Published in final edited form as:
Nat Genet. 2018 April ; 50(4): 591–602. doi:10.1038/s41588-018-0073-4.

DNA methylation loss in late-replicating domains is linked to mitotic cell division

Wanding Zhou[1,5], **Huy Q. Dinh**[2,5], **Zachary Ramjan**[3], **Daniel J. Weisenberger**[4], **Charles M. Nicolet**[4], **Hui Shen**[1,6], **Peter W. Laird**[1,6], and **Benjamin P. Berman**[2,6]

[1]Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI

[2]Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA

[3]Van Andel Institute, Grand Rapids, MI

[4]USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA

Abstract

DNA methylation loss occurs frequently in cancer genomes, primarily within lamina-associated, late-replicating regions termed Partially Methylated Domains (PMDs). We profiled 39 diverse primary tumors and 8 matched adjacent tissues using Whole-Genome Bisulfite Sequencing (WGBS), and analyzed them alongside 343 additional human and 206 mouse WGBS datasets. We identified a local CpG sequence context associated with preferential hypomethylation in PMDs. Analysis of CpGs in this context ("Solo-WCGWs") revealed previously undetected PMD hypomethylation in almost all healthy tissue types. PMD hypomethylation increased with age, beginning during fetal development, and appeared to track the accumulation of cell divisions. In cancer, PMD hypomethylation depth correlated with somatic mutation density and cell-cycle gene expression, consistent with its reflection of mitotic history, and suggesting its application as a mitotic clock. We propose that late replication leads to lifelong progressive methylation loss, which acts as a biomarker for cellular aging and which may contribute to oncogenesis.

Loss of 5-methylcytosine in both benign and malignant neoplasms was discovered more than thirty years ago[1–4], yet the mechanisms that lead to this hypomethylation and its role in disease remain poorly understood. Genomic studies[5–9] established that hypomethylation occurs in only about half the genome, coinciding with megabase-scale domains of repressive chromatin characterized by low gene density, low GC-density, late replication timing, localization at the nuclear lamina, and Hi-C "B" domains[10,11]. These regions were termed

[6]Correspondence should be addressed to: Benjamin.Berman@csmc.edu (B.P.B), Peter.Laird@vai.org (P.W.L), Hui.Shen@vai.org (H.S.).

[5]These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

HS, PWL, and BPB conceived the study. CMN, PWL, and BPB oversaw the data generation and data quality control, with assistance from DJW. ZR automated the NGS analysis and quality control steps, and submission of data to NCI repositories. WZ, HQD, HS, and BPB performed computational analysis and produced figures. WZ, HS, PWL, and BPB wrote the manuscript, with significant contributions from HQD. HS, PWL, and BPB supervised the project.

"Partially Methylated Domains" (PMDs), and were contrasted with "Highly Methylated Domains" (HMDs) that make up the remainder of the genome[12]. PMDs have been confirmed as a common feature of most epithelial cancers[13], and other cancer types such as pediatric medulloblastoma[14].

Conflicting evidence suggests that PMD hypomethylation could provide tumors with a growth advantage or alternatively may represent only a side effect of cancer[15,16]. An understanding of the earliest origins of this process could help elucidate a potential role of PMD hypomethylation in cancer initiation, yet results in pre-cancer cell types have been conflicting. Since the 1980s, long-term cell culture has been known to result in significant DNA hypomethylation[17], which was later discovered to occur primarily in PMD domains[8,12,18,19] and to accumulate stochastically in culture[20,21]. In primary uncultured tissues, one study showed the existence of PMDs in a few highly proliferative tissues such as peripheral white blood cells and placenta, but not in slowly dividing tissues like kidney, lung, or brain[9]. Other studies have shown the presence of global hypomethylation in placenta[22] and more differentiated B cells[23] and T cells[24], but not in early stage B cells or T cells nor in myelocytes[23,24]. The largest whole-genome bisulfite sequencing (WGBS) study of normal tissues concluded that PMDs were *undetectable* in 17 of 19 human tissue types studied (34 of 37 total samples), with the only exceptions being placenta and pancreas[25]. This reinforced the prevailing view that PMD hypomethylation may be restricted to a very limited set of normal cell types, or only initiated upon exposure to environmental factors such as carcinogens[26]. Our group and one other group detected a small degree of PMD hypomethylation in normal mucosa adjacent to colon tumors[5,6], but could not rule out a pre-cancer "field effect" in these adjacent tissues.

Here, we have analyzed the largest and most diverse set of WGBS experiments to date, including new tumor and adjacent normal data from 8 common cancer types. By identifying a local sequence signature that defined the most strongly hypomethylated CpGs within PMDs, we were able to determine that most PMDs are shared by cancers and nearly all healthy human and mouse tissue types starting from fetal development. This allowed us for the first time to investigate the dynamics of hypomethylation across a large number of normal and malignant tissues, and define the relationship between PMDs, other chromatin features, and genomic mutational processes.

RESULTS

Solo-WCGW CpGs are prone to hypomethylation

We sequenced TCGA tumors and adjacent normal samples using paired-end WGBS at ~15× sequence depth, to compile a set of 40 *core tumor* samples and 9 *core normal* samples (Online Methods and Supplementary Table 1). We first defined a set of *shared PMDs* and *HMDs* across the majority of our 49 core sample set using an existing HMM-based method, MethPipe[27] (Supplementary Fig. 1a and Online Methods). Previous studies have suggested that DNA methylation is associated with local sequence context, including local CpG density[28,29] and nucleotides directly flanking the CpG[29]. We used the shared MethPipe PMD set (excluding CpG islands) to determine local CpG density and tetranucleotide sequence contexts most predictive of DNA hypomethylation.

Low CpG density within windows of +/−35 bp was optimal for predicting PMD-specific hypomethylation (Supplementary Fig. 1b). Additionally, CpGs flanked by an A:T ("W") on both sides (WCGW tetranucleotides) were consistently more prone to DNA hypomethylation than those flanked by a C:G ("S") on either (SCGW) or both (SCGS) sides (Fig. 1a, Supplementary Fig. 1c). In colon tumors and adjacent normal tissues, low CpG density and the WCGW context contributed additively to hypomethylation (Fig. 1b, upper). The most hypomethylation-prone sequence context was at CpGs with the combination of zero neighboring CpGs ("solo") and the WCGW motif. In two "adjacent normal" colon samples, only these solo-WCGW CpGs showed significant hypomethylation (Fig. 1b, upper). These same sequence dependencies were apparent in a colorectal tumor and normal colon tissue from mice (Fig. 1b, lower). They were consistent within all other tumor and adjacent normal samples in the *core* set, using either the WGBS data (Supplementary Fig. 2a) or matched Illumina Infinium HumanMethylation450 (HM450) microarray data (Supplementary Fig. 2b). An additional 390 human and 206 mouse WGBS samples examined later exhibited the same pattern (Supplementary Fig. 3a–b), with the exception of three germ cell samples (Supplementary Fig. 3c).

We focused all subsequent analyses on solo-WCGWs, representing 13% of all CpGs in the human genome. While they represent only the extreme of a hypomethylation process that affects other CpGs, focusing on solo-WCGWs alone enhanced the signal of PMD/HMD structure, especially in normal adjacent tissues and weakly hypomethylated tumors such as COAD-3518 (Fig. 1c). The relatively shallow hypomethylation in COAD-3518 could not be attributed to a greater fraction of non-cancer cells in this sample, as the cancer cell fraction in this sample was estimated (by ABSOLUTE[30]) to be 80%, compared to 51% for the more strongly hypomethylated COAD-A00R; this suggested that PMD depth was quantitative and driven by an independent property of the cancer cells.

In addition to enhancing the PMD/HMD signal in high coverage WGBS data, solo-WCGW CpGs allowed accurate PMD structure to be determined with average genomic read coverage as low as 0.05× in down-sampled bulk WGBS data (Supplementary Fig. 4a), and in low-coverage single-cell WGBS data[31] (Supplementary Fig. 4b), suggesting a possible application for low coverage or single-cell WGBS studies.

Most PMDs are shared across cancer and normal tissues

Genomic plots of solo-WCGW CpG mean methylation revealed strong concordance between PMD locations in all samples in the *core* set (Fig. 2a). Comparing the average solo-WCGW methylation of the core tumors vs the core normals in multi-scale plots (Fig. 2b) confirmed that PMDs ranging from 100 kb to 5 mb[32] were mostly overlapping between tumors and normals, but less hypomethylated in the normals.

Given the high variability of solo-WCGW PMD hypomethylation across samples (Fig 2a), we compared the standard deviation (SD) of 100-kb bins across our core normal tissues and across core tumors, finding that PMDs had higher SD than HMDs within each group (Fig. 2c). Genome-wide, SD was bimodally distributed within 100-kb bins in both normal and tumor core groups (Fig. 2d), unlike mean methylation (Supplementary Fig. 5) and all other features examined (not shown). While the highly variable nature of hypomethylation in

PMDs has been noted previously[5,7], it has not been used as a method for identifying PMDs. Using the bimodal SD peaks as a classifier resulted in a segmentation of the genome into HMDs and PMDs, with PMDs covering 63% of the genome in the core tumors (SD>0.125), and 66% of the genome in the core normals (SD>0.07). Strikingly, this simple method resulted in 100-kb bin classifications that were 83% concordant between the normal and tumor groups (Fig. 2d). These PMDs covered 95% of the base pairs in PMDs previously reported in colorectal cancer[6], and 93% of PMDs in the IMR90 fibroblast cell line[12] (Supplementary Fig. 6). This SD-based classification of PMDs allowed us to rescale methylation values for individual samples based on their sample-specific degree of PMD hypomethylation (Fig. 2e–f), further illustrating the high degree of concordance in PMD/HMD structure across tumor and normal samples.

Most PMDs are shared across developmental lineages

We investigated solo-WCGW PMD structure by combining our TCGA dataset with 343 previously published human and 206 mouse WGBS samples (Supplementary Table 1), examining solo-WCGW methylation averages with human samples arranged into 6 groups (Fig. 3) and mouse samples into 4 groups (Fig. 4). As in the *core* set, the overall degree of hypomethylation varied widely, but PMD structure was largely shared for 5 of the 6 categories. Common PMDs overlapped lamina-associated regions (LADs)[33] and late replicating domains, as expected (Fig. 3a and Fig. 4, bottom). The germline and embryo (GE) category was the only exception, with only some samples sharing PMDs (Fig. 3a, Group GE, Fig. 4, Group GE). Immortalized cell lines (cancer and non-cancer), with the exception of pluripotent embryonic cells, generally showed strongly hypomethylated PMDs that were shared with other groups (Fig. 3a, Group CL, Fig. 4, Group ESC). More discussion on methylation maintenance in embryonic and induced pluripotent stem cells is given in the Supplementary Note and Supplementary Fig. 7a.

In agreement with the TCGA tumor-adjacent “normals”, most disease-free post-natal tissues showed PMD structure shared with tumors and other groups (Fig. 3a, Group PN and Fig. 4, Group PN). The normal human samples from Schultz *et al.*[25] made up the majority of non-brain samples in our PN group and clearly had shared PMDs in our solo-WCGW analysis, while the original analysis of Schultz *et al.* identified PMDs in only 3 of these 37 samples. Most brain samples in the PN group were from a different study[34], and these stood out as the one post-natal tissue type without clearly detectable PMDs in our analysis, possibly attributable to *de novo* DNA methylation in post-mitotic brain cells[34]. Tissue types with high stem cell turnover[35] including liver, colon, skin, and pancreas displayed the strongest PMD hypomethylation.

All nucleated blood cell types showed shared PMD structure, in contrast to an earlier analysis of many of the same WGBS datasets[41] that found PMD hypomethylation to be limited to the lymphoid lineage (Fig. 3a, Group PB). Both B cells and T cells could generally be divided into subgroups of strong vs. weak hypomethylation. Those subtypes having undergone antigen presentation and activation (e.g., memory B/T cells, regulatory T cells, germinal center B cells, and plasma cells) fell into the strongly hypomethylated class, while naïve B and T cells fell into the weakly hypomethylated class, consistent with earlier reports showing that B and T cell hypomethylation increased during maturation[23,24]. However, unlike these earlier reports, our solo-WCGW analysis showed that PMD hypomethylation was already clearly evident by the naïve stage (Figure 3a and Supplementary Fig. 7b). Lymphocyte activation involves clonal expansion (proliferation of individual B/T cells to produce large numbers of daughter cells with the same antigen specificity)[36], and the dramatic hypomethylation that occurs after activation strengthens the notion that methylation loss accumulates during successive rounds of cell division (as shown explicitly in long term cultures[21]). Our solo-WCGW analysis provided the first demonstration that PMDs occur across all cell types of the myeloid lineage and are largely shared with other cell types (Figure 3a and Supplementary Fig. 7c).

The tumor group (TM) consisted of 50 solid tumors (largely made up of the 40 *core* tumors shown previously), plus 50 hematopoietic malignancies (Fig. 3a, Group TM). Interestingly, while hematopoietic tumors had more strongly hypomethylated PMDs than normal hematopoietic samples, they generally followed the trend established by their developmental origin: those derived from myeloid cells (AML) had shallower PMDs than those derived from lymphoid cells (CLL, MCL, TPLL, MM) (one-way Wilcoxon test, p=9.69e-7). The notable exception among lymphoid-derived tumors was ALL, which had hypomethylation levels similar to normal lymphoid cells. The lower degree of hypomethylation in ALL (derived from childhood cases) may reflect the generally lower degree of hypomethylation in cells from younger individuals, a topic investigated below.

For five of the six cell type groups (excluding group “GE”), mean methylation across samples in the group (Fig. 3b), as well as SD (Fig. 3c–d), revealed largely shared PMD structure. SD was bimodally distributed across the genome in all five groups (Fig. 3e), and could thus be used to define PMD regions. For all of the five sample groups, the majority of PMDs defined by high-SD bins were substantially overlapping PMDs defined earlier from the core tumor group (Fig. 3e and Supplementary Fig. 8). For example, 82% of high-SD bins were overlapping between the post-natal non-blood group (PN) and the core tumor group, and 84% were overlapping between the post-natal blood group (PB) and the core tumor group. Our findings reinforce the idea that a large set of cell-type-invariant PMDs dominate the hypomethylation landscape in most tissues.

PMD hypomethylation emerges during embryonic development

The presence of PMD hypomethylation in multiple fetal tissue types led us to further investigate solo-WCGW methylation in gametes and early developmental stages (Fig. 5a–c). Human sperm was highly methylated, with little discernable PMD structure aside from the peri-centromeric region (Fig. 5a, Group I), while mouse methylomes displayed consistent PMD structures throughout spermatogenesis (Supplementary Fig. 9). Human germinal vesicle oocytes had deep PMD hypomethylation (Fig. 5a, Group II), although a subset of PMD boundaries appeared to differ from somatic tissues. The rapid and global demethylation that occurs within the Inner Cell Mass (ICM) is thought to be an active process, attributable to a different mechanism than PMD-associated hypomethylation[37]. Interestingly, while ICM and blastocyst samples were strongly de-methylated, they did retain weak PMDs with boundaries resembling those of oocytes rather than those of later somatic cell types (Fig. 5a, Group III). Primordial germ cells (PGCs), which are set aside from the soma soon after implantation, showed an even more extreme erasure of DNA methylation than blastocysts, precluding any discernable PMD structure (Fig. 5a, Group IV).

Embryonic somatic tissues (Fig, 5a, Group V) were rapidly re-methylated genome-wide, and PMD structure could not be readily resolved, in contrast to more mature fetal samples (Fig. 5a, Group VI). Tissues sampled at different developmental stages revealed a progressive emergence of PMD/HMD structure along organismal development (Fig. 5c). This analysis revealed a substantial degree of similarity between PMD structure in brain tissues and PMD structure in other lineages, something that was not apparent from genomic plots. The substantial similarity of PMD structure detected between ICMs, ESCs, embryonic (<8 weeks) stages, and post-natal samples, suggests that PMD hypomethylation may begin at the earliest stages of development. This interpretation is strengthened by the observation that the degree of hypomethylation observed at the fetal and postnatal stages for each cell type largely mirror the lineage-specific hypomethylation rate within the same embryonic cell type.

PMD hypomethylation is associated with chronological age

To investigate the link between PMD-associated hypomethylation and cumulative numbers of cell divisions, we tested whether solo-WCGW methylation level within common PMDs was associated with donor age in different primary cell types. A strong age association was evident from the WGBS profile of sorted CD4+ T cells from a newborn vs. those from a 103-year-old individual, with the latter being closer to a T cell derived leukemia than to the newborn sample (Fig. 6a). To investigate age-related properties within larger studies only performed using the HM450 platform, we used the common PMDs derived from all WGBS samples to define a standard set of solo-WCGW PMD probes represented on HM450 (Online Methods). In these larger studies, PBMC samples from newborns had significantly less PMD hypomethylation than those from elderly donors (Fig. 6b left), and fetal liver samples had significantly less PMD hypomethylation than adult liver samples (Fig. 6b, right). Strikingly, fetal tissues from four different developmental lineages showed nearly linear accumulation of hypomethylation from 9 weeks post-gestation to 22 weeks post-gestation (Fig. 6c). Despite small sample sizes, this was statistically significant for 3 of the 4 fetal tissue types. A similar association was observed between PMD hypomethylation and gestational age in multiple mouse fetal tissue types (Supplementary Fig. 10).

An earlier study used the HM450 platform to investigate the effects of environmental (UV) exposure on PMD hypomethylation in human skin samples[26]. While the earlier study described PMD hypomethylation as only occurring within the sun-exposed samples of the epidermal layer, our re-analysis of solo-WCGWs revealed that both dermal and epidermal cells exhibited age-associated PMD hypomethylation without sun exposure, but that this process was dramatically accelerated specifically in epidermal cells upon sun exposure (Fig. 6d). This suggests that while PMD hypomethylation is a nearly universal process in aging, the degree of hypomethylation is a reflection of the complete mitotic history of the cell, including proliferation associated with normal development and tissue maintenance, plus additional cell turnover occurring as a consequence of environmental insults.

HM450 datasets showed that diverse hematopoietic cell types had a significant association between donor age and degree of hypomethylation, with the myeloid lineage (Fig. 6e) having a much slower rate of age-associated loss compared to the lymphoid lineage (Fig. 6f). This finding is consistent with the overall lower degree of methylation observed in myeloid cell types from WGBS data. While the rate of loss within the myeloid lineage was extremely low, the association to donor age was highly significant within the large human monocyte dataset (Fig. 6e). This finding contradicts an earlier analysis based on many of the same samples, which found that monocytes lacked PMD hypomethylation and age-associated hypomethylation[24].

PMD hypomethylation is linked to mitotic cell division in cancer

We studied the landscape of cancer hypomethylation in 9,072 tumors from 33 cancer types included in TCGA, using the HM450 solo-WCGWs located within common PMDs (Fig. 7a). PMD hypomethylation was nearly universal but showed extensive variation both within and across cancer types. Comparison to 749 adjacent normals from TCGA showed that the relative degree of hypomethylation across cancer types was correlated with that of the disease-free tissue of origin (Supplementary Fig. 11–13). This association was reduced in cancer types for which the normal adjacent specimens contained low fractions of relevant cell types representing putative cells of origin for the tumor.

Somatic mutation events are known to display mitotic clock-like properties[38]. Within TCGA tumors, we found that higher genome-wide somatic mutation densities were significantly associated with deeper PMD hypomethylation, suggesting that mitotic turnover may underlie both somatic mutation and PMD hypomethylation (Fig. 7b). This association was consistent using different purity thresholds (Supplementary Fig. 13c), indicating that it was not the result of confounding due to differential detection sensitivity related to purity.

PMD hypomethylation was also associated with somatic copy number aberration density (Supplementary Fig. 13d). Activation and insertion of LINE-1 endogenous retro-transposable elements is a common event in human cancer and can induce structural alterations, copy number alterations, and induction of oncogenes[39–41]. Using somatic LINE-1 insertions identified from Whole Genome Sequencing (WGS) of TCGA tumors[41], we found that LINE-1 insertion breakpoints were preferentially enriched in PMD regions (Fig. 7c), in agreement with an earlier study[39]. Intriguingly, tumors with deeper PMD hypomethylation had more LINE-1 insertions in 8 of 9 cancer types, with the only exception being endometrial cancer (Fig. 7d, Supplementary Fig. 14). While the mechanisms controlling LINE-1 insertion density in cancer are not well understood, they may be stochastically linked to the number of cell divisions (like SNVs), and/or require de-repression of “hot” LINE-1 elements, a process which may be linked to DNA hypomethylation[42,43].

We reasoned that tumors highly proliferative at the time of specimen collection may also reflect an extensive history of past cell division. Using TCGA samples with matched gene expression data, we identified the 60 genes most strongly associated with PMD hypomethylation, finding that these genes were most enriched in Gene Ontology functional terms associated with proliferation and mitotic cell division (Fig. 7e). In further support of this link between ongoing cell proliferation and PMD hypomethylation, the genes with the greatest association to PMD hypomethylation were strongly enriched within a list of 350 cell-cycle dependent genes from Cyclebase[44] (Fig. 7f). Ranking tumor samples by their degree of PMD hypomethylation showed that this association involved most cell-cycle dependent genes across different mitotic stages (Fig. 7g). Remarkably, proliferative tumors had deep PMD hypomethylation despite having higher levels of both *DNMT1* and *DNMT3A/B,* which are expressed as part of a general DNA replication program (Supplementary Note). The most hypomethylated tumors also had high expression of *UHRF1* (a contributor to DNMT1 methylation maintenance activity), underscoring that PMD hypomethylation accumulates despite strong expression of the DNA methylation maintenance machinery. We also investigated whether overexpression of TET genes, which participate in active DNA demethylation, might contribute to PMD hypomethylation. None of the three TET genes were highest in the tumors with strongly hypomethylated PMDs, indicating that TET enzymes are not responsible for DNA methylation loss in PMD regions (in contrast to promoters and CpG islands, where extensive evidence exists for TET-mediated demethylation). All of our tumor mutation and expression results suggest cumulative mitotic cell divisions as the major driving force behind PMD hypomethylation accumulation.

Replication timing and H3K36me3 both affect methylation

We used the one cell type with publicly available data for all relevant histone and topological marks, IMR90, to systematically analyze our solo-WCGW based PMD definition. This analysis confirmed previous findings[6,7] that HMD/PMD structure coincided with nuclear architecture, as characterized by Hi-C A/B compartments, Lamin B1 distribution and replication timing (Fig. 8a). At the single CpG scale, Solo-WCGW CpG methylation was most strongly correlated with replication timing, followed by the histone mark H3K36me3 (Supplementary Fig. 15a). The *de novo* methyltransferase DNMT3B has recently been shown to be guided to transcribed gene bodies via a direct interaction with the H3K36 methylation mark[45]. Active genes marked by H3K36me3 are overwhelmingly located in early replicating regions, and it has been suggested that both active transcription of gene bodies and early replication timing contribute to differential methylation throughout the genome[9]. To disentangle the contributions of H3K36me3 and replication timing to genome-wide DNA methylation levels and PMDs, we performed a stratified analysis of all solo-WCGW CpGs in the genome (Fig. 8b–c). We found that the 14% of Solo-WCGWs overlapping H3K36me3 were highly methylated, irrespective of position relative to gene annotations or replication timing (Fig. 8b, left). The remaining 86% of Solo-WCGWs (those not overlapping an H3K36me3 peak) had lower methylation across all contexts, but were strongly replication-timing dependent (Fig. 8b, right). In IMR90 cells, the degree of methylation maintenance associated with early replication timing was even greater than the degree associated with H3K36me3 (Fig. 8b, right). The relative contribution of replication timing vs. H3K36me3 was reversed in the H1 (hESC) cell line (Fig. 8c), a cell type with exceptionally high DNMT3A/B activity that makes them one of the few cell types able to survive loss of Dnmt1 function[46,47]. Because most somatic cell types had detectably hypomethylated PMDs like IMR90 (and unlike H1), our observations support a model in which highly effective methylation maintenance at H3K36me3-marked regions is achieved through a process mediated by the direct recruitment of DNMT3B through its PWWP domain[45]. Consistent with earlier observations[9], this H3K36me3-linked maintenance appears to act independently from the effect of replication timing on PMD methylation loss (Fig. 8d).

DISCUSSION

In this study, we identified four distinct features influencing DNA methylation levels in large portions of the human and mouse genomes: *First*, the local sequence context of the CpG dinucleotide; *second*, the timing of DNA replication; *third*, the presence of the H3K36me3 histone mark; and *fourth*, the accumulated number of cell divisions. The sequence context, replication timing, and H3K36me3 marks each confer differential susceptibility to replication-associated DNA methylation loss, and thus collectively shape PMD/HMD structure, while the degree of PMD hypomethylation is a function of the cumulative number of cell divisions from the earliest stages of embryonic development.

We showed that two local sequence features (CpG density and the WCGW sequence context) exert a strong influence on the rate of DNA methylation loss at individual CpGs within PMDs, and that these influences are consistent across cell types and species. The bulk of DNA methylation maintenance is performed by DNMT1 and augmented by DNMT3A/B[48]. DNMT1 has been shown to act processively, with increased efficiency in the presence of multiple CpG sites in close proximity[49], a feature consistent with the poorer methylation maintenance of “solo” CpGs (Fig. 8e). *In vitro* biochemical studies have yielded conflicting findings regarding the role of the immediate CpG flanking positions on DNMT1 activity, with one study suggesting higher affinity for G/C rich flanking sequences[50], and another suggesting higher affinity for A/T rich sequences[51]. The *in vivo* effects of the WCGW motif described here on methylation maintenance efficiency should be followed up with careful mechanistic studies to identify the causative factor or factors. The discovery of the Solo-WCGW signature largely allowed for our improved analysis of HMD/PMD structure, which may lead to better characterization of not just the “common PMDs” studied here but also important classes of cell-type-specific PMDs[6,7,14,52] (Supplementary Note).

Most Solo-WCGW were not marked by H3K36me3, and we identified replication timing as the major determinant for methylation levels at these H3K36me3-negative CpGs. We propose that replication late in S phase provides the cell with less time for re-methylation of newly synthesized daughter strands during DNA replication (Fig. 8f). This *re-methylation window model* is supported by a recent study that reconstructed methylation gains and losses at individual CpGs upon clonal expansions of individual somatic cells in culture[21], showing that progressive methylation loss was most pronounced at late-replicating domains. Further strengthening the re-methylation window model, biochemical studies have shown that re-methylation during mitosis is in fact relatively slow and not fully completed until after the S-G2 checkpoint[53,54]. Therefore, re-methylation efficiency is likely dependent on the time window between daughter strand synthesis and the beginning of M-phase. This is consistent with the mitotic clock-like PMD methylation loss we observe specifically within late-replicating regions (Fig. 8f).

The presence of H3K36me3 appeared to override this late-replication associated methylation loss at Solo-WCGW CpGs (Fig. 8d). Genetic evidence suggests that maintenance of DNA methylation at H3K36me3-marked CpGs is mediated by the direct recruitment of DNMT3B to H3K36me3-marked nucleosomes[45,55]. The independent contributions of replication timing and H3K36me3 are consistent with earlier findings based on actively transcribed gene bodies[9], and help to resolve the long-standing paradox concerning positive associations between actively transcribed gene bodies and DNA methylation[56]. This would also explain why head and neck squamous cell carcinomas with *NSD1* mutations, which exhibit significant reductions in H3K36me2 and H3K36me3 levels[57], have substantial loss of DNA methylation in the HMD compartment (Supplementary Fig. 15b). It is important to note that the two major genomic contexts we found to contribute to hypomethylation, are strongly associated with specific nuclear territories (Fig. 8g). As the heterochromatin likely represents a distinct compartment separated by a physical boundary, we cannot rule out other compositional differences of this compartment contributing to the less efficient DNA methylation maintenance observed there.

A number of studies have identified specific CpGs predictive of chronological age[58–60] as well as gestation age at birth[61]. *These signatures are largely non-overlapping with PMDs, as shown in earlier work*[26] *and with the PMD solo-WCGWs identified here*. We believe this is because PMD hypomethylation captures underlying mitotic dynamics, which are only loosely associated with chronological age *per se*. Organismal aging and the associated physiological changes affect transcriptional regulation of various genes and pathways, and many or most of the loci identified on the basis of age alone[58–60] likely represent transcriptionally-coupled chromatin changes at these genes (for example, changes to Somatostatin which regulated growth hormone[58]). As we have shown, PMD hypomethylation is likely a more direct clock-like readout of mitotic age, which is generally correlated with chronological age but can be accelerated by environmental factors or processes that promote cell turnover, such as cellular damage, wounding, inflammation, etc.

DNA hypomethylation has long been proposed to allow the aberrant expression and transposition of retroelements that can play a role in cancer by inducing chromosomal aberrations at the point of insertion[62–66]. Genetically engineered *Dnmt1* hypomorphism in mouse was shown to cause lymphomas frequently harboring retrotranspon-induced Notch1 activation events[43]. Whole-genome sequencing has shown that approximately 50% of human tumors contain somatic retrotranspositions of LINE-1 elements, and that these often lead to structural alterations[39,40,67,68] enriched within PMDs[39]. In one study, human lung tumors exhibiting mobilization of LINE-1 elements shared a common DNA hypomethylation signature[42]. Across a large TCGA cohort, we showed that tumors with higher degrees of PMD hypomethylation are more likely to have LINE-1 insertions (Fig. 7c–d). This evidence is correlative in nature, and it is certainly possible that LINE-1 activity is caused by a methylation-independent event. However, our new results are consistent with the genetic models cited above, and it is tempting to hypothesize that LINE-1 activity could be accelerated by PMD hypomethylation.

The methylation loss process described here affects a sizeable fraction of all CpGs in the genome, and thus could exert a significant influence on methylation-dependent mutational processes, most importantly CpG to TpG substitutions driven by methylation-dependent deamination of CpGs. This mutational signature accounts for a large fraction of single nucleotide mutations observed in both evolution and cancer, and thus we might expect systematic DNA methylation changes to influence the rate of these mutations. A simplistic model would predict that hypomethylated solo-WCGWs within late replicating PMDs would be protected from deamination and thus have a lower CpG to TpG mutation rate. Indeed, we observed some evidence in support of this model for both somatic mutations (from tumor sequencing) and *de novo* mutations in the human germline (from whole-genome trio sequencing). This analysis, described in detail in Supplementary Fig. 16 and the Supplementary Note, does not take into account other important factors such as the rate of transcription-coupled DNA repair, and should be confirmed in future studies.

URLs

Roadmap Epigenomics data is downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/roadmapepigenomics/.

BLUEPRINT epigenome project data is downloaded from ftp://ftp.ebi.ac.uk/pub/databases/blueprint/

ENCODE data project is downloaded from www.encodeproject.org

The Bis-SNP easy run procedure is detailed at http://people.csail.mit.edu/dnaase/bissnp2011/stepByStep.html

Our entire customized work flow ECWorkflows is hosted and freely available at https://github.com/uec/ECWorkflows.

Picard tools was downloaded from http://broadinstitute.github.io/picard

ONLINE METHODS

Whole Genome Bisulfite Sequencing

Cases for the WGBS assay was selected from 8 of the most common cancer types (Lung squamous cell carcinoma, Lung adenocarcinoma, Breast, Colorectal, Endometrial, Stomach, Bladder, Glioblastoma). For at least one tumor from each cancer type, we also sequenced its adjacent histologically normal tissue; for the rest, only the tumor was profiled. We combined these samples with one tumor and matched normal colon cancer pair from our earlier study[6], yielding a *core* set of 40 well characterized tumors and 9 adjacent normal samples (Supplementary Table 1). These tumors and normal samples are referred to as *core tumors* and *core normals* in the text. Paired-End WGBS-PE protocol was adapted from earlier protocols developed by our group[6]. Briefly, sample genomic DNA (2 μg) is sonicated using a Diagenode Bioruptor and will be size-selected to a range of 400–500bp. Sodium bisulfite conversion of all DNA samples is performed using the EZ DNA Methylation Kit (Zymo Research). All libraries are quality controlled by Agilent Bioanalyzer examination and quantified using the Kapa Biosystems kit. Cluster generation and paired-end sequencing are performed according to Illumina guidelines for the HiSeq 2000, utilizing the latest version reagents and software updates.

External Data

The external human WGBS data consists of 19 germ cells and pre-implantation embryonic tissues, 13 post-implantation embryonic and fetal tissues, 37 cell lines, 59 non-blood normal primary tissues (including normal adjacent tissues of tumors as well as disease-free samples), 154 blood or blood component samples, 11 solid tumors and 50 blood malignancies (Supplementary Table 1). The 206 mouse WGBS data sets are constituted by 13 ES cells, 17 germ cells and embryonic tissues, 123 primary fetal tissues and 53 primary postnatal normal samples. Human postnatal normals were retrieved from Roadmap Epigenomics Project (see URLs). Sorted blood WGBS and blood malignancies were downloaded from the BLUEPRINT epigenome project (see URLs). Mouse fetal WGBS samples were downloaded from the ENCODE project (see URLs). Other postnatal and fetal WGBS samples were downloaded from MethBase[27]. For MethBase samples, we only included data sets that passed the Q/C standard of the database. We list the relevant citations and sources of the WGBS data sets used in this work in Supplementary Table 1. HM450 datasets and the corresponding meta-information used for age association were obtained from Gene Expression Omnibus by downloading the following datasets: GSE30870, GSE35069, GSE56046, GSE59065, GSE51954, GSE61278, GSE56515. Mutation prevalence for TCGA tumor samples were obtained from the Broad Institute TCGA Genome Data Analysis Center (2016): MutSigCV v0.9 cross-sample somatic mutation rate estimates (Jan 28th 2016 release). Tumors that have POLE or APOBEC family mutations, or classified as with microsatellite instability, were annotated to be hypermutator tumors. When hypermutator samples were excluded, samples without annotation were also excluded. Numbers of somatic LINE-1 insertions in 1-mb bins were downloaded from an earlier report[41].

Alignment and Extraction of Methyl-cytosine Levels

Reads were aligned to the genome (build GRCh37) using BSmap[71] under the following parameters "-p 27 -s 16 -v 10 -q 2 -A AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT". We marked duplicated reads using Picard tools (see URLs, version 1.38). DNA methylation rates and SNP information were called using Bis-SNP[72], using the default easy-run procedure (see URLs). Bis-SNP allows for distinguishment of C->T mutation from bisulfite conversion by investigating the complementary strand. CpGs with fewer than 10 reads' coverage are excluded from analysis.

Genomic binning

To show megabase-scale HMD/PMD structures, we chose a 100-kb window size so that the segments would contain a sufficient number of solo-WCGWs to give reliable methylation averages (Supplementary Fig. 17 and Supplementary Note), without losing resolution to detect the majority of PMD positions, which fall within PMDs of 500 kb or greater[6].

Definition of Preliminary PMD/HMD Domains Based on All CpGs

We used WGBS at ~15× coverage to profile methylation patterns of 40 tumors (39 new TCGA samples and one from our prior study[6]) from 8 of the most common cancer types, and tumors were selected on the basis of high cancer cell content (Supplementary Table 1). For one case from each of the 8 cancer types, we profiled both the tumor and adjacent normal tissue; for the rest, only the tumor was profiled. Most of our tumor samples had a high degree of hypomethylation, so we first used an existing HMM-based tool, MethPipe[27] using a window size setting of 10 kb, to identify PMDs in each sample individually (Supplementary Fig. 1a). While the fraction of the genome covered by PMDs in different samples differed by two to three folds (Supplementary Fig. 1b), there was sufficient overlap to define a *shared MethPipe PMD* set of 417 PMDs (covering 13% of the genome) that was shared among at least 21 of the 30 tumors. As a comparison group, we defined a *shared MethPipe HMD* (highly methylated domain) set that was not covered by PMDs in any tumor sample, and included 830 regions (covering 32% of the genome).

Final Definition of PMDs/HMDs Based on Standard Deviation of solo-WCGW Methylation

Every 100-kb bins are dichotomized into PMD/HMD using a Gaussian mixture model (implemented in the R package mixtools) based on cross-sample SD of beta values from our core tumor samples (N=40). The Gaussian mixture model assumes two subpopulations of 100-kb bins---those located in PMDs with higher cross-sample SDs and those located in HMDs with lower cross-sample SDs. The final threshold of cross-sample SD for classifying PMDs from HMDs is determined to be 0.125. The more conservative sets of “common PMDs” and “common HMDs” are defined by the criteria that SD>0.15 and SD<0.10 respectively. Overlap of PMD boundaries of two samples were measured in the percentage of 100-kb bins identified as both in PMDs and in HMDs in the two samples respectively. The mouse PMDs/HMDs were defined in the same way using 32 postnatal non-brain WGBS samples (Supplementary Table 1). The SD threshold for classifying PMDs from HMDs in mouse is determined to be 0.09.

HM450 Analysis

For TCGA HM450 data sets, raw IDATs were preprocessed by first applying background subtraction[73] and then linear dye-bias correction matching the signal intensities of the two detection channels. Probe signals with detection p-value <0.05, as well as probes overlapping common SNPs and putative repetitive elements which cause potential cross-hybridization were then masked[74]. For external data sets where raw IDATs were unavailable we use processed beta values downloaded from GEO. Based on our WGBS analysis, we classified HM450 probes according to the number of neighboring CpGs and the tetranucleotide sequence context. Only probes targeting solo-WCGW CpGs are retained. We also removed probes falling into annotated CpG Islands or is unmethylated (beta < 0.2) in at least 20 of the 749 matched normal tissue samples included in TCGA. This results in 6,214 probes in common PMDs and 9,040 probes in common HMDs. Four letter acronyms for cancer types were taken following the official TCGA nomenclature. We used the difference of methylation between the mean methylation of solo-WCGW probes located in common PMDs and those in common HMDs to measure the degree of PMD-associated DNA hypomethylation in each sample. This method avoids confounding in the case of cancer types derived from globally demethylated cell types such as primordial germ cells (Supplementary Fig. 12–13).

Analysis of The IMR90 Epigenome

Features are clustered using $1 - |\rho|$ as distance where ρ is the Spearman's correlation coefficient. Centromeres are excluded from IMR90 analysis. IMR90 epigenome data was downloaded from the ENCODE project data center (accessions listed in Supplementary Table 1). Wavelet-transformed signals for replication timing were downloaded from GEO (GSM923447)[75]. Histone mark signal was quantified using percentage of base overlaps of each window with gapped peaks downloaded from the Roadmap Epigenome Consortium. Gene bodies were extracted from GENCODE transcript annotation version 26. Base overlap was used as the gene body signal. RNA-seq signal is log2 transformed number of reads overlapping with each window using bedtools[76]. Only the protein-coding gene annotation from the HAVANA team was used for genic analysis in Fig. 8d. Intergenic regions exclude all transcript annotation from all sources. Solo-WCGW CpGs LaminB1 ChIP and HiC data were downloaded from GEO under the accession GSE53331 and GSE35156 respectively.

Rescaling Based on PMD Methylation

We calculated the distribution of methylation values within common PMD 100-kb bins. We then trimmed the top and bottom 20% of this distribution for each sample setting low values to 0 and high values to 1, and linearly rescaled all values between 20% and 80% to the range [0,1] (Fig. 2e). The same genomic region of chr16p is visualized in Fig. 2f.

Stratified Analysis of Solo-WCGW CpGs in The Genome

The Solo-WCGW CpGs were first classified (Fig. 8b–c) by their overlap with H3K36me3 into H3K36me3-positive (left) and H3K36me3-negative (right) categories, then by relative position to gene structures and placement in one of the four replication timing bins quartiles (colors, with threshold ≤ 40, (40,60],(60,75], >75.for IMR90 Repli-Seq and ≤ −0.5, (−0.5,0.4], (0.4,1.15],>1.15 for H1 Repli-ChIP). For Solo-WCGWs residing within +/− 10 kb of an annotated gene, metagene plots (Fig. 8b–c) are used to show average methylation levels across all genes in relation to the Transcription Start Site (TSS) and the Transcription Termination Site (TTS). For all other Solo-WCGWs (intergenic), we showed the distribution of methylation values together for each replication timing group as a single violin plot.

Statistics

Except for when described explicitly in the text, P-values for two-group comparison were calculated using one-tailed Wilcoxon's Rank Sum test. Correlation coefficients were computed with Spearman's method, with the exact P-values calculated in R using algorithm AS 89, otherwise via asymptotic t-approximation when exact computation was not feasible.

Data availability

The WGBS data is available in Genome Data Commons (GDC) under the TCGA project with IDs and file names listed in Supplementary Table 1.

Code availability

Our customized work flow for preprocessing WGBS sequencing data is freely accessible (see URLs).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Toshinori Hinoue and Dr. Houtan Noushmehr for help selecting TCGA WGBS samples based on analysis of other TCGA data types, and Dr. Helen Goodridge and Dr. De-Chen Lin for useful discussions and comments. We thank The Cancer Genome Atlas Research program office, especially Dr. Kenna Shaw for helping to get the WGBS project up and running. We also thank the members of the TCGA Research Network, along with the dozens of other research groups that generated the published datasets that were used here to gain novel insights. This project was supported by the Van Andel Research Institute, and the Cedars-Sinai Center for Bioinformatics and Functional Genomics and Samuel Oschin Comprehensive Cancer Institute, and the University of Southern California USC Epigenome Center. The work was funded by the following grants: National Institutes of Health/ National Cancer Institute grants U24 CA143882 (PWL, BPB, HQD, and HS); R01 CA170550 (PWL); U01 CA184826 (BPB); U24 CA 210969 (PWL, BPB and HS) and National Institutes of Health/National Human Genome Research Institute grant R01 HG006705 (BPB).

REFERENCES FOR MAIN TEXT

1. Ehrlich M, Wang RY. 5-Methylcytosine in eukaryotic DNA. Science. 1981; 212:1350–7. [PubMed: 6262918]
2. Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983; 301:89–92. [PubMed: 6185846]
3. Gama-sosa MA, et al. The 5-methykytosine content of DNA from human tumors. Nucleic Acids Res. 1983; 11:6883–6894. [PubMed: 6314264]
4. Goelz S, Vogelstein B, Feinberg A. Hypomethylation of DNA from benign and malignant human colon neoplasms. Science (80-). 1985; 228:187–190.
5. Hansen KD, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011; 43:768–775. [PubMed: 21706001]
6. Berman BP, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2012; 44:40–46.
7. Fortin JP, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. 2015; 16:180. [PubMed: 26316348]
8. Weber M, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005; 37:853–62. [PubMed: 16007088]
9. Aran D, Toperoff G, Rosenberg M, Hellman A. Replication timing-related and gene body-specific methylation of active human genes. Hum Mol Genet. 2011; 20:670–680. [PubMed: 21112978]
10. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. 2013; 20:274–281. [PubMed: 23463312]
11. Quante T, Bird A. Do short, frequent DNA sequence motifs mould the epigenome? Nat Rev Mol Cell Biol. 2016; 17:257–62. [PubMed: 26837845]
12. Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]
13. Timp W, et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med. 2014; 6:61. [PubMed: 25191524]
14. Hovestadt V, et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature. 2014; 510:537–541. [PubMed: 24847876]

15. Baylin S, Bestor TH. Altered methylation patterns in cancer cell genomes: Cause or consequence? Cancer Cell. 2002; 1:299–305. [PubMed: 12086841]
16. Brennan K, Flanagan JM. Is there a link between genome-wide hypomethylation in blood and cancer risk? Cancer Prev Res (Phila). 2012; 5:1345–57. [PubMed: 23135621]
17. Ehrlich M, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res. 1982; 10:2709–21. [PubMed: 7079182]
18. Lister R, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011; 471:68–73. [PubMed: 21289626]
19. Hansen KD, et al. Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. Genome Res. 2014; 24:177–184. [PubMed: 24068705]
20. Landan G, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet. 2012; 44:1207–1214. [PubMed: 23064413]
21. Shipony Z, et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature. 2014; 513:115–119. [PubMed: 25043040]
22. Schroeder DI, et al. The human placenta methylome. Proc Natl Acad Sci U S A. 2013; 110:6037–42. [PubMed: 23530188]
23. Kulis M, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nat Genet. 2015; 47:746–56. [PubMed: 26053498]
24. Durek P, et al. Epigenomic Profiling of Human CD4(+) T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. Immunity. 2016; 45:1148–1161. [PubMed: 27851915]
25. Schultz MD, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015; 523:212–6. [PubMed: 26030523]
26. Vandiver AR, et al. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. Genome Biol. 2015; 16:80. [PubMed: 25886480]
27. Song Q, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One. 2013; 8:e81148. [PubMed: 24324667]
28. Edwards JR, et al. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res. 2010; 20:972–80. [PubMed: 20488932]
29. Gaidatzis D, et al. DNA Sequence Explains Seemingly Disordered Methylation Levels in Partially Methylated Domains of Mammalian Genomes. PLoS Genet. 2014; 10
30. Carter SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012; 30:413–421. [PubMed: 22544022]
31. Farlik M, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell. 2016; 19:808–822. [PubMed: 27867036]
32. Knijnenburg, Ta, et al. Multiscale representation of genomic signals. Nat Methods. 2014; 11:689–94. [PubMed: 24727652]
33. Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453:948–51. [PubMed: 18463634]
34. Lister R, et al. Global Epigenomic Reconfiguration During Mammalian Brain Development. Science. 2013; 341:629–643.
35. Tomasetti C, Vogelstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science (80-). 2015; 347:78–81.
36. Burnet FM. A modification of Jerne’s theory of antibody production using the concept of clonal selection. CA Cancer J Clin. 1976; 26:119–21. [PubMed: 816431]
37. Wu H, Zhang Y. Reversing DNA methylation: Mechanisms, genomics, and biological functions. Cell. 2014; 156:45–68. [PubMed: 24439369]
38. Alexandrov LB, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015; 47:1402–7. [PubMed: 26551669]
39. Lee E, et al. Landscape of Somatic Retrotransposition in Human Cancers. Science (80-). 2012; 337:967–971.
40. Tubio JMC, et al. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science (80-). 2014; 345:1251343–1251343.

41. Rodriguez-Martin B, et al. Pan-cancer analysis of whole genomes reveals driver rearrangements promoted by LINE-1 retrotransposition in human tumours. bioRxiv. 2017; 179705doi: 10.1101/179705

42. Iskow RC, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010; 141:1253–1261. [PubMed: 20603005]

43. Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene. 2008; 27:404–8. [PubMed: 17621273]

44. Santos A, Wernersson R, Jensen LJ. Cyclebase 3.0: A multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res. 2015; 43:D1140–D1144. [PubMed: 25378319]

45. Baubec T, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015; 520:243–7. [PubMed: 25607372]

46. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992; 69:915–26. [PubMed: 1606615]

47. Li Z, et al. Distinct roles of DNMT1-dependent and DNMT1-independent methylation patterns in the genome of mouse embryonic stem cells. Genome Biol. 2015; 16:115. [PubMed: 26032981]

48. Jones, Pa, Liang, G. Rethinking how DNA methylation patterns are maintained. Nat Rev Genet. 2009; 10:805–811. [PubMed: 19789556]

49. Hermann A, Goyal R, Jeltsch A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. J Biol Chem. 2004; 279:48350–9. [PubMed: 15339928]

50. Flynn J, Azzam R, Reich N. DNA binding discrimination of the murine DNA cytosine-C5 methyltransferase. J Mol Biol. 1998; 279:101–16. [PubMed: 9636703]

51. Bashtrykov P, Ragozin S, Jeltsch A. Mechanistic details of the DNA recognition by the Dnmt1 DNA methyltransferase. FEBS Lett. 2012; 586:1821–1823. [PubMed: 22641038]

52. Johann PD, et al. Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. Cancer Cell. 2016; 29:379–393. [PubMed: 26923874]

53. Liang G, et al. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol Cell Biol. 2002; 22:480–91. [PubMed: 11756544]

54. Schermelleh L, et al. Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation. Nucleic Acids Res. 2007; 35:4301–12. [PubMed: 17576694]

55. Neri F, et al. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017; 543:72–77. [PubMed: 28225755]

56. Jones PA. The DNA methylation paradox. Trends Genet. 1999; 15:34–7. [PubMed: 10087932]

57. Papillon-Cavanagh S, et al. Impaired H3K36 methylation defines a subset of head and neck squamous cell carcinomas. Nat Genet. 2017; 49:180–185. [PubMed: 28067913]

58. Hannum G, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol Cell. 2013; 49:359–367. [PubMed: 23177740]

59. Horvath S. DNA methylation age of human tissues and cell types. Genome biol. 2013; 14:R115. [PubMed: 24138928]

60. Slieker RC, et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. Genome Biol. 2016; 17:191. [PubMed: 27654999]

61. Knight AK, et al. An epigenetic clock for gestational age at birth based on blood methylation data. Genome Biol. 2016; 17:206. [PubMed: 27717399]

62. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet. 1998; 20:116–7. [PubMed: 9771701]

63. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature. 2004; 431:96–99. [PubMed: 15318244]

64. Trinh BN, Long TI, Nickel AE, Shibata D, Laird PW. DNA methyltransferase deficiency modifies cancer susceptibility in mice lacking DNA mismatch repair. Mol Cell Biol. 2002; 22:2906–17. [PubMed: 11940649]

65. Eden A. Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. Science (80-). 2003; 300:455–455.

66. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009; 1:239–259. [PubMed: 20495664]

67. Solyom S, et al. Pathogenic orphan transduction created by a nonreference LINE-1 retrotransposon. Hum Mutat. 2012; 33:369–371. [PubMed: 22095564]

68. Helman E, et al. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res. 2014; 24:1053–63. [PubMed: 24823667]

69. Amendola M, van Steensel B. Nuclear lamins are not required for lamina-associated domain organization in mouse embryonic stem cells. EMBO Rep. 2015; 16:610–7. [PubMed: 25784758]

70. Hiratani I, et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res. 2010; 20:155–69. [PubMed: 19952138]

71. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009; 10:232. [PubMed: 19635165]

72. Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 2012; 13:R61. [PubMed: 22784381]

73. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013; 41

74. Zhou W, Laird PWPW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017; 45:e22. [PubMed: 27924034]

75. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010; 107:139–44. [PubMed: 19966280]

76. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

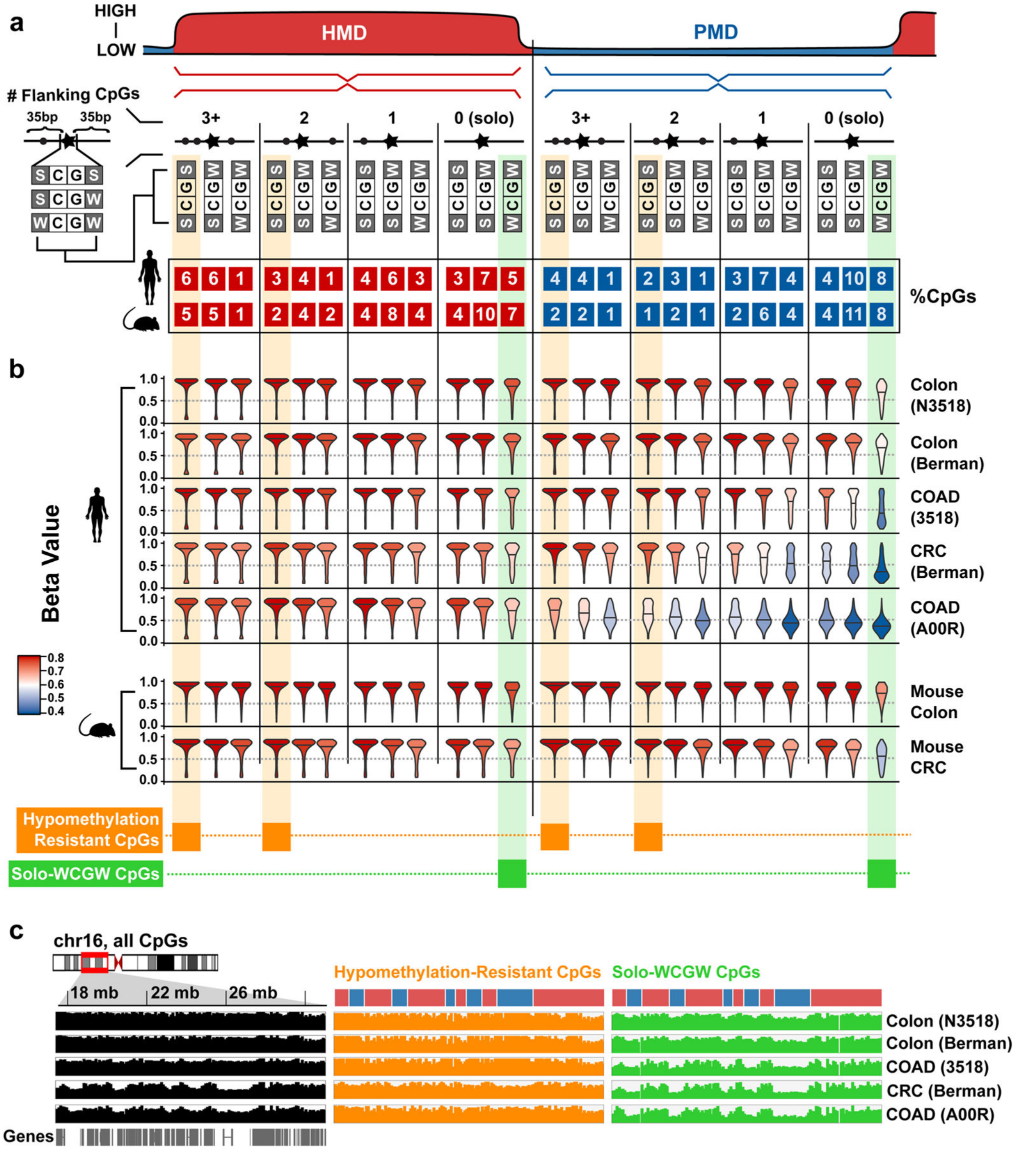

Figure 1. Solo-WCGW CpGs are prone to hypomethylation

(**a**) Each genomic CpG dinucleotide was placed into one of four CpG density categories (0, 1, 2, or 3+, depending on the number of additional CpGs within a +/− 35 bp window), and one of the three flanking nucleotide categories (SCGS, SCGW and WCGW, with “S” being C:G and “W” being A:T). Because CpGs are palindromic, WCGS and SCGW were combined. Each of the 4×3=12 possible contexts are shown as columns for CpGs within common HMDs (left) or common PMDs (right). In the illustrations, a star indicates the target CpGs, and solid circles indicate all neighboring CpGs within the window. The number of CpGs in each context is shown as a percentage of all genomic CpGs; for instance, the first column shows that 6% of all CpGs in the human genome are within HMDs, have 3+ flanking CpGs, and SCGS tetranucleotide context. (**b**) Violin plots show beta value distributions for CpGs in each context, for five human tissues (two normal colon tissues and three colon tumors) and two mouse tissues (one normal colon tissue and one colon tumor).

Violin color indicates mean beta value. Columns shaded orange and green indicate the most hypomethylation-resistant and most hypomethylation-prone categories, respectively. (**c**) Average methylation values (non-overlapping 100-kb bins) across a 12-mb section of chr16p, for the human colon samples. Values were calculated using all CpGs (left), only hypomethylation resistant CpGs (orange, middle), or only Solo-WCGW CpGs (green, right). CpG islands were removed in all analyses.

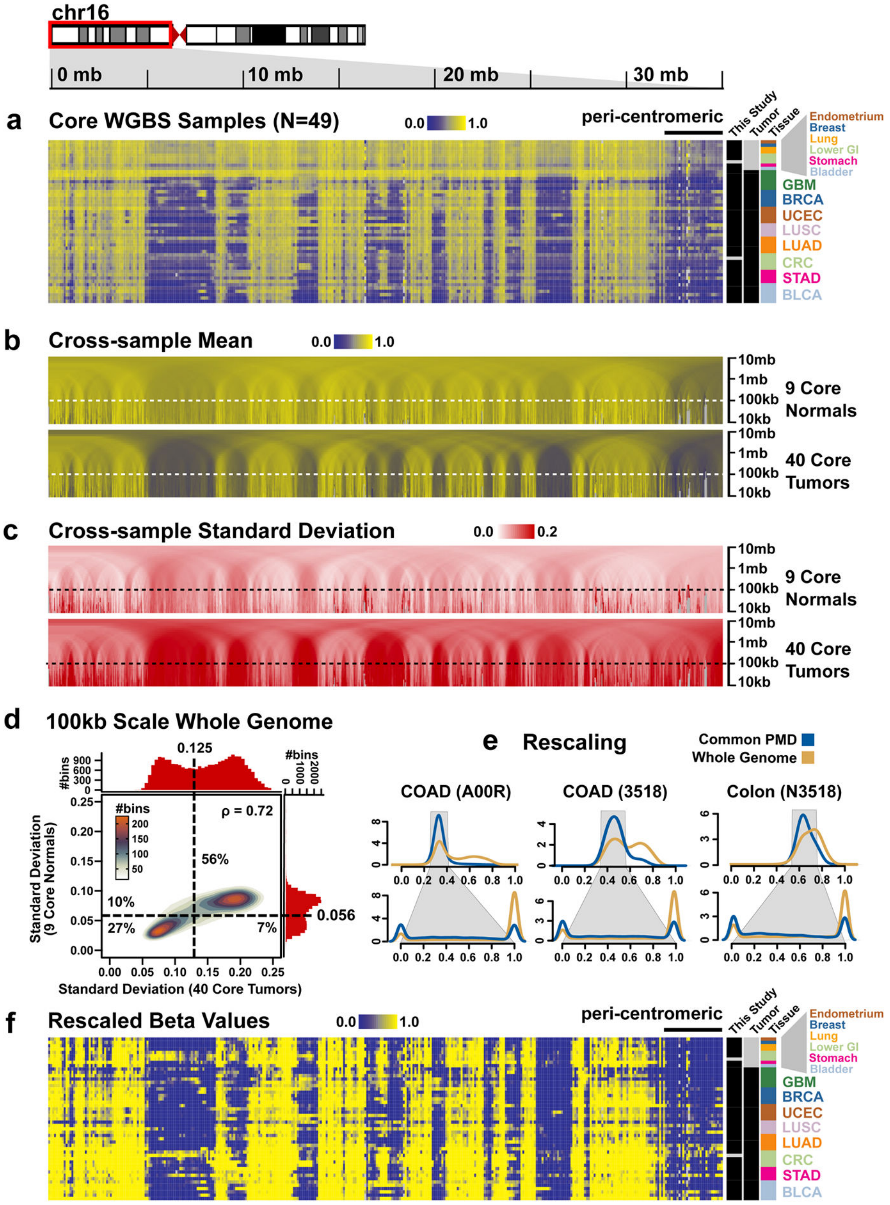

Figure 2. Most PMDs are shared across cancer and normal tissues

(**a**) Average methylation values (non-overlapping 100-kb bins) for chr16p, shown for the *core* tumor/normal dataset. The “tumor” field indicates tumors (black) vs. adjacent normals, and “this study” field indicates samples that were newly sequenced as part of this study (black). Within both normal and tumor classes, tissue types are grouped and ordered by average methylation level of samples from the group. For instance, “endometrium” is the first normal group because it has the highest methylation among normal groups, and likewise for “GBM” among tumor groups. (**b**) Average methylation across all normal (upper) or tumor samples (lower), calculated for multiple window sizes from 10 kb to 10 mb (“multi-scale plot”). (**c**) SD across all normal or tumor samples as multi-scale plots. (**d**) 100-kb SD values for the all non-overlapping genomic bins, plotted for tumors (red histogram, X-axis) vs. normals (blue histogram, Y-axis). Bimodal peaks for each were identified via a

Gaussian mixture model, and cutoffs dividing low and high SD values are indicated by dashed lines for each axis. A scatter cloud shows the correlation between SD values between the tumors and normals, indicating the percentage of 100-kb bins falling into each of the four quadrants as well as Spearman's ρ. (**e**) Illustration of method used to rescale each sample's methylation values based on genome-wide levels within a common set of PMDs (Online Methods). (**f**) Same data as panel (a), but using rescaled methylation values.

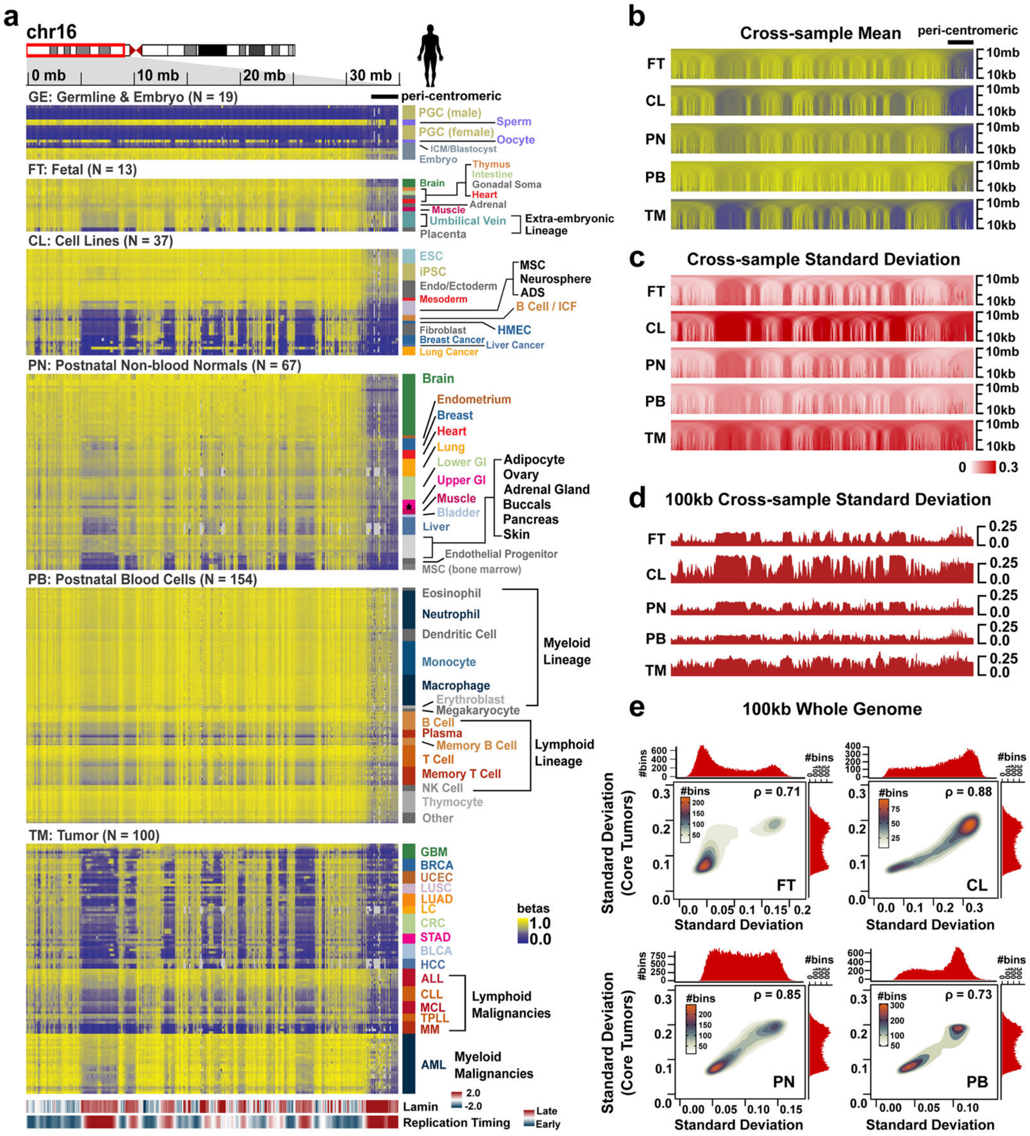

Figure 3. Most PMDs are shared across developmental lineages

(**a**) Average solo-WCGW methylation levels were plotted along chromosome 16p for 390 WGBS samples, organized into 6 groups: Germline and preimplantation embryo (GE). Post-implantation embryonic/fetal samples (FT), grouped first by embryonic vs. extra-embryonic, then by average methylation. Cell lines (CL). Post-natal non-blood normal tissue samples (PN). Post-natal blood-derived samples (PB). Primary tumors (TM). Within each of the 6 groups, samples were organized by cell type (labeled with color codes). Lamin B1 signal and replication timing of IMR90 lung fibroblast are shown below methylation heatmaps (bottom). (**b**) Mean methylation levels within each of the 5 major groups (excluding group GE), plotted as in Fig. 2b. (**c**) SD within each of the 5 major groups, plotted as in Fig. 2c. (**d**) SDs shown for the 100-kb scale alone. (**e**) Distribution of SD for all non-overlapping 100-kb genomic bins across all samples of the core tumor group (from panel (d)) are plotted on the Y-axis, compared to each of four major groups (FT, CL, PN, and PB), shown on the X-axis. Group GE is omitted due to lack of PMD structure.

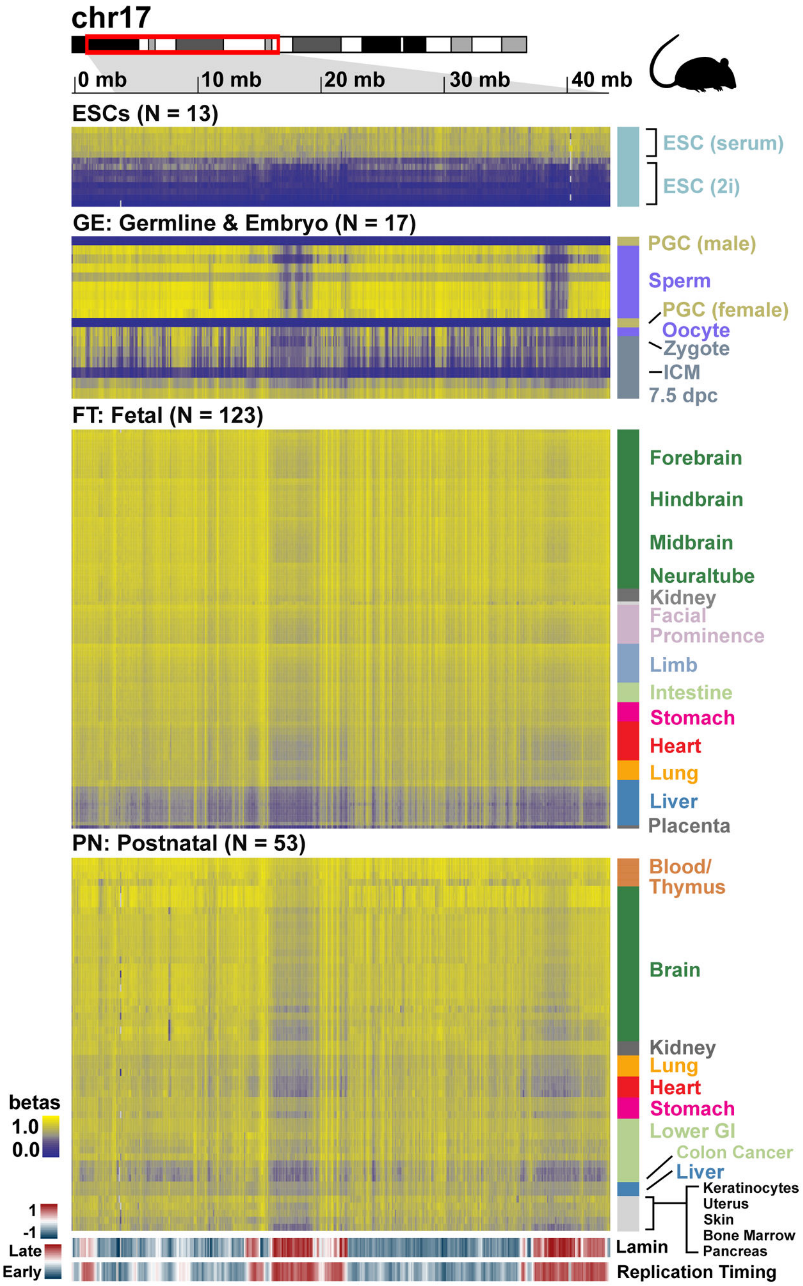

Figure 4. Most PMDs are shared across developmental lineages in mouse

Average solo-WCGW methylation levels were plotted along a representative 30-mb region of chromosome 17 in mouse. 206 WGBS samples are organized into four groups: Embryonic Stem Cells (ESC); Germline and embryos (GE); Fetal tissues (FT); Postnatal tissues (PN); Grouping and ordering of samples were performed as described in Fig. 3. Lamin and replication timing are shown on the bottom of the heatmap. Lamin A DamID from wild type mouse ESCs were downloaded from GEO with accession GSE62683[69]. Replication timing of day 9 differentiated ESCs were downloaded from GEO with accession GSE17983[70].

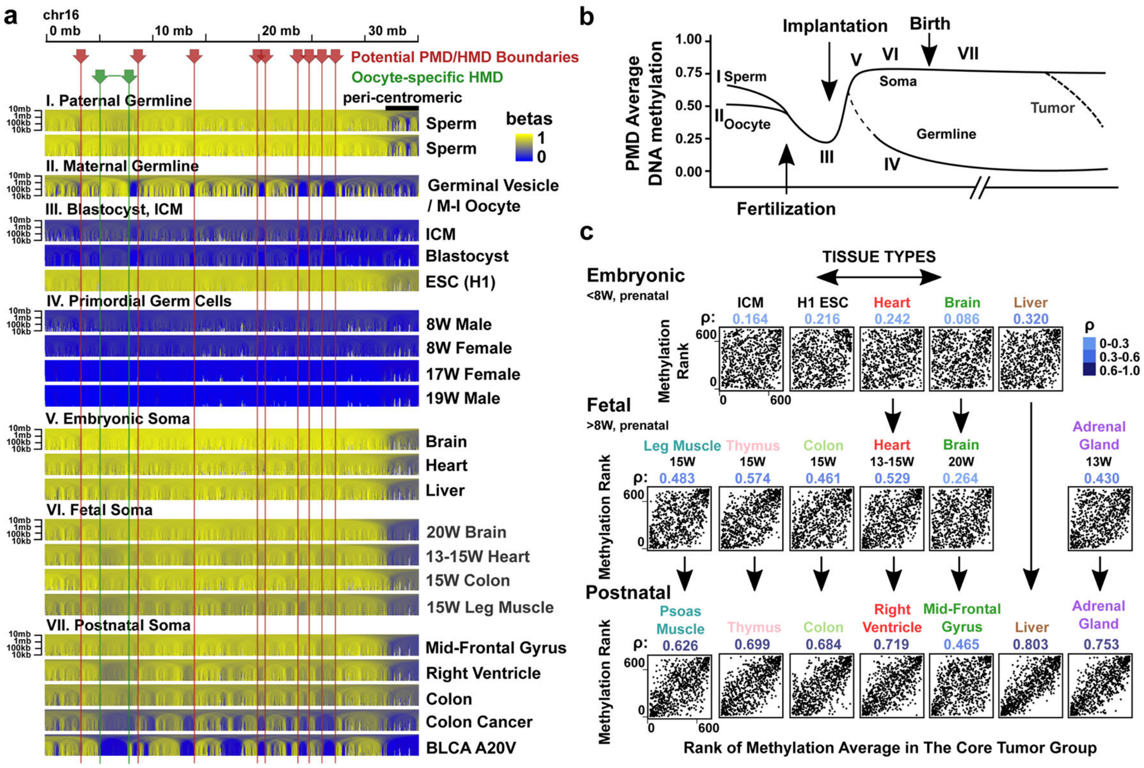

Figure 5. PMD hypomethylation emerges during embryonic development

(**a**) Multi-scale solo-WCGW average plots are shown for samples divided into seven developmental stages, as diagrammed in (**b**): paternal (I) and maternal (II) germ cells, implantation-related tissues (III), primordial germ cells (IV), embryonic soma (V), fetal soma (VI) and postnatal soma (VII). (**c**) Rank-based analysis of the 792 genomic 100-kb bins from chr16, comparing methylation ranks of the core tumors (Y-axis) to each developmental sample (X-axis), with each axis going from a rank of 1 (lowest methylation) to the rank of the highest methylation (excluding bins with missing value from either of the samples). Greater correlations (indicated by the Spearman's correlation coefficient ρ) indicated stronger HMD/PMD structure.

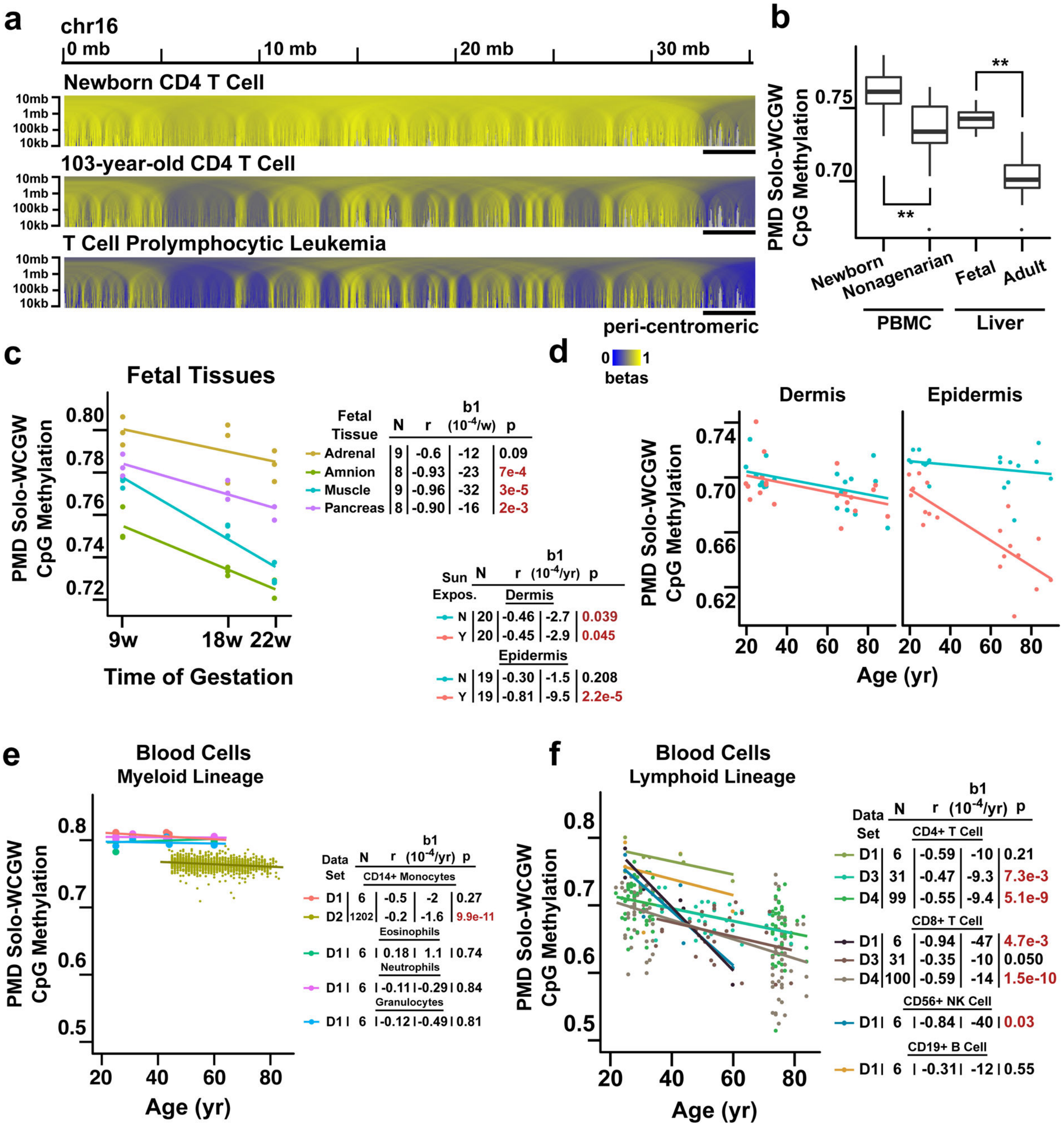

Figure 6. PMD hypomethylation is associated with chronological age

(**a**) Multi-scale solo-WCGW average plots are shown for newborn CD4 T cell, 103-year-old CD4 T cell (GSE31438) and T cell prolymphocytic Leukemia (BLUEPRINT accession S016KWU1). (**b–f**) Summarization of average PMD hypomethylation in HM450-based samples, by averaging beta values for 6,214 solo-WCGW probes mapped to common PMDs (Online Methods). Peripheral Blood Mononuclear Cell (PBMC) in newborns and nonagenarians (left, from GSE30870, p=8.8e-5, one-way Wilcoxon Rank Sum test), and disease-free fetal and adult liver tissue (right, from GSE61278). Center lines of the box plots indicate median, and the lower and upper bounds indicate lower and upper quartiles. The lower and upper whiskers indicate smallest and largest methylation values. ** p <= 0.001 from Wilcoxon Rank Sum test. **(c–f)** HM450-based solo-WCGW averages vs. age for individual donors for several tissue types. N is the number of donors/samples, r is Pearson's product moment correlation, b1 is the estimated rate of methylation loss, and p is the p-value based on Pearson correlation test. (**c**) Four fetal tissue types during three pre-natal time points (from GSE56515). (**d**) Sun-exposed and sun-protected dermis and epidermis (from GSE51954). (**e**) Sorted blood cells of the myeloid lineage (D1: GSE35069; D2: GSE56046). (**f**) Sorted blood cells of lymphoid lineage (D1: GSE35069; D3: GSE71955; D4: GSE59065).

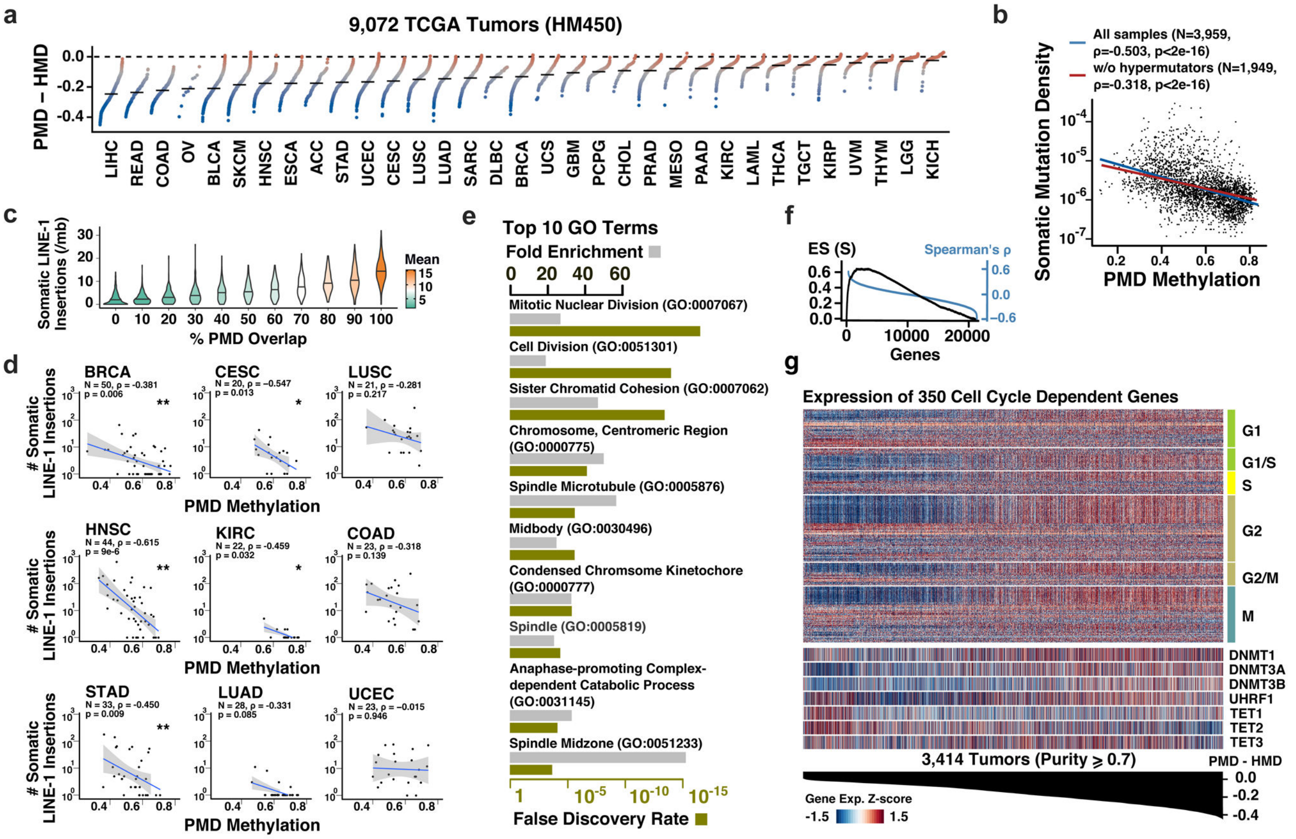

Figure 7. PMD hypomethylation is linked to mitotic cell division in cancer

(**a**) PMD-HMD solo-WCGW methylation difference for 9,072 tumors from TCGA HM450 data. Each sample is ordered within cancer type by PMD-HMD difference, and cancer types are ordered by average PMD-HMD difference. (**b**) PMD methylation (X-axis) vs. somatic mutation density (Y-axis) for all 3,959 high purity TCGA cases (purity>=0.7), with Spearman's ρ indicated. The blue line represents the regression line for all samples, while the red regression line excludes "hypermutator" samples (Online Methods). (**c**) Density of somatic LINE-1 insertions (violin plot elements) in non-overlapping 1-mb genomic bins (N=3,053), stratified by percent of bin overlapping common PMDs (only cases with whole-genome sequencing are included). (**d**) PMD methylation (X-axis) vs. LINE-1 insertion counts (Y-axis) for nine TCGA cancer types having substantial LINE-1 insertion counts. * ($\rho < 0.05$) and ** ($\rho <= 0.01$) indicate Spearman's test significance. (**e**) The 10 most significantly enriched Gene Ontology (GO) terms for the 60 genes with the most strongly correlated expression vs. PMD hypomethylation in TCGA tumors, showing fold enrichment (grey) and false discovery rate (olive). (**f**) Gene Set Enrichment Analysis (GSEA) for 350 cell-cycle-dependent genes from Cyclebase[44], ranking all genes according to degree of expression vs. PMD hypomethylation correlation. (**g**) Normalized expression (Z-scores) of cell-cycle-dependent genes from Cyclebase (categorized by cell cycle phase) in 3,414 high purity TCGA tumor samples (purity >= 0.7), ordered by PMD-HMD methylation difference.

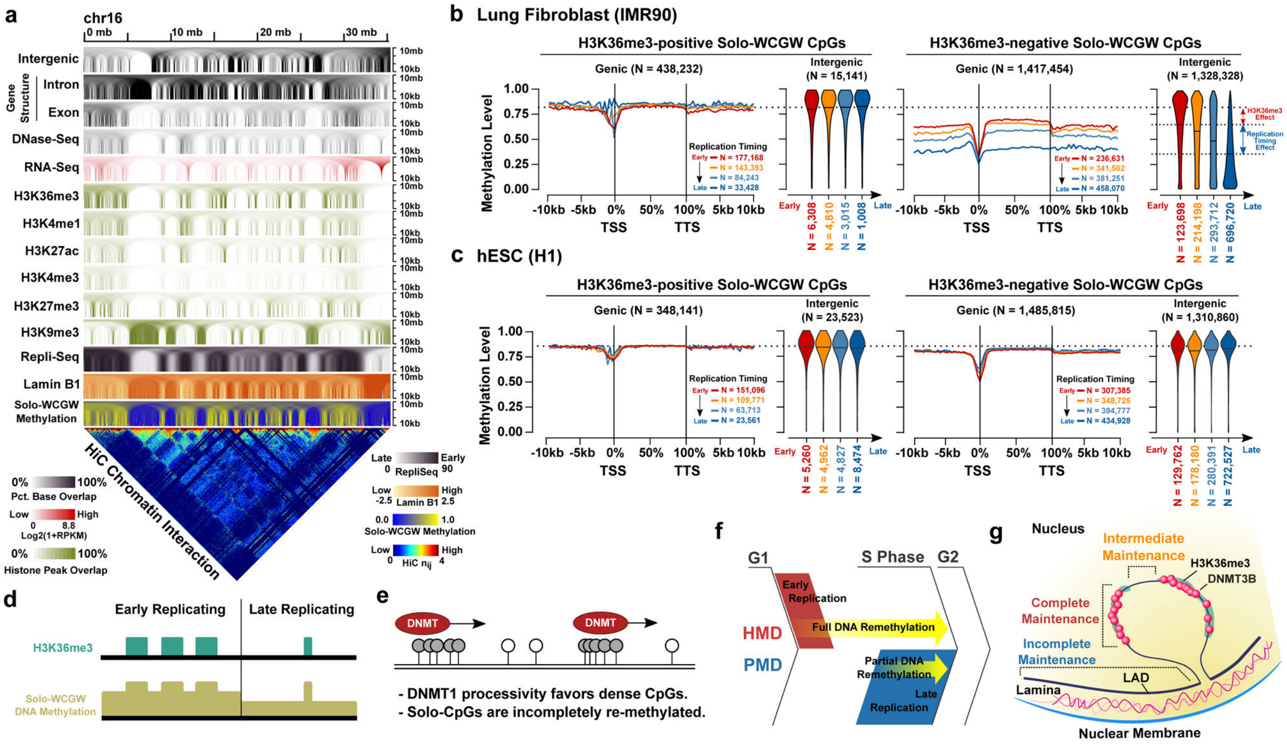

Figure 8. Replication timing and H3K36me3 contribute independently to methylation maintenance

(**a**) Multi-scale plot of chr16p showing similarity between solo-WCGW methylation and other chromatin marks in the IMR90 fibroblast cell line. (**b**) Average methylation level of all genomic solo-WCGWs in IMR90, stratified by (1) overlap with H3K36me3 peaks (left vs. right), (2) context relative to gene annotations ("Genic" vs. "Intergenic"), and (3) Repli-seq replication timing bin (red, yellow, light blue, dark blue). For Solo-WCGWs residing within +/− 10 kb of an annotated gene (Genic), meta-gene plots show methylation averages in relation to the Transcription Start Site (TSS) and the Transcription Termination Site (TTS). For all other Solo-WCGWs (Intergenic), each replication timing group is shown as a single violin plot. **(c)** The same representation of data plotted for the H1 hESC cell line (using Repli-chip data rather than Repli-seq). (**d**) Schematic summary, showing Solo-WCGW CpG methylation loss primarily determined by replication timing domain but locally protected by H3K36me3. (**e**) Schematic model illustrating DNMT1 processivity favoring dense CpGs and leading to incomplete re-methylation of Solo CpGs. (**f**) Schematic illustration of the "re-methylation timing model" where genomic regions synthesized earlier in S-phase (HMDs) spend more time exposed to methylation maintenance machinery and thus more complete methylation maintenance than PMDs. (**g**) Illustration of the relationship between major determinants of hypomethylation and 3D nuclear topology, with Lamina Associated Domains (LADs) occupying a distinct heterochromatic nuclear compartment.