Supporting Information Appendix

Comprehensive experimental fitness landscape and evolutionary network for small RNA

Jose I. Jiménez, Ramon Xulvi-Brunet, Gregory Campbell, Rebecca Turk-MacLeod, Irene A. Chen

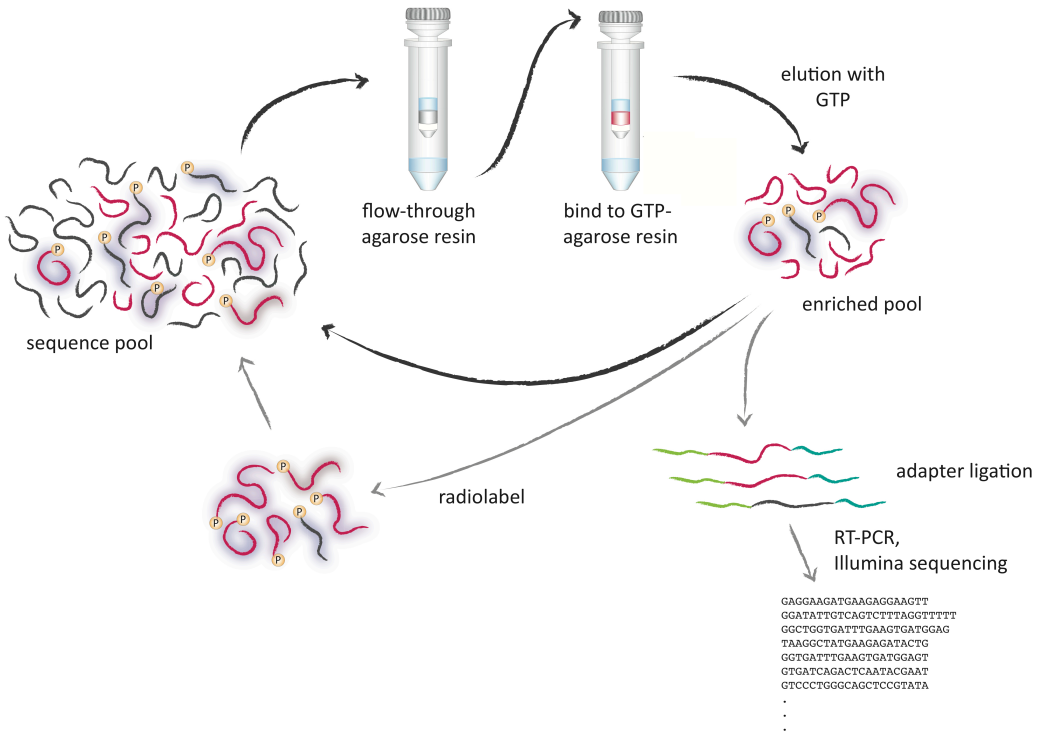


Figure S1. Diagram of selection procedure. The RNA pool is loaded onto an agarose resin and the flow-through fractions are collected and loaded onto the GTP-agarose resin. RNA sequences that bind to the column are eluted by GTP in free solution. Fractions of the enriched pool are taken for Illumina sequencing and for radiolabeling. The bulk of the pool is returned to the cycle (without PCR amplification between rounds).

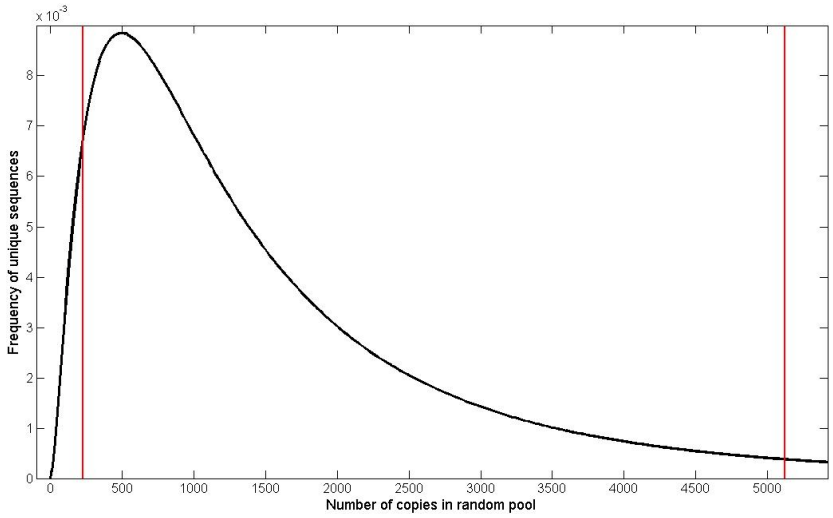


Figure S2. Coverage of sequence space by the initial library. We estimated the representation of sequences in the initial pool using a model for synthesis described in the Methods. The figure shows a histogram of the number of times sequences were estimated to be present in the pool. 90% of the sequences were estimated to be present ~250-5100 times (indicated by red lines), with >99.99% present at least once.

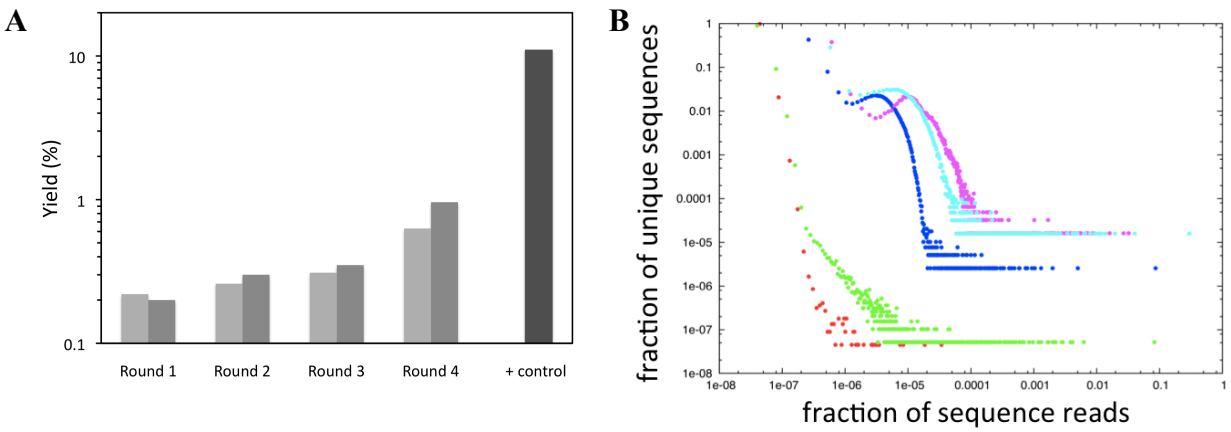


Figure S3. Selection progress. (A) Yield of RNA from each round of selection, measured by radiolabeling. The yield is the percentage of initial counts loaded onto the agarose resin that were subsequently obtained after elution from the GTP-agarose resin, desalting, and concentration. Data for two independent replicate selections are shown. Yield of a known GTP-binding aptamer is also shown (+control). **(B)** Enrichment of a subset of unique sequences during selection. HTS data after each round of selection were directly analyzed by counting the number of sequence reads corresponding to each unique sequence identified in the sample. Both axes were normalized within each sample. Fewer unique sequences comprised a greater fraction of the sequence reads as the selection progressed, seen as an upward-rightward shift in the curves (initial pool is red; round 1 is green; round 2 is blue; round 3 is purple; round 4 is cyan). See Figures S4, S17 for replicate experiment and post-correction data.

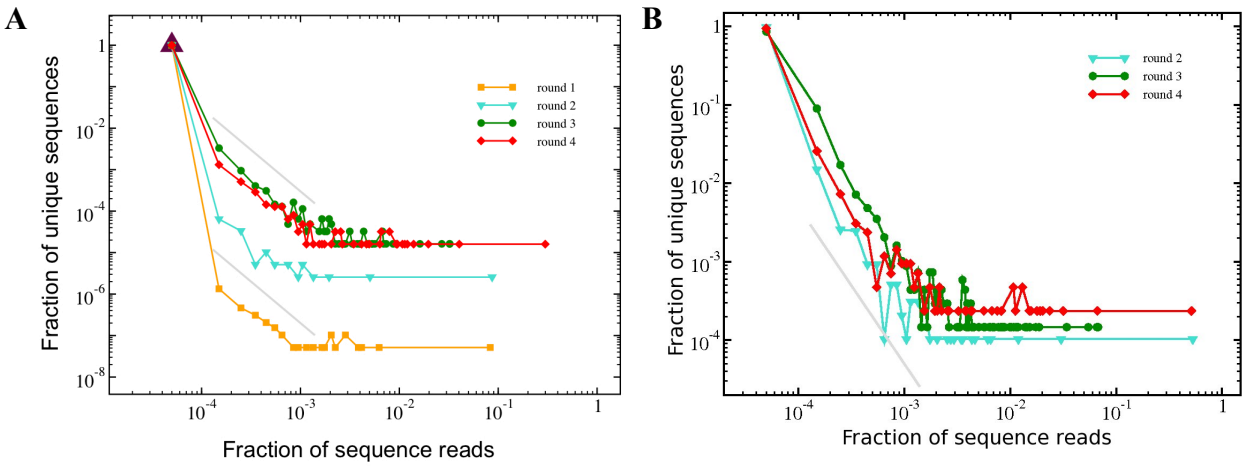


Figure S4. Histograms for corrected vs. uncorrected data. Corrections for sequencing errors, ligation biases, and non-random synthesis were applied to HTS data as described. Only data from rounds 2-4 could be corrected in this way because correcting for sequencing errors was computationally intractable for the initial pool and round 1, due to the large diversity of sequences. The correction did not affect the qualitative shape of the curves, as shown by comparison between the raw data (**A**) and corrected data (**B**). Data shown are the initial pool (brown triangle), round 1 (orange), round 2 (blue), round 3 (green), and round 4 (red). The gray line is a power law with exponent -2, provided to facilitate visual comparison between panels. The *x*-axis was divided into 10,000 evenly spaced bins.

d	# of isolated sequences	Range of fitness values
7	5	3458-6271
8	4	3567-6271
9	2	4349-6271

Table S1. Few isolated sequences were found to be above the threshold for significance. Because peaks are typically surrounded by sequences that form dead-end evolutionary tendrils (see Figure 3B in the main text), isolation in sequence space was defined as fulfilling two criteria: (1) no other sequence within a distance of three mutations, and (2) no central peak sequence within a distance *d*, where *d*=7, 8, or 9. In Replicate 1, no isolated sequences were found. In Replicate 2, a handful of isolated sequences were identified, each with relatively low fitness.

		Replicate 1, round			Replicate 2, round		
Peak name	Sequence	2	3	4	2	3	4
m01j03	GGCUGGUGAUUUGAAGUGAUGGAG		2.31E-1 (2.11E7)			1.43E-1 (1.07E6)	
m02j01	GAGGAAGAUGAAGAGGAAGUU		1.30E-1 (1.18E7)			1.68E-1 (1.26E6)	
m03j02	GGAUAUUGUCAGUCUUUAGGUUUUU	5.39E-1 (1.70E6)	1.75E-2 (1.59E6)	3.15E-3 (3.44E5)	8.67E-1 (1.51E7)	1.60E-1 (1.20E6)	8.31E-1 (2.36E7)
m04j04	UAAGGCUAUGAAGAGAUACUG		5.95E-3 (5.41E5)	5.30E-1 (5.78E5)		3.63E-2 (2.72E5)	1.26E-2 (3.58E5)
m05j10	GCCAUGUACACGAGGAAGGAAU		5.90E-3 (5.37E5)			4.32E-3 (3.24E4)	
m06j06*	GGUGAUUUGAAGUGAUGGAGUUGG		4.81E-3 (4.38E5)			6.94E-3 (5.22E4)	
m07j07	GUGAUCAGACUCAAUACGAAU		3.72E-3 (3.38E5)	6.73E-3 (7.34E4)		5.93E-3 (4.45E4)	9.90E-4 (2.81E4)
m10j11	GAAGUGAUGGAGUUGGCCAGCC		1.03E-3 (9.35E4)			2.53E-3 (1.90E4)	
m14j12	CCUAAAGACUGACAAUAUCCAAAAA		4.18E-4 (3.80E4)			2.02E-3 (1.51E4)	
m15j18	AUUCGUAUUGAGUCUGAUCACAC		3.38E-4 (3.08E4)			4.10E-4 (3.08E3)	
m16j20	AAUUCUCCGAACGUGUCACGU		3.37E-4 (3.07E4)			3.91E-4 (2.94E3)	
m17j08	ACCGGCAAAGAAGCGAUGCUU		3.30E-4 (3.00E4)			5.19E-3 (3.90E4)	
m18j13	UUAAUUAAAGACUUCAAGCCC		3.05E-4 (2.77E4)			1.85E-3 (1.39E4)	
m19j09	GUCCCUGGGCAGCUCCGUAUA		3.00E-4 (2.72E4)			4.42E-3 (3.32E4)	1.27E-3 (3.61E4)
m20j22	GGGGACUCAUGGAGAACAG		2.97E-4 (2.70E4)			3.71E-4 (2.79E3)	
m08	GAGCCGAGACAAUCUCUG		3.10E-3 (2.82E5)				
m09	GGAUAUUGUCAGUCUUUAGGU		1.86E-3 (1.69E5)				
m11	AAGAAGGAUCGAGCACCAGAAC		7.92E-4 (7.21E4)				
m12	GGCGGGACACGAGAAAUCCUGUCUG		4.79E-4 (4.35E4)				
m13	GGUGAUUGGAAGGGAGGGAGGUG		4.43E-4 (4.03E4)				
j05	CUCACUCUGCUGCGAGAAGUG					1.41E-2 (1.06E5)	
j14	GGACGGAUCGCGGAAGGUU					1.26E-3 (9.48E3)	
j15	AAGGCCGAGAGCUGAUGACGAGUUAU					6.59E-4 (4.95E3)	
j16	GAAUAGGAUGUCAUAGAGGGUG					5.39E-4 (4.04E3)	
j17	GAUUGGAAGGGAGGGAGGUGGCCA					4.95E-4 (3.72E3)	
j19	UACGGAAUUCGCCAGAGAUGGCUUAG					4.02E-4 (3.02E3)	
j21	GAGGACAGGAAGAUGAAGAGAAGU					3.88E-4 (2.91E3)	
r4m03	GGCUGGUGAUUGGAAGGGAGGGAGG			1.76E-2 (1.92E5)			
r4m04	GGAGGGGAUGAGCUGUGGAUAGGGGU			1.12E-2 (1.22E5)			
r4j03	UUAAUUAAAGACUUCAAGCUU						3.11E-3 (8.82E4)

Table S2. Peak sequences detected in rounds 2-4. Fitness values are given normalized to the total fitness observed in the sequence reads, with the absolute fitness value in parentheses. The absolute values are not meaningful except for the establishment of thresholds for significance (27,000 in Replicate 1; 2,790 in Replicate 2; corresponding to the fitness of peak m20j22).

* indicates that the highest fitness sequence of the peak was slightly different in the other replicate (i.e., lacking the three terminal nucleotides (UGG) on the 3′ end).

Peak	slope	*r*
all peaks	0.600 +/- 0.013	0.7939
m01j03	0.720 +/- 0.056	0.7830
m02j01	0.669 +/- 0.098	0.9148
m03j02	0.910 +/- 0.085	0.9055
m04j04	0.775 +/- 0.122	0.8864
m05j10	0.454 +/- 0.010	0.8353
m06j06	0.880 +/- 0.058	0.8812
m07j07	1.28 +/- 0.17	0.9829
m10j11	0.635 +/- 0.069	0.8556
m14j12	0.848 +/- 0.115	0.9333
m15j18	0.706 +/- 0.160	0.9111
m16j20	0.769 +/- 0.104	0.9821
m17j08	1.018 +/- 0.134	0.9831
m18j13	0.771 +/- 0.112	0.8506
m19j09	1.125 +/- 0.121	0.9830
m20j22	only one point	

Table S3. Reproducibility of fitness in different experiments. Correlation coefficients and the slope of the line of best fit for the log-transformed data for all peaks are given in the table. These slopes would be equal to one in the ideal case; the deviation from one suggests that the experiments differed slightly in some way, such as the strength of selection. Standard deviations are given for the slope values.

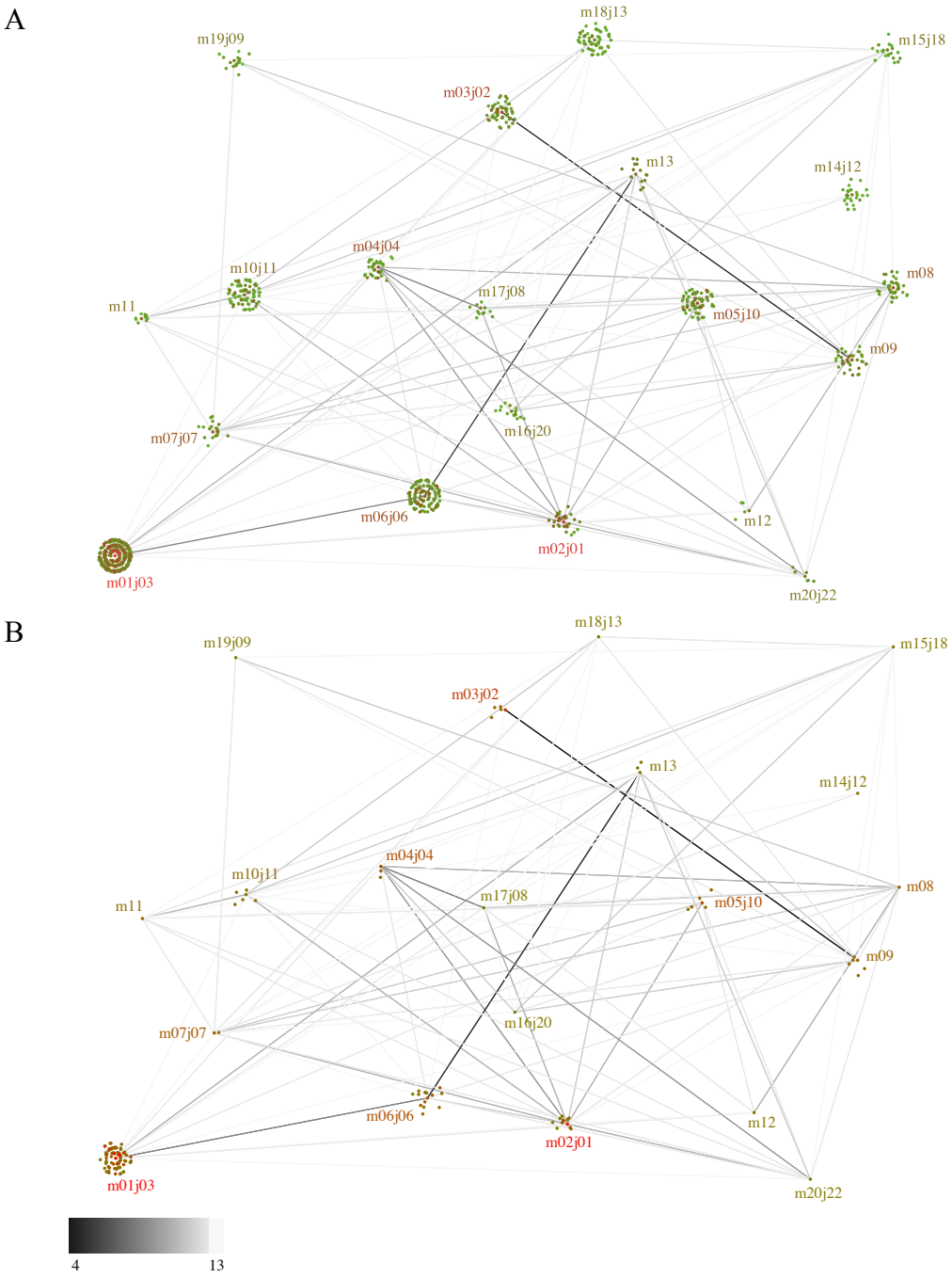


Figure S5. Fitness network for Round 3, with labels (one replicate). Threshold for peak detection was set as described in the Methods to maximize overlap between replicate experiments. Given that the peak sequence met the threshold, within each peak either (**A**) all sequences with adjusted abundance >0 were plotted, or (**B**) all sequences above the threshold were plotted. Low to high fitness is shown by color (green to red) and edit distances are indicated by the darkness of the line (legend at bottom).

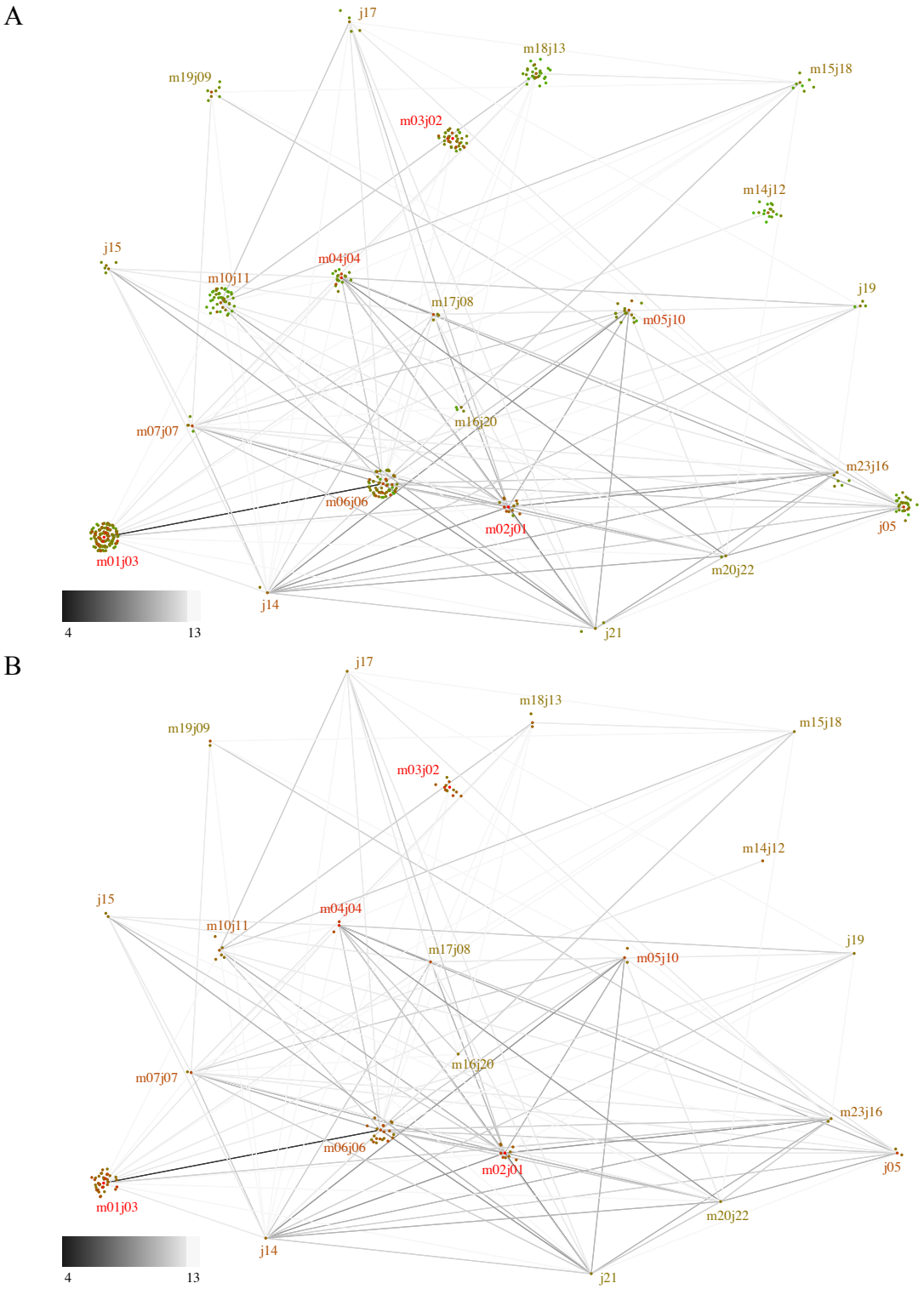
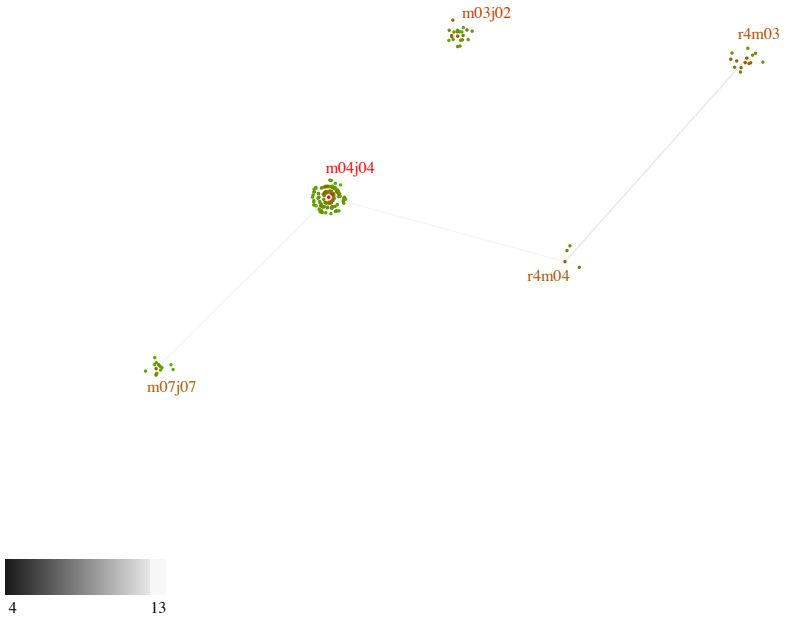


Figure S6. Fitness network for Round 3, with labels (another replicate).

A



B

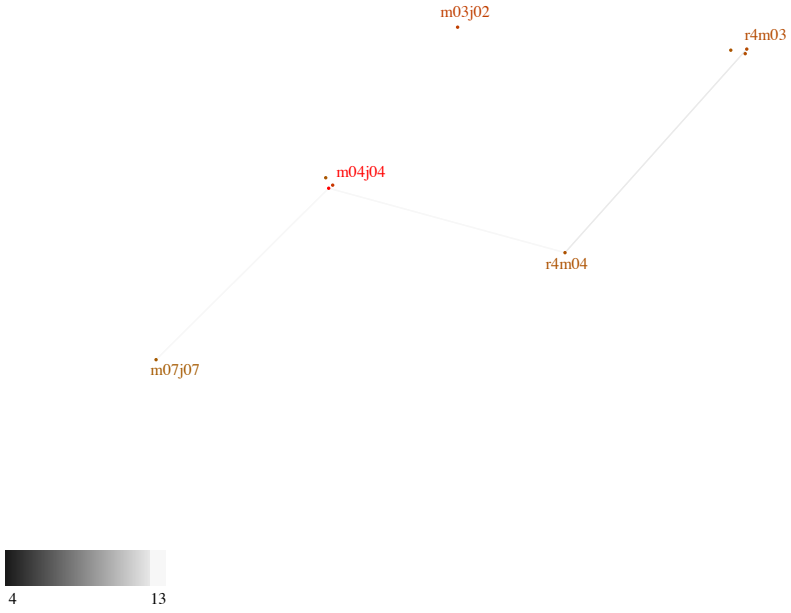
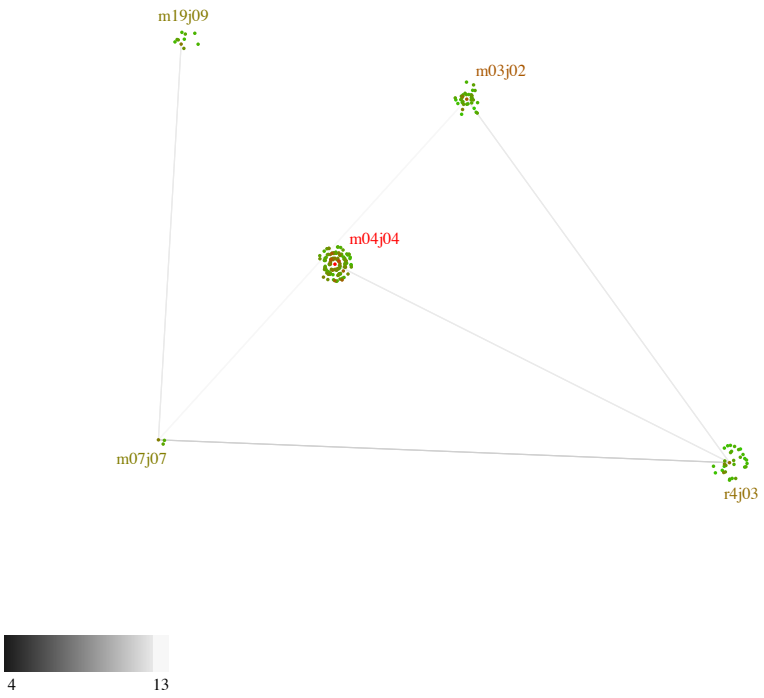


Figure S7. Fitness network for Round 4, with labels (one replicate).

A



B

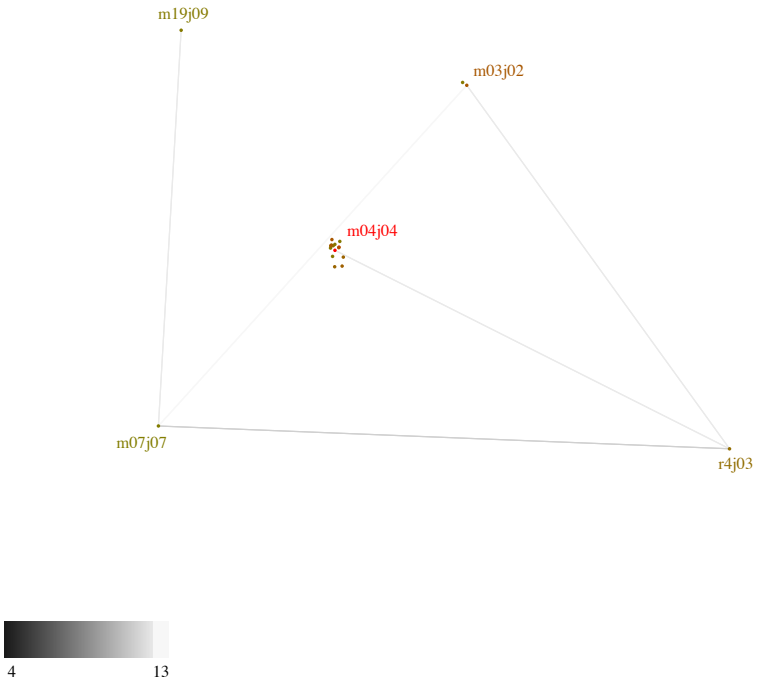


Figure S8. Fitness network for Round 4, with labels (another replicate).

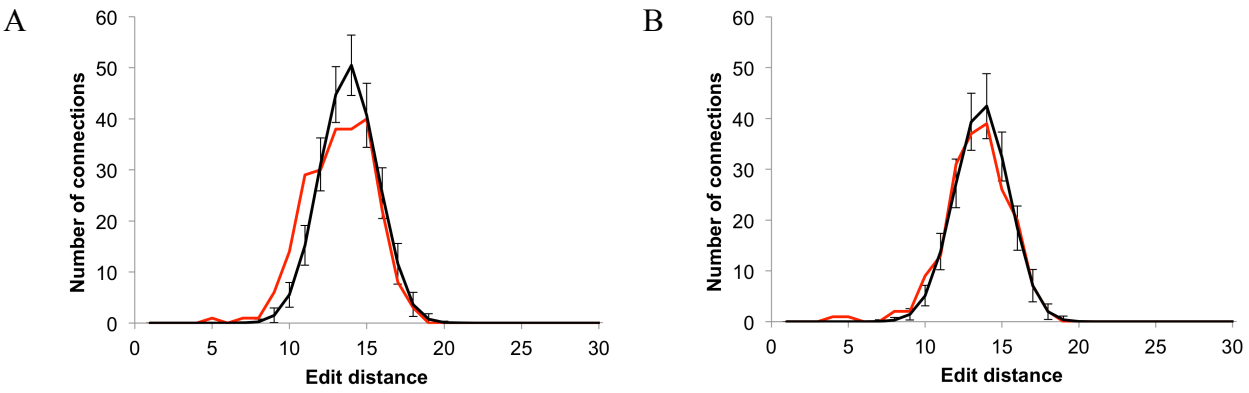


Figure S9. Network connectivity of fitness landscape compared with randomly distributed peaks in sequence space. See Figure 3A of main text; shown here are the analogous analyses for individual replicate experiments (**A** and **B**); error bars show the standard deviation.

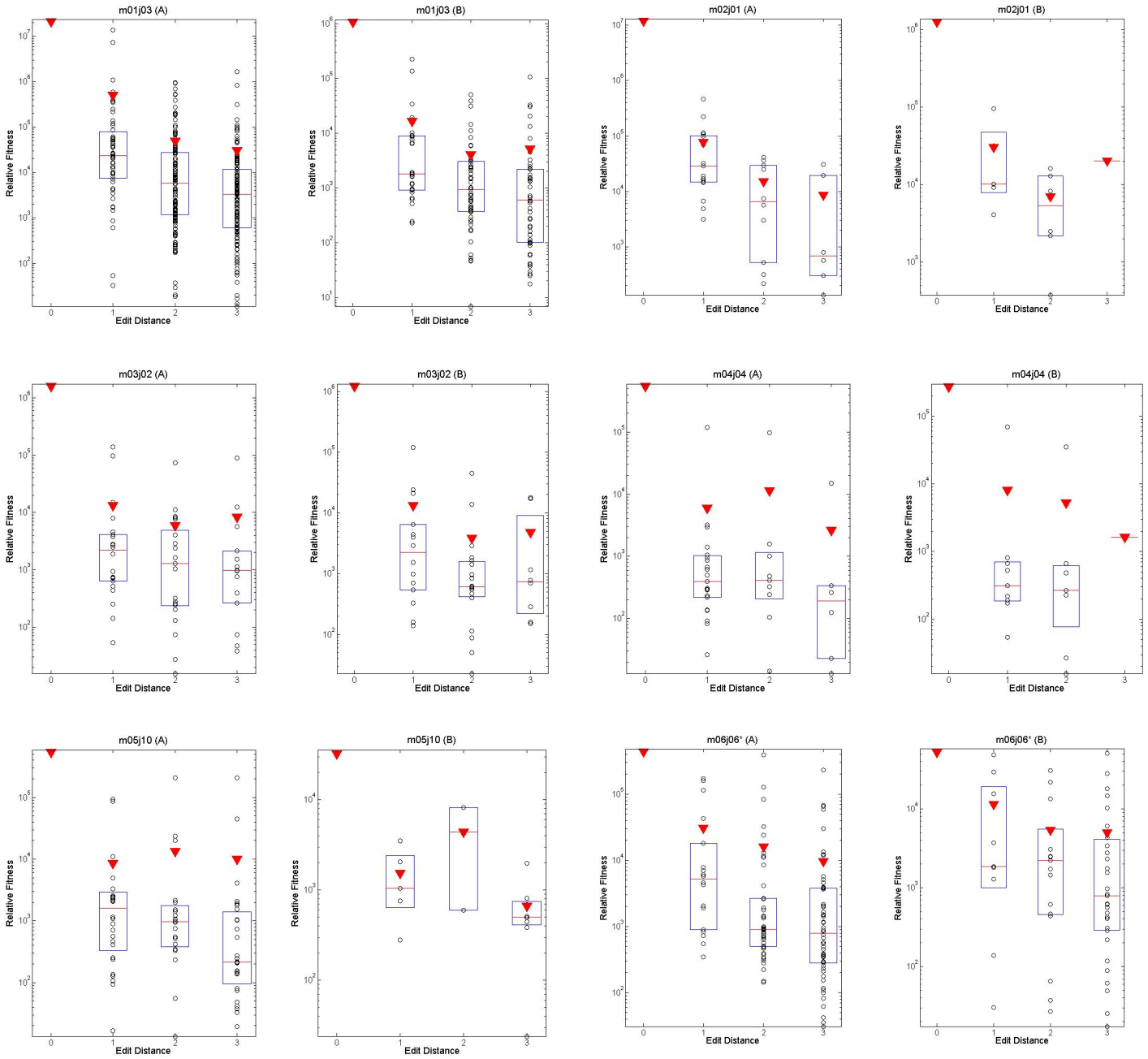


Figure S10. Fitness decreases with increasing edit distance from peak sequence. Plots are shown for all peaks having sequences detected at distance 1, 2, and 3. Box plots show relative fitness for peaks as a function of edit distance from the peak. Red triangles are the mean; red line is the median and box edges are the quartiles. Both replicates (A and B) are shown.

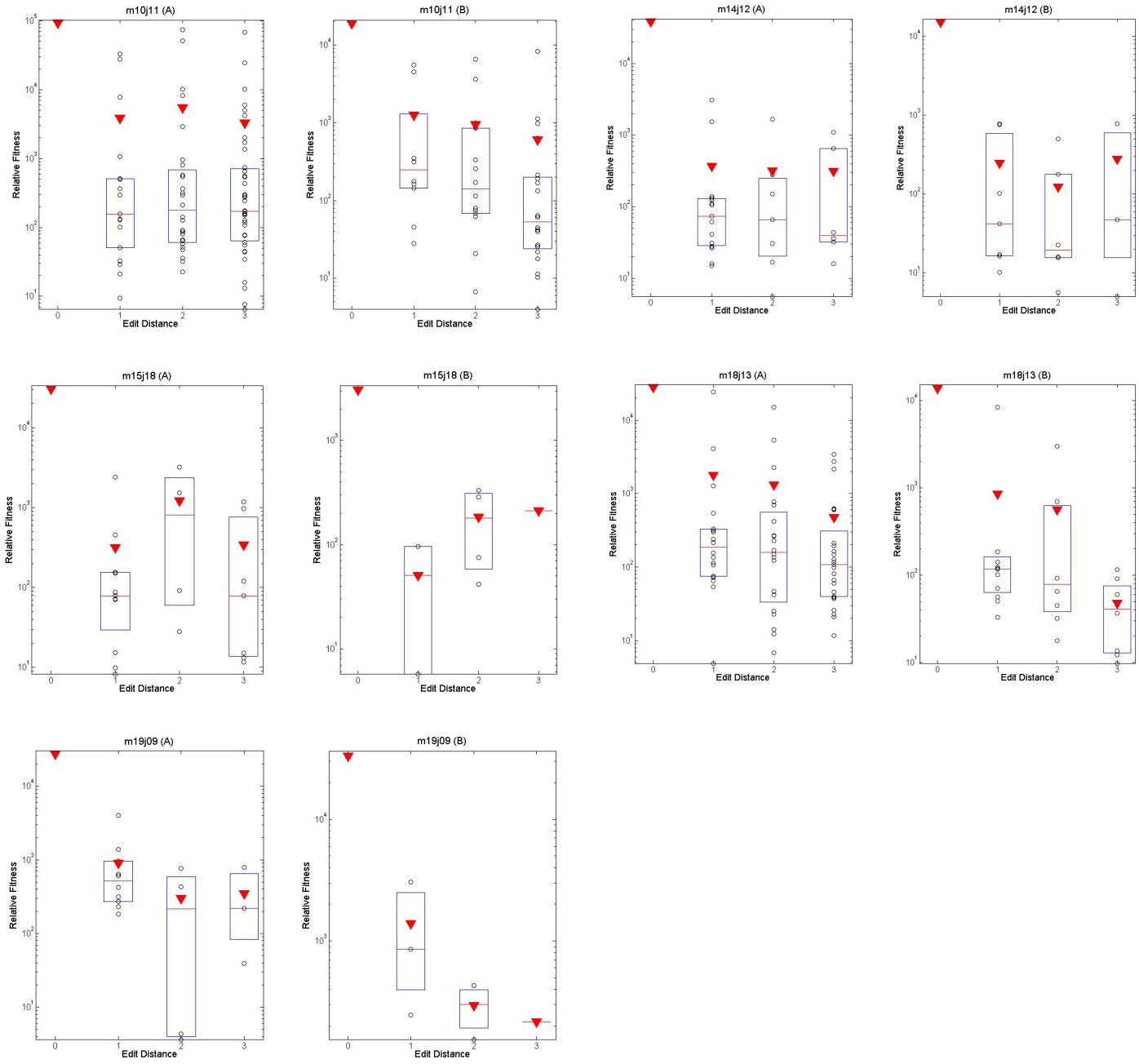
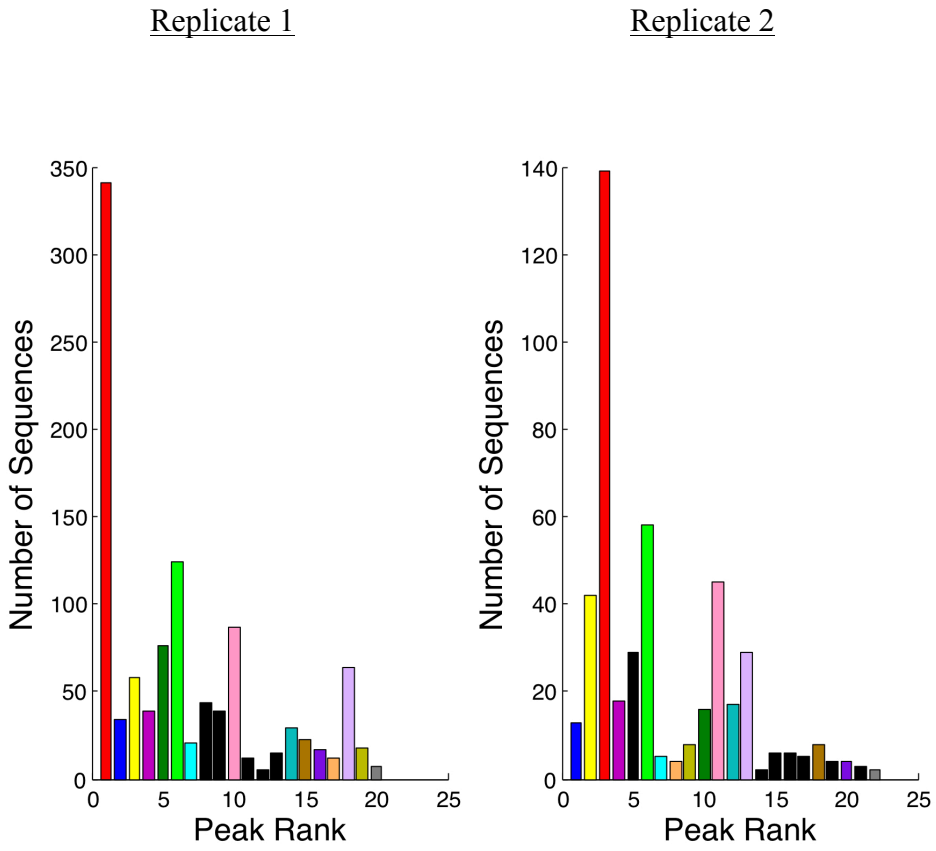


Figure S10 continued.



Replicate	Spearman's ρ	p-value
1	-0.6637	4.1×10^{-4}
2	-0.5082	3.4×10^{-5}

Figure S11. Histogram of number of sequences per peak for both replicate experiments. Data are from round 3. Identical peak sequences are given by the same color in both replicates (left and right). Sequences only detected in one replicate are shown by black bars. The Spearman correlation was calculated between the fitness rank and number of sequences in the peak for both replicates.

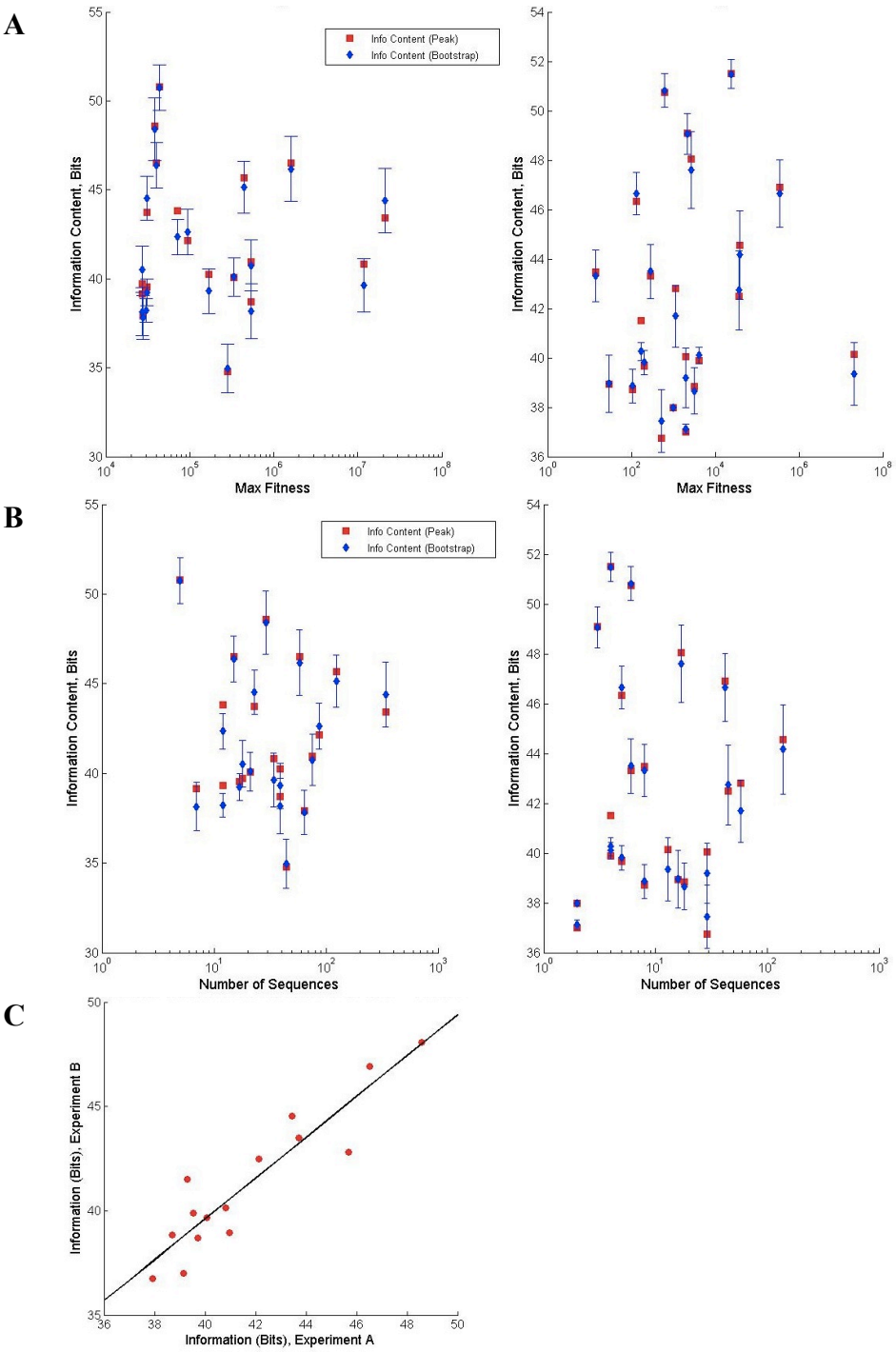


Figure S12. Functional information and (A) peak fitness or (B) peak size, i.e., number of sequences detected in the peak, calculated for the 15 peaks found in common between the two experiments (left and right). The red is the functional information calculated for the entire peak; in blue is the average value and standard deviation for subsamples taken from the peak. No obvious pattern was seen in these data. **(C)** Functional information estimated for the same peak from the two replicate data sets were similar ($r = 0.922$).

m01j03	5′ GGCUGGUGAUUUGAAGUGAUGGAG 3′
m02j01	5′ GAGGAAGAUGAAGAGGAAGUU 3′
m03j02	5′ GGAUAUUGUCAGUCUUUAGGUUUUU 3′
m04j04	5′ UAAGGCUAUGAAGAGAUACUG 3′
m05j10	5′ GCCAUGUACACGAGGAAGGAAU 3′
m06j06	5′ UGAUUUGAAGUGAUGGAGUUGG 3′
m07j07	5′ GUGAUCAGACUCAAUACGAAU 3′
m10j11	5′ GAAGUGAUGGAGUUGGCCAGCC 3′
m14j12	5′ CCUAAAGACUGACAAUAUCCAAAAA 3′
m15j18	5′ AUUCGUAUUGAGUCUGAUCACAC 3′
m18j13	5′ UUAAUUAAAGACUUCAAGC 3′
m19j09	5′ GUCCCUGGGCAGCUCCGUAUA 3′
m09	5′ GGAUAUUGUCAGUCUUUAG 3′
j05	5′ CUCACUCUGCUGCGAGAAGUG 3′
RP	5′ NNNNNNNNNNNNNNNNNNNNNNN 3′
C1	5′ CCCUCGCUACUGAAACAUAUUAAA 3′
C2	5′ GGUAAUCUUCACAAACAUCUUUUU 3′
C3	5′ UACUUCCGUCAUGUUUUUGUCCUC 3′
polyU	5′ UUUUUUUUUUUUUUUUUUUUUUUUU 3′

Table S4. List of individual sequences assayed biochemically. The name of the sequence identifies the peak to which it belongs; ‘m’ and ‘j’ refer to the two replicate experiments, and the number is the fitness rank of the peak in that experiment. For example, m01j03 means the sequence represents the peak from the most abundant peak from replicate ‘m’, which was also the 3rd most abundant peak from replicate ‘j’. Note that for m09 and m18j13, the sequence assayed biochemically is slightly different from the sequence ultimately identified as having the highest fitness within the peak (Figure 2).

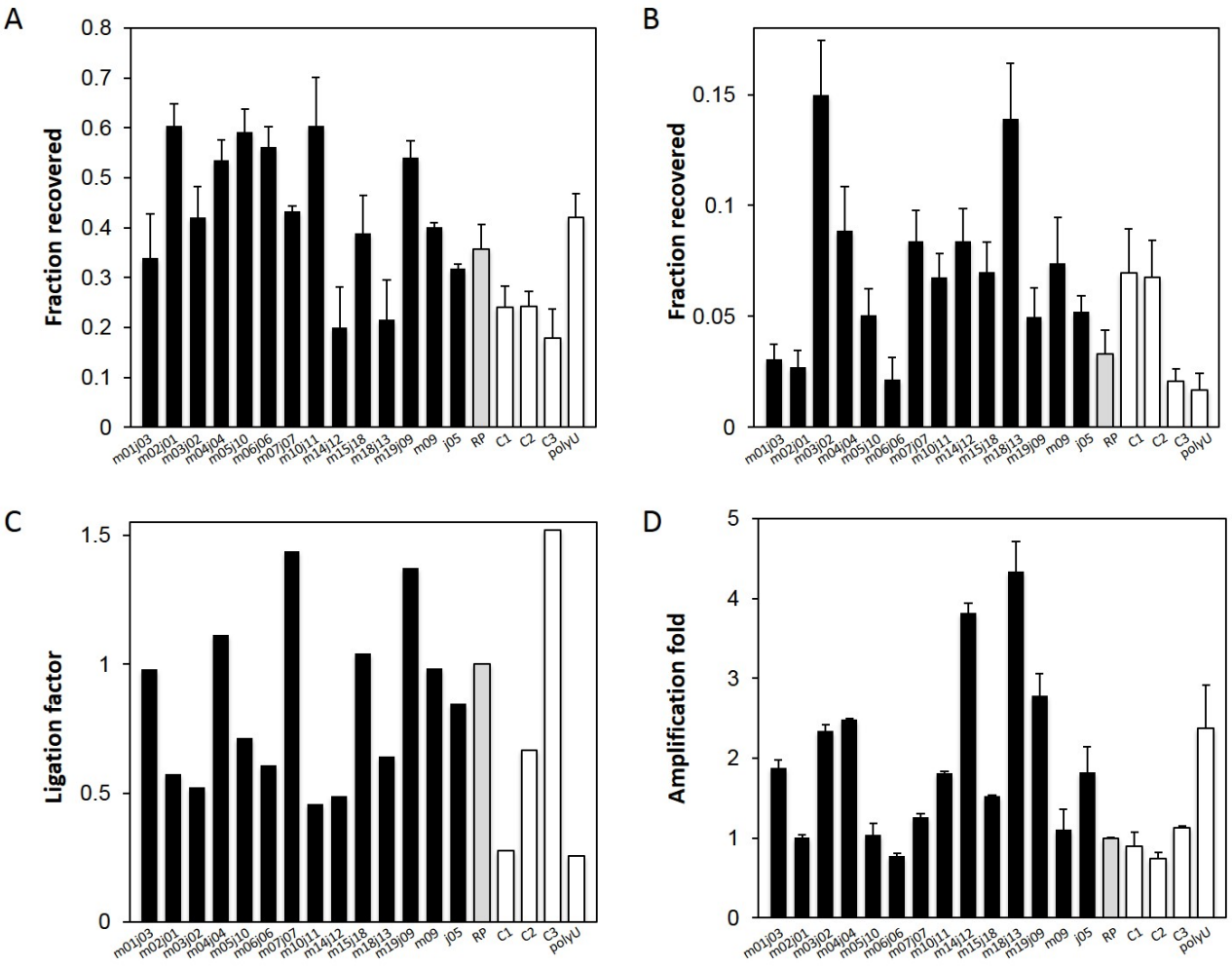


Figure S13. Efficiency of the different steps of the selection and preparation of the samples prior to deep sequencing. The efficiency of each sequence through each step was tested by: measuring the fraction of each sequence recovered in the flow-through after binding to agarose resin in a background of random pool sequences (**A**); determining the fraction of each sequence recovered after binding to GTP-agarose resin, washing and elution with free GTP in a background of random pool sequences (**B**); the factor applied to each of the sequences to correct for the ligation bias, calculated from our analysis of the random pool (**C**); and the relative PCR amplification of each sequence compared to the random pool (**D**). See Methods section for experimental details. Error bars show the standard deviation of at least 4 independent measurements.

sequence	yield from agarose-resin (%)	std. dev. (%)	yield from GTP-agarose (%)	std. dev. (%)	relative PCR yield	std. dev.	relative ligation efficiency	relative composite score	std. dev.
m01j03	33.87	8.85	3.08	0.69	1.88	0.10	0.98	0.019	0.007
m02j01	60.35	4.42	2.72	0.74	1.01	0.03	0.57	0.009	0.002
m03j02	42.10	6.11	15.00	2.45	2.34	0.08	0.52	0.077	0.009
m04j04	53.54	3.99	8.88	1.97	2.48	0.02	1.11	0.131	0.034
m05j10	59.17	4.65	5.04	1.21	1.04	0.14	0.71	0.022	0.005
m06j06	56.20	3.99	2.17	0.96	0.77	0.03	0.61	0.006	0.002
m07j07	43.24	1.14	8.38	1.41	1.26	0.04	1.44	0.065	0.016
m10j11	60.35	9.77	6.76	1.09	1.81	0.02	0.46	0.034	0.004
m14j12	19.93	8.23	8.39	1.46	3.82	0.12	0.49	0.031	0.007
m15j18	38.94	7.46	6.99	1.38	1.52	0.01	1.04	0.043	0.012
m18j13	21.58	7.92	13.93	2.50	4.33	0.38	0.64	0.084	0.022
m19j09	54.03	3.35	4.98	1.31	2.78	0.28	1.37	0.103	0.041
m09	40.13	0.95	7.40	2.08	1.11	0.25	0.98	0.032	0.011
j05	31.81	0.93	5.23	0.68	1.82	0.32	0.85	0.026	0.005
RP	35.78	4.93	3.33	1.06	1.00	0.00	1.00	0.012	0.004
C1	24.00	4.22	6.97	1.97	0.90	0.17	0.28	0.004	0.000
C2	24.29	3.01	6.78	1.64	0.74	0.07	0.67	0.008	0.002
C3	17.91	5.73	2.08	0.54	1.13	0.02	1.52	0.006	0.004
polyU	41.99	4.88	1.70	0.75	2.38	0.53	0.26	0.004	0.001

Table S5. Yields of individual steps for selected sequences and random pool (RP). Standard deviations (light gray columns) are given from replicate experiments for the measurement (dark gray columns) to the left in the table. Relative ligation efficiency was estimated from our study of ligation bias. The relative composite score is the product of the individual measurements. Also see Figure S13.

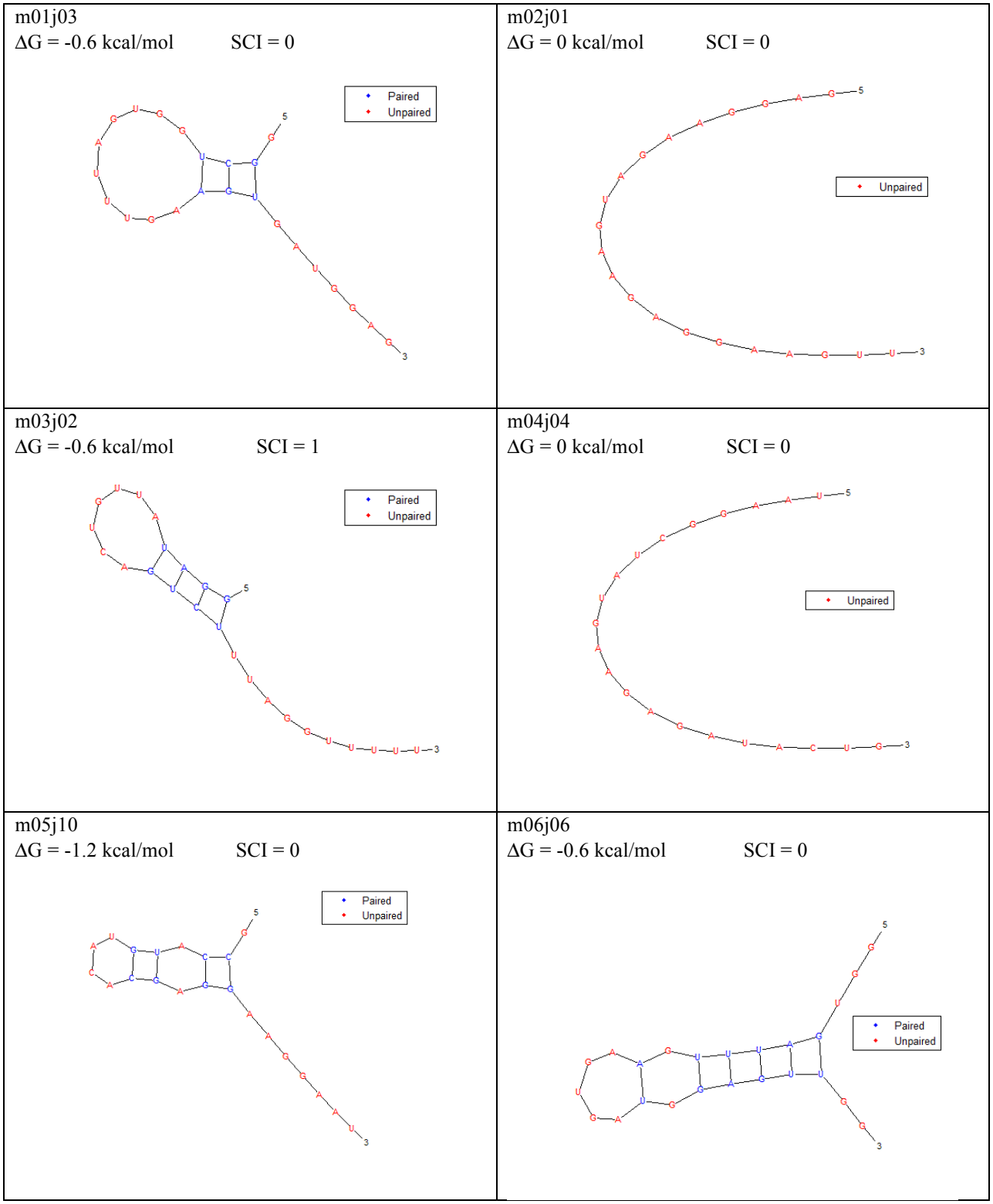
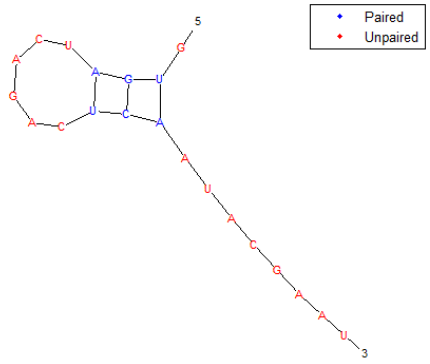
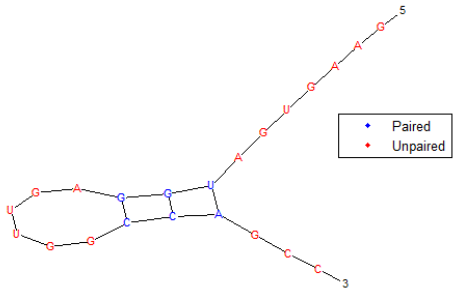


Figure S14. Secondary structures and structural conservation index predicted for peak sequences.

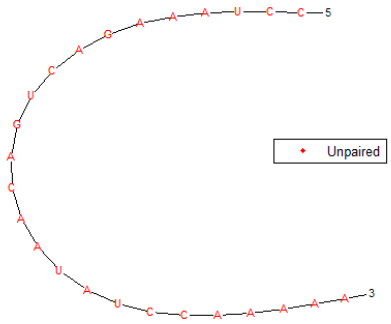
m07j07
ΔG = -0.4 kcal/mol SCI = 0



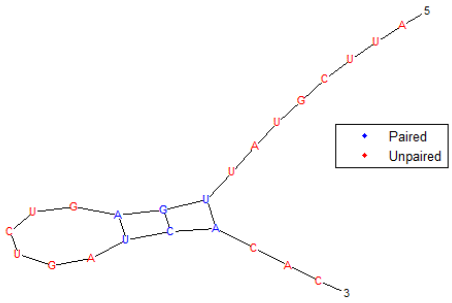
m10j11
ΔG = -2.4 kcal/mol SCI = 1



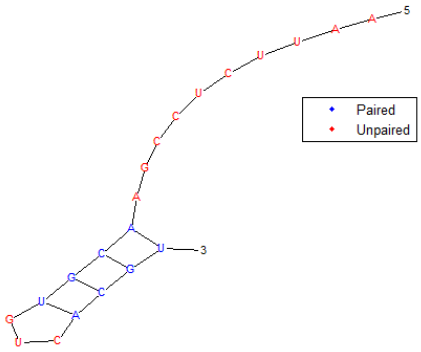
m14j12
ΔG = 0 kcal/mol SCI = 0



m15j18
ΔG = -0.9 kcal/mol SCI = 1



m16j20
ΔG = -1.4 kcal/mol SCI = 0



m17j08
ΔG = -1.9 kcal/mol SCI = 0.71

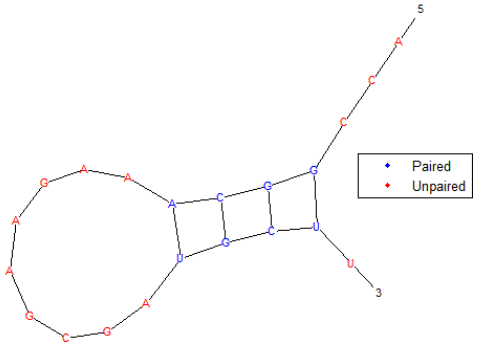
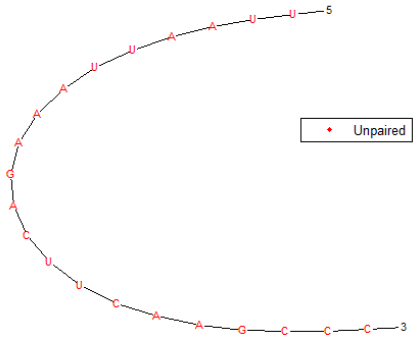
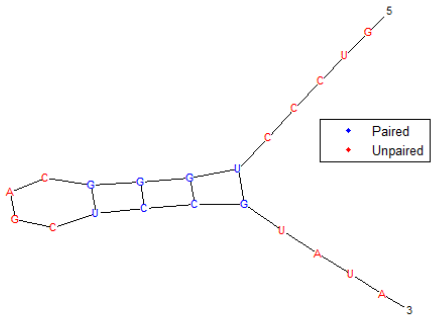


Figure S14 continued.

m18j13
ΔG = 0 kcal/mol SCI = 0



m19j09
ΔG = -3.7 kcal/mol SCI = 0.71



m20j22
ΔG = -1 kcal/mol SCI = 0.7

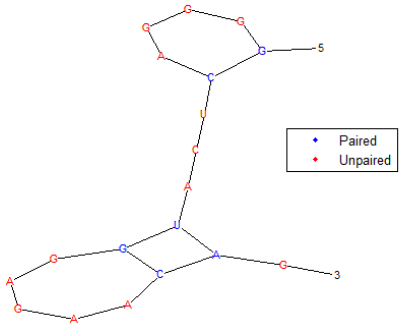


Figure S14 continued.

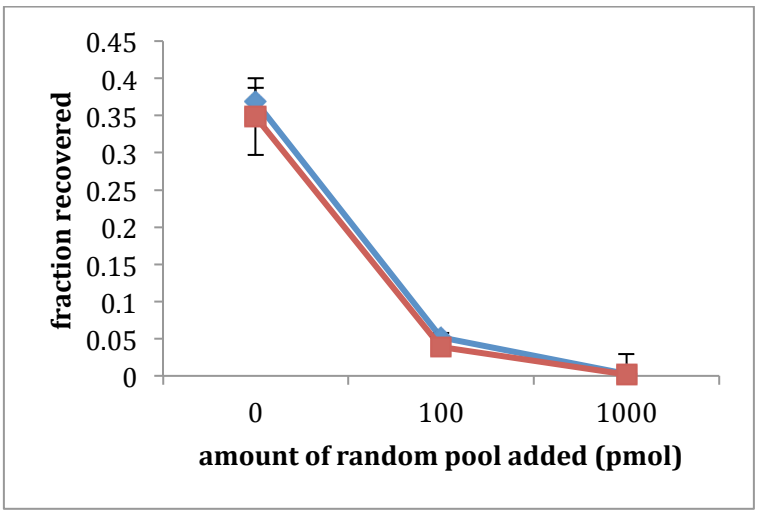


Figure S15. Dependence of recovery yield on total RNA loaded onto the column. The binding and release from the GTP-agarose resin was tested in the presence of different amounts of random pool. 100 pmol of radiolabeled RNA sequence (red: random pool; blue: j05) were mixed with 0, 100 or 1000 pmol of cold random pool and incubated with the resin following the protocol for binding assays described previously. The recovery of radiolabeled material is expressed as the fraction of cpm eluted by free GTP over the total amount loaded in the column; standard deviations are given by the error bars.

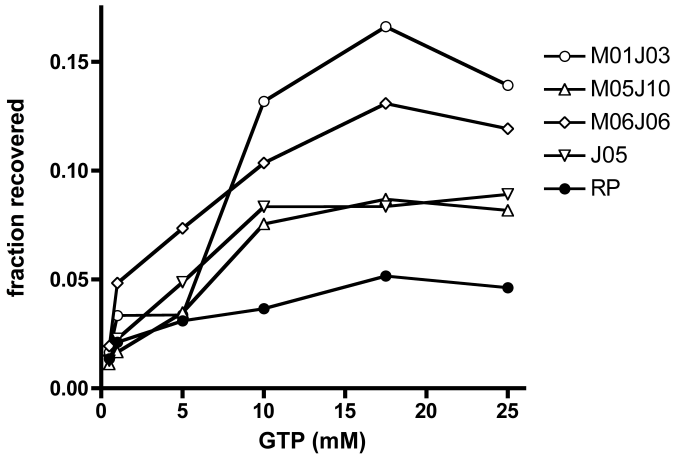


Figure S16. Elution yield from GTP-agarose column using different concentrations of GTP. Radiolabeled sequences were mixed with 100 pmol of the same unlabeled sequence before loading on the column.

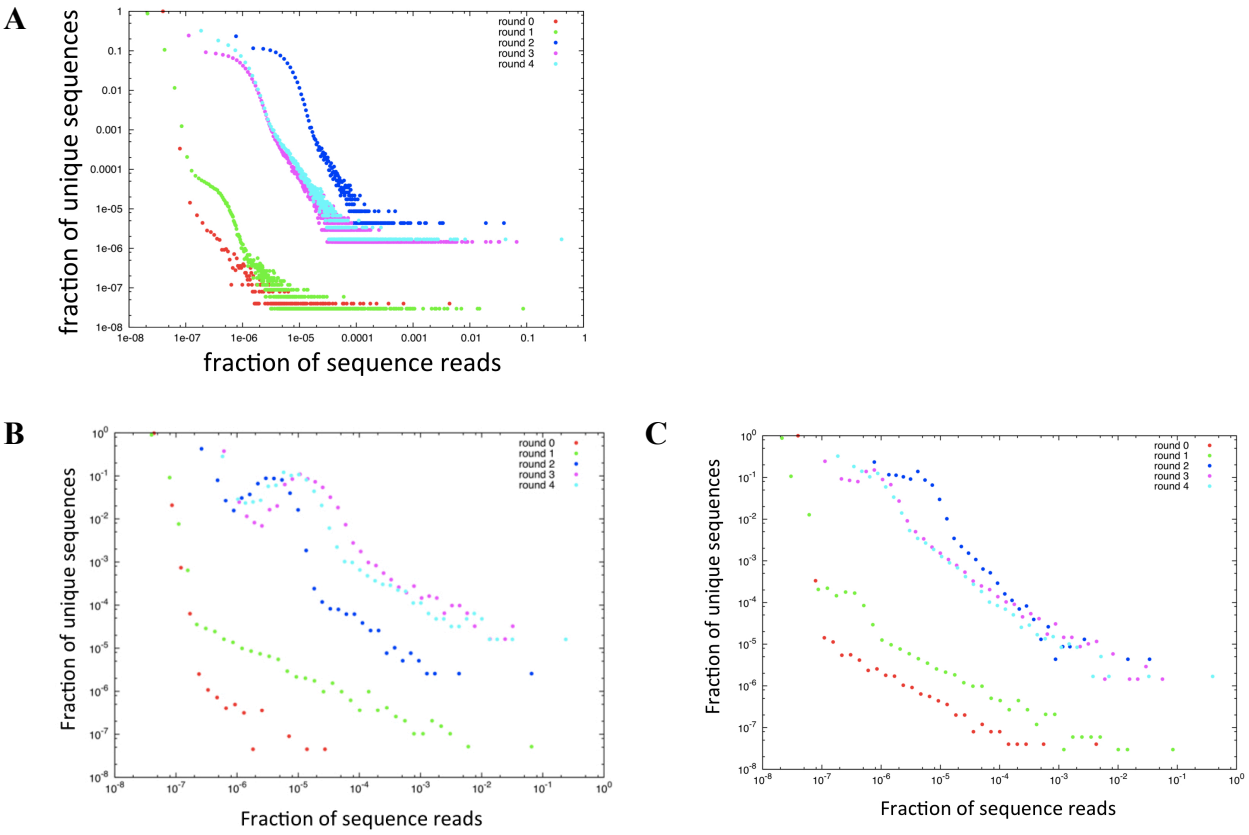


Figure S17. Enrichment histogram for raw data. Histogram of the fraction of unique sequences, illustrating enrichment, for the raw, uncorrected data. Data from another replicate is shown in panel (**A**), analogous to Figure S3B showing one replicate; no binning was applied. Panels (**B**) and (**C**) (two replicates) show the same data binned along the *x*-axis on a log scale. 'Round 0' refers to the initial pre-selection pool.

Text S1. Additional Methods.

High-throughput sequencing (HTS). Aliquots obtained in the selection were tagged with different 3′ adapters in ligation reactions. Samples were mixed with 5 μM adapter sequence for 6 h at 20°C in a 20 μL ligation reaction using 10 U T4 RNA ligase (New England Biolabs). The adapters used were as follows (lower case indicates RNA, upper case indicates DNA):

Initial pool tag (TAG1): 5′-ucgTGTCGTATGCCGTCTTCTGCTTGTddC

Round 1 tag (TAG2): 5′-ucgCATCGTATGCCGTCTTCTGCTTGTddC

Round 2 tag (3-TAG1): 5′-ucgTACTCGTATGCCGTCTTCTGCTTGTddC

Round 3 tag (3-TAG2): 5′-ucgACATCGTATGCCGTCTTCTGCTTGTddC

Round 4 tag (3-TAG3): 5′-ucgCTATCGTATGCCGTCTTCTGCTTGTddC

Products of ligation were purified by denaturing PAGE, ethanol precipitated, phosphorylated by PNK (New England Biolabs), extracted with phenol and chloroform, and gel purified again. The resulting preparations were ligated to the 5′ adapter (5′-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACaguccgacgauc; 5 μM) in a 20 μL reaction with 10 U of T4 RNA ligase for 6 h at 20°C. Samples obtained after this step were gel purified, ethanol precipitated and denatured for 2 min at 80°C. Retrotranscription was performed for 1 h at 48°C in a 20 μl reaction using 200 U of SuperScript III RT (Invitrogen), dNTPs (0.5 mM), 10 mM DTT and 5 μM of primer RT3 (5′-CAAGCAGAAGACGGCATACGA). The resulting cDNA was PCR amplified for 30 cycles using primers RT3 and PCR5 (5′-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA). Products were gel-

purified and their concentration determined using the fluorescent Qubit assay (Invitrogen). Concentrations were also estimated by comparison to the Low Molecular Weight DNA Ladder (New England Biolabs) on a gel stained with Sybr Gold (Invitrogen); these estimates were usually within ~2-fold of the Qubit assay. Gel image acquisitions were performed using a Typhoon 9400 scanner (GE Healthcare). Samples were submitted for analysis on an Illumina HiSeq 2000 for short, single end reads (core facility at the FAS Center for Systems Biology at Harvard). Sequencing of the random pool and unrelated samples using the same reagents used for sequencing rounds 3 and 4 did not yield sequences identified as peaks in these experiments, indicating that peak sequences were not artifacts of sample preparation.

Binding assays. RNA oligonucleotides were obtained from UCDNA (Calgary, Canada) at 1 μmol scale synthesis. Oligos were gel purified upon arrival. 100 pmol of each oligonucleotide were radioactively labeled as explained previously, ethanol precipitated, washed thoroughly with 100% ethanol and resuspended in 70 μL of water. Gel electrophoresis in pilot studies indicated that ^{32}P-γ-ATP was removed through this procedure. 100 pmol of either unlabeled RNA sequence (identical to the labeled sequence) or of random pool RNA was mixed with 1 μL of the radiolabeled preparation (~50,000 cpm). 4 μL of 5x binding buffer was added with water for a total volume of 20 μL (5 μM cold RNA). The mixture was heated at 65°C for five minutes, cooled for 5 min on the benchtop at room temperature, loaded onto 20 μL of the appropriate agarose resin in spin column format and incubated for 15 min at room temperature. To estimate the yield of recovery of the negative selection, the samples were loaded into the agarose resin as indicated and the flow-through of the column was recovered in one centrifugation. The amount of radioactivity in the flow-through fraction was determined using a Beckman Coulter LS6500

scintillation counter. To estimate yield of survival during positive selection on GTP-agarose resin, an analogous procedure was followed, but the flow-through was discarded and the resin washed 10 times with 50 μL of 1x binding buffer. Elution was performed in ten steps using 20 μL of 1x binding buffer supplemented with GTP at the desired concentration (0.1-25 mM). The elution profile of the sequences was determined by measuring the amount of radioactivity present in the eluates by scintillation counting. The values obtained were summed and expressed as a fraction of the initial counts loaded into the resin.

Assay for efficiency of PCR amplification. To mimic the cDNA used as template for PCR amplification prior to deep sequencing, template oligonucleotides were designed as the reverse complement of the RNA peaks identified in the selection, flanked by the two adapters used for sequencing, following the scheme 5′-CAAGCAGAAGACGGCATACGA-*sequence*-GATCGTCGGACTGTAGAACTCTGAAC (IDT DNA technologies). The templates were gel-purified and PCR was carried out using primers PCR5 and RT3 (0.2 μM each). In particular, 2 and 20 ng of each template were dissolved in 20 μl of water and amplified in 30 cycles of PCR with 0.2 mM each dNTP and 0.2 U of Taq DNA Polymerase in the standard Taq Mg-free buffer supplemented with 1.5 mM $MgCl_2$ (New England Biolabs). The amount of product formed was measured from denaturing PAGE, comparing the intensity of PCR product bands stained with SYBR gold (Invitrogen) to a band of known concentration corresponding to 100 ng of the template sequence. The intensity of the PCR product band was generally close to that of the standard bands. Gel analysis was performed using the Image-J software package.

Estimation of fitness from observed sequence frequency. All algorithms for preparing, analyzing, and sorting the sequence data were performed in Fortran (code available on request). Sequences related to the flanking adapters (within edit distance of 1) were assumed to be artifacts and removed from the analysis (see SI Appendix, Table S6 for list of sequences). For analysis of the initial pool, sequences longer than 12 nucleotides were identified for further analysis. For post-selection rounds, only sequences of length 12-33 were further analyzed; most were of length 17-24 (see SI Appendix, Fig. S18). For sequences in the post-selection analysis, the fitness was estimated from its observed frequency in the sequence reads, corrected for three effects.

First, HTS is known to introduce sequencing errors at a rate of approximately 1% per base, so an abundant sequence could produce a 'halo' of closely related sequences due to sequencing errors alone. We made a first-order correction on the number of observed sequences to account for this effect by assuming a 1% chance of error at each position. Starting with the most abundant sequence in the pool, an expected number of copies (n_{exp}) was calculated for each of its related sequences. This number was then compared to the observed abundance (n_{obs}) for each related sequence. For a given sequence, if $n_{exp} < n_{obs}$, then we recorded the corrected number of copies (n_c) for that sequence as $n_c = n_{obs} - n_{exp}$; if $n_{exp} > n_{obs}$, then that sequence was removed from the analysis. Additionally, the corrected copy number (n_c) for the most abundant sequence was obtained by $n_c = n_{obs} + \Sigma_j (n_{exp})$ where j = the set of related sequences. This process was repeated for the next most abundant sequence, and so forth, until the correction for every sequence had been calculated. During each step of this process for a given sequence, n_c either increased or decreased; if it decreased such that $n_c < n_{obs}$ after a given number of steps, then that sequence was removed from the analysis. The abundances were updated upon the completion of

this process. Once the copy number for each sequence was corrected as described, a frequency (f_{c1}) was assigned to each sequence by dividing its corrected abundance by the updated total number of copies of all sequences. This heuristic algorithm was able to recover starting frequencies using simulated data for two control scenarios, 50 related sequences (plus errors) and 5 unrelated sequences (plus errors) that spanned several orders of magnitude in fitness.

Second, f_{c1} could be skewed by differences in the efficiency of ligation to the adapter sequences. We noticed that the overall nucleotide composition at the 5′ and 3′ ends (especially the first and last 6 nucleotides) in the initial pool differed substantially from the composition in the middle, suggesting a bias from ligation. In order to estimate this bias, we modeled oligonucleotide synthesis based on a framework of Markov conditional probabilities, which were inferred from the sequence reads obtained from the initial pool. In our model of synthesis, sequences are built from 3′ to 5′ (as they are in the actual synthesis), and the rate of addition of a particular nucleotide (A, C, G, or U) depends on the identity of the incoming nucleotide and on the last two positions at the 5′ end of the growing sequence (64 rate constants). We found that considering the last two 5′ nucleotides was sufficient to construct a successful model that accurately reproduced the most relevant statistical features of the pool. The first nucleotide was assumed to be present in even proportion (25% of each nucleotide, which is consistent with the indications of Core DNA services, who synthesized the pools); the relative rate of addition of the second nucleotide was estimated from the conditional probability of each nucleotide following the first (16 rate constants, estimated in a fashion analogous to the estimation of the 64 rate constants). The 64 relative rates (for the 3rd position onward) were estimated from the corresponding conditional probabilities found in the HTS data from the initial pool, using positions 7 to L-6 within the sequence. Ignoring the ends of the sequences avoided the

confounding effect of ligation bias. This process allowed us to calculate the probability of any sequence appearing in the synthesized pool (f_i), as well as the probability of any 6-mer subsequence appearing at the 3′ or 5′ end. By comparing the calculated probabilities for the 3′ and 5′ 6-mer subsequences with the frequencies at which those 6-mers appeared at the termini of sequences in the initial pool, we calculated a correction factor (c_{lig}) for each sequence to account for the bias due to ligation. Applying this correction factor yielded the ligation-corrected frequency (f_{c2}) according to the equation, $f_{c2} = f_{c1} \cdot c_{lig}$. Third, we computed the relative fitness (f_r, relative to other sequences in the experiment) by dividing each sequence's ligation-corrected frequency by the probability of that sequence appearing in the initial pool, as given by the model of synthesis above, $f_r = f_{c2} / f_i$. This accounts for the fact that the abundance of a sequence in the initial pool could be higher or lower than average, which would otherwise result in a corresponding overestimate or underestimate of fitness.

Calculation of functional information. Sequences within a given peak were aligned using Matlab R2011b (Bioinformatics Toolbox) or by the R-Coffee online server, with manual corrections for the larger peaks. Gaps were sometimes introduced during the alignment process. The occurrence of each nucleotide was recorded at each position in the aligned sequence set, resulting in a 4-by-L matrix where each row represents one of the 4 nucleotides (A, C, G, U) and each column gives the number of times the base appeared at the corresponding position in the sequence set. The frequency of base i (F_i) at each position was calculated, ignoring gaps, such that the sum of each column was 1. The matrix of frequencies was used to calculate the Shannon uncertainty at each position according to the equation $H = -\Sigma_i [F_i \cdot \log_2(F_i)]$ where i = A, C, G, and U. Each element in the resulting vector of H values (i.e. the Shannon entropy at each

position in the sequence set) was then subtracted from the maximum per-site information content (2 bits), to yield the information content at each site. The information content at each site was weighted by the occupancy of the position (i.e., the proportion of sequences containing any base rather than a gap). To confirm that the specific alignment, treatment of gaps, and the possibility of undersampling of peaks were not major sources of error, we used a Monte Carlo approach to subsample the data for each peak and calculate the functional information content as described. For each peak, 7 sequences were randomly chosen with replacement and aligned using Matlab R2011b (Bioinformatics Toolbox; no manual correction needed given the small sample size). This process was repeated 100 times for each peak and the average and standard deviation were calculated.

Motif analysis and secondary structure prediction. The Gibbs Motif Sampler (1) was used to detect motifs common to multiple fitness peaks. The highest fitness sequence of each peak was used as input for the comparison. The searched motif length was varied from 7 to 9 bases. Secondary structures and the corresponding minimum free energies for peak sequences were predicted using the *rnafold* function in the Matlab 2011b Bioinformatics toolbox, based on a nearest-neighbor thermodynamic model (2, 3). The structure conservation index (SCI, i.e., the ratio of a peak's consensus minimum free-energy (MFE) to the average of individual sequence MFEs) for each peak was calculated using RNAz (4). Sequences with fitness <10% of the peak sequence were generally not included in SCI calculations, with the proviso that the top five sequences of each peak were always included.

TCGCTATCGTATGC

AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGACGATC

TCGTGTCGTATGCCGTCTTCTGCTTGT

TCGCATCGTATGCCGTCTTCTGCTTGT

TCGTACTCGTATGCCGTCTTCTGCTTGT

TCGACATCGTATGCCGTCTTCTGCTTGT

TCGCTATCGTATGCCGTCTTCTGCTTGT

Table S6. Sequences removed from the analysis. These sequences are related to the adapter sequences and were assumed to be artifacts of the HTS process. Sequences containing a 13 nucleotide subsequence derived from the above, including sequences within an edit distance of 1, were removed from the analysis.

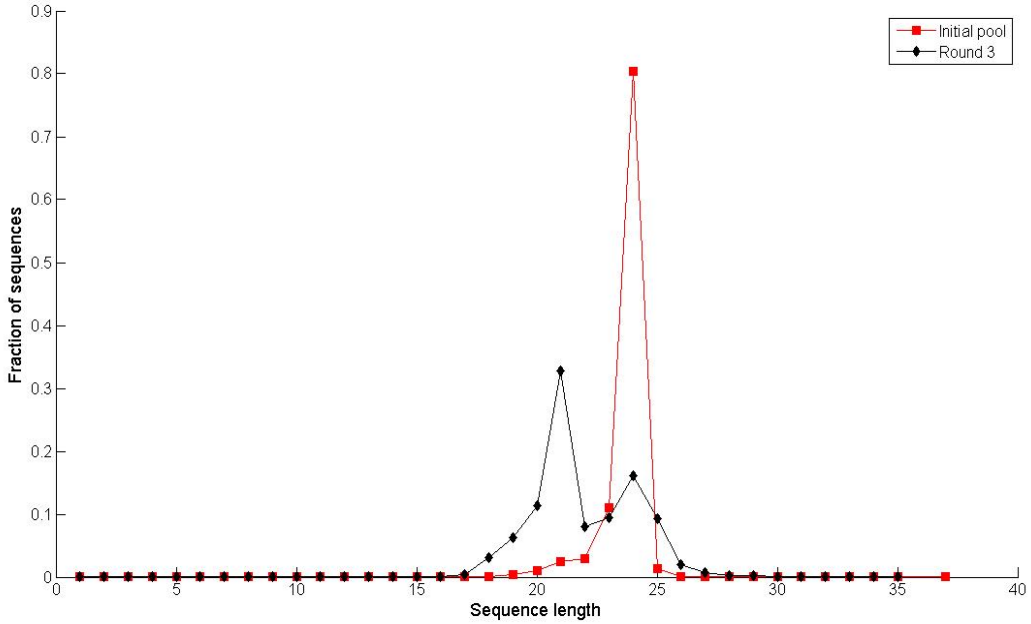


Figure S18. Distribution of sequence lengths. Shown are the fraction of sequences of a given length in the initial random pool and after 3 rounds of selection.

SI References

1. Thompson WA, Newberg LA, Conlan S, McCue LA, & Lawrence CE (2007) The Gibbs Centroid Sampler. *Nucleic Acids Res* 35(Web Server issue):W232-237.
2. Mathews DH, Sabina J, Zuker M, & Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288(5):911-940.
3. Wuchty S, Fontana W, Hofacker IL, & Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49(2):145-165.
4. Washietl S, Hofacker IL, & Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 102(7):2454-2459.