Spring 2022

Witnessing the Evolution of RNA

Xavier Portillo
Yale University Graduate School of Arts and Sciences, jportillo@me.com

Abstract

Witnessing the Evolution of RNA

Xavier Portillo

2022

The RNA World theory postulates that at a certain point around 4 billion years ago during an RNA World period, a molecule, possibly RNA, developed the ability to self-replicate via polymerization. The theory also suggests that all extant life evolved from this initial RNA-based predecessor. No geological evidence from this RNA World period remains, and a self-replicating RNA polymerase has yet to be discovered or created. However, there is ample evidence from extant life pointing to an RNA-based predecessor. For example, the ribosome is a ribozyme, RNA functions as precursor signaling molecules, riboswitches regulate transcription and translation, and many more. For billions of years and even today, natural selection and the physical laws of nature direct and select molecules deemed fit for the pressure present in their environment. In this way, RNA has been evolving in darkness, unwitnessed by the very entities that are comprised of and still in part regulated by it. With advancements in next-generation sequencing and nucleic acid selections, scientists now witness novel RNA functions that sometimes are accompanied by large structural changes to the RNA secondary structure. Using techniques like *in vitro* and *in vivo* selections, we can now control the selection pressure on these molecules. We can now evaluate selections with next-generation sequencing and machine learning to witness RNA evolution in action.

In the past thirty years, scientists have made major contributions towards the selection of a self-replicating polymerase molecule. In 1995, scientists used *in vitro* selections to create the class I ligase, an RNA that could join two separate strands of RNA through covalent phosphodiester linking. The ligase was selected from a pool of random RNA sequences, much like there would have been on the early earth during this RNA World period. Through further selection, the class I ligase evolved into the molecule termed R18, which had the ability to polymerize from an RNA primer on an RNA template. From the R18 molecule, multiple research groups developed branching RNA polymerase ribozyme lineages, all with the common goal of selecting for a self-replicating molecule. In addition to these branching lineages, non-enzymatic assembly of polynucleotides has also been developed. Despite the significant effort placed in selecting for a self-replicating RNA, such a molecule remains elusive. To understand the role that RNA evolution has played in the development of extant life, we must first understand how RNA evolved to encompass all the roles it serves in the multitude of functions in life today.

My colleagues designed a modified version of the R18 polymerase ribozyme, deemed "WT," which served as the starting sequence for their selections. To evolve the WT polymerase, they developed selection strategies that utilized functional RNAs such as aptamers and self-cleaving ribozymes. They then carried out 52 rounds of either aptamer or self-cleaver selections on this WT population. Every few rounds, a small number of polymerase variants were cloned out of the evolved population. An even smaller number of polymerase variants were biochemically validated to determine if they had increased in polymerization rate. Rather than validating a few cloned sequences, I developed a bioinformatic pipeline that resulted in the ability to tally, align, and cluster all variant sequences in a given selection population. I used the bioinformatic

pipeline on every few rounds of selection within the WT lineage, tallying and tracking the frequency of RNA sequence variations over 52 rounds of selection. I then used this method to validate a novel RNA secondary structure pseudoknot rearrangement, termed P8, in the polymerase population.

I subsequently validated the novel secondary structure rearrangement by using in line probing, an *in vitro* biochemical technique used to determine an RNA's secondary structure or interaction with another molecule or ligand. The secondary structure for six variant sequences that were pulled from the 52 rounds of selection allowed us to witness how the novel secondary structure pseudoknot gradually evolved to a greater fitness peak. Ribozymes, in particular the RNA polymerase ribozyme, are thought to occupy high and isolated fitness peaks that are tied to the molecule's secondary structural elements. Because these secondary structural elements are tightly associated with the ribozyme's optimized fitness peak, exploring alternative structures generally leads to severe negative consequences for fitness. With the bioinformatics pipeline mentioned above, I clustered highly represented variants by the sequence of their P8 region. The P8 pseudoknot structure spontaneously emerged during the evolution process and was optimized and conserved after 28 rounds of selection. Next, I transplanted the novel P8 pseudoknot from the 52-2 variant into the WT sequence. The results of that experiment show that the P8 was necessary, but not sufficient to improve the WT catalytic activity to the 52-2 variant's capacity. The results showed that the novel P8 region was indeed a jump to a higher fitness local.

To my knowledge and after thorough analysis of the literature of the field, this is the first RNA secondary structure remodeling that has been validated and witnessed mid-evolution in a synthetically evolved RNA. Additionally, no such secondary structure remodeling of a natural RNA has been observed. Witnessing the evolution of RNA either synthetic or from nature

provides a powerful means of control and understanding our RNA ancestors, our current RNA components, and any future RNA evolution target we select.

Contained within this document I provide a review of instances where RNA evolution has been witnessed, starting 4 billion years ago following the proposed end of an RNA World transition from RNA- to DNA-protein based life, to the present time. Advances in *in vitro/vivo* selections and next-generation sequencing reveal RNA evolution in action today. Described are instances where scientists have witnessed natural RNA evolution and synthetic RNA evolution, providing evidence for a prehistoric RNA World and a path forward for future RNA evolution advancements. From this breadth of literature, it would appear that the RNA World continues today.

Following this review, I outline my discovery of an RNA polymerase ribozyme that underwent the first observed structural rearrangement of a synthetic RNA, which resulted in an increase in its activity. Furthermore, the RNA polymerase can now synthesize a full length, active copy of its ancestral molecule the class I ligase. While there are other examples of RNA polymerase lineages from other research groups that are mid-evolution, this lineage that I present is the first to catalog a structural rearrangement. I developed bioinformatic means to track the evolution of the RNA polymerase ribozyme. This bioinformatic pipeline can be developed further to track any synthetic or natural RNA evolution over many generations and it provides the foundation to work toward a self-replicating RNA by enabling scientists to design more informed selections. The inevitable discovery of a self-replicating RNA will serve as incontrovertible evidence that RNA has the capacity to initiate Darwinian evolution and may demonstrate a possible route to the discovery of the origins of life as we understand it on earth.

Witnessing the Evolution of RNA

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

In Candidacy for the Degree of

Doctor of Philosophy

by

Xavier Portillo

Dissertation Director: Ronald R. Breaker

May 2022

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Ronald Breaker. His advice and guidance has been critical to my growth as a professional. The passion that Ron has for his work is unmatched, and all student scholars that have the fortune to interact with him, in whatever capacity that may be, have seen and felt Ron's dedication to his students and to science. Whether it is a 3-minute lighting talk, or an undergraduate basic science course, or a professional research fellowship or a graduate program or postdoctoral work, Ron's ability to inspire love for an incredibly abstract topic is second to none. I am incredibly fortunate to have experienced each level described above. This experience has shaped who I am and how I think about science and mentorship, and it serves as the driving force that keeps me moving forward in my career.

In addition to Ron, I would like to thank Jerry Joyce. I have had the opportunity of a lifetime to bridge the proverbial legacy found within academic lineages. This is fitting due to the nature of the research described within this dissertation. Jerry's personality is in many regards similar to mine. Long had I worried that my eight plus years experience in sales and business may at worst hinder and at best not apply to my ability to do good science. But Jerry is the proof. Jerry is arguably one of the best writers I have ever read, and he's an even better scientist. His ability to advocate and communicate his work to any audience is of the highest caliber. I am grateful that Jerry would allow me to work with him not only once, but twice during a global pandemic. I am grateful for Jerry's time, attention, and for his belief in me.

In addition to my two primary advisors, I would like to acknowledge all of the secondary and tertiary advisors that I had the great honor of working beside. Adam

Roth's guidance and patience when I first entered my thesis lab is precisely what I needed at the time. Having just switched research fields (neuroscience to RNA sciences) I felt behind on not only the literature but also on understanding the incredible complexities of biochemistry. Adam never made me feel small, and he helped me realize that there is no project too crazy or ambitious. This work and any work created by me in the foreseeable future will have his influence.

On even keel with Adam is David Horning. David's appreciation for the mastery of his work is unparalleled, and for that, I admire him the most as an equal level colleague. The level of detail and awareness that David shows at the lab bench or in his writing is astonishing. In science, as well as in other fields I'm sure, are held up by the giant shoulders of people like David. He is not only a colleague. but he is also a friend.

There are other lab members that I would like to acknowledge that by no means, are secondary or tertiary to the names mentioned above. These people were the day-to-day north star for my journey. Seth Lyon, Grant Bare, and Neil White all served as professional and personal support when tunnel vision threatened to overwhelm. I could rely on each one of them to, at a moment's notice, drop what they were doing in lab just to talk to me about the latest comedy show or cartoon we had all been binge watching. Comedy references shouted across lab spaces were a staple to maintaining the homeostasis of mental health that is required to succeed through such a challenging endeavor.

Lab members that served a similarly important role in my life are Suddasan, Sidd, Harini, Ankana, Keith, and Megan. Their unwavering support and company made the

journey through PhD seem more like the whimsical storylines from the comedy shows mentioned above.

It must be acknowledged that the final two years of my PhD experience have been under the devastating effects of the global pandemic brought on by the SARS-CoV2. The mental metamorphosis experienced by a scholar throughout their doctoral training is possibly the most mentally challenging journey a person can go through. Combined, the societal/civil uncertainty brought on by the pandemic, civil unrest, and the impact of coming from a disadvantaged background only added insult to injury during this process. Even, as I write this document today, society seems shaken to its core, uncertain of what is to manifest as we all prepare to return to a sense of "normalcy," whatever normalcy can mean after two years of anything but. This process has never felt normal to me, not now, not even when I entered my program pre-pandemic. I am guided by the love of my work. Occasionally distracted but rarely defeated by obstacles still unclear to me, I carry on through these difficult times. With this pandemic, there seems to be a universal bandwidth drain affecting everyone at every level. I am however one of the fortunate ones that gets to immortalize my account in the scriptures of academia. For that I am grateful.

Next I would like to thank my department and many professors including my committee (Josien van Wolfswinkel, Karla Neugebauer, and Matthew Simon) for their unrelenting support of my research and their incredible ability to foster community. Additionally, I would like to thank the professors that did everything they could in and out of the academy to help me through challenging times, Daniel Colón-Ramos, Jeremy Lee, Anton Bennett, Anthony Koleske and Thomas Pollard. I would not have gotten this

far without them. To the students in my department or neighboring departments, I would like to say thank you for easing the pain of the hard times, as well as for your time, attention, and support. Meaghan McGeary witnessed much of the pain. Her support is paramount to my moving onwards and upwards.

Finally, I would like to thank my family and closest loved ones. First, I thank Lydia my partner and best friend. Lydia and I survived the global pandemic together, we started dating at the beginning of the pandemic during a time of great fear and uncertainty in the world, because of you, the long-isolated days sequestered at home were brighter and more beautiful. Thank you for the inspiration and the never-ending deep conversations.

Jorge my best friend from back home has been there through every transition. Jorge has always been quick to send just the right *The Simpson's* quote, quotes that were not only accurate but extremely relevant for the struggle I was facing at the times he sent them. I hope Jorge finds and joins his own Stone Cutter's society one day. I want to thank my sister and brother, Sofy and Alfred. Their never-ending love and belief in me helped me transition from one impossibility to the next. In every way a disadvantaged household was challenged, we were, and at an early age, we worked to help and support our family financially. Sofy has been one of my biggest inspirations, even though she may not believe it. As children, we made many promises about the future to each other. Though growing up in poverty made these promises seem impossible, I'm proud to say that I feel myself now accomplishing one, a feat that would not have been impossible if I had not first dreamed with her.

I am most grateful for my parents, Sofia Ramirez and Javier Portillo Sr. I have always said that my parents' journey to achieve what they have in life will always be a greater

and more precarious than anything I've ever been through, including this. Both immigrated to the US with no formal education or promise of work. Both persevered beyond any reasonable expectations. My parents are the wisest and hardest working people I have ever met. They never pushed me to work on something I didn't want to do. They fostered an environment filled with love and the belief that we should follow our creativity no matter where it took us. Their influence to work hard was seen on the day-to-day. I owe all my success to them. Mamá, Papá, los quiero mucho. Les dedico estos esfuerzos a ustedes.

CHAPTER ONE

Introduction: Witnessing RNA Evolution

Abstract

In this chapter I review the literature on the RNA World theory which postulates that all extant life evolved from and left behind an RNA-based predecessor. 4 billion years following the supposed transition from RNA- to DNA-based life, *in vitro/vivo* selections and next-generation sequencing reveal RNA evolution in action today. Here, we describe instances where scientists have witnessed natural RNA evolution and synthetic RNA evolution, providing evidence for a prehistoric RNA World and a path forward with future RNA evolution advancements.

Introduction

Over the past 40 years, biochemists and other life scientists have developed the concept of an RNA World preceding extant life on earth (Gilbert 1986, Joyce 1989). In addition to postulating RNA as the precursor to DNA-based life, RNA World theory posits that, for the past 4 billion years, RNA may have been a key driver in Darwinian evolution (Joyce 2002). Even though geological records of the prehistoric RNA World may be long gone (Orgel 1968), we observe contemporary clues with riboswitches (Breaker 2012), RNA-derived signaling molecules from a lost chemical language (Nelson and Breaker 2017) and catalytic relics (Breaker 2020) that point to this RNA past and suggest the RNA World's continued relevance today. For example, the ribosome which serves as the catalyst for biological protein synthesis has been determined to have critical RNA components (Yusupov et al. 2001, Ban et al. 2000). Rather than leaving the RNA World behind, we have instead continued to evolve within it. For the past ~4 billion years, the RNA World has been and is still in motion.

Recent advancements in next-generation sequencing, bioinformatics, and machine learning have provided us biochemical and bioinformatic tools to witness RNA evolution in action. For example, ribosomal RNA sequencing reveals the evolutionary linkage between prokaryotes and eukaryotes (Petrov et al. 2015, Bokov and Steinberg 2009, Melnikov et al. 2012). Additionally, many RNAs that once only evolved through natural selection are now subjected to lab-directed selection (Ellington and Szostak 1990, Hammerling et al. 2020). Today, RNAs are used as selection modules (Portillo et al. 2021, Trachman et al. 2019) and as tools to edit and analyze genomes themselves (Liu et al. 2021, Shams et al. 2021). Machine learning has made it possible to use highly targeted guide RNAs to predict genome wide changes (Koblan et al. 2021, Xiang et al. 2021). Illuminating RNA evolution helps elucidate a path from a primordial RNA World to modern day.

While there have been many advancements in RNA biology, there is still much to be known about RNA-driven biological functions. For example, RNA-protein complexes drive many biological functions which are still being investigated today, such as the ribosome, spliceosome, CRISPR-Cas systems, and many other ribonucleoproteins which serve in gene regulation. However, we've only recently discovered that RNA plays a critical role in the examples above but there may also be many undiscovered biological functions. Using biochemical techniques such as *in vitro* selections and bioinformatics, these hidden RNA functions can not only be discovered but also engineered to serve a greater purpose

Now dawns an era where RNA machines witness and control our evolutionary destiny. The following text provides an overview of *in vitro/vivo* and bioinformatic

observations of RNA evolution, as well as a contemporary look into what kinds of RNAs are currently being selected for in future tech. Rather than viewing the RNA World and extant life in separation, this text asserts the RNA World's continued reality. Though it is only recently that we have begun to witness its evolution, these instances of an evolving RNA molecule indicate that life was and is still in many ways directed by RNA.

Evolving in an RNA World

In the pre-RNA World, before nucleic-acid polymers could withstand natural selections, prebiotic chemistry drove the surveying of chemical processes that inevitably led to the formation of a chemically active polymer capable of serving as both data storage and catalyst (Gilbert 1986, Joyce 1989). In that time ~4 billion years ago (Joyce 2002), the environment may have been composed of building blocks of amino acids, hydroxy acids, sugars, purines, pyrimidines and fatty acids (Ferris, Sanchez and Orgel 1968) (Figure 1). From these monomeric units, polymers may have formed that were capable of withstanding selective pressures found in nature. These pressures may have been driven by natural thermocycles, and hydration/dehydration cycles (Admiraal and Herschlag 1999). The energy fluctuations may have led to a natural rearrangement of these polymers. Though little evidence corroborates this analysis of early earth conditions (~4 billion years ago), without archaeological records, much of early RNA evolution occurred in the dark unwitnessed.

From this pre-RNA World, through successive and astronomically vast surveys of different polymer arrangements, a polymer capable of self-replication may have formed (Orgel and Lohrmann 1974). To drive Darwinian evolution, the polymer would need to be able to do two things: create copies of itself and polymerize other RNAs with

mutations to their sequence. The RNA World theory postulates that during this time, an RNA polymerase ribozyme developed that was capable of self-replication. This molecule would serve both as catalyst and template for exponential replication. Mutations, insertions, and deletions into the RNA polymerase ribozyme's sequence would propel Darwinian evolution.

Natural and synthetic RNA selection

The ribosome is one such example that has undergone natural evolution and is now the subject of many selection strategies (Hammerling et al. 2020, Petrov et al. 2015) (Figure 2). With advances in technology, scientists today mimic natural selection in a lab, evolving RNA molecules by carrying out iterative rounds of selection while selecting for specific functions. In these synthetic RNA selections, pressure can be placed on either a random pool of RNA or on a partly degenerative sequence pool (Robertson and Joyce 1990). The survival of each RNA variant in the lab is linked to its fitness to withstand the pressure being selected for (Ellington and Szostak 1990). These synthetic selections are analogous to natural selection acting on RNA molecules deemed fit for the pressure present in their environment. Though RNA for the most part has been selected in nature, unwitnessed by scientists, there are many natural examples of RNA evolution found in extant life today (Breaker 2011, McCown et al. 2017).

Many natural and synthetic instances of RNA evolution have been observed in the laboratory (Portillo et al. 2021). As sequencing technology and bioinformatic tools advance, the evolution of synthetic RNA through the iterative generations of selection can be observed in real time and with increased precision. By pairing these technological advancements with RNA selections, the synthetic evolution of RNA can in a sense,

mirror the natural evolution of RNA, but with the benefit of observation and targeted application.

Novel RNA functions

Selection of RNAs can either be carried out *in vitro* or *in vivo,* resulting in the gain of novel function (Figure 3). The former is commonly referred to as "directed evolution." This method of selection relies on the selectable action of an RNA molecule that is generally created synthetically and does not rely on a living organism. The latter, *in vivo* selections are dependent on a living organism, generally a simple single cell organism. The RNAs selected *in vivo* are much larger in size and are generally RNA-protein machines like the spliceosome (Butt et al. 2019), ribosome, CRISPR gRNAs (Nalefski et al. 2021). In the laboratory, *in vitro* RNA evolution has yielded several novel functions. To list a few, RNA polymerization (Tjhung et al. 2020, Horning and Joyce 2016, Attwater et al. 2018, Cojocaru and Unrau 2021), RNA methylation (Scheitl et al. 2020), RNA catalyzed nucleotide synthesis (Unrau and Bartel 1998), and RNAs with peptidyl transferase activity (Zhang and Cech 1997).

Both *in vitro* and *in vivo* selection techniques allow us to direct and observe RNA evolution in ways that resemble the kind of selection nature places on RNA and that lead to similar introductions of novel functions and increases of function. However, even in a lab-controlled environment, discoveries of novel RNA function occur in the same darkness that much of RNA natural evolution has. We are limited in our ability to witness evolution precisely, as biochemical assays and biophysical snap shots provide evidence but not a sharp enough resolution to witness with base-pair precision RNA in evolutionary motion. Without this precision to witness RNA as it evolves on a molecular

level, scientists can miss secondary structural rearrangements and other key changes as selection pressures reshape the molecule.

Illuminating with next-generation sequencing

The development of genomic sequencing (Church and Gilbert 1984) dawned a new era in RNA evolution. For the first time, the field was able to catalog species based on the rRNA (Kettler et al. 2007), the discovery of CRISPR-Cas9 was possible (Haurwitz et al. 2010), and our understanding of human diseases rapidly advanced. As time progresses, next-generation sequencing becomes more affordable and easily accessible. The number of bioinformatic tools available increases more and more. Each one of these tools allows scientists to witness the evolution of genetic material on a molecular level.

Laboratories are now using artificial intelligence and machine learning to predict the effects brought on by changes in genetic information (Koblan et al. 2021) (Figure 4). Together, these advancements add an additional level of resolution in the witnessing of RNA evolution.

Discussion

Even though the proposed RNA World period may have long passed, RNA has both evolved into and continues to evolve alongside extant life today. RNA vestiges of a time past remain relevant in biology today. As we better understand how RNA functions within DNA-protein based systems, one cannot help but suspect that the RNA World period never ended. RNA is still very much in charge, serving a pivotal role behind the DNA and protein components inside living cells and viruses. A chapter in the prehistoric RNA World ends, and a new chapter into a brave new RNA World begins.

Altogether, advancements in RNA selection, next-generation sequencing, and bioinformatics, provide a path forward for greater control and design of RNA evolution in the laboratory. Even now, advancements are under way using CRISPR-guide RNAs, RNA-regulated synthetic genomes and many other smaller RNAs (Table 1). Technologies using RNA evolution are at the forefront of many fields in medicine. Personalized medicine now relies on sequencing a patient's genetic code to better understand whether a therapeutic will be effective. In the midst of a global pandemic, an mRNA therapeutic was humanity's silver bullet against a devastating single-stranded RNA genome virus, SARS-CoV-2. Understanding RNA evolution both past and present is a critical lesson we must learn so that we can better respond to whatever challenges extant life might face in the future.

What follows is a description of my work to better understand RNA evolution by witnessing it in action. An RNA polymerase ribozyme capable of self-replication could have been the catalyst to initiate Darwinian evolution. By using conventional biochemistry, coupled with powerful next-generation sequencing, the evolution of the most advanced RNA polymerase ribozyme was witnessed. RNA is generally thought to reach and stay in desirable fitness locales. However, through sequence variation and selection, a structural rearrangement found immediately adjacent to the RNA polymerase ribozyme's catalytic core has allowed an RNA to pierce past its current fitness locale. My work described here catalogs what could be one of the first critical steps in RNA evolution.

References

Kirby, A. J. & Younas, M. The reactivity of phosphate esters. Diester hydrolysis. J. Chem. Soc. B 510–513 (1970).

Admiraal, S. J. & D. Herschlag (1999) Catalysis of phosphoryl transfer from ATP by amine nucleophiles. *J. Am. Chem. Soc.,* 121.

Attwater, J., A. Raguram, A. S. Morgunov, E. Gianni & P. Holliger (2018) Ribozyme-catalysed RNA synthesis using triplet building blocks. *elife,* 7.

Ban, N., P. Nissen, J. Hansen, P. B. Moore & T. A. Steitz (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science,* 289**,** 905-20.

Bokov, K. & S. V. Steinberg (2009) A hierarchical model for evolution of 23S ribosomal RNA. *Nature,* 457**,** 977-80.

Breaker, R. R. (2011) Prospects for riboswitch discovery and analysis. *Mol. Cell,* 43.

Breaker, R. R. (2012) Riboswitches and the RNA World. *Cold Spring Harb. Perspect. Biol.,* 4.

Breaker, R. R. (2020) Imaginary ribozymes. *ACS Chem. Biol.,* 15.

Butt, H., A. Eid, A. A. Momin, J. Bazin, M. Crespi, S. T. Arold & M. M. Mahfouz (2019) CRISPR directed evolution of the spliceosome for resistance to splicing inhibitors. *Genome Biol,* 20**,** 73.

Church, G. M. & W. Gilbert (1984) Genomic sequencing. *Proc Natl Acad Sci U S A,* 81**,** 1991-5.

Cojocaru, R. & P. J. Unrau (2021) Processive RNA polymerization and promoter recognition in an RNA World. *Science,* 371**,** 1225-1232.

Ellington, A. D. & J. W. Szostak (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature,* 346.

Ferris, J. P., R. A. Sanchez & L. E. Orgel (1968) Studies in prebiotic synthesis III. Synthesis of pyrimidines from cyanoacetylene and cyanate. *J. Mol. Biol.,* 33.

Gilbert, W. (1986) The RNA world. *Nature,* 319.

Hammerling, M. J., B. R. Fritz, D. J. Yoesep, D. S. Kim, E. D. Carlson & M. C. Jewett (2020) In vitro ribosome synthesis and evolution through ribosome display. *Nat Commun,* 11**,** 1108.

Haurwitz, R. E., M. Jinek, B. Wiedenheft, K. Zhou & J. A. Doudna (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science,* 329**,** 1355-8.

Horning, D. P. & G. F. Joyce (2016) Amplification of RNA by an RNA polymerase ribozyme. *Proc Natl Acad Sci U S A,* 113**,** 9786-91.

Joyce, G. F. (1989) RNA evolution and the origins of life. *Nature,* 338**,** 217-24.

Joyce, G. F. (2002) The antiquity of RNA-based evolution. *Nature,* 418**,** 214-21.

Kettler, G. C., A. C. Martiny, K. Huang, J. Zucker, M. L. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, C. Steglich, G. M. Church, P. Richardson & S. W. Chisholm (2007) Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet,* 3**,** e231.

Koblan, L. W., M. Arbab, M. W. Shen, J. A. Hussmann, A. V. Anzalone, J. L. Doman, G. A. Newby, D. Yang, B. Mok, J. M. Replogle, A. Xu, T. A. Sisley, J. S. Weissman, B. Adamson & D. R. Liu (2021) Efficient C*G-to-G*C base editors developed using CRISPRi screens, target-library analysis, and machine learning. *Nat Biotechnol,* 39**,** 1414-1425.

Liu, T. Y., G. J. Knott, D. C. J. Smock, J. J. Desmarais, S. Son, A. Bhuiya, S. Jakhanwal, N. Prywes, S. Agrawal, M. Diaz de Leon Derby, N. A. Switz, M. Armstrong, A. R. Harris, E. J. Charles, B. W. Thornton, P. Fozouni, J. Shu, S. I. Stephens, G. R. Kumar, C. Zhao, A. Mok, A. T. Iavarone, A. M. Escajeda, R. McIntosh, S. Kim, E. J. Dugan, I. G. I. T. Consortium, K. S. Pollard, M. X. Tan, M. Ott, D. A. Fletcher, L. F. Lareau, P. D. Hsu, D. F. Savage & J. A. Doudna (2021) Accelerated RNA detection using tandem CRISPR nucleases. *Nat Chem Biol,* 17**,** 982-988.

McCown, P. J., K. A. Corbino, S. Stav, M. E. Sherlock & R. R. Breaker (2017) Riboswitch diversity and distribution. *RNA,* 23.

Melnikov, S., A. Ben-Shem, N. Garreau de Loubresse, L. Jenner, G. Yusupova & M. Yusupov (2012) One core, two shells: bacterial and eukaryotic ribosomes. *Nat Struct Mol Biol,* 19**,** 560-7.

Nalefski, E. A., N. Patel, P. J. Y. Leung, Z. Islam, R. M. Kooistra, I. Parikh, E. Marion, G. J. Knott, J. A. Doudna, A. M. Le Ny & D. Madan (2021) Kinetic analysis of Cas12a and Cas13a RNA-Guided nucleases for development of improved CRISPR-Based diagnostics. *iScience,* 24**,** 102996.

Nelson, J. W. & R. R. Breaker (2017) The lost language of the RNA World. *Sci Signal,* 10.

Orgel, L. E. (1968) Evolution of the genetic apparatus. *J Mol Biol,* 38**,** 381-93.

Orgel, L. E. & R. Lohrmann (1974) Prebiotic chemistry and nucleic acid replication. *Acc. Chem. Res.,* 7.

Petrov, A. S., B. Gulen, A. M. Norris, N. A. Kovacs, C. R. Bernier, K. A. Lanier, G. E. Fox, S. C. Harvey, R. M. Wartell, N. V. Hud & L. D. Williams (2015) History of the ribosome and the origin of translation. *Proc Natl Acad Sci U S A,* 112**,** 15396-401.

Portillo, X., Y. T. Huang, R. R. Breaker, D. P. Horning & G. F. Joyce (2021) Witnessing the structural evolution of an RNA enzyme. *elife,* 10.

Robertson, D. L. & G. F. Joyce (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature,* 344.

Scheitl, C. P. M., M. Ghaem Maghami, A. K. Lenz & C. Hobartner (2020) Site-specific RNA methylation by a methyltransferase ribozyme. *Nature,* 587**,** 663-667.

Shams, A., S. A. Higgins, C. Fellmann, T. G. Laughlin, B. L. Oakes, R. Lew, S. Kim, M. Lukarska, M. Arnold, B. T. Staahl, J. A. Doudna & D. F. Savage (2021) Comprehensive deletion landscape of CRISPR-Cas9 identifies minimal RNA-guided DNA-binding modules. *Nat Commun,* 12**,** 5664.

Tjhung, K. F., M. N. Shokhirev, D. P. Horning & G. F. Joyce (2020) An RNA polymerase ribozyme that synthesizes its own ancestor. *Proc Natl Acad Sci U S A,* 117**,** 2906-2913.

Trachman, R. J., 3rd, A. Autour, S. C. Y. Jeng, A. Abdolahzadeh, A. Andreoni, R. Cojocaru, R. Garipov, E. V. Dolgosheina, J. R. Knutson, M. Ryckelynck, P. J. Unrau & A. R. Ferre-D'Amare (2019) Structure and functional reselection of the Mango-III fluorogenic RNA aptamer. *Nat Chem Biol,* 15**,** 472-479.

Unrau, P. J. & D. P. Bartel (1998) RNA-catalysed nucleotide synthesis. *Nature,* 395**,** 260-3.

Xiang, X., G. I. Corsi, C. Anthon, K. Qu, X. Pan, X. Liang, P. Han, Z. Dong, L. Liu, J. Zhong, T. Ma, J. Wang, X. Zhang, H. Jiang, F. Xu, X. Liu, X. Xu, J. Wang, H. Yang, L. Bolund, G. M. Church, L. Lin, J. Gorodkin & Y. Luo (2021) Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat Commun,* 12**,** 3238.

Yusupov, M. M., G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate & H. F. Noller (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science,* 292**,** 883-96.

Zhang, B. & T. R. Cech (1997) Peptide bond formation by in vitro selected ribozymes. *Nature,* 390**,** 96-100.

Figures

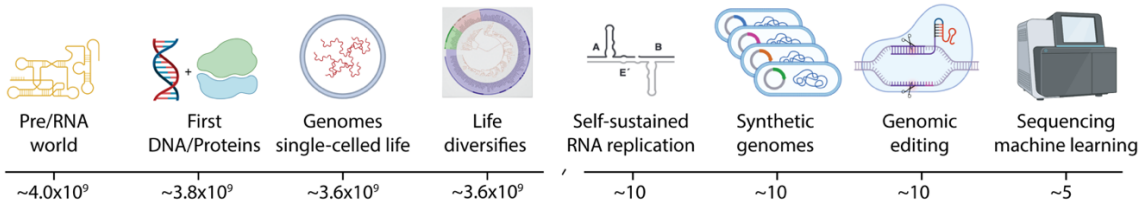


Figure 1-1. Timeline of RNA/DNA evolution. Timeline splits from life's early history on earth (~4 billion years ago) to the present day. Approximate dates shown in years.

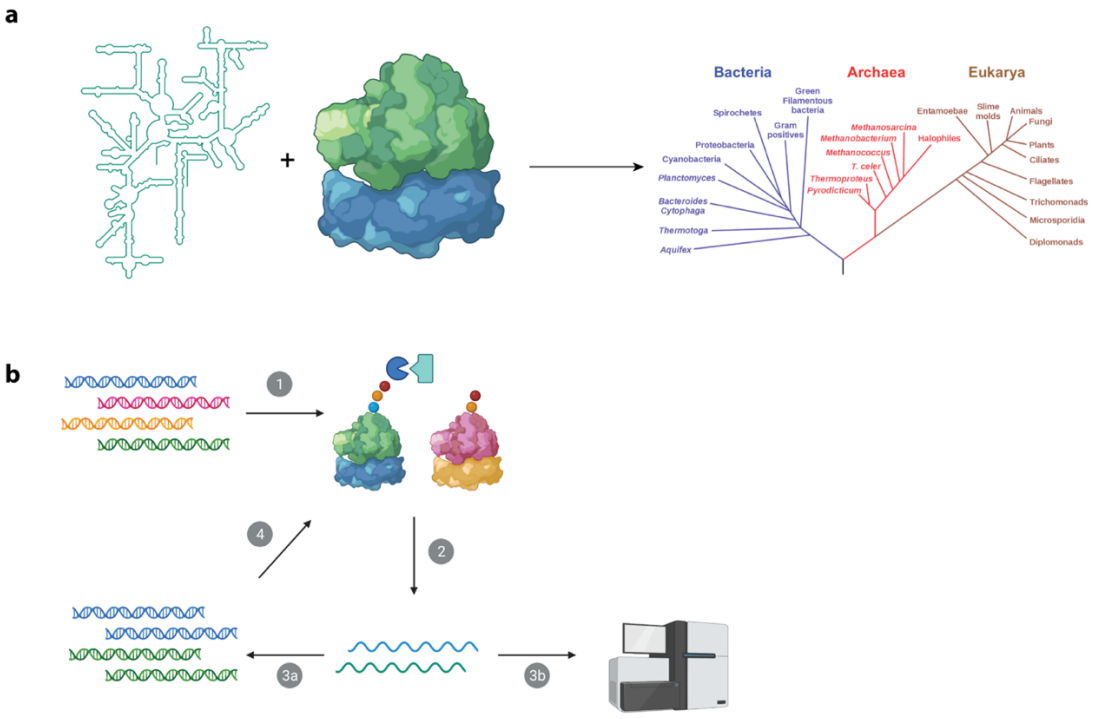


Figure 1-2. Ribosome evolution natural and synthetic. a, Ribosome evolution in nature. b, Synthetic ribosome evolution (1) Starting ribosome sequence variant pool is subjected to selection pressure, (2) Selected ribosome variant RNA isolated, (3a) Selected ribosome variant RNA amplified, (3b) Selected ribosome variant RNA sequenced, (4) Selected ribosome sequence variant pool is subjected to selection pressure.

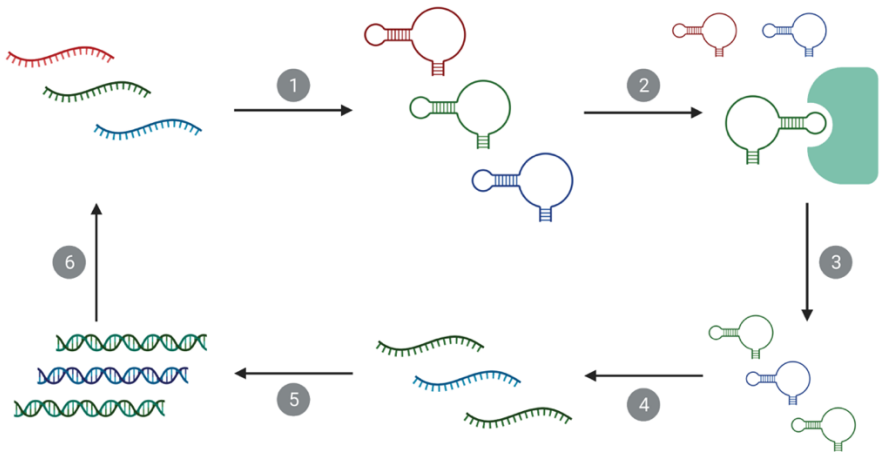


Figure 1-3. Example of *In vitro* selection of RNA. (1) Random population of RNA to be selected adopts secondary structure. (2) Selection module is introduced to the pool of RNA. (3) Selected RNA variants are isolated out of the population. (4) Selected variants are purified. (5) Selected variants are amplified. (6) Enriched population re-enters selection scheme.

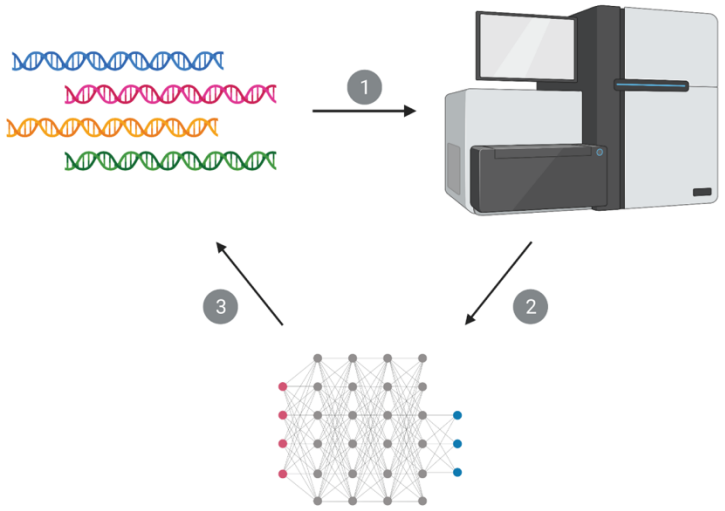


Figure 1-4. Next-generation sequencing bioinformatics pathway for RNA selections. (1) Selected RNA variants from selection study are sequenced. (2) Machine learning identifies rising variant populations. (3) Sequential populations optimized prior to subsequent selection.

Tables

RNA components selected	Type of advancement	Selection type
I. CRISPR-Cas systems	Medicine	*In vivo*
II. RNA selected aptamers	Medicine	*In vitro*
III. Synthetic genomes	Application	*In vitro*
IV. Spliceosome	Medicine	*In vitro*
V. RNA polymerase ribozyme	Application	*In vitro*

Table 1-1. Recent advancements in selected RNA systems

CHAPTER TWO

Witnessing the Structural evolution of an RNA polymerase enzyme

Largely adapted from the following publication:

Portillo, X., Huang, Y., Breaker, R. R., Horning, D.P., Joyce, G.F. (2021) Witnessing the Structural evolution of an RNA polymerase enzyme. *eLife*

Co-author Contributions

The subsequent section's (Chapter 2), text was largely adapted from the published manuscript titled "Witnessing the Structural Evolution of an RNA Enzyme" published in *eLife* on September 23, 2021. Yu-Ting Huang assisted in designing methodology, data analysis and carried out the experiments for the site-directed mutagenesis studies. David Horning carried out the additional 14 rounds of selection from the previously published "38-6" RNA polymerase ribozyme molecule, used to identify the 52-2 variant. David also carried out biochemical kinetic studies. I wrote the first draft and assisted in reviewing and editing subsequent drafts. Gerald Joyce and David Horning wrote, edited, and reviewed subsequent drafts including the final submitted manuscript. Ronald Breaker and Gerald Joyce advised and supervised the project. Deep sequencing of selected populations and polymerized products (used in fidelity studies) were carried out at Yale Center for Genome Analysis and the Salk Next Generation Sequencing Core, respectively. I developed the bioinformatics pipeline, generated the RNA sequencing libraries, performed biochemical kinetic assays and carried out the In-line probing assays used to identify and validate the structural rearrangement defined as the novel P8 pseudoknot. Figure legends contain specific contributions from co-authors.

Abstract

In Chapter 2, my colleagues and I discovered that an RNA polymerase ribozyme that has been the subject of extensive directed evolution efforts has attained the ability to synthesize complex functional RNAs, including a full-length copy of its own evolutionary ancestor. During the course of evolution, the catalytic core of the ribozyme has undergone a major structural rearrangement, resulting in a novel tertiary structural element that lies in close proximity to the active site. Through a combination of site-directed mutagenesis, structural probing, and deep sequencing analysis, the trajectory of evolution was seen to involve the progressive stabilization of the new structure, which provides the basis for improved catalytic activity of the ribozyme. Multiple paths to the new structure were explored by the evolving population, converging upon a common solution. Tertiary structural remodeling of RNA is known to occur in nature, as evidenced by the phylogenetic analysis of extant organisms, but this type of structural innovation had not previously been observed in an experimental setting. Despite prior speculation that the catalytic core of the ribozyme had become trapped in a narrow local fitness optimum, the evolving population has broken through to a new fitness locale, raising the possibility that further improvement of polymerase activity may be achievable.

Introduction

Directed evolution in the laboratory has proven to be a powerful means for obtaining proteins and nucleic acids with desired functional properties. Starting from molecules that contain either regions of random sequence or a defined sequence that has been diversified, one carries out iterative rounds of selection and amplification to obtain ever more fit variants in pursuit of the desired phenotype. This process is analogous to Darwinian evolution in biology, except that the fitness criteria are imposed by the experimenter, rather than being the result of natural selection.

Whereas biological evolution has been operating on Earth for billions of years across highly diverse environments, directed evolution experiments typically involve only a dozen 'generations' and are confined to narrowly defined reaction conditions. Thus, it is perhaps not surprising that the trajectory of evolution in the laboratory tends to follow a narrow path. Once a functional motif has been defined, subsequent rounds of evolution typically refine rather than remodel that motif. In some cases, the experimenter intervenes by appending a new region of random or defined sequence that can evolve into a new structural domain (Jaeger et al., 1999; Johnston et al., 2001; Ikawa et al., 2004) or evolves a different function that emerges together with a new structural motif (Lorsch and Szostak, 1994; Schultes and Bartel, 2000; Huang and Szostak, 2003). However, more extensive evolution appears to be required to achieve tertiary structural remodeling while maintaining the same function throughout the evolutionary process.

Here, from the perspective of 52 consecutive rounds of directed evolution under progressively more demanding selection constraints, an RNA polymerase ribozyme was seen to undergo a tertiary structural change, similar to changes that are inferred to have

occurred in nature based on phylogenetic analyses (Gutell et al., 1994; Williams and Bartel, 1996; Pfingsten et al., 2007). Multiple evolutionary pathways were explored by the evolving population of RNA molecules as they transitioned from one structural configuration to another, ultimately converging upon a new fold that results in improved catalytic activity. The details of this transition are witnessed by a combination of structural, biochemical, and deep sequencing analyses, providing a clear-eyed view of the molecular evolution of structural innovation.

The RNA-catalyzed polymerization of RNA has received special attention because it is thought to be the central function of the 'RNA World', a time in the early history of life, prior to the emergence of DNA and proteins, when RNA served as both the genetic material and the chief agent of catalytic function (Crick, 1968; Gilbert, 1986; Joyce, 2002). An RNA enzyme that catalyzes the RNA templated copying of RNA could, in principle, generate additional copies of itself and thus serve as the basis for self-sustained Darwinian evolution. No such enzyme currently exists, although diligent efforts by several laboratories have used directed evolution to isolate an RNA ligase ribozyme from a population of random-sequence RNAs (Bartel and Szostak, 1993; Ekland et al., 1995), then drive the ribozyme to function as an ever more efficient RNA polymerase (Johnston et al., 2001; Zaher and Unrau, 2007; Wochner et al., 2011; Horning and Joyce, 2016; Cojocaru and Unrau, 2021), now with the ability to synthesize RNAs as complex as the parental ligase (Attwater et al., 2018; Tjhung et al., 2020).

All of the previously described descendants of the original class I ligase ribozyme retain the same catalytic core. In 2001, Bartel and colleagues appended 76 random-sequence nucleotides to the 3' end of the ligase and selected for its ability to catalyze the

polymerization of nucleoside 5'-triphosphates (NTPs). This effort resulted in a novel 'accessory domain' that enables the addition of up to 14 successive NTPs on the most favorable templates (Johnston et al., 2001). Following further evolutionary optimization, the maximum length of extension by the polymerase was increased to 20 NTPs (Zaher and Unrau, 2007). Holliger and colleagues then added 48 random-sequence nucleotides to the 5' end of the ribozyme and selected for its ability to catalyze multiple NTP additions. This procedure resulted in the discovery of a 'processivity tag' that forms a region of Watson-Crick pairing between the 5' end of the ribozyme and the 5' end of the template, enabling addition of up to 95 NTPs on a template that contains multiple repeats of an especially favorable sequence (Wochner et al., 2011). Throughout these many rounds of evolution, only a single point mutation became fixed within the core ligase domain, converting a G-C pair to a G-U wobble pair. However, the combination of the added accessory domain and processivity tag, hereafter referred to as the 'wild type', provided a more robust polymerase that made it possible to impose more stringent selection criteria going forward.

In two subsequent studies, the wild-type polymerase ribozyme was further evolved by requiring it to synthesize functional RNAs, with selection of the ribozyme being dependent on the function of the synthesized product. In the first study, the ribozyme was required to synthesize two different RNA aptamers, each involving the copying of a challenging template (Horning and Joyce, 2016). The resulting '24-3' polymerase, obtained after 24 rounds of evolution, has substantially improved activity compared to its predecessors, especially when copying structured templates with heterogeneous base composition. It is able to synthesize the entire 33-nucleotide hammerhead ribozyme,

which became the requirement for selection in the second study. Another 14 rounds of evolution were then carried out, culminating in the '38-6' polymerase, which is ~10 -fold more active than the 24-3 polymerase and can more efficiently synthesize complex RNA products, such as yeast phenylalanyl-tRNA (Tjhung et al., 2020).

This lineage continues in the present study, which began with a population of variants of the 38-6 polymerase that had been randomized at a frequency of 10 % per nucleotide position, and entailed 14 additional rounds of evolution that sought to improve both the activity and fidelity of the polymerase. The resulting '52-2' polymerase is indeed further improved, but also reveals that the ribozyme underwent structural rearrangement of its catalytic core, enabled by 11 substitution, 2 insertion, and 2 deletion mutations that accumulated over the course of evolution starting from the wild-type polymerase. An existing stem element became shortened while a new stem element was formed, together creating a pseudoknot structure that lies in close proximity to the ribozyme's active site. This new structure became stabilized over time through the sampling and fixation of successive mutations, providing a compelling demonstration of the blind inventiveness of Darwinian evolution. The new structure also shows that the ribozyme was not trapped on a local fitness peak, but instead is actively evolving, with the opportunity to explore novel regions of sequence space.

In this chapter, my colleagues and I present evidence that confirm that the RNA polymerase variant 52-2 has undergone a secondary structural rearrangement that has resulted in an increase in catalytic activity. Of note, this is the first account in which a synthetically selected RNA has been observed to have structurally evolve. Most importantly, the 52-2 RNA polymerase now has the ability to synthesize a full-length

copy of its evolutionary ancestor, this gets the field one step closer to finding a molecule capable of self-replication.

Results

Advanced evolution of an RNA polymerase ribozyme

The 38-6 polymerase ribozyme contains 182 nucleotides, with 20 substitution, 4 insertion, and 2 deletion mutations compared to the wild-type ribozyme (Tjhung et al., 2020). David Horning introduced random mutations throughout the 38-6 polymerase at a frequency of 10 % per nucleotide position, excluding 14 nucleotides at the 5' end and 15 nucleotides at the 3' end that served as primer binding sites for amplification of the selected RNAs. A starting population of approximately three copies each of 4 × 1014 different RNAs was used to initiate subsequent rounds of directed evolution, requiring the polymerase to synthesize a functional hammerhead ribozyme and further requiring the polymerase to operate under conditions of reduced Mg2+ ion concentration (Supplementary file 1). The latter constraint sought to provide conditions that may be conducive to increased polymerase fidelity (Eckert and Kunkel, 1990; Achuthan et al., 2014) and to the reduced degradation of RNA. Fourteen rounds of evolution were carried out, performing error-prone PCR during most rounds to maintain genetic diversity in the population, although not during the final two rounds so that the population could converge on the fittest variants.

Following the 14th round, 52 rounds in total relative to the wild-type polymerase, 30 individuals were cloned from the population and sequenced. The majority of these individuals, including the 52-2 polymerase that dominated the final population, exhibited increased activity compared to the 38-6 polymerase in the presence of either standard (200 mM) or reduced (50 mM) concentrations of Mg2+. A dominant clone, termed '52-2', was chosen for further study. It contains four mutations relative to the 38-6

polymerase, is 3-fold more efficient in synthesizing the hammerhead ribozyme (Figure 1A), and is 23-fold more efficient in synthesizing the class I ligase (Figure 1B). The synthesized ligase is catalytically active, with an observed rate of RNA-templated RNA ligation of 0.31 ± 0.02 hr–1, which corresponds to a rate acceleration of 1500-fold compared to the uncatalyzed reaction (Figure 1C). However, this rate is substantially lower than that of the class I ligase synthesized by T7 RNA polymerase, which has an observed rate of 6.3 ± 0.6 min–1 under the same reaction conditions.

Based on the modest fidelity of the 24-3 and 38-6 polymerases (Tjhung et al., 2020), it is likely that ligase molecules synthesized by the 52-2 polymerase contain multiple mutations, which may reduce or eliminate catalytic activity. Deep sequencing was carried out to analyze both the hammerhead and class I ligase ribozymes synthesized by the 52-2 polymerase. For the hammerhead, synthesized in the presence of either 200 or 50 mM Mg2+, the average fidelity per nucleotide position was 91.7 % or 94.4%, respectively (Supplementary file 2). The most common mutations are the result of G•U wobble pairing, with all types of mutations being less frequent in the presence of the lower concentration of Mg2+. For the class I ligase, synthesized in the presence of 200 mM Mg2+, the average fidelity was 84.1 %. There are an average of 12 mutations per copy of the ligase, which explains its reduced activity compared to that of the protein-synthesized material.

Sequence changes over the course of evolution

The 24-3, 38-6, and 52-2 polymerases represent distinct points along a lineage that has diverged substantially from the wild type. Overall, the 52-2 polymerase differs by 26 mutations compared to the wild type, which corresponds to 14 % of its total sequence

(Figure 2—figure supplement 1). Fifteen of these mutations are within the catalytic core. Two core mutations that arose between the 38-6 and 52-2 polymerase are notable because they change A and U residues at positions 15 and 85 to C and G, respectively. Position 15 is the first nucleotide beyond the fixed primer binding site and had been thought to be part of a single-stranded region (termed J1/3) that helps to position a catalytic Mg2+ ion within the active site of the parental ligase ribozyme (Shechner et al., 2009; Shechner and Bartel, 2011). Position 85 lies within what was thought to be a loop region that closes the P7 stem of the catalytic core, but otherwise has no functional importance (Ekland and Bartel, 1995). Yet, these two mutations arising in concert raised suspicion that they might form a Watson-Crick pair, which would require a very different structural arrangement within the catalytic core.

Neither the J1/3 nor the P7 portion of the ribozyme was mutated during the many rounds of evolution leading from the ligase to the wild-type polymerase. However, both of these regions accumulated numerous mutations during the subsequent rounds of evolution, with seven changes in the 24-3 polymerase, six more in the 38-6 polymerase, and two more in the 52-2 polymerase (Figure 2B). Taken together, these mutations suggest that a new stem, termed P8, has evolved. One strand of this stem contains six nucleotides derived from the J1/3 region, while the complementary strand contains six nucleotides derived from the P7 stem-loop. The proximal portion of the P7 stem appears to remain intact. The new P8 stem would result in a pseudoknot structure that alters the orientation of the P7-derived nucleotides in a manner that is mutually exclusive to the prior core structure of the ribozyme.

Mutagenesis studies in support of the novel structure

Site-directed mutagenesis studies were carried out by Yu Ting Huang to investigate the hypothesized structural rearrangement of the catalytic core. Putative base pairs within the P7 and P8 stems were mutated to identify disruptive and compensatory mutations that may be indicative of Watson-Crick pairing. Each of the six base pairs within the proposed P8 stem was mutated on each of the two strands, together with their combined mutation that would restore complementarity. All 18 of these constructs were evaluated for polymerase activity using a moderately challenging template that required synthesis of the sequence 5'-GUGUGGAGUGACCUCUCCUGUGUGAGUG-3'. On this template, the 52-2 polymerase extends a primer to form full-length products in 20 % yield after 30 min. Each of the single mutations reduced this activity by at least 50-fold, whereas most of the corresponding double mutations restored activity (Figure 3A).

For three central pairs of the stem (C12-G88, A13-U87, and C14-G86), activity was nearly fully restored in the corresponding double mutant (G12-C88, U13-A87, and G14-C86). For the two adjacent pairs (G11-C89 and C15-G85), activity was substantially increased in the corresponding double mutant (C11-G89 and G15-C85), although not fully to the level of the 52-2 polymerase. For the most distal pair (A16-U84), mutation of A16 was highly disruptive and activity could not be restored through compensatory mutation. In the class I ligase, the nucleotide corresponding to A16 is known to make a base-specific contact with the template-primer duplex (Shechner and Bartel, 2011), and therefore may be required to play a similar role in the polymerase. Nonetheless, sequence covariation in support of this distal pairing is observed among other evolved variants

(e.g., G16-C84 and C16-G84), perhaps requiring accompanying mutations to compensate for substitution of A16.

These data supporting the existence of the P8 stem need to be reconciled with the consequences for the P7 stem. Nucleotides 84–89 were previously required to form the P7 stem-loop, with U87, G88, and C89 engaging in Watson-Crick pairs to close one end of the stem. However, mutations that would be expected to disrupt the pairing of either U87-A82 or G88-C81 did not have a deleterious effect, nor was there a beneficial effect of mutations that would be expected to provide additional pairing of C89-G80 (Figure 3—figure supplement 1). Thus, the distal portion of the P7 stem-loop no longer appears to form, with some of those nucleotides instead helping to form the new P8 stem. Kinetic studies were carried out to assess more quantitatively the effect of disruptive and compensatory mutations within the P8 stem. The 52-2 polymerase was compared to the C12G and G88C single mutants, as well as the corresponding double mutant, in a reaction involving an 11-nucleotide templating region that enables measurement of both the rate of the first NTP addition and the average rate of NTP additions across the entire template sequence. The full-length extension product of this reaction has the sequence 5'-UGCGAAGCGUG-3'.

David Horning and I determined the reaction kinetics of the 52-2 polymerase. The polymerase exhibits first-order kinetics, with a kobs of 0.031 min–1. However, there is a substantial burst phase, with 18 % of the template-bound primers extended within the first 10 s of the reaction (Figure 3B). Multiple nucleotide additions are seen during this short burst phase, including full-length products, suggesting that there is a subpopulation of molecules with a very rapid rate of reaction. The average rate of NTP addition across

the entire template during the first 30 s of the reaction is 3.1 min–1, which is the fastest rate measured for a polymerase ribozyme (Figure 3C). The C12G and G88C mutant polymerases each have substantially lower activity, with a kobs of 0.0039 and 0.0056 min–1, and an average rate of NTP addition during the initial phase of the reaction of 0.033 and 0.45 min–1, respectively. The amplitude of the burst phase and the rate of the first NTP addition are also substantially lower for the two single mutants. For the compensatory double mutant, however, all of these rates are restored to nearly that of the 52-2 polymerase, with a kobs of 0.030 min–1, burst-phase amplitude of 17%, and average rate of NTP addition during the first 30 s of the reaction of 2.3 min–1.

The 52-2 polymerase was tested with substantially longer templates to determine the extent to which NTP addition can continue in the burst phase. These templates encoded either 5 or 10 repeats of the sequence 5'-UGCGAAGCGUG-3', which is known to be especially favorable for synthesis by the wild-type polymerase (Wochner et al., 2011). Again with a burst amplitude of ~20%, the burst phase was found to continue, with detectable full-length products after 5 min for the template with 5 repeats and after 10 min for the template with 10 repeats (Figure 3—figure supplement 2). By 20 min, the yield of full-length products was 12.9% and 2.3%, respectively.

The mutagenesis studies suggest that the new pseudoknot structure is necessary, but not whether it is sufficient, for the improved catalytic activity of the 52-2 polymerase, which contains 15 additional mutations outside the region of the pseudoknot (Figure 2A). I designed a chimeric molecule by 'transplanting' the pseudoknot onto the wild-type polymerase background, but otherwise maintaining the wild-type sequence (Figure 2—figure supplement 2A,B). This chimeric molecule has comparable activity to that of the

52-2 polymerase, including the initial burst phase behavior for NTP addition, which is not seen for the wild type (Figure 2—figure supplement 2C). It is notable that for the chimeric molecule, and especially for the 52-2 polymerase, primer extension continues a few nucleotides beyond the templating region and into the oligoadenylate spacer that links the templating region to the processivity tag.

Structural probing of the wild-type and evolved polymerases

I mapped the secondary structure of the wild-type, 24-3, 38-6, and 52-2 polymerases by in-line probing, a technique that measures the susceptibility of each phosphodiester linkage to spontaneous cleavage (Soukup and Breaker, 1999; Regulski and Breaker, 2008). Unstructured single-stranded regions of RNA, such as loops and linkers, are more susceptible to spontaneous cleavage because their greater backbone flexibility allows the ribose 2'-hydroxyl to access an in-line geometry with regard to the adjacent phosphate, as is required for the cleavage event.

Comparison of the wild-type and 52-2 polymerases revealed that the latter is less susceptible to cleavage at nucleotide positions 11–14, which correspond to one of the two strands of the P8 stem (Figure 4A). Nucleotides 86–88, which correspond to the other strand of P8, are protected in both polymerases, although these nucleotides would be part of the P7 stem in the wild-type polymerase. Conversely, nucleotides 79–81 are more susceptible to cleavage in the 52-2 polymerase, these nucleotides no longer being part of the P7 stem. The retained portions of the P7 stem, nucleotides 73–78 and 91–96, are well protected from cleavage in both the wild-type and 52-2 polymerases. Note that there is a two-nucleotide insertion in the 52-2 polymerase at positions 89–90, which lies between the 3' end of the P8 stem and 5' end of the P7 stem, and there is strong cleavage at the

unpaired nucleotide A90 in the 52-2 polymerase. All of these data are consistent with a rearrangement of the catalytic core that results in formation of a novel pseudoknot structure involving the P8 stem.

In-line probing of the 24-3 and 38-6 polymerases showed that these ribozymes have intermediate structural features relative to the wild-type and 52-2 polymerases (Figure 4—figure supplement 1). For all four polymerases, the degree of spontaneous cleavage was measured for each nucleotide in the P7 and P8 regions and mapped onto both the original and evolved structures (Figure 4B). In the 24-3 polymerase, there is reduced susceptibility to cleavage at nucleotide positions 11–14 and enhanced cleavage at positions 79–81, both of which are more pronounced in the 38-6 polymerase. The two-nucleotide insertion first appears in the 38-6 polymerase, with some susceptibility to cleavage at position 90, and this susceptibility becomes more pronounced in the 52-2 polymerase. The retained portion of the P7 stem, nucleotides 73–78 and 91–96, is well protected from spontaneous cleavage for all of the ribozymes in the evolutionary lineage.

Sequence variation over the course of evolution

I developed a deep sequencing analysis pipeline to investigate the population dynamics that underly the emergence of the novel structure. Sequences were obtained after rounds 6, 8, 11, 12, 14, 16, 18, 21, 24, 27, 28, 31, 34, 36, 38, 43, 46, 49, and 52. These sequences were aligned and clustered based on the region encompassing the P8 stem (Figure 5; Supplementary file 3). Clusters representing >1% of the population in any given round were identified and those that were present in only a single round at <5 % frequency were ignored. This analysis resulted in a total of 105 clusters, representing ~95 % of all sequence reads. Looking across all 52 rounds, there are 18 highly

represented clusters that differ with regard to the sequence of the P8 stem, which include the wild-type, 24-3, 38-6, and 52-2 polymerases. These 18 variants are shown in Figure 5, together with their frequency of occurrence over time.

During the early rounds of evolution, the wild-type sequence continued to dominate, but became extinct after round 14. Other clusters that appeared early and lacked complementarity in the region of the P8 stem also became extinct after round 14. Most of the clusters that arose subsequently over the course of evolution had five Watson-Crick pairs in the region of the P8 stem. The sequences of the 24-3, 38-6, and 52-2 polymerases first became apparent at rounds 11, 31, and 46, respectively, and were the dominant cluster at the time they were first isolated from the population, following rounds 24, 38, and 52, respectively.

Among the distinct forms of the P8 stem that became abundant over the course of evolution, most contained five base pairs, but some contained either four or six base pairs. All of the variability among these sequences occurs at the distal end of the P8 stem, whereas the proximal nucleotides G11, C12, A13, and C14 (and their pairing partners) are universally conserved. This is not surprising because nucleotides 11–14 are part of the primer binding site that remains fixed during each round of selective amplification. Nucleotides 15 and 16 are free to vary, and do so, so long as complementarity is maintained with the corresponding nucleotides of the opposing strand.

The 5’ and 3’ strands of the P8 stem were also considered individually to determine whether the frequency of occurrence of particular variants of the two strands is correlated over time. Because sequence variation within the 5’ strand is limited to two nucleotides, there was insufficient variation to track all 18 highly represented clusters, but this could

be done for variants corresponding to the wild-type, 24-3, 38-6, and 52-2 polymerases (Figure 5—figure supplement 1). For each of the major polymerase species, there is a high degree of correlation for the occurrence of paired variants of the 5' and 3' strands, indicating that they rose and fell together.

The sequence of the P8 stem that occurs in the 52-2 polymerase first emerged at round 46, and by round 49 constituted 98 % of the population. During that same interval, the 38-6 form of the polymerase fell to extinction. Although unlikely to be the last chapter in the evolution of the polymerase ribozyme, the 52-2 polymerase strongly consolidates the pseudoknot structure within the catalytic core, providing robust activity for the RNA-catalyzed synthesis of complex RNAs.

Discussion

The class I ligase ribozyme was evolved from a starting pool of random-sequence RNAs nearly 30 years ago (Bartel and Szostak, 1993), and has been subjected to more rounds of directed evolution, as either a ligase or the core component of a polymerase, than any other ribozyme. It has proven to be a remarkably stable motif, perhaps due to the high catalytic efficiency of the parental ligase, which is comparable to that of RNA ligase proteins and approaches the physical limit of substrate recognition through Watson-Crick base pairing (Bergman et al., 2000). As a polymerase, however, the ribozyme has considerably lower activity due to its poor affinity for the primer-template complex (Lawrence and Bartel, 2003). This limitation has been partially circumvented by adding a processivity tag that enables the ribozyme to bind tightly to the template though Watson-Crick pairing (Wochner et al., 2011). More recently, the polymerase was evolved to

recognize a ‘promoter’ sequence on the template through a clamp-like mechanism, which enables it to operate in a more processive manner (Cojocaru and Unrau, 2021). Extensive rounds of directed evolution have also been used to increase the catalytic efficiency and sequence generality of the polymerase, to the level that it is now capable of synthesizing the class I ligase and other complex functional RNAs (Tjhung et al., 2020).

It has been suggested that the ribozyme occupies a high and isolated fitness peak, whereby its structural elements are so tightly interwoven that any exploration of alternative structures would have severe negative consequences for fitness (Ellington, 2008). Similar isolation in sequence space has been observed for smaller artificial ribozymes (Pitt and Ferré-D’Amaré, 2010; Blanco et al., 2019), although it could be argued that the more complex structure of the ligase affords more degenerate tertiary interactions that support greater evolvability compared to simpler structures (Edelman and Gally, 2001).

The new pseudoknot structure emerged spontaneously during the directed evolution process and clearly contributes to the improved fitness of the polymerase ribozyme, demonstrating that the ligase core can indeed access alternative structures in response to stringent selection pressure. Notably, the only major structural changes that this motif had undergone previously were also the result of selection pressures that aimed to improve polymerase activity, although those new structures arose from regions of random-sequence nucleotides that were appended to the ends of the motif (Johnston et al., 2001; Wochner et al., 2011; Cojocaru and Unrau, 2021).

Pseudoknots are compact and informationally economical structures that need not alter the global architecture of an RNA (Gutell et al., 1994). Thus a pseudoknot was able

to evolve within the catalytic core of the polymerase ribozyme without significantly perturbing other structural features. The more processive polymerase variant that was evolved by Cojocaru and Unrau, 2021, does not contain, nor could it accommodate, the pseudoknot structure, demonstrating that there are alternative solutions to achieve improved catalytic activity.

Both structural probing and sequence analysis of the evolving population over the course of 52 rounds revealed that the new core structure did not appear suddenly as a 'hopeful monster' (Goldschmidt, 1940; Gould, 1977), but rather was the result of gradual remodeling of the core through a succession of variants along multiple mutational pathways. Prior studies have shown that the nucleotides that gave rise to the P8 stem are relatively tolerant of mutation (Petrie and Joyce, 2014). However, the structural stasis of the ligase fold was broken only when selection required a very challenging enzymatic activity, involving the accurate copying of 10–30 nucleotides from structured RNA templates. Presumably, this phenotype could not have been achieved by more subtle modification of the prior structure. Comparing the core sequence of the 52-2 polymerase to that of the wild type, there are only four substitution and two deletion mutations outside the region of the P7 and P8 stems (Figure 2), none of which would alter the secondary and presumed tertiary structure of the ribozyme.

Due to the requirement to provide a primer binding site for selective amplification, the 14 nucleotides at the 5' end of the polymerase were immutable throughout the evolution process. The last four of these nucleotides became part of the new P8 stem, which necessitated a C-to-U mutation at position 87 and a C insertion at position 89 to achieve complementary pairing with the fixed nucleotides. It is tempting to wonder how

the evolutionary solution might have been different, perhaps better, if this constraint had not been in place. Clearly, there is sequence flexibility within the region that connects the 3' end of the processivity tag to the 5' end of the P8 stem, suggesting that the nucleotides both upstream and within the 5' half of the P8 stem should be allowed to vary in future rounds of evolution. The new topology of the catalytic core also suggests locations where the insertion of random-sequence nucleotides would be tolerated and may provide an opportunity for further evolutionary improvement.

The new pseudoknot structure results in more than a change of primary and secondary structure, also having remodeled the tertiary structure of the catalytic core. One strand of the P8 stem derives from the former J1/3 region and the other from the distal portion of the P7 stem-loop. Based on the X-ray crystal structure of the class I ligase, the P7 stem-loop had previously been oriented away from the active site (Shechner et al., 2009), but the new topology draws those nucleotides back toward the active site. The crystal structure also shows that the J1/3 region lies in direct contact with the minor groove of the primer-template duplex, with three adenosine residues of J1/3 forming A-minor interactions with nucleotides located 3 and 4 positions upstream of the ligation junction. One of those adenosines has been deleted in the 52-2 polymerase. Other residues of J1/3, which are retained in the 52-2 polymerase, coordinate a Mg2+ ion that helps to catalyze the phosphoester transfer reaction (Shechner and Bartel, 2011).

The comparative in-line probing studies show that there has been subtle alteration of the polymerase structure in regions beyond the P7 and P8 stems (Figure 4—figure supplement 1), most notably in the P3 and P6 stems that form a coaxial stack with P7 (Shechner et al., 2009). There also were changes in peripheral regions of the ribozyme,

perhaps secondary adaptations to the changes that occurred within the catalytic core. The burst-phase kinetics of the 52-2 polymerase suggests that it can adopt multiple folded states, some that are highly active and might be stabilized by further evolution through a combination of core and peripheral mutations.

Further improvement of the polymerase, especially with regard to template processivity and copying fidelity, will be required to develop a general RNA replicase. Fidelity is a major obstacle if the aim is to synthesize functional products as long as the polymerase itself. A previous study demonstrated that there is a trade-off between product length and fidelity, especially when copying challenging templates (Tjhung et al., 2020). The more time that is required to complete the synthesis, the more opportunity there is to extend a mismatched terminus and thereby incorporate a mutation among the full-length materials. In synthesizing the hammerhead ribozyme in the presence of 200 mM Mg2+, the 24-3, 38-6, and 52-2 polymerases all have a fidelity of ~92 % per nucleotide position. The 52-2 polymerase is able to operate in the presence of 50 mM Mg2+, and under that condition the fidelity of hammerhead synthesis improves to 94.4 %. But when synthesizing the class I ligase, which requires 200 mM Mg2+ to achieve good yield, the fidelity of the 52-2 polymerase is only 84.1 %. The evolutionary path to substantially improved polymerase fidelity likely will entail both improved catalytic activity and an ability to operate under conditions that are less conducive to base mismatch.

With the heritage of 52 successive generations, it has been illuminating to follow the trajectory of evolution as the population sifted through an astronomical number of possibilities to find those that confer selective advantage. Directed evolution is a highly

reductionistic process compared to biological evolution, but has few unseen variables and can provide a detailed picture of how novel sequence begets novel structure and corresponding novel function. The class I ligase motif is old by the standard of ribozymes evolved in the laboratory, but vastly younger than ribozymes found in nature, and thus has not been shaped by long-term selection for evolvability. Nonetheless, through sustained selection for novel function, structural novelty emerged as the population escaped the prior fitness peak and entered a new and more promising fitness regime.

Materials and methods

Materials

All oligonucleotides used in this study are listed in Supplementary file 4. Synthetic oligonucleotides were either purchased from IDT (Coralville, IA) or prepared by solid-phase synthesis using an Expedite 8909 DNA/RNA synthesizer, with reagents and phosphoramidites from either Chemgenes (Wilmington, MA) or Glen Research (Sterling, VA). RNA templates were prepared by *in vitro* transcription of synthetic DNA. Polymerase ribozymes were prepared by *in vitro* transcription of dsDNA that was generated by either PCR amplification of the corresponding plasmid DNA or by PCR assembly of synthetic oligodeoxynucleotides. All RNA primers, templates, and ribozymes were purified by denaturing polyacrylamide gel electrophoresis (PAGE) and ethanol precipitation prior to use. His-tagged T7 polymerase was prepared from Escherichia coli strain BL21 containing plasmid pBH161 (kindly provided by W McAllister, SUNY Downstate Medical Center, Brooklyn, NY). Hot Start OneTaq was obtained from New England BioLabs (Ipswich, MA), rAPid alkaline phosphatase was from SigmaAldrich (St. Louis, MO), T4 polynucleotide kinase was from New England Biolabs (Ipswich, MA), and QIAprep Spin Miniprep Kit was from Qiagen (Germantown, MD). MyOne C1 streptavidin magnetic beads, PureLink PCR cleanup kit, TOPO TA cloning kit, SuperScript IV reverse transcriptase, Turbo DNase, and RNase T1 all were from Thermo Fisher Scientific (Waltham, MA). NTPs were from ChemImpex International (Wood Dale, IL) and all other chemical reagents were from Sigma-Aldrich. The pH of Tris-HCl was adjusted at 23 °C.

Assembly PCR

Polymerase ribozymes containing specific mutations were prepared by PCR assembly of synthetic oligodeoxynucleotides (Supplementary file 4), followed by *in vitro* transcription. The polymeraseencoding DNA was provided as six fragments, each overlapping by 20–22 base pairs. The fragments were assembled and amplified by PCR, using 0.5 μM each of the two outermost fragments and 0.005 μM each of the four internal fragments, and 0.025 U/μL OneTaq Hot Start polymerase, carried out for 25 thermal cycles. The PCR products were used directly in the *in vitro* transcription reaction.

***In vitro* transcription**

RNA templates and ribozymes were prepared by *in vitro* transcription in a mixture containing 5–20 ng/ μL template DNA, 5 mM each NTP, 15 U/μL T7 RNA polymerase, 0.002 U/μL inorganic pyrophosphatase, 25 mM MgCl2, 2 mM spermidine, 10 mM DTT, and 40 mM Tris-HCl (pH 8.0), which was incubated at 37 °C for 2 hr. The template DNA was then digested by adding 0.1 U/μL Turbo DNase and incubating at 37 °C for 1 hr.

***In vitro* evolution**

A starting pool of DNA templates was prepared by solid-phase synthesis, based on the sequence of the 38-6 polymerase and introducing random mutations at a frequency of 10 % per nucleotide for all positions between the two primer binding sites (nucleotides 15–167). The DNA was made doublestranded by primer extension using SuperScript IV reverse transcriptase, including 1.5 mM MnCl2 in the reaction mixture to promote extension through lesions that arose during DNA synthesis (Chaput et al., 2003). The dsDNA was amplified linearly by eight cycles of PCR using only the upstream primer, which introduced the T7 RNA polymerase promoter sequence. The DNA products were

purified using the PureLink PCR cleanup kit, then 650 pmol dsDNA was used to prepare 2 nmol RNA to initiate the first round of evolution (round 39; see Supplementary file 1). The starting population consisted of an average of three copies each of 4 × 1014 different RNAs. In all subsequent rounds, the size of the RNA population was 200 pmol.

In vitro evolution was carried out as described previously (Tjhung et al., 2020). The polymerase ribozymes were tethered at their 5' end to an RNA primer that was annealed to a complementary RNA template and extended by the ribozyme using the four NTPs. The resulting materials were subjected to the selection protocols described below, then reverse-transcribed, PCR-amplified, and forward transcribed to yield progeny RNAs to begin the next round of evolution. Error-prone PCR (Cadwell and Joyce, 1992) was performed after rounds 43–50.

During rounds 39 and 40, the RNA template was biotinylated and all template-bound materials were captured on streptavidin-coated magnetic beads, which were washed twice with a solution of 8 M urea, 1 mM EDTA, 10 mM Tris-HCl (pH 8.0), and 0.05 % Tween-20. The extended products were then eluted from the template with a solution containing 25 mM NaOH, 1 mM EDTA, and 0.05 % Tween-20, neutralized with HCl, and precipitated with ethanol. The wash and elution conditions were optimized to exclude polymerases that failed to extend the attached primer, while retaining those that had extended the primer to yield full-length products.

In all subsequent rounds, selection was based on the ability of the polymerase to synthesize a functional hammerhead ribozyme. In those rounds, the 5' end of the polymerase was tethered to the 5' end of an RNA primer via a synthetic linker that contained both a biotin moiety and a substrate for the hammerhead ribozyme. The primer

was then annealed through Watson-Crick pairing to a separate template encoding the sequence of the hammerhead. Following extension of the primer by the polymerase to generate the hammerhead ribozyme, the full-length products were purified by PAGE, then bound to streptavidin beads in the presence of 1 mM EDTA, 300 mM NaCl, 10 mM Tris-HCl (pH 8.0), and 0.05 % Tween-20, which prevented premature cleavage of the substrate by the hammerhead. The beads were washed with this same solution, then incubated in the presence of 20 mM MgCl2 at 23 °C for 30 min. Under the latter conditions, active hammerhead ribozymes cleaved the attached RNA substrate, thereby releasing the corresponding polymerase from the beads.

Over the course of evolution, both the time allotted for RNA polymerization and the concentration of MgCl2 were reduced (Supplementary file 1). Following round 52, the PCR-amplified DNA was cloned into E. coli using the TOPO-TA cloning kit, and the cells were grown at 37 °C for 16 hr on LB agar plates containing 50 μg/mL kanamycin. Individual colonies were picked and grown in 3 mL of LB medium with 50 μg/mL kanamycin at 37 °C for 16 hr. Plasmid DNA was harvested using the QIAprep Spin Miniprep Kit and sequenced by Eton Bioscience (San Diego, CA).

RNA-catalyzed polymerization of RNA

RNA polymerization reactions used 100 nM ribozyme, 80 nM fluoresceinand biotin-labeled RNA primer, and 100 nM RNA template, which were annealed by heating at 80 °C for 30 s and then cooling to 17 °C. The annealed RNAs were added to a reaction mixture containing 4 mM each NTP, either 50 or 200 mM MgCl2, 25 mM Tris-HCl (pH 8.3), and 0.05 % Tween-20, which was incubated at 17 °C. The reaction was quenched by manually adding an equal volume of a solution containing 250 mM EDTA (pH 8.0), 500

mM NaCl, 5 mM Tris-HCl (pH 8.0), and 0.025 % Tween-20, then mixed with 5 μg streptavidin magnetic beads per pmol biotinylated RNA primer, and incubated with gentle agitation at room temperature for 30 min. Prior to use, the beads had been washed according to the manufacturer's protocol, then incubated with 1 mg/mL tRNA in a solution containing 1 M NaCl, 1 mM EDTA, and 10 mM Tris-HCl (pH 8.0) for 30 min. The RNA template was removed from the bead-bound materials by two washes with a solution containing 25 mM NaOH, 1 mM EDTA, and 0.05 % Tween-20, followed by two washes with a solution containing 8 M urea, 1 mM EDTA, and 10 mM Tris-HCl (pH 8.0). Then the reaction products were eluted from the beads by incubating in 95 % formamide and 10 mM EDTA at 95 °C for 10 min, and were analyzed by PAGE. For fast-reaction kinetics, all reaction components other than the NTPs were pre-incubated at 17 °C for 5 min, then the NTPs were added and the solution was rapidly mixed to initiate the reaction.

RNA-catalyzed ligation of RNA

The 52-2 polymerase was used to synthesize the class I ligase ribozyme by extending a 20-nucleotide RNA primer on a complementary RNA template. The resulting full-length products were purified by PAGE and subsequent ethanol precipitation. RNA ligation reactions were performed as described previously (Tjhung et al., 2020), using the same oligonucleotide substrates employed in previous kinetic studies (Bergman et al., 2000). The reaction mixture contained 20 μM 5'-substrate that had been fluorescently labeled with Cy5, 80 μM 3'-substrate that had been chemically triphosphorylated, either no or 1 μM ligase ribozyme, 60 mM MgCl2, 200 mM KCl, 0.6 mM EDTA, and 50 mM Tris-HCl (pH 8.3), which was incubated at 23 °C for 24 hr. The reaction was quenched

by adding four volumes of 95 % formamide and 20 mM EDTA and the products were analyzed by PAGE.

Analysis of polymerase fidelity

The hammerhead and class I ligase ribozymes were synthesized by the 52-2 polymerase under standard reaction conditions. For the hammerhead, all partial and full-length products, obtained in the presence of either 50 or 200 mM MgCl2, were analyzed. For the ligase, only full-length products obtained in the presence of 200 mM MgCl2 were analyzed. The products were converted to dsDNA molecules for Illumina sequencing, as described previously (Tjhung et al., 2020). Sequencing was carried out by the Salk Next-Generation Sequencing Core on an Illumina MiniSeq, with either a 75or 150-cycle paired-end run for the hammerhead or ligase, respectively.

The sequence data were processed to categorize all mutations relative to the expected sequence, as described previously (Tjhung et al., 2020). For both the hammerhead and ligase ribozyme, a custom JavaScript (source code in Tjhung et al., 2020) was used to calculate the number of matches, mismatches, deletions, and insertions as a function of template position and read length along the reference sequence. For the ligase, the distribution of Levenshtein distances from the reference sequence was determined directly from the alignment. The resulting data were manually processed to generate fidelity tables and position-specific data plots for the full-length products. HTS data, scripts, and related files are archived at the Dryad Digital Repository: https://doi.org/105061/dryadc866t1g78.

In-line probing

The 5’ end of the ribozyme was dephosphorylated using rAPid alkaline phosphatase, then [5’-32 P]-labeled with [γ-32P]ATP using T4 polynucleotide kinase, both according to the manufacturer’s protocol. In-line probing (Soukup and Breaker, 1999) of 5’-labeled ribozymes was performed under the same conditions as the polymerization reaction, including the RNA primer, template, and four NTPs in the mixture. After 3, 6, 12, or 24 hr, the reaction was quenched with EDTA and the products were analyzed by PAGE. Individual bands in the gel were quantitated using ImageQuant 8.2. The raw counts were corrected by subtracting background counts, then scaled to the nucleotide position within the region of interest that had the highest level of cleavage.

Analysis of the evolving population by deep sequencing

Sequencing of PCR products obtained after various rounds of evolution was performed at the Yale Center for Genome Analysis on an Illumina NovaSeq 6000, which generated ~20 million paired reads for each round that was sampled. The sequence datasets were quality-filtered and trimmed using the paired-end read merger program PEAR (Zhang et al., 2014). The data were then filtered to include only reads of >150 nucleotides with a Phred score >33. Individual sequences were enumerated and converted to the fastq file format using a custom Python script (Portillo et al., 2021). The file sizes were reduced by removing sequences with <10 reads for rounds 16 and 31, <10,000 reads for round 27, and <1000 reads for all other rounds. The fastq file entries were then aligned using MUSCLE (Edgar, 2004). The aligned reads were trimmed to the region encompassing the P7 and P8 stems (nucleotides 9–17 and 83–95) using AliView (Larsson, 2014), then clustered using cd-hit-est (Weizhong and Godzik, 2006), with a

clustering threshold of 100 % identity (-c 1.0), maximum unmatched length of two nucleotides (-U 2), and length difference cutoff of two nucleotides (-S 2). Clusters with >1% representation in any given round were identified. The insertion/deletion of A residues between nucleotides 17 and 18, and the presence of single mutations outside positions 11–16 and 84–89, were treated as representing the same cluster. HTS data, scripts, and related files are archived at the Dryad Digital Repository: *https://doi.org/105061/dryadc866t1g78*.

Acknowledgements

The authors thank Adam Roth and Noah Setterholm for helpful discussions, and James Knight of the Yale Center for Genome Analysis for consultation on bioinformatics. X.P. was supported by the National Science Foundation Graduate Research Fellowship Program (DGE1752134). R.R.B. and X.P. were supported by an NIH grant (P01GM022778). R.R.B. is also supported by the Howard Hughes Medical Institute. This work was also supported by grants to G.F.J. from the National Aeronautics and Space Administration (NSSC19K0481) and a Simons Foundation grant (287624).

References

Achuthan V, Keith BJ, Connolly BA, DeStefano JJ. 2014. Human immunodeficiency virus reverse transcriptase displays dramatically higher fidelity under physiological magnesium conditions *in vitro*. *Journal of Virology* 88: 8514–8527. DOI: https://doi.org/10.1128/JVI.00752-14, PMID: 24850729

Attwater J, Raguram A, Morgunov AS, Gianni E, Holliger P. 2018. Ribozyme-catalysed RNA synthesis using triplet building blocks. *eLife* 7: e35255. DOI: https://doi.org/10.7554/eLife.35255, PMID: 29759114

Bartel DP, Szostak JW. 1993. Isolation of new ribozymes from a large pool of random sequences. *Science* 261:1411–1418. DOI: https://doi.org/10.1126/science.7690155, PMID: 7690155

Bergman NH, Johnston WK, Bartel DP. 2000. Kinetic framework for ligation by an efficient RNA ligase ribozyme. *Biochemistry* 39: 3115–3123. DOI: https://doi.org/10.1021/bi992654u, PMID: 10715133

Blanco C, Janzen E, Pressman A, Saha R, Chen IA. 2019. Molecular fitness landscapes from high-coverage sequence profiling. *Annual Review of Biophysics* 48: 1–18. DOI: https://doi.org/10.1146/annurev-biophys-052118-115333, PMID: 30601678

Cadwell RC, Joyce GF. 1992. Randomization of genes by PCR mutagenesis. *PCR Methods* Applic 2: 28–33. DOI:https://doi.org/10.1101/gr.2.1.28

Chaput JC, Ichida JK, Szostak JW. 2003. DNA polymerase-mediated DNA synthesis on a TNA template. *Journal of the American Chemical Society* 125: 856–857. DOI: https://doi.org/10.1021/ja028589k, PMID: 12537469

Cojocaru R, Unrau PJ. 2021. Processive RNA polymerization and promoter recognition in an RNA world. *Science* 371: 1225–1232. DOI: https://doi.org/10.1126/science.abd9191, PMID: 33737482

Crick FHC. 1968. The origin of the genetic code. *Journal of Molecular Biology* 38: 367–379. DOI: https://doi. org/10.1016/0022-2836(68)90392-6, PMID: 4887876

Eckert KA, Kunkel TA. 1990. High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Research* 18: 3739–3744. DOI: https://doi.org/10.1093/nar/18.13.3739, PMID: 2374708

Edelman GM, Gally JA. 2001. Degeneracy and complexity in biological systems. *PNAS* 98: 13763–13768. DOI: https://doi.org/10.1073/pnas.231499798, PMID: 11698650

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797. DOI: https://doi.org/10.1093/nar/gkh340, PMID: 15034147

Ekland EH, Bartel DP. 1995. The secondary structure and sequence optimization of an RNA ligase ribozyme. *Nucleic Acids Research* 23: 3231–3238. DOI: https://doi.org/10.1093/nar/23.16.3231, PMID: 7667099

Ekland EH, Szostak JW, Bartel DP. 1995. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* 269: 364–370. DOI: https://doi.org/10.1126/science.7618102, PMID: 7618102

Ellington AD. 2008. Man versus machine versus ribozyme. *PLOS Biology* 6: e132. DOI: https://doi.org/10.1371/ journal.pbio.0060132, PMID: 18507505

Gilbert W. 1986. The RNA world. *Nature* 319: 618.

Goldschmidt R. 1940. *The Material Basis of Evolution.* Yale University Press.

Gould SJ. 1977. The return of hopeful monsters. *Natural History* 86: 22–30.

Gutell RR, Larsen N, Woese CR. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews* 58: 10–26. DOI: https://doi.org/10.1128/mr.58.1.10-26.1994, PMID: 8177168

Horning D, Joyce GF. 2016. Amplification of RNA by an RNA polymerase ribozyme. *PNAS* 113: 9786–9791. DOI: https://doi.org/10.1073/pnas.1610103113, PMID: 27528667

Huang Z, Szostak JW. 2003. Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer. *RNA* 9: 1456–1463. DOI: https://doi.org/10.1261/rna.5990203, PMID: 14624002

Ikawa Y, Tsuda K, Matsumura S, Inoue T. 2004. De novo synthesis and development of an RNA enzyme. *PNAS* 101: 13750–13755. DOI: https://doi.org/10.1073/pnas.0405886101, PMID: 15365187

Jaeger L, Wright MC, Joyce GF. 1999. A complex ligase ribozyme evolved *in vitro* from a group I ribozyme domain. *PNAS* 96: 14712–14717. DOI: https://doi.org/10.1073/pnas.96.26.14712, PMID: 10611278

Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* 292: 1319–1325. DOI: https://doi.org/10.1126/science.1060786, PMID: 11358999

Joyce GF. 2002. The antiquity of RNA-based evolution. *Nature* 418: 214–221. DOI: https://doi.org/10.1038/418214a, PMID: 12110897

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30: 3276–3278. DOI: https://doi.org/10.1093/bioinformatics/btu531, PMID: 25095880

Lawrence MS, Bartel DP. 2003. Processivity of ribozyme-catalyzed RNA polymerization. *Biochemistry* 42:8748–8755. DOI: https://doi.org/10.1021/bi034228l, PMID: 12873135

Lorsch JR, Szostak JW. 1994. *In vitro* evolution of new ribozymes with polynucleotide kinase activity. *Nature* 371:31–36. DOI: https://doi.org/10.1038/371031a0, PMID: 7521014

Petrie KL, Joyce GF. 2014. Limits of neutral drift: lessons from the *in vitro* evolution of two ribozymes. *Journal of Molecular Evolution* 79: 75–90. DOI: https://doi.org/10.1007/s00239-014-9642-z, PMID: 25155818

Pfingsten JS, Costantino DA, Kieft JS. 2007. Conservation and diversity among the three-dimensional folds of the Dicistroviridae intergenic region IRESes. *Journal of Molecular Biology* 370: 856–869. DOI: https://doi.org/10.1016/j.jmb.2007.04.076, PMID: 17544444

Pitt JN, Ferré-D'Amaré AR. 2010. Rapid construction of empirical RNA fitness landscapes. *Science* 330:376–379. DOI: https://doi.org/10.1126/science.1192001, PMID: 20947767

Portillo X, Huang YT, Breaker RR, David P, Joyce GF. 2021. Deep Sequencing Datasets from: Witnessing the Structural Evolution of an RNA Enzyme. Dryad Repository. DOI: https://doi.org/10.5061/dryad.c866t1g78

Regulski EE, Breaker RR. 2008. In-line probing analysis of riboswitches. *Methods in Molecular Biology* 419:53–67. DOI: https://doi.org/10.1007/978-1-59745-033-1_4, PMID: 18369975

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* 289: 448–452. DOI: https://doi.org/10.1126/science.289.5478.448, PMID: 10903205

Shechner DM, Grant RA, Bagby SC, Koldobskaya Y, Piccirilli JA, Bartel DP. 2009. Crystal structure of the catalytic core of an RNA-polymerase ribozyme. *Science* 326: 1271–1275. DOI: https://doi.org/10.1126/science.1174676,PMID: 19965478

Shechner DM, Bartel DP. 2011. The structural basis of RNA-catalyzed RNA polymerization. *Nature Structural & Molecular Biology* 18: 1036–1042. DOI: https://doi.org/10.1038/nsmb.2107, PMID: 21857665

Soukup GA, Breaker RR. 1999. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* 5: 1308–1325. DOI: https://doi.org/10.1017/s1355838299990891, PMID: 10573122

Tjhung KF, Shokhirev MN, Horning DP, Joyce GF. 2020. An RNA polymerase ribozyme that synthesizes its own ancestor. *PNAS* 117: 2906–2913. DOI: https://doi.org/10.1073/pnas.1914282117, PMID: 31988127

Weizhong L, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659. DOI: https://doi.org/10.1093/bioinformatics/btl158, PMID: 16731699

Williams KP, Bartel DP. 1996. Phylogenetic analysis of tmRNA secondary structure. *RNA* 2: 1306–1310 PMID: 8972778.,

Wochner A, Attwater J, Coulson A, Holliger P. 2011. Ribozyme-catalyzed transcription of an active ribozyme. *Science* 332: 209–212. DOI: https://doi.org/10.1126/science.1200752, PMID: 21474753

Zaher HS, Unrau PJ. 2007. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* 13: 1017–1026. DOI: https://doi.org/10.1261/rna.548807, PMID: 17586759

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina paired-end read merger. *Bioinformatics* 30: 614–620. DOI: https://doi.org/10.1093/bioinformatics/btt593, PMID: 24142950

Figures

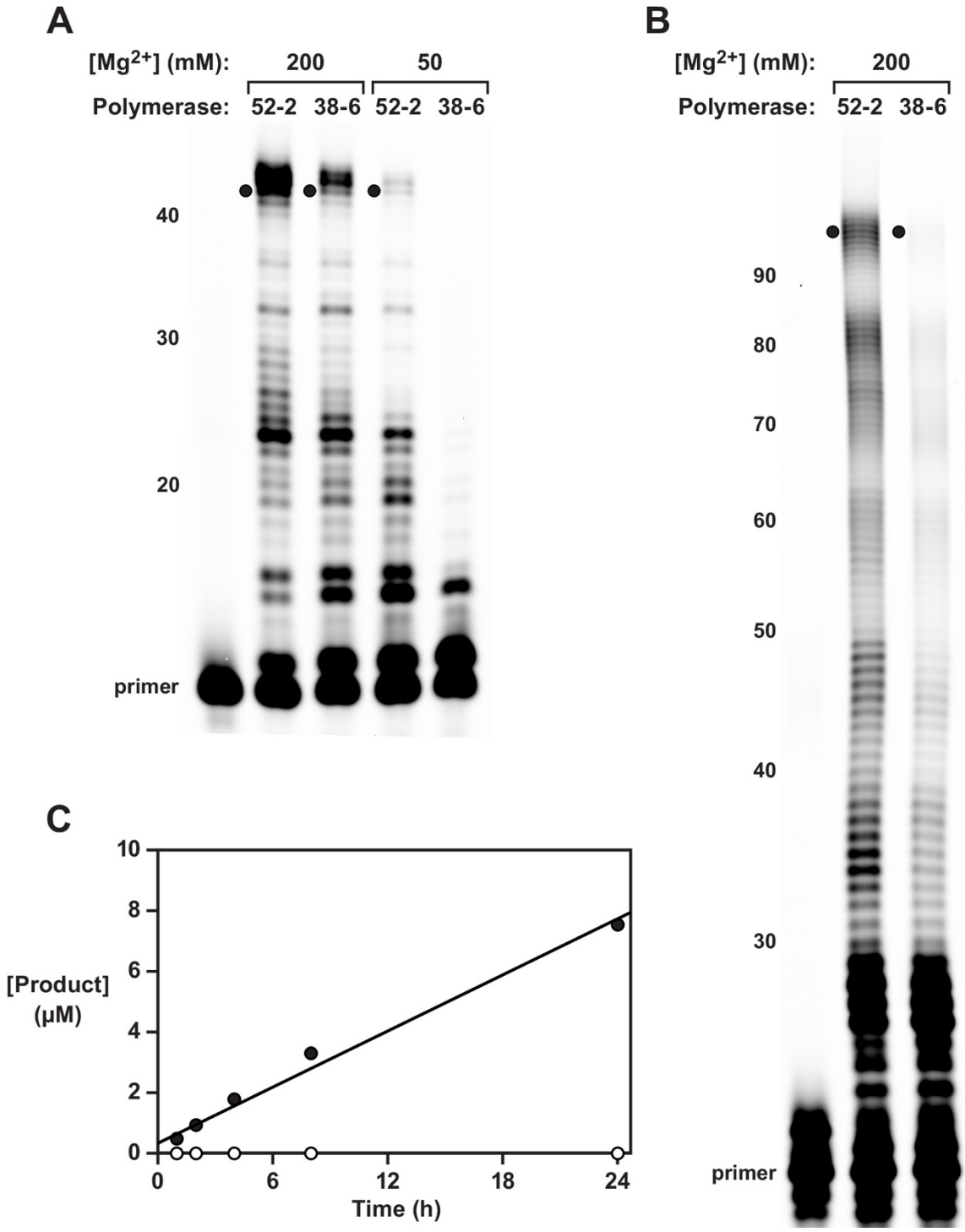

A
[Mg2+] (mM): 200 50
Polymerase: 52-2 38-6 52-2 38-6
40
30
20
primer
B
[Mg2+] (mM): 200
Polymerase: 52-2 38-6
90
80
70
60
50
40
30
primer
C
[Product] (μM)
10
8
6
4
2
0
0
6
12
18
24
Time (h)

Figure 2-1. Synthesis of functional RNA molecules by the 38-6 and 52-2 polymerases. (**A**) Synthesis of the hammerhead ribozyme in the presence of either 50 or 200 mM Mg^{2+} after 1 hr. (**B**) Synthesis of the class I ligase ribozyme in the presence of 200 mM Mg^{2+} after 24 hr. Reaction conditions for (**A**) and (**B**): 100 nM polymerase, 80 nM primer, 100 nM template, 4 mM each NTP, and either 50 or 200 mM $MgCl_2$ at pH 8.3 and 17 °C. Intermediate-length products are numbered at the left. Black dots indicate full-length products. (**C**) Time course of RNA ligation catalyzed by the class I ligase ribozyme that had been synthesized by the 52-2 polymerase, as shown in (**B**), comparing the RNA-catalyzed (black circles) and the uncatalyzed (white circles) reactions, which have a rate of 0.31 and 0.00021 hr^{-1}, respectively. Reaction conditions: ±1 μM ligase ribozyme, 20 μM 5'-substrate, 80 μM 3'-substrate, 60 mM $MgCl_2$, 200 mM KCl, and 0.6 mM EDTA at pH 8.3 and 23 °C. David Horning contributed to the analysis, methodology, and design used to generate these data.

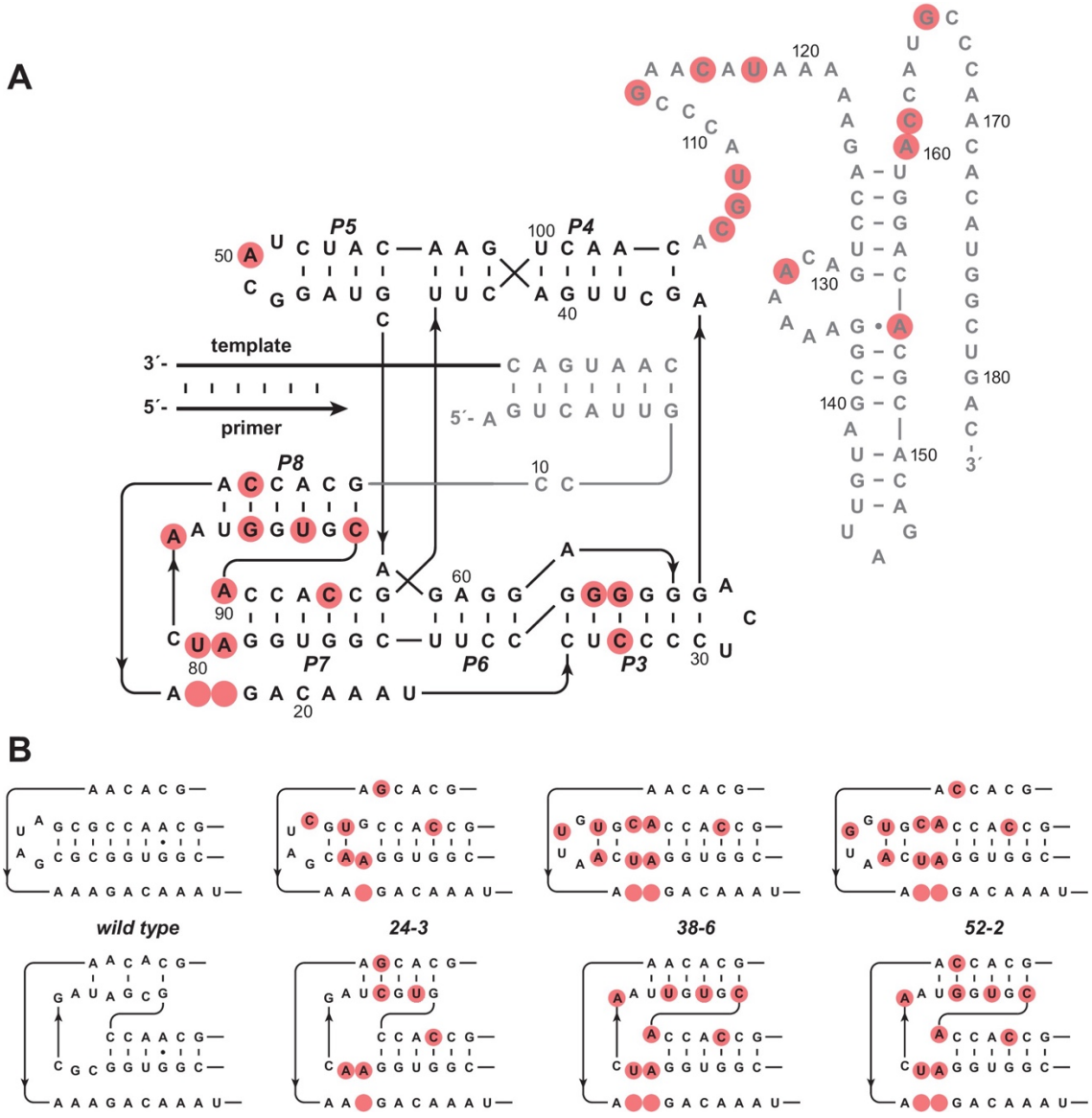

A
P5
P4
P8
P7
P6
P3
template
primer
3´-
5´-
5´- A
B
wild type
24-3
38-6
52-2

Figure 2-2. Evolution of the novel pseudoknot structure.

(**A**) Sequence and secondary structure of the 52-2 polymerase. Red circles indicate mutations relative to the wild-type polymerase. Regions outside the ligase core are shown in gray. Paired regions and nucleotides are numbered according to the 52-2 polymerase, with corresponding nucleotides numbered similarly for all polymerase variants. (**B**) Progressive mutation of the region encompassing the P7 and P8 stems, mapped onto both the old structure (*top*) and the new pseudoknot structure (*bottom*). Blank circles indicate deletions.

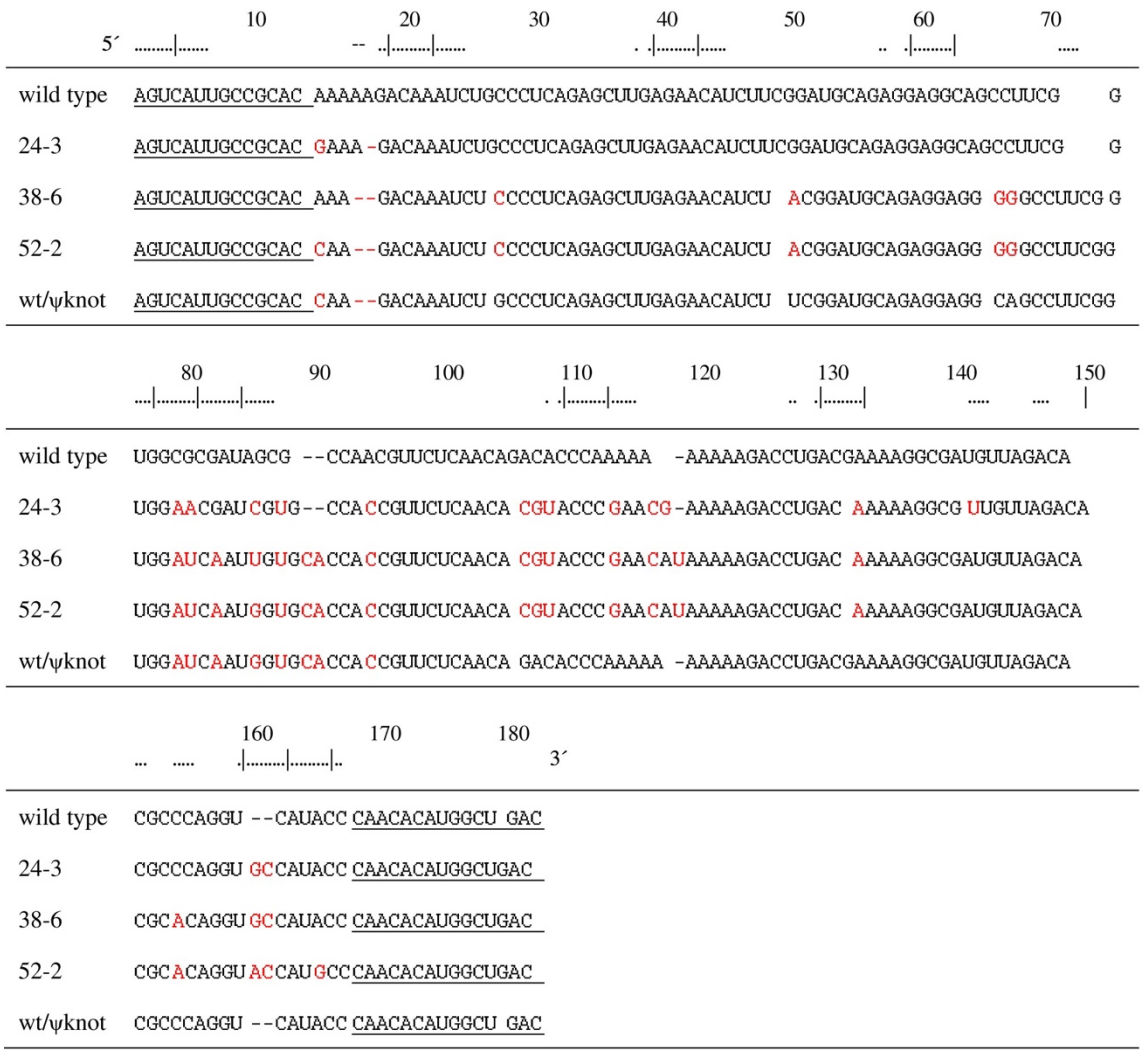


Figure 2-2-supplement 1. Aligned sequences of the named polymerase ribozymes.

Nucleotide positions are numbered according to the 52-2 polymerase (*Figure 2A*). Fixed primer binding sites are underlined. Mutations relative to the wild type are shown in red. wt/ψknot is a chimeric molecule that contains the pseudoknot structure of the 52-2 polymerase transplanted onto the wild-type background.

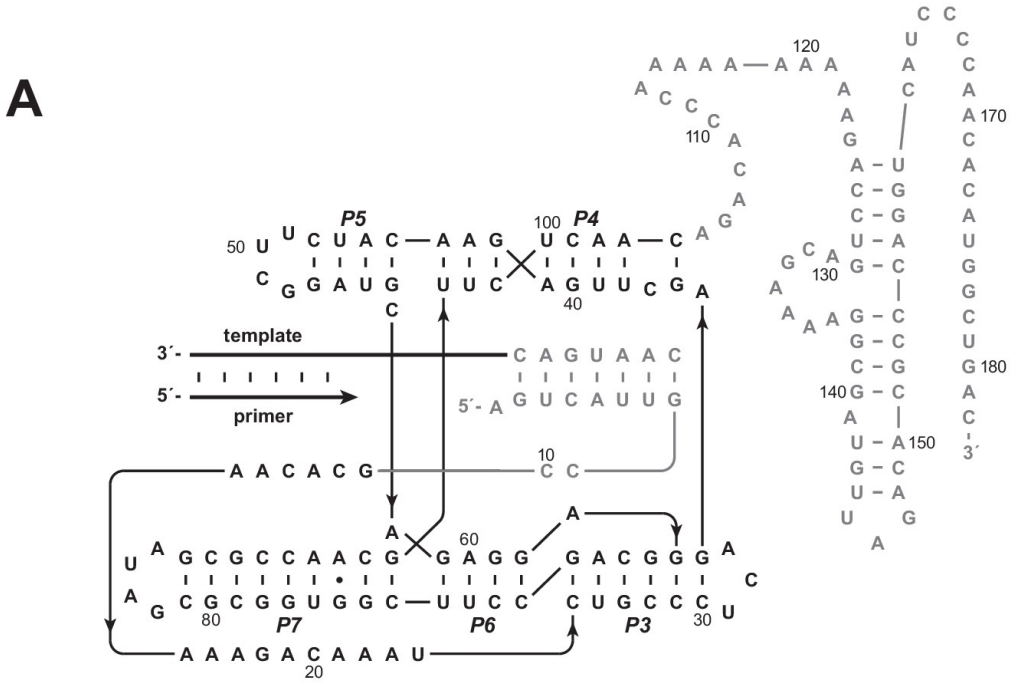

A
P5
P4
P7
P6
P3
template
primer
3′-
5′-
5′- A
B
C
52-2
wild type
chimera
Time (min):
– 0 0.25 2 15 120 0 0.25 2 15 120 0 0.25 2 15 120
primer

Figure 2-2-supplement 2. Installing the pseudoknot structure on the wild-type background.

(**A**) Sequence and secondary structure of the wild-type polymerase. Paired regions and nucleotides are numbered according to the 52-2 polymerase (see *Figure 2A*). (**B**) Region encompassing the P7 and P8 stems for both the wild-type (*left*) and 52-2 (*right*) polymerase. (**C**) Comparative activity of the 52-2, wild-type, and chimeric polymerases, the latter containing the pseudoknot structure (P7 and P8 region) from the 52-2 polymerase, but otherwise having the sequence of the wild type. Black dots indicate full-length products. Reaction conditions: 100 nM polymerase, 100 nM template encoding the sequence 5'-UGCGAAGCGUG-3', 80 nM primer, 4 mM each NTP, and 200 mM $MgCl_2$ at pH 8.3 and 17°C.

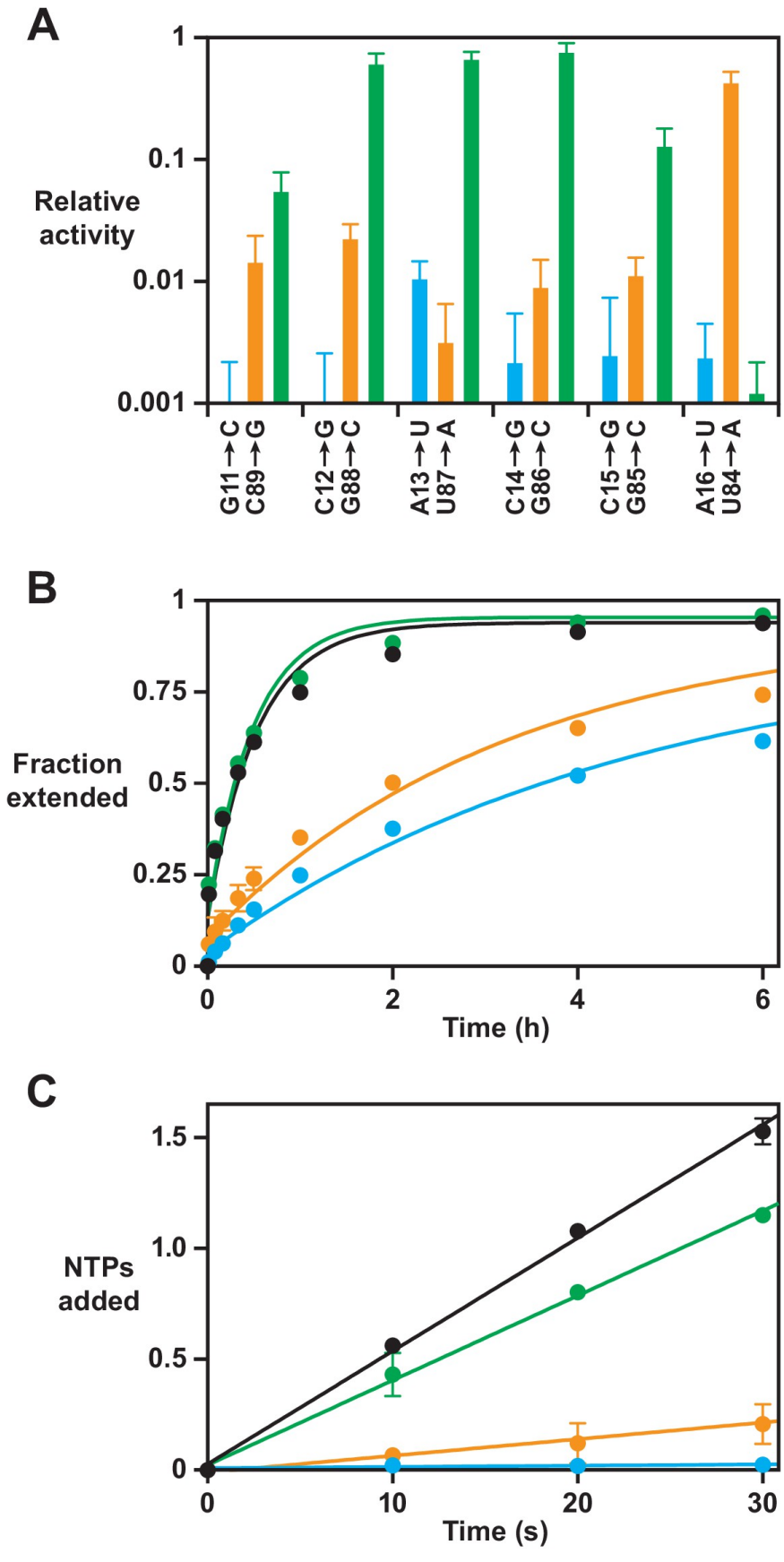
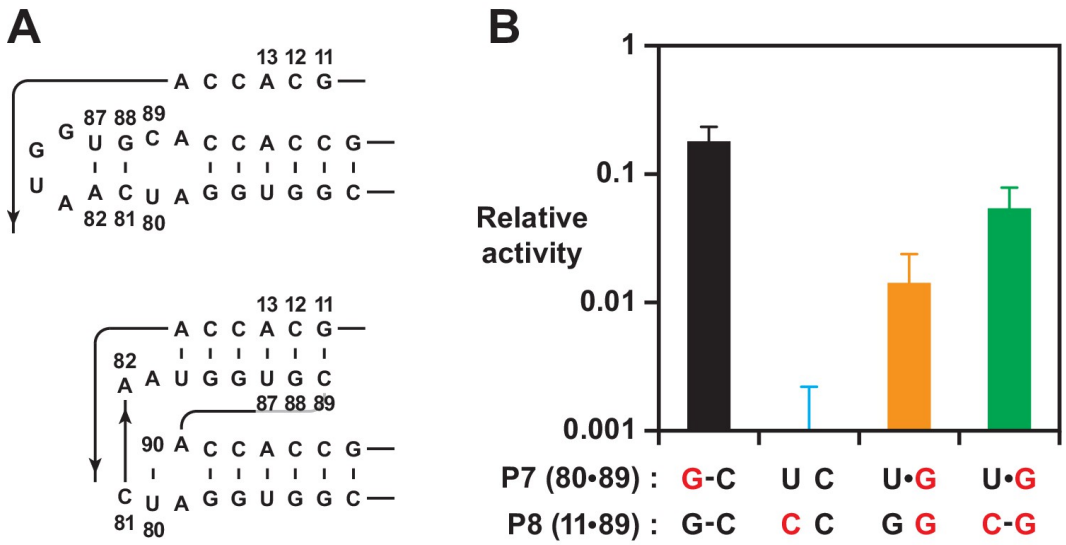
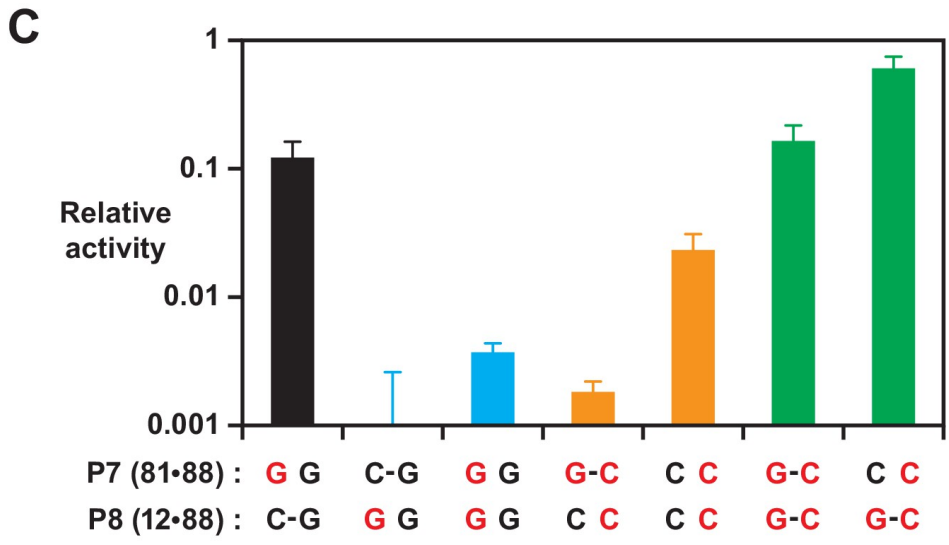

A
Relative activity
1
0.1
0.01
0.001
G11→C C89→G
C12→G G88→C
A13→U U87→A
C14→G G86→C
C15→G G85→C
A16→U U84→A
B
Fraction extended
1
0.75
0.5
0.25
0
0
2
4
6
Time (h)
C
NTPs added
1.5
1.0
0.5
0
0
10
20
30
Time (s)

Figure 2-3. Effect on polymerase activity of disruptive and compensatory mutations within the P8 stem. (**A**) Yield of full-length RNA relative to that of the 52-2 polymerase in a 30 min reaction requiring the addition of 28 nucleoside 5'-triphosphates (NTPs) (note the logarithmic scale). At each position within the P8 stem, a transversion mutation was made in either the 5' strand (blue) or the 3' strand (gold), or the two mutations were combined to restore complementarity (green). Values are the average of at least three replicates with standard deviation. (**B**) Time course of primer extension by at least one nucleotide on an 11-nucleotide template, comparing the 52-2 polymerase (black), C12G mutant (blue), G88C mutant (gold), and C12G/G88C double mutant (green). The data were fit to a single exponential rise to maximum, allowing for an initial burst phase. (**C**) Average number of nucleotides added during the first 30 s of the reactions depicted in (**B**). The data were fit to a linear equation. For both (**B**) and (**C**), values are the average of two replicates with standard deviation. Reaction conditions: 100 nM polymerase, 100 nM template, 80 nM primer, 4 mM each NTP, and 200 mM MgCl2 at pH 8.3 and 17 °C. Yu-Ting Huang contributed to the methodology and design used to generate these data.


A
13 12 11
A C C A C G
87 88 89
G U G C A C C A C C G
G
U
A A C U A G G U G G C
82 81 80
13 12 11
A C C A C G
82
A
A U G G U G C
87 88 89
90 A C C A C C G
C U A G G U G G C
81 80
B
1
0.1
0.01
0.001
Relative activity
P7 (80•89) : G-C U C U•G U•G
P8 (11•89) : G-C C C G G C-G


C
1
0.1
0.01
0.001
Relative activity
P7 (81•88) : G G C-G G G G-C C C G-C C C
P8 (12•88) : C-G G G G G C C C C G-C G-C

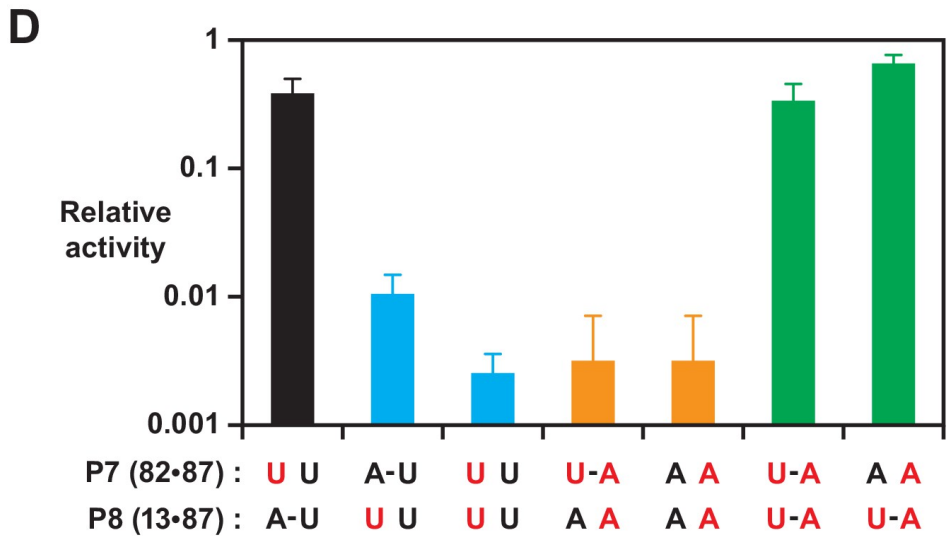

D
1
0.1
0.01
0.001
Relative activity
P7 (82•87) : U U A-U U U U-A A A U-A A A
P8 (13•87) : A-U U U U U A A A A U-A U-A

Figure 2-3-supplement 1. Effect on polymerase activity of mutations within the P7 and P8 stems.

(**A**) Secondary structure of the region encompassing the P7 and P8 stems, mapped onto both the old (*top*) and new (*bottom*) structures. Nucleotide positions that were subject to mutation are numbered. (**B**) Activity relative to the 52-2 polymerase due to mutations (red) that would have different consequences for the old vs. new structure, affecting the U80-C89 pair of P7 and/or G11-C89 pair of P8. Bars indicate activity following mutations that did not alter the P8 pair (black), introduced a transversion mutation in either the 5' (blue) or 3' (gold) strand, or introduced compensatory mutations to restore complementarity (green). (**C**) Activity determined as above for mutations affecting the C81-G88 pair of P7 and/or C12-C88 pair of P8. (**D**) Activity determined as above for mutations affecting the A82-U87 pair of P7 and/or A13-U87 pair of P8. For (**B**), (**C**), and (**D**), relative activity is the yield of full-length RNA compared to that of the 52-2 polymerase in a 30 min reaction requiring the addition of 28 nucleoside 5'-triphosphates (NTPs). Values are the average of at least three replicates with standard deviation. Reaction conditions as in Figure 3. Yu-Ting Huang and David Horning contributed to the analysis, methodology, and design used to generate these data.

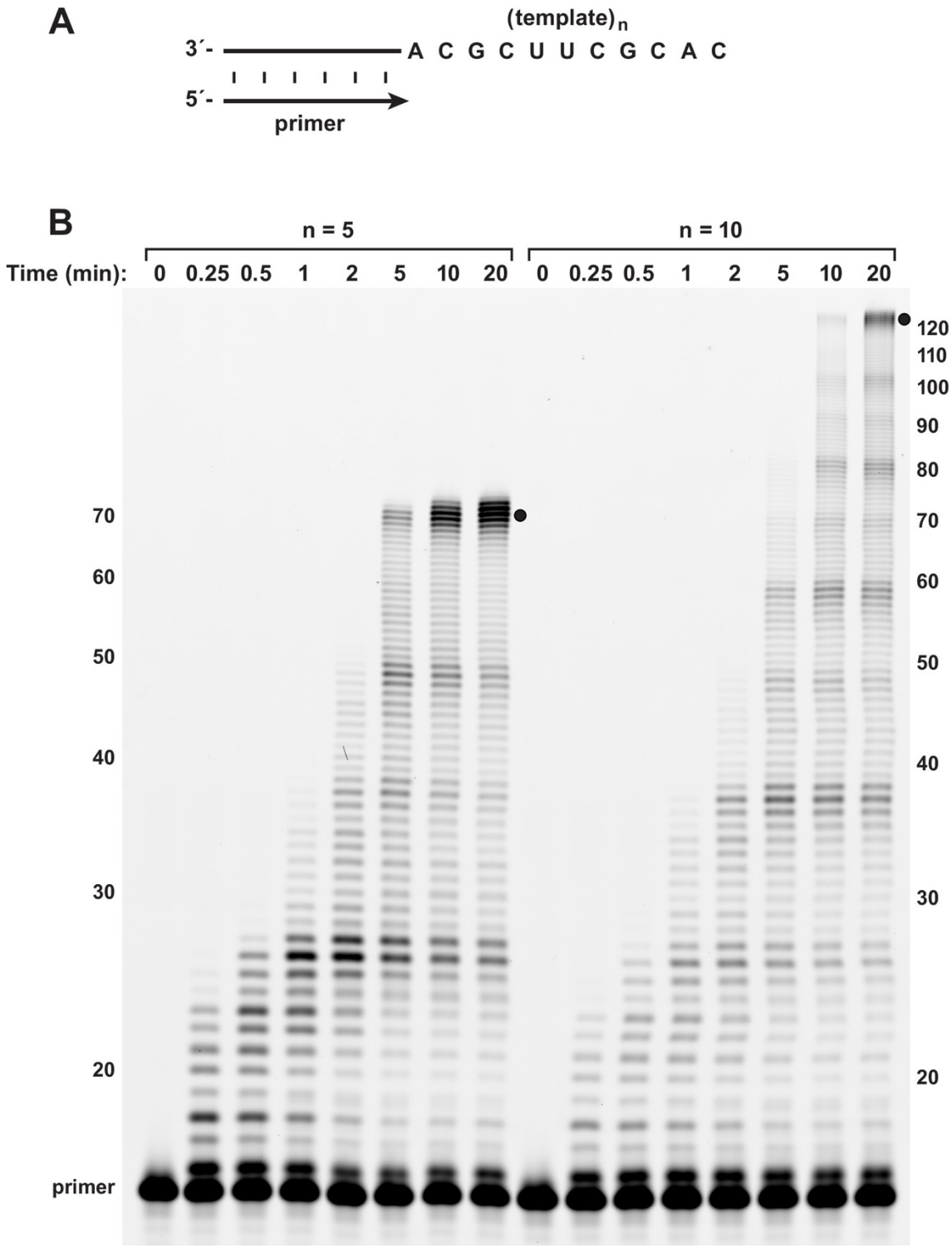
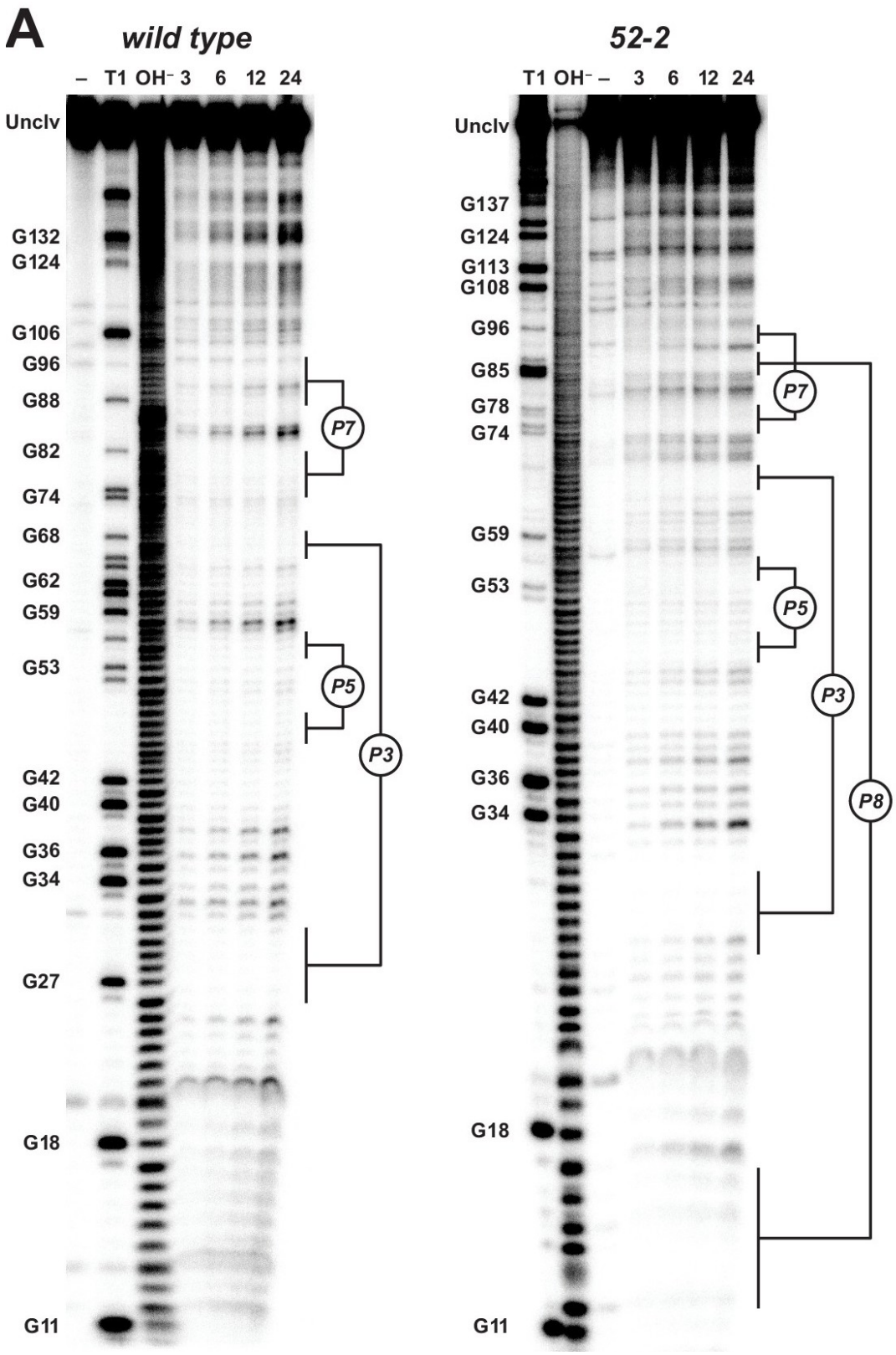

A
(template)n
3´-
A C G C U U C G C A C
5´-
primer
B
n = 5
n = 10
Time (min):
0 0.25 0.5 1 2 5 10 20 0 0.25 0.5 1 2 5 10 20
120
110
100
90
80
70
60
50
40
30
20
70
60
50
40
30
20
primer

Figure 2-3-supplement 2. Burst-phase synthesis on long RNA templates.

(**A**) The template encoded either 5 or 10 repeats of the sequence 5'-UGCGAAGCGUG-3'. (**B**) Time course of primer extension on the long templates, demonstrating persistence of the burst phase to reach full-length products (indicated by black dots) by 5 or 10 min, respectively. Reaction conditions as in Figure 3. David Horning contributed to the analysis, methodology, and design used to generate these data.


A
wild type
– T1 OH⁻ 3 6 12 24
Unclv
G132
G124
G106
G96
G88
G82
G74
G68
G62
G59
G53
G42
G40
G36
G34
G27
G18
G11
P7
P5
P3
52-2
T1 OH⁻ – 3 6 12 24
Unclv
G137
G124
G113
G108
G96
G85
G78
G74
G59
G53
G42
G40
G36
G34
G18
G11
P7
P5
P3
P8

B

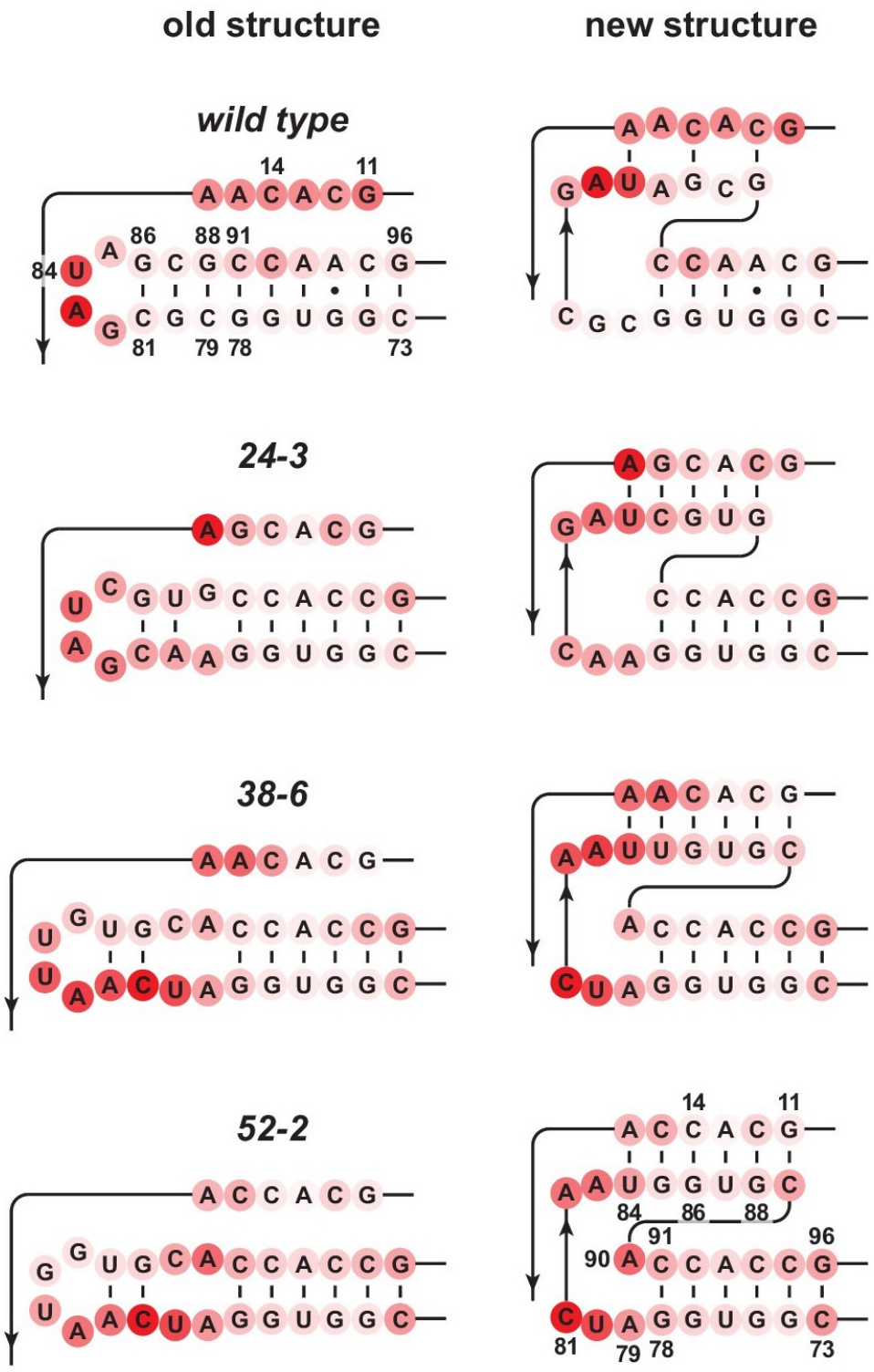


Figure 2-4. Analysis of polymerase structure by in-line probing.

(**A**) Polyacrylamide gel electrophoresis analysis of [5'-^{32} P]-labeled wild-type and 52-2 polymerases, incubated under polymerization conditions for 3, 6, 12, or 24 hr, in comparison to unincubated material (–) and material that had been subjected to partial digestion with either RNase T1 (cleaves after G residues) or NaOH. Full-length polymerase (unclv) and various G residues are labeled at the left; stem regions are labeled at the right. (**B**) Sensitivity to in-line cleavage mapped onto the P7 and P8 stems for both the old structure (*left*) and the new pseudoknot structure (*right*). For each polymerase, red circles of varying intensity indicate % cleavage after 24 hr at each nucleotide position relative to the position with the highest level of cleavage within the region shown.

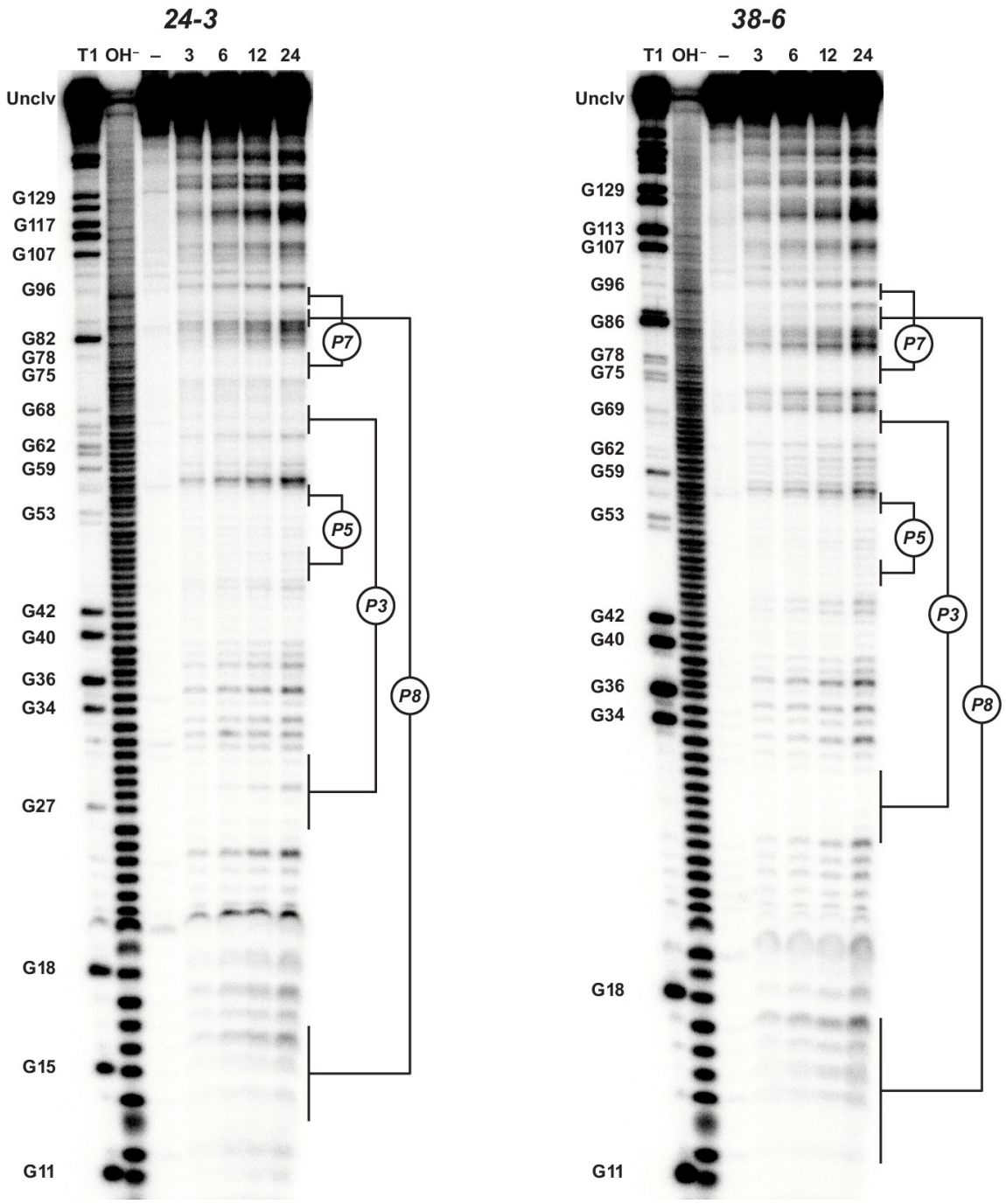

24-3
T1 OH⁻ – 3 6 12 24
Unclv
G129
G117
G107
G96
G82
G78
G75
G68
G62
G59
G53
G42
G40
G36
G34
G27
G18
G15
G11
P7
P5
P3
P8
38-6
T1 OH⁻ – 3 6 12 24
Unclv
G129
G113
G107
G96
G86
G78
G75
G69
G62
G59
G53
G42
G40
G36
G34
G18
G11
P7
P5
P3
P8

Figure 2-4-supplement 1. Analysis of polymerase structure by in-line probing. Polyacrylamide gel electrophoresis analysis of [5'-^{32} P]-labeled 24-3 and 38-6 polymerases, incubated under polymerization conditions for 3, 6, 12, or 24 hr, in comparison to unincubated material (–) and material that had been subjected to partial digestion with either RNase T1 (cleaves after G residues) or NaOH. Full-length polymerase (unclv) and various G residues are labeled at the left; stem regions are labeled at the right.

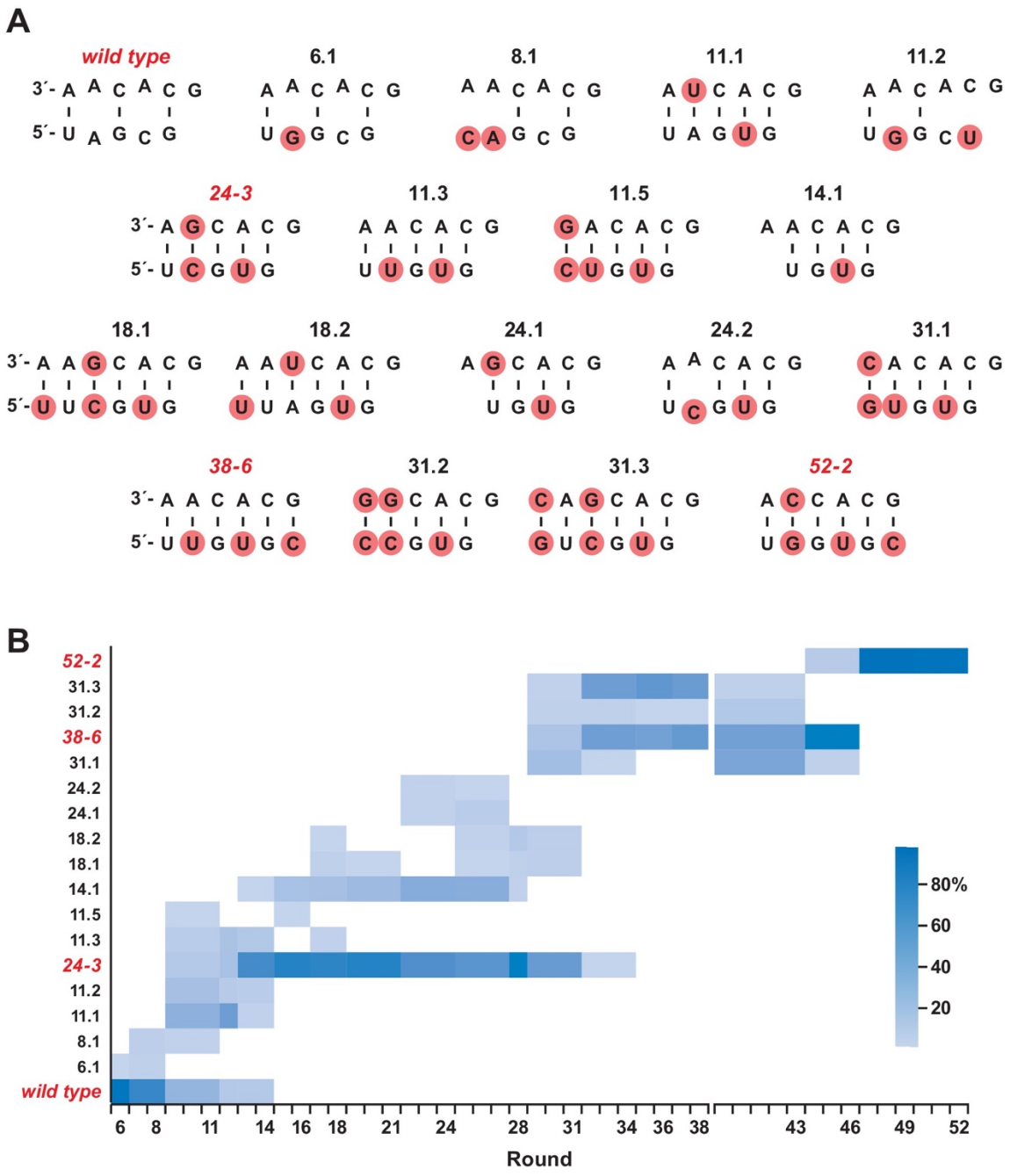

A
wild type
6.1
8.1
11.1
11.2
24-3
11.3
11.5
14.1
18.1
18.2
24.1
24.2
31.1
38-6
31.2
31.3
52-2
B
52-2
31.3
31.2
38-6
31.1
24.2
24.1
18.2
18.1
14.1
11.5
11.3
24-3
11.2
11.1
8.1
6.1
wild type
80%
60
40
20
6 8 11 14 16 18 21 24 28 31 34 36 38 43 46 49 52
Round

Figure 2-5. Composition of the evolving population.

(**A**) The 18 most highly represented sequence clusters for the P8 stem. Clusters that include the wild-type, 24-3, 38-6, and 52-2 polymerases are named for those polymerases; all other clusters are named according to the round in which they first appeared at a frequency of >1%. Red circles indicate mutations relative to the wild type. Note that for clusters 14.1 and 24.1, an A residue has been deleted in the bottom strand. (**B**) Heatmap depicting the representation of the 18 clusters over the course of evolution (scale bar at right). Axis break after round 38 indicates that the 38-6 polymerase was isolated from the population and mutated at a frequency of 10% per nucleotide position before resuming directed evolution.

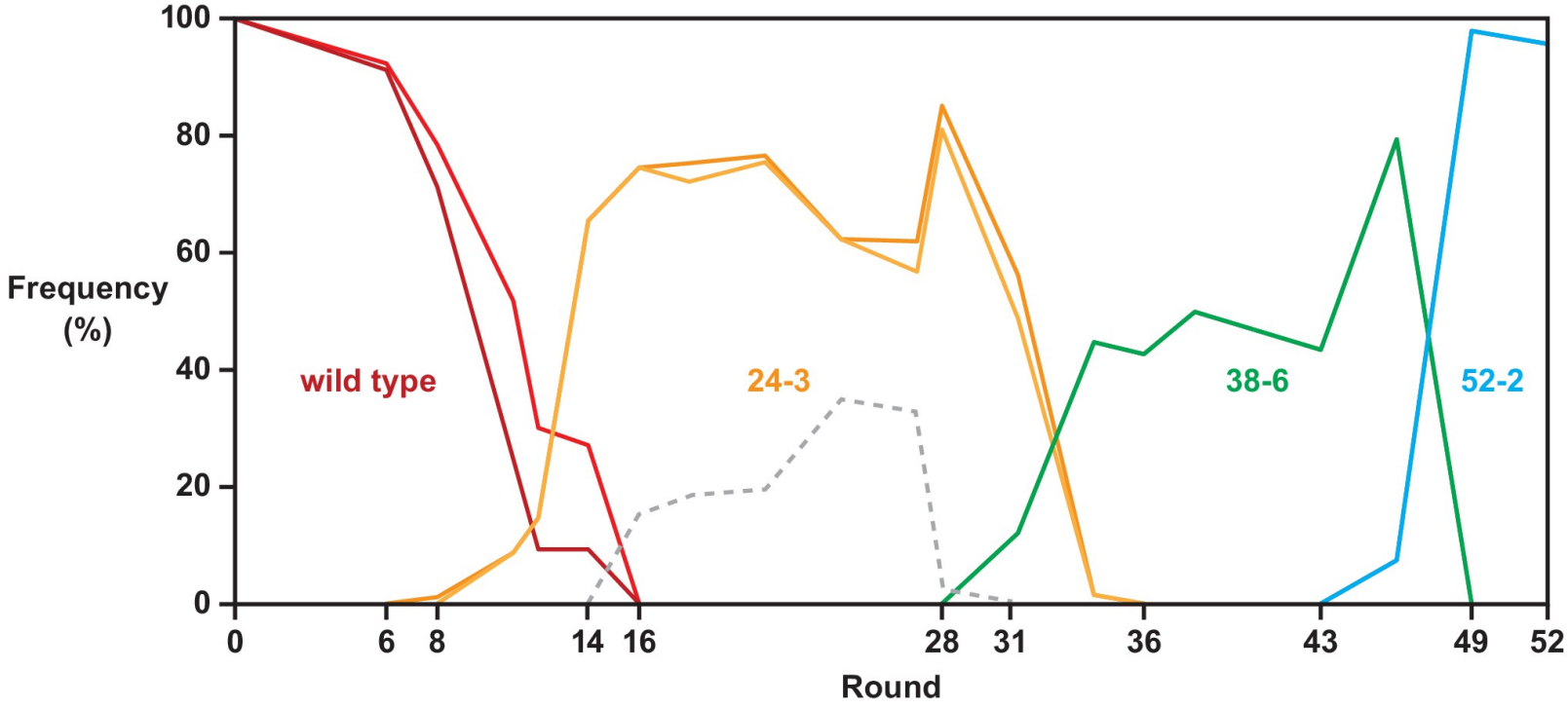

100
80
60
40
20
0
Frequency
(%)
wild type
24-3
38-6
52-2
0
6
8
14
16
28
31
36
43
49
52
Round

Figure 2-5-supplement 1. Composition of each strand of the P8 stem over the course of evolution.

Individual frequency of occurrence of variants of the 5’ and 3’ strands of P8 (darker and lighter colors, respectively) corresponding to those that occur in the wild-type (red), 24-3 (gold), 38-6 (green), and 52-2 (blue) polymerases. For the 38-6 and 52-2 polymerases, the frequencies were identical for the two strands. The first four nucleotides of the 5’ strand are fixed as part of the primer binding site, thus limiting sequence variability to two nucleotides. For this reason, some variants appearing in rounds 16–28 (gray) could not be assigned to a named polymerase because they contain the same two nucleotides in the 5’ strand as both the wild-type and 38-6 polymerases. Gerald Joyce contributed to the analysis, methodology, and design used to generate these data.

Tables

Table 2-Supplementary file 1. Parameters for directed evolution of polymerase ribozymes.

Round	Selection method	+NTPs	Time	[Mg^{2+}]	Mutagenesis
1	gel-shift	20	24	200	
2	"	"	"	"	
3	"	"	"	"	
4	"	"	"	"	+
5	vitamin B_{12} aptamer	12	4	"	
6	"	"	"	"	
7	"	"	"	"	
8	gel-shift	18	24	"	
9	"	"	"	"	
10	"	"	"	"	
11	"	"	"	"	
12	"	"	"	"	+
13	"	30	2	"	
14	"	"	"	"	
15	"	"	"	"	
16	"	"	"	"	
17	gel-shift & GTP aptamer	"	"	"	+
18	gel-shift	32	6	"	

19	"	"	1	"	
20	"	"	0.25	"	
21	"	40	"	"	
22	gel-shift & GTP aptamer	30	"	"	+
23	"	"	"	"	
24	"	"	"	"	
25	hammerhead	33	2	"	
26	"	"	0.33	"	
27	"	"	"	"	
28	"	"	"	"	
29	"	"	"	"	
30	"	"	"	"	
31	gel-shift & hammerhead	"	"	"	
32	"	"	"	"	
33	"	"	"	"	
34	"	"	0.08	"	
35	"	"	"	"	
36	"	"	"	"	
37	"	"	"	"	
38	"	"	"	"	10%
39	urea wash	28	15	"	
40	"	"	2.5	"	

41	gel-shift & hammerhead	33	20	"	
42	"	"	2	"	
43	"	"	"	"	+
44	"	"	1	"	+
45	"	"	0.4	50	+
46	"	"	4	"	+
47	"	"	1	"	+
48	"	"	0.25	"	+
49	"	"	"	"	+
50	"	"	"	"	+
51	"	"	1	25	
52	"	"	24	50	

The third column indicates the number of NTPs to be added in order to meet the selection criterion. “+” indicates PCR mutagenesis; “10%” indicates random mutagenesis of the 38-6 polymerase at 10% degeneracy per position. David Horning contributed to the analysis, methodology, and design used to generate this table.

Table 2-Supplementary file 2. Fidelity of the 52-2 polymerase.

Product [Mg^{2+}]	Expected	Observed					
		A	C	G	U	Del	Ins
Hammerhead 200 mM	A	**85.9**	1.6	8.2	0.4	3.8	0.3
	C	0.3	**90.1**	0.7	8.4	0.6	0.2
	G	0.4	0.5	**98.3**	0.1	0.7	0.2
	U	0.6	3.9	1.3	**92.9**	1.4	0.2
Hammerhead 50 mM	A	**90.4**	0.8	5.4	0.2	3.2	0.1
	C	0.1	**93.4**	0.2	6.0	0.3	0.0
	G	0.1	0.2	**99.2**	0.1	0.4	0.1
	U	0.3	2.5	0.9	**94.9**	1.4	0.1
Ligase 200 mM	A	**79.3**	3.7	9.9	1.1	5.9	0.4
	C	0.5	**89.8**	1.1	6.0	2.6	0.5
	G	0.8	0.4	**97.2**	0.3	1.2	0.4
	U	2.8	6.8	5.7	**72.4**	12.3	0.4

The 52-2 polymerase was used to synthesize the hammerhead and class I ligase ribozymes (see supplementary file 4 for sequences). The hammerhead was synthesized in the presence of either 200 or 50 mM Mg^{2+} and the ligase was synthesized in the presence of 200 mM Mg^{2+}. The full-length products were analyzed by deep sequencing to obtain the frequencies of each type of mutation. The average fidelity was calculated as the geometric mean of the fidelities for each templating nucleobase, which gave values of 91.7% or

94.4% for the hammerhead with either 200 or 50 mM Mg^{2+}, respectively, and 84.1% for the ligase. Deletions were included as mutations and insertions were treated as a single mutation at the immediately upstream position. David Horning and I generated analysis, methodology, and design used to generate this table.

Table 2-Supplementary file 3. Prevalent sequence clusters over the course of evolution.

Round	Cluster	Nucleotides 9–17	Nucleotides 83–95	Frequency
6	wild type	CCGCACAAA	AUAGCG--CCAAC	91.2
6	6.1	CCGCACAAA	AUGGCG--CCAAC	1.2
8	wild type	CCGCACAAA	AUAGCG--CCAAC	64.0
8	8.1	CCGCACAAA	CAAGCG--CCAAC	4.8
8	6.1	CCGCACAAA	AUGGCG--CCAAC	3.5
8	wild type	CCGCACAAA	AUAGCG--CCAUC	2.1
8	8.2	CCGCACAGA	AUGGCG--CCAAC	1.6
8	wild type	CCGCACAAA	ACAGCG--CCAGC	1.6
8	8.3	CCGCACAGA	AUAGCG--CCAAC	1.5
8	8.4	CCGCACAAA	AUAGCG--CCAGC	1.5
8	8.5	CCGCACGAA	AUAGCG--CCAAC	1.1
8	wild type	CCGCACAAA	AUAGCG--UCAAC	1.1
11	11.1	CCGCACUAA	AUAGUG--CCAAC	27.1
11	wild type	CCGCACAAA	AUAGCG--CCAAC	24.8
11	11.2	CCGCACAAA	AUGGCU--CGAAG	16.1
11	24-3	CCGCACGAA	AUCGUG--CCACC	7.4
11	11.3	CCGCACAAA	AUUGUG--CCACC	6.1
11	8.1	CCGCACAAA	CAAGCG--CCAAC	2.4
11	11.1	CCGCACUAA	AUAGUG--CCACC	1.5

11	24-3	CCGCACGAA	AUCGUG--CCAAC	1.3
11	11.4	CCGCACAAA	AUCGCG--CCACC	1.2
11	11.5	CCGCACAGA	ACUGUG--CCACC	1.1
11	11.4	CCGCACAAA	AUCGCG--CCAAC	1.0
12	11.1	CCGCACUAA	AUAGUG--CCAAC	45.2
12	11.3	CCGCACAAA	AUUGUG--CCACC	13.7
12	24-3	CCGCACGAA	AUCGUG--CCACC	13.6
12	wild type	CCGCACAAA	AUAGCG--CCAAC	7.2
12	11.2	CCGCACAAA	AUGGCU--CGAAG	7.1
12	wild type	CCGCACAAA	AUAGCG--CCAUC	2.1
12	11.1	CCGCACUAA	AUAGUG--CCACC	1.6
12	24-3	CCGCACGAA	AUCGUG--CCAAC	1.1
14	24-3	CCGCACGAA	AUCGUG--CCACC	63.4
14	11.3	CCGCACAAA	AUUGUG--CCACC	10.1
14	wild type	CCGCACAAA	AUAGCG--CCAAC	9.3
14	11.2	CCGCACAAA	AUGGCU--CGAAG	6.6
14	11.1	CCGCACUAA	AUAGUG--CCAAC	2.5
14	24-3	CCGCACGAA	AUCGUG--CCAAC	2.1
14	14.1	CCGCACAAA	-U-GUG--CCACC	1.3
16	24-3	CCGCACGAA	AUCGUG--CCACC	73.1
16	14.1	CCGCACAAA	-U-GUG--CCACC	14.9
16	11.5	CCGCACAGA	ACUGUG--CCACC	1.7

16	24-3	CCGCACGAA	AUCGUG--CCAUC	1.4
18	24-3	CCGCACGAA	AUCGUG--CCACC	69.9
18	14.1	CCGCACAAA	-U-GUG--CCACC	16.2
18	24-3	CCGCACGAA	AUCGUG--CCAUC	2.2
18	11.3	CCGCACAAA	AUUGUG--CCAAC	2.1
18	18.1	CCGCACGAA	UUCGUG--CCACC	2.1
18	18.2	CCGCACUAA	UUAGUG--CCACC	1.0
18	18.1	CCGCACGAA	UUCGUG--CCAUC	1.0
21	24-3	CCGCACGAA	AUCGUG--CCACC	75.4
21	14.1	CCGCACAAA	-U-GUG--CCACC	18.3
21	14.1	CCGCACAAA	AU-GUG--CCACC	1.4
21	18.1	CCGCACGAA	UUCGUG--CCACC	1.1
24	24-3	CCGCACGAA	AUCGUG--CCACC	60.2
24	14.1	CCGCACAAA	AU-GUG--CCACC	31.3
24	24.1	CCGCACGAA	AU-GUG--CCACC	2.1
24	24.2	CCGCACAAA	AUCGUG--CCACC	2.0
24	14.1	CCGCACAAA	---GUG--CCACC	1.1
27	24-3	CCGCACGAA	AUCGUG--CCACC	55.2
27	14.1	CCGCACAAA	-U-GUG--CCACC	31.1
27	24.1	CCGCACGAA	-U-GUG--CCACC	5.7
27	18.2	CCGCACUAA	UUAGUG--CCACC	2.1
27	24.2	CCGCACAAA	AUCGUG--CCACC	1.6

27	18.1	CCGCACGAA	UUCGUG--CCACC	1.1
28	24-3	CCGCACGAA	AUCGUG--CCACC	81.1
28	18.2	CCGCACUAA	UUAGUG--CCACC	8.5
28	18.1	CCGCACGAA	UUCGUG--CCACC	4.0
28	14.1	CCGCACAAA	-U-GUG--CCACC	2.6
31	24-3	CCGCACGAA	AUCGUG--CCACC	48.8
31	31.1	CCGCACACA	AGUGUG--CCACC	17.8
31	38-6	CCGCACAAA	AUUGUGCACCACC	12.0
31	18.1	CCGCACGAA	UUCGUG--CCACC	4.8
31	31.2	CCGCACGGA	ACCGUG--CCACC	4.2
31	18.2	CCGCACUAA	UUAGUG--CCACC	3.6
31	31.3	CCGCACGAC	GUCGUG--CCAGC	2.7
31	31.4	CCGCACGAA	----------ACC	1.4
31	18.2	CCGCACUAA	UUAGUG--CCAAC	1.2
31	31.5	CCGCACGAA	----UG--CCACC	1.2
34	31.3	CCGCACGAC	GUCGUG--CCAGC	46.1
34	38-6	CCGCACAAA	AUUGUGCACCACC	44.7
34	31.2	CCGCACGGA	ACCGUG--CCACC	3.5
34	31.1	CCGCACACA	AGUGUG--CCACC	1.8
34	24-3	CCGCACGAA	AUCGUG--CCACC	1.6
36	31.3	CCGCACGAC	GUCGUG--CCAGC	53.1
36	38-6	CCGCACAAA	AUUGUGCACCACC	42.7

36	31.2	CCGCACGGA	ACCGUG--CCACC	1.7
38	38-6	CCGCACAAA	AUUGUGCACCACC	50.0
38	31.3	CCGCACGAC	GUCGUG--CCAGC	47.0
38	31.2	CCGCACGGA	ACCGUG--CCACC	1.3
43	38-6	CCGCACAAA	AUUGUGCACCACC	43.5
43	31.1	CCGCACACA	AGUGUG--CCACC	38.3
43	31.2	CCGCACGGA	ACCGUG--CCACC	9.4
43	31.3	CCGCACGAC	GUCGUG--CCAGC	3.2
46	38-6	CCGCACAAA	AUUGUGCACCACC	79.4
46	52-2	CCGCACCAA	AUGGUGCACCACC	7.4
46	31.1	CCGCACACA	AGUGUG--CCACC	3.5
46	46.1	CCGCACUAA	AUAGUGCACCACC	2.6
49	52-2	CCGCACCAA	AUGGUGCACCACC	75.6
49	"	CCGCACCAA	AUGGUGCACCUCC	17.5
49	"	CCGCACCAA	AUGGUGCCCCACC	3.7
49	"	CCGCACCAA	AUGGUGCUCCACC	1.1
52	52-2	CCGCACCAA	AUGGUGCACCACC	80.8
52	"	CCGCACCAA	AUGGUGCACCUCC	9.6
52	"	CCGCACCAA	AUGGUGCCCCACC	3.6
52	"	CCGCACCAA	AUGGUGCUCCACC	1.7

Clusters representing >1% of the population in a given round are shown.

Clusters that differ by only a single nucleotide outside the P8 stem are given

the same name. Clusters that include the wild type, 24-3, 38-6, or 52-2 polymerase are given the name of that polymerase. Gerald Joyce and I generated analysis, methodology, and design used to generate this table.

Table 2-Supplementary file 4. Sequences of RNA and DNA molecules used in this study.

Type	Name	R or DNA	Source	Sequence (5′→3′)
PCR primer	Fwd	DNA	com	GGACTAATACGACTCACTATTAGTCATTGCCGCAC
	Rev	DNA	com	GTCAGCCATGTGTTG
Splint	C1	DNA	com	GACGTATAAGCGTCCGTGCTTTTGCACGTGTGGAGTG
Polymerase primer	SP1	RNA	syn	hex-$(PEG)_4$-pcl-PEG-FAM-PEG-GGAGCGAGAA (urea)
	SP2	RNA	syn	hex-$(PEG)_4$-CGCUUAUACGUC-PEG-biotin-PEG-pcl-PEG-FAM-CACUCCACAC (hammerhead)
	P1	RNA	syn	FAM-biotin-CACUCCACAC (hammerhead)
	P2	RNA	syn	FAM-biotin-GGAAAAGACAAAUCUGCCCU (ligase)
	P3	RNA	syn	FAM-biotin-UUGCUACUACACGAC (T1 and R8)
Template	urea selection	RNA	ivt	GACAAUGACAAAAAACACUCACACACACUCCACACGGAGAGGUUUCUCGCUC

	hammerhead	RNA	ivt	GACAAUGACAAAAAACGCUUUUCGGCCUUUCGGCCUCAUCAGUACGUCGUGUGGAGUG
	ligase	RNA	ivt	GACAAUGACAAAAAAUCACUAUUGUUGAGAACGUUGGCGUUAAAGCCACCGGGGGCUGCCUCCCCUGCAUCCGAAGAUGUUCUCAAGCUCUGAGGGCAGAUUUGUCUUUUCC
	T1	RNA	ivt	GACAAUGACAAAAAACACGCUUCGCAGUCGUGUAGUAGCAA
	T5	RNA	ivt	GACAAUGACAAAAAA-(CACGCUUCGCA)$_5$-GUCGUGUAGUAGCAA
	T10	RNA	ivt	GACAAUGACAAAAAA-(CACGCUUCGCA)$_{10}$-GUCGUGUAGUAGCAA
	R8	RNA	ivt	GACAAUGACAAAAAACACUCACACAGGAGAGGUCACUCCACACGUCGUGUAGUAGCAA
Ligation	b1-207t	RNA	ivt	GGAAAAGACAAAUCUGCCCUCAGAGCUUGAGAACAUCUUCGGAUGCAGGGGAGGCAGCCCCCGGUGGCUUUAACGCCAACGUUCUCAACAAUAGUGA

	S1	R/DNA	syn	Cy5-d(AAA)-r(CCAGUC)
	S2	RNA	syn	pppGGAACAUUAUACGACUGGCACCAU
Sequencing	Fwd2	DNA	com	GGGGGGATGCTACATG
	Fwd3	DNA	com	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTACAGGGCACTCCACAC
	Fwd4	DNA	com	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGAAAAGACAAATCTGCC
	Rev2	DNA	com	ATTGATGGTGCCTACAG
	Rev3	DNA	com	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTGATGGTGCCTACAG
PCR assembly fragments	F1 (52-2)	DNA	com	GGACTAATACGACTCACTATTAGTCATTGCCGCACCAAGACAAATC
	F1 (G11C)	DNA	com	GGACTAATACGACTCACTATTAGTCATTGCCCCACCAAGACAAATC
	F1 (C12G)	DNA	com	GGACTAATACGACTCACTATTAGTCATTGCCGGACCAAGACAAATC
	F1 (A13U)	DNA	com	GGACTAATACGACTCACTATTAGTCATTGCCGCTCCAAGACAAATC
	F1 (C14G)	DNA	com	GGACTAATACGACTCACTATTAGTCATTGCCGCAGCAAGACAAATC

	F1 (C15G)	DNA	com	GGACTAATACGACTCACTATTAGTCAT TGCCGCACGAAGACAAATC
PCR assembly fragments	F1 (A16U)	DNA	com	GGACTAATACGACTCACTATTAGTCAT TGCCGCACCTAGACAAATC
	F1 (wt)	DNA	com	GGACTAATACGACTCACTATTAGTCAT TGCCGCACAAAAAGACAAATCTGCCCT CAGAGC
	F1 (wt/yknot)	DNA	com	GGACTAATACGACTCACTATTAGTCAT TGCCGCACCAAGACAAATCTGCCCTCA GAGCTT
	F2 (52-2)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTTGGTGCGGCAA
	F2 (G11C)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTTGGTGGGGCAA
	F2 (C12G)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTTGGTCCGGCAA
	F2 (A13U)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTTGGAGCGGCAA
	F2 (C14G)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTTGCTGCGGCAA
	F2 (C15G)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTTCGTGCGGCAA

	F2 (A16U)	DNA	com	GTAGATGTTCTCAAGCTCTGAGGGGAG ATTTGTCTAGGTGCGGCAA
	F2 (wt)	DNA	com	GCATCCGAAGATGTTCTCAAGCTCTGA GGGCAGATTTGTC
	F3 (52-2)	DNA	com	CAGAGCTTGAGAACATCTACGGATGC AGAGGAGGGGGCCTTCGGTGGATCAA
	F3 (U80G)	DNA	com	CAGAGCTTGAGAACATCTACGGATGC AGAGGAGGGGGCCTTCGGTGGAGCAA
	F3 (C81G)	DNA	com	CAGAGCTTGAGAACATCTACGGATGC AGAGGAGGGGGCCTTCGGTGGATGAA
	F3 (A82U)	DNA	com	CAGAGCTTGAGAACATCTACGGATGC AGAGGAGGGGGCCTTCGGTGGATCTA
	F3 (wt)	DNA	com	TTGAGAACATCTTCGGATGCAGAGGA GGCAGCCTTCGGTGG
	F4 (52-2)	DNA	com	GTGTTGAGAACGGTGGTGCACCATTGA TCCACCGAAGGCCCCC
	F4 (C89G)	DNA	com	GTGTTGAGAACGGTGGTCCACCATTGA TCCACCGAAGGCCCCC
	F4 (G88C)	DNA	com	GTGTTGAGAACGGTGGTGGACCATTGA TCCACCGAAGGCCCCC
	F4 (U87A)	DNA	com	GTGTTGAGAACGGTGGTGCTCCATTGA TCCACCGAAGGCCCCC

	F4 (G86C)	DNA	com	GTGTTGAGAACGGTGGTGCAGCATTGA TCCACCGAAGGCCCCC
	F4 (G85C)	DNA	com	GTGTTGAGAACGGTGGTGCACGATTGA TCCACCGAAGGCCCCC
	F4 (U84A)	DNA	com	GTGTTGAGAACGGTGGTGCACCTTTGA TCCACCGAAGGCCCCC
	F4 (U80G)	DNA	com	GTGTTGAGAACGGTGGTGCACCATTGC TCCACCGAAGGCCCC
	F4 (C81G)	DNA	com	GTGTTGAGAACGGTGGTGCACCATTCA TCCACCGAAGGCCCC
	F4 (C81G/G88C)	DNA	com	GTGTTGAGAACGGTGGTGGACCATTCA TCCACCGAAGGCCCCC
	F4 (A82U)	DNA	com	GTGTTGAGAACGGTGGTGCACCATAG ATCCACCGAAGGCCCCC
	F4 (A82U/U87A)	DNA	com	GTGTTGAGAACGGTGGTGCTCCATAGA TCCACCGAAGGCCCCC
	F4 (wt)	DNA	com	TTGGGTGTCTGTTGAGAACGTTGG--CG CTATCGCGCCACCGAAGGCTGCCTCCT C
	F4 (wt/yknot)	DNA	com	TTGGGTGTCTGTTGAGAACGGTGGTGC ACCATTGATCCACCGAAGGCTGCCTCC TC

	F5 (52-2)	DNA	com	ACCACCGTTCTCAACACGTACCCGAACATAAAAAGACCTGACAAAAAGGCGATGTTA
	F5 (wt)	DNA	com	CGTTCTCAACAGACACCCAAAAA-AAAAAGACCTGACGAAAAGGCGATGTTAGA
	F6 (52-2)	DNA	com	GTCAGCCATGTGTTGGGCATGGTACCTGTGCGTGTCTAACATCGCCTTTTTGTCAG
	F6 (wt)	DNA	com	GTCAGCCATGTGTTGGGTATG--ACCTGGGCGTGTCTAACATCGCCTTTTCGTC

The molecules were synthesized in-house (syn), purchased from IDT (com), or prepared by *in vitro* transcription (ivt). The T7 RNA polymerase promoter sequence is underlined. Sequences in green indicate the complementary tag on the ribozyme and templates used to improve processivity. Sequences in blue indicate the primer binding site for RNA-templated RNA polymerization. Nucleotides in red are mutations relative to the 52-2 polymerase, that were introduced during PCR assembly. hex, hexynyl group for click chemistry; PEG, polyethylene glycol linker; pcl, photocleavable linker; FAM, 6-fluorescein label; Cy5, cyanine 5-methine label; ppp, 5′-triphosphate added by chemical synthesis. David Horning and I generated analysis, methodology, and design used to generate this table.