# Decoding the function of bivalent chromatin in development and cancer

Dhirendra Kumar, Senthilkumar Cinghu, Andrew J Oldfield, et al.

*Genome Res.* published online October 19, 2021
Access the most recent version at doi:10.1101/gr.275736.121

# 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

*Correspondence: dhirendra.kumar@nih.gov (D.K.), jothi@nih.gov (R.J.)

Corresponding address

Raja Jothi
NIEHS, NIH
111 TW Alexander Drive
Bldg 101; Room A314
Research Triangle Park, NC, 27709, USA
Phone: 984-287-3696

**RUNNING TITLE:** Decoding the function of bivalent chromatin

**KEY WORDS**: Bivalent Chromatin, Transcriptional Regulation, Chromatin, Epigenetics, Stem Cells, Cancer

## ABSTRACT

**Bivalent chromatin is characterized by the simultaneous presence of H3K4me3 and H3K27me3, histone modifications generally associated with transcriptionally active and repressed chromatin, respectively. Prevalent in embryonic stem cells (ESCs), bivalency is postulated to poise/prime lineage-controlling developmental genes for rapid activation during embryogenesis while maintaining a transcriptionally repressed state in the absence of activation cues; however, this hypothesis remains to be directly tested. Most gene promoters DNA-hypermethylated in adult human cancers are bivalently marked in ESCs, and it was speculated that bivalency predisposes them for aberrant *de novo* DNA methylation and irreversible silencing in cancer, but evidence supporting this model is largely lacking. Here we show that bivalent chromatin does not poise genes for rapid activation but protects promoters from *de novo* DNA methylation. Genome-wide studies in differentiating ESCs reveal that activation of bivalent genes is no more rapid than that of other transcriptionally silent genes, challenging the premise that H3K4me3 is instructive for transcription. H3K4me3 at bivalent promoters—a product of the underlying DNA sequence—persists in nearly all cell types irrespective of gene expression and confers protection from *de novo* DNA methylation. Bivalent genes in ESCs that are frequent targets of aberrant hypermethylation in cancer are particularly strongly associated with loss of H3K4me3/bivalency in cancer. Altogether, our findings suggest that bivalency protects reversibly repressed genes from irreversible silencing and that loss of H3K4me3 may make them more susceptible to aberrant DNA methylation in diseases such as cancer. Bivalency may thus represent a distinct regulatory mechanism for maintaining epigenetic plasticity.**

## INTRODUCTION

The DNA in eukaryotic cells is organized into chromatin consisting of repeating units of nucleosome, an octamer of histone proteins wrapped with approximately 147 base pairs of DNA. Consequently, chromatin plays a central role in regulating accessibility to DNA in many DNA-templated processes including transcription. Histone modifications and DNA methylation are key epigenetic mechanisms that modulate chromatin structure and thus regulate gene expression programs controlling cell fate decisions and cell identity during development (Jaenisch and Bird 2003; Kouzarides 2007; Li et al. 2007; Allis and Jenuwein 2016).

Histones are subject to a vast array of post-translational modifications including acetylation and methylation (Kouzarides 2007; Li et al. 2007). Whereas histone acetylation is generally associated with gene activation, histone methylation, depending on the residue modified, is associated with either activation or repression. Trimethylation of histone H3 on lysine 4 (H3K4me3) and lysine 27 (H3K27me3) are two of the most extensively studied histone modifications associated with transcriptionally active and repressed chromatin, respectively (Barski et al. 2007). H3K4me3 and H3K27me3, respectively, are catalyzed by the Trithorax group (TrxG) and Polycomb group (PcG) of proteins (Simon and Kingston 2009; Margueron and Reinberg 2011; Di Croce and Helin 2013; Piunti and Shilatifard 2016; Schuettengruber et al. 2017). Because TrxG and PcG proteins act antagonistically to regulate, respectively, the activated and repressed states of gene expression, H3K4me3 and H3K27me3 were thought to be mutually exclusive. But this assumption was challenged by the discovery of bivalent domains—genomic regions characterized by the simultaneous presence of H3K4me3 and H3K27me3—found predominantly at developmentally

regulated gene promoters in embryonic stem cells (ESCs) (Azuara et al. 2006; Bernstein et al. 2006; Voigt et al. 2012; Harikumar and Meshorer 2015; Shema et al. 2016; Blanco et al. 2020). Although H3K4me3 and H3K27me3 occupy essentially non-overlapping regions within bivalent domains (Barski et al. 2007), with H3K27me3 domains typically flanking a H3K4me3 domain, it was later established that nucleosomes that bear both "active" H3K4me3 and repressive H3K27me3 do exist *in vivo*, albeit on opposite H3 tails in nearly all cases (Voigt et al. 2012; Shema et al. 2016), consistent with direct allosteric inhibition of PRC2 activity by H3K4me3 (Schmitges et al. 2011). Despite the presence of H3K4me3, bivalently marked promoters are transcriptionally inactive, if not expressed at very low levels (Mikkelsen et al. 2007). This initial observation led to the elegant and inherently appealing concept that bivalency poises/primes lineage-controlling developmental genes for rapid activation during embryogenesis while maintaining a repressed state in the absence of activation cues (Azuara et al. 2006; Bernstein et al. 2006; Voigt et al. 2013); yet, this hypothesis remains to be directly tested, and the function of bivalent domains in development remains a mystery.

Bivalency was initially hypothesized to be an ESC-specific chromatin state. During ESC differentiation, bivalency is thought to resolve into either H3K4me3-only or H3K27me3-only state depending on whether the gene is activated or silenced, respectively. Later observations, however, confirmed the existence of bivalent domains in terminally differentiated cell types (Barski et al. 2007; Mikkelsen et al. 2007; Mohn et al. 2008), raising the question about its need and functional relevance in cell types with no differentiation potential.

DNA methylation, as an heritable epigenetic mark, adds an additional level of stability by serving as an enduring 'lock' to reinforce a previously silenced state by subjecting genes to irreversible transcriptional silencing even in the presence of all of the factors required for their expression (Deaton and Bird 2011; Jones 2012; Bestor et al. 2015; Schubeler 2015). Most gene promoters DNA-hypermethylated in adult human cancers are bivalently marked in ESCs, and it was speculated that bivalency predisposes these genes for aberrant *de novo* DNA methylation and irreversible silencing in cancer (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007), but evidence supporting this model is largely lacking. Here we set out to decode the function of bivalent chromatin in physiological and pathological settings.

## RESULTS

### Bivalent chromatin does not poise genes for rapid activation

To assess whether bivalent chromatin represents a distinct epigenetic state and/or a regulatory mechanism, we investigated genes with bivalently marked promoters in pluripotent human embryonic stem cells (ESCs) (**Supplemental Table 1; Methods**); henceforth, we use 'bivalent genes' to refer to genes whose promoters are bivalently marked in ESCs, unless stated otherwise. To gain insight into the functional significance of bivalent chromatin, we first sought to determine the chromatin and expression status of bivalent genes in lineage-restricted multipotent and terminally differentiated cells. Using publicly available NIH Roadmap Epigenomics project data from a large and diverse collection of human tissues and progenitor cells (http://www.roadmapepigenomics.org/data/), we noted that H3K4me3 enrichment at bivalently marked gene promoters in ESCs persists in nearly all other cell types irrespective of transcriptional activity (**Fig. 1A-C; Supplemental Fig. S1**). In contrast, H3K27me3 enrichment is more dynamic

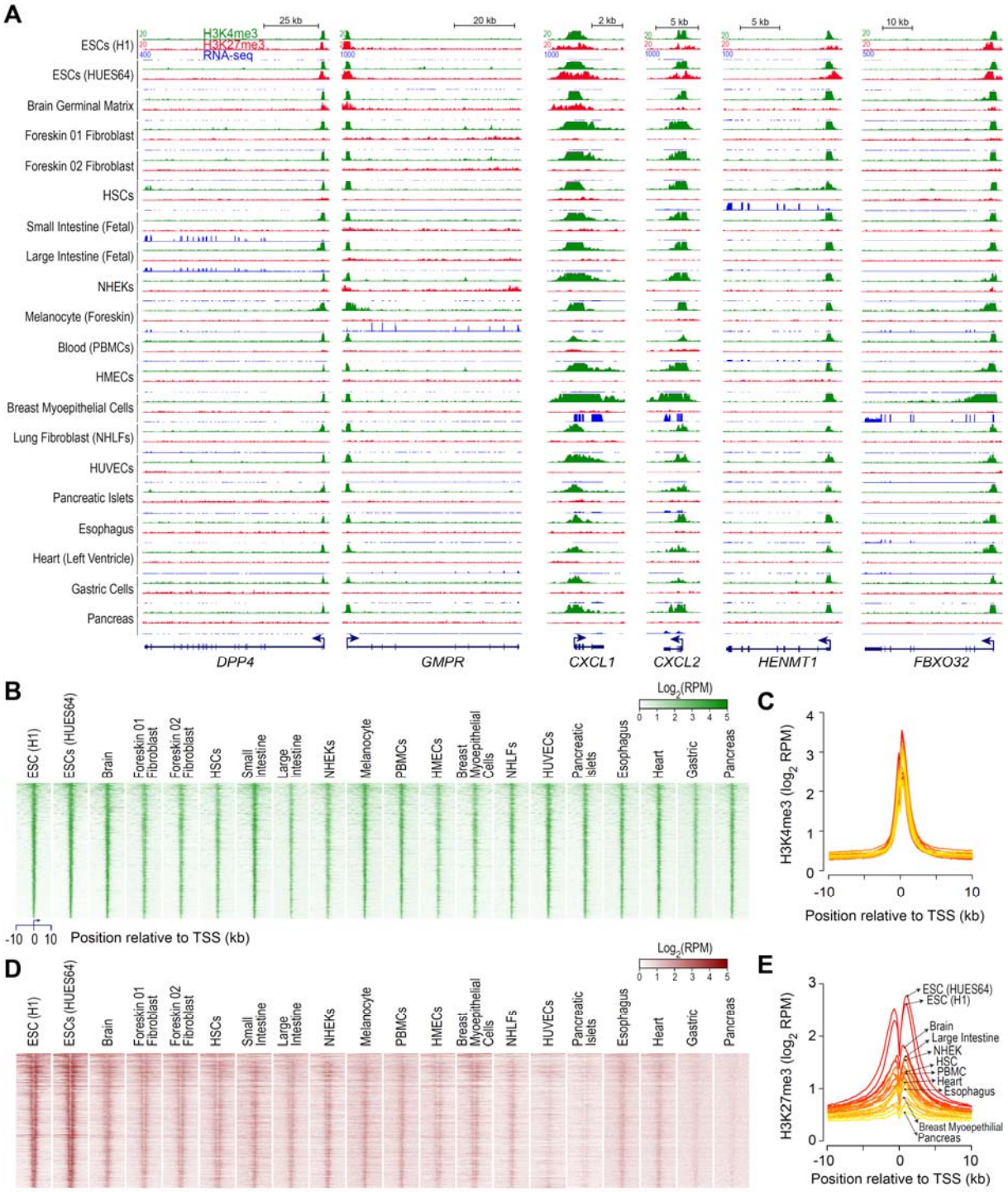



**Figure 1.** H3K4me3, observed at bivalent promoters in ESCs, persists in nearly all cell types irrespective of gene expression. (***A***) Genome browser shots of select genes, bivalently marked in human embryonic stem cells (ESCs), showing ChIP-seq read density profiles for H3K4me3 (green) and H3K27me3 (red) in various cell types. Also shown are read density profiles for gene expression (blue; RNA-seq). HSCs, hematopoietic stem cells; PBMCs, peripheral blood mononuclear cells; HMECs, human mammary epithelial cells; NHEKs, normal human embryonic kidneys, NHLFs, normal human lung fibroblasts; HUVECs, human umbilical vein endothelial cells. (***B***, ***D***) Heatmap representation of H3K4me3 (B) and H3K27me3 (D) ChIP-seq read density, in various cell types, near transcription start sites (TSSs) of genes bivalently marked in human ESCs. Genes were ordered by decreasing order of H3K4me3 signal in ESCs (top to bottom). Read density is represented as RPKM (reads per million mapped reads). (***C***, ***E***) Average H3K4me3 (C) and H3K27me3 (E) ChIP-seq read density, in various cell types, near TSSs of genes shown in B. Shades of color represent individual cell types. Select cell types, out of the twenty plotted, are

across cell types (**Fig. 1A,D,E; Supplemental Fig. S1**), with its absence not necessarily accompanied by gene activation. Resolution of bivalent state in ESCs to H3K4me3-only state in lineage-restricted cell types coupled with no guarantees of transcription—in the absence of H3K27me3 (**Supplemental Fig. S1**)—calls into question the premise that the H3K4me3 component of bivalent chromatin poises genes for rapid activation.

To determine whether H3K4me3 at bivalent genes confers them rapid or higher activation potential compared with other transcriptionally silent genes that lack the H3K4me3 mark, we sought to investigate the transcriptional fate of bivalent genes during early embryonic development using a previously validated differentiation system (Hayashi et al. 2011; Buecker et al. 2014; Shirane et al. 2016; Yang et al. 2019) (**Fig. 2A**), wherein naïve mouse ESCs—representing the pre-implantation mouse embryo from approximately embryonic day E3.75-E4.5—can be induced to epiblast-like cells (EpiLCs), which most closely resemble the early post-implantation epiblast (E5.5-E6.5). Using the data that we previously generated from chromatin immunoprecipitation sequencing (ChIP-seq) analyses of the chromatin from naïve mouse ESCs using antibodies against histone modifications H3K4me3 and H3K27me3 (Yang et al. 2019), we identified 2,163 genes with bivalently marked promoters (**Fig. 2B-D; Supplemental Table 2; Methods**). For comparison purposes, we also identified genes whose promoters are enriched for H3K4me3 but not H3K27me3 (H3K4me3-only) and vice-versa (H3K27me3-only). Genes with neither H3K4me3 nor H3K27me3 enrichment at their promoters were grouped as 'unmarked'. Consistent with H3K27me3's role in maintaining the transcriptionally repressed chromatin state (Riising et al. 2014), nearly all of the bivalent (94%) and H3K27me3-only (99%) genes are transcriptionally inactive, if not expressed at very low levels (<1 FPKM), in ESCs (**Fig. 2E**). We note that although the expression of bivalent genes is ~2-3 fold higher

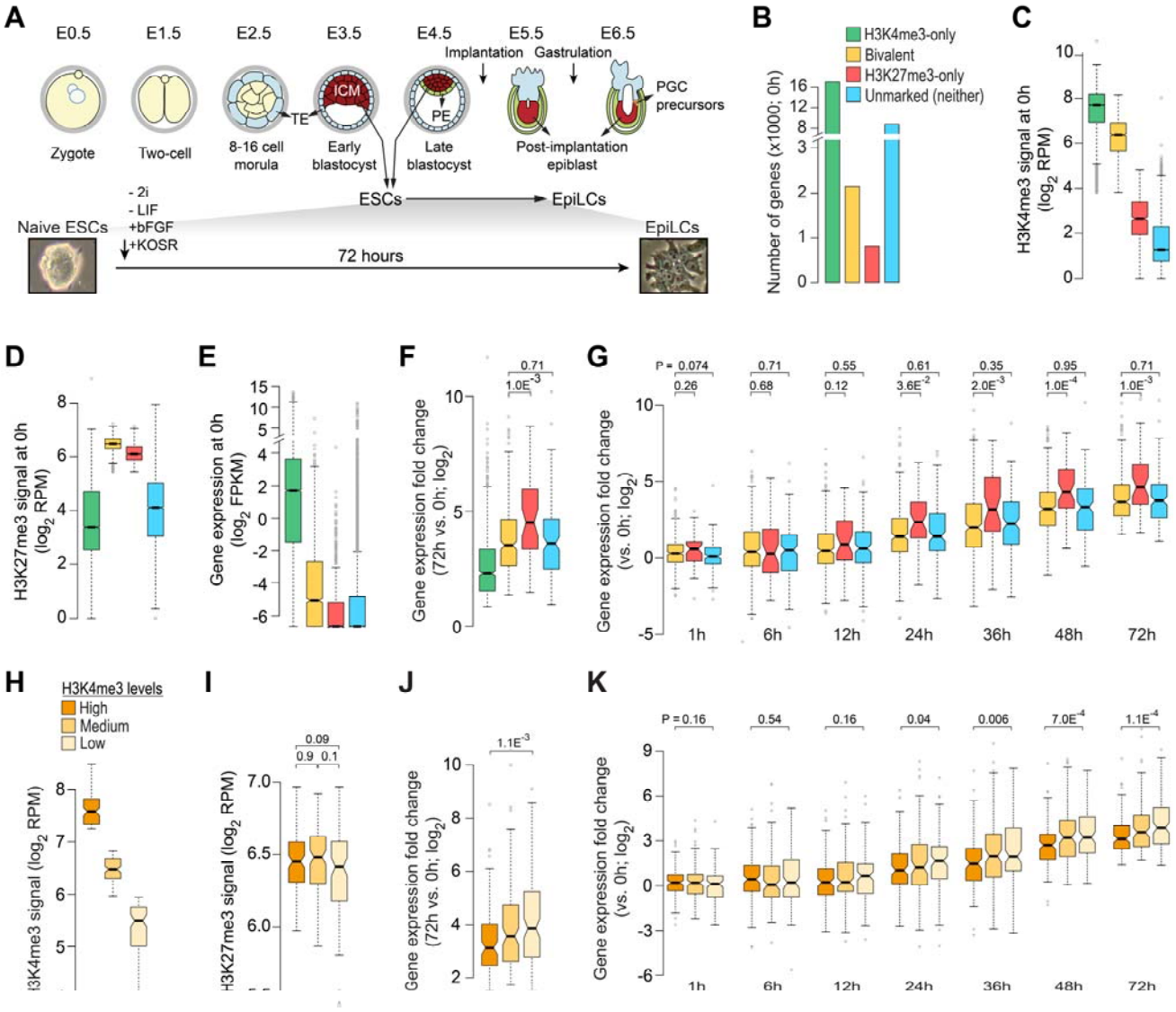



**Figure 2.** Bivalent chromatin does not poise genes for rapid activation. (***A***) *Top*: Developmental events during early embryogenesis in mouse embryos. *Bottom*: Schematic showing *in vitro* differentiation of naïve ESCs to EpiLCs. ICM, inner cell mass; ESCs, embryonic stem cells; TE, trophectoderm; PE, primitive endoderm; EpiLC, post-implantation epiblast-like cells; PGCs, primordial stem cells. (***B***) 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 naïve ESCs (0h) (Yang et al. 2019). 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. (***C-E***) Boxplots showing the distribution of ChIP-seq read densities for H3K4me3 (*C*) and H3K27me3 (*D*) at gene promoters and gene expression (*E*) in naïve ESCs for each of the four gene classes defined in *B*. RPM, reads per million mapped reads; FPKM, fragments per kilobase per million mapped reads. (***F***) Boxplot showing the distribution of gene expression fold changes for genes upregulated in EpiLCs (72h vs 0h; *n* = 1,372) (Yang et al. 2019). Genes grouped based on their chromatin states in naïve ESCs (0h). (***G***) Boxplot showing the distribution of gene expression fold changes over time (vs 0h) for genes upregulated in EpiLCs (Yang et al. 2019). Genes grouped based on their chromatin states in naïve ESCs (0h). (***H***, ***I***) Bivalent genes upregulated in EpiLCs (72h vs 0h; *n* = 1,372) were binned into three equal-sized sets based on H3K4me3 enrichment at gene promoters in naïve ESCs (0h). Box plots showing the distribution of ChIP-seq read densities for H3K4me3 (*H*) and H3K27me3 (*I*) in naïve ESCs for each of the three defined sets. (***J***) Boxplot showing the distribution of gene expression fold changes (72h vs 0h) for bivalent genes upregulated in EpiLCs. Genes grouped based the three sets defined in *H*. (***K***) Boxplot showing the distribution of gene expression fold changes over time (72h to 0h) for bivalent genes upregulated in EpiLCs (72h vs 0h). Genes grouped based on high/low H3K4me3 signal, as defined in *H*. All the *P* values were calculated using two-sided Wilcoxon rank-sum test. See also Supplemental

compared to transcriptionally silent H327me3-only or unmarked genes, it is still very low (<1 FPKM) to be considered expressed, consistent with mutual exclusivity between PRC2 activity and active transcription (Brookes et al. 2012; Riising et al. 2014).

To evaluate whether bivalent genes are activated any faster or any more than other transcriptionally silent genes (H3K27me3-only or unmarked), we focused on genes upregulated (*q*-value < 0.05) during ESC to EpiLC differentiation (**Supplemental Fig. S2A; Supplemental Table 3**). Our analysis revealed that upregulated bivalent genes are no more activated, measured as either fold change or absolute difference in expression, compared to upregulated H3K27me3-only or unmarked genes (**Fig. 2F; Supplemental Fig. S2B**). Next, to address whether bivalent chromatin confers rapid activation potential, we examined gene expression changes at various time points (0, 1, 6, 12, 24, 36, 48, and 72h) during ESC to EpiLC differentiation. We found that activation of upregulated bivalent genes is no more rapid than that of upregulated H3K27me3-only or unmarked genes (**Fig. 2G; Supplemental Fig. S2C,D**), challenging the notion that H3K4me3 at bivalent promoters poises them for rapid activation. This observation holds true irrespective of how stringently upregulated genes are defined (**Supplemental Fig. S3**). Nonetheless, if it is true that H3K4me3 at bivalent genes poises them for rapid activation, we reasoned that bivalent genes with higher levels of H3K4me3 must be activated much sooner or much more compared to those with relatively lower levels of H3K4me3. Our analysis of the activation dynamics of upregulated bivalent genes, divided into three equal-sized groups based on their H3K4me3 levels (**Supplemental Fig. S4A,B; Fig. 2H,I**) revealed that activation of bivalent genes with higher levels of H3K4me3 is neither greater nor faster compared to those with lower levels of H3K4me3 (**Fig. 2J,K; Supplemental Fig. S4C-E**). Grouping upregulated bivalent genes slightly differently based on the distribution of

H3K4me3 levels yielded the same result (**Supplemental Fig. S4F-J**). Together, these results indicate that bivalent chromatin does not poise genes for rapid activation any more than chromatin marked with just H3K27me3 or chromatin marked with neither H3K4me3 nor H3K27me3.

To assess the fate of bivalent genes in later developmental stages, we analyzed gene expression profiles of representative lineage-restricted populations from the three germ layers (ectoderm, mesoderm, and endoderm), derived through directed differentiation of human ESCs (Gifford et al. 2013). Consistent with our findings from ESC to EpiLC differentiation, upregulated bivalent genes in lineage-restricted cells are no more activated compared to upregulated H3K27me3-only or unmarked genes (**Supplemental Fig. 5**).

**Transcriptional competence of 'poised' RNA Polymerase II at bivalent genes**

Although about two-thirds of the bivalent genes lack transcriptionally engaged RNA Polymerase II (RNAPII) (Williams et al. 2015), the presence of 'poised' RNAPII, preferentially phosphorylated at serine 5 but not serine 2 and serine 7, at a subset of bivalent gene promoters has lent some credence to the conceptually appealing notion that bivalency poises genes for rapid activation while keeping them repressed (Stock et al. 2007; Brookes et al. 2012; Ferrai et al. 2017). Phosphorylation of serine 5 on RNAPII , largely mediated by the TFIIH complex, promotes transcription initiation, whereas phosphorylation of serine 2 on RNAPII mediates transition of RNAPII from initiation into productive elongation (Phatnani and Greenleaf 2006). At a cohort of PRC2-targeted developmental genes, it is MAPK1 (also known as ERK2) but not TFIIH that phosphorylates serine 5 on RNAPII, and it was hypothesized that, at these PRC2 target genes, MAPK1's phosphorylation of Serine 5 on RNAPII establishes a poised/stalled form of RNAPII that is competent for transcription (Tee et al. 2014).

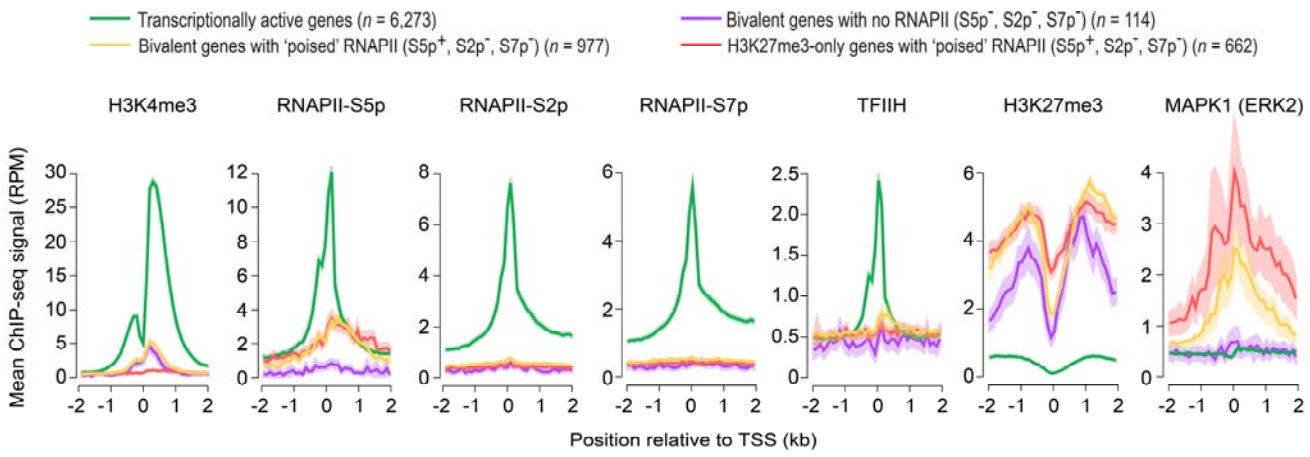



**Figure 3.** Bivalent promoters with 'poised' RNA Polymerase II are enriched for MAPK1 (also known as ERK2) but not TFIIH. ChIP-seq read density profiles of histone modifications H3K4me3 and H3K27me3 (Marks et al. 2012), phosphorylated forms (S2p, S5p, or S7p) of RNA polymerase II (RNAPII) (Brookes et al. 2012), general transcription factor TFIIH (ERCC3) (Tee et al. 2014) and mitogen-activated protein kinase MAPK1 (Tee et al. 2014) near transcription start site (TSS) of indicated gene classes in mouse ESCs grown in serum-containing medium. Mean ChIP-seq signal is shown. Confidence interval (95%) of the mean is shown as a cloud. RPM, reads per million mapped reads. See also Supplemental Fig. S6.

To gain further insight into the transcriptional competence and poising potential of RNAPII observed at bivalent promoters, we focused on bivalent genes that harbor 'poised' RNAPII (**Supplemental Table 4**), defined as RNAPII phosphorylated at serine 5 ($S5p^+$) but not serine 2 ($S2p^-$) and serine 7 ($S7p^-$) (Brookes et al. 2012). We noted that the expression of bivalent genes with 'poised' RNAPII is no more than that of bivalent genes with no RNAPII (**Supplemental Fig. S6A**). Unlike transcriptionally active gene promoters, which are enriched for TFIIH but not MAPK1, bivalent gene promoters with 'poised' RNAPII are enriched for MAPK1 but not TFIIH (**Fig. 3**). Consistent with this observation, RNAPII-S5p levels at active promoters correlate with TFIIH levels (R = 0.61) whereas RNAPII-S5p levels at bivalent promoters correlate with MAPK1 levels (R = 0.58). During early stages of transcription, serine 5 phosphorylation on RNAPII, together with the PAF complex, is known to recruit the histone methyltransferase, Set1/COMPASS, to tri-methylate H3K4 (Shilatifard 2012), which is reflected in the correlation between RNAPII-S5p and H3K4me3 levels at active genes (R = 0.6) (**Supplemental Fig. S6B,C**). However, RNAPII-S5p levels at bivalent promoters with 'poised'

RNAPII exhibit no such correlation with H3K4me3 levels (R = 0.13) (**Supplemental Fig. S6B,C**); instead, RNAPII-S5p levels correlate with repression-associated H3K27me3 levels (R = 0.4), which is in marked contrast to the inverse correlation observed between RNAPII-S5p and H3K27me3 at transcriptionally active genes (R = -0.31), but is similar to the correlation observed between RNAPII-S5p and H3K27me3 at H3K27me3-only genes with 'poised' RNAPII (R = 0.57). Given that (i) MAPK1 binds exclusively to a subset of PRC2 targets and phosphorylates serine 5 on RNAPII (Tee et al. 2014), (ii) bivalent promoters with 'poised' RNAPII are devoid of TFIIH and thus any serine 5 phosphorylation on RNAPII at bivalent promoters with 'poised' RNAPII is attributable to MAPK1 (**Fig. 3**), and (iii) PRC2 binding and active transcription are mutually exclusive (Riising et al. 2014), these data suggest that the MAPK1-mediated serine 5 phosphorylation on RNAPII at 'poised' bivalent genes may be incompatible with transcription.

Next, to assess whether the presence of 'poised' RNAPII at bivalent genes confers them rapid or higher activation potential during differentiation, we examined their expression during ESC to EpiLC differentiation. Our analysis revealed that the activation dynamics of bivalent genes with 'poised' RNAPII is no greater or no more rapid than those of bivalent genes with no RNAPII (**Supplemental Fig. S6D,E**), raising questions about the poising potential and transcriptional competence of 'poised' RNAPII at bivalent genes.

**Bivalent chromatin is a product of PRC2 activity at CpG-rich sequences**

Enrichment for developmental genes and regulators among bivalent genes is yet another characteristic that has been used to make the case for the biological relevance of bivalent chromatin in poising lineage-controlling genes for rapid activation during early embryogenesis

(Bernstein et al. 2006). Although it is true that bivalent genes are enriched for genes associated with developmental processes, this characteristic is not unique to bivalent genes as it also holds true for H3K27me3-only genes (**Supplemental Fig. S7; Supplemental Table 5**), making it a general feature of genes targeted by PRC2.

With bivalent chromatin conferring no more poising or activation potential than chromatin decorated with just H3K27me3 (**Fig. 2G**), we asked why some promoters targeted by PRC2 are also co-enriched for H3K4me3, making them bivalent, while others are not. Because H3K4me3 enrichment at bivalent promoters in ESCs persists in nearly all cell types irrespective of transcriptional activity (**Fig. 1**), we reasoned that H3K4me3 at bivalent chromatin is perhaps a product of the underlying DNA sequence features. Indeed, analysis of dinucleotide frequency at H3K27me3-enriched promoters (bivalent and H3K27me3-only) revealed a high correlation (R = 0.88) between CpG density and H3K4me3 levels (**Fig. 4A,B; Supplemental Table 2**), which is somewhat expected given that CpG-rich sequences, when unmethylated, *per se* are sufficient to establish H3K4me3 domains (Thomson et al. 2010). This observation indicates that CpG density can discriminate between PRC2 targets that are bivalent versus those that are H3K27me3-only. Moreover, this association between CpG density and H3K4me3 levels holds true even for gene promoters that are not PRC2 targets (**Fig. 4C**), suggesting that CpG density alone can be an excellent predictor of H3K4me3 levels. A corollary to this conclusion would be that CpG density, together with H3K27me3 levels, can predict bivalent chromatin (**Fig. 4B,D**). Indeed, a machine learning approach based on multinomial logistic regression of dinucleotide frequencies and H3K27me3 levels predicted, with >90% accuracy, chromatin status of promoters into one of the four classes (bivalent,

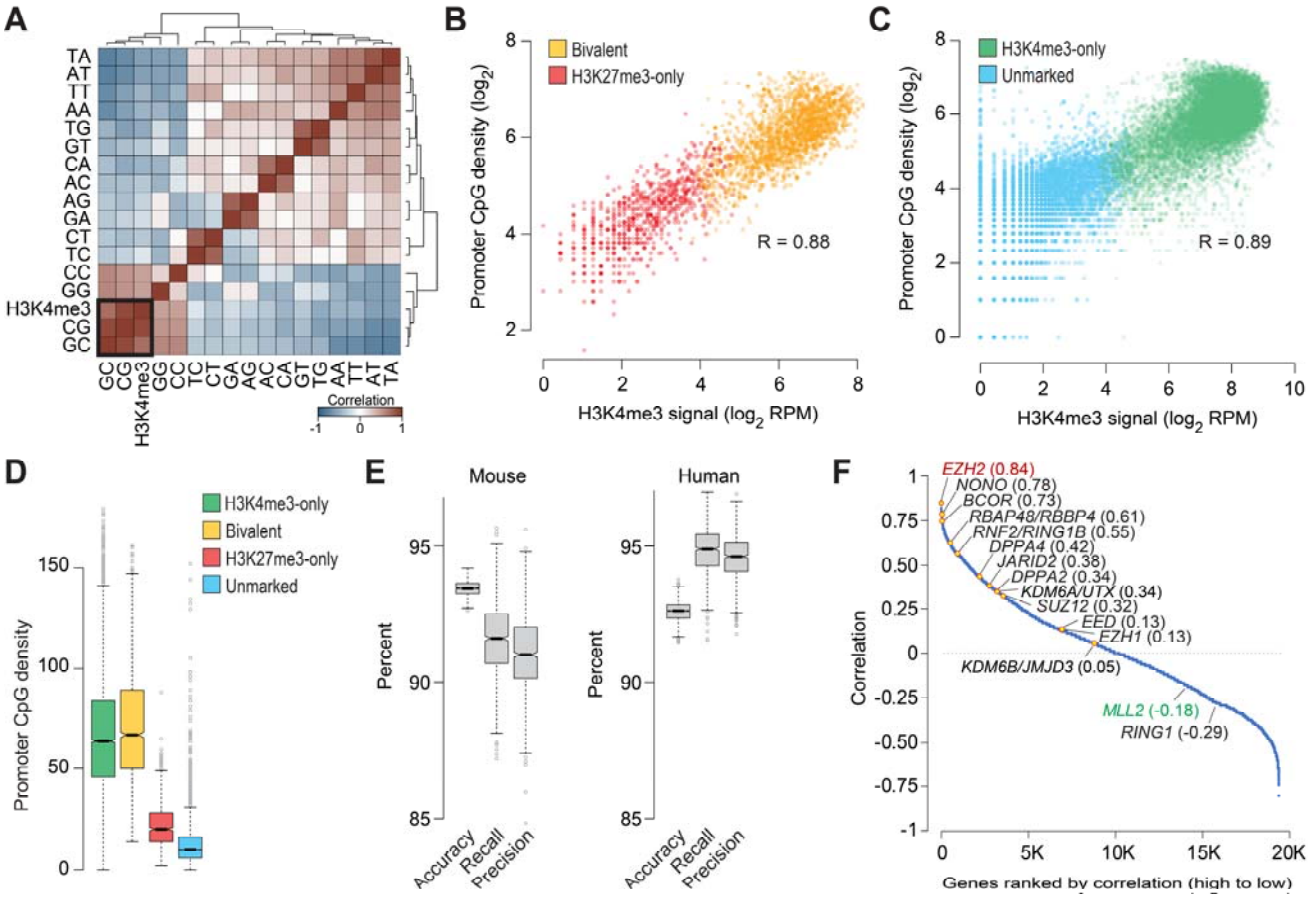



**Figure 4.** Bivalent chromatin is a product of PRC2 activity at CpG-rich sequences. (***A***) Heatmap showing unsupervised hierarchical clustering of pairwise Pearson's correlations between H3K4me3 levels and dinucleotide frequency (5' to 3') at the promoters (+/- 500 bp of TSS) of genes enriched for H3K27me3 in naïve mouse ESCs. (***B, C***) Scatter plot showing correlation between CpG dinucleotide frequency and H3K4me3 read density at promoters (+/- 500bp of TSS) of genes with (*B*) or without (*C*) H3K27me3 enrichment in naïve mouse ESCs. (***D***) Boxplots showing the distribution of CpG dinucleotide frequency at promoters (+/- 500 bp of TSS) of genes within each the four classes defined in Fig. 2B. (***E***) Statistics summarizing the performance of the multinomial logistic regression-based machine learning method for predicting chromatin state of mouse or human gene promoters (Bivalent, H3K27me3-only, H3K4me3-only, or unmarked) using only H3K27me3 data and dinucleotide frequencies at gene promoters. Boxplots show the distribution of indicated performance measures over 1,000 models. Accuracy, fraction of predictions that are correct; Recall (sensitivity), fraction of bivalent genes correctly predicted as such; Precision (positive predictive value), fraction of predicted bivalent genes that are correct. (***F***) Plot showing Pearson's correlation between number of bivalent genes and expression of individual genes, calculated based on data from various cell types. Genes, denoted as individual data points, are sorted (left to right, x-axis) based on their correlation values (y-axis). See also Supplemental Fig. S7.

H3K27me3-only, H3K4me3-only, or unmarked) (**Fig. 4E; Supplemental Methods**). Promoters with bivalent chromatin are predicted with >90% sensitivity (recall) and precision.

With DNA sequence features and CpG density in particular seeming to have such an outsized influence on H3K4me3 levels and thus establishment of bivalent domains, we next sought to

212 understand why bivalent domains are more prevalent in pluripotent ESCs compared to terminally
213 differentiated or lineage-restricted multipotent cells (**Supplemental Fig. S8A**). Because bivalently
214 marked promoters across various cell types are CpG-rich (**Supplemental Fig. S8B**) but not *vice versa*
215 (**Fig. 4B-D**) and because they almost always harbor H3K4me3 (**Fig. 1; Supplemental Fig. S8C**), we
216 reasoned that the prevalence of bivalent domains in a given cell type is presumably a function of the
217 extent of PRC2 activity and/or targeting. Consistent with this notion, analysis of gene expression
218 across various cell types revealed that the number of bivalent genes in a given cell type correlates
219 the best with expression levels of *EZH2* (R = 0.84), the catalytic subunit of the PRC2 complex (**Fig. 4F;**
220 **Supplemental Table 6**). Expression of *MLL2* (*KMT2D*), the catalytic subunit of the Set1/MLL complex
221 chiefly responsible for H3K4me3 at bivalent domains (Hu et al. 2013; Denissov et al. 2014),
222 correlates negatively (R = -0.18), if at all, with the number of bivalent genes (**Fig. 4F; Supplemental**
223 **Fig. S8D**). Given that CpG-rich sequences, when unmethylated, *per se* are sufficient to establish
224 H3K4me3 domains (Thomson et al. 2010) and that EZH2 gain/loss-of-function strongly correlates
225 with number of bivalent genes (Shema et al. 2016), these data suggest that bivalent domains are
226 likely a product of PRC2 activity at genomic regions with high CpG density.

227 **H3K4me3 at bivalent chromatin protects promoters from *de novo* DNA methylation**
228 Gene promoters targeted by PRC2 in ESCs—mostly bivalent genes—are often found to be DNA-
229 hypermethylated in adult human cancers, and it was speculated that bivalent chromatin and/or the
230 presence of Polycomb proteins might predispose these genes for aberrant *de novo* DNA methylation
231 and irreversible silencing in cancer (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al.
232 2007). Acquisition of promoter DNA methylation at these genes is thought to lock in stem cell
233 phenotypes—at the expense of ability to respond to appropriate lineage commitment and

differentiation cues—and initiate abnormal clonal expansion and thereby predispose to cancer (Jones and Baylin 2007; Widschwendter et al. 2007; Easwaran et al. 2014).

To determine whether bivalent genes in ESCs are more susceptible to *de novo* DNA methylation during normal development, we investigated promoter DNA methylation levels in naïve mouse ESCs and EpiLCs. ESCs represent the pre-implantation epiblast (~E3.75-E4.5), and EpiLCs most closely resemble early post-implantation epiblast (E5.5-E6.5) (Hayashi et al. 2011; Yang et al. 2019) (**Fig. 2A**). This time period spanning pre- to post-implantation epiblast differentiation during early embryonic development is noteworthy because naïve ESCs are associated with global DNA hypomethylation, with a major wave of global *de novo* methylation occurring after implantation (~E5.0) (Okano et al. 1999; Smith et al. 2012; Leitch et al. 2013; Shirane et al. 2016), when *de novo* methylatransferases *Dnmt3a* and *Dnmt3b*—not expressed in naïve ESCs—get induced by about ~500-1000 fold (**Supplemental Fig. S9A**).

Our analysis of DNA methylation and H3K4me3 levels in naïve ESCs revealed that promoters enriched for H3K4me3 are devoid of DNA methylation (**Supplemental Fig. S9B**). Focusing on the 910 genes whose promoters are hypermethylated in EpiLCs compared to naïve ESCs (**Supplemental Fig. S9C,D**), we noted that bivalent genes in ESCs are significantly under-represented among genes hypermethylated in EpiLCs (**Fig. 5A; Supplemental Fig. S9E; Supplemental Table 7**). In contrast, H3K27me3-only genes in ESCs are significantly over-represented among the hypermethylated

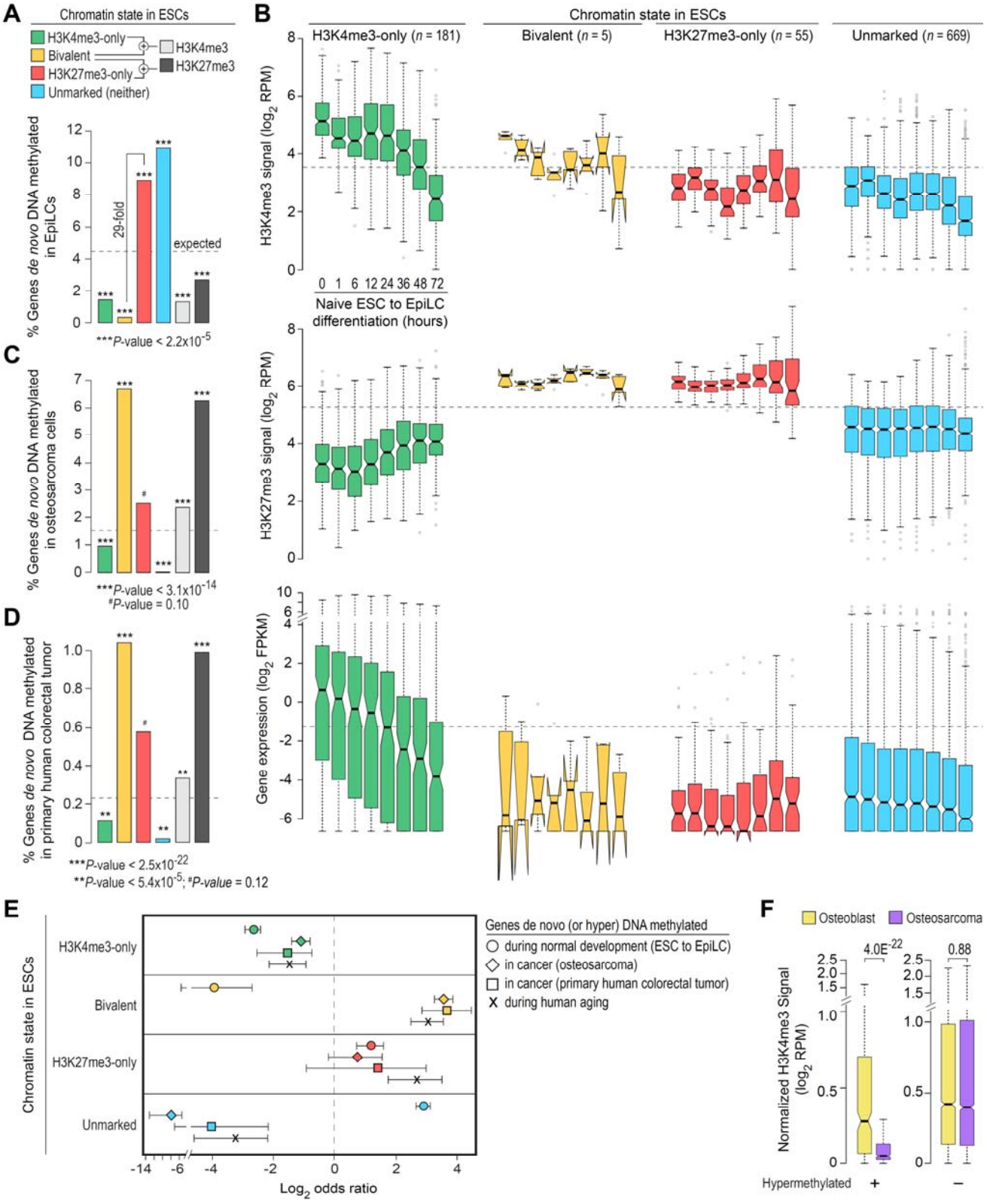



**Figure 5.** Bivalent chromatin protects promoters from *de novo* DNA methylation. (***A***) Percentage of genes, within each of the four classes of genes defined in naïve mouse ESCs, whose promoters are DNA hypermethylated in EpiLCs (Shirane et al. 2016) are shown. See also Supplemental Fig. S9C, D.

*(legend continued on next page)*

(***B***) Boxplots showing changes in promoter H3K4me3/H3K27me3 levels and gene expression (Yang et al. 2019), for each of the four classes of genes shown in *A*, during ESC to EpiLC differentiation. FPKM, fragments per kilobase per million mapped reads. RPM, reads per million mapped reads. (***C, D***) Percentage of genes, within each of the four classes of genes defined in human ESCs, whose promoters are aberrantly DNA hypermethylated in human osteosarcoma (U2OS) (Easwaran et al. 2012) (*C*) or primary human colorectal tumor (Widschwendter et al. 2007) (*D*). (***E***) $Log_2$ odds ratio, with 95% confidence intervals, of promoter *de novo* DNA methylation during normal development (mouse ESC to EpiLC differentiation; circle), in cancer (human osteosarcoma and colorectal cancer; diamond and square, respectively), and during aging (Rakyan et al. 2010) (X mark) based on their chromatin state in ESCs. (***F***) 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 H3K4me3 levels at these gene promoters in human osteoblasts (yellow) and osteosarcoma (purple) (Easwaran et al. 2012). All the *P* values were calculated using Fisher's exact test (panels A,C,and D) or two-sided Wilcoxon rank-sum test (panels F and E). See also Supplemental Fig. S9.

genes. Although H3K27me3-only genes in ESCs are 29-fold more likely to be hypermethylated compared with bivalent genes, PRC2 targets in ESCs (bivalent and H3K27me3-only), as a group, are still under-represented among hypermethylated genes (**Fig. 5A**), indicating that PRC2 targets are less susceptible to *de novo* DNA methylation at least during early stages of embryonic development. Moreover, because genes hypermethylated in EpiLCs are relatively CpG-poor (**Supplemental Fig. S9F**), over-representation of H3K27me3-only genes among the hypermethylated set is more likely a reflection of their CpG-poor promoters—known to undergo extensive and dynamic methylation and de-methylation during normal development (Meissner et al. 2008)—than their being PRC2 targets. Although the generally inverse correlation between CpG density and hypermethylation (**Fig. 4D; Fig. 5A**) suggests that protection from *de novo* methylation is perhaps a direct function of the local CpG density, studies in mouse have shown that, for most promoters, high CpG density alone cannot account for their unmethylated state *in vivo* (Lienert et al. 2011), underscoring our limited understanding of the mechanisms that protect most CpG island (CGI) promoters from *de novo* methylation.

To gain insight into mechanisms that underlie protection of bivalent promoters from *de novo* DNA methylation, we examined chromatin and expression changes that accompany hypermethylation in

EpiLCs. High-temporal resolution profiles of H3K4me3, H3K27me3, and gene expression during ESC to EpiLC differentiation (Yang et al. 2019) revealed no major changes in H3K4me3, H3K27me3, or expression levels for hypermethylated H3K27me3-only and unmarked genes (**Fig. 5B**), suggesting that *de novo* DNA methylation merely reinforces the previously silenced state at these genes. In contrast, hypermethylated H3K4me3-only genes exhibit a gradual decrease in gene expression and H3K4me3 levels (**Fig. 5B**), which would be consistent with the notion that transcriptional activity protects promoters from DNA methylation and that hypermethylation of these genes likely reflects consequence rather than cause of transcription inactivation (Jones 2012; Bestor et al. 2015; Schubeler 2015). Although almost all of the bivalent genes in ESCs are protected from *de novo* DNA methylation (**Fig. 5A; Supplemental Fig. S9E**), the very few that get hypermethylated in EpiLCs also exhibit a gradual loss of H3K4me3 and thus bivalency, but no change in expression (which was negligible to begin with) or H3K27me3 (**Fig. 5B**). Given that (i) *de novo* methylases cannot act on H3K4me3 modified nucleosomes *in vitro* (Ooi et al. 2007), (ii) complete erasure of H3K4 methylation elevates DNA methylation levels, and restoration of H3K4 methylation levels reduces DNA methylation levels back to wild-type levels (Hu et al. 2009), and (iii) reversible DNA methylation has no impact on H3K4me3 levels at gene promoters (King et al. 2016), these findings suggest that H3K4me3 at bivalent promoters, either directly or indirectly, protects transcriptionally repressed yet permissive CpG-rich promoters from *de novo* DNA methylation (also see **Supplemental Text**).

**Loss of H3K4me3 at bivalent promoters is associated with aberrant DNA hypermethylation in cancer**

To address whether bivalent chromatin predisposes genes for (or protects genes from) aberrant *de novo* methylation in adult cancer, we analyzed genes hypermethylated in osteosarcoma and

colorectal tumor (Widschwendter et al. 2007; Easwaran et al. 2012). About two-thirds of the genes hypermethylated in cancer are bivalent in ESCs (**Supplemental Fig. S9G**). Unlike in EpiLCs, genes bivalent in ESCs are significantly over-represented among genes hypermethylated in cancer (**Fig. 5A,C,D**). Whereas genes that are either H3K4me3-only or H3K27me3-only in ESCs are equally susceptible to hypermethylation during normal development and in cancer, genes bivalent in ESCs are more likely to be hypermethylated in cancer (odds ratio [OR] 11.87, 95% CI: 9.66-14.65 for osteosarcoma, and OR 12.48, 95% CI 7.21-22.34 for colorectal tumor) than during normal development (OR 0.063, 95% CI: 0.02-.15) (**Fig. 5E**).

To determine whether predisposition of bivalent genes in ESCs to acquire aberrant methylation in cancer might be due to loss of their bivalent status, we examined H3K4me3 and H3K27me3 levels for genes hypermethylated in osteosarcoma. We found that bivalent genes in ESCs that are hypermethylated in osteosarcoma, as opposed to those that are not, exhibit significantly reduced levels of H3K4me3 (**Fig. 5F**). No such specificity was observed for H3K27me3 levels; all bivalent genes in ESCs exhibit reduced levels of H3K27me3 irrespective of their methylation status (**Supplemental Fig. S9H**). Altogether, these data suggest that bivalent chromatin protects promoters from *de novo* DNA methylation and irreversible silencing while maintaining a reversibly repressed state, and that loss of H3K4me3 may make these genes more susceptible to aberrant DNA methylation in cancer.

Our analysis of hypermethylated genes in 14 other cancer types further corroborated our observation that genes bivalent in ESCs, which are CpG-rich (**Supplemental Fig. S8B**) and less likely to be methylated during normal development (**Fig. 5A**), are significantly over-represented among

genes hypermethylated in cancer (**Fig. 6A,C; Supplemental Table 8**). In contrast, bivalent genes in ESCs are not among those hypomethylated in cancer (**Fig. 6B,C; Supplemental Table 8**). Genes unmarked in ESCs, which are CpG-poor and more likely to be methylated during normal development (**Fig. 5A**), are significantly over-represented among genes hypomethylated in cancer.

Age is the single biggest risk factor for most diseases including cancer (Niccoli and Partridge 2012). Because genes frequently hypermethylated and silenced in many adult human cancers exhibit aging-associated hypermethylation and because aging-associated hypermethylation occurs predominantly at promoters bivalently marked in ESCs (Rakyan et al. 2010), we surmised that bivalent genes in ESCs are more susceptible to hypermethylation during the aging process. Indeed, our analysis of aging-associated hypermethylated genes (**Supplemental Table 7**) revealed that genes bivalent in ESCs are more likely to be hypermethylated during aging—to the same extent as in cancer—than during ESC to EpiLC differentiation (**Fig. 5E; Supplemental Fig. S9I,J**). Moreover, unlike in EpiLCs, promoters of bivalent genes in ESCs that get hypermethylated in cancer and/or during aging are CpG-rich (**Supplemental Fig. S9F,K**). Given that CpG-rich promoters are mostly unmethylated in all cell types at all stages of development, even when transcriptionally inactive (Deaton and Bird 2011; Jones 2012; Bestor et al. 2015; Schubeler 2015), these findings suggest that aging-associated hypermethylation of genes that are bivalently marked in ESCs can serve as a potential biomarker for carcinogenesis in the elderly.

**Establishment and fate of bivalent chromatin**

During ESC differentiation and embryonic development, bivalent chromatin is postulated to resolve into either H3K4me3-only or H3K27me3-only state depending on whether the gene is activated or

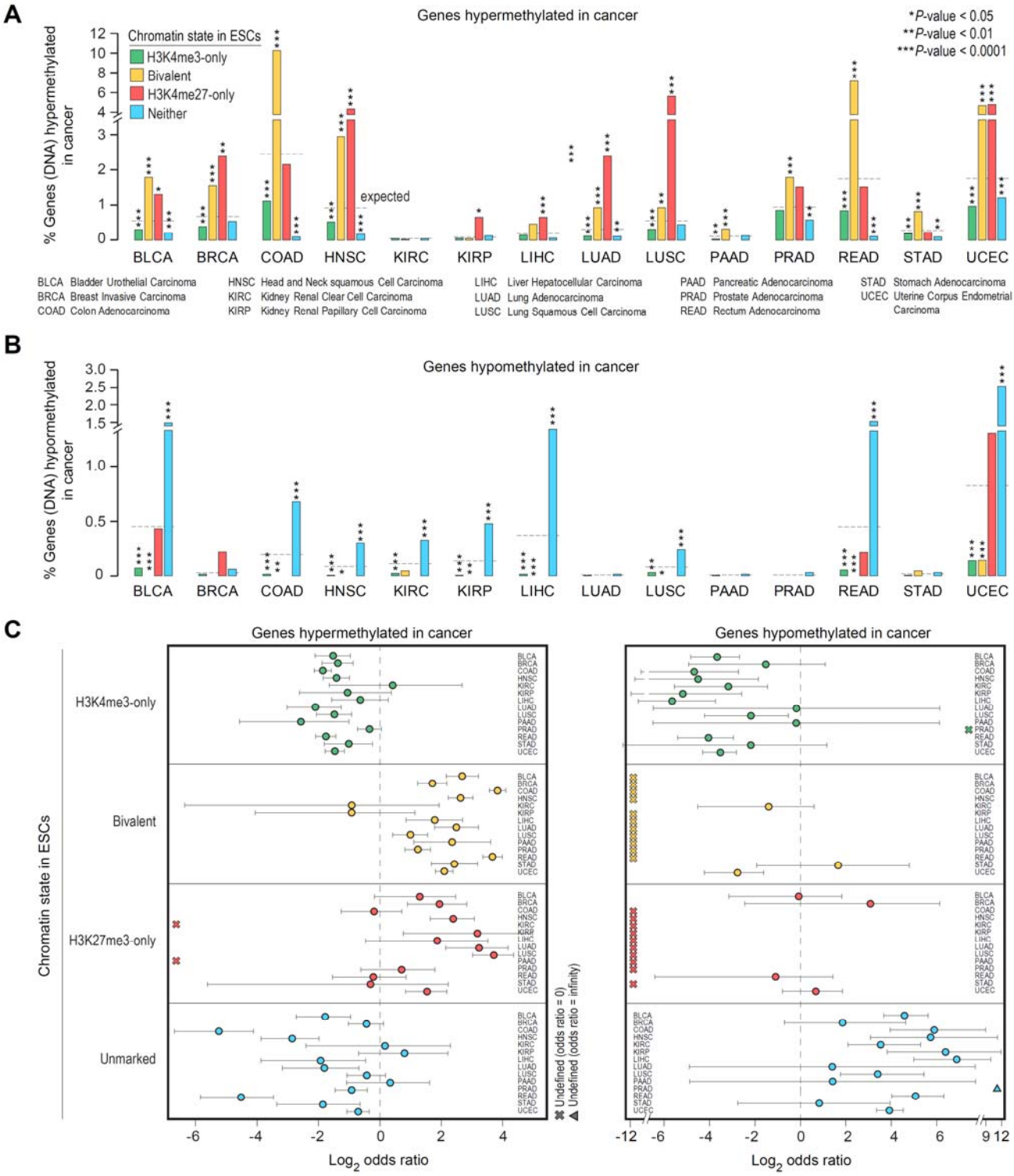
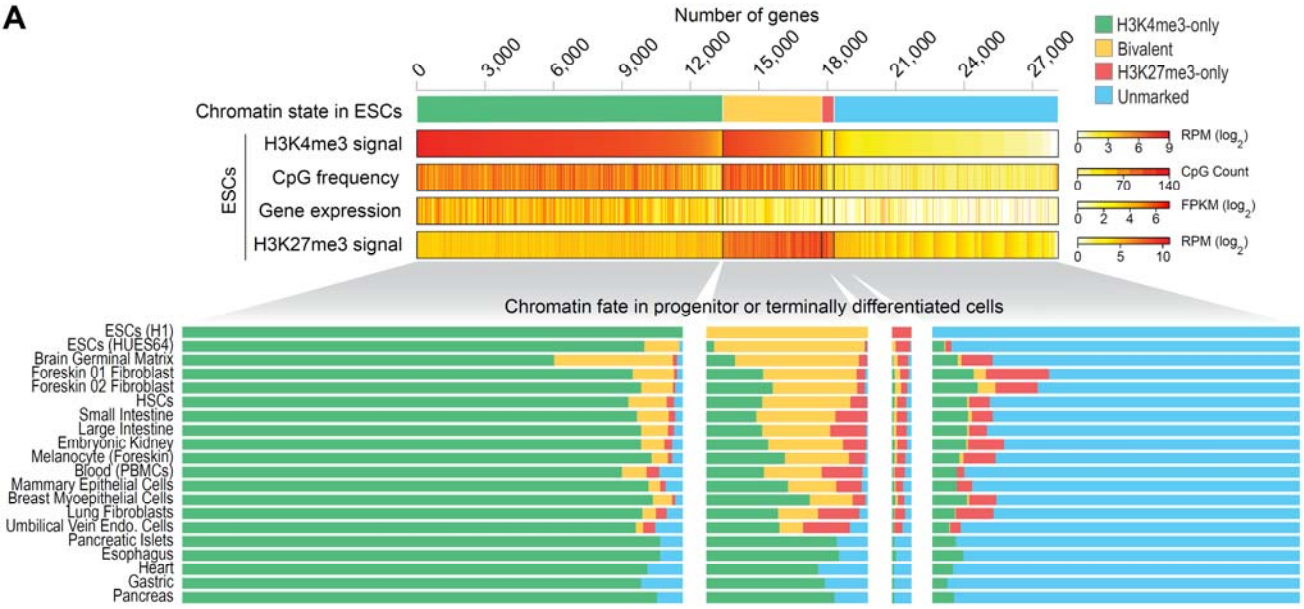
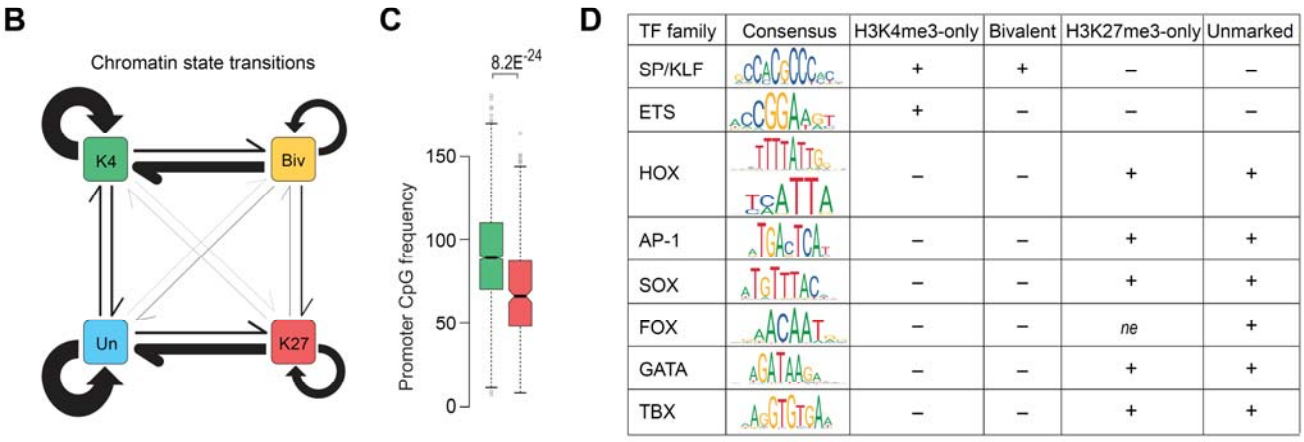



**Figure 6.** Genes hypermethylated in cancer are more likely to be PRC2 targets in ESCs. (***A, B***) Percentage of genes, within each of the four classes of genes defined in human ESCs, whose promoters are aberrantly DNA hypermethylated (A) or hypomethylated (***B***) in various cancer types. All the *P* values were calculated using Fisher's exact test. (***C***) $Log_2$ odds ratio, with 95% confidence intervals, of promoter DNA hypermethylation (left) or hypomethylation (right) in various cancer types based on their chromatin state in human ESCs.

334 silenced, respectively (Azuara et al. 2006; Bernstein et al. 2006; Voigt et al. 2013). To definitively
335 determine the fates of bivalent genes, we examined chromatin states of gene promoters across
336 various cell types. Our analysis revealed that a vast majority (83%) of genes that are bivalent in ESCs
337 retain H3K4me3 in other cell types, with 52% resolving into H3K4me3-only chromatin state and 31%
338 remaining bivalent (**Fig. 7A; Supplemental Table 9**). Only a small fraction (10%) of bivalent genes in
339 ESCs resolve into H3K27me3-only state in other cell types. Because bivalent promoters are CpG-rich
340 (**Fig. 4D; Supplemental Fig. S8B**) and because CpG-rich sequences, when unmethylated, are
341 sufficient to establish H3K4me3 (Thomson et al. 2010), these data suggest that bivalent genes have
342 a predilection to resolve into their presumably default H3K4me3-only state in the absence of PRC2
343 activity (**Fig. 7B**). Consistent with this conclusion, we find bivalent promoters that resolve into
344 H3K4me3-only state in most cell types are more CpG-rich compared to those that resolve into
345 H3K27me3-only state in most cell types (**Fig. 7C; Supplemental Table 9**).

346 Next, to understand the establishment of the bivalent chromatin state, we focused on genes that
347 acquire bivalency in cell types with restricted potency and found that an overwhelming majority
348 (81%) of these genes are H3K4me3-only in ESCs and are CpG-rich (**Fig. 7A**). Our analysis of
349 chromatin fates during reconstituted pre- to post-implantation epiblast differentiation in mouse
350 revealed similar results (**Supplemental Fig. S10A**). Specifically, nearly all the genes that acquire
351 bivalency in EpiLCs were H3K4me3-only previously. These data further support our conclusion that
352 bivalent chromatin is the culmination of PRC2 activity at regions with high CpG density. Our findings
353 are consistent with studies linking EZH2 to the acquisition of 910 (out of 1,026; 88%) new bivalent
354 genes in germinal center B cells, almost all of which were previously H3K4me3-only in naive B cells
355 (Beguelin et al. 2013).




D

| TF family | Consensus | H3K4me3-only | Bivalent | H3K27me3-only | Unmarked |
|---|---|---|---|---|---|
| SP/KLF | | + | + | – | – |
| ETS | | + | – | – | – |
| HOX | | – | – | + | + |
| AP-1 | | – | – | + | + |
| SOX | | – | – | + | + |
| FOX | | – | – | *ne* | + |
| GATA | | – | – | + | + |
| TBX | | – | – | + | + |

**Figure 7.** Chromatin fate and sequence characteristics of bivalent promoters. (***A***) *Top*: Genes are grouped into four classes based on their chromatin state, defined based on H3K4me3 (+/- 500 bp of TSS) and/or H3K27me3 (+/- 2 kb of TSS) enrichment at gene promoters in human ESCs. H3K4me3 and H3K27me3 levels, CpG density (+/- 500 bp of TSS), and gene expression are shown. *Bottom*: Chromatin states of the same four gene classes in various cell types. (***B***) Schematic summarizing chromatin state transitions of gene promoters from one state to another. Arrows represent state transitions. The thicker the arrow, the more frequently observed that transition is. K4, H3K4me3-only; Biv, bivalent; K27, H3K27me3-only; Un, unmarked. (***C***) Boxplot showing distribution of CpG density at promoters of bivalent genes (in ESCs) that predominantly resolve into H3K4me3-only (green, left) or H3K27me3-only (red, right) state in other cell types. *P* value calculated using two-sided Wilcoxon rank-sum test. (***D***) Binding motifs for various transcription factor (TF) families and their enrichment within promoters (+/- 500 bp of TSS) of the four genes classes defined in *A*. +/- denotes over/under-enrichment, '*ne*' denotes no enrichment. See also Supplemental Fig. S10 and Supplemental Table 10.

Lastly, to gain further insight into sequence features—besides CpG density—that underlie bivalent promoters, we explored transcription factor (TF) binding motifs over-represented in H3K4me3-only, bivalent, H3K27me3-only and unmarked promoter classes in ESCs (**Supplemental Methods**). We

found motifs for ubiquitously expressed SP/KLF family of TFs generally over-represented (~2- to 5-fold) within mostly CpG-rich H3K4me3-only and bivalent promoters compared with mostly CpG-poor H3K27me3-only and unmarked promoters (**Fig. 7D; Supplemental Fig. S10B,C; Supplemental Table 10**). An exception to this is a small fraction of bivalent promoters—relatively CpG poor—that mostly resolve into H3K27me3-only state in other cell types (**Fig. 7C**); they exhibit no such enrichment for a subset of SP/KLF TF motifs (**Supplemental Fig. S10D**). Unlike CpG-rich promoters, the largely CpG-poor H3K27me3-only and unmarked promoters are characterized by over-representation of motifs recognized by families of TFs that are tissue-specific (e.g., HOX, AP-1, SOX, FOX, GATA, and TBX) (**Fig. 7D; Supplemental Fig. S10B,C**; **Supplemental Table 10**). Our analysis also revealed that H3K4me3-only but not bivalent promoters are characterized by over-representation of motifs for ETS family of TFs (**Supplemental Fig. S10B**), known to activate genes associated with variety of cellular house-keeping processes including cell cycle control, cell proliferation, and cellular differentiation. These data suggest that CpG-rich promoters that are enriched for motifs for SP/KLF but not ETS factors, when transcriptionally inactive, provide a fertile ground for PRC2 activity and establishment of bivalent chromatin, consistent with a causal role for GC-rich sequences—lacking activating TF motifs—in PRC2 recruitment (Mendenhall et al. 2010).

## DISCUSSION

Bivalent genes, by virtue of their exhibiting features of both transcriptionally active and repressed chromatin, are posited as being in a poised state—enabling them to be rapidly activated upon appropriate activation cues during development—while maintaining a transcriptionally repressed state (Azuara et al. 2006; Bernstein et al. 2006; Voigt et al. 2013).

Collectively, our studies reveal that bivalency does not poise genes for rapid activation but protects promoters from *de novo* DNA methylation. Activation of bivalent genes is neither greater nor more rapid than that of other transcriptionally silent genes that lack H3K4me3 at their promoters (**Fig. 2G**), challenging the premise that H3K4me3 at bivalent promoters is instructive for rapid activation of transcription. We find that promoter H3K4me3 levels are a product of the underlying CpG-rich DNA sequence, so much so that CpG density alone can predict H3K4me3 levels/enrichment reasonably accurately (**Fig. 4B,C**). This likely explains why H3K4me3 at bivalent promoters in one cell type persists in nearly all other cell types irrespective of gene expression (**Fig. 1**) and why unmethylated CGI promoters harbor H3K4me3 even when transcriptionally inactive (Guenther et al. 2007; Mikkelsen et al. 2007). Our findings are consistent with studies showing that high CpG-rich sequences, when unmethylated, *per se* are sufficient to establish H3K4me3 domains (Thomson et al. 2010)—even in the absence of RNAPII and sequence specific TFs (Vastenhouw et al. 2010)—but insufficient to induce transcriptional activity on chromatin (Hartl et al. 2019). Because bivalently marked promoters across various cell types overlap CpG-rich sequences (**Supplemental Fig. S8B**), which inherently are devoid of DNA methylation (Deaton and Bird 2011) and almost always positive for H3K4me3 (**Supplemental Fig. S8C**), establishment (as well as dissolution) of bivalent domains likely boil down to PRC2 activity (inactivity, respectively) at genomic regions with high CpG density. Supporting this notion, the number of bivalent genes in a given cell type strongly correlate with *EZH2* expression (catalytic subunit of PRC2) (**Fig. 4F**), with gain- or loss-of-function EZH2 mutation, respectively, associated with increased or decreased bivalency (Shema et al. 2016).

H3K4me3 and H3K27me3 are loosely referred to as “activating” and “repressive” marks respectively, but neither has been firmly established to play a causative role in the regulation of

gene expression. To the contrary, it was shown that PRC2/H3K27me3 is not required for the initiation of transcriptional repression of its targets, but is only required for the maintenance of the repressed state (Riising et al. 2014). Despite the general correlation between H3K4me3 and gene expression, it remains unclear as to whether H3K4me3 is instructive for transcription (Howe et al. 2017). Our results showing that the mere presence or the extent of H3K4me3 at bivalent genes does not confer an added advantage when it comes to rapid or higher activation potential **(Fig. 2)** suggest that H3K4me3 is not instructive for transcription activation. Consistent with this conclusion, deletion of *Mll2*—chiefly responsible for H3K4me3 at bivalent chromatin—in mouse ESCs resulted in no substantial disruption in the responsiveness of gene activation after retinoic acid treatment despite the almost complete loss of H3K4me3 and concomitant gain of H3K27me3 at bivalent promoters (Hu et al. 2013; Denissov et al. 2014). Our findings are also consistent with studies in yeast demonstrating that loss of H3K4me3 has no effect on the levels of nascent transcription and, conversely, loss of RNAPII has no effect on H3K4me3 levels (Murray et al. 2019). Together, these observations indicate that H3K4me3 is neither instructive for nor informed by transcription.

Besides its ability to predict transcription or chromatin states, the precise role(s) of H3K4me3 still remains elusive (Piunti and Shilatifard 2016). Our findings suggest that H3K4me3 is a better predictor of unmethylated CpGs than transcriptional activity and may be a general mechanism to maintain the hypomethylated state of CGIs, even when transcriptionally inactive. This would be consistent with studies showing that H3K4me3 repulses *de novo* methyltransferases *in vitro* (Ooi et al. 2007) and that complete erasure of H3K4me3 elevates DNA methylation levels (Hu et al. 2009; Rose and Klose 2014) but reversible DNA methylation has no impact on H3K4me3 levels (King et al. 2016). About 70% of mammalian promoters overlap with CGIs (Deaton and Bird 2011). Although

CpG dinucleotides are substrates for DNA methyltransferases, few CGI promoters gain methylation—even when transcriptionally inactive—during normal development (Deaton and Bird 2011; Jones 2012; Bestor et al. 2015; Schubeler 2015). However, most genes that are hypermethylated in cancer have CGI promoters and are bivalently marked in ESCs, which led to the proposition that bivalency predisposes them for aberrant *de novo* DNA methylation and irreversible silencing in cancer (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007). Recent studies have also shown an association between hypermethylation of bivalent promoters in cancer and acquired resistance to chemotherapy (Curry et al. 2018). Our findings suggest reveal that bivalency and H3K4me3 in particular protects promoters from *de novo* methylation during pre- to post-implantation epiblast differentiation and that aberrant hypermethylation in cancer may be explained by the loss of H3K4me3/bivalency (**Fig. 5**). In other words, it may be that it is not the bivalency but the loss of bivalency that make bivalent genes more susceptible to aberrant DNA methylation in diseases such as cancer. This would be consistent with studies showing cancer cell lines exhibiting a general loss of bivalency (Bernhart et al. 2016), and bivalent promoters with high H3K27me3:H3K4me3 ratio being targets for DNA hypermethylation in cancer (Dunican et al. 2020).

CGI promoters, the superset containing bivalent promoters, are relatively nucleosome-deficient, intrinsically accessible without the need for ATP-dependent nucleosome displacement, and transcriptionally permissive (Ramirez-Carrozzi et al. 2009). So, what keeps transcriptionally repressed bivalent promoters from getting transcribed? Most bivalent promoters lack transcriptionally engaged RNAPII (Williams et al. 2015) but harbor what is referred to as 'poised' RNAPII (preferentially phosphorylated at serine 5 but not serine 2), and it has been suggested that 'poised' RNAPII primes bivalent genes for rapid activation (Stock et al. 2007; Brookes et al. 2012;

Tee et al. 2014; Ferrai et al. 2017). Because promoter-proximal pausing of RNAPII is not a common mechanism employed at bivalent genes (Min et al. 2011; Williams et al. 2015), it is less likely that the 'poised' RNAPII at bivalent promoters represents some form of paused/stalled RNAPII competent for rapid transcription re-activation. Our analyses reveal that bivalent promoters are devoid of TFIIH and that any serine 5 phosphorylation on RNAPII at bivalent promoters is attributable to MAPK1 (**Fig. 3)**, known to bind exclusively to a subset of PRC2 targets and phosphorylate serine 5 on RNAPII (Tee et al. 2014). Because MAPK1 and TFIIH are mutually exclusive at their target promoters and because RNAPII-S5p levels at 'poised' bivalent promoters correlate with H3K27me3 levels (and not H3K4me3, as observed at active promoters), it is conceivable that MAPK1-mediated phosphorylation of serine 5 on RNAPII (or MAPK1's mere presence on chromatin) is refractory to transcription. In this scenario, MAPK1 and/or the substrate it modifies on RNAPII (one or more CTD heptad repeats) may antagonize TFIIH, and activation of transcription likely occurs only upon loss of MAPK1 binding and/or MAPK1-mediated phosphorylation of serine 5 on RNAPII, which may be followed by binding of appropriate transcription factors at promoters and/or enhancers. Further studies are required to ascertain any potential antagonism between MAPK1 and TFIIH.

*Mll2* is dispensable for maintaining ESC self-renewal, but *Mll2* deficiency is embryonic lethal. *Mll2* knock-out (KO) mice exhibit growth defects as early as ~E6.5 and die at ~E10.5 (Glaser et al. 2006), suggesting that MLL2 is not required until after implantation, right when *Dnmt3a* and *Dnmt3b* get induced to carry out global *de novo* methylation in early post-implantation embryo. Furthermore, *in vitro* differentiation of *Mll2* KO ESCs results in impaired embryoid body formation, with many bivalent genes with key functions in embryonic development and differentiation failing to activate

or exhibiting delayed activation kinetics (Lubitz et al. 2007; Mas et al. 2018), indicating an essential role for MLL2 during ESC differentiation. Our finding that H3K4me3 at transcriptionally repressed bivalent promoters, catalyzed primarily by MLL2 (Hu et al. 2013; Denissov et al. 2014), confers protection against *de novo* DNA methylation during pre- to post-implantation epiblast differentiation (**Fig. 5A,B,E**) suggests that the requirement for MLL2 after implantation—when it is no longer the major H3K4 trimethyltransferase (Glaser et al. 2006)—at least in part might have to do with its role in implementing H3K4me3 at bivalent genes in order to maintain epigenetic plasticity by protecting against *de novo* DNA methylation and thus irreversible silencing. Because ESCs do not express *Dnmt3a* or *Dnmt3b* and are DNA hypomethylated, this could perhaps explain why MLL2 is dispensable in mouse ESCs. Moreover, a recent study showed that MLL2—which also is responsible for H3K4me3 at a vast majority of transcriptionally active genes (Denissov et al. 2014)—protects about 2% of MLL2 -dependent active genes from DNMT1-mediated maintenance methylation (Douillet et al. 2020), highlighting MLL2's multifaceted role in regulating gene expression.

In summary, our findings suggest a unifying model (**Fig. 8**) wherein bivalency maintains epigenetic plasticity by protecting gene promoters from irreversible silencing while maintaining a reversibly repressed state, and that loss of H3K4me3 may make them more susceptible to aberrant DNA methylation in diseases such as cancer. One limitation of our study is that the ESC to EpiLC differentiation model we used to investigate bivalency only recapitulates events during pre- to post-implantation epiblast development (~E3.75-5.75) and does not provide sufficient time for all of bivalency to resolve. Assessing the fate of all bivalent genes through all stages of development

would require further investigation using differentiation models that recapitulate later developmental stages through directed differentiation of ESCs to specific lineages.

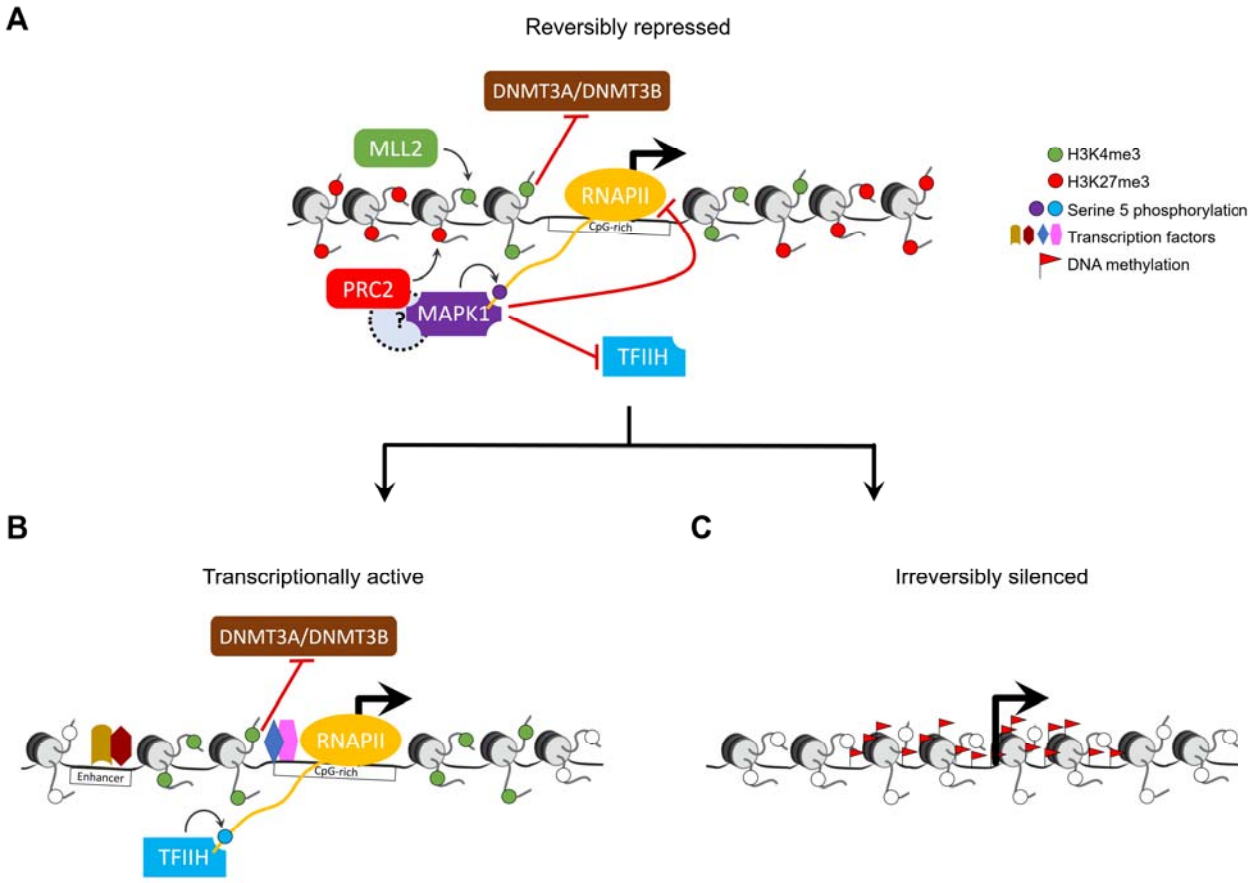



**Figure 8.** Model for bivalent chromatin maintaining epigenetic plasticity by protecting gene promoters from irreversible silencing while maintaining a reversibly repressed state. (***A***) H3K4me3, catalyzed primarily by the MLL2/COMPASS complex, protects CpG-rich bivalent promoters from DNA methylation by repelling *de novo* methyltransferases Dnmt3a and Dnmt3b. Bivalent promoters, by virtue of their overlapping CGIs, are more averse to assembling into nucleosomes compared to other genomic DNA (Ramirez-Carrozzi et al. 2009); consequently, bivalent promoters are relatively nucleosome-deficient, intrinsically accessible, and transcriptionally permissive (Deaton and Bird 2011; Mas et al. 2018). The assembly of transcription machinery at the promoter triggers PRC2 to recruit— either directly or indirectly through yet-to-be-determined mechanism—reinforcement in the form of MAPK1 (ERK2), which, in lieu of the usual TFIIH, phosphorylates serine 5 on a particular (or a set of) RNAPII CTD heptad repeat(s) (Tee et al. 2014). The presence of MAPK1 and/or the ensuing serine 5 phosphorylation is refractory to transcription as it antagonizes the recruitment of the TFIIH complex, which, besides its role in phosphorylating serine 5 on RNAPII, is necessary to unwind promoter DNA to form transcription bubble for RNA synthesis. (***B***) Activation of transcription at bivalent genes likely occurs only upon loss of PRC2 activity and thus MAPK1 activity, followed by binding of appropriate transcription factors at promoters and/or enhancers. (***C***) Loss of H3K4me3 at CpG-rich bivalent promoters make them more susceptible to aberrant DNA methylation during aging and in diseases such as cancer.

Poised enhancers, characterized by the co-occurrence of histone modifications generally associated with positive (H3K4me1) and negative (H3K27me3) transcriptional states, are another class of bivalent regions (Blanco et al. 2020). Similarly to bivalent promoters, they are considered to be

primed for future activation during embryo development, with loss of H3K27me3 and gain of H3K27ac associated with enhancer activation. Although some of our findings on bivalent promoters may be applicable to poised enhancers, further studies are needed to precisely characterize the functional relevance and mechanism of poised enhancers.

## METHODS

***Data sources.*** All data sets pertaining to this manuscript were obtained from previously published studies (for details, see **Supplemental Methods**).

***Histone modification enrichment at gene promoters.*** For each gene promoter, read densities (RPM, reads per million mapped reads) of individual histone modifications (H3K4me3/H3K27me3) and corresponding genomic input were calculated. Promoters were defined as the region spanning TSS ± 500 bp for H3K4me3 and TSS ± 2 kb for H3K27me3. A promoter is deemed to be enriched for a particular histone modification only if its ChIP signal (RPM) is at least (i) 3-fold greater than its input signal (RPM), and (ii) greater than a threshold (1% FDR), estimated as the lowest RPM value at which the number of qualifying promoters (RPM greater than or equal to the threshold) based on the input signal (RPM) is less than 1% of the number of qualifying promoters based on the ChIP signal (RPM). Promoters enriched for both H3K4me3 and H3K27me3 were defined as bivalent. Promoters enriched for H3K4me3 but not H3K27me3 were defined as H3K4me3-only, and those that are enriched for H3K27me3 but not H3K4me3 were defined as H3K27me3-only. Promoters with neither H3K4me3 nor H3K27me3 enrichment were defined as 'unmarked'. Custom scripts used to compute H3K4me3/H3K27me3 enrichment at gene promoters, define promoter chromatin state, and

generate data for making tag density plots and heatmaps are available as **Supplemental Code** and at GitHub (https://github.com/DhirKumar/Biv_ChIP).

***RNA-seq data analysis.*** RNA-seq reads were aligned to the mouse genome (mm9) using STAR aligner (Dobin et al. 2013), allowing up to three mismatches, retaining only reads that align to unique genomic locations. Alignment files were used to quantitate expression (FPKM) of genes and isoforms annotated in the mm9 genome (source: NCBI RefSeq) using the cuffdiff tool (Trapnell et al. 2013) with default parameters and library type defined as fr-firststrand. The RefSeq-GTF file (downloaded from the UCSC database) for the mm9 genome build was supplied to cuffdiff tool as the reference. The resultant "isoforms.fpkm_tracking" file from the cuffdiff run was used to infer differentially expressed genes ($q$-value < 0.05) and fold changes between two cell-types/time-points of interest. In the case of multiple promoters giving rise to alternative isoforms, we used one-to-one association of promoter chromatin state and isoform expression. For human, processed RNA-seq signal (FPKM) for various cell types were obtained from NIH Roadmap Epigenomics project (Roadmap Epigenomics Consortium et al. 2015) (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz). RefSeq accession (NM/NR ID) was used to integrate gene expression and chromatin-level data.

***Mouse DNA methylation analysis.*** Processed promoter DNA methylation data from naïve mouse ESCs and EpiLCs were obtained from a previously published study (Shirane et al. 2016). liftOver tool was used to map gene coordinates between mm9 and mm10 assemblies. A gene promoter is defined as hypermethylated in EpiLCs compared to naïve ESCs only if (i) its methylation level in

EpiLCs is at least 50%, and (ii) its methylation level in EpiLCs is at least 2-fold greater than that in naïve ESCs.

***Human DNA methylation analysis.*** Genes DNA-hypermethylated (based on promoter DNA methylation levels) in primary colorectal tumor was obtained from a previously published study (Widschwendter et al. 2007). 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**). Genes DNA hypermethylated/hypomethylated in 14 TCGA solid epithelial cancer types were inferred from DNA-methylation (Illumina 450K array) beta values for 6,129 tumors and respective control tissues, generated by the TCGA Research Network (for details, see **Supplemental Methods**). Genes DNA-hypermethylated during aging were inferred by mapping processed probe-level methylation data (Rakyan et al. 2010) to gene promoter coordinates using Infinium manifest file.

## COMPETING INTEREST STATEMENT

The authors declare no competing interests.

## ACKNOWLEDGMENTS

We thank members of the Jothi Lab for insightful discussions and comments on the manuscript. We thank J. Rodriguez, P.A. Wade, and S. Yellaboina for critical comments on the manuscript. This work was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (1ZIAES102625 to R.J.).

## AUTHOR CONTRIBUTIONS

D.K. and R.J. conceived and designed the study; D.K. performed the research, with contributions from S.C., A.J.O., and P.Y; D.K. and R.J. analyzed the data and wrote the manuscript; all co-authors reviewed and edited the manuscript.

## REFERENCES

Allis CD, Jenuwein T. 2016. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17**: 487-500.

Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, John RM, Gouti M, Casanova M, Warnes G, Merkenschlager M et al. 2006. Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**: 532-538.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823-837.

Beguelin W, Popovic R, Teater M, Jiang Y, Bunting KL, Rosen M, Shen H, Yang SN, Wang L, Ezponda T et al. 2013. EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* **23**: 677-692.

Bernhart SH, Kretzmer H, Holdt LM, Juhling F, Ammerpohl O, Bergmann AK, Northoff BH, Doose G, Siebert R, Stadler PF et al. 2016. Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Sci Rep* **6**: 37393.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315-326.

Bestor TH, Edwards JR, Boulard M. 2015. Notes on the role of dynamic DNA methylation in mammalian development. *Proc Natl Acad Sci U S A* **112**: 6796-6799.

Blanco E, Gonzalez-Ramirez M, Alcaine-Colet A, Aranda S, Di Croce L. 2020. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends Genet* **36**: 118-131.

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.

Buecker C, Srinivasan R, Wu Z, Calo E, Acampora D, Faial T, Simeone A, Tan M, Swigut T, Wysocka J. 2014. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**: 838-853.

Curry E, Zeller C, Masrour N, Patten DK, Gallon J, Wilhelm-Benartzi CS, Ghaem-Maghami S, Bowtell DD, Brown R. 2018. Genes Predisposed to DNA Hypermethylation during Acquired Resistance to Chemotherapy Are Identified in Ovarian Tumors by Bivalent Chromatin Domains at Initial Diagnosis. *Cancer Res* **78**: 1383-1391.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010-1022.

Denissov S, Hofemeister H, Marks H, Kranz A, Ciotta G, Singh S, Anastassiadis K, Stunnenberg HG, Stewart AF. 2014. Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development* **141**: 526-537.

Di Croce L, Helin K. 2013. Transcriptional regulation by Polycomb group proteins. *Nat Struct Mol Biol* **20**: 1147-1155.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Douillet D, Sze CC, Ryan C, Piunti A, Shah AP, Ugarenko M, Marshall SA, Rendleman EJ, Zha D, Helmin KA et al. 2020. Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. *Nat Genet* **52**: 615-625.

Dunican DS, Mjoseng HK, Duthie L, Flyamer IM, Bickmore WA, Meehan RR. 2020. Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells. *BMC Biol* **18**: 25.

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.

Easwaran H, Tsai HC, Baylin SB. 2014. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. *Mol Cell* **54**: 716-727.

Ferrai C, Torlai Triglia E, Risner-Janiczek JR, Rito T, Rackham OJ, de Santiago I, Kukalev A, Nicodemi M, Akalin A, Li M et al. 2017. RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation. *Mol Syst Biol* **13**: 946.

611 Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek AK, Kelley DR, Shishkin AA,
612 Issner R et al. 2013. Transcriptional and epigenetic dynamics during specification of human
613 embryonic stem cells. *Cell* **153**: 1149-1163.

614 Glaser S, Schaft J, Lubitz S, Vintersten K, van der Hoeven F, Tufteland KR, Aasland R, Anastassiadis K,
615 Ang SL, Stewart AF. 2006. Multiple epigenetic maintenance factors implicated by the loss of
616 Mll2 in mouse development. *Development* **133**: 1423-1432.

617 Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and
618 transcription initiation at most promoters in human cells. *Cell* **130**: 77-88.

619 Harikumar A, Meshorer E. 2015. Chromatin remodeling and bivalent histone modifications in
620 embryonic stem cells. *EMBO Rep* **16**: 1609-1619.

621 Hartl D, Krebs AR, Grand RS, Baubec T, Isbel L, Wirbelauer C, Burger L, Schubeler D. 2019. CG
622 dinucleotides enhance promoter activity independent of DNA methylation. *Genome Res* **29**:
623 554-563.

624 Hayashi K, Ohta H, Kurimoto K, Aramaki S, Saitou M. 2011. Reconstitution of the mouse germ cell
625 specification pathway in culture by pluripotent stem cells. *Cell* **146**: 519-532.

626 Howe FS, Fischl H, Murray SC, Mellor J. 2017. Is H3K4me3 instructive for transcription activation?
627 *Bioessays* **39**: 1-12.

628 Hu D, Garruss AS, Gao X, Morgan MA, Cook M, Smith ER, Shilatifard A. 2013. The Mll2 branch of the
629 COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat Struct Mol*
630 *Biol* **20**: 1093-1097.

631 Hu JL, Zhou BO, Zhang RR, Zhang KL, Zhou JQ, Xu GL. 2009. The N-terminus of histone H3 is required
632 for de novo DNA methylation in chromatin. *Proc Natl Acad Sci U S A* **106**: 22187-22192.

633 Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates
634 intrinsic and environmental signals. *Nat Genet* **33 Suppl**: 245-254.

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

637 Jones PA, Baylin SB. 2007. The epigenomics of cancer. *Cell* **128**: 683-692.

638 King AD, Huang K, Rubbi L, Liu S, Wang CY, Wang Y, Pellegrini M, Fan G. 2016. Reversible Regulation
639 of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem
640 Cells. *Cell Rep* **17**: 289-302.

641 Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**: 693-705.

642 Leitch HG, McEwen KR, Turp A, Encheva V, Carroll T, Grabole N, Mansfield W, Nashun B, Knezovich
643 JG, Smith A et al. 2013. Naive pluripotency is associated with global DNA hypomethylation. *Nat*
644 *Struct Mol Biol* **20**: 311-316.

645 Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* **128**: 707-719.

646 Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D. 2011. Identification of genetic
647 elements that autonomously determine DNA methylation states. *Nat Genet* **43**: 1091-1097.

648 Lubitz S, Glaser S, Schaft J, Stewart AF, Anastassiadis K. 2007. Increased apoptosis and skewed
649 differentiation in mouse embryonic stem cells lacking the histone methyltransferase Mll2. *Mol*
650 *Biol Cell* **18**: 2356-2366.

651 Margueron R, Reinberg D. 2011. The Polycomb complex PRC2 and its mark in life. *Nature* **469**: 343-
652 349.

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

656 Mas G, Blanco E, Ballare C, Sanso M, Spill YG, Hu D, Aoi Y, Le Dily F, Shilatifard A, Marti-Renom MA
657 et al. 2018. Promoter bivalency favors an open chromatin architecture in embryonic stem cells.
658 *Nat Genet* **50**: 1452-1462.

659 Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum
660 C, Jaffe DB et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated
661 cells. *Nature* **454**: 766-770.

662 Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M, Bernstein BE. 2010. GC-rich
663 sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet* **6**: e1001244.

664 Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK,
665 Koche RP et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-
666 committed cells. *Nature* **448**: 553-560.

667 Min IM, Waterfall JJ, Core LJ, Munroe RJ, Schimenti J, Lis JT. 2011. Regulating RNA polymerase
668 pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**: 742-754.

669 Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D. 2008. Lineage-
670 specific polycomb targets and de novo DNA methylation define restriction and potential of
671 neuronal progenitors. *Mol Cell* **30**: 755-766.

672 Murray SC, Lorenz P, Howe FS, Wouters M, Brown T, Xi S, Fischl H, Khushaim W, Rayappu JR, Angel
673 A et al. 2019. H3K4me3 is neither instructive for, nor informed by, transcription. *bioRxiv*
674 **709014**.

675 Niccoli T, Partridge L. 2012. Ageing as a risk factor for disease. *Curr Biol* **22**: R741-752.

676 Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu
677 W et al. 2007. A stem cell-like chromatin pattern may predispose tumor suppressor genes to
678 DNA hypermethylation and heritable silencing. *Nat Genet* **39**: 237-242.

679 Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential
680 for de novo methylation and mammalian development. *Cell* **99**: 247-257.

681 Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin SP, Allis CD et al.
682 2007. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA.
683 *Nature* **448**: 714-717.

684 Phatnani HP, Greenleaf AL. 2006. Phosphorylation and functions of the RNA polymerase II CTD.
685 *Genes Dev* **20**: 2922-2936.

686 Piunti A, Shilatifard A. 2016. Epigenetic balance of gene expression by Polycomb and COMPASS
687 families. *Science* **352**: aad9780.

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

691 Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M,
692 Smale ST. 2009. A unifying model for the selective regulation of inducible transcription by CpG
693 islands and nucleosome remodeling. *Cell* **138**: 114-128.

694 Riising EM, Comet I, Leblanc B, Wu X, Johansen JV, Helin K. 2014. Gene silencing triggers polycomb
695 repressive complex 2 recruitment to CpG islands genome wide. *Mol Cell* **55**: 347-360.

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

699 Rose NR, Klose RJ. 2014. Understanding the relationship between DNA methylation and histone
700 lysine methylation. *Biochim Biophys Acta* **1839**: 1362-1372.

701 Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-
702 Shushan E, Reubinoff BE et al. 2007. Polycomb-mediated methylation on Lys27 of histone H3
703 pre-marks genes for de novo methylation in cancer. *Nat Genet* **39**: 232-236.

704 Schmitges FW, Prusty AB, Faty M, Stutzer A, Lingaraju GM, Aiwazian J, Sack R, Hess D, Li L, Zhou S et
705 al. 2011. Histone methylation by PRC2 is inhibited by active chromatin marks. *Mol Cell* **42**: 330-
706 341.

707 Schubeler D. 2015. Function and information content of DNA methylation. *Nature* **517**: 321-326.

708 Schuettengruber B, Bourbon HM, Di Croce L, Cavalli G. 2017. Genome Regulation by Polycomb and
709 Trithorax: 70 Years and Counting. *Cell* **171**: 34-57.

710 Shema E, Jones D, Shoresh N, Donohue L, Ram O, Bernstein BE. 2016. Single-molecule decoding of
711 combinatorially modified nucleosomes. *Science* **352**: 717-721.

712 Shilatifard A. 2012. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in
713 development and disease pathogenesis. *Annu Rev Biochem* **81**: 65-95.

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

717 Simon JA, Kingston RE. 2009. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat*
718 *Rev Mol Cell Biol* **10**: 697-708.

719 Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, Meissner A. 2012. A unique regulatory
720 phase of DNA methylation in the early mammalian embryo. *Nature* **484**: 339-344.

721 Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, Koseki H, Brockdorff N, Fisher AG, Pombo A.
722 2007. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent
723 genes in mouse ES cells. *Nat Cell Biol* **9**: 1428-1435.

724 Tee WW, Shen SS, Oksuz O, Narendra V, Reinberg D. 2014. Erk1/2 activity promotes chromatin
725 features and RNAPII phosphorylation at developmental promoters in mouse ESCs. *Cell* **156**: 678-
726 690.

727 Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James
728 KD et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1.
729 *Nature* **464**: 1082-1086.

730 Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of
731 gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**: 46-53.

732 Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J, Schier AF. 2010. Chromatin
733 signature of embryonic pluripotency is established during genome activation. *Nature* **464**: 922-
734 926.

735 Voigt P, LeRoy G, Drury WJ, 3rd, Zee BM, Son J, Beck DB, Young NL, Garcia BA, Reinberg D. 2012.
736 Asymmetrically modified nucleosomes. *Cell* **151**: 181-193.

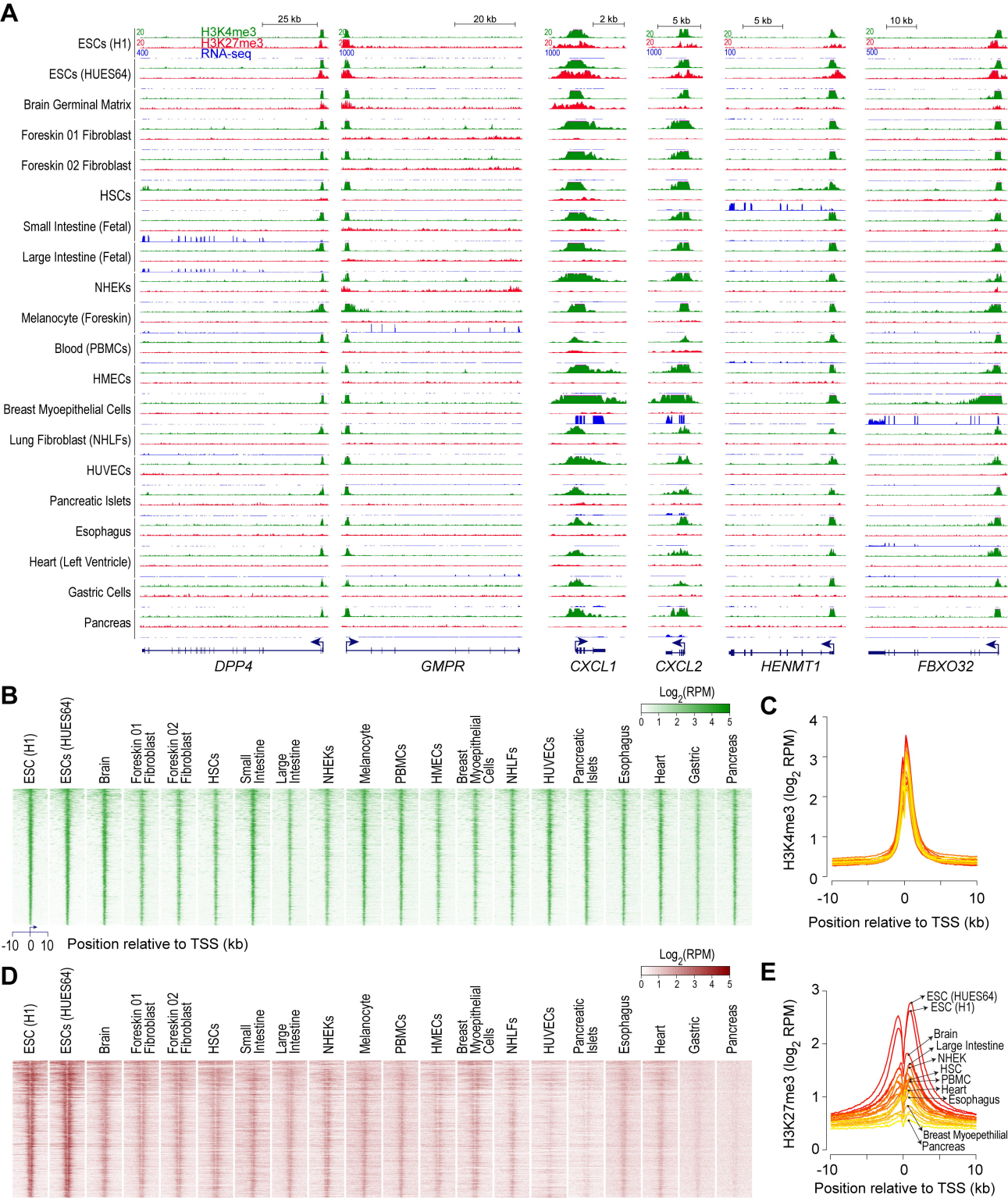
737 Voigt P, Tee WW, Reinberg D. 2013. A double take on bivalent promoters. *Genes Dev* **27**: 1318-1338.

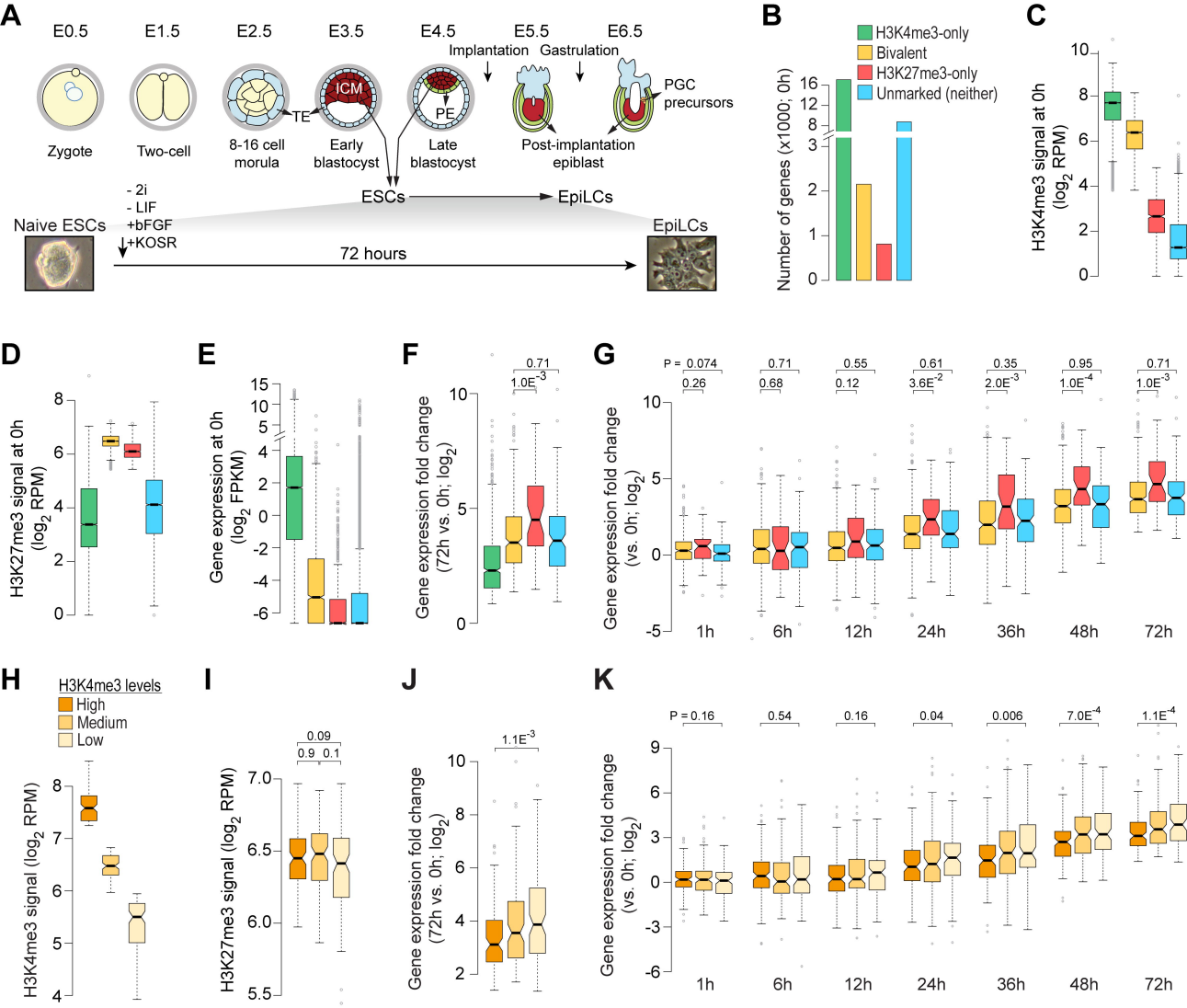
738 Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisenberger DJ, Campan
739 M, Young J, Jacobs I et al. 2007. Epigenetic stem cell signature in cancer. *Nat Genet* **39**: 157-158.

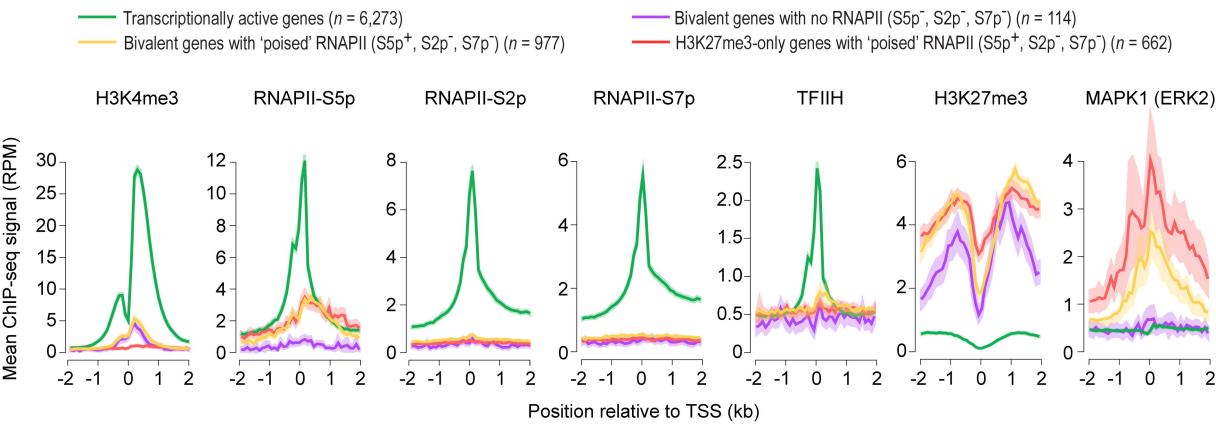
740 Williams LH, Fromm G, Gokey NG, Henriques T, Muse GW, Burkholder A, Fargo DC, Hu G, Adelman
741 K. 2015. Pausing of RNA polymerase II regulates mammalian developmental potential through
742 control of signaling networks. *Mol Cell* **58**: 311-322.

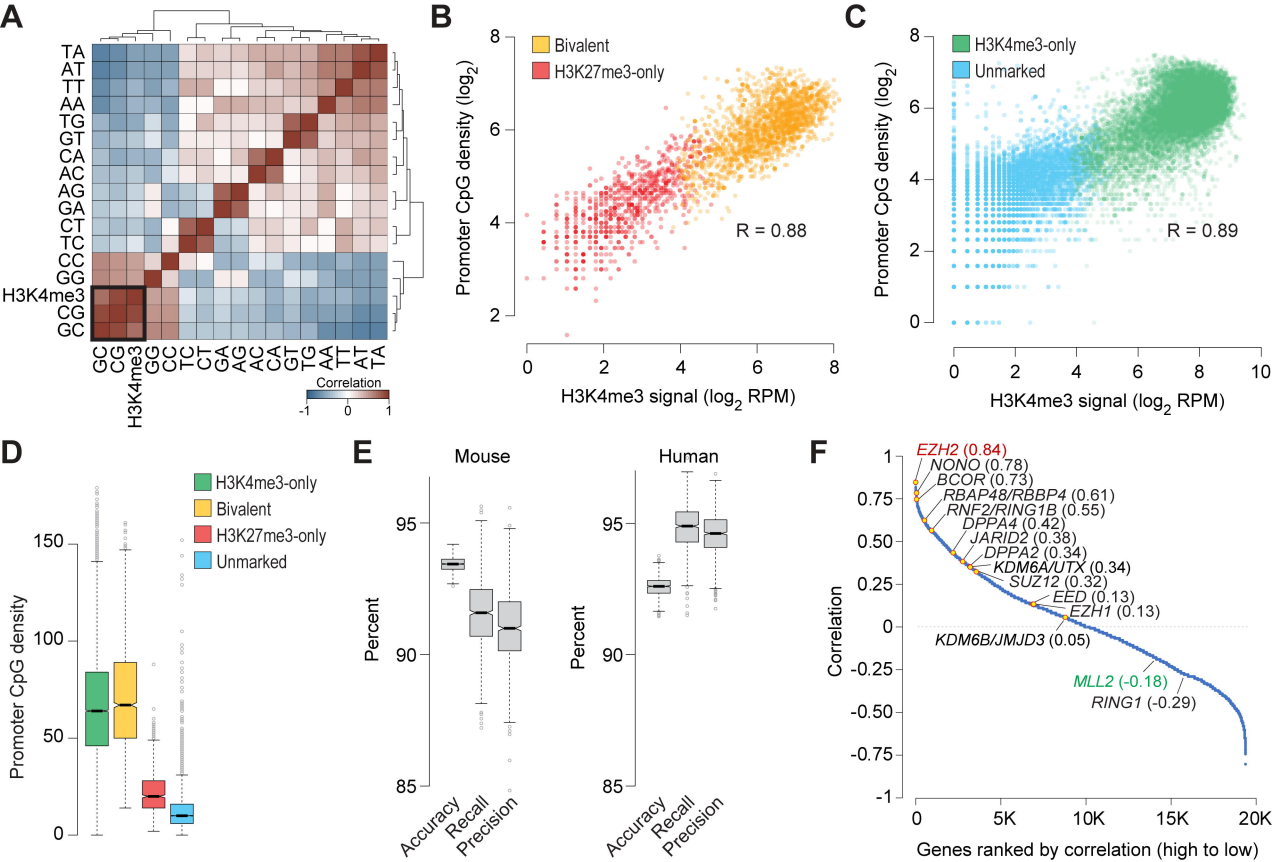
743 Yang P, Humphrey SJ, Cinghu S, Pathania R, Oldfield AJ, Kumar D, Perera D, Yang JYH, James DE,
744 Mann M et al. 2019. Multi-omic Profiling Reveals Dynamics of the Phased Progression of
745 Pluripotency. *Cell Syst* **8**: 427-445 e410.

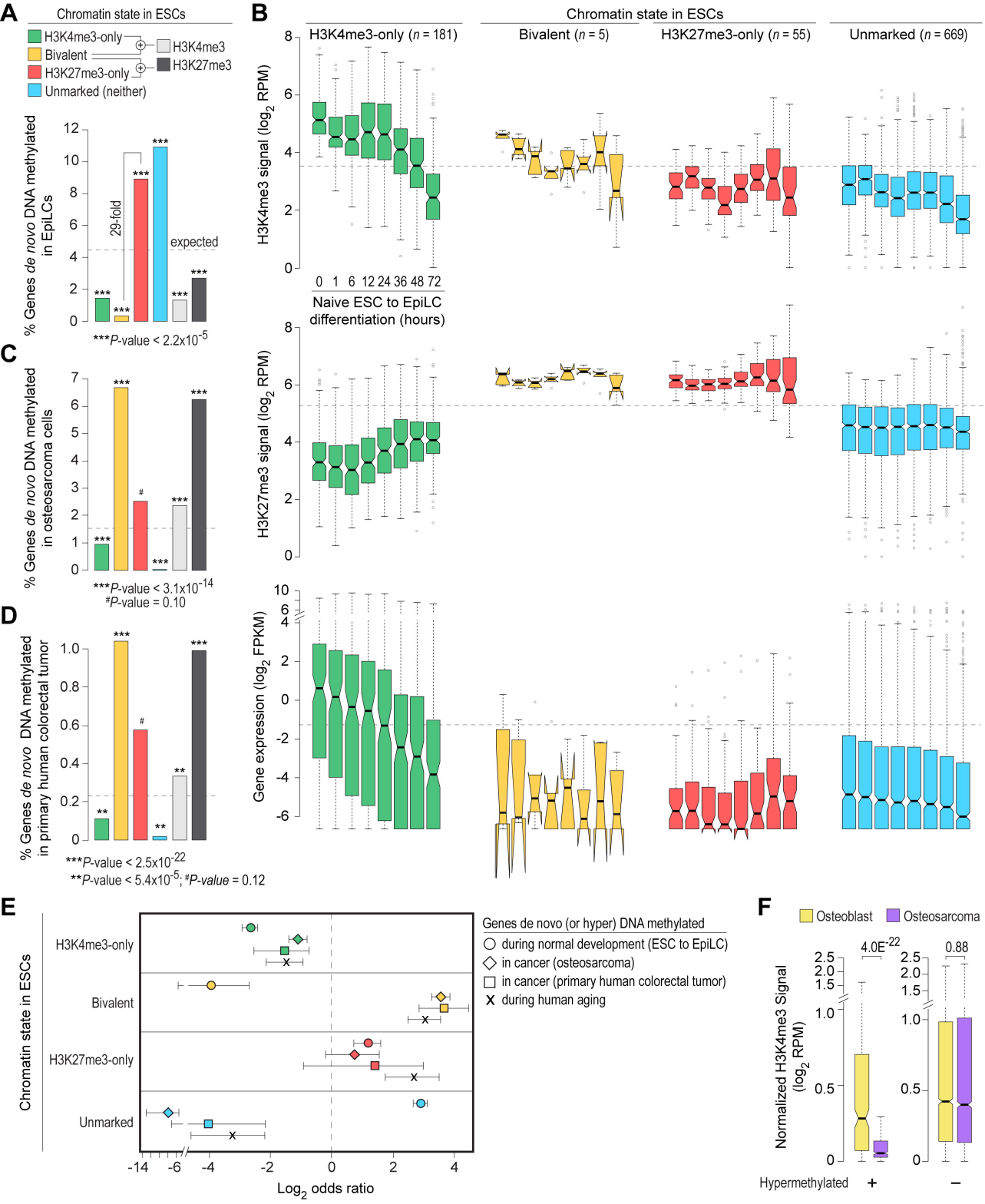
746

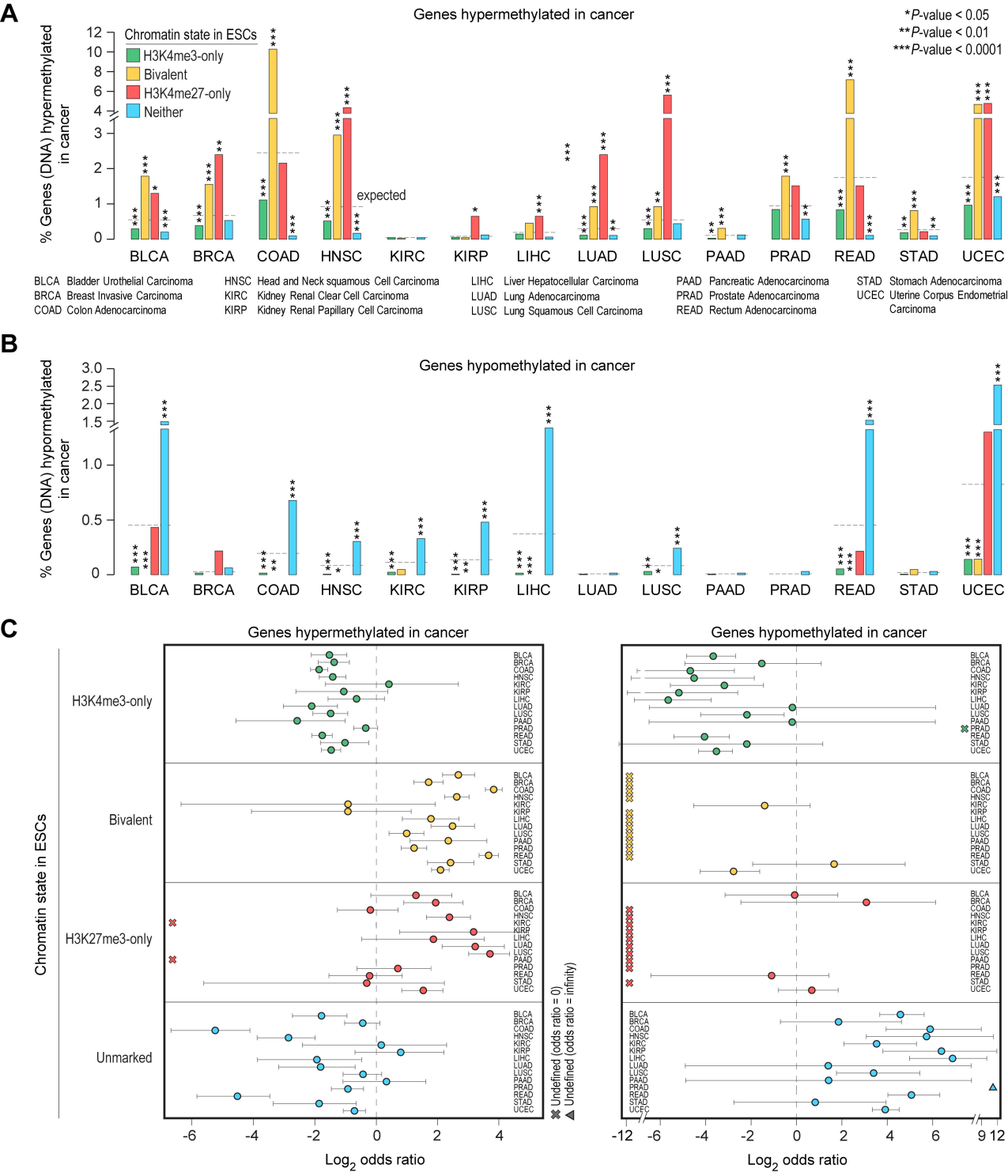

A
H3K4me3
H3K27me3
RNA-seq
25 kb
20 kb
2 kb
5 kb
5 kb
10 kb
ESCs (H1)
ESCs (HUES64)
Brain Germinal Matrix
Foreskin 01 Fibroblast
Foreskin 02 Fibroblast
HSCs
Small Intestine (Fetal)
Large Intestine (Fetal)
NHEKs
Melanocyte (Foreskin)
Blood (PBMCs)
HMECs
Breast Myoepithelial Cells
Lung Fibroblast (NHLFs)
HUVECs
Pancreatic Islets
Esophagus
Heart (Left Ventricle)
Gastric Cells
Pancreas
DPP4
GMPR
CXCL1
CXCL2
HENMT1
FBXO32
B
Log2(RPM)
ESC (H1)
ESCs (HUES64)
Brain
Foreskin 01 Fibroblast
Foreskin 02 Fibroblast
HSCs
Small Intestine
Large Intestine
NHEKs
Melanocyte
PBMCs
HMECs
Breast Myoepithelial Cells
NHLFs
HUVECs
Pancreatic Islets
Esophagus
Heart
Gastric
Pancreas
-10 0 10
Position relative to TSS (kb)
C
H3K4me3 (log2 RPM)
Position relative to TSS (kb)
D
Log2(RPM)
E
H3K27me3 (log2 RPM)
ESC (HUES64)
ESC (H1)
Brain
Large Intestine
NHEK
HSC
PBMC
Heart
Esophagus
Breast Myoepithelial
Pancreas
Position relative to TSS (kb)

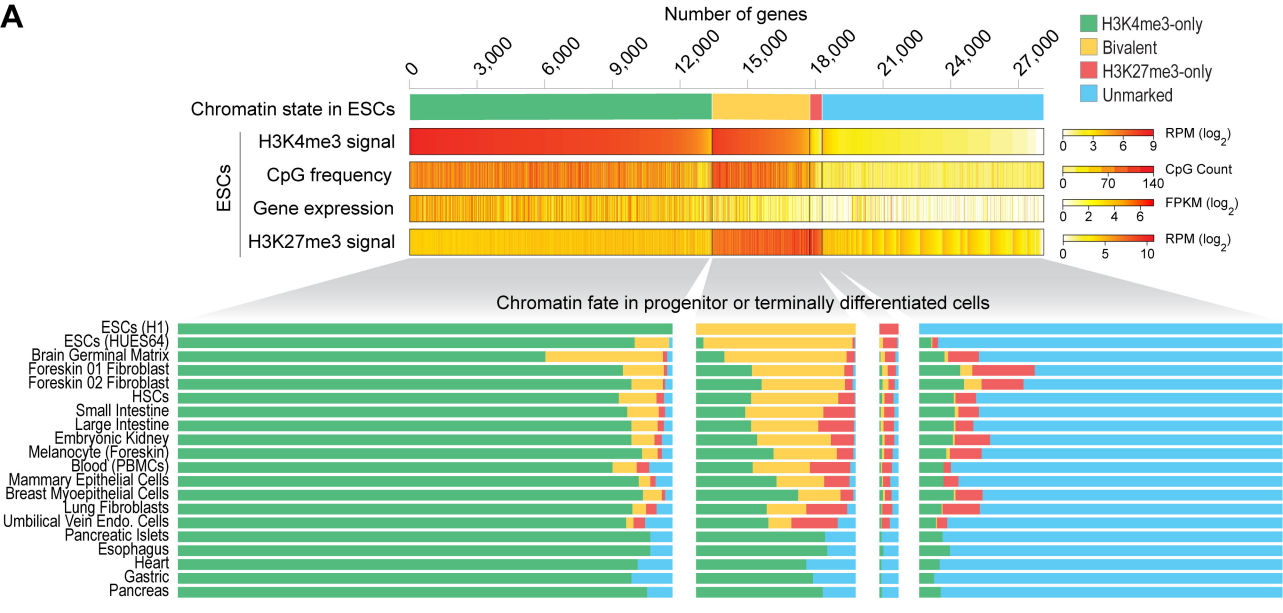

A
E0.5
E1.5
E2.5
E3.5
E4.5
Implantation
E5.5
Gastrulation
E6.5
Zygote
Two-cell
8-16 cell morula
TE
ICM
Early blastocyst
PE
Late blastocyst
Post-implantation epiblast
PGC precursors
ESCs
EpiLCs
Naive ESCs
- 2i
- LIF
+bFGF
+KOSR
72 hours
EpiLCs
B
H3K4me3-only
Bivalent
H3K27me3-only
Unmarked (neither)
Number of genes (x1000; 0h)
C
H3K4me3 signal at 0h (log2 RPM)
D
H3K27me3 signal at 0h (log2 RPM)
E
Gene expression at 0h (log2 FPKM)
F
0.71
1.0E-3
Gene expression fold change (72h vs. 0h; log2)
G
P = 0.074
0.26
0.71
0.68
0.55
0.12
0.61
3.6E-2
0.35
2.0E-3
0.95
1.0E-4
0.71
1.0E-3
Gene expression fold change (vs. 0h; log2)
1h
6h
12h
24h
36h
48h
72h
H
H3K4me3 levels
High
Medium
Low
H3K4me3 signal (log2 RPM)
I
0.09
0.9
0.1
H3K27me3 signal (log2 RPM)
J
1.1E-3
Gene expression fold change (72h vs. 0h; log2)
K
P = 0.16
0.54
0.16
0.04
0.006
7.0E-4
1.1E-4
Gene expression fold change (vs. 0h; log2)
1h
6h
12h
24h
36h
48h
72h

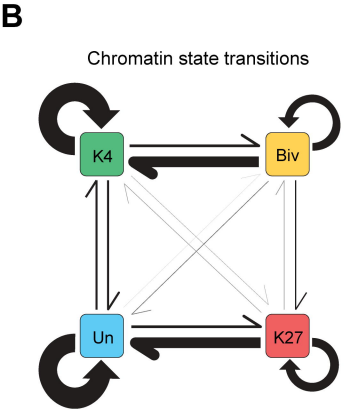

Transcriptionally active genes (n = 6,273)
Bivalent genes with 'poised' RNAPII (S5p+, S2p-, S7p-) (n = 977)
Bivalent genes with no RNAPII (S5p-, S2p-, S7p-) (n = 114)
H3K27me3-only genes with 'poised' RNAPII (S5p+, S2p-, S7p-) (n = 662)
H3K4me3
RNAPII-S5p
RNAPII-S2p
RNAPII-S7p
TFIIH
H3K27me3
MAPK1 (ERK2)
Mean ChIP-seq signal (RPM)
Position relative to TSS (kb)

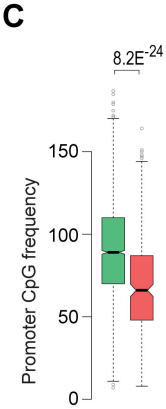

A
TA
AT
TT
AA
TG
GT
CA
AC
AG
GA
CT
TC
CC
GG
H3K4me3
CG
GC
GC
CG
H3K4me3
GG
CC
TC
CT
GA
AG
AC
CA
GT
TG
AA
TT
AT
TA
Correlation
-1
0
1
B
Bivalent
H3K27me3-only
Promoter CpG density (log2)
H3K4me3 signal (log2 RPM)
R = 0.88
C
H3K4me3-only
Unmarked
Promoter CpG density (log2)
H3K4me3 signal (log2 RPM)
R = 0.89
D
H3K4me3-only
Bivalent
H3K27me3-only
Unmarked
Promoter CpG density
E
Mouse
Human
Percent
Accuracy
Recall
Precision
F
EZH2 (0.84)
NONO (0.78)
BCOR (0.73)
RBAP48/RBBP4 (0.61)
RNF2/RING1B (0.55)
DPPA4 (0.42)
JARID2 (0.38)
DPPA2 (0.34)
KDM6A/UTX (0.34)
SUZ12 (0.32)
EED (0.13)
EZH1 (0.13)
KDM6B/JMJD3 (0.05)
MLL2 (-0.18)
RING1 (-0.29)
Correlation
Genes ranked by correlation (high to low)


A
Chromatin state in ESCs
H3K4me3-only
Bivalent
H3K27me3-only
Unmarked (neither)
H3K4me3
H3K27me3
% Genes de novo DNA methylated in EpiLCs
29-fold
expected
***P-value < 2.2x10-5
B
Chromatin state in ESCs
H3K4me3-only (n = 181)
Bivalent (n = 5)
H3K27me3-only (n = 55)
Unmarked (n = 669)
H3K4me3 signal (log2 RPM)
0 1 6 12 24 36 48 72
Naive ESC to EpiLC differentiation (hours)
H3K27me3 signal (log2 RPM)
Gene expression (log2 FPKM)
C
% Genes de novo DNA methylated in osteosarcoma cells
***P-value < 3.1x10-14
#P-value = 0.10
D
% Genes de novo DNA methylated in primary human colorectal tumor
***P-value < 2.5x10-22
**P-value < 5.4x10-5; #P-value = 0.12
E
Chromatin state in ESCs
H3K4me3-only
Bivalent
H3K27me3-only
Unmarked
Log2 odds ratio
Genes de novo (or hyper) DNA methylated
during normal development (ESC to EpiLC)
in cancer (osteosarcoma)
in cancer (primary human colorectal tumor)
during human aging
F
Osteoblast
Osteosarcoma
4.0E-22
0.88
Normalized H3K4me3 Signal (log2 RPM)
Hypermethylated
+
−


A
Genes hypermethylated in cancer
*P-value < 0.05
**P-value < 0.01
***P-value < 0.0001
Chromatin state in ESCs
H3K4me3-only
Bivalent
H3K4me27-only
Neither
% Genes (DNA) hypermethylated in cancer
expected
BLCA BRCA COAD HNSC KIRC KIRP LIHC LUAD LUSC PAAD PRAD READ STAD UCEC
BLCA Bladder Urothelial Carcinoma
BRCA Breast Invasive Carcinoma
COAD Colon Adenocarcinoma
HNSC Head and Neck squamous Cell Carcinoma
KIRC Kidney Renal Clear Cell Carcinoma
KIRP Kidney Renal Papillary Cell Carcinoma
LIHC Liver Hepatocellular Carcinoma
LUAD Lung Adenocarcinoma
LUSC Lung Squamous Cell Carcinoma
PAAD Pancreatic Adenocarcinoma
PRAD Prostate Adenocarcinoma
READ Rectum Adenocarcinoma
STAD Stomach Adenocarcinoma
UCEC Uterine Corpus Endometrial Carcinoma
B
Genes hypomethylated in cancer
% Genes (DNA) hypermethylated in cancer
C
Genes hypermethylated in cancer
Genes hypomethylated in cancer
Chromatin state in ESCs
H3K4me3-only
Bivalent
H3K27me3-only
Unmarked
Undefined (odds ratio = 0)
Undefined (odds ratio = infinity)
Log2 odds ratio
Log2 odds ratio



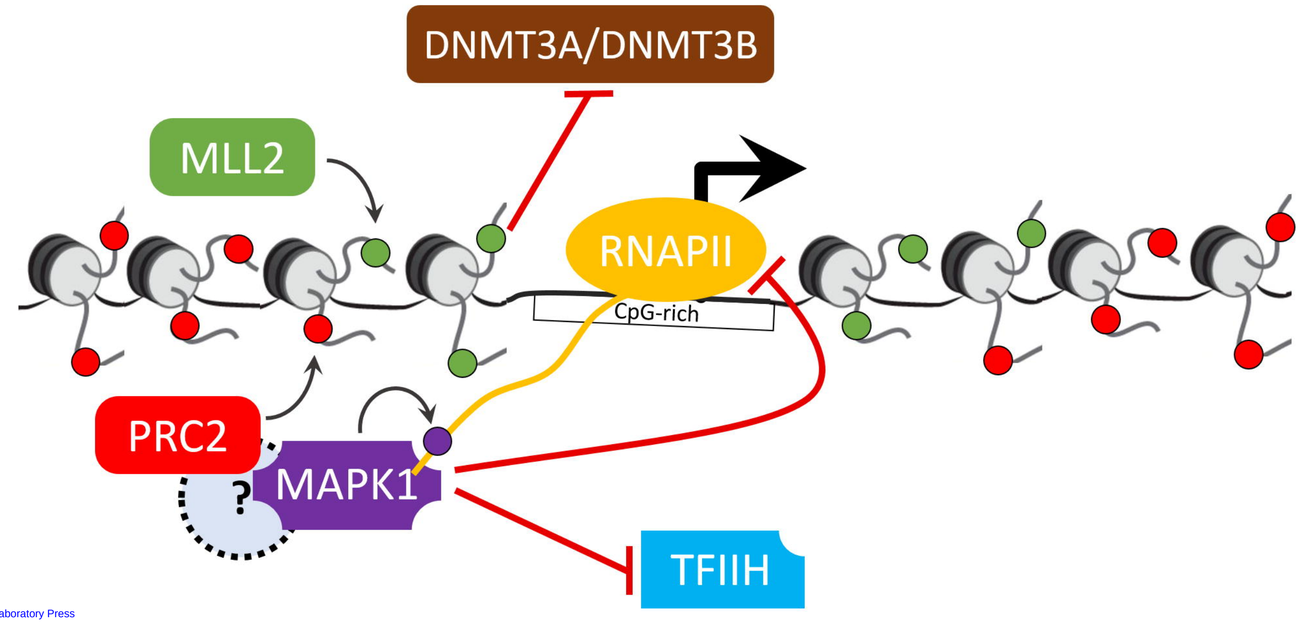



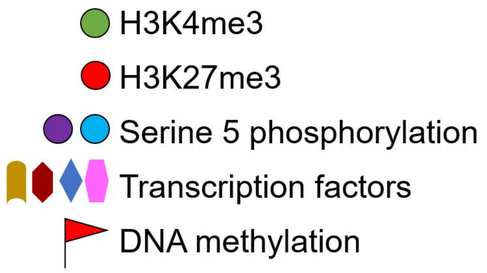

D

| TF family | Consensus | H3K4me3-only | Bivalent | H3K27me3-only | Unmarked |
|---|---|---|---|---|---|
| SP/KLF |  | + | + | – | – |
| ETS |  | + | – | – | – |
| HOX |  | – | – | + | + |
| AP-1 |  | – | – | + | + |
| SOX |  | – | – | + | + |
| FOX |  | – | – | *ne* | + |
| GATA |  | – | – | + | + |
| TBX |  | – | – | + | + |

**A**

Reversibly repressed

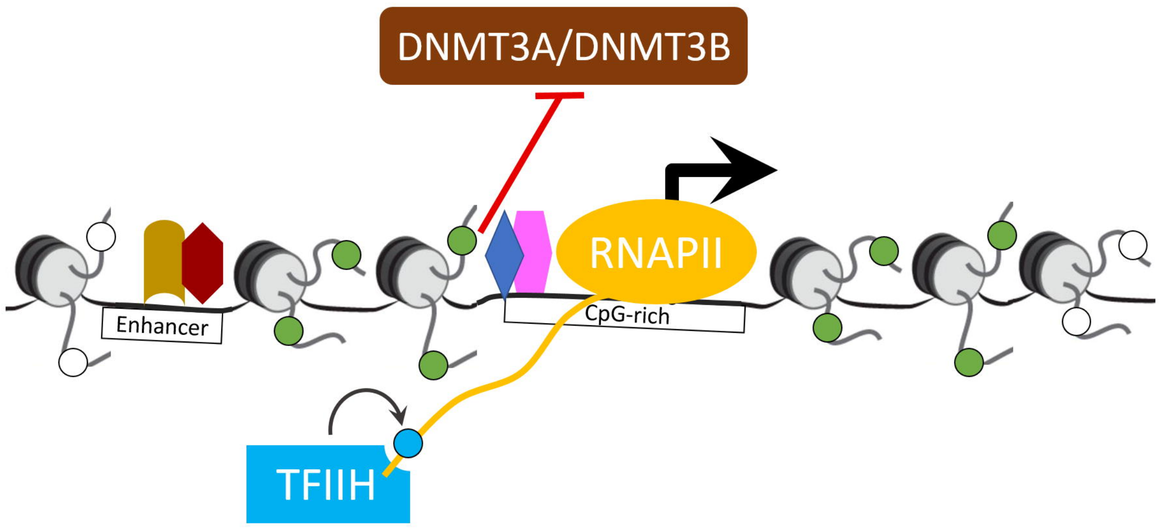





**B**

Transcriptionally active




**C**

Irreversibly silenced