Sequence characteristics distinguish transcribed enhancers from promoters and predict their breadth of activity

Supplementary Material

Laura L. Colbran,* Ling Chen† and John A. Capra*,†,‡,1
* Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA
† Department of Biological Sciences, Vanderbilt University, Nashville, TN, 37235, USA
‡ Center for Structural Biology, Departments of Biomedical Informatics and Computer Science, Vanderbilt University, Nashville, TN, 37235, USA

SUPPLEMENTARY TEXT

GC and CpG content are associated with the activity of regulatory regions

To further investigate the sequence differences between broadly and narrowly active promoters and enhancers, we performed principle components analysis (PCA) on the 6-mer spectra of the regions. Both promoter and enhancer sequences cluster based on their breadth of activity and show similar relative relationships (Figure S16). However, enhancers have much greater overlap when projected on the first two PCs; this reflects the greater difficulty of distinguishing broadly from narrowly active enhancers compared to broadly vs. narrowly active promoters, and suggests that relatively simple sequence features may distinguish promoter activity. To explore this, we examined the PCs for the promoter and enhancer PCA plots. We found that both GC content and CpG content are strongly correlated with the PCs (Figure S15). Specifically, GC content had a Spearman's rho of 0.95 with PC1 from the enhancer PCA and 0.92 for the promoter PC1. The correlation of GC content with PC2 was smaller in magnitude for both enhancers and promoters (ρ = –0.22, 0.31, respectively; $P < 2.2E{-}16$ for all). Since GC and CpG content are correlated with each other, we calculated a semipartial Spearman correlation (sr_s^2) for CpG content. For both enhancers and promoters, CpG content explains little variance in PC1 beyond that explained by GC content ($sr_s^2 = 0.028$, $P = 1.2E{-}26$; $sr_s^2 = 0.076$, $P = 5.3E{-}27$, respectively). In contrast, CpG content explains significant variance in PC2 beyond that explained by GC content for both enhancers and promoters ($sr_s^2 = 0.49, 0.44$, respectively, $P < 2.2E{-}308$ for both).

The presence of CpG islands (CGIs) is associated with broad promoter activity [27–30]. To more directly investigate the contribution of CGIs to the ability to identify promoters and enhancers and predict their breadth of activity, we stratified them into CGI-containing and Non-CGI-containing regions based on the *cpgIslandExt* track from the UCSC Genome Browser. This yielded 15,291 CGI promoters, 12,008 Non-CGI promoters, 1,590 CGI enhancers, and 36,948 Non-CGI enhancers.

PCA on the stratified sets revealed that, while there is significant overlap between enhancers and promoters in PCA space, they cluster most strongly by region type rather than by activity or CGI status (Figure S16). Despite their similar characteristics, this supports the presence of sequence differences between promoters and enhancers, even in the context of the first two PCs. Enhancers and promoters both separated by CGI status rather than activity in PCA space (Figure S15). However, Non-CGI promoters

overlapped more with CGI promoters than did Non-CGI enhancers. This suggests that CpGs play a more active role in Non-CGI promoters than in Non-CGI enhancers.

CpG islands influence classifier performance and generalization

We next applied the SVM breadth of activity classification approach to the CGI-stratified region sets. The classifiers were still able to accurately distinguish broadly active promoters and enhancers from narrowly active regions and the genomic background, but the presence of CGIs modulated performance. For example, the promoter classifiers performed extremely well at distinguishing CGI-containing promoters from genomic background regions, even when matched for GC content (ROC AUCs = 1.0 and 0.95, respectively; Figures S17 and S18). The Non-CGI promoter classifiers trained against genomic background and GC-matched regions also performed very well, but not as strongly as for the CGI promoters (ROC AUCs = 0.96 and 0.91). However, distinguishing broadly active CGI promoters from narrowly active promoters is more challenging (ROC AUC = 0.84), presumably because both positive and negative regions contain CGIs by definition. In contrast, distinguishing broad and narrow Non-CGI promoters was easier than for those that contained CGIs (ROC AUC = 0.95, Figures S17 and S18). The general trends observed for promoters were also true for the enhancer classifiers, but performance was lower for all tasks compared to the corresponding promoter classifiers (Figures S17 and S18). Together, these results suggest that, as expected, CGIs contribute to the ability of our classifiers to predict both broadly active promoters and enhancers; however, other sequences play a significant role in making accurate classifications.

Next, we performed the cross-region evaluation using models trained on the CGI-stratified region sets. As for the non-stratified models, the classifiers trained on enhancers and applied to promoters performed better on every group than when the promoter-trained classifiers were evaluated on enhancer prediction tasks (Figure S6).

CpG dinucleotides are important for promoter activity, but less so for enhancers

Motivated by the role of GC and CpG content in the differences in sequence patterns in promoters and enhancers, we quantified their contributions to the prediction models learned by the SVM classifiers. We analyzed the weights assigned to the 6-mer motifs stratified by GC content. The weights assigned to individual 6-mers in both promoter and enhancer classifiers versus genomic background had a positive correlation with 6-mer GC content (Spearman's $\rho = 0.42$ and 0.31 respectively, $P < 2.2E{-}16$ for both). The positive correlation remained when we split the sets by CGI status (Figures S19 and S20). However, while the positive correlation between 6-mer weight and GC content was maintained in the Non-CGI region classifiers trained against narrow regions, this was not the case for the CGI region classifiers (Figure S8). This suggests that while high GC content is generally indicative of broadly active regulatory regions, in regions with CGIs, this is mainly due to the CGIs themselves.

Since CGI status is a binary classification based on a threshold that could potentially mask subtler effects of CpG dinucleotides, we also analyzed the relationship between the number of CpGs present in 6-mers and their weight. In CGI and Non-CGI promoter classifiers trained against the genomic background, CpG count explained a large amount of variance in 6-mer weight (Figures S8 and S19; $R^2 = 0.46$ for CGI and

0.31 for Non-CGI; $P < 2.2E{-}16$ for both). On the other hand, in enhancers, CpG content explained notable variance for CGI enhancers (Figures S8 and S20; $R^2 = 0.37$, $P < 2.2E{-}16$), but not Non-CGI enhancers ($R^2 = 0.050$, $P < 2.2E{-}16$). This suggests that outside the small number of CGI enhancers, CpG content is less important for enhancer activity. The greater importance of the CpG-containing 6-mers in Non-CGI promoters compared to Non-CGI enhancers also held for the classifiers trained on broad vs. narrow regions (Figure S9; $R^2 = 0.23$ vs. 0.070, $P < 2.2E{-}16$ for both).

SUPPLEMENTARY FIGURES

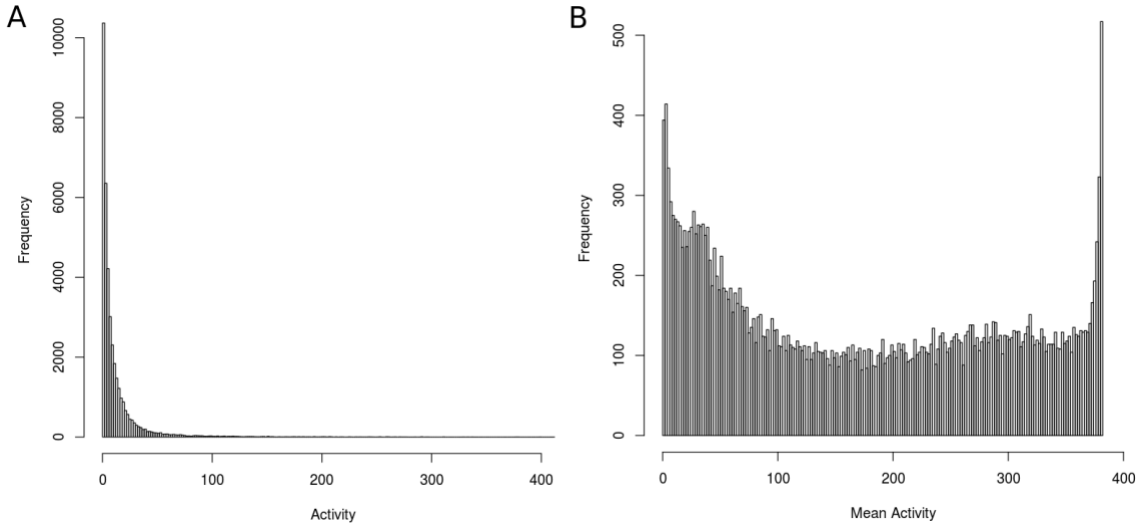

Figure S1. Distribution of activity across all A) enhancers and B) promoters from all 411 contexts in FANTOM. Where multiple promoters overlapped, we took the mean of their activity.

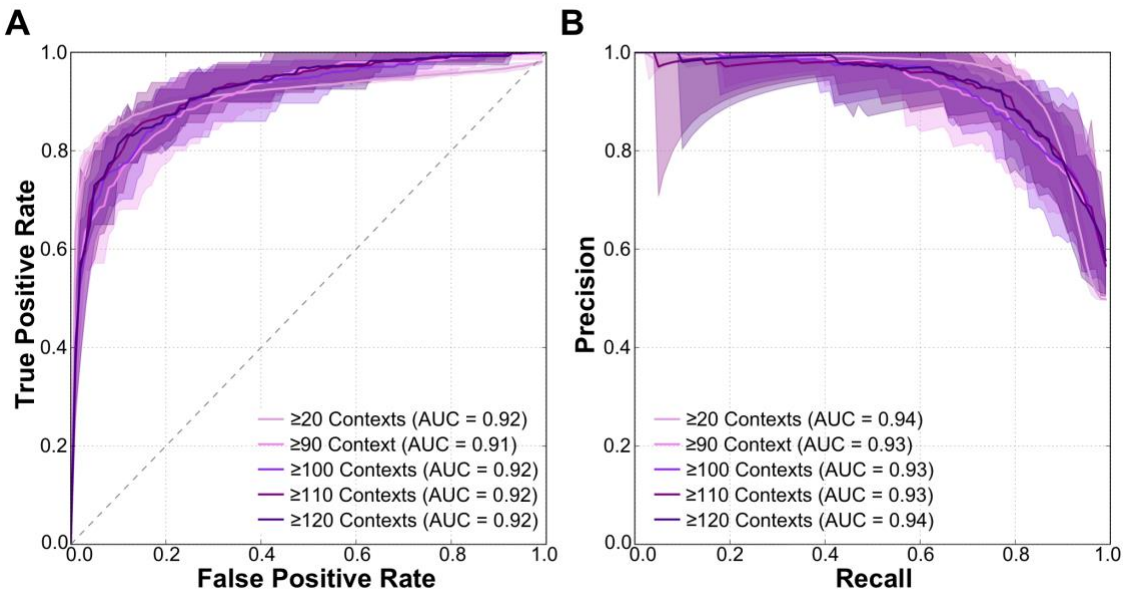

Figure S2. A) ROC curves and B) PR curves of an SVM classifier trained using all possible 6-mers as features. The positive training sets were broadly active CAGE enhancers defined as those active in at least 90, 100, 110, or 120 contexts, while the negative training set for each was an equal number of random length-matched non-enhancer genomic regions. The solid curves represent the mean over 10-fold cross-validation. Shaded regions represent the minimum and maximum curves.

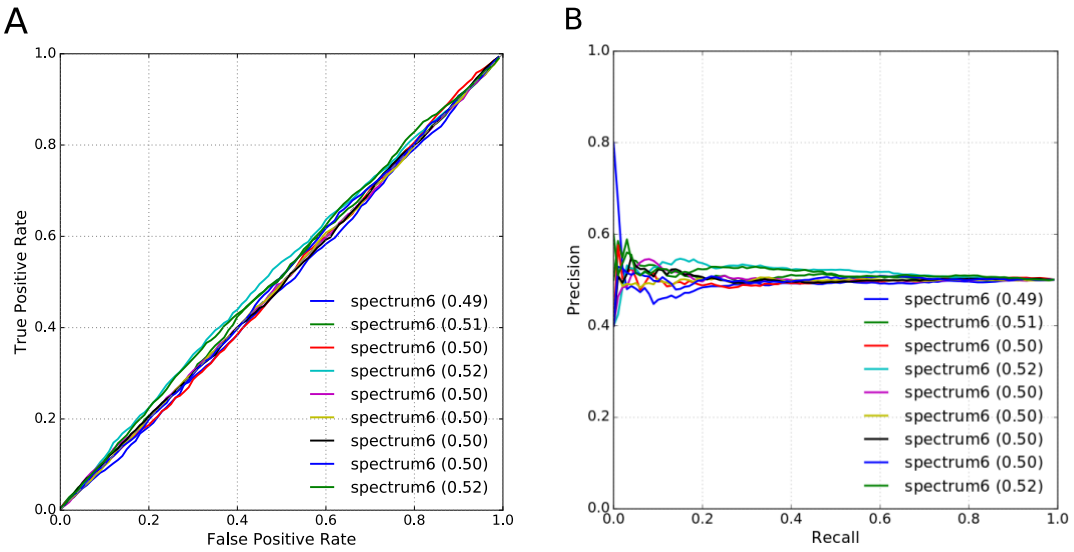

Figure S3. A) ROC curves and B) PR curves for 9 sets of randomized sets. Sets consist of the same regulatory regions as in the enhancer vs. promoter classifiers, but the enhancer/promoter labels were scrambled.

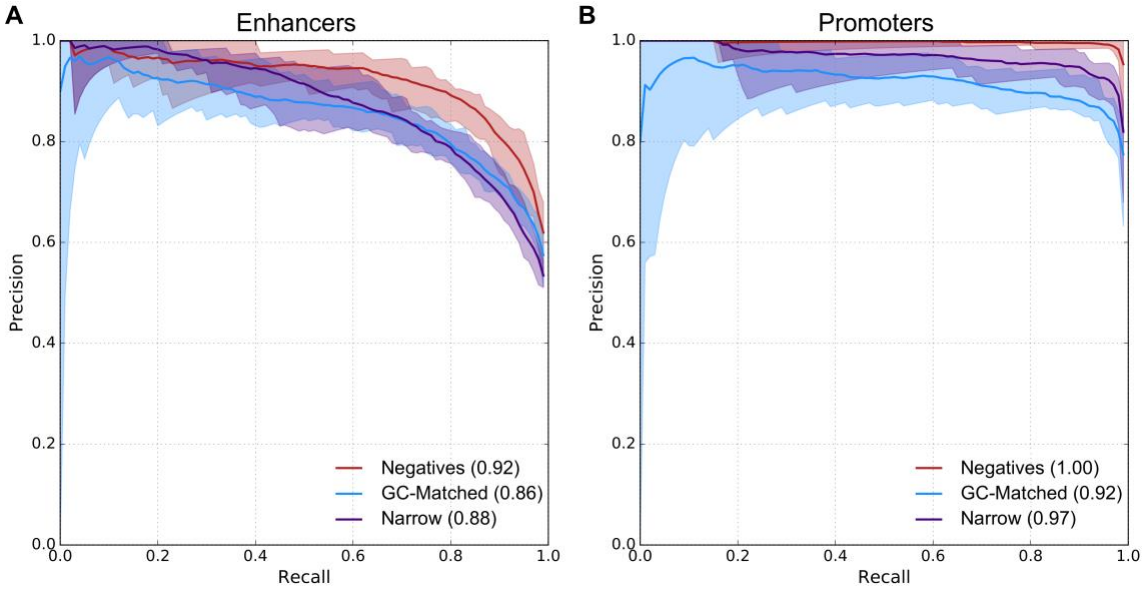

Figure S4. PR curves for (A) Enhancer and (B) Promoter SVM classifiers in Figure 2. Classifiers were trained against length-matched genomic background regions (red), GC-matched genomic background regions (blue), or corresponding narrowly-active regions (purple).

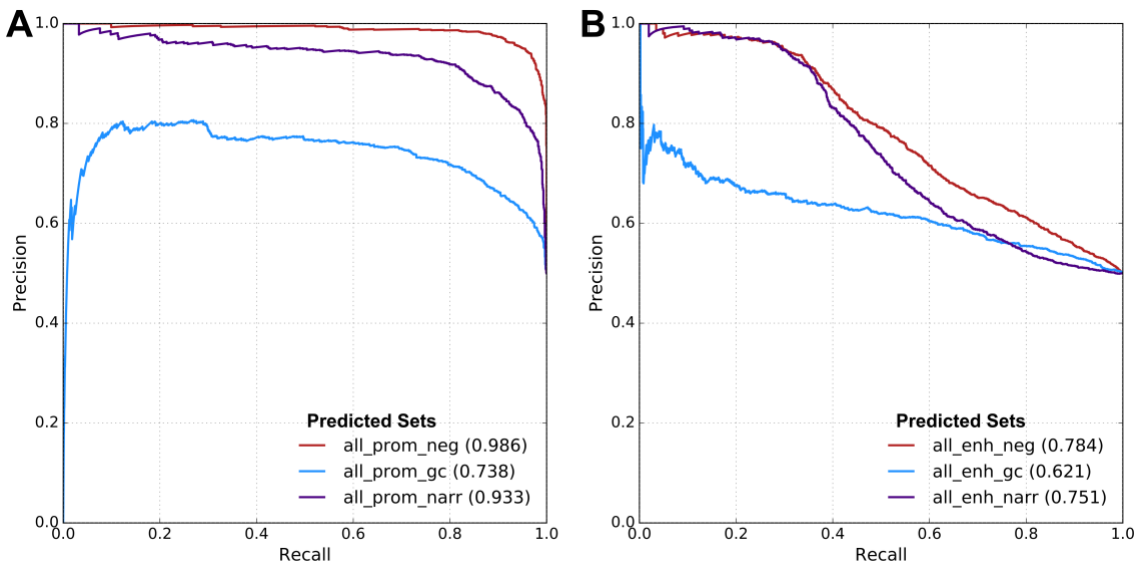

Figure S5. PR curves for the classifiers in Figure 3. Classifiers were trained on (A) broadly active enhancers and then used to classify the corresponding set of broadly active promoters, and (B) trained on broadly active promoters to predict broadly active enhancers. Three different negative sets were considered for each set: genomic background (red), GC-matched background (blue), or narrow regions (purple).

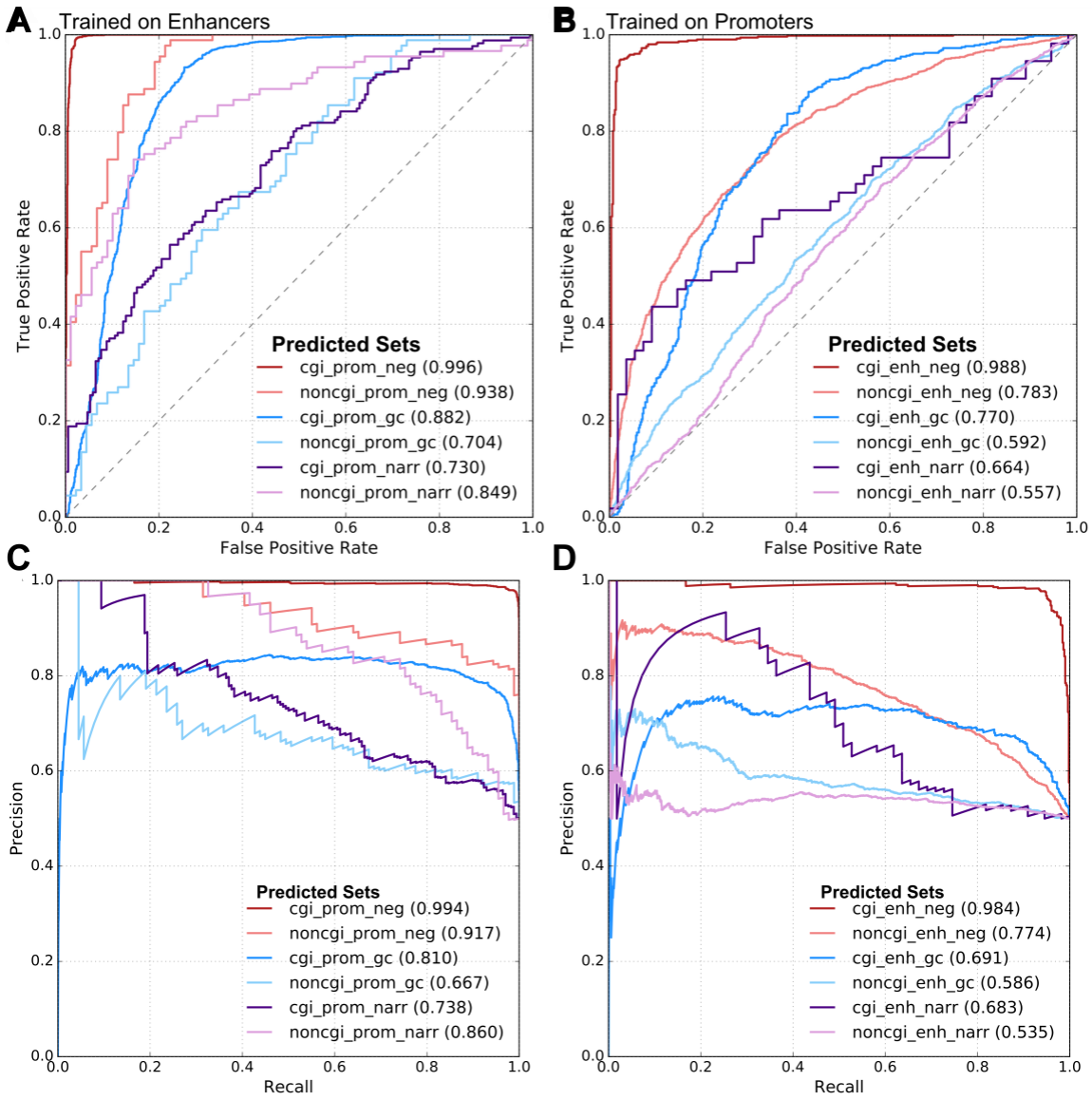

Figure S6. Cross-region prediction performance curves for regions with CGI and Non-CGI combined. For (A) and (C) models trained on enhancers were used to classify the corresponding set of promoters, and for (B) and (D) trained on promoters to predict enhancers. Classifiers were trained on CGI regions (dark) or Non-CGI regions (light) against genomic background (red), GC-matched background (blue), or narrow regions (purple.

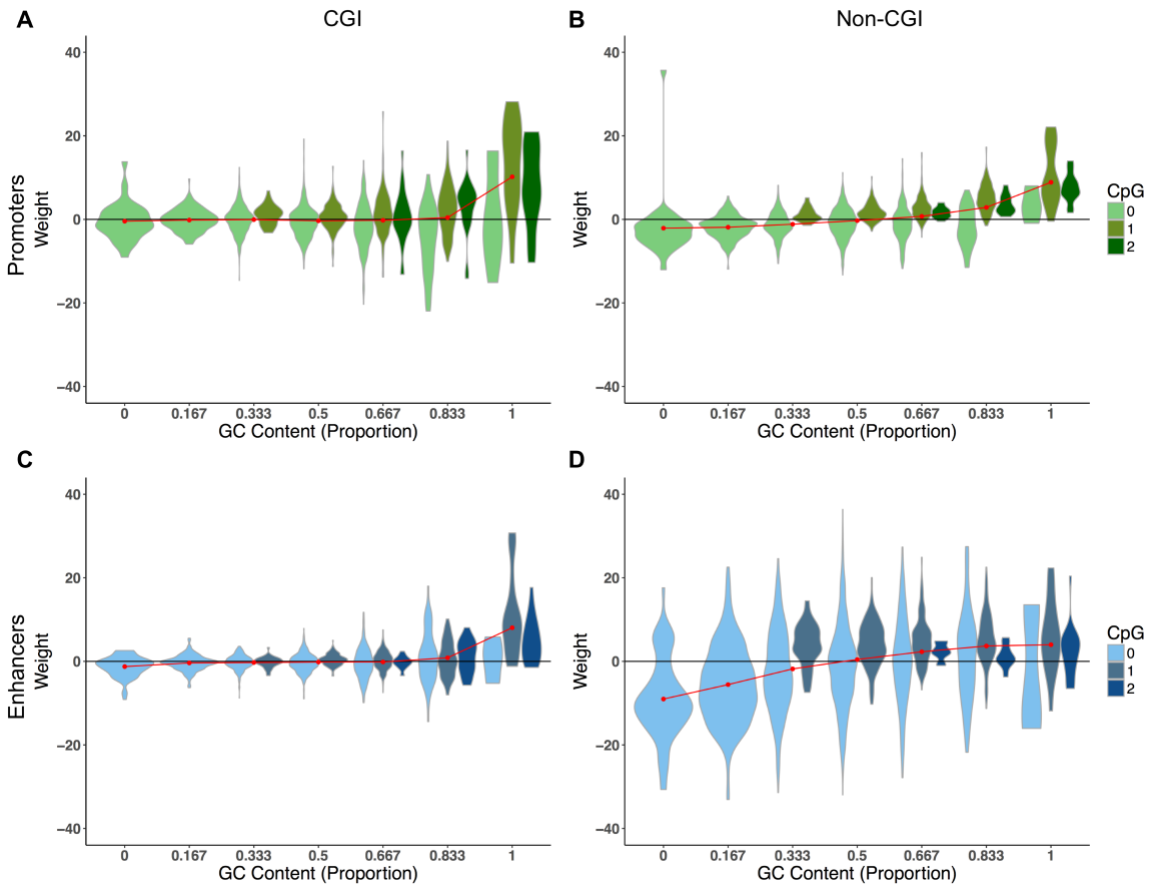

Figure S7. Weight assigned each 6-mer in the models trained with broad vs. narrow (A) CGI promoters, (B) Non-CGI promoters, (C) CGI enhancers, and (D) Non-CGI enhancers, plotted against GC content and stratified by the number of CpG dinucleotides in each 6-mer. 0, 1 and 2-3 CpGs per 6-mer are possible (light to dark). The red points mark the mean weight for each GC content bin. Spearman rho for GC content vs. weight: 0.097 CGI Broad/Neg (P = 5.2E-10); 0.28 Non-CGI Broad/Neg (P < 2.2E-16); -0.147 CGI Broad/GC (P < 2.2E-16); -0.0073 (P = 0.64).

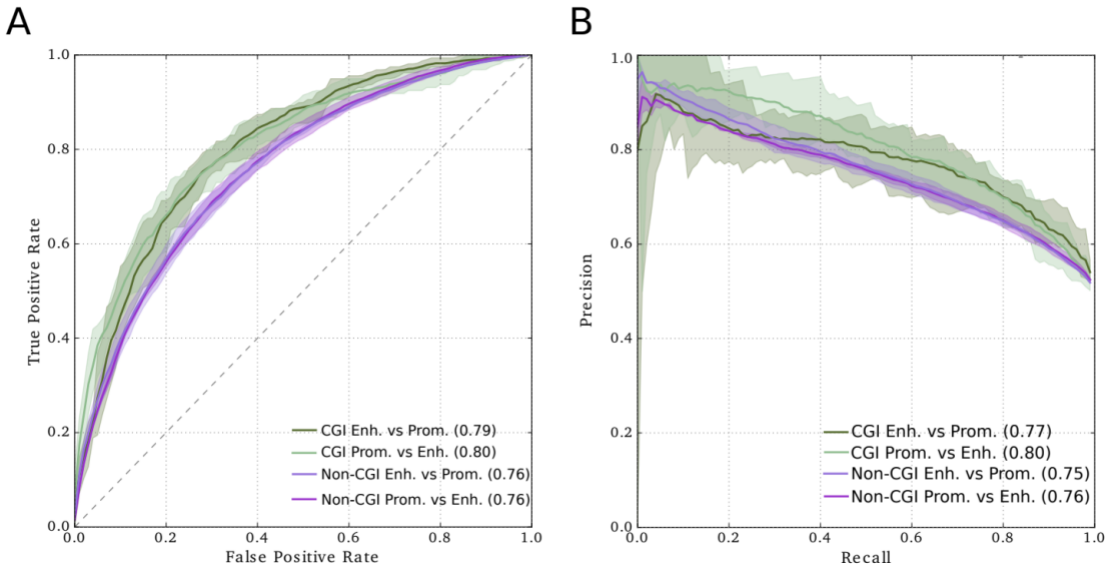

Figure S8. (A) ROC and (B) PR curves for SVM classifiers distinguishing all enhancers and promoters, stratified by overlap with CGIs.

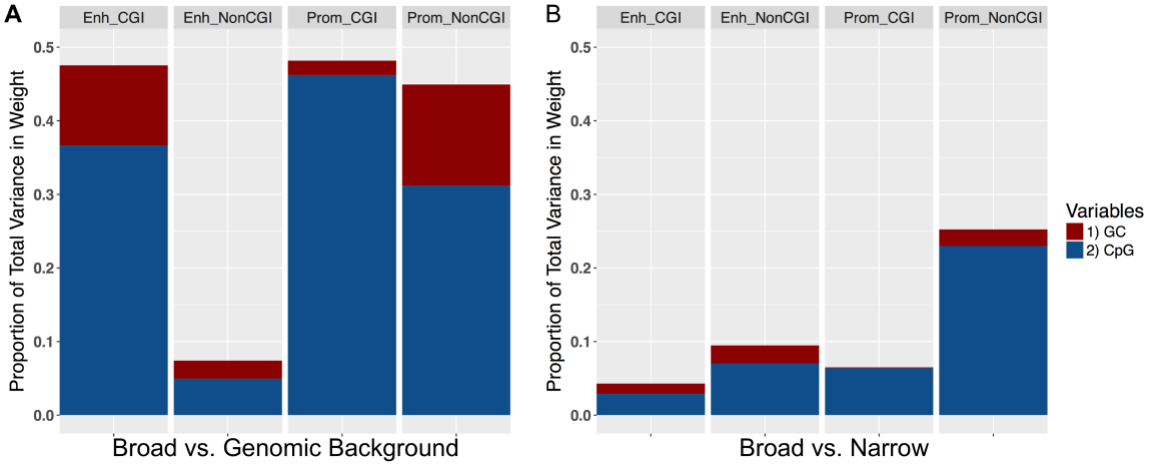

Figure S9. The number of CpG dinucleotides in a 6-mer is more strongly associated with its predictive weight in promoters more than enhancers. (A) We computed the proportion of total variance in 6-mer weights explained by CpG count (blue) and further explained by GC count (red) for classifiers trained to distinguish broadly active regulatory regions from genomic background stratified by the presence or absence of CGIs. The variance explained by CpG content is quantified as the Pearson R^2, and for GC content as the squared semipartial Pearson correlation (sr_p^2). 6-mer CpG count is strongly correlated with predictive weight for both CGI and Non-CGI promoters; however, it is only strongly associated with weight for the CGI enhancers, which make up 4% of all enhancers. (B) The same analyses on classifiers trained to distinguish broad vs. narrow regions. CpG count continued to be more strongly correlated with weight in promoters than enhancers. All correlations were significant ($P < 4E{-}14$) except the GC sr_p^2 for CGI promoters ($P = 0.025$). Weight distributions are plotted in Figure S16.

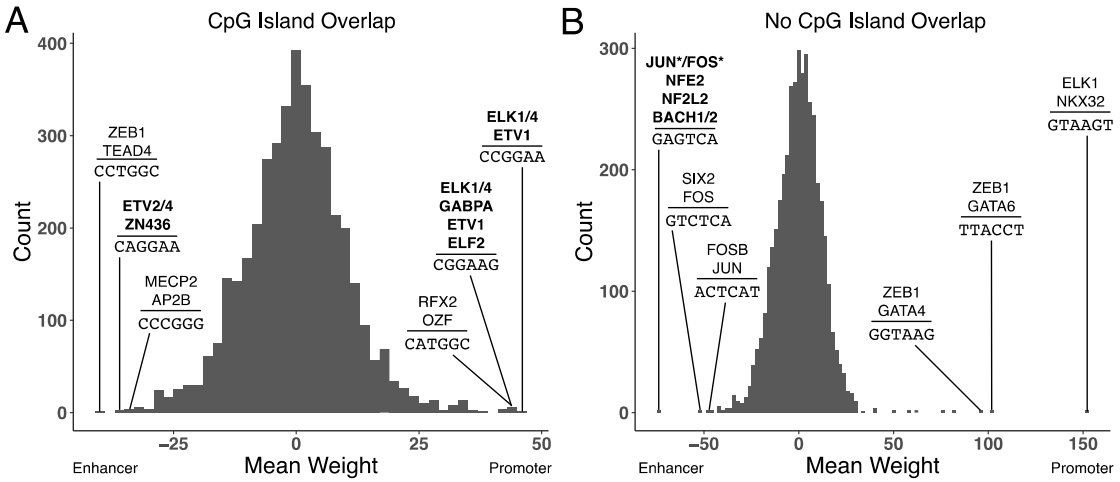

Figure S10. Distribution of weights assigned to 6-mers for promoter vs. enhancer classifiers split by (A) CGI and (B) Non-CGI Status. The most extreme 6-mers are annotated with the top matching TF motifs using the HOCOMOCO v11 CORE database.

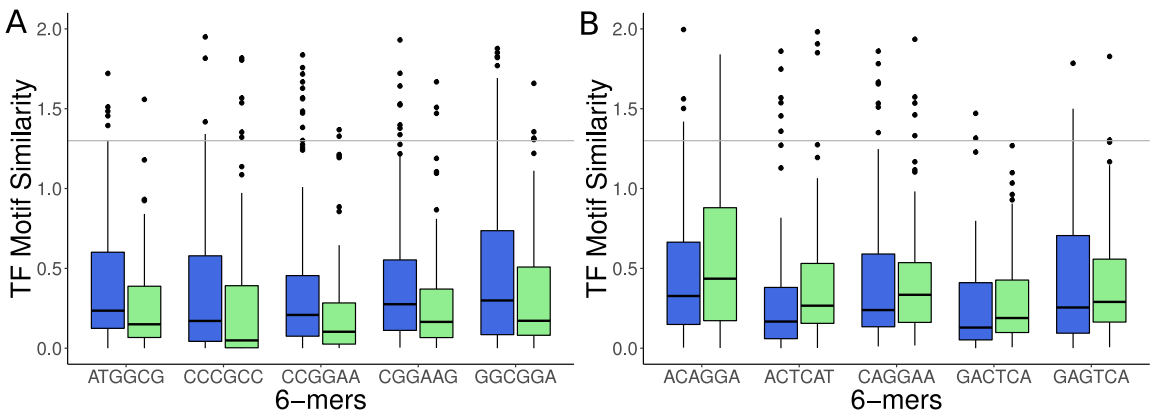

Figure S11. (A) The most highly weighted DNA 6-mers by the SVMs for distinguishing broadly active promoters from narrowly active promoters match different sets of TFs. Sequence patterns predictive of broadly active promoters are enriched for similarity to broadly active TF motifs (blue), in particular ETS family members (CCGGAA, CGGAAG, GGCGGA). (B) Same as (A), but for SVMs trained to distinguish broadly active enhancers from narrowly active enhancers. These are the same plots as Figure 5B and 5C, respectively, with outliers added.

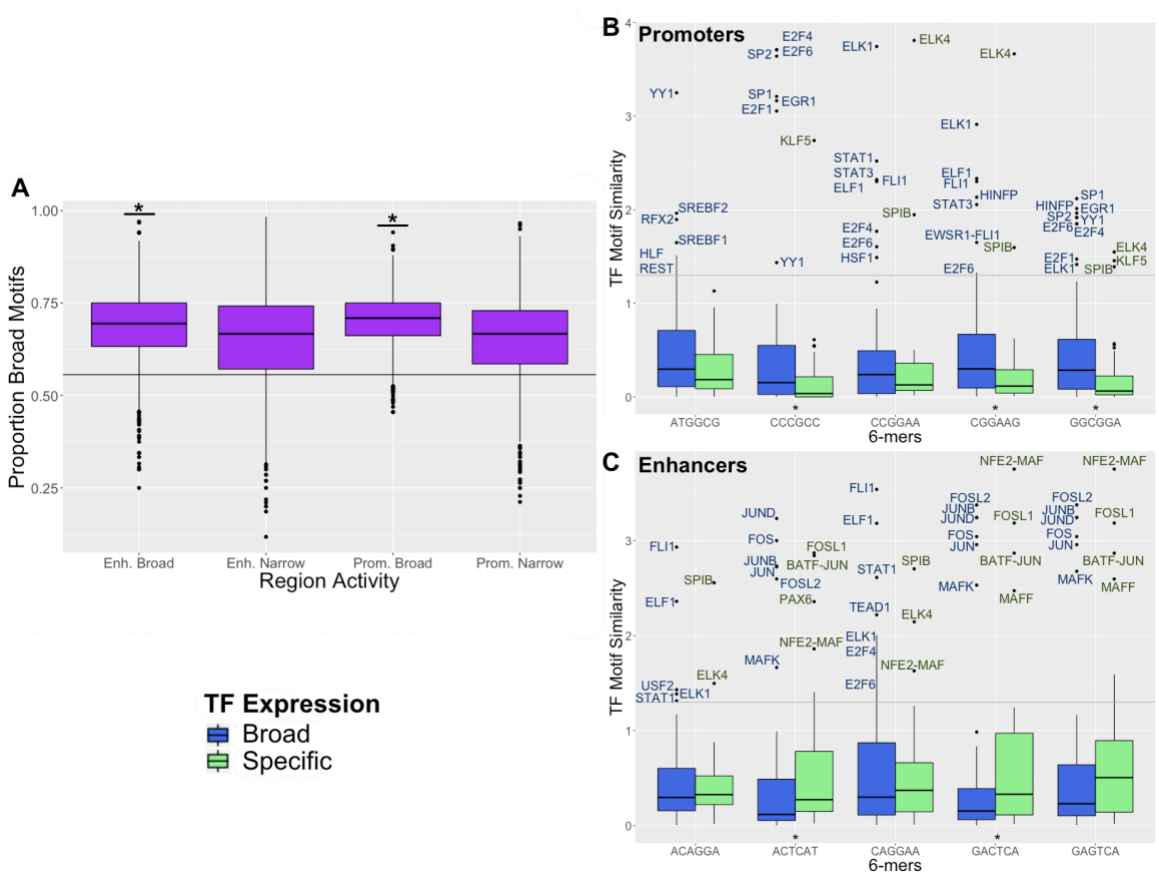

Figure S12. Patterns observed in the TF motifs in the HOCOMOCO database are replicated in the JASPAR database. (A) All promoter and enhancer sets are enriched for motifs for broadly-active TF motifs (P < 2.2E-16 for all, binomial test), while broadly-active promoters and enhancers are further enriched compared to their narrowly-active counterparts (P = 3.5E-41 and P = 2.7E-18, respectively, Wilcoxon Rank Sum test).

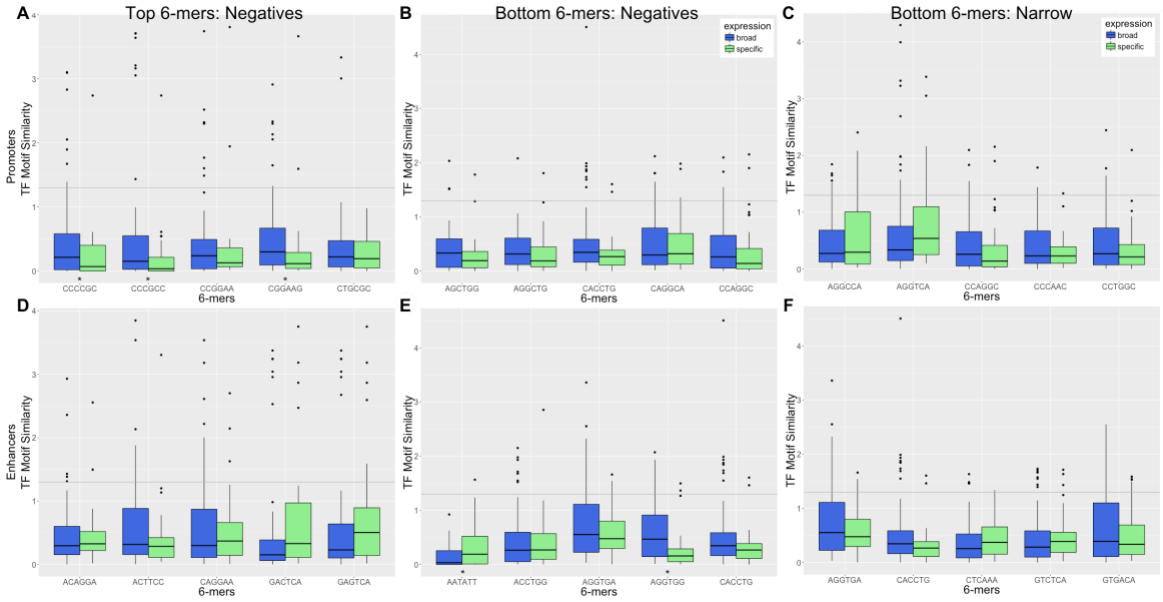

Figure S13. DNA patterns most predictive of (A) broadly active promoters (compared to genomic background regions) match the binding motifs of broadly expressed transcription factors (blue) more strongly than context-specific TFs (green). The most negatively weighted 6-mers from (B) the genomic background classifier and (C) the narrow promoter classifier. (D-F) same but for broadly-active enhancers. Motif similarity to a 6-mer is quantified as the $-\log_{10}(P\text{-value})$ of the match, with values >1.3 significant. The box plots show the median and 1st/3rd quartiles. $*P < 0.05$, Wilcoxon rank-sum test. Motifs were from the JASPAR 2016 database.

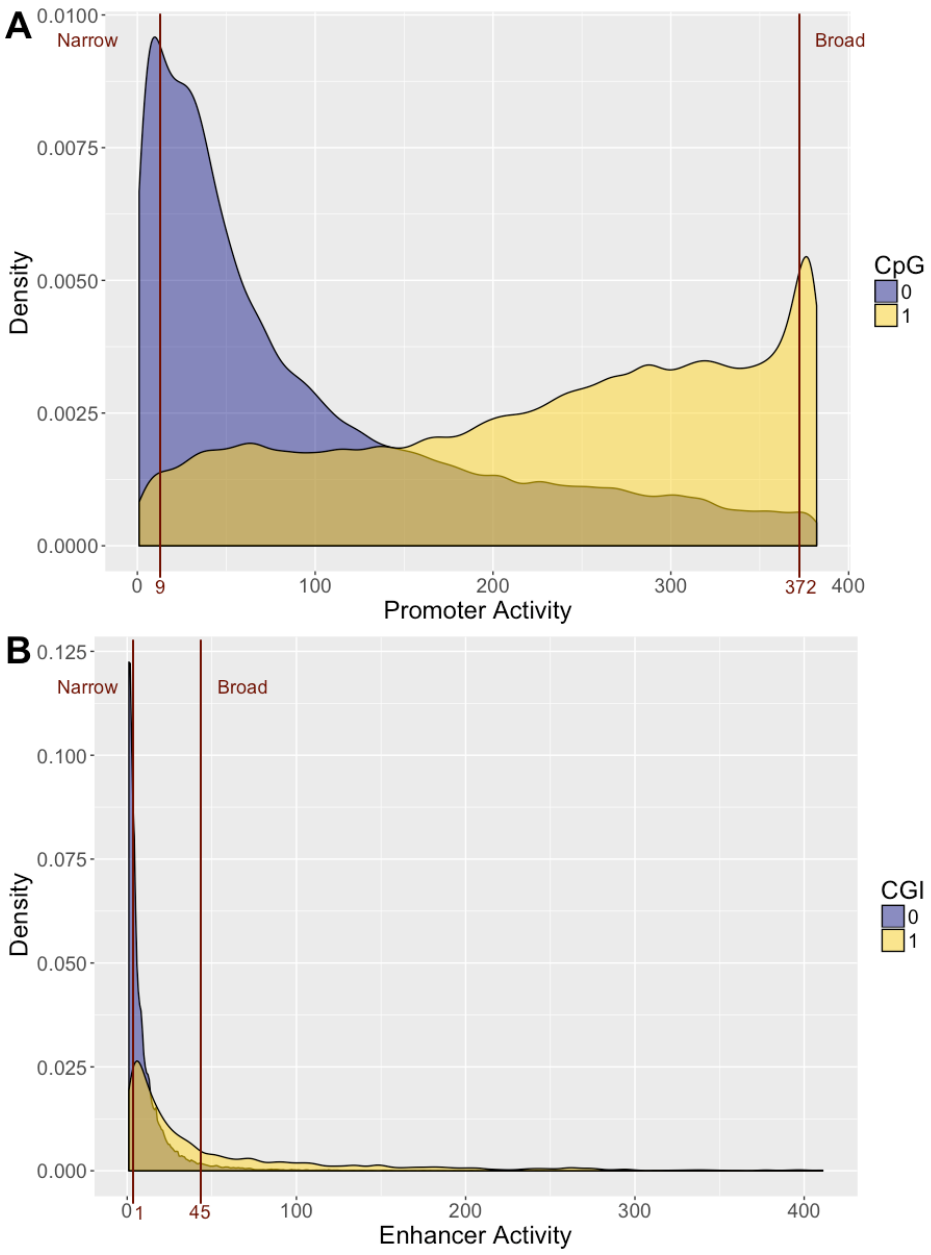

Figure S14. Activity density distributions of CAGE (A) promoters and (B) enhancers divided by CGI status (Yellow: CGI, Blue: Non-CGI)

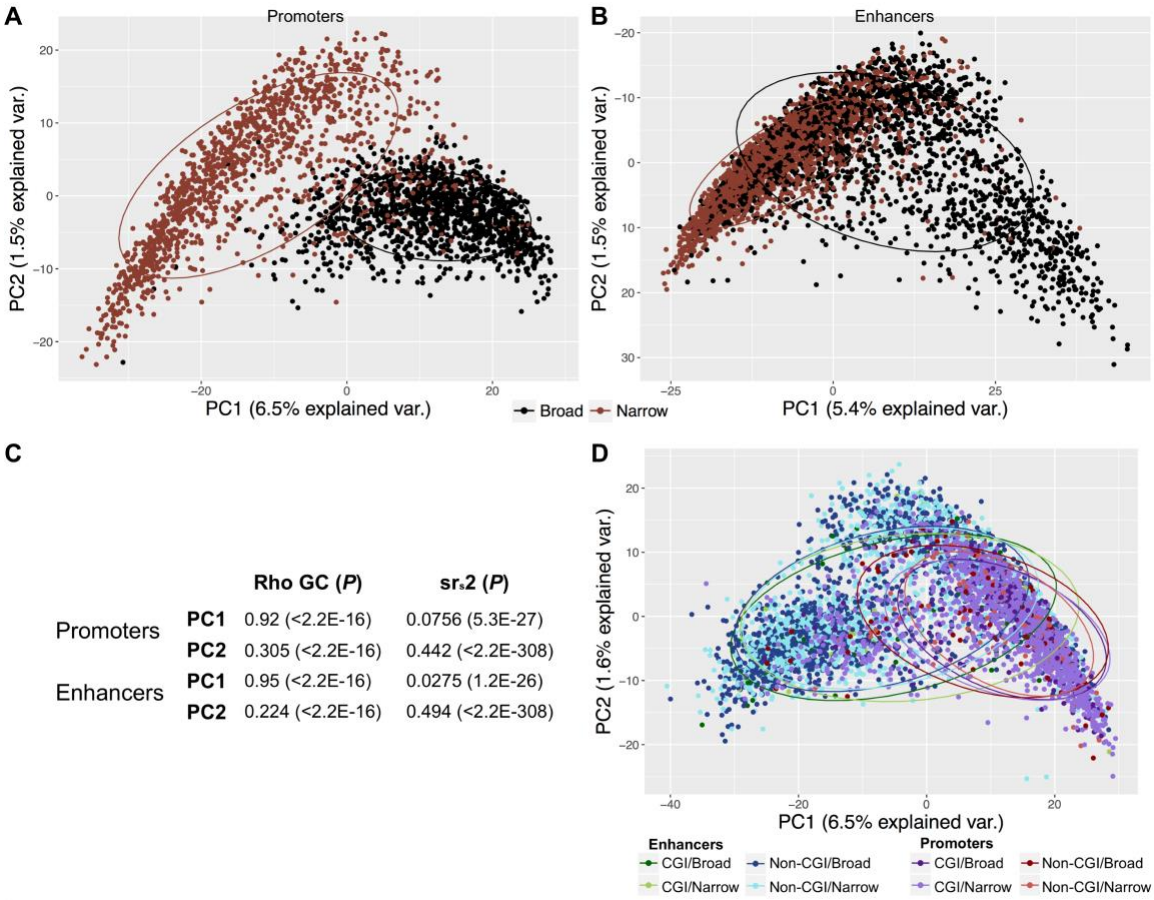

Figure S15. Broadly active enhancers and promoters can be separated from narrowly active regions by simple combinations of sequence characteristics. PCA plot based on of counts of all 6-mers in (A) promoters and (B) enhancers split by activity (black: broad, brown: narrow). The ellipses give 95% confidence regions based on the multivariate normal distribution. (C) Proportions of variance in PC1 and PC2 explained by GC content (Spearman's ρ) and of CpG content beyond GC content (semipartial Spearman correlation, sr_s^2). GC content strongly correlates with the first PCs; CpG content is strongly correlated with the second PCs beyond the contribution of GC content. (D) PCA of all promoters and enhancers together colored by activity (dark: broad; light: narrow) and CGI status (purple/green: CGI; red/blue: Non-CGI).

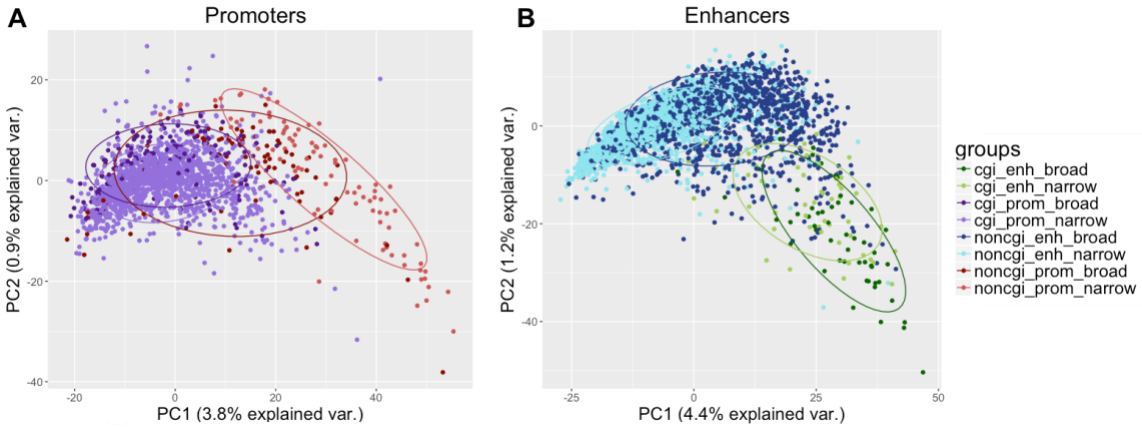

Figure S16. CGI and Non-CGI enhancers separate more cleanly than the corresponding promoters. PCA plot of counts of all 6-mers in (A) promoters and (B) enhancers colored by activity and CGI status (Purple/Green- CGI; Red/Blue- Non-CGI). Broad/Narrow regions are dark/light colors, respectively.

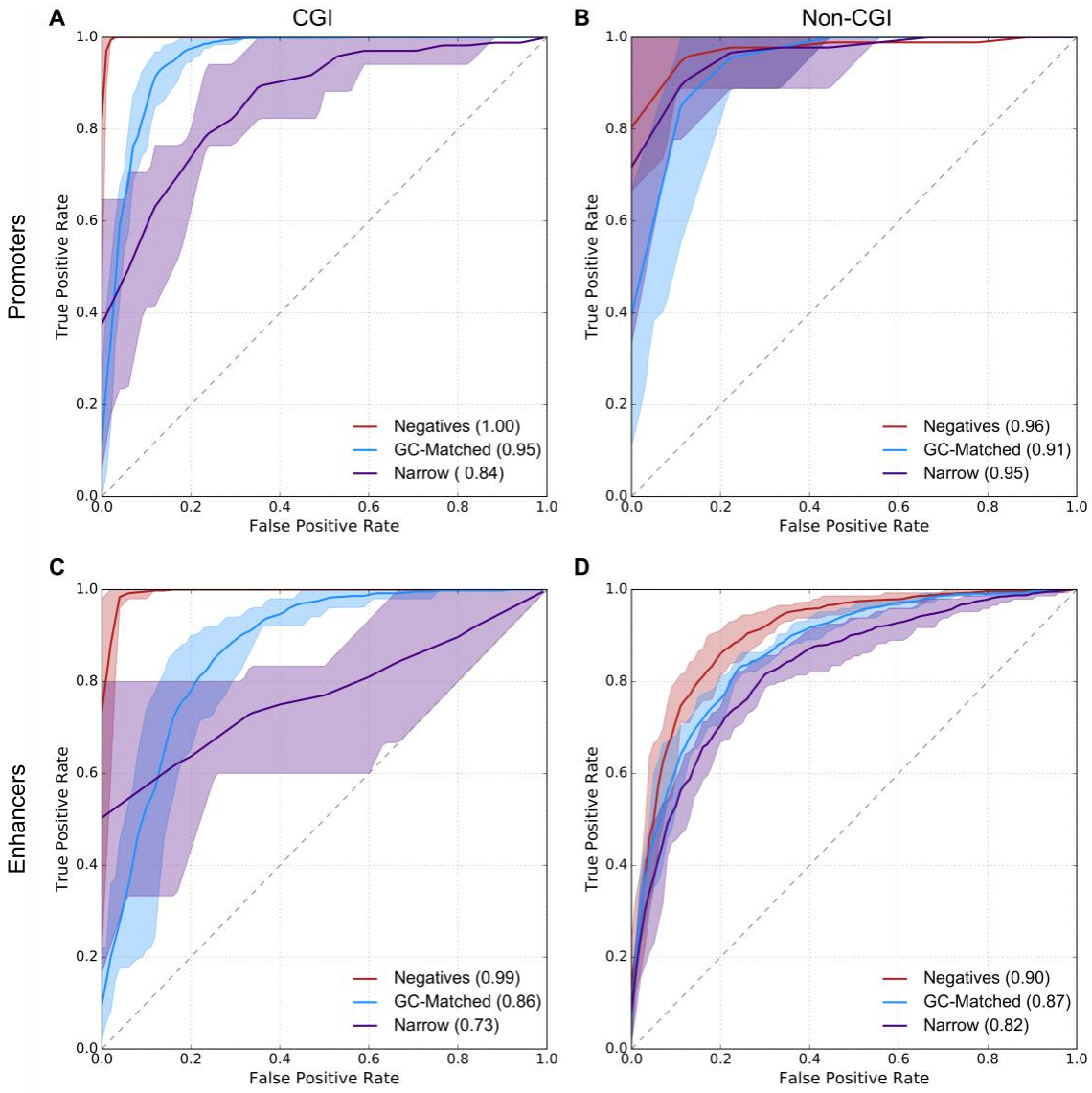

Figure S17. SVM classifiers can distinguish broadly active promoters and enhancers regardless of CGI status. ROC curves for (A) CGI-containing promoters, (B) Non-CGI promoters, (C) CGI enhancers, and (D) Non-CGI enhancers trained against length-matched genomic background regions (red), GC-matched genomic background regions (blue), or narrowly-active CGI/Non-CGI regions, respectively (purple).

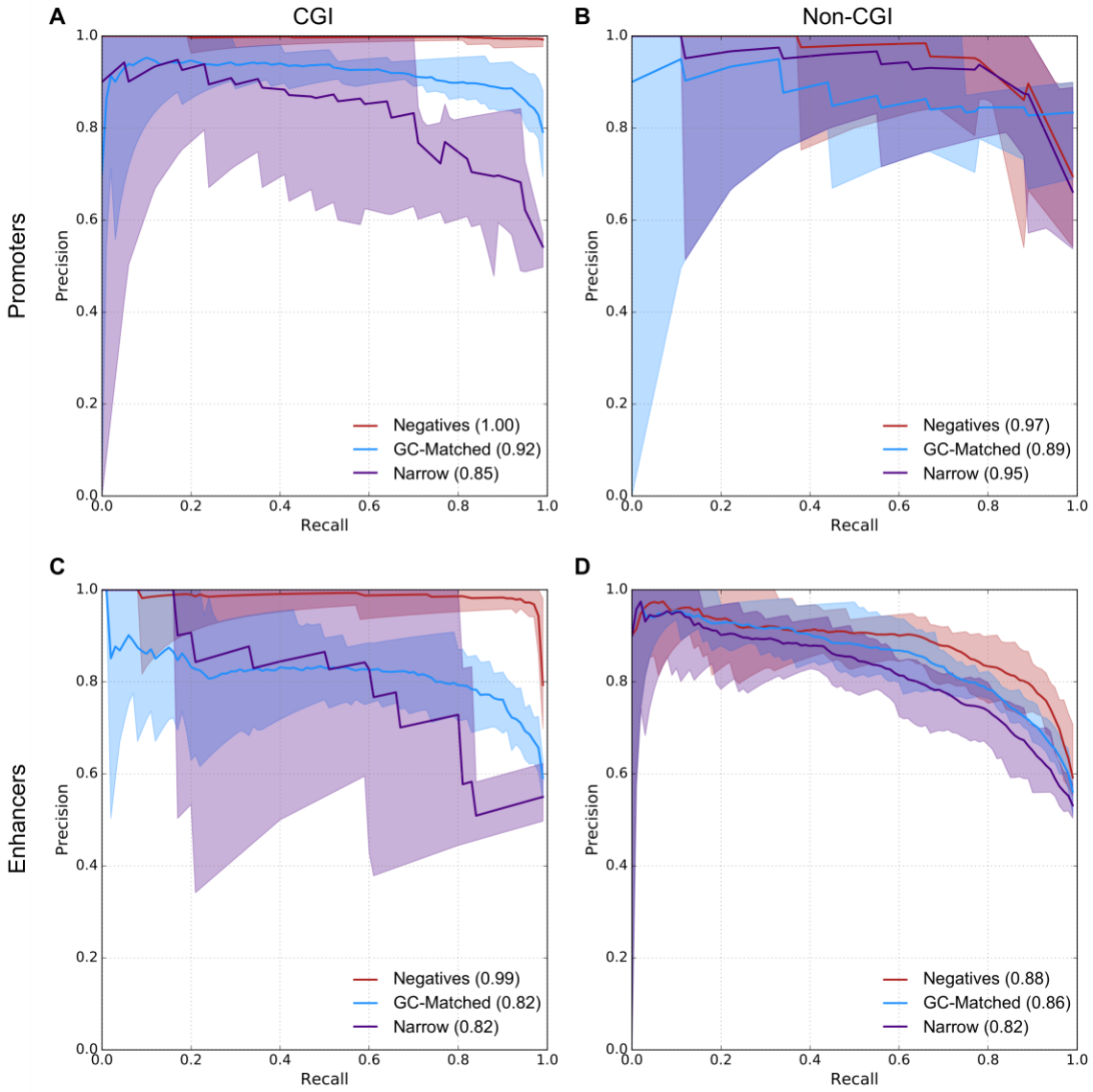

Figure S18. PR curves for classifiers in Figure S6. (A) CGI-containing promoters, (B) Non-CGI promoters, (C) CGI enhancers, and (D) Non-CGI enhancers trained against length-matched genomic background regions (red), GC-matched genomic background regions (blue), or narrowly-active CGI/Non-CGI regions, respectively (purple).

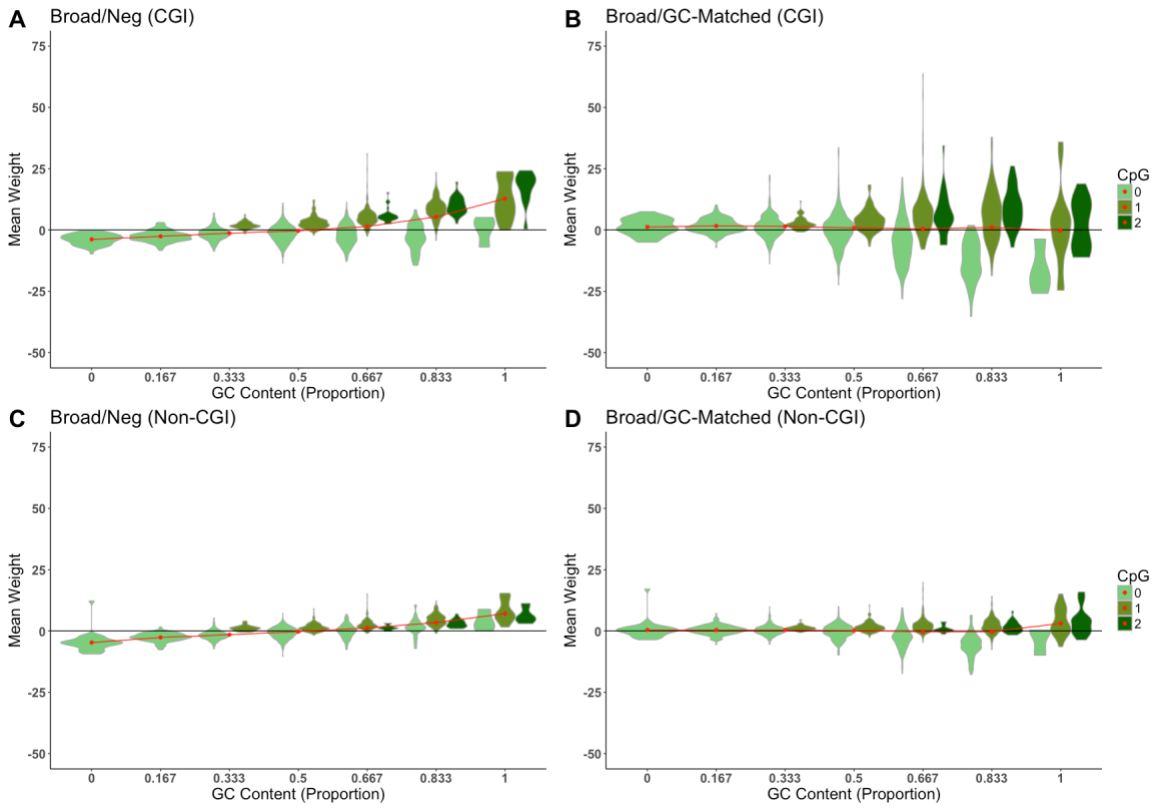

Figure S19. Weight assigned to each 6-mer in the models trained using (A) broad CGI promoters vs. genomic background, (B) broad CGI promoters vs. GC-matched background, and (C) broad Non-CGI promoters vs. genomic background and (D) broad Non-CGI promoters vs. GC-matched background plotted against GC content and stratified by the number of CpG dinucleotides in each 6-mer. 0, 1 or 2-3 CpGs per 6-mer are possible (light to dark green). The red points mark the mean weight for each GC content bin. Spearman rho for GC content vs. weight: 0.42 CGI Broad/Neg (P < 2.2E-16); 0.60 Non-CGI Broad/Neg (P < 2.2E-16); -0.036 CGI Broad/GC (P = 0.020); -0.00788 (P = 0.61).

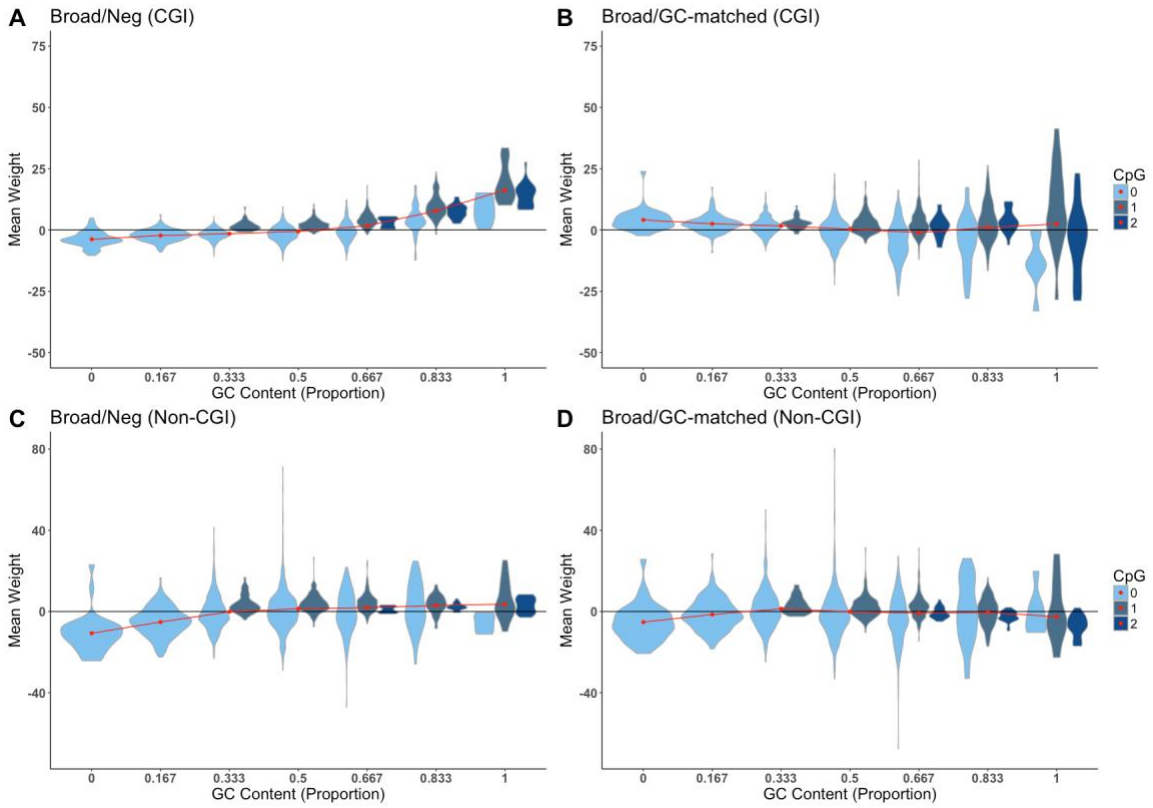

Figure S20. Weight assigned each 6-mer in the models trained against (A) broad CGI enhancers vs. genomic background, (B) GC-matched background, and (C) broad Non-CGI enhancers vs. genomic background and (D) GC-matched background plotted against GC content and stratified by the number of CpG dinucleotides in each 6-mer. 0, 1 and 2-3 CpGs per 6-mer are possible (light to dark blue). The red points mark the mean weight for each GC content bin. Spearman rho for GC content vs. weight: 0.54 CGI Broad/Neg (P < 2.2E-16); 0.28 Non-CGI Broad/Neg (P < 2.2E-16); -0.147 CGI Broad/GC (P < 2.2E-16); -0.0073 (P = 0.64).

SUPPLEMENTARY TABLES

Table S1. Summary of positive and negative training data used in each analysis.
* Results averaged over four classifiers trained with different random negative sets

Figures	Classifier	Positive Set	Negative	#Pos/Neg	Subsampling
1	Promoter/Enhancer	All Promoters	All Enhancers	3000	yes
2,S4	*Enh. Broad/Neg	Broad Enhancers	Random Background Regions	1961	no
2,S4	*Enh. Broad/GC	Broad Enhancers	GC-matched background regions	1961	no
2,S4	Enh. Broad/Narrow	Broad Enhancers	Narrow Enhancers	1961	yes (narrow)
2,S4	*Prom. Broad/Neg	Broad Promoters	Random Background Regions	1362	no
2,S4	*Prom. Broad/GC	Broad Promoters	GC-matched background regions	1362	no
2,S4	Prom. Broad/Narrow	Broad Promoters	Narrow Promoters	1362	yes (broad)
S17,S18	*CGI Enh. Broad/Neg	Broad CGI Enhancers	Random Background Regions	507	no
S17,S18	*CGI Enh. Broad/GC	Broad CGI Enhancers	GC-matched background regions	507	no
S17,S18	CGI Enh. Broad/Narrow	Broad CGI Enhancers	Narrow CGI Enhancers	55	yes (broad)
S17,S18	*Non-CGI Enh. Broad/Neg	Broad Non-CGI Enhancers	Random Background Regions	1454	no
S17,S18	*Non-CGI Enh. Broad/GC	Broad Non-CGI Enhancers	GC-matched background regions	1454	no
S17,S18	Broad Non-CGI Enhancers	Broad Non-CGI Enhancers	Narrow Non-CGI Enhancers	1454	yes (narrow)
S17,S18	*CGI Prom. Broad/Neg	Broad CGI Promoters	Random Background Regions	1273	no
S17,S18	*CGI Prom. Broad/GC	Broad CGI Promoters	GC-matched background regions	1273	no
S17,S18	CGI Prom. Broad/Narrow	Broad CGI Promoters	Narrow CGI Promoters	170	yes (broad)
S17,S18	*Non-CGI Prom. Broad/Neg	Broad Non-CGI Promoters	Random Background Regions	89	no
S17,S18	*Non-CGI Prom. Broad/GC	Broad Non-CGI Promoters	GC-matched background regions	89	no
S17,S18	Non-CGI Prom. Broad/Narrow	Broad Non-CGI Promoters	Narrow Non-CGI Promoters	89	yes (narrow)
S8	CGI Enhancer/Promoter	CGI Enhancers	CGI Promoters	1590	yes (prom.)
S8	Non-CGI Enhancer/Promoter	Non-CGI Enhancers	Non-CGI Promoters	6000	yes (enh.)

Table S2. List of significant matches between TF motifs and 6-mers for Figures 5 and S11-13. Separate File.