Loss of 5hmC identifies a new type of aberrant DNA hypermethylation in glioma

Agustin F. Fernandez[1,†,*], Gustavo F. Bayón[1,†], Marta I. Sierra[1], Rocio G. Urdinguio[2], Estela G. Toraño[1,2], Maria García[1,2], Antonella Carella[1,2], Virginia Lopez[2], Pablo Santamarina[1,2], Raúl F. Pérez[1,2], Thalía Belmonte[1,2], Juan Ramon Tejedor[1], Isabel Cobo[1,3], Pablo Menendez[3,4], Cristina Mangas[1], Cecilia Ferrero[1], Luis Rodrigo[5], Aurora Astudillo[6], Ignacio Ortea[7], Sergio Cueto Díaz[8], Pablo Rodríguez-Gonzalez[9], J. Ignacio García Alonso[9], Manuela Mollejo[10], Bárbara Meléndez[10], Gemma Dominguez[11], Felix Bonilla[11], and Mario F. Fraga[2,*]

[1]Institute of Oncology of Asturias (IUOPA), ISPA-HUCA, Universidad de Oviedo, Oviedo, Spain.

[2]Nanomaterials and Nanotechnology Research Center (CINN-CSIC)-Universidad de Oviedo-Principado de Asturias, Spain.

[3]Josep Carreras Leukemia Research Institute and Department of Biomedicine, School of Medicine, University of Barcelona, Barcelona, Spain

[4]Institució Catalana de Recerca i Estudis Avançats (ICREA) and Centro de Investigacion Biomedica en Red en Cancer CIBER-ONC, ISCIII, Barcelona, Spain.

[5]Department of Gastroenterology, Hospital Universitario Central de Asturias (HUCA), Oviedo, Spain.

[6]Department of Pathology, Hospital Universitario Central de Asturias and Instituto Universitario de Oncología del Principado de Asturias, Oviedo, Spain.

[7]Proteomics Unit, IMIBIC, Maimonides Institute for Biomedical Research, Córdoba, Spain.

[8]Mass Spectrometry Unit, University of Oviedo.

[9]Department of Physical and Analytical Chemistry, University of Oviedo

[10]Department of Pathology, Hospital Virgen de la Salud, Toledo, Spain. Avd. Barber 30, Toledo 45005

[11]Servicio de Oncología Médica, Hospital Universitario Puerta de Hierro. Majadahonda, Facultad de Medicina, Universidad Autónoma de Madrid, Madrid, Spain

*Corresponding Authors:

Mario F. Fraga: mffraga@cinn.es. Avda. de la Vega, 4 – 6. 33940 El Entrego. Asturias. Spain. Tel. +34 985733644

Agustin F Fernandez: affernandez@hca.es Avenida de Roma s/n 33011 Oviedo. Asturias. Spain. Tel. +34 985652411

†Same contribution.

Abstract

Aberrant DNA hypermethylation is a hallmark of cancer although the underlying molecular mechanisms are still poorly understood. To study the possible role of 5-hydroxymethylcytosine (5hmC) in this process we analyzed the global and locus-specific genome-wide levels of 5hmC and 5mC in human primary samples from 12 non-tumoral brains and 53 gliomas. We found that the levels of 5hmC identified in non-tumoral samples were significantly reduced in gliomas. Strikingly, hypo-hydroxymethylation at 4,627 (9.3%) CpG sites was associated with aberrant DNA hypermethylation and was strongly enriched in CpG island (CGI) shores. The DNA regions containing these CpG sites were enriched in H3K4me2 and presented a different genuine chromatin signature to that characteristic of the genes classically aberrantly hypermethylated in cancer. As this 5mC gain is inversely correlated with loss of 5hmC and has not been identified with classical sodium bisulfite-based technologies, we conclude that our data identifies a novel 5hmC-dependent type of aberrant DNA hypermethylation in glioma.

Introduction

DNA methylation at the fifth position of cytosine (5mC) has been one of the most studied epigenetic modifications in mammals to date. 5mC is involved in the regulation of multiple physiological and pathological processes, including cancer, and when located at gene promoters, it is usually linked to transcriptional repression.

As distinctive features of tumorigenesis, local DNA hypermethylation and global hypomethylation have been attributed to changes in 5mC levels (1, 2). However, the discovery a few years ago, of 5-hydroxymethylcytosine (5hmC), a new epigenetic mark resulting from 5mC oxidation, is reshaping our view of the cancer epigenome (3, 4). This 5mC to 5hmC conversion in mammals is mediated by ten-eleven translocation proteins (TET1, TET2, and TET3), a family of α-ketoglutarate (αKG) and Fe(II)-dependent dioxygenases (4, 5). Global levels of 5hmC in the genome fluctuate considerably according to tissue type, and are consistently around 10-fold lower than those of 5mC, though it is interesting that the highest levels of both marks are found in brain (6-11).

Several studies have shown that 5hmC is an intermediate of DNA demethylation (7, 12, 13), and that it is also associated with cancer (14-18). In this same vein, loss of 5hmC has been reported across a wide range of human cancers including melanoma, glioma, breast, colon, gastric, kidney, liver, lung, pancreatic, and prostate cancers (9, 14, 18-22). Although none of these studies actually demonstrate that changes in 5hmC are directly affecting cancer biology independent of its role in DNA demethylation, other studies identifying 5hmC-specific protein binders suggest that this epigenetic mark might have its own biological role (23, 24). The fact that there are now methods available that distinguish 5mC and 5hmC positions at single-base resolution within the genome prompted us to reassess the role of DNA methylation status in tumorigenesis from a 5hmC perspective. The method used here allowed us to describe global and genome-wide locus-specific 5mC and 5hmC patterns in brain samples, to identify a specific chromatin signature associated with changes of these epigenetic marks in glioma and, most importantly, to describe a novel non-canonical type of aberrant DNA hypermethylation.

Results

Global changes of 5mC and 5hmC in cancer

To evaluate the role of 5hmC in the changes of DNA methylation observed in glioma, we first analyzed the levels of 5hmC and 5mC at DNA repeats in 11 normal and 50 tumor samples. We used oxidative bisulfite conversion (oxBS) of DNA to discriminate between 5mC and 5hmC (see Methods) and bisulfite pyrosequencing was used to determine the level of both epigenetic modifications in 4 different types of repeated DNA: the retrotransposons LINE-1 and AluYb8, and the pericentromeric tandem repeats Sat-alpha and NBL-2 (25). These 4 DNA regions contain most of the genomic methylation and, consequently, global DNA methylation level is highly dependent on their 5mC content (26). As expected, 5mC levels at repeated DNA in healthy tissue were high but were reduced in tumor samples, a change which was statistically significant in LINE-1 and Sat-alpha (linear model, $p<0.001$) (**Fig. 1a**). In contrast, levels of 5hmC at repeated DNA in healthy tissue were low, while tumoral tissue showed even lower levels of 5hmC in the same DNA regions, although this was only statistically significant in LINE-1 (linear model, $p< 0.05$). (**Fig. 1b**).

5mC and 5hmC profiling in brain tissue

As changes in 5hmC at repeated DNA were not able to explain the global loss of this epigenetic mark previously observed by mass spectrometry (14, 20, 27, 28), we hypothesized that these changes primarily occur at single copy sequences. To investigate this possibility in more detail, we first used 450K Infinium methylation arrays to determine the level and genomic distribution of 5mC and 5hmC in 5 healthy brain tissue samples. A preliminary examination of the data revealed that the beta values of the oxidized samples (true 5mC) were lower than their non-oxidized counterparts (5mC+5hmC) (Wilcoxon rank sum test; $p<0.001$; W=2.34e13) (**Fig. 2a**). Specifically, 111,633 hydroxymethylated CpG sites (5hmC sites) distributed throughout the whole genome were identified, irrespective of the chromosome analyzed (**Fig. 2a, and Supplementary Table 1**).

To further validate the consistency of our results we compared our data with published data for 5hmC in human normal brain (29) obtained by Tet-Assisted Bisulfite Sequencing (TAB-Seq), an alternative technique that directly measures the 5hmC rather than inferring it, and we found that more than 75% of the 5hmC-enriched CpG sites identified in our study overlapped with 5hmC-enriched CpG sites analyzed by the TAB-Seq (**Supplementary Fig. S1)**.

The analysis of the genomic distribution of the 5hmC sites showed that hydroxymethylation is enriched at the low CpG-density regions interrogated by the array (Wilcoxon non-parametric test; p<0.001, D=-0.29, and p<0.001, D=-0.5, respectively) (**Fig. 2b**). Consequently, the 5hmC sites were enriched in non-CpG islands (non-CGI) (chi-square test; p<0.001; OR=1.93, and p<0.001, OR=3.45, respectively) and infrequent in CGIs (chi-square test; p<0.001, OR=0.14, and p<0.001, OR=0.13) (**Fig. 2c**). With respect to genes, 5hmC sites were enriched in introns (chi-square test; p<0.001, OR=1.82, and p<0.001, OR=1.76, respectively), but were less frequent than expected in gene promoters (chi-square test; p<0.001, OR=0.58, and p<0.001, OR=0.6) (**Fig. 2d**).

To identify possible chromatin marks associated with 5hmC sites, we compared these CpG sites with previously published data on a range of histone modifications and chromatin modifiers in 10 different cell types (see Methods) (**Fig. 2e**). This approach identified statistically significant associations (Fisher´s exact test; p<0.05) between the 5hmC sites in brain and the active histone marks H3K4me1, H3K36me3, and H4K20me1 (**Fig. 2e**). A similar framework was used to test for the enrichment of our selected probes over the computer-generated chromatin segmentation states from the ENCODE ChromHMM project (see Methods). In total, fifteen states were used to segment the genome, and these were then grouped and colored to highlight predicted functional elements. This approach showed that the hmC sites were significantly enriched in states associated with enhancers and transcription (Fisher´s exact test; p<0.05) (**Fig. 2f**). These associations were further corroborated by using available ChIP-seq tracks from epigenomes of 8 brain cell types obtained from the NIH Roadmap Epigenome consortia (30) (Fisher´s exact test; p<0.05) (**Supplementary Figure S2**).

Locus-specific alterations of 5hmC in glioma

To identify differentially hydroxymethylated CpG sites (d5hmC) at single copy sequences in cancer, we used 450K methylation arrays to analyze 9 primary tumors obtained from patients with glioma (see Methods). A total of 49,601 CpG sites that were hypo-hydroxymethylated were identified in gliomas, but almost no hyper-hydroxymethylated sites were found (see Methods) (**Fig. 3a and Supplementary Table 2**). To validate the results obtained with the methylation arrays with an alternative technique nondependent on the oxidative bisulfite conversion of DNA, we randomly selected five of the 100 most variable sequences previously identified and analyzed their

5hmC status using a hydroxymethylated DNA Immunoprecipitation (hMeDIP) Kit (Epigentek; see Methods) in 5 brain samples and 5 gliomas**.** The results corroborated the massive loss of 5-hydroxymethylation in glioma in all the candidate sequences (Mann Whitney test; p<0.05) (**Supplementary Figure S3**).

Hierarchical clustering using the differentially hydroxymethylated CpG sites showed the correct classification of normal and tumor samples (**Fig. 3b**). The analysis of the genomic distribution of the hypo-hydroxymethylated CpG sites in gliomas showed an enrichment at low CpG density regions (Wilcoxon rank sum test, p<0.001, D=-0.41), and consequently at non-CpG islands (chi-squared test, p<0.001, OR=2.53) (**Fig. 3c**). With respect to gene location, hypo-hydroxymethylation was more frequent in introns (chi-squared test, p<0.001, OR=1.77) (**Fig. 3c**).

To identify possible chromatin signatures associated with DNA hypo-hydroxymethylation in gliomas, we compared our list of hypo-hydroxymethylated CpG sites with previously published data on a range of histone modifications and chromatin modifiers in 10 different cell types (see Methods) (**Fig. 3d**). Interestingly, this approach showed an enrichment of hypo-hydroxymethylation at chromatin regions marked with the activating histone PTMs H3K4me1, H3K36me3, H4K20me1 and H3K79me2 (Fisher's exact test, p<0.05) (**Fig. 3d**), but not with the repressive histone modification H3K27me3, which has been previously shown to be associated with aberrant DNA hypermethylation in cancer (31, 32) (**Fig. 3d**). A similar framework was used to test for the enrichment of our selected probes over the computer-generated chromatin segmentation states from the ENCODE ChromHMM project. Using this approach, we found that hypohydroxymethylated CpG sites were significantly associated with transcription regulation and enhancers (Fisher's exact test; p < 0.05) (**Fig. 3e**). These associations were further corroborated using available ChIP-seq tracks from the epigenomes of 8 types of brain cell obtained from the NIH Roadmap Epigenome consortia (30) (Fisher´s exact test; p<0.05) (**Supplementary Figure S4**).

DNA hypo-hydroxymethylation identifies a novel type of non-canonical aberrant DNA hyper-methylation in glioma

To study the relationship between changes in 5mC and 5hmC in glioma, we first identified aberrantly methylated CpG (d5mC) sites. The comparison of the methylation data between tumoral and control samples (see Methods) identified 2,727 hypo- and 12,050 hyper-methylated CpG sites in gliomas (**Supplementary Tables 3 and 4**). Next, we compared these d5mC sites with the previously identified hypo-hydroxymethylated CpG sites (**Fig. 3a, Supplementary Table 2**). This approach showed that 4,627 (38.4%) of the CpG sites aberrantly hypermethylated in gliomas also lost 5hmC (**Fig. 4a, Supplementary Table 5**). Interestingly, those CpG sites were those that showed the highest values of 5hmC in normal tissue (**Supplementary Fig. S5)** (Wilcoxon rank sum test, p< 0.001).

To verify that these hypermethylated CpG sites that also lose 5hmC in gliomas (hyper5mC-hypo5hmC) had not been identified in previous studies, owing to no distinction being made between 5mC and 5hmC, we compared the DNA methylation values of our samples with those of gliomas obtained from the same type of methylation array available in TCGA (33). We observed that not separating 5mC and the 5hmC resulted in many false negatives for hypermethylation in gliomas since the gain of 5mC in tumors was masked by the high levels of 5hmC in normal brain (**Supplementary Fig. S6a**). However, irrespective of the issue of separating or not 5mC and 5hmC, similar results were found when we performed the same comparisons using hypermethylated CpG sites that showed no changes in 5hmC in gliomas, since, in this case, these CpGs showed very low levels of 5hmC in normal brain (**Supplementary Fig. S6b**).

To investigate, at a functional genomic level, the characteristics of these two classes of aberrantly hypermethylated CpG sites in gliomas we first analyzed their genomic distribution in relation to density of CpG sites and we found that the hypermethylated CpG sites that lose 5hmC (hyper5mC-hypo5hmC) were enriched in low density CpG regions (Wilcoxon rank sum test, p<0.001, D=-0.11) as compared with the hypermethylated CpG sites that showed no changes in 5hmC (hyper5mC) (Wilcoxon rank sum test, p<0.001, D=-0.23) (**Fig. 4b, Supplementary Tables 5 and 6**). Furthermore, hyper5mC-hypo5hmC sites were strongly depleted from CGIs (chi-squared test, p<0.001, OR=0.42) and enriched in CGI shores (chi-squared test, p<0.001, OR=2.03) (**Fig. 4b**). Hierarchical clustering using the differentially methylated CpG sites showed that the hyper5mC-hypo5hmC sites were slightly more methylated in control brain samples than the hyper5mC sites, and that they were more uniformly hypermethylated in glioma (**Fig. 4c**). To further corroborate our results, we took advantage of recently published data on the whole-genome bisulfite sequencing (WGBS), albeit data was from one glioblastoma patient only (34). We found that, in addition to a large percentage of CpGs (n: 4,051; 88%) showing the same patterns of change as in our methylation arrays, the WGBS analysis identified more than 10^6 new hyper5mC-hypo5hmC sites, thus confirming that this is a frequent event in glioma (**Supplementary Fig. S7**).

Next, to identify possible chromatin signatures associated with the two classes of aberrantly hypermethylated CpG sites in gliomas, we compared our data with previously published data on a range of histone modifications and chromatin modifiers in 10 different cell types (see Methods) (**Fig. 5a**). This approach confirmed the association between hyper5mC and the repressive histone marks H3K9me3 and H3K27me3 (Fisher's exact test, $p<0.05$) (31, 32, 35). The hyper5mC-hypo5hmC sites showed a completely different chromatin signature, with enrichment in the activating histone PTMs H3K4me1, H3K36me3, H3K79me2 and H4K20me1 (Fisher's exact test, $p<0.05$) (**Fig. 5a**). Notably, as compared with the chromatin signature of the whole set of hypo-hydroxymethylated CpGs in glioma, these CpG sites were particularly enriched at the H3K4me2 histone mark (Fisher's exact test, $p<0.001$, OR in [1.19, 1.78] for all cell lines in the Broad Histone project) (**Fig. 5b**).

These results indicate that the hyper5mC sites behave like the aberrantly hypermethylated canonical CpG sites in cancer (i.e., enriched in CGIs and repressive histone marks), whilst the hyper5mC-hypo5hmC sites represent a novel and functionally different non-canonical type of aberrantly methylated DNA sequence in glioma (**Fig. 5a**, **5b**, **Supplementary Tables 5 and 6**). In support of this notion, experiments focused on the computational prediction of functional elements confirmed the enrichment of canonical aberrant hypermethylation in promoters and repressed sequences and revealed a completely different pattern for non-canonical hypermethylation, one which is more closely associated with enhancers and transcriptional regulation (Fisher’s exact test; $p < 0.05$) (**Supplementary Fig. S8**). These associations were also corroborated by using available ChIP-seq tracks from epigenomes of 8 types of brain cell obtained from the NIH Roadmap Epigenome consortia (30) (Fisher´s exact test; $p<0.05$) (**Supplementary Figure S9**).

Distinct functional role of canonical and non-canonical aberrant hypermethylation in glioma

To identify possible differences between the functional role of canonical and non-canonical aberrant DNA hypermethylation in glioma we first ascribed CpG sites to specific genes and then used HOMER to carry out gene ontology analyses of each group of genes (see methods). Using this approach, we identified 1,921 genes displaying canonical hypermethylation, 2,042 displaying non-canonical hypermethylation and 938 displaying both types of aberrant hypermethylation (**Fig. 6a, Supplementary Tables 7, 8 and 9**). As expected, GO analyses showed an enrichment of development and differentiation processes in canonical genes (36) (**Fig. 6a, Supplementary Table 10**). In contrast, non-canonical genes were enriched in cell signaling and protein processing pathways (**Fig. 6a, Supplementary Table 11**).

To further investigate the functional role of canonical and non-canonical hypermethylation in cancer, we compared our methylation data with previously published gene expression data in the same type of tumor (see Methods). Results showed that 681 (23.8%) of the canonical and 585 (19.6%) of the non-canonical aberrantly hypermethylated genes were repressed in gliomas (**Fig. 6b**).

Genomic distribution analysis of both types of aberrant hypermethylation confirmed the enrichment of canonical hypermethylation in exons (chi-squared test, p<0.001, OR=1.79 for general exons, OR=2.01 for first exons), while non-canonical hypermethylation was more frequent in introns (chi-squared test, p<0.001, OR=1.7) (**Fig. 6c**). The genes frequently downregulated in glioma, *SLC14A* and the *SMAD7*, represent two bona fide examples of this pattern of non-canonical aberrant hypermethylation (**Fig. 6d, Supplementary Fig. S10**).

Taken as a whole, these results indicate that both types of aberrant hypermethylation have a similar effect on gene expression, but that they affect different types of genes and gene regions.

Discussion

During recent decades, it has largely been accepted that aberrant genomic DNA methylation is a hallmark of cancer (1, 2) and the best-known DNA methylation alterations in tumors were the aberrant hypermethylation of CpG island promoters, and global DNA hypomethylation. In both cases, the alterations were mostly attributed to changes in the overall content and genomic distribution of 5mC (1, 2).

The vast majority of studies on DNA methylation and cancer have been based on the sodium bisulfite modification of the genomic DNA, a chemical reaction that allows C and 5mC to be distinguished by polymerase chain reaction (37). However, this approach cannot distinguish between 5mC and 5-hydroxymethylcytosine (5hmC), the latter being a chemical modification of the cytosine first identified in bacteriophages in 1952 (38), and which has recently been found to be quite abundant in specific mammalian tissue (3). 5hmC is synthesized from 5mC by the Ten-eleven Translocation (Tet) Enzymes, a family of proteins that can also catalyze the successive conversion of 5hmC to 5-formylcytosine and then to 5-carboxylcytosine, both of which can be transformed to unmodified C (39). Although 5hmC was originally described as simply a demethylation intermediate of C (7, 12, 13), recent data suggest that this may be an epigenetic mark in its own right (40, 41). Thus, as most previous studies did not distinguish between 5mC and 5hmC, and it appears that DNA hydroxymethylation might play a specific role in cancer, in this work we aimed to re-evaluate changes in DNA methylation in cancer, paying special attention to the specific contribution of 5hmC.

To identify the DNA regions affected by hydroxymethylation changes in cancer, we first focused on four types of repeated DNA (LINE1, Sat-alpha, NBL2 and AluYb8). Among them, the LINE1 repeat is of particular interest because it contains almost 20% of the genomic 5mC, and it has been proposed to be a surrogate of global DNA methylation (26). Our results confirmed that tumors lose 5mC at repeated DNA (42). However, the level of 5hmC at repeated DNA in healthy samples was very low and no significant differences were observed compared to tumors, which indicates that the global DNA hypo-hydroxymethylation previously observed in cancer (14, 20, 27, 28, 43) does not principally occur at repeated DNA. As changes in 5hmC at repeated DNA could not explain the global differences previously observed by mass spectrometry (14, 20, 27, 28), we decided to study the possible contribution of single copy sequences. Genome-wide profiling of 5mC and 5hmC of healthy tissue has identified more than 100,000 CpG sites frequently hydroxymethylated in brain, providing evidence that the level of this epigenetic mark is very abundant in this tissue (3, 7-11). Moreover, 5hmC was enriched in specific regions, i.e. those with low CpG density and in introns, indicating that 5hmC is not simply a demethylation intermediate (7, 12, 13). Interestingly, 5hmC co-localized in regions marked with the activating histone PTM H3K4me1. This histone mark has been previously associated with gene enhancers (44, 45), which suggests that DNA hydroxymethylation might play a role in gene regulation in trans. Moreover, we have recently found an association between H3K4me1 and DNA hypomethylation during aging in stem and differentiated cells (46), which may represent an interesting link between aging and cancer at these genomic regions.

The cell type(s) from which glioblastomas originate is not well understood at this moment in time, although there is some evidence that they might be neuronal stem cells or a glial precursor (47, 48). In our study, normal brain cells were obtained from the frontal cortex, which principally comprises neural and glial cells, thus we cannot rule out some differences found in our analysis possibly reflecting differences between normal cell types. That said, the great number of hypo-hydroxymethylated single CpG sites in glioma could explain the global differences previously observed by mass spectrometry (14, 20, 27, 28) and suggests that, in contrast to 5mC, most DNA hypo-hydroxymethylation in brain tumors occurs at single copy sequences.

The behavior of 5hmC led us to next identify two types of CpG sites aberrantly hypermethylated in glioma: i. aberrantly hypermethylated CpG sites that showed no changes in 5hmC; and ii. hypermethylated CpG sites that lose 5hmC. This negative correlation between changes in 5mC and in 5hmC agrees with that previously found in cancer differentially methylated regions (c-DMRs) of both liver and lung tumors (49).

The former sites display similar chromatin signatures to previously described genes aberrantly hypermethylated in cancer (i.e. enrichment in the repressive histone marks H3K9me3 and H3K27me3) (31, 32, 35). In contrast, the latter type of aberrantly hypermethylated CpG sites were enriched in the activating histone PTMs H3K4me1, H3K36me3, H3K79me2, H4K20me1 and H3K4me2. As these CpG sites present a genuine chromatin signature which is different to the repressive chromatin signature of the classical genes aberrantly hypermethylated in cancer (31, 32, 35), we conclude that they represent a novel 5hmC-dependent non-canonical class of aberrant DNA hypermethylation in glioma (**Fig. 7**). As this gain in 5mC is inversely correlated with loss of 5hmC, it was not possible to identify this significant alteration in previous studies using the classical sodium bisulfite-based technologies, since they are not able to distinguish between the two chemical modifications.

Aberrant DNA hypermethylation in cancer was discovered more than 30 years ago, but the underlying molecular mechanisms are still poorly understood. For example, it has been proposed that genes enriched in bivalent histone modifications (H3K4me3 and H3K27me3) and polycomb group proteins during embryo development are prone to become aberrantly hypermethylated in cancer (31, 32, 35) but the molecular basis of this is unknown. Our data suggest that tumor cells might in fact acquire aberrant DNA methylation through various different pathways. Moreover, in the case of the non-canonical hypermethylation, the previous loss of 5hmC suggests that aberrant hypermethylation at these DNA regions could be due to an attempt by the cell to reverse or repair the loss of 5hmC at functionally sensible loci. This possibility is supported by the fact that the non-canonical aberrant hypermethylation described here seems to play an important role in gene regulation. Intriguingly, 5hmC at gene promoters has also been proposed to protect from aberrant hypermethylation in colorectal cancer (28). Thus, although it seems that 5hmC plays an important role in the regulation of the DNA methylation changes in cancer, more research is needed to fully understand its role.

The non-canonical aberrant hypermethylation described here seems to have a similar overall effect on gene expression as classical canonical hypermethylation, although the type of genes and the genomic regions affected are very different. Previous research has shown that the repression of developmental genes affected by canonical aberrant hypermethylation promotes tumorigenesis (36). However, the possible functional role of disruption of cell signaling and protein processing pathways affected by the non-canonical hypermethylation described in this study remains to be elucidated. Future research is thus needed to address this issue, and to determine whether the two types of aberrant DNA hypermethylation have distinct functional roles in cancer.

Methods

Normal samples and primary tumors

Brain samples analyzed in this study were collected at the Hospital Universitario Central de Asturias (HUCA), the Hospital Virgen de la Salud, Toledo, and the Hospital Universitario Puerta de Hierro, Madrid. The samples studied comprised 12 normal brains and 53 glioblastomas. The study was approved by the Clinical Research Ethics Committee and all the individuals involved provided written informed consent.

Pyrosequencing assays

5mC and 5hmC patterns at repetitive sequences (LINE1, AluYb8, Sat-alpha and NBL2) were analyzed by pyrosequencing using previously described primers (25). To calculate 5hmC levels, each sample was analyzed using two methods performed in parallel; an oxidative bisulfite conversion (oxBS) and a bisulfite-only conversion (BS), in accordance with the TrueMethyl® Array Kit User Guide (CEGX, Version 2) with some modifications. Briefly, DNA samples were cleaned using Agencourt AMPure XP (Beckman Coulter) then oxidated with 1 μL of a KRuO4 (Alpha Aeser) solution (375 mM in 0.3 M NaOH), after which bisulfite conversion was performed using EpiTect bisulfite kit (Qiagen®).

After PCR amplification of the region of interest in oxBS and BS samples, pyrosequencing was performed using PyroMark Q24 reagents, and vacuum prep workstation, equipment and software (Qiagen®). To avoid negative methylation values due to the substraction of the oxBS and BS signals, 5mC and 5hmC estimations were calculated by means of a maximum likelihood model using the OxyBS R CRAN package (version 1.5) (50). Briefly, percentages of CpG methylation obtained from the PyroMark Q24 software were used as beta values for the BS or the oxBS treated samples, and signal intensities from the oxBS and BS experiments were obtained from the peak height signals of the corresponding nucleotides measured in the pyrosequencing reaction.

Genome-wide DNA methylation analysis with high-density arrays

Microarray-based DNA methylation profiling was performed with the HumanMethylation 450 BeadChip (51). Oxidative bisulfite (oxBS) and bisulfite-only (BS) conversion was performed using the TrueMethyl® protocol for 450K analysis (Version 1.1, CEGX) following the manufacturer's recommended procedures.

Processed DNA samples were then hybridized to the BeadChip (Illumina), following the Illumina Infinium HD Methylation Protocol. Genotyping services were provided by the Spanish Centro Nacional de Genotipado (CEGEN-ISCIII) (www.cegen.org). Array data were deposited in ArrayExpress accession numbers E-MTAB-6003.

HumanMethylation450 BeadChip data preprocessing

Raw IDAT files were processed using the R/Bioconductor package minfi (52) (version 1.14.0), implementing the SWAN algorithm (53) to correct for differences in the microarray probe designs. No background correction or control probe normalization was applied. Probes where at least two samples had detection p-values > 0.01, and samples where at least 5500 probes had detection p-values > 0.01 were filtered out. M-values and beta values were computed as the final step in the preprocessing procedure. In line with a previously published methodology (54), M-values were used for the statistical analyses and beta values for effect size thresholding, visualization and report generation.

Batch effect correction

In order to detect whether there was any batch effect associated with technical factors, the visualization technique of multidimensional scaling (MDS) was employed to highlight any strange interaction affecting the different samples. Where necessary, posterior adjustment of the samples was performed by means of the SVA method (55) implemented in the R/Bioconductor sva package (version 3.14.0).

Computation of hydroxymethylation levels

Beta values from oxBS samples were subtracted from their corresponding BS treated pairs, generating an artificial dataset representing the level of 5hmC for each probe and sample as per a previously published methodology (56). One further dataset was created to represent the 5mC levels using beta values from oxBS samples.

Detection of differentially methylated probes

Differential methylation and hydroxymethylation of an individual probe was determined by a moderated t-test implemented in the R/Bioconductor package limma (57). A linear model, with methylation or hydroxymethylation levels as response and the sample group (normal/tumoral) as the principal covariate of interest, was then fitted to the methylation or hydroxymethylation data. Surrogate Variables generated using SVA were also included in the model definition but excluding those found to be correlated to the phenotype of interest. P values were corrected for multiple testing using the Benjamini-Hochberg method for controlling false discovery rate (FDR). An FDR threshold of 0.001 was employed to determine differentially methylated and hydroxymethylated probes. Additionally, these probes were filtered according to their effect size, keeping only those probes with methylation or hydroxymethylation changes between-groups which exceeded the median of all differences for the same comparison. The probes without no significant 5hmC signal on control samples were filtered out from the set of hypo-hydroxymethylated probes in glioma.

In order to describe the genomic distribution of 5hmC in brain, we used Hilbert curves (58), which are especially suited to the visualization of simple measurements, such as the location of the 5hmC enriched probes, over large scales. These curves allow for compact representation on a genomic scale and have an interesting property by which two points which are near each other in the one-dimensional genomic location space are also closely located in the two-dimensional transformation generated by the curve. The converse may not always be true.

Identification of hydroxymethylated probes

In order to identify those probes representing the regions where the 5hmC mark is located, a differential hydroxymethylation analysis was performed as described previously (59) using a dataset containing both oxBS and BS versions of the control samples. Probes with significant differences in beta values between the BS and oxBS samples were considered to be enriched for the 5hmC mark. An FDR threshold of 0.001 was employed. No filtering on effect size was applied in this case.

5-hydroxymethylcytosine immunoprecipitation-qPCR assay

Immunoprecipitation of 5hmC was carried out using the EpiQuik Hydroxymethylated DNA Immunoprecipitation (hMeDIP) Kit (Epigentek), according to the manufacturer's instructions.

Input, non-specific IgG- and 5hmC-enriched fractions were obtained from eleven samples corresponding to five normal brains, five glioma tumors and one glioma cell line. All these fractions were amplified by qPCR with oligonucleotides specific for the CpGs detailed in **Supplementary Table 12.** After confirming there were no significant differences between input DNAs, 5hmC relative enrichment was calculated as a Fold Change relative to Input Ct Mean.

Histone enrichment analysis

In order to analyze the enrichment of histone marks on a subset of probes, we used the information contained in the UCSC Genome Browser Broad Histone tracks from the ENCODE Project (10 different cell types) and the NIH Roadmap Epigenome consortium (8 brain cell types, but less available ChIP-seq data). Histone mark peaks were downloaded for every combination of cell line and antibody. For each track, a 2x2 contingency table was built to represent the partition of the whole set of possible probes in the microarray with respect to the membership of the subset of interest and the overlap between the probes and the histone peaks. A Fisher's exact test was used to determine whether there was significant enrichment of the selected histone mark for the subset of interest. P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method for controlling FDR. A significance level of 0.05 was used to determine whether the given combination of histone mark and cell line presented a significant change in proportion. Additionally, the base-2 logarithm of the Odds Ratio (OR) was used as a measure of effect size.

Chromatin segment enrichment analysis

Data from the BROAD ChromHMM Project were downloaded from the UCSC Genome Browser and the NIH Roadmap Epigenome consortium. Each of the tracks comprising these datasets represents a different segmentation (15 and 18 chromatin states respectively) generated by a Hidden Markov Model (HMM) using Chip-Seq signals from the Broad Histone Project as inputs. The segmentations were later curated and labelled according to their functional status (60, 61). In order to detect any significant enrichment in the proportion of probes in a given subset of interest belonging to one functional category, an analysis strategy similar to the one employed for the detection of histone enrichment was performed. In this case, a 2x2 contingency table was built using segments of a given functional status rather than antibodies. A Fisher's exact test was employed, and significant combinations were detected using a FDR threshold of 0.05 (Benjamini-Hochberg procedure). Again, the base-2 logarithm of the OR was used as a measure of effect size.

Genomic region analysis

The probes in the microarray were assigned to a genomic region according to their position relative to the transcript information extracted from the R/Bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene (package version 3.1.2). A probe was said to be in a promoter region if it was located in a region up to 2kb upstream of the transcription start site (TSS) of any given transcript. Similarly, a set of mutually exclusive regions were defined inside the transcripts, namely 5UTR, 3UTR, First Exon, Exon and Intron. A probe could only belong to one category, hence if the location of a probe overlapped with two or more regions in different transcripts, it was assigned to the region with a higher level of precedence (i.e. in the order stated above, earlier mention indicates higher precedence). If a probe was not assigned to any of these special regions, it was labelled by default as Intergenic. A contingency table was built for each of the subsets, partitioning the whole set of probes according to membership to a given category and the subset of interest. A Pearson's $\chi 2$ test was used to determine whether there was any significant change in proportion between the number of probes marked as belonging to a given region inside and outside the subset of interest. A significance level of 0.05 was employed, and effect size measured by OR.

CGI status analysis

Similar to the genomic region analysis, probes were labelled according to their relative position to CpG-islands (CGIs), the locations of which were obtained from the R/Bioconductor package FDb.InfiniumMethylation.hg19 (package version 2.2.0). The generation procedure of these CGIs is described by (62), i.e. ‘CpG shores’ were defined as the 2kbp regions flanking a CGI. ‘CpG shelves’ were defined as the 2kbp regions either upstream of or downstream from each CpG shore. Probes not belonging to any of the regions thus far mentioned were assigned to the special category ‘non-CGI’ with each probe being assigned to only one of the categories. A 4x2 contingency table was constructed for each subset of probes in order to study the association between the given subset and the different CGI categories. A $\chi 2$ test was used to determine whether any of the categories had a significant association with the given subset. For each of the CGI status levels, a 2x2 contingency table was defined and another $\chi 2$ test used to independently evaluate the association of the given subset with each status level, a significance level of 0.05 being employed for all tests. Effect size was reported as the OR for each of the individual tests.

Analysis of CpG density

For each of the probes in the HumanMethylation450 microarray, CpG density was measured as the number of CG 2-mers present divided by the number which would be theoretically possible in a 2kbp window with the CpG under study at its centre. A Wilcoxon non-parametric test was used to determine if any significant difference existed between the CpG density of each subset of interest and that of the array probes in the background. A significance level of 0.05 was employed for all tests. Effect size was measured using Cliff's Delta (D).

Gap distance analysis

Distance to both the centromere and telomere was measured for each of the probes in the HumanMethylation450 microarray. In order to find significant differences between the probes within the subset of interest and those in the background, a Wilcoxon non-parametric test was used. Once again, a significance level of 0.05 was employed for all tests, and Cliff's Delta (D) was used as a measure of effect size.

Microarray background correction

Although it is sometimes referred to as a genome-wide solution, the HumanMethylation450 BeadChip only covers a fraction of the entire genome. In its 27K predecessor, the probes were mainly located at gene promoter regions, while the newer HumanMethylation450 BeadChip additionally includes probes located inside genes and in intergenic regions (63).

The irregular distribution of probes can however lead to unwanted biases when studying whether a selected subset of probes is enriched with respect to any functional or clinical mark. For this reason, here a reference to the background distribution of features was included in all statistical tests performed in order to prevent our conclusions from being driven by the irregular distribution of probes. In qualitative tests (CGI status, genomic region, or histone mark enrichment), the contingency matrix was built to represent the background distribution of the microarray. In quantitative tests (CpG density, distance to centromeres and telomeres) the corresponding metric was compared between the subset of interest and the remaining probes in the microarray. Thus, any significant result would indicate a departure from the fixed background distribution and ignore any bias inherent in the test.

Gene ontology analysis and annotation

Probe sets were converted to gene sets by using the annotation information from the R/Bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene (version 3.1.2). A probe was assigned to a gene if the probe was contained within the overlap of all the genomic regions represented by the different transcripts belonging to that gene, or in a 2kbp region upstream of the corresponding TSS. Probes converted this way can be assigned to one or more genes, or to zero (i.e. intergenic probes).

After gene conversion, each subset of interest was analyzed using the HOMER software tool (64). The software was configured to use the whole set of genes represented in the HumanMethylation450 architecture as a background. HOMER tested the genes in each subset of interest against 21 different databases, including the Gene Ontology (GO) Biological Process, Molecular Function and Cellular Component ontologies, as well as KEGG and Reactome pathway databases, among many others.

Circular visualization and track smoothing

In order to plot the CpG and histone peak information on the circular genome-wide and example graphs, smoothing was applied to the data. CpG enrichment information for canonical and non-canonical hypermethylation was generated by partitioning the genome into intervals of 10kbp and assigning to each a score corresponding to the average coverage of the selected CpGs in the interval.

Whole-genome bisulfite sequencing (WGBS) datasets

Tet-assisted bisulfite sequencing (TAB-Seq) data, corresponding to an adult brain prefrontal cortex tissue sample (GSM1135082) (29), was used as a validation dataset for the location of 5hmC in controls.

Additionally, TrueMethyl (ox-BS) Whole Genome data referenced in (34) (E-MTAB-5171), obtained from a single glioblastoma patient, was used as a validation dataset. Previously processed data in the form of quantified methylation for each CpG measured in both strands of the genome was downloaded and filtered. The resulting dataset comprised only two samples (normal and tumoral), hence a descriptive strategy was used to distinguish the different types of probe according to their methylation status.

For both the TAB-Seq and TrueMethyl-seq validation datasets, hydroxymethylated probes were identified as those having a 5hmC measure higher than 0.1. In the case of

WGBS, differentially methylated probes were defined as those having an absolute difference in their methylation values between the control and tumor samples which was above a given threshold (0.2 for 5mC and 0.1 for 5hmC). Only methylation measures from CpGs having a total read count higher than 10 were retained.

The validation datasets may contain either one or two methylation measures for each CpG in the genome as they measure methylation in both strands. Strand- agnostic CpG regions representing the CpG dinucleotides with at least one measure were defined in order to compute the degree of intersection between the WGBS and methylation arrays results.

The Cancer Genome Atlas (TCGA) expression and methylation dataset

In order to analyze changes in gene expression, samples of glioblastoma multiforme (GBM) were selected from among the data generated by the TCGA Research Network (http://cancergenome.nih.gov). DNA Methylation data for GBM was additionally obtained from TCGA for visualization and validation purposes.

Expression Level-3 pre-processed data was obtained for 572 GBM samples (10 controls and 562 tumors). The moderated t-test approach in the R/Bioconductor package *limma* was used to assess the differential expression status of each gene in the TCGA datasets. The normalized expression ratio in the TCGA datasets was used as the response variable, and the sample group (normal/tumoral) as the covariate of interest. No adjustment for possible confounders was performed in this case. An FDR threshold of 0.001 was used to correct for multiple hypotheses. No filtering on effect size was applied in this case.

DNA Methylation Level-1 raw data for the Illumina 450k architecture was obtained for 162 GBM samples (33). The raw values were normalized using the SWAN algorithm. No additional filtering was performed on the samples.

Data analysis workflow

All the necessary steps for upstream and downstream analyses were defined and implemented using the Snakemake tool (65), which helps data scientists to generate a reproducible and inherently parallel processing pipeline. Individual workflow tasks were implemented in R (version 3.2.2) and Python (version 3.4.3).

Acknowledgments

We thank Ronnie Lendrum for editorial assistance. We also thank the Tumor Bank of the Hospital Virgen de la Salud (BioB-HVS, Toledo, Spain) for providing tumor samples. This work has been financially supported by: the Plan Nacional de I+D+I 2013-2016/FEDER (PI15/00892 to M.F.F. and A.F.F.; RTC-2015-3393-1 to A.F.F.); the ISCIII-Subdirección General de Evaluación y Fomento de la Investigación, and the Plan Nacional de I+D+I 2008-2011/FEDER (CP11/00131 to A.F.F.); IUOPA (to G.F.B. and M.S); the Fundación Científica de la AECC (to R.G.U.); the Fundación Ramón Areces (to M.F.F); FICYT (to E.G.T., M.G.G. and A.C.); and the Asturias Regional Government (GRUPIN14-052 to M.F.F.). Work in P.M. lab is supported by the European Research Council (CoG-2014-646903), the Spanish Ministry of Economy-Competitiveness (SAF-SAF2013-43065), the Obra Social La Caixa-Fundaciò Josep Carreras, and the Generalitat de Catalunya. P.M. is an investigator in the Spanish Cell Therapy cooperative network (TERCEL). The IUOPA is supported by the Obra Social Cajastur-Liberbank, Spain.

Competing interests

The authors declare that they have no competing interests.

References

1 Esteller, M. (2005) Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol*, **45**, 629-656.

2 Feinberg, A.P. and Tycko, B. (2004) The history of cancer epigenetics. *Nat Rev Cancer*, **4**, 143-153.

3 Kriaucionis, S. and Heintz, N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, **324**, 929-930.

4 Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930-935.

5 Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. and Zhang, Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129-1133.

6 Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S. and Reik, W. (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, **473**, 398-402.

7 Globisch, D., Munzel, M., Muller, M., Michalakis, S., Wagner, M., Koch, S., Bruckl, T., Biel, M. and Carell, T. (2010) Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One*, **5**, e15367.

8 Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M. *et al.* (2012) 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nat Struct Mol Biol*, **19**, 1037-1043.

9 Li, W. and Liu, M. (2011) Distribution of 5-hydroxymethylcytosine in different human tissues. *J Nucleic Acids*, **2011**, 870726.

10 Nestor, C.E., Ottaviano, R., Reddington, J., Sproul, D., Reinhardt, D., Dunican, D., Katz, E., Dixon, J.M., Harrison, D.J. and Meehan, R.R. (2012) Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res*, **22**, 467-477.

11 Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X. *et al.* (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol*, **29**, 68-72.

12 Klug, M., Schmidhofer, S., Gebhard, C., Andreesen, R. and Rehli, M. (2013) 5-Hydroxymethylcytosine is an essential intermediate of active DNA demethylation processes in primary human monocytes. *Genome Biol*, **14**, R46.

13 Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A. and Balasubramanian, S. (2014) 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem*, **6**, 1049-1055.

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

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

16 Putiri, E.L., Tiedemann, R.L., Thompson, J.J., Liu, C., Ho, T., Choi, J.H. and Robertson, K.D. (2014) Distinct and overlapping control of 5-methylcytosine and 5-hydroxymethylcytosine by the TET proteins in human cancer cells. *Genome Biol*, **15**, R81.

17 Sun, M., Song, C.X., Huang, H., Frankenberger, C.A., Sankarasharma, D., Gomes, S., Chen, P., Chen, J., Chada, K.K., He, C. *et al.* (2013) HMGA2/TET1/HOXA9 signaling pathway regulates breast cancer growth and metastasis. *Proc Natl Acad Sci U S A*, **110**, 9920-9925.

18 Yang, H., Liu, Y., Bai, F., Zhang, J.Y., Ma, S.H., Liu, J., Xu, Z.D., Zhu, H.G., Ling, Z.Q., Ye, D. *et al.* (2013) Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene*, **32**, 663-669.

19 Haffner, M.C., Chaux, A., Meeker, A.K., Esopi, D.M., Gerber, J., Pellakuru, L.G., Toubaji, A., Argani, P., Iacobuzio-Donahue, C., Nelson, W.G. *et al.* (2011) Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget*, **2**, 627-637.

20 Kraus, T.F., Globisch, D., Wagner, M., Eigenbrod, S., Widmann, D., Munzel, M., Muller, M., Pfaffeneder, T., Hackner, B., Feiden, W. *et al.* (2015) Low values of 5-hydroxymethylcytosine (5hmC), the "sixth base," are associated with anaplasia in human brain tumors. *Int J Cancer*, **131**, 1577-1590.

21 Kudo, Y., Tateishi, K., Yamamoto, K., Yamamoto, S., Asaoka, Y., Ijichi, H., Nagae, G., Yoshida, H., Aburatani, H. and Koike, K. (2012) Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. *Cancer Sci*, **103**, 670-676.

22 Liu, C., Liu, L., Chen, X., Shen, J., Shan, J., Xu, Y., Yang, Z., Wu, L., Xia, F., Bie, P. *et al.* (2013) Decrease of 5-hydroxymethylcytosine is associated with progression of hepatocellular carcinoma through downregulation of TET1. *PLoS One*, **8**, e62828.

23 Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S. and Reik, W. (2013) A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol*, **14**, R119.

24 Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F. *et al.* (2013) Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, **152**, 1146-1159.

25 Urdinguio, R.G., Bayon, G.F., Dmitrijeva, M., Torano, E.G., Bravo, C., Fraga, M.F., Bassas, L., Larriba, S. and Fernandez, A.F. (2015) Aberrant DNA methylation patterns of spermatozoa in men with unexplained infertility. *Hum Reprod*, **30**, 1014-1028.

26 Weisenberger, D.J., Campan, M., Long, T.I., Kim, M., Woods, C., Fiala, E., Ehrlich, M. and Laird, P.W. (2005) Analysis of repetitive element DNA methylation by MethyLight. *Nucleic Acids Res*, **33**, 6823-6836.

27 Kraus, T.F., Kolck, G., Greiner, A., Schierl, K., Guibourt, V. and Kretzschmar, H.A. (2012) Loss of 5-hydroxymethylcytosine and intratumoral heterogeneity as an epigenomic hallmark of glioblastoma. *Tumour Biol*, **36**, 8439-8446.

28 Uribe-Lewis, S., Stark, R., Carroll, T., Dunning, M.J., Bachman, M., Ito, Y., Stojic, L., Halim, S., Vowler, S.L., Lynch, A.G. *et al.* (2015) 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. *Genome Biol*, **16**, 69.

29 Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C. *et al.* (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol*, **15**, R49.

30 Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317-330.

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

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

33 Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray, B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M. *et al.* (2016) Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, **164**, 550-563.

34 Raiber, E.-A., Beraldi, D., Martinez Cuesta, S., McInroy, G.R., Kingsbury, Z., Becq, J., James, T., Lopes, M., Allinson, K., Field, S. *et al.* (2017) Base resolution maps reveal the importance of 5-hydroxymethylcytosine in a human glioblastoma. *npj Genomic Medicine*, **2**, 6.

35 McGarvey, K.M., Fahrner, J.A., Greene, E., Martens, J., Jenuwein, T. and Baylin, S.B. (2006) Silenced tumor suppressor genes reactivated by DNA demethylation do not return to a fully euchromatic chromatin state. *Cancer Res*, **66**, 3541-3549.

36 Easwaran, H., Johnstone, S.E., Van Neste, L., Ohm, J., Mosbruger, T., Wang, Q., Aryee, M.J., Joyce, P., Ahuja, N., Weisenberger, D. *et al.* (2012) A DNA hypermethylation module for the stem/progenitor cell signature of cancer. *Genome Res*, **22**, 837-849.

37 Herman, J.G., Jen, J., Merlo, A. and Baylin, S.B. (1996) Hypermethylation-associated inactivation indicates a tumor suppressor role for p15INK4B. *Cancer Res*, **56**, 722-727.

38 Wyatt, G.R. and Cohen, S.S. (1952) A new pyrimidine base from bacteriophage nucleic acids. *Nature*, **170**, 1072-1073.

39 Plongthongkum, N., Diep, D.H. and Zhang, K. (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet*, **15**, 647-661.

40 Chapman, C.G., Mariani, C.J., Wu, F., Meckel, K., Butun, F., Chuang, A., Madzo, J., Bissonette, M.B., Kwon, J.H. and Godley, L.A. (2015) TET-catalyzed 5-hydroxymethylcytosine regulates gene expression in differentiating colonocytes and colon cancer. *Sci Rep*, **5**, 17568.

41 Jiang, D., Zhang, Y., Hart, R.P., Chen, J., Herrup, K. and Li, J. (2015) Alteration in 5-hydroxymethylcytosine-mediated epigenetic regulation leads to Purkinje cell vulnerability in ATM deficiency. *Brain*, **138**, 3520-3536.

42 Ehrlich, M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239-259.

43 Orr, B.A., Haffner, M.C., Nelson, W.G., Yegnasubramanian, S. and Eberhart, C.G. (2012) Decreased 5-hydroxymethylcytosine is associated with neural progenitor phenotype in normal brain and shorter survival in malignant glioma. *PLoS One*, **7**, e41036.

44 Hon, G.C., Song, C.X., Du, T., Jin, F., Selvaraj, S., Lee, A.Y., Yen, C.A., Ye, Z., Mao, S.Q., Wang, B.A. *et al.* (2014) 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol Cell*, **56**, 286-297.

45 Torano, E.G., Bayon, G.F., Del Real, A., Sierra, M.I., Garcia, M.G., Carella, A., Belmonte, T., Urdinguio, R.G., Cubillo, I., Garcia-Castro, J. *et al.* (2016) Age-associated hydroxymethylation in human bone-marrow mesenchymal stem cells. *J Transl Med*, **14**, 207.

46 Fernandez, A.F., Bayon, G.F., Urdinguio, R.G., Torano, E.G., Garcia, M.G., Carella, A., Petrus-Reurer, S., Ferrero, C., Martinez-Camblor, P., Cubillo, I. *et al.* (2014) H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. *Genome Res*, **25**, 27-40.

47 Agnihotri, S., Munoz, D., Zadeh, G. and Guha, A. (2011) Brain tumor-initiating cells and cells of origin in glioblastoma. *Translational Neuroscience*, **2**, 331.

48 Alcantara Llaguno, S.R. and Parada, L.F. (2016) Cell of origin of glioma: biological and clinical implications. *Br J Cancer*, **115**, 1445-1450.

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

50 Houseman, E.A., Johnson, K.C. and Christensen, B.C. (2016) OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. *Bioinformatics*, **32**, 2505-2507.

51 Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288-295.

52 Fortin, J.P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M. and Hansen, K.D. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol*, **15**, 503.

53 Maksimovic, J., Gordon, L. and Oshlack, A. (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol*, **13**, R44.

54 Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

55 Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**, 1724-1735.

56 Stewart, S.K., Morris, T.J., Guilhamon, P., Bulstrode, H., Bachman, M., Balasubramanian, S. and Beck, S. (2015) oxBS-450K: a method for analysing hydroxymethylation using 450K BeadChips. *Methods*, **72**, 9-15.

57 Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**, e47.

58 Hilbert, D. (1891) Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen* **38**, 2.

59 Field, S.F., Beraldi, D., Bachman, M., Stewart, S.K., Beck, S. and Balasubramanian, S. (2015) Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). *PLoS One*, **10**, e0118202.

60 Ernst, J. and Kellis, M. (2011) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, **28**, 817-825.

61 Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43-49.

62 Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A. and Feinberg, A.P. (2010) Redefining CpG islands using hidden Markov models. *Biostatistics*, **11**, 499-514.

63 Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C. and Fuks, F. (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, **3**, 771-784.

64 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576-589.

65 Koster, J. and Rahmann, S. (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520-2522.

Figure legends

Figure 1. 5mC and 5hmC levels at repetitive DNA sequences in glioma. 5mC (a) and 5hmC (b) values of several repetitive regions (AluYb8, LINE-1, NBL-2, and Sat-alpha) measured by pyrosequencing in controls and gliomas. Individual CpG site values for each repeat are displayed, and a linear model including both “sample group” and “CpG site” as covariates was fitted. Significant p-values for any repetitive region are shown.

Figure 2. Characterization of DNA 5hmC in normal brain. (a) Bean plots showing differences between average Beta values of 5mC+5hmC (BS) and true 5mC (OxBS) values in normal brain. The diamond inside the bean indicates the mean value. Hilbert curve showing the amount and genomic distribution of 5hmC in brain. A level-9 Hilbert curve was used. Each region delimited by black lines represents a chromosome. A point in the image represents a genomic segment of approximately 950bp. A blue point indicates presence of at least one 5hmC enriched CpG in the given segment. A Hilbert curve including all CpG sites analyzed in the methylation array (450K) is also shown (b) Associations between 5hmC and CpG density. (c) Distribution of 5hmC CpG sites relative to CpG island status and compared to the array background (450K). (d) Distribution of 5hmC CpG sites relative to different genomic regions. (e) Heatmaps showing significant enrichment of the 5hmC CpG sites identified in brain, with different histone marks contained in the UCSC Browser Broad Histone track from the ENCODE project. Color code indicates the significant enrichment based on log2 odds ratio (OR). (f) Heatmaps showing significant enrichment of 5hmC CpG sites with fifteen “chromatin states” generated by a Hidden Markov Model (HMM). Color codes indicate the significant enrichment based on log2 odds ratio (OR).

Figure 3. Alterations of 5hmC in glioma. (a) Bar plot showing the number of d5hmC sites in glioma. (b) Unsupervised hierarchical clustering and heatmap including CpG sites with 5hmC loss in glioma (3,000 random probes). (c) Associations between 5hmC loss in glioma and density of CpGs (upper panel), CpG island status (middle panel), and different genomic regions (lower panel). (d) Heatmaps showing significant enrichment of hypo 5hmC CpGs identified in glioma with different histone marks contained in the UCSC Browser Broad Histone track from the ENCODE project. (e) Heatmaps showing significant enrichment of hypo 5hmC CpGs in gliomas with fifteen “chromatin states”

generated by a Hidden Markov Model (HMM) (right panel). Color codes indicate the significant enrichment based on log2 odds ratio (OR).

Figure 4. Relationships between changes in 5mc and 5hmc in glioma. (a) Euler diagram illustrating overlap of CpGs that lose 5hmC (hypo 5hmC) and gain 5mC (hyper 5mC) in glioma. (b) Associations between hypermethylated CpG sites that lose (or not) 5hmC and CpG density (upper panel) and CpG island status (lower panel), compared to the array background (450K). (c) Unsupervised hierarchical clustering and heatmap including 3000 randomly chosen CpG sites with 5mC changes (hyper- and hypomethylation) in glioma. Hypo- (purple) and non-hypo (orange) 5hmC overlapped CpGs are indicated by colored lines on the annexed track. Average beta methylation values are displayed from 0 (blue) to 1 (yellow).

Figure 5. Canonical and non-canonical hypermethylation in glioma. (a) Heatmaps showing significant enrichment of CpG sites in glioma which exclusively gain 5mC (canonical hypermethylation) (upper panel), and both lose 5hmC and gain 5mC (non-canonical hypermethylation) (lower panel), with different histone marks contained in the UCSC Browser Broad Histone track from the ENCODE project. Histone PTMs related to activation and repression are distinguished by colors as indicated in the key. (b) Circular representation of two representative chromosomes (12 and 17), indicating genomic location of canonical (orange) and non-canonical (purple) hypermethylation in glioma. Inner tracks display chromatin marks (H3K9me3, H3K27me3, and H3K4me2), generated for NH-A cells. Two examples of genes showing canonical and non-canonical hypermethylation associated with specific chromatin signatures are displayed below.

Figure 6. Functional role of canonical and non-canonical hypermethylation in glioma. (a) Euler diagrams showing number of genes associated with canonical hypermethylation, non-canonical hypermethylation, or both. On the right are representative gene ontology terms (Biological process) of genes associated with canonical (orange) and non-canonical (purple) hypermethylation, ranked by Q-value, and enrichment score (relative risk). (b) Euler diagram showing overlap of canonical and non-canonical hypermethylated genes with down-regulation. (c) Associations of canonical and non-canonical hypermethylation in glioma with different genomic regions. (d) Representative example of one gene (*SLC1A4*) showing non-canonical

hypermethylation in glioma (orange frame). Organization of the gene, locations of CpGs included in the methylation array (black dots), and transcription start site (TSS) are shown below. 5mC hypermethylation (blue to yellow) and 5hmC loss (gray to blue) in glioma are shown above. Whole genome bisulfite sequencing (WGBS) data (34) including all the CpG sites in the same region are shown on the right. The associated change in gene expression is displayed below.

Figure 7. Schematic representation of genomic regions and related histone marks associated with canonical and non-canonical DNA hypermethylation in glioma. CpG sites that suffered canonical hypermethylation are overrepresented in CpG islands (CGI) and in poised promoters and repressed regions enriched in H3K27me3 and H3K9me3. In contrast, non-canonical hypermethylated CpG sites are enriched in CGI shores, and enhancers and transcribed regions characterized for activating histone marks.

Abbreviations

5hmC: 5-hydroxymethylcytosine

5mC: 5-methylcytosine

CGI: CpG island

TET1, TET2, and TET3: ten-eleven translocation proteins 1, 2, and 3

TAB-Seq: Tet-Assisted Bisulfite Sequencing

OR: Odd ratio

ChIP-seq: Chromatin Immunoprecipitation Sequencing

d5hmC: differentially hydroxymethylated CpG sites

hMeDIP: hydroxymethylated DNA Immunoprecipitation

PTMs: Post-translational modifications

TCGA: The Cancer Genome Atlas

WGBS: whole-genome bisulfite sequencing

GO: Gene ontology

c-DMRs: differentially methylated regions

oxBS: oxidative bisulfite conversion

BS: bisulfite conversion

SWAN: Subset-quantile Within Array Normalization

MDS: multidimensional scaling

SVA: Surrogate variable analysis

FDR: false discovery rate

GBM: glioblastoma

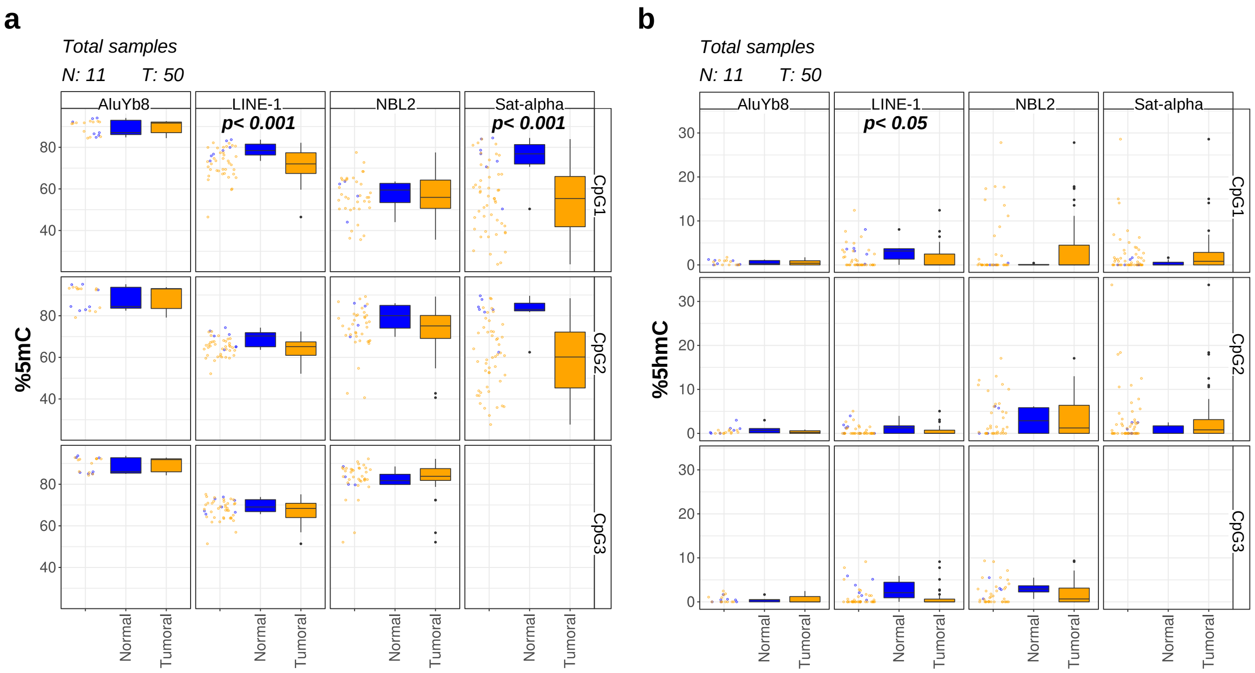
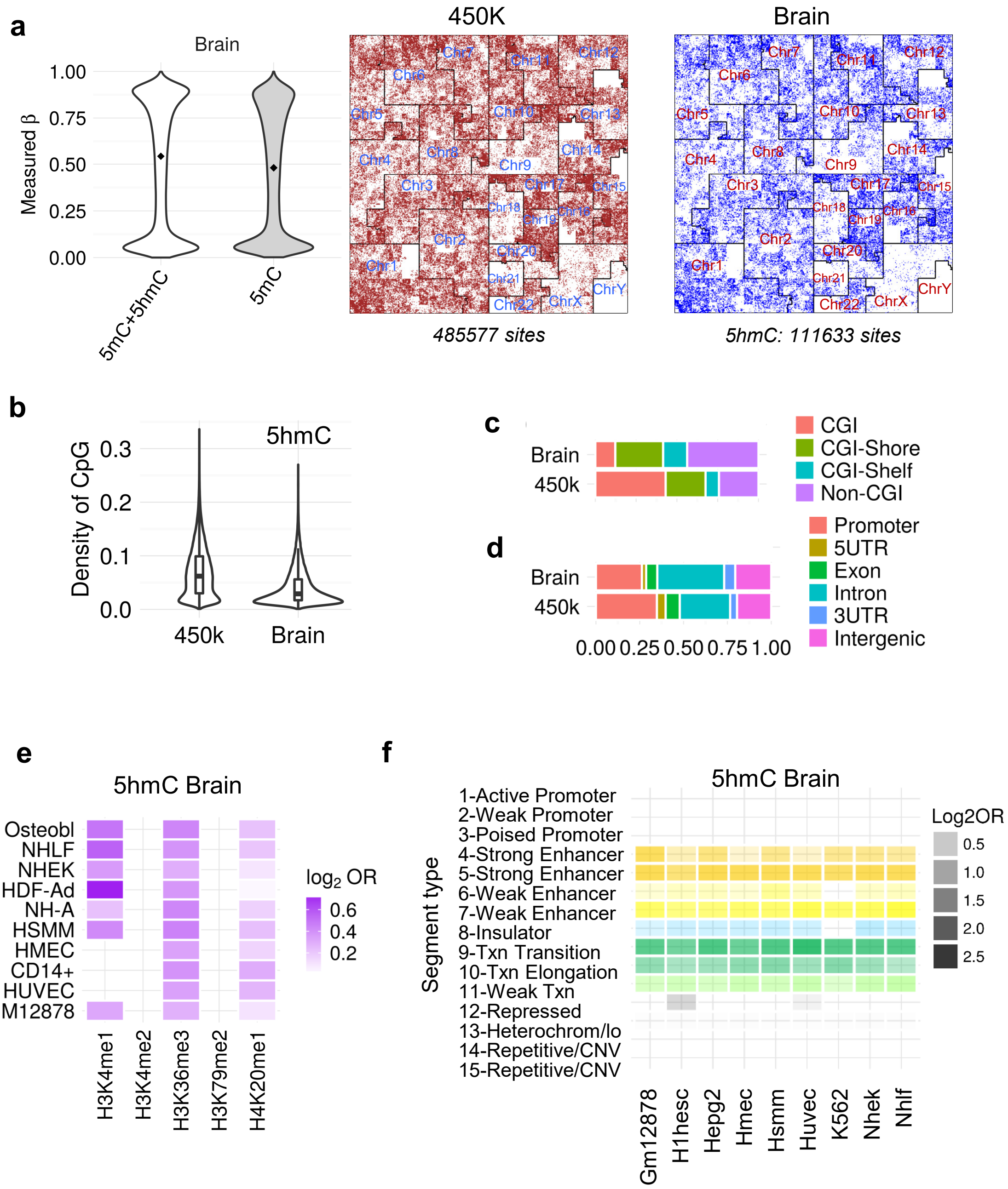
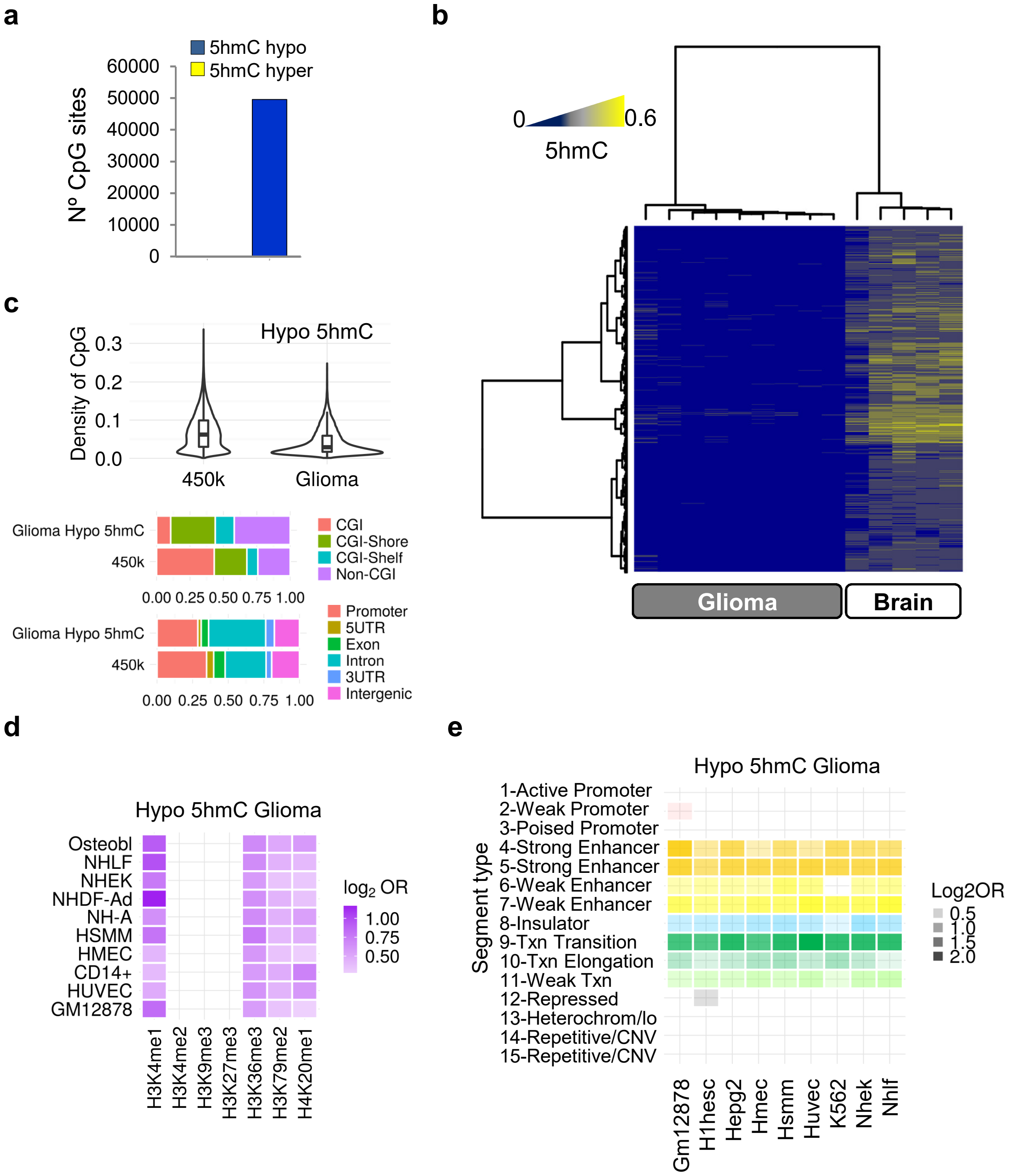

a
Total samples
N: 11 T: 50
AluYb8
LINE-1
NBL2
Sat-alpha
p< 0.001
p< 0.001
%5mC
80
60
40
CpG1
CpG2
CpG3
Normal
Tumoral
b
Total samples
N: 11 T: 50
AluYb8
LINE-1
NBL2
Sat-alpha
p< 0.05
%5hmC
30
20
10
0
CpG1
CpG2
CpG3
Normal
Tumoral

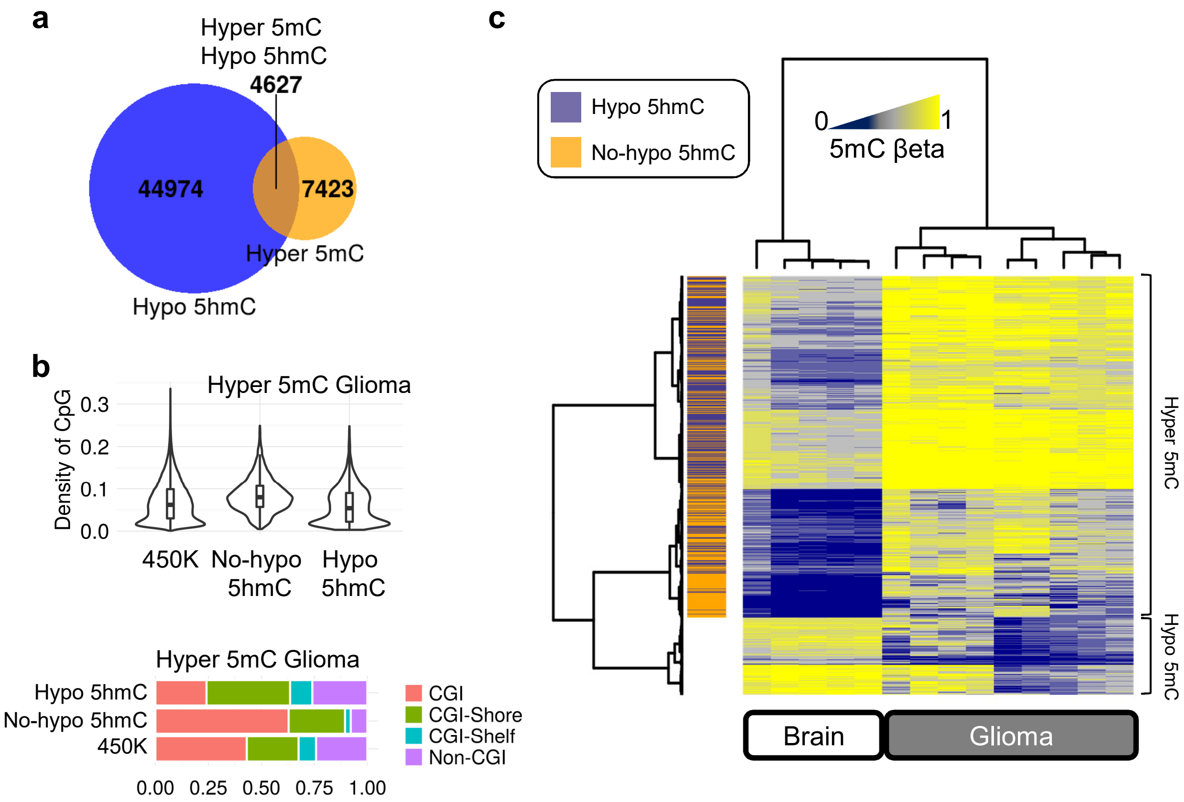
Figure 2

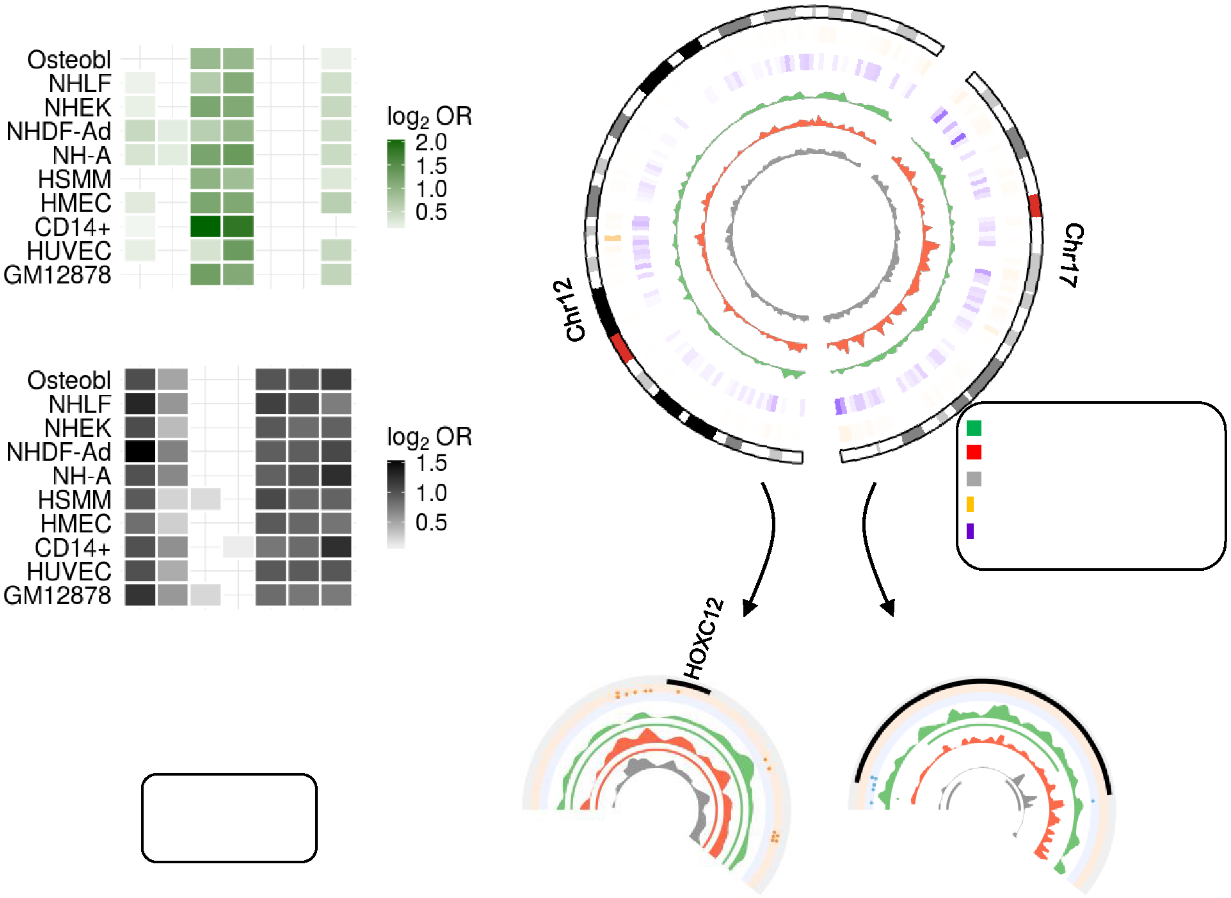


a
Brain
Measured β
1.00
0.75
0.50
0.25
0.00
5mC+5hmC
5mC
450K
485577 sites
Brain
5hmC: 111633 sites
b
5hmC
Density of CpG
0.3
0.2
0.1
0.0
450k
Brain
c
Brain
450k
CGI
CGI-Shore
CGI-Shelf
Non-CGI
d
Brain
450k
0.00
0.25
0.50
0.75
1.00
Promoter
5UTR
Exon
Intron
3UTR
Intergenic
e
5hmC Brain
Osteobl
NHLF
NHEK
NHDF-Ad
NH-A
HSMM
HMEC
CD14+
HUVEC
GM12878
H3K4me1
H3K4me2
H3K36me3
H3K79me2
H4K20me1
log2 OR
0.6
0.4
0.2
f
5hmC Brain
Segment type
1-Active Promoter
2-Weak Promoter
3-Poised Promoter
4-Strong Enhancer
5-Strong Enhancer
6-Weak Enhancer
7-Weak Enhancer
8-Insulator
9-Txn Transition
10-Txn Elongation
11-Weak Txn
12-Repressed
13-Heterochrom/lo
14-Repetitive/CNV
15-Repetitive/CNV
Gm12878
H1hesc
Hepg2
Hmec
Hsmm
Huvec
K562
Nhek
Nhlf
Log2OR
0.5
1.0
1.5
2.0
2.5

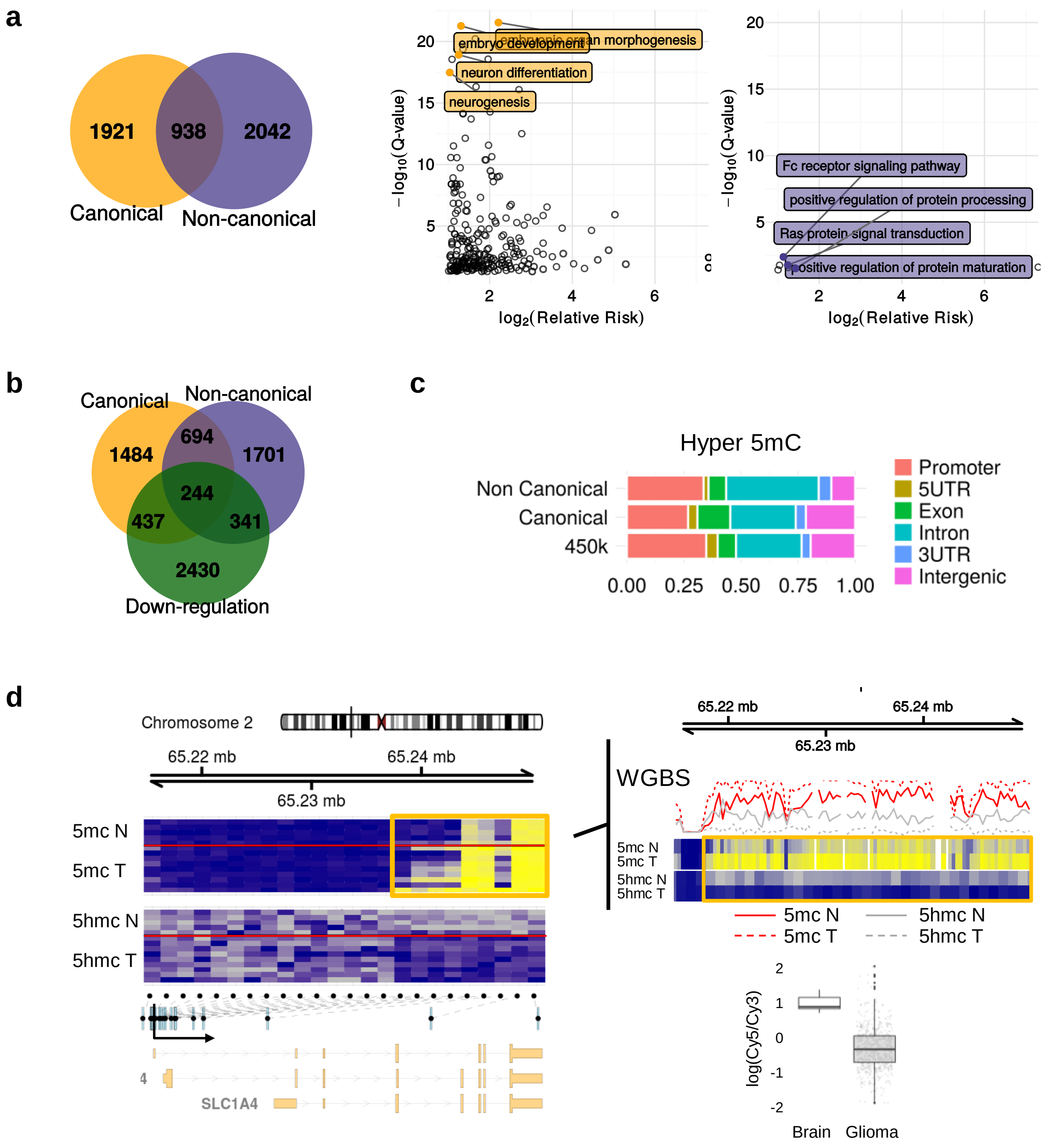
Figure 3

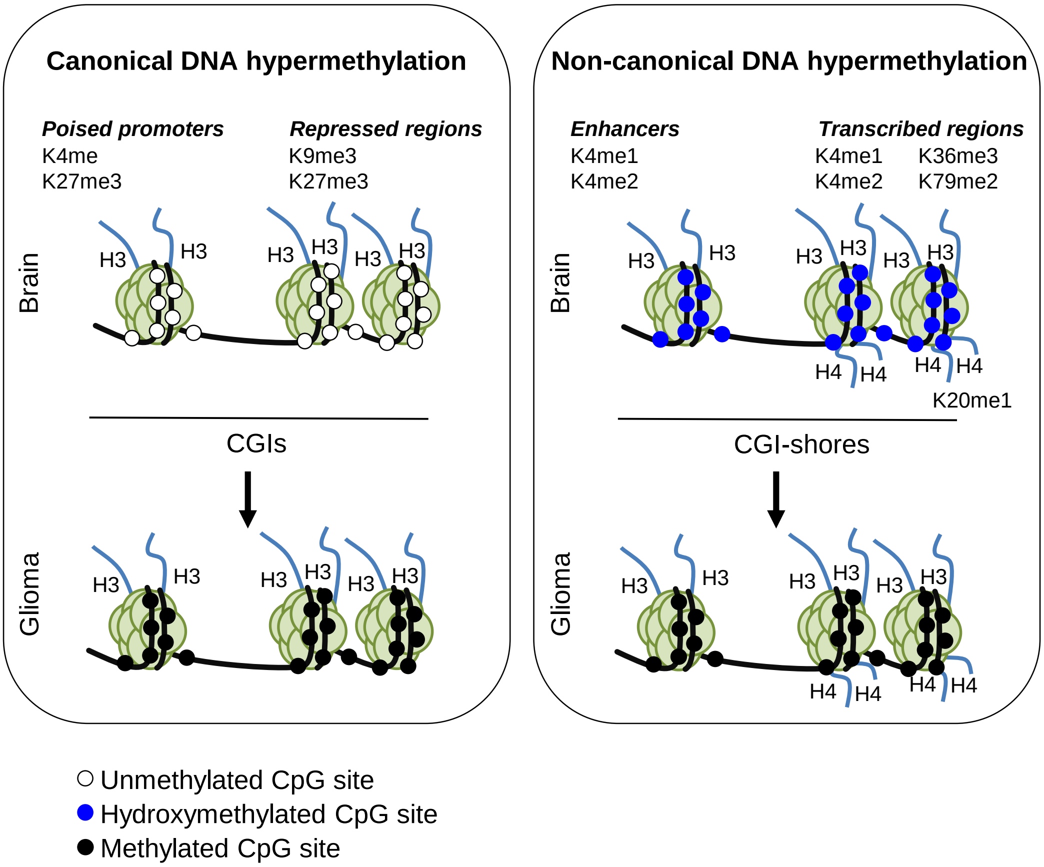

a
5hmC hypo
5hmC hyper
Nº CpG sites
60000
50000
40000
30000
20000
10000
0
b
0
0.6
5hmC
Glioma
Brain
c
Hypo 5hmC
Density of CpG
0.3
0.2
0.1
0.0
450k
Glioma
Glioma Hypo 5hmC
450k
0.00 0.25 0.50 0.75 1.00
CGI
CGI-Shore
CGI-Shelf
Non-CGI
Promoter
5UTR
Exon
Intron
3UTR
Intergenic
d
Hypo 5hmC Glioma
Osteobl
NHLF
NHEK
NHDF-Ad
NH-A
HSMM
HMEC
CD14+
HUVEC
GM12878
H3K4me1
H3K4me2
H3K9me3
H3K27me3
H3K36me3
H3K79me2
H4K20me1
log2 OR
1.00
0.75
0.50
e
Hypo 5hmC Glioma
Segment type
1-Active Promoter
2-Weak Promoter
3-Poised Promoter
4-Strong Enhancer
5-Strong Enhancer
6-Weak Enhancer
7-Weak Enhancer
8-Insulator
9-Txn Transition
10-Txn Elongation
11-Weak Txn
12-Repressed
13-Heterochrom/lo
14-Repetitive/CNV
15-Repetitive/CNV
Gm12878
H1hesc
Hepg2
Hmec
Hsmm
Huvec
K562
Nhek
Nhlf
Log2OR
0.5
1.0
1.5
2.0


a
Hyper 5mC
Hypo 5hmC
4627
44974
7423
Hyper 5mC
Hypo 5hmC
b
Hyper 5mC Glioma
Density of CpG
0.3
0.2
0.1
0.0
450K
No-hypo 5hmC
Hypo 5hmC
Hyper 5mC Glioma
Hypo 5hmC
No-hypo 5hmC
450K
0.00
0.25
0.50
0.75
1.00
CGI
CGI-Shore
CGI-Shelf
Non-CGI
c
Hypo 5hmC
No-hypo 5hmC
0
1
5mC βeta
Hyper 5mC
Hypo 5mC
Brain
Glioma



Osteobl
NHLF
NHEK
NHDF-Ad
NH-A
HSMM
HMEC
CD14+
HUVEC
GM12878
log$_2$ OR
2.0
1.5
1.0
0.5
Osteobl
NHLF
NHEK
NHDF-Ad
NH-A
HSMM
HMEC
CD14+
HUVEC
GM12878
log$_2$ OR
1.5
1.0
0.5
Chr12
Chr17
HOXC12


a
1921
938
2042
Canonical
Non-canonical
20
15
10
5
−log10(Q-value)
embryonic organ morphogenesis
embryo development
neuron differentiation
neurogenesis
2
4
6
log2(Relative Risk)
Fc receptor signaling pathway
positive regulation of protein processing
Ras protein signal transduction
positive regulation of protein maturation
b
Canonical
Non-canonical
1484
694
1701
244
437
341
2430
Down-regulation
c
Hyper 5mC
Non Canonical
Canonical
450k
0.00
0.25
0.50
0.75
1.00
Promoter
5UTR
Exon
Intron
3UTR
Intergenic
d
Chromosome 2
65.22 mb
65.24 mb
65.23 mb
5mc N
5mc T
5hmc N
5hmc T
SLC1A4
WGBS
5mc N
5mc T
5hmc N
5hmc T
2
1
0
-1
-2
log(Cy5/Cy3)
Brain
Glioma


Canonical DNA hypermethylation
Poised promoters
K4me
K27me3
Repressed regions
K9me3
K27me3
Brain
H3
H3
H3
H3
H3
H3
CGIs
Glioma
H3
H3
H3
H3
H3
H3
Non-canonical DNA hypermethylation
Enhancers
K4me1
K4me2
Transcribed regions
K4me1
K4me2
K36me3
K79me2
Brain
H3
H3
H3
H3
H3
H3
H4
H4
H4
H4
K20me1
CGI-shores
Glioma
H3
H3
H3
H3
H3
H3
H4
H4
H4
H4
○ Unmethylated CpG site
● Hydroxymethylated CpG site
● Methylated CpG site