1 TITLE

2 Mechanisms of DNA repair have evolved to minimise the probability of nonsense mutations.

3 AUTHORS

4 George Powell^{*1,2}, Helen S Long^{1,2}, Thomas M M Versluys³, Ann-Marie Mallon², Cecilia M
5 Lindgren^{1,4,5}, Michelle M Simon²

6 *Denotes corresponding author

7

8 AFFILIATIONS

9 1. Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,
10 University of Oxford, Oxford, UK

11 2. MRC Harwell Institute, Mammalian Genetics Unit, Oxfordshire, UK, OX11 0RD

12 3. Department of Life Sciences, Silwood Park Campus, Imperial College London,
13 Ascot, UK, SL5 7PY

14 4. Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

15 5. Medical and Population Genetics Program, Broad Institute of MIT and Harvard,
16 Cambridge, MA, USA.

17

18

19 ABSTRACT

20

Variation in sequence mutability has important implications for evolutionary models and predicting disease occurrence, and is driven in part by evolutionary divergence in mechanisms of DNA repair. The aim of this study was twofold: first, to assess the effect of local sequence context on substitution rates in the mouse lineage; second, to investigate the relationship between sequence mutability and selection. We show that the 7-mer context (i.e. three bases either side of the base of interest) explains more variation in substitution rates between chromosomes in the mouse lineage than either the 3-mer, 5-mer, or 9-mer contexts. Furthermore, we also show that 7-mer substitutions with the potential to cause nonsense mutations when they occur in translated sequences occur at a lower rate across the genome than 7-mer substitutions with the potential to cause synonymous mutations. We propose that mechanisms of DNA repair have evolved to prioritise substitutions that are more likely to be deleterious to fitness.

33

INTRODUCTION

35

Germline mutations are the basis of evolution and cause of inherited Mendelian and complex disease. Most mutations are either neutral or deleterious to fitness, and therefore organisms have evolved mechanisms of repairing DNA to maintain sequence fidelity (Harris and Pritchard 2017). These mechanisms are complex, involving hundreds of genes in eukaryotes (Lindahl and Wood 1999; Ségurel, Wyman, and Przeworski 2014), and are known to diverge between lineages over evolutionary time, resulting in variation in sequence mutability (Harris 2015; Kumar and Subramanian 2002; Lindsay et al. 2019). Quantifying this variation in sequence mutability is important for parameterising evolutionary models, interpreting patterns of variation, and identifying mutational signatures and pathogenic variants

45 (Aggarwala and Voight 2016; Samocha et al. 2014; Karczewski et al. 2020; Helleday et al.
46 2008).

47

48 Variation in sequence mutability is influenced by local sequence context. For example,
49 substitution rates are affected by a base's k-mer context, which describes the composition of
50 its flanking nucleotides (Harris and Pritchard 2017; Aikens, Johnson, and Voight 2019;
51 Schroeder et al. 2016). In human populations, the 7-mer context (ie three bases either side of
52 the base of interest) has been shown to account for more variation in substitution rates than
53 either the 3-mer or 5-mer context (Aggarwala and Voight 2016). Variation in mutability
54 between k-mers is partly driven by the chemical properties of bases and their epigenetic state
55 (Ségurel, Wyman, and Przeworski 2014; Gonzalez-Perez, Sabarinathan, and Lopez-Bigas
56 2019; Makova and Hardison 2015). For example, the higher mutability of C-G relative to
57 A-T base pairs is partly attributed to the difference in their number of connecting hydrogen
58 bonds ((Petruska and Goodman 1985; Hardison et al. 2003); (Ségurel, Wyman, and
59 Przeworski 2014)), and C > T substitutions are enriched amongst CG dinucleotides due to
60 methylation and spontaneous cytosine deamination (Shen, Rideout, and Jones 1994; Duncan
61 and Miller 1980). However, variation in k-mer mutability is also affected by mechanisms of
62 DNA repair, which can diverge over evolutionary time. For example, the TCC > TTC
63 substitution rate has increased by 50% in the human European lineage since Europeans began
64 diverging from Asians as a result of changes to DNA repair mechanisms (Harris 2015). The
65 cause of these changes remains unclear, but they are tentatively linked to differences in UV
66 exposure. TCC > TTC is the substitution most frequently induced by UV and present in
67 melanoma skin cancers, and it is speculated that decreased exposure to UV in the European
68 lineage may have reduced the selection pressure on mechanisms for repairing TCC > TTC
69 substitutions, resulting in their relative increase in rate of occurrence (Harris 2015).

70 Nevertheless, determining how selection may have affected sequence mutability through the
71 evolution of mechanisms of DNA repair remains poorly understood and is difficult to test
72 empirically.

73

74 Here, we assess the effect of local sequence context on substitution rates across the mouse
75 genome using genetic variation between mice sampled from wild populations (Harr et al.
76 2016). Mice are the most commonly used mammalian model organism for biomedical
77 research, and offer important insights into human biology (Brown et al. 2018). As mutation
78 rates diverge over evolutionary time, quantifying sequence mutability in the mouse lineage
79 can be used to parametrize mouse-specific models for inferring functional genomic regions
80 under negative selection, and identify mutational signatures and pathogenic variants
81 (Karczewski et al. 2020; Helleday et al. 2008; di Iulio et al. 2018). Our research has two main
82 aims. First, we assess the contribution of flanking nucleotides to the ability to predict
83 substitution mutations in the mouse lineage. We show that the 7-mer context explains more
84 variation in substitution occurrence between chromosomes than either the 3-mer, 5-mer, or
85 9-mer contexts. Second, we assess the relationship between sequence mutability and
86 selection. Specifically, we hypothesise that substitutions that are more likely to be deleterious
87 to fitness have lower base rates of occurrence. In support of this hypothesis, we show that
88 7-mer substitutions with the potential to cause nonsense mutations when they occur in
89 translated sequences occur at a lower rate across the genome than 7-mer substitutions with
90 the potential to cause synonymous mutations. We propose that this variation in mutability is
91 the result of selection on mechanisms of DNA repair.

92

93 RESULTS

94 The effect of local sequence context on substitution rates.

95

96 We assessed the effect of local sequence context on the ability to predict substitution
97 mutations in the mouse lineage. We hypothesised that considering a greater number of
98 flanking nucleotides either side of the base of interest (ie expanding the sequence context)
99 would improve the ability to predict substitutions across the mouse genome.

100

101 We considered six different k-mer models to account for local sequence context
102 (supplementary table 1): the 1-mer, 3-mer, 5-mer, 7-mer, 9-mer and A-CG models. The 1-mer
103 model is the simplest model and does not consider the local sequence context of the base of
104 interest. It describes 12 substitution rates, as each of the four bases has three possible
105 substitutions. The 3-mer model considers one flanking nucleotide either side of the base of
106 interest, making a 3-mer. It describes 192 substitution rates as there are 64 possible 3-mers
107 and the middle base in each 3-mer has three possible substitutions. The 5-mer, 7-mer, and
108 9-mer models consider 2, 3, and 4 flanking nucleotides either side of the base of interest,
109 respectively. The number of substitution rates that models describe increases exponentially
110 as the number of flanking nucleotides expands, with the 9-mer model describing 786,432
111 substitution rates as there are 262,144 possible 9-mers and the middle base in each 9-mer has
112 three possible substitutions. Substitution probabilities are known to be largely driven by the
113 presence of CG dinucleotides, which have a high mutation rate from cytosine to thymine (M.
114 Li and Chen 2011). We therefore expanded our simple 1-mer model to consider CG context,
115 and named this the A-CG model. The A-CG model describes 18 substitution rates as there are
116 6 possible contexts (A, T, C (non-CG), C (CG), G (non-CG), and G (CG)), each with three

possible substitutions. For each k-mer model we additionally calculated the probability of any mutation as the sum of each potential substitution for each k-mer.

To estimate substitution rates we considered autosomal single nucleotide variants (SNVs) from populations of wild mice (*Mus musculus* sp. and *Mus spretus*) (Harr et al. 2016), and inferred the ancestral state of each allele from alignment with the *Mus caroli* reference genome. We filtered out 75.3% of the mouse genome, including genomic regions with low sequence coverage, repetitive elements, and regions with functional annotation such as coding and regulatory features that are likely to be under selection (see methods for more detail). We assessed the accuracy of each k-mer model to predict substitution occurrence using train/test resampling by chromosome. In brief, fifteen autosomes were randomly selected to calculate k-mer probabilities of substitution and the remaining four autosomes were used to test predictive accuracy using the absolute error between the observed substitution rates (test autosomes) and expected substitution rates (training autosomes). This was repeated 100 times to provide the mean absolute error (MAE) and 95% confidence intervals. Each ancestral k-mer and k-mer change has a complement on the reverse strand. For example, the ancestral 3-mer TAG on the forward strand must have the complementary k-mer CTA on the reverse strand, and a TAG > TTG substitution must result in a complementary CTA > CAA substitution. We therefore totaled the counts of each forward and reverse k-mer and k-mer change to calculate substitution probabilities, and report the substitution probabilities for adenine and cytosine only. This results in six substitution classes under consideration (noted as A>C, A>G, A>T, C>A, C>G, and C>T), plus the occurrence of any adenine or cytosine substitution (noted as A>* and C>*).

141 Of the k-mer models considered, the 7-mer context provides the greatest increase in
 142 predictive accuracy relative to the 1-mer and A-CG models for each substitution class in the
 143 mouse lineage (figure 1). The 7-mer model reduces mean absolute error (MAE) by $73.7\% \pm$
 144 0.3% relative to the 1-mer model for cytosine substitutions ($C > *$). This is largely driven by
 145 the presence of CG dinucleotides and the frequency of $CG > TG$ substitutions, with the A-CG
 146 model reducing MAE by $67.0\% \pm 0.3\%$ for cytosine substitutions relative to the 1-mer model
 147 (figure 1). The 7-mer model, however, reduces MAE for cytosine substitutions by 20.5%
 148 $\pm 0.9\%$ relative to the A-CG model, indicating the additional nucleotides either side of a base
 149 of interest affect its substitution rate. This pattern is consistent for adenine substitutions ($A >$
 150 $*$), with the 7-mer model reducing MAE by $34.5\% \pm 0.7\%$ relative to the 1-mer model.

151

152 The 9-mer model explains less variance than the 7-mer model, suggesting that a further
 153 increase in bases does not provide further information to account for substitution occurrence
 154 (figure 1). However, it is possible that due to the increased number of parameters in the 9-mer
 155 model there is insufficient statistical power to detect the improvement. To test this, we
 156 simulated a null dataset where the counts of each 9-mer on each chromosome is equal to the
 157 observed counts, and the probability of substitution is dependent on the 9-mer context (see
 158 methods). Given this simulated dataset, we found the MAE was lower for the 9-mer than the
 159 7-mer model (supplementary figure 2) for all substitution classes except $C > T$. This suggests
 160 we were not lacking statistical power to detect an effect of the simulated magnitude for each
 161 substitution class other than $C > T$. The distributions of k-mer and k-mer substitution counts
 162 are provided in supplementary table 8.

163

164 Genomic regions with high GC content have higher mutation rates (Ségurel, Wyman, and
 165 Przeworski 2014). This is hypothesised to result from decreased efficiency of exonuclease

166 activity in separating C-G base pairs relative to A-T in order to excise erroneous nucleotides,
 167 because A-T base pairs are bound by two hydrogen bonds whereas C-G base pairs are bound
 168 by three (Petruska and Goodman 1985; Hardison et al. 2003). We fit a generalised linear
 169 model to predict the number of cytosine or guanine bases in the 7-mer as a function of its
 170 substitution rate, and found a significant positive relationship (estimate = 0.82 ± 0.03 , $z = 25.7$,
 171 $p = 2.66e-167$) (supplementary figure 4). To test whether this relationship is inflated by CG
 172 deamination we fit another model which excluded all CG>TG substitutions, and found a
 173 significant positive relationship remained (estimate = 1.90 ± 0.12 , $z = 15.7$, $p = 1.18e-55$).
 174 These results highlight that the relationship between GC content and mutability persists at a
 175 finer scale, and are consistent with the hypothesis that higher substitution rates are caused in
 176 part by decreased efficiency of exonuclease activity in separating C-G base pairs relative to
 177 A-T in order to excise erroneous nucleotides.

178

Figure 1 -- A) 1-mer substitution rates by class. C > T substitutions occur at the greatest, largely driven by the mutability of CG > TG dinucleotides. B) Mean absolute error (MAE) per kb relative to the 1-mer model for predicting adenine substitutions. The 7-mer model provides the greatest improvement in MAE relative to the 1-mer model for all substitution classes, with the greatest improvement for A > G substitutions (36.5% $\pm$ 0.7%). C) Mean absolute error (MAE) per kb relative to the 1-mer model for predicting cytosine substitutions. The A-CG model decreases MAE by 67.0% $\pm$ 0.3% for C > T substitutions relative to the 1-mer model by considering the CG context. D) Mean absolute error (MAE) per kb relative to the A-CG model for predicting cytosine substitutions. The 7-mer model provides the greatest improvement in MAE relative to the A-CG model for all substitution classes, with

190 the greatest improvement ($22.7\% \pm 0.9\%$) shown for C > T substitutions. Any potential
191 substitution is denoted with a *.

192

193 Methylation and CG dinucleotide substitution rates.

194

195 Methylation state affects the probability of substitutions within CG dinucleotides (figure 2).
196 CG dinucleotides are highly mutable across the genome, particularly C > T substitutions
197 resulting from deamination at methylated sites (Shen, Rideout, and Jones 1994; Duncan and
198 Miller 1980). We quantified the effect of methylation state on the CG mutability in the mouse
199 lineage. We considered all CG dinucleotides across unmasked regions of the genome, and
200 inferred methylation status in the germline from bisulfite sequencing in embryonic stem cells,
201 which show a similar pattern of methylation to sperm cells in mice (Popp et al. 2010). In
202 total, 72.9% of CGs are classified as methylated, 6.9% as unmethylated, and 20.2% were
203 unclassified due to insufficient coverage. The C > T substitution rate in methylated CG
204 dinucleotides is 0.37 compared with 0.04 in unmethylated CGs, an 8.3 fold increase. By
205 comparison, the C > T substitution rate is 0.35 across all CGs independent of methylation
206 status. This is closer to the substitution rate for methylated CGs as there is a higher
207 proportion of methylated CGs relative to unmethylated CGs in our sample. Accounting for
208 methylation state improves prediction accuracy for all CG substitution classes, with the effect
209 greatest for C > T substitutions (figure 2). MAE is reduced by $82.5\% \pm 0.25\%$ when
210 predicting the occurrence of C > T substitutions across methylated CGs, and $99.7\% \pm 0.1\%$
211 across unmethylated CGs, relative to a null model where methylation status is not considered.
212 The improvement in MAE is greater for unmethylated CGs relative to methylated CGs
213 because the null substitution rate (calculated across all CGs independent of methylation

status) has a higher proportion of methylated CGs relative to unmethylated CGs. The improvement in MAE by considering methylation status is greater than consideration of the 7-mer context alone, which improves prediction accuracy of C > T substitutions by 22.7%±0.9% relative to the A-CG model, further highlighting the importance of methylation state to predicting C > T substitution occurrence.

Figure 2 -- A) Substitution rates of methylated (CG+), unmethylated (CG-), and all (null) dinucleotides by class. Methylated CG dinucleotides have a C > T substitution rate of 0.37 compared with 0.04 for unmethylated CG dinucleotides. In total, we classified 72.9% of CG dinucleotides across the unmasked genome as methylated, 6.9% are unmethylated, and 20.2% are unclassified because of low coverage. B) Mean absolute error (MAE) per kb relative to the null model, in which methylation status was not considered, for methylated (CG+) and unmethylated (CG-) dinucleotides. Considering methylation status reduces MEA for each substitution class. The greatest decreases are for C > T substitutions, with a MAE decrease of -82.5%±0.2.5% and -99.7%±0.1% for methylated and unmethylated CG dinucleotides, respectively.

231

232 Mutational effect and probability of occurrence

233

234 Mechanisms for DNA repair affect rates of substitution and can diverge over evolutionary
235 time (Harris 2015; Harris and Pritchard 2017). We hypothesise that k-mer substitutions that
236 are more likely to be deleterious to fitness will occur at lower frequencies across the genome
237 resulting from selection on mechanisms of DNA repair. For example, if $AAC > ACC$
238 substitutions are more likely to be deleterious to fitness than $AAG > ACG$, there would be a
239 stronger selection pressure on DNA repair mechanisms to repair $AAC > ACC$ compared with
240 $AAG > ACG$, resulting in a relative decrease of $AAC > ACC$ substitutions. To assess this we
241 looked at the potential of 7-mer substitutions to cause nonsense and synonymous mutations if
242 they occur in translated regions of the genome, as nonsense mutations tend to be under strong
243 negative selection relative to synonymous mutations. Specifically, we tested three
244 hypotheses. First, 7-mer substitutions with the potential to cause nonsense mutations if they
245 occur in translated sequences have lower substitution rates across the genome than 7-mers
246 substitutions with the potential to cause synonymous mutations. Second, the number of
247 potential nonsynonymous mutations for each 7-mer substitution is negatively correlated with
248 its substitution rate, and conversely the number of potential synonymous mutations for each
249 7-mer substitution is positively correlated with the substitution rate. Finally, rates of 7-mer
250 substitutions with the potential to cause nonsense mutations will be more conserved between
251 species than rates of 7-mer substitution with the potential to cause synonymous mutations.

252

253 We first tested whether 7-mer substitutions with the potential to cause nonsense mutations if
254 they occur in translated sequences have lower substitution rates across the genome (ie across

translated and untranslated regions) than 7-mers substitutions with the potential to cause synonymous mutations. To test this hypothesis, we grouped 7-mer substitutions by their potential to cause nonsense, missense, or synonymous mutations if they occur in translated sequences (supplementary figure 3) and tested the difference in observed substitution rates between groups. Substitution rates were calculated from the unmasked regions of the genome as previously described and excluded translated sequences (see methods). Nonsense or missense mutations are more likely to be under negative selection than synonymous mutations, and we therefore expect 7-mer substitutions with the potential to cause nonsense or missense mutations in translated sequences to have lower substitution rates than 7-mers substitutions with the potential to cause synonymous mutations. 7-mer substitutions with the potential to cause nonsense mutations when in translated regions have significantly lower rates of occurrence across the genome than those with the potential to cause synonymous mutations ($W = 124021674$, $p = 5.8e-124$)(figure 3). When considered across all substitution classes, this relationship is largely driven by the inability of A > G substitutions, which have on average higher substitution rates than either A > C or A > T substitutions, to cause nonsense mutations (figure 3). Differences do, however, remain when controlling for substitution class (ie A>C, A>G, A>T, C>A, C>G, and C>T) (supplementary table 9). For each substitution class, 7-mer substitutions with the probability of causing nonsense mutations have significantly lower probabilities of occurrence than 7-mer substitutions with the probability of causing synonymous mutations (supplementary table 9) ($p < 0.001$). There are more potential missense substitutions ($n = 415$) than nonsense ($n = 23$) or synonymous ($n = 138$), and all 7-mer substitutions ($n = 24576$) have the potential to cause at least one missense mutation. Potential to cause a missense substitution therefore describes the distribution of all 7-mer substitution rates.

279

280 7-mer substitutions have the potential to cause multiple different mutation types when they
 281 occur in translated sequences (supplementary figure 3). For example, each 7-mer substitution
 282 can have six different consequences as they have the potential to change each of the three
 283 bases in a codon, and this applies to both the forward and reverse strands. We expect 7-mer
 284 substitutions with a higher probability of causing a nonsynonymous mutation to have lower
 285 rates of substitution, and 7-mer substitutions with a higher probability of causing a
 286 synonymous mutation to have higher rates of substitution. To test this we counted the number
 287 of potential synonymous and nonsynonymous mutations for each 7-mer substitution
 288 (supplementary figure 3). We grouped missense and nonsense substitutions to increase our
 289 power to detect a relationship, as the maximum number of nonsense mutations for any 7-mer
 290 substitution is two. We fit poisson generalised linear models to predict the number of each
 291 mutation type as a function of substitution rate for each 7-mer substitution. 7-mer
 292 substitutions with greater potential to cause nonsynonymous mutations have significantly
 293 lower substitution rates (coefficient = -0.67 ± 0.05 , $z = -14.1$, $p = 4.2e-45$), and 7-mer
 294 substitutions with greater potential to cause synonymous mutations have significantly higher
 295 substitution rates (coefficient = 0.85 ± 0.06 , $z = 13.3$, $p = 1.5e-40$).
 296
 297 Finally, we reasoned that if 7-mer substitution rates are under selection (or rather mechanisms
 298 of DNA repair mediating substitution rate) resulting from their potential to cause different
 299 mutation types in translated sequences, we would expect to see less variation in substitution
 300 rates between species for 7-mers with the potential to cause nonsense mutations relative to
 301 synonymous. This is because we expect negative selection to have minimised the substitution
 302 rates of potentially nonsense causing 7-mers in both lineages, whereas the substitution rates
 303 of potentially synonymous causing 7-mers if under weaker selection would be more likely to
 304 vary due to genetic drift. To test this we calculated rank correlation between human and

305 mouse substitution rates for two groups: 7-mers with the potential to cause nonsense
306 mutations and 7-mers with the potential to cause synonymous substitutions. For each
307 substitution class, there is a significant correlation ($p < 3.78e-126$) and the correlation
308 coefficient is greater for 7-mers with the potential to cause nonsense mutations than for
309 7-mers with the potential to cause synonymous substitutions (supplementary table 10). This
310 highlights less variation in 7-mer substitution rates between species for those with the
311 potential to cause nonsense mutations relative to synonymous.
312

Figure 3 -- 7-mer substitutions with the potential to cause nonsense and missense mutations when they occur in translated sequences have significantly lower probabilities of occurrence than 7-mer substitutions with the potential to cause synonymous mutations (A). This is largely driven by the inability of A > G substitutions, which have a higher substitution rate (B), to cause nonsense mutations (C). A) Distributions of substitution rates for 7-mers with

319 the potential to cause synonymous, missense, and nonsense variation. “A > *” and “C > *”
 320 denote all possible adenine and cytosine substitutions and their complementary thymine and
 321 guanine substitutions, respectively. “(A|C) >” encompasses all possible substitutions.
 322 Differences between groups were assessed with Wilcoxon signed-rank tests, with ***
 323 denoting a p value less than 0.0001. Outliers, defined by being outside 1.5 times the
 324 interquartile range above the upper quartile and below the lower quartile, are not shown. B)
 325 7-mer substitution rates by class. Outliers, defined by being outside 1.5 times the interquartile
 326 range above the upper quartile and below the lower quartile, are not shown. C) Fraction of
 327 7-mer substitutions by type with the potential to cause nonsense, missense, and synonymous
 328 variants. All 7-mer substitutions have the potential to cause a missense mutation.
 329

330 DISCUSSION

331
 332 We have assessed the effect of local sequence context on substitution rates across the mouse
 333 genome, and the relationship between sequence mutability and selection. We have two main
 334 results: first, the three flanking nucleotides either side of the base of interest (its 7-mer
 335 context) explain the most variability in substitution rates between chromosomes in the mouse
 336 lineage; second, 7-mer substitutions that are more likely to be highly deleterious to fitness
 337 because they have the potential to cause nonsense mutations when they occur in translated
 338 sequences are less likely to occur across the genome.

339
 340 The importance of 7-mer context to substitution occurrence has important implications for
 341 evolutionary models. In particular, models that aim to quantify the strength of selection from
 342 the relative enrichment or depletion of variation must control for regional differences in

343 mutability (Samocha et al. 2014; di Iulio et al. 2018). Failure to do this may lead to false
 344 inferences of selection. For example, regions that are less mutable are likely to contain fewer
 345 variants, independent of selection. Our finding that 7-mer context improves predictive
 346 accuracy of substitution occurrence in mice is consistent with human studies reporting similar
 347 findings (Aggarwala and Voight 2016). This suggests that the causes of 7-mer context
 348 importance have been conserved between the species. These causes remain unclear, but they
 349 could involve both fundamental properties of DNA and DNA repair mechanisms. For
 350 example, sequence context may affect the ability of DNA polymerase to separate DNA in
 351 order to remove and replace an erroneous base. Erroneous bases that are surrounded by C or
 352 G nucleotides may be less likely to be corrected, as C-G nucleotides are joined by three
 353 hydrogen bonds and therefore require more energy to separate than A-T nucleotides, which
 354 are bound by two (Petruska and Goodman 1985; Hardison et al. 2003). The 7-mer context
 355 could account for the distance from an erroneous base by which DNA must be separated in
 356 order for repair. It remains to be determined whether the 7-mer context holds the same
 357 importance for non-mammalian species that are more distantly related to humans, such as
 358 drosophila.

359

360 Variation in sequence mutability between lineages captured by local sequence context
 361 suggests evolutionary divergence in mechanisms of DNA repair, with some lineages repairing
 362 specific substitutions more or less effectively than others (Ségurel, Wyman, and Przeworski
 363 2014; Harris 2015; Kumar and Subramanian 2002). Evolutionary theory suggests the
 364 complexity of DNA proofreading and repair mechanisms in eukaryotes may limit the impact
 365 of selection on shaping the efficiency of individual genes in performing these processes
 366 (Sung et al. 2012; Lynch 2011). For example, if a variant changes the mutation rate in a given
 367 sequence context by a small amount then the effect on fitness may fall below the threshold at

which selection is effective. Therefore, the fixation of the variant in the population would be dependent on genetic drift. This line of reasoning has led to the hypothesis that drift is the predominant cause of variation in mutability between species (Sung et al. 2012). We have shown, however, that 7-mer substitutions that are more likely to be deleterious to fitness (by having the potential to cause nonsense mutations in translated sequences) are less likely to occur across the genome. This pattern is consistent across each substitution class, and between humans and mice, indicating that when the strength of selection is strong enough, natural selection affects mutability. The genetic code that determines nonsense and synonymous mutations is universal to almost all species (Knight, Freeland, and Landweber 2001), but understanding whether selection drives further variation in mutability between lineages requires further research. For example, it would be interesting to identify k-mers known to be under lineage-specific negative selection, and determine whether this is reflected in their substitution rates.

381

382 METHODS

383 Data download and analysis software

384

We considered genetic variation between 67 mice (*Mus musculus* sp. and *Mus spretus*), sampled from 11 wild populations (Harr et al. 2016). Variant (vcf files) and coverage (bam files) information were downloaded from <http://wwwuser.gwdg.de/~evolbio/evolgen/wildmouse/>. The soft-masked mouse reference genome (GRCm38), *Mus musculus* to *Mus caroli* pairwise alignment (Herrero et al. 2016), GenCode exonic annotations (Harrow et al. 2012), and the Ensembl Regulatory Build

391 (Zerbino et al. 2015) were all downloaded from Ensembl ftp
 392 (<ftp://ftp.ensembl.org/pub/release-101/>) (v101). GERP conserved regions were calculated by
 393 Ensembl across 111 mammalian species alignments, and were downloaded from Ensembl ftp
 394 (v101) (Herrero et al. 2016). Genome-wide methylation of CG dinucleotides in mouse
 395 embryonic stem cells was quantified by Stadler et al (2011) using bisulfite sequencing, and
 396 were downloaded from Ensembl ftp
 397 ([http://ftp.ensembl.org/pub/data_files/mus_musculus/GRCm38/dna_methylation_feature/NP](http://ftp.ensembl.org/pub/data_files/mus_musculus/GRCm38/dna_methylation_feature/NP_C_5mC_Publication_Stadler2011_Pmid22170606/)
 398 [C_5mC_Publication_Stadler2011_Pmid22170606/](http://ftp.ensembl.org/pub/data_files/mus_musculus/GRCm38/dna_methylation_feature/NP_C_5mC_Publication_Stadler2011_Pmid22170606/)) (Stadler et al. 2011). Human 7-mer
 399 substitution rates were downloaded from the supplementary information of Aggarwala and
 400 Voight (2016).

401

402 All analyses were conducted using custom scripts written in R (v3.6.2) (R Core Team 2017)
 403 and bash, and are available on the following github repository:
 404 https://github.com/powege/Thesis_workflow.git.

405

406 Filtering variants and genomic regions

407

408 We built a mask for the mouse genome (GRCm38) to filter low coverage regions, repetitive
 409 elements, and regions with functional annotation that are likely to be under selection. We
 410 filtered low coverage loci defined as having less than 10X coverage in more than 10% of
 411 individuals in the sample. Read depth at each genomic position for each mouse in the sample
 412 was calculated from the bam coverage files using samtools (H. Li et al. 2009). We filtered
 413 repetitive elements and ambiguous (N) bases by excluding all soft-masked regions classified
 414 using RepeatMasker by Ensembl (Zerbino et al. 2018). We also filtered regions likely to be

under selection. These included functional regions annotated as either “promoter”, “promoter flanking”, “enhancer”, “CTCF binding site”, “TF binding site”, or “open chromatin” by the Ensembl Regulatory Build (v101) (Zerbino et al. 2015); all exonic regions annotated by GenCode (Harrow et al. 2012); and GERP evolutionarily conserved regions (Herrero et al. 2016). In total we masked 75.3% of the mouse genome.

We considered all 41,000,862 single nucleotide variants (SNVs) with “PASS” filter status across unmasked regions of the genome.

Local sequence context models and probabilities of substitution

We considered six k-mer models for determining the effect of local sequence context on substitution probabilities (supplementary table 1): the 1-mer, 3-mer, 5-mer, 7-mer, 9-mer and A-CG models. The 1-mer model is the simplest model and does not consider the local sequence context of the base of interest. It has 12 parameters as each of the four bases has three possible substitutions. The 3-mer model considers one flanking nucleotide either side of the base of interest, making a 3-mer. It has 192 parameters as there are 64 possible 3-mers (4^3) and the middle base in each 3-mer has three possible substitutions. The 5-mer, 7-mer, and 9-mer models consider 2, 3, and 4 flanking nucleotides either side of the base of interest, respectively. The number of parameters increases exponentially as the number of flanking nucleotides considered expands, with the 9-mer model having 786,432 parameters as there are 262,144 possible 9-mers (4^9) and the middle base in each 9-mer has three possible substitutions. Substitution probabilities are largely driven by the presence of CG dinucleotides which have a high mutation rate from cytosine to thymine (M. Li and Chen

2011). We therefore expanded our simple 1-mer model to consider CG context, and named this the A-CG model. The A-CG model has 18 parameters as there are 6 possible contexts (A, T, C (non-CG), C (CG), G (non-CG), and G (CG)), each with three possible substitutions.

We inferred the mutational direction of each SNV (ie the ancestral and mutant states) using Ensembl's alignment of the *Mus musculus* (GRCm38) and *Mus caroli* reference genomes. For each SNV, the ancestral state was assumed to be the reference allele unless the alternate allele was shared with *Mus caroli*.

We calculated the substitution probability for each k-mer by dividing the counts of each k-mer change by the counts of each ancestral k-mer across the unmasked genome (supplementary tables 2 to 7). Each ancestral k-mer and k-mer change has a complement on the reverse strand. For example, the ancestral 3-mer TAG on the forward strand must have the complementary k-mer CTA on the reverse strand, and a TAG > TTG substitution must result in a complementary CTA > CAA substitution. We therefore totaled the counts of each forward and reverse k-mer and k-mer change to calculate substitution probabilities, effectively halving the number of parameters for each model, and report the substitution probabilities for adenine and cytosine only.

Methylation states and probabilities of substitution

Genome-wide methylation of CG dinucleotides in mouse embryonic stem cells was quantified by Stadler et al (2011) using bisulfite sequencing, and were downloaded from Ensembl (Stadler et al. 2011; Zerbino et al. 2018) (v101). Genome-wide methylation data is

not available for mouse germline cells. However, similar patterns of methylation have been observed between mouse embryonic stem cells and sperm cells (Popp et al. 2010), and we therefore used mouse embryonic stem cells as a proxy for the germline.

We classified all CG dinucleotides across the unmasked mouse genome as methylated if they have a coverage greater than or equal to 5 and a percentage methylated greater than or equal to 60, and unmethylated if they have a coverage greater than or equal to 5 and a percentage methylated less than or equal to 20. We excluded CG dinucleotides with coverage less than 5, or a percentage methylated more than 20 and less than 60. The distribution of methylation percentages is shown in supplementary figure 1. In total we classified 72.9% of CG dinucleotides as methylated, 6.9% as unmethylated, and excluded 20.2% from our analysis.

We calculated the probability for cytosine substitutions across three CG dinucleotide groups: all CG dinucleotides, unmethylated CG dinucleotides, and methylated CG dinucleotides. For each unmasked CG dinucleotide we determined the substitution class by the alternate allele (ie C>A, C>G, C>T, G>A, G>C, G>T). We calculated probabilities of substitution for each CG group by dividing the counts of each substitution class in each CG group, by the number of CGs in each CG group. We then combined counts between complementary strands to leave three substitution probabilities: C>A, C>G, and C>T.

Testing the prediction accuracy of substitution models

We used resampling by chromosome to test the prediction accuracy of local sequence context models. For each iteration, we split the data into training and testing datasets. Four autosomes

487 were randomly selected to be the testing data, and the remaining fifteen autosomes were used
 488 to train the models. Due to differences in chromosome size, this meant that the testing data
 489 could range between 28.0% and 14.1% of the total dataset. Probabilities of substitution were
 490 calculated for each k-mer model using the training data, following the methodology described
 491 above. To minimise bias introduced by the composition of test chromosomes (ie some kmers
 492 appearing at a much higher frequency than others) we gave each 9-mer an equal weighting in
 493 calculating substitution occurrence. This means the observed number of substitutions in the
 494 test data were calculated by totaling the observed substitution rates for each 9-mer. The
 495 absolute error was then calculated for each local sequence context model and substitution
 496 class (ie A>C, A>G, A>T, C>A, C>G, C>T) as the sum of the absolute difference between
 497 the observed (testing data) and predicted (training data) number of each substitution class for
 498 each 9-mer. Resampling was conducted 100 times and the mean absolute error and 95%
 499 confidence interval for each k-mer model and substitution class were calculated across the
 500 iterations.

501

502 We tested the accuracy to predict CG dinucleotide substitutions given different methylation
 503 states using resampling by chromosome, with four chromosomes randomly sampled for the
 504 test dataset in each iteration. Probabilities of substitution were calculated for each CG
 505 dinucleotide group in the training dataset following the methodology described above.
 506 Probabilities of substitution for methylated and unmethylated CG dinucleotides were
 507 calculated with the test dataset. The prediction accuracy of each CG group was determined as
 508 the difference between the observed (testing data) and predicted (training data) fraction for
 509 each substitution class.

510

511 Power testing

512

513 The number of potential k-mers (ie the potential combinations of A, C, G, and T) increases
 514 exponentially with k-mer length (supplementary table 1). This reduces the average count for
 515 each k-mer and k-mer substitution in the sample (supplementary table 8), and therefore the
 516 statistical power to accurately determine probabilities of substitution. We therefore tested
 517 whether we have the power to detect improvements in predictive accuracy between
 518 chromosomes, given the observed number of polymorphic sites and expanding k-mer sizes.
 519 To do this we created two simulated datasets of substitution counts: one where the true
 520 probability of substitution for each k-mer is defined by its 3-mer context, and the other where
 521 it is defined by its 9-mer context. We then tested prediction accuracy for the different k-mer
 522 models using these simulated datasets to determine whether we would detect an improvement
 523 in prediction accuracy.

524

525 In the simulated data, the number of substitutions for each 9-mer in each chromosome was
 526 determined from a binomial distribution, where the number of trials was equal to the number
 527 of each 9-mer, and the probability of success was defined by the observed substitution rates
 528 given the 3-mer or 9-mer context. For example, 3-mer and 9-mer probabilities of substitution
 529 were calculated from the observed data using all chromosomes, and then these probabilities
 530 were used to define the probabilities of success for each trial within chromosomes. Given
 531 these simulated datasets, prediction accuracy was tested as mean absolute error calculated
 532 from train test resampling between chromosomes as described above.

533

534 Mutational effect and probability of occurrence

535

536 We consider the coding consequences for all possible point mutations across all possible
 537 codons. In total there are 64 codons and 576 possible point mutations. Of these, 23 cause
 538 nonsense mutations, 415 cause missense mutations, and 138 cause synonymous mutations.
 539 Each codon point mutation has 256 possible 7-mers and 7-mer substitution rates
 540 (supplementary figure 3). We tested the difference in mutation rates between 7-mer
 541 substitutions with the potential to cause nonsense, missense, and synonymous mutations with
 542 Wilcoxon signed-rank tests. We tested the relationship between the number of potential
 543 nonsense, missense, and synonymous mutations as a function of 7-mer substitution rate using
 544 poisson regression. Human 7-mer substitution rates were downloaded from the
 545 supplementary information of Aggarwala and Voight (2016). We calculated the Spearman
 546 correlation between human and mouse 7-mer substitution rates in three groups: 7-mers
 547 substitutions with the potential to cause nonsense mutations, 7-mers substitutions with the
 548 potential to cause missense mutations, and 7-mers substitutions with the potential to cause
 549 synonymous mutations.

550

551 References

- 552 Aggarwala, Varun, and Benjamin F. Voight. 2016. “An Expanded Sequence Context Model
 553 Broadly Explains Variability in Polymorphism Levels across the Human Genome.”
 554 *Nature Genetics* 48 (4): 349–55.
- 555 Aikens, Rachael C., Kelsey E. Johnson, and Benjamin F. Voight. 2019. “Signals of Variation
 556 in Human Mutation Rate at Multiple Levels of Sequence Context.” *Molecular Biology*

557 *and Evolution* 36 (5): 955–65.

558 Brown, Steve D. M., Chris C. Holmes, Ann-Marie Mallon, Terrence F. Meehan, Damian

559 Smedley, and Sara Wells. 2018. “High-Throughput Mouse Phenomics for Characterizing

560 Mammalian Gene Function.” *Nature Reviews. Genetics* 19 (6): 357–70.

561 Duncan, B. K., and J. H. Miller. 1980. “Mutagenic Deamination of Cytosine Residues in

562 DNA.” *Nature* 287 (5782): 560–61.

563 Gonzalez-Perez, Abel, Radhakrishnan Sabarinathan, and Nuria Lopez-Bigas. 2019. “Local

564 Determinants of the Mutational Landscape of the Human Genome.” *Cell* 177 (1):

565 101–14.

566 Hardison, Ross C., Krishna M. Roskin, Shan Yang, Mark Diekhans, W. James Kent, Ryan

567 Weber, Laura Elnitski, et al. 2003. “Covariation in Frequencies of Substitution, Deletion,

568 Transposition, and Recombination during Eutherian Evolution.” *Genome Research* 13

569 (1): 13–26.

570 Harr, Bettina, Emre Karakoc, Rafik Neme, Meike Teschke, Christine Pfeifle, Željka Pezer,

571 Hiba Babiker, et al. 2016. “Genomic Resources for Wild Populations of the House

572 Mouse, *Mus Musculus* and Its Close Relative *Mus Spretus*.” *Scientific Data*.

573 Harris, Kelley. 2015. “Evidence for Recent, Population-Specific Evolution of the Human

574 Mutation Rate.” *Proceedings of the National Academy of Sciences of the United States*

575 *of America* 112 (11): 3439–44.

576 Harris, Kelley, and Jonathan K. Pritchard. 2017. “Rapid Evolution of the Human Mutation

577 Spectrum.” *eLife* 6 (April). <https://doi.org/10.7554/eLife.24284>.

578 Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix

579 Kokocinski, Bronwen L. Aken, et al. 2012. “GENCODE: The Reference Human

580 Genome Annotation for The ENCODE Project.” *Genome Research* 22 (9): 1760–74.

581 Helleday, Thomas, Eva Petermann, Cecilia Lundin, Ben Hodgson, and Ricky A. Sharma.

582 2008. “DNA Repair Pathways as Targets for Cancer Therapy.” *Nature Reviews. Cancer*
583 8 (3): 193–204.

584 Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel
585 Pignatelli, Albert J. Vilella, et al. 2016. “Ensembl Comparative Genomics Resources.”
586 *Database: The Journal of Biological Databases and Curation* 2016 (February).
587 <https://doi.org/10.1093/database/bav096>.

588 Iulio, Julia di, Istvan Bartha, Emily H. M. Wong, Hung-Chun Yu, Victor Lavrenko,
589 Dongchan Yang, Inkyung Jung, et al. 2018. “The Human Noncoding Genome Defined
590 by Genetic Diversity.” *Nature Genetics*, February.
591 <https://doi.org/10.1038/s41588-018-0062-7>.

592 Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica
593 Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint
594 Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

595 Knight, R. D., S. J. Freeland, and L. F. Landweber. 2001. “Rewiring the Keyboard:
596 Evolvability of the Genetic Code.” *Nature Reviews. Genetics* 2 (1): 49–58.

597 Kumar, Sudhir, and Sankar Subramanian. 2002. “Mutation Rates in Mammalian Genomes.”
598 *Proceedings of the National Academy of Sciences of the United States of America* 99 (2):
599 803–8.

600 Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth,
601 Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing
602 Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.”
603 *Bioinformatics* 25 (16): 2078–79.

604 Li, Mingkun, and Su-Shing Chen. 2011. “The Tendency to Recreate Ancestral CG
605 Dinucleotides in the Human Genome.” *BMC Evolutionary Biology* 11 (January): 3.

606 Lindahl, T., and R. D. Wood. 1999. “Quality Control by DNA Repair.” *Science* 286 (5446):

1897–1905.

Lindsay, Sarah J., Raheleh Rahbari, Joanna Kaplanis, Thomas Keane, and Matthew E. Hurles. 2019. “Similarities and Differences in Patterns of Germline Mutation between Mice and Humans.” *Nature Communications* 10 (1): 4053.

Lynch, Michael. 2011. “The Lower Bound to the Evolution of Mutation Rates.” *Genome Biology and Evolution* 3 (August): 1107–18.

Makova, Kateryna D., and Ross C. Hardison. 2015. “The Effects of Chromatin Organization on Variation in Mutation Rates in the Genome.” *Nature Reviews. Genetics* 16 (4): 213–23.

Petruska, J., and M. F. Goodman. 1985. “Influence of Neighboring Bases on DNA Polymerase Insertion and Proofreading Fidelity.” *The Journal of Biological Chemistry* 260 (12): 7533–39.

Popp, Christian, Wendy Dean, Suhua Feng, Shawn J. Cokus, Simon Andrews, Matteo Pellegrini, Steven E. Jacobsen, and Wolf Reik. 2010. “Genome-Wide Erasure of DNA Methylation in Mouse Primordial Germ Cells Is Affected by AID Deficiency.” *Nature* 463 (7284): 1101–5.

R Core Team. 2017. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Samocha, Kaitlin E., Elise B. Robinson, Stephan J. Sanders, Christine Stevens, Aniko Sabo, Lauren M. McGrath, Jack A. Kosmicki, et al. 2014. “A Framework for the Interpretation of de Novo Mutation in Human Disease.” *Nature Genetics* 46 (9): 944–50.

Schroeder, Jeremy W., William G. Hirst, Gabriella A. Szewczyk, and Lyle A. Simmons. 2016. “The Effect of Local Sequence Context on Mutational Bias of Genes Encoded on the Leading and Lagging Strands.” *Current Biology: CB* 26 (5): 692–97.

Ségurel, Laure, Minyoung J. Wyman, and Molly Przeworski. 2014. “Determinants of

Mutation Rate Variation in the Human Germline.” *Annual Review of Genomics and Human Genetics* 15 (June): 47–70.

Shen, J. C., W. M. Rideout 3rd, and P. A. Jones. 1994. “The Rate of Hydrolytic Deamination of 5-Methylcytosine in Double-Stranded DNA.” *Nucleic Acids Research* 22 (6): 972–76.

Stadler, Michael B., Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, et al. 2011. “DNA-Binding Factors Shape the Mouse Methylome at Distal Regulatory Regions.” *Nature* 480 (7378): 490–95.

Sung, Way, Matthew S. Ackerman, Samuel F. Miller, Thomas G. Doak, and Michael Lynch. 2012. “Drift-Barrier Hypothesis and Mutation-Rate Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (45): 18488–92.

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46 (D1): D754–61.

Zerbino, Daniel R., Steven P. Wilder, Nathan Johnson, Thomas Juettemann, and Paul R. Flicek. 2015. “The Ensembl Regulatory Build.” *Genome Biology* 16 (March): 56.

Author contributions

GP conceived of the idea, conducted the analyses, and wrote the manuscript. HSL, TMMV, AM, CML, and MMS contributed to the manuscript. AM, CML, and MMS helped supervise the project.