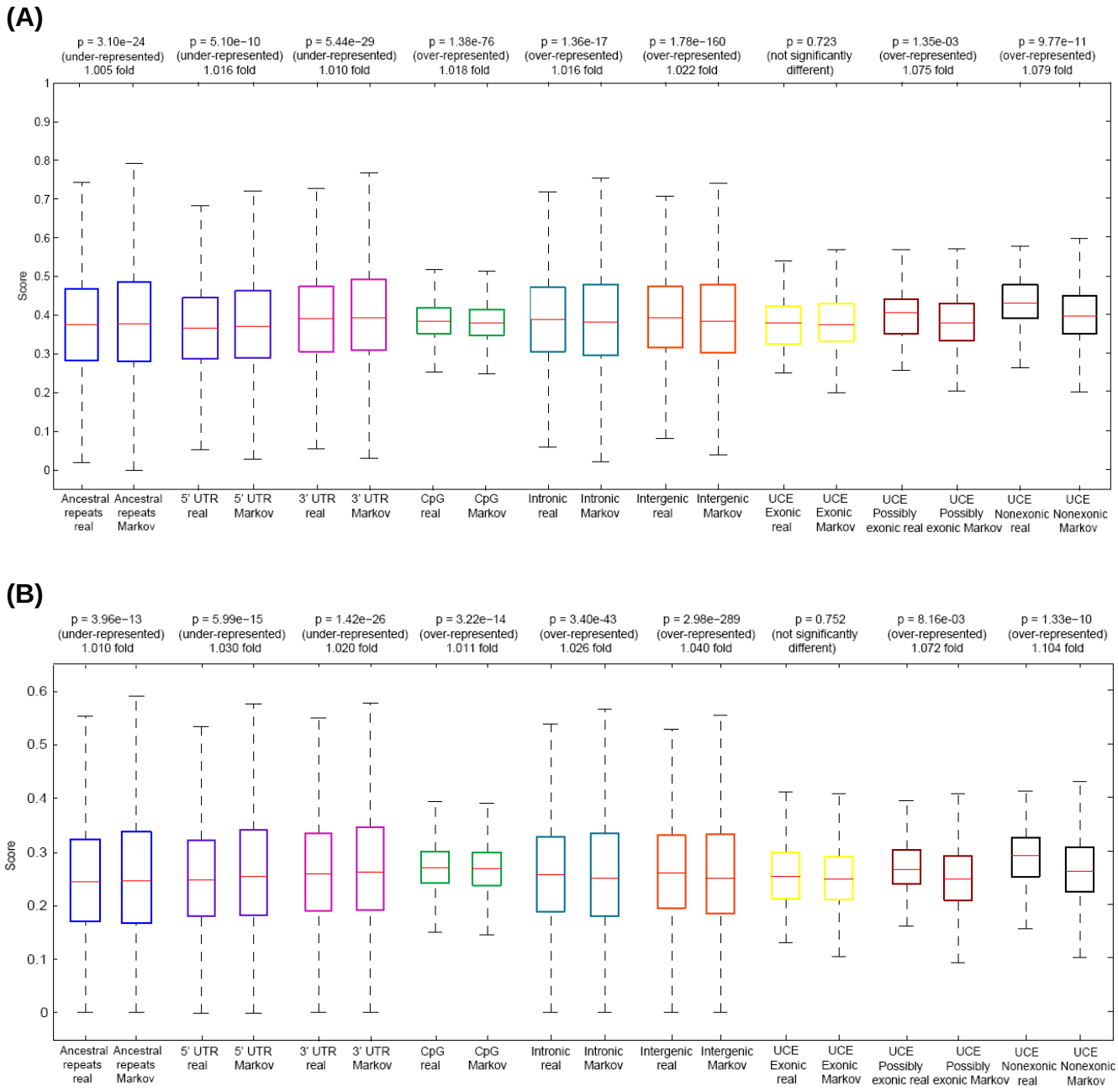
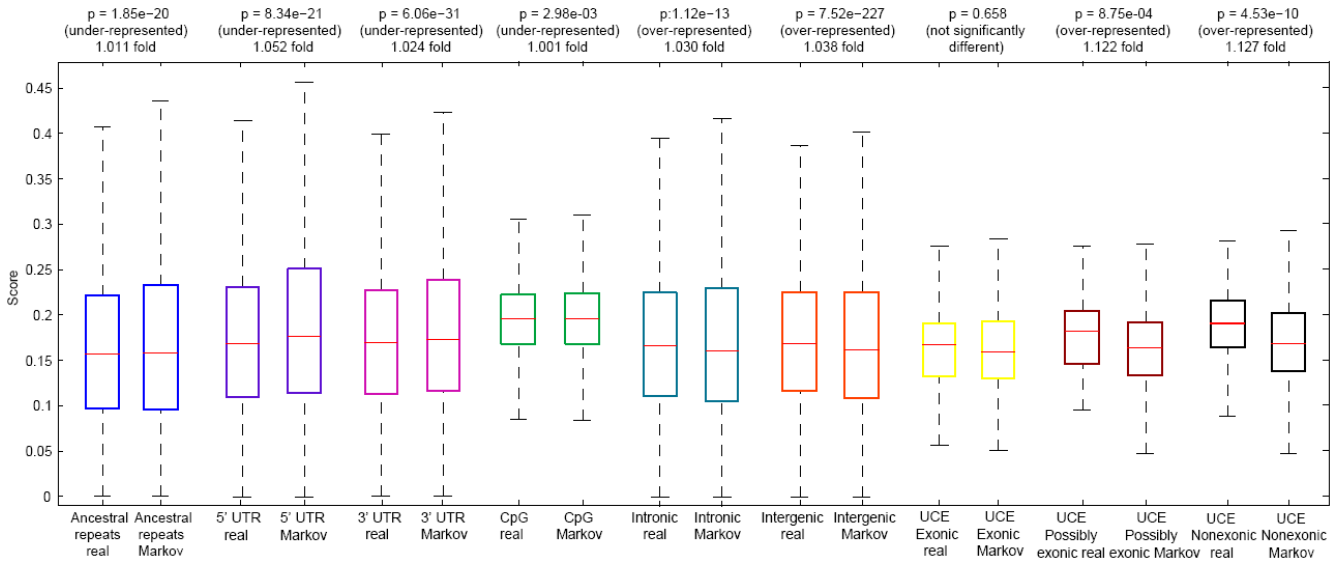
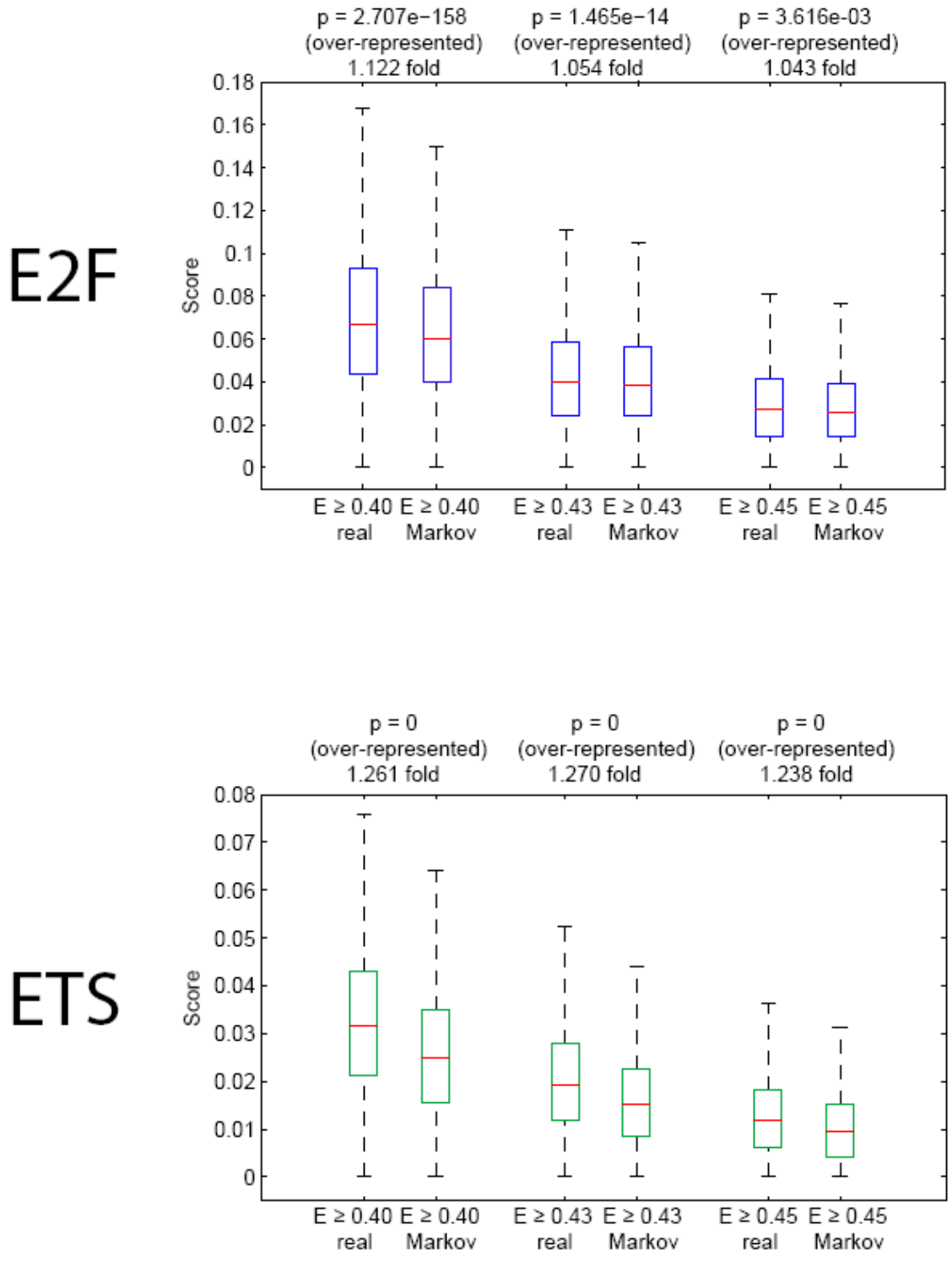


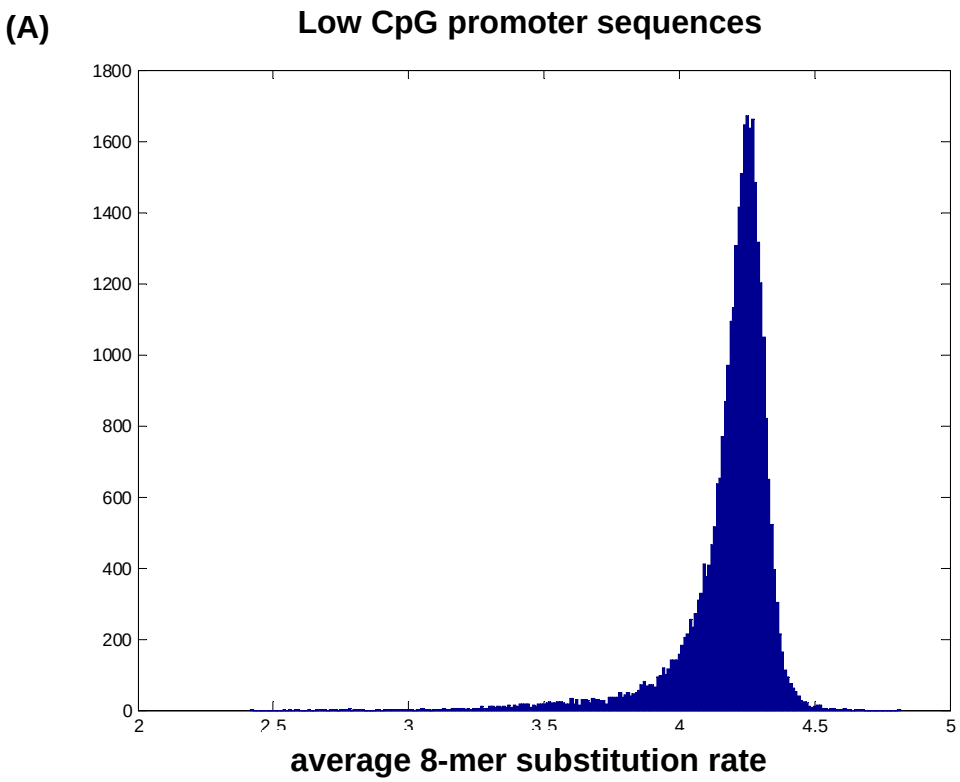
Supplementary Figure S1: Box plots for enrichment of TF binding site *k*-mers within various genomic regions. All *k*-mers bound at **(A)** E ≥ 0.40, (B) E ≥ 0.43, or **(C)** E ≥ 0.45 were considered for each TF. A 1st order Markov model was used to create shuffled control sequences that control for dinucleotide bias in each given class of genomic regions. Each sequence in each set of genomic regions and control sequences was scanned in overlapping 8-bp windows with a step size of 1 bp using custom Perl scripts. "Score" indicates the fraction of windows with matches to 8-mers with E ≥ 0.40, 0.43, or 0.45 for any TFs of a particular structural class for each sequence. Under/over-representation *P*-values (Wilcoxon-Mann-Whitney test) were calculated by comparing 8-mers within each set of genomic regions ("real") to their corresponding shuffled control sequences ("Markov"). In each box plot, the central bar indicates the median, the edges of the box indicate the 25th and 75th percentiles, and the whiskers extend to the most extreme data points not considered outliers.

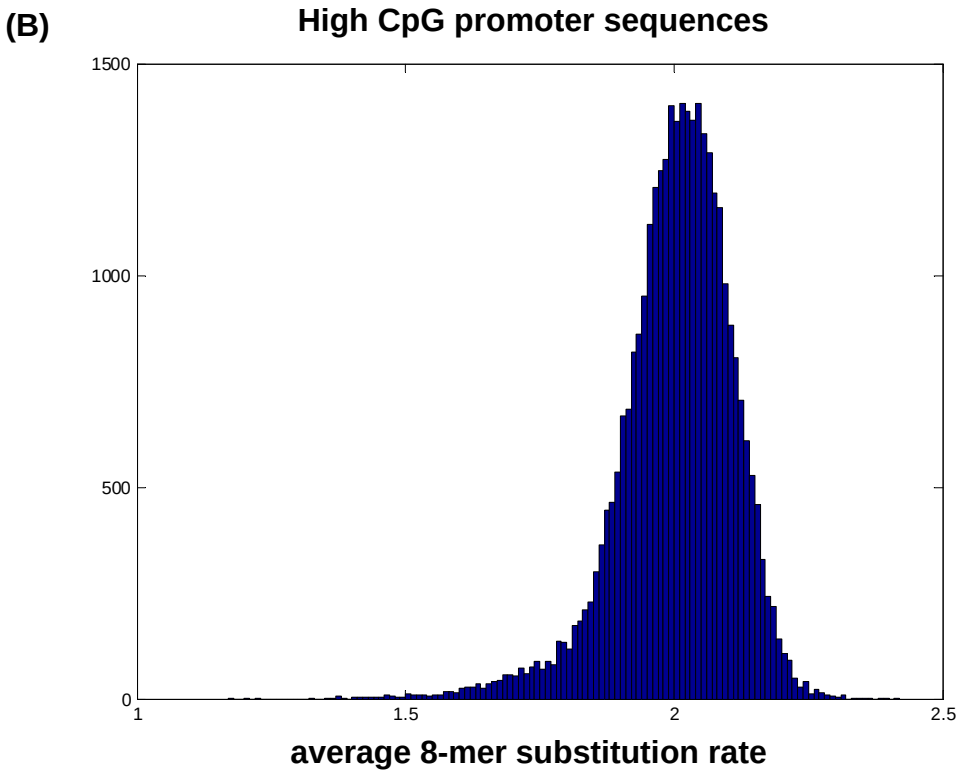
(C)


p = 1.85e−20 (under-represented) 1.011 fold
p = 8.34e−21 (under-represented) 1.052 fold
p = 6.06e−31 (under-represented) 1.024 fold
p = 2.98e-03 (under-represented) 1.001 fold
p:1.12e−13 (over-represented) 1.030 fold
p = 7.52e−227 (over-represented) 1.038 fold
p = 0.658 (not significantly different)
p = 8.75e-04 (over-represented) 1.122 fold
p = 4.53e−10 (over-represented) 1.127 fold
Score
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Ancestral repeats real
Ancestral repeats Markov
5' UTR real
5' UTR Markov
3' UTR real
3' UTR Markov
CpG real
CpG Markov
Intronic real
Intronic Markov
Intergenic real
Intergenic Markov
UCE Exonic real
UCE Exonic Markov
UCE Possibly exonic real
UCE Possibly exonic Markov
UCE Nonexonic real
UCE Nonexonic Markov

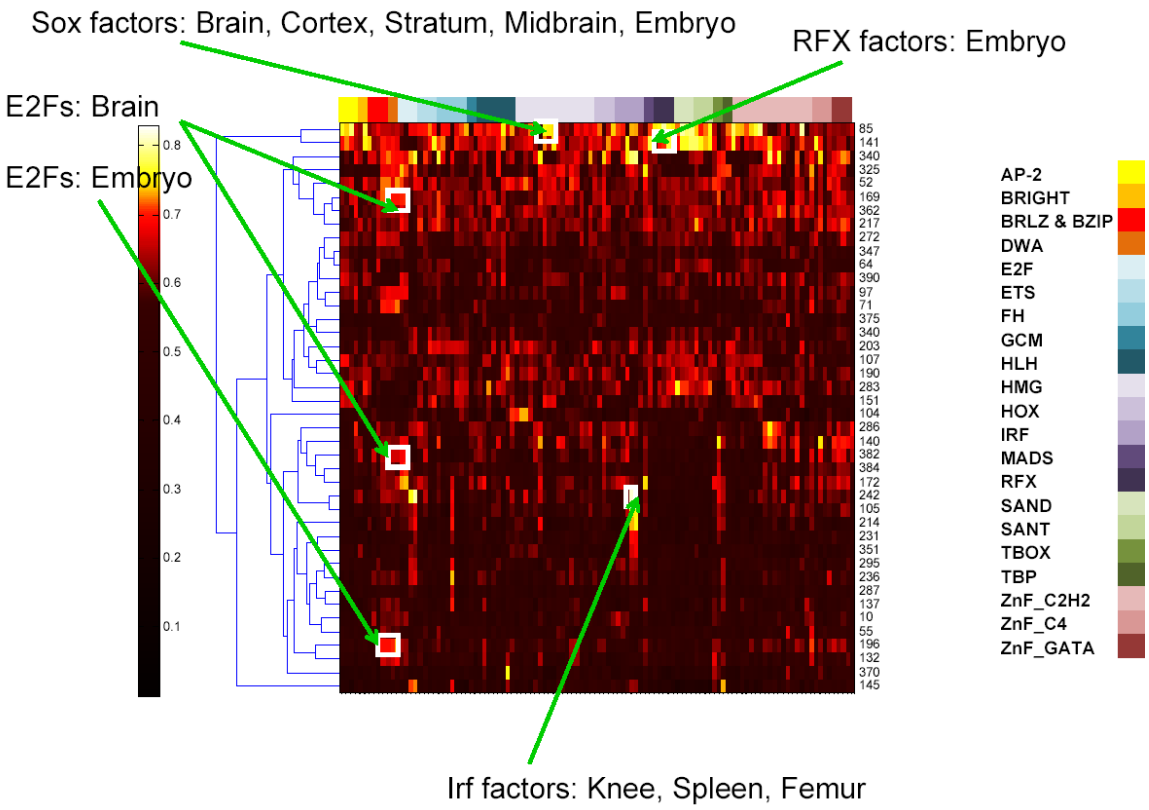


Supplementary Figure S2: CpG islands are enriched for PBM 'bound' *k*-mers for E2F and ETS proteins.

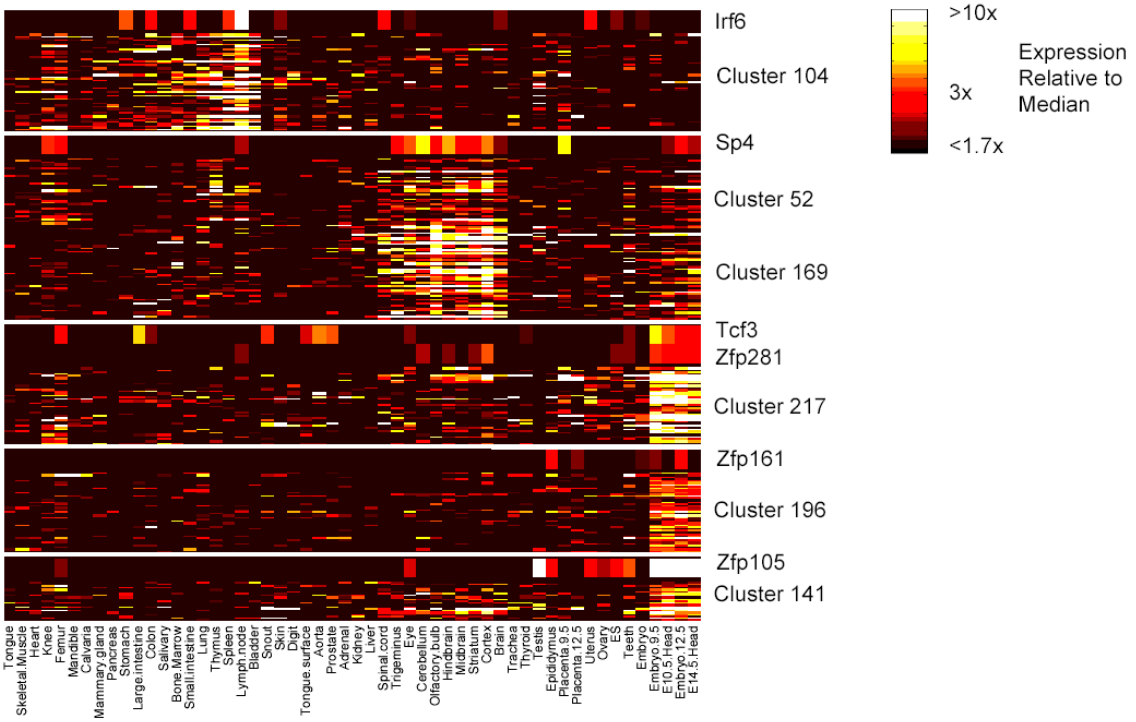




Supplementary Figure S3. Distribution of the mean substitution rates of all ungapped 8-mers over (A) low CpG promoter sequences and (B) high CpG promoter sequences.



Supplementary Figure S4: Results from Lever screen (AUC ≥ 0.6, $Q \leq 0.05$) of tissue gene expression clusters. Gene expression is from Zhang *et al.*, *J. Biol.*, 2004.



Supplementary Figure S5: Relative gene expression levels of TFs and clusters that they are associated with in the results from Lever analysis. Expression data is from Zhang *et al.*, *J. Biol.*, 2004. Irf6 is associated with a gene expression cluster appearing in tissues that contain immune cells (e.g. spleen, lymph node, bone marrow, lung, etc.) (AUC = 0.720, $Q = 0.017$). Sfpi1, an ETS protein involved in myeloid cell differentiation (Scott *et al.*, *Science*, 1994), is specifically associated with this GO category in our analyses (AUC = 0.752; $Q \leq 0.01$ unless otherwise indicated). Other associations represent potential new findings: Sp4, a ZnF protein, is associated with a gene expression cluster expressed primarily in brain (AUC = 0.715), and while its GO annotations do not reflect this, reports in the literature (Zhou *et al.*, *Mol. Psychiatry*, 2005) indicate that mutants have defects in memory, and the gene itself is expressed primarily in brain; Tcf3, which is best known for its involvement in Wnt signaling (Korswagen *et al.*, *Cold Spring Harb Symp Quant Biol,* 1999), is associated in our analysis with an expression cluster that is highest embryos (AUC = 0.622, $Q = 0.32$), which is where it is expressed itself, and it is also associated with the GO category 'segmentation' (AUC = 0.718), consistent with a recently described role in restricting induction of the anterior-posterior axis (Merrill *et al.*, *Development*, 2004); Zfp161, Zfp281, and Zfp105, of which Zfp161 and Zfp281 were previously uncharacterized and of unknown function while Zfp105 (ZF5) is a repressor of the human fragile X-mental retardation 1 (FMR1) gene (Orlov *et al.*, *FEBS J*, 2007), are each associated with gene expression clusters primarily in the embryo (Zfp161, AUC = 0.711, $Q = 0.024$; Zfp281, AUC = 0.627, $Q = 0.27$; Zfp105, AUC = 0.615, $Q = 0.35$).