Enhancer cell line	Factor	Conser-vation	Original	Scramble	Removal	Max 1-bp decrease	Least 1-bp change	Max 1-bp increase	Random 1-bp change	Total
HepG2	HNF1	high	154^a	160	15	15	15	15	30	400
		ignored	158^a	160	15	15	15	15	30	406
	HNF4	high	160	160	15	15	15	15	30	408
		ignored	160	160	15	15	15	15	30	409
	FOXA	high	160	160	15	15	15	15	30	407
		ignored	160	160	14^b	15	15	15	30	406
	GATA	high	18	18						36
		ignored	18	18						36
	NFE2L2	high	18	18						36
		ignored	18	18						36
	ZFP161	high	10^d	10^d						20
		ignored	18	18						36
	GFI1	high	160	160	15	15	15	15	30	408
		ignored	160	160	15	15	15	15	30	410
K562	HNF1	high	18	18						36
		ignored	18	18						36
	HNF4	high	18	18						36
		ignored	18	18						36
	FOXA	high	18	18						36
		ignored	17^a	17^a						34
	GATA	high	160	160	15	15	15	15	30	408
		ignored	160	160	15	15	15	15	30	408
	NFE2L2	high	159^b	159^b	14^c	14^c	14^c	14^c	28^c	400
		ignored	160	160	12^c	12^c	12^c	12^c	24^c	392
	ZFP161	high	51^d	51^d	15	15	15	15	30	191
		ignored	105^d	105^d	15	15	15	15	30	299
	GFI1	high	18	18						36
		ignored	18	18						36
Total			2104	2112	203	204	204	204	412	5418

Table S1: Precise number of tested sequences. Number fewer than indicated in Table 1 due to ^{a}identical sequences being found at different locations or due to matches on opposite strands, ^{b}the creation of a restriction site at boundary of tested region, ^{c}the motif instance being the best possible match to the desired motif, and consequently excluded from all non-scramble manipulations, or ^{d}having too few matches in the desired enhancer type. Totals indicate the number of distinct tested sequences and thus differ from sum of the corresponding rows or columns due to reuse of sequences across modifications (e.g. a random mutation matching one of the directed 1-bp modifications) or reuse of a position (e.g. a conserved instance tested even when conservation was not taken into account).

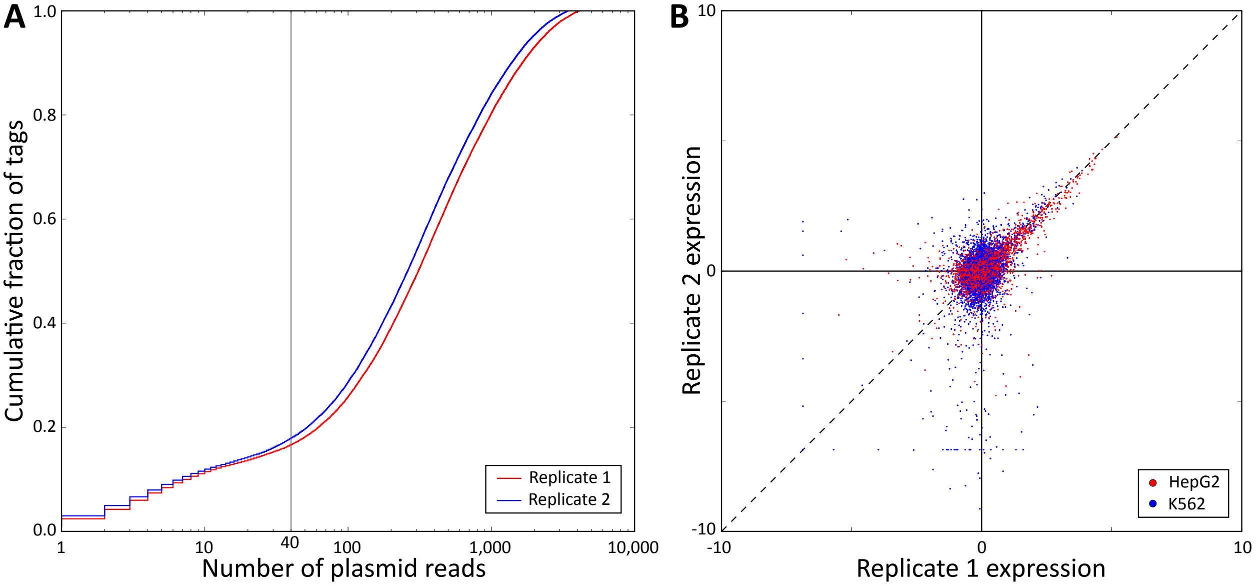

Figure S1: Data quality measures of data.

(A) Cumulative distribution of the plasmid read counts for the two replicates. 17.6% and 16.5% of tags for replicates 1 and 2, respectively did not pass the criteria of having at least 40 reads. However, because each tested sequence was tested with 10 unique tags per replicate, only 46 of 5,418 (0.85%) tested sequences had no passing tags in both replicates.

(B) Comparison of the two replicates using the same averaging procedure as was employed when combining all data (see Methods). Pearson correlation is 0.69 and 0.36 for HepG2 and K562 data, respectively. These correlations are within the 95% confidence intervals when comparing tags 1-5 to 6-10 across both replicates (0.63-0.69, and 0.27-0.36, respectively), indicating we cannot attribute a significant portion of the difference to biological variance. 60 of the 5,418 tested sequences were excluded due to not having any plasmids with sufficient (40) plasmid reads in at least one replicate.

Correlation level is largely influenced by the vast majority of the tested sequences being inactive, leading to expression values that are dominated by noise. The testing of fewer activator motifs along with the higher difficulty in transfection (leading to a raise in the threshold for activity detection) explain to the lower correlation seen for K562. Active sequences, however, are readily detectable (Table S2).

Sequences selected with conserved instances:

Factor	Original	Scramble	Removal	Max 1-bp decrease	Least 1-bp change	Max 1-bp increase	Random 1-bp change
HNF1	80/154 (52%)	21/160 (13%)	2/15 (13%)	2/15 (13%)	6/15 (40%)	11/15 (73%)	8/30 (27%)
HNF4	80/160 (50%)	15/160 (9%)	3/15 (20%)	4/15 (27%)	13/15 (87%)	13/15 (87%)	20/30 (67%)
FOXA	43/160 (27%)	17/160 (11%)	1/15 (7%)	0/15 (0%)	2/15 (13%)	2/15 (13%)	3/30 (10%)
GATA	70/160 (44%)	11/160 (7%)	1/15 (7%)	4/15 (27%)	6/15 (40%)	5/15 (33%)	6/30 (20%)
NFE2L2	55/159 (35%)	10/159 (6%)	1/14 (7%)	0/14 (0%)	3/14 (21%)	6/14 (43%)	4/28 (14%)
Activators	328/793 (41%)	74/799 (9%)	8/74 (11%)	10/74 (14%)	30/74 (41%)	37/74 (50%)	41/148 (28%)
ZFP161	4/51 (8%)	3/51 (6%)	3/15 (20%)	4/15 (27%)	1/15 (7%)	0/15 (0%)	4/30 (13%)
GFI1	10/160 (6%)	10/160 (6%)	0/15 (0%)	1/15 (7%)	1/15 (7%)	1/15 (7%)	2/30 (7%)
Repressors	14/211 (7%)	13/211 (6%)	3/30 (10%)	5/30 (17%)	2/30 (7%)	1/30 (3%)	6/60 (10%)

Sequences selected ignoring instance conservation:

Factor	Original	Scramble	Removal	Max 1-bp decrease	Least 1-bp change	Max 1-bp increase	Random 1-bp change
HNF1	42/158 (27%)	16/160 (10%)	2/15 (13%)	1/15 (7%)	2/15 (13%)	4/15 (27%)	5/30 (17%)
HNF4	46/160 (29%)	19/160 (12%)	1/15 (7%)	2/15 (13%)	3/15 (20%)	5/15 (33%)	3/30 (10%)
FOXA	30/160 (19%)	12/160 (8%)	2/14 (14%)	2/15 (13%)	3/15 (20%)	2/15 (13%)	7/30 (23%)
GATA	39/160 (24%)	7/160 (4%)	1/15 (7%)	2/15 (13%)	3/15 (20%)	8/15 (53%)	5/30 (17%)
NFE2L2	42/160 (26%)	8/160 (5%)	2/12 (17%)	1/12 (8%)	0/12 (0%)	5/12 (42%)	1/24 (4%)
Activators	199/798 (25%)	62/800 (8%)	8/71 (11%)	8/72 (11%)	11/72 (15%)	24/72 (33%)	21/144 (15%)
ZFP161	13/105 (12%)	6/105 (6%)	1/15 (7%)	3/15 (20%)	0/15 (0%)	1/15 (7%)	4/30 (13%)
GFI1	18/160 (11%)	15/160 (9%)	3/15 (20%)	3/15 (20%)	2/15 (13%)	2/15 (13%)	7/30 (23%)
Repressors	31/265 (12%)	21/265 (8%)	4/30 (13%)	6/30 (20%)	2/30 (7%)	3/30 (10%)	11/60 (18%)

Table S2: Number of tested sequences that can individually be identified as being expressed ($P \leq 0.05$; cell types matched as in Figure 3B). One-tailed Mann-Whitney p-values are computed by comparing each of the up to 20 replicate values to those from all sequences whose motifs are scrambled. P-values are available in Table S3.

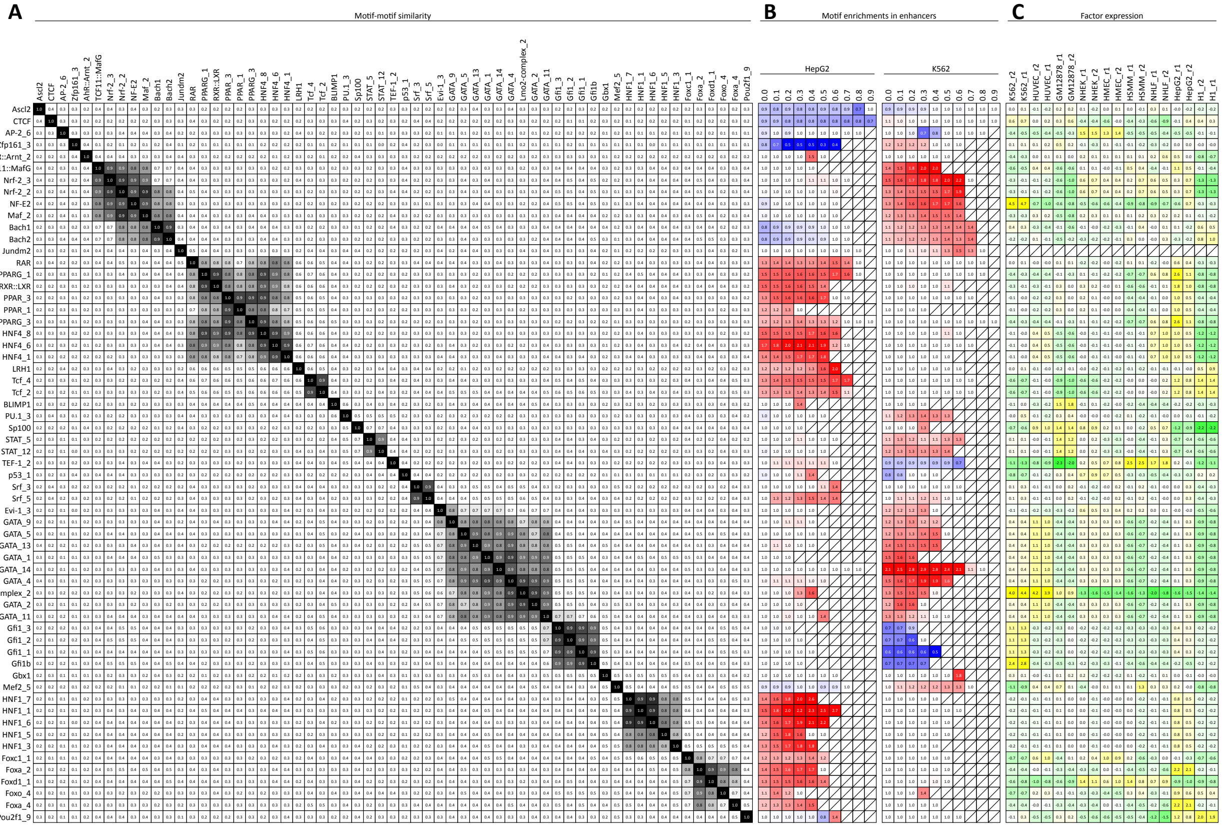

Figure S2: Analysis of all motifs that are enriched in enhancers for HepG2 or K562.

(A) The motif-motif correlation (at all shifts including reverse complement). Different variants of literature motifs are numbered.

(B) The enrichment (red) or depletion (blue) of the motif at different conservation cutoffs in enhancers (see Methods) for the indicated cell line.

(C) The relative expression (log2) of the factor corresponding to the motif in the cell lines (Ernst and Kellis, 2010). Where several proteins exist for a given motif we average the expression values. We ultimately choose HNF1_1 (MA0046.1; Sandelin et al., 2004), Foxa_2 (MA0047.2; Sandelin et al., 2004), HNF4_6 (M00158; Matys et al., 2003), Zfp161_3 (ZFP161; Badis et al., 2009), Nrf-2_3 (MA0150.1; Sandelin et al., 2004), GATA_14 (MA0140.1; Sandelin et al., 2004), and Gfi1_1 (M00250; Matys et al., 2003) on the basis of their enrichment/depletion, expression and sequence-level uniqueness (shown in detail in Figure 1A). We note that because our analysis is unable to distinguish between factors which share a motif, we may not capture the expression of all potential alternative factors.

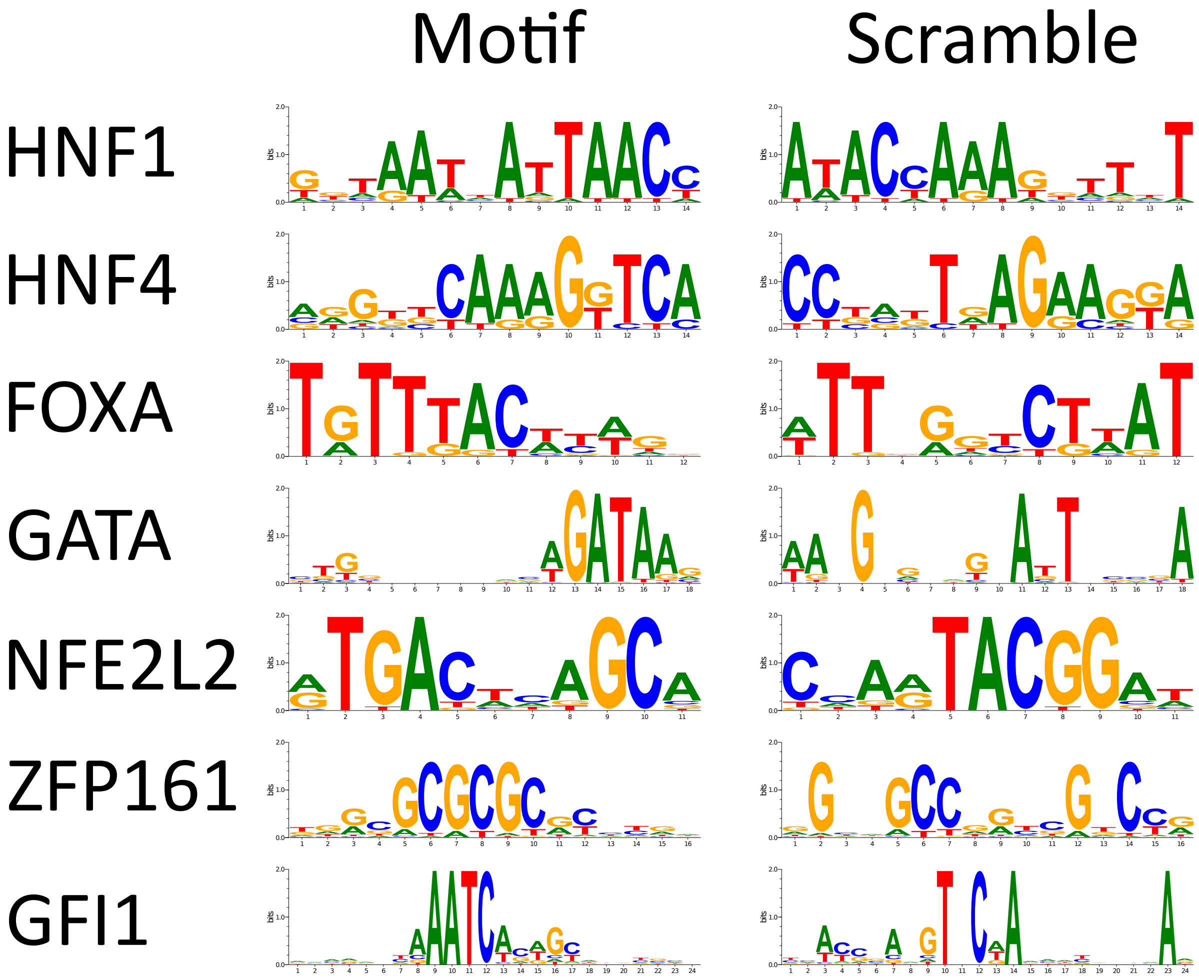

Figure S3: Logos of the specific motifs selected for analysis and the permutation order used for scrambling applied to the matrix (see methods). The corresponding PWMs and score cutoffs are included in Data S2. Motifs are taken as-is from the corresponding databases and consequently may include flanking bases.

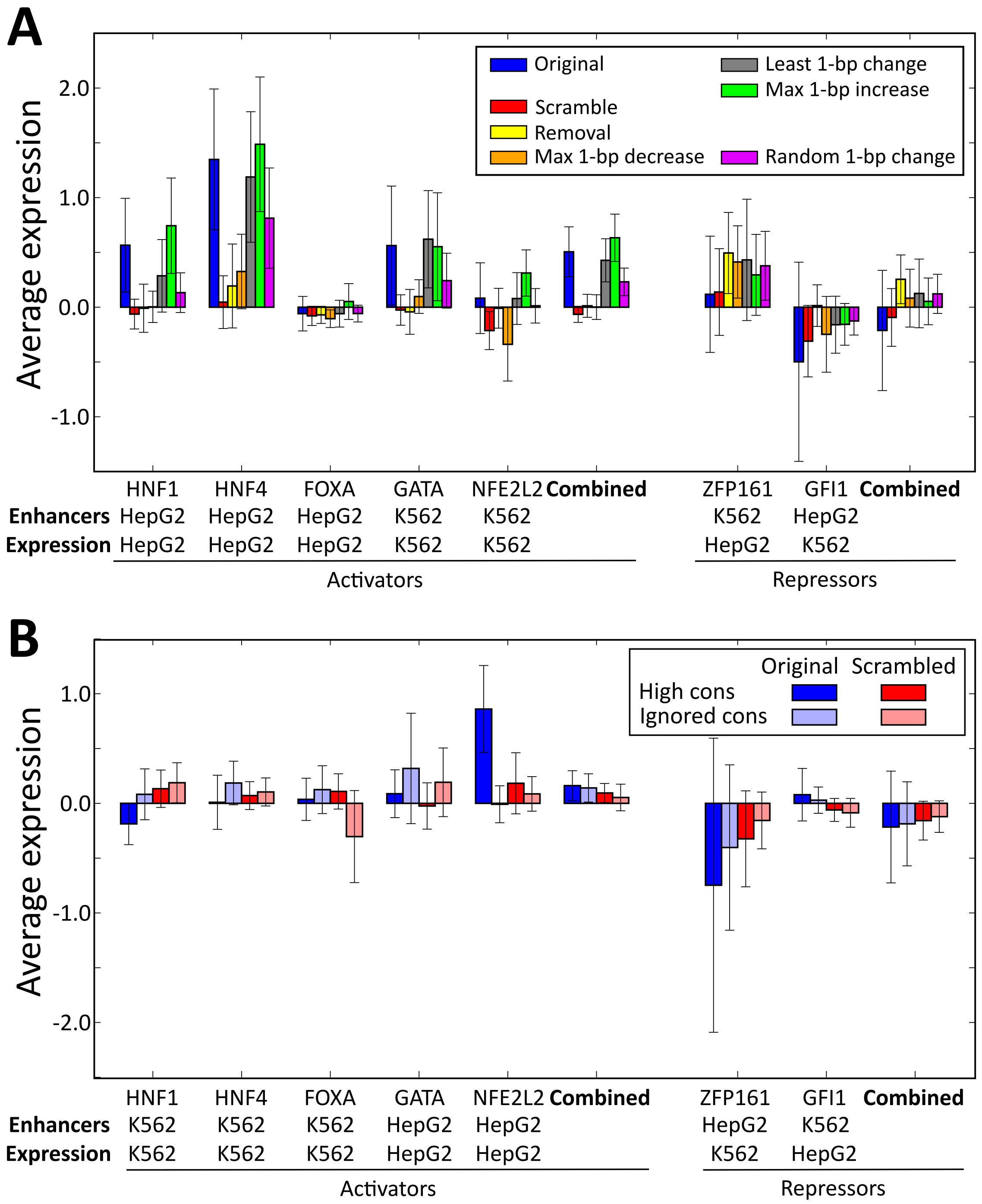

Figure S4: Additional bar plots.

(A) Bar plots showing differences amongst locations tested for additional modifications. Because these are restricted to only these regions, the original and scramble value differs from those in Figure 3B.

(B) As with Figure 3B, except with flipped cell types for both enhancer selection and cell type (using the 18 control enhancers tested per factor). Lack of significant difference is seen for every factor except NFE2L2.

A

					High conservation						Ignored conservation					
		Enhancer cell line	Expression cell line	High vs. ignored conserv. (160)^{a}	Scramble (160)^{b}	Removal (15)^{b}	Max 1-bp decrease (15)^{b}	Least 1-bp change (15)^{b}	Max 1-bp increase (15)^{b}	Random 1-bp change (30)^{b}	Scramble (160)^{b}	Removal (15)^{b}	Max 1-bp decrease (15)^{b}	Least 1-bp change (15)^{b}	Max 1-bp increase (15)^{b}	Random 1-bp change (30)^{b}
Activators	HNF1	HepG2	HepG2	2.3 x 10^{-4}	**1.7 x 10^{-13}**	0.0409	0.0038	0.0146	(0.0783)	8.9 x 10^{-5}	**1.9 x 10^{-6}**	0.8647	(0.7333)	(0.6496)	(0.0199)	(0.5038)
	HNF4	HepG2	HepG2	2.1 x 10^{-4}	**8.0 x 10^{-15}**	0.0018	0.0012	0.4955	(0.0609)	1.3 x 10^{-4}	**4.4 x 10^{-4}**	(0.9096)	0.9096	0.9096	(0.0090)	(0.8130)
	FOXA	HepG2	HepG2	0.1618	0.0035	(0.8647)	(0.7764)	(0.4955)	(0.1398)	(0.7343)	0.5947	0.0736	0.3003	0.7299	0.4326	0.3164
	GATA	K562	K562	1.7 x 10^{-6}	**1.9 x 10^{-18}**	0.1118	0.1398	(0.9547)	0.9547	0.0196	**9.3 x 10^{-6}**	0.4265	0.7333	(0.4955)	(0.0054)	(0.5857)
	NFE2L2	K562	K562	0.0352	**4.3 x 10^{-11}**	0.4326	0.0303	0.4326	(0.2209)	0.3745	**6.3 x 10^{-7}**	0.3739	0.1307	0.5937	(0.1823)	0.0015
	Combined			9.0 x 10^{-13}	**2.8 x 10^{-54}**	**1.5 x 10^{-4}**	**1.7 x 10^{-6}**	**0.0814**	**(0.0056)**	1.6 x 10^{-7}	**5.1 x 10^{-17}**	0.1031	0.1663	(0.9557)	(2.0 x 10^{-4})	0.1831
Repressors	ZFP161	K562	HepG2	0.8714	0.9325	(0.2213)	(0.7007)	(0.0281)	(0.5829)	(0.1919)	0.7643	(0.3882)	(0.3739)	(1.0000)	0.7213	(0.5272)
	GFI1	HepG2	K562	(0.0417)	**(0.0369)**	(0.4955)	0.3942	(0.5701)	0.3942	0.7499	0.9001	(0.1556)	(0.1252)	0.9547	(0.7333)	0.1714
	Combined			(0.0490)	(0.0815)	(0.1648)	0.7327	(0.0517)	0.9425	(0.4212)	0.8254	(0.1023)	(0.0819)	0.8689	(0.9036)	0.6570

B

					High conservation						Ignored conservation					
		Enhancer cell line	Expression cell line	High vs. ignored conserv. (160)^{a}	Scramble (160)^{b}	Removal (15)^{b}	Max 1-bp decrease (15)^{b}	Least 1-bp change (15)^{b}	Max 1-bp increase (15)^{b}	Random 1-bp change (30)^{b}	Scramble (160)^{b}	Removal (15)^{b}	Max 1-bp decrease (15)^{b}	Least 1-bp change (15)^{b}	Max 1-bp increase (15)^{b}	Random 1-bp change (30)^{b}
Activators	HNF1	HepG2	K562	0.3728	0.1671	(1.0000)	0.5701	0.5701	0.0995	0.4165	0.8338	(0.3066)	(0.3635)	(0.4955)	(0.0535)	(0.3820)
	HNF4	HepG2	K562	(0.3290)	(0.4116)	(0.7333)	(0.2330)	(0.5321)	0.3942	(0.6143)	(0.8352)	0.2115	0.2330	(0.0995)	(0.7764)	(0.9754)
	FOXA	HepG2	K562	(0.9747)	0.8902	(0.1556)	(0.2560)	(0.1252)	(0.0468)	(0.0020)	(0.1788)	0.9750	(0.7299)	0.6378	0.9750	(0.7847)
	GATA	K562	HepG2	0.0155	0.0440	(0.4603)	0.9547	(0.6092)	(0.3635)	(0.3286)	(0.2652)	(0.5701)	(0.2115)	0.1118	0.7333	0.6884
	NFE2L2	K562	HepG2	0.0316	1.1 x 10^{-12}	0.2719	0.2209	0.9250	(0.0029)	0.0305	6.5 x 10^{-8}	0.7221	0.6566	0.9292	(0.2477)	0.0240
	Combined			0.0465	3.3 x 10^{-6}	(0.5024)	(0.7939)	(0.4237)	(0.1349)	(0.6268)	0.1269	(0.9697)	(0.7058)	0.9697	(0.1108)	0.5944
Repressors	ZFP161	K562	K562	(0.2057)	0.0529	0.1520	0.7532	0.1579	0.1579	0.7366	0.0534	(0.3465)	(0.5337)	(0.8589)	(0.7213)	(0.7164)
	GFI1	HepG2	HepG2	(0.2884)	**(0.5798)**	(0.2115)	(0.6496)	0.9096	(0.1398)	(0.3493)	(0.6216)	(0.1398)	(0.3066)	(0.2330)	(0.1252)	(0.0207)
	Combined			(0.0295)	0.5343	0.6987	(1.0000)	0.3246	0.7731	(0.9332)	0.3398	(0.0926)	(0.2087)	(0.4237)	(0.1919)	(0.0629)

C

					High conservation	Ignored conservation
		Enhancer cell line	Expression cell line	High vs. ignored conserv. (18)^{a}	Scramble (18)^{b}	Scramble (18)^{b}
Activators	HNF1	K562	K562	(0.1249)	(0.0347)	(0.5566)
	HNF4	K562	K562	(0.3038)	0.9133	0.5862
	FOXA	K562	K562	(0.4290)	(0.6475)	0.1446
	GATA	HepG2	HepG2	(0.8993)	0.7112	0.9133
	NFE2L2	HepG2	HepG2	3.5 x 10^{-4}	0.0018	(0.3958)
	Combined			0.6972	0.4579	0.5828
Repressors	ZFP161	HepG2	K562	(0.6316)	(0.7989)	0.8446
	GFI1	K562	HepG2	(0.7517)	0.9058	0.2145
	Combined			(0.7047)	0.9808	0.3223

p = 10^{-10} p > 0.05 p = 10^{-10}

higher in modifed lower in modifed

Figure S5: P-values comparing enhancer source and modifications across cell types.

(A) P-values showing a consistent and strongly significant reduction in expression specific to disruption of motif binding sites. P-values are computed using the ^{a}Wilcoxon signed-rank and ^{b}Mann-Whitney U two-tailed non-parametric tests (details in Methods). Headers indicate the maximum number of tested sequences per factor in parenthesis (precise numbers are in Table S1). P-values in parenthesis indicate an increased value in either the ^{a}instances ignoring conservation or for the ^{b}modified sequence (this is redundant with the blue/red color). Values used in in the text are highlighted in bold. Corresponding to bar plots in Figure 3B and Figure S4A.

(B) Same as (A), except with expression cell line reversed to demonstrate cell type specificity. A notable exception is NFE2L2.

(C) Same as (A), except with both expression and enhancer cell lines switched. Only 18 such sites were selected per factor, and they were only tested with the scramble modification. Corresponding to bar plots in Figure S4B.

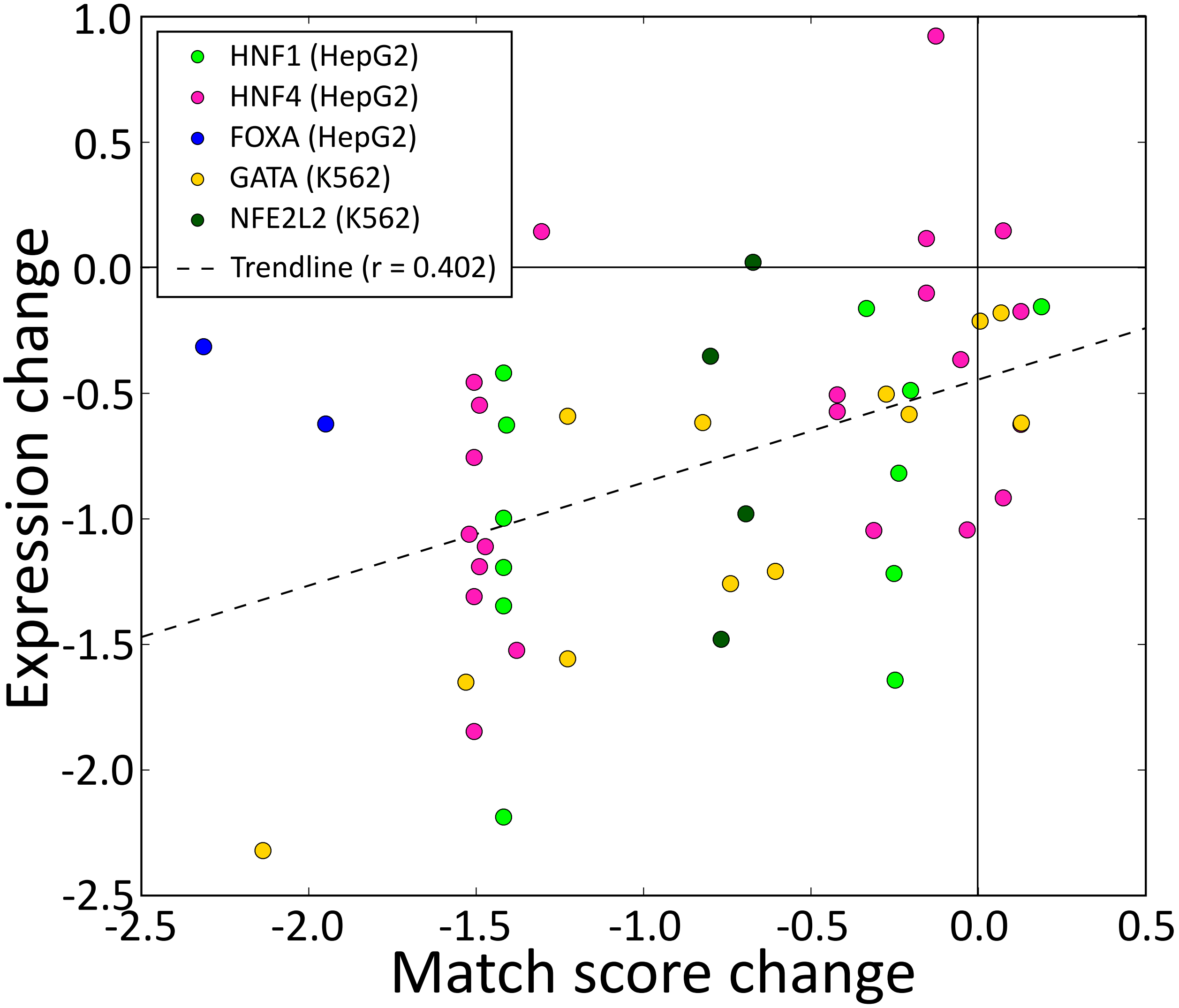

Figure S6: Change in expression for random manipulations of conserved instances with wild type expression score of at least 0.5 (52 of the 148 random manipulations). A correlation is seen between the change in match score strength and the corresponding change in expression ($r = 0.402$; random permutation $P = 2.8 \times 10^{-3}$). Match score is normalized for each motif such that 0 is the weakest possible permitted match and 1 is a perfect match (scores can be less than 0 after the manipulations).

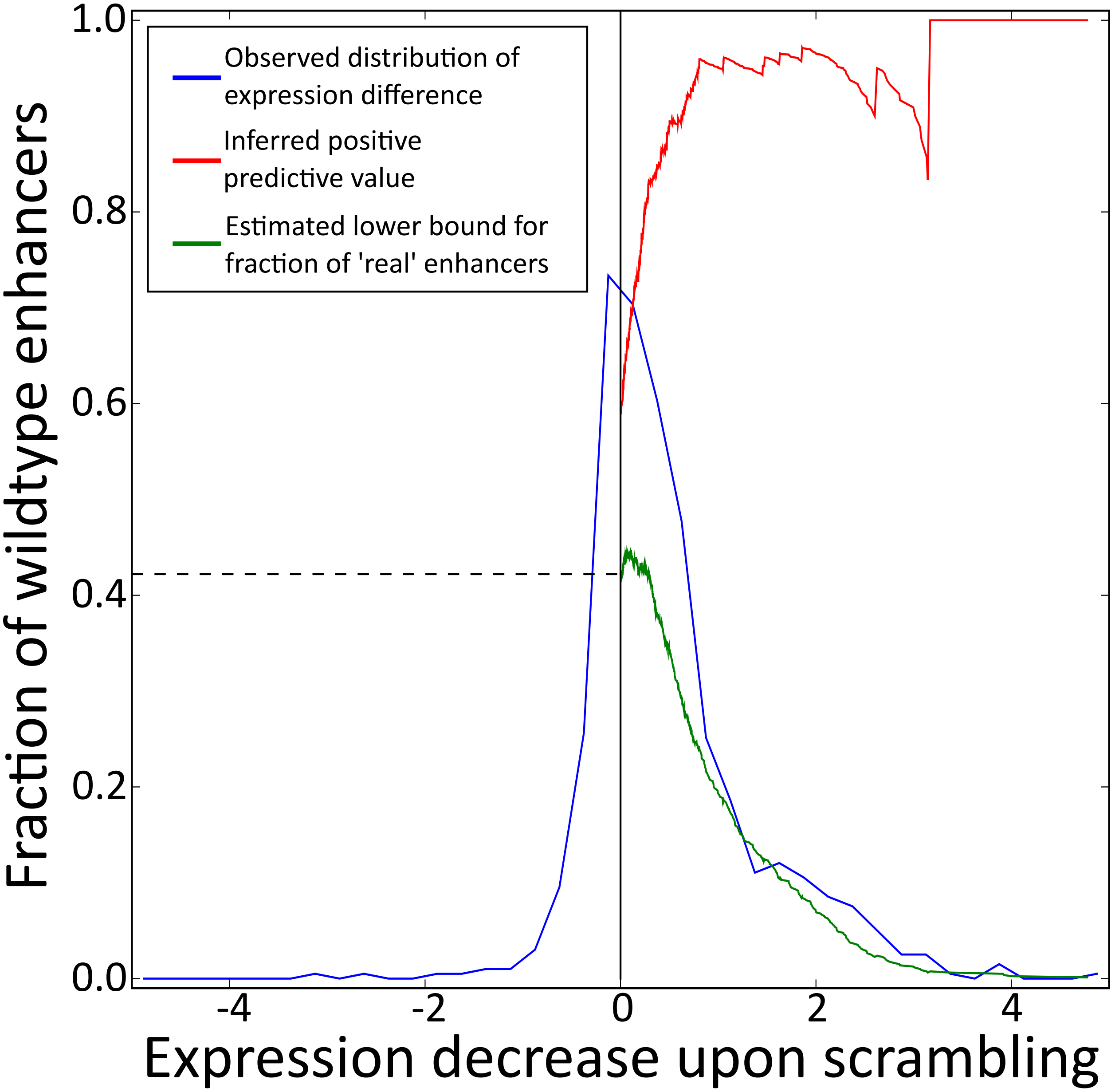

Figure S7: The change in reporter expression upon motif disruption allows us to estimate that $T_{\mathrm{true}}=T_{\mathrm{pos}}-T_{\mathrm{false}}=42\%$ (see Methods) of enhancers (assuming that $T_{\mathrm{false}}=1-T_{\mathrm{pos}}=29\%$) tested show wild-type reporter expression using only a 145-bp sequence centered at conserved motifs, a fraction similar to previous studies (Pennacchio et al., 2006; Heintzman et al., 2009; Attanasio et al., 2008). The estimated lower bound for fraction of 'real' enhancers is computed by subtracting the fraction of enhancers with expression greater than the value indicated on the x-axis by those with expression less than negative that value. The inferred positive predictive value is then this value divided by the fraction of enhancers with expression greater than the x-axis.

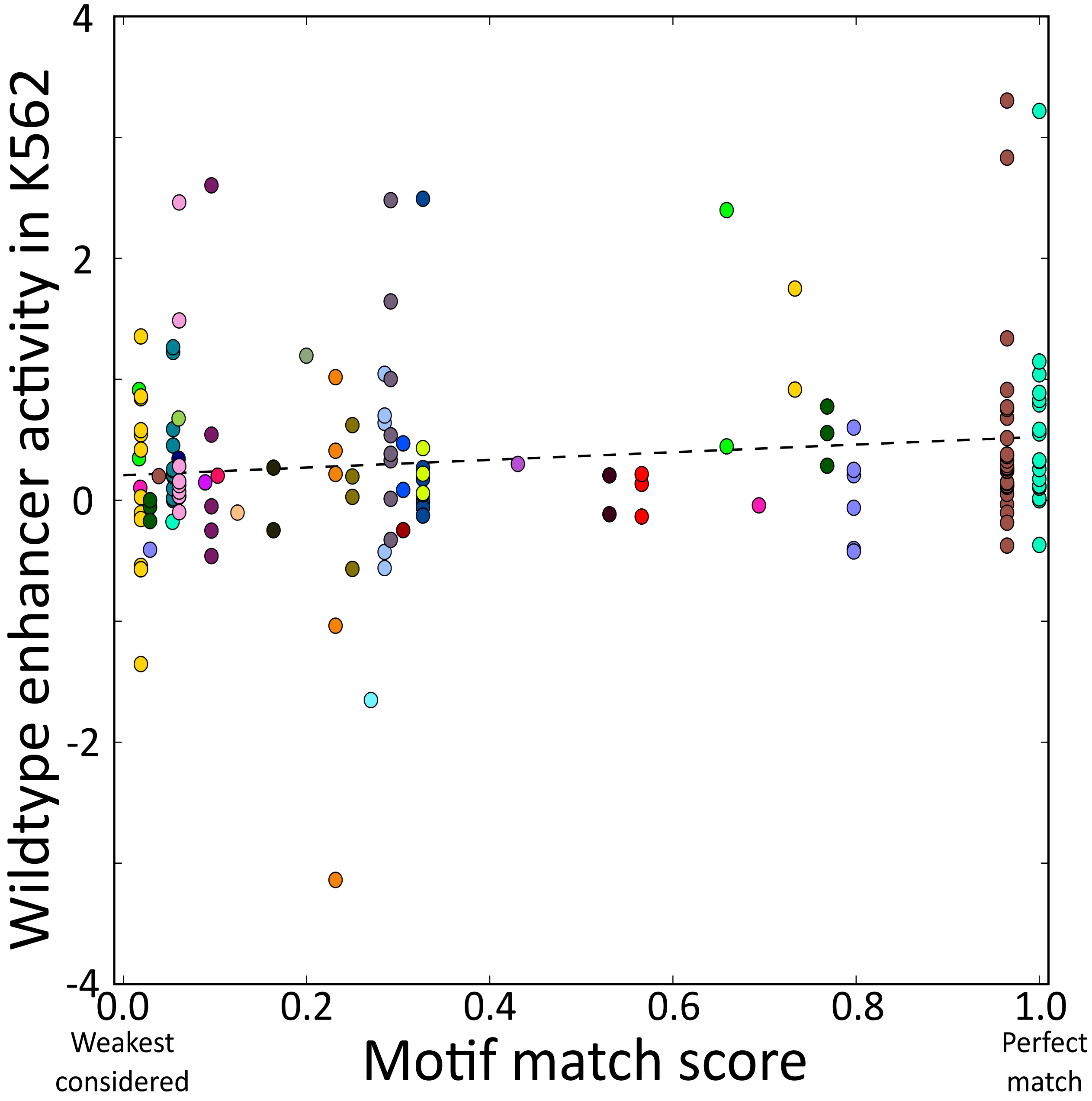

Figure S8: The average wild-type reporter expression (y-axis) for 160 enhancers in K562 centered on conserved NFE2L2 motifs shows dramatic variation based on context even for enhancers that contain identical NFE2L2 motif sequences (vertical columns), emphasizing that motif match score (x-axis) is necessary but not sufficient for enhancer activity. Colors are used only to emphasize identical sequences (but may be reused when unambiguous).

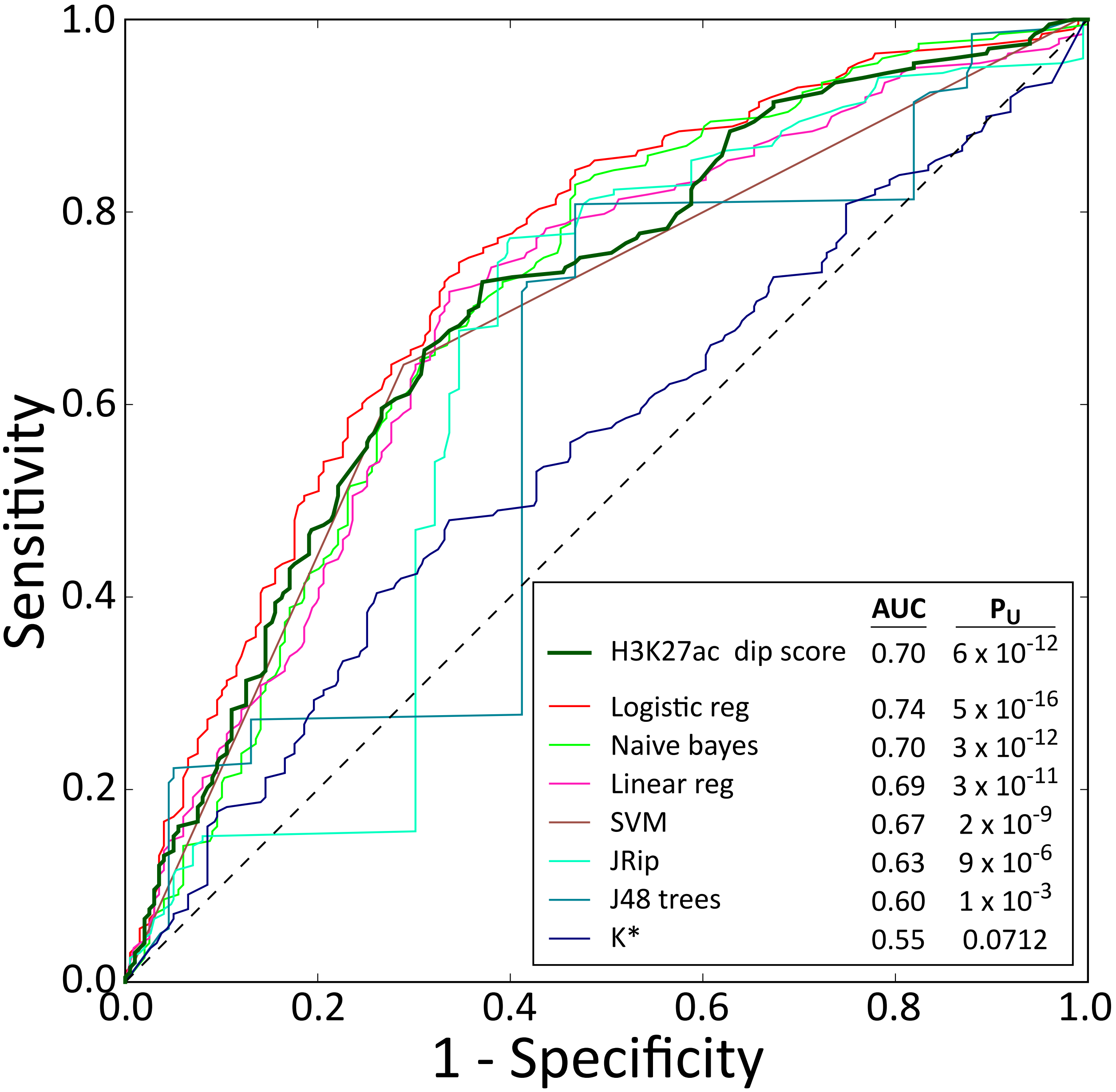

Figure S9: Combination of the properties shown in Figure 5C-D using various machine learning techniques and comparison to best individual feature (H3K27ac dip). We employed 7 machine learning techniques implemented by WEKA (v3.6.4; Hall et al., 2009) and scored each of top and bottom 25% of sequences for each activator dataset in a leave-one-out cross-validation framework. Three pairs of sequences overlapped and were placed in the same cross-validation group. We find that the logistic regression algorithm outperforms the best individual feature (AUC 0.74 vs. 0.70).

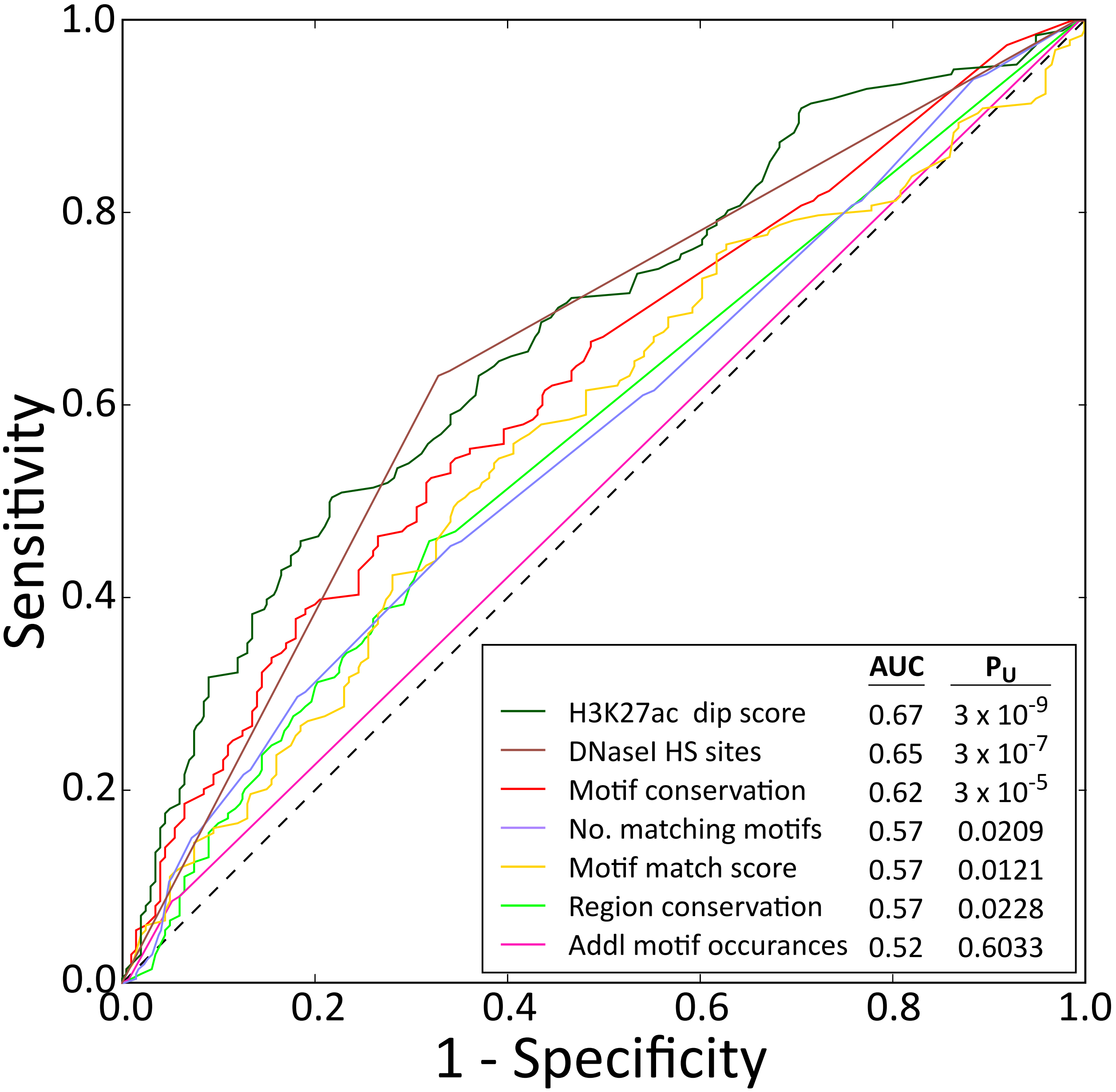

Figure S10: Predictive power analogous to Figure 5B, except separating the top and bottom 25% of sequences in terms of change in expression upon motif scrambling (rather than expression value). AUC values for each figure are similar to those in Figure 5B.

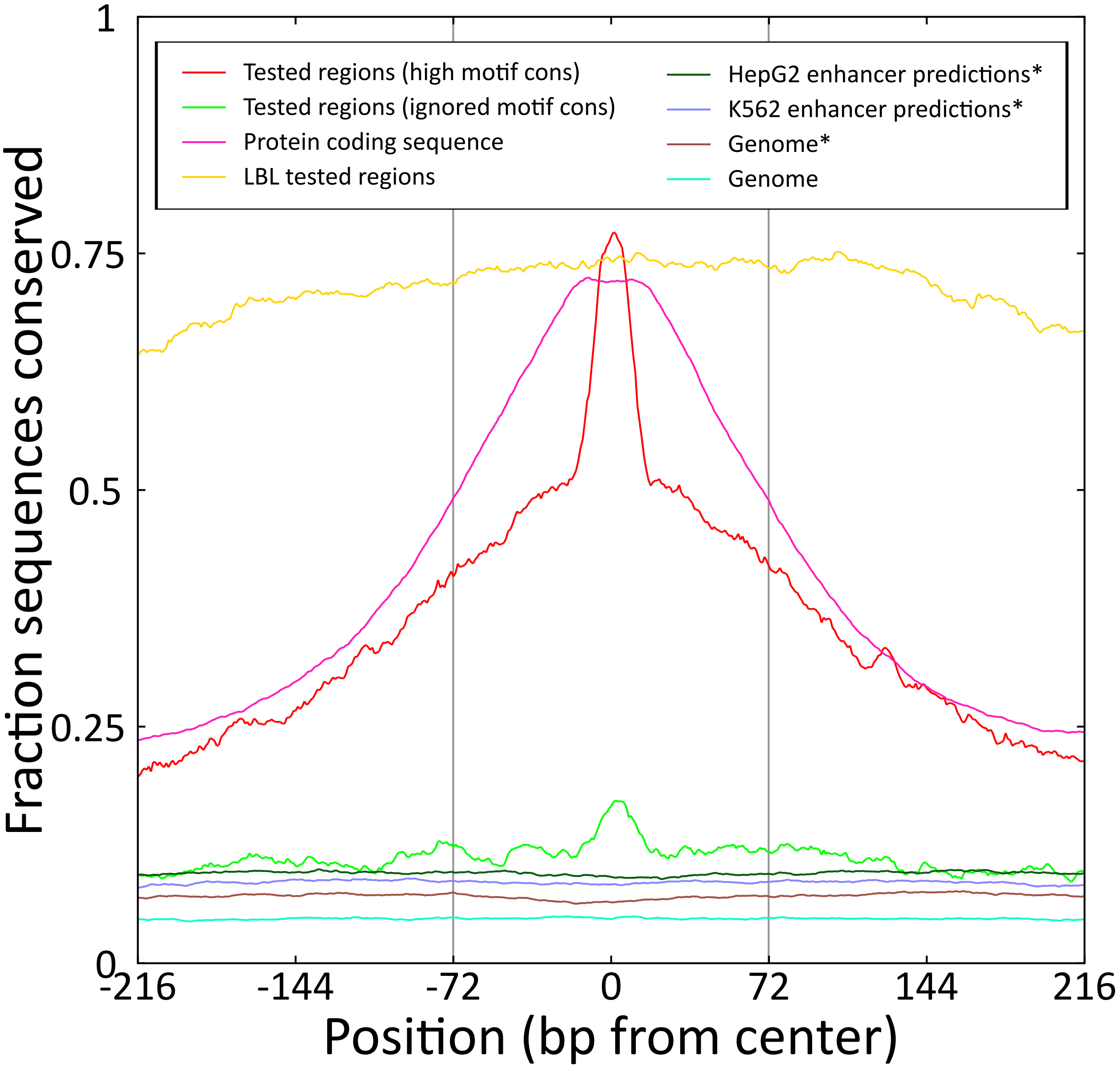

Figure S11: Comparison of the conservation (as assessed by SiPhy-ω on 29 mammals; Lindblad-Toh et al. 2011) for: the tested putative enhancer regions, human LBL sequences (Pennacchio et al. 2006), and 25,000 randomly selected 15mers from protein coding exons, the input enhancer regions, and the genome. The tested regions, including those not selected using a conserved motif instance, have a higher level of conservation than the chromatin-based input enhancer regions, particularly at the location of the motif, but are on average less conserved than the LBL tested sequences. Sequences aligned by center. *Indicates regions were confined to those scanned for motifs (excluding coding exons, $3'$ UTRs, and repeats; see Methods).

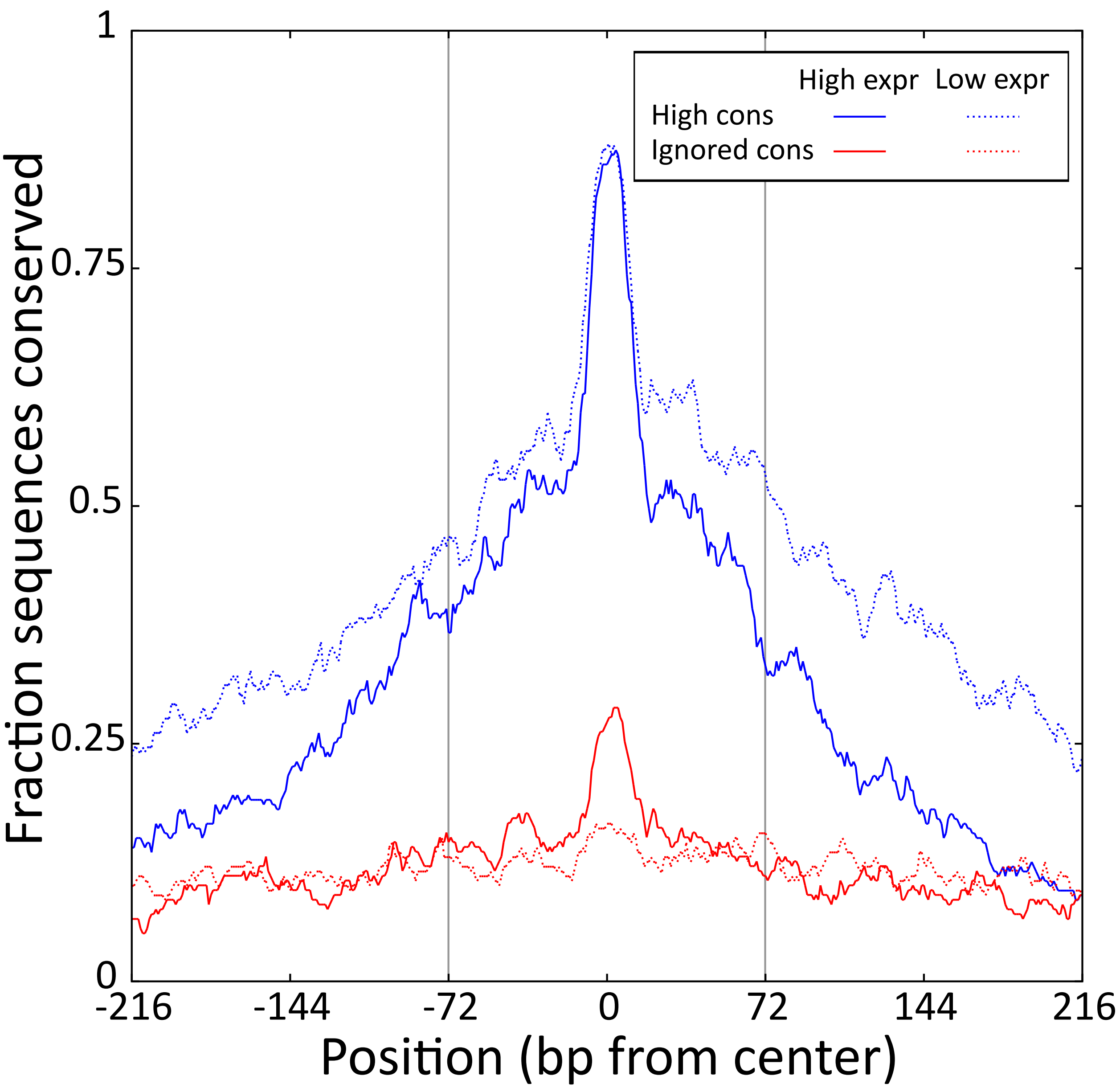

Figure S12: Comparison of conservation level (as in Figure S11) of the top vs. bottom 25% most expressed sequences containing an activator motif in the matched cell line. For sequences selected with a conserved motif, lower average conservation is seen outside the motif, indicating that specificity of conservation to the motif instance is indicative of expression. For sequences selected ignoring conservation, motif conservation is higher for the more highly expressed sequences.

1. Tested sequence number, matching Data S1
2. Name of tested transcription factor
3. Cell line enhancer was identified in
4. Conservation level (either IgnoredCons or HighCons)
5. Position of motif match (in hg18)
   (a) Chromosome
   (b) Motif start (1-indexed, inclusive)
   (c) Motif end (1-indexed, inclusive)
   (d) Strand
   (e) Sequence start (1-indexed, inclusive)
   (f) Sequence end (1-indexed, inclusive)
6. Original (subcolumns repeated for engineered sequences): (a) Tested sequence[a] (b) Sequence ID (as used in GEO) (c) HepG2 normalized expression (d) HepG2 expressed p-value[b] (e) HepG2 repressed p-value[b] (f) K562 normalized expression (g) K562 expressed p-value[b] (h) K562 repressed p-value[b]
7. Scramble
8. Removal
9. Max 1-bp decrease
10. Least 1-bp change
11. Max 1-bp increase
12. Random 1-bp change #1
13. Random 1-bp change #2

Table S3: Tab separated file with the following columns (subcolumns are indicated by letters and are semi-colon separated). Ordered by column 1, which is in decreasing order of maximal expression level in either cell line for the wild-type original sequence. Raw data (e.g. individual replicates) are available in Gene Expression Omnibus (GEO) under accession GSE33367 and also in the supplemental materials. [a]synthesized sequence may have been reverse complemented; compare to sequence indicated for ID in GEO [b]computed using a one-tailed Mann-Whitney test comparing the values across all replicates (increased for expressed, decreased for repressed) for this sequence versus all values for sequences with scrambled motifs in the same cell type. Table is available as a separate file in the supplemental materials.

Seq Num	Factor	Cell Type	MPRA (SV40 promoter)			Luciferase (SV40 promoter)			Luciferase (TATA promoter)		
			Original	Scramble	P-value	Original	Scramble	P-value	Original	Scramble	P-value
10	FOXA	HepG2	3.80	-0.12	3.1×10^{-15}	-1.87	-4.56	3.5×10^{-4}	-8.91	-11.23	5.4×10^{-3}
		K562	0.34	0.38	0.963	1.11	1.33	0.374	-6.53	-7.43	0.230
15	HNF1	HepG2	3.47	0.10	6.3×10^{-16}	0.29	-2.95	3.8×10^{-5}	-1.89	-6.08	3.3×10^{-5}
		K562	0.32	0.13	0.143	0.89	1.55	0.087	-6.87	-6.97	0.712
25	HNF4	HepG2	3.21	0.05	2.4×10^{-17}	-2.79	-4.81	9.6×10^{-4}	-6.98	-8.99	9.9×10^{-6}
		K562	0.26	-0.38	0.175	1.35	0.94	0.074	-4.27	-4.88	0.058
40	GATA	HepG2	0.01	-0.27	0.880	-3.48	-3.62	0.259	-6.77	-7.11	0.234
		K562	2.88	-0.14	3.3×10^{-8}	5.47	1.90	8.6×10^{-7}	0.74	-4.89	2.1×10^{-3}
66	NFE2L2	HepG2	1.11	0.08	1.5×10^{-3}	-1.40	-4.12	1.3×10^{-3}	-5.06	-11.16	1.0×10^{-5}
		K562	2.48	0.53	8.1×10^{-5}	4.88	2.26	3.4×10^{-4}	0.12	-6.70	3.5×10^{-7}
129	HNF1	HepG2	1.88	-0.00	1.1×10^{-14}	0.19	-3.49	1.1×10^{-5}	-1.63	-7.69	8.3×10^{-6}
		K562	-0.30	0.08	0.073	1.09	1.46	5.2×10^{-4}	-7.03	-6.82	0.426
344	HNF4	HepG2	0.91	-0.20	0.090	-3.72	-4.49	0.020	-7.41	-8.65	2.7×10^{-3}
		K562	-0.13	-0.58	0.058	1.08	1.14	0.702	-6.51	-6.94	0.139
1476	ZFP161	HepG2	-0.55	0.58	0.523	-4.58	-4.58	0.984	-8.94	-8.70	0.055
		K562	0.10	-0.33	0.414	2.25	1.89	0.033	-3.93	-4.33	0.110
1929	HNF1	HepG2	-0.16	0.13	0.245	-2.25	-2.70	0.091	-5.66	-6.39	0.024
		K562	-0.07	0.10	0.703	NA	0.94	NA	-5.54	-6.71	0.021
2302	GFI1	HepG2	0.37^a	0.18	0.668^a	-3.04	-4.45	1.6×10^{-3}	-10.87	-10.37	0.345
		K562	-1.21^a	-0.10	0.020^a	0.11	1.31	1.1×10^{-3}	-7.48	-7.10	0.518

Table S4: Summary of results from luciferase validation and comparison to MPRA expression values. All expression values are $\log_2$ (including those used in t-tests; see Methods). P-values are computed using two-tailed unpaired, unequal variance t-tests comparing the replicates for the original sequence to the sequence with the motif scrambled (see Methods). 'NA' indicates sequences for which all replicates failed. ^{a}Replaced with values for 1-bp neutral manipulation of wild-type sequence because only one replicate was available (insufficient for computing t-test p-values). Full data from validation available in Table S5.

1. Tested sequence number, matching Data S1 and Table S3
2. Name of tested transcription factor
3. Position of motif match (in hg18)
    (a) Chromosome
    (b) Motif start (1-indexed, inclusive)
    (c) Motif end (1-indexed, inclusive)
    (d) Strand
    (e) Sequence start (1-indexed, inclusive)
    (f) Sequence end (1-indexed, inclusive)
4. Cell line expression was measured in (one line for each of HepG2/K562)
5. Original MPRA experiment values with 145bp sequence:
    (a) Average expression for WT sequence
    (b) Average expression for scramble sequence
    (c) Two-tailed t-test p-value
    (d) Two-tailed Mann-Whitney p-value
6. Luciferase experiment with SV40 promter (a-d as for MPRA; subcolumns repeated for TATA promoter):
    (e) Replicate values for WT (comma separated; 4 per sequence/promoter; $\log_2$)
    (f) Replicate values for scramble
7. Luciferase experiment with TATA promoter

Table S5: Tab separated file with the following columns (subcolumns are indicated by letters and are semi-colon separated). Expression values are computed as described in the methods (separately for MPRA and luciferase validation) and are in both cases $\log_2$. P-values are computed on the individual replicate values for each pair of sequences that is being compared. The t-test and Mann-Whitney p-values are similar for MPRA: all t-test significant p-values are also significant with Mann-Whitney ($P < 0.05$) and only HNF4 (#344) in HepG2 is significant with Mann-Whitney but not the t-test. Generally, t-test p-values are smaller for the positive luciferase results because on four replicates the two-sided Mann-Whitney has a minimum of $P_U = 0.0286$. Six (out of 160) failed replicates for luciferase experiments are indicated with 'NA'. Table is available as a separate file in the supplemental materials.