## Supplemental Data

# Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation

Kyong-Rim Kieffer Kwon, Zhonghui Tang, Ewy Mathe, Jason Qian, Myong-Hee Sung, Guoliang Li, Wolfgang Resch, Songjoon Baek, Nathanael Pruett, Lars Grøntved, Laura Vian, Steevenson Nelson, Hossein Zare, Ofir Hakim, Deepak Reyon, Arito Yamane, Hirotaka Nakahashi, Alexander L. Kovalchuk, Jizhong Zou, J. Keith Joung, Vittorio Sartorelli, Chia-Lin Wei, Xiaoan Ruan, Gordon L. Hager, Yijun Ruan, & Rafael Casellas

## Supplemental Experimental Procedures

- *Genome editing*
- *qPCR*
- *ChIP-Seq*
- *DHS-Seq*
- *RNA-Seq*
- *Bi-Seq*
- *ChIA-PET*
- *Bioinformatics software*
- *Bioinformatics analysis:*
    - *DHS-Seq, ChIP-Seq, and RNA-Seq processing*
    - *Alignment of DHS-Seq and ChIP-Seq reads*
    - *Alignment of RNA-Seq reads*
    - *PolII ChIA-PET processing*
    - *Bi-Seq processing*
    - *Definition of DHS hotspots and footprints*
    - *Annotation of DHS regions into promoters and enhancers*
    - *ChIA-PET cluster interaction: definitions and annotation*
    - *Classification of PET clusters*
    - *Defining transcription models from chromatin interactions*
    - *Reproducibility of ChIA-PET PolII peaks and interactions*
    - *Identification of lincRNAs associated genes*

- *CpG methylation changes across cell types*

Index of Supplemental Materials

- ***Table S1, relate to Figure 1:*** *Deep-sequencing statistics and SRA/GEO numbers*
- ***Table S2, relate to Figure 4:*** *Characterization of genes linked to B or ES specific enhancers*
- ***Table S3, relate to Methods section:*** *Primers and PCR conditions used for TALEN-mediated genome editing*
- ***Figure S1, relate to Figure 1****: Properties of DHS-Seq and ChIA-PET datasets*
- ***Figure S2, relate to Figure 1****: Analysis of PolII long-range interactions via ChIA-PET*
- ***Figure S3, relate to Figure 1****: Examples of ChIA-PET connectivity*
- ***Figure S4, relate to Figure 2****: Validation of ChIA-PET connections via genome editing*
- ***Figure S5, relate to Figure 3****: Characterization of ChIA-PET clusters*
- ***Figure S6, relate to Figure 4****: Changes in CpG methylation during B cell lymphopoiesis.*
- ***Figure S7, relate to Figure 7****: Distribution of transcription factor footprints at promoters and enhancers.*

## Supplementary figure legends

**Supplementary Figure 1: Properties of DHS-seq and ChIA-PET datasets.** (**A**) Reproducibility of DHS-seq signal at defined hotspots in two biological replicates of activated B cells. Correlation was calculated via Spearman's coefficient $\rho$. (**B**) Overlap in B cell DHS regions vis-à-vis Med12, p300, and Nipbl occupancy. (**C**) Bar graph showing the percentage of ChIA-PET interactions overlapping with DHS or non-DHS genomic domains. (**D**) Box plot representing the transcription levels of genes associated or not associated with ChIA-PET connections at promoters ($P$ < 2e-16). (**E**) Box plot representing PolII density (RPKM) at promoters associated or not associated with ChIA-PET connections ($P$ < 2e-16). (**F**) Percentage of H3K27Ac$^{+}$ (active) enhancers within ChIA-PET anchored and not anchored groups. (**G**) DHS signal intensity at enhancers and promoters associated with 0, 2, or more than 5 PolII long-range interactions. (**H**) Schematics showing the classification of PolII ChIA-PET interactions (PETs) as intragenic,

extragenic, intergenic, or enhancer-enhancer. (**I**) Representation of direct and indirect connections between promoters and enhancers as determined by ChIA-PET. Percentages were calculated for B and ES cell ChIA-PET datasets combined.

**Supplementary Figure 2: Analysis of PolII long-range interactions via ChIA-PET** (**A**) Schematics showing details of the ChIA-PET protocol, including ChIP pull-down and chromosome conformation capturing steps. To facilitate their representation throughout the main text, individual PolII long-range interactions or PETs within 500bp from each other were grouped into PET clusters. By definition (see Methods), two elements are considered to be connected if they are linked by a PET cluster of 2 or more individual PETs. As in previous ChIA-PET publications, singletons were excluded from the analysis. (**B**) Reproducibility of PolII ChIA-PET peaks at biological replicates from activated B cells and ES cells. (**C**) Overlapping PET clusters from two B cell and ES cell ChIA-PET biological replicates. The panel depicts the reproducibility of interacting PET clusters where each dot represents PET counts from replicates and dot sizes are proportional to the number of overlapping PET clusters.

**Supplementary Figure 3:** Regulatory domain interactions determined by ChIA-PET at *Sox2* (**A**) and *Igh* (**B**) gene loci from ES cells and LPS+IL-4 activated B lymphocytes respectively. ChIA-PET and DHS datasets in each cell type are provided for both loci. At *Igh*, the number of PETs anchoring the constant domain and the 3'E$\alpha$ enhancer are provided. mRNA expression is also provided as RPKM values (+ strand transcription in green, - strand in blue).

**Supplementary Figure 4: Validation of ChIA-PET connections via genome editing.** (**A**) Custom FLASH TALENs were engineered to selectively delete specific DHS enhancer domains in CH12 B cells or ES cells. The donor construct contains a loxP-PGK promoter/puromycin-T2A-thymidine kinase/PolyA-loxP cassette to select positive clones. PCR primers (Table S3) specific for genomic and construct sequences were used to verify the insertion of knockout constructs at the desired targeted genomic locus. The cassette was removed by Cre-mediated recombination and clones were selected with Ganciclovir and verified by PCR. (**B**) Targeted deletion of AID enhancer E1 results in a marked decrease in PolII occupancy at AID gene regulatory domains but not at the *Foxj2-Necap1* locus (~190Kb downstream of *Apobec1*). (**C**) *Pou2af1* enhancer E3 was deleted and *Pou2af1* expression levels were measured by qPCR (bar graph) in non-activated (N.A.), or CH12 cells (WT, black bars; and $\Delta$E3, yellow bars) activated in the presence of IL-6, $\alpha$CD40, TGF$\beta$, and IL4. *P* values were 0.15 (N.A.), 0.008 (IL6), and 0.03 ($\alpha$CD40+TGF$\beta$+IL4). To demarcate regulatory domains both DHS (black) and H3K4me3 (red) profiles are provided. (**D**) Enhancers 1 (E1) and 2 (E2) at the *Cd79a* locus were deleted and transcription of genes within (*Rps19*, *Cd79a*, and *Arhgef1*) and outside the cluster (*Pou2f2* and *Rps3*) was assessed by qPCR (right bar graph) in WT (black bars), $\Delta$E2 (grey bars), and $\Delta$E1 (yellow bars) CH12 cells. Data represent the mean +/- SEM (n = 6). *P* values were 0.009 (*Cd79a*, ΔE2), 0.0001 (*Cd79a*, ΔE1), 0.05 (*Rps19*,

ΔE2), and 0.04 (*Arghef1*, ΔE1). ***Note***: We point out that most PolII long-range interactions link E1 to E2, instead of tethering the enhancers to the *Cd79a* promoter. One interesting possibility is that these interactions help establish the higher-order chromatin structure of the two loci, which in turn may regulate promoter activity as has been shown at the mouse β-globin locus within the context of CTCF binding (Phillips and Corces, 2009). Notably, E1 and E2 enhancers in question display clear CTCF occupancy, which might facilitate their association (depicted with a semi circle). A similar scenario applies at the Pax5 locus (see Figure S7D below).

**Supplementary Figure 5: Characterization of PET gene clusters.** (**A**) Distribution of the cluster span (intra-cluster distance) for single- and multiple-promoter gene clusters. Distances are provided in megabases (Mb). (**B**) PET cluster associated with *Mir290-295* locus. (**C**) Heat map: PolII recruitment at promoters where no ChIA-PET interactions were detected (not-anchored) or at promoters from single- or multiple-gene clusters (for definitions see Figure 3 main text). Bar graph: transcription levels, as measured by RNA-Seq (FPKM values), of the three promoter groups. (**D**) Clustering of the lymphoid signaling *Gimap* gene family. The enhancer located downstream of the Gmap6 gene (boxed) is unique among B cell gene regulatory domains in that it interacts with 7 different *Gmap* promoters (red arrows). (**E**) PET interactions between *Bcl11a* and lincRNA *E123592*. DHS islands are provided to delineate regulatory elements.

**Supplementary Figure 6: Changes in CpG methylation during B cell development.** (**A**) Long-range interactions and DHS profiles at loci containing the B cell-specific *Cd79b* gene. (**B**) Box plot showing CpG methylation levels (y-axis) relative to number of PolII long-range interactions in B cells (x-axis). (**C**) Comparison of methylation levels (%) at DHS enhancer regions present only in ES cells (blue line), activated B cells (red line), and in both cell types (black line). (**D** and **E**) Relative demethylation profiles at *Pim1* and *Aicda* gene loci in activated (a) and resting (r) B cells, as well as CLP, and KSL bone marrow progenitors relative to ES cells (the methylome of ES cells was subtracted from that of each cell type as in Figure 5B of the main text). Enhancers and promoters are highlighted on top of each graph.

**Supplementary Figure 7: Distribution of transcription factor footprints at promoters and enhancers.** (**A**) Bar graph representing the enrichment of transcription factor DNA motifs within footprints, DHS islands, or genomic DNA not associated with DHS islands in B cells. Motif enrichment was calculated based on the percentage of nucleotides within the particular domain that overlapped with DNA motifs. (**B**) Extended view (+/-1Kb) of digital footprint signatures for Sp1 and Erf1. (**C**) Distribution of ES cell transcription factor motifs at the three enhancer groups defined in Figure 7A of the main text (enhancers present in both cell types and associated with shared promoters (light blue), cell-type specific enhancers associated with promoters active in B and ES cells (grey), and cell-type specific enhancers associated with cell-type specific promoters (blue)). (**D**)

Long-range interactions and DHS profiles at the B cell-specific *Pax5* gene. Two linked enhancers ~250Kb from the *Pax5* TSS are highlighted. **Note**: As discussed within the context of the *Cd79a* locus (Figure S4D), the *Pax5* interacting enhancers are recruit substantial CTCF, which might facilitate their tethering in 3D space (depicted with a semi-circle).

## EXTENDED EXPERIMENTAL PROCEDURES

### Genome editing

TALENs for specific loci were designed and assembled using protocols described in (Reyon et al., 2012). The donor vector included 500-1500 bp homology arms and a loxP-flanked puromycin-T2A-thymidine kinase cassette to select for targeted clones (Figure S3). Activated B or E14 cells were nucleofected with the donor cassette and the TALEN plasmid pair using Nucleofector Kit V according to the manufacturer's instructions (Lonza). After 72 h, limiting dilution was performed in media containing 0.5-1 ug/ml puromycin (Sigma) and incubation was continued for 6-8 days. Individual clones were picked and genomic DNA was extracted (Promega). Genotyping was done by nested PCR using locus-specific external and vector internal primers (Table S3) under the following conditions: 98°C for 30 sec; 35 x (98°C for 10 s, 63°C for 30 s, 72°C for 30 s); 72°C for 3 min; hold at 4°C. PCR products were run on 1% agarose gel. Positive clones were expanded and nucleofected with a plasmid expressing CMV- or EF1α-driven Cre recombinase (Addgene). At 72 h, limiting dilution was performed in the presence of 0.5-2 ug/ml ganciclovir (Sigma) for 6–8 days. Once again, single clones were picked and genotyped by PCR using primers amplifying deleted loci (Figure S3).

### qPCR

RNA from locus-deleted clones was extracted with RNAqueous-Micro Kit (Ambion). cDNA was synthesized with SuperScript III First-Strand Synthesis SuperMix (Invitrogen). qPCR samples were mixed with SyBrgreener (Invitrogen) and run on BioRad CFX96. qPCR primers are listed in Table S3.

### ChIP-Seq

Cultured cells were fixed with 1% formaldehyde (Sigma) for 10' at 37°C. Fixation was quenched by addition of glycine (Sigma) at a final concentration of 125 mM. Twenty million fixed cells were washed with PBS and resuspended in 1 ml of RIPA buffer (10 mM Tris [pH 7.6], 1 mM EDTA, 0.1% SDS, 0.1% sodium deoxycholate, 1% Triton X-100, 1× Complete Mini EDTA free proteinase inhibitor (Roche)) or stored at −80°C until further processing. Sonication was performed using Covaris S2 sonicator at duty cycle 20%, intensity 5, cycle/burst 200 for 30 min. Five to ten micrograms of anti-NIPBL (Bethyl, A301-779A), anti-p300 (Santa Cruz, SC-584), anti-Med12 (Bethyl, A300-774A), anti-H3K27Ac (Abcam ab4279-100), anti-H2AZ (Abcam, ab4174-100), anti-H3K4me1 (Abcam ab8895-50), anti-H3K4me3 (Millipore 04-745), and anti-PU1 (Santa Cruz, SC-352) was incubated with 40 µl of Dynabeads Protein A (or G) for 40 min at room temperature. Antibody-bound beads were added to 500 µl of sonicated chromatin, incubated at 4°C overnight, and

washed twice with RIPA buffer, twice with RIPA buffer containing 0.3M NaCl, twice with LiCl buffer (0.25 M LiCl, 0.5% Igepal-630, 0.5% sodium deoxycholate), once with TE (pH 8) plus 0.2% Triton X-100, and once with TE (pH 8.0). Crosslinking was reversed by incubating the beads at 65°C for 4 hr in the presence of 0.3% SDS and 1 mg/ml Proteinase K. ChIP DNA was purified by phenol-chloroform extraction followed by ethanol precipitation. DNA was subsequently blunt-ended with End-It DNA end repair kit (Epicenter) and A-tailed with Taq DNA polymerase (Invitrogen) in the presence of 200mM of dATP for 40 min at 70°C. Samples were purified by phenol-chloroform extraction after each reaction. Illumina compatible adaptors (Illumina or Bioo Scientific) were then ligated with T4 DNA ligase (Enzymatics), and the reaction was purified once with AMpure XP magnetic beads (Beckman Coulter). Samples were PCR amplified for 18 cycles with KAPA HiFi DNA polymerase mix (KAPA Biosystems) and run on a 2% agarose gel and size-selected at 200–300 bp. Thirty-six or 50 bp of sequencing data (Table S1) were acquired on the Illumina GAII or HiSeq2000 (Illumina).

**DHS-Seq**

Digital DNase I mapping was performed as described in reference (Sekimata et al., 2009). Briefly, we pelleted $1\times10^8$ B cells, washed them with cold PBS and resuspended them in Buffer A (15 mM Tris-Cl (pH 8.0), 15 mM NaCl, 60 mM KCl, 1 mM EDTA (pH 8.0), 0.5 mM EGTA (pH 8.0), 0.5 mM spermidine, 0.15 mM spermine) to a final concentration of $2\times10^6$ cells/ml. Nuclei were isolated by dropwise addition of an equal volume of Buffer A containing 0.04% NP-40, followed by incubation on ice for 10'. Nuclei were centrifuged at 1,000*g* for 5 min and then resuspended and washed with 25 ml of cold Buffer A. Nuclei were resuspended in 2 ml of Buffer A at a final concentration of $1\times10^7$ nuclei/ml. We performed DNase I (Roche, 10–80 U/ml) digests for 3' at 37 °C in 2 ml volumes of DNase I buffer (13.5 mM Tris-HCl pH 8.0, 87 mM NaCl, 54 mM KCl, 6 mM CaCl2, 0.9 mM EDTA, 0.45 mM EGTA, 0.45 mM Spermidine). Reactions were terminated by addition of an equal volume (2 ml) of stop buffer (1 M Tris-Cl (pH 8.0), 5 M NaCl, 20% SDS and 0.5 M EDTA (pH 8.0), 10 μg/ml RNase A, Roche) and incubated at 55°C. After 15', we added Proteinase K (25 μg/ml final concentration) to each digest reaction and incubated for one hour at 55°C. After DNase I treatment, phenol-chloroform extractions were performed. Control (untreated) samples were processed as above except for the omission of DNase I. DNase I double-cut fragments and sequencing libraries were constructed as described in the ChIP-Seq protocol with the exception of the size-selection of 300-400 bp.

**RNA-Seq**

Total RNA from $10^6$ ES, resting, or activated B cells was isolated by Trizol extraction. To obtain more precise measurements of transcription, RNA spike-ins were used by adding 1μl of 1/10 dilution of Ambion's ERCC RNA Spike-in Mix (catalog number 4456740) to total RNA. mRNA was then isolated and the standard RNA-Seq library preparation was performed following Illumina's RNA-Seq protocol v2.

**Bi-Seq**

Genomic DNA was isolated from $5x10^6$ cells using Qiagen DNeasy blood and tissue kit. Libraries were prepared following whole-genome bisulfite sequencing for methylation analysis guide from Illumina (15021861_B) with slight modifications. Briefly, 5 μg of genomic DNA was sheared and blunt-ended with End-It DNA end repair kit (Epicenter) and A-tailed with Taq DNA polymerase (Invitrogen) in the presence of 200mM of dATP for 40 min at 70°C. Illumina compatible adaptors (5' P-GATXGGAAGAGXGGTTXAGXAGGAATGXXGAG, 5' AXAXTXTTTXXXTAXAXGAXGXTXTTXXGATXT where X is a methylated cytosine) were then ligated with T4 DNA ligase (Enzymatics). Adapter-ligated DNA of 275-350 bp was isolated by 2% agarose gel electrophoresis, and sodium bisulfite conversion performed on it using the Epitect Bisulfite kit (Qiagen). Bisulfite converted DNA was divided in three tubes and PCR amplified for 6 cycles by PfuTurbo Cx hotstart DNA polymerase (Stratagene). The reaction products were purified using the MinElute PCR purification kit (Qiagen) then separated by 2% agarose gel electrophoresis and purified from the gel using the MinElute gel purification kit (Qiagen).

**ChIA-PET**

RNA PolII ChIA-PET was performed as previously described (Fullwood et al., 2009; Goh et al., 2012; Li et al., 2012). Briefly, B or ES cells (up to $3x10^9$ cells) were treated with 1% formaldehyde at room temperature for 10 min and then neutralized using 0.2 M glycine. The cross-linked chromatin was subjected to fragmentation with an average length of 300 bp by sonication. The anti-PolII monoclonal antibody 8WG16 (Covance, MMS-126R) was used to enrich PolII-bound chromatin fragments. A portion of ChIP DNA was eluted off from antibody-coated beads for concentration quantification using Picogreen fluorimetry and for enrichment analysis using quantitative PCR. For ChIA-PET library construction ChIP DNA fragments from two biological replicates were end-repaired using T4 DNA polymerase (NEB) and ligated to either linker A or linker B. Other than four nucleotides in the middle of the linkers that were used as nucleotide barcode, the two linkers share the same nucleotide sequences. After linker ligation, the two samples were combined for proximity ligation in diluted conditions. During the proximity ligation, DNA fragments within the same ChIP complex with the same linker were ligated, which generated the ligation products with homodimer linker composition. However, chimeric ligations between ChIP fragments that are bound in different chromatin complexes could also occur, thus producing ligation products with heterodimer linker composition. These heterodimer linker composition products were used to assess the frequency of non-specific ligations and were then removed bioinformatically. As shown in Figure S2A, all heterodimer linker ligations are by definition non-specific. Because random intermolecular associations in the test tube are expected to be comparable for linkers A and B, the frequency of random homo and heterodimer linker ligations must also be equivalent. In our PolII ChIA-PET libraries, only 3.8% of pair-end ligations (PETs) involved heterodimer linkers. Thus, we estimate that less than 5% of total homodimer ligations are non-specific. Even though this represents but a small number of PETs, we reduced this number even further by discarding singleton PETs. In other words, we only report PolII long-range interactions when two or more pair-end reads create a PET cluster (Figure S2A). The strategy is based upon

the fact that random heterodimer ligations rarely form PET clusters. In the PolII ChIA-PET libraries for instance, of 487,981 heterodimer PETs, only 26 PET clusters were obtained. Conversely, 9M homodimer PETs created ~15,000 PET clusters. Thus, while there are ~20 times more homodimer than heterodimer PETs, we obtain ~600 times more homodimer than heterodimer PET clusters. Following proximity ligation, the Paired-End-Tag (PET) constructs were extracted from the ligation products and the PET templates were subjected to paired-end sequencing using Illumina GAII.

**Bioinformatics Software**

- ABI Solid whole transcriptome alignment pipeline (WTP) and Bioscope 1.0
- BatMis 3.0 (Tennakoon et al., 2012)
- Bedtools 2.17.0 (Quinlan and Hall, 2010)
- Bowtie 0.12.8 (Langmead et al., 2009)
- ChIA-PET Tool (Li et al., 2010)
- Cufflinks 2 (Trapnell et al., 2010)
- Cytoscape 2.8.2 (Shannon et al., 2003)
- DNase2Hotspots (Baek et al., 2012)
- Fastqc 0.10.0 and 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- ggplot2 (H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.)
- gsnap 2012-07-20 (Wu and Nacu, 2010)
- htseq 0.5.3p9 (Simon Anders, http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html)
- Illumina Casava (versions 1.3 to 1.8.2)
- Macs2 2.0.10 (Zhang et al., 2008)
- R 2.15 ("*R: A Language and Environment for Statistical Computing*", R Core Team, http://www.R-project.org)
- Samtools 0.1.18 (Li et al., 2009)

**Bioinformatics analyses**

**DHS-Seq, ChIP-Seq, and RNA-Seq processing**

DHS-Seq and ChIP-Seq were sequenced using the Illumina Genome Analyzer II and the Illumina HiSeq 2000, following the manufacturer's instructions. The standard Illumina pipeline (versions 1.3 to 1.8.2) was used for image analysis and base calling. The software Fastqc was applied to the fastq files to ensure that the quality of each read position was no less than 28.

**Alignment of DHS-Seq and ChIP-Seq reads**

Reads were aligned to the National Center for Biotechnology Information mouse genome data (July 2007; NCBI37/mm9). The alignment software Bowtie version 0.12.8 was used with the following options: --best --all --strata -m1 -n2 –l[read length]. These options report the reads that align uniquely to the best stratum and allowing 2 mismatches.

**Alignment of RNA-Seq reads**

Strand-specific B and E14 cell RNA-Seq reads were mapped onto the mouse reference genome (mm9) with SOLiD WTP and analyzed by ABI SOLiD Bioscope (version 1.0) analysis pipeline. Expression values were determined in terms of reads per kilobase per million mapped reads (RPKM).

For increased precision in transcription measurements, ERCC RNA spike-in mix was added to the total RNA before standard RNA-Seq library preparation (as described above). Reads were aligned against the mouse genome (build mm9) with gsnap using only known splice sites obtained from the Refseq annotation as present in the USCS genome browser database in January 2012 with –novelsplicing=0. Subsequently, the same reads were also aligned to the ERCC RNA standard, also with gsnap but not looking for splice sites. The number of reads matching each Refseq gene was then determined using htseq-count while the number of reads matching ERCC standard RNAs was determined by a simple line count. Spike and mRNA counts were then read into R where counts were normalized by library size and exonic size of each gene to obtain RPKM (reads per kb per million aligned reads). A linear model was fit to the ERCC spike data to relate the known copy number to the measured RPKM and cell type: lm(log10(known copy number) ~ log10(RPKM) + cell, data = counts). The linear model was then used to estimate the copy number of each expressed gene based on the cell type and measured RPKM.

**PolII ChIA-PET processing**

The pipeline for ChIA-PET sequencing processing is described in (Li et al., 2010). Briefly, the redundant sequences were collapsed into a non-redundant PET sequence set based on sequence content. The non-redundant PET sequences were analyzed for linker barcode composition and identified as sequences with homodimer linker derived from specific ligation products, or sequences with heterodimer linker derived from nonspecific ligation products. The linker composition information was used to evaluate the noise in the ChIA-PET library and sequences with heterodimer linkers were removed. After trimming the linkers, PET sequences were mapped to the mouse reference genome (mm9) using customized BatMis 3.0 and only perfectly and uniquely aligned PETs were retained. PETs uniquely aligned and similarly mapped (within +/- 2bp) were merged into one PET as they were considered to be derived from the same DNA-fragment ligation with variations in MmeI enzyme digestion. Based on the mapping characteristics, each PET was categorized as a self-ligation PET (two ends of the same DNA fragment) or inter-ligation PET (two ends from two different DNA fragments in the same chromatin complex). The PET categories were established by evaluating the mapping orientation and genomic span between two tags of a PET. The inter-ligation PETs were further categorized into intrachromosomal PETs, where the two tags from a PET lay on the same chromosome, and interchromosomal PETs, where the two tags lay on different chromosomes. The self-ligation PETs were utilized as a proxy for ChIP fragments since they provide two defined end points, as described above. The coverage of all self-ligated PET sequences across the genome reflects the specific PolII binding sites, which is analogous to ChIP-Seq mapping for protein binding sites. The local summits of the sequence coverage were called as potential

peaks if there were multiple self-ligation PETs overlapping in that region. Assuming multiple self-ligation PETs would not occur by random chance, the random background was set as the maximum of the global tag density, local tag densities at 10 kb and 20 kb windows around the peak, and the Poisson distribution was applied to estimate the p-values for the peaks. P values were then adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate method (B-H FDR method). FDRs less than 0.05 were used as the criteria for final peak calling. Throughout the text, PET clusters with 2 or more PETs are referred to as *interactions* or *connections*.

**Bi-Seq processing**

In silico conversion mimicking complete bisulfite conversion was performed on the National Center for Biotechnology Information mm9 mouse genome. Specifically, in silico C → T and G → A mm9 genomes were constructed by converting all C or Gs into T or As. Sequence reads were C → T converted and aligned to the C → T and G → A genomes, resulting in two alignments per sample. Alignments were performed with Bowtie version 0.12.8 with the options "--best -n2 -k2", which report the top 2 best alignments. Alignments against the C → T converted genome had the additional option –norc (only alignments against the forward strand were reported) and alignments against the G → A converted genome was performed with the additional option –nofw (only alignments against the reverse strand were reported). The two alignments per sample were merged and sorted using Samtools, and only reads that had a unique alignment with the minimum number of mismatches were retained. Going through all the sorted unique alignments, the number of methylated Cs (C nucleotide in the original read and C nucleotide in the genome) and the number of unmethylated Cs (T nucleotide in the original read and C nucleotide in the genome) were determined for every position covered. The context of each C was determined and each C was classified as CpG, CHH, or CHG, where H is either A, T, or C nucleotide. For the CpGs, base positions that were consecutive and on different strands were merged. To ensure proper quality of the methylation calls, a minimal base quality score of 20 and a minimum of 5 uniquely aligned reads covering a given position were required for that position to be considered. For each C nucleotide with sufficient quality and coverage then, we calculated the percentage of methylated C nucleotides with respect to the sum of methylated and unmethylated C nucleotides. For each CG, the difference in percent methylation between KSL or CLP or resting B cells or activated B cells and ES cells was calculated.

**Definition of DHS hotspots and footprints**

DHS hotspots were detected using the software DNase2Hotspots with the following parameters: background window size 200,000 bp, target window size 250 bp, mergeable gap: 0, z-score threshold 2, FDR 0%. Reads that mapped to chromosome M were removed prior to hotspot detection. The software DNase2Hotspots uses an algorithm that identifies local enrichment of tags in a 250 bp target window relative to a local background window spanning 200Kb. To remove artifacts, reads that overlapped satellites, long interspersed repetitive elements, and short tandem repeats were removed. With the remaining reads, a z-score, $(n - \mu)/ \sigma$ is calculated

for each target window where n is the original number of tags observed in the target window, μ is the expected number of tags overlapping that target window based on the number of mappable reads in the background window, and σ is the standard deviation of the expectation. A false discovery rate is estimated for each z-score by calculating z-scores using randomly sampled uniquely mapped reads from the observed dataset. Hotspots with 0% FDR were selected here. Finally, DHS hotspots located within 1Kb of TSSs were merged into promoter domains, while all other DHS were merged into enhancer domains if they were within 2Kb from each other.

Transcription factor footprints were called by applying 'DNase2TF', a program that implements a new footprint detection algorithm (Baek, Hager G., Sung, unpublished). Briefly, the algorithm looks for regions of local depletion in DNaseI cutting within each hotspot, taking the enzyme's dinucleotide bias and mappability of sequence reads into account. Candidate intervals are obtained by considering binomial z-scores that reflect the depletion in cutting relative to a background window 3-fold the width of the candidate region and then by merging consecutive intervals. Candidate intervals whose z-scores correspond to FDR < 5% were retained for subsequent analyses.

**Annotation of DHS regions into promoters and enhancers**

The Bedtools software suite was utilized to map the ChIP-Seq reads onto the DHS regions. First, the fragment sizes for each ChIP-Seq experiment was estimated using macs2. The 3' end of the aligned ChIP-Seq reads were then lengthened to match these predicted fragment sizes. Next, the read counts overlapping each DHS region was calculated using bedtools intersect. These overlapping read counts were then normalized to the total number of reads for each epigenetic mark and p-values based on the negative binomial distribution of the negative controls (a combination of negative pull-downs for B-cells and IgG for ES cells) were calculated for each region. DHS regions with a Benjamini-Hochberg FDR-adjusted p-value less than 1% were considered positive. DHS regions that were within 1 Kb of a TSS were denoted DHS-promoter. The remaining regions were classified as enhancers if they were p300 or Med12 or Nipbl positive.

**ChIA-PET cluster interactions: definitions and annotation**

Inter-ligation PETs reflect long-range chromatin interactions. However, there is technical noise from various sources, which should be inevitably considered. To determine if an inter-ligation PET represents a specific interaction event between two DNA fragments that are bound together in close spatial proximity by a PolII protein complex, we reasoned that the multiple inter-ligation PETs would be enriched by the ChIP procedure between the same DNA fragments. To identify such chromatin interactions, both ends of the inter-ligation PETs were extended by 500 bp along the reference genome, and PETs overlapping at both ends were clustered together as one PET cluster. The number of PETs in a PET cluster therefore reflects the frequency of an interaction between two genomic regions. To determine whether the observed number of PETs in a PET cluster was significantly different from background noise or weak interactions represented by singleton inter-ligation PETs, we evaluated p-values from a hyper-geometric distribution with the tag

counts from both anchor regions of PET clusters and the sequencing depth as input. P values were corrected using the B-H FDR method for multiple hypothesis testing and the FDR cutoff is 0.05.

**Classification of PET clusters**

In this study, we only considered the intra-chromosomal PET clusters with span less than 1 Mb on chromosomes and used the defined DHS regions to annotate PET cluster interactions. DHS regions were extended by 1.5 kb in both directions and interaction PET clusters were considered anchored if they overlapped the extended DHS regions by at least 1 base. Interaction PET clusters were then annotated into the following 4 types, according to the type of DHS region they overlapped: intragenic (promoter to gene internal region), extragenic (promoter to enhancer), and intergenic (promoter to distal promoter), and enhancer-enhancer (neither anchor of a PET cluster overlaps with a DHS promoter region). In some cases, interaction PET clusters overlap with both DHS promoters and enhancers. If the interaction PET cluster overlapped with a DHS promoter region and a DHS enhancer region located in the same gene internal region, the interaction was denoted as intragenic interaction. If the interaction PET cluster overlapped with a DHS promoter and a distal DHS enhancer, the interaction was denoted as an extragenic interaction as shown in Figure S1H.

**Defining transcription models from chromatin interactions**

PET interaction clusters were grouped with other interaction clusters based on what type of DHS region they were anchored to. We then defined three transcription models based on how the gene promoters were involved in these complex chromatin interactions: multigene (MG) interaction model, single gene (SG) interaction model, and basal promoter (BP) model. The MG model comprises intergenic promoter-promoter interactions grouped in an interaction cluster that could also include intragenic and extragenic enhancer-promoter interactions. The SG model consists of single or multiple enhancer interactions with only one gene promoter, whereas the BP model includes genes with PolII binding but no chromatin interaction. MG and SG interaction cluster models were visualized with Cytoscape.

**Reproducibility of ChIA-PET PolII peaks and interactions**

Since the sonicated fragment size of ChIA-PET library is less than 1 kb, PolII peaks between biological replicates were considered overlapping if the distance from peak center to peak center was within 1 Kb. To account for the sonication fragment size anchors of interaction PET clusters were extended 1 kb at both ends. Extended PET clusters were considered overlapping if they shared at least one nucleotide.

**Identification of lncRNAs associated genes**

The lncRNA-associated genes were identified by PET clusters, which connected the DHS promoter regions of lnRNA and their target gene. Ensembl version NCBI36.67 was utilized for annotation. When a lncRNA associated with multiple genes, the gene with highest connection frequency was selected as the lnRNA-associated gene.

Expression levels of lnRNAs and associated genes were measured as reads per kilobase per million reads (RPKM) from RNA-Seq sequencing by using Cufflinks 2 (Trapnell et al., 2013).

**CpG methylation changes across cell types**

RPKM values were calculated for genes annotated using Refseq annotation as found in the USCS genome browser database (January 2012). Using R, a two-component mixture model was fit to the log-transformed RPKM values and genes with expression levels exceeding the 95$^{th}$ percentile were considered active while genes with expression levels below the 1$^{st}$ percentile were considered silent. Methylation levels was averaged over each gene (TSS +/- 200 bp) and the distribution of methylation levels in silent and active genes were depicted with violin plots (Figure 5A). To assess global demethylation in enhancers across cellular development (Figure 5D), we considered DHS enhancer regions anchored by PETs in activated B cells only. A subset of these regions were noted as demethylated (average methylation ≤ 40%) in activated B cells and the number of these regions is depicted in Figure 5D.

**References**

Baek, S., Sung, M.H., and Hager, G.L. (2012). Quantitative analysis of genome-wide chromatin remodeling. Methods Mol Biol *833*, 433-441.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H.*, et al.* (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58-64.

Goh, Y., Fullwood, M.J., Poh, H.M., Peh, S.Q., Ong, C.T., Zhang, J., Ruan, X., and Ruan, Y. (2012). Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. Journal of visualized experiments : JoVE.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y.B., Ooi, H.S., Tennakoon, C.*, et al.* (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol *11*, R22.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J.*, et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84-98.

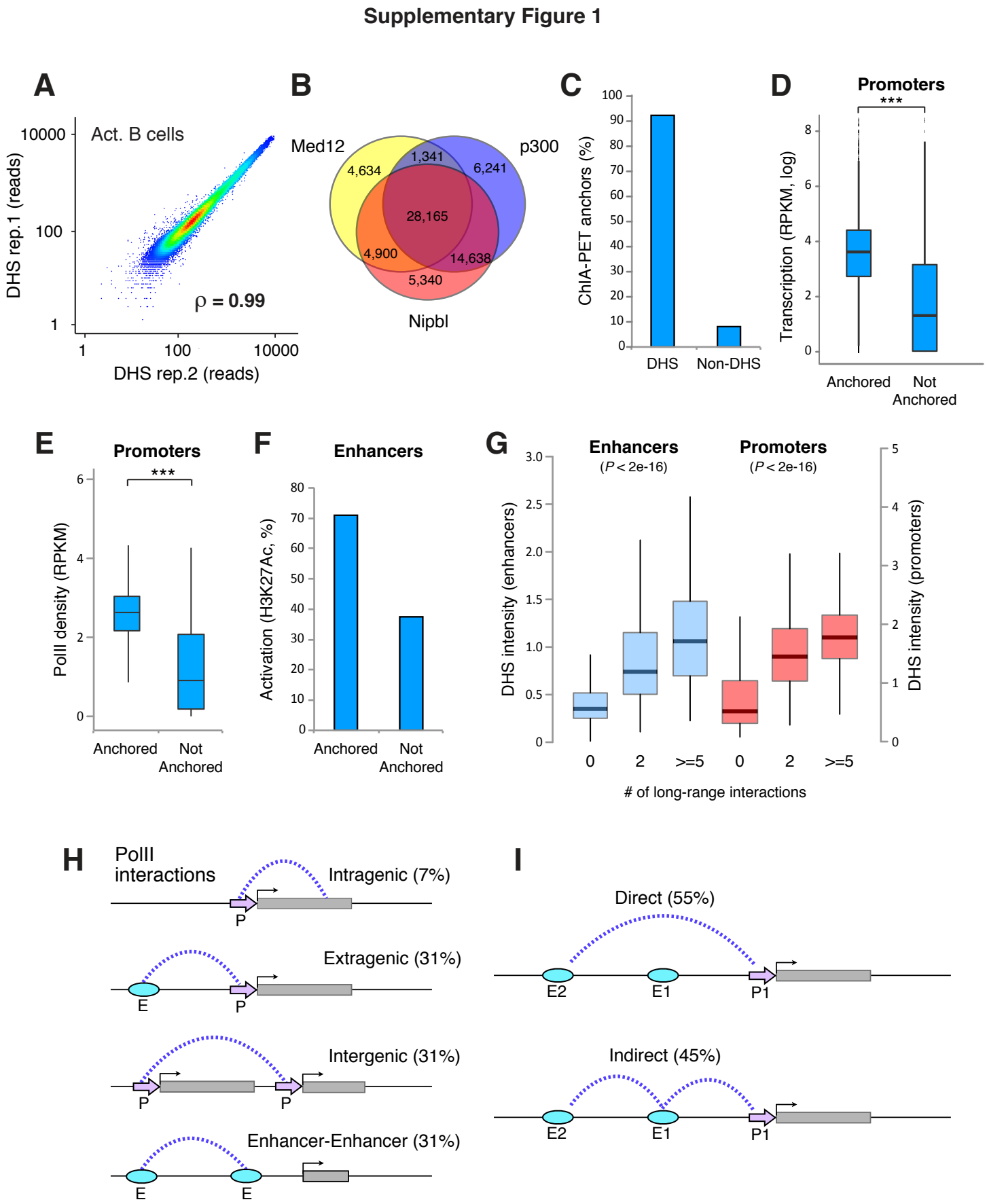
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194-1211.

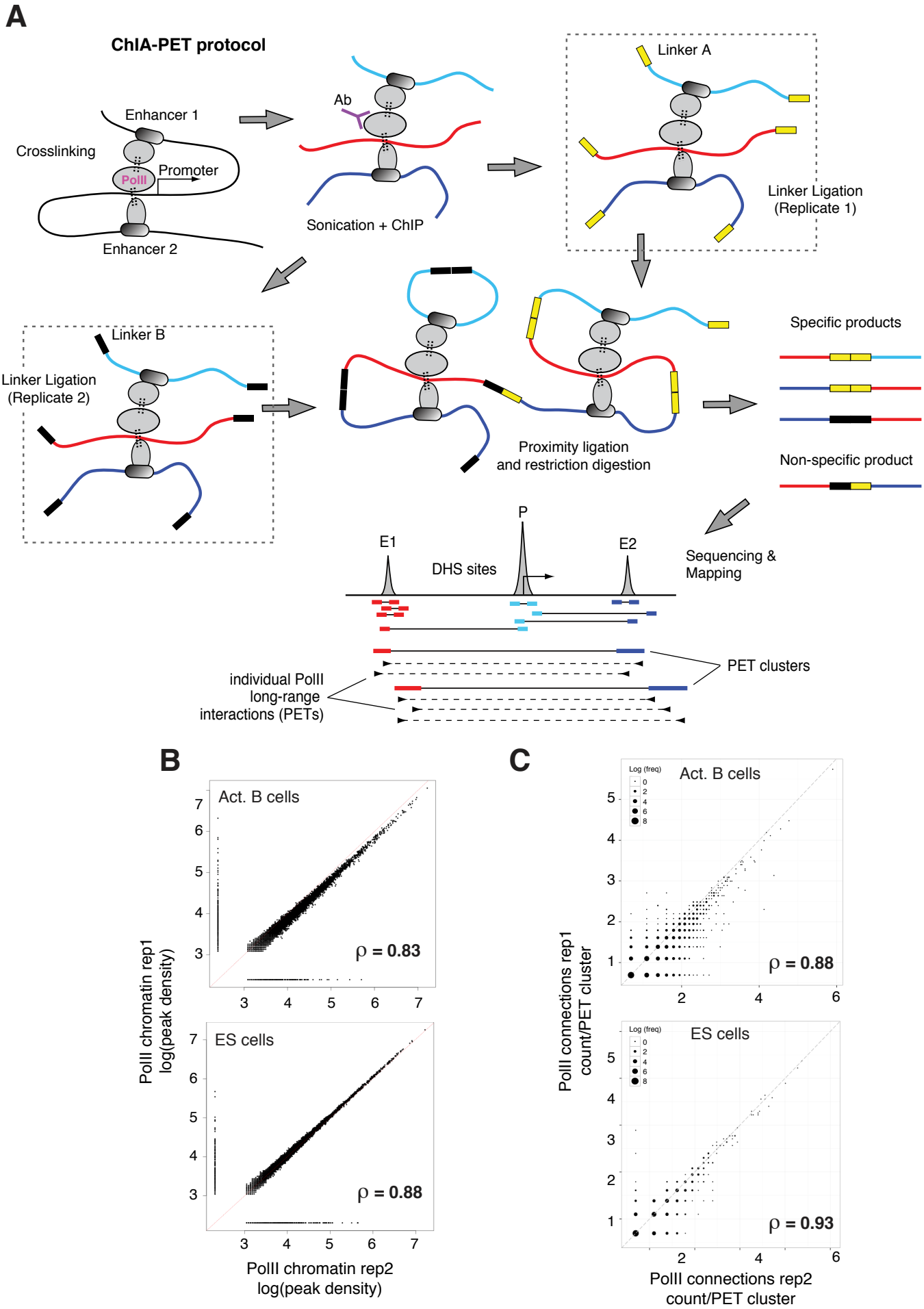
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nat Biotechnol *30*, 460-465.
Sekimata, M., Perez-Melgosa, M., Miller, S.A., Weinmann, A.S., Sabo, P.J., Sandstrom, R., Dorschner, M.O., Stamatoyannopoulos, J.A., and Wilson, C.B. (2009). CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. Immunity *31*, 551-564.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Tennakoon, C., Purbojati, R.W., and Sung, W.K. (2012). BatMis: a fast algorithm for k-mismatch mapping. Bioinformatics *28*, 2122-2128.
Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.
Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W*., et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

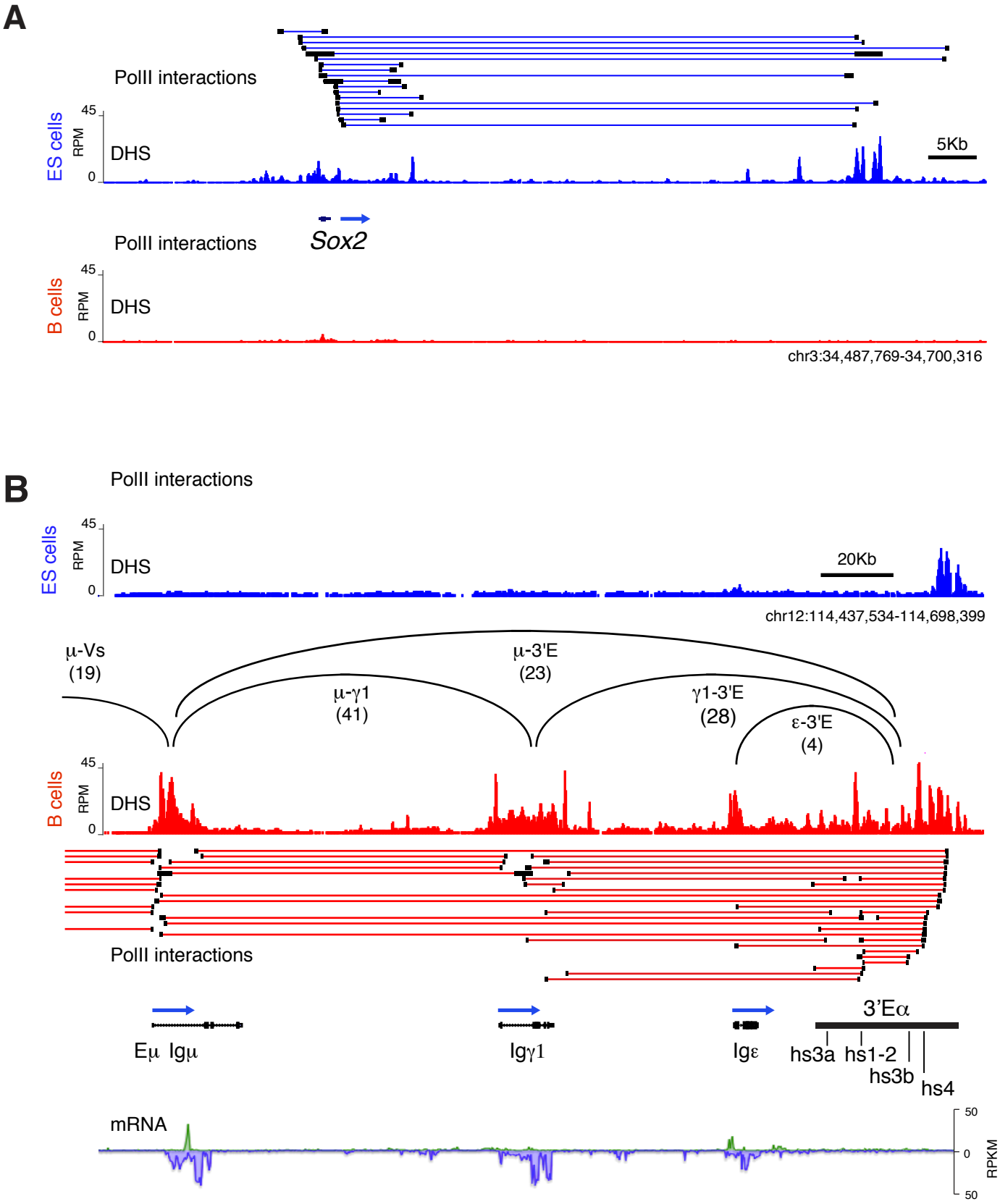
Supplemental Figure 1



# Supplementary Figure 2

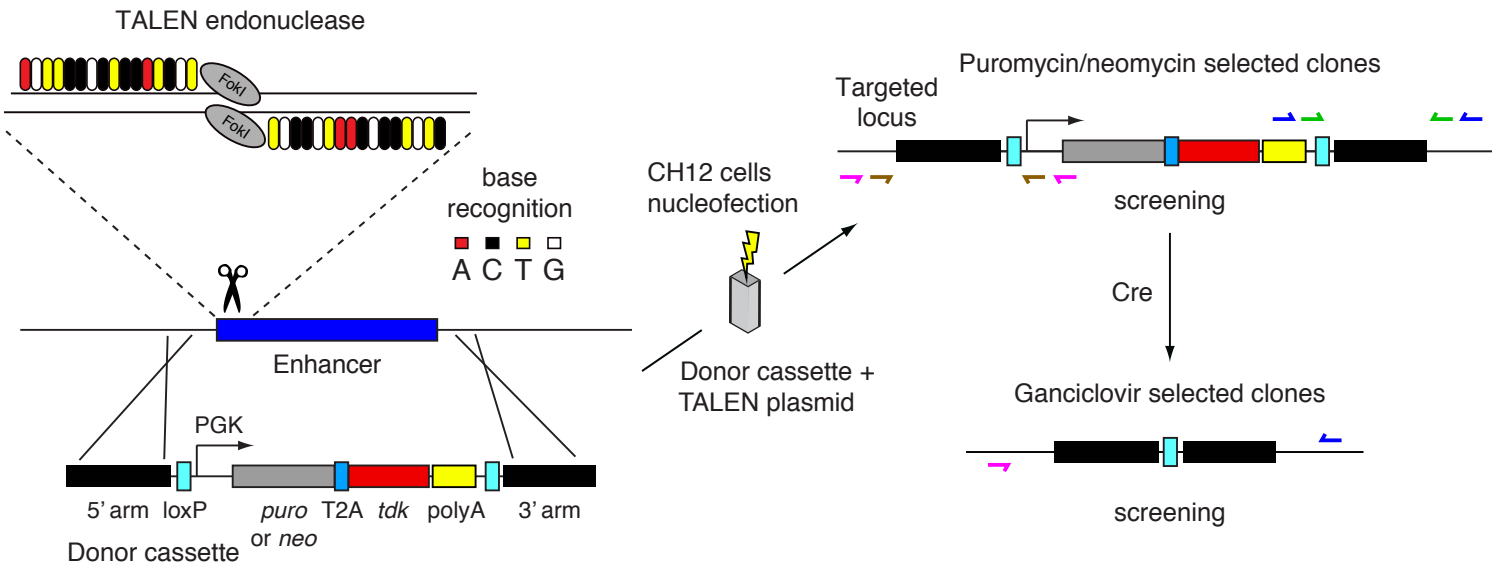


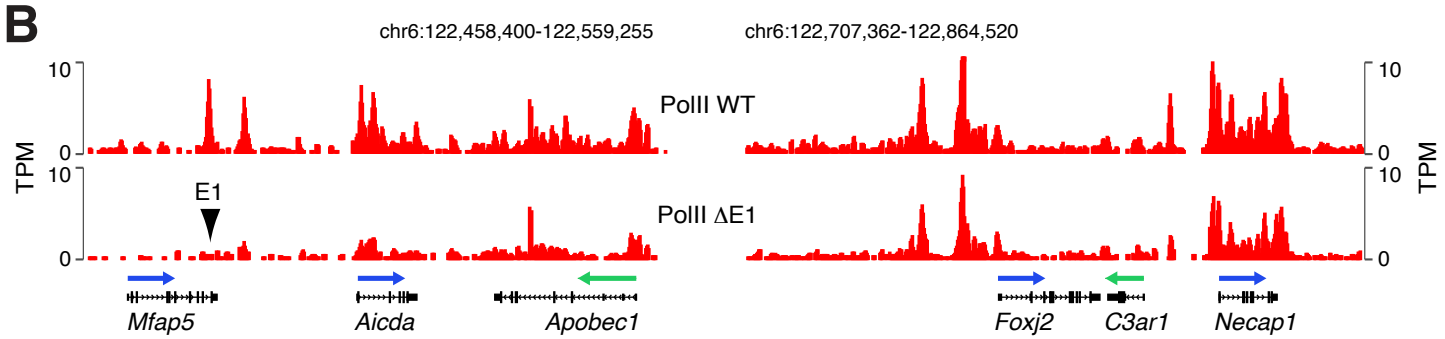
# Supplementary Figure 3

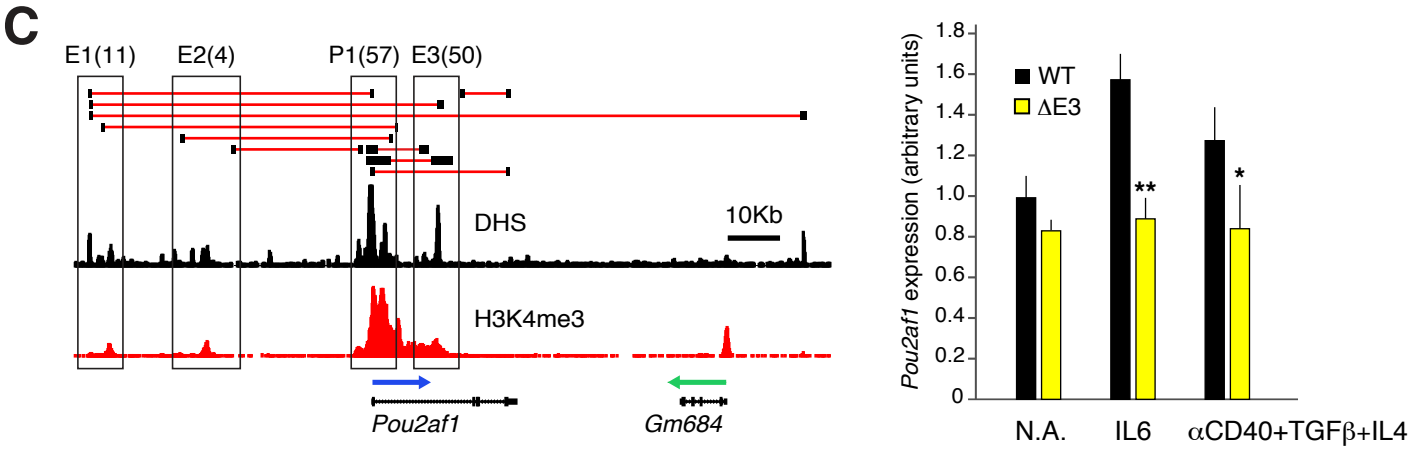


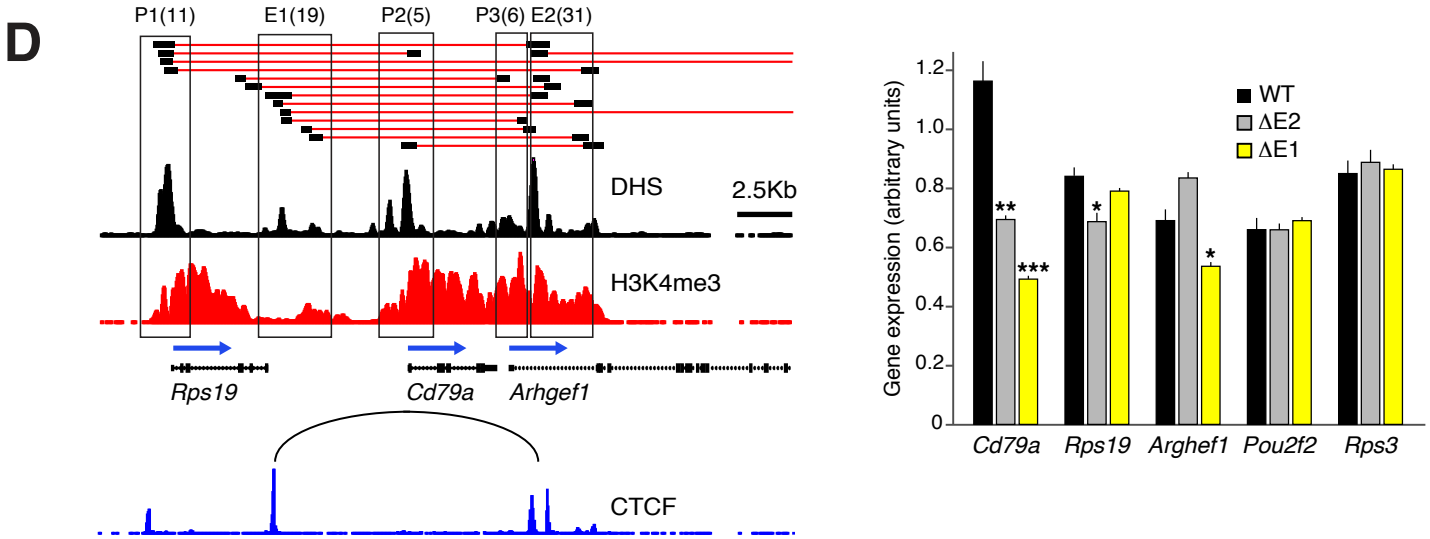
Supplemental Figure 4

# Supplementary Figure 4

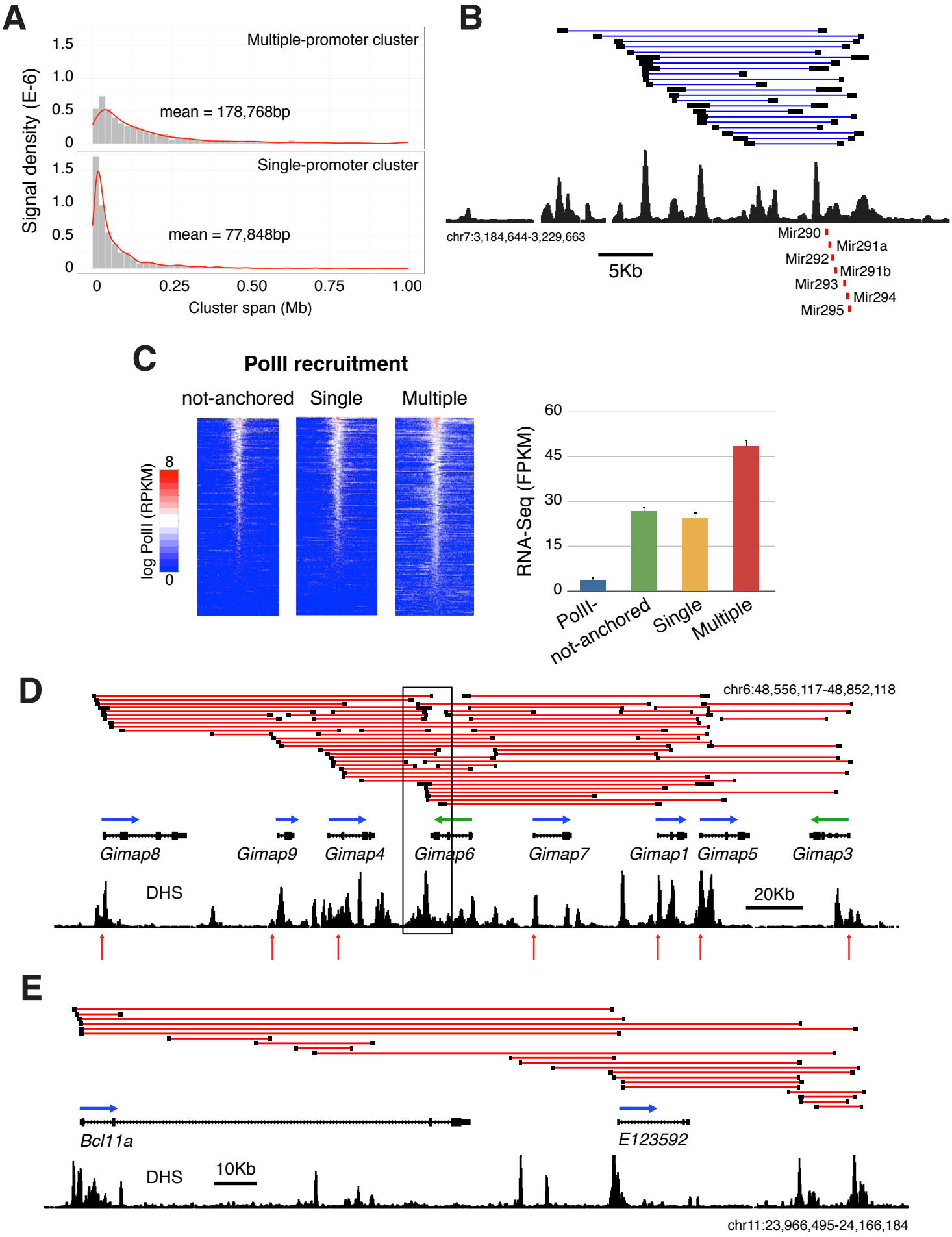


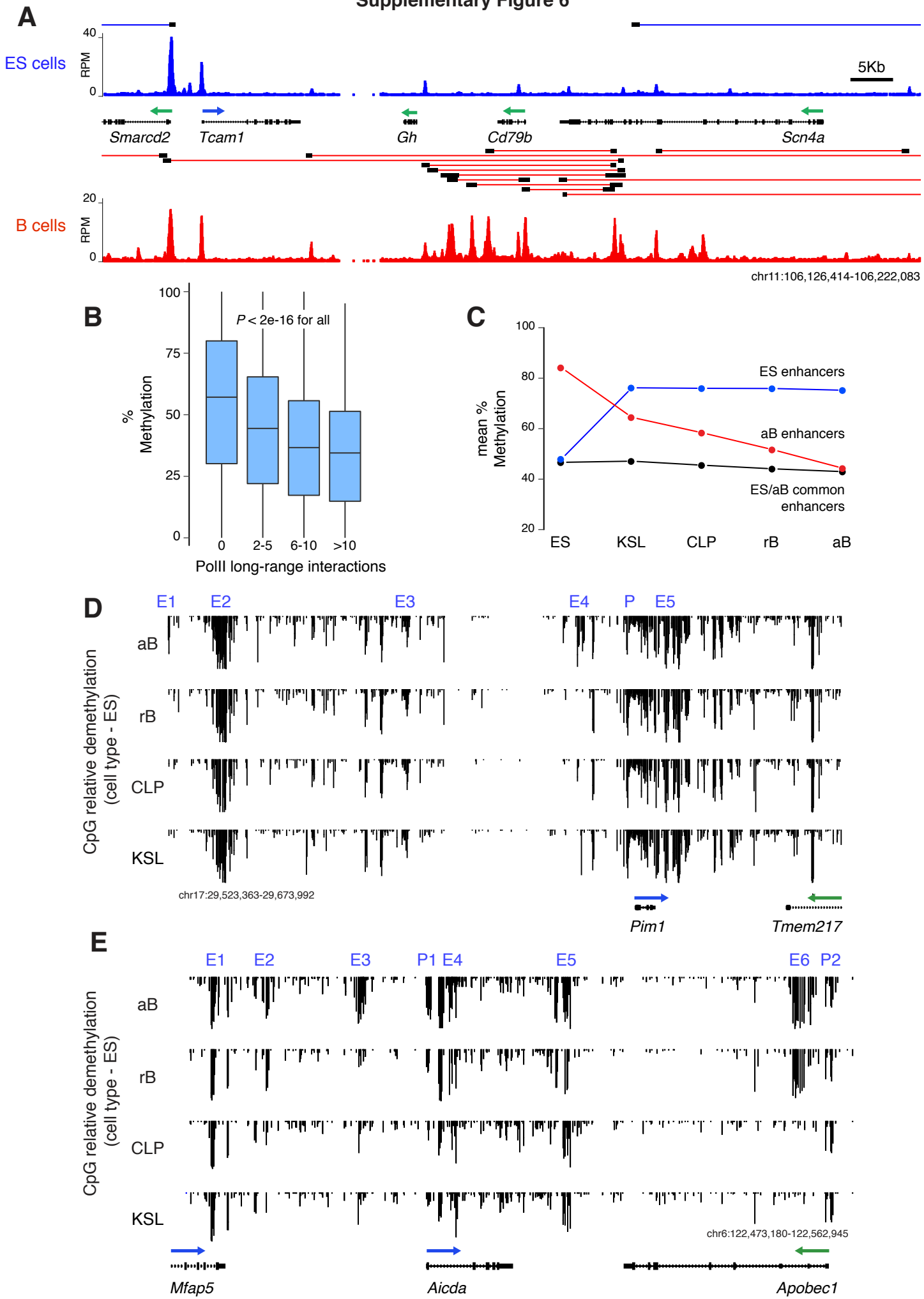



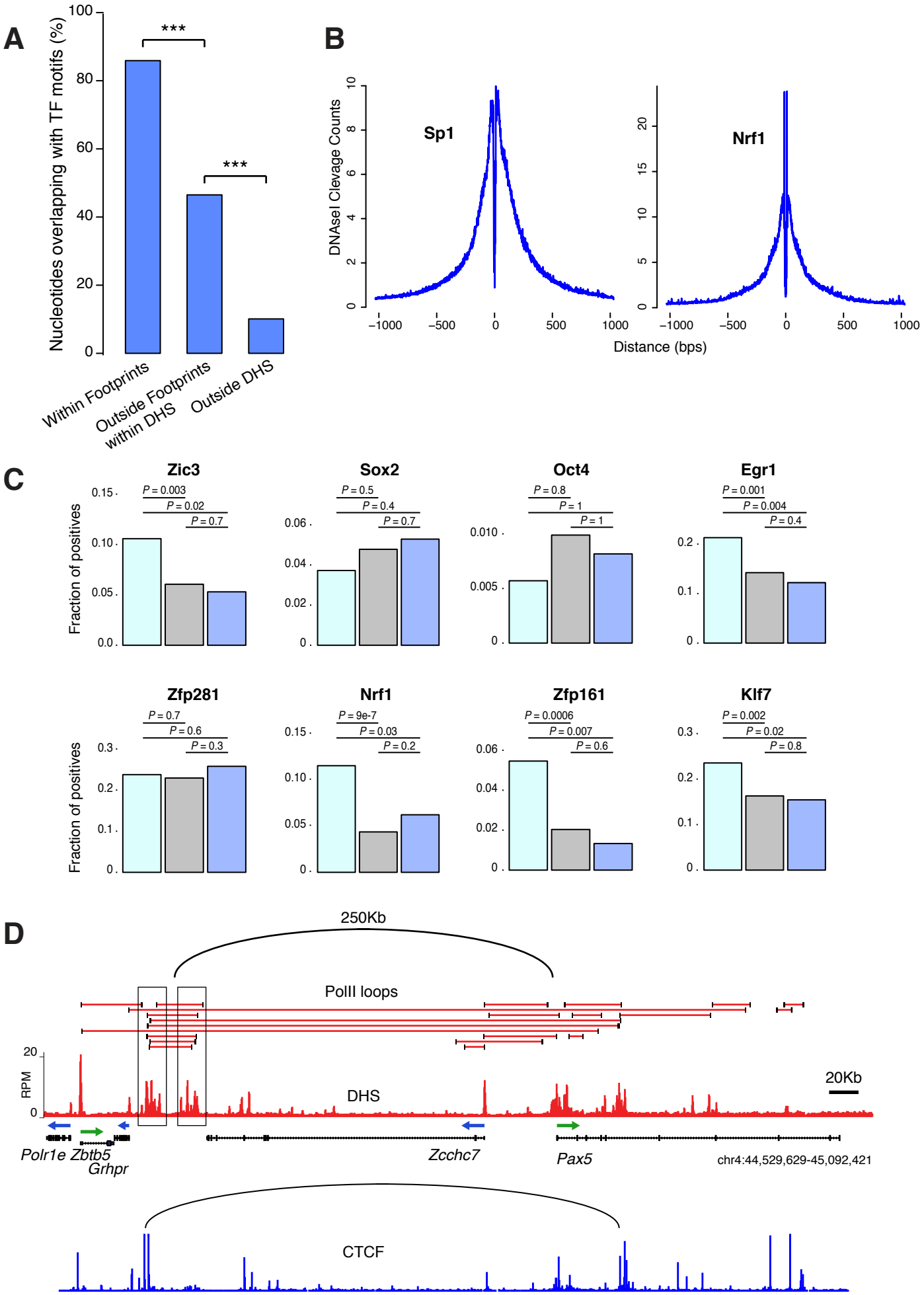
# Supplementary Figure 5



Supplemental Figure 6

# Supplementary Figure 6



Supplemental Figure 7

# Supplementary Figure 7