Supplementary Figures for
"Diversity and Complexity in DNA Recognition by Transcription Factors"

Figure S1: Cloning strategy. **p. 3**

Figure S2: Comparison of PBM data for DBD versus full-length constructs for 5 TFs. **(A)** Motif logo comparisons and *k*-mer correlations, and **(B)** *k*-mer PBM enrichment score scatter plots. **p. 4-5**

Figure S3: Comparison of TFs overexpressed and purified from *E. coli* versus expressed by *in vitro* transcription and translation. **(A)** Motif logos, and **(B)** *k*-mer correlation plots, from PBM experiments. **p. 7**

Figure S4: PBM data reproducibility. **(A)-(D)** Clustergram of *k*-mers for all PBM data prior to combining data from array designs #1 and #2, showing that array designs #1 and #2 cluster together for each protein. **(E)** Reproducibility of E-scores and Z-scores from array designs #1 and #2. **p. 7-11**

Figure S5: Agreement of PBM *k*-mer data with prior motif data, in general. **p. 12-15**

Figure S6: Comparison of PBM data versus K_d data for Max and for the yeast TF Cbf1. **p. 16-17**

Figure S7: Confirmation of PBM-derived motifs by EMSAs for three newly characterized proteins and one recently characterized protein. **p. 18**

Figure S8: Binding profiles of specific TF DBD structural classes. **(A)** HMG/SOX, **(B)** AP-2, **(C)** ARID/BRIGHT, **(D)** bZIP, **(E)** ZnF_C4, **(F)** E2F, **(G)** ETS, **(H)** Forkhead, **(I)** GATA, **(J)** HLH, **(K)** homeodomain, **(L)** IRF, **(M)** RFX, **(N)** SAND DNA-binding domains. **p. 19-33**

Figure S9: Confirmation of secondary motifs by EMSAs for 6 TFs: Hnf4a, Nkx3.1, Mybl1, Foxj3, Rfxdc2 and Myb. **p. 34-35**

Figure S10: Primary, secondary, and tertiary Seed-N-Wobble motifs identified in PBM data for the human POU homeodomain TF Oct-1. **p. 36**

Figure S11: High-scoring *k*-mers belonging to the Jundm2 secondary motif are not bound as well by the related bZIP protein Atf1. **p. 37**

Figure S12: RFX protein-DNA recognition positions. **p. 38**

Figure S13: Graphs showing $\log_{10}$(1-AUC) (area under ROC curve) (*y*-axis) versus $\log_{10}$(number of positives) (*x*-axis) for Hnf4a. **p. 39-40**

Figure S14: Enrichment of primary versus secondary motif 8-mers bound *in vitro* within genomic regions bound *in vivo* for **(A, C, D)** Hnf4a and **(B, E, F)** Bcl6b. **p. 41-45**

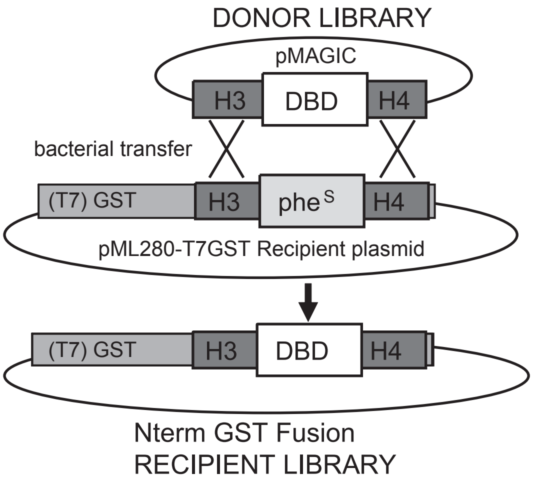


Figure S1

"MAGIC" system to express GST fusion proteins.
DNA-binding domains (DBDs) were cloned into a pMAGIC Donor vector, enabling a bacterial transfer of DBDs into pML280-T7GST , by "mating-assisted genetically integrated cloning" (MAGIC, see Li et al. 2005), generating a recipient library expressing N-term GST fusion-DBD.

(A)

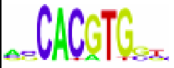
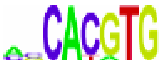
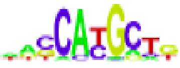
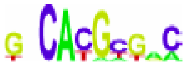
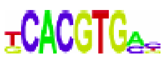
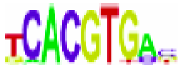
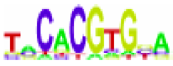
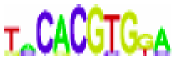
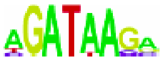
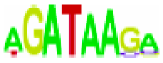
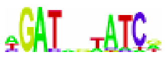
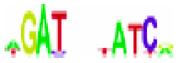
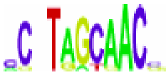
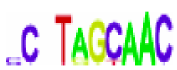
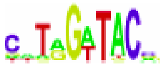
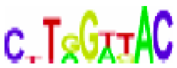
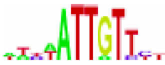
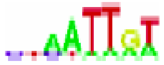
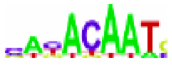
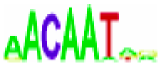
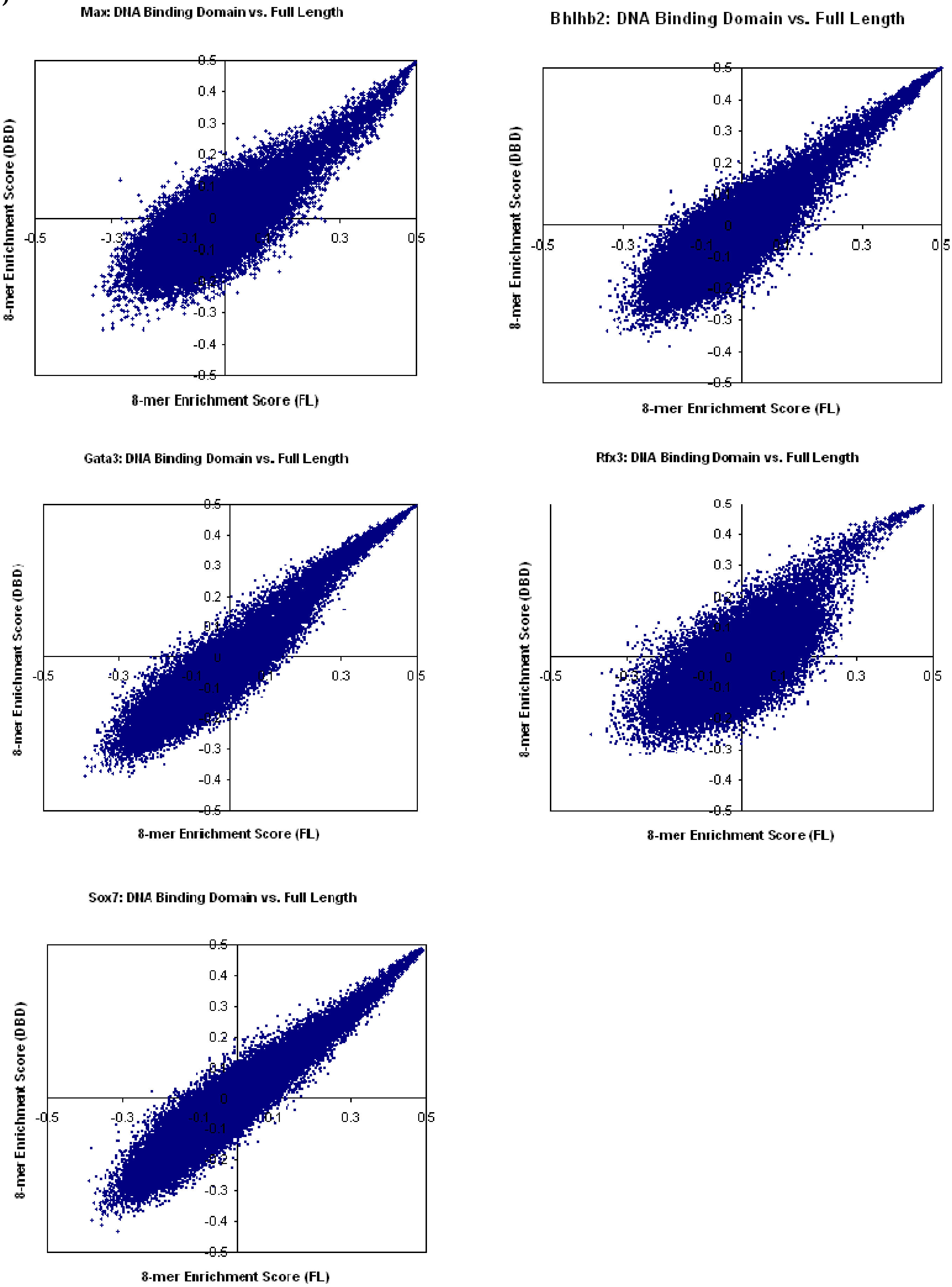
Protein	Primary Motif DNA Binding Domain	Primary Motif Full Length	Secondary Motif DNA Binding Domain	Secondary Motif Full Length	8-mer E-score Pearson (R)	8-mer E-score Spearman (R')
Max					0.81	0.72
Bhlhb2					0.88	0.80
Gata3					0.94	0.90
Rfx3					0.72	0.67
Sox7					0.94	0.93

Figure S2: Comparison of PBM data for DNA binding domain versus full-length protein. We created two constructs for five transcription factors: one encompassing just the DNA binding domain, and one spanning the entire protein. Each protein was applied to two PBMs of independent sequence designs, and we compared the motifs and 8-mer scores after combining the data from these arrays. **(A)** Primary and secondary motifs from Seed-and-Wobble, and correlations of 8-mer enrichment scores (E-scores) for DNA binding domain and full-length proteins. Both constructs produced essentially identical motifs by the Seed-and-Wobble algorithm and highly correlated E-scores across all 8-mers. **(B) (next page)** Scatter plots of 8-mer E-scores for the two constructs (DNA binding domain versus full-length) of these five proteins.

(B)



Max: DNA Binding Domain vs. Full Length

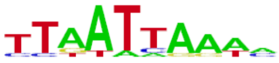
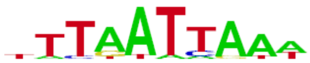
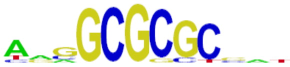
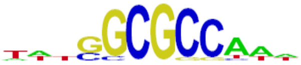
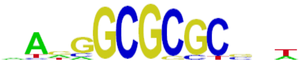
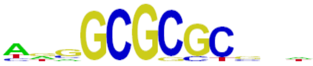
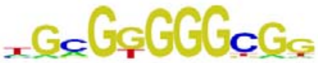
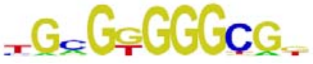
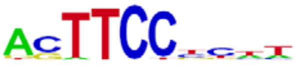
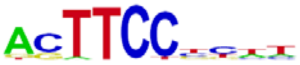
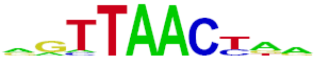
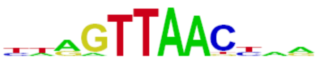
Bhlhb2: DNA Binding Domain vs. Full Length

Gata3: DNA Binding Domain vs. Full Length

Rfx3: DNA Binding Domain vs. Full Length

Sox7: DNA Binding Domain vs. Full Length

(A)

Protein	**Motif** *E. coli* purification	**Motif** *in vitro* purification	**8-mer E-score** **Pearson (R)**	**8-mer E-score** **Spearman (R')**
Arid3a			0.85	0.80
E2F2			0.92	0.85
E2F3			0.94	0.88
Egr1			0.89	0.82
Sfpi1			0.91	0.89
Tcf1			0.85	0.80

(B)

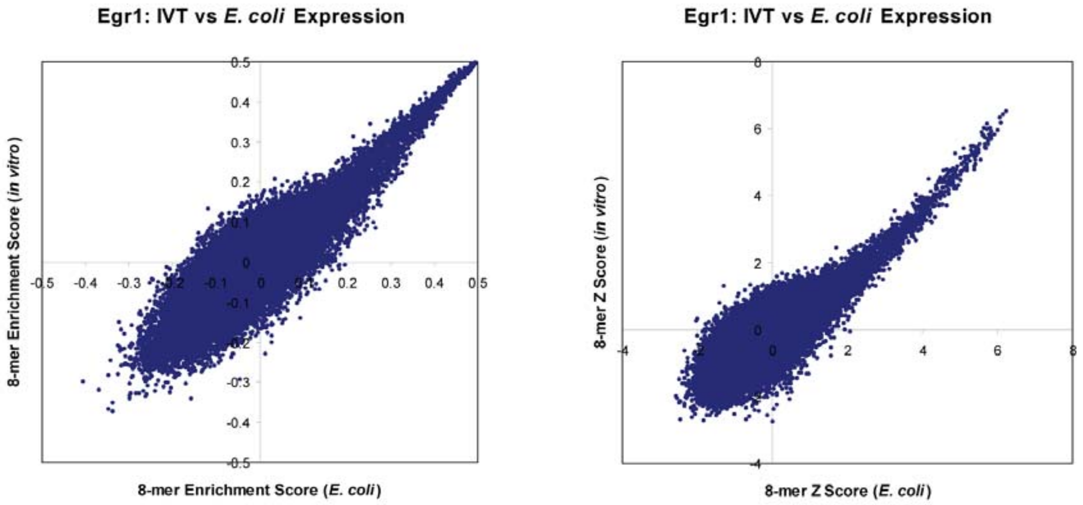
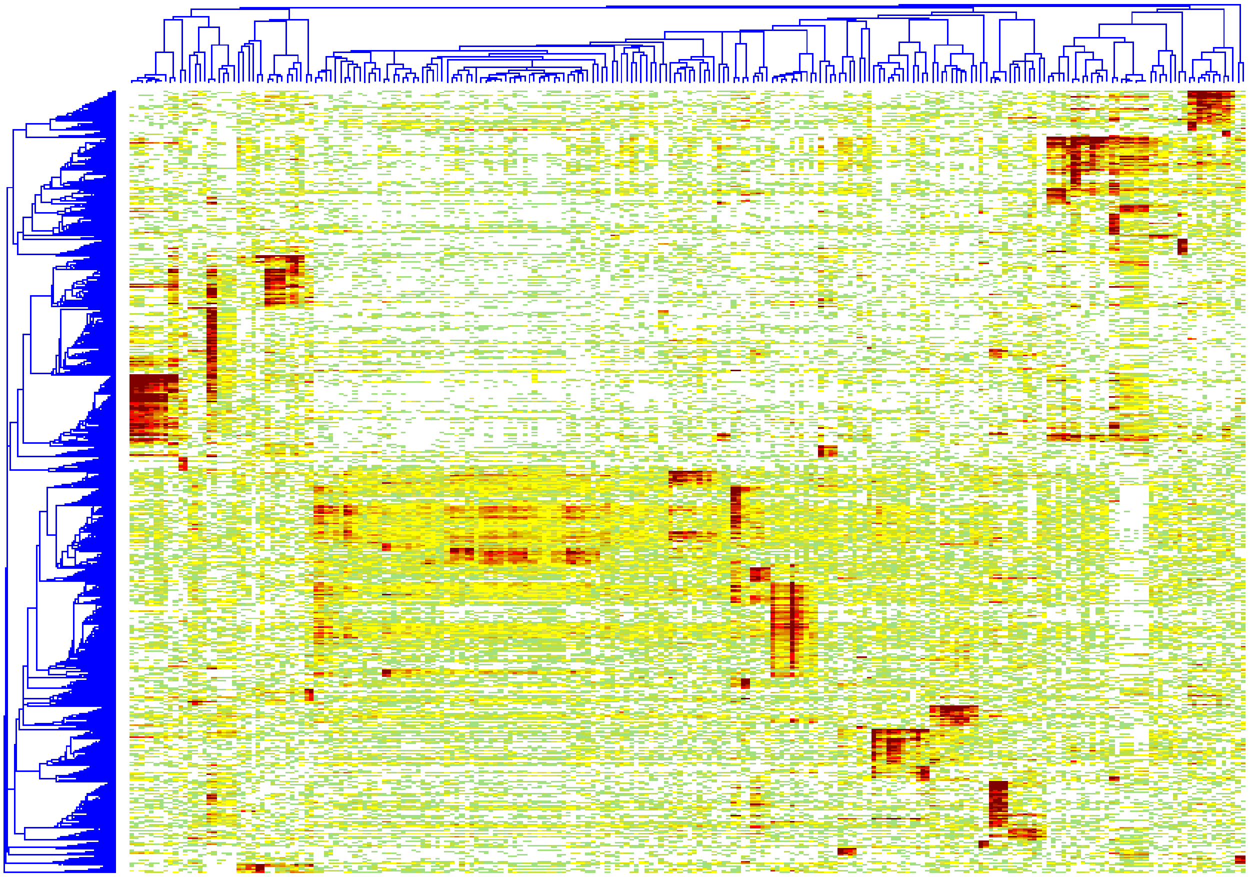


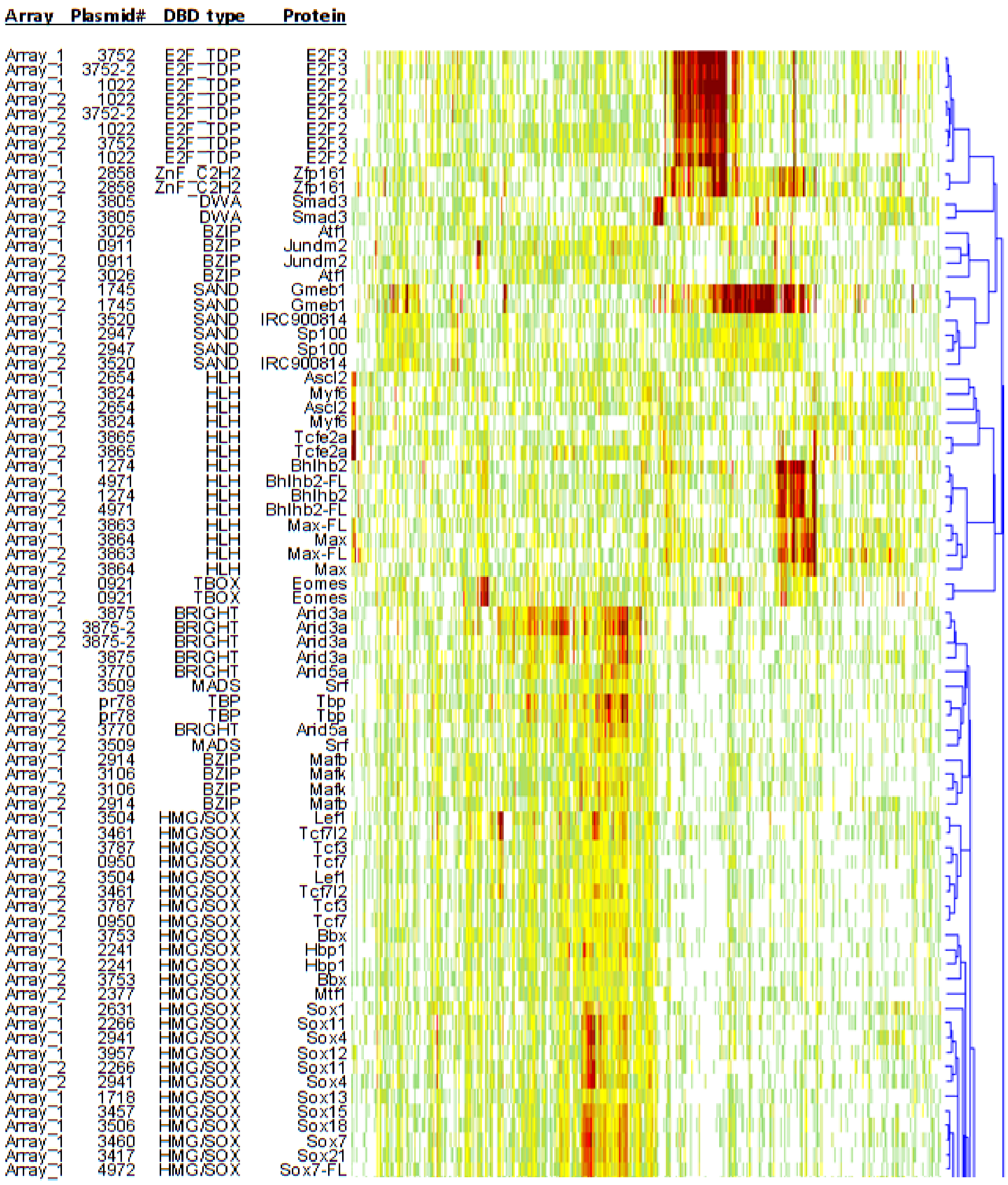
Figure S3: *E. coli in vivo* versus *in vitro* protein expression. We expressed six proteins both in *E. coli* (*in vivo*) and *in vitro* (see **Methods**) and performed PBM experiments to determine the data reproducibility for different methods of protein production. Proteins expressed in vivo were purified by GST affinity chromatography (see **Methods**). Each individual protein sample was applied to two PBMs of independent sequence designs, and we compared the motifs and 8-mer scores after combining the data from both arrays. **(A)** Both methods of protein expression produced essentially identical motifs by the Seed-and-Wobble algorithm and highly correlated Enrichment scores (E-scores) across all 8-mers. **(B)** Correlation of 8-mer E-scores (left) and Z-scores (right) for the C_2H_2 zinc finger protein, Egr1.

Figure S4. PBM data reproducibility. Panels **A-D** show that replicate arrays cluster together. We combined the 8-mer Z-scores from the two replicate arrays into a single file, with each replicate retained as a separate column and each 8-mer in a separate row. To minimize the impact of noise, we reduced this data structure to the 14,873 8-mers that have a Z-score of 6 or greater in at least one experiment, and set entries less than zero to zero. We clustered these data using Pearson correlations and hierarchical agglomerative linkage. Panel **A** shows the full clustering analysis. Panels **B**, **C**, and **D** show zoom-ins of the left, middle, and right of Panel **A**. Panel **E** shows the reproducibility of 8-mer E-scores (Pearson correlation coefficient r=0.65) and Z-scores (Pearson correlation coefficient r=0.85) for replicate PBMs for a single transcription factor (Esrra).

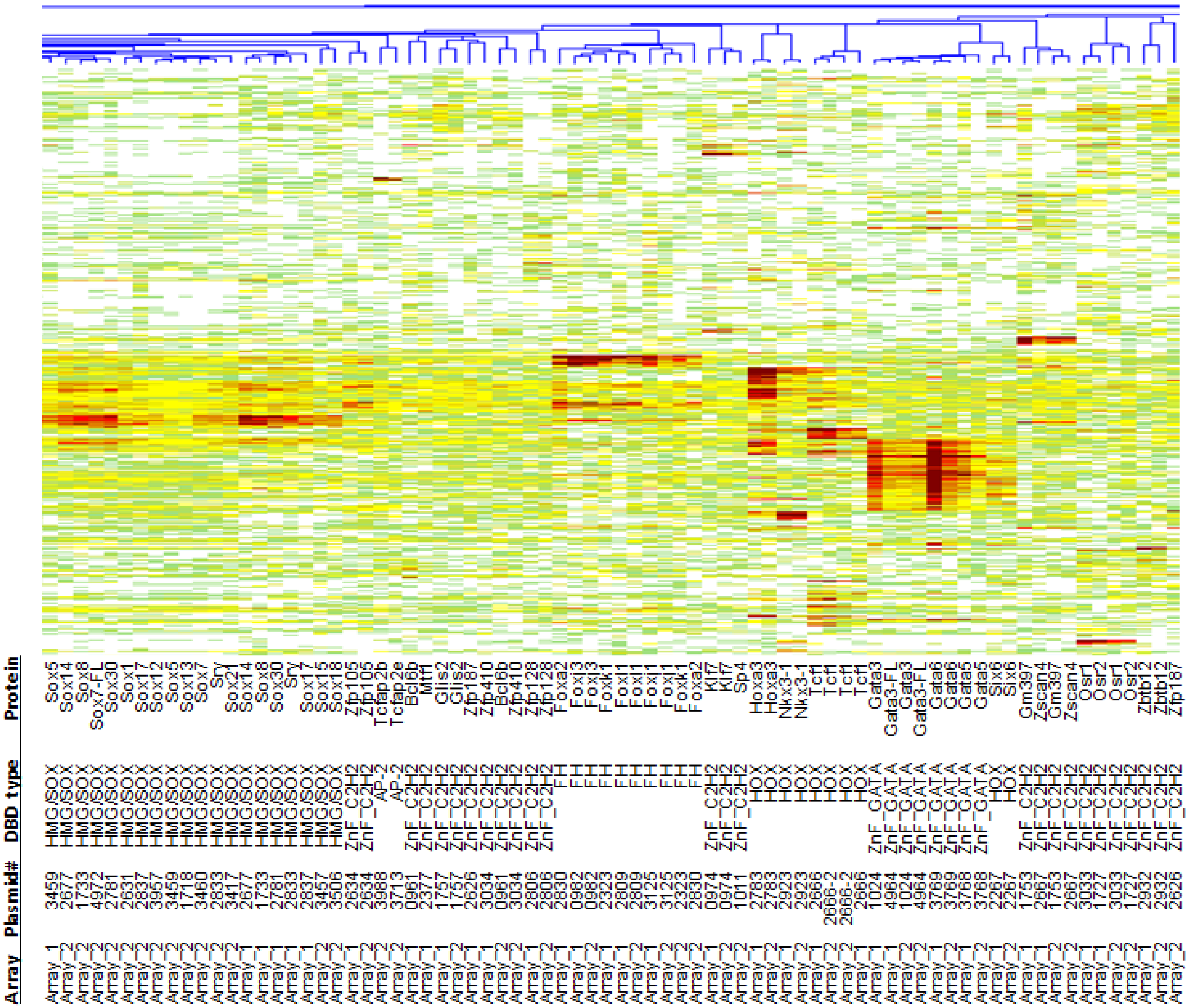
A.



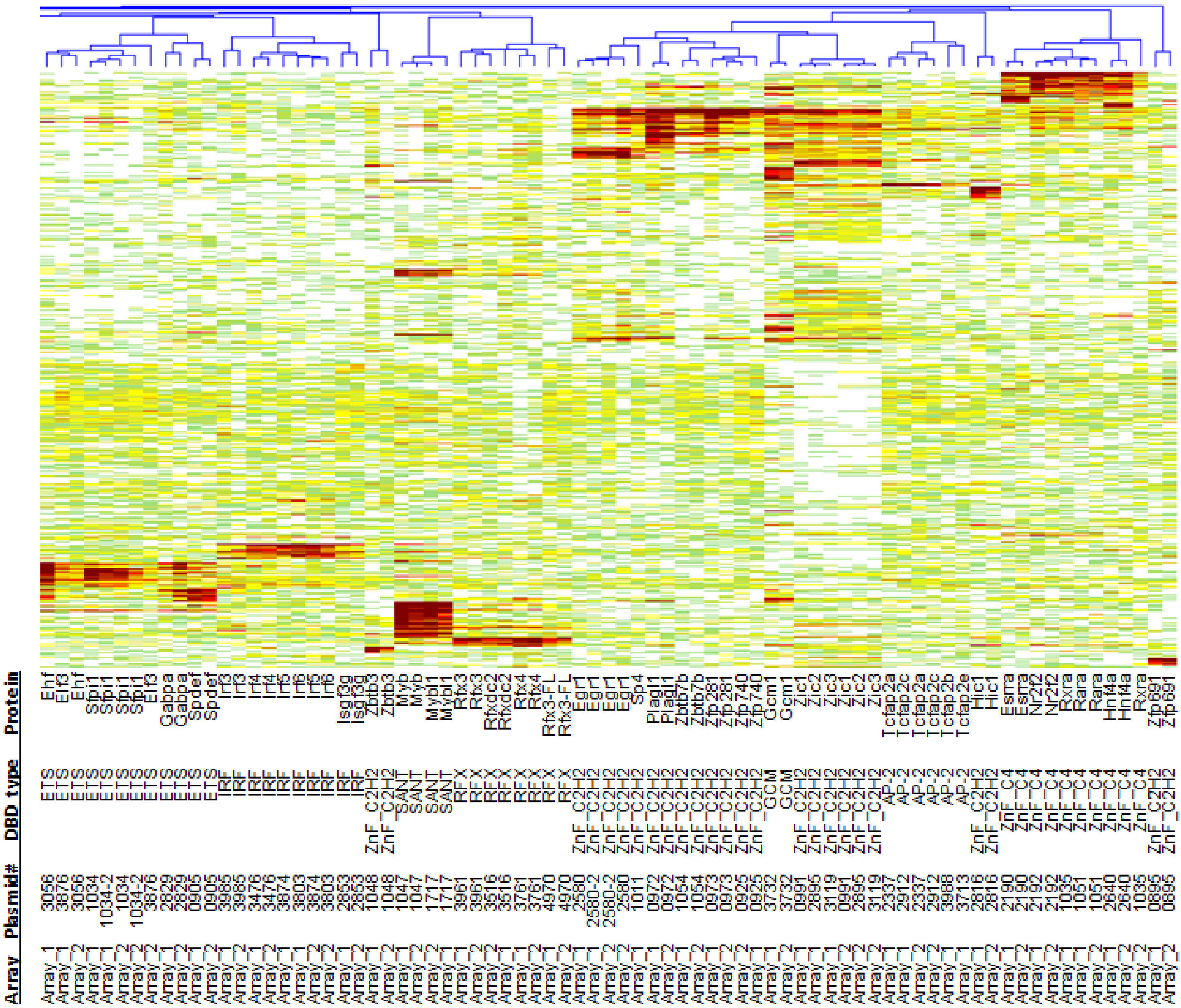
B.


Array Plasmid# DBD type Protein
Array_1 3752 E2F_TDP E2F3
Array_1 3752-2 E2F_TDP E2F3
Array_1 1022 E2F_TDP E2F2
Array_2 1022 E2F_TDP E2F2
Array_2 3752-2 E2F_TDP E2F3
Array_2 1022 E2F_TDP E2F2
Array_2 3752 E2F_TDP E2F3
Array_1 1022 E2F_TDP E2F2
Array_1 2858 ZnF_C2H2 Zfp161
Array_2 2858 ZnF_C2H2 Zfp161
Array_1 3805 DWA Smad3
Array_2 3805 DWA Smad3
Array_1 3026 BZIP Atf1
Array_1 0911 BZIP Jundm2
Array_2 0911 BZIP Jundm2
Array_2 3026 BZIP Atf1
Array_1 1745 SAND Gmeb1
Array_2 1745 SAND Gmeb1
Array_1 3520 SAND IRC900814
Array_1 2947 SAND Sp100
Array_2 2947 SAND Sp100
Array_2 3520 SAND IRC900814
Array_1 2654 HLH Ascl2
Array_1 3824 HLH Myf6
Array_2 2654 HLH Ascl2
Array_2 3824 HLH Myf6
Array_1 3865 HLH Tcfe2a
Array_2 3865 HLH Tcfe2a
Array_1 1274 HLH Bhlhb2
Array_1 4971 HLH Bhlhb2-FL
Array_2 1274 HLH Bhlhb2
Array_2 4971 HLH Bhlhb2-FL
Array_1 3863 HLH Max-FL
Array_1 3864 HLH Max
Array_2 3863 HLH Max-FL
Array_2 3864 HLH Max
Array_1 0921 TBOX Eomes
Array_2 0921 TBOX Eomes
Array_1 3875 BRIGHT Arid3a
Array_2 3875-2 BRIGHT Arid3a
Array_2 3875-2 BRIGHT Arid3a
Array_1 3875 BRIGHT Arid3a
Array_1 3770 BRIGHT Arid5a
Array_1 3509 MADS Srf
Array_1 pr78 TBP Tbp
Array_2 pr78 TBP Tbp
Array_2 3770 BRIGHT Arid5a
Array_2 3509 MADS Srf
Array_1 2914 BZIP Mafb
Array_1 3106 BZIP Mafk
Array_2 3106 BZIP Mafk
Array_2 2914 BZIP Mafb
Array_1 3504 HMG/SOX Lef1
Array_1 3461 HMG/SOX Tcf7l2
Array_1 3787 HMG/SOX Tcf3
Array_1 0950 HMG/SOX Tcf7
Array_2 3504 HMG/SOX Lef1
Array_2 3461 HMG/SOX Tcf7l2
Array_2 3787 HMG/SOX Tcf3
Array_2 0950 HMG/SOX Tcf7
Array_1 3753 HMG/SOX Bbx
Array_1 2241 HMG/SOX Hbp1
Array_2 2241 HMG/SOX Hbp1
Array_2 3753 HMG/SOX Bbx
Array_2 2377 HMG/SOX Mtf1
Array_1 2631 HMG/SOX Sox1
Array_1 2266 HMG/SOX Sox11
Array_1 2941 HMG/SOX Sox4
Array_1 3957 HMG/SOX Sox12
Array_2 2266 HMG/SOX Sox11
Array_2 2941 HMG/SOX Sox4
Array_1 1718 HMG/SOX Sox13
Array_1 3457 HMG/SOX Sox15
Array_1 3506 HMG/SOX Sox18
Array_1 3460 HMG/SOX Sox7
Array_1 3417 HMG/SOX Sox21
Array_1 4972 HMG/SOX Sox7-FL

C.

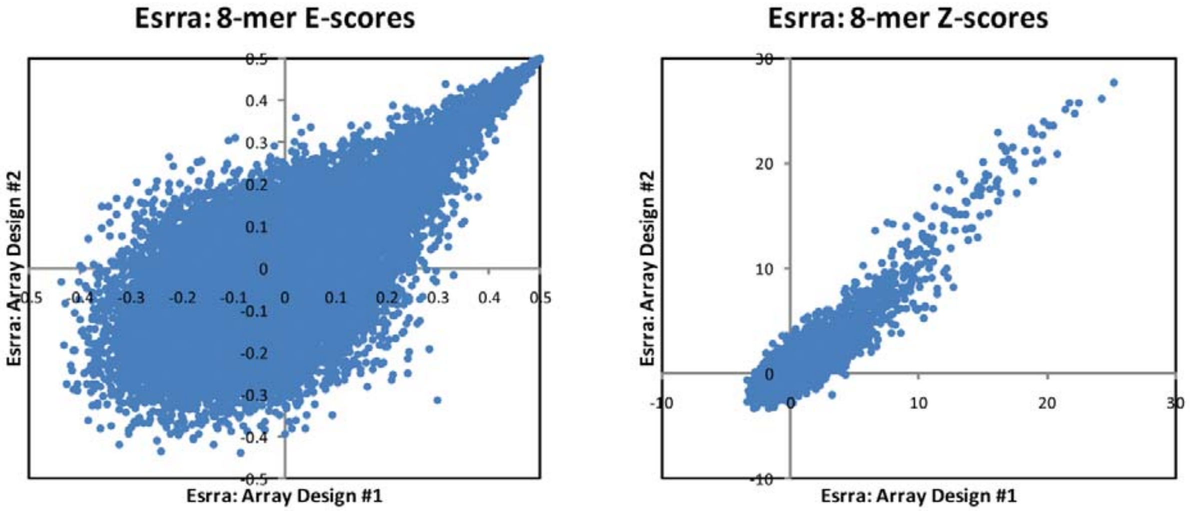

Array
Plasmid#
DBD type
Protein
Array_1 3459 HMG/SOX Sox5
Array_2 2677 HMG/SOX Sox14
Array_2 1733 HMG/SOX Sox8
Array_2 4972 HMG/SOX Sox7-FL
Array_2 2781 HMG/SOX Sox30
Array_2 2631 HMG/SOX Sox1
Array_2 2837 HMG/SOX Sox17
Array_2 3957 HMG/SOX Sox12
Array_2 3459 HMG/SOX Sox5
Array_2 1718 HMG/SOX Sox13
Array_2 3460 HMG/SOX Sox7
Array_2 2833 HMG/SOX Sry
Array_2 3417 HMG/SOX Sox21
Array_1 2677 HMG/SOX Sox14
Array_1 1733 HMG/SOX Sox8
Array_1 2781 HMG/SOX Sox30
Array_1 2833 HMG/SOX Sry
Array_1 2837 HMG/SOX Sox17
Array_2 3457 HMG/SOX Sox15
Array_2 3506 HMG/SOX Sox18
Array_1 2634 ZnF_C2H2 Zfp105
Array_2 2634 ZnF_C2H2 Zfp105
Array_2 3988 AP-2 Tcfap2b
Array_2 3713 AP-2 Tcfap2e
Array_1 0961 ZnF_C2H2 Bcl6b
Array_1 2377 ZnF_C2H2 Mtf1
Array_1 1757 ZnF_C2H2 Glis2
Array_2 1757 ZnF_C2H2 Glis2
Array_1 2626 ZnF_C2H2 Zfp187
Array_1 3034 ZnF_C2H2 Zfp410
Array_2 0961 ZnF_C2H2 Bcl6b
Array_2 3034 ZnF_C2H2 Zfp410
Array_1 2806 ZnF_C2H2 Zfp128
Array_2 2806 ZnF_C2H2 Zfp128
Array_1 2830 FH Foxa2
Array_1 0982 FH Foxj3
Array_2 0982 FH Foxj3
Array_1 2323 FH Foxk1
Array_1 2809 FH Foxj1
Array_2 2809 FH Foxj1
Array_1 3125 FH Foxj1
Array_2 3125 FH Foxj1
Array_2 2323 FH Foxk1
Array_2 2830 FH Foxa2
Array_1 0974 ZnF_C2H2 Klf7
Array_2 0974 ZnF_C2H2 Klf7
Array_2 1011 ZnF_C2H2 Sp4
Array_1 2783 HOX Hoxa3
Array_2 2783 HOX Hoxa3
Array_1 2923 HOX Nkx3-1
Array_2 2923 HOX Nkx3-1
Array_1 2666 HOX Tcf1
Array_2 2666-2 HOX Tcf1
Array_2 2666-2 HOX Tcf1
Array_1 2666 HOX Tcf1
Array_1 1024 ZnF_GATA Gata3
Array_1 4964 ZnF_GATA Gata3-FL
Array_2 1024 ZnF_GATA Gata3
Array_2 4964 ZnF_GATA Gata3-FL
Array_1 3769 ZnF_GATA Gata6
Array_2 3769 ZnF_GATA Gata6
Array_1 3768 ZnF_GATA Gata5
Array_2 3768 ZnF_GATA Gata5
Array_1 2267 HOX Six6
Array_2 2267 HOX Six6
Array_1 1753 ZnF_C2H2 Gm397
Array_1 2667 ZnF_C2H2 Zscan4
Array_2 1753 ZnF_C2H2 Gm397
Array_2 2667 ZnF_C2H2 Zscan4
Array_1 3033 ZnF_C2H2 Osr1
Array_1 1727 ZnF_C2H2 Osr2
Array_2 3033 ZnF_C2H2 Osr1
Array_2 1727 ZnF_C2H2 Osr2
Array_1 2932 ZnF_C2H2 Zbtb12
Array_2 2932 ZnF_C2H2 Zbtb12
Array_2 2626 ZnF_C2H2 Zfp187

D.



Array
Plasmid#
DBD type
Protein
Array_1 3056 ETS Ehf
Array_1 3876 ETS Elf3
Array_2 3056 ETS Ehf
Array_1 1034 ETS Spi1
Array_1 1034-2 ETS Spi1
Array_2 1034 ETS Spi1
Array_2 1034-2 ETS Spi1
Array_2 3876 ETS Elf3
Array_1 2829 ETS Gabpa
Array_2 2829 ETS Gabpa
Array_1 0905 ETS Spdef
Array_2 0905 ETS Spdef
Array_1 3985 IRF Irf3
Array_2 3985 IRF Irf3
Array_1 3476 IRF Irf4
Array_2 3476 IRF Irf4
Array_1 3874 IRF Irf5
Array_1 3803 IRF Irf6
Array_2 3874 IRF Irf5
Array_2 3803 IRF Irf6
Array_1 2853 IRF Isgf3g
Array_2 2853 IRF Isgf3g
Array_1 1048 ZnF_C2H2 Zbtb3
Array_2 1048 ZnF_C2H2 Zbtb3
Array_1 1047 SANT Myb
Array_2 1047 SANT Myb
Array_1 1717 SANT Mybl1
Array_2 1717 SANT Mybl1
Array_1 3961 RFX Rfx3
Array_2 3961 RFX Rfx3
Array_1 3516 RFX Rfxdc2
Array_2 3516 RFX Rfxdc2
Array_1 3761 RFX Rfx4
Array_2 3761 RFX Rfx4
Array_1 4970 RFX Rfx3-FL
Array_2 4970 RFX Rfx3-FL
Array_1 2580 ZnF_C2H2 Egr1
Array_1 2580-2 ZnF_C2H2 Egr1
Array_2 2580-2 ZnF_C2H2 Egr1
Array_2 2580 ZnF_C2H2 Egr1
Array_1 1011 ZnF_C2H2 Sp4
Array_1 0972 ZnF_C2H2 Plagl1
Array_2 0972 ZnF_C2H2 Plagl1
Array_1 1054 ZnF_C2H2 Zbtb7b
Array_2 1054 ZnF_C2H2 Zbtb7b
Array_1 0973 ZnF_C2H2 Zfp281
Array_2 0973 ZnF_C2H2 Zfp281
Array_1 0925 ZnF_C2H2 Zfp740
Array_2 0925 ZnF_C2H2 Zfp740
Array_1 3732 GCM Gcm1
Array_2 3732 GCM Gcm1
Array_1 0991 ZnF_C2H2 Zic1
Array_1 2895 ZnF_C2H2 Zic2
Array_1 3119 ZnF_C2H2 Zic3
Array_2 0991 ZnF_C2H2 Zic1
Array_2 2895 ZnF_C2H2 Zic2
Array_2 3119 ZnF_C2H2 Zic3
Array_1 2337 AP-2 Tcfap2a
Array_1 2912 AP-2 Tcfap2c
Array_2 2337 AP-2 Tcfap2a
Array_2 2912 AP-2 Tcfap2c
Array_1 3988 AP-2 Tcfap2b
Array_1 3713 AP-2 Tcfap2e
Array_1 2816 ZnF_C2H2 Hic1
Array_2 2816 ZnF_C2H2 Hic1
Array_1 2190 ZnF_C4 Esrra
Array_2 2190 ZnF_C4 Esrra
Array_1 2192 ZnF_C4 Nr2f2
Array_2 2192 ZnF_C4 Nr2f2
Array_1 1035 ZnF_C4 Rxra
Array_1 1051 ZnF_C4 Rara
Array_2 1051 ZnF_C4 Rara
Array_1 2640 ZnF_C4 Hnf4a
Array_2 2640 ZnF_C4 Hnf4a
Array_2 1035 ZnF_C4 Rxra
Array_1 0895 ZnF_C2H2 Zfp691
Array_2 0895 ZnF_C2H2 Zfp691

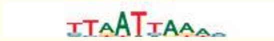
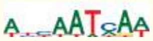
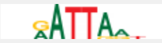
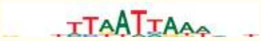
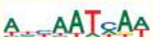
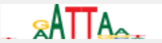
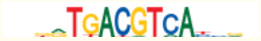
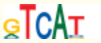
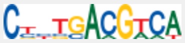
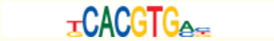
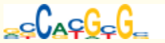
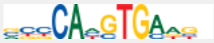
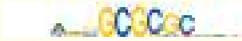
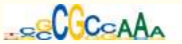
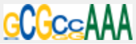
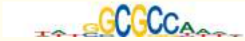
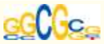
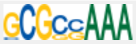
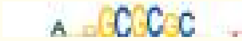
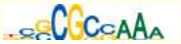
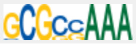
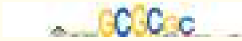
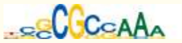
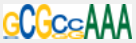
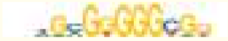
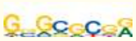
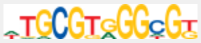
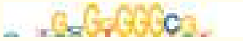
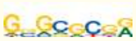
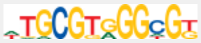
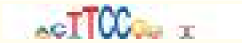
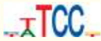
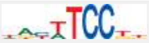
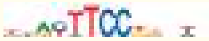
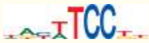
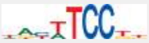
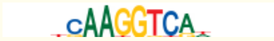
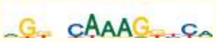
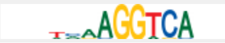
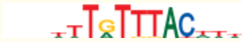
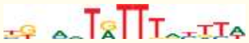
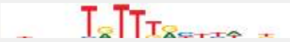
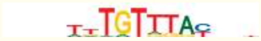
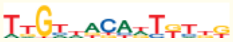
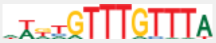
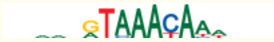
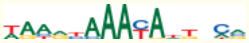
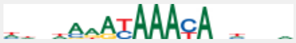
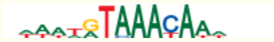
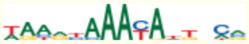
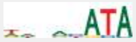
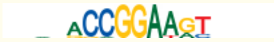
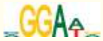
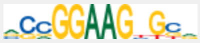
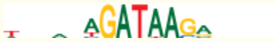
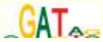
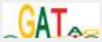
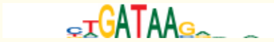
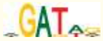
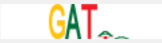
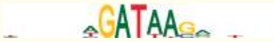
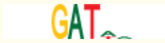
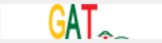
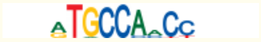
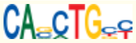
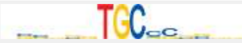
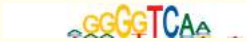
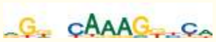
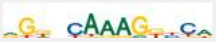
E.

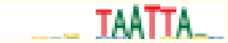
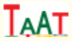
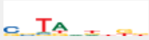
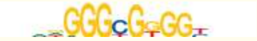
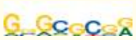
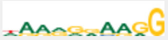
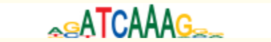
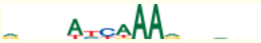
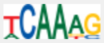
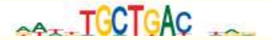
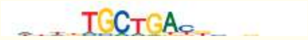
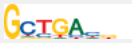
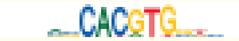
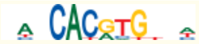
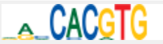
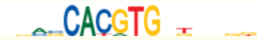
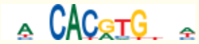
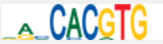
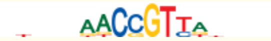
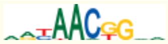
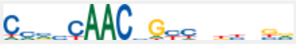
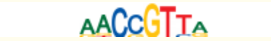
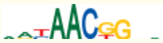
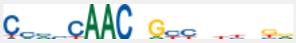
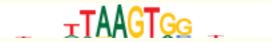
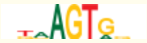
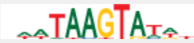
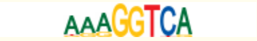
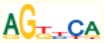
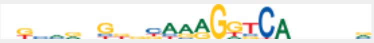
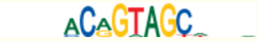
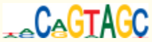
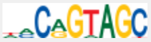
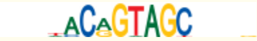
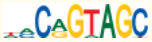
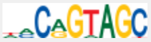
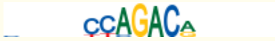
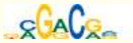
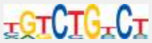
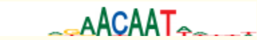
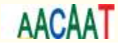
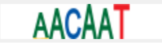
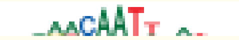
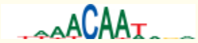
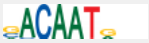
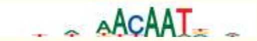
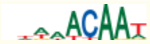
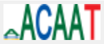
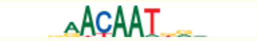
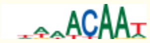
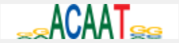
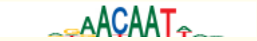
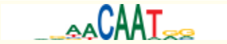
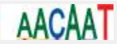
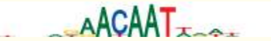
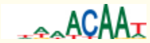
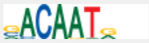
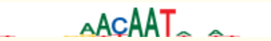
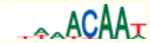
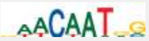
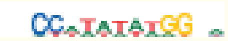
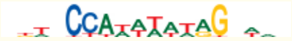
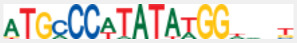
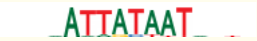
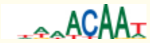
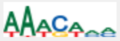
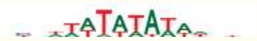
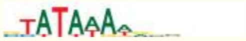
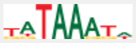


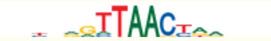
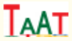
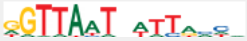
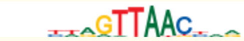
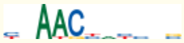
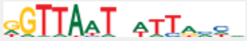
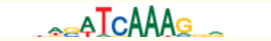
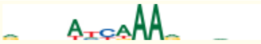
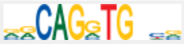
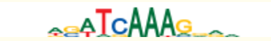
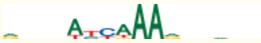
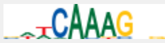
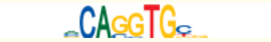
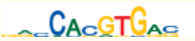
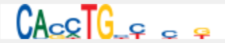
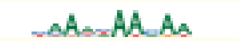
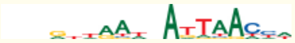
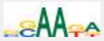
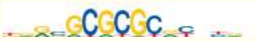
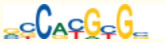
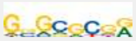
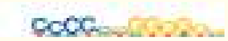
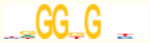
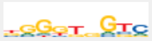
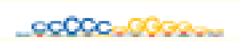
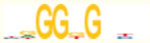
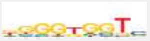
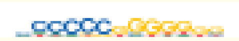
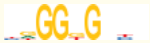
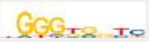
Esrra: 8-mer E-scores

Esrra: 8-mer Z-scores

Figure S5: **Agreement of PBM *k*-mer data with prior motif data, in general.** Comparisons were performed as described in **Materials and Methods**. 44 of the 50 proteins (88%) in rings 1, 2, or 3 had their top AUC matches to members of their structural families; 5 of these 44 proteins had their top AUC match to the expected protein (the exact match, paralog, or ortholog referenced by the ring system). Full comparison results (AUC ≥ 0.8 and $Q \leq 0.01$) are provided in **Table S3**.

PBM TF	Top Lever Match	AUC	Same Struct Class?	Closest Previously Annotated Match	Ring	AUC
Arid3a_3875.1	Pbx-1 (V$PBX1_01)	0.965695	No	dri (I$DRI_01)	ring 3	0.920001
						
Arid3a_3875.2	Pbx-1 (V$PBX1_01)	0.978981	No	dri (I$DRI_01)	ring 3	0.934148
						
Atf1_3026.3	TCF11-MafG (MA0089)	0.962233	No	ATF1 (V$ATF1_Q6)	ring 1	0.780575
						
Bhlhb2_1274.3	c-Myc:Max (V$MYCMAX_B)	0.869423	Yes (HLH)	DEC (V$DEC_Q1)	ring 3	0.648959
						
E2F2_1022.2	E2F (V$E2F_Q4_01)	0.961466	Yes (E2F family)	E2f1 (MA0024)	ring 3	0.895325
						
E2F2_1022.4	E2F (V$E2F_Q2)	0.966291	Yes (E2F family)	E2f1 (MA0024)	ring 3	0.901104
						
E2F3_3752.1	E2F (V$E2F_Q4_01)	0.959812	Yes (E2F family)	E2f1 (MA0024)	ring 3	0.893595
						
E2F3_3752.2	E2F (V$E2F_Q4_01)	0.960145	Yes (E2F family)	E2F1 (MA0024)	ring 3	0.890967
						
Egr1_2580.1	ZF5 (V$ZF5_01)	0.939128	Yes (Znf_C2H2)	Egr-1 (V$EGR1_01)	ring 1	0.642253
						
Egr1_2580.2	ZF5 (V$ZF5_01)	0.936849	Yes (Znf_C2H2)	Egr-1 (V$EGR1_01)	ring 1	0.639174
						
Ehf_3056.2	ETS1 (MA0098)	0.988278	Yes (ETS)	ELF5 (MA0136)	ring 2	0.984428
						
Elf3_3876.1	ELF5 (MA0136)	0.97288	Yes (ETS)	ELF5 (MA0136)	ring 2	0.97288
						
Esrra_2190.2	HNF4A (MA0114)	0.89013	Yes (ZnF_C4)	ERR alpha (V$ERR1_Q2)	ring 1	0.682352
						
Foxa2_2830.2	HNF3beta (V$HNF3B_01)	0.959604	Yes (Forkhead)	HNF3 (V$HNF3_Q6_01)	ring 1	0.947694
						
Foxj1_3125.2	DMRT7 (V$DMRT7_01)	0.961358	No	FOXJ1 (V$HFH4_01)	ring 1	0.858688
						
Foxj3_0982.2	HNF3beta (V$HNF3B_01)	0.963847	Yes (Forkhead)	FOXJ2 (V$FOXJ2_01)	ring 2	0.905563
						
Foxl1_2809.2	HNF3beta (V$HNF3B_01)	0.979563	Yes (Forkhead)	FOXL1 (MA0033)	ring 3	0.889422
						
Gabpa_2829.2	ETS1 (MA0098)	0.984335	Yes (ETS)	GABP (V$GABP_B)	ring 1	0.656266
						
Gata3_1024.3	GATA3 (MA0037)	0.95315	Yes (ZnF_Gata)	GATA3 (MA0037)	ring 3	0.95315
						
Gata5_3768.1	GATA3 (MA0037)	0.985313	Yes (ZnF_Gata)	GATA-6 (V$GATA6_01)	ring 2	0.935301
						
Gata6_3769.1	GATA-6 (V$GATA6_01)	0.937566	Yes (ZnF_Gata)	GATA-6 (V$GATA6_01)	ring 1	0.937566
						
Hic1_2816.2	myogenin (V$MYOGENIN_Q6)	0.833216	No	HIC1 (V$HIC1_02)	ring 3	0.68262
						
Hnf4a_2640.2	HNF4A (MA0114)	0.918195	Yes (ZnF_C4)	HNF4A (MA0114)	ring 1	0.918195
						

Hoxa3_2783.2	Ubx (MA0094)	0.986339	Yes (Homeodomain)	HOXA3 (V$HOXA3_01)	ring 1	0.736896
						
Klf7_0974.2	ZF5 (V$ZF5_01)	0.93137	Yes (Znf_C2H2)	Klf4 (MA0039)	ring 2	0.682812
						
Lef1_3504.1	TCF (I$TCF_Q6)	0.887154	Yes (HMG)	LEF1 (V$LEF1_Q2)	ring 1	0.761938
						
Mafb_2914.2	c-Maf (V$CMAF_01)	0.934102	Yes (bZIP)	Mafb (MA0117)	ring 3	0.58046
						
Max_3863.1	c-Myc:Max (V$MYCMAX_02)	0.884495	Yes (HLH)	MAX (MA0058)	ring 3	0.621124
						
Max_3864.1	c-Myc:Max (V$MYCMAX_02)	0.931824	Yes (HLH)	MAX (MA0058)	ring 3	0.605609
						
Myb_1047.3	v-Myb (V$VMYB_01)	0.910701	Yes (SANT)	c-Myb (V$CMYB_01)	ring 2	0.795148
						
Mybl1_1717.2	v-Myb (V$VMYB_01)	0.920978	Yes (SANT)	c-Myb (V$CMYB_01)	ring 2	0.7907
						
Nkx3-1_2923.2	Bapx1 (MA0122)	0.918855	Yes (Homeodomain)	Nkx3-1 (V$NKX3A_01)	ring 1	0.749729
						
Nr2f2_2192.2	HNF4 (V$HNF4_Q6_02)	0.917819	Yes (ZnF_C4)	COUPTF (V$COUPTF_Q6)	ring 1	0.727204
						
Osr1_3033.2	Odd-skipped (Wolfe et al., 2005)	0.947458	Yes (Znf_C2H2)	Odd-skipped (Wolfe et al., 2005)	ring 3	0.947458
						
Osr2_1727.2	Odd-skipped (Wolfe et al., 2005)	0.974839	Yes (Znf_C2H2)	Odd-skipped (Wolfe et al., 2005)	ring 3	0.974839
						
Smad3_3805.1	MAD (I$MAD_Q6)	0.802327	Yes (MAD)	SMAD3 (V$SMAD3_Q6)	ring 1	0.757946
						
Sox13_1718.2	Sox5 (MA0087)	0.980609	Yes (HMG)	SOX5 (V$SOX5_01)	ring 2	0.975989
						
Sox17_2837.2	SRY (V$SRY_02)	0.946124	Yes (HMG)	Sox17 (MA0078)	ring 1	0.84448
						
Sox18_3506.1	SRY (MA0084)	0.968292	Yes (HMG)	SOX17 (V$SOX17_01)	ring 2	0.958906
						
Sox30_2781.2	SRY (MA0084)	0.948422	Yes (HMG)	Sox30 (Osaki et al., 1999)	ring 1	0.753482
						
Sox5_3459.1	SOX9 (V$SOX9_B1)	0.972955	Yes (HMG)	Sox5 (MA0087)	ring 1	0.955712
						
Sox7_3460.1	SRY (MA0084)	0.962653	Yes (HMG)	Sox17 (MA0078)	ring 2	0.887095
						
Sox8_1733.2	SRY (MA0084)	0.946788	Yes (HMG)	SOX9 (MA0077)	ring 3	0.92127
						
Srf_3509.1	AGL3 (P$AGL3_01)	0.99214	Yes (MAD)	SRF (V$SRF_01)	ring 1	0.82962
						
Sry_2833.2	SRY (MA0084)	0.970784	Yes (HMG)	SRY (V$SRY_01)	ring 1	0.871343
						
Tbp_pr781.1	TATA (V$TATA_01)	0.979028	Yes (TBP)	TBP (V$TBP_01)	ring 1	0.951961
						
Tcf1_2666.2	Ubx (MA0094)	0.893147	Yes (Homeodomain)	HNF1 (V$HNF1_01)	ring 3	0.834492

						
Tcf1_2666.3	C1 (P$C1_Q2)	0.917045	Yes (Homeodomain)	HNF1 (V$HNF1_01)	ring 3	0.854438
						
Tcf3_3787.1	TCF (I$TCF_Q6)	0.950095	Yes (Homeodomain)	E12 (V$E12_Q6)	ring 3	0.266878
						
Tcf7_0950.2	TCF (I$TCF_Q6)	0.955304	Yes (Homeodomain)	LEF1 (V$LEF1_Q2_01)	ring 1	0.750827
						
Tcfe2a_3865.1	USF (V$USF_Q6_01)	0.885149	Yes (bHLH)	E2A (V$E2A_Q2)	ring 3	0.711049
						
Zfp105_2634.2	HNF1 (V$HNF1_Q6)	0.982651	No	Znf35 (Pengue et al., 1993)	ring 3	0.57543
						
Zfp161_2858.2	c-Myc:Max (V$MYCMAX_B)	0.915214	No	ZF5 (V$ZF5_01)	ring 1	0.88187
						
Zic1_0991.2	Macho-1 (MA0118)	0.898683	Yes (ZnF_C2H2)	Zic1 (V$ZIC1_01)	ring 1	0.76883
						
Zic2_2895.2	Macho-1 (MA0118)	0.926914	Yes (ZnF_C2H2)	Zic2 (V$ZIC2_01)	ring 1	0.686375
						
Zic3_3119.2	Macho-1 (MA0118)	0.899988	Yes (ZnF_C2H2)	Zic3 (V$ZIC3_01)	ring 1	0.792524
						

(A)

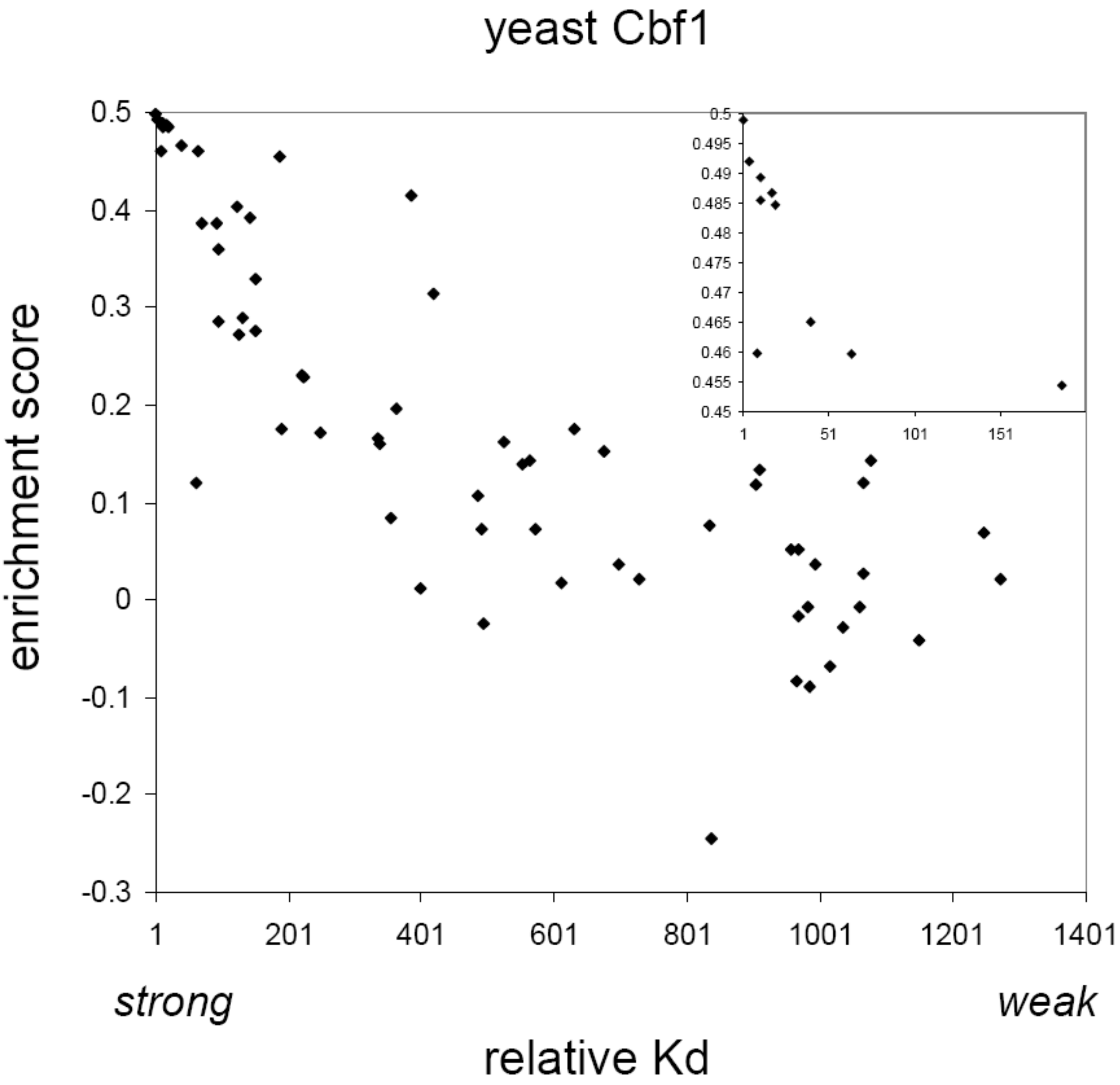
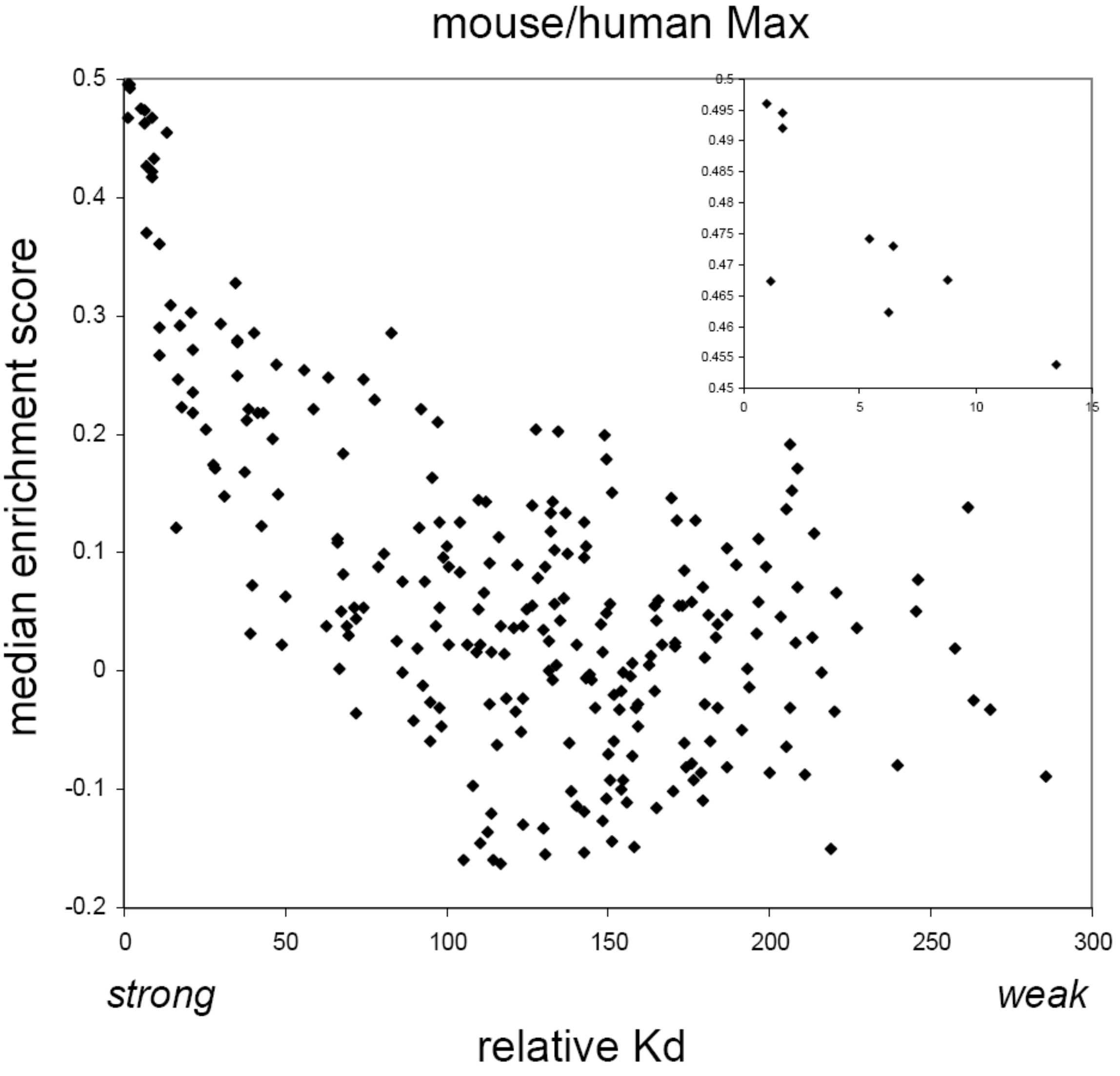


Figure S6. Comparison of PBM data versus K$_d$ data. *k*-mers with higher median signal intensity are of higher DNA binding affinity, as shown in PBM enrichment score versus relative K$_d$ plots for **(A)** yeast Cbf1(data shown for 8-mers analyzed by Maerkl and Quake, *Science* (2007)) and **(B) (next page)** murine/human Max (data shown for median of all 8-mers that contain each 7-mer analyzed by Maerkl and Quake, *Science* (2007)). Yeast Cbf1 PBM data are from Berger *et al.*, *Nature Biotechnology* (2006). Max PBM data are for murine Max from this paper. K$_d$ data were calculated from ddG data from Maerkl and Quake, *Science* (2007), and correspond to affinities for the highest affinity sequences, of 16.6 nM for Cbf1 and 67.0 nM for human MAX isoform A. The lower limit of detection of the MITOMI assays was ~18 uM, as reported in that study. Note: Maerkl and Quake, *Science* (2007) examined human Max protein. Additional comparisons of PBM versus K$_d$ data were shown previously in Berger *et al.*, *Nature Biotechnology* (2006) for Egr1 (Zif268).

(B)


mouse/human Max
median enrichment score
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0
50
100
150
200
250
300
strong
weak
relative Kd
0.5
0.495
0.49
0.485
0.48
0.475
0.47
0.465
0.46
0.455
0.45
0
5
10
15

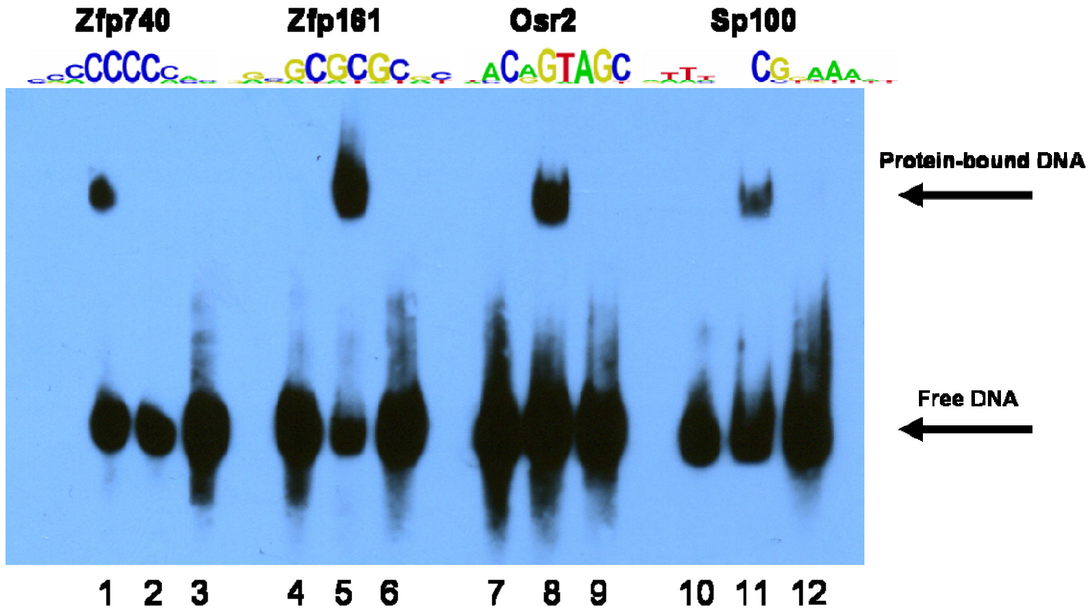


Figure S7. Confirmation of PBM-derived motifs by EMSAs for three newly characterized proteins (Zfp740, Osr2, Sp100) and one recently characterized protein (Zfp161, also known as ZF5 (Orlov *et al*., *FEBS J, 2007*)). Electrophoretic mobility shift assays were performed to verify select motifs which were determined by PBM. Lane 1: Zfp740 protein + C_8 probe; lane 2: Zfp740 protein + $(GC)_5$ probe; lane 3: Zfp740 protein + $(GGCC)_2$ probe; lane 4: Zfp161 protein + C_8 probe; lane 5: Zfp161 protein + $(GC)_5$ probe; lane 63: Zfp161 protein + $(GGCC)_2$ probe; lane 7: Osr2 positive probe; lane 8: Osr2 protein + Osr2 positive probe; lane 9: Osr2 protein + Sp100 positive probe; lane 10: Sp100 positive probe; lane 11: Sp100 protein + Sp100 positive probe; lane 12: Sp100 protein + Osr2 positive probe. Lanes 1-6 were designed to examine the specificity of the protein to its PBM-derived motif by testing each protein with two other probe sequences of similar GC content (Zfp740 positive control probe containing C_8, Zfp161 positive control probe containing $(GC)_5$, or probe containing $(GGCC)_2$); see **Materials and Methods** for the complete probe sequences. Lanes 7-12 validate binding by testing the protein both to its PBM-derived motif and to a probe designed to test a different protein, as a negative control.

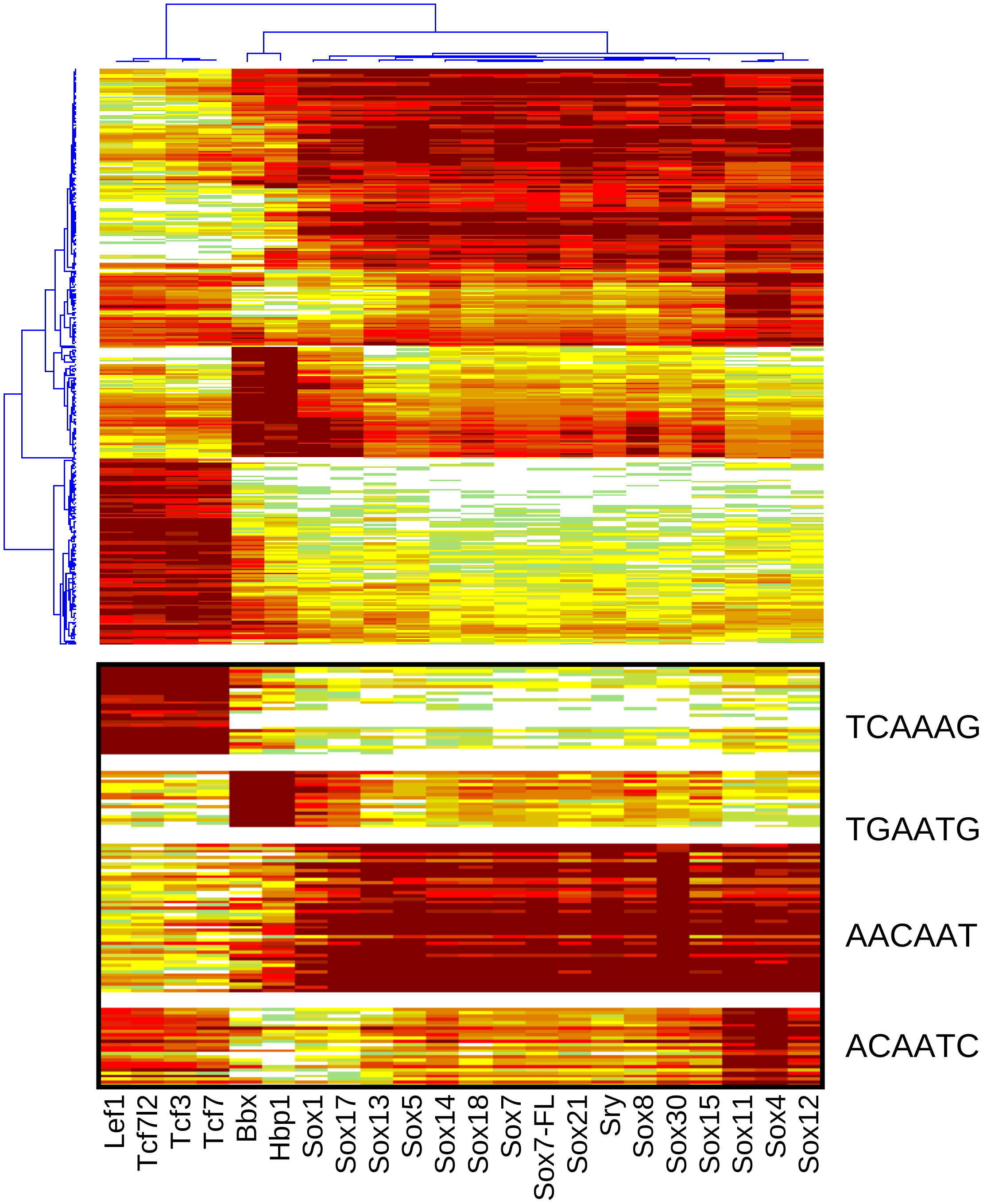
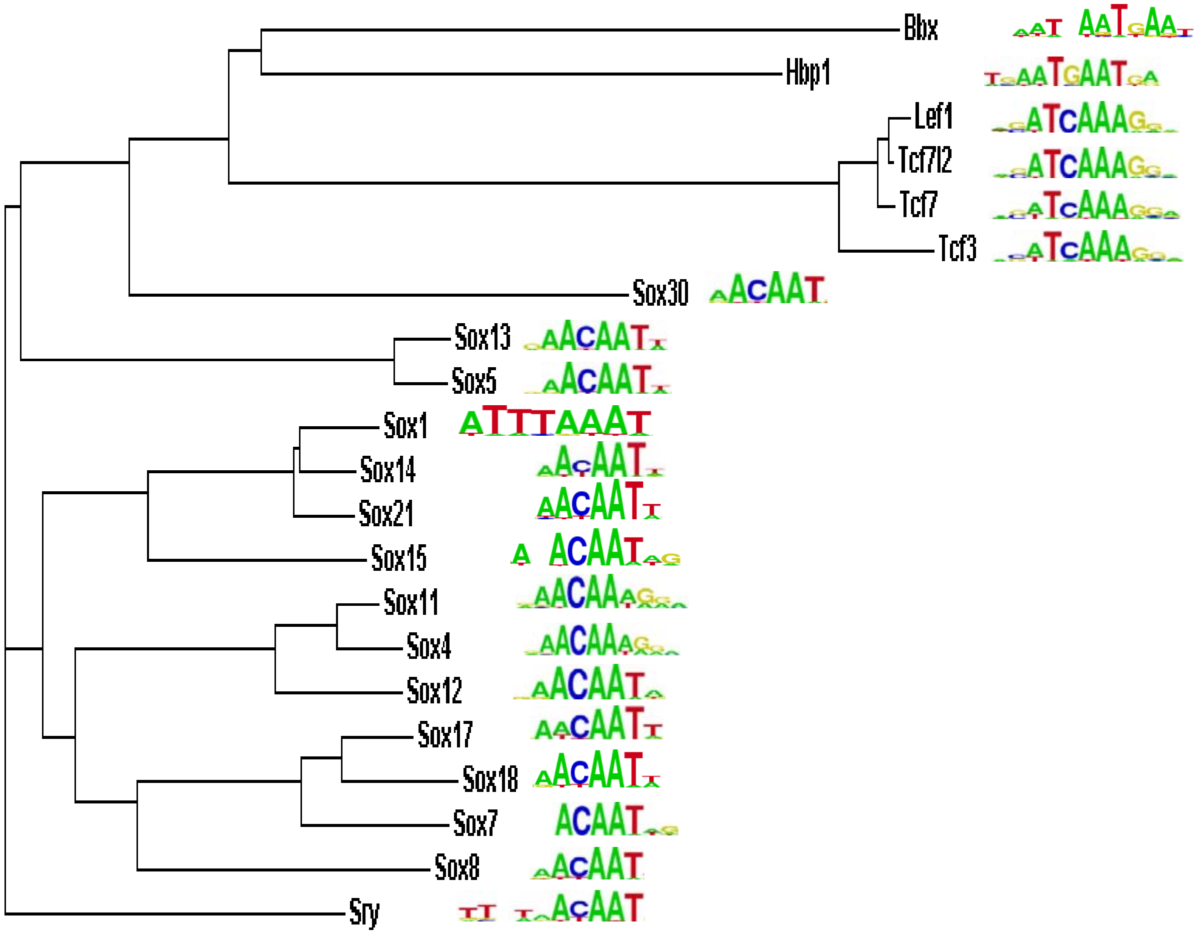


Figure S8. (A) HMG/SOX DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 310 8-mers x 21 HMG/SOX DNA-binding domains (with Sox7 as both DBD and FL). The 310 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the DBDs shown. Each of the 310 8-mers was then given a rank score (between 1 and 310) within each column, and the ranks were analyzed here, in order to compensate for any overall differences in magnitude of the E-scores. ***Bottom,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Next page***, Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.


Bbx
AAT AATGAAT
Hbp1
TGAATGAATGA
Lef1
ATCAAAG
Tcf7l2
ATCAAAG
Tcf7
ATCAAAG
Tcf3
ATCAAAG
Sox30
AACAAT
Sox13
AACAAT
Sox5
AACAAT
Sox1
ATTTAAAT
Sox14
AACAAT
Sox21
AACAAT
Sox15
A ACAAT
Sox11
AACAAAG
Sox4
AACAAAG
Sox12
AACAAT
Sox17
AACAAT
Sox18
AACAAT
Sox7
ACAAT
Sox8
AACAAT
Sry
TT TAACAAT

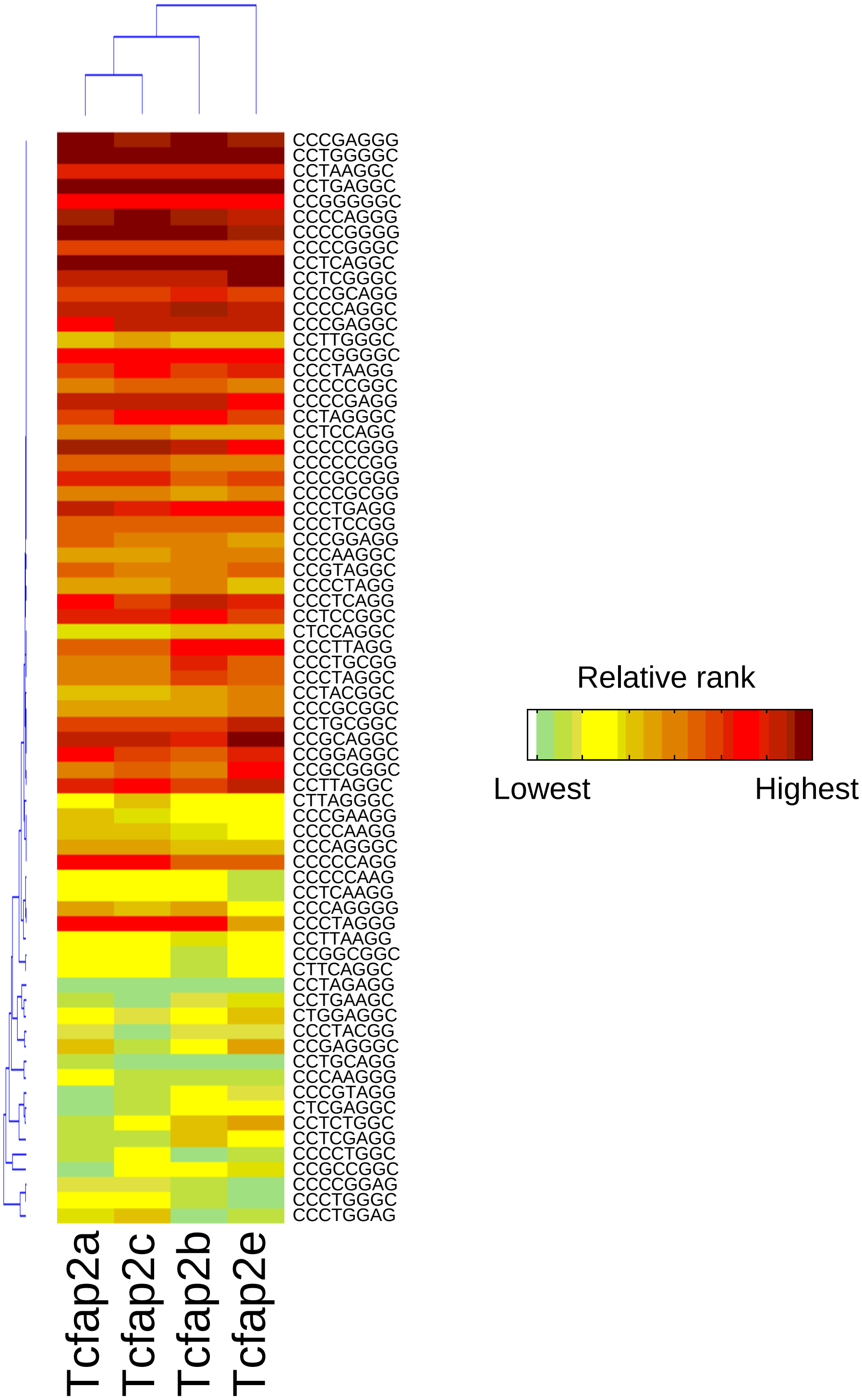


Figure S8. (B) AP-2 DNA-binding domains. 2-D Hierarchical agglomerative clustering analysis of relative ranks for 71 8-mers x 4 AP-2 DNA-binding domains. The 71 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 71 8-mers was then given a rank score (between 1 and 71) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores.

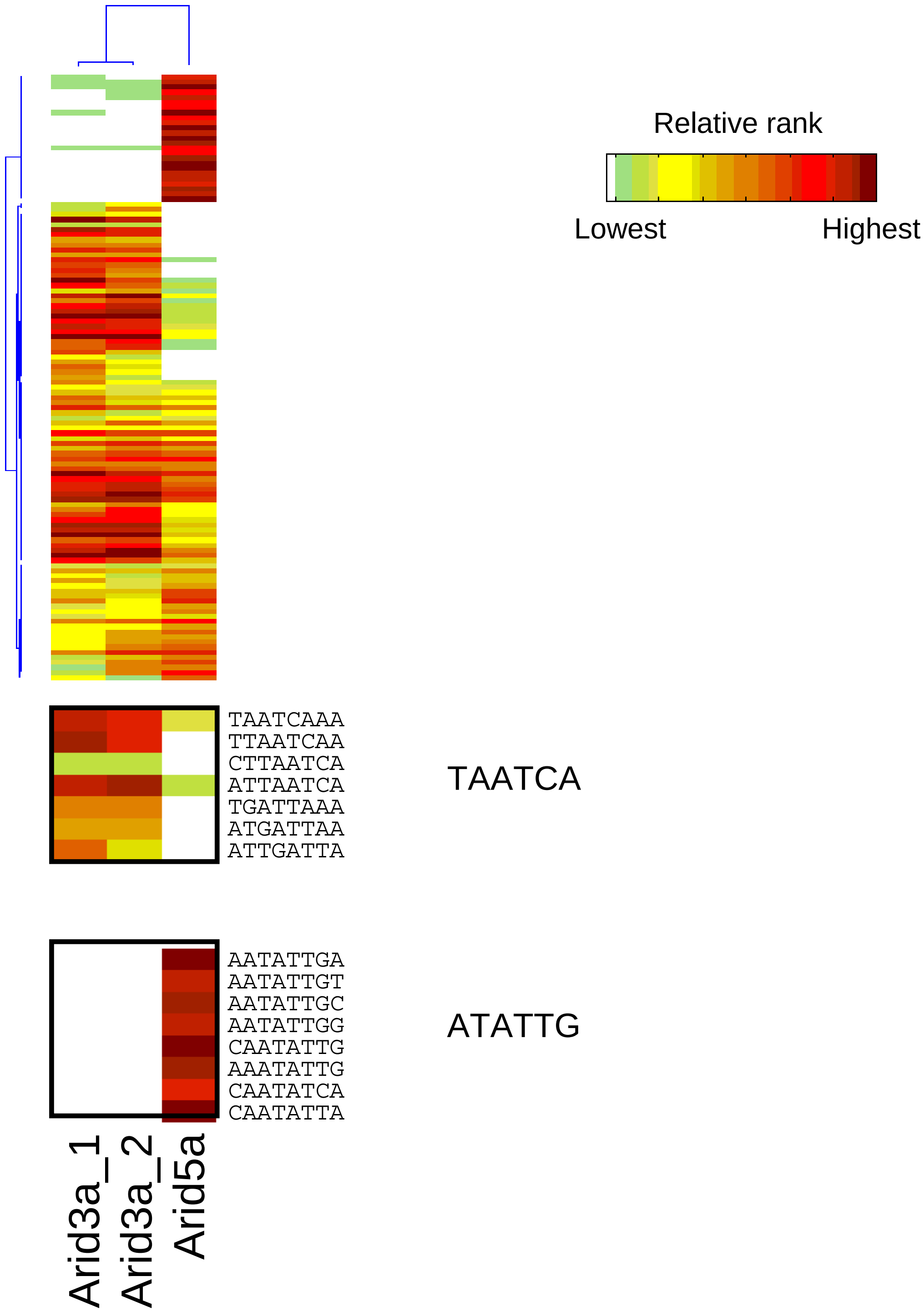


Figure S8. (C) ARID/BRIGHT DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 119 8-mers x 3 ARID/BRIGHT DNA-binding domains. The 119 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 119 8-mers was then given a rank score (between 1 and 119) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Bottom,*** 6mer sequences that are preferred within the 8-mers shown in the top panel.

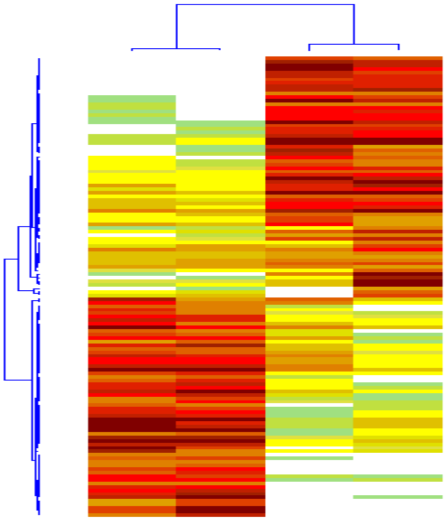
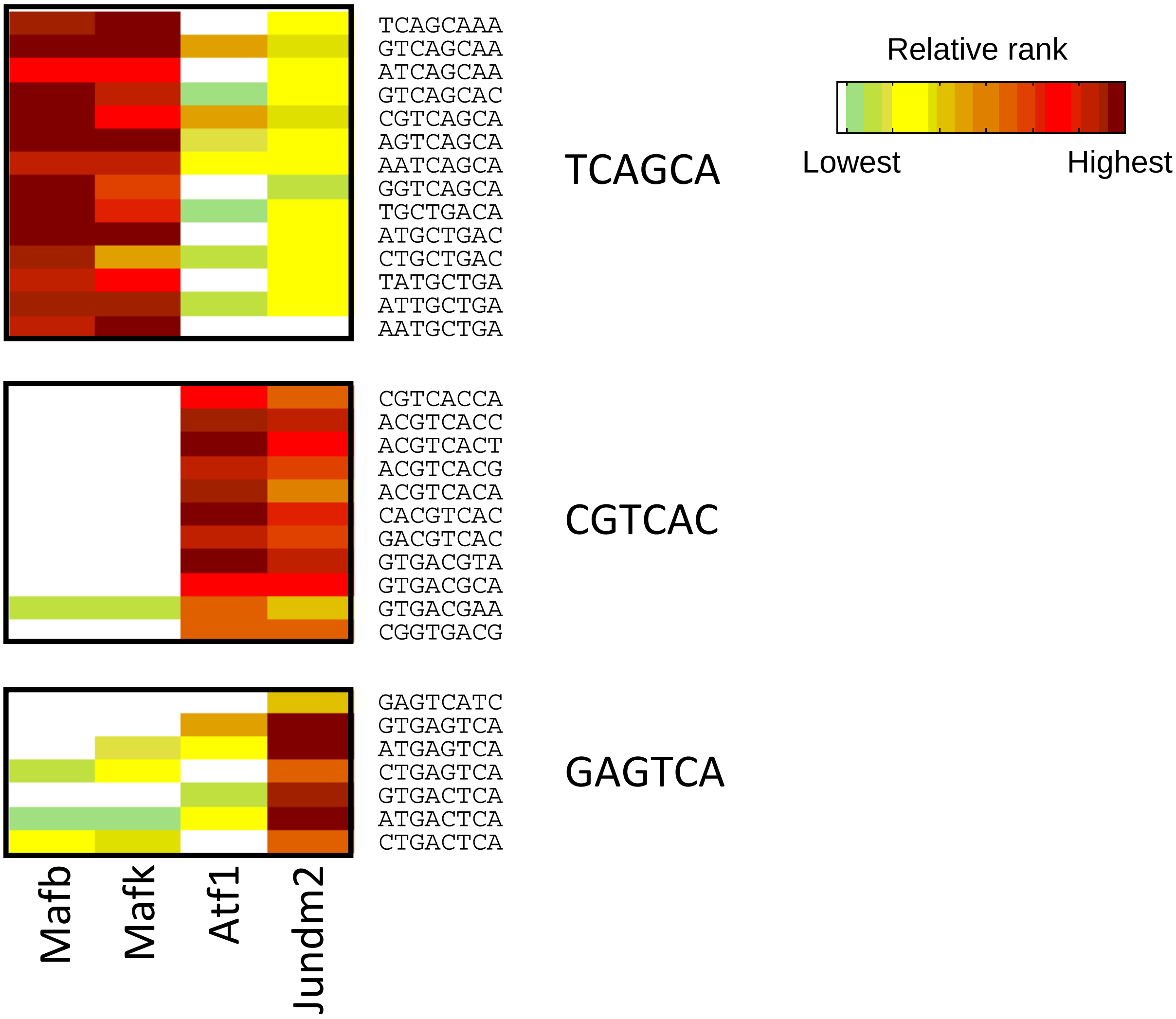


Figure S8. (D) BZIP DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 130 8-mers x 4 BZIP DNA-binding domains. The 130 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 130 8-mers was then given a rank score (between 1 and 130) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.



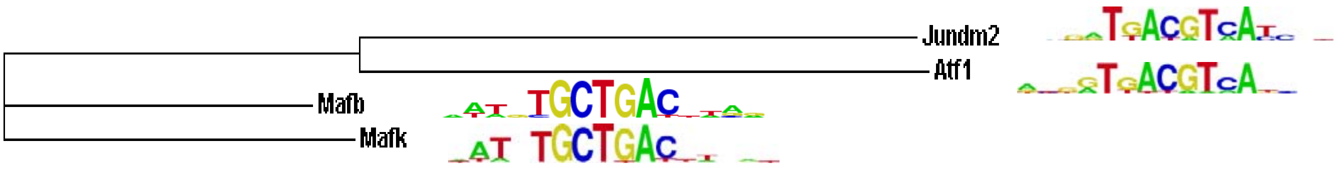
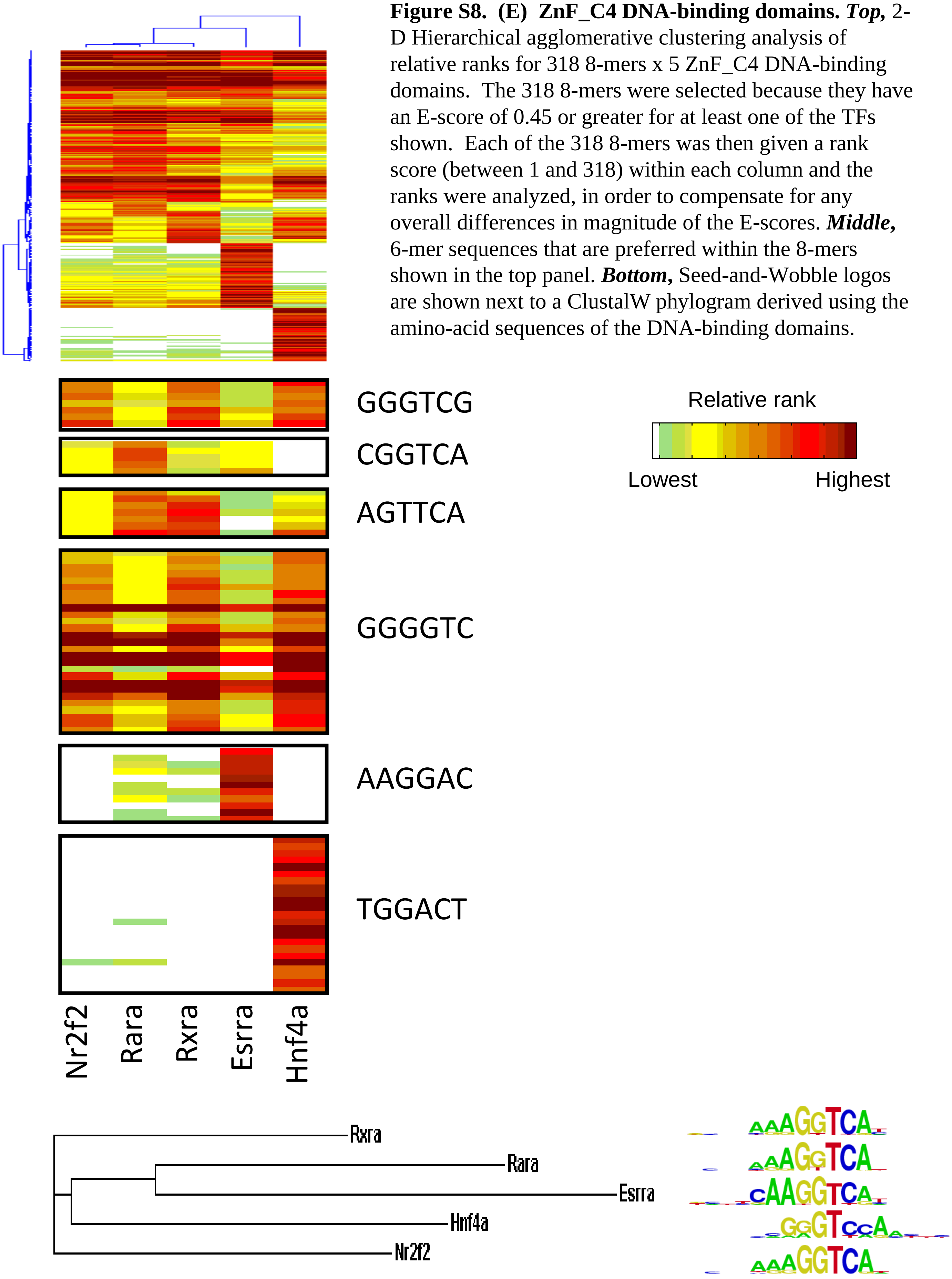


Figure S8. (E) ZnF_C4 DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 318 8-mers x 5 ZnF_C4 DNA-binding domains. The 318 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 318 8-mers was then given a rank score (between 1 and 318) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.



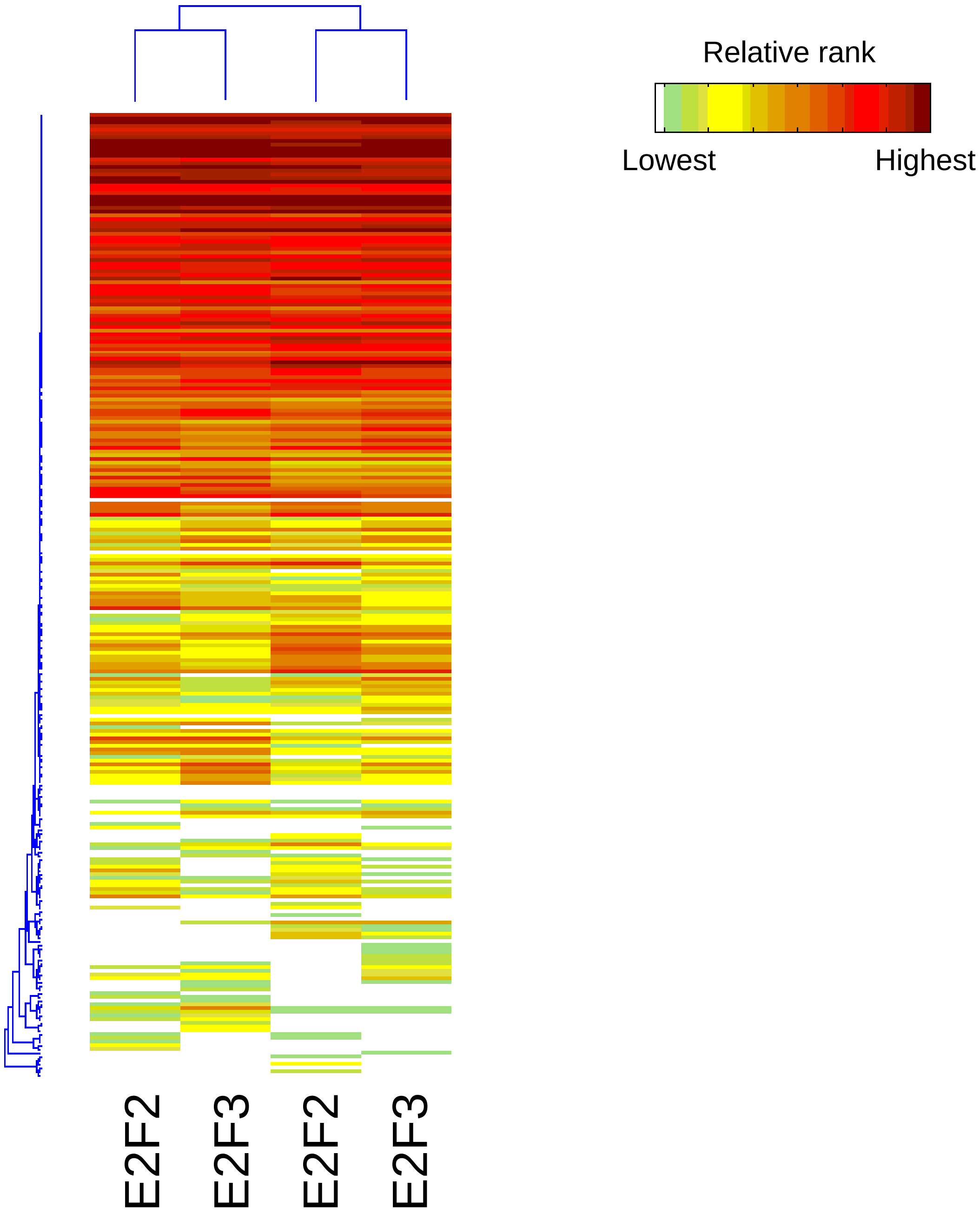
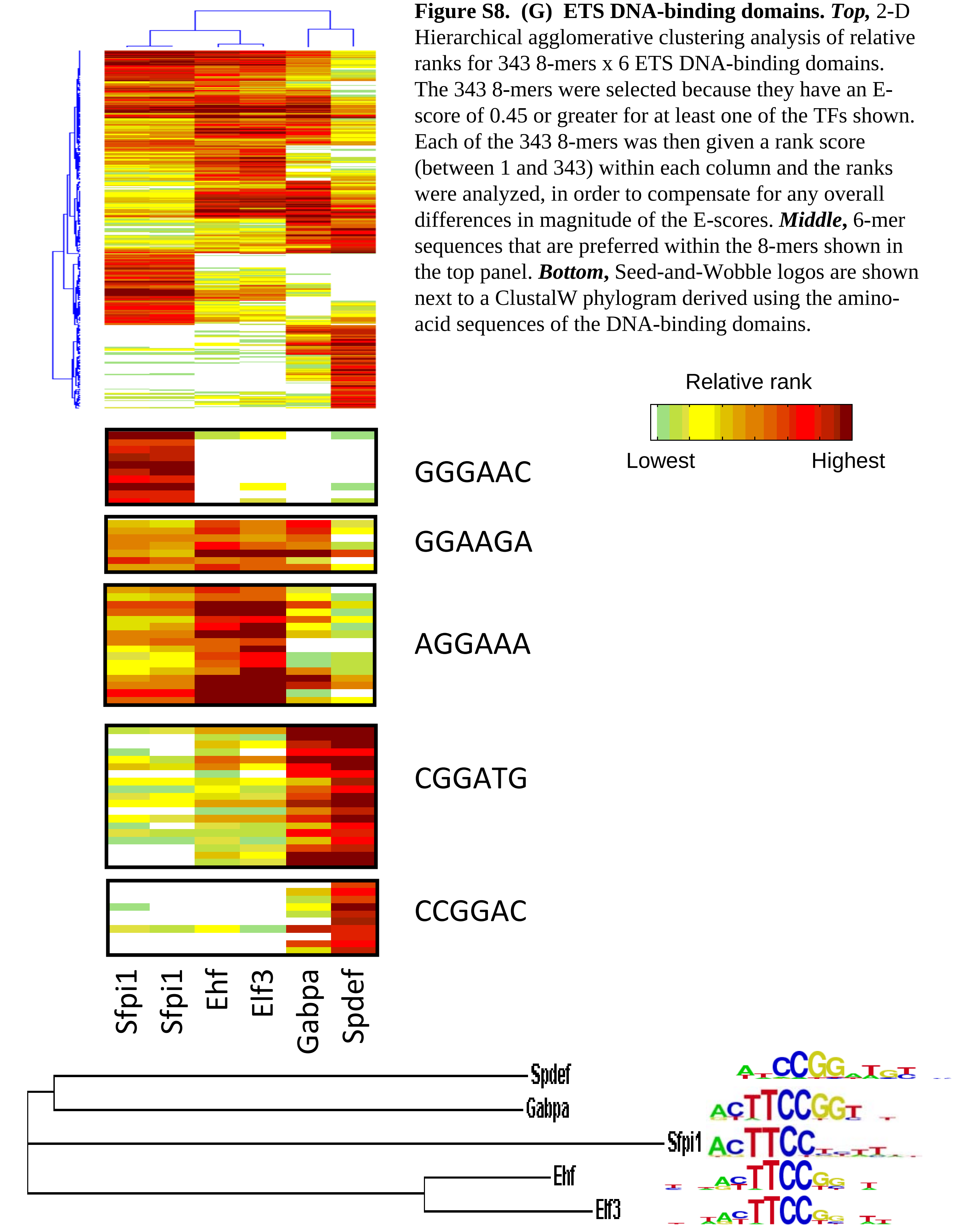


Figure S8. (F) E2F DNA-binding domains. 2-D Hierarchical agglomerative clustering analysis of relative ranks for 260 8-mers x 4 E2F DNA-binding domains. The 260 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 260 8-mers was then given a rank score (between 1 and 260) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores.

Figure S8. (G) ETS DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 343 8-mers x 6 ETS DNA-binding domains. The 343 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 343 8-mers was then given a rank score (between 1 and 343) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.



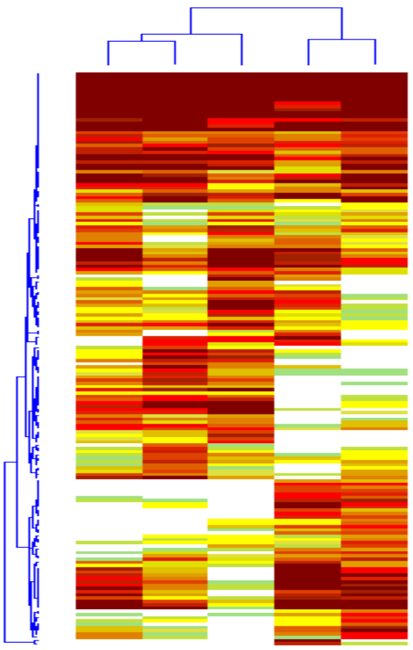
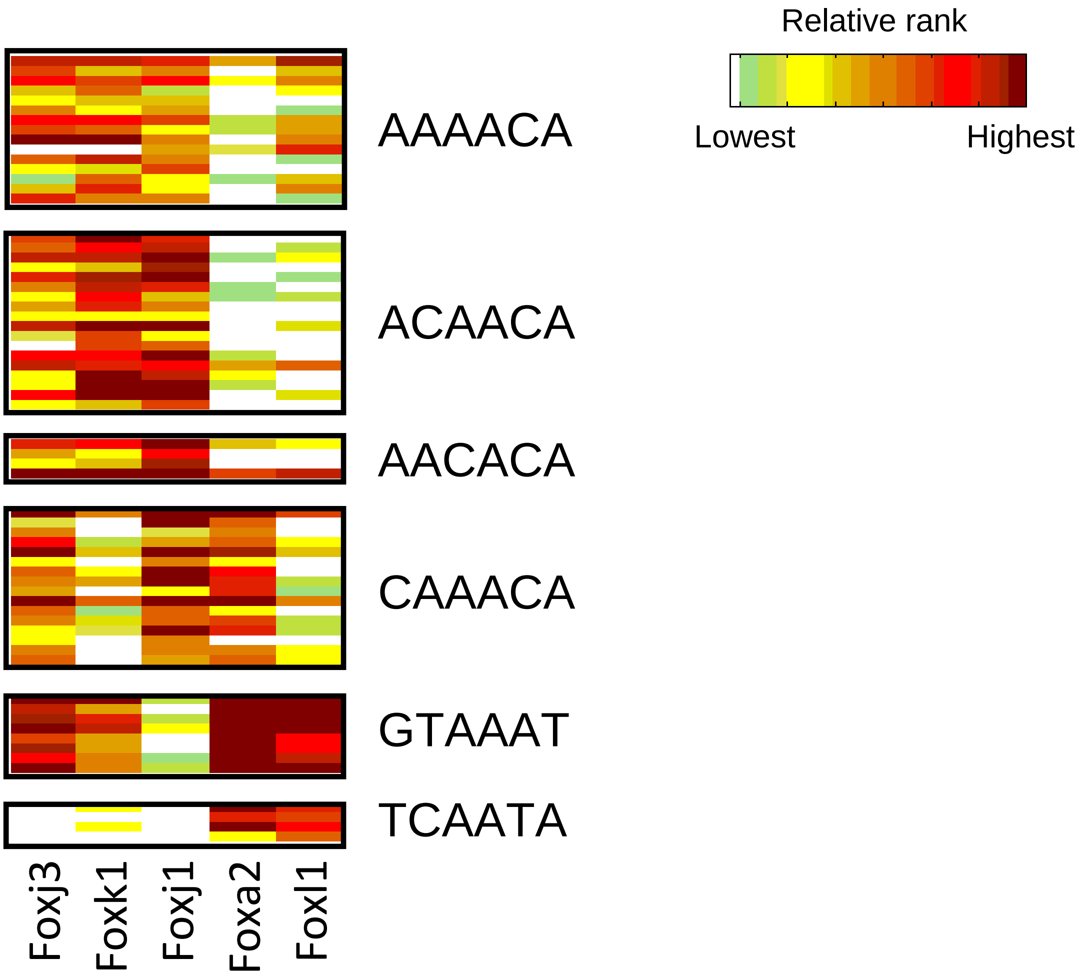


Figure S8. (H) Forkhead (FH) DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 176 8-mers x 5 FH DNA-binding domains. The 176 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 176 8-mers was then given a rank score (between 1 and 176) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.



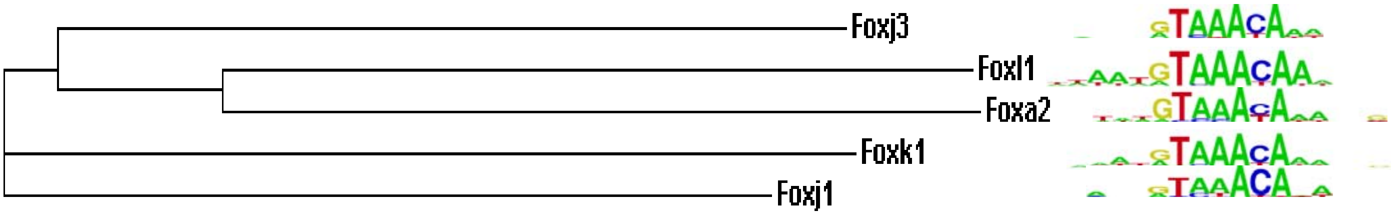
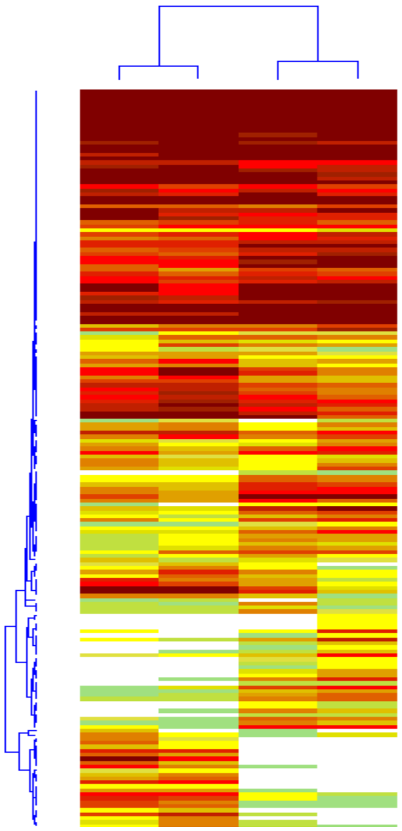
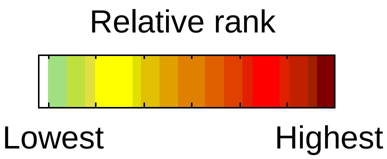
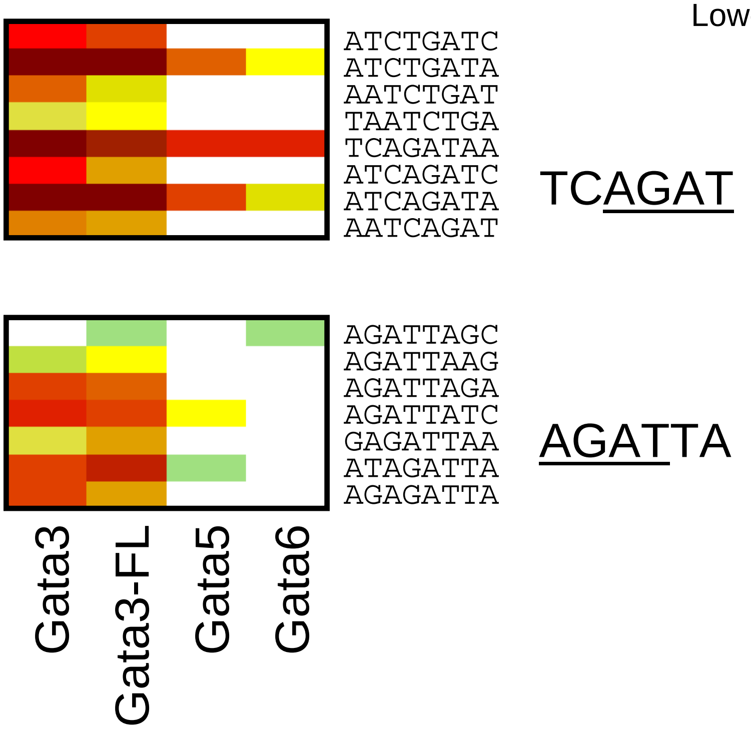
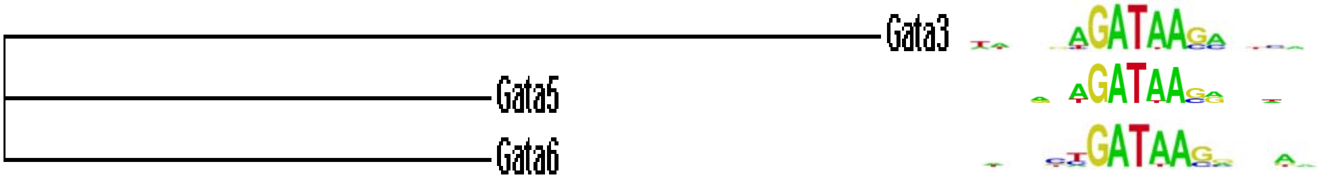


Figure S8. (I) GATA DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 186 8-mers x 3 GATA DNA-binding domains (with Gata3 as both DBD and FL). The 186 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 186 8-mers was then given a rank score (between 1 and 186) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.









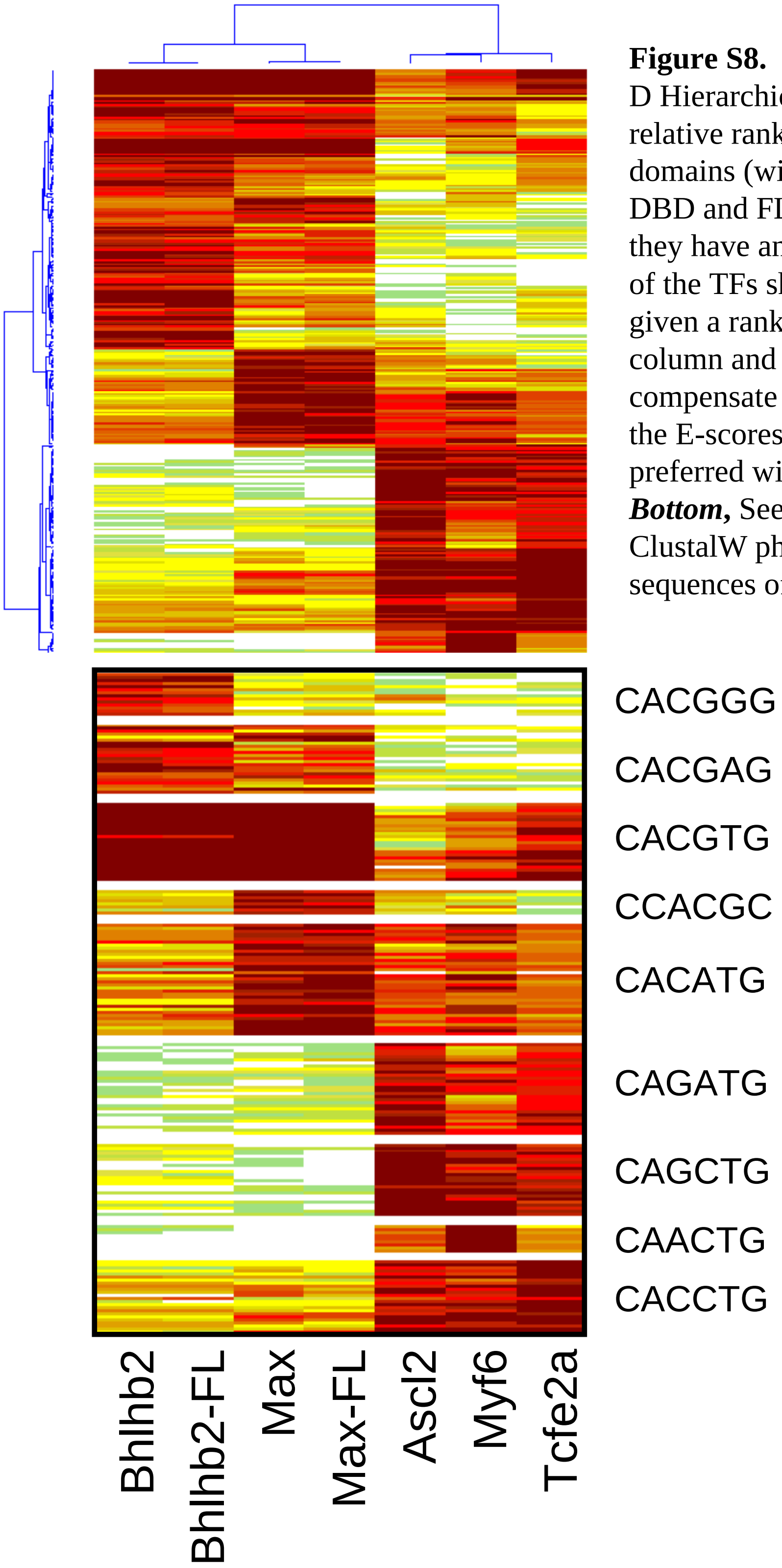
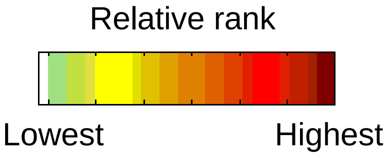
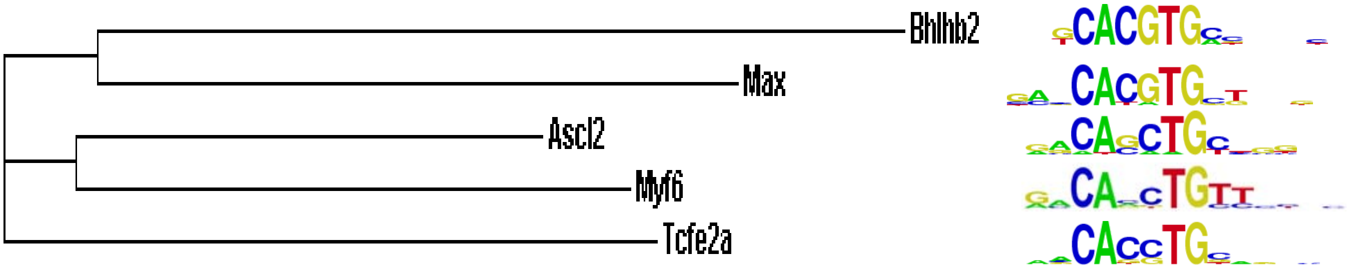


Figure S8. (J) HLH DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 320 8-mers x 6 HLH DNA-binding domains (with Max in duplicate and Bhlhb2 including DBD and FL). The 320 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 320 8-mers was then given a rank score (between 1 and 320) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.





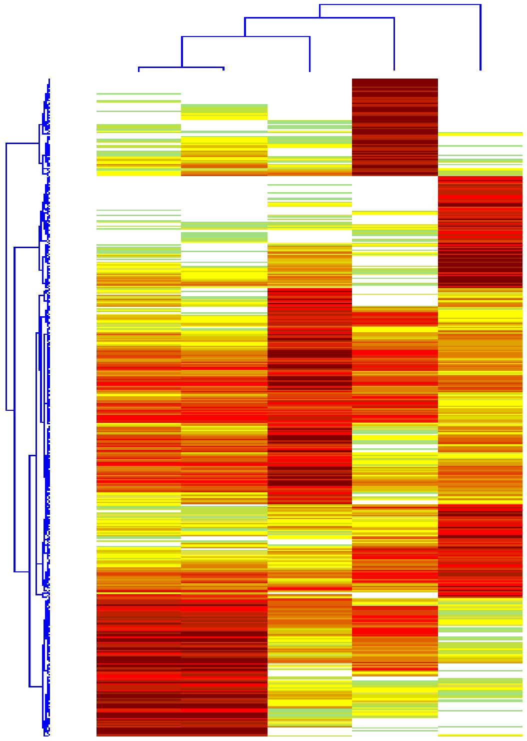
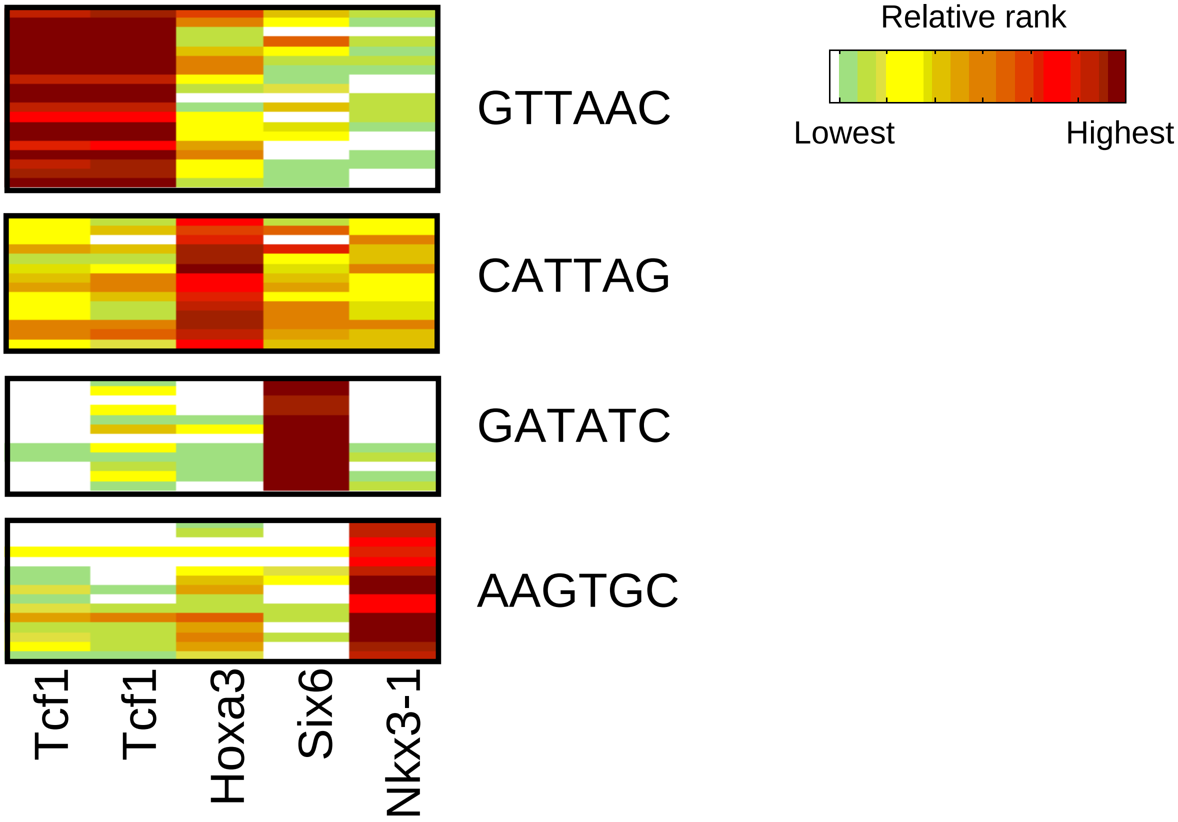
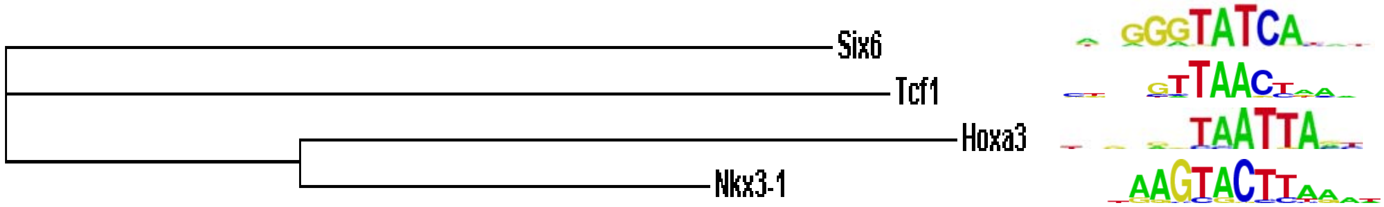


Figure S8. (K) Homeodomain (HOX) DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 514 8-mers x 4 HOX DNA-binding domains (with Tcf1 in duplicate). The 514 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 514 8-mers was then given a rank score (between 1 and 514) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.





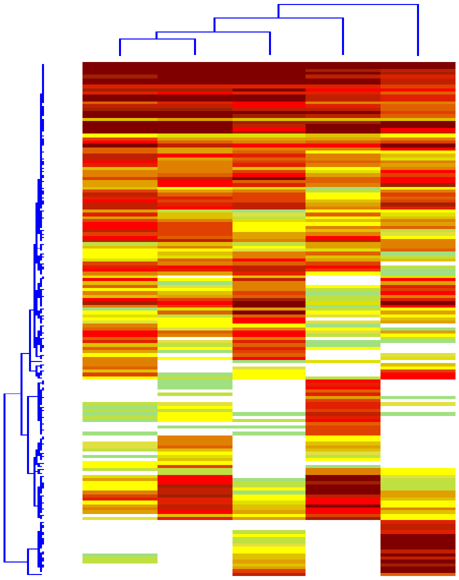


Figure S8. (L) IRF DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 157 8-mers x 5 IRF DNA-binding domains. The 157 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 157 8-mers was then given a rank score (between 1 and 157) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.

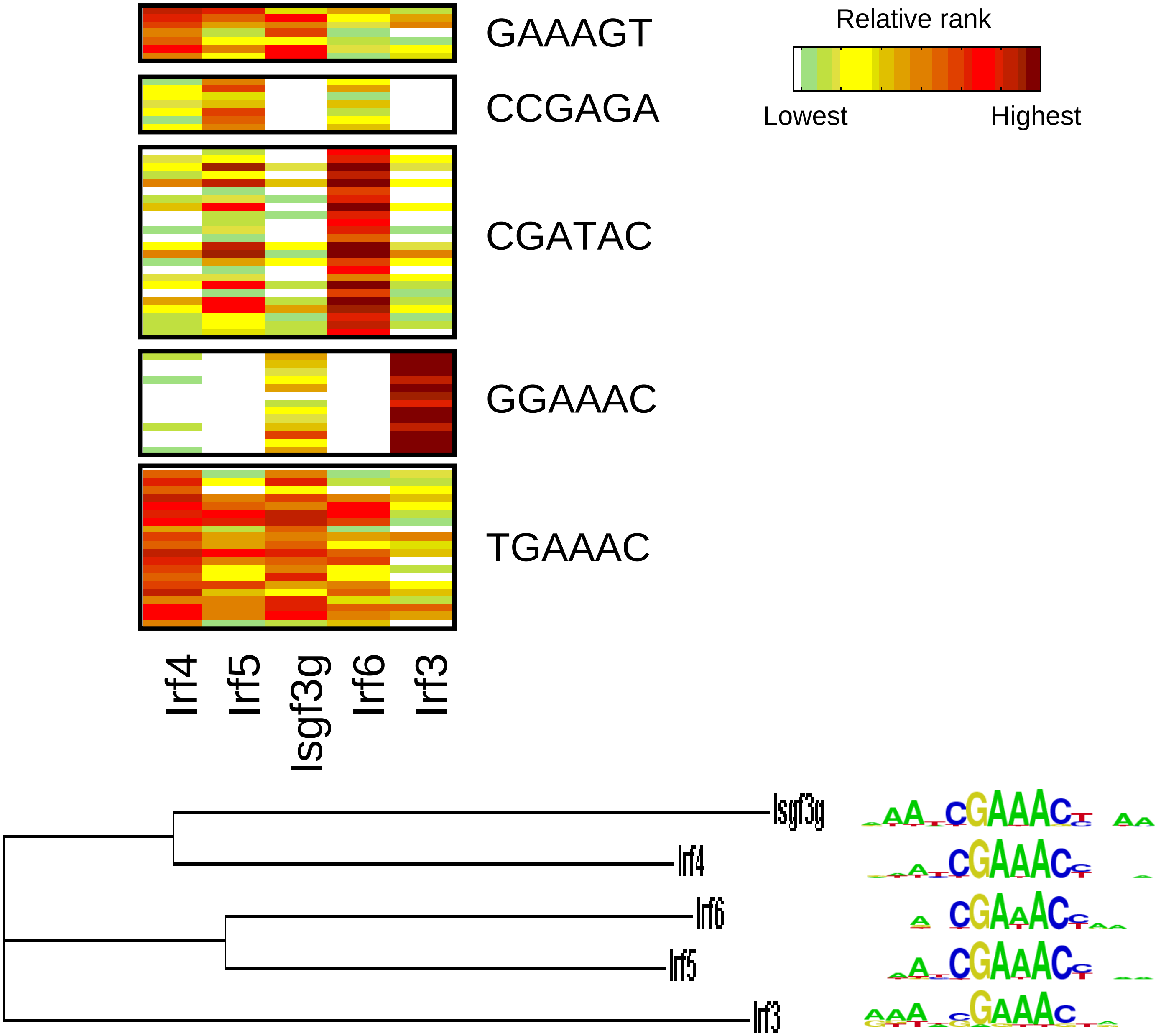
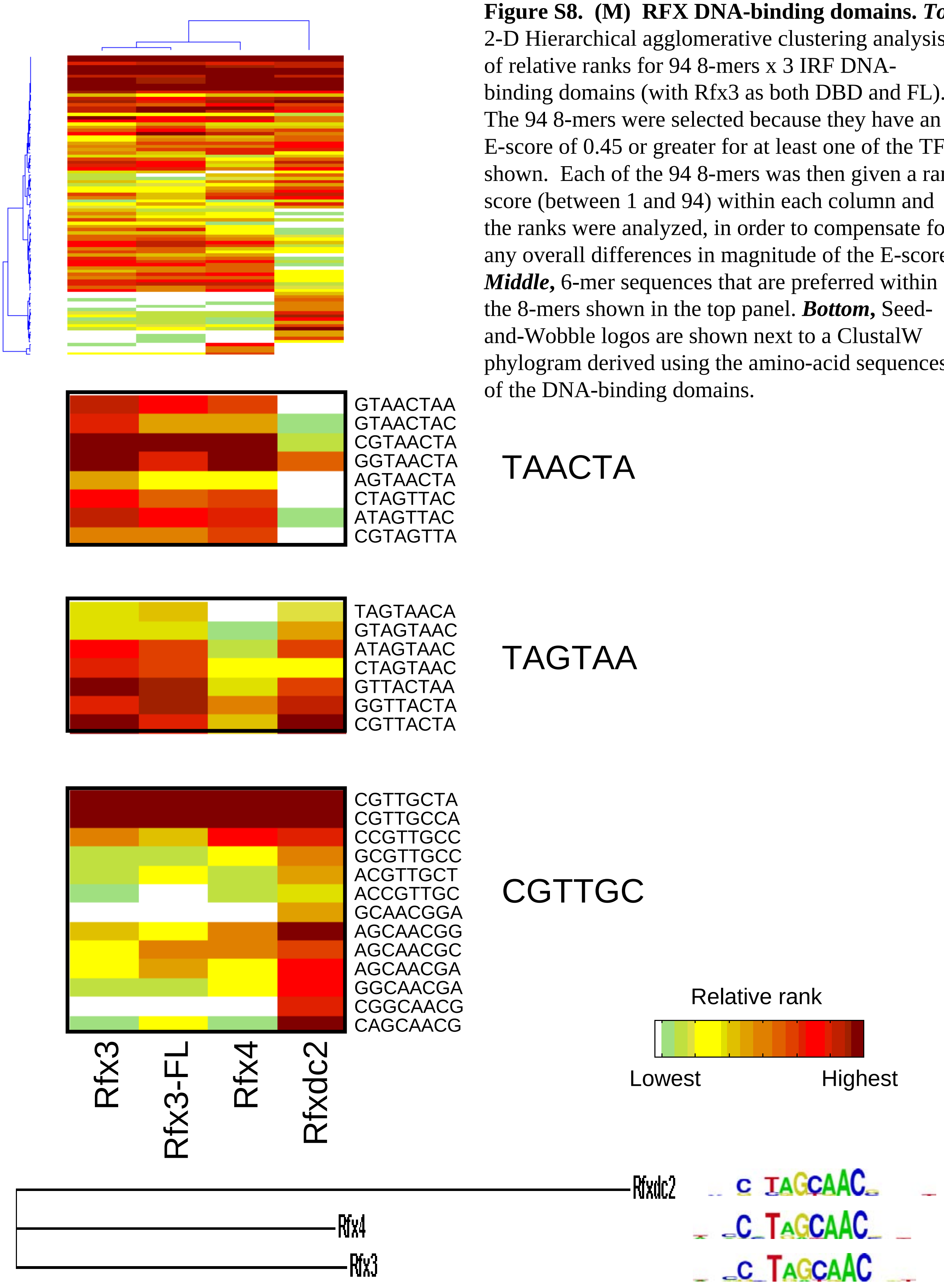


Figure S8. (M) RFX DNA-binding domains. ***Top,*** 2-D Hierarchical agglomerative clustering analysis of relative ranks for 94 8-mers x 3 IRF DNA-binding domains (with Rfx3 as both DBD and FL). The 94 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 94 8-mers was then given a rank score (between 1 and 94) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.



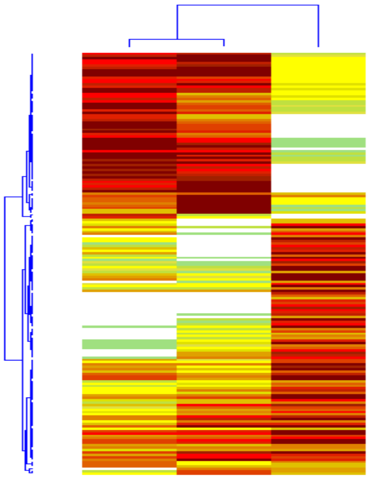
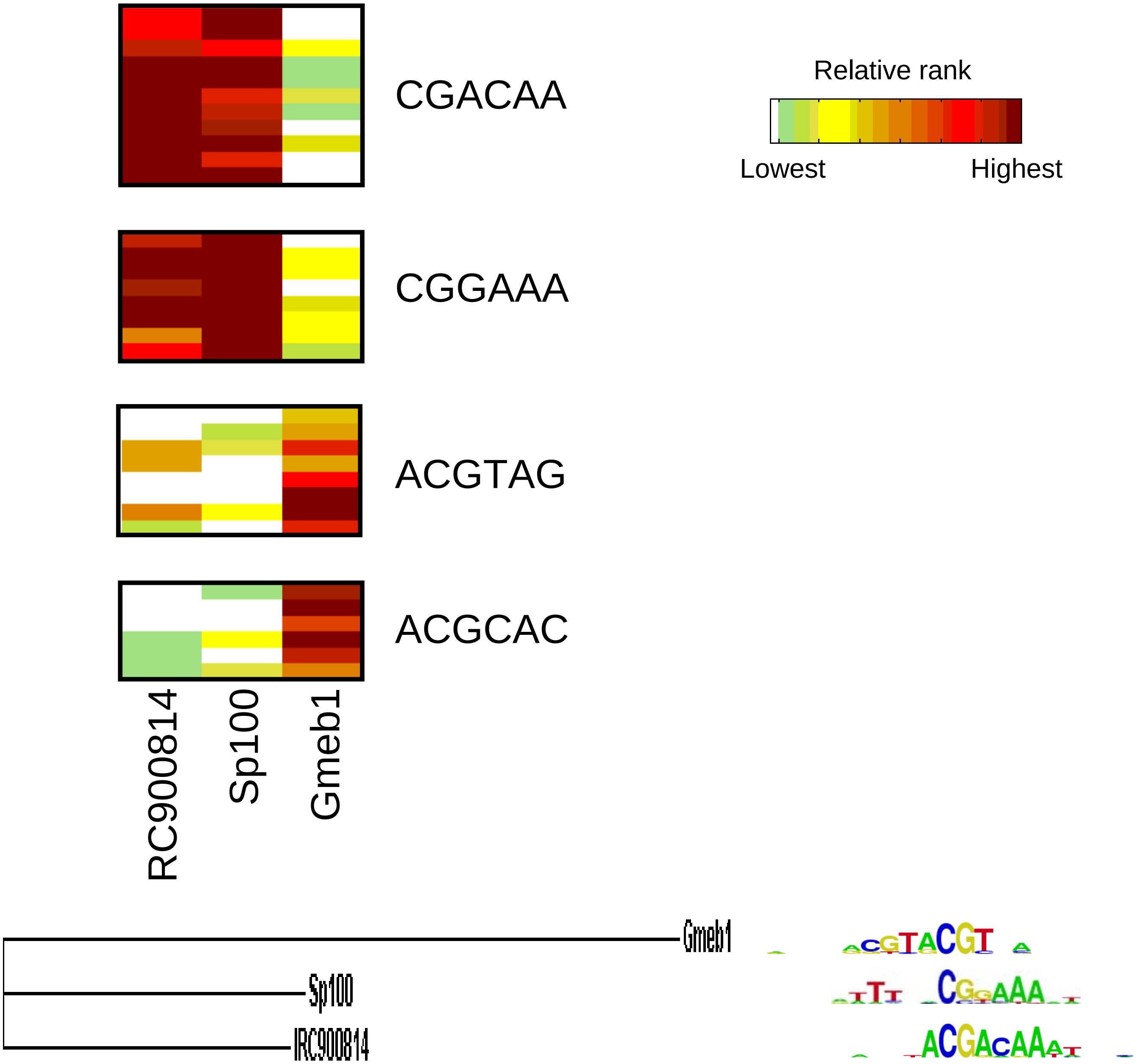


Figure S8. (N) SAND DNA-binding domains. *Top,* 2-D Hierarchical agglomerative clustering analysis of relative ranks for 178 8-mers x 3 SAND DNA-binding domains. The 178 8-mers were selected because they have an E-score of 0.45 or greater for at least one of the TFs shown. Each of the 178 8-mers was then given a rank score (between 1 and 178) within each column and the ranks were analyzed, in order to compensate for any overall differences in magnitude of the E-scores. ***Middle,*** 6-mer sequences that are preferred within the 8-mers shown in the top panel. ***Bottom,*** Seed-and-Wobble logos are shown next to a ClustalW phylogram derived using the amino-acid sequences of the DNA-binding domains.



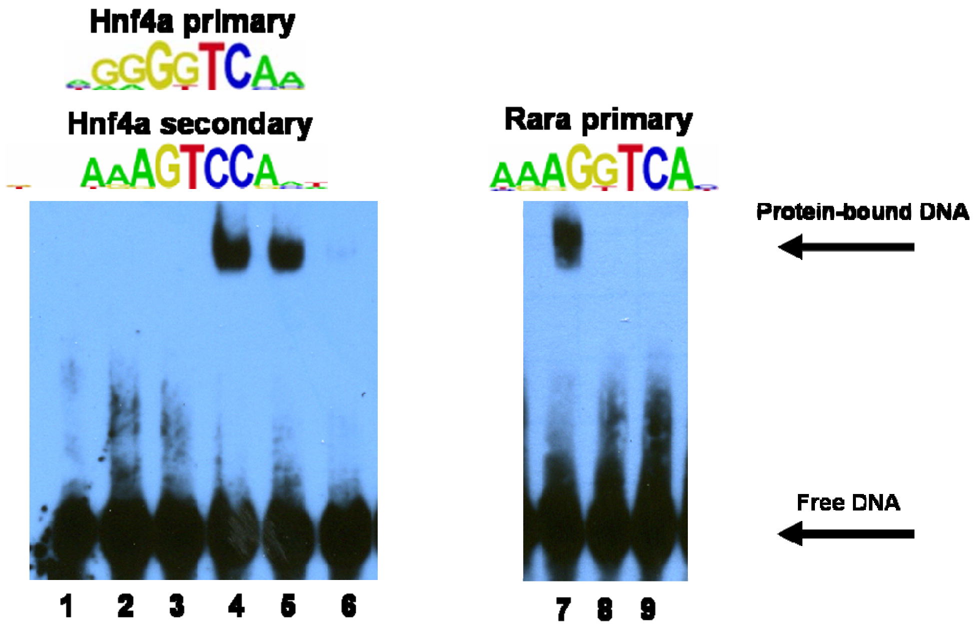


Figure S9. EMSA confirmation of secondary motifs. EMSAs were performed to validate binding to secondary motifs, as determined by the Seed-and-Wobble algorithm (Berger *et al.*, *Nature Biotechnology*, 2006) for Hnf4a. Lane 1: Hnf4a primary probe alone; lane 2: Hnf4a secondary probe alone; lane 3: GGTCCCA probe; lane 4: Hnf4a protein + Hnf4a primary probe; lane 5: Hnf4a protein + Hnf4a secondary probe; lane 6: Hnf4a protein + GGTCCCA probe; lane 7: Rara protein + Hnf4a primary probe; lane 8: Rara protein + Hnf4a secondary probe; lane 9: Rara protein + GGTCCCA probe. Lanes 1-6 show that Hnf4a binds to both the primary and secondary motifs derived by PBM, and very weakly to a third probe containing the sequence GGTCCCA; see **Materials and Methods** for the complete probe sequences. Hnf4a is the only C4 class of zinc finger proteins assayed in this study which showed a preference for this secondary motif (GGTCCA secondary, GGTCA primary). To validate that this secondary motif is specific to Hnf4a, we ran the same probes against another C4 zinc finger protein, Rara (lanes 7-9). Rara can bind to the Hnf4a primary motif sequence (GGTCA), but not the secondary motif of Hnf4a (GGTCCA), or to a probe containing the sequence (GGTCCCA); Rara did not yield a significant secondary Seed-and-Wobble PBM motif. All probe sequences are provided in the **Materials and Methods**.

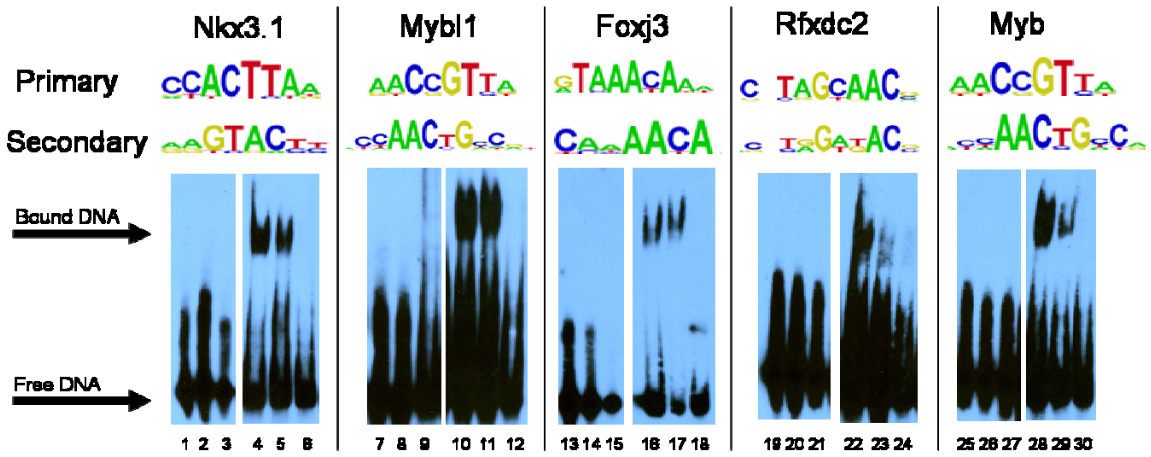
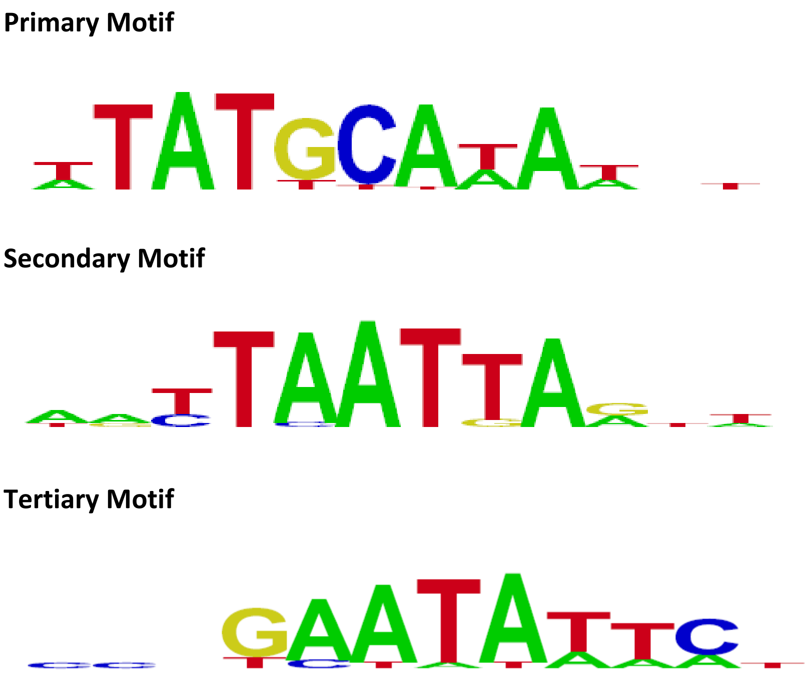


Figure S9 (continued). EMSA confirmation of secondary motifs. EMSAs were performed to validate binding to secondary motifs, as determined by the Seed-and-Wobble algorithm (Berger *et al.*, *Nature Biotechnology*, 2006) Lane 1: Nkx3.1 primary probe alone; lane 2: Nkx3.1 secondary probe alone; lane 3: Foxj3 primary probe alone; lane 4: Nkx3.1 protein + Nkx3.1 primary probe; lane 5: Nkx3.1 protein + Nkx3.1 secondary probe; lane 6: Nkx3.1 protein + Foxj3 primary probe; lane 7: Mybl1 primary probe alone; lane 8: Mybl1 secondary probe alone; lane 9: Foxj3 primary probe alone; lane 10: Mybl1 protein + Mybl1 primary probe; lane 11: Mybl1 protein + Mybl1 secondary probe; lane 12: Mybl1 protein + Foxj3 primary probe; lane 13: Foxj3 primary probe alone; lane 14: Foxj3 secondary probe alone; lane 15: Nkx3.1 primary probe alone; lane 16: Foxj3 protein + Foxj3 primary probe; lane 17: Foxj3 protein + Foxj3 secondary probe; lane 18: Foxj3 protein + Nkx3.1 primary probe; lane 19: Rfxdc2 primary probe alone; lane 20: Rfxdc2 secondary probe alone; lane 21: Mybl1 primary probe alone; lane 22: Rfxdc2 protein + Rfxdc2 primary probe; lane 23: Rfxdc2 protein + Rfxdc2 secondary probe; lane 24: Rfxdc2 protein + Mybl1 primary probe; lane 25: Myb primary probe alone; lane 26: Myb secondary probe alone; lane 27: Rfxdc2 secondary probe alone; lane 28: Myb protein + Myb primary probe; lane 29: Myb protein + Myb secondary probe; lane 30: Myb protein + Rfxdc2 secondary probe. All probe sequences are provided in the **Materials and Methods**.



Primary Motif

Secondary Motif

Tertiary Motif

Construct	SELEX Consensus Site
POU	TATGCAAAT
POU_{HD}	RTAATNA
POU_{S}	GAATATKC

Verrijzer, et al., EMBO Journal (1992), 11:4993-5003

R = A or G; K = T or G; N = A, C, G, or T

Figure S10: Primary, secondary, and tertiary Seed-and-Wobble motifs for the human POU homeodomain Oct-1. We searched for secondary and tertiary motifs in previously generated universal PBM data [Berger, *et al*., *Nature Biotechnology* (2007), 24:1429-1435] using our modified Seed-and-Wobble algorithm [Berger, *et al*., *Nature Biotechnology* (2007), 24:1429-1435] described in **Materials and Methods**. For one protein, human Oct-1, which has a bipartite POU DNA-binding domain, another group had already determined the consensus binding sites by *in vitro* selection (SELEX) for three separate constructs: the entire POU domain, the POU-specific subdomain (POU_S), and the POU-type homeodomian (POU_{HD}) [Verrijzer, *et al*., *EMBO Journal* (1992), 11:4993-5003]. The three motifs we derived from our universal PBM data correspond exactly to the previously-identified binding sites for these three constructs, suggesting to us that we can capture multiple modes of DNA-protein interactions *in vitro* from a single experiment.

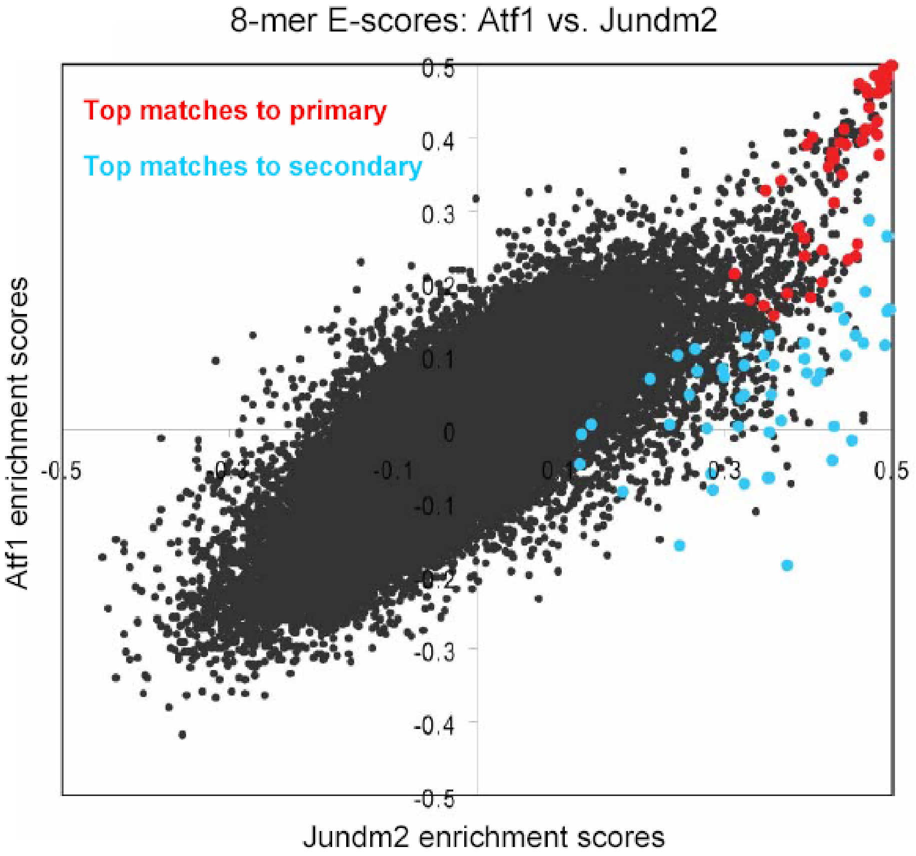


Figure S11. High-scoring *k*-mers belonging to the Jundm2 secondary motif are not bound as well by the related bZIP protein Atf1. Scatter plot comparing 8-mer enrichment scores for closely related TFs.

```
CLUSTAL W (1.83) multiple sequence alignment


RFX3-IVT             TLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLG 60
RFX3-purified        HLQWLLDNYETAEGVSLPRSTLYNHYLRHCQEHKLDPVNAASFGKLIRSIFMGLRTRRLG 60
hRFX1                TVQWLLDNYETAEGVSLPRSTLYCHYLLHCQEQKLEPVNAASFGKLIRSVFMGLRTRRLG 60
RFX4-IVT             TLQWLEENYEIAEGVCIPRSALYMHYLDFCEKNDTQPVNAASFGKIIRQQFPQLTTRRLG 60
RFXDC2-purified      AFSWIRNTLEEHPETSLPKQEVYDEYKSYCDNLGYHPLSAADFGKIMKNVFPNMKARRLG 60
                      ..*: :. *    ..:*:. :* .*  .*::   .*:.**.***:::. *  : :****

RFX3-IVT             TRGNSKYHYYGIRVKPDSPLNR 82
RFX3-purified        TRGNSKYHYYGIRVKPDSPLN- 81
hRFX1                TRGNSKYHYYGLRIKASSPLLR 82
RFX4-IVT             TRGQSKYHYYGIAVKESSQYY- 81
RFXDC2-purified      TRGKSKYCYSGLRKKAFVHMP- 81
                     ***:*** * *:  *
```

Figure S12. RFX family protein-DNA recognition positions. It is likely that RFX3, RFX4, and RFXDC2 all use the same mechanism of alternative modes of DNA recognition as RFX1 (Gajiwala *et al.*, *Nature*, 2000), because seven out of nine residues involved in direct or water-mediated DNA contacts (highlighted in red) are identical among these proteins, while the other two residues have conservative substitutions.

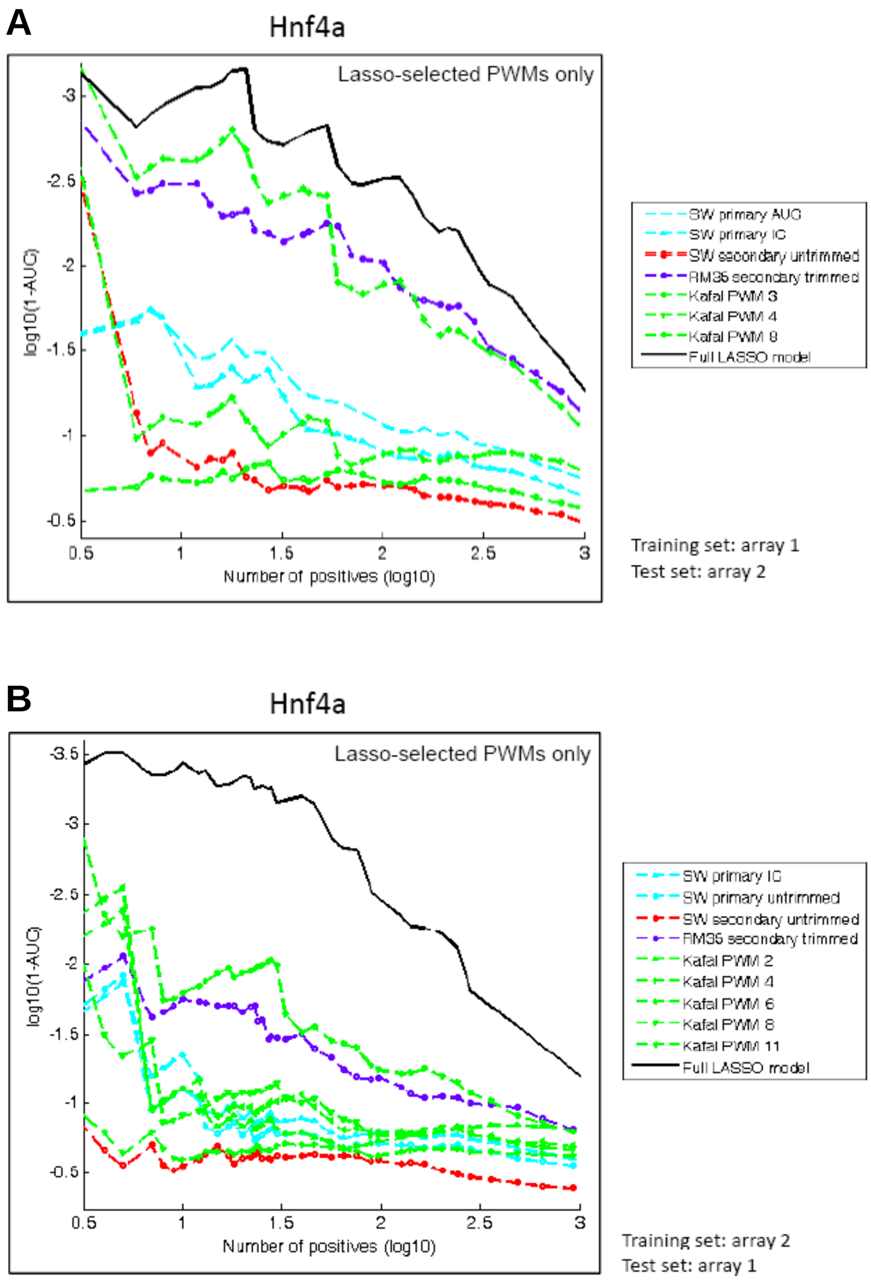
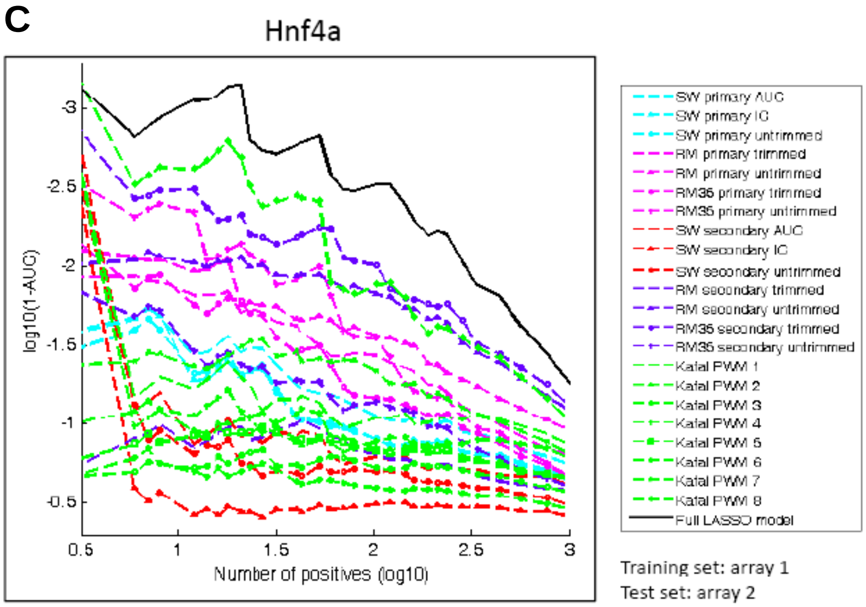
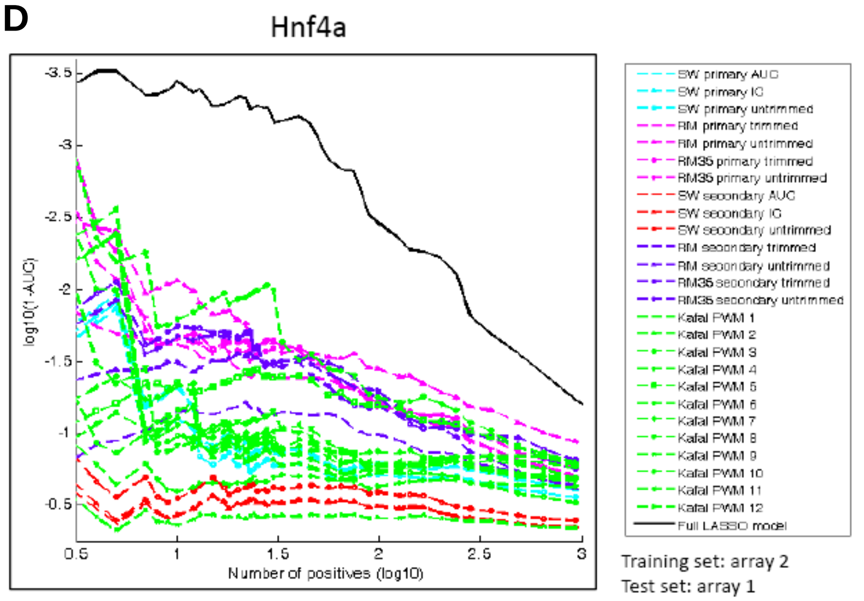


Figure S13: Graphs showing $\log_{10}$(1-AUC) (area under ROC curve) (*y*-axis) versus $\log_{10}$(number of positives) (*x*-axis) for Hnf4a. Log_{10}(1-AUC) is shown to highlight differences between the methods, all of which have an AUC near 1. Graphs were generated using Array 1 as training and Array 2 as test data (panels **A,C**; *this and next page*), and separately using Array 2 as training and Array 1 as test data **(**panels **B,D**; *this and next page*). The solid black line ("Full Lasso model") indicates performance of the multiple motif model; all other lines indicate performance of various other individual motifs identified by other motif finding algorithms (see **Materials and Methods**). For clarity, only data for the Lasso-selected PWMs are shown in panels **A,B**; plots showing data from all motifs considered are shown in panels **C,D**.



Training set: array 1
Test set: array 2



Training set: array 2
Test set: array 1

(A) Hnf4a

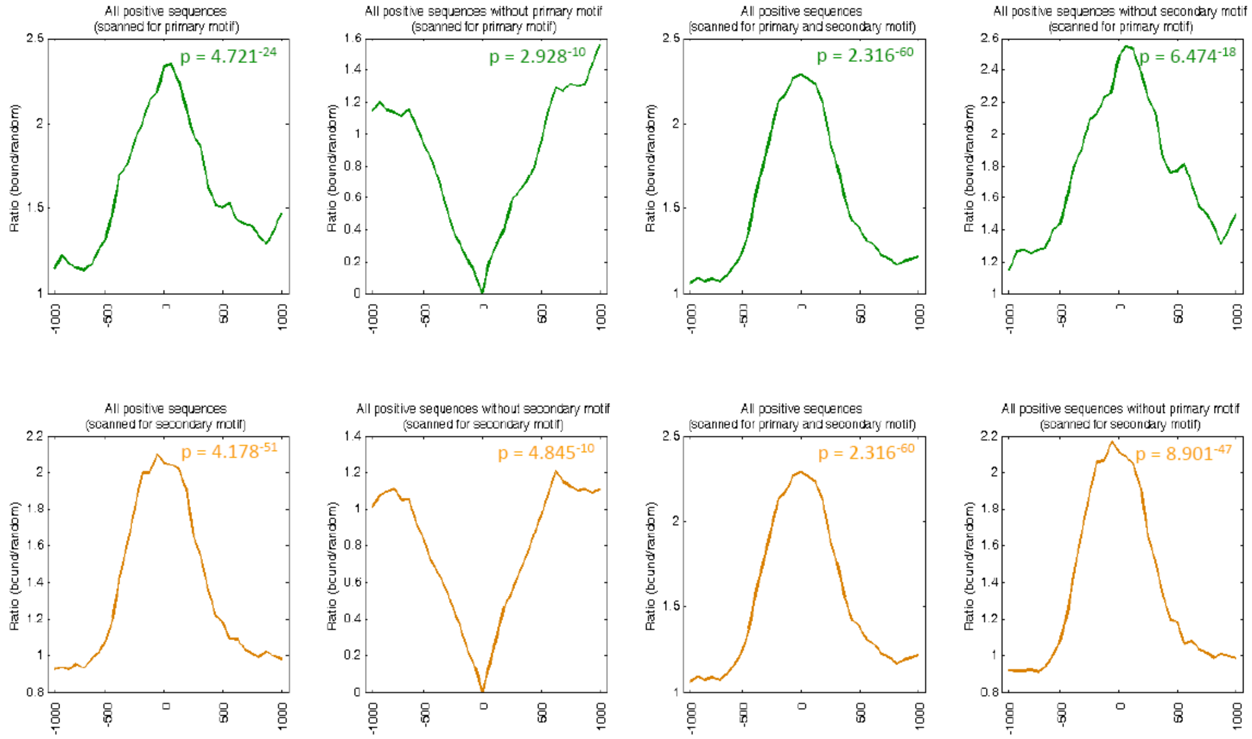
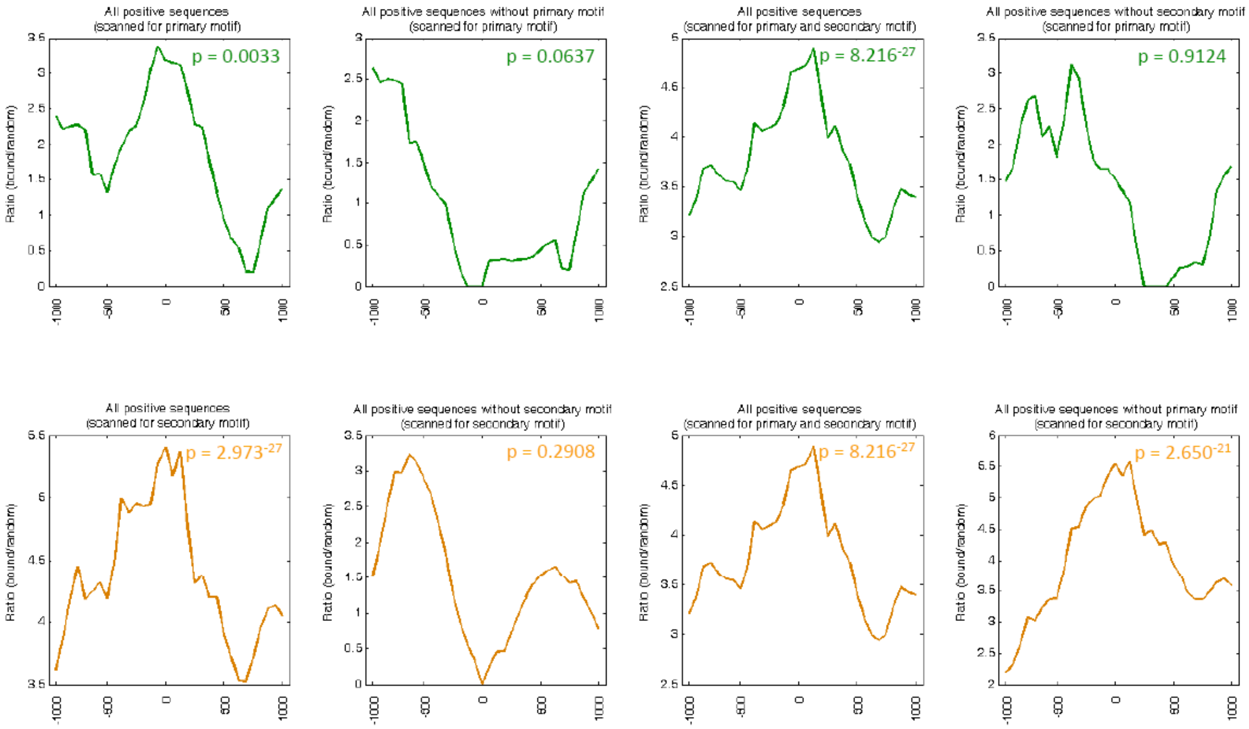
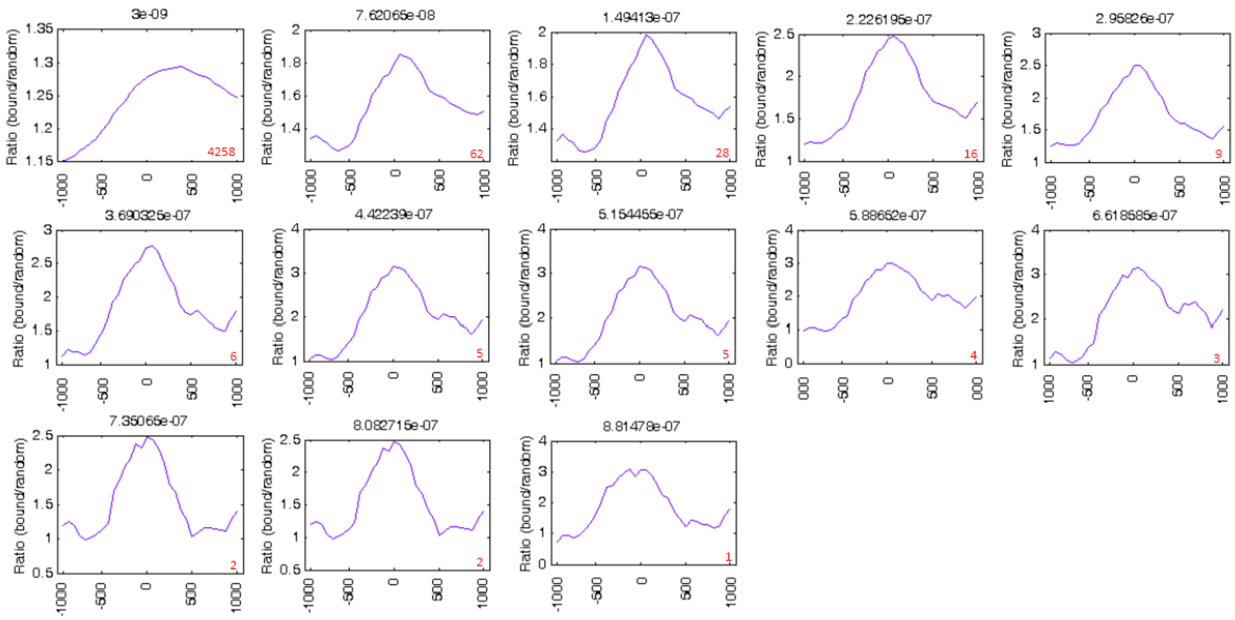


Figure S14: Enrichment of primary versus secondary motif 8-mers bound *in vitro* within genomic regions bound *in vivo*. Relative enrichment of *k*-mers corresponding to the primary versus secondary Seed-and-Wobble motifs within bound genomic regions in ChIP-chip data as compared to randomly selected sequences was calculated (see **Materials and Methods**) for **(A, C, D)** Hnf4a (Neilsen *et al.*, submitted; GEO accession #GSE7745) and **(B, E, F) (next page)** Bcl6b (*34*) (GEO accession #GSE7673). ChIP-chip 'bound' regions were identified according to the criteria of the respective studies (*34*)(Neilsen *et al.*, submitted). A window size of 500 bp with a step size of 100 bp was used. Either all 'bound' regions (far left, upper and lower rows), 'bound' regions lacking primary motif *k*-mers (second from left, upper row; far right, lower row) or 'bound' regions lacking secondary motif *k*-mers (far right, upper row; second from left, lower row) were considered for matches to primary motif *k*-mers (far left, second from left, and far right in upper row), secondary motif *k*-mers (far left, second from left, and far right in lower row), or either primary or secondary motif *k*-mers (second from right, upper and lower rows). The coarseness of the Bcl6 distributions is due to a smaller sample size of ChIP-chip 'bound' regions. The GOMER thresholds used in **(A)** are 2.958 x 10^{-7} and 8.419 x 10^{-7}, corresponding to 9 primary and 20 secondary 8-mers scanned, respectively for Hnf4a. The GOMER thresholds used for the data shown in **(B)** correspond to 1.513 x 10^{-6} and 3.294 x 10^{-7} corresponding to 4 primary and 17 secondary 8-mers scanned, respectively, for Bcl6b. *P*-values for enrichment of 8-mers within the bound genomic regions shown in each panel were calculated for the interval −250 to +250 by the Wilcoxon-Mann-Whitney rank sum test, comparing the number of occurrences per sequence in the bound set versus the background set. Enrichment plots at varying GOMER score thresholds (indicated above each plot in panels **C-F, next pages**) are shown in **(C, D)** for Hnf4a and **(E, F)** for Bcl6b for primary **(C, E)** versus secondary **(D, F)** motifs using a window size of 500 bp and a step size of 50 bp. Enrichment is generally observed across varying GOMER thresholds, with the exception that at permissive GOMER thresholds enrichment can be lost. Number of *k*-mers included at each GOMER threshold is indicated in red on each plot in panels **C-F**.

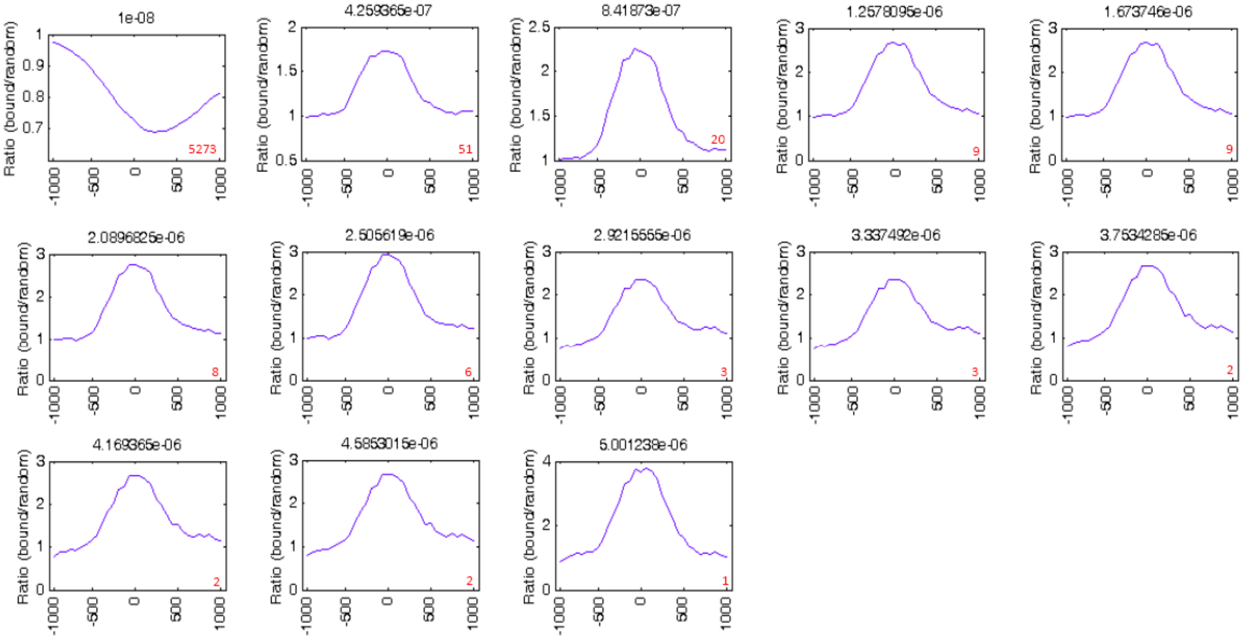
(B) Bcl6b


All positive sequences
(scanned for primary motif)
p = 0.0033
All positive sequences without primary motif
(scanned for primary motif)
p = 0.0637
All positive sequences
(scanned for primary and secondary motif)
p = 8.216^{-27}
All positive sequences without secondary motif
(scanned for primary motif)
p = 0.9124
All positive sequences
(scanned for secondary motif)
p = 2.973^{-27}
All positive sequences without secondary motif
(scanned for secondary motif)
p = 0.2908
All positive sequences
(scanned for primary and secondary motif)
p = 8.216^{-27}
All positive sequences without primary motif
(scanned for secondary motif)
p = 2.650^{-21}
Ratio (bound/random)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
-1000
-500
0
500
1000

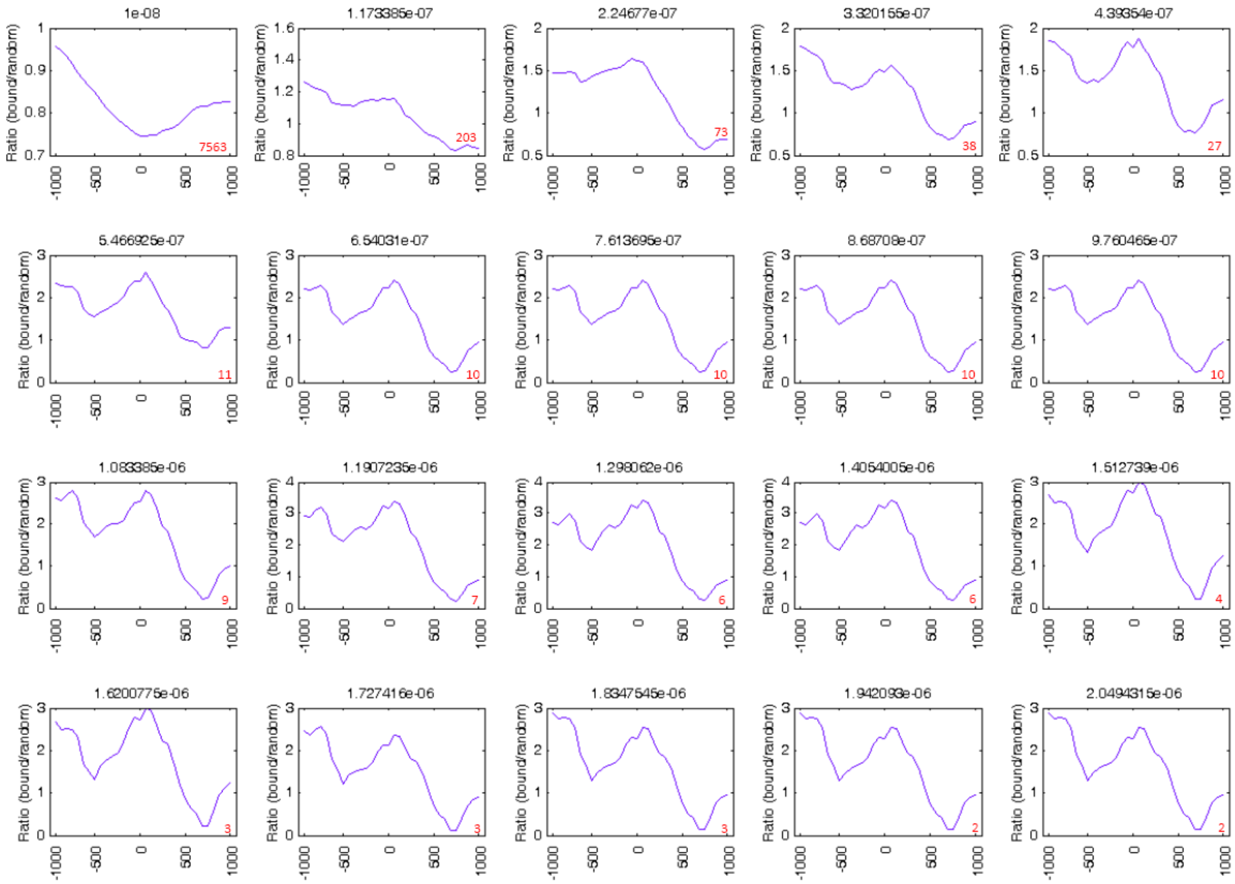
(C) Hnf4a primary motif enrichment within ‘bound’ genomic regions



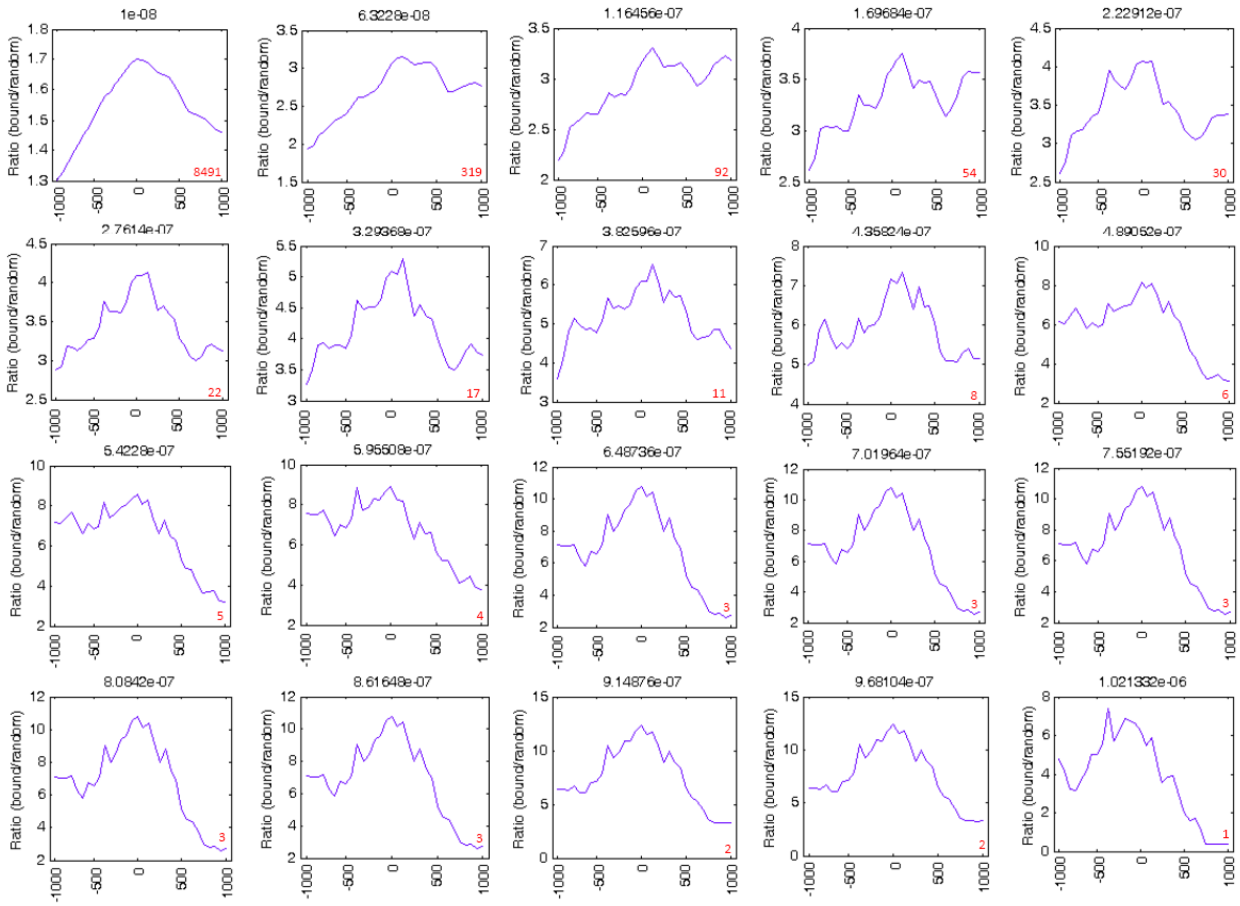
(D) Hnf4a secondary motif enrichment within ‘bound’ genomic regions



(E) Bcl6b primary motif enrichment within ‘bound’ genomic regions



(F) Bcl6b secondary motif enrichment within ‘bound’ genomic regions


1e-08
6.3228e-08
1.16456e-07
1.69684e-07
2.22912e-07
2.7614e-07
3.29369e-07
3.82596e-07
4.35824e-07
4.89052e-07
5.4228e-07
5.95508e-07
6.48736e-07
7.01964e-07
7.55192e-07
8.0842e-07
8.61648e-07
9.14876e-07
9.68104e-07
1.021332e-06
Ratio (bound/random)