Sequence context and methylation interact to shape germline mutation rate variation at CpG sites

Sheel Chandra[1], Ziyue Gao[2*]

[1] Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

[2] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

* Corresponding author

E-mail: ziyuegao@pennmedicine.upenn.edu

Abstract

A prominent example of sequence context-dependent mutation rate variation is the elevated transition rate at CpG sites, which is largely attributed to cytosine methylation. CpGs with different flanking sequences also exhibit mutation rate variation, but this variation is only partially correlated with context-specific methylation level. Here, we quantify the CpG mutation rate and mutagenic effect of methylation across sequence contexts. Using a regression framework that accounts for recurrent mutations, we analyze human polymorphisms from the gnomAD dataset to estimate mutation rates of unmethylated and methylated CpGs in each unique 4-mer or 6-mer context. We find that CpG mutation rate variation in the human genome is shaped by methylation at the focal cytosine, the flanking nucleotides, and interactions between them. Our analysis reveals distinct context-dependent mutation patterns for unmethylated and methylated cytosines, driven by largely independent effects of upstream and downstream sequences. Notably, an upstream adenine markedly increases CpG mutation rates regardless of methylation status or downstream sequences. Furthermore, upstream and downstream sequences have qualitatively similar effects in chimpanzee and rhesus macaque, indicating that some conserved, intrinsic sequence features shape CpG mutability. On the other hand, some inter-species differences, which are especially pronounced at methylated sites, point to recent evolutionary changes, possibly in context-specificity of proteins governing DNA demethylation and repair processes.

Author Summary

The DNA sequence surrounding a nucleotide strongly influences how likely it is to mutate. An extreme example is the CpG dinucleotide: cytosines in CpGs mutate far more frequently than other sites in the human genome. This is related to DNA methylation, a chemical modification that occurs almost exclusively at CpGs in vertebrates and makes cytosines more prone to mutations. However, CpGs in different sequence contexts also vary in their mutation rates, and methylation level alone cannot explain this variation. To gain insight into what processes drive this variation, we estimate mutation rates for methylated and unmethylated CpGs in different sequence contexts using human genetic variation data. We find that methylation and neighboring bases interact to influence CpG mutation rates, and that the DNA sequence on either side of the CpG exerts largely independent effects. Extending our analysis to other primates reveals both conserved and species-specific patterns, with differences being especially pronounced at methylated sites. Together, our results suggest that while intrinsic DNA sequence features underlie some conserved context effects, other differences reflect recent evolutionary changes in the mechanisms that regulate DNA demethylation and repair.

Introduction

Mutation rate varies across the human genome, even at the scale of individual nucleotides [1]. Quantifying and accounting for this mutation rate variation is important for making inferences about natural selection [2] and identifying functional coding [3–6] and non-coding regions [7]. Among the many predictors of mutation rate, one of the strongest is the local sequence context—the nucleotides flanking the mutated site [8,9]. Although context effects have been attributed to several mechanisms, including local DNA shape [10,11], biases in DNA polymerase fidelity [12], and differential repair efficiency across sequence environments [13,14], the relative importance of these mechanisms and whether they fully account for observed patterns remain open questions. Addressing these questions requires systematically characterizing how local sequence shapes site-specific mutation rate variation, which can help disentangle these contributions and shed light on the mechanistic basis of mutagenesis.

Among sequence contexts, the CpG dinucleotide is both exceptionally mutable and mechanistically informative. Across vertebrates, CpG sites experience C>T mutation rates an order of magnitude higher than those at other sites in the genome [1,15,16]. This hypermutability is directly linked to DNA methylation, as supported by two observations: first, in vertebrates, cytosine methylation occurs almost exclusively at CpG sites [17], and second, in humans, de novo C>T mutation rates at CpGs correlate strongly with germline methylation levels [18–20].

A widely accepted mechanism underlying the mutagenic effect of methylation is the spontaneous deamination of 5-methylated cytosine (5mC), which occurs at a much higher rate than for unmethylated cytosine [21]. Deamination of 5mC produces thymine, generating a T:G mismatch that can be converted to a T:A base pair if the mismatch is erroneously repaired or

74 left unrepaired until the next round of DNA replication. In contrast, deamination of unmethylated
75 cytosine yields uracil, a non-canonical base that is more reliably recognized and removed [22].
76 However, not all C>T mutations at CpG sites arise from deamination and faulty/inefficient repair.
77 Recent work on human DNA polymerase ε suggests that mis-incorporation of nucleotides
78 during replication can also contribute to CpG>TpG mutations, as this enzyme shows a bias
79 toward incorporating adenine opposite 5mC and even unmethylated cytosine in a CpG context
80 [12,23].
81
82 CpG sites also exhibit substantial mutation rate variation across sequence contexts [8,24].
83 Aggarwala and Voight [8] reported such variation and a significant yet modest correlation (R^2 =
84 0.33) between methylation in sperm and the probability of C>T substitutions in 7-*mer* CpG
85 contexts, suggesting that variation in methylation level cannot fully explain CpG mutation rate
86 variation across sequence contexts. A key question, then, is what makes some CpG sites, by
87 virtue of their sequence context, more mutable than others. Intrinsic features of the local DNA
88 structure may influence repair efficiency of T:G mismatches [14,25,26], the rate of deamination
89 [27], or replication fidelity [12,23], offering distinct mechanistic explanations. In addition, context-
90 specific DNA-binding proteins such as transcription factors can also modulate mutation rate by
91 interfering with repair machinery [28–30]. To evaluate these mechanisms, we need a mutation
92 model that explicitly considers both methylation state and sequence context for each CpG site.
93 Such a model would enable us to disentangle the contribution of a well-defined biochemical
94 mechanism (i.e., 5mC deamination) from additional context effects that may arise from DNA
95 replication, damage, repair, or protein binding.
96
97 Here, we introduce a framework for modeling CpG mutation rates across sequence contexts
98 and methylation levels. Unlike existing models that ignore methylation [31,32] or treat it as a
99 categorical variable [3,33], our approach incorporates methylation as a continuous predictor,

enabling us to disentangle the contributions of methylation and local sequence context while leveraging information from CpG sites across different methylation levels. Applying this approach to human polymorphism data from the Genome Aggregation Database (gnomAD v4.0), we estimate mutation rates for both unmethylated and methylated CpGs in different 4-mer and 6-mer contexts. We also extend our analysis to two additional primate species to identify conserved and divergent patterns in CpG mutation rate variation. Overall, our work shows that unmethylated and methylated cytosines represent fundamentally different mutational substrates with distinct context dependencies that together shape CpG mutation rate variation across the genome.

Results

Interaction between methyl group and flanking nucleotides on CpG mutation rate

We hypothesized that the variation in CpG mutation rates across contexts is shaped by both differences in baseline mutation rates of unmethylated cytosines and differential mutagenic effects of methylation (i.e. the difference in mutation rate between methylated and unmethylated cytosines). We started by calculating the CpG transition polymorphism rate for each tetranucleotide (4-mer) context at varying methylation levels, using single nucleotide polymorphisms (SNPs) as a proxy for mutation events in the ancestors of the sampled genomes. For primary analysis, we used SNPs in the inter-genic regions of the human genome from gnomAD [34], excluding singletons due to concerns of high false positive rates, and polarized SNPs based on the inferred ancestral human genome (Methods); inter-genic SNPs from the 1000 Genomes (1KG) Project including singletons were used for replication [35].

124 To approximate methylation levels in the male germline, we used bisulfite sequencing from
125 human sperm [36]. We recognize that germline mutations can arise in both the male and female
126 germlines, as well as during early embryonic development, so the most relevant methylation
127 metric should be a weighted average of methylation levels across these different stages.
128 However, we chose to use sperm bisulfite sequencing data as a proxy for germline methylation
129 level for two reasons. First, the site-level methylation intensity in sperm shows the strongest
130 correlation with CpG mutation rate and has much greater predictive power than data from any
131 other developmental stages [37]. Second, whole-genome bisulfite sequencing data for tissues
132 other than testis and/or sperm are much more limited and not corrected for genotypes of the
133 assayed individuals, which creates biases in methylation level measurement (see Methods).
134
135 As expected, we found that the polymorphism rate increases with germline methylation level for
136 each 4-mer context (Fig 1A). However, the increase in polymorphism rate is approximately
137 linear only up to ~25% methylation, after which the curves become flatter. This non-linear
138 relationship between polymorphism rate and methylation can be explained by higher rates of
139 recurrent mutations at highly mutable sites in large samples [18,38]. This suggests that, at the
140 current sample size, the polymorphism probability of CpG sites no longer reflects their mutation
141 rate proportionately.

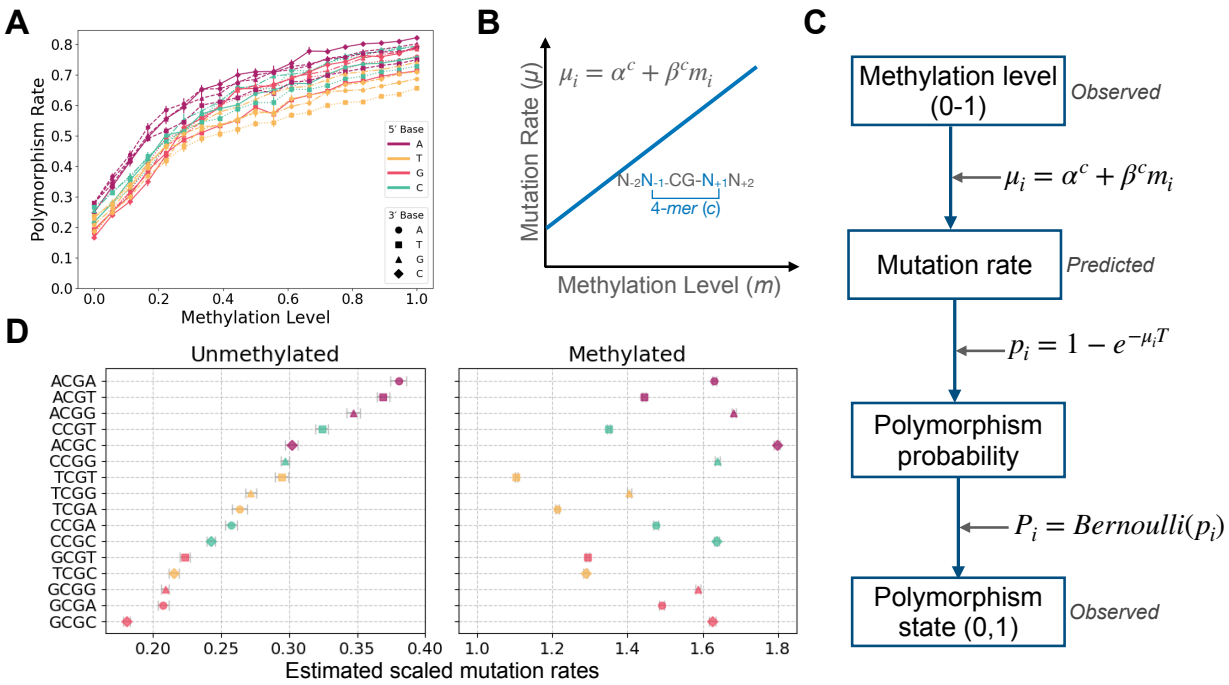

Fig 1. Interaction between sequence context and methylation shapes mutation rate variation at CpG sites. A. Polymorphism rate for each 4-mer context across increasing methylation level. Polymorphism rate was calculated across 20 bins of methylation level from 0 to 1. Each point represents the proportion of SNPs observed across the total number of sites within a methylation bin, along with 95% confidence intervals assuming binomial sampling. Each context is represented by a color denoting the nucleotide at the 5' position and a shape denoting the nucleotide at the 3' position **B.** Proposed mutation rate model where mutation rate μ_i for a given cytosine i in a CpG within a 4-mer sequence context c is a linear function of its methylation level m_i . α^c is the context-specific baseline rate when methylation is zero, and β^c captures the increase in mutation rate from unmethylated to fully methylated. **C.** Schematic showing how mutation rates are related with the observed methylation level m_i and polymorphism state at every site i. **D.** Scaled mutation rate estimates for unmethylated and methylated sites in each 4-mer context. Contexts are sorted by descending mutation rate when unmethylated. Shape and color are coded in the same way as in (A). Error bars represent 95% confidence intervals for the estimates.

To account for recurrent mutations, we related the polymorphism probability p to the underlying per generation mutation rate μ using an exponential transformation as $p = 1 - e^{-\mu T}$, where T represents the total branch length of the coalescent tree connecting all sampled genomes. This approach corrects for recurrent mutations to a first approximation and produces a sample-scaled mutation rate, μT. We treated the observed polymorphism state (i.e., presence or absence of polymorphism) at each site as a Bernoulli trial with success probability p (Fig 1C). To model the effects of sequence context, methylation, and their interaction, we assumed that the mutation rate increases linearly with methylation level, allowing both the intercept and slope to vary across sequence contexts (Fig 1B,C).

We implemented this model in a regression framework and estimated the scaled mutation rates of unmethylated and methylated CpGs in each 4-mer context using the gnomAD polymorphism data. We observed that, in terms of scaled mutation rates, the order of the 16 4-mer contexts was completely different between unmethylated and methylated states, demonstrating an interaction between the methyl group at the focal cytosine and flanking bases on CpG mutation rate (Fig 1D). Similar analysis of polymorphisms from the 1KG project [35] recovered highly concordant orders of most to least mutable contexts (S1A Fig).

Independent effects of upstream and downstream bases

By organizing the 4-mer contexts by their 5’ and 3’ bases in rows and columns and visualizing the inferred scaled mutation rates in heatmaps, we observed clear marginal effects of each flanking nucleotide on the mutation rate of unmethylated and methylated CpGs, as demonstrated by the row- and column-wise patterns (Fig 2A). Specifically, we found that 5’A has a strong, positive marginal effect on CpG mutability at both unmethylated and methylated

183 sites, which recapitulates previously reported hypermutable CpG motifs with A upstream to the
184 focal C [1,8,31]. We also observed a negative effect of 5'G on CpG mutability at unmethylated
185 sites and a strong negative effect of 5'T on methylated sites. For the downstream base, 3'C has
186 a negative effect on unmethylated CpGs, while 3'T exerts a negative impact on methylated
187 cytosines. These inferred marginal effects of flanking bases are all replicated qualitatively using
188 1KG polymorphism data (S1B Fig).
189
190 These consistent marginal effects—where upstream bases have similar effects regardless of
191 the downstream bases and vice versa— suggested that the flanking bases on the two sides of
192 the CpG act largely independently. We therefore asked to what degree the effect of each 4-mer
193 context could be captured by the combined effect of the upstream and downstream bases. To
194 test this, we explicitly defined a model where the upstream and downstream nucleotides
195 independently affect the mutability of the focal cytosine (hereafter referred to as the up_1+$down_1$
196 model) (Table 1). This model contrasts with the 4-mer model presented above, which implicitly
197 allows for interactions between the two flanking bases.
198
199 **Table 1. Descriptions of the various models specified in our regression framework.**

Model	Formula	Variable	Variable description	Assumption
All		*m*	Methylation level as measured in sperm	
4-mer	*4mer * m*	*4mer*	4-mer including the focal CpG and one nucleotide on either side	Each 4-mer context may have unique baseline mutation rate and methylation effect
up_1+$down_1$	*U1 + D1 + m + U1:m + D1:m*	*U1*	Nucleotide one position upstream of focal C	Each nucleotide upstream or downstream to CpG affects the baseline mutation rate and mutagenic effect of methylation independently
		D1	Nucleotide one position downstream of focal G	

6-mer	*6mer * methylation*	*6mer*	6-mer including the focal CpG and two nucleotides on either side	Each 6-mer context may have unique baseline mutation rate and methylation effect
up_{21}+$down_{12}$	*U21 + D12 + m + U21:m + D12:m*	*U21*	Two nucleotides upstream of focal C	Each dimer upstream and downstream to CpG may have a unique baseline mutation rate and methylation effect
		D12	Two nucleotides downstream of focal G	
up_2+up_1+ $down_1$+$down_2$	*U2 + U1 + D1+ D2 + m + U2:m + U1:m + D1:m + D2:m*	*U2*	Nucleotide two positions upstream of focal C	Each nucleotide up to two positions upstream or downstream to CpG affects the baseline mutation rate and mutagenic effect of methylation independently
		U1	Nucleotide one position upstream of focal C	
		D1	Nucleotide one position downstream of focal G	
		D2	Nucleotide two positions downstream of focal G	

We found that the scaled mutation rates in the 16 unique 4-mer contexts predicted from the up_1+$down_1$ model are highly concordant with those predicted from the 4-mer model (Pearson's r=0.98 for both methylated and unmethylated CpG; Fig 2A), confirming largely independent effects of the upstream and downstream nucleotides. Statistically, the 4-mer model is still significantly better than the up_1+$down_1$ model based on Akaike Information Criterion (AIC), but the gain in variance explained is small (12.148% vs. 12.124%) (S1 Table), suggesting that the simpler up_1+$down_1$ model largely captures variation in CpG mutation rates, albeit with fewer than half the parameters (14 vs. 32).

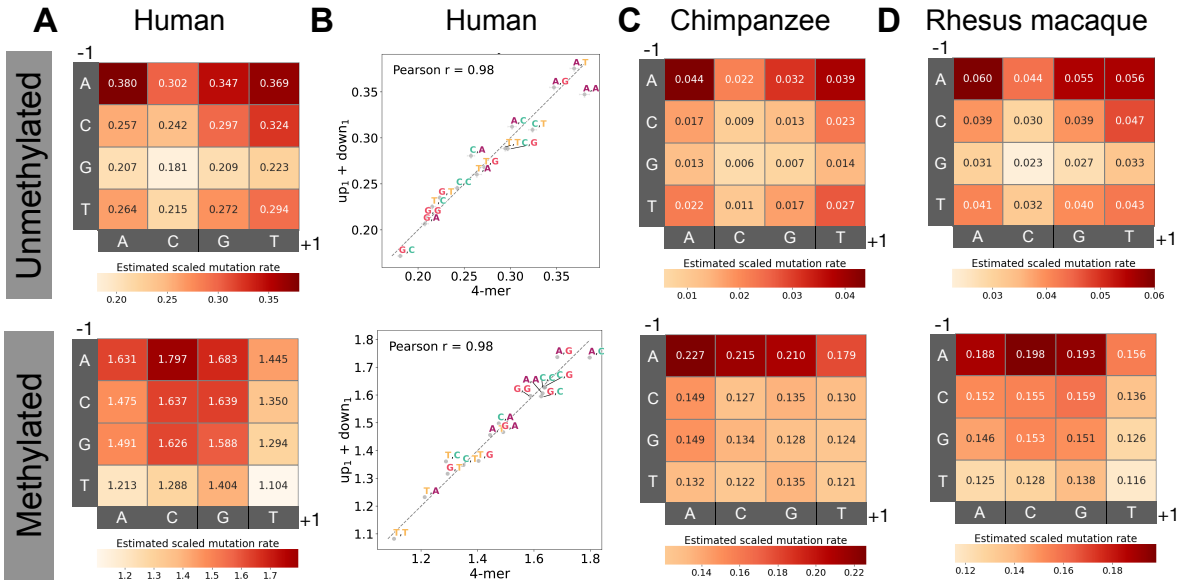

Fig 2. Independent effects of upstream and downstream nucleotides on CpG mutation rates and shared context effects across primates. **A.** Scaled mutation rates in humans estimated from the 4-mer model, which assumes interactions between the upstream and downstream bases. **B.** Concordance between human mutation rates estimated from the 4-mer model and from the up_1+$down_1$ model, which assumes independent effects of upstream and downstream bases. Dashed line indicates *x*=*y*. Points are labeled by the nucleotides upstream and downstream of the CpG, with error bars representing 95% confidence intervals for the estimates. **C, D.** Same as **A** but for chimpanzee and rhesus macaque. Each heatmap shows relative mutation rates for that panel; color scales are not comparable across panels.

Effects of sequence context on CpG mutation rate in other primate species

We then applied our regression framework to two additional primate species, chimpanzee and rhesus macaque, and recovered many similar observations as in humans (Fig 2C and Fig 2D). First, the rank order of 4-mer contexts in mutation rate differs between methylated and unmethylated states (S2 Fig). Second, the mutation rates predicted by the up_1+$down_1$ model

226 and 4-mer model are highly concordant, indicating largely independent effects of upstream and
227 downstream bases on focal CpG mutability (chimpanzee: Pearson's r=0.97 for unmethylated
228 and r = 0.98 for methylated CpG; rhesus macaque: Pearson's r=0.98 for both unmethylated and
229 methylated CpG; S2 Fig). Some of the marginal effects of flanking bases are also conserved
230 across the studied primates. In all three species, a 5'A increases the mutation rate of both
231 unmethylated and methylated CpGs, and a 5'G or 3'C decreases the mutation rate of
232 unmethylated CpGs (Fig 2A,C,D). Together, these findings suggest considerable conservation
233 in context-dependent mutation patterns of CpG sites across the three primate species.
234
235 Despite these similarities in overall patterns, there are clear inter-species differences in the
236 context-dependency of CpG mutation rate, demonstrated by the imperfect correlation in scaled
237 mutation rates across contexts (Fig 3). In general, 4-mer contexts show greater between-
238 species divergence at methylated sites than at unmethylated sites, with the greatest divergence
239 observed between humans and chimpanzees (Pearson's r = 0.61). Notably, discrepancies in
240 context effects between humans and chimpanzees are most pronounced in contexts with G/C at
241 both sites flanking the CpG, especially at methylated CpGs: such "G/C rich" contexts are
242 associated with relatively high mutation rates in humans but low mutation rates in chimpanzees
243 (Fig 3A). Together, our results show that context-specific CpG mutability, particularly at
244 methylated sites, is more diverged between these two closely related species, relative to each
245 with rhesus macaques (Fig 3B,C).

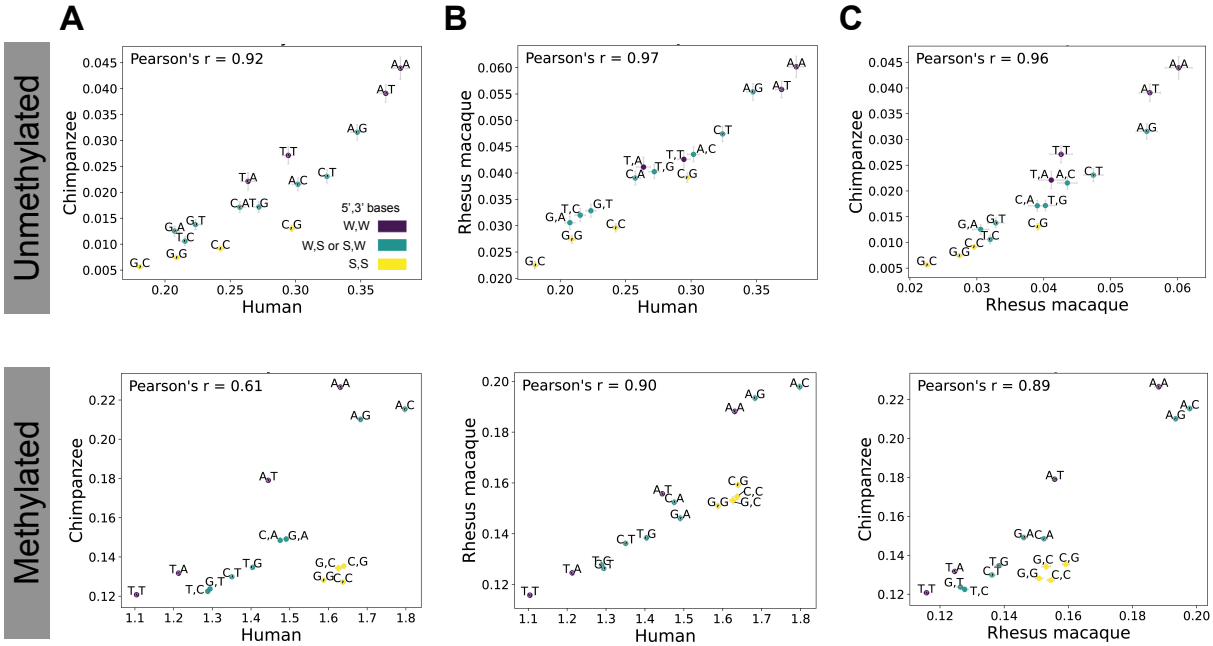

Fig 3. Similarity and differences in context effects on CpG mutability across primates . A. Comparison of scaled mutation rate estimates between human and chimpanzee at unmethylated (upper) and methylated sites (lower). Colors indicate whether the nucleotides flanking the focal CpG at the 5' and 3' positions are strong (S; C or G) or weak (W; A or T) bases. Error bars represent 95% confidence intervals for the estimates on both axes. **B.** Same as **A.** for comparisons between human and rhesus macaque. **C.** Same as **A.** for comparison between rhesus macaque and chimpanzee.

Mutational asymmetry within the same CpG dinucleotide

Given that the upstream and downstream nucleotides have asymmetric influences on the mutability of the focal cytosine, we wondered whether the two adjacent C:G base-pairs within the same CpG dinucleotide differ in mutation rate. To test this, we compared mutation rates of reverse complement 4-mer contexts (e.g., ACGC and GCGT). Since our modeling and analysis always measure mutation rate at the C position of each CpG, reverse complement contexts correspond to the two adjacent C:G base pairs within the same CpG dinucleotide (Fig 4).

For humans, we found that at both methylated and unmethylated sites, all context pairs showed significant mutational asymmetry between the two adjacent base pairs (Fig 4). The direction of mutational asymmetry is concordant for unmethylated and methylated CpGs in four out of six 4-mer context pairs. Notably, a cytosine directly flanked by an upstream A is more mutable than the neighboring G:C base pair within the same CpG dinucleotide, consistent with the strong mutagenic effect of 5'A (Fig 2). Two exceptions occur in the CCGA:TCGG and GCGA:TCGC pairs, where methylation status determines the direction of mutational asymmetry: the C:G base pair flanked by A:T is more mutable than its counterpart when unmethylated but less mutable when methylated.

Consistent with the findings in humans, for most contexts, adjacent C:G base pairs within the same CpG also differed in mutability in chimpanzees and rhesus macaques (Fig 4). Specifically, in all three species, the C:G base pair flanked by an A:T is more mutable than the neighboring G:C base pair regardless of methylation state. Importantly, when significant asymmetry is observed for a context pair, the direction of asymmetry is usually consistent across all three species at both unmethylated and methylated sites, reflecting the well-conserved effects of certain flanking bases (Fig 2), except for CCGA:TCGG, where chimpanzee and human show opposite directionality in asymmetry.

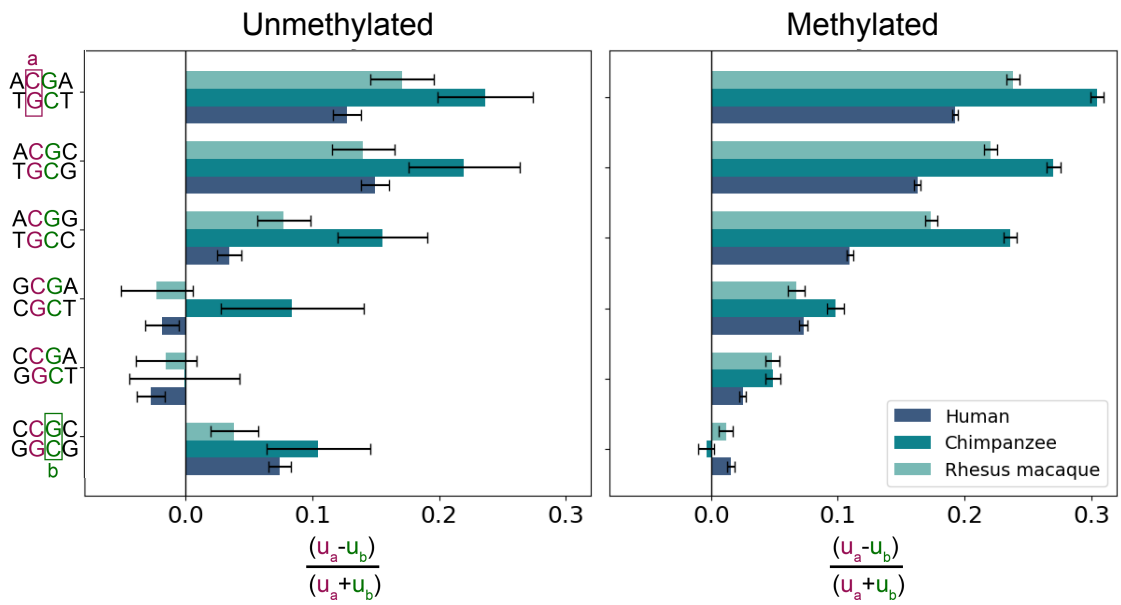

Fig 4. Mutational asymmetry within a CpG site. Each bar represents a pair of reverse complement 4-mer contexts (e.g., ACGC and GCGT) that contain the same CpG dinucleotide, with color indicating the species (human, chimpanzee, or rhesus macaque). Within the same CpG dinucleotide, adjacent C:G base pairs *a* and *b* often show different mutation rates in both unmethylated and methylated states. Error bars represent 95% confidence intervals calculated using propagation of uncertainty.

Effects of expanded sequence context on CpG mutation rate in primates

The expanded sequence context, beyond bases immediately adjacent to the mutated site, is also a known predictor of mutation rate across the genome [8,31,33]. We therefore wanted to test to what degree expanded contexts could explain variation in CpG mutation rates. We applied our regression framework to 6-mer contexts, now including the nucleotides two positions upstream of the cytosine and two positions downstream of the guanine. The 6-mer model assumes all possible interactions between the four nucleotides flanking the CpG. Consistent with observations based on a 4-mer context, effect of the same 6-mer context differs between

unmethylated and methylated sites, indicating interactions between flanking nucleotides and the methyl group (Fig 5).

Using the predicted mutation rates from the 6-mer model, we explored the marginal effects of upstream and downstream dimers on the mutability of the focal CpG in humans (Fig 5A). Echoing the strong mutagenic effect of upstream A detected at the 4-mer scale, our 6-mer model revealed that an upstream TA dimer has the strongest, positive marginal effect on CpG mutability at unmethylated sites, whereas an upstream GA dimer has the strongest, positive marginal effect at methylated sites. We also found that at methylated sites, irrespective of the nucleotide at the +1 downstream position, a T at the downstream +2 position (T_{+2}) is consistently associated with the lowest mutation rates.

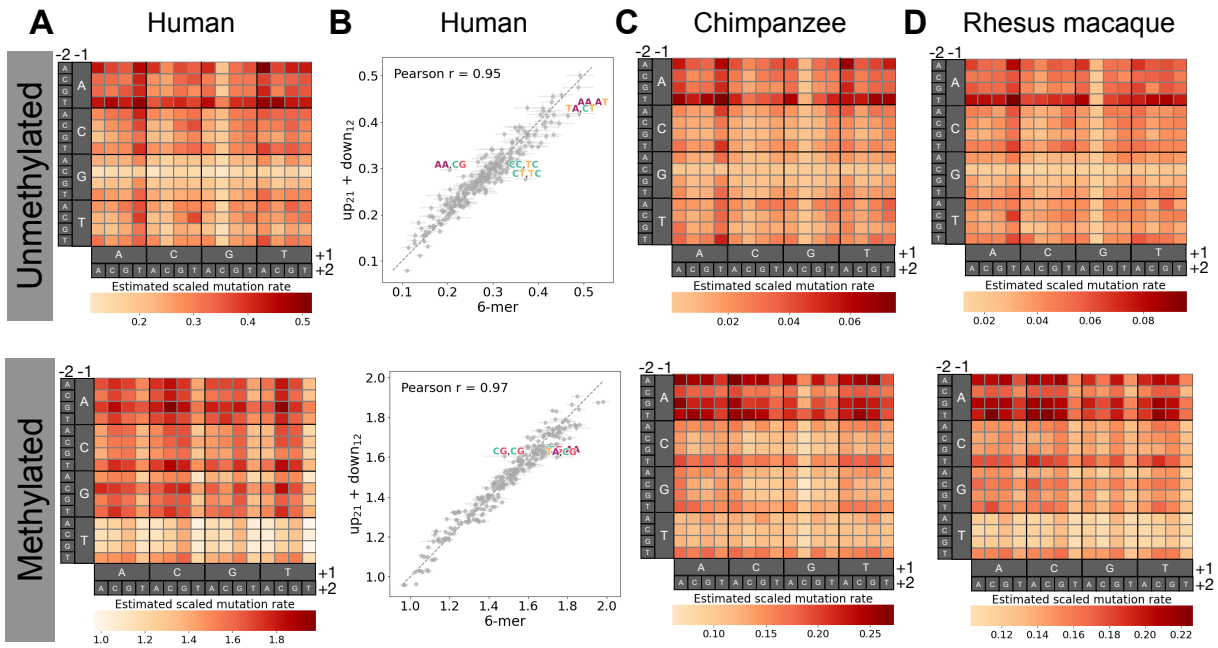

Fig 5. Shared and diverged effects of upstream and downstream bases on CpG mutability across primates at 6-mer resolution. A. Scaled mutation rates estimated from the 6-mer model for human at unmethylated (upper) and methylated sites (lower). **B.** Concordance between scaled mutation rates estimated from the 6-mer model and the up_{21}+$down_{12}$ model,

which assumes interactions only between bases within the upstream and downstream dimers. Dashed line indicates x=y. Points are labeled by the dimers upstream and downstream of the CpG. Only contexts for which the difference between the two model estimates exceeds 2.5 standard deviations from the mean difference are labeled. Error bars represent 95% confidence intervals for the estimates. **C, D.** Same as **A** but for chimpanzee and rhesus macaque. Each heatmap shows relative mutation rates for that panel; color scales are not comparable across panels.

Inspired by the distinct row- and column-specific patterns in the heatmaps showing results generated by the 6-mer model, we tested a simpler model: the up_{21}+$down_{12}$model, which only includes interactions between adjacent bases within the upstream and downstream dimers but no between upstream and downstream sequences. We found high concordance between scaled mutation rates predicted from the 6-mer model and the up_{21}+$down_{12}$model (Pearson's r=0.95 and 0.97 for unmethylated and methylated CpG, respectively; Fig 5B), indicating that upstream and downstream sequences exert mostly independent effects on mutation rate of the focal CpG.

We then asked how an even simpler model, where all flanking nucleotides contributed independently to CpG mutability, would perform. This model, called the up_2+up_1+$down_1$+$down_2$ model, showed weaker concordance with the up_{21}+$down_{12}$ model (Pearson's r=0.93 and 0.92 for unmethylated and methylated CpG, respectively; S3 Fig), suggesting that interactions within the upstream or downstream dimers contribute to CpG mutation rate variation, even though interactions between upstream and downstream sequences play less important roles. However, complex models with many more parameters do not perform considerably better than simpler models (S1 Table). Overall, the up_{21}+$down_{12}$ model seems to offer the best balance by capturing most context effects without the heavy parametrization of the 6-mer model.

When extending this analysis to include chimpanzee and rhesus macaque, we identified several conserved context effects at the 6-mer resolution as well as some differences. We found that the upstream dimer CG and downstream dimer GC were consistently associated with low mutation rates at unmethylated sites in all three species (Fig 5A,C,D). As in humans, a 5'TA dimer has a strong, positive effect on unmethylated CpG mutability in chimpanzees and rhesus macaques. However, at methylated sites, differences in context effects become more noticeable among the three species. Notably, the striking, negative effect of downstream T_{+2} on methylated CpG in humans is not seen in chimpanzees. Nonetheless, in chimpanzees and rhesus macaques, echoing our observations in humans, the most complex 6-mer model did not perform considerably better than simpler models, and that the up_{21}+$down_{12}$ model provided the best tradeoff between model complexity and variance explained (S1 Table). Overall, across all primates, interactions within but not between upstream and downstream dimers can capture most mutation rate variation at CpG sites.

Effects of sequence context on CpG mutation rate in an insect species with minimal DNA methylation

Our analysis thus far has inferred the mutagenic effects of methylation by modeling it as a continuous spectrum. However, sites classified as "unmethylated" in our analysis may still experience methylation in other tissues or developmental stages that is not captured by sperm bisulfite sequencing. Therefore, species with minimal methylation can function as a biological control and can better approximate a truly unmethylated state. We extended our analysis to the insect *Bombyx mori* (domestic silkworm), where less than 1% of the genome is methylated, with methylated cytosines largely in gene bodies [39]. We modified our regression framework to eliminate the effect of methylation, modeling the mutation rate as a function of sequence context

366 alone. As in our primate analyses, we also tested simpler additive models without higher-order
367 interactions between flanking bases.
368
369 In *B. mori*, we observed high concordance between mutation rate estimates from the 4-*mer*
370 model and the up_1+$down_1$ model with nearly identical AIC (Pearson's r = 0.98; S4B Fig; S1
371 Table), indicating that flanking bases influence CpG mutability mostly independently. Some of
372 the marginal effects of flanking nucleotides are similar between silkworm and primates at
373 unmethylated sites. Specifically, silkworms exhibit a positive effect of 5'A and a negative effect
374 of 3'C on CpG mutability (S4A Fig), mirroring the marginal effects at unmethylated sites in all
375 three primates (Fig 2). These shared patterns suggest context effects that are independent of
376 DNA methylation and may instead reflect intrinsic features of the local sequence itself.
377
378 However, when we extended the analysis to 6-mer contexts, we found less distinctive column-
379 or row-wise patterns (S4C Fig) and lower concordance between the 6-mer model and the
380 up_{21}+$down_{12}$ model (Pearson's r = 0.87, S4D Fig), suggesting that independent effects of
381 adjacent dimers only partially capture the variation. Interestingly, although the complex 6-mer
382 model showed a significant increase in the proportion of variance explained compared to the
383 up_{21}+$down_{12}$ and the up_2+up_1+$down_1$+$down_2$ models, it has a greater AIC suggesting worse
384 performance given increased model complexity (S1 Table). Together, these results suggest
385 diminishing returns with increased model complexity: while more complex models capture more
386 variance, the improvement in predictive power does not offset the cost of including additional
387 parameters.
388

389 # Discussion

390 Here, we present a regression framework that estimates CpG mutation rates in different 4-mer
391 and 6-mer contexts at unmethylated and methylated sites. Our model introduces two key
392 innovations: treating methylation level as a continuous predictor and explicitly modeling
393 recurrent mutations. This approach avoids the arbitrary binning of methylation levels into
394 discrete categories (e.g., high, medium, low) and reduces systematic underestimation of
395 mutation rates at hypermutable sites.
396
397 While sperm bisulfite sequencing provides a practical approximation of methylation levels in the
398 male germline, it has clear limitations. Sperm-based methylation data may not accurately reflect
399 levels in spermatogonial stem cells, and they fail to capture methylation profiles in other cell
400 types relevant to germline mutagenesis (e.g., female germ cells, early embryonic tissues). A
401 particular concern is that some sites classified as “unmethylated” in our analysis may in fact be
402 methylated in these other contexts. To strengthen our findings on unmethylated sites, we
403 applied our model to silkworm, a species with very low levels of DNA methylation that is largely
404 concentrated in gene bodies [39]. The recovery of consistent effects of certain contexts (e.g.,
405 positive effect of upstream A) in this system supports the robustness of our findings for
406 unmethylated CpGs.
407
408 Unlike several prior mutation models, we explicitly account for effects of recurrent mutations on
409 observed polymorphisms. At hypermutable sites in large samples, recurrent mutations make the
410 observed polymorphism rate a poor linear proxy for mutation rate (Fig 1A). To address this, we
411 used a standard exponential transformation to relate polymorphism probability to mutation rate,
412 assuming that the total length of the coalescent tree (T) is constant across the genome.
413 However, genealogical histories certainly vary among genomic loci due to recombination and
414 coalescent stochasticity, as well as systematic factors like background selection, all of which
415 cause heterogeneity in T [40,41]. This assumption is especially problematic for species with

very small sample sizes, such as chimpanzees, where sampling stochasticity in the coalescent process is greater. The over-simplifying assumption of constant T may distort our quantitative estimates of context-specific mutation rates and introduce biases in between-species comparisons. Nonetheless, because such heterogeneity should largely average out across CpG sites within each context, it is unlikely to affect the relative ranking of contexts by mutability. We thus focus on qualitative rather than quantitative comparisons across contexts and species.

Applying this modeling framework to population polymorphism datasets, we show that CpG mutation rate variation is shaped by the interaction between the local sequence context and methylation at the focal cytosine. Specifically, the effects of flanking sequences are often distinct on unmethylated and methylated cytosines, even though such effects are shared across species, supporting the view that these two cytosine states act as fundamentally different mutational substrates. Our analysis also demonstrates that the upstream and downstream bases flanking the CpG exert largely independent influences on mutability, a pattern that holds for both 4-mer and 6-mer sequence contexts. Lastly, we identify several strong marginal effects of adjacent nucleotides on CpG sites, including a strong mutagenic effect of an upstream A on both unmethylated and methylated CpGs, a negative effect of flanking CG—either upstream or downstream—on focal unmethylated CpGs, and a positive effect of the 5′TA dimer on unmethylated CpGs.

Importantly, the context and methylation effects we report here can help generate hypotheses and guide interpretations about the underlying mechanisms of germline CpG>TpG mutagenesis. For example, recent work has shown that transcription factor (TF) binding sites may act as mutation hotspots by interfering with DNA repair [30], with TF motifs typically spanning 6-10 nucleotides [50]. By contrast, our results indicate that CpG mutability is primarily explained by largely independent effects of upstream and downstream sequences. This

suggests that the molecular mechanisms underlying context effects may operate in a modular fashion—upstream and downstream sequences likely influence distinct steps in mutagenesis rather than requiring complex long-range structural interactions. This argues against TF occupancy as a dominant driver of context-dependent CpG mutation rate variation in the germline. Future work should test whether this modularity extends to CpG transversions or other single-nucleotide substitutions. If confirmed, this modularity has practical implications: simpler additive or multiplicative models can achieve nearly the same predictive power as fully parameterized models while remaining interpretable and computationally tractable.

Notably, an upstream A substantially increases mutability at both unmethylated and methylated sites relative to other bases, which recapitulates previously reported hypermutable motifs [1,8,31] and appears conserved across species, including chimpanzee, rhesus macaque, and even silkworm. This deep conservation suggests that the ACG motif has intrinsic biophysical properties—perhaps related to DNA shape—that elevate CpG mutation rates independent of methylation status. Intriguingly, previous work shows that C>T mutation rates are also elevated at non-CpG sites flanked by an upstream A [31], indicating that hypermutability may be a general property of ACN sequence, regardless of CpG or non-CpG sites, which calls for investigation into the mechanism.

In contrast, an upstream G does not elevate mutation rate at methylated sites and consistently reduces mutation rate at unmethylated CpGs in all three primate species (Fig 2). This observation contrasts with recent reports of elevated error rate of wild-type DNA polymerase ε in GCG contexts [12]. Together, this suggests that although replication-related CpG>TpG mutations occur, they cannot explain the observed context-dependent mutation rate variation and are unlikely to represent the major source of germline CpG mutations .

468 Cross-species comparisons of the CpG mutation patterns offer additional mechanistic insight.
469 Within primates, divergence in context effects between species appears more pronounced at
470 methylated sites than at unmethylated sites, pointing to recent evolutionary changes related to
471 methylated CpGs. These may reflect shifts in the context-specificity of proteins involved in
472 demethylation (e.g., TET enzymes [51]) or in DNA repair (e.g., thymine DNA glycosylases [52]).
473

474 Methods

475 Extracting CpG sites

476 We retrieved the hard-masked reference genomes for human (hg38), chimpanzee (panTro5),
477 rhesus macaque (Mmul10), and silkworm [53] from UCSC genome browser and Silkbase (see
478 Data sources). To minimize the effects of selection, we excluded genic regions based on NCBI
479 RefSeq annotations for all species studied; for human, we also removed phylogenetically
480 conserved regions defined by *phastcons* (see Data sources). For each species, we scanned the
481 reference genome to identify every occurrence of “CG” and recorded the position of the dimer
482 along with four nucleotides upstream and four nucleotides downstream to get the “10-mer”
483 sequence context for the CpG. Each CpG dinucleotide was treated as two independent
484 cytosines: one on the forward strand in the extracted context and the other on the reverse
485 strand in the reverse complement sequence context. From these 10-mers, we extracted 4-mer
486 and 6-mer contexts as needed.
487

488 Since reliable ancestral genome reconstructions were not available for all species studied, we
489 used each species’ reference genome to identify CpG sites. We note that this approach will
490 miss ancestral CpG sites that are no longer present in the reference due to mutations (e.g.,
491 ancestral CpG sites where the reference sequences are TpG or CpA). To assess the potential
492 impact of this limitation, we repeated the analysis by extracting CpG sites from the inferred

493 human ancestral genome and observed qualitatively similar results (S5 Fig). To test whether
494 excluding repetitive regions using the hard-masked reference genome could technically bias our
495 results, we repeated the analysis by extracting CpG sites from the unmasked human reference
496 genome and again observed qualitatively similar results (S6 Fig), suggesting that the reported
497 context effects on CpG mutability are likely robust to the choice of reference versus ancestral
498 context and hard-masked versus unmasked reference genomes.
499

500 ## Data sources

501

Reference Genome	Human: https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.masked.gz Chimpanzee: https://hgdownload.soe.ucsc.edu/goldenPath/panTro5/bigZips/panTro5.fa.masked.gz Rhesus: https://hgdownload.soe.ucsc.edu/goldenPath/rheMac10/bigZips/rheMac10.fa.masked.gz Silkworm: https://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/download.cgi
Conserved Regions	Human: https://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/phastConsElements30way.txt.gz

Gene Annotation	Human:http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/ncbiRefSeq.txt.gz Chimpanzee: https://hgdownload.soe.ucsc.edu/goldenPath/panTro5/database/ncbiRefSeq.txt.gz Rhesus: https://hgdownload.soe.ucsc.edu/goldenPath/rheMac10/database/ncbiRefSeq.txt.gz Silkworm: https://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/download.cgi
Methylation	Human: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138598
LiftOver	Human: https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz Chimpanzee: https://hgdownload.soe.ucsc.edu/goldenPath/panTro6/liftOver/panTro6ToPanTro5.over.chain.gz Rhesus: https://hgdownload.soe.ucsc.edu/goldenPath/rheMac8/liftOver/rheMac8ToRheMac10.over.chain.gz

Polymorphism	Human: gnomAD v4: https://gnomad.broadinstitute.org/data 1000 Genomes: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_raw_GT_with_annot/ Chimpanzee: https://www.ebi.ac.uk/ena/browser/view/PRJEB39475?show=analyses Rhesus: https://zenodo.org/records/3515522 Silkworm: https://ftp.cngb.org/pub/CNSA/data5/CNP0002456/Variation/CVF0092651/
Ancestral Reference Genome	Human: https://ftp.ensembl.org/pub/release-106/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38.tar.gz Chimpanzee: https://ftp.ensembl.org/pub/release-106/fasta/ancestral_alleles/pan_troglodytes_ancestor_Pan_tro_3.0.tar.gz Rhesus:

	https://ftp.ensembl.org/pub/release-106/fasta/ancestral_alleles/macaca_mulatta_ancestor_Mmul_10.tar.gz

502

503

504 ## Methylation data

505 For human, we downloaded bisulfite sequencing data from Chen et al. [36], which reported
506 sperm WGBS data from nine healthy individuals. When a C/T SNP exists at a CpG site,
507 methylation estimates are biased in individuals carrying one or two T alleles, as the T allele will
508 be mistaken for a bisulfite-converted unmethylated C. For instance, even if the CpG allele is
509 "fully" methylated in sperm, the presence of the TpG allele results in incorrect reporting of only
510 50% methylation in heterozygotes and 0% methylation in T/T homozygotes. To reduce this bias,
511 we calculated the average methylation level at each CpG by pooling methylated and
512 unmethylated read counts across samples. This approach effectively reduces biases for most
513 CpGs where the derived T allele is rare in the nine individuals but still underestimates
514 methylation levels at sites where the T allele is at high frequency. Additionally, this pooling
515 strategy assumes that methylation patterns are relatively consistent across individuals at most
516 CpG sites, which may not hold for regions with high inter-individual variability or imprinted loci.
517 We then lifted over the average methylation level from hg19 to hg38.

518

519 For chimpanzee and rhesus macaque, we retrieved testes EM-Seq methylation calls averaged
520 between two individuals and corrected for genotype (A. Stoliarova, personal communications) at
521 each CpG dinucleotide. Chimpanzee methylation data was lifted over from panTro6 to panTro5.

522

523 ## Filtering and polarization of polymorphism

We downloaded autosomal polymorphism data for human [34,35], rhesus macaque [54], chimpanzee [55], and silkworm [56] (see Data sources). For rhesus macaque, variant coordinates were lifted over from Mmul8 to Mmul10. Singletons were excluded in gnomAD (v4) and the silkworm datasets to reduce false-positive variant calls [57]. To polarize the substitution that gave rise to each SNP, we used ancestral allele reconstructions based on the 10 primate EPO alignment [58] for human (hg38), chimpanzee (panTro5), and rhesus macaque (Mmul10). For silkworm, we used *est-sfs* [59], which estimates the probability that the major allele of the focal species (*Bombyx mori*) is the ancestral state using polymorphism data from the focal species and alignment with a closely related outgroup (*Bombyx mandarina*). We then annotated each SNP with the type of substitution and retained only C>T and G>A mutations at CpG sites in the reference genome (extracted as described above, see S2 Table).

Modeling effects of methylation and sequence contexts on CpG mutation rate

To quantify the effects of local sequence context and methylation level on CpG mutability, we modeled the mutation rate μ_i at a given cytosine i (within a CpG dinucleotide) as a linear function of its methylation level m_i, allowing both the intercept and slope to vary with sequence context:

$$\mu_i = \alpha^c + \beta^c m_i$$

where, α^c represents the baseline mutation rate for context c and β^c denotes the effect of full methylation within that context.

For a site with mutation rate of μ per generation, the probability of having polymorphism in a sample was modeled as $p = 1 - e^{-\mu T}$, where T is the total coalescent branch length (in generations) connecting all sampled genomes. For simplicity, we assumed that T was constant

across all CpG sites in non-genic, non-conserved regions. The presence or absence of polymorphism at each site was then modeled as a Bernoulli trial. i.e., drawn from a binomial distribution $Binom(1, p)$ with $p = 1 - e^{-\mu T}$.

Regression framework

To implement this model, we used the *glm* function in R, specifying the binomial family and a custom link function, which maps the polymorphism probability (p) to a linear predictor $\eta = -\mu T = \log(1 - p)$ as follows:

- Link function: $link(p) = \log(1 - p)$
- Inverse link function: $link^{-1}(\eta) = 1 - e^{\eta}$
- Derivative of the inverse link: $\frac{d\mu}{d\eta} = -e^{\eta}$

We fitted the generalized linear model to site-level data, where each record corresponded to one base-pair in a CpG site annotated with its methylation level, sequence context, and a binary indicator of polymorphism status (0 = monomorphic, 1 = polymorphic). The regression model was specified differently depending on the context length and type of interaction, as described in Table 1.

After fitting the regression models to the polymorphism data, we estimated the scaled mutation rates (μT) for each sequence context and methylation status (*x*=1 for methylated; *x*=0 for unmethylated) using the predict() function in R with type = "link" to obtain the linear predictor (η).

To evaluate model performance, we extracted summary statistics from each fitted regression and compared models using the Akaike Information Criterion (AIC), which balances between

model fit and complexity. We also measured explanatory power as the proportion of variance in polymorphism status explained, calculated as the difference between null and residual deviance, standardized by the null deviance.

Acknowledgments

We thank Anastasia Stoliarova and Molly Przeworski for kindly sharing sperm methylation data for chimpanzee and rhesus macaque before publication; Iain Mathieson, Fabian Ramos-Almodovar, and other members of the Gao and Mathieson laboratories for helpful discussions.

References

1. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 2011;12:756–66. https://doi.org/10.1038/nrg3098

2. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5:e1000471. https://doi.org/10.1371/journal.pgen.1000471

3. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43. https://doi.org/10.1038/s41586-020-2308-7

4. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, Samocha KE, et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat Genet. 2017;49:806–10. https://doi.org/10.1038/ng.3831

5. Agarwal I, Fuller ZL, Myers SR, Przeworski M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. eLife. 2023;12:e83172. https://doi.org/10.7554/eLife.83172

6. Zeng T, Spence JP, Mostafavi H, Pritchard JK. Bayesian estimation of gene constraint from an evolutionary model with gene features. Nat Genet. 2024;56:1632–43. https://doi.org/10.1038/s41588-024-01820-9

7. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625:92–100. https://doi.org/10.1038/s41586-023-06045-0

8. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016;48:349–55. https://doi.org/10.1038/ng.3511

9. Blake RD, Hess ST, Nicholson-Tuell J. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol. 1992;34:189–200. https://doi.org/10.1007/BF00162968

10. Karolak A, Levatić J, Supek F. A framework for mutational signature analysis based on DNA shape parameters. Korolev S, editor. PLOS ONE. 2022;17:e0262495. https://doi.org/10.1371/journal.pone.0262495

11. Liu Z, Samee MAH. Structural underpinnings of mutation rate variations in the human genome. Nucleic Acids Res. 2023;51:7184–97. https://doi.org/10.1093/nar/gkad551

12. Tomkova M, McClellan MJ, Crevel G, Shahid AM, Mozumdar N, Tomek J, et al. Human DNA polymerase ε is a source of C>T mutations at CpG dinucleotides. Nat Genet. 2024;56:2506–16. https://doi.org/10.1038/s41588-024-01945-x

13. Donigan KA, Sweasy JB. Sequence context-specific mutagenesis and base excision repair. Mol Carcinog. 2009;48:362–8. https://doi.org/10.1002/mc.20497

617 14. Mitra R, Pettitt BM, Blake RD. Conformational states governing the rates of spontaneous
618 transition mutations. Biopolymers. 1995;36:169–79. https://doi.org/10.1002/bip.360360206

619 15. Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, et al. Evolution of
620 the germline mutation rate across vertebrates. Nature. 2023;615:285–91.
621 https://doi.org/10.1038/s41586-023-05752-y

622 16. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de
623 novo mutations and the importance of father's age to disease risk. Nature. 2012;488:471–5.
624 https://doi.org/10.1038/nature11396

625 17. Bird AP. CpG islands as gene markers in the vertebrate nucleus. Trends Genet.
626 1987;3:342–7. https://doi.org/10.1016/0168-9525(87)90294-0

627 18. Agarwal I, Przeworski M. Mutation saturation for fitness effects at human CpG sites. eLife.
628 2021;10:e71513. https://doi.org/10.7554/eLife.71513

629 19. Chen C, Qi H, Shen Y, Pickrell J, Przeworski M. Contrasting Determinants of Mutation
630 Rates in Germline and Soma. Genetics. 2017;207:255–67.
631 https://doi.org/10.1534/genetics.117.1114

632 20. Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, et al. Overlooked roles
633 of DNA damage and maternal age in generating human germline mutations. Proc Natl Acad Sci.
634 2019;116:9491–500. https://doi.org/10.1073/pnas.1901259116

635 21. Shen J-C, Rideout WM, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in
636 double-stranded DNA. Nucleic Acids Res. 1994;22:972–6. https://doi.org/10.1093/nar/22.6.972

637 22. Schmutte C, Yang AS, Beart RW, Jones PA. Base excision repair of U:G mismatches at a
638 mutational hotspot in the p53 gene is more efficient than base excision repair of T:G
639 mismatches in extracts of human colon tumors. Cancer Res. 1995;55:3742–6.

640 23. Tomkova M, McClellan M, Kriaucionis S, Schuster-Böckler B. DNA Replication and
641 associated repair pathways are involved in the mutagenesis of methylated cytosine. DNA
642 Repair. 2018;62:1–7. https://doi.org/10.1016/j.dnarep.2017.11.005

643 24. Harpak A, Bhaskar A, Pritchard JK. Mutation Rate Variation is a Primary Determinant of the
644 Distribution of Allele Frequencies in Humans. Eyre-Walker A, editor. PLOS Genet.
645 2016;12:e1006489. https://doi.org/10.1371/journal.pgen.1006489

646 25. Beard BC, Wilson SH, Smerdon MJ. Suppressed catalytic activity of base excision repair
647 enzymes on rotationally positioned uracil in nucleosomes. Proc Natl Acad Sci. 2003;100:7465–
648 70. https://doi.org/10.1073/pnas.1330328100

649 26. Sibghat-Ullah null, Day RS. DNA-substrate sequence specificity of human G:T mismatch
650 repair activity. Nucleic Acids Res. 1993;21:1281–7. https://doi.org/10.1093/nar/21.5.1281

651 27. Frederico LA, Kunkel TA, Shaw BR. A sensitive genetic assay for the detection of cytosine
652 deamination: determination of rate constants and the activation energy. Biochemistry.
653 1990;29:2532–7. https://doi.org/10.1021/bi00462a015

654 28. Hara R, Mo J, Sancar A. DNA Damage in the Nucleosome Core Is Refractory to Repair by
655 Human Excision Nuclease. Mol Cell Biol. 2000;20:9173–81.
656 https://doi.org/10.1128/MCB.20.24.9173-9181.2000

657 29. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, et al. Reduced
658 local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat
659 Biotechnol. 2014;32:71–5. https://doi.org/10.1038/nbt.2778

660 30. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. Nucleotide
661 excision repair is impaired by binding of transcription factors to DNA. Nature. 2016;532:264–7.
662 https://doi.org/10.1038/nature17661

663 31. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, et al. Extremely rare
664 variants reveal patterns of germline mutation rate heterogeneity in humans. Nat Commun.
665 2018;9:3753. https://doi.org/10.1038/s41467-018-05936-5

666 32. Adams CJ, Conery M, Auerbach BJ, Jensen ST, Mathieson I, Voight BF. Regularized
667 sequence-context mutational trees capture variation in mutation rates across the human
668 genome. Balding D, editor. PLOS Genet. 2023;19:e1010807.
669 https://doi.org/10.1371/journal.pgen.1010807

670 33. Seplyarskiy V, Koch EM, Lee DJ, Lichtman JS, Luan HH, Sunyaev SR. A mutation rate
671 model at the basepair resolution identifies the mutagenic effect of polymerase III transcription.
672 Nat Genet. 2023;55:2235–42. https://doi.org/10.1038/s41588-023-01562-0

673 34. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational
674 constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
675 https://doi.org/10.1038/s41586-020-2308-7

676 35. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource
677 (IGSR) collection of open human genomic variation resources. Nucleic Acids Res.
678 2020;48:D941–7. https://doi.org/10.1093/nar/gkz836

679 36. Chen X, Lin Q, Wen J, Lin W, Liang J, Huang H, et al. Whole genome bisulfite sequencing
680 of human spermatozoa reveals differentially methylated patterns from type 2 diabetic patients. J
681 Diabetes Investig. 2020;11:856–64. https://doi.org/10.1111/jdi.13201

682 37. Zhou Y, He F, Pu W, Gu X, Wang J, Su Z. The Impact of DNA Methylation Dynamics on the
683 Mutation Rate During Human Germline Development. G3 GenesGenomesGenetics.
684 2020;10:3337–46. https://doi.org/10.1534/g3.120.401511

685 38. Exome Aggregation Consortium, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E,
686 et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
687 https://doi.org/10.1038/nature19057

688 39. Xiang H, Zhu J, Chen Q, Dai F, Li X, Li M, et al. Single base-resolution methylome of the
689 silkworm reveals a sparse epigenomic map. Nat Biotechnol. 2010;28:516–20.
690 https://doi.org/10.1038/nbt.1626

691 40. Bromham L, Penny D. The modern molecular clock. Nat Rev Genet. 2003;4:216–24.
692 https://doi.org/10.1038/nrg1020

41. Duchêne DA, Tong KJ, Foster CSP, Duchêne S, Lanfear R, Ho SYW. Linking Branch Lengths across Sets of Loci Provides the Highest Statistical Support for Phylogenetic Inference. Kosakovsky Pond S, editor. Mol Biol Evol. 2020;37:1202–10. https://doi.org/10.1093/molbev/msz291

42. Kobayashi H, Sakurai T, Miura F, Imai M, Mochiduki K, Yanagisawa E, et al. High-resolution DNA methylome analysis of primordial germ cells identifies gender-specific reprogramming in mice. Genome Res. 2013;23:616–27. https://doi.org/10.1101/gr.148023.112

43. Bergeron LA, Besenbacher S, Bakker J, Zheng J, Li P, Pacheco G, et al. The germline mutational process in rhesus macaque and its implications for phylogenetic dating. GigaScience. 2021;10:giab029. https://doi.org/10.1093/gigascience/giab029

44. Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci. 2012;109:15716–21. https://doi.org/10.1073/pnas.1211740109

45. Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, et al. The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. Genome Res. 2016;26:1651–62. https://doi.org/10.1101/gr.204255.116

46. Wroblewski EE, Murray CM, Keele BF, Schumacher-Stankey JC, Hahn BH, Pusey AE. Male dominance rank and reproductive success in chimpanzees, Pan troglodytes schweinfurthii. Anim Behav. 2009;77:873–85. https://doi.org/10.1016/j.anbehav.2008.12.014

47. Venn O, Turner I, Mathieson I, De Groot N, Bontrop R, McVean G. Strong male bias drives germline mutation in chimpanzees. Science. 2014;344:1272–5. https://doi.org/10.1126/science.344.6189.1272

48. Gao Z, Wyman MJ, Sella G, Przeworski M. Interpreting the Dependence of Mutation Rates on Age and Time. Barton NH, editor. PLOS Biol. 2016;14:e1002355. https://doi.org/10.1371/journal.pbio.1002355

49. Martinez G, Garcia C. Sexual selection and sperm diversity in primates. Mol Cell Endocrinol. 2020;518:110974. https://doi.org/10.1016/j.mce.2020.110974

50. Stewart AJ, Hannenhalli S, Plotkin JB. Why Transcription Factor Binding Sites Are Ten Nucleotides Long. Genetics. 2012;192:973–85. https://doi.org/10.1534/genetics.112.143370

51. Ravichandran M, Rafalski D, Davies CI, Ortega-Recalde O, Nan X, Glanfield CR, et al. Pronounced sequence specificity of the TET enzyme catalytic domain guides its cellular function. Sci Adv. 2022;8:eabm2427. https://doi.org/10.1126/sciadv.abm2427

52. Maiti A, Morgan MT, Pozharski E, Drohat AC. Crystal structure of human thymine DNA glycosylase bound to DNA elucidates sequence-specific mismatch recognition. Proc Natl Acad Sci. 2008;105:8890–5. https://doi.org/10.1073/pnas.0711061105

53. Kawamoto M, Jouraku A, Toyoda A, Yokoi K, Minakuchi Y, Katsuma S, et al. High-quality genome assembly of the silkworm, Bombyx mori. Insect Biochem Mol Biol. 2019;107:53–62. https://doi.org/10.1016/j.ibmb.2019.02.002

54. Harris RA, Raveendran M, Worley KC, Rogers J. Unusual sequence characteristics of human chromosome 19 are conserved across 11 nonhuman primates. BMC Evol Biol. 2020;20:33. https://doi.org/10.1186/s12862-020-1595-9

55. Fair BJ, Blake LE, Sarkar A, Pavlovic BJ, Cuevas C, Gilad Y. Gene expression variability in human and chimpanzee populations share common determinants. eLife. 2020;9:e59929. https://doi.org/10.7554/eLife.59929

56. Tong X, Han M-J, Lu K, Tai S, Liang S, Liu Y, et al. High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. Nat Commun. 2022;13:5619. https://doi.org/10.1038/s41467-022-33366-x

57. Johnston HR, Hu Y, Cutler DJ. Population Genetics Identifies Challenges in Analyzing Rare Variants. Genet Epidemiol. 2015;39:145–8. https://doi.org/10.1002/gepi.21881

58. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18:1814–28. https://doi.org/10.1101/gr.076554.108

59. Keightley PD, Jackson BC. Inferring the Probability of the Derived vs. the Ancestral Allelic State at a Polymorphic Site. Genetics. 2018;209:897–906. https://doi.org/10.1534/genetics.118.301120

Supporting Information

S1 Fig. Concordance between mutation rates for 4-mer CpG contexts estimated from gnomAD (v4) and 1000 Genomes (1KG) polymorphism datasets. A. Scaled mutation rate estimates from regression for unmethylated and methylated sites for each 4-mer context using 1KG data. The contexts are sorted in descending order of mutability when unmethylated. **B.** Heatmap of estimated scaled mutation rates using the 4-mer model, which assumes an interaction between the upstream and downstream bases, at unmethylated and methylated sites, using polymorphism data from 1KG. Each heatmap shows relative mutation rates for that panel; color scales are not comparable across panels. **C.** Comparison of mutation rates estimated using polymorphisms from gnomAD and 1KG at unmethylated and methylated sites, respectively. Error bars in panel A and C represent 95% confidence intervals for the estimates.

761 **S2 Fig. Independent effects of upstream and downstream bases on CpG mutability in**
762 **chimpanzee and rhesus macaque. A.** Concordance between mutation rates as estimated
763 from 4-mer model and from up_1 + $down_1$ model which independently models the effects of
764 upstream and downstream bases in chimpanzee. **B.** Same as (A), for rhesus macaque. Error
765 bars represent 95% confidence intervals for the estimates on both axes.
766
767 **S3 Fig. Interactions within upstream and downstream dimers are still needed to explain**
768 **mutation rate variation at CpG sites for human polymorphisms in gnomAD dataset.**
769 Comparison of mutation rates as estimated from the up_{21}+$down_{12}$ model, which assumes
770 interactions between bases within the dimer upstream and downstream to the CpG and the
771 up_2+up_1+$down_1$+$down_2$ model, which assumes independent effects of every flanking base. Only
772 contexts for which the difference between the two models' estimates are more than 2.5 standard
773 deviations away from the average difference across all contexts are labelled. Error bars
774 represent 95% confidence intervals for the estimates on both axes.
775
776
777 **S4 Fig. Independent effects of upstream and downstream bases on CpG mutability in the**
778 **insect species, *Bombyx mori* (silkworm). A.** Heatmap of estimated scaled mutation rates
779 using the 4-mer model, which assumes an interaction between the upstream and downstream
780 bases. Each heatmap shows relative mutation rates for that panel; color scales are not
781 comparable across panels. Mutation rate is modelled as a function of sequence context alone.
782 **B.** Concordance between mutation rates as estimated from 4-mer model and from up_1+$down_1$
783 model which independently models the effects of upstream and downstream. **C.** Heatmap of
784 scaled mutation rates from the 6-mer model which assumes all possible interactions between
785 flanking nucleotides for the two species. **D.** Concordance between scaled mutation rates as
786 estimated from 6-mer model and those estimated from up_{21}+$down_{12}$, which only assumes

interactions between bases within the dimer upstream and downstream to the CpG. Only contexts for which the difference between the two model estimates are more than 2.5 standard deviations away from the average difference across all contexts are labelled.

S5 Fig. Order of mutability for 4-mer CpG contexts predicted by the regression is similar when CpG contexts are extracted from the reference genome or the ancestral genome in humans. A. Concordance between mutation rates as estimated from 4-mer model using the human reference genome versus the human ancestral genome inferred from the EPO 10-primate alignment (see Methods) for identification of CpG sites and flanking sequence. **B.** Concordance between mutation rates as estimated from 4-mer model and from up_1+$down_1$ model which independently models the effects of upstream and downstream bases at unmethylated and methylated sites, using the human ancestral genome. Error bars represent 95% confidence intervals for the estimates on both axes. **C.** Heatmap of estimated scaled mutation rates using the 4-mer model, which assumes an interaction between the upstream and downstream bases, at unmethylated and methylated sites, using the human ancestral genome. Each heatmap shows relative mutation rates for that panel; color scales are not comparable across panels.

S6 Fig. Order of mutability for 4-mer CpG contexts predicted by the regression is similar when CpG contexts are extracted from the masked or unmasked human reference genome. A. Concordance between mutation rates as estimated from 4-mer model using the hard-masked human reference genome versus the unmasked reference genome for identification of CpG sites and flanking sequence. **B.** Concordance between mutation rates as estimated from 4-mer model and from up_1+$down_1$ model which independently models the effects of upstream and downstream bases at unmethylated and methylated sites, using the unmasked human reference genome. Error bars represent 95% confidence intervals for the estimates on

both axes. **C.** Heatmap of estimated scaled mutation rates using the 4-mer model, which assumes an interaction between the upstream and downstream bases, at unmethylated and methylated sites, using the unmasked human reference genome. Each heatmap shows relative mutation rates for that panel; color scales are not comparable across panels.

S1 Table. Model comparison for human and non-human species. Summary statistics of each fitted regression model for each of the four species in our analysis. The proportion variance explained is calculated using the difference between null and residual deviance, standardized by the null deviance. In silkworm, the mutation rate is modeled as a function of sequence context alone.

S2 Table. Summary of sites and SNPs for each species. Genic sites are excluded in our analysis using NCBI *RefSeq* annotations. For human, phylogenetically conserved regions as defined by *phastcons* are also excluded. We identify C>T SNPs by polarizing the substitution that gave rise to each SNP using ancestral allele reconstruction based on the 10 primate EPO for human, chimpanzee, and rhesus macaque, and *est-sfs*, a tool that infers the ancestral state probabilities at polymorphic sites using information from a focal and outgroup species, for silkworm.