Enhanced Functional Potential of Nucleic Acid Aptamer Libraries Patterned to Increase Secondary Structure

Karen M. Ruff, Thomas M. Snyder, and David R. Liu*

Supporting Information

Library Design

After choosing a pattern in which purine-rich positions pair with pyrimidine-rich positions (see the main text) we next considered possible arrangements of purines and pyrimidines in the library. One possibility would be a long stretch of purines that would find and base pair with a long stretch of pyrimidines (R_nY_n). A second possibility would be a track of alternating purines and pyrimidines that would find and base pair with another track of alternating purines and pyrimidines ($(RY)_n$). One thousand randomly chosen sequences matching these two possible patterns were generated and folded *in silico* using the Oligonucleotide Modeling Platform (OMP, DNA Software), and the average folding energies of the two patterns were compared (Figure S1). The $(RY)_n$ library members exhibited a significantly higher average predicted folding energy than R_nY_n library members, possibly because any RY track can pair with any other RY track, while in an R_nY_n sequence, purine stretches can only pair with pyrimidine stretches, and not other purine stretches. We therefore chose to use library patterns in which purines and pyrimidines alternate.

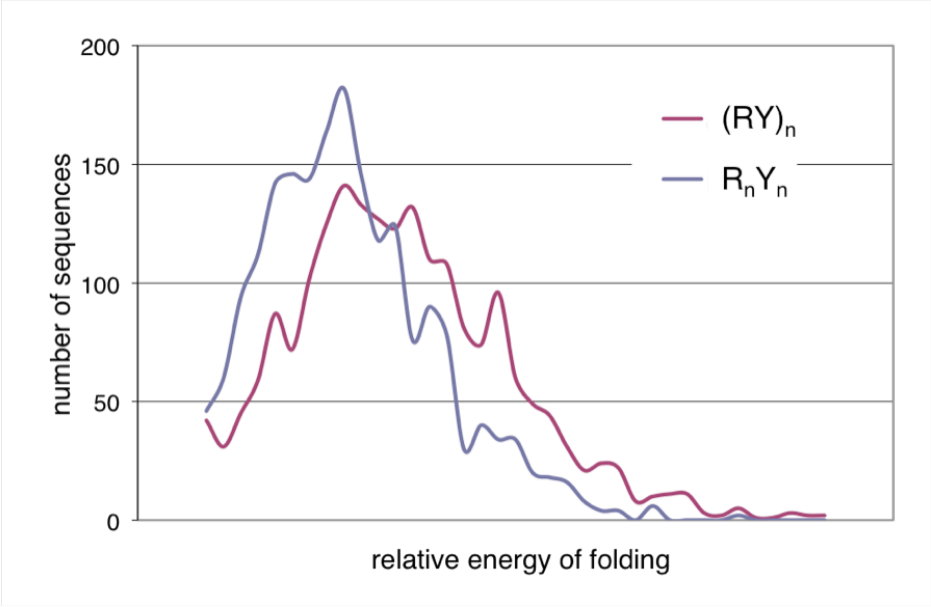

Figure S1. Predicted folding energies of two possible patterns. $(RY)_n$ = $(RY)_4NNN(RY)_5NNNN(RY)_5NNN(RY)_5NNNN(RY)_4$; $R_nY_n = R_8NNNY_{10}NNNNR_{10}NNNY_{10}NNNNR_8$

We focused on library designs that contain several alternating patterned and N_m stretches to maximize the ways in which each patterned region can interact with multiple other patterned or random regions. Sixty-base variable regions were chosen to allow the inclusion of multiple pattern and N_m regions while still maintaining the ability of the libraries to be synthesized as a single degenerate oligonucleotide without requiring enzymatic ligation. In addition, Knight, Yarus, and coworkers found that the optimal variable region length for selecting a simple, well-studied isoleucine aptamer was between 50 and 70 nucleotides.[1]

Choice of Primer-Binding and Tag Sequences

Following the design of the variable regions of the N_{60}, R*Y*, and RY libraries, we calculated the predicted average folding energy of 5,000 arbitrarily chosen members of each library, without any constant sequences. Next we identified a 6-base tag sequence for each library that did not perturb the relative folding energies of the libraries or their spreads (standard deviations). These tags were chosen to be cleavable by restriction endonucleases. Several possible tag sequences were screened computationally to identify tags that would not affect the predicted folding energy distribution of the libraries. Similarly, primer sequences were chosen that meet the usual requirements of similar and sufficiently high melting temperatures and no mutual or self-complementarity at their 3' ends, and that also preserve the relative folding energies of the libraries (Figure 1 in the main text and Table S1). The final primer and tag sequences are given below. Note that the constant sequences increased the folding energy of all libraries because the extra sequences increase the opportunities to form internal base pairing, even though the predicted relative folding energies of the three libraries and the standard deviation of these energies remain similar to that of the libraries lacking these constant regions.

library	average predicted ΔG of folding ± standard deviation (kcal/mol)		
	no primer	streptavidin primer set	IgE/VEGF primer set
N	-11.2 ± 3.2	-12.0 ± 3.3	-16.3 ± 3.3
R*Y*	-13.6 ± 3.5	-14.3 ± 3.6	-20.2 ± 3.8
RY	-16.1 ± 3.7	--	-22.6 ± 4.0

Table S1. Predicted folding energies of libraries with and without primers.

Streptavidin selection library sequences:
N_{60} (5'-CGGTGCTCCTTGCGGTC-GGATCC-N_{60}-GCACCAGACCACACGG), where N = 25:25:25:25 A:C:G:T
R*Y* (5'-CGGTGCTCCTTGCGGTC-CAGCTG-$(R*Y*)_4N_4(R*Y*)_5N_3(R*Y*)_5N_4(R*Y*)_5N_3(R*Y*)_4$GCACCAGACCACGACGG), where N = 25:25:25:25 A:C:G:T, R*= 45:5:45:5, and Y*= 5:45:5:45

IgE and VEGF selection library sequences:
N_{60} (5' TGTCGCTGCGTCGCCTG-GGATCC-N_{60}-CACCGGAAGACGCACGC), where N is a mixture that couples at 25:25:25:25 A:C:G:T
R*Y* (5' TGTCGCTGCGTCGCCTG-CAGCTG-$(R*Y*)_4N_4(R*Y*)_5N_3(R*Y*)_5N_4(R*Y*)_5N_3(R*Y*)_4$-CACCGGAAGACGCACGC), where N = 25:25:25:25 A:C:G:T; R* = 45:5:45:5; Y* = 5:45:5:45
RY (5' TGTCGCTGCGTCGCCTG-GCTAGC-$(RY)_4N_4(RY)_5N_3(RY)_5N_4(RY)_5N_3(RY)_4$-CACCGGAAGACGCACGC), where N = 25:25:25:25 A:C:G:T; R = 50:0:50:0; and Y = 0:50:0:50

DNA Library Synthesis and Analysis

Libraries were synthesized using phosporamidite mixtures as described in the main text. Because different bases couple at different efficiencies, mixtures with molar ratios based on desired target ratios were not sufficient to achieve desired amounts of incorporation of the bases into the libraries. Instead, mixtures were optimized empirically using an HPLC assay that quantifies dimers of form [5'-mix-C]. After the mixtures were adjusted to achieve the desired coupling ratios, the N_{60}, R*Y*, and RY libraries were synthesized. Approximately 30 members of each library were cloned and sequenced. The target nucleotide ratios, the ratios observed by HPLC, and the ratios observed by DNA sequencing

at each type of position in all three libraries were in good agreement (Figure S2). The ratios observed by DNA sequencing were used to calculate binding motif probabilities and the folding energies of the libraries used in the selections.

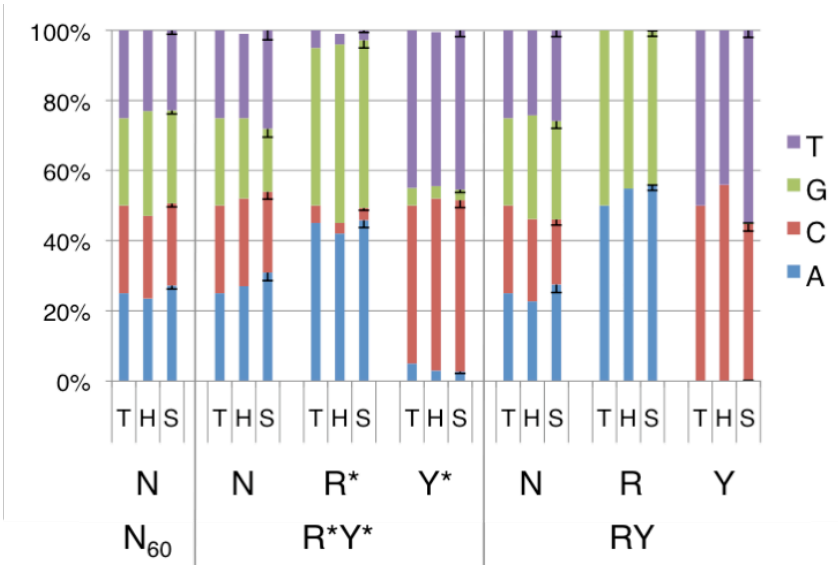

Figure S2. Base ratios for IgE and VEGF starting libraries. Target (T), HPLC–determined (H), and sequencing-determined (S) nucleotide ratios for each type of position in all three libraries are shown. Error bars reflect the standard error from sequencing-determined ratios.

We generated 3,000 randomly chosen members from each library using the observed nucleotide ratios and determined their predicted energy of folding with the Oligonucleotide Modeling Platform (DNA Software) (Table S2). The N_{60} library base ratios and average energy were very close to the intended values. The R*Y* library had slightly lower incorporation of off-pattern bases than desired (3% instead of 5% each off-pattern base), but the A:G and C:T ratios were balanced and the average predicted folding energy was similar to the intended value. The experimental RY library had a significantly lower average predicted folding energy than a theoretical library with the intended ratios ($\Delta\Delta G$ = 2 kcal/mol), such that its average was the same as that of the R*Y* library (standard deviation of the mean = 0.1 kcal/mol). We believe that the slightly higher abundance of A and T lowered the overall average folding energy. Because the experimental RY library is predicted to have the same folding energy average and standard deviation as that of the R*Y* library, they can be directly compared to determine the importance of incorporating a small fraction of off-pattern bases at each patterned position.

	energy of folding (kcal/mol)	
library	average	standard deviation
N_{60}	-16.2	3.3
R*Y*	-20.6	3.6
RY	-20.6	3.6

Table S2. Experimental folding energies of IgE and VEGF starting libraries.

The starting library for the streptavidin selection used different primer-binding sequences and was synthesized separately with different phosphoramidite mixes. The sequences of 25 clones from the

input library were aligned with the pattern and the experimental ratios of the bases at each position were determined (Table S3).

	A	C	G	T
N	26	28	18	28
R*	52	4	41	3
Y*	2	40	2	56

Table S3. Observed base ratios (in percent) in the starting library for the streptavidin selection.

While some of these variations in composition were wider than desired, the patterning was successful with predominantly purines/pyrimidines at the appropriate positions in the structured library. We remodeled the folding energies of the N_{60} and R*Y* libraries using our estimates of the compositions given above. The average energy of our synthesized N_{60} library is -9.8 ± 3.2 kcal/mol, and the average energy of R*Y* is -13.0 ± 3.3 kcal/mol. Thus the average predicted folding energy difference between the libraries was 3.2 kcal/mol, one standard deviation of the N_{60} library (comparable to the one standard deviation difference in average energy of the theoretical libraries).

PCR Amplification Efficiency Tests

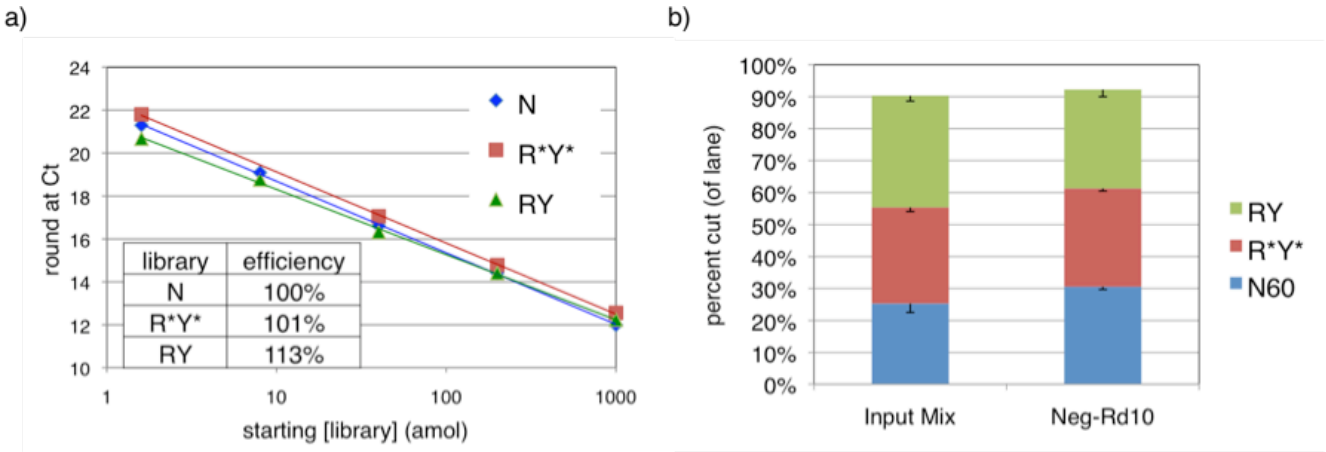

Figure S3. DNA libraries amplify with comparable efficiency when tested (a) separately under the same conditions with common primers, and (b) in one solution carried through ten successive cycles of 128-fold dilution and PCR amplification. Error bars represent the standard error of three digestions. For (a), we generated standard curves for each library by qPCR, fit the C_T values to a line based on the natural log of the known starting amount of DNA, and determined the amplification efficiencies according to the formula, efficiency = [e^(-1/slope) – 1] * 100%. The three libraries amplify with similar efficiencies.

Restriction Digestion of Tag Sequences

We used tag digestions to determine the bulk library ratios in our selection pools. These digestions were performed under conditions that resulted in complete cleavage of the target library, but no off-target cleavage of the other two libraries (Figure S4).

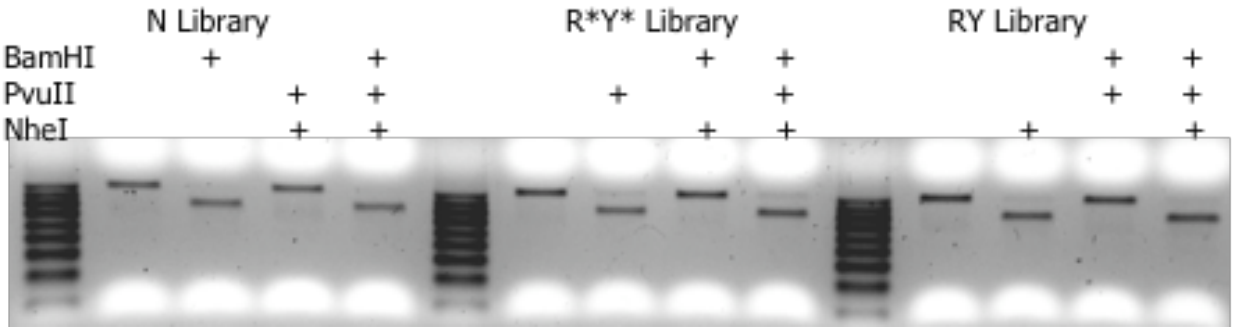

Figure S4. Library digestions by tag-specific endonucleases. For each library, the corresponding enzyme causes complete digestion, while a combination of the other two enzymes results in no digestion.

Analysis of Selection Results

Library Ratios by Restriction Enzyme Digestion

We digested the pool material from each round of the three selections under the same conditions used in Figure S4. The library ratios across each of the three selections are depicted in Figure S5. For both the streptavidin and VEGF selections, the presence of significant uncut DNA especially in later rounds (representing up to ~30% of either selection's pool) obscured interpretation. We attribute the uncut DNA to mutations in tag sequences, as was explicitly observed in sequences of R*Y*-derived clones in both the streptavidin and VEGF selections, or to inefficient digestion of particularly well-folded DNA sequences.

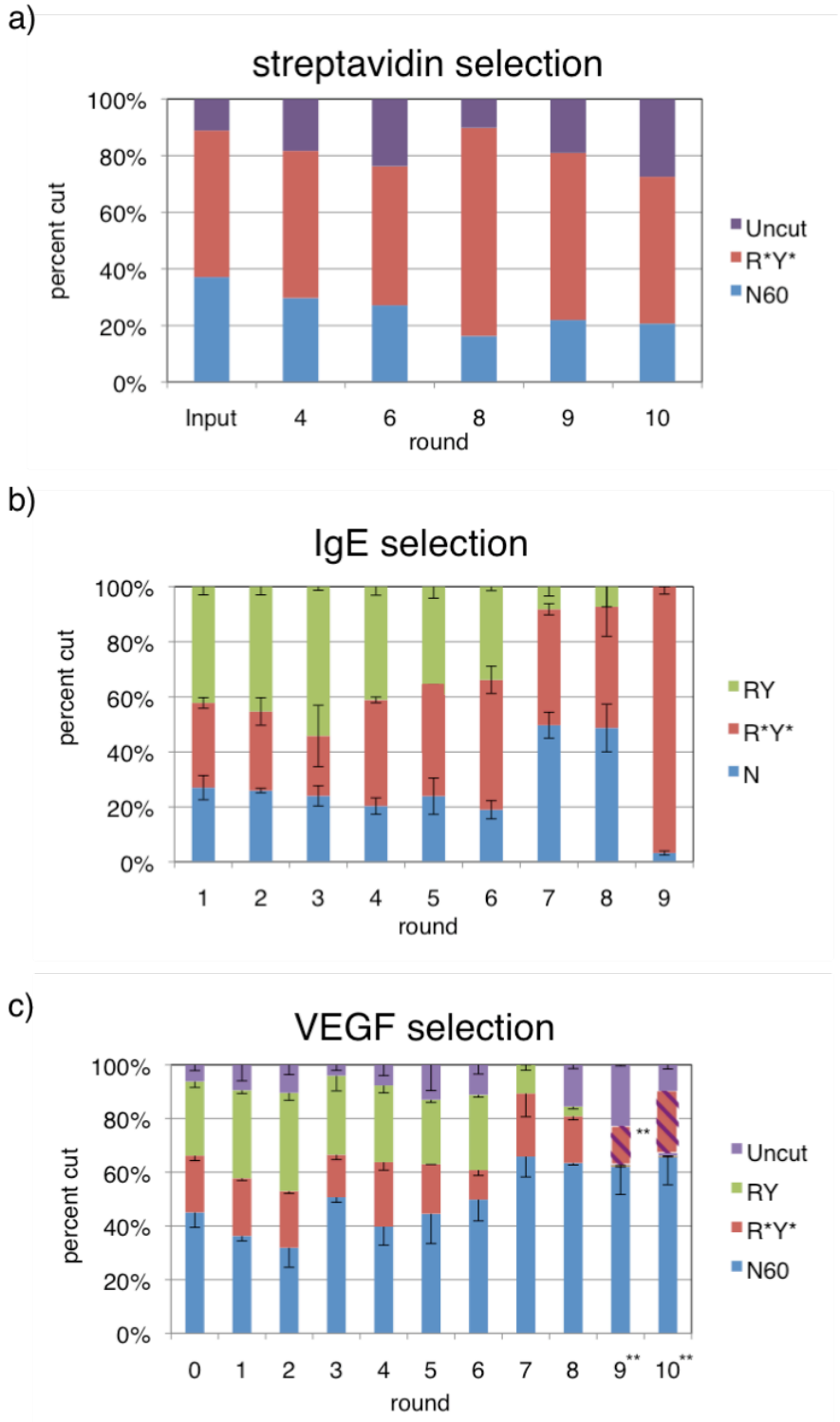

Figure S5. Tag cleavage ratios for selections against (a) streptavidin, (b) IgE, and (c) VEGF. Error bars indicate the standard error of three replicate digestions; ** the striped bars indicate the normalized percentage of clone V9-103 (which was from the R*Y* library but which contained a mutation in its tag sequence, preventing digestion), as determined by sequence-specific digestion in Figure S7. Due to the significant presence of uncut DNA in the streptavidin and VEGF selections, tag sequences were

used primarily to ascertain a library member's origins rather than to calculate library abundances by digestion.

Streptavidin Selection Sequences

For the streptavidin selection, the round 10 pool was sequenced, and 37 distinct sequences were determined (Table S4a). As described in the text, 27 of these 37 sequences contained a common hexaloop motif (bulge/loop in bold; stems underlined in Table S4a). Additionally, four sequences contained related motifs with expanded bulges or loops. Several clones, including the most frequent R*Y* clone, contained mutations in the library tag that preclude digestion (mutated tags italicized in Table S4a). The standard motif occurred in the same frame in all of the R*Y* clones (R*Y* pattern shown for comparison). The sequences' frequency and motif characteristics are described in Table S4b.

clone	sequence (variable region; full-length = GGTGCTCCTTGCGGTC-variable- GCACCAGACCACGACGG)
	RYRYRYRY NNNN RYRYRYRYRY NNN R**YRY**RYRY**RY** **NNNN** RYRYRYRYRY NNN RYRYRYRY
S10-317	CAGCTG AGAAGCGC CAGG GTGTACATGC ACA A**CGC**GTAC**GC** **CGCA** GTACTTGTGC TTA ACATGTGT
S10-134	*CAAGCTG*ACGGCACGCCGAA GTATGCACAC TTA A**CGC**GTAC**GC** **TGCA** GTACTTAAGT TGT GTATCGCC
S10-171	CAGCTG ACGCACGC TAAA ACATACGAAT ATG A**CGC**GTAC**GC** **TGCA** GTACTCATAT TCC GCGCATAT
S10-338	CAGCTG ACACGCAT CGAG AACTACGCAT ATG A**CGC**GTAC**GC** **TGCA** GTACTCATAT GCT ATAAACGT
S10-132	CAGCTG GTACGCAT CCAC ACGTATGCGC AGA A**CGC**ACAT**GT** **CGCA** ATGTTTATGC GCT ATGCGTGT
S10-104	CAGCTG ACACGTAT CCGG TCGTACATAT TTA G**CGC**GTGT**GT** **CGCA** ACACCTATAT ATG ACGCGTGT
S10-164	CAGCTG ATACACGT GAGG ATACATACAT TTG A**CGC**GTGC**GT** **CGCA** GCACTCAAAT TGT ACACGTAT
S10-102	CAGCTG ATACAGCAAACTG ACACATACAC ATG G**CGC**GTGT**GC** **CGCA** ACACCCATGT AGT ACGTGTGT
S10-303	CAGCTG ACGCGCAACGTAT GTACTGGCAT GCG A**CGC**ATGT**GT** **CGCA** ACATTCGCAG GAA GTATATAC
S10-321	CAGCTG ATATAGAT TTGT GTATGCATAC GTA A**CGC**ATAC**GC** **CGCA** GTATTTACGT ATG GTATACGC
S10-201	CAGCTG ACAGTGTACATTA GTGCACATAC GAA A**CGC**ATAT**GT** **CGCA** ATATTTTCGT AGT ATGCCCGC
S10-122	CAGCTG GCACGTAG GATA GTATAAATGG GAA A**CGC**ATAC**GC** **CGCA** GTATTTGCCC GGA ATATATGT
S10-161	CAGCTG AATATATTCTGTC GCTTATATGT ACG A**CGC**GTAC**GT** **CGCA** GTACTCGTAC TAA ACACACAT
S10-353	CAGCTG ACGCACAGTGCTT GCATACATAT ACA A**CGC**GTAC**GT** **CGCA** GTACTCGTAC TAA GCACACAC
S10-340	CAGCTG ACAGACAT GCGT CCTTACACAC GAA G**CGC**GTGT**GT** **CGCA** ACACCTTCGT TTG ACAGGCGC
S10-123	CAGCTG GCGCGCAT AACC ACGCACGTAC ATT T**CGC**GTAC**GT** **CGCA** GTACAAATGT TTG ACACGCGT
S10-175	CAGCTG GCGTACGT GGTG ATACGTATGC ATG G**CGC**GTGT**GT** **CGCA** ACACCCATGT CTA ATGTACGT
S10-323	*CAAGCTG*GTGCACGC GGTA GCATATACGT GTG A**CGC**GTGT**GC** **CGCA** ACACTCATAC TAT GTATATGC
S10-342	*CAAGCTG*GCATACGT AATC ATGCATTCAT ATA A**CGC**GTGT**GT** **CGCA** ACACTTATAT GTA GTATATGC
S10-101	CAGCTG ACATACGC ACTG GTGCAGTATACGTA A**CGC**ATAC**GC** **CGCA** GTATTTATGT TAT GCACGTGC
S10-215	CAGCTG GTAGGCAC TCAT GCGTGCACAT ATG A**CGC**ATAC**GT** **CGCA** GTATTCATAT TTC ACATGCGC
S10-217	CAGCTG GCATCCGT ACGA GCACATACAC GTG A**CGC**ATAT**GT** **CACA** ATATTCACGT ATT ATGTGCGC
S10-355	CAGCTG GTACACCT TCGG AGATATACGT A**TC** **GCT**CACAC**GC** **CGCA** GTGTGTACGT AAT ATCTACAT
S10-124	CAGCTG ATACACAT **TTGC** **A**CCTC**GACGC** **A**GA GGATGTGTGT CCGC ATGGACGTAC GGG ATGCATGT
S10-103	GGGATCCAATAATAGGACAAA**CGC**ACAC**GGCGCA**GTGTTTTGACCTACTTAGCCGACGCTGTCCCCG
S10-173	GGATCC TGTTACTATAAA**CGC**TACT**GTCGCA**AGTATTTATAGTCACTTACTGACCACTCAGCCTGC
S10-205	GGATCC ATCGTAATTCAATTCACTGATAACA**TGC**CATC**GTCGCA**GATGTGTTATTCTGATATTGCC
S10-346	GGATCC ATCTCCTATAAA**TGC**CTAT**GCCGCA**ATAGTTTATAGCCTCTGCAACTGGCTCGTCTGCCCC
S10-314	GGATCC AAAGCTTGACCGTCATGTACAAAACACCCATGA**TGC**CAAT**GCCGCA**ATTGTCATGGACGT
S10-115	*GGACC* CGCGGTAAGTTGTGT**TTGCT**CCCC**GACGCA**GGGGACACAAATACCCTACTGTCTCTCGCT
S10-113	GGATCC ATTTGAAGATTAGCAAACCT**CGC**GCC**GATTGCA**GGCAGGTTTGATTGATGACCTGGCCCC

S10-114	CAGCTG ACACGCGT TATT GGGTAACTGT CCA CCACGCATGA GGGC TCACATACGT ACA GTGTATGG
S10-111	CAGCTG ACCTATCT GGTC GGGTTTATAT TAA ACATATACAT TAGA ATACGTTTGT ACC GCGCGTAT
S10-162	CAGCTG GTATACGATCAGA GTACGCATAT CAT ACGCGGGTGT TGCT CCCTATGTAC TAT GCGCACGT
S10-144	CAGCTG GCATGCGC AGAT GTATCCACAC GTA ACGCACACAT CGGT GTTTACGGGC GGT ACTCGTGC
S10-155	CAGCTG ACGCATAT CCGT ACGTATGCAT GGA GTGAATGAGT CGCA ACGCGTATTT GGG ATACTTGC
S10-121	GGATCC GGCCAGCACTCTGTTACGCGTAATTGGGTTACTAACATATCCTGGGACTCTCGTAGCCCT

Table S4a. Streptavidin selection round 10 sequences.

clone	percent of round 10	library	motif	major predicted fold displays motif?	predicted energy of folding (kcal/mol)
S10-317	1%	R*Y*	standard	yes	-17.9
S10-134	15%	R*Y*	standard	yes	-17.1
S10-171	1%	R*Y*	standard	yes	-16.8
S10-338	1%	R*Y*	standard	yes	-20.1
S10-132	1%	R*Y*	standard	yes	-20.2
S10-104	3%	R*Y*	standard	yes	-15.9
S10-164	1%	R*Y*	standard	yes	-17.5
S10-102	1%	R*Y*	standard	yes	-19.2
S10-303	3%	R*Y*	standard	yes	-20.8
S10-321	1%	R*Y*	standard	yes	-14.1
S10-201	12%	R*Y*	standard	yes	-14.0
S10-122	1%	R*Y*	standard	no	-13.0
S10-161	1%	R*Y*	standard	no	-14.5
S10-353	1%	R*Y*	standard	no	-13.6
S10-340	1%	R*Y*	standard	no	-20.6
S10-123	1%	R*Y*	standard	no	-17.5
S10-175	1%	R*Y*	standard	no	-19.9
S10-323	1%	R*Y*	standard	no	-21.3
S10-342	3%	R*Y*	standard	no	-15.5
S10-101	5%	R*Y*	standard	no	-18.3
S10-215	5%	R*Y*	standard	no	-17.1
S10-217	3%	R*Y*	*GTCACA	yes	-19.0
S10-355	1%	R*Y*	*5-5-6	yes	-16.1
S10-124	1%	R*Y*	*5-4-6	yes	-20.2
S10-103	11%	N_{60}	standard	yes	-14.8
S10-173	1%	N_{60}	standard	yes	-13.2
S10-205	1%	N_{60}	*TGC	yes	-9.0
S10-346	1%	N_{60}	*TGC	yes	-11.9
S10-314	1%	N_{60}	*TGC	yes	-17.5
S10-115	1%	N_{60}	*5-4-6	yes	-19.1
S10-113	4%	N_{60}	*3-3-7	no	-16.4
S10-114	1%	R*Y*	unknown		
S10-111	1%	R*Y*	unknown		
S10-162	1%	R*Y*	unknown		
S10-144	1%	R*Y*	unknown		

S10-155	1%	R*Y*	unknown		
S10-121	3%	N_{60}	unknown		

Table S4b. Summary of streptavidin selection round 10 clones.

We synthesized minimized forms of five sequences from among the round 10 clones, including several with the standard bulge/loop sizes and also one each with an expanded bulge and expanded loop. Because the motif was apparent by inspection, minimal sequences were analyzed by CE (Table S5, bulge/loop sequences are shown in italics).

clone	sequence
S10-101	CACGTAA*CGC*ATAC*GCCGCA*GTATTTATGTG
S10-103	TAGGACAAA*CGC*ACAC*GGCGCA*GTGTTTTGACCTA
S10-104	CATATTTAG*CGC*GTGT*GTCGCA*ACACCTATATATG
S10-113	CAAACCT*CGC*GCC*GATTGCA*GGCAGGTTTG
S10-115	CTTGTGT*TTGCT*CCCC*GACGCA*GGGGACACAAG
consensus	CGCTGA*CGC*GTAC*GTCGCA*GTACTCAGCG
NRR	CTGTGAGACGA*CGC*ACC*GGTCGCA*GGTTTTGTCTCACAG

Table S5. Minimal streptavidin binding motif analyzed by capillary electrophoresis.

IgE Selection Sequences

The round 9 pool from the IgE selection was sequenced, and 23 distinct sequences were found (Table S6). I9-102 occurred 60 times, while the other 22 sequences were unique.

clone	library	sequence (variable region; full length = TGTCGCTGCGTCGCCTG-variable-CACCGGAAGACGCACGC)
I9-101	R*Y*	CAGCTGACGTGTAATGTTGATGCAAACATCGTACACACGCGTTCTGGCCCGACGCATGCGGCACGCGC
I9-102	R*Y*	CAGCTGACGTACGTGCATGGCAAACACACTTCATCCGTACCTTCTAGTGGGTGTGTAGCAAGCGCGC
I9-103	R*Y*	CAGCTGACATGTATCCCGATACATGCCTGATCCACACGCATCACGACATACGCGCCACGNNCGGG
I9-104	R*Y*	CAGCTGATGTATCTTCGTGCGGGCACAAATCATATATGCGTCAATGACACGTAAGCGGCATGTGTGC
I9-107	R*Y*	CAGCTGGGGGCCTATTGCAGTAAGCACGCCGCGCGCATACGCCNTCGGATGGACGCAAGACACGTGT
I9-110	RY	GCTAGGCGTGTACGTTGTGGGCACACACACTAAATGTACATGCGGAGACACACGCGCGCGACACATGC
I9-111	RY	GCTAGCGCGTACGCTACGACATGTGCGCGGAATGTGTATATCTAGGCACACGTGCCAAATATGTAT
I9-113	N	GGATCCTCGTCGGAAACAAAACCCCGTTTCGGTGATTGGGGATCAAGGGCGACTCAGGGAGAGCATA
I9-115	R*Y*	CAGCTGATAGGCTATTCTTGTACATGCAGTAAAACGTACGCACTCCCACCCATATATCGCGCATATAT
I9-116	R*Y*	CAGCTGATGTGTGTTCCTGTATGCGAGAACACCATCTACGCTCGGGTAGACGCATTGGGCACGTGC
I9-117	R*Y*	CAGCTGCCATCCATCGCCACGCCCGTACCCCACGCACACTCTCCCATCTACCTACGCCACACATTT
I9-119	N	GGATCCGATTGGATCATAGGTAAGAAGGCAGGGAGATGCGCTTATGTAGGGGGACCCGCGGGTGG
I9-202	N	GGATCCTTCGGGACGTCGGAAGCCAGGTTAAGATGATCCGAGGCACACCATACTCACACAAGACCGTG
I9-204	N	GGATCCATTAATCCGTTCTTTATCCTCCACCCTCTCAATTCCCTAGTATTTACCCATCAGGCCTACG
I9-211	N	GGATCCGTTCCTGGTAAGAGTTGTAAACATAACTGAAACTGAGGGGGGAGAACACAGGGACGCGCG
I9-216	R*Y*	CAGCTGATACGTGGGAGATGGTGTAGACACCCCATACGCGTGATAAGGGCTGATGTGTGAGACGCGCGC
I9-219	R*Y*	CAGCTGATGCACGCGAGGGTGCTAATGCGGCAAGCGCAGGCAGGGACACGTGCACATGGTACATGT
I9-224	N	GGATCCCACTTCCCTTCCACCATCCCCCGATCTTTGTTACCTCAACGTAGGCCTCCCAGCAACACC
I9-230	RY	GCTAGCGTATATGTCCACACATGCGTGTGGAACGTGTACGTAGAAATGCGTATGCAGCGCGCGTAT
I9-301	RY	GGCTAGCACACGCACGGATACGCACGTATTCGGTGTGTGCATCCATATACGTACGCGCGACACGTAC
I9-309	N	GGATCCATATCTCCCGTTCTTTGCCCCTCGCCCTTACTCCTCTTTTCGGTCCCATAAATCTTCACT

I9-313	R*Y*	CAGCTGATGCACGCTAGAGGTGCTACACGCGTAACACATATTCAGCTGTGCATATGCATGACGCAGAT
I9-317	N	GGATCCAAACGTGCTGCGGCACCGAGCGATCACTGATTTACGACCTTGGGACAGCATGGAATCGACG

Table S6. IgE selection round 9 sequences.

IgE Binding Analysis

In contrast with the streptavidin selection results, no common motifs were apparent by inspection. Therefore, all 23 clones were analyzed for binding to IgE-linked beads. Only those clones with binding activity of ~1% or more are shown in Figure S6.

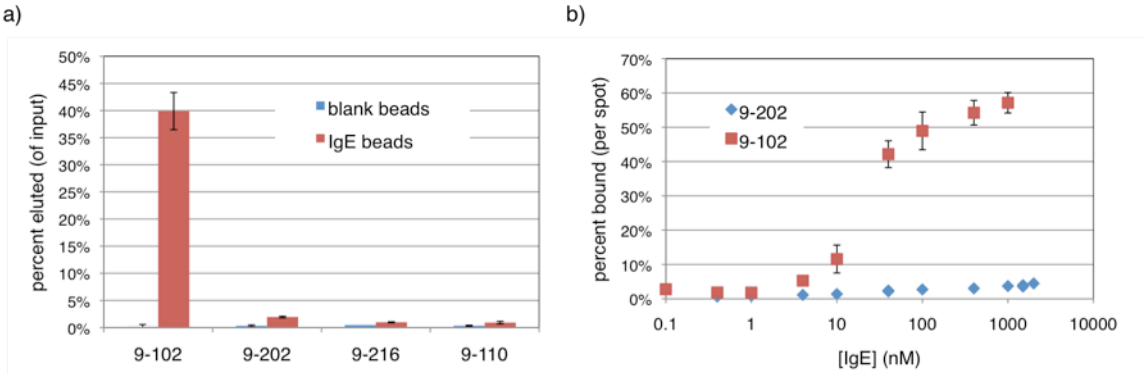

Figure S6. Binding activity of the most active IgE clones to (a) immobilized IgE, and (b) free IgE by nitrocellulose filter binding. Error bars reflect the standard error of two replicates, except bead binding of I9-102, which is three replicates.

Binding constants were determined from nitrocellulose binding data by fitting to the curve:

$$\text{signal} = min + max \times \left(\frac{K_d + DNA + IgE - \sqrt{(K_d + DNA + IgE)^2 - 4DNA \times IgE}}{2DNA} \right)$$

where the DNA concentration was constant and much less than the K_d; the IgE concentration was the independent variable; and min and max signals are parameters optimized by the fit. The standard deviation of the K_d was determined using a statistics package to determine the variance matrix of the parameters based on all of the independent data points.

VEGF Selection Sequences

The round 9 and 10 pools surviving VEGF selection were sequenced, and 26 distinct sequences were determined. Several of these sequences occurred multiple times. Sequences and abundances are given in Table S7a and b. Mutated library tag sequences are underlined in Table S7a.

clone	sequence (variable region; full length = TGTCGCTGCGTCGCCTG-variable-CACCGGAAGACGCACGC)
V9-101	GGATCCGCTGCCTGTCGCGTGGGTCCGGATGGCGCAAGGTTTGCTTCGCGGCAGCTTATTGGGAA
V9-103	CAGCTTGCGTAGTGAGTCCGAATGGGTGCACAAAGTGAGCGTATGCCAGTGCGCGCGCCATATACATAC
V9-104	GGAATCCCTGCAGGCCCGGGCCAAAACACTGAAATCCGTACTTGCGGTGGAAGTCCGAATGGGTGTC
V9-105	GGATCCCATATGGTTTACGTATTCCTGGTGCTACCCTGACGCCGGCACTCAGGCGCAGCCGGAAAG
V9-110	GGGATCCGATTATGCTTAACAGCAGAGAGCCTGGCAGACTAGAGTGCAGCAGCGACTAACTTTAATA
V9-112	GGATCCGCGTCCGAATGGCGCACTGAACCCAACGCACGACATTTCGCGAGAAACAGCCATCATCTA
V9-114	CAGCTGAAGCAAGTCCGAATGGGTGTGTTCAGCATATGCACTGCGACGTGCGTACACCATACACGC
V9-132	CAGCTTGGTATGCATACTAACTTACATACATGGTATATGAGTCCGAATGGGTGCATATAATATGTAC
V9-201	GGATCCACGTAGACGCGTATCCAACGTTGACACTCGACTGCATCATCTGAGCTAGCAATGCTAGT

V9-204	GGATCCCGGCCGGCATAAGAGTCCGAATGGGTGCTTACAGTCTCGAATGGGCTTGCGATAGGGGAC
V9-209	GGATCCTTTTCAGATACCCGGCACACCTCTGCAATTGCGGAGGCAAGGTCTAACTTCGACCAGGCA
V9-215	GGATCCGAATGGGTCACACATGGGCTCACACAATAAGAGGTCCGAAGGGGACTCTTTTGAATCGCT
V9-218	GGATCCGCGTGCAGGAAAAGGGCCGTGACGCGGCGCGAGCGTTCTTCATGGACCTGAACGCCAAAC
V9-224	CAGCTGAGTACGTACTTTAATGCACAGCATGGTATATACTTGTAGGTGTACGCACACCCGACACGTGT
V9-228	GGATCCCAGCGACTAATAATGCTACCCCGCAGCGCGTTAATATTTGCTGCTAGCACATTTTCAACTA
V9-229	CAGCTGCTGTTTGTTTATACGTGTATACTCAATACGCACGCAACTGCAAACGCATACGATGTGTGC
V10-122	GGAATCCAGCTGAGTCCGAATGGGTGCAGCCGGGCCAGGCAAAGCGAGTCCACCGGCCATTCATAAA
V10-208	GGATCCCAACAAAATGGTCCGGATGGGTCAGTGCTTGGGGTCATGTCCGCATCCAGGCGACACGCG
V10-247	GGATCCTCTGTTCGTCCAAACTACCGTGGACCTGTCGGTTTTGGACTAGAGGGCAGATACGGGGGA
V10-254	GGATCCCCGTCCGAATGGCAGTCTCACTCTGTGTACGTGGGGTTAAGGCAACCAGCGGGCTCATCG
V10-209	GGATCCGCTCTATGAAATTATTTTAAACGTATGTTAAAAATCGCCGCGCAGCCAGAGAGCTCAGG
V10-229	GGATCCGCCTCCTCCCGCCGGGTGTTTGTTGAGTCCGAATGGGTGCCAAACGAGCGCGACACTGTC
V10-232	GGATCCACGGTGTGCTTTGGTGTACAGCCCGTCGAAGACAAGAGCGCAGGGCTATCAGACCATGCA
V10-235	GGATCCCATGGAGCTCAGATCAGGAAGGGACGCGGGGAGAATTGTGACGTATCCGGCTAAGGTACAT
V10-238	GGATCCGTAAGCCGGTAGCCACGTCCGAATGGTGTGGTGTACGTCCGAATGGCGGAAGGGATGAGAT
V10-245	GGATCCGCCTGTCGAATGAGTCCGAATGGGTGCAATCGTCGTTACCAATATTTCGCAAATCCCTCTA

Table S7a. VEGF selection round 9 and round 10 sequences.

clone	library	% of round 9 pool	% of round 10 pool
V9-101	N_{60}	13%	28%
V9-103	R*Y*	40%	38%
V9-104	N_{60}	3%	0%
V9-105	N_{60}	16%	13%
V9-110	N_{60}	2%	0%
V9-112	N_{60}	2%	0%
V9-114	R*Y*	2%	1%
V9-132	R*Y*	2%	0%
V9-201	N_{60}	2%	0%
V9-204	N_{60}	5%	4%
V9-209	N_{60}	2%	0%
V9-215	N_{60}	6%	0%
V9-218	N_{60}	2%	0%
V9-224	R*Y*	2%	0%
V9-228	N_{60}	2%	0%
V9-229	R*Y*	2%	0%
V10-122	N_{60}	0%	1%
V10-208	N_{60}	0%	2%
V10-247	N_{60}	0%	2%
V10-254	N_{60}	0%	2%
V10-209	N_{60}	0%	1%
V10-229	N_{60}	0%	1%
V10-232	N_{60}	0%	1%
V10-235	N_{60}	0%	1%

V10-238	N_{60}	0%	1%
V10-245	N_{60}	0%	1%

Table S7b. Summary of VEGF selection round 9 and round 10 clones.

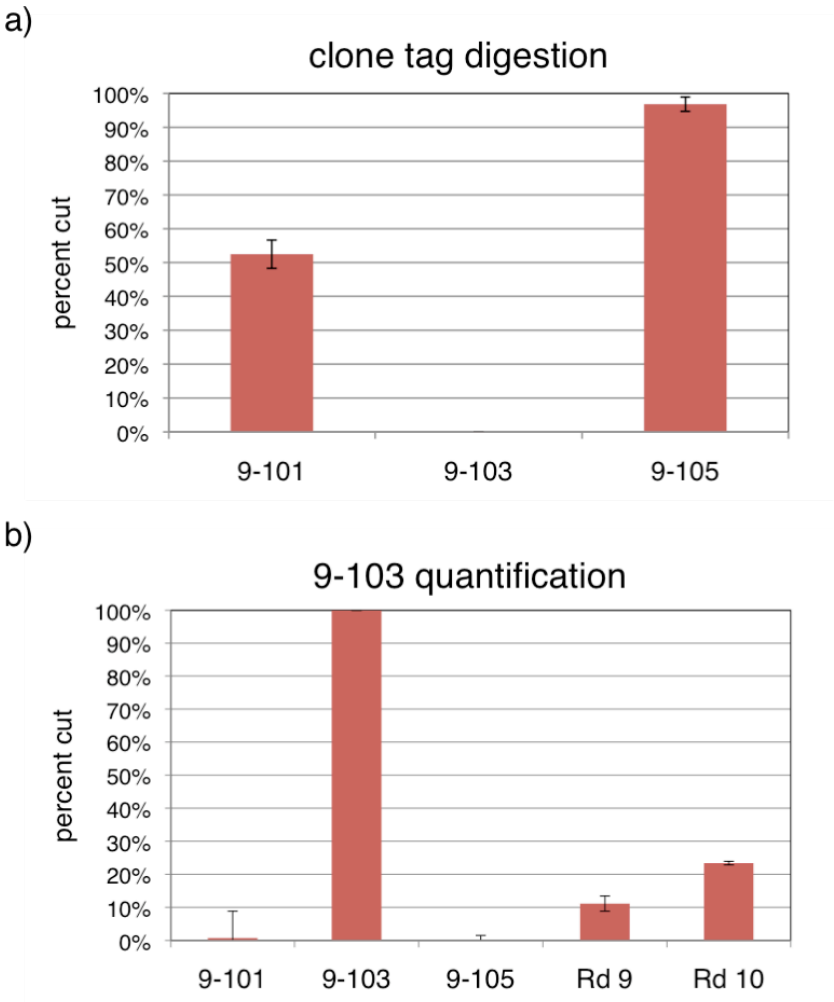

Figure S7. Tag mutation among clones in the VEGF selection. (a) Digestion of common VEGF clones with cognate library tagging enzymes; (b) quantification of V9-103 by digestion with *Apa*LI.

As indicated in Table S7, several of the VEGF clones included insertions in the library tagging sequence that prevent restriction enzyme digestion, including the most common R*Y* clone, V9-103. The round 10 pool of the VEGF selection (Figure S5c) included a substantial amount of uncut dsDNA in the samples treated with all three enzymes (30% of the lane). In addition, 15-20% of the pool was cut in the presence of all three enzymes but not in the presence of any one enzyme alone. The three most common clones were analyzed individually by tag digestion (Figure S7a). As predicted by sequencing, clone V9-105 was fully digested by *Bam*HI, and clone V9-103 was unaffected by *Pvu*II. However, clone V9-101, despite its sequence indicating an intact tag, was only partially digested by *Bam*HI.

In order to independently quantify the amount of V9-103 in the pool by digestion, we used the *Apa*LI restriction enzyme, which is predicted to digest clone V9-103 within its variable region but is not predicted to cut any of the other repeated clones listed in Table 5. We experimentally confirmed its specificity for V9-103 by showing that it completely digested this clone while not digesting the

other two most common clones, V9-101 and V9-105 (Figure S7b). Finally, we used this enzyme to quantify the amount of V9-103 in the VEGF selection round 9 and 10 pools. By digestion, the round 10 pool includes 23% clone V9-103 (Figure S7b).

VEGF Binding Analysis

No common structural motifs were apparent by inspection, so we analyzed full-length VEGF clones. We tested those clones that occurred more than once in rounds 9 and 10, and all of the tested clones bound VEGF beads (Figure S8). Given the prevalence of binding activity among these clones, it is probable that some or all of the unique clones also bind VEGF. We assume that the best binders are those which have enriched appreciably over the course of the selection and which therefore occur multiple times.

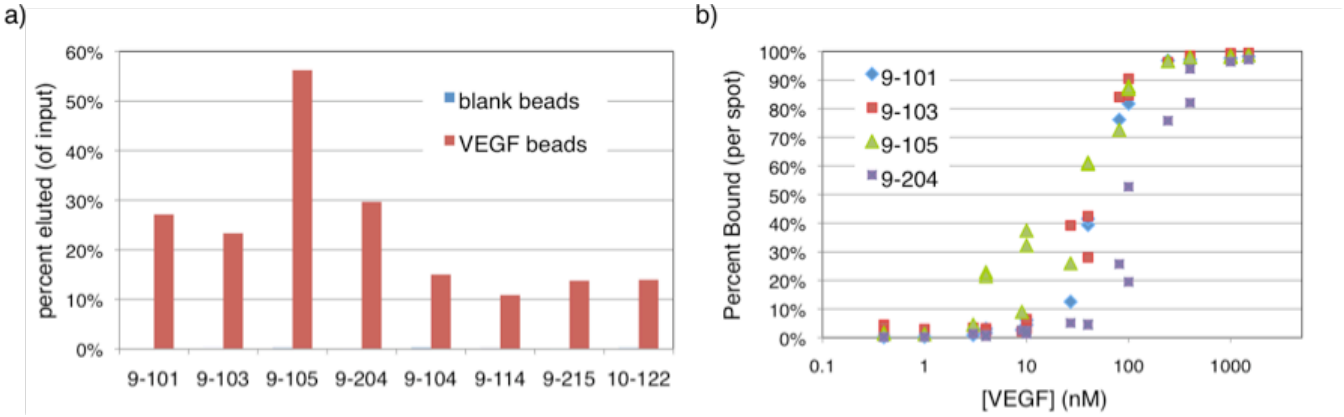

Figure S8. Analysis of VEGF clones observed multiple times during DNA sequencing by (a) VEGF-linked bead binding assay, and (b) free VEGF binding assay by nitrocellulose filter binding.

Streptavidin Binding Motif Probability Calculation

Because individual sequence survival is stochastic during the early rounds of a selection, and because an input of 10^{14} molecules covers <1% of all sequences possible in an N_{60} library, the results of any given selection must be considered only one amongst many possible outcomes. An alternative approach to evaluating the functional potential of a nucleic acid library is to calculate the expected probabilities of a known motif occurring in the patterned and standard libraries.

The most common binding motif from the selection for streptavidin binding was:

XXXXXX-YGC-XXXX-GNYGCA-XXXX-XXXXXX,

where X indicates a position that must form a base pair across a stem. A single frame in the patterned library fits the bulge/loop motif better than any other frame. This frame is shown relative to the pattern:

XXXXXX-YGC-XXXX-GNYGCA-XXXX-XXXXXX
RYNNNR-YRY-RYRY-RYNNNN-RYRY-RYRYRY

Alignment of the round 10 sequences from the R*Y* library with the pattern revealed that, indeed, all 22 instances of the standard motif in the structured library occurred in this predicted frame (Table S4a). In contrast, one R*Y* sequence (C4) with the expanded 5-base bulge –TTCGW- contained the motif in a different frame. However several other frames of the R*Y* pattern fit the bulge/loop motif slightly less well but still add significantly to the overall likelihood of the motif occurring in the

library. The binding motif was compared to every register of the R*Y* pattern and the probability of motif occurrence was calculated, as discussed in the following pages.

Unlike R*Y*, every frame of N_{60} is alike, so the motif would be predicted to occur with equal frequency in any frame of the variable region. Indeed, the five round 10 sequences from the N_{60} library include the motif in different registers. Therefore, in considering the overall likelihood of the motif, the N_{60} library has 32 registers that could include a 29-base motif.

For any given frame in a library variable region, the probability of motif occurrence is the product of the frequency of the required base(s) at each bulge/loop position times the product of the probability of ten base pairs. These values depend on the base frequencies observed in each mix from sequencing the input library (Table S3). As an example, the bulge/loop motif probability is calculated for the observed frame from the R*Y* library and for any frame in the N_{60} library (Table S8).

	frame	Y	G	C	G	N	Y	G	C	A	TOTAL
N_{60}	any	N-.56	N-.18	N-.28	N-.18	N-1.0	N-.56	N-.18	N-.28	N-.26	3.72×10^{-5}
R*Y*	observed	Y*-.96	R*-.41	Y*-.40	R*-.41	Y*-1.0	N-.56	N-.18	N-.28	N-.26	4.74×10^{-4}

Table S8. Probability of the bulge/loop motif in the observed frames in the N_{60} and R*Y* libraries used for the streptavidin selection.

The stem probability is equal to the likelihood of ten base pairs forming, where each base pair could be either of the two Watson-Crick pairs or a G-T wobble.

For the N_{60} library:

Likelihood of any two 'N' making a pair	AT	0.26 x 0.28	0.0728
	TA	0.28 x 0.26	0.0728
	CG	0.28 x 0.18	0.0504
	GC	0.18 x 0.28	0.0504
	GT	0.18 x 0.28	0.0504
	TG	0.28 x 0.18	0.0504
	Total		0.3472

For the R*Y* library:

Likelihood of any R* and Y* making a pair	AT	0.52 x 0.56	0.2912
	GC	0.41 x 0.40	0.1640
	GT	0.41 x 0.56	0.2296
	TA	0.03 x 0.02	0.0006
	CG	0.04 x 0.02	0.0008
	TG	0.03 x 0.02	0.0006
	Total		0.6868

By analogous calculations,

Likelihood of R* and R* making a pair	0.0886
Likelihood of Y* and Y* making a pair	0.0608
Likelihood of Y* and N making a pair	0.3352
Likelihood of R* and N making a pair	0.3956
Likelihood of N and N making a pair	0.3472

As an example, for the observed frame in the R*Y* library, the probability of ten base pairs forming would be:

Probability of (6 R*-Y* + 1 R*-R* + 2 Y*-N + 1 R*-N) base pairs =

$(0.6868)^6$ x (0.0886) x $(0.3352)^2$ x (0.3956) = 4.13×10^{-4}

Therefore, the total probability for the observed frame in R*Y* would be:

Loop probability x stem probability = 4.74×10^{-4} x 4.13×10^{-4} = 1.96×10^{-7}

Streptavidin Binding Motif Total Probability

The frequency of the required base(s) at each bulge/loop position and the probability of formation of each base pair were determined for every frame in both the N_{60} and R*Y* libraries using a computer program to assign a value based on the base ratios in the appropriate position in the pattern. These values were then multiplied to determine the overall probability of the motif occurring in each frame (Table S9). The probability in each frame was summed to determine the total probability for the library (Table S9). We treated the frames as independent from each other because the likelihood of the motif occurring in any given frame was so small as to have an insignificant effect on overlapping frames.

frame	N_{60}	R*Y*	RY
1	9.49×10^{-10}	5.99×10^{-10}	3 loop off
2	9.49×10^{-10}	7.64×10^{-11}	4 loop off
3	9.49×10^{-10}	**1.53×10^{-7}**	1 loop off
4	9.49×10^{-10}	2.61×10^{-12}	4 loop off; 1 stem off
5	9.49×10^{-10}	6.96×10^{-11}	1 loop off; 2 stem off
6	9.49×10^{-10}	8.84×10^{-12}	2 loop off; 3 stem off
7	9.49×10^{-10}	3.39×10^{-15}	4 loop off; 4 stem off
8	9.49×10^{-10}	1.11×10^{-10}	4 stem off
9	9.49×10^{-10}	1.98×10^{-17}	6 loop off; 4 stem off
10	9.49×10^{-10}	5.33×10^{-10}	4 stem off
11	9.49×10^{-10}	1.10×10^{-17}	7 loop off; 3 stem off
12	9.49×10^{-10}	7.92×10^{-9}	2 stem off
13	9.49×10^{-10}	6.80×10^{-16}	7 loop off; 1 stem off
14	9.49×10^{-10}	**1.39×10^{-7}**	1 loop off
15	9.49×10^{-10}	3.02×10^{-13}	6 loop off
16	9.49×10^{-10}	**1.51×10^{-7}**	1 loop off
17	9.49×10^{-10}	2.39×10^{-11}	4 loop off
18	9.49×10^{-10}	6.09×10^{-9}	2 loop off
19	9.49×10^{-10}	7.08×10^{-9}	2 loop off
20	9.49×10^{-10}	2.60×10^{-11}	4 loop off
21	9.49×10^{-10}	**1.96×10^{-7}**	1 stem off
22	9.49×10^{-10}	2.94×10^{-13}	4 loop off; 2 stem off
23	9.49×10^{-10}	9.81×10^{-12}	2 loop off; 3 stem off
24	9.49×10^{-10}	8.16×10^{-13}	3 loop off; 3 stem off
25	9.49×10^{-10}	2.22×10^{-13}	4 loop off; 2 stem off
26	9.49×10^{-10}	3.79×10^{-11}	3 loop off; 1 stem off
27	9.49×10^{-10}	2.16×10^{-11}	4 loop off
28	9.49×10^{-10}	5.99×10^{-10}	4 loop off
29	9.49×10^{-10}	7.64×10^{-11}	4 loop off
30	9.49×10^{-10}	**1.53×10^{-7}**	1 loop off
31	9.49×10^{-10}	2.61×10^{-12}	4 loop off; 1 stem off

32	9.49×10^{-10}	6.96×10^{-11}	2 loop off; 2 stem off
total	3.04×10^{-8}	8.16×10^{-7}	impossible

Table S9. Streptavidin-binding motif likelihood in each library.

Note that five frames contribute significantly to the overall motif probability for the R*Y* library (probability greater than 1×10^{-7}). These likely R*Y* frames have two or fewer mismatches for the binding motif bulge/loop positions and one or fewer mismatched pairing positions in the stem. The observed frame (21) is the most likely frame in R*Y*, and is 6.4-fold more likely to contain the binding motif than the cumulative probability for every frame of N_{60}. Including every frame in both libraries, the streptavidin-binding hexaloop motif is 27-fold more likely in the R*Y* library than in the N_{60} library.

A comparison of the binding motif to the RY pattern indicates that every frame contains at least one mismatched bulge/loop position or one mismatched base pair. Although the RY patterned library was not included in the selection for streptavidin binding, motif probability indicates that it would have failed to result in aptamers with the consensus motif.

IgE Aptamer 9-102 Minimization and Motif Analysis

9-102 Minimization and Mutational Analysis

We identified the important region for binding IgE by synthesizing a series of sequences based on truncations and mutations of I9-102 (sequences shown in Table S10, loop sequences in italics, mutations in minimized sequence stems underlined). The results are summarized in Figure S9 and Figure 5a in the text.

construct	sequence
102_29-83	CGTGCATGGCAAACACAC*TTCATCCGTACCTTCTAGTGG*GTGTGTAGCAAGCGCG
102_29-83m	CGTGCATGGCAAACACAC*TTCAT*GTGTGTAGCAAGCGCG
102_1-26	TGTCGCTGCGTCGCCTGCAGCTGACG
102minA	GCACACAC*TTCATCCGTACCTTCTAGTGG*GTGTGTGC
102minB	GCACACAC*CCGTACCTTCTAGTGG*GTGTGTGC
102minD	GCACACAC*TTCATCCGTACCTTCTAGTCC*GTGTGTGC
102minE	GCACACAC*TTCATGGGTACCTTCTAGTCC*GTGTGTGC
102minG	GCACACTC*TTCATCCGTACCTTCTAGTGG*CTCTGTGC
102minH	GCACAGTG*TTCATCCGTACCTTCTAGTGG*CACTGTGC
102minI	GCACAGAG*TTCATCCGTACCTTCTAGTGG*GAGTGTGC

Table S10. Sequences for I9-102 minimization and mutational analysis.

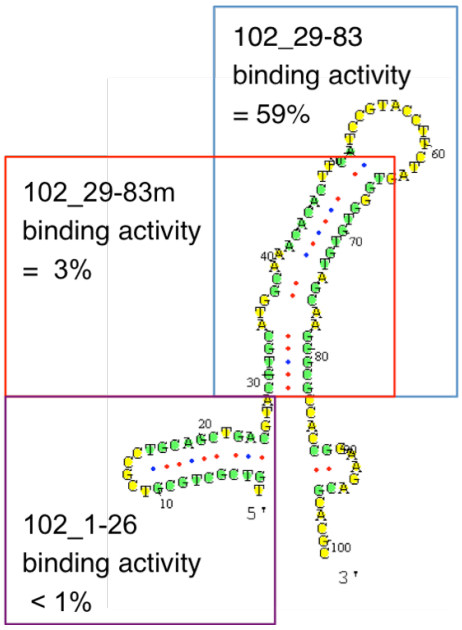

Figure S9. Predicted fold and binding activities of minimized IgE clone 9-102 variants.

102min Reselection

Having isolated the IgE binding activity in the large loop (bases 46-67 by full length I9-102 numbering), we wished to determine which of the positions in the loop were important for binding, and which, if any, were flexible. We designed and synthesized a library that was based on 102minA, but which included variability at each position. The library sequence was: 5'ACCTATCGTATCCTACCGATTTgcacacacttcatccgtaccttctagtgggtgtgtgcTTTGTGGAGTAAGGT AGACTCA), where lower case bases are a mixture that gives (79% indicated, 7% each of the other bases). Fifty sequences from the input library were analyzed, and the base ratios are indicated in Figure S10.

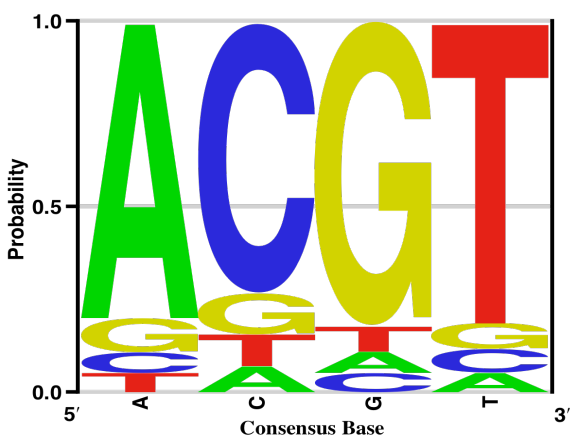

Figure S10. Base ratios by DNA sequencing of the 102min starting library, where each position was intended to be 79% consensus base and 7% each of the other three bases. Created with enoLOGOS.[2]

Selection for IgE binding starting with the 102min library resulted in binding activity after four rounds, and sequencing showed that the pool had largely converged back to the 102min sequence (Figure 6 in the main text). The resulting 59 total sequences, containing 42 unique sequences, were analyzed first for stem requirements, then for loop requirements for binding.

102min Motif Analysis

While many individual mutations were observed in the stem-forming region, wobble and co-variation preserved base pairing across the stem. Of the 42 unique sequences, all eight base pairs were maintained in 25 cases, and seven of the eight were preserved in another 16 cases. All told, at least seven out of eight base pairs were preserved in 41 out of 42 cases. In addition, the outermost and two innermost base pairs were preserved in all 42 unique sequences. The significance of this preservation depends on the probability of maintaining these 5 G:C base pairs and 3 A:T base pairs by chance in the starting library.

The probability of maintaining a base pair at any given pairing position is calculated from the frequencies of each base in the starting library. The probability of maintaining a base pair at a position that is A:T or T:A in the starting library is given below as an example:

$Prob(AT_{pair})$ =

AT =	0.801 x 0.815 =	0.6528
GT =	0.092 x 0.815 =	0.0750
CG =	0.056 x 0.070 =	0.0039
GC =	0.092 x 0.064 =	0.0059
TA =	0.052 x 0.052 =	0.0027
TG =	0.052 x 0.070 =	0.0036
		0.7439

By an analogous calculation, the probability of maintaining a base pair at a position that is G:C or C:G in the starting library is $Prob(GC_{pair}) = 0.6986$. Therefore, the probability of losing a base pair at a position that is A:T in the starting library is: $Prob(\sim AT_{pair}) = 1 - 0.7439 = 0.2561$ and the probability of losing a base pair at a position that is G:C in the starting library is: $Prob(\sim GC_{pair}) = 0.3014$.

Using these probabilities, the probability of seven out of the eight base pairing positions in the stem being maintained in any sequence from the starting library can be calculated by:

The probability of all 8 base pairs being maintained

$(P(GC))^5 \times (P(AT))^3 = (0.6986)^5 \times (0.7439)^3 = 0.16640 \times 0.41166 = 0.0685$

The probability of 7/8 base pairs being maintained (with the three outer pairs fixed)

Prob(outer) x Prob(variable) =

$[P(GC))^2 x(P(AT)] x [3x(P(\sim GC))x (P(GC))^2 x (P(AT))^2 + 2x(P(\sim AT))x (P(GC))^3 x P(AT))]$

$[(0.6986)^2(0.7439)] \times [3(0.3014)(0.6986)^2(0.7439)^2 + 2(0.2561)(0.6986)^3(0.7439)] = 0.1358$

The probability of *At least 7 pairs* by chance in any one sequence

= Prob(8) + Prob(7/8) = 0.0685 + 0.1358 = 0.2043

The probability of at least seven base pairs being maintained in any given sequence from the starting library is 0.2043. Therefore, the binomial probability of at least seven base pairs occurring in 41 out of 42 sequences in the output pool by chance is less than 0.0001 (p-value < 0.0001). The 9-102 motif for IgE binding is considered to be a stem with at least seven out of eight base pairs, with the loop described below.

The loops of the 42 unique clones in the round 4 pool are summarized in Table S11 (position numbering based on 102minA). White columns indicate entirely conserved positions, while increasing blue color indicates increasing flexibility at the position.

position:		9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
original		T	T	C	A	T	C	C	G	T	A	C	C	T	T	C	T	A	G	T	G	G
#	A	0	0	0	42	0	0	0	0	0	23	1	0	0	0	0	0	42	0	0	0	0
	C	0	0	29	0	0	42	42	0	0	4	41	42	0	5	41	1	0	0	0	0	0
	G	0	0	5	0	0	0	0	42	0	4	0	0	0	0	0	0	0	42	0	42	42
	T	42	42	8	0	42	0	0	0	42	11	0	0	42	37	1	41	0	0	42	0	0

Table S11. Summary of 102min reselection loop sequences in the round 4 pool

For the conserved (white) positions, the significance of all 42 clones maintaining the original base can be calculated by determining the binomial probability of 'at least 42/42' given the base ratios in the starting library. The ratios and significances for the conserved (white) positions are given in Table S12. All of the conserved (white) positions can be considered fixed in the binding motif with a significance > 0.999.

original base	frequency in input	*p*-value prob(at least 42/42 in input)
A	0.801	0.0001
C	0.734	< 0.0001
G	0.824	0.0003
T	0.815	0.0002

Table S12. Significance of highly conserved residues in the 102min binding loop.

For the semi-conserved (light blue) positions, the significance of 41 of 42 clones maintaining the original base can be calculated by determining the binomial probability of 'at least 41/42,' given the base ratios in the starting library. The ratios and significances for the semi-conserved positions are given in Table S13. All of the semi-conserved (light blue) positions can be considered fixed in the binding motif with a significance > 0.99.

position	original base	frequency in input	*p*-value prob(at least 41/42 in input)
19, 23	C	0.734	< 0.0001
24	T	0.815	0.0020

Table S13. Significance of semi-conserved residues in the 102min binding loop.

Determining the requirements for the flexible positions (dark blue - positions 11, 18, and 22) requires the calculation of 'at most' and 'at least' binomial probabilities for each base. In each case, the frequencies imply a preference, although in most cases there are no statistically significant requirements. For example, for position 22, the observed number of Cs is higher than would be expected, given the starting ratios, but only barely significant (~0.95). The preferences indicated in Table S14 were employed as the position requirements for the binding motif in the motif probability calculations below.

position	base	frequency in input	k (# in output)	*p*-value		preference
				Prob(at least k/42) in input	Prob(at most k/42) in input	
11	A	0.070	0	1	0.0475	not A
	C	0.734	29	0.7945	0.3140	
	G	0.111	5	0.5056	0.6788	C, G, or T all acceptable
	T	0.086	8	0.0250	0.9916	
18	A	0.801	23	> 0.9999	0.0002	dislikes A
	C	0.056	4	0.2073	0.9159	C or G acceptable
	G	0.092	4	0.5480	0.6569	prefers T
	T	0.052	11	< 0.0001	> 0.9999	
22	A	0.052	0	1	0.1062	C or T acceptable
	C	0.064	5	0.1288	0.9503	
	G	0.070	0	1	0.0475	
	T	0.815	37	0.1854	0.9097	

Table S14. Base preferences of flexible residues.

Reselection identifies only one position (18) that is not ideal in the starting sequence. Because the original clone showed activity, however, and all other bases at this position survived the reselection at least as well as the original base, the requirement at this position is 'any base.'

Taken together, the IgE 102min reselection results indicate that the binding motif isolated in the IgE selection is a loop with the sequence indicated in Table S15 flanked by a stem with at least seven out of eight base pairs. The likelihood of this motif occurring in each library is discussed on the following pages.

position:	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
original	T	T	C	A	T	C	C	G	T	A	C	C	T	T	C	T	A	G	T	G	G
consensus	T	T	C G T	A	T	C	C	G	T	A C G T	C	C	T	C or T	C	T	A	G	T	G	G

Table S15. IgE 102min binding motif loop consensus sequence.

IgE Binding Motif Probability Calculation

The probability of the above motif occurring in each library can be calculated as: (loop probability x stem probability), summed for all frames. Because every 37 base frame in N_{60} is the same as every other, the calculation for motif probability for N_{60} can be done once and multiplied by 24 to indicate that it can occur in any register. However, every register along the R*Y* pattern is different, and the motif probability must be calculated separately for each.

IgE Motif Loop Probability

The loop motif probability will be the product of the likelihoods of the required base(s) occurring at each position along the motif. The individual position probabilities reflect the frequencies of the bases in each mix in the starting N_{60} or R*Y* libraries, based on sequencing (Table 1 in the main text; the relevant percentages are reproduced in Table S16).

library	mix	A	C	G	T
N_{60}	N	27.2	23.4	26.6	22.8
R*Y*	N	30.9	23.0	18.1	28.1
	R*	45.8	3.5	47.9	2.8
	Y*	2.7	48.8	3.0	45.6

Table S16. Base ratios at each position in the N_{60} and R*Y* libraries.

The loop probability was determined for every frame in both the N_{60} and R*Y* libraries, using a computer program to assign a value at each bulge/loop position based on the base ratios in the appropriate position in the pattern. These values were then multiplied to determine the overall probability of the motif occurring in each frame (Table S17). In addition, the eight base pairs flanking the loop motif for each frame were qualitatively classified in order to rank the most likely frames (Table S17).

frame	N_{60} loop probability	R*Y* loop probability	R*Y* stem fit	R*Y* likelihood rank	RY loop mismatch
1	2.80×10^{-12}	1.92×10^{-16}	6 on; 2-N		7 off
2	2.80×10^{-12}	6.05×10^{-14}	4 on; 4-N	3	5 off
3	2.80×10^{-12}	6.21×10^{-19}	2 on; 6-N		9 off
4	2.80×10^{-12}	8.48×10^{-15}	1 on; 7-N		6 off
5	2.80×10^{-12}	8.14×10^{-20}	2 on; 4-N; 2 N-N		10 off
6	2.80×10^{-12}	1.19×10^{-15}	4 on; 4 N-N		7 off
7	2.80×10^{-12}	2.02×10^{-19}	2 on; 4-N; 2 N-N		10 off
8	2.80×10^{-12}	3.39×10^{-17}	1 on; 7-N		8 off
9	2.80×10^{-12}	3.87×10^{-17}	2 on; 6-N		7 off
10	2.80×10^{-12}	6.34×10^{-16}	4 on; 4-N		6 off
11	2.80×10^{-12}	1.83×10^{-16}	6 on; 2-N		6 off
12	2.80×10^{-12}	7.16×10^{-17}	7 on; 1-N		7 off
13	2.80×10^{-12}	**4.34×10^{-14}**	**8 on**	**2**	**5 off**
14	2.80×10^{-12}	7.22×10^{-18}	7 on; 1-N		8 off
15	2.80×10^{-12}	**1.06×10^{-11}**	**6 on; 2-N**	**1**	**3 off**
16	2.80×10^{-12}	4.07×10^{-19}	4 on; 4-N		9 off
17	2.80×10^{-12}	5.25×10^{-14}	3 on; 5-N	4	5 off
18	2.80×10^{-12}	5.03×10^{-20}	2 on; 6-N		10 off
19	2.80×10^{-12}	1.05×10^{-14}	4 on; 2-N; 2 N-N	4	6 off
20	2.80×10^{-12}	9.29×10^{-20}	4 on; 2-N; 2 N-N		10 off
21	2.80×10^{-12}	6.04×10^{-16}	2 on; 6-N		7 off
22	2.80×10^{-12}	4.50×10^{-19}	3 on; 5-N		9 off
23	2.80×10^{-12}	9.04×10^{-15}	4 on; 4-N		5 off
24	2.80×10^{-12}	1.28×10^{-17}	6 on; 2-N		7 off

Table S17. Loop occurrence probabilities for every frame in the N_{60} and R*Y* libraries. In addition, stem fits are described to identify the most likely frames.

Of all of the registers along the R*Y* pattern, frame 15, in which the motif was observed in the original selection, fits it the best:

XXXXXXXX-TTCATCCGTACCTTCTAGTGG-XXXXXXXX
RYRYRYRY-NNNRYRYRYRYRYNNNNRYRY-RYRYRYNN

3 loop motif bases off pattern, 6 stem base pairs on pattern/2 stem base pairs with N

The next most likely frame, 13, which lies two positions to the left along the pattern, is:

XXXXXXXX-TTCATCCGTACCTTCTAGTGG-XXXXXXXX
RYRYRYRY-RYNNNRYRYRYRYRYNNNNRY-RYRYRYRY

5 loop motif bases off pattern, 8 stem base pairs on pattern

The stem fits this register better, because all eight base pairs are between an R* and a Y* position, but the loop fits significantly worse, with five off-pattern positions, instead of three.

Comparison of the loop motif to the other 22 registers in R*Y* reveals that every other frame will fit the loop motif less well than these two frames (greater than or equal to five loop motif bases off pattern, less than eight stem base pairs on pattern), and the motif likelihoods in those registers will not be significant in comparison to those of the two most likely frames. Because every frame in the patterned library contains at least three off-pattern positions, this binding motif is not found in the pure RY library.

IgE Motif Stem Probability

Having identified the most likely frames in which the motif could appear and calculated the loop motif probability for each, the stem probabilities were calculated. These calculations were performed for each frame individually.

For the N_{60} library:

Likelihood of any two 'N' making a pair

AT	0.272 x 0.228	= 0.0620
TA	0.228 x 0.272	= 0.0620
CG	0.234 x 0.266	= 0.0622
GC	0.266 x 0.234	= 0.0622
GT	0.266 x 0.228	= 0.0606
TG	0.266 x 0.228	= 0.0606
Total		0.3696

Likelihood of any two 'N' not making a pair 0.6304

Therefore, the probability of at least seven out of eight base pairs in a stem in the N_{60} library

Prob(7/8 base stem) = Prob(3 fixed base pairs) x Prob(4/5 variable base pairs) =
$P(pair)^3 \times [5 \times P(\sim pair) \times P(pair)^4] =$
$(0.3696)^3 \times [5(0.6304)(0.3696)^4] = 0.00297$

Prob(8 base stem) =
$P(pair)^8 = (0.3696)^8 = 0.00035$

Prob(stem) = 0.00297 + 0.00035 = **0.00332**

For the R*Y* library:

Likelihood of any R* and Y* making a pair

AT	0.458 x 0.456	= 0.2088
GC	0.479 x 0.488	= 0.2338
GT	0.479 x 0.456	= 0.2184
TA	0.028 x 0.027	= 0.0008

CG	0.035 x 0.030	= 0.0011
TG	0.028 x 0.030	= 0.0008
Total		0.6781

Likelihood of an R* and Y* not making pair 0.3219

Equivalent calculations determine:

Likelihood of an R* and N making a pair 0.3936

Likelihood of an R* and N not making pair 0.6064

Likelihood of a Y* and an N making a pair 0.3346

Likelihood of a Y* and an N not making a pair 0.6654

Therefore the probability of at least seven out of eight base pairs in a stem in the R*Y* library is:

In the observed frame:

Prob(8 bp stem)

$P(R*N) \times P(Y*N) \times P(R*Y*)^6 = 0.3936 \times 0.3346 \times (0.6781)^6 = 0.0128$

Prob(7/8 bp stem) = Prob(3 fixed base pairs)x Prob(4/5 variable base pairs)

$[P(R*N)xP(R*Y*)^2]x\ [4xP(\sim R*Y*)x\ P(R*Y*)^3x\ P(Y*N) + P(R*Y*)^4x\ P(\sim Y*N)]$

$[(0.3936)(0.6781)^2]\ x\ [4(0.3219)(0.6781)^3(0.3346) + (0.6781)^4(0.6654)] = 0.0498$

Prob(stem) = 0.0128 + 0.0498 = **0.0626**

In the second most likely frame:

Prob(8 bp stem) = $P(R*Y*)^8$ =

$(0.6781)^8 = 0.0447$

Prob(7 bp stem) = Prob(3 fixed base pairs)x Prob(4/5 variable base pairs) =

$P(R*Y*)^3 \times [5x\ P(\sim R*Y*)x\ P(R*Y*)^4)$ =

$(0.6781)^3 \times 5(0.3219)(0.6781)^4 = 0.1061$

Prob(stem) = 0.0447 + 0.1061 = **0.1508**

IgE Binding Motif Total Probability

Based on these values, the overall probability of the full binding motif occurring in the library is calculated in Table S18.

library	frame	loop motif probability	stem probability	registers	total probability
N_{60}	Any	$2.80x10^{-12}$	0.00297	24	$2.0x10^{-13}$
R*Y*	Current	$1.06x10^{-11}$	0.0626	1	$6.6x10^{-13}$
R*Y*	Second	$4.34x10^{-14}$	0.1508	1	$6.5x10^{-15}$

Table S18. IgE-binding motif likelihood in each library

Therefore, the likelihood of the full motif (21 base loop surrounded by a stem with at least seven out of eight base pairs) occurring in the N_{60} library is $2.0x10^{-13}$, while the likelihood of it occurring in the R*Y* library is approximately $6.7x10^{-13}$. The motif is ~3.4-fold more likely in the R*Y* library than in the standard N_{60} library.

Katilius/Woodbury Motif Comparison

D17.4 TTTATCCGTCCCTCCTAGTGG

D17 (original consensus) TTYATCCGTYHCTCYYAGTGG

102 (original) TTCATCCGTACCTTCTAGTGG

102 (reselection)	TTBATCCGTNCCTYCTAGTGG
D17.4 (microarray)	TTBATCCGTHYCTYYYAGTGG

The D17.4 binding motif as determined by Katilius *et al* using a microarray screen is very similar to the 102min motif determined here. They assigned position 18 (102minA numbering) as 'not G' instead of 'N' and assigned all of the light blue semi-conserved positions as 'Y' instead of 'C.' In general, the two techniques identified the same positions as fixed, conserved positions, but differed slightly in the assignment of the semi-conserved or variable positions.

VEGF Binding Motif

Alignment of the variable regions of the 16 unique sequences from round 10 of the VEGF selection revealed that twelve share a common sequence (GTCCGGAATGG-$N_{(0-4)}$-GTGC). In contrast with the streptavidin and IgE cases, however, this consensus sequence was not predicted by OMP to occur in a common context, and variations from the consensus within the 16 unique clones do not correlate with changes in VEGF binding affinity. The motif was therefore considered not sufficiently conserved to enable rigorous probability calculations.

None of the VEGF clones selected here contained the previously published VEGF DNA aptamers: ATACCAGTCTATTCAATTGGGCCCGTCCGTATGGTGGGTGTGCTGGCCAGATAGTATGTGCAATCA[3] or TGTGGGGGTGGACGGGCCGGGTAGA[4]

Effects of Divalent Cation Concentration on Aptamer Function and Predicted Structure

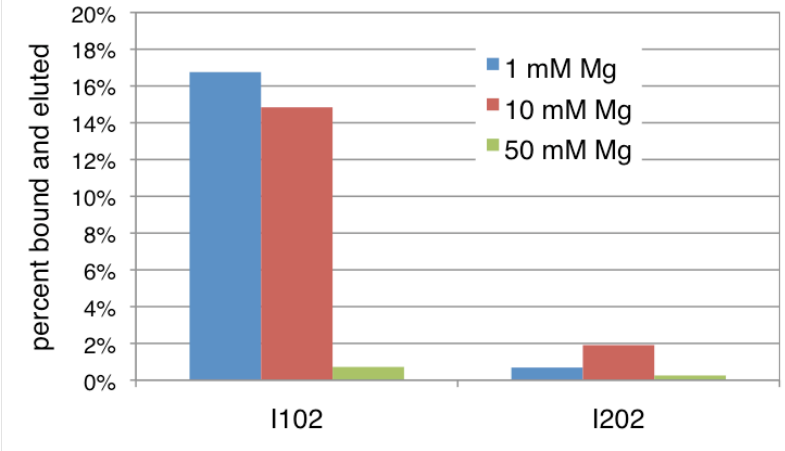

Figure S11. Binding of the best R*Y* IgE aptamer and the best N_{60} IgE aptamer to immobilized IgE is comparable at 1 mM versus 10 mM $MgCl_2$. No significant binding was observed in the presence of 50 mM $MgCl_2$. Clones were bound to IgE beads as described in the main text in PBS supplemented with 1, 10, or 50 mM $MgCl_2$.

	energy of folding (average ± standard deviation in kcal/mol)	
library	1 mM Mg	10 mM Mg
N_{60}	-16.2 ± 3.3	-18.3 ± 3.3
R*Y*	-20.6 ± 3.6	-22.4 ± 4.0
ΔΔG	-4.4	-4.1
Z-score	1.33	1.24

Table S19. Predicted folding energies using OMP of N_{60} and R*Y* libraries in 1 mM versus 10 mM $MgCl_2$. Z-score is defined as (ΔΔG/standard deviation).

	1 mM Mg			10 mM Mg		
	N_{60}	R*Y*	ratio	N_{60}	R*Y*	ratio
top 5% N_{60}	5%	34%	6.8	5%	31%	6.2
I102	25%	69%	2.8	22%	66%	3.0
Top 3 VEGF	8%	41%	5.1	9%	43%	4.8

Table S20. Percentage of N_{60} and R*Y* libraries that are predicted to be at least as structured as aptamers in 1 mM versus 10 mM $MgCl_2$.

References Cited

(1) Legiewicz, M.; Lozupone, C.; Knight, R.; Yarus, M. *RNA* **2005**, *11*, 1701-9.
(2) Workman, C. T.; Yin, Y.; Corcoran, D. L.; Ideker, T.; Stormo, G. D.; Benos, P. V. *Nucleic Acids Res* **2005**, *33*, W389-92.
(3) Hasegawa, H.; Sode, K.; Ikebukuro, K. *Biotechnol Lett* **2008**, *30*, 829-34.
(4) Nonaka, Y.; Sode, K.; Ikebukuro, K. *Molecules*, *15*, 215-25.