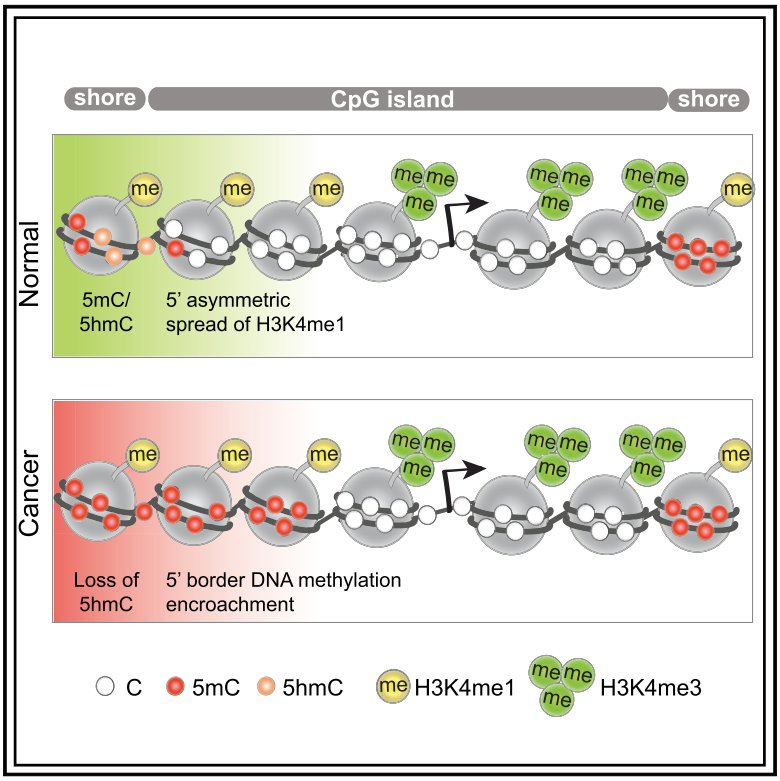
Article

Cancer Cell

DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns

Graphical Abstract



Authors

Ksenia Skvortsova,
Etienne Masle-Farquhar,
Phuc-Loi Luu, ...,
Christopher C. Goodnow,
Clare Stirzaker, Susan J. Clark

Correspondence

c.stirzaker@garvan.org.au (C.S.),
s.clark@garvan.org.au (S.J.C.)

In Brief

Skvortsova et al. identify promoter CpG islands that display partial methylation encroachment across one or both borders, which is often associated with gene suppression, and show that the pattern of H3K4me1 at CpG island borders in normal cells predicts patterns of cancer CpG island hypermethylation.

Highlights

- Promoter CpG islands display asymmetric border methylation encroachment in cancer
- 5hmC is enriched in normal cells at CpG island shores prone to methylation spread
- H3K4me1 patterns at CpG island borders are associated with the mode of encroachment
- H3K4me1 disruption results in DNA methylation alterations at island borders

Skvortsova et al., 2019, Cancer Cell *35*, 297–314
February 11, 2019
https://doi.org/10.1016/j.ccell.2019.01.004

CellPress

Cancer Cell
Article

DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns

Ksenia Skvortsova,[1,15] Etienne Masle-Farquhar,[2] Phuc-Loi Luu,[1] Jenny Z. Song,[1] Wenjia Qu,[1] Elena Zotenko,[1] Cathryn M. Gould,[1] Qian Du,[1] Timothy J. Peters,[1] Yolanda Colino-Sanguino,[1] Ruth Pidsley,[1] Shalima S. Nair,[1] Amanda Khoury,[1] Grady C. Smith,[1] Lisa A. Miosge,[3] Joanne H. Reed,[2] James G. Kench,[4,5,14] Mark A. Rubin,[6,7,8,9,10] Lisa Horvath,[5,11,12,13,14] Ozren Bogdanovic,[11,15] Sue Mei Lim,[16,17,18] Jose M. Polo,[16,17,18] Christopher C. Goodnow,[2,11] Clare Stirzaker,[1,11,19,*] and Susan J. Clark[1,11,19,20,*]

[1]Epigenetics Research Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, 384 Victoria St, Sydney, NSW 2010, Australia
[2]Immunogenomics Laboratory, Immunology Division, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
[3]Immunogenomics Laboratory, John Curtin School of Medical Research, Australian National University, Canberra, ACT 2601, Australia
[4]Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Sydney, NSW 2010, Australia
[5]Sydney Medical School, University of Sydney, Sydney, NSW 2010, Australia
[6]Caryl and Israel Englander Institute for Precision Medicine, New York Presbyterian Hospital-Weill Cornell Medicine, New York 10021, USA
[7]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York 10065, USA
[8]Sandra and Edward Meyer Cancer Center at Weill Cornell Medicine, New York 10065, USA
[9]Department for Biomedical Research, University of Bern, Bern, CH-3012, Switzerland
[10]Bern Center for Precision Medicine, Inselspital, Bern University Hospital, Bern, CH-3012, Switzerland
[11]St Vincent's Clinical School, UNSW, Sydney, NSW 2010, Australia
[12]Department of Medical Oncology, Chris O'Brien Lifehouse, Sydney, NSW 2050, Australia
[13]Royal Prince Alfred Hospital, Sydney, NSW 2050, Australia
[14]Cancer Division, The Kinghorn Cancer Centre, Sydney, NSW 2010, Australia
[15]Developmental Epigenomics Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, NSW 2010, Australia
[16]Department of Anatomy & Developmental Biology, Monash University, Melbourne, VIC 3800, Australia
[17]Development and Stem Cells Program, Monash Biomedicine Discovery Institute, Melbourne, VIC 3800, Australia
[18]Australian Regenerative Medicine Institute, Monash University, Melbourne, VIC 3800, Australia
[19]These authors contributed equally
[20]Lead Contact
*Correspondence: c.stirzaker@garvan.org.au (C.S.), s.clark@garvan.org.au (S.J.C.)
https://doi.org/10.1016/j.ccell.2019.01.004

SUMMARY

Promoter CpG islands are typically unmethylated in normal cells, but in cancer a proportion are subject to hypermethylation. Using methylome sequencing we identified CpG islands that display partial methylation encroachment across the 5′ or 3′ CpG island borders. CpG island methylation encroachment is widespread in prostate and breast cancer and commonly associates with gene suppression. We show that the pattern of H3K4me1 at CpG island borders in normal cells predicts the different modes of cancer CpG island hypermethylation. Notably, genetic manipulation of *Kmt2d* results in concordant alterations in H3K4me1 levels and CpG island border DNA methylation encroachment. Our findings suggest a role for H3K4me1 in the demarcation of CpG island methylation borders in normal cells, which become eroded in cancer.

Significance

How CpG islands are maintained in an unmethylated state in normal cells and what triggers the erosion of CpG island methylation borders in tumors is still enigmatic. We show that the pattern of H3K4me1-marked nucleosomes in embryonic stem cells and normal epithelial cells predicts the mode of promoter CpG island hypermethylation in cancer. Depletion or enrichment of H3K4me1 levels at the borders of CpG islands results in loss or gain of DNA methylation encroachment. Prostate cancers harboring *KMT2D* mutations display reduced erosion of CpG island border methylation. While H3K4me1 has been mainly considered as a mark of enhancers, our findings establish a role for H3K4me1 in determining promoter CpG island DNA methylation status in cancer.

INTRODUCTION

The human cancer epigenome is largely reorganized, including hypermethylation of CpG islands in gene promoters and genome-wide hypomethylation (Stirzaker et al., 2014) (Figure 1A). CpG island methylation and its relationship to chromatin modifications and gene expression have been widely studied. H3K27me3 co-exists with the active H3K4me3 to mark bivalent chromatin at ~22% of promoter CpG islands and is required for suspended transcription prior to the lineage-specific differentiation (Azuara et al., 2006; Bernstein et al., 2006; Mikkelsen et al., 2007; Ohm et al., 2007). A number of seminal papers reported that pre-marking of bivalent chromatin at CpG islands in embryonic stem cells (ESCs) (Ohm et al., 2007; Schlesinger et al., 2007; Widschwendter et al., 2007) and adult stem cells (Easwaran et al., 2012) renders CpG islands susceptible to DNA methylation in cancer cells. H3K27me3 and H3K4me3 chromatin signature at these CpG islands is commonly lost in cancer and replaced with DNA methylation (Gal-Yam et al., 2008). However, the mechanisms responsible for this switch are still unresolved.

The presence of bivalent chromatin at a subset of promoter CpG islands in ESCs, however, does not fully account for the variability of CpG islands that become hypermethylated in cancers. Furthermore, the vast majority of bivalent chromatin domains in ESCs resolve upon differentiation into monovalent H3K4me3-active or H3K27me3-repressive domains (Bernstein et al., 2006), resulting in a smaller proportion of bivalent promoter CpG islands in cancer precursor cells (Easwaran et al., 2012). It has been proposed that the gene expression status in cancer precursor cells can potentially define the susceptibility to promoter CpG island hypermethylation in cancer (Sproul et al., 2011, 2012). That is, genes prone to hypermethylation are expressed in a more tissue-specific manner, being either lowly expressed or repressed in matching normal tissues, suggesting the existence of further characteristic chromatin signatures at these promoter CpG islands.

Distinct roles for H3K4 histone methyltransferases (HMTs) have been observed in the regulation of gene expression (Rao and Dou, 2015). The family of H3K4 HMTs includes six main members, KMT2A to KMT2G. While H3K4me3, at the promoters of active genes, is predominantly implemented by KMT2F and KMT2G (Clouaire et al., 2012), KMT2B mediates H3K4me3 at bivalent and lowly expressed gene promoters (Denissov et al., 2014; Hu et al., 2013b). Furthermore, KMT2C/D promote conditional repression of a subset of genes through the deposition of histone H3K4 monomethylation (H3K4me1) at their promoters (Cheng et al., 2014). Importantly, H3K4 HMTs possess distinct substrate specificity with KMT2F/G being capable of mono-, di-, and trimethylation, KMT2A/B of mono- and dimethylation, and KMT2C/D of predominantly monomethylation (Rao and Dou, 2015). H3K4me3 marked nucleosomes are reported to be important for the establishment and maintenance of unmethylated CpG islands at the promoters of active genes as this modification is agnostic to DNA methylation (Weber et al., 2007). Indeed the DNA methyltransferase DNMT3A/L shows no detectable binding to H3K4me2/3 but displays the highest binding affinity to nucleosomes marked with H3K4me0, and interestingly only a ~2-fold reduction in binding affinity to H3K4me1 (Noh et al., 2015; Ooi et al., 2007). This raises the possibility that H3K4 monomethylation may play a role in shaping DNA methylation patterns at regulatory elements.

RESULTS

Cancer-Associated Promoter CpG Island DNA Methylation Exhibits Distinct Patterns

To address the pattern of cancer-associated CpG island hypermethylation in more detail, we first performed a methylome analysis at single-nucleotide resolution using whole-genome bisulfite sequencing (WGBS) in normal prostate epithelial cells (PrECs) and the prostate cancer cell line LNCaP with >30× coverage. We compared the average change in DNA methylation of promoter-associated CpG islands (n = 8,835) and showed that, while ~70% of CpG islands remain unmethylated, ~30% of the CpG islands gain different amounts of DNA methylation in LNCaP cells (Figure 1B). Notably, in addition to the "extensive" (50%–100%) methylation gain, there were CpG islands that also acquired "moderate" levels (10%–50%) of aberrant DNA methylation (Figure 1B). To further interrogate the nature of "extensive" versus "moderate" hypermethylation changes, we compared methylation patterns across CpG islands using k-means clustering (Figure S1A) and Pearson's correlation coefficient (Figures S1B and 1C). We found that 71.5% of CpG islands remained unmethylated (Figure 1C; un) and 7.3% of CpG islands were hypermethylated across the entire CpG island in LNCaP cells compared with PrECs and human ESCs (hESCs) (Figure 1C; hyper). Notably, we also observed a mode of aberrant hypermethylation (13.0%) that involved partial methylation encroachment across the 5′ and/or 3′ borders of the CpG island and the creation of new discrete DNA methylation borders and smaller unmethylated islands (Figure 1C; 5′, 3′, 5′-3′). The observed methylation patterns are not affected by CpG island length (Figure S1C). Examples of 5′ and 3′ asymmetric methylation encroachment are shown in Figure 1D. We next asked whether the gain in methylation across the CpG island borders corresponds to different methylation levels at the adjacent CpG island shores (500 bp upstream/downstream of CpG island boundary). We found that CpG islands with asymmetrical methylation encroachment across the 5′ or 3′ border have elevated methylation in normal cells only at the adjacent 5′ or 3′ shores, respectively, whereas islands that are extensively methylated or undergo bidirectional methylation encroachment have elevated methylation levels at both shores (Figure 1E).

The DNA methylation profile can also be used to define functional regulatory elements. For example, undermethylated regions (UMRs) and lowly methylated regions (LMRs) are indicative of active regulatory regions. MethylseekR partitions methylomes into CpG-poor LMRs, corresponding to distal regulatory sites, and CpG-rich UMRs at proximal regulatory sites (Burger et al., 2013; Stadler et al., 2011). We therefore applied MethylSeekR to the WGBS data of PrECs and identified 10,380 UMRs, of which 64% harbor a CpG island (Figure S1D), spanning ~2.5 kbp (Figure S1E). Notably, the UMRs also displayed similar modes of aberrant DNA methylation encroachment (Figure S1F).

DNA Hypermethylation Modes Are Uniform across Clinical Cancer Samples

We next asked whether methylation encroachment across CpG islands is a general characteristic of cancer. We

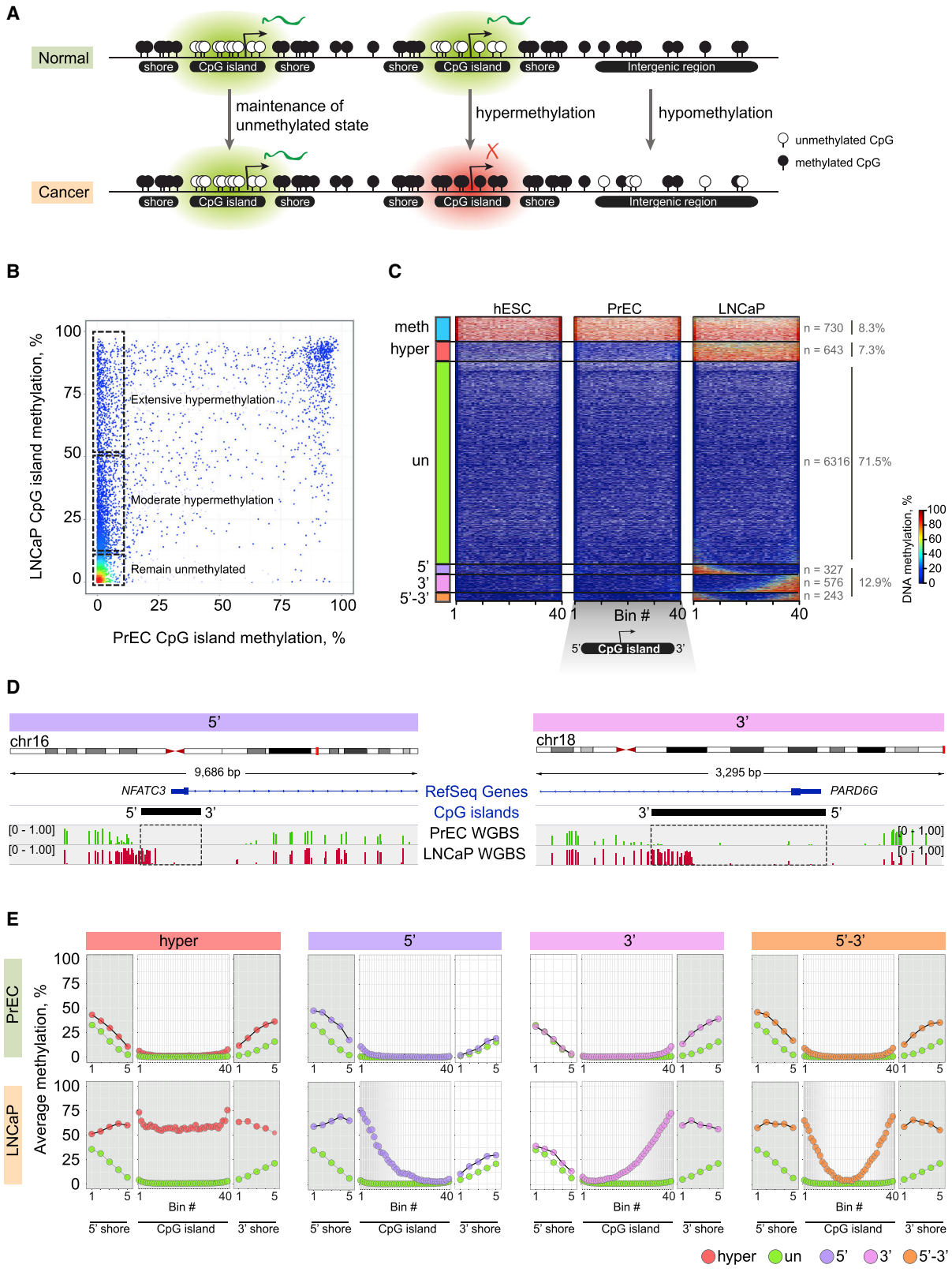


Figure 1. Aberrant DNA Methylation Changes in Cancer Cells

(A) Schematic depicting promoter CpG island hypermethylation in cancer.

(B) Scatterplot showing average promoter CpG island DNA methylation in PrECs and LNCaP cells. Each dot represents average methylation of a single CpG island.

(legend continued on next page)

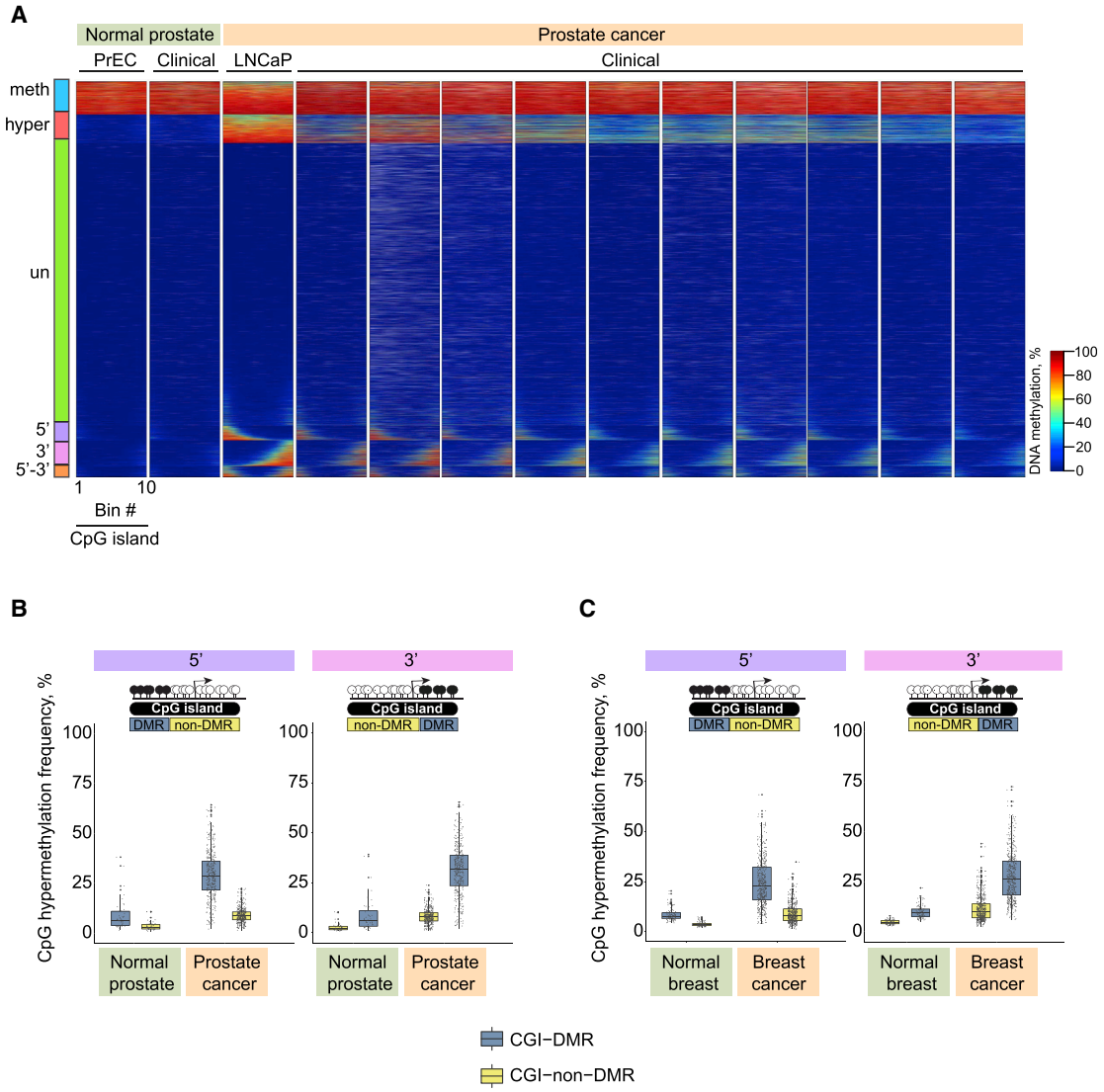


Figure 2. DNA Methylation Changes in Clinical Prostate and Breast Cancer

(A) Heatmaps showing methylation patterns of promoter CpG islands in normal prostate tissue (n = 1 shown) and prostate cancer tissues (n = 10 shown). Each line represents a single CpG island. The separation into groups is as in Figure 1C.

(B and C) Boxplots showing DNA hypermethylation frequency of CpG sites residing within promoter CpG islands with DNA methylation encroachment in TCGA prostate (B) and breast (C) cancer and normal tissues. Each dot represents a single tissue sample. Boxplot boundaries indicate the first and third quartiles of the data points, with the median value within the box. The whiskers extend to 1.5 interquartile range (IQR) from the boundaries. Data points lying outside of 1.5 IQR are shown as closed circles.

See also Figure S2 and Table S1.

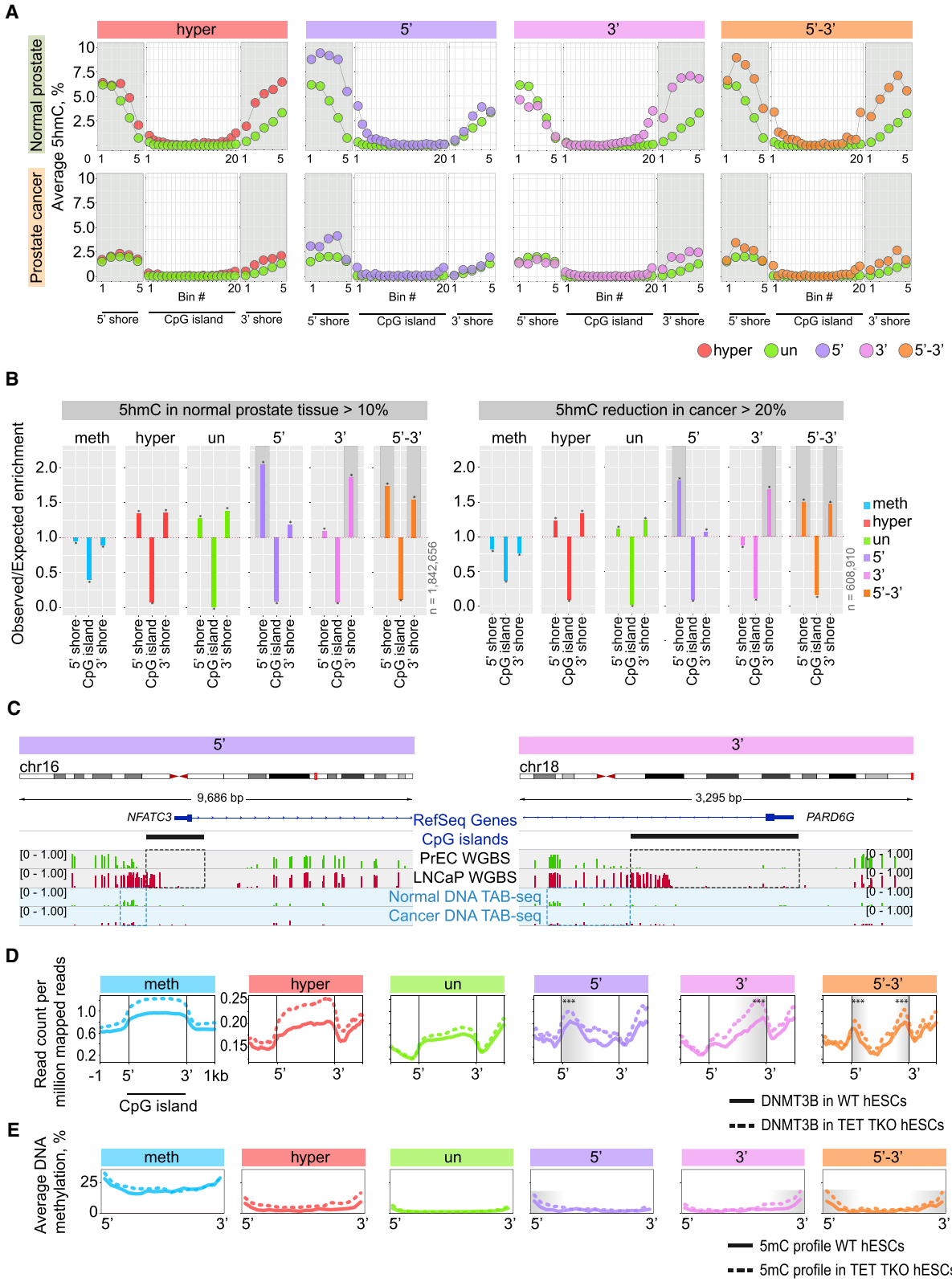
performed WGBS on clinical prostate cancer and adjacent normal prostate samples (Table S1) and found that the promoter CpG islands clustered into similar modes of DNA hypermethylation patterns as defined in LNCaP cells (Figure 2A). We then analyzed public WGBS data from primary normal and breast cancer tissues and cell lines, and found that methylation encroachment was associated with the same subset of CpG islands (Figure S2A).

(C) Heatmaps showing methylation patterns of promoter CpG islands in hESC, PrEC, and LNCaP cells. Each line represents a single CpG island. Each CpG island was binned into 40 equally sized bins and average methylation per bin was calculated. Bins without CpG sites are depicted in white.

(D) Integrative Genomics Viewer (IGV) browser track showing WGBS methylation data for promoter CpG islands of *NFATC3* and *PARD6G* in PrEC and LNCaP cells with 5′ and 3′ DNA methylation encroachment, respectively, in the cancer cells.

(E) Average methylation patterns across CpG islands separated into groups as in (C) and their shores in PrEC and LNCaP cells. Color key denotes the six groups.

See also Figure S1.



(legend on next page)

We further extended our analysis to prostate and breast cancer HumanMethylation 450K data available from The Cancer Genome Atlas (TCGA) consortium. We divided CpG islands (with 5′ or 3′ methylation encroachment) into two classes, namely internal borders that are prone to DNA methylation encroachment in LNCaP cells (CGI-DMR, blue) or internal borders that remain unmethylated (CGI-non-DMR, yellow). Next, we defined CpG probes overlapping CGI-DMRs and CGI-non-DMRs (Figure S2B) and calculated the percentage of CpG probes that display >40% DNA methylation for each tissue sample. The data confirmed elevated DNA methylation at the CGI-DMRs compared with the CGI-non-DMRs in cancer tissues (Figures 2B and 2C).

DNA Hydroxymethylation and DNMT3B Are Enriched at the Borders of CpG Islands with Methylation Encroachment

5-Hydroxymethylation (5hmC) has been shown to play a role in DNA demethylation and the maintenance of unmethylated CpG islands in normal tissues (Jin et al., 2014; Li et al., 2016). Moreover a global reduction of 5hmC in cancer tissues (Jin et al., 2011; Lian et al., 2012) suggests that 5hmC depletion may also be involved in CpG island hypermethylation. To further address whether 5hmC is associated with demarcation of methylation at CpG island borders, we performed whole-genome 5hmC profiling and 5mC WGBS from matching prostate cancer and normal tissue samples (n = 3 pairs, Table S1). The data confirmed similar modes of CpG island 5mC DNA hypermethylation patterns (Figure S3A). The analysis identified 2.4%–9.7% of CpG sites that were hydroxymethylated in normal prostate tissues with a significant 5hmC depletion in prostate cancer (0.2%–1.3%) (Figures S3B and S3C). Next, we examined the distribution of 5hmC associated with different modes of DNA methylation gain. In normal prostate tissues, we observed increased average 5hmC levels at the CpG island shores (Figure 3A), in agreement with other cancer types (Jin et al., 2014; Li et al., 2016). Notably, 5hmC sites are enriched at the 5′ or 3′ shores of CpG islands in normal cells that are prone to asymmetric methylation encroachment in cancer (Figures 3B and S3D). CpG sites that, in turn, lose 5hmC in cancer, are enriched at the 5′ or 3′ shores of CpG islands that undergo asymmetric methylation encroachment in cancer (Figures 3B and S3E). Figure 3C shows examples of genes with CpG island DNA methylation encroachment accompanied by 5hmC reduction at the corresponding shores. DNMT3B is also enriched at 5′ and 3′ internal CpG island boundaries that are susceptible to DNA methylation encroachment in cancer, and this enrichment is elevated in TET triple-knockout (TKO) hESCs (Figure 3D). Notably, we found that DNA methylation is also elevated in the TKO cells at these internal CpG island borders that display 5′ and 3′ encroachment in cancer (Figure 3E).

Relationship between Promoter CpG Island DNA Methylation Encroachment, Gene Expression, and Intrinsic DNA Features

To investigate whether the different modes of DNA hypermethylation in cancer cells result in altered gene expression, we analyzed RNA sequencing (RNA-seq) data from PrECs and LNCaP cells (Figure S4A). We found that genes prone to cancer-associated CpG island hypermethylation, compared with those without, had reduced expression in normal cells (Figure 4A). Moreover, genes that displayed 3′ versus 5′ cancer methylation border encroachment were notably less expressed in normal cells (Figure 4A) and showed even greater repression in the cancer cells (Figure 4B). In total, 20%–40% of genes undergoing methylation encroachment showed statistically significant reduction in expression (Figure S4B); however, 12%–27% of such genes showed elevated expression (Figure S4B) and increased enrichment of the active histone mark H3K4me3 at promoter CpG islands (Figure S4C). DNA hypermethylation and H3K4me3 enrichment in cancer has previously been associated with aberrant expression of particular gene isoforms (Bert et al., 2013) and alterations in the balance of gene isoforms (Zhao et al., 2016). We found that downregulation of gene isoforms was also associated with DNA methylation encroachment across transcription start sites (TSSs) (Figure 4C). In addition, cap analysis gene expression sequencing (CAGE-seq) data shows a reduction of transcription at TSSs associated with DNA methylation encroachment (Figure 4D). Figure 4E shows examples of promoter CpG islands with cancer-associated methylation encroachment accompanied by gene isoform repression.

Interestingly we observed that in PrECs, the CpG island borders susceptible to DNA methylation encroachment in cancer generally show reduced enrichment of active TSSs (Figure S4D), regulatory factor binding sites (Figure S4E), and display higher nucleosome occupancy (Figure S4F) and lower CpG density (Figure S4G). Intriguingly, G-quadruplex (G4) DNA structures, as measured by G4 chromatin immunoprecipitation sequencing (ChIP-seq) data (Mao et al., 2018), are also depleted at the internal borders of CpG island that are susceptible to DNA methylation encroachment, in contrast to the observed enrichment for G4 structures at the internal CpG island regions that remain unmethylated in cancer (Figure S4H).

Figure 3. DNA Hydroxymethylation at Promoter CpG Island Shores in Normal Prostate and Prostate Cancer Tissues

(A) Average 5hmC patterns across CpG islands separated into groups and their shores in normal prostate and prostate cancer tissues (sample #24023). Gray shading indicates the highest 5hmC enrichment in normal prostate at the shores of CpG islands susceptible to DNA methylation encroachment in prostate cancer.

(B) Observed over expected enrichment of CpGs hydroxymethylated in normal prostate tissue and CpGs with reduced hydroxymethylation in cancer at CpG islands and their shores (sample #24023). Gray shading highlights CpG island shores with the highest 5hmC enrichment in normal prostate tissue.

(C) IGV browser track depicting DNA hydroxymethylation enrichment in normal prostate tissue at the shores of CpG islands prone to DNA methylation encroachment in cancer.

(D and E) Average DNMT3B ChIP-seq enrichment (D, ***$p = 2.2 \times 10^{-16}$, Wilcoxon rank test) and WGBS DNA methylation levels (E) in WT and TKO human ESCs (GSE89728).

See also Figure S3 and Table S1.

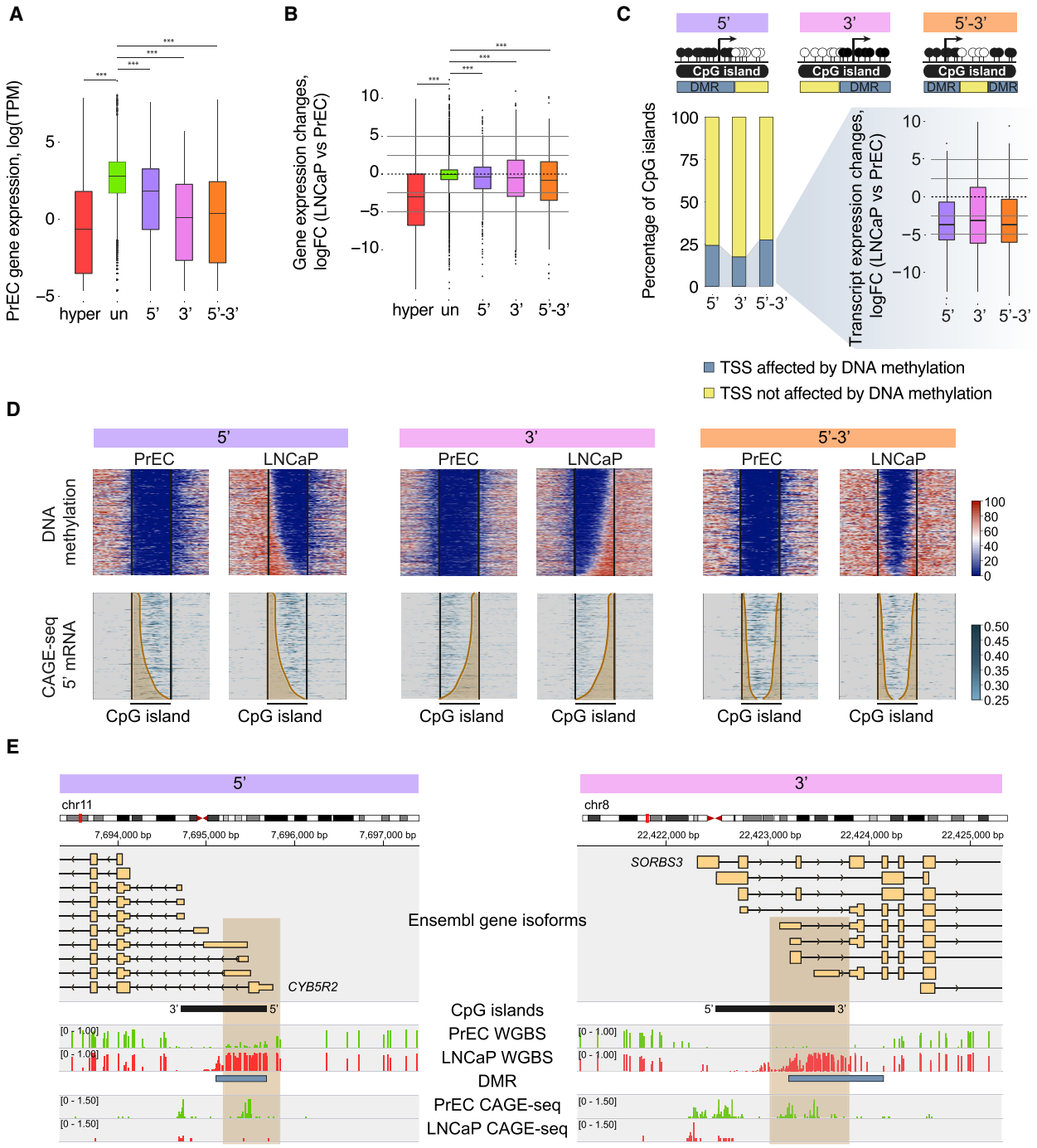


Figure 4. DNA Methylation Encroachment in Cancer and Gene Expression Changes

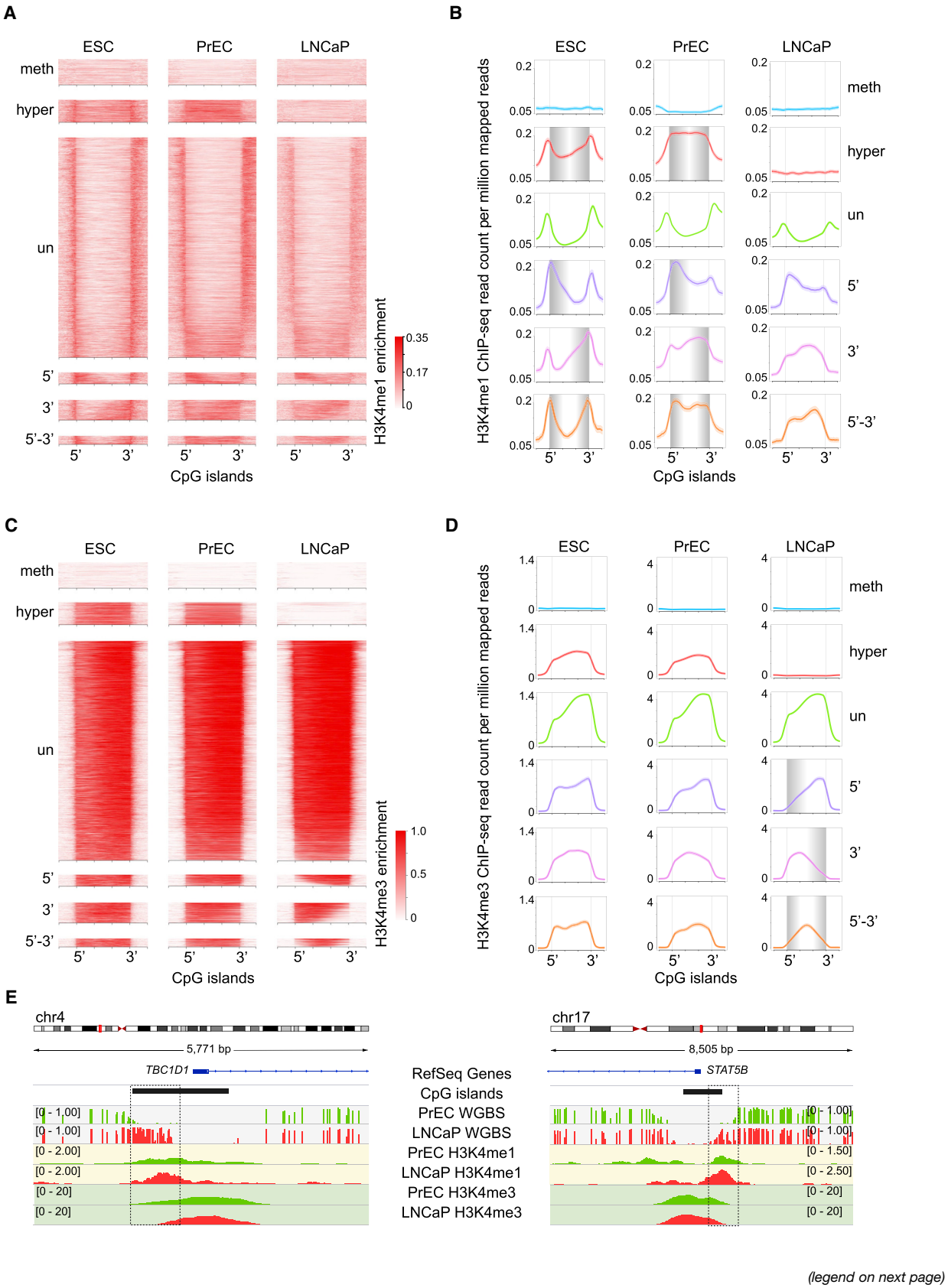
(A and B) Expression in PrECs (A) and expression differences between LNCaP and PrECs (B) of genes that display different modes of DNA methylation change at their promoter CpG islands in cancer (***$p < 0.01$, Wilcoxon rank test). Boxplot boundaries indicate the first and third quartiles of the data points, with the median value within the box. The whiskers extend to 1.5 IQR from the boundaries. Data points lying outside of 1.5 IQR are shown as closed circles.

(C) Expression changes of transcript isoforms (logFC) affected by DNA methylation encroachment. Bar plots indicate the percentage of CpG islands with known transcripts affected (blue) and not affected (yellow) by DNA methylation encroachment. Boxplot boundaries indicate the first and third quartiles of the data points, with the median value within the box. The whiskers extend to 1.5 IQR from the boundaries. Data points lying outside of 1.5 IQR are shown as closed circles.

(D) Heatmaps showing DNA methylation patterns and CAGE-seq 5′ end mRNAs enrichment in PrECs and LNCaP cells at promoter CpG islands undergoing DNA methylation encroachment in cancer. Brown line indicates the extent of DNA methylation encroachment.

(E) IGV browser track depicting CAGE-seq isoform-specific 5′ end mRNA signal changes at promoter CpG islands upon 5′ and 3′ DNA methylation encroachment in LNCaP cells. Brown shading indicates the loss of CAGE-seq-derived 5′ end mRNA signal in LNCaP cells upon DNA methylation encroachment.

See also Figure S4.



(legend on next page)

Finally, we found that promoter CpG islands susceptible to methylation encroachment are enriched for genes in cancer-associated pathways (Figure S4I).

Role for Histone Mark H3K4me1 at the Borders of CpG Islands

To determine whether there is a relationship between different modes of DNA methylation gain and chromatin signatures, we first analyzed ChIP-seq data of H3K4me1, H3K4me3, and H3K27me3 from the ENCODE project for hESCs. We identified similar H3K4me1 and H3K27me3 chromatin enrichment patterns between the different modes of CpG island hypermethylation in cancer (Figures 5A and S5A). We observed higher enrichment of H3K4me1 and H3K27me3 in hESCs across the body of the CpG islands that become extensively hypermethylated in cancer. In contrast, CpG islands that remain unmethylated in cancer show a bimodal enrichment of H3K4me1 and H3K27me3 at the CpG island borders in hESCs and depletion in H3K4me1 and H3K27me3 across the island (Figures 5A, 5B, S5A, and S5B). Notably, the bimodal H3K4me1 enrichment at the CpG island borders is maintained in normal PrECs (Figures 5A and 5B) whereas the H3K27me3 bimodal mark is essentially depleted (Figures S5A and S5B).

To our surprise, we found an asymmetric pattern of H3K4me1 and H3K27me3 enrichment in hESCs at the corresponding 5′ or 3′ internal CpG island borders prone to DNA methylation encroachment in cancer (Figures 5A, 5B, S5A, and S5B). Importantly, H3K4me1 enrichment is maintained in PrECs at the internal borders of these CpG islands (Figures 5A and 5B), in contrast to H3K27me3, which is significantly depleted (Figures S5A and S5B).

H3K4me3, in contrast to H3K4me1 and H3K27me3, is highly enriched across the body of unmethylated CpG islands in both hESCs and PrECs but completely absent from the CpG island borders (Figures 5C and 5D). Interestingly, only CpG dense enhancer regions that are susceptible to DNA hypermethylation in cancer show similar dynamics with more H3K4me1 and less H3K4me3 in normal cells (Figure S5C). In cancer cells, H3K4me3 is also lost from DNA hypermethylated islands and is notably absent from the internal borders of the islands that undergo DNA methylation encroachment (Figures 5C and 5D). Figure 5E depicts candidate CpG islands showing 5′ and 3′ methylation encroachment in cancer pre-marked by H3K4me1 in normal prostate cells and H3K4me3 enrichment across the region of the CpG island that remains unmethylated. This mutually exclusive pattern of H3K4me3 and H3K4me1 enrichment suggests a transition or critical interface between H3K4me1- and H3K4me3-marked nucleosomes in demarcating the boundaries of CpG island DNA methylation in normal cells and DNA methylation encroachment in cancer.

H3K4me1/H3K4me3 Ratio in Normal Cells Predicts the Extent of DNA Methylation Encroachment in Cancer

We next asked whether chromatin and genetic features at promoter CpG islands in normal cells could computationally predict DNA methylation gain in cancer. We separated CpG islands that remain unmethylated in prostate and breast cancer from those that become hypermethylated. For CpG islands prone to DNA methylation encroachment, we separated CpG sites that acquire methylation in cancer from those that remain unmethylated. For each CpG site we calculated the enrichment of histone marks in normal prostate and primary mammary epithelial cells (HMECs) using blkbox (Guennewig et al., 2017). blkbox identified random forests as the best-performing algorithm with receiver-operating characteristic area under the curve of ~0.88–0.98 (Figure S6A). The analysis revealed that the H3K4me1/H3K4me3 ratio in both normal cells shows the highest prediction importance in distinguishing CpG islands that become hypermethylated in cancer from those that remain unmethylated, followed by H3K27me3/H3K4me3 ratio in prostate and H3K4me1/H3K27me3 ratio in breast cells (Figure S6B). Notably, the H3K4me1/H3K4me3 ratio in both PrECs and HMECs is also statistically significantly higher at CpG islands that become hypermethylated in cancer compared with CpG islands that remain unmethylated (Figure S6C). Similarly, internal 5′ and 3′ CpG island borders that undergo DNA methylation encroachment in cancer can be distinguished by a higher H3K4me1/H3K4me3 ratio in normal prostate and breast cells from the internal CpG island borders that remain unmethylated (Figures S6B and S6C).

CpG "Seeding" Methylation at H3K4me1-Marked Island Borders that Become Hypermethylated in Cancer

To address the relationship of DNA molecules marked by H3K4me1 and their DNA methylation state, we performed H3K4me1 ChIP followed by bisulfite treatment and sequencing (BisChIP-seq) (Statham et al., 2012) on PrECs and LNCaP cells. The data revealed that in PrECs, H3K4me1-marked DNA possess elevated levels of CpG methylation at the internal borders of CpG islands that undergo aberrant DNA methylation encroachment in cancer, whereas CpG islands that remain unmethylated show the lowest level of DNA methylation (Figure 6A). In LNCaP cells, H3K4me1-marked DNA show substantial levels of aberrant methylation at CpG islands and border regions (Figure 6A). H3K4me1 BisChIP-seq sequencing reads depict CpG methylation "seeding" at H3K4me1-marked nucleosomal DNA in PrECs that acquire high levels of CpG methylation in the cancer cells (Figure 6B). Similarly, these regions are prone to spurious DNA methylation in aging (Figure 6C). Overall, H3K4me1-marked nucleosomal DNA at CpG islands borders carry low levels of cytosine methylation in PrECs and

Figure 5. Enrichment of H3K4me1- and H3K4me3-Marked Nucleosomes at Promoter CpG Islands

(A and B) Heatmaps (A) and average plots (B) showing H3K4me1 enrichment at different groups of CpG islands in hESCs, PrECs, and LNCaP cells. Gray shading in (B) highlights H3K4me1 enrichment in hESC and PrEC cells across the entire CpG island (hyper) or internal CpG island borders (5′, 3′, 5′-3′) undergoing DNA methylation gain in cancer.

(C and D) Heatmaps (C) and average plots (D) showing H3K4me3 enrichment at different groups of CpG islands in hESC, PrECs, and LNCaP cells. Gray shading in (D) highlights the reduction of H3K4me3 enrichment at the internal CpG island borders upon DNA methylation encroachment in LNCaP cells.

(E) IGV browser track depicting H3K4me1 in PrECs pre-marking internal borders of CpG islands that undergo DNA methylation encroachment in LNCaP cells.

See also Figures S5 and S6.

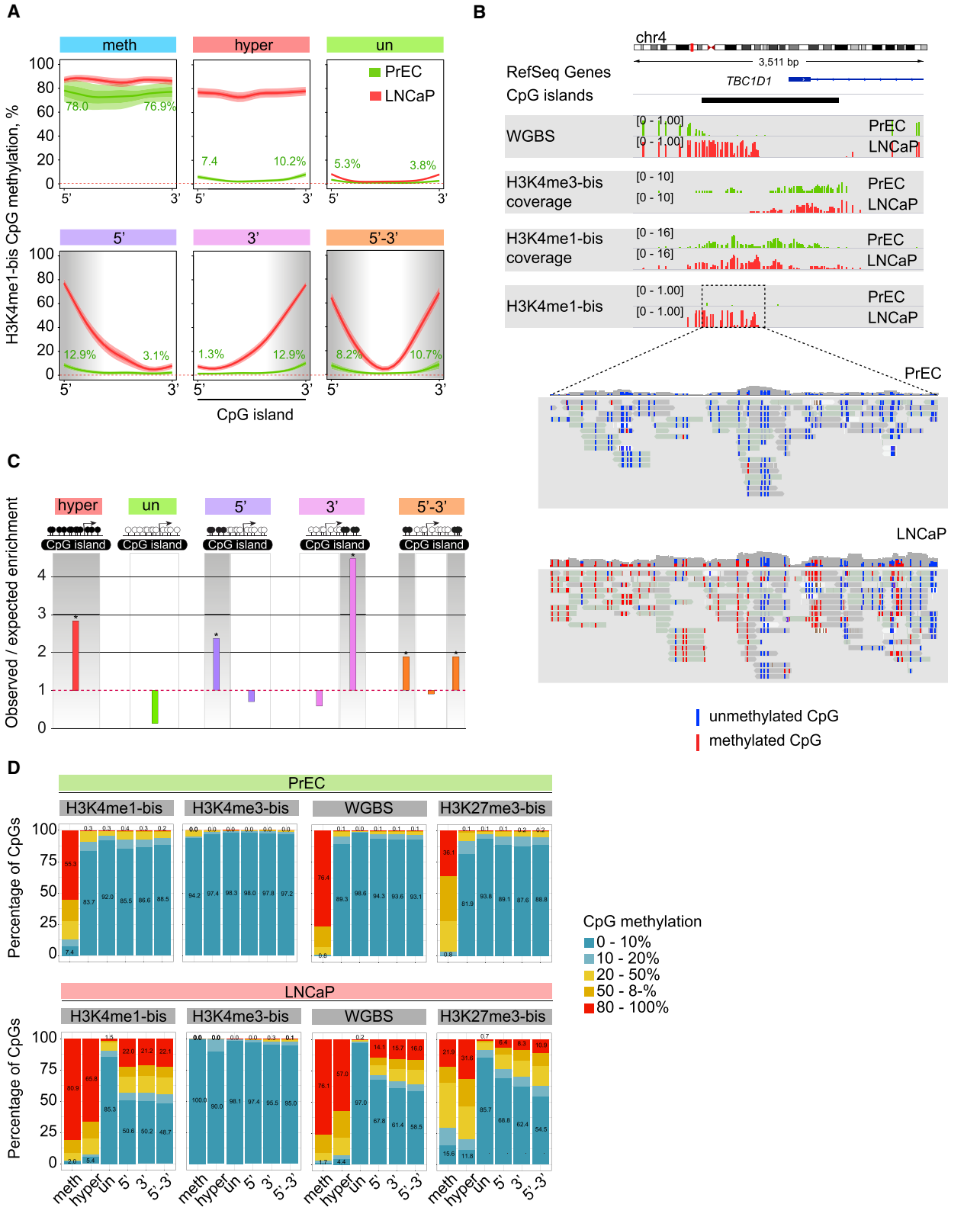


Figure 6. DNA Methylation of H3K4me1-Bound DNA Molecules

(A) Average methylation patterns in PrECs and LNCaP cells of H3K4me1-bound DNA across the different groups of CpG islands. Gray shading highlights internal CpG island borders susceptible to 5′, 3′, or 5′-3′ DNA encroachment.

(legend continued on next page)

display a greater shift toward higher methylation levels in LNCaP cells in comparison with H3K27me3-bound DNA (Figure 6D), whereas H3K4me3-marked nucleosomal DNA at CpG islands is mutually exclusive with CpG methylation, in agreement with other studies (Noh et al., 2015; Ooi et al., 2007; Weber et al., 2007).

Alterations in H3K4me1 Enrichment at Promoter CpG Islands Is Accompanied by Alterations in DNA Methylation Encroachment

We next analyzed the functional relationship of H3K4 monomethylation and DNA methylation at CpG island borders in mouse models harboring loss of function in *Kmt2c* and *Kmt2d*. We first analyzed mouse ESCs (mESCs) with either a point mutation in the catalytic SET domain (dCD) of *Kmt2c* and *Kmt2d* or loss of *Kmt2c* and *Kmt2d* (dKO) (Dorighi et al., 2017). We focused our analysis on promoter CpG-rich regions to explore the potential impact of H3K4me1 alterations on the DNA methylation at CpG island borders. Due to the fact that CpG islands are significantly shorter than the corresponding unmethylated regions in the mouse genome (Figure S7A), we employed experimentally defined clusters of unmethylated CpG sites in mESCs (here referred to as unmethylated islands [UMIs]) (Long et al., 2013). Surprisingly, we observed enrichment of H3K4me1 at UMI borders in dKOs relative to wild-type (WT) mESCs and an even stronger enrichment spanning borders and encroaching into the UMI in the dCD cells (Figures 7A and S7B). This is in contrast to its depletion from active enhancers, but in agreement with the reported H3K4me1 accumulation at poised enhancers in the mutant cells (Dorighi et al., 2017). We also found KMT2C/D enrichment was internal to UMIs in WT cells and did not change enrichment levels in dCD and dKOs (Figure S7C). An increase of H3K4me1 at UMIs upon *Kmt2c/d* deletion is therefore potentially due to compensatory activity of other H3K4 methyltransferases. Indeed, enrichment in H3K4me1 at TSSs has been reported in *Drosophila* S2 cells upon the depletion of TRR, a *Drosophila* homolog of KMT2C and KMT2D (Herz et al., 2012), and in human colorectal cancer HCT116 cells upon KMT2D depletion (Hu et al., 2013a).

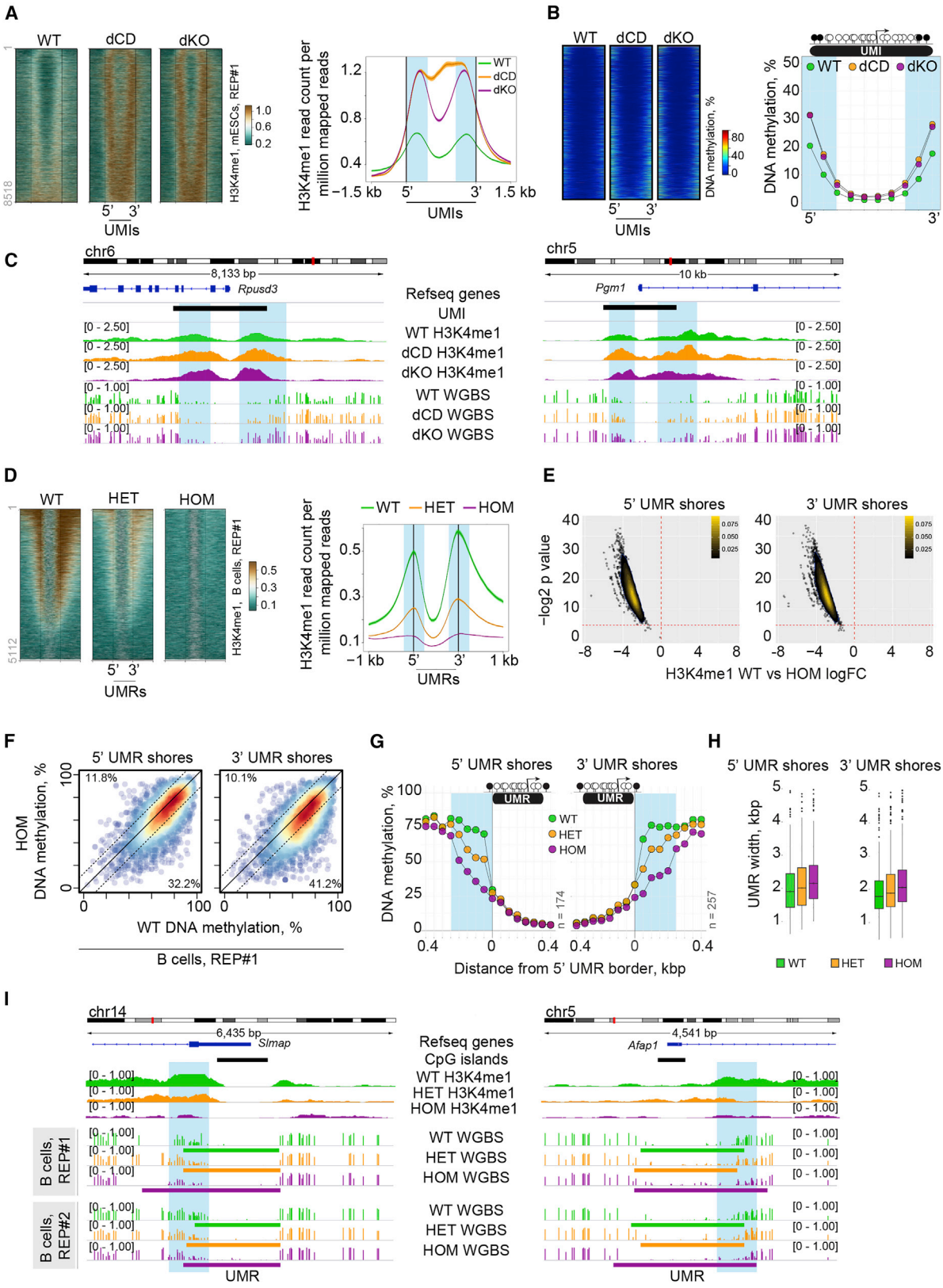
To determine whether the alterations in H3K4me1 patterns we observed at UMIs are also associated with alterations in DNA methylation in the *Kmt2c/d* mutant cells, we performed WGBS in WT, dCD, and dKO mESCs. Genome-wide we observed an overall hypermethylation in dCD and a lesser, but significant, trend toward a gain in methylation in the dKO cells (Figures S7D and S7E). More specifically, average methylation levels at UMIs were increased in both dCD and dKO (Figure S7F). We also plotted DNA methylation patterns across the promoter UMIs sorted by the H3K4me1 enrichment changes in dCD mESCs. The data revealed DNA methylation encroachment at the UMI internal borders in both dCD and dKO cells relative to the WT mESCs (Figure 7B), coinciding with the elevated levels of H3K4me1 at these regions (Figure 7A). Figure 7C shows examples of concordant spreading of H3K4me1 and DNA methylation and at the borders of the UMIs in dCD and dKO cells.

We next analyzed the relationship of H3K4 monomethylation and promoter DNA methylation in a mouse model harboring a germline loss-of-function mutation of *Kmt2d*. Here we restricted our analyses to splenic B cells of WT, heterozygous (HET), and homozygous (HOM) *Kmt2d* mutants due to the previously reported B cell phenotype (Ortega-Molina et al., 2015; Zhang et al., 2015). We observed a global reduction of H3K4me1 in HOM and HET mutants relative to WT B cells by western blot analysis (Figure S8A). To determine the genomic landscape of H3K4me1 associated with loss of function of KMT2D, we performed H3K4me1 ChIP-seq for the WT, HET, and HOM mutants. We binned the genome into 1,000-bp sliding windows and ran csaw differential enrichment analysis (Lun and Smyth, 2016), confirming a gradual reduction of H3K4me1 enrichment in HET and HOM *Kmt2d* mutants compared with WT (Figure S8B). Next, we performed WGBS on sorted B cells (in biological replicates) and total blood (B cell-depleted) from HOM, HET, and WT mice (see Table S2). We observed a trend toward genome-wide DNA hypomethylation in HET mutants, becoming more pronounced in HOM mutants (Figure S8C).

We focused our analysis on promoter undermethylated regions (UMRs, defined based on WGBS data using MethylSeekR), due to unavailability of experimental UMI profile in mouse B cells. We found that in WT cells, both up- and downstream UMR shores possess enrichment of H3K4me1, undergoing a reduction in *Kmt2d* HET and HOM mutant cells (Figures 7D and S8D), with the majority of H3K4me1-marked 5′ and 3′ UMR shores showing depletion of H3K4me1 in HOM cells (Figure 7E). We also found that H3K4me3 was reduced in the body of the UMR in HET and HOM cells (Figure S8E). Concordant with the reduction of H3K4me1 enrichment at UMR shores, depletion of DNA methylation also occurs at promoter UMR shores (Figure S8F).

We observed that the average methylation of UMR shores (predominantly downstream) is reduced in HOM mutant B cells and total blood (Figures 7F and S8G) compared with WT cells, with hypomethylation $\geq$10% encompassing up to 40% of UMR shores. Next, we focused on a subset of 5′ and 3′ UMR shores that showed the most pronounced level of hypomethylation ($\geq$40%) in HOM mutants compared with WT and interrogated their DNA methylation patterns. We expanded promoter UMR shores to 400 bp and plotted average DNA methylation per each bin, and found a gradual reduction of DNA methylation at the promoter UMR shores in HET and HOM *Kmt2d* mutant cells (Figures 7G and S8H). This coincided with statistically significant expansion of the width of the UMR in HET (~100 bp wider) and HOM (~200 bp wider) compared with the WT and an erosion of the promoter UMR borders (Figure 7H). Figure 7I

(B) IGV browser track depicting CpG methylation values and bisulfite (bis) sequencing reads of H3K4me1-bound DNA (H3K4me1-bis), and sequencing coverage (bis coverage) of H3K4me1-bis and H3K4me3-bis at the promoter CpG islands in PrECs undergoing 5′ DNA methylation encroachment in LNCaP cells.
(C) Observed over expected enrichment of CpGs that gain methylation during aging (Heyn et al., 2012) at different groups of CpG islands.
(D) Percentage of CpG sites in CpG islands with 0%–10%, 10%–20%, 20%–50%, 50%–80%, 80%–100% methylation values within H3K4me1-, H3K4me3-, or H3K27me3-bound fractions of DNA (H3K4me1-bis, H3K4me3-bis, and H3K27me3-bis, respectively) and within the total population of DNA molecules (WGBS) in PrECs and LNCaP cells.



(legend on next page)

shows example genes showing H3K4me1 reduction at the borders of promoter-associated unmethylated regions concomitant with expansion of the unmethylated regions. RNA-seq analysis in WT, HET, and HOM *Kmt2d* mutant B cells revealed virtually no changes in expression (Figure S8I), suggesting that changes to DNA methylation are not the result of gene expression changes. Overall, the data highlight a role for KMT2D and H3K4me1 in delineating DNA methylation boundaries at CpG island-associated gene promoters.

Finally, we asked whether *KMT2D* mutant prostate cancers in the TCGA cohort similarly possess lower levels of DNA methylation encroachment compared with those with the WT *KMT2D* (Table S3). We combined CpG probes overlapping 5′, 3′ and 5′-3′ CGI-DMRs and calculated the percentage of CpG probes displaying >40% hypermethylation for each case. The data revealed reduced DNA methylation encroachment at the internal CpG island borders in *KMT2D* mutant tumors (Figure S8J).

DISCUSSION

Despite decades of research, it is still unclear how the discrete patterns of DNA methylation at CpG island borders is established and maintained in normal cells and, importantly, why some islands are susceptible to cancer-associated DNA hypermethylation whereas others remain resistant. Here, we describe a mode of CpG island methylation that involves the asymmetric or symmetric spread of DNA methylation across CpG island boundaries, which results in smaller unmethylated islands and new discrete methylation island borders. Moreover, we show that the different modes of CpG island hypermethylation in cancer are distinguished by the patterns of H3K4me1 enrichment at CpG island borders in normal cells.

H3K4me1 is primarily thought to be an enhancer mark (Calo and Wysocka, 2013), although there is increasing evidence of its role at gene promoters (Cheng et al., 2014; Vavouri and Lehner, 2012). We found that the distinct pattern of H3K4me1 at promoter CpG island and UMR boundaries in hESCs and normal PrECs is a key determinant in distinguishing the modes of cancer-associated CpG island hypermethylation. Concomitant with the H3K4me1 patterns, H3K27me3 enrichment patterns in hESC also distinguish between different modes of CpG island DNA hypermethylation and DNA methylation encroachment in cancer. However, unlike H3K27me3 enrichment patterns that are substantially reduced during differentiation, the H3K4me1 enrichment pattern is retained at CpG islands or internal CpG island borders that gain DNA methylation in cancer.

H3K4me1 and H3K4me3 have opposing footprints, as H3K4me3 is substantially reduced at CpG islands prone to DNA hypermethylation but enriched at CpG islands that remain unmethylated. While the H3K4me1/H3K4me3 ratio has been shown to differentiate gene promoters and enhancers (Calo and Wysocka, 2013; Heintzman et al., 2007), we found that the H3K4me1/H3K4me3 ratio at promoter CpG islands in normal epithelial cells is highly predictive in distinguishing the different modes of aberrant DNA methylation in prostate and breast cancer, even more so than the bivalent H3K27me3/H3K4me3 mark. H3K4me1 at gene promoters has previously been shown to constrain the recruitment of H3K4me3-interacting "reader" proteins regulating the activity of corresponding genes (Cheng et al., 2014). Importantly, the balance between H3K4me1 and H3K4me3 is determined by an interplay between active removal of H3K4me3 by lysine demethylases KDM5A and KDM5B (Dahl et al., 2016; Kidder et al., 2014; Zhang et al., 2016) and the promotion of histone methylation by histone methyltransferases (Rao and Dou, 2015). Thus, the transition between H3K4me1 and H3K4me3 at gene promoters may be critical for the regulation of transcriptional activity and the predisposition for aberrant DNA hypermethylation in cancer.

We observed an accumulation of H3K4me1 at the borders of UMIs, which was accompanied by DNA methylation encroachment and reduction of the UMI in *Kmt2c/d* dCD and dKO mouse ESCs. Conversely, we found a depletion of H3K4me1 at the borders of promoters, accompanied by a reduction of DNA methylation at a subset of these regions and expansion of UMRs in mutant *Kmt2d* cells. Similar erosion of the UMR boundaries and the expansion of the unmethylated state have been described upon DNMT3A depletion (Jeong et al., 2014). It is unclear why the two *Kmt2* models confer opposing H3K4me1 states, but potentially this may be due to different compensation processes between the *in vitro* ECS cell model

Figure 7. H3K4me1 and DNA Methylation Changes in Mouse ESCs and Mouse Model Harboring Mutation or Complete Loss of KMT2C and KMT2D

(A) Heatmaps and average line plots showing enrichment of H3K4me1 at the promoter-associated UMIs in WT, dCD, and dKO mESCs. Heatmaps were sorted by the degree of DNA methylation change (low to high) in dCD cells compared with WT cells (see Figure 7B).

(B) Heatmaps and average plots showing DNA methylation profiles at promoter UMIs in WT, dCD, and dKO mESC.

(C) IGV browser track depicting UMIs undergoing DNA methylation encroachment (blue shading) in dCD and dKO mESCs and H3K4me1 accumulation.

(D) Heatmaps and average line plots showing enrichment of H3K4me1 in the B cells of WT and HET and HOM *Kmt2d* mutant mice.

(E) Volcano plot showing H3K4me1 enrichment changes (logFC) at promoter 5′ and 3′ UMR shores in HOM *Kmt2d* mutant B cells compared with the WT B cells. Each dot represents a single H3K4me1 peak overlapping with the promoter UMR shore.

(F) Scatterplot showing average promoter UMR shores methylation in WT and HOM *Kmt2d* mutant B cells. Each dot represents a single UMR shore.

(G) Average methylation patterns across promoter UMR shores in WT and HET and HOM *Kmt2d* mutant B cells. The shores shown possess ≥40% average DNA methylation reduction in HOM mutant B cells.

(H) Boxplots showing promoter UMR width in HET and HOM *Kmt2d* mutant B cells compared with WT B cells. Boxplot boundaries indicate the first and third quartiles of the data points, with the median value within the box. The whiskers extend to 1.5 IQR from the boundaries. Data points lying outside of 1.5 IQR are shown as dots.

(I) IGV browser track depicting H3K4me1 reduction in HET and HOM *Kmt2d* mutant B cells and DNA methylation reduction (blue shading) at the promoter UMR borders.

See also Figures S7 and S8; Tables S2 and S3.

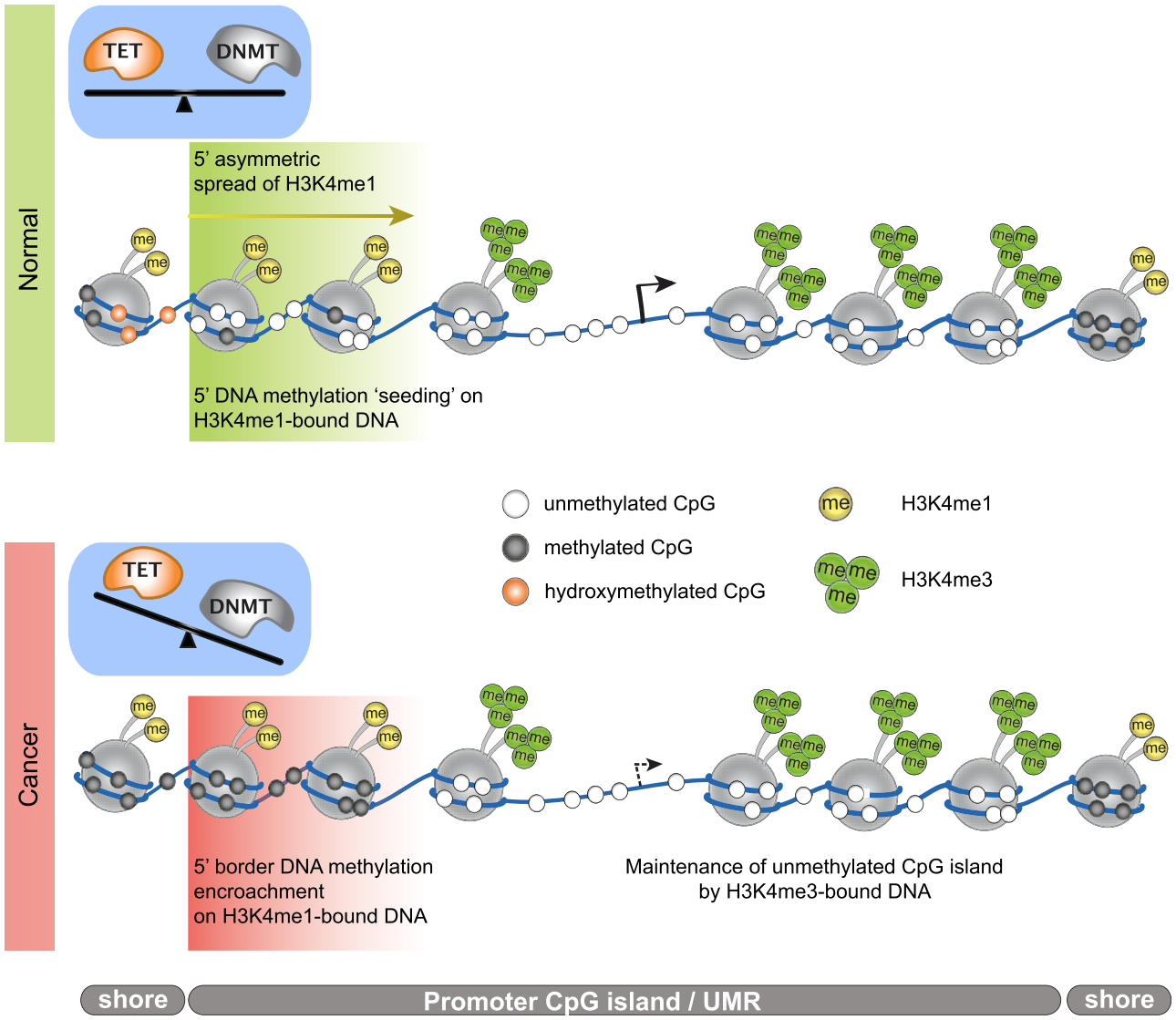


Figure 8. A Model Linking Promoter CpG Island DNA Methylation Encroachment in Cancer and H3K4me1 Distribution

The asymmetric distribution of H3K4me1 in normal cells is associated with mode of CpG island DNA methylation encroachment in cancer. The H3K4me1-bound regions are more densely packed and harbor "seeds" of DNA methylation. In normal cells, there is an enrichment of both DNMTs and TET enzymes at CpG island shores, which together help to maintain CpG islands in an unmethylated state by the opposing DNA methylation and DNA demethylation activities via 5hmC. In cancer, however, there is a reduction in TET and loss of 5hmC from the CpG island shores, which results in aberrant DNA methylation encroachment in the subset of islands where H3K4me1 has breached the CpG island interior in normal cells (shown here by 5′ asymmetric encroachment) and establishment of a new DNA methylation CpG island border in cancer and reduction in gene expression.

and the *in vivo* B cell model. It is possible that the proposed long-range co-activator function of *Kmt2c/d* (Yan et al., 2018) may also account for the observed differences in the two *Kmt2* models. Regardless, both models support a direct association between H3K4me1 levels and DNA methylation encroachment at CpG island borders.

Based on our data, we propose a model whereby CpG islands that are prone to hypermethylation encroachment in cancer are marked by the spread of H3K4me1 at the internal CpG island borders that harbor "seeds" of DNA methylation and are more nucleosome dense (Figure 8). Such cytosine methylation "seeding" may serve to recruit DNA methyl binding domain proteins and *de novo* DNA methyltransferases (Stirzaker et al., 2017), resulting in aberrant DNA methylation spread in cancer at the CpG island borders. Interestingly the CpG island border regions marked by H3K4me1 are also depleted in G4 DNA structures, in contrast to the regions that remain unmethylated. This is in agreement with the recent report that G4 DNA is involved in sequestering DNMT1 and thereby protecting CpG islands from methylation (Mao et al., 2018). We further propose that in normal cells spurious DNA methylation at CpG island borders is prevented from spreading by opposing DNA demethylation. Indeed, we find enrichment of 5hmC in normal cells at the shores of CpG islands that are prone to methylation encroachment,

concomitant with the presence of H3K4me1. The function of 5hmC as an intermediate in DNA demethylation (Bogdanovic et al., 2016; Lu et al., 2014) may play a critical role in enforcing DNA methylation boundaries of H3K4me1-marked CpG islands in normal cells. Manzo et al. (2017) found 5hmC flanking the H3K27me3-marked UMRs in ESCs. We propose that depletion of 5hmC from the CpG island borders in cancer is permissive for methylation encroachment to occur, which is further supported by studies showing that depletion of TET1 promotes methylation spreading into the CpG islands in differentiated cells (Jin et al., 2014). Similarly, TET TKO in human ESCs results in DNA hypermethylation, specifically at bivalent promoter CpG islands (Verma et al., 2018). In contrast, H3K4me3-marked internal CpG island border remains unmethylated due to the potential repulsion of the methylation machinery by H3K4me3-marked chromatin (Ooi et al., 2007; Otani et al., 2009). Additionally, reduction of DNA methylation upon DNMT KO in mESCs did not affect H3K4me3 levels, but H3K27me3 and H3K4me1 were diminished at bivalent promoters (King et al., 2016), suggesting a direct association between DNA methylation and H3K27me3 and H3K4me1 levels.

Altogether, our findings highlight a role for H3K4me1 in methylation "border security" at promoter CpG islands. Recently it has been reported that H3K4me1 facilitates recruitment of cohesion complex (Yan et al., 2018) and associated proteins include chromatin-remodeling complexes (Local et al., 2018). However, further work is required to determine whether there is also a direct instructive role for H3K4me1 in promoting DNA methylation CpG island encroachment in cancer.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - PrEC and LNCaP
  - Human Subjects
  - Mouse Embryonic Stem Cells
  - Mouse B Cells
- METHOD DETAILS
  - Whole-Genome Bisulphite Sequencing (WGBS)
  - Whole-Genome TET-Assisted Bisulphite Sequencing (TAB-seq)
  - ChIP-seq for H3K4me1, H3K4me3, H3K27me3 in PrEC and LNCaP Cells
  - Bis-ChIP-seq for H3K4me1 and H3K4me3 in PrECs and LNCaP Cells
  - Western Blot for H3K4me1 in Mouse B Cells
  - ChIP-seq for H3K4me1 in Mouse B Cells
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data Analysis
  - Definition of CpG Islands and Under-Methylated Regions (UMRs)
  - Computational Definition of CpG Island DNA Methylation Patterns
  - Observed Over Expected Enrichment
  - Gene Set Enrichment Analysis (GSEA)
  - TCGA HM450K Array DNA Methylation Data
  - ChIP-seq Enrichment Profiles
  - H3K4me1 ChIP-seq Differential Binding
  - Gene Expression and Regulatory Factors Analysis
  - Machine Learning
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures and three tables and can be found with this article online at https://doi.org/10.1016/j.ccell.2019.01.004.

ACKNOWLEDGMENTS

We thank the Wysocka Laboratory for the generous gift of wild-type and mutant mESCs (Dorighi et al., 2017) and Roger Daly for financial support of prostate cancer WGBS (Cancer Institute NSW, Australia, program grant). This work was funded by National Health and Medical Research Council, Australia (NHMRC) project grant (#1088144) (SJC); NHMRC Fellowship (#1063559) (S.J.C.); Cancer Australia, Australia, project grant (#1044458) (S.J.C.); NHMRC Program grant (#1113904) and Fellowship (#1081858) (C.C.G.); The Bill and Patricia Ritchie Foundation, Australia (C.C.G.); UNSW Sydney University International Postgraduate Award (UIPA), Australia (K.S.). The contents of the published material are solely the responsibility of the administering institution and individual authors and do not reflect the views of the NHMRC.

AUTHOR CONTRIBUTIONS

Conception and design, K.S., C.S., and S.J.C.; Data analysis, K.S., P.-L.L., C.M.G., Y.C.-S., Q.D., O.B.; WGBS, J.Z.S., W.Q., S.S.N., A.K., and G.C.S.; TAB-seq, K.S.; ChIP-seq, J.Z.S. and E.M.-F.; *Kmt2d* ENU mice experiments, E.M.-F., L.A.M., J.H.R., and C.C.G.; *Kmt2c/d* KO mESC cell culture, S.M.L. and J.M.P.; TCGA data curation, R.P. and E.Z.; WGBS clinical sample curation, R.P., T.P., J.G.K., L.H., and M.A.R.; Writing and review of manuscript, K.S., C.S., and S.J.C. All authors have read and approved the final manuscript.

DECLARATION OF INTERESTS

M.A.R. receives royalties on licensing agreements for prostate cancer diagnostics and has received research support from Janssen Pharma, Eli-Lilly, Millenium, and Sanofi-Aventis, United States. All other authors declare no competing interests.

Received: December 1, 2017
Revised: November 14, 2018
Accepted: January 7, 2019
Published: February 11, 2019

REFERENCES

Akalin, A., Franke, V., Vlahovicek, K., Mason, C.E., and Schubeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics *31*, 1127–1129.

Andrews, T.D., Whittle, B., Field, M.A., Balakishnan, B., Zhang, Y., Shao, Y., Cho, V., Kirk, M., Singh, M., Xia, Y., et al. (2012). Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. Open Biol. *2*, 120061.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jorgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nat. Cell Biol. *8*, 532–538.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., 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.

Bert, S.A., Robinson, M.D., Strbenac, D., Statham, A.L., Song, J.Z., Hulf, T., Sutherland, R.L., Coolen, M.W., Stirzaker, C., and Clark, S.J. (2013). Regional activation of the cancer genome by long-range epigenetic remodeling. Cancer Cell *23*, 9–22.

Bogdanovic, O., Smits, A.H., de la Calle Mustienes, E., Tena, J.J., Ford, E., Williams, R., Senanayake, U., Schultz, M.D., Hontelez, S., van Kruijsbergen, I., et al. (2016). Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat. Genet. *48*, 417–426.

Brinkman, A., Nik-Zainal, S., Simmer, F., Rodriguez-Gonzalez, F.G., Smid, M., Alexandrov, L.B., Butler, A., Martin, S., Davies, H., Dominik, G., et al. (2018). Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. bioRxiv. https://doi.org/10.1101/305193.

Burger, L., Gaidatzis, D., Schubeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. *41*, e155.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2017). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. *46*, D267–D275.

Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S., and Dynlacht, B.D. (2014). A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. Mol. Cell *53*, 979–992.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev. *26*, 1714–1728.

Coolen, M.W., Stirzaker, C., Song, J.Z., Statham, A.L., Kassir, Z., Moreno, C.S., Young, A.N., Varma, V., Speed, T.P., Cowley, M., et al. (2010). Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. Nat. Cell Biol. *12*, 235–U237.

Cunha, S., Lin, Y.C., Goossen, E.A., DeVette, C.I., Albertella, M.R., Thomson, S., Mulvihill, M.J., and Welm, A.L. (2014). The RON receptor tyrosine kinase promotes metastasis by triggering MBD4-dependent DNA methylation reprogramming. Cell Rep. *6*, 141–154.

Dahl, J.A., Jung, I., Aanes, H., Greggains, G.D., Manaf, A., Lerdrup, M., Li, G., Kuan, S., Li, B., Lee, A.Y., et al. (2016). Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. Nature *537*, 548–552.

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

Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N.V., Scruggs, B.S., Nady, N., Still, C.D., 2nd, Garcia, B.A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 facilitate enhancer RNA synthesis and transcription from promoters independently of H3K4 monomethylation. Mol. Cell *66*, 568–576.e4.

Du, Q., Bert, S.A., Armstrong, N.J., Caldon, C.E., Song, J.Z., Nair, S.N., Gould, C.M., Luu, P.-L., Peters, T.J., Khoury, A., et al. (2019). Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. Nat. Commun. *10*, 416.

Easwaran, H., Johnstone, S.E., Van Neste, L., Ohm, J., Mosbruger, T., Wang, Q.J., Aryee, M.J., 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.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Egelhofer, T.A., Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A.A., Cheung, M.S., Day, D.S., Gadel, S., Gorchakov, A.A., et al. (2011). An assessment of histone-modification antibody quality. Nat. Struct. Mol. Biol. *18*, 91–93.

Feng, X., Grossman, R., and Stein, L. (2011). PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics *12*, 139.

Gal-Yam, E.N., Egger, G., Iniguez, L., Holster, H., Einarsson, S., Zhang, X., Lin, J.C., Liang, G., Jones, P.A., and Tanay, A. (2008). Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc. Natl. Acad. Sci. U S A *105*, 12979–12984.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

Griffon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S., and Ballester, B. (2015). Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. *43*, e27.

Guennewig, B., Pinese, M., and Cooper, A.A. (2017). blkbox: Integration of multiple machine learning approaches to identify disease biomarkers. bioRxiv. https://doi.org/10.1101/123430.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y.T., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C.X., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Herz, H.M., Mohan, M., Garruss, A.S., Liang, K., Takahashi, Y.H., Mickey, K., Voets, O., Verrijzer, C.P., and Shilatifard, A. (2012). Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. Genes Dev. *26*, 2604–2620.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., et al. (2012). Distinct DNA methylomes of newborns and centenarians. Proc. Natl. Acad. Sci. U S A *109*, 10522–10527.

Heyn, H., Vidal, E., Ferreira, H.J., Vizoso, M., Sayols, S., Gomez, A., Moran, S., Boque-Sastre, R., Guil, S., Martinez-Cardus, A., et al. (2016). Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. *17*, 11.

Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. *22*, 246–258.

Hu, D., Gao, X., Morgan, M.A., Herz, H.M., Smith, E.R., and Shilatifard, A. (2013a). The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. Mol. Cell. Biol. *33*, 4745–4754.

Hu, D., Garruss, A.S., Gao, X., Morgan, M.A., Cook, M., Smith, E.R., and Shilatifard, A. (2013b). The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1093–1097.

Jeong, M., Sun, D.Q., Luo, M., Huang, Y., Challen, G.A., Rodriguez, B., Zhang, X.T., Chavez, L., Wang, H., Hannah, R., et al. (2014). Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat. Genet. *46*, 17–23.

Jin, S.G., Jiang, Y., Qiu, R., Rauch, T.A., Wang, Y., Schackert, G., Krex, D., Lu, Q., and Pfeifer, G.P. (2011). 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. Cancer Res. *71*, 7360–7365.

Jin, C.L., Lu, Y., Jelinek, J., Liang, S.D., Estecio, M.R.H., Barton, M.C., and Issa, J.P.J. (2014). TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. Nucleic Acids Res. *42*, 6956–6971.

Kidder, B.L., Hu, G., and Zhao, K. (2014). KDM5B focuses H3K4 methylation near promoters and enhancers during embryonic stem cell self-renewal and differentiation. Genome Biol. *15*, R32.

King, A.D., Huang, K., Rubbi, L., Liu, S., Wang, C.Y., Wang, Y., Pellegrini, M., and Fan, G. (2016). Reversible regulation of promoter and enhancer histone

landscape by DNA methylation in mouse embryonic stem cells. Cell Rep. *17*, 289–302.

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

Li, X., Liu, Y., Salz, T., Hansen, K.D., and Feinberg, A. (2016). Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. Genome Res. *26*, 1730–1741.

Lian, C.G., Xu, Y.F., Ceol, C., Wu, F.Z., Larson, A., Dresser, K., Xu, W.Q., Tan, L., Hu, Y.G., Zhan, Q., et al. (2012). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. Cell *150*, 1135–1146.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Liberzon, A. (2014). A description of the molecular signatures database (MSigDB) Web site. Methods Mol. Biol. *1150*, 153–160.

Lin, J.L., Lee, W.I., Huang, J.L., Chen, P.K., Chan, K.C., Lo, L.J., You, Y.J., Shih, Y.F., Tseng, T.Y., and Wu, M.C. (2015). Immunologic assessment and KMT2D mutation detection in Kabuki syndrome. Clin. Genet. *88*, 255–260.

Local, A., Huang, H., Albuquerque, C.P., Singh, N., Lee, A.Y., Wang, W., Wang, C., Hsia, J.E., Shiau, A.K., Ge, K., et al. (2018). Identification of H3K4me1-associated proteins at mammalian enhancers. Nat. Genet. *50*, 73–82.

Long, H.K., Sims, D., Heger, A., Blackledge, N.P., Kutter, C., Wright, M.L., Grutzner, F., Odom, D.T., Patient, R., Ponting, C.P., et al. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. Elife *2*, e00348.

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S., and Zhang, Y. (2014). Role of Tet proteins in enhancer activity and telomere elongation. Genes Dev. *28*, 2103–2119.

Lun, A.T., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. *44*, e45.

Manzo, M., Wirz, J., Ambrosi, C., Villasenor, R., Roschitzki, B., and Baubec, T. (2017). Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands. EMBO J. *36*, 3421–3434.

Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martinez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hansel-Hertsch, R., Tannahill, D., and Balasubramanian, S. (2018). DNA G-quadruplex structures mold the DNA methylome. Nat. Struct. Mol. Biol. *25*, 951–957.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Nair, S.S., Luu, P.L., Qu, W., Maddugoda, M., Huschtscha, L., Reddel, R., Chenevix-Trench, G., Toso, M., Kench, J.G., Horvath, L.G., et al. (2018). Guidelines for whole genome bisulphite sequencing of intact and FFPET DNA on the Illumina HiSeq X Ten. Epigenetics Chromatin *11*, 24.

Noh, K.M., Wang, H., Kim, H.R., Wenderski, W., Fang, F., Li, C.H., Dewell, S., Hughes, S.H., Melnick, A.M., Patel, D.J., et al. (2015). Engineering of a histone-recognition domain in Dnmt3a alters the epigenetic landscape and phenotypic features of mouse ESCs. Mol. Cell *59*, 89–103.

Ohm, J.E., McGarvey, K.M., Yu, X., Cheng, L.Z., Schuebel, K.E., Cope, L., Mohammad, H.P., Chen, W., Daniel, V.C., Yu, W., et al. (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat. Genet. *39*, 237–242.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717.

Ortega-Molina, A., Boss, I.W., Canela, A., Pan, H., Jiang, Y., Zhao, C., Jiang, M., Hu, D., Agirre, X., Niesvizky, I., et al. (2015). The histone lysine methyltransferase KMT2D sustains a gene expression program that represses B cell lymphoma development. Nat. Med. *21*, 1199–1208.

Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M., and Shirakawa, M. (2009). Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. EMBO Rep. *10*, 1235–1241.

Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S.J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. *17*, 208.

Pidsley, R., Lawrence, M.G., Zotenko, E., Niranjan, B., Statham, A., Song, J., Chabanon, R.M., Qu, W., Wang, H., Richards, M., et al. (2018). Enduring epigenetic landmarks define the cancer microenvironment. Genome Res. *28*, 625–638.

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Rao, R.C., and Dou, Y.L. (2015). Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. Nat. Rev. Cancer *15*, 334–346.

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

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

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. Bmc Genomics *15*, 284.

Shinsky, S.A., Hu, M., Vought, V.E., Ng, S.B., Bamshad, M.J., Shendure, J., and Cosgrove, M.S. (2014). A non-active-site SET domain surface crucial for the interaction of MLL1 and the RbBP5/Ash2L heterodimer within MLL family core complexes. J. Mol. Biol. *426*, 2283–2299.

Song, J.Z., Stirzaker, C., Harrison, J., Melki, J.R., and Clark, S.J. (2002). Hypermethylation trigger of the glutathione-S-transferase gene (GSTP1) in prostate cancer cells. Oncogene *21*, 1048–1061.

Song, Q., Decato, B., Hong, E.E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., and Smith, A.D. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One *8*, e81148.

Sproul, D., Nestor, C., Culley, J., Dickson, J.H., Dixon, J.M., Harrison, D.J., Meehan, R.R., Sims, A.H., and Ramsahoye, B.H. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. Proc. Natl. Acad. Sci. U S A *108*, 4364–4369.

Sproul, D., Kitchen, R.R., Nestor, C.E., Dixon, J.M., Sims, A.H., Harrison, D.J., Ramsahoye, B.H., and Meehan, R.R. (2012). Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. Genome Biol. *13*, R84.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Statham, A.L., Robinson, M.D., Song, J.Z., Coolen, M.W., Stirzaker, C., and Clark, S.J. (2012). Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. Genome Res. *22*, 1120–1127.

Stempor, P., and Ahringer, J. (2016). SeqPlots—interactive software for exploratory data analyses, pattern discovery and visualization in genomics. Wellcome Open Res. *1*, 14.

Stirzaker, C., Taberlay, P.C., Statham, A.L., and Clark, S.J. (2014). Mining cancer methylomes: prospects and challenges. Trends Genet. *30*, 75–84.

Stirzaker, C., Song, J.Z., Ng, W., Du, Q., Armstrong, N.J., Locke, W.J., Statham, A.L., French, H., Pidsley, R., Valdes-Mora, F., et al. (2017). Methyl-CpG-binding protein MBD2 plays a key role in maintenance and spread of DNA methylation at CpG islands and shores in cancer. Oncogene *36*, 1328–1338.

Taberlay, P.C., Achinger-Kawecka, J., Lun, A.T., Buske, F.A., Sabir, K., Gould, C.M., Zotenko, E., Bert, S.A., Giles, K.A., Bauer, D.C., et al. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. *26*, 719–731.

Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J., and Jones, P.A. (2014). Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. *24*, 1421–1432.

Vavouri, T., and Lehner, B. (2012). Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biol. *13*, R110.

Verma, N., Pan, H., Dore, L.C., Shukla, A., Li, Q.V., Pelham-Webb, B., Teijeiro, V., Gonzalez, F., Krivtsov, A., Chang, C.J., et al. (2018). TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. Nat. Genet. *50*, 83–95.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

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

Yan, J., Chen, S.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., Rivera, C.M., Wang, C., Ye, Z., et al. (2018). Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. Cell Res. *28*, 387.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137.

Zhang, J., Dominguez-Sola, D., Hussein, S., Lee, J.E., Holmes, A.B., Bansal, M., Vlasevska, S., Mo, T., Tang, H., Basso, K., et al. (2015). Disruption of KMT2D perturbs germinal center B cell development and promotes lymphomagenesis. Nat. Med. *21*, 1190–1198.

Zhang, B., Zheng, H., Huang, B., Li, W., Xiang, Y., Peng, X., Ming, J., Wu, X., Zhang, Y., Xu, Q., et al. (2016). Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. Nature *537*, 553–557.

Zhao, W., Hoadley, K.A., Parker, J.S., and Perou, C.M. (2016). Identification of mRNA isoform switching in breast cancer. BMC Genomics *17*, 181.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-H3K4me3	Abcam	Cat#ab8580; RRID:AB_306649
Rabbit polyclonal anti-H3K4me1	Active Motif	Cat#39297; RRID:AB_2615075
Rabbit polyclonal anti-H3K27me3	Millipore	Cat#07-449; RRID:AB_310624
Rabbit polyclonal anti-H3K4me3	Active Motif	Cat#39159; RRID:AB_2615077
Mouse anti-mouse beta-actin	Sigma-Aldrich	Cat#A2228; RRID:AB_476697
Goat anti-mouse IgG IRDye® 680RD	LI-COR®Biosciences	Cat#925-68070; RRID:AB_2651128
Goat anti-Rabbit IgG IRDye® 800RD	LI-COR® Biosciences	Cat#925-32211; RRID:AB_2651127
Biological Samples		
Clinical prostate tissue samples	This paper (Garvan/St Vincent's Hospital)	
Critical Commercial Assays		
TrueMethyl Whole Genome kit v 3.1	CEGX	Cat#CEGXTMWG
AllPrep DNA/RNA kit	Qiagen, USA	Cat#80204
EZ DNA Methylation-Gold	Zymo Research, USA	Cat#D5005
EZ DNA Methylation-Lightning Kit	Zymo Research, USA	Cat#D5030
TET-assisted bisulphite treatment kit	WiseGene, USA	Cat#K001
Illumina TruSeq Chip Library Prep Kit	Illumina	Cat#15034288
Deposited Data		
Data generated in this study have been deposited in NCBI's Gene Expression Omnibus	This paper	GSE104791
WGBS: 11 clinical prostate cancer and 3 clinical benign prostate samples	This paper	GSE104789
TAB-seq data of 3 clinical prostate cancer and 3 clinical benign prostate samples	This paper	GSE104780
WGBS data of KMT2D wild type, heterozygous and homozygous mutant mouse B cells (biological replicates #1 and #2) and mouse blood (B cell-depleted)	This paper	GSE104781
H3K4me1 ChIP-seq data of KMT2D wild type, heterozygous and homozygous mutant mouse B cells (biological replicates #1 and #2) and mouse blood (B cell-depleted)	This paper	GSE104533
WGBS data of *Kmt2c*/ *Kmt2d* WT, dCD and dKO mouse ESCs	This paper	GSE118314
Experimental Models: Organisms/Strains		
Human: PrEC cells	Cambrex Bio Science	Cat#CC-2555
Human: LNCaP cells	ATCC	Cat#CRL-1740
Mouse: Wild-type ESCs (WT) cells	Wysocka Laboratory	Dorighi et al., 2017
Mouse: *Kmt2c*/*Kmt2d* (*Mll3*/*Mll4*) (dKO) ESCs	Wysocka Laboratory	Dorighi et al., 2017
Mouse: SET domain mutants (dCD) ESC cells	Wysocka Laboratory	Dorighi et al., 2017
Mouse: $Kmt2d^{+/+}$ Wild type (WT) bulk B cells / total blood (B-cell depleted)	Australian National University	Andrews et al., 2012
Mouse: C57BL/6NCrl ENU-induced A>T point mutation, creating a KMT2D Ile5482Asn substitution; $Kmt2d^{I5430N/+}$ (HET) bulk B cells/ total blood (B-cell depleted)	Australian National University	Andrews et al., 2012
Mouse: $Kmt2d^{I5430N/I5430N}$ (HOM) bulk B cells / total blood (B-cell depleted)	Australian National University	Andrews et al., 2012

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Meth10X	Nair et al., 2018	https://github.com/luuloi/Meth10X
bowtie v.1.1.0	Langmead et al., 2009	https://github.com/BenLangmead/bowtie/
MACS2	Zhang et al., 2008	https://github.com/taoliu/MACS
MethylSeekR	Burger et al., 2013	https://bioconductor.org/packages/release/bioc/html/MethylSeekR.html
genomation	Akalin et al., 2015	https://github.com/BIMSBbioinfo/genomation
ngs.plot.r	Shen et al., 2014	https://github.com/shenlab-sinai/ngsplot
deepTools2	Ramirez et al., 2016	https://deeptools.readthedocs.io/en/develop/
SeqPlots	Stempor and Ahringer, 2016	https://github.com/Przemol/seqplots
edgeR	Robinson et al., 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
blkbox	Guennewig et al., 2017 (biorxiv)	https://cran.r-project.org/web/packages/blkbox/README.html
Rsubread	Liao et al., 2014	https://www.rdocumentation.org/packages/Rsubread/versions/1.22.2/topics/featureCounts
Other		
Histone modifications ChIP-seq data for human ESC cells was downloaded from the ENCODE project	ENCODE	(http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeBroadHistone/
DNMT3B ChIP-seq data of wild type and TET triple knockout human ESCs	Verma et al., 2018	GSE89728
WGBS data from primary breast cancer samples	Brinkman et al., 2018	https://zenodo.org/record/1217427#.W1gB1SN7GFA
WGBS from primary normal breast tissues	Lin et al., 2015; Heyn et al., 2016	E-MTAB-2014; GSE52271
WGBS from breast cancer cell line MCF7	Cunha et al., 2014	GSE52688
WGBS from the primary mammary epithelial cells HMEC	Hon et al., 2012	GSE29127
CAGE-seq from PrEC and LNCaP cells	Bert et al., 2013	GSE38685
G-quadruplex from K562	Mao et al., 2018	GSE107690
Regulatory Factor Analysis	Cheneby et al., 2017; Griffon et al., 2015	http://pedagogix-tagc.univ-mrs.fr/remap/
DNase-seq from PrEC	Du et al., 2019	GSE98732
H3K4me3 from PrEC/LNCaP	Bert et al., 2013	GSE38685
H3K4me1 from PrEC/LNCaP	Taberlay et al., 2014	GSE57498
H3K27me3 from PrEC/LNCaP	Bert et al., 2013	GSE38685
H3K27ac from PrEC	Taberlay et al., 2014	GSE57498
H3K36me3 from PrEC	Du et al., 2019	GSE98732
ChIP-seq features from HMEC	ENCODE	https://www.encodeproject.org/reference-epigenomes/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Susan Clark, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, NSW 2010. Email: s.clark@garvan.org.au; tel: +61-2-92958315

EXPERIMENTAL MODEL AND SUBJECT DETAILS

PrEC and LNCaP

Genomic DNA from normal prostate cells PrECs and prostate cancer cell line LNCaP was extracted using QIAamp DNA Mini kit (Qiagen, USA). LNCaP prostate cells were cultured as described previously (Song et al., 2002). PrECs (Cambrex Bio Science, cat

no CC-2555) were cultured according to the manufacturer's instructions in Prostate Epithelial Growth Media (PrEGM Cambrex Bio Science, cat no CC-3166) as previously described (Coolen et al., 2010).

Human Subjects

All human protocols were reviewed and approved by the St Vincent's Hospital (Sydney) Human Research Ethics Committee (HREC) (SVH File Number 12/231). Informed consent was obtained from all subjects involved in this study. Clinical prostate samples were identified from the Garvan Institute/St Vincent's Prostate Cancer biobank. Formalin-fixed, paraffin-embedded radical prostatectomy blocks were retrieved and H&E sections reviewed by a specialist prostate cancer pathologist. Five 2 mm core biopsies were taken from 1) each of the dominant nodule in areas of >90% cancer cell density, and 2) adjacent normal prostate tissue. DNA was extracted using the AllPrep DNA/RNA kit (Qiagen, USA).

Mouse Embryonic Stem Cells

Mouse wild-type mESCs (WT), *Kmt2c*/*Kmt2d* (*Mll3*/*Mll4*) mutants (dKO) and catalytic SET domain mutant (dCD) ESCs were kindly provided by the Wysocka Laboratory (Dorighi et al., 2017). Mouse ESCs were cultured as previously described (Dorighi et al., 2017). DNA was extracted using QIAamp DNA Mini kit (Qiagen, USA).

Mouse B Cells

A C57BL/6NCrl mouse with an ENU-induced A > T point mutation at chr15: 98,835,228 A > T (GRCm38), creating a KMT2D Ile5482Asn substitution, was identified by exome sequencing first generation offspring of ENU-treated mice (Andrews et al., 2012). KMT2D Ile5482 in mouse corresponds to human KMT2D Ile5431, an absolutely conserved residue in the Kabuki Interaction Surface (KIS) (Shinsky et al., 2014) of the SET domain. Mutations of the adjacent residue, R5432, have been previously identified as a germline mutation in Kabuki syndrome (Lin et al., 2015), and a somatic mutation in lymphoma (Zhang et al., 2015). Heterozygous mutant mice were propagated on the B6 background, outcrossed to BALB/c, and F1 heterozygotes intercrossed to produce homozygous mutant, heterozygous and wild-type F2 littermates for harvesting spleen B cells. All of the mice were housed in specific pathogen-free (SPF) conditions at the Australian National University (ANU) Australian Phenomics Facility (APF). All procedures performed were approved by the ANU Animal Ethics and Experimentation Committee (A2014/62). Mice were sacrificed by cervical dislocation, followed by harvest of the spleen into cold complete RPMI (cRPMI) and transport on ice to the Garvan Institute.

A single cell suspension of the spleen was obtained by gently mashing the solid tissue through a 70 μm nylon mesh filter. Bulk B cell purification was performed by cell staining based on the Miltenyi Biotec Pan-B Cell Isolation Kit Mouse II protocol, followed by manual MACS cell separation by negative depletion. The resulting B cell purity of the negative fraction was approximately 95%, as assessed by flow cytometric analysis. The cellularity of the purified B cell fraction was determined using the Scepter^{TM} 2.0 Handheld Automated Cell Counter. 2 x 10^6 cells were pelleted by centrifugation. gDNA was extracted in parallel using the AllPrep DNA/RNA Mini Kit (Qiagen, USA). The gDNA extracted was eluted in 30 μl and the gDNA concentration was determined using the Qubit dsDNA HS Assay Kit.

METHOD DETAILS

Whole-Genome Bisulphite Sequencing (WGBS)

Human PrECs and LNCaP cell lines: WGBS of PrECs and LNCaP genomic DNA was performed as previously described (Pidsley et al., 2016). Briefly, genomic DNA was spiked with 0.5% unmethylated λDNA (Promega, USA, cat no D1521) and sheared to an average size of 150-300 bp using Covaris S2 (Covaris, USA). Library preparation was performed according to the Illumina Paired-End DNA Sample Prep Kit protocol (Illumina, USA). 1 μg of sheared DNA fragments was end-repaired and adenylated prior to the ligation of Illumina adaptors following by gel size selection (260-330 bp) using Qiagen Gel extraction kit (Qiagen, USA, cat no 28704). Bisulphite treatment was performed using EZ DNA Methylation-Gold Kit (Zymo Research, USA, cat no D5005) according to the manufacturer's instructions for 4 hrs at 55°C. After desulfonation and clean-up DNA was resuspended in 50 μl H_2O. The adaptor-ligated bisulphite treated DNA was enriched by performing 10 PCR cycles in 5 independent reactions followed by their pooling and clean-up using the MinElute PCR purification kit and elution in 20 μl Qiagen EB buffer. Library quality was assessed using Agilent 2100 Bioanalyzer High-sensitivity DNA kit (Agilent, CA, USA). The DNA library was quantified using KAPA Library Quantification kit (KAPA Biosystems, USA, cat no KK4835). Paired-end 100 bp sequencing was performed on the Illumina HiSeq 2500 platform using TruSeq v3 cluster kits and SBS kits.

Clinical Samples

For library preparation for patient specimens, DNA (250 ng) from FFPE prostate tumour (n = 11) and normal prostate tissue (n = 3) was bisulphite treated using the EZ DNA methylation Gold kit (Zymo Research) following the manufacturer's protocol. Whole genome bisulphite sequencing libraries were prepared using the EpiGnome Methyl-Seq kit (Epicentre/Illumina, cat no EGMK81312) according to the manufacturer's protocol. Briefly, the single-stranded bisulphite-treated DNA was randomly primed and tagged to generate double-stranded DNA molecules with sequence tags at both ends. The tagged DNA was cleaned up using AMPure XP Beads, eluted in 22.5 μl of nuclease-free water and then enriched by performing PCR for 10 cycles in a volume of 50 μl per reaction. PCR products were cleaned up using AMPure XP Beads and eluted in 20 μl of nuclease-free water. Library quality was assessed with the Agilent 2100 Bioanalyzer using the High-sensitivity DNA kit (Agilent, CA, USA). DNA was quantified using the

KAPA Library Quantification kit by quantitative PCR (KAPA Biosystems) and sequenced using 70 bp paired end sequencing on the Illumina HiSeq 2500 platform.

Kmt2c/Kmt2d dCD, dKO ESCs; mouse *Kmt2d* WT, HET and HOM B cells: mouse genomic DNA was sheared to an average size of 800 bp using Covaris S2 (Covaris, USA). 200 ng of sheared DNA was used for the bisulphite reaction, library preparation and indexing carried out according to the manufacturer's instructions (CEGX TrueMethyl Seq, UK). Library quality was assessed with the Agilent 2100 Bioanalyzer using the High-sensitivity DNA kit (Agilent, CA, USA). DNA was quantified using the KAPA Library Quantification kit by quantitative PCR (KAPA Biosystems, cat no KK4835). Paired-end 150 bp sequencing was performed for each library on the Illumina HiSeqX platform using the HiSeq X™ Ten Reagent Kit v2.

Whole-Genome TET-Assisted Bisulphite Sequencing (TAB-seq)

TET-assisted bisulphite treatment of 3 pairs of matched normal-tumour samples was performed using the 5hmC TAB-seq Kit (WiseGene, USA, cat no K001) according to the manufacturer's instruction. Briefly, 1 μg genomic DNA was sonicated to the size of approximately 2 kbp. After sonication, DNA was spiked with 10 ng (1%) M.SssI λDNA control and 10 ng (1%) 5hmC pUC18 control DNA. To protect 5hmC, β-Glucosetransferase (GT)-based reaction was performed at 37°C for 1 hr and the DNA purified using QIAquick PCR Purification Kit (Qiagen, USA, cat no 28106) according to the protocol and eluted in 27 μl water. The eluted DNA was split into two separate reactions to ensure no more than 300 ng DNA per TET1-based oxidation reaction. The TET1 oxidation reaction was performed at 37°C for 1 hr, followed by the treatment of 1 μl of Proteinase K (20 mg/ml) at 50°C for 1 hr. The oxidised DNA was purified using QIAquick PCR purification kit (Qiagen, USA, cat no 28106) and eluted in 50 μl water. Library preparation of TET1-oxidised DNA was performed using the TrueMethyl Whole-Genome kit from Cambridge Epigenetix according to the manufacturer's instructions. 10 cycles of PCR amplification were performed. Library quality was assessed with the Agilent 2100 Bioanalyzer using the High-sensitivity DNA kit (Agilent, CA, USA). DNA was quantified using the KAPA Library Quantification kit (KAPA Biosystems, USA, cat no KK4835). Paired-end 150 bp sequencing was performed for each library on the Illumina HiSeqX platform using the HiSeq X™ Ten Reagent Kit v2.

ChIP-seq for H3K4me1, H3K4me3, H3K27me3 in PrEC and LNCaP Cells

ChIP assays were carried out according to the manufacturer's protocol (Upstate Biotechnology) as described previously (Coolen et al., 2010). Briefly, ~ 1 x 10^6 cells, in a 10 cm dish, were fixed by adding formaldehyde at a final concentration of 1% and incubating for 10 minutes at 37°C. The cells were washed twice with ice cold PBS containing protease inhibitors (1 mM phenylmethylsulfonyl fluoride (PMSF), 1 μg/ml aprotinin and 1μg/ml pepstatin A), harvested and treated with SDS lysis buffer for 10 min on ice. Resulting lysates were sonicated to shear the DNA to fragment lengths of 200 to 500 bp. Complexes were immunoprecipitated with antibodies specific for H3K4me3 (Abcam #ab8580), H3K4me1 (Active Motif, #39297), and H3K27me3 (Millipore #07-449). Antibody specificity was validated to show no cross-reactivity using a panel of modified peptides on a dot blot assay (Egelhofer et al., 2011). http://compbio.med.harvard.edu/antibodies/antibodies/74. For H3K4me1, specificity to mono-methylated lysine 4 was assessed using a peptide array with 384 unique histone modifications by the manufacturer. Binding to H3K4me2 and H3K4me3 were both less than 0.1% of the signal of H3K4me1. 10 μl of antibody was used for each immunoprecipitation. No antibody controls were also included for each ChIP assay and no precipitation was observed by quantitative Real-Time PCR (qPCR) analysis. Input samples were processed in parallel. Antibody/protein complexes were collected by either salmon sperm DNA/protein A agarose slurry or Protein A/G PLUS agarose beads (Santa Cruz, cat no sc-2003) and washed several times. Immune complexes were eluted with 1% SDS and 0.1 M $NaHCO_3$ and samples treated with proteinase K for 1 hr, DNA was purified by phenol/chloroform extraction, ethanol precipitation and resuspended in 30 μl H_2O. Libraries were prepared with the Illumina TruSeq Chip Library Prep Kit and sequenced on an Illumina HiSeq2500.

Bis-ChIP-seq for H3K4me1 and H3K4me3 in PrECs and LNCaP Cells

Chromatin immunoprecipitation was performed as described above. Two ChIP reactions were pooled together to gain sufficient amount of H3K4me1- and H3K4me3-bound DNA for the subsequent bisulphite treatment and library preparation. 60–70 ng of chromatin immunoprecipitated DNA was bisulphite treated using EZ DNA Methylation-Lightning™ Kit (Zymo Research, USA, cat no D5030). The DNA was eluted in final volume of 9 μl. Library preparation of the bisulphite-treated ChIP DNA was performed using the Illumina TruSeq DNA Methylation Kit according to the manufacturer's instructions. 10 PCR cycles were performed. The ChIP-bisulphite library was resuspended in 20 μl nuclease-free water. Libraries quality was assessed using Agilent 2100 Bioanalyzer High-sensitivity DNA kit (Agilent, CA, USA). The DNA libraries were quantified using KAPA Library Quantification kit (KAPA Biosystems, USA, cat no KK4835). Paired-end 70 bp sequencing was performed for each library on the Illumina HiSeq 2500 platform.

Western Blot for H3K4me1 in Mouse B Cells

$KMT2D^{+/+}$, $KMT2D^{I5430N/+}$ and $KMT2D^{I5430N/I5430N}$ bulk B cells were purified by MACS manual separation, as described above. The cellularity of the purified B cell fraction was determined using the Scepter™ 2.0 Handheld Automated Cell Counter. Approximately 5 x 10^6 B cells were pelleted by centrifugation, all supernatant removed and stored as pellets at -80°C. In preparation for Western Blot analysis, cell pellets were thawed on ice, and resuspended in 500 μl NP-40 buffer (ddH_2O / 150 mM NaCl / 1% IGEPAL CA-630 / 10 mM Tris-HCl pH 7.8 / 0.1% Sodium Azide) and protease inhibitors. The protein concentration of the solutions was determined using the Pierce BCA Protein Assay Kit (Thermo Scientific). Reducing/loading solution, made up of NuPAGE® LDS

Sample Buffer (4x) (Life Technologies, cat no NP0007) and 200 μM DTT diluted in PBS, was mixed with 2-5 μg protein, followed by denaturation at 95°C for 10 minutes. The lysate solutions were loaded into each well of a NuPAGE Novex 4-12% Bis-Tris Protein Gel (1.0 mm, 10-well), followed by gel electrophoresis and transfer onto a PVDF membrane. The PVDF membrane was incubated with mouse anti-mouse beta-actin (Sigma, cat no A2228-200UL, 1/10000) and rabbit anti-mouse H3K4-me1 (Active Motif, cat no 39297, 1/10000) or H3K4-me3 (Active Motif cat no 39159, 1/10000), in Odyssey® Blocking Buffer, followed by multiple wash steps, and incubation with IRDye® 680RD Goat anti-Mouse IgG (LI-COR®, cat no 925-68070, 1/20000) and IRDye® 800RD Goat anti-Rabbit IgG (LI-COR®, cat no 925-32211, 1/20000). The membrane was then imaged using the LI-COR Odyssey Clx. Fluorescence intensities were determined using Image Studio™ Lite.

ChIP-seq for H3K4me1 in Mouse B Cells

Sorted splenic B cells were fixed with 1% formaldehyde at room temperature for 10 mins followed by quenching with 0.125 M glycine (final concentration). The cells were washed twice with ice cold PBS containing protease inhibitors (1 mM phenylmethylsulfonyl fluoride (PMSF), 1 μg/ml aprotinin and 1μg/ml pepstatin A), harvested and treated with SDS lysis buffer for 10 min on ice. Resulting lysates were sonicated to shear the DNA to fragment lengths of 200 to 500 bp. Chromatin was quantitated after reversing the cross-links as described above. For each ChIP, 20-30 μg chromatin was used. 10 μl of histone H3K4me1 polyclonal antibody (Active Motif, cat no 39297, lot no 01714002). Antibody/protein complexes were collected by either salmon sperm DNA/protein A agarose slurry or Protein A/G PLUS agarose beads (Santa Cruz, USA, cat no sc-2003) and washed several times. Immune complexes were eluted with 1% SDS and 0.1 M $NaHCO_3$ and samples treated with proteinase K for 1 hr, DNA was purified by phenol/chloroform extraction, ethanol precipitation and resuspended in 30 μl H_2O. 10 ng chromatin immunoprecipitated DNA was used for each library prep using Illumina Truseq Chip Library Prep Kit (Sample Prep 48 Samples - Set A: Ref#15034288 Lot 20023472). Libraries were sequenced on an Illumina HiSeq2500.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Analysis

Data processing and alignment was performed using in-house computational pipelines. Statistical analyses were conducted in the R statistical software.

Whole Genome Bisulphite Sequencing Data

Bisulphite reads were aligned to the human (hg19) or mouse (mm10) genomes, using version 1.2 of an internally developed pipeline Meth10X (Nair et al., 2018), publicly available for download from https://github.com/luuloi/Meth10X. Sequencing metrics are specified in the Tables S1 and S2.

Whole Genome TET-Assisted Bisulphite (TAB) Sequencing Data

Sequencing reads from TAB-seq data were aligned to the human genome using version 1.0 of an internally developed pipeline, publicly available for download from https://github.com/luuloi/Meth10X. Briefly, adaptor sequences and poor-quality bases were removed using cutadapt (https://github.com/marcelm/cutadapt) in paired-end mode with default parameters. Bwa-meth version 0.10 (https://github.com/brentp/bwa-meth) was then used to align reads to hg19 using bwa version 0.7.9a (https://github.com/lh3/bwa). Sequencing metrics are specified in the Table S1. CpG sites with sequencing coverage more than 10x in both matching normal and tumour prostate tissues were used in the analysis. Proportion test was run using R prop.test function to determine the probability of the 5hmC level at a given CpG site being greater than 5%. Multiple comparisons correction was undertaken using the Benjamini-Hochberg procedure. CpG sites with adjusted p value $\leq$ 0.05 were considered significantly hydroxymethylated. For the observed over expected enrichment analysis, CpGs hydroxymethylated in normal prostate tissue (5hmC > 10%, p value $\leq$ 0.05) and CpGs with 5hmC reduction in cancer (5hmC normal – 5hmC cancer > 20%, p value $\leq$ 0.05) were used.

ChIP Sequencing Data

Sequenced reads from ChIP-seq and input control PrECs and LNCaP samples were mapped to the reference human genome (hg19) with bowtie v.1.1.0 (Langmead et al., 2009) allowing up to three mismatches. Reads mapping to multiple locations and/or deemed as PCR duplicates were filtered out. For H3K4me1 ChIP-seq peaks were called using MACS2 (Zhang et al., 2008). For H3K27me3 ChIP-seq peaks were called using PeakRanger (version 1.16) (Feng et al., 2011).

Sequenced reads from H3K4me1 ChIP-seq and input control WT, HET and HOM *Kmt2d* mutant B cells samples were mapped to the reference mouse genome (mm10) with bowtie v.1.1.0 (Langmead et al., 2009) allowing up to three mismatches.

Definition of CpG Islands and Under-Methylated Regions (UMRs)

Computationally predicted human and mouse genome CpG islands were downloaded from UCSC genome browser. To define human genome promoter-associated CpG islands, UCSC CpG islands were overlapped with the Ensembl transcription start sites (5' ends of 'Ensembl Genes') obtained from UCSC genome browser. To define mouse genome promoter-associated CpG islands, UCSC CpG islands were overlapped with the Gencode-derived transcription start sites ('GENCODE VM9 (Ensembl 84)' 5' ends) obtained from UCSC genome browser. CpG island shores were derived from CpG island coordinates by taking 500 bp flanking regions up- and downstream of the CpG island according to the direction of transcription of a corresponding gene.

To define promoter-associated UMRs we applied MethylSeekR to the WGBS data of normal human PrECs as well as WT, HET and HOM *Kmt2d* mouse B cells (Burger et al., 2013) and selected UMRs directly overlapping Ensembl transcription start sites for human

promoter UMRs and the Gencode-derived transcription start sites for mouse promoter UMRs. UMR shores in the analyses of DNA methylation in *Kmt2d* mutant B cells were derived from WT UMR coordinates by taking 200 bp-flanking regions up- and downstream of the UMR according to the direction of transcription of a corresponding gene.

Computational Definition of CpG Island DNA Methylation Patterns

CpG islands overlapping with the TSSs were split into 40 equally sized bins. For each bin across the CpG island from the 5' to 3' end as defined by the direction of transcription, average methylation of PrEC, LNCaP and Delta (LNCaP-PrEC) was calculated using ScoreMatrixList function from R genomation package. Average methylation of bins without CpG sites was imputed to zero. First, we applied k-means clustering algorithm that revealed five major groups of promoter CpG islands based on the methylation changes between PrECs and LNCaP cells including DNA methylation encroachment from the 5' or 3' CpG island border (Figure S1A). Since k-means clustering algorithm is randomised in its starting centres, we were willing to develop a non-randomised algorithm to define distinct groups of CpG islands hypermethylation patterns in LNCaP cells. To this aim, we used Pearson correlation coefficient between Delta (LNCaP-PrEC) average methylation at each bin and the corresponding bin number (Figures S1B and 1C). Thus, CpG islands with Pearson correlation of $\rho < -0.5$ and average methylation in PrEC $\leq$ 10% and Delta (LNCaP-PreC) $\geq$ 10% were considered as undergoing 5' methylation encroachment (Figures 1C and S1B, 5'). Conversely, CpG islands with Pearson correlation of $\rho > 0.5$ and average methylation in PrEC $\leq$ 10% and Delta (LNCaP-PreC) $\geq$ 10% were considered as undergoing 3' methylation encroachment (Figures 1C and S1B, 3'). To define CpG islands with bidirectional (5'-3') methylation encroachment, Pearson correlation was calculated separately for the 20 bins at the 5' half and the 20 bins at the 3' half of a CpG island - between the bin number and corresponding LNCaP CpG methylation. Thus, CpG islands with Pearson correlation of $\rho < -0.5$ for the 20 bins at 5' end and $\rho > 0.5$ for the 20 bins at 3' end together with average methylation in PrEC $\leq$ 10% and Delta (LNCaP-PreC) $\geq$ 10% were considered as undergoing bidirectional methylation encroachment (Figures 1C and S1B, 5'-3').

CpG islands with Pearson correlation of $-0.5 < \rho < 0.5$ and average methylation in PrEC and LNCaP being $\leq$ 10% were considered as remaining unmethylated (un). CpG islands with Pearson correlation of $-0.5 < \rho < 0.5$ and average methylation in PrEC $\leq$ 10% and LNCaP > 50% were considered as hypermethylated (hyper) and CpG islands with Pearson correlation of $-0.5 < \rho < 0.5$ and average methylation in PrEC and LNCaP > 50% were considered as constantly methylated (meth) in both PrEC and LNCaP cells.

To ensure that the representation of observed DNA methylation patterns is not affected by distinct CpG islands lengths (and therefore different bin length for different CpG islands), we selected the 5' end of each CpG island as an anchor, extended it 1 kbp upstream and 3 kbp downstream and plotted DNA methylation patterns. We binned these 4 kbp-wide regions into 40 equally sized bins and plotted average DNA methylation per bin as a heatmap (Figure S1C). Such representation confirmed the distinct patterns of DNA methylation in LNCaP cells including DNA methylation encroachment.

Observed Over Expected Enrichment

Genomic regions of interest and the background regions bed files were overlapped with the annotation file using bedtools annotateBed function with the fraction of each genomic region of interest and background region covered by each annotation file (observed and expected, respectively) being calculated. The probability of getting an observed/expected enrichment value was calculated using hypergeometric distribution phyper R function.

Gene Set Enrichment Analysis (GSEA)

GSEA of the genes harbouring promoter CpG islands that undergo DNA methylation encroachment or remain unmethylated in cancer was performed using GSEA_MSIGDB_v5.0 from the Broad Institute at MIT (Liberzon, 2014). All genes harbouring promoter CpG islands were used as a background. Hypergeometric distribution test was used to calculate p values. Fold enrichment over background of gene sets with FDR < 0.05 with minimum 3 genes per gene set was plotted.

TCGA HM450K Array DNA Methylation Data

Prostate Cancer Samples

Raw IDAT files and corresponding clinical and specimen data for Prostate Adenocarcinoma samples were downloaded from the TCGA Data Portal website (TCGA: http://tcga-data.nci.nih.gov/tcgafiles) on 26th May 2015. Samples and probes filtering were performed as previously described (Pidsley et al., 2018). The resulting datasets comprised 414,133 CpG sites from 392 tumour and 45 normal samples. β values were calculated from unmethylated (U) and methylated (M) signal: M/(U + M + 100) and ranged from 0 to 1 (0 to 100% methylation).

Breast Cancer Samples

Raw IDAT files for breast cancer samples were downloaded from Genomics Data Commons (GDC) portal (https://gdc-portal.nci.nih.gov/legacy-archive/search/f) in September 2016. β values were calculated from unmethylated (U) and methylated (M) signal: M/(U + M + 100) and ranged from 0 to 1 (0 to 100% methylation). DNA methylation data spanned 515 tumour and 95 normal samples.

To interrogate whether DNA methylation encroachment is a widespread feature in prostate and breast cancer cohorts, we employed TCGA HM450K array DNA methylation data and identified DNA methylation profiles at the CpG islands found to undergo DNA methylation encroachment in prostate cancer LNCaP cells. Specifically, we separated internal CpG island borders prone to DNA methylation encroachment from the internal CpG island borders resistant to DNA methylation in cancer. To do this, we split CpG islands undergoing DNA methylation encroachment into to regions: 1) overlapping with the differentially methylated regions

(CGI-DMRs) identified in LNCaP cells using MethPipe pipeline (Song et al., 2013) and 2) not overlapping with the DMR (CGI-non-DMRs). Then we separated HM450K array CpG probes overlapping CGI-DMRs from the CpG probes overlapping CGI-non-DMRs and calculated the percentage of hypermethylated (>40%) CpG probes of all CpG probes per each cancer or normal tissue sample.

To interrogate whether prostate cancer samples carrying *KMT2D* mutation possess reduced DNA methylation at the internal CpG island borders undergoing DNA methylation encroachment, TCGA prostate cancer samples carrying *KMT2D* mutations were identified using cBioPortal (Cerami et al., 2012; Gao et al., 2013). Samples carrying missense *KMT2D* mutations and gain of copy number were excluded from the analysis. Then, we calculated the percentage of hypermethylated (> 40%) CpG probes of all CpG probes overlapping internal CpG island borders undergoing 5', 3' or 5'-3' DNA methylation encroachment per each prostate cancer sample carrying *KMT2D* mutation (n = 16) or wild-type *KMT2D* (n = 376).

ChIP-seq Enrichment Profiles

Chromatin modifications enrichment profiles across genomic regions of interest were generated using ngs.plot.r (Shen et al., 2014), deeptools2 (Ramirez et al., 2016) and SeqPlots (Stempor and Ahringer, 2016). PrECs and LNCaP cell lines H3K4me3, H3K4me1, H3K27me3 ChIP-seq profiles were previously generated in our lab (Bert et al., 2013); hESC H3K4me3, H3K4me1, H3K27me3 ChIP-seq profiles were downloaded from UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone).

H3K4me1 ChIP-seq Differential Binding

Differential H3K4me1 binding between WT, HET and HOM *Kmt2d* mutant B cells was calculated using csaw package (Lun and Smyth, 2016). H3K4me1 enrichment counts were calculated per 1000 bp windows (200 bp sliding) normalised against input ChIP-seq signal with minimum of 5 reads per window. Blacklisted regions as specified by ENCODE were excluded from window read counts. Windows were filtered to retain the ones with the average abundance of $\geq$3-fold increase above the background. Significant H3K4me1 depletion in HET and HOM *Kmt2d* mutants was defined as logFC < -2 and p value $\leq$ 0.05.

Gene Expression and Regulatory Factors Analysis

Differential gene expression analysis (DGE) of polyA RNA-seq data for PrECs and LNCaP cell lines previously generated in our lab (Taberlay et al., 2016) was performed using edgeR (Robinson et al., 2010) to interrogate whether DNA methylation encroachment coincides with changes in gene expression. Genes that displayed ±2 logFC (p value $\leq$ 0.05) between LNCaP and PrEC cells were considered as significantly differentially expressed. To determine changes in gene isoforms expression, Ensembl-derived transcription start sites were overlapped with internal CpG island borders prone to DNA methylation encroachment (CGI-DMRs) and their DGE was plotted. To interrogate the patterns of the 5' ends of the messenger RNAs in PrECs and LNCaP cells we used CAGE-seq data (cap analysis gene expression), previously generated in our lab (Bert et al., 2013). To interrogate the distribution of regulatory factors binding sites across CpG islands prone to DNA methylation encroachment we used publically available dataset of 237 transcription factors (TF) (Cheneby et al., 2017; Griffon et al., 2015). To this aim, we calculated a number of TF overlapping internal CpG island borders prone to DNA methylation encroachment (CGI-DMRs) and adjacent internal CpG island borders resistant to DNA methylation encroachment (CGI-non-DMRs). Then the number of overlaps was normalised to the length of CGI-DMRs or CGI-non-DMRs, respectively.

Machine Learning

To interrogate whether CpG island DNA hypermethylation including CpG island methylation encroachment could be computationally predicted we applied seven machine learning algorithms using blkbox (Guennewig et al., 2017) R package. To this aim we separated CpG sites that belong to CpG islands prone to DNA hypermethylation in cancer from the CpG sites that belong to CpG islands resistant to the DNA methylation gain. For CpG islands prone to 5' or 3' DNA methylation encroachment, we separated CpG sites that belong to the internal CpG island borders prone to DNA methylation encroachment from the CpG sites that belong to the internal CpG island borders resistant to DNA methylation encroachment. Then, for each CpG site we calculated H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K36me3, LaminA, Lamin B logCPM (counts per million) enrichment as well as H3K4me1:H3K4me3 ratio in normal PrECs using featureCounts R function (Liao et al., 2014) from Rsubread package. For each CpG site we also calculated the distance to neighboring CpG site and evolutionary sequence conservation score (phastCons, downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/). Then seven machine learning algorithms were run using blkbox (Guennewig et al., 2017) R package using 10-fold cross-validation. A measure of feature importance was calculated for each fold and averaged across 10 folds. Random forests algorithm was identified as the best performing using Receiver Operating Characteristic (ROC) and MeanDecreaseGini index depicting feature importance was plotted for each feature.

DATA AND SOFTWARE AVAILABILITY

The data generated in this study have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE104791 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104791). WGBS data of 11 clinical prostate cancer and 3 matched normal prostate samples have been submitted under accession number

GSE104789. TAB-seq data of 3 prostate cancer and 3 matched normal prostate samples have been submitted under accession number GSE104780. WGBS and H3K4me1 ChIP-seq data of *Kmt2d* wild-type, heterozygous and homozygous mutant mouse B cells (biological replicates #1 and #2) and mouse blood (B cell-depleted) have been submitted under accession numbers GSE104781 and GSE104533, respectively. WGBS data of *Kmt2c/d* WT, dCD and dKO mouse ESCs have been submitted under accession numbers GSE118314. Histone modifications ChIP-seq data for human ESC cells was downloaded from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeBroadHistone/). DNMT3B ChIP-seq data of wild-type and TET triple knockout human ESCs was downloaded from GSE89728 (Verma et al., 2018). WGBS data from primary breast cancer samples was downloaded from https://zenodo.org/record/1217427#.W1gB1SN7GFA. WGBS from primary normal breast tissues was downloaded from E-MTAB-2014 and GSE52271. WGBS from breast cancer cell line MCF7 was downloaded from GSE52688 and re-mapped using Meth10X pipeline (see below). WGBS from the primary mammary epithelial cells HMEC was downloaded from GSE29127.