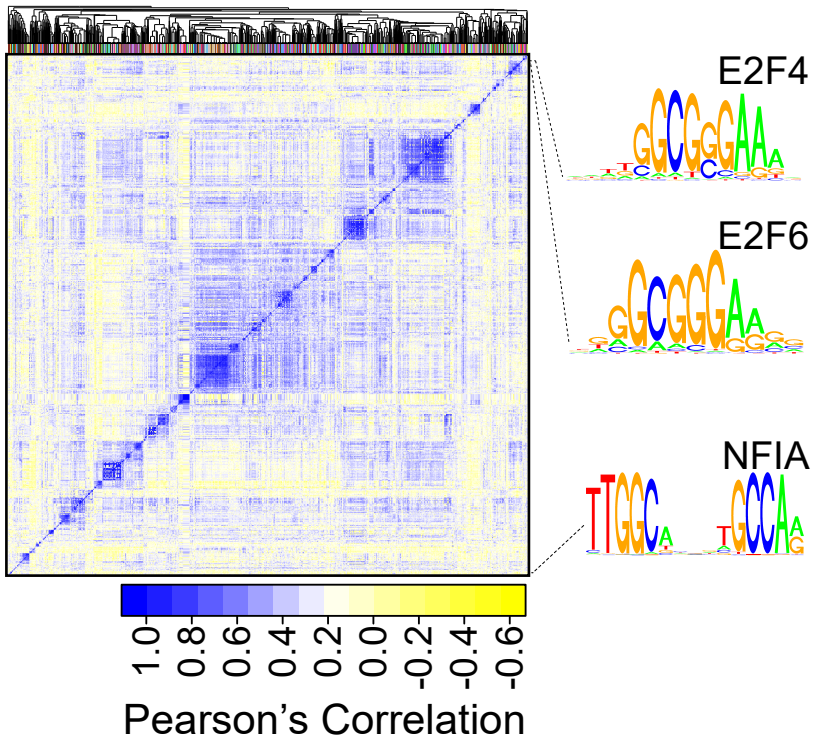
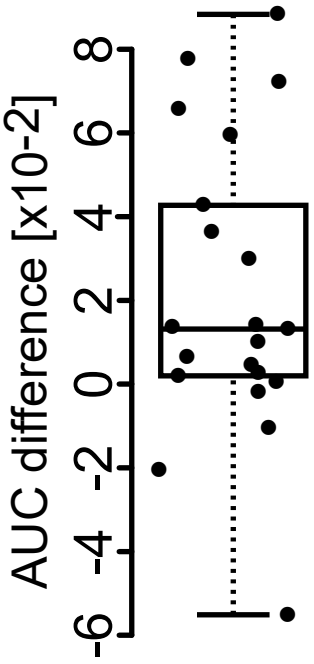
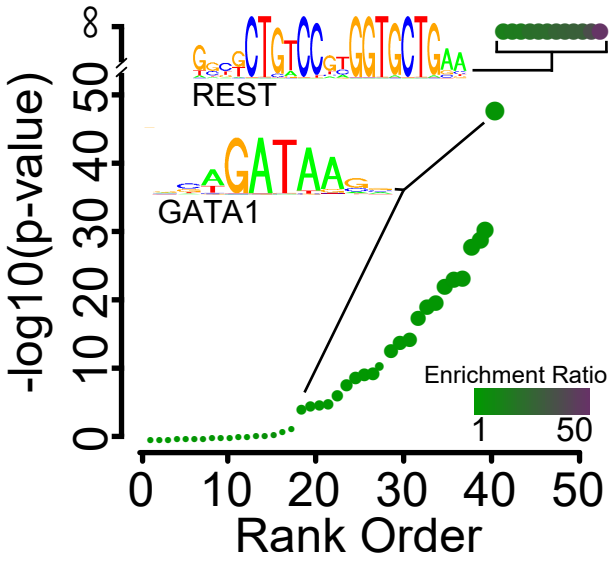


Supplementary Fig. 1: Correlation between motifs with similar DNA binding specificities. A correlation matrix grouping 1,964 motifs recognized by human TFs into 625 clusters using an affinity propagation clustering algorithm. The color scale represents Perason's correlation between the DNA sequence specificity represented by each motif.



Supplementary Fig. 2: Accuracy of motif classification using 1 or 4 GC content groups. DNase-1 peaks in K562 cells were classified as either 'bound' or 'unbound' to 21 TFs using the motif match log-odds score. We computed the area under the receiver operating characteristic curve (AUC) using ChIP-seq data in K562 cells as a gold-standard set, either with or without dividing sequences into 4 separate GC content groups before scoring. The difference between AUCs when dividing sequences into either 1 or 4 separate GC content groups is indicated on the Y-axis. Individual points are shown, and the box-and-whiskers plot denotes the median, 25th and 75th percentile, and maximum values.

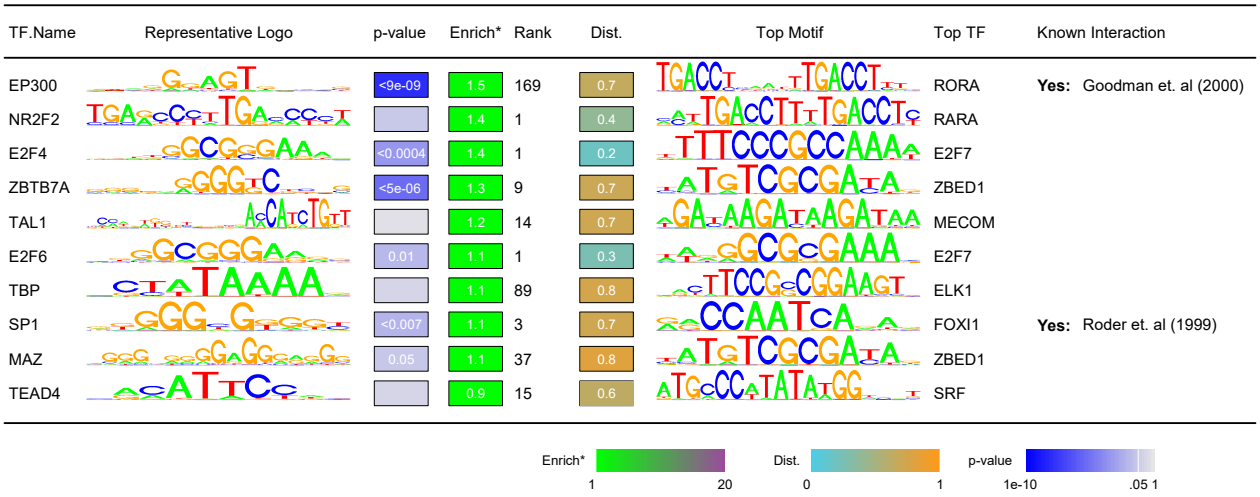


Supplementary Fig. 3: Enrichment of motifs in REST ChIP-seq peaks. The -log-10 p-value (Y-axis) as a function of the p-value rank order (X-axis) illustrates motifs enriched in ChIP-seq peaks binding the transcriptional repressor REST. The magnitude of enrichment is shown by the color scale and by the size of each point.

TF	AUC (GC1)	AUC (GC4)	accurancy	Motif Length	GC content
ZBTB7A	0.775	0.687	12.9	14.67±2.52	0.70±0.06
E2F6	0.777	0.7	11.1	11.20±0.45	0.70±0.01
MEF2A	0.81	0.738	9.8	15.64±4.78	0.28±0.04
CTCFL	0.748	0.689	8.6	15.00±0.00	0.80±0.00
GATA2	0.866	0.8	8.2	11.90±3.54	0.39±0.07
SP2	0.846	0.803	5.3	15.67±1.15	0.70±0.06
EGR1	0.812	0.775	4.7	15.00±3.34	0.75±0.06
GABPA	0.753	0.723	4.1	12.43±1.81	0.63±0.03
ELF1	0.771	0.757	1.9	12.12±1.81	0.59±0.05
SP1	0.865	0.852	1.6	12.50±3.33	0.77±0.05
FOSL1	0.863	0.85	1.5	12.00±1.73	0.48±0.01
ZBTB33	0.824	0.814	1.2	13.50±2.12	0.58±0.04
ATF3	0.663	0.657	1	10.00±1.15	0.57±0.08
USF1	0.838	0.833	0.5	12.45±2.50	0.61±0.05
MAX	0.69	0.687	0.4	11.75±2.14	0.61±0.05
REST	0.643	0.641	0.3	17.17±3.07	0.61±0.02
SPI1	0.723	0.722	0.1	16.50±2.17	0.43±0.03
ETS1	0.765	0.767	-0.3	14.44±5.03	0.55±0.05
YY1	0.716	0.726	-1.4	14.25±2.63	0.57±0.05
TBP	0.689	0.744	-7.4	10.80±2.59	0.24±0.10

Supplementary Table 1: Accuracy of motif classification using 1 or 4 GC content groups. Related to the box and whiskers plot (Supplementary Fig. 2). The table shows the TF name, the area under the receiver operating characteristic curve (AUC) for 1 and 4 GC content groups, the differences in accuracy between groups (percent difference), the motif length, and the GC content. Errors associated with the motif length and GC content group show standard errors between different motifs associated with each TF.

TF.Name	Representative Logo	p-value	Enrich*	Rank	Dist.	Top Motif	Top TF	Known Interaction
SRF		<6e-49	60.4	1	0		SRF	
ZNF274		<5e-37	48.7	1	0		ZNF274	
REST		<=0	48.4	1	0		REST	
CEBPB		<=0	41.7	1	0		CEBPG	
SPI1		<2e-66	25.1	1	0		SPI1	
MEF2A		<3e-06	17.6	1	0		MEF2A	
CTCF		<1e-128	16.6	1	0		CTCF	
MAFF		<1e-52	14.8	1	0		MAFF	
NRF1		<6e-130	14.1	1	0		NRF1	
USF1		<=0	13.5	1	0.1		ARNTL	
YY1		<6e-267	11.5	1	0		YY1	
EGR1		<2e-26	11.3	1	0		EGR1	
NFYB		<7e-272	11.1	1	0.1		NFYA	
NFYA		<2e-40	10.9	1	0		NFYA	
ELK1		<1e-06	10.2	1	0		ELK1	
GATA1		<=0	10.1	7	0.8		SOX18	**Yes:** Murakami et. al. (2004). Kuwahara et. al. (2012)
ZBTB33		<5e-160	10	1	0		BRCA1	
FOSL1		<=0	9.8	1	0.1		NFE2	
FOS		<=0	9.6	1	0.1		NFE2	
MAFK		<1e-157	9.5	1	0.1		MAFF	
USF2		<=0	9	1	0.2		ARNTL	
BACH1		<1e-54	8.7	1	0		BACH1	
ETS1		<6e-27	7.8	1	0.1		SMARCC2	
ELF1		<9e-51	7.7	1	0		ELF3	
BHLHE40		<2e-88	7.6	4	0		BHLHE41	
JUNB		<7e-147	5.4	1	0.1		NFE2	
JUN		<4e-131	5.4	1	0.1		ATF3	
NFE2		<2e-123	5	1	0.2		BACH1	
MAX		<3e-21	4.7	4	0.2		HES7	
SIX5		<9e-49	4.5	1	0.4		ETS1	
JUND		<6e-30	4.2	1	0.1		BATF3	
GATA2		<2e-155	3.9	11	0.5		MECOM	
GABPA		<4e-66	3.6	1	0.3		ELK1	
THAP1		<5e-08	3.2	1	0		THAP1	
ATF1		<5e-24	2.7	5	0.2		BATF3	
RFX5		<1e-185	2.7	9	0.6		BATF3	**Yes:** Lochamy et. al. (2007)
NR2C2		<1e-13	2.7	1	0.3		NR1H3	
MXI1		<5e-20	2.4	1	0.2		RFX5	**Possible:** Neph et. al. (2012)
CTCFL		<4e-68	2.2	1	0.2		CTCF	
ZNF143		<2e-07	2.1	1	0.1		SMARCC2	
MYC		<1e-43	2.1	25	0.7		E2F1	**Possible:** Wanzel et. al. (2003)
ATF3		<2e-33	2.1	1	0.2		ARNTL	
STAT5A		<3e-08	1.9	3	0.8		SRF	**Possible:** Engblom et. al. (2007)
ARID3A		<6e-06	1.9	32	0.7		POU3F2	**Yes:** Rhee et. al. (2014)
ZNF263		<1e-40	1.7	6	0.9		ZNF423	
SP2		<3e-47	1.7	20	0.6		NFYA	**Yes:** Roder et. al (1999)



Supplementary Table 2: Motifs discovered in 56 ENCODE ChIP-seq datasets. The table shows the ChIP-seq TF name (TF.Name), expected motif logo (Representative Logo), one-sided p-value of enrichment for the expected motif (p-value), enrichment of the expected motif (Enrich), rank of the cluster containing the expected motif (Rank), distance between the expected and most enriched motif (Dist.), most enriched motif logo (Top motif), and TF name corresponding to the most enriched motif (Top TF). The final column denotes evidence for known interactions between the TF analyzed by ChIP-seq and the most enriched motif. **Yes** denotes known interactions between this pair of TFs, or close family members, and provides the reference. **Possible** denotes evidence that the ChIP TF is involved in tethering interactions with another TF (though not necessarily the motif discovered).

Supplementary References:

Engblom,D. et al. (2007) Direct glucocorticoid receptor–Stat5 interaction in hepatocytes controls body size and maturation-related gene expression. Genes Dev., 21, 1157–1162.

Goodman,R.H. and Smolik,S. (2000) CBP/p300 in cell growth, transformation, and development. Genes Dev., 14, 1553–1577.

Kuwahara,M. et al. (2012) The transcription factor Sox4 is a downstream target of signaling by the cytokine TGF-[beta] and suppresses TH2 differentiation. Nat. Immunol., 13, 778–786.

Lochamy,J. et al. (2007) CREB and phospho-CREB interact with RFX5 and CIITA to regulate MHC class II genes. Mol. Immunol., 44, 837–847.

Murakami,A. et al. (2004) SOX7 and GATA-4 are competitive activators of Fgf-3 transcription. J. Biol. Chem., 279, 28564–28573.

Neph,S. et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489, 83–90.

Rhee,C. et al. (2014) Arid3a is essential to execution of the first cell fate decision via direct embryonic and extraembryonic transcriptional regulation. Genes Dev., 28, 2219–2232.

Roder,K. et al. (1999) Interaction between the two ubiquitously expressed transcription factors NF-Y and Sp1. Gene, 234, 61–69.

Wanzel,M. et al. (2003) Transcriptional repression by Myc. Trends Cell Biol., 13, 146–150.