Purifying selection acts on germline methylation to modify the CpG mutation rate at promoters

Leandros Boukas[1,2], Hans T. Bjornsson[1,3,4,5,*], and Kasper D. Hansen[1,2,*]

[1]Department of Genetic Medicine, Johns Hopkins University School of Medicine
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[3]Department of Pediatrics, Johns Hopkins University School of Medicine
[4]Faculty of Medicine, University of Iceland
[5]Landspitali University Hospital
[*]Correspondence to khansen@jhsph.edu (KDH), hbjorns1@jhmi.edu (HTB)

ABSTRACT

A fundamental component of molecular evolution are the rules that govern when, and why, a given change (allele) is deleterious or neutral. The ability to define such rules for epialleles – analogous to the rules at the DNA sequence level – would thus have profound implications for our understanding of epigenetic variation and evolution. Here, we focus on promoter methylation in the male human germline, which – apart from its role in gene regulation – is also known to greatly increase the mutation rate of CpG dinucleotides. We first develop a simple but general approach for detecting selection on epialleles, which does not require population-scale data. We then show that germline promoter methylation is deleterious at loss-of-function intolerant genes, but neutral at loss-of-function tolerant ones. In concordance with this, a human-mouse comparative analysis of sperm methylomes reveals strong suppression of methylation acquisition at loss-of-function intolerant promoters. We demonstrate that this selection is neither a secondary consequence of germline gene expression levels, nor of promoter H3K4me3 levels. Rather, the deleteriousness of promoter methylation is explained by its mutagenic effect on the underlying CpGs. Our results thus address a long-standing open question in molecular evolution, providing the first demonstration of selection acting on an epigenetic mutation rate modifier to locally dictate the mutation rate in humans. They also suggest the existence of a mechanism that preferentially protects loss-of-function intolerant promoters from methylation in the germline. Finally, they directly refute the prevailing dogma that CpG islands are not under active selection.

INTRODUCTION

At the DNA sequence level, there is a rich understanding of rules that describe which changes are deleterious and which are neutral, especially within coding sequences. However, there are several "non-nucleotide" features, such as epigenetic marks (hereafter referred to as epialleles), whose high conservation across species (Woo and Li, 2012; Long, Sims, et al., 2013; Qu et al., 2018) suggests they are largely genetically determined, but for which deleteriousness and neutrality rules have not been clearly delineated. This fundamentally limits our understanding of their molecular evolution. It also hinders our ability to interpret intra- and inter-species variation.

An important class of epialleles in humans, and vertebrates in general, are the DNA methylation states of promoters in the germline. Apart from their role in gene regulation, these methylation states are intimately linked to the underlying promoter CpG densities. The reason is that the mutation rate of CpG dinucleotides is elevated by ∼15-fold when they are methylated in the germline (Coulondre et al., 1978; Wang et al., 1982; Cooper and Youssoufian, 1988). This methylation-induced hypermutability is believed to explain the relative

paucity of CpGs in most of the genome. At many promoters, in contrast, an unusually high density of CpGs is encountered (Bird, 1987; Deaton and Bird, 2011), and has been linked to transcriptional activity (Thomson et al., 2010; White et al., 2013; Clouaire et al., 2012; Wachter et al., 2014; Hartl et al., 2019). However, it is currently believed that such CpG islands are evolving neutrally (Cohen et al., 2011). This view has been adopted by the genetics community (Antequera and Bird, 2018), and has important ramifications; it implies that the biochemical properties CpG-richness confers to promoters have a negligible contribution to organismal fitness.

In Cohen et al. (2011), the approach to selection inference on CpG islands is predicated on the unverified assumption that germline promoter methylation is a neutral epiallele. But this need not be the case. It has long been proposed that selection can act on regional mutation rate modifiers, to reduce the mutation burden at locations where mutations are deleterious (Sturtevant, 1937; Leigh, 1970; Kimura, 1967; Kondrashov, 1995). Critically, a modifier acting on many such locations simultaneously would be able to overcome stochastic genetic drift (Lynch et al., 2016). This phenomenon has not been demonstrated for any vertebrate species. However, empirical evidence in support of it at coding sequences has recently surfaced in bacteria (Martincorena et al., 2012; Chen and J Zhang, 2013; Martincorena and Luscombe, 2013), and plants (Monroe et al., 2020), albeit without identification of the specific modifiers under selection.

RESULTS

An approach for inferring the deleteriousness of an epiallele without population-scale data

To explore whether germline promoter methylation is under selection, we designed a versatile test for detecting selection on any epiallele, provided it can be linked to a gene. This problem can be formulated as the inverse of that addressed by recent large-scale analyses of exome sequences (Petrovski et al., 2013; Lek et al., 2016; Cassa et al., 2017; Karczewski et al., 2020). These studies relied on prior knowledge about which alleles are deleterious – by virtue of being loss-of-function (LoF) – and which are neutral, to estimate the selective pressure against coding LoF alleles at a given gene.

Inverting this approach, we here assume that the heterozygous presence of an epiallele at a given gene leads to a reduction in fitness equal to ws_{het}, where s_{het} is the gene-level selection coefficient against heterozygous coding LoF alleles, and w is unknown (Methods). Critically, because the epiallele is the same across genes, we can assume that w is gene-independent, ie. that the epiallele's effect on fitness is a constant fraction of the effect of a coding LoF allele.

Testing if the epiallele is deleterious thus reduces to determining whether $w = 0$. An implication of our model is that the fitness effect of the epiallele increases as s_{het} (ie. genic LoF-intolerance) increases. This implies a monotone relationship between the population frequency of the epiallele and s_{het} if $w > 0$, and no relationship if $w = 0$. Importantly, s_{het} is now known for most human genes (Cassa et al., 2017). We describe (Methods) that it is feasible to test for $w = 0$ *without* population-scale data, in two simple steps. First, we pool genes with similar s_{het}'s together. Then, we test if the epiallele is significantly less likely to be encountered at genes where s_{het} is greater. Finally, other metrics reflective of LoF-intolerance which are approximations of s_{het} can be used as well (Methods); here we use LOEUF from gnomAD (Karczewski et al., 2020).

The absence of DNA methylation from promoters of loss-of-function intolerant genes in the male germline indicates its deleteriousness.

We applied our approach to promoter DNA methylation in the male germline, where most methylation-induced mutations have been shown to arise (Reik et al., 2001; Gao et al., 2019). We investigated its distribution with respect to downstream gene LoF-intolerance, and discovered a clear relationship: greater LoF-intolerance is associated with a smaller probability of having a methylated promoter (Figure 1a; $p < 2.2 \cdot 10^{-16}$). This increasing trend is highly consistent with what our model predicts if the presence of methylation were to be deleterious. At the most LoF-intolerant promoters (bottom 10% LOEUF), methylation is almost universally absent, with only 0.9% of promoters being methylated.

We used LOEUF as the metric of LoF-intolerance, as provided by gnomAD; smaller LOEUF values indicate greater LoF-intolerance (Karczewski et al.,

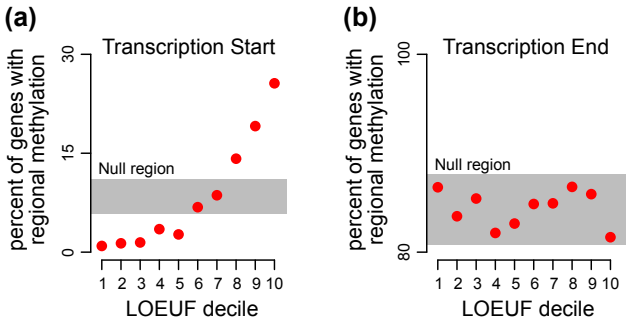


Figure 1. Testing for the deleteriousness of promoter DNA methylation in the male germline. (a) The percentage of genes with methylated promoters ($\geq$ 80% methylated CpGs in the 4kb region centered around the transcriptional start site) across LOEUF deciles. **(b)** The percentage of genes with methylated 3' ends ($\geq$ 80% methylated CpGs in the 4kb region centered around the transcriptional stop site) across LOEUF deciles. In both (a) and (b), the shaded grey area corresponds to the rejection region (two-sided, Bonferroni-adjusted for multiple testing) of the permutation null distribution obtained by randomly resampling a gene set of size equal to a LOEUF decile (10,000 times, Methods).

2020). We examined whether our result is sensitive to the choice of metric, and found that the same trend across gene deciles is observed when genes are ranked using s_{het} (Cassa et al., 2017) instead of LOEUF (Supplementary Figure S1a; $p < 2.2 \cdot 10^{-16}$). Hereafter, we use LOEUF for our analyses, but we note that the high rank correlation between the two metrics (Supplementary Figure S1b; Spearman's $\rho = -0.85$) ensures that the results are robust to this choice.

As a negative control, we tested whether germline methylation around the transcriptional stop site of the same genes is deleterious. As expected, we did not find any evidence for selection in that case (Figure 1b). The percentage of methylated stop sites was essentially constant across LOEUF deciles, consistent with our model, with minor fluctuations that did not deviate from what would be expected by random chance alone ($p = 0.61$).

To establish the above, we used a set of 7,460 human genes with reliable LOEUF and s_{het} estimates, and high-confidence promoter annotation (Methods). Using whole-genome bisulfite sequencing data from human sperm (Molaro et al., 2011; Qu et al., 2018), we grouped promoters in a binary fashion into methylated ($\geq$ 80% CpGs methylated), and hypomethylated ($\leq$ 40% CpGs methylated) (Methods). This binary categorization was motivated by the bimodal distribution of the percentage of methylated CpGs across promoters (Supplementary Figure S2).

Human-mouse comparative analysis of male germline methylomes reveals strong purifying selection on promoter methylation at loss-of-function intolerant genes.

An orthogonal test for the deleteriousness of germline promoter methylation can be conducted via a between-species comparison. A key aspect of our result is that promoter methylation is only deleterious at genes with selective pressure against LoF alleles. Therefore, hypomethylation-to-methylation changes should be suppressed by purifying selection at such LoF-intolerant genes. It is possible to distinguish this suppression from the neutral "epimutation rate", by comparing the rate of hypomethylation-to-methylation changes at LoF-intolerant genes to the same rate at LoF-tolerant genes. This approach is analogous to the classical dN/dS test for selection in coding sequences (Kimura, 1977; Kryazhimskiy and Plotkin, 2008).

We performed a human-mouse comparative analysis, using whole-genome bisulfite sequencing from mouse sperm (Hammoud et al., 2014; Qu et al., 2018). While 93.9% of promoters shared the same

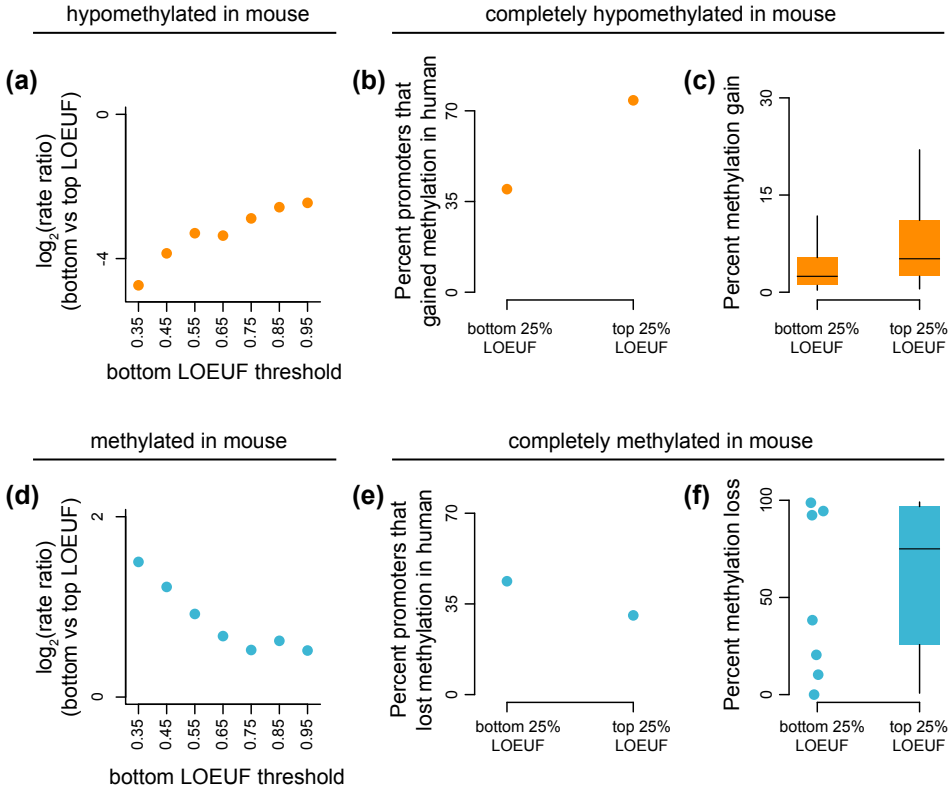


Figure 2. Testing for purifying and positive selection on promoter DNA methylation in the male germline. (a) The ratio of the rate at which promoters of loss-of-function intolerant gene promoters that are hypomethylated in the mouse germline ($\leq$ 40% methylated CpGs) have acquired the methylated state in humans, versus the corresponding rate for loss-of-function tolerant promoters, at different LOEUF thresholds for loss-of-function intolerance; loss-of-function tolerance was fixed at LOEUF $\geq$ 1 **(b)** The percentage of completely hypomethylated (0% methylated CpGs) promoters in mouse, for which at least one CpG is methylated in humans. **(c)** The distribution of the percentage of methylated CpGs in the human germline, for promoters that are completely hypomethylated in mouse and have at least one methylated CpG in human. In (b) and (c), promoters are stratified according to downstream gene loss-of-function intolerance in human. **(d)** Like (a) but for promoters methylated in the mouse germline ($\geq$ 80% methylated CpGs) that have acquired the hypomethylated state in human. **(e,f)** Like (b,c) but considering loss of methylation in human for promoters completely methylated (100% methylated CpGs) in mouse.

promoter methylation state in the two species, we identified 80 promoters methylated in mouse but hypomethylated in human, and 65 with the reverse pattern (Methods; Supplemental Figure S3).

We then focused on promoters hypomethylated in mouse ($\leq$ 40% CpGs methylated). We discovered that hypomethylation-to-methylation changes have been suppressed $\sim$ 20-fold at highly LoF-intolerant promoters (LOEUF $\leq$ 0.35, corresponding to bottom 25%) compared to LoF-tolerant ones (LOEUF $\geq$ 1, corresponding to top 25%; Figure 2a; rate ratio = 0.04, $p = 0$ after 10,000 permutations; Methods). To assess the robustness of this result, we repeatedly estimated the rate ratio after progressively relaxing the LOEUF cutoff for LoF-intolerance. As expected, the rate ratio increases as the LOEUF cutoff becomes more lenient (Figure 2a).

We next asked if this selection is operating not only at the whole-promoter level, but at the single CpG level as well. We focused on promoters completely hypomethylated in mouse (1,781 promoters with 0% CpGs methylated). At 962 of these (54%), at least one CpG has gained methylation in human. However, we found that such gains have occurred much less frequently at LoF-intolerant promoters compared to LoF-tolerant ones (Figure 2b; probability of gain = 39.8% vs 74%, $p = 0$ after 10,000 permutations). When methylation gain did occur at LoF-intolerant promoters, it was for a lower number of CpGs (Figure 2c; median = 2.45 vs 5.16, $p = 3.25 \cdot 10^{-12}$).

Finally, we turned to promoters methylated in mouse ($\geq$ 80% CpGs methylated). We discovered that methylation-to-hypomethylation changes are encountered at a higher rate at highly LoF-intolerant promoters compared to LoF-tolerant ones (Figure 2d; rate ratio = 2.82, $p = 0.0021$ after 10,000 permutations). The rate ratio progressively decreases as the LOEUF cutoff for LoF-intolerance becomes less stringent (Figure 2d). This is consistent with positive selection favoring the loss of the methylated state at these promoters. We note, however, that this positive selection appears substantially weaker than the purifying selection we describe above, as is also reflected in the rate of methylation loss of single CpGs Figure 2e, f). This may indicate that it is only happening at a subset of methylated promoters.

The absence of methylation from loss-of-function intolerant gene promoters is not explained by germline expression levels or promoter H3K4me3 levels.

Our results so far establish that germline methylation at LoF-intolerant promoters is under purifying selection. But they do not establish that the reason for its deleteriousness is its mutagenic effect on CpGs. To test if this is indeed the underlying cause, we first considered two alternative possibilities: a) that the relationship between genic LoF-intolerance and promoter germline methylation merely reflects the higher expression of LoF-intolerant genes in the germline, and b) that it is a secondary consequence of the high H3K4me3 levels at LoF-intolerant promoters in the germline, as promoter H3K4me3 has been shown to locally repel the DNA methylation machinery (Ooi et al., 2007; Vavouri and Lehner, 2011; Lesch and Page, 2014).

If either of these two expalanations is true, the association between LOEUF and promoter methylation should no longer be present when comparing genes matched with respect to their expression level, or with respect to the presence of H3K4me3 at their promoter. In contrast, we saw that the relationship persists in both cases (Figure 3a). This was true with RNA-seq measurements from both sperm and testis ($p = 1.7 \cdot 10^{-8}$, $1.5 \cdot 10^{-5}$, for genes not expressed and $p = 3.3 \cdot 10^{-16}$, $2 \cdot 10^{-8}$, for genes highly expressed; Methods), and when quantifying the presence of H3K4me3 in two different ways using ChIP-seq in sperm (Supplemental Figure S4; $p < 2.2 \cdot 10^{-16}$ for promoters without H3K4me3 and $p < 2.2 \cdot 10^{-16}$, $1.7 \cdot 10^{-11}$ for promoters with H3K4me3; Methods).

The deleteriousness of germline promoter methylation is explained by its mutagenic effect on CpGs.

We subsequently noticed that methylation leaves 48% of the genome-wide variation in promoter CpG density unexplained (Supplemental Figure S5). This suggests that additional forces acting on CpGs have generated the remaining variation, as previously suggested (Panchin et al., 2016; Goldmann et al., 2018). We thus hypothesized that, if selection is acting on methylation to indirectly maintain CpG density, then these other forces should also be selective in nature (either direct selection

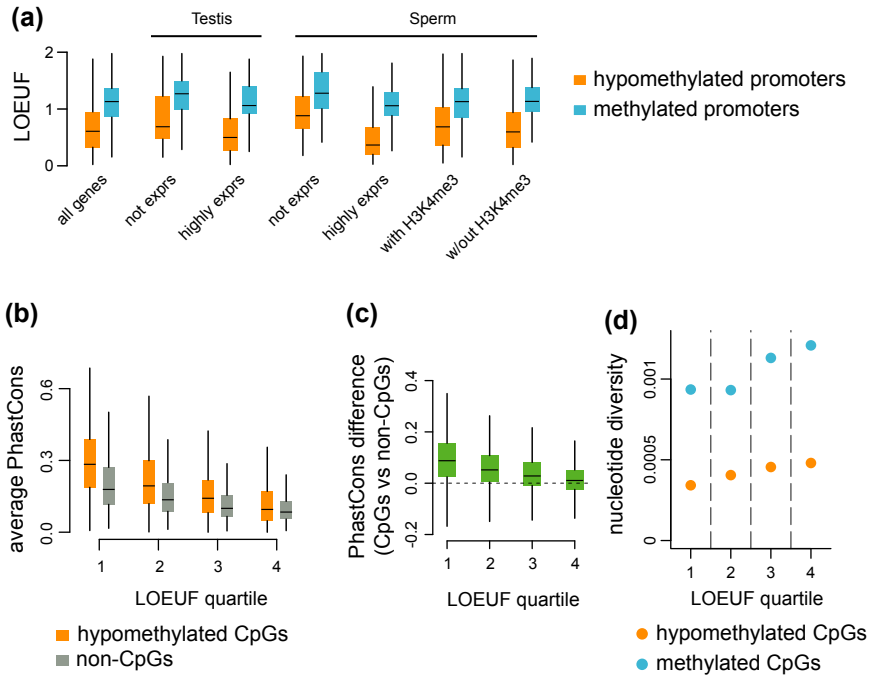


Figure 3. Evaluating whether the increased CpG mutation rate mediates the deleteriousness of promoter methylation in the male germline. (a) The LOEUF distributions of genes grouped according to promoter methylation state, and conditional on either expression level or the presence of promoter H3K4me3 in the male germline (testis and sperm; see Methods). **(b)** The per promoter average across-vertebrate conservation (PhastCons score) of hypomethylated CpG sites, stratified according to loss-of-function intolerance of the downstream gene, compared to the conservation of non-CpG sites in the same promoters. **(c)** The distribution of per-promoter differences in the average conservation of CpG sites versus that of non-CpG sites, stratified according to loss-of-function intolerance of the downstream gene. **(d)** The nucleotide diversity of promoter CpG sites in TOPMed (62,854 individuals), stratified according the their methylation status and the loss-of-function intolerance of the downstream gene.

on CpGs, or selection on other mutation rate modifiers). Importantly, they should be acting with less intensity at LoF-intolerant promoters.

To assess this, we partitioned individual CpGs according to their germline methylation state, and focused only on the 1,810,550 hypomethylated CpGs. Using the PhastCons score across 100 vertebrates (Methods), we found that the per promoter average CpG conservation increases as downstream gene LoF-intolerance increases (Figure 3b; $p < \cdot 10^{-16}$). This increase does not merely reflect generic selection on promoter sequence, because it is accompanied by a progressively larger difference between the average per promoter conservation of CpG sites versus that of non-CpG sites (Figure 3b, c; $p < \cdot 10^{-16}$). For LoF-tolerant genes (top 25% LOEUF), there is essentially no difference between CpGs and non-CpGs (median = 0.01). These results are consistent with selection operating specifically on CpGs via pathways other than methylation.

Finally, we sought to obtain a sense of how strong these other selective forces are relative to methylation. We leveraged a compendium of single-nucleotide variants from 62,874 individuals in TOPMed ((Taliun et al., 2019); Methods), and compared the intra-human genetic variability of hypomethylated versus methylated promoter CpG sites, stratified by downstream gene LOEUF. Unsurprisingly, we observe that methylated CpGs show much higher variation compared to hypomethylated ones (Figure S6d). In addition, the influence of methylation clearly dominates over LoF-

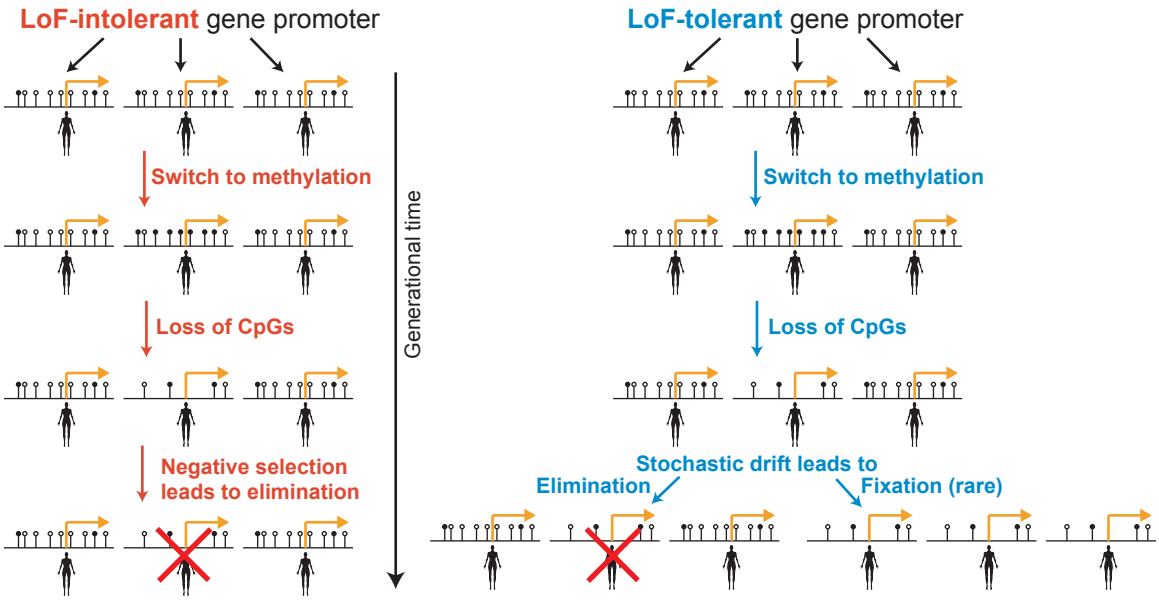


Figure 4. A model for the molecular evolution of promoter methylation in the male germline. Two scenarios are depicted. In the first (left), the promoter of a loss-of-function intolerant gene acquires the methylated state. Over time, the methylated state becomes deleterious, because it leads to loss of promoter CpGs, and is thus eliminated from the population by negative selection. In the second scenario (right), the methylated state is acquired by a previously hypomethylated promoter of a loss-of-function tolerant gene. Loss of CpGs again ensues, but this time it is neutral. Therefore, the spread of the methylated state in the population is determined by stochastic drift, which can lead either to elimination, or (rarely) to fixation. CpGs are here depicted as lollipops, with the filled-in circle indicating methylation.

intolerance (Figure S6d). This emphasizes that, as expected, the dominant factor determining the fate of CpGs is methylation, and without its exclusion a high CpG density is unlikely to be efficiently preserved.

DISCUSSION

We propose that the molecular evolution of germline DNA methylation at human promoters is governed primarily by the following principles (Figure 4). First, its effect on fitness is exerted through the elevation of CpG mutation rate and the ensuing gradual reduction in CpG density. We note that this does not preclude other consequences of methylation, for example on germline gene expression, but it means that such other consequences have no appreciable impact on fitness. Second, low promoter CpG density, and therefore the presence of germline methylation, are deleterious at LoF-intolerant promoters. It follows that hypomethylation-to-methylation changes are deleterious at such promoters, whereas methylation-to-hypomethylation changes are advantageous. At LoF-tolerant promoters, both of these changes are neutral. Finally, other factors, which have yet to be clearly identified, have a relatively small (compared to DNA methylation) but detectable contribution to the preferential maintenance of CpG density at LoF-intolerant promoters.

This model can explain two recent pieces of genetic evidence that were hard to reconcile with the notion that CpG islands are not subject to selection. First, that genes encoding for CxxC-domain-containing proteins, which bind unmethylated CpGs (Lee et al., 2001; Long, Blackledge, et al., 2013), are highly intolerant to LoF variation (Boukas et al., 2019). Second, that a high promoter CpG density is predictive of downstream gene LoF-intolerance, even more so than exonic or promoter

across-species conservation (Boukas et al., 2020).

The above description highlights two differences between the molecular evolution of germline methylation and that of the DNA sequence. The first is that a deleterious nucleotide variant occurring in the germline (e.g. in the coding or regulatory sequence of a gene) typically has effects in the next generation, and immediately becomes subject to selection. In contrast, the effects of promoter methylation on fitness likely only become non-trivial when the CpG density drops below a certain threshold. Thus, selection begins to operate on promoter methylation after a large number of generations; until then, its spread in the population is determined by stochastic drift alone. The second difference is that the promoter methylation "epimutation rate", which is equal to the rate at which the methylation state changes at LoF-tolerant promoters, appears much lower than the mutation rate at DNA sequences. This is evidenced by the fact that only $\sim$ 20% of LoF-tolerant promoters are methylated, and echoes what was recently seen in a distant unicellular species (Catania et al., 2020).

Out model raises two main questions. First, what is the biological mechanism that renders high CpG density advantageous at LoF-intolerant promoters? While it is well-described that many genes with high-CpG-density promoters show a housekeeping, or developmentally regulated, expression pattern (Saxonov et al., 2006; Lenhard et al., 2012), our previous work suggests this is unlikely to be the answer (Boukas et al., 2020). A growing body of evidence indicates the mechanism does, however, relate to the effects of CpG density on transcription (Thomson et al., 2010; White et al., 2013; Wachter et al., 2014; Morgan and Marioni, 2018; Hartl et al., 2019). Going forward, it will be necessary to more precisely characterize these effects, ultimately *in vivo*. Second, how is methylation targeted away from LoF-intolerant promoters in the germline? A number of intriguing possibilities exist, implicating transcription factor binding (Krebs et al., 2014), and a GC-rich base composition (Wachter et al., 2014). We note that the phenomenon here is analogous, albeit much more widespread in the genome, to imprinting. Recent clues into the establishment of imprinting control regions may therefore also provide useful insights (Takahashi et al., 2019).

Finally, the framework we employed to infer the deleteriousness of methylation at LoF-intolerant gene promoters is both simple and general. We thus anticipate its application to yield insights into the action of natural selection on other important epialleles and "non-nucleotide" features, including mutation rate modifiers such as replication timing.

METHODS

A simple and general framework for inferring selection on epialleles without population-scale data.

Consider a single gene with a biallelic locus with alleles A and D in the population. Let D correspond to a loss-of-function state, with the selection coefficient against heterozygotes equal to s_{het}. The fitness f is then given by

$$\begin{aligned} f(AA) &= 1 \\ f(AD) &= 1 - s_{het} \end{aligned}$$

with $0 \leq s_{het} \leq 1$. We ignore the DD genotype, since selection happens predominantly through heterozygotes (Falconer and Mackay, 1996; Fuller et al., 2019).

We extend this model with an additional biallelic locus in the same gene, with states U and M. Now, U and M are epialleles. Since D is a loss-of-function allele, we will assume that the fitness effect of the D, U and D, M combinations is the same when they reside on the same chromosome, and just use D to designate them. We assume that the fitness f is now:

$$\begin{aligned} f(AU, AU) &= 1 \\ f(AU, AM) &= 1 - ws_{het} \\ f(AM, AM) &= \max(0, 1 - kws_{het}) \\ f(AU, D) &= 1 - s_{het} \\ f(AM, D) &= \max(0, 1 - s_{het} - ws_{het}) \end{aligned} \qquad \text{(Eq. 1)}$$

with $0 \leq w \leq 1$, and $k \geq 1$. In other words, U has no fitness effect, and the fitness effect of M is a fraction of the effect of D. We note that we here present the framework in its general form. Our specific application on germline promoter methylation imposes some special features due to differences between the male and female germlines, which we discuss below (see section "Special features of the above framework in the case of germline promoter methylation").

We now consider this model across the genome, ie. across G genes indexed by $g = 1, \ldots, G$. Each gene will have its own heterozygous selection coefficient $s_{het,g}$. The critical aspect is that the epialleles M or U are the same across all genes. We can thus assume that the fitness effect of M is always the same fraction of $s_{het,g}$; in other words, w is shared amongst genes. We note that since the fitness effect of M ultimately depends on $s_{het,g}$ through the product $ws_{het,g}$, the M allele has no effect on fitness for genes with $s_{het,g} = 0$.

Thus, under the above model, testing whether the epiallele is deleterious is equivalent to testing whether $w = 0$.

Without loss of generality, we can assume that genes are ranked according to their selection coefficient, ie. $s_{het,1} \leq \ldots \leq s_{het,G}$ and therefore $ws_{het,1} \geq \ldots \geq ws_{het,G}$. Each inequality is only an equality if $w = 0$ or if $s_{het,i} = s_{het,i+1}$. Given that (epi)alleles with greater effects on fitness will have lower frequencies in the population, this implies that $\pi_1 \leq \ldots \leq \pi_G$, where π_g is the frequency of the M epiallele at gene g. Thus, we can in principle test whether $w = 0$ by assessing whether these population frequencies are equal or not.

However, because the M epiallele is the same amongst genes, it is in fact possible to test this hypothesis without population-scale data on π_g. To do so, we group the genes into K groups $(B_k, k = 1, \ldots, K)$, according to these selection coefficients (here we use deciles). Within each group, genes have similar selection coefficients (ie. for $g, g' \in B_k : s_{het,g} \approx s_{het,g'}$). Hence, the population frequencies of their M epiallele will also be similar, which we term π_{B_k}. The hypothesis $w = 0$ thus becomes equivalent to $\pi_{B_1} = \ldots = \pi_{B_K}$, and the alternative hypothesis is $\pi_{B_1} \leq \ldots \leq \pi_{B_K}$ with at least 1 strict inequality. Therefore, in the absence of population-level data on the frequencies of the M epiallele, we can estimate $\hat{\pi}_{B_k}$, the proportion of genes with the M epiallele in B_k. We then test the null hypothesis of no trend across the groups using the Cochran-Armitage test (Cochran, 1954; Armitage, 1955), with other choices possible.

In addition, to aid in the visual interpretation of the results, we obtain a null region as follows (Figure 1 and Supplemental Figure S1a, shaded grey area). We permute genes and then split them into K groups of equal size as the groups we had before. We then compute estimates $\pi_{B_1}^{\text{perm}}, \ldots, \pi_{B_K}^{\text{perm}}$ of the proportion of genes with the M epiallele in each permuted group. This is the same as essentially using the marginal (across all genes) distribution of the M epiallele, accounting for the group split. We then use this permutation distribution to define the rejection region. Because we have K groups, we use a Bonferroni adjustment to set the α level.

Note: it is not necessary to know the exact values of s_{het} for our approach to be valid; it suffices to know the rank ordering of genes. As a result, any metric serving as an approximation of s_{het} that preserves gene ranking can be used. Finally, to avoid circularity, it is essential that these metrics have been estimated without any information about the presence of the epiallele.

Special features in the case of germline promoter methylation

In the case of germline promoter methylation, it is important to clarify that M and U (in Eq. 1) correspond to the presence or absence, respectively, of DNA methylation at the promoter of the gene in the germlines of the *parents*. That is, we view the individual as containing an extra "cell type" serving as precursor to the zygote, which consists of the two parental germlines.

Our results show that the deleteriousness of promoter methylation is due to its mutagenic effect on CpGs. This implies when the M epiallele is present only in the female germline, then it is likely that $w \approx 0$, and when it is present in both germlines then $k \approx 1$. This is because the female germline becomes methylated just shortly before ovulation. Thus, methylation-induced mutations only have a short time window in which they can occur, in contrast to the male germline where methylation is present from sex determination during the embryonic development of the parent, until the conception of the offspring (Reik et al., 2001; Kobayashi et al., 2013). This is confirmed by patterns of CpG>TpG *de novo* mutation transmission in humans (Gao et al., 2019). Notwithstanding these considerations, the parameter of interest determining the presence of selection on the M epialle is still w, and inference on whether $w = 0$ can proceed as described in the above section.

Promoter coordinates

We defined promoters as 4kb regions centered around the transcriptional start site (TSS). While we recognize that there is not a universally agreed upon definition of a promoter, our previous work suggests that this is a suitable interval (Boukas et al., 2020). Importantly, even though this definition leads to the inclusion of a few exonic CpGs in some cases, we have shown that the ability of this 4kb-interval CpG density to predict downstream gene loss-of-function intolerance is driven by the CpGs in proximity to the TSS, and not by the exonic ones (Boukas et al., 2020). We obtained a set of 11,059 promoters with high-confidence GENCODE TSS annotation provided in Boukas et al. (2020). As described therein, this set does not contain subtelomeric promoters (within 2 Mb of chromosome ends), as the CpG islands of such promoters have distinct characteristics (they are organized in clusters, and are thought to be maintained principally by GC-biased gene conversion (Cohen et al., 2011)). Also excluded are promoters of genes on the sex chromosomes, for which loss-of-function intolerance estimates have a different interpretation than for autosomal genes, because of hemizygosity in males and X inactivation in females.

We here further restricted to promoters where the downstream transcript had $\geq$ 10 expected loss-of-function variants, in order to ensure that our rank ordering of genes according to the selective pressure against loss-of-function heterozygotes is not severely corrupted by genes not adequately powered for LOEUF, or s_{het}, estimation. Subsequently, bidirectional promoters were handled exactly as described in Boukas et al. (2020), yielding a set of 7,518 promoters that we used for our analyses.

Whole-genome bisulfite sequencing data from human and mouse sperm

We used processed whole-genome bisulfite sequencing data from human and mouse sperm (Qu et al., 2018). These data were accessed though the DNA methylation trackhub at the UCSC genome browser (Song et al., 2013), and consisted of methylation level (defined as the proportion of reads supporting the methylated state) and coverage. The raw experimental data consisted of two biological replicates from human (Molaro et al., 2011), and four biological replicates from mouse (Hammoud et al., 2014). For human, sequencing reads were mapped to hg19. For mouse, reads were first mapped to mm10 to generate methylation and coverage level, and the mouse CpG were subsequently aligned to their homologous position in hg19 (which need not be a CpG; see Qu et al. (2018) for details).

We only considered CpGs with at least 10x coverage. We labeled a given CpG as methylated if its methylation level was $\geq$ 80%, and hypomethylated if its methylation level was $\leq$ 20% (we note that this is a different threshold from the one used to classify a *promoter* as methylated). As orthogonal support for the methylation state of the human CpG sites, we examined their nucleotide diversity and minor allele frequency spectrum in TOPMed; reassuringly, methylated CpGs are substantially more variable than hypomethylated ones (Supplemental Figure S6a, b). Finally, we discarded CpGs with intermediate methylation level (that is, between 20% and 80%), and restricted to promoters with $\geq$ 10 CpGs (Supplemental Figures S2b, S3b show the distribution of the total number of CpGs – either methylated or hypomethylated - for all promoters in human and mouse). In total, this yielded 7,460 promoters with $\geq$ 10 CpGs in human, and 7,085 promoters with $\geq$ 10 CpGs in both human and mouse.

Genetic variation data from TOPMed

We downloaded a VCF file containing human variation data from dbSNP (version 151, hg19 assembly). We then used bedtools (Quinlan and Hall, 2010) to restrict to variants within our set of 7,518 promoter regions. We used the allele frequencies of these variants in TOPMed (freeze 5; 62,874 individuals), and only considered biallelic sites with single nucleotide variants (that is, we excluded multiallelic sites and sites with indels). We further restricted to sites where the reference allele was the major allele (allele frequency $\geq$ 0.5). These filters resulted in a total of 4,568,818 sites, of which 707,637 were CpGs. Following Asthana et al. (2007), the nucleotide diversity (π) for a given set of sites was estimated as

$$\sum_i \frac{\text{maf}_i(1-\text{maf}_i)N/(N-1)}{L}$$

where L is the total number of sites, maf_i is the minor allele frequency in site i, and N is the total num-

ber of individuals (which in our case is equal to 62,874). For sites with no variant allele, the minor allele frequency was taken to be 0.

Between species nucleotide conservation

We quantified nucleotide conservation across 100 vertebrates species with the PhastCons score (Siepel et al., 2005). We obtained these scores for nucleotides in promoters with the phastCons100way.UCSC.hg19 R package. For Figures 3b, c, we only considered promoters with ≥ 30 CpGs.

RNA-seq expression data

For expression in testis, we downloaded the gene-level TPM expression values from the GTEx v7 release (GTEx Consortium, 2017), from the GTEx portal. We restricted to the genes downstream of our 7,518 promoters. Subsequently, for each gene we computed the median (across individuals) expression in testis (in $\log_2(TPM+1)$ scale). We then labeled genes with median expression $\leq$ 0.26 (2.5th percentile, 80 and 98 genes with methylated and hypomethylated promoter, respectively) as not expressed, and genes with median expression >= 5.33 (75th percentile, 51 and 1,809 genes with methylated and hypomethylated promoter, respectively) as highly expressed.

For expression in sperm, we downloaded the raw sequencing reads from 8 healthy individuals (Swanson et al., 2020). We pseudoaligned these reads to a fasta file containing all human cDNA sequences (Homo_sapiens.GRCh37.75.cdna.all.fa; downloaded from ENSEMBL), and subsequently quantified transcript abundances with Salmon (version 0.10.0) (Patro et al., 2017). We then obtained normalized gene-level counts from the transcript abundances using the tximport R package (Soneson et al., 2015), with the "countsFromAbundance" parameter equal to "lengthScaledTPM".

We then labeled genes with median (across individuals) expression equal to 0, and maximum expression $\leq$ 0.25 as not expressed (43 and 61 genes with methylated and hypomethylated promoter, respectively), and genes with median expression $\geq$ 7.5 and minimum expression $\geq$ 5 as highly expressed (25 and 791 genes with methylated and hypomethylated promoter, respectively). We observed concordance between the genes labeled as not expressed or highly expressed in testis, and their counterparts in sperm (Supplemental Figure S7), despite the low mapping rate of sperm reads (median = 25%).

H3K4me3 ChIP-seq data

We downloaded the raw ChIP-seq reads (both from the sample treated with the H3K4me3 antibody and the input control sample) from Hammoud et al. (2009). Reads were mapped to hg19 using Bowtie2 (Langmead and Salzberg, 2012). Subsequently, we used Picard (http://broadinstitute.github.io/picard/) to remove duplicate reads, with the function MarkDuplicates. We then called peaks using MACS2 (Y Zhang et al., 2008), with the "keep-dup" parameter equal to "all".

For Figure 3a, the presence of H3K4me3 is quantified by the presence of at least one peak whose coordinates overlap the promoter coordinates. Out of the promoters with peaks, 103 were methylated and 5981 were hypomethylated. Out of the promoters without a peak, 525 were methylated and 772 were hypomethylated. We verified that the persistence of the relationship between germline promoter methylation and LoF-intolerance when conditioning on the presence or absence of promoter H3K4me3 is not an artifact of the specific parameters used for peak calling. Specifically, we used Rsubread (Liao et al., 2019) to map the reads from the antibody-treated and input control samples into our promoter regions (excluding chimeric fragments and multimapping reads), and obtained the same result using the ratio of antibody-to-control reads (Supplemental Figure S4).

Code availability

The code used for the analyses and figures is available at https://github.com/hansenlab/cpg_selection_methylation_paper_repro.

Acknowledgements

Funding: Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM121459. HTB is funded

by the Louma G. Foundation, The Icelandic Research Fund (#195835-051, #206806-051) and the Icelandic Technology Development Fund (#2010588-0611). LB was partly supported by the Maryland Genetics, Epidemiology and Medicine (MD-GEM) training program, funded by the Burroughs-Wellcome Fund.
Conflict of Interest: None.

REFERENCES

Antequera, F and Bird, A (2018). CpG Islands: A Historical Perspective. Methods in Molecular Biology *1766*, 3–13. DOI: 10.1007/978-1-4939-7768-0_1.

Armitage, P (1955). Tests for Linear Trends in Proportions and Frequencies. Biometrics *11*, 375–386. DOI: 10.2307/3001775.

Asthana, S, Noble, WS, Kryukov, G, Grant, CE, Sunyaev, S, and Stamatoyannopoulos, JA (2007). Widely distributed noncoding purifying selection in the human genome. PNAS *104*, 12410–12415. DOI: 10.1073/pnas.0705140104.

Bird, A (1987). CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics *3*, 342–347. DOI: 10.1016/0168-9525(87)90294-0.

Boukas, L, Bjornsson, HT, and Hansen, KD (2020). Promoter CpG density predicts downstream gene loss-of-function intolerance. bioRxiv, 2020.02.15.936351. DOI: 10.1101/2020.02.15.936351.

Boukas, L, Havrilla, JM, Hickey, PF, Quinlan, AR, Bjornsson, HT, and Hansen, KD (2019). Coexpression patterns define epigenetic regulators associated with neurological dysfunction. Genome Research *29*, 532–542. DOI: 10.1101/gr.239442.118.

Cassa, CA, Weghorn, D, Balick, DJ, Jordan, DM, Nusinow, D, Samocha, KE, O'Donnell-Luria, A, MacArthur, DG, Daly, MJ, Beier, DR, et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nature Genetics *49*, 806–810. DOI: 10.1038/ng.3831.

Catania, S, Dumesic, PA, Pimentel, H, Nasif, A, Stoddard, CI, Burke, JE, Diedrich, JK, Cook, S, Shea, T, Geinger, E, et al. (2020). Evolutionary Persistence of DNA Methylation for Millions of Years after Ancient Loss of a De Novo Methyltransferase. Cell *180*, 263–277.e20. DOI: 10.1016/j.cell.2019.12.012.

Chen, X and Zhang, J (2013). No gene-specific optimization of mutation rate in Escherichia coli. Molecular Biology and Evolution *30*, 1559–1562. DOI: 10.1093/molbev/mst060.

Clouaire, T, Webb, S, Skene, P, Illingworth, R, Kerr, A, Andrews, R, Lee, JH, Skalnik, D, and Bird, A (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes & Development *26*, 1714–1728. DOI: 10.1101/gad.194209.112.

Cochran, WG (1954). Some Methods for Strengthening the Common χ^2 Tests. Biometrics *10*, 417–451. DOI: 10.2307/3001616.

Cohen, NM, Kenigsberg, E, and Tanay, A (2011). Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell *145*, 773–786. DOI: 10.1016/j.cell.2011.04.024.

Cooper, DN and Youssoufian, H (1988). The CpG dinucleotide and human genetic disease. Hum. Genet. *78*, 151–155.

Coulondre, C, Miller, JH, Farabaugh, PJ, and Gilbert, W (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775–780.

Deaton, AM and Bird, A (2011). CpG islands and the regulation of transcription. Genes & development *25*, 1010–1022. DOI: 10.1101/gad.2037511.

Falconer, DS and Mackay, TFC (1996). *Introduction to Quantitative Genetics*. Pearson.

Fuller, ZL, Berg, JJ, Mostafavi, H, Sella, G, and Przeworski, M (2019). Measuring intolerance to mutation in human genetics. Nature Genetics *51*, 772–776. DOI: 10.1038/s41588-019-0383-1.

Gao, Z, Moorjani, P, Sasani, TA, Pedersen, BS, Quinlan, AR, Jorde, LB, Amster, G, and Przeworski, M (2019). Overlooked roles of DNA damage and maternal age in generating human germline mutations. PNAS *116*, 9491–9500. DOI: 10.1073/pnas.1901259116.

Goldmann, JM, Seplyarskiy, VB, Wong, WSW, Vilboux, T, Neerincx, PB, Bodian, DL, Solomon, BD, Veltman, JA, Deeken, JF, Gilissen, C, et al. (2018). Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. Nature Genetics *50*, 487–492. DOI: 10.1038/s41588-018-0071-6.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. DOI: 10.1038/nature24277.

Hammoud, SS, Low, DHP, Yi, C, Carrell, DT, Guccione, E, and Cairns, BR (2014). Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell *15*, 239–253. DOI: 10.1016/j.stem.2014.04.006.

Hammoud, SS, Nix, DA, Zhang, H, Purwar, J, Carrell, DT, and Cairns, BR (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478. DOI: 10.1038/nature08162.

Hartl, D, Krebs, AR, Grand, RS, Baubec, T, Isbel, L, Wirbelauer, C, Burger, L, and Schübeler, D (2019). CG dinucleotides enhance promoter activity independent of DNA methylation. Genome Research *29*, 554–563. DOI: 10.1101/gr.241653.118.

Karczewski, KJ, Francioli, LC, Tiao, G, Cummings, BB, Alföldi, J, Wang, Q, Collins, RL, Laricchia, KM, Ganna,

A, Birnbaum, DP, et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. DOI: 10.1038/s41586-020-2308-7.

Kimura, M (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature *267*, 275–276.

Kimura, M (1967). On the evolutionary adjustment of spontaneous mutation rates*. Genetics Research *9*, 23–34. DOI: 10.1017/S0016672300010284.

Kobayashi, H, Sakurai, T, Miura, F, Imai, M, Mochiduki, K, Yanagisawa, E, Sakashita, A, Wakai, T, Suzuki, Y, Ito, T, et al. (2013). High-resolution DNA methylome analysis of primordial germ cells identifies gender-specific reprogramming in mice. Genome Research *23*, 616–627. DOI: 10.1101/gr.148023.112.

Kondrashov, AS (1995). Modifiers of mutation-selection balance: general approach and the evolution of mutation rates. Genetics Research *66*, 53–69. DOI: 10.1017/S001667230003439X.

Krebs, AR, Dessus-Babus, S, Burger, L, and Schübeler, D (2014). High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. eLife *3*, e04094. DOI: 10.7554/eLife.04094.

Kryazhimskiy, S and Plotkin, JB (2008). The population genetics of dN/dS. PLOS Genetics *4*, e1000304. DOI: 10.1371/journal.pgen.1000304.

Langmead, B and Salzberg, SL (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359. DOI: 10.1038/nmeth.1923.

Lee, JH, Voo, KS, and Skalnik, DG (2001). Identification and characterization of the DNA binding domain of CpG-binding protein. The Journal of Biological Chemistry *276*, 44669–44676. DOI: 10.1074/jbc.M107179200.

Leigh, EG (1970). Natural Selection and Mutability. The American Naturalist *104*, 301–305.

Lek, M, Karczewski, KJ, Minikel, EV, Samocha, KE, Banks, E, Fennell, T, O'Donnell-Luria, AH, Ware, JS, Hill, AJ, Cummings, BB, et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291. DOI: 10.1038/nature19057.

Lenhard, B, Sandelin, A, and Carninci, P (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics *13*, 233–245. DOI: 10.1038/nrg3163.

Lesch, BJ and Page, DC (2014). Poised chromatin in the mammalian germ line. Development *141*, 3619–3626. DOI: 10.1242/dev.113027.

Liao, Y, Smyth, GK, and Shi, W (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research *47*, e47. DOI: 10.1093/nar/gkz114.

Long, HK, Blackledge, NP, and Klose, RJ (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochemical Society Transactions *41*, 727–740. DOI: 10.1042/BST20130028.

Long, HK, Sims, D, Heger, A, Blackledge, NP, Kutter, C, Wright, ML, Grützner, F, Odom, DT, Patient, R, Ponting, CP, et al. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. eLife *2*, e00348. DOI: 10.7554/eLife.00348.

Lynch, M, Ackerman, MS, Gout, JF, Long, H, Sung, W, Thomas, WK, and Foster, PL (2016). Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics *17*, 704–714. DOI: 10.1038/nrg.2016.104.

Martincorena, I and Luscombe, NM (2013). Response to No gene-specific optimization of mutation rate in Escherichia coli. arXiv: 1305.1436 [q-bio.GN].

Martincorena, I, Seshasayee, ASN, and Luscombe, NM (2012). Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature *485*, 95–98. DOI: 10.1038/nature10995.

Molaro, A, Hodges, E, Fang, F, Song, Q, McCombie, WR, Hannon, GJ, and Smith, AD (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell *146*, 1029–1041. DOI: 10.1016/j.cell.2011.08.016.

Monroe, JG, Srikant, T, Carbonell-Bejerano, P, Exposito-Alonso, M, Weng, ML, Rutter, MT, Fenster, CB, and Weigel, D (2020). Mutation bias shapes gene evolution in Arabidopsis thaliana. bioRxiv, 2020.06.17.156752. DOI: 10.1101/2020.06.17.156752.

Morgan, MD and Marioni, JC (2018). CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness. Genome Biology *19*, 81. DOI: 10.1186/s13059-018-1461-x.

Ooi, SKT, Qiu, C, Bernstein, E, Li, K, Jia, D, Yang, Z, Erdjument-Bromage, H, Tempst, P, Lin, SP, Allis, CD, et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717. DOI: 10.1038/nature05987.

Panchin, AY, Makeev, VJ, and Medvedeva, YA (2016). Preservation of methylated CpG dinucleotides in human CpG islands. Biol. Direct *11*, 11. DOI: 10.1186/s13062-016-0113-x.

Patro, R, Duggal, G, Love, MI, Irizarry, RA, and Kingsford, C (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods *14*, 417–419. DOI: 10.1038/nmeth.4197.

Petrovski, S, Wang, Q, Heinzen, EL, Allen, AS, and Goldstein, DB (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLOS Genetics *9*, e1003709. DOI: 10.1371/journal.pgen.1003709.

Qu, J, Hodges, E, Molaro, A, Gagneux, P, Dean, MD, Hannon, GJ, and Smith, AD (2018). Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. Genome Research *28*, 145–158. DOI: 10.1101/gr.225896.117.

Quinlan, AR and Hall, IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. DOI: 10.1093/bioinformatics/btq033.

Reik, W, Dean, W, and Walter, J (2001). Epigenetic reprogramming in mammalian development. Science *293*, 1089–1093. DOI: 10.1126/science.1063443.

Saxonov, S, Berg, P, and Brutlag, DL (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America *103*, 1412–1417. DOI: 10.1073/pnas.0510310103.

Siepel, A, Bejerano, G, Pedersen, JS, Hinrichs, AS, Hou, M, Rosenbloom, K, Clawson, H, Spieth, J, Hillier, LW, Richards, S, et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research *15*, 1034–1050. DOI: 10.1101/gr.3715005.

Soneson, C, Love, MI, and Robinson, MD (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research *4*, 1521. DOI: 10.12688/f1000research.7563.2.

Song, Q, Decato, B, Hong, EE, Zhou, M, Fang, F, Qu, J, Garvin, T, Kessler, M, Zhou, J, and Smith, AD (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PloS ONE *8*, e81148. DOI: 10.1371/journal.pone.0081148.

Sturtevant, AH (1937). Essays on Evolution. I. On the Effects of Selection on Mutation Rate. The Quarterly Review of Biology *12*, 464–467.

Swanson, GM, Estill, M, Diamond, MP, Legro, RS, Coutifaris, C, Barnhart, KT, Huang, H, Hansen, KR, Trussell, JC, Coward, RM, et al. (2020). Human chromatin remodeler cofactor, RNA interactor, eraser and writer sperm RNAs responding to obesity. Epigenetics *15*, 32–46. DOI: 10.1080/15592294.2019.1644880.

Takahashi, N, Coluccio, A, Thorball, CW, Planet, E, Shi, H, Offner, S, Turelli, P, Imbeault, M, Ferguson-Smith, AC, and Trono, D (2019). ZNF445 is a primary regulator of genomic imprinting. Genes & Development *33*, 49–54. DOI: 10.1101/gad.320069.118.

Taliun, D, Harris, DN, Kessler, MD, Carlson, J, Szpiech, ZA, Torres, R, Gagliano Taliun, SA, Corvelo, A, Gogarten, SM, Kang, HM, et al. (2019). "Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program".

Thomson, JP, Skene, PJ, Selfridge, J, Clouaire, T, Guy, J, Webb, S, Kerr, ARW, Deaton, A, Andrews, R, James, KD, et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086. DOI: 10.1038/nature08924.

Vavouri, T and Lehner, B (2011). Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. PLoS Genetics *7*, e1002036. DOI: 10.1371/journal.pgen.1002036.

Wachter, E, Quante, T, Merusi, C, Arczewska, A, Stewart, F, Webb, S, and Bird, A (2014). Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. eLife *3*, e03397. DOI: 10.7554/eLife.03397.

Wang, RY, Kuo, KC, Gehrke, CW, Huang, LH, and Ehrlich, M (1982). Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. Biochim. Biophys. Acta *697*, 371–377.

White, MA, Myers, CA, Corbo, JC, and Cohen, BA (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. PNAS *110*, 11952–11957. DOI: 10.1073/pnas.1307449110.

Woo, YH and Li, WH (2012). Evolutionary conservation of histone modifications in mammals. Molecular Biology and Evolution *29*, 1757–1767. DOI: 10.1093/molbev/mss022.

Zhang, Y, Liu, T, Meyer, CA, Eeckhoute, J, Johnson, DS, Bernstein, BE, Nusbaum, C, Myers, RM, Brown, M, Li, W, et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology *9*, R137. DOI: 10.1186/gb-2008-9-9-r137.

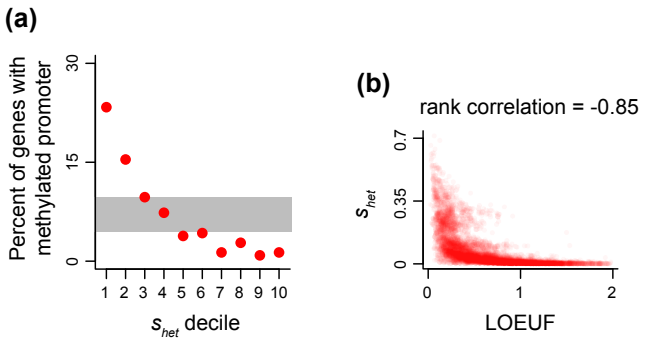
Supplementary Materials

Purifying selection acts on germline methylation to modify the CpG mutation rate at promoters

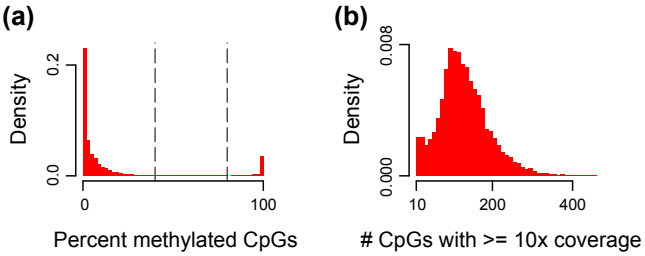
Leandros Boukas, Hans T. Bjornsson*, Kasper D. Hansen*

*Correspondence to khansen@jhsph.edu, hbjorns1@jhmi.edu

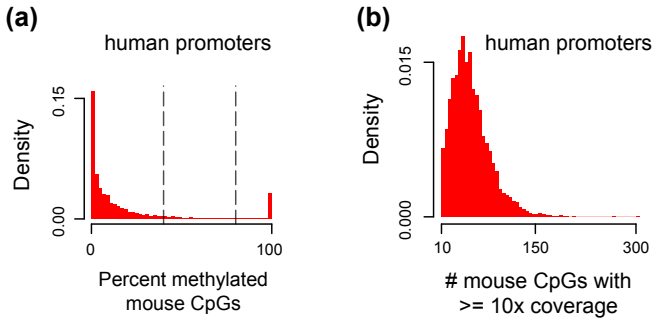
Contents



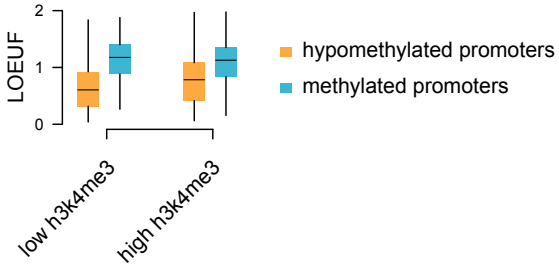
Supplementary Figure S1. Testing for the deleteriousness of promoter DNA methylation in the male germline using s_{het} estimates. (a) Like Figure 1a, but with s_{het} instead of LOEUF. **(b)** Scatterplot of LOEUF against s_{het}. Each point corresponds to a gene.



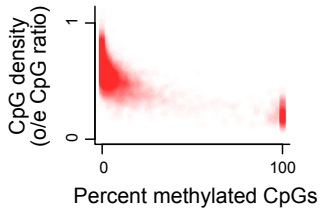
Supplementary Figure S2. The methylation state of CpGs at human promoters in the male germline. (a) The distribution of the percentage of CpGs that are methylated ($\geq$ 80% of bisulfite sequencing reads supporting the methylated state) across human promoters. The vertical lines correspond to the cutoffs used to group human promoters as hypomethylated or methylated ($\leq$ 40% and $\geq$ 80% CpGs methylated, respectively). **(b)** The distribution of the number of CpGs that were either methylated or hypomethylated ($\geq$ 80% and $\leq$ 20% of bisulfite sequencing reads supporting the methylated state, respectively), and had more than 10x coverage, across human promoters.



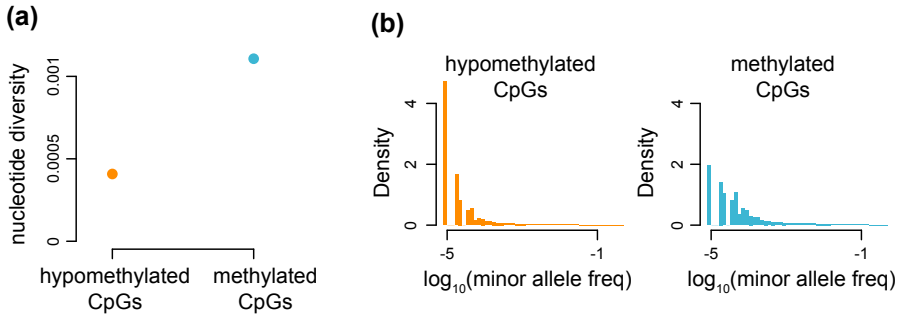
Supplementary Figure S3. The methylation state of mouse CpGs that align to human promoters in the male germline. (a) The distribution of the percentage of mouse CpGs that are methylated ($\geq$ 80% of bisulfite sequencing reads supporting the methylated state) across human promoters. For each CpG, the methylation status was quantified in mouse, and these CpGs were subsequently aligned to the human genome. The vertical lines correspond to the cutoffs used to group human promoters as hypomethylated or methylated in mouse ($\leq$ 40% and $\geq$ 80% CpGs methylated, respectively). **(b)** The distribution of the number of mouse CpGs that were either methylated or hypomethylated ($\geq$ 80% and $\leq$ 40% of bisulfite sequencing reads supporting the methylated state, respectively), and had more than 10x coverage, across human promoters.



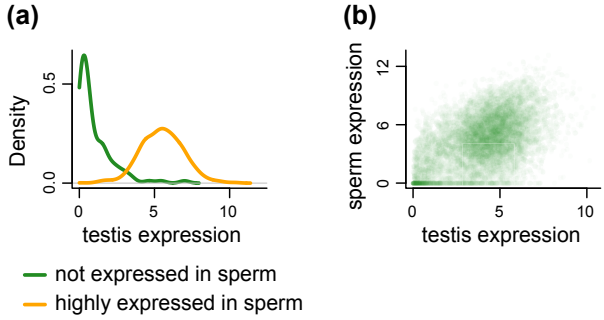
Supplementary Figure S4. The relationship between genic loss-of-function intolerance and promoter methylation in the male germline conditional on the presence of H3K4me3. Low H3K4me3/input ratio corresponds to values $\leq$ 0.45 (10th percentile), while high H3K4me3/input ratio corresponds to values $\geq$ 6.57 (90th percentile). See Methods for details on the calculation of the ratio. This figure complements Figure 1c, where the presence of H3K4me3 was determined using peak calling (Methods).



Supplementary Figure S5. The relationship between human promoter CpG density (o/e CpG ratio) and methylation in the male germline. Each point corresponds to a promoter. A given CpG was classified as methylated if $\geq$ 80% of bisulfite sequencing reads supported the methylated state. Only CpGs with $\geq 10x$ coverage and methylation level $\geq$ 80% or $\leq$ 20% were considered. The CpG density of a given promoter was computed as described in Boukas et al. (2020).



Supplementary Figure S6. Genetic diversity of promoter CpG sites in TOPMed. (a) The nucleotide diversity of all promoter CpG sites, stratified according to whether they are methylated or hypomethylated in the male germline ($\geq$ 80% and $\leq$ 20% of bisulfite sequencing reads supporting the methylated state, respectively). Nucleotide diversity was estimated as described in methods. Only CpGs with $\geq 10x$ coverage are considered. **(b)** The minor allele frequency spectrum of the CpGs used in (a).



Supplementary Figure S7. Assessing the concordance between gene expression in testis and sperm. (a) The distribution of gene expression (median across individuals in $log_2(TPM+1)$ scale) in testis, for genes either not expressed, or highly expressed, in sperm. **(b)** Scatterplot of expression in testis against expression in sperm. Each point corresponds to a gene. Testis expression data were obtained from the GTEx consortium, and sperm expression data (raw RNA-seq reads) were obtained from Swanson et al. (2020) (Methods).