674–683 *Nucleic Acids Research, 2021, Vol. 49, No. 2* *Published online 24 December 2020*
doi: 10.1093/nar/gkaa1238

In vitro selections with RNAs of variable length converge on a robust catalytic core

Milena Popović[1,2,3], Alexander Q. Ellingson[3], Theresa P. Chu[3], Chenyu Wei[1,2,4], Andrew Pohorille[1,2,4] and Mark A. Ditzler[1,2,*]

[1]Center for the Emergence of Life, NASA Ames Research Center, Moffett Field, CA 94035, USA, [2]Exobiology Branch, Space Science and Astrobiology Division, NASA Ames Research Center, Moffett Field, CA 94035, USA, [3]Blue Marble Space Institute of Science, Seattle, WA 98145, USA and [4]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94158, USA

Received October 11, 2020; Revised December 05, 2020; Editorial Decision December 07, 2020; Accepted December 10, 2020

ABSTRACT

***In vitro* selection is a powerful tool that can be used to understand basic principles of molecular evolution. We used *in vitro* selection to understand how changes in length and the accumulation of point mutations enable the evolution of functional RNAs. Using RNA populations of various lengths, we performed a series of *in vitro* experiments to select for ribozymes with RNA ligase activity. We identified a core ribozyme structure that was robust to changes in RNA length, high levels of mutagenesis, and increased selection pressure. Elaboration on this core structure resulted in improved activity which we show is consistent with a larger trend among functional RNAs in which increasing motif size can lead to an exponential improvement in fitness. We conclude that elaboration on conserved core structures is a preferred mechanism in RNA evolution. This conclusion, drawn from selections of RNAs from random sequences, is consistent with proposed evolutionary histories of specific biological RNAs. More generally, our results indicate that modern RNA structures can be used to infer ancestral structures. Our observations also suggest a mechanism by which structural outcomes of early RNA evolution would be largely reproducible even though RNA fitness landscapes consist of disconnected clusters of functional sequences.**

INTRODUCTION

The evolution of functional RNAs is marked by changes in sequence, size, and complexity. Despite these changes, orthologous RNAs harbor highly conserved structural cores, which can remain unchanged for billions of years. The ribosome offers perhaps the most striking example of RNA structure preservation, with a structural core composed of multiple strands and thousands of nucleotides, which is conserved in all known cellular life (1–3). Other ancient RNAs such as RNase P and the signal recognition particle also have structural cores that are highly conserved throughout the tree of life (3–6). The processes by which core RNA structures are selected and subsequently preserved or lost, and the roles of chance and necessity in guiding those processes are of great interest and have been the subject of sustained theoretical and experimental inquiry.

Imperfect fidelity in templated RNA synthesis by polymerase enzymes (7) or ribozymes (8) can introduce molecular diversity into RNA populations in the form of point mutations. The evolutionary consequences of these point mutations have been the focus of many studies. Both theoretical (9–16) and experimental models (17,18) have been used to argue that populations of sufficiently long RNAs can traverse arbitrarily long distances on a fitness landscape—a genotype-fitness map—via a series of neutral or near-neutral point mutations. If these arguments are correct, then evolutionary outcomes should be largely independent of the set of initial sequences, thereby limiting the role of chance (13–16). In contrast to the studies cited above, multiple *in vitro* selection experiments have identified significant barriers to the broad exploration of fitness landscapes and to the evolution of new structures through accumulation of point mutations (19–21), and it has been argued that such barriers contribute to an increased role of chance in evolutionary outcomes (19).

In addition to point mutations, spontaneous recombination (22,23) and aberrant replication (7,24) can introduce molecular diversity by inserting or deleting stretches of sequence in functional RNAs. Insertions and deletions frequently disrupt preexisting structures and functions, but in specific cases they are accommodated by preexisting structures and can even enhance function (25–27). Competing models for the early evolution of RNA have been used to make inferences about primordial life by invoking either

*To whom correspondence should be addressed. Tel: +1 650 604 1058; Email: mark.a.ditzler@nasa.gov

structural rearrangement (28,29) or structural preservation (1–3) in response to sequence insertion events.

A better understanding of the evolutionary consequences of both changes in length and point mutations is needed to resolve conflicting models of RNA evolution. To this end, we determined how these two phenomena impact RNA evolution by analysing a series of *in vitro* selection experiments in which we evolved RNAs that catalyse RNA ligation. We selected catalytic RNAs from high diversity RNA populations of various lengths, and each population was designed to be independent of biological sequences and structures. In this way, we were able to learn how changes in polymer length impact RNA evolution independently of the specific historical events that shaped modern biological RNAs. By comparing our results with evolutionary models built from the analysis of biological RNAs, we conclude that elaboration on conserved core structures is a preferred mechanism in RNA evolution and, consequently, that the ancestral forms of RNAs are largely retained as embedded structures within modern RNAs.

MATERIALS AND METHODS

Oligonucleotides

DNA oligonucleotides were synthesized by Integrated DNA Technologies (IDT). For all selections double-stranded DNA starting libraries were generated by heat annealing two partially complementary DNA oligonucleotides, followed by extension by DNA polymerase I Klenow fragment (Thermo Scientific). All transcriptions were carried out in transcription buffer (50 mM Tris–HCl pH 7.5, 10 mM NaCl, 30 mM $MgCl_2$, 2 mM spermidine, 40 mM DTT) with 5 mM of each NTP and T7 RNA polymerase (Promega) for 12–18 h at 37°C. RNAs were purified using denaturing polyacrylamide gel electrophoresis (PAGE), recovered from gels through electro-elution, ethanol precipitated, and re-suspended in deionized water. RNA substrates for selections and ligation assays were synthesized by Dharmacon and IDT.

20N selections

The 65 base pair DNA starting library included 20 fully degenerate positions (20N), flanked by a promoter and 3′ constant region. The DNA library contains the T7 promoter sequence GCC *ATG TAA TAC GAC TCA CTA TAG G*. The RNA version of the promoter sequence GCC *AUG UAA UAC GAC UCA CUA UA* served as the substrate for the 20N selections. The 3′ constant sequence is AGA TCG GAA GAG CGT CGT GT. The DNA library of $\sim 6 \times 10^{14}$ molecules, with a maximum diversity of 4^{20} ($\sim 10^{12}$) unique sequences, was transcribed to generate a population of $\sim 10^{17}$ RNA molecules. The preparation of the 20N library and associated 20N selection experiments were initiated approximately three months after completion of the 80N selections, which are described in the next sub-section. Separating the 20N and 80N selections in time combined with careful sample handling (e.g. exclusive use of barrier pipette tips, frequent cleaning/decontamination of equipment), lowered the probability of any inadvertent contamination of the 20N selections.

For the first round in the 20N selections, 1.2×10^{16} RNA molecules ($\sim$20 nmol) were suspended in 1x folding buffer (100 mM NaCl, 100 mM KCl, 50 mM 3-(*N*-morpholino)propanesulfonic acid (MOPS) pH 7.5) to a final concentration of 1 μM RNA. RNA populations were refolded by heating to 92°C for 2 min, then cooled to 25°C for 30 min in folding buffer. After cooling, magnesium chloride was added to a final concentration of 5 mM $MgCl_2$. Ligation was initiated by the addition of the substrate oligomer and allowed to react for 96 h. Ligation reactions were stopped by the addition of an equimolar amount of ethylenediaminetetraacetic acid (EDTA; 5 mM final concentration), and then an equal volume of 8 M urea pH 5. The samples were then ethanol precipitated and resuspended in deionized water. The active RNA molecules were isolated by denaturing PAGE, only the RNA molecules that ligated to the substrate were cut out of the gel. The catalytically active sequences were recovered from the gel through electro-elution (Biorad), precipitated and resuspended in deionized water. The resuspended sample was then reverse transcribed using ImProm-II reverse transcriptase (Promega) and amplified via PCR using Taq DNA polymerase (Thermo Scientific). Finally, the PCR products were transcribed *in vitro* to generate the RNA population used in the next round of evolution. This process was repeated for two more rounds, either with the same 1x folding buffer with 5 mM $MgCl_2$ or with 2× folding buffer and 5 mM $MgCl_2$ (Supplementary Figure S1).

80N selections

The 125 base pair DNA starting library included 80 fully degenerate positions (80N). The promoter and 3′ constant sequence used in the 80N selections were the same as those used in the 20N selection, in this way the only difference between the RNA populations in 80N and 20N selections was the length of the randomized region. For the last three rounds of selection, the substrate was the same as the 20N selections, and in the first five rounds the substrate was the same with the exception of a biotin modification on the 5′ end. The DNA starting library of $\sim 6 \times 10^{14}$ molecules was transcribed to generate a population of $\sim 10^{17}$ RNA molecules, from which aliquots of 1.2×10^{16} molecules ($\sim$20 nmol) were used in the first round of the two series of 80N selections. The two 80N starting populations, though not identical, are therefore expected to have been highly similar and most of the sequences that were present in one are predicted to have been present in the other. The selection experiments with the 80N populations were completed approximately three months prior to the 20N selections, which are described in the previous sub-section. Completing the 80N selection first, ensured that it was impossible for any material from the 20N selections to influence the outcome of the 80N selections. This is important because the 80N selections would have been very sensitive to even vanishingly small levels of contamination from 20N selections especially in the early rounds when copy number was low and even a single copy of a sequence generated from the mis-priming of a partially extended DNA primer from the 20N selections could have substantially influenced the outcome of the selection.

For each round of selection the 80N RNA populations were suspended in 1× folding buffer and refolded by heating to 92°C for 2 min, then cooled to 25°C for 30 min. After cooling, magnesium chloride was added to a final concentration of 5 mM $MgCl_2$, and for some selection steps, either 20% polyethylene glycol or 20% dextran was added (Supplementary Figure S1). Ligation was initiated by the addition of the substrate oligomer and allowed to react for 24 h. Ligation reactions were stopped by the addition of an equimolar amount of EDTA, then an equal volume of 8 M urea pH 5, ethanol precipitated, and resuspended in deionized water. In the first five rounds, ribozymes ligated to the biotinylated substrate were captured on streptavidin magnetic beads (Invitrogen) by using the following protocol. 1.2 ml of magnetic beads were washed three times with binding buffer (10 mM 2-(*N*-morpholino)ethanesulfonic acid (MES) pH 5, 0.01 mM EDTA, 200 mM NaCl) and 0.1% Tween-20, in preparation for binding of the biotinylated target. Material precipitated from ligation reactions was incubated with magnetic beads for 60 min. Unbound RNA was removed by three washes with 5 M urea-containing binding buffer, followed by three washes in high salt buffer (10× binding buffer), and finally three washes in binding buffer. Bound catalytic RNAs were eluted by heating in deionized water at 70°C for 1 min. The eluted material was ethanol precipitated, treated with DNAase I (Thermo Scientific) to remove any DNA, reverse transcribed using ImProm-II reverse transcriptase (Promega), and amplified *via* polymerase chain reaction (PCR) using Taq DNA polymerase (Thermo Scientific).

After five rounds of selection, we used an error-prone (EP)-PCR protocol (30) to introduce random mutation into the variable region of the round 5 DNA templates. In brief, $MnCl_2$ was added to the PCR mixture to promote mutations and dNTP concentrations were adjusted to generate an unbiased distribution of mutations. Through repeated dilution and amplification steps the DNA was taken through ∼60 doublings. Following (EP)-PCR, we transcribed 1.2×10^{14} DNA molecules (0.2 nmol). $\sim 6 \times 10^{14}$ RNA molecules were used as starting populations for the last three rounds of selection. For the last three rounds, the active sequences within the RNA population that ligated to the substrate were separated from inactive RNA using denaturing PAGE.

Reselection

Double-stranded DNA libraries were generated for the populations pIL, pILint and pILter. All three DNA libraries contained the same promoter used in the 20N and 80N selections. For all three libraries the 3′ constant region was changed to the sequence AAGCAATATTGGAT GGATAAGATG. The reselection experiments were carried out after the 20N and 80N selections, the altered 3′ constant sequence helped to limit the potential for inadvertent contamination by material from the prior selections. The pIL library contained 40 partially randomized positions, based on 18% mutagenesis per position of the parent sequence, AGCAAACTTCGGACAATCGAAAGGA AACAAGCTCATGGTG. The pILint library had 20 fully randomized positions inserted between positions 16 and 17 of the full RNA sequence (underlined positions in parent sequence above), and the parent sequence was mutagenized at 6% per position. The pILter library had 20 fully randomized positions inserted between the 3′ end of the parent sequence and the 3′ constant sequence), and the parent sequence was mutagenized at 6% per position. Each library was transcribed *in vitro*. Each series of selections was initiated by allowing 1.2×10^{16} molecules (∼20 nmol) to react with the substrate for 1 h. The reselections were carried out for five, four and five rounds for the pIL, pILint and pILter, respectively. In every round, the active RNA molecules were isolated by denaturing PAGE.

Sequencing and analysis

RNA libraries were prepared for high-throughput sequencing using PCR with primers designed to add index sequences that allow multiplexing of multiple populations on a single sequencing lane. The prepared DNA libraries were sequenced on an Illumina 4000 HiSeq instrument. Only sequence that were the same length as their respective starting populations, matched perfectly in both directions of the two paired-end reads, and were present in more than three reads were used for our primary analysis.

The FASTAptamer toolkit (31) was used to count reads for all sequences, compare abundance of sequences between populations, and determine edit distances between sequences. Sequences selected from the 20N and pIL populations were assigned to separate networks by generating connection maps. Each sequence was considered as a vertex, and two sequences were connected by an edge if and only if they differed by a single point mutation. Within a single network, every pair of vertices was connected by a sequence of edges. This means that every sequence in a network could evolve to another sequence in the same network through a series of sequences, all belonging to the same network, that differ only by a single point mutation. For all selections, RNA secondary structures were predicted through a combination of both manual inspection of sequences and the use of RNA folding software. In the latter case, individual sequences were evaluated using RNAfold to generate secondary structure predictions and multiple-sequences alignments were evaluated using RNAalifold to generate secondary structure predictions based on conservation and covariation among related sequences (32). Sequences selected from the 80N, pILint and pILter populations were assigned to families on the basis of sequence similarities. For the 80N selections, families were defined as sequences that differ by a maximum of 15 edits, and for the pILint and pILter selections families were defined as sequences that differ by a maximum of six edits.

The motif sizes used in figure 4C were determined by using a previously published convention for determining motif size (33). Motif size was determined as the sum of all nucleotides that make up loops and bulges needed for activity, the intervening base-paired stems that separated those loops and bulges, and the flanking base-paired stems that define those loops and bulges (this includes nucleotides from the substrate that contribute to forming the active structure). For the flanking base-paired stems, only the first 5 bp are included in the total count for determining motif length. The loops and bulges needed for activity were determined using functional and structural assays. In this work,

we established that the IL motif is the functional part of the ribozyme and that the terminal loop is part of the TL-motif by showing that the deletion of the loop results in a clear drop in activity. We also tested the impact of deleting either 3nt or 9nt from the most abundant sequence in network 4 from the 20N selection, sequence 20N.4.1, to form the sequences 20N.4.1(–3) and 20N.4.1(–9), respectively. 20N.4.1(–9) is predicted to favor a pseudoknot secondary structure and 20N.4.1(–3) is predicted to favor a non-pseudoknot structure. The higher activity observed for 20N.4.1(–9) supports the prediction that the pseudo-knot structure is the functional structure and so the pseudo-knot secondary structure was used to determine the motif size.

Ligation assays

Substrate oligomers were 5′-end labeled with ^{32}P. After labeling, the substrate was passed serially through two size exclusion columns and then precipitated using ethanol precipitation. ^{32}P-labeled substrate oligomer was resuspended in ligation buffer. In 1× folding buffer, individual ribozymes were refolded by heating to 92°C for 2 min and then cooled to 25°C for 30 min. After cooling, magnesium chloride was added to a final concentration of 5 mM $MgCl_2$. Reactions were initiated by the addition of the substrate and terminated by the addition of 2 volumes of gel running buffer (94% formamide, 30 mM EDTA). Products were separated on denaturing PAGE and quantified using ImageQuant software to determine the amount of signal from each band. The extent of ligation was calculated based on the signal from the band corresponding to the ligated RNA and that of the unligated RNA. Ligation time courses were, in most cases, fit to a single exponential $y(t) = A(1 - e^{(-k*t)})$, and in one case, to a double exponential $y(t) = A_1(1 - e^{(-k1*t)}) + A_2(1 - e^{(-k2*t)})$ using MyCurveFit (MyAssays Ltd.). Because ligation reactions with sequences derived from the 20N selections remain linear even at long time points it was not possible to determine the amplitudes of those reactions. We therefore determined only initial rates for the 20N derived ribozymes using a linear fit.

Normalization of ligation rate constants

The normalization factors k_{deg} used to normalize ligation rate constants between earlier studies and those measured here were predicted using an equation that predicts k_{deg} using pH, temperature and ion concentrations as parameters

$$k_{deg} = k_{background} \times 10^{(0.983(pH-6))} \times 10^{(-0.24(3.16-[K]-[Na]))} \times 69.3 \times [Mg]^{0.8} \times 3.57 \times ([K]+[Na])^{-0.419} \times 10^{(0.07(T-23))}$$

the equation is a slight modification from that used by (34). $k_{background} = 1.30 \times 10^{-9}$ min^{-1}, pH is pH, [K] is potassium concentration, [Na] is sodium concentration, [Mg] is magnesium concentration and T is temperature in Celsius.

RESULTS

Parallel *in vitro* selection experiments with short or long RNA populations

For *in vitro* selection experiments with short RNAs, we generated an RNA library 42 nucleotides in length with 20 fully randomized positions (20N) (Figure 1A). We used this 20N library to perform selections for ligase activity (Supplementary Figure S1A). In the first round of the *in vitro* selection, the library was incubated in our 1× selection buffer (5 mM $MgCl_2$, 100 mM KCl, 100 mM NaCl and 50 mM MOPS pH 7.5) with an RNA substrate 23 nucleotides in length (Figure 1A). After 96 h, the sequences that had ligated the substrate to their 5′ ends were isolated. The isolated sequences were reverse transcribed into DNA and then amplified through PCR. Finally, the PCR products were converted back into RNA through *in vitro* transcription and this new population of RNAs was subjected to additional rounds of selection. For subsequent rounds, the new RNA population was split and used to carry out two series of se-

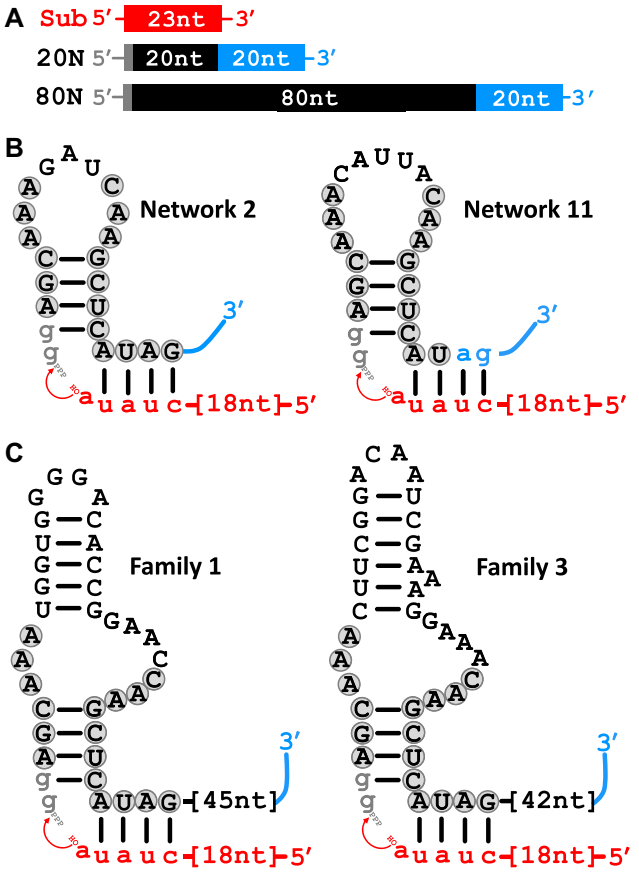

Figure 1. Independent *in vitro* selection experiments starting with RNA populations of two different lengths converged on a shared motif. (**A**) The substrate (red) and two starting populations. The positions of the fully randomized regions (black), 3′ constant regions (light blue), and 2 Gs on the 5′ end (grey) are indicated. (**B**) Secondary structures for two sequences evolved from the 20N population. Representatives of the two most abundant sequence networks that conform to the TL motif (a simple motif defined by a specific ligation junction and conserved terminal loop) are shown bound to the substrate. Shared sequence elements evolved from the randomized regions of both the 20N and 80N populations are indicated with grey circles. (**C**) Secondary structures for sequences evolved from the 80N starting populations. Representatives of the two most abundant sequence families that conform to the IL motif (a more complex motif with the same ligation junction as the TL motif and an internal loop that shares sequence elements with the TL motif's terminal loop) are shown bound to the substrate.

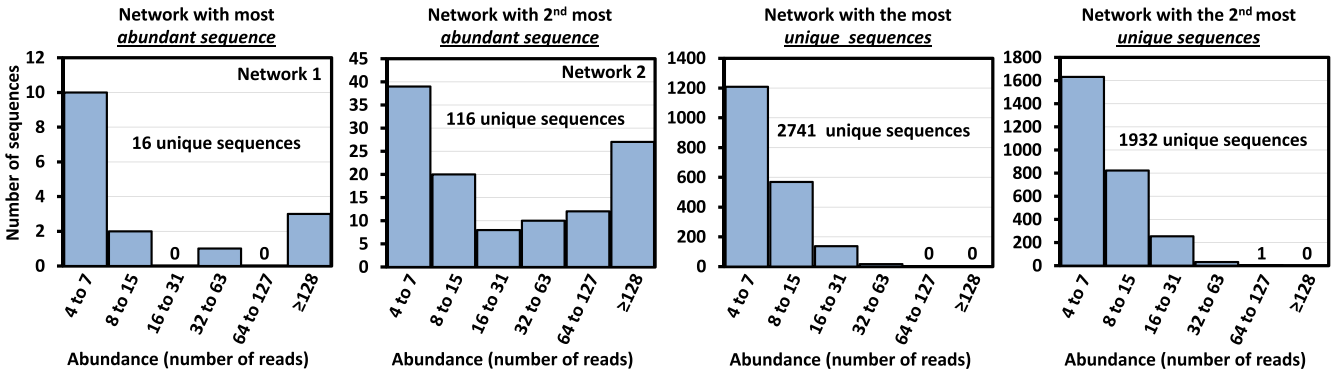

Figure 2. Networks that include the most abundant sequences are small, and the largest networks do not include highly abundant sequences. Distributions of sequence abundances within four different networks from population 20N.1x.r3 (Supplementary Figure S1) are shown. Distributions are shown for the networks to which the first and second most abundant sequences belong, which are small networks that contain other high abundance sequences. Distributions are also shown for the networks that contain the most and second most unique sequences, which do not contain high abundance sequences.

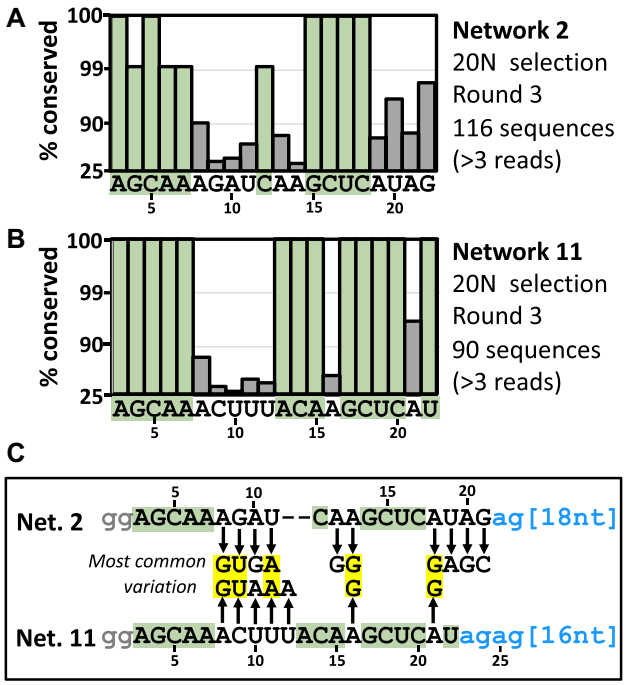

Figure 3. Conservation and variation within multiple sequence networks selected from the 20N population follow similar patterns. Network 2 and Network 11 have similar patterns of conservation and variation, and all sequences in both networks contain the TL motif. (**A**) Among the 116 unique sequences in Network 2, the percentage of sequences that match the consensus sequence is indicated for each position. A log scale is used so differences between highly conserved positions can be seen clearly. Positions with 99% conservation are highlighted in green. (**B**) Among the 90 unique sequences in Network 11 selected from the 20N population, the percentage of sequences that match the consensus sequence is indicated for each position. (**C**) The consensus sequences for Networks 2 and 11 are aligned. Between the two sequences, for each position that is less than 99% conserved, arrows point to the second most common nucleotide found in that position.

lections. In one series of selections, subsequent rounds of selections were carried out in 1× selection buffer, and for the other series, the concentrations of monovalent salts were doubled, providing a slightly more permissive environment for ligation. The two series of selections were both limited to just three total rounds of selection (including the shared first round), which was sufficient to provide clear enrichment while limiting the potential for artifacts due to the amplification steps in each round. The resulting populations were sequenced. The most abundant sequences in each of the two final populations are also present in the other final population; however, the most abundant sequences selected in 1× buffer make up a larger fraction of the total population than those selected with higher ionic strength (Supplementary Figure S2). This difference in the final populations is consistent with the higher ionic strength lowering the selection pressure. Sequences in the final populations were clustered into networks in which every sequence is only one mutation away from at least one other sequence in the same network. The most abundant sequences belong to small, disconnected networks and the largest, most expansive networks are composed of less abundant sequences (Figure 2).

In the 20N selections, abundance is a reasonable proxy for fitness because the random region is sufficiently short to provide a comprehensive sampling, in the starting population ($\sim 10^{16}$ RNAs generated from $\sim 6 \times 10^{14}$ randomly generated DNA templates), of the 4^{20} ($\sim 10^{12}$) possible sequences (20). We therefore focused our analysis on the most abundant sequences and their associated sequence networks (Supplementary Figure S3). After three rounds of selection, many of the most abundant sequences (e.g. 19 of the 50 most abundant in population 20N.1x.r3) belong to networks in which all sequences share a motif that has a specific junction and a conserved terminal loop (TL motif) (Figures 1B and 3 and Supplementary Figure S3). In addition to their abundance, the enrichment of TL motif sequences in round 3 relative to round 2 provides further evidence for their relative fitness (Supplementary Figure S4). Both the junction and the loop that make up the TL motif are formed by the 20N random region of the selected RNAs. The junction establishes proximity between the 3′ end of the substrate and the 5′ end of the ribozyme and undoubtably contributes to ligation of the two RNAs. To determine the contribution of the loop to activity we assayed the ligation activity of a junction-only construct, which contains the ligation junc-

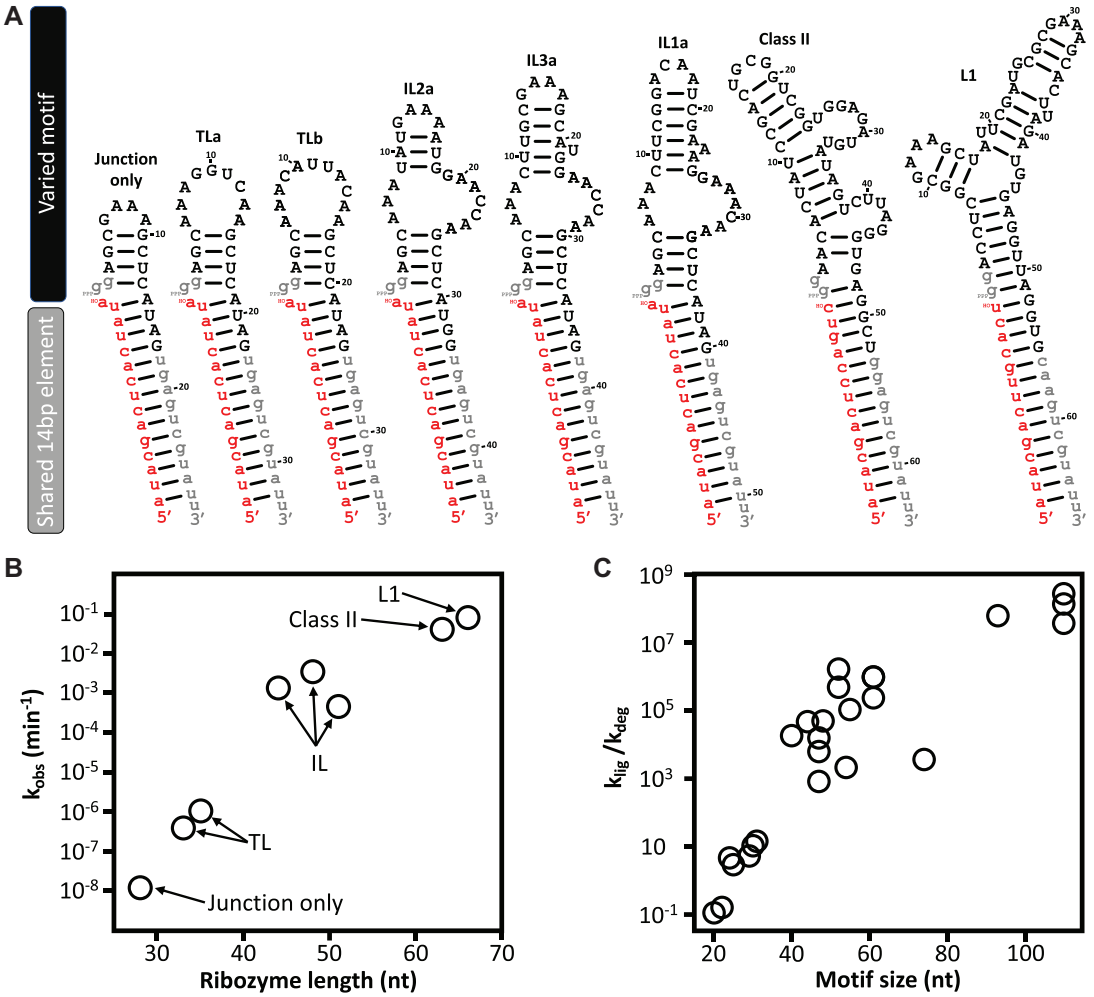

Figure 4. Ligase activity improves with increased length and complexity. (**A**) Secondary structures of ribozymes assayed in a constant structural background of a stem that forms 14 consecutive base-pairs between the ribozyme and the substrate. (**B**) Rate constants for the ribozymes in panel A are plotted as a function of ribozyme length. Rate constants are plotted for representative ribozymes that have TL and IL motifs along with the isolated ligation junction used by these ribozymes. The rate constants measured for the Class II and L1 ligase ribozymes when assayed under our conditions are also shown. (**C**) Normalized rate constants are plotted as a function of motif length for ribozymes assayed in this work and ribozymes assayed in previous independent studies (Supplementary Table S1). To account for different conditions used to measure activity, the rate constants for ligation are normalized to the predicted degradation rate of a flexible RNA phosphodiester bond under the assay conditions used for each ribozyme. The degradation rate was chosen as a normalization factor because ligase ribozymes are in competition with degradation, which destroys both the ligation product and eventually the ribozyme.

tion of the TL motif, but lacks the TL motif's terminal loop. In the junction-only construct the terminal loop sequence of the TL motif was replaced with an unrelated, stable 4-nucleotide loop. The ligation assays showed that the conserved terminal loop sequences present within the TL motif contribute to activity. The 9-nucleotide loop of TLa and the 11-nucleotide loop of TLb increase the rate of ligation relative to the junction-only construct 32-fold and 83-fold, respectively (Figure 4A, B and Supplementary Figure S5). Among the representative sequences derived from the 20N selections that were tested for ligase activity the TL motif-containing sequence, TLb was the most active (Figure 4A and Supplementary Figures S5 and S6).

For an independent set of selections with longer RNAs, we generated an RNA library 102 nucleotides in length with 80 fully randomized positions (80N) (Figure 1A). We initiated two series of 80N selections in which the starting populations for both were drawn from this shared multi-copy RNA library. (Supplementary Figure S1B; 80N selections were performed prior to the 20N selections, see Materials and Methods). For both 80N series, the populations were incubated with the substrate for 24 h in each round of selection and passed through a total of eight rounds. In one series of selections, the first five rounds were carried out in 1x selection buffer, and in the other series, the first five rounds were in 1x selection buffer with the addition of 20% polyethylene glycol to simulate the effects of macromolecular crowding in cellular environments. After the fifth round, both populations were subjected to error-prone PCR and then each was split into three separate populations for the final three rounds, resulting in six final populations (Supplementary Figure S1B). In the last three rounds, selections were carried out in either 1x selection buffer, 1x selection buffer with 20% polyethylene glycol, or 1× buffer with 20%

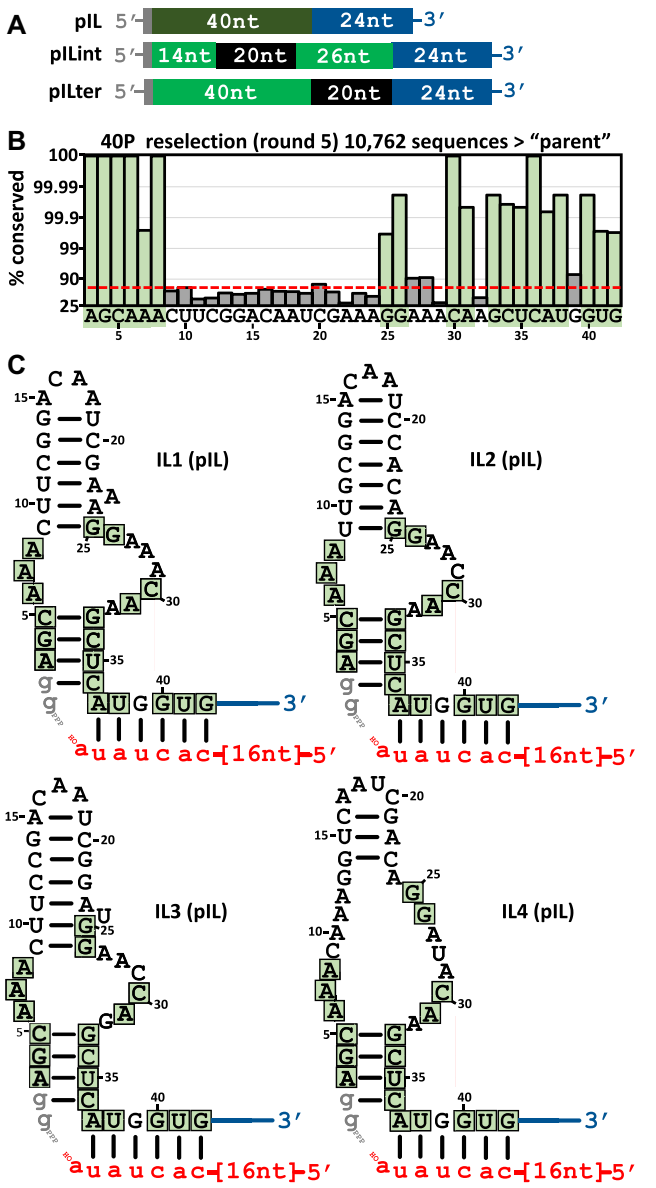

Figure 5. Sequence elements of the TL motif present in the IL motif are highly conserved following reselection. (**A**) The starting populations used in reselection experiments. The partially randomized regions are represented in dark green (18% mutagenesis) and light green (6% mutagenesis), fully randomized regions are in black, 3′ constant regions in blue, and 2 Gs on the 5′ ends are in grey. (**B**) Among the 10,762 unique sequences in the pIL population that were more enriched during reselection than the parent sequence, the percentage of unique sequences that match the parent sequence (% conserved) at each position is indicated. Positions with >99% conservation are highlighted in green. A log scale is used so differences between highly conserved positions can be seen clearly. A dashed red line indicates percent of sequences that matched the parent sequence in the starting population. (**C**) Secondary structures are shown for four sequences that evolved from the pIL population. Representatives of the four most common variants of the IL motif are shown. Positions >99% conserved among the enriched sequences in panel B are indicated with green boxes.

dextran as an alternative crowding agent. Populations were sequenced and sequences were grouped into families on the basis of sequence similarity (fewer than 15 differences). Given the sparse sampling in the 80N starting populations ($\sim 10^{15}$ out of 10^{48} possible sequences), all sequences in a family are likely related by descent from a common ancestor. The most abundant sequences in each of the six final populations are also present in all the other final populations, which is perhaps not surprising since all populations ultimately began from a shared starting library; however, sequence abundances vary considerably between the six final populations. The presence of polyethylene glycol had a relatively large impact on the outcome of the selections, but the presence of dextran did not (Supplementary Figure S7). This result indicates that crowding alone did not have a large impact on the outcome of the selections, but rather, molecular interactions specific to the crowding agent polyethylene glycol were responsible for the differences in the outcomes.

Due to sparse sampling and low copy number in the 80N starting populations, stochastic processes influence sequence abundance, and therefore abundance does not accurately reflect fitness. Rather, the enrichment between rounds is a better indicator of fitness (19,35). We therefore focused on families of ribozymes that were consistently enriched from round five to round eight. Sequences within nine families were enriched >10-fold relative to round five (after error-prone PCR) in at least four of the six final populations. These nine enriched families adopt five distinct structures at the ligation junction (Supplementary Figure S8), and in all cases the predicted ligation junction is formed by sequences within the 80N random region. The most common ligation junction is present in four of these families, three of which converged on a larger structural motif with a specific asymmetric internal loop (IL motif) (Figure 1C and Supplementary Figure S8). The structural elements of the IL motif illustrated in figure 1C are solely responsible for catalysing ligation. The activity of a full-length sequence that contains the IL motif (80N.3.1 (Supplementary Figure S6)) is the same within error as the activity of a truncated and modified version of the sequence (IL1a (Supplementary Figure S5)) that only includes the junction, internal loop and flanking stems (Figure 4A). The presence of the IL motif in multiple families that were highly enriched in multiple selection experiments indicates that this is the most common of the evolutionarily fit motifs within the 80N populations.

Both 20N and 80N *in vitro* selection experiments converge on a shared core motif

The TL motif selected from the 20N population and the IL motif independently evolved from the 80N population are strikingly similar. The two motifs share the same ligation junction, which in both cases is separated from a conserved loop by a 4 bp stem. Also, the sequence on the 5′ side of the IL motif's asymmetric internal loop is the same as the first three nucleotides on the 5′ end of the TL motif's loop, and the sequence of the last three nucleotides at the 3′ end of the 3′ side of the IL motif's internal loop is the same as that of the last three nucleotides on the 3′ end of the TL motif's loop. In brief, the TL motif appears as a discontinuous (with

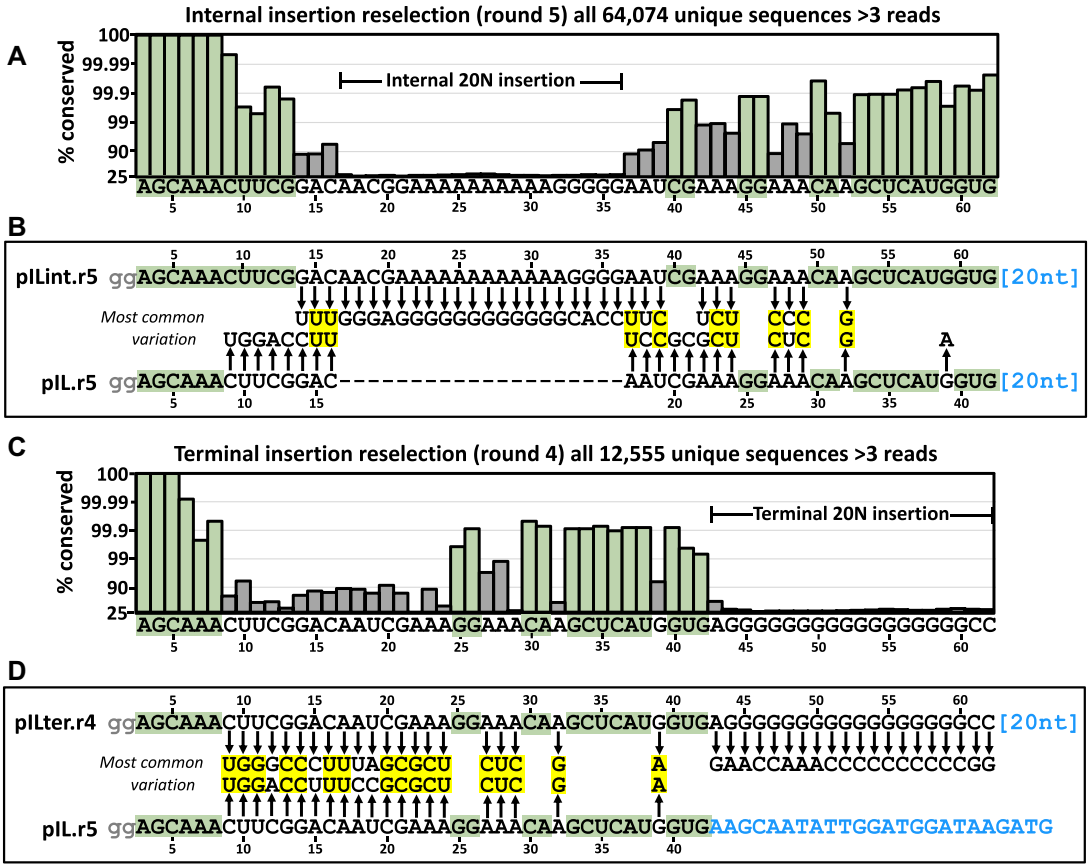

Figure 6. Conservation and variation within the IL motifs selected from the pIL, pILint, and pILter populations follow similar patterns in all populations. (**A**) Among the 64,074 unique sequences present after five rounds of selection with the pILint populations (pILint.r5), the percentage of sequences that match the most commonly observed nucleotide at each position is indicated. Positions with >99% conservation are highlighted in green. A log scale is used so differences between highly conserved positions can be seen clearly. (**B**) Among the pILint.r5 sequences, the most commonly observed nucleotide at each position is aligned to the parent sequence used for reselection in the pIL population. Between the two sequences, for each position that is <99% conserved, arrows point to the second most common nucleotide found in that position. (**C**) Among the 12,555 unique sequences present after four rounds of selection with the pILter populations (pILter.r4), the percentage of sequences that match the most commonly observed nucleotide at each position is indicated. (**D**) Among the pILter.r4 sequences, the most commonly observed nucleotide at each position is aligned to the parent sequence used for reselection in the pIL population. Between the two sequences, for each position that is <99% conserved, arrows point to the second most common nucleotide found in that position.

respect to the linear sequence) component of the larger IL motif (Figure 1B and C). The similarity between these structures in short and long ribozymes reveals the potential for elaboration upon the simpler TL structure, through the insertion of a wide range of sequences within the TL motif's terminal loop, that would result in the formation of an IL motif sequence. The larger size of the IL motif provides for a substantial improvement in activity relative to the TL motif (Figure 4A and B and Supplementary Figure S5) An insertion event that converts a TL motif into an IL motif, even if it was initially converted to a suboptimal IL variant, could therefore confer a significant selective advantage upon the ribozyme. Conversely, a deletion within the TL motif resulting in the smaller junction-only structure leads to diminished activity. To determine whether this length-activity relationship represents a larger trend among ligases, we tested the activity of other previously evolved, truncated, and optimized ligase motifs (36,37) under the same conditions and in the same structural background (Figure 4A) used to assay our TL and IL motifs, and junction-only sequence. The activities of these unrelated ligases conform to the roughly exponential length–activity relationship among our truncated constructs (Figure 4B). A comparison of motif size and activity for ligases evolved here, along with reported values for previously evolved ligases, further supports this trend (Figure 4C and Supplementary Table S1). A similar relationship between length and function has also been reported for GTP-binding RNA motifs (33) (Supplementary Figure S9). It therefore appears that the potential for exponential improvement of fitness in response to increasing motif size is a general feature of functional RNAs over the size range considered above (~20–110 nucleotides).

Reselection experiments demonstrate robustness of the core motif

To probe the plasticity of the core structure of the TL/IL motif, we performed selection experiments with three additional populations derived from an IL motif sequence (Figure 5A): (i) a partially randomized IL sequence with 40 nucleotides mutagenized at 18% per position (pIL); (ii) the same IL sequence mutagenized at 6% per position with

an additional 20-nucleotide, fully randomized sequence inserted between positions 16 and 17 (pILint) and (iii) the same IL sequence mutagenized at 6% per position with an additional 20-nucleotide fully randomized sequence inserted at the 3′ end of the motif (pILter). The populations were subjected to four (pILter) or five (pIL and pILint) rounds of selection with the selection pressure increased relative to the original selections by limiting the incubation with the substrate to 1 h (Supplementary Figure S1C). In all cases, the core sequence elements of the IL motif were strongly reselected (Figures 5B-C and 6). Specifically, features that were present in the TL motif were preserved within the IL structures. Patterns of conservation and variation are similar, both among the reselected populations (Figures 5B and 6) and between the reselected populations and the TL sequences from the 20N selection (Figures 3, 5B and 6). For example, in all reselected populations and among the TL sequences, the last three nucleotides at the 3′ end of the loop follow the same pattern of conservation, with the C in the first position consistently the most conserved, the A in the second position the second most conserved, and the A in the final position the least conserved, G being the most common variant in that last position in all four cases. Elements outside the shared central core varied in the reselected populations—four different peripheral elements were selected (Figure 5C and Supplementary Figure S10).

We identified conceptual paths by which our ligases can improve function through a combination of sequence insertions and point mutations. Sequence insertion allows for elaboration upon the smaller, simpler, and far less fit TL motif to form any of the four IL variants. Given sufficient length, IL1, IL2 and IL3 can be interconverted through a continuous series of functional point mutants that were enriched in the selected populations (Figure 7). Point mutations can therefore provide an avenue for IL1 and IL2 to evolve to the more active IL3. Within our data set, we do not observe a continuous series of functional point mutants that connect IL1, IL2 or IL3 sequences with IL4 sequences. Aside from the TL motif, motifs in the 20N population show little similarity to the structures predicted for the 80N ribozymes (Supplementary Figures S3 and S8). This indicates that there are fewer ways to increase fitness by elaboration upon these other 20N motifs. The TL sequences are therefore not only relatively fit, but also more likely to benefit from sequence insertion than other short sequences.

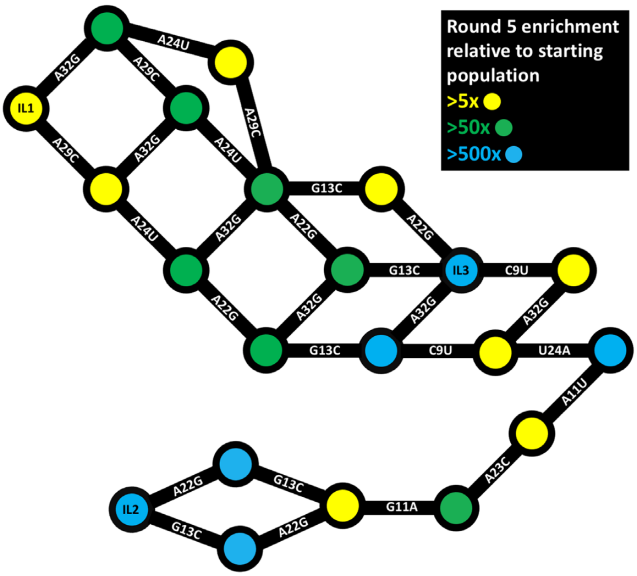

Figure 7. IL1 and IL2 sequences can evolve to more active IL3 sequences by passing through a continuous series of fit mutants separated by single point mutations. Among the sequences in the pIL.r5 population that were enriched at least as much as the parent sequence, the shortest continuous series of point mutations connecting the most abundant IL1 and IL2 sequences to the most abundant IL3 sequence are shown. Circles represent sequences that are enriched at least as much as the parent sequence. Lines between circles are point mutations that connect the sequences in sequence space. The color of the circles indicates the magnitude of the enrichment as indicated in the key.

DISCUSSION

For the 20N populations examined here and RNA populations of similar length used in prior experiments (19,20), fit RNA sequences belong to networks of functional point mutations that are small and disconnected. Therefore, the absence of viable paths (via the accumulation of point mutations) between disconnected clusters of fit sequences appears to be a universal feature among short RNAs. For longer RNAs evolved from the pIL population, we observe a strict conservation of the core motif and no evidence for evolution of new global structures. Prior reselection experiments also consistently yield sequences predicted to conform to the original parent structure (21,36,38). This is at variance with the prediction that longer sequences should form highly interconnected networks of functional structures (12). Additionally, insertions in our ligases do not support the emergence of new structures. Both pILint and the pILter only give rise to sequences in which the original core motif remains conserved. Prior *in vitro* selection experiments with recombined populations also result in preservation of the original core structural motifs within new contexts (25–27). Taken together, these results suggest that even though point mutations or insertions could, in principle, lead to remodeling of core structural elements, elaboration upon a stable core is the more likely evolutionary mechanism to improve function.

The advantages conferred by increasing length and complexity (33) (Figure 4), the frequency of spontaneous recombination between RNA molecules (22,23), and barriers to the exploration of sequence space through point mutations (19–21) (Figures 2, 3 and 5) are factors that, when considered in combination, support the idea that the evolution of biological RNAs is characterized by elaboration without disruption of preexisting structures. This, in turn, means that ancestral forms can be inferred from the analysis of modern RNA structures (1–3).

Our results have implications for the roles of chance and necessity in the early evolution of life. If the sampling of short RNA sequences was sufficient, which is a reasonable scenario in early evolution of RNA, then our results suggest that the outcomes of RNA evolution were not significantly influenced by chance. Short, functional RNA motifs with structures that were predisposed to improve fitness through elaboration would have, following elaboration, reliably outcompeted the other motifs. In our experiments, rel-

atively fit 20N sequences adopt the TL motif, which, unlike the other sequences, can become significantly more fit through the insertion of a wide variety of sequences. Reselection and recombination experiments indicate that once simple structures, such as the TL motif, are elaborated upon to form more fit structures, such as the IL motif, the emergence of new and potentially better structural solutions are unlikely either through point mutations or insertions. This means that the outcome of early RNA evolution may not have been significantly influenced by chance, even though sequence space was only explored locally.

DATA AVAILABILITY

NCBI's Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra). Sequence data are deposited with the NCBI's Sequence Read Archive. Bioproject accession number PRJNA668406.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Aeronautics and Space Administration's Planetary Science Division Research Program; NASA's Exobiology Program [cooperative agreements NNX16AN16A, NNX15AM88A]. Funding for open access charge: National Aeronautics and Space Administration (NASA).
Conflict of interest statement. None declared.

REFERENCES

1. Petrov,A.S., Gulen,B., Norris,A.M., Kovacs,N.A., Bernier,C.R., Lanier,K.A., Fox,G.E., Harvey,S.C., Wartell,R.M., Hud,N.V. *et al.* (2015) History of the ribosome and the origin of translation. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 15396–15401.
2. Petrov,A.S., Bernier,C.R., Hsiao,C., Norris,A.M., Kovacs,N.A., Waterbury,C.C., Stepanov,V.G., Harvey,S.C., Fox,G.E., Wartell,R.M. *et al.* (2014) Evolution of the ribosome at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 10251–10256.
3. Gray,M.W. and Gopalan,V. (2020) Piece by piece: building a ribozyme. *J. Biol. Chem.*, **295**, 2313–2323.
4. Ellis,J.C. and Brown,J.W. (2009) The RNase P family. *RNA Biol*, **6**, 362–369.
5. Wu,J., Niu,S., Tan,M., Huang,C., Li,M., Song,Y., Wang,Q., Chen,J., Shi,S., Lan,P. *et al.* (2018) Cryo-EM structure of the human ribonuclease P holoenzyme. *Cell*, **175**, 1393–1404.
6. Rosenblad,M.A., Larsen,N., Samuelsson,T. and Zwieb,C. (2009) Kinship in the SRP RNA family. *RNA Biol*, **6**, 508–516.
7. Kunkel,T.A. (2004) DNA replication fidelity. *J. Biol. Chem.*, **279**, 16895–16898.
8. Tjhung,K.F., Shokhirev,M.N., Horning,D.P. and Joyce,G.F. (2020) An RNA polymerase ribozyme that synthesizes its own ancestor. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 2906–2913.
9. Schuster,P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.*, **255**, 279–284.
10. Ancel,L.W. and Fontana,W. (2000) Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.*, **288**, 242–283.
11. Fontana,W. and Schuster,P. (1998) Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, **194**, 491–515.
12. Gavrilets,S. and Gravner,J. (1997) Percolation on the fitness hypercube and the evolution of reproductive isolation. *J. Theor. Biol.*, **184**, 51–64.
13. Gravner,J., Pitman,D. and Gavrilets,S. (2007) Percolation on fitness landscapes: effects of correlation, phenotype, and incompatibilities. *J. Theor. Biol.*, **248**, 627–645.
14. Fontana,W. and Schuster,P. (1998) Continuity in evolution: on the nature of transitions. *Science*, **280**, 1451–1455.
15. Gavrilets,S. (2004) In: *Fitness Landscapes and the Origin of Species*. Princeton Univ. Press, Princeton.
16. Kauffman,S. and Levin,S. (1987) Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, **128**, 11–45.
17. Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
18. Bendixsen,D.P., Collet,J., Ostman,B. and Hayden,E.J. (2019) Genotype network intersections promote evolutionary innovation. *PLoS Biol.*, **17**, e3000300.
19. Pressman,A.D., Liu,Z., Janzen,E., Blanco,C., Muller,U.F., Joyce,G.F., Pascal,R. and Chen,I.A. (2019) Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for Self-Aminoacylating RNA. *J. Am. Chem. Soc.*, **141**, 6213–6223.
20. Jimenez,J.I., Xulvi-Brunet,R., Campbell,G.W., Turk-MacLeod,R. and Chen,I.A. (2013) Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 14984–14989.
21. Petrie,K.L. and Joyce,G.F. (2014) Limits of neutral drift: lessons from the in vitro evolution of two ribozymes. *J. Mol. Evol.*, **79**, 75–90.
22. Smail,B.A., Clifton,B.E., Mizuuchi,R. and Lehman,N. (2019) Spontaneous advent of genetic diversity in RNA populations through multiple recombination mechanisms. *RNA*, **25**, 453–464.
23. Mutschler,H., Taylor,A.I., Porebski,B.T., Lightowlers,A., Houlihan,G., Abramov,M., Herdewijn,P. and Holliger,P. (2018) Random-sequence genetic oligomer pools display an innate potential for ligation and recombination. *eLife*, **7**, e43022.
24. Kaessmann,H., Vinckenbosch,N. and Long,M. (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.*, **10**, 19–31.
25. Burke,D.H. and Willis,J.H. (1998) Recombination, RNA evolution, and bifunctional RNA molecules isolated through chimeric SELEX. *RNA*, **4**, 1165–1175.
26. Wang,Q.S. and Unrau,P.J. (2005) Ribozyme motif structure mapped using random recombination and selection. *RNA*, **11**, 404–411.
27. Plebanek,A., Larnerd,C., Popovic,M., Wei,C., Pohorille,A. and Ditzler,M.A. (2019) Big on change, small on innovation: evolutionary consequences of RNA sequence duplication. *J. Mol. Evol.*, **87**, 240–253.
28. Fujishima,K. and Kanai,A. (2014) tRNA gene diversity in the three domains of life. *Front. Genet.*, **5**, 142.
29. Widmann,J., Di Giulio,M., Yarus,M. and Knight,R. (2005) tRNA creation by hairpin duplication. *J. Mol. Evol.*, **61**, 524–530.
30. Cadwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl.*, **2**, 28–33.
31. Alam,K.K., Chang,J.L. and Burke,D.H. (2015) FASTAptamer: a Bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Mol. Ther. Nucleic Acids*, **4**, e230.
32. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
33. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, **126**, 5130–5137.
34. Li,Y. and Breaker,R.R. (1999) Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2′-Hydroxyl group. *J. Am. Chem. Soc.*, **121**, 5364–5372.
35. Cho,M., Xiao,Y., Nie,J., Stewart,R., Csordas,A.T., Oh,S.S., Thomson,J.A. and Soh,H.T. (2010) Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 15373–15378.
36. Pitt,J.N. and Ferre-D'Amare,A.R. (2010) Rapid construction of empirical RNA fitness landscapes. *Science*, **330**, 376–379.
37. Robertson,M.P., Hesselberth,J.R. and Ellington,A.D. (2001) Optimization and optimality of a short ribozyme ligase that joins non-Watson-Crick base pairings. *RNA*, **7**, 513–523.
38. Diaz Arenas,C. and Lehman,N. (2010) Quasispecies-like behavior observed in catalytic RNA populations evolving in a test tube. *BMC Evol. Biol.*, **10**, 80.