Cell Systems, Volume *6*

Supplemental Information

Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential *In Vivo* Binding

Ning Shen, Jingkang Zhao, Joshua L. Schipper, Yuning Zhang, Tristan Bepler, Dan Leehr, John Bradley, John Horton, Hilmar Lapp, and Raluca Gordan

Supplemental Information for:

Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding

Authors:

Ning Shen[1,2,3], Jingkang Zhao[1,3,4], Joshua L. Schipper[1,3], Yuning Zhang[1,4], Tristan Bepler[1], Dan Leehr[1], John Bradley[1], John Horton[1,3], Hilmar Lapp[1], Raluca Gordan[1,3,5,6]*

[1]Center for Genomic and Computational Biology, [2]Department of Pharmacology and Cancer Biology, [3]Department of Biostatistics and Bioinformatics, [4]Program in Computational Biology and Bioinformatics, [5]Department of Computer Science, Duke University, [6]Department of Molecular Genetics and Microbiology, Duke University Durham, NC 27708, USA

Email address of corresponding author and lead contact:

* Corresponding author and lead contact. E-mail: raluca.gordan@duke.edu

Supplemental Figures

Figure S1. Related to Figure 1. DNA motifs derived from *in vivo* binding data. PWM motif models derived from ChIP-seq data (ENCODE(Consortium, 2012)) using the MEME-ChIP software tool (Machanick and Bailey, 2011). For TF E2f3, which does not have ChIP-seq data in human cells, the motif was downloaded from the TRANSFAC database, and it was derived from a small, curated set of E2f3 genomic binding sites (Matys et al., 2006).

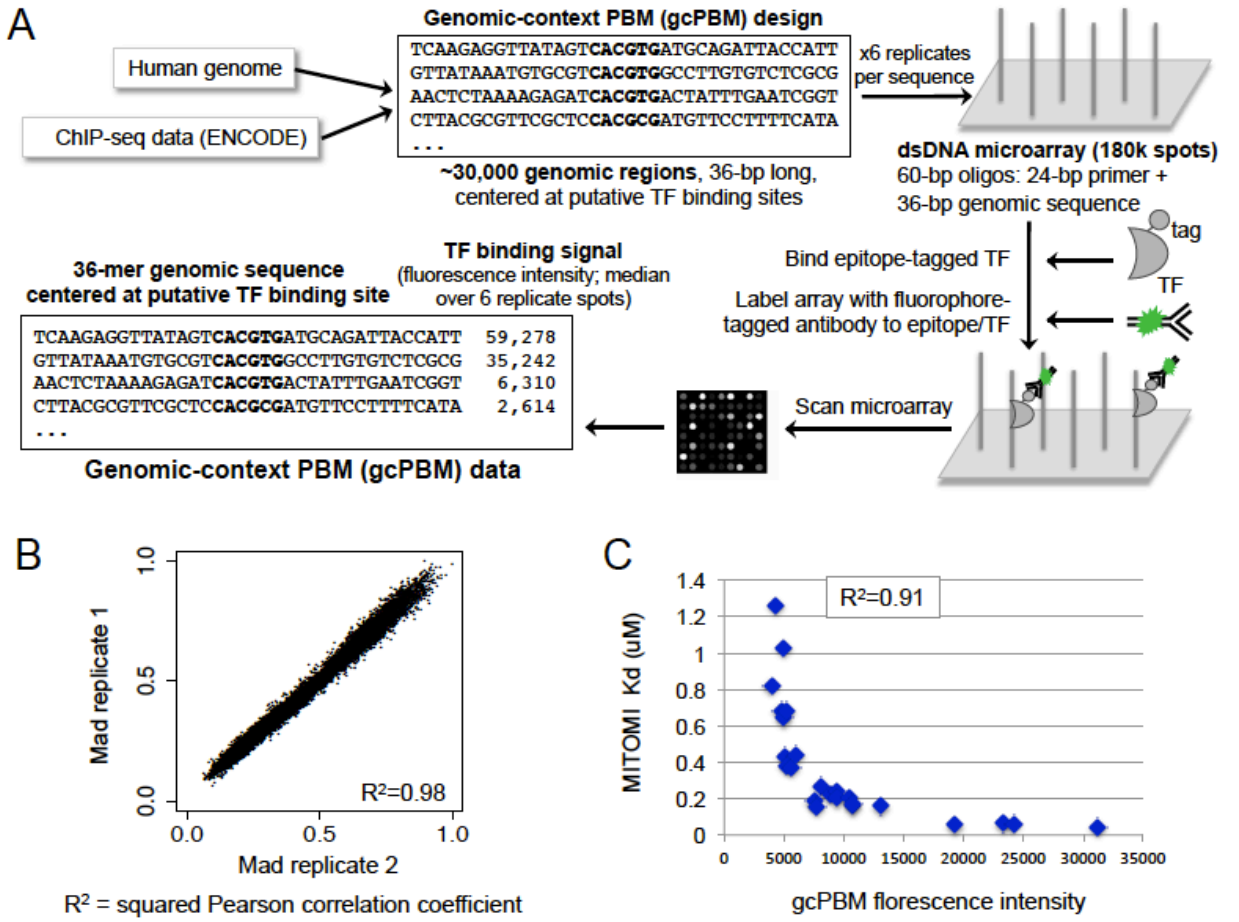


Figure S2. Related to Figure 1. The genomic-context protein-binding microarray (gcPBM) assay. **(A)** gcPBM design. A set of ~30,000 distinct genomic sequences, centered on putative TF binding sites, are selected according ChIP-seq data (see Methods for details). The sequences are synthetized de novo on microarrays (Agilent, 4x180k array format), with each genomic sequence being represented in 6 replicate DNA spots randomly distributed across the microarray. After DNA double stranding (Berger and Bulyk, 2009), the microarray is incubated with an epitope-tagged TF, and labeled with a fluorophore-conjugated antibody to the epitope. The array is gently washed to remove non-specifically bound proteins, and then scanned to quantify the amount of protein bound to each DNA spot. **(B)** Representative example of the agreement between replicate gcPBM data set. Plots show the natural logarithm of the gcPBM fluorescence intensity (median over 6 replicate spots, randomly distributed across the array) for human TF Mad. Higher values correspond to higher affinity binding sites. **(C)** Correlation between gcPBM and MITOMI data. High affinity TF binding sites have high PBM fluorescent intensities and low equilibrium dissociation constants (Kd). We note that binding affinity data is only available for a few human TFs and a few DNA sites. The only medium-scale Kd data available for human TFs is the MITOMI data for TF Max (Maerkl and Quake, 2007). We analyzed 24 TTGnnnnGTGGGTG DNA sites that were tested both by MITOMI and by gcPBM, and found a remarkable correlation of gcPBM intensities with MITOMI affinities. For comparison, the agreement between the Max affinity data and predictions made using existing Max-DNA binding models is much lower ($R^2 < 0.59$), as reported in previous work (Mathelier and Wasserman, 2013). Importantly, gcPBM and MITOMI data correlate over a wide affinity range (40nM - 1.2uM), which includes medium and low affinity sites. This is important for characterizing TF binding because low and medium affinity sites can play important regulatory roles, especially during development.

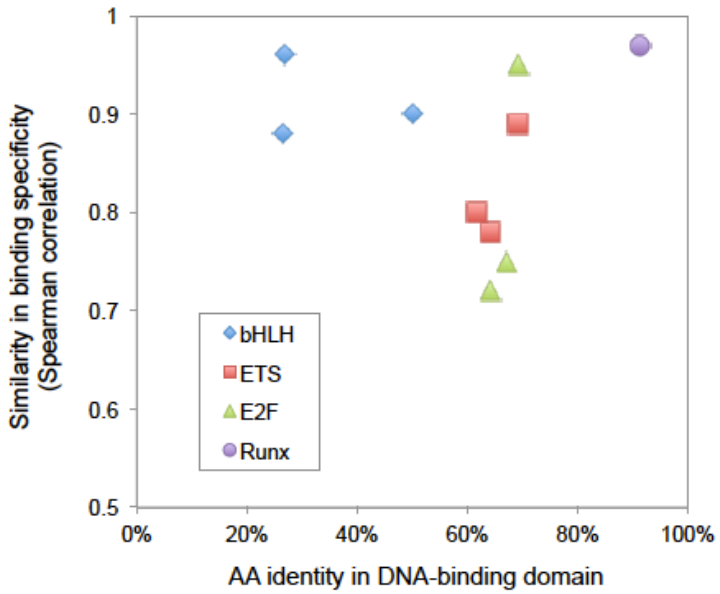


Figure S3. Related to Figure 1. Correlation between amino-acid similarity and DNA binding specificity similarity for paralogous TF. Scatter plot shows that the % identify in the amino-acid sequence of the DNA-binding domains of paralogous TFs does not correlate well with their similarity in DNA-binding specificity.

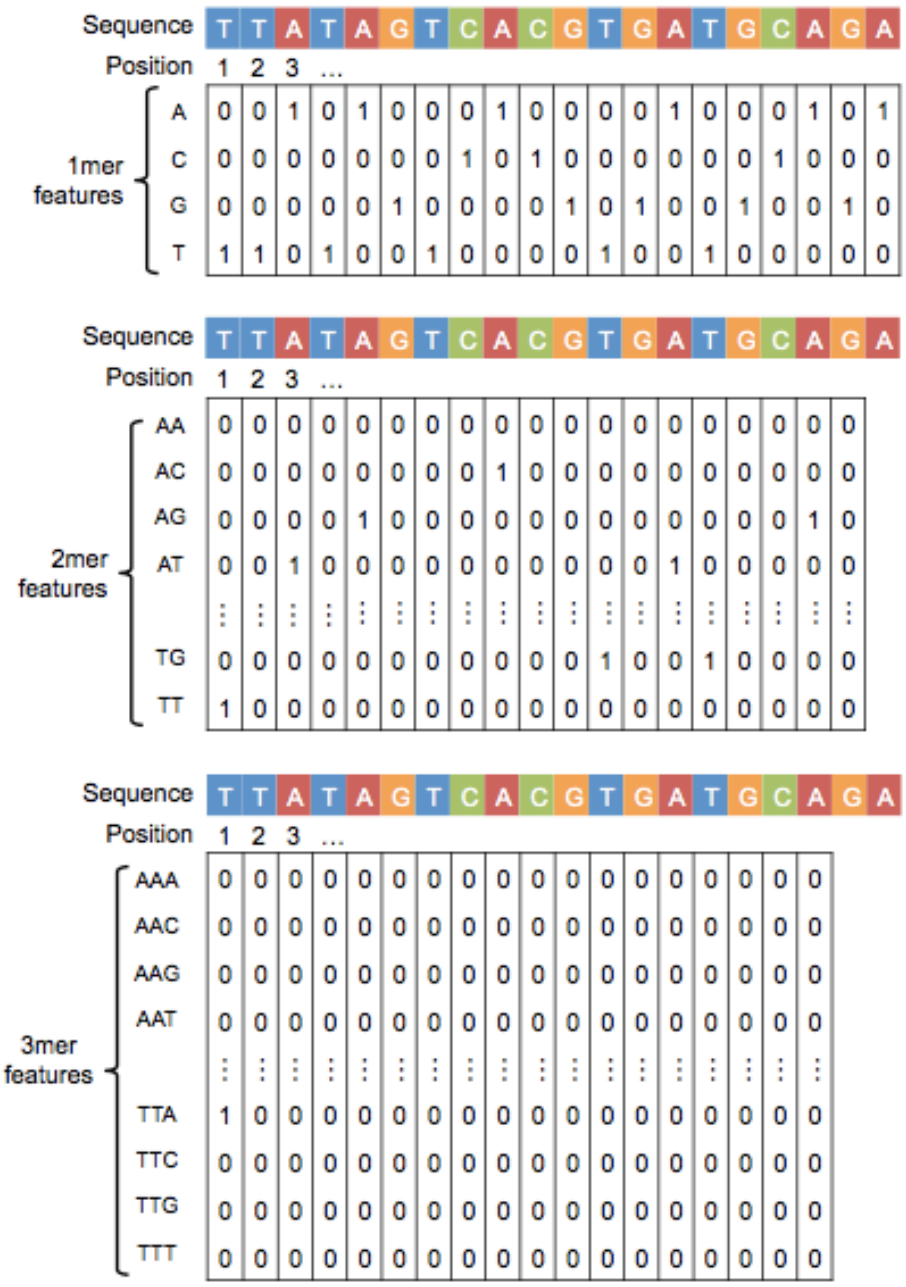


Figure S4. Related to Figure 2. Feature derivation for a 20-bp DNA sequence centered on a putative TF binding site.

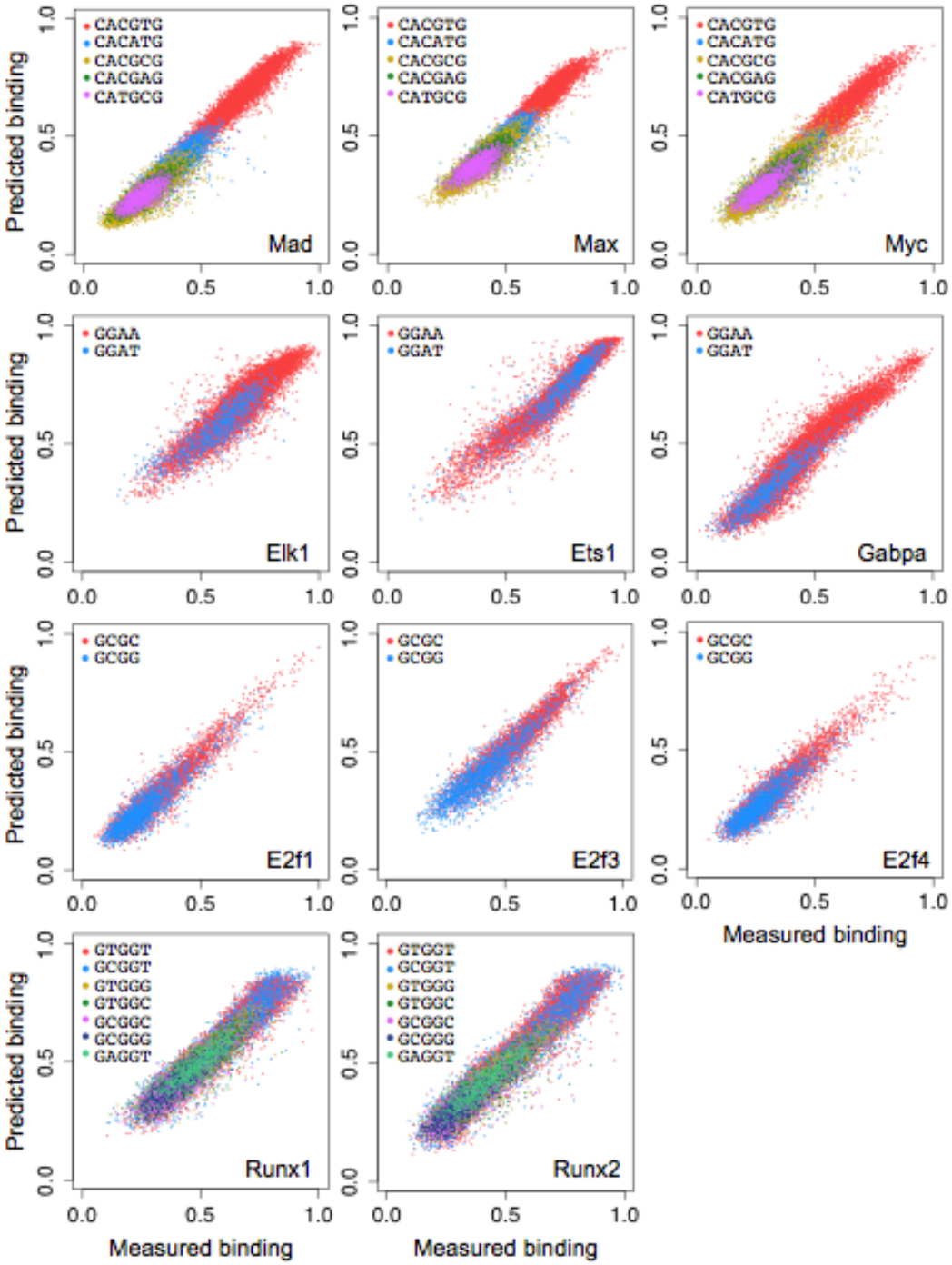


Figure S5. Related to Figure 2. Accuracy of core-stratified SVR models. Scatter plots show measured versus predicted DNA binding scores for sequences not used during for model training. For each TF, sequences containing different cores are shown in different colors.

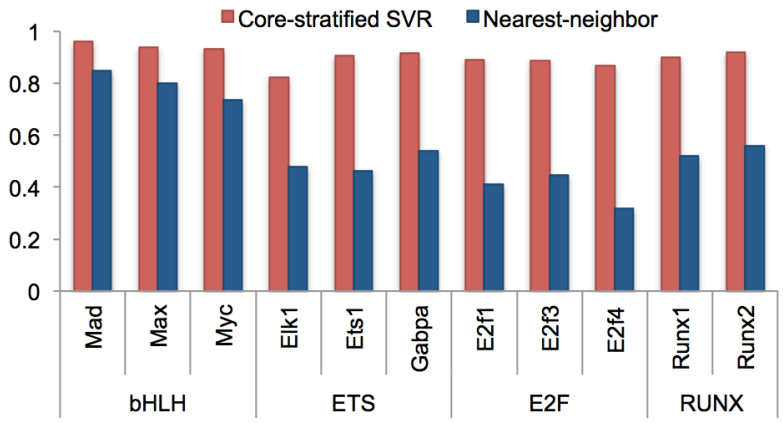


Figure S6. Related to Figure 2. Comparison of the prediction accuracies for core-stratified SVR models versus nearest-neighbor models. Both types of models were evaluated on 20-mer data, split by cores. For each core, embedded 5-fold cross-validation was performed, using the exact same folds (**Table S1**) for both types of models.

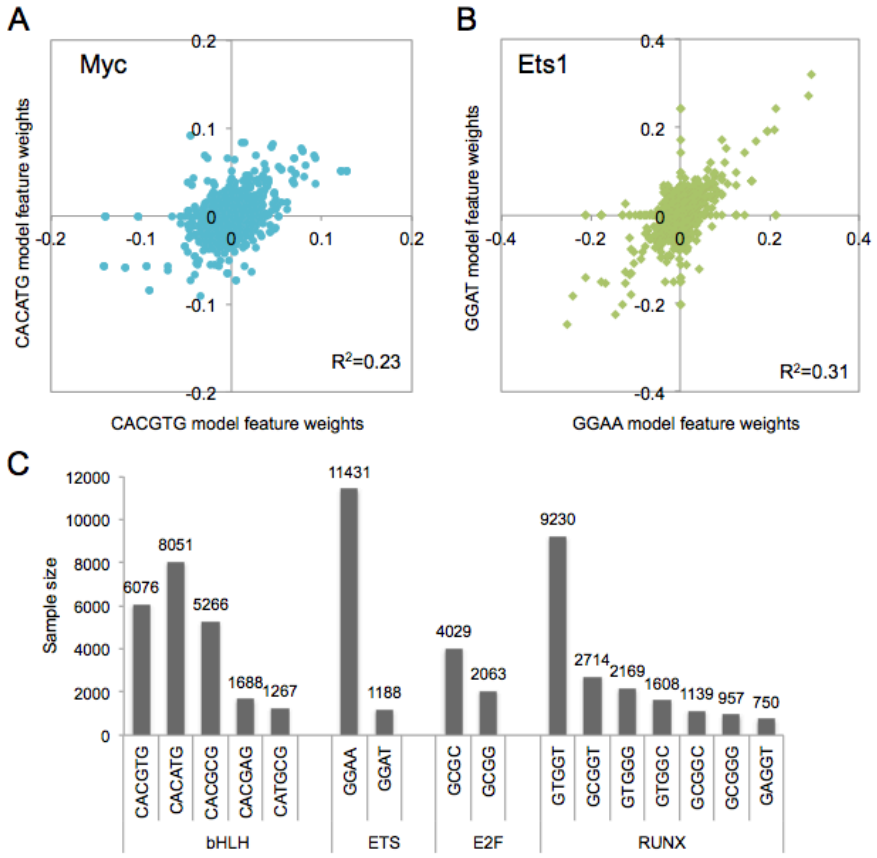


Figure S7. Relate to Figure 2. Results and data supporting our core-stratified approach for modeling TF-DNA binding specificity. (A) For TF Myc, flanking sequence features contribute differently to Myc-DNA binding specificity for sites with the CACGTG versus the CACATG core. Scatter plot shows the feature weights for each 1mer, 2mer, and 3mer features in the flanking regions, according to linear kernel SVR models trained on sequences with each core. **(B)** Similar to (A), for TF Ets1 and core sequences GGAA and GGAT. **(C)** Sample sizes for individual cores in our gcPBM data. Plot shows that we have sufficient data to train a separate SVR model for each core.

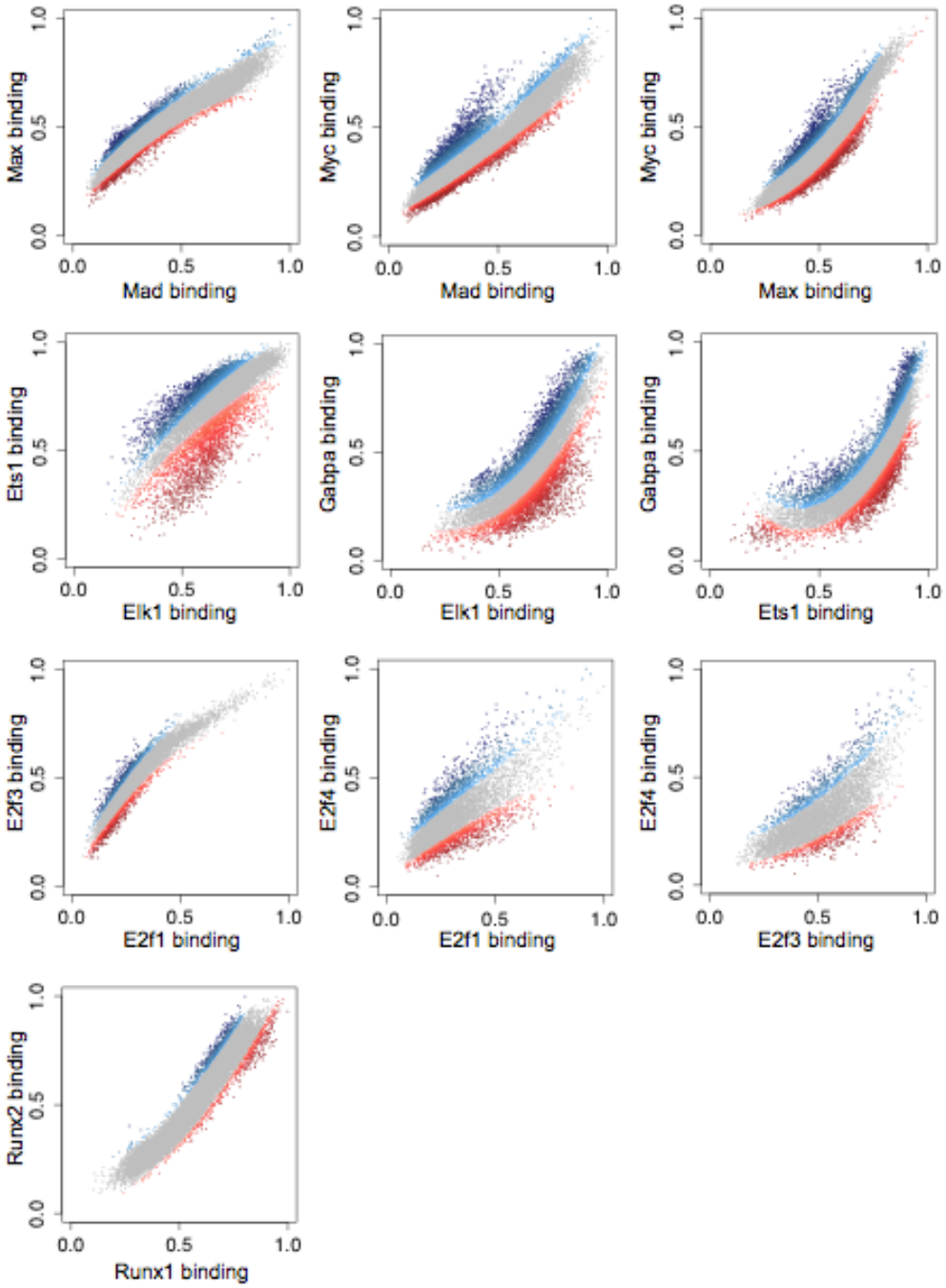


Figure S8. Related to Figure 3. DNA sites differentially preferred by paralogous TFs. Scatter plots show, in grey, the 99% prediction intervals for each pair of TFs, computed using the WLSR model (see **Figure 3** and related text). The sequences inside the prediction interval are not differentially preferred by the paralogous TFs. The sequences outside the interval, shown in red or blue, are preferred by one of the two TFs being compared. Color intensity shows the magnitude of the preference score (see Methods).

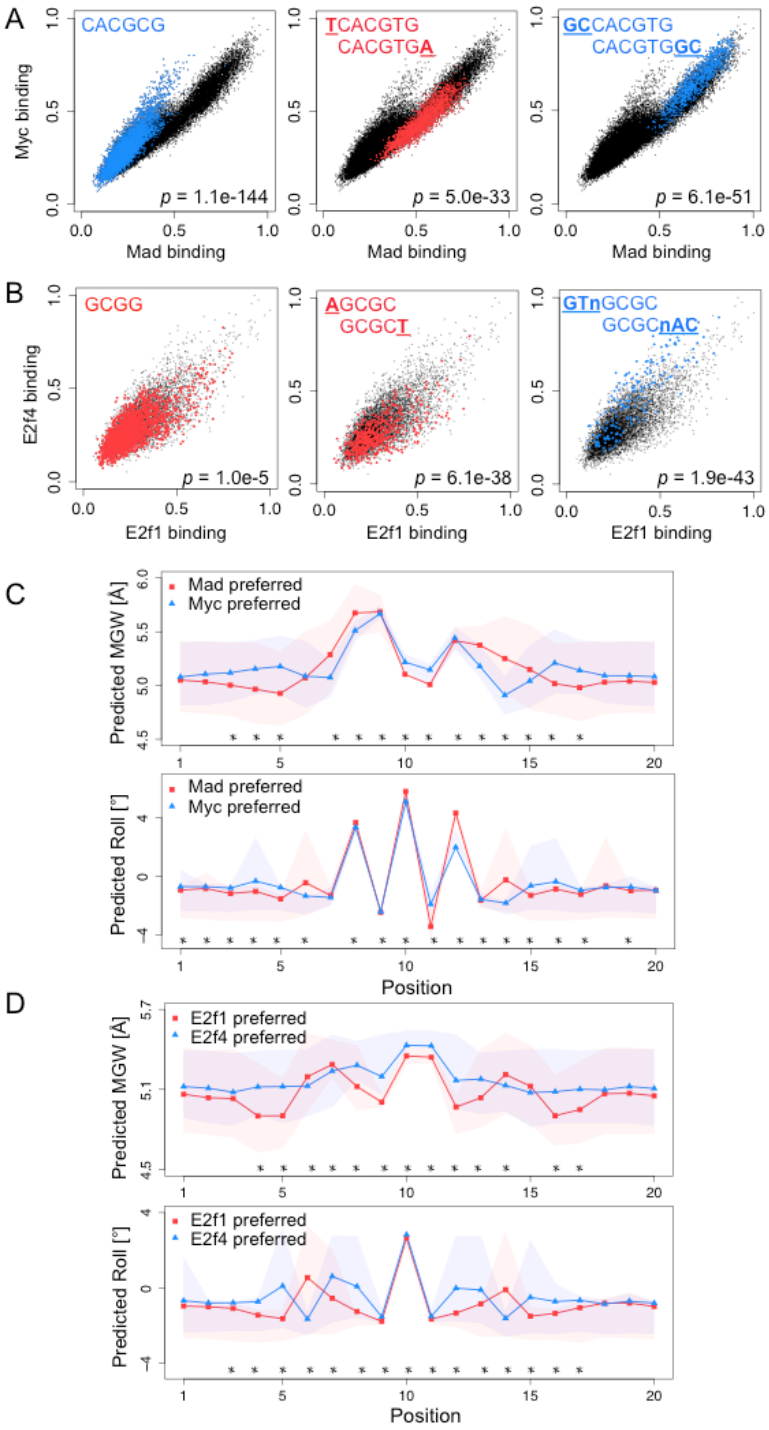


Figure S9. Related to Figure 4. Sequence and shape features of DNA sites differentially bound by paralogous TFs. (A) Scatter plots highlighting the most differentially preferred core, as well as the most differentially preferred 1-mer and 2-mer features for the canonical core CACGTG, for paralogous bHLH factors Mad versus Myc. *P*-value shows the enrichment in Myc-preferred or Mad-preferred sites, according to a Mann-Whitney *U* test (Methods). **(B)** Similar to (A), but for E2F factors E2f1 versus E2f4. **(C)** MGW and roll profiles for genomic sites preferred by Mad vs. Myc. Stars (*) mark the positions within the binding sites (core or flanking region) that are significantly different between the two profiles (*p*-value < 10^{-5} according to Mann-Whitney *U* test). Shaded regions show the 25th-75th percentile ranges at each position. **(D)** Similar to (C), but for E2F factors E2f1 versus E2f4.

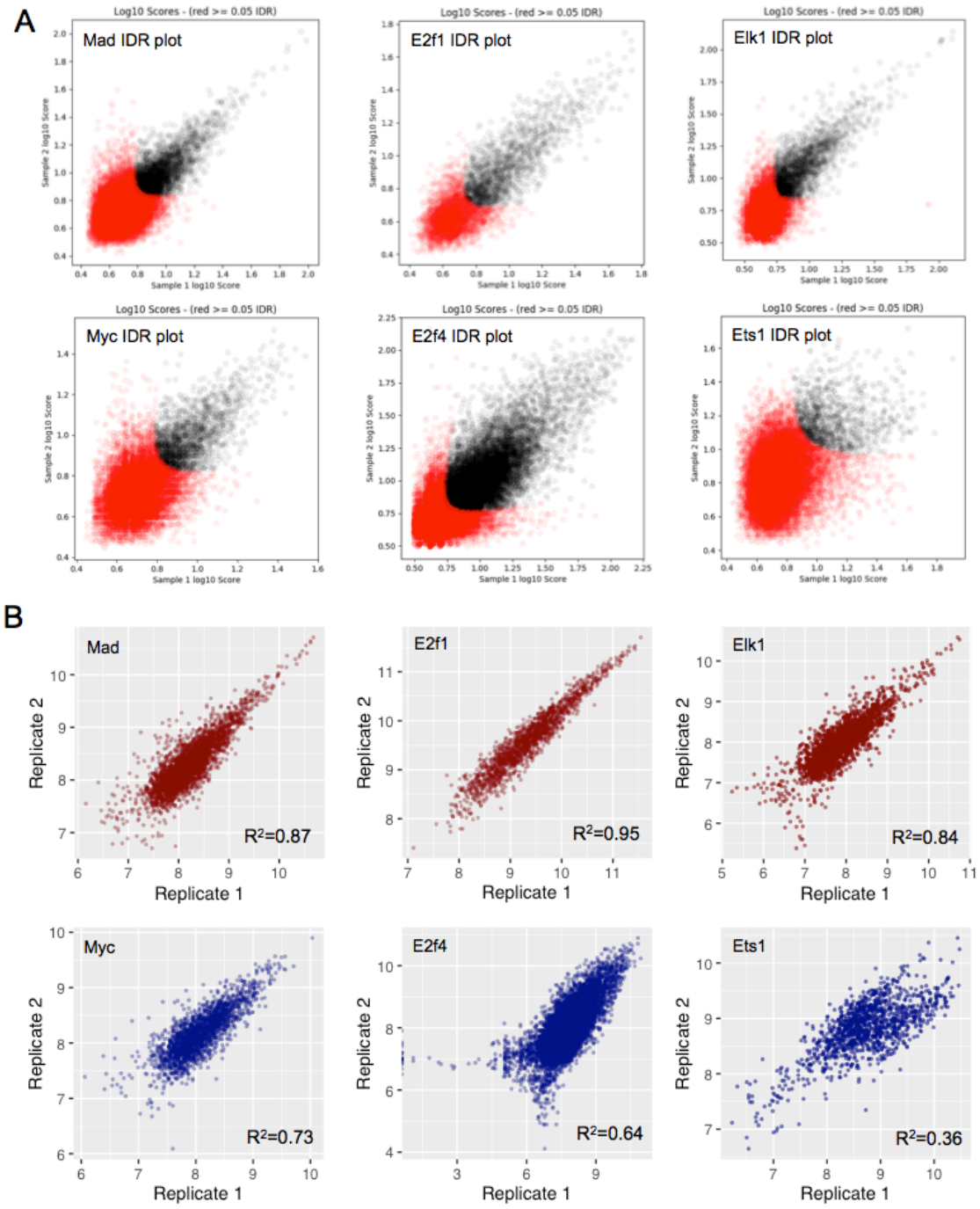


Figure S10. Related to Figure 5. ChIP-seq data quality. (A) Scatter plots illustrate the quality of the ChIP-seq data, as evaluated using the IDR pipeline (Li et al., 2011) (Methods). Plots show the correlation between ChIP-seq replicate data sets, before data normalization. Black points show the "reproducible" peaks identified by the IDR analysis. Red points show the peaks considered "irreproducible". All data sets used in this analysis have high correlation for the reproducible peaks, except for the Ets1 ChIP-seq data (lower-right panel). **(B)** Scatter plots illustrate the agreement between ChIP-seq replicates after data normalization (Methods). Axes show the natural logarithm of the ChIP-seq pileup signal (i.e. number of reads) in the peak regions. All data sets show good correlation, except for the Ets1 ChIP-seq data (lower-right panel).

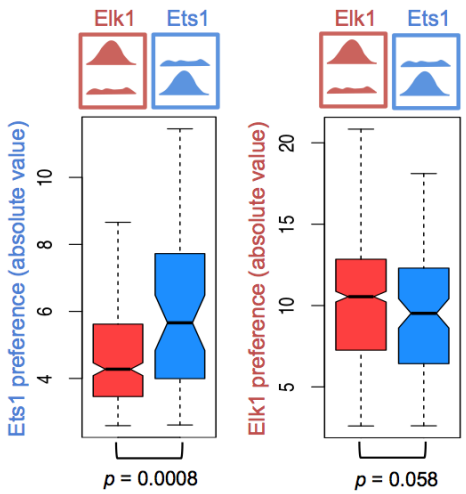


Figure S11. Related to Figure 5. Genomic sites with differential *in vitro* specificity are enriched in differentially bound *in vivo* regions. Left: genomic sites with higher Ets1 preference, as predicted by iMADS, are enriched in Ets1-specific *in vivo* targets (as determined using DESeq2; n=736 and 106, respectively; see Methods). Right: genomic sites with higher Elk1 preference are enriched in Elk1-specific *in vivo* targets (n=318 and 51, respectively). *P*-values were computed using one-sided Man-Whitney *U*-test.

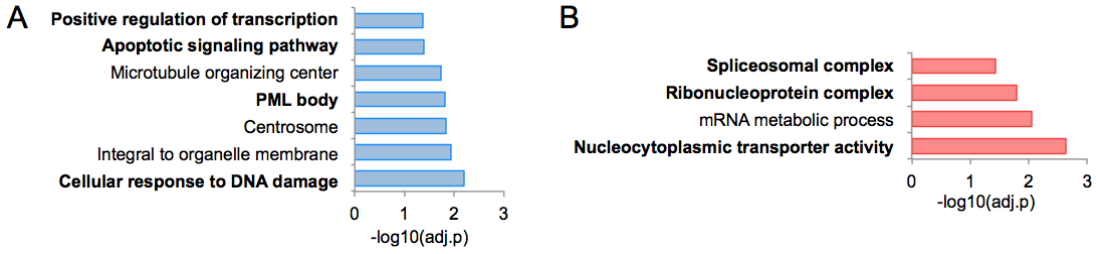
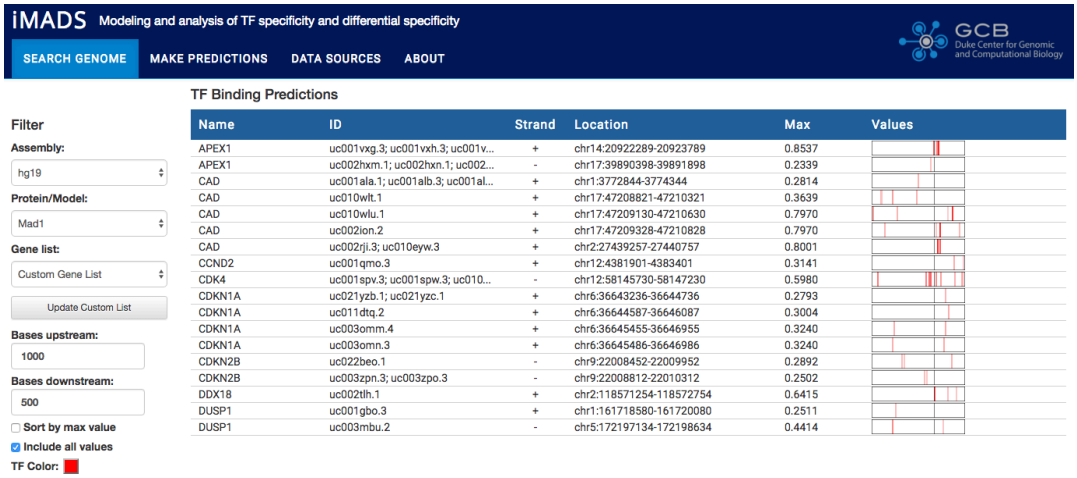


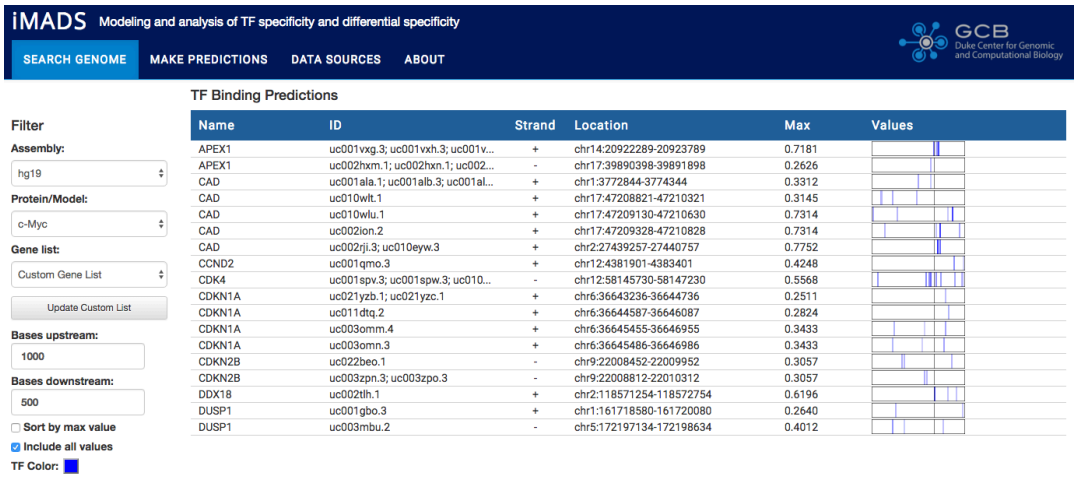
Figure S12. Related to Figure 5. GO enrichment analysis. GO categories enriched in the set of genes with nearby **(A)** Ets1-preferred or **(B)** Ets1-preferred genomic sites. Categories with literature support (Alberstein et al., 2007, Boros et al., 2009, Bories et al., 1995, Teruyama et al., 2001) are highlighted in bold.

A



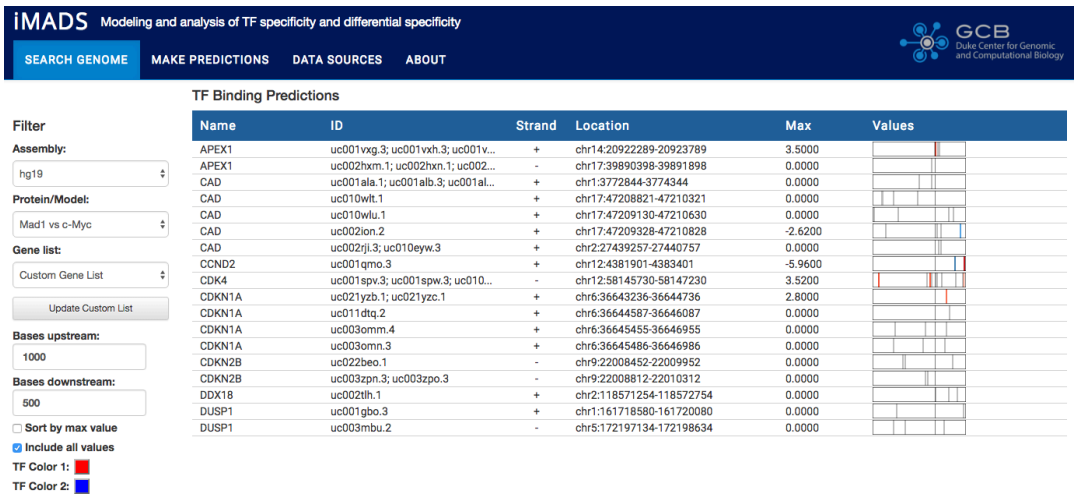
Name	ID	Strand	Location	Max	Values
APEX1	uc001vxg.3; uc001vxh.3; uc001v...	+	chr14:20922289-20923789	0.8537	
APEX1	uc002hxm.1; uc002hxn.1; uc002...	-	chr17:39890398-39891898	0.2339	
CAD	uc001ala.1; uc001alb.3; uc001al...	+	chr1:3772844-3774344	0.2814	
CAD	uc010wlt.1	+	chr17:47208821-47210321	0.3639	
CAD	uc010wlu.1	+	chr17:47209130-47210630	0.7970	
CAD	uc002ion.2	+	chr17:47209328-47210828	0.7970	
CAD	uc002rji.3; uc010eyw.3	+	chr2:27439257-27440757	0.8001	
CCND2	uc001qmo.3	+	chr12:4381901-4383401	0.3141	
CDK4	uc001spv.3; uc001spw.3; uc010...	-	chr12:58145730-58147230	0.5980	
CDKN1A	uc021yzb.1; uc021yzc.1	+	chr6:36643236-36644736	0.2793	
CDKN1A	uc011dtq.2	+	chr6:36644587-36646087	0.3004	
CDKN1A	uc003omm.4	+	chr6:36645455-36646955	0.3240	
CDKN1A	uc003omn.3	+	chr6:36645486-36646986	0.3240	
CDKN2B	uc022beo.1	-	chr9:22008452-22009952	0.2892	
CDKN2B	uc003zpn.3; uc003zpo.3	-	chr9:22008812-22010312	0.2502	
DDX18	uc002tlh.1	+	chr2:118571254-118572754	0.6415	
DUSP1	uc001gbo.3	+	chr1:161718580-161720080	0.2511	
DUSP1	uc003mbu.2	-	chr5:172197134-172198634	0.4414	

B



Name	ID	Strand	Location	Max	Values
APEX1	uc001vxg.3; uc001vxh.3; uc001v...	+	chr14:20922289-20923789	0.7181	
APEX1	uc002hxm.1; uc002hxn.1; uc002...	-	chr17:39890398-39891898	0.2626	
CAD	uc001ala.1; uc001alb.3; uc001al...	+	chr1:3772844-3774344	0.3312	
CAD	uc010wlt.1	+	chr17:47208821-47210321	0.3145	
CAD	uc010wlu.1	+	chr17:47209130-47210630	0.7314	
CAD	uc002ion.2	+	chr17:47209328-47210828	0.7314	
CAD	uc002rji.3; uc010eyw.3	+	chr2:27439257-27440757	0.7752	
CCND2	uc001qmo.3	+	chr12:4381901-4383401	0.4248	
CDK4	uc001spv.3; uc001spw.3; uc010...	-	chr12:58145730-58147230	0.5568	
CDKN1A	uc021yzb.1; uc021yzc.1	+	chr6:36643236-36644736	0.2511	
CDKN1A	uc011dtq.2	+	chr6:36644587-36646087	0.2824	
CDKN1A	uc003omm.4	+	chr6:36645455-36646955	0.3433	
CDKN1A	uc003omn.3	+	chr6:36645486-36646986	0.3433	
CDKN2B	uc022beo.1	-	chr9:22008452-22009952	0.3057	
CDKN2B	uc003zpn.3; uc003zpo.3	-	chr9:22008812-22010312	0.3057	
DDX18	uc002tlh.1	+	chr2:118571254-118572754	0.6196	
DUSP1	uc001gbo.3	+	chr1:161718580-161720080	0.2640	
DUSP1	uc003mbu.2	-	chr5:172197134-172198634	0.4012	

C



Name	ID	Strand	Location	Max	Values
APEX1	uc001vxg.3; uc001vxh.3; uc001v...	+	chr14:20922289-20923789	3.5000	
APEX1	uc002hxm.1; uc002hxn.1; uc002...	-	chr17:39890398-39891898	0.0000	
CAD	uc001ala.1; uc001alb.3; uc001al...	+	chr1:3772844-3774344	0.0000	
CAD	uc010wlt.1	+	chr17:47208821-47210321	0.0000	
CAD	uc010wlu.1	+	chr17:47209130-47210630	0.0000	
CAD	uc002ion.2	+	chr17:47209328-47210828	-2.6200	
CAD	uc002rji.3; uc010eyw.3	+	chr2:27439257-27440757	0.0000	
CCND2	uc001qmo.3	+	chr12:4381901-4383401	-5.9600	
CDK4	uc001spv.3; uc001spw.3; uc010...	-	chr12:58145730-58147230	3.5200	
CDKN1A	uc021yzb.1; uc021yzc.1	+	chr6:36643236-36644736	2.8000	
CDKN1A	uc011dtq.2	+	chr6:36644587-36646087	0.0000	
CDKN1A	uc003omm.4	+	chr6:36645455-36646955	0.0000	
CDKN1A	uc003omn.3	+	chr6:36645486-36646986	0.0000	
CDKN2B	uc022beo.1	-	chr9:22008452-22009952	0.0000	
CDKN2B	uc003zpn.3; uc003zpo.3	-	chr9:22008812-22010312	0.0000	
DDX18	uc002tlh.1	+	chr2:118571254-118572754	0.0000	
DUSP1	uc001gbo.3	+	chr1:161718580-161720080	0.0000	
DUSP1	uc003mbu.2	-	chr5:172197134-172198634	0.0000	

Figure S13. Related to Figure 5. Screen shots illustrating the use of the iMADS web server to analyze gene promoters. (A) Using iMADS to identify Mad binding sites in the genomic regions 1000-bp upstream and 500-bp downstream of the transcription start sites of eight genes known to be regulated by Myc (Zeller et al., 2003). **(B)** Similar to (A), but showing Myc binding sites. **(C)** Similar to (A), but showing preference scores reflecting the differential DNA binding specificity between Mad and Myc.

Step1. Identify variant of interest

rs786205688 (chr5:74893909)

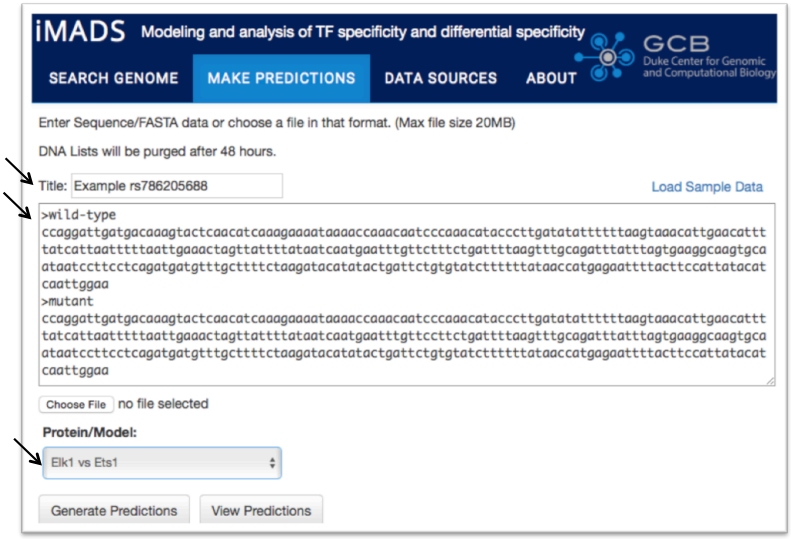
Step2. Extract genomic region around the variant

To obtain the 300-bp regions centered around rs786205688, we can use
http://genome.ucsc.edu/cgi-bin/das/hg19/dna?segment=chr5:74893759,74894059

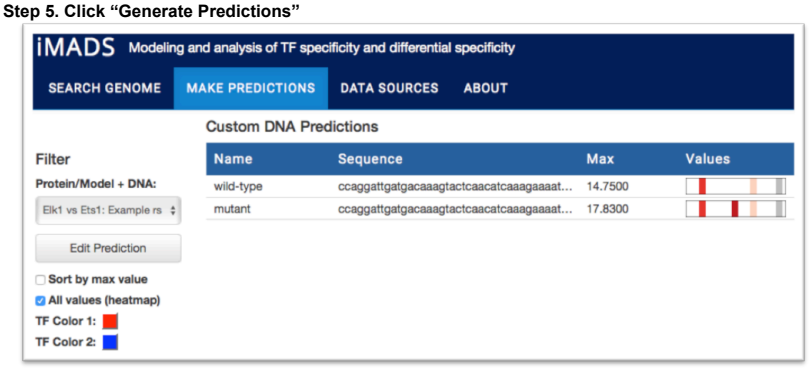
Step3. Generate the sequences of interest in a fasta format

>wild-type
ccaggattgatgacaaagtactcaacatcaaagaaaataaaaccaaacaatcccaaacataccttgatatatttttaagtaaacattgaacattttatcattaattttta
attgaaactagttatttataatcaatgaatttgttctttctgattttaagtttgcagatttatttagtgaaggcaagtgcaataatccttcctcagatgatgtttgcttttctaagata
catatactgattctgtgtatctttttataaccatgagaattttacttccattatacatcaattggaa
>mutant
ccaggattgatgacaaagtactcaacatcaaagaaaataaaaccaaacaatcccaaacataccttgatatatttttaagtaaacattgaacattttatcattaattttta
attgaaactagttatttataatcaatgaatttgttccttctgattttaagtttgcagatttatttagtgaaggcaagtgcaataatccttcctcagatgatgtttgcttttctaagata
catatactgattctgtgtatctttttataaccatgagaattttacttccattatacatcaattggaa

Step 4. Input sequences of interest into the iMADS PREDICTION tool, set the title for the job (e.g. "Example rs786205688"), and set the model



Step 5. Click "Generate Predictions"



Step 6. Click on the prediction heatmap in the "Values" column to see details of the predictions

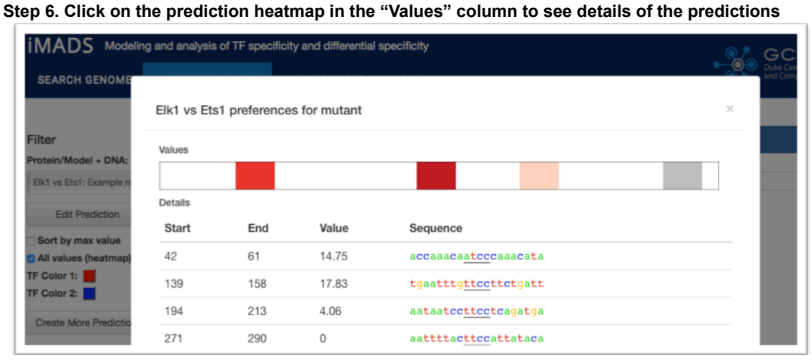


Figure S14. Related to Figure 6. Example of iMADS analysis of non-coding variants.

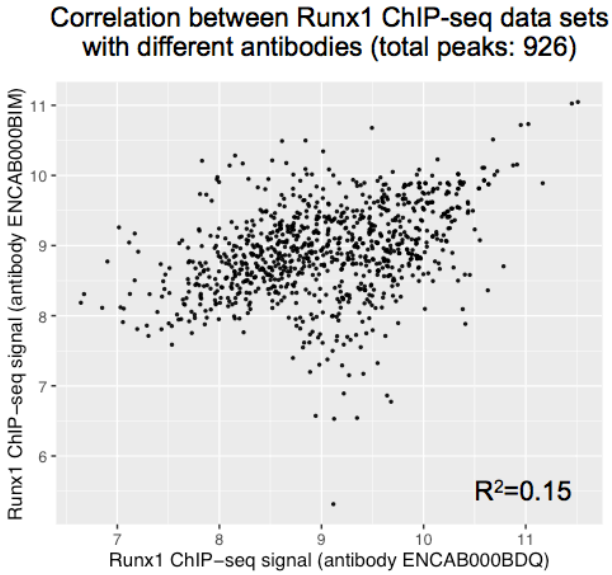


Figure S15. Related to Figure 5. Agreement between ChIP-seq data sets obtained for different antibodies. Among the 11 proteins tested in our study, Runx1 is the only protein for which ChIP-seq data is available, in the same cell line (K562), for 2 different primary antibodies (ENCAB000BIM and ENCAB000BDQ) that both gave ChIP-seq data of acceptable quality according to ENCODE guidelines. The total number of peaks, after combining the two ChIP-seq experiments, is 926. R^2 = squared Pearson correlation coefficient**.**