Published online 21 June 2017

Nucleic Acids Research, 2017, Vol. 45, No. 14 ***8167–8179***
doi: 10.1093/nar/gkx540

NAR Breakthrough Article

Analysis of *in vitro* evolution reveals the underlying distribution of catalytic activity among random sequences

Abe Pressman[1,2], Janina E. Moretti[3], Gregory W. Campbell[1,4], Ulrich F. Müller[3,*] and Irene A. Chen[1,4,*]

[1]Department of Chemistry and Biochemistry 9510, University of California, Santa Barbara, CA 93106, USA, [2]Program in Chemical Engineering, University of California, Santa Barbara, CA 93106, USA, [3]Department of Chemistry and Biochemistry, University of California, San Diego, CA 92093, USA and [4]Program in Biomolecular Sciences and Engineering, University of California, Santa Barbara, CA 93106, USA

Received April 24, 2017; Revised June 08, 2017; Editorial Decision June 09, 2017; Accepted June 12, 2017

ABSTRACT

The emergence of catalytic RNA is believed to have been a key event during the origin of life. Understanding how catalytic activity is distributed across random sequences is fundamental to estimating the probability that catalytic sequences would emerge. Here, we analyze the *in vitro* evolution of triphosphorylating ribozymes and translate their fitnesses into absolute estimates of catalytic activity for hundreds of ribozyme families. The analysis efficiently identified highly active ribozymes and estimated catalytic activity with good accuracy. The evolutionary dynamics follow Fisher's Fundamental Theorem of Natural Selection and a corollary, permitting retrospective inference of the distribution of fitness and activity in the random sequence pool for the first time. The frequency distribution of rate constants appears to be log-normal, with a surprisingly steep dropoff at higher activity, consistent with a mechanism for the emergence of activity as the product of many independent contributions.

INTRODUCTION

In vitro evolution of RNA from random sequence pools has a long history of success in identifying sequences with novel chemical function, such as catalytic RNAs (1–3). By exploring a large sample of sequence space, *in vitro* evolution also probes the underlying distribution of molecular fitness among random sequences of nucleic acids. Insofar as fitness correlates with activity, a fitness distribution also reflects the corresponding chemical activity distribution (e.g. a ribozyme's k_{cat}) over a sequence space (4). The shape of such frequency distributions is a fundamental open issue for understanding the emergence and evolution of an RNA World during the early stages of life (5). Knowledge of activity distributions would yield insight into the likelihood, repeatability, and activity level of ribozyme emergence during prebiotic scenarios as well as during *in vitro* selection of new catalysts. Knowledge of the related distribution of catalytic activation energies may also offer suggestions on an underlying mechanism for the emergence of function from sequence. From a practical perspective, knowledge of the fitness and/or activity distribution underlying a selection is also necessary to accurately design the best conditions for selection (6).

Despite the high degree of interest (6,7), prior work estimating the underlying fitness and chemical activity distributions of any *in vitro* selection has been relatively limited. Studies on selections from starting pools with different sequence complexity suggested a power-law relation between pool size and aptamer affinity or ribozyme activity, though only a small number of measurement points were available (7,8). In contrast, theoretical considerations suggested a log-normal distribution of K_D values (and a normal distribution of binding energies) in sequence space for nucleic acid aptamers (9,10), as well as a normal distribution of activation energies of RNA melting (11). Previous experimental work has measured fitness over limited sequence spaces, e.g. ribozymes in which several nucleotides (or pep-

*To whom correspondence should be addressed. Tel: +1 805 893 8364; Fax: +1 805 893 4120; Email: chen@chem.ucsb.edu
Correspondence may also be addressed to Ulrich F. Müller. Tel: +1 858 534 6823; Fax: +1 858 534 6255; Email: ufmuller@ucsd.edu

tides) were randomized (12–14) or aptamers containing a small enough region of variation to explore within the span of a microarray (15,16), or in special scenarios in which DNA sequencing itself can be used as an assay of function (e.g. for self-cleaving activity) (17). Thus, empirical measurement of fitness and chemical activity distributions over a highly diverse, random pool has been lacking.

Extracting distributions from *in vitro* evolution has been historically hampered by the low throughput of sequencing data. Now, high-throughput sequencing (HTS) has become a useful tool for addressing deep evolutionary questions (12,18,19), e.g. enabling quantitative analysis of the fitness 'landscape' in sequence space, including the distribution of fitness peaks and neutral evolutionary pathways (13,20,21), as well as revealing the importance of cryptic variation in rapid adaptation (21,22). On the practical side, HTS of *in vitro* evolution of RNA is useful for identifying high-fitness sequences (23,24). To estimate fitness, HTS analyses typically count sequences present at a single round of selection, although more recent analyses use the relative enrichment of sequences before and after a final round (23–25), or follow a specific ribozyme and its variants over several rounds (26). Integrative analysis of HTS data over the whole course of an *in vitro* evolution from random sequence would take advantage of more data to provide more accurate estimation of the fitness of optimal sequences, while also providing insight into the consistency of selection and the accuracy of such estimation. Here we develop a method to estimate fitness from multiple rounds of *in vitro* evolution, translate fitness into chemical activity, and infer the underlying distributions of an initially random pool of RNA. We apply this analysis to a previously performed *in vitro* evolution of triphosphorylation ribozymes (27) and validate inferences on catalytic activity. Through this analysis, we identify substantially improved ribozymes and discover the approximately log-normal shape of the underlying frequency distribution of ribozyme activities.

MATERIALS AND METHODS

Ribozyme selection and sequencing

Triphosphorylating ribozymes using cyclic trimetaphosphate (Tmp) were selected in a previous study (27) (Figure 1A). Ribozymes were selected for converting their 5′-hydroxyl group to a 5′-triphosphate by reacting with Tmp. In brief, the selection consisted of: (i) processing of an RNA construct, including a 150-nt randomized region, with a hammerhead ribozyme to create a 5′-OH, (ii) incubation of the pool with Tmp such that active ribozymes generate a 5′-triphosphate, (iii) selection of the 5′-triphosphorylated molecules by reaction with a biotinylated RNA oligonucleotide and a ligase ribozyme and (iv) RT-PCR to generate the pool for the subsequent selection round. Selection rounds used constant reaction conditions for this triphosphorylation reaction step, with the following exception: Rounds 1–4 were carried out using a 3-h incubation with Tmp, while Rounds 5–8 were split into two selection branches, one with 3 h (branch '3h') and one with 5 min (branch '5m') of Tmp incubation.

DNA samples from Rounds 1–8 (including both branches) as well as the initial pool (Round 0) were submitted for high-throughput sequencing on the Illumina MiSeq platform with paired-end 150 bp reads. Mutagenic PCR during amplification between Rounds 5–8 would result in an expected mutation rate of 1.7% per base (28,29), substantially greater than the expected error rate from paired-end sequencing. In addition, because of the sparse coverage of sequence space, clusters are expected to be highly distinct, and thus a correction for sequencing errors (20) was not performed. (See further details in Supplemental Text S1.)

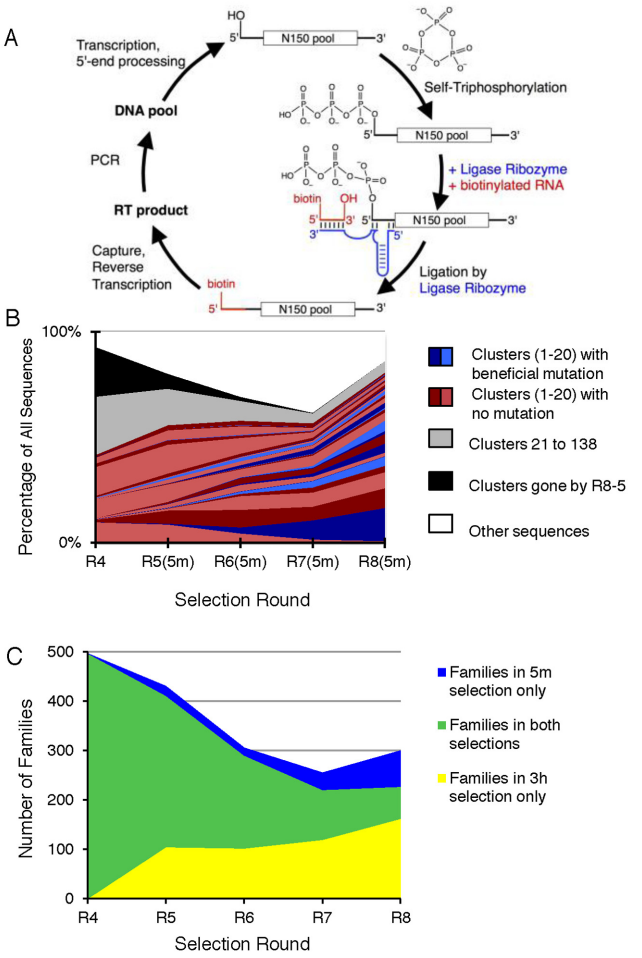

Figure 1. Selection scheme and evolutionary dynamics of ribozyme families and clusters during Rounds 4–8. (**A**) Abbreviated depiction of selection scheme for triphosphorylation ribozymes (27). (**B**) The abundance of clusters over time in the 5m branch of selection: red areas represent clusters close to the original centers of the top 20 families of Round 8 (5m), blue areas represent new clusters whose central sequence diverged from the original family center, gray area represents clusters from families ranked 20–138 in abundance, solid black area represents families and clusters that disappeared by Round 8 (5m), white area represents other low abundance sequences. (**C**) The 5m and 3h selection branches were separated after Round 4; green areas represent families present in both branches, yellow areas represent families present only in the 3h branch, blue areas represent families present only in the 5m branch.

Experimental determination of ribozyme activities

Experimental determination of ribozyme activities was performed as previously described (27), allowing direct com-

parison to the previously reported ribozyme activities. The equivalence of assay conditions was confirmed by re-measuring the activity of four clones from the previous study, which resulted in the same values within error (Supplemental Figure S1). In brief, the RNA sequence containing a 5′-hydroxyl group was incubated with Tmp, generating a 5′-triphosphorylated RNA product, and then reacted with R3C ligase ribozyme (30) and a radiolabelled substrate oligonucleotide. Ligated products were separated from unreacted oligonucleotides by denaturing PAGE and quantified by phosphorimaging, and each reaction was measured at least three times. The activity assay conditions are the same as in the *in vitro* selection, thereby measuring fitness as experienced during the evolution procedure. (See further details in Text S1.) Apparent first-order rate constants measured this way are termed k_{app}.

Assignment of sequences to families

Sequences were sorted by similarity using custom software running on the Galaxy bioinformatics platform (20,31), with the highest-abundance sequence in each family defined as the center, for each round. In general, families were separated by large Hamming distances from each other due to the long length of the random region, allowing unambiguous assignment of sequences to families. Some families displayed a shift in center sequence across rounds, typically consisting of 1–2 nucleotide mutations (termed 'notable' mutations). For some analyses, similarity to one of multiple variant centers was used to 'split' families into two or more new clusters, treated independently, as the notable mutations impacted fitness. We use the term 'cluster' to refer to a group of related sequences for which such splits were performed when appropriate (and 'cluster' is equivalent to 'family' when no split was appropriate). Further details on sequence-grouping procedures are described in previous work (20,31) and in Text S1.

Best method to estimate fitness from HTS data

We define the *abundance* (A_R) of a sequence (or cluster) in a given round (R) as the frequency of its reads in that round, ranging from 0 to 1. Each round contained at least 1.5 million sequence reads, so the abundance for a single read is slightly less than 10^{-6}. We define the *enrichment* (E_R) of a sequence (or cluster) in round R as $A_R/A_{R\text{-}1}$. During an incubation period t, we define F as the fraction of a particular sequence that has reacted and is carried forth to the next round's sequence pool. F is the absolute viability of a particular sequence, ranging from 0 to 1, which should remain constant under unchanging selection conditions and represents the chemical activity of the sequence. If differences in the amplification tendency (or fecundity) of the sequences are neglected, then F is also the *relative fitness* (also called 'w' in the evolutionary biology literature). For any sequence, E_R is expected to be proportional to F by a round-dependent scaling factor (S_R) that primarily expresses the inverse of the average F of the pool in that round ($E_R/S_R = F$). (E_R values are not expected to be constant across rounds, as they depend on the fitness of the other sequences in the pool in a given round.) Several methods were used to estimate a single enrichment value (E_e, expected to scale with estimated viability F_e) from E_R values from multiple rounds, for each cluster or each sequence. Method 1 of fitness estimation used $E_{R\text{-}1}$ (normalized by average enrichment for round R-1) as a predictor of normalized E_R; Method 2 used a geometric mean of multiple E_R values to estimate E_e; Methods 3–5 used different modifications of summation of squares to build linear estimates of E_e; and Method 6 used Maximum Likelihood Estimation to estimate E_e. The weighted coefficient of determination (r^2) between E_R and E_e was used to test the consistency of each estimation method. Because F is directly proportional to E_R, the value of r^2 between E_R and F is the same as between E_R and E_e, allowing the best method for estimation of F to be determined through analysis of E_e (see further details in Supplemental Text S1, Figure S2).

Conversion of evolutionary fitness to catalytic activity

To estimate ribozyme activity from HTS data alone (i.e. independent of biochemical assays), we made the following approximation. We assumed that triphosphorylation follows pseudo-first-order kinetics, as observed for ribozymes previously isolated from this selection,(27) such that $F = L * (1 - e^{-kt})$ where k is the rate constant (k_{cat}[Tmp]), and L is the maximum extent of triphosphorylation followed by ligation. The constant L includes the fraction of molecules that fold into the active structure for triphosphorylation catalysis and their ability to act as a substrate for the ligase ribozyme. To calculate F from E_e, an overall scaling factor (S_e, resulting from a combination of S_R) was estimated as described below for each branch of selection, as $E_e/S_e = F_e$. For the purpose of comparing 5m and 3h HTS data to determine the scaling factors S_e, in order to minimize the number of fitting parameters, we approximate average $L = 0.5$ across sequences. For any given ribozyme, L may vary from 1, as sequences vary with respect to their optimality as a ligation substrate. The approximation of average $L = 0.5$ can be justified (i) because the previous study (27) found that most amplitudes were in the range of 0.5, (ii) because an incorrect estimate of average L does not affect the relative scaling of L for individual ribozymes, but only affects the absolute scale of estimated L values and (iii) *post hoc* because the resulting activity estimates conform well with values measured biochemically (see Results). Setting k equal at two t measurements gave the mathematical relation $E_e(t_1)/L = S_e(t_1)(1 - [1 - \frac{E_e(t_2)/L}{S_e(t_2)}]^{t_1/t_2})$ which was fit to the HTS-derived E_e for 5m and 3h selection branches (using sequence clusters), thereby obtaining the two scaling factors S_e(5m) and S_e(3h), using the Matlab curve-fitting toolbox. S_e(5m) was used to calculate an absolute estimate of F_e(5m) for clusters from the 5-min selection, and similarly for S_e(3h) and F_e(3h). The ratio F_e(5m)/F_e(3h) was used to calculate an estimated k for each sequence as $\frac{F_e(5m)}{F_e(3h)} = \frac{(1-e^{-5k})}{(1-e^{-180k})}$, where k has units min^{-1}. The estimated value of L for an individual ribozyme cluster was calculated using estimated k and F_e in the relation $F = L * (1 - e^{-kt})$. If the approximation of average $L = 0.5$ was inaccurate, the estimated L values would be affected in proportion (see further

details in Supplemental Text S1, Figure S3). L and k estimated in this way are termed L_{est} and k_{est}. As both L and k were expected to affect sequence enrichment, we used estimated initial rate ($L_{est}k_{est}$) to compare measured and predicted sequence activities.

Analysis of evolutionary noise using HTS data

To analyze the noise present in observations of fitness during the selection experiments, a distribution of noise was calculated from the absolute difference between E_R and the expectation of E_R generated from F (i.e. E_e) for each sequence in each round. (See further details in Text S1.)

Evaluation of selection behavior using Fisher's Fundamental Theorem of Natural Selection and its corollary

Fisher's Fundamental Theorem of Natural Selection (FFTNS; also called the 'growth rate theorem') (32–35) states that, under ideal haploid conditions, change in population mean fitness should equal normalized population variance, such that mean(F, R+1) – mean(F, R) = Var(F, R)/mean(F, R), with mean(F, R) being abundance-weighted mean F distributed over all sequences in the Round R pool, and Var(F, R) being abundance-weighted variance of F in the same round. This relation, and a corollary to FFTNS derived in the Supplemental Text S1, which relates the change in variance to the skewness of the fitness distribution, were used to determine whether the shape of estimated fitness distribution evolved in a fashion consistent with evolutionary theory. (See further details in Text S1.)

Retrospective inference of underlying fitness and activity distributions

The 3h selection branch was chosen for further analysis of the fitness distribution, having undergone identical selection conditions from Round 1 through 8. F_e values were binned into an empirical probability distribution function of estimated fitness for Rounds 4–8, $P_R(F_e)$ (with P_R denoting probability distribution at round R), normalized to integrate to 1. Fitness distributions for Rounds 1–3 were calculated retrospectively, with $P_{R-1}(F_e) = \frac{P_R(F_e)/F_e}{\int_0^1 (P_R(F_e)/F_e) dF_e}$, as the relative abundance of a sequence in a preceding round is expected to be its relative abundance in the current round divided by F. (See further details in Supplemental Text S1).

The retrospectively inferred distribution of F in the random pool (i.e. $P_0(F_e)$) was converted to a distribution of catalytic rate constants k by the relation $F_e = (1 - e^{-kt})\ L$. For this analysis, the assumption of $L = 0.5$ was used, justified as described above. To focus on fitting the long tail of the initial distribution, curve fitting was performed with log-transformed variables. To account for difficulty in estimating the precise area under the low-fitness end of the distribution curve, each theoretical probability distribution function was multiplied by a constant parameter that was allowed to vary during curve fitting. The initial distribution of rate constants was compared to candidate distribution shapes using the Matlab curve-fitting toolbox and standard Trust-Region fitting parameters. We investigated fit to an exponential, Pareto (scale-free), and log-normal distribution, fitting with parameters selected to maximize r^2 for log-log-transformed values of F_e and $P_R(F_e)$.

Any data and code not previously described (20,31) will be made available by the authors upon request.

RESULTS

Evolutionary dynamics of ribozyme families

High-throughput sequencing (HTS) of Rounds 0–8 provided 1.6–4.8 million usable sequences per round (Table 1). Sequences were sorted into distinct families based on sequence similarity. Because the initial RNA pool represented a minute fraction of sequence space for random 150-mers, families are expected to be highly distinct, and detectable sequence similarity almost certainly indicates relatedness by descent. Indeed, the Hamming distance between all sequences in the pool shows a maximum ∼110, as expected for completely random sequences with a length of 150 nucleotides (Supplementary Figure S4). Additionally, the number of families was robust to the value of the pairwise Hamming distance cutoff from 60 to 90, with a distance of 70 chosen to allow unambiguous family assignment and alignment across rounds (Supplementary Figure S4).

Analyzing multiple rounds of the ribozyme selection permitted tracking of the sequence families and clusters, their distribution, and their relative abundance. Sequence families could not be reliably identified in Rounds 1 and 2 due to the large number of unique sequences, but 829 ribozyme families were identified in Round 3. These were gradually winnowed over subsequent rounds. Over a hundred unique families were present at the end of Round 8 of both branches of selection, and several of the major families were best analyzed after splitting into clusters based on the presence or absence of notable mutations (see below). The top 20 families comprised ∼80% of the pool (Figure 1B). The presence of many ribozyme families indicated that a low-throughput approach would not be sufficient to identify the fittest sequences (Supplementary Figure S5A). Indeed, previous analysis of the selection—in which 16 colonies selected from transformants from Round 5 and 20 colonies selected from transformants from Round 8, were Sanger sequenced and assayed for activity—had identified members of some, but not all, high-abundance families, consistent with the diversity of families we observed (Supplementary Figure S5B and previous work (27)). Beginning in Round 4, ribozyme selection was dominated by competition among dozens of sequence clusters. In Round 4, the top five clusters composed a larger percentage of the pool than in Round 8 for either selection branch, illustrating the emergence of many clusters of similarly high fitness toward the end of the selection. Many high-abundance clusters in Round 8(5m) were extremely scarce at Round 4 (Figure 1B), suggesting that these clusters had a fast relative enrichment rate and thus high fitness. While each cluster tended to either increase or decline in relative abundance over multiple rounds, abundances taken at a single round would be a poor predictor of overall evolutionary fitness. Indeed, abundances did not correlate with observed ribozyme activity previously measured for isolated clones (Supplementary Figure S6).

Table 1. Number of raw sequence reads, sequences extracted based on presence of known adapter sequences, and unique sequence families present in each round of ribozyme selection. 5m and 3h indicate the two branches of selection

Sample	# of sequence reads (# of extracted sequences)	# of sequence families
Round 0 (initial pool)	2 376 803 (2 182 213)	n.d.
Round 1	2 721 407 (2 623 400)	2
Round 2	2 682 133 (2 572 500)	6
Round 3	2 569 736 (2 486 543)	829
Round 4	2 309 537 (2 256 055)	497
Round 5(3h)	1 815 219 (1 758 685)	411
Round 5(5m)	1 615 498 (1 563 259)	327
Round 6(3h)	1 932 193 (1 869 175)	290
Round 6(5m)	1 719 683 (1 658 222)	204
Round 7(3h)	1 624 442 (1 560 875)	220
Round 7(5m)	1 710 580 (1 648 206)	136
Round 8(3h)	4 781 335 (4 516 812)	227
Round 8(5m)	3 989 007 (3 771 315)	138

The 5m and 3h selection branches showed high overlap between families initially, as expected since they were both derived from the same earlier rounds. The similarity then decreased such that about half of the families from the 5m branch were also found in the 3h branch in Round 8 (Figure 1C). The higher-stringency 5m selection branch experienced a significantly faster loss of pool diversity and had fewer sequence families at the end of selection, and therefore a larger fraction of its families were shared compared to the 3h branch.

Best method for estimation of fitness

For all sequences present in our selection pools, ribozyme fitness was estimated under two categories: individual sequences and sequence clusters. In the clustered approach, each family was generally assigned a single estimated fitness value based on the total abundance of that family. However, some families underwent notable changes in the center sequence (i.e. the sequence of highest abundance) during the selection, usually corresponding to a single nucleotide mutation initially appearing in a few unique sequences and then increasing to become the majority of a family towards later rounds. We refer to these mutations as 'notable'. Such families were identified by a change in center sequence between Rounds 4 and 8 (amplification in these rounds included mutagenic PCR). In these cases, families could be split into two or more clusters based on similarity to the old or new center.

As many prominent clusters and sequences were not present at detectable abundance until Round 4, estimated fitness values (F_e) and the corresponding enrichment values (E_e) were calculated for all sequences and clusters in Rounds 4–8 from the 5m and 3h selection branches using Methods 1–6. These methods integrate information from multiple rounds in different ways, depicted graphically in Supplementary Figure S2. Coefficient of determination (r^2), calculated between E_e and observed E_R (weighted by sequence or cluster abundances), was used to evaluate the methods. In general, the correlations appeared linear, as expected, and estimation of an overall fitness was more descriptive of all rounds compared to fitness from any individual round (Figure 2A shows one example, with all comparisons detailed in Supplementary Figure S7). Correlations based on clusters were substantially better than correlations based on in-

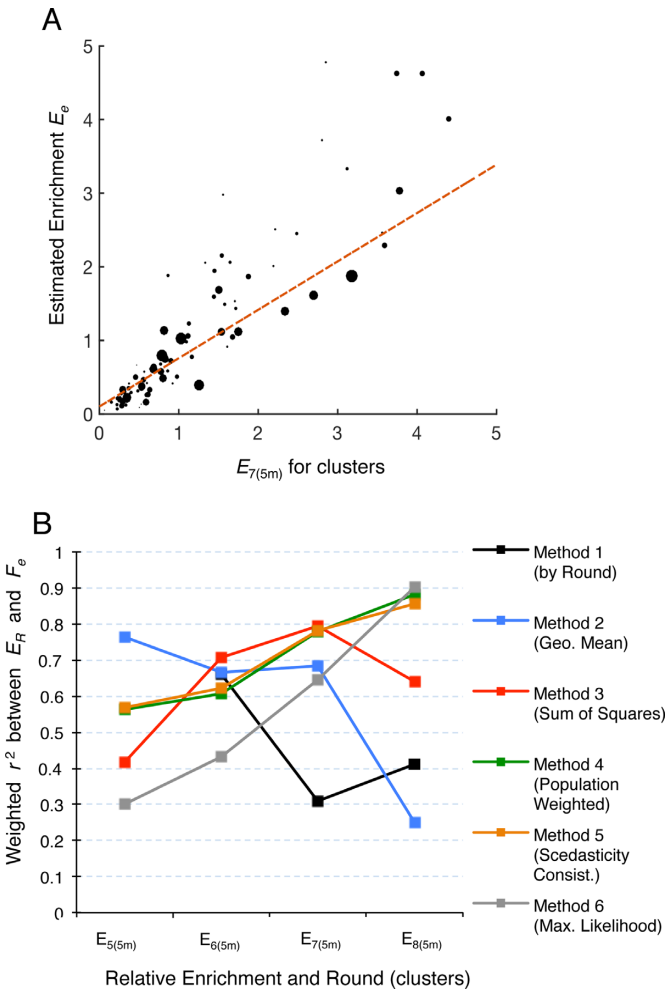

Figure 2. Estimation of evolutionary fitness across the selection. (**A**) E_e (Method 4) correlates with observed enrichment of clusters in the 5m branch at Round 7 (correlation trend line shown with weighted $r^2 = 0.77$; similar comparisons occur in other rounds); E_e was initially scaled to the same range as E_7 (as described in Materials and Methods). The area of each dot is proportional to the relative abundance of the cluster. (**B**) Comparison of different methods of fitness estimation using the weighted r^2 coefficient between estimated fitness of clusters (F_e, using Methods 1–6) and observed enrichment (E_R) at Rounds 5–8 (5m); r^2 correlation does not depend on scaling or normalization.

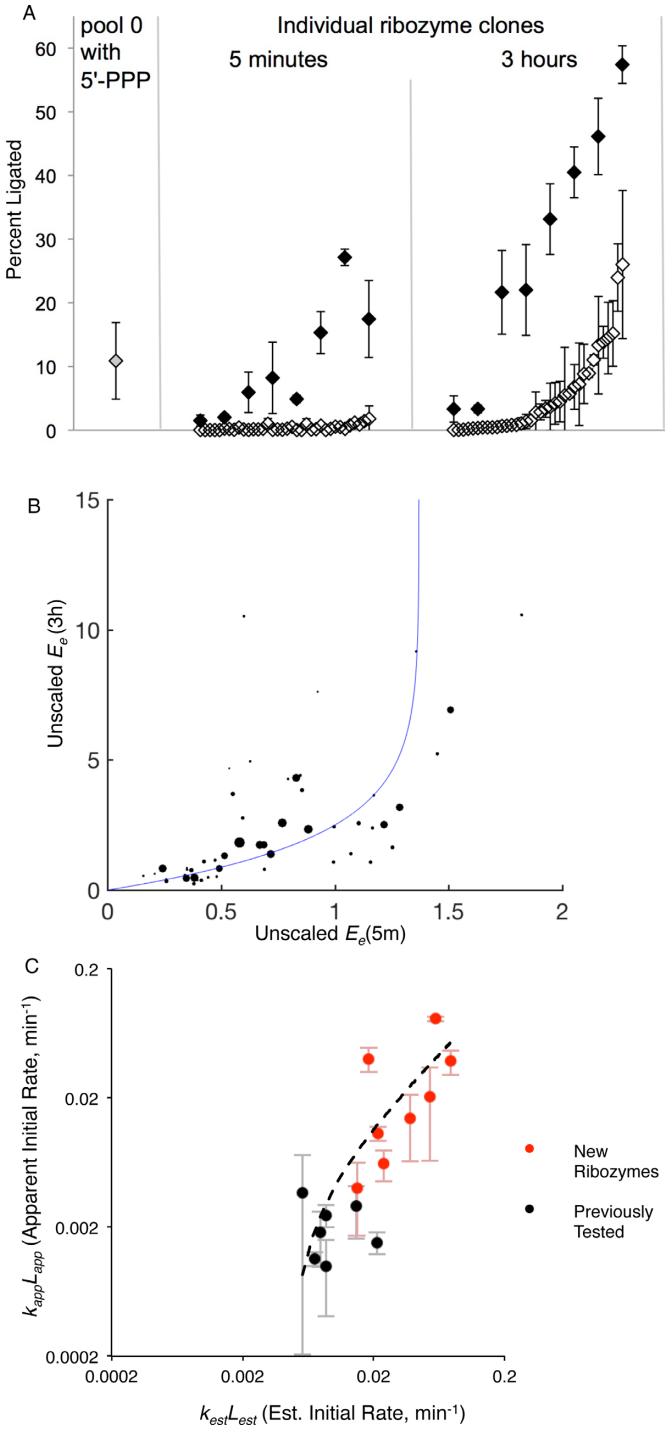

Figure 3. HTS analysis used to identify high activity ribozymes and predict ribozyme activity from evolutionary fitness**. (A)** Triphosphorylation and ligation rates for radiolabeled RNA. Left section (gray diamond) indicates the activity of a control RNA transcribed carrying a 5′-triphosphate; ligation higher than this shows adaptation for increased ligation efficiency, corresponding to higher L. Other sections show activity of individual sequences, assayed after 5 min or 3 h; white diamonds denote clones chosen previously (27), and black diamonds denote sequences identified by HTS data. **(B)** Comparison of fitness estimated from the 5m and 3h branches (weighted $r^2 = 0.87$), fit by setting k equal at two time points; the fitted parameters provide scaling constants S_e(5m) and S_e(3h). The area of each dot is proportional to relative abundance of each cluster. **(C)** Comparison of $k_{est}L_{est}$ estimated from HTS fitness (Method 4, split clusters along 5m branch), with $k_{app}L_{app}$ observed from isolated sequences, for points described in Table 2. Black points correspond to previously identified and tested sequences; red points correspond to eight sequences found to have high fitness and tested biochemically in the present study. Observed values of $k_{app}L_{app}$ were obtained in triplicate, with error bars corresponding to standard deviation. The error ranges for $k_{est}L_{est}$ for individual sequences are expected to be on the order of ±50% (Supplementary Text S2, Figure S11). Overall, these points (with the linear trend line shown as a dotted black line) show an r^2 correlation of 0.52.

dividual sequences (Supplementary Figure S8). We therefore focused further analysis primarily on clusters. Enrichment ratios using clusters split according to notable mutations showed a greater consistency across rounds compared to families (Supplementary Figure S8), so clusters were used in further analysis, with cluster enrichment and fitness used to predict the fitness and activity of their center sequences. The change of cluster fitness can be visualized over several rounds (Video S1). Overall, Methods 4 and 5 (certainty-weighted linear combination of E_R, with or without scedasticity correction) showed the highest correlation across rounds of selection (Figure 2B). As Method 5 adds an extra level of complexity compared to Method 4, Method 4 was chosen as the best method of fitness prediction in this study, and used in further calculations.

Identification of new highly active ribozymes

We sought to determine whether the HTS fitness analysis could identify higher activity ribozymes than those previously identified through the arbitrary sampling and Sanger sequencing of 36 clones from Rounds 5 and 8. We chose eight sequences with high F_e and high prediction confidence (see Supplementary Table S1 for details), and tested their activities by reaction with Tmp and ligation of the triphosphorylated ribozyme by a ligase ribozyme, thus mimicking the conditions of the selection procedure (27). Indeed, these sequences reached considerably higher experimental activity than the ribozymes previously identified from the same selection (Figure 3A; Table 2). Six of the eight high-F_e sequences showed activity greater than or equal to the best previously identified ribozymes, by a factor of up to 10–20. Overall, this approach identified ribozymes with substantially greater activity while testing fewer individual sequences (8 tested sequences versus 36 tested previously).

The 5m and 3h selection branches gave consistent estimates of fitness (Figure 3B), which fits the relationship expected according to first-order kinetics (Text S1 (Equation 7)). The fit of the data to this equation yields the parameters S_e(5m) and S_e(3h), which are needed to convert E_e into F_e, which is directly tied to kinetic parameters (see Methods and Text S1). This provides an *absolute* prediction of ribozyme activities from analysis based solely on evolutionary dynamics. The predicted activity (kL) of most high-fitness sequences was found to be within a factor of 3 of the values determined by experimentation on isolated sequences, with overall correlation ($r^2 = 0.52$) across all tested sequences (Figure 3C). The estimation errors in both HTS-derived prediction and empirical measurement likely limit the observable correlation (Supplementary Figure S9). The

Table 2. Ribozyme activity assayed experimentally, comparing newly identified ribozymes with the best previously identified (27)

Sequence Name	Initial rate estimated by HTS ($k_{est}L_{est}$) ($min^{-1}M^{-1}$)	R8(5m) Abundance	Initial rate measured experimentally ($k_{app}L_{app}$) ($min^{-1}M^{-1}$)	Previously Identified?
1-S	0.476 ± 0.238	0.1581	0.124 ± 0.033	No
2-S	0.761 ± 0.381	0.0449	0.278 ± 0.149	No
3-S	0.369 ± 0.185	0.0229	0.807 ± 0.171	No
4-S	0.431 ± 0.215	0.0435	0.212 ± 0.026	No
6-S	1.547 ± 0.774	0.0497	0.770 ± 0.166	No
8	0.298 ± 0.149	0.0759	0.080 ± 0.46	No
11-S	1.192 ± 0.596	0.0125	1.638 ± 0.078	No
22-S	1.072 ± 0.536	0.0060	0.412 ± 0.281	No
R5_3C21	0.115 ± 0.058	0.0060	0.074 ± 0.083	Yes
R8_35C18A	0.143 ± 0.072	0.0097	0.023 ± 0.003	Yes
R8_35C18B	0.173 ± 0.087	0.0130	0.020 ± 0.001	Yes
R8_55C10	0.423 ± 0.212	0.0102	0.030 ± 0.001	Yes
R8_35C10	0.173 ± 0.087	0.0130	0.050 ± 0.001	Yes
R8_55C18	0.158 ± 0.079	0.0030	0.036 ± 0.001	Yes
R8_35C16	0.295 ± 0.148	0.0457	0.025 ± 0.001	Yes

Estimated and measured values were calculated as described in Supplementary Text. Errors given are ± 1 standard deviation, as described in Methods and Supplementary Text, with the standard deviation of k_{est} calculated from cluster abundance as a scale-variant error of ±50% for cluster enrichment.

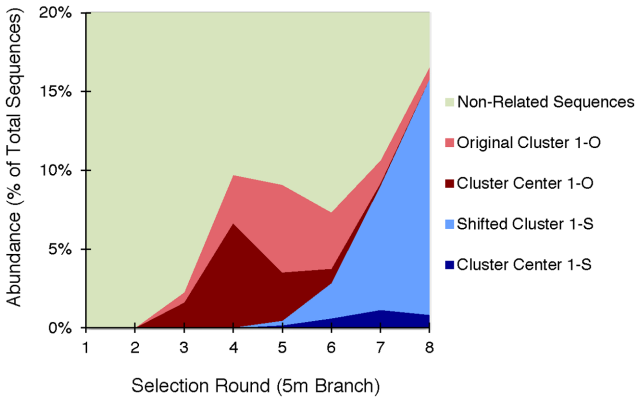

Figure 4. Area plot of sequence distribution over successive rounds. The highest-count family of Round 8 (5m) (Family 1, with original center '1-O') begins the selection as a single center sequence at low count, with a cloud of similar sequences appearing in subsequent rounds (red). At R4, a strong beneficial mutation (referred to as '1-S') appears, gradually becoming surrounded by a cluster of similar mutants (blue) after mutagenic PCR is introduced at R5 (see Supplementary Figure S10 for more details).

slope of the line of best fit was 0.73, close to the value of 1 expected from perfect prediction of absolute rates.

Beneficial mutations within clusters

Most of the clusters of highest estimated fitness carried notable mutations, in that they contained sequences that outcompeted the original highest-count sequence in the family between Rounds 4 and 8. In the more stringent 5m selection branch, 34 out of the 59 highest-abundance peaks at Round 8 displayed at least one notable mutation from the central sequence of Round 4, such that a large portion of the pool consisted of sequences similar to these mutants. In such these cases, the notable mutation appeared to demonstrate significantly increased survival fitness, with the new cluster rapidly enriching to outpace the old sequence center, with one such sweep shown in Figure 4. For clusters with high abundance at Round 4, notable mutations (that would dominate in later rounds) typically each accounted for less than 1% of the cluster population at Round 4; thus, out-competing the original center over the next four rounds of selection would require a mutation with at least $100^{1/4}$ times (∼3×) the fitness of the original center. To determine the effect of notable mutations on ribozyme activity, four sequences (1-S, 2-S, 6-S, 11-S) that clustered with previously tested clones, but also possessed notable mutations, were among those assayed experimentally. These sequences were independently chosen on the basis of high fitness. Three of the four mutants exhibited higher activity compared to the best previously identified clone from the same cluster (up to a five-fold increase), indicating that the notable mutations were usually beneficial. One sequence (2-S) showed a five-fold increase in activity despite a difference of only a single nucleotide (Supplementary Table S2).

The observed bulk mutation rate during error-prone PCR (following Rounds 4–8) was ∼1.7 mutations per sequence, on average, or a fidelity q of 0.989 per base, which is consistent with values from the original protocol (28). This gives an estimated 18% (q^{150}) of sequences surviving free of mutations each round, such that the center sequence of a cluster is expected to enrich at 18% of the rate of its cluster if all mutations are neutral. For the top 20 clusters in the pool, a single round of mutagenic PCR is expected to generate at least one copy of all possible single mutants, most double mutants, and a substantial fraction of triple mutants (though any individual sequence would appear at a much lower count than the center sequence). HTS analysis showed cluster center sequences enriching at approximately half the rate of their overall clusters (Supplementary Figure S10). This implies that cluster centers have ∼2.8-fold greater fitness than the average of their mutants within a given peak (50% versus 18%).

Ideality of *in vitro* evolution behavior

The observation that enrichment varies from round to round, leading to inexact estimates of fitness, prompted further study of the noise inherent to the selection. Error

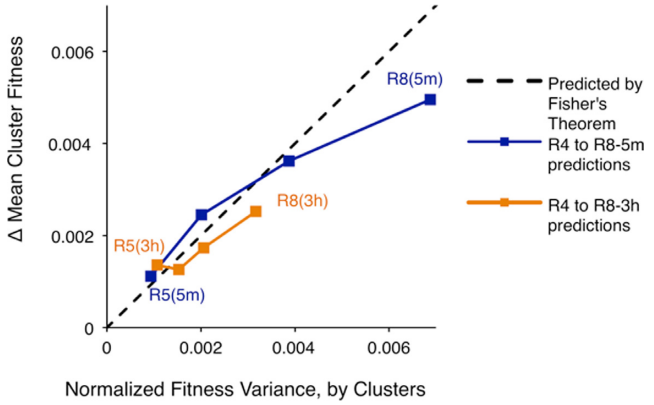

Figure 5. The ribozyme selection follows Fisher's Fundamental Theorem of Natural Selection (FFTNS). Clusters were analyzed for variance of fitness in each round and for the change in mean $\bar{F}_e$ (using Method 4). The dotted line ($y = x$) represents behavior expected by FFTNS. Data are shown for Round 4 and later, since earlier rounds were too heterogeneous to support fitness calculation given the depth of sequencing performed.

distributions were calculated from the normalized difference between E_e and E_R. Analysis suggested that lower-abundance sequences enriched with greater noise, but enrichment noise did not drop below a certain proportional threshold for high-abundance sequences. This suggests that abundance-dependent noise (e.g. genetic drift) dominates in early rounds of selection and abundance-independent noise or error (e.g. experimental variations) has a greater effect on the enrichment of high-abundance sequences in later rounds of selection. For lower-abundance clusters, scale-dependent noise impacted individual sequences more than clusters, suggesting that cluster enrichment might be a better predictor of the fitness of individual sequences (compared to observed enrichment of individual sequences) (Supplementary Text S2, Figure S11).

To evaluate whether the evolution experiment would be suitable for retrospective analysis (described below), we measured the extent to which the selection as a whole followed ideal behavior as predicted by Fisher's Fundamental Theorem of Natural Selection (FFTNS) (33,34,36). In general terms, FFTNS states that, over a round of natural (or artificial) selection (assuming that each allele's fitness does not change), the change in average fitness of a population should equal the variance of sequence fitness in the population (normalized by mean fitness). Obedience to FFTNS by an evolving population implies that the evolutionary dynamics are well-behaved and governed by rules of natural selection. Indeed, the evolutionary dynamics were consistent with FFTNS for both clusters and individual sequences (Figure 5, Supplementary Figure S12A), suggesting that selection is the primary factor driving changes in the estimated fitness distribution of the pool. This concurrence indicates that that the selection behaved predictably and confirms that fitness mean and variance were accurately estimated. Sequence clusters followed FFTNS more closely than individual sequences, consistent with our earlier observation that a cluster-based analysis is subject to less noise.

As expected, estimated fitness $\bar{F}_e$ increased during the selection, with the higher stringency 5m selection branch resulting in approximately 2-fold greater $\bar{F}_e$ than the 3h branch by Rounds 7–8. Interestingly, the variance of fitness also increased over time in both selection branches. Intuitively, this increase is expected during a selection from random sequence space, as the distribution of fitness is initially sharply centered near zero, and then spreads to include higher fitness values. To quantify this effect, we derived a corollary to FFTNS, that the change in fitness variance between rounds is expected to equal the mean-scaled skewness of the fitness distribution minus the change in average fitness squared: $(\sigma^2{}_{R+1} - \sigma^2{}_R) = E(F_R - \bar{F}_R)^3/\bar{F}_R - (\bar{F}_{R+1} - \bar{F}_R)^2$ (Text S1). The fit of the data to this corollary reflects whether the shape of the fitness distribution, as captured by the first through third moments, obeys expected dynamics. As skewness is a higher-order shape factor than mean or variance, this relation is expected to be more sensitive to noise or inaccuracies in the estimated shape of the fitness distribution. Clusters followed this corollary well, although individual sequences did not (Supplementary Figure S12). Therefore, F_e based on clusters gives a reasonably accurate estimate of skewness, and the shape of the fitness distribution based on clusters behaved in a predictable manner.

Calculating the underlying distribution of ribozyme activity

Our ability to estimate fitness values, and therefore ribozyme rate constants, with reasonable accuracy allows an analysis of the overall distributions of fitnesses and rate constants and their dynamics over the course of selection. Analysis focused on the distribution of these values for clusters, since the fitness of individual sequences was subject to a larger degree of noise and was less consistent with expected dynamics. A fitness distribution, $P_R(F_e)$, was calculated from the HTS data for Rounds 4–8. As expected, $P_R(F_e)$ shifts toward the right as the selection progresses (Figure 6A), as sequences with high F_e increase in abundance while those with low F_e decrease.

To infer the underlying fitness distribution of the initial pool ($P_0(F)$), 3h selection data were used, as the 3h selection experienced identical selection conditions throughout Rounds 1–8. While the scarcity of conserved sequences and clusters prior to Round 4 prevented direct measurement of their distributions from sequencing data, FFTNS analysis indicated that the selection exhibited predictable behavior. We therefore employed a retrospective analysis to infer the $P_R(F)$ of early rounds. As F represents a sequence's survivability before amplification, the selection process essentially multiplies a fitness distribution $P_{R-1}(F)$ by F to produce $P_R(F)$ (with appropriate normalization; see Materials and Methods). Beginning at Round 4, we divided $P_R(F_e)$ by F_e (and renormalized) to obtain the estimated distribution $P_{R-1}(F)$, until $P_0(F)$ (the initial pool) was obtained. The validity of this process was verified by the observation that $P_R(F_e)/P_{R-1}(F_e)$ was linearly related to F_e (for later rounds in which this comparison could be made) (Supplementary Figure S13), in addition to the concordance of the experiment with FFTNS.

The qualitative and quantitative shape of the distribution of catalytic activities, not only fitnesses, in a pool of random RNA is of fundamental interest. We therefore translated F_e

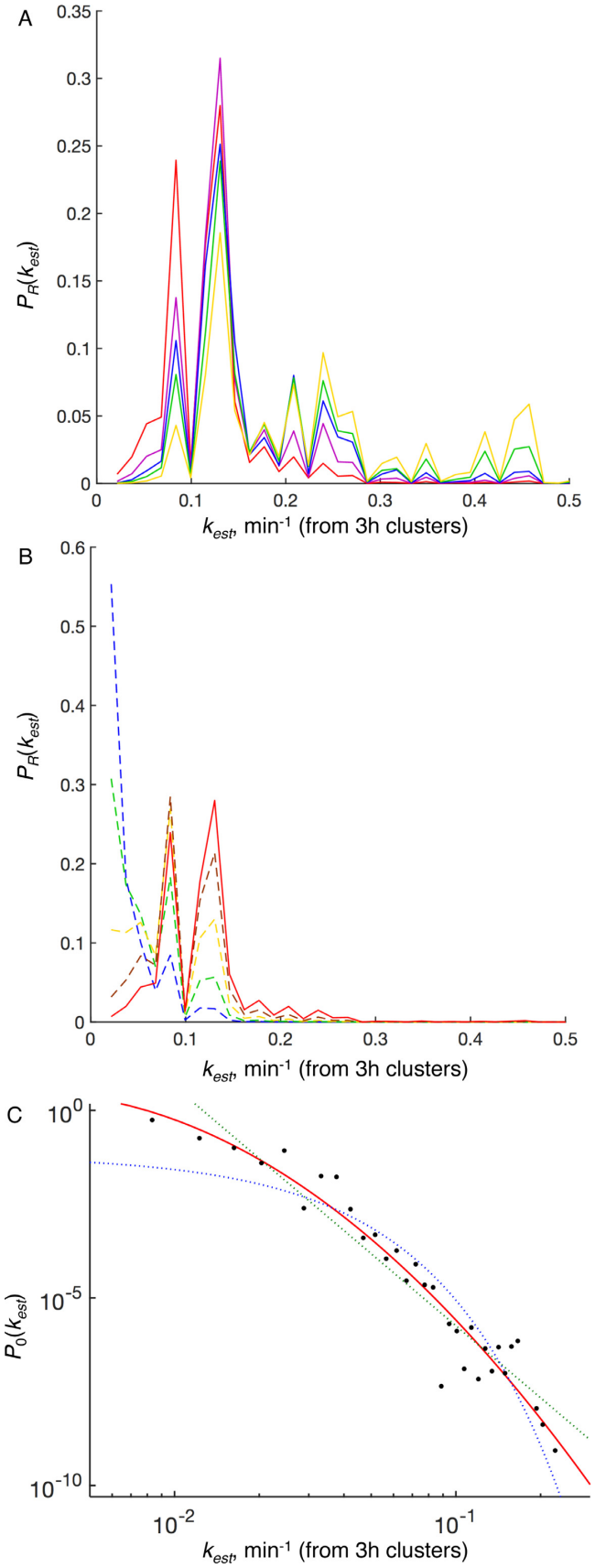

Figure 6. The distributions of fitness and rate constants in a pool of RNA. (**A**) Distribution of estimated rate constants (k_{est}) for sequence clusters over multiple rounds of selection; lines correspond to distributions measured from HTS data of Rounds 4–8 in the 3h selection branch (Yellow: Round 8; Green: Round 7; Blue: Round 6; Purple: Round 5; Red: Round 4). (**B**) Inferred distribution of k_{est} for sequence clusters in earlier rounds of selection; solid red line corresponds to Round 4 distribution, while dashed lines represent the retrospectively inferred distributions for Rounds 0–3 (Brown: Round 3; Yellow: Round 2; Green: Round 1; Blue: Round 0, i.e. initial pool). (**C**) The high-activity (right-sided) tail of the initial distribution of rate constants (k) is fit by a log-normal distribution (Red line: log-normal distribution $y = \frac{1}{\sigma k\sqrt{2\pi}}\, exp[-\frac{(ln\ k-\mu)^2}{2\sigma^2}]$ with $\sigma = 0.665$, $\mu = -5.375$, pdf scaling factor = 54.3, and nonlinear $r^2 = 0.933$; dotted green line corresponds to fit scale-free distribution $y = \frac{6.38*[27.9]^{6.38}}{k^{7.38}}$, with pdf scaling factor = 0.00807 and nonlinear $r^2 = 0.910$; dotted blue line corresponds to fit exponential distribution $y = 89.08e^{-89.08k}$, with pdf scaling factor = 1390 and nonlinear $r^2 = 0.859$.

(an absolute metric) into absolute rate constants by assuming first-order kinetics with an average L value across the population, as described in the Methods. The approriateness of this translation was suggested by the correspondence of absolute rate constants estimated by HTS data and rate constants measured on isolated sequences of RNA (Figure 3C). Our retrospective analysis applied to rate constants showed that the initial distribution of both fitness values and rate constants had a tall peak near zero, as expected for random RNA (Figure 6). Since the right-sided 'tail' of the distribution would correspond to functional ribozymes, we examined this region more closely. This tail was fit to three possible distributions: log-normal, exponential, and scale-free (Pareto). The log-normal distribution fit well to the ribozyme tail for both fitness and rate constant, with a greater correlation coefficient than the scale-free and exponential distributions, whose fits also showed non-random patterns of residuals (Figure 6C; Supplementary Figure S14). Simulations verified that stochastic noise in abundance data would not interfere with the retrospective analysis to identify the distribution of the initial pool (Supplementary Figure S15).

Our method of translating fitness to rate constants had made the approximation of a constant L (maximum extent of triphosphorylation and ligation), i.e. that variations in k (rather than L) dominate variations in fitness. At short times (e.g. 5 min), the rate of a first order reaction is approximately Lk, and we would expect that the amount of product conversion $F \approx Lkt$. If k is dominant, then k should correlate with F; conversely, if L is dominant, then L should correlate with F. We found that k correlates well with F_e ($r^2 = 0.79$) while L does not ($r^2 = 0.03$), supporting our approximation (Supplementary Figure S16A and B). In addition, we found that k and L were uncorrelated, suggesting that these kinetic parameters vary independently for ribozymes (Supplementary Figure S16C).

DISCUSSION

The origin of a proposed RNA World depended on the emergence of relatively rare, functional sequences from abiotic synthesis. A critical, but poorly understood, aspect of this emergence (and other molecular selection processes) is the distribution of activity in random sequence space. We used HTS data to follow the evolutionary dynamics of an

in vitro selection for ribozymes capable of triphosphorylation, and we thereby inferred the distribution of fitness and the underlying distribution of catalytic activity in the random pool of RNA.

This ribozyme selection began with a random pool (N150), whose effective complexity (1.7×10^{14} starting sequences from a theoretically possible set of 2.0×10^{90}) far exceeded the capacity of HTS. The selection thus illuminates many random 'pinpoints' in sequence space (37), and the high complexity of the pool allowed essentially unambiguous assignment of sequences to families. Although the complexity of this pool prevents detailed mapping of the complete fitness landscape (27), our selection yielded hundreds of high-fitness sequence clusters, which provided data suitable for generating a potential overall probability distribution of fitness. In this selection, RNA fitness is a combination of catalytic activity for the triphosphorylation reaction as well as suitability as a substrate for ligation. Catalytic activity is reflected by the rate constant (k), while both catalytic and ligation substrate activity are reflected in the total extent of reaction (L). In practice, we found that fitness correlated well with k_{est} but not with L_{est}, suggesting that catalytic activity was the most important determinant of fitness in this selection (Supplementary Figure S16).

Although sequence abundance is a convenient, and frequently used, metric for fitness (23,38,39), abundance is subject to unpredictable effects that distort a direct relationship to ribozyme activity (Text S3). Fitness estimated from enrichment over multiple rounds indeed improved the fitness metric. We attempted multiple methods to estimate true fitness from enrichment data, including the previously-reported geometric mean (36), as well as maximum likelihood estimation, and variations of least-squares methods. The best metric appeared to be a modified least-squares regression (Method 4), in which enrichment observations were weighted by the absolute sequence abundance in each round for each sequence, essentially giving a larger 'vote' to the rounds having greater observational certainty.

We identified eight high-fitness sequences that had not been previously identified or tested during a conventional analysis (in which 36 clones were assayed for activity) (27). Our analysis proved to be powerful, in that six of the eight clones had greater activity than the most active ribozyme previously found, improving materially upon conventional analysis (Text S4). As noted in other contexts (e.g. (23), HTS also readily identified individual beneficial mutations ('notable' mutations), some of which were verified experimentally.

The underlying distribution of fitness, and the accompanying distribution of ribozyme activity, across populations of random RNA are subjects of special interest for understanding both the RNA World hypothesis of early life and the practical *in vitro* evolution of ribozymes. However, little is known empirically about these—or any—evolutionary fitness landscapes over large nucleotide sequence spaces. An experimentally tractable pool of random RNA sequences cannot sample its whole sequence space if the random region is longer than ~28 nucleotides. However, in principle, the underlying distribution of fitness, i.e. the probability distribution function, can still be inferred given an incomplete sample of sufficient size, which our analysis of hundreds of individual sequence families likely provided. We found that changes in fitness distribution during *in vitro* evolution followed Fisher's Fundamental Theorem of Natural Selection (FFTNS) and a corollary of FFTNS describing higher moments of the distribution (Text S5). We thus verified that our selection was well-behaved and predictable in the forward direction.

A well-ordered selection propagates its fitness distribution through each round according to an applied survival function. The inverse function could in principle be applied retrospectively to later rounds, giving an inferred fitness distribution of the preceding rounds. We found a linear relationship between the fitness distributions of consecutive rounds, as expected, confirming the validity of this approach. The inferred fitness distribution of the random pool of RNA had a large component near zero, as expected, and a high-fitness tail that represents sequences of particular interest. We converted the fitness distribution of the right-sided tail into a distribution of absolute ribozyme rates, scaled using a comparison of sequence fitness between the two selection branches. One may expect that accurate prediction of activity for each ribozyme cluster is likely to be challenging, due to experimental variations during selection and measurement of ribozyme kinetics, as well as variations in ligase efficiency. The accuracy of our conversion was evaluated for several individual ribozymes by comparing rate constants predicted from HTS data with rate constants measured biochemically, resulting in prediction of real values to within a factor of around 3 ($r^2 = 0.52$ for overall correlation). In addition, the conversion makes two simplifying assumptions, namely assuming a pseudo-first-order reaction, and neglecting variations in maximum activity (L), which may be influenced by multiple factors (e.g. ligation efficiency and folding stability) (40) (Supplementary Figure S17). The assumption of pseudo-first-order kinetics is consistent with earlier work in which eight isolated ribozymes were found to have kinetics that fit well to this model (27). The assumption of constant L is consistent with the *post hoc* observation that L is essentially uncorrelated with F_e (in contrast, k correlates well with F_e); while individual sequences' L values are expected to vary, we assume its average will hold roughly constant over the larger overall fitness distribution. It is important to note that the rate constants were predicted from fitness on an *absolute* scale (set by the reaction times of the selection branches), with no free parameters needed to achieve workable agreement with experimental numbers.

No previous work has measured or approximated the distribution of a catalytic activity over random molecular space of this scale; here, analysis of HTS data allowed for the conversion of fitness information into kinetic parameters. Despite experimental and theoretical interest, there is little consensus in the literature on the nature and shape of any such distribution for any ribozyme function. The affinity of aptamers has been posited to be log-normally distributed, based on a model and experimental data for dsDNA-protein interactions in which individual base pairs contribute independently to overall binding energy (9,10,41). However, energetic contributions that are correlated along the sequence, such as from DNA bending, could alter this distribution (42). In the case of RNA fold-

ing, a theoretical model suggests that the activation energies of melting follow a Gaussian distribution (11), while folding simulations suggest that the distribution of minimum free energies for random RNAs is non-Gaussian (43). In microbial populations, the distributions of fitness effects from new mutations have been fit to a variety of distributions, including normal and exponential (44). In addition, extreme value theory indicates that a high-value tail can be approximated as a Pareto (scale-free) distribution if the value is comprised of independent random variables, a trend observed in previous comparisons of pool size and activity (8). We attempted to fit the empirically derived ribozyme rate constant distribution to log-normal, exponential, and scale-free distributions.

We found that the inferred distribution of ribozyme rates in a random pool fit well to a log-normal distribution. A log-normal distribution of catalytic constants k for ribozymes across sequence space could indicate a normal distribution of the corresponding activation energies. Although the observed distribution cannot be quantitatively translated into activation energies without knowledge of the Arrhenius pre-exponential factor (A), we note that the probability density drops precipitously as rate increases, such that high-fitness ribozymes occur in the population as extremely rare events. A normal distribution of $\log(k)$ implies a normal distribution of $E_a/RT + \log(A)$. If we posit that A is similar for most ribozymes in the population, the standard deviation of $\log(k)$ should equal the standard deviation of E_a/RT, such that our calculated $\sigma_{\log(k)} = 0.665$ corresponds to $\sigma_{\mathrm{Ea}} = 1.6$ kJ/mol. This distribution of activation energies is surprisingly steep; under such a distribution, a ribozyme cluster with 100-fold higher activity than the mean would occur only once in a pool of 10^{11} sequences. The steep drop-off of the distribution can be put in terms of the expected activity of the best ribozyme in a pool of a given size (Supplementary Figure S18), which shows that even very large increases in pool complexity result in relatively small gains in activity. While various possible initial distributions of ribozyme activity have been proposed, the fit of rate constants to a single log-normal distribution (and the fit of activation energies to a normal distribution) suggest that the ribozymes, despite the large heterogeneity of sequence, share an underlying mechanism for the emergence of function. One possible interpretation is that a normal distribution of activation energies reflects the energetic contributions of many independent interactions with finite variance, with the ribozymes each using a similar number of interactions. The apparent independence of small contributions could be tested by combining mutations (45). HTS could be used to expand and systematize this approach.

A few caveats should be considered regarding the use of HTS data to infer the distribution of rate constants. First, while sequence clusters exhibited behavior that was consistent with evolutionary theory, individual sequences were less well-behaved. The effect of this limitation can be seen in the somewhat imperfect correlation between kinetics estimated from HTS and kinetics determined biochemically for individual sequences. This limits the retrospective analysis to sequence clusters, and the relationship of $P_R(F)/P_{R\text{-}1}(F)$ to F should be used to verify the validity of such an analysis. Given that an earlier effort analyzing a mutational library of the Tetrahymena group I ribozyme did not find a good match to FFTNS (46), the success in the present instance may be due to a combination of the data-rich, robust metric for cluster fitness, the relatively large variation in fitness, and the low background of the experimental selection, for which all tested sequences have exhibited catalytic activity (i.e. low false positive rate). Second, any conversion of fitness to rate constants would rely on assumptions about the kinetic model. In this case, the assumption of constant L is known to be a simplification. L is presumably influenced by RNA folding and activity as a ligase substrate, in addition to the triphosphorylation reaction. While it may be justified as discussed above, based on the ribozymes observable in later rounds, it is possible that the statistical properties of L differ for lower activity ribozymes. In that case, it would not be possible to deconvolve the effect of k and L on fitness, although F would still have biochemical meaning as the fraction reacted. Another assumption made here is that RT-PCR amplification was not a substantial influence on fitness; the plausibility of this assumption should be considered as a source of error. With respect to the distribution itself, it should be noted that there is presumably an upper limit to activity that truncates any probability distribution function in reality. That is, increasingly precise arrangements of nucleotides at the active site, and correspondingly higher catalytic rates, are presumably limited by the RNA's ability to arrange around a catalytic site. Structural and/or chemical limits in RNA would provide an upper limit for the possible catalytic rate enhancement, affecting the distribution at very high activity. Finally, it is unknown whether conclusions drawn from *in vitro* evolution for triphosphorylation ribozymes would apply to other selections (Text S6).

In summary, an evolutionary analysis of a ribozyme selection for triphosphorylation activity identified improved ribozymes as well as crucial beneficial mutations, and thus this procedure may be useful for identifying the highest activity ribozymes (or aptamers or protein enzymes) in a selection pool with minimal experimental testing. In addition, estimated ribozyme fitness across the selection largely obeyed evolutionary dynamics as expected by Fisher's theorem. Here the translation relied on normalization using the two selection branches (5m and 3h), but normalization could also be achieved using varying substrate concentrations or knowledge of the amplification required between rounds. A retrospective analysis allowed inference of the distribution of catalytic activities in the initial pool of random RNA, providing, for the first time, a look at the underlying distribution of rate constants and activation energies for an *in vitro* selection. The analysis suggests a Gaussian distribution of activation energies, perhaps reflecting the summation of many independent energetic contributions. The surprisingly steep drop-off in the frequency of high-activity ribozymes, and the accompanying flatness of the expected maximum k vs. complexity curve (Supplementary Figure S18), suggests that the ribozyme activity level is largely determined by the nature of the function, not the complexity or diversity of the library. For some functions, a relatively low complexity pool of RNA may thus possess ribozymes of biochemically significant activity, suggesting that the emergence of such functions may not be uncommon in an RNA world.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Shreyas Athavale for submission of samples for sequencing; Tetsuya Yomo and Ramon Xulvi-Brunet for discussions. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

FUNDING

The Center for Scientific Computing from the CNSI, as well as MRL: an NSF MRSEC [DMR-1121053] and NSF [CNS-0960316] provided computational resources used in this project; Sequencing was performed by the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant [1S10OD010786-01]; Simons Foundation [290356 to I.A.C.]; NASA [NNX16AJ32G]; Searle Scholars Program; Hellman Faculty Fellows Program; Institute for Collaborative Biotechnologies [W911NF-09-0001] from the U.S. Army Research Office; National Aeronautics and Space Administration [NNX13AJ09G to J.E.M and U.F.M.] issued through the Science Mission Directorate (ROSES-2011), in Astrobiology/Exobiology (to U.F.M.). Funding for open access charge: Simons Foundation (grant no. 290356 to IAC).
Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. and Szostak,J.W. (1993) Isolation of new ribozymes from a large pool of random sequences [see comment]. *Science*, **261**, 1411–1418.
2. Blain,J.C. and Szostak,J.W. (2014) Progress toward synthetic cells. *Annu. Rev. Biochem.*, **83**, 615–640.
3. Klussmann,S. (2006) *The Aptamer Handbook: Functional Oligonucleotides and their Applications*. John Wiley & Sons.
4. Pitt,J.N. and Ferre-D'Amare,A.R. (2010) Rapid construction of empirical RNA fitness landscapes. *Science*, **330**, 376–379.
5. Pressman,A., Blanco,C. and Chen,I.A. (2015) The RNA World as a model system to study the origin of life. *Curr. Biol.*, **25**, R953–R963.
6. Wang,J., Rudzinski,J.F., Gong,Q., Soh,H.T. and Atzberger,P.J. (2012) Influence of target concentration and background binding on in vitro selection of affinity reagents. *PLoS One*, **7**, e43940.
7. Joyce,P., Rokyta,D.R., Beisel,C.J. and Orr,H.A. (2008) A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics*, **180**, 1627–1643.
8. 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.
9. Vant-Hull,B., Gold,L. and Zichi,D.A. (2000) Theoretical principles of in vitro selection using combinatorial nucleic acid libraries. *Curr. Protoc. Nucleic Acid Chem.*, doi:10.1002/0471142700.nc0901s00.
10. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–743.
11. Tacker,M., Fontana,W., Stadler,P.F. and Schuster,P. (1994) Statistics of RNA melting kinetics. *Eur. Biophys. J.*, **23**, 29–38.
12. Kobori,S., Nomura,Y., Miu,A. and Yokobayashi,Y. (2015) High-throughput assay and engineering of self-cleaving ribozymes by sequencing. *Nucleic Acids Res.*, **43**, e85.
13. Wu,N.C., Dai,L., Olson,C.A., Lloyd-Smith,J.O. and Sun,R. (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, **5**, e16965.
14. Jalali-Yazdi,F., Lai,L.H., Takahashi,T.T. and Roberts,R.W. (2016) High-throughput measurement of binding kinetics by mRNA display and next-generation sequencing. *Angew. Chem. Int. Ed. Engl.*, **55**, 4007–4010.
15. Fischer,N.O., Tok,J.B. and Tarasow,T.M. (2008) Massively parallel interrogation of aptamer sequence, structure and function. *PLoS One*, **3**, e2720.
16. Rowe,W., Platt,M., Wedge,D.C., Day,P.J., Kell,D.B. and Knowles,J. (2010) Analysis of a complete DNA-protein affinity landscape. *J. R. Soc. Interface*, **7**, 397–408.
17. Kobori,S. and Yokobayashi,Y. (2016) High-throughput mutational analysis of a twister ribozyme. *Angew. Chem.*, **55**, 10354–10357.
18. de Visser,J.A. and Krug,J. (2014) Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.*, **15**, 480–490.
19. Ameta,S., Winz,M.-L., Previti,C. and Jäschke,A. (2014) Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Res.*, **42**, 1303–1310.
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. Hayden,E.J., Ferrada,E. and Wagner,A. (2011) Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, **474**, 92–95.
23. Hoinka,J., Berezhnoy,A., Dao,P., Sauna,Z.E., Gilboa,E. and Przytycka,T.M. (2015) Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res.*, **43**, 5699–5707.
24. Thiel,W.H., Bair,T., Peek,A.S., Liu,X., Dassie,J., Stockdale,K.R., Behlke,M.A., Miller,F.J. Jr and Giangrande,P.H. (2012) Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection. *PLoS One*, **7**, e43836.
25. Schütze,T., Wilhelm,B., Greiner,N., Braun,H., Peter,F., Mörl,M., Erdmann,V.A., Lehrach,H., Konthur,Z. and Menger,M. (2011) Probing the SELEX process with next-generation sequencing. *PLoS One*, **6**, e29604.
26. Hayden,E.J., Bratulic,S., Koenig,I., Ferrada,E. and Wagner,A. (2014) The effects of stabilizing and directional selection on phenotypic and genotypic variation in a population of RNA enzymes. *J. Mol. Evol.*, **78**, 101–108.
27. Moretti,J.E. and Muller,U.F. (2014) A ribozyme that triphosphorylates RNA 5′-hydroxyl groups. *Nucleic Acids Res.*, **42**, 4767–4778.
28. Cadwell,R.C. and Joyce,G.F. (1994) Mutagenic PCR. *Genome Res.*, **3**, S136–S140.
29. Cadwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *Genome Res.*, **2**, 28–33.
30. Rogers,J. and Joyce,G.F. (2001) The effect of cytidine on the structure and function of an RNA ligase ribozyme. *RNA*, **7**, 395–404.
31. Xulvi-Brunet,R., Campbell,G.W., Rajamani,S., Jiménez,J.I. and Chen,I.A. (2016) Computational analysis of fitness landscapes and evolutionary networks from in vitro evolution experiments. *Methods*, **106**, 86–96.
32. Gustafsson,L. (1986) Lifetime reproductive success and heritability: empirical support for Fisher's fundamental theorem. *Am. Naturalist*, **128**, 761–764.
33. Fisher,R.A. (1930) *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford University Press.
34. Lessard,S. (1997) Fisher's fundamental theorem of natural selection revisited. *Theor. Popul. Biol.*, **52**, 119–136.
35. Edwards,A.W. (1994) The fundamental theorem of natural selection. *Biol. Rev.*, **69**, 443–474.
36. Ichihashi,N., Aita,T., Motooka,D., Nakamura,S. and Yomo,T. (2015) Periodic pattern of genetic and fitness diversity during evolution of an artificial cell-like system. *Mol. Biol. Evol.*, msv189.
37. Athavale,S.S., Spicer,B. and Chen,I.A. (2014) Experimental fitness landscapes to understand the molecular evolution of RNA-based life. *Curr. Opin. Chem. Biol.*, **22**, 35–39.
38. Schutze,T., Wilhelm,B., Greiner,N., Braun,H., Peter,F., Morl,M., Erdmann,V.A., Lehrach,H., Konthur,Z., Menger,M. *et al.* (2011)

Probing the SELEX process with next-generation sequencing. *PLoS One*, **6**, e29604.
39. Thiel,W.H., Bair,T., Peek,A.S., Liu,X., Dassie,J., Stockdale,K.R., Behlke,M.A., Miller,F.J. Jr and Giangrande,P.H. (2012) Rapid identification of cell-specific, internalizing RNA aptamers with bioinformatics analyses of a cell-based aptamer selection. *PLoS One*, **7**, e43836.
40. Schmitt,T. and Lehman,N. (1999) Non-unity molecular heritability demonstrated by continuous evolution in vitro. *Chem. Biol.*, **6**, 857–869.
41. Slutsky,M. and Mirny,L.A. (2004) Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.*, **87**, 4021–4035.
42. Slutsky,M., Kardar,M. and Mirny,L.A. (2004) Diffusion in correlated random potentials, with applications to DNA. *Phys. Rev. E*, **69**, 061903.
43. Wolfsheimer,S. and Hartmann,A. (2010) Minimum-free-energy distribution of RNA secondary structures: entropic and thermodynamic properties of rare events. *Phys. Rev. E*, **82**, 021902.
44. Bataillon,T. and Bailey,S.F. (2014) Effects of new mutations on fitness: insights from models and data. *Ann. N. Y. Acad. Sci.*, **1320**, 76–92.
45. Ekland,E.H., Szostak,J.W. and Bartel,D.P. (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science*, **269**, 364–370.
46. Lehman,N. and Joyce,G.F. (1993) Evolution in vitro of an RNA enzyme with altered metal dependence. *Nature*, **361**, 182–185.