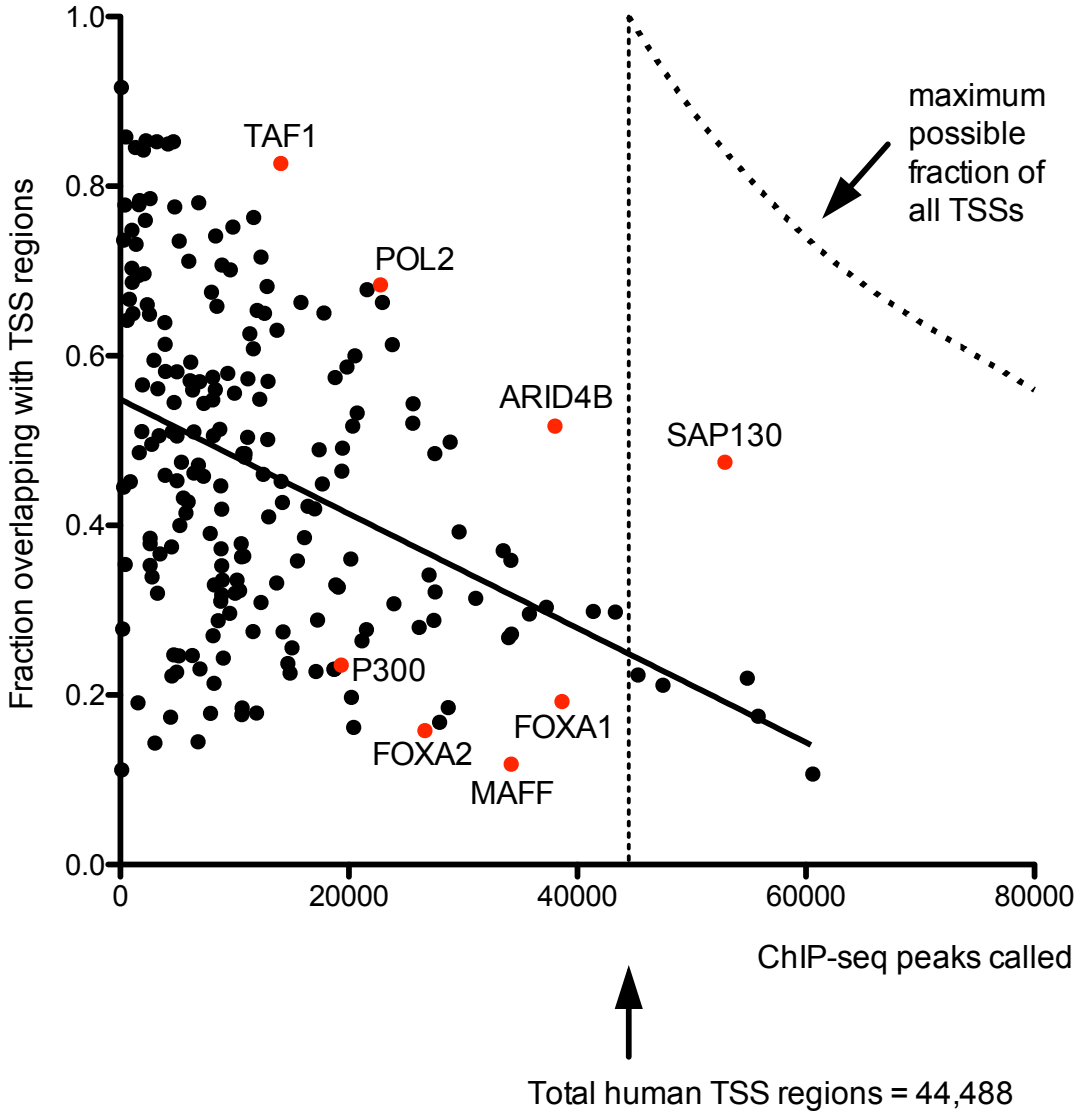
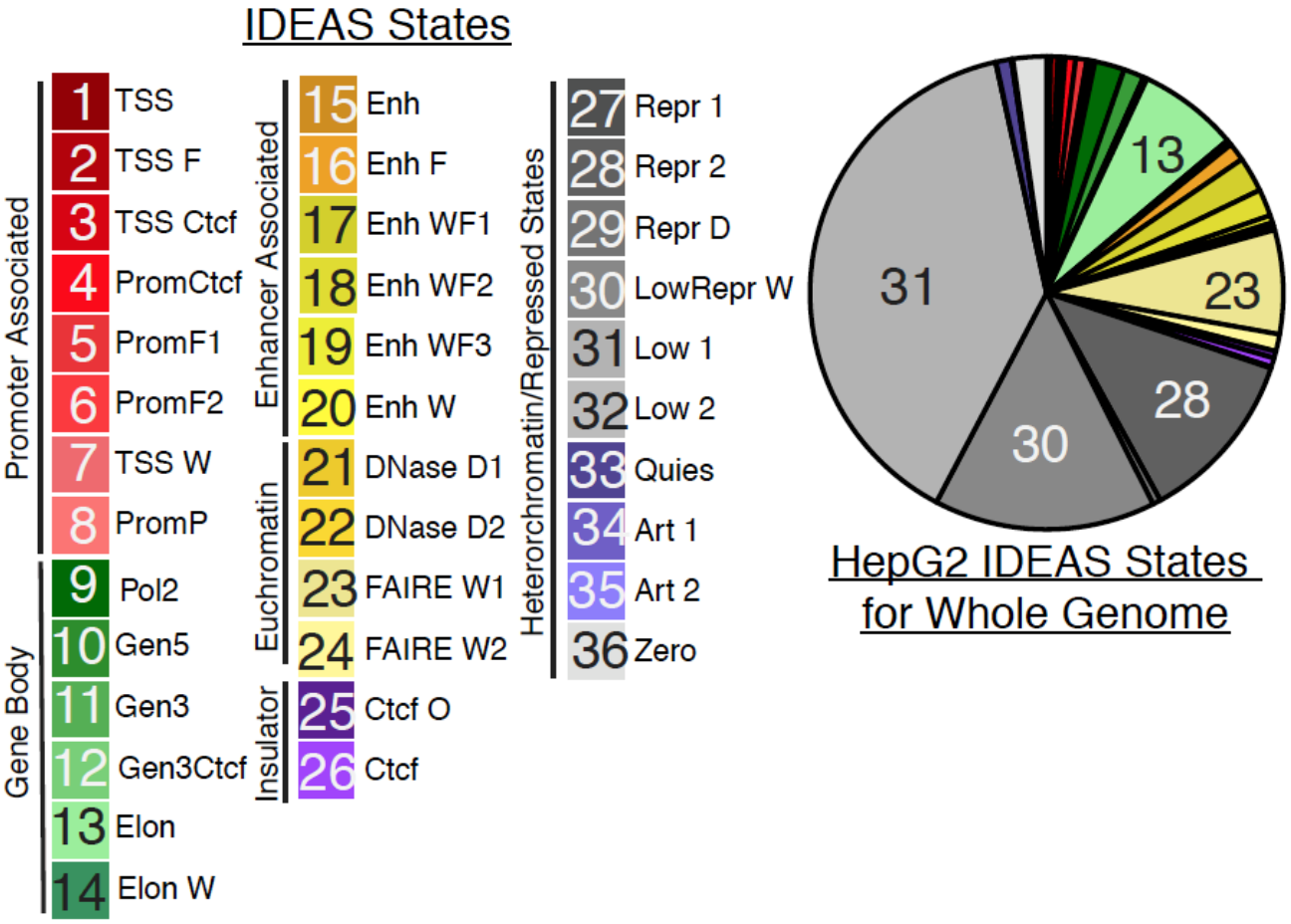
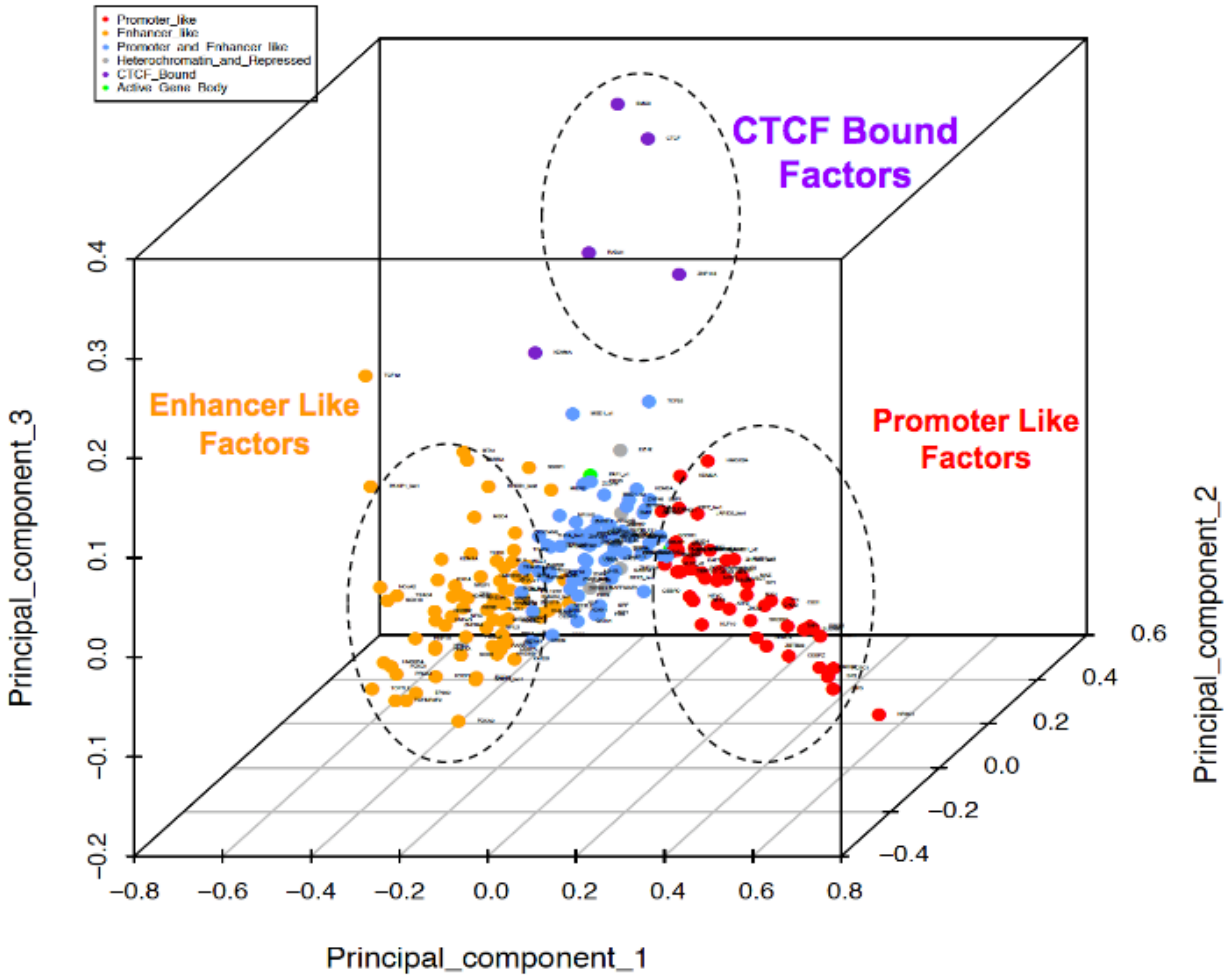


Supplementary Figure 1. 208 ChIP-seq/CETCh-seq experiments plotted by number of peaks called in each experiment (x-axis) vs fraction of peaks overlapping with any of 44,488 TSSs in the human genome (peaks +/- 3 kb of TSS). Selected individual factors are labelled. Solid line is linear regression through all points; dotted lines represent number of total TSS regions and maximum possible fraction of TSSs.

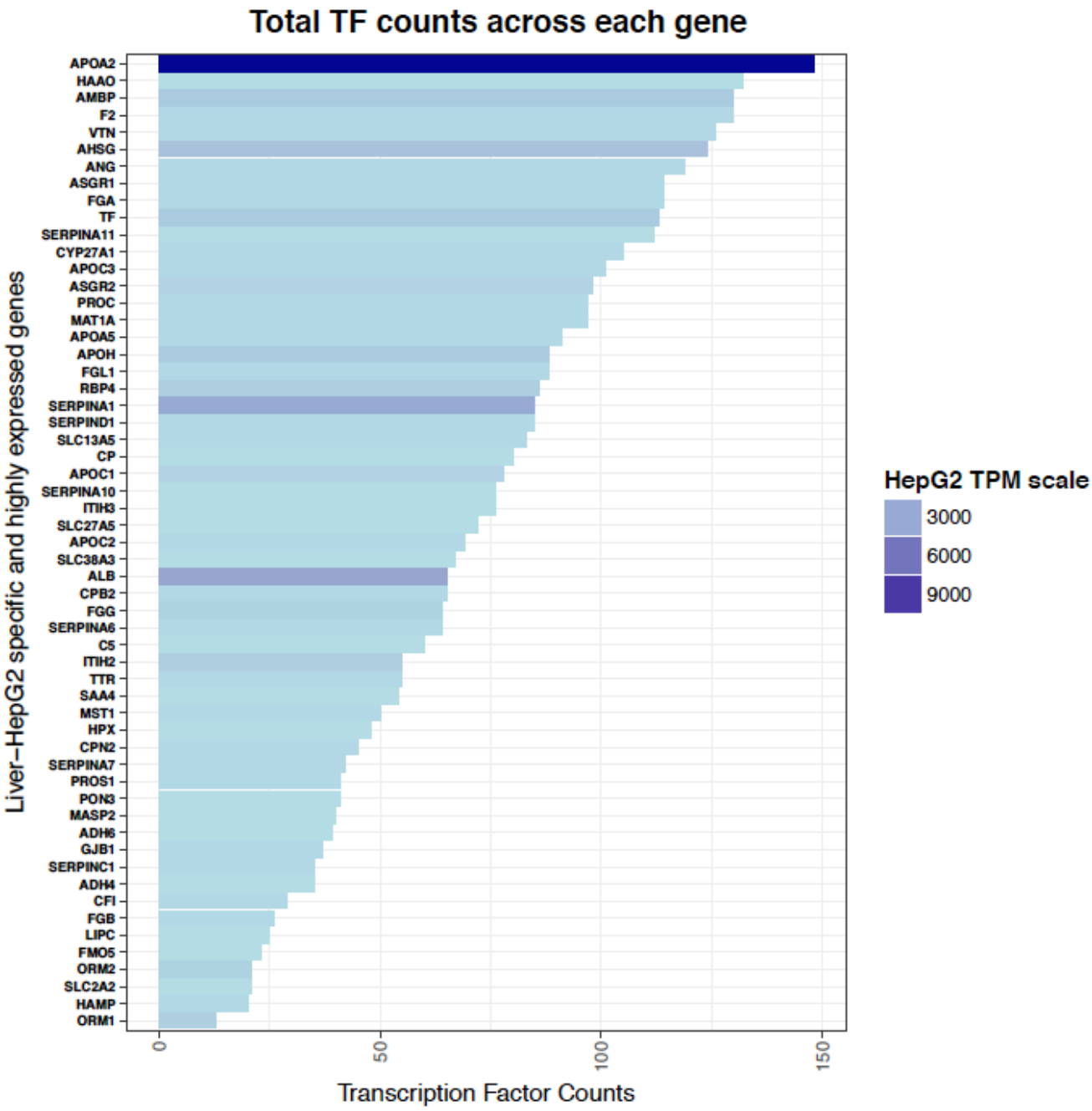


Supplementary Figure 2. IDEAS segmentation of HepG2 genome. Left: color key for all IDEAS states. Right: Pie chart indicating fraction of HepG2 genome associated with each state.

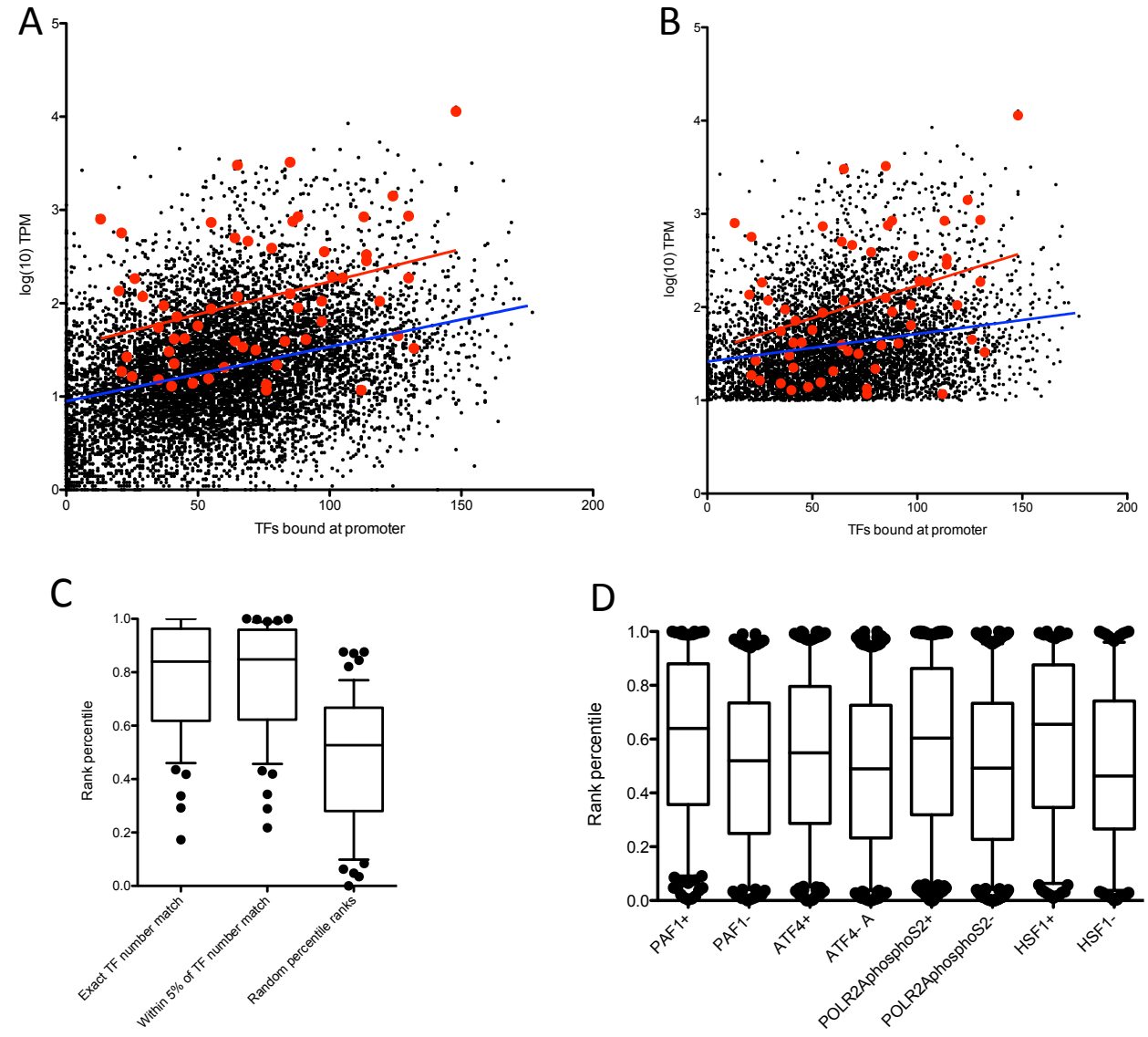
- Promoter_like
- Enhancer_like
- Promoter_and_Enhancer_like
- Heterochromatin_and_Repressed
- CTCF_Bound
- Active_Gene_Body



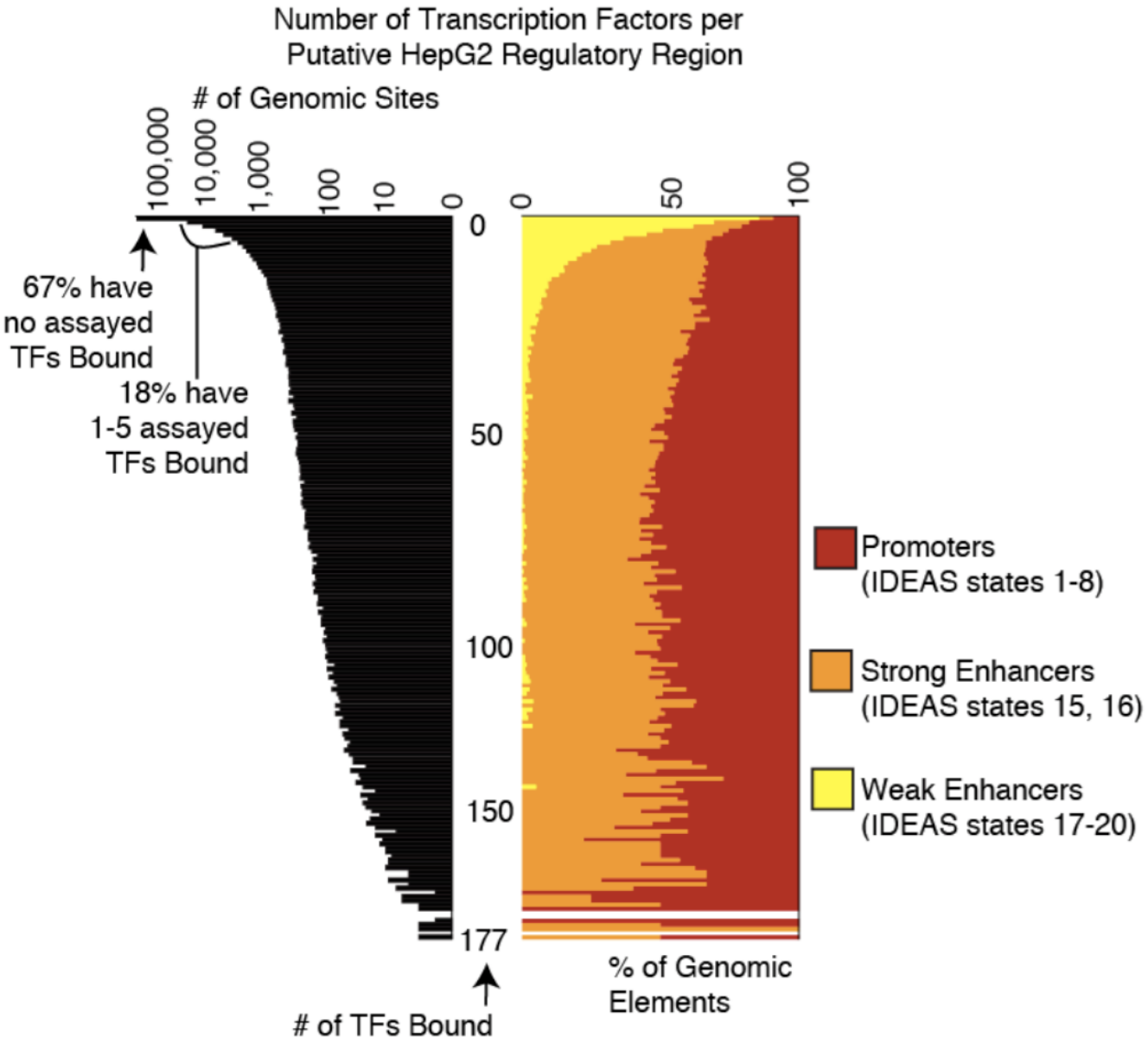
Supplementary Figure 3. Clustering of DNA associated factors based on chromatin states recapitulating the assigned cluster with PC1 (63.50%), PC2(16.51%) and PC3(6.48%) variances explained.



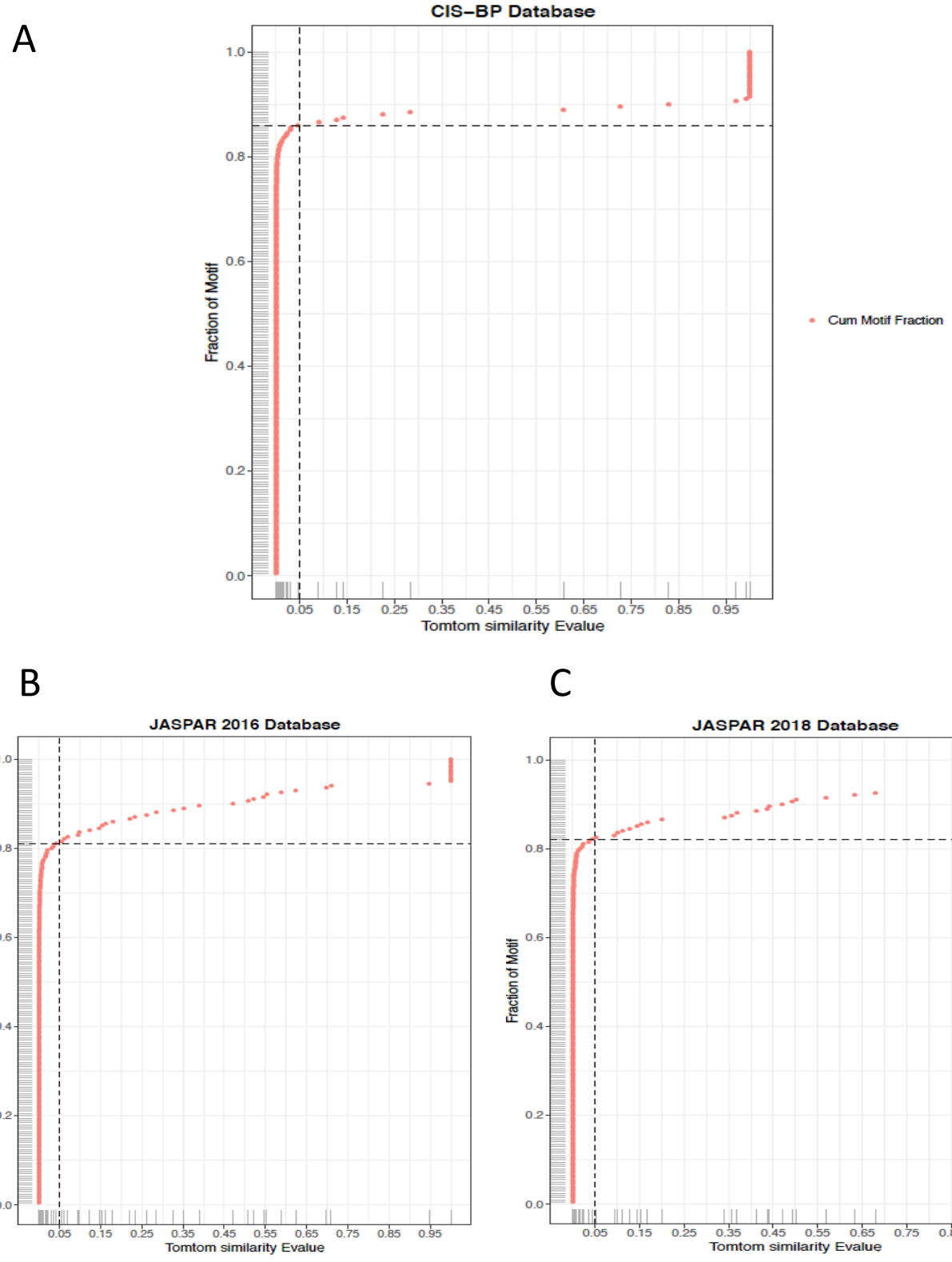
Supplementary Figure 4. Number of factors with called peaks within 2 kb of the TSSs of 57 liver specific genes with expression levels in HepG2 shown by bar color. TPM = Transcripts Per Million.



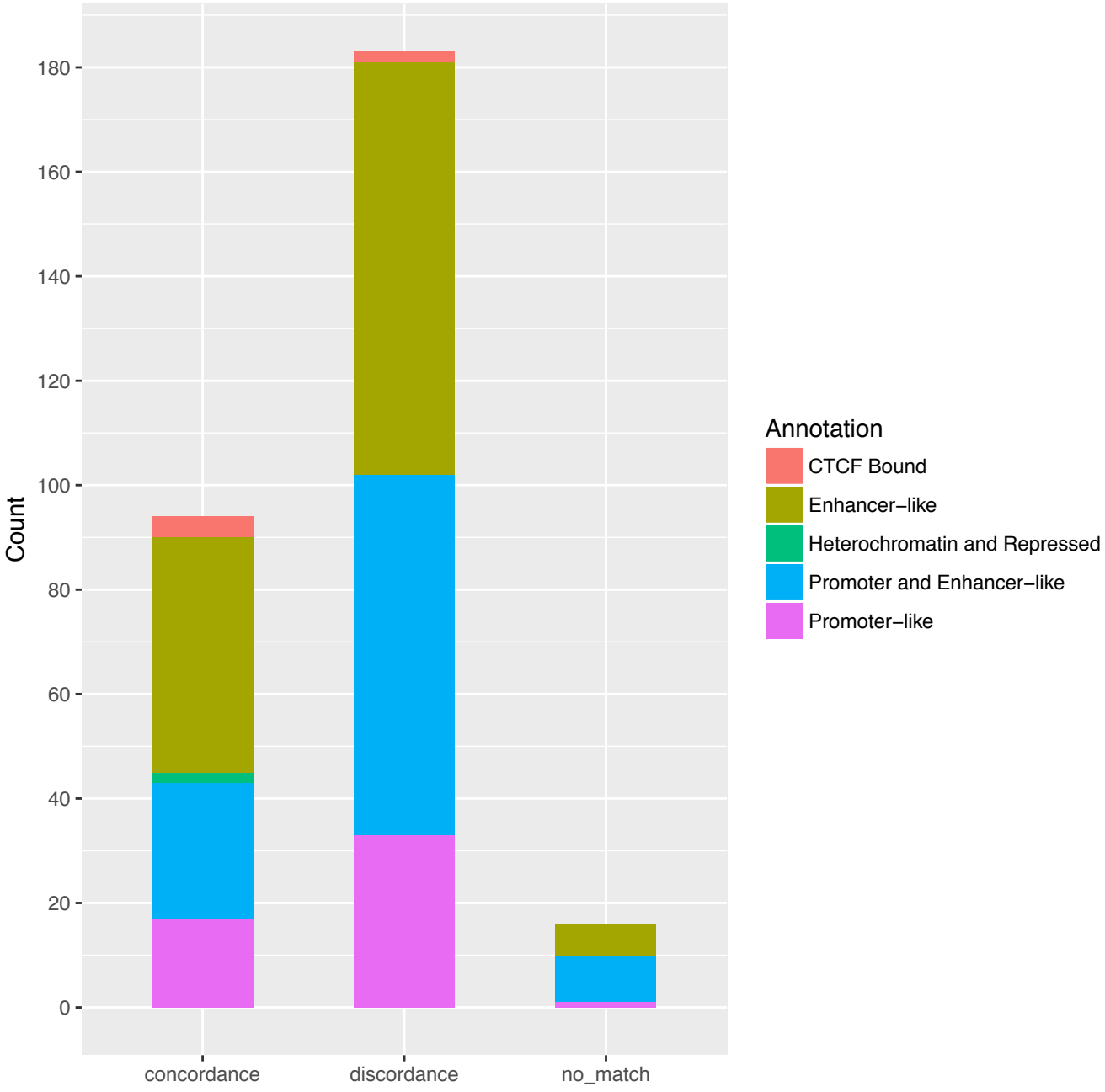
Supplementary Figure 5. A. Scatterplot of all genes (black points), showing log10 TPM in HepG2 (y-axis) vs number of unique TFs with called peak +/- 2 kb of gene TSS (x-axis). Blue line indicates linear regression through black points. Red points represent 57 liver-specific genes; red line indicates linear regression through red points. **B.** same as A, but all genes expressed below 10 TPM are removed; blue line indicates linear regression through only these >10 TPM genes. **C.** Distribution of rank percentiles for expression of 57 liver-specific genes, compared to exactly matching number of TFs (left box) and to within 5% of number of TFs (center box); random rank percentile for comparison is shown in right box (Mann-Whitney p-value < 0.0001 for both exact and 5% match when tested against random). **D.** Rank percentile of expression of all genes with specific TF's presence compared to rank percentile of equal number of random matched genes with within 5% of same number of TFs but without specific TF. TFs analyzed are PAF1 (M-W p<0.0001) , ATF4 (M-W p=0.0093), POLR2AphosphoS2 (M-W p<0.0001), and HSF1 (M-W p=0.0002). TPM = Transcripts Per Million. Boxes indicate quartiles, whiskers are drawn to 5-95% quartiles.



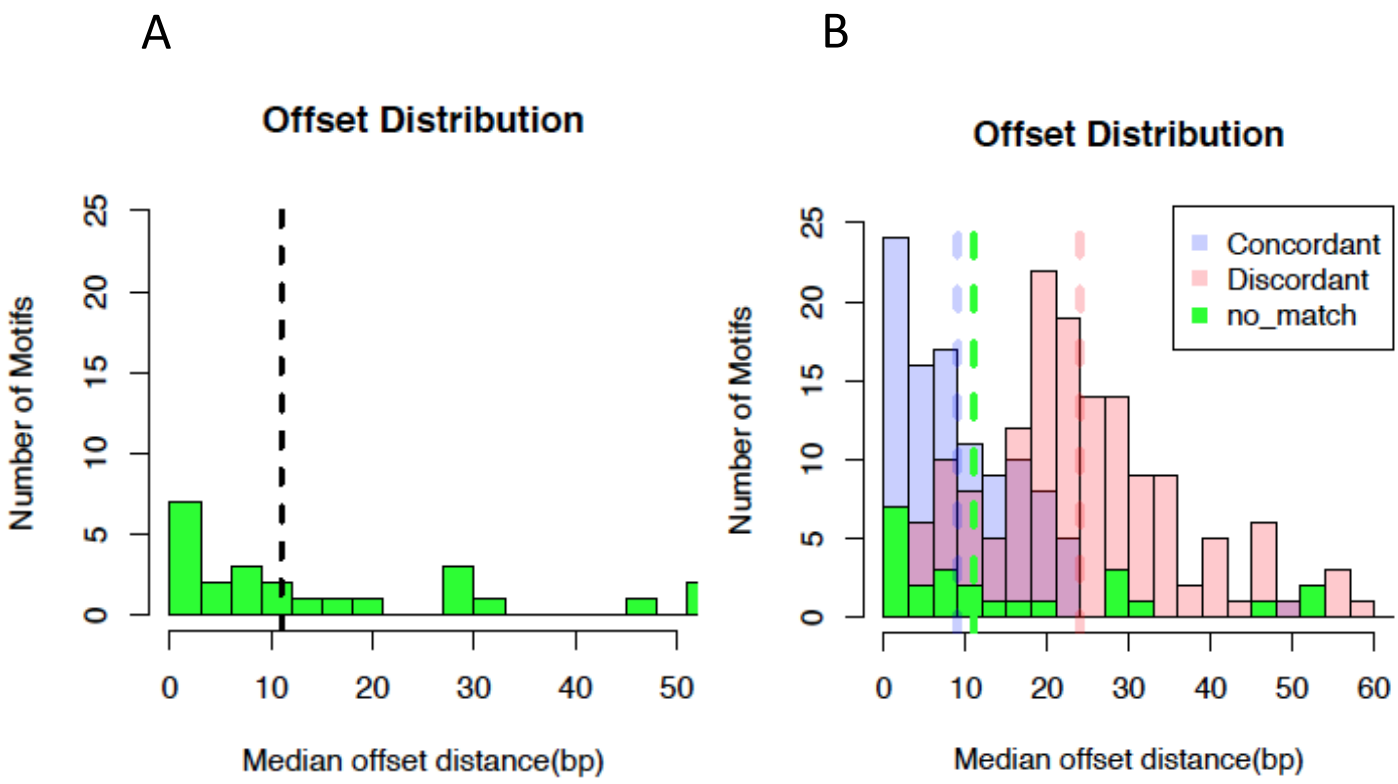
Supplementary Figure 6. Distribution of regulatory regions by number of associated TFs (left). Distribution of horizontally matched sites by IDEAS states (right).



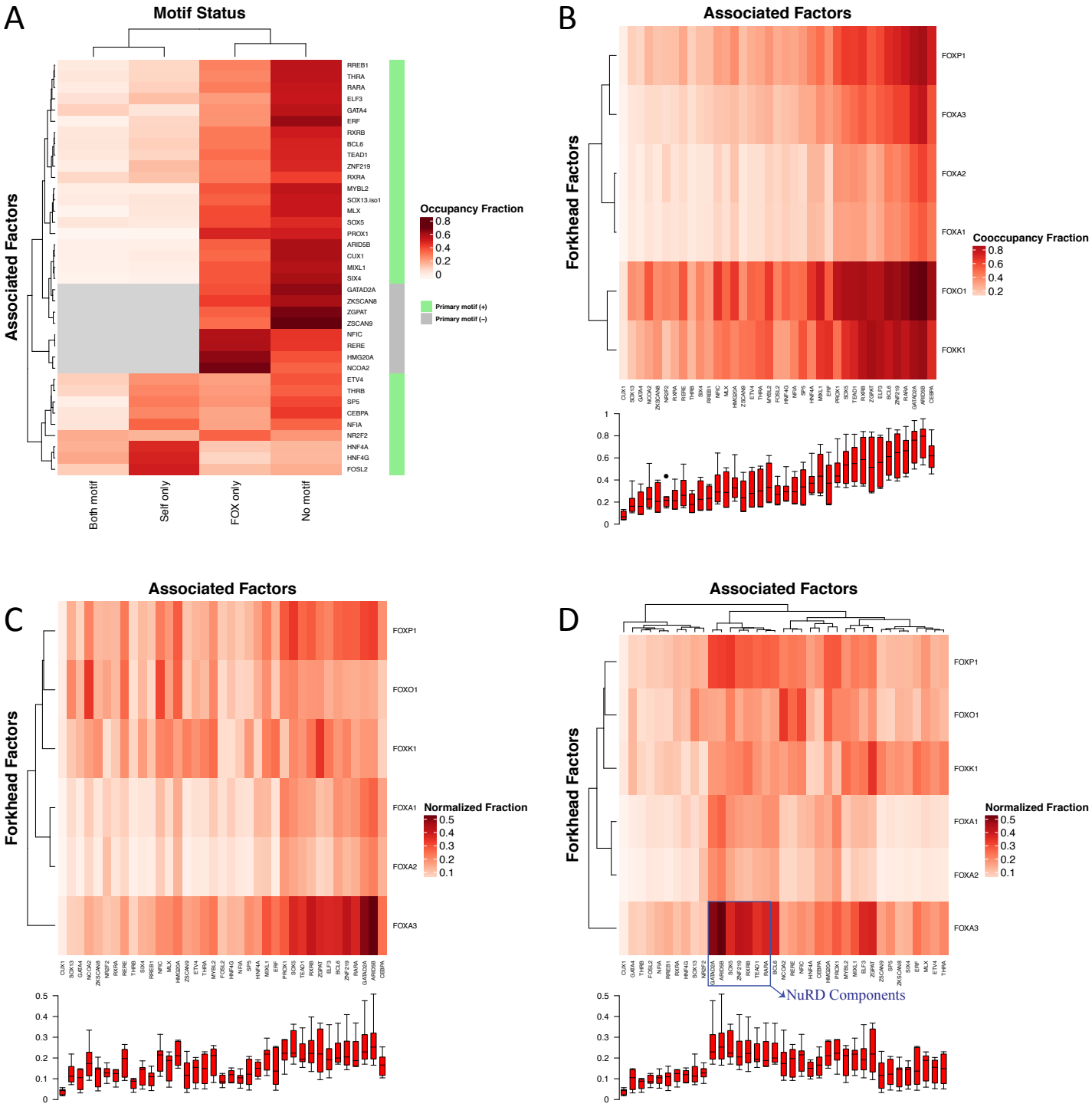
Supplementary Figure 7. Cumulative fraction of called motifs in our data compared to motifs in databases as scored by TomTom similarity E-value in **A)** CISBP (build 1.02) Homo sapiens database, **B)** JASPAR 2016 vertebrate database, and **C)** JASPAR 2018 vertebrate database.



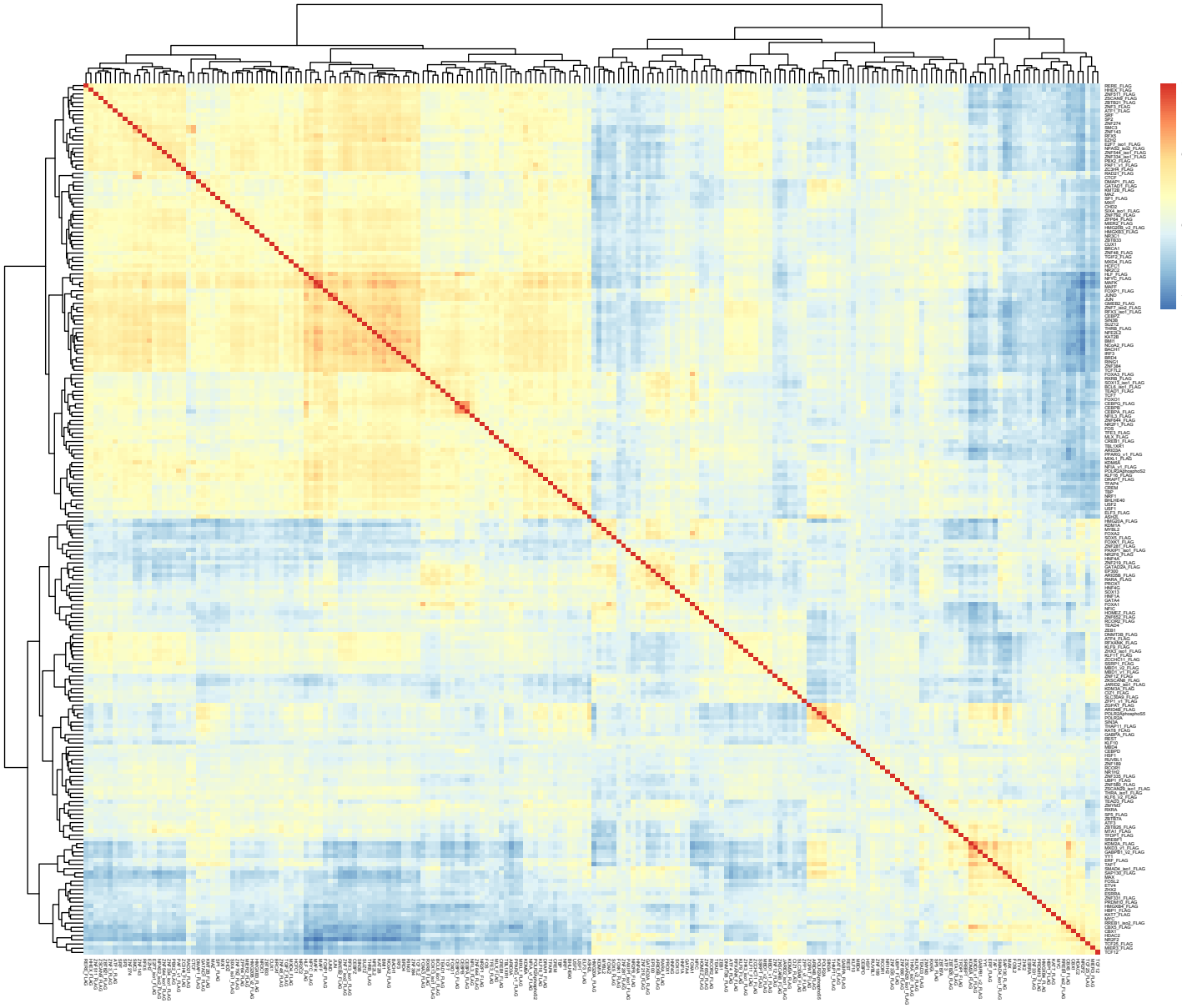
Supplementary Figure 8. Distribution of TF motifs by concordance (matching expected TF), discordance (matching different TF), and no match in CIS-BP database. Stacked bar plots are colored by main TF groups from previous unsupervised clustering.



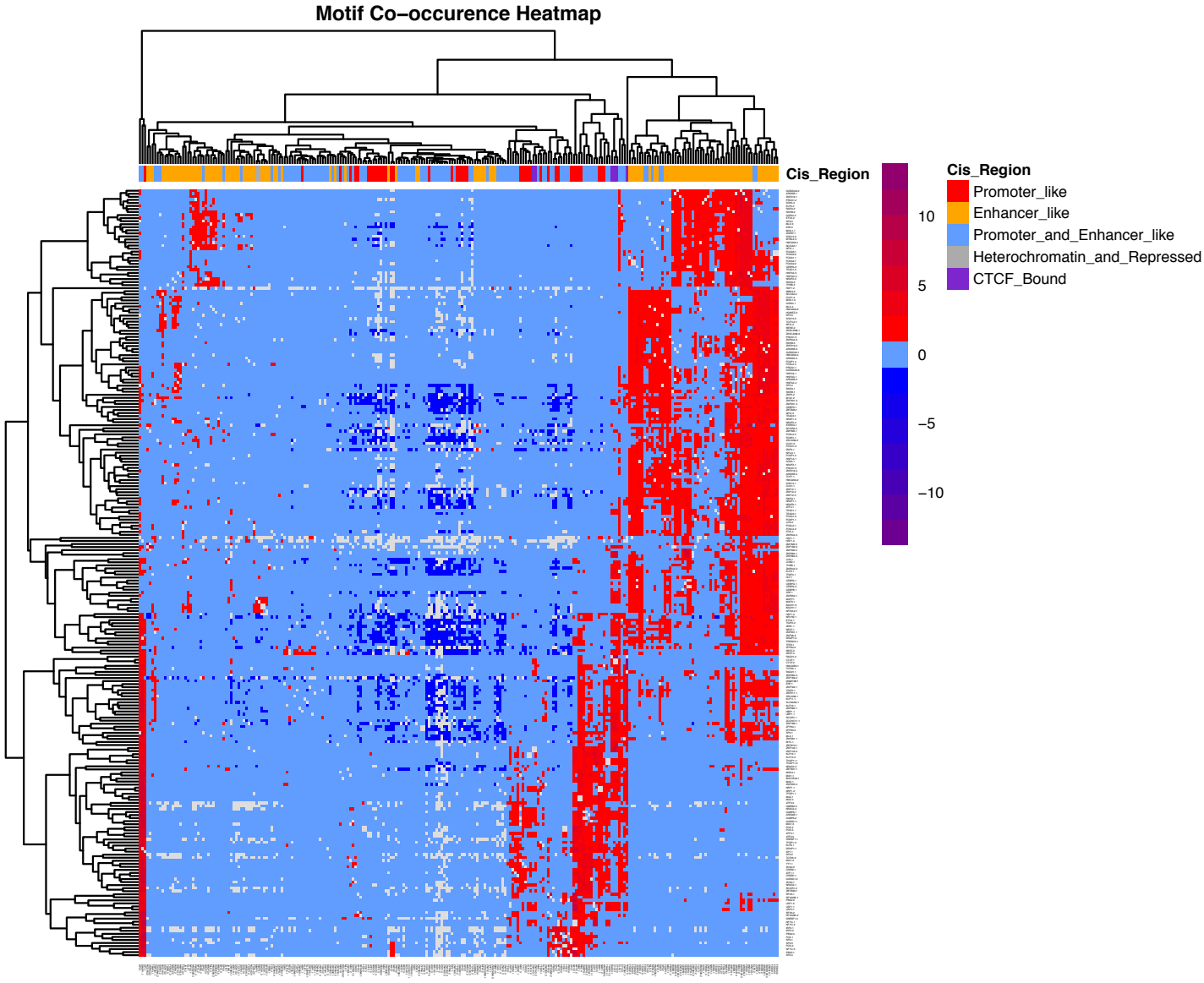
Supplementary Figure 9. A) Distribution of TF motifs highly dissimilar to all motifs in CIS-BP (y-axis) and their median offset distance from the center of peaks (x-axis). **B)** Stacked distribution of highly dissimilar motifs (no match; green) with similar (concordant; blue) and motif called for secondary factor (discordant; red) and their median offset distances from the peak center (x-axis).



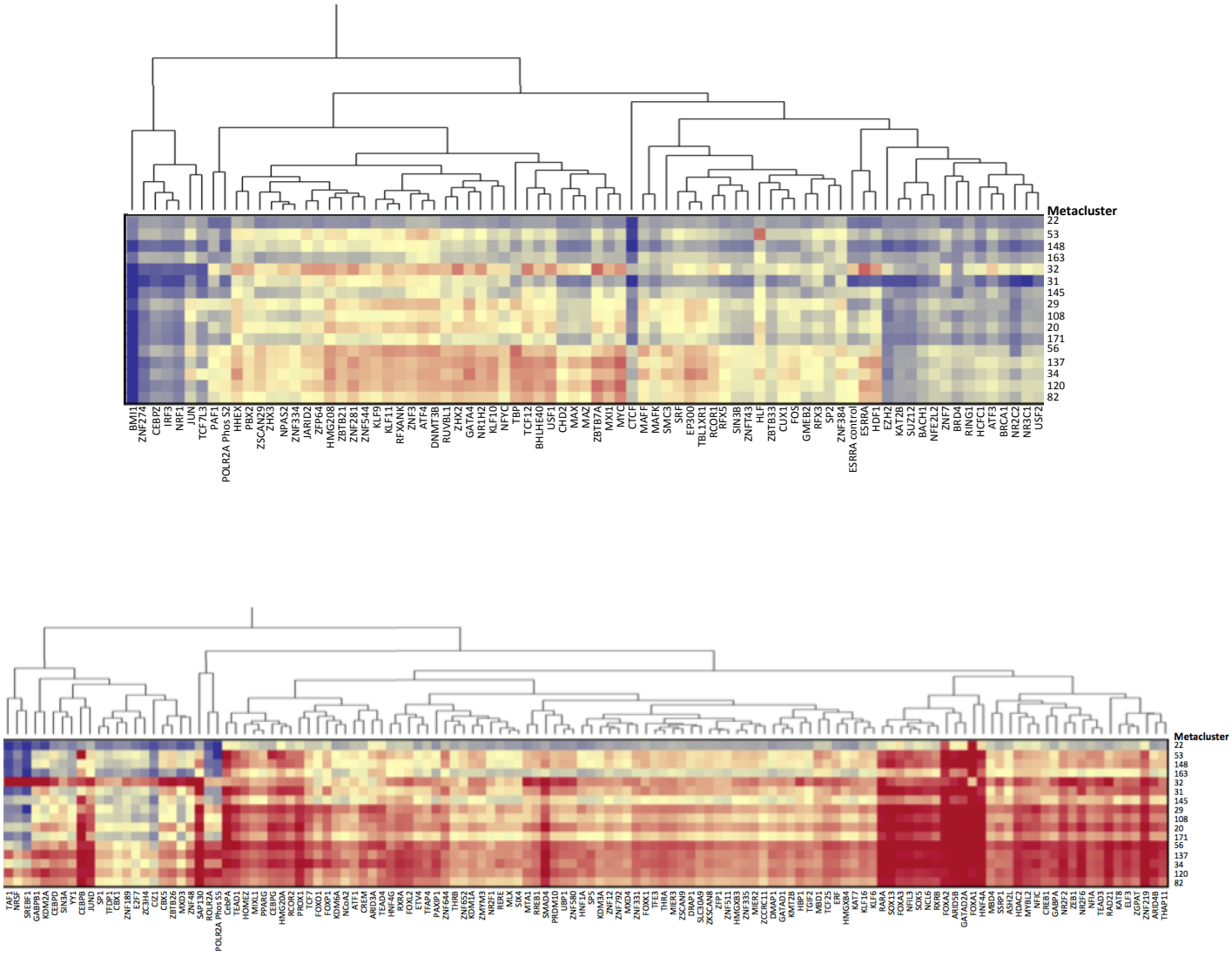
Supplementary Figure 10. A. 37 non-FOX TFs with a called Forkhead motif, with heatmap denoting fraction of called peaks with both a primary (matched to specific TF) motif and a FOX motif, with a primary motif but not FOX motif, with a FOX motif but no primary motif, and with neither a primary nor a FOX motif. The eight TFs with gray boxes do not have a known primary motif. **B.** Peak overlaps between the 37 TFs and six FOX factors for which we obtained ChIP-seq data; bar plots represent distribution of all FOX overlaps for each of the 37 factors. **C.** Same as B, but normalized for peak counts of each of the 37 factors. **D.** Same as C, but clustered vertically, revealing NuRD component clustering. Boxes indicate quartiles, whiskers are drawn to minimum/maximum values.



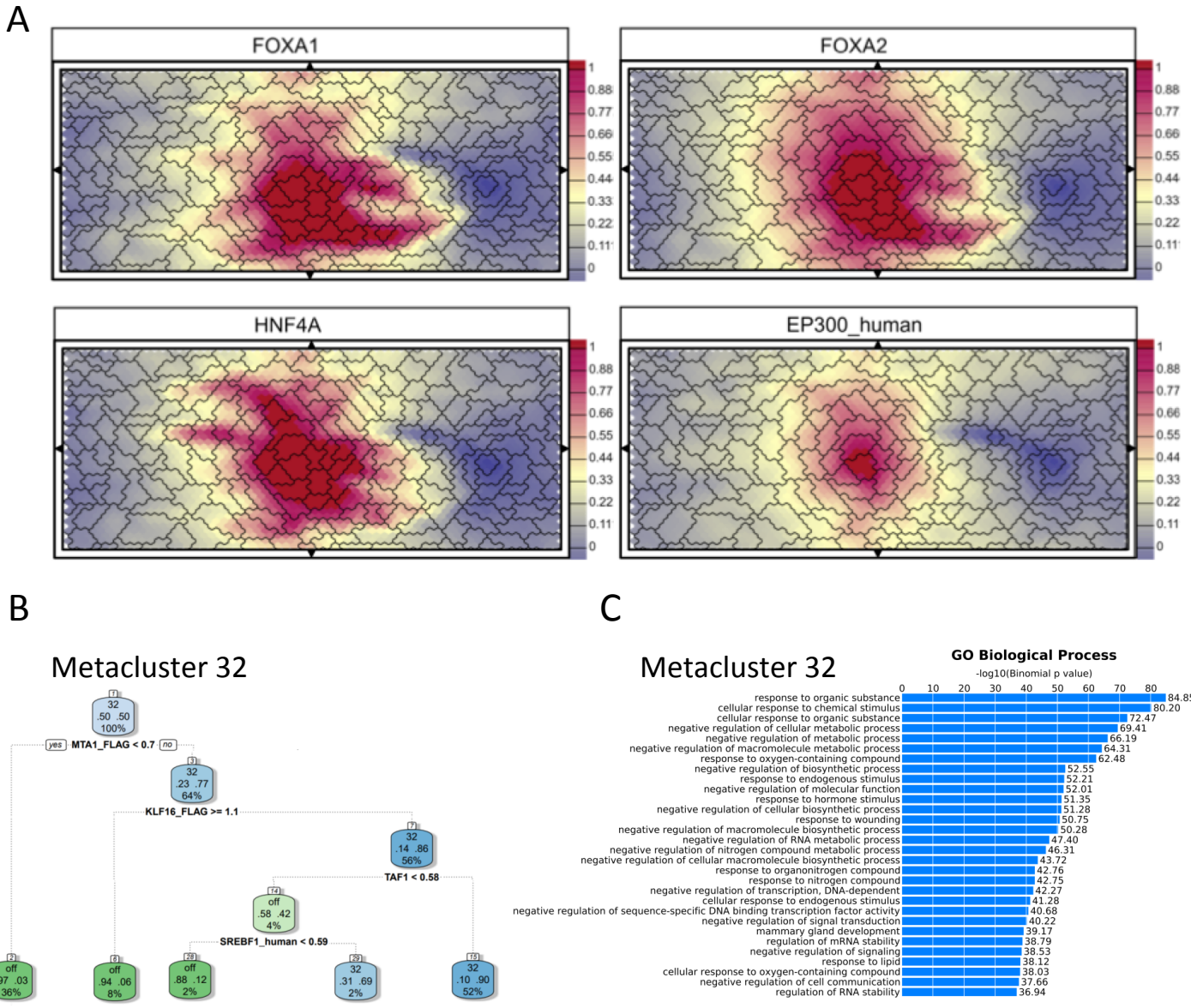
Supplementary Figure 11. Read count correlations between all 208 assayed factors, mean centered and squared, with unsupervised clustering.



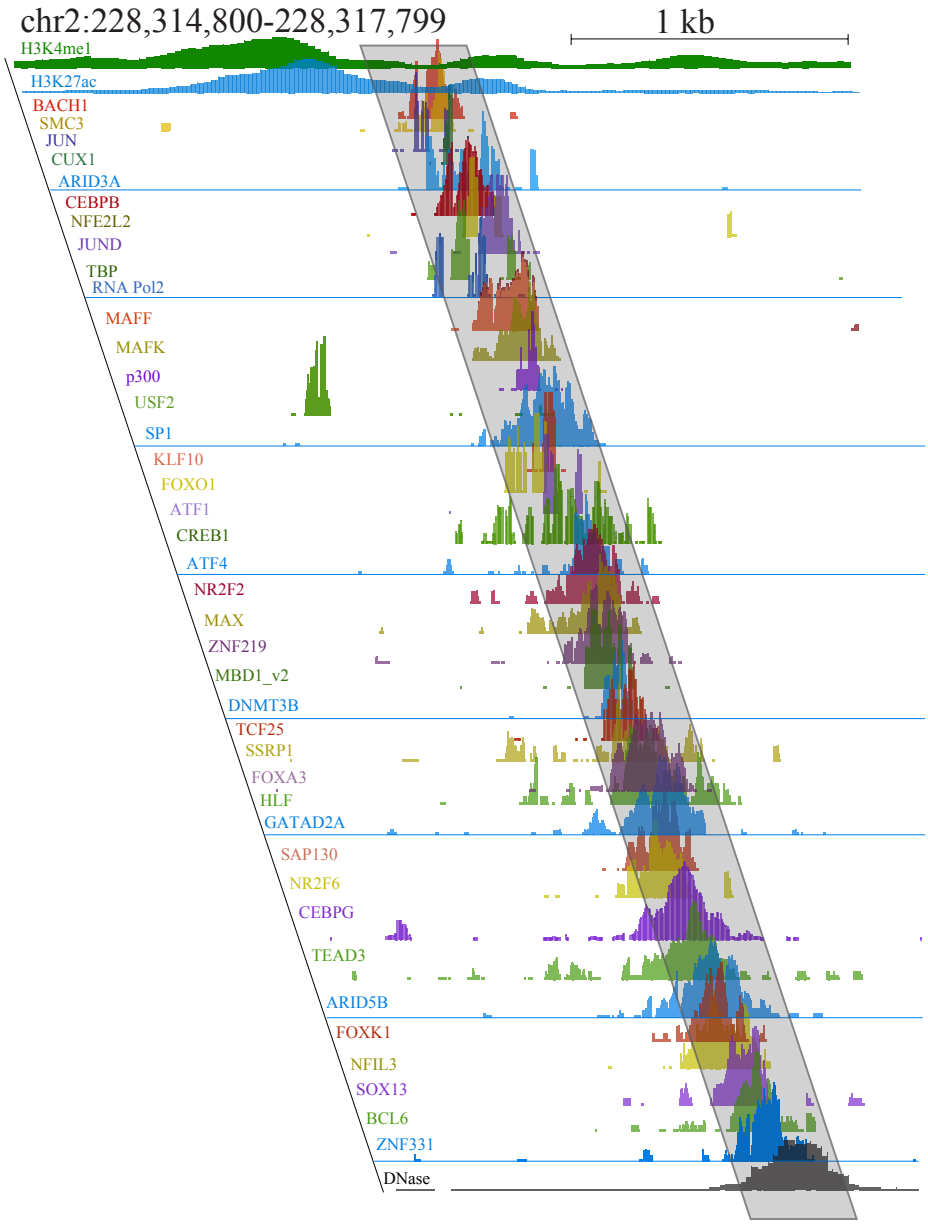
Supplementary Figure 12. Directional co-occurrence of motifs in ChIP-seq called peaks.



Supplementary Figure 13. Example heatmaps showing factor enrichment in 16 key metaclusters.



Supplementary Figure 14. A. SOMs for FOXA1, FOXA2, HNF4A, and EP300. **B.** Example decision tree showing presence/absence of factors for Metacluster 32. **C.** GREAT analysis of Metacluster 32 assigned genes likely regulated in this metacluster, and GO term analysis for these genes.



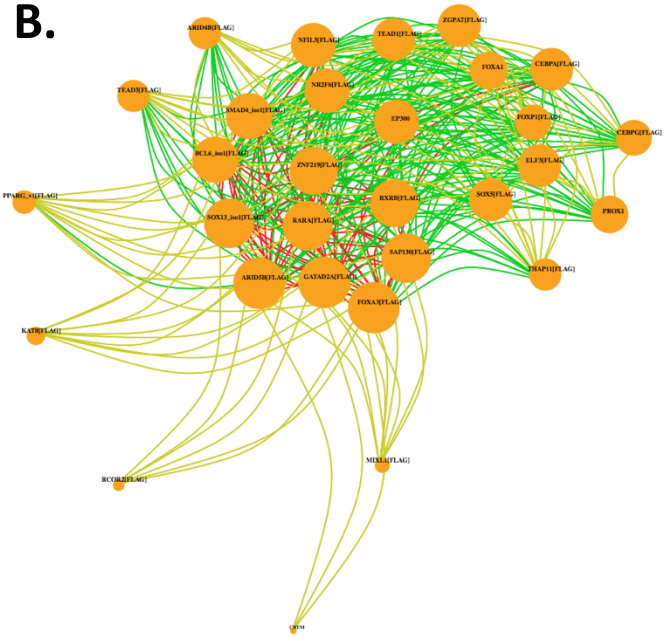
Supplementary Figure 15. Example of genomic site with many associated factors. Each track shows aligned ChIP-seq reads, and is slightly offset to better show peaks for each experiment.

A.

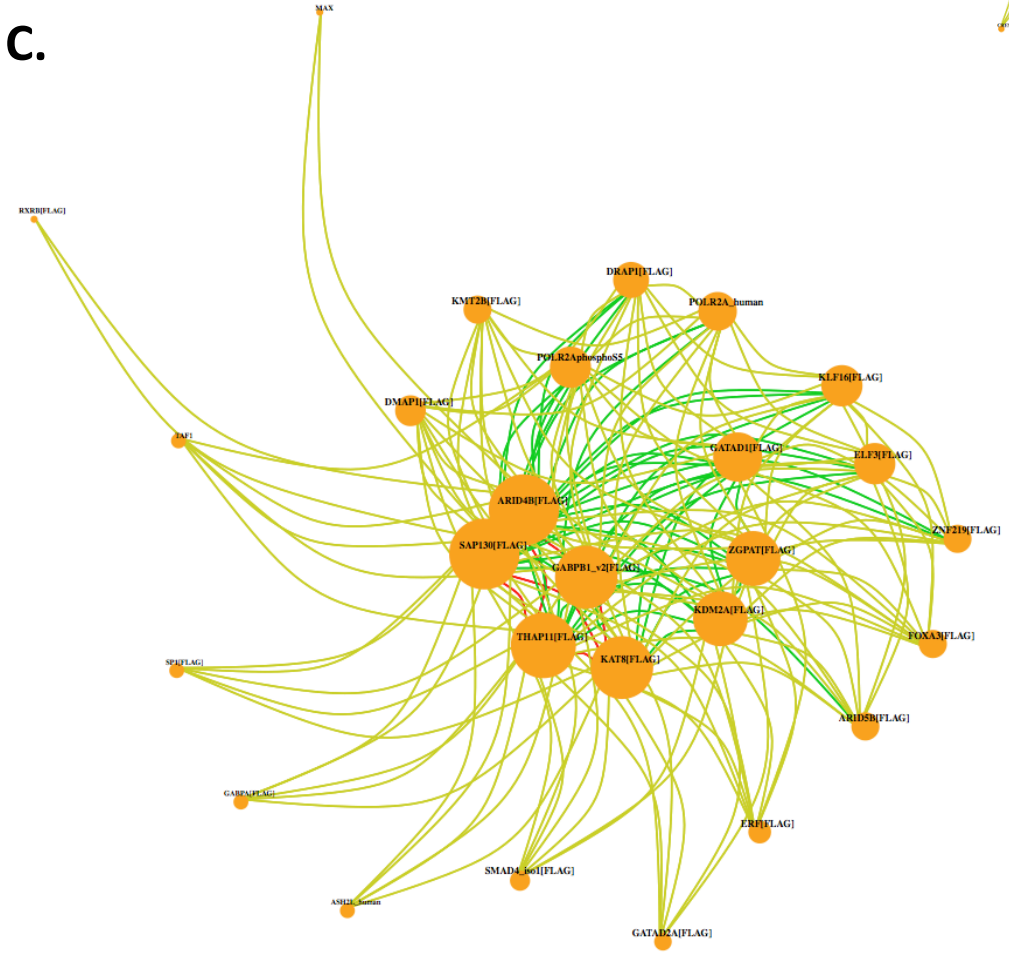
Highly Bound Regions at Promoters		
Top MSigDB Term	**Binomial P-Value**	**Fold Enrichment**
Metabolism of RNA	1.6e-147	50.5
Metabolism of Proteins	1.07e-146	34.0
Cell Cycle	2.29e-102	27.8
Cell Cycle, Mitotic	2.11e-90	30.0

Highly Bound Regions at Enhancers		
Top MSigDB Term	**Binomial P-Value**	**Fold Enrichment**
Metabolism of Lipids and Lipoproteins	2.42e-37	3.1
Fatty acid, triacylglycerol and ketone body metabolism	1.87e-24	4.0
FOXA2/FOXA3 TF networks	5.68e-17	7.6
HIF-1-alpha TF Network	8.86e-15	5.4

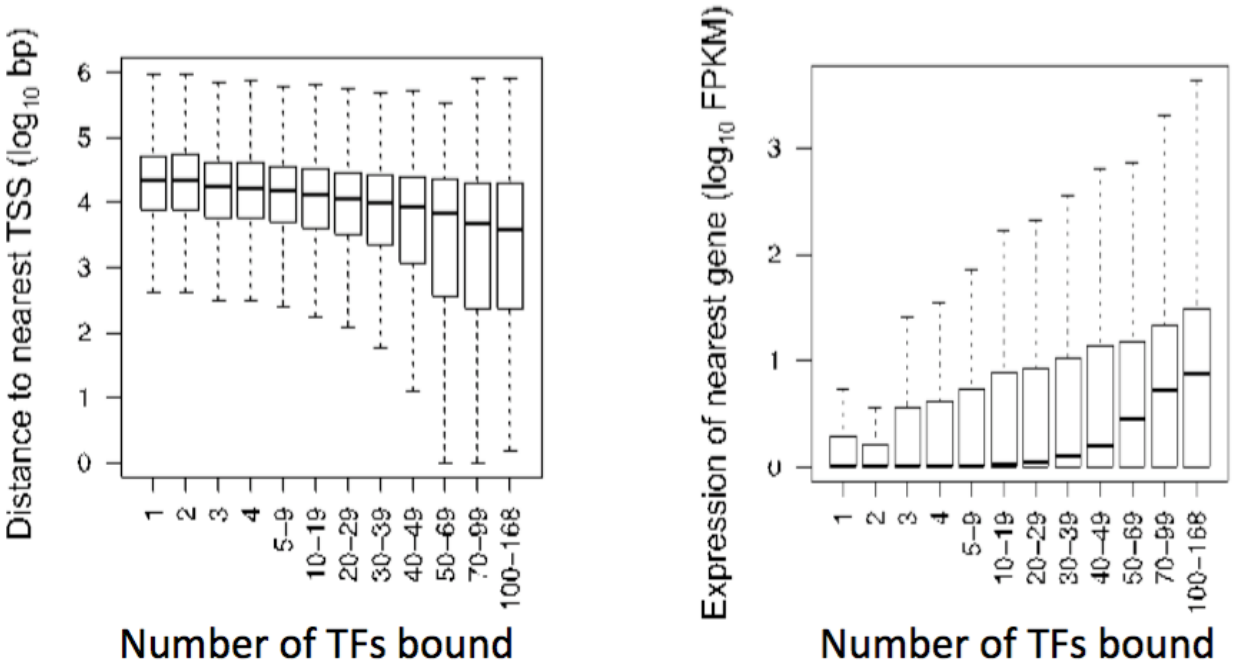
B.



C.

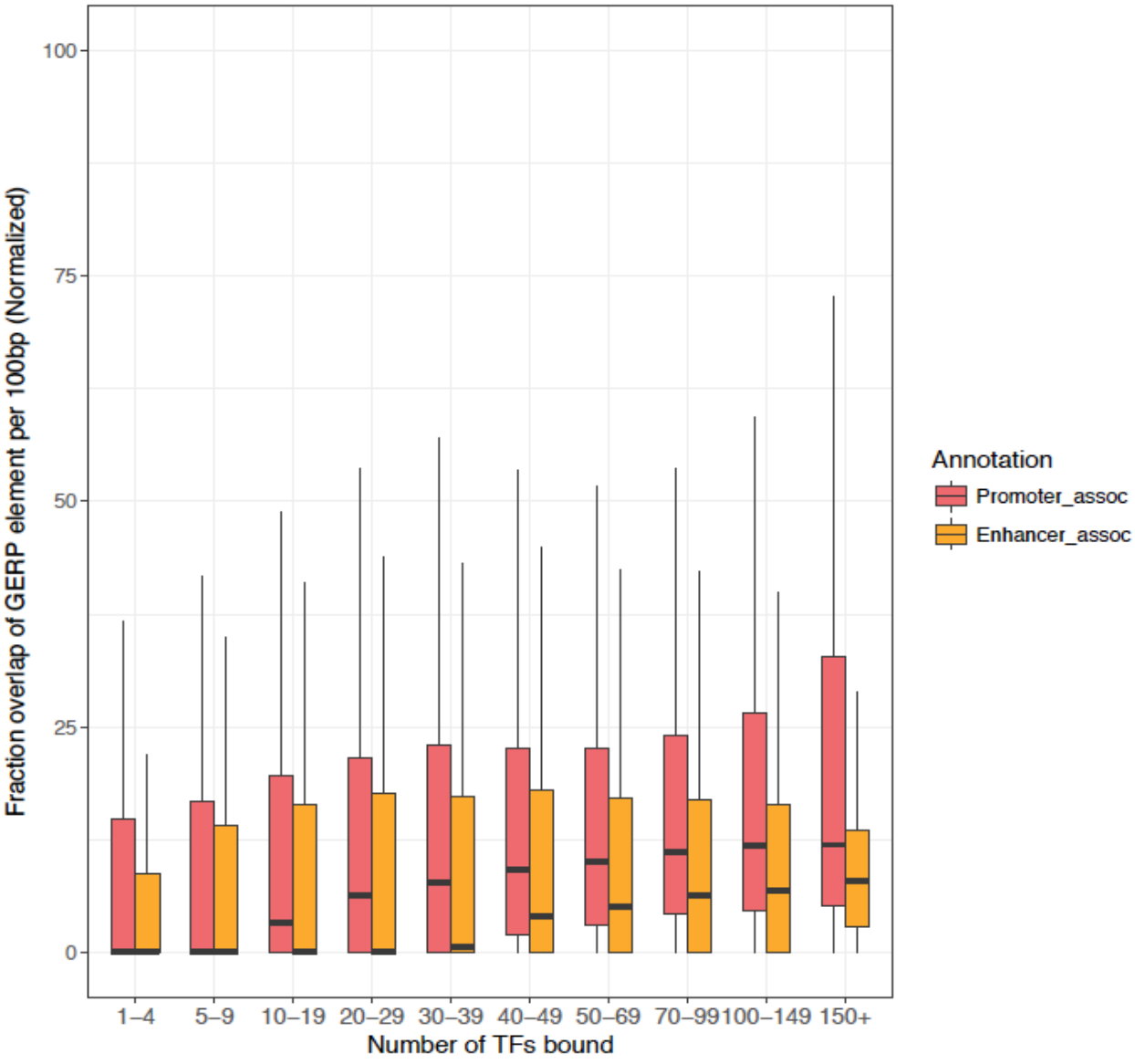


Supplementary Figure 16. A. Enrichment of biological pathways at HOT regions near enhancers or promoters. **B.** Co-binding analysis of factors bound to HOT regions classified as enhancer-like by IDEAS. **C.** Co-binding analysis of factors bound to HOT regions classified as promoter-like by IDEAS. Connecting lines indicate high overlap of peaks, with percentages shown by color of lines. For enhancers: yellow = 75-79%, green = 80-89%, red = 90+%. For promoters: yellow = 70-79%, green = 80-89%, red = 90+%.

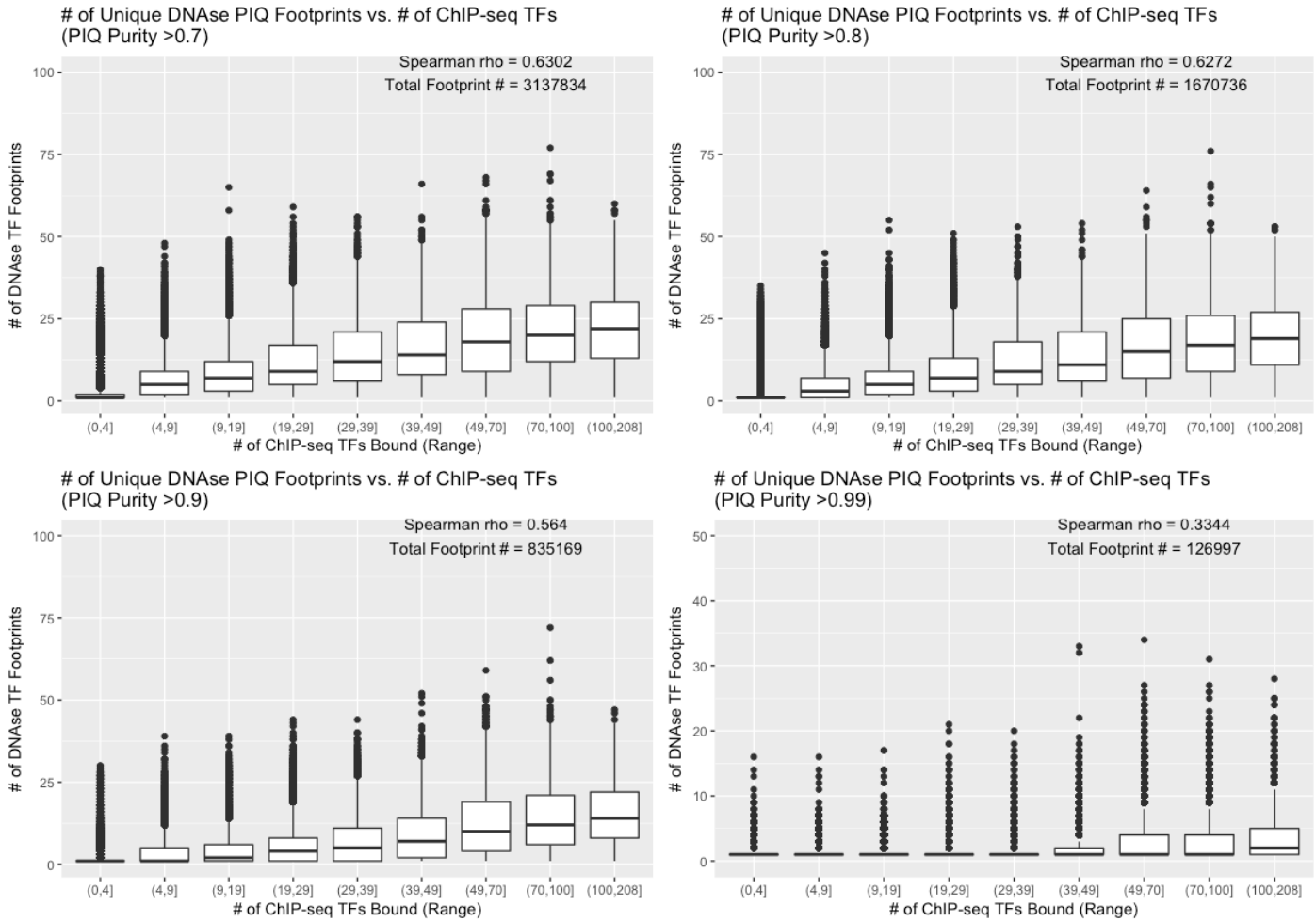


Supplementary Figure 17. Increasing numbers of factors bound at genomic sites (less than 2 kb in size) associate with decreasing distance to nearest TSS (left) and with increasing expression of nearest gene (right). Boxes represent middle two quartiles, whiskers are 1.5X inter-quartile range. TSS = Transcription Start Site, FPKM = Fragments Per Kilobase of transcript per Million reads.

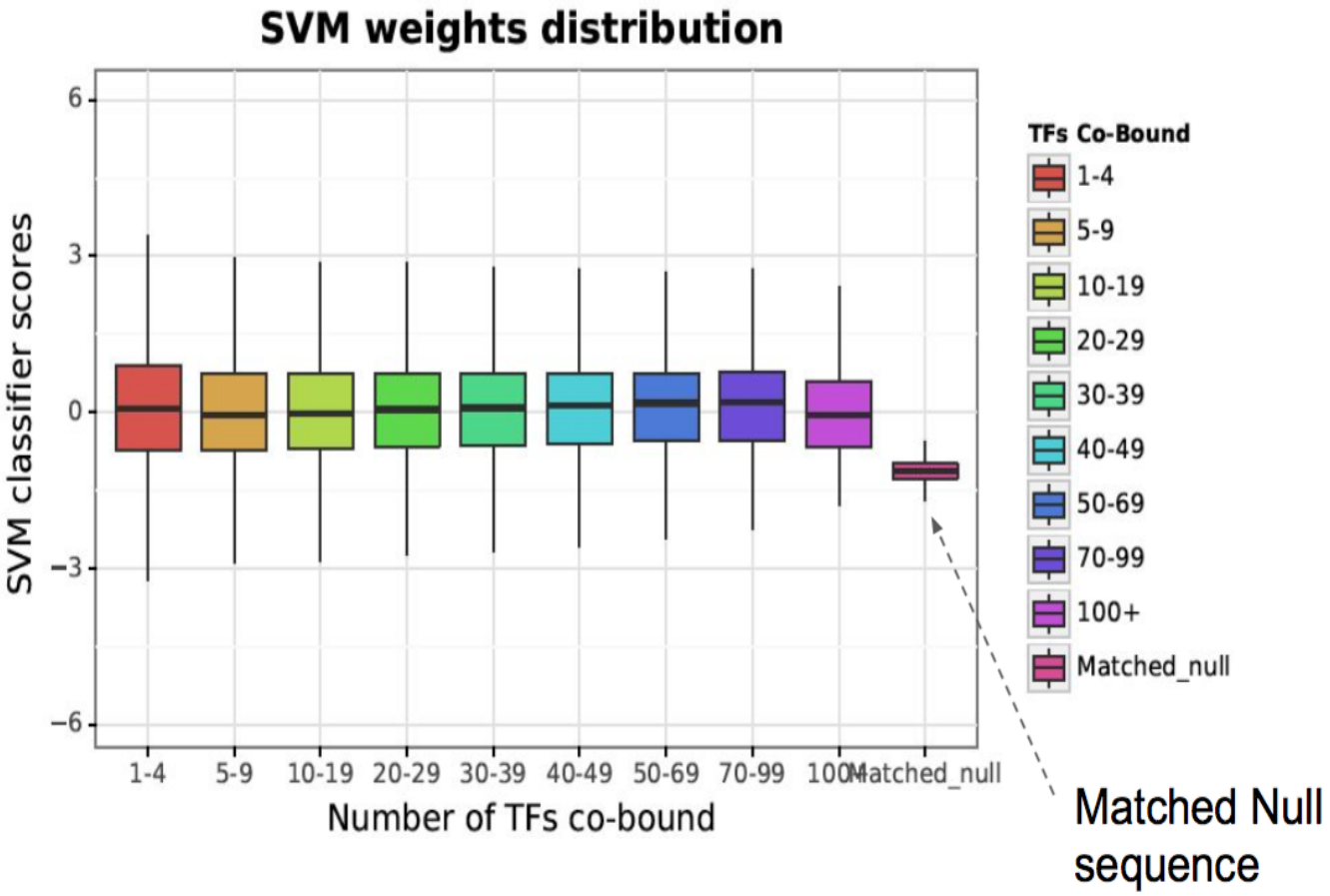
GERP Highly Constrained Element overlap with Number of TFs bound



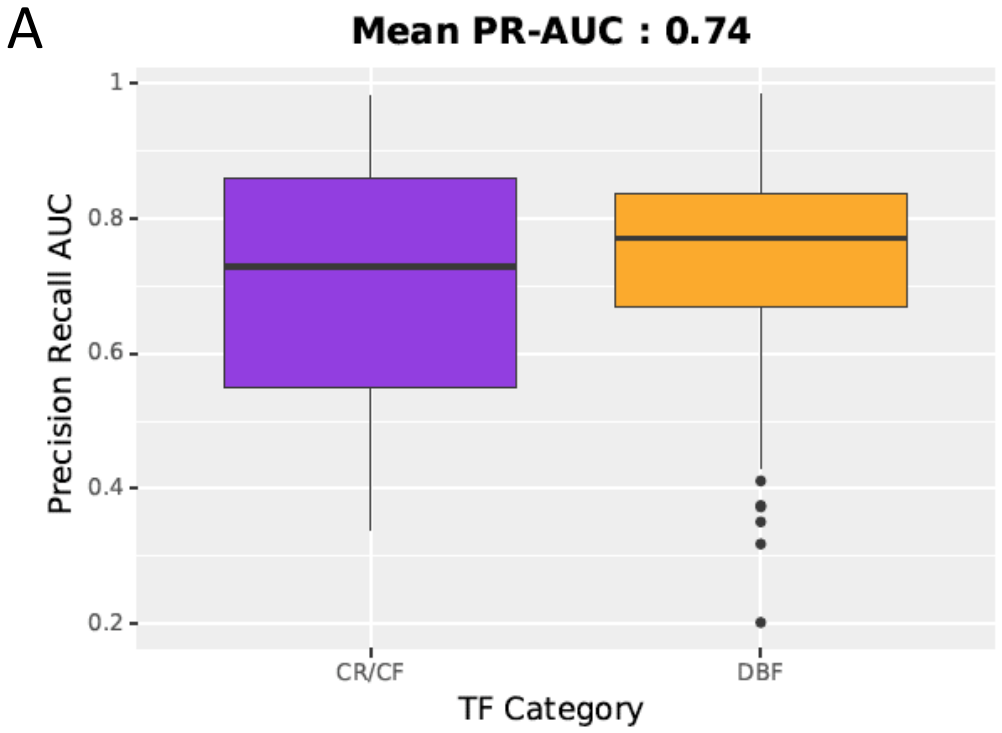
Supplementary Figure 18. Increasing numbers of factors bound at genomic sites correlate with increased evolutionary constraint as measured by GERP (Genomic Evolutionary Rate Profiling) showing incremental fraction overlap of highly constrained elements with factor-associated sites, for both promoter regions (red) and enhancer regions (orange). Boxes indicate quartiles, whiskers are drawn to maximum value.

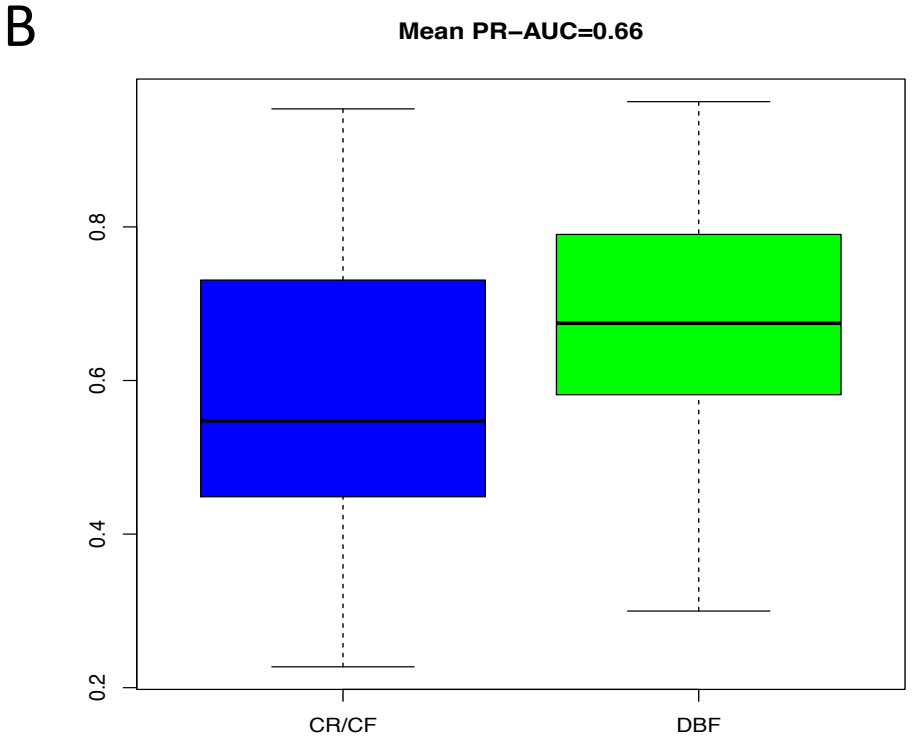


Supplementary Figure 19. Number of unique DNase PIQ footprints (y-axis) plotted by sites with varying number of associated factors (x-axis), plotted for varying thresholds of PIQ purity scores (upper left: > 0.7; upper right: > 0.8; lower left: > 0.9; lower right: > 0.99).

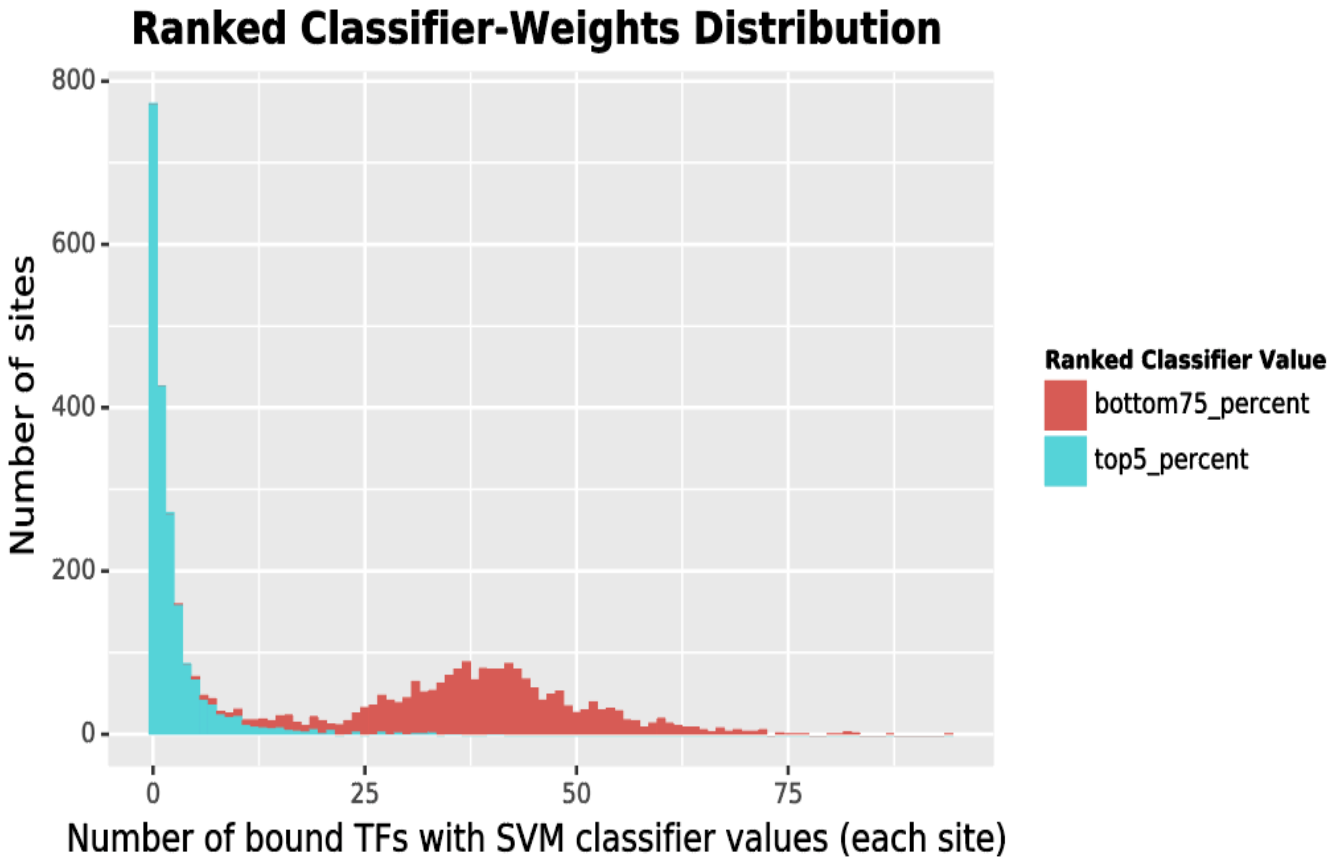


Supplementary Figure 20. Distribution of SVM classifier scores (y-axis) in sites with varying numbers of associated factors (x-axis). The scores remain relatively constant across sites and are significantly higher than the scores of classifier values in matched null sites.

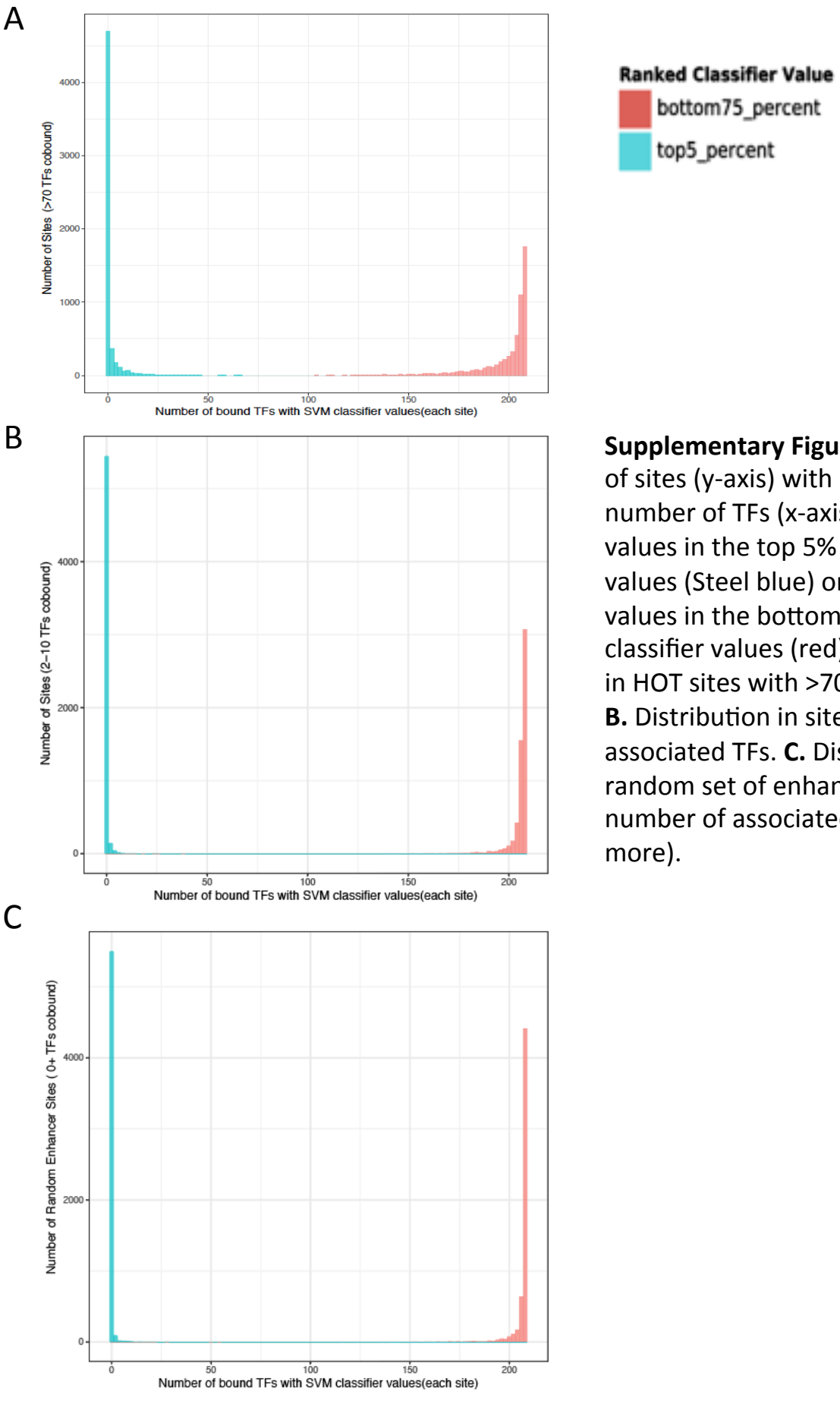




Supplementary Figure 21. SVM PR-AUC (Precision Recall Area Under Curve) scores for chromatin regulators and cofactors (CR/CF) and for DNA-binding transcription factors (DBF) **A)** Motif Level mean PR-AUC (0.74) **B)** Peak Level mean PR-AUC (0.66)

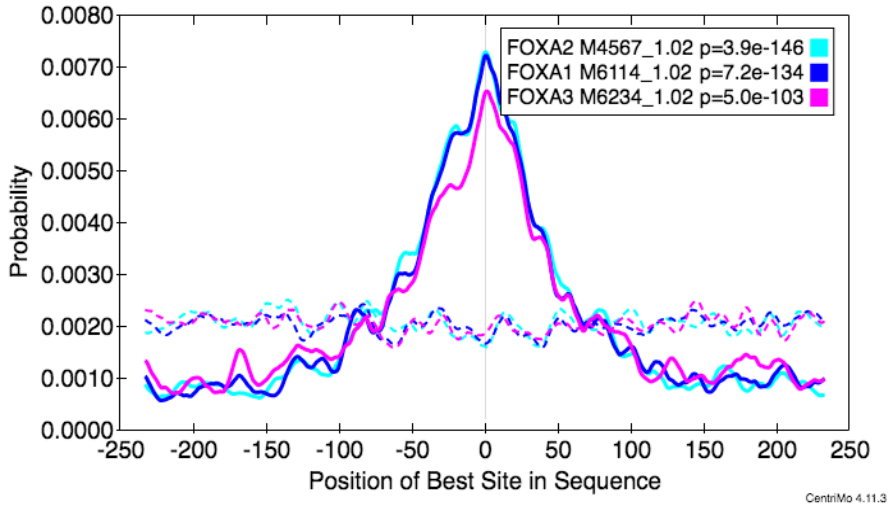


Supplementary Figure 22. Number of sites (y-axis) with measured number of TFs (x-axis) with classifier values in the top 5% of all classifier values (steel blue) or with classifier values in the bottom 75% of all classifier values (red) in highly bound regions, based on SVM scores of factor peaks associated to highly bound regions.

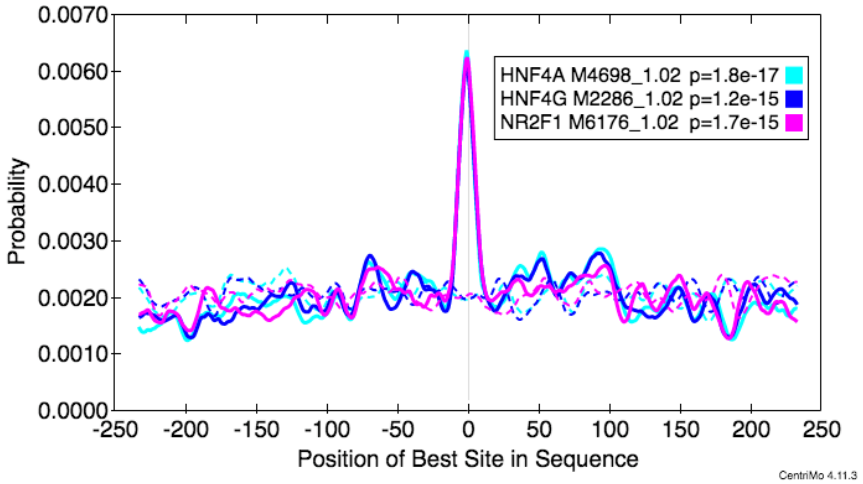


Supplementary Figure 23. Number of sites (y-axis) with measured number of TFs (x-axis) with classifier values in the top 5% of all classifier values (Steel blue) or with classifier values in the bottom 75% of all classifier values (red). **A.** Distribution in HOT sites with >70 associated TFs. **B.** Distribution in sites with 2-10 associated TFs. **C.** Distribution in random set of enhancers with any number of associated TFs (0 or more).

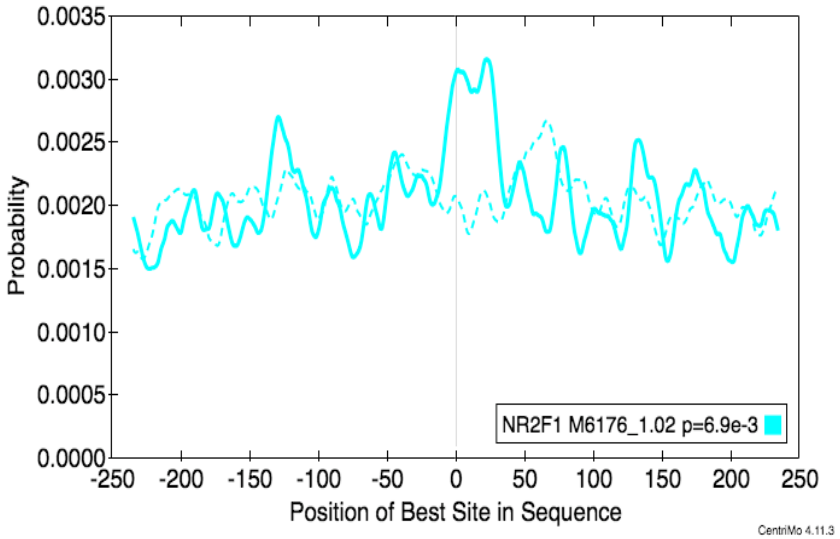
A



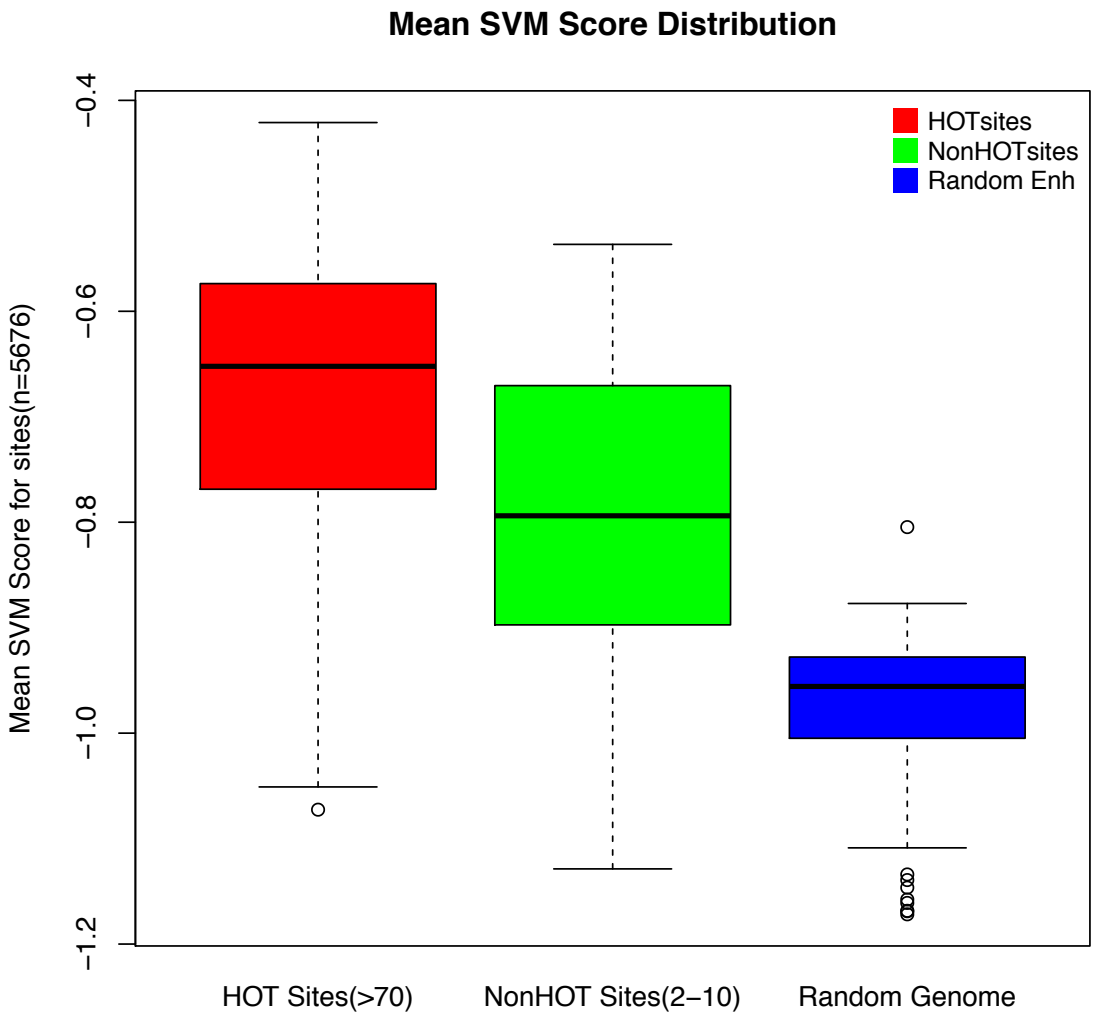
B



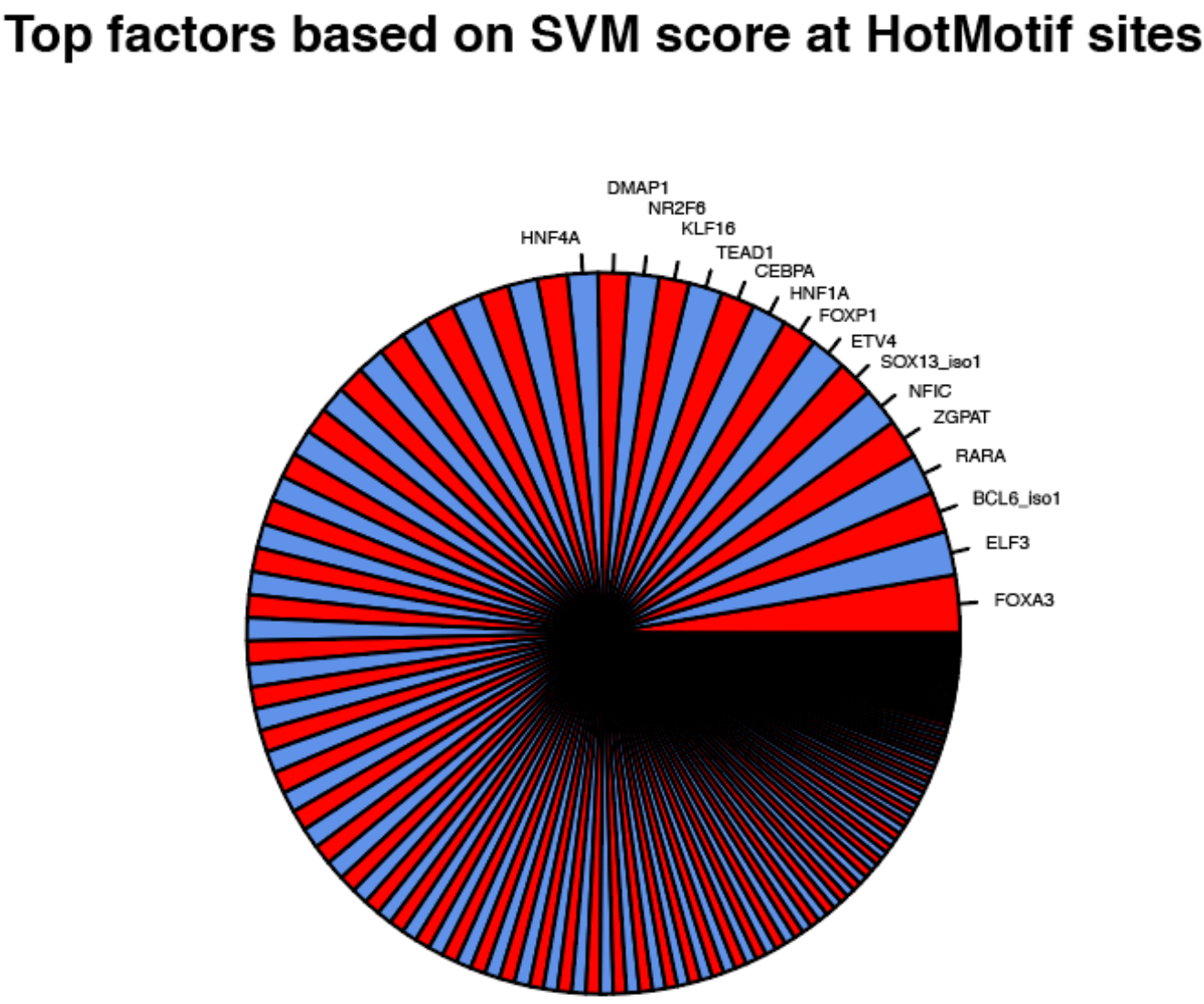
C



Supplementary Figure 23. Degree of Motif enrichment in highly bound regions for all HepG2 expressed factors with available motifs (n=365) for 3 Categories **A)** Top 3 motifs enriched in highly bound sites with 50+ TF (highest p-value = 3.9e-146) **B)** Top 3 motifs in Enhancer with 2-10 TFs (highest p-value = 1.8e-17) **C)** Top motif in Random Genome Enhancer with 0+ TFs (highest p-value=6.9e-3)



Supplementary Figure 24. Distribution of all SVM scores (y-axis) for HOT sites with >70 associated TFs (red), for sites with 2-10 associated TFs (green), and for random enhancer sites with 0+ TFs (blue).



Supplementary Figure 25. Pie chart showing fraction of HOT sites in which each factor has the highest SVM classifier value, indicating the strongest motif present.