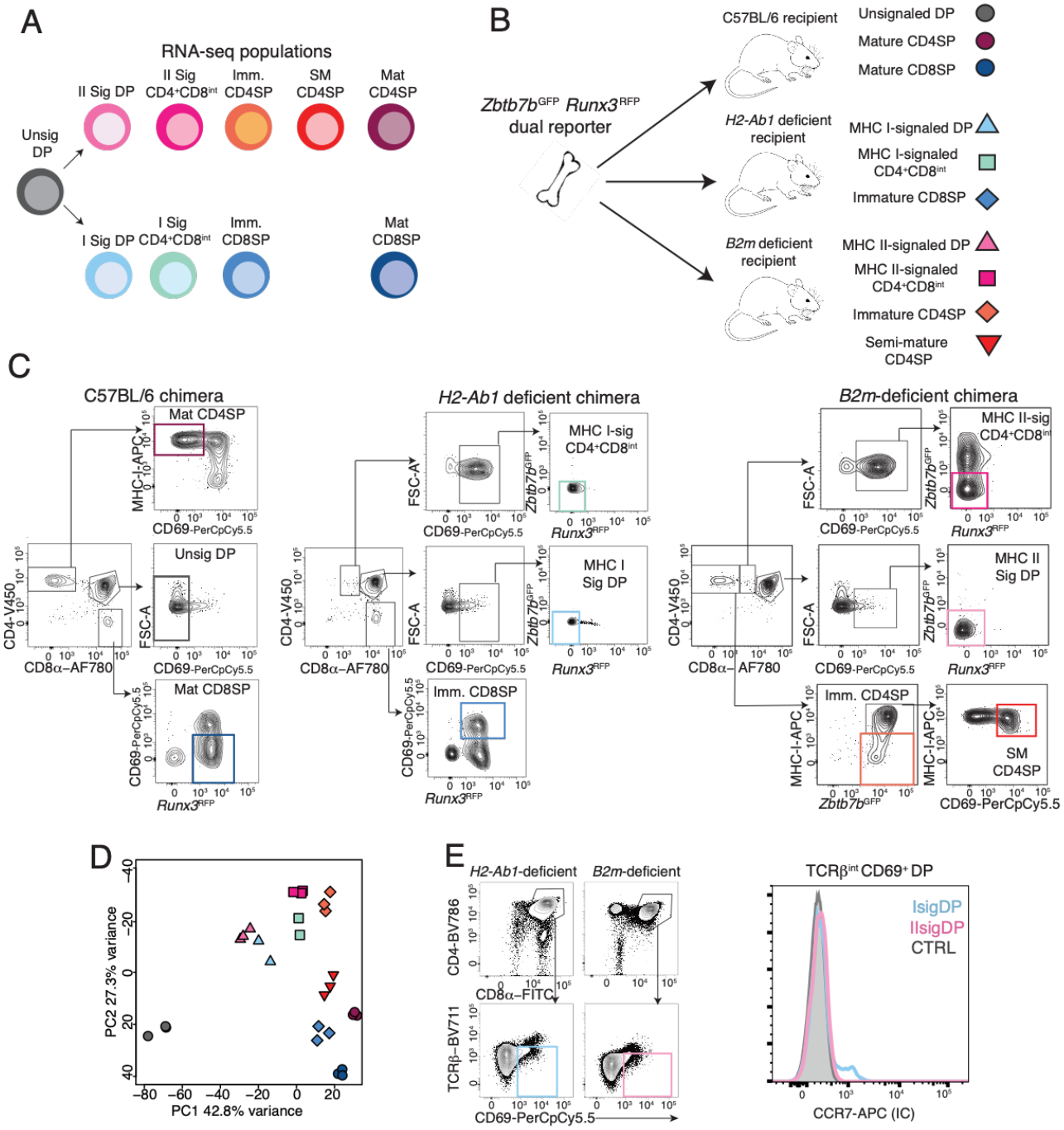

Figure S1 (related to Figure 1): Population transcriptomic analyses of mouse thymocytes

(A) Schematic showing developmental intermediates sorted for population-based RNA-seq of developing mouse thymocytes.

(B) Congenic *Zbtb7b*^{GFP} *Runx3*^{RFP} bone marrow was transplanted into lethally irradiated C57BL/6, *H2-Ab1*-deficient (MHC II), and *B2m*-deficient mice to obtain unsignaled (grey), MHC I- (blue symbols),

and MHC II- (red symbols) signaled thymocyte subsets. Two distinct sets of chimeras were generated in two independent experiments for a total of 2-3 biological replicates of each population.

(C) Flow cytometry sorting strategy for obtaining the populations indicated in (B). Populations shown were gated as $CD44^{low}$, $CD45.2^{+}$. Final sort gates are surrounded by colored boxes corresponding to the colors shown in (B).

(D) Principal-component analysis (PCA) displays cell subsets according to the first and second components. Each symbol [defined in (B)] represents an individual biological replicate.

(E) (Left) Flow cytometry strategy for staining MHC-I and MHC-II signaled DP thymocytes from germline *H2-Ab1*-deficient and *B2m*-deficient mice. Populations pre-gated as $CD44^{low}$. As previously reported, *H2-Ab1*-deficient thymocytes exhibit elevated TCRβ surface expression. (Right) Overlaid histograms of CCR7 measured by intracellular flow cytometry on MHC-I and MHC-II signaled DP thymocytes. CTRL is C57BL/6-derived unsignaled DP thymocytes. Representative of 3 individual mice of each genotype, in two independent experiments.

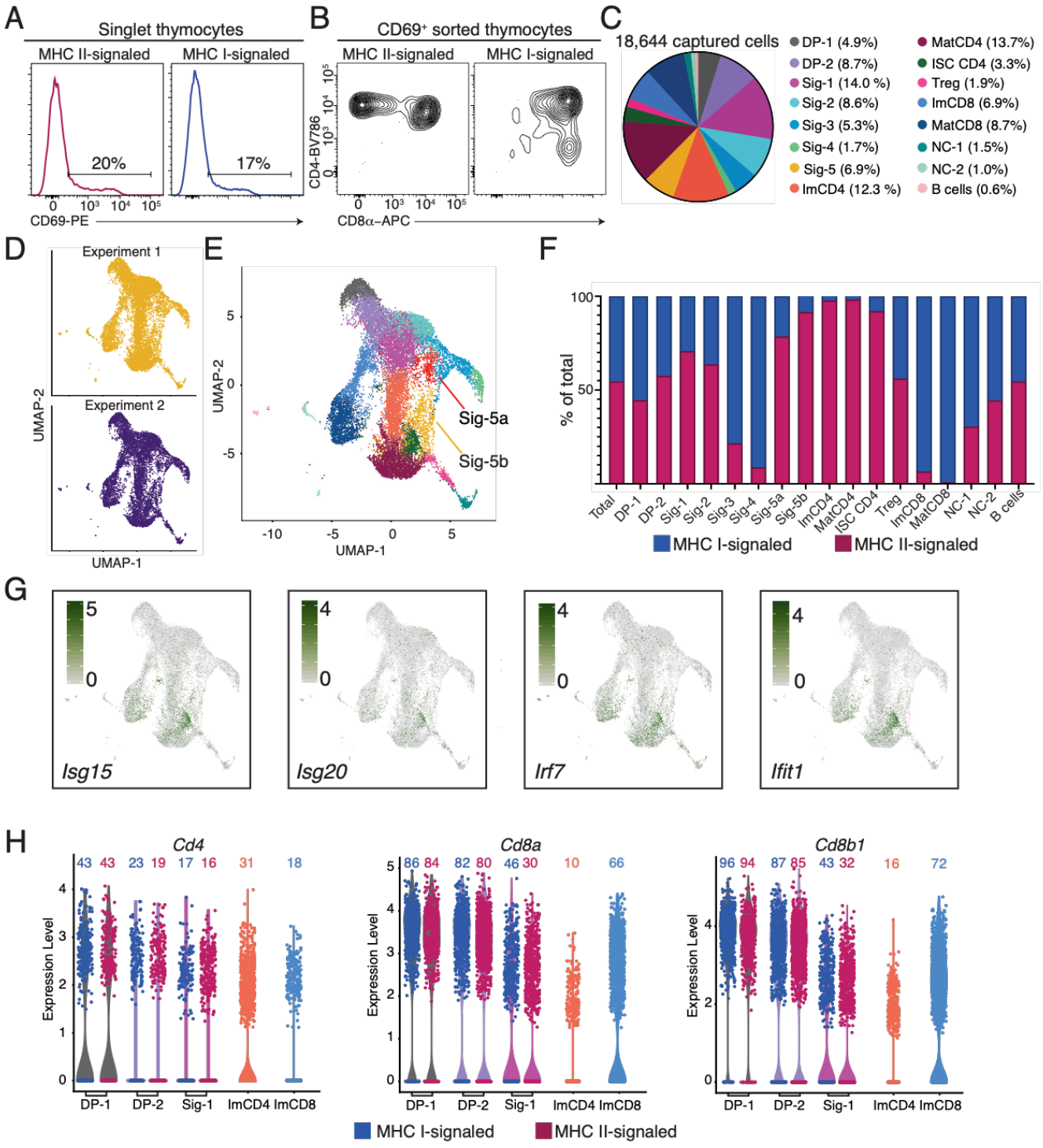

Figure S2 (related to Figure 2): Single cell transcriptomic analysis of mouse αβ T cell development

(A-H) MHC I- and MHC II-signaled thymocytes (from *H2-Ab1*- and *B2m*-deficient mice, respectively) were sorted, processed for scRNAseq using the 10x Chromium technology and analyzed with the R

Seurat package. Data integrates two biological replicates from each genotype from two distinct experiments.

(A) Flow cytometry plot showing expression of CD69 on thymocytes from *H2-Ab1*- and *B2m*-deficient mice. The indicated sort gate allows for inclusion of some $CD69^{low}$ cells in the population.

(B) Flow cytometry plots show CD4/CD8α profile of sorted cells from *H2-Ab1*- and *B2m*-deficient thymi prior to loading onto the 10X Chromium system.

(C) The distribution of all analyzed thymocytes into clusters (defined in Fig. 2A). Each slice of the pie corresponds to a cluster, and the percentage indicates the number of cells in that cluster divided by the total number of cells analyzed.

(D) UMAP dimensional reduction plot generated as in Fig. 2A displaying cells color-coded according to experimental replicates.

(E) UMAP dimensional reduction plot generated as in Fig. 2A displaying all analyzed thymocytes. Cluster Sig-5 was split into two clusters based on their location on the UMAP. The top half of the cluster is low for *Ccr7* expression, and is designated Sig-5a, while Sig-5b is positive for *Ccr7.*

(F) Bar chart depicting the proportion of MHC-I and MHC-II signaled cells within each cluster defined as in Fig. 2A and S2E.

(G) Scaled expression of indicated genes is shown on UMAP plots generated as in Fig. 2A.

(H) Violin plots depicting expression of *Cd4*, *Cd8a,* and *Cd8b1,* split by MHC-specificity within each indicated cluster. Numbers above violins indicates percentage of expressing cells within the corresponding cell population.

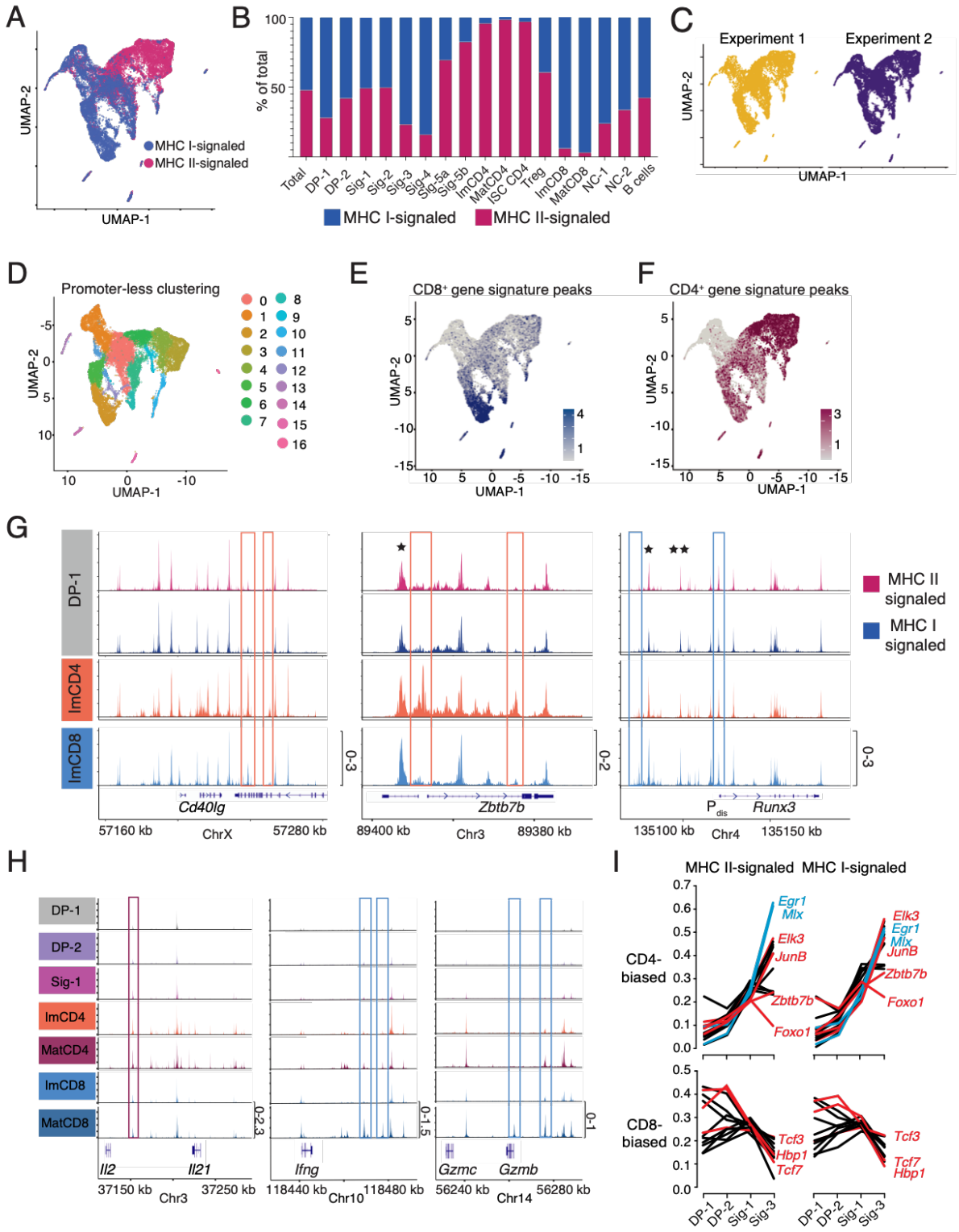

Figure S3 (related to Figure 3): Single cell epigenomic analysis of mouse αβ T cell development

(A-H) MHC I- and MHC II-signaled thymocytes (from *H2-Ab1*- and *B2m*-deficient mice, respectively) were prepared (as in Fig. S2AB), processed for scATACseq using the 10x Chromium technology and

analyzed with the Signac extension of the R Seurat package. Data integrates two biological replicates from each genotype from two distinct experiments.

(A) Same UMAP plot as in Fig. 3A, with cells colored according to their MHC restriction.

(B) Bar chart depicting the proportion of MHC-I and MHC-II signaled cells within each population of cells with a shared inferred ID defined as in Fig. 3C.

(C) UMAP plot as in Fig. 3A displaying cells color-coded according to experimental replicates.

(D) Peaks mapping to gene promoters [designation from Homer peak annotation] were removed from the peaks used for dimensional reduction and clustering. The UMAP plot displayed was generated from this "promoter-less" peak set. Cells are color-coded according to their distribution into clusters defined by unsupervised clustering.

(E, F) Scaled accessibility scores for individual cells at peaks near $CD8^+$- (E) or $CD4^+$- (F) lineage signature genes (Figs. 1EF and Table S1) are projected onto the same UMAP plot as in Fig. 3A.

(G) Genome browser tracks show scATACseq signals at indicated genes (bottom), displayed as scaled sequence read density averaged for all cells sharing the indicated inferred identity (left). DP-1 thymocyte tracks are shown and color-coded separately according to MHC restriction (color code on the right). Boxes indicate lineage specific peaks. Stars indicate the *Zbtb7b* silencer and two *Runx3* enhancers identified by mutagenesis (He et al., 2008; Kojo et al., 2017; Setoguchi et al., 2008).

(H) Genome browser tracks show scATACseq signals at indicated genes (bottom), displayed as scaled sequence read density averaged for all cells sharing the indicated inferred identity (left). Boxes indicate lineage specific peaks.

(I) Line graphs depict the scores of the top 15 significantly CD4-biased transcriptional activities (top) and the 12 significantly CD8-biased transcriptional activities (bottom) across MHC-II signaled (left) and MHC-I signaled (right) DP-1, DP-2, Sig-1, and Sig-3 cells.

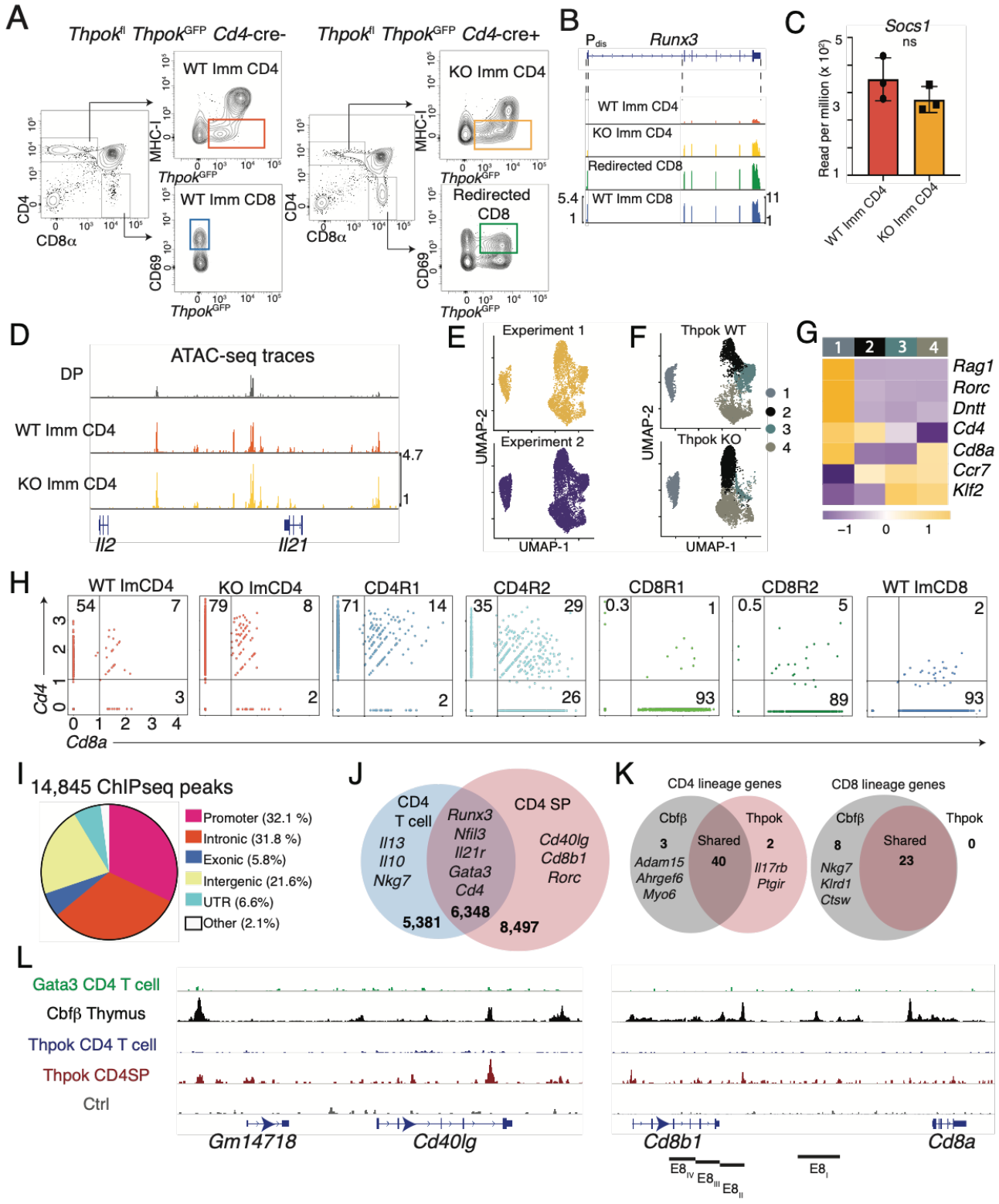

Figure S4 (related to Figure 4): Transcriptomic and genomic footprint of Thpok

(A) Thymocyte populations were sorted from wild-type (*Zbtb7b*$^{fl/fl}$ *Cd4*-cre negative, WT) or *Zbtb7b*-deficient (*Zbtb7b*$^{fl/fl}$ *Cd4*-cre positive, KO) mice, both carrying a *Zbtb7b*^{GFP} BAC reporter. Flow

cytometry sorting strategy for obtaining WT immature $CD4^+$ SP and $CD8^+$ SP (left), and KO immature $CD4^+$ SP and redirected $CD8^+$ SP (right). All cells pre-gated as singlet $CD44^{low}$. Three independent biological replicates of each population were sorted and processed for population RNA-seq.

(B) Genome browser tracks show RNAseq signals at *Runx3*, displayed as scaled counts per million sequence read density.

(C) Bar graph showing reads per million normalized expression of *Socs1* in wild-type and *Zbtb7b*-deficient immature $CD4^+$ SP sorted as in Fig. S4A. Significance calculated using a two-tailed student's t-test; ns = no significance. Error bars represent SEM.

(D) Population ATAC-seq was performed on wild-type or *Zbtb7b*-deficient immature $CD4^+$ SP thymocytes sorted as in Fig. S4A, and unsignaled wild-type DP thymocytes ($CD4^+CD8^+CD69^–$). Genome browser tracks show ATAC-seq signals at the *Il2-Il21* gene locus. Representative of three biological replicates per genotype.

(E-H) scRNAseq comparison of (i) $Zbtb7b^{GFP+}$, $CD8^+$ SP, and $CD69^–$ unsignaled DP thymocytes from wild-type ($Zbtb7b^{fl/fl}$ *Cd4*-$cre^–$) $Zbtb7b^{GFP}$ BAC reporter mice and (ii) $Zbtb7b^{GFP+}$ and unsignaled DP thymocytes from Thpok-deficient ($Zbtb7b^{fl/fl}$ *Cd4*-cre^+) $Zbtb7b^{GFP}$ BAC reporter mice. Data integrates two biological replicates from each genotype from two distinct experiments.

(E) UMAP dimensional reduction plot displaying cells color-coded according to experimental replicates.

(F) UMAP dimensional reduction plot generated as in (E), color-coded according to cell distribution into clusters defined by low-resolution unsupervised clustering.

(G) Heatmap shows row-standardized expression (Z-scores of average values per cluster, scale at bottom) of relevant genes in indicated clusters. Gene expression patterns identify cluster 1 as unsignaled DP thymocytes, cluster 2 as immature $CD4^+$ SP thymocytes and clusters 3 and 4 as mature $CD4^+$ and $CD8^+$ SP thymocytes, respectively. Note the contribution of cells of both genotypes to clusters 1, 2 and 4.

(H) Scatter plots of *Cd4* and *Cd8a* expression in indicated clusters defined as shown in Fig. 4CD. Numbers indicate the frequency of cells within each quadrant.

(I-L) ChIP-seq on sorted $CD4^+$ SP thymocytes sorted from $Zbtb7b^{Bio/+}$ $Rosa26^{BirA+}$ (Thpok) and $Zbtb7b^{+/+}$ $Rosa26^{BirA+}$ (Ctrl) mice. ChIPseq data representative of three biological replicates.

(I) Pie chart shows genome-wide distribution of Thpok binding [annotated from Homer].

(J) Thpok binding sites in $CD4^+$ SP thymocytes were compared to Thpok binding sites in peripheral CD4 T cells (GSE116506) by using the MergePeaks function of Homer. Venn diagram depicts numbers of shared and unique binding sites (peaks) with example nearby genes shown.

(K) Thpok binding sites (from thymus $CD4^+$ SP) were compared to Cbfβ binding sites (thymus, GSE90794) near genes of the $CD4^+$ (left) and $CD8^+$ (right) signatures defined as in Fig. 1EF and Table S1. Venn diagram depicts shared and specific binding near these genes.

(L) Genome browser tracks show counts per million normalized ChIP-seq signals for indicated proteins on the *Cd40lg* and *Cd8b1-Cd8a* gene bodies. Relevant CD8 enhancers are depicted. Data for Gata3 ChIP-seq on CD4 T cells from GSE20898, and Cbfβ ChIP-seq on thymocytes from GSE90794.

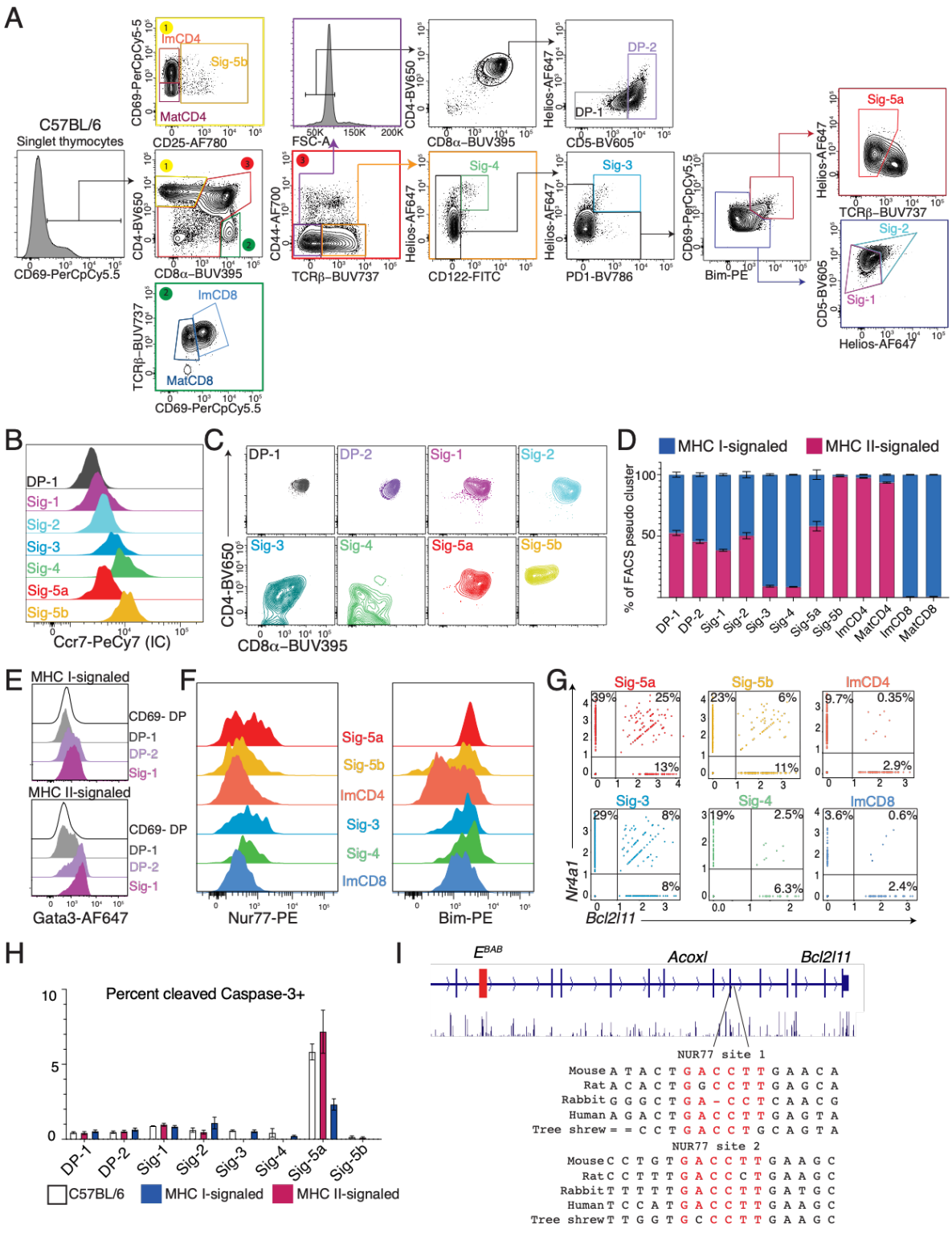

Figure S5 (related to Figures 2 and 5): Flow cytometric validation of scRNAseq clustering

(**A**) Flow cytometric analysis of thymocytes from C57BL/6 mice identifying cell subsets corresponding to scRNAseq clusters identified in Fig. 2. Column 1 is the starting population for all downstream gates.

Column 2 separates $CD4^+$ SP (1) and $CD8^+$ SP (2) cells, from other thymocytes analyzed in downstream gates (3). Unsignaled cells with high forward light scatter were excluded because their large size confounds analyses of low expression intra-cellular proteins. Representative of 6 independent experiments on 6 mice.

(B) Intracellular Ccr7 staining on indicated thymocyte subsets, defined as in (A).

(C) Expression of CD4 and CD8α on indicated thymocyte subsets, defined as in (A).

(D) Bar chart depicting the respective contribution of MHC I- and MHC II-signaled cells within each flow cytometric subset defined in (A), as determined by staining of thymocytes from *H2-Ab1*- and *B2m*-deficient mice. Data depicts the results from 3 mice of each genotype analyzed side-by-side in 3 independent experiments. Error bars represent SEM.

(E) Intracellular Gata3 staining on indicated thymocyte subsets, gated as in (A), from *H2-Ab1*- and *B2m*-deficient thymi, defining MHC I- and MHC II-signaled cells, respectively.

(F) Intra-cellular staining for Nur77 (left) and Bim (right) on indicated thymocyte subsets, defined as in (A).

(G) Scatter plots show *Nr4a1* and *Bcl2l11* mRNA expression in indicated scRNAseq clusters (defined in Fig. 2A and S2E). Percentages indicate frequency of cells within each gate.

(H) Bar chart depicting frequency of Cleaved Caspase-3^+ cells within thymocyte subsets, defined as in (A). Data depicts the results from 3 mice of each genotype analyzed side-by-side in 3 independent experiments. Error bars represent SEM.

(I) Schematic view of the mouse *Bcl2l11* (Bim) locus. Mammalian sequence conservation indicated by blue bars beneath the gene body. Conserved Nur77 binding sites within open chromatin regions are shown. Red letters show the consensus sequence.

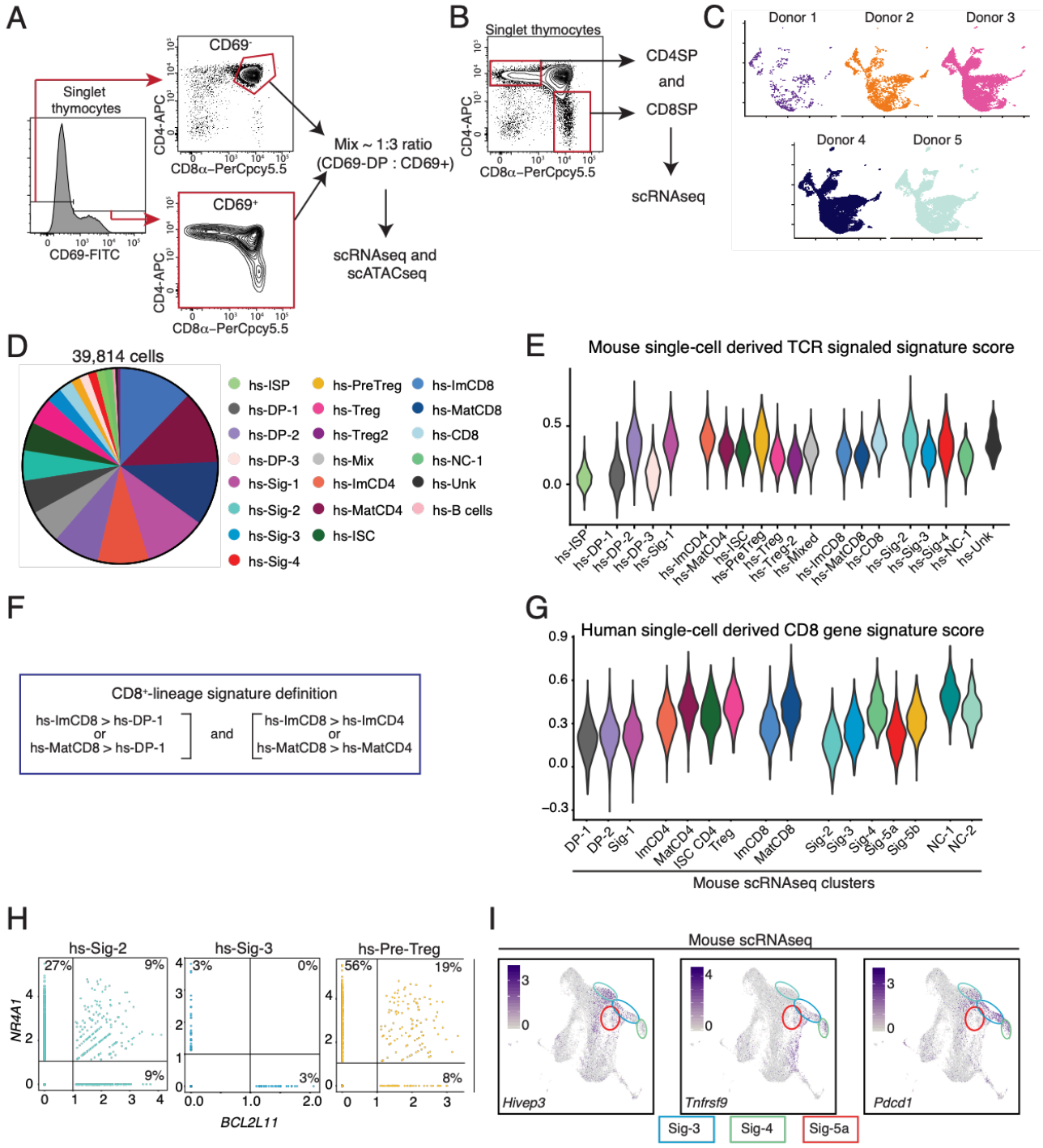

Figure S6 (related to Figure 6): Single-cell RNAseq on human thymocytes

(A-E) Thymocytes sorted from human donors (Table S7) were sorted, processed for scRNAseq using the 10x Chromium technology and analyzed with the R Seurat package. Data integrates nine distinct samples from five individual donors

(A) Sort strategy for enriching signaled thymocytes from human donors for processing by scRNAseq and scATACseq. $CD69^{low}$ DP cells were mixed with $CD69^{+}$ cells in a 1:3 ratio prior to processing for scRNAseq and scATACseq.

(B) $CD4^{+}$ SP and $CD8^{+}$ SP thymocytes were sorted from two donors, in addition to samples sorted as in (A). Singlet thymocytes were sorted as $CD4^{+}$ SP and $CD8^{+}$ SP, regardless of CD69 levels. Populations were captured separately from each other and from the mixture sorted as in Fig. S6A.

(C) Same UMAP plot as in Fig. 6A, showing cells color-coded according to donor origin. Each dot represents a cell. Note the higher contribution of donors 4 and 5 to clusters corresponding to $CD4^{+}$ and $CD8^{+}$ SP populations, which were separately sorted for these donors (as described in experimental procedures).

(D) The distribution of all analyzed thymocytes into clusters defined as in Fig. 6A. Each slice of the pie corresponds to a cluster.

(E) Violin plots show expression scores (cluster-based averages of scores for each individual cell) of mouse TCR-induced signature genes (Fig. 2F and Table S1) in cell clusters defined as in Fig 6A.

(F) Definition of a human $CD8^{+}$ gene signature score, with Log2 fold change cutoff of +0.3.

(G) Violin plots show expression scores (cluster-based averages of scores for each individual cell) of the $CD8^{+}$ gene signature defined in F, in mouse cell clusters defined in Fig. 2A and S2E.

(H) Scatter plots show expression of *Nr4a1* and *Bcl2l11* in indicated clusters as defined in Fig. 6A. Percentages indicate frequency of cells within each gate.

(I) Scaled expression of indicated genes is shown on mouse UMAP plots generated as in Fig. 2A.

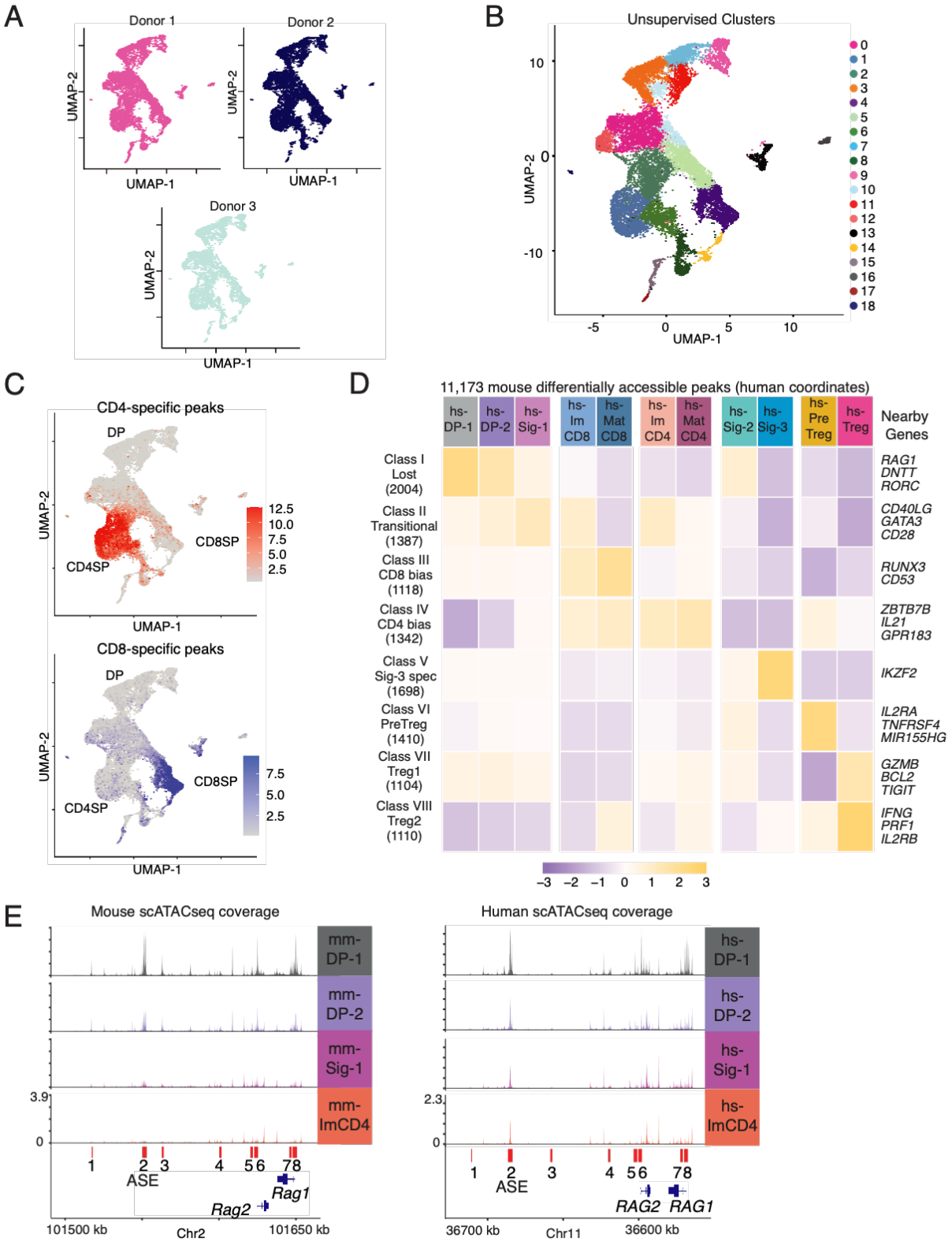

Figure S7 (related to Figure 7): Evolutionary conservation of epigenomic programs

(A-D) Thymocytes sorted from human donors (Table S7) were prepared as in Fig. S6A, processed for

scATACseq using the 10x Chromium technology, and analyzed with the Signac extension of the R Seurat package. Data integrates thymocytes from three distinct donors.

(A, B) UMAP dimensional reduction plot of scATACseq data displaying all analyzed human thymocytes (each dot is an individual cell), color-coded according to their donor origin (A) or to their distribution into clusters defined by unsupervised clustering (B).

(C) Scaled accessibility scores for individual cells at $CD4^+$- and $CD8^+$-lineage specific peaks are projected onto the same UMAP plot as in (A).

(D) 11,173 human genomic regions syntenic to differentially accessible regions from murine thymocytes (defined as in Fig. 3D) were identified using the UCSC LiftOver function. Heatmap shows row-standardized k-means classified accessibility. Population names are indicated at the top of the figure. Gene assignment of open chromatin regions performed by Homer, which identifies the closest transcription start site. Examples of nearby genes are indicated to the right of the heatmap.

(E) Genome browser tracks show scATACseq signals at the mouse *Rag1-Rag2* (left) and human *RAG1-RAG2* (right) gene loci displayed as scaled sequence read density averaged for all cells sharing the indicated inferred identity (right). Red boxes above gene names indicate sequence-conserved open chromatin regions, and the location of the Anti-Silencer Element is indicated (Yannoutsos et al., 2001).