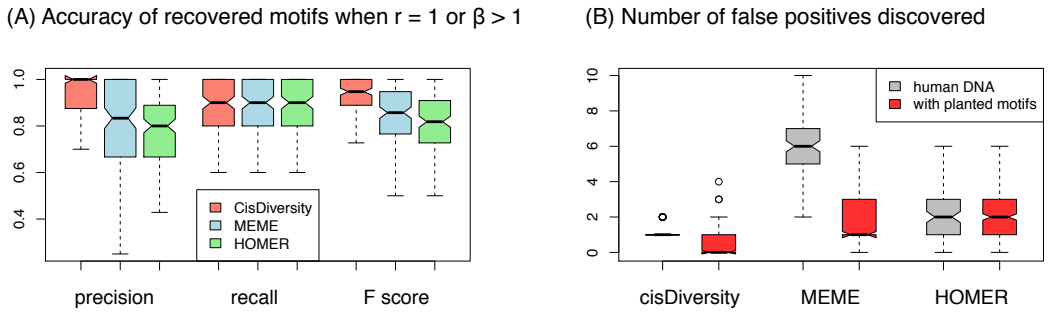
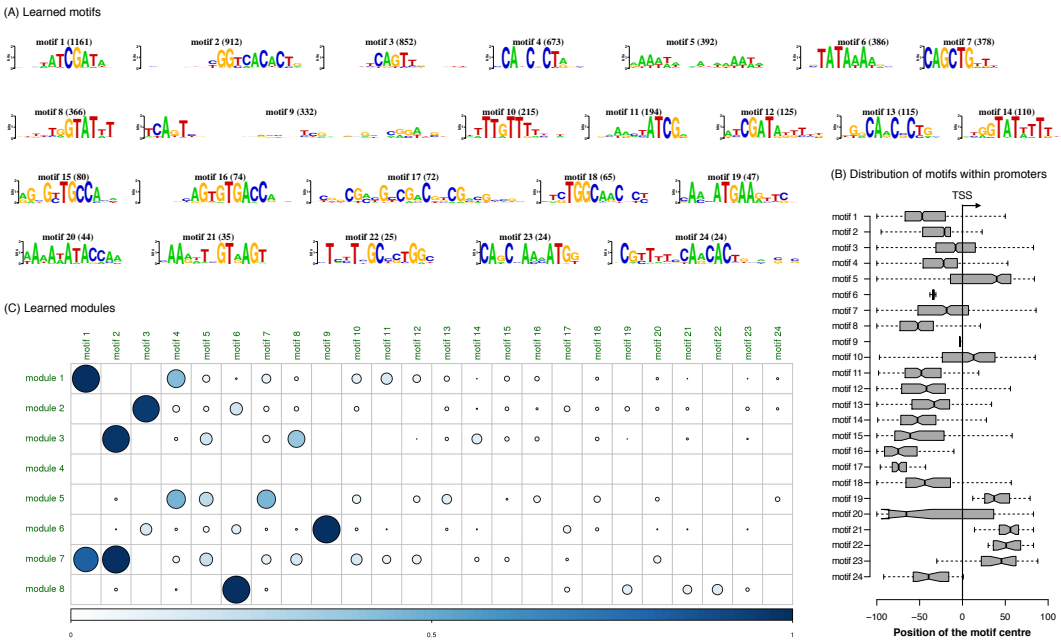
Assay	Cell-type	Accession (Source)
Paired-End Analysis of TSSs	mixed stage embryos (fly)	processed set (Ni et al., 2010)
CTCF ChIP-seq	white pre-pupa (fly)	GSM602326 (GEO)
CTCF ChIP-seq	H1 ESCs (human)	ENCFF991GTC (ENCODE)
GR ChIP-seq	treated A459 (human)	ENCFF835HHK (ENCODE)
EP300 ChIP-seq	treated A459 (human)	ENCFF774PLX (ENCODE)
eRNAs from GRO-cap	K562 (human)	processed set of eRNAs (Azofeifa et al., 2018) from SRR1552480 (SRA)
EP300 ChIP-seq	K562 (human)	ENCFF821GNB (ENCODE)
STARR-seq	K562 (human)	processed (Lee et al., 2020) from ENCFF717VJK & ENCFF394DBM (ENCODE)
dELSs (ENCODE III)	K562 (human)	dELS from ENCFF464BRU (ENCODE)
pELSs (ENCODE III)	K562 (human)	pELS from ENCFF464BRU (ENCODE)
scATAC-seq	2–4hr after egg-laying (fly)	Table S1 from Cusanovich et al. (2018)
DNase-seq	H1 ESCs (human)	ENCFF030XPN (ENCODE)
ATAC-seq	20 mouse tissues	processed set (Liu et al., 2019)

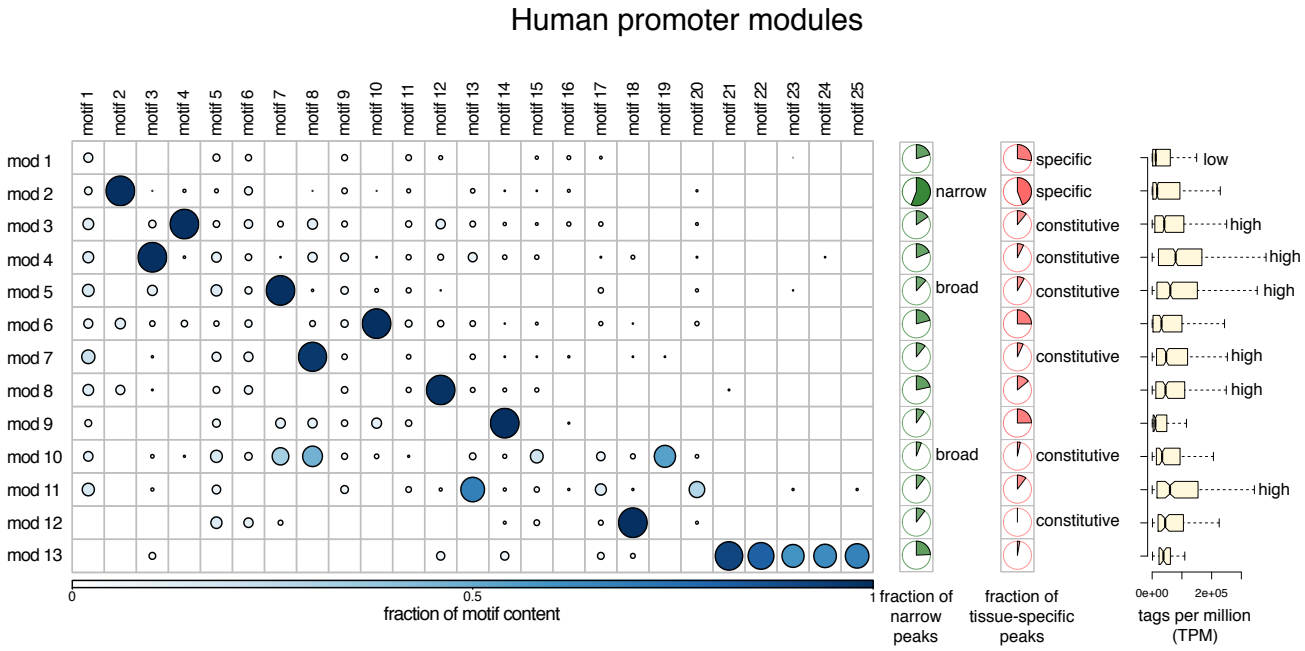
Supplemental Table S1: Datasets on which cisDiversity was run in this study. Processed data/peaks were used from the original study in each case.



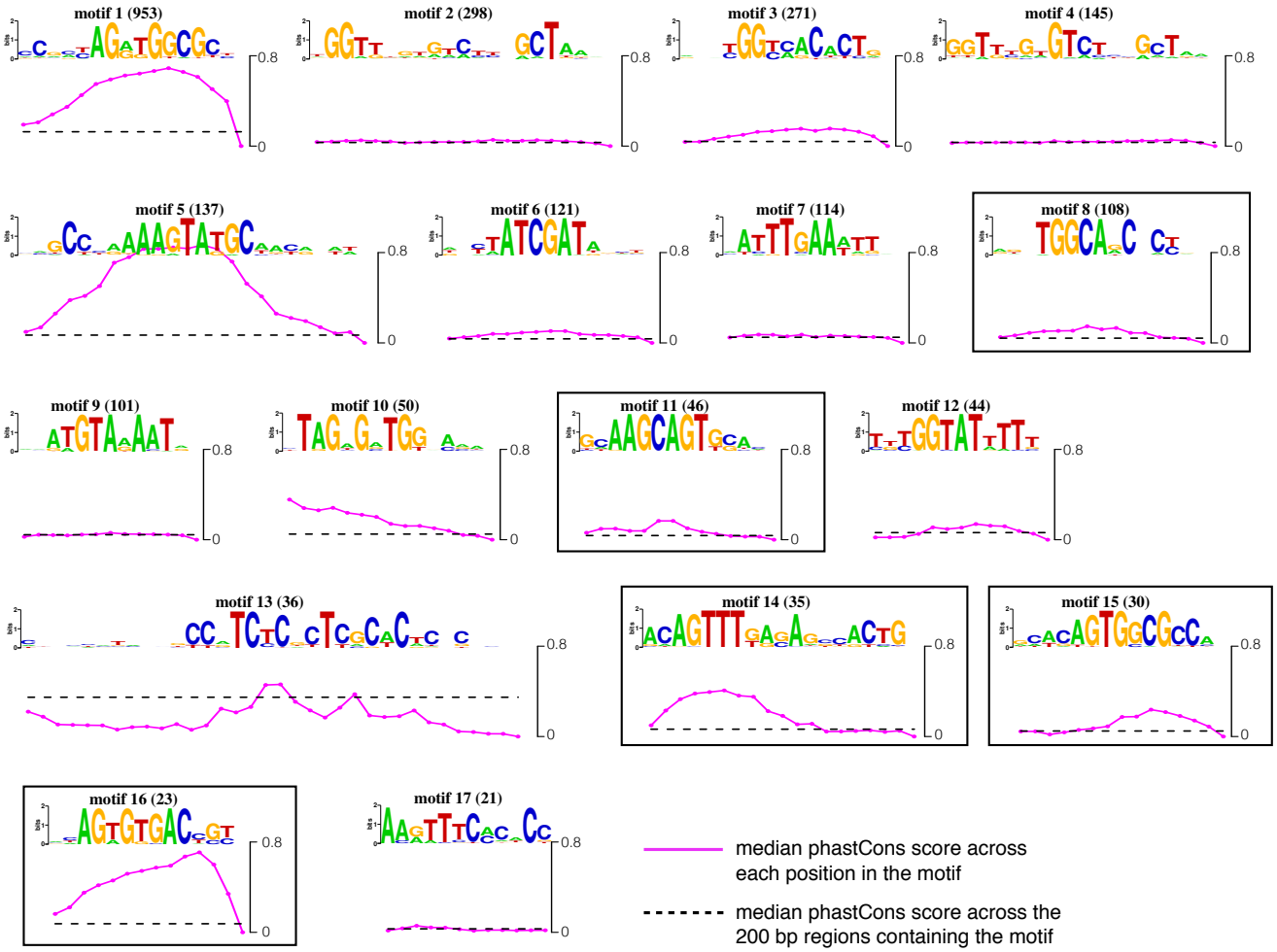
Supplemental Figure S1: When there is a single module ($r = 1$) or the modules are less separable ($\beta = 10$), the performance of cisDiversity in terms of recovering planted motifs is still better or comparable to that of MEME or HOMER.



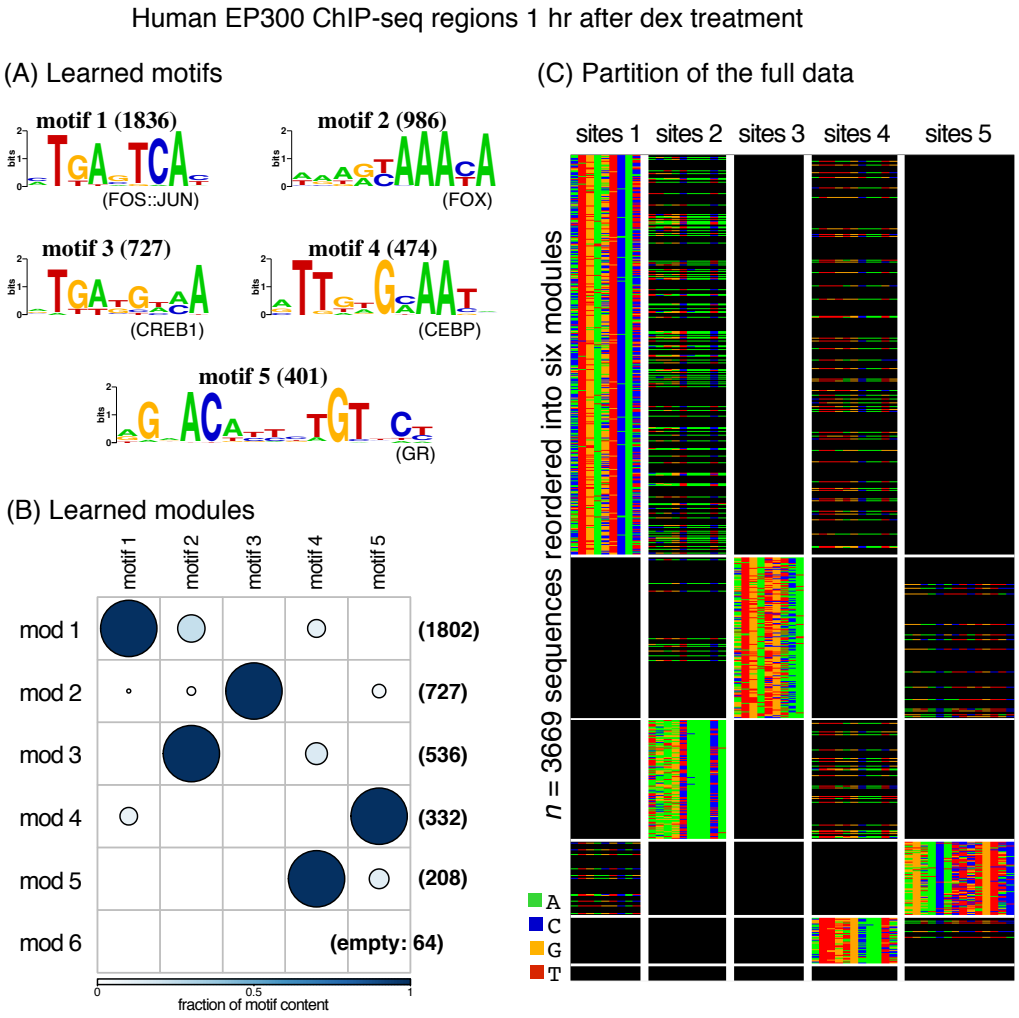
Supplemental Figure S2: The complete model learned on 4159 fly promoters.



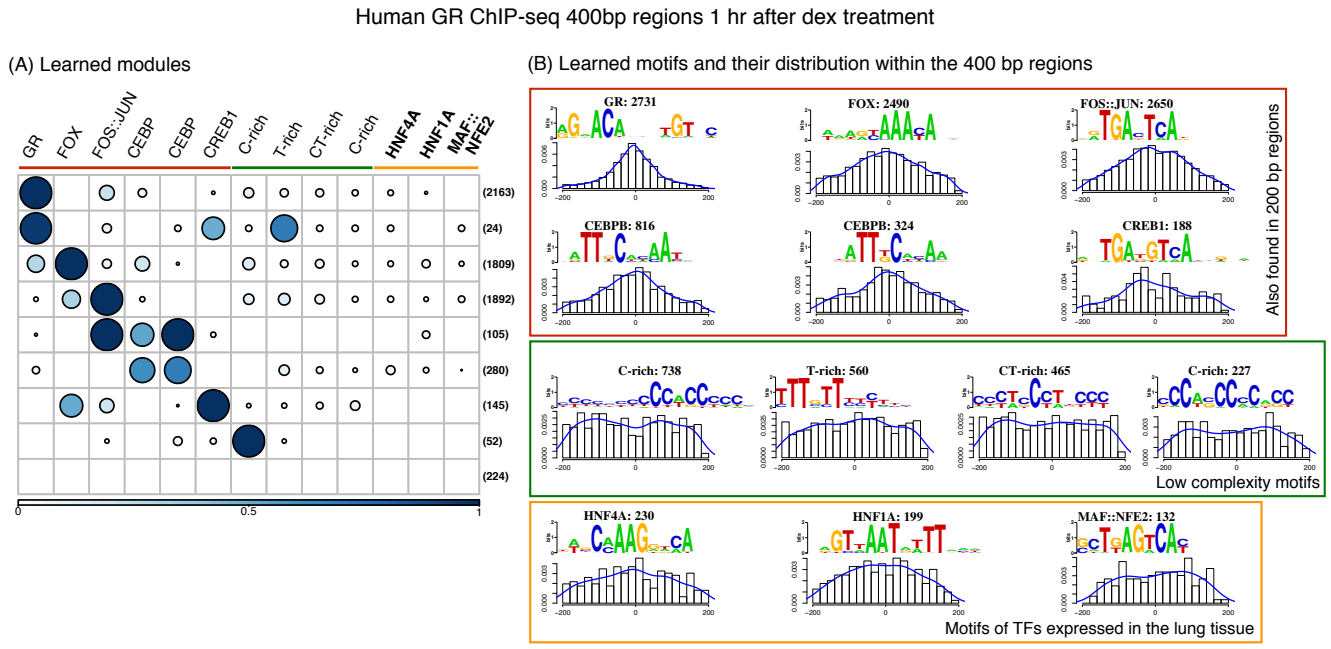
Supplemental Figure S3: Modules learned on the pooled human CAGE data. The motifs are part of the main text. Promoters are defined as narrow if the interquantile range of the tags (Frith and the FANTOM consortium, 2014) falls below three bases. Promoters are defined as tissue-specifically expressed if their Shannon entropy falls below 6. The boxplot of pooled tag counts is shown for each module. If a module is significantly enriched or depleted (corrected $p < 10^{-10}$) with one of the three properties, the property is displayed.



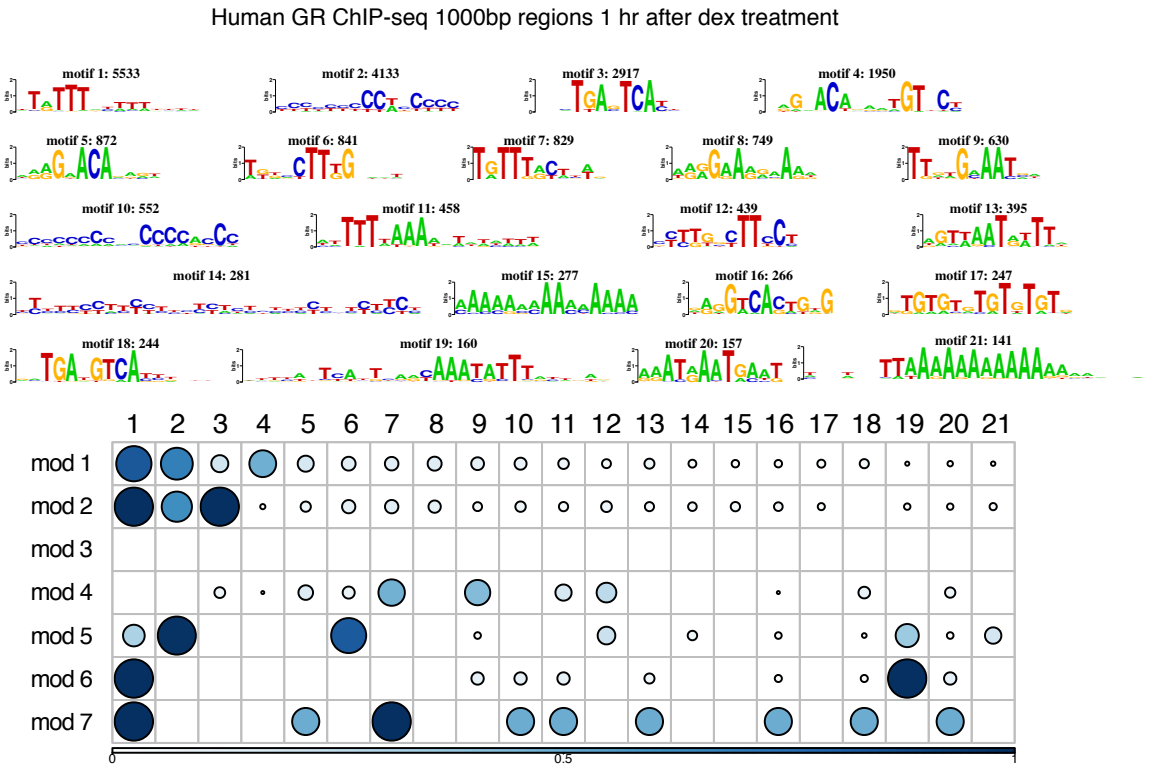
Supplemental Figure S4: Median phastCons scores across 15 fly genomes are shown in pink at all positions within each motif. The dotted line shows the overall median score across each region that contains the particular motif. Su(Hw) and CTCF are the most conserved. In contrast, the recently discovered Pita motifs (motifs 2&4) and the Ibf motif (motif 9) are not conserved. Unknown motifs, but which are relatively well-conserved, are shown in black boxes. These warrant further investigation.



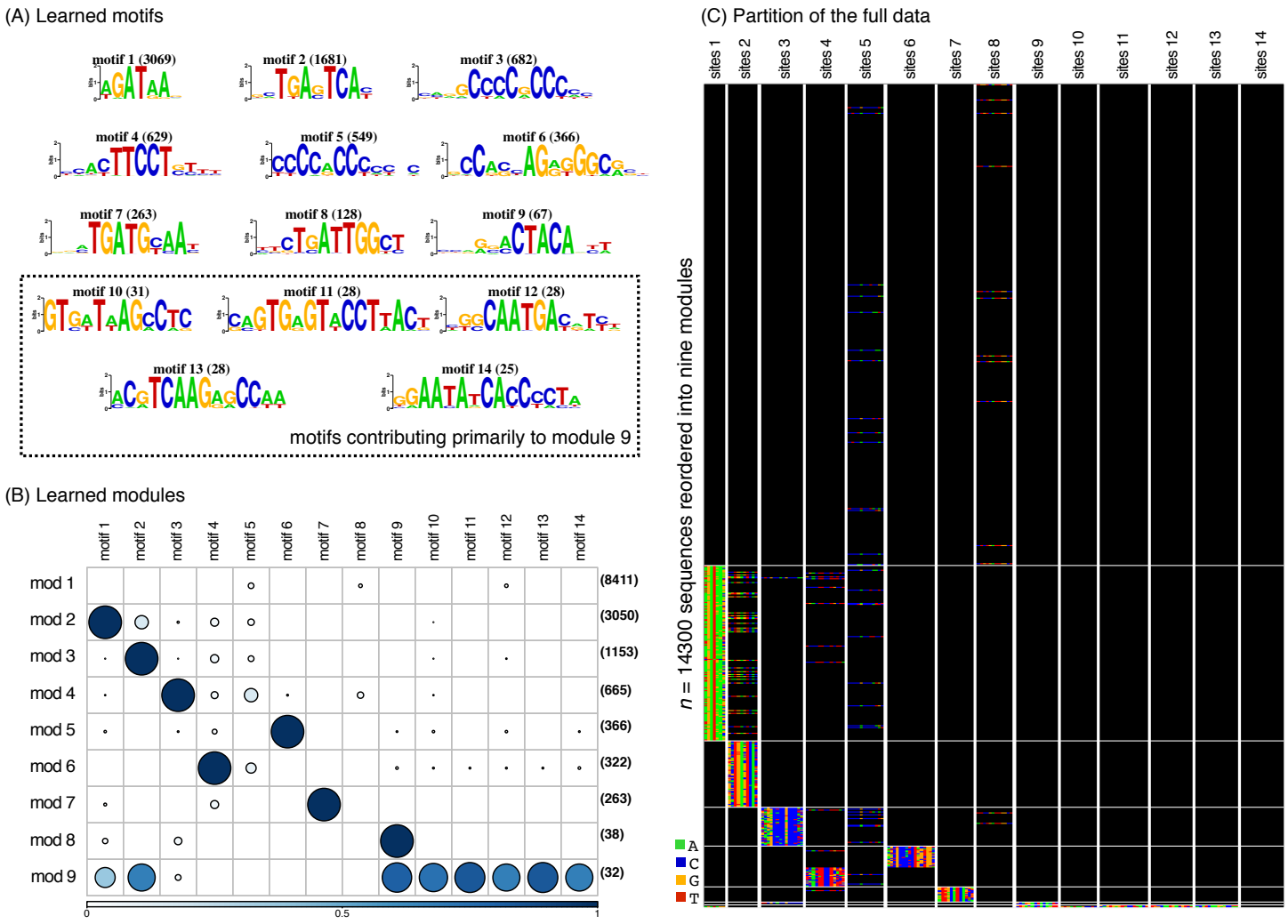
Supplemental Figure S5: Results of CISDIVERSITY on the EP300 ChIP-seq data in A549 cells 1 hr after being treated with dex. The motifs are near-identical to those learned on GR ChIP-seq data in the same cell-type, but the distribution of the modules is different.



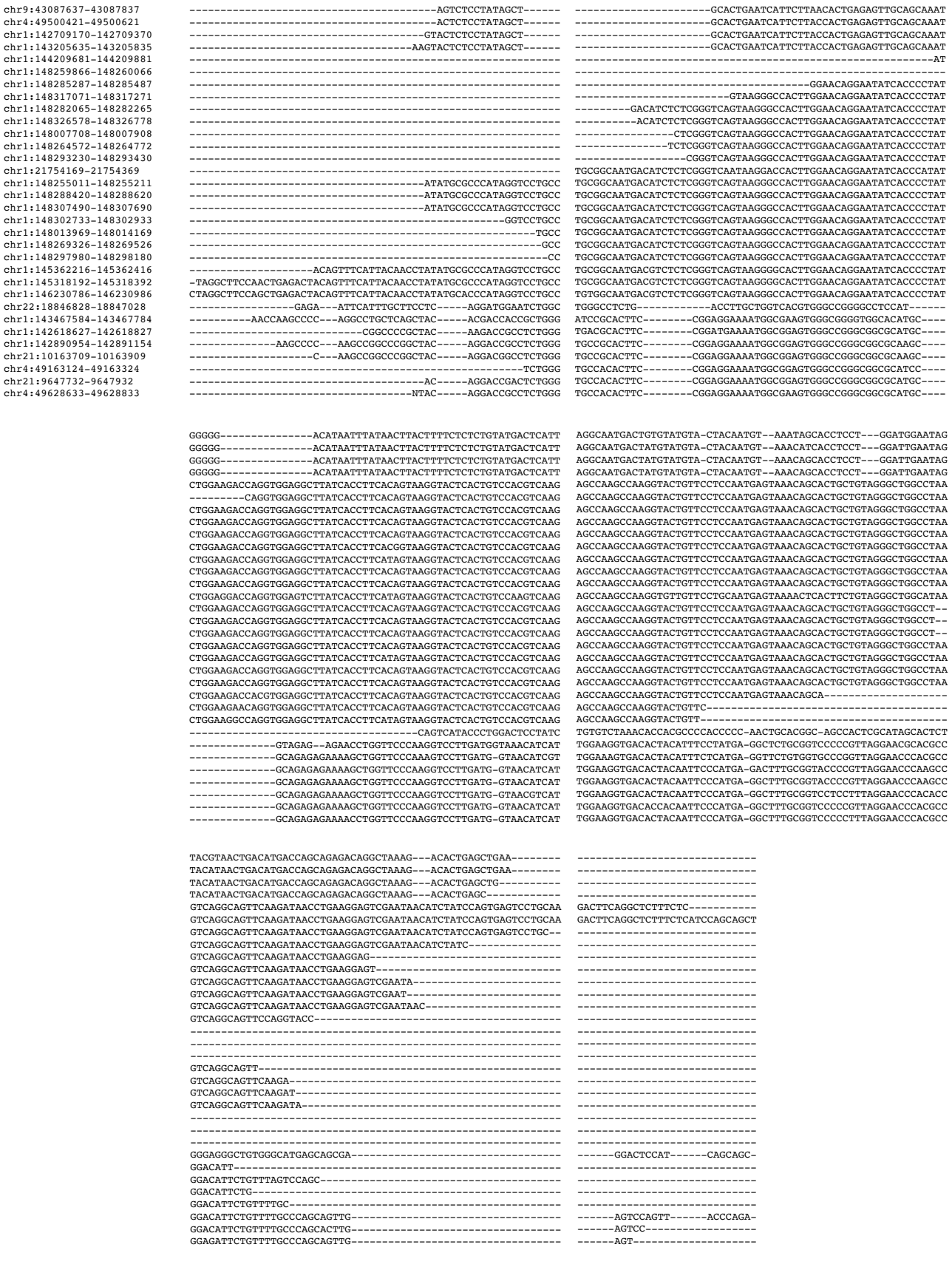
Supplemental Figure S6: Results of CISDIVERSITY on a larger 400bp neighborhood of GR-bound summits. The modules and motifs are reordered based on the motifs that are also found in the 200bp neighborhood (red box), those that are more spread out along the region and have low-complexity (green box), and those that resemble motifs of TF active in the lung tissue (orange box). The histogram depicts the distribution of the binding sites of each motif along the 400bp region.



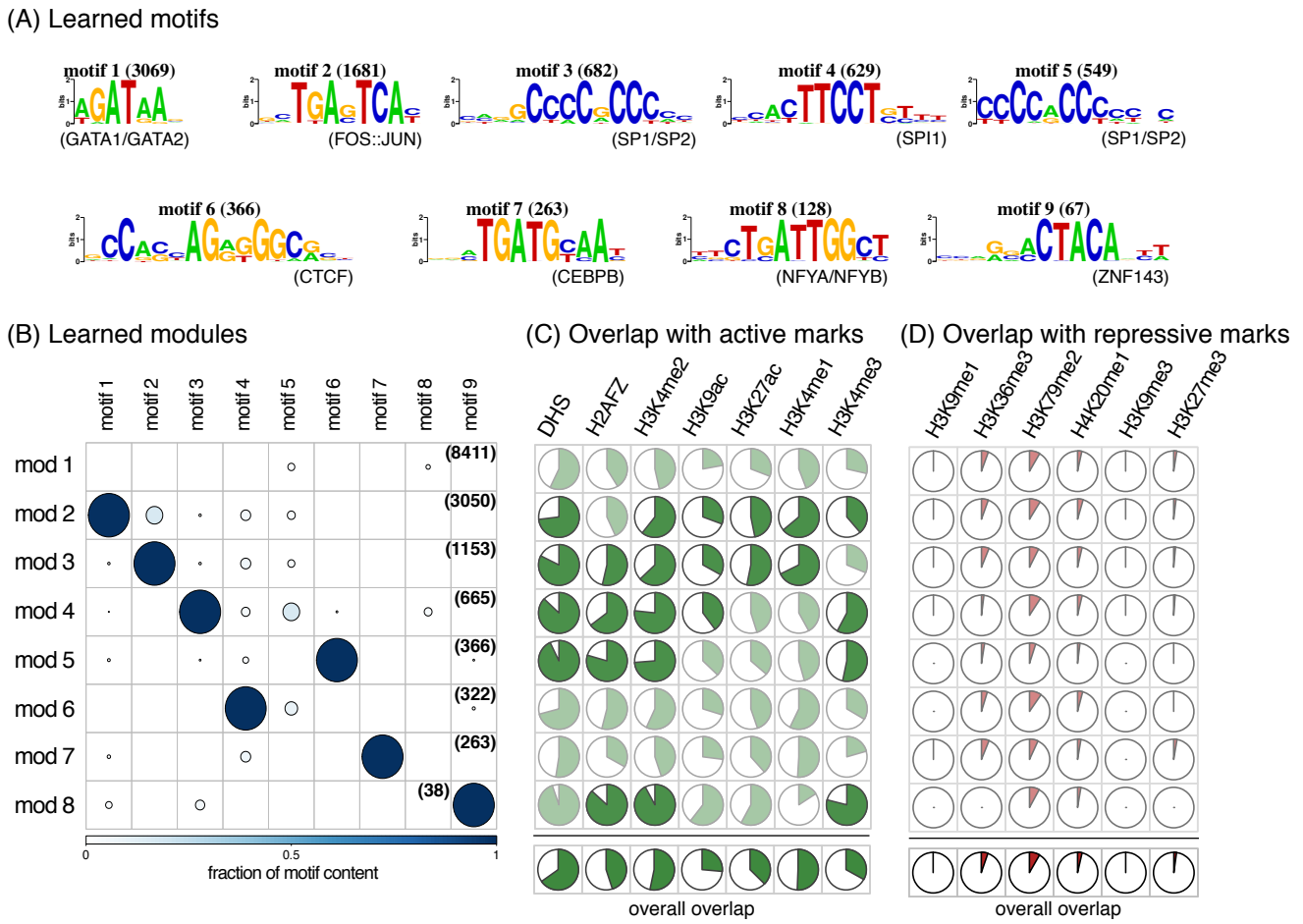
Supplemental Figure S7: Results of cisDiversity on an even larger 1000bp neighborhood of GR-bound summits. GR motif is no longer the most abundant motif. Instead, we notice low-complexity and co-factors being more frequent.



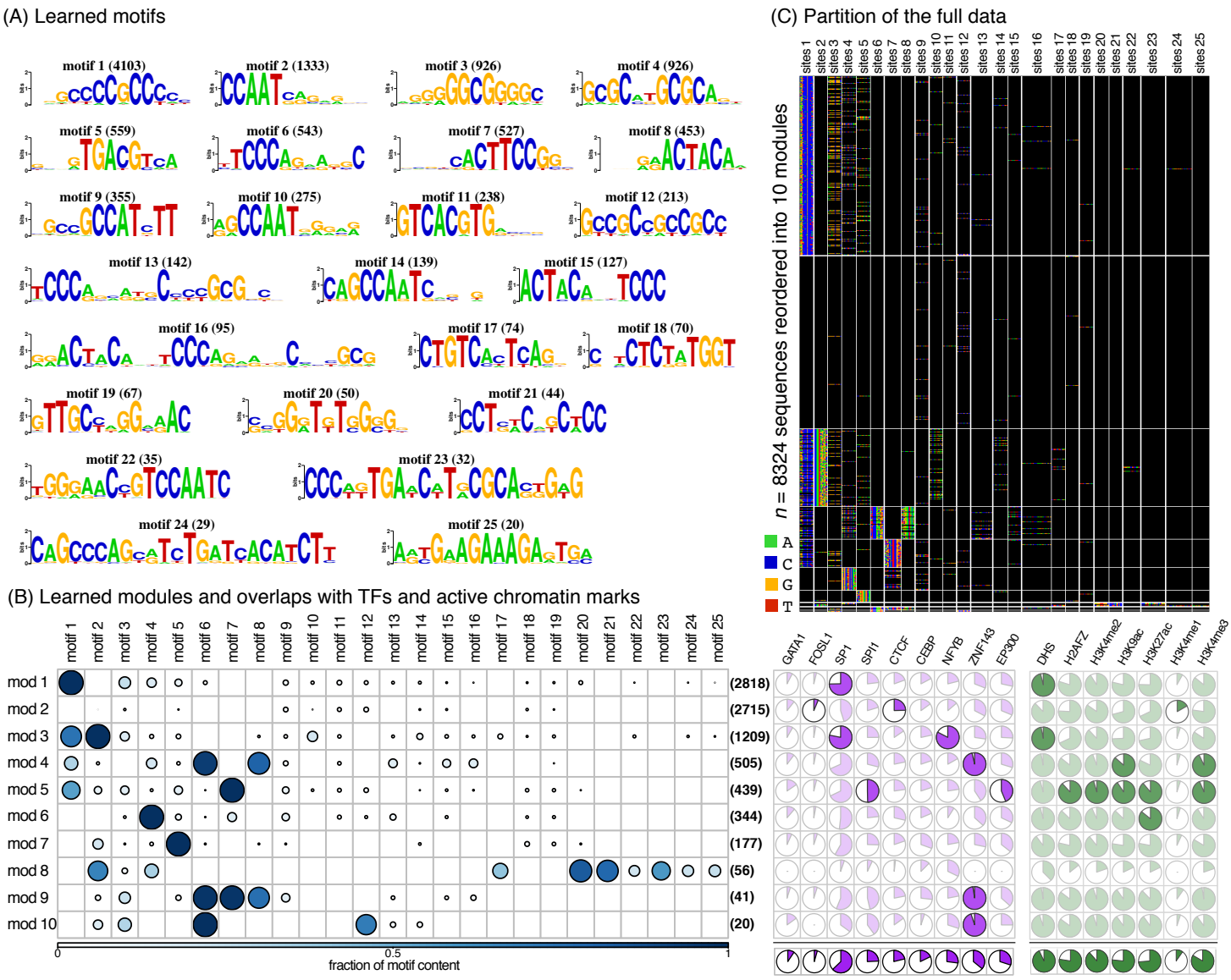
Supplemental Figure S8: Results of of CISDIVERSITY on the distal eRNA dataset. The motifs in the dotted box characterize the ninth module. All regions in this module are very similar and very likely result from segmental duplications.



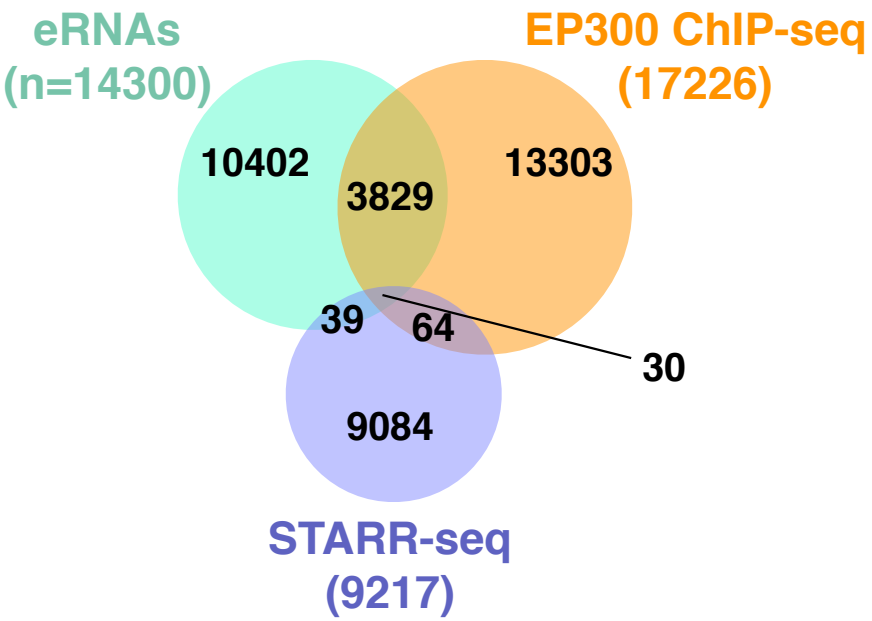
Supplemental Figure S9: Results of running CUSTAL on the regions of module 9 of the distant eRNAs.



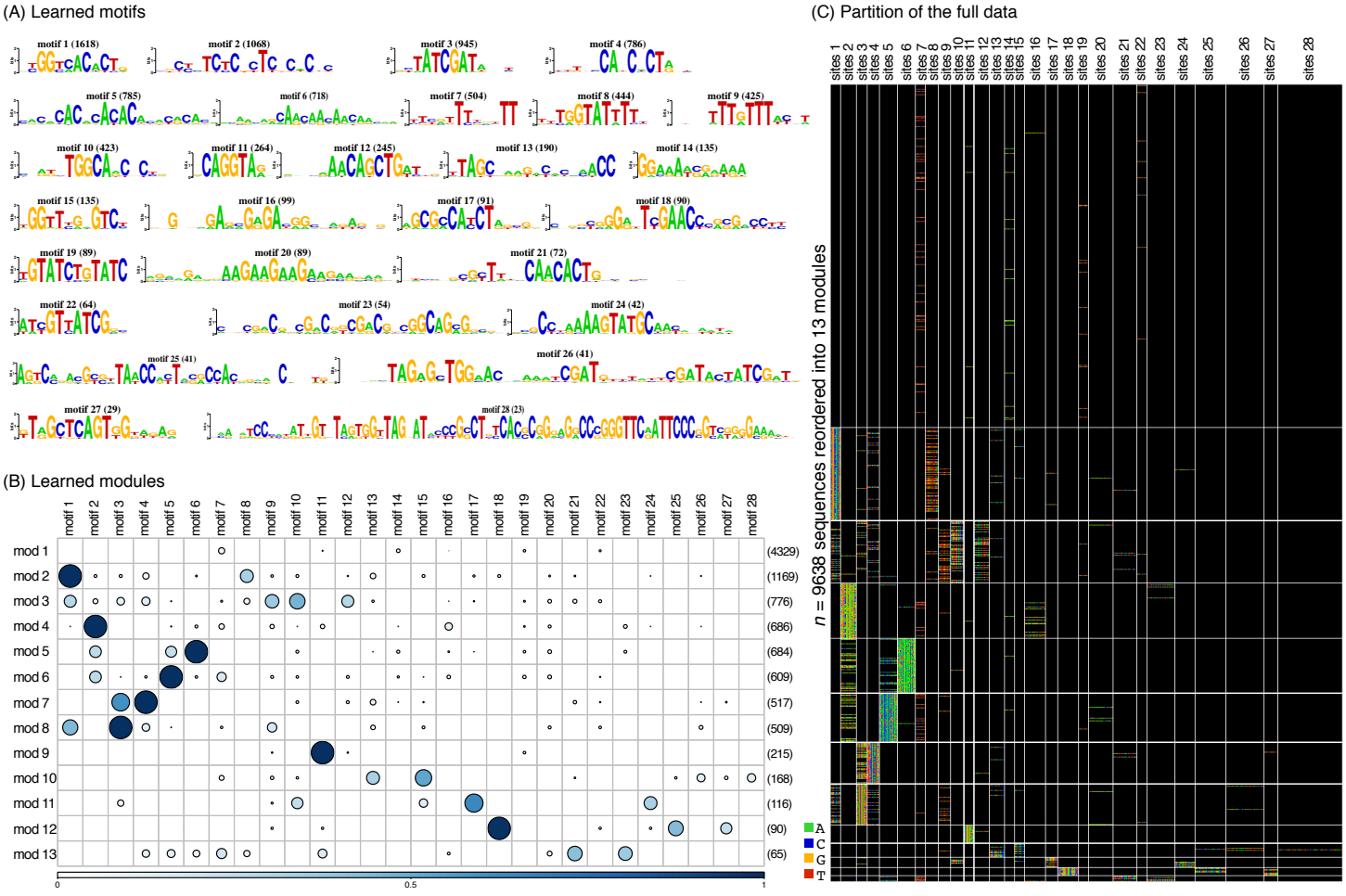
Supplemental Figure S10: Similar to Figure 6A, overlaps with active histone marks and repressive marks are shown in green and red. None of the red marks are enriched in any module, but the active marks are differentially enriched (hypergeometric $p < 10^{-4}$, shown in bold).



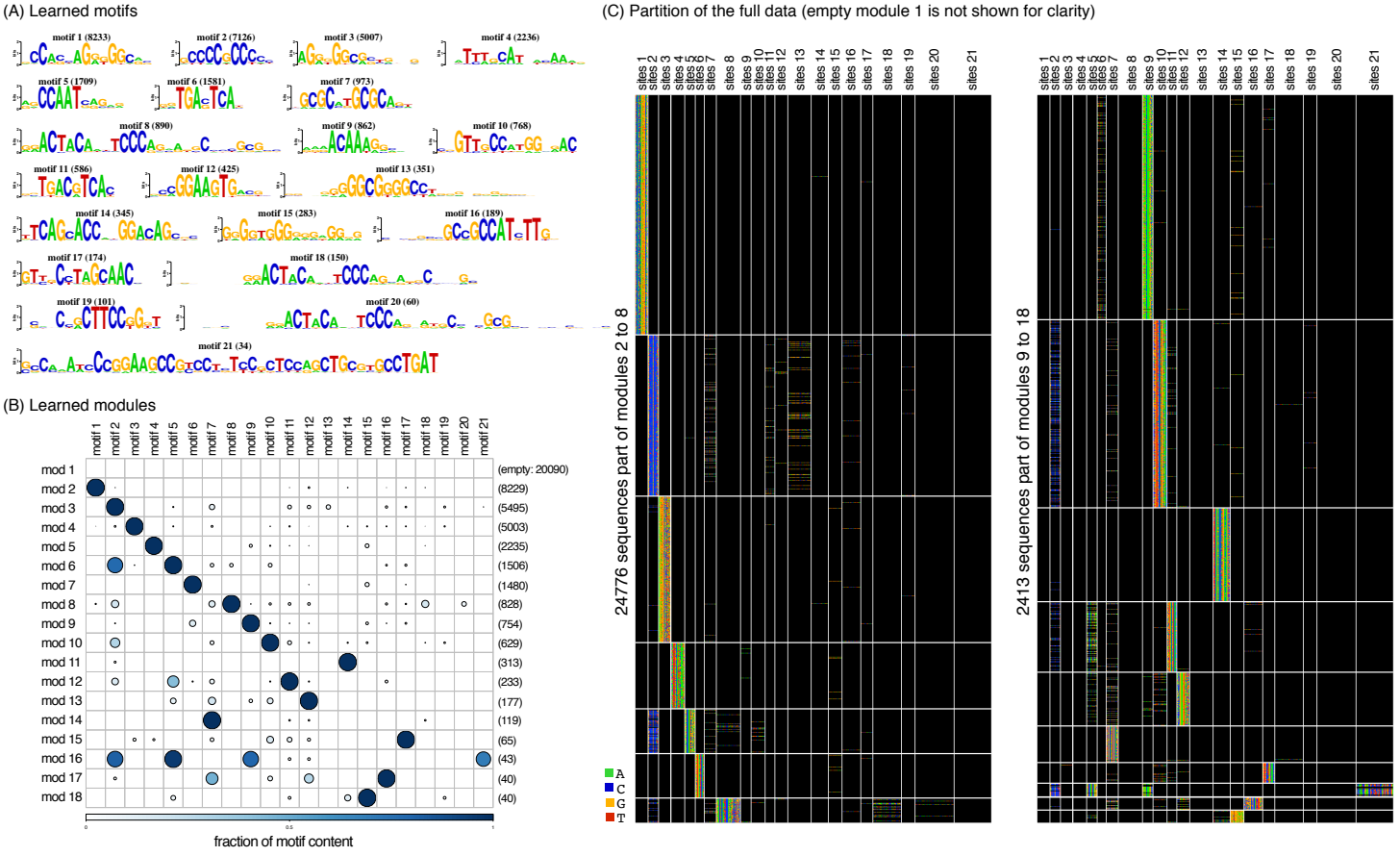
Supplemental Figure S11: Results of CISDIVERSITY on bidirectional transcripts which overlap with an annotated TSS. Also shown is the overlap with the TF binding data used in Figure 6A and active histone marks (hypergeometric $p < 10^{-4}$, shown in bold).



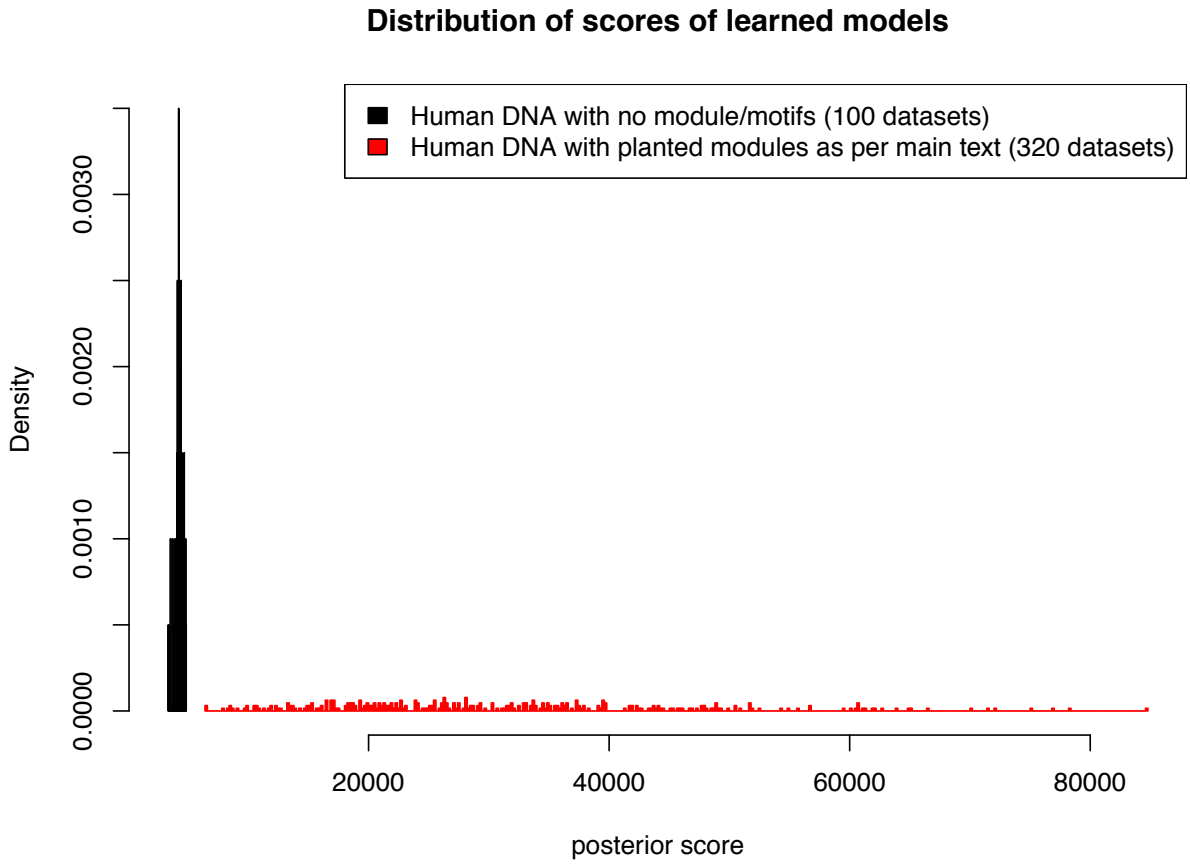
Supplemental Figure S12: Sequences that are common across three assays that measure enhancer activity in K562.



Supplemental Figure S13: Complete results of cisDiversity run on ATAC-seq regions in fly at 2–4 hrs after embryogenesis.



Supplemental Figure S14: Complete results of cisDiversity run on DNase-seq regions in human H1 ESCs.



Supplemental Figure S15: Distribution of posterior scores of models learned on the simulated datasets are shown in black (100 datasets containing human non-repetitive genomic regions with no planted motifs) and red (320 datasets with planted motifs and modules as per main text).