bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation

AUTHORS

Arie B. Brinkman[1*], Serena Nik-Zainal[2,3], Femke Simmer[1#], F. Germán Rodríguez-González[4], Marcel Smid[4], Ludmil B. Alexandrov[2,6,7], Adam Butler[2], Sancha Martin[2], Helen Davies[2], Dominik Glodzik[2], Xueqing Zou[2], Manasa Ramakrishna[2], Johan Staaf[5], Markus Ringnér[5], Anieta Sieuwerts[4], Anthony Ferrari[8], Sandro Morganella[9], Thomas Fleischer[10], Vessela Kristensen[10,11,12], Marta Gut[13] Marc J. van de Vijver[14], Anne-Lise Børresen-Dale[10,11], Andrea L. Richardson[15,16], Gilles Thomas[8], Ivo G. Gut[13], John W.M. Martens[4], John A. Foekens[4], Mike Stratton[2], Hendrik G. Stunnenberg[1*]

AFFILIATIONS

1 Radboud University, Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Nijmegen, Netherlands

2 Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK

3 Academic Department of Medical Genetics, University of Cambridge, Cambridge CB2 0QQ, UK

4 Erasmus MC Cancer Institute and Cancer Genomics Netherlands, Erasmus University Medical Center, Department of Medical Oncology, Rotterdam, The Netherlands

5 Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden

6 Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7 Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America

8 Synergie Lyon Cancer,Centre Léon Bérard, 28 rue Laënnec, Lyon Cedex 08, France

9 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus,Hinxton, Cambridgeshire, CB10 1SD

10 Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, 0310 Norway

11 K.G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo, 0316 Norway

12 Department of Clinical Molecular Biology and Laboratory Science (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, 1478 Norway

13 Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Barcelona, Spain

14 Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands

15 Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115 USA

16 Dana-Farber Cancer Institute, Boston, MA 02215 USA

[#] Current address: Department of Pathology, Radboud University Nijmegen Medical Centre, P.O. Box 9101, Nijmegen 6500 HB, The Netherlands

[*] Corresponding authors: A.B. Brinkman (a.brinkman@science.ru.nl), H.G. Stunnenberg (h.stunnenberg@ncmls.ru.nl)

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SUMMARY

Global loss of DNA methylation and CpG island (CGI) hypermethylation are regarded as key epigenomic aberrations in cancer. Global loss manifests itself in partially methylated domains (PMDs) which can extend up to megabases. However, the distribution of PMDs within and between tumor types, and their effects on key functional genomic elements including CGIs are poorly defined. Using whole genome bisulfite sequencing (WGBS) of breast cancers, we comprehensively show that loss of methylation in PMDs occurs in a large fraction of the genome and represents the prime source of variation in DNA methylation. PMDs are hypervariable in methylation level, size and distribution, and display elevated mutation rates. They impose intermediate DNA methylation levels incognizant of functional genomic elements including CGIs, underpinning a CGI methylator phenotype (CIMP). However, significant repression effects on cancer-genes are negligible as tumor suppressor genes are generally excluded from PMDs. The genomic distribution of PMDs reports tissue-of-origin of different cancers and may represent tissue-specific 'silent' regions of the genome, which tolerate instability at the epigenetic, transcriptomic and genetic level.

Global loss of methylation was among the earliest recognized epigenetic alterations of cancer cells[1]. It is now known to occur in large genomic blocks that partially lose their default hypermethylated state, termed partially methylated domains (PMDs)[2–6]. PMDs have been described for a variety of cancer types and appear to represent repressive chromatin domains that are associated with nuclear lamina interactions, late replication and low transcription. PMDs are not exclusive to cancer cells and have also been detected in some normal tissues[2,7–11], but not in pluripotent cells and brain tissue[12,13]. PMDs can comprise up to half of the genome[3,4], and it has been suggested that PMDs in different tissues are largely identical[3]. PMDs have been shown to harbor 'focal' sites of hypermethylation that largely overlap with CGIs[3]. Questions remain as to what instigates such focal hypermethylation, whether loss of methylation inside PMDs is linked to repression of cancer-relevant genes and whether the genomic distribution of PMDs is invariant throughout primary tumors of the same type, perhaps determined by tissue-of-origin. In breast cancer, PMDs have been detected in two cultured cancer cell lines[5], but their extent and variation

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in primary tumors is hitherto unknown. A major limitation of most DNA methylation studies is that only a small subset of CpGs are interrogated. This prevents accurate determination of the extent and location of PMDs. Few samples of a certain tissue/tumor have typically been analyzed using whole-genome bisulfite sequencing (WGBS). Thus, observations cannot be extrapolated to individual cancer types, let alone generalized to other cancers. Here, we analyzed DNA methylation profiles of 30 primary breast tumors at high resolution through WGBSsThis allowed us to, and delineate PMD characteristics in detail. We show that PMDs define (breast) cancer methylomes and are linked to other key epigenetic aberrations such as CGI hypermethylation.

RESULTS

Primary breast tumors display variable loss of DNA methylation

To study breast cancer epigenomes we performed WGBS encompassing ~95% of annotated CpGs (Suppl. Fig. 1A, Suppl. Table 1). For 25/30 and 24/30 of these tumors we previously analyzed their full genomes[14,15] and transcriptomes[16], respectively. Of the 30 tumors, 25 and 5 are ER-positive and ER-negative, respectively (Suppl. Fig. 1B).

To globally inspect aberrations in DNA methylation patterns we generated genome-wide and chromosome-wide methylome maps by displaying mean methylation in consecutive tiles of 10 kb (see Methods). These maps revealed extensive inter-tumor variation at genome-wide scale (Fig. 1A), that lacked obvious association with ER-status (p=0.15, *t*-test, Suppl. Fig. 2A). At chromosome level, we observed stably hypermethylated regions next to regions that were hypomethylated to various extents and across tumors (Fig. 1B). Chromosomes 1 and X were exceptionally prone to methylation loss. At megabase scale (Fig. 1C) DNA methylation profiles showed that the widespread loss of methylation occurred in block-like structures previously defined as PMDs[2]. Across primary breast tumor samples, DNA methylation levels and genomic sizes of PMDs differ extensively between tumors and PMDs do appear as separate units in some tumors and as merged or extended in others, underscoring the high variation with which methylation loss occurs. Despite this variation, however, we observed common PMD boundaries as well.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Given the variation between tumors, we asked whether the patterns of methylation loss were associated with distribution of copy-number variations (CNVs) throughout the genome. We found no evidence for such association (Pearson R=0.17), although we noticed that chromosomes with the most pronounced loss of methylation (chr1, chrX, chr8-p) frequently contained amplifications (Suppl. Fig. 1C). Next, we asked whether loss of methylation was associated with aberrant expression of genes involved in writing, erasing, or reading the 5-methylcytosine modification. However, we found no such correlation (Suppl. Fig. 1D).

To provide a reference for the observed patterns of methylation loss we compared WGBS profiles of primary breast tumors to that of 72 normal tissues (WGBS profiles from Roadmap Epigenomics Project and[10], Suppl. Fig. 2A,B). In sharp contrast to breast cancer, most normal tissues were almost fully hypermethylated (except for pancreas and skin), with heart, thymus, embryonic stem cell(-derived), induced pluripotent stem cells and brain having the highest levels of methylation. Importantly, inter-tissue variation was much lower as compared to breast tumors ($p < 2.2e-16$, MWU-test on standard deviations). Thus, breast tumors show widespread loss of DNA methylation in PMDs, and the extent and patterns appear to be hypervariable between tumor samples. In line with this, principal component analysis confirmed that methylation inside PMDs is the primary source of variation across full-genome breast cancer DNA methylation profiles (Fig. 1D): the first principal component (PC1) is strongly associated with mean PMD methylation (p=6.8e-07). The second-largest source of variation, PC2, is associated with ER status (p=1.9e-06, Fig. 1D), while successive PCs were not significantly associated with any clinicopathological feature. It should be noted that with 30 tumors only very strong associations can achieve statistical significance. Taken together, breast tumor whole-genome DNA methylation profiles reveal global loss of methylation due to PMDs, the extent of which is hypervariable across tumors and represent the major source of variation between tumors.

Distribution and characteristics of breast cancer PMDs

We set out to further characterize breast cancer PMDs and their variation (see Methods: data access). The genome fraction covered by PMDs varies greatly across our WGBS cohort of 30 tumors, ranging between 10% and 50% across tumors, covering 32% of the genome on average

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Fig. 2A). We define 'PMD frequency' as the number of tumors in which a PMD is detected. A PMD frequency of 30 (PMDs common to all 30 cases) occurs in only a very small fraction of the genome (2%), while a PMD frequency of 1 (representing the union of all PMDs from 30 cases) involves 70.2% of the genome (Fig. 2B). We tested to which extent PMD distribution is random, by counting PMD borders in 30-kb genomic tiles (Fig. 2C). Randomly shuffled PMDs yield a normal distribution centered at a PMD frequency of four. In contrast, observed PMDs show a skewed distribution: the mode was for a PMD frequency of 0 suggesting that many tiles (23,492, 25%) do not coincide with any PMD borders. The majority of tiles (62%) had a low PMD border frequency (1-10). The tail represents low numbers of tiles with up to maximal PMD frequency of 30. We conclude that PMD distribution is not random: part of the genome appears not to tolerate PMDs while PMDs occur in a large fraction of the genome with varying frequencies.

PMDs have been shown to coincide with lamin-associated domains (LADs)[3,4]: large repressive domains that preferentially locate to the nuclear periphery[17]. LADs are characterized by low gene density and late replication[17,18]. Accordingly we found that PMDs show reduced gene densities (Fig. 2E, Suppl. Fig. 3A), have high LaminB1 signals (associated with LADs[17], Fig. 2D), are late replicating (ENCODE data, Fig. 2D) and have a low frequency of (Hi-C) 3D loops[19], an indicator of lower levels of transcription. Finally, we observed a local increase in binding of the transcription factor CTCF at the borders of PMDs (Fig. 2D) as shown in previous reports[3,17,20–22].

We previously analyzed the full transcriptomes (RNA-seq) in a breast cancer cohort of 266 cases[16] from which our WGBS cohort is a subset. We determined the mean expression of genes as a function of PMD frequency. Genes inside PMDs are expressed at consistently lower levels than genes outside of PMDs (Fig. 2F, $p < 2.2e-16$, *t*-test), with a tendency towards lower expression in highly-frequent PMDs ($p < 2.2e-16$, linear regression). Given the variable nature of DNA methylation patterns of PMDs, we also determined the variation in gene expression as a function of PMD frequency and found higher variation for genes inside PMDs (Fig. 2F, $p < 2.2e-16$, MWU-test). Even when restricting this analysis to only the subset of 24 overlapping cases from the transcriptome and WGBS cohort we observed the same trends, with similar statistical significance (Suppl. Fig. 3B, $p < 2.2e-16$, *t*-test for expression; $p < 2.2e-16$, MWU-test for variation). Given the observed variability of DNA methylation and gene expression inside PMDs,

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

we asked whether genetic stability, i.e. the number of somatic mutations, was also altered within PMDs. In our cohort of 560 full breast cancer genomes[14], substitutions, insertions, and deletions occur more frequently within than outside PMDs, with a clear increase in highly frequent PMDs ($p < 2.2e-16$ for each mutation type, logistic regression, Fig. 2G). In contrast, rearrangements are more abundant outside of PMDs ($p < 2.2e-16$, logistic regression), in keeping with the hypothesis that regions with higher transcriptional activity are more susceptible to translocations[23]. As above, a restrictive analysis of only the 25 overlapping cases from the full genomes and WGBS cohorts revealed the same trends except for insertions ($p < 2.2e-16$ (substitutions); $p = 0.362$ (insertions), $p = 1.7e-05$ (deletions), $p = 1.1e-09$ (rearrangements), logistic regressions, Suppl. Fig. 3C). Taken together, breast cancer PMDs share key features of PMDs including low gene density, low gene expression, and colocalization with LADs, suggesting that they reside in the 'B' (inactive) compartment of the genome[24]. Importantly, in addition to epigenomic instability, breast cancer PMDs also tolerate transcriptomic variability and genomic instability.

Relationship between CpG island methylation and PMDs in breast cancer

To determine how PMDs affect methylation of functional genomic elements we accordingly stratified all CpGs from all tumors and assessed the methylation distribution in these elements (Fig. 2H). We found that the normally observed near-binary methylation distribution is lost inside PMDs; the hypermethylated bulk of the genome and hypomethylated CGIs/promoters acquire intermediate levels of DNA methylation inside PMDs. DNA methylation deposition inside PMDs thus appears incognizant of genomic elements, resulting in intermediate methylation levels regardless of the genomic elements' functions. Among all elements, the effect of incognizant DNA methylation deposition is most prominent for CGIs as they undergo the largest change departing from a strictly hypomethylated state. This has been described also as focal hypermethylation inside PMDs[3].

We further focused on methylation levels of CGIs. When indiviual PMDs are regarded, CGIs inside of them lose their strictly hypomethylated state and become more methylated to a degree that varies between tumors (Fig. 3A). Across all tumors and all CGIs, this effect is extensive (Fig.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3B,C), affecting virtually all CGIs inside PMDs: on average 92% of CGIs lose their hypomethylated state and gain some level of methylation (Fig. 3B, left panel). Outside of PMDs only 25-30% of the CGIs is hypermethylated, although to a higher level (Fig. 3B, right panel). Thus, incognizant deposition of DNA methylation inside PMDs results in extensive hypermethylation of virtually all PMD-CGIs.

Concurrent hypermethylation of CGIs in cancer has been termed CIMP[25], and in breast cancer this phenomenon has been termed B-CIMP[26–28]. To determine whether CIMP is directly related to PMD variation we defined B-CIMP as the fraction of CGIs that are hypermethylated (>30% methylated), and determined its association with the fraction of CGIs inside PMDs. Regression analysis (see Methods) showed that this association is highly significant (Fig. 3D, p=2.1e-08, R^2=0.51, n=30). The fraction of hypermethylated CGIs is generally higher than the fraction of hypermethylated CGIs in PMDs, suggesting that CGI hypermethylation is not solely dependent on PMD occurrence. However, CGI methylation levels outside PMDs are far more stable than inside PMDs (Fig. 3E), which likely represents an invariably methylated set of CGIs (Suppl. Table 2).

We applied the same regression analysis to other tumor types (TCGA,[29], Fig. 3F). Although sample sizes were small, we found significant CIMP-PMD associations for lung adenocarcinoma (LUAD), rectum adenocarcinoma (READ), uterine corpus endometrial carcinoma (UCEC) and bladder urothelial carcinoma (BLCA). We did not find significant associations for Burkitt's lymphoma (BL), lung squamous cell carcinoma (LUSC), follicular lymphoma (FL), and glioblastoma (GBM), even though G-CIMP has been previously described[30]. Taken together, we conclude that PMD occurrence is an important determinant for CIMP.

PMD demethylation effects on gene expression

To assess whether widespread hypermethylation of CGI-promoters within PMDs instigates gene repression we analyzed expression as a function of gene location inside or outside of PMDs. Overall, CGI-promoter genes showed a mild but significant downregulation when inside PMDs (p=4.5e-12, *t*-test), while strong downregulation was specifically restricted to low-frequency PMDs (Fig. 3G). For non-CGI-promoter genes this trend was very weak or absent (Suppl. Fig.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4A). As healthy controls were not included in transcriptome analysis of our cohort[16] we used gene expression (RNA-seq) profiles from breast tumors (769) and normal controls (88) from TCGA. Similar to our cohort (see Fig. 2F) we found that overall gene expression for the TCGA tumors is lower inside PMDs, with lowest expression for genes inside high-frequent PMDs (Fig. 3H, $p < 2.2e-16$, linear regression). However, the expression of genes in tumor PMDs is very similar to healthy control samples ($p = 0.807$, linear regression). To analyze this in more detail we selected normal/tumor matched pairs (i.e. from the same individuals, n=86) and analyzed the fold change over the different PMD frequencies (Fig. 3I). As in our cohort, downregulation is restricted to genes with low PMD-frequency ($p < 2.2e-16$ for PMD frequency 1-3, linear regression). No obvious changes occur in high-frequency PMD genes, nor in non-CGI-promoter genes (Suppl. Fig. 4B). Taken together, widespread cancer-associated repression of all genes inside PMDs is limited: downregulation is restricted to low-frequency (i.e. the more variable) PMDs and affects only CGI-promoter genes, which undergo widespread hypermethylation inside PMDs.

Given the widely accepted model of hypermethylated promoter-CGIs causing repression of tumor suppressor genes (TSGs) we determined whether breast cancer PMDs overlap with these genes to instigate such repression. For non-TSGs as a reference we found that 64% (14,037) are located outside of PMDs (Fig. 3J), while 36% are located inside, (see also Fig. 2E). Strikingly, TSGs (Cancer Gene Census) overlap poorly with PMDs: most TSGs (218/254, 86%) are located outside of PMDs. Only 14% overlap with mostly low-frequency PMDs, implying exclusion of TSGs from PMDs ($p=8.8e-16$, hypergeometric test). When we specifically focused on breast cancer-related TSGs (Cancer Gene Census), this exclusion was even stronger: practically all (27/28, 96%) breast cancer TSGs are located outside of PMDs ($p=3.5e-06$, hypergeometric test). Similarly, from our previously identified set of genes containing breast cancer driver mutations[14]: 86/93 (92%) were located outside of PMDs ($p=2.0e-11$, hypergeometric test). Alltoghether, only 31 breast cancer-mutated genes were not excluded from PMDs. We assessed whether these genes are downregulated in tumors when inside PMDs. 24/31 (74%) genes were downregulated (Suppl. Fig. 5A,B), and an overall negative correlation between CGI-promoter methylation and expression was evident (Suppl. Fig. 5C). For 16 out of these 24 genes we confirmed that significant downregulation also takes place in cancer relative to normal in an independent breast cancer expression dataset (TCGA, Suppl. Fig. 5D and data not shown). Among the

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

downregulated genes in PMDs are EGFR (epidermal growth factor receptor) and PDGFRA (platelet-derived growth factor receptor α) that have tumor promoting mutations (Suppl. Fig. 5A,B,C). Paradoxically, both genes are significantly downregulated in our as well as the TCGA breast cancer dataset (Suppl. Fig. 5D). Taken together, despite the large number of hypermethylated CpG islands inside breast cancer PMDs (13,013 CGIs; 47%, Suppl. Fig. 4C), these CGIs do not generally co-occur with TSGs and other breast cancer-relevant genes. Repression of these genes through classical promoter-hypermethylation in PMDs does not occur at large scale, and is likely limited to a few genes.

We next identified genes that are downregulated when inside PMDs regardless of any documented TSG function or mutation in breast cancer. 400 genes were downregulated at least 2.5 log2-fold (Suppl. Table 3). Gene set enrichment analysis showed that these genes were involved in processes such as signaling and adhesion (Suppl. Fig. 6A). In addition, there is a significant enrichment of genes downregulated in luminal B breast cancer (and upregulated in basal breast cancer)[31]. This suggests that PMDs are involved in downregulation of luminal B-specific genes. Examples of luminal B-downregulated genes include CD3G, encoding the gamma polypeptide of the T-cell receptor-CD3 complex (gene sets 'signaling' and 'adhesion'), and RBP4, encoding retinol binding protein 4 (gene set 'signaling') (Suppl. Fig. 6B). Stratification of tumors according to low and high median expression of the 400 PMD-downregulated genes revealed significant differences in overall survival of the corresponding patients (p=2.6e-03, *chi*-square test, Suppl. Fig. 6C), suggesting clinical significance of PMD-associated gene repression. Taken together, downregulation of genes inside PMDs occurs rarely and is restricted to low-frequency PMDs. However, these rare cases include genes relevant to breast cancer given the overlap with previously identified luminal B breast cancer-relevant genes and differential overall survival.

PMDs are not unique to cancers, but reduced DNA methylation in PMDs is a feature of many cancers

To assess the generality of PMD occurrence in cancer, we extended our analysis to other cancer types and normal tissues. We performed PMD detection in a total of 134 WGBS profiles (57

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tumors from TCGA,[29] and 77 normal tissues from the Roadmap Epigenomics Project and[5]). PMDs are detectable in virtually all tumors, but also in 30% of normal tissues (Fig. 4A, see Methods: data access). However, mean DNA methylation inside detected PMDs is much lower in tumors as compared to normal tissues (Fig. 4B, Suppl. Fig. 7A, p=1.0e-18, *t*-test), and is not associated with tumor tissue origin: upon ranking the samples according their mean PMD methylation, tumors of the same type are dispersed rather than clustered together (Fig. 4B). Thus, PMDs are not unique to tumors per se but the overall loss of methylation inside PMDs is consistently greater in tumors. Still, the absolute loss does not typify tumor tissue origin, underscoring the variable nature of methylation within PMDs. To assess whether CGI hypermethylation in PMDs is as extensive in these additional tumor types as in breast cancer, we analyzed CGI methylation of the 57 tumor samples in total (Suppl. Fig. 7B, see Methods: data access). As in breast cancer, widespread hypermethylation of CGIs inside PMDs was consistent in most tumor types. The levels of hypermethylation in Burkitt's lymphoma (BL)[29] were among the highest of all tested tumors, while hypermethylation levels in lung adenocarcinoma/squamous cell carcinoma (LUAD/LUSC) were slightly lower than in other tumors. Possibly, these differences are linked to tumor cellularity of the samples. In two glioblastoma multiforme tumors, CGI hypermethylation was not restricted to PMDs, which is suggestive of inaccurate PMD detection due to high methylation inside glioblastoma PMDs (see Fig. 4B). Importantly, these results extend the observed tendency of CGI hypermethylation inside PMDs to other tumors.

Lastly, to assess whether the distribution of tumor PMDs reflects tissue of origin we scored the presence of PMDs in genomic tiles of 30 kb and subsequently clustered the resulting binary profiles. The analysis showed that the majority of tumors of the same type clustered together, although not fully accurately (Fig. 4C), suggesting that the genomic distribution of PMDs is linked to tissue of origin. Thus, even though methylation levels of PMDs are independent of tissue-of-origin (Fig. 4B), the distribution of PMDs associates with tissue of origin, likely reflecting differences in the genomic parts that tolerate PMDs.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

In this study we analyzed breast cancer DNA methylation profiles to high resolution. The main feature of breast cancer epigenomes is the extensive loss of methylation in PMDs and their hypervariability. Directly linked to this is the concurrent CGI hypermethylation, for which PMDs appear to be a major driver or even causal. Although various features of PMDs have been described before, our study is the first to include a larger WGBS cohort from one tumor type, while integrating sparse WGBS data from other tumor types. PMDs may be regarded as tissue-type-specific inactive constituents of the genome: the distribution shows tissue-of-origin specificity, gene expression inside PMDs is low and they are late replicating. Inside PMDs the accumulation of breast cancer mutations is higher than outside of them. The resulting domain-like fluctuation in mutation density is likely related to the fluctuating mutational density along the genome in cancer cells observed by others[32–34]. The phenomena observed in breast cancer extend to tumors of at least 10 additional tissue types underscoring the generality of our findings. We conclude that loss of methylation in PMDs and concurrent CGI hypermethylation is a general hallmark of most tumor types with the exception of AMLs (data not shown).

The phenomena that we describe for breast cancer have remained elusive in genome-scale studies that only assessed subsets of the CpGs; the sparsity of included CpGs does not allow accurate PMD detection. Typical analysis strategies include tumor stratification by clustering of the most highly variable CpGs which at least in our breast cancer cohort are located in PMDs. In effect such approaches are biased towards CGIs due to their design and consequently, the hypermethylation groups represent tumors in which PMDs are highly abundant (e.g.[30,35–43]). It is very likely that for some tumor types hypermethylation groups associate with clinicopathological features, amongst which a positive association with tumor cellularity is recurrent[36,40–42]. This suggests that PMDs are more pronounced in tumor cells than in the non-tumor tissue of a cancer sample. This makes hypermethylated CGIs useful diagnostic markers but less likely informative as prognostic markers informing about tumor state, progression and outcome.

Since PMDs are domains in which instability at the genetic, epigenetic, and transcriptome level is tolerated, they may provide plasticity that is beneficial for the heterogeneity of tumor cells.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS

Data access

Tables containing CpG methylation values (bigwig), genomic coordinates and mean methylation values of PMDs and CGIs are available via DOI 10.5281/zenodo.1217427 or DOI 10.17026/dans-276-sda6. Raw data for whole-genome bisulfite sequencing of the 30 breast tumor samples of this study is available from the European Genome-phenome Archive (https://www.ebi.ac.uk/ega) under dataset accession EGAD00001001388.

Sample selection, pathology review and clinical data collection

Sample selection, pathology review and clinical data collection for this study has been described in[14].

Processing of whole-genome bisulfite sequencing data

WGBS library preparation, read mapping, and methylation calling was done as described before[44]. The genome build used for mapping of bisulfite sequencing reads, and throughout this study was hg19 (GRCh37).

Principal component analysis of WGBS data

For principal component analysis (PCA) of WGBS profiles, CpGs with coverage of at least 10 were used. Subsequently, the top 5% most variable CpGs were selected. We used the FactoMineR package[45] for R to perform PCA and to determine association of principal components with clinicopathological features.

Detection of PMDs

Detection of partially methylated domains (PMDs) in all methylation profiles throughout this study was done using the MethylSeekR package for R[46]. Before PMD calling, CpGs overlapping common SNPs (dbSNP build 137) were removed. The alpha distribution[46] was used to determine whether PMDs were present at all, along with visual inspection of WGBS profiles. After PMD calling, the resulting PMDs were further filtered by removing regions overlapping with centromers (undetermined sequence content).

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Mean methylation in PMDs and genomic tiles

Wherever mean methylation values from WGBS were calculated in regions containing multiple CpGs, the 'weighted methylation level'[47] was used. Calculation of mean methylation within PMDs or genomic tiles involved removing all CpGs overlapping with CpG island(-shores) and promoters, as the high CpG densities within these elements yield unbalanced mean methylation values, not representative of PMD methylation. For genome/chromosome-wide visualizations (Fig. 1), 10-kb tiles were used. For visualization, the samples were ordered according hierarchical clustering of the tiled methylation profiles, using 'ward.D' linkage and [1-Pearson correlation] as a distance measure.

Clustering on PMD distribution

For each sample, the presence of PMDs was binary scored (0 or 1) in genomic tiles of 5 kb. Based on these binary profiles, a distance matrix was calculated using [1-Jaccard] as a distance metric, which was used in hierarchical clustering using complete linkage.

Tumor suppressor genes and driver mutations

For overlaps with tumor suppressor genes, Cancer Gene Census (http://cancer.sanger.ac.uk/census, October 2017) genes were used. Overlaps with genes containing breast cancer driver mutations were determined using the list of 93 driver genes as published previously by us[14].

CIMP

To determine the association between B-CIMP (fraction of CGIs that are hypermethylated, >30% methylated) and PMD occurrence we used beta-regression using the 'betareg' package in R[48].

Survival analysis

Survival analysis of patient groups stratified by expression of genes downregulated in PMDs. For each tumor sample of our breast cancer transcriptome cohort (n=266,[16]), the median expression of all PMD-downregulated genes (Suppl. Table 3) was calculated. The obtained distribution of these medians was used to stratify patient groups, using a two-way split over the median of this

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

distribution. Overall survival analysis using these groups was done using the 'survival' package in R, with *chi*-square significance testing.

ACKNOWLEDGEMENTS

This work has been funded through the ICGC Breast Cancer Working group by the Breast Cancer Somatic Genetics Study (BASIS), a European research project funded by the European Community's Seventh Framework Programme (FP7/2010-2014) under the grant agreement number 242006; the Triple Negative project funded by the Wellcome Trust (grant reference 077012/Z/05/Z). For contributions towards specimens and collections: Tayside Tissue Bank, OSBREAC consortium, Icelandic Centre for Research (RANNIS), Swedish Cancer Society, Swedish Research Council, Fondation Jean Dausset-Centre d'Etudes du polymorphisme humain, Icelandic Cancer Registry, Brisbane Breast Bank, Breast Cancer Tissue and Data Bank and ECMC (King's College London), NIHR Biomedical Research Centre (Guy's and St Thomas's Hospitals), Breakthrough Breast Cancer, Cancer Research UK. We thank E.M. Janssen-Megens, K. Berentsen, H. Kerstens and K.J. Francoijs for technical support. We thank H. Kretzmer and R. Siebert for providing processed data files of the lymphoma dataset. Funding to A.B.B. was through the Dutch Cancer Foundation (KWF) grant KUN 2013-5833. SN-Z is personally funded by a CRUK Advanced Clinician Scientist Award (C60100/A23916). M.S. was supported by the EU-FP7-DDR response project. L.B.A. is supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. A.L.R. is partially supported by the Dana-Farber/Harvard Cancer Center SPORE in Breast Cancer (NIH/NCI 5 P50 CA168504-02). J.A.F. was funded through an ERC Advanced Grant (ERC-2012-AdG-322737) and ERC Proof-of-Concept Grant (ERC-2017-PoC-767854). A.S. was supported by Cancer Genomics Netherlands (CGC.nl) through a grant from the Netherlands Organisation of Scientific research (NWO). We received additional support from the Dutch national e-infrastructure (SURF Foundation). Finally, we would like to acknowledge all members of the ICGC Breast Cancer Working Group.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FIGURE LEGENDS

Figure 1 | Visualization of inter-tumor variation at genome-wide scale.

(A), Genome-wide and **(B)**, chromosome-wide maps of WGBS DNA methylation profiles from 30 breast tumor samples. Mean methylation is displayed in consecutive tiles of 10 kb (see Methods). Ordering of tumor samples is according clustering of the tiled profiles. **(C)**, WGBS DNA methylation visualization at megabase-scale. Pink coloring indicates common methylation loss (PMDs), although tumor-specific PMD borders vary. A scale bar (100 kb) is shown at the top of each panel. CpG islands are indicated in green. **(D)**, Principal component analysis of WGBS DNA methylation profiles (see Methods). Each tumor sample is represented with its estrogen-receptor (ER) status (point shape) and mean PMD methylation (point color).

Figure 2 | Characterization of breast cancer PMDs.

(A), Fraction of the genome covered by PMDs. Each dot represents one tumor sample, the boxplot summarizes this distribution. **(B)**, Fraction of the genome covered by PMDs that are common between breast tumors. PMD frequency: the number of tumors in which a genomic region or gene is a PMD. **(C)**, Breast cancer PMDs are not distributed randomly over the genome. The genome was dissected into 30-kb tiles, PMD frequency (number of boundaries) was calculated for each tile. The same analysis was done after shuffling the PMDs of each tumor sample. **(D)**, Average profiles of LaminB[17], repliSeq (DNA replication timing, ENCODE), 3D chromatin interaction loops (HiC[21], and CTCF (ENCODE) over PMD borders. If available, data from the breast cancer cell line (MCF7) and mammary epithelial cells (HMEC) was used, otherwise data from fibroblasts (IMR90, Tig3) was used. **(E)**, Gene distribution inside PMDs (as a fraction of all annotated genes). **(F)**, Gene expression inside PMDs. Gene expression (top) and standard deviation (bottom) was plotted as a function of PMD frequency. **(G)**, Somatic mutations inside PMDs. Substitutions, insertions, deletions, and rearrangements were calculated for each of the 560 fully sequenced breast cancer genomes[14], and plotted as a function of PMD frequency. **(H)**, Distribution of DNA methylation over functional genomic elements, inside and outside PMDs. CpGs were classified according PMD status and genomic elements, and the distribution of DNA methylation within each element was plotted.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3 | CpG island hypermethylation inside PMDs.

(A), Example of a genomic region with CGI hypermethylation inside PMDs. Red bars, PMDs for each tumor sample; below, CGI methylation for each tumor sample (same ordering). Green bars, CGIs. **(B)**, Distribution of CGI methylation, represented as the fraction of all CGIs (x-axis). Each horizontal bar represents one tumor sample. **(C)**, Average profile of methylation over all CGIs inside (red) or outside (black) PMDs, over all 30 tumor samples. Black/red lines, median; grey/pink area, 1st and 3rd quartiles. **(D)**, Regression analysis of B-CIMP (y-axis) as a function of the fraction CGIs inside PMDs (x-axis). B-CIMP is defined as the genome-wide fraction of hypermethylated CGIs (>30% methylation). **(E)**, Variation of CGI methylation (standard deviation) as a function of PMD frequency. **(F)**, Summary of regression analysis as in (D), including additional cancer types. n, the number of samples for each type. For abbreviations of cancer type names, see Fig. 4C. **(G)**, Expression change of CGI-promoter genes inside vs. outside of PMDs, as a function of PMD frequency. **(H)**, Gene expression levels as a function of PMD frequency in an independent breast cancer dataset (TCGA). PMD frequency for each gene was taken from our own dataset. **(I)**, Expression change of CGI-promoter genes of tumor vs. normal, as a function of PMD frequency. From the TCGA breast cancer dataset, matched tumor/normal pairs were selected. PMD frequency for each gene was taken from our own dataset. **(J)**, Tumor-suppressor genes (TSGs) are excluded from PMDs. For each TSG its PMD frequency was determined and the resulting distribution was plotted. Main plot, relative distribution; inset, absolute number of genes. 'Non-TSGs', genes not annotated as TSGs; 'TSGs all cancers', genes annotated as TSGs regardless of cancer type; 'TSGs breast cancer', genes annotated as TSG in breast cancer; 'Nik-Zainal breast cancer driver mutations', genes with driver mutations in breast cancer[14].

Figure 4 | PMD methylation in normal tissues and tumors of various tissues.

(A), PMDs are detectable in virtually all tumors, but only in a fraction of normal tissues. WGBS data was used for PMD detection. **(B)**, Mean PMD methylation of normal tissues and tumors of various tissues. Each dot represents one sample. Arrows represent breast tumor samples from this study. **(C)**, Hierarchical clustering of tumor samples based on genomic distribution of their PMDs.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Figure 1

(A), CpG coverage in WGBS DNA methylation profiles of 30 breast tumor samples used in this study (see also Suppl. Table 1). **(B)**, Clinicopathological features of the 30 tumor samples. **(C)**, Mean copy-number profiles of 25/30 tumor samples used in this study. Copy-number data was taken from our previous work[14]. **(D)**, Association between mean PMD methylation and expression of genes involved in writing, erasing, or reading the 5-methylcytosine modification. Each dot represents one tumor sample. Linear regression was used to determine the variation explained (R^2) and the p-value of the association. Expression data was taken from our previous work[16].

Supplemental Figure 2

Visualization of inter-tumor variation at genome-wide scale, as in main Figure 1, but including WGBS data from 72 additional, non-tumor tissues (Roadmap Epigenomics Project and ref.[10]). **(A)**, Genome-wide and **(B)**, chromosome-wide maps. Mean methylation is displayed in consecutive tiles of 10 kb (see Methods). For breast tumors of this study, the ER-status is indicated at the right (A).

Supplemental Figure 3

(A), Gene coding density plotted as a function of PMD frequency. **(B)**, Gene expression as a function of PMD frequency, as in main Figure 2F, but here restricted to the 24 cases overlapping between our WGBS cohort and the breast tumor (RNA-seq) transcriptomes cohort[16]. Top, gene expression; bottom, standard deviation. **(C)**, Somatic mutations plotted as a function of PMD frequency, as in main Figure 2G, but here restricted to the 25 cases overlapping between our WGBS cohort and the breast tumor full genomes cohort[14].

Supplemental Figure 4

(A), Expression change of non-CGI-promoter genes inside vs. outside of PMDs, as a function of PMD frequency. **(B)**, Expression change of non-CGI-promoter genes of tumor vs. normal, as a function of PMD frequency. From the TCGA breast cancer dataset, matched tumor/normal pairs were selected. PMD frequency for each gene was taken from our own dataset. **(C)**, Number of

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CGIs inside and outside of breast cancer PMDs. CGIs are classified as 'in' when inside a PMD in at least one tumor sample.

Supplemental Figure 5

(A), Expression change of TSGs/breast cancer driver mutated genes when inside PMDs. 31 of such genes are located inside PMDs in a subset of tumor samples. 'TSGs all cancers', genes annotated as TSGs regardless of cancer type; 'TSGs breast cancer', genes annotated as TSG in breast cancer; 'Nik-Zainal breast cancer driver mutations', genes with driver mutations in breast cancer[14]. **(B)**, Examples of genes from panel (A) being repressed when inside PMDs. Blue line, DNA methylation (WGBS); green bars, CGIs; red bars, PMDs. Gene expression (RNA-seq) of the corresponding gene is represented at the right of each panel. **(C)**, Pearson correlation between CGI-promoter methylation and expression. Gene classes are indicated as in panel (A). **(D)**, Expression changes (RNA-seq) of genes in panel (B), breast tumor vs. normal. Data is from an independent cohort (TCGA). Left panels, non-matched normal (n=88) and tumor samples (n=769); right panels, matched normal/tumor samples (n=86). p-values were calculated using a *t*-test.

Supplemental Figure 6

(A), Gene set enrichment analysis (GSEA) of genes downregulated when inside PMDs (>2.5 log2-fold, 400 genes, Suppl. Table 3). **(B)**, Examples of downregulated genes inside PMDs. CD3D encodes the gamma polypeptide of the T-cell receptor-CD3 complex (gene sets 'signalling', 'adhesion', and 'breast cancer luminal B down'); RBP4 encodes retinol binding protein 4 (gene set 'signalling', and 'breast cancer luminal B down'). Blue line, DNA methylation (WGBS); green bars, CGIs; red bars, PMDs. Gene expression (RNA-seq) of the corresponding gene is represented at the right of each panel. **(C)**, Overall survival of patient groups stratified according expression of the 400 PMD-downregulated genes (see Methods).

Supplemental Figure 7

(A), Boxplot summarizing mean PMD methylation of normal tissues and tumors of various tissues (summary of Fig. 4B). **(B)**, Distribution of CGI methylation, represented as the fraction of all CGIs (x-axis). Each horizontal bar represents one tumor sample (WGBS). Top panel, tumor

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

samples other than breast cancer (TCGA and ref.[29], abbreviations are given below); bottom panel, repeated from main Figure 3B for comparison.

Supplemental Table 1

Quality metrics and global methylation values from whole-genome bisulfite sequencing (WGBS) of 30 breast tumor samples from this study.

Supplemental Table 2

PMD frequency of all annotated CpG islands (. For each CGI, PMD frequency indicates the number of tumors in which the CGI is inside a detected PMD.

Supplemental Table 3

Genes that are downregulated when inside PMDs. 400 genes are downregulated at least 2.5 log2-fold.

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

REFERENCES

1. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301,** 89–92 (1983).

2. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).

3. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature genetics* **44,** 40–6 (2012).

4. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics* **43,** 768–775 (2011).

5. Hon, G. C. G. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome research* **22,** 246–258 (2012).

6. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510,** 537–41 (2014).

7. Schroeder, D. I., Lott, P., Korf, I. & LaSalle, J. M. Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Research* **21,** 1583–1591 (2011).

8. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nature reviews. Cancer* **13,** 497–510 (2013).

9. Schroeder, D. I. *et al.* The human placenta methylome. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 6037–42 (2013).

10. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–6 (2015).

11. Kulis, M. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nature genetics* **47,** 746–756 (2015).

12. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471,** 68–73 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

13. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (New York, N.Y.)* **341,** 1237905 (2013).

14. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 1–20 (2016).

15. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nature Communications* **7,** 11383 (2016).

16. Smid, M. *et al.* Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nature Communications* **7,** 12910 (2016).

17. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

18. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 139–144 (2010).

19. Rao, S. S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680 (2014).

20. Sexton, T. & Cavalli, G. The Role of Chromosome Domains in Shaping the Functional Genome. *Cell* **160,** 1049–1059 (2015).

21. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–80 (2012).

22. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell* **48,** 471–484 (2012).

23. Roukos, V. & Misteli, T. The biogenesis of chromosome translocations. *Nature cell biology* **16,** 293–300 (2014).

24. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

25. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Medical Sciences* **96,** 8681–8686 (1999).

26. Fang, F. *et al.* Breast Cancer Methylomes Establish an Epigenomic Foundation for Metastasis. *Science Translational Medicine* **3,** 75ra25 (2011).

27. Conway, K. *et al.* DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival. *Breast Cancer Research* **16,** 450 (2014).

28. Roessler, J. *et al.* The CpG island methylator phenotype in breast cancer is associated with the lobular subtype. *Epigenomics* **7,** 1–13 (2014).

29. Kretzmer, H. *et al.* DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nature genetics* **47,** 1316–25 (2015).

30. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17,** 510–522 (2010).

31. Smid, M. *et al.* Subtypes of breast cancer show preferential site of relapse. *Cancer Research* **68,** 3108–3114 (2008).

32. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–7 (2012).

33. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521,** 81–84 (2015).

34. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–4 (2015).

35. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

36. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

37. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163,** 1011–1025 (2015).

38. Holm, K. *et al.* An integrated genomics analysis of epigenetic subtypes in human breast tumors links DNA methylation patterns to chromatin states in normal mammary cells. *Breast Cancer Research* **18,** (2016).

39. Johann, P. D. *et al.* Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. *Cancer Cell* **29,** 379–393 (2016).

40. Burk, R. D. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543,** 378–384 (2017).

41. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171,** 540–556 (2017).

42. Raphael, B. J. *et al.* Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32,** 185–203.e13 (2017).

43. Ally, A. *et al.* Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169,** 1327–1341 (2017).

44. Habibi, E. *et al.* Whole-Genome Bisulfite Sequencing of Two Distinct Interconvertible DNA Methylomes of Mouse Embryonic Stem Cells. **13,** 360–369 (2013).

45. Le, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. of Statistical Software* **25,** 1–18 (2008).

46. Burger, L., Gaidatzis, D., Schübeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research* **41,** (2013).

47. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28,** 583–585 (2012).

48. Cribari-Neto, F. & Zeileis, A. Beta Regression in R. *Journal of Statistical Software* **34,** 1–24 (2010).

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

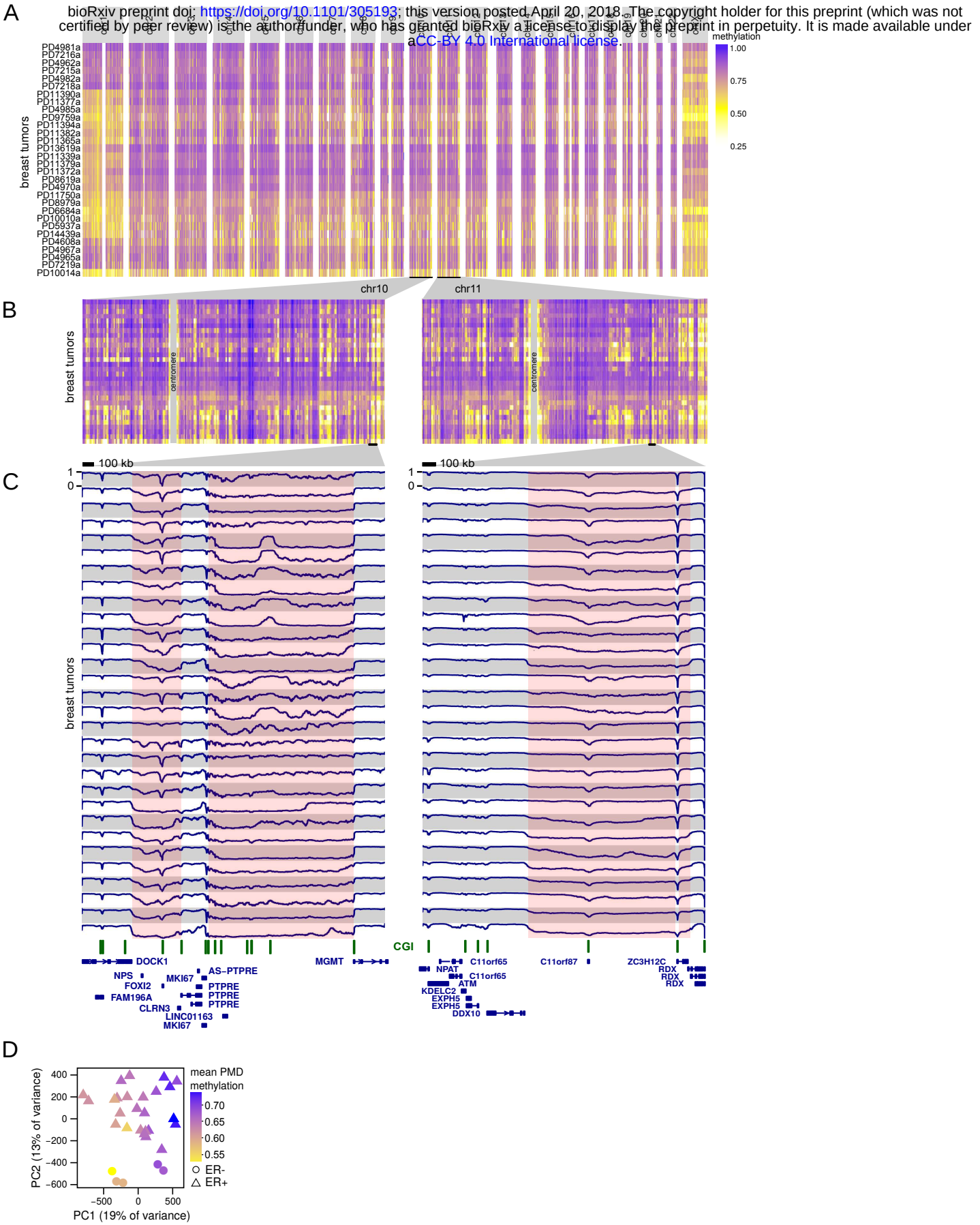


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

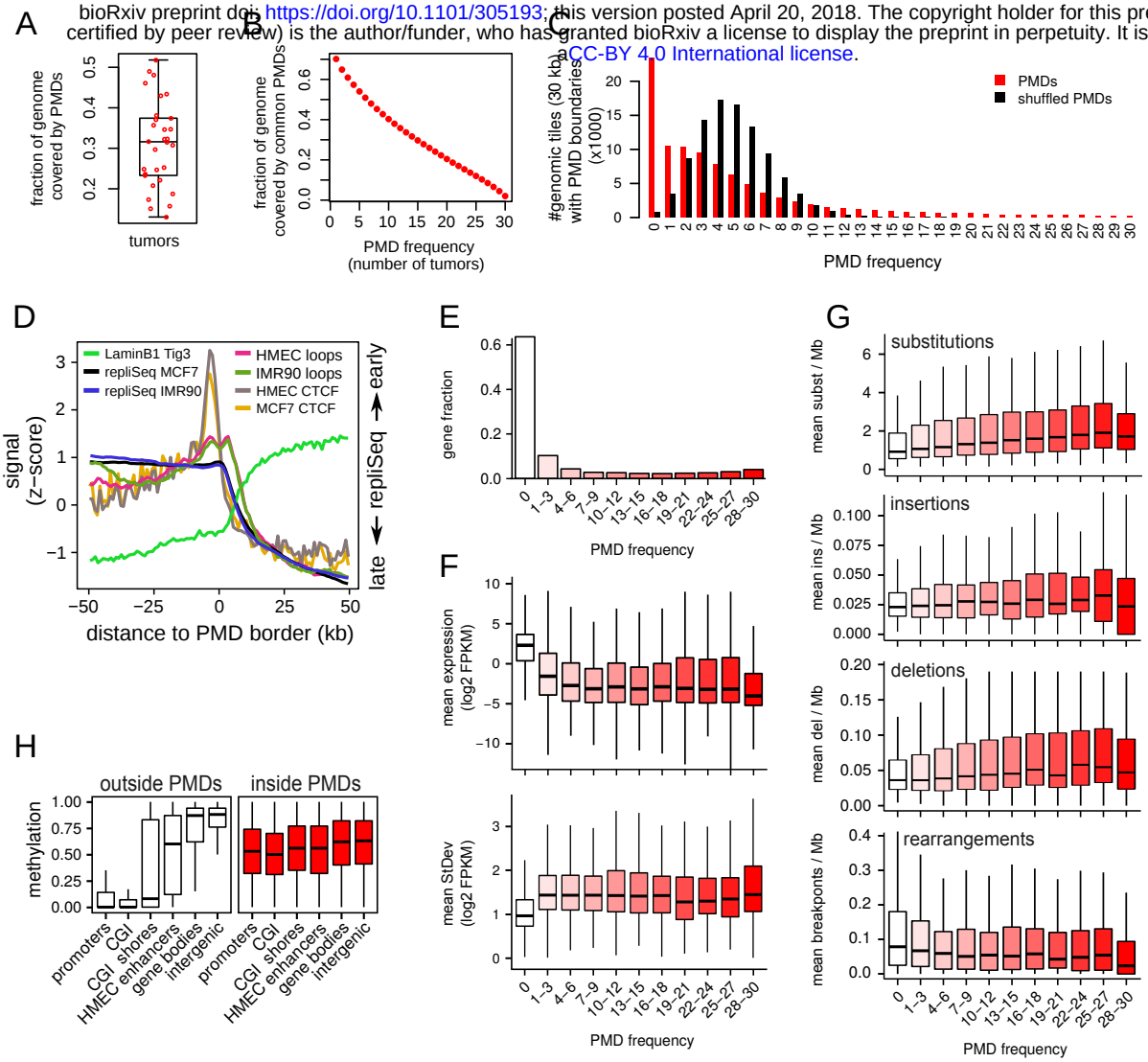


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

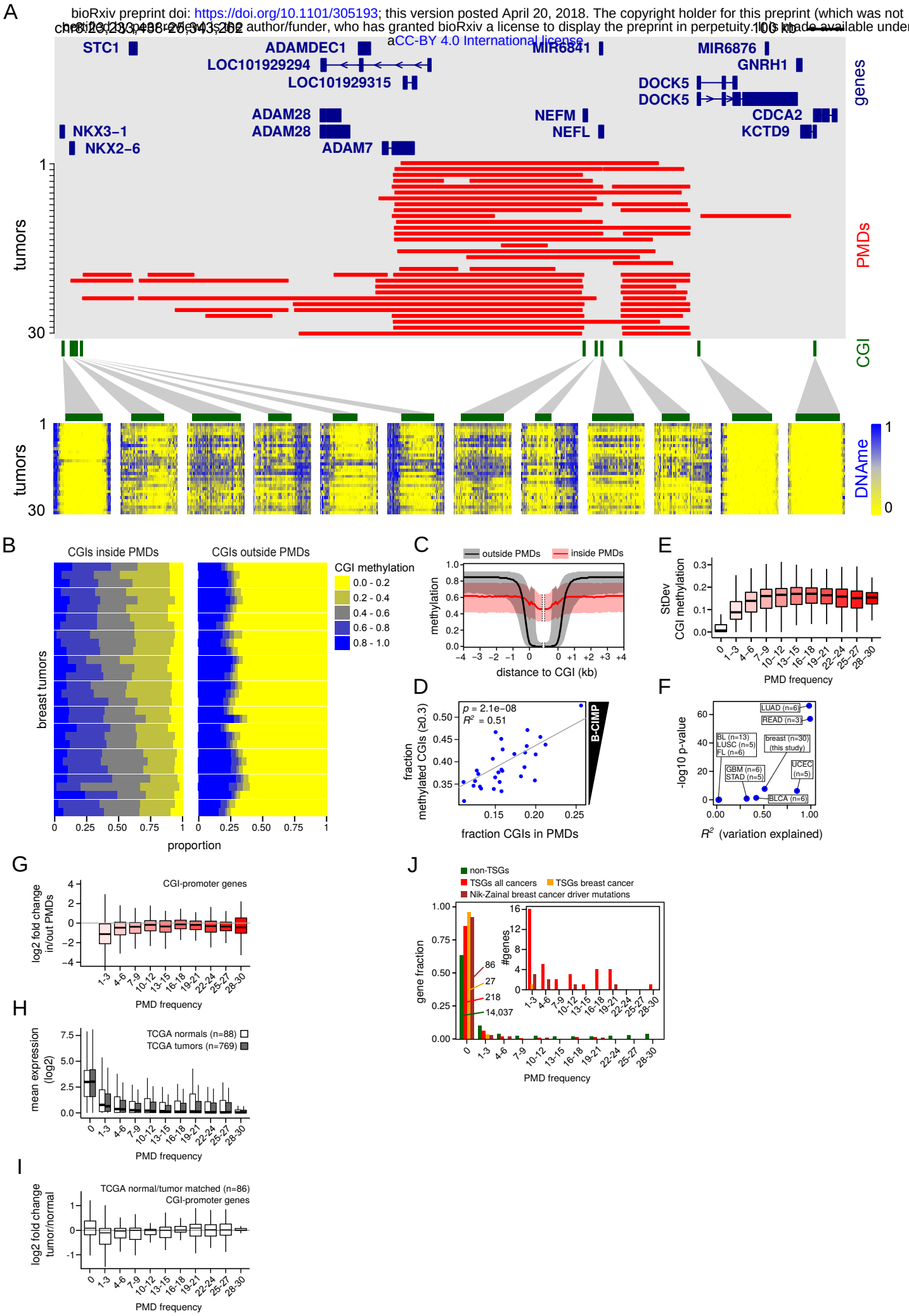
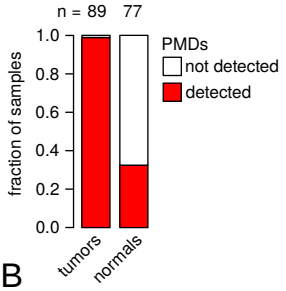


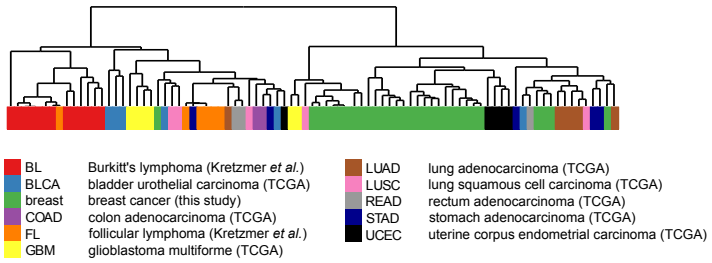
Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

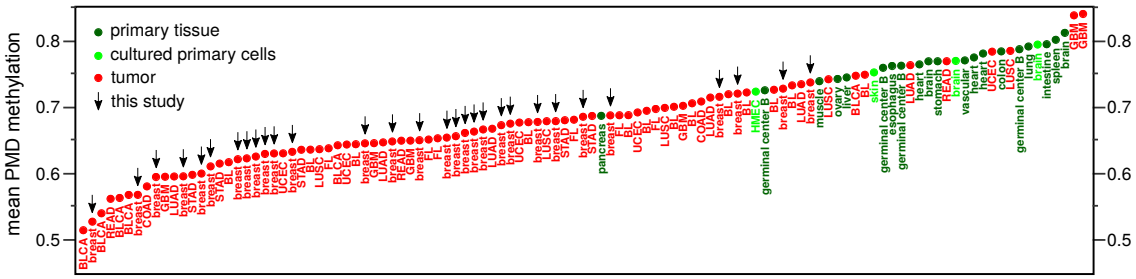
A



C



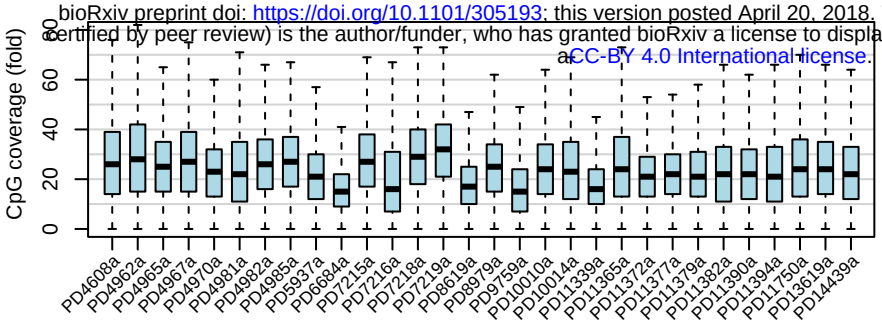
B



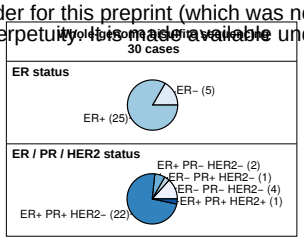
Supplementary Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

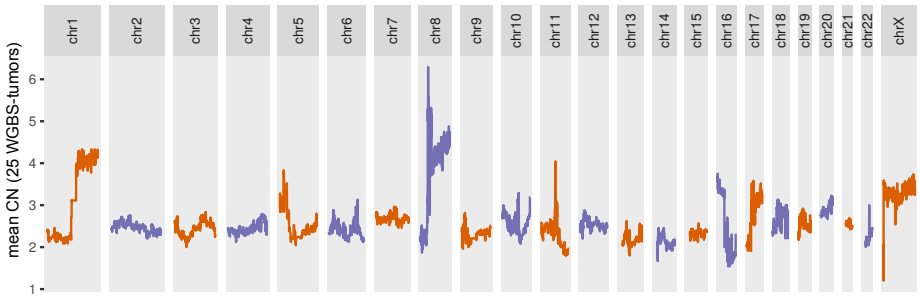
A



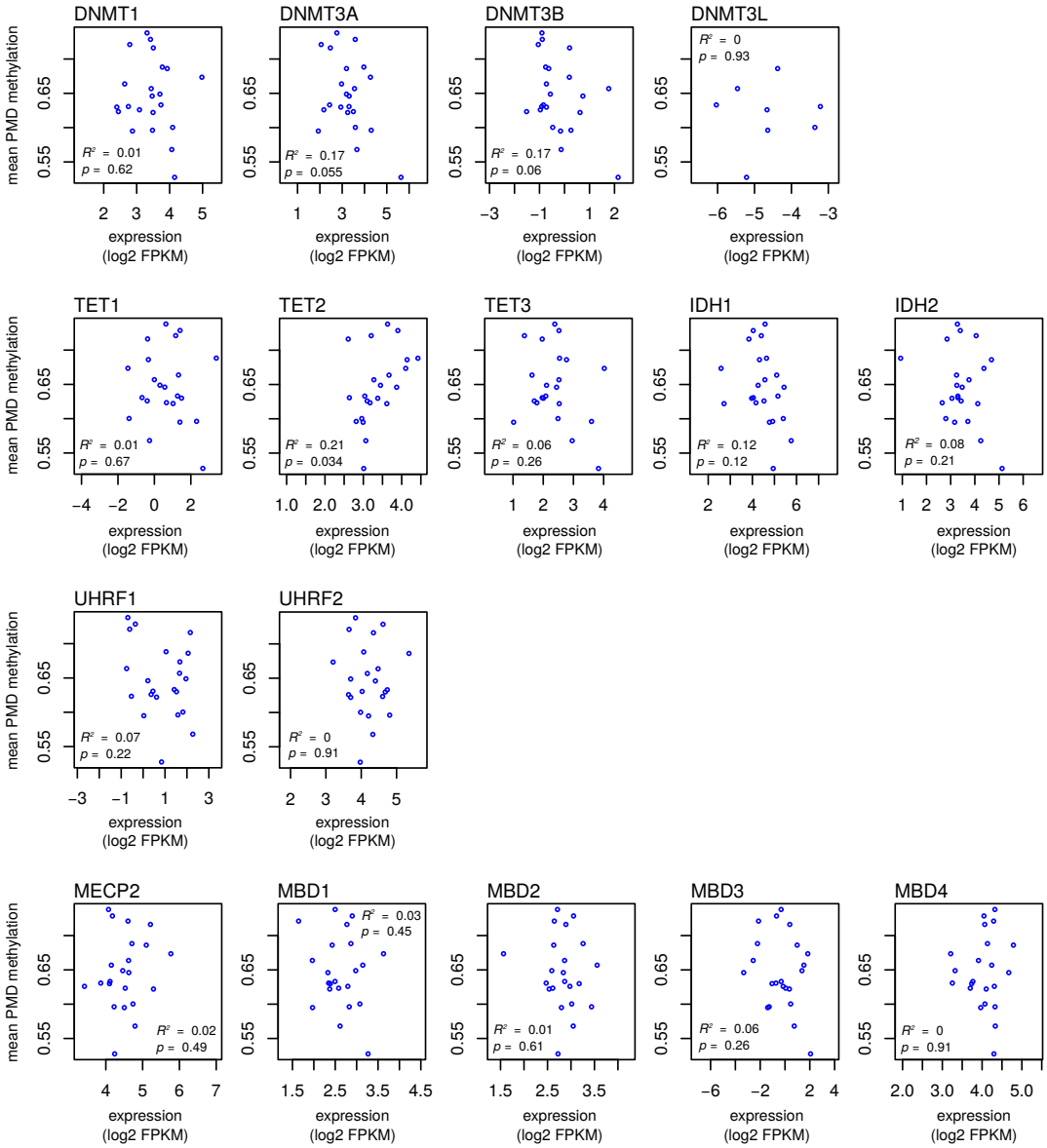
B



C



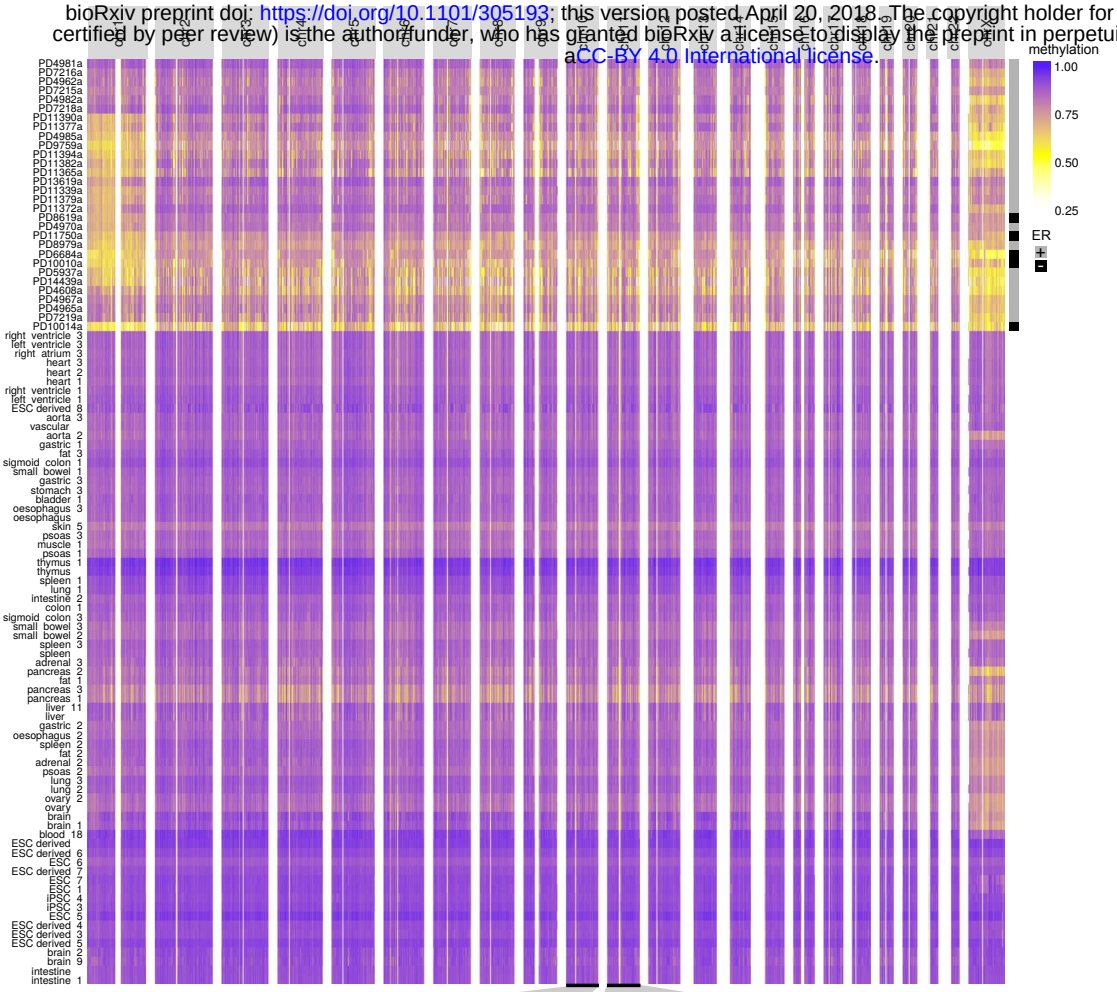
D



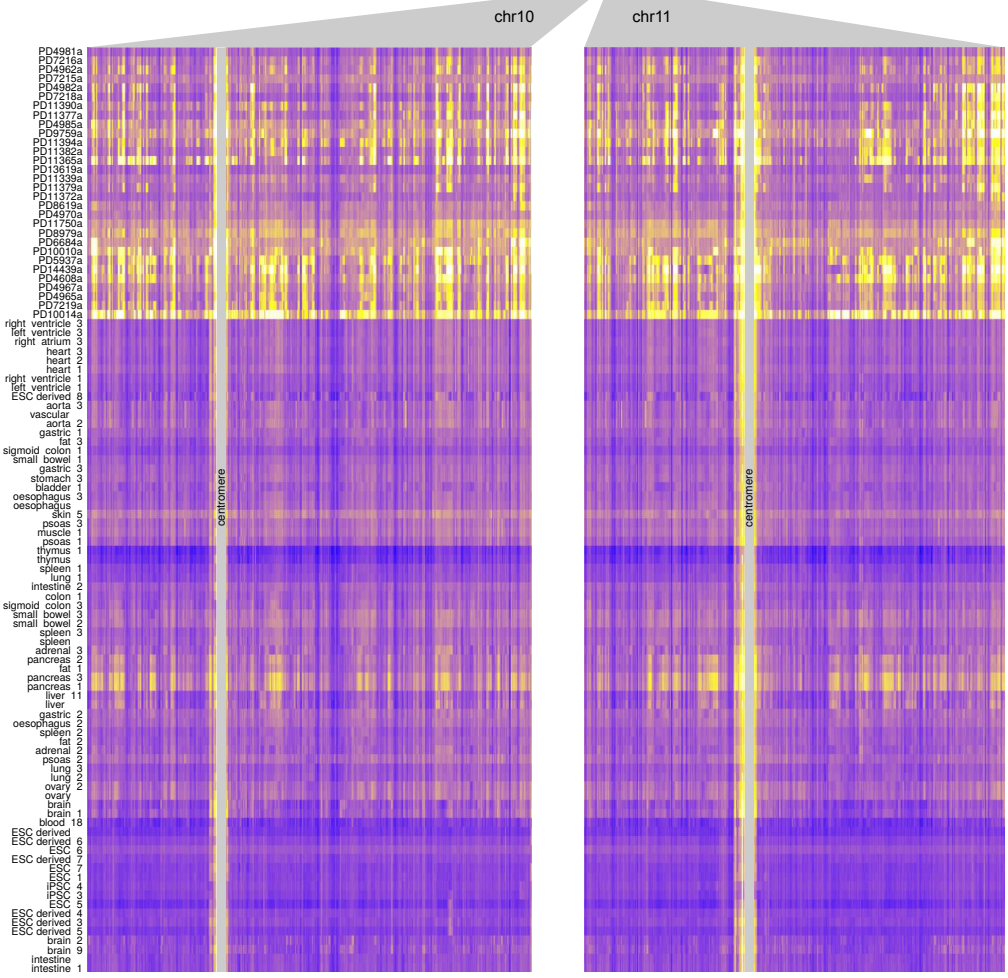
Supplemetary Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A

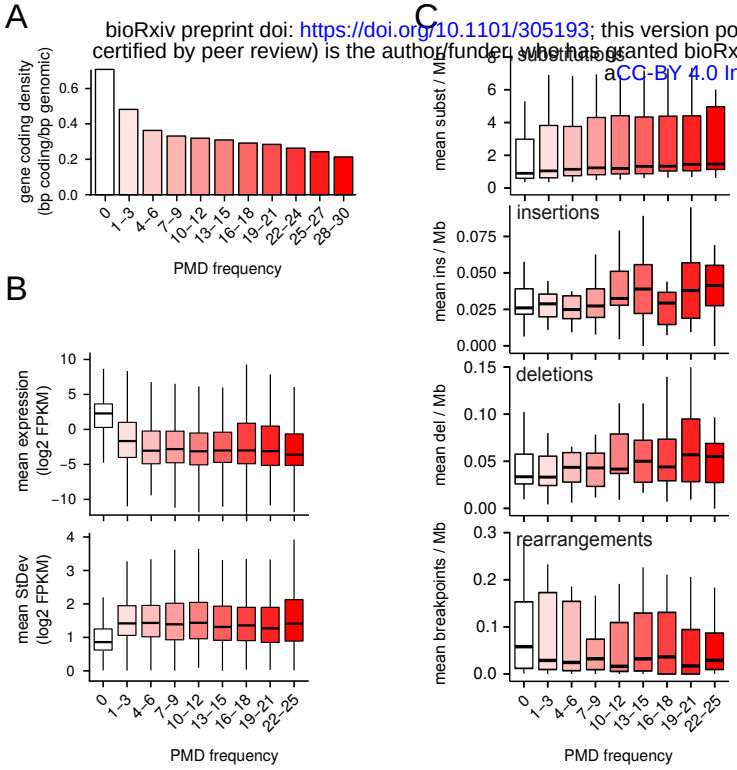


B



Supplementary Figure 3

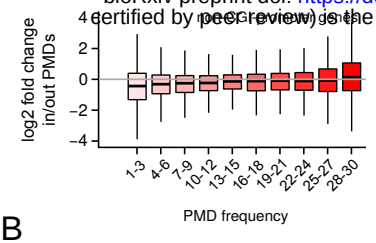
bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



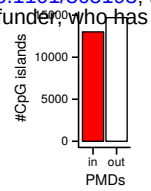
Supplementary Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

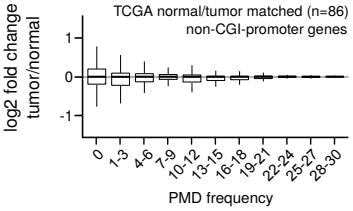
A



C



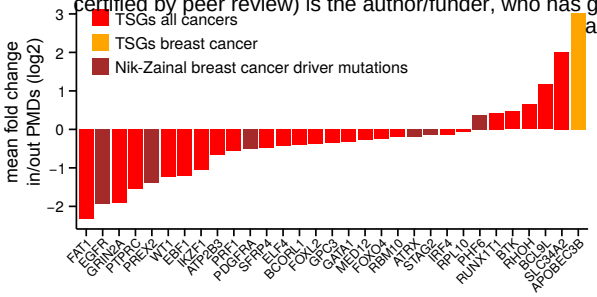
B



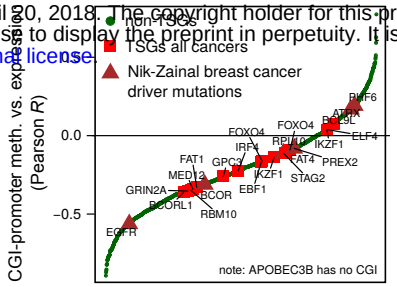
Supplementary Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

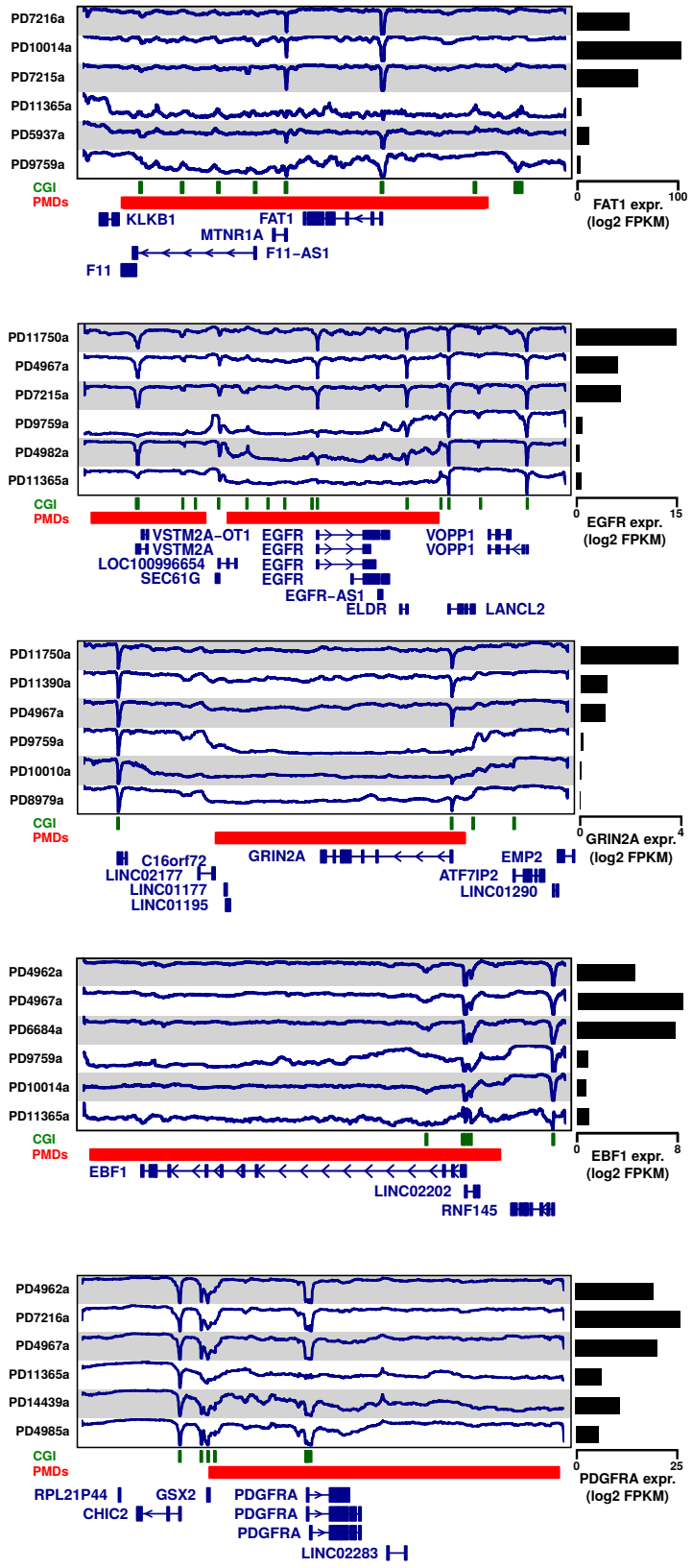
A



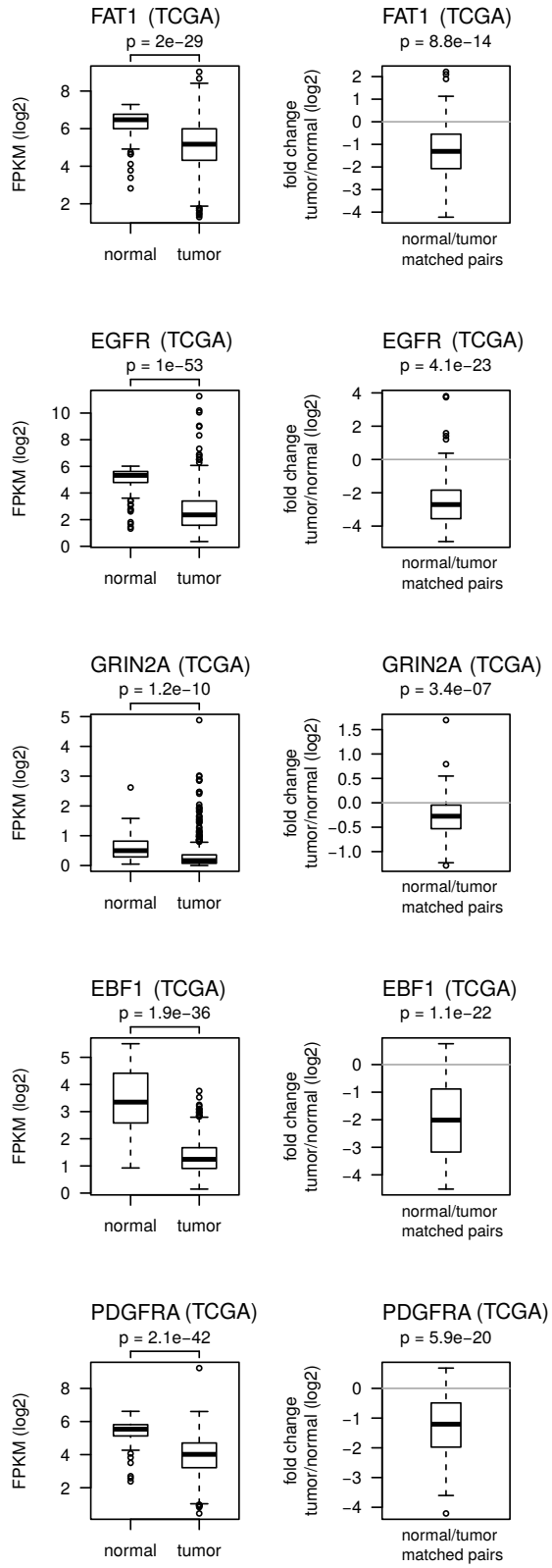
C



B



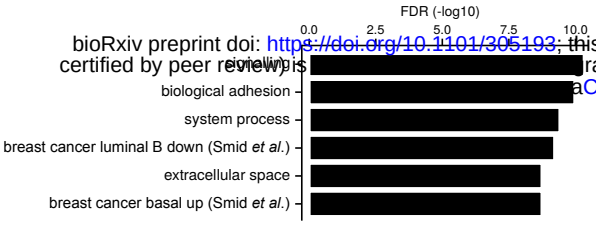
D



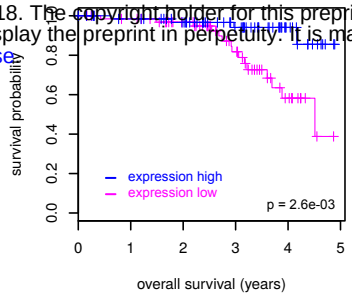
Supplementary Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is [illegible] granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

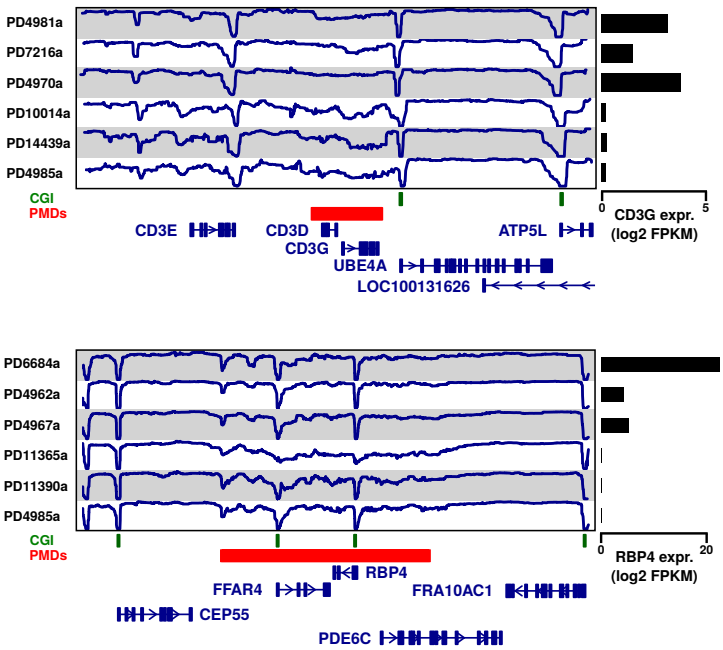
A



C



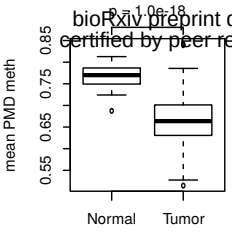
B



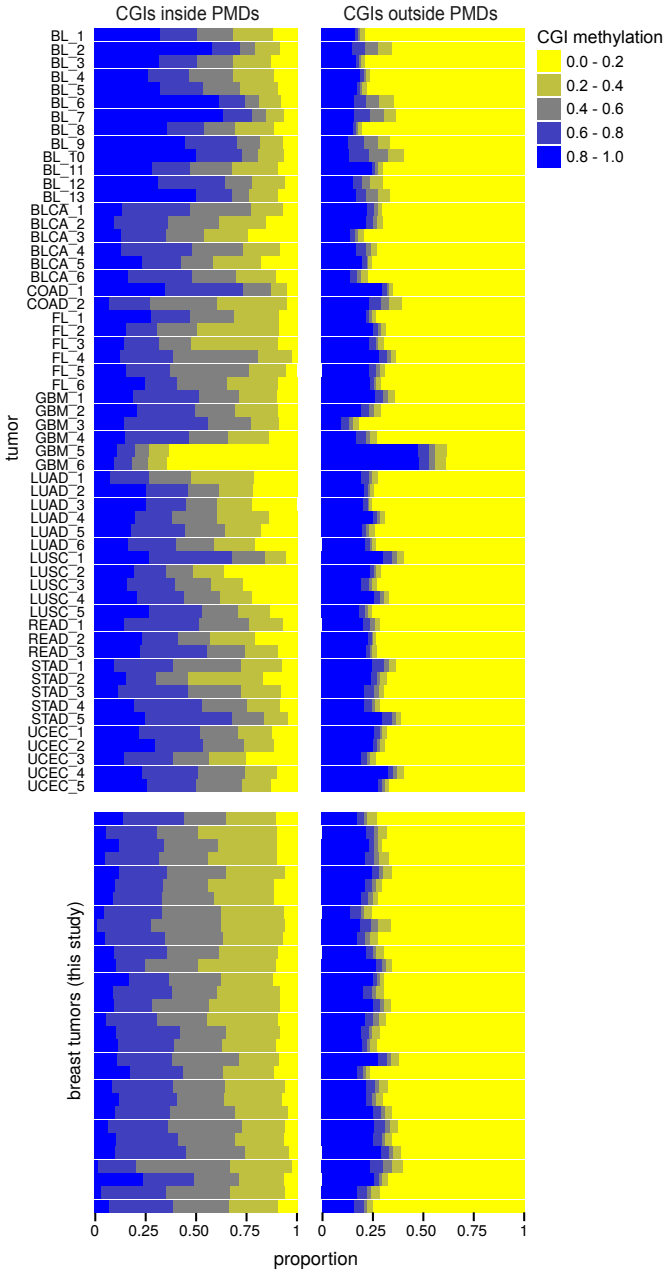
Supplementary Figure 7

bioRxiv preprint doi: https://doi.org/10.1101/305193; this version posted April 20, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A



B



BL	Burkitt's lymphoma (Kretzmer *et al.*)
BLCA	bladder urothelial carcinoma (TCGA)
COAD	colon adenocarcinoma (TCGA)
FL	follicular lymphoma (Kretzmer *et al.*)
GBM	glioblastoma multiforme (TCGA)
LUAD	lung adenocarcinoma (TCGA)
LUSC	lung squamous cell carcinoma (TCGA)
READ	rectum adenocarcinoma (TCGA)
STAD	stomach adenocarcinoma (TCGA)
UCEC	uterine corpus endometrial carcinoma (TCGA)