Supporting Information

Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA

Abe D. Pressman, Ziwei Liu, Evan Janzen, Celia Blanco, Ulrich F. Muller, Gerald F. Joyce, Robert Pascal, and Irene A. Chen

Supporting Text S1

Two replicates of the selection (RS1 and RS2) were performed at different coverage of the initial DNA library. In RS1 and RS2, the initial DNA library contained ~5.5 x 10^{12} or ~8.8 x 10^{13} molecules (1.25x or 20x coverage), respectively. Assuming a Poisson distribution, in each replicate, ~70% or >99.99% of all possible sequences would be represented, respectively. After transcription for each round, the amount of RNA used was >40 times the number of possible sequences (4^{21}).

Supporting Text S2

The catalytic ratio k_sA_s/k_0A_0 underestimates the true rate enhancement at an internal 2'-OH site due to the ribozyme. There are 73 potential sites for aminoacylation (70 internal 2'-OH groups, the vicinal diol, and the 5'-triphosphate). If all sites had equal reactivity, the true background rate at a single site would be $k_0A_0/73$. However, the true background rate at an internal site is even lower because the background reaction is known to be dominated by reaction at the vicinal diol and 5' triphosphate [1]. For an oligonucleotide model, the rate of reaction at an internal 2'-OH was undetectable; assuming a conservative limit of detection of 10%, the rate of non-catalyzed modification at a specific internal site would be (at least) 10-fold lower than that at the vicinal diol or 5' triphosphate. This background rate is not expected to vary by much for different sequences. Thus it is likely that the true background rate at a specific internal site is $< k_0A_0/(73 \text{ x } 10)$. In other words, the true catalytic rate enhancement at a specific internal site would be at least 730 times greater than the r_s ratios reported here.

Supporting Text S3

We first consider a ribozyme reaction with BYO, neglecting BYO degradation. In analogy to enzyme kinetics, the reaction is assumed to occur through an initial binding step followed by reaction, as follows:

$$RNA\ +\ BYO\ \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}}\ RB\ \xrightarrow{k_2}\ P$$

where RB is the noncovalent complex and P is the aminoacylated RNA. The rate of formation $d[P]/dt = k_2[RB]$. The steady-state approximation of $d[RB]/dt = 0$ gives:

$$[RB] = \frac{k_1[RNA][BYO]}{k_{-1} + k_2}$$

Therefore $d[P]/dt = k_{ss}[RNA][BYO]$, where $k_{ss} = k_1k_2/(k_{-1}+k_2)$. Since $[BYO]>>[RNA]$, $[BYO]$ is taken to be constant for now (see next paragraph to consider degradation by hydrolysis), and $[RNA] + [P]$ = a constant D. Then $d[P]/dt = k_{ss}[BYO](D-[P])$. Alternatively, the pre-equilibrium approximation would give $[RB] = (k_1/k_{-1})[RNA][BYO]$, leading to $d[P]/dt = k_{pe}[BYO](D-[P])$, where $k_{pe} = k_1k_2/k_{-1}$. Both approximations yield the same form of the rate law, namely the solution $P = D(1-e^{-k[BYO]t})$. Expressing P and D as fractions gives $F_s =\ A_s(1 - \mathrm{e}^{-k_s[BYO]t})$, as used in the remainder of this paper.

BYO hydrolysis occurs on the timescale of minutes to hours, which can influence measurement of rate constants. Given the high molar excess of BYO compared to RNA, the change in [BYO] over time is dominated by BYO degradation. The integrated rate law therefore has the form $F_s(t) = A_s(1-e^{-k[\mathrm{BYO}]\exp(-k't)})$, where [BYO] is the initial concentration of BYO, k' is the BYO degradation rate, and k is a combined rate constant. In the experiments here, t = a constant (90 min), so the effective rate law has the form $F_s = A_s(1-e^{-k[\mathrm{BYO}]C})$, where C is a constant that depends on t and k'. Intuitively, this rate law could be compared to a first-order rate law $F_s =\ A_s(1 - \mathrm{e}^{-k_s[BYO]t})$ at constant t, in which [BYO] has been adjusted by a correction factor α that accounts for BYO degradation, and k_s is the effective rate constant. To determine α, the half-life of BYO was determined as described in the Methods, giving a half-life for BYO of 36.5 min, or $k' = 0.019$ min^{-1}. From k', α was calculated as the mean fraction of initial [BYO] over a 90 minute incubation ($\alpha = 0.479$). The pseudo-first-order rate law with k_s and α is not mechanistically correct but is used as a convenient formalism for fitting reaction rates for ribozymes. Note that the dimensionless catalytic ratio is not affected by use of this formalism (i.e., the catalytic ratio of k and k_s are equal).

Name	**Sequence (random region)**	**A_s** (by gel)	**A_s** (by *k*-Seq)	**k_s** (by gel) (min^{-1}M^{-1})	**k_s** (by *k*-Seq) (min^{-1}M^{-1})	**r_s** (by gel)	**r_s** (by *k*-Seq)	**Reason chosen**
S-2.1-a	ATTACCCTGGT CATCGAGTGA	0.450 ± 0.012	0.161 ± 0.007	1570 ± 260	779 ± 21	1100 ± 240	1010 ± 100	Most abundant sequence, most abundant family
S-2.1-t	ATTACCCTGGT CATCGAGTGT	0.446 ± 0.072	0.158 ± 0.007	890 ± 267	729 ± 28	620 ± 240	930 ± 130	Mid-low-abundance mutant of most abundant family
S-1A.1-a	CTACTTCAAAC AATCGGTCTG	0.708 ± 0.008	0.283 ± 0.069	303 ± 29	121 ± 11	330 ± 27	280 ± 54	Most abundant sequence, second-most abundant family
S-1B.1-a	CCACACTTCAA GCAATCGGTC	0.708 ± 0.124	0.865 ± 0.185	247 ± 49	46.2 ± 17.6	270 ± 45	320 ± 83	Most abundant sequence, third-most abundant family
S-1B.2-a	CCGCTTCAAGC AATCGGTCGC	0.704 ± 0.238	0.669 ± 0.275	112 + 23	47.3 ± 11.5	120 ± 20	260 ± 54	Most abundant sequence, fourth-most abundant family
S-1B.3-a	CCGAGTTTCAA GCAATCGGTC	0.700 ± 0.064	0.458 ± 0.313	194 ± 19	71.2 ± 20.6	210 ± 17	260 ± 98	Expected to have medium-high activity
S-3.1-a	AAGTTTGCTAA TAGTCGCAAG	0.825 ± 0.006	0.134 ± 0.013	169 ± 12	142 ± 3	220 ± 11	150 ± 14	Most abundant sequence, most abundant family of motif 3
S-2.2-a	ATTCACCTAGG TCATCGGGTG	0.404 ± 0.050	0.132 ± 0.019	355 ± 75	197 ± 9	220 ± 69	210 ± 45	Most abundant sequence from second-most abundant family from motif 2
S-1A.1-n	CTCTTCAAACA ATCGGTCTTC	0.719 ± 0.249	0.251 ± 0.145	127 ± 860	74.9 ± 5.2	180 ± 790	150 ± 24	Expected to have medium-low activity
S-1C.1-a	CTCTTCAATAA TCGGTTGCGT	0.516 ± 0.048	1.000 ± 0.000	81.5 ± 20.4	6.65 ± 0.75	63 ± 19	54 ± 4	Most abundant sequence, least abundant submotif
Baseline Activity	random RNA	**k_0A_0** (measured by gel) (min^{-1}M^{-1})			**k_0A_0** (used for comparison to *k*-Seq data) (min^{-1}M^{-1})			background activity
		0.645 ± 0.283			0.124 (see caption)			

Supporting Table 1. Sequences chosen for gel-shift assay. Ten sequences were chosen for gel-based activity testing; nine were the highest-abundance centers of a range of different sequence families (± indicates standard deviation of triplicates). *k*-Seq activity estimates were not adjusted for expected loss due to column binding and recovery (see Methods). A linear fit of data (Supporting Figure 3A) gives a correction factor for loss $l = 80.7\%$. Since the baseline activity measurement was made by gel shift assay ($k_0A_0 = 0.645$ min^{-1} M^{-1}), we calculate the effective *k*-Seq baseline $k_0A_0 = (1\text{-}l)(0.645) = 0.124$ min^{-1} M^{-1}, used when calculating catalytic ratio r_s from *k*-Seq measurements.

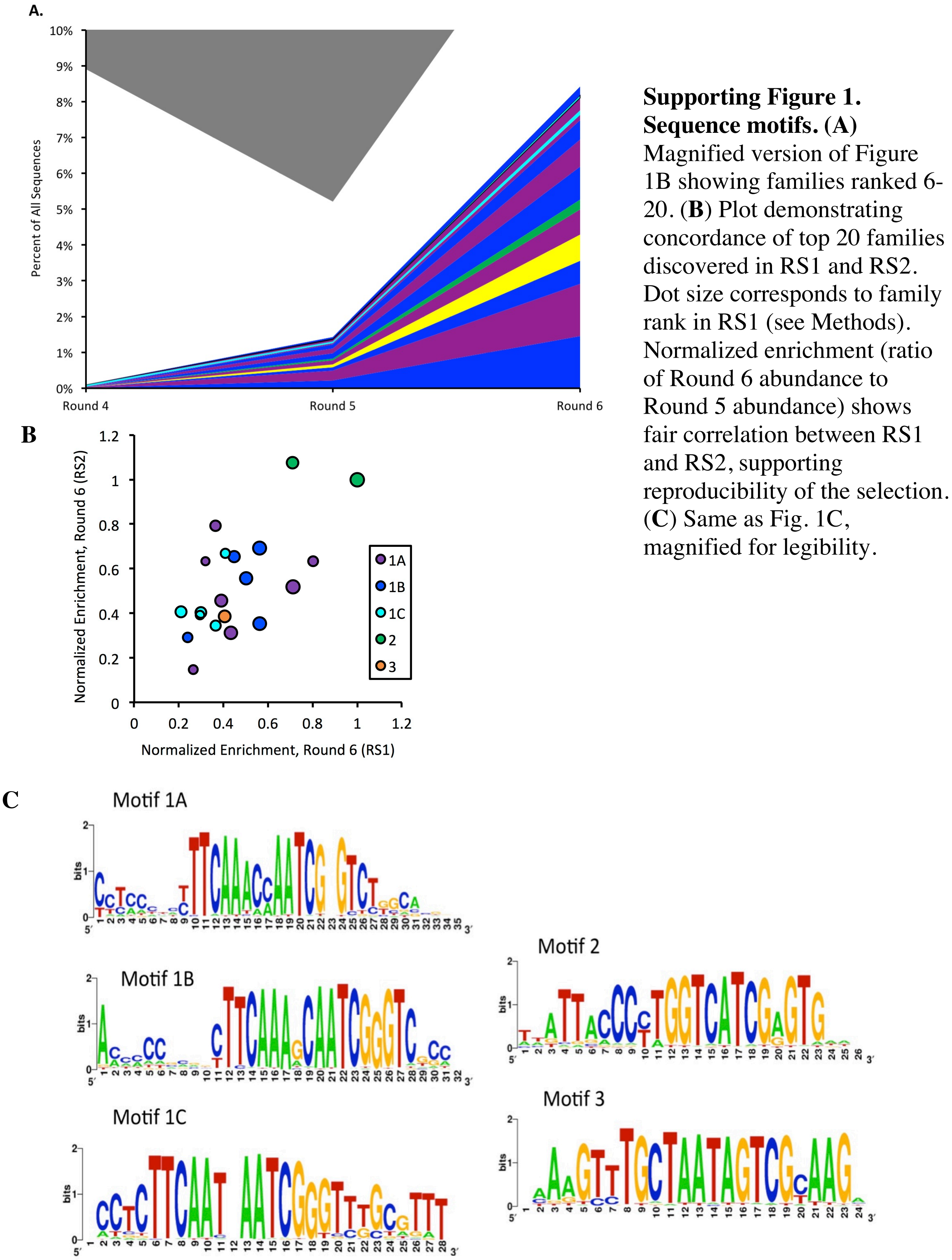
Start Sequence	**End Sequence**	**Name**	**Total Path Length**	**Path Steps**	**Largest Step**	**Minimum Count**
S-1A.1-a	**S-1B.1-a**	**1A-1B:1**	7	7	1	4
		1A-1B:2	7	7	1	4
		1A-1B:3	7	7	1	4
		1A-1B:4	7	7	1	4
		1A-1B:5	7	7	1	4
		1A-1B:6	6	5	2	12
		1A-1B:7	6	5	2	12
		1A-1B:8	6	5	2	9
		1A-1B:9	6	5	2	9
		1A-1B:10	6	5	2	7
S-1A.1-a	**S-1C.1-a**	**1A-1C:1**	8	6	2	4
		1A-1C:2	8	6	2	4
		1A-1C:3	8	6	2	4
		1A-1C:4	8	6	2	3
		1A-1C:5	8	6	2	3
		1A-1C:6	7	5	3	13
		1A-1C:7	7	5	3	10
		1A-1C:8	7	5	3	7
		1A-1C:9	7	5	3	7
		1A-1C:10	7	5	3	3
S-1B.1-a	**S-1C.1-a**	**1B-1C:1**	12	9	2	2
		1B-1C:2	12	9	2	2
		1B-1C:3	12	8	2	2
		1B-1C:4	12	8	2	2
		1B-1C:5	12	8	2	2
		1B-1C:6	11	7	3	3
		1B-1C:7	11	7	3	3
		1B-1C:8	11	7	3	3
		1B-1C:9	11	7	3	2
		1B-1C:10	11	7	3	2
S-1A.1-a	**S-2.1-a**	**1A-2:1**	24	11	4	2
		1A-2:2	24	11	4	2
		1A-2:3	24	11	4	2
		1A-2:4	24	11	4	2
		1A-2:5	24	11	4	2
		1A-2:6	15	7	5	2
		1A-2:7	15	7	5	2
		1A-2:8	15	7	5	2
		1A-2:9	15	7	5	2
		1A-2:10	15	7	5	2
S-1A.1-a	**S-3.1-a**	**1A-2:1**	27	11	4	2
		1A-2:2	27	11	4	2
		1A-2:3	27	11	4	2
		1A-2:4	27	11	4	2
		1A-2:5	27	11	4	2
		1A-2:6	12	9	5	2
		1A-2:7	12	9	5	2
		1A-2:8	12	9	5	2
		1A-2:9	12	9	5	2

		1A-2:10	12	9	5	2
S-2.1-a	**S-3.1-a**	**1A-2:1**	21	9	4	2
		1A-2:2	21	9	4	2
		1A-2:3	21	9	4	2
		1A-2:4	21	9	4	2
		1A-2:5	21	9	4	2
		1A-2:6	18	7	5	2
		1A-2:7	18	7	5	2
		1A-2:8	18	7	5	2
		1A-2:9	18	7	5	2
		1A-2:10	18	7	5	2
S-2.1-a	**S-2.2-a**	**2-2.1:1**	5	5	1	2
		2-2.1:2	5	5	1	2
		2-2.1:3	5	5	1	2
		2-2.1:4	5	5	1	2
		2-2.1:5	5	5	1	2
		2-2.1:6	5	4	2	40
		2-2.1:7	5	4	2	13
		2-2.1:8	5	4	2	13
		2-2.1:9	5	4	2	13
		2-2.1:10	5	4	2	3

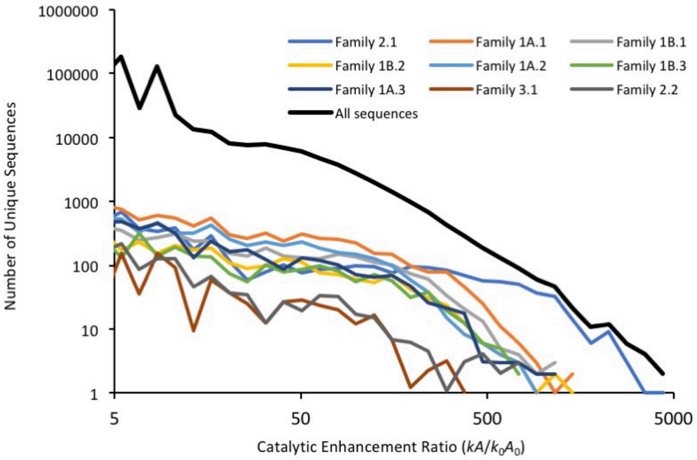
Supporting Table 2. Pathways between sequence peaks. For each pathway found between peak centers of interest, 10 pathways were found as described in Methods. For each path, “largest step” corresponds to the edit distance of the largest step present within the pathway, and “minimum count” is the lowest Round count (# of sequence reads) of any sequence that the pathway passes through.

Sequence	Nucleotide position (Supp. Fig. 6B)	Motif position (Fig. 1C)	Predicted secondary structure (Supp. Fig. 6B)	Observation (Fig. 1C)	Supports predicted secondary structure
S-1A.1-a	30	9	stem	C and U tolerated across G	yes
	29	8	loop	lack of conservation	yes
S-1B.1-a	27-28	5-6	stem	U not tolerated across G	maybe
	29-31	7-10	loop	lack of conservation	yes
S-3.1-a	28	4	stem	G tolerated across U	yes
	31	7	stem	C tolerated across G	yes
	44	20	loop	C or U	maybe
S-2.1-a	27	3	stem	G not tolerated more than C across U	maybe
	30	6	loop	A conserved	maybe
	31	7	stem	U not tolerated more than A across G	maybe
	43	20	stem	G tolerated across U	yes
	47	24	loop	lack of conservation	yes

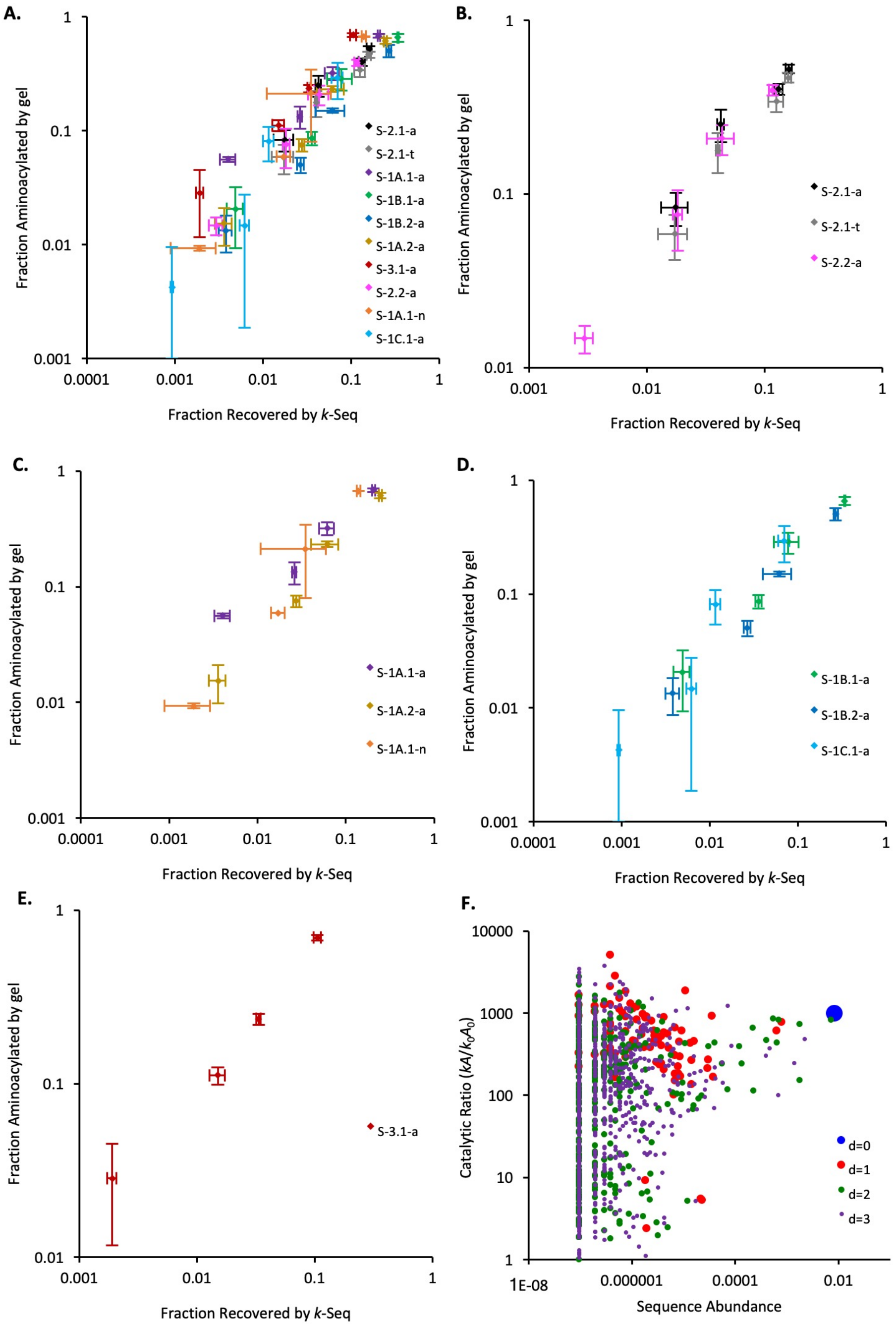
Supporting Table 3. **Comparison of predicted secondary structure elements and sequence conservation patterns**. Sites of relatively low conservation (information content < 1 bit, shown in blue in Supp. Fig. 6B) were inspected to determine whether their predicted presence in a stem or loop was supported by the observed mutation pattern (Fig. 1C). Such sites that could be identified unambiguously in the motif were included here. In 10 of 16 sites among the 4 sequences, the conservation pattern supports the predicted secondary structure. Regarding the remaining sites, because these sites may have a functional role in addition to a structural role, lack of the expected conservation pattern cannot be interpreted as disagreement with the predicted secondary structure.



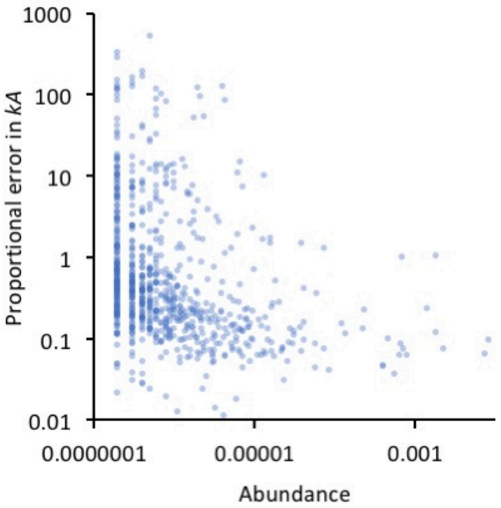
Supporting Figure 1. Sequence motifs. (A) Magnified version of Figure 1B showing families ranked 6-20. (**B**) Plot demonstrating concordance of top 20 families discovered in RS1 and RS2. Dot size corresponds to family rank in RS1 (see Methods). Normalized enrichment (ratio of Round 6 abundance to Round 5 abundance) shows fair correlation between RS1 and RS2, supporting reproducibility of the selection. (**C**) Same as Fig. 1C, magnified for legibility.



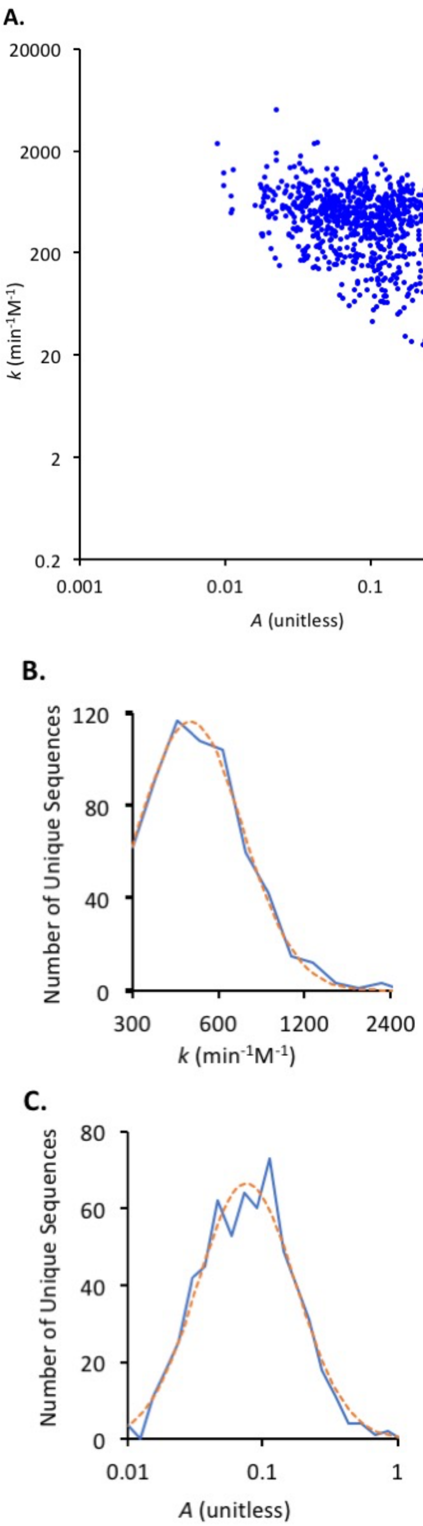
Supporting Figure 2. Frequency distribution of ribozymes over activity. Distributions for the highest activity families are plotted individually. The overall curve, including all families (black), is driven at higher activities by contributions from the families shown, which make up a significant portion of high-activity sequences.



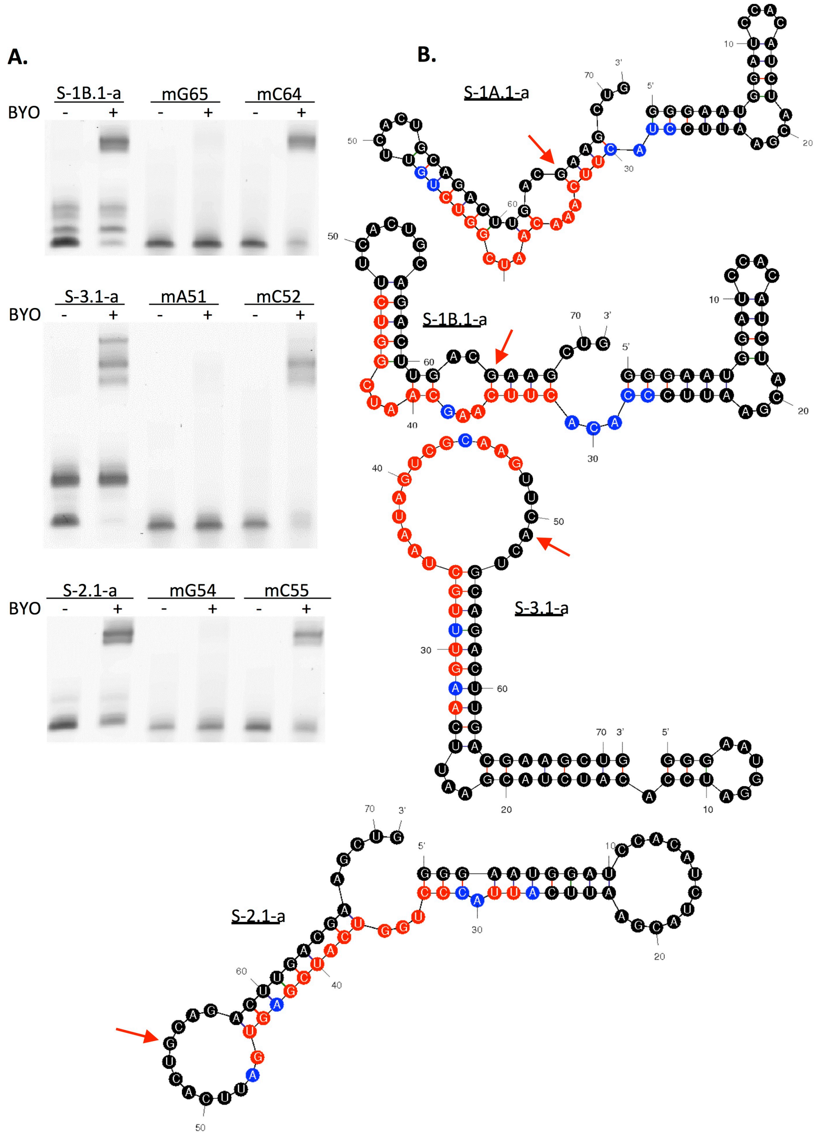
Supporting Figure 3. Sequence abundance and activity. (A) All data points from gel-shift assays (10 sequences, 4 concentrations of BYO, 2-3 replicates), compared to *k*-Seq measurements of sequence recovery for the same sequences. Error bars are standard deviation for triplicates (*k*-Seq) or 2-3 replicates (gel assay). The same data are separated by motif (**B-E**). The four data points for each ribozyme correspond to the four concentrations of BYO tested (2, 10, 50, and 250 µM); in each case, the average fraction aminoacylated increases with increasing BYO concentration. **(F)** Catalytic rate enhancement (*k*-Seq) vs Round 6 abundance for Family 2.1. Sequences sorted by distance from peak center (*d* = 0,1,2,3, colored as blue, red, green, and purple, respectively). No correlation is observed between sequence abundance and catalytic enhancement.



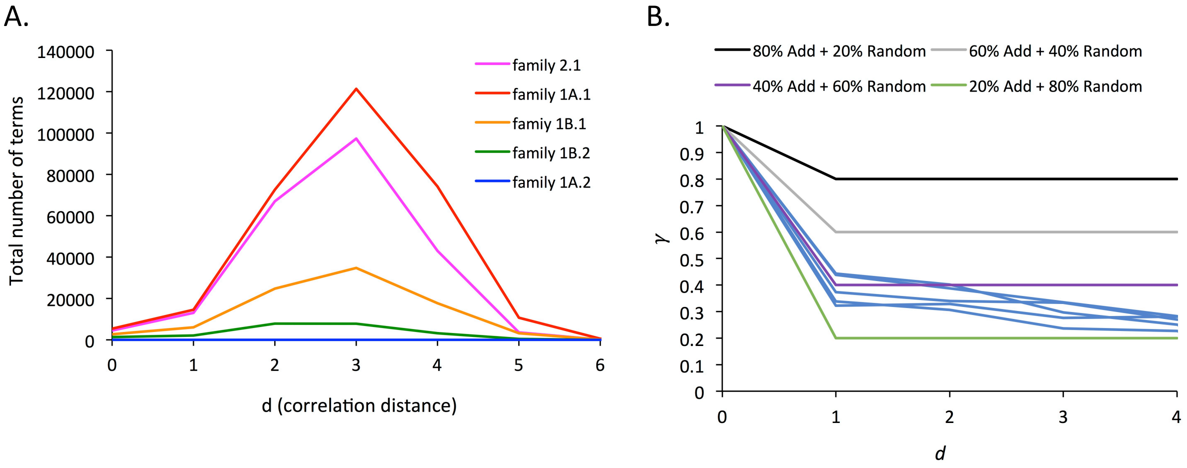
Supporting Figure 4. Standard error and limit of detection of *k*-Seq for Family 2.1. Proportional standard error is higher at very low abundance than at higher abundance. Most sequences having abundance >10^{-6} (count ~10) and nearly all sequences having abundance >10^{-5} (count ~ 100) have proportional error < 1 (i.e., <2-fold error). These values are consistent with noise due to stochastic sampling of sequencing reads, consistent with the vertical structure (corresponding to integer number of reads) seen in this plot.



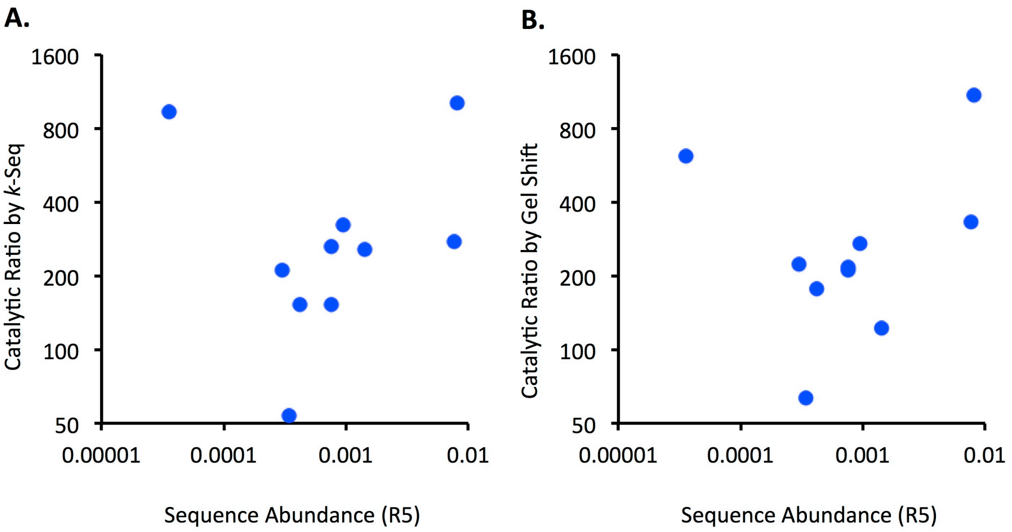
Supporting Figure 5. The distribution of k_s and A_s in Family 2.1. (A) k_s vs. A_s for all sequences in Family 2.1. Whether k_s and A_s can be estimated well separately depends on whether the concentration series adequately samples into the saturation regime for a particular ribozyme. For sequences of high $k_{s,}$ the saturation phase of the curve is captured and a reasonable fit can be obtained. However, for sequences of low k_s, a fit with A_s=1 is found; $k_s A_s$ is still expected to be accurately estimated, but k_s and A_s cannot be estimated separately with accuracy as the curve appears linear in this range. If k_s and A_s estimation for lower activity sequences is desirable, a reaction at high concentration could be added to the *k*-Seq experiment. For sequences of higher k_s, there appears to be little correlation between k_s and A_s. The distributions of k_s **(B)** and A_s **(C)** fit well to log-normal distributions (dotted red line, nonlinear R^2=0.99 for k_s , and R^2=0.97 for A_s).



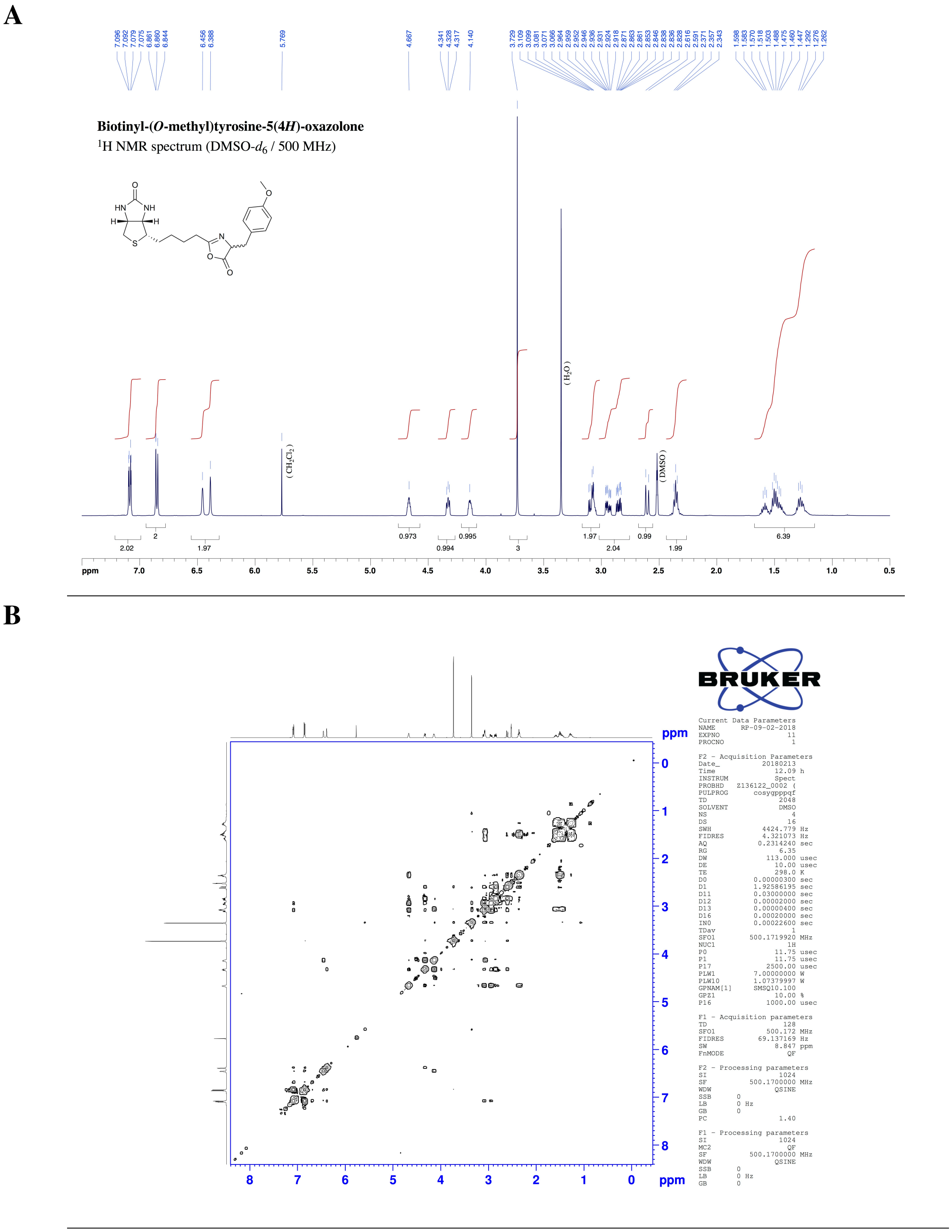
Supporting Figure 6. Catalytically reactive nucleotide and predicted structures. (A) Streptavidin gel shift assay for RNA sequences S-1B.1-a, S-3.1-a, and S-2.1-a in analogy to Figure 3. **(B)** Minimum free energy secondary structures for the sequences indicated, predicted by mfold [2]. Note that these structures have not been experimentally verified in this work (see Supporting Table 3 for discussion). Black denotes constant regions. Sites in the selected region conserved with information content < 1 bit are shown in blue; sites with information content > 1 bit are shown in red (also see Fig. 1C). Red arrows indicate the observed aminoacylation site.



Supporting Figure 7. Fitness correlation calculations. (A) Number of terms contributing to the calculation of γ_d across various edit distance *d* for five ribozyme families (2.1, 1A.1, 1B.1, 1B.2, 1A.2). The number of possible comparisons increased up to $d = 3$, with a substantial number of terms present up to $d = 4$, suggesting that γ_d is likely to be accurate up to $d = 4$. **(B)** Average correlation of fitness effects for the Rough Mt. Fuji model, calculated with different amounts of additivity vs. randomness (see legend). The observed data (Figure 3C) is shown in blue.



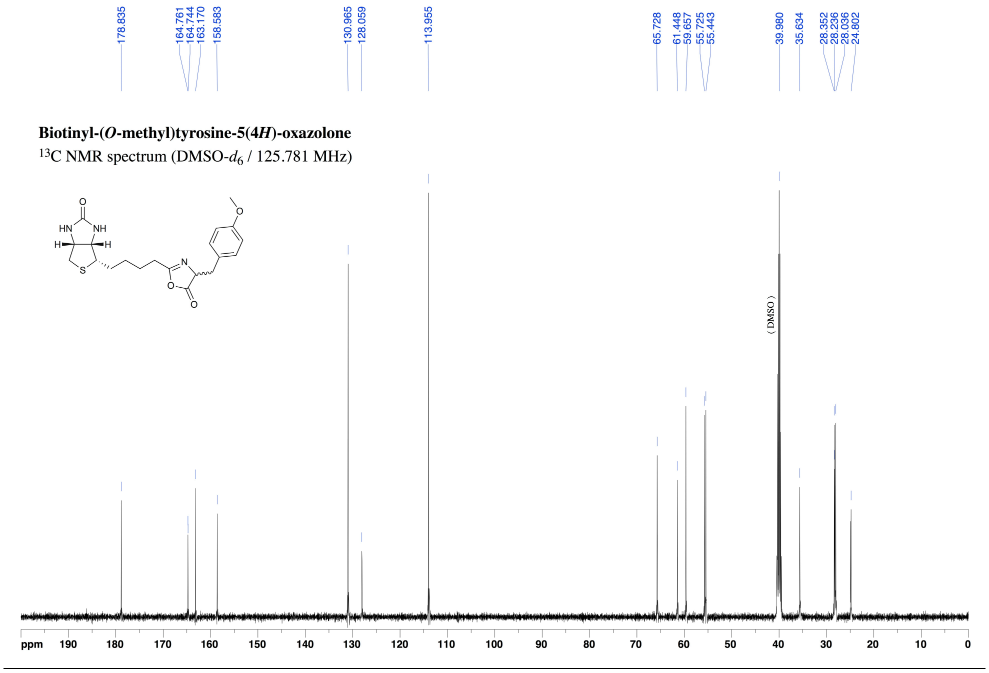
Supporting Figure 8. Abundance does not correlate with catalytic activity. The catalytic ratio measured by *k*-Seq (**A**) or gel shift (**B**) is shown against the abundance of the sequence in Round 5 (RS1). This lack of correlation underscores the need for measurement of catalytic activity rather than abundance.



```
Current Data Parameters
NAME      RP-09-02-2018
EXPNO                11
PROCNO                1

F2 - Acquisition Parameters
Date_          20180213
Time              12.09 h
INSTRUM           Spect
PROBHD   Z136122_0002 (
PULPROG       cosygpppqf
TD                 2048
SOLVENT            DMSO
NS                    4
DS                   16
SWH            4424.779 Hz
FIDRES         4.321073 Hz
AQ            0.2314240 sec
RG                 6.35
DW              113.000 usec
DE                10.00 usec
TE                298.0 K
D0           0.00000300 sec
D1           1.92586195 sec
D11          0.03000000 sec
D12          0.00002000 sec
D13          0.00000400 sec
D16          0.00020000 sec
IN0          0.00022600 sec
TDav                  1
SFO1        500.1719920 MHz
NUC1                 1H
P0                11.75 usec
P1                11.75 usec
P17             2500.00 usec
PLW1         7.00000000 W
PLW10        1.07379997 W
GPNAM[1]     SMSQ10.100
GPZ1              10.00 %
P16             1000.00 usec

F1 - Acquisition parameters
TD                  128
SFO1            500.172 MHz
FIDRES        69.137169 Hz
SW                8.847 ppm
FnMODE               QF

F2 - Processing parameters
SI                 1024
SF          500.1700000 MHz
WDW               QSINE
SSB      0
LB       0 Hz
GB       0
PC                 1.40

F1 - Processing parameters
SI                 1024
MC2                  QF
SF          500.1700000 MHz
WDW               QSINE
SSB      0
LB       0 Hz
GB       0
```

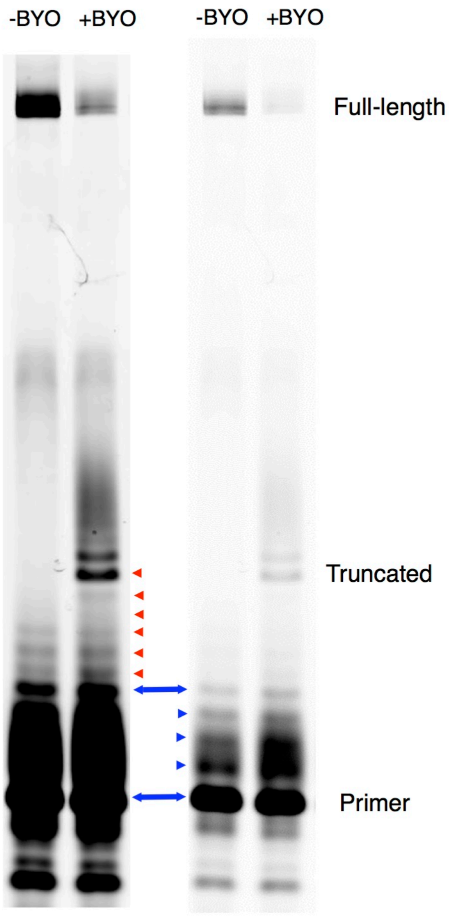
Supporting Figure 9. NMR spectra of biotinyl-(*O*-methyl)tyrosine-5(4*H*)-oxazolone. NMR spectra obtained from a Bruker Avance HD spectrometer (500MHz), equipped with a helium BBO cryoprobe, for biotinyl-(O-methyl)tyrosine oxazolone. Shown are (**A**) ^{1}H, (**B**) 2D-COSY, and (**C**) ^{13}C spectra.

C

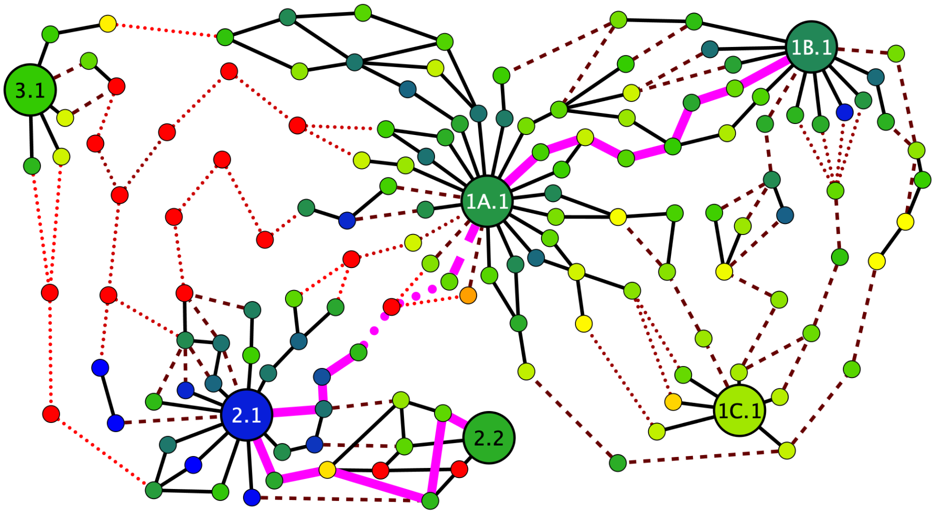


Biotinyl-(*O*-methyl)tyrosine-5(4*H*)-oxazolone
^{13}C NMR spectrum (DMSO-d_6 / 125.781 MHz)

Supporting Figure 9, continued.



Supporting Figure 10. High contrast version (left) of Figure 3A (right), showing single-nucleotide resolution of banding used to identify the suspected site of aminoacylation. Blue marks indicate bands terminating within the ligated adapter; red marks indicate bands terminating within the ribozyme sequence. In this case, the main reverse transcription stall occurs immediately before the 7th position from the end of ribozyme sequence, implicating G65.



Supporting Figure 11. Evolutionary network shown in Figure 4B, with the pathway of Figure 4A highlighted in fuchsia.

Supporting References

1. Liu, Z.; Rigger, L.; Rossi, J.C.; Sutherland, J.D.; Pascal, R. *Chemistry* **2016**, *22*, 14940; Liu, Z.; Hanson, C.; Ajram, G.; Boiteau, L.; Rossi, J.C.; Danger, G.; Pascal, R. *Synlett* **2017**, *28*, 73.
2. Zuker, M. *Nucleic Acids Res* **2003**, *31*, 3406.