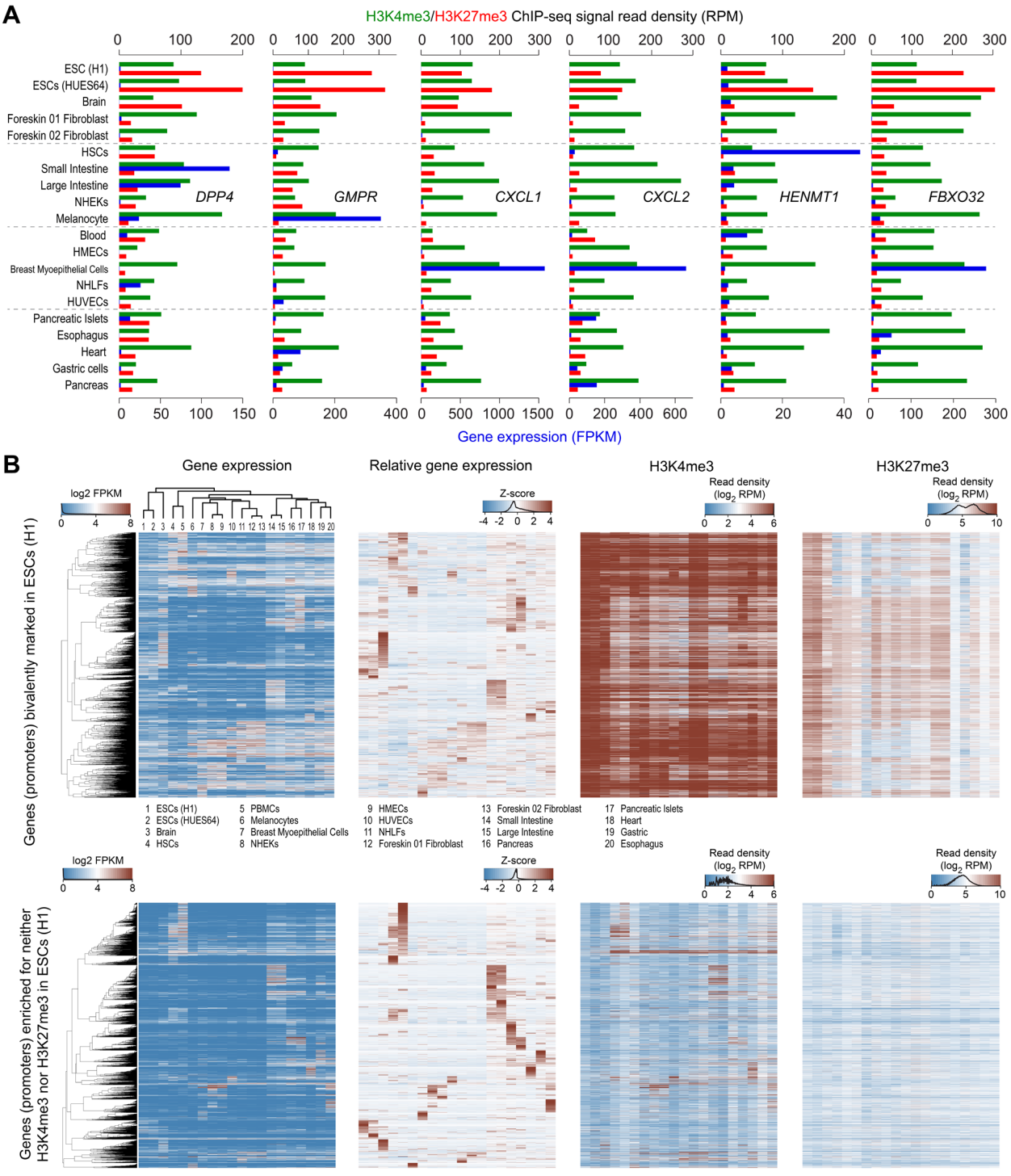
Supplemental Material

Decoding the function of bivalent chromatin in development and cancer

Dhirendra Kumar[1], Senthilkumar Cinghu[1], Andrew J Oldfield[1,2], Pengyi Yang[1,3], and Raja Jothi[1]

[1]Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA
[2]Present address: Institute of Human Genetics, CNRS, University of Montpellier, Montpellier, 34396, France
[3]Present address: Charles Perkins Centre and School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia

Table of Contents



C

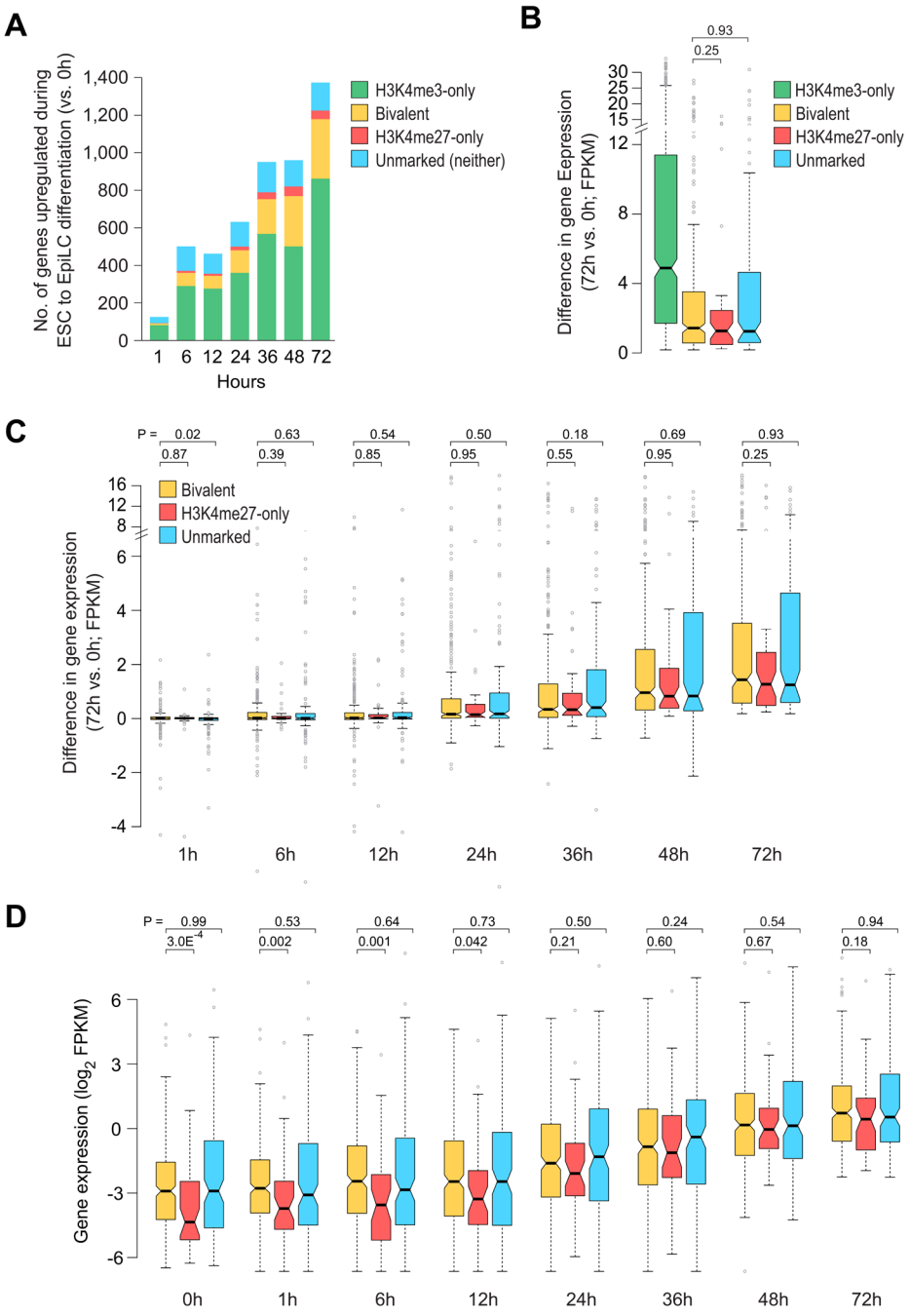
Cell type ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
All genes	0.69	0.68	0.63	0.71	0.66	0.63	0.61	0.64	0.65	0.63	0.69	0.68	0.65	0.59	0.67	0.61	0.70	0.67	0.67	0.71
Bivalent	0.48	0.39	0.52	0.63	0.51	0.52	0.53	0.62	0.48	0.60	0.62	0.62	0.61	0.51	0.67	0.54	0.64	0.61	0.58	0.64

Supplemental Fig. S1. H3K4me3, observed at bivalent promoters in ESCs, persists in nearly all cell types irrespective of gene expression.

(A) Expression (FPKM) of genes (blue), shown in Fig. 1A, in various cell types. FPKM, fragments (RNA-seq) per kilobase per million mapped reads. Also shown are ChIP-seq read densities (RPM) for H3K4me3 (green) and H3K27me3 (red) at gene promoters in various cell types. RPM, reads per million mapped reads. Promoters were defined as the region spanning TSS ± 500 bp for H3K4me3 and TSS ± 2 kb for H3K27me3.

(B) *Top-row (left to right)*: Unsupervised hierarchical clustering of bivalently marked genes in human ESCs (H1) (y-axis) based on their expression across various cell types (x-axis). Relative gene expression (row-normalized), H3K4me3 and H3K27me3 ChIP-seq read density at the promoters of genes (+/- 500 bp of TSS for H3K4me3 and +/- 2 kb of TSS for H3K27me3) shown in the top-left panel across various cell types (ordering of genes and cell types is same as in the top-left panel). For comparison purposes, bottom panels show data for (unmarked) genes whose promoters are enriched for neither H3K4me3 nor H3K27me3 in ESCs (H1). Ordering of cell types (x-axis) is same as in the top-left panel.

(C) Pearson's correlation between gene expression and promoter H3K4me3 levels in each of the 20 cell types, as shown in B.

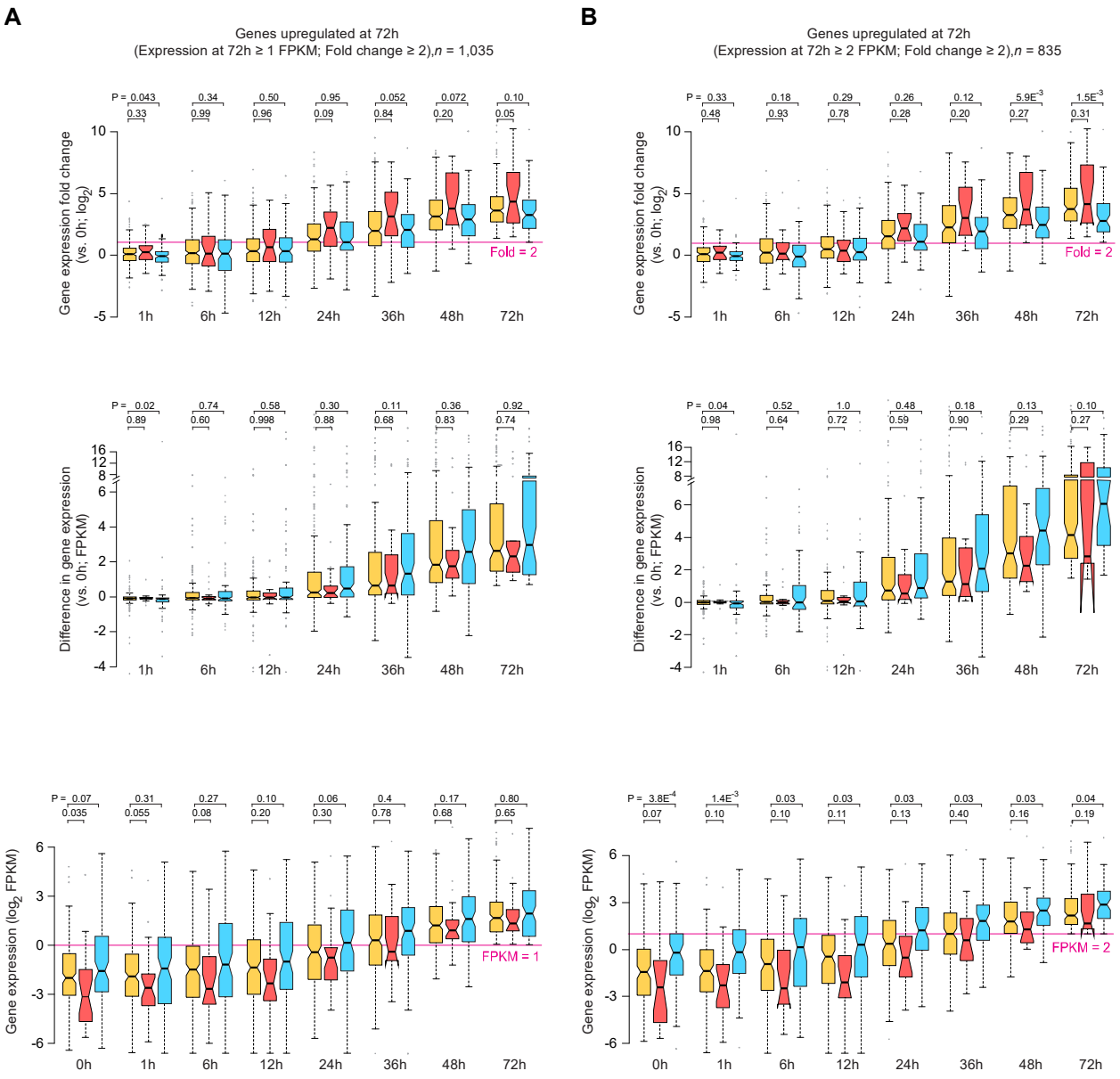


Supplemental Fig. S2. Bivalent chromatin does not poise genes for rapid activation.

(A) Barplot showing the distribution of the number of genes upregulated during differentiation of naïve ESC (0h) to EpiLC (72h). Genes grouped based on their chromatin states in naïve ESCs (0h).

(B) Boxplot showing the distribution of absolute differences (delta) in gene expression (72h vs 0h) for genes upregulated in EpiLCs. Genes grouped based on their chromatin states in naïve ESCs (0h).
(C) Boxplot showing the distribution of absolute differences (delta) in gene expression over time (compared to 0h) for genes upregulated in EpiLCs. Genes are grouped based on their chromatin states in naïve ESCs (0h).
(D) Same as in H, except that the boxplot shows the distribution of absolute gene expression over time for genes upregulated in EpiLCs (72h vs 0h).

All the *P* values were calculated using two-sided Wilcoxon rank-sum test.

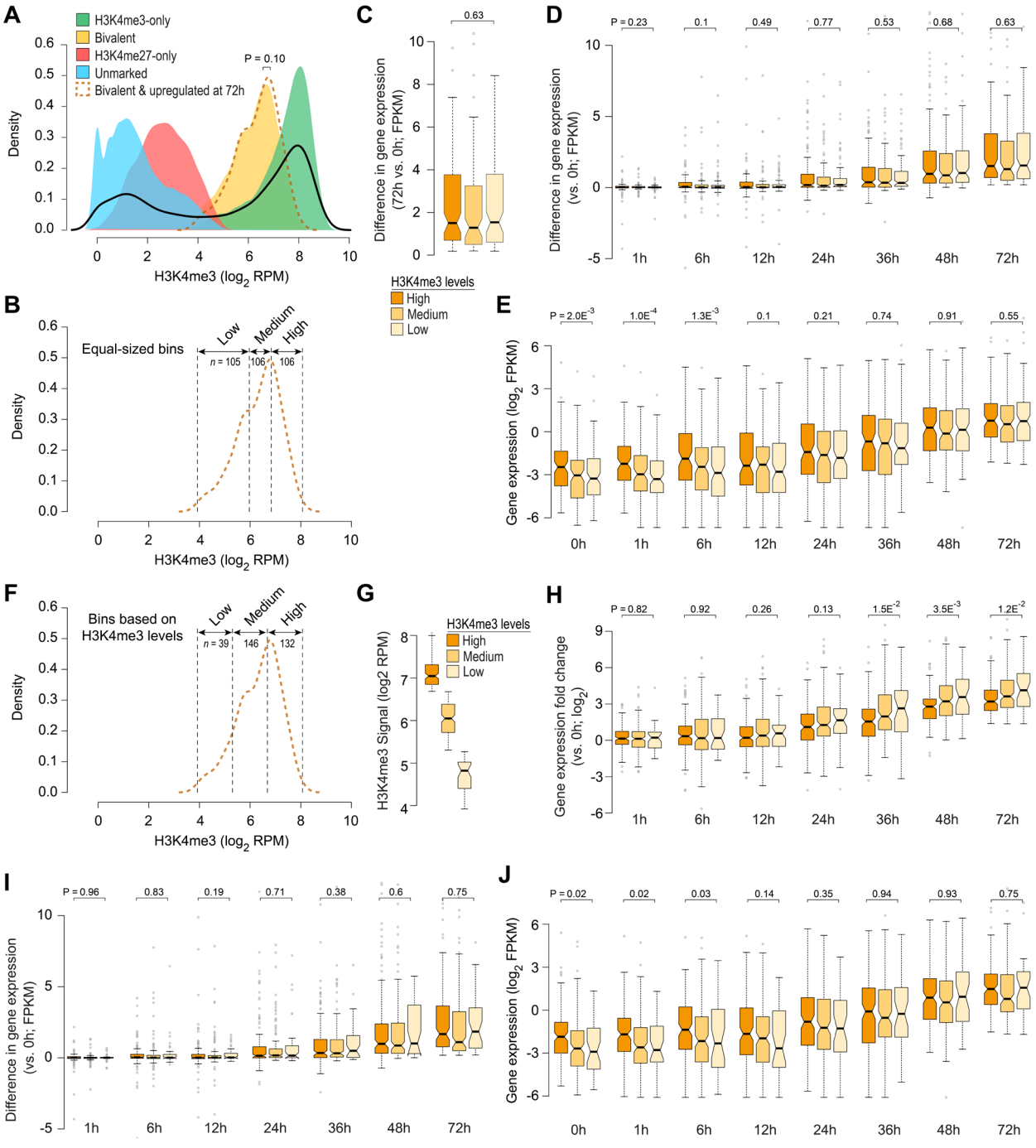


Supplemental Fig. S3. Expression changes of genes upregulated during ESC (0h) to EpiLC (72h) differentiation.

(A) *Top*: Boxplot showing the distribution of gene expression fold changes over time for genes upregulated in EpiLCs (72h vs 0h; expression at 72h ≥ 1 FPKM, fold change ≥ 2). *Middle*: Boxplot showing the distribution of absolute differences (delta) in gene expression over time for the same set of upregulated genes. *Bottom*: Boxplot showing the distribution of absolute gene expression over time for the same set of upregulated genes.

(B) Same as in A, except that genes upregulated in EpiLCs are defined using a more stringent criteria (72h vs 0h; expression at 72h ≥ 2 FPKM, fold change ≥ 2).

All the *P* values were calculated using two-sided Wilcoxon rank-sum test.



Supplemental Fig. S4. Activation of bivalent genes with higher levels of H3K4me3 is neither greater nor faster than that of those with lower levels of H3K4me3.

(A) Density plot showing the distribution of H3K4me3 signal at various classes of gene promoters (+/-500bp of TSS) in naïve ESCs (0h).

(B) Density plot showing the distribution of ESC (0h) H3K4me3 signal at the promoters of bivalent genes upregulated in EpiLCs (72h vs 0h; *n* = 1,372). Also shown are cut-offs (dotted

vertical lines) used to divide this group of genes into three equal-sized bins (low, medium, high).

(C) Boxplot showing the distribution of absolute differences (delta) in gene expression (72h vs 0h) for genes upregulated in EpiLCs. Genes are grouped based on high, medium, or low H3K4me3 signal, as defined in B.

(D) Boxplot showing the distribution of absolute differences (delta) in gene expression over time (compared to 0h) for genes upregulated in EpiLCs. Genes are grouped based on H3K4me3 enrichment at promoters in naïve ESCs (0h), as defined in B.

(E) Same as D, except that the boxplot shows the distribution of absolute gene expression over time for genes upregulated in EpiLCs (72h vs 0h).

(F) Same as in B, except that the genes are divided into three bins (low, medium, high) based on their H3K4me3 levels in ESCs (0h).

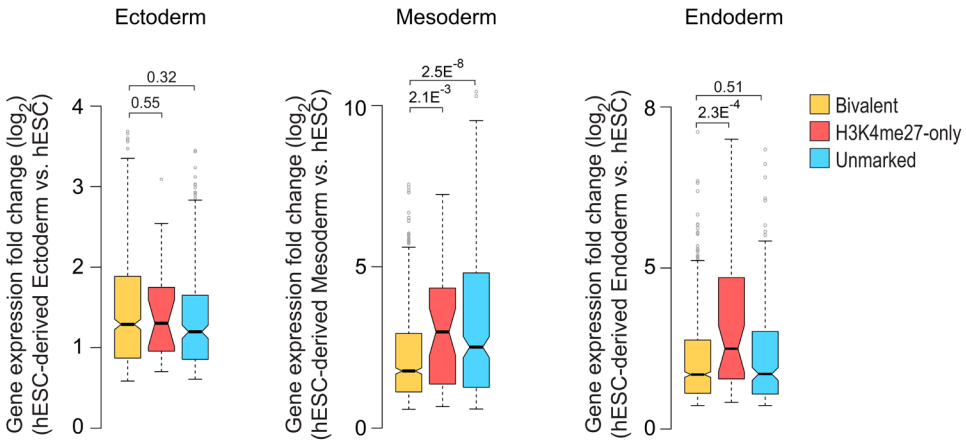
(G) Box plots showing the distribution of H3K4me3 ChIP-seq read densities for genes classes shown in F

(H) Boxplot showing the distribution of gene expression fold changes over time (72h to 0h) for bivalent genes upregulated in EpiLCs (72h vs 0h). Genes are grouped based on high, medium, or low H3K4me3 signal, as defined in F.

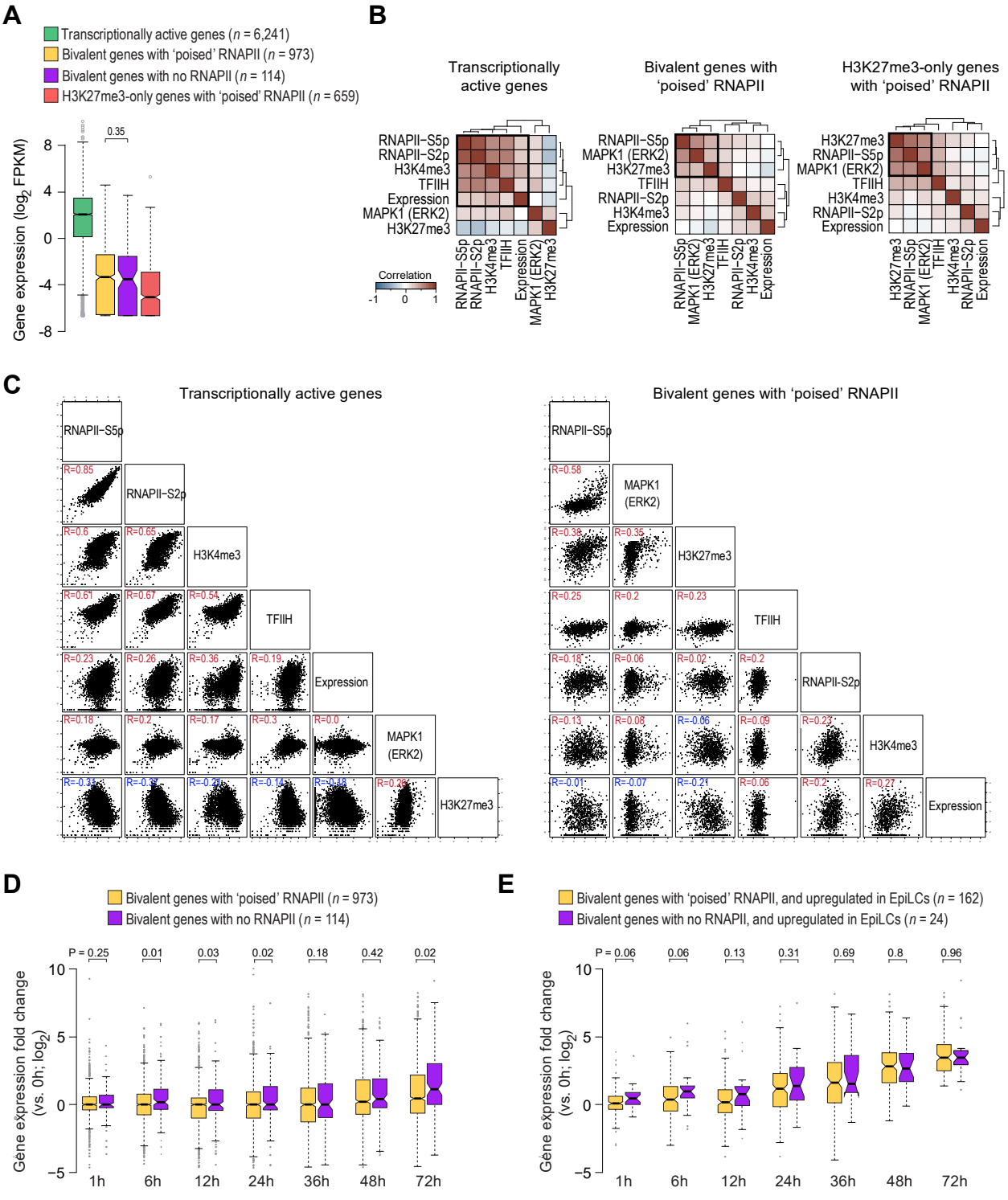
(I) Same as in H, except that the boxplot shows the distribution of absolute differences (delta) in gene expression (72h vs 0h) for bivalent genes upregulated in EpiLCs.

(J) Same as in H, except that the boxplot shows the distribution of absolute gene expression over time for bivalent genes upregulated in EpiLCs (72h vs 0h).

All the *P* values were calculated using two-sided Wilcoxon rank-sum test.

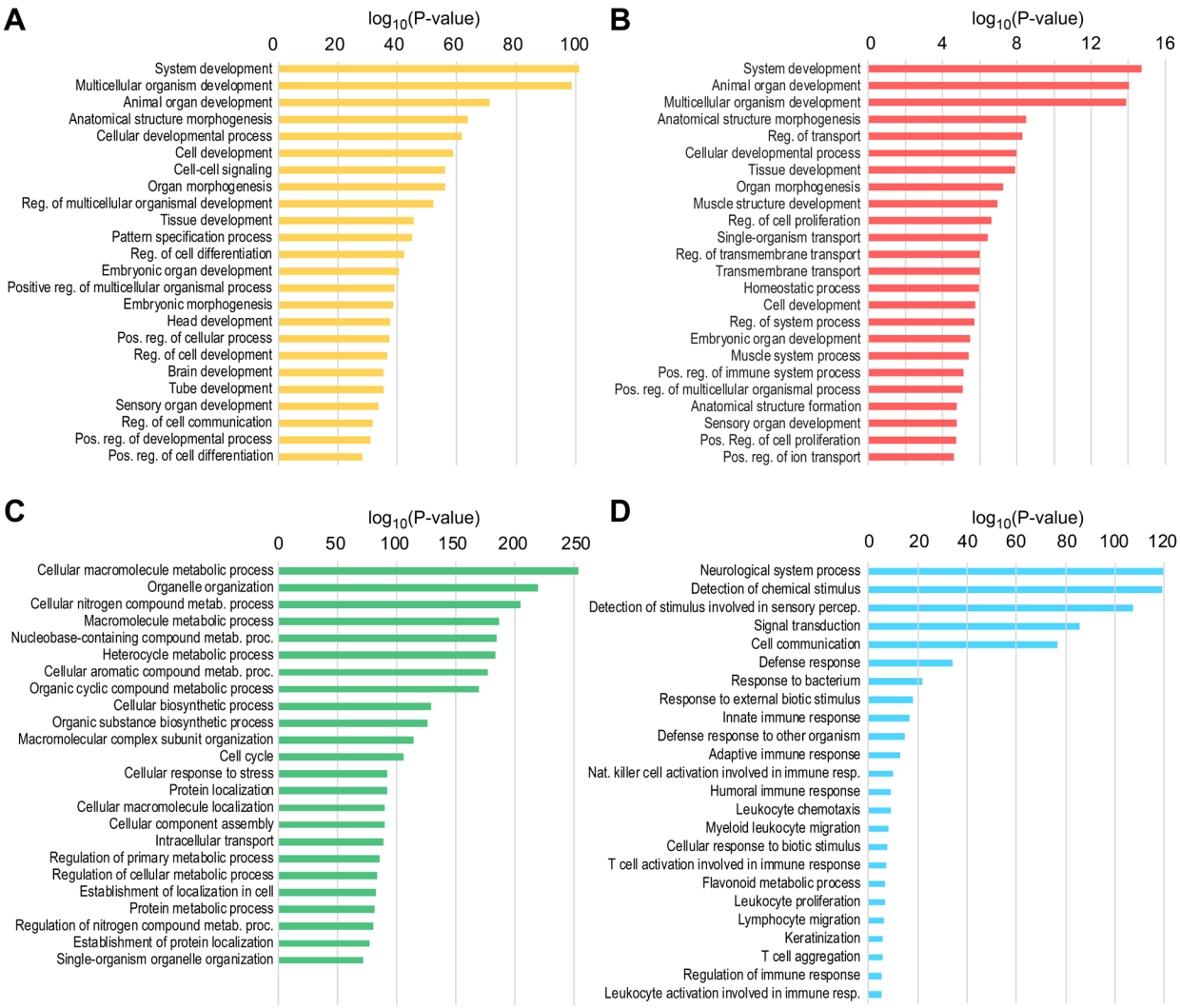


Supplemental Fig. S5. Upregulated bivalent genes in lineage-restricted cells are no more activated compared to upregulated H3K27me3-only or unmarked genes. Barplots showing the distribution of gene expression fold changes for genes upregulated in hESC-derived Ectoderm, Mesoderm, or Endoderm compared to hESCs. hESC, human ESCs. Genes grouped based on their chromatin states in hESCs. All the *P* values were calculated using two-sided Wilcoxon rank-sum test.



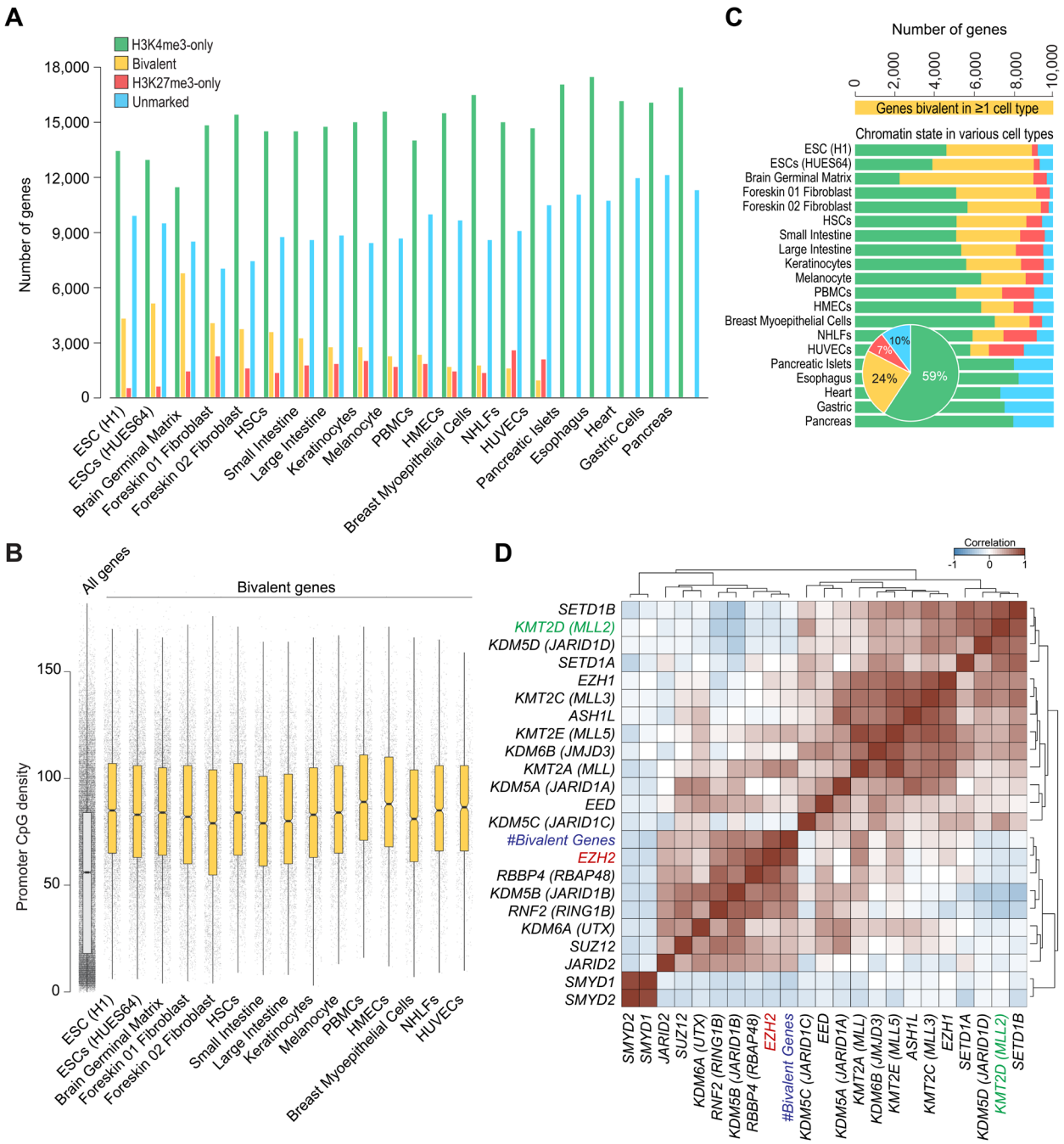
Supplemental Fig. S6. 'Poised' RNA polymerase II at bivalent genes is incompatible with transcription.

(A) Boxplot showing the distribution of gene expression in mouse ESCs grown in serum-containing medium (Marks et al. 2012) for the four gene classes shown in Fig. 3. FPKM, fragments per kilobase per million mapped reads.
(B) Heatmaps showing unsupervised hierarchical clustering of pairwise Pearson's correlations between (log_2 transformed) signals near TSSs from indicated ChIP-seq datasets and gene expression (Brookes et al. 2012; Marks et al. 2012; Tee et al. 2014).
(C) Scatter plots showing pairwise Pearson's correlations between (log_2 transformed) signals near TSSs from indicated ChIP-seq datasets and gene expression (Brookes et al. 2012; Marks et al. 2012; Tee et al. 2014) for transcriptionally active genes (left) and bivalent genes with poised RNA Polymerase II (right). Positive and negative correlations are noted in red and blue, respectively.
(D) Boxplot showing the distribution of gene expression fold changes over time (vs 0h) for bivalent genes with 'poised' or no RNAPII.
(E) Same as in C, but only for genes upregulated in EpiLCs (q-value < 0.05).



Supplemental Fig. S7. Gene ontology (GO) enrichment analysis.

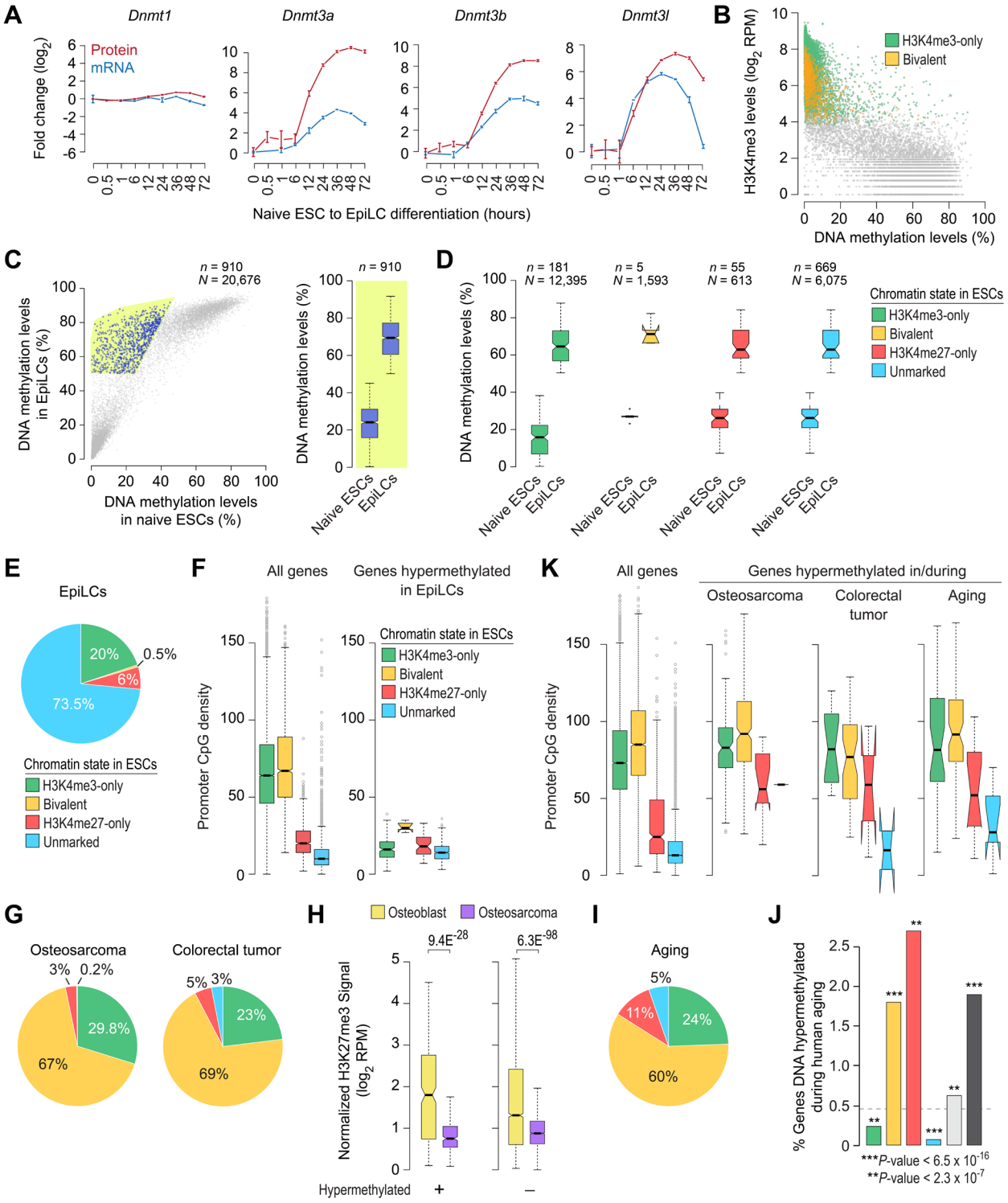
(A-D) Top GO categories (biological processes) from enrichment analysis of bivalent (A), H3K27me3-only (B), H3K4me3-only (C), and unmarked (D) genes, as defined in naïve mouse ESCs. See Supplemental Table 4 for a complete list with additional details.



Supplemental Fig. S8. Characterization of bivalent promoters across various cell types.

(A) Number of genes within each of the four classes, defined based on H3K4me3 (+/-500bp of TSS) and/or H3K27me3 (+/-2 kb of TSS) enrichment at gene promoters in various human cell types. Bivalent, positive for H3K4me3 and H3K27me3; H3K4me3-only, positive for H3K4me3 and negative for H3K27me3; H3K27me3-only, positive for H3K27me3 and negative for H3K4me3; Unmarked, negative for both H3K4me3 and H3K27me3.

(B) Boxplot showing the distribution of CpG dinucleotide frequency at promoters (+/- 500 bp of TSS) of all genes (left-most) and genes bivalently marked in various human cell types.
(C) Genes bivalently marked in one or more cell types (*n* = 10,042) and their chromatin state in various cell types (bottom). Inset: pie-chart summarizing the proportional breakdown of chromatin states across all cell types. Color scheme same as in A.
(D) Heatmap showing unsupervised hierarchical clustering of Pearson's correlations between number of bivalent genes and expression levels of H3K4me3 or H3K27me3 methylases and demethylases across cell types.



Supplemental Fig. S9. Bivalent chromatin protects promoters from *de novo* DNA methylation.

(A) Relative protein and mRNA expression levels (compared to 0 h) of DNA methyltransferases (DNMT1, DNMT3A, DNMT3B, DNMT3L) during naïve mouse ESC to EpiLC differentiation (Yang et al. 2019). Error bars represent SEM.

(B) Scatter plot showing DNA methylation levels (Shirane et al. 2016) (x-axis) and H3K4me3 levels (Yang et al. 2019) (y-axis) at gene promoters (*N* = 20,676) in naïve mouse ESCs. Individual data points represent gene promoters. H3K4me3-only and bivalent promoters are highlighted in green and orange, respectively.

(C) *Left*: Scatter plot showing DNA methylation levels at gene promoters (*N* = 20,676) in naïve mouse ESCs (x-axis) and EpiLCs (y-axis) (Shirane et al. 2016). Individual data points represent gene promoters. Promoters that are hypermethylated in EpiLCs compared to ESCs (*n* = 910) are highlighted in blue. *Right*: Boxplot showing the distribution of DNA methylation levels at promoters hypermethylated in EpiLCs compared to ESCs.

(D) Promoters hypermethylated in EpiLCs, shown in C (right panel), are divided into four groups based on their chromatin state in naïve mouse ESCs. Boxplots show the distribution of DNA methylation levels for each group of promoters hypermethylated in EpiLCs compared to ESCs. *n*, number of hypermethylated genes within each category; *N*, total number of genes within each category.

(E) Pie-chart showing proportion of genes hypermethylated in EpiLCs based on their promoter chromatin status in naïve mouse ESCs (Shirane et al. 2016).

(F) Boxplots showing the distribution of CpG dinucleotide frequency at promoters (+/- 500 bp of TSS) of all genes (left) and those that are hypermethylated in EpiLCs (right). Genes are divided into four groups based on their chromatin state in naïve mouse ESCs.

(G) Same as in E, but for genes hypermethylated in adult human cancers (Widschwendter et al. 2007; Easwaran et al. 2012). Genes are grouped based on their promoter chromatin status in human ESCs.

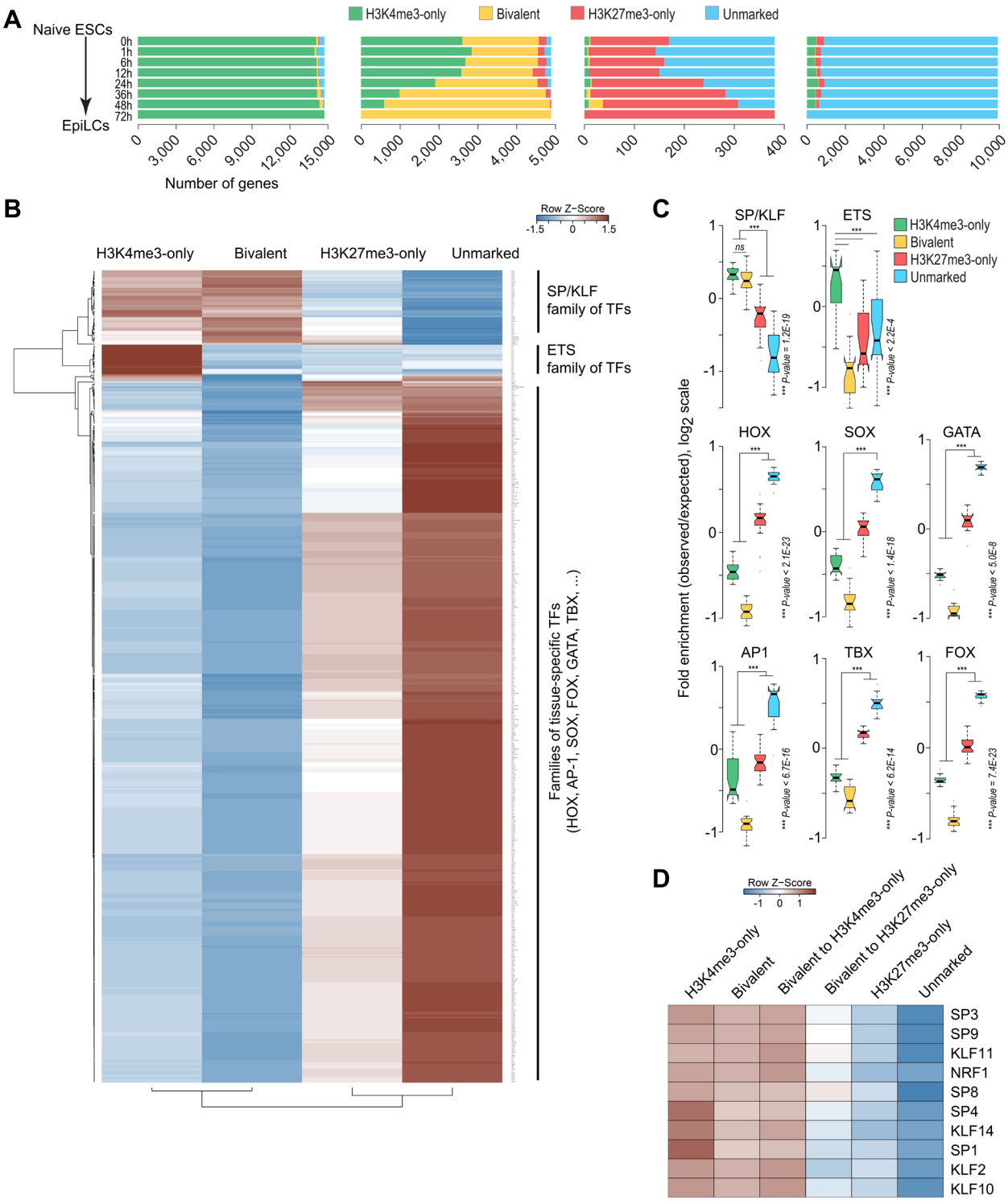
(H) Genes bivalently marked in human ESCs were divided into those that are aberrantly DNA hypermethylated in human osteosarcoma (*left*) and those that are not (*right*). Boxplots show the distribution of H3K27me3 levels at these gene promoters in human osteoblasts (yellow) and osteosarcoma (purple) (Easwaran et al. 2012).

(I) Same as in G, but for genes DNA hypermethylated during human aging (Rakyan et al. 2010).

(J) Percentage of genes, within each of the four classes of genes defined in human ESCs, whose promoters are DNA hypermethylated during human aging. Light and dark gray bars respectively denote genes enriched for H3K4me3 (H3K4me3-only and bivalent) and H3K27me3 (bivalent and H3K27me3-only). Dotted gray line denotes expected frequency.

(K) Same as in F, but for genes hypermethylated in osteosarcoma, colorectal tumor, and during aging.

All the *P* values were calculated using two-sided Wilcoxon rank-sum test.



Supplemental Fig. S10. Chromatin fate and sequence characteristics of bivalent promoters.

(A) Genes are grouped into four classes based on their chromatin state in mouse EpiLCs (72h), defined based on H3K4me3 (+/- 500 bp of TSS) and/or H3K27me3 (+/- 2 kb of TSS)

enrichment at gene promoters, and their chromatin states during mouse ESC to EpiLC transition are shown (top to bottom).

(B) Heatmap showing relative enrichment for binding motifs for various transcription factors (TFs) within promoters (+/- 500 bp of TSS) of the four genes classes defined based on the chromatin state of promoters in human ESCs. See also Supplemental Table 9.

(C) Boxplots showing the distribution of fold enrichment (observed/expected) values for binding motifs for various families of TFs (SP/KLF, ETS, HOX, SOX, GATA, AP1, TBX, FOX) within promoters (+/- 500 bp of TSS) of four classes of genes defined based on their chromatin state in mouse ESCs. All the P values were calculated using two-sided Wilcoxon rank-sum test. See also Supplemental Table 9.

(D) Same as in B, for select TFs. Also shown (middle two columns) are relative enrichment within bivalent promoters that mostly resolve into H3K4me3-only or H3K27me3-only state in other cell types.

SUPPLEMENTAL TEXT

Intragenic CGIs and bivalent chromatin. Although our study focused only on bivalently modified promoters, bivalent chromatin has also been observed elsewhere in the genome. Intragenic CGIs—when unmethylated—are marked by bivalent chromatin in ESCs and are associated with key developmental regulators (Lee et al. 2017). It should be noted here that, unlike promoter CGI methylation (which is associated with gene repression), DNA methylation of intragenic CGIs (and gene bodies in general) is associated with active transcription (Jones 2012). Cell type-specific methylation of bivalently modified intragenic CGIs, which is required for gene activation, is linked to loss of both the H3K4me3 and H3K27me3 marks, which is consistent with our conclusion that bivalency confers promoter CGIs protection against DNA methylation.

SUPPLEMENTAL METHODS

Mouse ChIP-seq and RNA-seq data sources. ChIP-seq datasets for mouse ESC to EpiLC differentiation (H3K4me3, H3K27me3, and corresponding genomic input) (Yang et al. 2019) and mouse ESCs grown in serum-containing medium (H3K4me3 and H3K27me3 (Marks et al. 2012); RNAPII-S5p, RNAPII-S2p, and RNAPII-S7p (Brookes et al. 2012); MAPK1 and TFIIH/ERCC3 (Tee et al. 2014)) were obtained from NCBI GEO portal and processed the same way for uniformity. Briefly, single-end reads from the ChIP-seq and genomic input were mapped to the mouse genome (mm9 assembly) using Bowtie (Langmead et al. 2009), allowing for up to three mismatches, retaining only reads that align to unique genomic locations (bowtie -m 1 -v 3 -p 3 --chunkmbs 256 mm9). RNA-seq datasets for mouse ESC to EpiLC differentiation (Yang et al. 2019) and mouse ESCs grown in serum containing medium (Marks et al. 2012) were obtained from NCBI GEO portal and processed the same way for uniformity. List of genes that are (a) transcriptionally active, (b) bivalent and harbor 'poised' RNAPII, (c) bivalent but harbor no RNAPII, or (d) H3K27me3-only but harbor poised RNAPII were derived from a previous study (Brookes et al. 2012).

***Human ChIP-seq and RNA-seq data sources*.** Uniformly processed and consolidated Human Epigenome Roadmap data, mapped to the human genome (hg19), for various cell/tissue types were downloaded from the Washington University portal (https://egg2.wustl.edu/roadmap/web_portal/). Only those normal cell/tissue-types for which RNA-seq, H3K4me3 ChIP-seq, H3K27me3 ChIP-seq, and genomic input (control) datasets were available were considered for analysis. To ensure that only the highest quality ChIP-seq datasets are used for downstream analysis, only those that satisfied the following criteria were retained for further analysis: (i) number of mapped reads is at least 10 million, (ii) reported signal-to-noise ratio (SNR) (Roadmap Epigenomics Consortium et al. 2015)—the degree to which reads are concentrated in peaks versus the background—for the H3K4me3 (or H3K27me3) dataset from a given cell/tissue type is (a) greater than that for corresponding genomic input and (b) at least (mean – 1 STD) of SNRs of all H3K4me3 (or H3K27me3, respectively) datasets across all cell/tissue types. (iii) reported SNR for the genomic input from a given cell/tissue type is at most (mean + 1 STD) of SNRs of all input datasets across all cell/tissue types. H3K4me3 and H3K27me3 ChIP-seq data for osteoblasts and osteosarcoma (Easwaran et al. 2012) were obtained from NCBI GEO portal, aligned to the human genome (hg19) using Bowtie (as described above), and promoter H3K4me3/H3K27me3 signal normalized by the average of all promoter signals.

ChIP-seq read density plots and heatmaps. For a gene-set of interest, genes were first aligned (5' to 3') relative to their TSSs, and average read density near TSSs (within +/-*N* kb of the TSSs) was plotted by calculating normalized mean read densities (RPM) within each of 2*N*/100 100bp non-overlapping windows spanning *N* kb (**Fig. 1; Fig. 3**). Ninety-five percent confidence interval of the mean read density (**Fig. 3**) was estimated as 1.96 times standard error of the mean (S.E.M.). Heatmaps were generated using normalized read densities (RPM) for each 100 bp window (column) for each gene (row). The order of genes in all heatmaps is the same, determined by H3K4me3 signal (within +/- 500 bp of TSS) in H1-ESCs in decreasing order (**Fig. 1B, D**) or unsupervised hierarchical clustering of genes based on their expression (**Supplemental Fig. S1B**). For visualization purposes, the maximum signal threshold was set as

the RPM value at the 98th percentile, and values higher than the threshold were set to the threshold value. Row-normalized heatmaps (Fig. S1B) represent gene expression as a *Z*-score, a relative value.

UCSC Genome Browser Tracks. Consolidated bigWig files for human ChIP-seq or RNA-seq data (containing strand-specific normalized RNA signal) were obtained from the NIH Roadmap Epigenomics project (Roadmap Epigenomics Consortium et al. 2015) (https://egg2.wustl.edu/roadmap/data/byDataType/rna/signal/normalized_bigwig/stranded/) and visualized on the UCSC Genome Browser.

Functional enrichment analysis. For each set of genes, Gene Ontology (GO) functional enrichment analysis was performed using DAVID (https://david.ncifcrf.gov/) (Huang da et al. 2009).

Machine learning approach for predicting bivalent chromatin. Multinomial log-linear models *via* neural networks (multinom function in R (R Core Team 2020)) were generated using (a) promoter (±500 bp of TSS) dinucleotide frequencies (5' to 3') and (b) promoter (±2 kb of TSS) H3K27me3 tag density (RPM) and enrichment (ChIP/input). A 5-fold cross validation was employed to train and test a model for its ability to classify promoters into one of the four chromatin states: H3K4me3-only, H3K27me3-only, bivalent, and unmarked. At least 1,000 models were generated by randomly sampling the dataset for training and testing. Accuracy of a four-class classification was estimated from the confusion matrix for each of the models. Precision and recall for prediction of the bivalent class were estimated from the confusion matrix of the four-class model, where promoters of the bivalent class were considered as positives and those from the other three classes were considered as negatives.

Human DNA methylation analysis. Genes DNA-hypermethylated in human osteosarcoma was inferred from processed probe-level methylation data (Easwaran et al. 2012) using criteria outlined in the original study (for details, see **Supplemental Methods**). Probe-level methylation

data for gene promoters (defined as -1000 bp upstream and +200 bp downstream of TSS) in osteosarcoma, osteoblasts, and mesenchymal stem cells (MSCs) were obtained from this study (Easwaran et al. 2012). Genomic co-ordinates for hypermethylated probes were obtained from Infinium manifest file and were mapped to the hg19 RefSeq annotations. Genes were considered hypermethylated in osteosarcoma if at least one probe within its promoter has methylation ratio >=0.75 in osteosarcoma and methylation ratio <=0.25 in both osteoblasts and mesenchymal stem cells (MSCs).

DNA-methylation (Illumina 450K array) beta values for 6,129 tumors and respective control tissues from 14 TCGA solid epithelial cancer types, generated by the TCGA Research Network, were downloaded using the TCGAbiolinks R package (Colaprico et al. 2016; R Core Team 2020). For each cancer type, probe-specific differential DNA-methylation levels were calculated by comparing methylation beta values between tumor samples (sample_type: “Primary Tumor”) and control samples (sample_type: “Solid Tissue Normal”) using the function TCGAanalyze_DMC from the TCGAbiolinks package. All qualified probes (with p-value < 0.05) that are within +/-500bp of TSS of an annotated transcript (RefSeq annotation for the hg19 genome build) were assigned to that transcript (NR/NM ids). In cases where a probe can be assigned to two or more transcripts, the probe was assigned to the transcript whose TSS is closer to the probe. For a given condition (tumor or control), a promoter’s methylation level was computed by taking the average of the methylation values of all probes assigned to the transcript promoter. A transcript was considered DNA-hypermethylated (or DNA-hypomethylated) in tumor compared to control if its methylation level in tumor (control, respectively) is at least 0.4 and 2-fold greater than that in control (tumor, respectively). RefSeq accession (NM/NR ids) was used to integrate human ESC chromatin state and methylation data. Odds Ratios and *P*-values were calculated using Fisher’s exact test as implemented in fisher.test() function in R (R Core Team 2020).

Motif analysis. Occurrences of TF binding motifs within gene promoters (+/- 500 bp of TSS) were inferred using the FIMO tool using default parameters (Grant et al. 2011). A collection of

746 known motifs from the non-redundant JASPAR CORE 2020 database (Fornes et al. 2020) of vertebrate TF motifs was used as queries. Background frequencies of a motif's occurrence at gene promoters, required to run FIMO, were determined based on all promoter sequences annotated in the RefSeq annotation for the hg19 genome build. Motif occurrences with p-value $< 10^{-4}$ were deemed significant and were considered for further analysis.

Note on genome assemblies used. In this study, we used the GRCm37 (mm9) and GRCh37 (hg19) assemblies to map all sequencing reads from mouse and human origin, respectively. We do not expect changes to our conclusions if we used the more recent version of the mouse or human genome assembly (GRCm38/mm10 or GRCh38, respectively) as our analysis is focused on RefSeq-annotated gene promoters, known to be highly conserved regions outside of repeats. These non-repetitive regions (promoters) were already well sequenced in mm9 and hg19.

REFERENCE FOR SUPPLEMENTAL MATERIAL

Brookes E, de Santiago I, Hebenstreit D, Morris KJ, Carroll T, Xie SQ, Stock JK, Heidemann M, Eick D, Nozaki N et al. 2012. Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell* **10**: 157-170.

Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I et al. 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**: e71.

Easwaran H, Johnstone SE, Van Neste L, Ohm J, Mosbruger T, Wang Q, Aryee MJ, Joyce P, Ahuja N, Weisenberger D et al. 2012. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Res* **22**: 837-849.

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranasic D et al. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**: D87-D92.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**: 484-492.

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

Lee SM, Lee J, Noh KM, Choi WY, Jeon S, Oh GT, Kim-Ha J, Jin Y, Cho SW, Kim YJ. 2017. Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. *Proc Natl Acad Sci U S A* **114**: E1885-E1894.

Marks H, Kalkan T, Menafra R, Denissov S, Jones K, Hofemeister H, Nichols J, Kranz A, Stewart AF, Smith A et al. 2012. The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**: 590-604.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM et al. 2010. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res* **20**: 434-439.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.

Shirane K, Kurimoto K, Yabuta Y, Yamaji M, Satoh J, Ito S, Watanabe A, Hayashi K, Saitou M, Sasaki H. 2016. Global Landscape and Regulatory Principles of DNA Methylation Reprogramming for Germ Cell Specification by Mouse Pluripotent Stem Cells. *Dev Cell* **39**: 87-103.

Tee WW, Shen SS, Oksuz O, Narendra V, Reinberg D. 2014. Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs. *Cell* **156**: 678-690.

Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisenberger DJ, Campan M, Young J, Jacobs I et al. 2007. Epigenetic stem cell signature in cancer. *Nat Genet* **39**: 157-158.

Yang P, Humphrey SJ, Cinghu S, Pathania R, Oldfield AJ, Kumar D, Perera D, Yang JYH, James DE, Mann M et al. 2019. Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. *Cell Syst* **8**: 427-445 e410.