Cell Genomics, Volume *2*

Supplemental information

Functional inference of gene regulation

using single-cell multi-omics

Vinay K. Kartha, Fabiana M. Duarte, Yan Hu, Sai Ma, Jennifer G. Chew, Caleb A. Lareau, Andrew Earl, Zach D. Burkett, Andrew S. Kohlway, Ronald Lebofsky, and Jason D. Buenrostro

Supplemental Information

Supplemental Figures

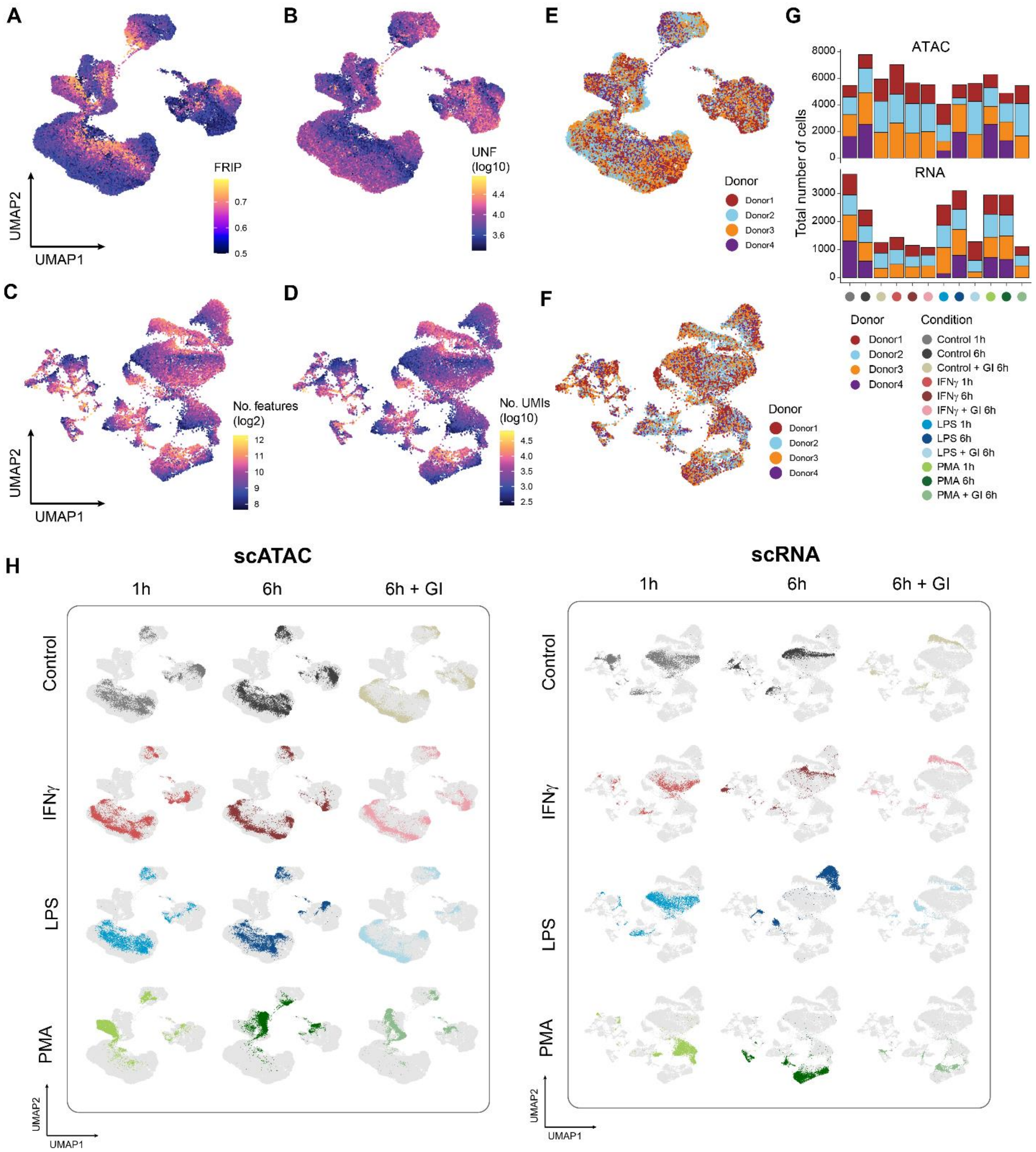

Figure S1. Quality metrics associated with scATAC-seq and scRNA-seq profiling of resting and stimulated PBMCs (related to Figure 1). **A-B.** UMAP projection of scATAC-seq cells colored by fraction of reads in peaks (FRIP) (A) or total number of unique nuclear Tn5 insertion fragments (B). **C-D.** UMAP projection of scRNA-seq cells colored by total number of detected features (C) or total number of unique molecular identifiers (UMIs) per feature (D). **E-F.** UMAP projection of scATAC-seq (E) and scRNA-seq (F) stimulation data colored by Donor. **G.** Number of cells passing quality filtering for scATAC-seq and scRNA-seq stimulation data per donor per condition. **H.** UMAP of scATAC-seq (left) and scRNA-seq (right) cells profiled, with cells for each condition highlighted on the background of all cells.

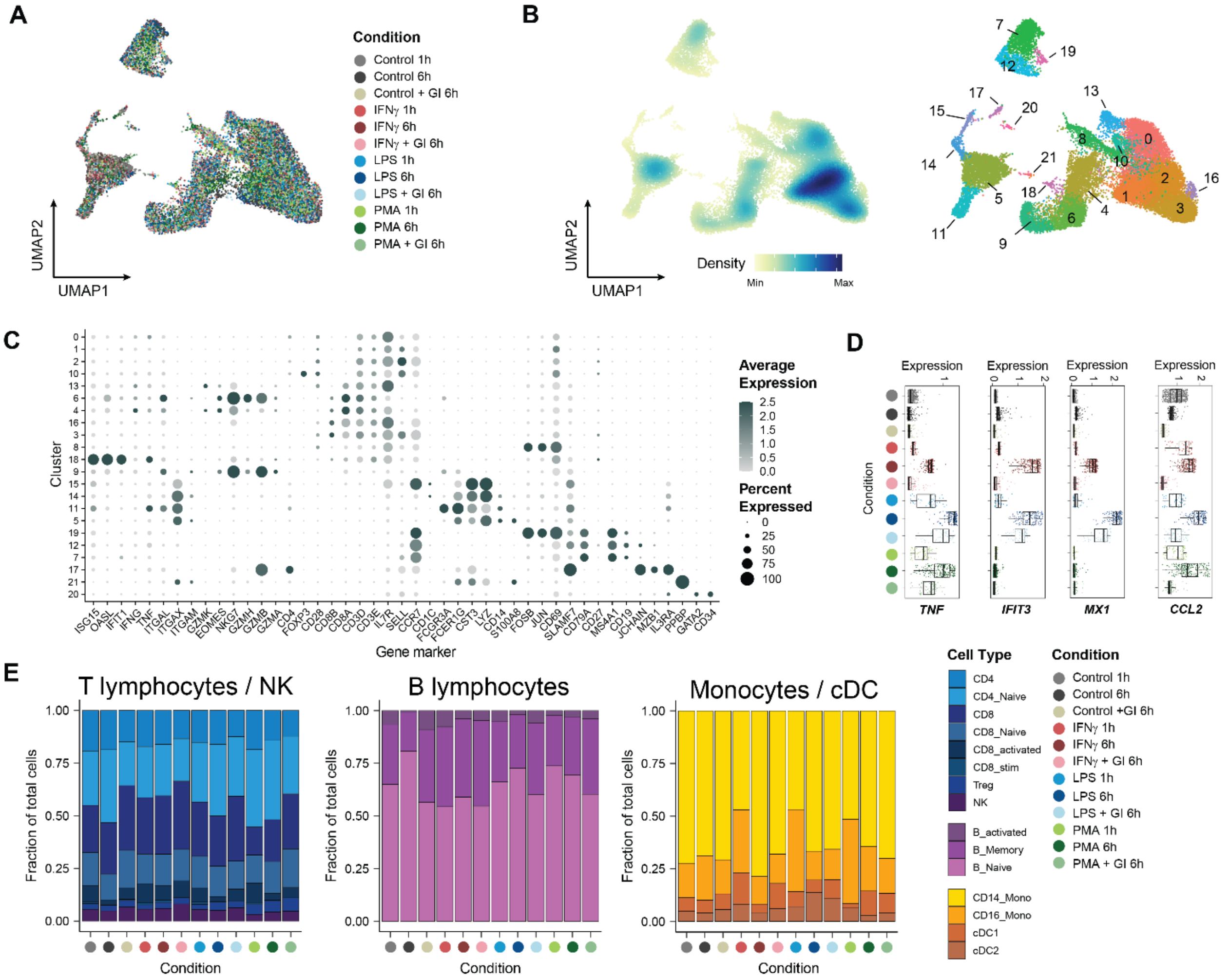

Figure S2. Alignment, cell clustering and annotation of scRNA-seq data across stimulation conditions (related to Figure 1). **A.** UMAP of scRNA-seq cells (aligned) after adjusting for treatment condition **B.** UMAP of aligned scRNA-seq cells in A, colored by density (left) or cell cluster based on Leiden clustering (right). **C.** Dotplot of gene expression markers highlighting cluster specific expression (used for scRNA-seq cell annotation) **D.** Smoothed RNA expression distribution in CD14 Monocytes across conditions for specific stimulus response genes **E.** Fraction of total scRNA-seq CD4/CD8/NK cells (left), B lymphocytes (middle) and Monocytes / cDCs (right) grouped per condition and cell type annotation.

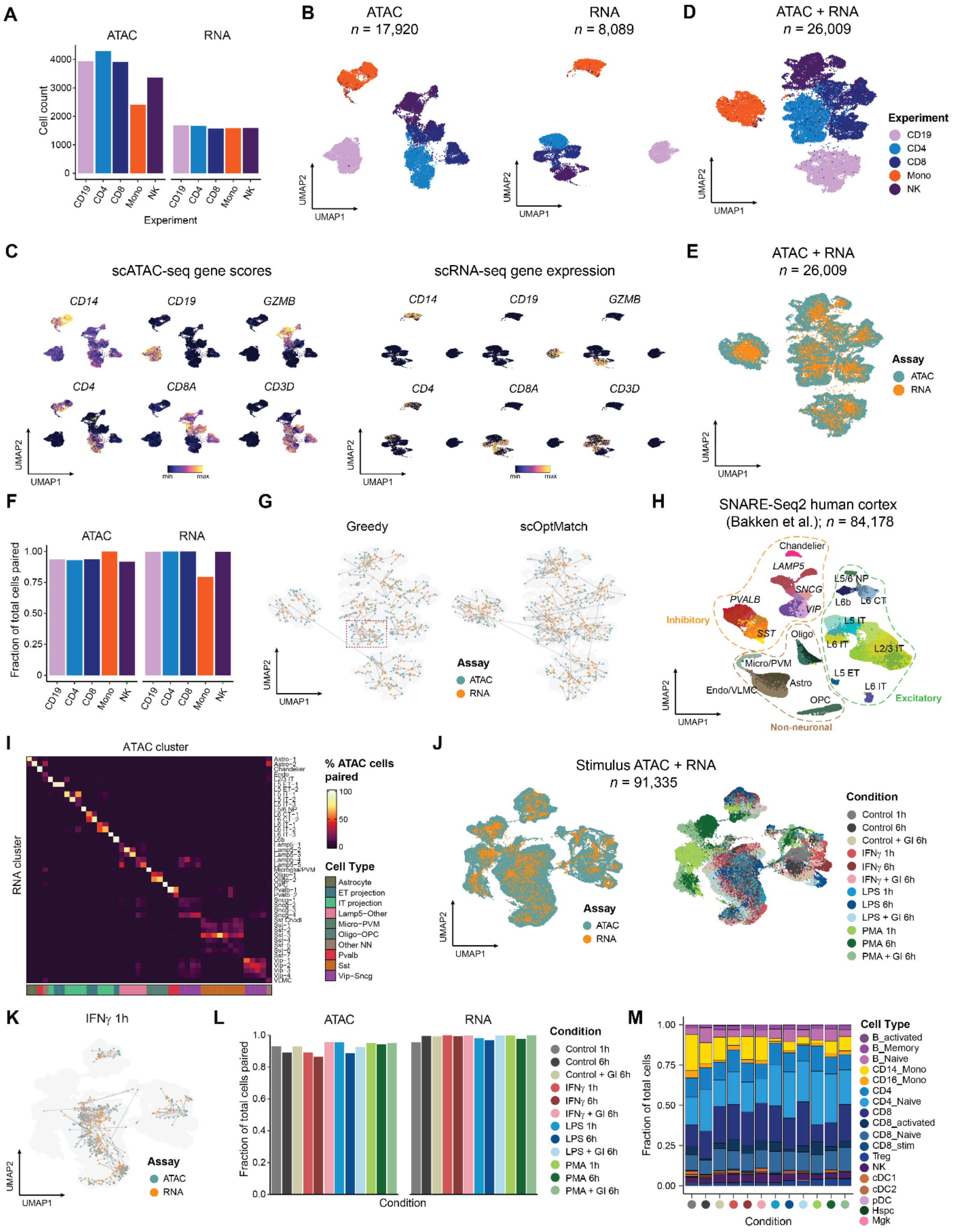

Figure S3. Applications of scOptMatch to pair scATAC-seq and scRNA-seq cells from enriched PBMC subtypes, SNARE-Seq2 data from human cortex, and extension to stimulated scATAC-seq and scRNA-seq PBMC data (related to Figure 2). **A.** Total number of cells assayed and

passing QC for scATAC-seq and scRNA-seq from bead enriched cells. **B.** UMAP plots of bead enriched PBMCs showing single cell projections for scATAC (left) or scRNA cells (right), based on peak accessibility or gene expression, respectively. **C.** UMAPs of scATAC or scRNA cells (from B) colored by smoothed gene activity scores or RNA gene expression of cell type marker genes, respectively **D.** UMAP of both scATAC and scRNA cells based on CCA co-embedding using union of top variable scATAC gene scores and top variable scRNA gene expression, with cells colored by enriched sub-population **E.** Same as in D, with cells colored by assay. **F.** Fraction of total scATAC and scRNA-seq cells paired using our OptMatch approach per isolate cell type assayed. **G.** CCA-based UMAP of cells from D, highlighting computational pairing (300 pairs shown at random) between scATAC-seq and scRNA-seq cells using a greedy approach (scRNA cell with maximum Pearson r for each scATAC cell; left) versus our scOptMatch constrained pairing method (right). Red box highlights multiple RNA cells mapping to the same ATAC cell. **H.** UMAP projection of single cells (*n*=84,178) for SNARE-Seq2 human primary cortex data (Bakken et al.)[1], with cells colored by previously established cell type annotations (*n*=43 clusters). Labels indicate predetermined cell type annotations or gene expression markers (italicized). **I.** Heatmap highlighting scOptMatch pairing accuracy when applied to SNARE-Seq2 human brain multi-modal data (*n*=84,178 cells; Fig S3H). Percentages indicate percent cells in each ATAC cluster paired with the corresponding RNA cell cluster, using previously annotated clusters to group cells by. Color bar indicates broader cell cluster groupings of neuronal and non-neuronal cell types, as described in Bakken et al.[1] **J.** UMAP clustering based on CCA co-embedding of stimulation data cells colored by assay (left) or by stimulus condition (right). **K.** Computational pairing (300 pairs shown at random) between scATAC-seq and scRNA-seq cells for the IFNɣ 1h stimulation condition. **L.** Fraction of total scATAC and scRNA-seq cells paired using our OptMatch approach per stimulus condition assayed. **M.** Distribution of scATAC cells (*n*=62,219) based on paired annotation obtained from pairing to scRNA-seq cells, per stimulus condition. CT: Corticothalamic cells; ET: Extratelencephalic cells; IT: Intratelecenphalic cells; NP: Near-projecting; OPC: Oligodendrocyte precursors; Oligo: Oligodendrocyte; Micro: Microglia; PVM: Perivascular macrophages; Astro: Astrocytes; Edo: Endothelial cells; VLMC: vascular and leptomeningeal cells.

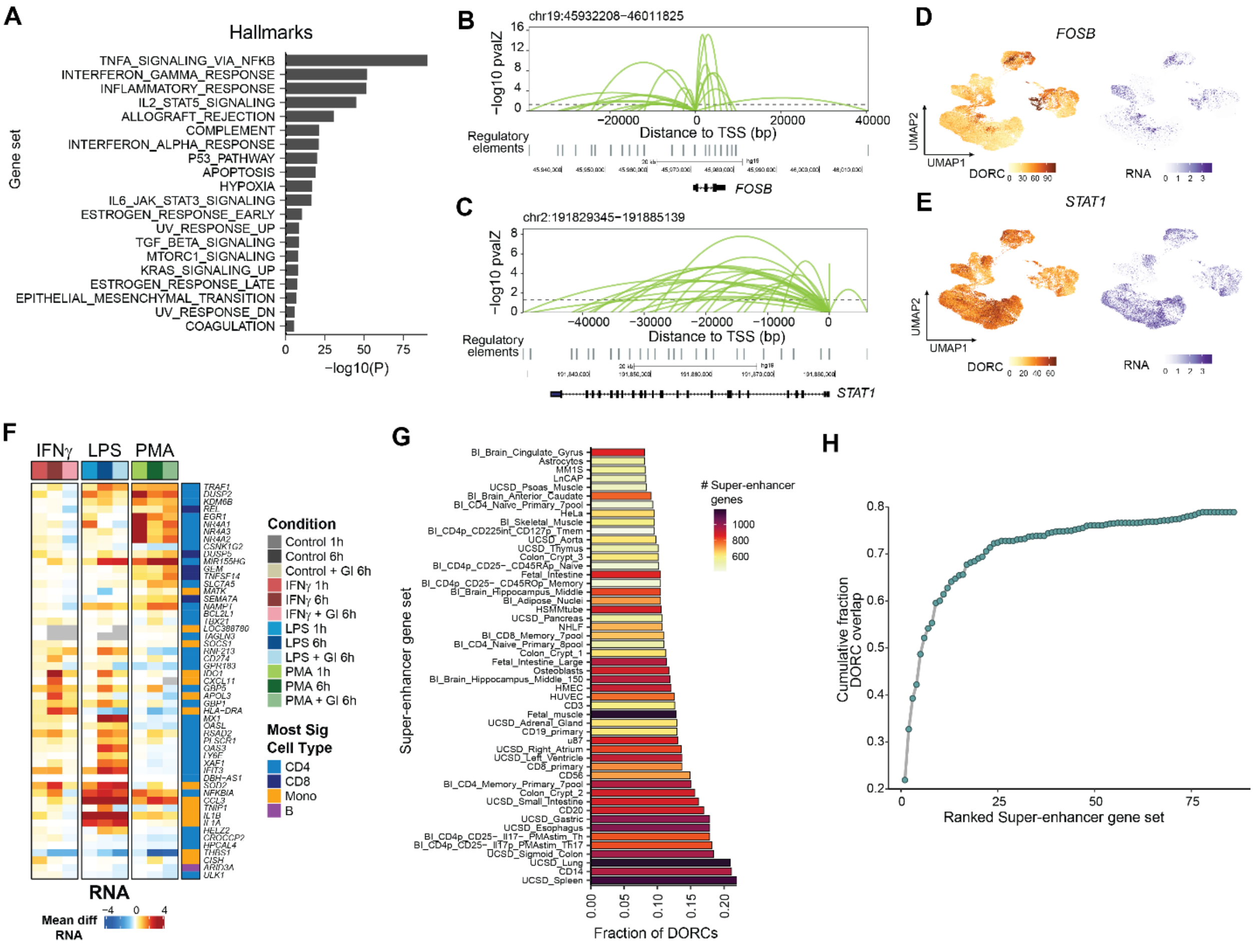

Fig S4. Peak-gene associations determine domains of regulatory chromatin associated with stimulus response (related to Figure 3). **A.** Gene set hyper-enrichment testing among DORCs for Hallmark gene sets (*n*=50). **B-C.** Loop plots highlighting significant gene-peak associations for *FOSB* (B) *and STAT1* (C). **D-E**. UMAPs of paired cells (*n*=62,219) highlighting scATAC DORC scores (left) or scRNA expression (right) for *FOSB* (D) and *STAT1* (E). **F.** Heatmap of the mean difference in single cell RNA expression for the union of the top 10 differential DORCs across conditions and cell types (*n*=53 genes; see Fig 3G). Cell type color bar represents the cell group having the most significant change across all conditions, for that assay. Gray indicates undetected RNA for that condition. **G.** Fraction of DORC genes (*n*=1,128) overlapping genes previously linked to super-enhancer regions under different cellular contexts (*n*=86). Only the top 50 gene sets are shown. **H.** Cumulative fraction of super-enhancer associated genes overlapping DORC genes with the addition of each cellular context

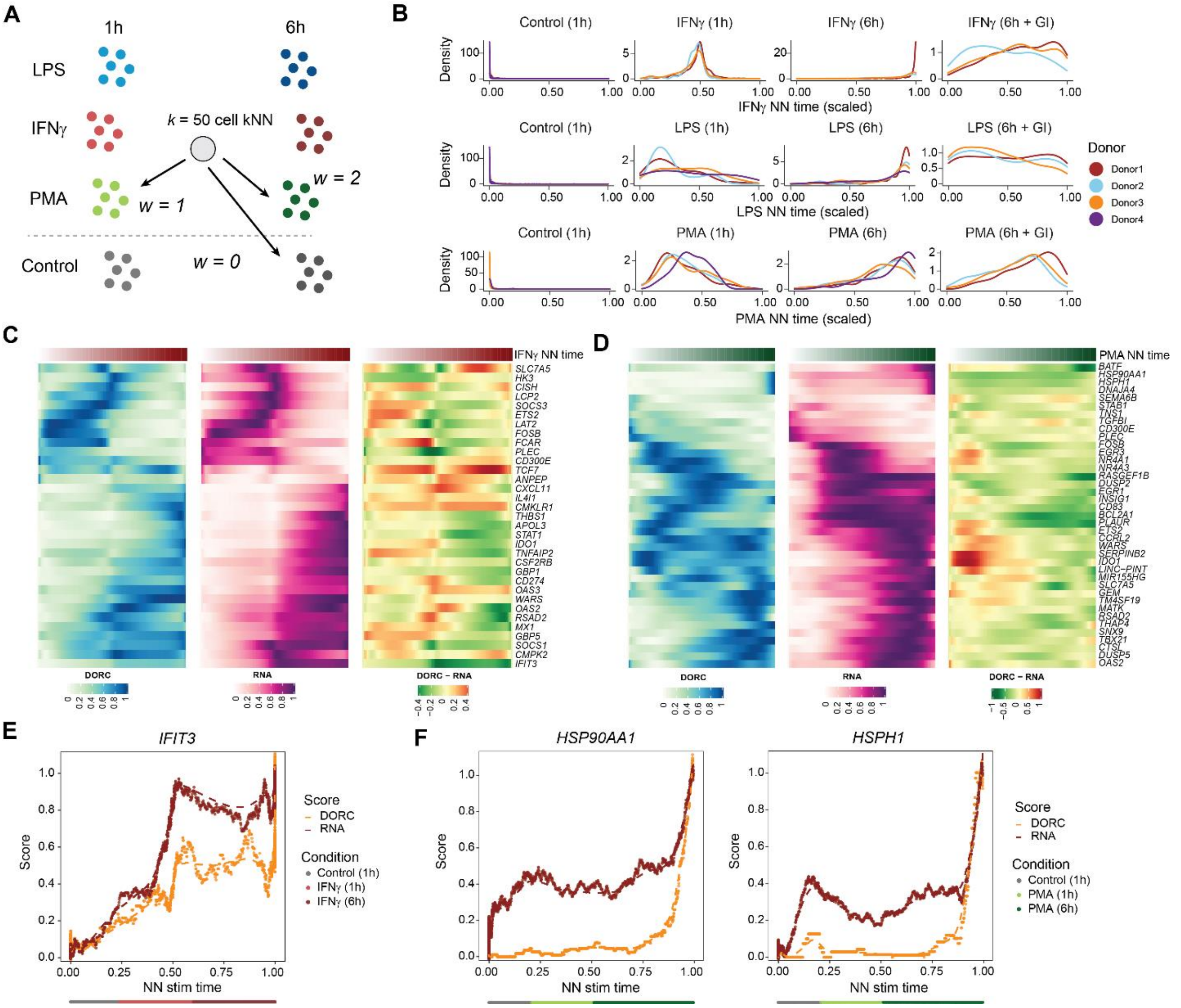

Figure S5. Stimulation nearest-neighbor (NN) time determination and DORC features associated with IFNɣ and PMA NN time in monocytes (related to Figure 4). **A.** Schematic of stimulation nearest neighbor (NN) stimulation time estimation using the weighted average of cell k-nearest neighbors (kNN), per condition. **B.** Cell density distributions of NN stimulation time shown for scATAC monocytes. **C.** Heatmaps highlighting smoothed normalized DORC accessibility, RNA expression and residual (DORC - RNA) levels for DORC genes with respect to IFNɣ NN stimulation time for Control 1h and IFNɣ 1h/6h monocyte cells (related to Fig 4C). **D.** Same as in C., but for PMA NN stimulation time-associated DORCs in monocytes. **E.** Same as in Fig 4D, but for IFNɣ NN stimulation time in monocytes. **F.** Same as in E., but for heat shock protein encoding genes *HSP90AA1* and *HSPH1* with respect to PMA NN stim time.

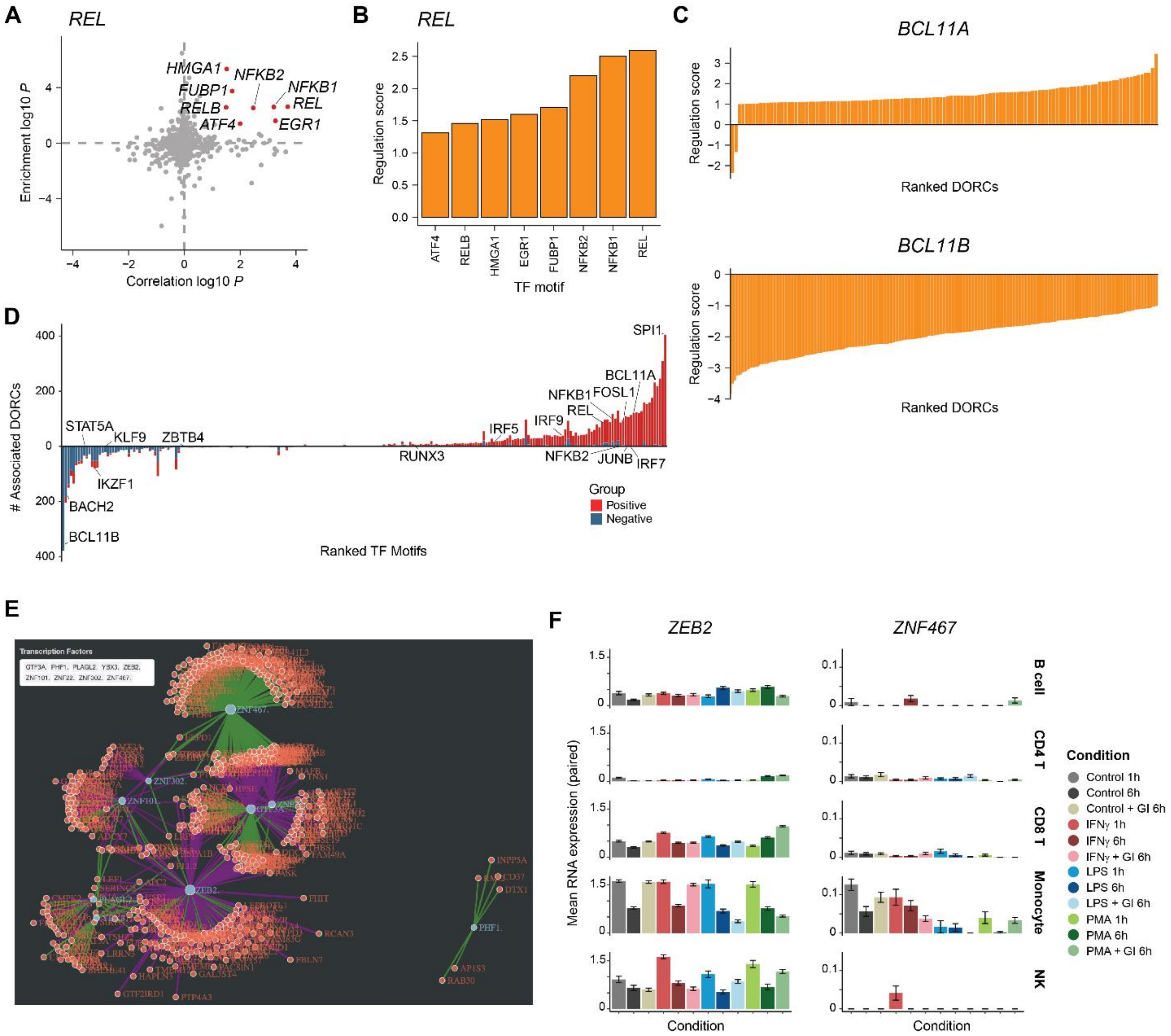

Figure S6. FigR GRN application to PBMC stimulation data (related to Figure 5). **A.** Scatter plot of TF motif enrichment among DORC KNN peaks versus TF RNA expression correlation with DORC accessibility for DORC *REL*. Red points indicate candidate drivers of *REL* (absolute(regulation score) ≥ 1). **B.** Regulation scores for candidate drivers of *REL* shown in A. **C.** Regulation score distributions for all putative DORCs (absolute(regulation score) ≥ 1) driven by *BCL11A* (top) *or BCL11B* (bottom). **D.** Barplot showing total number of positively (red) vs negatively (blue) associated DORCs for each TF (absolute(regulation score) ≥ 1) highlighting select TF activators and repressors also highlighted in 5E. **E.** FigR GRN visualization of TFs and their associated DORCs (absolute(regulation score) ≥ 1), highlighting select TFs (blue points) detected as having activating (green lines) or repressive (purple lines) associations with DORCs (red points) discovered through incorporating a newly curated motif database with FigR. Visualized using https://buenrostrolab.shinyapps.io/stimFigR/. **F.** Bar plot of paired single cell RNA expression levels for TFs *ZEB2* and *ZNF467* shown per condition per cell type annotation (mean +/- standard error).

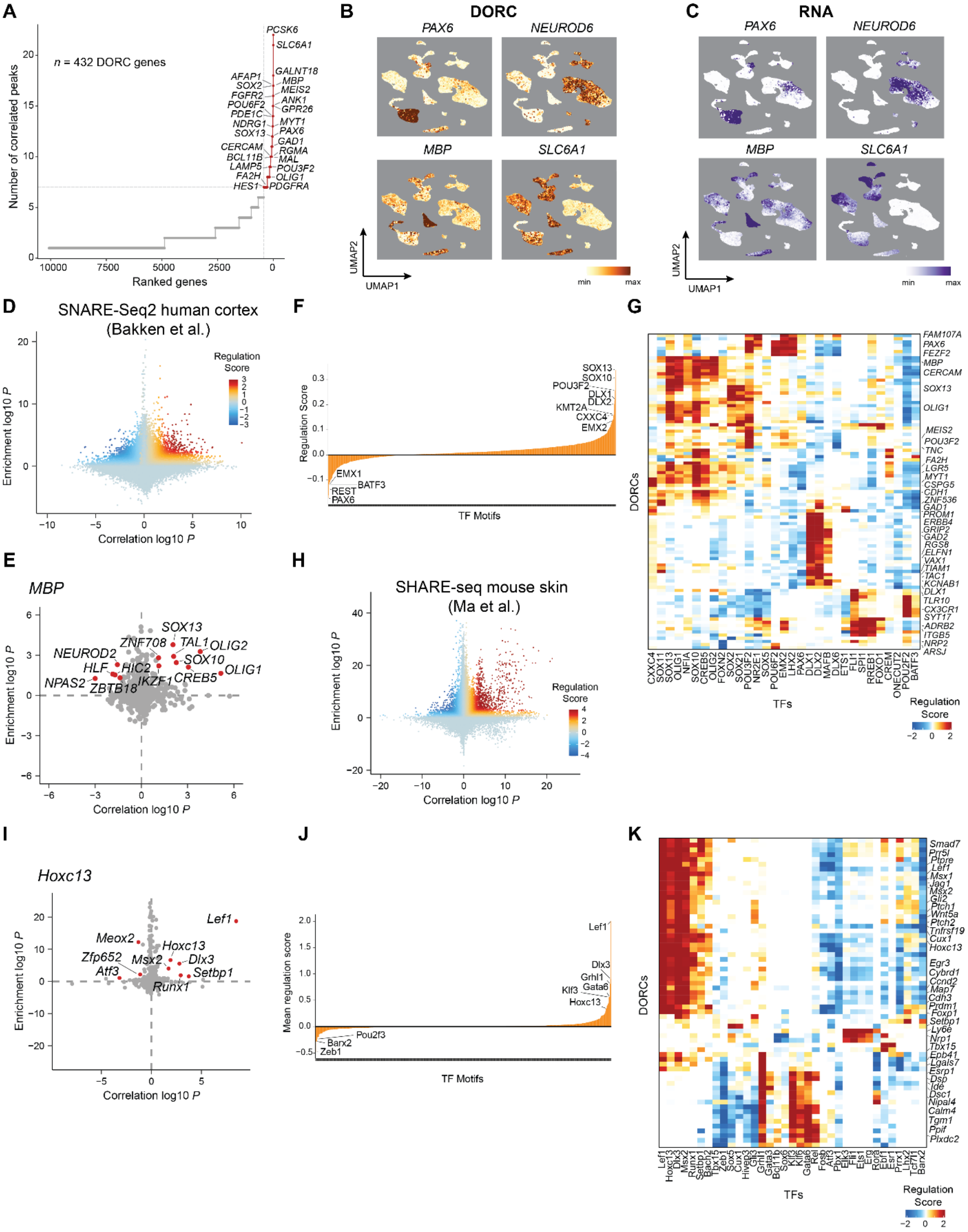
A
n = 432 DORC genes
Number of correlated peaks
Ranked genes
10000
7500
5000
2500
0
20
15
10
5
PCSK6
SLC6A1
GALNT18
AFAP1
MBP
SOX2
MEIS2
FGFR2
ANK1
POU6F2
GPR26
PDE1C
MYT1
NDRG1
PAX6
SOX13
GAD1
CERCAM
RGMA
BCL11B
MAL
LAMP5
POU3F2
FA2H
OLIG1
HES1
PDGFRA
B
DORC
PAX6
NEUROD6
MBP
SLC6A1
UMAP2
UMAP1
min
max
C
RNA
PAX6
NEUROD6
MBP
SLC6A1
UMAP2
UMAP1
min
max
D
SNARE-Seq2 human cortex
(Bakken et al.)
Enrichment log10 P
Correlation log10 P
Regulation Score
3
2
1
0
−1
−2
−3
−10
−5
0
5
10
20
10
0
E
MBP
SOX13
ZNF708
TAL1
OLIG2
NEUROD2
SOX10
HLF
HIC2
OLIG1
CREB5
NPAS2
IKZF1
ZBTB18
Enrichment log10 P
Correlation log10 P
6
3
0
−3
−6
F
Regulation Score
TF Motifs
0.3
0.2
0.1
0.0
−0.1
SOX13
SOX10
POU3F2
DLX1
DLX2
KMT2A
CXXC4
EMX2
EMX1
BATF3
REST
PAX6
G
DORCs
TFs
FAM107A
PAX6
FEZF2
MBP
CERCAM
SOX13
OLIG1
MEIS2
POU3F2
TNC
FA2H
LGR5
MYT1
CSPG5
CDH1
ZNF536
GAD1
PROM1
ERBB4
GRIP2
GAD2
RGS8
ELFN1
VAX1
TIAM1
TAC1
KCNAB1
DLX1
TLR10
CX3CR1
SYT17
ADRB2
ITGB5
NRP2
ARSJ
CXXC4
SOX11
SOX13
OLIG1
NFIA
SOX10
CREB5
OLIG2
FOXN2
SOX2
POU3F2
NR2E1
SOX5
POU6F2
EMX2
LHX2
PAX6
DLX1
DLX2
MAFB
DLX6
ETS1
FLI1
SPI1
RREB1
FOXO1
CREM
ONECUT2
POU2F2
BATF3
Regulation Score
−2
0
2
H
SHARE-seq mouse skin
(Ma et al.)
Enrichment log10 P
Correlation log10 P
Regulation Score
4
2
0
−2
−4
−20
−10
0
10
20
40
20
0
−20
I
Hoxc13
Lef1
Meox2
Hoxc13
Zfp652
Msx2
Dlx3
Atf3
Setbp1
Runx1
Enrichment log10 P
Correlation log10 P
20
10
0
−10
−20
−5
0
5
J
Mean regulation score
TF Motifs
2.0
1.5
1.0
0.5
0.0
−0.5
Lef1
Dlx3
Grhl1
Gata6
Klf3
Hoxc13
Pou2f3
Barx2
Zeb1
K
DORCs
TFs
Smad7
Prr5l
Ptpre
Lef1
Msx1
Jag1
Msx2
Gli2
Ptch1
Wnt5a
Ptch2
Tnfrsf19
Cux1
Hoxc13
Egr3
Cybrd1
Ccnd2
Map7
Cdh3
Prdm1
Foxp1
Setbp1
Ly6e
Nrp1
Tbx15
Epb41
Lgals7
Esrp1
Dsp
Ide
Dsc1
Nipal4
Calm4
Tgm1
Ppif
Plxdc2
Lef1
Hoxc13
Dlx3
Msx2
Runx1
Setbp1
Bach2
Tbx15
Zeb1
Sox5
Cux1
Hivep3
Gli3
Grhl1
Gata3
Bcl11b
Sox9
Klf3
Klf6
Gata6
Rel
Fosb
Atf3
Pbx1
Elk3
Fli1
Ets1
Erg
Rora
Ebf1
Esr1
Prrx1
Lhx2
Tcf7l1
Barx2
Regulation Score
−2
0
2

Figure S7. FigR GRN application to human cortex and mouse skin multi-omic data (related to Figure 5). **A.** Scatter plot highlighting the number of significant peak-gene associations (permutation $P \leq 0.05$) determined for human cortex cells profiled using SNARE-Seq2 (Bakken et al.[1]). Select DORCs are highlighted among all detected DORCs (n=432 DORCs with $\geq$ 7 peaks; red points). **B-C.** UMAP of human cortex cells (Fig S3H) with cells colored by DORC accessibility scores **(B)** or RNA expression **(C)** shown for *PAX6*, *NEUROD6*, *MBP* and *SLC6A1*. **D.** Scatter plot showing TF-DORC associations based on TF motif enrichment among DORC peaks, and TF RNA correlation to DORC accessibility, respectively, colored by FigR's regulation score. **E.** Same as in D., but restricted for putative TF drivers (absolute(regulation score) $\geq$ 1) of *MBP* (red points). **F.** Ranked plot of the mean regulation score across all DORCs per TF (similar to Fig 5E), highlighting top activating and repressive TFs across all DORCs determined for human cortex cells. **G.** Heatmap of regulation scores between filtered TFs and DORCs (absolute(regulation score) $\geq$ 2.5; similar to Fig 5F) determined for human cortex cells. **H.** Same as in D, but run using mouse skin SHARE-seq data (Ma et al.[2]). **I.** Same as in H, but restricted for putative TF drivers (absolute (regulation score) $\geq$ 1) of DORC *Hoxc13*. **J.** Same as in F, but corresponding to associations shown in H for mouse skin. **K.** Heatmap of regulation scores between top TFs and a filtered subset of previously determined DORCs (log10 absolute(regulation score) $\geq$ 2) for mouse skin SHARE-seq data.

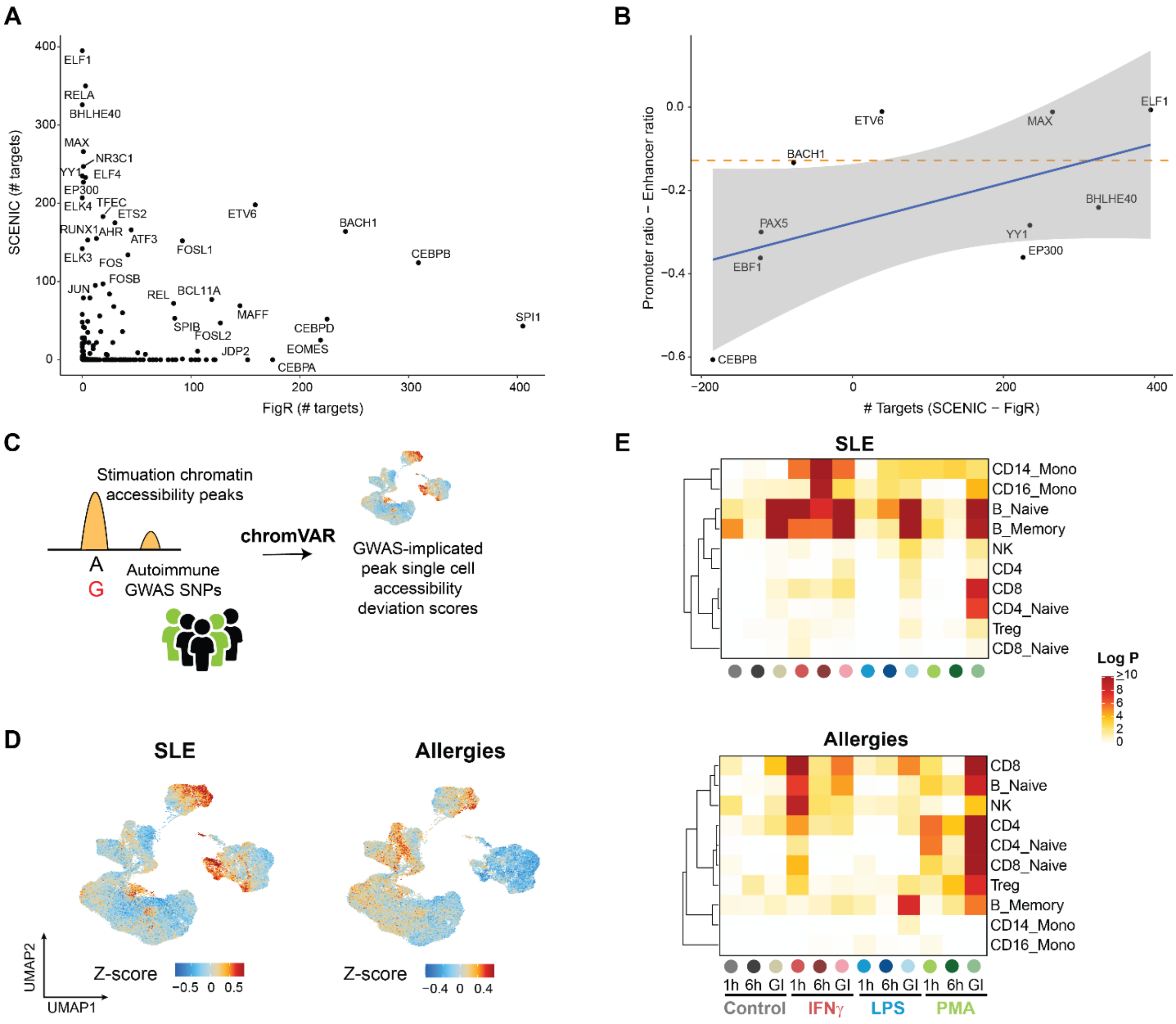

Figure S8. Comparison of FigR and SCENIC and interrogation of GWAS-variant linked accessible elements across single cells and stimulus conditions (related to Figure 5). **A.** FigR and SCENIC comparison. Scatter plot of the total number of targets associated with TFs determined using SCENIC's co-expression framework (y-axis) versus FigR's approach (x-axis) using PBMC stimulation data. **B.** Scatter plot of the top SCENIC or top FigR TFs from A (*n*=5 each) that also have TF ChIP-seq data (ENCODE GM12878), plotting the difference between the fraction of ChIP-seq-determined binding sites that fall in promoter elements and the fraction that overlaps distal enhancer elements (y-axis), versus the total number of inferred target genes in SCENIC minus that in FigR (x-axis). Dotted line represents the mean (expected) difference in promoter ratio and enhancer ratio determined across all TFs with ChIP-seq data (*n*=84). Gray shading represents the 95% confidence interval of the linear fit. **C.** Overview of our approach to integrate GWAS variants for 14 autoimmune/inflammatory diseases with stimulation scATAC-seq data. **D.** UMAP of scATAC-seq cells (*n*=62,219) colored by accessibility Z-scores based on peak-SNP overlaps for Systemic Lupus Erythematosus (SLE) and Allergies GWAS variants. Scores shown are smoothed among *k*=50 cell nearest neighbors, and thresholded at +/- 3 s.d. for visualization.

E. Heatmap of significance estimates for peak-SNP overlap Z-score combined per condition and per cell type using Fisher's method, shown for SLE and Allergies GWAS variants.

Supplemental References

[1] T. E. Bakken *et al.*, "Comparative cellular analysis of motor cortex in human, marmoset and mouse," *Nature*, vol. 598, no. 7879, pp. 111–119, Oct. 2021.

[2] S. Ma *et al.*, "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin," *Cell*, vol. 183, no. 4, pp. 1103–1116.e20, Nov. 2020.