EMPIRICAL INVESTIGATIONS OF RNA FITNESS LANDSCAPES: HARNESSING THE POWER OF HIGH-THROUGHPUT SEQUENCING AND EVOLUTIONARY SIMULATIONS

by

Devin Pratt Bendixsen

A dissertation
submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Biomolecular Sciences
Boise State University

August 2018

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the dissertation submitted by

Devin Pratt Bendixsen

Dissertation Title: Empirical Investigations of RNA Fitness Landscapes: Harnessing the Power of High-Throughput Sequencing and Evolutionary Simulations

Date of Final Oral Examination: 11 July 2018

The following individuals read and discussed the dissertation submitted by student Devin Pratt Bendixsen, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Eric J. Hayden, Ph.D.	Chair, Supervisory Committee
Matthew L. Ferguson, Ph.D.	Member, Supervisory Committee
Elton Graugnard, Ph.D.	Member, Supervisory Committee

The final reading approval of the dissertation was granted by Eric J. Hayden, Ph.D., Chair of the Supervisory Committee. The dissertation was approved by the Graduate College.

DEDICATION

I dedicate this to my amazing, beautiful wife and my two perfect kids: Owen and Fynnley. Without you this never would have happened.

ACKNOWLEDGEMENTS

I would like to foremost thank my wife Melissa for her hard-work and countless sacrifices that has made this possible. She dealt with my long hours and numerous sleepless nights for years without complaint. She prioritized our family and made it look effortless. This dissertation is just as much a fruit of her labor and sacrifice as it is mine. I would also like to thank my two wonderful children: Owen and Fynnley. You’ve given me the drive and desire to work hard and progress in my field. I hope to provide a bright future for you and our family.

I would also like to acknowledge and thank my advisor, Dr. Eric Hayden. He taught me everything that I know about RNA and evolutionary theory. He gave me the opportunity to develop my own research directions and have real ownership of my research. I would also like to thank the other members of my committee: Matthew Ferguson and Elton Graugnard. Their encouragement and research advice were essential to this research.

Lastly, I would like to thank the Boise State University Biomolecular Sciences PhD program. It was a great opportunity to be admitted to the program while it was in its infancy and be able to contribute to the culture of the program. Also, I would like to thank Beth Gee for her incredible help throughout this process.

This research was supported by Boise State University (Biomolecular Sciences Graduate Programs), National Science Foundation Directorate for Biological Sciences (Grant No. MCB-1413664), National Science Foundation Office of Integrative Activities

(Grant No. OIA-1738865), and National Aeronautics and Space Administration (Grant No. 80NSSC17K0738).

DISSERTATION ABSTRACT

Fitness landscapes or adaptive landscapes represent the mapping of genotype (sequence) to phenotype (function or fitness). Originally proposed as a metaphor to envision evolutionary processes and mutational interactions, the fitness landscape has recently transitioned from theoretical to empirical. This is due in part to advances in DNA synthesis and high-throughput sequencing. This allows for the construction and analysis of empirical fitness landscapes that encompass thousands of genotypes. These landscapes provide tractable insight into mutational pathways, the predictability of evolution or even the evolution of life. RNA enzymes (ribozymes) are an attractive model system for the construction of empirical fitness landscapes. Ribozymes function as both a genotype (primary RNA sequence) and a phenotype (catalytic function). To construct and characterize empirical RNA fitness landscapes, two high-throughput functional assays (self-cleavage and self-ligation), including a technique to improve data recovery from high-throughput sequencing using phased nucleotide inserts (Appendix A), were developed and implemented. Following fitness landscape construction, a stochastic evolutionary model was developed and employed based on the Wright-Fisher model. This model follows the principles of Darwinian evolution and allows a population to explore the fitness landscape by means of mutation and selection. These newly developed tools allowed for a novel approach to important evolutionary questions.

Chapter 1 explored the evolution of innovation at the intersection of two ribozyme functions: self-cleavage and self-ligation. Evolutionary innovations are qualitatively

novel traits that emerge through evolution. Theories have suggested that innovations can occur where two genotype networks are in close proximity. However, only isolated examples of intersections have been investigated. The fitness landscape between the two ribozyme functions was explored by determining the ability of numerous neighboring RNA sequences to catalyze two different chemical reactions. This revealed that there was extensive functional overlap, and over half the genotypes can catalyze both functions to some extent. Data-driven evolutionary simulations found that these numerous points of intersection facilitated the discovery of a new function, yet the rate of optimization depended upon the starting location in the genotype network. This study constructed a fitness landscape where genotype networks intersect and uncovered the implications for evolutionary innovations.

Chapter 2 determined the effect of higher sequence space complexity and dimensionality on evolutionary adaptation in RNA fitness landscapes. The complexity and dimensionality of landscapes scale with the length of the RNA molecule. For this study, complexity was defined as the size of the genotype space and dimensionality as the number of edges connecting each genotype (node) to other genotypes that differ by a single mutation. Low-dimensional ‘direct’ landscapes consisting of only two possible nucleotides at various positions were compared to higher-dimensional ‘indirect’ landscapes that had all four nucleotides at the same positions. Indirect pathways contributed to the ruggedness and navigability of landscapes. Increased dimensionality in RNA fitness landscapes had the potential to circumvent fitness valleys, however indirect pathways also harbored stasis genotypes isolated by reciprocal sign epistasis.

Chapter 3 applied ancestral sequence resurrection and fitness landscape construction to naturally evolved ribozymes. The CPEB3 ribozyme is highly conserved in mammals and has been linked to episodic memory. By predicting, ‘resurrecting’ and functionally characterizing ancient gene sequences, hypotheses about gene function or selection can be empirically tested in an evolutionary context. Using the extant ribozyme sequences found in a range of mammalian species as a basis for inference of ancestral sequences, a phylogenetic fitness landscape was experimentally resurrected and reconstructed. A single high-activity *ancestral* sequence was found to be highly conserved and purifying selection is expected to have reduced the accumulation of mutations through geologic time. Many of the extant mammalian ribozyme sequences had high ribozyme activity, however a few had relatively low activity. Yet, given the local fitness landscape, a selective pressure for functional ribozyme sequences was seen. A single nucleotide polymorphism (SNP) found in humans, reduced co-transcriptional ribozyme activity *in vitro* and might alter our understanding of the CPEB3 ribozyme’s biological function.

Chapter 4 analyzed epistatic interactions in four published RNA fitness landscapes generated from high-throughput analyses. Two of the landscapes were assessed *in vivo* and two were assessed *in vitro*. Epistasis occurs when the effects of some mutations are dependent on the presence or absence of other mutations. The data allowed for an analysis of the distribution of fitness effects of individual mutations as well as combinations of two or more mutations. Two different approaches to measuring epistasis in the data both revealed a predominance of negative epistasis, such that higher combinations of two or more mutations are typically lower in fitness than expected from

the effect of each individual mutation. This finding differed from studies using computationally predicted RNA but is similar to mutational experiments in protein enzymes.

The work presented here represents a significant contribution to our ability to construct and empirically characterize RNA fitness landscapes. The development of two high-throughput ribozyme assays opens the door for further empirical landscape construction. The implementation of data-driven stochastic evolutionary modeling allows for a clearer evolutionary characterization of the landscape. Understanding the connection between genotype and phenotype in RNA systems is important for designing RNA functions, improving *in vitro* selections and understanding the origins and evolution of new RNA functions (innovations). Applying these advances yielded valuable information about evolutionary innovations, the effects of higher dimensionality, evolution of extant ribozymes and the prevalence of epistasis in RNA fitness landscapes. Construction and analysis of empirical RNA fitness landscapes provides tractable insight into evolutionary processes, mutational pathways and the predictability of evolution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DISSERTATION INTRODUCTION

Fitness landscapes or adaptive landscapes are often called genotype-phenotype maps because they link the genetic sequence (genotype) to its corresponding function or fitness (phenotype). Landscapes are often depicted as three-dimensional surfaces with genotype space on the x- and y-axes and fitness on the z-axis. This results in a hilly topography where high-fitness genotypes occupy peaks and low-fitness genotypes occupy valleys. Catalytic RNA molecules (ribozymes) are an attractive model system for the construction of empirical fitness landscapes. Fitness landscapes offer powerful insight into evolutionary adaptation and mutational interactions. However, the construction and characterization of RNA fitness landscapes is difficult and limited on many fronts. This required the expansion of the *fitness landscape toolbox* by the development of novel high-throughput tools and evolutionary modeling techniques (Fig. 1). The work presented here developed new tools in the *fitness landscape toolbox* and applies these tools to answer important evolutionary questions.

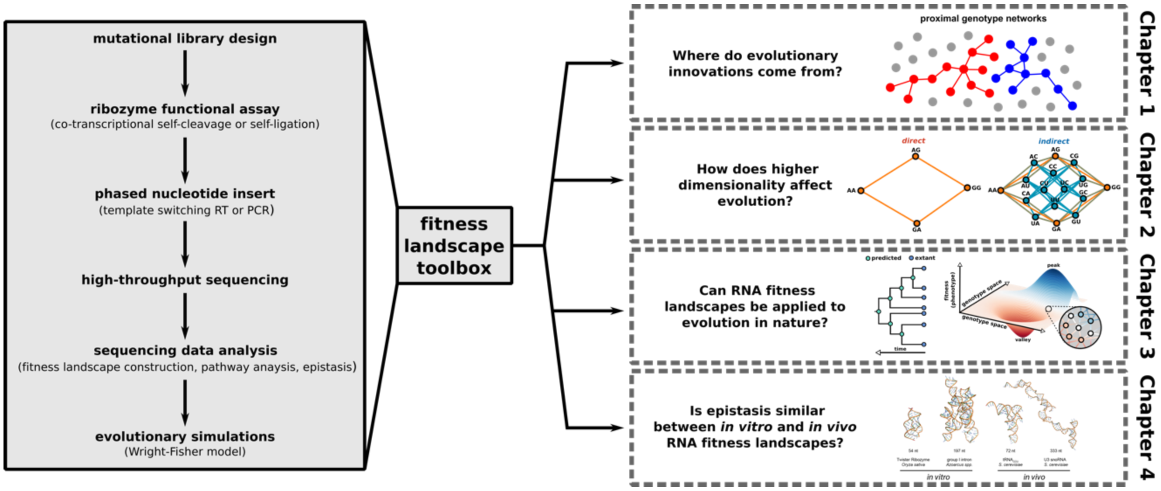

mutational library design
ribozyme functional assay
(co-transcriptional self-cleavage or self-ligation)
phased nucleotide insert
(template switching RT or PCR)
high-throughput sequencing
sequencing data analysis
(fitness landscape construction, pathway anaysis, epistasis)
evolutionary simulations
(Wright-Fisher model)
fitness landscape toolbox
Where do evolutionary innovations come from?
proximal genotype networks
Chapter 1
How does higher dimensionality affect evolution?
direct
indirect
AA
AG
GA
GG
AC
CG
CC
AU
CU
UC
UG
CA
UU
GC
UA
GU
Chapter 2
Can RNA fitness landscapes be applied to evolution in nature?
predicted
extant
time
fitness (phenotype)
genotype space
genotype space
peak
valley
Chapter 3
Is epistasis similar between *in vitro* and *in vivo* RNA fitness landscapes?
54 nt
197 nt
72 nt
333 nt
Twister Ribozyme Oryza sativa
group I intron Azoarcus spp.
tRNA S. cerevisiae
U3 snoRNA S. cerevisiae
in vitro
in vivo
Chapter 4

Figure 1. Dissertation overview: Contributions to and application of the *fitness landscape toolbox*.

The architecture and structure of fitness landscapes are a key determinant of evolutionary exploration and navigation (Beerenwinkel et al. 2006; de Visser and Krug 2014). When the metaphor of a fitness landscape was first proposed in the 1930s, it was highly theoretical and the range of genotype space that was able to be characterized was severely limited (Wright 1932; Pigliucci 2008). However, due to advances in DNA manipulation during synthesis and major strides in high-throughput or deep sequencing, researchers are able to construct highly-complex fitness landscapes using empirical data (Hietpas et al. 2011; Jimenez et al. 2013; Bank et al. 2015). These complex fitness landscapes hold valuable information and can inform us about molecular and population genetic mechanisms that drive evolution (Hartl 2014). Fitness landscapes offer tangible glimpses into the complex nature of evolution and provide tractable insight into mutational pathways (Poelwijk et al. 2007; Kogenaru et al. 2009; Franke et al. 2011; Wu et al. 2016), the predictability of evolution (de Visser and Krug 2014; Lässig et al. 2017; Gorter et al. 2018) or the evolution of life (Athavale et al. 2014; Kun and Szathmáry 2015). This information can be utilized in designing RNA functions, improving *in vitro* selections, understanding the origins and evolution of new functions or even to predict or forecast future evolutionary directions.

The construction of fitness landscapes requires selecting a model system. RNA is the only macromolecule that has the ability to function as both genotype and phenotype. For ribozymes, the genotype is the primary RNA sequence and the phenotype is the catalytic function that the ribozyme can perform. This allows for straightforward rapid construction of RNA fitness landscapes (Pitt and Ferré-D'Amaré 2010). The two most common ribozyme phenotypes are self-cleavage and self-ligation. Ribozymes have been shown to evolve from random genotype space (Ameta et al. 2014) and are often selected for using *in vitro* selection (Robertson and Joyce 1990; Bartel and Szostak 1993; Pressman et al. 2017). This supports the premise that life originated through the RNA World (Alberts et al. 2002; Orgel 2004; Pressman et al. 2015). Furthermore, synthetic DNA libraries can be designed that contain >25,000 unique sequences that can be transcribed and functionally characterized. This capability primes the ribozyme model system for the high-throughput construction of fitness landscapes.

However, the construction and characterization of RNA fitness landscapes can be difficult and presents several obstacles that must be overcome. Due to the high complexity of fitness landscapes and the corresponding genotype space, high-throughput functional assays for ribozyme fitness were needed. These assays often rely on harnessing the power of high-throughput sequencing. Using high-throughput sequencing to determine the reacted state (cleaved/ uncleaved or ligated/ unligated) requires that the ribozymes chosen perform either self-ligation or self-cleavage reactions. Ribozymes that perform either 5' self-cleavage or 5' self-ligation were chosen for this work. One major issue that had to be overcome is that the self-cleavage or self-ligation reaction leaves the 5’ end of the RNA molecule variable due to the presence or absence of cleavage product or ligation substrate. This is important because in order to prepare the reverse-transcribed cDNA sample for high-throughput sequencing on Illumina platforms, PCR binding sites were needed on each end. To combat this issue, two high-throughput functional ribozyme assays were developed (Fig. 2).

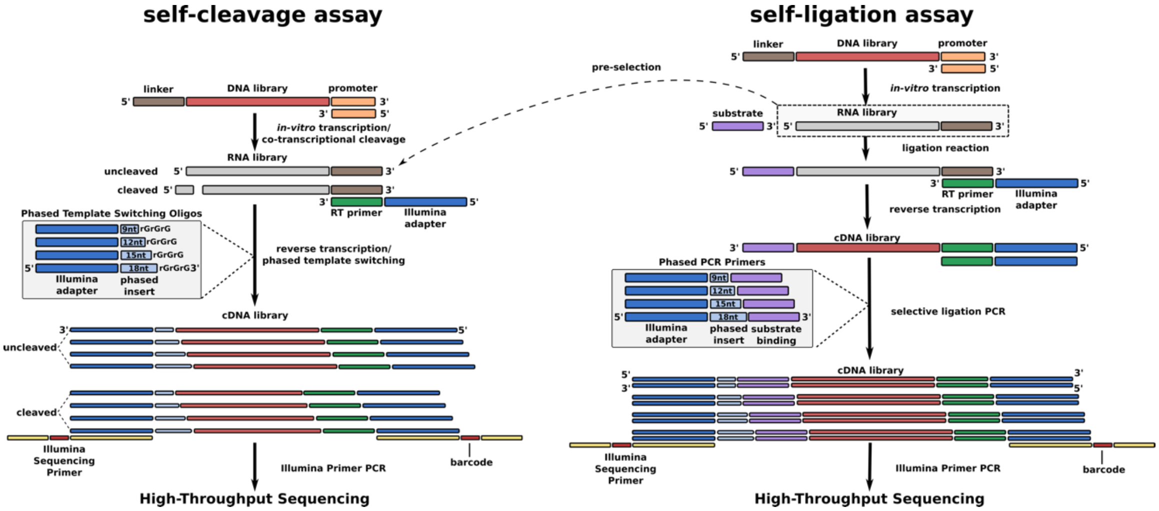


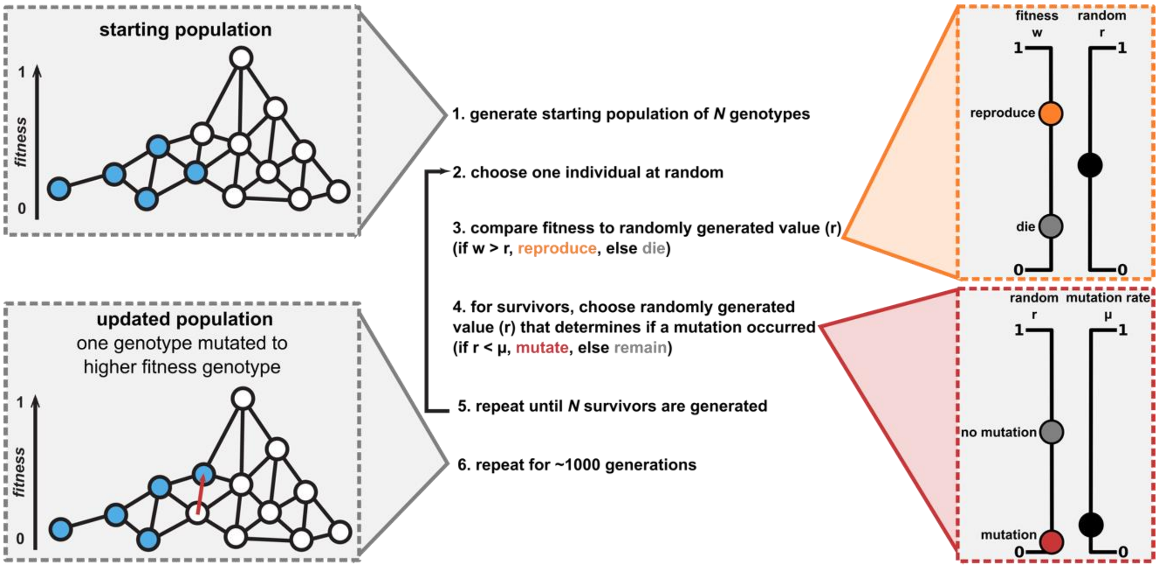
Figure 2. Overview of two high-throughput ribozyme functional assays.

The self-cleavage assay allows for co-transcriptional self-cleavage followed by a unique template-switching reverse-transcription. This reverse-transcriptase allows for the ligation of a substrate to the 5’ end of the RNA during reverse-transcription. This substrate can then be used as a PCR primer binding site for the addition of Illumina adapters. The self-ligation assay, which was developed with a colleague and co-author James Collet, uses a post-transcriptional ligation reaction. The sample is then reverse-transcribed and is followed by a selective ligation PCR, which amplifies only sequences that successfully ligated the substrate. The product then goes through a low-cycle PCR to add on the Illumina adapter. In order to calculate the fold enrichment of those that successfully ligated, the RNA library in its entirety must also be sequenced. This is accomplished by taking a pre-selection sample and using the protocol developed for the self-cleavage assay. By overcoming this obstacle, two high-throughput functional ribozyme assays were added to the *fitness landscape toolbox*.

The next barrier to the construction of fitness landscapes was that the cDNA samples generated using the high-throughput ribozyme assays were all of very low-diversity. This is due to the limitations of high-throughput platforms. Although advances have improved the capacity of sequencing platforms, we are still limited in the amount of mutations that we can design into mutational libraries. This results in samples that only differ at a limited amount of positions. Low-diversity samples cause sequencing errors due to inability to locate clonal clusters on the sequencing plate. To combat this issue, custom phased nucleotide inserts were developed (Fig. 2). These inserts can be inserted into PCR primers or template-switching oligos and results in the sequences becoming phased from one another and alleviates the low-diversity issue. The application of this technique and its potential to reduce wasting sequencing space on the addition of PhiX is reported in great detail in Appendix A.

The next issue encountered was the downstream sequence read analysis. High-throughput sequencing generates 100s of millions of sequencing reads that need to assessed and analyzed carefully. Custom sequencing analysis pipelines were developed to clean the raw sequencing reads and generate useable ribozyme activity or fitness measurements for each genotype. This was done primarily using custom Python scripts. Once fitness measurements were determined, empirical RNA fitness landscapes were able to be constructed. Using pathway analyses from custom Python scripts, important information about the accessibility of mutational pathways and epistatic interactions could be determined. This added more tools to the *fitness landscape toolbox*.

The final major contribution made to the *fitness landscape toolbox* was the ability to simulate evolving populations on the newly constructed empirical fitness landscapes (Fig. 3). Using a Wright-Fisher model (Donnelly and Weber 1985), simulated evolution is stochastic and allows for evolutionary exploration. As populations evolve, higher fitness genotypes are more likely to survive, and the average fitness of the population increases. Furthermore, fitness valleys can be crossed by stochastic events. This results in a more accurate assessment of the dynamics of natural evolution.



1. generate starting population of *N* genotypes

2. choose one individual at random

3. compare fitness to randomly generated value (r) (if w > r, reproduce, else die)

4. for survivors, choose randomly generated value (r) that determines if a mutation occurred (if r < μ, mutate, else remain)

5. repeat until *N* survivors are generated

6. repeat for ~1000 generations

Figure 3. Evolutionary simulations using a Wright-Fisher model.

With powerful tools in the *fitness landscape toolbox*, these tools were then applied to important evolutionary questions (Fig. 1). Chapter 1 explored the intersection of two RNA functions: self-cleavage and self-ligation. A previous study suggested that at-least a single genotype had the capability to perform both functions (Schultes and Bartel 2000). Theory suggested that genotype network intersections could give rise to evolutionary innovations (novel function). Chapter 2 determined the effect of higher dimensionality on evolutionary adaptation. Recent work in protein landscapes suggested that indirect, non-parsimonious pathways facilitated adaptation (Wu et al. 2016). This analysis studied the effect of indirect pathways in RNA fitness landscapes. Chapter 3 applied the principles of phylogenetics and fitness landscapes to naturally evolved ribozymes. The CPEB3 ribozyme is highly conserved in mammals and affects episodic memory (Vogler et al. 2009; Webb and Lupták 2011). The phylogenetic fitness landscape was resurrected and the local landscape was assessed where evolution occurred. Chapter 4 applied the Python analysis from the *fitness landscape toolbox* to four published fitness landscape datasets: two *in vitro* and two *in vivo*. Epistatic interactions for each dataset were calculated and compared. Overall this work represents a significant contribution to the field of evolutionary biology, both in terms of new tools in the *fitness landscape toolbox* and new insights into evolutionary processes.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2002. The RNA World and the Origins of Life. Available from: http://www.ncbi.nlm.nih.gov/books/NBK26876/

Ameta S, Winz M-L, Previti C, Jäschke A. 2014. Next-generation sequencing reveals how RNA catalysts evolve from random space. Nucleic Acids Res 42:1303–1310.

Athavale SS, Spicer B, Chen IA. 2014. Experimental fitness landscapes to understand the molecular evolution of RNA-based life. Current Opinion in Chemical Biology 22:35–39.

Bank C, Hietpas RT, Jensen JD, Bolon DNA. 2015. A Systematic Survey of an Intragenic Epistatic Landscape. Mol Biol Evol 32:229–238.

Bartel DP, Szostak JW. 1993. Isolation of new ribozymes from a large pool of random sequences. Science 261:1411–1418.

Beerenwinkel N, Pachter L, Sturmfels B. 2006. Epistasis and Shapes of Fitness Landscapes. arXiv:q-bio/0603034 [Internet]. Available from: http://arxiv.org/abs/q-bio/0603034

Donnelly P, Weber N. 1985. The Wright-Fisher model with temporally varying selection and population size. J. Math. Biology 22:21–29.

Franke J, Klözer A, Visser JAGM de, Krug J. 2011. Evolutionary Accessibility of Mutational Pathways. PLOS Computational Biology 7:e1002134.

Gorter FA, Aarts MGM, Zwaan BJ, Visser JAGM de. 2018. Local Fitness Landscapes Predict Yeast Evolutionary Dynamics in Directionally Changing Environments. Genetics 208:307–322.

Hartl DL. 2014. What Can We Learn From Fitness Landscapes? Curr Opin Microbiol 0:51–57.

Hietpas RT, Jensen JD, Bolon DNA. 2011. Experimental illumination of a fitness landscape. PNAS 108:7896–7901.

Jimenez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen IA. 2013. Comprehensive experimental fitness landscape and evolutionary network for small RNA. Proceedings of the National Academy of Sciences 110:14984–14989.

Kogenaru M, Vos MGJ de, Tans SJ. 2009. Revealing evolutionary pathways by fitness landscape reconstruction. Critical Reviews in Biochemistry and Molecular Biology 44:169–174.

Kun Á, Szathmáry E. 2015. Fitness Landscapes of Functional RNAs. Life 5:1497–1517.

Lässig M, Mustonen V, Walczak AM. 2017. Predicting evolution. Nature Ecology & Evolution 1:0077.

Orgel LE. 2004. Prebiotic Chemistry and the Origin of the RNA World. Critical Reviews in Biochemistry and Molecular Biology 39:99–123.

Pigliucci M. 2008. Sewall Wright's adaptive landscapes: 1932 vs. 1988. Biology & Philosophy 23:591–603.

Pitt JN, Ferré-D'Amaré AR. 2010. Rapid Construction of Empirical RNA Fitness Landscapes. Science 330:376–379.

Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445:383–386.

Pressman A, Blanco C, Chen IA. 2015. The RNA World as a Model System to Study the Origin of Life. Current Biology 25:R953–R963.

Pressman A, Moretti JE, Campbell GW, Müller UF, Chen IA. 2017. Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. Nucleic Acids Res. 45:8167–8179.

Robertson DL, Joyce GF. 1990. Selection In Vitro of an RNA Enzyme That Specifically Cleaves Single-Stranded DNA. Nature; London 344:467–468.

Schultes EA, Bartel DP. 2000. One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science 289:448–452.

de Visser JAGM, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet 15:480–490.

Vogler C, Spalek K, Aerni A, Demougin P, Müller A, Huynh K-D, Papassotiropoulos A, de Quervain DJ-F. 2009. CPEB3 is Associated with Human Episodic Memory. Front Behav Neurosci [Internet] 3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2691156/

Webb C-HT, Lupták A. 2011. HDV-like self-cleaving ribozymes. RNA Biol 8:719–727.

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress of Genetics 1:356–366.

Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife Sciences 5:e16965.

CHAPTER ONE: GENOTYPE NETWORK INTERSECTIONS PROMOTE EVOLUTIONARY INNOVATION

Devin P. Bendixsen[1], James Collet[2], Bjørn Østman,[3] and Eric J. Hayden[1,2]

[1]Biomolecular Sciences Graduate Program, Boise State University, Boise, ID 83725
[2]Department of Biological Science, Boise State University, Boise, ID 83725
[3]Keck Graduate Institute, Claremont, CA, 91711

Abstract

Evolutionary innovations are qualitatively novel traits that emerge through evolution and increase biodiversity. The genetic mechanisms of innovation remain poorly understood. A systems view of innovation requires the analysis of genotype networks – the vast networks of genetic variants that produce the same phenotype. Innovations can occur at the intersection of two different genotype networks. Here, we study the fitness landscape between the genotype networks of two catalytic RNA molecules (ribozymes) by determining the ability of numerous neighboring RNA sequences to catalyze two different chemical reactions. We find extensive functional overlap, and over half the genotypes can catalyze both functions to some extent. We demonstrate through evolutionary simulations that these numerous points of intersection facilitate the discovery of a new function, yet the rate of optimization depends upon the starting location in the genotype network. The study reveals the properties of a fitness landscape where genotype networks intersect, and the consequences for evolutionary innovations.

Significance Statement

It has been proposed that evolutionary innovations can occur where two genotype networks are in close proximity. However, only isolated examples of intersection sequences have been demonstrated. Here we show that the genotype networks of two different RNA enzymes overlap extensively through numerous intersection genotypes. We applied two high-throughput RNA assays to the same sequences and found that over half of the 16,384 mutational neighbors studied perform both functions to some extent. We conducted data-driven evolutionary simulations, which show that adaptation rate depends upon the specific starting locations in the genotype networks. The extensive overlap suggest that functional divergence is likely to precede gene duplication in the evolution of RNA innovations, and that many new functional intersections are awaiting discovery.

Introduction

The mechanisms by which evolution produces new functions has intrigued biologists since the earliest formulations of evolutionary theory (1, 2). Random genetic changes and natural selection would seem to prevent novelty by keeping populations near genotypes at the peaks of fitness landscapes, preserving existing forms at the expense of novel mutants (3–5). Models to explain the origins of new functions often invoke gene duplication events, which create redundancy needed to allow either copy to eventually evolve toward a new function (6–9). However, the fitness landscape between old and new functions has been difficult to study largely because of the vast number of possible genetic variants for any given gene. As a result, models of innovation differ in the relative importance of neutral drift, environmental changes, the timing and type of selection pressure, and the high-dimensional nature of sequence space (10). Our understanding of innovations will benefit from direct observations of the evolution of new functions (11–16).

Macromolecular phenotypes such as enzymes can tolerate changes to their primary sequence without necessarily changing structure or function. As a consequence of this robustness to mutations, many genotypes have the same phenotype (17, 18). Natural populations of both organisms and macromolecules that appear the same phenotypically still harbor many genetic differences. Genotype networks are the collection of all genotypes with the same phenotype that are interconnected by mutational steps (19). Populations occupy finite regions of these vast networks, and it has been suggested that innovations can occur where two genotype networks are in close proximity (20) (Fig. 1.1A). To evaluate various models of molecular innovation, it is necessary to characterize the number of mutations that separate two networks and the fitness consequences of the mutational changes needed to move from one network to the other.

Here, we report an experimentally constructed intersection of two genotype networks. For our study system we have chosen two distinct RNA phenotypes. The RNA molecules are ribozymes, structured RNA molecules that catalyze chemical reactions. One ribozyme phenotype is the naturally-occurring self-cleaving HDV ribozyme. The second phenotype is the class III ligase ribozyme that was discovered through artificial selection in a lab (Fig. 1.1B) (21). The two ribozymes share no evolutionary history, catalyze different chemical reactions, and fold into very different structures. Despite the differences between the two ribozymes, it was previously shown that the two genotype networks come in close proximity, and very few mutations could convert one ribozyme into the other (22). This provides an experimentally tractable example of a molecular innovation. To characterize the fitness landscape between the two genotype networks we developed two high-throughput sequencing based assays to quantify both ribozyme phenotypes. We analyzed 16,384 neighboring sequence variants using both assays. For each sequence, we determined the *ribozyme fitness* for both activities, defined as the catalytic activity relative to a reference sequence. With these fitness values, we analyzed the billions of mutational trajectories between the two genotype networks, and used computational simulations to explore how these genotype networks facilitate or inhibit evolutionary innovations.

Results and Discussion

We obtained fitness measurements for all 16,384 RNA sequences for both RNA phenotypes. For visualization of the resulting genotype networks, we plot the data as a network graph, where each node is a unique sequence, nodes are connected if they differ by a single mutation, and the fitness is represented by the size of the node (Fig. 1.2A). Each node is colored based on the dominant activity, with HDV in red and Ligase in blue. Fitness values were normalized such that *fitness* = 1 for the reference ribozyme, previously referred to as the “prototype” (22). This representation of the data allows a visual appraisal of the proximity of the two genotype networks. In general, both networks are characterized by a decrease in fitness with distance from the reference. The region where the two networks are in closest proximity contains sequences with low activity for either function. Still, we find that numerous genotypes in the two networks are proximal, and numerous distance measurements are required to characterize the mutational distance between the networks.

To quantify the average distance between the two genotype networks, we measured the distance between every genotype on one network and the nearest genotype on the other network (Fig. 1.2B). We find that this distance depends upon whether or not a lower bound is set for genotypes to be considered a member of the genotype network. We find that the average distance between the networks decreases as the fitness cut-off is lowered (Fig. 1.2B). For example, if “wild-type” activity is required (*fitness* > 1), the two networks are separated by ~7 mutations on average. However, if molecules with 10% of wild-type activity or better are considered part of the network, then most genotypes are only 1-2 mutations from the other network.

Surprisingly, if we do not set any fitness cut-off, and count all genotypes as being a part of a network as long as they were detected as catalytically active in all three replicates of our assay, we find that over half the molecules (9,032) can actually perform both functions (Fig 2C). Most of these dual-function intersection sequences have very low fitness for both functions, and not surprisingly, no single sequence had higher than wild-type fitness for both functions ($\log_{10}(\textit{fitness}) > 0$). However, several sequences do show detectable levels of activity for one function and higher than wild-type fitness for the other function. Under many evolutionary scenarios, these genotypes would be the most likely to facilitate a molecular innovation because they would be favored if selection was acting on only one function, yet would already provide the new function as a suboptimal promiscuous function (23, 24). These results demonstrate that the genotype networks have substantial overlap with numerous intersection sequences.

Next, we set out to evaluate the implications of these genotype networks for the evolution of molecular innovations. The networks are in fact high-dimensional, which limits any intuitive interpretation. We therefore turned to computational simulations of populations of RNA molecules evolving on the networks. We modeled evolution using a Wright-Fisher model (25) with a fixed population size, a fixed mutation rate, and selection determined by the differences in the fitness of neighboring genotypes (see Materials and Methods). To simulate evolutionary innovations, we imagined the naturally occurring HDV genotype as the established function and the *in vitro* selected Ligase activity as the “new” function. We modeled a situation where the enzymatic function of the HDV ribozyme is first under selection, but gene duplication allows a copy of the gene to evolve under selection for Ligase activity. We therefore apply immediate selection pressure using the Ligase fitness measurements, with no further consequence for the changes in HDV activity. For these simulations, it is useful to think of the genotype networks as a three-dimensional fitness landscape, where the height of the landscape is determined by the fitness (Fig. 1.3A and Movie S1). Evolving populations will tend to move uphill towards the peaks in such a landscape. We started multiple simulations from different genotypes on the HDV network and challenged the populations to evolve on the Ligase fitness landscape. We recorded these simulations as movies to observe the process of evolution toward the *new* Ligase function (Fig. 1.3B and Movie S2-S5).

We noticed that many of the individual simulations had periods where the population plateaus at a specific, often low average fitness for many generations (Fig. 1.3B). To evaluate the average contribution of these periods of stasis, we measured the average fitness of the evolving population over time (Fig. 1.4A and Fig. S1.1) and did so for 100 replicate simulations from each of the different starting genotypes (Fig. 1.4B). We find that different genotypes on the HDV network result in different average rates of adaptation to the *new* Ligase function (Fig. 1.4C). The fact that some genotypes promote very rapid adaptation supports the idea that *neutral* evolution that enables a population to explore a genotype network can facilitate evolutionary innovations (20, 26).

Additionally, we find that there exist specific genotypes on the Ligase fitness landscape that cause these periods of stasis and slower average rates of adaptation (Fig. 1.4D and Fig. S1.2-S3). These genotypes are characterized by very few pathways to higher fitness. Importantly, the genotypes that cause the slowest adaptation are characterized by extensive reciprocal sign epistasis, meaning that achieving higher fitness requires two or more mutational steps, but *every* initial step is deleterious. These genotypes are local fitness peaks with not a single beneficial one-mutation-neighbor in our data set. Different starting genotypes on the HDV network frequently stall at the same intermediate fitness level indicating that they are likely to encounter a specific stasis genotype. These results are encouraging for efforts aimed at forecasting evolutionary outcomes in cases where the underlying fitness landscape can be measured or accurately estimated (27, 28).

Our results show that at regions of genotype space where two phenotypes intersect, there exist numerous evolutionary trajectories between functions. We demonstrate that this region enables rapid evolution of innovation. Mutational walks that maintain one function while approaching a new function are abundant, and dual-function sequences permeate this region of sequence space. The decrease in the fitness of both functions at this interface suggests that intermediate forms are disfavored over the sequences that can do one function well (10). The evolution of innovation in this sequence space is not only possible, but probable. However, it remains unknown whether these characteristics are peculiar to these specific phenotypes. Further research advancements will be required to understand how functional intersections change over larger expanses of genotype space, and if historic evolutionary innovations found in natural systems have properties like the model system studied here. The high probability of finding a dual-function sequence at this intersection encourages the search for more genotype network intersections and motivates future research on the forecasting of evolutionary innovations.

Materials and Methods

Library Design

For our experiments, we first identified an HDV and a Ligase reference sequence (Fig. 1.2). For this purpose, we chose sequence variants that were expected to have near wild-type ribozyme fitness and that were 14 mutations apart (29). We then set out to construct a library of ribozyme sequences that contained all the possible presence-absence combinations of these 14 nucleotide differences. These sequence variants represent all the parsimonious intermediates on the evolutionary trajectories between the two reference sequences. Library construction was accomplished by chemically synthesizing a degenerate DNA oligonucleotide that would serve as a template for in vitro transcription with T7 RNA polymerase. At each position where the Ligase and HDV reference ribozymes differed, the synthesis used equal mixtures of two nucleotide phosphoramidites, generating approximately equal probability of both sequence variants. This creates 2^{14} = 16,384 ribozyme variants. We synthesized two such libraries, one “HDV-library” with a 5’-leader sequence that is cleaved by variants with the HDV phenotype, and a second “Ligase-library” that begins at the 5’-end of the Ligase ribozyme, so that variants with the Ligase phenotype could react with a separate substrate oligonucleotide (30). A common sequence was added to the 3’-end of both libraries to serve as a universal primer binding site for reverse transcription (31). Oligonucleotides used in this experiment are listed in Table S2.1.

Co-transcriptional Cleavage Assay

The sample preparation was done entirely in triplicate yielding three biological replicates. The ssDNA ultramer cleavage library used for in vitro transcription of the ribozyme mutants was annealed to the T7-TOP+ primer. 20 picomoles each of DNA template and primer were heated for 5 mins at 98°C in 10 μL final volume of custom T7 Mg10 buffer (500 μL 1M Tris pH 7.5, 50 μL 1M DTT, 20 μL 1M Spermidine, 100 μL 1M $MgCl_2$, 330 μL RNase-free water). The template and primer were then diluted 10-fold and cooled to room temperature. 2 μL of template and primer were then transcribed in vitro in a 50 μL reaction with 5 μL T7 Mg10 buffer, 1 μL rNTP (25 mM, NEB), 1 μL T7 RNA polymerase (200 units, Thermo Scientific) and 41 μL RNase free water (Ambion) at 37°C for 20 mins. The transcription was then terminated by adding 15 μL of 50 mM EDTA. Although the total amount of cleaved RNA increases during transcription, the ratio of cleaved to uncleaved remains the same, as long as the rate of transcription is constant, which is true for moderately short transcription times before reagents become limited (32). 20 mins was determined to be the optimal time for transcription by transcribing the library at multiple time points and measuring RNA levels using denaturing PAGE. 20 mins was selected as optimal because it was still during linear growth before reaching a plateau. The transcription reaction was then cleaned and concentrated with Direct-zol RNA MicroPrep w/ TRI-Reagent (Zymo Research) to 7 μL.

The concentration of the RNA sample was then determined using a spectrophotometer (ThermoFisher NanoDrop) and the samples were normalized to 5 μM. The transcribed and cleaned RNA (5 picomoles) was mixed with 20 picomoles of RT-library primer (Table S2.1) in a volume of 10 μL and was heated at 72 °C for 3 mins and then cooled on ice. 4 μL SMARTScribe 5x First-Strand Buffer (Clontech), 2 μL dNTP (10 mM), 2 μL DTT (20 mM), 2 μL phased template switching oligo mix (10 μM), 1 μL water and 1 μL SMARTScribe Reverse Transcriptase (10 units, Clontech) were then added to the RNA template and RT primer. The phased template switching oligo mix consisted of four oligonucleotides that were phased by the addition of 9, 12, 15 or 18 nucleotides (Table S2.1). The mixture was then incubated at 42 °C for 90 mins. The reaction was stopped and the RNA degraded by heating the sample to 72 °C for 15 mins. The cDNA was then purified using DNA Clean & Concentrator-5 (Zymo Research) and eluted into 7 μL water.

Ligation Assay

The ssDNA ultramer ligation library used for in vitro transcription of the ribozyme mutants was annealed to the T7-TOP+ primer. 20 picomoles each of DNA template and primer were heated for 5 mins at 98°C in 10 μL water. The template and primer were then transcribed in vitro in a 30 μL reaction with 12 μL rNTP (25mM, NEB), 3 μL MEGAshortscript T7 Reaction Buffer (10X, Thermo Fisher) and 3 μL MEGAshortscript T7 RNA Polymerase (Thermo Fisher) at 37 °C for 2 hours. The DNA was then degraded using 2 μL TURBO DNase (2 units/μL, Thermo Fisher) and incubating at 37 °C for 15 mins. The transcription reaction was then cleaned and concentrated with Direct-zol RNA MicroPrep w/ TRI-Reagent (Zymo Research) to 7 μL.

The concentration of the RNA sample was then determined using a spectrophotometer (ThermoFisher NanoDrop) and the samples were normalized to 5 μM. To assess the starting abundance of each genotype prior to in vitro selection, a portion of each sample was aliquoted and reverse transcribed using the template switching protocol identical to what was used for the HDV-library. The transcribed and cleaned RNA (25 picomoles) was mixed with 200mM Tris pH 7.5 in a volume of 10 μL and heated at 65 °C for 2 minutes and then cooled to room temperature. 500 picomoles of ligation substrate (Table S2.1) were then added with 4 μL $MgCl_2$ (50mM) for a total volume of 20 μL. The mixture was then incubated for 2 hours at 37 °C. To reverse transcribe the samples, 10 μL of the ligation reaction were heated with 40 picomoles of RT-library primer and heated to 72 °C for 3 mins and then cooled on ice. 4 μL SMARTScribe 5x First-Strand Buffer (Clontech), 2 μL dNTP (10 mM), 2 μL DTT (20 mM), 1 μL water and 1 μL SMARTScribe Reverse Transcriptase (10 units, Clontech) were then added to the RNA template and RT primer. The mixture was then incubated at 42 °C for 90 mins. The reaction was stopped and the RNA degraded by heating the sample to 72 °C for 15 mins. The cDNA was then purified using DNA Clean & Concentrator-5 (Zymo Research) and eluted into 10 μL water. To amplify the cDNA that had performed the ligation reaction a mix of phased selective ligation PCR primers were used. The PCR reaction consisted of 1 μL purified cDNA, 12.5 μL KAPA HiFi HotStart ReadyMix (2X, KAPA Biosystems), 2.5 μL selective ligation primer, 2.5 μL RT primer and 5 μL water. To prevent bias during the PCR amplification, multiple cycles of PCR were examined using gel electrophoresis and an appropriate PCR cycle was chosen because it was still in linear growth (Fig S4). Each PCR cycle consisted of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 30 s. The PCR cDNA product was then cleaned using DNA Clean & Concentrator-5 (Zymo Research) and eluted in 12 μL water.

Illumina Adapter PCR

In preparation for high-throughput sequencing, Illumina adapter sequences were added to the cDNA using PCR. Each of the nine samples (3 HDV, 3 ligated, 3 unligated) were each assigned a unique combination of The PCR reaction consisted of 1 μL purified cDNA, 12.5 μL KAPA HiFi HotStart ReadyMix (2X, KAPA Biosystems), 2.5 μL forward, 2.5 μL reverse primer (Illumina Nextera Index Kit) and 5 μL water. To prevent bias during the PCR amplification, multiple cycles of PCR were examined using gel electrophoresis an appropriate PCR cycle was chosen because it was still in linear growth (Fig. S1.4). Each PCR cycle consisted of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 30 s. The PCR cDNA product was then cleaned using DNA Clean & Concentrator-5 (Zymo Research) and eluted in 30 μL water. The final product was then verified using gel electrophoresis.

High-Throughput Sequencing

In preparation for high-throughput sequencing, the three cleavage replicates, three ligated replicates and three unligated replicates each with unique Illumina adapter barcodes were pooled and sent to the University of Oregon Genomics and Cell Characterization Core Facility. The samples were sequenced using Illumina NextSeq 500 Single End 150 with 25% PhiX addition. This generated ~125 million reads (Cluster PF Yield) across the nine samples.

Data Analysis

Sequencing data were analyzed using custom Python scripts. These scripts identified a universally conserved 3’ handle, determined the reacted state (ligated/ unligated or cleaved/ uncleaved) and isolated the 14 mutational nucleotides to determine genotype. This process was repeated for each replicate. A genotype was considered to be a part of the corresponding genotype network only if detected as catalytically active in all three replicates and had a catalytic rate above the uncatalyzed cleavage or ligation rate. The uncatalyzed cleavage rate is estimated to be 7e-7 min^{-1} (33). The rates of template-directed, nonenzymatic oligonucleotide ligation are estimated to 2.4e-10 min^{-1} for 2’,5’-linkage and 1.5e-8 min^{-1} for 3’,5’-linkage (34, 35). To determine the reproducibility of the sequencing, the three replicates were correlated with each other (Fig. S1.5) with high correlation coefficients. The distribution of HDV and Ligase sequencing read counts were also determined to verify sequencing quality (Fig. S1.6).

Fitness Calculations from Sequence Data

Fitness values for each genotype were determined from the sequence data. Fitness values for the HDV genotypes were calculated from the fraction of each genotype found in the cleaved form divided by the total reads of that genotype. These fraction cleaved values were normalized by dividing by the fraction cleaved for the HDV reference genotype, resulting in the *HDV fitness* values reported. The Ligase fitness was determined by the level of enrichment between the unligated and ligated samples. The relative abundance of each genotype was determined by dividing the reads corresponding to that genotype by the total number of reads in that replicate sample. The change in abundance was determined by taking the relative abundance of a specific genotype in the ligated replicate sample and dividing it by the relative abundance in the unligated sample. This value was normalized by dividing by the change in abundance for the Ligase reference sequence, resulting in the *Ligase Fitness* values reported. We observed detectable Ligase activity for all 16,384 sequences. We note that even the lowest fitness Ligase genotypes were still observed as ligated more than 4 separate times in a given replicate, and more than 32 times across all three replicates. We detected HDV activity for 9,032 of the sequences. The least frequent genotypes in our data that showed HDV activity were observed as cleaved more than once in all three replicates, and uncleaved more than 108 times. Genotypes that were not detected as cleaved in a single replicate were not considered active. This approach provides a conservative estimate for genotypes belonging to a given network.

Genotype Network and Fitness Landscape Construction

In order to visualize the highly complex network of genotypes presented in this study, a genotype network (Fig. 1.2A) and a three-dimensional fitness landscape (Fig. 1.3A) were constructed using Gephi software (36). Each node represents a unique genotype and edges connecting genotypes represent a single mutation. ForceAtlas 2 was used to approximate genotype repulsion using a Barnes-Hut calculation. The z-axis in the fitness landscape was generated using the Network Splitter 3D plugin.

Evolutionary Simulations

In order to computationally simulate the evolution of populations of RNA molecules on the Ligase genotype network, we used custom Python scripts that model evolution based on the Wright-Fisher approach[14,25]. The simulation started with 1000 individuals of the same genotype. Every generation (update) a new population of 1000 genotypes was generated in the following way. First, a parent genotype from the population was selected at random. The fitness of the genotype was compared to a randomly selected value from a fitness range (between 0 and 1). If the genotype fitness was less than the random value, the genotype was not placed in the new generation. If the genotype fitness was greater than or equal to the random value, it was placed in the new generation, with a chance of mutating at a single, randomly chosen nucleotide position. The probability of mutation was proportional to the mutation rate that was set at the beginning of the simulation ($\mu = 0.01$) and remained constant. This process was repeated until 1000 individuals were placed in the new generation. The simulation then repeated this process for 1000 generations. We repeated the simulation starting from all genotypes with HDV fitness ≥ 1 and did so for a total of 100 replicates for each genotype (n=17, Fig. S1.1). The 100 replicates for each starting genotype were averaged (Fig 4b) and the initial rate of adaptation and unique genotypes explored for each starting genotype were calculated (Fig. 1.4c). For each simulation, *initial rate* was determined by subtracting the population fitness at *generation* = 0 from the population fitness at *generation* = 200 and dividing this value by the 200 *generations*. This results in a per generation rate of initial adaptation.

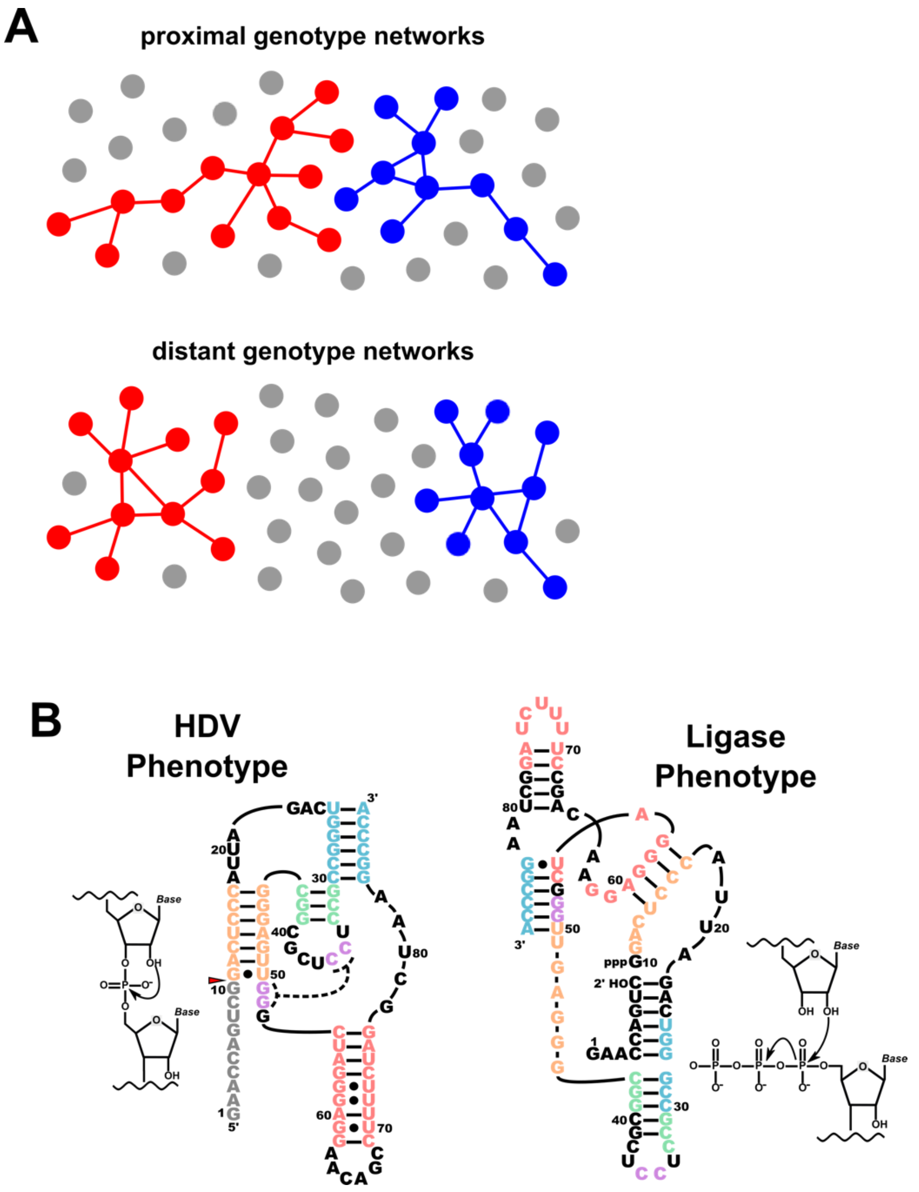


Fig. 1.1 Evolutionary innovation from a network perspective.

a, Each node represents a genotype. Genotypes with the same phenotype (genotype networks) have the same color and are interconnected by mutational steps (edges). Gray nodes are non-functional. Proximal genotype networks have neighboring genotypes with different phenotypes. Distant genotype networks have neighbors with the same function or that are non-functional. b, The two RNA phenotypes used in this study. Phenotypes are represented by the structure diagram of the antigenomic HDV ribozyme (HDV phenotype) and the class III ligase ribozyme (Ligase phenotype). Each phenotype is detected by the ability of a genotype to catalyze each specific chemical reaction that is shown beside each structure, which results in the removal (HDV) or addition (Ligase) of a short sequence (gray letters). These changes in length after a reaction can be detected in nucleotide sequence data. The two structures have the same nucleotide sequence and nucleotides are colored based on the secondary structure of the HDV phenotype.

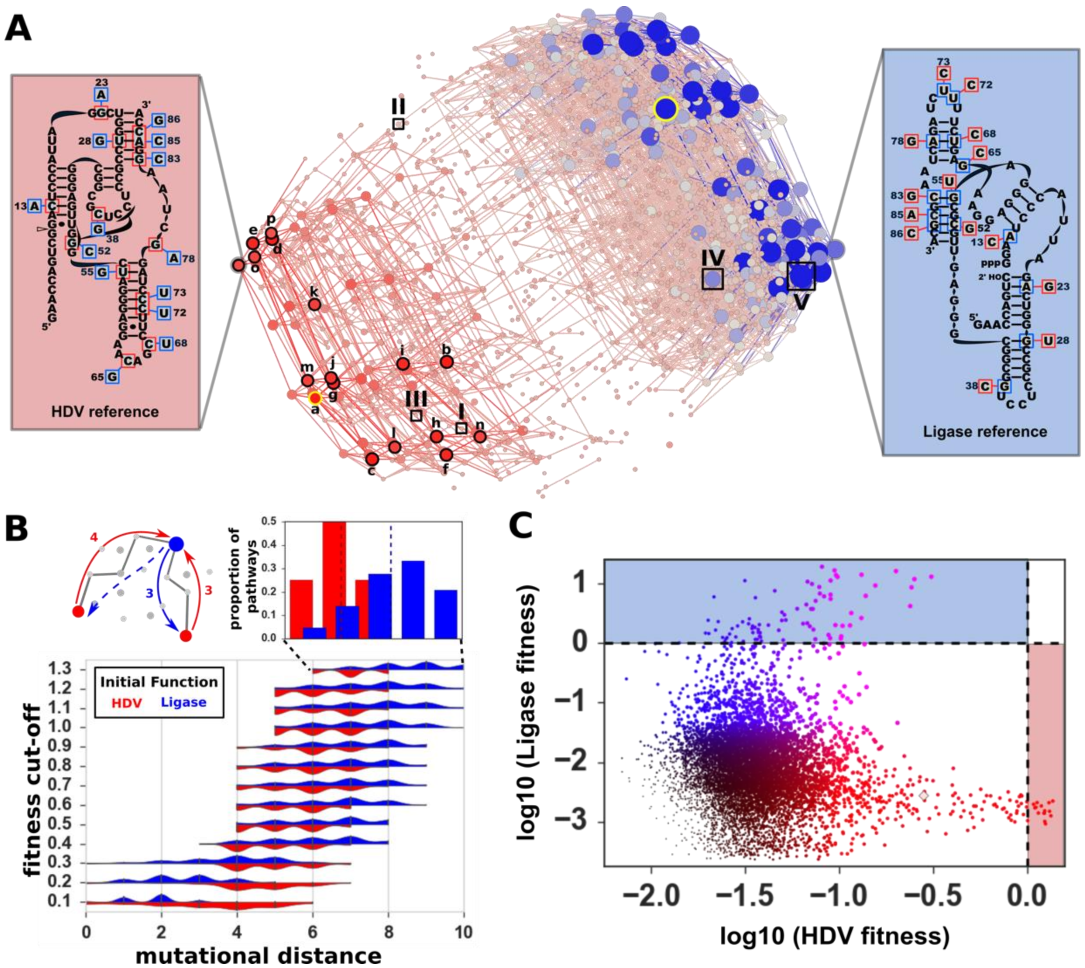


Fig. 1.2 The experimental fitness landscape at the intersection of two genotype networks.

a, The overlay of the HDV and Ligase genotype networks. Nodes represent individual sequences, and sequences are connected by an edge if they are different by a single nucleotide change. Nodes are colored based on their dominant activity (red = HDV; blue = Ligase), and the fitness is indicated by the size of the node. Boxes on the left (HDV reference) and right (Ligase reference) show the secondary structure for the reference genotypes, and all the mutational changes that were analyzed. The mutations in blue boxes convert the HDV reference to the Ligase reference. The mutations in red boxes convert the Ligase reference to the HDV reference. Genotypes used to start evolutionary simulations are indicated (a-p). Examples of stasis genotypes that were shown to impede evolution on the Ligase fitness landscape are indicated (I-IV). b, Distributions of shortest *mutational distance* between genotypes on different networks as a function of *fitness cut-off* (blue = Ligase to HDV distances; red = HDV to Ligase distances). Inset shows the distribution at *fitness cut-off* = 1.3 as histograms; dashed lines indicate the sample means. The diagram illustrates the measurement of distance between the two functions. c, Intersection sequences with detectable activity for both functions. Color indicates the ratio of ligation fitness (blue) to HDV fitness (red).

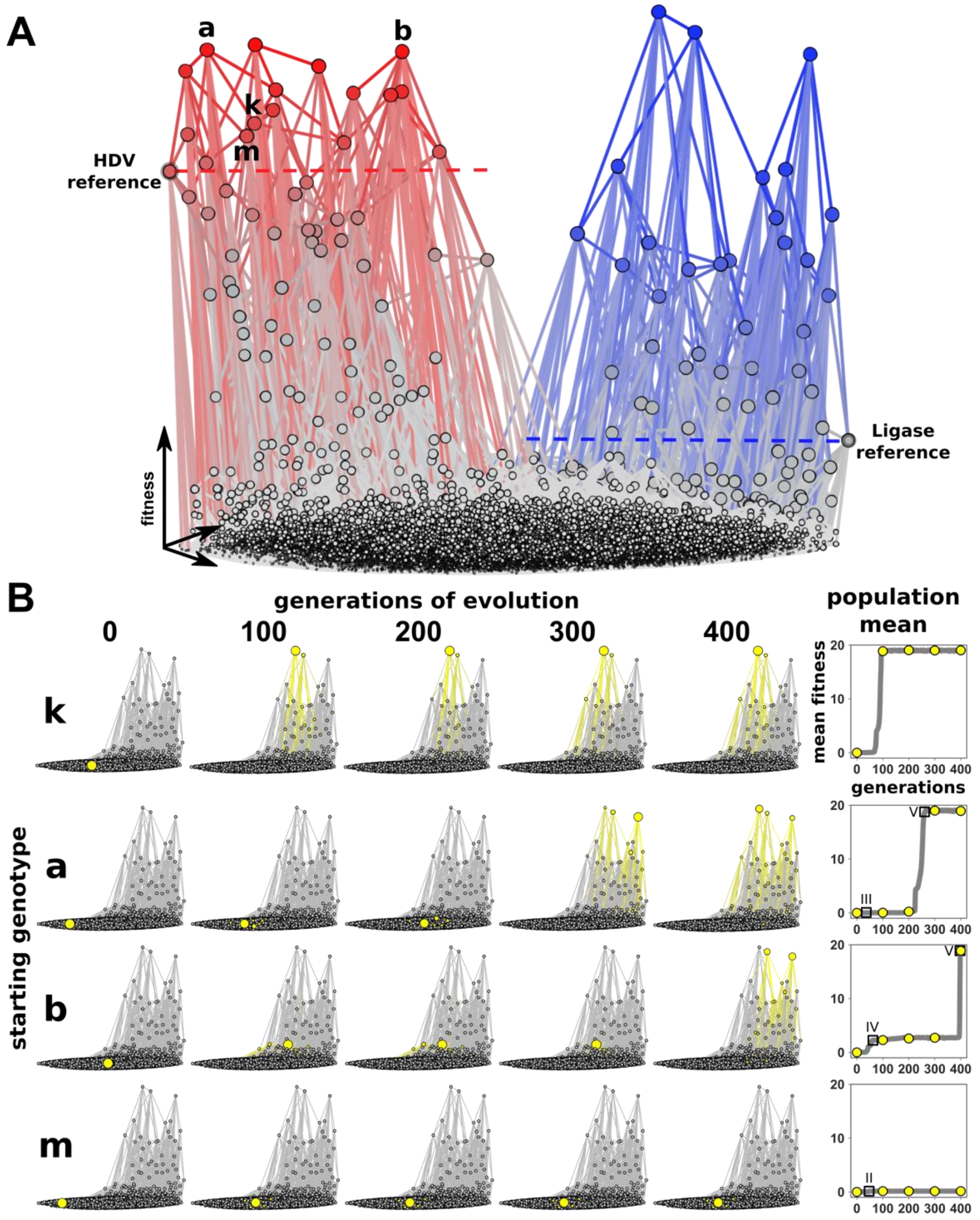


Fig. 1.3 Computational simulation of evolutionary innovation reveal periods of stasis.

a, Three-dimensional fitness landscape for both genotype networks. The height of each node indicates the relative fitness for the HDV phenotype (red) and the Ligase phenotype (blue). Fitness are normalized so that both graphs are similar heights. Nodes are connected if they are different at one nucleotide position. Starting genotypes (a,b,k,m) are indicated as examples that show different rates of Ligase adaptation. b, Frames from simulations of evolving populations. Genotypes present in the population (yellow nodes and edges) change over generation time due to mutation and selection. The corresponding mean fitness of these genotypes experience periods of stasis, followed by rapid increase in fitness. During simulations the population size (N = 1000) and mutation rate (μ = 0.01) were constant.

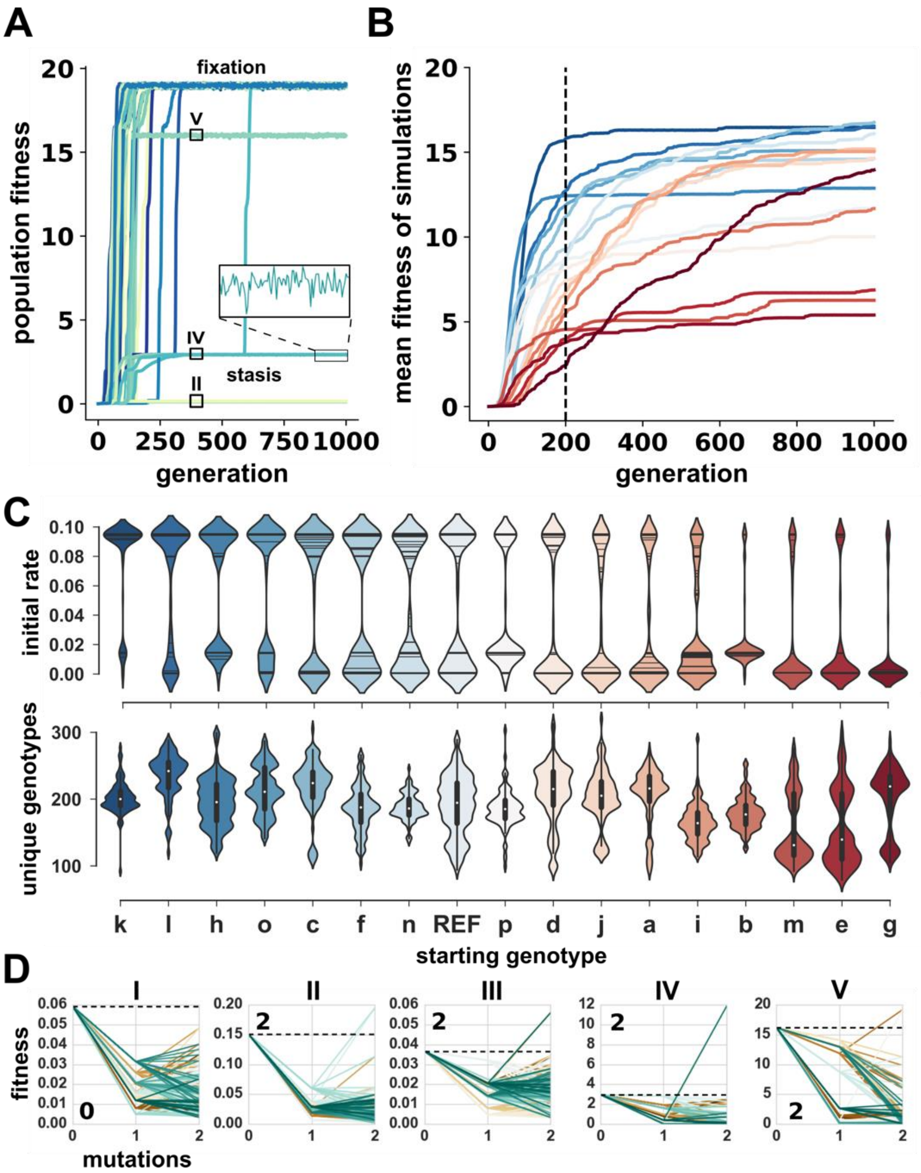


Fig. 1.4 Starting genotypes result in different rates of adaptation. a, Rates of Ligase adaptation from a single HDV genotype.

Each trace shows the average fitness as a function of generation time for a separate simulation of 1000 individuals each. Inset shows minor fluctuations during periods of stasis. b, Average rates for multiple evolutionary simulations from different starting genotypes. Each trace represents a different starting genotype and shows the mean fitness of 100 simulations such as in a plotted as a function of generation time. c, Distributions of initial rates of adaptation and unique genotypes explored during simulations. Initial rate is determined as the per generation fitness increase over the first 200 generations (see Methods). Unique genotypes represent the total number of genotypes encountered during a simulation. d, The local fitness landscape of genotypes that cause periods of stasis show sign epistasis. The fitness of the stasis genotype is plotted at mutations = 0 and marked with a dashed line. The fitness of neighboring genotypes that differ by 1 or 2 mutations are shown. The roman numeral above each graph corresponds to Fig. 1.2.

References

1. Darwin CR (1859) *On the Origin of Species by means of natural selection, or the preservation of favoured races in the struggle for life.* (John Murray, London). 1st Ed.

2. Pigliucci M (2008) What, if Anything, Is an Evolutionary Novelty? *Philosophy of Science* 75(5):887–898.

3. Tracewell CA, Arnold FH (2009) Directed enzyme evolution: climbing fitness peaks one amino acid at a time. *Current Opinion in Chemical Biology* 13(1):3–9.

4. Wright S (1988) Surfaces of selective value revisited. *The American Naturalist* 131(1):115–123.

5. Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 128(1):11–45.

6. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* 11(2):97.

7. Ohno S (1970) *Evolution by gene duplication.* (Springer-Verlag, Berlin; New York).

8. Bergthorsson U, Andersson DI, Roth JR (2007) Ohno’s dilemma: evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences* 104(43):17004–17009.

9. Zhang J (2003) Evolution by gene duplication: an update. *Trends in ecology & evolution* 18(6):292–298.

10. Pigliucci M, Kaplan J (2014) *Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology* (University of Chicago Press, Chicago, UNITED STATES) Available at:

http://ebookcentral.proquest.com/lib/boisestate/detail.action?docID=485956 [Accessed January 17, 2018].

11. Meyer JR, et al. (2012) Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. *Science* 335(6067):428–432.

12. Ross BD, et al. (2013) Stepwise evolution of essential centromere function in a Drosophila neogene. *Science* 340(6137):1211–1214.

13. Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Reviews Genetics* 8(9):675–688.

14. Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* 489(7417):513–518.

15. Näsvall J, Sun L, Roth JR, Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. *Science* 338(6105):384–387.

16. Voordeckers K, et al. (2012) Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol* 10(12):e1001446.

17. Wagner A (2011) The molecular origins of evolutionary innovations. *Trends in Genetics* 27(10):397–410.

18. Takeuchi N, Poorthuis PH, Hogeweg P (2005) Phenotypic error threshold; additivity and epistasis in RNA evolution. *BMC Evolutionary Biology* 5(1):9.

19. Wagner A (2011) Genotype networks shed light on evolutionary constraints. *Trends in Ecology & Evolution* 26(11):577–584.

20. Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* 9(12):965–974.

21. Bartel DP, Szostak JW (1993) Isolation of new ribozymes from a large pool of random sequences [see comment]. *Science* 261(5127):1411–1418.

22. Schultes EA, Bartel DP (2000) One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. *Science* 289(5478):448–452.

23. Khanal A, McLoughlin SY, Kershner JP, Copley SD (2015) Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution. *Mol Biol Evol* 32(1):100–108.

24. O'Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* 6(4):R91–R105.

25. Donnelly P, Weber N (1985) The Wright-Fisher model with temporally varying selection and population size. *J Math Biology* 22(1):21–29.

26. Hayden EJ, Ferrada E, Wagner A (2011) Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* 474(7349):92–95.

27. Lobkovsky AE, Koonin EV (2012) Replaying the tape of life: quantification of the predictability of evolution. *Front Genet* 3:246.

28. Otwinowski J, Plotkin JB (2014) Inferring fitness landscapes by regression produces biased estimates of epistasis. *PNAS* 111(22):E2301–E2309.

29. Schultes EA, Bartel DP (2000) One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. *Science* 289(5478):448–452.

30. Bartel DP, Szostak JW (1993) Isolation of new ribozymes from a large pool of random sequences. *Science* 261(5127):1411–1418.

31. Wilkinson KA, Merino EJ, Weeks KM (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc* 1(3):1610–1616.

32. Long DM, Uhlenbeck OC (1994) Kinetic characterization of intramolecular and intermolecular hammerhead RNAs with stem II deletions. *Proc Natl Acad Sci U S A* 91(15):6977–6981.

33. Li Y, Breaker RR (1999) Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2‘-Hydroxyl Group. *J Am Chem Soc* 121(23):5364–5372.

34. Rohatgi R, Bartel DP, Szostak JW (1996) Nonenzymatic, Template-Directed Ligation of Oligoribonucleotides Is Highly Regioselective for the Formation of 3‘−5‘ Phosphodiester Bonds. *J Am Chem Soc* 118(14):3340–3344.

35. Rohatgi R, Bartel DP, Szostak JW (1996) Kinetic and Mechanistic Analysis of Nonenzymatic, Template-Directed Oligoribonucleotide Ligation. *J Am Chem Soc* 118(14):3332–3339.

36. Mathieu Bastian, Heymann S, Jacomy M (2009) Gephi: An Open Source Software for Exploring and Manipulation Networks. *Third International ICWSM Conference*.

Acknowledgments

We would like to thank A. Wagner (University of Zurich) and J. Payne (ETH) for helpful discussions. Funding: This study was supported by Boise State University (Biomolecular Sciences Graduate Programs), National Science Foundation Directorate for Biological Sciences (Grant No. MCB-1413664), National Science Foundation Office of Integrative Activities (Grant No. OIA-1738865), and National Aeronautics and Space Administration (Grant No. 80NSSC17K0738).

Author contributions

Conceptualization – EJH, Methodology – DPB JC, Software – DPB BØ, Formal Analysis – EJH DPB BØ, Investigation – DPB JC, Writing (Original Draft Preparation) – EJH DPB, (Review and Editing) – EJH DPB JC BØ, Visualization – DPB EJH.

Supporting Information

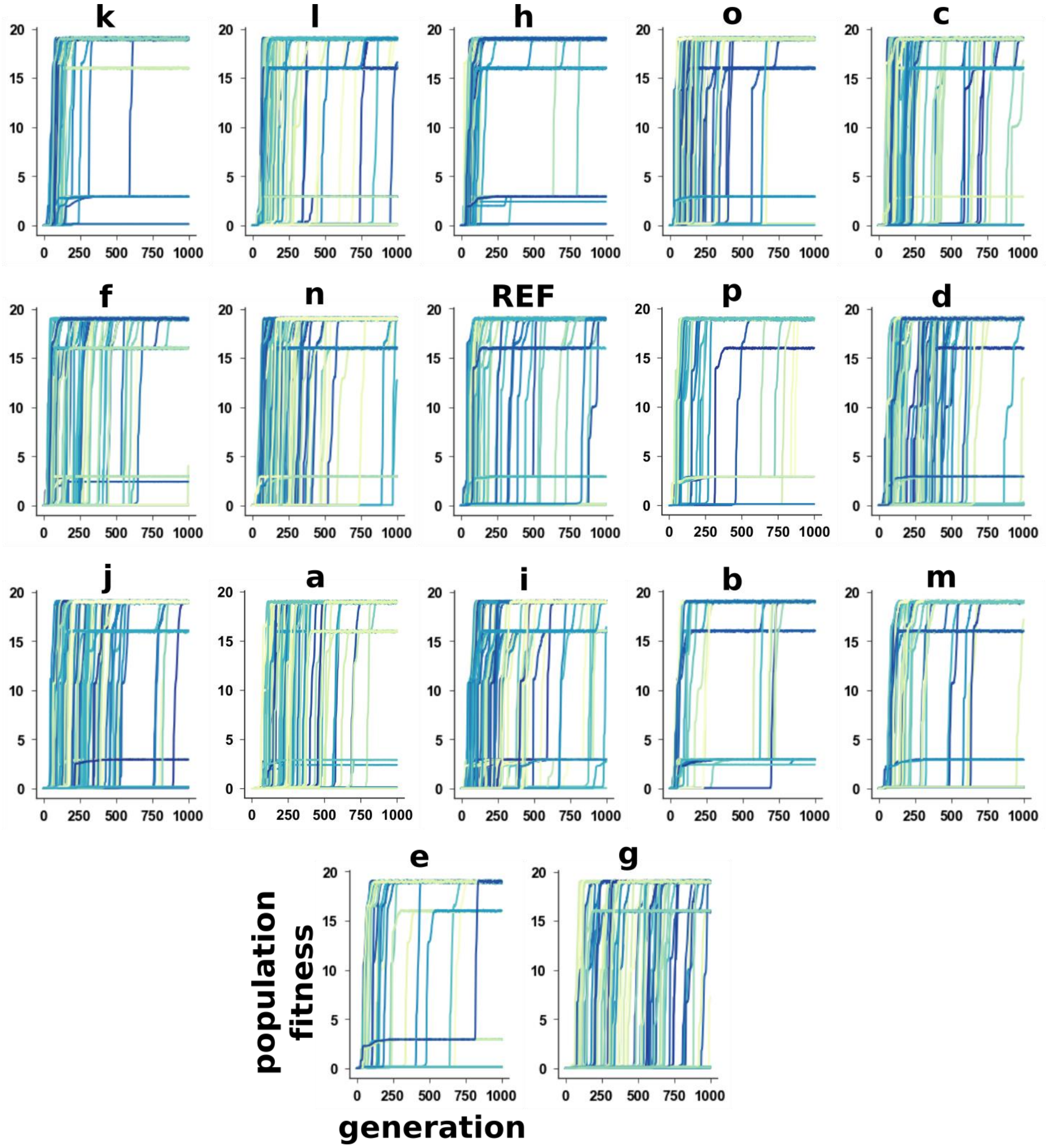


Fig. S1.1 Rate of adaptation for populations starting from different genotypes.

Each trace shows the increase in population fitness over generation time for a single simulation of 1000 individuals. Each plot shows 100 simulations starting from the same genotype. All starting genotypes has *HDV fitness* ≥ 1. The letter above each subplot indicates the starting point from the network as shown in Fig. 1.3a. Letters were assigned alphabetically based on highest to lowest HDV fitness and genotype a represents the genotype with the highest measured HDV fitness. The graphs are ordered from fastest to slowest initial rates (Fig. 1.4a).

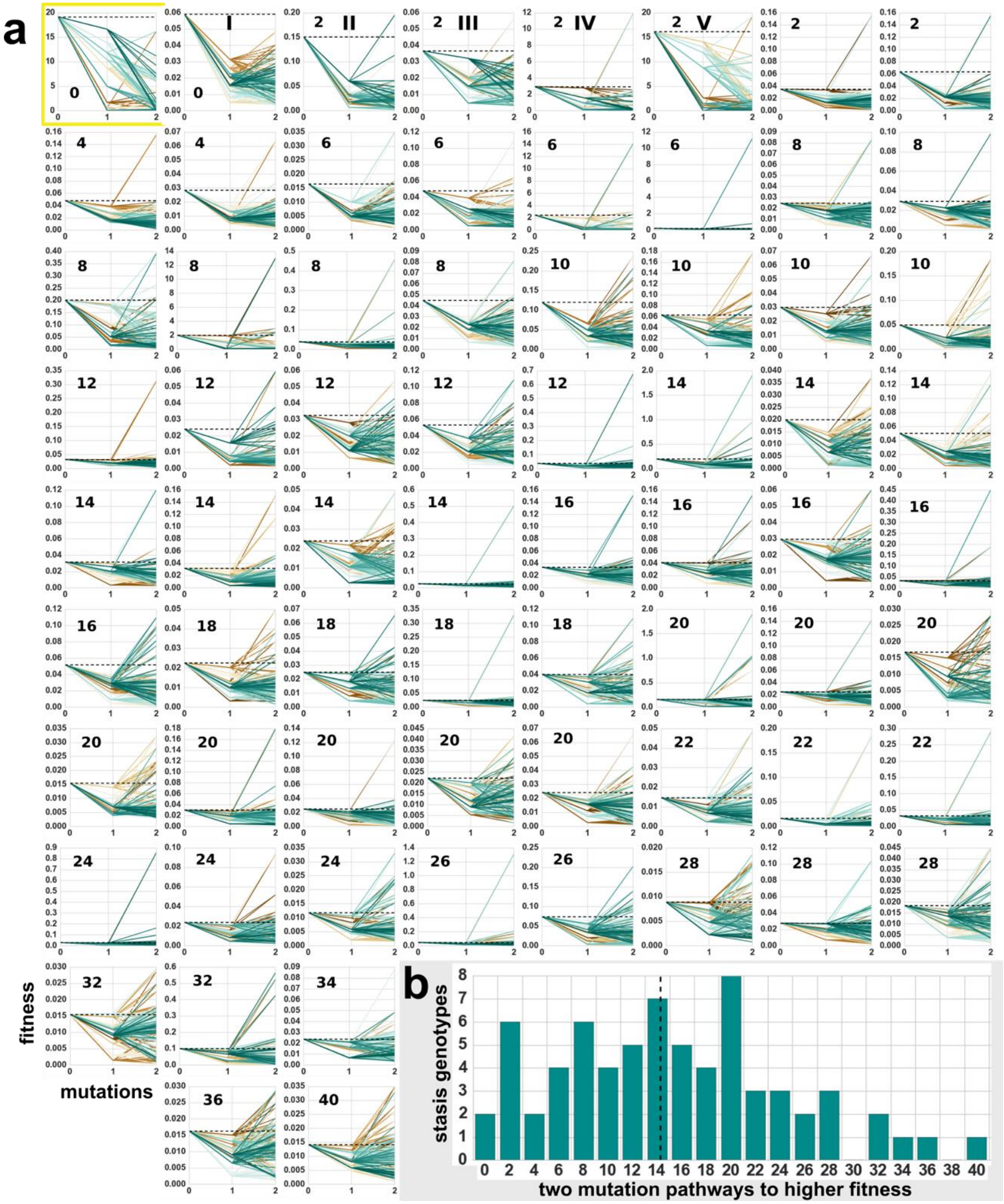


Fig. S1.2 Trajectories away from *stasis genotypes.*

a, Each line leads from the stasis genotype (*Mutations* = 0) to one and two mutations away. All 69 stasis genotypes (peaks) in the Ligase fitness landscape are depicted. The number on each graph represents the number of two mutation pathways to higher fitness from each stasis genotype. Yellow box indicates the genotype with the highest measured Ligase fitness. b, The distribution of two mutation pathways to higher fitness genotypes from each stasis genotypes in the Ligase landscape. The dotted vertical line indicates the mean of the distribution.

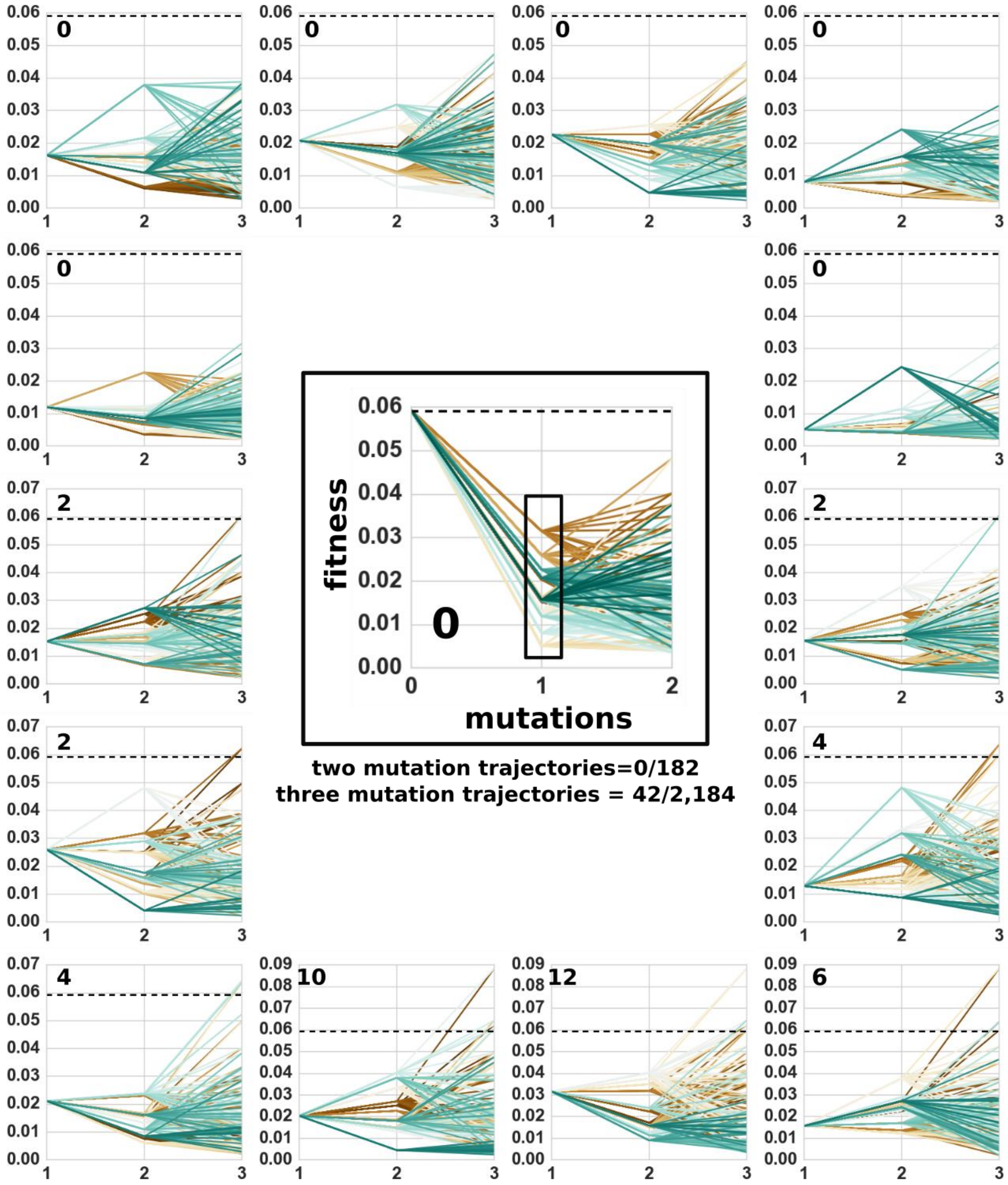


Fig. S1.3 Characterization of *stasis genotype* I.

Stasis genotype I from Fig. 1.4d is depicted in the center with each of the two mutation trajectories. None of the 182 two mutation trajectories lead to higher fitness than the stasis genotype (mutation = 0). The pathways two mutations from each of the 14 genotypes that are a single mutation away from the stasis genotype are individually depicted. In total, 42 out of a possible 2,184 three mutation trajectories yield a higher fitness than the initial stasis genotype (dashed line).

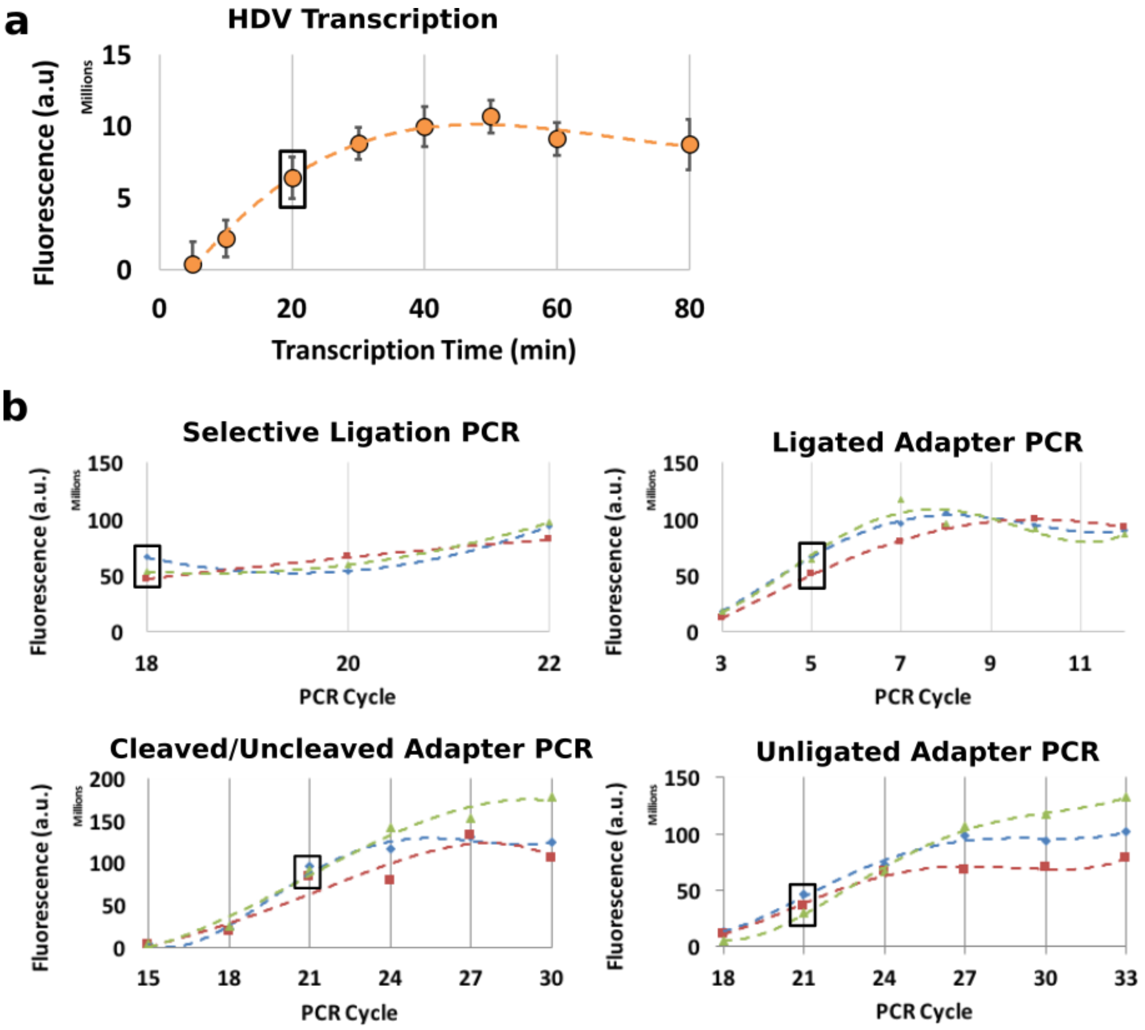


Fig. S1.4 Time-course PCR for sample optimization.

a, Time-course transcription for total RNA yield using the developed co-transcriptional cleavage assay. Data points indicate the mean RNA yield of five replicates. Error bars are standard error of the mean. Samples were run on 10% denaturing polyacrylamide gel, visualized with GelRed (Biotium), and quantified by densitometry. The time chosen as optimal (20 mins) is indicated with a box. b, Time-course PCR was performed for the selective ligation PCR and each Illumina adapter PCR for each replicate (blue, green, red). Samples were run on 2% agarose gel, visualized with GelRed (Biotium), and quantified by densitometry. The black box indicates the PCR cycle that was determined to be optimal for each PCR reaction.

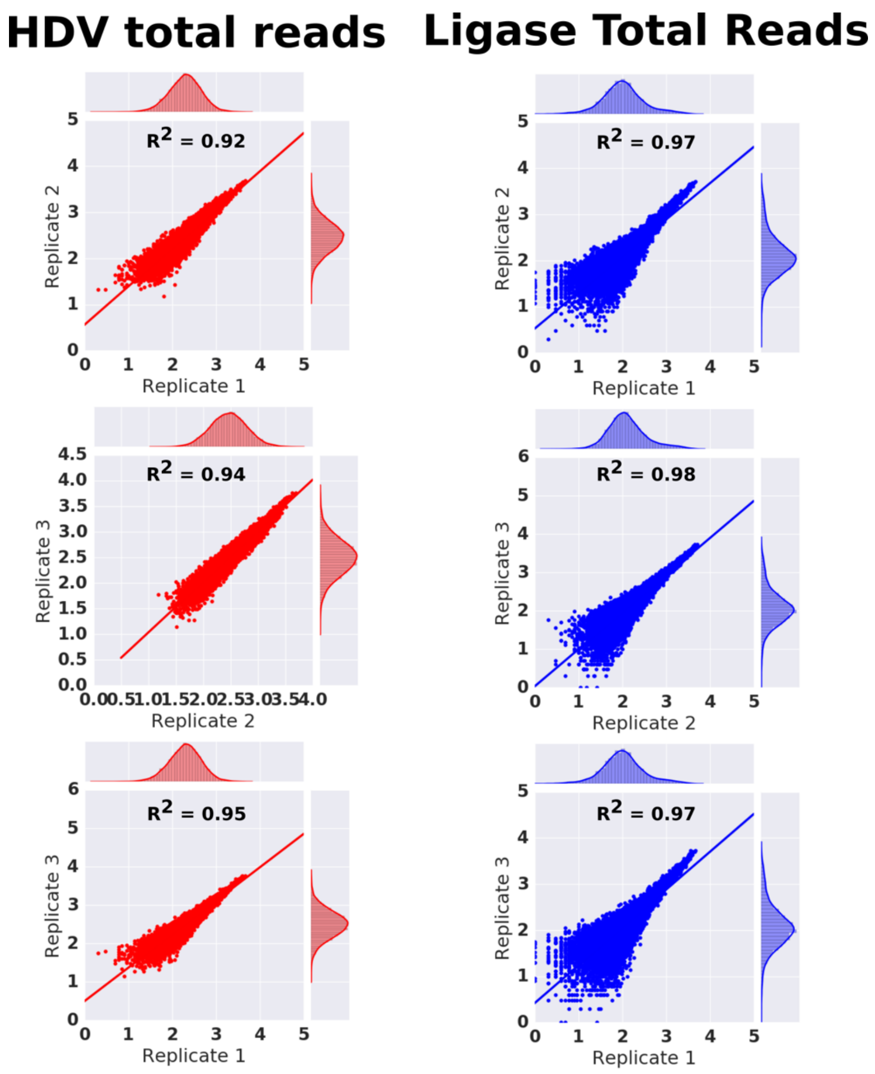


Fig. S1.5 Correlation of high-throughput sequencing replicates.

Correlation of total HDV and Ligase reads for each of the three replicates. Each figure consists of all 16,384 genotypes presented in this study. Each data point represents the frequency that a specific sequence was observed in a particular replicate (y-axis) vs. another replicate (x-axis). Sequence kernel density estimation is also reported from each replicate in the jointplot (Seaborn). The number of reads on the x and y-axis are log10 transformed.

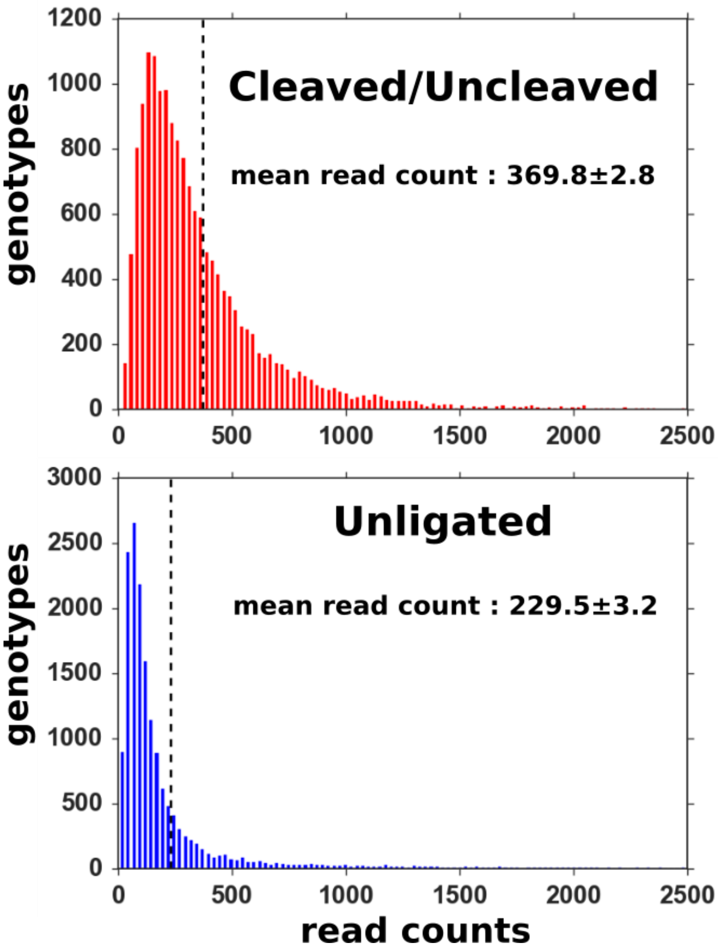


Fig. S1.6 Distribution of sequencing read counts.

Histograms indicating the average read counts for each individual genotype for the HDV and Ligase samples. The mean read count for each genotype in HDV and Ligase replicates was 369 and 230, respectively.

Table S1.1 Oligonucleotides used in this study.

Name	Sequence (5' - 3')	Notes
HDV-library	GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**TGSKCSTTAGYCTAGRRAAGRCTSTTCCTCCCTMGCSCAACTCCCGCCGCSAGGAGGCGGMCCAGTCTAATGGGAKTC**GAATGGTCCTATAGTGAGTCGTATTAGCCG	HDV template oligonucleotide. Ribozyme sequence is bolded. Cleaved sequence is in red.
Ligase-library	GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**TGSKCSTTAGYCTAGRRAAGRCTSTTCCTCCCTMGCSCAACTCCCGCCGCSAGGAGGCGGMCCAGTCTAATGGGAKTC**CTATAGTGAGTCGTATTAGCCG	Ligase template oligonucleotide. Ribozyme sequence is bolded.
T7-TOP+ primer	CGGCTAATACGACTCACTATAG	T7 transcription primer.
RT-library primer	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAACCGGACCGAAGCCCG	Reverse transcription primer
Phased TSO 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GCATGCATGCATGCATGC**rGrGrG	Phased template switching oligonucleotides. Phased insert is bolded. rG indicates RNA bases
Phased TSO 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**TGCATGCATGCATGC**rGrGrG	
Phased TSO 3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**ATGCATGCATGC**rGrGrG	
Phased TSO 4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CATGCATGC**rGrGrG	
LIG-substrate	AAGCATCTAAGCATCTCAAGCrArArArCrCrArGrUrC	Substrate for ligation reaction. rN indicates RNA bases.
Phased LIG primer 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GCATGCATGCATGCATGC**AAGCATCTAAGCATCTCAAGCAAACCAG	Phased selective ligation primers. Phased insert is bolded. rG indicates RNA bases
Phased LIG primer 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**TGCATGCATGCATGC**AAGCATCTAAGCATCTCAAGCAAACCAG	
Phased LIG primer 3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**ATGCATGCATGC**AAGCATCTAAGCATCTCAAGCAAACCAG	
Phased LIG primer 4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CATGCATGC**AAGCATCTAAGCATCTCAAGCAAACCAG	

Table S1.2. Starting Genotypes (HDV fitness ≥ 1) used in Evolution Simulations.

Genotypes are represented by the unique combination of nucleotides in the 14 variable positions of the library. Binary Genotypes represent the nucleotides of the reference HDV genotype as “0”, and a nucleotide change as “1”.

Starting Point	Genotype	Binary Genotype	HDV Fitness
REF	**CGUCGUCCCCGGAC**	00000000000000	1
a	**CGUCGUGUCCGGAC**	00000011000000	1.37
b	**CAGCGUGUUCGGCC**	01100011100010	1.34
c	**CGUCGUGUCUGGAC**	00000011010000	1.31
d	**CAUCGUGCCCGGAC**	01000010000000	1.3
e	**CAUCGUCCCCGGAC**	01100001100010	1.25
f	**CGGCGUGUCUGGCC**	00100011010010	1.24
g	**CAUCGUGUCCGGAC**	01000011000000	1.24
h	**CAGCGUCUCUGGCC**	01100001010010	1.23
i	**CAGCGUCUUCGGCC**	01100001100010	1.23
j	**CAUCGUCUUCGGAC**	01000001100000	1.19
k	**CAUCGUGCCUGGAC**	01000010010000	1.15
l	**CAUCGUGUCUGGAC**	01000011010000	1.13
m	**CAUCGUCUCCGGAC**	01000001000000	1.11
n	**CAGCGUGUCUGGCC**	01100011010010	1.1
o	**CGUCGUCCUCGGAC**	00000000100000	1.08
p	**CAUCGUCCUCGGAC**	01000000100000	1

CHAPTER TWO: EVOLUTIONARY CONSTRAINT FROM HIGHER DIMENSIONALITY IN AN RNA FITNESS LANDSCAPE

Devin P. Bendixsen[1], James Collet[2], Eric J. Hayden[1,2]
[1]Biomolecular Sciences Graduate Programs, Boise State University, Boise, ID, USA.
[2]Department of Biological Science, Boise State University, Boise, ID, USA.

Abstract

Fitness landscapes are a useful metaphor to understand and visualize the highly complex nature of sequence space. Sequence space is comprised of all possible genotypes and the phenotype corresponding to each genotype is often measured in terms of fitness. Fitness landscapes contain areas of higher fitness (peaks) which are generally separated from each other by areas of lower fitness (valleys). Due to difficulty in comprehensively characterizing empirical fitness landscapes, one important aspect that has yet to be fully understood is how does higher dimensionality affect evolution? Here, we study the evolutionary consequences of altering the dimensionality of an RNA enzyme (ribozyme) fitness landscape. We found that indirect pathways can contribute significantly to the ruggedness and navigability of a local fitness landscape. Our study suggests that increased dimensionality in RNA fitness landscapes has the potential to circumvent fitness valleys, however indirect pathways might also harbor stasis genotypes isolated by reciprocal sign epistasis.

Introduction

Fitness landscapes are a graphical representation of the mapping from genotype and phenotype (1,2). They have served as a tool that allows tangible glimpses into the highly complex nature of evolutionarily available sequence space. Sequence space is comprised of all possible genotypes and the phenotype is often measured in terms of fitness. For whole genomes the fitness can easily be derived as the probability of survival and reproduction. For catalytic biomolecules, such as enzymes or ribozymes (RNA enzymes), genotype fitness is often measured in terms of biochemical rate or activity. Fitness landscapes inherently contain areas of higher fitness (peaks) separated from each other by areas of lower fitness (valleys). Adaptive evolution combined with random mutations leads populations to explore these landscapes and potentially reach fitness peaks. These adaptive walks through evolutionary trajectories are dependent on the complexity, dimensionality and ruggedness of the fitness landscape (3–5).

It is convenient to visualize fitness landscapes as three dimensional ‘surfaces’, however they are in fact high-dimensional objects. The complexity and dimensionality of landscapes grows significantly with the length (l) of the genome or biomolecule. For this study, we defined complexity as the size of the genotype space and dimensionality as the number of edges connecting each genotype (node) to other genotypes that differ by a single mutation. For example, for RNA with four possible nucleotides at each position, the complexity of genotype space is equal to 4^l and dimensionality is equal to $3l$. With advances in both sequencing and DNA manipulation capabilities, we can synthesize and characterize ever expanding regions of sequence space (6–9). However, there still exists limitations on the amount of sequence space that can be readily and accurately characterized in a given study (2,6,10). For RNA molecules greater than ~25 in length, the high complexity and dimensionality preclude a comprehensive characterization of genotype space. Due to the uncomprehensive nature of empirical fitness landscapes, one important aspect that has yet to be fully understood is how does higher dimensionality affect evolution?

Theory has predicted that fitness valleys may become unimportant once higher dimensional spaces are considered because evolution will find indirect pathways to higher fitness that go through nearly neutral or incrementally beneficial mutational steps (11,12). In agreement with this, a recent study found that indirect pathways facilitate adaptation in a protein fitness landscape (13). Direct pathways are governed by the law of parsimony (14,15) and imply that the shortest mutational path between two genotypes is the most probable. However, direct, equally parsimonious pathways have been shown to differ significantly in their evolutionary likelihood, suggesting that parsimony alone is not always an evolutionary determinant (16). Direct pathways in the protein landscape were shown to be constrained by pairwise epistasis, however indirect pathways allowed for escape from epistatic traps (13). Epistasis or non-additive mutational interactions are a driving force in the structure of fitness landscapes and the accessibility of mutational pathways (17–22). Pairwise epistasis can be categorized into three increasingly severe classes: magnitude epistasis, simple sign epistasis and reciprocal sign epistasis (23,24). Simple sign epistasis and reciprocal sign epistasis cause the ruggedness of a fitness landscape, which directly affects its navigability by evolutionary exploration (23).

Here, we study the evolutionary consequences of altering the dimensionality of an RNA fitness landscape (Fig. 2.2.1a). Specifically, we distinguish between a ‘direct’ landscape that includes only two possible nucleotides at various positions (2^n), and an 'indirect' landscape that has all four possible nucleotides at the same positions (4^n). We chose the synthetic class III ligase ribozyme as our model system (25). In concept, indirect pathways in an RNA fitness landscape could facilitate adaptation by allowing for the escape from epistatic traps (as seen in protein) or could hinder adaptation by introducing more epistasis and creating additional evolutionary traps. We also sought to determine if the effects of higher dimensionality were the same for all regions of genotype space. We therefore identified two diverse regions of genotype space to position mutational libraries, each containing a direct and indirect landscape. Each mutational library started from a single reference genotype that had previously been shown to be at the intersection of two phenotypes (self-ligating and self-cleaving) and therefore could be considered a theoretical ancestral sequence (26). One region of sequence space was expected to contain a high-fitness peak and was therefore termed the *Peak* library. The second region was designed to occupy a region of sequence space with relatively low fitness genotypes and was therefore termed the *Valley* library. Each library encompassed 7 mutational positions, such that each direct landscape contained 128 (2^7) unique genotypes and each indirect landscape contained 16,384 (4^7) genotypes. We assigned a ribozyme fitness to each genotype using a previously developed high-throughput sequencing based assay. In this assay, we define fitness as the enrichment rate of a ribozyme sequence during a single round of in vitro selection. This enrichment rate is dependent upon ribozyme activity and is calculated from how much more (or less) frequent a genotype is after a round of selection, analogous to the growth rate of competing microbial genotypes. We characterized each empirically derived fitness landscape and used computational simulations of evolving populations to assess the effects of indirect pathways on evolutionary adaptation in these RNA fitness landscapes.

Results

Direct vs Indirect Pathway Landscapes in The *Peak* Library

We obtained relative ligase fitness measurements for all 16,384 unique genotypes in the *Peak* library (Materials and Methods). We isolated the 128 genotypes that encompass the direct pathways present between the two anchor genotypes: the reference genotype and the genotype 7 mutations in the direction of the 'wild-type' ligase ribozyme (Fig. 2.2.1b). We used network graphs to represent both two-dimensional (Fig. 2.2.1c) and three-dimensional (Fig. 2.2a) representations of the fitness landscapes for both the direct (2^n) and indirect (4^n) pathway landscapes. Each node in the visualization indicates a single genotype and edges connect genotypes different by a single mutation. The size and color saturation of the node is representative of the relative fitness of that genotype, with high fitness values represented by large size and fully saturated color. The z-axis in the three-dimensional fitness landscapes represents the ribozyme fitness (Fig. 2.2a). Nodes that are found in the direct pathway landscape are on a grey-orange-red color scale, while those exclusively on the indirect pathway landscape are on a grey-green-blue scale. There are 128 genotypes in the direct landscape and 16,384 genotypes in the indirect landscape making the indirect landscape 128 times more complex. Similarly, the indirect landscape has a three-fold increase in dimensionality. Each genotype in the direct landscape is connected to 7 other genotypes, whereas in the indirect landscape the number of connections increased to 21.

The direct landscape contained of two fitness peaks with a total of 9 genotypes with fitness ≥0.1 (Figure 2.1-figure supplement 1). A fitness peak is defined as a genotype where every connecting genotype has lower fitness and therefore represents a downward step in the landscape. The indirect landscape had a total of 20 peaks and 464 genotypes with fitness ≥0.1. The highest fitness genotype, or *summit* of the direct landscape had a relative ligation fitness of 2.47. Higher dimensionality in the indirect landscape introduced a new summit with a fitness of 3.11. The change in the summit value and genotype might have significant impacts on the evolutionary adaptation. Starting from each genotype in the two landscapes, we were able to computationally determine all accessible mutational paths to the summit of the respective landscape (23,27). A path was deemed accessible to the summit if the genotypes on the path had increasingly higher fitness from the start genotype to the summit genotype. Therefore, each mutation on the path is an upward step in fitness and downward steps are not allowed. We found that the direct and indirect pathways allowed for a proportionally equal number of genotypes within the landscape to have mutational access to the summit (67% for direct, 69% for indirect). Therefore, although the indirect pathway landscape has significantly more connections and results in billions of unique pathways (as compared to ~5,000 in the direct landscape), the new pathways still result in ~30% of genotypes being isolated from the summit by downward steps in fitness.

Higher Dimensionality Introduced More Severe Pairwise Epistasis in The *Peak* Library

Pairwise epistasis was found to be prevalent in both the direct and indirect landscapes in the *Peak* library. Two precise mutations can occur in either order, and are represented in our landscapes by subgraphs of four connected genotypes that we refer to as “squares”. Analysis of the squares revealed that the distributions of epistasis severity in the direct and indirect landscapes were very similar, with the majority of epistatic interactions being of relatively low severity (Fig. 2.2b). The indirect landscape showed a proportionally higher amount of interactions with magnitudes less than 0.5 (~90% as compared to 79%) and the direct landscape had a higher proportion of interactions in the range of 0.5 to 1.5 (~18% as compared to 9%). However, a striking difference between the two distributions is the maximum magnitude of epistasis encountered. In the direct landscape the highest magnitude encountered was ~2.8, however in the indirect landscape there are more than 750 squares that exhibited higher epistasis than this and the maximum encountered was 3.9.

The squares in the direct pathway landscape exhibited significantly higher amounts of pairwise epistasis. In fact, 91% of squares in the direct landscape exhibited epistasis, as compared to 63% in the indirect landscape (Table 2.1, Figure 2.2-supplement figure 1). The pairwise epistatic interactions were categorized into three classes: magnitude epistasis, simple sign epistasis, and reciprocal sign epistasis. 62% of mutation pairs in the direct landscape exhibited magnitude epistasis as compared to only 52% in the indirect landscape. Simple sign epistasis was found to be three times as prevalent in the direct landscape (~24%) compared to the indirect landscape (~8%). Likewise, reciprocal sign epistasis was twice as high in the direct landscape (~6%) as compared to the indirect landscape (~3%). We found that given the high prevalence of pairwise epistasis in the direct landscape, it was more rugged (0.35) as compared to the indirect pathway landscape (0.14). Therefore, although epistatic interactions may be more prevalent in the direct landscape making it more rugged, as compared to the indirect, the severity of epistasis increases with higher dimensionality in the indirect landscape.

Reciprocal Sign Epistasis in a Structurally Important Base Pair

The base-pairing within RNA structures is expected to be a common source of epistatic interactions (28). As an illustrative example, we will describe the consequence of breaking and reforming a base pair through direct and indirect pathways in our data. First, reciprocal sign epistasis was observed between peak 1 and peak 2 in the direct pathway landscape. The two peaks are isolated from each other by two mutations that break and then repair a canonical Watson-Crick base pair. *Peak 1* utilizes a C-G base pair in the terminal 3' stem at positions 86 and 52, while peak 2 utilizes a G-C base pair at these positions (Fig. 2.2c). The two direct intermediates between the peaks break the base pair interaction and cause a decrease in relative fitness (Fig. 2.2d). Using indirect pathways, the number of genotypes between peak 1 and peak 2 increases from 2 to 14 and the number of unique pathways increases from 2! = 2 to 4! = 24. The indirect pathways contain four new genotypes that are greater than the fitness of *peak 1* (Fig. 2.2d). The four new high-fitness genotypes contain either a canonical A-U base pair or a G-U wobble base pair and therefore retain the stem structure. The four genotypes appear to form a bridge between the two isolated peaks, however it should be noted that a downward step in fitness is still required in order to reach peak 2. The four genotypes in the direct landscape form a single mutation pair or square and the magnitude of epistasis within this square was calculated to be ~2.7 (Fig. 2.2e). The addition of 12 new genotypes available on the indirect pathways resulted in 71 new mutation pairs with a wide range of epistatic values all of which are less severe than in the direct landscape.

Therefore, the downward step that is required to reach peak 2 does not have to be as detrimental in the indirect landscape as compared to the direct landscape. This premise is emphasized in the local fitness landscape of peak 1 (Fig. 2.2f). Using direct pathways there exist only a pair of two-mutation pathways to higher fitness, as compared to 7 in the indirect landscape.

However, it is important to note that not all base pair interactions exhibited predictable reciprocal sign epistasis that could be alleviated by indirect pathways. Using a different base pair on the terminal 3' stem of the ribozyme at position 83 and 55 we see that the direct pathways exhibits only simple sign epistasis and not reciprocal sign epistasis (Figure 2.2-figure supplement 2). In this example, indirect pathways find a new local peak (U-A) with pathways that contain reciprocal sign epistasis.

Higher Dimensionality Hindered More Than Facilitated Adaptation

Stochastic evolutionary modeling offers novel insight into the navigability of genotype space. Modeling of evolution is computationally expensive and difficult to implement and therefore receives less attention in fitness landscape studies. However, as compared to static pathway analyses, evolutionary simulations are an improved representation of evolution in nature. This is due in part to the ability of simulated populations to traverse fitness valleys according to the valley depth and severity of epistatic interaction. We employed a Wright-Fisher model (29) to simulate populations of RNA molecules evolving on the direct and indirect pathway landscapes. We maintained a fixed population size, a fixed mutation rate, and allowed selection to be guided by the differences in the relative ligation fitness of the neighboring genotypes. We identified 84 genotypes that were present in the direct and indirect landscapes that had a relative ligation fitness ≤0.01. Using each one of these genotypes as starting points for a population of RNA molecules, we simulated adaptive evolution on either the direct pathway landscape or the indirect pathway landscape for 1,000 generations. Simulations were replicated 100 times for each starting genotype. During the simulation several metrics were tracked and recorded including, final fitness, final diversity, number of beneficial and deleterious mutations encountered, beneficial substitutions and the number of unique genotypes explored.

In order to evaluate the rate of adaptation during simulations on the different landscapes, we recorded the mean population fitness of the population at each generation. Both the final fitness achieved at 1000 generations as well as the rate that the average fitness increases are metrics of how easily the population achieves higher fitness. By comparing these average fitness values from both landscapes, we can evaluate whether indirect pathways facilitate or hinder adaptation (Fig. 2.3a). We found that, on average, evolutionary adaptation in ~38% of genotypes were hindered or slowed by higher dimensionality in the indirect landscape. For ~45% of genotypes adaption saw no significant difference between direct and indirect pathways. Only ~17% of genotypes significantly improved adaptation as a result of using indirect pathways. The mean population fitness that encompassed the average of the 100 replicates, indicated that adaptation occurred rapidly on both landscapes and revealed several periods where the population plateaus at a specific fitness (Fig. 2.3a, b). For the direct landscape these plateaus in adaptation corresponded to the two fitness peaks in the landscape (~1.8 and ~2.5). However, in the indirect landscape we saw several plateaus at lower levels of fitness. These plateaus corresponded to new fitness peaks in the indirect landscape that are isolated by reciprocal sign epistasis. These stasis peaks impeded adaptation and prevented simulated populations from attaining higher fitness. This was displayed when following the mean population fitness for each replicate from a single starting genotype on the direct and indirect landscapes (Fig. 2.3c). 79 out of 100 replicates were able to obtain the summit (fitness≅2.5) on the direct landscape. However, none of the 100 replicates starting from the same genotype on the indirect landscape were able to reach a fitness ≥2. Most (90%) of the replicates were isolated on a low-fitness stasis peak with reciprocal sign epistasis (Figure 2.3-figure supplement 1). Similar trends can be seen for each of the starting 84 genotypes by examining the final population fitness distributions on the direct (orange-red) and indirect (green-blue) landscapes (Fig. 2.3d). Of the 84 starting genotypes for our simulations, 32 genotypes on average obtained higher fitness using only direct pathways, 38 genotypes showed no significant difference and only 14 genotypes were significantly improved by the use of indirect pathways (Figure 2.3-figure supplement 2).

To validate that the observed slower adaptation in the indirect landscape was interconnected to the cases of higher epistasis severity, we developed a simple simulation model that challenged populations to ‘escape’ from a sub-optimal peak isolated by reciprocal sign epistasis (Figure 2.3-figure supplement 7). We found a predominant trend that as the magnitude of the epistatic interaction increased, the amount of successful escapes from the sub-optimal peak decreased. This was especially true as the epistatic values transition from 3 to 4. Interestingly, for the populations that did escape, there existed no correlation between epistatic value and the generation of escape. This validates the stochastic nature of the simulations. Importantly, the squares with high epistatic values are still possible to successfully escape, however as expected the success rate is significantly lower than those with lesser epistatic values.

Valley Library Epistasis and Evolutionary Simulations

Given the vastness of genotype space and the complex structure of ribozymes, it can be expected that a significant portion of sequence space is dominated by low-fitness genotypes. Therefore, from an evolutionary standpoint, the majority of evolutionary exploration by natural selection is through low-fitness valleys. To determine the effects of higher dimensionality in a region of sequence space with only low-fitness genotypes, a second library containing a direct and indirect landscape was designed (Fig. 2.4a). These two landscapes were equal in size and complexity to the previously discussed direct and indirect landscapes in the *Peak* library. The direct landscape in the *Valley* and *Peak* libraries start from the same reference genotype and overlap for 16 genotypes (Fig. 2.4b). We obtained relative ligation measurements for all 16,384 genotypes contained in the *Valley* library. The direct pathway landscape contained only a single low-fitness peak (fitness=0.08), while the indirect pathway landscape encompassed a total of 68 peaks, with a global peak or summit of fitness=0.12 (Fig. 2.4c, Table 2.1). As expected both landscapes are composed of very low-fitness genotypes with only 5 genotypes in the direct landscape and 53 genotypes in the indirect landscape with a relative ligation fitness >0.01 or 1% of wild-type (Figure 2.1-figure supplement 1).

Examining the pairwise epistasis in each landscape revealed that, similar to the *Peak* library, epistasis was proportionally more prevalent in the direct landscape (~88% of squares), as compared to the indirect landscape (~69%, Table 2.1, Figure 2.2-figure supplement 1). Magnitude epistasis was the most prevalent occurring in ~68% and ~50%

of squares in the direct and indirect landscapes, respectively. Simple sign epistasis was also slightly more proportionally prevalent in the direct landscape (~17% vs ~13%). However, reciprocal sign epistasis was approximately twice as prevalent in the indirect landscape (~6% vs ~3%). This resulted in the two landscapes having approximately the same amount of ruggedness (0.23-0.24). Interestingly, we found that a significantly lower proportion of genotypes in the direct landscape had mutational access to the summit genotype using a pathway with increasingly higher fitness (~23%). This number is significantly increased when using indirect pathways (~91%). Similar to the *Peak* library, the distributions of epistasis severity were similar in the direct and indirect landscapes, with the majority of epistatic interaction being of low severity (Fig. 2.4d). However, similar to the *Peak* library, more severe epistasis was encountered in the indirect pathway landscape (max~4) than the direct pathway landscape (max~2).

Simulating adaptive evolution on the direct and indirect landscapes of the *Valley* library yielded similar results to the direct and indirect landscapes in the *Peak* library. All simulations on the direct landscape attained the summit within the first 100 generations of evolution (Fig. 2.4e). This is particularly interesting because only ~23% of genotypes had mutational access to the summit using only beneficial mutations, therefore fitness valleys were easily traversed during the simulations. Adaptation was enhanced by indirect pathways in many instances, however for several starting genotypes simulated populations were hindered by sub-optimal peaks isolated by epistatic interactions (Figure 2.3-figure supplement 2). As expected and similar to the *Peak* library, we also found higher amounts of final population diversity, beneficial mutations, deleterious mutations and unique genotypes explored in the indirect landscape as compared to the direct landscape (Figure 2.3-figure supplement 3-6).

Discussion

Using empirical data, we constructed RNA fitness landscapes that encompass direct and indirect pathways of adaption. By simulating adaptive evolution on these landscapes, we found that indirect pathways can facilitate adaptation, however they can also significantly hinder it. Similar to protein fitness landscapes, we found regions where ‘extra-dimensional bypass’ using indirect pathways allowed for escape from isolated peaks (13). However, unlike protein, due to increased epistatic severity along indirect pathways, many populations were precluded from further evolutionary exploration and thus never attained the landscape summit. These sub-optimal stasis genotypes were caused by epistatic interactions, which were found to be prevalent in both landscapes.

Epistasis has long been shown to have significant impacts on the ruggedness and navigability of fitness landscapes (18–20,23). Negative epistasis, where the combination of two mutations results in a lower fitness than expected from the effect of each individual mutation is prevalent in RNA fitness landscapes (30). Without epistasis, fitness landscapes would be smooth, only contain a single peak and would be easily traversed by evolutionary exploration. However, epistasis increases landscape ruggedness and results in multi-peak landscapes with several local fitness maximums (3,19,21). Our analysis was limited to only pairwise epistatic interactions, however it has been shown that the magnitude or severity of epistasis is shown to significantly alter evolutionary trajectories or pathways much more than the order of epistasis (31). The number of peaks in a landscape is linked to reciprocal sign epistasis and directly affects the potential for evolutionary adaptation (22,32). The indirect landscapes in this study encompassed significantly more peaks compared to the direct landscapes. Indirect pathways not only introduced numerous novel pathways but also resulted in many new high fitness genotypes, some of which represented new peaks. The number of peaks in a landscape is a strong metric of its navigability (23). The findings of our study agree that the more peaks present in a landscape the less navigable it becomes.

The differences observed between RNA and protein fitness landscapes might be explained due to the different approaches used to assess adaptive evolution. Previous work in protein limited adaptation to sequential steps of single mutations of increasing fitness (13). This metric is similar to the approach taken in our study wherein the number of genotypes with accessible mutational pathways to the summit was assessed for each landscape. Although this metric can contribute to evolutionary potential, it makes one important assumption: evolution cannot cross fitness valleys. However, studies have shown that recombination and genetic drift allow for evolution to traverse fitness valleys (18,33,34). In fact, it has been shown in digital organisms that deleterious mutations can play an important role in adaptive evolution (35). By using a Wright-Fisher model to simulate adaptive evolution, we allow for fitness valleys to be crossed according to the valley depth and the severity of the epistatic interaction. This results in a dynamic evolutionary model that better represents the accessibility of mutational pathways in empirical fitness landscapes. The differences in the structural nature of protein and RNA might also contribute to the differences observed. Constraints caused by base pairing of RNA nucleotides produce structural epistasis which exhibit deep fitness valleys (28). The globular structure of proteins causes less pairwise structural interactions. It's also important to note that the number of pathways present in protein landscapes are significantly larger due to expanded mutational options. For example, when examining a single mutational pair, for both protein and RNA, the direct landscape consists of four (2^2) genotypes, where each genotype has two connections (Fig. 2.2.1a). However, in the indirect pathway landscapes, the network for RNA consists of 16 (4^2) genotypes as compared to 400 (20^2) genotypes for protein. The number of mutational connections for each genotype scales accordingly as well (RNA=6, protein=38). Therefore, it is possible that indirect pathways offer greater connectivity and thus facilitate evolution better in protein than in RNA fitness landscapes.

Our study suggests that increased dimensionality using indirect pathways in RNA fitness landscapes has the potential to circumvent fitness valleys, however indirect pathways might also harbor stasis genotypes isolated by reciprocal sign epistasis. Testing the law of parsimony, we found that indirect pathways can contribute significantly to the ruggedness and navigability of a local fitness landscape. It is important to note that these genotypes are only isolated within the scale of the fitness landscape studied. Our study only investigated seven mutational nucleotides, however the true dimensionality of a fitness landscape is much more complex and scales with length. Therefore, it is possible that a large RNA molecule might be able to circumvent all stasis genotypes and avoid evolutionary stasis (11,36). However, a recent comprehensive fitness landscape for a small RNA (*length*=24) found that fitness peaks were largely isolated from one another (36). We also found that higher dimensionality had similar effects in regions of high-fitness (*Peak*) or low-fitness (*Valley*). In conclusion, we found that higher dimensionality in empirical RNA fitness landscapes can allow for escape from epistatic traps, however it can also constrain evolution by the introduction of more severe epistasis.

Materials and Methods

Direct and Indirect Landscape Design

For our experiments, we used a previously developed assay to assess the relative ligation of a synthetic class III ligase ribozyme (Chapter 1). This ribozyme was first isolated from a large pool of random sequences using *in vitro* selection and yields a 2’,5’-phosphodiester bond by means of a terminal 2’ hydroxyl attack on the 5’ triphosphate of another RNA molecule (25). This causes displacement of the pyrophosphate and ligates the two RNA molecules together. The genotype network of this ribozyme was previously shown to intersect with a self-cleaving ribozyme network resulting in the identification of a single ‘intersection’ sequence that had the ability to both self-ligate and self-cleave (26). Using this ‘intersection’ sequence as a reference sequence, we identified 7 nucleotides (mut) that were different than the ‘prototype’ sequence initially isolated. Aligning these two sequences, we designed the direct library to encompass only the two nucleotides (n=2) present in either the ‘intersection’ sequence or the ‘prototype’ sequence. Therefore, they represent the most direct, parsimonious pathways from the ‘intersection’ to the ‘prototype’ sequence. This direct pathway landscape consists of 128 ($n^{mut} = 2^7$) unique genotypes and ~5,000 ($((n-1) \times (mut))! = (2-1) \times (7))!$) unique pathways from the ‘intersection’ to the ‘prototype’ sequence. The indirect pathway landscape encompasses the same 7 mutational nucleotides, however at each location all four nucleotide options (A, C, U, G) are allowed (n=4). This results in a significantly bigger library with 16,384 (4^7) unique genotypes and billions ($((4-1) \times$

(7))!)) of unique pathways. A region of the direct pathway landscape was previously characterized and was shown to contain a high fitness peak (Chapter 1); therefore, this landscape was called the *Peak* landscape (Fig. 2.2.1b). Library construction was accomplished by chemically synthesizing a degenerate DNA oligonucleotide that would serve as a template for in vitro transcription. A conserved sequence was added to the 3’-end of the library to serve as a universal primer binding site for reverse transcription (37). A single DNA library was synthesized that consisted of both the direct and indirect pathways. A similar approach was used in a different region that was expected to have only relatively low fitness genotypes. The intersection sequence was once again used as a reference sequence and a direct landscape and indirect landscape were designed. Due to the expected low fitness genotypes of the region, this landscape was called the *Valley* landscape (Fig. 2.4a).

Ligation Assay

The assay used to assess relative ligation was previously described (Chapter 1). In brief, the *Peak* and *Valley* libraries were ordered as ssDNA ultramers and were used for *in vitro* transcription of the ribozyme mutants. 20 picomoles of each ultramer library was annealed to a T7 primer by brief heating in 10 µL water. The template was then transcribed with rNTP (25mM, NEB), MEGAshortscript T7 Reaction Buffer (10X, Thermo Fisher) and MEGAshortscript T7 RNA Polymerase (Thermo Fisher) at 37 °C for 2 hours. The DNA was then degraded by DNase treatment, and the RNA was purified and normalized to 5µM. 25 picomoles of the RNA were then placed in 200mM Tris pH 7.5 and heated and cooled. 500 picomoles of ligation substrate were then added with $MgCl_2$ (50mM). The mixture was then incubated for 2 hours at 37 °C. The ligation reaction was then reverse transcribed using SMARTscribe Reverse Transcriptase (10 units, Clontech) by incubating at 42 °C for 90 mins. The reaction was stopped and the RNA was degraded by heating the sample. To selectively amplify the cDNA that successfully performed the ligation reaction, a mix of phased insert selective PCR primers were used (Chapter 1). The PCR reaction consisted of purified cDNA, KAPA HiFi HotStart ReadyMix (2X, KAPA Biosystems), phased insert selective ligation primers, and reverse transcription primer. To prevent PCR bias, multiple cycles of PCR were assessed using gel electrophoresis and an appropriate cycle was chosen that was still in linear growth. The selective PCR product was then purified. To determine the pre-selection starting abundance of each genotype, a portion of each sample was reverse transcribed using the template switching capabilities of SMARTscribe reverse transcriptase. In preparation for high-throughput sequencing, Illumina adapter sequences were added to the cDNA using low-cycle PCR. Each library was done entirely in triplicate yielding a total of twelve samples (3 *Peak* pre-selection, 3 *Peak* post-selection, 3 *Valley* pre-selection, 3 *Valley* post-selection). Each sample was assigned a unique Illumina adapter barcode that was added during the adapter PCR.

High-Throughput Sequencing

In preparation for high-throughput sequencing the twelve samples, each with unique Illumina adapter barcodes were pooled and sent to the University of Oregon Genomics and Cell Characterization Core Facility. The samples were sequenced using Illumina HiSeq 4000 Single End 150 with 25% PhiX addition. This generated ~241 million reads (Cluster PF Yield) between the twelve samples with a mean quality score of 39.5.

Sequencing Data Analysis

The high-throughput sequencing was analyzed using custom Python scripts. The scripts identified the conserved 3' sequence, determined if the sequence was ligated or unligated and extracted the sequence at the 7 mutational nucleotides to identify the read genotype. This process was repeated for each of the three replicates. A genotype was considered to be a part of the ligase genotype network only if it was detected as ligated in all three replicates and had a catalytic rate above the uncatalyzed ligation rate (38,39). The three replicates for pre-selection and post-selection for each library (*Peak* or *Valley*) were then correlated with each other to verify accurate and precise measurements (Figure 2.1-supplement figure 2).

The level of enrichment between the pre-selection and the post-selection samples (Figure 2.1-supplement figure 1) was used to determine the relative ligation fitness for each genotype. The relative abundance of each genotype was normalized by dividing the read count for each genotype by the total number of sequencing reads in that sample. The level of enrichment was calculated by dividing the relative abundance of a genotype in the post-selection sample by the relative abundance in the pre-selection sample. We observed detectable ligase activity for all 16,384 genotypes in the *Peak* and *Valley* landscapes.

Empirical Fitness Landscape Construction and Characterization

The highly complex nature of the genotype networks presented here are difficult to visually interpret and assess. Therefore, to aid in visualization, three-dimensional fitness landscapes were constructed for both the direct and indirect pathways in the *Peak* (Fig. 2.2a) and *Valley* (Fig. 2.4b) landscapes. Due to the fact that the direct pathways are encompassed within the indirect pathway landscape, we first used custom Python scripts to isolate the 128 genotypes from the sequencing libraries that correspond to the direct pathway landscapes. We then determined all of the edges within the genotypes in the direct and indirect pathway landscapes. An edge connects two genotypes (nodes) that can be interconverted by a single mutation. The corresponding fitness landscapes were then constructed using Gephi software (40), with ForceAtlas 2 used as the layout to approximate genotype repulsion using a Barnes-Hut calculation. The z-axis was generated using the Network Splitter 3D plugin and indicates the relative ligation fitness.

To characterize the highly complex fitness landscapes we used custom Python scripts and the Genonets Server (27). Genonets allows for in-depth analysis of genotype networks including peak, summit, robustness, and epistasis calculations. The epistasis calculation includes the assessment and categorizing of every square present in the network. A square represents a pair of mutations and consists of a total of 4 genotypes. For example, the simple network presented in Figure 1a direct represents a single square. There exist 672 squares in the direct pathway landscapes and 774,144 in the indirect pathway landscapes. Each square is categorized as either containing magnitude epistasis, simple sign epistasis, reciprocal sign epistasis or no epistasis depending on the sign of the individual mutations and of the combination of the mutations (23). For each class of epistasis, the proportion of all squares belonging to that class was determined. This allowed for a thorough assessment of the prevalence of each class of epistasis in the direct and indirect pathway landscapes. Using the prevalence of pairwise epistasis we calculated an estimate of fitness landscape ruggedness as $f_{sign}+2f_{reciprocal}$, where f_{type} indicates the fraction of all the squares that fell into each epistasis class (13). However, because not all epistatic interactions are equal in their severity, we used custom Python scripts to calculate the magnitude of each epistatic interaction. Epistatic values for each square of mutation pairs was calculated as $\varepsilon = \log_{10} (W_{AB}*W_{wt} / W_{A}*W_{B})$, where W_A and W_B are the fitness of RNA variants with a single mutation, W_{AB} is the fitness of the variant with both mutations, and W_{wt} is the fitness of the wild-type. In this case, the wild-type indicates the variant in the square with no mutations. The distribution of epistatic magnitudes was then determined for the direct and indirect pathway landscapes (Fig. 2.2b, Fig. 2.4d).

Simulated Adaptive Evolution on Empirical Fitness Landscapes

In order to simulate the adaptive evolution of populations of RNA molecules on the fitness landscapes, we used previously custom Python scripts that were previously described (Chapter 1). In brief, the simulation models evolution based on the Wright-Fisher approach (18,29). Each simulation begins with 1000 individuals of the same genotype. Each generation of the simulation, a new population of 1000 individuals was populated. The genotypes of the new population were the result of stochasticity (introduced by random number generation) and dependent on the relative fitness of the parent genotype. First, a genotype was randomly chosen form the parent population. The fitness of the genotype was normalized (0-1) and then ‘competed’ against a randomly selected value between 0 and 1. If the parent genotype fitness was lower than the random value, the genotype was not placed in the offspring population. If the parent genotype fitness was greater than or equal to the random value, it was placed in the offspring population. The genotype in the offspring population then had chance to mutate at a randomly determined nucleotide. The probability of a single mutation occurring was dependent on a predetermined mutation rate. For our simulations we used a constant mutation rate (μ) of 0.01. This process of survival and potential mutation was repeated until the offspring population consisted of 1000 individuals. This constituted a single generation and the simulation was repeated for 1000 generations. As starting genotypes for simulated evolution on the *Peak* landscape we selected the 84 genotypes on the direct pathway landscape with relative fitness ≤0.01. As starting genotypes for simulated evolution on the *Valley* landscape we selected the 75 genotypes on the direct pathway landscape with relative fitness ≤0.0001. The simulation was started from each genotype and ran on both the direct and indirect pathway landscapes for 100 replicates each. The mean population fitness for each 100 replicates for each starting genotype were averaged and plotted as a function of generational time (Fig. 2.3a, b, Fig. 2.4e). For each simulation several metrics were tracked and recorded, namely: final fitness (Fig. 2.3c, d), final diversity, number of beneficial, and deleterious mutations encountered, beneficial substitutions and the number of unique genotypes explored (Figure 2.3-figure supplement 2-6).

To validate the consequences of epistatic interactions on adaptive evolution, we also developed a model that demonstrates the difficulty of ‘escaping’ from highly epistatic genotypes. The model consists of a single square of mutation pairs, resulting in a total of four genotypes. The wild-type sequence is held constant with a fitness (W_{wt}) of 1. The two variants with a single mutant (A, B) were varied from a fitness (W_{A}, W_{B}) of 0.1 to 1. The variant with two mutations (AB) was varied from a fitness (W_{AB}) of 1 to 10. Therefore, AB was the optimal peak in the square. This generated 4,756 unique fitness combinations and a wide distribution of epistatic values (Figure 2.3-figure supplement 7), calculated using $\varepsilon = \log_{10} (W_{AB}*W_{wt} / W_{A}*W_{B})$. Each square indicates a case of reciprocal sign epistasis. Simulating adaptive evolution on each square begins with a population of 1000 individuals on the wild-type genotype. The evolutionary simulation then proceeded as previously described. The simulation continued for up to 100 generations or until 50% of the population had successfully traversed the intermediate valley and made it to the optimal peak. If the population made it to the optimal peak within the 100 generations, the generation of successful 'escape' was recorded. This was repeated for 100 replicates of each square. The average number of generations to escape was then plotted as a function of the epistasis value calculated for that square (Figure 2.3-figure supplement 7).

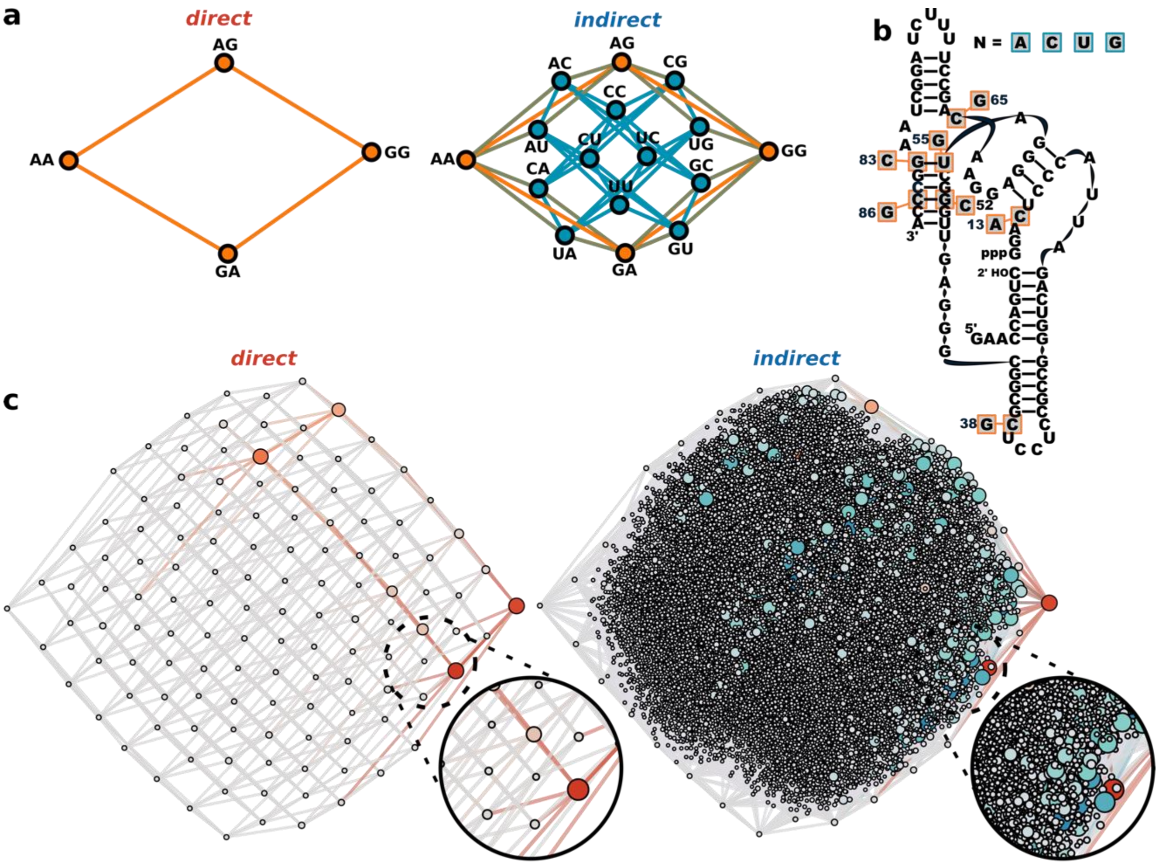


Figure 2.1: Direct and indirect pathways in an empirical RNA fitness landscape.

(a) Example of direct (orange) and indirect (blue) pathways leading from a single genotype ‘AA’ to ‘GG’. Nodes represent a unique genotype and edges connect nodes that differ by a single mutation. (b) Secondary structure of ligase ribozyme depicting the library design for the *Peak* library. Nucleotides in orange indicate the direct parsimonious nucleotide mutations included in the direct landscape. The indirect landscape contains all four nucleotides (A, C, U, G) at these mutations. (c) The genotype network of the direct (left) and indirect (right) landscapes of the *Peak* library are depicted. Node size and color indicate the relative ligation fitness.

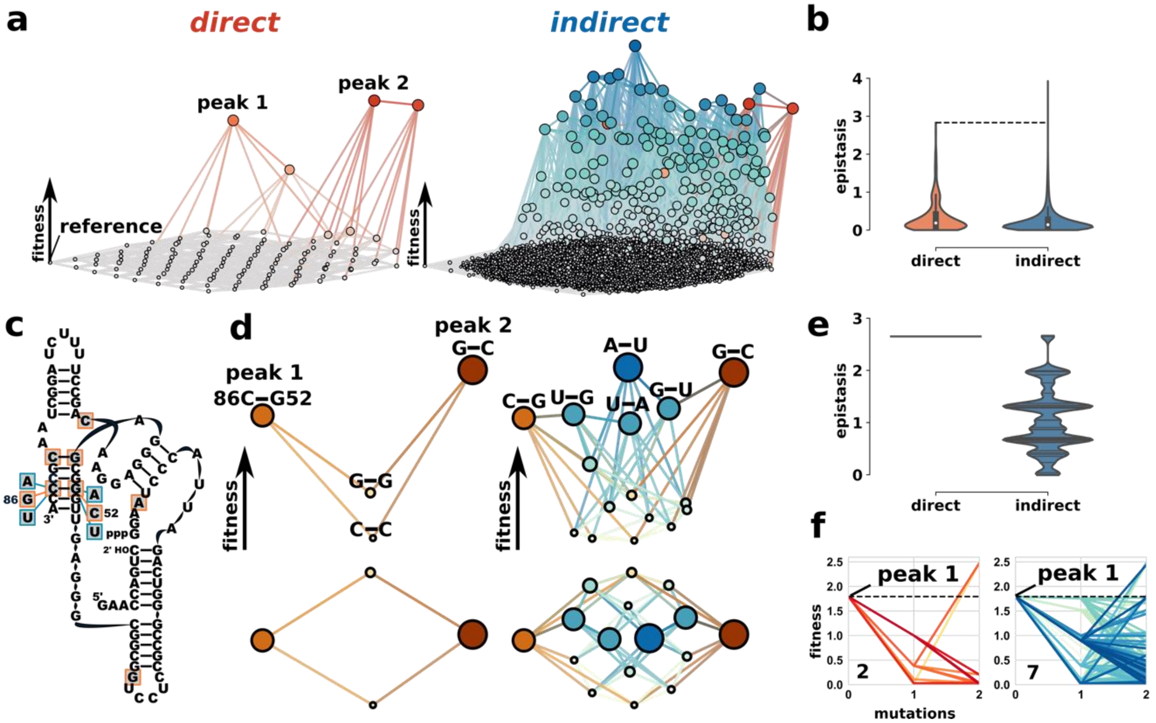


Figure 2.2: Characterization of epistasis in direct and indirect pathways in the *Peak* library fitness landscape.

(a) Three-dimensional fitness landscapes depicting the direct and indirect pathway landscapes in the Peak library. Nodes represent a single genotype and edges connect nodes that differ by a single mutation. The height of the node represents the relative ligase fitness. (b) Distribution of epistatic values for pairs of mutations in the direct (orange) and indirect (blue) landscapes. Epistatic values were calculated as ε = log10 (WAB*Wwt / WA*WB), where WA and WB are the fitness of RNA variants with a single mutation, WAB is the fitness of the variant with both mutations, and Wwt is the fitness of the wild-type. (c) Secondary structure of the ligase ribozyme indicating the genetic background and nucleotides that are mutated between peak 1 and peak 2 in the direct landscape. (d) An example of reciprocal sign epistasis between the two peaks in the direct library. The height and size of each node represents relative ligation fitness. Nodes in the direct landscape are orange-red and nodes unique to the indirect landscape are green-blue. The nucleotides above each node indicate the two nucleotides present at position 86 and 52. (e) Distribution of epistatic values for pairs of mutation in the subgraph presented in panel d. Direct values are depicted in orange and indirect in blue. (f) Local fitness landscapes of peak 1 in the direct landscape. The fitness of the peak genotype is plotted at mutations=0 and marked with a dashed line. Lines depict the two-mutation pathways (orange=direct, blue=indirect) away from this genotype. The number on each graph represents the total number of two-mutation pathways that lead to higher fitness.

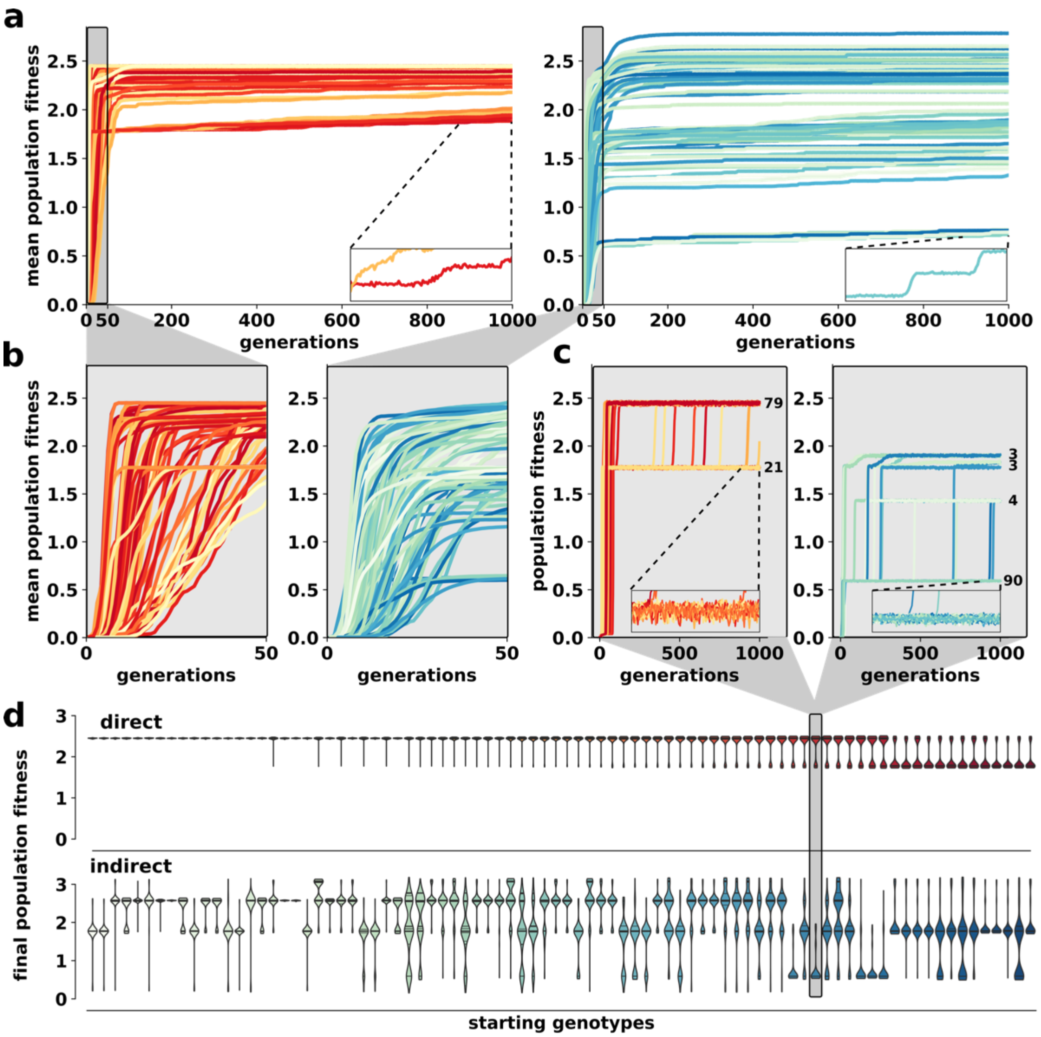


Figure 2.3: Computational simulation of adaptive evolution on the direct and indirect landscapes in the *Peak* library.

(a) Average rates for multiple evolutionary simulations starting from the 84 starting genotypes in the direct landscape with fitness <0.01. Each trace represents a different starting genotype and shows the mean fitness of 100 simulations plotted as a function of generation time. 100 simulations were performed on the direct (orange) or indirect (blue) pathway landscapes for each starting genotype. (b) Enhanced view of the first 50 generations of adaptive evolution from panel a. (c) Rates of adaptation on the direct and indirect landscapes starting from a single starting genotype. Each trace shows the average population fitness as a function of generation time for a separate simulation consisting of 1000 individuals. Numbers indicate the number of overlapped simulations. (d) Final population fitness following 1000 generations of adaptive evolution on the direct (orange-red) and indirect (blue-green) landscapes starting from the 84 starting genotypes in the direct landscape with fitness <0.01. Each plot represents the distribution of 100 replicates. The distributions are ordered by the mean final fitness on the direct landscape and the corresponding simulations on the indirect landscape are shown directly below.

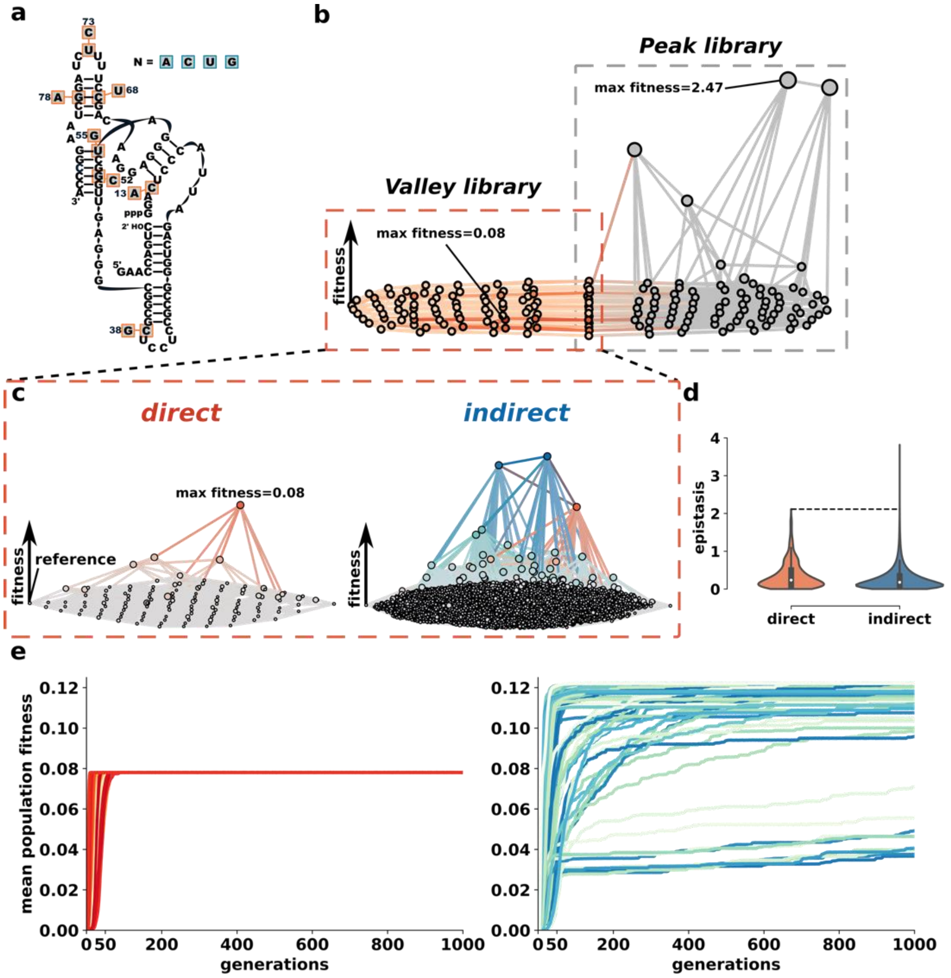


Figure 2.4: *Valley* library fitness landscape characterization.

(a) Secondary structure of ligase ribozyme depicting the library design for the *Valley* library. Nucleotides in orange indicate the direct parsimonious nucleotide mutations included in the direct landscape. The indirect landscape contains all four nucleotides (A, C, U, G) at these mutations. (b) Three-dimensional fitness landscape depicting the direct landscapes from the *Peak* (grey) and *Valley* (orange) library. Each node indicates a unique genotype and the edges connect genotypes that are one mutation apart. (c) Three-dimensional landscapes depicting the direct and indirect pathway landscapes in the *Valley* library. Nodes represent a single genotype and edges connect nodes that differ by a single mutation. The height of the node represents the relative ligase fitness. (d) Distribution of epistatic values for pairs of mutations in the direct (orange) and indirect (blue) landscapes. Epistatic values were calculated as $\varepsilon = \log_{10} (W_{AB}*W_{wt} / W_A*W_B)$, where W_A and W_B are the fitness of RNA variants with a single mutation, W_{AB} is the fitness of the variant with both mutations, and W_{wt} is the fitness of the wild-type. (e) Average rates for multiple evolutionary simulations starting from the 75 starting genotypes in the direct landscape with fitness <0.0001. Each trace represents a different starting genotype and shows the mean fitness of 100 simulations plotted as a function of generation time. 100 simulations were performed on the direct (orange) or indirect (blue) pathway landscapes for each starting genotype.

References

1. de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 2014 Jul;15(7):480–90.

2. Athavale SS, Spicer B, Chen IA. Experimental fitness landscapes to understand the molecular evolution of RNA-based life. Current Opinion in Chemical Biology. 2014 Oct;22:35–9.

3. Kondrashov DA, Kondrashov FA. Topological features of rugged fitness landscapes in sequence space. Trends in Genetics. 2015 Jan;31(1):24–33.

4. Kauffman S, Levin S. Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol. 1987 Sep 7;128(1):11–45.

5. Østman B, Hintze A, Adami C. Critical properties of complex fitness landscapes. arXiv:10062908 [q-bio] [Internet]. 2010 Jun 15 [cited 2017 Oct 1]; Available from: http://arxiv.org/abs/1006.2908

6. Hayden EJ. Empirical analysis of RNA robustness and evolution using high-throughput sequencing of ribozyme reactions. Methods. 2016 Aug 15;106:97–104.

7. Kobori S, Yokobayashi Y. High-Throughput Mutational Analysis of a Twister Ribozyme. Angew Chem Int Ed. 2016 Aug 22;55(35):10354–7.

8. Steinberg B, Ostermeier M. Environmental changes bridge evolutionary valleys. Science Advances. 2016 Jan 22;2(1):e1500921–e1500921.

9. Pitt JN, Ferré-D’Amaré AR. Rapid Construction of Empirical RNA Fitness Landscapes. Science. 2010 Oct 15;330(6002):376–9.

10. Jimenez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen IA. Comprehensive experimental fitness landscape and evolutionary network for small RNA. Proceedings of the National Academy of Sciences. 2013 Sep 10;110(37):14984–9.

11. Gavrilets S. Evolution and speciation on holey adaptive landscapes. Trends Ecol Evol (Amst). 1997 Aug;12(8):307–12.

12. Kaplan J. The end of the adaptive landscape metaphor? Biology & Philosophy. 2008 Nov;23(5):625–38.

13. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife Sciences. 2016 Jul 8;5:e16965.

14. Steel M, Penny D. Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. Mol Biol Evol. 2000 Jun 1;17(6):839–50.

15. Fong JH, Geer LY, Panchenko AR, Bryant SH. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol. 2007 Feb 9;366(1):307–15.

16. Lee YH, DSouza LM, Fox GE. Equally parsimonious pathways through an RNA sequence space are not equally likely. J Mol Evol. 1997 Sep;45(3):278–84.

17. Franke J, Klözer A, Visser JAGM de, Krug J. Evolutionary Accessibility of Mutational Pathways. PLOS Computational Biology. 2011 Aug 18;7(8):e1002134.

18. Ostman B, Hintze A, Adami C. Impact of epistasis and pleiotropy on evolutionary adaptation. Proc Biol Sci. 2012 Jan 22;279(1727):247–56.

19. Beerenwinkel N, Pachter L, Sturmfels B. Epistasis and Shapes of Fitness Landscapes. arXiv:q-bio/0603034 [Internet]. 2006 Mar 29 [cited 2015 Apr 17]; Available from: http://arxiv.org/abs/q-bio/0603034

20. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006 Dec 14;444(7121):929–32.

21. Whitlock MC, Phillips PC, Moore FB-G, Tonsor SJ. Multiple Fitness Peaks and Epistasis. Annual Review of Ecology and Systematics. 1995;26(1):601–29.

22. Poelwijk FJ, Tănase-Nicola S, Kiviet DJ, Tans SJ. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. Journal of Theoretical Biology. 2011 Mar 7;272(1):141–4.

23. Aguilar-Rodríguez J, Payne JL, Wagner A. A thousand empirical adaptive landscapes and their navigability. Nature Ecology & Evolution. 2017 Jan 23;1:0045.

24. Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. Empirical fitness landscapes reveal accessible evolutionary paths. Nature. 2007 Jan 25;445(7126):383–6.

25. Ekland EH, Szostak JW, Bartel DP. Structurally complex and highly active RNA ligases derived from random RNA sequences. Science. 1995 Jul 21;269(5222):364–70.

26. Schultes EA, Bartel DP. One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science. 2000 Jul 21;289(5478):448–52.

27. Khalid F, Aguilar-Rodríguez J, Wagner A, Payne JL. Genonets server—a web server for the construction, analysis and visualization of genotype networks. Nucl Acids Res. 2016 Jul 8;44(W1):W70–6.

28. Dutheil JY, Jossinet F, Westhof E. Base Pairing Constraints Drive Structural Epistasis in Ribosomal RNA Sequences. Mol Biol Evol. 2010 Aug 1;27(8):1868–76.

29. Donnelly P, Weber N. The Wright-Fisher model with temporally varying selection and population size. J Math Biology. 1985 Jun 1;22(1):21–9.

30. Bendixsen DP, Østman B, Hayden EJ. Negative Epistasis in Experimental RNA Fitness Landscapes. J Mol Evol. 2017 Nov 10;1–10.

31. Sailer ZR, Harms MJ. High-order epistasis shapes evolutionary trajectories. PLOS Computational Biology. 2017 May 15;13(5):e1005541.

32. Kvitek DJ, Sherlock G. Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. PLoS Genet. 2011 Apr 28;7(4):e1002056.

33. Weissman DB, Desai MM, Fisher DS, Feldman MW. The rate at which asexual populations cross fitness valleys. Theor Popul Biol. 2009 Jun;75(4):286–300.

34. Weissman DB, Feldman MW, Fisher DS. The Rate of Fitness-Valley Crossing in Sexual Populations. Genetics. 2010 Dec 1;186(4):1389–410.

35. Covert AW, Lenski RE, Wilke CO, Ofria C. Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. PNAS. 2013 Aug 20;110(34):E3171–8.

36. Gavrilets S, Gravner J. Percolation on the fitness hypercube and the evolution of reproductive isolation. J Theor Biol. 1997 Jan 7;184(1):51–64.

37. Wilkinson KA, Merino EJ, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc. 2006;1(3):1610–6.

38. Rohatgi R, Bartel DP, Szostak JW. Nonenzymatic, Template-Directed Ligation of Oligoribonucleotides Is Highly Regioselective for the Formation of 3‘−5‘ Phosphodiester Bonds. J Am Chem Soc. 1996 Jan 1;118(14):3340–4.

39. Rohatgi R, Bartel DP, Szostak JW. Kinetic and Mechanistic Analysis of Nonenzymatic, Template-Directed Oligoribonucleotide Ligation. J Am Chem Soc. 1996 Jan 1;118(14):3332–9.

40. Mathieu Bastian, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulation Networks. In: Third International ICWSM Conference. 2009.

Acknowledgements

This study was supported by Boise State University (Biomolecular Sciences Graduate Programs), National Science Foundation Directorate for Biological Sciences (Grant No. MCB-1413664), National Science Foundation Office of Integrative Activities (Grant No. OIA-1738865), and National Aeronautics and Space Administration (Grant No. 80NSSC17K0738).

Supporting Information

Table 2.1 Genotype network analysis of direct and indirect fitness landscapes

Genotype Network Analysis	Peak Library		Valley Library	
	Direct	Indirect	Direct	Indirect
peaks	2	20	1	68
genotype network sizes	128	16384	128	16384
genotypes accessible to summit	86 (67%)	11320 (69%)	29 (23%)	14991 (91%)
number of squares	672	774144	672	774144
magnitude epistasis	62.1%	51.6%	67.6%	50.3%
simple sign epistasis	23.7%	8.2%	16.7%	13.1%
reciprocal sign epistasis	5.5%	2.8%	3.3%	5.7%
total epistasis	91.2%	62.6%	87.5%	69.1%
ruggedness	0.346718	0.13749	0.232	0.244634

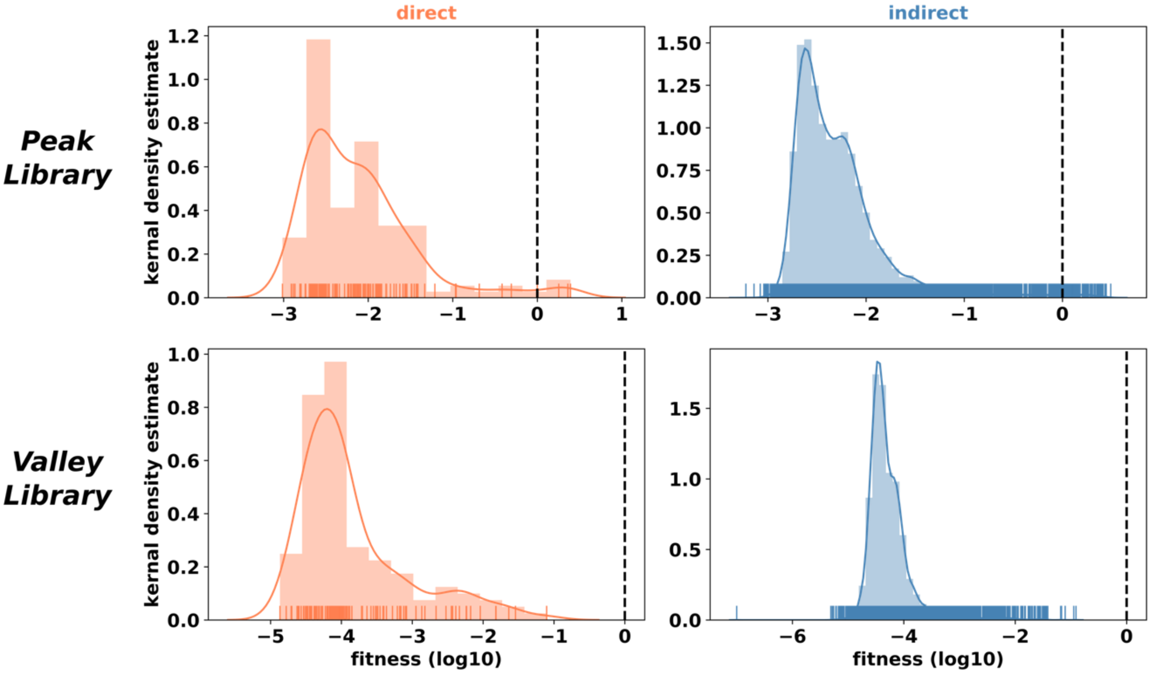


Figure 2.1-supplement figure 1: Distribution of relative ligation fitnesses for the *Peak* and *Valley* libraries.

The distributions of relative fitness in the direct (orange) and indirect (blue) pathway landscapes are plotted as histograms with the kernel density estimate on the y-axis and the log10 fitness on the x-axis. The dashed black line indicates wild-type fitness.

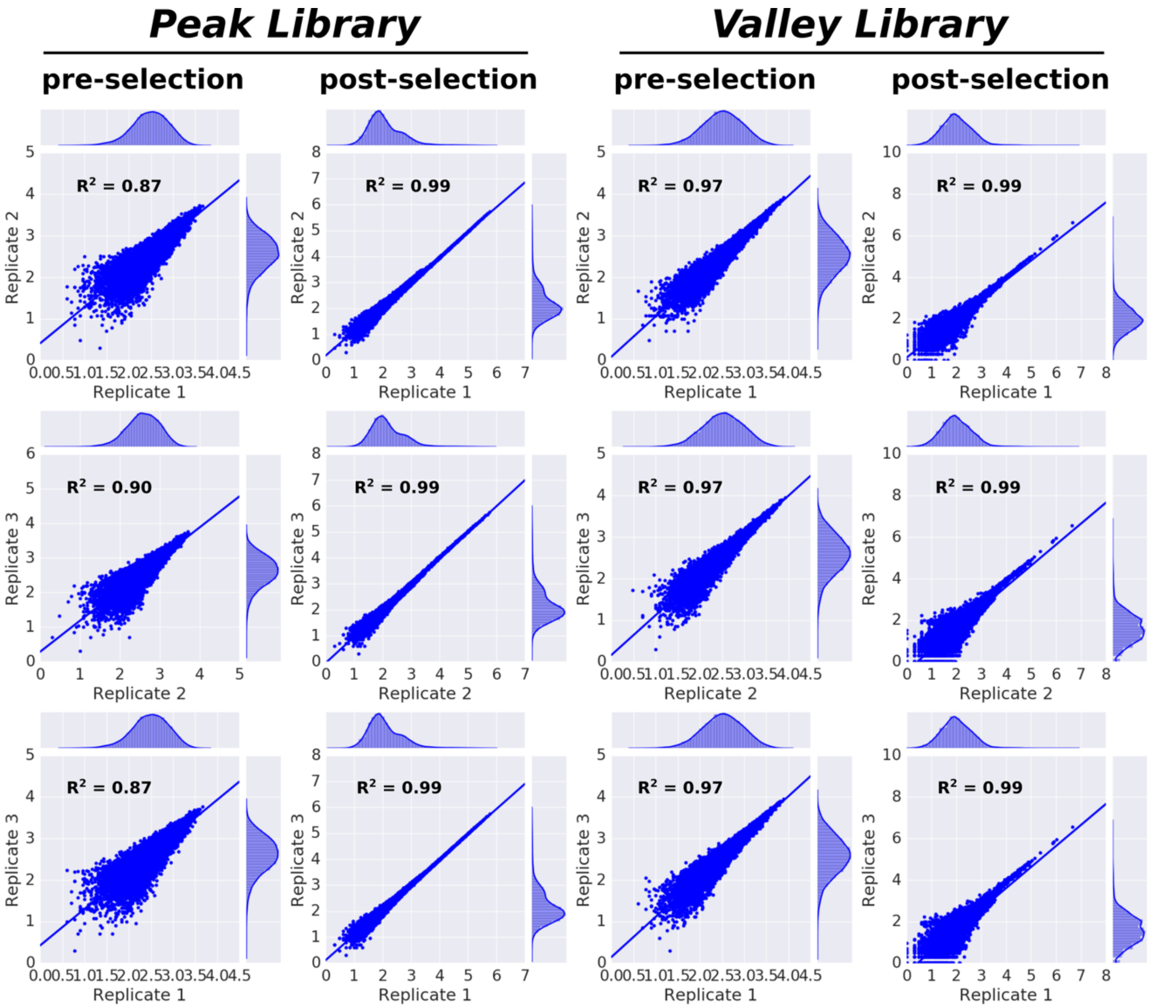


Figure 2.1-figure supplement 2: Correlation of high-throughput sequencing replicates for the *Peak* and *Valley* libraries.

Correlation of sequencing reads for pre-selection and post-selection for three replicates. Each figure displays all 16,384 genotypes and indicates the frequency that a specific genotype was observed in a particular replicate (x-axis) vs. another replicate (y-axis). Sequence kernel density estimation is also displayed for each replicate in the jointplot. The number of reads on the x and y-axis are log10 transformed.

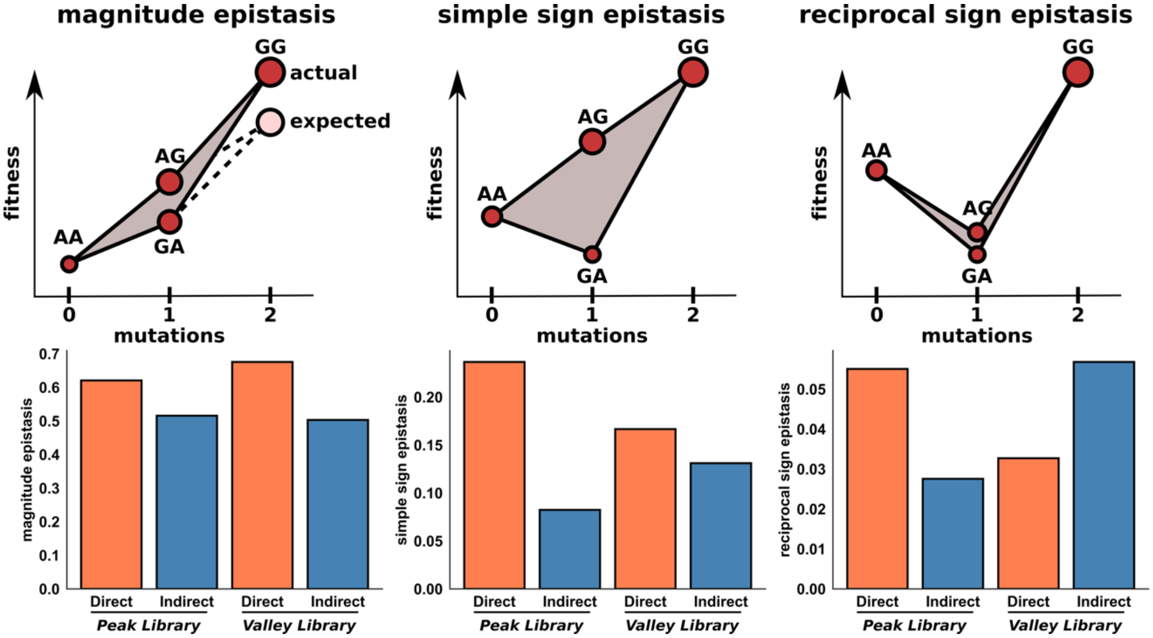


Figure 2.2-supplement figure 1: Prevalence of pairwise epistasis in direct and indirect pathway landscapes.

Examples of magnitude, simple sign and reciprocal sign epistasis, with the relative prevalence of these classes of epistasis in the direct (orange) and indirect (blue) landscapes of the *Peak* and *Valley* libraries. The y-axis indicates the proportion of all squares (mutational pairs) that correspond to each epistasis class.

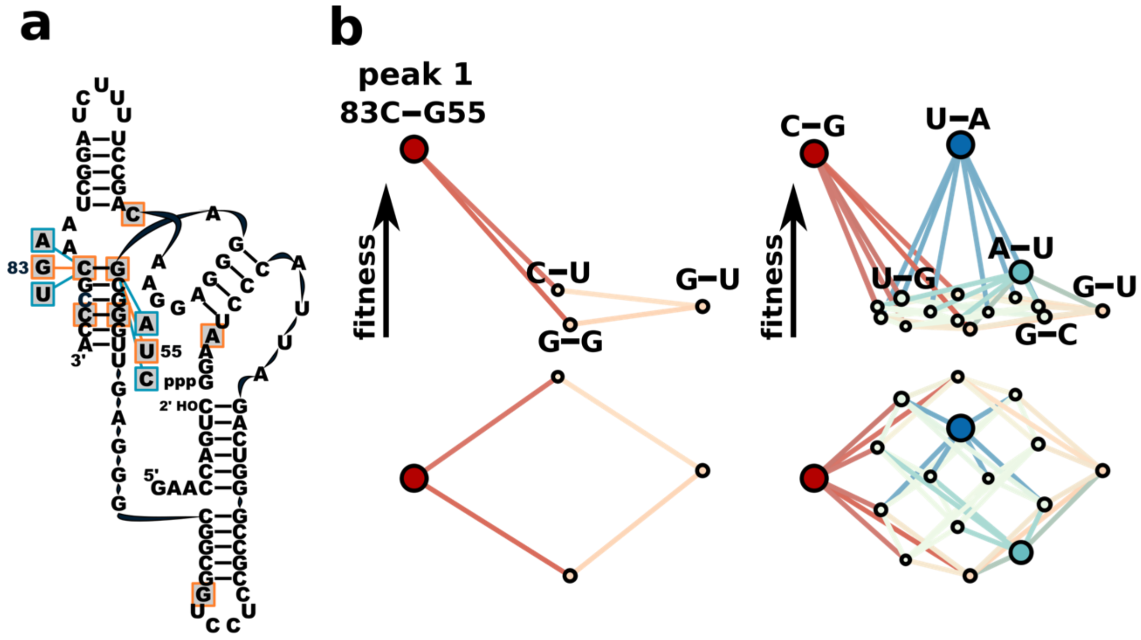


Figure 2.2-figure supplement 2: Epistasis in the base pair of terminal 3' stem of the ligase ribozyme.

(a) Secondary structure of the ligase ribozyme indicating the nucleotides and genetic background involved in the direct and indirect pathway landscape. (b) Local fitness landscape of a base pair. The height and size of each node represents relative ligation fitness. The nucleotides above each node indicate the two nucleotides present at position 83 and 55.

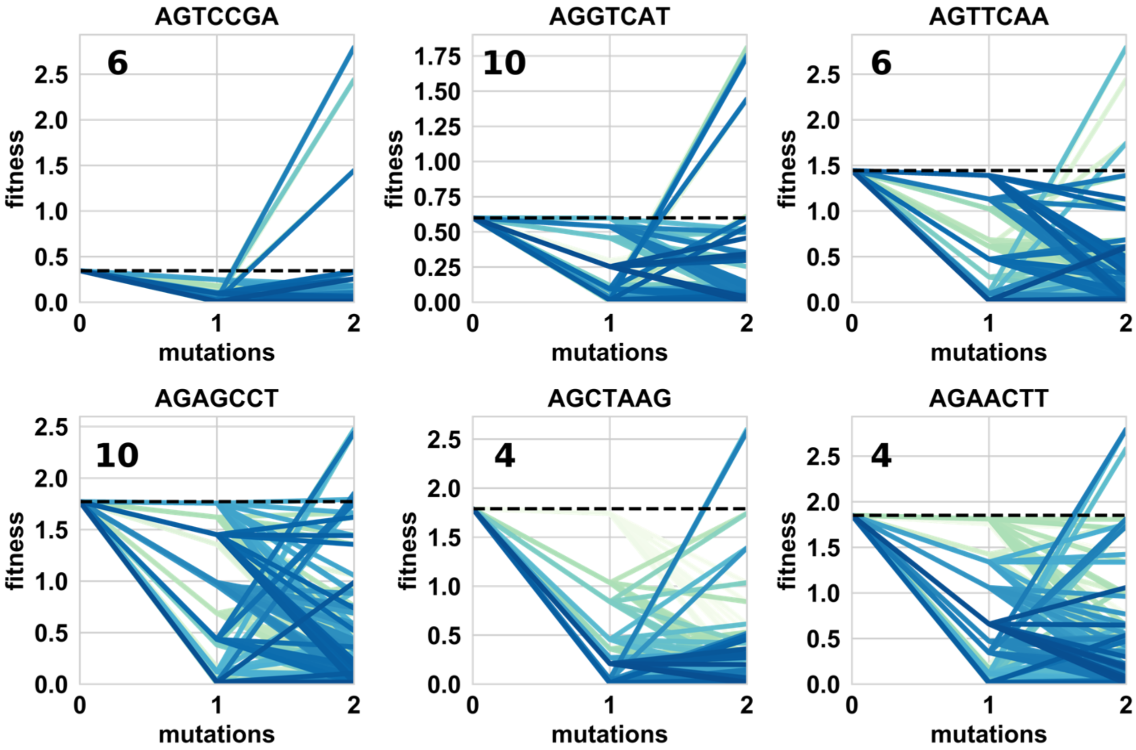


Figure 2.3-figure supplement 1: Local fitness landscapes of stasis peak genotypes in the indirect landscape.

The fitness of the stasis peak genotype is plotted at mutations=0 and marked with a dashed line. Lines depict the two-mutation pathways away from this genotype. Titles of each subplot indicate the seven-nucleotide genotype of the stasis peak. The number on each subplot indicates the number of two-mutation paths to higher fitness from the stasis peak.

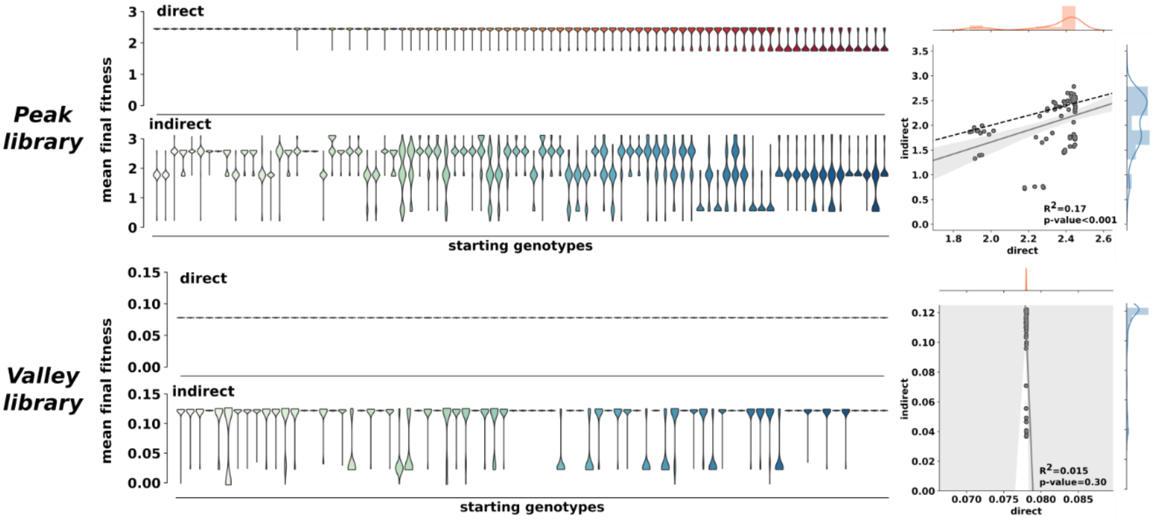


Figure 2.3-figure supplement 2 Mean population final fitness following adaptive evolution on direct and indirect pathway landscapes.

Final population fitness following 1000 generations of adaptive evolution on the direct (orange-red) and indirect (blue-green) landscapes of the *Peak* and *Valley* library. Each plot represents the distribution of 100 replicates of simulated evolution for each unique starting genotype. The distributions are ordered by the mean final fitness on the direct landscape and the corresponding simulations on the indirect landscape are shown directly below. The correlation of mean final fitness values from the direct and indirect landscapes for each of the starting points is also depicted.

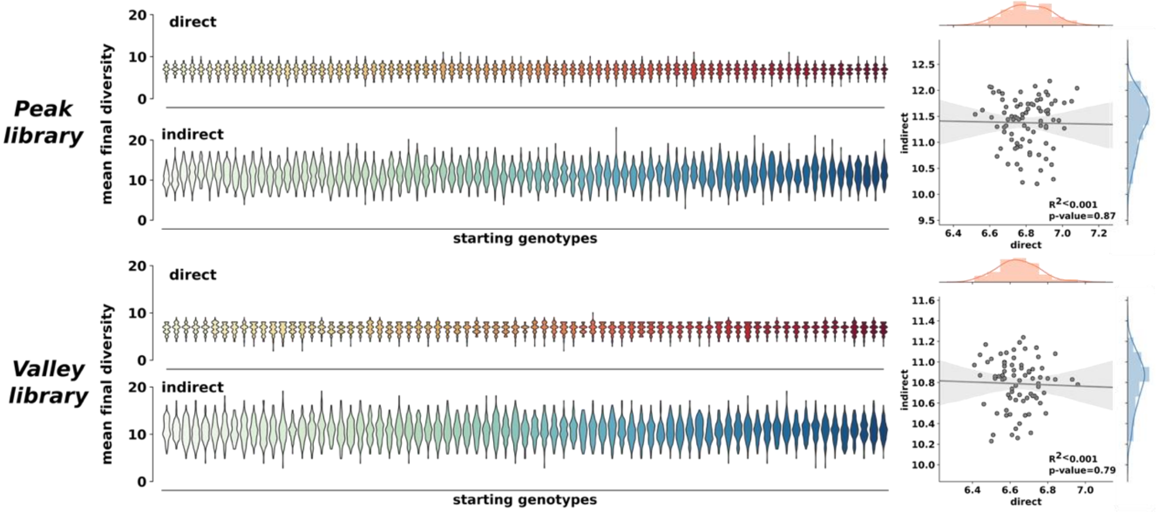


Figure 2.3-figure supplement 3: Mean population final diversity following adaptive evolution on direct and indirect pathway landscapes.

Final population diversity following 1000 generations of adaptive evolution on the direct (orange-red) and indirect (blue-green) landscapes of the Peak and Valley library. Each plot represents the distribution of 100 replicates of simulated evolution for each unique starting genotype. The distributions are ordered by the mean final fitness on the direct landscape and the corresponding simulations on the indirect landscape are shown directly below. The correlation of mean final diversity values from the direct and indirect landscapes for each of the starting points is also depicted.

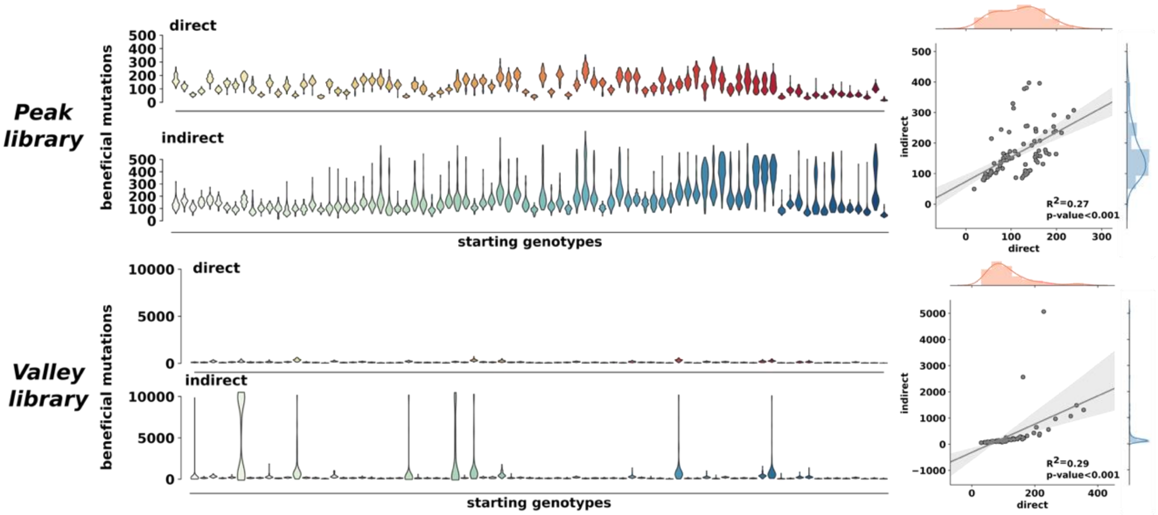


Figure 2.3-figure supplement 4: Mean beneficial mutations following adaptive evolution on direct and indirect pathway landscapes.

Mean beneficial mutations following 1000 generations of adaptive evolution on the direct (orange-red) and indirect (blue-green) landscapes of the *Peak* and *Valley* library. Each plot represents the distribution of 100 replicates of simulated evolution for each unique starting genotype. The distributions are ordered by the mean final fitness on the direct landscape and the corresponding simulations on the indirect landscape are shown directly below. The correlation of mean beneficial mutation values from the direct and indirect landscapes for each of the starting points is also depicted.

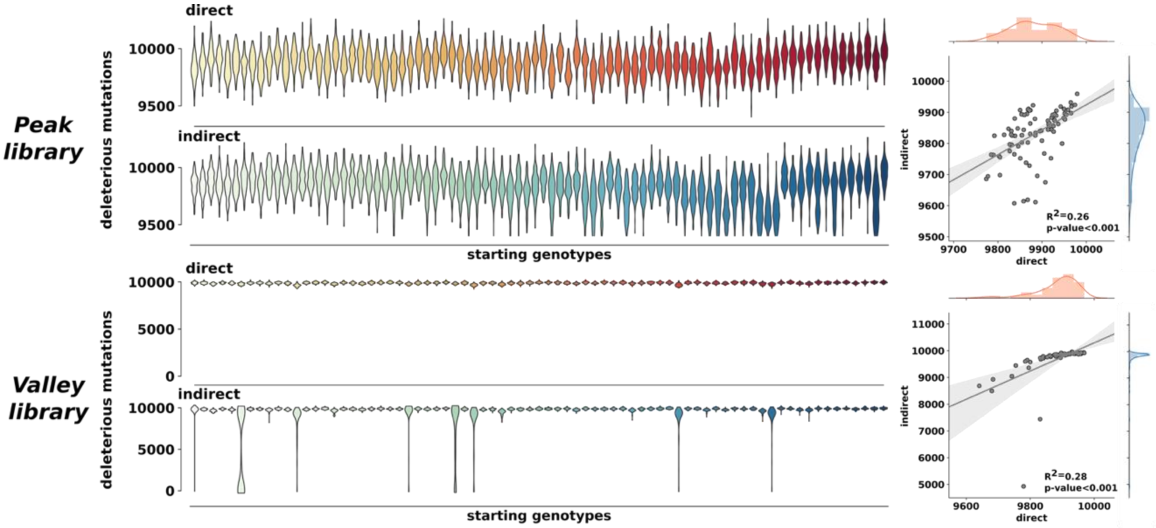


Figure 2.3-figure supplement 5: Mean deleterious mutations following adaptive evolution on direct and indirect pathway landscapes.

Mean deleterious mutations following 1000 generations of adaptive evolution on the direct (orange-red) and indirect (blue-green) landscapes of the *Peak* and *Valley* library. Each plot represents the distribution of 100 replicates of simulated evolution for each unique starting genotype. The distributions are ordered by the mean final fitness on the direct landscape and the corresponding simulations on the indirect landscape are shown directly below. The correlation of mean deleterious mutation values from the direct and indirect landscapes for each of the starting points is also depicted.

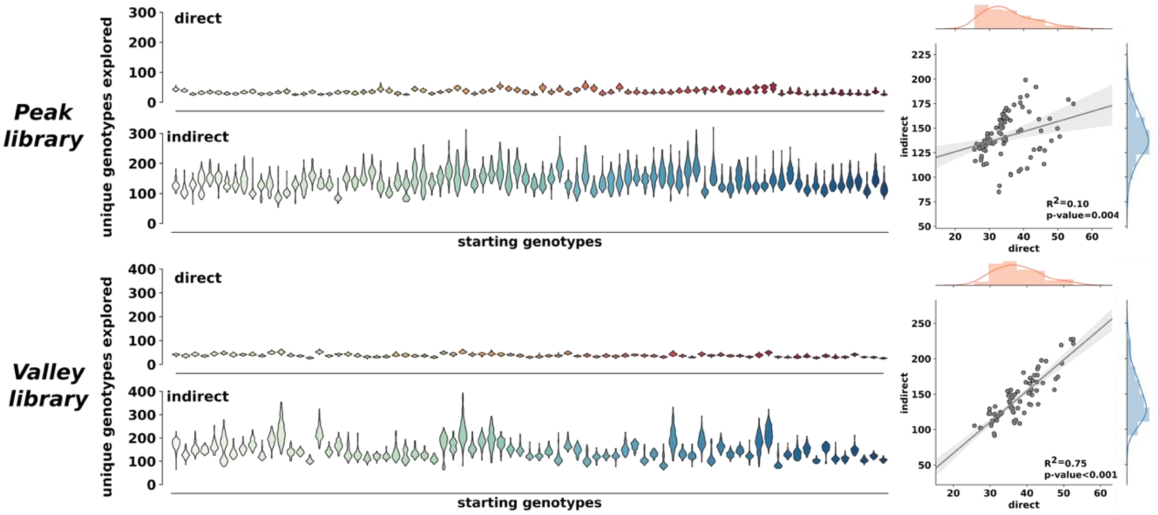


Figure 2.3-figure supplement 6: Unique genotypes explored during adaptive evolution on direct and indirect pathway landscapes.

The number of unique genotypes explored during 1000 generations of adaptive evolution on the direct (orange-red) and indirect (blue-green) landscapes of the *Peak* and *Valley* library. Each plot represents the distribution of 100 replicates of simulated evolution for each unique starting genotype. The distributions are ordered by the mean final fitness on the direct landscape and the corresponding simulations on the indirect landscape are shown directly below. The correlation of the number of unique genotypes explored from the direct and indirect landscapes for each of the starting points is also depicted.

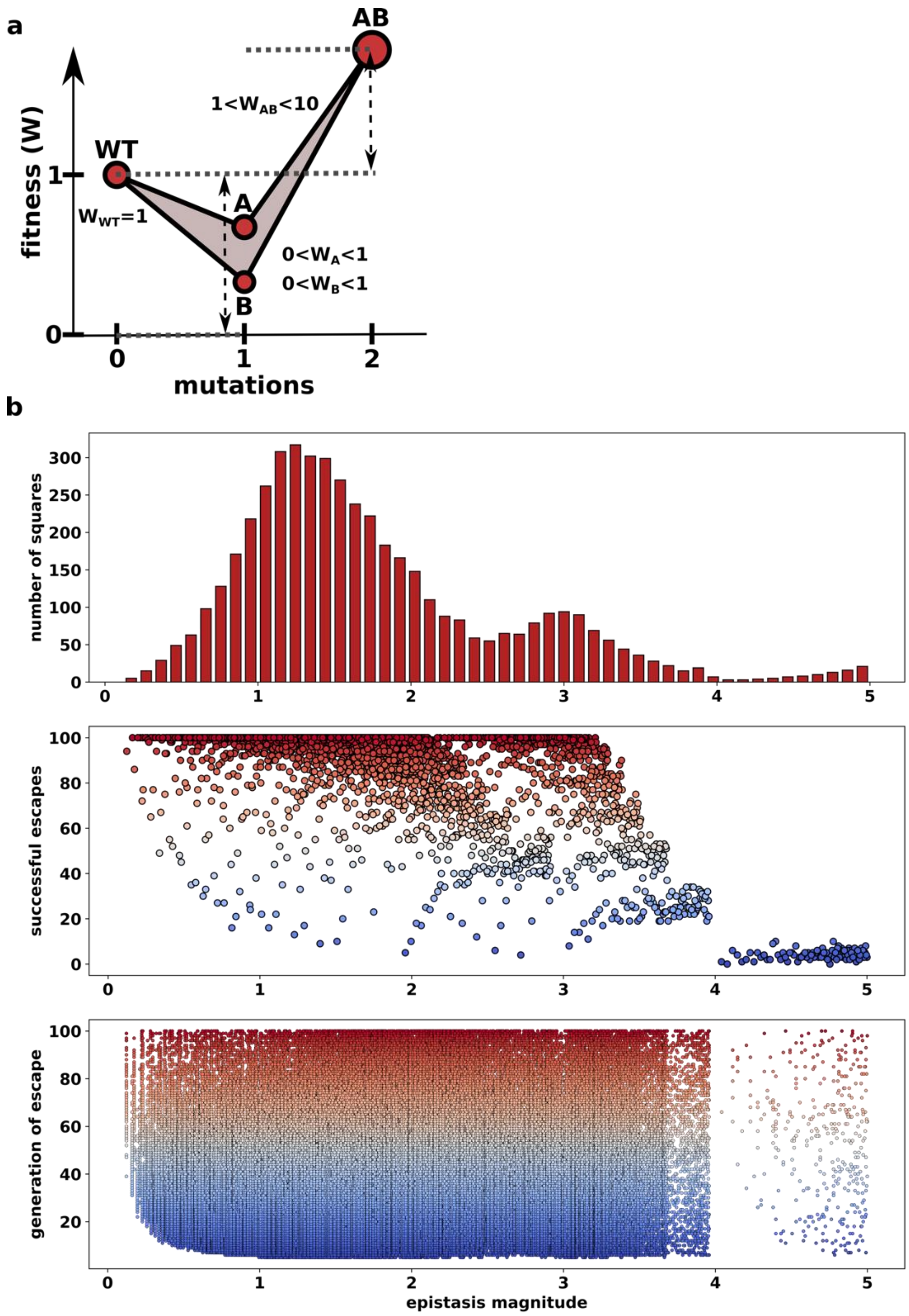


Figure 2.3-figure supplement 7: Escape from stasis genotypes isolated by reciprocal sign epistasis.

(a) Model depicting the mutational pairs (squares) that are used in the simulated evolution. The wild-type (WT) sequence had a constant fitness=1, while the fitness of single mutant genotypes A and B fluctuated between 0 and 1. The fitness of the double mutant (AB) fluctuated between 1 and 10. (b) Top graph depicts the distribution of epistatic values explored during the simulations. Middle graph depicts the total number of ‘successful’ escapes from the stasis genotype (WT) out of 100 replicates as a function of epistasis magnitude. Bottom graph displays the generation of each ‘successful’ escape.

CHAPTER THREE: RESURRECTION OF AN RNA PHYLOGENETIC FITNESS LANDSCAPE

Devin P. Bendixsen[1], Tanner B. Pollock[2] and Eric J. Hayden[1,2]
[1]Biomolecular Sciences Graduate Programs, Boise State University, Boise, ID, USA.
[2]Department of Biological Science, Boise State University, Boise, ID, USA.

Abstract

Phylogenetic analyses combined with ancestral sequence reconstruction and a high-throughput mutational assay can resurrect a phylogenetic fitness landscape. By predicting, ‘resurrecting’ and functionally characterizing ancient gene sequences, hypotheses about gene function or selection can be empirically tested in an evolutionary context. Recent advances in DNA synthesis and next-generation sequencing allow for the high-throughput assessment of activity for thousands of sequence variants. Here, we report an experimentally resurrected and reconstructed phylogenetic fitness landscape for the naturally occurring CPEB3 self-cleaving RNA enzyme (ribozyme). This ribozyme is highly-conserved in mammals and has been associated with episodic memory. We found that a single high-activity *ancestral* sequence was highly conserved and purifying selection is expected to have reduced the accumulation of mutations through geologic time. Many of the extant mammalian ribozyme sequences had high ribozyme activity, however a few had relatively low activity. Yet, given the local fitness landscape, a selective pressure for functional ribozyme sequences was seen. We showed that the

single nucleotide polymorphism (SNP) found in humans reduced co-transcriptional ribozyme activity *in vitro* and might alter our understanding of the CPEB3 ribozyme's biological function.

Introduction

Phylogenetic analyses of genes or gene products (RNA, protein) yield valuable insights into evolutionary theory, processes and mechanisms. Extant sequences from a range of taxa can be used to predict the sequences of ancestral genes (Fig. 3.3.1a). It is an important application of phylogenetics and has been widely used to test hypotheses about gene function and structure (Fournier and Alm 2015; Natarajan et al. 2016; Stern et al. 2017) or to guide the design of novel biomolecules (Zakas et al. 2017; Alva and Lupas 2018). In addition to ancestral sequence prediction, phylogenetics can be used to recover a measurable characteristic (phenotype) of a sequence (Joy et al. 2016). However, the accuracy of predicted phenotypes (i.e. activity or gene function) for ancestral sequences is uncertain and often difficult to validate. Ancestral sequences that code for gene products, such as protein and RNA enzymes (ribozymes), can be readily assessed and characterized resulting in accurate fitness measurements. By predicting, 'resurrecting' and functionally characterizing ancient gene sequences, hypotheses about gene function or selection can be empirically tested in an evolutionary context (Thornton 2004).

Recent advances in DNA synthesis and next-generation sequencing allow for the high-throughput assessment of activity or fitness for thousands of sequence variants (Dupont et al. 2015; Hayden 2016; Kobori and Yokobayashi 2016). The activity or fitness measurements (phenotype) for each sequence (genotype) can be constructed into an empirical fitness landscape with valuable evolutionary insights (Fig. 3.3.1b). Fitness landscapes are a classical approach for visualizing the relationship between genotype and phenotype (McCandlish 2011) and can be used to determine the accessibility of mutational pathways or even predict or forecast evolution (Kogenaru et al. 2009; Franke et al. 2011; Lobkovsky et al. 2011; de Visser and Krug 2014). DNA libraries for the construction of fitness landscapes are often designed by identifying nucleotide positions of interest or variable nucleotide positions between two anchoring sequences. The library is then synthesized with equal probability of each variable nucleotide at those positions resulting in a library that contains the anchoring sequences, as well as the parsimonious intermediate between them. This allows for elucidation of the possible evolutionary trajectories or pathways in the immediate sequence space (Hayden 2016). By combining this approach with ancestral sequence resurrection, we can use extant sequences as anchors and design a library that contains not only extant sequences and their predicted ancestors, but also every other combination of the mutations that differ between the chosen extant sequences. This novel approach can provide new insight into gene function, conservation and evolutionary selection pressure.

Here, we report an experimentally resurrected and reconstructed phylogenetic fitness landscape for the naturally occurring CPEB3 self-cleaving RNA enzyme (ribozyme). This ribozyme is found in an intron of the cytoplasmic polyadenylation element-binding 3 gene (*CPEB3*). The ribozyme folds into a complex nested double-pseudoknot secondary structure (Fig. 3.3.1c), very similar to the Hepatitis delta virus (HDV) ribozyme, and exhibits 5’ co-transcriptional self-cleavage (Salehi-Ashtiani et al. 2006; Chadalavada et al. 2010; Webb and Lupták 2011; Skilandat et al. 2016). The 67-nt core ribozyme is located in the second intron of the human *CPEB3* (*hsCPEB3*) gene that codes for a functional prion with an important role in synaptic plasticity and long-term memory (Fig. 3.3.1d, Stephan et al. 2015). This ribozyme is conserved in mammals and has been associated with episodic memory (Vogler et al. 2009; Webb and Lupták 2011). Alignment and comparison of the ribozyme sequences identified across mammalian taxa shows a high level of sequence conservation as compared to the human CPEB3 sequence (Fig. 3.3.1d). The true role of the CPEB3 ribozyme is unclear, however hypotheses suggest that it might play an important role in co-transcriptional processing of the CPEB3 pre-mRNA (Webb and Lupták 2011). For our experiments reported here, we chose CPEB3 ribozyme sequences from 25 mammalian species that differed at 13 mutational positions and designed a DNA library that encompassed 27,648 genotypes containing all combinations of these 13 mutations. We determined the relative activity for each unique sequence through the deep sequencing of in vitro co-transcriptional self-cleavage reactions. We used these data to reconstruct the phylogenetic fitness landscape of the CPEB3 ribozyme, offering a novel glimpse into the evolution of this ribozyme. A phylogenetic fitness landscape encompasses not only extant sequences and predicted ancestral sequences, but also the local genotype space, consisting of all local mutational combinations. Using high-throughput sequencing we obtained ribozyme activity measurements for all 27,648 unique genotypes, which includes all the extant ribozyme sequences found in 25 species of mammals and several predicted ancestral sequences.

Results and Discussion

CPEB3 Ribozyme Ancestral Sequence is Highly Conserved Through Time

By mapping the extant ribozyme sequences onto the tree of life, we predicted that the ancestral sequence has been completely conserved in 13 of the extant lineages (Fig.

3.3.2). Interestingly, we found that this ancestral sequence shows very high self-cleavage activity. The high level of conservation found throughout mammalian sequences is indicative of purifying selection, which is widespread in mammalian RNA structures (Smith et al. 2013). To test this hypothesis, a Tajima's neutrality test based on the alignment of extant ribozyme sequences was performed and resulted in a negative D-value (-1.34), which is indicative of a low frequency of mutations and polymorphisms (Tajima 1989). This is often due to population expansion following a bottleneck event or, as suspected in this case, purifying selection.

A Range of Ribozyme Activity is Found in Extant Ribozyme Sequences

We next sought to determine if high ribozyme activity was conserved throughout extant sequences suggesting a strong evolutionary selection for ribozyme function. Ribozyme activity was calculated as the fraction cleaved during the 20-minute co-transcriptional assay. We found that 21 of the 36 extant species contained ribozyme sequences with ribozyme activity ≥0.77 (Fig. 3.3.2). Interestingly, the remaining 15 extant species contained or were predicted to contain ribozyme sequences with ribozyme activity ≤0.33. Four of these ribozyme sequences were measured using high-throughput sequencing in the phylogenetic mutational library. Three of these ribozyme sequences were assessed using standard lab assays combined with denaturing gel electrophoresis to determine ribozyme activity. The remaining eight extant sequences were not measured, but are predicted to have low ribozyme activity due to mutations at the G1 position, which results in low ribozyme activity in our data and in previous reports (Salehi-Ashtiani et al. 2006; Webb and Lupták 2011).

Phylogenetic Fitness Landscape Revealed Selective Pressure

Next, we set out to use the activity measurements of all mutational combinations to better understand the fitness landscape that shaped the evolution of this ribozyme. (Fig. 3.3a). Each node in the fitness landscape indicates a unique ribozyme sequence and nodes that differ by a single mutation are connected by edges. Pairwise epistasis was prevalent in the landscape and was found in ~57% of all mutation pairs (Fig. S3.1). This is consistent with earlier reports on the amounts of epistasis in RNA fitness landscapes (Bendixsen et al. 2017). However, the severity of epistatic interactions was not extreme ($\varepsilon<1.6$), suggesting that evolution might have been able to traverse the landscape without high levels of stasis. We found that the ribozyme activities were not uniformly distributed across the activity range, but rather ~96% of all ribozyme sequences assessed had low ribozyme activity (≤0.2, Fig. S3.2). In contrast, less than 1% of sequences in the phylogenetic mutational library had high ribozyme activity (≥0.8). However, we find that 80% of extant species (20 of 25) in the mutational library have sequences with high ribozyme activity. Only three species have low ribozyme activity and interestingly, only two sequences (human and human-SNP) exhibit intermediate ribozyme activity ($0.2>\mu<0.8$). Of the three extant ribozyme sequences with low activity, giant panda (*Ailuropoda melanoleuca*) exhibited the lowest ribozyme activity (~0.03). Although this ribozyme activity is relatively low, it is still higher than ~61% (16,960) of sequences that were available to evolution in this local sequence space. This suggests that although the majority of sequences in the local sequence space have low ribozyme activity, evolution selected for many sequences that exhibit high or at least functional activity.

We then assessed the effects of the mutations in the 13 variable positions within the phylogenetic mutational library. We determined the average ribozyme activity for each nucleotide at each position independent of genetic background and mapped their effects to the ribozyme secondary structure (Fig. 3.3.2b, Fig. S3.3). Many of the mutations that were accumulated in extant species showed very little average effect on ribozyme activity. In particular, L4 of the ribozyme was very tolerant to mutations and consistently had no detectable impact on function. Not surprisingly and in agreement with previous studies, the most detrimental mutations occurred at position 1, which is directly upstream of the cleavage site (Salehi-Ashtiani et al. 2006; Webb and Lupták 2011). Notably, the nucleotides found in the highly-conserved, highly-active *ancestral* sequence were found to be optimal for 12 out of the 13 variable positions. The exception, position 9, was the only mutation that on average had a higher ribozyme activity when the ancestral C was mutated to a U. This mutation is found exclusively in marsupials and motivated a deeper exploration of the local fitness landscape found in marsupials (Fig. 3.3c). We identified the five mutational positions present in the marsupial clade and isolated ribozyme sequences that differ only at these positions. These sequences were then built into a fitness landscape with extant ribozyme sequences and the *ancestral* sequence in the inner circle and all other combinations in the outer circle. All of the 48 ribozyme sequences found in this landscape show high ribozyme activity. This is, however, expected because the four mutations other than position 9, are all in L4 which is highly robust. In contrast, using a similar approach for the primate clade results in a fitness landscape with very few sequences with high ribozyme activity (Fig. 3.3c). However, it should be noted that five of the seven mutations within this clade are found exclusively in the marmoset (*Callithrix jachus*) which has low ribozyme activity. Interestingly, a closer examination of the potential evolutionary trajectories from the ancestral sequence to the marmoset reveals that four of the five mutations incur ribozyme activity loss (Fig. S3.4). Next, we plotted the 27,648 unique sequences in the phylogenetic mutational library as a function of mutational distance from the *ancestral* sequence and ribozyme activity (Fig. 3.3d). We found that up to six mutations could be added to the *ancestral* sequence before a significant reduction in ribozyme activity. However, on average the ribozyme activity quickly decreased as the sequence mutated away from the *ancestral* sequence (dotted line). This suggests that the *ancestral* sequence is highly optimized and that there exists selective pressure to maintain ribozyme activity.

Single Nucleotide Polymorphism (SNP) in Humans Reduces Ribozyme Activity

An important and biologically relevant mutation within the CPEB3 ribozyme is a single nucleotide polymorphism (SNP) found in the human sequence at position 36. Humans that are homozygous in this SNP show poorer episodic memory. This nucleotide directly interacts with position 1 of the ribozyme which we found to be very sensitive to mutations. The wild-type human sequence contains a U at this position forming a non-canonical G•U wobble base pair. A subpopulation of humans contain a C at this position forming a canonical Watson-Crick G-C base pair. We found that the wild-type human sequence exhibited relatively high ribozyme activity (0.72), however the ribozyme sequence containing the SNP (U36C) exhibited significantly lower ribozyme activity (0.23). These two ribozyme sequence variants were ordered individually and validated in the lab using a co-transcriptional assay followed by denaturing gel electrophoresis (Fig. S3.5). A previous study had suggested that the SNP mutation increased ribozyme activity three-fold (Salehi-Ashtiani et al. 2006). However, this study assayed ribozyme activity following transcription after RNA isolation and purification, not co-transcriptionally as presented in this study. Another research group assessed the human wild-type ribozyme and found that it was co-transcriptionally fast-reacting (Chadalavada et al. 2010). In fact, they determined that the CPEB3 human ribozyme cleaves 50-fold faster than previously reported. This is not surprising because RNA structures have been shown to fold extensively during transcription and in some instances this significantly improves the efficiency of cleavage (Pan and Sosnick 2006). The HDV ribozyme, which has high amounts of similarity to CPEB3, has been shown to be highly dependent on co-transcriptional folding (Chadalavada et al. 2000; Diegelman-Parente and Bevilacqua 2002; Chadalavada et al. 2007). These data agree with our findings that the human wild-type ribozyme has high co-transcriptional ribozyme activity.

However, our findings still disagree with previous findings that introducing the U36C SNP mutation into the human ribozyme improves ribozyme activity (Salehi-Ashtiani et al. 2006). This might be explained by the difference in assay approaches (co-transcriptional or following purification) as previously mentioned. It might also be attributed to differences in the assay conditions, for example our study used a higher Mg^{2+} concentration (10mM) than was previously used (5mM). The wobble base pair is a binding site for $Mg^{2+}$and therefore might be sensitive to these differences (Skilandat et al. 2014; Skilandat et al. 2016). Recent investigations of Mg^{2+} interaction with the CPEB3 ribozyme suggested that Mg^{2+} is necessary for proper native folding and that the optimized Mg^{2+} concentration to correspond to physiological conditions is 10mM Mg^{2+} (Strulson et al. 2013; Skilandat et al. 2016). Although the physiological concentration of

Mg^{2+} is approximately 20 to 100-fold less (0.1-0.5mM) than the 10mM used in this study, 10mM Mg^{2+} mimics the molecular crowding found in cells. Therefore, the Mg^{2+} concentration used in our study most closely resembles the environmental conditions encountered by the ribozyme *in vivo*. Further supporting our findings that the U36C SNP mutations reduces activity in the human ribozyme are recent investigations on the HDV ribozyme. The HDV and CPEB3 ribozymes have a very similar secondary structure and exhibit many nucleotides or base pairs that are conserved (Salehi-Ashtiani et al. 2006; Skilandat et al. 2016). One base pair found in both wild-type ribozyme sequences is the G•U wobble base pair immediately downstream of the cleavage site. The G•U base pair in the HDV ribozyme had been shown to be important to ribozyme activity and when substituted with most base pair combinations caused significant activity loss (Wu et al. 1993; Perrotta and Been 1996). However, mutations that retain a base pair interaction and a purine at position 1 were tolerated (Cerrone-Szakal et al. 2008; Chen et al. 2009). A recent mutational analysis suggested that the G•U wobble base pair played a specific role in the stabilization of the active sites (Sripathi et al. 2015). The wobble base pair stabilizes high in-line conformation and the hydrogen bond between the G at position 1 and the catalytic nucleotide (C75). Although this data is not entirely conclusive, it does suggest that the G•U wobble base pair might play an important stabilization role in the CPEB3 ribozyme, similar to the HDV ribozyme.

The effects of the SNP mutation may have strong implications for human episodic memory. An association test found that individuals homozygous for the SNP mutation performed poorer in episodic memory tests (Vogler et al. 2009). The effect was particularly prevalent when the material had an emotional valence. At the time it was believed that the SNP mutation increased ribozyme activity and conclusions were drawn from these data. However, in this study we found that the SNP mutation actually reduces the co-transcriptional activity of the ribozyme by ~3 fold. Although the true function of the ribozyme in human memory is unclear, it was proposed that the ribozyme might play a role in co-transcriptional processing of the CPEB3 pre-mRNA (Webb and Lupták 2011). Rapid ribozyme self-cleavage may prevent further processing of the pre-mRNA before the next exon is synthesized and tagged for splicing. Due to the fast co-transcriptional CPEB3 ribozyme activity in many mammalian species it is believed that it can affect the stability of the pre-mRNA on timescales relevant to gene expression (Webb and Lupták 2011). Given the findings in this study it can be proposed that human episodic memory improves with the rate of co-transcriptional ribozyme activity, although the mechanism needs further exploration.

Random Mutagenic Library Identified Key Nucleotides for Ribozyme Activity

The phylogenetic mutational library of the CPEB3 ribozyme offered valuable insight into the local fitness landscape encountered by evolution. However, due to limitations in sequencing capabilities not all of the mutations that exist in extant species were able to be exhaustively explored. To identify other key nucleotides that had a significant impact on co-transcriptional ribozyme activity, we designed a random mutagenic library (18% mutation rate per position). This library contained a distribution of random mutations from the highly-conserved highly active *ancestral* sequence (Fig. 3.4a, grey). In total >30 million unique sequences were found in the library and only <90,000 were found to be cleaved. This low level of active ribozyme sequences was expected due to the distribution of mutations in the library being centered around approximately 15 mutations away from the *ancestral* sequence. As expected, the population of cleaved sequences showed a slight shift towards the *ancestral* ribozyme sequence as compared to the library (Fig. 3.4a, teal). By comparing the sequences that were found to be cleaved against the mutagenic library we were able to identify several nucleotides that were highly conserved among cleaved sequences (Fig. 3.4b). Importantly the catalytic nucleotide (C57) was highly conserved among cleaved sequences. This agrees with our prediction that extant species with a mutation at this location (*Dasypus novemcinctus*, *Felis catus*, *Myotis lucifigus* and *Dipodomys ordii*) have low ribozyme activity. The first three nucleotides in the ribozyme sequence which form the P1 stem were found also found to be highly conserved. This was expected due to the catalytic importance of the stability of the P1 helical structure (Salehi-Ashtiani et al. 2006; Webb and Lupták 2011) and further supports our observation and prediction that extant sequences with mutations at these positions have low activity. The only mutational difference between the *ancestral* sequence and the human wild-type ribozyme sequence is a G30A mutation. Interestingly the G found in the *ancestral* sequence was enriched, which agrees with our findings that this mutation reduces ribozyme activity by ~20%. We also found that at position 36 of the ribozyme activity the human wild-type nucleotide was enriched as compared to the human-SNP U36C mutation. This further supports our finding that the SNP mutation reduces ribozyme activity in humans. Three other notable conserved nucleotides (positions 10,11 and 67) are found in the P2 stem structure, which is important for stability of the ribozyme structure.

Conclusion

Using a high-throughput mutational assay and ancestral sequence resurrection, we resurrected and reconstructed a phylogenetic fitness landscape for the CPEB3 ribozyme. We found that a single *ancestral* sequence was highly conserved and purifying selection is expected to have reduced the accumulation of mutations through geologic time. However, high ribozyme activity was not universally conserved in extant mammalian ribozyme sequences. Yet, given the local fitness landscape encountered by evolution, a selective pressure for functional ribozyme sequences was seen. Due to the proposed important role of the CPEB3 ribozyme in episodic memory, it is possible that *in vivo* co-evolved regulatory elements in each mammalian species help to account for reduced measured ribozyme activity. We showed that the single nucleotide polymorphism (SNP) found in humans reduces co-transcriptional ribozyme activity *in vitro* and might alter our understanding of the CPEB3 ribozyme's biological function. In agreement with previous findings, we also identified key nucleotides that are responsible for ribozyme activity. The biochemical assessment of extant and ancestral CPEB3 ribozyme variants, coupled with the characterization of the local fitness landscape motivates further research into the true biological function of the CPEB3 ribozyme.

Materials and Methods

Mammalian CPEB3 Ribozyme Phylogenetic Mutational Library Design

In order to build a phylogenetic mutational library, the 36 identified mammalian CPEB3 ribozyme sequences (Webb and Lupták 2011) were aligned and 13 mutational positions were identified that maximized phylogenetic coverage. For this study, only the mutations that occurred in the length of the ribozyme (67nt) were considered. Of the 36

CPEB3 ribozyme sequences, 25 species had only differences within these 13 mutational positions. Of the 25 species in this group, 13 of these species had the same identical ribozyme sequence. Library construction was accomplished by chemically synthesizing a degenerate DNA oligonucleotide that served as temple for *in vitro* transcription. At each of the 13 identified mutational positions, only nucleotides that were present in the group were considered. At each mutational position, the DNA library was synthesized with equal mixtures of two or three nucleotide phosphoramidites, generating approximately equal likelihood of each sequence variant. At 10 of the mutational positions only two nucleotides were allowed, and at three of the mutational positions only three nucleotides were allowed. This resulted in a library that consisted of 27,648 ($2^{10} \bullet 3^{3}$) unique sequences. This library consisted of the CPEB3 ribozyme sequences for 25 mammalian species, as well as all possible intermediates and combinations. For assay purposes a T7 promoter sequence was added to the end of the ribozyme sequence and a common sequence was added to the 3’-end to act as a universal primer binding site during reverse transcription (Wilkinson et al. 2006).

CPEB3 Ribozyme Random Mutagenesis Library Design

In order to identify and assess key nucleotides for ribozyme activity that were not included in the phylogenetic fitness landscape library, we designed a randomly mutagenic library. At each position in the CPEB3 ribozyme (length=67), the DNA library was synthesized with an 82:6:6:6 mixture of all four nucleotide phosphoramidites. At each position the nucleotide found in the wild-type human sequence was maintained at 82%. This generated a synthetic DNA library with a random mutation rate of ~18%. For the high-throughput assay a T7 promoter sequence was added to the end of the ribozyme sequence and a common sequence was added to the 3'-end to act as a universal primer binding site during reverse transcription (Wilkinson et al. 2006).

Co-Transcriptional Self-Cleavage Assay

Both CPEB3 mutational libraries (phylogenetic and random) were assayed entirely in triplicate yielding three biological replicates. Each replicate for each library were prepared in the same manner. The ssDNA template library was annealed with the T7-TOP+ primer by heating 20 picomoles of template and primer in custom T7 Mg10 buffer (10 µL 1M Tris pH 7.5, 50 µL 1M DTT, 20µL 1M spermidine, 100 µL 1M $MgCl_2$, 300 µL RNase-free water). The template and primer were then diluted 10-fold. 2 µL of template and primer were then transcribed in vitro in a 50 µL reaction with 5 µL T7 Mg10 buffer, 1 µL rNTP (25 mM, NEB), 1 µL T7 RNA polymerase (200 units, Thermo Scientific) and 41 µL RNase free water (Ambion) at 37°C for 20 mins. The transcription and co-transcriptional self-cleavage reaction was then terminated by adding 15 µL of 50 mM EDTA. The total amount of cleaved RNA increases during transcription, however the ratio of cleaved to uncleaved remains approximately the same, as long as the rate of transcription is constant. This holds true especially for moderately short transcription times before reagents become limited (Long and Uhlenbeck 1994). The transcription reaction was then cleaned and concentrated with Direct-zol RNA MicroPrep w/ TRI-Reagent (Zymo Research) to 7 µL. We then determined the concentration of the RNA sample using a spectrophotometer (ThermoFisher NanoDrop) and the samples were normalized to 5 µM. The cleaned RNA (5 picomoles) was mixed with 20 picomoles of reverse transcription primer in a volume of 10 µL and was heated at 72 °C for 3 mins and then cooled on ice. 4 µL SMARTScribe 5x First-Strand Buffer (Clontech), 2 µL dNTP

(10 mM), 2 μL DTT (20 mM), 2 μL phased template switching oligo mix (10 μM), 1 μL water and 1 μL SMARTScribe Reverse Transcriptase (10 units, Clontech) were then added to the RNA template and RT primer. The phased template switching oligo mix consisted of four oligonucleotides that were phased by the addition of 9, 12, 15 or 18 nucleotides. The mixture was then incubated at 42 °C for 90 mins. The reverse transcription reaction was stopped and the RNA degraded by heating the sample to 72 °C for 15 mins. The cDNA was then purified using DNA Clean & Concentrator-5 (Zymo Research) and eluted into 7 μL water. To prepare the mutational library for high-throughput sequencing, Illumina adapter sequences were added to the ends of the cDNA using PCR. Each of the replicates were assigned a unique combination of barcodes. The PCR reaction contained 1 μL purified cDNA, 12.5 μL KAPA HiFi HotStart ReadyMix (2X, KAPA Biosystems), 2.5 μL forward, 2.5 μL reverse primer (Illumina Nextera Index Kit) and 5 μL water. To ensure that the PCR didn't introduce bias, multiple cycles of PCR were examined using gel electrophoresis an appropriate PCR cycle was chosen that was still in linear amplification. Each PCR cycle consisted of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 30 s. The PCR cDNA product was then cleaned using DNA Clean & Concentrator-5 (Zymo Research) and eluted in 30 μL water. The final product was then verified using gel electrophoresis.

High-Throughput Sequencing

The three replicates for each mutational library (phylogenetic and random) were pooled together and sent to the University of Oregon Genomics and Cell Characterization Core Facility. The two libraries were sequenced using an Illumina HiSeq 4000 on separate lanes. The random library was pooled along with other samples not presented here. For each lane 25% PhiX was added to increase the nucleotide diversity during sequencing. The phylogenetic library yielded ~170 million reads passing filter with a mean quality score of 39.6. The random mutagenic library yielded ~57 million reads passing filter with a mean quality score of 39.1. Replicates were correlated with one another with a high amount of correlation (Fig. S3.6). The standard deviation (delta) of replicates were also calculated to be used in analyses (Fig. S3.7)

Sequencing Data Analysis

Sequencing data were analyzed using custom Python scripts. The scripts identified the universally conserved 3' handle and determined if the sequence was cleaved or uncleaved. For the phylogenetic mutational library, the nucleotides at the 13 variable mutational positions were then isolated. For each unique genotype in the library the number of cleaved and uncleaved sequences were counted. For each genotype, ribozyme activity (fraction cleaved) was calculated as: $f_{cleaved} = \frac{n_{cleaved}}{(n_{cleaved}+n_{uncleaved})}$. For the random mutagenesis library, the cleaved sequences were binned. For each genotype in the entire library (cleaved and uncleaved) the mutational distance (Hamming distance) was calculated between the genotype and the highly-conserved *ancestral* genotype. The distribution of mutational distances in the cleaved bin was compared to the entire library were calculated (Fig. 3.4a). For each position (p) in the CPEB3 ribozyme (length=67), fold enrichment in the cleaved bin was calculated as: $p_{enrichment} = \frac{\%cleaved_{ancestral}}{\%library_{ancestral}} = \frac{cleaved_{ancestral}/cleaved_{total}}{library_{ancestral}/library_{total}}$, where $x_{\text{ancestral}}$ indicates the number of sequences in the cleaved or library bin that have the *ancestral* nucleotide at that position.

CPEB3 Ribozyme Phylogenetic Tree Construction and Ancestral Sequence Resurrection

For the 35 mammalian species with known CPEB3 ribozyme sequences (Webb and Lupták 2011), we constructed a phylogenetic tree based on the tree-of-life and its evolutionary timescale (Hedges et al. 2015). We loaded the 35 species into TimeTree and generated a phylogenetic tree that showed the 35 extant species and ancestral progenitors in the class Mammalia (Kumar et al. 2017). Using this phylogenetic tree we used Molecular Evolutionary Genetics Analysis 7 (MEGA7) software to infer ancestral sequences (Kumar et al. 2016). Two statistical methods of inference (maximum likelihood and parsimony) yielded the same predicted ancestral sequences. All predicted sites were found to have >0.9 maximum probability. This process predicted the CPEB3 ribozyme sequences at the ancestral nodes within the phylogenetic tree. Mutations were added to edges connecting nodes in order to allow for easier interpretation. Node colors were determined by co-transcriptional self-cleavage assay followed by high-throughput sequencing or denaturing gel electrophoresis. For a subset of nodes, ribozyme activity was predicted based on known mutations in the sequence that result in low activity.

Figures

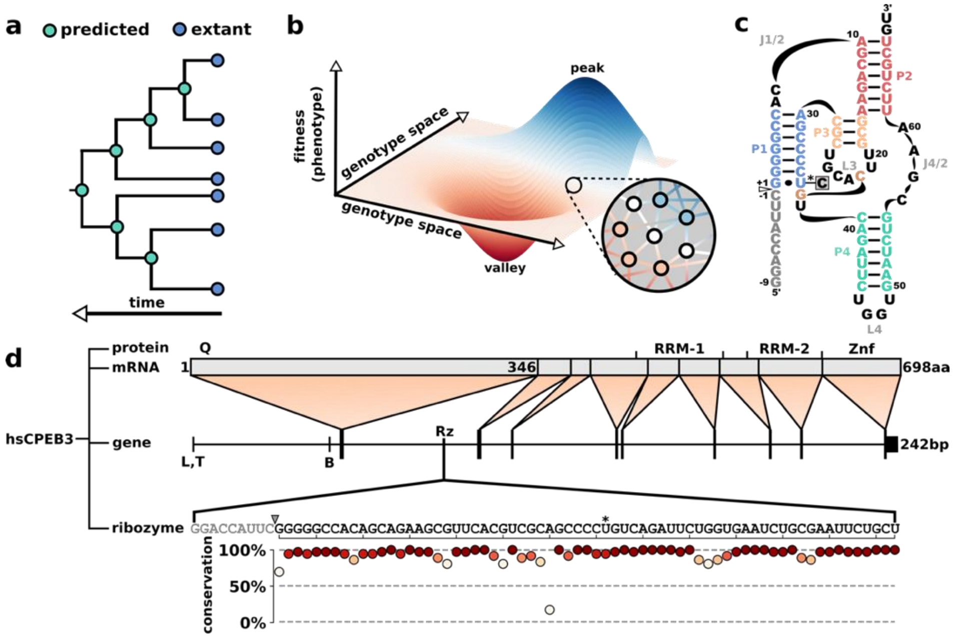


Figure 3.1: Overview of ancestral sequence resurrection, fitness landscapes and CPEB3 ribozyme.

(a) Simplified example of ancestral sequence resurrection. The tree is built based on the phylogenetic relationship between species containing extant sequences (blue). Statistical probability algorithms are then used to infer or predict ancestral sequences (green). (b) Simplified example of an empirical fitness landscape. Each node indicates a unique sequence (genotype) and nodes that differ by a single mutation are connected by edges. The x- and y-axes indicate the relative position in genotype sequence space and the z-axis indicates the relative fitness (phenotype). Node colors are indicative of the relative fitness of the given genotype. (c) Secondary structure of human (hs) CPEB3 ribozyme sequence. Triangle indicates self-cleavage site and grey letters indicate cleaved sequence. Base paired helices are distinguished by color. Asterisk indicates the location of a SNP (U36C) in the human ribozyme sequence. (d) Mapping and conservation of human (hs) CPEB3 ribozyme. Protein, mRNA and gene are adapted from Salehi-Ashtiani et al. 2006. Four notable domains are identified in the protein primary structure (Q=glutamine-rich domain, RRM=RNA-binding domains, Znf= zinc finger). Vertical dividers in the mRNA indicate splice sites. Tissue-specific untranslated exons are marked below the gene with letters (L=liver, T=testis, B=brain). Translated exons are indicated with large vertical lines. Self-cleaving CPEB3 ribozyme location is indicated as Rz in the second intron and the human CPEB3 sequence is shown. Grey triangle indicates self-cleavage site and grey letters indicate cleaved sequence. Asterisk indicates the location of a single nucleotide polymorphism (SNP) in the human CPEB3 ribozyme sequence. Plot indicates the conservation of each nucleotide in the human sequence as compared to the 36 identified mammalian CPEB3 sequences (Webb and Lupták 2011). Color indicates level of conservation.

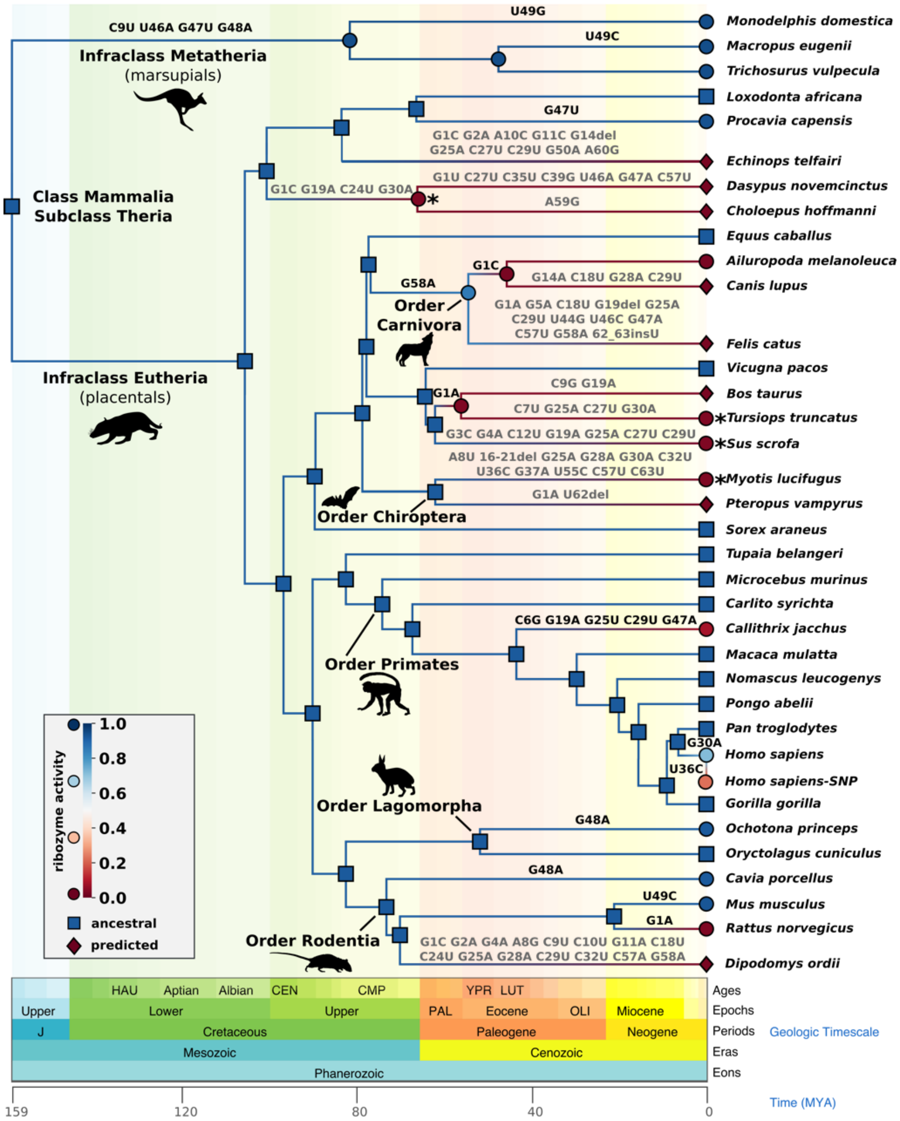


Figure 3.2: Phylogenetic tree and ancestral sequence resurrection of the mammalian CPEB3 ribozyme.

Phylogenetic tree derived from the 35 mammalian species with identified CPEB3 ribozyme sequences. Extant species are listed on the right. Each node indicates a ribozyme sequence that is either found in an extant species (right) or represents a predicted ancestral sequence. The color of the node indicates the self-cleaving ribozyme activity. Square nodes indicate a single highly-functional, highly-conserved *ancestral* sequence. Circle nodes indicate ribozyme sequences that were biochemically assessed using high-throughput sequencing. Circle nodes with an asterisk were assessed using co-transcriptional assay and denaturing gel electrophoresis. Diamond nodes indicate a sequence with predicted ribozyme activity based on known effects of mutations. Mutations listed above edges connecting nodes indicate the required mutations to convert from one ribozyme sequence to another. Black mutations indicate mutations that were included in the phylogenetic ribozyme library for high-throughput sequencing. Background color is colored by the Ages of the geologic timescale.

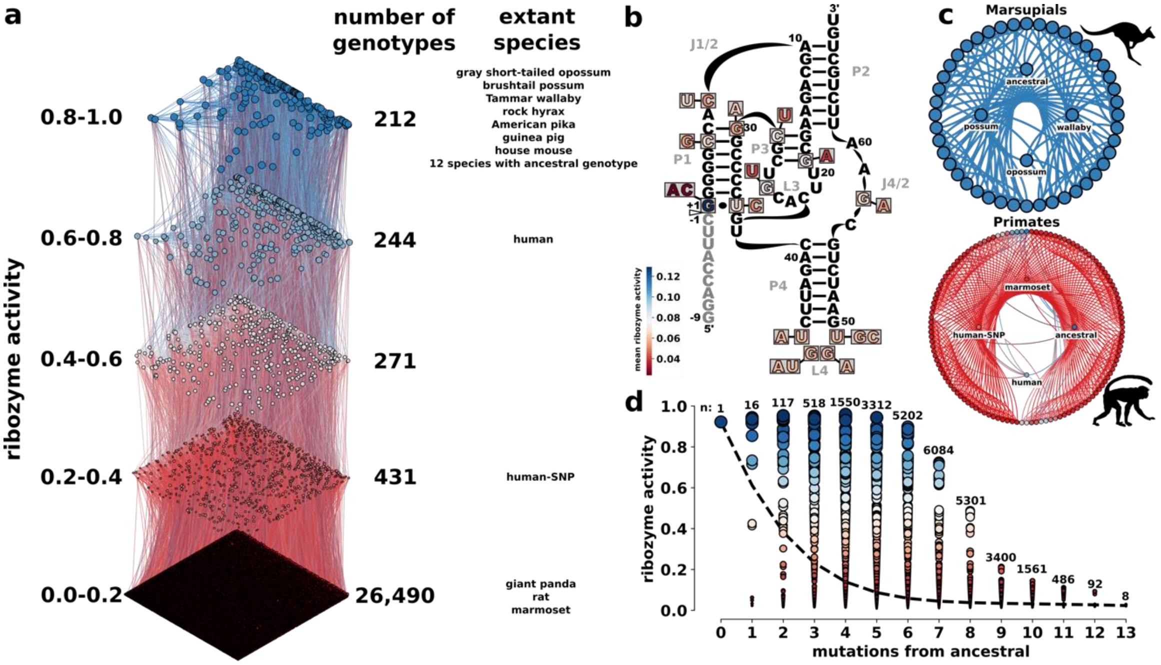


Figure 3.3: Phylogenetic fitness landscape of the mammalian CPEB3 ribozyme.

(a) Fitness landscape of the 27,648 unique ribozyme sequences in the phylogenetic mutational library. Each node indicates a unique sequence and nodes that differ by a single mutation are connected by an edge. The node size and color indicate the self-cleaving ribozyme activity of each sequence. For visualization purposes the fitness landscape is broken into five levels of ribozyme activity. The respective number of genotypes and a list of extant species found in each of the five levels are shown. (b) Secondary structure of the *ancestral* CPEB3 ribozyme with mutational nucleotides indicated. Each nucleotide at each mutational position is colored by the mean activity of all ribozyme sequences with that nucleotide. (c) Local fitness landscapes of two phylogenetic clades: marsupials and primates. Each node indicates a unique ribozyme sequence and two nodes are connected by an edge if they can be interconverted by a single mutation. Node color indicates the relative ribozyme activity. Nodes in the inner circle represent sequences identified in an extant species or the highly-conserved *ancestral* sequence. Nodes in the outer cycle indicate ribozyme sequences with combinations of mutations that have not been identified in extant species or predicted as an ancestral sequence. The combination of mutations is limited to mutations that are present in the clade. (d) Ribozyme activity of all 27,648 sequences in the phylogenetic mutational library plotted as a function of mutations from the highly-conserved *ancestral* sequence. Each node indicates a unique ribozyme sequence and the color and size of the node indicate ribozyme activity. The number of sequences (n) that correspond to each mutational distance from the *ancestral* sequence is shown. The dashed line indicates the average at each mutational distance.

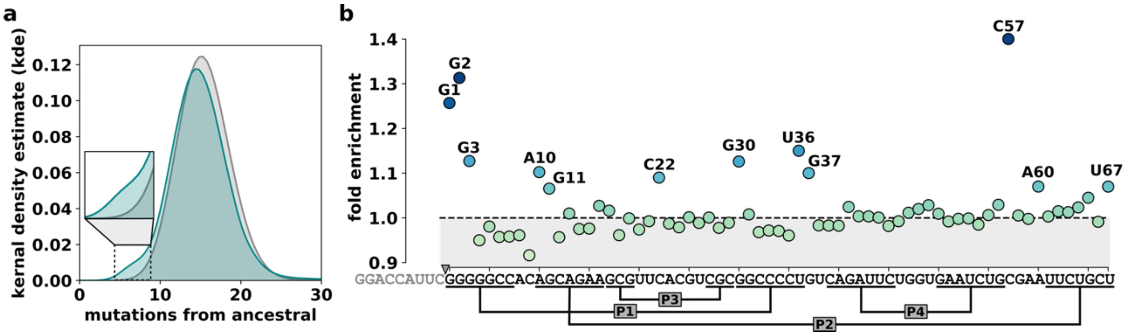


Figure 3.4: Random mutagenesis library of mammalian CPEB3 ribozyme.

(a) Distribution of mutational distances from the highly-conserved *ancestral* CPEB3 ribozyme sequence. Sequences that were detected as cleaved (teal) are shown compared to the distribution of the total mutagenic library (cleaved and uncleaved sequences, grey). (b) The fold enrichment for each nucleotide in the *ancestral* CPEB3 ribozyme sequence between the cleaved sequences and the library (cleaved and uncleaved). A fold enrichment value of 1.0 indicates that the nucleotide was not enriched and was equally represented in the cleaved population as compared to the library. A fold enrichment value of 1.3 indicates that the nucleotide was represented in the cleaved population at a level 30% higher than expected, given its proportion in the mutagenic library. The top twelve enriched or conserved nucleotides in the cleaved population are labeled. Node color indicates the level of enrichment. Grey triangle indicates the cleavage site and grey letters indicate the cleaved sequence. Base paired helical elements are shown and labeled as P1-P4.

References

Aguilar-Rodríguez J, Payne JL, Wagner A. 2017. A thousand empirical adaptive landscapes and their navigability. Nature Ecology & Evolution 1:0045.

Alva V, Lupas AN. 2018. From ancestral peptides to designed proteins. Current Opinion in Structural Biology 48:103–109.

Bendixsen DP, Østman B, Hayden EJ. 2017. Negative Epistasis in Experimental RNA Fitness Landscapes. J Mol Evol:1–10.

Cerrone-Szakal AL, Chadalavada DM, Golden BL, Bevilacqua PC. 2008. Mechanistic characterization of the HDV genomic ribozyme: The cleavage site base pair plays a structural role in facilitating catalysis. RNA 14:1746–1760.

Chadalavada DM, Cerrone-Szakal AL, Bevilacqua PC. 2007. Wild-type is the optimal sequence of the HDV ribozyme under cotranscriptional conditions. RNA 13:2189–2201.

Chadalavada DM, Gratton EA, Bevilacqua PC. 2010. The Human HDV-like CPEB3 Ribozyme Is Intrinsically Fast-Reacting. Biochemistry 49:5321–5330.

Chadalavada DM, Knudsen SM, Nakano S, Bevilacqua PC. 2000. A role for upstream RNA structure in facilitating the catalytic fold of the genomic hepatitis delta virus ribozyme. J. Mol. Biol. 301:349–367.

Chen J-H, Gong B, Bevilacqua PC, Carey PR, Golden BL. 2009. A catalytic metal ion interacts with the cleavage Site G.U wobble in the HDV ribozyme. Biochemistry 48:1498–1507.

Diegelman-Parente A, Bevilacqua PC. 2002. A mechanistic framework for co-transcriptional folding of the HDV genomic ribozyme in the presence of downstream sequence. J. Mol. Biol. 324:1–16.

Dupont DM, Larsen N, Jensen JK, Andreasen PA, Kjems J. 2015. Characterisation of aptamer-target interactions by branched selection and high-throughput sequencing of SELEX pools. Nucleic Acids Res. 43:e139.

Fournier GP, Alm EJ. 2015. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. J Mol Evol 80:171–185.

Franke J, Klözer A, Visser JAGM de, Krug J. 2011. Evolutionary Accessibility of Mutational Pathways. PLOS Computational Biology 7:e1002134.

Hayden EJ. 2016. Empirical analysis of RNA robustness and evolution using high-throughput sequencing of ribozyme reactions. Methods 106:97–104.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of Life Reveals Clock-Like Speciation and Diversification. Mol Biol Evol 32:835–845.

Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AFY. 2016. Ancestral Reconstruction. PLOS Computational Biology 12:e1004763.

Khalid F, Aguilar-Rodríguez J, Wagner A, Payne JL. 2016. Genonets server—a web server for the construction, analysis and visualization of genotype networks. Nucl Acids Res 44:W70–W76.

Kobori S, Yokobayashi Y. 2016. High-Throughput Mutational Analysis of a Twister Ribozyme. Angew. Chem. Int. Ed. 55:10354–10357.

Kogenaru M, Vos MGJ de, Tans SJ. 2009. Revealing evolutionary pathways by fitness landscape reconstruction. Critical Reviews in Biochemistry and Molecular Biology 44:169–174.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34:1812–1819.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33:1870–1874.

Lobkovsky AE, Wolf YI, Koonin EV. 2011. Predictability of evolutionary trajectories in fitness landscapes. PLoS Comput. Biol. 7:e1002302.

Long DM, Uhlenbeck OC. 1994. Kinetic characterization of intramolecular and intermolecular hammerhead RNAs with stem II deletions. Proc Natl Acad Sci U S A 91:6977–6981.

McCandlish DM. 2011. Visualizing Fitness Landscapes. Evolution 65:1544–1558.

Natarajan C, Hoffmann FG, Weber RE, Fago A, Witt CC, Storz JF. 2016. Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. Science 354:336–339.

Pan T, Sosnick T. 2006. RNA folding during transcription. Annu Rev Biophys Biomol Struct 35:161–175.

Perrotta AT, Been MD. 1996. Core sequences and a cleavage site wobble pair required for HDV antigenomic ribozyme self-cleavage. Nucleic Acids Res 24:1314–1321.

Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW. 2006. A Genomewide Search for Ribozymes Reveals an HDV-Like Sequence in the Human CPEB3 Gene. Science 313:1788–1792.

Skilandat M, Rowinska-Zyrek M, Sigel RKO. 2014. Solution structure and metal ion binding sites of the human CPEB3 ribozyme’s P4 domain. J Biol Inorg Chem 19:903–912.

Skilandat M, Rowinska-Zyrek M, Sigel RKO. 2016. Secondary structure confirmation and localization of Mg2+ ions in the mammalian CPEB3 ribozyme. RNA 22:750–763.

Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res 41:8220–8236.

Sripathi KN, Banáš P, Réblová K, Šponer J, Otyepka M, Walter NG. 2015. Wobble pairs of the HDV ribozyme play specific roles in stabilization of active site dynamics. Phys Chem Chem Phys 17:5887–5900.

Stephan JS, Fioriti L, Lamba N, Colnaghi L, Karl K, Derkatch IL, Kandel ER. 2015. The CPEB3 Protein Is a Functional Prion that Interacts with the Actin Cytoskeleton. Cell Rep 11:1772–1785.

Stern A, Yeh MT, Zinger T, Smith M, Wright C, Ling G, Nielsen R, Macadam A, Andino R. 2017. The Evolutionary Pathway to Virulence of an RNA Virus. Cell 169:35-46.e19.

Strulson CA, Yennawar NH, Rambo RP, Bevilacqua PC. 2013. Molecular Crowding Favors Reactivity of a Human Ribozyme Under Physiological Ionic Conditions. Biochemistry 52:8187–8197.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nature Reviews Genetics 5:366–375.

de Visser JAGM, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet 15:480–490.

Vogler C, Spalek K, Aerni A, Demougin P, Müller A, Huynh K-D, Papassotiropoulos A, de Quervain DJ-F. 2009. CPEB3 is Associated with Human Episodic Memory. Front Behav Neurosci [Internet] 3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2691156/

Webb C-HT, Lupták A. 2011. HDV-like self-cleaving ribozymes. RNA Biol 8:719–727.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat Protoc 1:1610–1616.

Wu HN, Lee JY, Huang HW, Huang YS, Hsueh TG. 1993. Mutagenesis analysis of a hepatitis delta virus genomic ribozyme. Nucleic Acids Res 21:4193–4199.

Zakas PM, Brown HC, Knight K, Meeks SL, Spencer HT, Gaucher EA, Doering CB. 2017. Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. Nature Biotechnology 35:35–37.

Acknowledgments

This work was supported by Boise State University (Biomolecular Sciences Graduate Programs), National Science Foundation Directorate for Biological Sciences (grant number MCB-1413664) National Science Foundation Office of Integrative Activities (grant number OIA-1738865) and National Aeronautics and Space Administration (grant number 80NSSC17K0738).

Supplementary Material

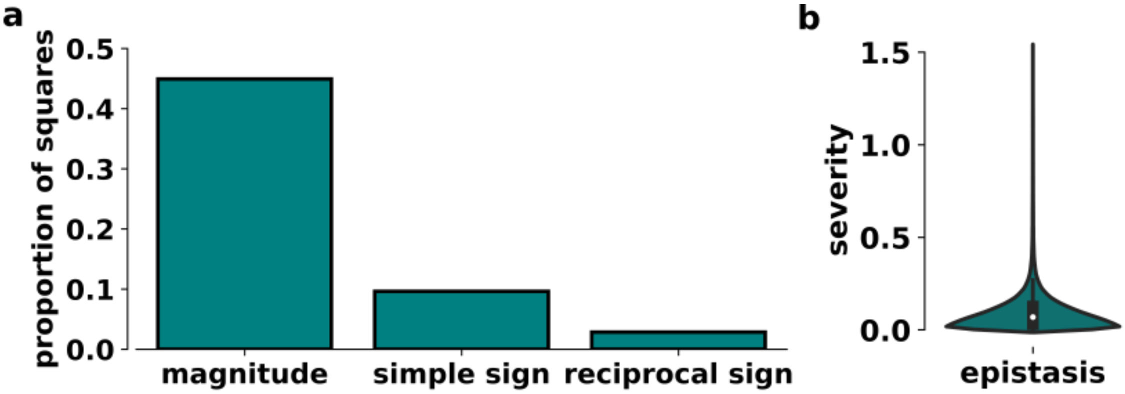


Figure S3.1: Pairwise epistasis in the phylogenetic fitness landscape

(a) Relative prevalence of three classes of pairwise epistasis (Khalid et al. 2016; Aguilar-Rodríguez et al. 2017). Pairwise epistasis was assessed using mutational pairs or squares. Two precise mutations can occur in either order, and are represented in our landscapes by subgraphs of four connected genotypes. Each square consists of a 'wild-type' genotype, two single mutants and a double mutant. Magnitude epistasis occurs when the magnitude of a mutations effect on ribozyme activity depends on the genetic background of the mutation. Simple sign epistasis occurs when a single mutant has lower activity than the wild-type and the double mutant. Reciprocal sign epistasis occurs when both double mutants have lower activity than the wild-type and double mutant. (b) The distribution of severity of epistatic interactions in the phylogenetic fitness landscape. Epistatic values for each square of mutation pairs was calculated as $\varepsilon = \log_{10} (W_{AB}*W_{wt} / W_A*W_B)$, where W_A and W_B are the fitness of RNA variants with a single mutation, W_{AB} is the fitness of the variant with both mutations, and W_{wt} is the fitness of the wild-type.

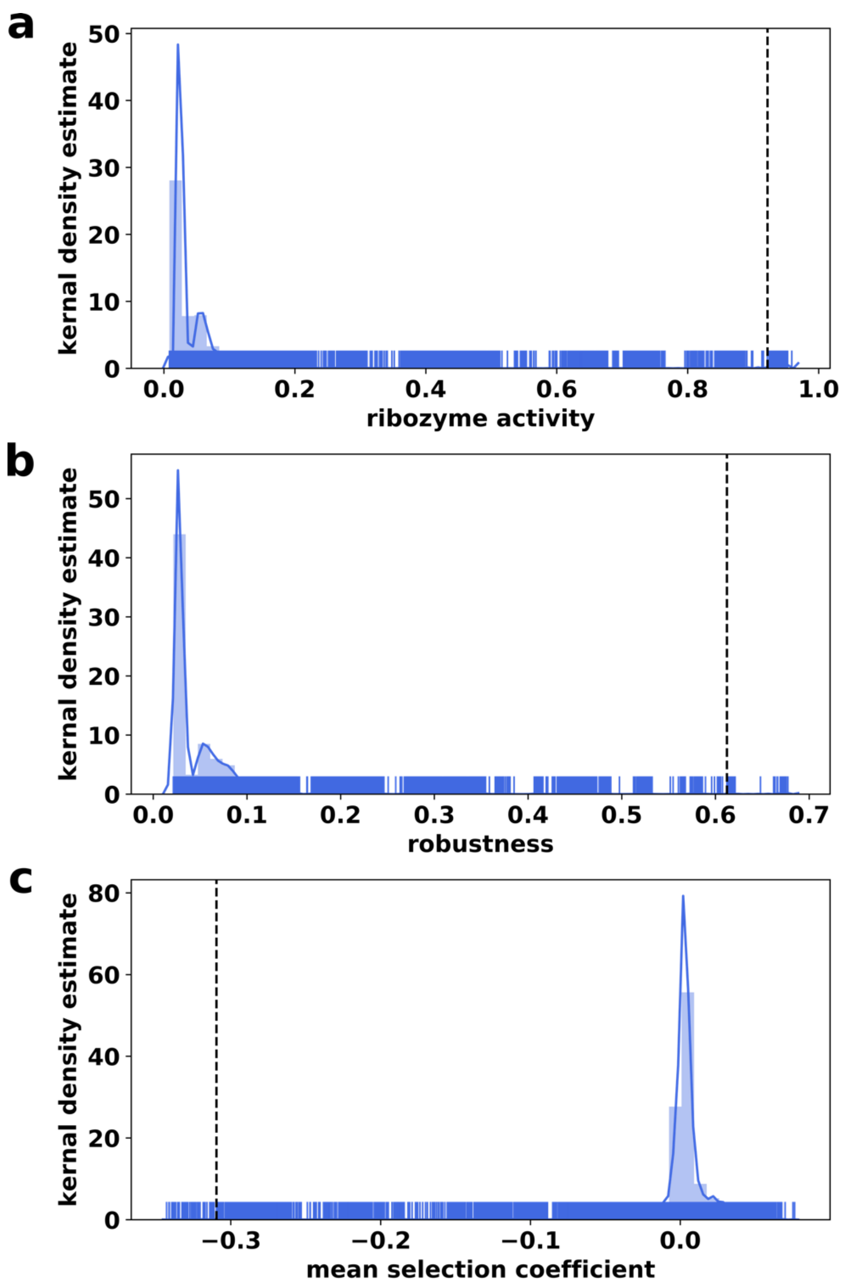


Figure S3.2: Distribution of ribozyme activity, robustness and selection coefficient in the phylogenetic fitness landscape.

(a) Distribution of ribozyme activity measured as fraction sequences cleaved during 20 minutes of transcription. The ribozyme activity of the *ancestral* sequence is indicated by the dotted line. (b) Distribution of mutational robustness. Robustness was calculated for each genotype as the average ribozyme activity of all mutational neighbors that differed by a single mutation. The robustness of the *ancestral* sequence is indicated by the dotted line. (c) Distribution of mean selection coefficients. Selection coefficients were calculated for each genotype as the difference in ribozyme activity between a genotype and its mutational neighbors that differed by a single mutation.

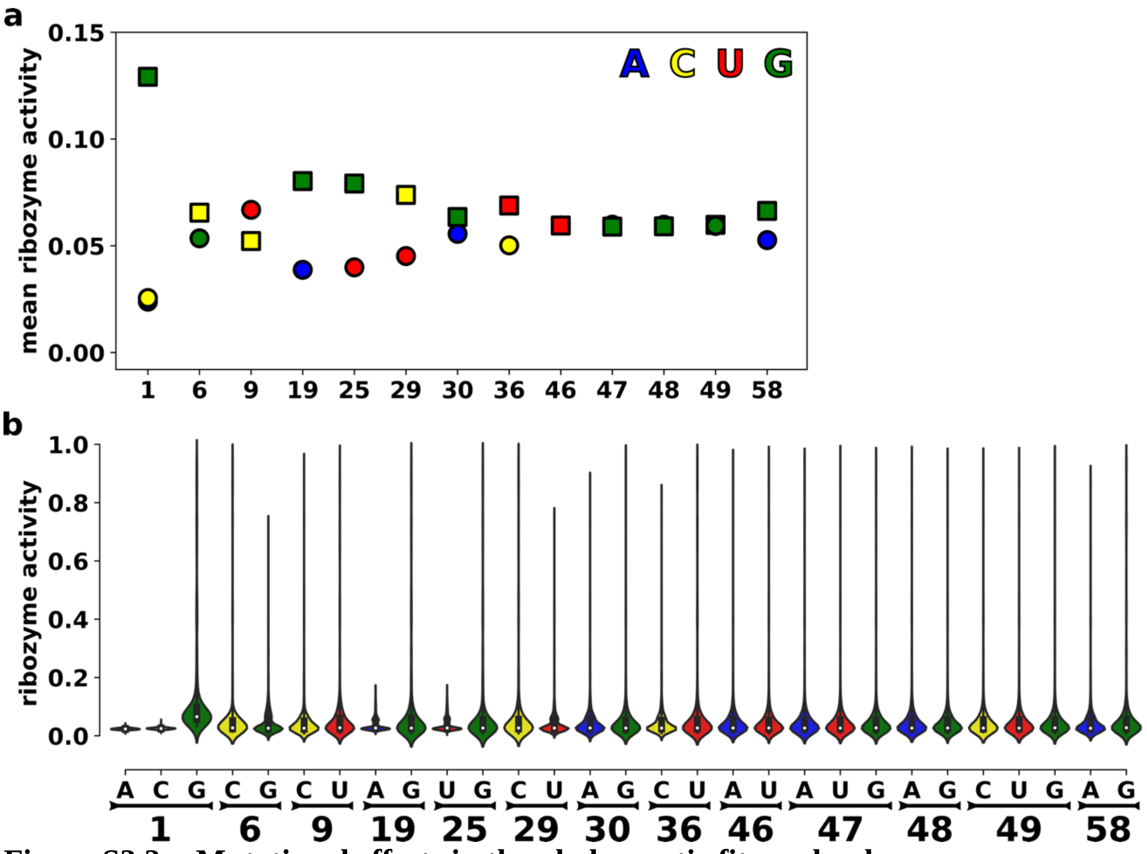


Figure S3.3: Mutational effects in the phylogenetic fitness landscape.

(a) The average of all genotypes with the corresponding mutation at the corresponding position. Position in the CPEB3 ribozyme are listed on the x-axis and mean ribozyme activity is on the y-axis. Color of node indicates the nucleotide at the position (A=blue, C=yellow, U=red, G=green). Square nodes represent nucleotides that are in the highly-conserved, highly-active *ancestral* sequence. (b) The distributions of ribozyme activities containing a given mutation at each position. Nucleotide and position in the CPEB3 ribozymes are indicated on the x-axis. Mean is indicated by white dot. Color corresponds to panel a.

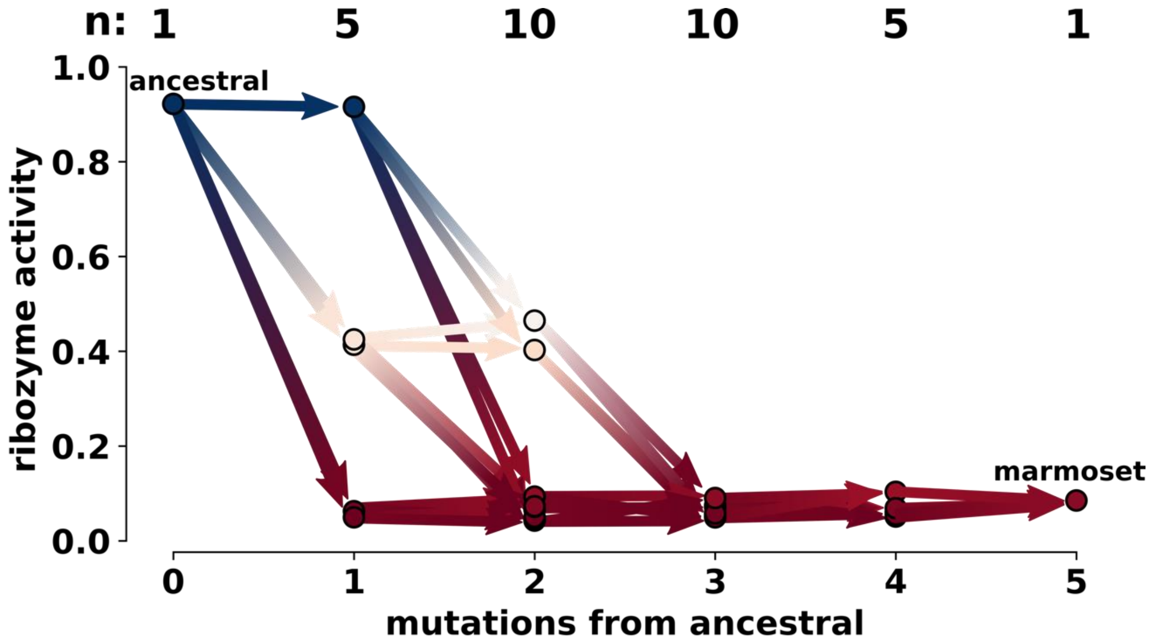


Figure S3.4: Potential evolutionary pathways from the *ancestral* sequence to the marmoset (*Callithrix jacchus*).

Hamming distance from the ancestral sequence is on the x-axis and measured ribozyme activity on y-axis. Color of nodes indicate ribozyme activity. Edges connect nodes that differ by a single mutation. Edge colors are dependent on the source and target node color. There are displayed 120 unique direct pathways that can be taken from *ancestral* sequence to the marmoset. The number of genotypes (n) at each mutational distance are indicated.

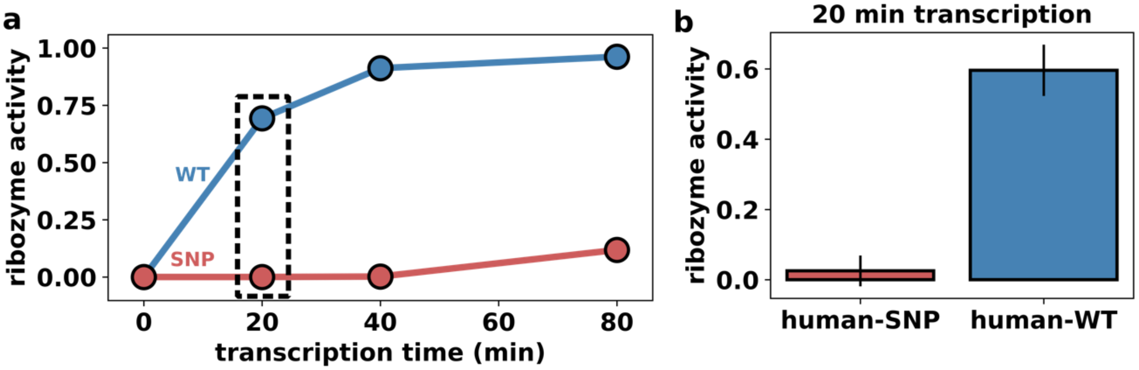


Figure S3.5: Co-transcriptional self-cleavage of human-WT and human-SNP CPEB3 ribozymes.

(a) Time-course transcription for ribozyme activity. Ribozyme activity is measured as the fraction cleaved at each timepoint (0, 20, 40 80 mins). Samples were allowed to co-transcriptionally cleave and were run on 10% denaturing polyacrylamide gel, visualized with GelRed (Biotium) and quantified by densitometry. Human-SNP is shown in red and human-WT is shown in blue. The two sequences only differ at the U36C SNP mutation.

(b) Mean of four replicates for the 20-minute co-transcriptional assay. This timepoint is what was used for high-throughput sequencing. Error bars indicate the standard deviation of the mean.

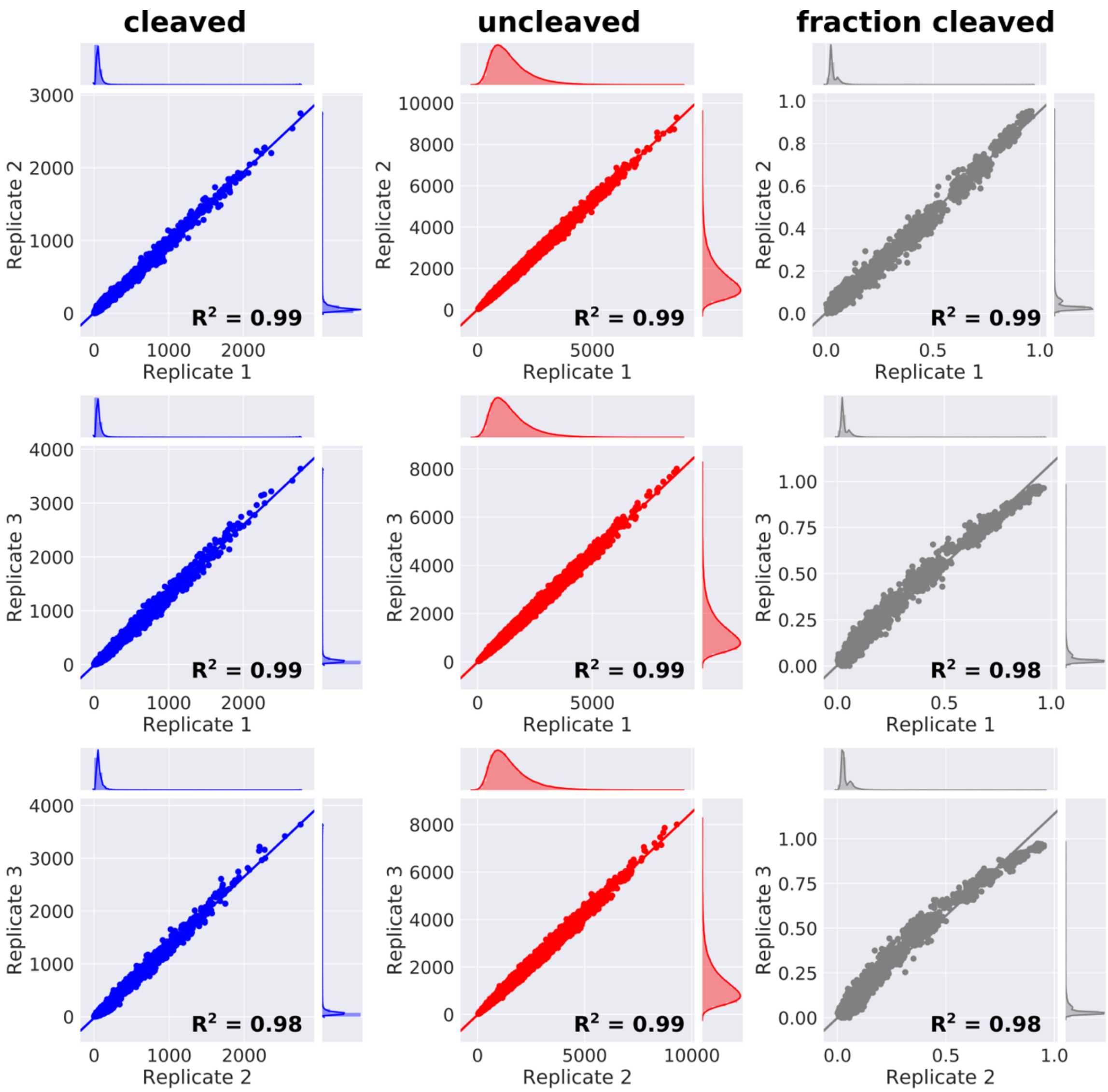


Figure S3.6: Correlation of high-throughput sequencing replicates for the phylogenetic fitness landscape library.

Correlation of number of cleaved reads, uncleaved reads and fraction cleaved (ribozyme activity) for each of the three replicates. Each figure consists of 27,648 genotypes present in the phylogenetic fitness landscape library. Each data point for cleaved and uncleaved represents the frequency that a specific genotype was observed in a particular replicate (x-axis) vs. another replicate (y-axis). The fraction cleaved is a product of observed cleaved and uncleaved reads for each genotype. Sequence kernel density is also reported from each replicate in the jointplot.

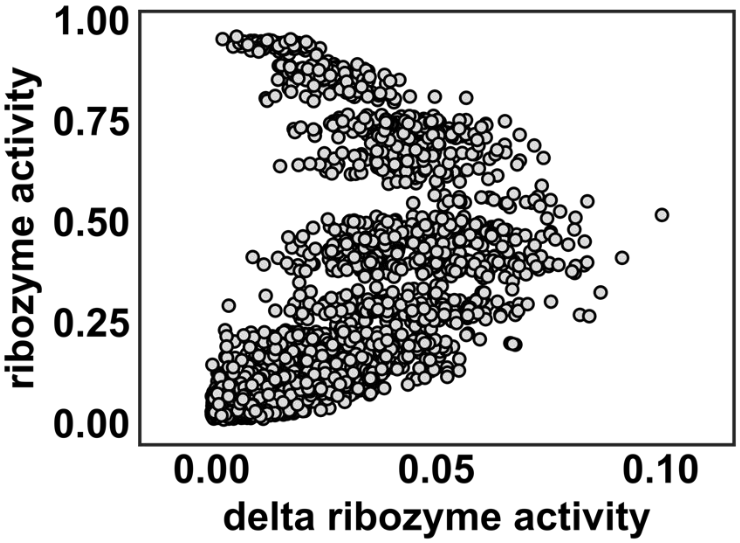


Figure S3.7: Distribution of delta ribozyme activities measured between sequencing replicates.

Delta ribozyme activity (x-axis) is measured as the standard deviation of the mean of ribozyme activities between the three biological replicates (y-axis). Delta measurements are taken into account when calculating epistasis and mutational pathways (Khalid et al. 2016)

CHAPTER FOUR: NEGATIVE EPISTASIS IN EXPERIMENTAL RNA FITNESS LANDSCAPES

Devin P. Bendixsen[1], Bjørn Østman[2,3] and Eric J. Hayden[1,4]

[1]Biomolecular Sciences Graduate Program, Boise State University
[2]Department of Ecology and Evolutionary Biology, UCLA, Los Angeles, CA 90095
[3]Department of Biomathematics, UCLA, David Geffen School of Medicine, Los Angeles, CA 90095
[4]Department of Biological Science, Boise State University
1910 University Dr., Boise, ID 83725

Abstract

Mutations and their effects on fitness are a fundamental component of evolution. The effects of some mutations change in the presence of other mutations, and this is referred to as epistasis. Epistasis can occur between mutations in different genes or within the same gene. A systematic study of epistasis requires the analysis of numerous mutations and their combinations, which has recently become feasible with advancements in DNA synthesis and sequencing. Here we review the mutational effects and epistatic interactions within RNA molecules revealed by several recent high-throughput mutational studies involving two ribozymes studied *in vitro*, as well as a tRNA and a snoRNA studied in yeast. The data allows an analysis of the distribution of fitness effects of individual mutations as well as combinations of two or more mutations. Two different approaches to measuring epistasis in the data both reveal a predominance of negative epistasis, such that higher combinations of two or more mutations are

typically lower in fitness than expected from the effect of each individual mutation. This data is in contrast to past studies of epistasis that used computationally predicted secondary structures of RNA that revealed a predominance of positive epistasis. The RNA data reviewed here is more similar to that found from mutational experiments on individual protein enzymes, suggesting a common thermodynamic framework may explain negative epistasis between mutations within macromolecules.

Introduction

It is often convenient to describe a mutation as deleterious, neutral or beneficial. In fact, there exists a continuous distribution of fitness effects caused by mutations, from lethal to highly beneficial, and every value in between. This distribution of fitness effects is important for many theories in the fields of genetics and evolution (Whitlock et al. 1995; Fenster et al. 1997; Ostman et al. 2012). For example, the rate at which a population adapts to an environment depends on the frequency of beneficial mutations relative to the more common deleterious and neutral mutations (Desai et al. 2007). In addition, whether a mutation is beneficial or deleterious depends on the presence of other mutations (genetic background). *Epistasis* is a term used broadly to describe instances when the effects of combinations of mutations are not easily predicted by the effect of each individual mutation. *Positive epistasis* is used to describe situations when combinations of mutations produce higher fitness than expected from their individual effects. *Negative epistasis* occurs when multiple mutations produce lower fitness together than expected from their individual effects (Fig. 4.1). In some situations, a mutation can even be beneficial in some genetic backgrounds, and deleterious in others. This is referred to as *sign epistasis*, and highlights the difficulty in classifying mutations as beneficial, deleterious or neutral. Sign epistasis has received considerable attention because it can make certain sequential combinations of mutations unlikely to persist in a population, even though they may be subsets of extremely beneficial combinations of mutations. The presence of sign epistasis means that some pathways to higher fitness are highly improbable (Weinreich et al. 2005; Weissman et al. 2009). Populations evolving in the laboratory, as well as in the wild, must navigate the peaks and valleys of fitness landscapes caused by selection pressure and epistatic interactions. Our ability to forecast evolutionary outcomes will require advancements in our understanding of epistasis within and between genes.

Two experimental approaches for detecting epistasis are commonly used (Fig. 4.1). One way is by determining the average fitness effects of increasing numbers of mutations relative to a reference. The genetic variants are often generated in the laboratory by mutation accumulation, utilizing extreme population bottlenecks to enhance genetic drift and force random mutation fixation (Halligan and Keightley 2009). Mutations are more frequently deleterious than beneficial, and the random accumulation thus causes a decline in the average fitness of individuals with increasing numbers of mutations. The decline can be expected to follow a simple exponential curve if epistasis is absent or if there exists a balance of positive and negative interactions (note that fitness decline will follow a linear curve for logarithmically transformed fitness values). Deviations from a simple exponential can identify epistasis when it is predominantly negative or positive (Fig. 4.1A). A second experimental approach to revealing epistasis is a pairwise comparison of mutational effects. For this, the effect of several individual mutations is determined and compared to pairs of these mutations. The epistasis between each pair of mutations can be determined by comparing their actual fitness effect relative to what is expected from each individual effect (Fig. 4.1). Both of these types of experiments have been conducted on the scale of individual genes (Bershtein et al. 2006; Hayden et al. 2015) and whole genomes (Kouyos et al. 2007; Halligan and Keightley 2009).

These experimental approaches can be thought of as ways to understand the complex mapping of genotypes to phenotypes. Ideally, a fitness value could be assigned to every possible genotype, generating a comprehensive mapping of genotype to phenotype. However, this is not possible because of the vast numbers of potential genotypes. For example, a genome of length 100 bp has $4^{100} > 10^{60}$ possible sequence variants. As the length of the genome increases, or the number of mutated sites increases, the effort required to identify, isolate or synthesize, and assay each variant increases exponentially. For this reason, individual gene products, including proteins and non-coding RNA (ncRNA) molecules, are an attractive model system (Jiménez et al. 2013). For the smaller genotype space of genes, a large subset of interesting mutations can be identified and many possible combinations can be produced and assayed. Understanding epistasis within genes has relevance to directed evolution approaches aimed at optimizing gene functions (Bloom and Arnold 2009). In addition, as the expressed components of genomes, the genotype to phenotype maps of individual genes may inform our understanding of epistasis at the level of whole genomes (Soskine and Tawfik 2010). A better understanding of epistasis may improve evolutionary theories required to assess the vast majority of genotype space that will not be studied with these experimental approaches. In the study of epistasis within a gene product, protein molecules have been the subject of the majority of empirical investigations. However, our growing understanding of the numerous and critical roles carried out by ncRNA molecules warrants investigation of how mutational effects contribute to the evolution of new RNA functions.

Several recent experiments have produced extensive data on epistatic interactions in ncRNA molecules (Fig. 4.2). Each study produced large numbers of mutational combinations and utilized a high-throughput assay for the effects of individual and combinations of mutations. Two studies used ribozymes (RNA enzymes), where fitness is defined as the ribozyme activity of a variant relative to that of a wild-type reference. In these experiments ribozyme fitness was determined *in vitro*, outside of a cellular context. Two other studies transformed yeast with libraries of specific ncRNA molecules, and used the RNA-dependent growth rate as the *in vivo* fitness metric. Here we characterize the epistatic mutational effects observed in each of these experimental systems. On average, all of the RNA molecules show a predominance of negative epistatic interactions between random mutations, whether the assay was carried out *in vivo* or *in vitro*. Underlying this average effect are similarities in the distribution of individual mutational effects and the distribution of pairwise mutational interactions. The predominance of negative epistasis is observed despite the fact that the RNA molecules reviewed have very different structures, and were studied with different experimental approaches. Similar epistasis has been seen in protein enzymes, suggesting that negative epistasis between mutations within genes may be a common property of biological parts (Bloom et al. 2004; Bershtein et al. 2006; Wylie and Shakhnovich 2011).

The RNA Molecules and Experimental Approaches

We have collected data from several recent experiments in order to facilitate direct comparison between different RNA molecules with different functions and between *in vitro* and *in vivo* experiments. We will briefly describe some of the pertinent details of each experimental system. The *in vitro* ribozyme data are from a 54-nt self-cleaving ribozyme that belongs to a structural family of ribozymes called Twister (Roth et al. 2014) and a 197-nt *Azoarcus* self-splicing group I intron ribozyme (Reinhold-Hurek and Shub 1992). For the Twister ribozyme, the mutants were generated by gene randomization during DNA synthesis, and therefore the analysis included all single and double mutants of the wild-type ribozyme, as well as a random sampling of sequences with three or more mutations (Kobori and Yokobayashi 2016). The fitness of each ribozyme was determined by the amount that each variant was able to self-cleave during *in vitro* transcription. All variants were transcribed simultaneously, and high-throughput sequencing was used to determine the fraction of each variant found in the cleaved form. Data was openly shared for the single and double mutants, enabling nearly exhaustive analysis of interactions between the effects of pairs of mutations. However, the average fitness of combinations of three or more mutations was not available. In contrast, the data for the *Azoarcus* ribozyme included average mutational effects even for high numbers of mutations, but the individual and pairwise effects were not exhaustively determined. For the *Azoarcus* ribozyme, average fitness effects were determined by producing separate populations of ribozymes with incrementally increasing numbers of mutations by consecutive rounds of error-prone PCR, and then determining the activity of each population relative to the wild-type ribozyme (Hayden et al. 2015). The number of mutations in each population was determined by high-throughput sequencing.

The ncRNA molecules studied *in vivo* include the Arginine $tRNA_{CCU}$, and the U3 small nucleolar RNA (snoRNA), both from *Saccharomyces cerevisiae*. The sequence variants, in both of these data sets were produced by randomization during DNA synthesis and contained most of the possible single and double mutants, and a random sampling of three or more mutations. The tRNA experiments used a yeast strain with this non-essential tRNA replaced by a tRNA sequence variant at its native genomic location (Li et al. 2016). The growth rate was determined for more than ~10^5 yeast strains, each with a different tRNA variant. This was accomplished by counting the population frequency of each tRNA before and after growth competition using high-throughput sequencing. The enrichment rate of each tRNA variant is assumed to arise from the growth advantage, or disadvantage, provided by the specific tRNA. The experiments were carried out at 37°C in YPD media, conditions that inhibited the growth rate of a $tRNA_{CCU}$ knockout strain to ~20% of the wild-type. The U3 snoRNA experiments used a yeast strain with the single copy of the U3 gene under a galactose inducible promoter (Puchta et al. 2016). This strain could grow in galactose media, but showed growth arrest when transferred to glucose. Growth in glucose was then recovered by transformation with a plasmid constitutively expressing the wild-type U3 sequence. Mutational variants were assayed by transforming this yeast strain with a library of U3 variants contained on a plasmid. Each plasmid also contained a unique 20nt DNA sequence or “barcode”. The barcodes linked to each U3 variant were verified in a separate sequencing reaction. This enabled only the barcodes to be sequenced during growth competitions. The population frequency of each barcode before and after growth competition was used to determine the fitness effect of each U3 variant associated with this barcode. For these *in vivo* experiments, the fitness distributions of mutations and epistasis were previously reported, but in different formats. The authors kindly shared their fitness data for this review, so that the data could be presented in a common format for comparison.

Fitness Declines to Accumulated Mutations Reveals Negative Epistasis and Robustness-Epistasis Link

Epistasis can be detected from the average effects of increasing numbers of mutations. Typically, this is done by categorizing many sequence variants based on the number of mutations per molecule n and then determining the average fitness at each value of n. The data is fit to the equation $w(n) = \exp(-\alpha n^{\beta})$ (Wilke and Adami 2001). Upon fitting this equation to the data, the parameter α indicates the average deleterious effects of mutations, and determines how rapidly fitness declines as more mutations are introduced. Lower values of α require that higher values of fitness are preserved upon mutation, a property referred to as mutational robustness (Wagner 2005). The parameter β indicates epistatic interactions in the following way. If there are no epistatic interactions, or a perfect balance of positive and negative interactions, then β=1 and the average fitness declines exponentially with increasing numbers of mutations. Values of β>1 indicates a predominance of negative epistasis (Wilke et al. 2003), and the deviation from a pure exponential is such that the fitness of genotypes with multiple mutations is lower than expected from the average fitness at n = 1 (Fig. 4.1). Values of β<1 indicates predominantly positive epistasis. In this case the decline in fitness is less rapid and multiple mutations are less deleterious than expected from the average fitness at n = 1.

In order to facilitate comparison, we have collected the data for each of the four RNA molecules, and plotted fitness as a dependent variable and number of mutations as the independent variable. We used non-linear least squares to fit the data to the above equation that models fitness decline as a function of the number of mutations (Fig. 4.3). All four of the RNA molecules analyzed in this way show a predominance of negative epistasis with β>1. For comparison, we have also plotted fitness decline curves without epistasis (β=1), but with similar deleterious effects of individual mutations (Fig. 4.3 dashed lines). Similar epistasis is observed for the *Azoarcus* ribozyme (β=1.3±0.20) and the U3 snoRNA (β=1.2±0.02), and the curves are nearly overlapping. More extreme average epistasis is seen in the Twister ribozyme (β=1.4±0.11) and the tRNA_{CCU} (β=2.7±0.02). Taken together, all molecules fit into the previously observed negative correlation between α and β (van Nimwegen et al. 1999; Wilke and Adami 2001; Bershtein et al. 2006), in that all molecules show relatively high robustness (α<0.6) and negative epistasis (β>1). It is important to point out that we are only noting general trends in the data, and have not evaluated the significance of the differences in α and β between the data sets. This is especially true for the Twister data, which comes from curve fitting to only three available data points of average fitness at n=0, n=1, and n=2.

Distributions of Individual Mutational Effects and Pairwise Interactions

Underlying the average epistatic effects described above is a range of fitness effects from individual mutations, and combinations of mutations. For example, mutational robustness (small α) could come from either many mutations with neutral effects, or a balance of beneficial and deleterious effects. Similarly, average negative epistasis could arise from a combination of positive and negative epistatic effects, if the negative epistatic effects are either more frequent or more extreme than the positive epistatic effects. Fortunately, three of the data sets (snoRNA, tRNA, and Twister) reported the fitness consequence of nearly every possible single and double mutation. This allows us to compare the distributions of fitness effects and pairwise epistatic interactions in each of these RNA molecules. The snoRNA data was reported as log(w), and we transformed this to w to facilitate comparison. In addition, we normalized all the data so that the variants with the lowest detected fitness had *fitness* = 0 and the wild-type variants all had *fitness* = 1.

We show the distribution of fitness effects caused by mutations as histograms (Fig. 4.4). The distributions of individual mutational effects are quite similar, despite the differences between the size and structures of these ncRNA molecules (Fig. 4.2), as well as the differences in experimental approaches. All three distributions are characterized by a large number of neutral mutations, indicated by the modal peak at *fitness* = 1. There is a long tail of deleterious mutational effects (0 < *Fitness* < 1). There are very few beneficial mutations (*Fitness* > 1), although more were detected for the U3 snoRNA. The U3 snoRNA shows the highest fraction of neutral mutations, which is consistent with the lowest α values obtained from curve fitting. In addition, the Twister ribozyme has a higher fraction of extremely deleterious effects (*Fitness < 0.2),* which is consistent with the larger values of α obtained by curve fitting. Taken together, the differences in the robustness of these ncRNA molecules to the effects of individual mutations is observable in the differences in the distributions of fitness effects.

We also plotted the distribution of effects caused by two mutations as lighter colored bar graphs in Figure 6.4. The distributions are clearly shifted to the left relative to the single-mutation distributions, which indicates that two mutations are typically more deleterious than one mutation. This is expected, and in itself does not indicate epistasis. However, the differences between the distributions of each RNA molecule are informative. We will describe the change that is observed for each ncRNA by comparing the distribution of single mutational effects to double mutational effects, i.e. from the dark to light histograms in Figure 6.4. The Twister data (Fig 6.4E gray) shows a dramatic increase in the number of non-functional variants, such that the modal fitness actually changes from *fitness*=1 to *fitness*=0 when comparing single to double mutations. In contrast, the most apparent change in the $\mathrm{tRNA}_{\mathrm{CCU}}$ data (Fig. 4.4A blue) is a dramatic decrease in the proportion of neutral variants (*fitness* = 1). The distribution of double-mutant fitness values for the $\mathrm{tRNA}_{\mathrm{CCU}}$ data is more broadly distributed over intermediate non-zero values, as compared to the Twister data. Finally, the U3 snoRNA data has the least pronounced change, with a slight increase in non-functional variants as well as intermediate low fitness variants, but the model peak near *fitness*=1 remains. The cause of differences in the fragility of the different RNA molecules to the effects of two mutations remains unknown, but may involve the thermodynamics of the specific structure, or differences in the environment, such as the presence of chaperone proteins *in vivo*, which will be discussed in more detail below.

To detect epistasis in this data requires a comparison between the measured effects of pairs of mutations and what would be predicted from the effects of each individual mutation (see Fig. 4.1 for explanation). The calculated pairwise epistatic values are plotted as histograms in Fig. 4.4. All distributions indicate that there exists both positive and negative epistasis. However, the mean and skew of the data support a predominance of negative epistasis. The distribution of effects in the Twister ribozyme is much broader than the other two distributions, with a heavy tail in the negative direction. The nearly balanced distribution of positive and negative effects in the pairwise interactions of U3 snoRNA emphasize the importance of higher-order epistatic interactions in this ncRNA (Weinreich et al. 2013; Sailer and Harms 2017), and is consistent with the curve fitting analysis to the average effects of higher numbers of mutations (Fig. 4.3).

Discussion

Laboratory and Computational Studies Find Predominantly Different Epistatic Interactions

The predominance of negative epistasis is notable because it differs from previous results from computational folding of RNA that uncovered a predominance of positive epistasis (Fig. 4.3 green curve). Specifically, Wilke et al. studied the computationally predicted folding of 76nt long RNA molecules (Wilke et al. 2003). They randomly generated 100 reference sequences of this length, and determined their computationally predicted secondary structure. Then they produced sequences at incrementally increasing numbers of mutations relative to each reference. They defined fitness as the fraction of sequences that folded into the same structure as the reference, and determined fitness for up to 10^6 sequences at each number of mutations. They fit this data to the epistasis equation used above, and extracted α and β parameters. In contrast to the laboratory studies reviewed here, all values of β fell into the positive epistasis range (β<1), and all molecules also showed a lower tolerance to mutations (α>0.6). One intriguing possibility for this difference is that natural selection has favored molecules with many mutational neighbors that maintain function, leading to the evolution of mutational robustness in naturally occurring RNA (Meyers et al. 2004; Kun et al. 2005). The random sequences used in the computational studies would not be expected to have mutational robustness if it is the product of evolution. Future high-throughput experiments with natural and artificially selected ribozymes could be designed to directly test this hypothesis. Another possible explanation for this discrepancy is that computationally folding RNA secondary structures are bimodal and predict either a fitness of 1 or 0 (Wilke et al. 2003). The empirical fitness measurements, on the other hand are continuous, and a large fraction of the fitness effects fall into intermediate values of fitness effects. This lack of intermediate values in computational structure prediction could lead to an overestimation of deleterious mutational effects, and an underestimate of negative epistasis.

As noted, several previous studies involving both theoretical prediction and experimental data found a negative correlation between the parameters α and β (Wilke and Adami 2001; Bershtein et al. 2006; Hayden et al. 2015). This suggests that epistasis becomes more predominantly negative as the average mutational effects are decreased. The data previously reported for the *Azoarcus* ribozyme involved three different conditions where selection pressure was intentionally decreased. As expected, decreased strength of selection resulted in increased β, and decreased α (Hayden et al. 2015). Therefore, it appears that the negative correlation between α and β holds whether the intensity of mutational effects is altered by changes in the specific molecule, such as the different RNA molecules reviewed here, or changes in the environment.

A direct comparison of *in vivo* fitness and *in silico* prediction was also previously reported for the specific Arginine tRNA_{CCU} (Li et al. 2016). The authors found that the propensity to fold properly is a poor indicator of fitness in their experiments. They found that their strain of *S. cerevisiae* is more robust to mutations in this particular tRNA than predicted from computational folding of these sequences. This could indicate that this tRNA can still function properly despite small deviations from a canonical structure, such as a single missing base pair. This would be somewhat surprising given the numerous interactions involving tRNA molecules during its lifetime in the cell (Maraia and Arimbasseri 2017). In addition, other tRNAs can also decode this codon, and it is possible that only a small amount of properly folded $\mathrm{tRNA}_{\mathrm{CCU}}$ is required to recover normal growth. The AGG codon that is decoded by $\mathrm{tRNA}_{\mathrm{CCU}}$ is the second most common Arginine codon in *S. cerevisiae*, representing ~20% of the codons for this amino acid (Cherry et al. 2012).

Another possibility is that RNA chaperones alter the folding such that some of sequence variants fold into the native structure even when it is not the most stable (MFE) structure. It is well established that an RNA chaperone called the La protein can assist the proper folding of several RNA molecules, including tRNAs, and has been shown to hide the deleterious effects of point mutations (Chakshusmathi et al. 2003). In addition, several RNA chaperones have been shown to facilitate the native folding of numerous RNA molecules (Herschlag 1995; Russell 2008), including several ribozymes (Herschlag et al. 1994; Halls et al. 2007; Sinan et al. 2011). The U3 snoRNA is a part of an RNA-protein complex (SSU processome) that processes ribosomal RNA from primary transcripts. Assembly of this complex has been shown to involve several RNA chaperones and helicases (Soltanieh et al. 2015; Hunziker et al. 2016). In addition to normal folding pathways, several chaperones have recently been shown to buffer the deleterious effects of many different mutations (Rudan et al. 2015). Similar to chaperones for protein folding, such as HSP90, RNA chaperones could enable a property referred to as phenotypic capacitance, where otherwise deleterious mutations are maintained in the population, providing the potential to produce novel adaptations upon environmental change (Rutherford and Lindquist 1998; Queitsch et al. 2002; Jarosz and Lindquist 2010). In fact, this has recently been proposed as a factor in tRNA diversification in Eukaryotes (Maraia and Arimbasseri 2017). Despite this critical role of RNA chaperones in the mapping of RNA genotypes to phenotypes, the role of RNA chaperones in promoting the evolution of novel ncRNA structures and functions remains poorly understood.

However, the prevalence in negative epistatic interactions for RNA molecules studied *in vitro* and *in vivo* suggests a common mechanism that cannot be explained by the different experimental environments. For example, besides the presence of chaperone proteins, negative epistasis *in vivo* could result from a cooperative destabilization of the multi-component complexes involving the studied RNA molecules. While this contribution is not ruled out, it cannot be the cause of negative epistasis for the ribozymes studied *in vitro* where only the RNA is present. This suggests that a property of the RNA structures themselves can account for negative epistasis. RNA structures may provide buffering against individual mutations, yet be sensitive to many mutations. We note that protein enzymes have also shown a predominance of negative epistasis that has been attributed to the crossing of a thermodynamic threshold (Bershtein et al. 2006). The combined results suggest a similar phenomenon is occurring in these RNA structures. A thermodynamic framework may provide a prediction of epistasis within gene products and multi-component RNA-protein complexes.

It is interesting to note that several previous studies have found a predominance of positive epistasis in the genomes of RNA viruses when they are randomly mutated away from the wild-type (Bonhoeffer et al. 2004; Sanjuán 2010). The distribution of individual mutational effects in RNA viruses has also been studied, and uncovered a very high fraction of mutations with a lethal effect (fitness = 0). For example, about 40% of mutations were found to be lethal in both the tobacco etch virus and vesicular stomatitis virus (Sanjuán et al. 2004; Iglesia and Elena 2007). These findings are consistent with the negative correlation between epistasis and robustness (larger α and $\beta<1$). More generally, the distribution of fitness effects in viral genomes (Sanjuán 2010) appears to be quite similar to findings in some cellular genomes (Eyre-Walker and Keightley 2007), in that both show large fractions of lethal mutations, and predominately positive epistatic interactions or no predominant direction of epistasis (Elena and Lenski 1997; He et al. 2010). Individual proteins, on the other hand, have fitness distributions very similar to those seen here in RNA molecules, with very few lethal mutations and negative epistasis (Soskine and Tawfik 2010). Given the apparent robustness of individual macromolecules, and the multiple forms of robustness in living organisms (Wagner 2011), the mutational targets that are the source of lethal mutations in the genomes of viruses and cellular organisms is not completely understood. It is possible that these few well-studied examples of proteins and RNA are not typical in their distribution of fitness effects, but this seems unlikely given the variety of forms and functions as well as the multiple ways they were assayed. Another possibility is that pleiotropic mutations that affect multiple traits or processes are more likely to be lethal. More research is needed to understand the mechanisms and implications of differences in the mutational effects in organisms and their genetically encoded biological parts.

The presence of negative epistasis is important for models that attempt to explain the evolution and maintenance of recombination and sexual reproduction (Kouyos et al. 2007). The RNA molecules reviewed here meet several of the requirements for this theory, namely a predominance of negative epistasis, correlation between epistasis and strength of selection, and probably tight linkage between mutations within the small RNA molecules. At first these small RNA genes seem unlikely to be a large enough target for recombination to explain the evolution of recombination at the organismal level. However, when one considers the pervasiveness of transcription, and the discovery of long non-coding RNA, the breaking up of linked mutations with negative epistasis within RNA molecules cannot be ruled out as a driving force for the evolution of recombination. Interestingly, it has been proposed that genetic recombination may have contributed to the origin of life (Lehman 2003). In fact, RNA recombinase ribozymes have been demonstrated in the laboratory as models of the RNA World versions of modern protein recombinase enzymes (Hayden et al. 2005; Vaidya et al. 2012; Pesce et al. 2016). The predominance of negative epistasis in the RNA molecules reviewed here suggest that the earliest RNA genomes could have benefited from recombination, and supports the theory that genetic recombination may be as old as life itself.

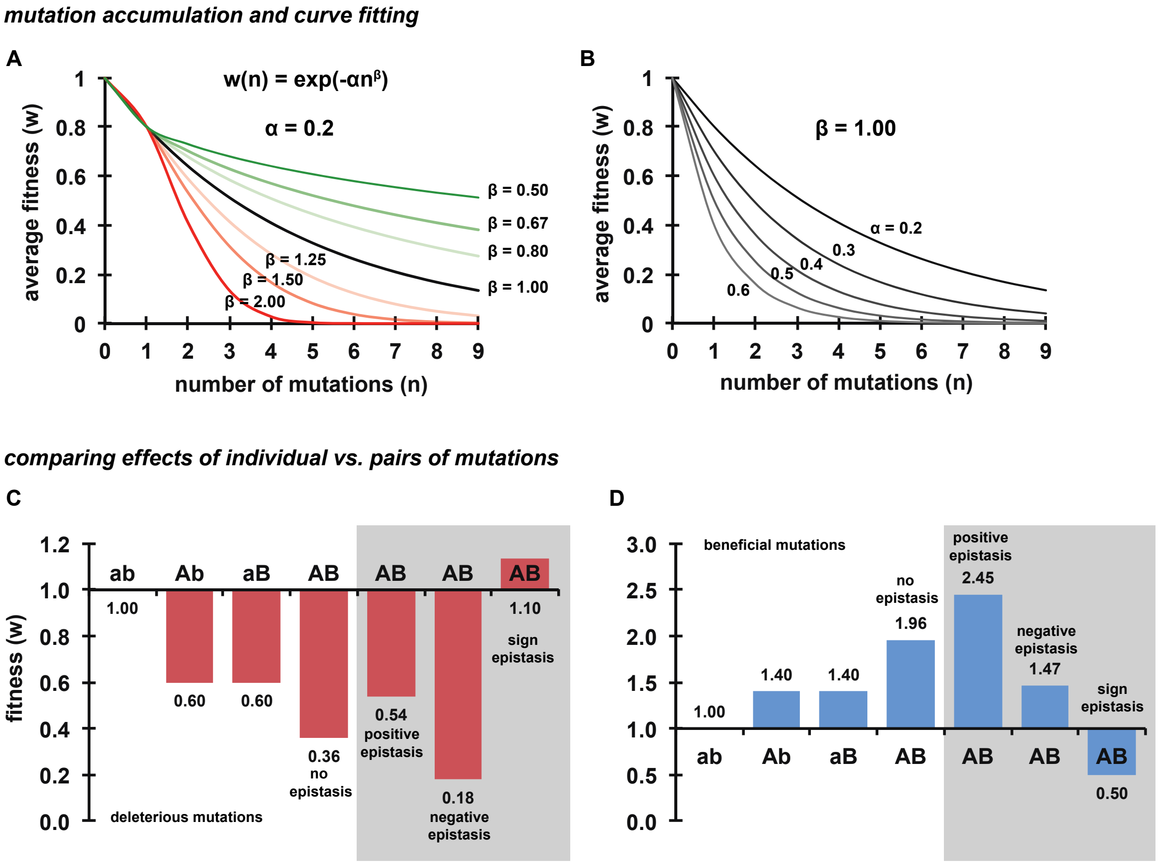


Figure 4.1 Experimental approaches to uncovering epistatic mutational interactions. ***Mutation accumulation and curve fitting.***

(A) Predominantly negative (red) or positive (green) epistasis can be detected by mutation accumulation followed by non-linear curve fitting. The average fitness (w) is determined for populations of genotypes with a given number of mutations (n). No epistasis is inferred by β=1 (black curve). Decreasing β results in positive epistasis, while increasing β results in negative epistasis. All curves have the same average mutation effect (α = 0.2). (B) Examples of curves with no epistasis (β = 1) and different values for α. *Comparing effects of individual vs. pairs of mutations.* Mutation effects are compared to a wild-type reference (ab). Two mutations are indicated as a to A and b to B. Assuming that the effects of each mutation is multiplicative, epistasis (grey box) is identified as a deviation from this prediction. Positive epistasis is observed as higher than expected fitness. Negative epistasis is observed as lower than expected fitness. (C) Epistasis between deleterious mutations (red). (D) Epistasis between beneficial mutations (blue).

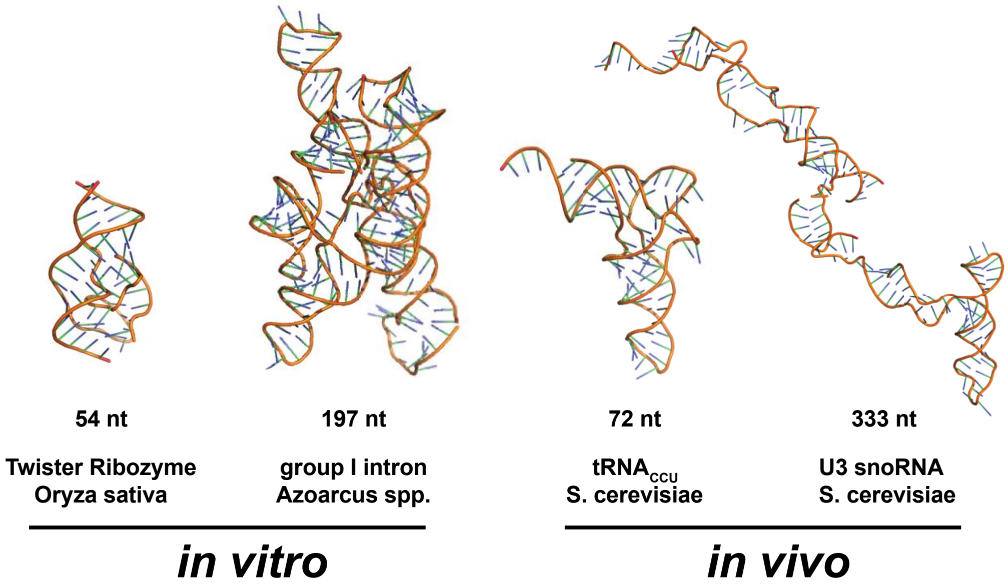


Figure 4.2 Structures of the RNA Molecules.

The size of the RNA, name, and genomic source is given below each structure. The assay conditions used to measure mutational effect (*in vitro* vs. *in vivo*) are also indicated. Structures were rendered in Pymol. Crystal structure coordinates are from the *Oryza sativa* Twister ribozyme (4OJI), the *Azoarcus* group I intron (1ZZN), an Asparagine tRNA from yeast (1VTQ). The U3 snoRNA structure is taken from the context of a cryo-EM structure of the 90S pre-ribosome (5JPQ). Structures are not to scale.

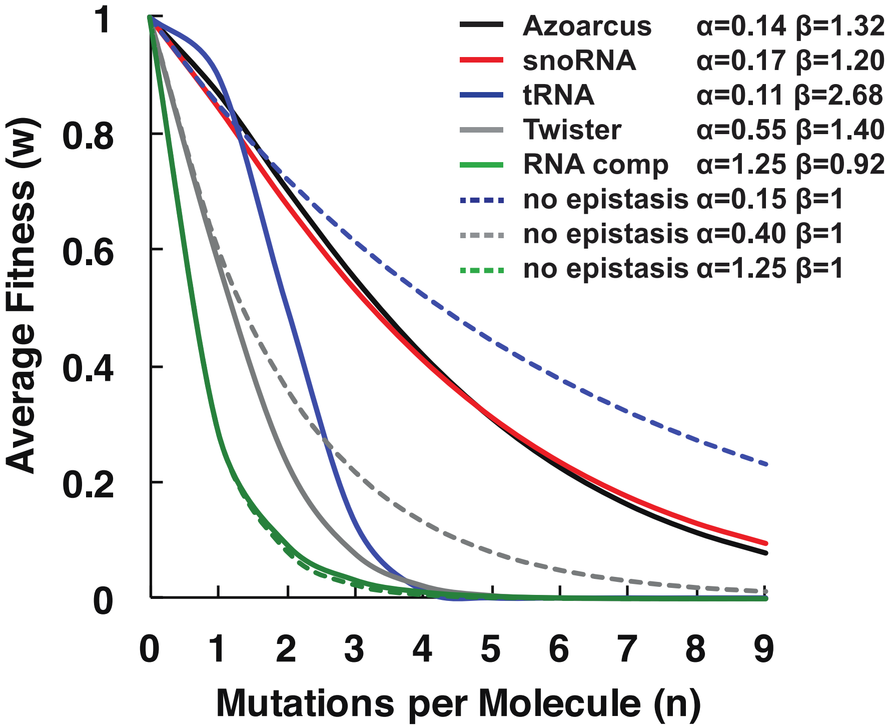


Figure 4.3 Decline in the average fitness of ncRNA variants caused by increasing numbers of mutations.

The average fitness *w* is plotted as a function of the number of mutations per molecule *n* determined by the number of nucleotide differences relative to a *wild-type* reference sequence. Solid lines represent epistatic equations of the form $w(n)=\exp(-\alpha n^{\beta})$, with parameters that produce the best-fit to the experimental data by non-linear least squares curve fitting (Python). Individual data points are excluded for visual clarity, however are included in the Supplemental Material (Fig. S4.2). For comparison, dashed lines show curves with no epistasis β=1, and an activity at *n*=1 similar to the twister ribozyme (gray dashed line) or the other three RNA molecules (blue dashed line). Data from the computational folding of RNA sequences (*RNA comp*) with positive epistasis are also shown for comparison (green).

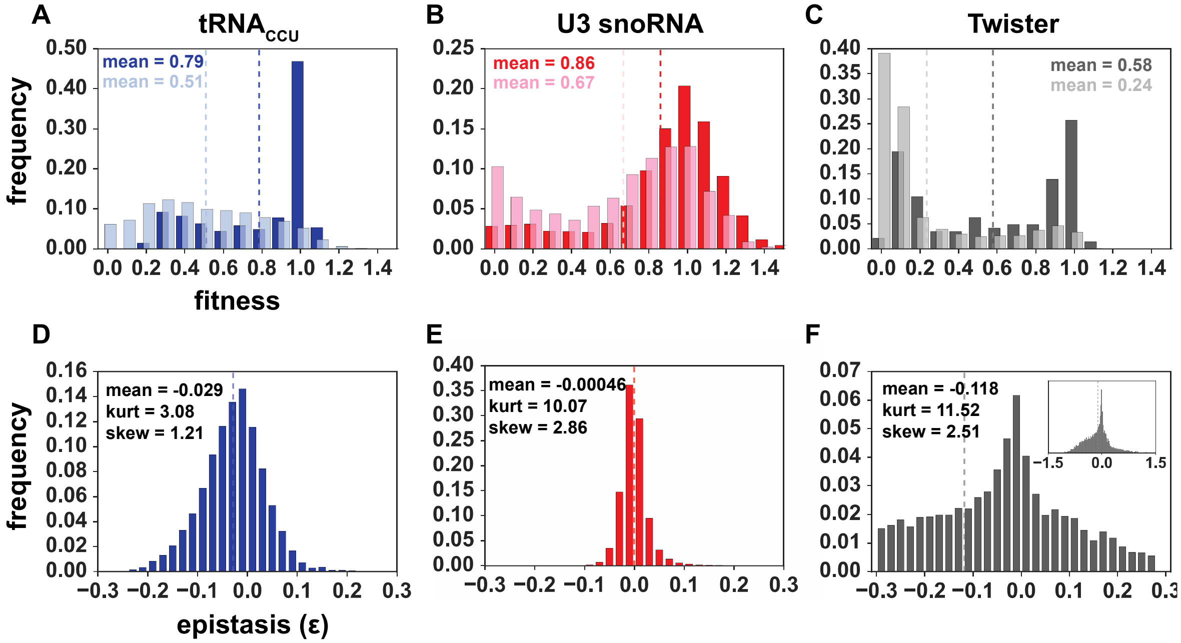


Figure 4.4 Distribution of fitness effects and pairwise epistasis.

(A) Distributions of fitness effects for individual mutations (dark blue) and pairs of mutations (light blue) for tRNA_{CCU} (Li et al. 2016). (B) Distributions of individual (dark red) and pairwise (light red) mutational effects in the U3 snoRNA (Puchta et al. 2016) (C) Distributions of individual (dark gray) and pairwise (light gray) mutational effects in the Twister ribozyme (Kobori and Yokobayashi 2016). (D) Distribution of epistatic values for pairs of mutations for the tRNA_{CCU}. (E) Distribution of epistatic values for pairs of mutations for the U3 snoRNA. (F) Distribution of epistatic values for pairs of mutations for the Twister ribozyme. Epistatic values were calculated as $\varepsilon = \log_{10} (W_{\text{AB}} * W_{\text{wt}} / W_{\text{A}} * W_{\text{B}})$, where W_{A} and W_{B} are the fitness of RNA variants with a single mutation, W_{AB} is the fitness of the variant with both mutations, and W_{wt} is the fitness of the wild-type. All distributions are set to the same scale on the x-axis. Inset in (F) shows the full distribution of epistatic effects in the Twister ribozyme.

References

Bershtein S, Segal M, Bekerman R, et al (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature 444:929–932. doi: 10.1038/nature05385

Bloom JD, Arnold FH (2009) Colloquium Papers: In the light of directed evolution: Pathways of adaptive protein evolution. Proc Natl Acad Sci 106:9995–10000. doi: 10.1073/pnas.0901522106

Bloom JD, Wilke CO, Arnold FH, Adami C (2004) Stability and the evolvability of function in a model protein. Biophys J 86:2758–2764. doi: 10.1016/S0006-3495(04)74329-5

Bonhoeffer S, Chappey C, Parkin NT, et al (2004) Evidence for positive epistasis in HIV-1. Science 306:1547–1550. doi: 10.1126/science.1101786

Chakshusmathi G, Kim SD, Rubinson DA, Wolin SL (2003) A La protein requirement for efficient pre-tRNA folding. EMBO J 22:6562–6572. doi: 10.1093/emboj/cdg625

Cherry JM, Hong EL, Amundsen C, et al (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40:D700-705. doi: 10.1093/nar/gkr1029

Desai MM, Fisher DS, Murray AW (2007) The Speed of Evolution and Maintenance of Variation in Asexual Populations. Curr Biol CB 17:385–394. doi: 10.1016/j.cub.2007.01.072

Elena SF, Lenski RE (1997) Test of synergistic interactions among deleterious mutations in bacteria. Nature 390:395–398. doi: 10.1038/37108

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8:610–618. doi: 10.1038/nrg2146

Fenster CB, Galloway LF, Chao L (1997) Epistasis and its consequences for the evolution of natural populations. Trends Ecol Evol 12:282–286.

Halligan DL, Keightley PD (2009) Spontaneous mutation accumulation studies in evolutionary genetics. Annu Rev Ecol Evol Syst 40:151–172.

Halls C, Mohr S, Del Campo M, et al (2007) Involvement of DEAD-box proteins in group I and group II intron splicing. Biochemical characterization of Mss116p, ATP hydrolysis-dependent and-independent mechanisms, and general RNA chaperone activity. J Mol Biol 365:835–855.

Hayden EJ, Bendixsen DP, Wagner A (2015) Intramolecular phenotypic capacitance in a modular RNA molecule. Proc Natl Acad Sci 201420902. doi: 10.1073/pnas.1420902112

Hayden EJ, Riley CA, Burton AS, Lehman N (2005) RNA-directed construction of structurally complex and active ligase ribozymes through recombination. RNA N Y N 11:1678–1687. doi: 10.1261/rna.2125305

He X, Qian W, Wang Z, et al (2010) Prevalent positive epistasis in E. coli and S. cerevisiae metabolic networks. Nat Genet 42:272–276. doi: 10.1038/ng.524

Herschlag D (1995) RNA chaperones and the RNA folding problem. J Biol Chem 270:20871–20874.

Herschlag D, Khosla M, Tsuchihashi Z, Karpel RL (1994) An RNA chaperone activity of non-specific RNA binding proteins in hammerhead ribozyme catalysis. EMBO J 13:2913.

Hunziker M, Barandun J, Petfalski E, et al (2016) UtpA and UtpB chaperone nascent pre-ribosomal RNA and U3 snoRNA to initiate eukaryotic ribosome assembly. Nat Commun 7:12090. doi: 10.1038/ncomms12090

Iglesia F de la, Elena SF (2007) Fitness Declines in Tobacco Etch Virus upon Serial Bottleneck Transfers. J Virol 81:4941–4947. doi: 10.1128/JVI.02528-06

Jarosz DF, Lindquist S (2010) Hsp90 and environmental stress transform the adaptive value of natural genetic variation. Science 330:1820–1824. doi: 10.1126/science.1195487

Jiménez JI, Xulvi-Brunet R, Campbell GW, et al (2013) Comprehensive experimental fitness landscape and evolutionary network for small RNA. Proc Natl Acad Sci 110:14984–14989. doi: 10.1073/pnas.1307604110

Kobori S, Yokobayashi Y (2016) High-Throughput Mutational Analysis of a Twister Ribozyme. Angew Chem Int Ed 55:10354–10357. doi: 10.1002/anie.201605470

Kouyos RD, Silander OK, Bonhoeffer S (2007) Epistasis between deleterious mutations and the evolution of recombination. Trends Ecol Evol 22:308–315.

Kun A, Santos M, Szathmary E (2005) Real ribozymes suggest a relaxed error threshold. Nat Genet 37:1008–1011. doi: 10.1038/ng1621

Lehman N (2003) A case for the extreme antiquity of recombination. J Mol Evol 56:770–777. doi: 10.1007/s00239-003-2454-1

Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. Science 352:837–840. doi: 10.1126/science.aae0568

Maraia RJ, Arimbasseri AG (2017) Factors That Shape Eukaryotic tRNAomes: Processing, Modification and Anticodon–Codon Use. Biomolecules 7:26. doi: 10.3390/biom7010026

Meyers LA, Lee JF, Cowperthwaite M, Ellington AD (2004) The Robustness of Naturally and Artificially Selected Nucleic Acid Secondary Structures. J Mol Evol 58:681–691. doi: 10.1007/s00239-004-2590-2

Ostman B, Hintze A, Adami C (2012) Impact of epistasis and pleiotropy on evolutionary adaptation. Proc Biol Sci 279:247–256. doi: 10.1098/rspb.2011.0870

Pesce D, Lehman N, de Visser JAGM (2016) Sex in a test tube: testing the benefits of in vitro recombination. Philos Trans R Soc Lond B Biol Sci. doi: 10.1098/rstb.2015.0529

Puchta O, Cseke B, Czaja H, et al (2016) Network of epistatic interactions within a yeast snoRNA. Science 352:840–844. doi: 10.1126/science.aaf0965

Queitsch C, Sangster TA, Lindquist S (2002) Hsp90 as a capacitor of phenotypic variation. Nature 417:618–624. doi: 10.1038/nature749

Reinhold-Hurek B, Shub DA (1992) Self-splicing introns in tRNA genes of widely divergent bacteria. Nature 357:173–176. doi: 10.1038/357173a0

Roth A, Weinberg Z, Chen AGY, et al (2014) A widespread self-cleaving ribozyme class is revealed by bioinformatics. Nat Chem Biol 10:56–60. doi: 10.1038/nchembio.1386

Rudan M, Schneider D, Warnecke T, Krisko A (2015) RNA chaperones buffer deleterious mutations in E. coli. eLife. doi: 10.7554/eLife.04745

Russell R (2008) RNA misfolding and the action of chaperones. Front Biosci J Virtual Libr 13:1–20.

Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396:336–342. doi: 10.1038/24550

Sailer ZR, Harms MJ (2017) Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. Genetics 205:1079–1088. doi: 10.1534/genetics.116.195214

Sanjuán R (2010) Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos Trans R Soc Lond B Biol Sci 365:1975–1982. doi: 10.1098/rstb.2010.0063

Sanjuán R, Moya A, Elena SF (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A 101:8396–8401. doi: 10.1073/pnas.0400146101

Sinan S, Yuan X, Russell R (2011) The Azoarcus group I intron ribozyme misfolds and is accelerated for refolding by ATP-dependent RNA chaperone proteins. J Biol Chem 286:37304–37312.

Soltanieh S, Osheim YN, Spasov K, et al (2015) DEAD-Box RNA Helicase Dbp4 Is Required for Small-Subunit Processome Formation and Function. Mol Cell Biol 35:816. doi: 10.1128/MCB.01348-14

Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nat Rev Genet 11:572–582. doi: 10.1038/nrg2808

Vaidya N, Manapat ML, Chen IA, et al (2012) Spontaneous network formation among cooperative RNA replicators. Nature 491:72–77. doi: 10.1038/nature11549

van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A 96:9716–9720.

Wagner A (2005) Robustness and Evolvability in Living Systems. Princton University Press

Wagner A (2011) The origins of evolutionary innovations: a theory of transformative change in living systems. OUP Oxford

Weinreich DM, Lan Y, Wylie CS, Heckendorn RB (2013) Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev 23:700–707. doi: 10.1016/j.gde.2013.10.007

Weinreich DM, Watson RA, Chao L (2005) Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evol Int J Org Evol 59:1165–1174.

Weissman DB, Desai MM, Fisher DS, Feldman MW (2009) The rate at which asexual populations cross fitness valleys. Theor Popul Biol 75:286–300. doi: 10.1016/j.tpb.2009.02.006

Whitlock MC, Phillips PC, Moore FB-G, Tonsor SJ (1995) Multiple Fitness Peaks and Epistasis. Annu Rev Ecol Syst 26:601–629. doi: 10.1146/annurev.es.26.110195.003125

Wilke CO, Adami C (2001) Interaction between directional epistasis and average mutational effects. Proc Biol Sci 268:1469–1474. doi: 10.1098/rspb.2001.1690

Wilke CO, Lenski RE, Adami C (2003) Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. BMC Evol Biol 3:3.

Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci 108:9916–9921. doi: 10.1073/pnas.1017572108

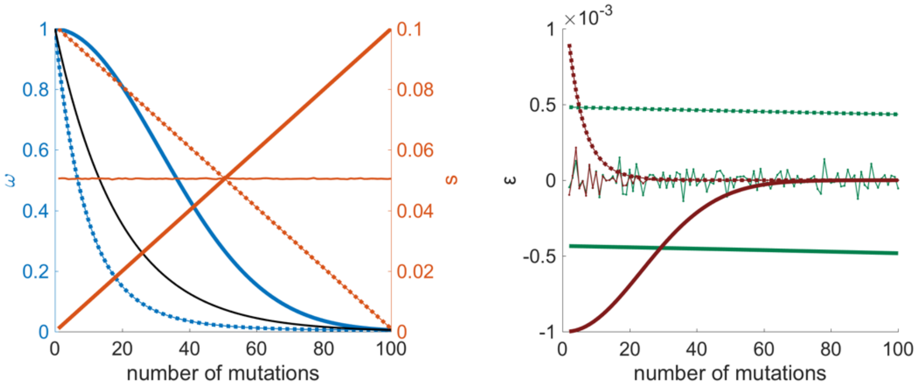
Supplemental material

Negative epistasis naturally arises when selection works on a linearly decreasing distribution of fitness effects. A population subject to just one of these (deleterious) mutations will consist of those with the smallest s, because selection will favor those mutants over others. With two mutations, the two smallest s mutations will go to fixation, and so on. The order of the mutations as they accumulate under selection will therefore be the smallest first (Fig. S4.1, solid lines), which results in negative epistasis.

Fitness-functions above the black line results in negative epistasis, and because this curve is close to the fitness-function of the decreasing s (dashed blue line), we can expect that most distributions will lie above this (black) line, and thus result in negative epistasis.

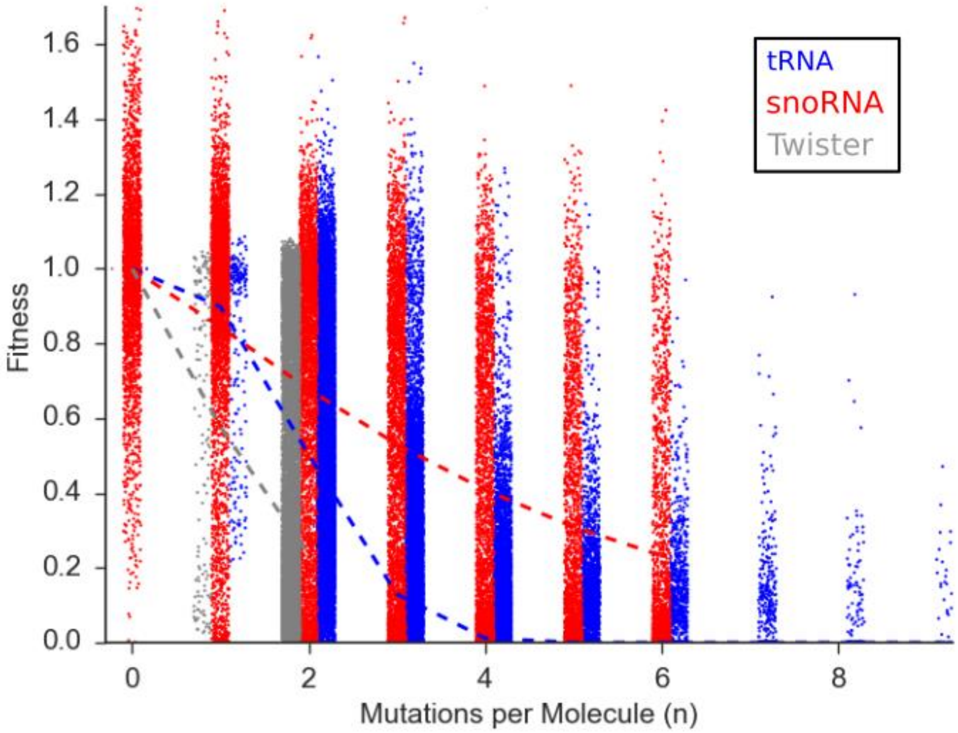
Epistasis is calculated as $\varepsilon = \log_{10}(W_0W_{AB}/W_AW_B)$ (green lines) and $\varepsilon = W_0W_{AB} - W_AW_B$ (brown lines), which both have the same sign, even if they show different magnitudes of epistasis. Here we assume that the effect of the other single-mutant is the same as the first, i.e., $W_A = W_B$.

The order of mutations (i.e., 0 → A → AB vs. AB → A → 0) makes no difference for the sign of epistasis, as interchanging W_0 and W_{AB} gives the same result. This means that for adaptive evolution with those same mutations in reverse, the (now beneficial) mutations with the largest *s* would most likely go to fixation first, and epistasis along this trajectory would therefore also be negative, resulting in diminishing returns epistasis.



Supplemental Figure 4.1. A linear distribution of fitness effects results in negative epistasis.

A linearly increasing selection coefficient distribution (orange solid line) results in a fitness-function that declines slowly as the number of mutations increase (blue solid line). A decreasing distribution of fitness effects (dashed orange line) results in a fitness function that decreases sharply with the first large-effect mutations (blue dashed line). The increasing s results in negative epistasis (green and brown solid lines), while the decreasing s results in positive epistasis (green and brown dashed lines). Random sampling of mutations over 100,000 replicates gives an intermediate fitness-function (black line), which results in a constant s = 0.5 (thin orange line) and zero epistasis (thin lines).



Supplemental Figure 4.2. Distribution of genotype fitness of ncRNA variants caused by increasing numbers of mutations.

The fitness of each genotype is plotted as a function of the number of mutations per molecule *n* determined by the number of nucleotide differences relative to a *wild-type* reference sequence. Each data point represents a single genotype from tRNA (blue), snoRNA (red) and Twister (gray) datasets. This is the normalized fitness data used to infer alpha and beta parameters in Figure 6.3. Dashed lines indicate the mean fitness for each mutational distance.

DISSERTATION CONCLUSION

Overall, the work presented here represents a significant contribution to our ability to construct and empirically characterize RNA fitness landscapes. The development of two high-throughput ribozyme assays opens the door for further empirical landscape construction. The phased nucleotide insert technique will allow for larger landscapes to be constructed using bigger mutational libraries without increasing the sequencing cost. The implementation of data-driven stochastic evolutionary modeling allows for a clearer evolutionary characterization of the landscape than straight-forward pathway analyses. Understanding the connection between genotype and phenotype in RNA systems is important for designing RNA functions, improving *in vitro* selections and understanding the origins and evolution of new RNA functions or even to predict or forecast future evolutionary directions. Applying the advances within this work yielded valuable information about evolutionary innovations, the effects of higher dimensionality, evolution of extant ribozymes and the prevalence of epistasis in RNA fitness landscapes. Overall this work represents a significant contribution to the field of evolutionary biology, both in terms of new tools in the *fitness landscape toolbox* and new insights into evolutionary processes.

APPENDIX A: PHASED NUCLEOTIDE INSERTS FOR SEQUENCING LOW-DIVERSITY SAMPLES FROM *IN VITRO SELECTION EXPERIMENTS*

Devin P. Bendixsen[a], Brent Townshend[b] & Eric J. Hayden[a,c]

[a]Biomolecular Sciences Graduate Programs, Boise State University, 1910 University Dr., Boise, ID, USA.

[b]Department of Bioengineering, Stanford University, 443 Via Ortega, Stanford, CA, USA

[c]Department of Biological Sciences, Boise State University, 1910 University Dr., Boise, ID, USA.

Abstract

In vitro selection combined with high-throughput sequencing is a powerful experimental approach with broad application in the engineering and characterization of RNA molecules. The diverse pools of starting sequences used for selection are often flanked by fixed sequences used as primer binding sites. The low nucleotide diversity at the beginning of the sequence causes reduced sequence output and poor sequence quality on Illumina platforms due to complications with fluorescence imaging algorithms. A common solution to this problem is the addition of fragmented bacteriophage PhiX genome. This increases diversity to allow for sequencing, but sacrifices a portion of the usable sequencing reads. An alternative approach to improve nucleotide balance is to insert nucleotides of variable length and composition at the beginning of each molecule when adding adaptors prior to sequencing. This approach preserves read depth by sacrificing several bases in the length of each read. Here, we test the ability of inserted nucleotides to replace PhiX in a low-diversity sample generated from a high-throughput sequencing based ribozyme activity screen. We designed an RNA library based on the twister ribozyme from *Oryza sativa* and screened the resulting 4,096 sequence variants for self-cleaving ribozyme activity. Phased nucleotides were inserted during reverse-transcription using four different template-switching oligos. The resulting cDNA was sequenced with and without the addition of PhiX DNA on a MiSeq platform. We found that libraries with the phased inserts produced equally high-quality sequence data whether or not PhiX DNA was added. In this experiment, the increase in read depth achieved without PhiX improved the consistency of activity measurements as compared to previously reported data. We conclude that phased inserts can be implemented following *in vitro* selection experiments to eliminate the use of PhiX when read length is not critical.

Introduction

The development of RNA molecules with desired functions has numerous applications in RNA research. While *de novo* rational design of sequences that provide a desired function remains an ongoing pursuit (Daher et al. 2017; Weenink et al. 2017), *in vitro* selection offers a proven experimental approach (Robertson and Joyce 1990; Ellington and Szostak 1990). *In vitro* selection starts with diverse pools of sequences and uses cycles of functional selection and amplification to enrich only the sequences with the desired function, while discarding unwanted sequences. Nucleotide sequence analysis is often used to monitoring the results of *in vitro* selections. Historically, sequencing has focused on the end-point of selection in order to find a small number of desired sequences that are in high abundance. More recently, advances in high-throughput sequencing have enabled a more quantitative analysis of selections over time and to provide a larger set of functional sequences. High-throughput sequence analysis has been applied to several common goals of *in vitro* selection, such as finding ligand specific aptamers (Dupont et al. 2015; Levay et al. 2015), catalytic RNA molecules (Ameta et al. 2014; Pitt and Ferré-D’Amaré 2010; Hayden 2016), and chimeric aptazymes, which are allosteric ribozymes that are engineered by combining aptamer and ribozyme sequences in a single molecule (Martini et al. 2015). The steps in the process are inspired by evolution, and the cycles of replication, mutation and selection have also been used to study the process of evolution as it unfolds in real-time in the laboratory (Joyce Gerald F. 2007; Hayden et al. 2011).

In vitro selections often begin with very diverse pools of sequences (libraries) that can include over 10^{15} different nucleotide sequences. These diverse pools enable the search for rare functions or beneficial combinations of mutations. However, these complex libraries often have no sequence diversity at their 5' and 3' ends because these sequence elements are used as primer binding sites for amplification by PCR or reverse-transcription PCR during the regeneration phase of a selection. Importantly, these primer binding sites are the first nucleotides to be sequenced by Illumina platforms, which causes a major challenge when the instrument's automated algorithms are trying to identify the precise location of individual sequence clusters that are subsequently monitor during rounds of sequencing by synthesis (Krueger et al. 2011). Ironically, because of the lack of nucleotide diversity at the beginning of each read, these complex libraries are considered "low-diversity" for the Illumina platforms. Low-diversity samples have been shown to result in poor sequence output and quality. One way to improve the sequencing of samples with low nucleotide diversity is to add randomly fragmented DNA from bacteriophage PhiX (Illumina Technical Report, https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-divers.html, April 7 2017). This PhiX addition changes the balance of juxtaposed fluorescent signals during early sequencing cycles which improves output and quality. However, this PhiX addition also consumes 5-50% of the sequencing reads effectively diverting a portion of the sequencing cost toward an unwanted target. For perspective, a typical HiSeq run with 15% PhiX results in the sequencing of the entire PhiX genome approximately 8,000 times. In addition to being wasteful, the abundant use of PhiX has also caused contaminated genome assemblies (Mukherjee et al. 2015). A brief survey of the literature shows recent *in vitro* RNA selection experiments use 8-30% PhiX addition resulting in high-quality data and significant data loss (Kobori and Yokobayashi 2016; Pressman et al. 2017; Kobori et al. 2017, 2015).

The loss of data at this level has consequences for the precision and accuracy of functional measurements from sequence data. Several experimental designs have used the change in abundance of each unique sequence over selection rounds to quantify function, such as binding affinity or ribozyme activity. The accuracy and precision of this approach depends upon sequencing depth, defined as the number of reads assigned to each unique sequence (Sims et al. 2014). In this approach, sequences with lower read depth show poor precision between replicates. Therefore, losing a substantial portion of sequencing reads can limit the ability to accurately quantify functions or properties of individual sequences, or fail to identify rare sequences. Alternatively, in order to improve read depth, precision and accuracy, it’s possible to increase the amount of sequences generated by moving to a higher throughput sequencing platform or simply doing multiple sequencing runs. However, this can significantly increase the cost of sequencing in an effort to obtain adequate usable data. Another approach is the use of custom read primers that overlap constant regions next to variable regions and improve diversity while reducing read lengths. This approach requires that there exist known constant regions in the library, which is not always the case, and requires that new custom primers be designed for each library. This approach also ignores any possible unexpected mutations or errors that might occur that are not in the variable region.

An alternative approach to improving low-diversity sequencing is to insert nucleotides of variable length and composition when preparing samples for sequencing. This has been achieved using a set of PCR primers that each add a different number of bases upstream of PCR amplicons that will be sequenced. These primers are often referred to as “phased primers” because they shift the cycle number in which the amplicon is sequenced such that neighboring clusters are often out of phase, and no longer produce identical or highly similar fluorescent signals in each sequencing cycle. This approach improves sequencing read depth but reduces the available sample read length. Therefore, this approach is best utilized when target molecules are shorter than the read length by more than the length of the phased nucleotide insert. This approach has been applied to the sequencing of ribosomal genes from microbial communities, and was shown to improve sequence throughput, as well as average base quality scores and effective read length. This prior work on microbial amplicon sequencing suggests that a similar phased primer design could also improve the sequencing of cDNA resulting from the *in vitro* selection of RNA.

Here, we test the use of phased nucleotide insertions for high-throughput screening of a library of RNA molecules for their ability to catalyze a self-cleaving ribozyme reaction. We designed an RNA library based on the twister ribozyme from the *Oryza sativa* genome (Roth et al. 2014) that randomized six nucleotide positions in two distal regions that interact in a tertiary structural element (Fig. A.1A). This ribozyme cleaves near the 5’-end of the RNA. Assessment of the relative activity of each sequence variant requires sequencing both uncleaved and cleaved molecules in order to quantify the fraction of each sequence that is in the cleaved vs. uncleaved form (fraction cleaved). For this purpose, we used a reverse transcription protocol that relies on the template switching property of the reverse transcriptase to produce single stranded cDNA with attached partial adapter sequences from all RNA molecules, regardless of their 5’-sequence identity (5’-RACE protocol). Our phased nucleotide inserts were introduced into the template-switching oligo such that they become the first nucleotides sequenced by the Illumina platform. The inserted sequences were of four different lengths in order to change the phasing of the subsequent ribozyme sequences. They were also designed to have a balanced nucleotide composition with approximately equal likelihood of A, C, G and T at each of the first nine inserted nucleotides to improve cluster identification at the critical initial cycles of sequencing (Fig. A.1B). To test our phased inserts, we carried out a co-transcriptional cleavage reaction with our ribozyme library and prepared this RNA for Illumina sequencing using our phased template switching oligos during reverse transcription. We then sequenced this sample on a MiSeq platform in two conditions. The first condition used the addition of 25% PhiX and the second condition used essentially no PhiX. 0.5% PhiX was added to the second condition in order to determine the sequencing error rates and did not significantly alter nucleotide diversity. Comparing these two conditions allows for a direct assessment of the efficacy of the phased insert nucleotides for the sequencing of low-diversity samples.

Results and Discussion

Simulated Sequencing Predicts That Phased Inserts Provide Better Nucleotide Balance Over Phix

To predict the capacity of our phased inserts to improve the nucleotide diversity and balance of the low-diversity ribozyme samples, we produced simulated sequence data to mimic the sequence diversity that would be produced under different library preparation protocols. For comparison, we determined the expected uncertainty of each position of sequencing reads (Fig. A.2). Our model used randomly generated collections of one million sequences taken from the twister ribozyme library sequences, then appended our phased inserts, or added random sequences to approximate the PhiX genome, or both. Each read was then appended with Illumina sequencing primers and the first 150 nucleotides that would be sequenced were analyzed. For each position (index) starting from the 5’ end, we used information theory to calculate the entropy (Adami Christoph 2012). Entropy is dependent on the nucleotide balance or relative proportion of each nucleotide at each position (Fig. AS.1). Entropy was calculated in units of bits, where a single completely random nucleotide has an entropy of 2 bits. Inversely, a nonrandom nucleotide has an entropy of 0. As a metric of nucleotide diversity along the entire read, the average positional entropy was calculated. For comparison, we also generated a low-diversity control that used only the 4,096 unique sequences of the twister library with no phased inserts and no PhiX addition. These sequences are identical for all but the six positions of the T1 pseudoknot that were randomized (Fig. A.2, grey). This control sample produces an average positional entropy of 0.08±0.39 ($\mu\pm\sigma$). The addition of 25% PhiX alone into the sequencing pool produces a modest improvement in the entropy, resulting in an average positional entropy of 1.03±0.19 (Fig. A.2, yellow). The addition of phased nucleotide insertions alone reduces the average nucleotide balance beyond what is achieved by PhiX, giving an average positional entropy of 1.43±0.36 (Fig. A.2, blue). Finally, having both phased inserts and 25% PhiX only modestly improves the nucleotide balance compared to phased nucleotide insertions alone, with an average positional entropy of 1.73±0.17 (Fig. A.2, red). We conclude from this simulated sequence run that the phased inserts are expected to improve nucleotide balance over PhiX alone and suggests that the addition of PhiX might no longer be needed.

Phased Inserts Produce High Quality Data Without Phix

We next tested our model prediction by comparing two sequencing runs of the twister library. Both runs used the same DNA sample that was generated by the same template switching reverse transcription reaction with phased nucleotide insertions. One sequencing run used a 25% PhiX addition, and the other did not. The two sequencing runs produced similar raw output quantity and quality, with >90% clusters passing filter, and >90% of base calls greater than Q30 (Table A.1). The run with 25% PhiX spike-in yielded slightly more clusters (1.10 million) compared to without PhiX (1.03 million), however this is within a normal expected range for the platform (MiSeq, Nano mode). The quality scores for each position in the sequencing read are similar for the two runs (Fig. A.3A). Each boxplot in Fig. A.3A is noticeably higher than a quality score of 28, which indicates a base call accuracy of 99.9%. This is typically considered very high-quality data. By averaging the quality scores for each nucleotide position in a sequencing read we can calculate a mean sequence quality score. The distributions of the mean and minimum sequence quality scores for sequencing with and without PhiX are also very similar and further supports the conclusion that the two runs yielded very similar, high-quality data (Fig. A.3B, 3C). To validate that the quality of the data was improved by the addition of phased nucleotide inserts, we compared it to a previous failed sequencing run without phased inserts and with 10% PhiX addition (Fig. AS.2). This library was similarly prepared but was based on a different self-cleaving ribozyme (HDV) and was sequenced using an Illumina HiSeq 3000 platform. This sequencing run resulted in significant amounts of failed cluster identification with only ~12% of clusters passing filter. The clusters that did pass filter were low quality, with a mean sequence quality of <28 (Fig. AS.2B). Although this sequencing run was on a different platform, both the MiSeq and the HiSeq platforms are known to suffer from similar low-diversity issues. These findings support the prediction that phased nucleotide insertions alone, without PhiX, remedy the low diversity of our ribozyme samples.

Eliminating PhiX Improves Sequence Read Depth

The extra data from eliminating PhiX results in an improvement in the sequencing depths for both cleaved and uncleaved reads of each genotype (Fig. A.4A). This is important because these counts are used to quantify the relative activity of each nucleotide sequence. The sequencing run without PhiX produced a total of 899k twister library reads. The sequencing run with PhiX produced only 689k ribozyme reads, a difference of ~210k reads (Table A.1). To a first approximation, phased inserts alone increases read depth by the amount of PhiX that was added, i.e. 25% in our case, albeit with shorter usable read length. The net effect, in terms of usable nucleotides sequenced, depends heavily on the read-length kits (70, 150, 300 cycles) and the length of the target sequence. If the length of the target sequence and the phased insert is less than the read-length kit then the net increase in reads is directly connected to the amount of PhiX used (25% in our case). This is because data output and quality were not reduced when PhiX was eliminated. We note that the reduction in coverage has a larger impact on the cleaved reads because they are in lower abundance than the uncleaved reads. For example, when comparing the read counts from the two runs, the cleaved reads show a lower correlation (R^2=0.48) than the uncleaved reads (R^2=0.97), illustrating the risk caused by PhiX addition for low abundance reads.

Genotype Fitness Validation with Previously Published Data

The relative ribozyme activity of a subset of our ribozyme library was previously reported in a high-throughput sequencing based mutation analysis (Kobori and Yokobayashi 2016). Specifically, the activity for 18 single and 81 double mutants within the T1 pseudoknot were previously reported. To facilitate a visual comparison, we present the previous data and our two data sets as heat maps (Fig. A.4C). A visual comparison shows a good general agreement between the two data sets. It is important to note that compared to our assay, the previously published data was collected with a longer co-transcription time (2 h), higher pH (8.0) and lower magnesium concentration (6 mM). The layout of the heat maps are such that the activities of sequence with a single mutation are on the far left column and bottom row. Along the diagonal are the activities of ribozymes with compensatory double mutations that convert one of the C-G base pairs (recall that the pseudoknot in the wild-type ribozyme is comprised of three consecutive C-G base pairs) to a different Watson-Crick base pair (G-C, A-U, or U-A). The previous data and our current data are very similar in that these compensatory mutations tend to maintain high relative ribozyme activity. The majority of the differences between the previous data and our current data lie off the diagonal. These positions in the heat map represent pairs of mutations that do not result in a Watson-Crick base pair at one position in the pseudoknot. We find that in general our new data reports higher activity for both G-U and A-C wobble base pairs. These base pairs may be stabilized slightly by the specific conditions of our experiment. Such data comparisons between experiments with different conditions may be used to understand genotype by environment interactions for RNA molecules (Adami 2004).

The decrease in reads associated with PhiX addition has real consequences in our data. Importantly, of the 4,096 unique sequences in our library, 4011 were observed in the cleaved state in data that was obtained with PhiX. In contrast, only 3933 sequences were found to be cleaved when PhiX was added. Every sequence was found in the uncleaved state with and without PhiX. This illustrates how data saved by our phased inserts results in better detection of low abundance reads. In addition, we look more closely at the deviations between our two data sets and the previously reported data (Fig. A.4D). We find that the differences in relative activity are more extreme between the previously published data and our data that used PhiX, and our data that was obtained with phased inserts only produced less extreme differences (Kruskal-Wallis H=6.42, p=0.01). We conclude that the data saved by the use of phased inserts increases the detection of low abundance reads as well as the consistency of activity measurements between experiments.

We find that phased nucleotide inserts are an effective method to improve sequencing yield of our ribozyme reactions by eliminating the need for PhiX addition. While our study system analyzed self-cleaving ribozymes, our results should hold for other *in vitro* selection experiments, including selections for RNA and DNA aptamers and aptazymes. Phased nucleotides could be introduced in several ways, depending upon how sequencing adaptors are added. For example, phased nucleotides could be introduced into oligos used during RNA ligation protocols, or into PCR primers. However, the approach does sacrifice read length for improving read depth. Replacing PhiX addition with phased inserts only makes sense in situations where sequenced molecules are shorter than the read length by more than the length of the inserted nucleotides. We conclude that the use of similar phased insert designs would improve numerous *in vitro* selection experiments.

Materials and methods

Simulated Sequencing Run Entropy Calculation

To determine the expected effect of phased nucleotide inserts as compared to the addition of a random genome such as PhiX, we used information theory to calculate the expected entropy at each position based on the alignment of 1 million generated reads. This was repeated for four different sequencing libraries: 1) no phased insert + no PhiX, 2) no phased insert + 25% PhiX, 3) phased insert + no PhiX and 4) phased insert + 25% PhiX. At each position in the aligned sequencing reads entropy (*H*) was defined as: $H(x) = -\sum_{i=1}^{N} p_i log_2 p_i$ (Adami Christoph 2012), where N=4 representing the four canonical DNA nucleotides and p_i indicates the relative proportion of that nucleotide at that position.

Phased Nucleotide Insert Design

Four template switching oligonucleotides (TSO) were designed with phased nucleotide inserts that added 9, 12, 15 or 18 nucleotides (Table A.2). The phased nucleotides were inserted between the partial Illumina adapter and three ribose guanines in a TSO needed for second strand synthesis by template switching. The four phased TSO were combined in equal concentrations and diluted to a concentration of 10 µM total oligonucleotides in Tris-EDTA pH 8.

Twister Library Design

The twister library was synthesized as a “machine-mixed” ssDNA oligonucleotide (IDT). The library was synthesized as the reverse complement to act as the template strand during *in vitro* transcription with T7 RNA polymerase. The minus strand of the T7 promoter was appended to the 3’-end, and a fixed sequence (linker) was added to the 5’-end of the DNA library to serve as a primer binding site for reverse transcription (Wilkinson et al. 2006) (Table A.2). The ribozyme sequence included 54 nucleotides taken from *Oryza sativa* (Osa-1-4) ribozyme except with randomized bases (N) at six nucleotide positions that correspond to the T1 pseudoknot (Fig. A.1A). This results in a library of 4^6 = 4,096 unique RNA sequences.

Co-Transcriptional Self-Cleavage Assay

The promoter region of the ssDNA library was made double stranded by annealing to the T7-TOP+ primer (Table A.2). Reactions containing 20 pmol of each oligonucleotide and 10X T7 buffer (300 mM Tris pH 7.5, 500 mM DTT, 200 mM Spermidine, 100 mM $MgCl_2$). Oligos were heated to 98°C for 5 mins then cooled to room temperature and diluted 10-fold. Transcription reactions used 8 µL of annealed library in a 200 µL reaction with 1X T7 buffer, 4 µL rNTP (25 mM, NEB), 8 µL T7 RNA polymerase (200 units, Thermo Scientific) and 160 µL RNase free water (Ambion) and were incubated at 37°C for 20 mins. The transcription and ribozyme self-cleavage was terminated by the addition of 15 µL of 50 mM EDTA. Protein and buffer were removed using Direct-zol RNA MicroPrep w/ TRI-Reagent (Zymo Research). The sample was eluted in 7 µL, quantified by UV absorbance, normalized to 5uM, and checked for quality by denaturing PAGE (10 % polyacrylamide, 8 M urea).

Reverse Transcription with Phased Template Switching

The purified RNA (5 pmol) was mixed with 20 pmol of RT-library primer (Table A.2) in a final volume of 10 µL and heated at 72°C for 3 mins and cooled on ice. A 10 µL mixture, consisting of 4 µL SMARTScribe 5x First-Strand Buffer (Clontech), 2 µL dNTP (10 mM), 2 µL DTT (20 mM), 1 µL water and 1 µL SMARTScribe Reverse Transcriptase (100 units, Clontech), were then added to the RNA template and RT primer. SMARTScribe Reverse Transcriptase was chosen for its template-switching activity which allows for the addition of a constant primer binding site onto the 3' end of the cDNA. The phased TSO mix (20 pmol) was added resulting in a 22 µL reverse transcription reaction. The mixture was then incubated at 42°C for 90 mins, followed by heating the mixture to 72°C to stop reverse transcription and degrade the RNA. The single stranded cDNA product was purified using a silica-based column kit (Zymo Research) eluted with 7 µL water.

Illumina Adapter PCR and High-Throughput Sequencing

In preparation for high-throughput sequencing, Illumina adapter sequences were added to each end of the cDNA library using low-cycle PCR. The PCR reaction consisted of a 1 µL cDNA library, 12.5 µL KAPA HiFi HotStart ReadyMix (2X, KAPA Biosystems), 2.5 µL forward, 2.5 µL reverse primer (Illumina Nextera Index Kit) and 5 µL water. Each PCR cycle consisted of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 30 s.

Multiple PCR cycles were analyzed using gel electrophoresis. A cycle with observable dsDNA but prior to saturation was chosen for sequencing. The PCR product was purified with a silica-based column kit (Zymo Research) and verified using gel electrophoresis. The sample was sent to the University of Oregon Genomics and Cell Characterization Core Facility for quality control, quantification by qPCR and sequencing on the Illumina MiSeq (Nano mode, single end 150bp). The same sample was run on the MiSeq sequencer twice, once with a 25% PhiX addition and once with a minimal 0.5% PhiX. The 0.5% PhiX does not significantly alter diversity but was added at the request of the core facility in order to ensure similar sequencing error rates between sequencing runs, which is determined by comparing the PhiX reads to the PhiX reference genome.

Data Analysis

Sequencing data were analyzed using Biopython (Cock et al. 2010, 2009) and custom Python scripts. Relative activity values (w) for each unique sequence, or genotype, were determined from the fraction of sequencing reads found in the cleaved form ($N_{cleaved}$) divided by the total reads of that genotype: $w=N_{cleaved}/(N_{cleaved}+N_{uncleaved})$. The fitness values were then normalized such that the wildtype sequence in each sequencing run was equal to 1 ($w_{norm}=w/w_{WT}$).

Tables

Table A.1. Sequencing metrics for phased nucleotide inserts.

Raw data incorporates all sequencing reads from each sequencing run. Twister data incorporates only sequencing reads that correspond to the twister ribozyme library for each run, thus excluding reads that mapped to the PhiX genome.

Data Type	Sequencing Metric	Phased Insert + 25% PhiX	Phased Insert
Raw Data	**Raw Clusters**	1,103,506	1,027,991
	Pass Filter Clusters	1,035,812	935,119
	% Pass Filter Clusters	93.87%	90.97%
	Yield (Mbases)	156	141
	% >= Q30 bases	94.12%	92.17%
	Mean Quality Score	36.54	35.96
Twister Data	**Pass Filter Clusters**	689,334	899,174
	Yield (Mbases)	104	136
	% >= Q30 bases	92.91%	92.76%
	Mean Quality Score	36.17	36.1
	% of lane	**66.55%**	**96.16%**

Table A.2. Oligonucleotides used in this study.

Name	Sequence (5' - 3')	Notes
Twister-library	GAACCGGACCGAAGCCCGATTTGGATCCGGCGAACCGGATCGA**CCGCCNNNTCCACTTTTATCCGGGCTTNNNACCGGCATTGGCAGTGTT**AGGCGGCCCTTTTCCTATAGTGAGTCGTATTAGCCG	HDV template oligonucleotide. Ribozyme sequence is bolded. Cleaved sequence is in red.
T7-TOP+ primer	CGGCTAATACGACTCACTATAG	T7 transcription primer.
RT-library primer	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAACCGGACCGAAGCCCG	Reverse transcription primer
Phased TSO 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**GCATGCATGCATGCATGC**rGrGrG	Phased template switching oligonucleotides. Phased insert is bolded. rG indicates RNA bases
Phased TSO 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**TGCATGCATGCATGC**rGrGrG	
Phased TSO 3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**ATGCATGCATGC**rGrGrG	
Phased TSO 4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG**CATGCATGC**rGrGrG	

Figures

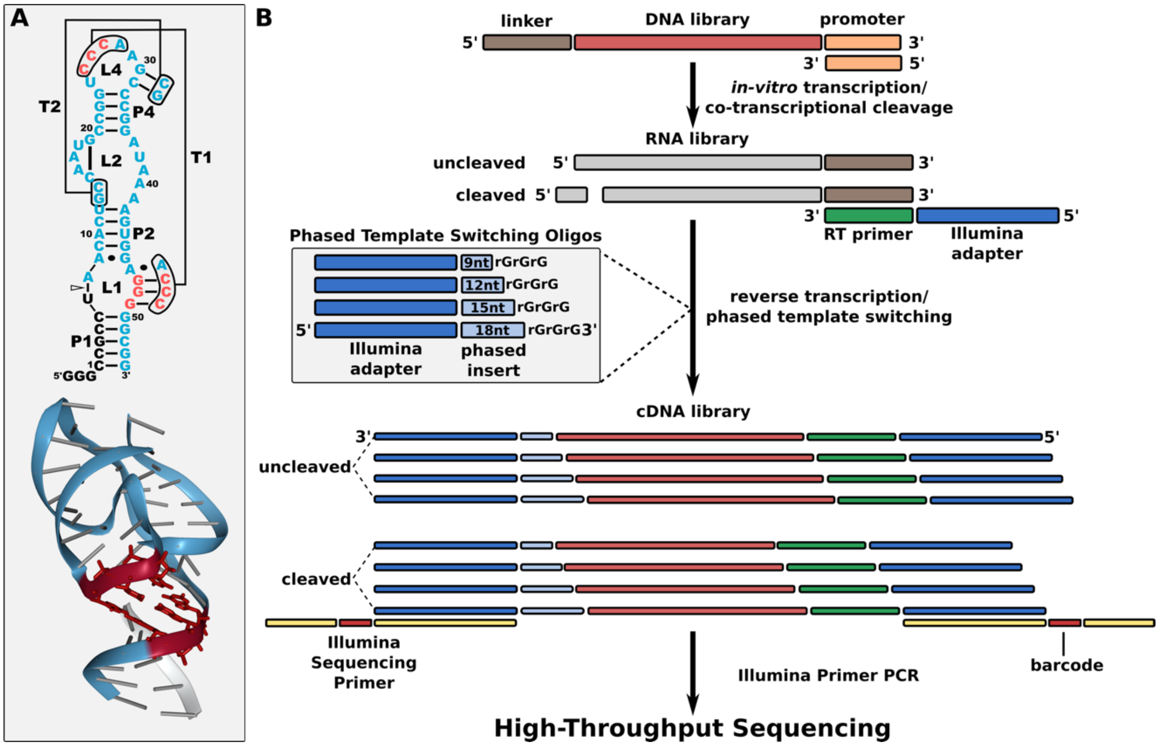


Figure A.1. The library design and *in vitro* protocol.

(A) Secondary and tertiary structure of the twister Osa-1-4 ribozyme (Liu et al. 2014; Rose and Hildebrand 2015). The library contained six randomized nucleotide positions indicated by the red nucleotides. The triangle in the secondary structure indicates the cleavage site, and the black nucleotides are the cleaved product. *(B)* Illustration of the protocol for co-transcriptional self-cleavage and phased nucleotide insertion during template switching reverse transcription.The DNA library is ordered as the template strand for transcription, with the T7 promoter at the 3'-end, and a primer binding sequence at the 5'-end (linker). Active variants self-cleave during transcrption by T7 RNA polymerase. Cleaved and uncleaved RNA products are reverse-transcribed with template switching using the linker sequence for primer binding, and a pool of four phased template switching oligonucleotides. These phased inserts are incorporated into the the cDNA transcripts during reverse transcription. The resulting single stranded cDNA products with phased inserts are amplified with index primers to add full adaptors for high-throughput sequencing (Illumina).

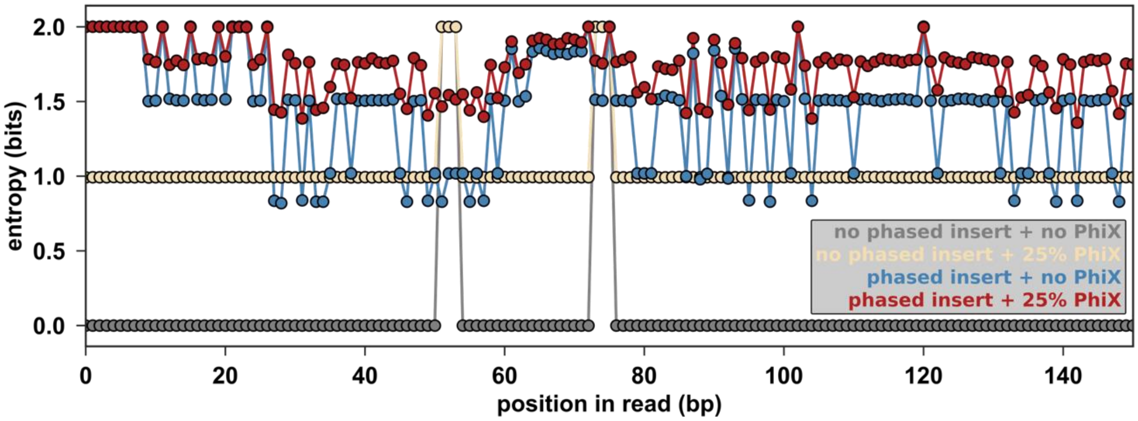


Figure A.2. Prediction of positional entropy from simulated sequence run.

Entropy was predicted for four simulated twister ribozyme library samples. Maximum entropy at a position is indicated by *entropy* = 2. Identical nucleotide for all sequences at a given position is indicated by entropy = 0. Predicted nucleotide balance is shown for a control library without phased nucleotide insertions or PhiX (grey), *only* 25% PhiX (tan), addition of only phased nucleotide insertions (blue), or *(D)* both phased insertions and PhiX (red).

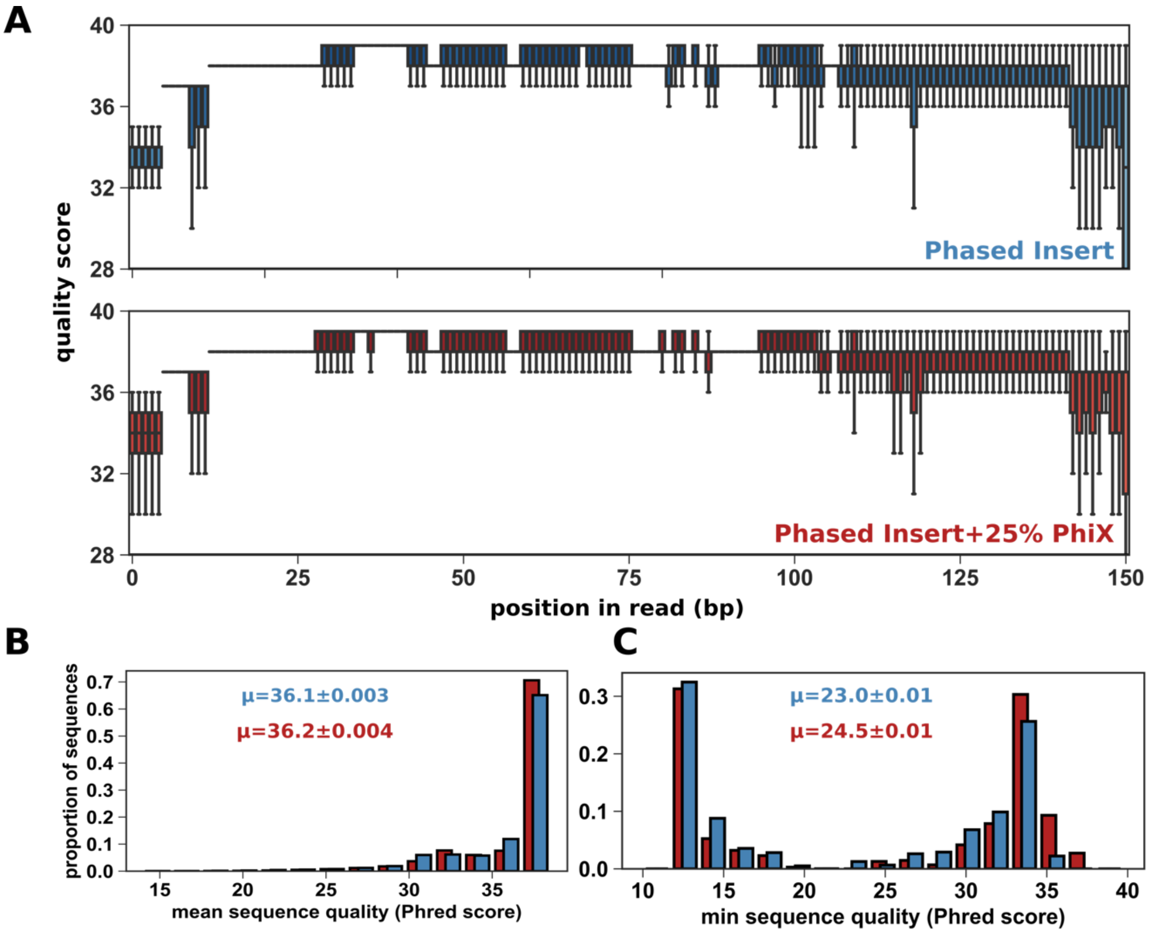


Figure A.3. Sequencing output and quality with phased nucleotide inserts.

(A) Sequencing quality scores per position in read for phased inserts with (red) and without PhiX (blue). Each boxplot represents the interquartile range (IQR) of the dataset and the whiskers extend to the minimum and maximum, excluding outliers (>3IQR difference). Outliers are not depicted. *(B)* Distribution of mean sequence quality (Phred score) for sequences with phased inserts with (red) and without PhiX (blue). *(C)* Distribution of minimum sequence quality (Phred score) for sequences with phased inserts with (red) and without PhiX (blue).

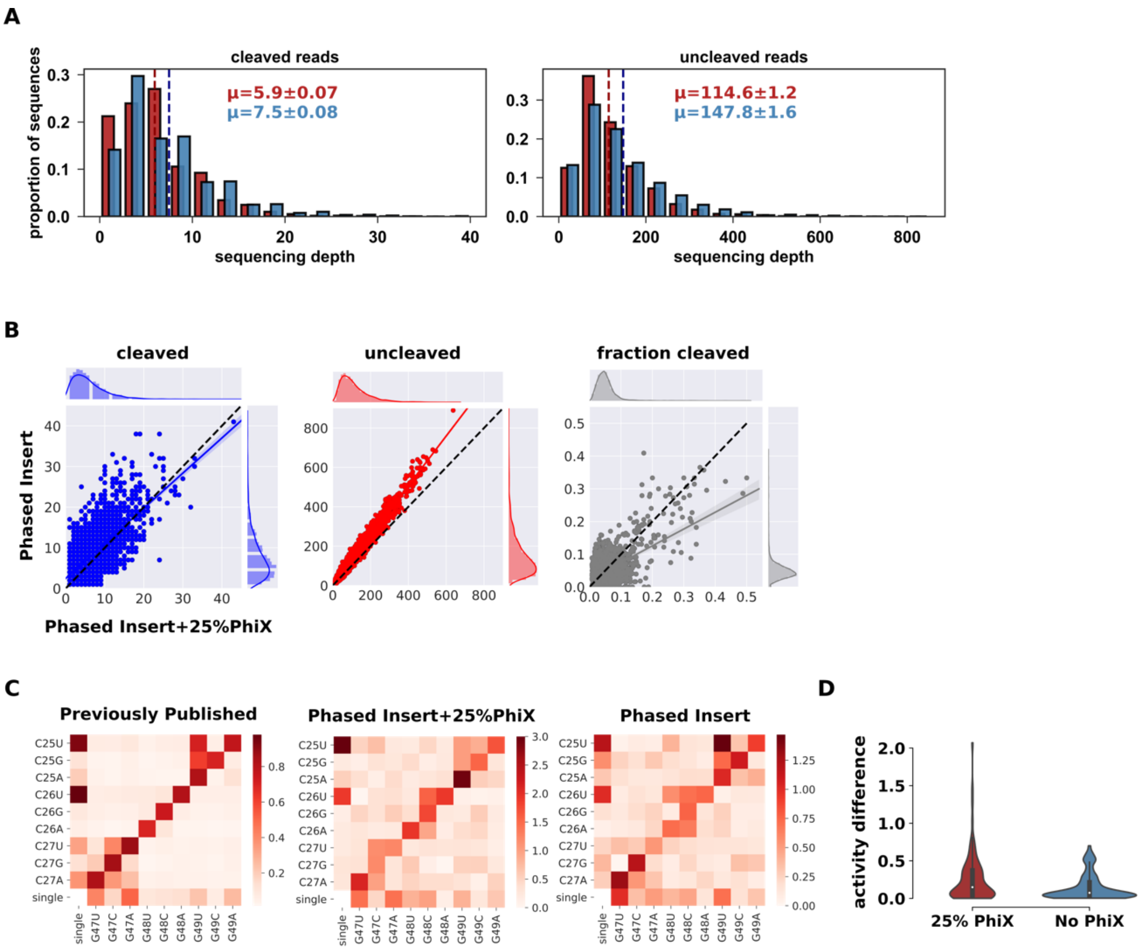


Figure A.4. Comparison of results between data sets.

(A) Distribution of sequencing depths of cleaved and uncleaved reads for phased insert sequencing runs with (red) and without (blue) PhiX. Dashed lines indicate mean sequencing depth. *(B)* Correlation between read counts with PhiX (x-axis) and without PhiX (y-axis). Each unique sequence from the library is plotted as the number of counts of cleaved (blue) and uncleaved (red) sequencing reads. The fraction cleaved for each sequence is also plotted (grey). The dashed line indicates a perfect correlation between the two sequencing runs. *(C)* Heat map visualization of the relative activity of the twister ribozyme mutants with previously published values (Kobori and Yokobayashi 2016). Nucleotide identities of mutations are shown as row and column labels. Double mutants are depicted at the intersection of two mutations. The diagonal contains compensatory double mutations that result in a new Watson-Crick base pair *(D)* Activity differences between previously published data and our current data with PhiX (red) or without PhiX (blue). Activity differences were determined as the absolute value after subtracting our data from the previously published data for each sequence variant in the heat map.

References

Adami C. 2004. Information theory in molecular biology. *Physics of Life Reviews* **1**: 3–22.

Adami Christoph. 2012. The use of information theory in evolutionary biology. *Annals of the New York Academy of Sciences* **1256**: 49–65.

Ameta S, Winz M-L, Previti C, Jäschke A. 2014. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Res* **42**: 1303–1310.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**: 1767–1771.

Daher M, Mustoe AM, Morriss-Andrews A, Brooks III CL, Walter NG. 2017. Tuning RNA folding and function through rational design of junction topology. *Nucleic Acids Res* **45**: 9706–9715.

Dupont DM, Larsen N, Jensen JK, Andreasen PA, Kjems J. 2015. Characterisation of aptamer-target interactions by branched selection and high-throughput sequencing of SELEX pools. *Nucleic Acids Res* **43**: e139.

Ellington AD, Szostak JW. 1990. In vitro selection of RNA molecules that bind specific ligands. *nature* **346**: 818.

Hayden EJ. 2016. Empirical analysis of RNA robustness and evolution using high-throughput sequencing of ribozyme reactions. *Methods* **106**: 97–104.

Hayden EJ, Ferrada E, Wagner A. 2011. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474**: 92–95.

Joyce Gerald F. 2007. Forty Years of In Vitro Evolution. *Angewandte Chemie International Edition* **46**: 6420–6436.

Kobori S, Nomura Y, Miu A, Yokobayashi Y. 2015. High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Research*. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv265 (Accessed July 13, 2015).

Kobori S, Takahashi K, Yokobayashi Y. 2017. Deep Sequencing Analysis of Aptazyme Variants Based on a Pistol Ribozyme. *ACS Synth Biol* **6**: 1283–1288.

Kobori S, Yokobayashi Y. 2018. Analyzing and Tuning Ribozyme Activity by Deep Sequencing to Modulate Gene Expression Level in Mammalian Cells. *ACS Synth Biol*. http://dx.doi.org/10.1021/acssynbio.7b00367 (Accessed January 24, 2018).

Kobori S, Yokobayashi Y. 2016. High-Throughput Mutational Analysis of a Twister Ribozyme. *Angew Chem Int Ed* **55**: 10354–10357.

Krueger F, Andrews SR, Osborne CS. 2011. Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. *PLOS ONE* **6**: e16607.

Levay A, Brenneman R, Hoinka J, Sant D, Cardone M, Trinchieri G, Przytycka TM, Berezhnoy A. 2015. Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment. *Nucleic Acids Res* **43**: e82.

Liu Y, Wilson TJ, McPhee SA, Lilley DMJ. 2014. Crystal structure and mechanistic investigation of the twister ribozyme. *Nat Chem Biol* **10**: 739–744.

Martini L, Meyer AJ, Ellefson JW, Milligan JN, Forlin M, Ellington AD, Mansy SS. 2015. In Vitro Selection for Small-Molecule-Triggered Strand Displacement and Riboswitch Activity. *ACS Synth Biol* **4**: 1144–1150.

Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. 2015. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences* **10**: 18.

Pitt JN, Ferré-D’Amaré AR. 2010. Rapid Construction of Empirical RNA Fitness Landscapes. *Science* **330**: 376–379.

Pressman A, Moretti JE, Campbell GW, Müller UF, Chen IA. 2017. Analysis of in vitro evolution reveals the underlying distribution of catalytic activity among random sequences. *Nucleic Acids Res* **45**: 8167–8179.

Robertson DL, Joyce GF. 1990. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344**: 467.

Rose AS, Hildebrand PW. 2015. NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res* **43**: W576–W579.

Roth A, Weinberg Z, Chen AGY, Kim PB, Ames TD, Breaker RR. 2014. A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nat Chem Biol* **10**: 56–60.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15**: 121–132.

Turner DH, Mathews DH. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* **38**: D280-282.

Weenink T, McKiernan RM, Ellis T. 2017. Rational Design Of RNA Structures That Predictably Tune Eukaryotic Gene Expression. *bioRxiv* 137877.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc* **1**: 1610–1616.

Wilson TJ, Liu Y, Domnick C, Kath-Schorr S, Lilley DMJ. 2016. The Novel Chemical Mechanism of the Twister Ribozyme. *Journal of the American Chemical Society* **138**: 6151–6162.

Acknowledgements

Funding: This work was supported by Boise State University (Biomolecular Sciences Graduate Programs); National Science Foundation Directorate for Biological Sciences [Grant No. MCB-1413664]; National Science Foundation Office of Integrative Activities

[Grant No. OIA-1738865]; and National Aeronautics and Space Administration [Grant No. 80NSSC17K0738].

Supplemental Notes

Failed HiSeq 3000 Sequencing Run Using Hepatitis Delta Virus (HDV) Ribozyme

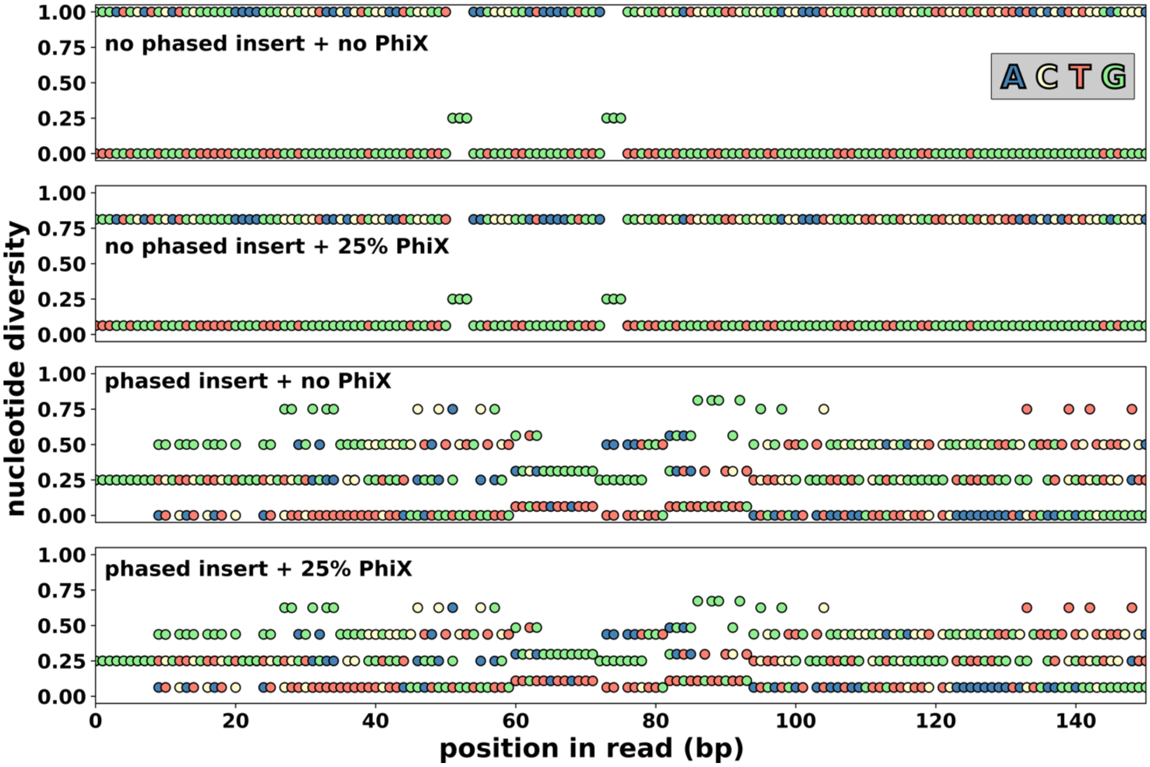
To determine the efficacy of using the phased nucleotide inserts during high-throughput sequencing, we compared our sequencing data from this study to a previously failed sequencing run. The library was based on the Hepatitis Delta Virus ribozyme which, similar to the twister ribozyme, exhibits 5’ self-cleavage activity. The library was prepared in an identical fashion to the twister library in this study, except without phased inserts in the template switching oligos. The sample was then sequenced using Illumina HiSeq 3000 platform with the addition of 10% PhiX. The cDNA library caused a significant amount of issues during cluster identification and resulted in only 12.26% of clusters passing the filter. The sequencing reads that did pass the filter were of low quality and had a mean quality score of 27.9±0.003 (Fig. AS.2). Although this library was sequenced on a HiSeq as compared to the MiSeq platform in this study, both platforms are known to have significant issues with low-diversity samples. Furthermore, it is recommended that PhiX be used for both platforms to increase the nucleotide diversity.

Comprehensive Mutational Analysis of Twister T1 Pseudoknot

In addition to using the twister ribozyme as a model system to validate the benefits of the phased nucleotide inserts, the data also enables a comprehensive evaluation of the T1 pseudoknot. The previously published data included only sequences with one or two mutations in the T1 pseudoknot (154 genotypes). Our current twister library consists of 3,942 additional genotypes, which includes sequences with combinations of 3-6 mutations relative to the wildtype sequence. To understand the relationship between pseudoknot thermodynamic stability and ribozyme activity, we first categorized the genotypes into subpopulations based on the presence of the number of base pairs at the three positions. We plotted the distributions of relative ribozyme activity for each category using our data set obtained without PhiX (Fig. AS.3A). As expected the 64 genotypes that form canonical Watson-Crick base pairs had the highest average relative fitness. This is followed by the three subpopulations that retain two Watson-Crick pairs and a single G-U wobble pair. We note that within this group, the position of the G-U wobble matters. There exists a non-canonical A-A base interaction in T1 that is conserved in >97% of all known twister ribozymes, and which is immediately adjacent to the general base required for the catalytic mechanism (G45) (Wilson et al. 2016). The relative activity of ribozymes decrease on average as the G-U wobble moves closer to the A-A interaction, suggesting that the G-U wobble has a more deleterious effect as it moves closer to the active site. A similar trend was noticed in a randomized stem loop in a HDV-like ribozyme (Kobori and Yokobayashi 2018). As the mismatch mutation came closer to the ribozyme core, the relative activity decreased.

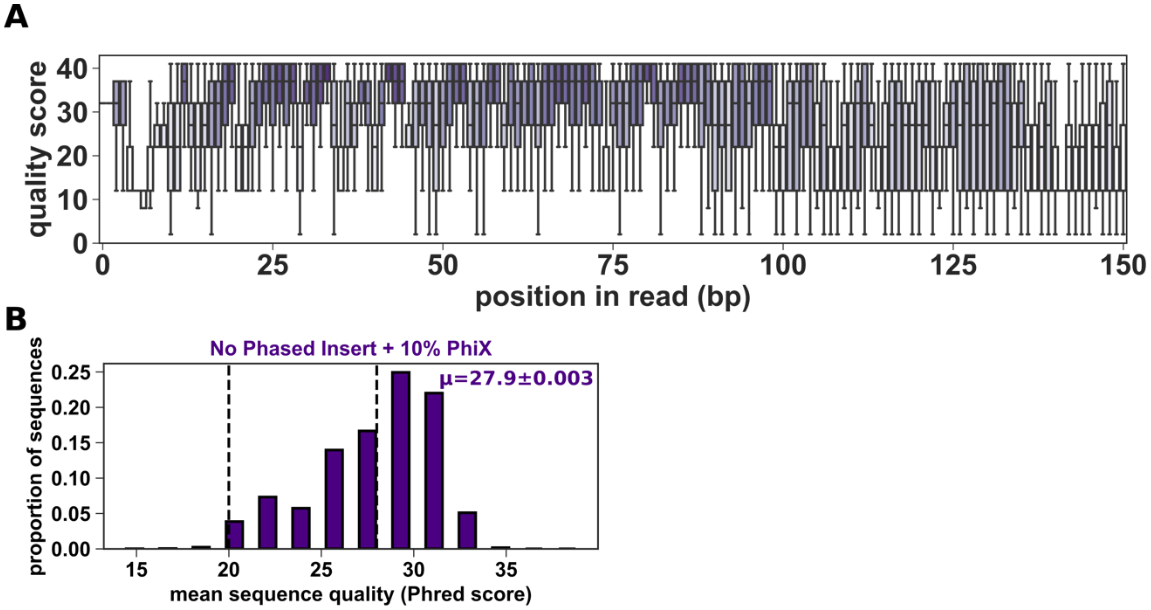
Next, in order to characterize the 64 genotypes that form canonical Watson-Crick base pairs, we calculated the Gibbs free energy for each of these T1 pseudoknots based on nearest-neighbor rules (Turner and Mathews 2010). We plotted ribozyme fitness as a function of free energy (Fig. 5B). The plot shows a negative correlation between the measured relative activity and the change in free energy ($R^2 = 0.23$, $p < 0.0001$, $n = 64$). This data confirms the importance of the stability of the T1 pseudoknot to the overall ribozyme structure and function (Fig. AS.3B). However, we note that the sequences with the highest ribozyme fitness do not have the lowest free energy. This indicates that specific interactions between each T1 sequence and the rest of the ribozyme are also important.

Supplemental Figures



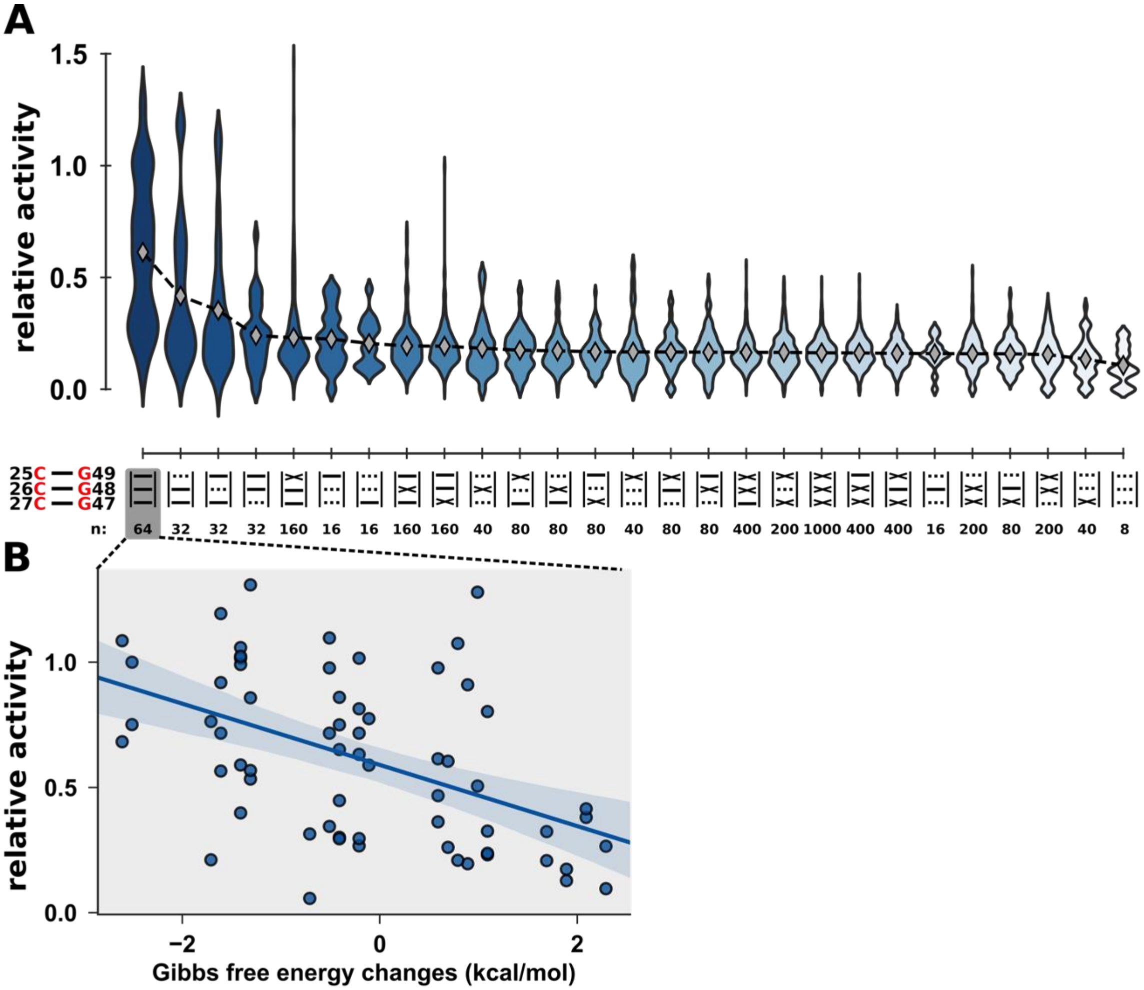
Supplementary Figure A.1. Prediction of nucleotide diversity from simulated sequence run.

Nucleotide diversity was predicted for four simulated twister ribozyme library samples. Predicted nucleotide diversity is shown for a control library without phased nucleotide insertions or PhiX, *only* 25% PhiX, addition only phased nucleotide insertions, or both phased insertions and PhiX. Each diamond indicates the relative proportion of that nucleotide that is present at the position.



Supplementary Figure A.2. Sequencing output and quality for failed sequencing run without phased nucleotide inserts.

(A) Sequencing quality scores per position in read for sequencing run without phased nucleotide inserts and the addition of 10% PhiX/ Each boxplot represents the interquartile range (IQR) of the dataset and the whiskers extend to the minimum and maximum, excluding outliers (>3IQR difference). Outliers are not depicted. *(B)* Distribution of mean sequence quality (Phred score) for sequencing reads.



Supplementary Figure A.3. Relative activities from ribozymes categorized by the composition of base pairs in the T1 pseudoknot.

(A) Symbols on the categorical axis indicate Watson-Crick base pair (solid line), G-U wobble pairs (dashed lines) or mismatch (X). The number *n* below indicates the number of variants in each subpopulation. Dashed line and grey diamonds indicate the mean of each subpopulation. Data is rank ordered by the mean of the relative activity for the category. *(B)* Gibbs free energy changes for the 64 sequences that form three canonical base pairs. Gibbs free energy is calculated from the Nearest Neighbor Database (Turner and Mathews 2010). Line indicates the regression model with 95% confidence interval.

APPENDIX B: EFFECTS OF POPULATION SIZE AND MUTATION RATE ON EVOLUTIONARY SIMULATION

Devin P. Bendixsen[a] & Eric J. Hayden[a,b]

[a]Biomolecular Sciences Graduate Programs, Boise State University, 1910 University Dr.,
Boise, ID, USA.

[b]Department of Biological Sciences, Boise State University, 1910 University Dr., Boise,
ID, USA.

Overview

Evolutionary simulations are a powerful tool for assessing the navigability and accessibility of sequence space. Evolutionary exploration by natural selection is central to Darwinian evolution and is often difficult to assess. Evolutionary simulations have many benefits over simple pathway analyses. Most notably simulations have the ability to cross fitness valleys and we believe most accurately depicts evolution in nature. The rate of adaptation and in particular the ability to escape stasis genotypes isolated by reciprocal sign epistasis can be greatly affected by two simulation parameters: mutation rate and population size. A recent study on the effect of population size on adaptation in empirical fitness landscapes, found that evolutionary dynamics cannot be fully explained by the population mutation rate ($N\mu$, Vahdati and Wagner 2017). Furthermore, contrary to some theoretical theories, even on the most rugged fitness landscapes, small population size was never advantageous over larger population sizes. Simulations on fitness landscapes derived from RNA folding showed that mutation rate (μ), population size (N) or the population mutation rate (μN) could not completely explain the rate of adaptation (Vahdati et al. 2017). This suggests that population size and mutation rate play a role in a very complex system.

The majority of our evolutionary simulations presented in this dissertation, use a constant population size (N) of 1000 individuals and a mutation rate (μ) of 0.01. This results in a population mutation rate of 10. To determine the effects of population size and mutation rate on evolutionary adaptation on empirical RNA fitness landscapes, we tested a range of mutation rates and population sizes. We used the two empirical fitness landscapes presented in Chapter 1. These landscapes represent the sequence space of the Hepatitis Delta Virus (HDV) ribozyme and the class III Ligase ribozyme. These two landscapes form an intersection of innovation and have extensive functional overlap. We selected a single starting genotype for each landscape to start simulations from. For the HDV landscape we chose the genotype that occupied the global peak on the Ligase landscape. And for the Ligase landscape we chose the genotype that occupied the global peak on the HDV landscape. We ran 100 replicates of evolutionary simulation using six different population sizes (25, 50, 125, 250, 500, 1000) and four mutation rates (0.0001, 0.01, 0.1, 1.0). This results in 24 unique combinations. We tracked the mean population fitness (Fig. B.1), mean final diversity (Fig. B.2), unique genotypes explored (Fig. B.3), mean final fitness (Fig. B.4), mean deleterious mutations, mean beneficial mutations (Fig. B.5), and initial rate (Fig. B.6). We report the distributions of these simulation metrics here.

Figures

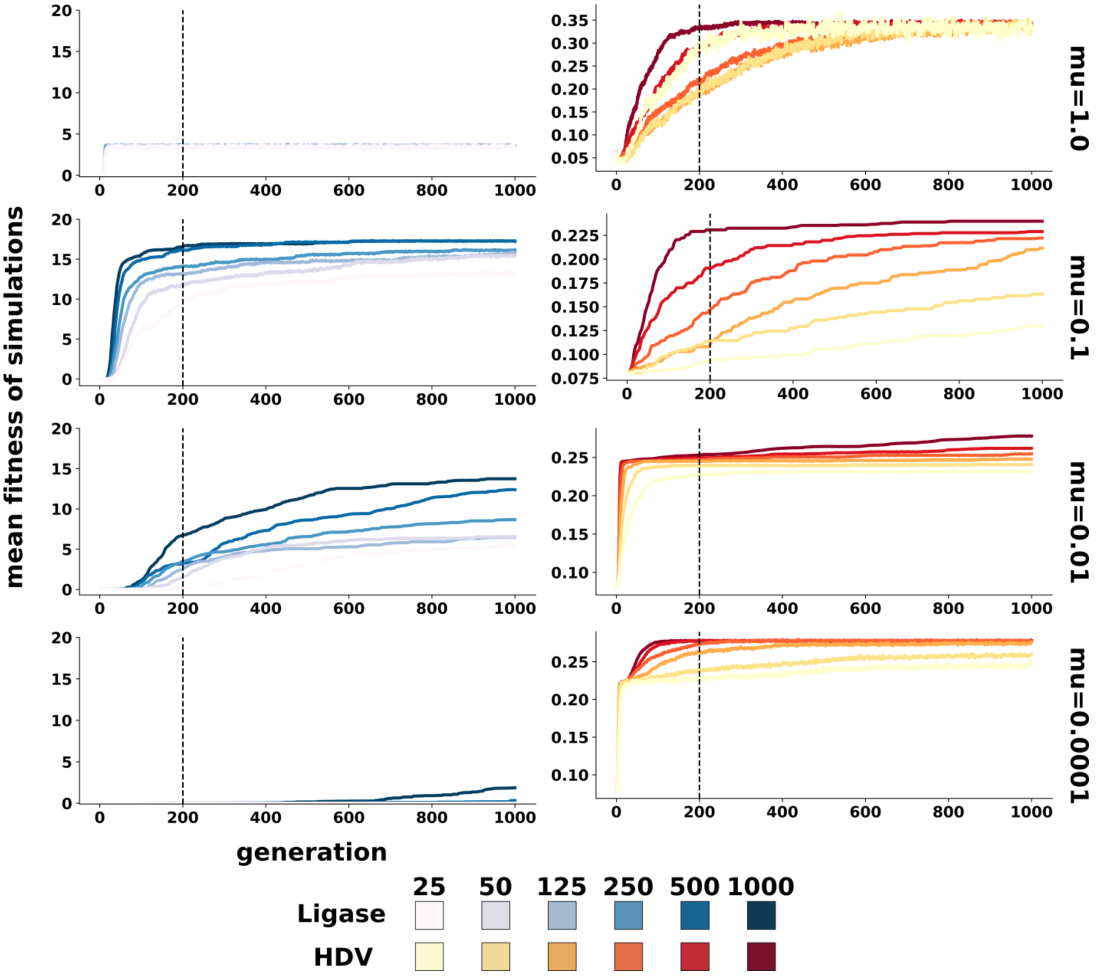


Figure B.1 Mean population fitness during evolutionary simulations.

Each line represents the mean of 100 simulation replicates. Each of the six population sizes are displayed in each plot and mutation rate increases as plot ascend. Simulations on the Ligase landscape are on the left. Simulations on the HDV landscape are on the right.

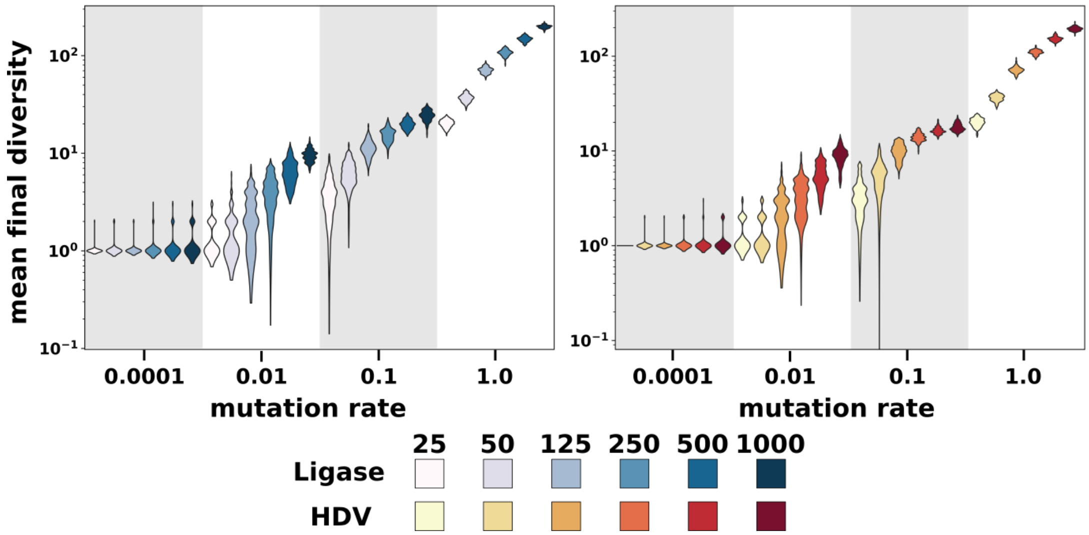


Figure B.2 Mean final diversity following 1000 generations of evolutionary simulations.

Each violin plot represents the distribution of 100 simulation replicates. Each of the six population sizes are displayed according to colors in the legend and mutation rates are on the x-axis.

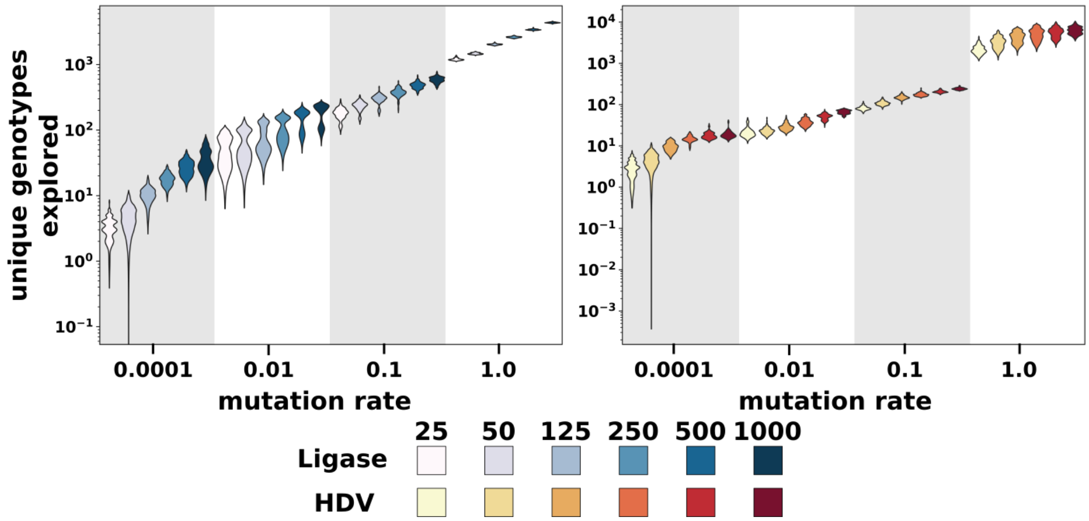


Figure B.3 Mean unique genotypes explored during 1000 generations of evolutionary simulations.

Each violin plot represents the distribution of 100 simulation replicates. Each of the six population sizes are displayed according to colors in the legend and mutation rates are on the x-axis.

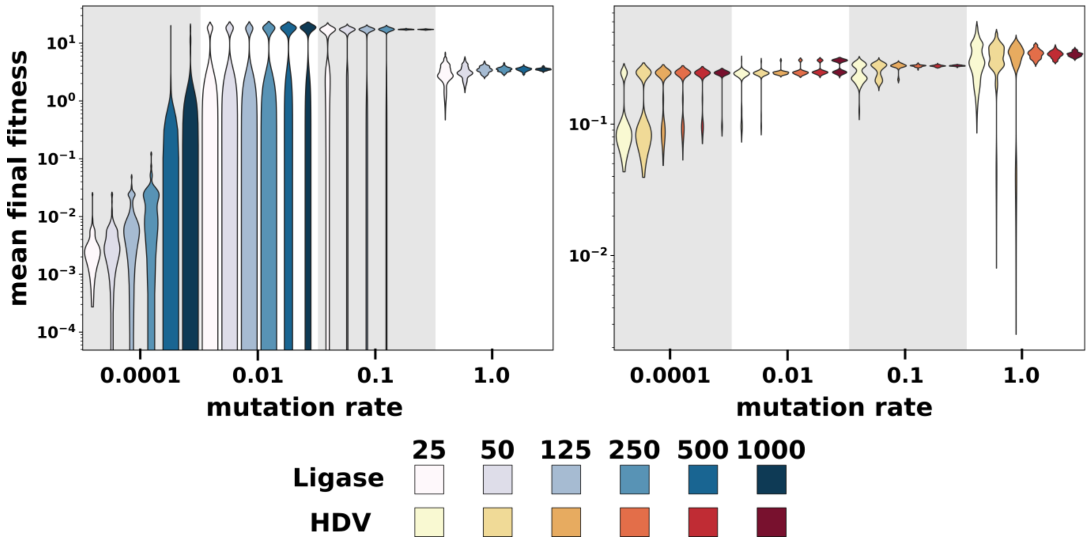


Figure B.4 Mean final fitness following 1000 generations of evolutionary simulations.

Each violin plot represents the distribution of 100 simulation replicates. Each of the six population sizes are displayed according to colors in the legend and mutation rates are on the x-axis.

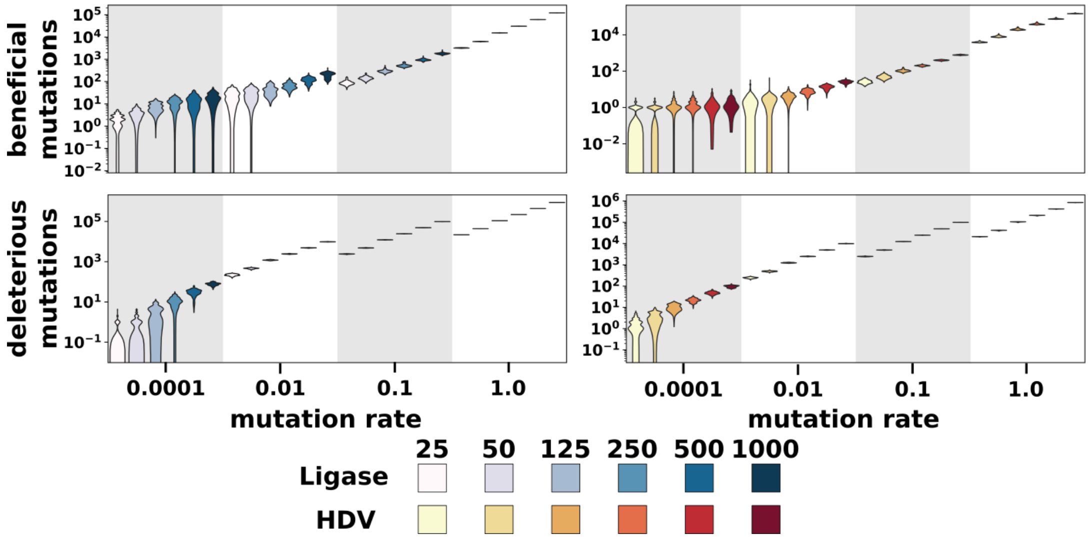


Figure B.5 Mean beneficial and deleterious mutations during 1000 generations of evolutionary simulations.

Each violin plot represents the distribution of 100 simulation replicates. Each of the six population sizes are displayed according to colors in the legend and mutation rates are on the x-axis.

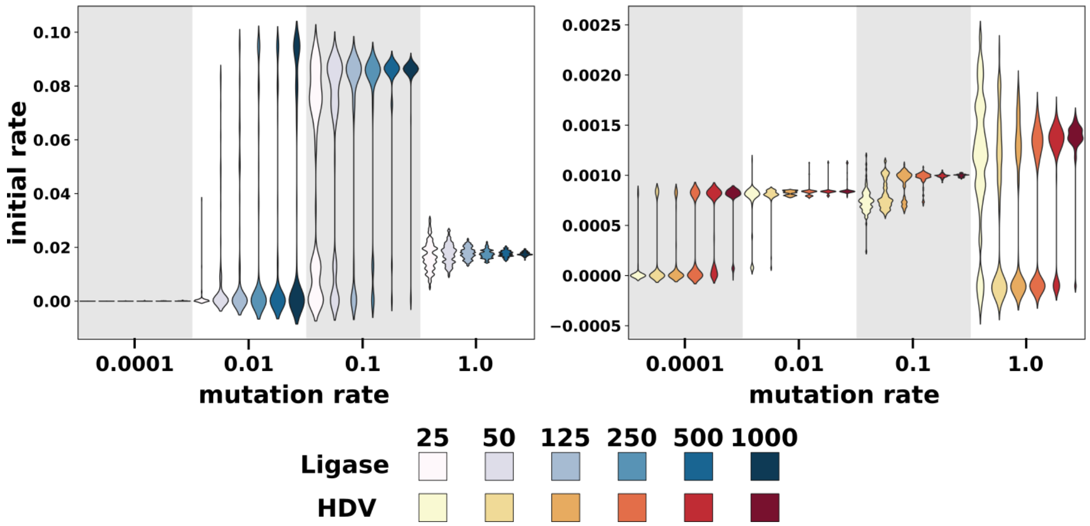


Figure B.6 Mean initial rate of adaptation during the first 200 generations of evolutionary simulations.

Each violin plot represents the distribution of 100 simulation replicates. Each of the six population sizes are displayed according to colors in the legend and mutation rates are on the x-axis.

References

Vahdati AR, Sprouffske K, Wagner A. 2017. Effect of Population Size and Mutation Rate on the Evolution of RNA Sequences on an Adaptive Landscape Determined by RNA Folding. International Journal of Biological Sciences 13:1138–1151.

Vahdati AR, Wagner A. 2017. Population Size Affects Adaptation in Complex Ways: Simulations on Empirical Adaptive Landscapes. Evolutionary Biology [Internet]. Available from: http://link.springer.com/10.1007/s11692-017-9440-9

APPENDIX C: MODELING THE EVOLUTION OF PROMISCUITY

Devin P. Bendixsen[a] & Eric J. Hayden[a,b]

[a]Biomolecular Sciences Graduate Programs, Boise State University, 1910 University Dr., Boise, ID, USA.

[b]Department of Biological Sciences, Boise State University, 1910 University Dr., Boise, ID, USA.

Overview

The evolution of novel function (innovation) was covered in Chapter 1 by closely examining the intersection of two ribozyme genotype networks: self-cleaving (HDV) and self-ligating (ligase). A potentially important aspect of innovations that is not fully covered is the role that promiscuous activity could play in evolving a new function. Functional promiscuity is when a gene or in our case a ribozyme is able to develop a second function albeit at low function while not significantly reducing the efficiency of the original function. Promiscuity is seen in proteins (Babtie et al. 2010; Espinosa-Cantú et al. 2015; Khanal et al. 2015) and can be perceived as being very advantageous in terms of evolutionary potential. If the environment changes and the new function is positively selected for, then it would have a head-start on optimizing the new function. This could also occur following a gene duplication event, where divergence occurs prior to the duplication (Ohno 1970; Taylor and Raes 2004; Bergthorsson et al. 2007; Andersson et al. 2015). Using the empirical fitness landscapes constructed in Chapter 1 we developed many evolutionary simulations to assess and model the evolution of promiscuity. We first determined the total fitness (W_{total}) of a sequence as a product of its ligase function and its HDV function using this equation : $W_{total} = \beta_{HDV} * \left(\frac{W_{HDV}}{\max(W_{HDV})}\right) + \beta_{LIG} * \left(\frac{W_{LIG}}{\max(W_{LIG})}\right)$. By weighting the beta values for HDV and LIG on a sliding scale, we generate 11 unique fitness landscapes with different topographies (Fig. C.1). We then identified a random starting population that was of equal distance from the HDV reference genotype and the LIG reference genotype. We then plotted these genotypes HDV and ligase function on the original HDV-LIG landscape presented in Chapter 1 (Fig. C.2). We then simulated evolution on each of the 11 unique co-selection landscapes. We repeated this simulation for 100 replicates per landscape and plotted the Ligase function, the HDV function and the total function (Fig. C.3). It appears that there was a *stasis* genotype at ~0.19 on the ligase landscape that often stopped evolutionary exploration. We also plotted the individual traces for the 100 replicates for three of the landscapes. The most extreme HDV landscape, the middle 50-50 landscape and the most extreme ligase landscape (Fig. C.4).

Figures

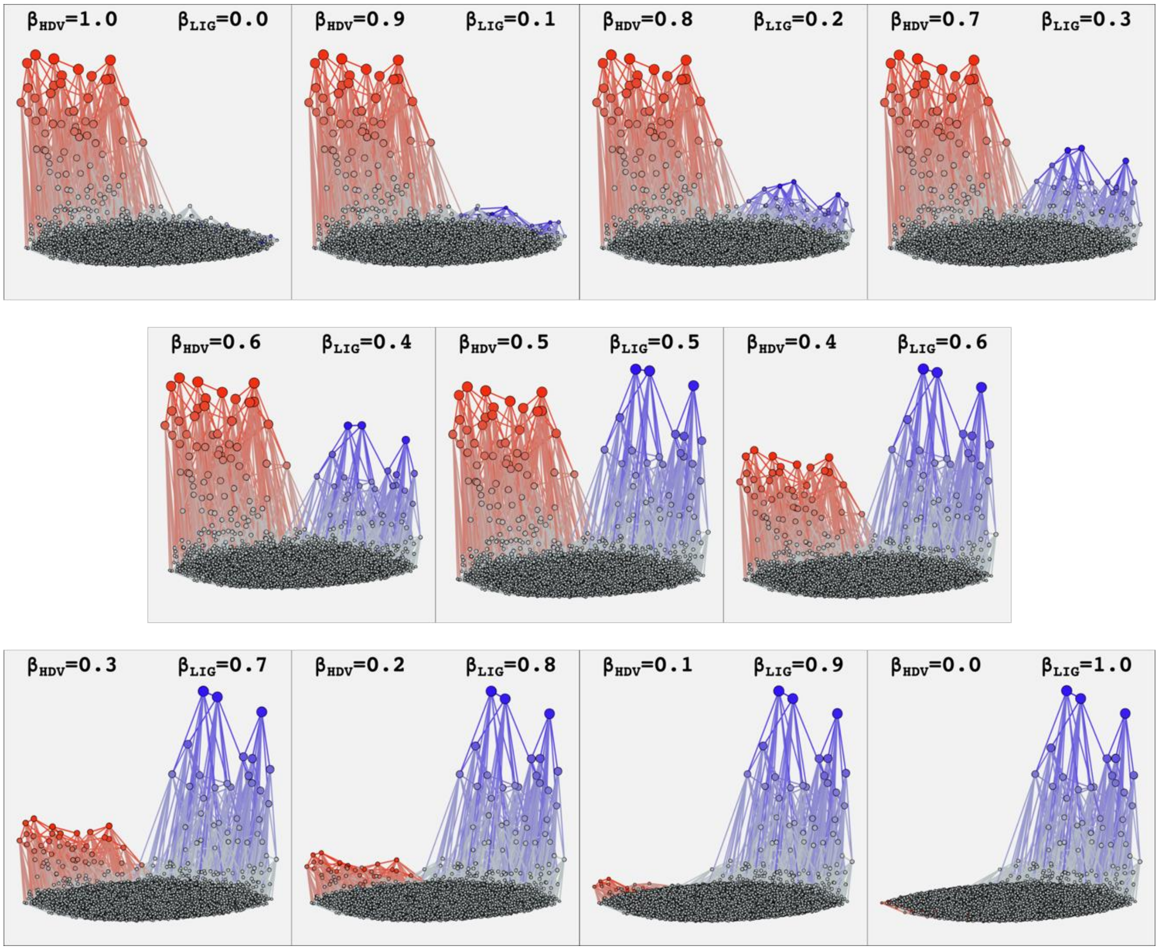


Figure C.1 HDV-LIG co-selection landscapes.

The 11 landscapes presented each have differing beta values for HDV and ligase function, resulting in conformational changes. The nodes with dominant HDV function are colored in red and the nodes with dominant ligase function are colored in blue. Nodes that differ by a single mutation are connected by an edge. Total fitness was calculated as a function of both functions: $\boldsymbol{W_{total}} = \boldsymbol{\beta_{HDV}} * \left(\frac{\boldsymbol{W_{HDV}}}{\mathbf{max}(\boldsymbol{W_{HDV}})}\right) + \boldsymbol{\beta_{LIG}} * \left(\frac{\boldsymbol{W_{LIG}}}{\mathbf{max}(\boldsymbol{W_{LIG}})}\right)$.

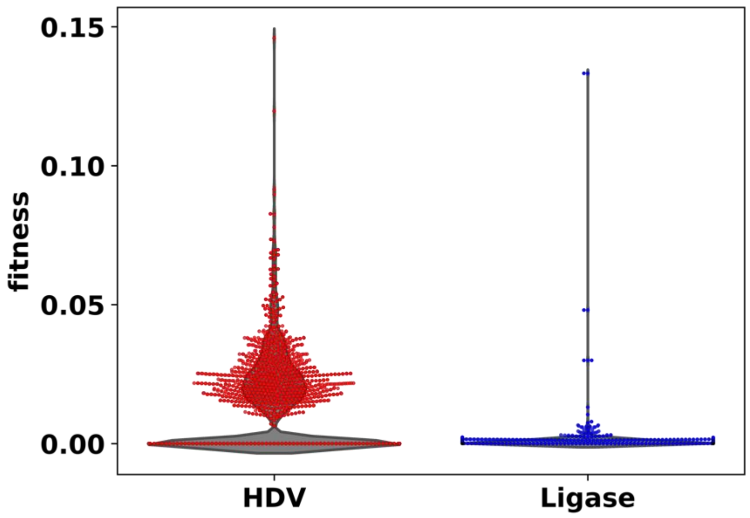


Figure C.2 Starting population for co-selection evolutionary simulations.

The random population was selected from all of the sequences that were of equal distance from the HDV reference genotype and LIG reference genotype.

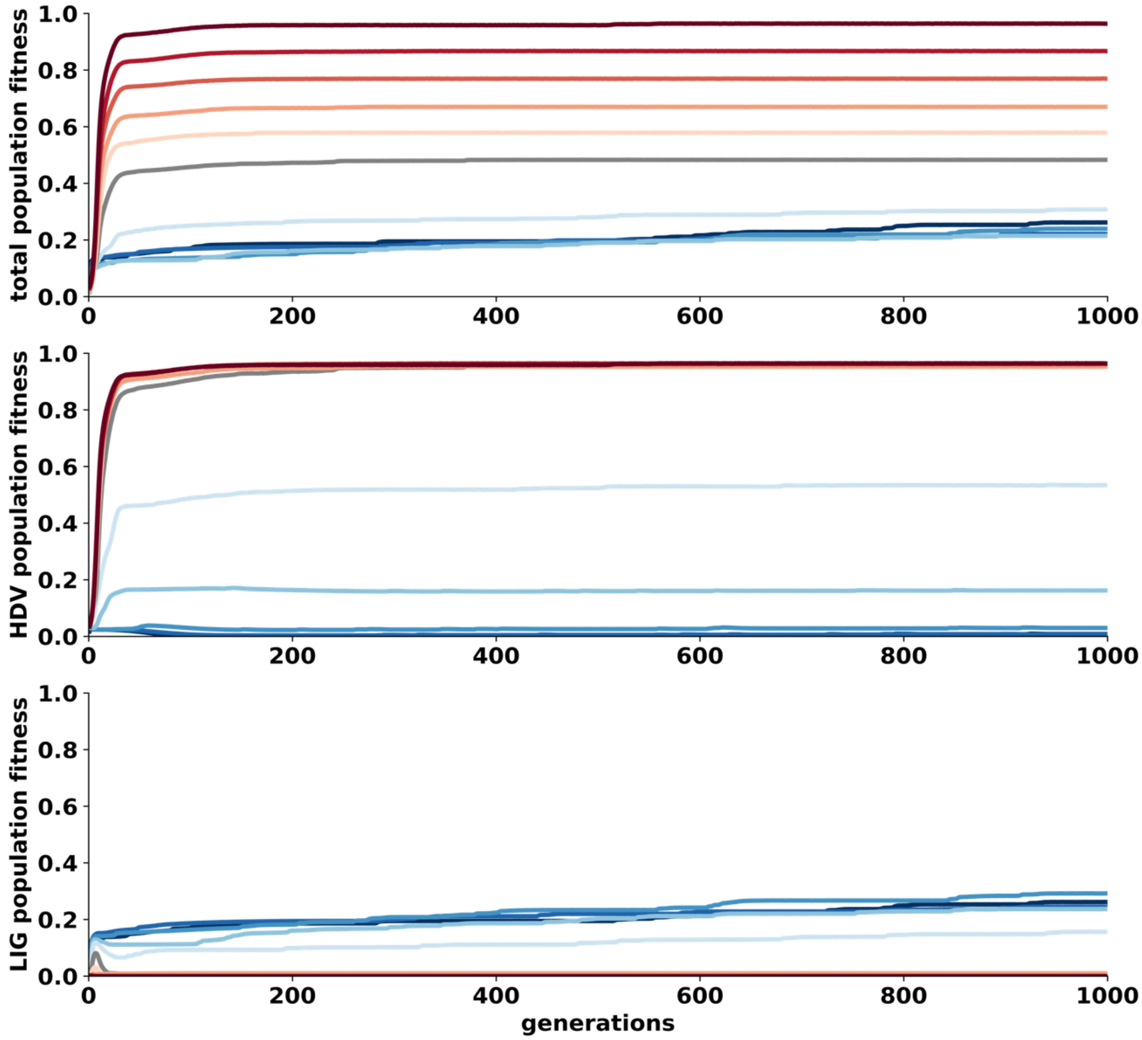


Figure C.3 Evolutionary simulations on the HDV-LIG co-selection landscapes.

The ligase, HDV and total fitness are displayed on the y-axis and the number of generations are on the x-axis. Each line indicates the average of 100 replicates. Red lines indicated the most extreme HDV landscape and blue lines indicated the most extreme ligase landscape.

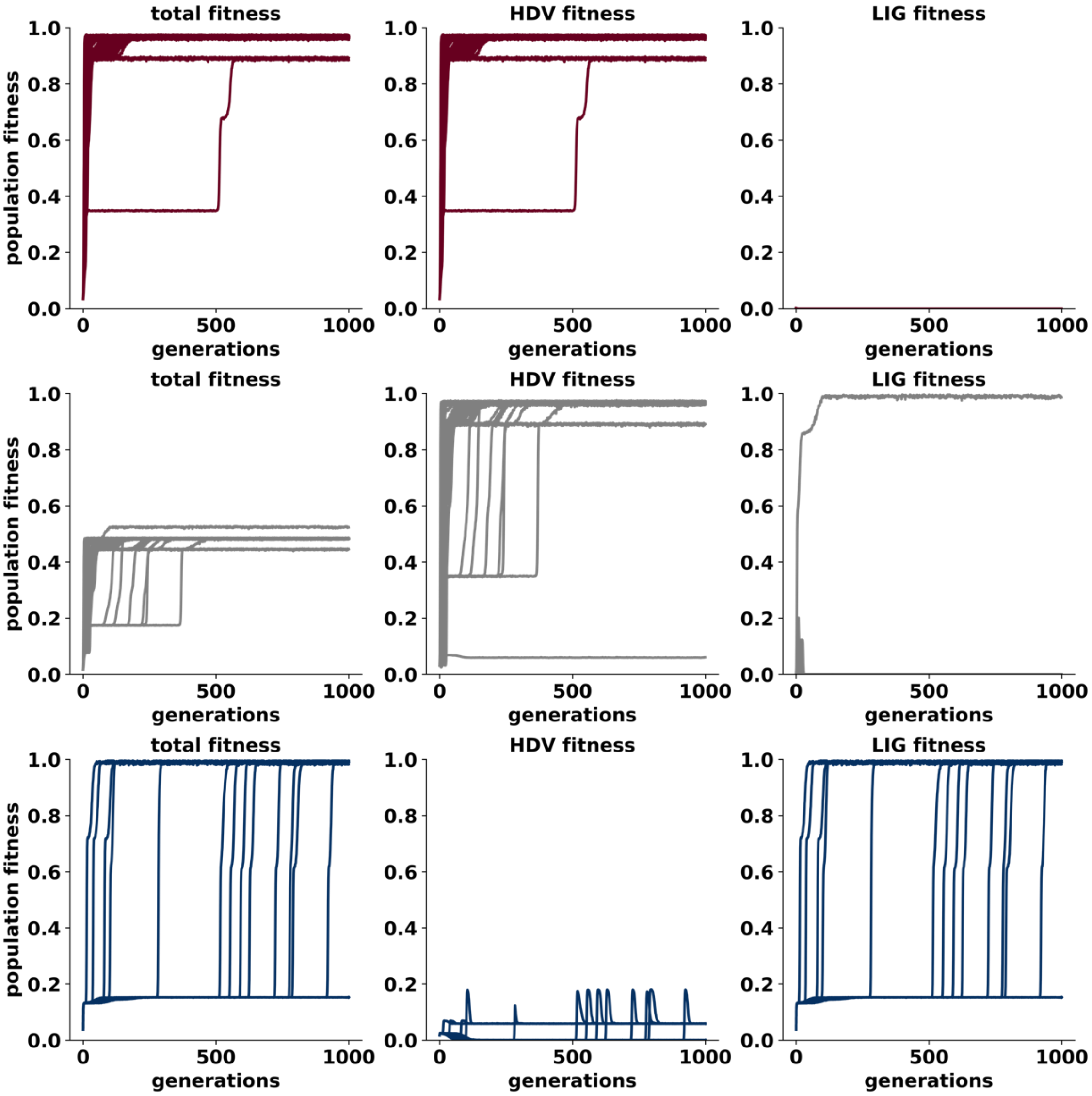


Figure C.4. Individual traces from evolutionary simulations.

The most extreme HDV landscape is shown in red, the 50-50 landscape is shown in grey and the most extreme ligase landscape is shown in blue. Total fitness, HDV fitness and Ligase fitness are shown for each trace.

References

Andersson DI, Jerlström-Hultqvist J, Näsvall J. 2015. Evolution of new functions de novo and from preexisting genes. Cold Spring Harb Perspect Biol 7.

Babtie A, Tokuriki N, Hollfelder F. 2010. What makes an enzyme promiscuous? Current Opinion in Chemical Biology 14:200–207.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno’s dilemma: Evolution of new genes under continuous selection. PNAS 104:17004–17009.

Espinosa-Cantú A, Ascencio D, Barona-Gómez F, DeLuna A. 2015. Gene duplication and the evolution of moonlighting proteins. Front. Genet. [Internet] 6. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2015.00227/full

Khanal A, McLoughlin SY, Kershner JP, Copley SD. 2015. Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution. Mol Biol Evol 32:100–108.

Ohno S. 1970. Evolution by Gene Duplication. In: Evolution by Gene Duplication. Springer, Berlin, Heidelberg. p. 1–2. Available from: https://link.springer.com/chapter/10.1007/978-3-642-86659-3_1

Taylor JS, Raes J. 2004. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annual Review of Genetics 38:615–643.