Cell Reports

Supplemental Information

Low-Cell-Number Epigenome Profiling Aids the Study of Lens Aging and Hematopoiesis

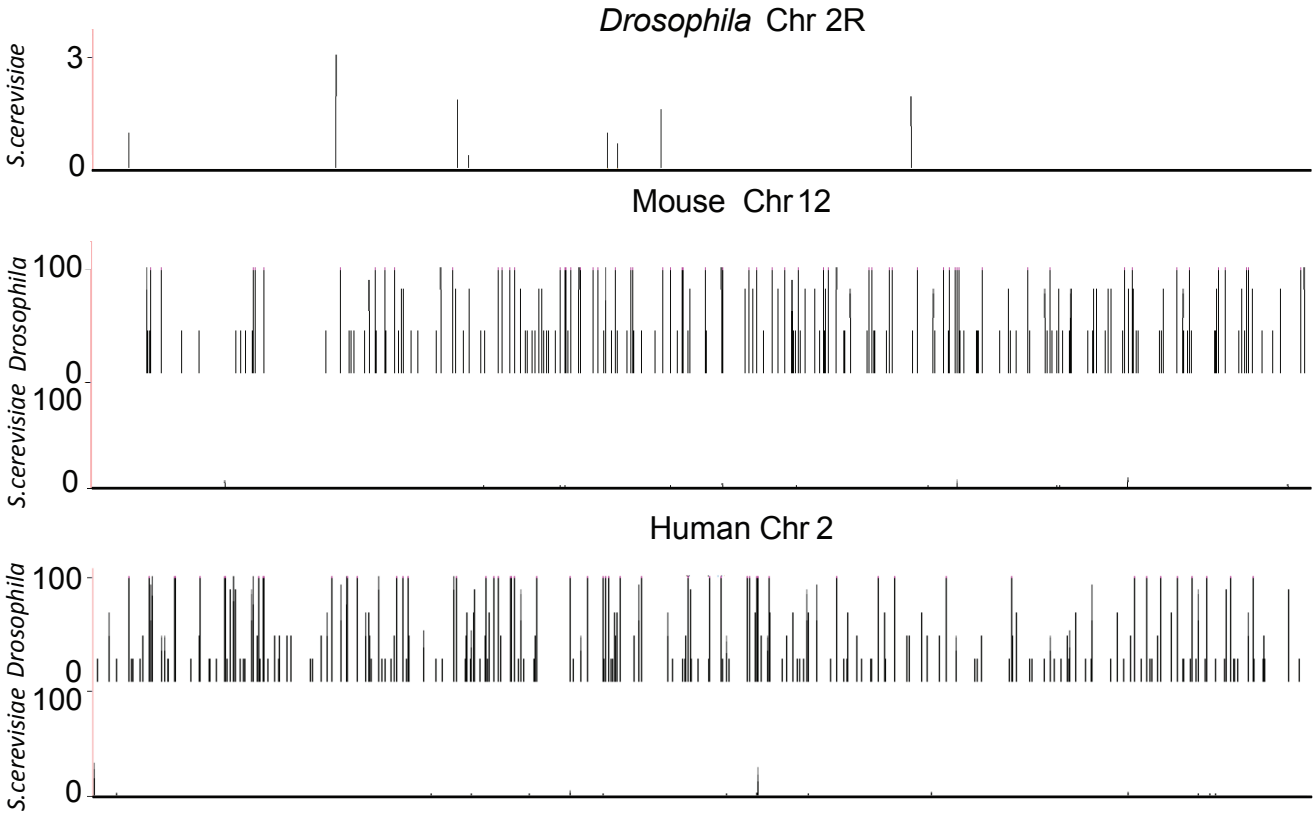
Xiaobin Zheng, Sibiao Yue, Haiyang Chen, Blake Weber, Junling Jia, and Yixian Zheng

Figure S1

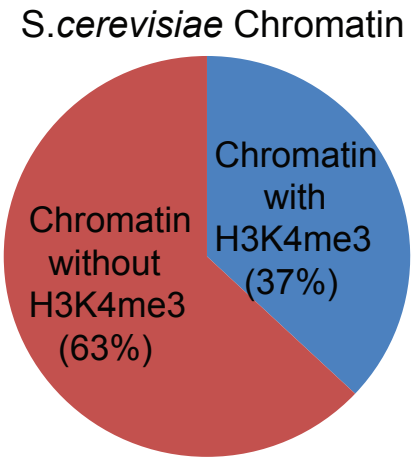
A

	Drosophila Genome	Mouse Genome	Human Genome
Drosophila	100%	50,684 (0.002%)	101,745 (0.003%)
S.cerevisiae	331 (0.0002%)	891 (0.00003%)	1,202 (0.00004%)
E.coli	0	0	8 (0.0000003%)
Synthetic biotin-DNA	0	0	0

B



C



D

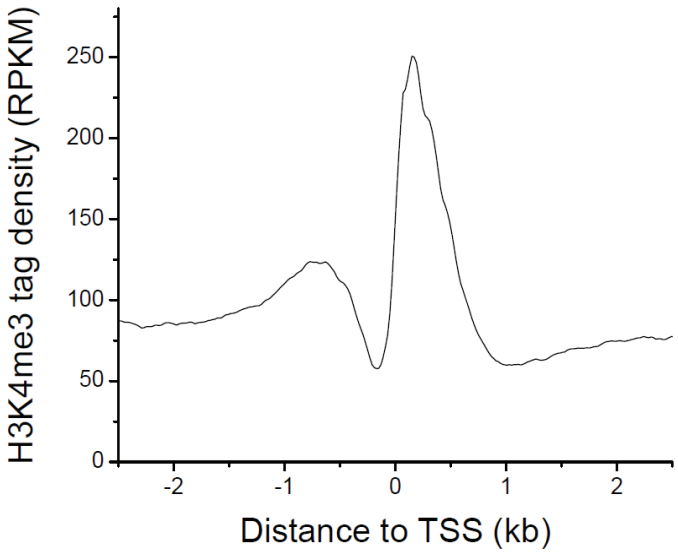


Figure S1. Analyses of carrier DNA (related to Figure 1). **A**. Whereas *Drosophila* genomic DNA sequences can be mapped to many regions in the *Mus musculus* (mouse) or *Homo sapiens* (human) genome, a very small number of *S. cerevisiae* or *E. coli* genomic sequences can be mapped to the genomes of *Drosophila*, mouse, or human. The left column lists the analyzed genomic or synthetic biotin-DNA that maps to the *Drosophila*, mouse, or human DNA. Both total mapped reads and percentages of target genome are shown. Synthetic biotin-DNA sequences are designed to not match to these genomes. Bowtie was used to map (with up to 3 mismatches) the theoretical 50 bp sequence tags covering *Drosophila*, *S. cerevisiae*, or *E. coli* genome with 1bp step-length. **B**. Browser views of theoretical yeast or *Drosophila* DNA sequence tags mapped onto selected chromosomes of indicated genomes. The peak-height corresponds to the number of simulated yeast or *Drosophila* reads mapped to each position. **C-D**. Analyses of yeast chromatin with H3K4me3 (pie chart, C). The normalized H3K4me3 densities are plotted within the 2.5 Kb up and downstream of TSS in yeasts (D).

Figure S2

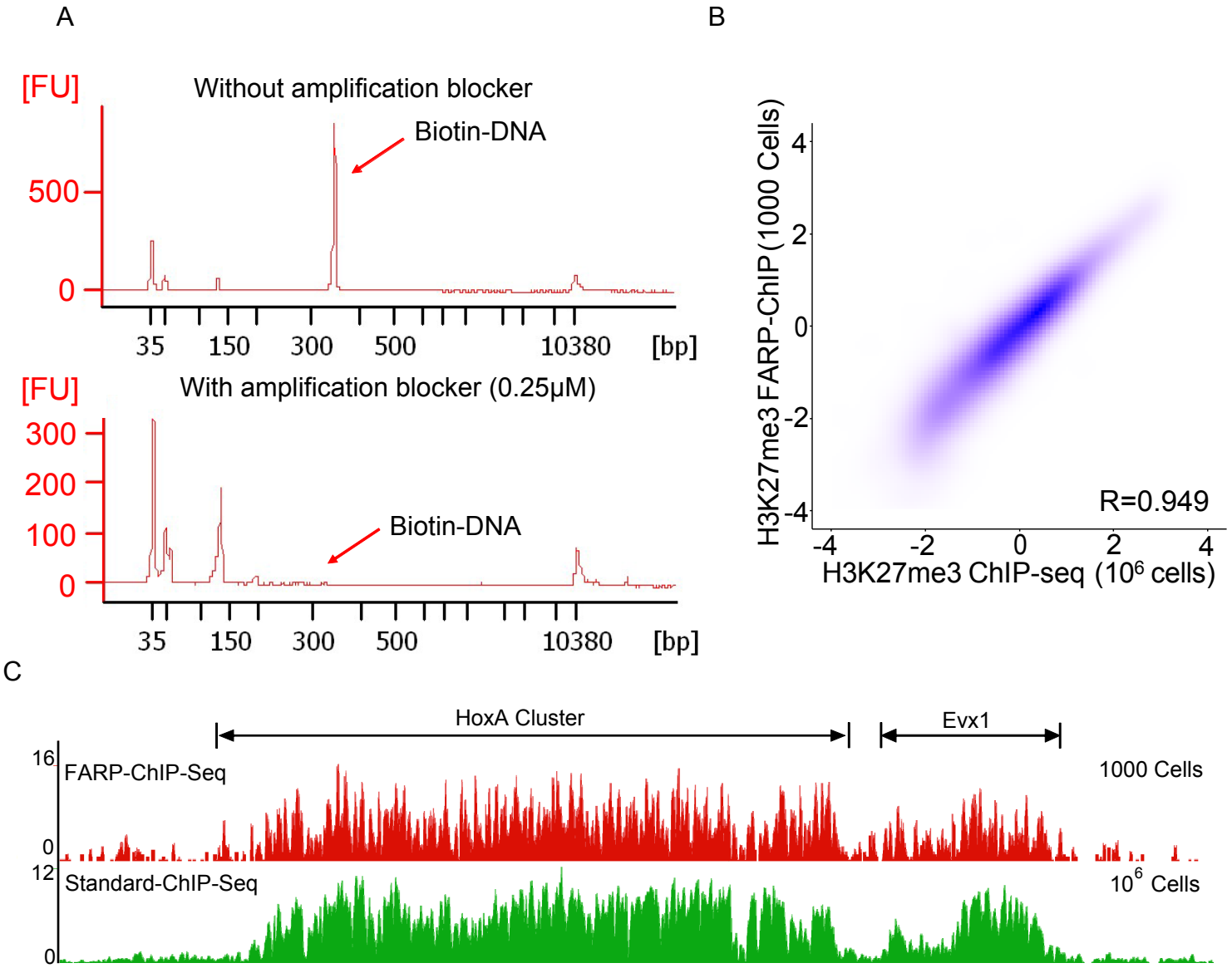


Figure S2. Quality control of FARP-ChIP-seq data (related to Figure 2). **A**. Bioanalyzer plots showing the amplification of the carrier biotin-DNA was reduced by ~99% by the blocker oligo compared to the reaction without the blocker during library building. Red arrows indicate the peak of biotin-DNA after amplification. Note the different scales on the Y-axis. **B**. Contour plot of correlation (Spearman correlation coefficient, R) of H3K27me3 enrichments (plotted as $\log_2$ of the average read density within the 10 Kb up and downstream of TSS) on PRC2-bound promoters mapped by FARP-ChIP-seq in 1000 mESCs and by the standard ChIP-seq in 10^6 mESCs. **C**. H3K27me3 enrichment at the HoxA cluster mapped by FARP-ChIP-seq in 1000 mESCs or by standard ChIP-seq in 1 million mESCs.

Figure S3

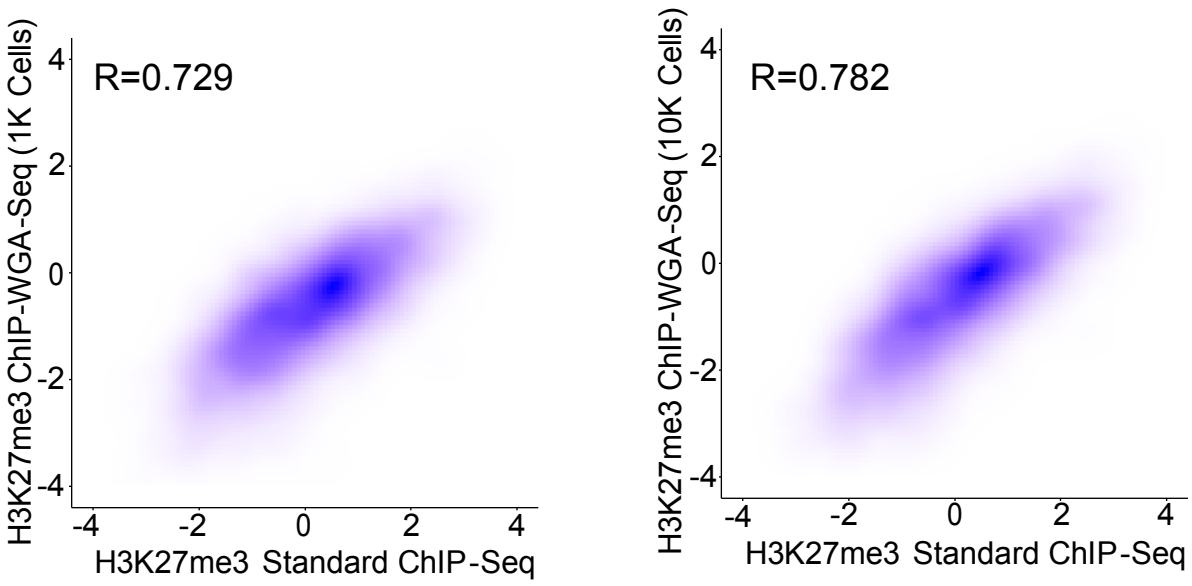


Figure S3. Comparisons of ChIP-WGA-seq of H3K27me3 in 1000 or 10000 mESCs with standard ChIP-seq (related to Figure 3). Contour plots of the correlation (Spearman correlation coefficient, R) of H3K27me3 enrichments on PRC2-bound promoters mapped by ChIP-WGA-seq in 1000 or 10,000 mESCs and by the standard ChIP-seq in 10^6 mESCs. H3K27me3 enrichments are plotted as $\log_2$ of the average read density within the 10 Kb up and downstream of TSS.

Figure S4

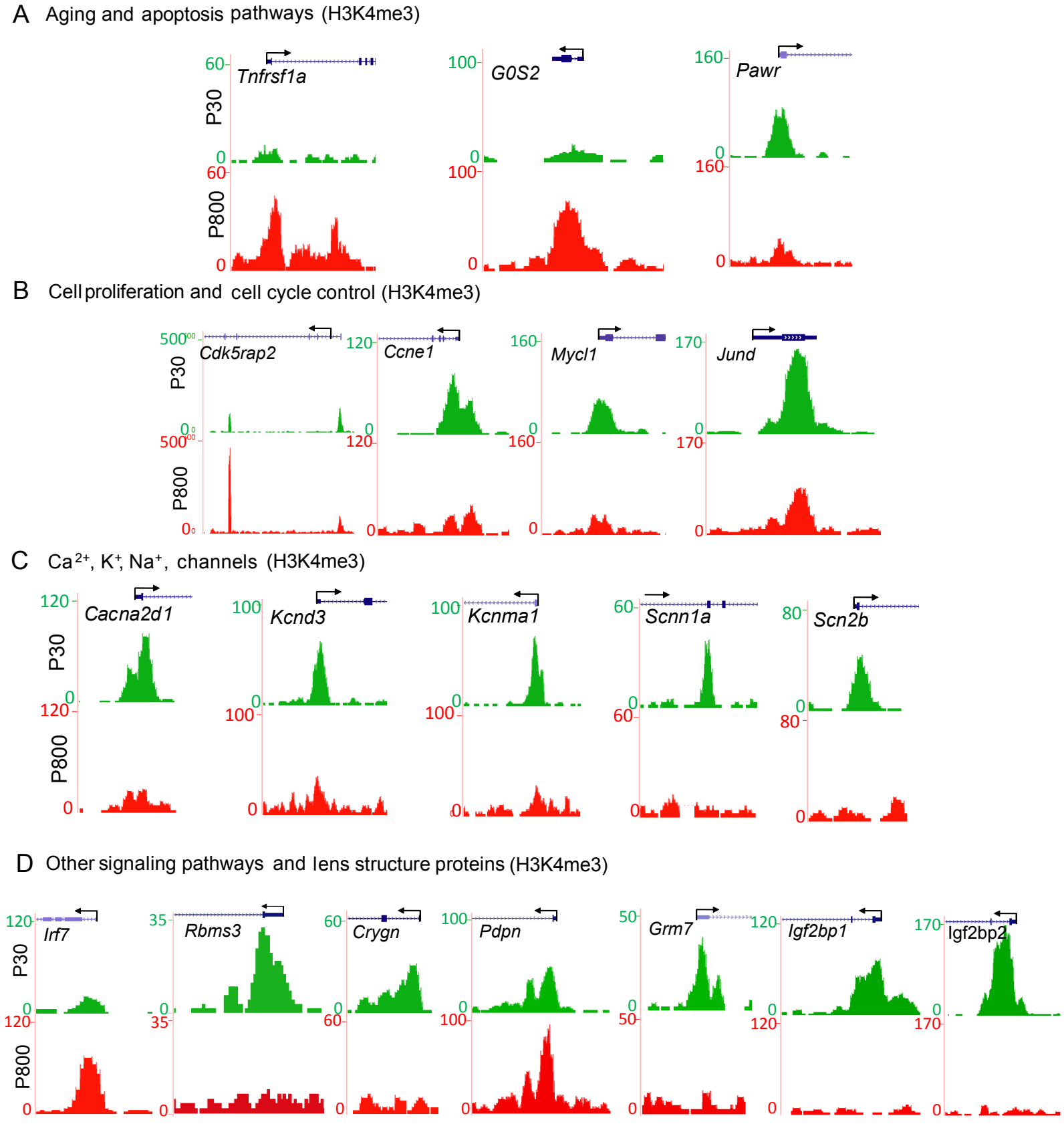


Figure S4. Browser views of aging-associated changes in lens H3K4me3 in different functional categories (related to Figure 4). **A**. Aging and apoptosis pathways. **B**. Cell proliferation and cell cycle control. **C**. Ion channels. **D**. Other signaling pathways and lens structural proteins.

Figure S5

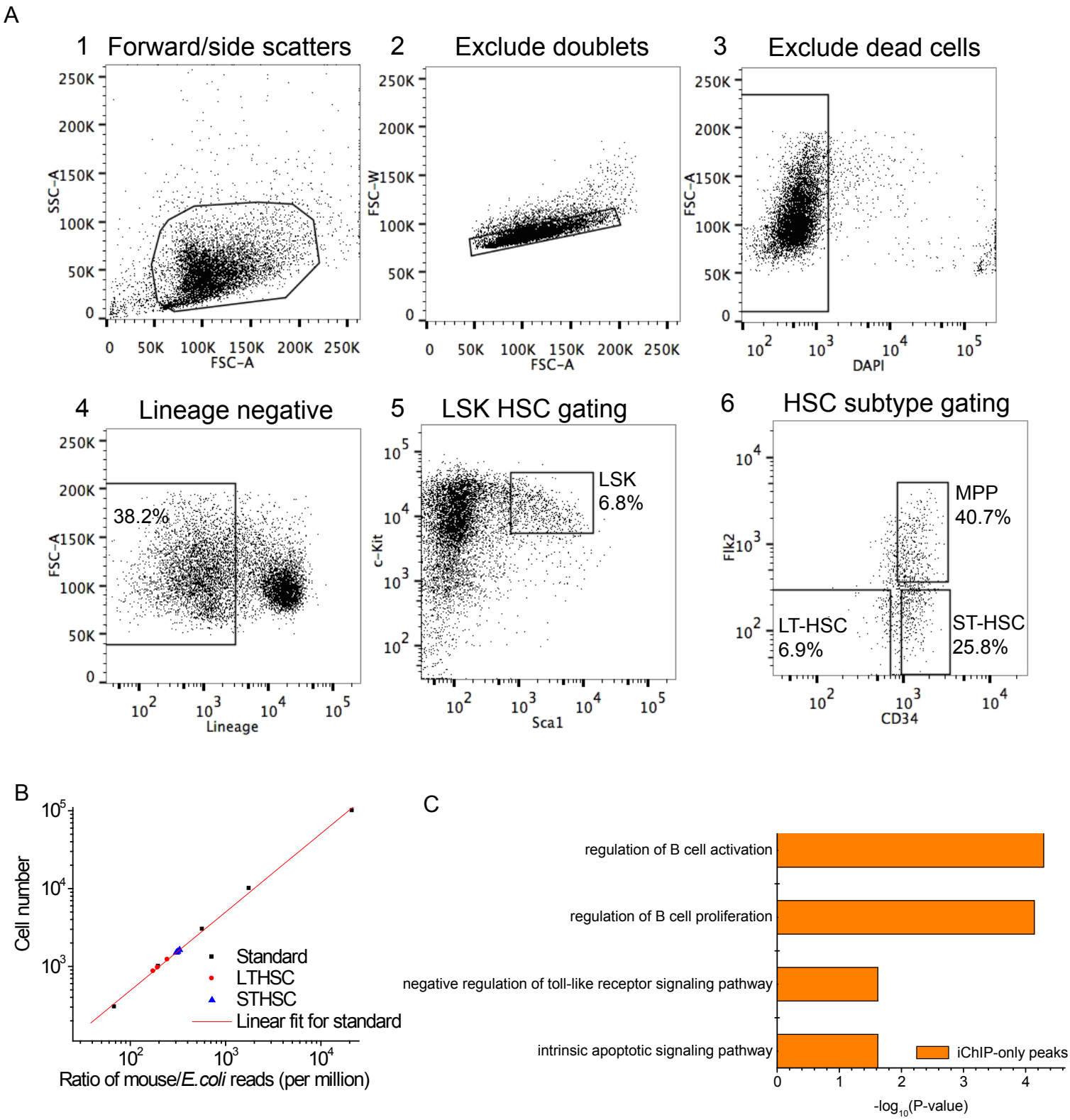


Figure S5. FARP-ChIP-seq of H3K4me3 using HSC subtypes (related to Figure 5). **A**. FACS plots showing the markers and gating used to isolate HSC subpopulations. **B**. Determining the cell number of sorted LT-HSCs and ST-HSCs used for FARP-ChIP-seq by low-depth genomic sequencing with spiked-in bacteria. **C**. GO-terms analysis of H3K4me3 peaks found in iChIP-seq of LT-HSCs.

Figure S6

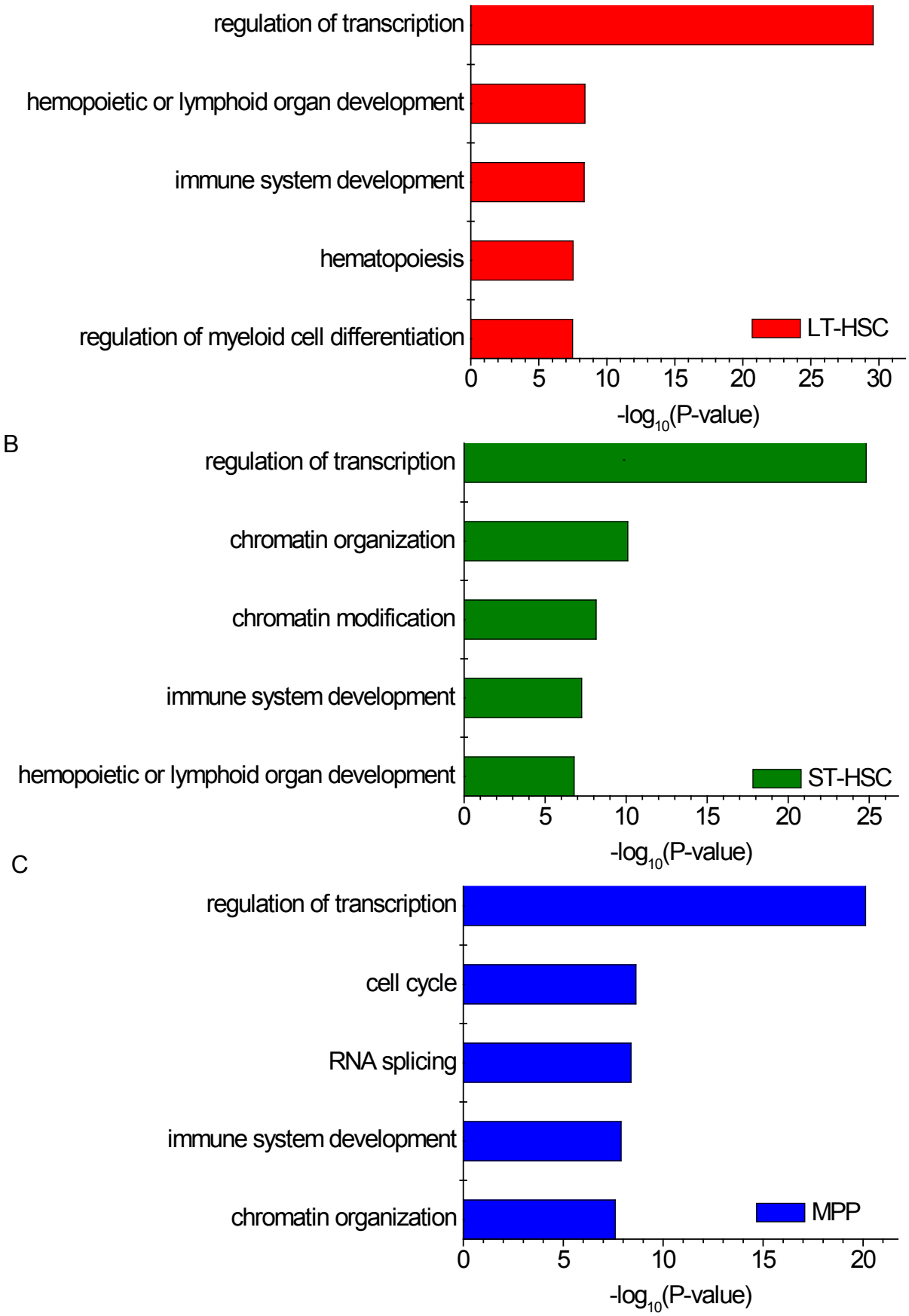


Figure S6. Go-term analyses of the top 800 broadest H3K4me3 peaks mapped by FARP-ChIP-seq in LT-HSCs (A), ST-HSCs (B), and MPPs (C) (related to Figure 6).

Figure S7

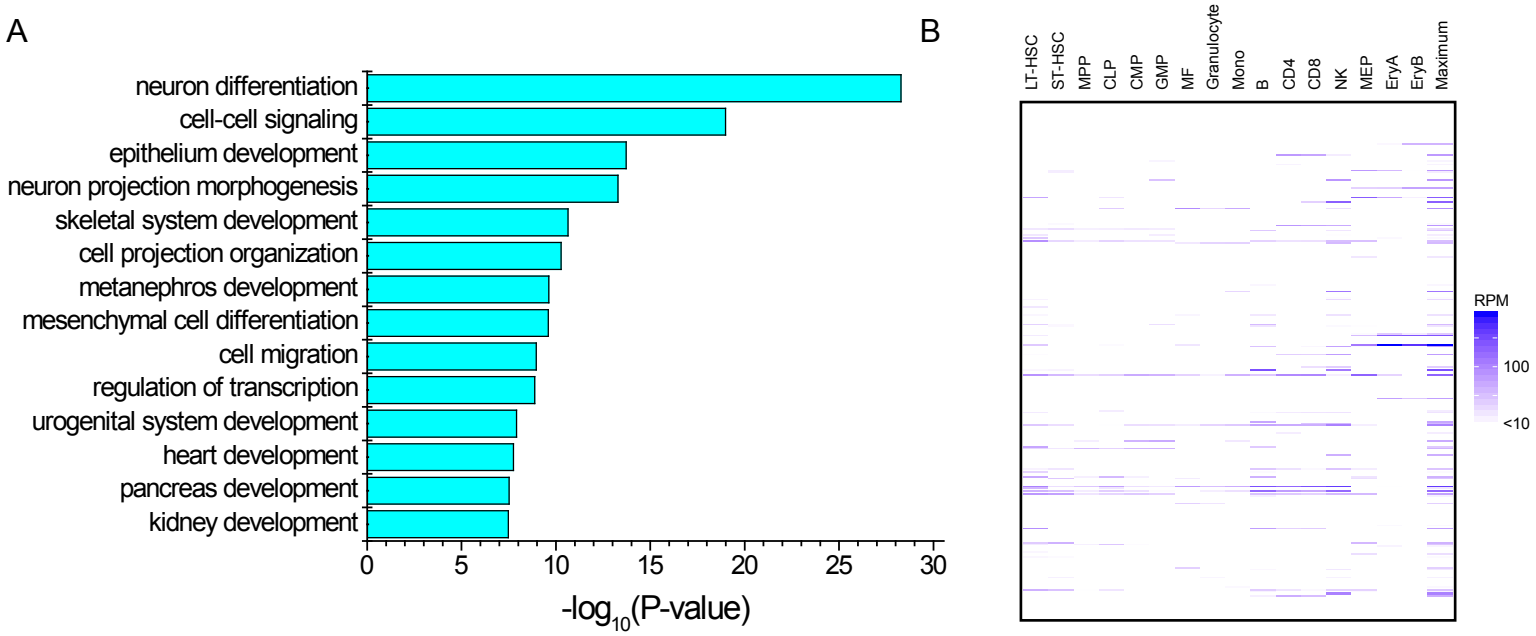


Figure S7. Analyses of H3K4me3 and H3K27me3 bivalent genes found in SP-CD150^{+}-LSK HSCs (related to Figure 7). **A**. GO-term analysis of the H3K4me3 and H3K27me3 bivalent genes. **B**. Heatmap showing the 3'-RNA-seq reads (reads per million, RPM) of the top 300 H3K4me3-enriched bivalent genes in the SP-CD150^{+}-LSK HSCs. Each row represents a gene and each column represents the RNA-seq for the indicated cell type. Only 72 out of 1940 genes were expressed in at least one of the 16 cell types. The last column shows the maximum expression of these genes.

Supplemental Table Legends

Supplemental Table S1. Statistics of the H3K4me3 and ERα libraries compared in Fig. 3A-B (support Figures 1, 2, and 3).

Supplemental Table S2. Numbers and total lengths of all the peaks under all peak-calling p-value thresholds for H3K4me3 and ERα libraries compared in Fig. 3A-B (related to Figures 1, 2, and 3). Benchmark datasets are highlighted in red. Yellow and blue shades are used to distinguish different datasets.

Supplemental Table S3. Peaks called from H3K4me3 mapping of young lens (related to Figure 4). Replicate datasets are pooled before peak calling.

Supplemental Table S4. Promoter peaks showing H3K4me3 changes between young and old lens (related to Figure 5).

Supplemental Table S5. Peaks called from FARP-ChIP-seq H3K4me3 map of LT-HSC, ST-HSC, and MPP (related to Figures 5 and 6). Replicate data are pooled before peak calling.

Supplemental Table S6. H3K4me3/H3K27me3 bivalency in FARP-ChIP-seq map of LT-HSC, ST-HSC, and MPP (related to Figure 7). All genes listed have promoter H3K4me3 enrichment in at least one cell type and H3K27me3 enrichment in at least one cell type.

Supplemental Table S7. GO-term analysis of H3K4me3/H3K27me3 bivalent genes in SP-CD150^{+}-LSK HSCs by Sun et al (see reference in the main text) (related to Figure 7).

Supplemental Experimental Procedures:

Standard ChIP-seq

Standard ChIPs were performed according to conditions suggested by manufacturers of ChIP-grade antibodies to H3K27me3 (Millipore, 07-449) and H3K4me3 (Cell Signaling, 9751). Briefly, formaldehyde was added to mESCs to crosslink proteins to DNA and the cells were lysed in 200 μl lysis buffer (50 mM HEPES, pH 7.5, 140 mM NaCl; 1 mM EDTA; 1% Triton X-100; 0.1% sodium deoxycholate; 0.1% sodium dodecyl sulfate). Cell lysates were sonicated using a Bioruptor ultrasonic cell disruptor (Diagenode) to shear genomic DNA to an average fragment size of 150 to 250 bp and ChIP was performed following protocol provided by Millipore ChIP Assay kit (Millipore, 17-295). After washing and de-crosslinking, the precipitated DNA was purified using a QIAquick PCR purification kit (QIAGEN).

RP-ChIP-seq

500 or 2000 mESCs or one mouse ocular lens (dissected from P30 and P800 C57/BL6J mice) were mixed with ~$5x10^7$ *S. cerevisiae* and fixed in 1% formaldehyde for 10 minutes at room temperature. After fixation and washes, the mixtures were sonicated with a 1/8 inch probe for a total of 15 min at 9 watts using a tip sonicator (Misonix sonicator 3000) to obtain genomic DNA fragments with an average size of 150-250 bp. Buffers and buffer volumes were according to standard ChIP procedures. After ChIP, libraries were prepared following Illumina True-Seq protocol with 10 PCR amplification cycles. The libraries were sequenced using an Illumina HiSeq2000 with single-end and 50 bp read length.

FARP-ChIP-seq

mESCs (500 or 1000) or sorted HSC subtypes were mixed with ~$5x10^8$ DH5α *E.coli* and fixed in 1% formaldehyde for 10 min at room temperature. After fixation and washes, the mixtures were sonicated with a 1/16-inch probe for 15 min at 3 watts using a tip sonicator (Misonix sonicator 3000) to obtain genomic DNA fragments with an

average size of 150-250bp. Buffers and buffer volumes were according to standard ChIP procedures. The lysates were incubated with 2 μl H3K27me3 antibody (Millipore, 07-449) or 1 μl H3K4me3 antibody (Cell Signaling, 9751S) overnight at 4^{o}C. At the meanwhile, 60 μl protein G magnetic beads (Life Technologies, 10004D) and 10 μl M-280 streptavidin beads (Life Technologies, 11206D) for each ChIP were prepared for the next step. To block any non-specific chromatin absorption, protein G beads and M-280 streptavidin beads were pre-incubated with ~5x10^{8} fixed and sonicated *E.coli* lysate overnight at 4^{o}C, respectively. After blocking step, 5 ng carrier biotin-DNA was then coupled to 10 μl M-280 streptavidin beads according to manufacturer's instruction (Life Technologies, 11206D). These treated protein G and streptavidin beads were combined and used to ChIP H3K27me3 or H3K4me3 in the sonicated *E.coli*+mESC or *E.coli*+HSC lysates for 4 h at 4^{o}C. The beads were then recovered using a magnetic stand. All washing, de-crosslinking were performed following standard ChIP procedures. After de-crosslinking, the precipitated genomic DNA and biotin-DNA were purified using Agencourt AMPure XP beads (Beckman Coulter, A63881). Library building steps were performed following Illumina True-Seq protocol with 12 PCR amplification cycles. After end repair, A-tailing, and adapter ligation steps, 0.25 μM blocker oligo was added at the final library amplification step.

The carrier biotin-DNA sequence:

Biotin-
5'atattaacgcttacaatttaggtggcacttttcggggaaatgtgcgcggaacccctatttgtttatttttctaaatacattcaaatatg
tatccgctcatgagacaataaccctgataaatgcttcaataatattgaaaaaggaagagtatgagtattcaacatttccgtgtcgccc
ttattccctttttgcggcattttgccttcctgttttt 3'

The amplification blocker sequence

5'
A*C*A*AATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTAA-3 carbon spacer 3'
* indicate the first three phosphorothioated DNA bases.

Isolation of HSC subtypes for FARP-ChIP-seq

Femora and tibiae were dissected from one C57BL/6J male mouse (8 weeks old) and bone marrow cells were flushed with FACS buffer (PBS with 0.3% BSA and 2 mM EDTA). The cells were enriched with CD117/c-Kit MicroBeads following the guidelines (Miltenyi Biotec, #130-091-224). c-Kit enriched cells were stained with lineage cell detection cocktail-biotin (Miltenyi Biotec, #130-092-613) for 20 min at 4^{o}C. Cells were then stained with DAPI, anti-Biotin, c-Kit, Sca1, CD34, and Flk2 antibodies (Biolegend) for 20 min followed by sorting with FACSAria^{TM} III cell sorter (BD Bioscience). The cell populations collected were: LT-HSC: Lin^{-}, c-Kit^{+}, Sca-1^{+}, CD34^{-}, Flk2^{-}; ST-HSC: Lin^{-}, c-Kit^{+}, Sca-1^{+}, CD34^{+}, Flk2^{-}; MPP: Lin^{-}, c-Kit^{+}, Sca-1^{+}, CD34^{+}, Flk2^{+}.

Determining the number of HSC subtypes used in FARP-ChIP-seq

Different numbers of BM cells, 300, 1000, 3000, 10,000 and 100,000, were mixed with 5x10^{8} bacteria and fixed in 1% formaldehyde for 10 min at room temperature. After fixation and washes, the mixtures were sonicated using the same conditions for HSC subtypes. The sample was then used to build library for low-depth genomic sequencing. A cell number standard was determined using the number of reads from mouse and bacteria by linear regression. To prevent the loss of sorted LT-HSCs, ST-HSCs, and MPPs during fixation and wash steps, each subtype was mixed with 5x10^{8} bacteria and fixed followed by washes to remove the fixative. After sonication, 5% of each sample was used for the same low-depth genomic sequencing as above. The mouse and bacteria reads obtained in each sample were used to determine the number of HSC subtypes based on the cell standard.

Quality-filtering and mapping of the ChIP-seq data

For RP-ChIP-seq and FARP-ChIP-seq libraries, reads were filtered using the quality of multiplexing indices before mapping to prevent contaminations by erroneous de-multiplexing. Reads with average index quality below 30 were discarded (Kircher et al., 2012). The reads were then mapped to mouse genome (mm9) by Bowtie allowing 2 mismatches per read. Only uniquely mapped reads were retained. The mapped reads were further mapped to carrier genome or biotin-DNA sequences by Bowtie allowing 3 mismatches, and reads mapped to the carrier genome or biotin-DNA were discarded.

Precision-recall analysis for comparisons between methods

To compare different ChIP-seq methods, we first defined the following sets of H3K4me3 peaks. *A* is the set of H3K4me3 peaks called from the standard ChIP-seq data from over 10^6 cells by the MACS software with p-value threshold of 10^{-5}. *B* is the top 40% of peaks in *A*, sorted by the p-value. *C* is the set of H3K4me3 peaks called from our ChIP-seq data by MACS (Zhang et al., 2008) with a certain p-value threshold. *D* is the top 40% of peaks in *C*, sorted by *p*-value. *E* is the set of peaks in *C* that overlaps *B*, and *F* is the set of peaks in *A* that overlaps *D*. Precision is defined as the size of *F* divided by the size of *D*, which represents the percentage of top H3K4me3 peaks by RP-ChIP-seq or FARP-ChIP-seq that are covered by the standard ChIP-seq peaks. Recall is defined as the size of *E* divided by the size of *B*, which represents the percentage of top H3K4me3 peaks by the standard ChIP-seq that are covered by our RP-ChIP-seq or FARP-ChIP-seq peaks. The precision and recall change in opposite directions when the *p*-value threshold used in calling the set *C* is changed. We draw a precision-recall curve by changing the *p*-value threshold from 0.05 to 10^{-5}. Curves for other published ChIP-seq methods were drawn in the same way. The H3K4me3 data by LinDA-seq from 1000 or 10,000 cells were compared to the H3K4me3 ChIP-seq data from 10^6 cells published in the same study. The ERα data obtained by using mRNA and histone as carriers from 10,000 MCF-7 cells were compared to the public ERα ChIP-seq data in 10^6 MCF-7 cells (Zwart et al., 2011). The definitions for precision and recall can be expressed as equations below.

$$A = \{\text{Standard ChIP-seq H3K4me3 Peaks}\}$$

$$B = \{\text{Top 40\% Peaks of A}\}$$

$$C = \{\text{RP-ChIP-seq H3K4me3 Peaks}\}$$

$$D = \{\text{Top 40\% Peaks of C}\}$$

$$E = \{\text{Peaks in C that overlap B}\}$$

$$F = \{\text{Peaks in A that overlap D}\}$$

$$\text{Precision} = \frac{\text{Size of F}}{\text{Size of D}}$$

$$\text{Recall} = \frac{\text{Size of E}}{\text{Size of B}}$$

Promoter correlation, whole-genome correlation, and ROC analysis

Promoter correlations for H3K4me3 enrichments are plotted as $\log_2$ of the average read density within 2 Kb up and downstream of TSS. Promoter correlations for H3K27me3 enrichments are plotted as $\log_2$ of the average read density within 5 Kb up

and downstream of TSS. We use different window size for H3K4me3 and H3K27me3 analyses because H3K4me3 has narrower peaks. The whole-genome correlation coefficients were calculated using non-overlapping 4-Kb windows covering the whole genome for H3K4me3 and 10 Kb windows for H3K27me3. For the H3K4me3 ROC analysis, we use top 25,000 2-Kb windows as "True" to mimic the ~25,000 peaks having ~50M genome coverage from the benchmark dataset (table S2). For H3K27me3 ROC analysis, we used top 25,000 5-Kb windows as "True" because H3K27me3 has broader peaks than H3K4me3.

Analyses of RP-ChIP-seq in young and old mouse lens

H3K4me3 peaks of young and old mouse lens epithelial cells are called by MACS with *p*-value threshold of 10^{-5} and parameter "--llocal=100000" to enable broad peak calling. Peaks are sorted based on their width and GO-terms for broad peaks were determined by DAVID (Huang da et al., 2009). Then the four sets of peaks (two replicates for young and two replicates for old) were merged to obtain a complete peak set, and the numbers of reads are counted for each peak region. We used edgeR (Robinson et al., 2010) to call differentially enriched H3K4me3 peaks between young and old data using standard edgeR procedures.

Analyses of FARP-ChIP-seq of HSC subtypes

H3K4me3 peaks of LT-HSC, ST-HSC, or MPPs are called by MACS with *p*-value threshold of 10^{-5} and parameter "llocal=100000" to enable broad peak calling. Then the six sets of peaks (two replicates for LT-HSC, ST-HSC, and MPP each) were merged to obtain a complete peak set, and the numbers of reads were counted for each peak region. We used edgeR (Robinson et al., 2010) to call differentially enriched H3K4me3 peaks between LT-HSC versus ST-HSC and between ST-HSC versus MPP data using standard edgeR procedure. GO-term was determined by DAVID (Huang da et al., 2009). For bivalency analysis, the reads densities for H3K4me3 and H3K27me3 were calculated within -2 kb to +2 kb regions around each promoter. We used 2 kb region to avoid reads from nearby promoters. H3K4me3 and H3K27me3 positive promoters were defined by promoter read density cutoffs that best separate the positive and negative clusters.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research *37*, 1-13.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research *40*, e3.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

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 biology *9*, R137.

Zwart, W., Theodorou, V., Kok, M., Canisius, S., Linn, S., and Carroll, J.S. (2011). Oestrogen receptor-co-factor-chromatin specificity in the transcriptional regulation of breast cancer. The EMBO journal *30*, 4764-4776.