Cell, Volume *139*

Supplemental Data

Profiling the Human Protein-DNA Interactome

Reveals ERK2 as a Transcriptional

Repressor of Interferon Signaling

Shaohui Hu, Zhi Xie, Akishi Onishi, Xueping Yu, Lizhi Jiang, Jimmy Lin, Hee-sool Rho, Crystal Woodard, Hong Wang, Jun-Seop Jeong, Shunyou Long, Xiaofei He, Herschel Wade, Seth Blackshaw, Jiang Qian, and Heng Zhu

Supplemental Experimental Procedures

Identifying Tissue-specific Motifs

We developed a program to identify tissue-specific motifs. We first defined sets of tissue-specific or tissue-enriched genes by examining their gene expression profiles across multiple tissues (Yu et al., 2006). We then calculated the most over-represented single motifs (8-mers, including a wide character) in the promoters of each set of tissue-specific genes. The program then enumerated all possible combinations of the top *n* motifs (e.g. *n* = 100). For each motif pair, the program recorded the occurrence of the motif pair in the promoter sequences. We then calculated the significance score for each motif pair, which was defined as the negative logarithm of the *p* value, -log(*p*). The motif pairs with scores above a specified threshold were considered putative TF binding motif pairs in the promoter sequences. With these predicted motif pairs, we could calculate a number of partners for each motif and select a certain number of top non-redundant motifs to be tested in the protein chip experiments.

Both the *p* values for a single motif and those for a motif pair were calculated using hypergeometric distribution. Here, we use a motif pair as an example to show the procedure. The *p* value of occurrence of the motif pair (*i*, *j*), $P_{occ}^{i,j}$, is calculated according to

$$P_{occ}^{i,j} = \sum_{k \ge g_{i,j}} \frac{C_n^k C_{N-n}^{G_{i,j}-k}}{C_N^{G_{i,j}}}, \tag{1}$$

where *N* is the number of all human promoters; *n* is the number of tissue-specific genes; $G_{i,j}$ is the number of human promoters that contain the motif pair (*i*,*j*), and $g_{i,j}$ is the number of tissue-specific promoters that contains the motif pair. C_n^k is the number of possible combinations, using *k* members from a set of size *n*.

Selection of DNA Motif Sequences

The total number of computationally predicted DNA motifs is 896, including 174 in (Xie et al., 2005), 233 in (Xie et al., 2007), 272 in (Elemento and Tavazoie, 2005), 73 in (Elemento et al., 2007), and 144 predicted in this study. To remove redundant DNA motifs that were highly similar, we compared the similarity scores among the 896 DNA motifs (Figure S1A). The sequence similarity (*S*) between two motifs, m1 and m2, is defined as

$$S_{m1,m2} = \frac{s(m1, m2)}{\min(length(m1), length(m2))}, \qquad (2)$$

where *s*(m1,m2) is the maximal number of matched nucleic acids between m1 and m2. The value of $S_{\mathrm{m1,m2}}$ is equal to one if m1 is identical to m2, or m1 is a part of m2 (or vice versa). The value of $S_{\mathrm{m1,m2}}$ is zero if m1 and m2 share no common nucleic acids.

We then compared the similarity between motif pairs and randomly removed one of the motifs if the similarity between the pair was greater than a defined cutoff value. This list consisted of 400 DNA motifs when we used a cutoff value of 0.9 (Figure S1B).

In addition to these predicted DNA motifs, we chose 60 DNA motifs from the TRNASFAC SITE (9.0) database (Wingender et al., 1996) that had known target TFs that were included in our protein chips.

Protein Annotation

To define known TFs, we first searched the GO database for the human proteins associated with the GO terms, including: transcription factor activity (0003700), RNA polymerase II TF activity (0003702), RNA polymerase III TF activity (0003709), transcription activator activity (0016563), and transcription repressor activity (0016564) (Ashburner et al., 2000). In addition, on the basis of extensive literature search by expert biologists, we added well-known TFs that were not included in the GO database.

Transcriptional coregulators were excluded from the TF list and were annotated as a separate functional category. Predicted TFs were defined as proteins containing TF DNA-binding domains that were annotated by the Pfam database but had not been established as TFs on the basis of any experimental evidence (Table S13) (Finn et al., 2006). Protein kinases were annotated on the basis of the list from www.kinase.com (original paper published in Science 2002, updated in Dec, 2007) (Manning et al., 2002). In addition, we added protein kinases that had been verified experimentally by our labs. RNA-binding proteins were annotated based on the GO term “RNA binding” (0003723) and its offspring terms. Nucleic acid-binding proteins were defined as proteins that were associated with the GO term “nucleic acid binding” (0003676) and its offspring terms but were not in the TF and RNA binding list. Chromatin-associated proteins were annotated based on the GO term “chromosome organization and biogenesis” (0051276) and its offspring terms. Mitochondrial proteins were proteins whose cellular location is in the mitochondrion (data obtained from P. Onyango, personal communication). Proteins that were not annotated into the groups listed above were grouped into “all other categories,” and their molecular functions are summarized in Table S3. The version of GO database used was that from February 2008. All the annotations were checked manually and were corrected after searching the literature if any protein was mistakenly annotated by the GO database.

Protein Microarray Data Analysis

Image scan: Protein microarray chips were scanned using GENEPIX PRO 5.0. We manually checked all the spots on the 460 chips and adjusted the size and position for the spots skewed by artifacts, such as dust or specks.

Background correction: To quantify the signal intensity for each spot, we calculated the signal intensity for each spot, which was defined as the foreground median intensity divided by its local background median intensity. A signal intensity close to 1 indicated that the protein in that spot did not bind to the DNA motif probe. The higher the signal intensity, the stronger the binding of that protein to the target DNA sequence.

Within-chip normalization: To eliminate spatial artifacts that can arise from uneven mixing of the probe or uneven washing and drying of the chips, we performed a within-chip normalization for each chip by assuming the signal distribution of all the blocks in a chip was consistent across the chip and the median signal intensity of each block was equal to 1. This assumption was based on the fact that the proteins were randomly printed on the chip, and only a small portion of the proteins (on average, <2%) bound to the target DNA sequences. Therefore, we normalized signal intensities (*I*) of a set of spots within a block in a chip by setting the median intensity of that block equal to one,

$$\hat{I}_{i,j} = I_{i,j} - median(I_j) + 1, \qquad (3)$$

where $\hat{I}$ is the signal intensity after within-chip normalization, *i* is the protein index in a block, and *j* is the block index in the chip.

Identifying positive hits: To identify proteins that bind to a DNA motif probe (positive hits), an intensity cutoff value needed to be assigned for each chip. A cutoff was defined as a number of standard deviation(s) (SD) away from the mean of the signal intensities for all the spots in a chip, and spots producing a signal greater than the cutoff were identified as “positive hits.” However, it has been frequently observed that some spots have very strong signals in protein chips. In such cases, a cutoff value defined by the method described above would produce arbitrarily high values and yield high false-negative rates. To tackle this problem, we generated a signal intensity distribution for proteins without DNA-binding activity and determined the SD from their distribution.

We first identified the proteins with signal intensities less than one (left-hand side of the mean of the blue curve in Figure S19). Symmetric pseudo-data for the right side of the mean were then generated to estimate the SD (right-hand side of the mean of the blue curve in Figure S19). Finally, we used a cutoff value of six SDs from the mean to identify positive hits (Table S4). Moreover, since each protein was printed in duplicate on a chip, a protein was counted as a positive hit only if both of its duplicated spots were identified as positive.

Non-specific binding filtering

We recognized that some proteins might bind to Cy5 directly and therefore produce signals in the absence of DNA motifs, and some proteins might bind to double-stranded T7 (the primer sequence) directly. To exclude these proteins from our list of “true” PDIs, we used four negative control experiments, assessing two chips probed with Cy5 only and two probed with T7 only. Any protein identified as a positive hit from one of these four experiments was filtered out from the target list for further data analysis. In total, 134 proteins were identified and eliminated on the basis of the negative control experiments.

DNA Motif Logo Discovery

We used AlignACE (Roth et al., 1998) to discover significant DNA motif logos. Multiple DNA logos were generated using a number of AlignACE parameters, including expect motif length or seed number, for each protein or for each protein family, in the case of generation of familial logos. The convergent logo was chosen. Degenerate DNA motif logos (significant nucleic acids were all separated in the logos) were excluded. Proteins bound to fewer than 30 motifs were considered “sequence-specific binding proteins” and were included in our further analysis.

DNA Binding Motif Analysis of ERK2

We first searched for significant DNA binding motifs among the 17 DNA sequences (with spacers) bound by ERK2 using AlignACE (Hughes et al., 2000), and we found a highly conserved position weight matrix (PWM), [G/C]AAA[C/G], comprising four possible variations: GAAAC, GAAAG, CAAAG, and CAAAC. To calculate whether these motifs were enriched in the promoter regions of the up-regulated genes identified by the ERK2-knockdown microarray, we retrieved promoter sequences of 82 genes (Xuan et al., 2005) in which the promoter region was defined as extending from -700 bp of the transcription start site (TSS) to 300 bp of the TSS. Enrichment analysis revealed that one of the ERK2 binding motifs, GAAAC, was highly enriched in the promoter regions of these up-regulated genes (p=1.5e-9, hypergeometric test with the whole human promoter regions as background), whereas GAAAG showed weak enrichment (p=0.014). On the other hand, CAAAG and CAAAC did not show any statistical enrichment (p=0.513 and 0.638, respectively). Application of MDscan (Liu et al., 2002) to the 82 promoter sequences revealed that GAAAC was the most significant potential DNA binding site, confirming the results from the enrichment analysis.

Construction of the Correlation Network

We first defined the distance between the DNA-binding profiles of two proteins. The distance (*D*) between the DNA-binding profiles of two proteins (A and B) was calculated according to

$$D_{A,B} = \left(\frac{\sum_{i=1,\ldots,m} \left(1 - \max_{j=1,\ldots,n}(S(i,j))\right)}{m} + \frac{\sum_{j=1,\ldots,n} \left(1 - \max_{i=1,\ldots,m}(S(i,j))\right)}{n} \right) / 2, \qquad (4)$$

where *S* is the similarity score defined by Eq.2, *m* is the number of motifs to which protein A binds, and *i* is its motif index, *n* is the number of motifs to which protein B binds, and *j* is its motif index.

We then calculated the pairwise distance between the DNA-binding profiles for all the proteins showing specific binding activity (binding motifs <30), including TFs and unconventional DNA binding proteins, according to Eq.4. The histogram of all the distances is shown in Figure S20. We arbitrarily chose a cutoff value of 0.1 to define proteins with highly correlated DNA binding profiles. All protein pairs with distances less than 0.1 were then used to construct the network. The network was visualized using Cytoscape 2.6.0 (Cline et al., 2007).

Supplemental Figures

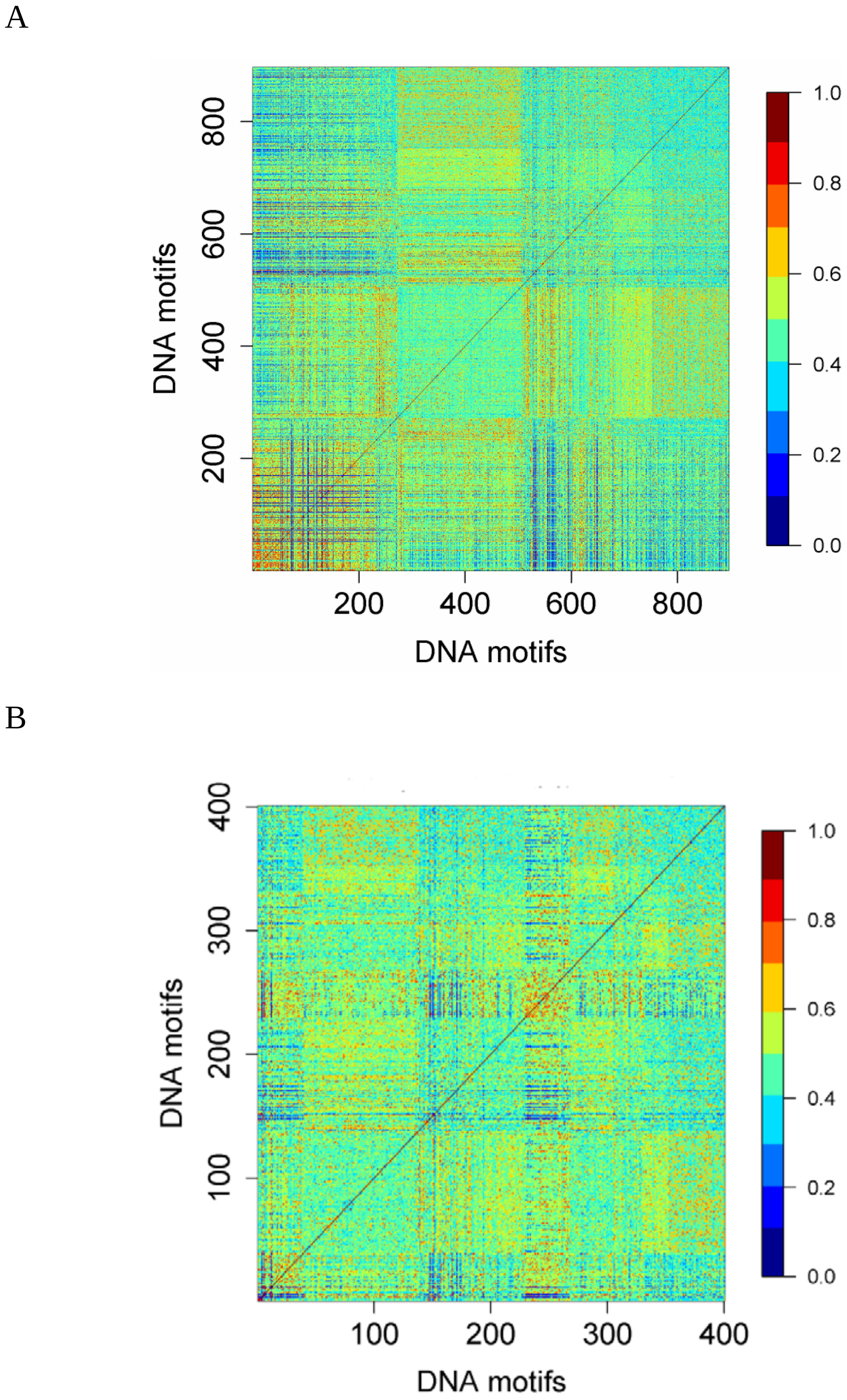


Figure S1. Heatmap of similarity scores between DNA motifs.

(A) Pairwise similarity scores for 896 input DNA motifs.

(B) Pairwise similarity scores for 400 DNA motifs after reduction.

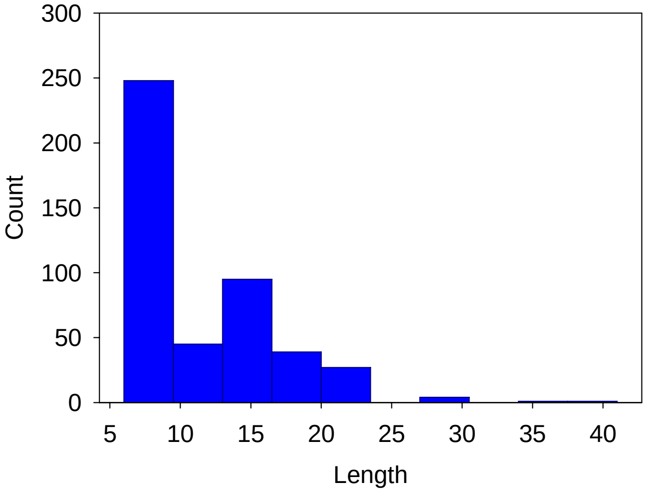


Figure S2. Histogram of motif length.

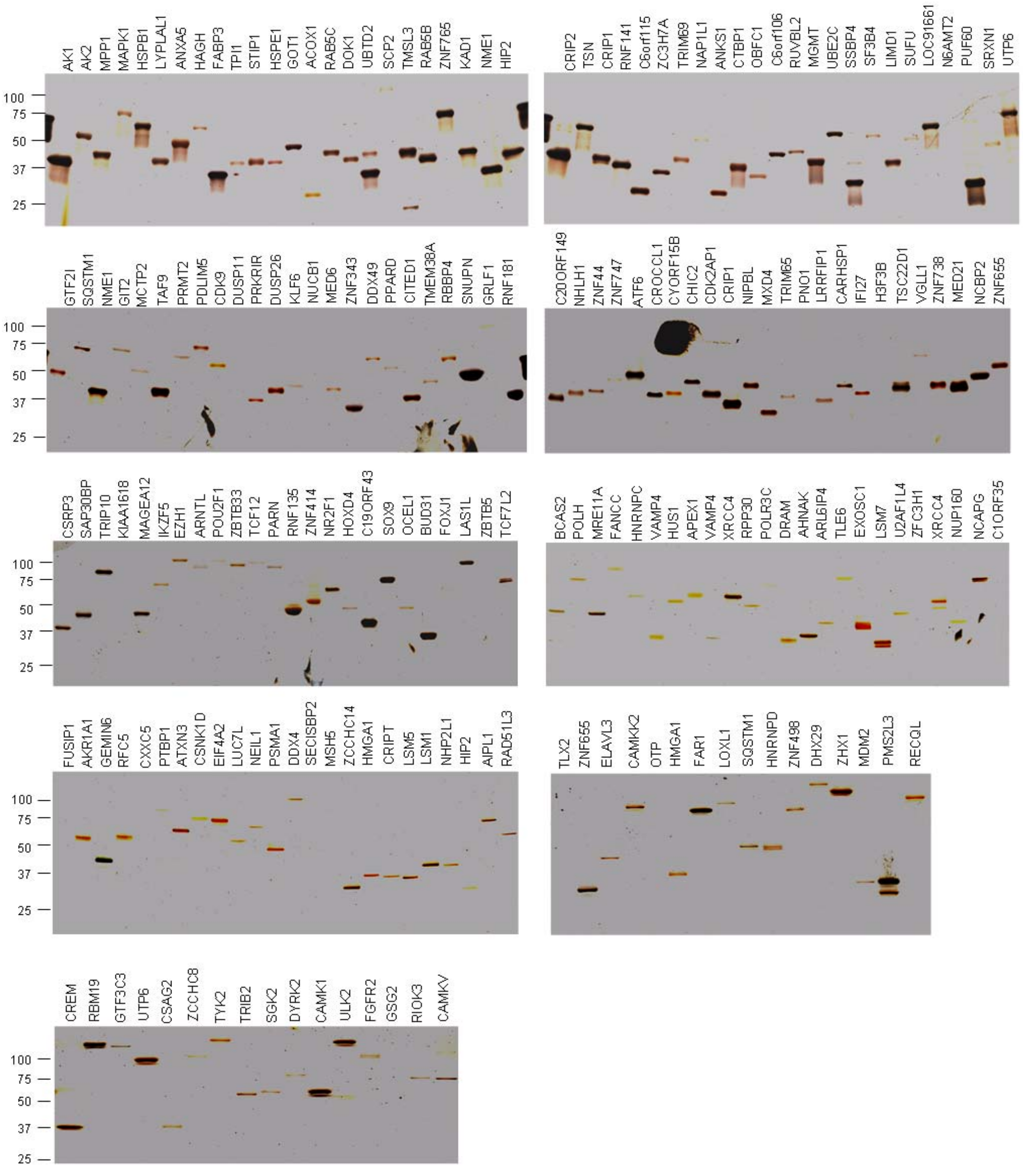

AK1 AK2 MPP1 MAPK1 HSPB1 LYPLAL1 ANXA5 HAGH FABP3 TPI1 STIP1 HSPE1 GOT1 ACOX1 RAB5C DOK1 UBTD2 SCP2 TMSL3 RAB5B ZNF765 KAD1 NME1 HIP2
CRIP2 TSN CRIP1 RNF141 C6orf115 ZC3H7A TRIM69 NAP1L1 ANKS1 CTBP1 OBFC1 C6orf106 RUVBL2 MGMT UBE2C SSBP4 SF3B4 LIMD1 SUFU LOC91661 N6AMT2 PUF60 SRXN1 UTP6
GTF2I SQSTM1 NME1 GIT2 MCTP2 TAF9 PRMT2 PDLIM5 CDK9 DUSP11 PRKRIR DUSP26 KLF6 NUCB1 MED6 ZNF343 DDX49 PPARD CITED1 TMEM38A RBBP4 SNUPN GRLF1 RNF181
C20ORF149 NHLH1 ZNF44 ZNF747 ATF6 CROCCL1 CYORF15B CHIC2 CDK2AP1 CRIP1 NIPBL MXD4 TRIM65 PNO1 LRRFIP1 CARHSP1 IFI27 H3F3B TSC22D1 VGLL1 ZNF738 MED21 NCBP2 ZNF655
CSRP3 SAP30BP TRIP10 KIAA1618 MAGEA12 IKZF5 EZH1 ARNTL POU2F1 ZBTB33 TCF12 PARN RNF135 ZNF414 NR2F1 HOXD4 C19ORF43 SOX9 OCEL1 BUD31 FOXJ1 LAS1L ZBTB5 TCF7L2
BCAS2 POLH MRE11A FANCC HNRNPC VAMP4 HUS1 APEX1 VAMP4 XRCC4 RPP30 POLR3C DRAM AHNAK ARL6IP4 TLE6 EXOSC1 LSM7 U2AF1L4 ZFC3H1 XRCC4 NUP160 NCAPG C1ORF35
FUSIP1 AKR1A1 GEMIN6 RFC5 CXXC5 PTBP1 ATXN3 CSNK1D EIF4A2 LUC7L NEIL1 PSMA1 DDX4 SECISBP2 MSH5 ZCCHC14 HMGA1 CRIPT LSM5 LSM1 NHP2L1 HIP2 AIPL1 RAD51L3
TLX2 ZNF655 ELAVL3 CAMKK2 OTP HMGA1 FAR1 LOXL1 SQSTM1 HNRNPD ZNF498 DHX29 ZHX1 MDM2 PMS2L3 RECQL
CREM RBM19 GTF3C3 UTP6 CSAG2 ZCCHC8 TYK2 TRIB2 SGK2 DYRK2 CAMK1 ULK2 FGFR2 GSG2 RIOK3 CAMKV
100
75
50
37
25

Figure S3. Silver staining analysis of 200 randomly selected human proteins purified from yeast. Molecular weights (kD) are indicated to the left.

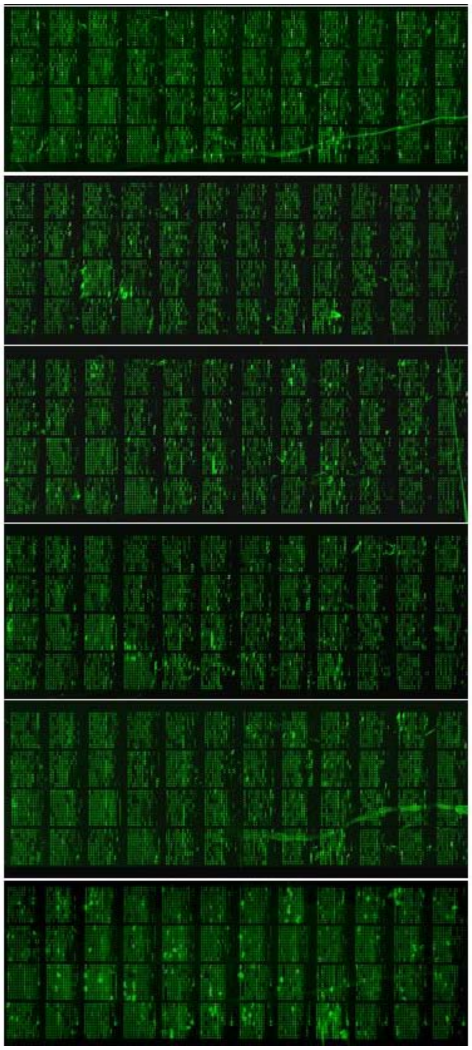


Figure S4. Protein microarrays probed with an anti-GST antibody. All the 4,191 non-redundant human proteins were printed in duplicates into 48 blocks. Anti-GST antibody was probed to check the quality of the microarrays. Proteins positively detected by the anti-GST antibody are represented in green and more than 98% of the spots on each microarray produced signals above background. Pairwise correlation coefficients of signal intensities between these slides ranged from 0.90–0.95. Each microarray contains 10,752 spots. The 4,191 proteins were printed in duplicate and occupied 8,382 spots. The rest spots either were printed with many control proteins (e.g., BSA, histones, IgGs, etc.) without GST tag, or left empty. Therefore, these spots were seen with extremely weak or no signal.

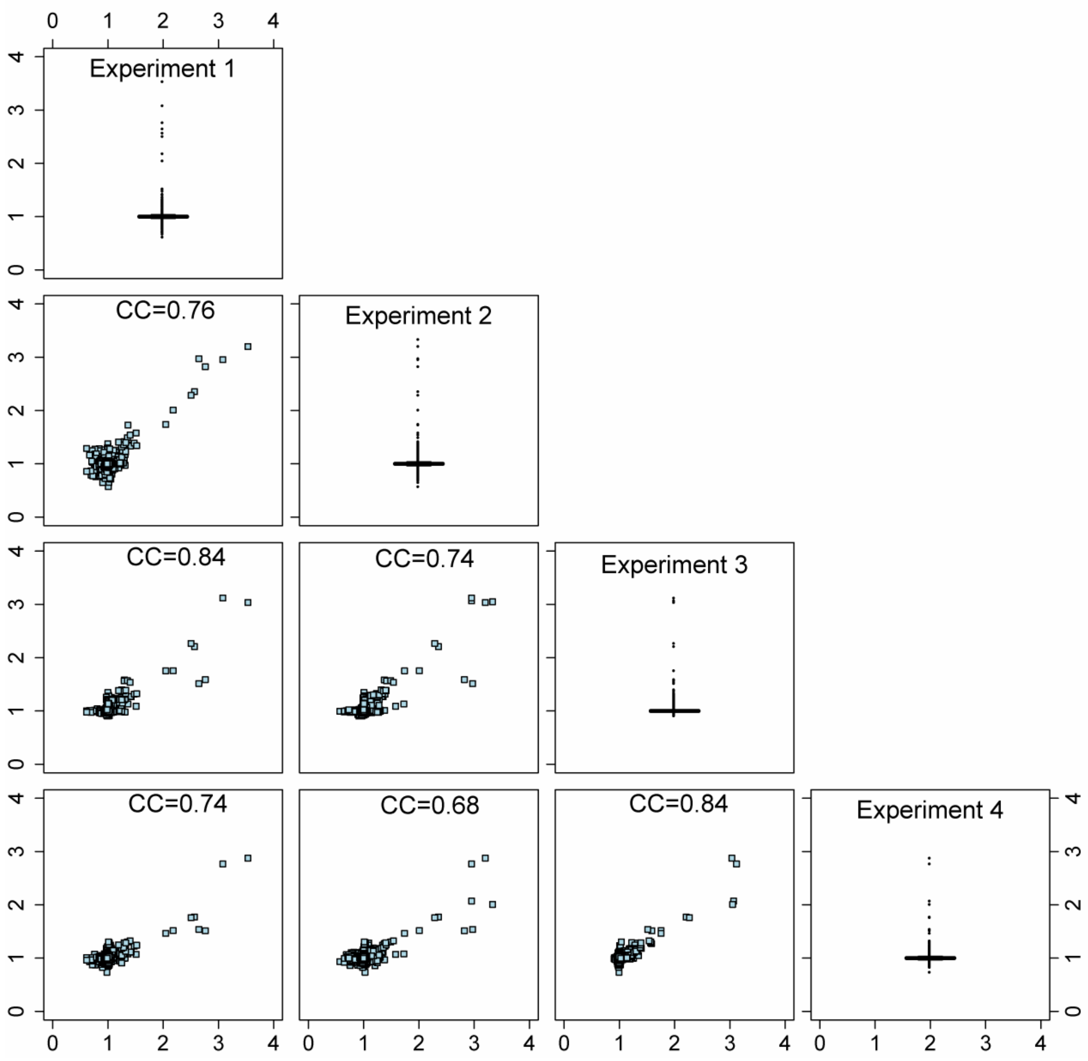


Figure S5. Boxplot and pairwise scatterplot of four replicated protein microarray experiments. Boxplot produces box-and-whisker plot of signal intensities (median foreground intensity / median background intensity) of a chip before normalization. Scatterplot compares the signal intensity of the spots between every two experiments. Each spot in the scatterplots represents one protein. X- and Y-axis are signal intensities. Note that the spots with high intensities are the positive hits. CC denotes correlation coefficient.

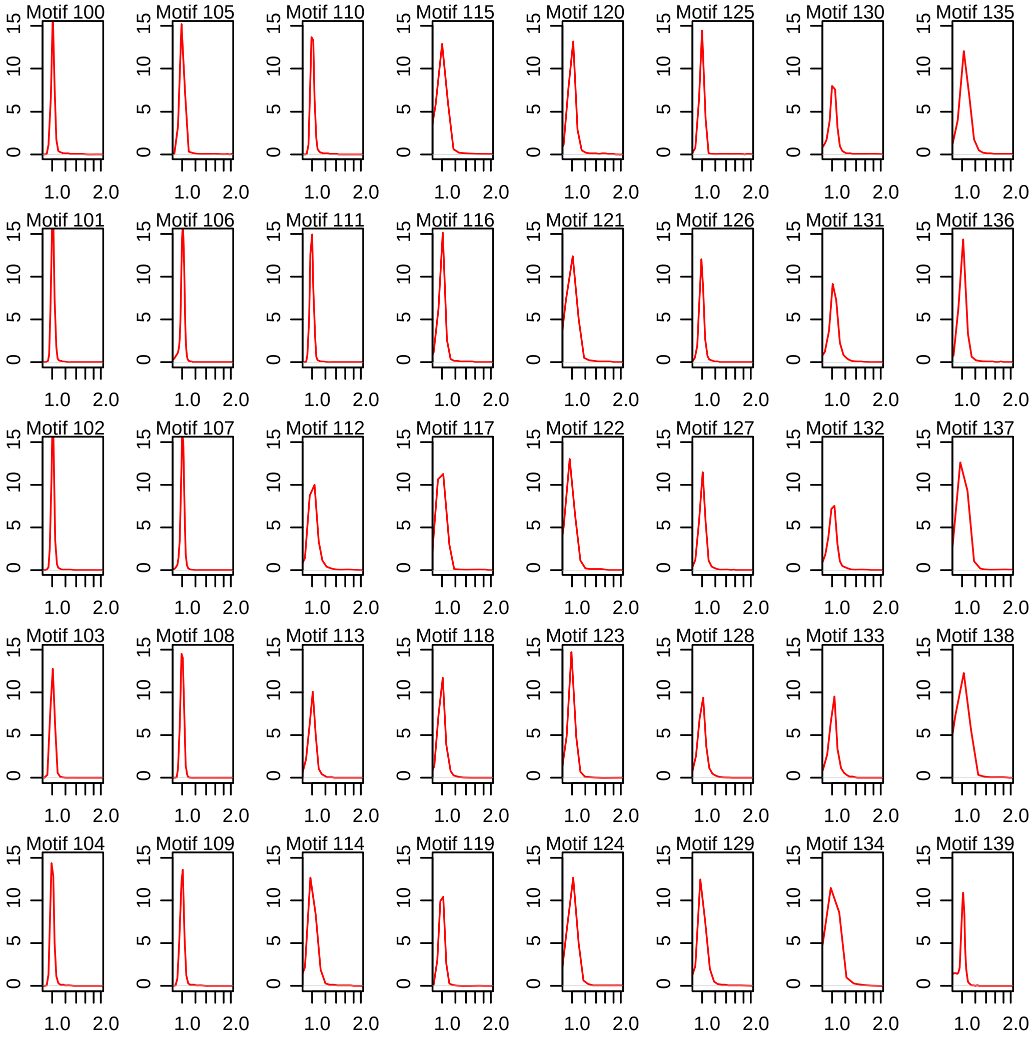


Figure S6. Density plots of signal intensity of 40 sample microarrays before normalization. The x-axis denotes signal intensity, and the y-axis denotes density of signal intensity.

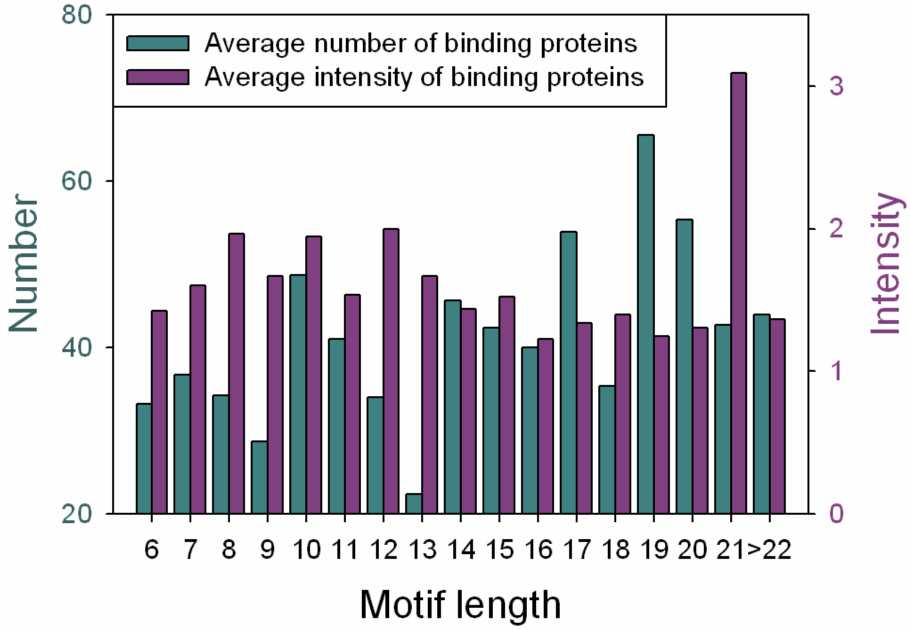


Figure S7. Motif length versus the number of binding proteins and the average signal intensity of binding proteins.

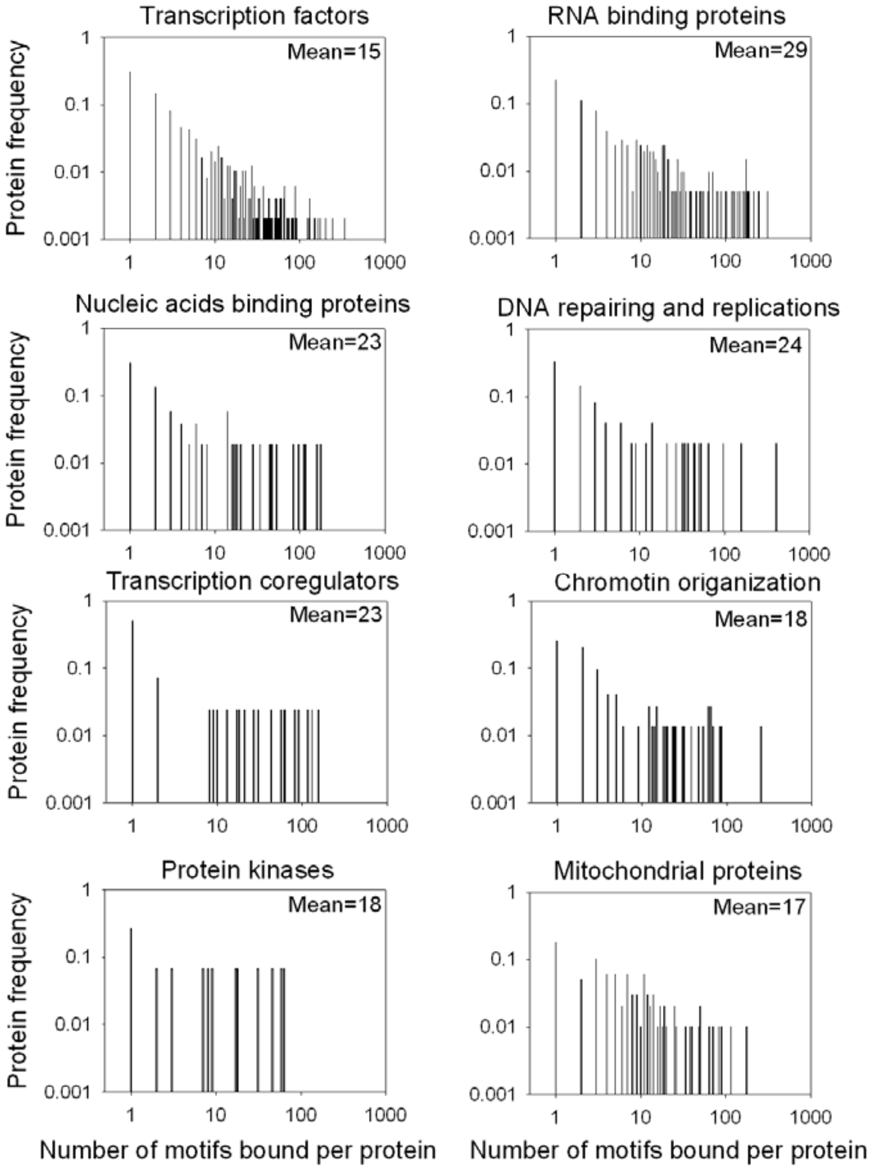


Figure S8. DNA binding specificity of different protein classes.

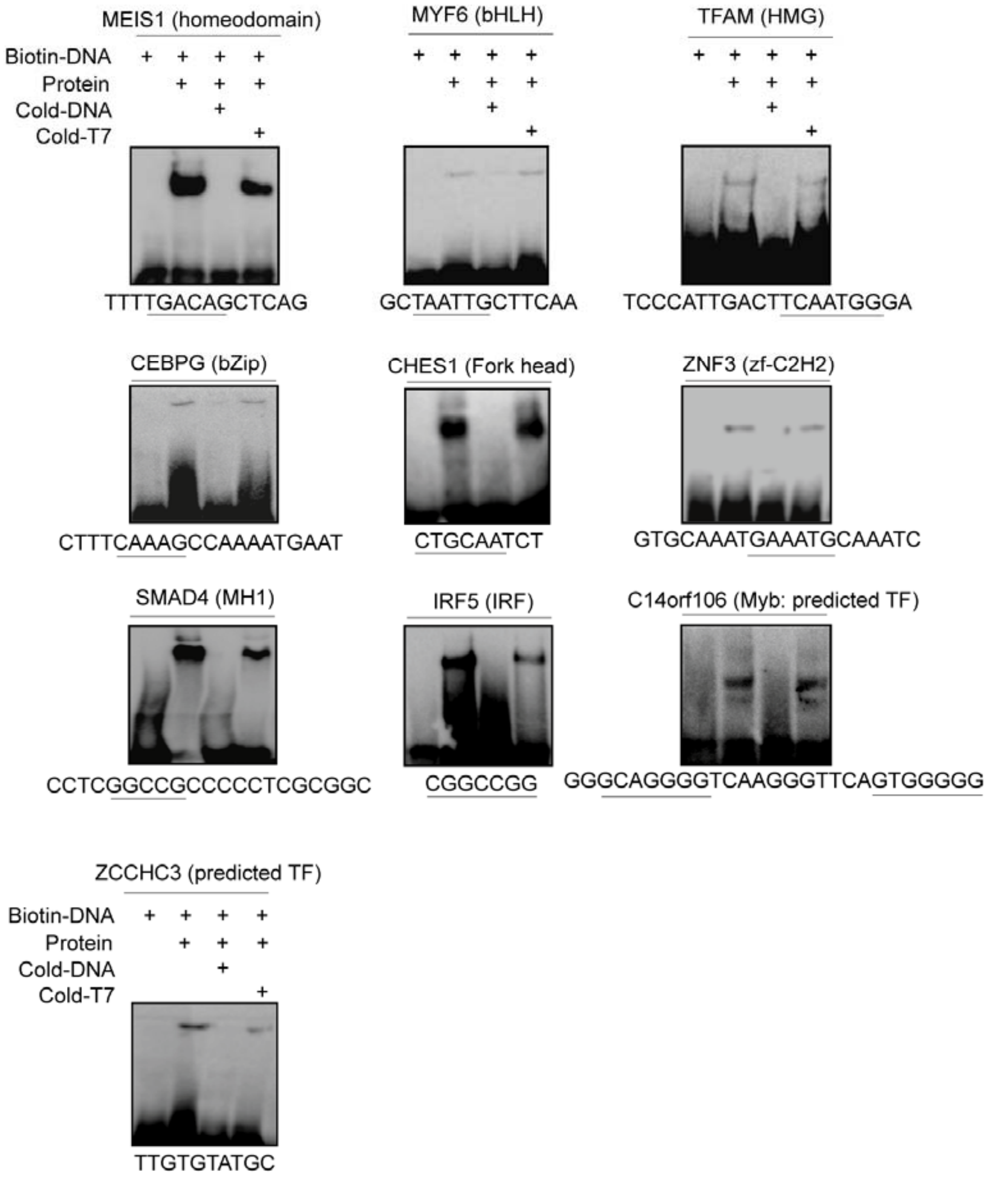


Figure S9. Validation of newly identified PDIs using EMSA analysis. Representative examples from the 9 subfamilies are shown, along with an example of a predicted TF that does not belong to any of these subfamilies.

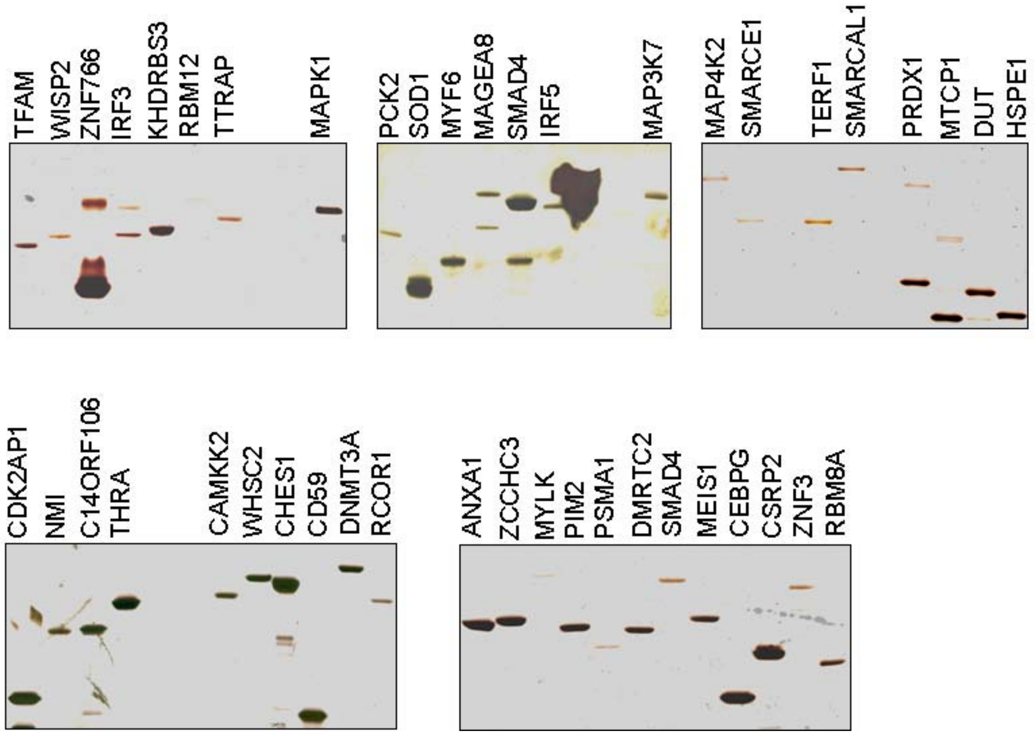


Figure S10. Silver staining images of proteins used in the EMSA assays.

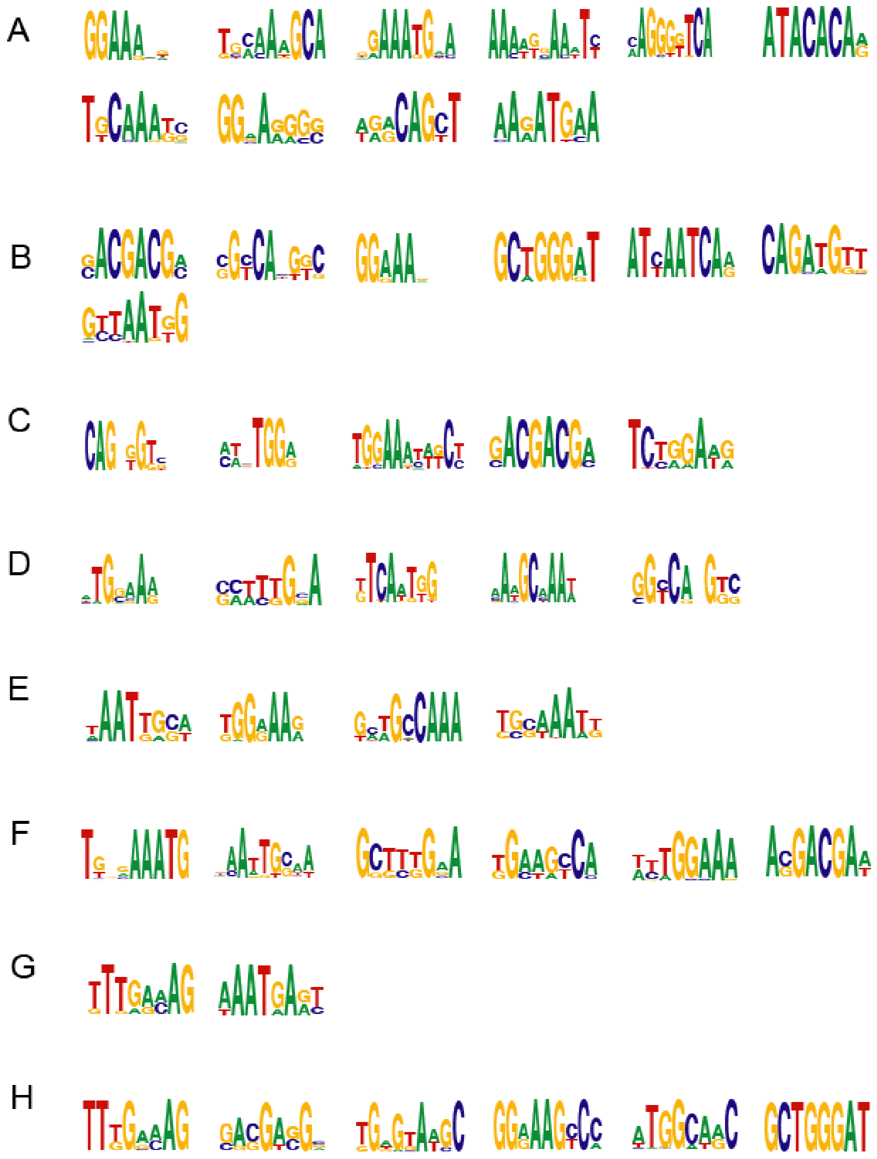


Figure S11. Significant familial logos of unconventional DNA binding proteins.
(A) RNA binding proteins.
(B) Mitochondria proteins.
(C) Chromatin associated proteins.
(D) Transcriptional coregulators.
(E) Proteins associated with DNA repairing and replications.
(F) Nucleic acid binding proteins.
(G) Protein kinases.
(H) All other categories.

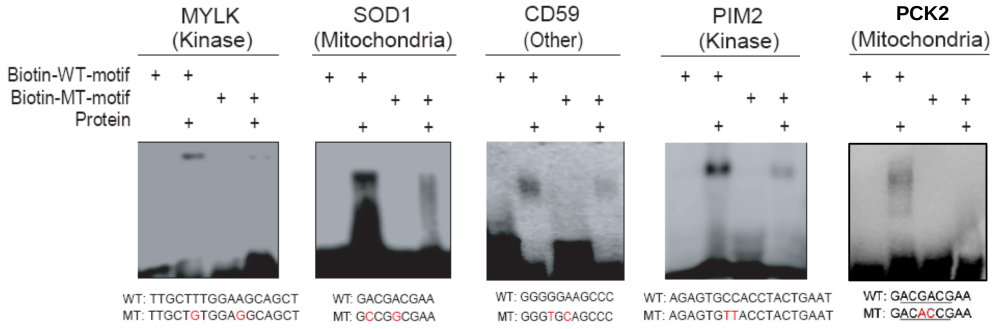


Figure S12. EMSA assays for four unconventional DNA-binding proteins. The mutant (MT) motifs for MYLK, SOD1, CD59, PIM2, and PCK2 showed significantly reduced binding activities compared to the wild-type (WT) motifs.

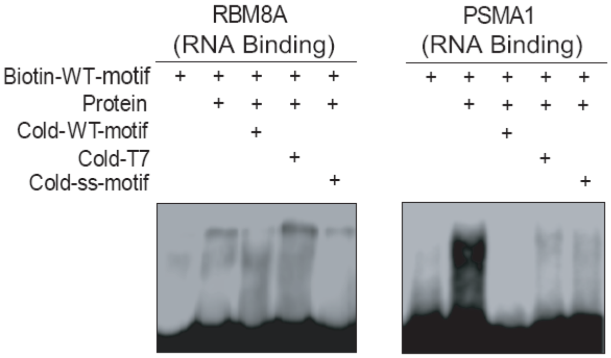


Figure S13. EMSA assays for RNA binding proteins RBM8A and PSMA1. Unlabeled dsDNA wild-type motifs efficiently competed for binding, while ssDNA had little effect.

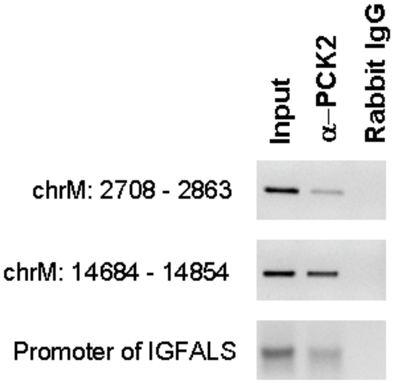


Figure S14. PCK2 is associated with DNA *in vivo* using ChIP coupled with PCR amplification. DNA fragments of PCK2-ChIPed mitochondrial DNAs are indicated as chrM: 2708 – 2863 and chrM: 14684 – 14854. PCK2 was also found to ChIP with the promoter of a chromosomal gene IGFALS.

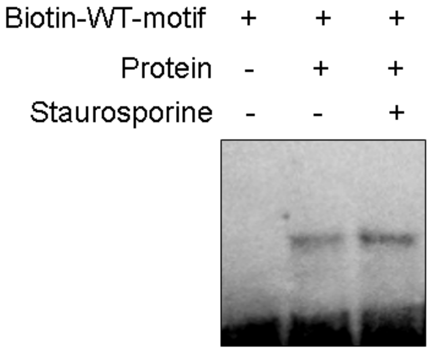


Figure S15. EMSA assay with *E. coli* purified ERK2 co-expressed with MEK1. The presence of staurosporine, a kinase inhibitor, did not affect the DNA binding activity of ERK2.

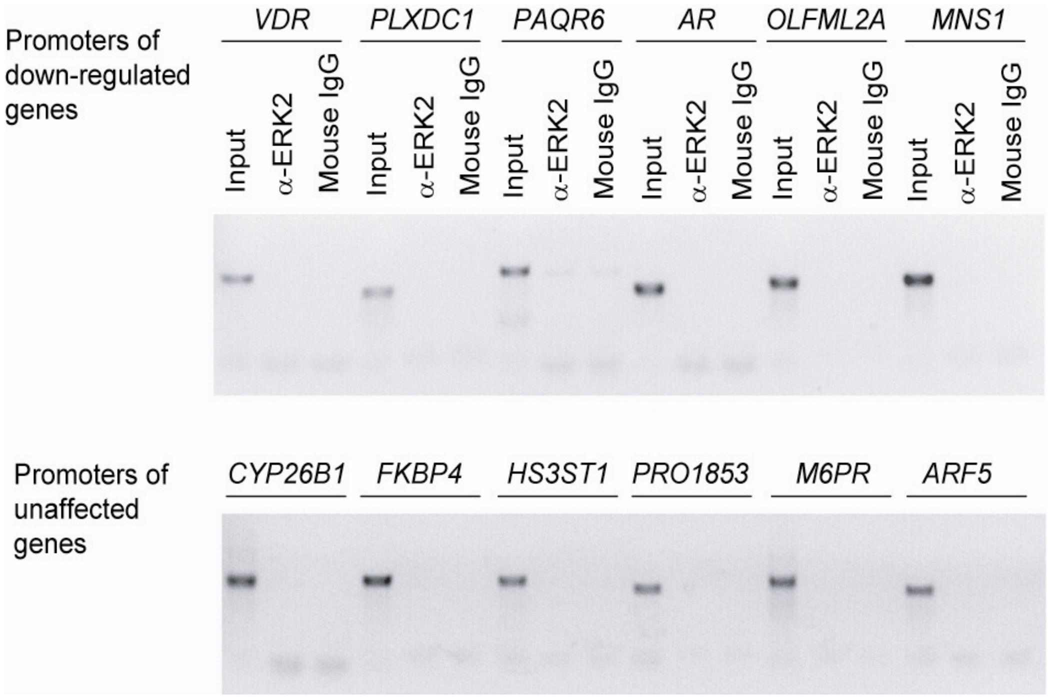


Figure S16. ChIP-PCR analysis of six down-regulated genes induced by ERK2 knockdown and six unaffected genes. The anti-ERK2 antibody did not show enrichment in any of these genes relative to the IgG control.

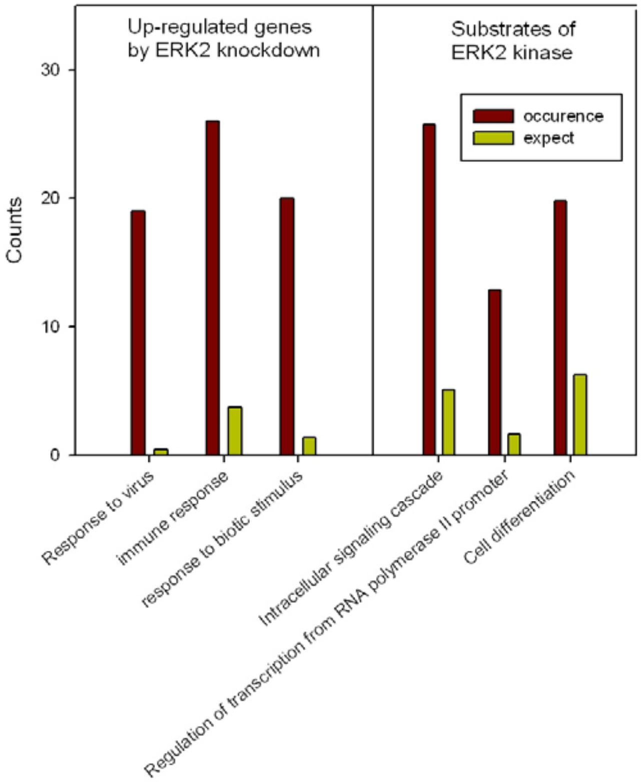


Figure S17. Significantly enriched GO terms of two gene sets, up-regulated genes by ERK2 knockdown and substrates of ERK2 kinase (*p* <0.001 using Fisher's exact test corrected for multiple testing).

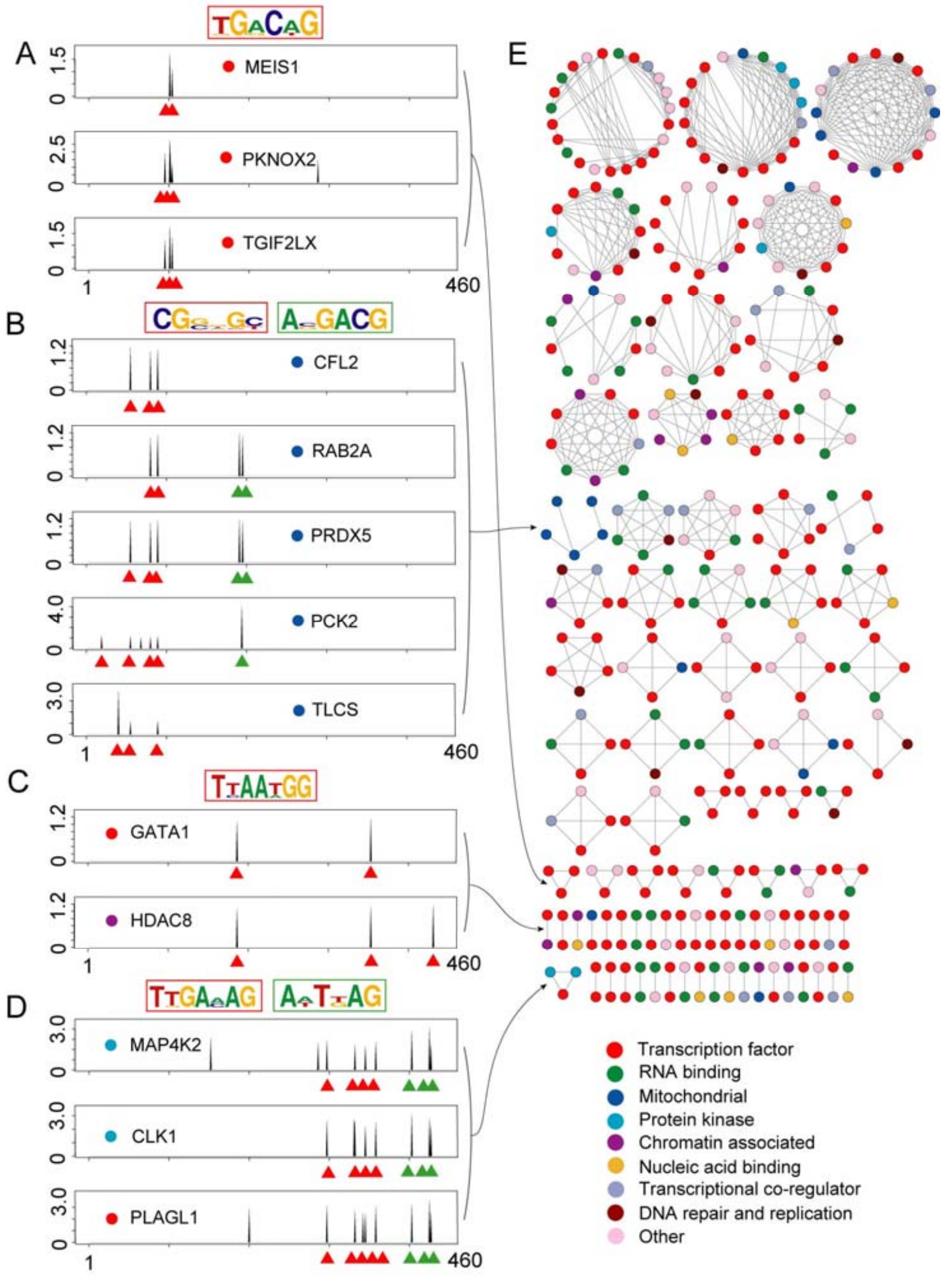


Figure S18. Correlation Network of the Target-Preference of All DNA-Binding Proteins Tested in the Study. (A-D) Examples of proteins sharing similar DNA binding profiles. Each peak represents the normalized signal intensity of a specific DNA motif probe, with individual motifs organized along the X-axis by sequence similarity. Binding peaks used to generate the major logo (outlined in red) are indicated by red triangles. For proteins that recognize more than one logo (outlined in green), binding peaks for the second logo are indicated in green. (E) Correlation network for proteins with highly similar DNA binding profiles (see Supplemental Data for construction of the network). Proteins of different function classes are color-coded. Proteins from different classes can share similar binding sites, indicating a potential crosstalk between unconventional DNA-binding proteins and annotated TFs.

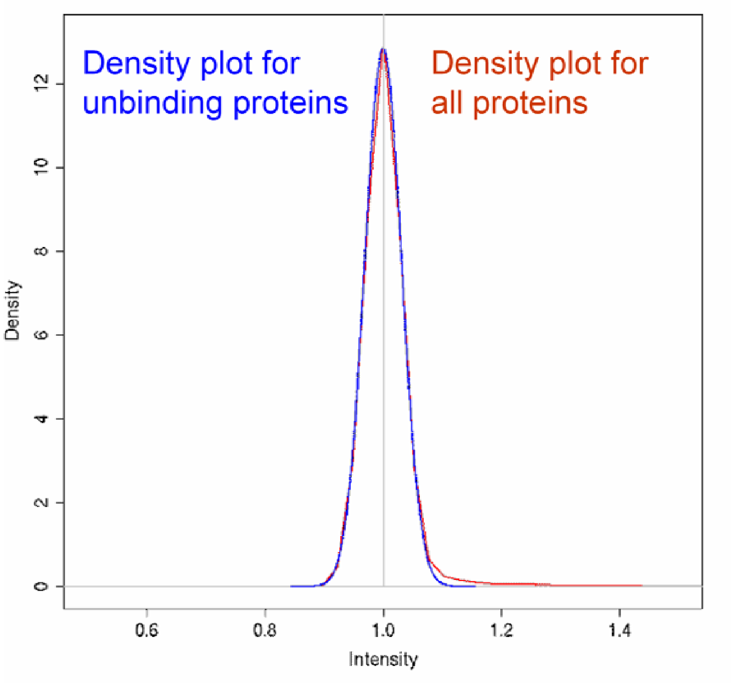


Figure S19. Density plot of signal intensity of all the spots in a protein microarray and that of negative hits in the microarray.

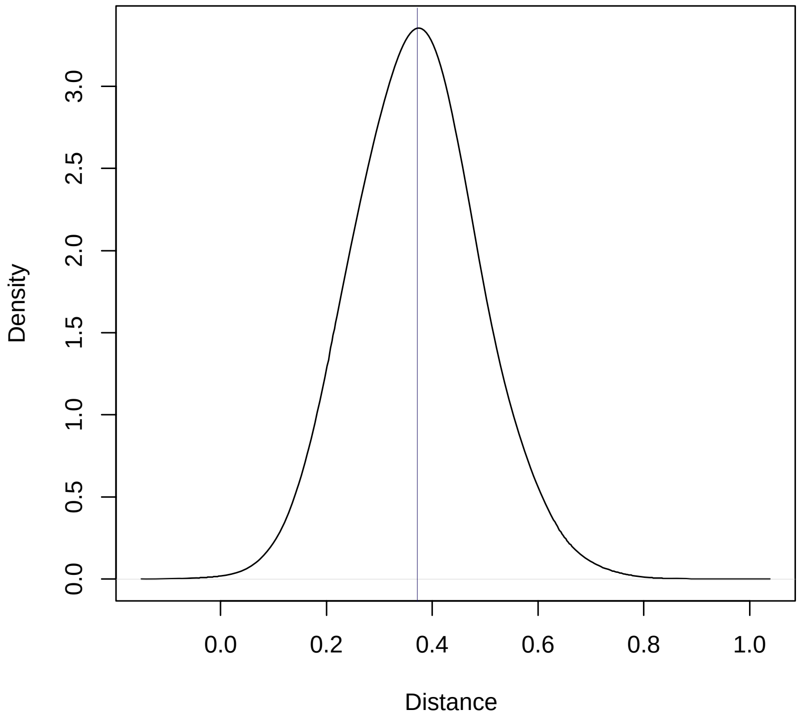


Figure S20. Histogram of the DNA-binding profile distance for all the proteins.

Supplemental Tables

Table S1 (DNA_motifs.xls) and Table S2 (Protein_annotation.xls) are uploaded separately.

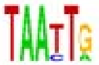
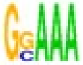
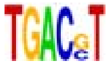
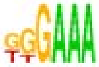
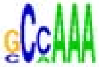
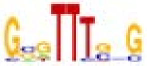
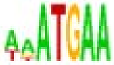
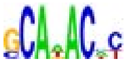
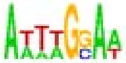
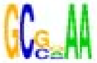
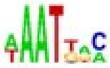
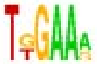
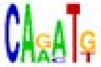
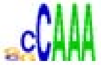
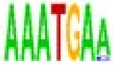
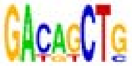
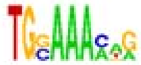
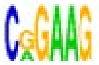
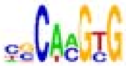
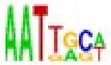
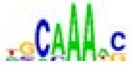
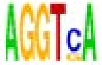
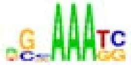
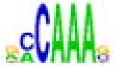
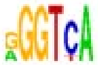
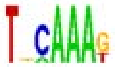
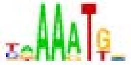
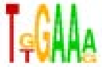
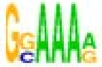
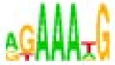
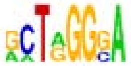
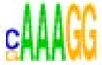
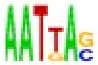
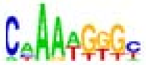
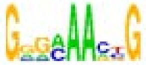
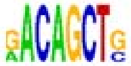
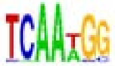
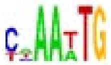
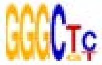
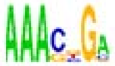
Table S3. Major molecular function categories of other classes of proteins annotated by the GO database. Note that some proteins have multiple GO terms.

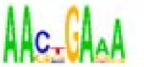
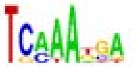
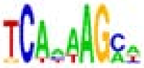
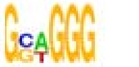
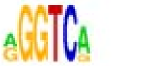
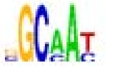
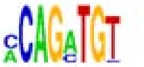
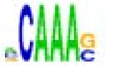
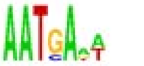
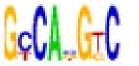
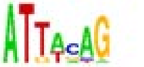
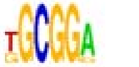
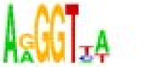
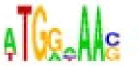
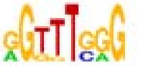
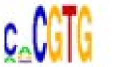
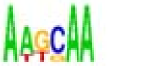
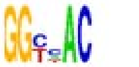
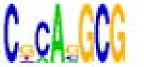
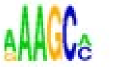
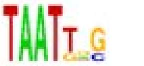
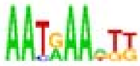
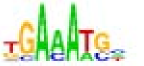
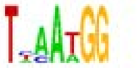
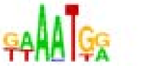
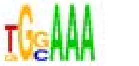
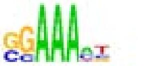
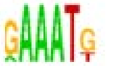
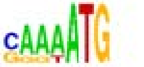
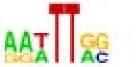
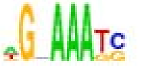
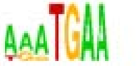
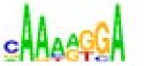
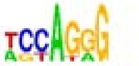
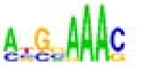
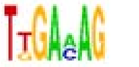
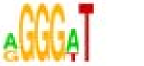
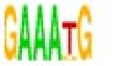
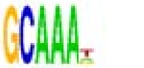
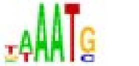
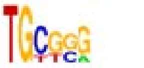
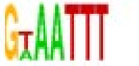
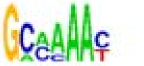
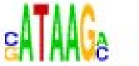
Molecular function	GO term	Number of proteins
Metal ion binding	GO:0046872	127
Receptor binding	GO:0005102	32
Catalytic activity	GO:0003824	138
Enzyme regulator activity	GO:0030234	44
Signal transducer activity	GO:0004871	46
Transporter acticity	GO:0005215	15
Other miscellaneous function protein		107
Molecular functions unclassified		181

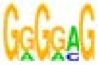
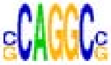
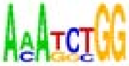
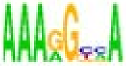
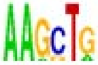
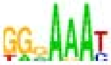
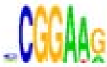
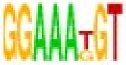
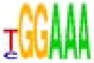
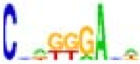
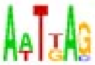
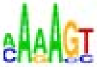
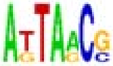
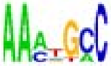
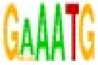
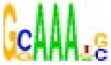
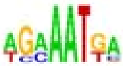
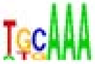
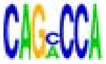
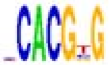
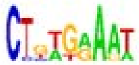
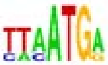
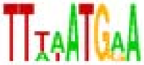
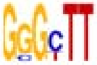
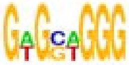
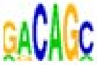
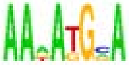
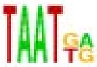
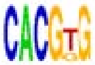
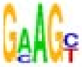
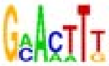
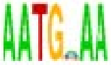
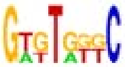
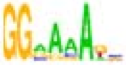
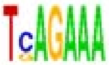
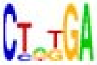
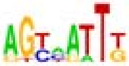
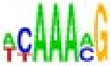
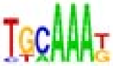
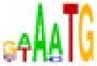
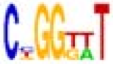
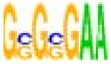
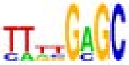
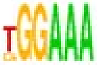
Table S4. Estimation of the true-positive rate. In addition to the 60 known PDIs retrieved from the TRANSFAC SITE database, 11 predicted PDIs were also found to have been experimentally verified previously. In total, 71 PDIs were used for positive control to estimate the true positive rate. A cutoff value of six SD was chosen to keep true-positive rate high while minimizing possible false negatives. The relatively low true-positive rate (42.3%) likely reflects the fact that not all proteins on the array are correctly folded and that many TFs lack necessary cofactors for DNA binding.

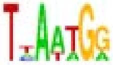
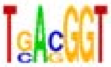
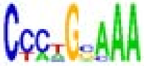
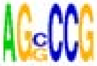
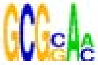
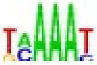
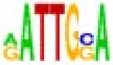
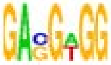
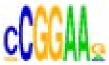
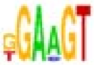
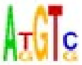
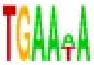
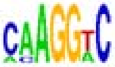
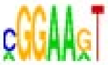
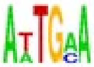
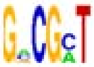
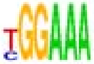
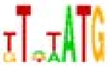
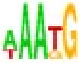
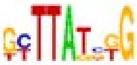
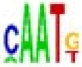
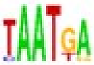
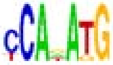
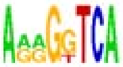
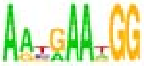
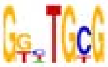
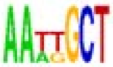
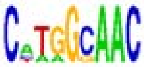
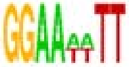
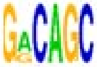
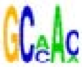
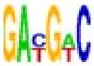
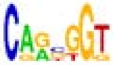
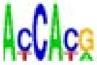
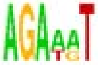
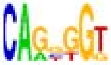
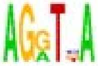
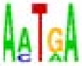
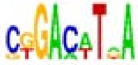
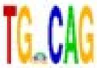
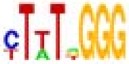
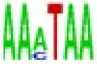
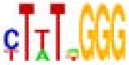
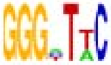
Standard deviation	3	4	5	**6**	7
Number of recovered known PDIs (71)	32	30	30	**30**	28
Recovery rate of known PDIs	0.451	0.423	0.423	**0.423**	0.394

Table S5. Consensus sequences (logos) identified for individual TFs

TF name	No. of binding sequences		Logos
	Protein chip	TRANSFAC SITE	
HOXB9	29	0	
SSX3	29	0	
CREB1	29	56	
RAB18	28	0	
ZNF26	27	0	
PSMC2	27	0	
TRMT1	27	0	
SMAD4	27	10	
TFAP2C	27	6	
TP73	27	4	
HHEX	26	1	
TFAM	26	1	
MYF6	25	0	
YEATS4	25	0	
RFX4	23	0	
MEIS3	23	0	
TFE3	23	3	
RARG	23	2	
MLX	23	1	
KLF3	22	1	
ZBTB4	21	0	
ZNF655	21	0	
CNOT6	21	0	
RFXANK	21	0	
RXRA	21	169	
JARID1D	20	0	
ZNF3	20	0	
LAS1L	20	0	
CPSF4	19	0	
TSNAX	18	0	
FHL2	18	0	
ZBTB25	18	0	
PHOX2A	18	5	
ZHX3	17	0	
VSX1	17	0	
JDP2	17	0	
ZBED1	17	0	
POU3F2	17	19	
GTF3C2	16	0	
RAX	15	0	

SOX14	15	0		ZNF124	11	0	
NME1	15	0		AFF4	11	0	
NR2F1	15	62		GTF2B	11	0	
ZNF238	15	22		ZNF131	11	0	
ENO1	15	3		HCLS1	11	0	
NKX2-3	14	0		HIP2	11	0	
ZNF695	14	0		TEAD1	11	11	
SND1	14	0		USF2	11	4	
SCAND2	14	0		THRA	11	3	
TRIM69	14	0		SOX13	11	1	
PRRX1	14	1		MEF2B	11	1	
OLIG3	13	0		ZNF76	10	0	
TCEAL2	13	0		EVX1	10	0	
IRF6	12	0		POU4F3	10	0	
ZNF205	12	0		PQBP1	10	0	
LARP1	12	0		CCDC16	10	0	
RAN	12	0		CHES1	10	0	
SNAPC5	12	0		PAX3	10	2	
ZNF160	12	0		BCL11A	9	0	
MYEF2	12	0		DLX6	9	0	
TGIF1	12	15		HOXD3	9	0	
ZNF326	11	0		ZNF720	9	0	

PDLIM5	9	0		PIR	6	0	
PURG	9	0		PRKRIR	6	0	
PAXIP1	9	0		TULP1	6	0	
ETV4	9	10		SSBP3	6	0	
NFATC3	9	2		KCNIP1	6	0	
PLAGL1	9	1		C19orf25	6	0	
NCALD	8	0		TAF1A	6	0	
SCMH1	8	0		ZNF250	6	0	
TCF3	8	67		FOXM1	6	14	
SMAD3	8	23		TFEB	6	4	
VAX2	7	0		MYOD1	6	10	
HOXB13	7	0		PITX1	5	0	
ZNF503	7	0		PKNOX2	5	0	
SSX2	7	0		LHX2	5	0	
USF1	7	68		ESX1	5	0	
HSF1	7	19		BARX1	5	0	
INSM1	7	11		FOXP4	5	0	
KLF4	7	3		CEBPG	5	0	
ZFP3	6	0		NMRAL1	5	0	
SNAPC4	6	0		MECP2	5	0	
MXD4	6	0		OTUD4	5	0	
DDX20	6	0		MAGED4	5	0	

MAGEF1	5	0		LASS4	4	0	
ZNF385	5	0		ZNF304	4	0	
HTATIP2	5	0		ZNF207	4	0	
ZNF706	5	0		THRAP6	4	0	
ELF2	5	7		ETS1	4	44	
NR4A1	5	6		IRF1	4	33	
ESRRA	5	5		FLI1	4	3	
NFIL3	5	25		RARA	4	50	
NFATC4	5	1		SMAD2	4	2	
CBFB	5	1		ARNTL	4	2	
HMG20A	4	0		LHX4	4	1	
OLIG1	4	0		ZNF71	3	0	
THAP5	4	0		FEZF2	3	0	
ZBTB46	4	0		RFX3	3	0	
ZBTB12	4	0		TGIF2LX	3	0	
BAD	4	0		ID2	3	0	
PDCD11	4	0		CREB3L1	3	0	
GTF2H3	4	0		JARID1A	3	0	
ZNF510	4	0		ZBTB43	3	0	
ZNF323	4	0		ZNF671	3	0	
TSC22D4	4	0		RUFY3	3	0	
ZNF192	4	0		HCFC2	3	0	

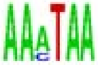
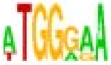
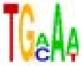
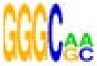
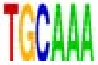
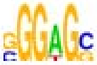
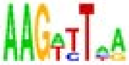
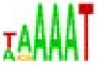
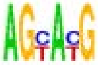
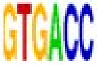
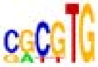
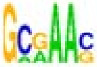
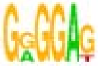
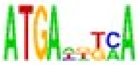
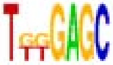
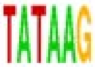
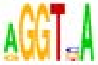
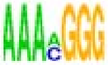
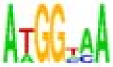
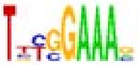
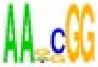
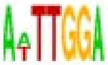
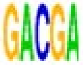
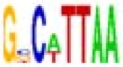
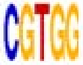
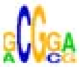
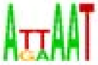
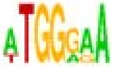
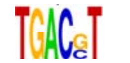
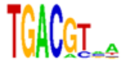
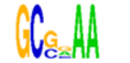
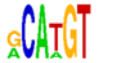
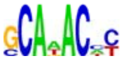
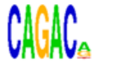
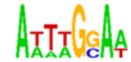
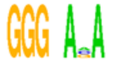
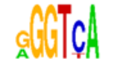
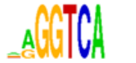
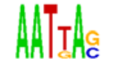
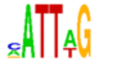
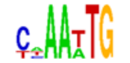
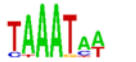
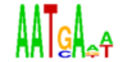
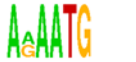
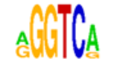
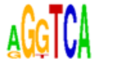
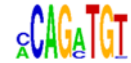
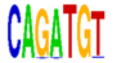
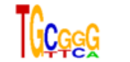
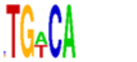
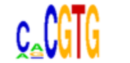
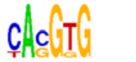
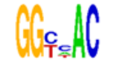
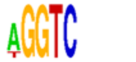
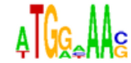
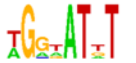
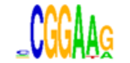
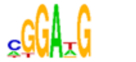
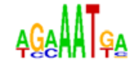
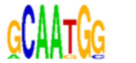
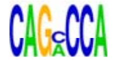
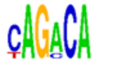
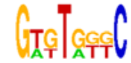
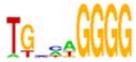
PHTF1	3	0		POLE3	3	0	
ZNF193	3	0		VPS4B	3	0	
NFIX	3	0		ZCCHC14	3	0	
GRHL1	3	0		SF1	3	0	
RBBP5	3	0		GTF3C5	3	0	
HES5	3	0		NFIB	3	0	
ASCC1	3	0		FOSL1	3	5	
CBFA2T2	3	0		RARB	3	9	
ZNF313	3	0		EBF1	3	9	
COBRA1	3	0		TFAP2A	3	196	
ZNF766	3	0		NR4A2	3	2	
TIMELESS	3	0		TBPL1	3	2	
TAF9	3	0		NRL	3	1	
HDAC8	3	0					
NUCB1	3	0					

Table S6. Comparison between TF binding logos identified in this study and those listed in TRANSFAC SITE database.

TF name	No. of binding motifs		DNA binding logo	
	Protein chip	Transfac site	Protein chip	Transfac site
CREB1	29	56		
TP73	27	4		
SMAD4	27	10		
TFAP2C	27	6		
RXRA	21	169		
PHOX2A	18	5		
POU3F2	17	19		
ENO1	15	3		
NR2F1	15	62		
ZNF238	15	22		
TGIF1	12	15		
USF2	11	4		
THRA	11	3		
TEAD1	11	11		
ETV4	9	10		
TCF3	8	67		
SMAD3	8	23		
INSM1	7	11		

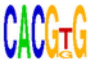
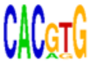
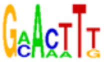
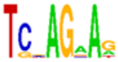
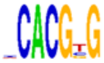
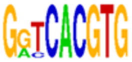
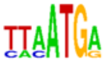
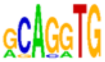
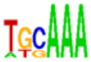
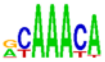
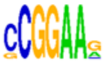
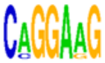
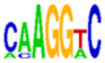
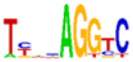
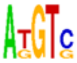
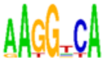
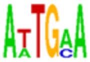
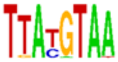
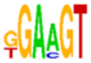
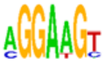
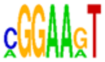
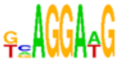
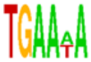
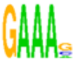
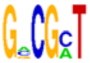
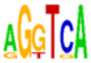
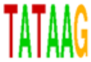
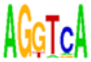
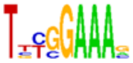
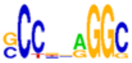
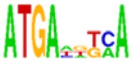
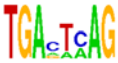
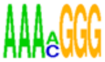
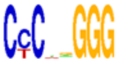
USF1	7	68		
HSF1	7	19		
TFEB	6	4		
MYOD1	6	10		
FOXM1	6	14		
ELF2	5	7		
ESRRA	5	5		
NR4A1	5	6		
NFIL3	5	25		
ETS1	4	44		
FLI1	4	3		
IRF1	4	33		
RARA	4	50		
RARB	3	9		
TFAP2A	3	196		
FOSL1	3	5		
EBF1	3	9		

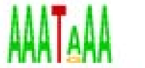
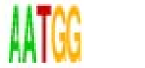
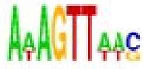
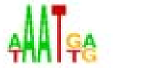
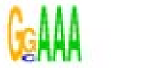
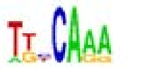
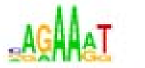
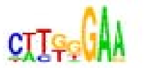
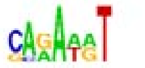
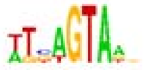
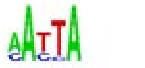
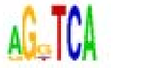
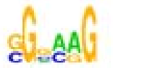
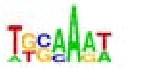
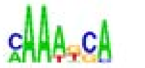
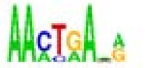
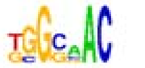
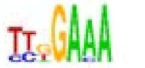
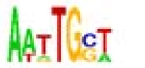
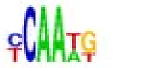
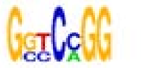
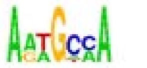
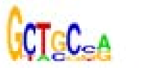
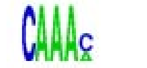
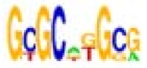
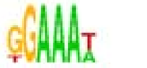
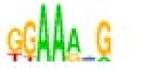
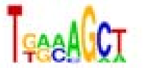
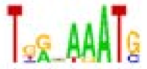
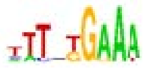
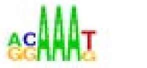
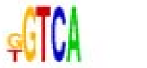
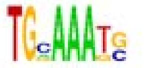
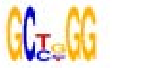
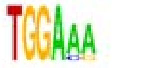
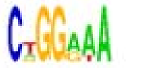
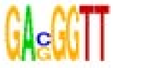
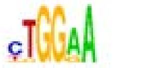
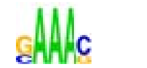
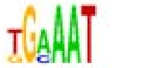
Table S7. Number of motifs shared by different TF subfamilies versus the expected numbers. Yellow background cells denote the number of motifs bound to the TF subfamily in the row. The number before "/" denotes the number of motifs shared. The number after "/" denotes the expected number of motifs shared. Green background cells indicate that shared motifs are over-represented by two subfamilies, where *, ** and *** denote *p*-values <0.01, <0.001, and <0.00001, respectively. *p* values were calculated using the hypergeometric test.

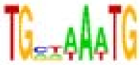
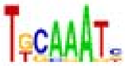
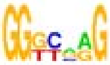
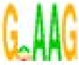
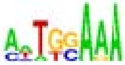
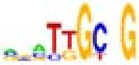
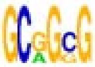
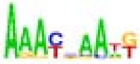
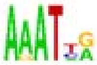
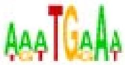
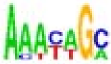
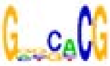
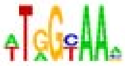
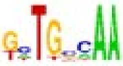
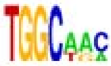
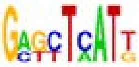
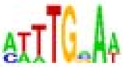
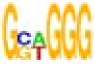
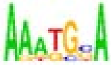
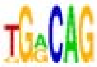
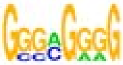
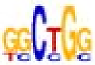
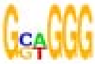
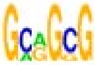
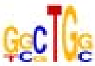
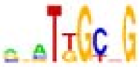
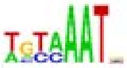
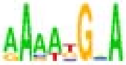
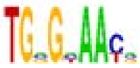
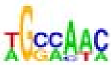
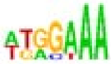
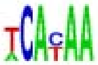
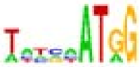
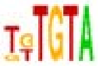
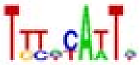
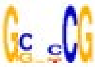
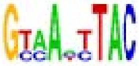
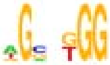
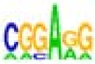
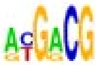
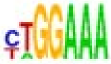
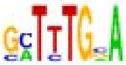
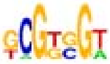
zf-C2H2	206											
Homeodomain	111/69.4 ***	155										
bHLH	53/43	45/32.3 *	96									
NHR	49/36.3 *	38/27.3 *	20/16.9	81								
bZIP	48/31.3 **	44/23.6 ***	20/14.6	20/12.3 *	70							
HMG	42/26 ***	43/19.5 ***	24/12.1 **	16/10.2	15/8.8	58						
MH	23/20.2	25/15.2 *	8/9.4	19/7.9 **	8/6.8	4/5.7	45					
Forkhead	18/10.7 *	12/8.1	8/5	8/4.2	4/3.7	6/3	4/2.3	24				
IRF	13/7.6 *	10/5.7	3/3.5	5/3	6/2.6	1/2.1	5/1.7	2/0.9	17			
Ets	6/5.4	6/4	2/2.5	3/2.1	4/1.8	2/1.5	1/1.2	0/0.6	0/0.4	12		
Myb	8/5.4	5/4	2/2.5	3/2.1	1/1.8	0/1.5	6/1.2 **	3/0.6	0/0.4	0/0.3	12	
RHD	8/4.5	7/3.4	3/2.1	4/1.8	2/1.5	2/1.3	2/1	2/0.5	2/0.4	1/0.3	0/0.3	10
	zf-C2H2	Homeodomain	bHLH	NHR	bZIP	HMG	MH	Forkhead	IRF	Ets	Myb	RHD

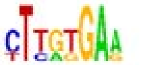
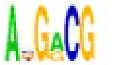
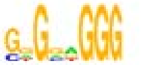
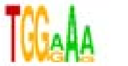
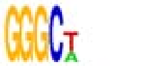
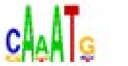
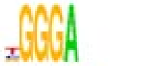
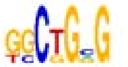
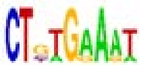
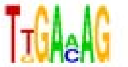
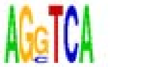
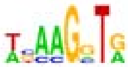
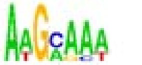
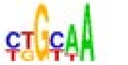
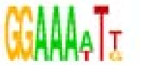
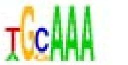
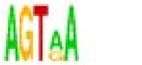
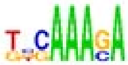
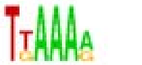
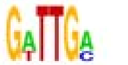
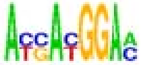
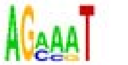
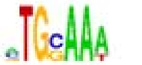
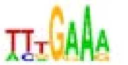
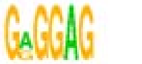
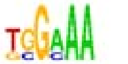
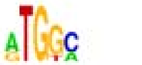
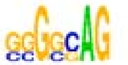
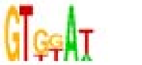
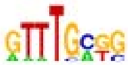
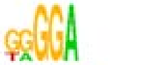
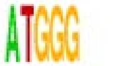
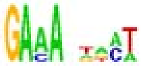
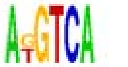
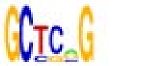
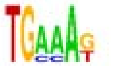
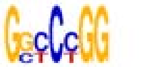
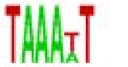
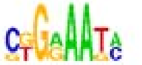
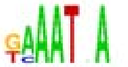
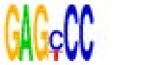
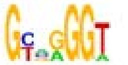
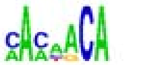
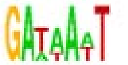
Table S8. EMSA result for 31 novel PDIs. PTF denotes predicted TFs, and RBP denotes RNA-binding proteins.

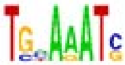
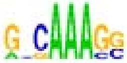
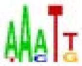
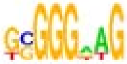
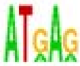
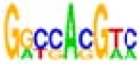
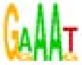
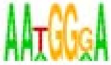
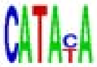
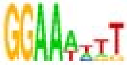
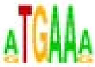
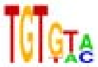
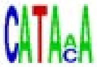
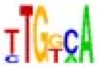
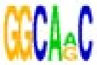
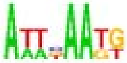
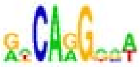
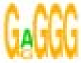
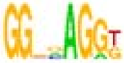
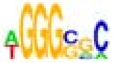
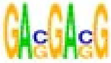
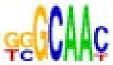
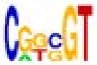
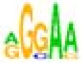
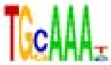
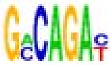
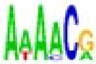
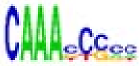
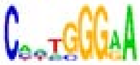
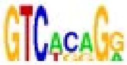
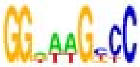
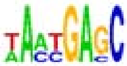
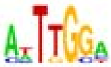
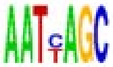
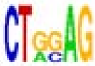
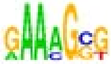
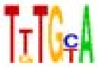
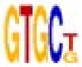
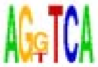
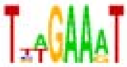
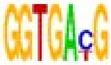
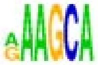
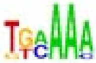
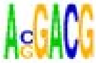
Gene symbol	Protein Class	DNA motif	EMSA results
TGIF2LX	TF	TTTTGACAGCTCAG	+
PKNOX1	TF	TTTTGACAGCTCAG	+
PKNOX2	TF	TTTTGACAGCTCAG	+
MEIS1	TF	TTTTGACAGCTCAG	+
MEIS2	TF	TTTTGACAGCTCAG	+
MEIS3	TF	TTTTGACAGCTCAG	+
SCML4	TF	TTTCCATCATAAATC	+
PAPD1	TF	ACTGAGCATGCTCAG	-
DSCR1	TF	GGAAAACTGAAAGGG	-
NRL	TF	CCCGTGACC	+
SMARCE1	TF	GGGCTTCCCCC	+
TTRAP	TF	CCCCTCCC	+
IRF3	TF	GACATCTGGTTGCAATTTG	+
CEBPG	TF	ATTCATTTTGGCTTTGAAAG	+
CHES1	TF	CTGCAATCT	+
ZNF3	TF	GATTTGCATTTCATTTGCAC	+
SNAPC4	TF	CCCCCACTGAACCCTTGACCCCTGCCC	-
MYF6	TF	TTGAAGCAATTAGC	+
SMAD4	TF	CCTCGGCCGCCCCCTCGCGGC	+
IRF5	TF	CCGGCCG	+
TFAM	TF	TCCCATTGACTTCAATGGGA	+
THRA	TF &RBP	CCCGTGACC	+
ZCRB1	PTF & RBP	TCTGTGTAT	+
RIPX	PTF	TCAAGTAACAGCAGGTGCAAAATAAAGT	+
ZCCHC3	PTF	TTGTGTATGC	+
TERF1	PTF	TTTCGCGC	-
FUBP3	PTF	GATTTCCTGTTGTG	+
ZNF261	PTF	GGGCTTCCCCC	+
ZNF765	PTF	GGGCTTCCCCC	+
C14orf106	PTF	CCCCCACTGAACCCTTGACCCCTGCCC	+
ZNF766	PTF	GATTTGCATTTCATTTGCAC	+

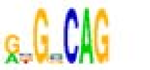
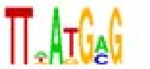
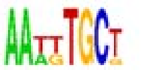
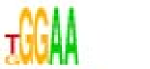
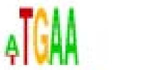
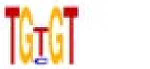
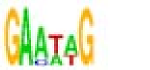
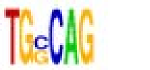
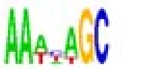
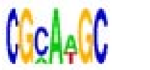
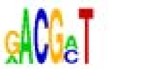
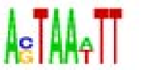
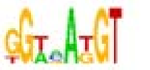
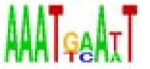
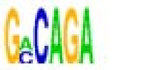
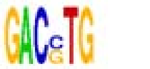
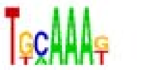
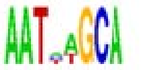
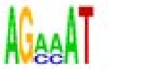
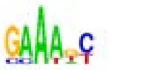
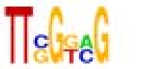
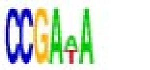
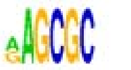
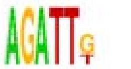
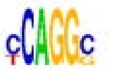
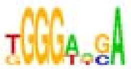
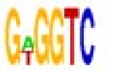
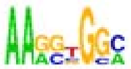
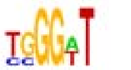
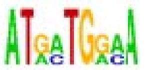
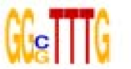
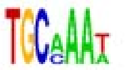
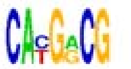
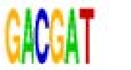
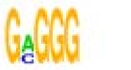
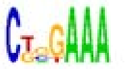
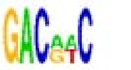
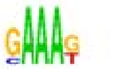
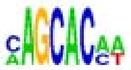
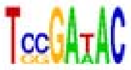
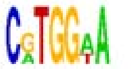
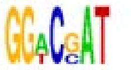
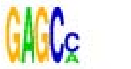
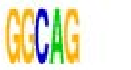
Table S9. Consensus sequences (logos) identified for uDBPs

Protein	No. of binding sequences	logo
CSTF2	29	
CDK2AP1	28	
STAU2	27	
RFC2	27	
DAZAP1	27	
DDX43	26	
CAT	25	
LARP4	25	
HIST2H2AB	24	
LRRFIP1	24	
RPL35	23	
CBX7	23	
TCEAL6	23	
SFT2D1	22	
HNRPC	21	
DTL	21	
FAM127B	21	
USP39	21	
SLC18A1	21	
C19orf40	20	
TAGLN2	20	
ZSWIM1	20	
DIABLO	19	
STUB1	19	
HIST1H2BN	19	
U2AF1	19	
DIS3	19	
RPP25	19	
RBM22	19	
HNRPA1	18	
TROVE2	18	
BRUNOL6	18	
IL24	18	
MTHFD1	18	
MYLK	18	
MAGEA8	18	
LOC653972	18	
HNRPH3	18	
ERK2	17	
ZMAT4	17	

MRPS25	17	
NMI	17	
SCC-112	17	
KIAA0907	16	
TSN	16	
SEMA4A	16	
ODC1	16	
EDN1	15	
CCDC25	15	
RKHD2	15	
MSI2	15	
TIMM8A	14	
TPPP	14	
APEX2	14	
C2orf52	14	
MAGOH	14	
RBM35B	14	
AKR1A1	14	
RFC3	14	
ZCCHC17	14	
PGAM2	14	
SMAP1L	13	
SPR	13	
NANOS1	13	
TRIM21	13	
H2AFY	13	
TRIP10	13	
MGC10433	13	
VAMP3	13	
ANXA1	13	
PSMA6	13	
GTPBP1	13	
ZDHHC15	12	
MSI1	12	
RUVBL1	12	
NNT	12	
DDEFL1	12	
NXPH3	12	
VIL2	12	
UQCRB	12	
HP1BP3	12	
RBM35A	12	
RAB14	11	

RPS4X	11		TMSL3	9	
GPD1	11		AVEN	9	
RBM17	11		RPL6	9	
UBB	11		C9orf156	9	
MRPL1	11		MAP4K2	9	
RPS10	10		FIP1L1	9	
TIA1	10		UTP18	9	
HNRPA0	10		NOC2L	8	
LOC51035	10		MBTPS2	8	
RBBP9	10		ASPSCR1	8	
HNRPLL	10		MORN1	8	
CENTG1	10		FLJ37078	8	
ANXA11	10		PHLDA2	8	
PPP5C	10		GRHPR	8	
BRUNOL5	9		UBE2V1	8	
PTPMT1	9		GPAM	8	
ADARB1	9		MSRB3	8	
RAB7A	9		CLK1	8	
SMPX	9		R3HDM2	8	
MDM2	9		RIOK2	7	
PIK3C3	9		TIMM44	7	
BOLL	9		PKM2	7	

LUZP2	7		DUSP26	6	
ZRSR2	7		LUZP1	6	
KIF22	7		SPAG7	6	
DDX4	7		DAB2	6	
RBM3	7		DHX36	6	
DUSP22	7		RBM8A	5	
CKMT1B	7		PICK1	5	
P4HB	7		MORG1	5	
MRPL2	7		ZDHHC5	5	
AGGF1	7		TOB2	5	
ETFB	7		HIRIP3	5	
PCK2	6		MCTP2	5	
DGCR8	6		SF3B1	5	
ACO1	6		CYCS	5	
H2AFZ	6		EIF5A2	5	
ZC3H7A	6		EWSR1	5	
WHSC2	6		IVD	5	
UGP2	6		TPI1	5	
ACF	6		CANX	5	
NUP133	6		SUCLG1	5	
HSPA5	6		WISP2	5	
GADD45A	6		PRDX5	5	

FGF19	5	
PDE6H	4	
XRCC1	4	
EXOSC3	4	
RNF138	4	
DDX53	4	
ECSIT	4	
HSPA1L	4	
C1orf176	4	
DNMT3A	4	
RAB2A	4	
SNRP70	4	
PTCD1	4	
GLYCTK	4	
PLG	4	
NCBP2	4	
SMCR7L	4	
RBMS1	4	
NOLA1	4	
ABCF2	4	
RNASEH2C	3	
PRNP	3	
POLI	3	
HHAT	3	
NAP1L1	3	
SOCS4	3	
DR-1	3	
SRP9	3	
YWHAZ	3	
XG	3	
NONO	3	
SRBD1	3	
GOT1	3	
MSRA	3	
ZMAT2	3	
H1FX	3	
RPS6KA5	3	
SPATS2	3	
SNRPB2	3	
CYB5R1	3	
SMUG1	3	
YWHAE	3	
SOD1	3	
HLCS	3	

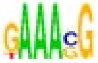
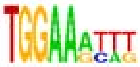
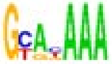
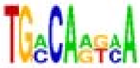
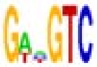
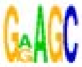
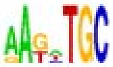
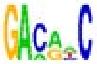
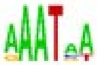
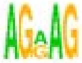
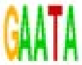
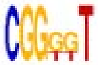
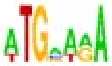
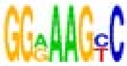
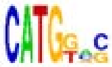
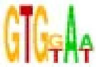
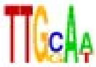
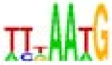
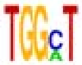
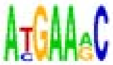
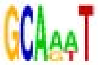
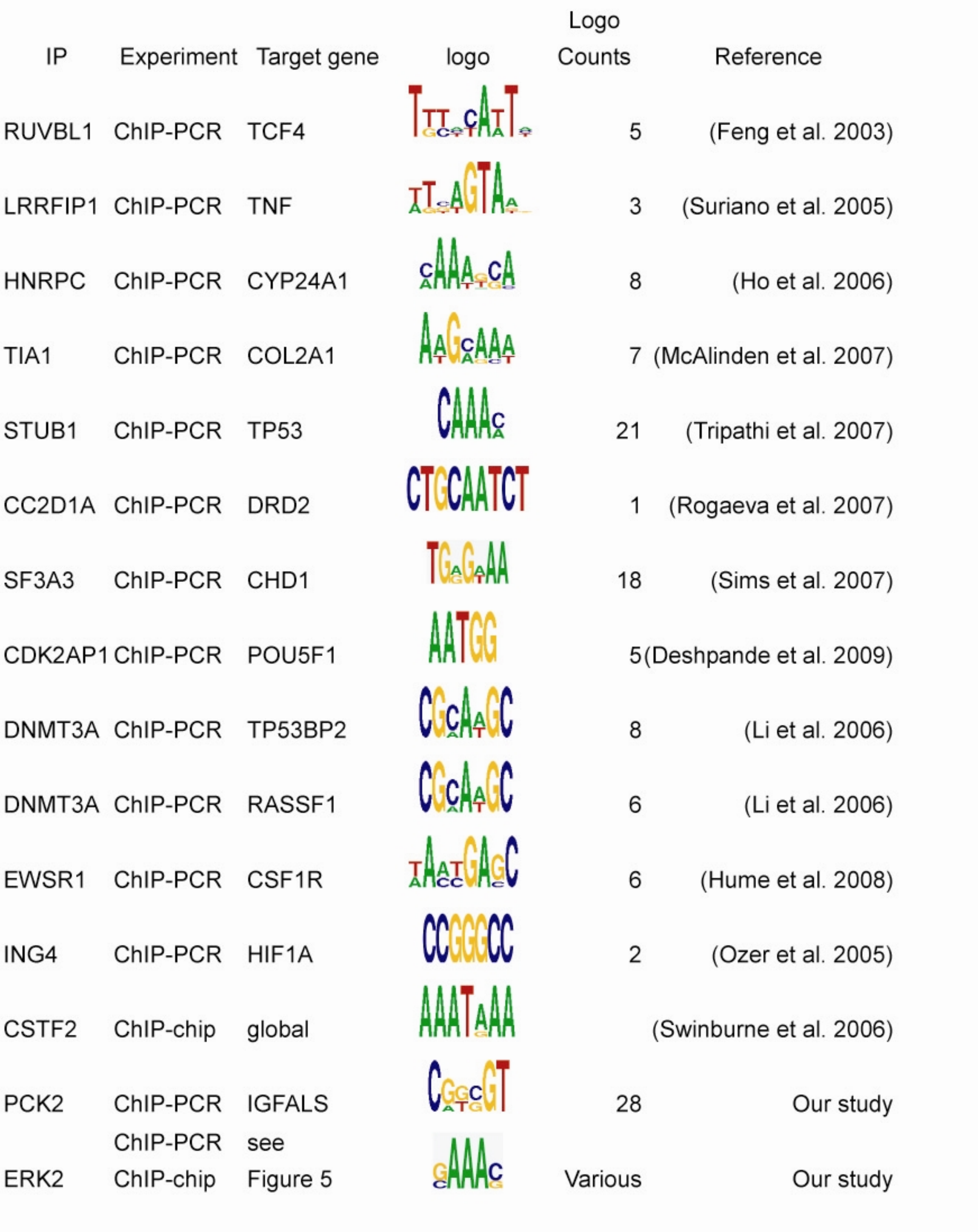
CSNK2B	3	
HIST2H2BE	3	
PPP2R3B	3	
EEF1D	3	
ING3	3	
MGC10334	3	
NUP107	3	
BAX	3	
FAM119B	3	
RBM7	3	
BAT4	3	
CFL2	3	
LSM6	3	
CD59	3	
ARFGAP1	3	
BRUNOL4	3	
GIT2	3	
GTPBP6	3	
DUS3L	3	
PPP1R10	3	
FEZ1	3	

Table S10. EMSA results for 45 uDBPs.

Gene symbol	Protein Class	DNA motif	EMSA results
SMARCA5	Chromatin	CCCCCACTGAACCCTTGACCCCTGCCC	-
JARID1D	Chromatin	CCCCCACTGAACCCTTGACCCCTGCCC	+
DNMT3A	Chromatin	CACATCTGGACAGATGTGGGCG	+
SMARCAL1	Chromatin	CCCCTCCC	+
CSRP2	Coregulator	CCCCTCCC	+
NMI	Coregulator	GCTCTGGAAATTTCCAG	+
MAGEA8	Coregulator	GCTCTGGAAATTTCCAG	+
RCOR1	Coregulator	CCCCCACTGAACCCTTGACCCCTGCCC	+
CD59	DNA Repair	GGGCTTCCCCC	+
WHSC2	DNA Repair	GGGCTTCCCCC	+
SPEG	Kianse&Coregulator	TTGTGTATGC	+
RIPK3	Kinase	GGGCTTCCCCC	+
MAP4K2	Kinase	GATTCATTTAGCAG	+
PIM2	Kinase	AGAGTGCCACCTACTGAAT	+
ERK2	Kinase	AAAGAGAAAG	+
MYLK	Kinase	TTGCTTTGGAAGCAGCT	+
CAMKK2	Kinase	GACGACGAA	+
MKNK2	Kinase	CCCTCCCG	-
MARK2	Kinase	CTTCCGC	-
ICK	Kinase	CTTCCGC	-
MAP3K7	Kinase	CTTCCGC	+
CLK1	Kinase	AATCATGTTTGAAAG	+
LYPLAL1	Mitochondrial	CCCCTCCC	+
MTHFD1	Mitochondrial	CCCTCCTC	+
MTCP1	Mitochondrial	GGGCTTCCCCC	+
HSPE1	Mitochondrial	GGGCTTCCCCC	+
PRDX1	Mitochondrial	TTGTGTATGC	+
MRPL55	Mitochondrial	TTGTGTATGC	+
DUT	Mitochondrial	CTGCCGC	+
PCK2	Mitochondrial	GACGACGAA	+
SOD1	Mitochondrial	GACGACGAA	+
CDK2AP1	Nucleic Acid Binding	TCATTTTGCAAGTGCAA	+
WISP2	Nucleic Acid Binding	GCGTGGAA	+
ANXA1	Other	TTGTGTATGC	+
ADPRTL3	Other	ACTTGCGCC	+
CSTF2	RNA Binding	TTTCCGGAAA	+
RBM12	RNA Binding	GGGCTTCCCCC	+
EIF4B	RNA Binding	GACATCTGGTTGCAATTTG	+
RNPC1	RNA Binding	TCTGTGTAT	+
PSMA1	RNA Binding	TTTCCATCATAAATC	+
KHDRBS3	RNA Binding	GGGCTTCCCCC	+
LARP7	RNA Binding	GGGCTTCCCCC	+
RBM19	RNA Binding	TTGTGTATGC	+
RBM8A	RNA Binding	TCTGTGTAT	+
NCL	RNA Binding	CCCCTCCC	+

Table S11. ChIP experiments of unconventional DNA binding proteins identified by the previous studies and our study. The counts of DNA logos in the promoter regions of target genes were calculated using “countPWM” function in Biostrings package of Bioconductor (Gentleman et al., 2004), where 85% of minimum score was used. For the counts of binding sequences of CC2D1A, CDK2AP1 and ING4, “countPattern” function was used, where exact match was used for CDK2AP1 and ING4 and one miss match was allowed for CC2D1A.



IP	Experiment	Target gene	logo	Logo Counts	Reference
RUVBL1	ChIP-PCR	TCF4		5	(Feng et al. 2003)
LRRFIP1	ChIP-PCR	TNF		3	(Suriano et al. 2005)
HNRPC	ChIP-PCR	CYP24A1		8	(Ho et al. 2006)
TIA1	ChIP-PCR	COL2A1		7	(McAlinden et al. 2007)
STUB1	ChIP-PCR	TP53		21	(Tripathi et al. 2007)
CC2D1A	ChIP-PCR	DRD2		1	(Rogaeva et al. 2007)
SF3A3	ChIP-PCR	CHD1		18	(Sims et al. 2007)
CDK2AP1	ChIP-PCR	POU5F1		5	(Deshpande et al. 2009)
DNMT3A	ChIP-PCR	TP53BP2		8	(Li et al. 2006)
DNMT3A	ChIP-PCR	RASSF1		6	(Li et al. 2006)
EWSR1	ChIP-PCR	CSF1R		6	(Hume et al. 2008)
ING4	ChIP-PCR	HIF1A		2	(Ozer et al. 2005)
CSTF2	ChIP-chip	global			(Swinburne et al. 2006)
PCK2	ChIP-PCR	IGFALS		28	Our study
ERK2	ChIP-PCR ChIP-chip	see Figure 5		Various	Our study

Table S12. Proteins showing identical DNA-binding profiles are grouped in each row.

ARMC6	CAMKK2	CCM2	CHGB	DNAJB2	NCAPH2	XRCC4
C19orf43	CC2D1A	MRLC2	NIPBL			
COQ6	CPSF1	ICK	MAP3K7	MARK2		
C8orf4	EIF2C2					
EIF1AX	EIF5	PANK1	RPL7L1			
CLIC1	FABP3	GINS2	RPA2	TOMM70A	TRFP	
EFTUD2	FKBP1B					
DDX25	GLE1L	POGK				
BANP	HAVCR2	INTS4				
DNMT2	HIST1H2BB	KLHL21	RPL12	SMARCA5		
C17orf79	ING4	OPA3	UBTD2			
EGLN2	JTV1					
HINT2	KIAA1509					
EIF4E2	LDB2	LSM4	MAGEC2	PCNA	RSRC2	
EIF4E	LHFP					
GPC5	LOXL1					
HUS1	MAGEB2					
DSE	MAGEB3	PAGE4	PPP2R5D	RTCD1		
DIS3L	NOL7	POM121	UTP11L			
CPEB4	PCQAP					
LRCH3	PLA2G1B					
DHX40	PRDM7					
CD80	PTGER3					
PSD	RNF10					
FMR1	RPP14	XRCC2				
FARS2	RPS14					
INTS7	TBC1D2					
ProSAPiP1	UBE2C					
FAS	UBE2I					
KIAA1429	UBE2V2					

Table S13. Human TF-DNA binding domain families listed in Pfam database

Zf-C2H2	MH	RFX	P53
Homeobox	E2F	AP-2	zf-C2HC
bZIP	STAT	bZIP-Maf	CBF-B/NFY-A
HLH	SRF	Head-Shock	zf-C4
Forkhead	Paired-box	Runt	GCM
HMG_box	T-box	TEA	HMG-I/HMG-Y
Ets	zf-GATA	ARID/BRIGHT	MBD
Hormone_recep	YL1	bZIP/zf-C2H2	PROX1
Myb	TIG	CBF-D/NFY-B	
IRF	CUT/Homeobox	HNF	
RHD	zf-CCHC	zf-NF-X1	

Supplemental References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., *et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. Nat Protoc *2*, 2366-2382.

Deshpande, A. M., Dai, Y. S., Kim, Y., Kim, J., Kimlin, L., Gao, K., and Wong, D. T. (2009). Cdk2ap1 is required for epigenetic silencing of Oct4 during murine embryonic stem cell differentiation. J Biol Chem *284*, 6043-6047.

Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. Mol Cell *28*, 337-350.

Elemento, O., and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol *6*, R18.

Feng, Y., Lee, N., and Fearon, E. R. (2003). TIP49 regulates beta-catenin-mediated neoplastic transformation and T-cell factor target gene induction via effects on chromatin remodeling. Cancer Res *63*, 8726-8734.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., *et al.* (2006). Pfam: clans, web tools and services. Nucleic Acids Res *34*, D247-251.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Ho, S. W., Jona, G., Chen, C. T., Johnston, M., and Snyder, M. (2006). Linking DNA-binding proteins to their recognition sequences by using protein microarrays. Proc Natl Acad Sci U S A *103*, 9940-9945.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol *296*, 1205-1214.

Hume, D. A., Sasmono, T., Himes, S. R., Sharma, S. M., Bronisz, A., Constantin, M., Ostrowski, M. C., and Ross, I. L. (2008). The Ewing sarcoma protein (EWS) binds directly to the proximal elements of the macrophage-specific promoter of the CSF-1 receptor (csf1r) gene. J Immunol *180*, 6733-6742.

Li, H., Rauch, T., Chen, Z. X., Szabo, P. E., Riggs, A. D., and Pfeifer, G. P. (2006). The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. J Biol Chem *281*, 19489-19500.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol *20*, 835-839.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912-1934.

McAlinden, A., Liang, L., Mukudai, Y., Imamura, T., and Sandell, L. J. (2007). Nuclear protein TIA-1 regulates COL2A1 alternative splicing and interacts with precursor mRNA and genomic DNA. J Biol Chem *282*, 24444-24454.

Ozer, A., Wu, L. C., and Bruick, R. K. (2005). The candidate tumor suppressor ING4 represses activation of the hypoxia inducible factor (HIF). Proc Natl Acad Sci U S A *102*, 7481-7486.

Rogaeva, A., Ou, X. M., Jafar-Nejad, H., Lemonde, S., and Albert, P. R. (2007). Differential repression by freud-1/CC2D1A at a polymorphic site in the dopamine-D2 receptor gene. J Biol Chem *282*, 20897-20905.

Roth, F. P., Hughes, J. D., Estep, P. W., and Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol *16*, 939-945.

Sims, R. J., 3rd, Millhouse, S., Chen, C. F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., Manley, J. L., and Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell *28*, 665-676.

Suriano, A. R., Sanford, A. N., Kim, N., Oh, M., Kennedy, S., Henderson, M. J., Dietzmann, K., and Sullivan, K. E. (2005). GCF2/LRRFIP1 represses tumor necrosis factor alpha expression. Mol Cell Biol *25*, 9073-9081.

Swinburne, I. A., Meyer, C. A., Liu, X. S., Silver, P. A., and Brodsky, A. S. (2006). Genomic localization of RNA binding proteins reveals links between pre-mRNA processing and transcription. Genome Res *16*, 912-921.

Tripathi, V., Ali, A., Bhat, R., and Pati, U. (2007). CHIP chaperones wild type p53 tumor suppressor protein. J Biol Chem *282*, 28441-28454.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res *24*, 238-241.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E. S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A *104*, 7145-7150.

Xuan, Z., Zhao, F., Wang, J., Chen, G., and Zhang, M. (2005). Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biology *6*, R72.

Yu, X., Lin, J., Zack, D. J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res *34*, 4925-4936.