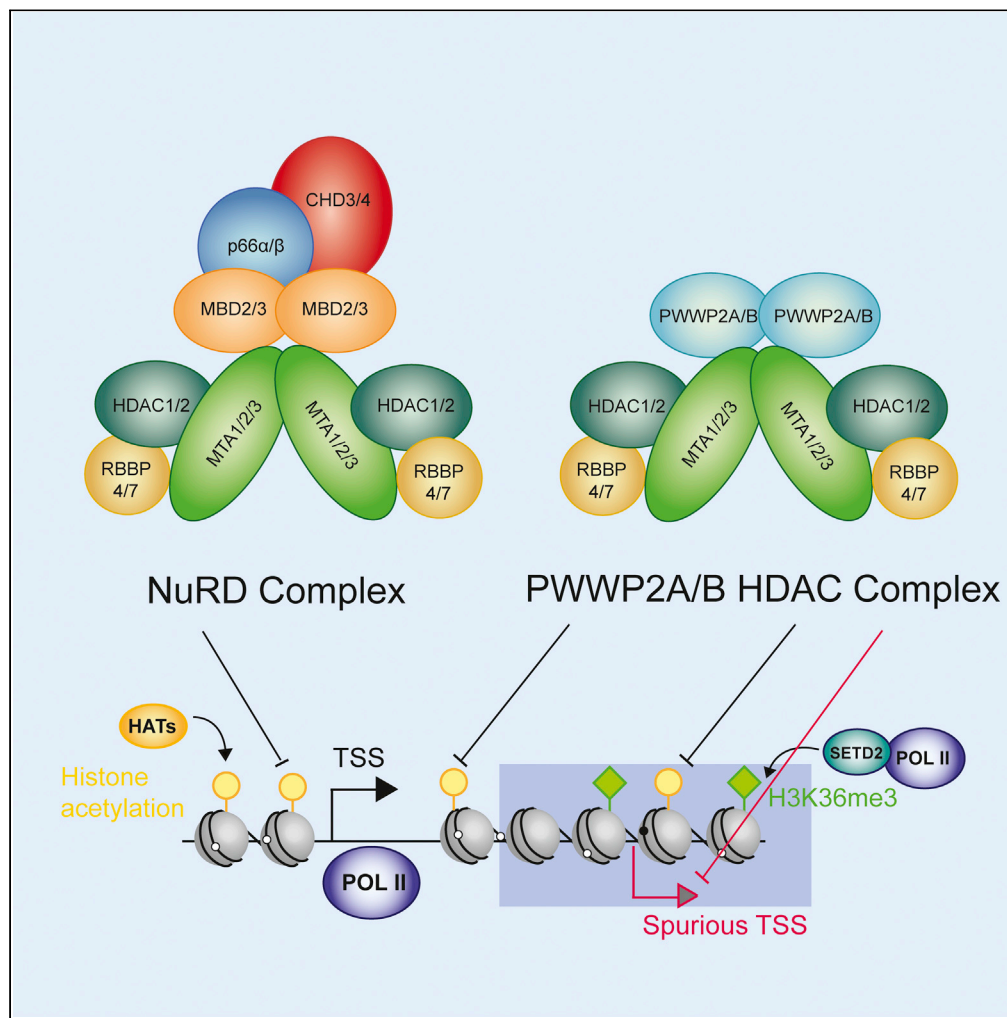


Article

The PWWP2A Histone Deacetylase Complex Represses Intragenic Spurious Transcription Initiation in mESCs



Guifeng Wei, Neil Brockdorff, Tianyi Zhang

neil.brockdorff@bioch.ox.ac.uk (N.B.)
tzhang17@ic.ac.uk (T.Z.)

HIGHLIGHTS

Loss of PWWP2A/B leads to increased levels of spurious transcription initiation

Spurious TSS sites are predominantly in the gene bodies of highly expressed genes

Spurious sites are marked with increased histone acetylation and initiating Pol II

PWWP2-spurious TSSs are distinct from those caused by DNMT3B loss



Article

The PWWP2A Histone Deacetylase Complex Represses Intragenic Spurious Transcription Initiation in mESCs

Guifeng Wei,¹ Neil Brockdorff,^{1,*} and Tianyi Zhang^{1,2,3,*}

SUMMARY

Transcriptional fidelity depends on accurate promoter selection and initiation from the correct sites. In yeast, H3K36me₃-mediated recruitment of the Rpd3S HDAC complex to gene bodies suppresses spurious transcription initiation. Here we describe an equivalent pathway in metazoans. PWWP2A/B is an H3K36me₃ reader that forms a stable complex with HDAC1/2. We used CAGE-seq to profile all transcription initiation sites in wild-type mESCs and cells lacking PWWP2A/B. Loss of PWWP2A/B enhances spurious initiation from intragenic sites present in wild-type mESCs, and this effect is associated with increased levels of initiating Pol-II and histone acetylation. Spurious initiation events in *Pwwp2a/b* DKO mESCs do not overlap in genomic location or chromatin features with spurious sites that arise in *Dnmt3b* KO mESCs, previously reported to function in the suppression of intragenic transcriptional initiation, suggesting these pathways function cooperatively in maintaining the fidelity of transcription initiation in metazoans.

INTRODUCTION

Trimethylation of lysine 36 on histone H3 (H3K36me₃) is a highly conserved posttranslational histone modification in eukaryotic organisms and is enriched over the bodies of actively transcribed genes (Bannister et al., 2005). H3K36me₃ is recognized by the PWWP domain, a motif that is present in many chromatin-modifying complexes (Dhayalan et al., 2010; Qin and Min, 2014; Tian et al., 2019; van Nuland et al., 2013; Wen et al., 2014; Xu et al., 2008). These readers are involved in a variety of biological processes, including intragenic DNA methylation, transcription elongation, DNA repair, alternative splicing, and repression of spurious transcription initiation (as reviewed in Huang and Zhu, 2018; Wagner and Carpenter, 2012). Furthermore, deregulation of H3K36 methylation and mutations of associated factors are linked to developmental disorders and cancers (as reviewed in Li et al., 2019; Zaghi et al., 2019).

Studies in yeast and mammalian cells have found that H3K36me₃ is important for maintaining transcriptional fidelity through repression of spurious transcription initiation from within coding regions. Mechanistically, H3K36me₃-mediated recruitment of effectors from two different epigenetic pathways, DNA methylation and histone deacetylation, has been linked to the suppression of spurious intragenic transcription initiation in actively transcribed genes. Human cancer cells treated with inhibitors against DNA methyltransferases and histone deacetylases show elevated levels of spurious transcription initiation (Brocks et al., 2017). Loss of the H3K36me₃-binding DNA methyltransferase DNMT3B in mouse embryonic stem cells (mESCs) leads to increased spurious transcription initiation (Neri et al., 2017). In *Saccharomyces cerevisiae*, H3K36me₃ recruits the Rpd3S HDAC complex to deacetylate histones across coding regions to repress spurious transcription initiation (Carrozza et al., 2005; Joshi and Struhl, 2005; Keogh et al., 2005), but mammalian equivalents of this pathway remained uncharacterized. In this study, we uncover a newly described PWWP2A/B-HDAC complex in mammalian cells that suppresses transcription initiation from the gene bodies of actively transcribed genes.

Recently, we and others have identified a variant nucleosome remodeling and deacetylase (NuRD) complex, in which the histone deacetylase subunits of NuRD (MTA1/2/3, HDAC1/2, RBBP4/7) form a highly stable and stoichiometric complex with the H3K36me₃-reader PWWP2A and its paralog PWWP2B (Figure 1A) (Link et al., 2018; Zhang et al., 2018). In this complex, PWWP2A/B, MTA1/2/3, and HDAC1/2 are present at a

¹Development Epigenetics, Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK

²Lymphocyte Development and Single Molecule Imaging, MRC London Institute of Medical Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK

³Lead Contact

*Correspondence: neil.brockdorff@bioch.ox.ac.uk (N.B.), tzhang17@ic.ac.uk (T.Z.)
<https://doi.org/10.1016/j.isci.2020.101741>



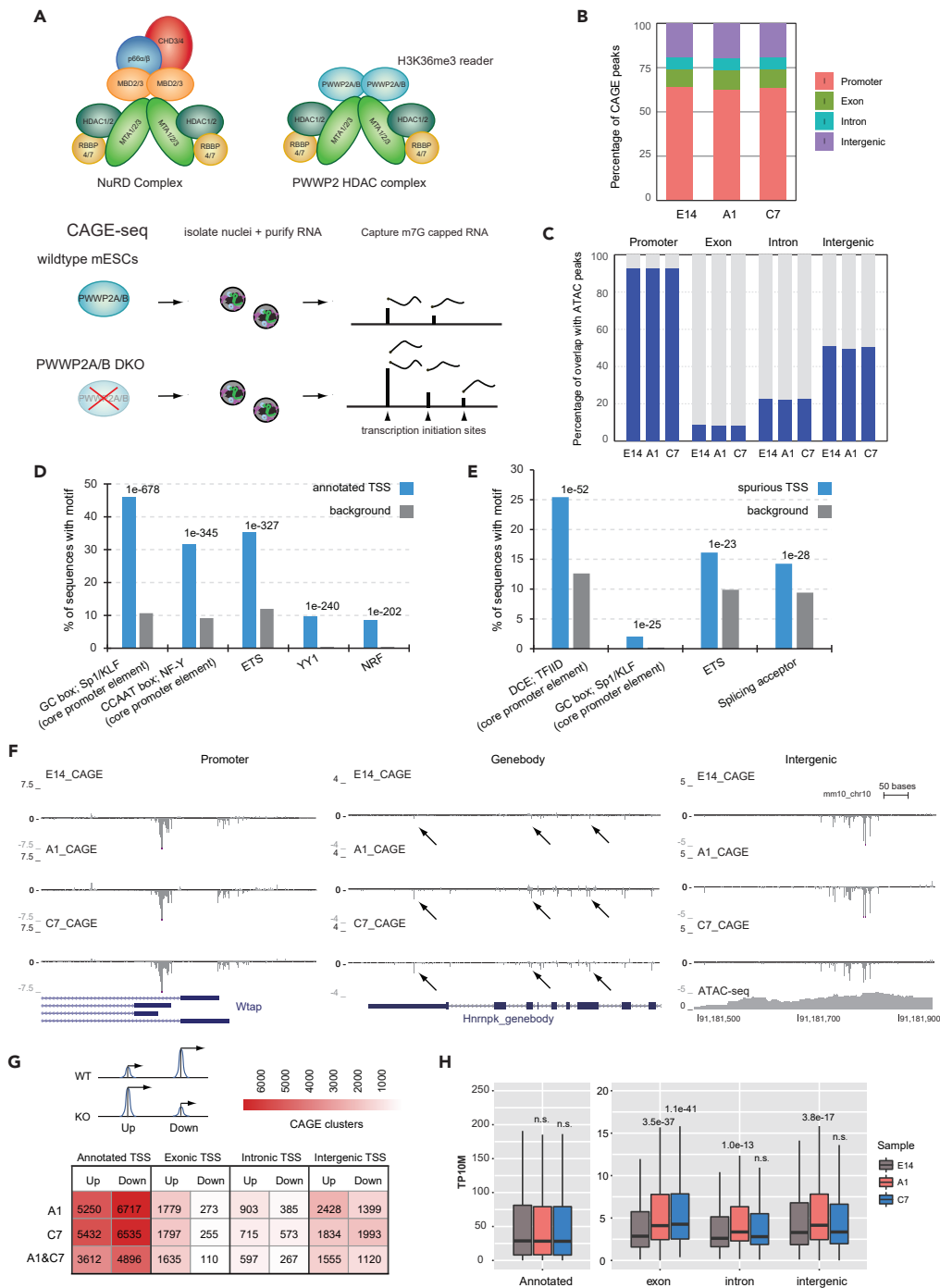


Figure 1. Loss of PWWP2A/B Leads to Increased Levels of intragenic Spurious Transcription Initiation in mESCs

(A) Comparison of the NuRD and PWWP2A/B HDAC complex, which contains the H3K36me3-binding PWWP2A/B protein bound to the MTA1/2/3:HDAC1/2:RBBP4/7 HDAC subcomplex (top). Schematic of nuclear RNA CAGE-seq from wild-type E14 and *Pwwp2a/b* DKO mESCs (bottom).

(B) The percentage of consensus TSS detected by CAGE-seq in wild-type E14 and *Pwwp2a/b* DKO clones A1 and C7 classified into four groups: annotated promoter, exonic, intronic, or intergenic TSS. See also [Table S1](#).

(C) The percentage of CAGE-seq TSSs in each group (annotated promoter, exon, intronic, intergenic) that overlap with ATAC-seq peaks for wild-type E14 and *Pwwp2a/b* DKO lines A1 and C7. See also [Figure S1](#).

(D) Annotated promoter TSSs (blue bars) detected by CAGE in mESCs are enriched in core promoter elements (GC box, CCAAT box) and in transcription factor motifs commonly found at promoters (ETS, YY1, NRF) compared with randomly

Figure 1. Continued

generated background sequences (gray bars). The top five most abundant and significant motifs found using Homer are shown with the significance value.

(E) Spurious TSSs (blue bars) from wild-type E14 mESCs are slightly enriched in core promoter elements (DCE, GC box) and transcription factor (ETS) sequences compared with randomly generated background sequences (gray bars). The top four most significant motifs found using Homer are shown with the significance value.

(F) UCSC genome browser view of the CAGE signal across several sites. Left: CAGE captures TSSs of different isoforms of the *Wtap* gene. Middle: transcription initiation peaks detected in the gene body of the *HnrnpK* gene where arrows indicate increased spurious transcription initiation events in *Pwwp2a/b* DKO lines. Right: intergenic transcription initiation event that overlaps with a region of high chromatin accessibility, possibly an enhancer. See also [Figure S2](#).

(G) CAGE-seq expression level heatmap for annotated promoter, exonic, intronic, and intergenic TSSs divided into two groups based on behavior: Up (increased CAGE expression in *Pwwp2a/b* DKO compared with wild-type) and Down (decreased CAGE expression in *Pwwp2a/b* DKO compared with wild-type). More spurious exonic and intronic TSSs go up than Down in *Pwwp2a/b* DKO cells. The bottom row shows the common events shared between the two clones.

(H) CAGE expression for annotated promoter, exonic, intronic, and intergenic TSSs in wild-type E14 and *Pwwp2a/b* DKO lines calculated using the CAGEr package in tags per 10 million (TP10M). The p values between wild-type E14 and DKO lines were calculated using the two-sided Mann-Whitney test. The exonic TSSs shows significantly more transcription initiation in the *Pwwp2a/b* DKO lines.

stoichiometry of 2:2:2 with PWWP2A/B being mutually exclusive with the canonical NuRD subunits MBD2/3 and CHD3/4 ([Figure 1A](#)). Although the NuRD complex has been found to play a role in cellular differentiation and cell type commitment through regulation of pluripotency and cellular reprogramming ([Luo et al., 2013](#); [Mor et al., 2018](#); [Reynolds et al., 2012](#)), the role of the PWWP2-HDAC complex is still not well understood. Unlike the NuRD complex subunits MBD3 and CHD4, which are enriched at active enhancers and promoters ([Bornelov et al., 2018](#); [Shimbo et al., 2013](#)), the PWWP2A variant NuRD complex is targeted to gene bodies. We previously showed that the PWWP2A localizes to gene bodies through binding to H3K36me3 via its C-terminal PWWP domain, where it regulates histone acetylation levels at highly transcribed genes ([Zhang et al., 2018](#)). Given that this novel HDAC complex was targeted to H3K36me3, we wondered whether the PWWP2-HDAC complex could fulfill the same function as the yeast Rpd3S HDAC complex in protecting against spurious transcription initiation. Cap analysis of gene expression (CAGE-seq) ([Takahashi et al., 2012](#)), a method widely applied to capture transcription initiation events ([Andersson et al., 2014](#); [Brocks et al., 2017](#); [Fort et al., 2014](#)), allowed us to precisely map and quantify all sites of transcription initiation at canonical promoters and spurious sites genome-wide in wild-type and *Pwwp2a/b* double knockout (DKO) mouse embryonic stem cells (mESCs). We used mESCs deleted for both PWWP2A and PWWP2B as our previous biochemical characterized showed these two paralogs can coexist in the same complex. Absence of PWWP2A/B in mESCs leads to increased levels of spurious transcription initiation particularly over the gene bodies of highly expressed genes. The profile of spurious initiation events upon PWWP2A/B depletion in mESCs contrasts with that in *Dnmt3b* knockout (KO) cells, suggesting that, although both gene body DNA methylation and histone deacetylation function downstream of H3K36me3 signaling, they have distinct roles in the suppression of spurious transcription initiation.

RESULTS**Loss of PWWP2A/B in mESCs Results in Increase Spurious Transcription Initiation within Intragenic Sites**

To investigate the role of the PWWP2A/B HDAC complex in spurious transcription initiation, we performed CAGE-seq in wild-type mESCs and two previously characterized *Pwwp2a/b* DKO mESC lines (clones A1 and C7) ([Zhang et al., 2018](#)). CAGE-seq using the nuclear RNA fraction as input was performed in triplicate for wild-type and *Pwwp2a/b* DKO C7 and in duplicate for DKO A1 ([Figure 1A](#)), and the biological replicates were merged owing to high reproducibility ([Figure S1A](#)). On average, each CAGE-seq library generated more than 20 million mapped reads ([Table S1](#)). The high sequencing depth allowed us to capture thousands of transcription initiation events (19134 CAGE peaks/clusters) in mESCs corresponding to annotated transcription initiation sites, spurious initiation sites, and transcription from enhancers. CAGE peaks were categorized as annotated promoter, intragenic (exonic and intronic), or intergenic TSSs (see [Transparent Methods](#)). Although the majority (~73%) of the reads and the consensus CAGE peaks (~63%) corresponded to GENCODE annotated promoter-TSSs ([Table S1](#) and [Figure 1B](#)), we detected thousands of initiation events (37% of CAGE peaks) that mapped to exonic, intronic, and intergenic regions ([Figure 1B](#)) in the wild-type and DKO mESCs.

To distinguish between spurious transcription initiation events and enhancers (eRNAs), all CAGE peaks were compared with ATAC-seq and H3K27ac ChIP-seq peaks, which mark active promoters and enhancers. As expected, the majority of annotated promoter CAGE peaks overlapped with regions of high chromatin accessibility as assayed by ATAC-seq and H3K27ac (Figures 1C and S1B). Approximately half of intergenic CAGE peaks overlapped with ATAC-seq peaks suggesting that these transcripts may correspond to eRNA production at enhancers (Figure 1C) (Andersson et al., 2014). The majority of exonic and intronic CAGE peaks did not overlap with enhancers (ATAC-seq and H3K27ac peaks) (Figures 1C and S1B) suggesting that the majority of intragenic TSSs are spurious transcription initiation events and not eRNAs.

We next examined the underlying sequence features of annotated promoter-TSSs and non-annotated spurious TSSs detected in wild-type E14 cells. As expected, canonical annotated promoter-TSSs are enriched in common core promoter elements (GC box, CCAAT box) and TF motifs commonly found at promoters (ETS, YY1, NRF) (Figure 1D). Spurious TSSs are less enriched for fewer common core promoter elements but still contain promoter-related motifs (DCE, GC box, ETS) compared with background (Figure 1E). CAGE peaks corresponding to spurious TSSs also tend to be much narrower than peaks at annotated promoter-TSSs (Figure S1C). These results suggest that by CAGE the majority of detectable peaks come from canonical eukaryotic promoters, but thousands of spurious transcription initiation events still arise from weak cryptic promoters even in wild-type mESCs.

To determine the effect of PWWP2A/B deletion in mESCs, we intersected all CAGE peaks from wild-type E14 with those from the DKO clones. The vast majority of CAGE peaks in wild-type and *Pwwp2a/b* DKO mESCs show extensive overlap (Figures 1F, S1D, and S1E), suggesting that PWWP2A/B loss does not have a significant impact on the number of transcription initiation sites in the genome. We next performed differential expression analysis to determine whether PWWP2A/B deletion would upregulate transcription initiation from existing spurious sites. TSS peaks that increase in expression in *Pwwp2a/b* DKO cells compared with wild-type are identified as “Up,” and peaks that decrease are “Down,” and we observed many intragenic TSSs being upregulated or induced in mESCs lacking PWWP2A/B. At intragenic TSSs, the number of Up sites is 14.8 and 2.2 times more than the number of Down sites for exonic and intronic TSSs, respectively (Figures 1G and Table S2). In particular, exonic TSSs show increased levels of spurious transcription initiation upon PWWP2A/B loss compared with wild-type mESCs (Figures 1F–1H and S2). Annotated TSS and intergenic TSS peaks did not show a skewed sensitivity to PWWP2A/B loss, with similar numbers of peaks increasing or decreasing in the level of transcription initiation (Figure 1G).

Spurious Sites Sensitive to PWWP2A/B Depletion Are Enriched at Intragenic Regions of Highly Expressed Genes

To further study the role of PWWP2A/B in regulating intragenic spurious initiation, we defined the subset of CAGE-seq peaks most sensitive to PWWP2A/B depletion. We removed intragenic TSS events that intersected with H3K27ac and ATAC-seq peaks, which may represent eRNAs, and classified the remaining 2,146 intragenic Up TSSs in both DKO clones as PWWP2-sensitive spurious TSSs (Figure 2A and Table S2). Owing to good correlation between the two clones A1 and C7 (Figures 2B and S3A), all downstream analysis was averaged for the two *Pwwp2a/b* DKO clones.

As discussed previously, motif analysis of the CAGE peaks for annotated and spurious TSSs in mESCs revealed that spurious TSSs are far less enriched for core promoter sequence motifs compared with annotated promoter-TSSs (Figures 1D and 1E). Furthermore, the majority (approximately 75%) of annotated promoter-TSSs overlapped with CpG islands and have high GC content (Figures 2C and 2D), whereas PWWP2-sensitive intragenic spurious transcription initiation peaks rarely overlap with CpG islands and are not GC-rich (Figures 2C and 2D). Lack of strong canonical promoter-like characteristics at spurious TSSs may explain the weaker transcription initiation capacity from these spurious intragenic cryptic promoter sites. Motif analysis comparing PWWP2-sensitive TSSs against annotated TSSs revealed that transcriptional repressors Tgif1/2 are significantly enriched at the flanking regions of PWWP2-sensitive spurious TSS in mESCs (Figure 2E) (Lee et al., 2015) and may suggest these cryptic promoters could be actively bound and repressed.

We next sought to characterize what type of genes were susceptible to PWWP2-sensitive spurious initiation. Gene ontology enrichment analysis revealed that genes harboring PWWP2-sensitive spurious TSSs are from a variety of biological pathways (Figure S4). Several RNA processing and surveillance pathways

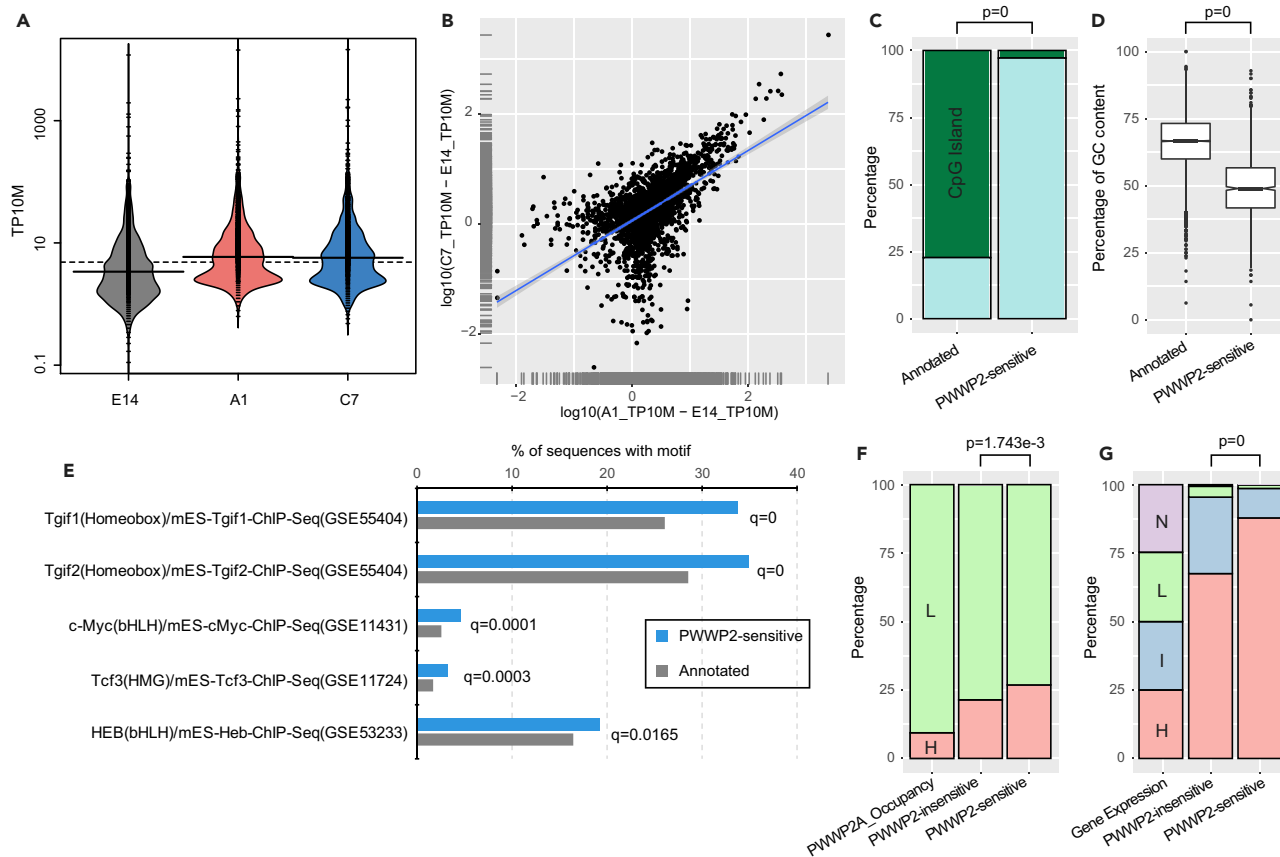


Figure 2. Intragenic Spurious Transcription Initiation Sites Arise Predominantly from Highly Expressed Genes

(A) Transcription initiation expression value (TP10M) of all intragenic PWWP2-sensitive spurious initiation sites (Up events $n = 2,146$) in E14 and *Pwwp2a/b* DKO lines. See also [Table S2](#).

(B) Scatterplot shows the correlation of PWWP2-sensitive spurious TSSs detected in *Pwwp2a/b* DKO lines A1 and C7. See also [Figure S3](#).

(C) Percentage of annotated TSSs and PWWP2-sensitive spurious TSSs ($n = 2,146$) that intersect (dark green) or do not intersect (cyan) with CpG islands. The p value was calculated by a chi-square test using the frequency.

(D) Boxplot of the %GC content for annotated TSSs ($n = 11,043$) and PWWP2-sensitive spurious TSSs (longer than 10 nt, $n = 1,224$). The p value was calculated using the two-sided Mann-Whitney test.

(E) The top most enriched motifs in PWWP2-sensitive spurious TSS (blue bars) compared with annotated TSSs (gray bars), as detected by the Homer motif search analysis.

(F) Percentage of spurious PWWP2-sensitive and PWWP2-insensitive TSSs occurring in genes with high (approximately top 10%) or low PWWP2A binding as defined by ChIP-seq ([Zhang et al., 2018](#)).

(G) Percentage of spurious PWWP2-sensitive and PWWP2-insensitive TSSs arising from genes with High, Intermediate, Low, or No expression (obtained by sorting all genes by expression level in wild-type E14 mESCs and dividing into four equal groups). A majority of spurious events (both PWWP2-sensitive and insensitive) occur from highly expressed gene group, which a greater fraction of PWWP2-sensitive spurious TSS residing in the most highly expressed group. The p value was calculated by chi-square test using the frequency. See also [Figure S4](#).

are also enriched, including RNA splicing, translation, and RNA nonsense-mediated decay ([Figure S4](#)). Previously we identified all genes bound by PWWP2A by ChIP-seq and classified them as being highly (9.4%) or lowly (90.6%) bound ([Zhang et al., 2018](#)). High-occupancy genes are more enriched for H3K36me3 than low-occupancy genes ([Figure S3B](#)). Here we see that all genes harboring intragenic spurious initiation sites are PWWP2A bound to some degree in mESCs, with a slightly greater proportion of PWWP2-sensitive spurious initiation events arising from high PWWP2A-occupied genes ([Figure 2F](#)). Strikingly, the greatest determinant correlated with the incidence of spurious transcription initiation is the level of gene expression, with the majority of spurious sites occurring within highly expressed genes ([Figure 2G](#)). Specifically, 88% of PWWP2A-sensitive spurious TSSs occur in the top 25% of genes ranked by expression ($n = 1,108$, see [Transparent Methods](#)) ([Figure 2G](#)). Together, these results indicate that PWWP2-sensitive spurious TSSs lack the underlying sequence features of canonical annotated promoters such as enrichment of many core

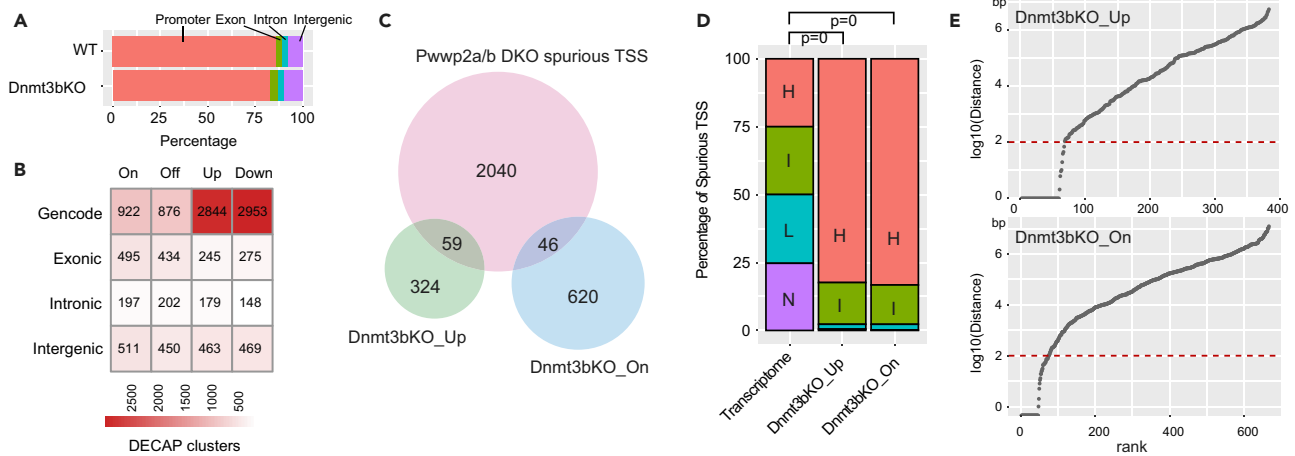


Figure 3. PWWP2A/B and DNMT3B Loss Have Different Effects on Spurious Transcription Initiation

(A) The percentage of consensus TSSs detected by DECAP-seq in wild-type and *Dnmt3b* KO mESCs from GSE72854 categorized as annotated promoter, exonic, intronic, intergenic TSSs.

(B) TSS peaks for wild-type and *Dnmt3b* KO cells categorized as Gencode annotated promoters, exonic, intronic, and intergenic TSSs and divided into four groups based on behavior: On (only present in *Dnmt3b* KO), Off (only present in wild-type), Up (upregulated in KO), and Down (downregulated in KO). See also Tables S1 and S3 and Figure S5.

(C) Overlap between PWWP2-sensitive spurious TSSs ($n = 2,146$) and DNMT3B-sensitive spurious TSSs (On $n = 666$ and Up = 383).

(D) Percentage of DNMT3B-sensitive spurious TSSs arising from genes with High, Intermediate, Low, or No expression (obtained by sorting all genes by expression level in wild-type E14 mESCs and dividing into four equal groups). The p values were calculated by chi-square test using the frequency.

(E) Distribution of the distance in base pairs of the closest PWWP2-sensitive spurious TSS relative to every DNMT3B-sensitive TSSs (Up and On). Dashed line indicates presence of a PWWP2-sensitive TSS within 100 bp.

promoter motifs, high GC and CpG island content, but contain weak promoter elements that are susceptible to spurious initiation events especially within highly expressed genes.

Loss of PWWP2A/B and DNMT3B Have Distinct Effects on Spurious Transcription Initiation

The DNA methyltransferase DNMT3B, like PWWP2A/B, is also selectively recruited to gene bodies through binding to H3K36me3 via its PWWP domain (Rondelet et al., 2016). A prior study found that loss of DNMT3B in mESCs leads to intragenic DNA hypomethylation and increased spurious transcription initiation events (Neri et al., 2017). Given that both of these pathways are downstream of H3K36me3 and contribute to regulating the epigenetic landscape at gene bodies, we sought to determine their relationship with one another. We applied our pipeline to analyze previously published data from *Dnmt3b* KO mESCs, which was performed using DECAP-seq, an alternative to CAGE-seq for profiling sites of transcription initiation. Consistent with the published study, we found that most DECAP TSSs correspond to Gencode annotated promoters (Figure 3A). We classified all DNMT3B TSSs following the terminology used in the previous DNMT3B study—TSSs absent in wild-type but present in *Dnmt3b* KO are categorized as “On,” TSSs present in wild-type cells but lost in *Dnmt3b* KO are “Off,” upregulated TSSs in *Dnmt3b* KO are “Up,” and downregulated TSSs in *Dnmt3b* KO are “Down” (Figures 3B and Table S3). Unlike in *Pwwp2a/b* DKO mESCs, where sites of spurious initiation almost perfectly overlap with wild-type but show different levels of transcription initiation, either increased (Up) or decreased (Down) (Figures 1G, S1D, and S1E), *Dnmt3b* KO mESCs showed many On and Off sites indicating that loss of DNMT3B-mediated DNA methylation led to a redistribution of spurious sites in mESCs (Figure 3B). Another distinction between the two pathways was that, *Pwwp2a/b* deletion led to preferential upregulation of intragenic spurious TSSs, whereas *Dnmt3b* deletion resulted in an equivalent number of Up and Down spurious sites.

We next identified DNMT3B-sensitive spurious TSSs as all intragenic On and Up TSSs after removal of sites that overlapped with ATAC-seq and H3K27ac peaks. The overlap between sites (genomic positions) of PWWP2-sensitive and DNMT3B-sensitive spurious initiation is very low (Figure 3C). However, like PWWP2-sensitive sites, DNMT3B-sensitive spurious initiation also occurs predominantly at highly expressed genes (Figure 3D) and displays similar promoter width as well as CpG Island and GC content as PWWP2-sensitive spurious TSSs (Figures 2C, 2D, and S5A–S5C). Despite both DNMT3B-sensitive and

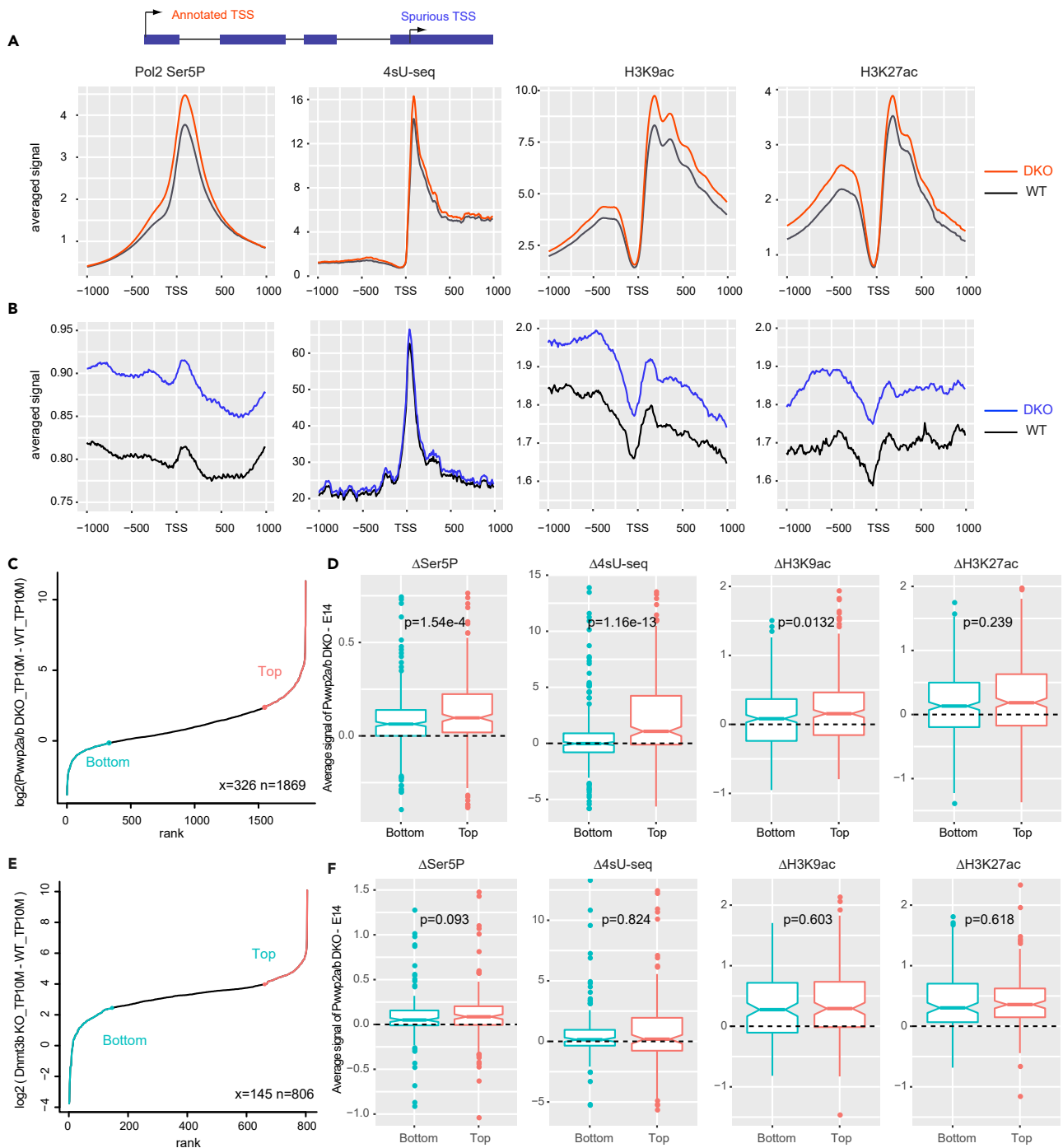


Figure 4. Elevated RNAPII and Histone Acetylation at Spurious Transcription Initiation Sites in *Pwpp2a/b* DKO mESCs

(A) Metagenes profiles of RNA Pol II Ser5P, 4sU-nascent transcription, H3K9ac, and H3K27ac at annotated TSSs ± 1 kb ($n = 8,164$) of wild-type and *Pwpp2a/b* DKO mESCs.

(B) Metagenes profiles RNA Pol II Ser5P, 4sU-nascent transcription, H3K9ac, and H3K27ac at the PWWP2-sensitive spurious TSSs ± 1 kb. Only TSSs that are 2 kb away from each other were considered ($n = 1,869$).

(C) PWWP2-sensitive spurious TSSs ranked by the level of upregulation in *Pwpp2a/b* DKO compared with wild-type. Only sites 2 kb away from each other were considered ($n = 1,869$). Bottom group: spurious sites that show small increase in CAGE expression upon PWWP2A/B loss (cyan). Top group: spurious sites that show large increase in CAGE expression upon PWWP2A/B loss (red). (Top and Bottom group: $x = 326$).

Figure 4. Continued

(D) Boxplots of the distribution of relative differences of the Top and Bottom PWWP2-sensitive spurious sites with respect to the level of Pol II Ser5P, nascent transcription, H3K9ac and H3K27ac. p Values were calculated and using the two-sided Mann-Whitney test between *Pwwp2a/b* DKO and wild-type E14 mESCs.

(E) DNMT3B-sensitive spurious TSSs ranked by the level of upregulation in *Dnmt3b* KO compared with wild-type. Only TSSs 2 kb away from annotated TSS and 100 bp away from PWWP2-sensitive TSSs were considered (n = 806). Red and cyan indicate top (most sensitive) and bottom group (least sensitive) sites, respectively (x = 145, the same proportion as in C).

(F) Boxplots of the distribution of relative differences between *Pwwp2a/b* DKO and E14 for Top and Bottom groups DNMT3B-spurious sites from (E) with respect to the level of Pol II Ser5P, nascent transcription, H3K9ac and H3K27ac. p Values were calculated and using the two-sided Mann-Whitney test *Pwwp2a/b* DKO and wild-type E14 mESCs. See also [Figure S6](#).

PWWP2-sensitive spurious residing within highly expressed genes, there is little overlap between sites ([Figure 3C](#)) with the vast majority of DNMT3B-sensitive spurious TSS more than 100 bp away from the closest neighboring PWWP2-sensitive spurious TSS ([Figure 3E](#)).

Transcription and Chromatin Signatures Underlying Spurious Transcription Initiation

We next explored the dynamics of Pol II binding, nascent transcription, and histone acetylation at PWWP2-sensitive and DNMT3B-sensitive spurious transcription initiation sites. Canonical transcription initiation sites are enriched for initiated RNA Polymerase II phosphorylated at serine 5 (RNA Pol II Ser5P) surrounding the TSS. Pol II Ser5P binding thus serves as an orthogonal approach to examine sites of spurious transcription initiation ([Phatnani and Greenleaf, 2006](#)). To reduce noise from neighboring peaks, only single peaks in a 2-kb window were considered for downstream analysis for both the annotated TSSs and spurious TSSs. By comparing Pol II Ser5P ChIP-seq in E14 and *Pwwp2a/b* DKO cells, we found that Pol II Ser5P signal slightly peaks just downstream of the spurious TSSs resembling annotated promoter-TSSs ([Figure 4A](#)), although at a much lower levels ([Figures 4A and 4B](#)). Pol II Ser5P levels are also increased in *Pwwp2a/b* DKO when compared with E14 cells ([Figure 4B](#)). These intragenic spurious initiation sites also showed a local increase of nascent RNA transcription compared with background as detected by 4sU nascent RNA-seq in wild-type E14 and *Pwwp2a/b* DKO lines ([Figure 4B](#)). In contrast to annotated TSSs, which display a strong asymmetric pattern around the peak indicative of a bias toward the sense direction and weak antisense transcription, the nascent RNA transcription around spurious transcription initiation sites is more balanced around the peak suggesting that the transcription driven by spurious TSSs is weakly expressed bidirectionally or may not produce long RNA products ([Figure 4B](#)).

Acetylation of H3K27 and H3K9 is enriched at active genes, peaking around the TSS. Previously we showed that loss of PWWP2A/B HDAC complex leads to increased histone acetylation at active gene promoters and gene bodies ([Zhang et al., 2018](#)). We assessed the pattern of H3K27ac and H3K9ac at PWWP2-sensitive spurious TSSs defined in this study and found the signature resembles that of annotated TSSs, with a trough at the spurious TSS and peaks at the +1 and -1 nucleosome positions ([Figures 4A and 4B](#)). Spurious TSSs show increased acetylation levels in *Pwwp2a/b* DKO cells compared with wild-type mESCs ([Figures 4B and S6B](#)). We calculated the difference in RNA Pol II Ser5P, 4SU-seq, H3K9ac, and H3K27ac levels between *Pwwp2a/b* DKO and wild-type E14 and found all of them to be enriched around the spurious TSSs in the DKO ([Figures S6A and S6B](#)).

We divided the PWWP2-sensitive spurious TSSs sites into two groups based on the degree of increase in spurious transcription initiation compared with wild-type, “Top” (large increase in DKO compared with E14) and “Bottom” (small increase in DKO compared with E14), using TP10M of each CAGE peak ([Figure 4C](#)). There are significantly higher levels of initiating Pol II Ser5P, nascent RNA transcription, and H3K9ac at Top sites compared with Bottom sites ([Figure 4D](#)). This suggests that increased histone acetylation upon PWWP2A/B loss correlates with increased initiation and transcription from spurious sites, and the degree of change correlates with sensitivity to PWWP2A/B loss. In parallel we performed the same analysis for DNMT3B-sensitive spurious TSSs and found that spurious sites more (Top) or less (Bottom) sensitive to DNMT3B loss did not show differences in the level of initiating Pol II, nascent RNA production, or histone acetylation ([Figures 4E, 4F, S6C, and S6D](#)). These features, which explain sensitivity to PWWP2A/B loss, do not explain differences between spurious sites arising from DNMT3B depletion.

In summary, our analysis of the chromatin environment at spurious TSSs show that these spurious initiation sites show increased recruitment of initiating Pol II Ser5P, elevated levels of nascent transcription, and histone acetylation relative to the surrounding background. Additionally, upon PWWP2A/B loss, increased

levels of spurious initiation are correlated with elevated histone acetylation and increased engaged RNA Pol II Ser5P and nascent transcription compared with wild-type. Although PWWP2A/B appears to regulate spurious transcription initiation at the level of histone acetylation and recruitment of initiating Pol II Ser5P, the DNMT3B pathway may regulate spurious initiation through altering binding of DNA-methylation sensitive factors (as discussed in [Neri et al., 2017](#)).

DISCUSSION

Transcription initiation is a highly coordinated process that establishes accurate patterns of gene expression and ensures the production of functional transcripts. Here we describe how cross talk between two epigenetic pathways H3K36me3 and histone acetylation suppresses spurious transcription initiation from the gene bodies of active genes. H3K36me3 is deposited co-transcriptionally with the elongating RNA Pol II and exclusively marks the gene bodies of actively transcribed genes ([Kizer et al., 2005](#)). Acetylation of H3K9 and H3K27 is enriched at active promoter regions and regulates RNA Pol II initiation and pause release ([Gates et al., 2017](#); [Stasevich et al., 2014](#)). Here we describe a pathway in metazoans, in which H3K36me3 recruitment of the PWWP2A/B HDAC complex represses spurious transcription initiation from the gene bodies of actively transcribed genes. This prevents RNA Pol II from initiating at cryptic promoters within coding regions and the production of aberrant transcripts, which could be harmful to cellular homeostasis.

PWWP2A/B binds H3K36me3 through a C-terminal PWWP domain, a module present in many other H3K36me3-binding proteins ([Qin and Min, 2014](#); [Vermeulen et al., 2010](#); [Vezzoli et al., 2010](#); [Wen et al., 2014](#)). Previously we and others showed that PWWP2A/B forms a stable complex HDAC1/2 and MTA1/2/3 ([Link et al., 2018](#); [Zhang et al., 2018](#)). In this study, we find that loss of PWWP2A/B in mESCs leads to increases in the level of spurious transcription initiation, especially over intragenic gene body regions.

Similar to studies in *S. cerevisiae*, we find that, although spurious transcripts are lowly expressed, they are highly prevalent in number within gene bodies in mESCs ([Lu and Lin, 2019](#); [Neil et al., 2009](#)). Spurious initiation sites contain some core promoter elements like the GC box, DCE, and ETS motifs but are far less enriched for these motifs than canonical eukaryotic promoters ([Haberle and Stark, 2018](#); [Hollenhorst et al., 2004](#); [Lee et al., 2005](#); [Roy and Singer, 2015](#)). Additionally, spurious initiation sites also lack the high GC and CpG island content associated with eukaryotic promoters ([Deaton and Bird, 2011](#); [Haberle and Stark, 2018](#); [Roy and Singer, 2015](#)). The vast majority of spurious transcription initiate from highly expressed genes, and these sites are marked by initiating Pol II Ser5P, which increases further upon PWWP2A/B loss. The increased level of spurious transcripts at highly expressed genes may reflect increased levels of hyperacetylation, increased chromatin accessibility and nucleosome turnover, and a high local concentration of coactivators and transcriptional machinery, which together lead to binding and initiation of RNA Pol II within intragenic sites. Our data support the model where promiscuous initiation by Pol II occurs more frequently at highly expressed genes and that spurious transcription initiation is an inevitable by-product of transcription ([Wade and Grainger, 2018](#)). Multiple mechanisms besides the H3K36me3-HDAC pathway have been found to play a role in countering spurious initiation and the production of spurious transcripts ([Gouot et al., 2018](#); [Neri et al., 2017](#); [Scandaglia et al., 2017](#)), with intragenic DNA methylation by DNMT3B as another pathway downstream of H3K36me3 that suppresses spurious transcription initiation.

We find that loss of the PWWP2A/B HDAC complex has a different profile than loss of DNMT3B, and PWWP2-sensitive and DNMT3B-sensitive spurious TSSs are largely non-overlapping. The only features common to both pathways was that intragenic spurious transcription initiate predominantly from the highly expressed genes. Our data suggest that PWWP2A/B suppresses transcription initiation through regulating histone acetylation and Pol II initiation, whereas DNA methylation may regulate the binding of methylation-sensitive TFs to suppress initiation at intragenic sites. Although PWWP2A/B loss leads to increased levels of initiation from pre-existing intragenic sites found in mESCs, loss of DNMT3B leads to redistribution of sites from which spurious transcripts initiate. It appears that in mammalian organisms these two epigenetic mechanisms cooperate downstream of H3K36me3 to control transcription fidelity by ensuring that initiation occurs predominantly at strong promoter regions and is actively suppressed from intragenic regions.

Limitations of the Study

There are two points that are outside the scope of this study and will require more experimental work to fully elucidate, these being whether PWWP2A and PWWP2B have redundant functions and the role of the PWWP2A/B-HDAC complex at intergenic regions.

1. Overlapping and paralog-specific differences of PWWP2A and PWWP2B

In this work, we used mESCs where both the *Pwwp2a* and *Pwwp2b* genes have been knocked out. Our justification for using the double knockout comes from biochemical characterization of the complex in our previous work (Zhang et al., 2018), which shows that two copies of PWWP2A bind to the MTA1-dimer (stoichiometry of 2:2), PWWP2A and PWWP2B can coexist in the same complex, and the PWWP domains have the same affinity for H3K36me3. However, despite high sequence conservation, there are differences in the domain composition of the two paralogs (see Figure S7) with PWWP2B lacking the N-terminal proline-rich region present in PWWP2A. Both paralogs are expressed in mESCs but could have differences in tissue-specific expression and functions. Further work is needed to address whether there are paralog-specific differences in this family of proteins.

2. Intergenic regions that show sensitivity to PWWP2A/B loss

Our characterization of spurious transcription initiation focused predominantly on intragenic spurious TSSs, which were sensitive to PWWP2A/B deletion. Although intragenic spurious TSSs were much more sensitive to loss of PWWP2A/B, some intergenic spurious TSSs were also induced. These intergenic sites are not enriched for H3K36me3, so we cannot subscribe these events to H3K36me3-mediated recruitment of PWWP2A/B. Aside from the H3K36me3-binding PWWP domain, there is a H2A.Z-binding domain, which could recruit PWWP2A/B to chromatin, but whether this could explain intergenic spurious TSSs events remains to be examined.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the corresponding author, Tianyi Zhang (tzhang17@ic.ac.uk).

Materials Availability

Materials and protocols used in this study are available from the authors upon request. This study did not generate new unique reagents.

Data and Code Availability

CAGE-seq and processed bigwig files generated in this study have been deposited in Gene Expression Omnibus (GEO) under GSE148382. Code used in this study has been deposited in github, <https://github.com/gui Fengwei/PWWP2-CAGE-seq>.

METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2020.101741>.

ACKNOWLEDGMENTS

Work in the Brockdorff lab is supported by the Wellcome Trust (grant no. 215513). T.Z. was supported by NSERC and the Clarendon Scholarship in Oxford and is now supported by the Sir Henry Wellcome Postdoctoral Fellowship (grant no. 215933). We thank Amanda Williams from Oxford Zoology for Next-seq sequencing and Joseph Bowness and Mafalda Almeida for discussion and critical reading of the manuscript.

AUTHOR CONTRIBUTIONS

T.Z. conceived and performed the experiments. G.W. performed all the bioinformatics analysis and generated all the figures. G.W. and T.Z. wrote and revised the manuscript with input from N.B. who supervised the whole study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 17, 2020

Revised: September 22, 2020

Accepted: October 23, 2020

Published: November 20, 2020

REFERENCES

- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461.
- Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.* 280, 17732–17736.
- Bornelov, S., Reynolds, N., Xenophontos, M., Gharbi, S., Johnstone, E., Floyd, R., Ralsler, M., Signolet, J., Loos, R., Dietmann, S., et al. (2018). The nucleosome remodeling and deacetylation complex modulates chromatin structure at sites of active transcription to fine-tune gene expression. *Mol. Cell* 71, 56–72 e54.
- Brocks, D., Schmidt, C.R., Daskalakis, M., Jang, H.S., Shah, N.M., Li, D., Li, J., Zhang, B., Hou, Y., Laudato, S., et al. (2017). DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* 49, 1052–1060.
- Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., et al. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* 123, 581–592.
- Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022.
- Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R.Z., Ragozin, S., and Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J. Biol. Chem.* 285, 26114–26120.
- Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., et al. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* 46, 558–566.
- Gates, L.A., Shi, J., Rohira, A.D., Feng, Q., Zhu, B., Bedford, M.T., Sagum, C.A., Jung, S.Y., Qin, J., Tsai, M.J., et al. (2017). Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *J. Biol. Chem.* 292, 14456–14472.
- Gouet, E., Bhat, W., Rufiange, A., Fournier, E., Paquet, E., and Nourani, A. (2018). Casein kinase 2 mediated phosphorylation of Sp76 modulates histone dynamics and regulates spurious transcription. *Nucleic Acids Res.* 46, 7612–7630.
- Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* 19, 621–637.
- Hollenhorst, P.C., Jones, D.A., and Graves, B.J. (2004). Expression profiles frame the promoter specificity dilemma of the ETS family of transcription factors. *Nucleic Acids Res.* 32, 5693–5702.
- Huang, C., and Zhu, B. (2018). Roles of H3K36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity. *Biophys. Rep.* 4, 170–177.
- Joshi, A.A., and Struhl, K. (2005). Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. *Mol. Cell* 20, 971–978.
- Keogh, M.C., Kurdastani, S.K., Morris, S.A., Ahn, S.H., Podolny, V., Collins, S.R., Schuldiner, M., Chin, K., Punna, T., Thompson, N.J., et al. (2005). Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* 123, 593–605.
- Kizer, K.O., Phatnani, H.P., Shibata, Y., Hall, H., Greenleaf, A.L., and Strahl, B.D. (2005). A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Mol. Cell Biol.* 25, 3305–3316.
- Lee, B.K., Shen, W., Lee, J., Rhee, C., Chung, H., Kim, K.Y., Park, I.H., and Kim, J. (2015). Tgif1 counterbalances the activity of core pluripotency factors in mouse embryonic stem cells. *Cell Rep.* 13, 52–60.
- Lee, D.H., Gershenzon, N., Gupta, M., Ioshikhes, I.P., Reinberg, D., and Lewis, B.A. (2005). Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol. Cell Biol.* 25, 9674–9686.
- Li, J., Ahn, J.H., and Wang, G.G. (2019). Understanding histone H3 lysine 36 methylation and its deregulation in disease. *Cell. Mol. Life Sci.* 76, 2899–2916.
- Link, S., Spitzer, R.M.M., Sana, M., Torrado, M., Volker-Albert, M.C., Keilhauer, E.C., Burgold, T., Punzeler, S., Low, J.K.K., Lindstrom, I., et al. (2018). PWWP2A binds distinct chromatin moieties and interacts with an MTA1-specific core NuRD complex. *Nat. Commun.* 9, 4300.
- Lu, Z., and Lin, Z. (2019). Pervasive and dynamic transcription initiation in *Saccharomyces cerevisiae*. *Genome Res.* 29, 1198–1210.
- Luo, M., Ling, T., Xie, W., Sun, H., Zhou, Y., Zhu, Q., Shen, M., Zong, L., Lyu, G., Zhao, Y., et al. (2013). NuRD blocks reprogramming of mouse somatic cells into pluripotent stem cells. *Stem Cells* 31, 1278–1286.
- Mor, N., Rais, Y., Sheban, D., Peles, S., Aguilera-Castrejon, A., Zviran, A., Elinger, D., Viukov, S., Geula, S., Krupalnik, V., et al. (2018). Neutralizing *gatad2a-chd4-mbd3*/NuRD complex facilitates deterministic induction of naive pluripotency. *Cell Stem Cell* 23, 412–425.e10.
- Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 457, 1038–1042.
- Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature* 543, 72–77.
- Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.* 20, 2922–2936.
- Qin, S., and Min, J. (2014). Structure and function of the nucleosome-binding PWWP domain. *Trends Biochem. Sci.* 39, 536–547.
- Reynolds, N., Latos, P., Hynes-Allen, A., Loos, R., Leaford, D., O'Shaughnessy, A., Mosaku, O., Signolet, J., Brennecke, P., Kalkan, T., et al. (2012). NuRD suppresses pluripotency gene expression to promote transcriptional heterogeneity and lineage commitment. *Cell Stem Cell* 10, 583–594.
- Rondelet, G., Dal Maso, T., Willems, L., and Wouters, J. (2016). Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *J. Struct. Biol.* 194, 357–367.
- Roy, A.L., and Singer, D.S. (2015). Core promoters in transcription: old problem, new insights. *Trends Biochem. Sci.* 40, 165–171.
- Scandaglia, M., Lopez-Atalaya, J.P., Medrano-Fernandez, A., Lopez-Cascales, M.T., Del Blanco, B., Lipinski, M., Benito, E., Olivares, R., Iwase, S., Shi, Y., et al. (2017). Loss of Kdm5c causes spurious transcription and prevents the fine-tuning of activity-regulated enhancers in neurons. *Cell Rep.* 21, 47–59.
- Shimbo, T., Du, Y., Grimm, S.A., Dhasarathy, A., Mav, D., Shah, R.R., Shi, H., and Wade, P.A. (2013). MBD3 localizes at promoters, gene bodies and enhancers of active genes. *PLoS Genet.* 9, e1004028.

Stasevich, T.J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., Tokunaga, M., Nagase, T., Nozaki, N., McNally, J.G., et al. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* 516, 272–275.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* 7, 542–561.

Tian, W., Yan, P., Xu, N., Chakravorty, A., Liefke, R., Xi, Q., and Wang, Z. (2019). The HRP3 PWWP domain recognizes the minor groove of double-stranded DNA and recruits HRP3 to chromatin. *Nucleic Acids Res.* 47, 5436–5448.

van Nuland, R., van Schaik, F.M., Simonis, M., van Heesch, S., Cuppen, E., Boelens, R., Timmers, H.M., and van Ingen, H. (2013). Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the

PSIP1-PWWP domain. *Epigenetics Chromatin* 6, 12.

Vermeulen, M., Eberl, H.C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K.K., Olsen, J.V., Hyman, A.A., Stunnenberg, H.G., et al. (2010). Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* 142, 967–980.

Vezzoli, A., Bonadies, N., Allen, M.D., Freund, S.M., Santiveri, C.M., Kvinlaug, B.T., Huntly, B.J., Gottgens, B., and Bycroft, M. (2010). Molecular basis of histone H3K36me3 recognition by the PWWP domain of Brpf1. *Nat. Struct. Mol. Biol.* 17, 617–619.

Wade, J.T., and Grainger, D.C. (2018). Spurious transcription and its impact on cell function. *Transcription* 9, 182–189.

Wagner, E.J., and Carpenter, P.B. (2012). Understanding the language of Lys36 methylation at histone H3. *Nat. Rev. Mol. Cell Biol.* 13, 115–126.

Wen, H., Li, Y., Xi, Y., Jiang, S., Stratton, S., Peng, D., Tanaka, K., Ren, Y., Xia, Z., Wu, J., et al. (2014). ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression. *Nature* 508, 263–268.

Xu, C., Cui, G., Botuyan, M.V., and Mer, G. (2008). Structural basis for the recognition of methylated histone H3K36 by the Eaf3 subunit of histone deacetylase complex Rpd3S. *Structure* 16, 1740–1750.

Zaghi, M., Broccoli, V., and Sessa, A. (2019). H3K36 methylation in neural development and associated diseases. *Front. Genet.* 10, 1291.

Zhang, T., Wei, G., Millard, C.J., Fischer, R., Konietzny, R., Kessler, B.M., Schwabe, J.W.R., and Brockdorff, N. (2018). A variant NuRD complex containing PWWP2A/B excludes MBD2/3 to regulate transcription at active genes. *Nat. Commun.* 9, 3798.

iScience, Volume 23

Supplemental Information

**The PWWP2A Histone Deacetylase
Complex Represses Intragenic Spurious
Transcription Initiation in mESCs**

Guifeng Wei, Neil Brockdorff, and Tianyi Zhang

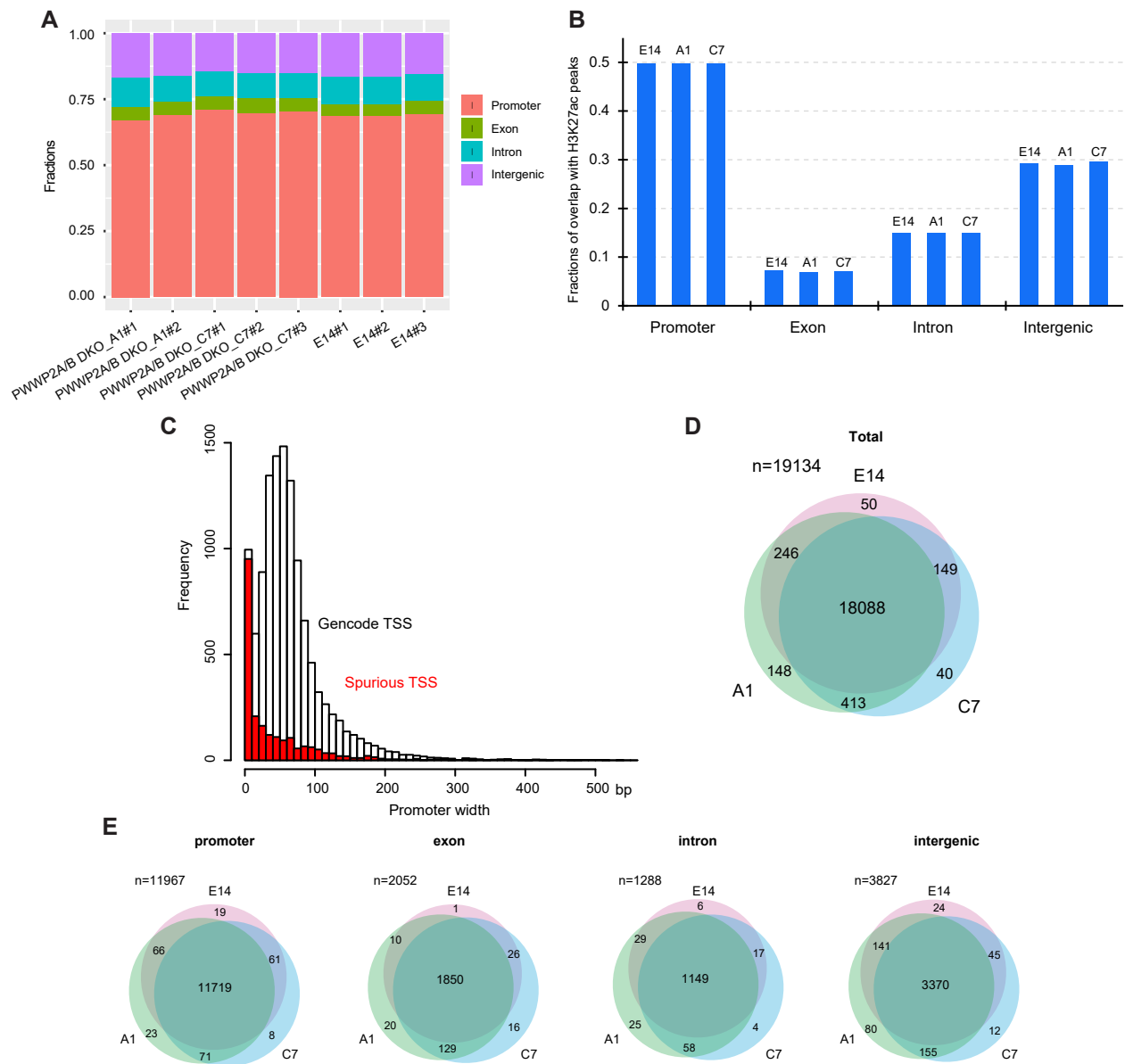


Figure S1. Annotation of CAGE-seq reads to GENCODE genomic features, related to Figure 1.

(A) Fraction of CAGE-seq reads in promoter, exon, intron, and intergenic regions in E14 wildtype and in *Pwwp2a/b* DKO mESC lines A1 and C7. Also see Table S1.

(B) The fraction of consensus TSS from all the categories that overlap with H3K27ac-seq peaks in mESC for E14 and *Pwwp2a/b* DKO lines.

(C) Histogram shows the width distribution for Gencode annotated TSSs (n=11967) and Spurious TSSs (n=2146) respectively. See also Table S2.

(D-E) Venn diagrams show the overlap of CAGE peaks (TP10M>1) between wildtype E14 and *Pwwp2a/b* DKO lines. All CAGE peaks are shown in (D) and peaks present in promoter, exonic, intronic, and intergenic sequences are shown separately in (E).

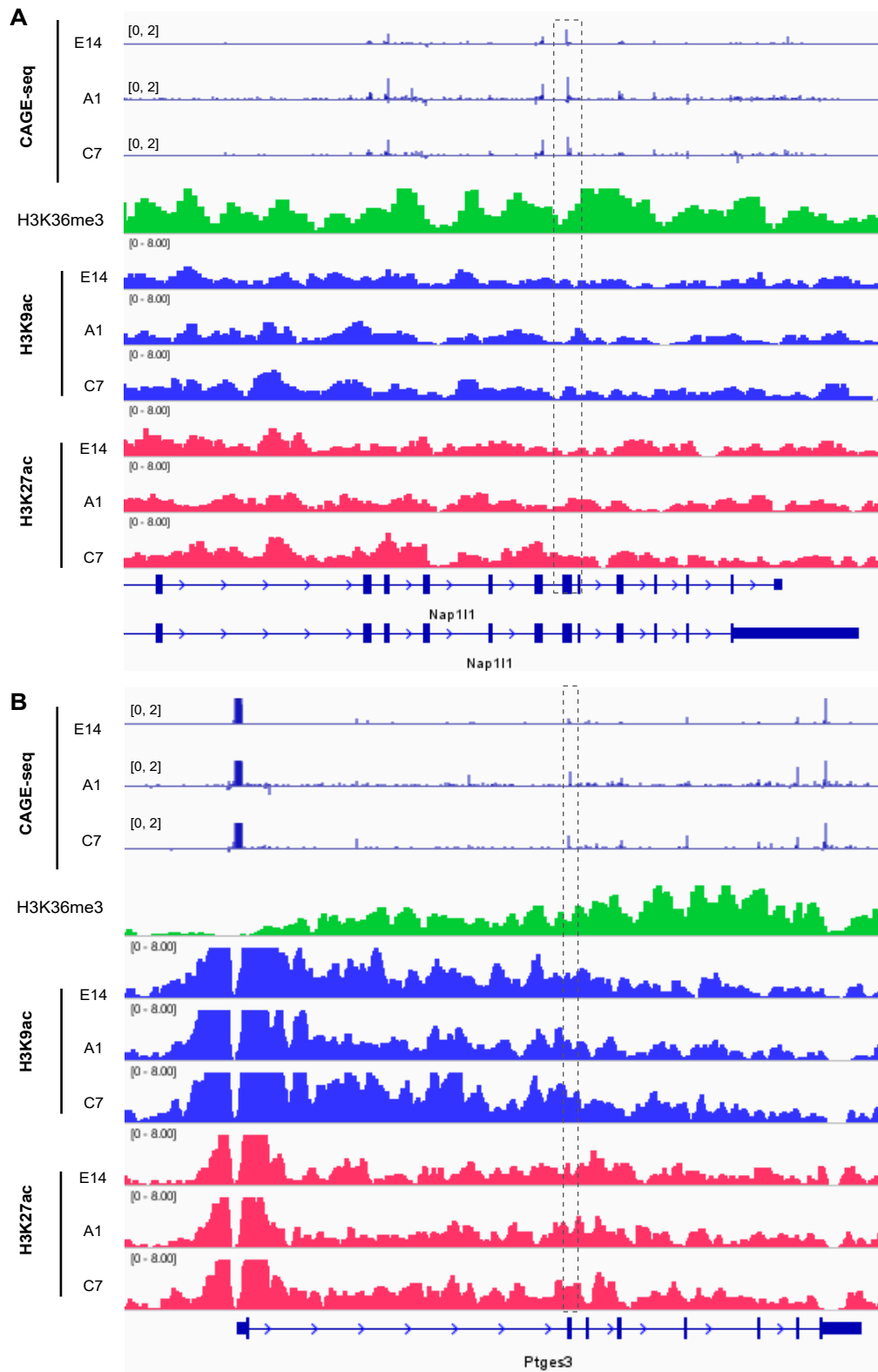


Figure S2. Chromatin landscape at representative PWWP2-sensitive spurious TSSs, related to Figure 1.

(A-B) IGV browser view of the profiles for CAGE, H3K36me3, H3K9ac, and H3K27ac data in E14 (WT) cells and *Pwwp2/b* DKO cells. Dashed box indicates the representative intragenic spurious TSS at (A) *Nap111* locus and (B) *Ptges3* locus. Same type of data is adjusted to same scale.

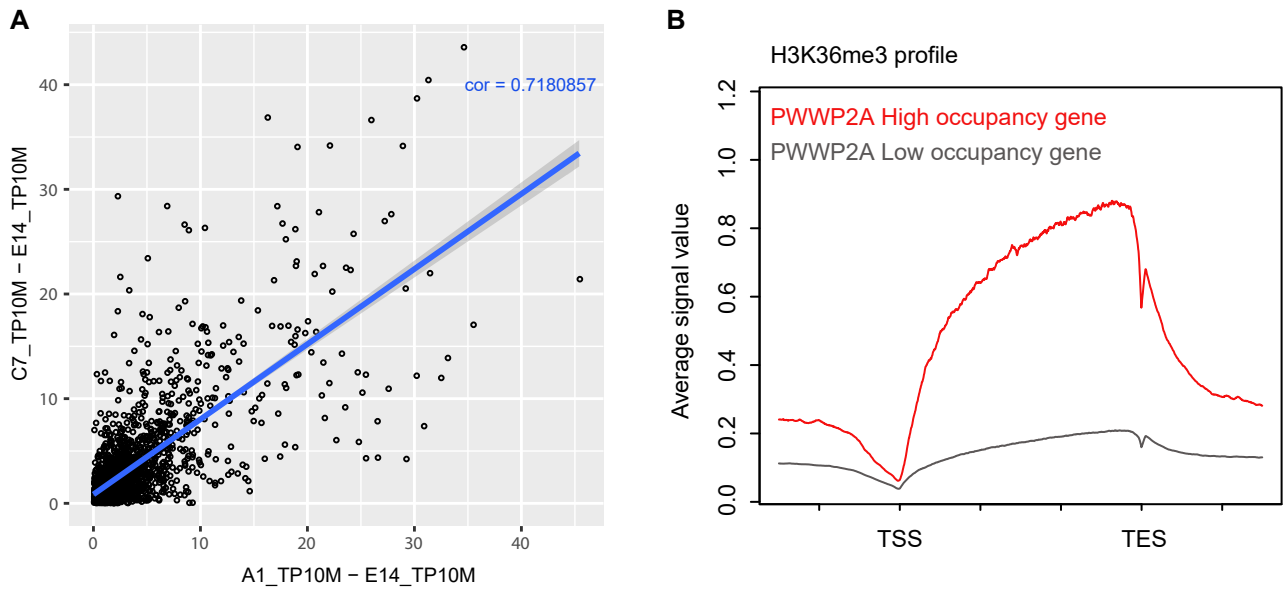
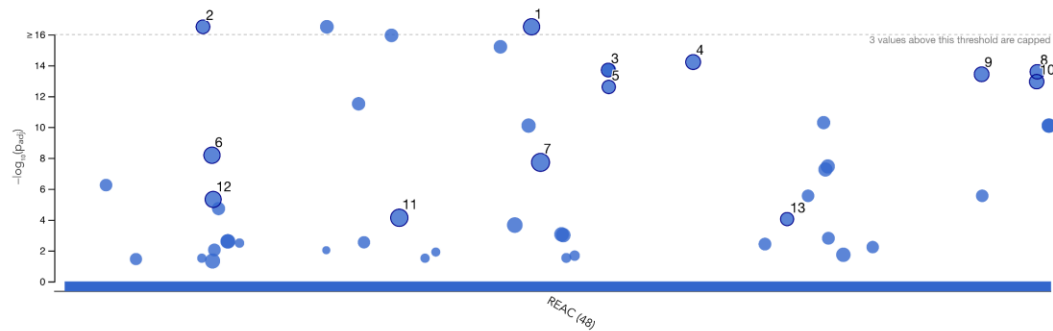


Figure S3. Correlation of CAGE-seq in between *Pwwp2a/b* lines, related to Figure 2.

(A) Correlation of the PWWP2-sensitive intragenic spurious TSS between *Pwwp2a/b* DKO clones A1 and C7 (excluding the extreme outliers) with the correlation score at top right. Pearson correlation was calculated based on non-log-transformed values. The dashed blue line indicates the linear regression. (B) Metaprofile of H3K36me3 level at high (red) and low (grey) PWWP2A occupancy genes as defined by ChIP-seq (Zhang et al. 2018).



ID	Source	Term ID	Term Name	P _{adj} (query_1)
1	REAC	REAC:R-MMU-8953854	Metabolism of RNA	2.264 × 10 ⁻³⁴
2	REAC	REAC:R-MMU-72737	Cap-dependent Translation Initiation	2.563 × 10 ⁻¹⁷
3	REAC	REAC:R-MMU-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	2.003 × 10 ⁻¹⁴
4	REAC	REAC:R-MMU-72203	Processing of Capped Intron-Containing Pre-mRNA	6.092 × 10 ⁻¹⁵
5	REAC	REAC:R-MMU-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	2.490 × 10 ⁻¹³
6	REAC	REAC:R-MMU-1640170	Cell Cycle	6.640 × 10 ⁻⁹
7	REAC	REAC:R-MMU-392499	Metabolism of proteins	1.924 × 10 ⁻⁸
8	REAC	REAC:R-MMU-72163	mRNA Splicing - Major Pathway	2.592 × 10 ⁻¹⁴
9	REAC	REAC:R-MMU-72766	Translation	3.779 × 10 ⁻¹⁴
10	REAC	REAC:R-MMU-72172	mRNA Splicing	1.135 × 10 ⁻¹³
11	REAC	REAC:R-MMU-74160	Gene expression (Transcription)	7.462 × 10 ⁻⁵
12	REAC	REAC:R-MMU-69278	Cell Cycle, Mitotic	4.826 × 10 ⁻⁶
13	REAC	REAC:R-MMU-450531	Regulation of mRNA stability by proteins that bind AU-rich elements	9.060 × 10 ⁻⁵

Figure S4. Biological pathways enrichment analysis, related to Figure 2.

Genes harbouring spurious transcription initiations were used for Gene Ontology enrichment analysis with online tools g:Profiler. Representative biological pathways and adjust p-values from Reactome are shown.

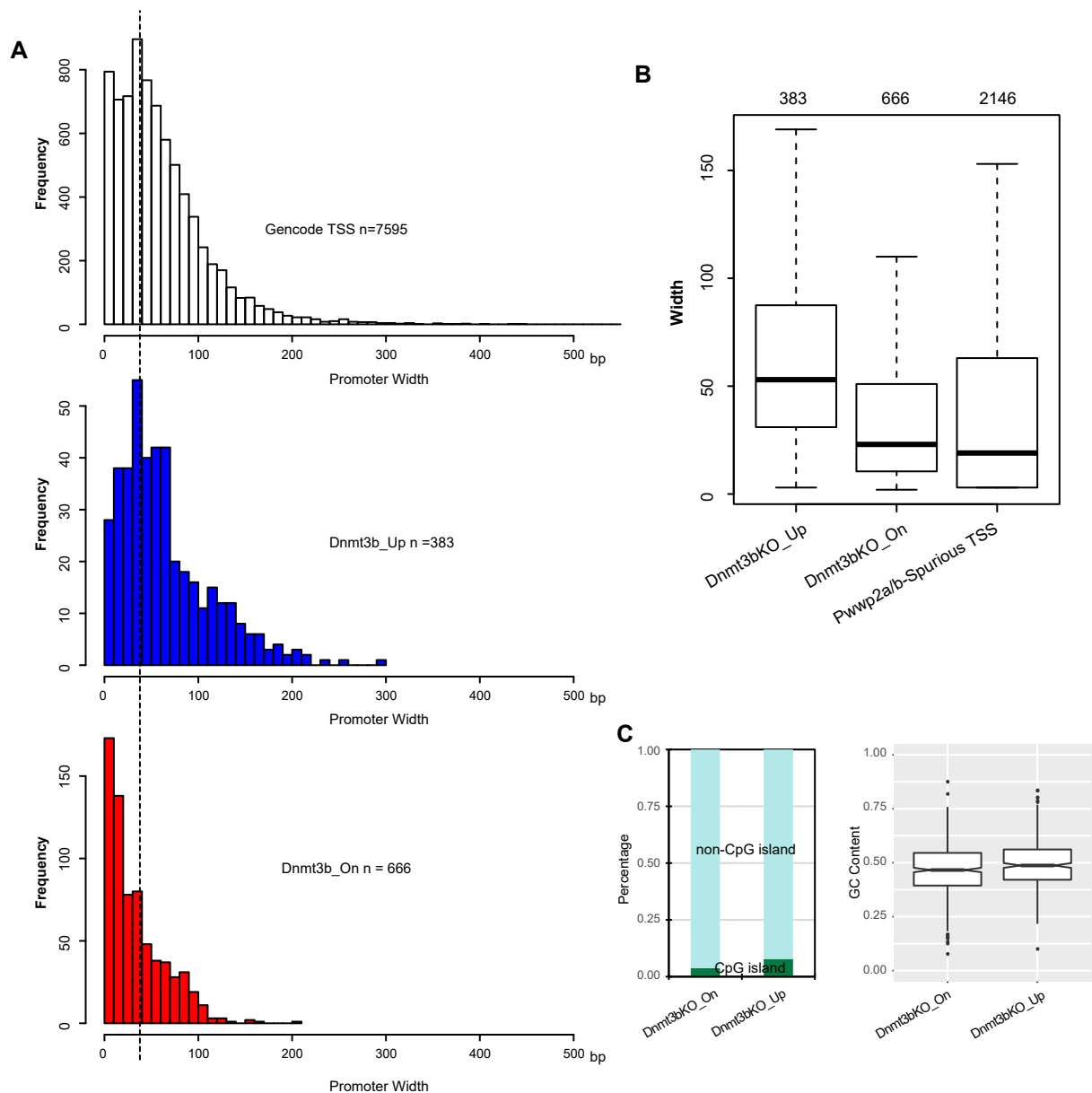


Figure S5. Width distribution and sequence content of DNMT3B-sensitive spurious TSSs, related to Figure 3.

(A) Histograms show width of consensus TSS called from DECAP-seq by CAGER pipeline. Top panel – Gencode annotated TSSs, middle panel – Dnmt3bKO UP TSSs, and bottom panel – Dnmt3bKO On TSS. TSS numbers are indicated and the dashed line shows the median width in Gencode TSSs. See also Table S3.

(B) Boxplot shows the TSS width comparison in Dnmt3bKO_Up, Dnmt3bKO_On, and PWWP2-sensitive Spurious TSSs. TSS numbers are shown on the top. Boxes indicate the median and IQRs, with whiskers indicating 1.5× the IQR, outliers are not shown.

(C) Percentage of Dnmt3bKO_On and Up TSSs that lie within a CpG island (Left), and their GC content (Right).

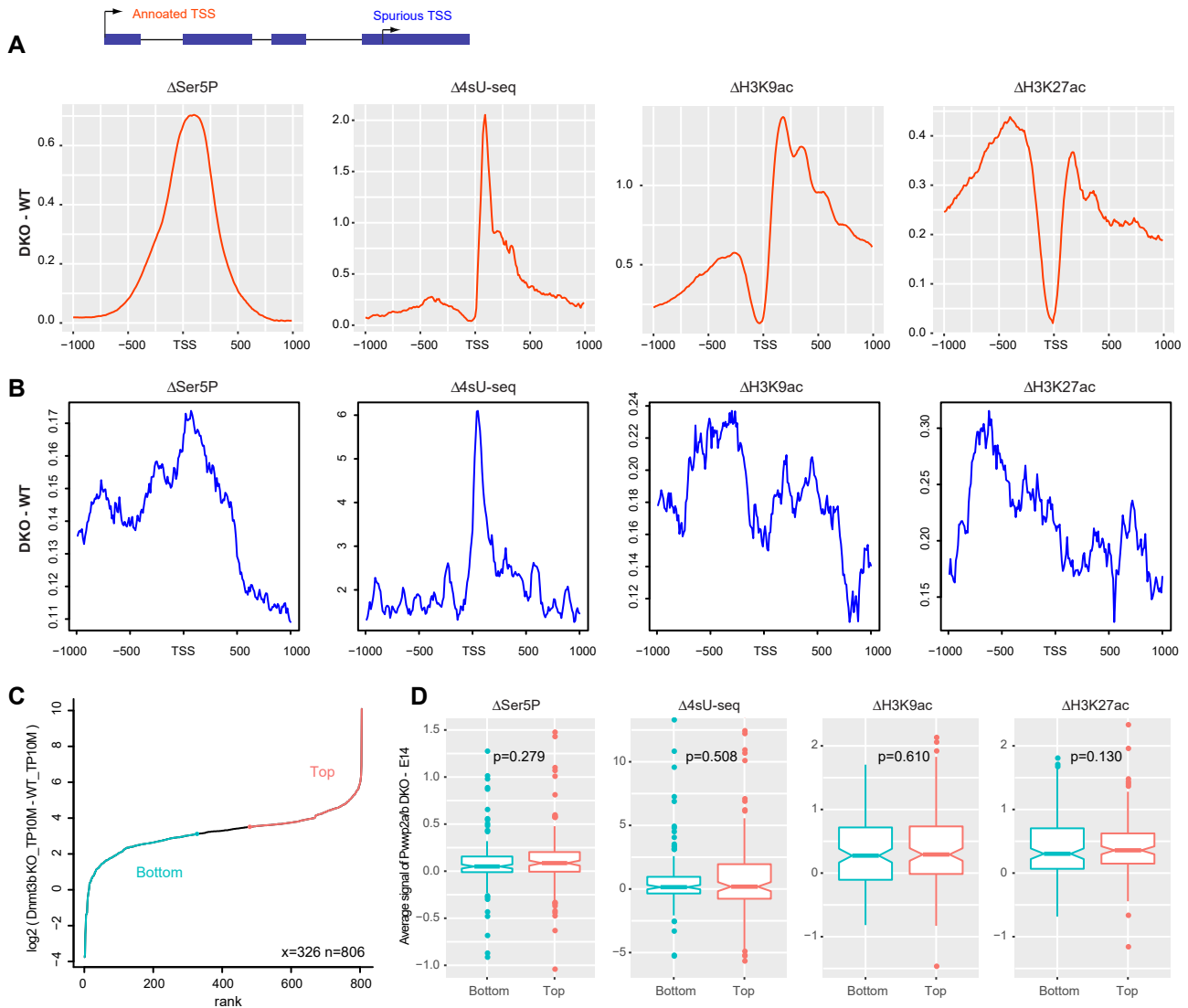


Figure S6. Changes of chromatin signatures and nascent transcription around PWWP2-sensitive spurious TSSs or DNMT3B sensitive spurious TSSs, related to Figure 4.

(A-B) Metaprofile of Pol II Ser5P, 4sU, H3K9ac, and H3K27ac of *Pwwp2a/b* DKO - E14 signal at the 1000 bp flanking annotated TSS (red) (A) and spurious TSSs (blue) (B), which all show a signal > 0 indicating greater occupancy in *Pwwp2a/b* DKO compared to E14.

(C-D) Same as Figure 4E,F but use the top and bottom 326 Dnmt3bKO_On and _Up spurious TSSs, matching the number in Figure 4C,D. In Figure 4C,D and 4E,F, the proportion is matched.

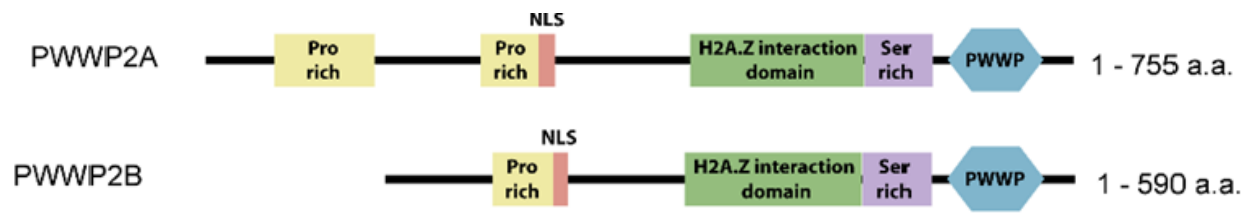


Figure S7. Domain composition of PWWP2A and PWWP2B, related to Figure 1 and Limitations of this study.

TRANSPARENT METHODS

Cell culture

Wildtype E14 and *Pwwp2a/b* DKO mouse embryonic stem cells were grown in Dulbecco's Modified Eagle Medium (DMEM, from Life Technologies) supplemented with 10% foetal calf serum (Seralab), 2 mM L-glutamine, 1X non-essential amino acids, 50 μ M β -mercaptoethanol, 50 g/mL penicillin/streptomycin (all from invitrogen) and 1000 U/mL of LIF in tissue culture dishes coated with PBS+1% gelatine. Cells were passaged using 0.05% trypsin-EDTA (Life Technologies) with 2% Chicken Serum (Life Technologies) and frozen in FCS +10% DMSO.

Nuclear RNA extraction

For the extraction of nuclear RNA, cells were lysed for 5 minutes on ice in lysis buffer (20 mM Tris pH 7.5, 150 mM NaCl, 5% glycerol, 0.5% NP-40, 1x protease inhibitor) then centrifuged at 2000 rpm to pellet nuclei. The nuclei were washed once in cold PBS then directly lysed in Trizol followed by standard RNA extraction. Nuclear RNA was prepared in biological replicates for wildtype and *Pwwp2a/b* DKO lines as the input RNA for CAGE-seq.

CAGE-seq library construction

Libraries for all samples (3 replicates for wildtype E14, 3 replicates for *Pwwp2a/b* DKO clone C7, and 2 replicates for DKO clone A1) were prepared using the nAnT-iCAGE-seq kit purchased from cage-seq.com. Briefly, 5 μ g of nuclear RNA for each sample was primed (random primers) and reverse transcribed to cDNA, followed by biotinylation of and capture of the 7-methylguanosine cap of all RNA Polymerase II transcripts. Adapters containing 3bp barcodes were ligated and samples were purified and multiplexed. As the nAnT-iCAGE-seq protocol involves no PCR amplification step, the multiplexed pool was quantified by HS DNA qubit denatured and loaded into the Illumina NextSeq 500 for single end 80bp sequencing.

CAGE-seq data analysis

Barcodes were extracted from pooled single-end RNA-seq reads and moved to the header line. The resulting fastq files were demultiplexed using the 3-letter barcodes, and then mapped to mm10 genome by STAR (v2.4.2a) (Dobin et al., 2013) with default parameters expect (`--outWigType bedGraph read1_5p --outFilterMultimapNmax 1 --outFilterMismatchNmax 4 --seedSearchStartLmax 15 --alignEndsType EndToEnd`). Only uniquely mapped reads were kept for further analysis. The

reads mapping summary for each sample are listed in Supplementary Table 1. CAGEr (1.28.0) (Haberle et al., 2015) was employed for CAGE-seq data analysis, including TSS detection, clustering, aggregation, promoter width (qLow=0.1 and qUp=0.9), and quantification (See attached codes). The CAGE-seq libraries were normalized (powerLaw) and TP10M (tags per million) scores were calculated as 10 million mapped reads in this study. TSS clusters shorter than 20bp from same strand were further aggregated to consensus peaks across replicates. GENCODE vM22 (comprehensive annotation) were used to annotate all CAGE peaks into four groups: promoter (Gencode), exonic, intronic, and intergenic. CAGE peaks were also compared for overlap with ATAC-seq and H3K27ac (high at promoters and enhancers). Differential expression analysis comparing CAGE expression in *Pwwp2a/b* DKO vs wildtype was performed for all TSSs, and classified as upregulated or downregulated (Brocks et al., 2017). To obtain the list of PWWP2-sensitive TSSs (n=2146), exonic and intronic sites were combined and sites potentially corresponding to eRNAs (sites which overlap with ATAC and H3K27ac peaks) were removed. Intragenic spurious sites which are not consistently upregulated in *Pwwp2a/b* DKO were regarded as PWWP2-insensitive spurious TSS (n=1108). The CAGE-seq overlap among samples in different groups (Figure S1D,E) were performed by comparing TSSs with TP10M>=1. Gene expression group were calculated from wildtype E14 RNA-seq and gene sets with high or low PWWP2A occupancy were obtained from previous ChIP-seq study (Zhang et al., 2018). CpG island annotation was retrieved from mm10 UCSC genome browser. The detailed scripts used for CAGE-seq analysis are available.

CAGE-seq mapped reads were split into positive and negative strands according to the flag filed in BAM file with Samtools (1.3) (Li et al., 2009), and then visualized in UCSC or IGV genome browser. Gene Ontology enrichment analysis of genes harbouring spurious TSSs from *Pwwp2a/b* DKO cells were performed using online tools g:Profiler (Raudvere et al., 2019). Bedtools (2.25.0) (Quinlan and Hall, 2010) were used to annotate TSSs detected from CAGE-seq. Annotated TSSs and PWWP2-sensitive spurious TSSs that are 2kb away from annotated TSS were taken for metagene profile, and DANPOS2 (Chen et al., 2013) was used for metagene profile with parameters setting (*--flank_up 1000 --flank_dn 1000 --heatmap 1 --bin_size 10 --excludeP 0.005*). Homer software (Heinz et al., 2010) was used to calculate motif enrichment. For discovery of motifs enriched in spurious TSS or annotated TSS compared to background, the 200bp (100nt upstream + 100nt downstream) region flanking spurious or annotated TSS were extracted as input, and random 200bp genomic sequence matched for GC content were served as background (*findMotifs.pl promoter.fa mouse output_dir -fasta random.fa*). For discovery of motifs enriched in spurious TSS over annotated TSSs, 200bp regions flanking spurious TSSs were used as input and 200bp regions flanking annotated TSSs were used as background (*findMotifs.pl spurious.promoter.fa mouse*

output_dir -fasta annotated.promoter.fa). R (3.6) package ggplot2 (3.2.1), dplyr (0.8.4), pheatmap (1.0.12), and beanplot (1.2) were used for statistics analysis and plot.

Reanalysis of *Dnmt3b* KO DECAP-seq and comparison with *Pwwp2/b* DKO data

The same CAGEr and gene expression group categorisation pipeline was applied to the DECAP-seq data generated from wildtype and *Dnmt3b* KO mESCs in the Neri et al. study, retrieved from GSE72854 (Neri et al., 2017). DECAP peaks were categorised as annotated promoter-TSS, exonic, intronic, or intergenic. DECAP peaks detected in wildtype cells were overlapped with DECAP peaks in *Dnmt3b* KO cells and sorted into four states as described in the Neri et al. study – “On”: peaks present only in KO, “Off”: peaks present only in wildtype, Up: upregulated in KO, and Down: downregulated in KO. *Dnmt3b*KO_Up and _On TSSs which overlapped neither with ATAC nor H3K27ac peaks were merged for meta-analysis, and only TSSs which are 2kb away from annotated TSSs and 100bp away from PWWP2-sensitive TSSs were taken for analysis. Intersection and distance between DNMT3B-sensitive and PWWP2-sensitive spurious TSSs were calculated using Bedtools (2.25.0) (Quinlan and Hall, 2010).

Public data sets used in this study

ATAC-seq for XY mESCs is from GSM2247119, H3K27ac and H3K36me3 ChIP-seq in E14 cells is from mouse ENCODE ENCSR000CGQ and ENCSR253QPK respectively. 4sU-RNA-seq, PWWP2A, Pol Ser5, H3K9ac, and H3K27ac ChIP-seq are from GSE112114. DECAP-seq data for E14 and its derived *Dnmt3b* knockout are from GSE72854.

Supplemental References

- Brocks, D., Schmidt, C.R., Daskalakis, M., Jang, H.S., Shah, N.M., Li, D., Li, J., Zhang, B., Hou, Y., Laudato, S., *et al.* (2017). DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat Genet* **49**, 1052-1060.
- Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* **23**, 341-351.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21.
- Haberle, V., Forrest, A.R., Hayashizaki, Y., Carninci, P., and Lenhard, B. (2015). CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res* **43**, e51.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.
- Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72-77.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.
- Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* **47**, W191-W198.
- Zhang, T., Wei, G., Millard, C.J., Fischer, R., Konietzny, R., Kessler, B.M., Schwabe, J.W.R., and Brockdorff, N. (2018). A variant NuRD complex containing PWWP2A/B excludes MBD2/3 to regulate transcription at active genes. *Nat Commun* **9**, 3798.