Cytosine methylation affects the mutability of neighbouring nucleotides in human, Arabidopsis, and rice

Vassili Kusmartsev[1,2] & Tobias Warnecke[1,2*]

[1]Medical Research Council London Institute of Medical Sciences, London, United Kingdom

[2]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, United Kingdom

[*]corresponding author (tobias.warnecke@lms.mrc.ac.uk)

Running title: Altered mutability around methylated cytosines

ABSTRACT

Methylated cytosines deaminate at higher rates than unmethylated cytosines and the lesions they produce are repaired less efficiently. As a result, methylated cytosines are mutational hotspots. Here, combining rare polymorphism and base-resolution methylation data in humans, *Arabidopsis thaliana*, and rice (*Oryza sativa*), we present evidence that methylation state affects mutation dynamics not only at the focal cytosine but also at neighbouring nucleotides. In humans, contrary to prior suggestions, we find that nucleotides in the close vicinity (±3nt) of methylated cytosines mutate less frequently. In contrast, methylation is associated with increased neighbourhood mutation risk in *A. thaliana* and rice. The difference in mutation risk associated with methylation is less pronounced further away from the focal CpG, is modulated by regional GC content, and enhanced in heterochromatic regions. Our results are consistent with a model where elevated risk at neighbouring bases is linked to lesion formation at the focal cytosine and subsequent long-patch repair. Our results provide evidence that cytosine methylation has a broader mutational footprints than commonly assumed. They also illustrate that methylation is not intrinsically associated with higher mutation risk for surrounding bases, but that mutagenic effects reflect evolved species-specific and lesion-specific predispositions to elicit error-prone long-patch DNA repair.

Keywords: Cytosine methylation, mutation, base excision repair, *Arabidopsis thaliana*

INTRODUCTION

5-methylcytosine (5mC) is found in bacteria, archaea (Blow *et al*. 2016) and diverse eukaryotes, including vertebrates (Goll and Halpern 2011; Li and Zhang 2014), many invertebrates (Regev *et al*. 1998; Bewick *et al*. 2017), plants (Zhang *et al*. 2018), and fungi (Bewick *et al*. 2019). The modification, introduced by methyltransferases that target specific nucleotide contexts, marks the underlying sequence for differential treatment, preventing the binding of some proteins or facilitating recruitment of others, often – though not always – in the context of transcriptional silencing. In mammals, for example, where cytosine methylation is found almost exclusively at CpG dinucleotides, methylation represses transcription at individual promoters and can act in conjunction with histone modifications to establish and maintain larger silent domains, including during X chromosome inactivation (Goto and Monk 1998). In plants, methylation is principally associated with silencing of transposable elements, but can also be found in the bodies of constitutively expressed genes and contributes to dynamic regulation of multiple individual loci (Zhang *et al*. 2018). In both mammals and plants, interfering with methylation is typically deleterious and affects development, survival, or the ability to respond to environmental challenges.

Despite its importance for genome regulation, cytosine methylation comes at a cost: 5mCs are more liable to spontaneous deamination than unmethylated cytosines (Coulondre *et al*. 1978; Duncan and Miller 1980; Wang *et al*. 1982; Ehrlich *et al*. 1986; Zhang and Mathews 1994; Shen *et al*. 1994) and deaminate to thymine rather than uracil. In addition, the resulting T:G mismatches are less efficiently repaired than U:G mismatches (Schmutte *et al*. 1995; Krokan and Bjoras 2013). Methylated cytosines therefore carry the double burden of a higher rate of lesion formation and less efficient repair and, consequently, are more likely to give rise to mutations, specifically C to T transitions (Lutsenko and Bhagwat 1999).

The elevated mutability of 5mC has had and continues to have repercussions for patterns of genetic variation and genome evolution. CpGs are more likely than other dinucleotides to be found polymorphic in mammalian populations (Barker *et al*. 1984; Xia *et al*. 2012) and CpG to TpG changes are disproportionately common amongst variants associated with disease (Cooper and Krawczak 1989; Denissenko *et al*. 1997; Mancini *et al*. 1997; Zemojtel *et al*. 2009; Cooper *et al*. 2010; Zemojtel *et al*. 2011). On the flipside, transitions at CpGs more

frequently provide the raw material required for adaptation to novel environments (Stoltzfus and McCandlish 2017; Storz *et al*. 2019). The mutational impact of methylation is also visible over longer evolutionary timescales: CpG to TpG transitions often dominate substitution profiles between species (Ebersberger *et al*. 2002; Hwang and Green 2004) and genomes where CpG methylation is common are depleted of CpGs (Josse *et al*. 1961; Russell *et al*. 1976; Salser 1978; Bird 1980; Simmen 2008). Importantly, higher transition rates at CpGs are also evident in data from parent-child trios (Kong *et al*. 2012; Francioli *et al*. 2015; Rahbari *et al*. 2016; Jónsson *et al*. 2017) somatic mutations in healthy tissues (Hoang *et al*. 2016; Martincorena *et al*. 2018), mutation accumulation lines (Ossowski *et al*. 2010; Lee *et al*. 2012; Weng *et al*. 2019), and when considering rare SNPs (Rahbari *et al*. 2016; Carlson *et al*. 2018), strongly supporting mutational processes as the driving force. Finally, whereas early studies had to rely on CpGs as a (reasonable) proxy for methylation, more recent analyses have tethered elevated rates directly to methylation by integrating base-resolution methylation maps with polymorphism/somatic mutation data and comparing rates of evolution or SNP incidence at methylated and unmethylated CpGs explicitly (Ossowski *et al*. 2010; Mugal and Ellegren 2011; Lee *et al*. 2012; Xia *et al*. 2012; Supek *et al*. 2014; Tomkova *et al*. 2016; Weng *et al*. 2019). There is, in short, overwhelming evidence that cytosine methylation strongly impacts the emergence of novel variants, the spectrum of standing genetic variation, genome-wide base composition, and longer-term patterns of genome evolution.

The focal point in understanding shorter- and longer-term effects of methylation on genome fragility and evolution has, quite naturally, been the methylated cytosine itself. Methylation, however, can cast a longer mutational shadow and affect the rates of lesion formation, recognition and repair beyond the focal cytosine. For specific mutational processes, this has been well documented. Notably, methylation increases UV-induced formation of pyrimidine dimers (Tommasi *et al*. 1997; Ikehata and Ono 2007; Banyasz *et al*. 2016) and slows subsequent repair (Tornaletti and Pfeifer 1996). 5mC also affects the formation and repair of other directly adjacent lesions, including oxidation damage at neighbouring guanines (Tomkova and Schuster-Böckler 2018). But might methylation cast a longer shadow still? Cytosine methylation alters the physico-chemical properties of sequence in which it is embedded, affecting helix stability, rigidity, and dynamics (Szer and Shugar 1966; Collins and Myers 1987; Severin *et al*. 2011; Ngo *et al*. 2016). It can induce slight displacement of

the surrounding bases to the minor groove of the helix (Heinemann and Hahn 1992; Marcourt *et al*. 1999; Derreumaux *et al*. 2001), perturb biological processes such as cruciform extrusion (Murchie and Lilley 1989), and might therefore influence damage surveillance and handling more broadly. Further, even if elevated lesion risk is initially limited to the methylated base, neighbouring nucleotides can become collateral damage whenever repair involves excision and re-synthesis around the focal lesion, as is the case for mismatch repair (MMR), nucleotide excision repair (NER) and long-patch modes of base excision repair (BER). The excess mutational risk here can derive, for example, from the use of error-prone polymerases for re-synthesis. Alternatively, it might simply come from the transient generation of single-stranded DNA, which is more sensitive to mutagenic insults or required as a substrate for mutagenic enzymes such as APOBEC deaminases. By monitoring the repair of engineered mismatches in reporter constructs, such excess risk at sites in the vicinity of a focal lesion has been demonstrated, both *in vivo* and *in vitro* (Peña-Diaz *et al*. 2012). But are elevated mutation rates at 5mCs a significant trigger for such events? More generally, can one detect footprints of altered mutability around methylated versus unmethylated cytosines in genomic data?

Previously, Qu and colleagues reported a ~1.5-fold higher incidence of SNPs ±10bp around methylated compared to unmethylated CpGs in both human and medaka fish (*Oryzias latipes*) (Qu *et al*. 2012), consistent with a role for methylation in increasing the mutability of nucleotides in its vicinity. The observation that, during primate evolution, non-CpG substitution rates positively track the density of CpG dinucleotides in L1 transposons (Walser *et al*. 2008) is further consistent with this scenario. Implicating methylation as the causative force behind increased mutation rate, however, requires careful control of context. Mutation rate varies at multiple scales across the genome, depending on local and regional sequence composition, chromatin state, replication timing, and functional context (Hodgkinson and Eyre-Walker 2011; Ségurel *et al*. 2014; Makova and Hardison 2015). Methylated cytosines are unevenly distributed across these contexts, which might show different mutation rates for reasons unrelated to methylation. For example, in *A. thaliana* and other plants, *de novo* methylation of a sizeable subset of cytosines occurs via a process that specifically targets transposable elements rather than the genome at large (Zhang *et al*. 2018). Taking the non-random distribution of methylated cytosines into account is paramount to dissect whether

methylation has left a mark on genome variation and evolution beyond the methylated base itself.

Here, we use data on rare polymorphisms in human, *A. thaliana*, and rice to quantify the mutational effect methylation has on adjacent bases. Rare SNPs constitute a better proxy for mutational processes than common SNPs or substitutions as the latter more strongly reflect longer-term selection and gene conversion (Rahbari *et al*. 2016; Zhu *et al*. 2017). Controlling for sequence context and chromatin state, and considering a range of potential confounders, we find, in contrast to previous results, that methylation is associated with *reduced* SNP incidence at CpG-neighbouring sites in human. In both *A. thaliana* and rice, on the other hand, methylation is positively associated with SNP incidence. In *A. thaliana* and human, excess mutability associated with methylation (or lack of methylation, respectively) appears confined to close neighbours (±3bp) and decays with distance to the methylated site, supporting a mechanism that is contingent on lesion formation at the focal CpG. Our work suggests that methylation casts a longer mutational shadow than commonly assumed, acting in a manner that depends on species-specific coupling between lesion formation and downstream choice of repair pathway.

RESULTS

Methylation is associated with decreased mutability of neighbouring bases in humans

To establish whether SNP incidence varies as a function of methylation at nearby CpGs, we combined data from large-scale surveys of population genomic variation with base-resolution methylation data. For human, we defined methylated (unmethylated) sites as those with >70% (<20%) methylated reads in H1 human embryonic stem cells (hESC) (Lister *et al*. 2009), previously shown to be a reasonable proxy for germline methylation (Prendergast *et al*. 2014; Supek *et al*. 2014). Analysis was limited to sites covered by at least ten reads (see Methods for further details). As in prior work (Supek *et al*. 2014), we then paired methylated and unmethylated CpGs according to multiple criteria, which were applied simultaneously:

First, as mutation rate strongly varies with local sequence context (Blake *et al*. 1992; Zhao and Boerwinkle 2002; Hwang and Green 2004; Carlson *et al*. 2018), we required the four

nucleotides either side (±4bp) of the focal CpG to be the same. Matching the sequence context in this manner also controls for local GC content, previously shown to correlate inversely with CpG mutability (as further discussed below), and sequence complexity, which is an important determinant of indel formation propensity. As heptanucleotide context was previously shown to account for more than 80% of variability in mutation rates in humans (Aggarwala and Voight 2016), we did not extend matching to even longer sequence motifs, which would have drastically reduced sample size.

Second, we required each member of the methylated/unmethylated pair to be in the same chromatin state as defined by a widely used hidden Markov model for H1, which integrates signals from multiple histone marks, methylation and DNA accessibility (Ho *et al*. 2014) (see Methods). Matching by chromatin state is important because mutation rates vary substantially with chromatin environment (Schuster-Böckler and Lehner 2012; Makova and Hardison 2015). In particular, heterochromatic regions accumulate more mutations compared to euchromatic, actively transcribed regions (Schuster-Böckler and Lehner 2012), which are generally more accessible to or specifically targeted by DNA repair machinery (Supek and Ben Lehner 2015; Frigola *et al*. 2017). Chromatin states also capture other determinants of mutation rate heterogeneity, including replication timing and transcriptional activity, the latter being important in the context of this work because deamination risk is more than two orders of magnitude higher in single- compared to double-stranded DNA (Frederico *et al*. 1990).

Requiring the same nucleotide and chromatin context, choosing the closest available match along the same chromosome, and excluding sequence contexts with CpGs other than the focal CpG, we obtained 60,589 pairs of matched sites. At each base surrounding the focal CpG, we then calculated the incidence of singleton SNPs as observed across 15,708 whole genomes from the gnomAD database (Lek *et al*. 2016; Karczewski *et al*. 2019) (see Methods). To avoid looking at compound effects of clustered methylation sites and to allow comparison with previous studies (see below), any SNP that was a T to C transition at a TpG or an A to G transition at a CpA, was excluded. Applying this protocol, we find a reduced incidence of SNPs adjacent to methylated CpGs compared to unmethylated CpGs, illustrated as the relative mutational risk associated with methylation, RR_{met}, in Figure 1A. Across all sites (±1bp to ±3bp from the focal CpG) and mutation types (transitions and transversion),

RR_{met} is 0.886 ($P=2.73*10^{-21}$, Z-test for proportions). In other words, there are 886 SNPs in the ±3bp neighbourhood of methylated CpGs for every 1000 SNPs surrounding unmethylated CpGs.

Prior contradictory results are linked to the use of HapMap variation data

The above results appear to contradict a prior analysis of genome-wide human polymorphism and methylation data, which found more mutations in the vicinity of methylated sites (Qu *et al*. 2012). The analytical pipeline of Qu and colleagues (hereafter simply referred to as Qu) exhibits several potentially important differences to our approach. First, Qu used methylation data from sperm (Molaro *et al*. 2012) rather than H1 hESC (Lister *et al*. 2009). Second, they considered SNPs of any frequency from the HapMap project (CEU population) while we consider rare SNPs from the much larger gnomAD database. Third, whereas we pursue the matching approach described above, Qu assigned nucleotides in the vicinity (±10bp) of a given CpG to one of two categories: joining blocks of overlapping CpG dinucleotides, they averaged methylation levels across CpGs in the resulting larger block and considered methylated (unmethylated) blocks to be those with overall CpG methylation levels ≥80% (≤20%). They then computed an overall SNP incidence rate for each category. Finally, Qu analysed sites with ≥5 reads.

To understand which – if any – of these analytical choices explain the discrepancy, we started by reimplemented our matching approach with the sperm methylation data used by Qu. We obtained very similar results (RR_{met}=0.877, $P=1.68*10^{-160}$, Figure 1B), suggesting that the different methylation datasets are not the source of the discrepancy. We also obtained very similar results when considering sites with ≥5 instead of ≥10 reads (overall RR_{met}=0.889, $P=3.19*10^{-81}$), which increases samples size to 162,156 pairs. Unsurprisingly, given that the distribution of methylation stoichiometries exhibits two modes at the extremes (i.e. towards 0% and 100% methylated, see Supek *et al*. 2014), there is also no substantive change when we require ≥80% rather than ≥70% read support to call a site methylated (not shown). Next, we sought to reproduce the original finding by implementing the approach of Qu in full, using HapMap (CEU) polymorphisms, sperm methylation data, sites covered by ≥5 read, a threshold of ≥80% for calling methylated sites, and calculating RR_{met} as the SNP incidence in blocks around focal CpGs as described above. Doing so, we can replicate their results, at

least qualitatively (RR_{met}=1.4, Figure 1C). Using H1 instead of sperm data again yielded very similar results (Figure 1C), re-confirming that differences in methylation data are not pertinent. We then checked whether inclusion of repeats or regions associated with less reliable SNP calls, might explain the difference. Excluding repeats or poorly accessible regions, as defined by the 1000 Genomes project (see Methods), however, has limited effect, with RR_{met} consistently >1 (Figure 1C). In contrast, we obtain dramatically different results when substituting HapMap (CEU) polymorphisms for singleton SNPs from the 1000 Genomes project (overall RR_{met}=0.971) or gnomAD (overall RR_{met}=0.87, Figure 1C).

One key difference between the variation datasets above is the relative prominence of common alleles, which is much greater in HapMap. One might therefore reasonably hypothesize that different RR_{met} estimates in the HapMap, 1000 Genomes, and gnomAD data are owing to differences in average allele frequency, perhaps because common alleles are a poorer proxy for mutational processes, reflecting selection and gene conversion to a greater extent than rare SNPs. However, allele frequency appears to be only a minor factor. We obtain similar results (i.e. RR_{met}<1) when we threshold 1000 Genomes and gnomAD data to only include SNPs present at ≥5% frequency in the given sample. We therefore conclude that difference in SNP quality/mapping in the original HapMap data likely underlie difference between our results and those of Qu. As rare SNPs provide better proxy for mutational processes, we further conclude that there is no evidence that methylation is associated with increased mutability at adjacent sites. Rather, cytosine methylation is associated with a significant reduction in the incidence of mutations in the neighbourhood of CpGs in humans. As we discuss in greater depth below, this is consistent with recent experimental data.

Methylation is associated with increased SNP incidence in plants

In contrast to humans, we find SNP-approximated mutability of CpG-neighbouring bases to be positively associated with methylation in two plants: *A. thaliana* and rice. Applying the same matching protocol (and confining analysis to repeat-masked sequence, see Methods), we find an overall RR_{met} of 1.28 ($P=2.06*10^{-89}$) in *A. thaliana* and 1.31 in *O. sativa* ($P=5.43*10^{-15}$), where estimates are noisier due to relatively smaller number of matched pairs (Figure 2A; N=121,774 pairs in *A. thaliana*, N=42,779 pairs in *O. sativa*). The strongest increase in mutational risk was associated with C to T transition SNPs (RR_{met}=1.38 in *A.*

thaliana; RR_{met}=1.71 in *O. sativa*). Interestingly, RR_{met} appears to level off as a function of distance from the focal CpG, with the greatest deviation from random expectation (RR_{met}=1) at the nucleotide directly adjacent to the CpG. A similar (albeit inverted) trend is also evident in humans (Figure 2A, Figure 1A/B).

Methylation is associated with an altered incidence of insertions and deletions

With a view to understanding the mechanism of altered neighbourhood mutability, we also investigated whether the presence of methylation affected the rates of insertions and deletions (indels) around the focal CpG, using the same matched pairs as above. As for SNPs, we focus on singleton indels and also confine analysis to single-base insertions and deletions, thus ensuring that the sequence context of the indel is comparable. As indels are rarer than SNPs, we compute only two RR_{met} estimates, for the CpG itself and all neighbouring bases upstream (±1-3bp) combined. In humans, echoing results from SNPs, we find a reduced incidence of indels for methylated sites, both at the CpG itself and the neighbouring sequence (Figure 2B). In *A. thaliana*, RR_{met} is higher across the board, with a tendency for RR_{met}>1 for deletions in CpG-proximal sites. There were too few events to carry out an informative analysis of indel patterns in rice.

Increased mutability around methylated sites in plants is not limited to the CpG context

As methylation in plants is not limited to CpGs but also occurs in CpHpG and CpHpH contexts (where H=A, C, or T), we independently collated matched pairs for these contexts. In both cases, there is a tendency for RR_{met} to be greater than 1 (Figure S1). However, as methylation at CpHpG and CpHpH sites is rarer (6% and 1.5% compared to 24% for CpG in *A. thaliana*, 21% and 2.2% compared to 59% for CpG in *O. sativa*) and more strongly skewed to certain functional contexts, sample sizes are much smaller. We therefore focus on CpGs below, but note that elevated RR_{met} across all contexts suggest that differential risk is likely not associated with machinery that acts exclusively at CpGs versus other contexts.

Methylation-associated changes in mutability are not sequence- or context-specific

To identify potential drivers of methylation-associated differences in mutability (and pinpoint difference between plants and humans), we stratified RR_{met} by sequence context, local/regional GC content, and chromatin state. First, we note that higher mutability around methylated CpGs is observed across most central sub-contexts (CpG ±1bp) that are sufficiently common (present in more than 20 pairs) to allow independent assessment of SNP rates (Figure 3A). This suggests that, although there is variability associated with the surrounding sequence, the effect is not driven by a specific nucleotide context or set of contexts. This also provides a first pointer that differential context composition does not explain why observations in plants differ from those in human. To rule this out explicitly, we subsampled matched pairs to achieve an identical representation of nucleotide contexts in human and *A. thaliana*. Doing so, RR_{met} remains >1 in *A. thaliana* and <1 in human (Figure S2).

We then considered to what extent RR_{met} is influenced by or robust to local and regional GC content. GC content has previously been suggested to impact focal deamination rates as DNA duplexes of high GC content are less likely to spontaneously become single-stranded (Leroy *et al*. 1988). In line with this model, observed CpG content was previously found to be lower than expected in regions of low GC but not in regions of high GC content (Adams and Eason 1984) and substitution rates at CpGs in primates and Arabidopsis correlate inversely with GC content (Fryxell 2004; Leah J DeRose-Wilson 2007). Considering local (motif-wide) and regional (±100bp either side of the focal CpG) GC content, we find that RR_{met} in plants is highest at low-to-intermediate GC content and becomes less pronounced at higher GC content (Figure 3B, Figure S3). This is consistent with a collateral damage model of mutagenesis, where initial deamination events at methylated focal CpGs put neighbouring sites at risk and do so more frequently in a low GC context, where spontaneous deamination is more likely to occur. In humans, we observe no such relationship, in line with the absence of 5mC deamination as the driver of mutational liability.

Next, we considered variability in RR_{met} across different chromatin states. Again, we find consistent signal RR_{met} >1 in plants (Figure 3C, Figure S4). However, some chromatin states emit much stronger signals than others. Notably, chromatin states characterized by the presence of transposable elements and/or histone marks that are generally associated with silencing (H3K27me3, H3K9me2), not only have higher absolute SNP incidence but also

notably higher RR_{met} compared to transcribed genic sequence. Some chromatin states marked by high accessibility (DNase hypersensitivity) also have high RR_{met}. However, this is only true for states (e.g. promoters) outside of transcribed regions, where the methylation-associated risk of mutation appears to be strongly reduced, perhaps as a result of transcription-coupled repair.

Sites accessible to the (de)methylation machinery show reduced relative risk

In both mammals and plants, methylated cytosines are subject to active demethylation. In mammals, ten-eleven translocation (TET) enzymes oxidize 5mC residues to yield 5-hydroxymethylcytosine, which – along with further-oxidized derivatives – can be excised by DNA glycosylases like TDG. In plants, active demethylation is more direct. Different glycosylases – such as DEMETER (DME) and REPRESSOR OF SILENCING 1 (ROS1) in *A. thaliana* – target and excise 5mC itself (Zhu 2009). Both direct or indirect demethylation pathways, being reliant on DNA glycosylase activity to generate abasic sites, effectively constitute instances of programmed lesion formation. For mammals, our lab and others have previously shown that cytosines that spend a larger fraction of their time in a hydroxymethylated state are subject to different mutation dynamics (Supek *et al*. 2014; Tomkova *et al*. 2016). In a similar vein, we wanted to know whether repeated activity of the *A. thaliana* demethylation machinery might similarly be associated with altered mutagenesis, which might also affect neighbouring sites. In particular, we hypothesized that sites that undergo frequent methylation/demethylation/remethylation cycles might suffer from a greater cumulative risk of mutation if repair following glycosylase activity – thought to be principally BER – is mutagenic. Sites that undergo such cycles are not uncommon in *A. thaliana*. ROS1 in particular has been implicated in preventing, through ongoing active removal of 5mCs, the spread of methylation from transposable elements into active genes (Zhang *et al*. 2018).

We therefore considered RR_{met} in the context ROS1 activity, classifying sites as ROS1 targets if they showed increased methylation in a *ros1* knock-out strain (see Methods). Similarly, we considered sites to be regular targets of the RNA-directed DNA methylation (RdDM) machinery, which is responsible for *de novo* methylation, if they showed decreased methylation in strains deleted for *nrpd1*, a polymerase IV subunit critical for methylation.

Here, in order to include effects of ROS1-targeted sites, we considered changes in methylation in *nrpd1*/*ros1* double mutants compared to the *ros1* mutant. We paired sites by sequence context and by ROS1/NRPD1 target status. As RR_{met} is >1 throughout different chromatin contexts, we jettison chromatin state pairing to maintain a sufficient samples size for analysis. We find that sites targeted by ROS1 experience a substantially lower excess mutation risk associated with methylation than sites that do not experience increased methylation upon *ros1* knock-out (Figure 4A). We observe very similar results for *nrpd1* (Figure 4A). Post hoc analyses suggests that these differences are also observed when chromatin state is controlled for (not shown). We note that this effect is quantitative: sites that experience greater gain (loss) in methylation in *ros1* (*nrpd1)* knock-outs, show lower RR_{met} (Figure 4B). Finally, cytosines that are dependent on the chromatin remodeler DDM1 for methylation show higher RR_{met} than sites where methylation can be established and subsequently maintained solely through RdDM and maintenance methyltransferase activity (Figure 4A, see Methods for how DDM1 dependency was defined).

Taken together, these results are inconsistent with a model where higher mutability of neighbouring bases is a consequence of frequent methylation/demethylation cycles. Rather, they reinforce the impression that RR_{met} is enhanced in regions that are less accessible to machinery that affects methylation, demethylation and presumably demethylation-coupled repair.

DISCUSSION

Our results suggest that methylation state affects mutability of neighbouring nucleotides in human, *A. thaliana* and rice. Notably, we find $RR_{met}>1$ in plants but $RR_{met}<1$ in human, which is not explained by differential representation of specific sequence or chromatin contexts. This suggests that differential effects of methylation across species cannot be traced back to difference between methylated/unmethylated sequence with respect to their biophysics, which is species-invariant, but must instead be caused by differential responses by cellular machinery to methylation or methylation-associated lesions. We therefore suggest that RR_{met} is not linked to methylation per se, but reflects the probability of lesion formation, the nature of the resulting lesion, and how the cellular machinery deals with that lesion.

In human and *A. thaliana*, the spatial pattern of RR_{met} around the focal CpG returns relatively quickly towards the baseline. Even though our matching approach curtails examination of longer-range effects, this suggests that the mutational effects are relatively local. We think that this locally confined signal is suggestive of BER, which has previously been tipped as a potential culprit behind neighbourhood mutation effects (Qu *et al*. 2012). Experiments in cell extracts have demonstrated the existence of different flavours of BER in both humans and plants (Córdoba-Cañero *et al*. 2009; Martínez-Macías *et al*. 2013), which differ in the number of bases excised and re-synthesized during the repair process. In short-patch BER (SP-BER) only a single nucleotide is added whereas during long-patch BER (LP-BER) multiple bases are excised and re-synthesized. In *A. thaliana*, tracking repair at U:G mismatches, repair tract lengths of up to 3bp have been observed (Córdoba-Cañero *et al*. 2009; Martínez-Macías *et al*. 2013). In mammals, tracts removed during LP-BER are similarly short, 2-13bp (Fortini and Dogliotti 2007; Krokan and Bjoras 2013). This contrasts with nucleotide excision and mismatch repair, where excision tract lengths are typically much longer, from >20bp for NER (Sancar 1996) to several hundred bases or more during MMR (Fang and Modrich 1993). The involvement of LP-BER is also consistent with increased single-nucleotide deletion rates (Bennett *et al*. 2001; Lyons and O'Brien 2010).

But why would trends be inverted for human and plants? We posit that RR_{met} reflects the combination of two distinct risk factors: the pathway chosen downstream of a given lesion (which determines sign; e.g. whether U:G or T:G is handled more efficiently), and the relative risk of lesion formation (which should be higher for T:G and affects amplitude). In humans, while some lesions, including 8-oxoguanine, are thought to mostly trigger SP-BER (Fortini *et al*. 1999), others have been associated with LP-BER. Importantly, this includes U:G mismatches, that result from deamination of unmethylated cytosines. Studying mouse embryonic fibroblasts, Bennett *et al*. showed that, where uracil removal is triggered by uracil DNA glycosylases (UNG), subsequent re-synthesis exceeded a single nucleotide in 80% of cases, suggesting that UNG, perhaps by virtue of interacting with PCNA, biases downstream repair pathway choice towards long-patch repair (Bennett *et al*. 2001; Fortini and Dogliotti 2007).

Importantly, such "uracil-initiated" (Bennett *et al*. 2001) LP-BER has already been associated with higher mutation risk at neighbouring sites: Chen and colleagues introduced mismatches into an SV40 episome capable of replicating in human cells to monitor mutagenic effects either side of the mismatch. While they found that both T:G and U:G mismatches were associated with collateral damage to the neighbourhood, liability was ~7-fold higher for U:G (Chen *et al*. 2014). Thus, even though, in a physiological context, the rate of lesion formation might be higher for T:G, the mutagenesis risk associated with repair might be greater for U:G. This is in line with our findings. The precise mechanism(s) behind increased rates surrounding unmethylated CpGs in our data, however, remains unclear. Chen *et al*. showed that both BER and MMR were required for elevated mutability and proposed a model inspired by events during somatic hypermutation at immunoglobulin genes, where BER-generated lesions are hijacked by the MMR machinery, as previously demonstrated for U:G repair (Schanz *et al*. 2009; Peña-Diaz *et al*. 2012) and known to occur in the context of active demethylation (Grin and Ishchenko 2016). Specifically, they suggested that excess mutability in their experimental model is consistent with APOBEC enzymes tagging along with the MMR machinery to attack and deaminate single-stranded cytosines. In our data, however, we find no enrichment of the tell-tale APOBEC mutational signature (C to U changes in a TpCpN context). APOBEC-related excess mutational risk might therefore be a specific manifestation of a more general liability of being more fragile in a single-stranded state. In plants, both LP- and SP- BER have been observed in *A. thaliana*. How different lesions (or associated glycosylases) predispose to either short- or long-patch repair remains poorly understood (Lee *et al*. 2014), but – as the repertoire of glycosylases differs substantially between humans and plants – it is certainly conceivable that U:G and T:G mismatches might be associated with a different propensity for LP- versus SP-BER to what is seen in humans.

We speculate in this regard that choice of repair (sub-)pathways in different species – and the ensuing mutational burden – might be evolved rather than random. Methylation in plants is more tightly linked to transposable elements than it is in mammals. Consequently, there is likely less selection in plants to prevent lesions at and around methylated sites. Where methylated CpGs are located in functionally important regions (gene bodies), they are relatively less affected, perhaps because of transcription-coupled repair (Figure 3C).
In mammals, on the other hand, there are more cases where methylation-based silencing is transient and the underlying sequence remains important, notably in the context of X

inactivation. Elevated mutation loads in the neighbourhood of methylated cytosines might therefore be much less well tolerated in mammals, so that better repair of these lesions evolved. It is interesting to note in this regard that, during post-replicative repair in mammals, G:T mismatches in the context of hemi-methylated sites are preferentially corrected to G:C with high (~90%) efficiency (Brown and Jiricny 1987; Bill *et al*. 1998), perhaps reflective of an evolved bias to counter frequent deamination at 5mC. In plants, no such bias has been observed although direct evidence is limited to tobacco protoblasts (Inamdar *et al*. 1992) and this issue deserves further investigation.

Irrespective of the mechanistic underpinnings of altered methylation-associated mutability, which remain to be resolved, our study provides strong evidence that the mutational impact of methylation extends beyond the methylated cytosine itself and shapes the emergence of novel variants across the genomes of different eukaryotes.

METHODS

Analysis of relative mutational risk associated with methylation in human

To isolate the specific impact of methylation on SNP incidence, we pursued a matched pairs approach, broadly as previously described (Supek *et al*. 2014). First, CpGs in the human genome were classified as methylated or unmethylated based on base-resolution methylation information in H1 human embryonic stem cells (Lister *et al*. 2009) (http://neomorph.salk.edu/human_methylome/data.html). Methylated sites were defined as cytosines with a ≥0.7 ratio of methylated to unmethylated reads; unmethylated sites as those with a ratio ≤0.2. CpGs with intermediate methylation (0.2>x<0.7) levels were excluded from further analysis. Note that, as methylation stoichiometry is strongly bimodal, this excludes relatively few sites (see Supek *et al*. 2014 for details). To provide robust classification of methylated/unmethylated sites, we only considered CpGs with ≥10 read coverage in the bisulfite sequencing data. As the X chromosome is differentially affected by methylation in males and females and because mutations that arise on sex chromosomes are subject to distinct evolutionary regimes, analysis was confined to autosomes. To enable intersection with polymorphism data, coordinates for all eligible methylated/unmethylated CpG sites were

converted to hg19 using the UCSC LiftOver utility (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Each unmethylated CpG was then paired with the nearest methylated CpG that matched the following criteria: a) identical ±4bp flanking sequence context in the reconstructed ancestral human genome (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/), b) identical chromatin state as defined by a hidden Markov model for H1 (https://personal.broadinstitute.org/anshul/projects/wfh/ihmm/chmmBED/human_H1/), c) located on the same chromosome. Subsequently, any pairs whose flanking sequence context contained one or more additional CpG dinucleotides were removed to avoid possible confounding effects of proximal CpGs on local mutation risk. Further, pairs that included CpGs not captured by the H1 HMM were also removed. For the surviving pairs of methylated and unmethylated CpGs, we then calculated the incidence of singleton SNPs from the gnomAD database (Karczewski *et al*. 2019) (https://gnomad.broadinstitute.org/). In calculating the incidence of transition SNPs, we removed T to C SNPs that had occurred in the TpG context and A to G SNPs that had occurred in the CpA context, principally to enable comparison to the analysis of Qu et al, where – given the data available at the time – it was prudent to exclude these sites because of potential polarization errors. Note that, above, we only consider SNPs at positions ±3bp from the focal CpG despite pairing for ±4bp. This is because at the ±4 position, the ±5 base is unknown, and so including this position could lead to skews due to unknown neighbouring bases.

Analysis of relative mutational risk associated with methylation in Arabidopsis and rice

The protocol outlined above was also applied to *A. thaliana* and rice. Classification of CpGs (but also CHG and CHH sites) into methylated and unmethylated sites was based on bisulfite data of Ws0 global stage seed for *A. thaliana* (GSM1664380) (Lin *et al*. 2017) , and data from 3-week old leaf tissue in rice (GSM1039487) (Stroud *et al*. 2013). As highlighted above, to calculate rare SNP incidence, we considered homozygous SNPs confined to a single accession in the selfing *A. thaliana* as singletons. HMM-defined chromatin states for both *A. thaliana* and rice were taken from the Plant Chromatin State Database (Liu *et al*. 2018) (http://systemsbiology.cau.edu.cn/chromstates/). The 1001 Arabidopsis genomes (https://1001genomes.org/data/GMI-MPI/releases/v3.1/) and the 3000 rice genomes projects (http://snp-seek.irri.org/), respectively, were the sources of polymorphism data.

Repeat sequences in both plants, as well as human, were masked using RepeatMasker v4.0.8 (www.repeatmasker.org).

Methylation/demethylation mutants

We obtained base-resolution methylation data for knock-out mutants of *ros1* (GSM1859475), *nrpd1* (GSM1859476), and a *ros1*/*nrpd1* (GSM1859478) double knock-out (Wierzbicki *et al*. 2012) as well as for DDM (GSM1014117) and DRD (GSM1014120) mutants (Zemach *et al*. 2013). Base calls were lifted over to TAIR10 as required. Sites affected by a given deletion were classified as follows: a site targeted by ROS1 was one where methylation was greater in the *ros1* mutant than in the WT; a site targeted by NRPD1 was one where methylation was lower in the *nrpd1*/*ros1* double mutant compared to the *ros1* deletion strain; a site requiring DDM for methylation was one where the *ddm* mutant had lower methylation and the *drd* mutant had no effect.

Replicating prior estimates of relative mutational risk associated with methylation

To track down divergent RR_{met} estimates compared to the study of Qu, we additionally considered base-resolution bisulfite sequencing data from sperm (Molaro *et al*. 2012) (GSE30340), lifted over to hg19, and re-implemented their original protocol based on the published methods and feedback provided by the lead author, Wei Qu. This included consideration of sites with ≥5 read coverage, with unmethylated/methylated CpGs defined as those with methylation stoichiometries of ≥0.8 and ≤0.2, respectively. Nucleotides flanking the focal CpG (±10bp) were considered part of methylated or unmethylated "blocks" based on the methylation status of the focal CpG. These site blocks, defined according to either H1 or sperm methylation data, were then overlapped with polymorphism data, either hg19-converted HapMap (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), 1000 Genomes (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), or gnomAD (https://gnomad.broadinstitute.org/) (Karczewski *et al*. 2019). As in the original study, SNPs that occurred in CpG/TpG/CpA dinucleotides were excluded.

Acknowledgements

The authors are grateful to Wei Qu for help in replicating her original protocol and members of the Molecular Systems group for discussions. This work was supported by UK Medical Research Council core funding to TW.

Author contributions

VK carried out all analyses. TW conceived the study. VK and TW designed analyses, interpreted results, and wrote the manuscript.

Competing financial interests

The authors declare that no competing financial interests exist.

References

Adams R. L. P., Eason R., 1984 Increased G + C content of DNA stabilizes methyl CpG dinucleotides. Nucleic Acids Research **12**: 5869–5877.

Aggarwala V., Voight B. F., 2016 An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet **48**: 349–355.

Banyasz A., Esposito L., Douki T., Perron M., Lepori C., Improta R., Markovitsi D., 2016 Effect of C5-Methylation of Cytosine on the UV-Induced Reactivity of Duplex DNA: Conformational and Electronic Factors. J. Phys. Chem. B **120**: 4232–4242.

Barker D., Schafer M., White R., 1984 Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. Cell **36**: 131–138.

Bennett S. E., Sung J. S., Mosbaugh D. W., 2001 Fidelity of uracil-initiated base excision DNA repair in DNA polymerase beta-proficient and -deficient mouse embryonic fibroblast cell extracts. The Journal of Biological Chemistry **276**: 42588–42600.

Bewick A. J., Hofmeister B. T., Powers R. A., Mondo S. J., Grigoriev I. V., James T. Y., Stajich J. E., Schmitz R. J., 2019 Diversity of cytosine methylation across the fungal tree of life. Nat Ecol Evol **3**: 479–490.

Bewick A. J., Vogel K. J., Moore A. J., Schmitz R. J., 2017 Evolution of DNA Methylation across Insects. Mol Biol Evol **34**: 654–665.

Bill C. A., Duran W. A., Miselis N. R., Nickoloff J. A., 1998 Efficient Repair of All Types of Single-Base Mismatches in Recombination Intermediates in Chinese Hamster Ovary Cells: Competition Between Long-Patch and G-T Glycosylase-Mediated Repair of G-T Mismatches. Genetics **149**: 1935–1943.

Bird A. P., 1980 DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research **8**: 1499–1504.

Blake R. D., Hess S. T., Nicholson-Tuell J., 1992 The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol **34**: 189–200.

Blow M. J., Clark T. A., Daum C. G., Deutschbauer A. M., Fomenkov A., Fries R., Froula J., Kang D. D., Malmstrom R. R., Morgan R. D., Posfai J., Singh K., Visel A., Wetmore K., Zhao Z., Rubin E. M., Korlach J., Pennacchio L. A., Roberts R. J., 2016 The Epigenomic Landscape of Prokaryotes (G Fang, Ed.). PLoS Genet. **12**: e1005854.

Brown T. C., Jiricny J., 1987 A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. Cell **50**: 945–950.

Carlson J., Locke A. E., Flickinger M., Zawistowski M., Levy S., Myers R. M., Boehnke M., Kang H. M., Scott L. J., Li J. Z., Zöllner S., 2018 Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nature Communications **9**: 3753.

Chen J., Miller B. F., Furano A. V., Walter J., 2014 Repair of naturally occurring mismatches can induce mutations in flanking DNA. eLife **3**: e02001.

Collins M., Myers R. M., 1987 Alterations in DNA helix stability due to base modifications can be evaluated using denaturing gradient gel electrophoresis. Journal of Molecular Biology **198**: 737–744.

Coulondre C., Miller J. H., Farabaugh P. J., Gilbert W., 1978 Molecular basis of base substitution hotspots in Escherichia coli. Nature **274**: 775–780.

Córdoba-Cañero D., Morales-Ruiz T., Roldán-Arjona T., Ariza R. R., 2009 Single-nucleotide and long-patch base excision repair of DNA damage in plants. The Plant Journal **60**: 716–728.

Derreumaux S., Chaoui M., Tevanian G., Fermandjian S., 2001 Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. Nucleic Acids Research **29**: 2314–2326.

Duncan B. K., Miller J. H., 1980 Mutagenic deamination of cytosine residues in DNA. Nature **287**: 560–561.

Ebersberger I., Metzler D., Schwarz C., Pääbo S., 2002 Genomewide comparison of DNA sequences between humans and chimpanzees. The American Journal of Human Genetics **70**: 1490–1497.

Ehrlich M., Norris K. F., Wang R. Y., Kuo K. C., Gehrke C. W., 1986 DNA cytosine methylation and heat-induced deamination. Bioscience Reports **6**: 387–393.

Fang W. H., Modrich P., 1993 Human strand-specific mismatch repair occurs by a bidirectional mechanism similar to that of the bacterial reaction. The Journal of Biological Chemistry **268**: 11838–11844.

Fortini P., Dogliotti E., 2007 Base damage and single-strand break repair: Mechanisms and functional significance of short- and long-patch repair subpathways. DNA Repair **6**: 398–409.

Fortini P., Parlanti E., Sidorkina O. M., Laval J., Dogliotti E., 1999 The type of DNA glycosylase determines the base excision repair pathway in mammalian cells. The Journal of Biological Chemistry **274**: 15230–15236.

Francioli L. C., Polak P. P., Koren A., Menelaou A., Chun S., Renkens I., van Duijn C. M., Swertz M., Wijmenga C., van Ommen G., Slagboom P. E., Boomsma D. I., Ye K., Guryev V., Arndt P. F., Kloosterman W. P., de Bakker P. I. W., Sunyaev S. R., 2015 Genome-wide patterns and properties of de novo mutations in humans. Nat Genet **47**: 822–826.

Frederico L. A., Kunkel T. A., Shaw B. R., 1990 A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29**: 2532–2537.

Frigola J., Sabarinathan R., Mularoni L., Muiños F., Gonzalez-Perez A., López-Bigas N., 2017 Reduced mutation rate in exons due to differential mismatch repair. Nat Genet **49**: 1684–1692.

Fryxell K. J., 2004 CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. Mol Biol Evol **22**: 650–658.

Goll M. G., Halpern M. E., 2011 DNA Methylation in Zebrafish. Progress in molecular biology and translational science **101**: 193–218.

Goto T., Monk M., 1998 Regulation of X-chromosome inactivation in development in mice and humans. Microbiol. Mol. Biol. Rev. **62**: 362–378.

Grin I., Ishchenko A. A., 2016 An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation. Nucleic Acids Research **44**: 3713–3727.

Heinemann U., Hahn M., 1992 C-C-A-G-G-C-m5C-T-G-G. Helical fine structure, hydration, and comparison with C-C-A-G-G-C-C-T-G-G. The Journal of Biological Chemistry **267**: 7332–7341.

Ho J. W. K., Jung Y. L., Liu T., Alver B. H., Lee S., Ikegami K., Sohn K.-A., Minoda A., Tolstorukov M. Y., Appert A., Parker S. C. J., Gu T., Kundaje A., Riddle N. C., Bishop E., Egelhofer T. A., Hu S. S., Alekseyenko A. A., Rechtsteiner A., Asker D., Belsky J. A., Bowman S. K., Chen Q. B., Chen R. A. J., Day D. S., Dong Y., Dose A. C., Duan X., Epstein C. B., Ercan S., Feingold E. A., Ferrari F., Garrigues J. M., Gehlenborg N., Good P. J., Haseley P., He D., Herrmann M., Hoffman M. M., Jeffers T. E., Kharchenko P. V., Kolasinska-Zwierz P., Kotwaliwale C. V., Kumar N., Langley S. A., Larschan E. N., Latorre I., Libbrecht M. W., Lin X., Park R., Pazin M. J., Pham H. N., Plachetka A., Qin B., Schwartz Y. B., Shoresh N., Stempor P., Vielle A., Wang C., Whittle C. M., Xue H., Kingston R. E., Kim J. H., Bernstein B. E., Dernburg A. F., Pirrotta V., Kuroda M. I., Noble W. S., Tullius T. D., Kellis M., MacAlpine D. M., Strome S., Elgin S. C. R., Liu X. S., Lieb J. D., Ahringer J., Karpen G. H., Park P. J., 2014 Comparative analysis of metazoan chromatin organization. Nature **512**: 449–452.

Hoang M. L., Kinde I., Tomasetti C., McMahon K. W., Rosenquist T. A., Grollman A. P., Kinzler K. W., Vogelstein B., Papadopoulos N., 2016 Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing.

Proceedings of the National Academy of Sciences of the United States of America **113**: 9846–9851.

Hodgkinson A., Eyre-Walker A., 2011 Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. **12**: 756–766.

Hwang D. G., Green P., 2004 Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America **101**: 13994–14001.

Ikehata H., Ono T., 2007 Significance of CpG Methylation for Solar UV-Induced Mutagenesis and Carcinogenesis in Skin‡. Photochemistry and Photobiology **83**: 196–204.

Inamdar N. M., Zhang X.-Y., Brough C. L., Gardiner W. E., Bisaro D. M., Ehrlich M., 1992 Transfection of heteroduplexes containing uracil · guanine or thymine · guanine mispairs into plant cells. Plant Mol Biol **20**: 123–131.

Josse J., Kaiser A. D., Kornberg A., 1961 Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. The Journal of Biological Chemistry **236**: 864–875.

Jónsson H., Sulem P., Kehr B., Kristmundsdottir S., Zink F., Hjartarson E., Hardarson M. T., Hjorleifsson K. E., Eggertsson H. P., Gudjonsson S. A., Ward L. D., Arnadottir G. A., Helgason E. A., Helgason H., Gylfason A., Jonasdottir A., Jonasdottir A., Rafnar T., Frigge M., Stacey S. N., Th Magnusson O., Thorsteinsdottir U., Masson G., Kong A., Halldorsson B. V., Helgason A., Gudbjartsson D. F., Stefansson K., 2017 Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature **549**: 519–522.

Karczewski K. J., Francioli L. C., Tiao G., Cummings B. B., Alföldi J., Wang Q., Collins R. L., Laricchia K. M., Ganna A., Birnbaum D. P., Gauthier L. D., Brand H., Solomonson M., Watts N. A., Rhodes D., Singer-Berk M., Seaby E. G., Kosmicki J. A., Walters R. K., Tashman K., Farjoun Y., Banks E., Poterba T., Wang A., Seed C., Whiffin N., Chong J. X., Samocha K. E., Pierce-Hoffman E., Zappala Z., O'Donnell-Luria A. H., Minikel E. V., Ben Weisburd, Lek M., Ware J. S., Vittal C., Armean I. M., Bergelson L., Cibulskis K., Connolly K. M., Covarrubias M., Donnelly S., Ferriera S., Gabriel S., Gentry J., Gupta N., Jeandet T., Kaplan D., Llanwarne C., Munshi R., Novod S., Petrillo N., Roazen D., Ruano-Rubio V., Saltzman A., Schleicher M., Soto J., Tibbetts K., Tolonen C., Wade G., Talkowski M. E., Consortium T. G. A. D., Neale B. M., Daly M. J., MacArthur D. G., 2019 Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv: 531210.

Kong A., Frigge M. L., Masson G., Besenbacher S., Sulem P., Magnusson G., Gudjonsson S. A., Sigurdsson A., Jonasdottir A., Jonasdottir A., Wong W. S. W., Sigurdsson G., Walters G. B., Steinberg S., Helgason H., Thorleifsson G., Gudbjartsson D. F., Helgason A., Magnusson O. T., Thorsteinsdottir U., Stefansson K., 2012 Rate of de novo mutations and the importance of father's age to disease risk. Nature **488**: 471–475.

Krokan H. E., Bjoras M., 2013 Base Excision Repair. Cold Spring Harbor Perspectives in Biology **5**: a012583.

Leah J DeRose-Wilson B. S. G., 2007 Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata*. BMC Evol. Biol. **7**: 66.

Lee H., Popodi E., Tang H., Foster P. L., 2012 Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proceedings of the National Academy of Sciences of the United States of America **109**: E2774–E2783.

Lee J., Jang H., Shin H., Choi W. L., Mok Y. G., Huh J. H., 2014 AP endonucleases process 5-methylcytosine excision intermediates during active DNA demethylation in Arabidopsis. Nucleic Acids Research **42**: 11408–11418.

Lek M., Karczewski K. J., Minikel E. V., Samocha K. E., Banks E., Fennell T., O'Donnell-Luria A. H., Ware J. S., Hill A. J., Cummings B. B., Tukiainen T., Birnbaum D. P., Kosmicki J. A., Duncan L. E., Estrada K., Zhao F., Zou J., Pierce-Hoffman E., Berghout J., Cooper D. N., Deflaux N., DePristo M., Do R., Flannick J., Fromer M., Gauthier L., Goldstein J., Gupta N., Howrigan D., Kiezun A., Kurki M. I., Moonshine A. L., Natarajan P., Orozco L., Peloso G. M., Poplin R., Rivas M. A., Ruano-Rubio V., Rose S. A., Ruderfer D. M., Shakir K., Stenson P. D., Stevens C., Thomas B. P., Tiao G., Tusie-Luna M. T., Weisburd B., Won H.-H., Yu D., Altshuler D. M., Ardissino D., Boehnke M., Danesh J., Donnelly S., Elosua R., Florez J. C., Gabriel S. B., Getz G., Glatt S. J., Hultman C. M., Kathiresan S., Laakso M., McCarroll S., McCarthy M. I., McGovern D., McPherson R., Neale B. M., Palotie A., Purcell S. M., Saleheen D., Scharf J. M., Sklar P., Sullivan P. F., Tuomilehto J., Tsuang M. T., Watkins H. C., Wilson J. G., Daly M. J., MacArthur D. G., 2016 Analysis of protein-coding genetic variation in 60,706 humans. Nature **536**: 285–291.

Leroy J. L., Kochoyan M., Huynh-Dinh T., Guéron M., 1988 Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. Journal of Molecular Biology **200**: 223–238.

Li E., Zhang Y., 2014 DNA Methylation in Mammals. Cold Spring Harbor Perspectives in Biology **6**: a019133.

Lin J.-Y., Le B. H., Chen M., Henry K. F., Hur J., Hsieh T.-F., Chen P.-Y., Pelletier J. M., Pellegrini M., Fischer R. L., Harada J. J., Goldberg R. B., 2017 Similarity between soybean and Arabidopsis seed methylomes and loss of non-CG methylation does not affect seed development. Proceedings of the National Academy of Sciences of the United States of America **114**: E9730–E9739.

Lister R., Pelizzola M., Dowen R. H., Hawkins R. D., Hon G., Tonti-Filippini J., Nery J. R., Lee L., Ye Z., Ngo Q.-M., Edsall L., Antosiewicz-Bourget J., Stewart R., Ruotti V., Millar A. H., Thomson J. A., Ren B., Ecker J. R., 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. Nature **462**: 315–322.

Liu Y., Tian T., Zhang K., You Q., Yan H., Zhao N., Yi X., Xu W., Su Z., 2018 PCSD: a plant chromatin state database. Nucleic Acids Research **46**: D1157–D1167.

Lutsenko E., Bhagwat A. S., 1999 Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. Mutat. Res. **437**: 11–20.

Lyons D. M., O'Brien P. J., 2010 Human base excision repair creates a bias toward -1 frameshift mutations. Journal of Biological Chemistry **285**: 25203–25212.

Makova K. D., Hardison R. C., 2015 The effects of chromatin organization on variation in mutation rates in the genome. Nat. Rev. Genet. **16**: 213–223.

Marcourt L., Cordier C., Couesnon T., Dodin G., 1999 Impact of C5-cytosine methylation on the solution structure of d(GAAAACGTTTTC)2. An NMR and molecular modelling investigation. Eur J Biochem **265**: 1032–1042.

Martincorena I., Fowler J. C., Wabik A., Lawson A. R. J., Abascal F., Hall M. W. J., Cagan A., Murai K., Mahbubani K., Stratton M. R., Fitzgerald R. C., Handford P. A., Campbell P. J., Saeb-Parsy K., Jones P. H., 2018 Somatic mutant clones colonize the human esophagus with age. Science **362**: 911–917.

Martínez-Macías M. I., Córdoba-Cañero D., Ariza R. R., Roldán-Arjona T., 2013 The DNA repair protein XRCC1 functions in the plant DNA demethylation pathway by stimulating cytosine methylation (5-meC) excision, gap tailoring, and DNA ligation. Journal of Biological Chemistry **288**: 5496–5505.

Molaro A., Hodges E., Fang F., Song Q., McCombie W. R., Hannon G. J., Smith A. D., 2012 Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates. Cell **146**: 1029–1041.

Mugal C. F., Ellegren H., 2011 Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. Genome Biol. **12**: R58.

Murchie A. I. H., Lilley D. M. J., 1989 Base methylation and local DNA helix stability: Effect on the kinetics of cruciform extrusion. Journal of Molecular Biology **205**: 593–602.

Ngo T. T. M., Yoo J., Dai Q., Zhang Q., He C., Aksimentiev A., Ha T., 2016 Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. Nature Communications **7**: 10813.

Ossowski S., Schneeberger K., Lucas-Lledó J. I., Warthmann N., Clark R. M., Shaw R. G., Weigel D., Lynch M., 2010 The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. Science **327**: 92–94.

Peña-Diaz J., Bregenhorn S., Ghodgaonkar M., Follonier C., Artola-Borán M., Castor D., Lopes M., Sartori A. A., Jiricny J., 2012 Noncanonical Mismatch Repair as a Source of Genomic Instability in Human Cells. Molecular Cell **47**: 669–680.

Prendergast J. G. D., Chambers E. V., Semple C. A. M., 2014 Sequence-level mechanisms of human epigenome evolution. Genome Biol Evol **6**: 1758–1771.

Qu W., Hashimoto S. I., Shimada A., Nakatani Y., Ichikawa K., Saito T. L., Ogoshi K., Matsushima K., Suzuki Y., Sugano S., Takeda H., Morishita S., 2012 Genome-wide

genetic variations are highly correlated with proximal DNA methylation patterns. Genome Research **22**: 1419–1425.

Rahbari R., Wuster A., Lindsay S. J., Hardwick R. J., Alexandrov L. B., Turki S. A., Dominiczak A., Morris A., Porteous D., Smith B., Stratton M. R., UK10K Consortium, Hurles M. E., 2016 Timing, rates and spectra of human germline mutation. Nat Genet **48**: 126–133.

Regev A., Lamb M. J., Jablonka E., 1998 The role of DNA methylation in invertebrates: Developmental regulation or genome defense? Mol Biol Evol **15**: 880–891.

Russell G. J., Walker P. M. B., Elton R. A., Subak-Sharpe J. H., 1976 Doublet frequency analysis of fractionated vertebrate nuclear DNA. Journal of Molecular Biology **108**: 1–20.

Salser W., 1978 Globin mRNA Sequences: Analysis of Base Pairing and Evolutionary Implications. Cold Spring Harbor Symposia on Quantitative Biology **42**: 985–1002.

Sancar A., 1996 DNA Excision Repair. Annu. Rev. Biochem. **65**: 43–81.

Schanz S., Castor D., Fischer F., Jiricny J., 2009 Interference of mismatch and base excision repair during the processing of adjacent U/G mispairs may play a key role in somatic hypermutation. Proceedings of the National Academy of Sciences of the United States of America **106**: 5593–5598.

Schmutte C., Yang A. S., Beart R. W., Jones P. A., 1995 Base excision repair of U:G mismatches at a mutational hotspot in the p53 gene is more efficient than base excision repair of T:G mismatches in extracts of human colon tumors. Cancer Res **55**: 3742–3746.

Schuster-Böckler B., Lehner B., 2012 Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature **488**: 504–507.

Severin P. M. D., Zou X., Gaub H. E., Schulten K., 2011 Cytosine methylation alters DNA mechanical properties. Nucleic Acids Research **39**: 8740–8751.

Ségurel L., Wyman M. J., Przeworski M., 2014 Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet **15**: 47–70.

Shen J.-C., Rideout W. M., Jones P. A., 1994 The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Research **22**: 972–976.

Simmen M. W., 2008 Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. Genomics **92**: 33–40.

Stoltzfus A., McCandlish D. M., 2017 Mutational Biases Influence Parallel Adaptation. Mol Biol Evol **34**: 2163–2172.

Storz J. F., Natarajan C., Signore A. V., Witt C. C., McCandlish D. M., Stoltzfus A., 2019 The role of mutation bias in adaptive molecular evolution: Insights from convergent changes in protein function. Philosophical Transactions of the Royal Society B: Biological Sciences **374**: 20180238.

Stroud H., Ding B., Simon S. A., Feng S., Bellizzi M., Pellegrini M., Wang G.-L., Meyers B. C., Jacobsen S. E., 2013 Plants regenerated from tissue culture contain stable epigenome changes in rice. eLife **2**: 245.

Supek F., Ben Lehner, 2015 Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature **521**: 81–84.

Supek F., Ben Lehner, Hajkova P., Warnecke T., 2014 Hydroxymethylated Cytosines Are Associated with Elevated C to G Transversion Rates (L Duret, Ed.). PLoS Genet. **10**: e1004585.

Szer W., Shugar D., 1966 The structure of poly-5-methylcytidylic acid and its twin-stranded complex with poly-inosinic acid. Journal of Molecular Biology **17**: 174–187.

Tomkova M., Schuster-Böckler B., 2018 DNA Modifications: Naturally More Error Prone? Trends in Genetics **34**: 627–638.

Tomkova M., McClellan M., Kriaucionis S., Schuster-Boeckler B., 2016 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. eLife **5**: 415.

Tommasi S., Denissenko M. F., Pfeifer G. P., 1997 Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. Cancer Res **57**: 4727–4730.

Tornaletti S., Pfeifer G. P., 1996 UV damage and repair mechanisms in mammalian cells. BioEssays **18**: 221–228.

Walser J.-C., Ponger L., Furano A. V., 2008 CpG dinucleotides and the mutation rate of non-CpG DNA. Genome Research **18**: 1403–1414.

Wang R. Y. H., Kuo K. C., Gehrke C. W., Huang L.-H., Ehrlich M., 1982 Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression **697**: 371–377.

Weng M.-L., Becker C., Hildebrandt J., Neumann M., Rutter M. T., Shaw R. G., Weigel D., Fenster C. B., 2019 Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in *Arabidopsis thaliana*. Genetics **211**: 703–714.

Wierzbicki A. T., Cocklin R., Mayampurath A., Lister R., Rowley M. J., Gregory B. D., Ecker J. R., Tang H., Pikaard C. S., 2012 Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. Genes & Development **26**: 1825–1836.

Xia J., Han L., Zhao Z., 2012 Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. BMC Genomics **13**: S7.

Zemach A., Kim M. Y., Hsieh P.-H., Coleman-Derr D., Eshed-Williams L., Thao K., Harmer S. L., Zilberman D., 2013 The Arabidopsis Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. Cell **153**: 193–205.

Zhang H., Lang Z., Zhu J.-K., 2018 Dynamics and function of DNA methylation in plants. Nature Publishing Group **19**: 489–506.

Zhang X., Mathews C. K., 1994 Effect of DNA cytosine methylation upon deamination-induced mutagenesis in a natural target sequence in duplex DNA. The Journal of Biological Chemistry **269**: 7066–7069.

Zhao Z., Boerwinkle E., 2002 Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. Genome Research **12**: 1679–1686.

Zhu J.-K., 2009 Active DNA Demethylation Mediated by DNA Glycosylases. Annu. Rev. Genet. **43**: 143–166.

Zhu Y. O., Sherlock G., Petrov D. A., 2017 Extremely Rare Polymorphisms in Saccharomyces cerevisiae Allow Inference of the Mutational Spectrum (SR Sunyaev, Ed.). PLoS Genet. **13**: e1006455.

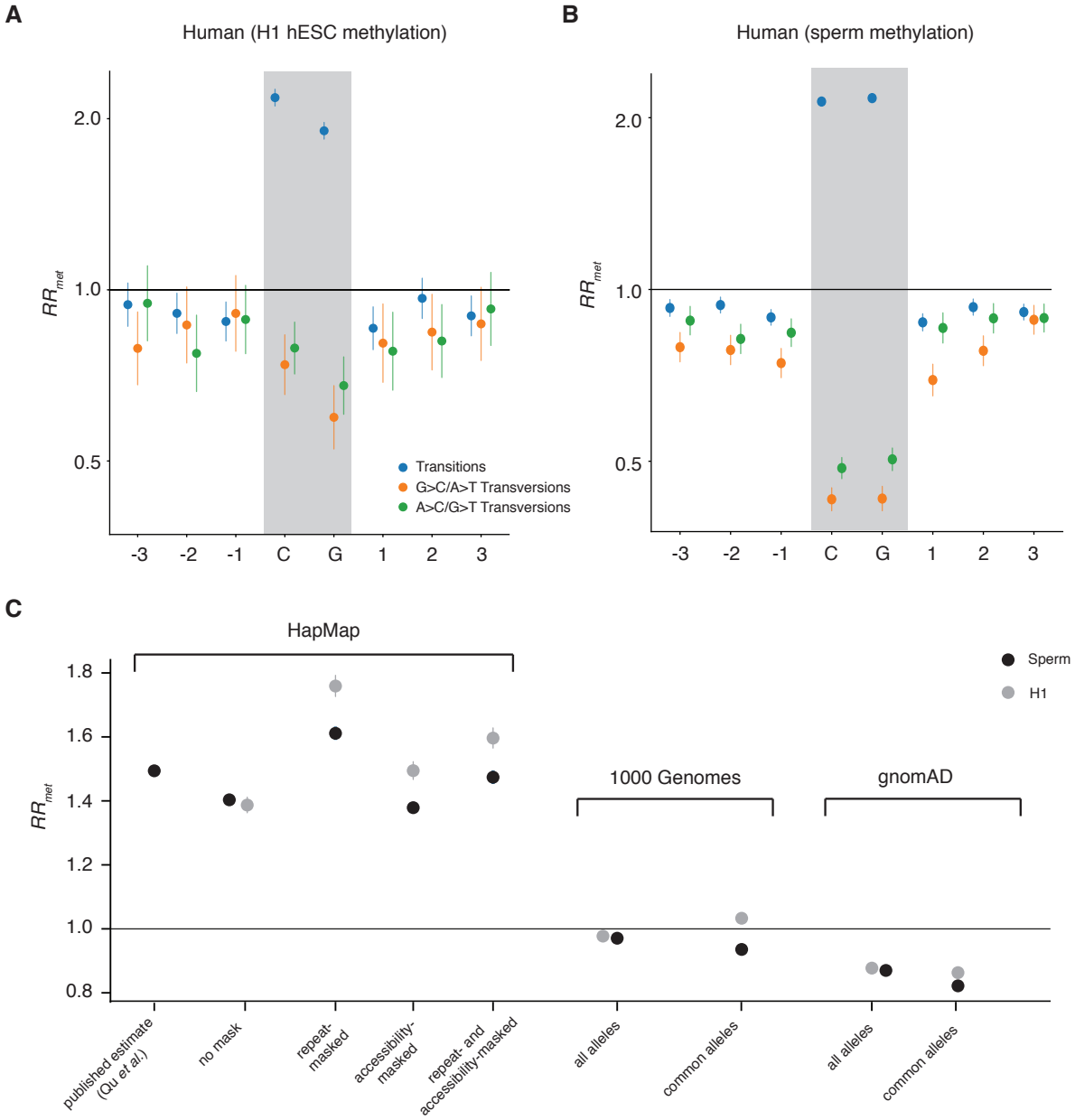


Figure 1. The relative mutational risk of methylation (RR_{met}) at CpGs and neighbouring nucleotides in human, when site pairing is based on methylation data from **(A)** H1 hESCs or **(B)** sperm and considering singleton SNPs from the gnomAD database. **(C)** The impact of analytical and dataset choice on RR_{met}estimates. Polymorphism data from the HapMap project (CEU) consistently lead to RR_{met}estimates >1, which are robust to inclusion of SNPs from poorly accessible or repeat regions. RR_{met}estimates are below one when considering SNPs from the 1000 Genomes Project or gnomAD. Common alleles are defined as alleles with a frequency of ≥5% in the population sample. All P<10-8. Vertical bars are confidence intervals on RR_{met}computed using the delta method.

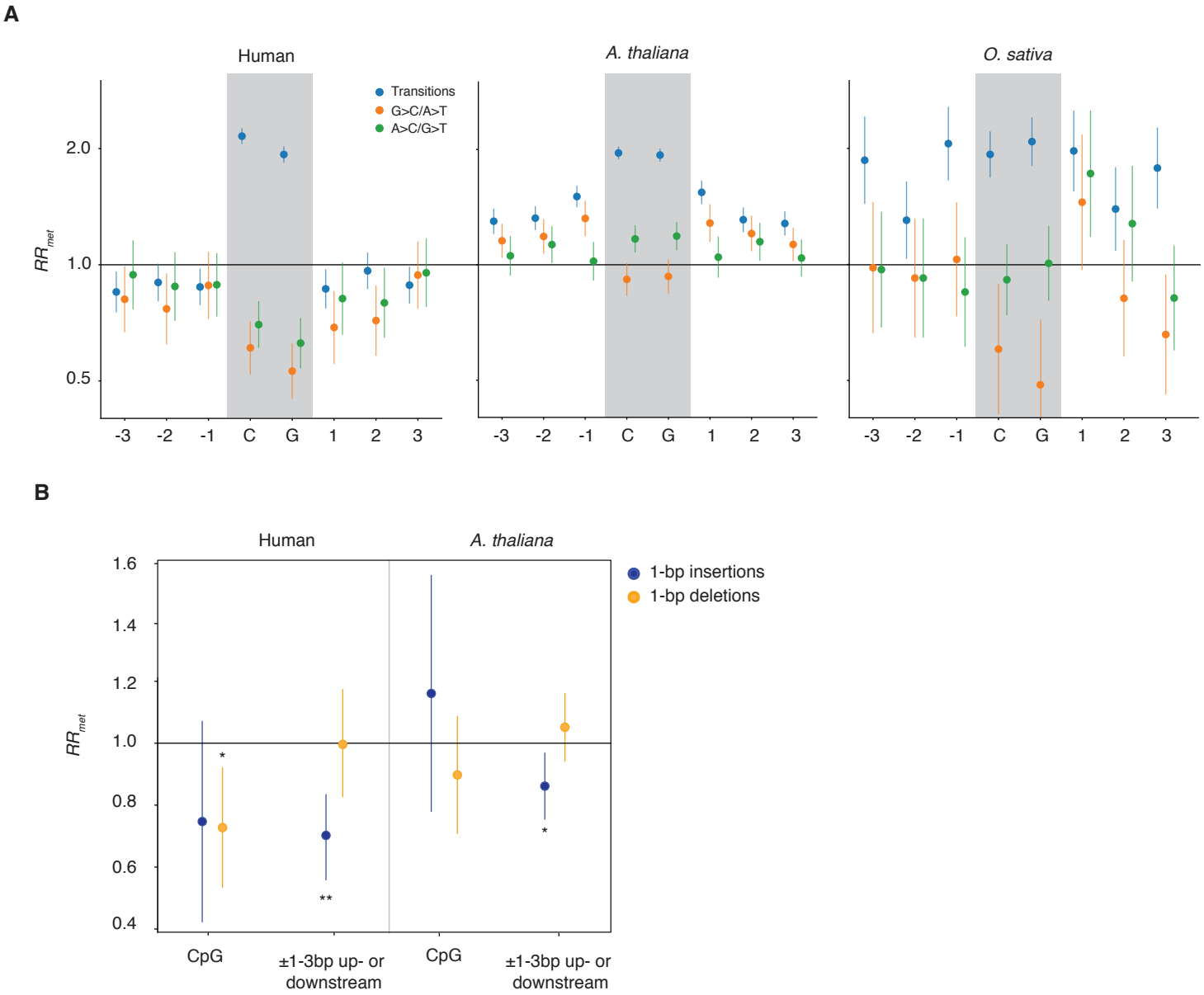


Figure 2. (**A**) The relative mutational risk of methylation (RR_{met}) at CpGs and neighbouring nucleotides in human, *A. thaliana*, and rice. For human, pairing is based on H1 hESC methylation data. In contrast to Figure 1, only SNPs in repeat-masked sequence was included for all species. (**B**) RR_{met} as applied to indels in human and *A. thaliana*. Vertical bars are confidence intervals on RR_{met} computed using the delta method. *P<0.05, **P<0.005

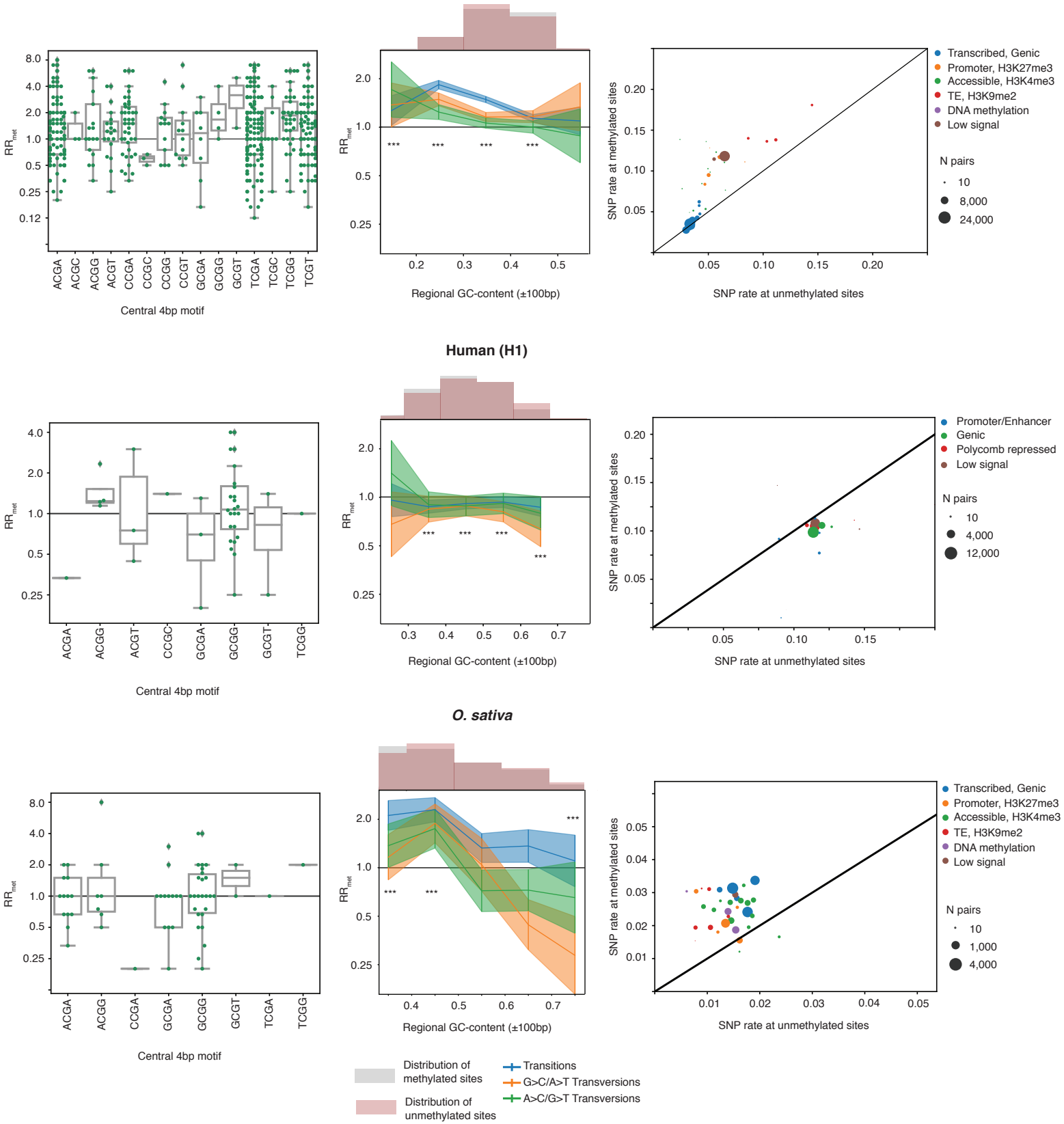


Figure 3. The relative mutational risk of methylation (RR_{met}) at CpG-neighbouring nucleotides in human, *A. thaliana*, and rice as a function of internal (±1bp) sequence context (left-hand panels), regional (±100bp) GC content (central panels), and chromatin state (right-hand panels). To calculate RR_{met} for motifs with a given internal sequence context, only motifs with a minimum number of 20 pairs were considered. To calculate dependency on GC content, methylated and unmethylated members of a given pair were independently binned based on their regional GC content and RR_{met} calculated based on these bins. The distribution of regional GC content for methylated and unmethylated sites is almost identical. ***P<0.001. Vertical bars are confidence intervals on RR_{met} computed using the delta method. RR_{met} for different chromatin states, deconstructed into SNP rates at methylated and unmethylated sites across pairs, is colour-coded according to broader overarching categories of similar states whose principal features are highlighted. See Figure S4 for fully annotated states.

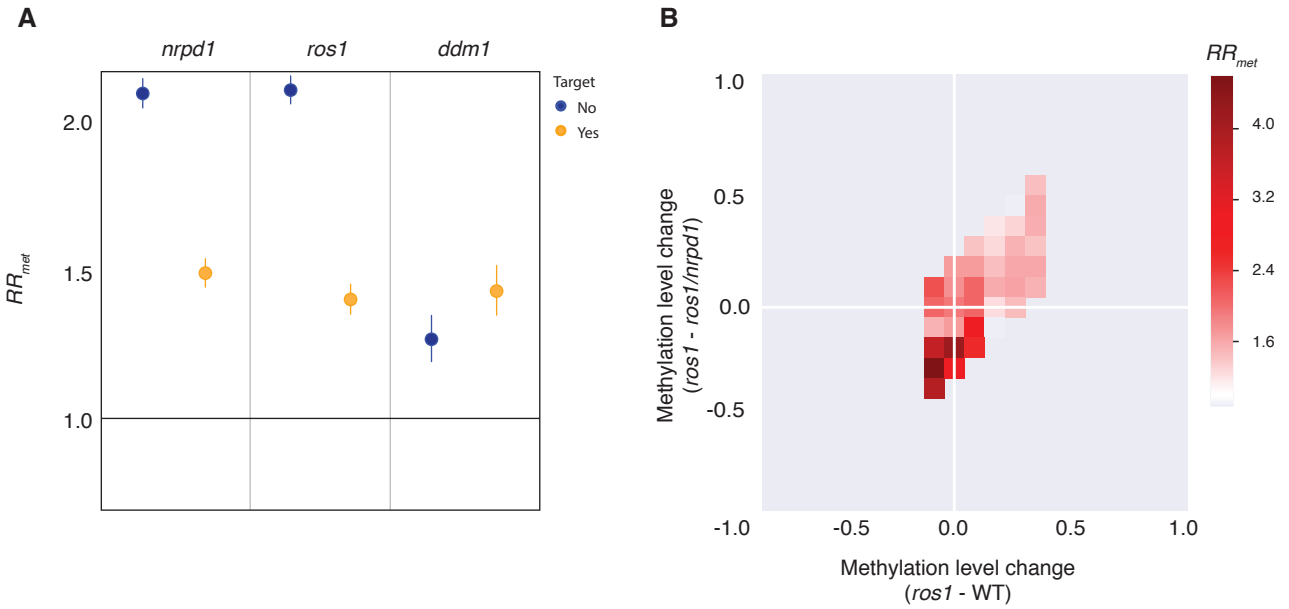


Figure 4. (**A**) The relative mutational risk of methylation (RR_{met}) at CpGs and neighbouring nucleotides in *A. thaliana* as a function of whether a given site is targeted by RNA-directed DNA methylation (which involves *nrpd1*), the DNA glycosylase ROS1, or the chromatin remodeler DDM1 (see main text for how targets and non-targets were defined). All $P<10^{-60}$. Vertical bars are confidence intervals on RR_{met} computed using the delta method. (**B**) RR_{met} as a function of relative methylation change in *ros1* deletion mutants compared to the corresponding wildtype strain and compared to the *nrpd1/ros1* double mutants. Sites with more methylation in *ros1* compared to the WT are sites of high ROS1 activity, where methylation is normally erased by ROS1 but increases in the *ros1* knock-out. Sites with more methylation in *ros1* compared to *nrpd1/ros1* are sites where the RNA-directed DNA methylation machinery is active, i.e. where deletion of *nrpd1* leads to further loss of methylation.

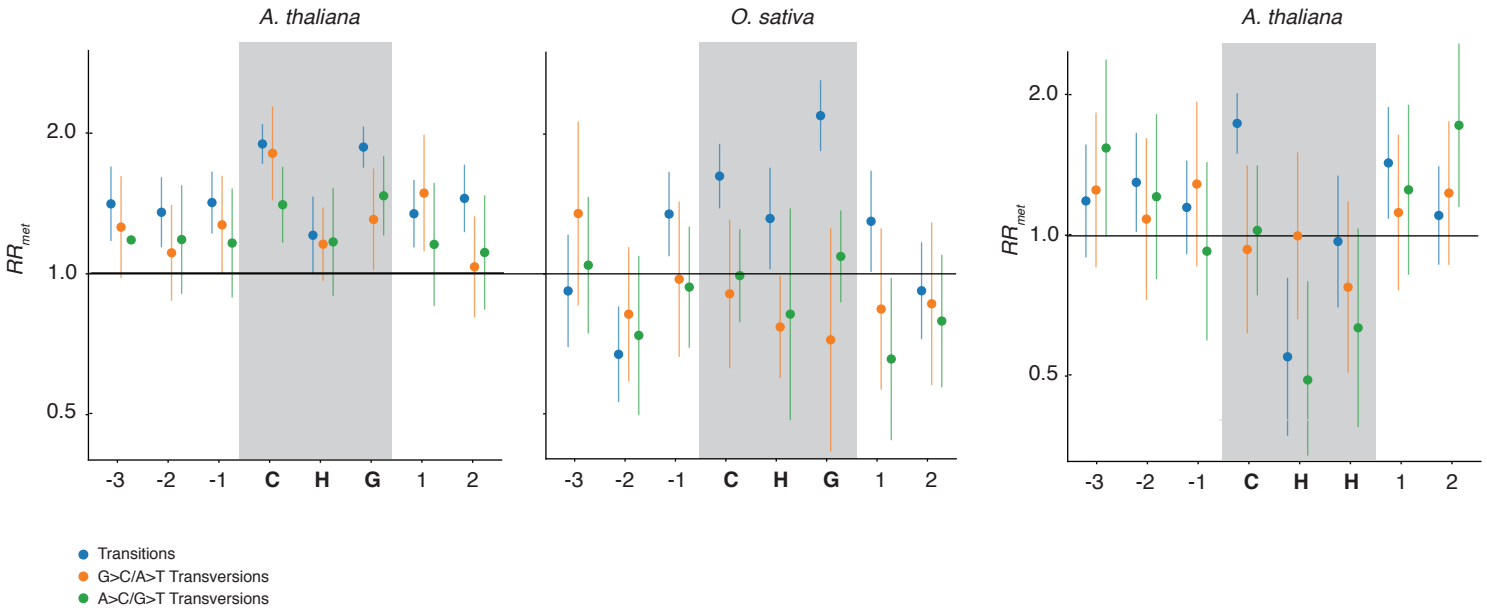


Figure S1. The relative mutational risk of methylation (RR_{met}) at CpHpG an CpHpH sites and neighbouring nucleotides in *A. thaliana* and rice. RR_{met} estimates for CpHpH in rice are too noisy to be informative and are therefore not shown. P-values for overall RR_{met} estimates are: CpHpG (*A. thaliana*): $P=3.8*10^{-19}$; CpHpH (*A. thaliana*): $P=1.8*10^{-8}$; CpHpG (*O. sativa*): P=0.043. Vertical bars are confidence intervals on RR_{met} computed using the delta method.

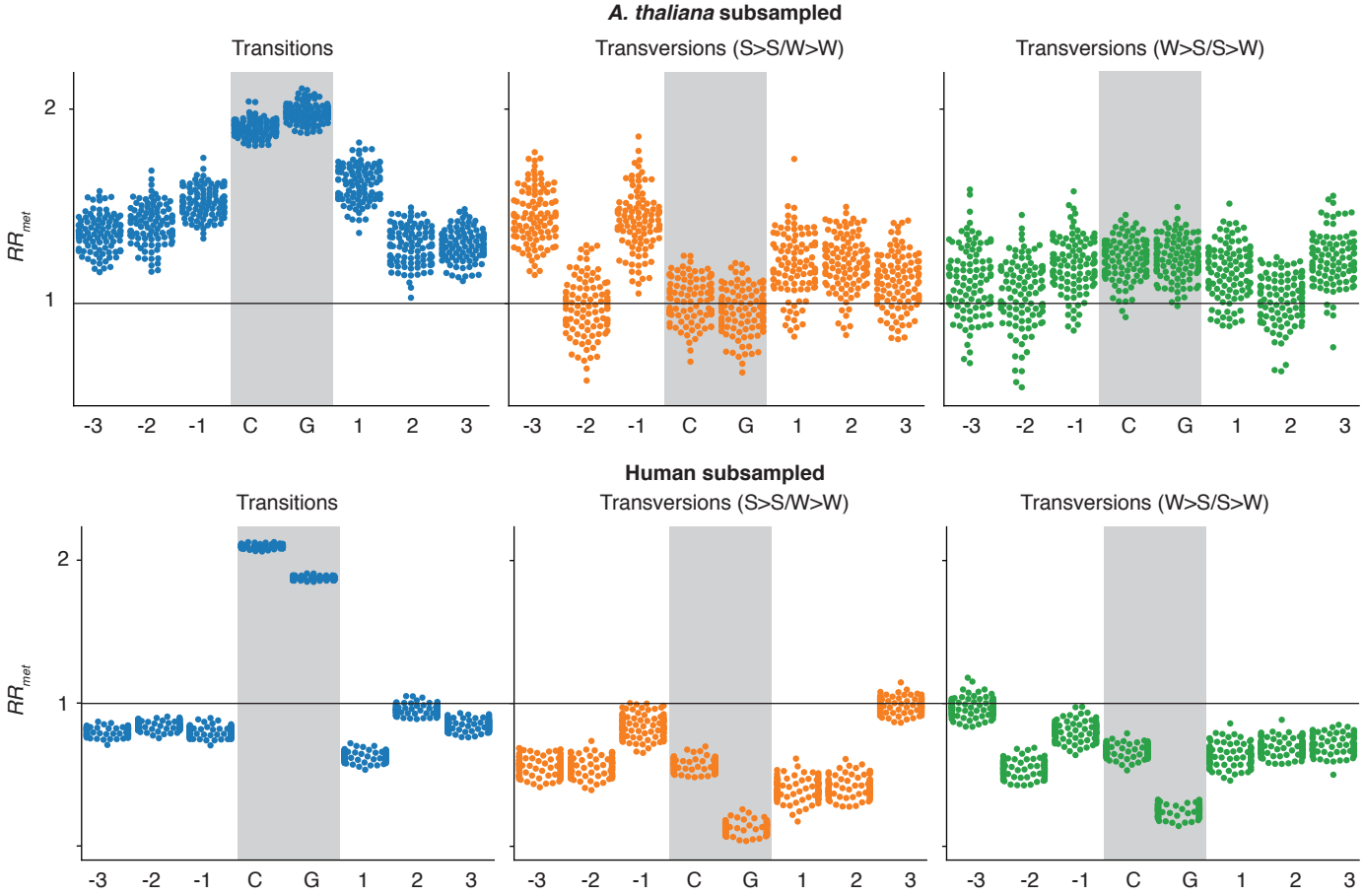


Figure S2. RR_{met} of subsampled motifs. The number of occurrences for each observed nucleotide motif was counted in both human and *A. thaliana* motif pairs. For each motif, the occurrences in both sets were compared, and the set with the larger number of said motif was randomly subsampled without replacement to contain the same number as the smaller sample. After doing this for all motifs, we have a reduced set of human and *A. thaliana* methylated and unmethylated motifs with identical motif distribution. RR_{met} was then calculated as described. This process was repeated 100 times, to account for biases in random sampling.

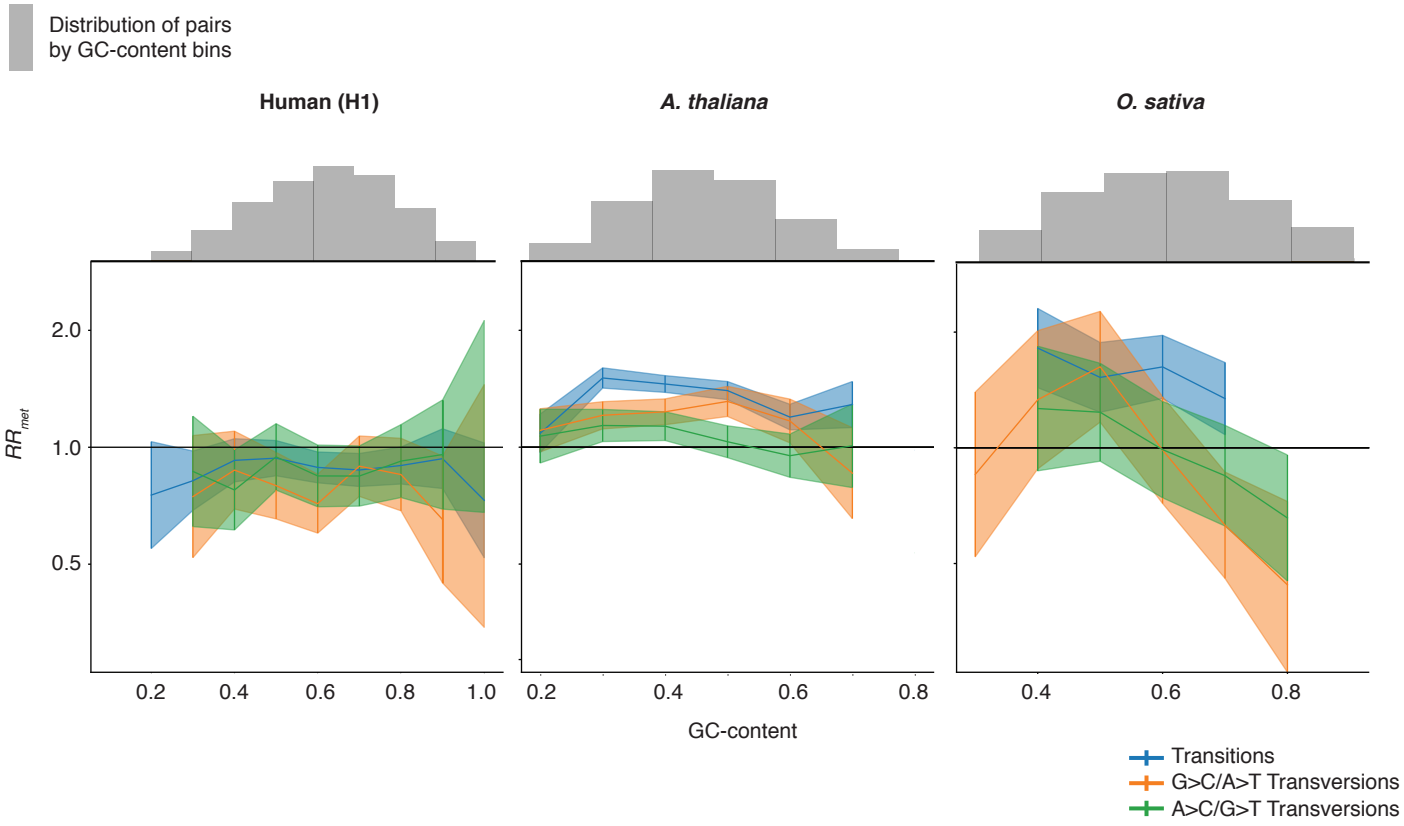
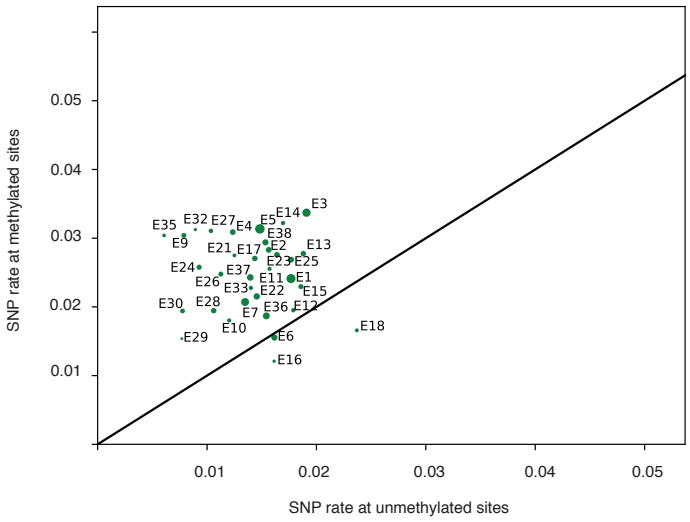


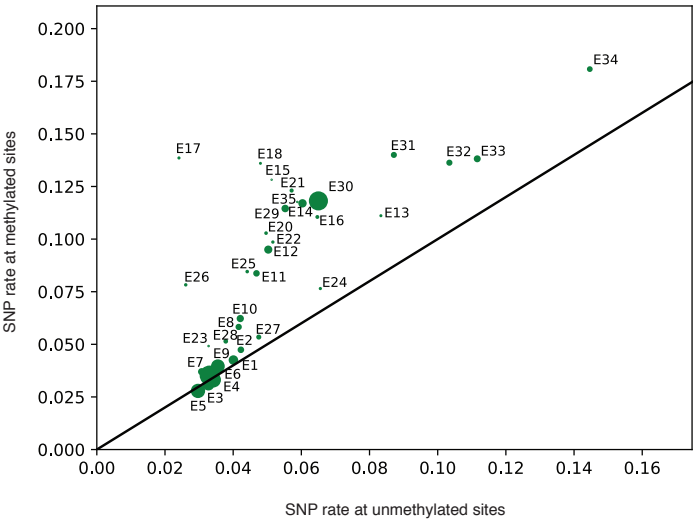
Figure S3. The relative mutational risk of methylation (RR_{met}) at CpG-neighbouring nucleotides in human, *A. thaliana*, and rice as a function of motif GC content. RR_{met} was calculated based on GC content bins. RR_{met} estimates are omitted where error bars were larger than the plotting range. Vertical bars are confidence intervals on RR_{met} computed using the delta method.

O. sativa



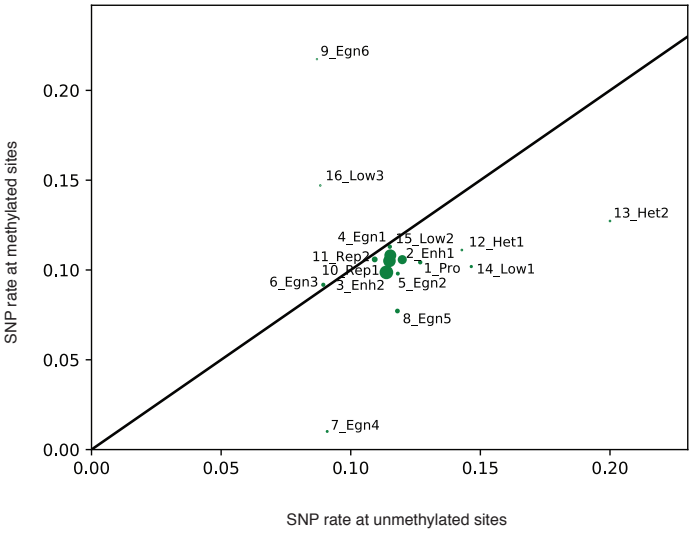
State	Preferential epigenetics marks	Preferential location
State 1	weak histone acetylation	intron,promoter,3'UTR,intergenic
State 2	H3K36me1,histone acetylation	promoter,3'UTR,intergenic
State 3	H3K36me2,histone acetylation	3'UTR
State 4	H3K36me2,H3K36me3,H3K4me2	exon
State 5	H3K36me3,H3K4me2	intron
State 6	H3K27me3	intergenic,intron,promoter
State 7	H3K27me3	intergenic,intron,promoter
State 8	H3K27me3	intergenic,intron,promoter
State 9	H3K27me3,H2A.Z	intergenic,intron,promoter
State 10	H3K27me3,H2A.Z,H3K9ac,H3K23ac,H3K4me3,H3K4me2	exon
State 11	H3K9ac,H3K23ac,H3K4me3,H3K4me2,H3K27me3,H2A.Z	exon,promoter
State 12	H2A.Z,histone acetylation,H3K4me2,H3K4me3	exon,promoter
State 13	histone acetylation,H3K4me3,H3K4me2,H3K36me3,H3K36me2,H2A.Z	5'UTR
State 14	H3K4me3,H3K4me3,H3K36me3,H2A.Z,histone acetylation	intron
State 15	H3K4me3,H3K4me3,H3K36me3	intron
State 16	H3K4me3,histone acetylation,H2A.Z	exon,promoter,5'UTR
State 17	histone acetylation,accessible DNA,H3K4me2,H3K4me3	promoter,exon,5'UTR
State 18	histone acetylation,accessible DNA,H2A.Z,H3K4me3	5'UTR,promoter,intergenic
State 19	DNA methylation,H3K4me3,H4K12ac,H4K16ac,OsNF-YB1,OsbZIP	centromere,intergenic
State 20	accessible DNA,H2A.Z,histone acetylation,H3K27me3,DNA methylation	intergenic,promoter,intron
State 21	accessible DNA	intergenic,promoter
State 22	accessible DNA	intergenic,promoter
State 23	accessible DNA	intergenic,promoter
State 24	accessible DNA	intergenic,promoter
State 25	H2A.Z,H3K4me2,H3K4me3,histone acetylation	promoter,intergenic,intron,3'UTR
State 26	DNA methylation,OsNF-YB1	intergenic,promoter,intron
State 27	H3K9me2,DNA methylation,OsNF-YB1	intergenic,TE,promoter,exon
State 28	H3K9me2,DNA methylation,Mnase	TE,intergenic,intron,promoter,exon
State 29	DNA methylation,H3K9me2,CENH3,Mnase	TE,exon,promoter,intron,intergenic
State 30	H3K9me2,DNA methylation	TE,exon,intron,promoter,intergenic
State 31	H3K9me2,CENH3,DNA methylation	TE,intron,intergenic,exon,promoter,
State 32	H3K9me2,CENH3,DNA methylation,Mnase	TE,intergenic,intron,exon,promoter
State 33	DNA methylation,H3K9me2	TE,exon,intergenic,promoter,intron
State 34	DNA methylation,H3K36,H3K4me2	intron
State 35	DNA methylation,H3K27me3,H3K36me1,H3K36me2,H3K36me3	intergenic,intron,promoter
State 36	DNA methylation,H3K36me1	intergenic,promoter,intron
State 37	DNA methylation	intergenic,intron,promoter
State 38	rare signal	intergenic,promoter,intron

A. thaliana



State	Preferential epigenetics marks	Preferential location
State 1	H3.3	3'UTR
State 2	H3.3, histone acetylation,H3K4me2,H2A.Z	CDS,3'UTR
State 3	H3K4me1,H3.3,H3.1	CDS
State 4	H3K4me1,H3.3	CDS,intron
State 5	H3K4me1,H3K36me3,H3.3,H3.1	CDS
State 6	H3K4me1,H3K36me3	intron
State 7	H3K4me1,H3K36me3,H3K4me2	CDS,intron
State 8	H3K4me1,H3K4me2,H2A.Z	CDS
State 9	H3K4me1	intron
State 10	H2A.Z	CDS,intron
State 11	H3K27me3,H2A.Z,H3K4me2	CDS
State 12	H3K27me3,H2A.Z	Promoter,CDS,intron,intergenic
State 13	H3K27me3,H2A.Z	Promoter,intergenic
State 14	H3K27me3	Promoter,intergenic
State 15	H3K27me3,accessible DNA	Promoter,intergenic
State 16	accessible DNA	Promoter,intergenic
State 17	accessible DNA	Promoter
State 18	accessible DNA	Promoter
State 19	accessible DNA	Promoter,snRNA
State 20	accessible DNA	Promoter
State 21	accessible DNA	Promoter
State 22	histone acetylation,H3K4me2	coding gene,miRNA,snoRNA
State 23	accessible DNA,H3K36ac,H3K56ac,H4K16ac,H3K4me3	Promoter,5'UTR
State 24	accessible DNA,histone acetylation,H3K4me3	5'UTR,snoRNA
State 25	histone acetylation,H3K4me3,H3K4me2,H2A.Z	intron
State 26	histone acetylation,H3K4me3,H3K4me2,H2A.Z	CDS
State 27	H3K4me2,histone acetylation,H3K4me3,H2A.Z	CDS
State 28	H3K4me3,H3K4me2,H2A.Z	intron
State 29	weak signal	intergenic
State 30	rare signal	intergenic
State 31	DNA methylation,H3K9me2,H3K27me3	intergenic,miRNA
State 32	DNA methylation,H3K9me2	intergenic,TE
State 33	H3K9me2,DNA methylation	TE
State 34	H3K9me2,DNA methylation,H3K27me1	TE,miRNA
State 35	H3K9me2,DNA methylation,H2A.X	intergenic,pericentromere
State 36	CENH3,H3K9me2,DNA methylation,accessible DNA	rRNA,tRNA,centromere

Human (H1)



State	Preferential location
State 1 Pro	Promoter
State 2 Enh1	Enhancer
State 3 Enh1	Enhancer
State 4 Egn1	Expressed gene
State 5 Egn2	Expressed gene
State 6 Egn3	Expressed gene
State 7 Egn4	Expressed gene
State 8 Egn5	Expressed gene
State 9 Egn6	Expressed gene
State 10 Rep1	Polycomb repressed
State 11 Rep2	Polycomb repressed
State 12 Het1	Heterochromatin
State 13 Het2	Heterochromatin
State 14 Low1	Low signal
State 15 Low2	Low signal
State 16 Low3	Low signal

Figure S4. RR_{met} in *A. thaliana*, human, and rice for different chromatin states, deconstructed into SNP rates at methylated and unmethylated sites across pairs. Individual chromatin states, classified as described in the Methods, are labelled. Text colour corresponds to the colouring of chromatin states in Figure 3.