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

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)

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 normal tissues[2,7–12], but are less pronounced in pluripotent cells and brain tissue[12–14]. PMDs can comprise up to half of the genome[3,4,12], and it has been suggested that PMDs in different tissues are largely identical[3,12]. 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

79 cancer cell lines[5], but their extent and variation in primary tumors is hitherto unknown. A major
80 limitation of most DNA methylation studies is that only a small subset of CpGs are interrogated.
81 This prevents accurate determination of the extent and location of PMDs. Few samples of a
82 certain tissue/tumor have typically been analyzed using whole-genome bisulfite sequencing
83 (WGBS). Thus, observations cannot be extrapolated to individual cancer types. Here, we
84 analyzed DNA methylation profiles of 30 primary breast tumors at high resolution through
85 WGBSs. This allowed us to delineate breast cancer PMD characteristics in detail. We show that
86 PMDs define breast cancer methylomes and are linked to other key epigenetic aberrations such
87 as CGI hypermethylation.

88

89 **RESULTS**

90 **Primary breast tumors display variable loss of DNA methylation**

91 To study breast cancer epigenomes we performed WGBS encompassing ~95% of annotated
92 CpGs (Suppl. Fig. 1A, Suppl. Table 1). For 25/30 of these tumors we previously analyzed their
93 full genomes[15,16] and transcriptomes[17], respectively. Of the 30 tumors, 25 and 5 are ER-positive
94 and ER-negative, respectively (Suppl. Fig. 1B).

95 To globally inspect aberrations in DNA methylation patterns we generated genome-wide and
96 chromosome-wide methylome maps by displaying mean methylation in consecutive tiles of 10
97 kb (see Methods). These maps revealed extensive inter-tumor variation at genome-wide scale
98 (Fig. 1A). At chromosome level, we observed stably hypermethylated regions next to regions
99 that were hypomethylated to various extents and across tumors (Fig. 1B). Chromosomes 1 and X
100 were exceptionally prone to methylation loss, the latter of which may be related to epigenetic
101 aberrations of the inactive X-chromosome in breast cancer observed by others[18]. At megabase
102 scale (Fig. 1C) DNA methylation profiles showed that the widespread loss of methylation
103 occurred in block-like structures previously defined as PMDs[2]. Across primary breast tumor
104 samples, DNA methylation levels and genomic sizes of PMDs differ extensively between tumors
105 and PMDs do appear as separate units in some tumors and as merged or extended in others,
106 underscoring the high variation with which methylation loss occurs. Despite this variation,
107 however, we observed common PMD boundaries as well.

108 Given the variation between tumors, we asked whether the patterns of methylation loss were
109 associated with distribution of copy-number variations (CNVs) throughout the genome. We
110 found no evidence for such association (Pearson R=0.17), although we noticed that chromosomes
111 with the most pronounced loss of methylation (chr1, chrX, chr8-p) frequently contained
112 amplifications (Suppl. Fig. 1C). Next, we asked whether loss of methylation was associated with
113 aberrant expression of genes involved in writing, erasing, or reading the 5-methylcytosine
114 modification. However, we found no such correlation (Suppl. Fig. 1D). Finally, we assessed
115 whether mean PMD methylation was associated with the fraction of aberrant cells within the
116 sample (ASCAT[19]). However, no such correlation was evident (Pearson R = -0.03, Suppl. Fig.
117 1E).

118 To provide a reference for the observed patterns of methylation loss we compared WGBS
119 profiles of primary breast tumors to that of 72 normal tissues (WGBS profiles from Roadmap
120 Epigenomics Project and[10], Suppl. Fig. 2A,B). In sharp contrast to breast cancer, most normal
121 tissues were almost fully hypermethylated (except for pancreas and skin), with heart, thymus,
122 embryonic stem cell(-derived), induced pluripotent stem cells and brain having the highest levels
123 of methylation. Importantly, inter-tissue variation was much lower as compared to breast tumors
124 ($p < 2.2e-16$, MWU-test on standard deviations). The variation observed among breast tumors
125 was also present when we reproduced Fig. 1ABC using only 'solo-WCGW' CpGs (CpGs
126 flanked by an A or T on both sides), which were recently shown to be more prone to PMD
127 hypomethylation[12] (Suppl. Fig. 3). Thus, breast tumors show widespread loss of DNA
128 methylation in PMDs, and the extent and patterns appear to be hypervariable between tumor
129 samples. In line with this, principal component analysis confirmed that methylation inside PMDs
130 is the primary source of variation across full-genome breast cancer DNA methylation profiles
131 (Fig. 1D): the first principal component (PC1) is strongly associated with mean PMD
132 methylation (p=6.8e-07, ref.[20], see Methods). The second-largest source of variation, PC2, is
133 associated with ER status (p=1.9e-06, Fig. 1D, see Methods) and to a lesser extent with
134 'intrinsic' AIMS subtypes (Absolute assignment of breast cancer Intrinsic Molecular Subtypes,
135 Suppl. Fig. 4A)[21,22], although the latter is likely confounded with ER status. Successive PCs were
136 not significantly associated with any clinicopathological feature. It should be noted that with 30
137 tumors only very strong associations can achieve statistical significance. Taken together, breast
138 tumor whole-genome DNA methylation profiles reveal global loss of methylation in features

139 known as PMDs, the extent of which is hypervariable across tumors and represent the major
140 source of variation between tumors.

141

142 **Distribution and characteristics of breast cancer PMDs**

143 We set out to further characterize breast cancer PMDs and their variation (see Methods: data
144 access). The genome fraction covered by PMDs varies greatly across our WGBS cohort of 30
145 tumors, ranging between 10% and 50% across tumors, covering 32% of the genome on average
146 (Fig. 2A). We define 'PMD frequency' as the number of tumors in which a PMD is detected. A
147 PMD frequency of 30 (PMDs common to all 30 cases) occurs in only a very small fraction of the
148 genome (2%), while a PMD frequency of 1 (representing the union of all PMDs from 30 cases)
149 involves 70.2% of the genome (Fig. 2B). Similar results were obtained with PMDs called on
150 only solo-WCGW CpGs[12] (Suppl. Fig. 4BC), and comparison of these 'solo-CpG' PMDs with
151 'all-CpG' PMDs revealed high overlap (92%) between their individual unions (Suppl. Fig. 4D).
152 We further compared our PMD calling with 'aggregate PMD calling' based on cross-sample
153 standard deviation (s.d.) of methylation in 100-kb genomic bins[12]. This method segments the
154 genome according to common PMDs across multiple samples, and we found that our PMDs are
155 all contained within this aggregate PMD track (Suppl. Fig. 4EF).

156 Given the inter-tumor variation of PMDs we tested to which extent PMD distribution is random
157 by counting PMD borders in 30-kb genomic tiles (Fig. 2C). Randomly shuffled PMDs yield a
158 normal distribution centered at a PMD frequency of four. In contrast, observed PMDs show a
159 skewed distribution: the mode was for a PMD frequency of 0 suggesting that many tiles (23,492,
160 25%) do not coincide with any PMD borders. The majority of tiles (62%) had a low PMD border
161 frequency (1-10). The tail represents low numbers of tiles with up to maximal PMD frequency of
162 30. We conclude that PMD distribution is not random: part of the genome appears not to tolerate
163 PMDs while PMDs occur in a large fraction of the genome with varying frequencies.

164 PMDs have been shown to coincide with lamin-associated domains (LADs)[3,4]: large repressive
165 domains that preferentially locate to the nuclear periphery[23]. LADs are characterized by low gene
166 density and late replication[23,24]. Accordingly we found that PMDs show reduced gene densities
167 (Fig. 2E), have high LaminB1 signals (associated with LADs[23], Fig. 2D), are late replicating

168 (ENCODE data, Fig. 2D) and have a low frequency of (Hi-C) 3D loops[25], an indicator of lower
169 levels of transcription. Finally, we observed a local increase in binding of the transcription factor
170 CTCF at the borders of PMDs (Fig. 2D) as shown in previous reports[3,23,26–28].

171 We previously analyzed the full transcriptomes (RNA-seq) in a breast cancer cohort of 266
172 cases[17] from which our WGBS cohort is a subset. We determined the mean expression of genes
173 as a function of PMD frequency in the overlapping subset of 25 tumors. Genes inside PMDs are
174 expressed at consistently lower levels than genes outside of PMDs (Fig. 2F, $p < 2.2e-16$, *t*-test),
175 with a tendency towards lower expression in highly-frequent PMDs ($p < 2.2e-16$, linear
176 regression). Given the variable nature of DNA methylation patterns of PMDs, we also
177 determined the variation (s.d.) in gene expression as a function of PMD frequency and found
178 higher variation for genes inside PMDs (Fig. 2F, $p < 2.2e-16$, MWU-test). When extending this
179 analysis to the full set of 266 cases from the transcriptome cohort we observed the same (Suppl.
180 Fig. 5A, $p < 2.2e-16$, *t*-test for expression; $p < 2.2e-16$, MWU-test for variation). Given the
181 observed variability of DNA methylation and gene expression inside PMDs, we asked whether
182 genetic stability, i.e. the number of somatic mutations, was also altered within PMDs. In the 25
183 overlapping cases between our WGBS cohort and the WGS cohort[15], substitutions, insertions,
184 and deletions occur more frequently within than outside PMDs ($p < 0.0005$ for each mutation
185 type, logistic regression), with a (slight) increase in highly frequent PMDs ($p < 2.2e-16$ for
186 substitutions, $p = 0.37$ for insertions, $p = 1.6e-05$ for deletions, logistic regression, Fig. 2G). In
187 contrast, rearrangements are more abundant outside of PMDs ($p = 1.1e-09$, logistic regression),
188 in keeping with the hypothesis that regions with higher transcriptional activity are more
189 susceptible to translocations[29]. We extended this analysis to the full cohort of 560 WGS tumor
190 samples[15], which confirmed these observations while showing much stronger effects in highly
191 frequent PMDs ($p < 2.2e-16$ for all mutation types and rearrangements, logistic regression, Suppl.
192 Fig. 5B). Taken together, breast cancer PMDs share key features of PMDs including low gene
193 density, low gene expression, and colocalization with LADs, suggesting that they reside in the
194 'B' (inactive) compartment of the genome[30]. Importantly, in addition to epigenomic instability,
195 breast cancer PMDs also tolerate transcriptomic variability and genomic instability.

196 **Relationship between CpG island methylation and PMDs in breast cancer**

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

207 We further focused on methylation levels of CGIs. When indiviual PMDs are regarded, CGIs
208 inside of them lose their strictly hypomethylated state and become more methylated to a degree
209 that varies between tumors (Fig. 3A). Across all tumors and all CGIs, this effect is extensive (Fig.
210 3B,C), affecting virtually all CGIs inside PMDs: on average 92% of CGIs lose their
211 hypomethylated state and gain some level of methylation (Fig. 3B, left panel). Outside of PMDs
212 only 25-30% of the CGIs is hypermethylated, although to a higher level (Fig. 3B, right panel).
213 Thus, incognizant deposition of DNA methylation inside PMDs results in extensive
214 hypermethylation of virtually all PMD-CGIs.

215 Concurrent hypermethylation of CGIs in cancer has been termed CIMP[31], and in breast cancer
216 this phenomenon has been termed B-CIMP[32–34]. To determine whether CIMP is directly related
217 to PMD variation we defined B-CIMP as the fraction of CGIs that are hypermethylated (>30%
218 methylated), and determined its association with the fraction of CGIs inside PMDs. Regression
219 analysis (see Methods) showed that this association is highly significant (Fig. 3F, p=2.1e-08,
220 R^2=0.51, n=30). The fraction of hypermethylated CGIs is generally higher than the fraction of
221 hypermethylated CGIs in PMDs, suggesting that CGI hypermethylation is not solely dependent
222 on PMD occurrence. However, CGI methylation levels outside PMDs are far more stable than
223 inside PMDs (Fig. 3E), which likely represents an invariably methylated set of CGIs (Suppl.
224 Table 2).

225 We applied the same regression analysis to 14 other tumor types (TCGA[35], BLUEPRINT[36],
226 refs.[37,38], Fig. 3G). Although sample sizes were small, we found significant CIMP-PMD
227 associations for lung adenocarcinoma (LUAD), rectum adenocarcinoma (READ), uterine corpus
228 endometrial carcinoma (UCEC) and bladder urothelial carcinoma (BLCA). We did not find
229 significant associations for other tumor types (ALL, BL, ALL, CLL, FL, LUSC, lung, TPL,
230 STAD, MCL, BLCA, see Fig. 4B for their abbreviations) and glioblastoma (GBM), even though
231 for the latter G-CIMP has been previously described[39]. Taken together, we conclude that PMD
232 occurrence is an important determinant for CIMP in breast cancer and a subset of other tumor
233 types.

234

235 **PMD demethylation effects on gene expression**

236 To assess whether widespread hypermethylation of CGI-promoters within PMDs instigates gene
237 repression we analyzed expression as a function of gene location inside or outside of PMDs.
238 Overall, CGI-promoter genes showed a mild but significant downregulation when inside PMDs
239 (p=4.5e-12, *t*-test), while strong downregulation was specifically restricted to low-frequency
240 PMDs (Fig. 3H). For non-CGI-promoter genes this trend was very weak or absent (Suppl. Fig.
241 6A). As healthy controls were not included in transcriptome analysis of our cohort[17] we used
242 gene expression (RNA-seq) profiles from breast tumors (769) and normal controls (88) from
243 TCGA. Similar to our cohort (see Fig. 2F) we found that overall gene expression for the TCGA
244 tumors is lower inside PMDs, with lowest expression for genes inside high-frequent PMDs (Fig.
245 3I, $p < 2.2e-16$, linear regression). However, the expression of genes in tumor PMDs is very
246 similar to healthy control samples ($p = 0.807$, linear regression). To analyze this in more detail
247 we selected normal/tumor matched pairs (i.e. from the same individuals, n=86) and analyzed the
248 fold change over the different PMD frequencies (Fig. 3J). As in our cohort, downregulation is
249 restricted to genes with low PMD-frequency ($p < 2.2e-16$ for PMD frequency 1-3, linear
250 regression). No obvious changes occur in high-frequency PMD genes, nor in non-CGI-promoter
251 genes (Suppl. Fig. 6B). Taken together, widespread cancer-associated repression of all genes
252 inside PMDs is limited: downregulation is restricted to low-frequency (i.e. the more variable)
253 PMDs and affects only CGI-promoter genes, which undergo widespread hypermethylation inside
254 PMDs.

255 Given the widely accepted model of hypermethylated promoter-CGIs causing repression of
256 tumor suppressor genes (TSGs) we determined whether breast cancer PMDs overlap with these
257 genes to instigate such repression. For non-TSGs as a reference we found that 64% (14,037) are
258 located outside of PMDs (Fig. 3K), while 36% are located inside, (see also Fig. 2E). Strikingly,
259 TSGs (Cancer Gene Census) overlap poorly with PMDs: most TSGs (218/254, 86%) are located
260 outside of PMDs. Only 14% overlap with mostly low-frequency PMDs, implying exclusion of
261 TSGs from PMDs (p=8.8e-16, hypergeometric test). When we specifically focused on breast
262 cancer-related TSGs (Cancer Gene Census), this exclusion was even stronger: practically all
263 (27/28, 96%) breast cancer TSGs are located outside of PMDs (p=3.5e-06, hypergeometric test).
264 Similarly, from our previously identified set of genes containing breast cancer driver mutations[15]:
265 86/93 (92%) were located outside of PMDs (p=2.0e-11, hypergeometric test). Alltoghether, only
266 31 breast cancer-mutated genes were not excluded from PMDs. We assessed whether these genes
267 are downregulated in tumors when inside PMDs. 24/31 (74%) genes were downregulated (Suppl.
268 Fig. 7A,B), and an overall negative correlation between CGI-promoter methylation and
269 expression was evident (Suppl. Fig. 7C). For 16 out of these 24 genes we confirmed that
270 significant downregulation also takes place in cancer relative to normal in an independent breast
271 cancer expression dataset (TCGA, Suppl. Fig. 7D and data not shown). Among the
272 downregulated genes in PMDs are EGFR (epidermal growth factor receptor) and PDGFRA
273 (platelet-derived growth factor receptor α) that have tumor promoting mutations (Suppl. Fig.
274 7A,B,C). Paradoxically, both genes are significantly downregulated in our as well as the TCGA
275 breast cancer dataset (Suppl. Fig. 7D). Taken together, despite the large number of
276 hypermethylated CpG islands inside breast cancer PMDs (13,013 CGIs; 47%, Fig. 3D), these
277 CGIs do not generally co-occur with TSGs and other breast cancer-relevant genes. Repression of
278 these genes through classical promoter-hypermethylation in PMDs does not occur at large scale,
279 and is likely limited to a few genes.

280 We next identified genes that are downregulated when inside PMDs regardless of any
281 documented TSG function or mutation in breast cancer. 400 genes were downregulated at least
282 2.5 log2-fold (Suppl. Table 3). Gene set enrichment analysis showed that these genes were
283 involved in processes such as signaling and adhesion (Suppl. Fig. 8A). In addition, there is a
284 significant enrichment of genes downregulated in luminal B breast cancer (and upregulated in
285 basal breast cancer)[40]. This suggests that PMDs are involved in downregulation of luminal B-

286 specific genes. Examples of luminal B-downregulated genes include CD3G, encoding the
287 gamma polypeptide of the T-cell receptor-CD3 complex (gene sets 'signaling' and 'adhesion'),
288 and RBP4, encoding retinol binding protein 4 (gene set 'signaling') (Suppl. Fig. 8B).
289 Stratification of tumors according to low and high median expression of the 400 PMD-
290 downregulated genes revealed significant differences in overall survival of the corresponding
291 patients (p=2.6e-03, *chi*-square test, Suppl. Fig. 8C), suggesting clinical significance of PMD-
292 associated gene repression. Taken together, downregulation of genes inside PMDs occurs rarely
293 and is restricted to low-frequency PMDs. However, these rare cases include genes relevant to
294 breast cancer given the overlap with previously identified luminal B breast cancer-relevant genes
295 and differential overall survival. We finally focused on expression changes of X-linked genes,
296 since the X-chromosome is exceptionally prone to methylation loss (Fig. 1A, Suppl. 3A). To
297 assess whether this is associated with altered expression of genes involved in the process of X-
298 inactivation (XCI) we regarded XIST and genes encoding PRC2 subunits. Multivariate
299 regression revealed that expression of XIST, EED, and EZH1/2 is associated with the fraction of
300 chrX inside PMDs (p = 4.8e-05, Suppl. Fig. 6CD). To further analyze the effect of PMDs on
301 expression on X-linked genes we stratified X-linked genes according their consensus X-
302 inactivation status (E, escape; S, subject to XCI; VE, variably escaping; PAR, pseudoautosomal
303 region)[41]. Notably, among these categories, escape (E) genes are strongly affected when inside
304 PMDs (Fig. 3L), suggesting a specific sensitivity of escape genes to become repressed when
305 inside PMDs. This was unrelated to altered copy number status of these genes (Suppl. Fig. 6E,
306 see also Suppl. Fig. 1C). Taken together, the fraction of chrX inside PMDs is associated with
307 expression levels of key XCI inactivation genes, and escape genes are specifically sensitive to
308 repression inside X-linked PMDs.

309

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

312 To assess the generality of PMD occurrence in cancer, we extended our analysis to other cancer
313 types and normal tissues. We performed PMD detection in a total of 320 WGBS profiles (133
314 tumors and 187 normals, from TCGA[35], BLUEPRINT[36], the Roadmap Epigenomics Project
315 (http://www.roadmapepigenomics.org), refs.[10,37,38]). Although PMDs are detectable in virtually

316 all tumors and normal tissues (see Methods: data access), mean DNA methylation inside PMDs
317 is much lower in tumors as compared to normal tissues (Fig. 4A, Suppl. Fig. 9A, $p < 2.2e-16$, *t*-
318 test). PMD methylation levels are not tumor tissue-type specific, as most types display the same
319 range of PMD methylation. However, some tumor tissue types have exceptional low methylation
320 inside PMDs (bladder urothelial carcinoma (BLCA), lung), or lack any loss of methylation
321 (glioblastoma multiforme (GBM), acute lymphoblastic leukemia (ALL), and acute myeloid
322 leukemia (ALL)). Thus, regardless of these extreme cases, absolute levels of PMD methylation
323 do not typify tumor tissue origin, underscoring the variable nature of methylation within PMDs.
324 To assess whether CGI hypermethylation in PMDs is as extensive in these additional tumor types
325 as in breast cancer, we analyzed CGI methylation of these 103 additional tumor samples (Suppl.
326 Fig. 9B, see Methods: data access). As in breast cancer, extensive hypermethylation of CGIs
327 inside PMDs was consistent in most tumor types, with levels of hypermethylation in Burkitt's
328 lymphoma (BL)[37] being among the highest of all tested tumors Possibly, these differences are
329 linked to tumor cellularity of the samples. In two GBM and some AML samples, CGI
330 hypermethylation was not restricted to PMDs, which is suggestive of inaccurate PMD detection
331 due to high methylation inside these tumors' PMDs (see Fig. 4A). Importantly, these results
332 extend the observed tendency of CGI hypermethylation inside PMDs to other tumors.

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

340

341 **DISCUSSION**

342 In this study we analyzed breast cancer DNA methylation profiles to high resolution. The main
343 feature of breast cancer epigenomes is the extensive loss of methylation in PMDs and their
344 hypervariability. Directly linked to this is the concurrent CGI hypermethylation, which inside

PMDs affects 92% of all CGIs. 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 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[42–44]. The phenomena observed in breast cancer extend to tumors of at least 16 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 AML, ALL and GBM.

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.[39,45–53]). 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[46,50–52]. 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.

370 **METHODS**

371 **Data access and code availability**

372 Tables containing CpG methylation values (bigwig), genomic coordinates and mean methylation
373 values of PMDs and CGIs are available via DOI 10.5281/zenodo.1467025 or DOI
374 10.17026/dans-276-sda6. Raw data for whole-genome bisulfite sequencing of the 30 breast
375 tumor samples of this study is available from the European Genome-phenome Archive
376 (https://www.ebi.ac.uk/ega) under dataset accession EGAD00001001388. All code for analyses
377 of this study is available on https://github.com/abbrinkman/brcancer_wgbs.git.

378 **Sample selection, pathology review and clinical data collection**

379 Sample selection, pathology review and clinical data collection for this study has been described
380 in[15].

381 **Processing of whole-genome bisulfite sequencing data**

382 WGBS library preparation, read mapping, and methylation calling was done as described
383 before[54]. The genome build used for mapping of bisulfite sequencing reads, and throughout this
384 study was hg19 (GRCh37).

385 **Principal component analysis of WGBS data**

386 For principal component analysis (PCA) of WGBS profiles, CpGs with coverage of at least 10
387 were used. Subsequently, the top 5% most variable CpGs were selected. We used the
388 FactoMineR package[20] for R to perform PCA, to determine association of principal components
389 with clinicopathological features, and to perform the corresponding significance testing.

390 **Detection of PMDs**

391 Detection of partially methylated domains (PMDs) in all methylation profiles throughout this
392 study was done using the MethylSeekR package for R[55]. Before PMD calling, CpGs overlapping
393 common SNPs (dbSNP build 137) were removed. The alpha distribution[55] was used to determine
394 whether PMDs were present at all, along with visual inspection of WGBS profiles. After PMD

395 calling, the resulting PMDs were further filtered by removing regions overlapping with
396 centromers (undetermined sequence content).

397 **Mean methylation in PMDs and genomic tiles**

398 Wherever mean methylation values from WGBS were calculated in regions containing multiple
399 CpGs, the 'weighted methylation level'[56] was used. Calculation of mean methylation within
400 PMDs or genomic tiles involved removing all CpGs overlapping with CpG island(-shores) and
401 promoters, as the high CpG densities within these elements yield unbalanced mean methylation
402 values, not representative of PMD methylation. For genome/chromosome-wide visualizations
403 (Fig. 1), 10-kb tiles were used. For visualization, the samples were ordered according
404 hierarchical clustering of the tiled methylation profiles, using 'ward.D' linkage and [1-Pearson
405 correlation] as a distance measure.

406 **Clustering on PMD distribution**

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

410 **Tumor suppressor genes and driver mutations**

411 For overlaps with tumor suppressor genes, Cancer Gene Census
412 (http://cancer.sanger.ac.uk/census, October 2017) genes were used. Overlaps with genes
413 containing breast cancer driver mutations were determined using the list of 93 driver genes as
414 published previously by us[15].

415 **CIMP**

416 To determine the association between B-CIMP (fraction of CGIs that are hypermethylated,
417 >30% methylated) and PMD occurrence we used beta-regression using the 'betareg' package in
418 R[57].

419 **Survival analysis**

420 Survival analysis of patient groups stratified by expression of genes downregulated in PMDs. For
421 each tumor sample of our breast cancer transcriptome cohort (n=266,[17]), the median expression

422 of all PMD-downregulated genes (Suppl. Table 3) was calculated. The obtained distribution of
423 these medians was used to stratify patient groups, using a two-way split over the median of this
424 distribution. Overall survival analysis using these groups was done using the 'survival' package
425 in R, with *chi*-square significance testing.

426

427 **ACKNOWLEDGEMENTS**

428 This work has been funded through the ICGC Breast Cancer Working group by the Breast
429 Cancer Somatic Genetics Study (BASIS), a European research project funded by the European
430 Community's Seventh Framework Programme (FP7/2010-2014) under the grant agreement
431 number 242006; the Triple Negative project funded by the Wellcome Trust (grant reference
432 077012/Z/05/Z). For contributions towards specimens and collections: Tayside Tissue Bank,
433 OSBREAC consortium, Icelandic Centre for Research (RANNIS), Swedish Cancer Society,
434 Swedish Research Council, Fondation Jean Dausset-Centre d'Etudes du polymorphisme humain,
435 Icelandic Cancer Registry, Brisbane Breast Bank, Breast Cancer Tissue and Data Bank and
436 ECMC (King's College London), NIHR Biomedical Research Centre (Guy's and St Thomas's
437 Hospitals), Breakthrough Breast Cancer, Cancer Research UK. We thank E.M. Janssen-Megens,
438 K. Berentsen, H. Kerstens and K.J. Francoijs for technical support. We thank H. Kretzmer and R.
439 Siebert for providing processed data files of the lymphoma dataset. Funding to A.B.B. was
440 through the Dutch Cancer Foundation (KWF) grant KUN 2013-5833. SN-Z is personally funded
441 by a CRUK Advanced Clinician Scientist Award (C60100/A23916). M.S. was supported by the
442 EU-FP7-DDR response project. L.B.A. is supported through a J. Robert Oppenheimer
443 Fellowship at Los Alamos National Laboratory. A.L.R. is partially supported by the Dana-
444 Farber/Harvard Cancer Center SPORE in Breast Cancer (NIH/NCI 5 P50 CA168504-02). J.A.F.
445 was funded through an ERC Advanced Grant (ERC-2012-AdG-322737) and ERC Proof-of-
446 Concept Grant (ERC-2017-PoC-767854). A.S. was supported by Cancer Genomics Netherlands
447 (CGC.nl) through a grant from the Netherlands Organisation of Scientific research (NWO). We
448 received additional support from the Dutch national e-infrastructure (SURF Foundation). Finally,
449 we would like to acknowledge all members of the ICGC Breast Cancer Working Group.

450 **FIGURE LEGENDS**

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

452 **(A)**, Genome-wide and **(B)**, chromosome-wide maps of WGBS DNA methylation profiles from
453 30 breast tumor samples. Mean methylation is displayed in consecutive tiles of 10 kb (see
454 Methods). Ordering of tumor samples is according clustering of the tiled profiles. **(C)**, WGBS
455 DNA methylation visualization at megabase-scale. Pink coloring indicates common methylation
456 loss (PMDs), although tumor-specific PMD borders vary. A scale bar (100 kb) is shown at the
457 top of each panel. CpG islands are indicated in green. **(D)**, Principal component analysis of
458 WGBS DNA methylation profiles (see Methods). Each tumor sample is represented with its
459 estrogen-receptor (ER) status (point shape) and mean PMD methylation (point color).

460 **Figure 2 | Characterization of breast cancer PMDs.**

461 **(A)**, Fraction of the genome covered by PMDs. Each dot represents one tumor sample, the
462 boxplot summarizes this distribution. **(B)**, Fraction of the genome covered by PMDs that are
463 common between breast tumors. PMD frequency: the number of tumors in which a genomic
464 region or gene is a PMD. **(C)**, Breast cancer PMDs are not distributed randomly over the genome.
465 The genome was dissected into 30-kb tiles, PMD frequency (number of boundaries) was
466 calculated for each tile. The same analysis was done after shuffling the PMDs of each tumor
467 sample. **(D)**, Average profiles of LaminB[23], repliSeq (DNA replication timing, ENCODE), 3D
468 chromatin interaction loops (HiC[27], and CTCF (ENCODE) over PMD borders. If available, data
469 from the breast cancer cell line (MCF7) and mammary epithelial cells (HMEC) was used,
470 otherwise data from fibroblasts (IMR90, Tig3) was used. **(E)**, Gene distribution inside PMDs
471 (top, as a fraction of all annotated genes; bottom, as gene coding density). **(F)**, Gene expression
472 inside PMDs. Gene expression (top) and standard deviation (bottom) for the 25 overlapping
473 cases of our WGBS and the transcriptome cohorts[17] was plotted as a function of PMD frequency.
474 **(G)**, Somatic mutations inside PMDs. Substitutions, insertions, deletions, and rearrangements
475 were calculated for the 25 overlapping cases of our WGBS and the breast tumor full genomes
476 cohorts[15], and plotted as a function of PMD frequency. **(H)**, Distribution of DNA methylation
477 over functional genomic elements, inside and outside PMDs. CpGs were classified according

478 PMD status and genomic elements, and the distribution of DNA methylation within each element
479 was plotted.

480 **Figure 3 | CpG island hypermethylation inside PMDs.**

481 **(A)**, Example of a genomic region with CGI hypermethylation inside PMDs. Red bars, PMDs for
482 each tumor sample; below, CGI methylation for each tumor sample (same ordering). Green bars,
483 CGIs. **(B)**, Distribution of CGI methylation, represented as the fraction of all CGIs (x-axis). Each
484 horizontal bar represents one tumor sample. **(C)**, Average profile of methylation over all CGIs
485 inside (red) or outside (black) PMDs, over all 30 tumor samples. Black/red lines, median;
486 grey/pink area, 1st and 3rd quartiles. **(D)**, Number of CGIs inside and outside of breast cancer
487 PMDs. CGIs are classified as 'in' when inside a PMD in at least one tumor sample. **(E)**,
488 Variation of CGI methylation (standard deviation) as a function of PMD frequency. **(F)**,
489 Regression analysis of B-CIMP (y-axis) as a function of the fraction CGIs inside PMDs (x-axis).
490 B-CIMP is defined as the genome-wide fraction of hypermethylated CGIs (>30% methylation).
491 **(G)**, Summary of regression analysis as in (F), including additional cancer types. n, the number
492 of samples for each type. For abbreviations of cancer type names, see Fig. 4B. **(H)**, Expression
493 change of CGI-promoter genes inside vs. outside of PMDs, as a function of PMD frequency. **(I)**,
494 Gene expression levels as a function of PMD frequency in an independent breast cancer dataset
495 (TCGA). PMD frequency for each gene was taken from our own dataset. **(J)**, Expression change
496 of CGI-promoter genes of tumor vs. normal, as a function of PMD frequency. From the TCGA
497 breast cancer dataset, matched tumor/normal pairs were selected. PMD frequency for each gene
498 was taken from our own dataset. **(K)**, Tumor-suppressor genes (TSGs) are excluded from PMDs.
499 For each TSG its PMD frequency was determined and the resulting distribution was plotted.
500 Main plot, relative distribution; inset, absolute number of genes. 'Non-TSGs', genes not
501 annotated as TSGs; 'TSGs all cancers', genes annotated as TSGs regardless of cancer type;
502 'TSGs breast cancer', genes annotated as TSG in breast cancer; 'Nik-Zainal breast cancer driver
503 mutations', genes with driver mutations in breast cancer[15]. **(L)**, Expression of X-linked genes
504 when inside or outside PMDs. Genes were grouped according their consensus X-inactivation
505 status (E, escape; S, subject to XCI; VE, variably escaping; PAR, pseudoautosomal region)[41].

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

507 **(A)**, Mean PMD methylation of normal tissues and tumors of various tissue types. Each dot
508 represents one sample. **(B)**, Hierarchical clustering of tumor samples based on genomic
509 distribution of their PMDs. For breast tumors (this study) the ER status is indicated.

510 **Supplemental Figure 1**

511 **(A)**, CpG coverage in WGBS DNA methylation profiles of 30 breast tumor samples used in this
512 study (see also Suppl. Table 1). **(B)**, Clinicopathological features of the 30 tumor samples. **(C)**,
513 Mean copy-number profiles of 25/30 tumor samples used in this study. Copy-number data was
514 taken from our previous work[15]. **(D)**, Association between mean PMD methylation and
515 expression of genes involved in writing, erasing, or reading the 5-methylcytosine modification.
516 Each dot represents one tumor sample. Linear regression was used to determine the variation
517 explained (R^2) and the p-value of the association. Expression data was taken from our previous
518 work[17].**(E)**, Mean PMD methylation (y-axis) is not associated with the fraction of aberrant cells
519 (ASCAT[19], x-axis).

520 **Supplemental Figure 2**

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

526 **Supplemental Figure 3**

527 **(A)**, Genome-wide and **(B)**, chromosome-wide maps of WGBS DNA methylation profiles from
528 30 breast tumor samples. Exactly as in Fig. 1AB, but using only solo-WCGW CpGs[12]. Mean
529 methylation is displayed in consecutive tiles of 10 kb (see Methods). Ordering of tumor samples
530 is the same as in Fig. 1. **(C)**, WGBS DNA methylation visualization at megabase-scale, exactly
531 as in Fig. 1C, but using only solo-WCGW CpGs. Pink coloring indicates common methylation
532 loss (PMDs) as in Fig. 1, although tumor-specific PMD borders vary. A scale bar (100 kb) is
533 shown at the top of each panel. CpG islands are indicated in green.

534 **Supplemental Figure 4**

535 **(A)**, Association between principal component 1 and 2 scores (PC1, PC2, see Fig. 1D) and major
536 pathological subtypes. Significance testing was done as described in ref.[20]. Left panels, ER-status;
537 right panels, AIMS 'intrinsic' subtypes[21]. **(B,C)**, The same analysis as in Fig. 2AB, but using
538 PMDs detected with only solo-WCGW CpGs[12]. **(B)**, Fraction of the genome covered by PMDs.
539 Each dot represents one tumor sample, the boxplot summarizes this distribution. **(C)**, Fraction of
540 the genome covered by PMDs that are common between breast tumors. PMD frequency: the
541 number of tumors in which a genomic region is a PMD. **(D)**, Venn-diagram showing the overlap
542 between the union of all breast cancer PMDs ('all-CpGs') and the union of all breast cancer solo-
543 WCGW PMDs. **(E)**, Bimodal distribution of cross-sample standard deviation of mean
544 methylation in 100 kb genomic windows. Only solo-WCGW CpGs were used to calculate
545 window means. As described in[12], a mixed gaussian was fitted to determine a cutoff for genome
546 segmentation. **(F)**, Overlaps between the cross-sample s.d. based PMDs (E) and PMDs called on
547 individual samples in this study, using all CpGs.

548 **Supplemental Figure 5**

549 **(A)**, Gene expression as a function of PMD frequency, as in main Figure 2F, but here extended
550 to all 266 cases of the breast tumor (RNA-seq) transcriptomes cohort[17]. Top, gene expression;
551 bottom, standard deviation. **(B)**, Somatic mutations plotted as a function of PMD frequency, as in
552 main Figure 2G, but here extended to all 560 cases of the breast tumor full genomes cohort[15].

553 **Supplemental Figure 6**

554 **(A)**, Expression change of non-CGI-promoter genes inside vs. outside of PMDs, as a function of
555 PMD frequency. **(B)**, Expression change of non-CGI-promoter genes of tumor vs. normal, as a
556 function of PMD frequency. From the TCGA breast cancer dataset, matched tumor/normal pairs
557 were selected. PMD frequency for each gene was taken from our own dataset. **(C,D)**,
558 Multivariate linear regression was performed with expression levels of genes involved in XCI as
559 explanatory variables and PMD abundance on chrX as response variable. The variable
560 importance of each XCI gene is plotted in (C), and their expression levels in two PMD
561 abundance bins is plotted in (D). **(E)**, Expression of X-linked genes when inside or outside
562 PMDs. Genes were grouped according their consensus X-inactivation status (E, escape; S,

563 subject to XCI; VE, variably escaping; PAR, pseudoautosomal region)[41] and further stratified
564 over their copy-number status (gain, loss, unchanged) as determined previously[15].

565 **Supplemental Figure 7**

566 **(A)**, Expression change of TSGs/breast cancer driver mutated genes when inside PMDs. 31 of
567 such genes are located inside PMDs in a subset of tumor samples. 'TSGs all cancers', genes
568 annotated as TSGs regardless of cancer type; 'TSGs breast cancer', genes annotated as TSG in
569 breast cancer; 'Nik-Zainal breast cancer driver mutations', genes with driver mutations in breast
570 cancer[15]. **(B)**, Examples of genes from panel (A) being repressed when inside PMDs. Blue line,
571 DNA methylation (WGBS); green bars, CGIs; red bars, PMDs. Gene expression (RNA-seq) of
572 the corresponding gene is represented at the right of each panel. **(C)**, Pearson correlation between
573 CGI-promoter methylation and expression. Gene classes are indicated as in panel (A). **(D)**,
574 Expression changes (RNA-seq) of genes in panel (B), breast tumor vs. normal. Data is from an
575 independent cohort (TCGA). Left panels, non-matched normal (n=88) and tumor samples
576 (n=769); right panels, matched normal/tumor samples (n=86). p-values were calculated using a *t*-
577 test.

578 **Supplemental Figure 8**

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

587 **Supplemental Figure 9**

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

591 samples other than breast cancer (TCGA, BLUEPRINT, the Roadmap Epigenomics Project,
592 refs.[10,37,38], abbreviations are given on the right); bottom panel, repeated from main Figure 3B for
593 comparison.

594 **Supplemental Table 1**

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

597 **Supplemental Table 2**

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

600 **Supplemental Table 3**

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

603

604 **REFERENCES**

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

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

609 3. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation
610 in colorectal cancer coincide with nuclear lamina-associated domains. *Nature genetics* **44,**
611 40–6 (2012).

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

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

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

618 7. Schroeder, D. I., Lott, P., Korf, I. & LaSalle, J. M. Large-scale methylation domains mark a
619 functional subset of neuronally expressed genes. *Genome research* **21,** 1583–91 (2011).

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

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

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

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

628 12. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell
629 division. *Nature Genetics* **50,** 591–602 (2018).

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

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

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

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

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

640 18. Chaligné, R. *et al.* The inactive X chromosome is epigenetically unstable and
641 transcriptionally labile in breast cancer. *Genome Research* **25,** 488–503 (2015).

642 19. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the*
643 *National Academy of Sciences* **107,** 16910–16915 (2010).

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

646 21. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene
647 expression data sets. *Proceedings of the National Academy of Sciences of the United States*
648 *of America* **100,** 8418–8423 (2003).

649 22. Paquet, E. R. & Hallett, M. T. Absolute assignment of breast cancer intrinsic molecular
650 subtype. *Journal of the National Cancer Institute* **107,** 1–9 (2015).

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

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

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

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

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

662 28. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene Density, Transcription, and Insulators
663 Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*
664 **48,** 471–484 (2012).

665 29. Roukos, V. & Misteli, T. The biogenesis of chromosome translocations. *Nature cell biology*
666 **16,** 293–300 (2014).

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

669 31. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Medical Sciences* **96,**
670 8681–8686 (1999).

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

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

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

678 35. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics*
679 **45,** 1113–1120 (2013).

680 36. Stunnenberg, H. G. *et al.* The International Human Epigenome Consortium: A Blueprint for
681 Scientific Collaboration and Discovery. *Cell* **167,** 1145–1149 (2016).

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

685 38. Li, X. Whole genome analysis of the methylome and hydroxymethylome in normal and
686 malignant lung and liver. 1730–1741 (2016). doi:10.1101/gr.211854.116.

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

689 40. Smid, M. *et al.* Subtypes of breast cancer show preferential site of relapse. *Cancer Research*
690 **68,** 3108–3114 (2008).

691 41. Balaton, B. P., Cotton, A. M. & Brown, C. J. Derivation of consensus inactivation status for
692 X-linked genes from genome-wide studies. *Biology of Sex Differences* **6,** 35 (2015).

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

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

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

699 45. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature*
700 **490,** 61–70 (2012).

701 46. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature*
702 **511,** 543–550 (2014).

703 47. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163,** 1011–
704 1025 (2015).

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

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

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

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

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

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

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

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

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

724 57. Cribari-Neto, F. & Zeileis, A. Journal of Statistical Software. *Journal of Statistical Software*
725 **34,** 1–24 (2010).

Figure 1

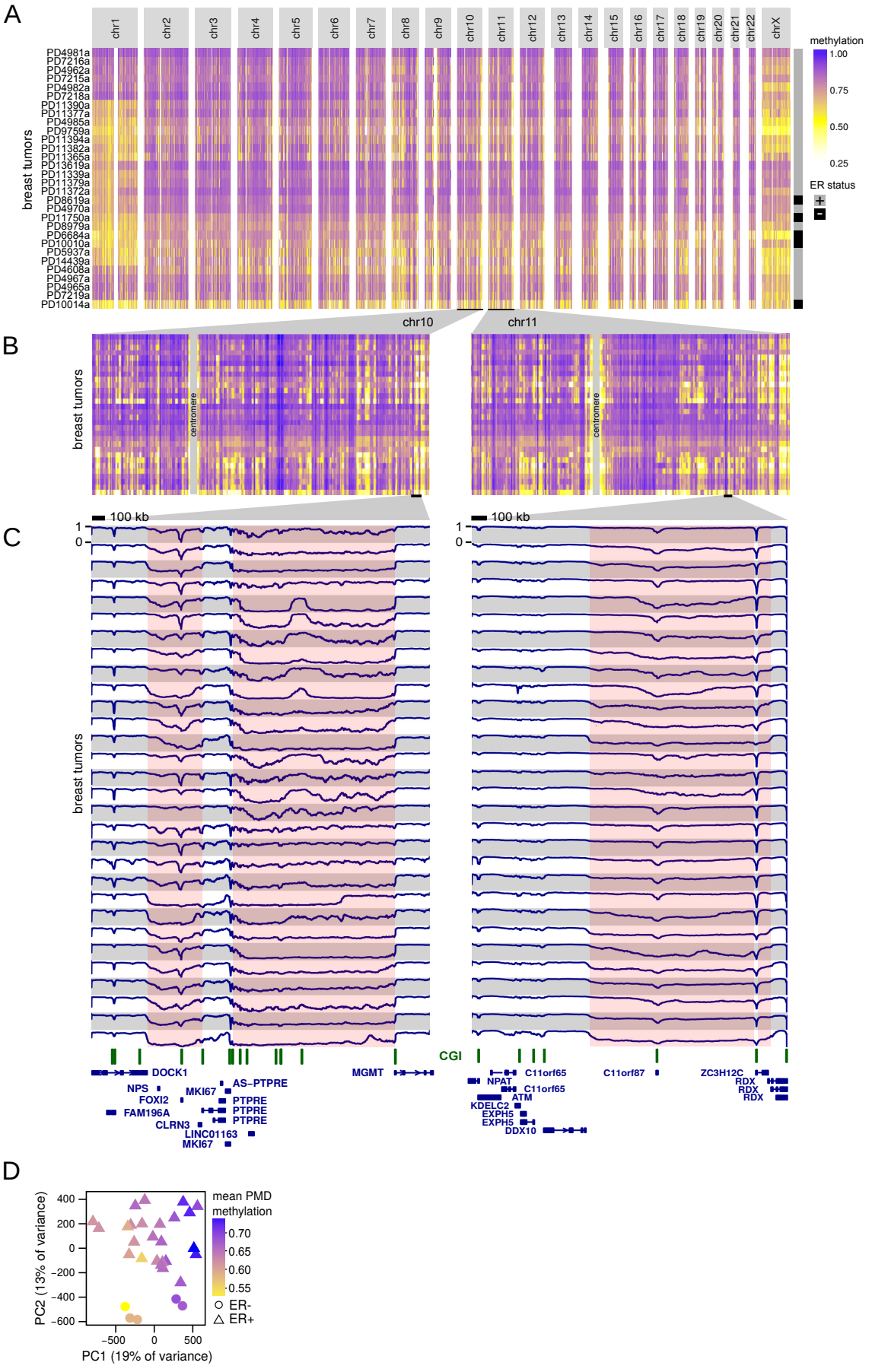

A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
breast tumors
methylation
1.00
0.75
0.50
0.25
ER status
+
-
B
chr10
chr11
breast tumors
centromere
C
100 kb
breast tumors
CGI
DOCK1
NPS
FOXI2
FAM196A
CLRN3
MKI67
LINC01163
AS-PTPRE
PTPRE
MGMT
NPAT
C11orf65
ATM
KDELC2
EXPH5
DDX10
C11orf87
ZC3H12C
RDX
D
PC2 (13% of variance)
PC1 (19% of variance)
mean PMD methylation
0.70
0.65
0.60
0.55
ER-
ER+

Figure 2

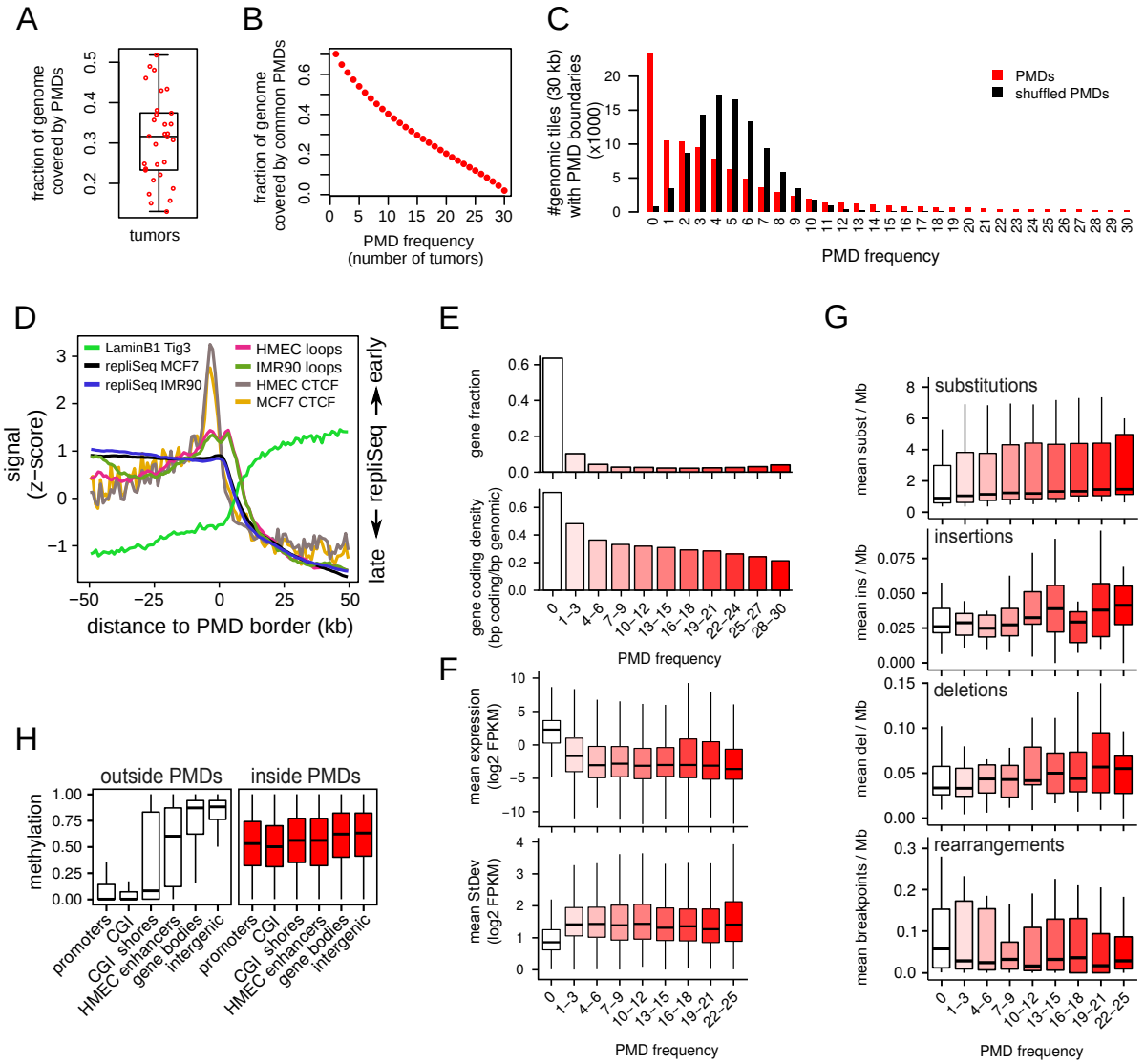

A
fraction of genome covered by PMDs
tumors
B
fraction of genome covered by common PMDs
PMD frequency (number of tumors)
C
#genomic tiles (30 kb) with PMD boundaries (x1000)
PMDs
shuffled PMDs
PMD frequency
D
LaminB1 Tig3
repliSeq MCF7
repliSeq IMR90
HMEC loops
IMR90 loops
HMEC CTCF
MCF7 CTCF
signal (z−score)
distance to PMD border (kb)
late ← repliSeq → early
E
gene fraction
gene coding density (bp coding/bp genomic)
PMD frequency
F
mean expression (log2 FPKM)
mean StDev (log2 FPKM)
PMD frequency
G
substitutions
mean subst / Mb
insertions
mean ins / Mb
deletions
mean del / Mb
rearrangements
mean breakpoints / Mb
PMD frequency
H
outside PMDs
inside PMDs
methylation
promoters
CGI
CGI shores
HMEC enhancers
gene bodies
intergenic

Figure 3

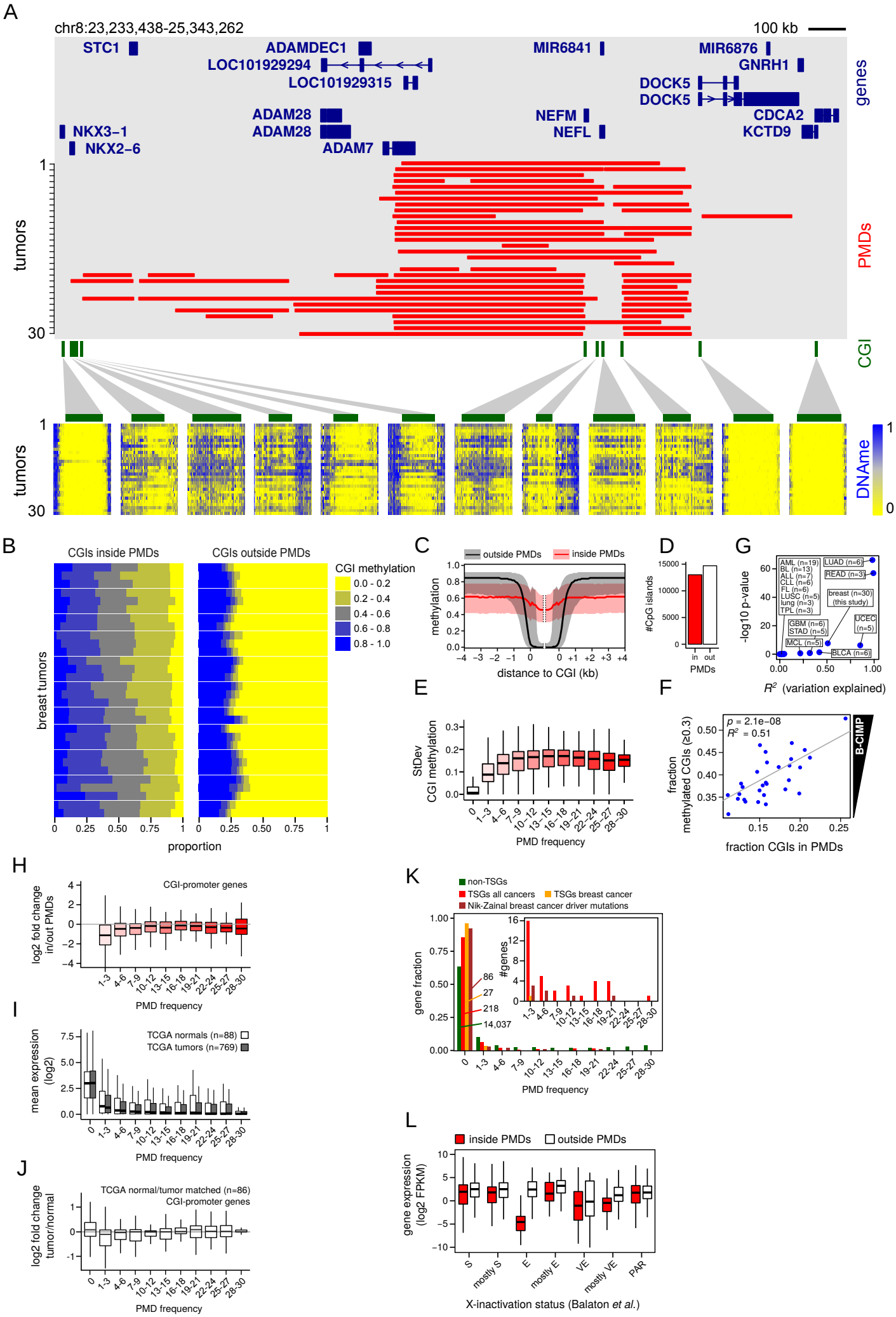
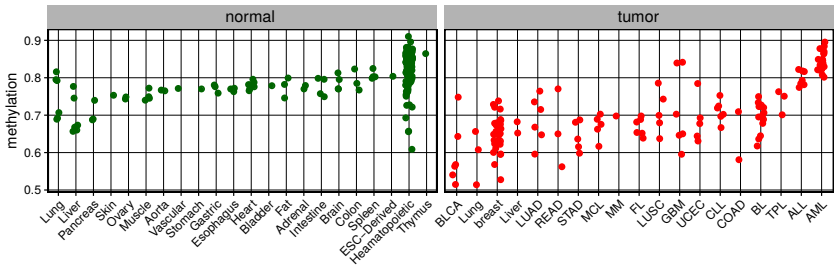
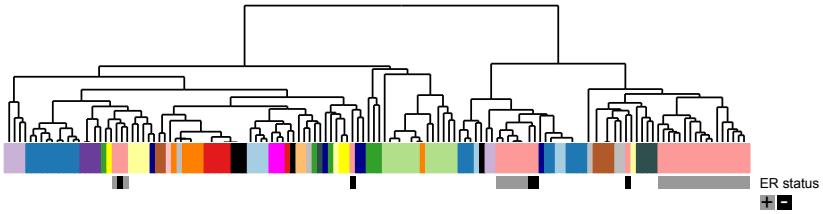

A
chr8:23,233,438-25,343,262
100 kb
STC1
ADAMDEC1
LOC101929294
LOC101929315
MIR6841
MIR6876
GNRH1
DOCK5
DOCK5
CDCA2
KCTD9
NEFM
NEFL
ADAM28
ADAM28
ADAM7
NKX3-1
NKX2-6
genes
tumors
PMDs
CGI
DNAme
B
CGIs inside PMDs
CGIs outside PMDs
CGI methylation
0.0 - 0.2
0.2 - 0.4
0.4 - 0.6
0.6 - 0.8
0.8 - 1.0
breast tumors
proportion
C
outside PMDs
inside PMDs
methylation
distance to CGI (kb)
D
#CpG islands
in
out
PMDs
G
-log10 p-value
AML (n=19)
BL (n=13)
ALL (n=7)
CLL (n=6)
FL (n=6)
LUSC (n=5)
lung (n=3)
TPL (n=3)
LUAD (n=6)
READ (n=3)
breast (n=30) (this study)
UCEC (n=5)
GBM (n=6)
STAD (n=5)
MCL (n=5)
BLCA (n=6)
R² (variation explained)
E
StDev CGI methylation
PMD frequency
F
p = 2.1e−08
R² = 0.51
B-CIMP
fraction methylated CGIs (≥0.3)
fraction CGIs in PMDs
H
CGI-promoter genes
log2 fold change in/out PMDs
PMD frequency
I
TCGA normals (n=88)
TCGA tumors (n=769)
mean expression (log2)
PMD frequency
J
TCGA normal/tumor matched (n=86)
CGI-promoter genes
log2 fold change tumor/normal
PMD frequency
K
non-TSGs
TSGs all cancers
TSGs breast cancer
Nik-Zainal breast cancer driver mutations
gene fraction
#genes
86
27
218
14,037
PMD frequency
L
inside PMDs
outside PMDs
gene expression (log2 FPKM)
S
mostly S
E
mostly E
VE
mostly VE
PAR
X-inactivation status (Balaton et al.)

Figure 4

A

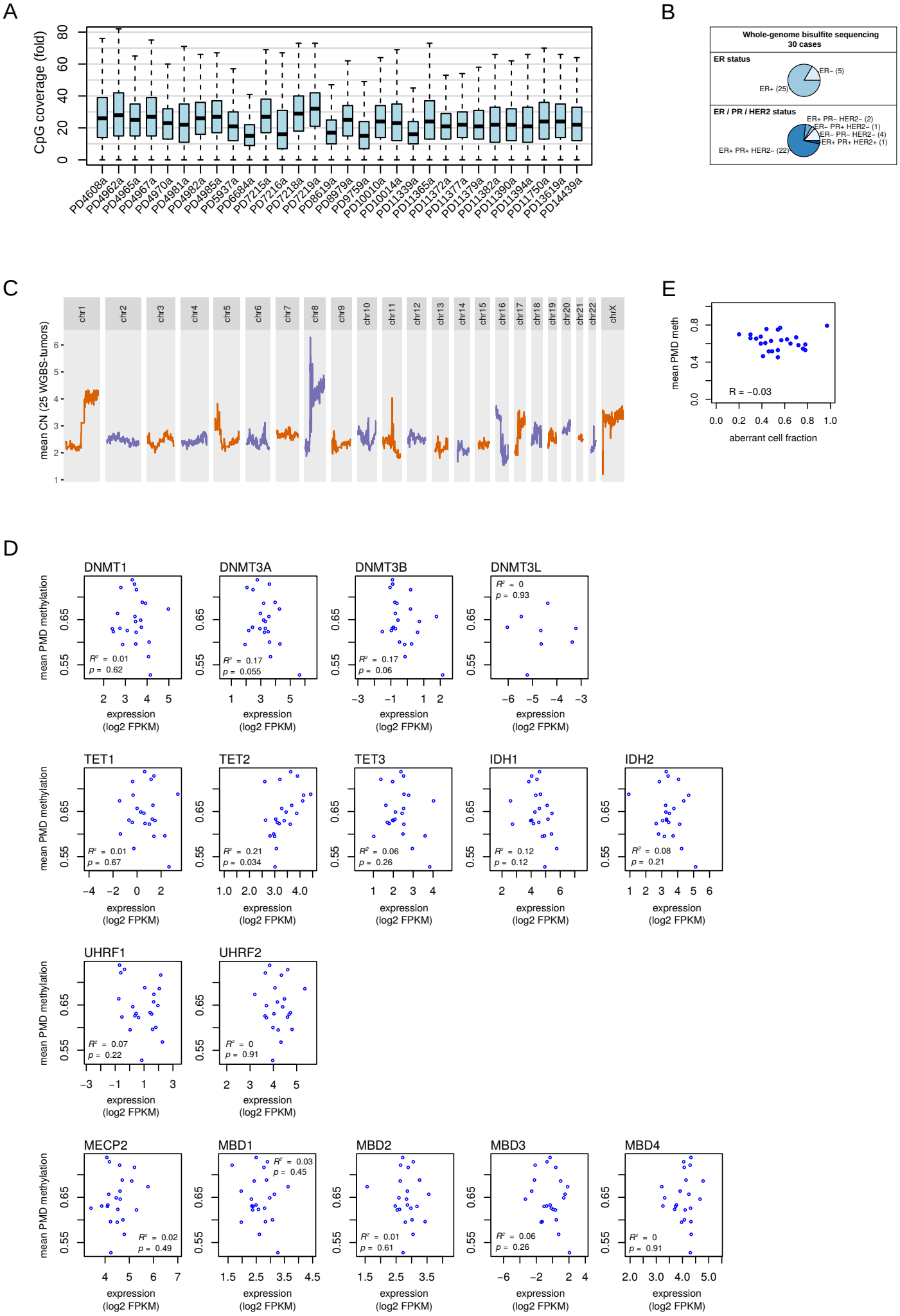


B

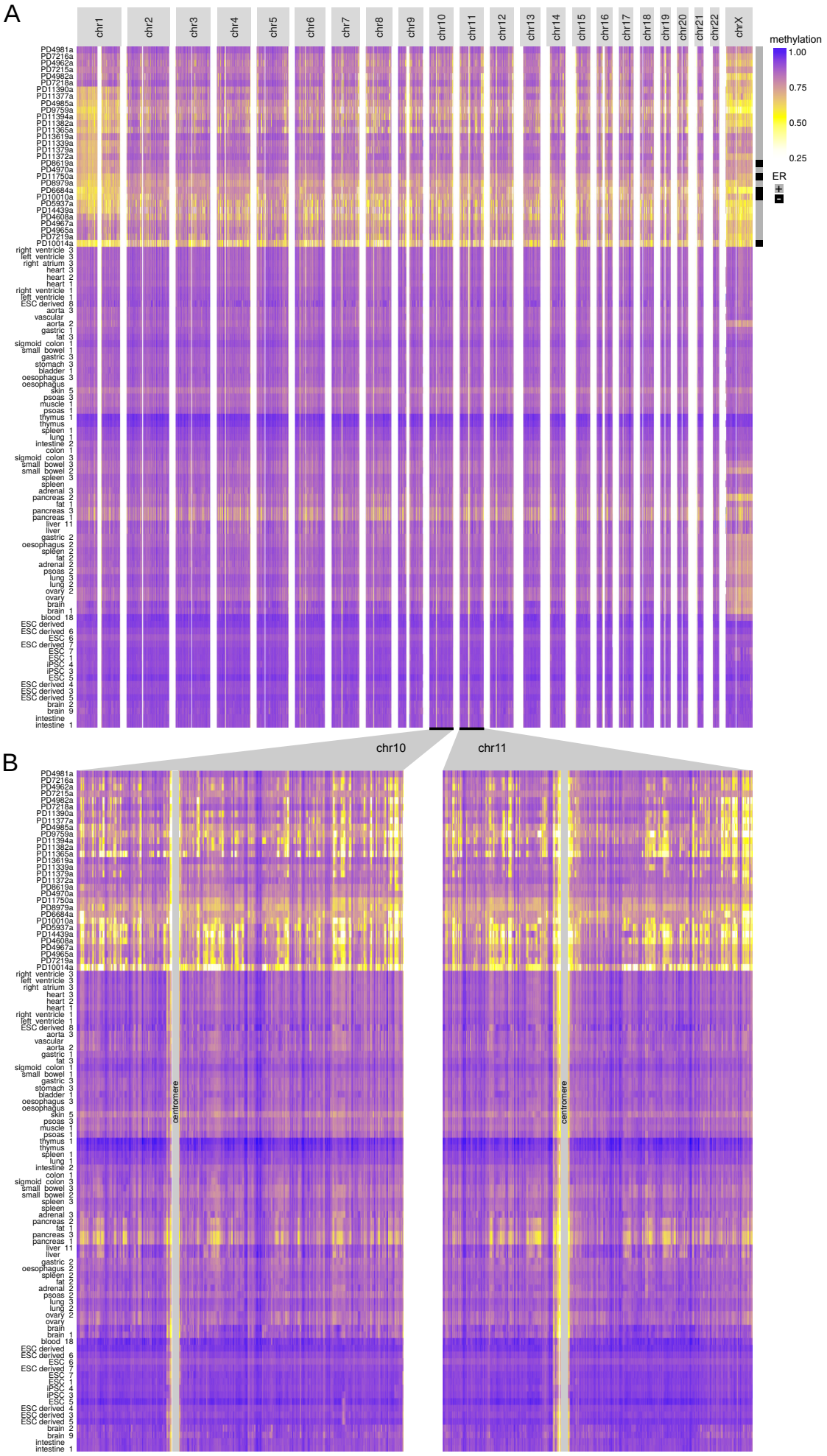


ALL	acute lymphoblastic leukemia (BLUEPRINT)
AML	acute myeloid leukemia (BLUEPRINT)
BL	Burkitt's lymphoma (Kretzmer *et al.*)
BLCA	bladder urothelial carcinoma (TCGA)
breast	**breast cancer (this study)**
CLL	chronic lymphocytic (BLUEPRINT)
COAD	colon adenocarcinoma (TCGA)
FL	follicular lymphoma (Kretzmer *et al.*)
GBM	glioblastoma multiforme (TCGA)
liver	liver tumor (Li *et al.*)
LUAD	lung adenocarcinoma (TCGA)
lung	lung tumor (Li *et al.*)
LUSC	lung squamous cell carcinoma (TCGA)
MCL	mantle cell lympoma (BLUEPRINT)
MM	multiple myeloma (BLUEPRINT)
READ	rectum adenocarcinoma (TCGA)
STAD	stomach adenocarcinoma (TCGA)
TPL	T-cell prolymphocytic leukemia (BLUEPRINT)
UCEC	uterine corpus endometrial carcinoma (TCGA)

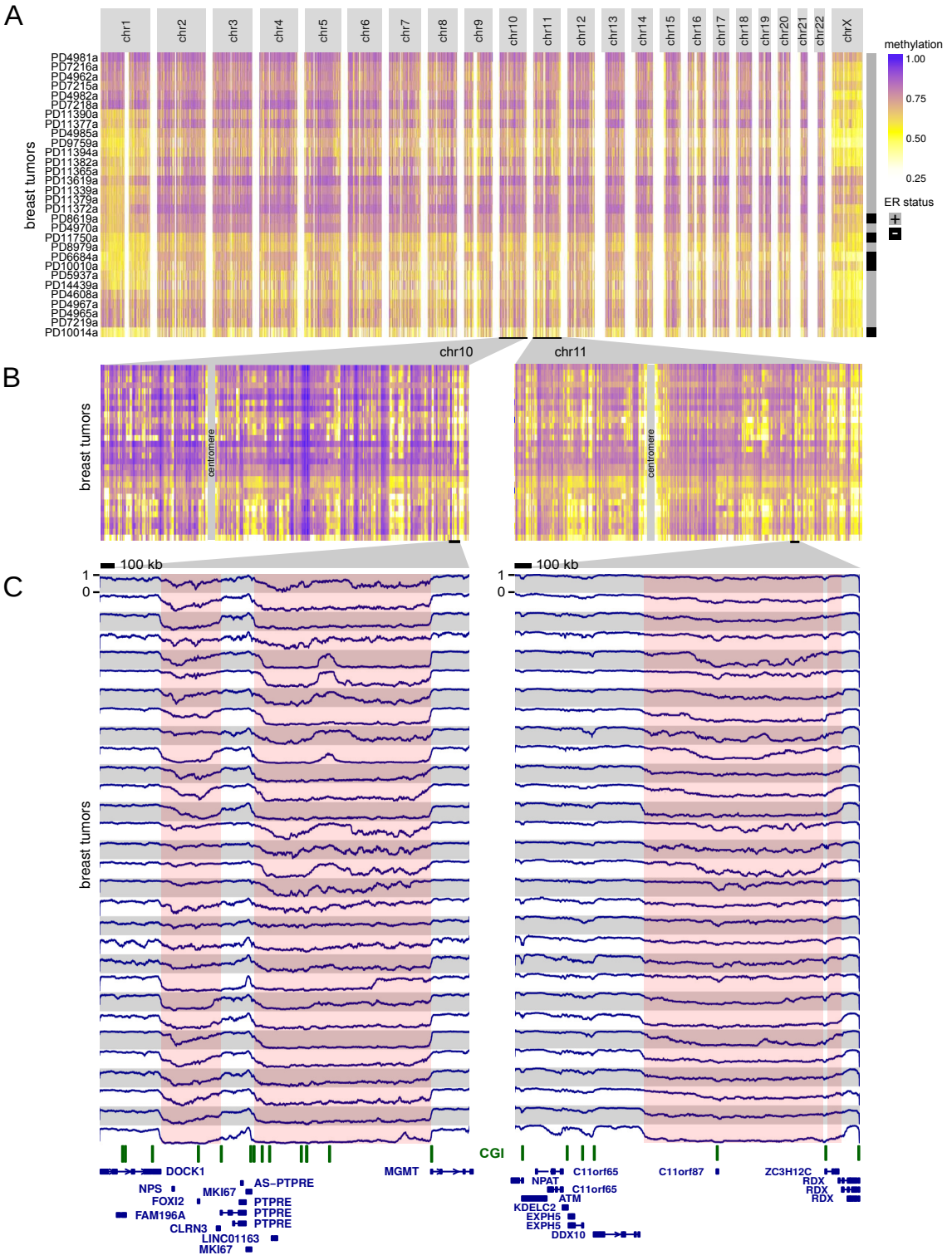
Supplementary Figure 1



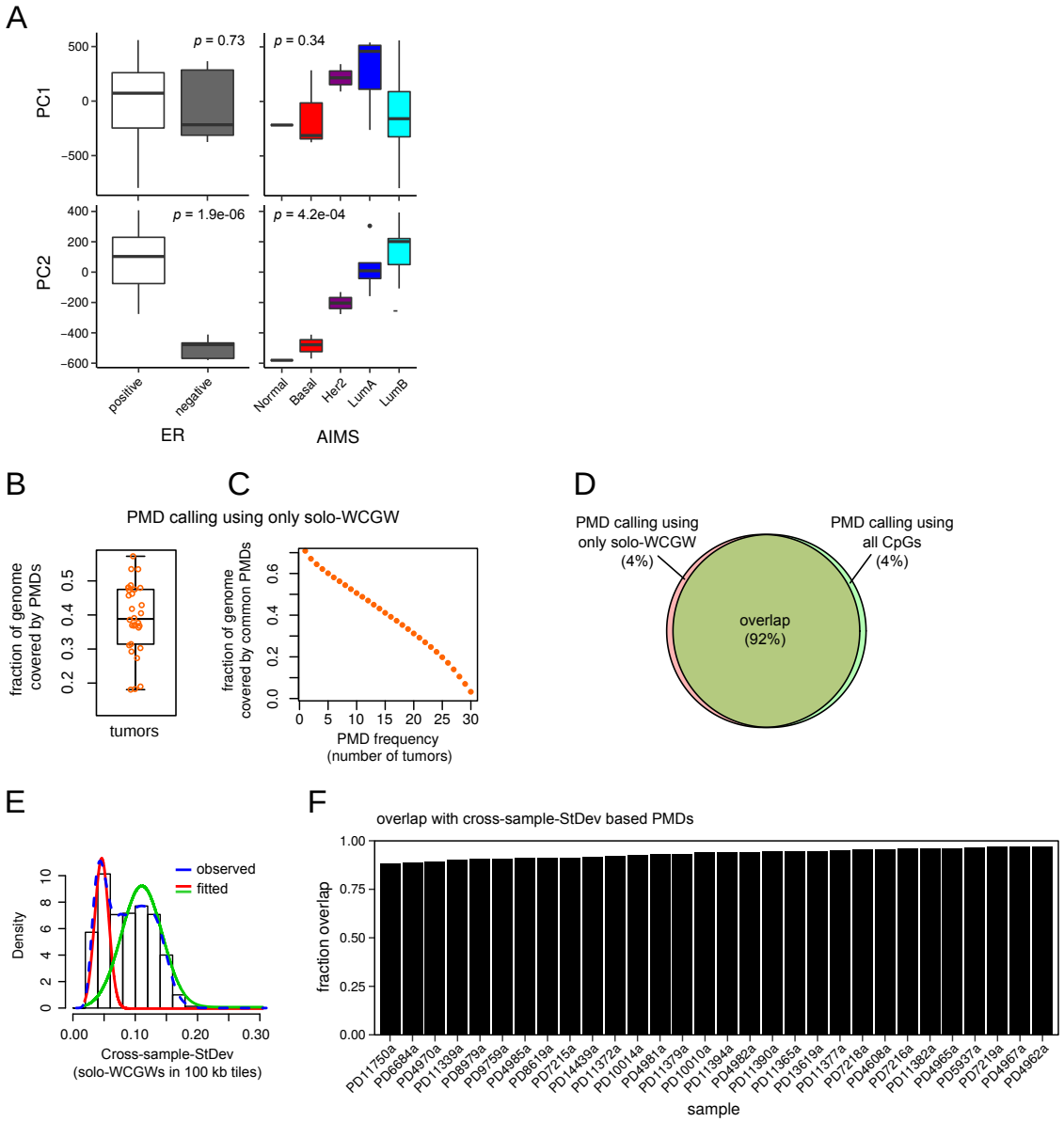
Supplemetary Figure 2



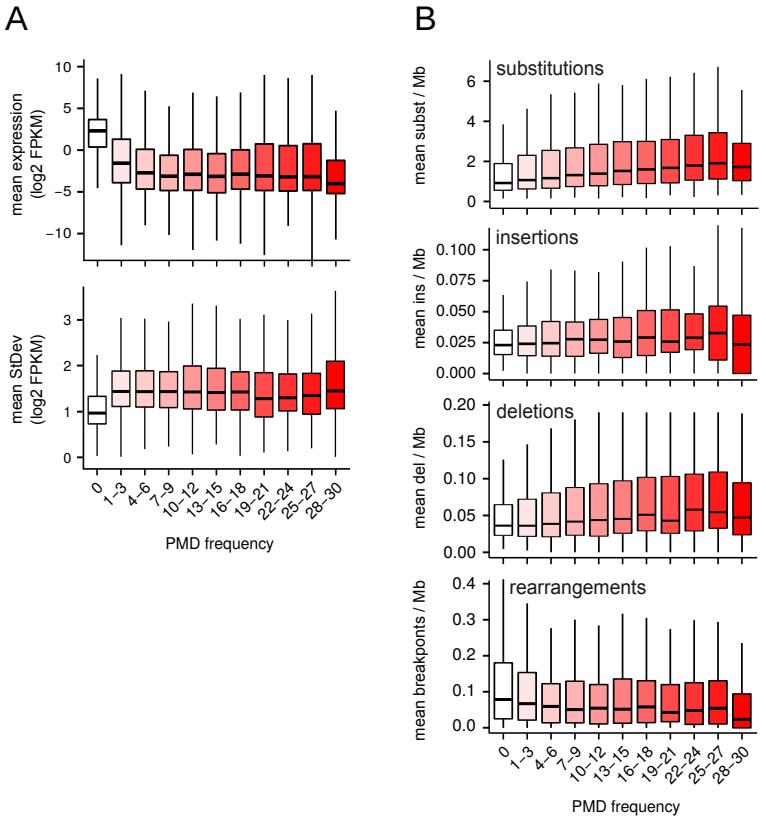
Supplemental Figure 3



Supplementary Figure 4

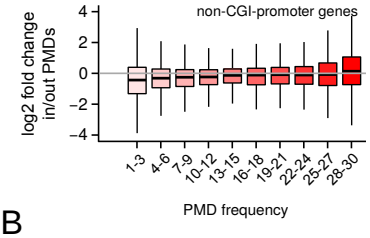


Supplementary Figure 5

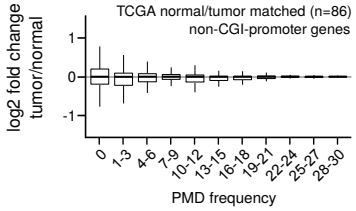


Supplementary Figure 6

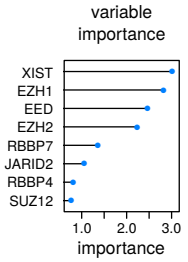
A



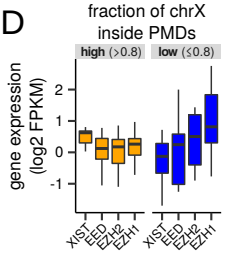
B



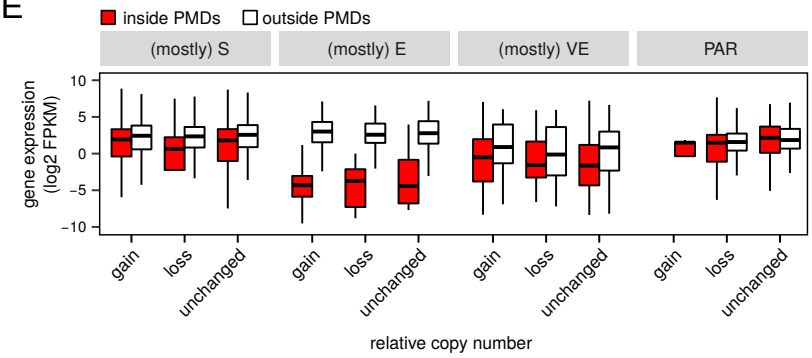
C



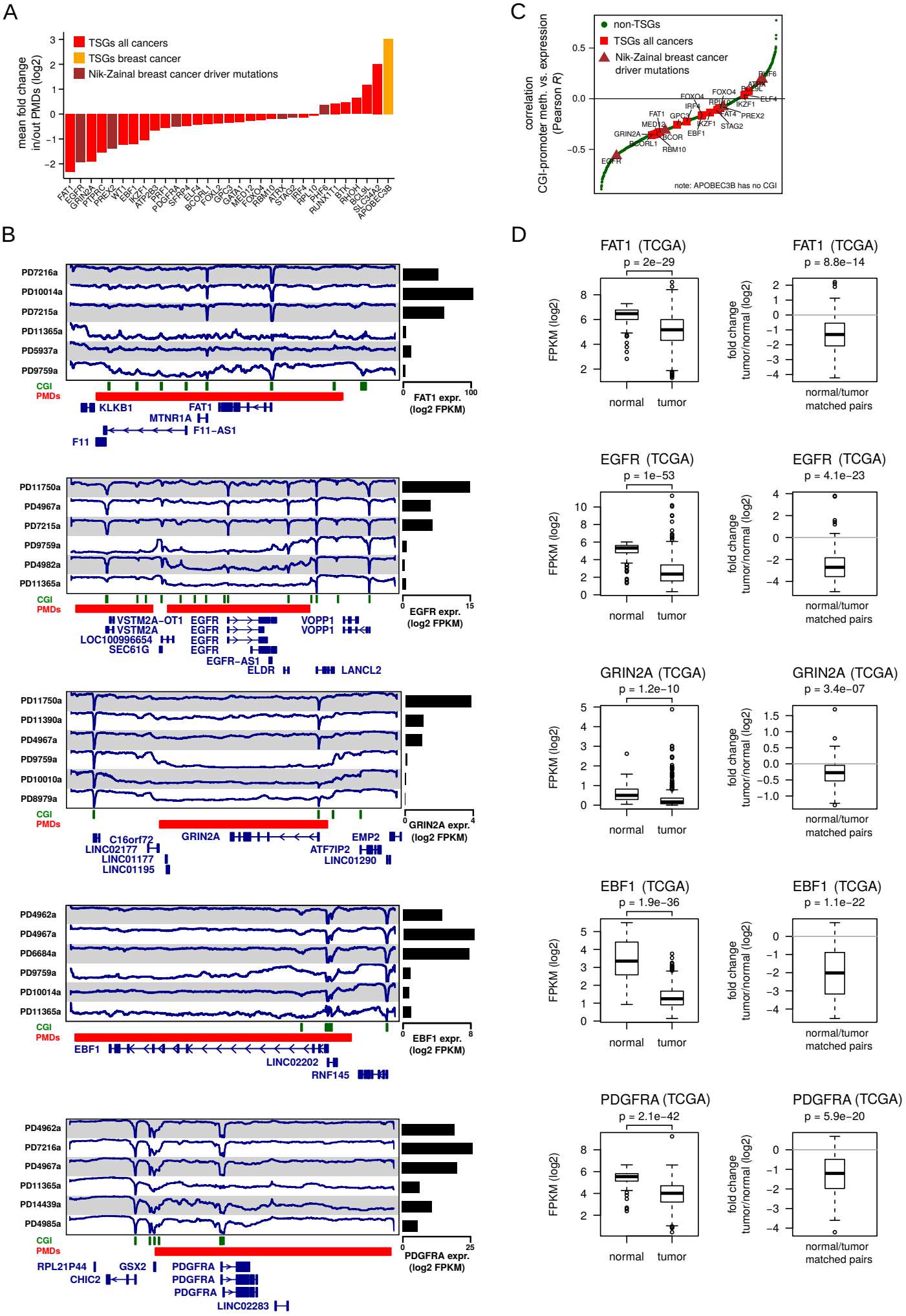
D



E

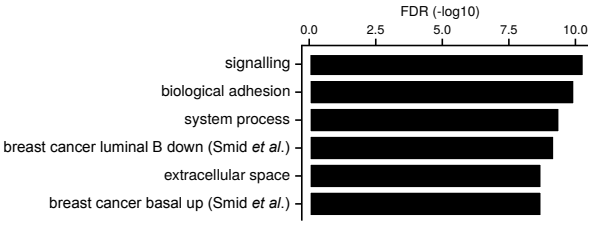


Supplementary Figure 7

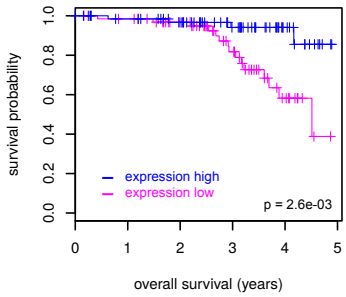


Supplementary Figure 8

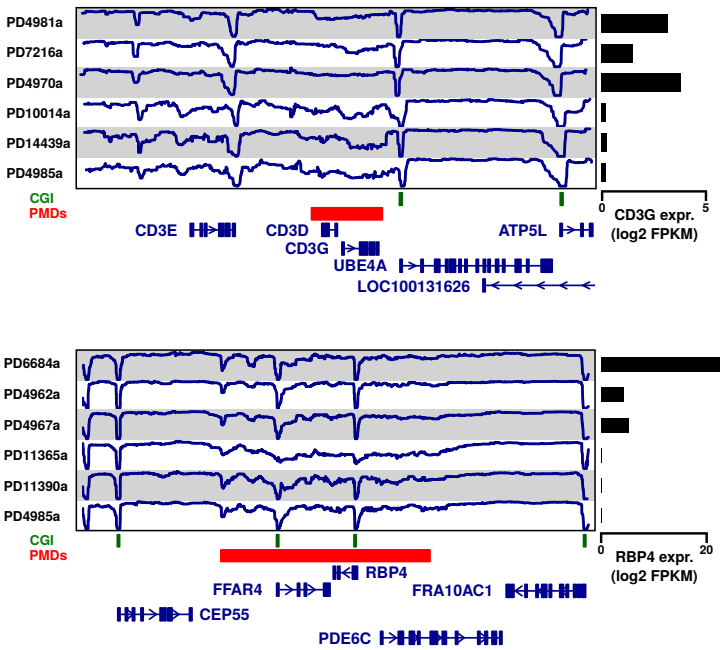
A



C

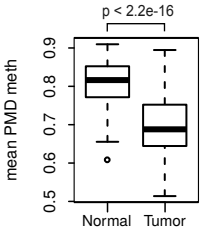


B

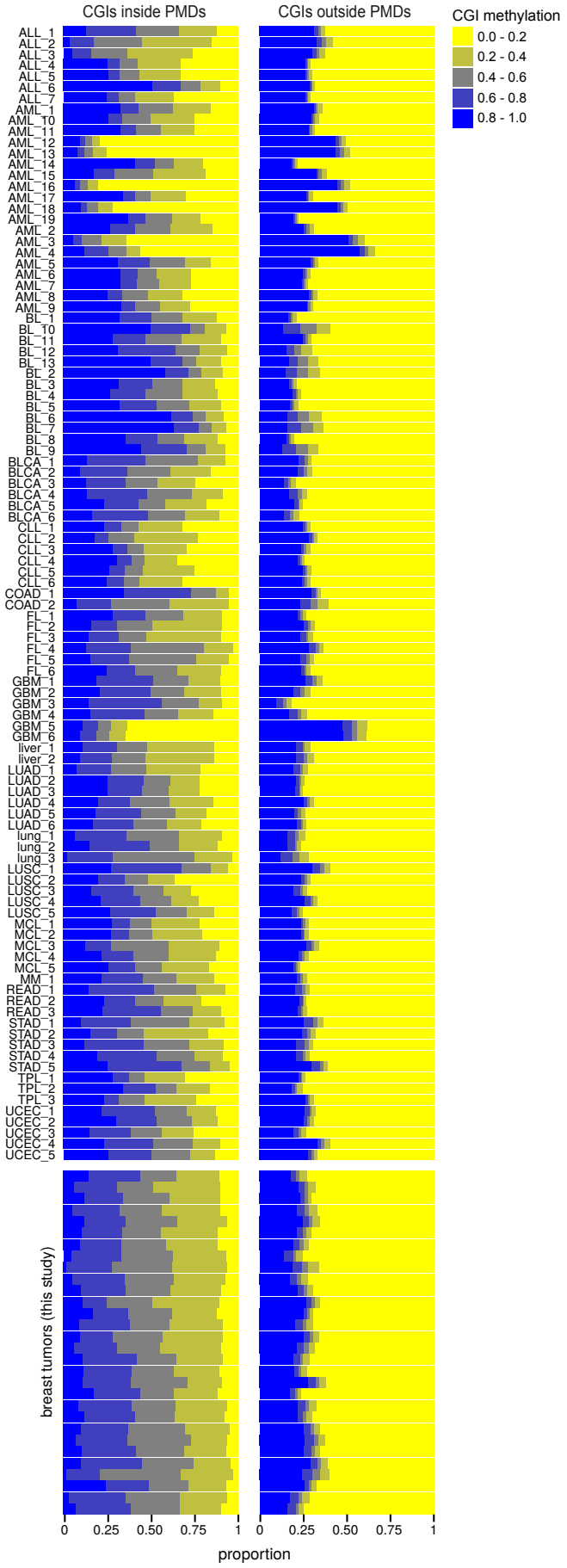


Supplementary Figure 9

A



B



CGI methylation

- 0.0 - 0.2
- 0.2 - 0.4
- 0.4 - 0.6
- 0.6 - 0.8
- 0.8 - 1.0

ALL	acute lymphoblastic leukemia (BLUEPRINT)
AML	acute myeloid leukemia (BLUEPRINT)
BL	Burkitt's lymphoma (Kretzmer *et al.*)
BLCA	bladder urothelial carcinoma (TCGA)
CLL	chronic lymphocytic (BLUEPRINT)
COAD	colon adenocarcinoma (TCGA)
FL	follicular lymphoma (Kretzmer *et al.*)
GBM	glioblastoma multiforme (TCGA)
liver	liver tumor (Li *et al.*)
LUAD	lung adenocarcinoma (TCGA)
lung	lung tumor (Li *et al.*)
LUSC	lung squamous cell carcinoma (TCGA)
MCL	mantle cell lympoma (BLUEPRINT)
MM	multiple myeloma (BLUEPRINT)
READ	rectum adenocarcinoma (TCGA)
STAD	stomach adenocarcinoma (TCGA)
TPL	T-cell prolymphocytic leukemia (BLUEPRINT)
UCEC	uterine corpus endometrial carcinoma (TCGA)