Supplementary Data

MPBind: A Meta-Motif Based Statistical Framework and Pipeline to Predict Binding Potential of SELEX-derived Aptamers

Peng Jiang[1], Susanne Meyer[2], Zhonggang Hou[1], Nicholas E. Propson[1], H. Tom Soh[3], James A. Thomson[1,2,4], Ron Stewart[1,*]

[1]Morgridge Institute for Research, Madison, WI 53707, USA

[2]Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, CA 93106

[3]Departments of Mechanical Engineering and Materials, University of California, Santa Barbara, CA 93106

[4]Department of Cell and Regenerative Biology, University of Wisconsin, Madison, WI 53706

Supplementary Methods

Given a motif length, MPBind will enumerate all possible n-mers (e.g., 4096 motifs for 6-mers) and calculate the frequency of each motif in the random sequence region from each round of SELEX-Seq. Then it will calculate four kinds of statistical tests:

Statistical Test 1:
We assume that high binding motifs should be enriched in the final SELEX round when compared to the control round. The control round can be either the initial library sequencing (R0) or sequencing rounds controlled by PCR cycles without target selection. MPBind will calculate the total number of occurrences of each motif (e.g., TGAGTT) in the final round as well as in the control round and compare these numbers to the total number of occurrences of all other motifs in these rounds. A one-sided Fisher's exact test (right tail) is calculated for each motif. For example, a motif has 100 total occurrences in the final round of SELEX-Seq and 50 occurrences in the control round. Assuming the total number of occurrences of all possible motifs in SELEX-Seq and Control-Seq are 1000 and 800, respectively, a 1-sided P-value is calculated for this motif based on a two by two table ([100, 1000-100] Vs. [50, 800-50]).

Statistical Test 2:
We assume that in the final round of SELEX-Seq, the percentage of reads which contain high binding n-mers should be enriched when compared to the control round. Thus for each motif, we calculate the number of motifs containing reads as well as the number of reads which do not contain this motif in the final SELEX-Seq and in the control round. A similar one-sided Fisher's exact test (right tail) is calculated.

Statistical Test 3:
We assume that the relative frequency of binding motifs should increase with each SELEX round. The relative frequency of each motif is defined as the total number of occurrences of motifs divided by the total number of unique motif positions within all random sequence regions for a given round. A one-sided Spearman correlation is calculated for each motif by the relative motif frequency against the SELEX round numbers.

Statistical Test 4:
We assume that the percentage of reads, which contain binding motifs, should increase with each SELEX round. A one-sided Spearman correlation is calculated for each motif based on the percentage of reads containing this motif against the SELEX round numbers.

For each p-value, we transform it to Z-Score:

$$Z = \Phi^{-1}(1-p) \qquad (1)$$

where Φ is the standard normal cumulative distribution function.

Thus for each motif, we have 4 kinds of Z-scores (Z1, Z2, Z3 and Z4). We further use Stouffer's method to combine these 4 Z-Scores into one motif level Z-score:

$$Z = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \qquad (2)$$

For any given aptamer sequence, we use an n-mer window to scan it with all potential n-mer motifs. The Meta-Z-Score is calculated as the aggregate of motif level Z-scores for all potential motifs across the aptamer using Stouffer's method (formula (2)). An overview of the MPBind method is shown in Figure S1.

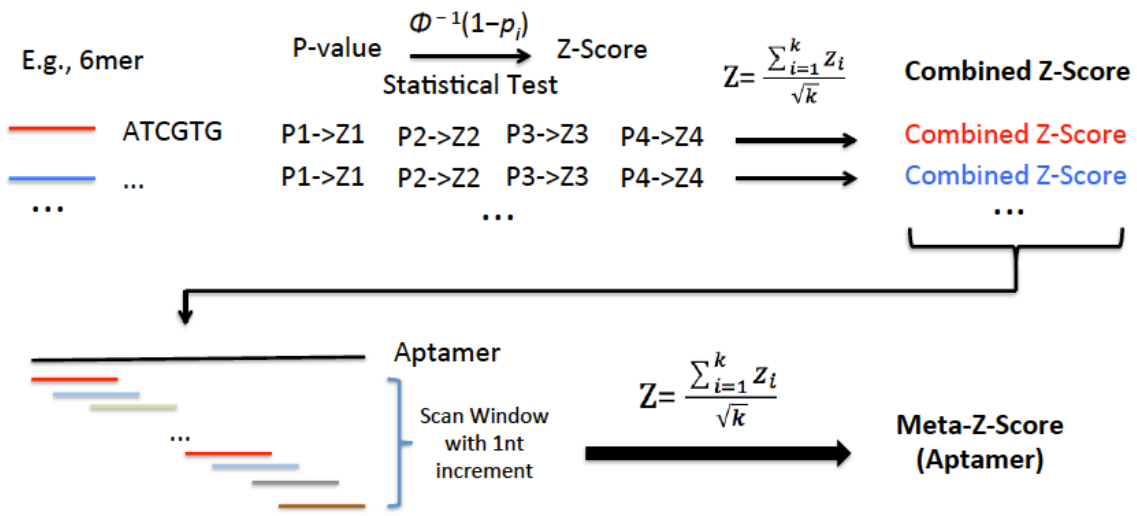

Figure S1. Overview of the MPBind method. A combined Z-Score is determined for each motif, then each aptamer is scanned with each motif-level Z-Score to arrive at a Meta-Z-Score for each entire aptamer.

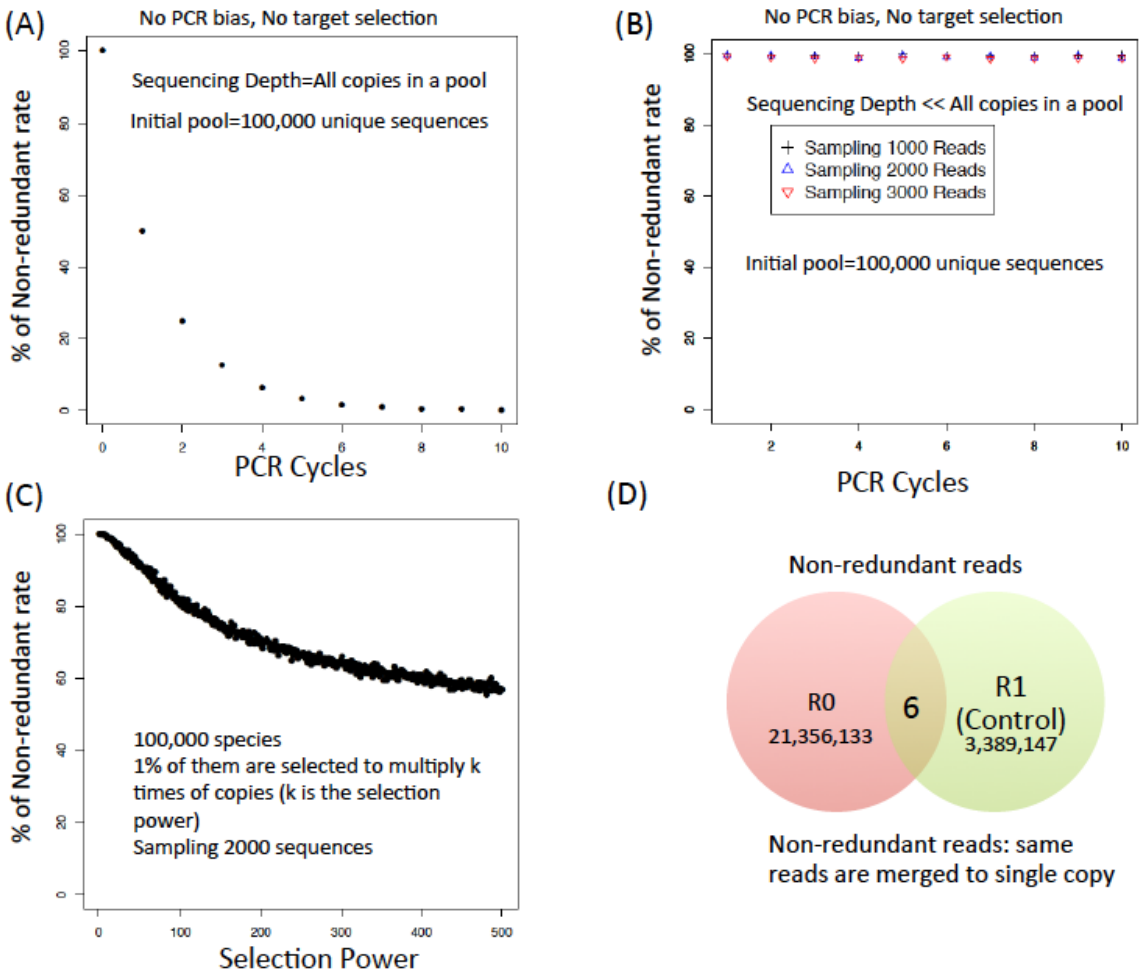

Figure S2. Simulations for factors affect read complexity. (A) If the sequencing depth is deep enough and equal to all copies of reads in a pool, the read complexity will drop with PCR cycle increase. This is because PCRs generated more redundant reads and the read complexity in this scenario indicates the overall redundancy in a pool. (B) If read depth is far less than all reads in a pool, the PCR cycles will not affect read complexity. This is because if there is no target selection and no PCR bias, PCR cycles may not change the relative abundance of each read. Thus the read complexity in this scenario indicates the equality of reads being sampled among species. In other words, a low read complexity might indicate some species are more likely being sampled than the others. (C) We add a selection power to 1% of species to simulate the scenario that some of reads are more likely being sampled than the others. The read complexity dropped with increasing the extend of selection power for those 1% favored reads (possibly by target selection or PCR bias). (D) The overlap of our initial sequencing pool (R0) and Control-Seq is minimal. Only 6 reads are present in both sequencing runs. It indicates that our sequencing read depth is far from enough to cover all the reads in a pool.

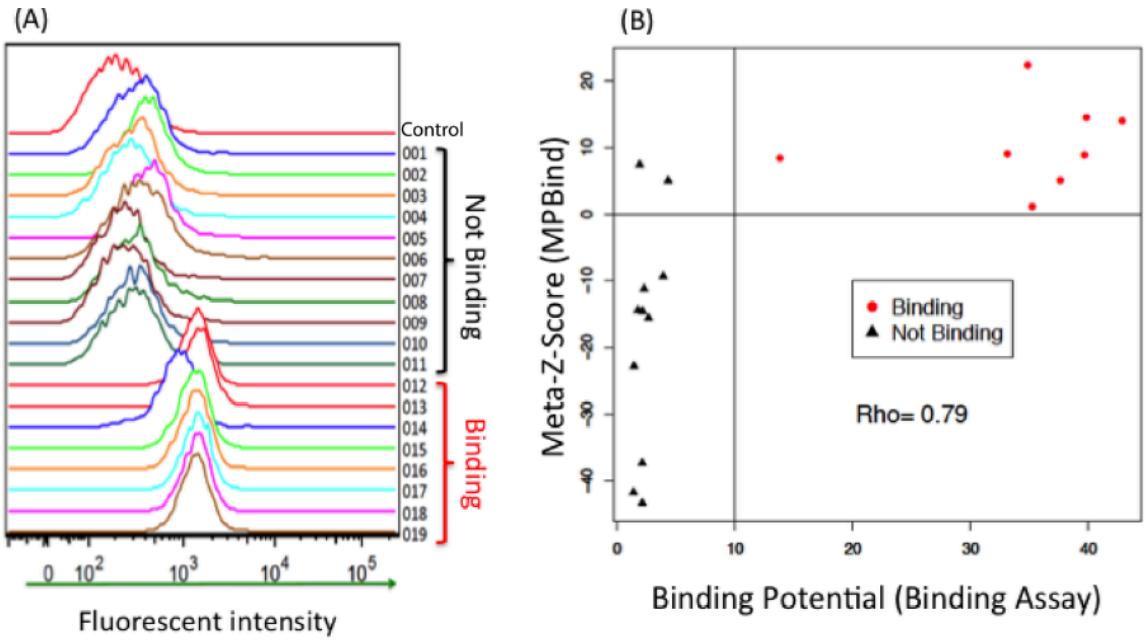

Figure S3. Experimental validation of MPBind prediction. (A) Binding assay: hESCs cells were individualized with Accutase and incubated with 500nM FAM labeled DNA aptamer for 30min in aptamer binding buffer (PBS (with Ca^{2+}/Mg^{2+}), 5mM $MgCl_2$, 0.45% glucose, 0.1% BSA). Cells were then washed with binding buffer and incubated with an AlexaFluor 488 labeled anti-FITC antibody. Fluorescent intensity of the cells was then analyzed by flow cytometry on BD FACSCantoII. (B) The correlation between Meta-Z-Score and binding potential. The binding potential of each aptamer is estimated by comparison of fluorescent intensities between aptamer and control (no aptamer). We did bootstrapping (10,000 times with replacement) of fluorescent intensities for each aptamer and control, respectively. Then we calculated the binding potential as the ratio of fluorescent intensities with aptamer>control versus aptamer<=control during the bootstrapping.

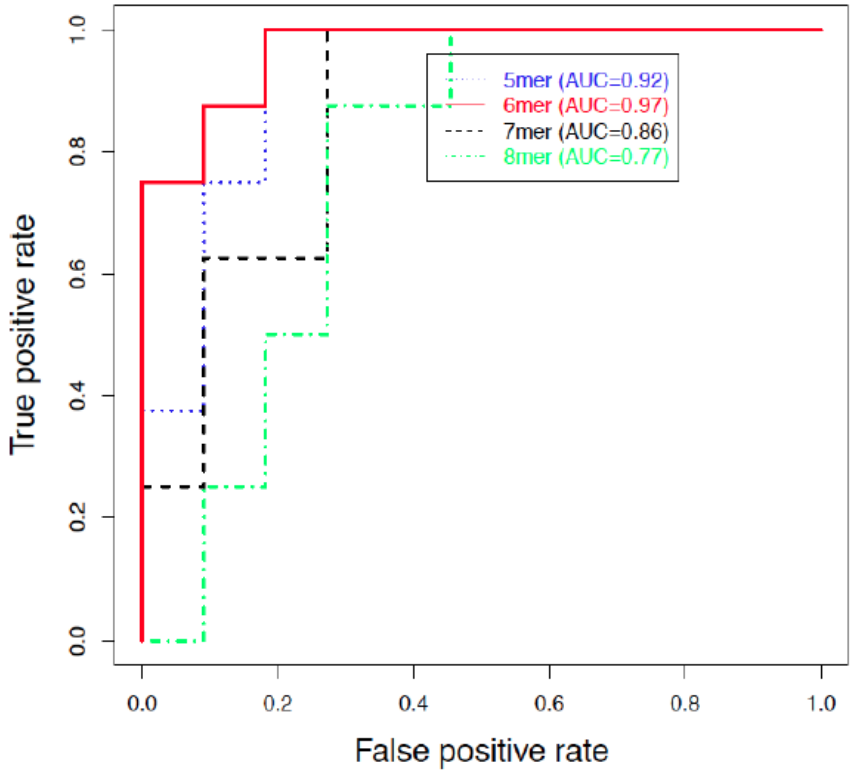

Figure S4. Receiver operating characteristic (ROC) curves for the aptamer level Meta-Z-Score using different motif lengths. Non-redundant reads are used in training MPBind. Areas under the curves (AUC) are given in the legend. 6-mers give the best results for the dataset used.

Table S1. Reads summary for each round

SELEX				Control			
	Total	Non-Redundant	%		Total	Non-Redundant	%
R0 (initial)	-	-	-	-	21,478,037	21,356,133	99.43
R1	4,199,837	2,139,527	50.94	Control-1	7,214,886	3,389,174	46.97
R2	21,315,209	19,739,762	92.61	Control-2	11,637,693	10,510,033	90.31
R3	24,850,609	24,364,510	98.04	Control-3	23,876,529	23,293,992	97.56
R4	25,253,924	19,332,976	76.55	Control-4	15,355,495	12,950,877	84.34
R5	15,327,604	4,381,160	28.58	Control-5	25,442,016	24,256,995	95.34

Non-Redundant: Number of reads where redundant reads have been removed. For instance, if there are 55 counts for the exact same read, this is considered as just 1 read.

Table S2. Top enriched aptamers in SELEX round 1 are also highly enriched in Control round 1

	Aptamer	Read Count
SELEX (R1) [Read Counts>80]	**CTAGATCGGAAGAGCTCGTATGCCGTCTT**	269
	GGAAGAGCTCGTATGCCGTCTTCTGCTTG	201
	TGCTGCTAGATCGGAAGAGCTCGTATGCC	130
	AGATCGGAAGAGCTCGTATGCCGTCTTCT	86
Control (R1) [Read Counts>80]	**GGAAGAGCTCGTATGCCGTCTTCTGCTTG**	474
	TGCTGCTAGATCGGAAGAGCTCGTATGCC	228
	AGATCGGAAGAGCTCGTATGCCGTCTTCT	121

Of the top four aptamers (read counts > 80) in round 1 SELEX-Seq, three are the top three aptamers in Control-Seq. Same reads are marked with the same color.

Table S3. Aptamer binding assay on human embryonic stem cells

ID	Aptamer	Meta-Z-Score	Binding
001	ACTTATTTGTCTTAAGTGGCGGGTCAATG	-11.25	No
002	GCAGGTGTGGTTTGCTGAGGTGGGCCCTG	-9.37	No
003	GTGGGCGCACTTAGACGGGGTGATCGTAA	-43.32	No
004	GGGTCCCTTCGGGGTGACGATGGTATCTA	-14.37	No
005	TTTGGTTTGCTGTATGGTGGGCTCTGTTA	4.96	No
006	GGTGTGGGGAGGGTCGTATTGTGTCCTGT	-15.55	No
007	TCGCTTGAACGGGGAACTACTCCAGACGT	-41.74	No
008	CTATTTGTTCTAGTGGCGGTCATCTAAGG	-14.51	No
009	GGTGAGGCGGACGTATCTTTTAGCAAATC	-22.83	No
010	GTGAGGGTGAGGACAGGTTAGCGTGGTGG	-37.31	No
011	CTTATTTGTGTTTAGTGGCGGGCGTTTGT	7.39	No
012	AGGGTATGGACTTCGACGTCTCGGCTGAA	14.59	Yes
013	AGGAGGGGGACTTAGGACTGGGTTTAGGG	14	Yes
014	CGCACAGGAAGGTATGGACTTCGACGTTT	8.4	Yes
015	TATCCGACTTGGATGGCTGAGCAAGGCTA	1.21	Yes
016	AGTATCTATCCGACTTGGATTTACGTTCG	8.91	Yes
017	GAAATATGGACTTCGATACGCCGGCTGAG	5.07	Yes
018	GGTATGGACTTCGACGTCTTCTGACCTAA	22.3	Yes
019	AGGAGGGGGACTTAGGACTGGGTTTATGA	9.01	Yes

Meta-Z-Score is calculated based on non-redundant reads from R0 to R5 with motif length set to 6nt.

Table S4. The prediction performance of using Meta-Z-Score or Z1, Z2, Z3, Z4

Prediction	AUC
Meta-Z-Score	**0.97**
Z1	0.95
Z2	0.95
Z3	0.93
Z4	0.93

Non-redundant reads are used in training MPBind. The motif length is set to 6nt.

Table S5. The prediction performance of MPBind on another SELEX-Seq data (ESRP1 SELEX-Seq)

	Aptamer	Meta-Z-Score	Binding Assay (EMSA)
WT A	CCGCGTGTGGGTGTGTCCGA	32.8	Bind
Mut A	CCGCGTGTAGATGTATCCGA	-23.17	Not Bind
WT B	CTCGTGTCGGTGTGGGGTAG	29.46	Bind
Mut B	CTCGTGTCGATGTAGAGTAG	-11.54	Not Bind
WT C	GTGGGTTCGGTGGTGGGTAG	37.3	Bind
Mut C	GTGAGTTCAGTAGTGAGTAG	-31.04	Not Bind
WT D	CCGGTGTGGGGTTGGGACGG	35.59	Bind
Mut D	CCGATGTGAGGCTAGGACGG	1.83	Not Bind

The ESRP1 SELEX reads were downloaded from the publication (Dittmar, et al., 2012). Dittmar, et al. generated 5 rounds of SELEX-Seq (ESRP1) data (R0, R2 R3 R6 and R7). For each round, we merged reads to unique reads (removed redundant reads) and trained MPBind with parameter n-mer=6. Four aptamers (WT A, WT B, WT C and WT D) are selected by Dittmar, et al. for binding validation using Electrophoretic Mobility Shift Assay (EMSA) analysis with increasing amounts of GST-ESRP1 (0 to 250ng). These four aptamers showed significant binding to ESRP1. Our MPBind prediction showed that these 4 aptamers have Meta-Z-Scores: 32.8, 29.46, 37.3 and 35.59, respectively. To further confirm the binding, Dittmar, et al., made 3-4 point mutations to each aptamer as controls. The EMSA did not show significant binding for those mutant aptamers. The predicted Meta-Z-Scores for these mutant aptamers (controls) are -23.17, -11.54, -31.04 and 1.83, respectively (Table S5). This indicates that MPBind can correctly predict aptamers that bind to ESRP1.

Reference

Dittmar, K.A., et al. (2012) Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing, *Molecular and Cellular Biology*, 32, 1468-1482.