Supplementary Figure Legends

Figure S1, related to Fig. 1. Cell purification and biological reproducibility.

A. $Lin^-c\text{-}Kit^+CD27^+$ fetal liver precursors were sorted from E13.5 to 14 C57BL/6 mouse embryos and co-cultured with OP9-DL1 stromal cells in the presence of IL-7 and Flt3-L. At day 4.5, FLDN1 cells were sorted from the co-culture as $Lin^-CD45^+cKit^{hi}Cd44^+$ $CD25^-$, FLDN2a as $Lin^-CD45^+cKit^{hi}Cd44^+CD25^+$. At day 8.5, FLDN2b cells were sorted as $Lin^-CD45^+cKit^{lo}Cd44^-$ $CD25^+$, and FLDN2a cells were also collected (see experimental procedure).

B. ThyDN3 cells were sorted from lineage-depleted thymocytes as $Lin^-c\text{-}Kit^-CD44^-CD25^+$ (see experimental procedure).

C. ThyDP cells were collected from B6 background $TCR\alpha^{-/-}$ thymi by depletion with Ter-119, CD19, F4/80, Gr-1, NK1.1, CD11c, c-Kit and CD44. After depletion, 90~95% of cells were $CD4^+CD8^+$(see experimental procedure). Cells lacking TCRα were used in order to enable DP cells to be generated normally, a process dependent only on TCRβ, but leaving them without the capacity to undergo positive selection to any later stages of T-cell development.

D. Global comparisons of RNA-seq between biological duplicates or triplicates (log_2 transformed) for individual cell populations (including FLDN1 sample I vs. II, FLDN2a sample I vs. II, FLDN2b sample I vs. II, FLDN2b sample II vs. III, ThyDN3 sample I vs. II, and ThyDP sample I vs. II). The Pearson's correlation coefficient for each comparison is indicated in the insert.

E. Global comparisons of histone modifications ChIP-seq experiments between biological duplicates (after $\log_2$ transformed) for individual cell populations (including FLDN1 experiment A vs. B, FLDN2b experiment A vs. B, ThyDN3 experiment A vs. B, and ThyDP experiment A vs. B). From top row to bottom are H3Ac, H3K4me2 and H3K27me3. The Pearson's correlation coefficient for each comparison is indicated in the insert.

Figure S2, related to Fig. 2. Global distribution of the three histone modifications

A. Length distribution of discrete genomic regions enriched with the specified histone modification in at least one cell population. First, more regions were enriched with H3K4me2 in at least one cell population than H3Ac or H3K27me3. Second, while the majority of enriched regions were within 500~5,000 bp (>95% for H3Ac and H3K4me2; 81.5% for H3K27me3), H3K27me3 enriched regions were relatively more diverse in length, including some regions of much greater extent.

B. Signal density plots of histone modifications at individual promoter-proximal or distal regions in each population. Signal densities were calculated for every discrete genomic region enriched with at least one histone modification in at least one population, then $\log_2$ transformed and normalized (see experimental procedure). X-axis indicates H3K27me3 value and y-axis indicates H3K4me2 value, while the color coding specifies the signal density of H3Ac in each element. In each stage, total regions were separated into four subgroups, based on the enrichment of H3K4me2 (≥4 RPM as positive) and H3K27me3 (≥2 RPM as positive). The percentage of regions for each group is indicated in the insert.

C. Distribution of histone modifications at the promoters with respect to transcriptional activity. Based on the transcriptional activity, total Refseq annotated genes were divided into subgroups of expressed genes (in red table, normalized RNA-seq value ≥ 1 RPKM) and silent genes (in

blue table, normalized RNA-seq value < 1 RPKM). Each number indicts total amount of promoters enriched with the specified histone modification or combination of histone modifications. The percentage of promoters enriched with the specified histone modification or combination of histone modifications within the respective subgroup is given in the parenthesis.

Figure S3, related to Figs. 2 & 3. Global histone modifications and gene expression profiles at both promoter regions and distal regions.

A. The normalized signal densities of histone modifications at the promoter region were aligned with the normalized mRNA data for each of 20,861 Refseq annotated genes, and one-dimensional (along genes) hierarchical clustering was performed using Ward linkage and Euclidean distance.

B. Each distal region was assigned to the closest gene, and one-dimensional hierarchical clustering, along genes, was performed as in Figure S3A.

Figure S4, related to Fig. 4. Dynamic histone modifications and transcription factor binding are linked to differentially regulated gene expression.

A. UCSC genome browser tracks depicting H3K4me2 and gene expression at *Notch1* locus. A *Notch1* distal alternative promoter or cis-regulatory element (black arrow, most 5' component of H3K4me2 peak) activity was upregulated at DN2b and DN3 stages.

B. UCSC genome browser tracks depicting GATA-3 binding, together with binding associated H3K4me2 and gene expression, at *Rag1/2* loci. Black arrows indicate the DP specific Rag1/2 enhancer, where GATA-3 occupancy was observed at a low level in DN2b stage and sharply increased in DP stages, as did binding associated H3K4me2. Due to the strong upregulation of

gene expression of *Rag1* and *Rag2* from DN3 to DP, note that in this panel the range of RNA-seq signal densities of Rag1 and Rag2 for DP (0.02 to 100 in red) is different from the one for other stages (0.02 to 16).

C. UCSC genome browser tracks showing GATA-3 binding, together with binding associated H3K4me2 at *Cd3e/d/g* loci.

D&E. UCSC genome browser tracks depicting histone modifications and gene expression at *Il2ra* (D) and *Ebf1* (E) loci.

F&G. UCSC genome browser tracks depicting PU.1 binding, together with binding associated histone modification(s) and gene expression (of *Flt3*), at *Flt3* (F) and *Hhex* (G) loci. PU.1 occupancies, as well as binding associated H3K4me2, at the promoters of *Flt3* and *Hhex* (black arrows) decreased from DN1 to DN2a stages, and completely disappeared in DN2b, in parallel with the gene expression pattern of two genes (*Hhex* expression and histone modification patterns are depicted in Figure 4E).

H&I. UCSC genome browser tracks depicting GATA-3 binding, together with binding associated histone modifications and gene expression, at *Tcf7* (H) and *Zbtb7b* (I) loci. GATA-3 occupancy at an upstream distal region of *Tcf7* (black arrow) increased from DN1 to DN2b, as gene expression of *Tcf7* was upregulated from DN1 to DN2b. Although repressed by H3K27me3 in DN2b and DP (pre-positive selection), *Zbtb7b* upstream distal region was constantly bound by GATA-3 (black arrow) in both stages. Note that the range of RNA-seq signal densities of Tcf7 for DP (0.02 to 80 in red) is different from the one for other stages (0.02 to 40).

Figure S5, related to Fig. 2 and Fig. 6. K−means clustering for differentially expressed genes.

3,697 differentially expressed genes were subjected to K-means clustering analysis that inferred to 25 differentially expressed patterns (see Experimental Procedures). Error bar represents the standard deviation of biological replicates or triplicates of individual genes at each stage. The genes in each cluster are listed in order in Table S4A.

Figure S6, related to Figs. 5 and 6. PU.1 occupancy associated with epigenetic modifications and gene expression during early T cell development. A. Cumulative distributions of changes in PU.1 occupancy between FLDN2b and FLDN1 among promoter-proximal sites (top) and promoter-distal sites (bottom). The total PU.1 binding sites in DN cells (as depicted in Figure 5B) were separated in two groups, promoter-proximal sites (≤1 kb from nearby TSS) and promoter-distal sites (>1 kb from nearby TSS). Based on the expression patterns of binding linked genes from FLDN1 to FLDN2b, each group was then divided into 4 subgroups: downregulated gene sites (linked to genes ≥2x downregulated, blue), upregulated gene sites (linked to genes ≥2x upregulated, red), stably expressed gene sites (linked to genes with <2x change in expression, green) and silent gene sites (linked to genes with <1 RPKM in both stages, black). Binding sites linked to downreulated genes, both distal and proximal, tend to lose PU.1 occupancy more rapidly than other groups of sites. K-S test was performed between stably expressed gene sites and each of the other three subgroups. The number of sites in each group and p value for each comparison are indicated in parentheses.

B. Cumulative distributions of changes in PU.1 binding associated H3K4me2 between FLDN2b and FLDN1 among promoter-proximal sites (top) and promoter-distal sites (bottom). H3K4me2 signal densities were calculated within ±1 kb of the summit of a given PU.1 bound region (as depicted in Figure 5B). K-S test was performed between stably expressed gene sites and each of

the other three subgroups as described in S6A. The number of sites in each group and p value for each comparison are indicated in parentheses.

C1&2. Heat maps of PU.1 occupancy and distribution of H3Ac, H3K4me2 and H3K27me3 surrounding ±2 kb of the binding summits for promoter-proximal regions (see experimental procedures). The PU.1 binding sites in DN cells were divided into four subgroups based on linked gene expression patterns as described in S6A. As comparison, a separate group of heat maps for promoter-proximal regions that were selected for much greater PU.1 binding in $E2A^{-/-}$ cells than in early T-lineage cells (as in Figure 5) are included.

D. Heat maps of PU.1 occupancy at promoter-distal regions are shown as in C1&2, correlated with distribution of H3Ac, H3K4me2 and H3K27me3 surrounding ±2 kb of the binding summits (see Experimental Procedures). RNA expression heat maps refer to the nearest linked gene and a single gene can be represented by more than one PU.1-bound distal region. The PU.1 binding sites in DN cells were divided into four subgroups as described in panel A. As comparison, a separate group of heat maps for promoter-distal regions that were specific for PU.1 binding in $E2A^{-/-}$ cells (as in Figure 5) are included.

Figure S7, related to Fig. 7. Characterization of sites of GATA-3 binding in early developing T cells.

A. Comparisons of GATA-3 DNA binding site distributions in ThyDP (Tcrα-/- DP, using Santa Cruz sc-268 antibody) and $CD3^{lo}$ DP (using BD biosciences #558686 antibody) (Wei et al. 2011). Pearson's correlation coefficient (r) is indicated.

B. Cumulative distributions of H3K4me2 enrichment over genomic regions within ±1 kb of PU.1 binding sites and GATA-3 binding sites in FLDN1 (top panels) and FLDN2b (bottom panels)

cells. In each stage, positive binding sites (≥2 RPM of PU.1 or GATA-3 enrichment, see table S5 or S7 respectively) were divided into two groups, promoter-distal sites and promoter-proximal sites. Each group of binding sites was further divided into two subgroups based on the expression level of binding associated genes (expressed and silent). Since PU.1 tends to bind at multiple sites of a single gene locus, it is possible that binding sites with low or no histone modifications of a particular gene locus are nonfunctional. PU.1 binding sites with the highest H3K4me2 enrichment at each gene locus were selected (sites in promoter-distal regions and promoter-proximal regions were selected separately), and plotted accordingly.

C. Cumulative distributions of H3K27me3 enrichment over genomic regions within ±1 kb of PU.1 binding sites and GATA-3 binding sites in FLDN1 (top panels) and FLDN2b (bottom panels) cells. In each stage, positive binding sites were divided similarly as in Figure S7B.

Legends to Supplementary Tables 1-7: included in Supplementary Tables**.**

TABLE 1. Lineage commitment status measurements for FLDN cells differentiated in vitro

TABLE 2. Genome-wide quantitation of histone modifications and gene expression

TABLE 3. Dynamic expression pattern of transcription factors during T-lineage commitment

TABLE 4. K-means clustering for differentially expressed genes and histone marking and transcriptome profiles of hematopoiesis genes

TABLE 5. Genome-wide PU.1 binding from FLDN1 to FLDN2b

TABLE 6: Distribution of PU.1 binding sites among differentially expressed gene clusters

TABLE 7. Genome-wide GATA-3 binding in FLDN1, FLDN2b and ThyDP

Complete Materials and Experimental Procedures

Cell Culture

Fetal liver (FL) cells from embryonic day 13.5 to 14 (E13.5-E14) C57BL/6 mouse embryos were first depleted of Gr-1$^+$, F4/80$^+$, Ter119$^+$, and CD19$^+$ ("Lin cocktail 1") cells using streptavidin-coupled magnetic microbeads (Miltenyi Biotec) against biotin-conjugated antibodies. Lin$^-$c-Kit$^+$CD27$^+$ multi-lineage precursors were then sorted from lineage-depleted FL cells by FACS and co-cultured with OP9-DL1 stromal cells as described previously (Taghon et al., 2005). 50~100 × 10^3 Lin$^-$c-Kit$^+$CD27$^+$ FL cells were plated on OP9-DL1 monolayers in 10 cm plates in the presence of 5 ng/mL Flt3-L and 5 ng/mL IL-7 (both from Peprotech). After 4.5 d of culture, half of the cells were harvested and sorted to isolate DN1 (FLDN1, Lin$^-$c-Kit^{hi}CD45$^+$CD44$^+$CD25$^-$) and DN2a cells (FLDN2a, Lin$^-$c-Kit^{hi}CD45$^+$CD44$^+$CD25$^+$) (using "Lin cocktail 2" = antibodies to Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11c, TCRγδ, TCRβ, CD3ε, CD8α). After 8.5 d of culture, the rest of the cells were harvested and sorted for FLDN2a and DN2b (FLDN2b, Lin$^-$c-Kit^{int}CD45$^+$CD44^{int}CD25$^+$). The FLDN2a samples used for analysis were each pools of day 4.5 and day 8.5 DN2a cells in approximately 2:1 ratio. For subsets from adult (4-6 weeks old) thymus, wild-type C57BL/6 mouse thymi were first depleted with antibodies to Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, CD11c, TCRγδ, TCRβ, CD3ε, CD4 and CD8α ("Lin cocktail 3"). Thymic DN3 (ThyDN3) cells were then sorted from lineage-depleted thymocytes as Lin$^-$c-Kit$^-$CD44$^-$CD25$^+$. Finally, to prepare ThyDP populations free of contaminating cells in early stages of TCR-dependent positive selection, while

maintaining viability of these fragile cells, ThyDP cells were collected from TCR$\alpha^{-/-}$ thymi (4-6 weeks old, The Jackson Laboratory, B6.129S2-Tcra^{tm1Mom}/J) by simple streptavidin-coupled magnetic microbead depletion with biotinylated antibodies against Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, CD11c, c-Kit, CD44 and CD25. After this depletion, 90~95% of cells were $CD4^+CD8^+$.

Antibodies used were from eBioscience and Biolegend, including anti-CD4 (GK1.5; biotin), anti-CD8α (53-6.7; biotin), anti-CD11b (M1/70; biotin), anti-CD11c (N418; biotin), anti-CD19 (eBio1D3; biotin), anti-CD122 (5H4; biotin), anti-Gr1 (RB6-8C5; biotin), anti-F4/80 (BM8; biotin), anti-TCRβ (H57-597; biotin), anti-TCRγδ (eBioGL3; biotin), anti-NK1.1 (PK136; biotin), anti-Ter119 (Ter-119; biotin), anti-c-Kit (2B8; PE, APC, biotin), anti-CD27 (LG.7F9; APC), anti-CD25 (PC61.5; APC-Alexa 750, APC-Alexa 780, biotin), anti-CD44 (1M7; Pacific Blue, eFluor 450, biotin), anti-CD45 (30-F11; APC), anti-CD3ε (145-2c11, PerCp-cy5.5). For detection of biotinylated antibodies, streptavidin–PerCp-Cy5.5 was used.

Lineage Commitment Assay

Samples of 25 FLDN1, FLDN2a or FLDN2b cells were each sorted into 96 well plates coated with either OP9-DL1 or OP9-Mig (control) monolayers in the presence of 5 ng/mL Flt3-L, 5 ng/mL IL-7, 5 ng/mL SCF, and either 200 units/mL IL2 or 5 ng/mL MCSF. After 7 days of co-culture, cells were harvested and subjected to FACS analysis, using NK1.1 and/or Dx5 as a marker for NK progeny, CD19 as a marker for B-cell progeny, and CD11b/CD11c as markers for myeloid and dendritic progeny. Results are shown in Table S1.

Chromatin Immunoprecipitation

Each histone modification ChIP was generated using 5 million cells and 20 μg of each of

the following antibodies: H3K(9,14)Ac (Millipore 06-599), H3K4me2 (Millipore 07-030) and H3K27me3 (Millipore 07-449). PU.1 ChIP was generated using 7.5~10 million cells and 5μg anti-PU.1 (Santa Cruz sc-352). GATA-3 ChIP was generated using 15~20 million cells and 2.5 μg anti-GATA-3 (Santa Cruz sc-268). In addition to FLDN1, FLDN2a and 2b samples, we performed PU.1 ChIP on a ThyDP sample, which naturally lacks PU.1 expression, as a negative control. ChIP was carried out essentially following the manufacturer's protocol (http://www.millipore.com/userguides/tech1/mcproto407), with the exceptions that protein A or G agarose beads were replaced by anti-rabbit or anti-mouse secondary antibody-coupled magnetic beads (Dynabeads, pre-washed with 1xPBS/0.5%BSA, 1 μg Ig/10 μL beads), and the pre-clear step was omitted. Independent biological replicates were generated for histone modification ChIP of FLDN1, FLDN2b, ThyDN3 and ThyDP. Different batches of histone modification antibodies (Millipore 06-599, Millipore 07-030, Millipore 07-449) were used among biological replicates. Purified ChIP DNA was subjected to end repairing, adaptor ligation, PCR amplification, size selection by gel electrophoresis (200~300 bp, insert plus adaptor and PCR primer sequences) and a second round of PCR amplification to generate each ChIP DNA library as described (Johnson et al. 2007) (Illumina ChIP-seq sample preparation kit #IP-102-1001).

mRNA Purification and cDNA Library Building

Total RNA was extracted from 2.5~20 million cells using Trizol (Invitrogen), and then subjected to two rounds of selection using Oligo-dT coupled magnetic beads (Dynabeads) according to the manufacturer's protocol. About 100 ng polyadenylated mRNA per sample was obtained after double selection. Independent biological replicates were generated for all five populations (triplicates for FLDN2b). cDNA library building was performed as described

(Mortazavi et al., 2008). Briefly, RNA was fragmented to an average length of 200 bp by Mg^{2+}-catalyzed hydrolysis and then converted into cDNA by random priming. cDNA was then subjected to end repairing, adaptor ligation (using Illumina ChIP-seq sample preparation kit #IP-102-1001), size selection and one round of PCR amplification.

High-Throughput Sequencing

Each ChIP DNA library or cDNA library was sequenced with the Illumina Genome Analyzer II or IIX following the manufacturer's protocols (http://www.illumina.com; Johnson et al., 2007, Mortazavi et al., 2008).

RNA-seq Data Analysis

Sequence reads from each cDNA library (38 bp, single-read) were trimmed to 32 bp long and mapped onto the mouse genome build NCBI37/mm9 using Bowtie (bowtie-0.12.1, http://bowtie-bio.sourceforge.net/index.shtml) with setting '-v 2 -k 11 -m 10 -t --best --strata'. The mappable data were then processed by the ERANGE v. 3.3 RNA-seq analysis program (Mortazavi et al, 2008). Assuming total transcriptional activity is comparable between different cell types, the obtained data (data units in RPKM, reads per kilobase exon model per million mapped reads) were first $\log_2$ transformed and linearly normalized between individual samples, then averaged among biological replicates or triplicates. At the same time, in order to find genes that were changed in expression between two populations to a statistically significant degree, ERANGE processed data were analyzed by the Bioconductor DEGseq program ((Wang et al., 2010) http://www.bioconductor.org/packages/2.6/bioc/html/DEGseq.html) (data units in RPM, reads per million mapped reads, method = "MARS", $p < 0.001$) (Figure 1A). This analysis yielded 3,697 DEGseq positive genes that had more than a twofold change in RNA-seq reads (after normalization and averaging), either between any two successive stages or between

FLDN1 and ThyDP, and these were defined as differentially expressed genes. To identify differentially regulated transcription factors, we did Gene Ontology analysis of this set with key term "DNA-dependent regulation of transcription" (GO:0006350), and the resulting list was then hand curated to remove cell surface receptors, cytokines, and other genes of questionable categorization. The final list used for alignment against our DEGseq set is presented as Table S3 part A.

Hierarchical clustering: To determine the overall tendencies of change in gene expression and the connection between different populations, we hierarchically clustered RNA-seq data of these 3,697 selected genes from all 11 samples (using normalized data, biological replicates and triplicates were treated independently, Figure 1B). Hierarchical clustering was performed along both dimensions with sample similarities clustered first, and then genes. Euclidean distance and complete linkage were used (MATLAB 7.10.0). Separately, two-dimensional hierarchical clustering was also performed on 379 differentially expressed transcription factors (Figure 1C).

K-means clustering: To profile and categorize the behavior of clusters of similarly regulated genes during early T cell development, we first normalized individual mRNA data for the 3,697 selected genes by the corresponding geometric mean of five stages, and then performed K-means clustering analysis on the results after $\log_2$ transformation (Figure S5). K was set at 25 and squared Euclidean distance was used (MATLAB 7.10.0).

ChIP-seq Data Analysis

Histone Modification ChIP-seq: DNA sequence reads from each ChIP-seq library (single-read) were trimmed and mapped onto NCBI37/mm9 using the same setting as for RNA-seq data, and uniquely mapped reads were used for further analysis. The data were processed by the ERANGE

v. 3.3 findall peak finder (Johnson et al. 2007) to identify enriched genomic regions. We used a stringent setting of '-spacing 100 -minimum 4 -ratio 4 -minPeak 0.5 -shift learn' for H3Ac and H3K4me2 ChIP-seq data, and a relatively less stringent setting of '-notrim -nodirectionality -spacing 100 -minimum 2 -ratio 4 -minPeak 0.25 -shift learn' for H3K27me3. The sequence data of the input DNA from the same cell type were used as background control. Since on average the total amount of mappable DNA reads of each H3K27me3 ChIP-seq data was about two times of that of each H3Ac and H3K4me2 ChIP-seq data, the minimum total DNA reads for called regions were comparable for all three histone modifications (that is, about minimum 60 to 80 enriched DNA reads per region).

All called regions (from all 27 samples) were pooled and merged if overlapping. Only resulting regions of at least 200 bp were considered for further analysis. This conservative approach treats any local change in peak height or spreading of histone modification as effects on a single region, thus providing a minimum estimate of the number of centers of regulatory change. Thus, for example, the change in shape factor of H3K4me2 commonly observed at active promoters is not considered to change the number of marked regions. We considered the positive regions overlapping ±1 kb from the TSSs of UCSC known genes (mm9, NCBI v.37) as promoter-proximal regions, and the rest as promoter-distal regions. Individual regions were then assigned to the nearest genes using ERANGE (200 kb as the maximum radius). Signal densities (number of DNA reads) were calculated using ERANGE v. 3.3 regionCounts, for each region of every histone modification dataset. For global histone modification status of promoter regions, we expanded every transcriptional starting site (TSS) of UCSC known genes to a window of ±1 kb, and calculated signal densities of each TSS regions using ERANGE. Assuming that total DNA enrichment of the same histone marker is comparable among different cell types, we

linearly normalized the read number (after $\log_2$ transformation) between samples from the same histone marker (i.e., based on slopes of correlation plots in Figure S1B). The mean for biological replicates was used for analysis. Since our RNA-seq data cannot accurately distinguish among isoforms, for genes that have multiple alternative promoters we selected one promoter that had the highest H3K4me2 level (or H3Ac if all had the same level of H3K4me2). Regions (both distal and promoter regions) that had more than 4 RPM in either H3Ac or H3K4me2, or more than 2 RPM in H3K27me3 were considered as positive for the particular histone modification(s) (Figures 2, S2). The processed data were plotted and visualized in MATLAB.

All RNA-seq and ChIP-seq sequencing tracks were generated in WIG file format and uploaded onto the UCSC genome browser for visualization. Publicly available data used in this study (Lin et al., 2010; Heinz et al., 2010; Wei et al., 2009) were downloaded as raw sequence data (http://www.ncbi.nlm.nih.gov/geo) and remapped onto NCBI37/mm9 using the same settings.

PU.1 and GATA-3 ChIP-seq: Since PU.1 ChIP enriched genomic regions were in general narrower than histone modification enriched regions, we used a setting of "-spacing 50 -minimum 2 -ratio 4 -minPeak 0.5 -shift learn -listPeak" for the ERANGE findall peak finder. The sequence data of the input DNA from the same cell type was used as background control. Publicly available PU.1 ChIP-seq and input data from $E2A^{-/-}$ pre-pro B cells, mature B cells and macrophages (Heinz et al., 2010) were downloaded as raw sequence data (http://www.ncbi.nlm.nih.gov/geo) and remapped using the same setting.

Called regions were pooled and merged if overlapping from each pair-wise or three-way comparison (from $E2A^{-/-}$ pre-pro B vs. FLDN1, B cell vs. FLDN1, macrophage vs. FLDN1, or FLDN1 vs. FLDN2a vs. FLDN2b). Individual regions were calculated for PU.1 enriched signal

densities and then assigned to the nearest genes (200 kb as the radius) using ERANGE. We next aligned the summits of all positive regions and calculated histone modification signal densities in a window of ±1 kb. All histone modification data were linearly normalized (using the parameters generated from global histone modification analysis). The mean for biological replicates was used for analysis. Scatter plots were generated and visualized in MATLAB.

To compare differential PU.1 binding with associated differential gene expression and H3K4me2 enrichment during early T cell development, we divided PU.1 binding linked genes into four subgroups: upregulated and downregulated genes (selected from the differentially expressed genes group and having more than 2-fold change in expression from FLDN1 to FLDN2b; see "RNA-seq Data Analysis"), stably expressed genes (less than 2-fold change in expression between FLDN1 and FLDN2b), and silent genes (<1 RPKM in both stages). The changes in PU.1 occupancy and in H3K4me2 enrichment (within ±1 kb of binding summits) between FLDN2b and FLDN1 were calculated and plotted separately as cumulative distribution for each group (Figure S6 A&B). To determine whether PU.1 binding sites linked to upregulated or downregulated genes were more likely differentially bound by PU.1 and enriched by H3K4me2 compared to sites linked to stably expressed genes, two-sample Kolmogorov-Smirnov test was performed between stably expressed gene sites and each of the other three subgroups (Figure S6 A&B).

To visualize histone modifications and degree of PU.1 occupancy surrounding the summits, we further expanded positive regions to a window of ±2 kb, and divided each window into 50 bins (80 bp each). Histone modification and PU.1 enrichment were calculated for each bin using the same method mentioned above. The data obtained were aligned with RNA-seq data of associated genes, and then hierarchically clustered (one dimensional clustering of binding

regions; using Euclidean distance and Ward linkage) and visualized as heat maps in MATLAB as shown in Figure S6C-D.

GATA-3 ChIP-seq data was processed similarly to PU.1 ChIP-seq data. To compare our findings with published results, raw sequence data for "CD3lo DP" cell samples (Wei et al., 2011) were downloaded from Gene Expression Omnibus and remapped using the same settings as used for our data, as described for comparing PU.1 results (Heinz et al., 2010) above.

De Novo Motif Analysis

We selected the top 1,000 PU.1 enriched peaks from each of the three subgroups ($E2A^{-/-}$ pre-pro B cells high, shared, and FLDN1 high), and performed MEME analysis on regions ±50 bp from the peaks by ERANGE v. 3.3 using the default setting to generate the position specific frequency matrix (PSFM) representation of the motifs. The PSFMs were mapped separately back to the three enriched regions subgroups at 85% match (Johnson et al. 2007).

All 1,652 enriched GATA-3 regions (pooled from FLDN1, FLDN2b and ThyDP) were subjected to MEME analysis. Since the consensus sequence motifs of GATA-3 binding sites were shorter than the ones of PU.1 binding sites, the PSFMs were mapped back to the 1,652 enriched GATA-3 regions and 1,652 random genomic regions at 90% match instead. Random genomic regions were comparable to the GATA-3 binding regions in both length and chromosomal distribution.

Supplemental References:

David-Fung, E.S., Yui, M.A., Morales, M., Wang, H., Taghon, T., Diamond, R.A., and Rothenberg, E.V. (2006). Progression of regulatory gene expression states in fetal and adult pro-T cell development. Immunol Rev. *209*, 212-236.

Heinzel, K., Benz, C., Martins, V.C., Haidl, I.D., and Bleul, C.C. (2007). Bone marrow-derived hemopoietic precursors commit to the T cell lineage only after arrival in the thymic microenvironment. J. Immunol. *178*, 858-868.

Masuda, K., Kakugawa, K., Nakayama, T., Minato, M., Katsura, Y., and Kawamoto, H. (2007). T cell lineage determination precedes the initiation of TCRb rearrangement. J. Immunol. *179*, 3699-3706.

Taghon,T.N., David,E.-S., Zúñiga-Pflücker,J.C., and Rothenberg,E.V. (2005). Delayed, asynchronous, and reversible T-lineage specification induced by Notch/Delta signaling. Genes Dev. 19, 965-978.

Wang, L., Feng, Z., Wang, X., Wang, X., Zhang, X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. *26*, 136-8.

Supplemental Figure 1

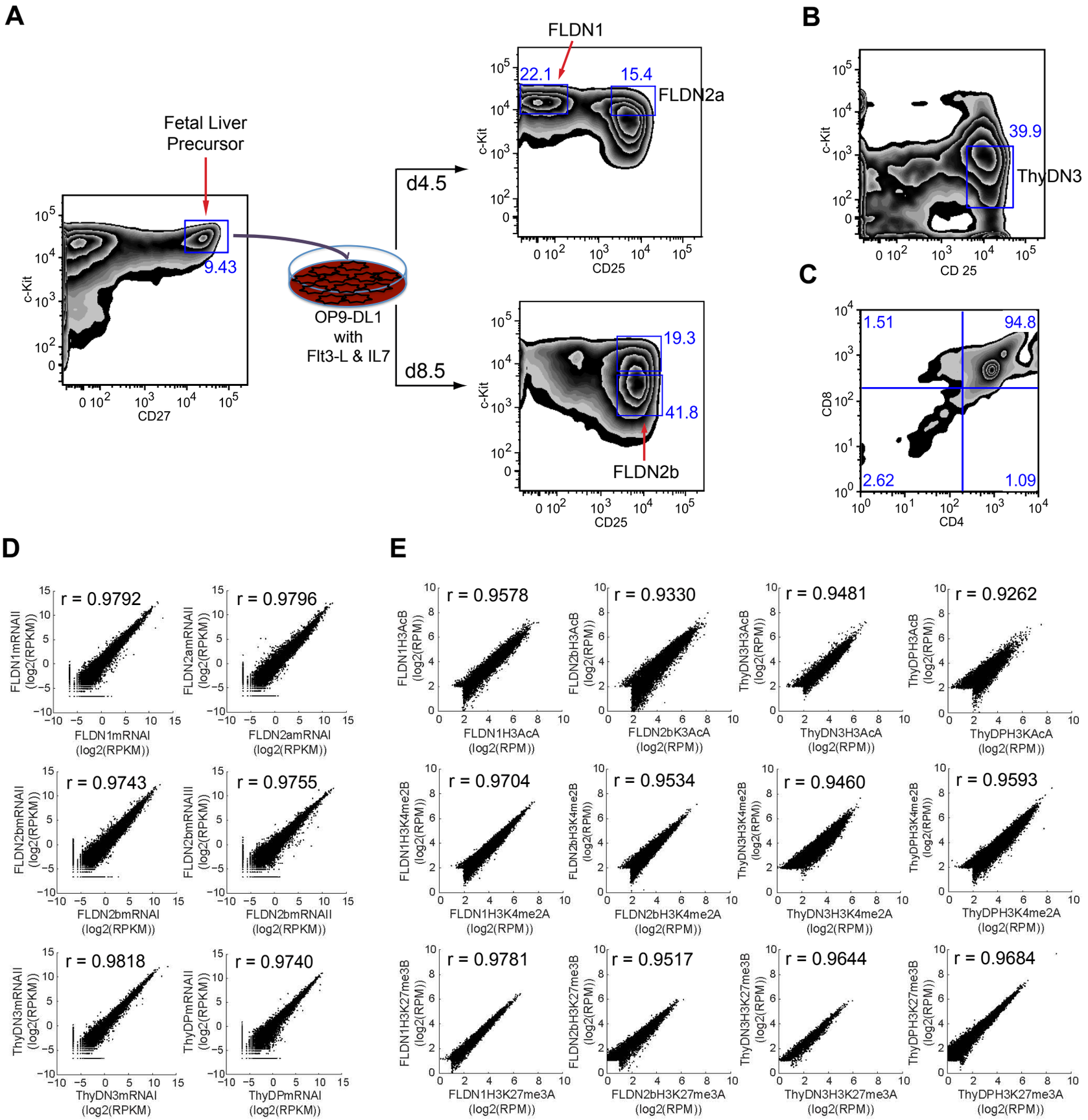
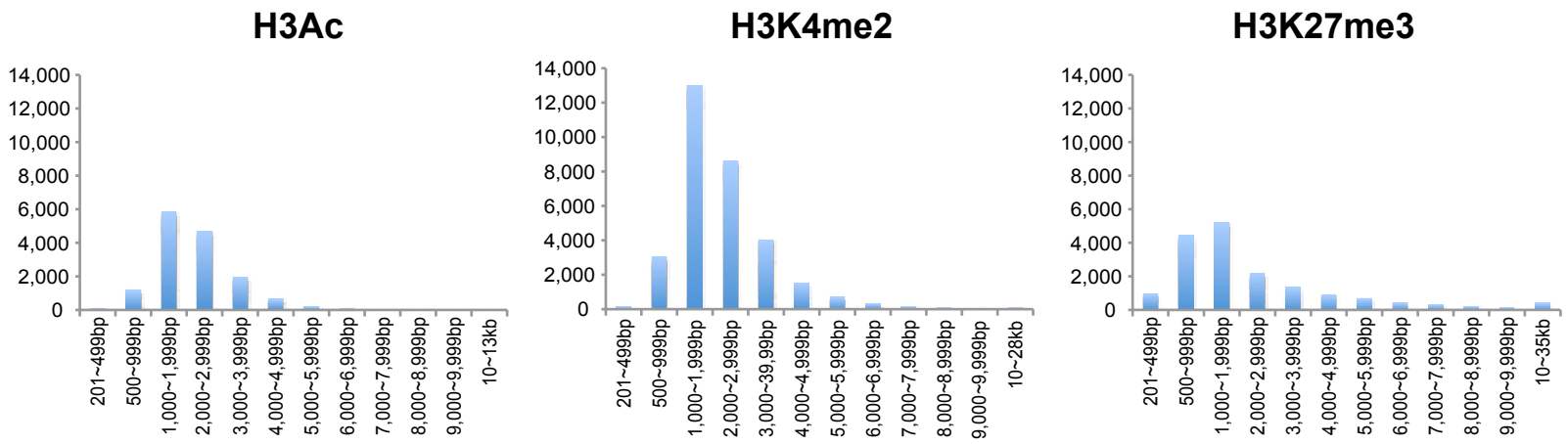


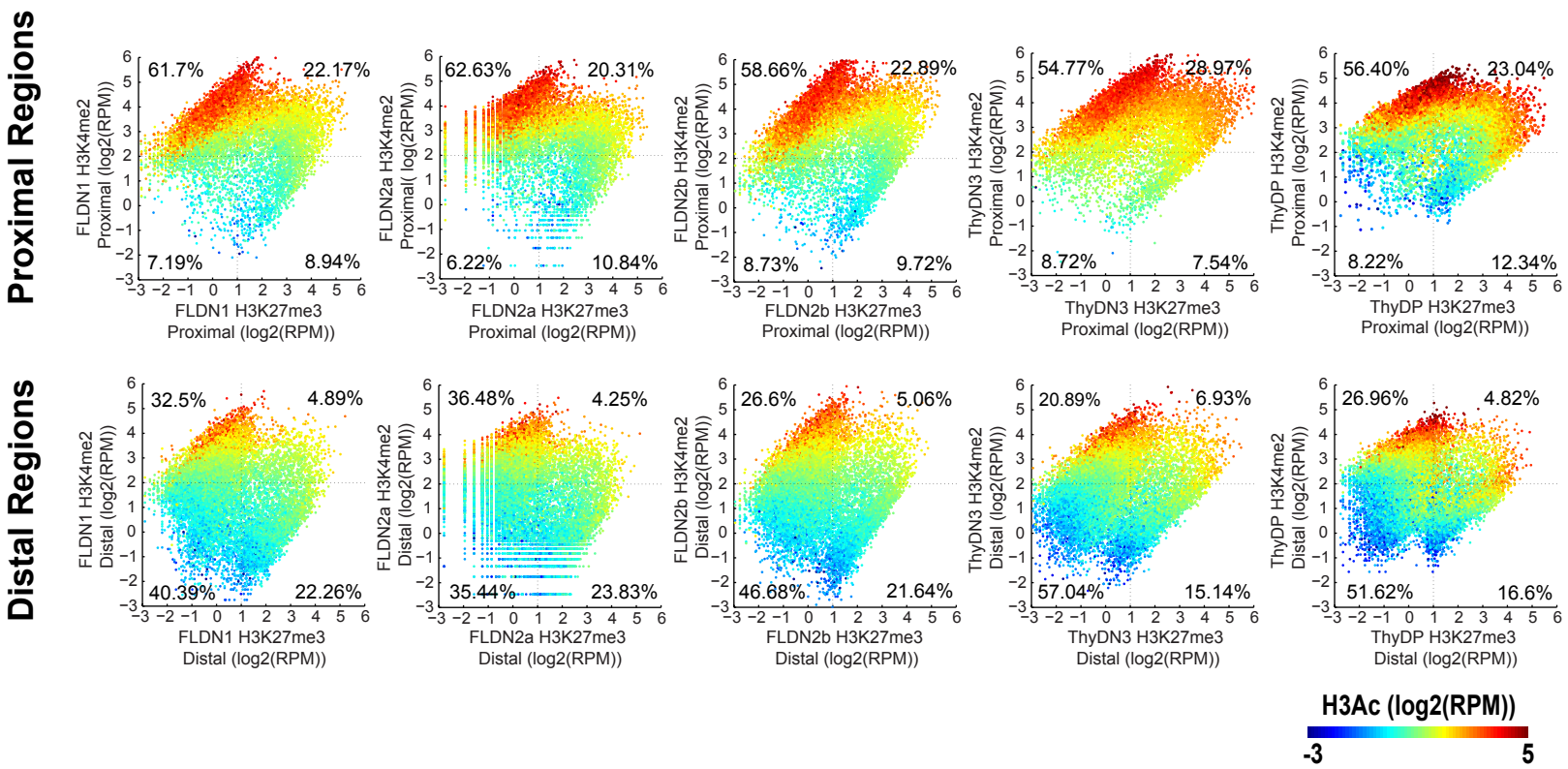
Fig. S1

Supplemental Figure 2

A



B



C

	FLDN1 (9,878*)	FLDN2a (9,531*)	FLDN2b (9,875*)	ThyDN3 (9,705*)	ThyDP (9,501*)
H3KAc+ (≥4 RPM)	8,469 (86%)	8,406 (88%)	8,603 (87%)	8,357 (86%)	7,948 (84%)
H3K4me2+ (≥4 RPM)	9,104 (92%)	8,913 (94%)	9,002 (91%)	8,963 (92%)	8,758 (92%)
H3K27me3+ (≥2 RPM)	366 (3.7%)	296 (3.1%)	437 (4.4%)	371 (3.8%)	494 (5.2%)
H3Ac-/ H3K4me2+/ H3K27me3+	137 (1.4%)	105 (1.1%)	106 (1.1%)	92 (0.95%)	121 (1.3%)
H3Ac+/ H3K4me2+/ H3K27me3-	8,261 (84%)	8,234 (86%)	8,274 (84%)	8,094 (83%)	7,612 (80%)
H3Ac+/ H3K4me2+/ H3K27me3+	179 (1.8%)	159 (1.7%)	274 (2.8%)	256 (2.6%)	332 (3.5%)
H3Ac-/ H3K4me2-/ H3K27me3-	695 (7.0%)	573 (6.0%)	761 (7.7%)	712 (7.3%)	699 (7.4%)

* Total number of expressed genes (mRNA ≥ 1 RPKM) in each stage

	FLDN1 (10,983*)	FLDN2a (11,330*)	FLDN2b (10,986*)	ThyDN3 (11,156*)	ThyDP (11,360*)
H3Ac+ (≥4 RPM)	850 (7.7%)	990 (8.7%)	1006 (9.2%)	1064 (9.5%)	841 (7.4%)
H3K4me2+ (≥4 RPM)	3,180 (29%)	3,177 (28%)	2,887 (26%)	3,256 (29%)	2,844 (25%)
H3K27me3+ (≥2 RPM)	4,134 (38%)	4,320 (38%)	4,199 (38%)	3,733 (33%)	3,734 (33%)
H3Ac-/ H3K4me2+/ H3K27me3+	1,829 (17%)	1,630 (14%)	1,521 (14%)	1,690 (15%)	1,495 (13%)
H3Ac+/ H3K4me2+/ H3K27me3-	647 (5.9%)	752 (6.6%)	607 (5.5%)	631 (5.7%)	685 (6.0%)
H3Ac+/ H3K4me2+/ H3K27me3+	200 (1.8%)	234 (2.1%)	399 (3.6%)	433 (3.9%)	299 (2.6%)
H3Ac-/ H3K4me2-/ H3K27me3-	5,695 (52%)	5,693 (50%)	5,820 (53%)	6,290 (56%)	6,574 (58%)

* Total number of silent genes (mRNA < 1 RPKM) in each stage

Fig. S2

Supplemental Figure 3

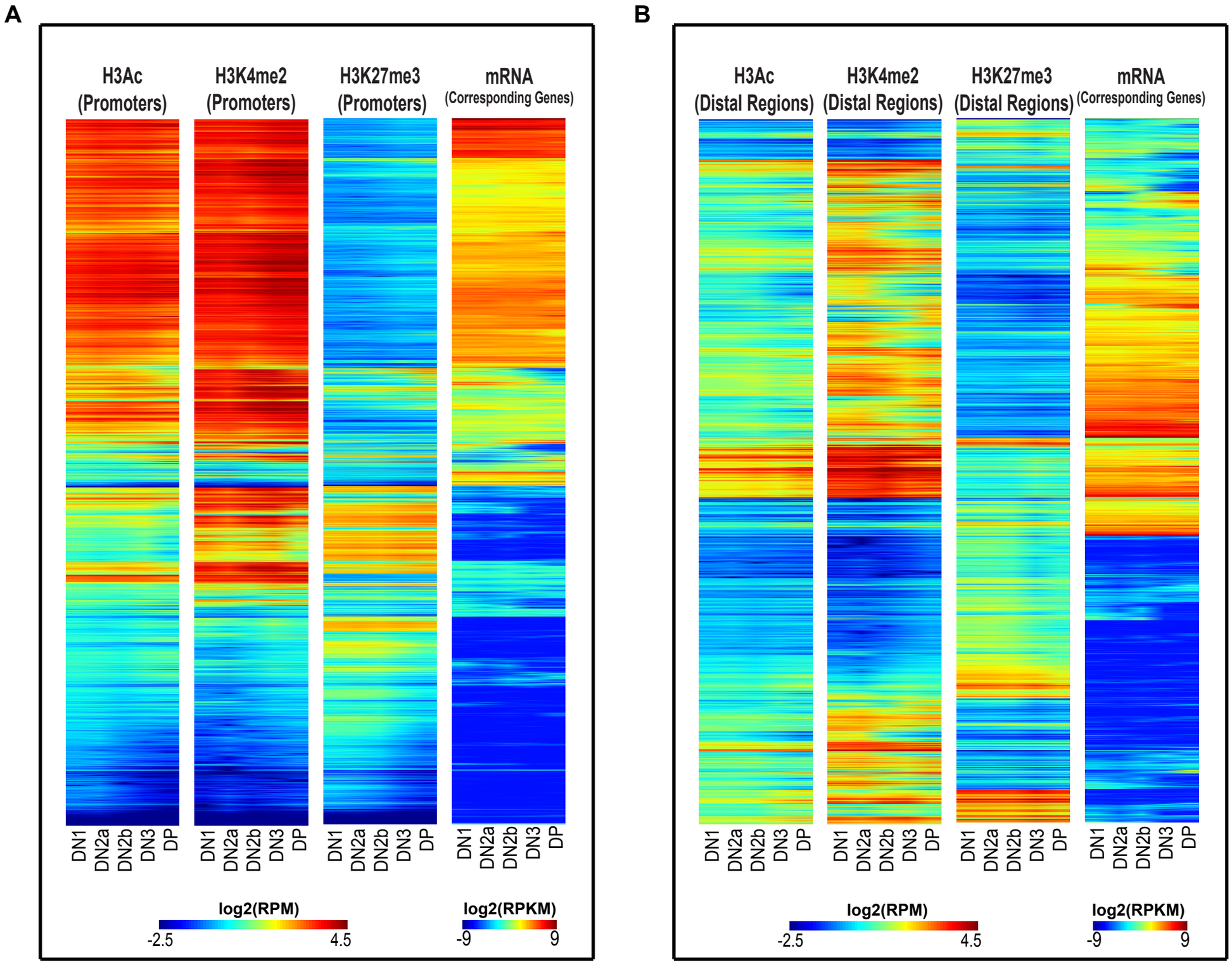


Fig. S3

Supplemental Figure 4

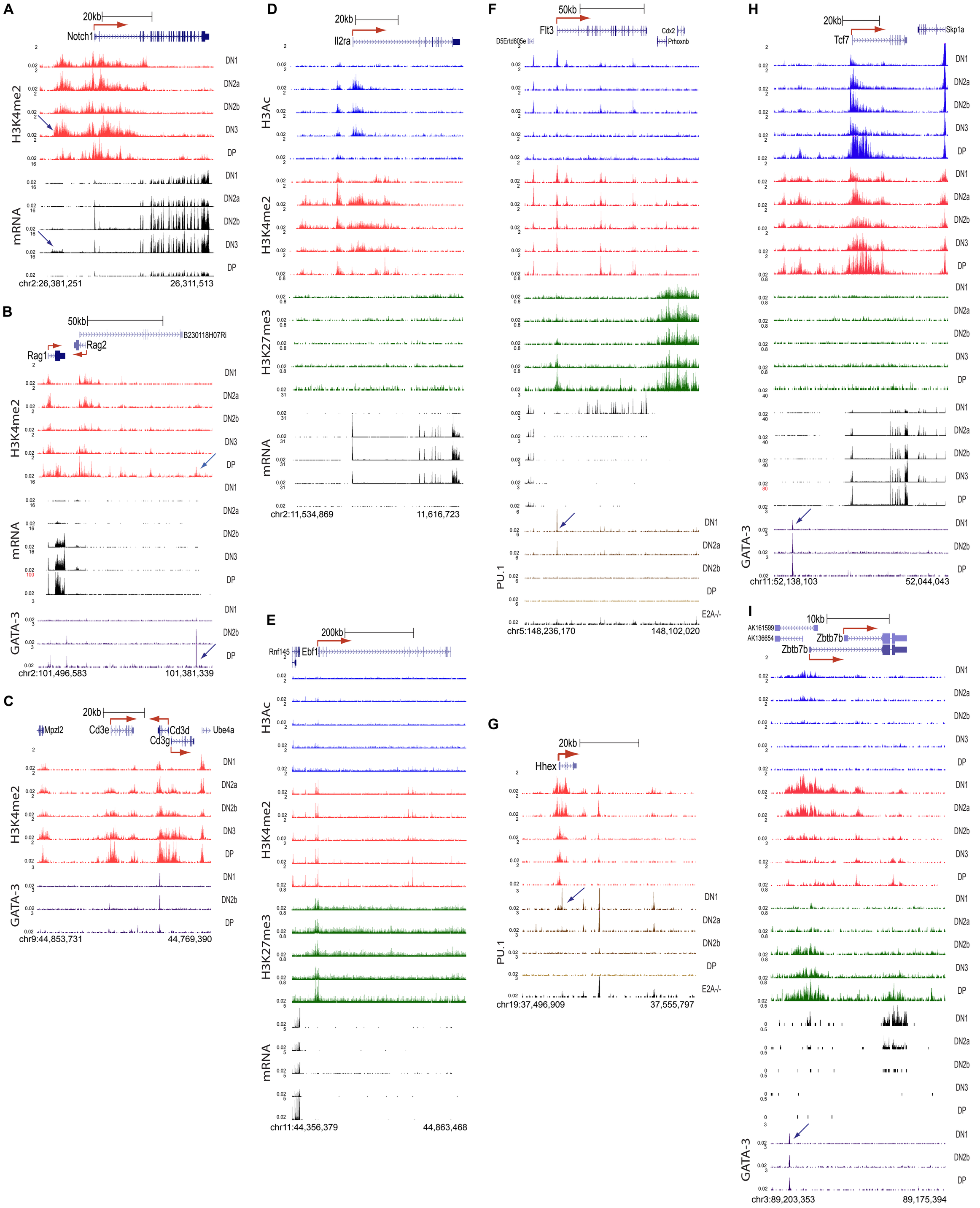


Fig. S4

Supplemental Figure 5

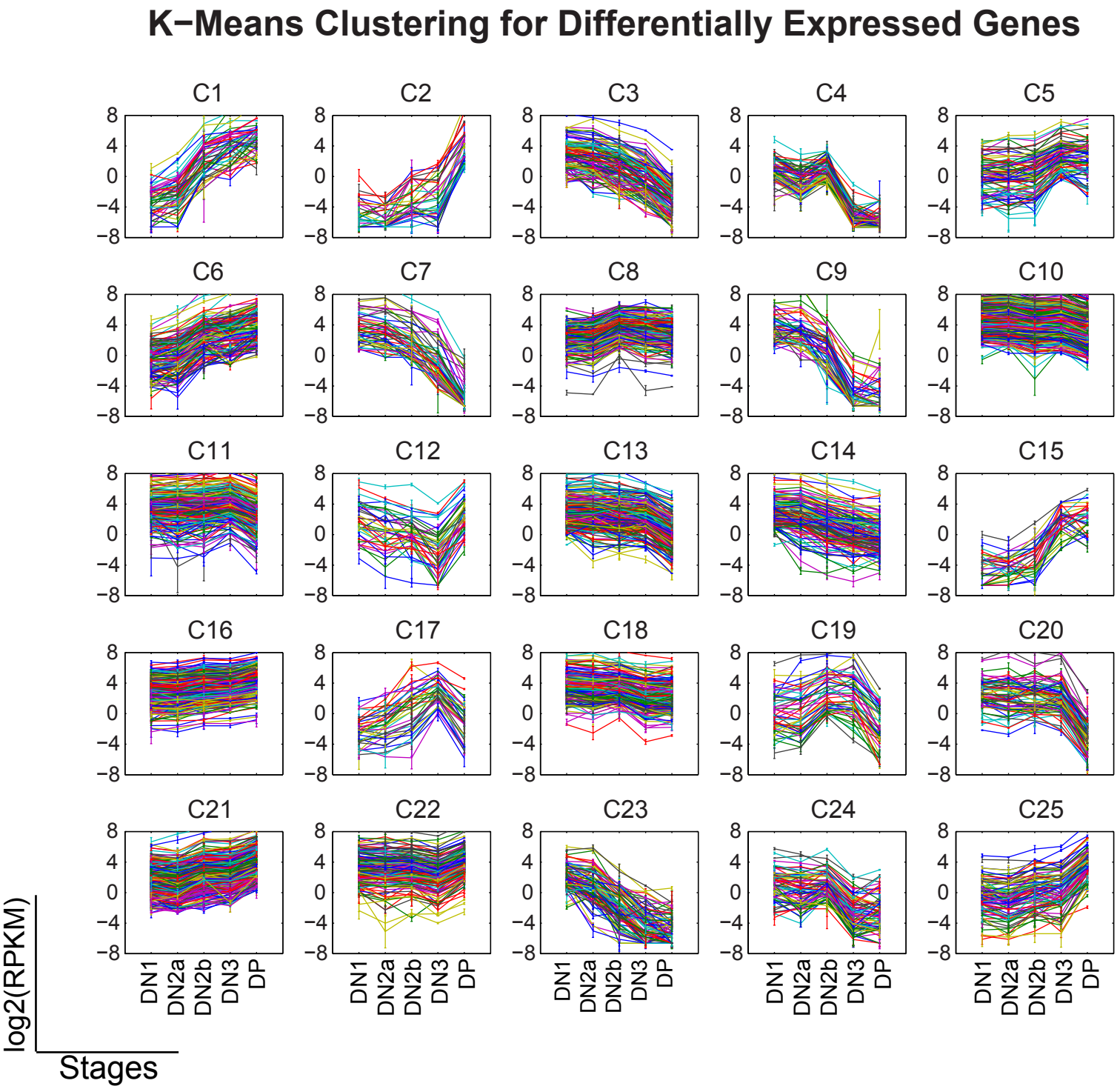
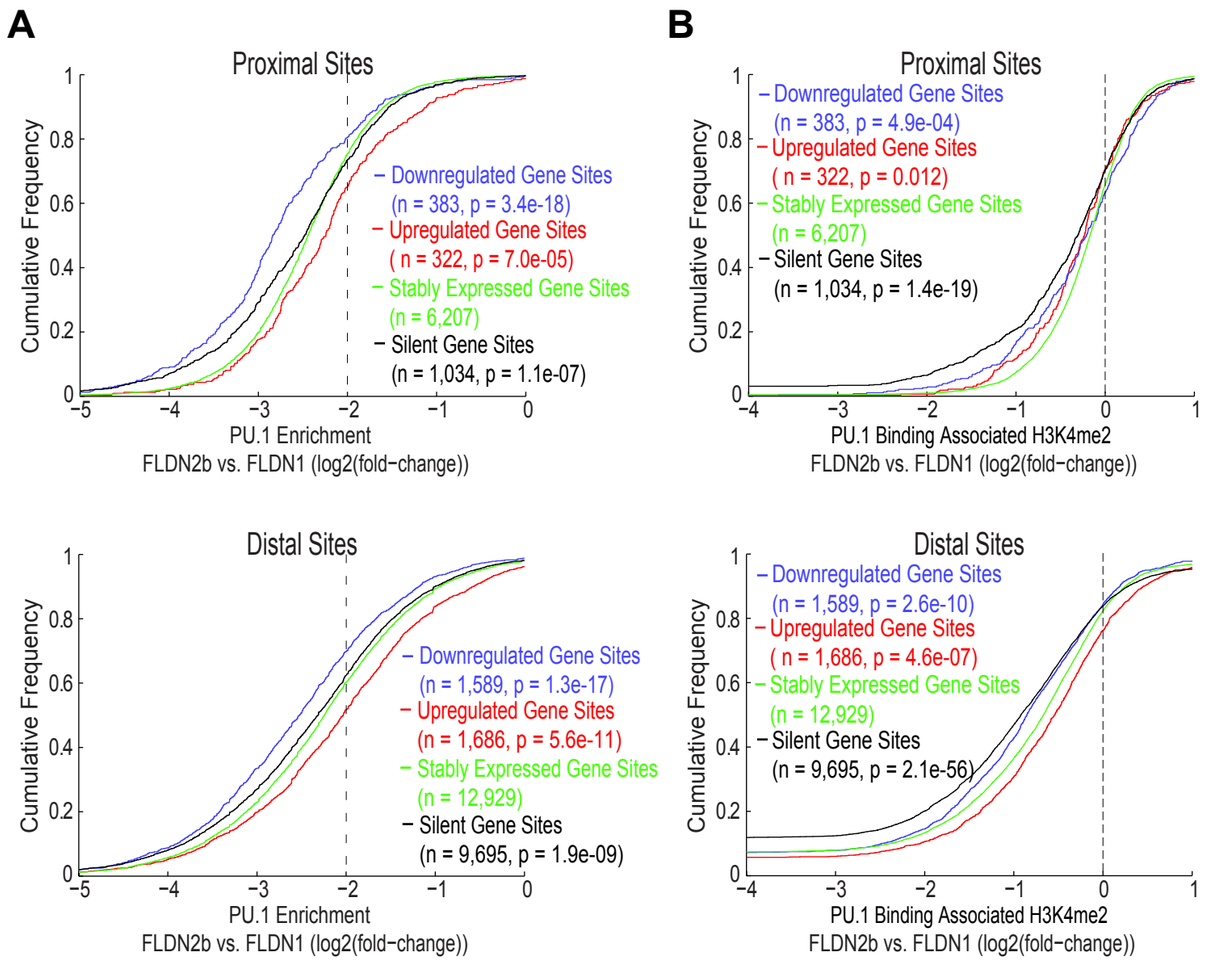
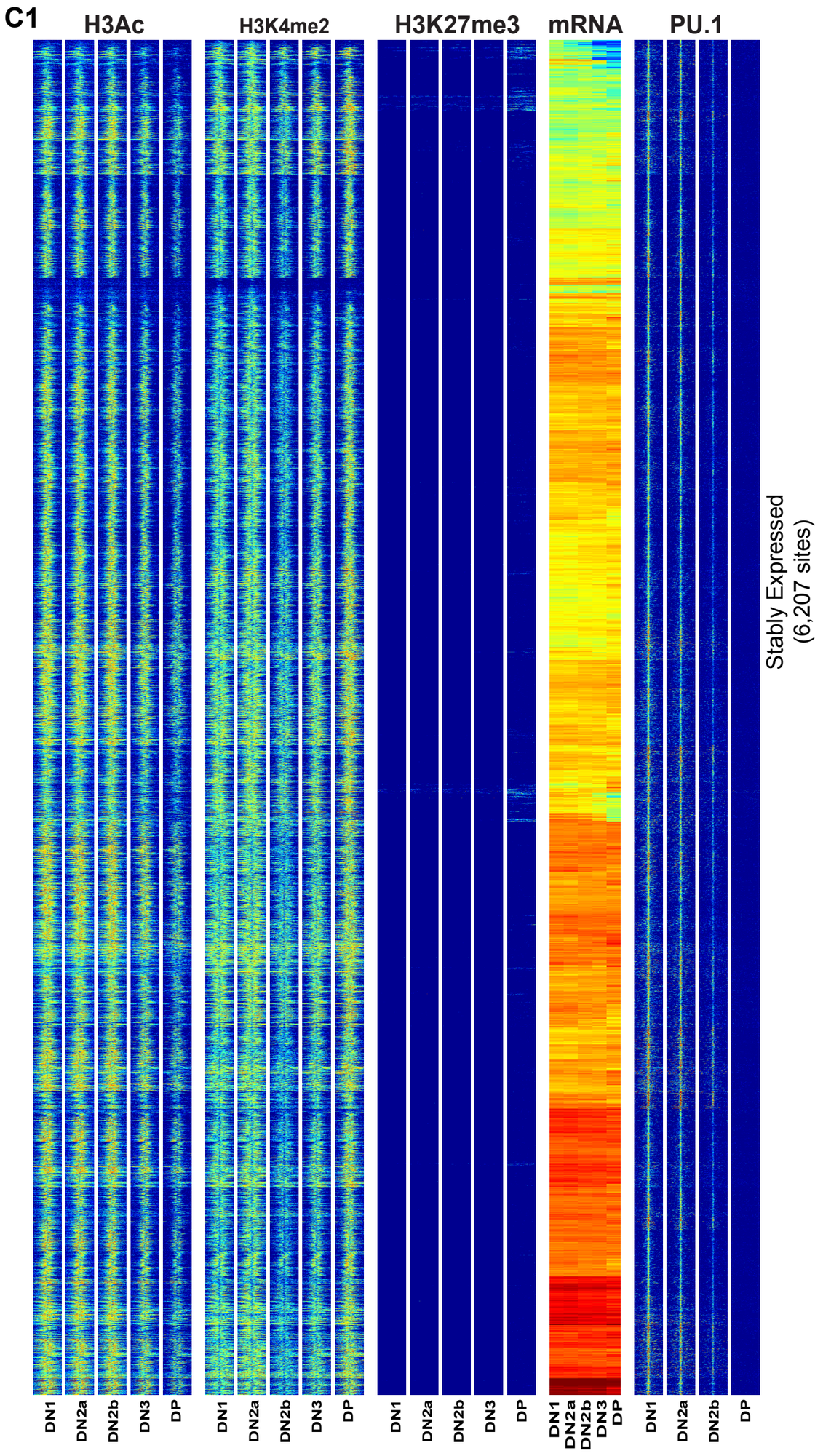
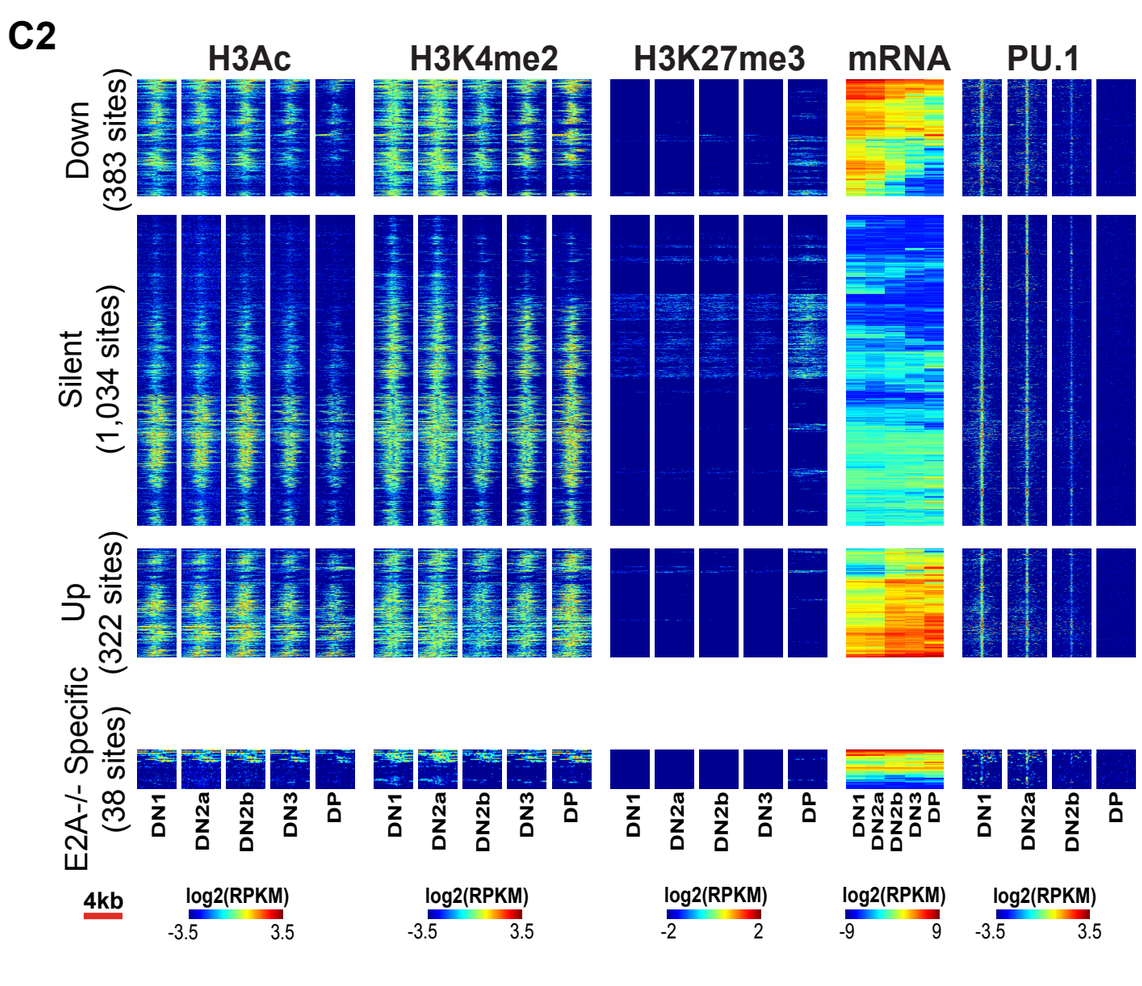


Fig. S5

Supplemental Figure 6







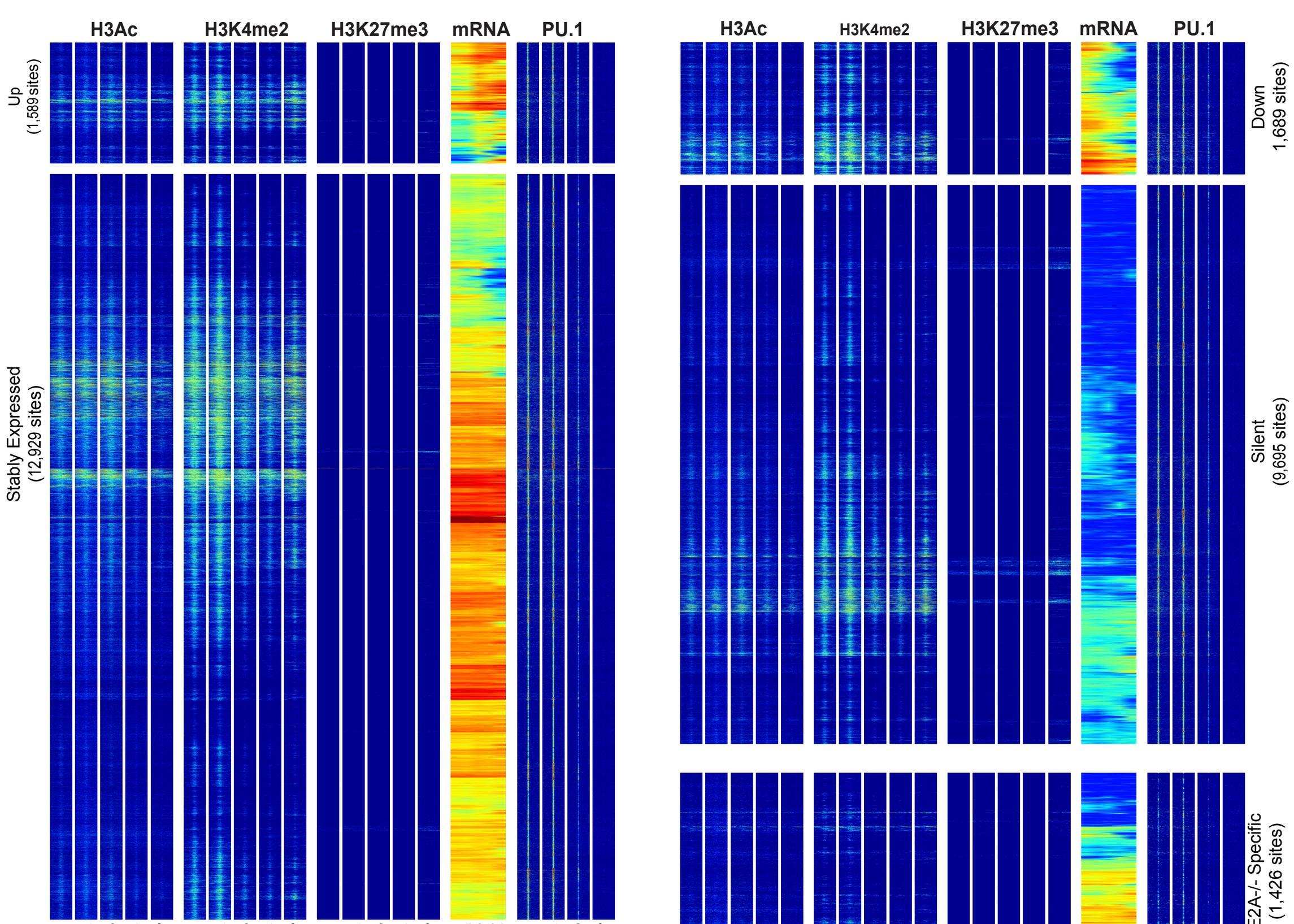
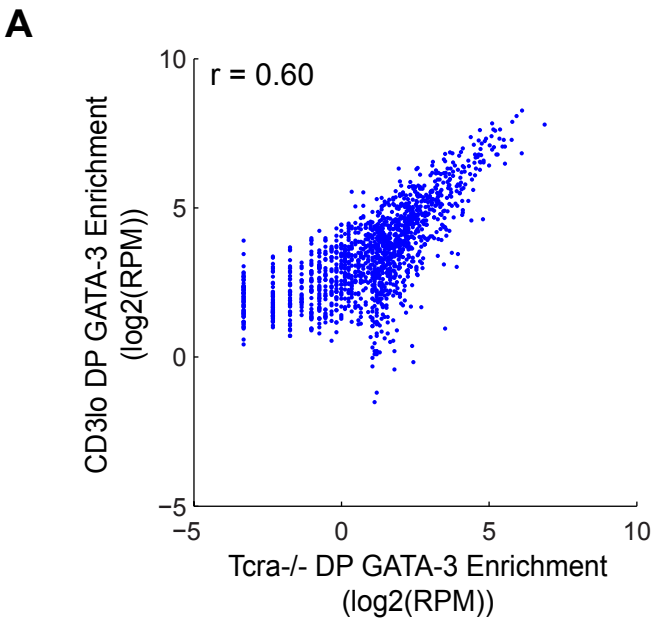
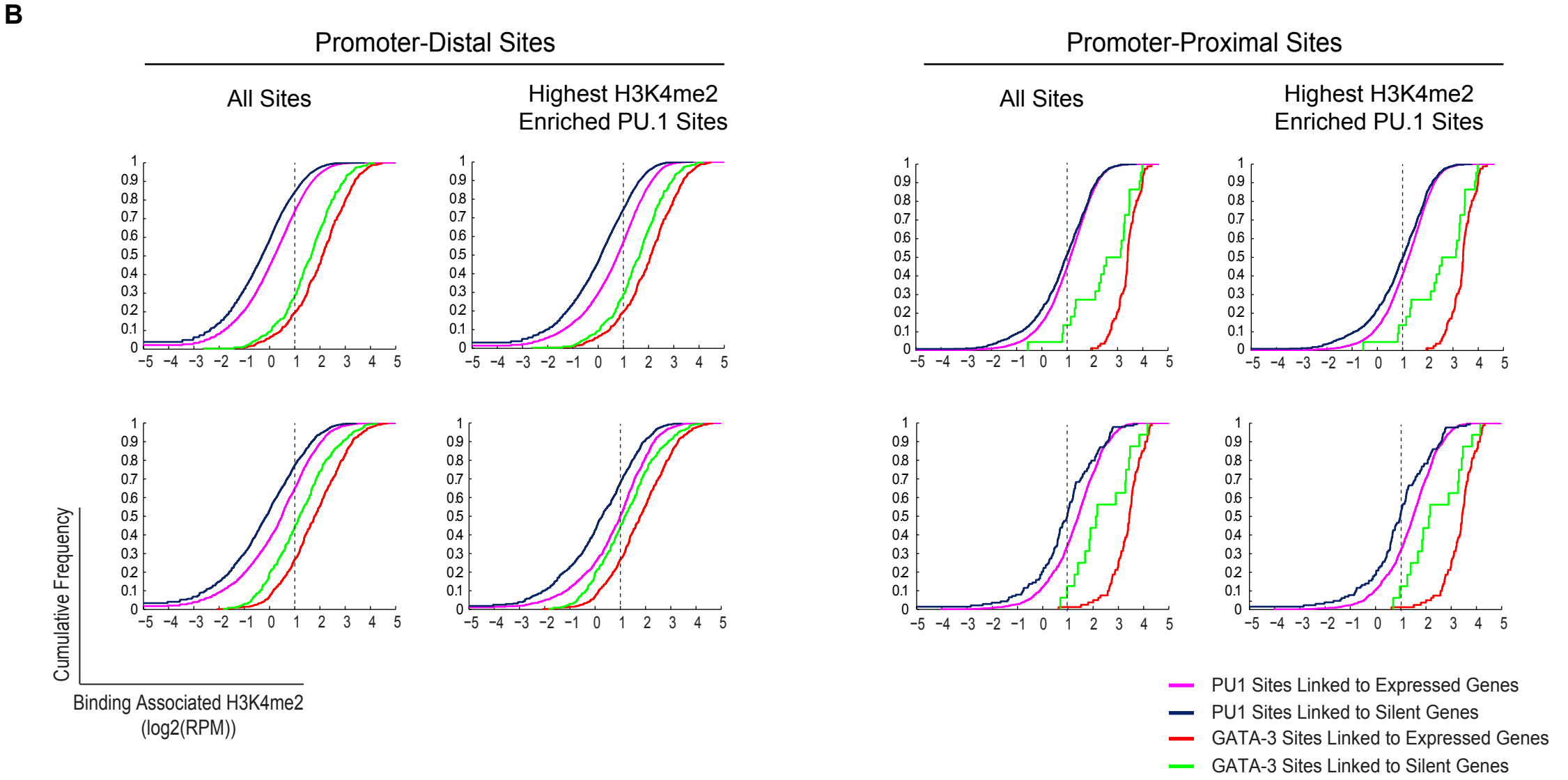


Fig. S6

Supplemental Figure 7





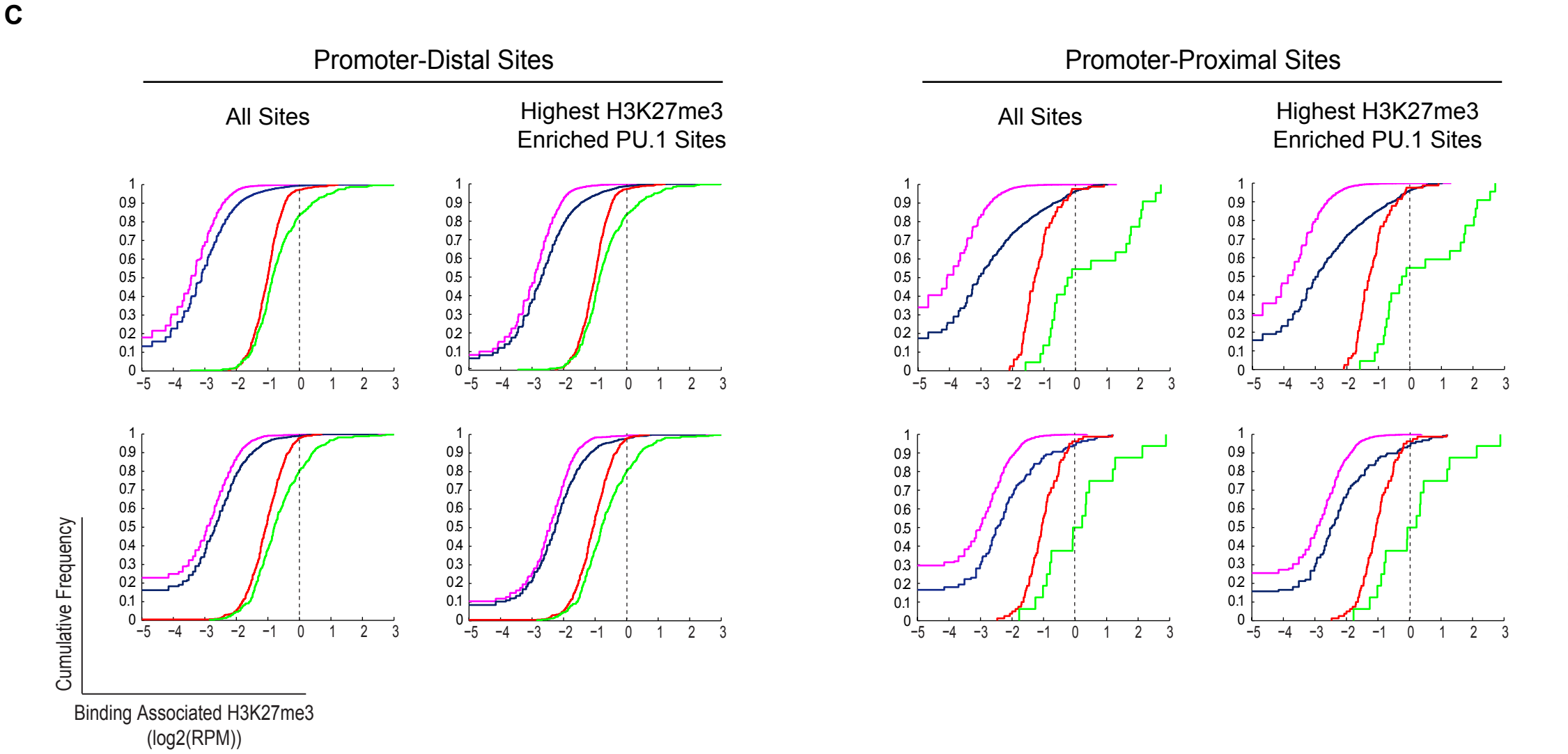


Fig. S7