Doroudi, Yae Lim Lee, Gregory Izatt, Sarah Wittman, Niranjan Srinivas, Damien Woods, et al. A cargo-sorting dna robot. *Science*, 357(6356):eaan6558, 2017.

[193] Ignacio Tinoco, Olke C. Uhlenbeck, and Mark D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, 1971.

[194] Jacob M Tome, Abdullah Ozer, John M Pagano, Dan Gheba, Gary P Schroth, and John T Lis. Comprehensive analysis of rna-protein interactions by high-throughput sequencing–rna affinity profiling. *Nature methods*, 11(6):683–688, 2014.

[195] Douglas H Turner, Naoki Sugimoto, and Susan M Freier. Rna structure prediction. *Annual review of biophysics and biophysical chemistry*, 17(1):167–192, 1988.

[196] Niels V Voigt, Thomas Tørring, Alexandru Rotaru, Mikkel F Jacobsen, Jens B Ravnsbæk, Ramesh Subramani, Wael Mamdouh, Jørgen Kjems, Andriy Mokhir, Flemming Besenbacher, et al. Single-molecule chemical reactions on dna origami. *Nature nanotechnology*, 5(3):200–203, 2010.

[197] Rolf HAM Vossen, Emmelien Aten, Anja Roos, and Johan T den Dunnen. High-resolution melting analysis (hrma)—more than just sequence variant screening. *Human mutation*, 30(6):860–866, 2009.

[198] G Terrance Walker, Melinda S Fraiser, James L Schram, Michael C Little, James G Nadeau, and Douglas P Malinowski. Strand displacement amplification—an isother-

mal, in vitro dna amplification technique. *Nucleic acids research*, 20(7):1691–1696, 1992.

[199] A E Walter, Douglas H. Turner, J Kim, Matthew H. Lyttle, Peter Müller, David H. Mathews, and Michael Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of rna folding. *Proceedings of the National Academy of Sciences of the United States of America*, 91 20:9218–22, 1994.

[200] Boya Wang, Cameron Chalk, and David Soloveichik. Simd||dna: single instruction, multiple data computation with dna strand displacement cascades. In *International Conference on DNA Computing and Molecular Programming*, pages 219–235. Springer, 2019.

[201] Boya Wang, Chris Thachuk, Andrew D. Ellington, and David Soloveichik. *The Design Space of Strand Displacement Cascades with Toehold-Size Clamps*, volume 10467 LNCS of *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, 2017.

[202] Boya Wang, Chris Thachuk, Andrew D Ellington, Erik Winfree, and David Soloveichik. Effective design principles for leakless strand displacement systems. *Proceedings of the National Academy of Sciences*, 115(52):E12182–E12191, 2018.

[203] Xiaofang Wang, Hyun Jeong Lim, and Ahjeong Son. Characterization of denaturation and renaturation of dna for dna hybridization. *Environmental health and toxicology*, 29, 2014.

[204] Maximilian Weitz, Jongmin Kim, Korbinian Kapsner, Erik Winfree, Elisa Franco, and Friedrich C Simmel. Diversity in the dynamical behaviour of a compartmentalized programmable biochemical oscillator. *Nature chemistry*, 6(4):295–302, 2014.

[205] Christopher M Wintersinger, Dionis Minev, Anastasia Ershova, Hiroshi Sasaki, Gokul Gowri, Jonathan Berengut, Franklin Eduardo Corea-Dilbert, Peng Yin, and William Shih. Multi-micron crisscross structures from combinatorially assembled dna-origami slats. *bioRxiv*, 2022.

[206] Brian R. Wolfe, Nicholas J. Porubsky, Joseph N. Zadeh, Robert M. Dirks, and Niles A. Pierce. Constrained Multistate Sequence Design for Nucleic Acid Reaction Pathway Engineering. *Journal of the American Chemical Society*, 139(8):3134–3144, March 2017.

[207] Stephen Wolfram. Statistical mechanics of cellular automata. *Reviews of Modern Physics*, 55(3):601–644, July 1983.

[208] Daniel J Wright, Jamie L Rice, Dawn M Yanker, and Brent M Znosko. Nearest neighbor parameters for inosine uridine pairs in rna duplexes. *Biochemistry*, 46(15):4625–4634, 2007.

[209] Yuhe R Yang, Yan Liu, and Hao Yan. Dna nanostructures as programmable biomolecular scaffolds. *Bioconjugate chemistry*, 26(8):1381–1395, 2015.

[210] SM Yazdi, Ryan Gabrys, and Olgica Milenkovic. Portable and error-free dna-based data storage. *Scientific reports*, 7(1):1–6, 2017.

[211] Kevin Yehl, Andrew Mugler, Skanda Vivek, Yang Liu, Yun Zhang, Mengzhen Fan, Eric R Weeks, and Khalid Salaita. High-speed dna-based rolling motors powered by rnase h. *Nature nanotechnology*, 11(2):184–190, 2016.

[212] Peng Yin, Rizal F Hariadi, Sudheer Sahu, Harry MT Choi, Sung Ha Park, Thomas H LaBean, and John H Reif. Programming dna tube circumferences. *science*, 321(5890):824–826, 2008.

[213] Peng Yin, Hao Yan, Xiaoju G Daniell, Andrew J Turberfield, and John H Reif. A unidirectional dna walker that moves autonomously along a track. *Angewandte Chemie*, 116(37):5014–5019, 2004.

[214] Y Whitney Yin and Thomas A Steitz. Structural basis for the transition from initiation to elongation transcription in t7 rna polymerase. *Science*, 298(5597):1387–1395, 2002.

[215] Y Whitney Yin and Thomas A Steitz. The structural mechanism of translocation and helicase activity in t7 rna polymerase. *Cell*, 116(3):393–404, 2004.

[216] Mingxu You, Fujian Huang, Zhuo Chen, Ruo-Wen Wang, and Weihong Tan. Building a nanostructure with reversible motions using photonic energy. *Acs Nano*, 6(9):7935–7941, 2012.

[217] YT Yu, M Di Shu, and JOAN A Steitz. A new method for detecting sites of 2'-o-methylation in rna molecules. *Rna*, 3(3):324, 1997.

[218] Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. Nupack: Analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1):170–173, 2011.

[219] Anton S Zadorin, Yannick Rondelez, Jean-Christophe Galas, and André Estevez-Torres. Synthesis of programmable reaction-diffusion fronts using dna catalyzers. *Physical review letters*, 114(6):068301, 2015.

[220] Anton S Zadorin, Yannick Rondelez, Guillaume Gines, Vadim Dilhas, Georg Urtel, Adrian Zambrano, Jean-Christophe Galas, and André Estevez-Torres. Synthesis and materialization of a reaction–diffusion french flag pattern. *Nature chemistry*, 9(10):990–996, 2017.

[221] A Zambrano, AS Zadorin, Y Rondelez, A Estévez-Torres, and J-C Galas. Pursuit-and-evasion reaction-diffusion waves in microreactors with tailored geometry. *The Journal of Physical Chemistry B*, 119(17):5349–5355, 2015.

[222] Yang Zeng, Olivia Young, Chris Wintersinger, Frances Anastassacos, James MacDonald, Maxence Dellacherie, Haiqing Bai, Amanda Graveline, Andyna Vernet, Melinda Sanchez, Derin Keskin, Catherine Wu, David Mooney, Ick Chan Kwon, Ju Hee Ryu, and William Shih. Optimizing cpg spatial distribution with dna origami for th1-polarized therapeutic vaccination. 27th International Conference on DNA Computing and Molecular Programming, September 2021.

[223] Chao Zhang, Yumeng Zhao, Xuemei Xu, Rui Xu, Haowen Li, Xiaoyan Teng, Yuzhen Du, Yanyan Miao, Hsiao-chu Lin, and Da Han. Cancer diagnosis with dna molecular computation. *Nature nanotechnology*, 15(8):709–715, 2020.

[224] David Yu Zhang. Towards Domain-Based Sequence Design for DNA Strand Displacement Reactions. In Yasubumi Sakakibara and Yongli Mi, editors, *DNA Computing*

*and Molecular Programming*, volume 6518, pages 162–175. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[225] David Yu Zhang and Erik Winfree. Control of dna strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47):17303–17314, 2009.

[226] Jinny X Zhang, John Z Fang, Wei Duan, Lucia R Wu, Angela W Zhang, Neil Dalchau, Boyan Yordanov, Rasmus Petersen, Andrew Phillips, and David Yu Zhang. Predicting dna hybridization kinetics from sequence. *Nature chemistry*, 10(1):91–98, 2018.

[227] Qian Zhang, Qiao Jiang, Na Li, Luru Dai, Qing Liu, Linlin Song, Jinye Wang, Yaqian Li, Jie Tian, Baoquan Ding, et al. Dna origami as an in vivo drug delivery vehicle for cancer therapy. *ACS nano*, 8(7):6633–6643, 2014.

[228] Chao Zhou, Xiaoyang Duan, and Na Liu. A plasmonic nanorod that walks on dna origami. *Nature communications*, 6(1):1–6, 2015.

[229] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415, 2003.