Cell Genomics, Volume *2*

Supplemental information

A *cis*-regulatory lexicon of DNA motif combinations mediating cell-type-specific gene regulation

Laura K.H. Donohue, Margaret G. Guo, Yang Zhao, Namyoung Jung, Rose T. Bussat, Daniel S. Kim, Poornima H. Neela, Laura N. Kellman, Omar S. Garcia, Robin M. Meyers, Russ B. Altman, and Paul A. Khavari

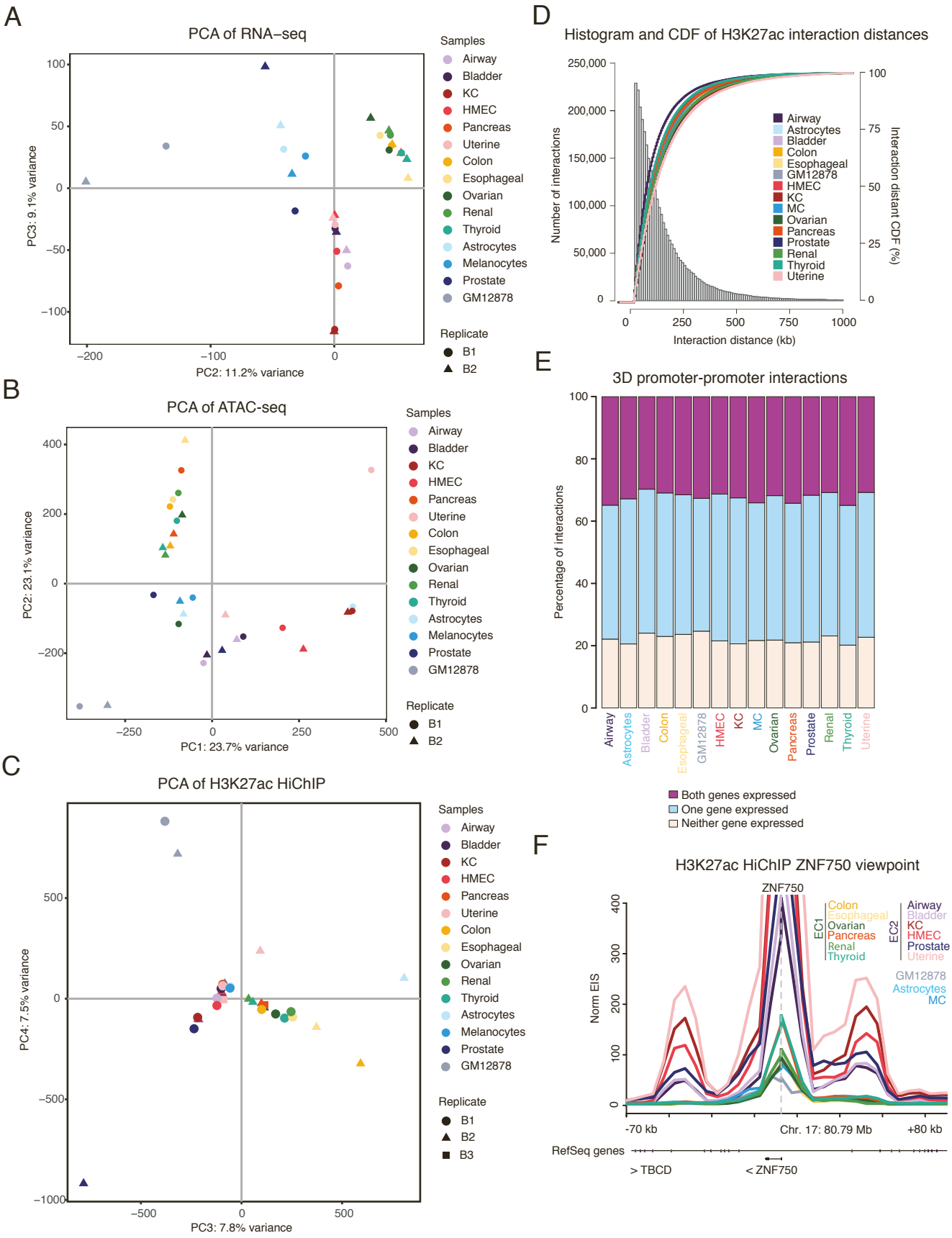


Figure S1. Additional transcriptomics and epigenomics correlation and statistics, Related to Figure 2.

(A) Scatter plot of the second and third principal components (PCs) of a PCA analysis done on RNA-seq expression data from the 15 different cell types.

(B) Scatter plot of the first and second PCs of a PCA analysis done on ATAC-seq peak data from the 15 different cell types.

(C) Scatter plot of the third and fourth PCs of a PCA analysis done on H3K27ac HiChIP loop data from the samples for the 15 different cell types.

(D) Histogram and cumulative distribution function (CDF) plot showing the distribution of interaction distance from H3K27ac HiChIP loops for the 15 different cell types.

(E) Bar plot depicting the distribution of promoter to promoter interactions found in each cell type.

(F) Virtual 4C visualization at 5 kb resolution centered at the ZNF750 TSS for all 15 cell types. > and < indicate RefSeq (NCBI Reference Sequence Database) gene orientation on the plus and minus DNA strand respectively.

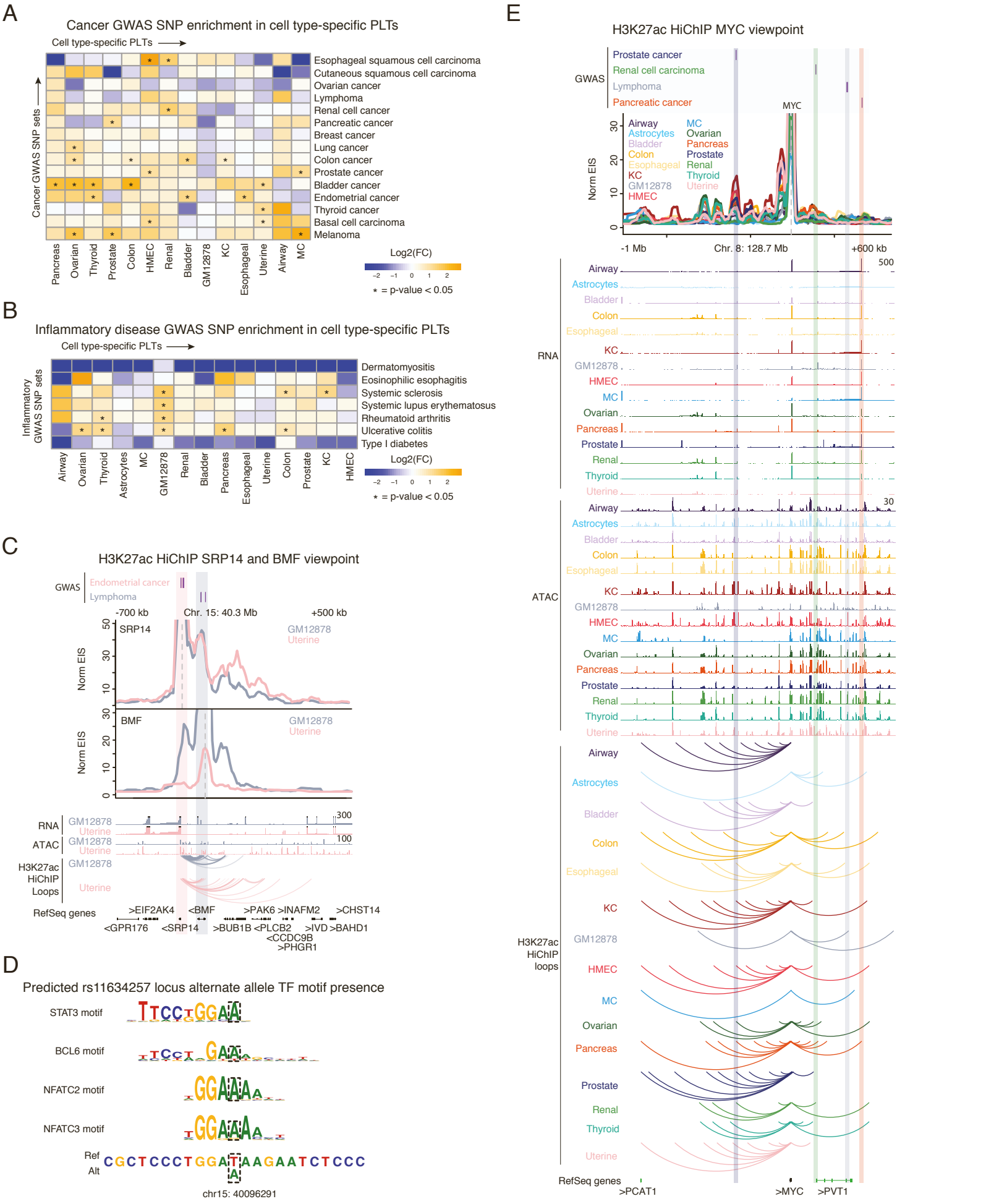
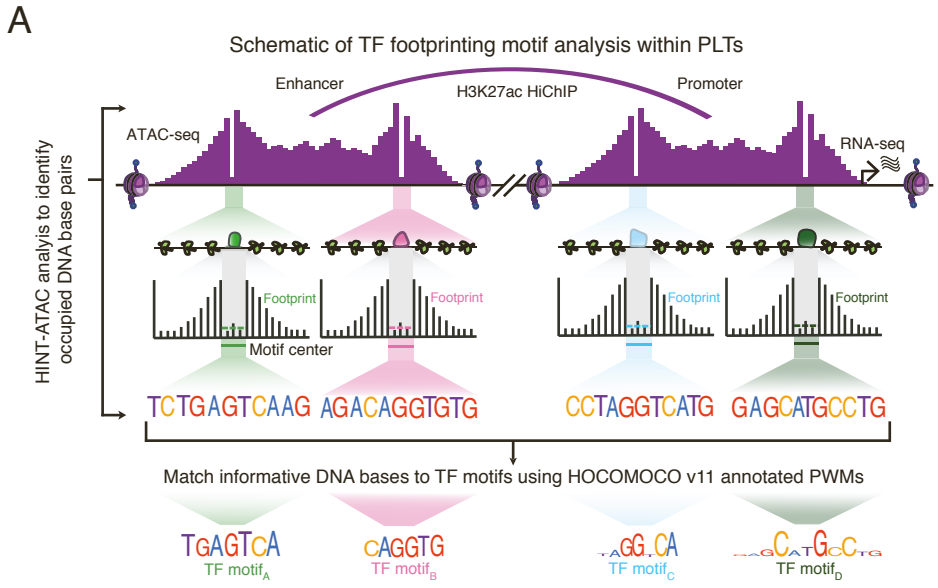
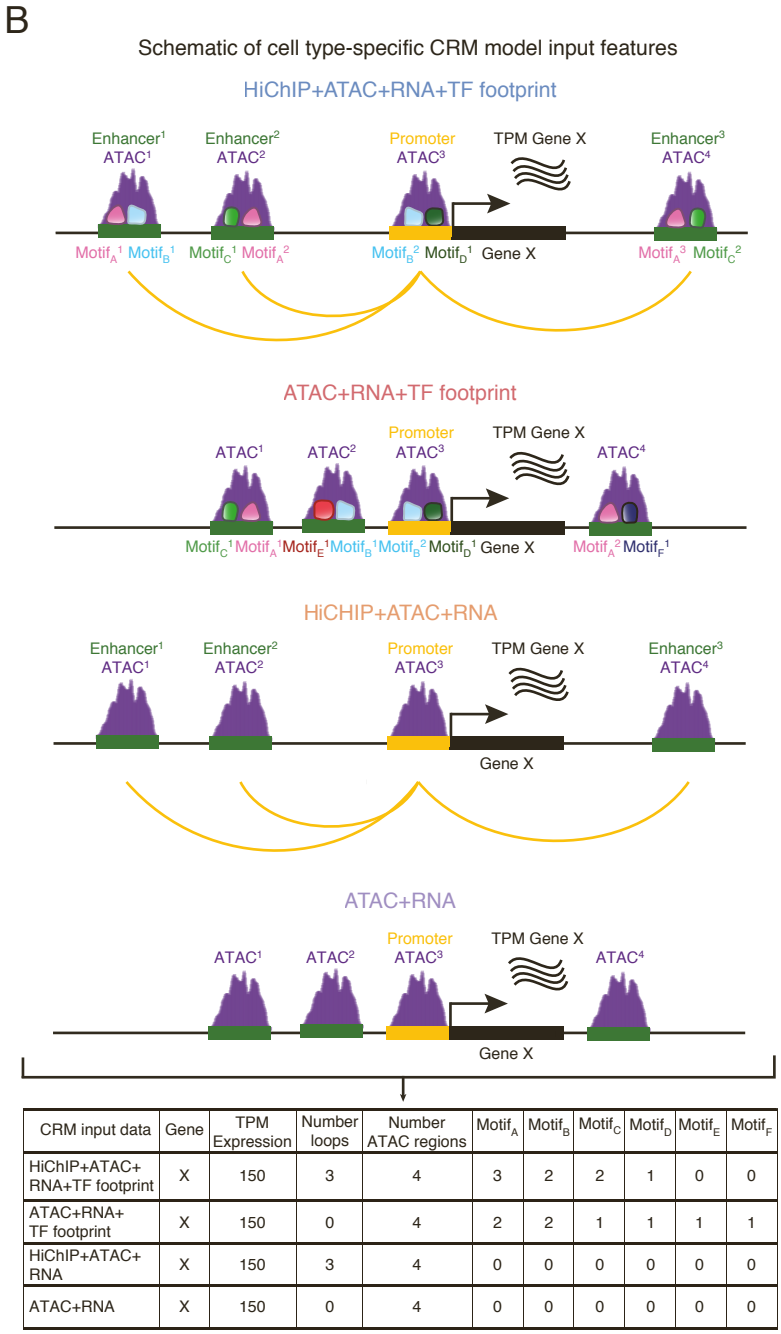


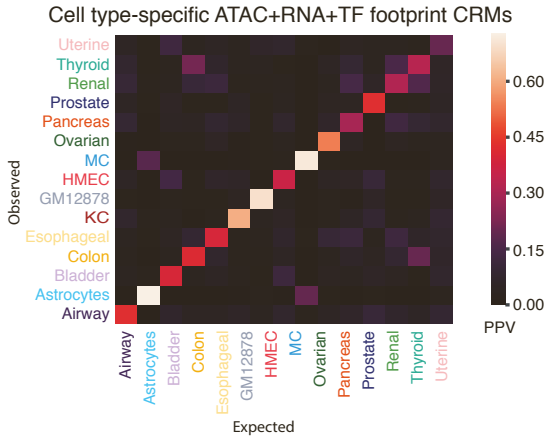
Figure S2. Human polygenic disease variant analysis in cell type-specific peak-loop-transcripts (PLTs), Related to Figure 2. (A) Heatmap depicting log2 fold-change enrichment of GWAS SNV sets associated with cancer types (rows) within cell type-specific PLTs (columns). * indicates p-value<0.05.

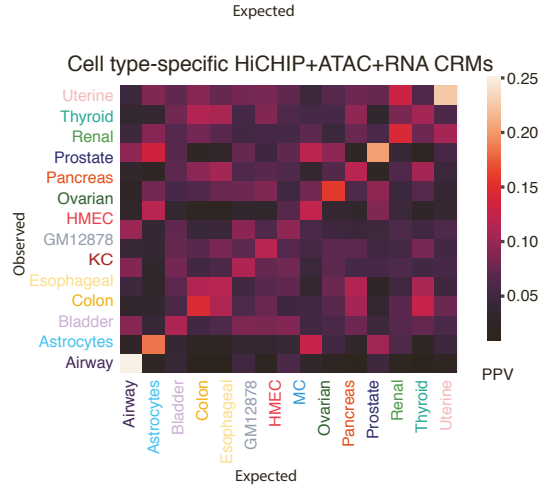
(B) Heatmap depicting log2 fold-change enrichment of GWAS SNV sets associated with inflammatory diseases (rows) within cell type-specific PLTs (columns). * indicates p-value<0.05.
(C) Top: Virtual 4C plot at 5 kb resolution and RNA, ATAC, and H3K27ac HiChIP looping tracks and endometrial cancer GWAS SNVs rs4924410, rs72731415, and rs9919974 and lymphoma GWAS SNVs rs11634257 and rs5812152, centered at the TSS of SRP14 and Bottom: BMF depicting a looping linkage to BUB1B. > and < indicate RefSeq (NCBI Reference Sequence Database) gene orientation on the plus and minus DNA strand respectively.
(D) Sequence logos of the motifs predicted to be present at the lymphoma GWAS SNV rs11634257 locus when the alternate, lymphoma-associated allele is present using motifBreakR126. Reference sequence shown is from hg19.
(E) Virtual 4C at 5 kb resolution and RNA, ATAC, and H3K27ac HiChIP looping tracks and Refseq genes at the MYC locus for all 15 cell types, depicting cell type specific looping, in particular to disease-specific SNV enhancer loci, prostate cancer GWAS SNV rs6983267, renal cancer GWAS SNV rs35252396, lymphoma GWAS SNVs rs13254990, rs13255292, and rs2720665 and pancreatic cancer GWAS SNV rs12675643. lncRNAs near the MYC locus are shown in green. > and < indicate RefSeq (NCBI Reference Sequence Database) gene orientation on the plus and minus DNA strand respectively.

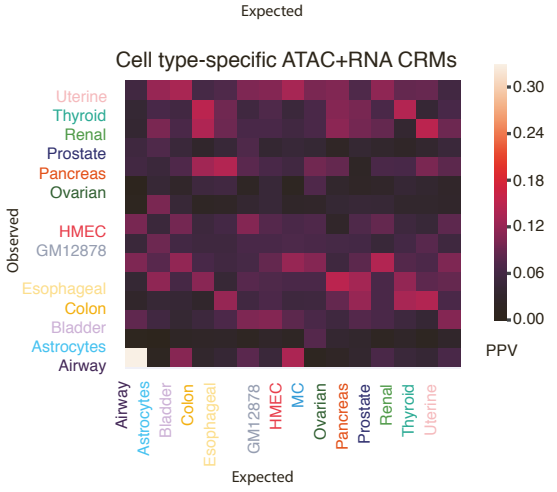




CRM input data	Gene	TPM Expression	Number loops	Number ATAC regions	Motif$_A$	Motif$_B$	Motif$_C$	Motif$_D$	Motif$_E$	Motif$_F$
HiChIP+ATAC+ RNA+TF footprint	X	150	3	4	3	2	2	1	0	0
ATAC+RNA+ TF footprint	X	150	0	4	2	2	1	1	1	1
HiChIP+ATAC+ RNA	X	150	3	4	0	0	0	0	0	0
ATAC+RNA	X	150	0	4	0	0	0	0	0	0







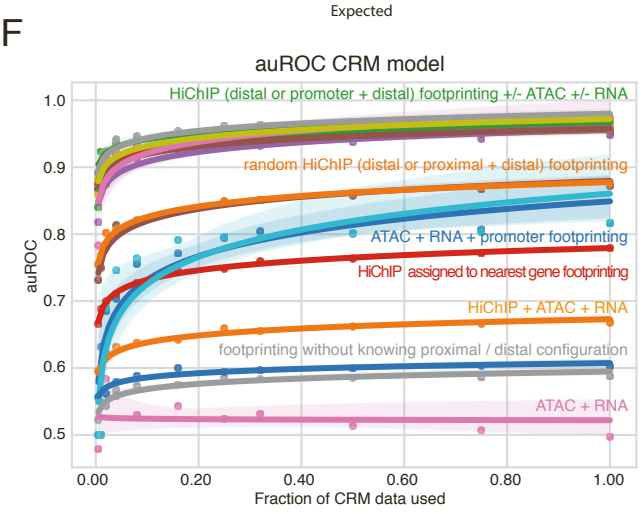


Figure S3. CRM model additional predictive metrics, Related to Figure 3.
(A) Schematic representation of TF footprinting analysis workflow.
(B) Schematic representation of gene-centric CRM data matrix input features to the iterative random-forest model.

(C-E) Confusion matrices depicting the positive predictive value (PPV) for the cell type prediction model for:
(C) ATAC + RNA.
(D) ATAC + RNA + TF footprinting.
(E) HiChIP + ATAC + RNA.
(F) Scatter plot showing auROC versus % of cis-regulatory modules learned over in our random-forest based cell type prediction model for all different types of data included as input to the model. Conditions with “random” indicate random HiChIP loops were created to loop to the promoter of interest, where the number of random loops was fixed to the number of actual loops to the given promoter. Condition with “assigned” indicates HiChIP anchor was assigned to the nearest TSS promoter.

A

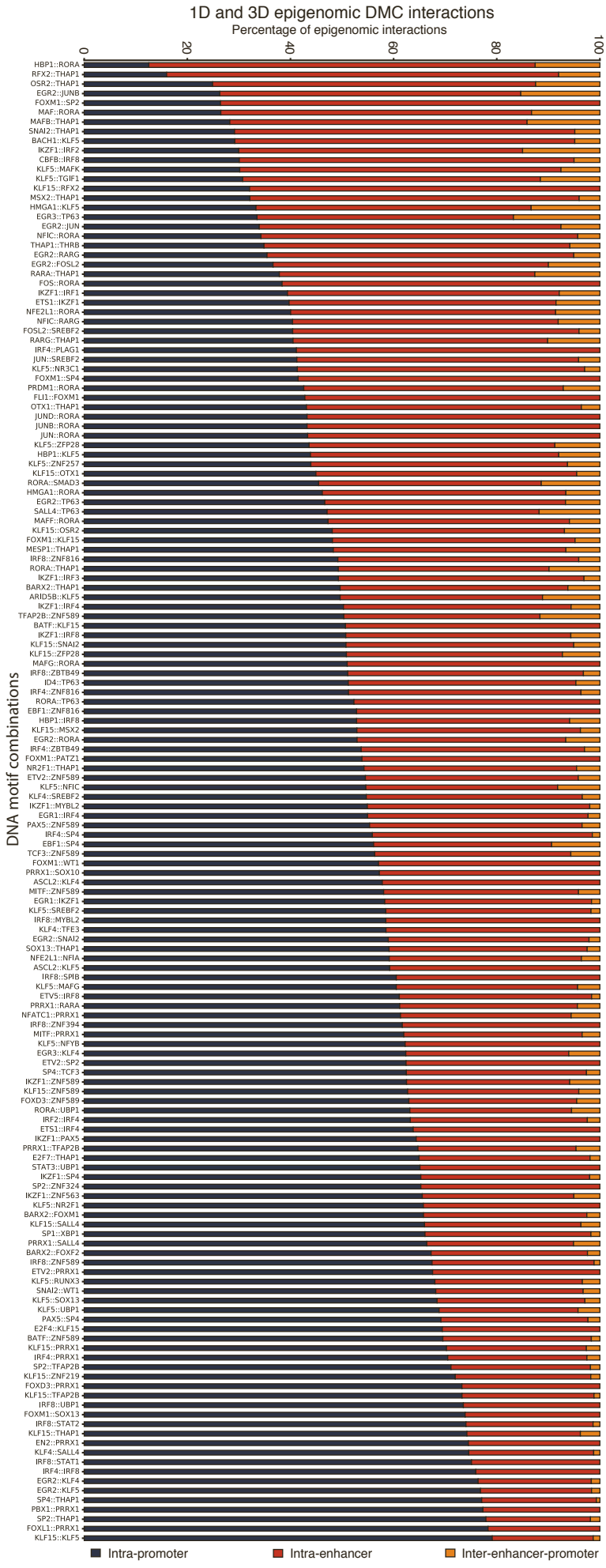


Figure S4. Regulatory DMC configurations, Related to Figure 4.

(A) Bar plot depicting the distribution of DMC configurations based on CRM epigenomic interactions for the 239 DMCs tested in MPRA.

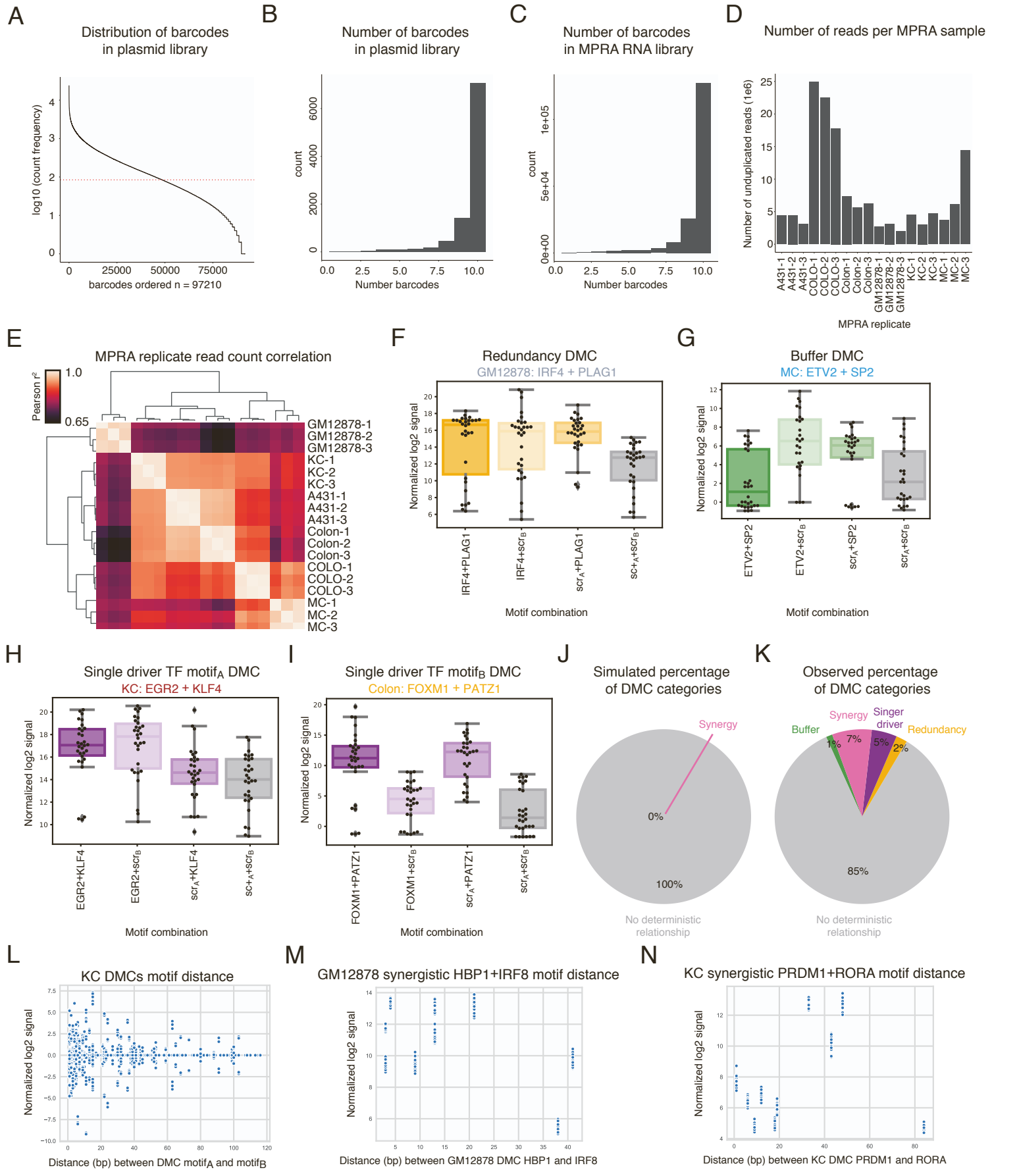


Figure S5. MPRA QC statistics, Related to Figure 5.

(A) Distribution of barcode count frequencies across the MPRA plasmid library where the mean number of barcode counts is indicated by the red dashed line at 298.5.

(B) Distribution of the number of barcodes per genomic instance of each DMC present in the plasmid library.

(C) Distribution of the number of barcodes per genomic instance of each DMC present in each tested MPRA RNA library replicate.

(D) Bar plot indicating the number of reads per replicate per cell type tested in MPRA.

(E) Heatmap depicting Pearson correlation of the different MPRA replicates to one another based on barcode counts.

(F-I) Box-and-whisker plots showing the normalized log2 MPRA signal (STAR Methods) for the different motifA-motfiB combinations. Each point on the plot represents the signal value in one genomic instance in one replicate. *s indicate p-value<0.05 (Mann-Whitney U test). These DMCs include:
(F) redundancy DMC IRF4+PLAG1 in GM12878.
(G) buffer DMC ETV2+SP2 in MC.
(H) single driver motifA DMC EGR2+KLF4 in KC.
(I) single driver motifB FOXM1+PATZ1 in colon.
(J-K) Pie chart of percent of each DMC category identified by:
(J) running simulations of MPRA values randomly assigned to a DMC configuration.
(K) observed matched MPRA values and DMC configurations.
(L-N) Scatterplot of distance between DMC motifA and motifB (x-axis) and normalized MPRA log2 signal (y-axis) in:
(L) all KC DMCs tested in KC.
(M) HBP1+IRF8 tested in GM12878s.
(N) PRDM1+RORA tested in KCs.

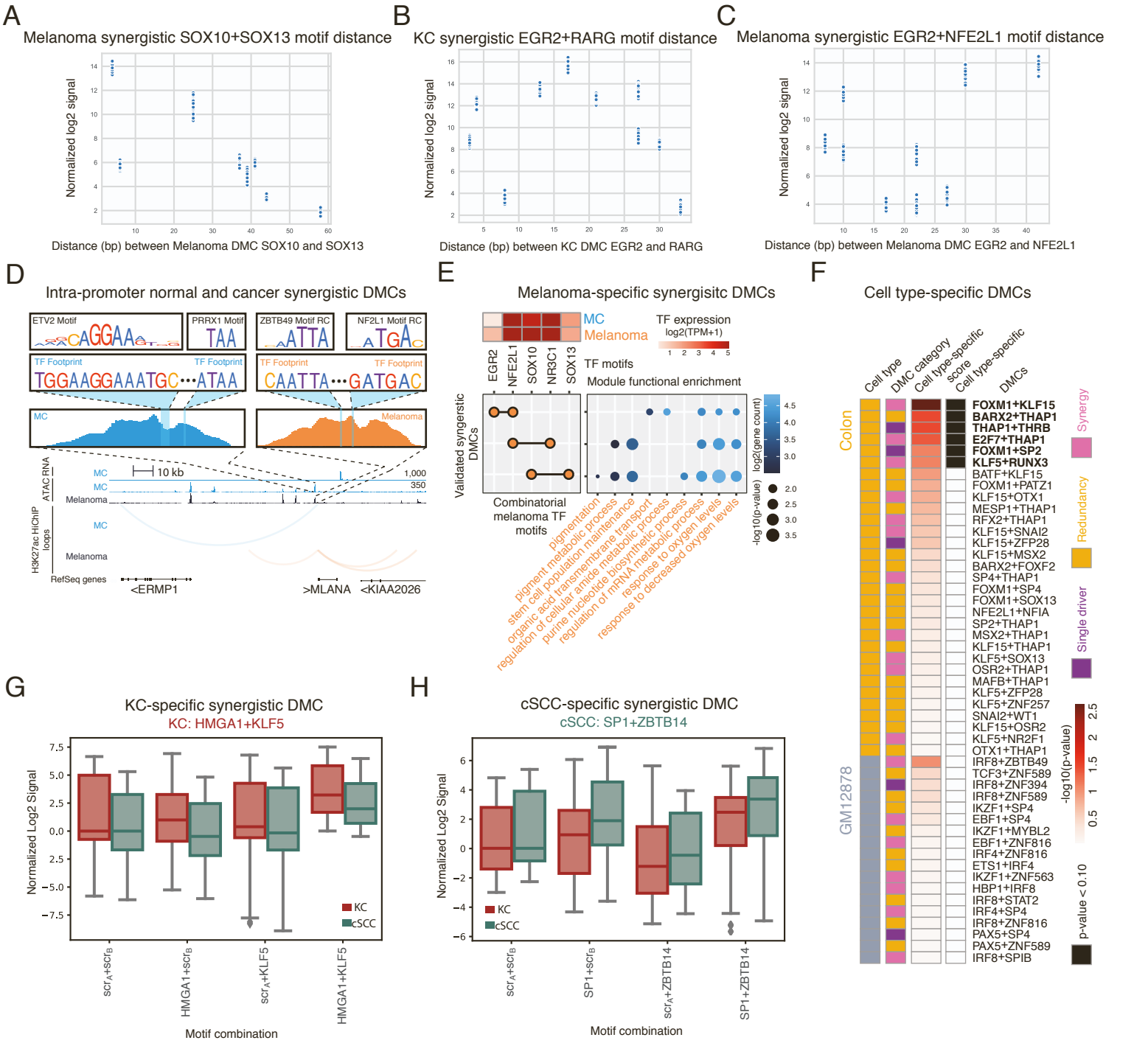


Figure S6. Cell state-specific DMCs, Related to Figure 6.

(A-C) Scatterplot of distance between DMC motifA and motifB (x-axis) and normalized MPRA log2 signal (y-axis) in:

(A) SOX10+SOX13 tested in MM COLO829 cells.

(B) EGR2+RARG tested in KCs.

(C) EGR2+NFE2L1 tested in MM COLO829 cells.

(D) Genomic instance of intra-promoter MC and MM DMCs including motif footprinting PWM, footprint sequence, and surrounding ATAC peak profile, and RNA, ATAC, and HiChIP tracks centered around gene MLANA. (RC= Reverse Complement).

(E) Top left: heatmap shows log2(TPM+1) values for TFs (columns) involved in functional MM-specific synergistic DMCs (Wilcoxon rank-sum test p<0.10). Left: combinatorial TFs of the DMC (rows). Motifs (columns) that make up the DMC are circles with a black line connecting them. Right: dot plot shows the GO biological processes that are enriched for target genes (x-axis) that utilize the DMC (y-axis). Dots are colored by log2(target gene count). Dot sizes are the -log10(p-value) of the GO enrichment.

(F) Left to right: panel colored by cell type/state; panel colored by functional DMC category; heatmap panel of -log10(p-value) cell type-/state-specificity score (STAR Methods); panel colored by cell-type- specific expression (Wilcoxon rank-sum test p-value<0.10).

(G) Box-and-whisker plot of the log2 MPRA signal, normalized to the double scramble condition, in the different combinations of motifA-motifB DMC scrambling for synergistic KC-specific DMC HMGA1+KLF5.

(H) Box-and-whisker plot of the log2 MPRA signal, normalized to the double scramble condition, in the different combinations of motifA-motifB DMC scrambling for synergistic cSCC-specific DMC SP1+ZBTB14.