Tomkova, M., McClellan, M., Kriaucionis, S., & Schuster-Böckler, B. (2018). DNA replication and associated repair pathways are involved in the mutagenesis of methylated cytosine. *DNA repair*, *62*, 1-7. https://doi.org/10.1016/j.dnarep.2017.11.005

Self-archived version of the accepted manuscript

DNA Replication and associated repair pathways are involved in the mutagenesis of methylated cytosine

Marketa Tomkova[1], Michael McClellan[1], Skirmantas Kriaucionis[1*] and Benjamin Schuster-Böckler[1*]

AFFILIATION

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

CORRESPONDING AUTHORS

* Co-corresponding authors

- Benjamin Schuster-Böckler
  E-mail: benjamin.schuster-boeckler@ludwig.ox.ac.uk (BSB)
- Skirmantas Kriaucionis
  E-mail: skirmantas.kriaucionis@ludwig.ox.ac.uk (SK)

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

ABSTRACT

Transitions of cytosine to thymine in CpG dinucleotides are the most frequent type of mutations observed in cancer. This increased mutability is commonly explained by the presence of 5-methylcytosine (5mC) and its spontaneous hydrolytic deamination into thymine. Here, we describe observations that question whether spontaneous deamination alone causes the elevated mutagenicity of 5mC. Tumours with somatic mutations in DNA mismatch-repair genes or in the proofreading domain of DNA polymerase ε (Pol ε) exhibit more 5mC to T transitions than would be expected, given the kinetics of hydrolytic deamination. This enrichment is asymmetrical around replication origins with a preference for the leading strand template, in particular in methylated cytosines flanked by guanines (GCG). Notably, GCG to GTG mutations also exhibit strand asymmetry in mismatch-repair and Pol ε wild-type tumours. Together, these findings suggest that mis-incorporation of A opposite 5mC during replication of the leading strand might be a contributing factor in the mutagenesis of methylated cytosine.

KEYWORDS

Mutagenesis; DNA methylation; DNA Replication; Cancer genomics

1. INTRODUCTION

C to T transitions in a CpG context (CpG>TpG) are the most frequently observed mutations in cancer and genetic disorders [1,2]. Two independent observations link these mutations to 5-methylcytosine (5mC), an epigenetic modification of cytosine. First, most cytosines in CpG dinucleotides are methylated in humans [3]. Moreover, the increased C>T mutagenicity in CpG dinucleotides is present specifically in cytosines that are methylated, compared to unmodified or hydroxymethylated cytosines [4]. Second, it was shown *in vitro* that methylated cytosine (5mC) has a four-fold higher rate of spontaneous deamination than unmodified cytosine [5]. The products of deamination can be repaired by base excision repair (BER). DNA glycosylases involved in BER of T:G mismatches (*TDG* and *MBD4*) excise T from the mismatch, leading to the restoration of C:G [6,7]. Notably, while the deamination of 5mC produces thymine, leading to a T:G mismatch, C and 5-hydroxymethylcytosine

(5hmC) deaminate into uracil and 5-hydroxymethyluracil, respectively. Since these bases do not normally occur in DNA, they are potentially more efficiently recognised and replaced by BER [8]. Moreover, deamination of 5hmC does not contribute to the steady-state levels of 5hmU in mouse embryonic stem cells, suggesting either infrequent deamination or very fast repair [9]. Failure to correct the T:G mismatch before replication results in a mutation in one daughter cell, due to the semiconservative nature of DNA replication. Thus replication of a T:G mismatch leads to a C:G>T:A mutation.

Mutations can also arise through mis-incorporation of bases during cell division. The fidelity of DNA replication relies on proofreading by the major replicative polymerases Pol ε and Pol δ, and on post-replicative DNA mismatch-repair (MMR) which removes errors from the newly synthesised DNA strand [10]. Deficiency in any of these protective mechanisms leads to an increase in the number of mutations. In particular, defects in MMR genes lead to "hypermutability" (10^4-10^5 mutations per Gbp), and mutations in the proofreading domain of Pol ε lead to "ultra-hypermutability", often exceeding 10^5 mutations per Gbp [11,12]. Moreover, defects in Pol ε and Pol δ proofreading cause tumours in mice [13] and germline mutations in *POLE* and *POLD1* (encoding the catalytic subunits of Pol ε and δ, respectively) and genes of the MMR pathway predispose to cancer in humans [10].

DNA polymerase proofreading and post-replicative MMR (in their canonical, replication-linked functions) are highly unlikely to play a role in the repair of 5mC deamination induced mutations, as they operate *after* parental strands have been separated during replication, at which point a 5mC to T deamination event is indistinguishable from other thymines. Therefore, although the total frequency of mutations due to unrepaired errors introduced during replication increases drastically in polymerase proofreading/MMR deficient samples, it would be expected that the frequency of CpG>TpG mutations should only increase by a small amount.

Contrary to this expectation, we observe that the frequency of CpG>TpG mutations in tumours with defective Pol ε or MMR is approximately six-fold higher than for other types of mutations. We show that the increased CpG>TpG mutation rate in Pol ε or MMR mutant cancers is linked to DNA

methylation, has a clear replication strand asymmetry, being enriched on the leading strand, with a preference for a GCG sequence context. We also detect weaker but consistent replication strand asymmetry of GCG>GTG mutations in Pol ε and MMR proficient samples. Together, our results suggest that a substantial fraction of C>T mutations at methylated cytosines is independent of spontaneous deamination, instead arising during DNA replication.

2. Materials and Methods

2.1. Somatic mutations

Cancer somatic mutations in 3442 whole-genome sequencing samples (Supplementary Table 1) were obtained from the data portal of The Cancer Genome Atlas (TCGA), the data portal of the International Cancer Genome Consortium (ICGC), and previously published data in peer-review journals [1,12,14–16]. MSI and *POLE*-MUT samples were combined from previous studies [11,12,17]. For the TCGA samples, aligned reads of paired tumour and normal samples were downloaded from the UCSC CGHub website under TCGA access request #10140 and somatic variants were called using Strelka (version 1.0.14) [18] with default parameters. Somatic mutations in autosomes only were taken into account.

2.2. DNA modification maps

Maps of cytosine modifications (Supplementary Table 2) were obtained from BS-Seq data sets from the data portals of The Cancer Genome Atlas (TCGA), Roadmap Epigenome, Blueprint, and from previously published data in peer-review journals [19–22] and where needed converted to hg19 using liftover tool. For brain, kidney, and prostate maps, raw reads were processed with Trim galore, Bismark[23] and Mark duplicates from Picard tools; and only sites covered with at least 5 reads were taken into account.

2.3. Mutation frequency with respect to modification levels

All cytosines in the CpG context were divided into 10 right-open intervals according to their modification levels (the number of unconverted reads divided by the number of all reads in BS-Seq):

[0-0.1), [0.1-0.2), …, [0.9-1]. In each bin, the frequency of mutations was computed and plotted for each sample. A linear regression model was fitted to the data (function `fitlm` in MatLab) and the offset, slope, and last value, and fold-change from first to last value were measured. When comparing CpG sites with low vs. intermediate vs. high modification levels, the thresholds (0.8 and 0.95) were chosen such that the three groups have approximately similar numbers of CpG sites in most tissues.

2.4. Direction of replication

Left- and right-replicating domains were taken from [17]. Each domain (called territory in the original source code and data) is 20kbp wide and annotated with the direction of replication and with replication timing.

2.5. Mutation frequency with respect to the direction of replication

First, transitions between left- and right-replicated domains were computed as in [17]. These transitions represent regions rich for replication origins. We computed the CpG>TpG mutation frequency in the 20kbp domains distant 0 to 1Mbp from the closest left-/right- transition, with respect to the strand (plus=Watson vs. minus=Crick) of the cytosine of the CpG. Template for the leading strand then corresponds to the plus strand in the left direction and minus strand in the right direction and *vice versa* for the lagging strand template. Finally, we annotated all cytosines in a CpG context whether they are on the leading or lagging strand, and computed CpG>TpG mutation frequency for the leading and lagging strand separately. `Signtest` was used for evaluating significance of CpG>TpG mutation frequency difference between the two strands.

2.6. Spontaneous deamination estimates

The number of years needed to reach the observed number of C>T mutations in methylated CpGs observed in *POLE*-MUT and MSI samples was based on the spontaneous deamination rate of 5mC in double-stranded DNA ($5.8 \cdot 10^{-13}$ s^{-1}) reported by Shen et al. [24], the number of seconds in a year (31556736), the observed frequency of GCG>GTG mutations (*i.e.*, GmCG>T/GmCG; for mC with a

129 modification level of at least 0.9) in MSI ($5.133 \cdot 10^{-4}$) and *POLE*-MUT ($1.785 \cdot 10^{-3}$) samples, and

130 computed as:

131 $$\text{MSI: } \frac{5.133 \cdot 10^{-4}}{5.8 \cdot 10^{-13} \cdot 31556736} = 28.05 \text{ years}$$

132 $$\textit{POLE}\text{-MUT: } \frac{1.785 \cdot 10^{-3}}{5.8 \cdot 10^{-13} \cdot 31556736} = 97.53 \text{ years}$$

133

134 ## 3. RESULTS

135 We explored the mutation spectra of 14 tumour samples with a mutation in Pol ε (*POLE*-MUT
136 samples), 19 samples with microsatellite-instability (MSI) deficient in MMR, and 3409 other cancer
137 samples (proficient; PROF). The median overall mutation frequency per base was $1.5 \cdot 10^{-6}$ (IQR
138 $0.6 \cdot 10^{-6}$–$3.5 \cdot 10^{-6}$) in PROF samples, $36.9 \cdot 10^{-6}$ (IQR $18.0 \cdot 10^{-6}$–$47.4 \cdot 10^{-6}$) in MSI samples, and $267.4 \cdot 10^{-6}$
139 (IQR $99.9 \cdot 10^{-6}$–$300.5 \cdot 10^{-6}$) in *POLE*-MUT samples (N>N in Fig. 1A–B). In PROF samples, the median
140 CpG>TpG mutation frequency (i.e., the number of CpG>TpG mutations relative to the number of
141 CpGs in the genome) was $7.4 \cdot 10^{-6}$ (IQR $3.7 \cdot 10^{-6}$–$16.8 \cdot 10^{-6}$), approximately 5 fold higher than the
142 overall mutation frequency (i.e., the number of all mutations relative to the number of all positions
143 in the genome). Notably, the CpG>TpG mutation frequency also increased in MSI and *POLE*-MUT
144 samples, compared to the overall mutation frequency (MSI: median $247.7 \cdot 10^{-6}$ per CpG, IQR
145 $162.7 \cdot 10^{-6}$–$367.3 \cdot 10^{-6}$; *POLE*-MUT: median $1559.8 \cdot 10^{-6}$ per CpG, IQR $707.9 \cdot 10^{-6}$–$2574.2 \cdot 10^{-6}$)
146 (CpG>TpG in Fig. 1A–B, Fig. 1-supplement 1). This observation is surprising, since neither MMR nor
147 proofreading during DNA replication by Pol ε are thought to be essential for effective repair of
148 deamination induced T:G mismatches [8].

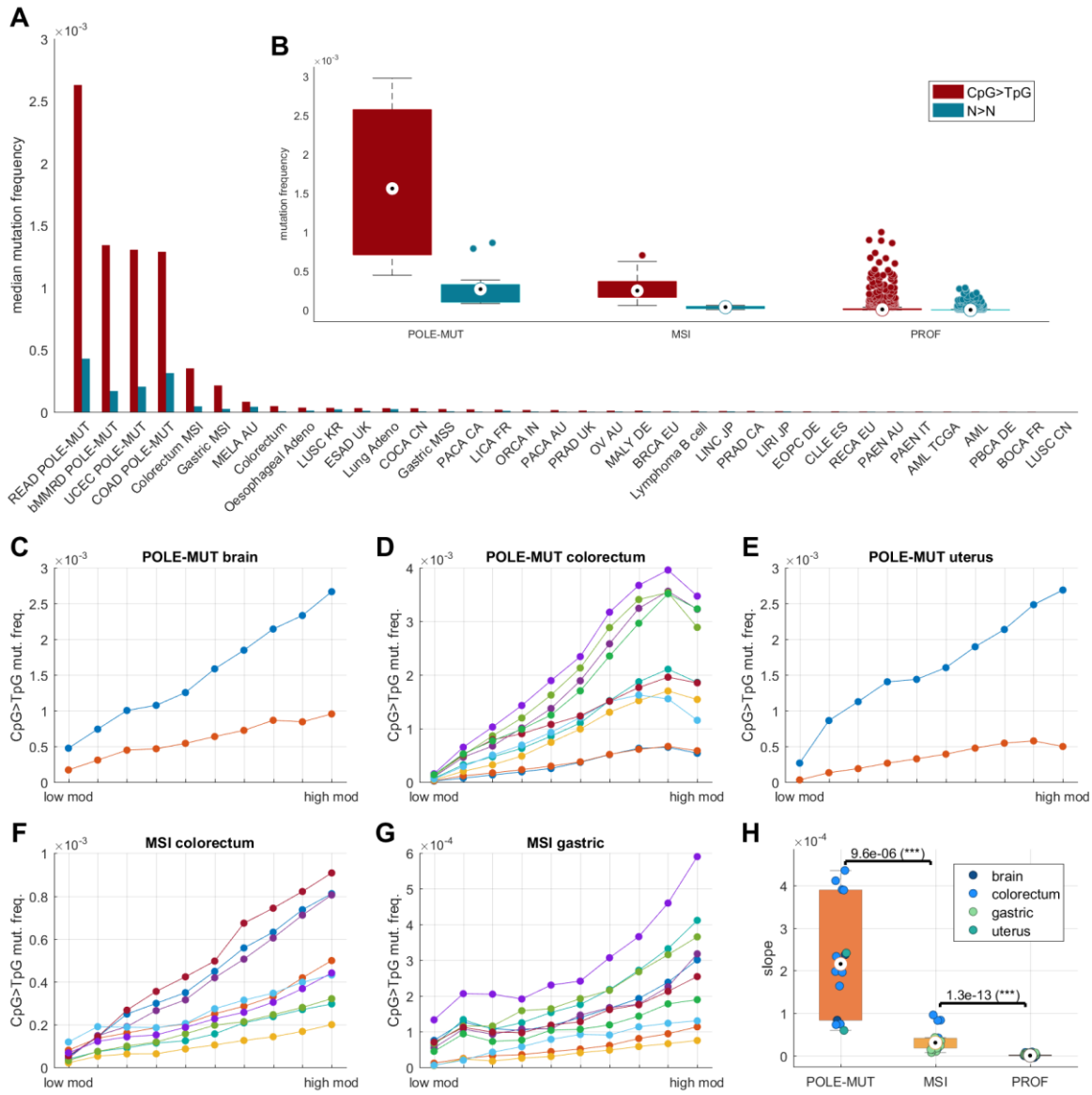


Fig. 1: Frequency of C to T mutations in a CpG context is unexpectedly high in *POLE*-MUT and MSI samples and correlates with DNA modification levels. A: Median CpG>TpG and N>N (overall) mutation frequency in each cancer type separately. **B**: Distribution of CpG>TpG and N>N mutation frequency in *POLE*-MUT, MSI, and PROF (other) samples. The white circle with the black dot inside denotes the median. **C-G**: Fraction of mutated CpG sites as a function of modification levels. The x-axis represents CpG sites grouped into 10 bins by their modification levels (0-0.1, …, 0.9-1.0). The y-axis represents C>T mutation frequency in each bin. Individual samples are plotted in different colours. **H:** Distribution of the slope of the linear relationship between DNA modification levels and CpG>TpG mutation frequency in four tissues (brain, colorectum, gastric, and uterus). The Wilcoxon ranksum test was used to evaluate differences between the groups (*POLE*-MUT, MSI, and PROF) of samples. See the distribution of offsets in Fig. 1-supplement 2.

We next used bisulfite-sequencing (BS-seq) derived DNA modification maps from normal tissue of the same organ as each cancer sample to explore whether DNA modifications play a role in the occurrence of CpG>TpG mutations in *POLE*-MUT and MSI samples. These maps represent levels of

both the more frequent 5mC as well as the less frequent 5hmC, since BS-seq alone cannot distinguish between these two modifications. In all *POLE*-MUT and MSI samples, the CpG>TpG mutation frequency was positively correlated with modification levels (Fig. 1C–G). Moreover, the slope of the correlation was significantly higher in *POLE*-MUT than in MSI, and in MSI than in tissue-matched PROF samples (Fig. 1H, 1-supplement 2). These results support the notion that the mechanism responsible for the elevated mutation rate of CpGs in *POLE*-MUT and MSI samples is linked to epigenetic DNA modifications.

It is unlikely that Pol ε or MMR, through their canonical, replication-linked activity, are used for the repair of deamination-induced T:G mismatches that happened before replication. However, it is possible that their non-canonical, replication unrelated, activity is involved in the repair of deamination induced mismatches. Conversely, the CpG>TpG mutations could be replication related, but independent of spontaneous deamination of 5mC. We therefore explored whether the CpG>TpG mutagenicity in *POLE*-MUT and MSI samples shows any replication-linked characteristics, to distinguish between the potential replication-unrelated repair of spontaneous deamination, and a – yet undescribed – replication-related source of CpG>TpG mutations.

In eukaryotic cells, DNA replication is initiated around replication origins (ORI) from where it proceeds in both directions, synthesizing the leading strand continuously and the lagging strand discontinuously. As Pol ε is the main leading strand DNA polymerase [25,26], mutations in *POLE*-MUT samples are distributed asymmetrically on the leading and lagging strands [11,17]. MSI samples also display replication strand bias across several types of mutations [17], presumably because MMR is involved in balancing the differences in fidelity of the leading and lagging polymerases [27]. In order to determine whether CpG>TpG mutations in *POLE*-MUT and MSI samples happened during or before replication, we computed the frequency of CpG>TpG mutations on the plus (Watson) and minus (Crick) strand around transitions between left- and right-replicating regions, as defined in [17]. The transitions correspond to regions enriched for replication origins.

In the *POLE*-MUT and MSI samples, we observed a strong enrichment of CpG>TpG mutations on the leading strand template (plus strand in the left direction, minus strand in the right direction) (Fig. 2). Moreover, the strand asymmetry was at least as strong or stronger in highly modified CpGs (top tertile) than in lowly modified CpGs (bottom tertile) (Fig. 2C–D). This effect was furthermore observed across cancer types and across modification levels (Fig. 2 supplement 1). It thus appears that DNA repair deficient cells accumulate more CpG>TpG mutations in cytosines that were modified on the template for the leading strand, suggesting that they are related to replication.

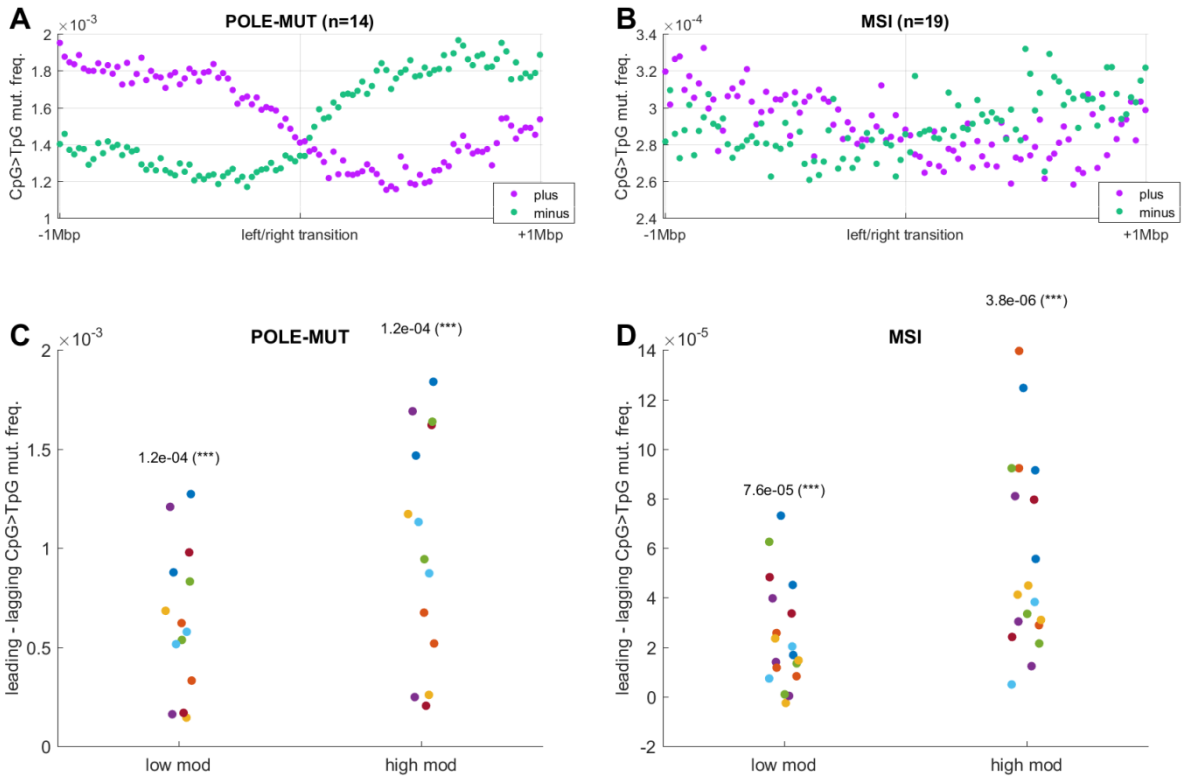


Fig. 2: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples is higher on the leading strand than on the lagging strand, especially in modified CpG sites. A-B: Mean CpG>TpG mutation frequency on the plus (Watson) and minus (Crick) strand around transitions between left- and right-replicating regions. The transitions correspond to regions enriched for replication origins. The leading strand template corresponds to the plus strand in the left direction and the minus strand in the right direction, whereas the lagging strand template corresponds to the minus strand in the left direction and the plus strand in the right direction. **C-D:** Difference in the leading and lagging CpG>TpG mutation frequency in each sample (signtest was used for evaluating significance between leading and lagging strand).

The link between C>T mutagenicity in methylated CpG sites and replication could either be a unique feature of *POLE*-MUT and MSI samples, or it could be present in all samples, but normally be suppressed by a combination of Pol ε proofreading and MMR. To explore the first option, we tested

the observed *POLE* and MMR mutations for signs of a "gain of function" mutation. A range of 9 different variants in the proofreading domain of *POLE* were present in the 14 *POLE*-MUT samples, all of them showing an increase of CpG>TpG mutations in modified cytosine (Fig. 3A). The positive correlation of CpG>TpG mutagenicity with methylation seems to be independent of the type of *POLE* mutation, cancer type or age at diagnosis, and is present in both *POLE*-MUT and MSI samples (Fig 3A). A gain-of-function mutation therefore seems unlikely.

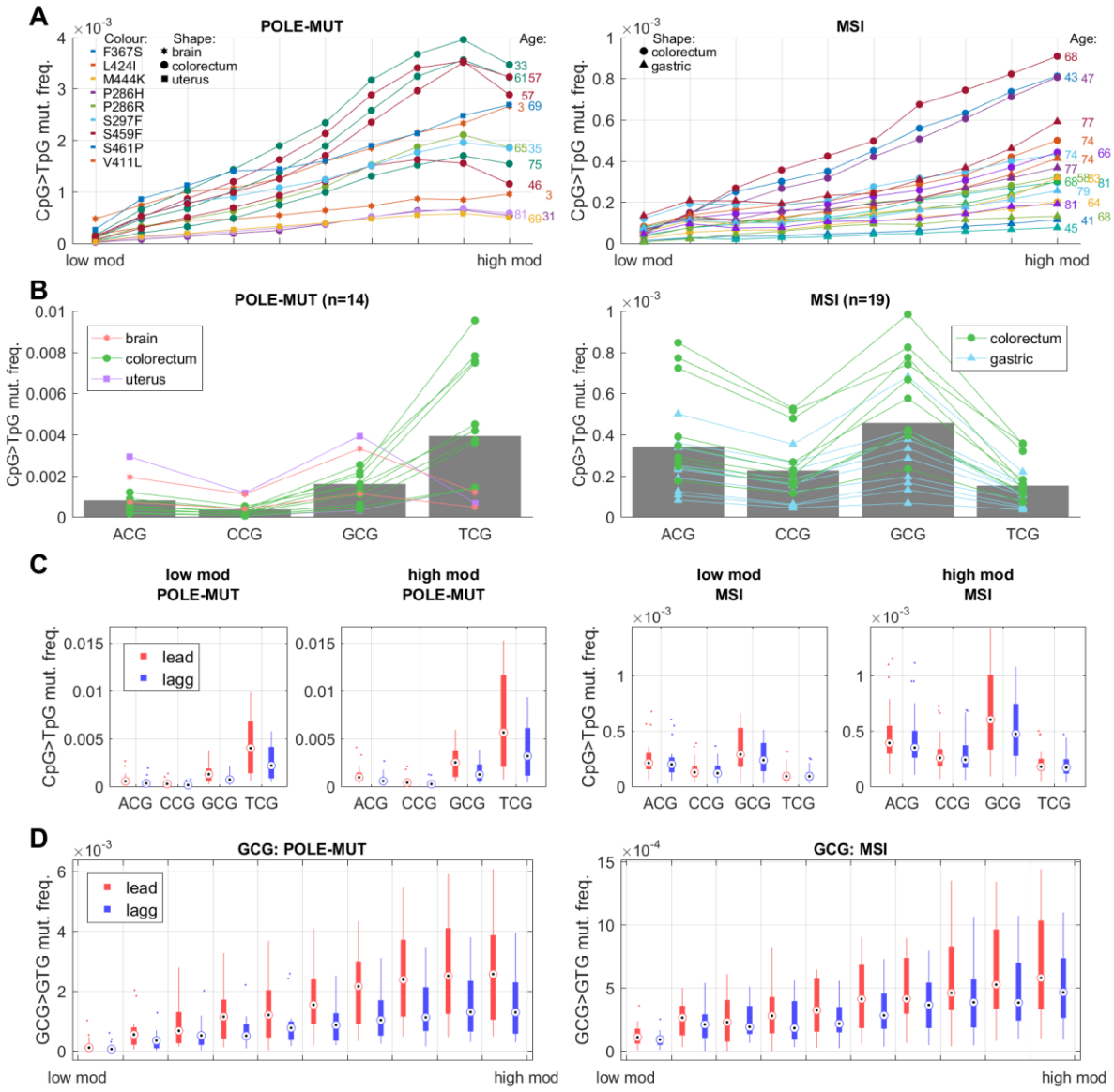


Fig. 3: Increase of C to T mutations in modified cytosine on the leading strand is most consistent in a GCG sequence context in *POLE*-MUT and MSI samples. A: C>T mutation frequency in CpG context binned by the tissue-matched modification levels (0-0.1, …, 0.9-1.0). Individual samples are plotted as separate traces. In *POLE*-MUT samples, the colour represents different variants of the *POLE* mutation. In both *POLE*-MUT and MSI samples, the shape of the marker represents different tissues. The age at diagnosis is shown next to the last value of the sample. **B:** CpG>TpG mutation

frequency stratified by the 5' flanking sequence context. The bars denote mean over samples and individual samples are shown as markers with shape and colour distinguishing the tissue type. **C:** C>T mutation frequency in CpG sites in the leading and lagging strands, in low mod (≤0.8) vs high mod (>0.95), and stratified by the 5' sequence context: ACG, CCG, GCG, and TCG. **D:** C>T mutation frequency in GCG context in leading and lagging strand binned by the tissue-matched modification levels (0-0.1, …, 0.9-1.0).

Interestingly, the frequency of C>T mutations was not only affected by the 3' sequence context, but also the 5' base of cytosine. We noticed that, while C>T mutations in a TCG context (TCG>TTG) dominate in colorectal *POLE*-MUT samples, both tissues with MSI samples and all tissues with *POLE*-MUT samples exhibited high levels of C>T mutations in a GCG context (GCG>GTG) (Fig.3B, Fig3-supplement 1). GCG>GTG mutations also showed particularly strong strand asymmetry and correlation with modification levels in all MSI and *POLE*-MUT samples (Fig. 3C, D, 3-supplement 2).

Our observations could be explained by a model of CpG>TpG mutagenesis in which 5mC is occasionally incorrectly paired with adenine by Pol ε during replication of the leading strand, potentially due to the structural similarity of 5mC and thymine. If such mismatches were not detected by the polymerase proofreading machinery or MMR, they would result in CpG>TpG mutations most frequently where 5mC occurred in the leading strand template. Under this model of decreased fidelity of wildtype Pol ε in replication of 5mC, we would expect that such errors could sometimes escape the polymerase proofreading and MMR even in *POLE*-WT and MMR proficient samples, resulting in a slight strand asymmetry of CpG>TpG mutations. To test this, we grouped PROF samples by tissue, and in each tissue measured the percentage of samples with a higher CpG>TpG mutation frequency on the leading than the lagging strand, while also distinguishing between all four sequence contexts. The majority of samples exhibited leading strand bias for GCG>GTG mutations in 13 out of 16 tissue types in lowly and intermediately modified CpGs (Fig. 4-supplement 1). This effect was even stronger (16 out of 16 tissues) when restricting the analysis to highly modified CpGs only (Fig. 4), supporting the hypothesis that CpG>TpG mutations can also be caused by errors during the replication of methylated cytosine by Pol ε.

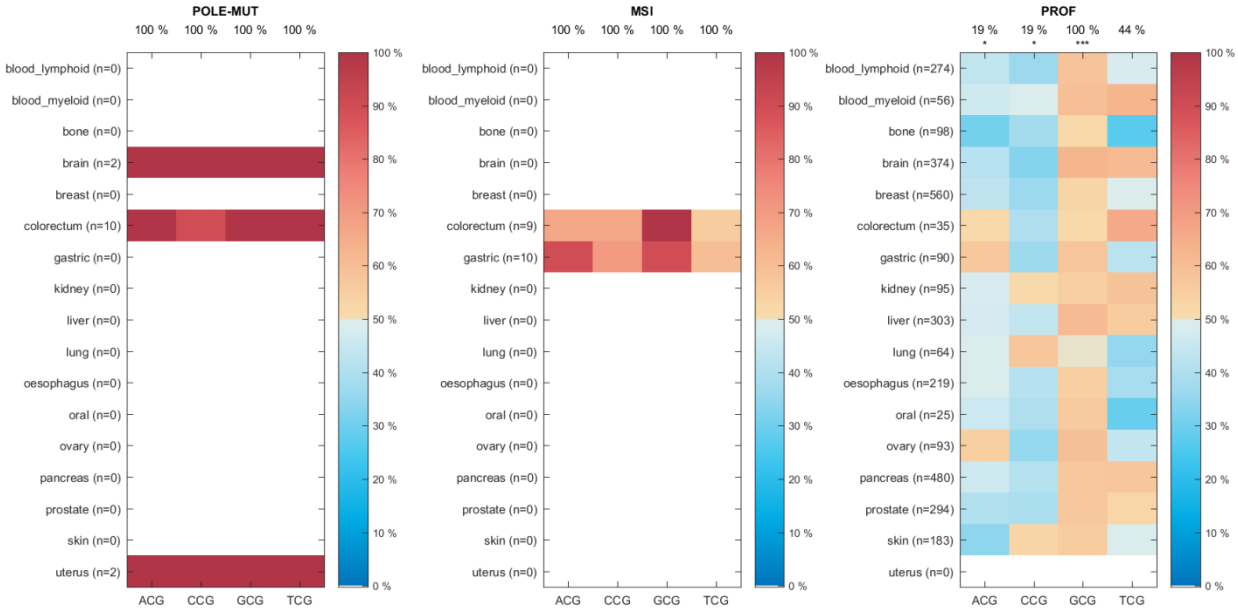


Fig. 4: GCG>GTG mutations are more frequent on the leading strand than on the lagging strand, even in Pol ε and MMR proficient samples. The heatmap shows the percentage of samples with higher C>T mutation frequency on the leading strand than on the lagging strand (only C>T mutations in highly modified (>0.95) CpG sites, using tissue-matched modification maps): white colour denotes no data, blue colour denotes more frequent lagging bias, and red denotes more frequent leading bias. The number above each column represents the percentage of cancer types with a leading strand bias in a majority of samples. Asterisks represent significance of the bias in each column (signtest; ***P < 0.001; **P < 0.01; *P < 0.05).

4. Discussion

The increased rate of C>T mutations at CpG dinucleotides across tissue types has been thought to primarily stem from spontaneous deamination of methylated cytosine. The fact that *POLE*-MUT and MSI samples exhibit high CpG>TpG mutation frequency is therefore surprising, since neither MMR nor proofreading by Pol ε are thought to be required for the repair of deamination damage. A similar increase of CpG>TpG mutations in MSI and POLE-MUT colorectal cancer samples has also been observed in another study that was published during the preparation of this manuscript [28], but the correlation of these mutations with methylation levels was not explored in much detail.

Three theoretical models could explain this observation. In the first model, MMR and Pol ε — through a non-canonical, replication-unrelated mechanism— are in fact essential for the repair of T:G mismatches created by spontaneous deamination of 5mC. For MMR, this is the model proposed

in a recent study [28]. However, the observed number of CpG>TpG mutations in MSI and *POLE*-MUT samples is difficult to reconcile with the known deamination kinetics of methylated cytosine in double-stranded DNA, even under the unrealistic assumption that no repair mechanisms at all are active in these samples. At 5.8×10^{-13} mutations per 5mC per second [24], it would take 28 years to reach the observed C>T mutation frequency in modified GCG sites of MSI samples, and 98 years for *POLE*-MUT samples (see Methods for calculations). These timescales are unlikely to represent the real time between the acquisition of the MMR or Pol ε mutation and the collection of the sample. Moreover, if spontaneous deamination was the source of CpG>TpG mutagenicity in MMR and Pol ε deficient samples, one would not expect to see replication strand asymmetry. However, CpG>TpG mutations are highly enriched on the leading strand in all these samples and therefore do not support this first model.

The second possible explanation is that the Pol ε and MMR mutations are gain of function mutations, causing a mutator phenotype that actively increases CpG>TpG mutagenicity during replication. This mechanism has been suggested by Poulos et al. [28] for the POLE-MUT samples and by Kane et al. [29] in *S. cerevisiae*, where an analog of the human P286R variant (but not other variants) in the yeast Pol ε produced a strong mutator phenotype, increasing the mutation rate beyond that of the proofreading-null allele. However, we observed a marked increase of C>T mutation frequency in modified CpG sites in a wide range of Pol ε variants (Fig. 3A). Furthermore, a strong correlation of GCG>GTG mutations with DNA modification levels was observed across *POLE*-MUT and MSI samples from multiple cancer types. It therefore seems unlikely that multiple different Pol ε and MMR mutations all result in the same mutator phenotype.

The third model posits that wildtype Pol ε has a slightly decreased fidelity when encountering 5mC, particularly in a GCG context, on the template strand and incorrectly pairs it with A, leading to 5mC:A mismatches. This could potentially be a consequence of the high structural similarity between 5mC and T, both of which present a methyl group at the same position of pyrimidine ring. If the resulting 5mC:A mismatches were not repaired before the next round of replication, for example

because of a lack of mismatch repair in MSI tumours, one would expect an enrichment of GCG>GTG mutations on the leading strand, as we observe in our data. Similarly, a lack of proofreading by Pol ε itself might overwhelm the capacity of downstream repair pathways and thus, too, lead to an increased CpG>TpG mutations rate. The fact that we also detected a leading strand bias for GCG>GTG mutations in a majority of Pol ε and MMR proficient tumours hints at the possibility that the mechanism described above does contribute to the overall CpG>TpG mutation burden. This model is also consistent with observations from samples with a mutation in the proofreading domain of *POLD1*, a gene encoding the catalytic subunit of Pol δ. *POLD1*-MUT samples are also highly mutated, but, unlike in POLE-MUT samples, CpG>TpG mutations form only a small percentage of the mutation burden [12]. This observation supports the notion that the CpG>TpG mutagenesis is specifically linked to the leading strand synthesis.

5. CONCLUSIONS

To conclude, we have presented evidence suggesting that replication of methylated cytosines is likely to contribute to the higher mutation rate of CpGs in the genome. This unanticipated finding changes the commonly accepted paradigm in the field, where spontaneous deamination has been proposed as the only reason for the mutagenicity of methylated CpG sites. While replication-linked CpG>TpG mutations dominate in Pol ε mutated or MMR deficient cells, the relative contribution of replication-linked mutations compared to deamination-induced mutations in repair-proficient cells is less clear. Pol ε proofreading and MMR both repair mutations originating during replication, while *MBD4* and *TDG* are glycosylases repairing lesions caused by spontaneous deamination of 5mC. Pol ε mutations increase CpG mutation rate by 210-fold in human cancers, while Mbd4 deficient mice exhibit an increase in mutation frequency by 3-fold [30], suggesting that replication might be more mutagenic at methylated CpGs than deamination, unless *TDG* plays a dominant role in repair of deamination lesions. Thus, Pol ε might even be the primary source of C>T mutations in methylated CpGs, which could also explain that cancers from tissues with higher turnover rates exhibit an

increased rate of CpG>TpG mutations [31]. Further experimental work will be required to fully elucidate the fidelity of Pol ε when replicating 5mC.

SUPPORTING INFORMATION

Fig. 1-supplement 1: Frequency of C to T mutations in a CpG context is unexpectedly high in *POLE*-MUT and MSI samples.

Fig. 1-supplement 2: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples correlates with DNA modification levels: comparison of linear models.

Fig. 2-supplement 1: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples is higher on the leading strand than on the lagging strand, especially in modified CpG sites.

Fig. 3-supplement 1: CpG>TpG mutation frequency in different sequence contexts.

Fig. 3-supplement 2: Increase of C to T mutations in modified cytosine on the leading strand is most consistent in a GCG sequence context in *POLE*-MUT and MSI samples.

Fig. 4-supplement 1: GCG>GTG mutations are more frequent on the leading strand than on the lagging strand, even in Pol ε and MMR proficient samples.

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

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

ACKNOWLEDGMENTS

We thank Dr. Mary Muers and Jakub Tomek for comments on the manuscript. 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 grant EP/F500394/1 and the Bakala Foundation.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest

REFERENCES

[1] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A.J.R. Aparicio, S. Behjati, A. V Biankin, G.R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A.P. Butler, C. Caldas, H.R. Davies, C. Desmedt, R. Eils, J.E. Eyfjörd, J.A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T.

Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger, D.T.W. Jones, D. Jones, S. Knappskog, M. Kool, S.R. Lakhani, C. López-Otín, S. Martin, N.C. Munshi, H. Nakamura, P.A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V Pearson, X.S. Puente, K. Raine, M. Ramakrishna, A.L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T.N. Schumacher, P.N. Span, J.W. Teague, Y. Totoki, A.N.J. Tutt, R. Valdés-Mas, M.M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L.R. Yates, J. Zucman-Rossi, P.A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S.M. Grimmond, R. Siebert, E. Campo, T. Shibata, S.M. Pfister, P.J. Campbell, M.R. Stratton, Signatures of mutational processes in human cancer., Nature. 500 (2013) 415–21. doi:10.1038/nature12477.

[2] M.S. Lawrence, P. Stojanov, P. Polak, G. V Kryukov, K. Cibulskis, A. Sivachenko, S.L. Carter, C. Stewart, C.H. Mermel, S.A. Roberts, A. Kiezun, P.S. Hammerman, A. McKenna, Y. Drier, L. Zou, A.H. Ramos, T.J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M.L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D.I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A.M. Dulak, J. Lohr, D.-A. Landau, C.J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S.A. McCarroll, J. Mora, R.S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S.B. Gabriel, C.W.M. Roberts, J.A. Biegel, K. Stegmaier, A.J. Bass, L.A. Garraway, M. Meyerson, T.R. Golub, D.A. Gordenin, S. Sunyaev, E.S. Lander, G. Getz, Mutational heterogeneity in cancer and the search for new cancer-associated genes., Nature. 499 (2013) 214–8. doi:10.1038/nature12213.

[3] A.P. Bird, M.H. Taggart, Variable patterns of total DNA and rDNA methylation in animals, Nucleic Acids Research. 8 (1980) 1485–1497. doi:10.1093/nar/8.7.1485.

[4] M. Tomkova, M. McClellan, S. Kriaucionis, B. Schuster-Boeckler, 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA, eLife. 5 (2016) 1–23. doi:10.7554/eLife.17082.

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

[6] K. Wiebauer, J. Jiricny, In vitro correction of GT mispairs to GC pairs in nuclear extracts from human cells, Nature. 339 (1989) 234–236. doi:10.1038/339234a0.

[7] B. Hendrich, U. Hardeland, H. Ng, J. Jiricny, A. Bird, The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites, Nature. 401 (1999) 525–525. doi:10.1038/35006691.

[8] A. Bellacosa, A.C. Drohat, Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites, DNA Repair. 32 (2015) 33–42. doi:10.1016/j.dnarep.2015.04.011.

[9] T. Pfaffeneder, F. Spada, M. Wagner, C. Brandmayr, S.K. Laube, D. Eisen, M. Truss, J. Steinbacher, B. Hackner, O. Kotljarova, D. Schuermann, S. Michalakis, O. Kosmatchev, S. Schiesser, B. Steigenberger, N. Raddaoui, G. Kashiwazaki, U. Müller, C.G. Spruijt, M. Vermeulen, H. Leonhardt, P. Schär, M. Müller, T. Carell, Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA., Nature Chemical Biology. 10 (2014) 574–81. doi:10.1038/nchembio.1532.

[10] E. Rayner, I.C. van Gool, C. Palles, S.E. Kearsey, T. Bosse, I. Tomlinson, D.N. Church, A panoply of errors: polymerase proofreading domain mutations in cancer, Nature Reviews Cancer. 16 (2016) 71–81. doi:10.1038/nrc.2015.12.

[11] E. Shinbrot, E.E. Henninger, N. Weinhold, K.R. Covington, A.Y. Göksenin, N. Schultz, H. Chao, H. Doddapaneni, D.M. Muzny, R.A. Gibbs, C. Sander, Z.F. Pursell, D.A. Wheeler, Exonuclease mutations in DNA Polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication., Genome Research. (2014) 1740–1750. doi:10.1101/gr.174789.114.

[12] A. Shlien, B.B. Campbell, R. de Borja, L.B. Alexandrov, D. Merico, D. Wedge, P. Van Loo, P.S. Tarpey, P. Coupland, S. Behjati, A. Pollett, T. Lipman, A. Heidari, S. Deshmukh, N. Avitzur, B. Meier, M. Gerstung, Y. Hong, D.M. Merino, M. Ramakrishna, M. Remke, R. Arnold, G.B. Panigrahi, N.P. Thakkar, K.P. Hodel, E.E. Henninger, A.Y. Göksenin, D. Bakry, G.S. Charames, H. Druker, J. Lerner-Ellis, M. Mistry, R. Dvir, R. Grant, R. Elhasid, R. Farah, G.P. Taylor, P.C. Nathan, S. Alexander, S. Ben-Shachar, S.C. Ling, S. Gallinger, S. Constantini, P. Dirks, A. Huang, S.W. Scherer, R.G. Grundy, C. Durno, M. Aronson, A. Gartner, M.S. Meyn, M.D. Taylor, Z.F. Pursell, C.E. Pearson, D. Malkin, P.A. Futreal, M.R. Stratton, E. Bouffet, C. Hawkins, P.J. Campbell, U. Tabori, Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers, Nature Genetics. 47 (2015) 257–262. doi:10.1038/ng.3202.

[13] T.M. Albertson, M. Ogawa, J.M. Bugni, L.E. Hays, Y. Chen, Y. Wang, P.M. Treuting, J.A. Heddle, R.E. Goldsby, B.D. Preston, DNA polymerase epsilon and delta proofreading suppress discrete mutator and cancer phenotypes in mice., Proceedings of the National Academy of Sciences of the United States of America. 106 (2009) 17101–4. doi:10.1073/pnas.0907147106.

[14] A.J. Bass, M.S. Lawrence, L.E. Brace, A.H. Ramos, Y. Drier, K. Cibulskis, C. Sougnez, D. Voet, G. Saksena, A. Sivachenko, R. Jing, M. Parkin, T. Pugh, R.G. Verhaak, N. Stransky, A.T. Boutin, J. Barretina, D.B. Solit, E. Vakiani, W. Shao, Y. Mishina, M. Warmuth, J. Jimenez, D.Y. Chiang, S. Signoretti, W.G. Kaelin, N. Spardy, W.C. Hahn, Y. Hoshida, S. Ogino, R.A. Depinho, L. Chin, L.A. Garraway, C.S. Fuchs, J. Baselga, J. Tabernero, S. Gabriel, E.S. Lander, G. Getz, M. Meyerson, Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion., Nature Genetics. 43 (2011) 964–8. doi:10.1038/ng.936.

[15] A.M. Dulak, P. Stojanov, S. Peng, M.S. Lawrence, C. Fox, C. Stewart, S. Bandla, Y. Imamura, S.E. Schumacher, E. Shefler, A. McKenna, S.L. Carter, K. Cibulskis, A. Sivachenko, G. Saksena, D. Voet, A.H. Ramos, D. Auclair, K. Thompson, C. Sougnez, R.C. Onofrio, C. Guiducci, R. Beroukhim, Z. Zhou, L. Lin, J. Lin, R. Reddy, A. Chang, R. Landrenau, A. Pennathur, S. Ogino,

J.D. Luketich, T.R. Golub, S.B. Gabriel, E.S. Lander, D.G. Beer, T.E. Godfrey, G. Getz, A.J. Bass, Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity., Nature Genetics. 45 (2013) 478–86. doi:10.1038/ng.2591.

[16] K. Wang, S.T. Yuen, J. Xu, S.P. Lee, H.H.N. Yan, S.T. Shi, H.C. Siu, S. Deng, K.M. Chu, S. Law, K.H. Chan, A.S.Y. Chan, W.Y. Tsui, S.L. Ho, A.K.W. Chan, J.L.K. Man, V. Foglizzo, M.K. Ng, A.S. Chan, Y.P. Ching, G.H.W. Cheng, T. Xie, J. Fernandez, V.S.W. Li, H. Clevers, P.A. Rejto, M. Mao, S.Y. Leung, Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer., Nature Genetics. 46 (2014) 573–82. doi:10.1038/ng.2983.

[17] N.J. Haradhvala, P. Polak, P. Stojanov, K.R. Covington, E. Shinbrot, J.M. Hess, E. Rheinbay, J. Kim, Y.E. Maruvka, L.Z. Braunstein, A. Kamburov, P.C. Hanawalt, D.A. Wheeler, A. Koren, M.S. Lawrence, G. Getz, Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair, Cell. 164 (2016) 538–549. doi:10.1016/j.cell.2015.12.050.

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

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

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

[21] R. Pidsley, E. Zotenko, T.J. Peters, M.G. Lawrence, G.P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, S.J. Clark, P. Jones, S. Baylin, Y. Ko, D. Mohtat, M. Suzuki, A. Park,

M. Izquierdo, S. Han, T. Dayeh, P. Volkov, S. Salo, E. Hall, E. Nilsson, A. Olsson, R. Pidsley, J. Viana, E. Hannon, H. Spiers, C. Troakes, S. Al-Saraj, C. Stirzaker, P. Taberlay, A. Statham, S. Clark, S. Clark, J. Harrison, C. Paul, M. Frommer, R. Lister, M. Pelizzola, R. Dowen, R. Hawkins, G. Hon, J. Tonti-Filippini, M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen, T. Hinoue, D. Weisenberger, C. Lange, H. Shen, H. Byun, D. Berg, L. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, V. Rakyan, T. Down, S. Maslau, T. Andrew, T. Yang, H. Beyan, M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. Le, T. Morris, S. Beck, Y. Chen, S. Choufani, D. Grafodatskaya, D. Butcher, J. Ferreira, R. Weksberg, Y. Chen, M. Lemire, S. Choufani, D. Butcher, D. Grafodatskaya, B. Zanke, H. Naeem, N. Wong, Z. Chatterton, M. Hong, J. Pedersen, N. Corcoran, T. Peters, M. Buckley, A. Statham, R. Pidsley, K. Samaras, R. V Lord, D. Wang, L. Yan, Q. Hu, L. Sucheston, M. Higgins, C. Ambrosone, C. Warden, H. Lee, J. Tompkins, X. Li, C. Wang, A. Riggs, M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, L. Siggens, K. Ekwall, S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, F. Fuks, R. Pidsley, Y.W. CC, M. Volta, K. Lunnon, J. Mill, L. Schalkwyk, A. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, N. Touleimat, J. Tost, R. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. Maurano, E. Haugen, R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, M. Ritchie, B. Phipson, D. Wu, Y. Hu, C. Law, W. Shi, M. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, M. Ziller, H. Gu, F. Müller, J. Donaghey, L.-Y. Tsai, O. Kohlbacher, S. Huang, B. Bao, T. Hour, C. Huang, C. Yu, C. Liu, S. Neuhausen, M. Slattery, C. Garner, Y. Ding, M. Hoffman, A. Brothman, R. Reams, K. Kalari, H. Wang, F. Odedina, K. Soliman, C. Yates, J. Song, C. Stirzaker, J. Harrison, J. Melki, S. Clark, M. Coolen, C. Stirzaker, J. Song, A. Statham, Z. Kassir, C. Moreno, M. Makrides, R. Gibson, A. McPhee, L. Yelland, J. Quinlivan, P. Ryan, M. Lawrence, R. Taylor, R. Toivanen, J. Pedersen, S. Norden, D. Pook, S. Clark, A. Statham, C. Stirzaker, P. Molloy, M. Frommer, A. Auton, L. Brooks, R. Durbin, E. Garrison, H. Kang, W. Kent, Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling, Genome Biology. 17 (2016) 208.

doi:10.1186/s13059-016-1066-1.

[22] A.R. Vandiver, R.A. Irizarry, K.D. Hansen, L.A. Garza, A. Runarsson, X. Li, A.L. Chien, T.S. Wang, S.G. Leung, S. Kang, A.P. Feinberg, Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin., Genome Biology. 16 (2015) 80. doi:10.1186/s13059-015-0644-y.

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

[24] J.-C. Shen, W.M. Rideout, P.A. Jones, The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA, Nucleic Acids Research. 22 (1994) 972–976. doi:10.1093/nar/22.6.972.

[25] B. Stillman, DNA Polymerases at the Replication Fork in Eukaryotes, Molecular Cell. 30 (2008) 259–260. doi:10.1016/j.molcel.2008.04.011.

[26] R.E. Georgescu, G.D. Schauer, N.Y. Yao, L.D. Langston, O. Yurieva, D. Zhang, J. Finkelstein, M.E. O'Donnell, Reconstitution of a eukaryotic replisome reveals suppression mechanisms that define leading/lagging strand operation, eLife. 2015 (2015) 1–20. doi:10.7554/eLife.04988.

[27] S.A. Lujan, J.S. Williams, Z.F. Pursell, A.A. Abdulovic-Cui, A.B. Clark, S.A. Nick McElhinny, T.A. Kunkel, Mismatch Repair Balances Leading and Lagging Strand DNA Replication Fidelity, PLoS Genetics. 8 (2012) e1003016. doi:10.1371/journal.pgen.1003016.

[28] R.C. Poulos, J. Olivier, J.W.H. Wong, The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes, Nucleic Acids Research. (2017) 1–10. doi:10.1093/nar/gkx463.

[29] D.P. Kane, P. V. Shcherbakova, A common cancer-associated DNA polymerase ε mutation causes an exceptionally strong mutator phenotype, indicating fidelity defects distinct from

loss of proofreading, Cancer Research. 74 (2014) 1895–1901. doi:10.1158/0008-5472.CAN-13-2892.

[30] C.B. Millar, Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice, Science. 297 (2002) 403–405. doi:10.1126/science.1073354.

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

6. SUPPLEMENTARY FIGURES

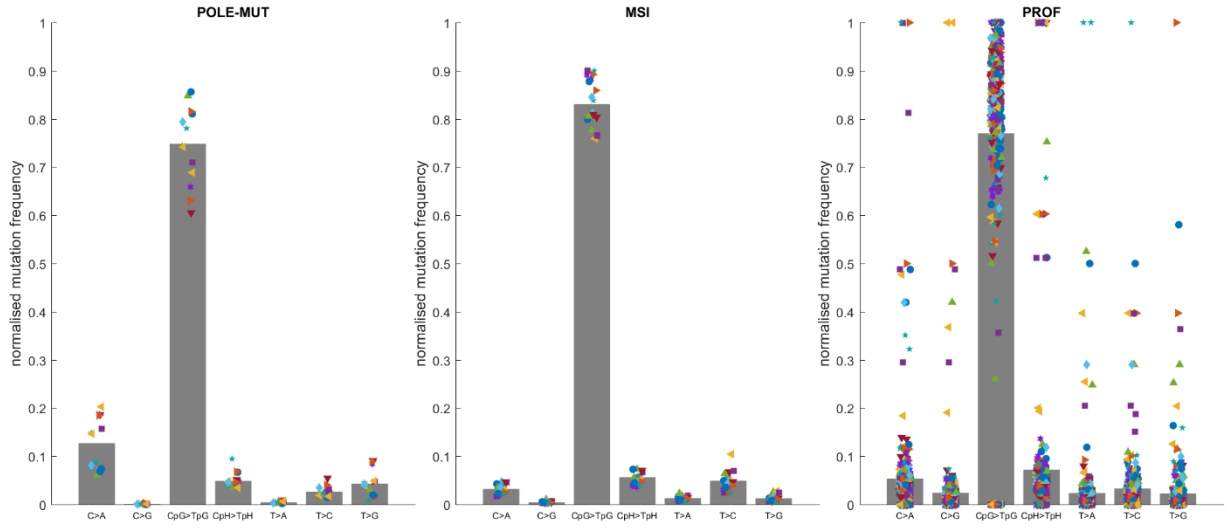
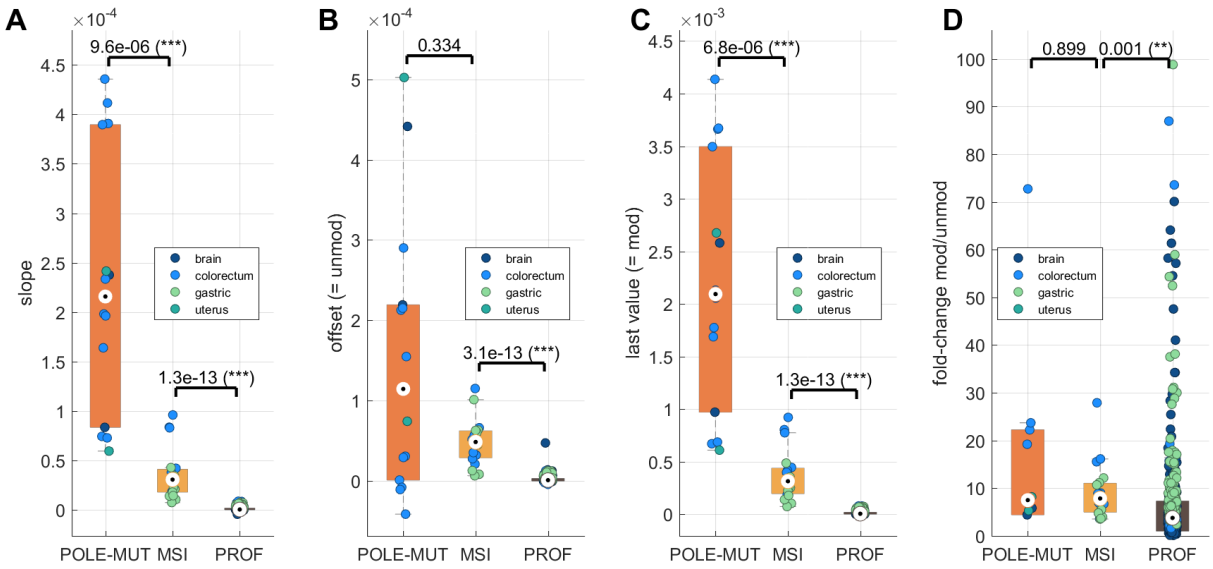
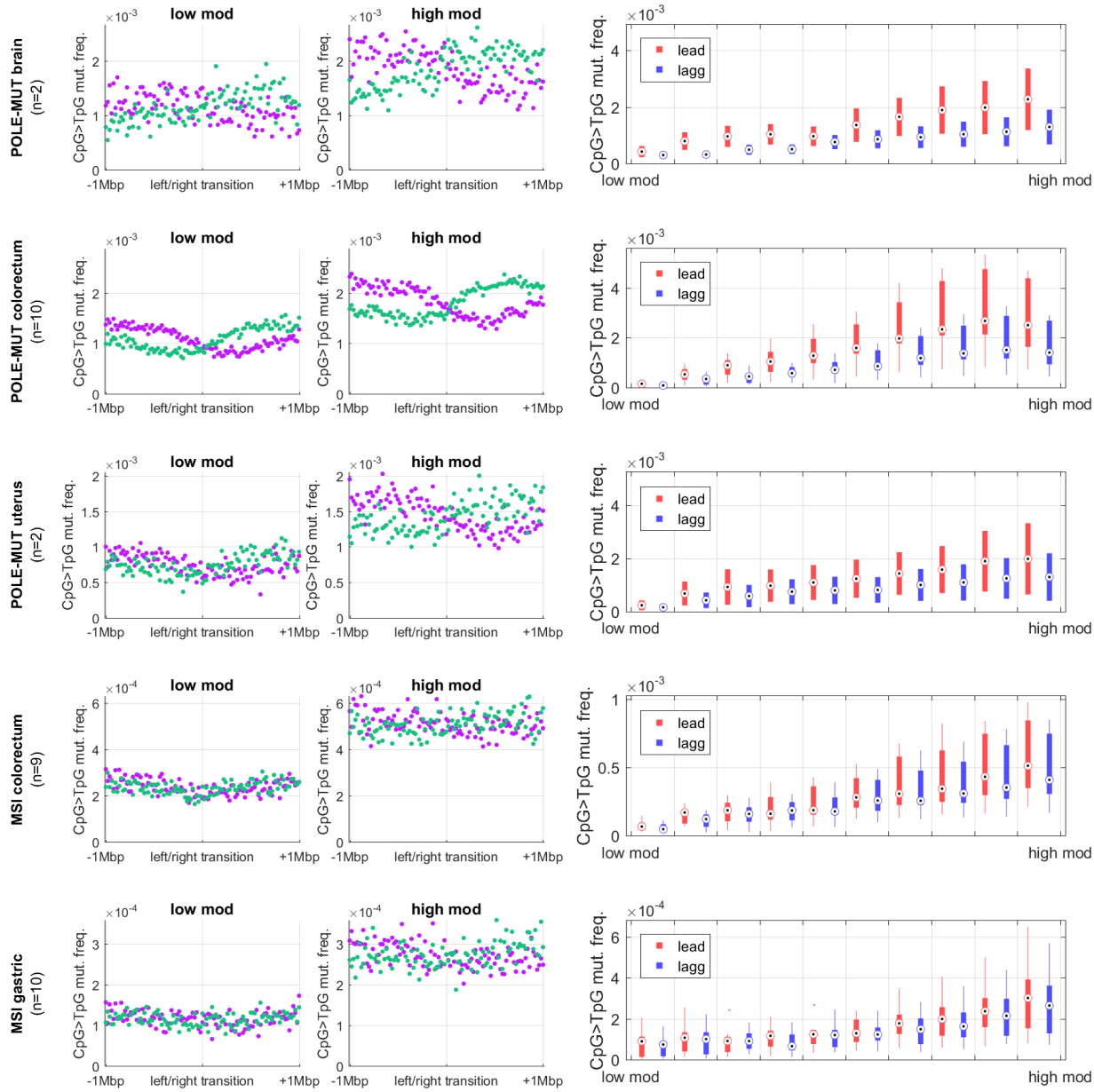


Fig. 1-supplement 1: Frequency of C to T mutations in a CpG context is unexpectedly high in *POLE*-MUT and MSI samples. Frequency of individual types of mutations in *POLE*-MUT, MSI, and tissue-matched PROF samples, normalised by the total sum in each sample. The bars denote mean over samples and individual samples are shown as markers in different shapes and colours.



503

504 **Fig. 1-supplement 2: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI**
505 **samples correlates with DNA modification levels: comparison of linear models.** In each sample, a
506 linear model was fitted on the data, representing CpG>TpG mutation frequency in different bins of
507 cytosine modification levels. The distribution of their parameters is compared: slope (**A**), offset, *i.e.*,
508 the value in unmodified cytosines (**B**), the last values, *i.e.*, the value in fully modified cytosines (**C**),
509 the fold-change from unmodified to fully modified cytosines (**D**) in MSI, *POLE*, and PROF samples in
510 four tissues (brain, colorectum, gastric, and uterus). The Wilcoxon ranksum test was used to
511 evaluate differences between the groups of samples.



512

513 **Fig. 2-supplement 1: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI**
514 **samples is higher on the leading strand than on the lagging strand, especially in modified CpG**
515 **sites. Left column:** Mean CpG>TpG mutation frequency on the plus (Watson) and minus (Crick)
516 strand around transitions between left- and right-replicating regions. The transitions correspond to
517 regions enriched for replication origins. Comparison of CpG sites with low modification levels (≤0.8)
518 and high modification levels (>0.95) is shown. Note the variation in the number of samples per
519 cohort (between 2 and 10). **Right column:** C>T mutation frequency in CpG sites in the leading and
520 lagging strand binned by their tissue-matched modification levels (0-0.1, …, 0.9-1.0).

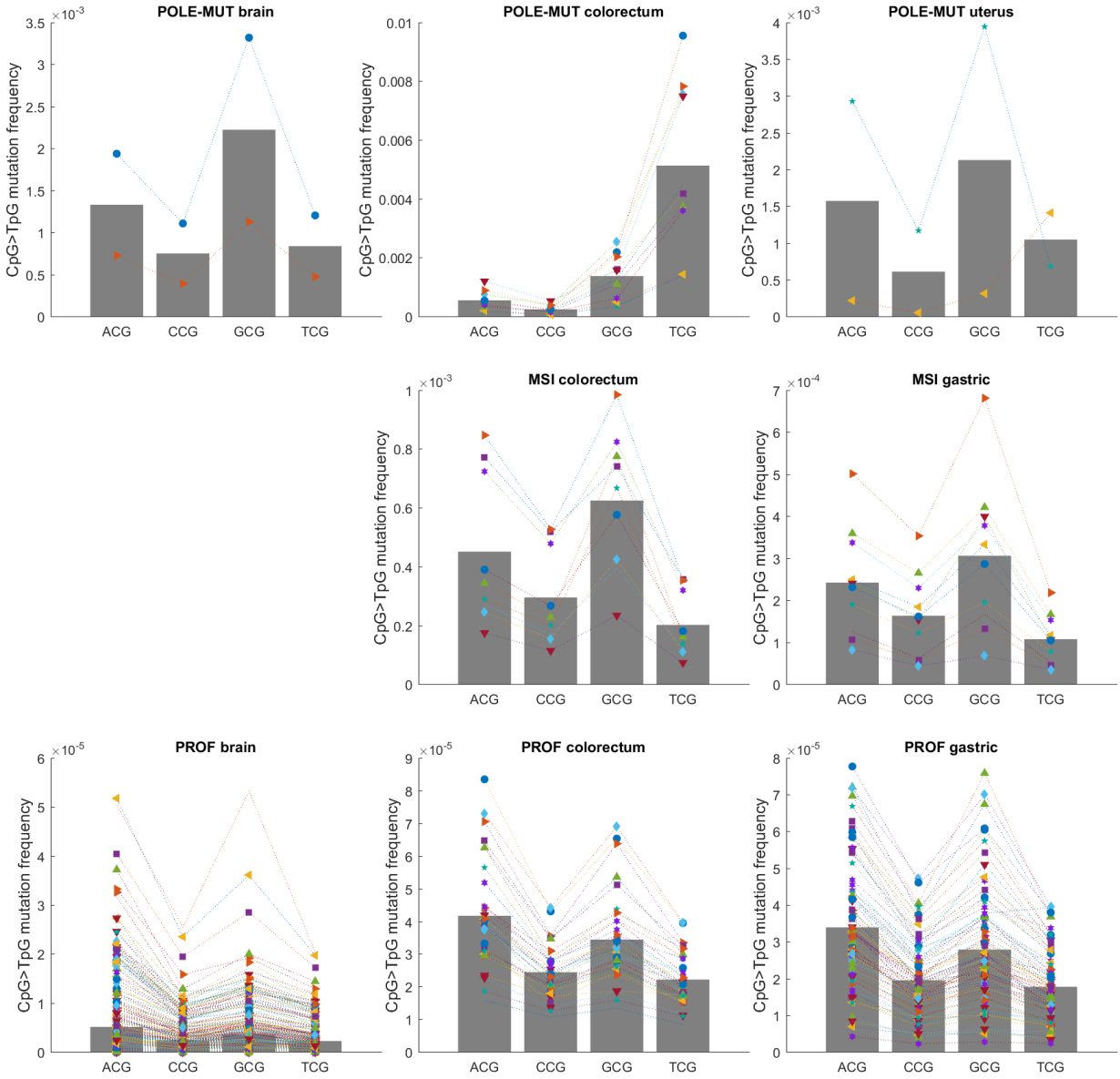
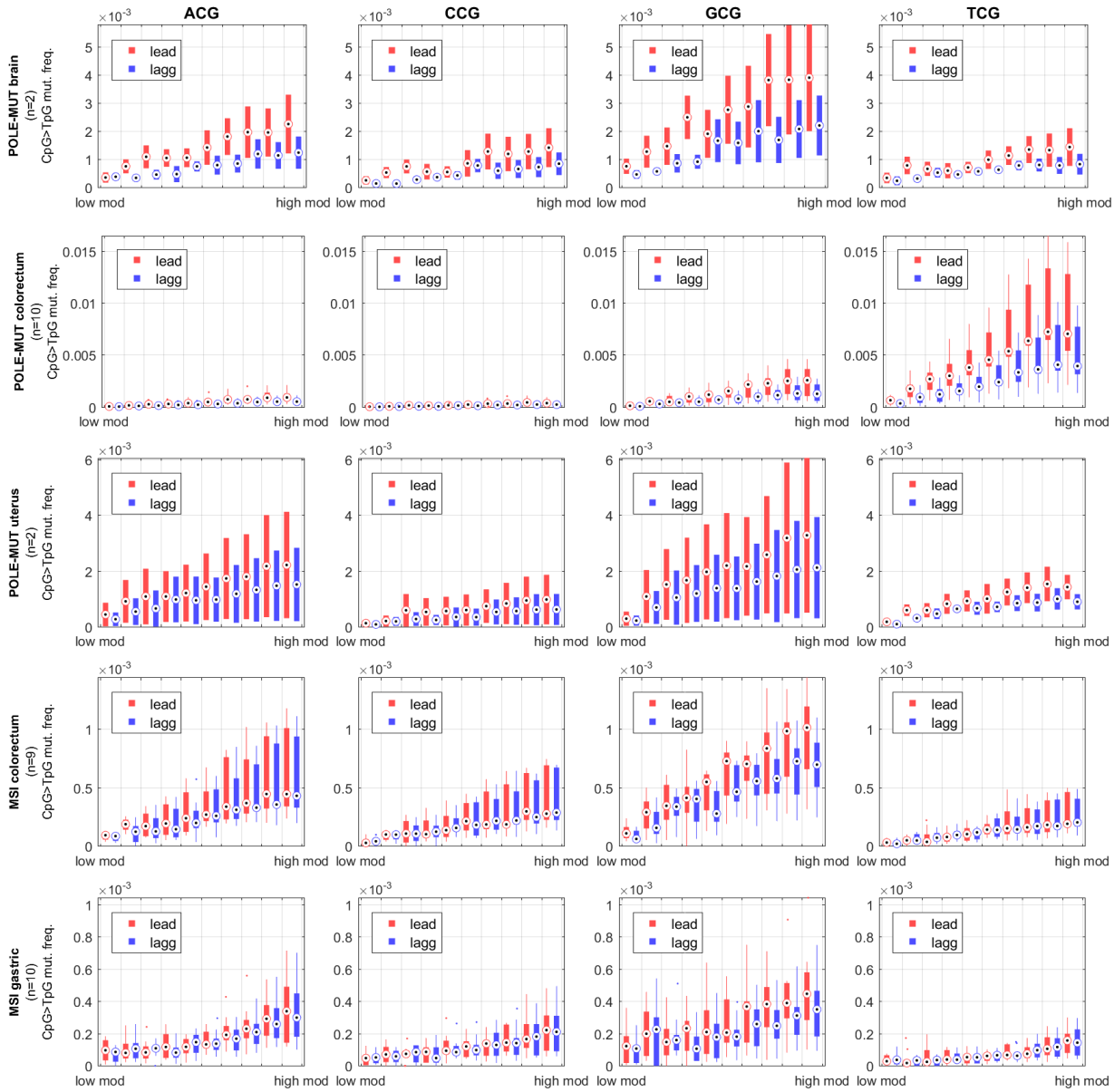
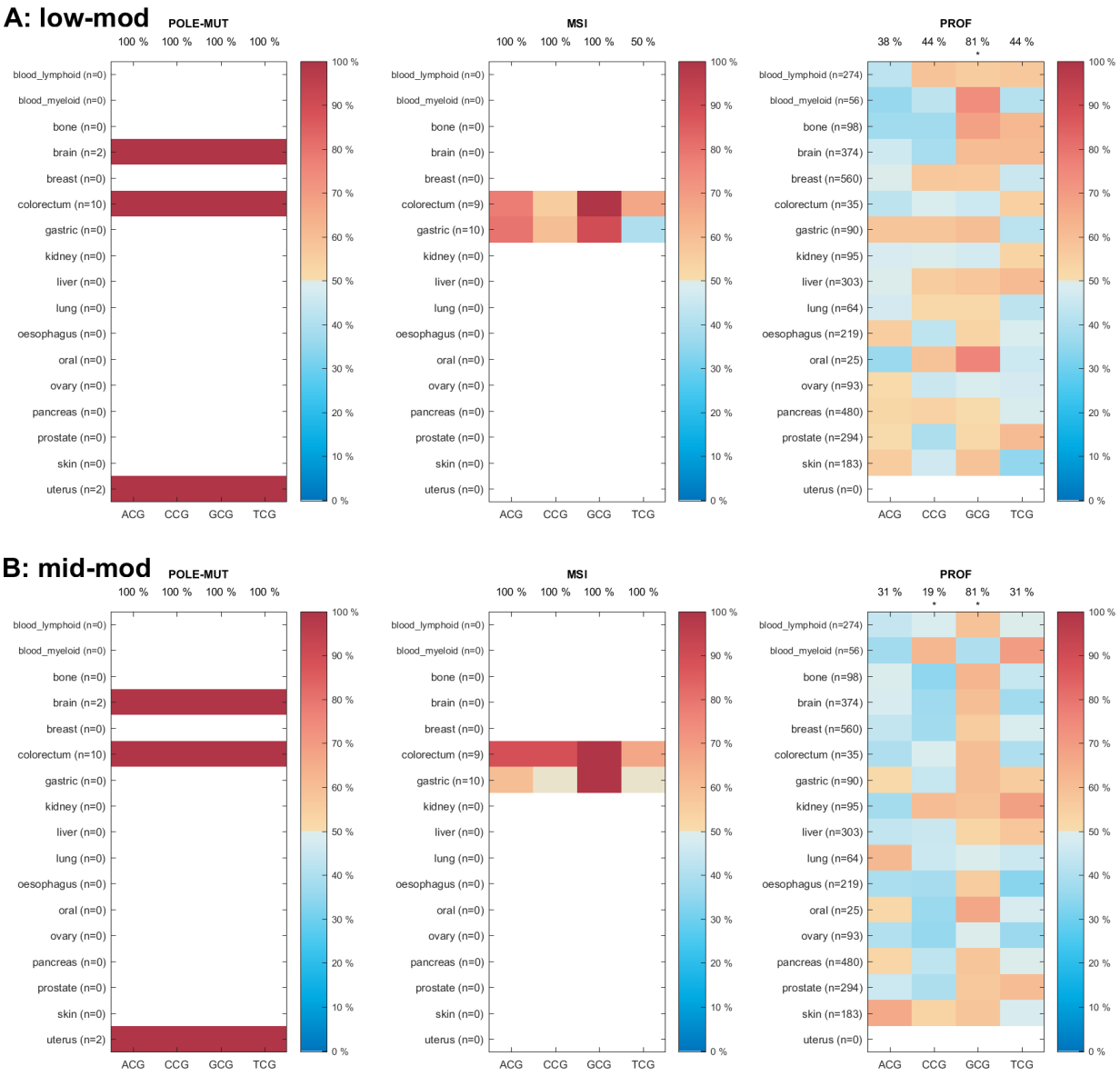


Fig. 3-supplement 1: CpG>TpG mutation frequency in different sequence contexts. CpG>TpG mutation frequency stratified by the 5' flanking sequence context and tissue type. The bars denote mean over samples and individual samples are plotted in different colours and markers.



525

526 **Fig. 3-supplement 2: Increase of C to T mutations in modified cytosine on the leading strand is**
527 **most consistent in a GCG sequence context in *POLE*-MUT and MSI samples.** C>T mutation
528 frequency in CpG sites in leading and lagging strand binned by their tissue-matched modification
529 levels (0-0.1, 0.1-0.2, …, 0.9-1.0) and sequence context: ACG (first column), CCG (second column),
530 GCG (third column), and TCG (fourth column).



531

532 **Fig. 4-supplement 1: GCG>GTG mutations are more frequent on the leading strand than on the**
533 **lagging strand, even in Pol ε and MMR proficient samples.** Percentage of samples with higher C>T
534 mutation frequency on the leading strand than on the lagging strand for CpG sites with low (≤0.8)
535 modification levels (A), and for sites with intermediate (between 0.8 and 0.95) modification levels
536 (B), using tissue-matched modification maps. White colour denotes no data, blue colour denotes
537 more frequent lagging strand bias, and red denotes more frequent leading strand bias. Asterisks
538 represent significance of the bias (signtest; ***P < 0.001; **P < 0.01; *P < 0.05).

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

Tissue	Method	Source	Link
blood lymphoid	BS-Seq	Blueprint	FTP
blood myeloid	BS-Seq	Blueprint	FTP
bone	BS-Seq	Blueprint	FTP
brain	BS-Seq	(Wen *et al.*, 2014)	SRR847423, SRR847424
brain	TAB-Seq	(Wen *et al.*, 2014)	SRR847425, SRR847426, SRR847427, SRR847428
breast	BS-Seq	Epigenome Roadmap	FTP
colorectum	BS-Seq	TCGA	TCGA-AA-3518-11A-01D-1518-05
gastric	BS-Seq	Epigenome Roadmap	FTP
kidney	BS-Seq	(Chen *et al.*, 2015)	SRR1654399, SRR1654400, SRR1654401
liver	BS-Seq	Epigenome Roadmap	FTP
lung	BS-Seq	Epigenome Roadmap	FTP
oesophagus	BS-Seq	Epigenome Roadmap	FTP
oral	BS-Seq	Blueprint	FTP
ovary	BS-Seq	Epigenome Roadmap	FTP
pancreas	BS-Seq	Epigenome Roadmap	FTP
prostate	BS-Seq	(Pidsley *et al.*, 2016)	FTP
skin	BS-Seq	(Vandiver *et al.*, 2015)	SRR1042910
uterus	BS-Seq	TCGA	TCGA-AX-A1CI-11A-11D-A17H-05

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

Cohort	Cancer type	samples	Source
Alexandrov_Ding_AML	Blood myeloid	7	(Alexandrov *et al.*, 2013)
Alexandrov_Imielinski_Lung_Adeno	Lung adenocarcinoma	24	(Alexandrov *et al.*, 2013)
Alexandrov_Lymphoma_B_cell	Blood lymphoid	24	(Alexandrov *et al.*, 2013)
Bass_Colon	Colorectum	9	(Bass *et al.*, 2011)
bMMRD	POLE-MUT brain	2	(Shlien *et al.*, 2015)
Dulak_Oesophagus	Oesophageal adenocarcinoma	16	(Dulak *et al.*, 2013)
ICGC_BOCA_FR	Bone	98	ICGC
ICGC_BRCA_EU	Breast	560	ICGC
ICGC_CLLE_ES	Blood lymphoid	152	ICGC
ICGC_COCA_CN	Colorectum	26	ICGC
ICGC_EOPC_DE	Prostate	62	ICGC
ICGC_ESAD_UK	Oesophagus adenocarcinoma	213	ICGC
ICGC_LICA_FR	Liver	14	ICGC
ICGC_LINC_JP	Liver	31	ICGC
ICGC_LIRI_JP	Liver	283	ICGC
ICGC_LUSC_CN	Lung squamous	10	ICGC
ICGC_LUSC_KR	Lung squamous	30	ICGC
ICGC_MALY_DE	Blood lymphoid	100	ICGC
ICGC_MELA_AU	Skin	199	ICGC

ICGC_ORCA_IN	Oral	25	ICGC
ICGC_OV_AU	Ovary	115	ICGC
ICGC_PACA_AU	Pancreas	252	ICGC
ICGC_PACA_CA	Pancreas	181	ICGC
ICGC_PAEN_AU	Pancreas	48	ICGC
ICGC_PAEN_IT	Pancreas	37	ICGC
ICGC_PBCA_DE	Brain	374	ICGC
ICGC_PRAD_CA	Prostate	124	ICGC
ICGC_PRAD_UK	Prostate	161	ICGC
ICGC_RECA_EU	Kidney clear cell	95	ICGC
TCGA_AML_Strelka	Blood myeloid	49	TCGA
TCGA_MSI_Strelka	MSI colorectum	9	TCGA
TCGA_POLE_COAD_Strelka	POLE colon	7	TCGA
TCGA_POLE_READ_Strelka	POLE rectum	3	TCGA
TCGA_POLE_UCEC_Strelka	POLE uterus	2	TCGA
Wang_Gastric_MSI	MSI gastric	10	(Wang *et al.*, 2014)
Wang_Gastric_MSS	Gastric	90	(Wang *et al.*, 2014)

542

543

544