SUPPLEMENTARY FIGURES

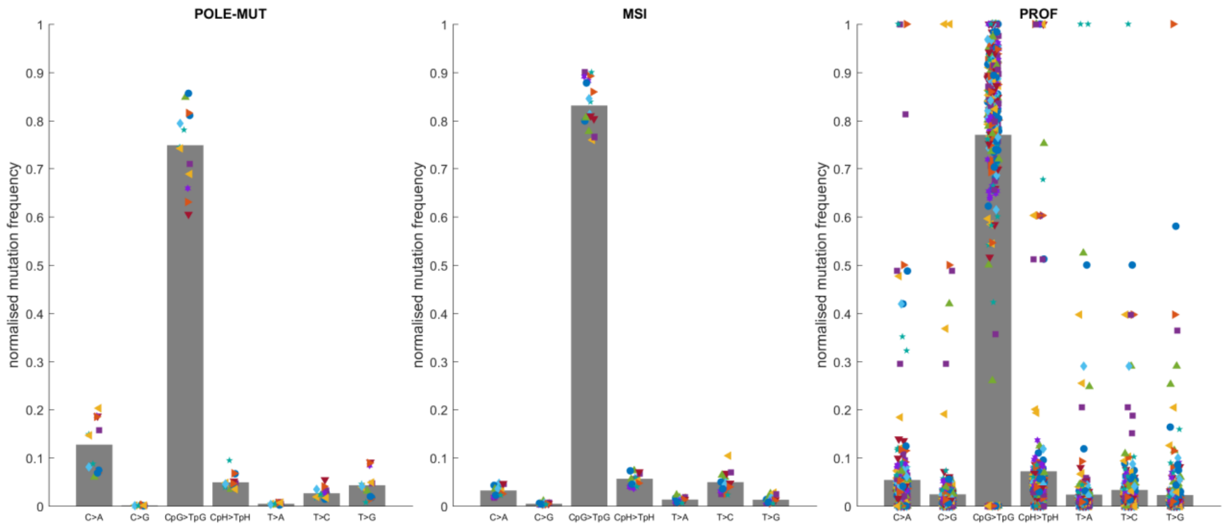

Fig. 1-supplement 1: Frequency of C to T mutations in a CpG context is unexpectedly high in *POLE*-MUT and MSI samples. Frequency of individual types of mutations in *POLE*-MUT, MSI, and tissue-matched PROF samples, normalised by the total sum in each sample. The bars denote mean over samples and individual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

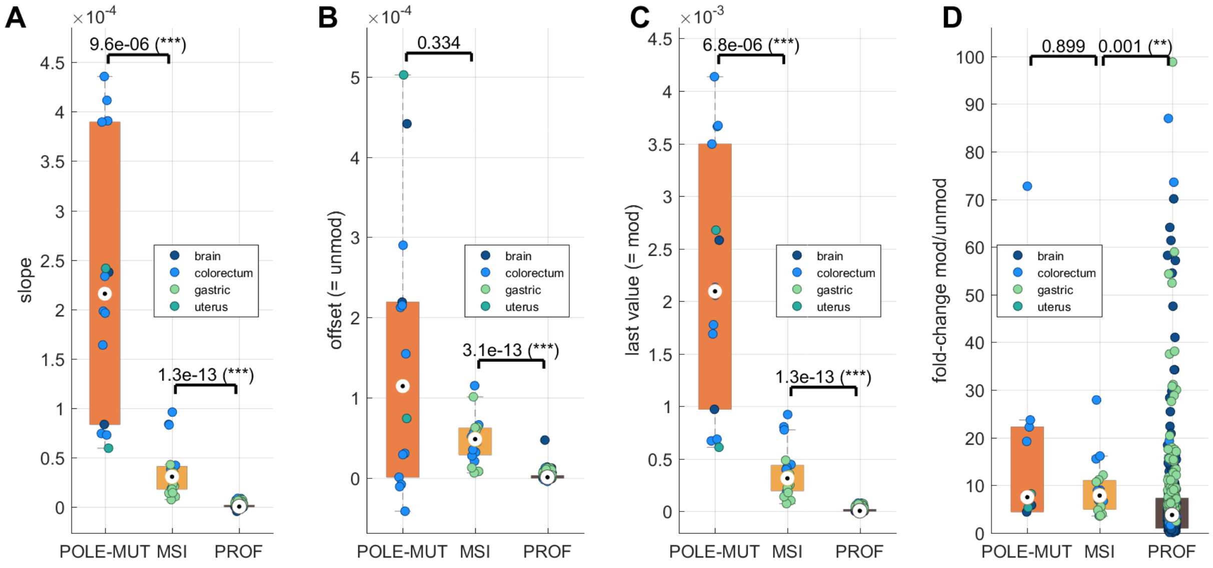

Fig. 1-supplement 2: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples correlates with DNA modification levels: comparison of linear models. In each sample, a linear model was fitted on the data, representing CpG>TpG mutation frequency in different bins of cytosine modification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

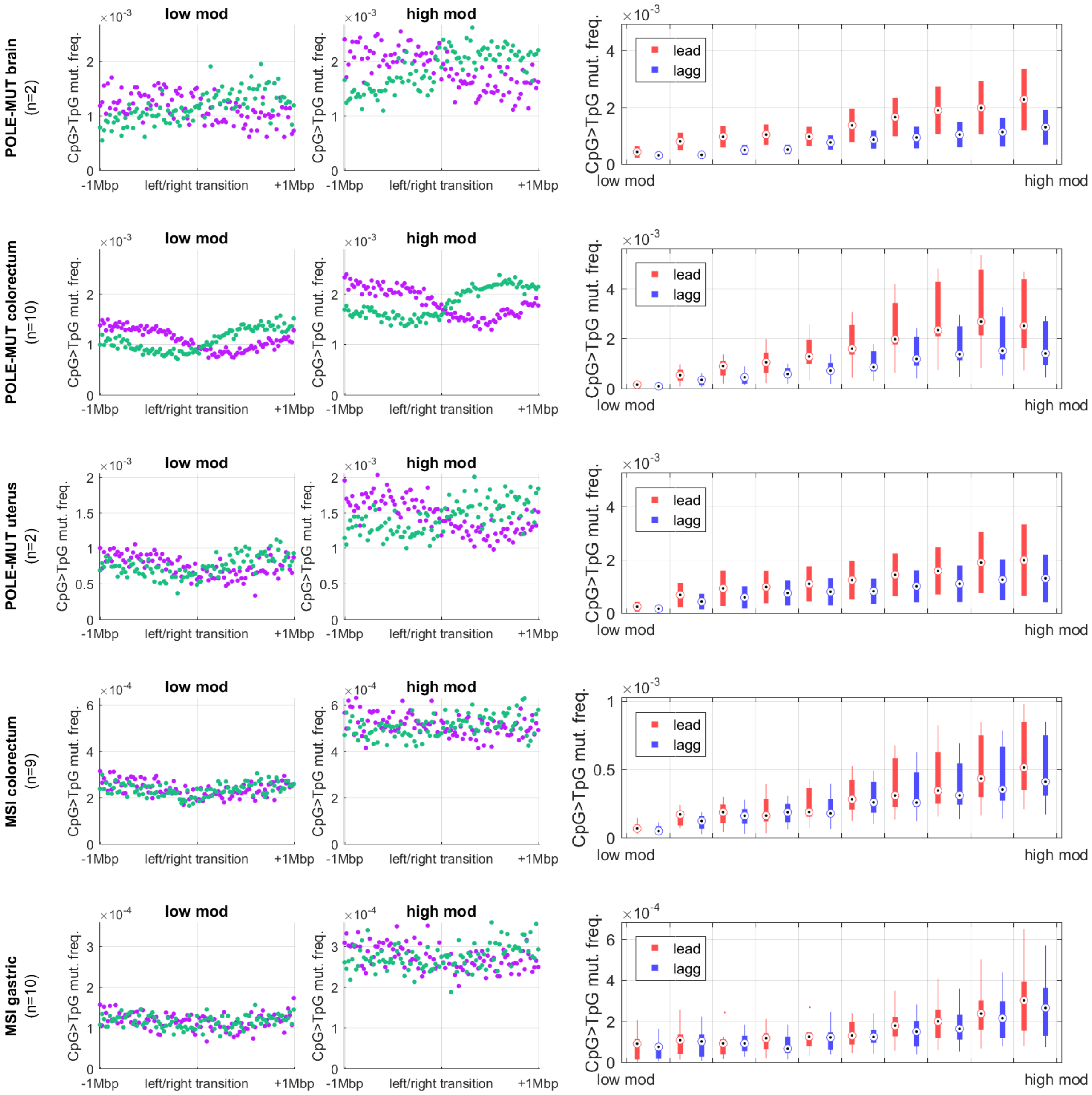

Fig. 2-supplement 1: Frequency of C to T mutations in a CpG context in *POLE*-MUT and MSI samples is higher on the leading strand than on the lagging strand, especially in modified CpG sites. Left column: Mean CpG>TpG mutation frequency on the plus (Watson) and minus (Crick) strand around transitions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

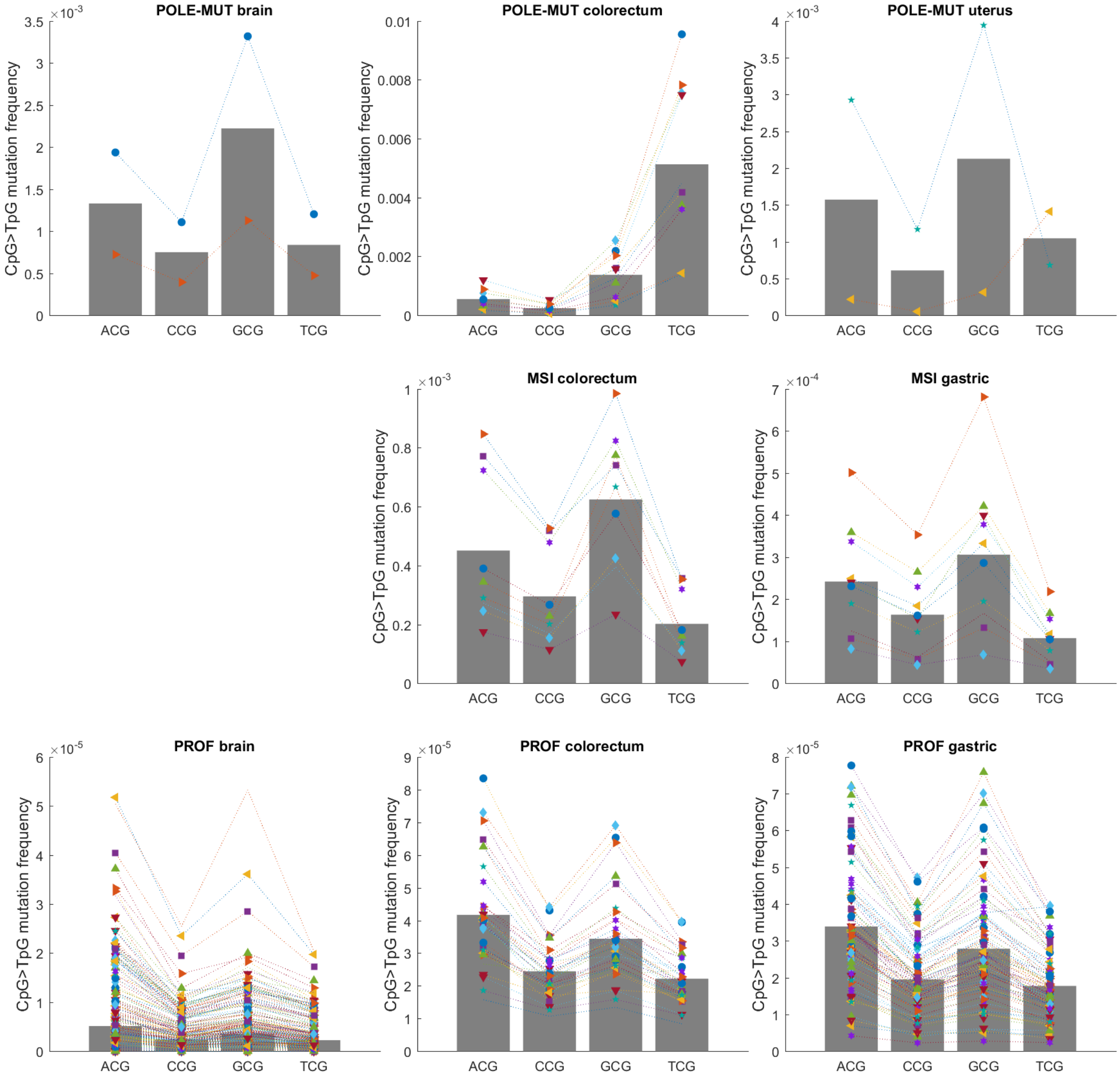

Fig. 3-supplement 1: CpG>TpG mutation frequency in different sequence contexts. CpG>TpG mutation frequency stratified by the 5’ flanking sequence context and tissue type. The bars denote mean over samples and individual samples are plotted in different colours and markers.

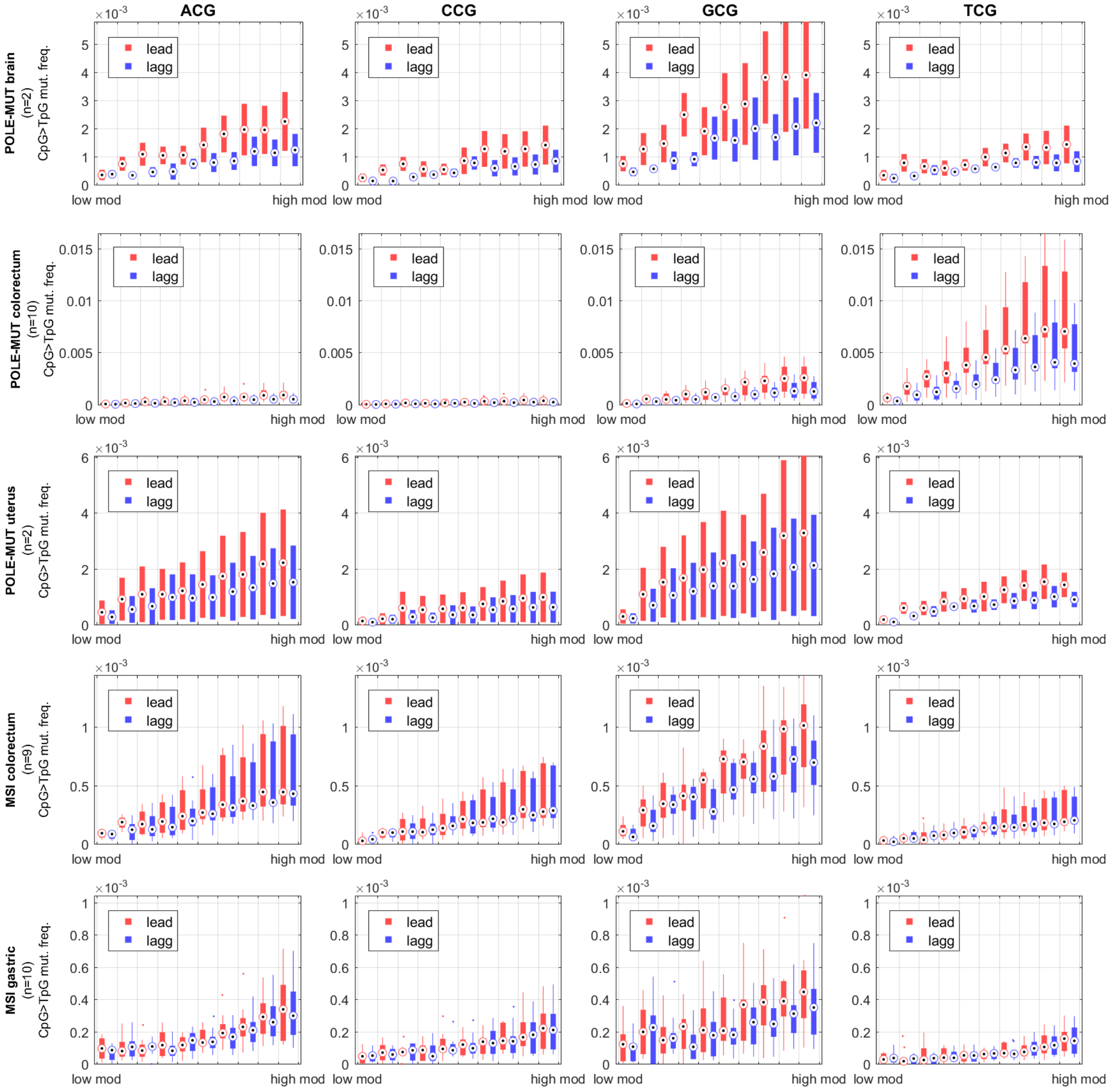

Fig. 3-supplement 2: Increase of C to T mutations in modified cytosine on the leading strand is most consistent in a GCG sequence context in *POLE*-MUT and MSI samples. C>T mutation frequency in CpG sites in leading and lagging strand binned by their tissue-matched modification levels (0-0.1, 0.1-0.2,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

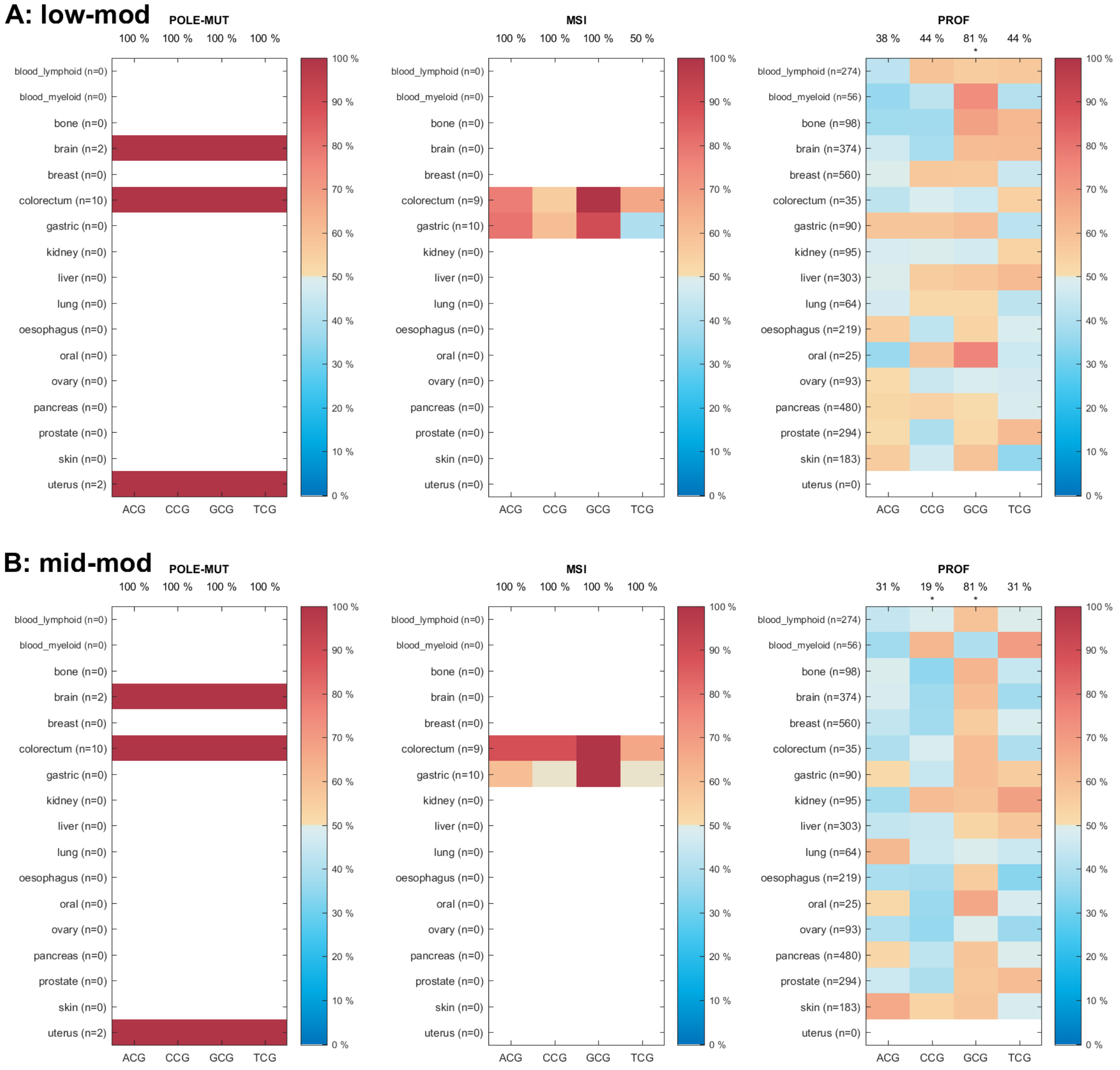

Fig. 4-supplement 1: GCG>GTG mutations are more frequent on the leading strand than on the lagging strand, even in Pol ε and MMR proficient samples. Percentage of samples with higher C>T mutation frequency on the leading strand than on the lagging strand for CpG sites with low (≤0.8) modification levels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

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)