DEVELOPING ULTRA-HIGH THROUGHPUT SEQUENCING BASED ASSAY FOR LIGASE RIBOZYMES FOR THE STUDY OF EVOLUTIONARY INNOVATIONS

by

James Collet

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Biology

Boise State University

December 2018

James Collet

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

James Collet

Thesis Title: Developing Ultra-High Throughput Sequencing Based Assay for Ligase Ribozymes for the Study of Evolutionary Innovations

Date of Final Oral Examination: 7 June 2018

The following individuals read and discussed the thesis submitted by student James Collet, 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
Kevin Feris, Ph.D.	Member, Supervisory Committee
Shane Panter, M.S.	Member, Supervisory Committee

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

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Eric Hayden, who made this study possible. Thank you for giving me the opportunity to work with you during the pioneer days of your BSU lab, when I could barely run a gel. You were always supporting, and never harsh, almost to a fault. You helped guide me through the rough times, and taught me many important lessons. Thank you for sharing your valuable time.

I would also like to thank my committee members Dr. Kevin Feris and Shane Panter, who agreed to give their time and advice to help me reach my goal. To my lab member and co-researcher Devin Bendixsen, thank you for your hard work and cheerful spirit. Also, thank you Dr. Bjorn Ostman for your contribution to the project. Thank you to all the other lab members with and without roles in the project.

Thank you, Boise State University Biomolecular Sciences Graduate Programs, National Science Foundation Directorate for Biological Sciences, National Science Foundation Office of Integrative Activities, and National Aeronautics and Space Administration for helping to fund the project.

Finally, I would like to thank my parents and siblings for their continuous support and tough love. They always pushed me, and although I sometimes resisted, I knew it was for the best. Hopefully, I will make everyone proud yet.

ABSTRACT

The study of evolutionary innovations, or novel traits, is integral to understanding evolution yet is poorly understood. By studying the regions between genotype networks that produce the same phenotype, we can better characterize the process by which innovation occurs. The overarching goal of this study is to assign fitness values to the overlapping genotype network of two catalytic RNA molecules, or ribozymes. Properly characterizing this region requires the study of thousands of individual sequences, which is achievable through the use of high-throughput sequencing analysis. This thesis focuses on developing assays for one of the ribozymes, the ligase ribozyme. Due to the low activity of this ribozyme, the best method for detection proved to be through qPCR amplification and fluorescence detection. This method allows for troubleshooting before sequencing, as well as validation afterward. It was adapted to create the sequencing assay for the ligase ribozyme, which turned out to be comprehensive in its coverage of the tested space, and reproducible between runs. Several ligase sequences were discovered with higher activities than the previous best ligase ribozymes. In addition, the utilization of phased primers proved to be very successful in increasing initial-read diversity in the sequencing pool. The work developing assays for the ligase ribozyme successfully contributed to the larger study on innovation.

Key words: ribozyme, fitness landscape, innovation, high-throughput sequencing, RNA, genotype network, phenotype, evolutionary trajectory

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEVELOPING ULTRA-HIGH THROUGHPUT SEQUENCING BASED ASSAY FOR LIGASE RIBOZYMES FOR THE STUDY OF EVOLUTIONARY INNOVATIONS

1. Introduction

It is common to describe evolution as an optimization process where continuous cycles of reproduction, mutation, and selection lead to the steady improvement of traits to match specific environments. However, sometimes anomalies arise that are unlike anything seen before. Occasionally, these anomalies can be classified as innovations – qualitatively novel traits that enable organisms to inhabit completely new environments (Moczek et al. 2011). The primary question of this thesis is not, “How are advantageous traits selected for?” but, “How do new traits arise to begin with?” The mechanisms of the evolution of innovation remain elusive to experimentation.

Innovation is easy to recognize, but difficult to define. Some definitions include “any newly acquired structure or property that permits the assumption of a new function,” (Mayr 1960) and “a qualitatively new structure with a discontinuous origin, marking a relatively abrupt deviation from the ancestral condition” (Müller and Nitecki 1990). The glowing light of a firefly and the eyespots on a moth’s wings are good examples of easily recognizable innovations in nature. The line gets fuzzy when considering the narwhal’s enlarged tooth or the protrusions on a Hercules beetles’ carapace (Moczek 2008). These could be considered modifications of preexisting structures, and may not be an abrupt enough deviation to be true innovation. When are they new structures, and when are they just glorified old ones?

In order to define novelty in biology, it is useful to explore the opposite yet more familiar concept of homology. Two structures are homologous if they have a shared genetic ancestry. The forelegs of beetles and flies are homologous to each other, and the front, middle, and back legs are serially homologous, but these are obvious examples of homology. In order to separate more ambiguous homologous traits from non-homologous traits, additional insight must be gathered, such as from developmental biology (Shubin et al. 2009). Looking at similarities between structures at different stages of an organism's development sheds light on which of these structures are homologous and which are not. Another place to look is gene expression. If similar genes are expressed to produce different structures, they are derived from a common gene and can be considered homologous. By using these tools to determine what is homology, we can also determine what is innovation.

The macroscopic novelties described above are generated by changes at the level of genes and their protein and RNA products. For example, regulatory circuits can be a tool that leads to innovation. Changes to macromolecular structures, such as transcription factors or their DNA targets, can result in the upregulation of genes in certain tissues of an organism. This can create novel structures, such as the eye spots on butterflies (Monteiro 2015). In addition, new environmental challenges can be overcome by simply mixing and matching elements currently present in the organism to produce a novel metabolic network (Dantas et al. 2008). Many bacteria have been studied that have created new pathways to utilize new carbon sources, or to break down hazardous compounds, by reshuffling a metabolic pathway (Rehmann and Daugulis 2008). Finally, even changes in simple macromolecules can produce significant innovations, such as antifreeze proteins that enable organisms to survive freezing (Fletcher et al. 2001). At the core, each of these examples requires a different type of change at the genetic level, which leads to innovation that is observable at larger, macroscopic scales.

For a discussion of genetic innovations, it is useful to define the concept of *genotype space*. Each organism has a *genotype*, or a specific DNA genome. Changing any DNA base in an organism's genome can be thought of as a step in the space of all possible DNA genomes, or *genotype space*. Many of these genotypes code for the same *phenotypes*, or physical characteristics. Sometimes there are genetic steps that result in new phenotypes. These are the regions of innovation (Figure 1). When selective pressures are added, the *fitness*, or environmental viability, of each phenotype must also be taken into account. The probability of a mutational path through genotype space toward a specific innovation depends upon the fitness of each intermediate mutational change (Bendixsen et al. 2018). Generally, in order to innovate, there must exist one or more paths through genotype space from one phenotype to another, and the associated fitness values along those paths must be high enough to survive the current environment. Although at first glance it seems unlikely that both of these requirements could be satisfied concurrently with any reasonable probability, the high dimensionality of genotype space and the genotypic redundancy of most phenotypes make it plausible.

Genotype space contains a mesh of entwining phenotypes with many points of contact, and these contact points enable innovation. Even a small gene that is only 150 nucleotides long has 2 x 10^{90} possible sequences, which is greater than the total number of atoms in the known universe. Many of these genotypes share the same phenotype, and not just those genotypes that are different by only a few mutations. Genotypes can often be more different than similar and still share the same phenotype, so each phenotype is sprawled far across genotype space with many contact points to other phenotypes (Wagner 2011). In other words, if a functional gene has many possible sequences, each of those sequences (or more precisely, sequence clusters) has different neighboring genes with different functions. Going back to our 150 nucleotide gene, each sequence in this space has 149 neighbors, allowing for a nearly infinite number of trajectories, or mutational pathways, from one sequence to another. This results in a vast, highly connected web of genotypes with phenotypes that can span most of the space. The main hinderance to nature's exploration of this space is the fitness requirements of survival. But as selection pressures relax or supporting structures such as chaperones evolve, lower fitness genotypes can survive, opening more trajectories through this space for exploration (Figure 2). When conditions relax, existing phenotypes gain contact points to other phenotypes, enabling innovation.

In order to study innovation on a large scale, a model system is required in which the fitness of numerous genotypes can be studied quickly and accurately. Catalytic RNA molecules, or ribozymes, offer a tractable system with both genotype and phenotypes. The genotype of RNA is defined as the order of nucleotides. Unlike DNA, RNA is generally single stranded, and thus can fold and base pair intramolecularly. The phenotype of RNA is defined as the structure resulting from these intramolecular interaction, or the biochemical function of this structure. Ribozymes are typically small RNA molecules that can be produced in large quantities through in-vitro transcription. Their functions can often be visualized through simple laboratory methods such as gel electrophoresis, qPCR, and next-generation sequencing, the last of which is the best tool available for exploring relatively large pieces of genotype space in high-throughput. These effects can be measured relative to each other, resulting in a “fitness” measurement. This ease of production, visualization, and comparison makes them excellent models to study innovation, without the maintenance and maturation requirements associated with studying living organisms. The study of ribozymes is important for understanding their roles in Nature (Ren et al. 2017), and also as a tool in biotechnology as gene regulators and reporter molecules (Lee et al. 2018). Thus, ribozymes lend themselves to the study of innovation.

When studying innovation, the best approach is to explore the boundaries between distinct phenotypes. In order to simplify the problem to a manageable project, we chose two previously characterized Ribozyme phenotypes of similar genotypic length (Schultes and Bartel 2000). The first was adapted from the self-cleaving ribozyme originally found in the Hepatitis Delta Virus (Figure 3). The second is a modified version of the class III ligase, a ribozyme artificially selected from a pool of random sequences. The HDV phenotype cleaves a piece of itself near its 5’ end, while the Ligase phenotype ligates a substrate RNA to itself. Although these functions are opposites, they require significantly different folds, and have different origins. The landscape of active sequences for each function come into close proximity. The goal of this research is to create as many relevant sequences as possible near the intersection, test their activities, and utilize this data to help characterize the boundaries of genetic innovation.

Studying the boundaries of innovation *in vitro* is novel and informative. Previous research studying fitness landscapes of ribozymes has been limited by technology. Researchers are forced either to limit the size of their ribozymes or their coverage of the landscape. Alternatively, computational models can give a theoretical idea of the space, but are only useful insofar as they are accurate. As next-generation sequencing technology improves, though, more of the space can be studied for larger ribozymes (Pitt and Ferré-D'Amaré 2010; Jiménez et al. 2013). This high-throughput process allows millions of sequences to be read in parallel, and with sufficient coverage of each genotype, is able to detect low activity ribozymes accurately. Although these strategies have been utilized before, exploring the space of innovation between two phenotypes is novel and may reveal insight into how innovation occurs. This requires new assays that can detect very low activity ribozymes accurately before sequencing. The objective of my thesis research is to create these assays, develop the sequence library, and analyze the resulting data, especially in regard to the ligase ribozyme.

2. Methods

2.1 Ribozyme synthesis by in vitro transcription

A ssDNA template library was ordered and diluted to 10 µM in TE8 buffer. The template contained a 92 base region adapted from the class III ligase with 14 ambiguous bases at specific locations. A non-interacting 30 base universal primer binding site was added to the 3' end of the template, and a T7 polymerase promoter sequence to the 5' end. The promoter sequence was made double stranded by annealing 20 pmol of a T7 primer to 20 pmol of the template by heating to 98°C for 2 minutes and cooling to room temperature. This template was transcribed to RNA using an Ambion Megashortscript kit. Transcription reactions were incubated at 37°C for 2h followed by a 20-minute DNase treatment, also from Ambion. The resulting RNA was purified using a silica-based column (Zymo), then split into triplicate groups.

2.2 Reaction and reverse transcription

500 pmol of a synthetic oligonucleotide substrate comprised of DNA except for the last 9 bases on the 3'-end was reacted with approximately 400 pmol of purified RNA for two hours at 37°C in a 10 mM magnesium and 80 µM Tris pH 7.5 buffer. Approximately 100 pmol of RNA from the completed reaction was then reverse transcribed at 42°C for 90 minutes using 20 pmol of the RT primer, which was complimentary to the 3'-end of the RNA (Table 1). The resulting DNA was purified using a silica-based column (Zymo).

2.3 Sequencing preparation

cDNA from reacted sequences was amplified in a PCR reaction using 10 pmol of a primer specific to the substrate sequence (Table 1). Illumina specific adapter sequences were added through PCR, then column purified and sent off for Illumina Nextseq 500 single-end 150 base sequencing at the University of Oregon. Both PCR reactions were stopped just before saturation.

2.4 Analysis of activity

Counts of each variant were tallied from the sequencing data by parsing fastq data using custom python scripts. The fraction reacted was calculated as the ratio of ligated sequences to the total for each.

2.5 Activity measurement via qPCR

Prior to an HTS run, key individual sequences' activities were measured via qPCR. Starting with cDNA for an individual variant, primers specific to either the substrate sequence or general to all sequences were used in a qPCR amplification. Fraction reacted was calculated from the difference in Cq between ligated and total groups, after accounting for primer efficiency. Ligated primer Efficiency (LE), Ligated ribozymes Cq (LCq), Unligated primer efficiency (UE), Unligated ribozymes Cq (UCq):

$$\frac{\mathrm{LE}^{\mathrm{LCq}}}{\mathrm{UE}^{\mathrm{UCq}}}$$

3. Results and Discussion

3.1 Library Design

We needed to design a library of sequences that would be large enough to cover a relevant portion of genotype space, while at the same time small enough to be analyzed through high-throughput sequencing. Ideally, we would have liked to study the entire genotype space between the wild-type HDV ribozyme and the original class III ligase ribozyme. Unfortunately, we were limited by current high-throughput sequencing technology, since a ribozyme of a hundred random bases contains 4^{100} or 1.6 x 10^{60} sequences, and current Illumina sequencing technology can read around 1.6 x 10^{8} sequences per run. Since we were mainly concerned with the region of sequence space near the intersection between the two functions, we rationally designed an endpoint candidate for each phenotype by studying the base-pairing of the ribozymes and the effects of previously studied individual mutations. These endpoints were separated by fourteen mutations, which would guarantee that the entire library could be captured in a single high-throughput sequencing run with sufficient coverage of each sequence. By randomizing these mutational differences between the bases from each sequence, we created a pool of 2^{14} = 16,384 unique sequences. Since our ribozyme had ambiguous regions near the ends, a primer binding site was required for reverse transcription, PCR amplification, and the addition of Illumina primers. We used a self-interacting universal primer binding site that would not interact with the rest of the ribozyme (Wilkinson et al. 2006). Since ordered DNA oligonucleotides can contain ambiguous bases, we could order our whole library in a single tube. Since current high-throughput sequencing provided around 100 million reads, and our library contained 16,384 sequences, each sequence had over 1,000 reads. Thus, by comparing sequencing runs measuring Ligase and HDV activity separately, we were able to sample a large portion of genotype space and study the region of innovation.

3.2 Ligase qPCR Assay allows low-activity detection

In order to test individual sequence activities, we developed a quantitative polymerase chain reaction, or qPCR, assay. After failing to visualize the Ligase ribozyme's activity on a gel, we attempted to use substrate oligonucleotide sequence modified with a fluorophore on the 5' end. This allowed us confirm the presence of ligated ribozymes, but did not allow quantification because the unligated RNA was not detectable. Thus, we decided to try reverse transcribing and amplifying the ligated and unligated fragments separately. Following end-point PCR, we were able to observe bands on gel corresponding to the size expected for ligated and unligated ribozyme PCR-products. This strategy turned out to be very successful, as can be seen by the easily detectable difference in size between the ligated and unligated lanes in Figure 4. To quantify the amount of each species, we proceeded to adapt the strategy to RT-qPCR to verify ligation activity of many sequences concurrently with good accuracy.

In order to generalize this strategy to other ribozyme variants, though, a generic reverse transcription adapter sequence was necessary. This way, rather than having to order a different reverse transcription primer for every new ribozyme, a consistent primer could be used. We tested a sequence previously reported to be non-interfering and self-interacting, two traits we were looking for (Wilkinson et al. 2006). We found the sequence to have negligible effect on ligation rate, with a log transformed relative ligation of 6.7 compared to the original primer's relative ligation of 5.8 (Figure 5a).

The final qPCR assay involved transcribing, ligating, and reverse transcribing sequences of interest, then splitting the resulting DNA into two qPCR amplifications performed using different sets of reverse primers. The first set attached inside the ribozyme and amplified all ribozymes (Figure 6a). The second attached to the substrate sequence of the ligase, and amplified only the ligated ribozymes (Figure 6b). Dividing the ligated by the total yields the percent of ribozymes that ligated for the sequence in question. For example, in Figure 7, the total RNA for the first construct (a) (13.7 Cq) compared to the ligated RNA from (A) (20.1 Cq) yields a ratio of 0.0176, or 1.76% ligation, after accounting for a primer efficiency of approximately 1.9 for both sets of primers. Similarly, the second construct (b) (17.9 Cq) versus (B) (23.7 Cq) yields a ratio of 0.0257, or 2.57% ligation.

One of the first applications of our qPCR assay was to test the effect of an extra 'G' base that was present at the beginning of the original Ligase sequence, but not the HDV sequence. We wished to measure the effect of removing this base on Ligase activity, as it would have been difficult to include in our sequence library. Using the qPCR assay, we were able to determine that its effect was minor, since ligation rates with and without the mutation were similar (Figure 5b). Thus, we were able to eliminate this indel mutation entirely, and limit our library to strictly point-mutations. This reduced the number of library sequences we had to order, and made sequence alignment more straightforward.

The qPCR assay allowed for proof of concept tests before the initial sequencing run, as well as validation of the sequencing run by comparing activities from each method. The substrate specific primer was also used in the actual high-throughput sequencing run, with Illumina-specific adapters attached. A template switching oligonucleotide was used during reverse transcription to add adapters to the pool of total ribozymes, since one of the fourteen mutations blocked the non-specific primer. Thus, the percent of ligated ribozymes could be calculated in a similar manner after sequencing.

3.3 Phased Primers

Phased primers dramatically improved clustering for our low-diversity library. Since high-throughput sequencing distinguishes individual sequencing “clusters” based on their differences early in the sequencing process, and in our initial sequencing run, our sequences were all nearly identical at the start, the sequencer failed to distinguish between neighboring clusters. It is possible to spike in random sequences to help clustering, generally fragments from the PhiX genome, but this was not sufficient to solve the problem. Thus, we developed phased primers with variations in the nucleotide sequence and length, with the goal of creating easily distinguishable clusters for HTS to progress (Table 1). These primers are identical except for a unique four base code that is repeated a varying number of times, and created enough differences at the start of the sequences for appropriate clusters to form. They also offset the rest of the ribozyme from its neighbors, such that all calls after the primer was read were not the same either. In this way, we were able to use high-throughput sequencing, a tool designed mainly for genomic sequencing, to sequence our ribozyme landscape.

3.4 All Sequences Detected with Tight Correlation Between Replicates

High-throughput sequencing provided over 100 million reads from our library, from which we were able to find every sequence in our library in abundance. As expected from previous studies on the ligase ribozyme, the reacted library contained large discrepancies between the counts for each sequence. Most sequences contained between 20 and 200 reads, but the most abundant had over 20,000. Without proper countermeasures, we could not assume that all these differences are due strictly to variation in ligation rate. Our rNTP mix was an even 25% for each base, which could perhaps favor the synthesis of sequences with a more even distribution of bases throughout (Krieg and Melton 1987). In addition, the buffer and temperature conditions could have favored synthesis of some sequences over others during transcription. To account for these differences, however minor, we compared our reacted library to a control library. By using the unreacted library as a baseline for differences in library synthesis, we could determine the real ligation rates for each sequence. We could also compare the data across the three replicates of each run, and across two sequencing runs. The ligase data was very consistent, with any two replicate runs having a coefficient of determination of 0.99 (Figure 8). Distinguishing between ribozymes with similar activities depends on the error in our measurements, so the smaller the variance in the fitness measurement, the better we can distinguish between RNA molecules with similar fitness. This tight correlation also allows detection of lower abundance molecules, which will allow us to expand our library size in the future.

After sequencing, we plotted the fitness values of the ligase sequences using Gephi (Figure 9). The right endpoint of the space is the ligase reference sequence, while the left endpoint of the space is the HDV reference sequence. The x and y axes represent sequence space, while the z axis represents fitness on a log scale. The high-activity ligase sequences are highly skewed toward the ligase reference side of the space. Although less than 1% of the total sequences have fitness values above the Ligase reference, many of those that do are substantially more fit than the reference. Since this space has not been explored before, this is potentially the first time these high-activity Ligase sequences have been discovered. The results from the ligase sequencing were later combined with the results from the HDV sequencing, resulting in a more comprehensive understanding of the region of innovation, discussed fully in the appendix below.

4. Conclusion

This thesis research demonstrates a new high-throughput sequencing based assay for the class III ligase ribozyme, in addition to the accessory assays leading up to sequencing. Before the qPCR assay, detection of low activity ribozymes such as the ligase ribozyme were difficult or impossible, as the resolution of a gel was often insufficient. The assay enabled the testing and comparison of many individual sequence variants, such that suitable endpoints for our library could be found. With endpoints, the library could be ordered, including the necessary extremities. The transition from qPCR preparation to sequencing preparation was reasonably straightforward. The sequencing assay was highly reproducible, and enabled detection of a wide range of activities for over 16,000 unique sequences simultaneously. After sequencing, the final analysis of the data in conjunction with the HDV data provided many insights into the boundaries of innovation. Even focusing on the ligase data alone, many variants were found with activities many times higher than the reference.

Although this research laid a solid foundation, there is still room for further study. Other regions of sequence space could be explored by allowing the mutations to vary among all four bases, or by choosing different mutations. Other types of mutations, such as insertions and deletions, could be included. One could study different cleavase and ligase molecules, or different catalytic reactions altogether. All of these alternatives require more room in the next-generation sequencer. In order to squeeze as many genotypes in as possible, the coverage of each genotype could be reduced by about ten-fold without negatively impacting the resolution. With time, next-generation sequencing will improve as well.

The work involving the ligase ribozyme was half of the necessary data to study innovation. The other half involves the HDV ribozyme, and was conducted by Devin Bendixsen. The Ligase activities reported in this thesis were combined with HDV self-cleaving activities. The results and discussion of this complete data set were prepared as a separate manuscript, which has been included in this thesis in the appendix.

5. References

Dantas G, Sommer MOA, Oluwasegun RD, Church GM. 2008. Bacteria Subsisting on Antibiotics. *Science* **320**: 100–103.

Bendixsen DP, Collet J, Østman B, Hayden EJ. 2018. Genotype network intersections promote evolutionary innovation. (included below)

Fletcher GL, Hew CL, Davies PL. 2001. Antifreeze Proteins of Teleost Fishes. *Annu Rev Physiol* **63**: 359–390.

Jiménez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen IA. 2013. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *PNAS* **110**: 14984–14989.

Krieg PA, Melton DA. 1987. [25] In vitro RNA synthesis with SP6 RNA polymerase. In *Methods in Enzymology*, Vol. 155 of *Recombinant DNA Part F*, pp. 397–415, Academic Press http://www.sciencedirect.com/science/article/pii/0076687987550273 (Accessed May 11, 2018).

Lee CH, Han SR, Lee S-W. 2018. Therapeutic applications of group I intron-based trans-splicing ribozymes. *Wiley Interdisciplinary Reviews: RNA* **9**: e1466.

Mayr E. 1960. The emergence of evolutionary novelties. *Evolution after Darwin* **1**: 349–380.

Moczek AP. 2008. On the origins of novelty in development and evolution. *Bioessays* **30**: 432–447.

Moczek AP, Sultan S, Foster S, Ledón-Rettig C, Dworkin I, Nijhout HF, Abouheif E, Pfennig DW. 2011. The role of developmental plasticity in evolutionary innovation. *Proceedings of the Royal Society of London B: Biological Sciences* **278**: 2705–2713.

Monteiro A. 2015. Origin, Development, and Evolution of Butterfly Eyespots. *Annu Rev Entomol* **60**: 253–271.

Müller GB, Nitecki MH. 1990. Developmental mechanisms at the origin of morphological novelty: a side-effect hypothesis. *Evolutionary innovations* 99–130.

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

Rehmann L, Daugulis AJ. 2008. Enhancement of PCB degradation by Burkholderia xenovorans LB400 in biphasic systems by manipulating culture conditions. *Biotechnol Bioeng* **99**: 521–528.

Ren A, Micura R, Patel DJ. 2017. Structure-based mechanistic insights into catalysis by small self-cleaving ribozymes. *Current Opinion in Chemical Biology* **41**: 71–83.

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

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. *Nature*. https://www.nature.com/articles/nature07891 (Accessed August 1, 2018).

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

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

Tables and Figures:

Table 1: Important Sequences

RT primer	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAACCGGACCGAAGCCCG
Substrate oligo	AAGCATCTAAGCATCTCAAGCrArArArCrCrArGrUrC
Selective, phased sequencing primer 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCATGCATGCATGCATGC AAGCATCTAAGCATCTCAAGCAAACCAG
Selective, phased sequencing primer 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGCATGCATGCATGC AAGCATCTAAGCATCTCAAGCAAACCAG
Selective, phased sequencing primer 3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATGCATGCATGC AAGCATCTAAGCATCTCAAGCAAACCAG
Selective, phased sequencing primer 4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CATGCATGC AAGCATCTAAGCATCTCAAGCAAACCAG
Non-selective qPCR primer	GAMTCCCATTAGRCTGG
Phased Template Switching Oligo (TSO) 1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCATGCATGCATGCATGC rGrGrG
Phased TSO 2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGCATGCATGCATGC rGrGrG
Phased TSO 3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATGCATGCATGC rGrGrG
Phased TSO 4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CATGCATGC rGrGrG

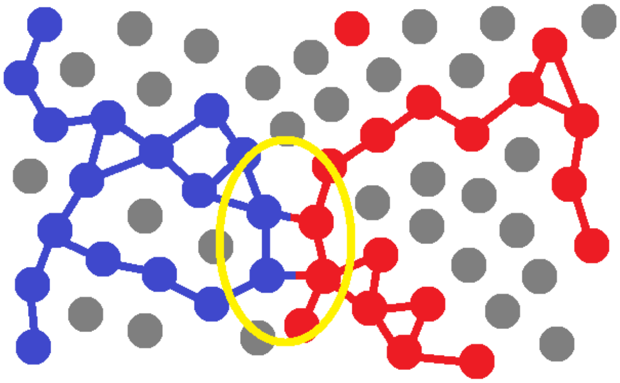

Figure 1. Innovation and Genotype Space. Dots represent individual genotypes, while colors represent phenotypes. The circled portion indicates a region of innovation in which a transition occurs from one phenotype to another.

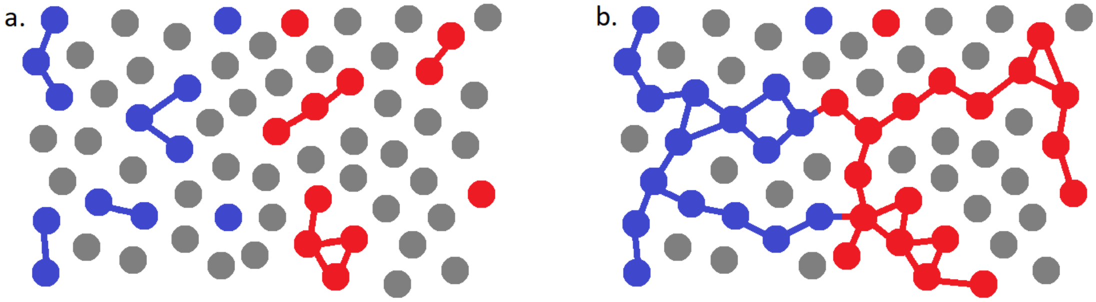

Figure 2. Genotype networks. Dots represent individual genotypes, while colors represent phenotypes. a. Under strict conditions, only the fittest genotypes can survive, leaving little room for exploration. b. When conditions relax, more genotypes are accessible, opening trajectories from one phenotype to another.

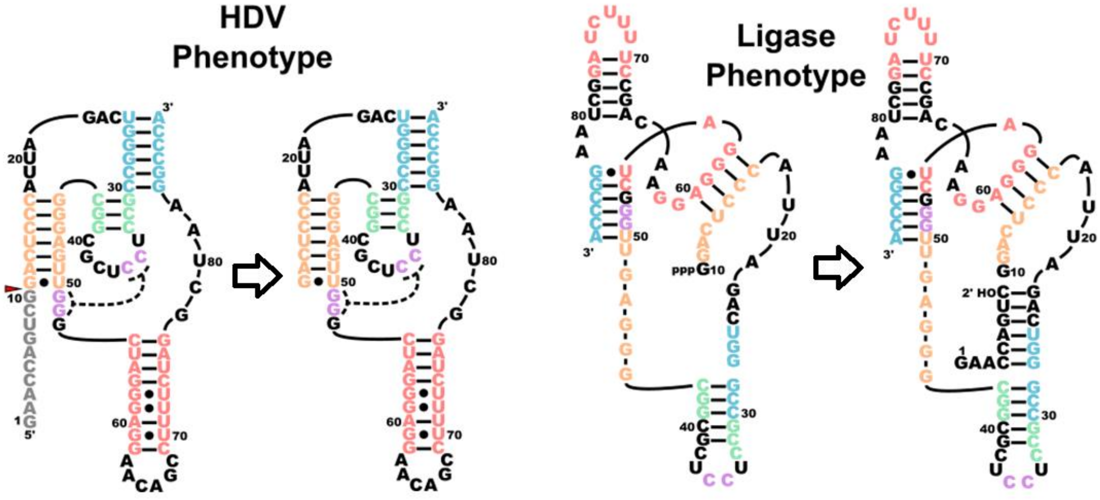

Figure 3. HDV and Ligase Ribozymes. The HDV ribozyme cleaves a portion of RNA from its 5’ end, whereas the Ligase ribozyme ligates a substrate to its 5’ end. The bases are color coded according to the base pairing of the HDV ribozyme to demonstrate that no two regions are paired the same.

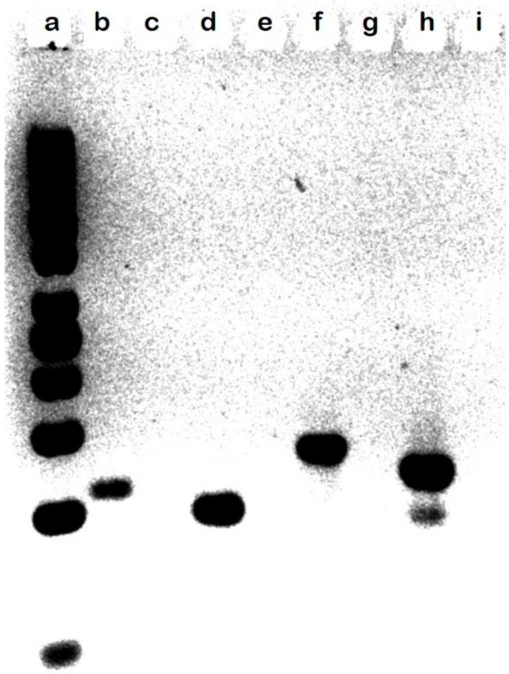

Figure 4. **RT-PCR amplification of ligated ribozymes. This gel demonstrated the viability of an RT-qPCR strategy for visualizing the low-activity ligase ribozyme's activity. Lane (a) contains the ladder, and (b)-(e) and (f)-(i) contain two variants of the ligase ribozyme. (b) and (f) are the ligated bands, while (d) and (h) contain total RNA (unligated and ligated RNA). (c) and (e) are the no-enzyme controls, while (g) and (i) are the no-template controls for each group.**

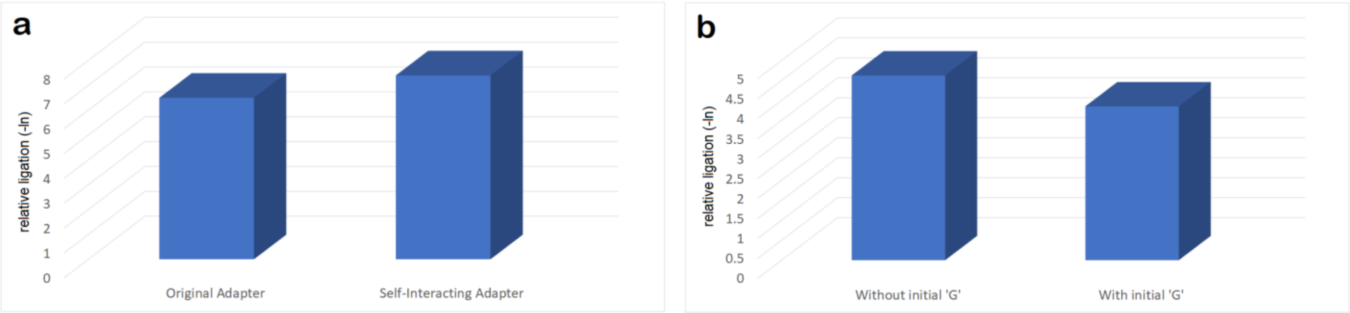

Figure 5. **qPCR Testing of Adapters and a Base Difference. Relative ligation after background subtraction of our original adapter vs the non-interfering self-interacting adapter (a), and of the ligase sequence with and without the initial 'G' base (b). Both graphs display similar ligation values, well within an order of magnitude.**

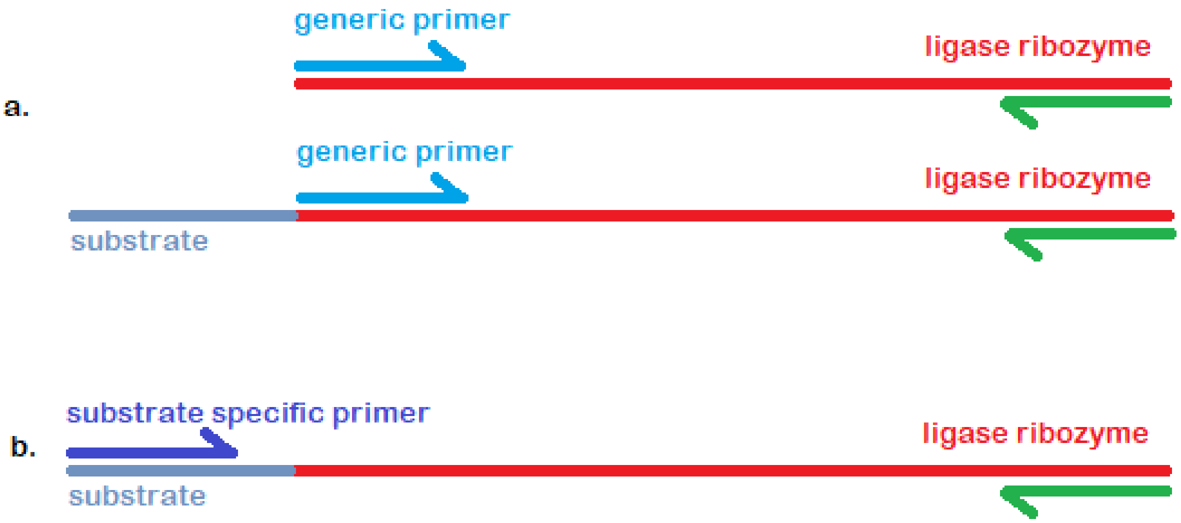

Figure 6. Detection of Ligase Activity. In order to detect ligase activity, two reverse primers were used. In (a), a generic primer amplifies all ribozymes. In (b), a substrate specific primer only amplifies ligated ribozymes.

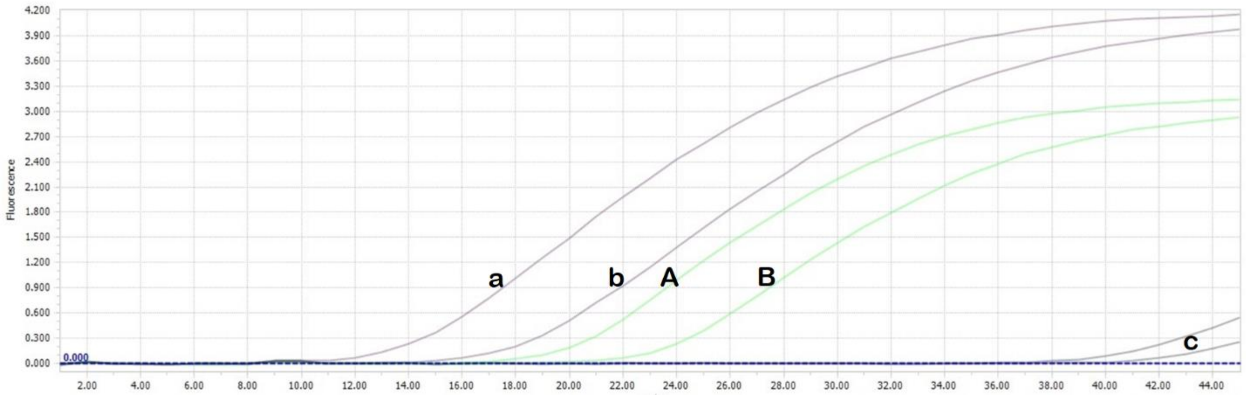

Figure 7. Quantification of Ligase Activity by qPCR. The plot shows DNA fluorescence (y-axis) as a function of PCR cycle number (x-axis) for two different ligase ribozyme constructs. This data was used to quantify the ligase activity of each construct. The total RNA concentration of the first construct is given by the Cq value (a), after taking into account the primer efficiency. The concentration of the ligated product of this construct is given by the Cq value of (A). The same is true for (b) and (B). The no-template negative control is (c).

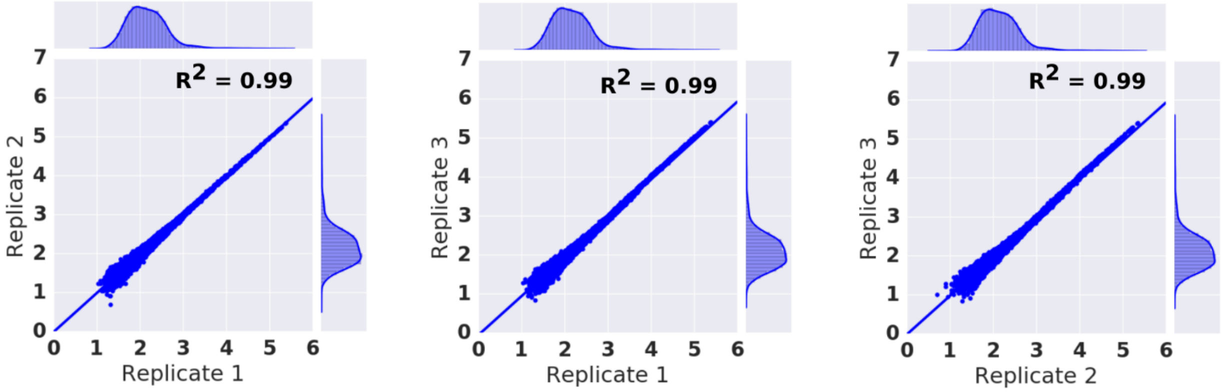

Figure 8. **Consistency of Ligase ribozyme frequencies in sequencing data from replicate ligation reactions. Each data point represents a unique sequence in the library. The axes are the $\log_{10}$ transformed number of reads for the indicated replicate, with a distribution of the density of those reads on the opposite side.**

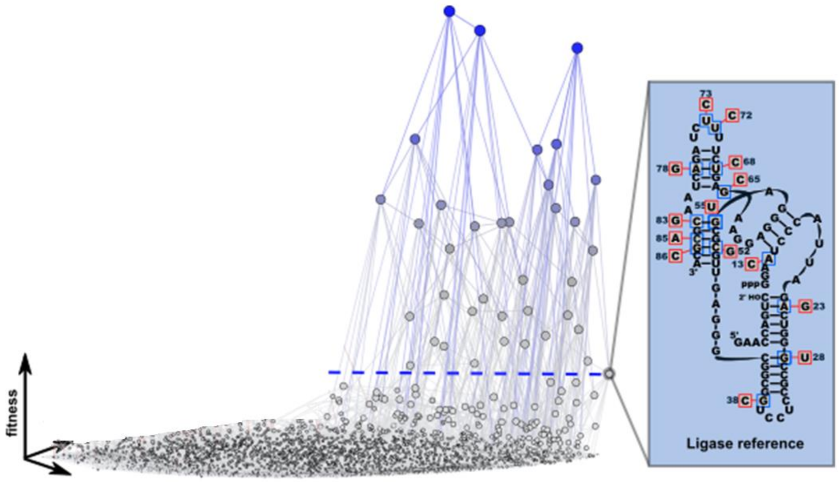

Figure 9. **Ligase Fitness. The x and y axes represent sequence space, while the z axis represents fitness. The Ligase reference ribozyme and its fitness is included.**

Application of Assay:

The work developing assays and analysis for the ligase ribozyme is only a portion of the project, which I focused on. This work was combined with Devin Bendixsen's data for the HDV cleavase ribozyme. Software for computer simulations in the analysis phase were provided in part by a collaborator, Dr. Bjorn Ostman. The entire project was overseen by Dr. Eric Hayden. The project can be found in the appendix below.

APPENDIX

Appendix: Genotype network intersections promote evolutionary innovation

Authors: Devin P. Bendixsen[1], James Collet[2], Bjørn Østman[3,4], Eric J. Hayden[1,2*]

Affiliations:

[1]Biomolecular Sciences Graduate Programs, Boise State University, Boise, ID, USA.

[2]Department of Biological Science, Boise State University, Boise, ID, USA.

[3]Department of Ecology and Evolutionary Biology, UCLA, Los Angeles, CA, USA.

[4]Department of Biomathematics, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA.

*Correspondence to: erichayden@boisestate.edu

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.

One Sentence Summary: Experimental fitness landscapes between two genotype networks provide the direct observation of the evolution of a novel function.

Main Text:

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. 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. 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.

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. 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. 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. 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 ($\log_{10}(fitness) > 0$). 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 Supplementary

Materials). 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. 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. 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. 3B). To evaluate the average contribution of these periods of stasis, we measured the average fitness of the evolving population over time (Fig. 4A and Fig. S1) and did so for 100 replicate simulations from each of the different starting genotypes (Fig. 4B). We find that different genotypes on the HDV network result in different average rates of adaptation to the *new* Ligase function (Fig. 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. 4D and Fig. S2-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 dual0function sequence at this intersection encourages the search for more genotype network intersections and motivates future research on the forecasting of evolutionary innovations.

References

1. C. R. Darwin, On the Origin of Species by means of natural selection, or the preservation of favoured races in the struggle for life. (John Murray, London, ed. 1, 1859).

2. M. Pigliucci, What, if Anything, Is an Evolutionary Novelty? Philos. Sci. **75**, 887–898 (2008).

3. C. A. Tracewell, F. H. Arnold, Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr. Opin. Chem. Biol. **13**, 3–9 (2009).

4. S. Wright, Surfaces of selective value revisited. Am. Nat. **131**, 115–123 (1988).

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

6. H. Innan, F. Kondrashov, The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. **11**, 97 (2010).

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

8. U. Bergthorsson, D. I. Andersson, J. R. Roth, Ohno’s dilemma: evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. **104**, 17004–17009 (2007).

9. J. Zhang, Evolution by gene duplication: an update. Trends Ecol. Evol. **18**, 292–298 (2003).

10. M. Pigliucci, J. Kaplan, Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology (University of Chicago Press, Chicago, UNITED STATES, 2014; http://ebookcentral.proquest.com/lib/boisestate/detail.action?docID=485956).

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

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

13. A. M. Dean, J. W. Thornton, Mechanistic approaches to the study of evolution: the functional synthesis. Nat. Rev. Genet. **8**, 675–688 (2007).

14. Z. D. Blount, J. E. Barrick, C. J. Davidson, R. E. Lenski, Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature. **489**, 513–518 (2012).

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

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

17. A. Wagner, The molecular origins of evolutionary innovations. Trends Genet. **27**, 397–410 (2011).

18. N. Takeuchi, P. H. Poorthuis, P. Hogeweg, Phenotypic error threshold; additivity and epistasis in RNA evolution. BMC Evol. Biol. **5**, 9 (2005).

19. A. Wagner, Genotype networks shed light on evolutionary constraints. Trends Ecol. Evol. **26**, 577–584 (2011).

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

21. D. P. Bartel, J. W. Szostak, Isolation of new ribozymes from a large pool of random sequences [see comment]. Science. **261**, 1411–1418 (1993).

22. E. A. Schultes, D. P. Bartel, One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science. **289**, 448–452 (2000).

23. A. Khanal, S. Y. McLoughlin, J. P. Kershner, S. D. Copley, 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 (2015).

24. P. J. O'Brien, D. Herschlag, Catalytic promiscuity and the evolution of new enzymatic activities. Chem. Biol. **6**, R91–R105 (1999).

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

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

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

28. J. Otwinowski, J. B. Plotkin, Inferring fitness landscapes by regression produces biased estimates of epistasis. Proc. Natl. Acad. Sci. **111**, E2301–E2309 (2014).

29. E. A. Schultes, D. P. Bartel, One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science. **289**, 448–452 (2000).

30. D. P. Bartel, J. W. Szostak, Isolation of new ribozymes from a large pool of random sequences. Science. **261**, 1411–1418 (1993).

31. D. M. Long, O. C. Uhlenbeck, Kinetic characterization of intramolecular and intermolecular hammerhead RNAs with stem II deletions. Proc. Natl. Acad. Sci. **91**, 6977–6981 (1994).

32. R. Rohatgi, D. P. Bartel, J. W. Szostak, Kinetic and mechanistic analysis of nonenzymatic, template-directed oligoribonucleotide ligation. J. Am. Chem. Soc. **118**, 3332–3339 (1996).

33. Y. Li, R. R. Breaker, Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2 ‘-hydroxyl group. J. Am. Chem. Soc. **121**, 5364–5372 (1999).

34. B. Ostman, A. Hintze, C. Adami, Impact of epistasis and pleiotropy on evolutionary adaptation. Proc. Biol. Sci. **279**, 247–256 (2012).

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. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All data, code and materials are available upon request.

Supplementary Materials:

Materials and Methods

Figures S1-S6

Tables S1-S2

Movies S1-S5

Figure 1

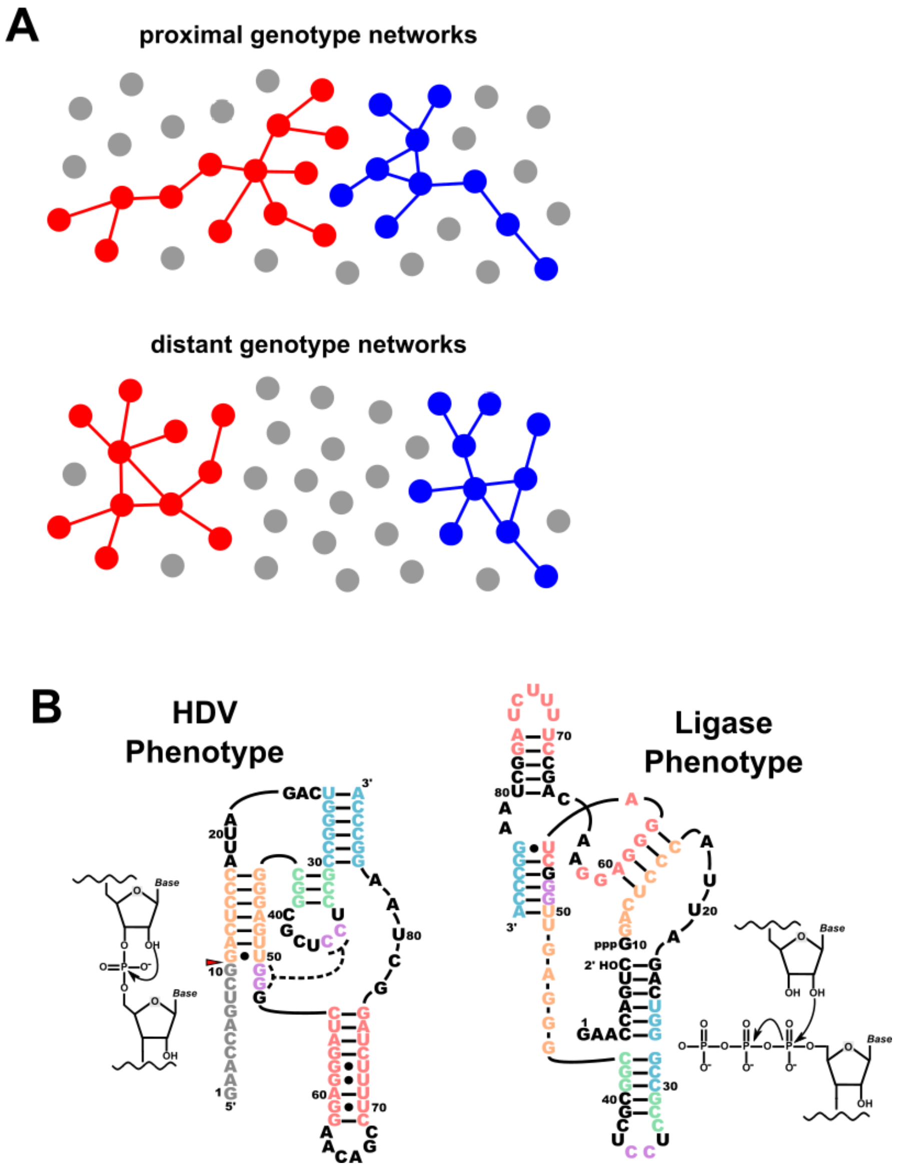

Fig. 1 Evolutionary innovation from a network perspective. (A) Each node represents a genotype. Genotypes with the same phenotype (genotype networks) have the same color and are interconnected by mutational steps (edges). Gray nodes are non-functional. Proximal genotype networks have neighboring genotypes 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). The two structures have the same nucleotide sequence and nucleotides are colored based on the secondary structure of the HDV phenotype. These sequence changes following catalysis can be detected in nucleotide sequence data.

Figure 2

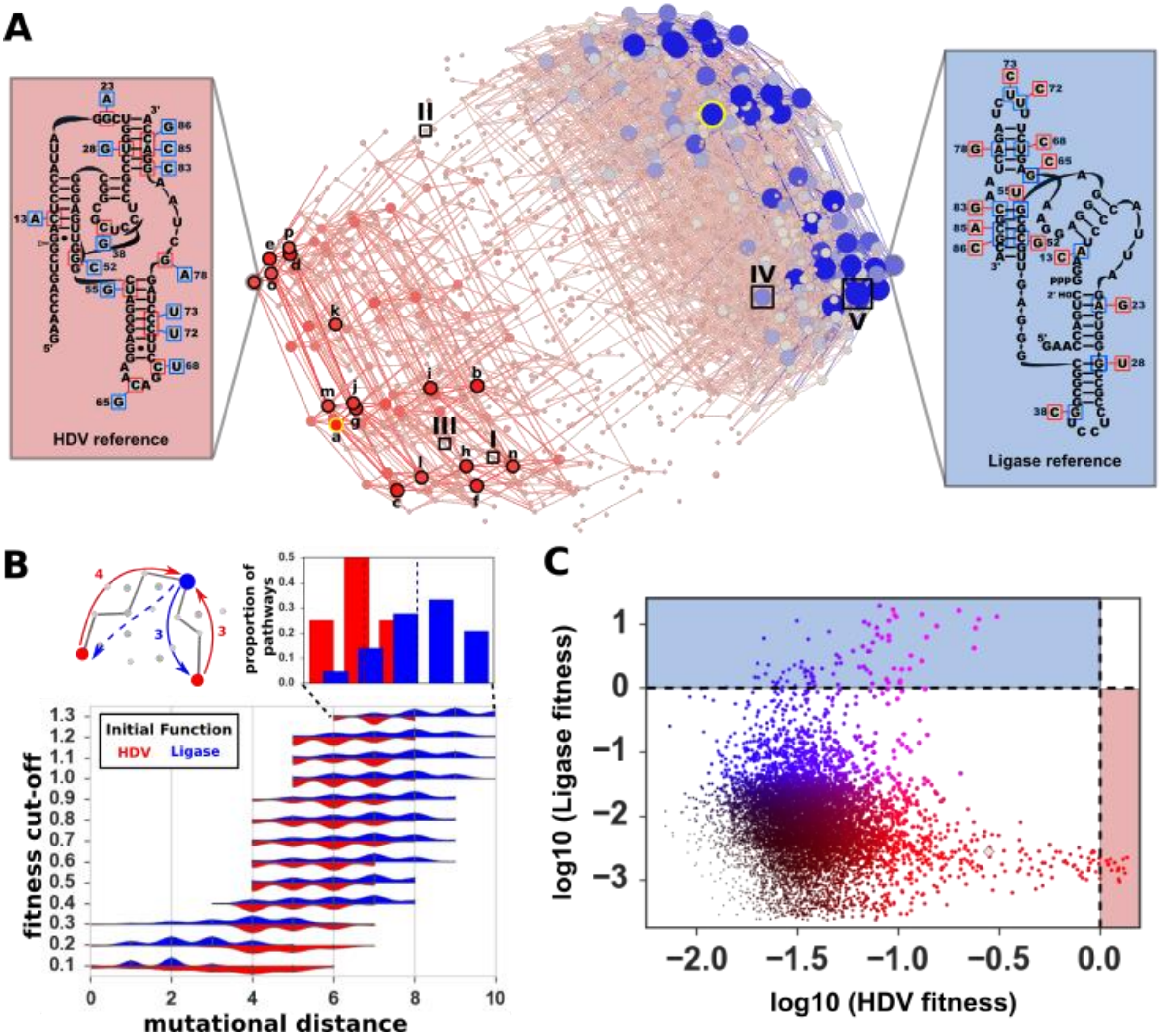

Fig. 2 The experimental fitness landscape at the intersection of two genotype networks. (A) The overlay of the HDV and Ligase genotype networks. Nodes represent

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).

Figure 3

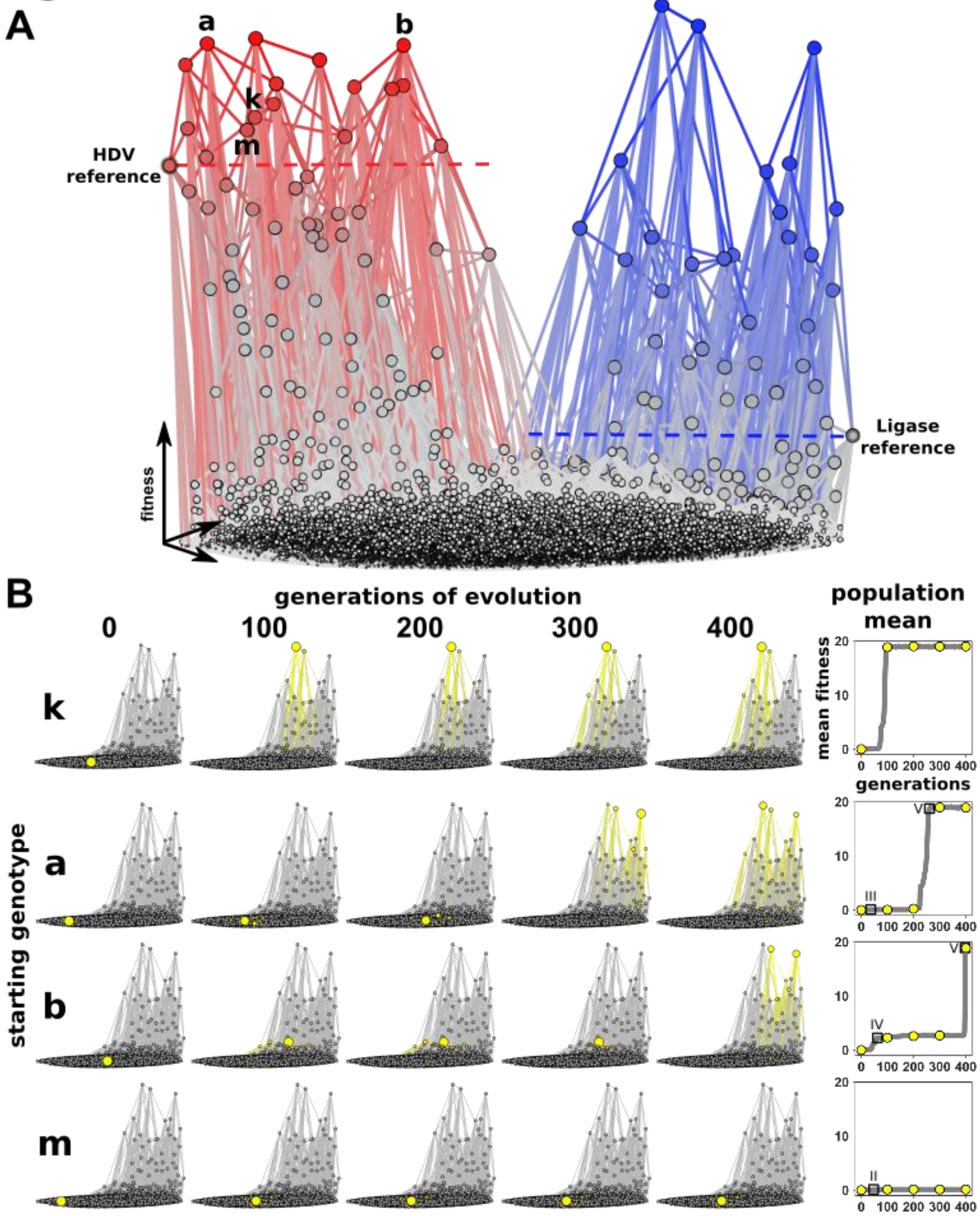

Fig. 3 Computational simulation of evolutionary innovation reveal periods of stasis. (A) Three-dimensional fitness landscape for both genotype networks. The height of each node indicates the relative fitness for the HDV phenotype (red) and the Ligase phenotype (blue). Fitness are normalized so that both 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.

Figure 4

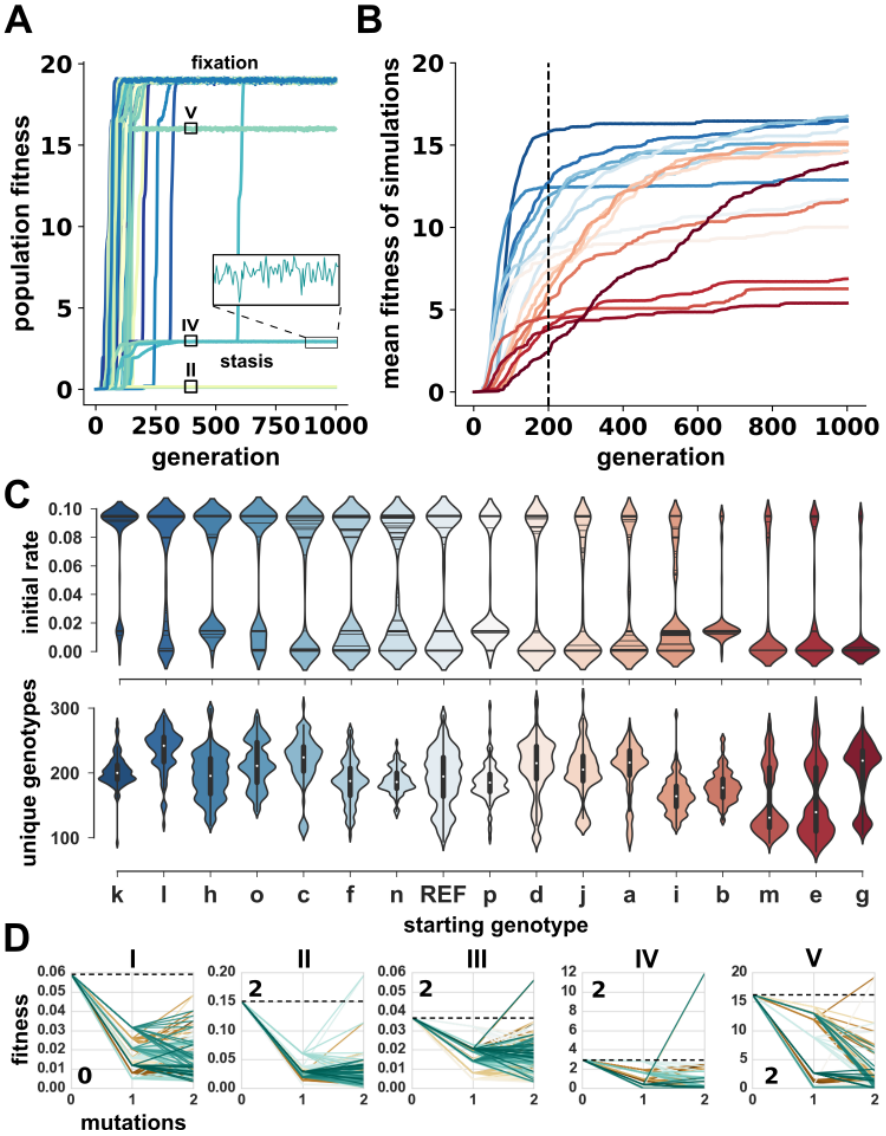

Fig. 4 Starting genotypes result in different rates of adaptation. (**A**) Rates of Ligase adaptation from a single HDV genotype. Each trace shows the average fitness as a function of generation time for a separate simulation of 1000 individuals each. Inset shows minor fluctuations during periods of stasis. Traces are colored from fastest (blue) to slowest (red) (**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 Supplementary Materials). 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. 2.