5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA

Marketa Tomkova[a], Michael McClellan[a], Skirmantas Kriaucionis[a1] and Benjamin Schuster-Böckler[a1]

AFFILIATION

a. Ludwig Cancer Research Oxford
University of Oxford
Old Road Campus Research Building
Oxford OX3 7DQ
United Kingdom

KEYWORDS

Somatic mutation;DNA methylation; Cancer genomics;Hydroxymethylcytosine

CORRESPONDING AUTHORS

1. Co-corresponding authors

- Benjamin Schuster-Böckler
benjamin.schuster-boeckler@ludwig.ox.ac.uk
- Skirmantas Kriaucionis
skirmantas.kriaucionis@ludwig.ox.ac.uk

Ludwig Cancer Research Oxford
University of Oxford
Old Road Campus Research Building
Oxford OX3 7DQ
United Kingdom

ABSTRACT

CpG dinucleotides are the main mutational hot-spot in most cancers. The characteristic elevated C>T mutation rate in CpG sites has been related to 5-methylcytosine (5mC), an epigenetically modified base which resides in CpGs and plays a role in transcription silencing. In brain nearly a third of 5mCs have recently been found to exist in the form of 5-hydroxymethylcytosine (5hmC), yet the effect of 5hmC on mutational processes is still poorly understood. Here we show that 5hmC is associated with an up to 53% decrease in the frequency of C>T mutations in a CpG context compared to 5mC. Tissue specific 5hmC patterns in brain, kidney and blood correlate with lower regional CpG>T mutation frequency in cancers originating in the respective tissues. Together our data reveal global and opposing effects of the two most common cytosine modifications on the frequency of cancer causing somatic mutations in different cell types.

BACKGROUND

Cancer genomics projects have revealed that the distribution of somatic mutations across the genome is not uniform (Lawrence et al., 2013). Apart from positive and negative selective pressure, a number of factors can influence regional mutation frequencies, such as chromatin organisation (Schuster-Böckler & Lehner, 2012), replication timing (Koren et al., 2012), metabolic load (Ames et al., 1993) and exposure to different mutagens (Poon et al., 2013). Furthermore, highly transcribed regions generally exhibit lower mutation frequencies due to the influence of transcription-coupled repair (Lawrence et al., 2013). In addition to the regional distribution of mutations, the local nucleotide contexts and mutation types (referred to as *mutational signatures*) have been investigated extensively since they provide critical clues about the mechanism of mutagenesis. For example, consensus motifs for cytidine deaminases (such as APOBEC and AID) were found enriched at mutational hot-spots, suggesting that activity of these enzymes could be the potential cause of those mutations (Nik-Zainal et al., 2012; Taylor et al., 2013). The most frequent mutational signature found in the majority of cancers is C to T transition in a CpG dinucleotide context (CpG>T)

(Alexandrov et al., 2013; Lawrence et al., 2013). This relates to the fact that cytosines in CpG dinucleotides are frequently methylated to form 5-methylcytosine (5mC). The rate of spontaneous deamination of 5mC into T is four fold higher than the rate of deamination of C into U (Lindahl & Nyberg, 1974). In the germline of vertebrates, this likely facilitated a general depletion of CpGs outside of CpG islands which are largely unmethylated.

The genomes of all examined vertebrate species feature DNA methylation, and loss of DNA methylation is incompatible with normal development in mice (E. Li et al., 1992; Okano et al., 1999). DNA methylation plays a role in gene expression, most notably by repressing one allele of imprinted genes. Moreover, it is involved in maintenance of genome stability, alternative splicing, X chromosome inactivation and suppression of retrotransposons (Klose & Bird, 2006; Jones, 2012). In 2009, 5-hydroxymethylcytosine (5hmC) was indisputably shown to exist in DNA of brain and other tissues (Kriaucionis & Heintz, 2009). It was concurrently shown that ten-eleven translocation (TET) enzymes are able to convert 5mC into 5hmC (Tahiliani et al., 2009). Unlike 5mC, which is observed at similar levels in many cell types, the abundance of 5hmC varies widely. 5hmC was observed to be particularly enriched in brain cells (Kriaucionis & Heintz, 2009; Lister et al., 2013) and detectable in embryonic stem cells and all examined tissues (Tahiliani et al., 2009; Globisch et al., 2010; Szwagierczak et al., 2010; H. Wu & Zhang, 2011). 5hmC and higher oxidised states of methyl-cytosine have been proposed to play a role in de-methylation via ineffective re-methylation after replication or directly by thymine DNA glycosylase (TDG) (Tahiliani et al., 2009; He et al., 2011; Maiti & Drohat, 2011; Shen et al., 2013; Hu et al., 2014). In addition to demethylation, 5hmC has been implicated in transcriptional regulation, and a number of DNA binding proteins recognising 5hmC have been identified (Mellén et al., 2012; Spruijt et al., 2013; Takai et al., 2014). 5hmC is found depleted in primary tumours and TET2 is frequently mutated in myelodysplastic syndrome, acute myelogenous leukaemia and T-cell lymphoma, indicating that 5hmC plays a role in carcinogenesis. However, the

molecular mechanism by which 5hmC affects carcinogenesis is poorly understood (Rasmussen & Helin, 2016).

5hmC is an important intermediate during demethylation in zygotes and ES cells (Tahiliani et al., 2009; Inoue & Zhang, 2011; Wossidlo et al., 2011), but the vast majority of 5hmC is found as a stable, long-lived modification in adult mouse tissue that undergoes little cell division (Bachman et al., 2014; Brazauskas & Kriaucionis, 2014). Thus, we hypothesised that – similar to 5mC – long-lived 5hmC could have a substantial influence on the mutability of DNA. Little is known about the mutational properties of 5hmC, in part because until recently there has been a lack of information on the precise location of 5hmC in the genome. With the development of techniques for single-nucleotide resolution mapping of 5mC and 5hmC (Yu et al., 2012; Booth et al., 2014), it is now possible to differentiate mutation rates at 5mC and 5hmC sites. Recently, Supek et al. (Supek et al., 2014) reported elevated C>G transversion rates at 5hmC sites, using 5hmC maps from human and mouse embryonic stem cells. However, these findings are limited by the fact that embryonic stem cells differ substantially from the somatic tissues in which mutations were observed (Schultz et al., 2015).

A large proportion of mutations observed in any cancer genome originate in its pre-cancerous cell of origin (Nik-Zainal et al., 2012; Stephens et al., 2012; Tomasetti et al., 2013; S. Wu et al., 2015) and will have been influenced by its epigenetic landscape. The publication of single-base resolution maps of 5mC and 5hmC occupancy in samples of human brain, kidney and blood (Wen et al., 2014; Chen et al., 2015; Pacis et al., 2015) now enables us to interrogate the tissue-specific effect of cytosine modifications on somatic mutation rates in unprecedented detail.

Since 5hmC has been shown to be most abundant in human brain (W. Li & Liu, 2011; Nestor et al., 2012), we have initially focussed on assessing the relationship between mutability and DNA modifications in brain cancers. Based on a DNA sequencing data from five brain cancer types

encompassing 665 patients, we show that the dominant mutational signature in brain cancers is CpG>T, which is modulated by the modification state of cytosine. Strikingly, the CpG>T mutation frequency of 5-hydroxymethylcytosine is reduced nearly two-fold compared to the methylated state. We find that the ratio of 5hmC to 5mC in 100kb genomic intervals correlates with CpG>T mutation frequency even after accounting for confounding factors like gene density or CpG islands. When we expand our analysis to include mutations and 5hmC maps from kidney and myeloid lineage of blood cells, we observe a clear tissue-specific effect of 5hmC on mutagenicity. Finally, we measured 5mC and 5hmC levels using methodology of high accuracy in eight different human tissue types and show that reduced 5hmC levels associate with an increased proportion of CpG>T mutations in cancers of the corresponding tissue. Together, our findings suggest that hydroxymethylation has a significant influence on the likelihood of mutations at CpG sites across many human tissue types.

RESULTS

We compiled base-resolution maps of 5mC and 5hmC frequency in brain, kidney and myeloid cells from publicly available sources (Wen et al., 2014; Chen et al., 2015; Pacis et al., 2015). All three data sets are based on bisulfite (BS) and "Tet-assisted bisulfite" (TAB) sequencing, respectively. BS-Seq detects any modified cytosine (i.e., does not distinguish 5mC and 5hmC) whereas TAB-Seq specifically detects 5hmC. The combination of the two methods allows an estimation of the levels of both 5mC and 5hmC for all sufficiently covered cytosines. As 5hmC predominantly occurs in a CpG context, we focussed the analysis on CpG sites. Sequencing reads come from heterogeneous populations of cells. Hence, a single locus usually cannot be assigned a single state (C, 5mC or 5hmC). Instead, we estimated the frequency of modification, hydroxymethylation and methylation per site using the percentage of BS-Seq reads that showed a modification (referred to as *mod level*), the percentage of TAB-Seq reads that showed hydroxymethylation (referred to as *5hmC level*) and their difference (*5mC level = mod level – 5hmC level*), respectively.

5hmC sites in brain exhibit lower frequency of CpG>T mutations than 5mC sites

Since brain tissue has been shown to exhibit particularly high levels of 5hmC (Fig. 1A), we first investigated the relationship between the regional distribution of 5hmC, 5mC and mutagenesis in brain tumours. We reasoned that this approach would provide the highest sensitivity to detect any correlation between 5hmC and mutation frequency.

We analysed 344370 somatic single nucleotide variants (SNVs) from 665 samples derived from exome and whole genome sequencing of the following cancer types: Glioblastoma (GBM), Glioma low grade (GLG), Neuroblastoma (NRB), Medulloblastoma (MDB) and Pilocytic astrocytoma (PA) (Alexandrov et al., 2013). The dominant point mutation type in these cancers were C>T transitions in a CpG context (Fig. 1B, 1C), similar to what was observed in combined analyses of all cancer types (Alexandrov et al., 2013).

Mutations and DNA modifications are not distributed uniformly along the chromosomes. Strikingly, 5hmC levels were visibly and significantly anti-correlated with the frequency of CpG>T (r=-0.71, chr3), while 5mC levels displayed a positive correlation (r=0.66, chr3, Fig. 1D, Fig. 1-figure supplements 1–2). This is not a simple consequence of the uneven distribution of genes, exons, CpG islands or levels of gene expression (Fig. 1-figure supplement 3 and additional analyses below).

Averaging over the entire genome, the frequency of C>T mutations differed substantially between $5mC_{high}$ ($5mC_{high}$: *mod level* > 10% and *5hmC level* / *mod level* <= 0.3) and $5hmC_{high}$ ($5hmC_{high}$: *mod level* > 10% and *5hmC level* / *mod level* >= 0.5) sites. The fraction of mutated $5hmC_{high}$ sites was significantly lower than the fraction of mutated $5mC_{high}$ sites (Fig. 1E). The lower mutation frequency was consistently observed in data derived from both exome and whole genome sequencing projects ($P<0.001$, Wilcoxon signed-rank test). Moreover, all brain cancer types individually displayed a significant (28–53%, $P<0.05$ in all types) reduction of C>T mutations in $5hmC_{high}$ sites (Fig. 2A, 2B).

146 It has been shown that CpG>T mutations are one of the two mutational signatures correlating with
147 age (Alexandrov et al., 2015), supporting the fact that these mutations were gathered during the
148 entire lives of the patients, not only after the origin of cancer. Here we observed that this correlation
149 is present in both methylated and hydroxymethylated sites (Fig. 2C). Moreover, the slope for 5mC
150 was steeper than for 5hmC, suggesting that even the mechanisms causing the difference of CpG>T
151 mutability between 5mC and 5hmC were present in the pre-cancerous cell of origin.

152 We also compared the fraction of mutated $5mC_{high}$ and $5hmC_{high}$ sites for the other two possible
153 types of mutations: C>A and C>G. As shown in Fig. 1E, C>A or C>G transversions are an order of
154 magnitude less frequent than C>T transitions in both 5mC and 5hmC sites. The relationship between
155 C>A and C>G mutations and 5hmC varied between cancer types. In GBM and LGG the frequency of
156 C>A mutations was significantly higher in $5mC_{high}$ compared to $5hmC_{high}$ sites, but in NRB, MDB and
157 PA we detected no significant difference. The frequency of C>G mutations in $5mC_{high}$ sites differed
158 significantly from $5hmC_{high}$ sites only in MDB, PA and GBM. In MDB and PA, $5hmC_{high}$ sites were
159 slightly enriched for C>G mutations, whereas in GBM an enrichment was observed at $5mC_{high}$ sites.
160 Since C>T transitions are the most common somatic mutation type in brain and the difference in C>T
161 mutations between $5mC_{high}$ and $5hmC_{high}$ sites is more consistent among cancer types than in the
162 C>A and C>G transversions, we focus mainly on C>T mutations in the remainder of this report.

163 We confirmed that C>T mutations are significantly depleted at 5hmC sites across a wide range of
164 thresholds in definitions of $5mC_{high}$ and $5hmC_{high}$ (Fig. 2-figure supplement 1A–F). In fact, more
165 stringent definitions of 5hmC (e.g., $5hmC_{high}$: *5hmC level* / *mod level* >= 0.7) result in even greater
166 differences (42–59%) in C>T mutation frequencies between $5mC_{high}$ and $5hmC_{high}$ sites (Fig. 2-figure
167 supplement 1G–I, Fig. 3-figure supplement 1A–D), but these definitions would reduce the number of
168 sites too much for our subsequent statistical analyses.

Reduced 5hmC mutability in brain is not accounted for by genomic regions or gene expression

We next examined whether the decreased frequency of C>T transitions at 5hmC vs. 5mC sites might be an indirect effect of 5hmC being associated with genomic regions of lower mutability. Levels of 5mC and 5hmC vary across genomic regions. For example, 5hmC density is elevated in highly expressed genes in brain (Mellén et al., 2012; Yu et al., 2012; Lister et al., 2013; Wen et al., 2014). The observed decrease in C>T mutation frequencies might therefore be attributable to higher gene expression, which would correlate with higher transcription coupled repair. We therefore performed the analysis described above separately for regions with high vs. low gene expression in human brain (see Methods). There was a lower overall mutation frequency in highly expressed genes (Fig. 3A–B), but both highly and lowly expressed genes exhibited significantly lower C>T transition rates at 5hmC sites compared to 5mC sites (Fig. 3A–D). This suggests that the observed effect is independent of transcription and thus not a result of transcription coupled repair.

Gene expression is only one of many possible region-related confounding factors. Hence, to correct for any regional variation, we performed the analysis on groups of sites generated by pairing the modified CpGs: each 5hmC site was paired with the nearest yet unpaired 5mC site from an equivalent genomic and sequence context (an approach adapted from (Supek et al., 2014), see Methods). Thereby we compared the mutation frequencies of two groups (one group comprising 5mC sites and one group comprising 5hmC sites) containing the same number of loci (6801374 cytosines in each group). As a result of this experimental setup, a substantial fraction of mutated 5mC sites were excluded, greatly reducing the statistical power of this “paired” analysis. Nevertheless, the frequency of C>T mutations in 5hmC remained significantly lower than in 5mC in both exomes and genomes (Fig. 3E–F), supporting that the difference between 5mC and 5hmC mutation frequency is not caused by regional differences.

To ensure that there is no confounding bias in the spatial distribution of mutations around 5mC or 5hmC sites, respectively, we plotted mutation frequencies in a 2kb radius up and downstream of

modified loci (Fig. 3-figure supplement 1G, Methods). The mutation frequency differed substantially in the aligned positions of DNA modifications but was indistinguishable in the surrounding area. In conclusion, regional mutation rate variability is unlikely to account for the difference in C>T mutational load in 5mC and 5hmC sites.

Relative 5hmC correlates with CpG>T mutation frequency

The 5mC and 5hmC frequency at each base reflect the prevalence of each modification within the sequenced cell population. We hypothesised that if 5hmC had a direct effect on C>T mutation likelihood, we would observe an increase in mutation frequency with decreasing 5hmC occupancy. To test this, we formally defined $5hmC_{rel}$ as the ratio of the proportion of reads supporting 5hmC, relative to the proportion of reads supporting any modification at a particular cytosine ($5hmC_{rel}$ = *5hmC level / mod level*). We then divided brain CpG sites into bins according to their $5hmC_{rel}$ level and computed the fraction of mutated sites in each bin (Fig. 4A). We observed a clear linear relationship between $5hmC_{rel}$ values and C>T mutation frequencies. Notably, the correlation was present in all the tested brain cancer types in exome- and whole genome-sequenced samples. A regression slope test confirmed significance of this relationship in all the cancer types. To confirm that the results are not influenced by an uneven distribution of information across bins, we performed quantile binning so that each bin contains an approximately equal number of positions (see Methods). The results of quantile bins were equivalent to evenly spaced bins (Fig. 4-figure supplement 1H).

For comparison, we also evaluated the relationship between $5hmC_{rel}$ and the frequency of C>A and C>G mutations (Fig. 4A). Consistent with our previous results, an increase in $5hmC_{rel}$ is associated with an increase in C>G mutations in whole genomes (from MDB and PA samples), but the relationship in other cancer types shows no significant trend. C>A mutations decrease with $5hmC_{rel}$ levels in GBM but exhibit no significant signal in the remaining tumour types.

This result supports the conclusion that the decrease in C>T mutation frequency at 5hmC sites is not an artefact of our chosen definition of 5mC or 5hmC. Even more importantly, it supports the notion that this decrease is directly caused by the properties of these DNA modifications.

Mutation load of 5hmC sites is similar to unmodified cytosines

The findings reported so far could be attributed to an elevated mutation rate in 5mC, to a lowered mutagenicity of 5hmC or a combination of the two. To investigate this question, we compared mutation frequencies at 5mC and 5hmC sites to that of unmodified cytosines. We divided all the sequenced CpG sites into 9x9 bins according to their levels of 5mC and 5hmC. We observed that the mutation frequency of unmodified cytosine is similar to 5hmC, whereas 5mC exhibited much higher mutation frequency (Fig. 4B). Further, we calculated the mutation frequency distribution in sites that exhibited almost no methylation or almost no hydroxymethylation, respectively. When methylated sites are excluded, the mutation frequency does not show any significant trend with increasing levels of 5hmC (Fig. 4C). Conversely, excluding hydroxymethylated sites leads to a significant gradient in mutation frequency with increasing levels of 5mC (Fig. 4D). When only fully modified sites (*mod level* >= 90%) are taken into account, increasing levels of 5hmC (i.e., decreasing levels of 5mC) are associated with a significant decrease in C>T mutation frequency (Fig. 4E).

5hmC is a predictor of CpG>T mutation frequency across the genome

To examine the exclusive impact of DNA modifications on regional frequencies of mutations, we split the genome into 100kb windows and fitted a generalised linear model to explain the observed per-window CpG>T mutation frequency from a combination of features including average 5mC and 5hmC levels, $5hmC_{rel}$, gene density, CpG island density amongst others. Only whole genome sequencing data were used for this analysis. To compare the resulting models, we calculated their respective “explained deviance” D^2, a generalisation of explained variance that is more appropriate for comparing generalised linear models (see Methods).

242 The best individual predictor of CpG>T mutation frequency was $5hmC_{rel}$ ($D^2 = 0.11$), outperforming all
243 other features (Fig. 5A). Interestingly, the sum of 5mC and 5hmC levels ("mod") performed worst,
244 suggesting that bisulfite sequencing measurements alone are a poor predictor of mutagenicity.
245 When combining all 11 features into one model, the total explained deviance for 100kb windows
246 was 16%.

247 Varying the chosen window size (3kbp–3Mbp; Fig. 5B, Fig. 5-figure supplement 1A–C) does not
248 substantially change the comparison of the predictive power of the respective features. In all cases,
249 5mC and $5hmC_{rel}$ were the two best predictors, with $5hmC_{rel}$ performing slightly better with smaller
250 windows. However, the total explained deviance increased with window size, reaching values as high
251 as 45% for univariate models and 60% for models with all predictors. This led us to question whether
252 the increasing predictive power of larger windows has a biological reason, or whether it is a
253 consequence of the lower data density in small windows.

254 Since many smaller windows contain no observed mutations, low D^2 values could simply reflect a
255 lack of data. To test this, we generated simulated mutations so that a "perfect" predictor was
256 linearly related to the mutation likelihood per window (see Methods). We then measured the effect
257 of window and sample size (number of patients) on the observed D^2, repeating the simulations 10
258 times. The resulting curves of the explained deviance resemble those measured in the real data (Fig.
259 5-figure supplement 1D). Moreover, in the simulated data, higher numbers of patients lead to higher
260 D^2 even for smaller window sizes, suggesting that lower D^2 values in smaller windows indeed are a
261 consequence of lower data density.

262 **Level of genic 5hmC correlates with decrease of CpG>T**

263 It has been reported that 5hmC is enriched in gene bodies, and several brain cancer sequencing data
264 sets. We therefore tested whether the relationship between 5hmC and mutations, which we
265 observed across the whole genome, is also detectable in exonic regions alone.

In line with our earlier results, we found that $5hmC_{rel}$ significantly contributes to the deviance explained by the model, beyond covariation with gene expression (Fig. 5C–D; F-test $p < 2e\text{-}100$). We hypothesised that this effect should be most pronounced when using modC>T and CpG>T as the response variable, whereas it should decrease when using definitions of mutations that include a progressively wider range of loci (C>T, C>N, N>N). Indeed, the unique contribution of $5hmC_{rel}$ to the explained gene mutation frequency decreased as the mutation sets became larger and more distant from modC>T (Fig. 5C–D, Fig. 5-figure supplements 2–3). Nevertheless, in all of the cases, $5hmC_{rel}$ significantly improved the fit of the model. Conversely, we confirmed that $5hmC_{rel}$ had no significant predictive power for the frequency of T>N mutations (Fig. 5C–D; column T>N), supporting the hypothesis that $5hmC_{rel}$ selectively affects mutations in modified cytosines.

Decreased CpG>T mutation frequency in 5hmC is not limited to brain tissue

Two recently published datasets allowed us to address the question of mutational properties of 5mC and 5hmC also in two other tissues: kidney (Chen et al., 2015) and blood (Pacis et al., 2015). For blood we used 174 sequencing samples from Acute Myeloid Leukaemia (AML) as the cancer type closest to the blood dendritic cells in which the BS-Seq and TAB-Seq measurements were performed. For kidney we combined 585 samples from four distinct sequencing projects, covering Kidney Clear Cell, Kidney Papillary and Kidney Chromophobe carcinomas.

Matching our findings in brain, 5hmC sites were mutated significantly less frequently than 5mC sites in both tissue types (Fig. 6B), irrespective of whether genome or exome sequencing data were used. Moreover, a similar difference was present in all available replicates of the BS-Seq and TAB-Seq measurements (6 for blood, 2 for kidney, Fig. 6-figure supplement 1A).

Genomic distribution of 5hmC differs substantially between the three tissue types (Fig. 6-figure supplement 2). Consequently, we expected the association between mutation frequency and 5hmC to be highest when mutation and modification data are derived from matching tissue types. To test

290 this hypothesis, we used a GLM on genomic windows of 100kbp to predict CpG>T mutation rate
291 from a combination of $5hmC_{rel}$ maps of all three tissues. Strikingly, for each cancer type, the best
292 predictor came from the same tissue type (Fig. 6A), suggesting that tissue differences in 5hmC are
293 reflected in the CpG>T mutation landscape. The same results were obtained in all available
294 replicates of the $5hmC_{rel}$ maps (Fig. 6-figure supplement 1B). Finally, we added a $5hmC_{rel}$ map
295 derived from embryonic stem cells (ESC) as an additional predictor, to compare our findings to
296 previously reported results (Supek et al., 2014). As we anticipated, the ESC-derived 5hmC levels have
297 substantially lower predictive power on CpG>T mutation rate than any of the tissue-derived maps.

298 While base-resolution maps of 5hmC for human tissue are still rare, there is a wide range of BS-Seq
299 data sets available in public databases. Given our findings thus far, we predicted that tissues with
300 high levels of 5hmC relative to 5mC would exhibit fewer CpG>T mutations in modified sites than
301 tissues with low total 5hmC. To test this hypothesis, we measured total levels of 5mC and 5hmC
302 using High Pressure Liquid Chromatography (HPLC-UV) in DNA of eight human tissue types for which
303 BS-Seq maps are publicly available (Fig. 6-figure supplement 3). In order to account for unrelated
304 cancer-type specific differences in CpG>T mutability, we normalised the mutation frequency in
305 modified sites by the mutation frequency in unmodified sites. The analysis of association between
306 genomic 5hmC and enrichment of CpG>T mutations revealed a strong negative correlation (Fig. 6C)
307 in all tissue types except lung. Interestingly, this difference seems to stem from smoking-related
308 effects. Lung cancer mutation data from heavy smokers revealed a markedly lower frequency of
309 CpG>T mutations in modified sites, relative to other mutations. It has been reported that the typical
310 C>A mutational signature associated with smoking was found significantly enriched in CpGs outside
311 CpG islands, suggesting that it preferentially occurs at modified CpG sites (Pleasance et al., 2010).
312 Accordingly, our data indicate that CpG>T mutations might also be differentially affected by
313 smoking-related mutagens.

DISCUSSION

Here we have established a link between the landscape of DNA modifications and the mutational profile of somatic human cells. Our measurements indicate that 5hmCs carry between 28 and 53% fewer mutations than methylated cytosines in brain. This results in a mutational load at 5hmC sites that is comparable to that of unmodified cytosines in CpG dinucleotides. This effect is not only observable in brain, but also in kidney cancers and myeloid leukaemias. The relationship between 5hmC and CpG>T mutation rate can be detected at the scale of the exome as well as genome-wide and is independent of other region-specific influences on mutation frequency. We show that the relative impact of hydroxymethylation on mutagenesis decreases proportionally to the level of 5hmC in the tissue, suggesting that it represents a general property of this DNA modification.

Two possible scenarios could explain the striking difference in mutability between 5mC and 5hmC. Firstly, spontaneous and enzymatic deamination reactions of 5hmC could be less favourable than 5mC. As a consequence, fewer new mutation events would be expected at 5hmC sites. Indeed, cytosine deaminases (namely, AID and APOBEC1-3) have 4.4–38x lower activity on sites with 5hmC compared to 5mC, supporting this possibility (Nabel et al., 2012; Rangam et al., 2012). Secondly, deamination of 5mC produces thymine whereas 5hmC deaminates to 5-hydroxymethyluracil (5hmU). This atypical base in DNA could be more efficiently recognised and replaced by the DNA glycosylases initiating base-excision repair (BER). Determining the relative contribution of DNA glycosylases to the lower mutation rate would be challenging, since some of these enzymes recognise several types of mismatches. TDG and MBD4 excise both T and 5hmU when mis-paired with G (Hardeland et al., 2003; Cortellino et al., 2011; Guo et al., 2011; Hashimoto et al., 2012; Moréra et al., 2012), whereas SMUG1 does not repair T:G but has a robust activity for 5hmU:G (Nilsen et al., 2001; Kemmerich et al., 2012). Therefore, there might be more efficient repair of 5hmU in the genome. Further genome sequencing efforts might identify patients with rare

inactivating mutations in BER and/or mismatch-repair pathways that could be valuable for future investigations of the relationship between DNA repair and cytosine mutability.

It has previously been suggested that 5hmC levels increase the frequency of C>G mutations (Supek et al., 2014). As part of this analysis, only a very small (albeit statistically significant) decrease of C>T mutations in 5hmC sites in both SNPs and cancer SNVs was observed. There are two factors that could explain why we observe very different effects sizes for C>T and C>G mutations in 5hmC sites. Firstly, Supek et al. consider all sites with as little as one 5hmC read to be hydroxymethylated, whereas we require the level of 5hmC to exceed 5mC. In fact, when examining the effect of variation in these thresholds (Fig. 2-figure supplement 1A–F), we noticed that the results for C>G fluctuate substantially across the range of tested cut-off values (see also Fig. 3-figure supplement 1E–F). Secondly, we present evidence that tissue-specific changes in 5hmC patterns have great influence on the extent of correlation between 5hmC and mutability (Fig. 6B). Specifically, 5hmC genomic localisation in embryonic stem cells was a poor predictor of CpG>T mutations in brain, kidney and blood, compared to the respective tissue-specific 5hmC patterns.

The best predictor of CpG>T mutations in any of the three tested tissues was the $5hmC_{rel}$ map from the corresponding anatomical site. This provides evidence that the slow accumulation of CpG>T mutations in the pre-cancerous tissue was strongly influenced by the DNA modification landscape. However, any bulk tissue sample encompasses a mixture of different cell types. Mounting evidence suggests that solid tumours originate from a defined subset of cells within any one tissue type. For example, glioblastomas were proposed to originate from stem or progenitor cell types enriched in the subventricular zone, while medulloblastomas have mixed cells of origin (Visvader, 2011). Those cell types are of low abundance in normal tissue biopsies. The fact that we observe a clear inverse relationship between CpG>T mutations and the location of 5hmC in multiple tissue types suggests that the DNA modification landscape in cancer-progenitor cells is sufficiently similar to the tissue average to be informative about the mutation frequencies in cancer.

Under this assumption we predict that the impact of DNA modifications on the frequency of CpG>T mutations is likely to be bigger than measured here, since the terminally differentiated cells that make up the bulk of the tissue may have diverged further from cancer-progenitors cells. Advancements in the identification of cancer origins and isolation of single cells, combined with single-cell bisulfite sequencing, will enable an improved assessment of the impact of DNA modifications on mutability.

The strong correlation between relative 5hmC levels in a tissue and the mutability of modified cytosine also points towards a shared underlying mutagenic process. The notable deviation of smoking-induced lung-cancers supports this hypothesis. We speculate that a yet undefined smoking-induced mutagenic mechanism preferentially affects unmethylated CpG sites. More experimental work will be needed to elucidate the biochemical causes for this phenomenon. In the future, the linear relationship between 5hmC levels and CpG>T mutation rate could thus be used to identify other environmental mutagens with a differential effect on modified cytosines.

METHODS

Code. Most of the analyses were performed using Matlab. Code and other required files are available on Figshare under doi 10.6084/m9.figshare.c.3249394 (Tomkova et al., 2016).

Mutation data. Cancer somatic mutations (see Table 2) were obtained from a dataset compiled by Alexandrov et al. (Alexandrov et al., 2013), complemented with whole genome samples from ICGC, (Wang et al., 2014), and TCGA. Briefly, aligned reads for 49 AML tumour and normal samples were downloaded from the UCSC CGHub website under TCGA access request #10140. Somatic variants were called using Strelka (Saunders et al., 2012) with default parameters. All variants were classified by the pyrimidine of the mutated Watson-Crick base pair (C or T) and the immediate 5' and 3' sequence context into 96 possible mutation types as described by Alexandrov et al. (2013).

Modification data. DNA modification information (see Table 1) for brain was extracted from supplementary information provided by Wen et al. (2014). Only sites with more than 5 TAB-Seq reads were taken into account. $5hmC_{high}$ and $5mC_{high}$ sites were defined based on values of *mod level* (unconverted/total BS-Seq reads) and *5hmC level* (unconverted/total TAB-Seq reads) per site:

- $5mC_{high}$: *mod level* > 10% and *5hmC level* / *mod level* ≤ $threshold_{5mC}$
- $5hmC_{high}$: *mod level* > 10% and *5hmC level* / *mod level* ≥= $threshold_{5hmC}$

Effects of the choice of both thresholds were explored and then the values of $threshold_{5mC}$ = 0.3 and $threshold_{5mC}$ = 0.5 were used. In blood, BS-Seq and TAB-Seq values in CpG sites were taken from supplementary files provided by (Pacis et al., 2015). For kidney and ESC maps, raw reads were processed with Trim galore, Bismark (Krueger et al., 2012) and Mark duplicates from Picard tools. Multiple replicates were processed both independently and together (adding the reads from the replicates together). Only sites with at least 10% *mod level* were taken into account to compute $5hmC_{rel}$.

To compute the number of modified sites inside the exome, the reference Illumina Truseq definition of exon loci was downloaded from the Illumina website. Overlapping exons were merged using bedtools so that each genomic site is covered by at most one exon. Two-sided paired Wilcoxon signed-rank test was used for testing significance between mutation frequency of $5mC_{high}$ and $5hmC_{high}$ sites. The same test was used for all the following statistics, if not stated otherwise.

Gene expression data. Gene expression (in FPKM) from RNAseq experiments on 630 brain tissue samples were downloaded from the GTEx human tissue expression project (http://www.gtexportal.org/home/).

Visualisation on genome. The following genomic features were computed in 100kbp windows: average 5hmC, 5mC, 5hmCrel (all from the supplementary information provided by Wen et al. (2014), i.e. mod ≥ 10%), average log(1+ gene expression), gene density, exon density, CpG density, modCpG density, CpG island (CGI) density, and average modification level (from raw BS-Seq reads). These features and CpG>T mutation frequency (from MDB and PA whole-genome sequencing datasets) were z-score normalised and plotted per chromosome after Gaussian smoothing with parameters n= 50, sigma = 2.5.

Mutation frequency in highly and lowly expressed genes. Genes were sorted according to their median expression values. The upper 50-percentile (9701 genes) were classified as highly expressed, the rest as lowly expressed. Introns were included only for whole genome samples.

Pairing of 5mC and 5hmC sites. For each $5hmC_{high}$ site in random order, the nearest not previously selected $5mC_{high}$ site was selected such that the 5mC-5hmC pair fulfilled the following conditions: both $5hmC_{high}$ and $5mC_{high}$ sites are inside an exon or both are outside exons, and both share the same context (CG, CHG, and CHH, where H is T, A or C). This resulted in 6801374 pairs with a median distance of 1 and 25^{th} and 75^{th} quantiles of -177 and +177, respectively.

Mutation frequency around aligned 5mC and 5hmC. Modified sites with no other modifications in a 2kb radius were selected (374 000 sites with 5mC and the same number of 5hmC sites), and the mutation frequency up to 2kbp upstream and downstream (in bins without other modifications) was plotted.

Gradients analysis. All modified cytosines (i.e., *mod level* > 10%) in the CpG context were divided into 9 right-open intervals according to their ratio of *5hmC level* to *mod level*. The leftmost bin contained cytosines where the major modification is 5mC, while the rightmost bin contained cytosines where the major modification is 5hmC. In each bin, the frequency of mutations was computed and plotted. A linear regression model was fitted to the data (function `fitlm` in MatLab) and the significance of the linear coefficient was tested using F-test for the hypothesis that the regression coefficient is zero (function `coefTest` in MatLab). For gradients with equal binning each interval contained approximately the same number of sites (apart from the first bin, which included all values with $5hmC_{rel}$=0).

Prediction of mutation frequency in genomic windows. CpG>T mutation frequency (response variable) and genomic features (predictors; same as above in *Visualisation on genome*) were computed in genomic windows of sizes 3kbp–3Mbp. Then a generalised linear model (`fitglm`) assuming Poisson distribution of the response variable was fitted with a linear model specification (i.e., intercept + linear term for each predictor) and `DispersionFlag` set to true. Model fits were compared in terms of D^2 and p-value (`model.devianceTest`), as recommended, e.g., in (Guisan & Zimmermann, 2000; Mittlböck & Heinzl, 2004).

Simulation of effects of number of patients on GLM. Each chromosome was split into windows of a given window size. For each window, all CpG sites were counted. A random predictor was generated in each window with a beta distribution (Beta(3,4)). For each patient, a random number of mutations in each window was generated as

$$Binomial(n = windowSize(iWindow), p = \frac{predictor(iWindow)}{coefficient}$$

where:

$$coefficient = \frac{\sum_{iWindow} windowSize(iWindow) * predictor(iWindow)}{174}$$

The coefficient was set so that the expected total number of mutations per patient summed to 174, the observed average number of CpG>T mutations in brain WGS data. The response variable was set as the average CpG>T mutation frequency over all patients. A GLM was fit on the given predictor and response variable and D^2 was measured. The process was repeated 10 times for each combination of window size and number of patients.

Gene-wise prediction of mutation frequency. Mutation frequency was modelled with two predictor variables: average $5hmC_{rel}$ per gene and $\log_e$-transformed gene expression. The following response variables computed in exons of each gene were compared:

- modC>T: number of C>T mutations in modified C sites / number of modified C sites
- CpG>T: number of C>T mutations in CpG sites / number of CpG sites
- C>T: number of C>T mutations / number of C sites
- C>N: number of mutations from C / number of C sites
- N>N: number of mutations / number of sites

- T>N: number of mutations from T / number of T sites

Genes with missing values in at least one of the predictors and genes classified as outliers in at least one response variable were excluded from the analysis. Outliers were classified in the following way: *y >= quantile(y, 0.999) + 2.5 * (quantile(y, 0.999) – quantile(y, 0.001))*. Out of 17,605 genes, 10 were classified and removed as outliers: ASPN, BBOX1, CCL4, ESPN, FOLH1, HLA-DPB1, IDH1, NLRP6, S100P, and TP53. The same GLM model as above was used. To calculate the relative contribution of one predictor variable over the other, two models were fitted with either one or both predictor variables and the difference in D^2 was used.

HPLC measurements of total 5hmC and 5mC in eight tissues. 10ug of genomic DNA (amsbio; D1234003, D1234004, D1234035, D1234086, D1234090, D1234122, D1234142, D1234148, D1234149, D1234152, D1234171, D1234188, D1234200, D1234206, D1234226, D1234227, D1234246, D1234248, D1234260, D1234274, HG-101) was treated with 1U RNase A (Thermo Scientific) , purified by phenol chloroform ethanol precipitation and incubated overnight in hydrolysis solution (45 mM NaCl, 9 mM MgCl2, 9 mM Tris pH 7.9, ≥250 U/ml Benzonase (sigma), 50 mU/ml Phosphodiesterase I, ≥20 U/ml Alkaline phosphatase, 46.8 ng/ml EHNA hydrochloride, 8.64 µM deferoxamine). Protein components were removed by centrifugation through Amicon centrifugal filter unit (3 kDa cut-off, Millipore) before samples were lyophilised and resuspended in buffer A. Nucleosides were resolved with an Agilent UHPLC 1290 instrument fitted with Eclipse Plus C18 RRHD 1.8 µm (2.1 × 150 mm column) and detected and quantified with Agilent 1290 DAD fitted with a Max-Light 60 mm cell. Buffer A was 100 mM ammonium acetate, pH 6.5; buffer B was 40% acetonitrile, and the flow rate 0.4 ml min−1. The gradient was between 1.8–100% of 40% acetonitrile with the following steps: 1–2 min, 100% A; 2–16 min 98.2% A, 1.8% B; 16–18 min 70% A, 30% B; 18–20 min 50% A, 50% B; 20–21.5 min 25% A, 75% B; 21.5–22.5 min 100% B; 22.5-24.5 min 100% A. Relative abundance of 5mC and 5hmC were established by detection of adenosine at 280nm allowing determination of total cytosine by extinction coefficient calculation using standards.

AUTHOR CONTRIBUTIONS

Conceptualization: S.K. and B.S.-B.; Investigation: M.T.; HPLC measurements: M.McM.; Writing – Original Draft: B.S.-B., S.K. and M.T.

ACKNOWLEDGEMENTS

We would like to thank Jakub Tomek, Pijus Brazauskas and David Severson for helpful discussions.

S.K. and B.S.-B. are funded by Ludwig Cancer Research. S.K. received funding from BBSRC grant BB/M001873/1. M.T. is funded by EPSRC (EP/F500394/1) and Bakala Foundation.

REFERENCES

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., & Peter, J. (2015). Clock-like mutational processes in human somatic cells. *Nature Publishing Group*, *47*(12), 1402–1407. doi:10.1038/ng.3441

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. a J. R., Behjati, S., Biankin, A. V, Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–21. doi:10.1038/nature12477

Ames, B. N., Shigenaga, M. K., & Hagen, T. M. (1993). Oxidants, antioxidants, and the degenerative diseases of aging. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(17), 7915–7922. doi:10.1073/pnas.90.17.7915

Bachman, M., Uribe-lewis, S., Yang, X., Williams, M., & Murrell, A. (2014). 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry*, *6*(12), 1049–1055. doi:10.1038/nchem.2064

Booth, M. J., Marsico, G., Bachman, M., Beraldi, D., & Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature Chemistry*, *6*(5), 435–40. doi:10.1038/nchem.1893

Brazauskas, P., & Kriaucionis, S. (2014). DNA modifications: Another stable base in DNA. *Nature Chemistry*, *6*(12), 1031–1033. doi:10.1038/nchem.2115

Chen, K., Zhang, J., Guo, Z., Ma, Q., Xu, Z., Zhou, Y., Xu, Z., Li, Z., Liu, Y., Ye, X., Li, X., Yuan, B., Ke, Y., He, C., Zhou, L., Liu, J., & Ci, W. (2015). Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Research*, 103–118. doi:10.1038/cr.2015.150

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D., Abramowitz, L. K., Bartolomei, M. S., Rambow, F., Bassi, M. R., Bruno, T., Fanciulli, M., Renner, C., … Bellacosa, A. (2011). Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*, *146*(1), 67–79. doi:10.1016/j.cell.2011.06.020

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M., &

Carell, T. (2010). Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS One*, *5*(12), e15367. doi:10.1371/journal.pone.0015367

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*, 147–186. doi:10.1016/S0304-3800(00)00354-9

Guo, J. U., Su, Y., Zhong, C., Ming, G. L., & Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, *145*(3), 423–434. doi:10.1016/j.cell.2011.03.022

Hardeland, U., Bentele, M., Jiricny, J., & Schär, P. (2003). The versatile thymine DNA-glycosylase: A comparative characterization of the human, Drosophila and fission yeast orthologs. *Nucleic Acids Research*, *31*(9), 2261–2271. doi:10.1093/nar/gkg344

Hashimoto, H., Zhang, X., & Cheng, X. (2012). Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: Structural basis and implications for active DNA demethylation. *Nucleic Acids Research*, *40*(17), 8276–8284. doi:10.1093/nar/gks628

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C., & Xu, G.-L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, *333*(6047), 1303–7. doi:10.1126/science.1210944

Hu, X., Zhang, L., Mao, S. Q., Li, Z., Chen, J., Zhang, R. R., Wu, H. P., Gao, J., Guo, F., Liu, W., Xu, G. F., Dai, H. Q., Shi, Y. G., Li, X., Hu, B., Tang, F., Pei, D., & Xu, G. L. (2014). Tet and TDG mediate DNA demethylation essential for mesenchymal-to- epithelial transition in somatic cell reprogramming. *Cell Stem Cell*, *14*(4), 512–522. doi:10.1016/j.stem.2014.01.001

Inoue, A., & Zhang, Y. (2011). Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. *Science*, *334*(6053), 194–194. doi:10.1126/science.1212483

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484–492. doi:10.1038/nrg3230

Kemmerich, K., Dingler, F. a., Rada, C., & Neuberger, M. S. (2012). Germline ablation of SMUG1 DNA glycosylase causes loss of 5-hydroxymethyluracil-and UNG-backup uracil-excision activities and increases cancer predisposition of Ung-/-Msh2-/- mice. *Nucleic Acids Research*, *40*(13), 6016–6025. doi:10.1093/nar/gks259

Klose, R. J., & Bird, A. P. (2006). Genomic DNA methylation: The mark and its mediators. *Trends in Biochemical Sciences*, *31*(2), 89–97. doi:10.1016/j.tibs.2005.12.008

Koren, A., Polak, P., Nemesh, J., Michaelson, J. J., Sebat, J., Sunyaev, S. R., & McCarroll, S. a. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics*, *91*(6), 1033–1040. doi:10.1016/j.ajhg.2012.10.018

Kriaucionis, S., & Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, *324*(5929), 929–30. doi:10.1126/science.1169786

Krueger, F., Kreck, B., Franke, A., & Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, *9*(2), 145–151. doi:10.1038/nmeth.1828

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–8. doi:10.1038/nature12213

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, *69*(6), 915–926. doi:10.1016/0092-8674(92)90611-F

Li, W., & Liu, M. (2011). Distribution of 5-hydroxymethylcytosine in different human tissues. *Journal of Nucleic Acids*, *2011*, 870726. doi:10.4061/2011/870726

Lindahl, T., & Nyberg, B. (1974). Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, *13*(16), 3405–3410. doi:10.1021/bi00713a035

Lister, R., Mukamel, E. a, Nery, J. R., Urich, M., Puddifoot, C. a, Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., … Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*, *341*(6146), 1237905. doi:10.1126/science.1237905

Maiti, A., & Drohat, A. C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *Journal of Biological Chemistry*, *286*(41), 35334–35338. doi:10.1074/jbc.C111.284620

Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S., & Heintz, N. (2012). MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*, *151*(7), 1417–1430. doi:10.1016/j.cell.2012.11.022

Mittlböck, M., & Heinzl, H. (2004). Pseudo R-squared measures for generalized linear models. In *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance* (pp. 71–80). Milan, Italy.

Moréra, S., Grin, I., Vigouroux, A., Couvé, S., Henriot, V., Saparbaev, M., & Ishchenko, A. a. (2012). Biochemical and structural characterization of the glycosylase domain of MBD4 bound to thymine and 5-hydroxymethyuracil-containing DNA. *Nucleic Acids Research*, *40*(19), 9917–9926. doi:10.1093/nar/gks714

Nabel, C. S., Jia, H., Ye, Y., Shen, L., Goldschmidt, H. L., Stivers, J. T., Zhang, Y., & Kohli, R. M. (2012). AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nature Chemical Biology*, *8*(9), 751–758. doi:10.1038/nchembio.1042

Nestor, C. E., Ottaviano, R., Reddington, J., Sproul, D., Reinhardt, D., Dunican, D., Katz, E., Dixon, J. M., Harrison, D. J., & Meehan, R. R. (2012). Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Research*, 467–477. doi:10.1101/gr.126417.111

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. a, Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979–93. doi:10.1016/j.cell.2012.04.024

Nilsen, H., Haushalter, K. a., Robins, P., Barnes, D. E., Verdine, G. L., & Lindahl, T. (2001). Excision of deaminated cytosine from the vertebrate genome: Role of the SMUG1 uracil-DNA glycosylase. *EMBO Journal*, *20*(15), 4278–4286. doi:10.1093/emboj/20.15.4278

Okano, M., Bell, D. W., Haber, D. a., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257. doi:10.1016/S0092-8674(00)81656-6

Pacis, A., Tailleux, L., Morin, A. M., Lambourne, J., MacIsaac, J. L., Yotova, V., Dumaine, A., Danckaert, A., Luca, F., Grenier, J., Hansen, K. D., Gicquel, B., Yu, M., Pai, A., He, C., Tung, J., Pastinen, T., … Barreiro, L. B. (2015). Bacterial infection remodels the DNA methylation landscape of human

603 dendritic cells. *Genome Research*, *25*(12), 1801–11. doi:10.1101/gr.192005.115

604 Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare,
605 D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordonez, G. R., Mudie, L. J.,
606 Latimer, C., Edkins, S., … Campbell, P. J. (2010). A small-cell lung cancer genome with complex
607 signatures of tobacco exposure. *Nature*, *463*(7278), 184–190. doi:10.1038/nature08629

608 Poon, S. L., Pang, S.-T., McPherson, J. R., Yu, W., Huang, K. K., Guan, P., Weng, W.-H., Siew, E. Y., Liu,
609 Y., Heng, H. L., Chong, S. C., Gan, A., Tay, S. T., Lim, W. K., Cutcutache, I., Huang, D., Ler, L. D., …
610 Teh, B. T. (2013). Genome-wide mutational signatures of aristolochic acid and its application as
611 a screening tool. *Science Translational Medicine*, *5*(197), 197ra101.
612 doi:10.1126/scitranslmed.3006086

613 Rangam, G., Schmitz, K. M., Cobb, A. J. a, & Petersen-Mahrt, S. K. (2012). AID enzymatic activity is
614 inversely proportional to the size of cytosine c5 orbital cloud. *PLoS ONE*, *7*(8), 3–8.
615 doi:10.1371/journal.pone.0043279

616 Rasmussen, K. D., & Helin, K. (2016). Role of TET enzymes in DNA methylation , development , and
617 cancer. *Genes & Development*, *30*, 733–750. doi:10.1101/gad.276568.115.maintain

618 Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka:
619 Accurate somatic small-variant calling from sequenced tumor-normal sample pairs.
620 *Bioinformatics*, *28*(14), 1811–1817. doi:10.1093/bioinformatics/bts271

621 Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. a., Leung, D., Rajagopal, N., Nery,
622 J. R., Urich, M. a., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski,
623 T. J., … Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation
624 variation. *Nature*, *523*(7559), 212–216. doi:10.1038/nature14465

625 Schuster-Böckler, B., & Lehner, B. (2012). Chromatin organization is a major influence on regional
626 mutation rates in human cancer cells. *Nature*, *488*(7412), 504–7. doi:10.1038/nature11273

627 Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A. C., Fung, H. L., Zhang, K., & Zhang, Y. (2013).
628 Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics.
629 *Cell*, *153*(3), 692–706. doi:10.1016/j.cell.2013.04.002

630 Spruijt, C. G., Gnerlich, F., Smits, A. H., Pfaffeneder, T., Jansen, P. W. T. C., Bauer, C., Münzel, M.,
631 Wagner, M., Müller, M., Khan, F., Eberl, H. C., Mensinga, A., Brinkman, A. B., Lephikov, K.,
632 Müller, U., Walter, J., Boelens, R., … Vermeulen, M. (2013). Dynamic readers for 5-
633 (hydroxy)methylcytosine and its oxidized derivatives. *Cell*, *152*(5), 1146–59.
634 doi:10.1016/j.cell.2013.02.004

635 Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Zainal, S. N.,
636 Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A.,
637 Cheverton, A., Gamble, J., Hinton, J., … Ottestad, L. (2012). The landscape of cancer genes and
638 mutational processes in breast cancer. *Nature*, *486*(7403), 400–404. doi:10.1038/nature11017

639 Supek, F., Lehner, B., Hajkova, P., & Warnecke, T. (2014). Hydroxymethylated cytosines are
640 associated with elevated C to G transversion rates. *PLoS Genetics*, *10*(9), e1004585.
641 doi:10.1371/journal.pgen.1004585

642 Szwagierczak, A., Bultmann, S., Schmidt, C. S., Spada, F., & Leonhardt, H. (2010). Sensitive enzymatic
643 quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Research*, *38*(19),
644 e181. doi:10.1093/nar/gkq684

645 Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M.,

Liu, D. R., Aravind, L., & Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, *324*(5929), 930–5. doi:10.1126/science.1170116

Takai, H., Masuda, K., Sato, T., Sakaguchi, Y., Suzuki, T., Suzuki, T., Koyama-Nasu, R., Nasu-Nishimura, Y., Katou, Y., Ogawa, H., Morishita, Y., Kozuka-Hata, H., Oyama, M., Todo, T., Ino, Y., Mukasa, A., Saito, N., … Akiyama, T. (2014). 5-Hydroxymethylcytosine Plays a Critical Role in Glioblastomagenesis by Recruiting the CHTOP-Methylosome Complex. *Cell Reports*, *9*(1), 48–60. doi:10.1016/j.celrep.2014.08.071

Taylor, B. J. M., Nik-Zainal, S., Wu, Y. L., Stebbings, L. a., Raine, K., Campbell, P. J., Rada, C., Stratton, M. R., & Neuberger, M. S. (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife*, *2013*(2), 1–14. doi:10.7554/eLife.00534

Tomasetti, C., Vogelstein, B., & Parmigiani, G. (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proceedings of the National Academy of Sciences*, *110*(6), 1999–2004. doi:10.1073/pnas.1221068110

Tomkova, M., McClellan, M., Kriaucionis, S., & Schuster-Böckler, B. (2016). Code and intermediate files for “5-hydroxymethylcytosine marks regions with reduced mutation frequency.” doi:10.6084/m9.figshare.c.3249394

Visvader, J. E. (2011). Cells of origin in cancer. *Nature*, *469*(7330), 314–322. doi:10.1038/nature09781

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H. N., Shi, S. T., Siu, H. C., Deng, S., Chu, K. M., Law, S., Chan, K. H., Chan, A. S. Y., Tsui, W. Y., Ho, S. L., Chan, A. K. W., Man, J. L. K., Foglizzo, V., … Leung, S. Y. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature Genetics*, *46*(6), 573–82. doi:10.1038/ng.2983

Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C., Yu, M., Liu, X., Zhu, P., Li, X., Hou, Y., Guo, H., Wu, X., … Qiao, J. (2014). Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biology*, *15*(3), R49. doi:10.1186/gb-2014-15-3-r49

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C. J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., & Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature Communications*, *2*, 241. doi:10.1038/ncomms1240

Wu, H., & Zhang, Y. (2011). Tet1 and 5-hydroxymethylation: A genome-wide view in mouse embryonic stem cells. *Cell Cycle*, *10*(15), 2428–2436. doi:10.4161/cc.10.15.16930

Wu, S., Powers, S., Zhu, W., & Hannun, Y. A. (2015). Substantial contribution of extrinsic risk factors to cancer development. *Nature*, *529*(7584), 43–47. doi:10.1038/nature16166

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J. H., Jin, P., Ren, B., & He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, *149*(6), 1368–1380. doi:10.1016/j.cell.2012.04.027

686 **FIGURE LEGENDS**

687 **Figure 1**

688 **C>T mutations are common in the genome but depleted in 5hmC sites compared to 5mC sites. A:**
689 Distribution of 5hmC in a CpG context in brain compared to kidney and blood. **B:** Frequency of SNVs
690 in brain cancer exomes, stratified by sequence context, normalised by frequency of trinucleotides. **C:**
691 Distribution of single-nucleotide variants (whole genomes) in brain cancer according to type, context
692 and modification state. **D:** CpG>T mutation frequency (black), 5hmC (blue) and 5mC (orange) density
693 in 100kbp windows of chromosome 3, smoothed with a Gaussian filter (n=50, sigma=2.5). **E:** Average
694 fraction of mutated sites for $5mC_{high}$ vs. $5hmC_{high}$ over all patient samples (CpG sites only; ***P <
695 0.001; **P < 0.01; *P < 0.05, see Methods).

696 **Figure 2**

697 **Differential mutation frequency between 5mC and 5hmC is present in all 5 brain cancer types and**
698 **correlates with age at diagnosis. A:** Average fraction of mutated CpG sites for $5mC_{high}$ vs. $5hmC_{high}$
699 computed separately for each cancer type. **B:** Box plot of C>T mutation frequency, as shown in A. **C.:**
700 Correlation of whole genome CpG>T mutation frequency with age at the time of diagnosis in
701 patients with Medulloblastoma and Pilocytic Astrocytoma.

702 **Figure 3**

703 **Depletion of C>T mutations in 5hmC sites is not explained by gene expression or regional mutation**
704 **rate variation. A–B:** Frequency of mutations in $5mC_{high}$ vs $5hmC_{high}$ sites within highly expressed (A)
705 or lowly expressed (B) genes (see Methods). **C–D:** Boxplot visualisation of C>T mutation frequency
706 for each cancer type. **E:** For each patient sample, the overall difference in mutations in paired sites
707 was calculated and compared using a Wilcoxon signed-rank test. Shown here is a histogram of
708 samples by the difference in mutations for paired 5mC and 5hmC sites (negative values shown blue,
709 positive in orange; see Methods for details). Mutations in 5mC sites exceed paired 5hmC sites,

710 causing a shift to the right. **F:** Same as E but using a more stringent definition of 5mC (only sites with
711 $threshold_{5mC} \leq 0.2$).

712 **Figure 4**

713 **Mutation frequency negatively correlates with $5hmC_{rel}$ level per base. A:** Fraction of mutated CpG
714 sites as a function of $5hmC_{rel}$ levels by mutation and cancer type. Bins to the left represent sites
715 predominantly methylated, while bins to the right contain increasingly hydroxymethylated sites.
716 Black line denotes linear regression fit (F-test for coefficient deviation from 0, see Methods). **B:**
717 Distribution of CpG>T mutation frequency by modification type. The top left bin contains cytosines
718 that are mostly unmodified, the bottom left bin contains exclusively methylated cytosines and the
719 top right bin contains cytosines that are mostly hydroxymethylated. **C:** Top row of B, *i.e.* distribution
720 of mutations in unmethylated sites. **D:** First column of B, *i.e.* distribution of mutations in sites
721 without 5hmC. **E:** Diagonal of B, i.e. distribution of mutations in highly modified sites.

722 **Figure 5**

723 **Predictors of mutations: $5hmC_{rel}$ compared to other genomic features. A:** Prediction of CpG>T
724 mutation frequency (using whole genome sequencing only) in 100kbp genomic windows. Predictors
725 are sorted according to the D^2 in a univariate model. The height of the k^{th} bar denotes the D^2 of a
726 model with the first *k* predictors. **B:** Comparison of the nine predictors of CpG>T mutation features
727 by D^2 in a univariate models, in a range of window sizes. **C:** Prediction of different types of mutation
728 frequency in genes. Increase in D^2 of a generalised linear model including $5hmC_{rel}$ over gene
729 expression (purple) or gene expression over $5hmC_{rel}$ (green) (see Methods). **D:** Significance of
730 observations in C (see Methods).

731 **Figure 6**

732 **Decreased CpG>T mutation frequency in 5hmC is not limited to brain tissue. A:** Predictions of
733 CpG>T mutation frequency in whole genome cancers in blood (AML), kidney and brain using $5hmC_{rel}$

maps from blood, kidney, brain and embryonic stem cells (ESC) in 100kbp genomic windows. The values are z-score normalised per rows in order to normalise for different number of patients and mutations in each cancer type (the original D^2 values are in parentheses); the higher values of D^2 (green), the better predictions. B: CpG>T mutation frequency in 5mC vs. 5hmC in kidney and blood. **C:** Correlation of total $5hmC_{rel}$ levels (measured with HPLC) with frequency of CpG>T mutations in modified cytosines normalised by the frequency in unmodified cytosines in different tissues (see Methods).

TABLE LEGENDS

Table 1: Overview of BS-Seq and TAB-Seq data used to generate modification maps.

Table 2: Overview of whole genome and exome sequencing data used for mutation information.

SUPPLEMENTARY FIGURE LEGENDS

Figure 1-figure supplements 1–2: **Distribution of CpG>T mutations vs modifications across all chromosomes**: CpG>T mutation frequency (black), 5hmC (blue) and 5mC (orange) density in 100kbp windows, smoothed with a Gaussian filter (n=50, sigma=2.5).

Figure 1-figure supplement 3: **Distribution of CpG>T mutations vs other genomic features**: CpG>T mutation frequency (black) and several genomic features in 100kbp windows on chromosome 3, smoothed with a Gaussian filter (n=50, sigma=2.5). CGIs: density of CpG islands, EXONs: density of exons, GENEs: density of genes, CpG: density of CpGs, modCpG: density of CpGs with *mod level* ≥ 10%; and average modification levels: mod, 5hmC, 5mC, and $5hmC_{rel}$.

Figure 2-figure supplement 1: **Depletion of C>T mutations in $5hmC_{high}$ is relatively insensitive to varying definitions of $5mC_{high}$ and $5hmC_{high}$. A–F:** Significance of a difference in mutation frequency in $5mC_{high}$ and $5hmC_{high}$, for a range of values of $threshold_{5mC}$ and $threshold_{5hmC}$ ($5mC_{high}$ is defined as sufficiently modified sites with $5hmC_{rel} \leq threshold_{5mC}$; $5hmC_{high}$ is defined as sufficiently modified

sites with $5hmC_{rel} \geq threshold_{5hmC}$). One-sided paired Wilcoxon sign-rank test was used. Red colour represents a significant increase of mutation frequency in $5mC_{high}$ (right tail test) whereas blue colour represents elevated mutations in $5hmC_{high}$ (left tail test). **G–I:** C>T mutation frequency for $5mC_{high}$ vs. $5hmC_{high}$ with $threshold_{5mC}$ =0.3 and $threshold_{5hmC}$ =0.7.

Figure 3-figure supplement 1: **Depletion of C>T mutations in $5hmC_{high}$ is relatively insensitive to varying definitions of $5mC_{high}$ and $5hmC_{high}$. A–D:** C>T mutation frequency for $5mC_{high}$ vs. $5hmC_{high}$ in highly vs. lowly expressed genes with $threshold_{5mC}$ =0.3 and $threshold_{5hmC}$ =0.7. **E–F:** C>G mutation frequency with $threshold_{5mC}$ =0.0 and $threshold_{5hmC}$ =0.5. **G:** Mutation frequency around aligned 5mC and 5hmC sites.

Figure 4-figure supplement 1: **CpG>T mutation frequency as a function of $5hmC_{rel}$ levels with equal binning** (each bin contains approximately the same number of sites).

Figure 5-figure supplement 1: **Genome-wide prediction of CpG>T mutation frequency: $5hmC_{rel}$ compared to other genomic features. A–C:** Comparison of nine predictors of CpG>T mutation frequency in a range of window sizes by p-value of univariate generalised linear models (**A**), Spearman correlation (**B**), and Pearson correlation (**C**). **D:** Effects of window size and patient numbers on D^2 of GLM with one response variable (simulated mutation frequency) generated proportionally from a single ideal predictor (see methods for details).

Figure 5-figure supplement 2: **Effects of $5hmC_{rel}$ levels on gene mutability.** Data for GLM with Poisson distribution (the fitted curve is in green). Genes defined as outliers in at least one definition of mutation frequency (above the red line) are plotted in red. For convenience, the mutation frequency is plotted on log-scale.

Figure 5-figure supplement 3: **Effects of $5hmC_{rel}$ levels on gene mutability. A–B:** GLM results fitted separately for $5hmC_{rel}$ (purple) and gene expression (green) and both of them together (yellow). **C–D:** Frequency of modC>T mutations of all genes (**C**) and gene density (**D**) in the space of $5hmC_{rel}$ and

gene expression. Briefly, for figures **C** and **D** the space was limited to [quantile(x, 0.05), quantile(x, 0.95)] on both axes and then binned into 100x100 bins. In each bin, the average mutation frequency (in the form of *log(mutFreq + min(mutFreq(mutFreq>0)))*) was computed. The resulting matrix was smoothed by applying a Gaussian filter (radius 5 bins, sigma 2) weighted by the number of genes in each bin (bins with >= 2/3 missing values in their neighbourhood were set to NaN) and plotted with `pcolor` (NaN bins are shown in black).

Figure 6-figure supplement 1: **Decreased CpG>T mutation frequency in 5hmC is present in three tissues consistently for different replicates of modification maps. A:** CpG>T mutation frequency in 5mC compared to 5hmC in blood and kidney using modification maps from different replicates merged together (A) and used separately (B). **C:** Predictions of CpG>T mutation frequency in whole genome cancers in blood (AML), kidney and brain using different replicates of $5hmC_{rel}$ maps from blood, kidney, brain and embryonic stem cells (ESC) in 100kbp genomic windows. The values are z-score normalised per rows in order to normalise for different number of patients and mutations in each cancer type (the original D^2 values are in parentheses); the higher values of D^2 (green colour), the better predictions.

Figure 6-figure supplement 2: **Comparison of 5hmC in 10kbp windows in blood, kidney (2 replicates), and brain**: distribution of 5hmC values in each map and Pearson correlation of pairs of maps.

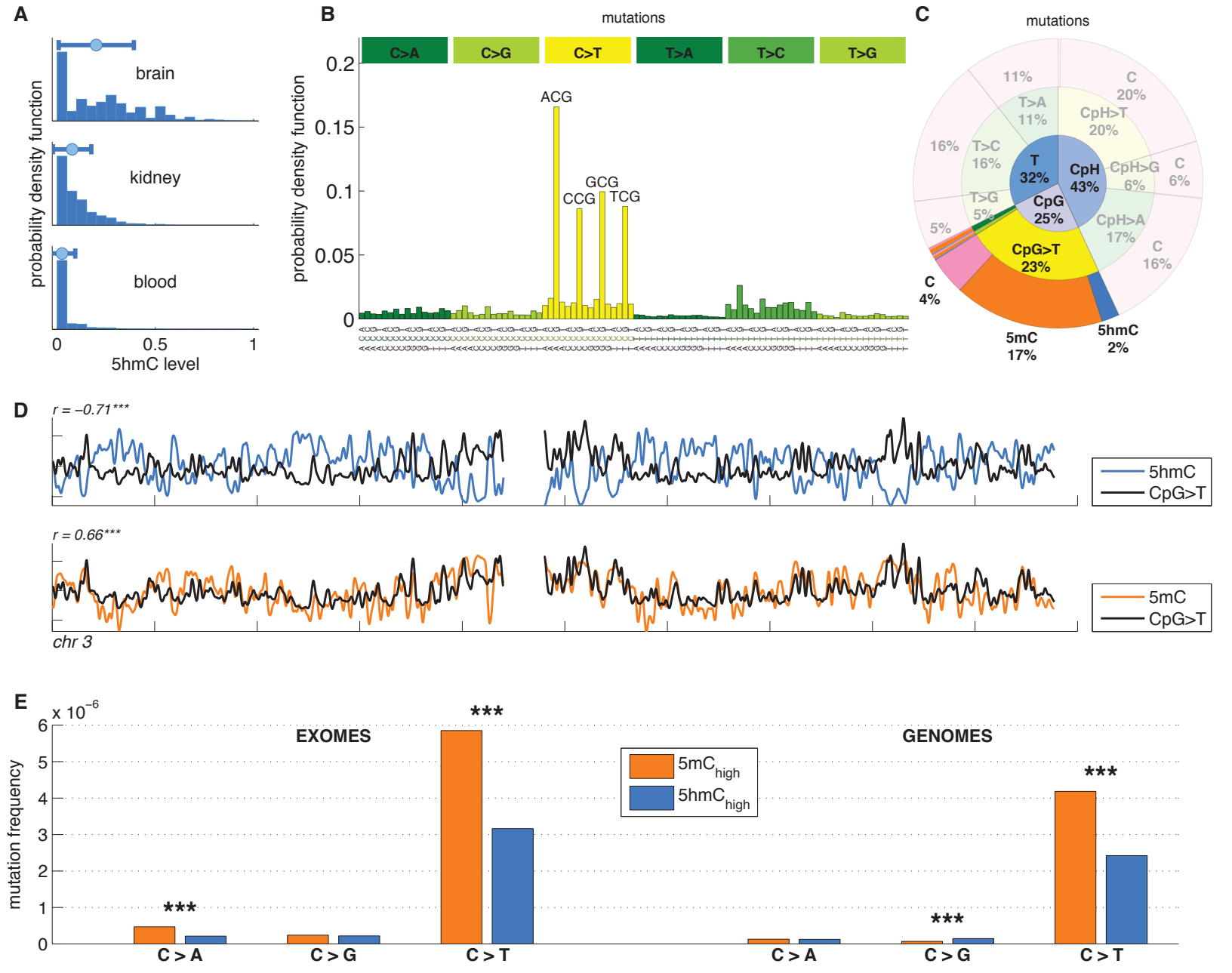
Figure 6-figure supplement 3: **HPLC measurements of total 5hmC and 5mC in eight tissues**: average values with standard deviation of 5mC and 5hmC (as a percentage of total cytosine).

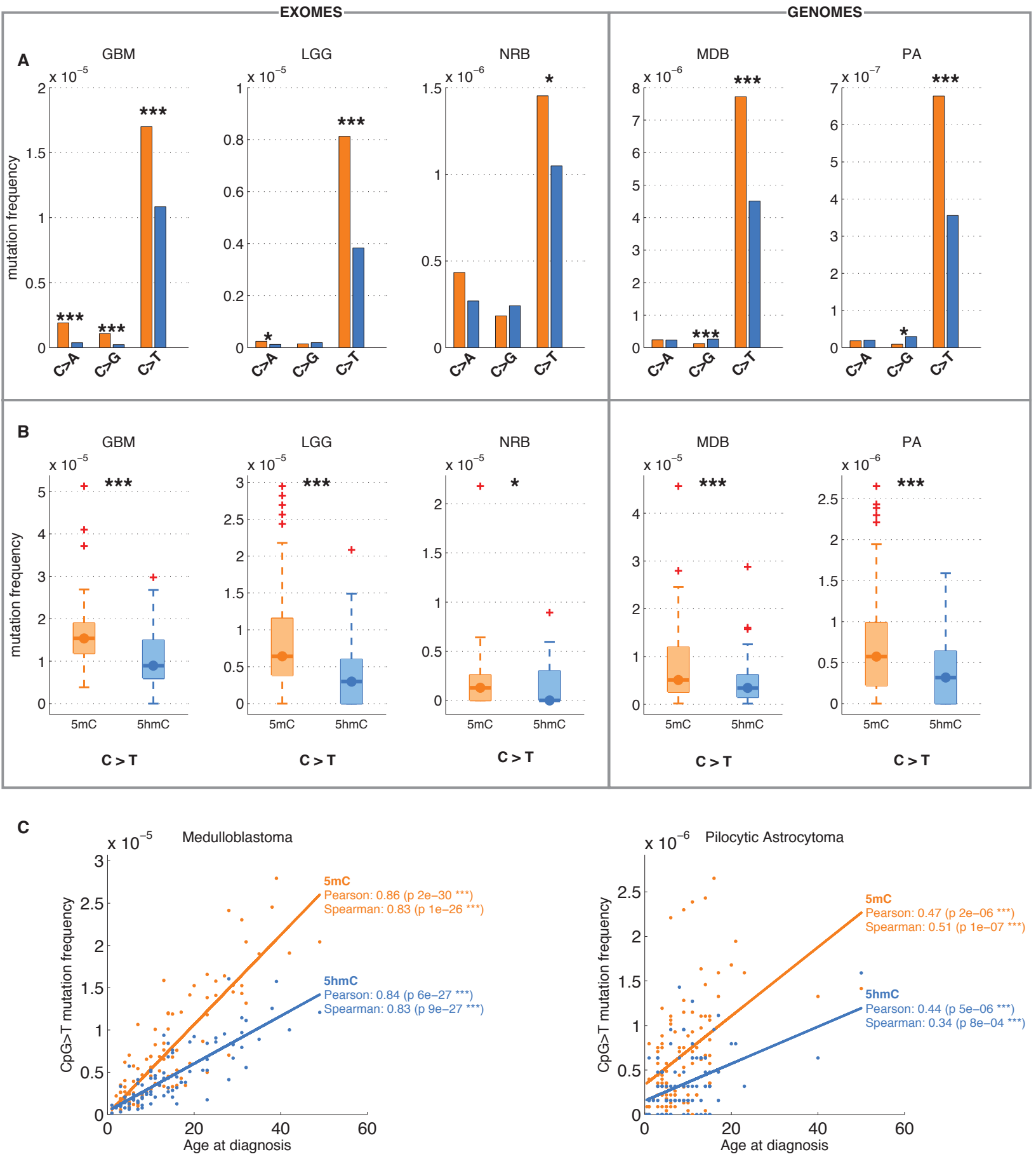
Cancer type	Exome		Whole genome		Tissue	Source
	Patients	SNVs	Patients	SNVs		
Glioblastoma	39	22954	-	-	brain	(Alexandrov et al., 2013)
Glioma Low Grade	215	9678	-	-	brain	(Alexandrov et al., 2013)
Medulloblastoma	-	-	100	139553	brain	(Alexandrov et al., 2013)
Neuroblastoma	210	5027	-	-	brain	(Alexandrov et al., 2013)
Pilocytic Astrocytoma	-	-	101	12989	brain	(Alexandrov et al., 2013)
Kidney Chromophobe	65	1646	-	-	kidney	(Alexandrov et al., 2013)
Kidney Clear Cell	325	30273	-	-	kidney	(Alexandrov et al., 2013)
Kidney Papillary	100	6479	-	-	kidney	(Alexandrov et al., 2013)
RECA-EU	-	-	95	488922	kidney	ICGC
Acute Myeloid Leukaemia	147	2214	7	3659	blood myeloid	(Alexandrov et al., 2013)
Acute Myeloid Leukaemia	-	-	49	176164	blood myeloid	TCGA
Acute Lymphoblastic Leukaemia	140	1869	1	7881	blood other	(Alexandrov et al., 2013)
Chronic Lymphoid Leukaemia	103	3998	28	54746	blood other	(Alexandrov et al., 2013)
Lymphoma B-cell	24	824	24	142753	blood other	(Alexandrov et al., 2013)
Myeloma	69	3973	-	-	blood other	(Alexandrov et al., 2013)
Breast	844	55731	119	647692	breast	(Alexandrov et al., 2013)
Liver	-	-	88	899445	liver	(Alexandrov et al., 2013)
Lung Adeno	636	248519	24	1505512	lung	(Alexandrov et al., 2013)
Lung Small Cell	70	17639	-	-	lung	(Alexandrov et al., 2013)
Lung Squamous	176	70485	-	-	lung	(Alexandrov et al., 2013)
Pancreas	98	5093	15	122787	pancreas	(Alexandrov et al., 2013)
Stomach	212	102110	0	0	stomach	(Alexandrov et al., 2013)
Stomach	0	0	100	1995618	stomach	(Wang et al., 2014)

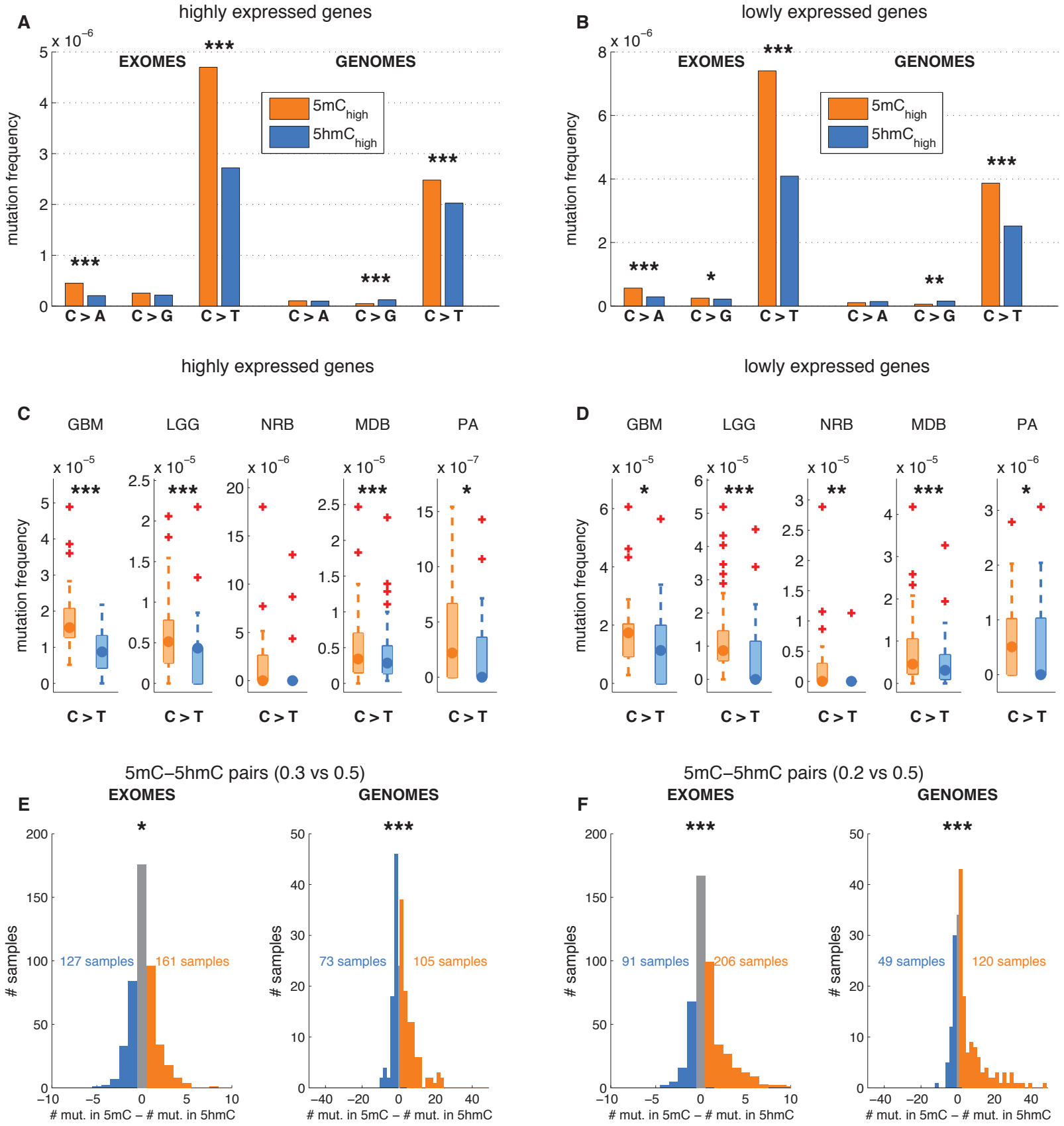
802

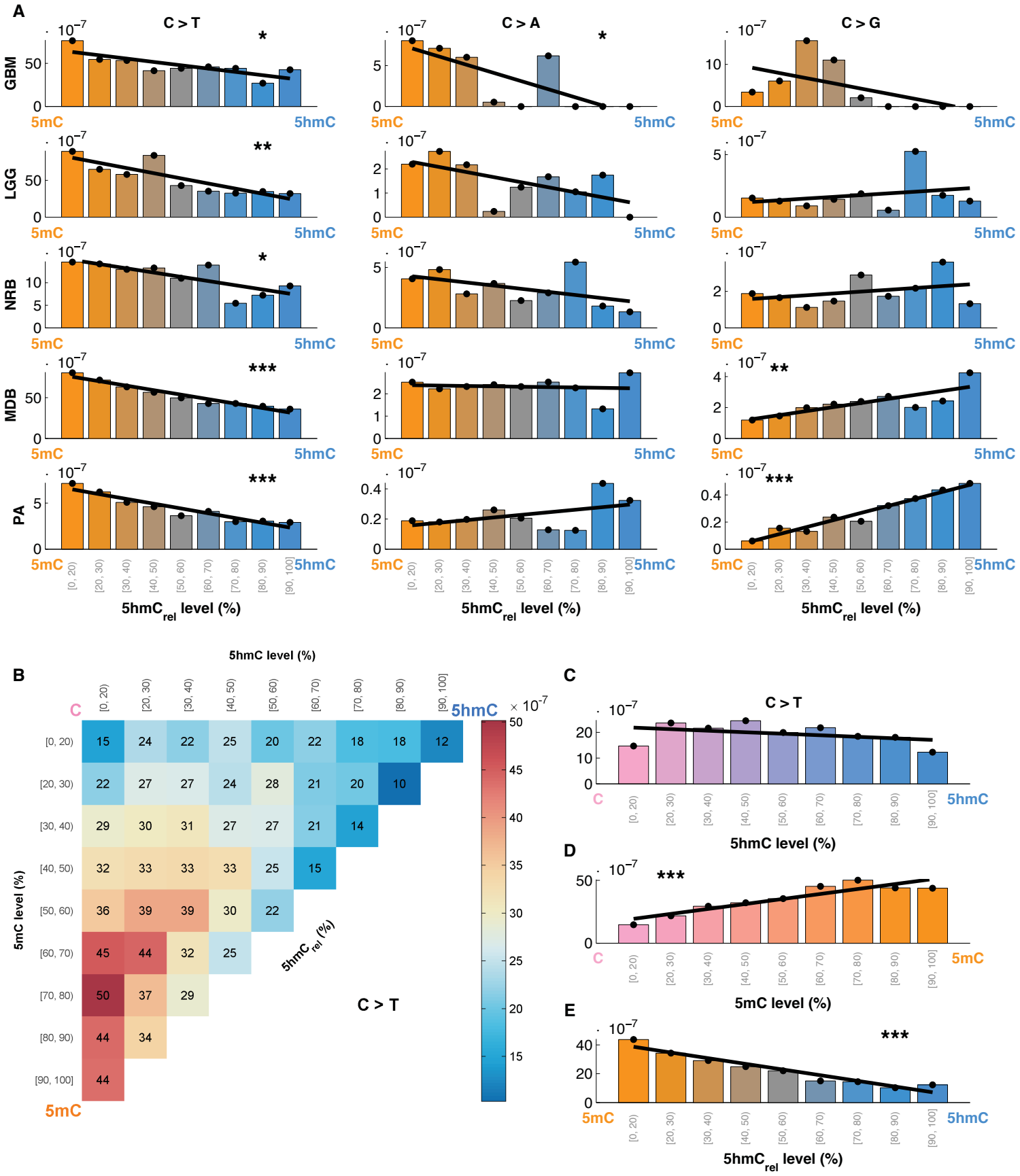
Tissue	BS-Seq		TAB-Seq		Source	Link
	CpGs	Average	CpGs	Average		
Brain	53388534	0.791200	53847986	0.221845	(Wen et al., 2014)	BS, TAB
Kidney r1	54117976	0.759816	46303252	0.088314	(Chen et al., 2015)	BS, TAB
Kidney r2	53861360	0.756454	54585341	0.094624	(Chen et al., 2015)	BS, TAB
Kidney total	54857866	0.757601	54928295	0.092868	(Chen et al., 2015)	
Blood dendritic r1	24586388	0.791371	-	-	(Pacis et al., 2015)	BS
Blood dendritic r2	23524704	0.786322	-	-	(Pacis et al., 2015)	BS
Blood dendritic r3	24419613	0.782467	-	-	(Pacis et al., 2015)	BS
Blood dendritic r4	24452547	0.787728	-	-	(Pacis et al., 2015)	BS
Blood dendritic r5	24399654	0.774058	-	-	(Pacis et al., 2015)	BS
Blood dendritic r6	24533745	0.790145	-	-	(Pacis et al., 2015)	BS
Blood dendritic total	25586845	0.789083	27754454	0.029103	(Pacis et al., 2015)	TAB
Breast	53222114	0.735273	-	-	Epigenome Roadmap	BS
Pancreas	54341922	0.697484	-	-	Epigenome Roadmap	BS
Lung	54236520	0.774050	-	-	Epigenome Roadmap	BS
Liver	51884076	0.757123	-	-	Epigenome Roadmap	BS
Stomach	54054176	0.762082	-	-	Epigenome Roadmap	BS
Blood (hematopoietic multipotent progenitor cell)	51822931	0.852170	-	-	Blueprint Epigenome	BS

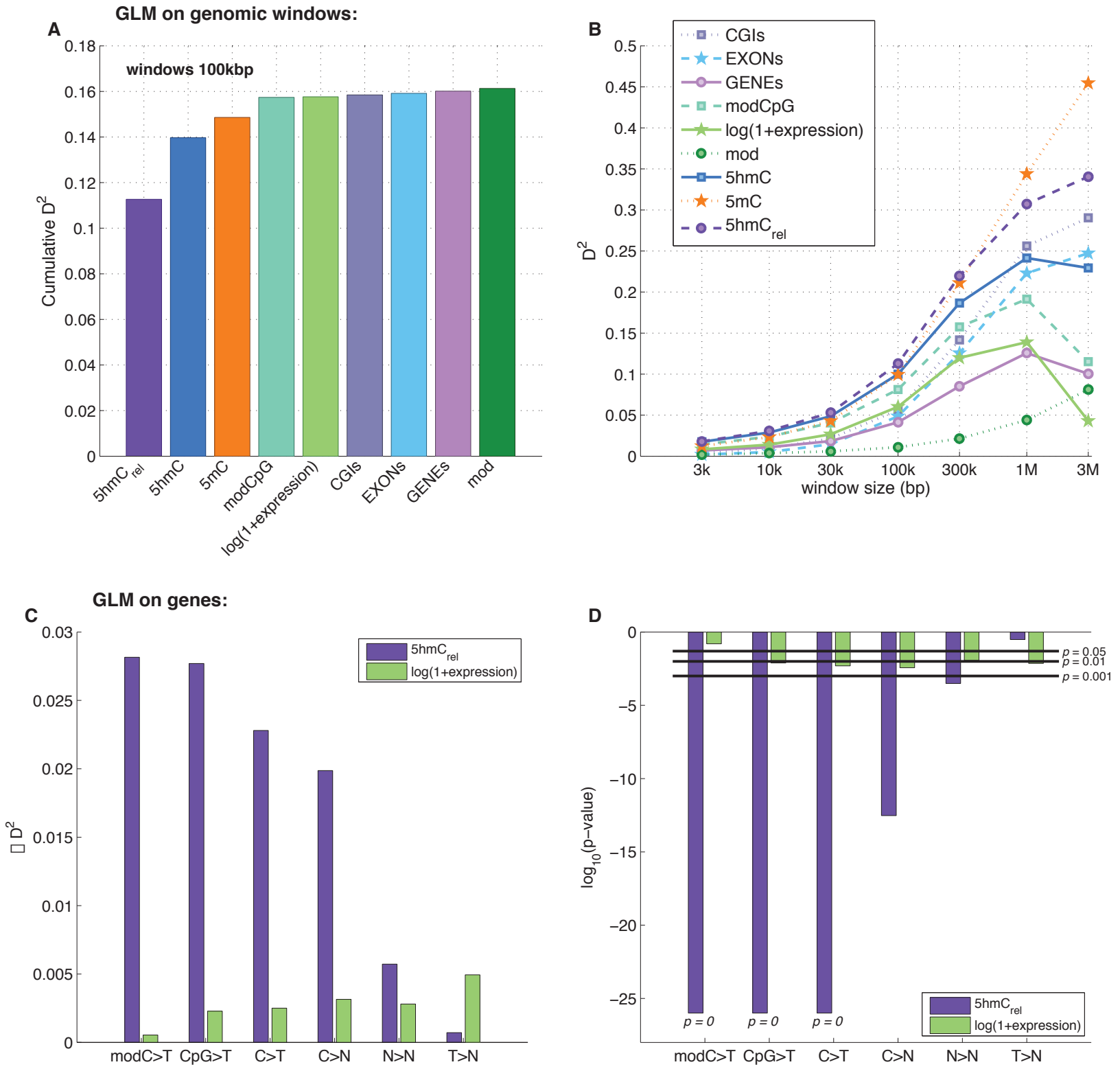
803

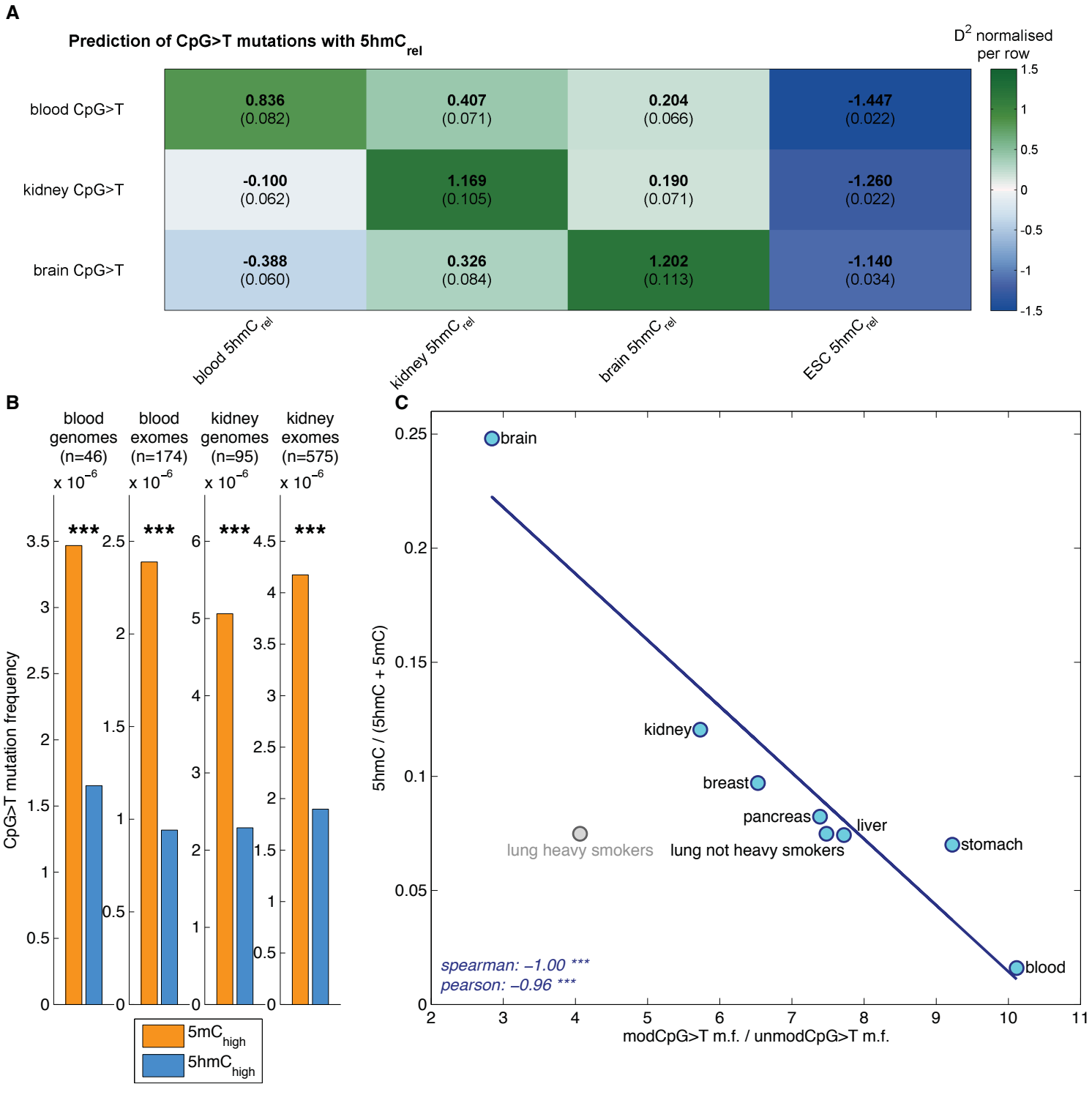

A
brain
kidney
blood
probability density function
0
0.5
1
5hmC level
B
mutations
C>A
C>G
C>T
T>A
T>C
T>G
0.2
0.15
0.1
0.05
0
probability density function
ACG
GCG
CCG
TCG
C
mutations
11%
C
20%
T>A
11%
CpH>T
20%
16%
T>C
16%
T
32%
CpH
43%
CpH>G
6%
C
6%
T>G
5%
CpG
25%
5%
CpH>A
17%
CpG>T
23%
C
16%
C
4%
5mC
17%
5hmC
2%
D
r = −0.71***
5hmC
CpG>T
r = 0.66***
5mC
CpG>T
chr 3
E
x 10−6
6
5
4
3
2
1
0
mutation frequency
EXOMES
GENOMES
5mChigh
5hmChigh
***
***
***
***
C > A
C > G
C > T
C > A
C > G
C > T


EXOMES
GENOMES
A
GBM
LGG
NRB
MDB
PA
mutation frequency
x 10^−5
x 10^−6
x 10^−7
C>A
C>G
C>T
***
*
B
5mC
5hmC
C > T
C
Medulloblastoma
Pilocytic Astrocytoma
CpG>T mutation frequency
Age at diagnosis
5mC
Pearson: 0.86 (p 2e−30 ***)
Spearman: 0.83 (p 1e−26 ***)
5hmC
Pearson: 0.84 (p 6e−27 ***)
Spearman: 0.83 (p 9e−27 ***)
Pearson: 0.47 (p 2e−06 ***)
Spearman: 0.51 (p 1e−07 ***)
Pearson: 0.44 (p 5e−06 ***)
Spearman: 0.34 (p 8e−04 ***)


A
highly expressed genes
x 10^−6
EXOMES
GENOMES
5mC high
5hmC high
mutation frequency
C > A
C > G
C > T
B
lowly expressed genes
C
highly expressed genes
GBM
LGG
NRB
MDB
PA
D
lowly expressed genes
E
5mC–5hmC pairs (0.3 vs 0.5)
127 samples
161 samples
73 samples
105 samples
samples
mut. in 5mC – # mut. in 5hmC
F
5mC–5hmC pairs (0.2 vs 0.5)
91 samples
206 samples
49 samples
120 samples


A
C > T
C > A
C > G
GBM
LGG
NRB
MDB
PA
5mC
5hmC
10^{-7}
*
**
***
[0, 20)
[20, 30)
[30, 40)
[40, 50)
[50, 60)
[60, 70)
[70, 80)
[80, 90)
[90, 100]
5hmC$_{rel}$ level (%)
B
5hmC level (%)
5mC level (%)
5hmC$_{rel}$ (%)
C > T
$\times 10^{-7}$
C
D
E
5mC level (%)


GLM on genomic windows:
A
windows 100kbp
Cumulative D2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5hmC rel
5hmC
5mC
modCpG
log(1+expression)
CGIs
EXONs
GENEs
mod
B
CGIs
EXONs
GENEs
modCpG
log(1+expression)
mod
5hmC
5mC
5hmC rel
D2
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3k
10k
30k
100k
300k
1M
3M
window size (bp)
GLM on genes:
C
5hmC rel
log(1+expression)
D2
0.03
0.025
0.02
0.015
0.01
0.005
0
modC>T
CpG>T
C>T
C>N
N>N
T>N
D
p = 0.05
p = 0.01
p = 0.001
log10(p−value)
0
−5
−10
−15
−20
−25
p = 0
p = 0
p = 0
5hmC rel
log(1+expression)
modC>T
CpG>T
C>T
C>N
N>N
T>N


A
Prediction of CpG>T mutations with 5hmC rel
D2 normalised per row
blood CpG>T
kidney CpG>T
brain CpG>T
0.836 (0.082)
0.407 (0.071)
0.204 (0.066)
-1.447 (0.022)
-0.100 (0.062)
1.169 (0.105)
0.190 (0.071)
-1.260 (0.022)
-0.388 (0.060)
0.326 (0.084)
1.202 (0.113)
-1.140 (0.034)
blood 5hmC rel
kidney 5hmC rel
brain 5hmC rel
ESC 5hmC rel
B
blood genomes (n=46)
blood exomes (n=174)
kidney genomes (n=95)
kidney exomes (n=575)
x 10−6
***
CpG>T mutation frequency
5mChigh
5hmChigh
C
brain
kidney
breast
pancreas
liver
stomach
lung heavy smokers
lung not heavy smokers
blood
spearman: −1.00 ***
pearson: −0.96 ***
5hmC / (5hmC + 5mC)
modCpG>T m.f. / unmodCpG>T m.f.