Hypermutable Non-Synonymous Sites are Under Stronger Negative Selection

Citation

Published Version
doi:10.1371/journal.pgen.1000281

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:4875873

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Hypermutable Non-Synonymous Sites Are under Stronger Negative Selection

Steffen Schmidt1,2*, Anna Gerasimova3,4, Fyodor A. Kondrashov5, Ivan A. Adzuhbei1, Alexey S. Kondrashov3,4, Shamil Sunyaev1*

1 Division of Genetics, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, United States of America, 2 Department of Biochemistry, Max Planck Institute for Developmental Biology, Tübingen, Germany, 3 Life Sciences Institute, University of Michigan, Ann Arbor, Michigan, United States of America, 4 Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America, 5 Section on Ecology, Behavior, and Evolution, Division of Biological Sciences, University of California San Diego, La Jolla, California, United States of America

Abstract

Mutation rate varies greatly between nucleotide sites of the human genome and depends both on the global genomic location and the local sequence context of a site. In particular, CpG context elevates the mutation rate by an order of magnitude. Mutations also vary widely in their effect on the molecular function, phenotype, and fitness. Independence of the probability of occurrence of a new mutation’s effect has been a fundamental premise in genetics. However, highly mutable contexts may be preserved by negative selection at important sites but destroyed by mutation at sites under no selection. Thus, there may be a positive correlation between the rate of mutations at a nucleotide site and the magnitude of that site’s effect on fitness. We studied the impact of CpG context on the rate of human–chimpanzee divergence and on intrahuman nucleotide diversity at non-synonymous coding sites. We compared nucleotides that occupy identical positions within codons of identical amino acids and only differ by being within versus outside CpG context. Nucleotides within CpG context are under a stronger negative selection, as revealed by their lower, proportionally to the mutation rate, rate of evolution and nucleotide diversity. In particular, the probability of fixation of a non-synonymous transition at a CpG site is two times lower than at a CpG site. Thus, sites with different mutation rates are not necessarily selectively equivalent. This suggests that the mutation rate may complement sequence conservation as a characteristic predictive of functional importance of nucleotide sites.

Introduction

The functional and phenotypic effects of mutations and, consequently, the strength of negative selection vary widely among nucleotide sites in any genome. At the opposite ends of the continuum, mutations at some sites are effectively neutral, while mutations at some other sites are lethal. Nucleotide sites can be subdivided, according to their molecular function, into classes with different typical strengths of negative selection. Generally, rapidly evolving segments of intergenic regions and introns, as well as most of synonymous coding sites, are controlled by only weak selection or even by no selection at all. Slowly evolving segments of intergenic regions and introns, as well as UTRs and non-synonymous coding sites, are under much stronger selection (e.g., [1–8]). However, even within such functional classes, the strength of negative selection varies widely among individual sites (e.g., 9–12).

The rate of spontaneous mutation is also not uniform across individual sites [13–15]. The standard deviation of the mutation rate at a site may be comparable to its mean. Moreover, some rare hot-spot sites may mutate much more frequently than an average site. Thus, the mutation rate at a site depends both on its local sequence context (e.g., [16–19]) and on its global location within the genome [13–15], although these dependencies are rather different in different groups of organisms [19,20]. In particular, in mammals the 5’CpG3’ context substantially increases the rate of transversions, and especially transitions [16–19,21].

Mutation and selection are generally thought to be independent evolutionary forces [22]. In other words, the rate with which a mutation occurs is routinely assumed to be independent of the effect of this mutation on fitness. Inferences of the strength of selection on specific genes and sites within genes usually rely on this assumption. Although selection for reduced mutability is stronger at sites where mutations are more deleterious [23], it is hard to imagine adaptive fine-tuning of mutation rates at the level of individual nucleotide sites. Thus, one might expect selective constraint and mutability to vary more or less independently across individual sites.

However, another phenomenon may lead to a seemingly counterintuitive association between stronger negative selection and higher mutation rates. Sites that are under weak or no selection are free to evolve and to get rid of hypermutable contexts. In contrast, negative selection will preserve such contexts at functionally important sites, provided that they confer a higher
**Author Summary**

Mutations occur in some sites in the genome more frequently than in others. Similarly, mutations in some sites have greater consequences than in others. The effect of mutations might not be independent of the frequency with which mutations occur. Indeed, sites where mutations happen frequently will be preserved if the effects of these mutations are severe or will otherwise be allowed to mutate if there are no consequences for the organism. We compared both human–chimpanzee differences and sequence variation among humans in protein coding genes. We found that highly mutable nucleotide sites, such as the dinucleotide CpG, are on average more important and more frequently preserved by natural selection. Using this information, together with other features such as sequence conservation, opens a new perspective to predict the effect of human mutations, including their potential involvement in diseases.

**Results**

If identical nucleotides at identical sites within codons of identical amino acids are under the same selection, regardless on whether they are located within or outside CpG context, then this context would equally impact the mutation rate, the rate of divergence between species, and the level of intraspecies nucleotide diversity. If, however, negative selection is stronger within CpG context, this context would elevate the level of nucleotide diversity and especially the rate of divergence, to a lesser extent than the mutation rate.

**Impact of CpG Context on Mutation Rates**

It is well known that in mammals CpG context substantially increases the mutation rate; however, the exact magnitude of this effect has not been established with certainty. We used three sources of information on the impact of CpG context on the rates of transitions and transversions: 1) direct data on Mendelian diseases in humans [18], 2) Bayesian Markov Chain Monte Carlo analysis of evolution of several species of mammals [19], and 3) parsimony-based analysis of human-chimpanzee-orangutan genome alignments (Table 1). The third analysis must underestimate the impact of CpG context on transversion and especially transition rates, because two nucleotide substitutions, one on the edge leading to a sister species (human or chimpanzee) and the other on the edge leading to the outgroup (orangutan), can happen within a CpG context. Such occurrences will lead to underestimation of the fraction of sites that were within CpG context in the common ancestor of human and chimpanzee and, thus, of the fraction of allele substitutions that destroy a CpG context. Indeed, this underestimation is evident from Table 1. Thus, below we will use the mean values of the first two estimates and will assume that in humans CpG context increases the rate of transitions by the factor of 14.5, and the rate of transversions by the factor of 3.5.

**Impact of CpG Context on the Rate of Evolution and Intraspecies Diversity at Non-Synonymous Sites**

We used human-chimpanzee-orangutan alignments of coding sequences to compare the rates of a particular nucleotide substitution that causes a particular amino acid replacement within vs. outside CpG context (CpG vs. ~CpG). For example, a P→L replacement, caused by a C→T transition, can occur within (CGG→CTG; the site of substitution is boldfaced) or outside (e.g., CCC→CTC) CpG context. The common ancestor of humans and chimpanzees, as revealed by the orangutan outgroup, carried, at all the loci we studied, TargetP→L CpG = 18,088 of CGG codons, and TargetP→L ~CpG = 185,826 of CCA, CCT, or CCC codons (Table 2). There were 215 and 284 P→L replacements (ReplacementsP→L CpG and ReplacementsP→L ~CpG), caused by C→T transitions, within CpG and outside CpG contexts, respectively. Thus the impact of CpG context on the rate of P→L replacements in the course of human-chimpanzee divergence is

\[
CpG_{\text{impact(P→L)}} = \frac{\text{ReplacementsP→L CpG}}{\text{TargetP→L CpG}} / \frac{\text{ReplacementsP→L ~CpG}}{\text{TargetP→L ~CpG}},
\]

\[= 7.78.\]

This analysis relies on the identification of the human-chimpanzee ancestral state using orangutan as outgroup. To test whether possible erroneous identifications affect our results, we repeated the same analysis using the macaque outgroup, which must lead to more errors, because macaque is about three times more distant from the human-chimpanzee last common ancestor than orangutan. Also, all the analyses were performed only for human and chimpanzee coding sequences, under the assumption

<table>
<thead>
<tr>
<th>Table 1. Estimates of the impact of CpG context on the mutation rates of transitions and transversions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Transition Cpg</td>
</tr>
<tr>
<td>Transition ~Cpg</td>
</tr>
<tr>
<td>Transversion Cpg</td>
</tr>
<tr>
<td>Transversion ~Cpg</td>
</tr>
</tbody>
</table>

The last column contains ratios computed using a (human-chimp)-orangutan alignment. doi:10.1371/journal.pgen.1000281.0001
<table>
<thead>
<tr>
<th>Divergence</th>
<th>Macaque as outgroup</th>
<th>Chimpanzee as outgroup</th>
<th>Orangutan as outgroup</th>
<th>Divergence</th>
<th>Macaque as outgroup</th>
<th>Chimpanzee as outgroup</th>
<th>Orangutan as outgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-L</td>
<td>368</td>
<td>481</td>
<td>11,714</td>
<td>12,877</td>
<td>8.41</td>
<td>318</td>
<td>3,353</td>
</tr>
<tr>
<td>A-T</td>
<td>197</td>
<td>217</td>
<td>11,451</td>
<td>12,423</td>
<td>5.60</td>
<td>176</td>
<td>1,544</td>
</tr>
<tr>
<td>G-C</td>
<td>480</td>
<td>582</td>
<td>26,206</td>
<td>31,879</td>
<td>6.20</td>
<td>345</td>
<td>4,082</td>
</tr>
<tr>
<td>A-G</td>
<td>177</td>
<td>238</td>
<td>14,783</td>
<td>18,403</td>
<td>5.05</td>
<td>150</td>
<td>1,544</td>
</tr>
<tr>
<td>A-C</td>
<td>124</td>
<td>205</td>
<td>12,561</td>
<td>13,213</td>
<td>6.82</td>
<td>96</td>
<td>1,013</td>
</tr>
<tr>
<td>A-T</td>
<td>119</td>
<td>236</td>
<td>16,236</td>
<td>19,121</td>
<td>5.94</td>
<td>119</td>
<td>8,722</td>
</tr>
<tr>
<td>E-K</td>
<td>190</td>
<td>307</td>
<td>21,984</td>
<td>28,322</td>
<td>6.27</td>
<td>106</td>
<td>9,129</td>
</tr>
<tr>
<td>S-L</td>
<td>162</td>
<td>79</td>
<td>14,345</td>
<td>14,950</td>
<td>8.14</td>
<td>122</td>
<td>2,969</td>
</tr>
<tr>
<td>D-H</td>
<td>235</td>
<td>401</td>
<td>20,988</td>
<td>29,606</td>
<td>8.41</td>
<td>215</td>
<td>2,244</td>
</tr>
<tr>
<td>R-I</td>
<td>368</td>
<td>368</td>
<td>11,714</td>
<td>12,877</td>
<td>8.41</td>
<td>318</td>
<td>3,353</td>
</tr>
<tr>
<td>M</td>
<td>197</td>
<td>217</td>
<td>11,451</td>
<td>12,423</td>
<td>5.60</td>
<td>176</td>
<td>1,544</td>
</tr>
<tr>
<td>T</td>
<td>352</td>
<td>582</td>
<td>26,206</td>
<td>31,879</td>
<td>6.20</td>
<td>345</td>
<td>4,082</td>
</tr>
<tr>
<td>S</td>
<td>177</td>
<td>238</td>
<td>14,783</td>
<td>18,403</td>
<td>5.05</td>
<td>150</td>
<td>1,544</td>
</tr>
<tr>
<td>G</td>
<td>124</td>
<td>205</td>
<td>12,561</td>
<td>13,213</td>
<td>6.82</td>
<td>96</td>
<td>1,013</td>
</tr>
<tr>
<td>C</td>
<td>119</td>
<td>236</td>
<td>16,236</td>
<td>19,121</td>
<td>5.94</td>
<td>119</td>
<td>8,722</td>
</tr>
<tr>
<td>E</td>
<td>190</td>
<td>307</td>
<td>21,984</td>
<td>28,322</td>
<td>6.27</td>
<td>106</td>
<td>9,129</td>
</tr>
<tr>
<td>S</td>
<td>162</td>
<td>79</td>
<td>14,345</td>
<td>14,950</td>
<td>8.14</td>
<td>122</td>
<td>2,969</td>
</tr>
<tr>
<td>D</td>
<td>235</td>
<td>401</td>
<td>20,988</td>
<td>29,606</td>
<td>8.41</td>
<td>215</td>
<td>2,244</td>
</tr>
</tbody>
</table>

Table 2. Non-synonymous substitutions in human-chimp divergence and human polymorphism data.

The nsSNP data shown here is based on the Applera data set using human-chimp alignments to determine the direction of the mutation.
that the proportion of CpG context within these sequences is at equilibrium. Estimates of the impact of CpG context on the rates of evolution obtained in this way were only slightly higher than estimates obtained using the orangutan outgroup (data not reported).

For intraspecies nucleotide diversity, the number of SNPs that involve a particular amino acid change within and outside CpG context were used in equation (1), instead of the corresponding numbers of substitutions (Table 2). The direction of an amino acid change associated with a particular SNP was determined by the orthologous chimpanzee sequence. We assumed that the ratio of CpG vs. non-CpG target sizes for a particular amino acid replacement was the same as for human-chimpanzee divergence. Indeed, the SNPs we used were obtained by resequencing of ~11,000 human loci [27] so that we can expect the nucleotide composition of this sample to be close to that of all protein-coding loci. The data on the impacts of CpG context on human-chimpanzee divergence and on intrahuman diversity are shown in Table 2 and in Figure 1. Thus, the impact of CpG context on the rate of divergence, i.e. the average ratio of the rates of divergence within vs. outside CpG contexts, was 7.1 for transitions and 2.5 for transversions. The average ratio of values of intrahuman diversities for non-synonymous SNPs within vs. outside CpG contexts was 11.2 for transitions and 2.4 for transversions (Table 3). If macaque instead of orangutan is used as an outgroup, the observed impacts of CpG context on the rates of divergence decline only slightly (6.8 instead of 7.1 for transitions, and 2.1 instead of 2.5 for transversions).

We applied several tests to evaluate the significance of the difference of the impact of CpG context on non-synonymous divergence and diversity. This difference is insignificant for transversions and highly significant for transitions, according to the $\chi^2$ test ($p=2.8 \times 10^{-16}$). However, the $\chi^2$ test does not stratify data according to amino acid replacements, which is essential in our case. We used two approaches to perform stratified analysis of contingency tables. First, we combined $p$-values of separate tests for each amino acid replacement, using Stouffer ($p<2.2 \times 10^{-16}$) and Fisher ($p=2.7 \times 10^{-16}$) methods. We also applied Cochran-Mantel-Haenszel test, the standard test for stratified analysis of contingency tables ($p=4.6 \times 10^{-16}$).

**Impacts of CpG Context at Synonymous and Non-Coding Sites**

We measured the impacts of CpG context on rates of evolution and nucleotide diversity at synonymous coding and at non-coding sites (Table 3). As it was the case for non-synonymous sites, we assumed parsimony. Thus, the data on rates of evolution at non-coding sites shown in Table 3 are taken from (human-chimpanzee)-orangutan comparison shown in Table 1.

We can see that the impacts of CpG context on non-coding human-chimpanzee divergence and intrahuman nucleotide diversity are rather close to the corresponding impacts on the mutation rate, which is consistent with effective neutrality of most of the non-coding DNA in humans. The figures in Table 3 are likely to be slightly underestimated, due to substitutions in the outgroup lineage.

In contrast to non-coding sites, at synonymous sites the impacts of CpG context on human-chimpanzee divergence and intrahuman nucleotide diversity due to transitions, but not to transversions, are substantially lower than the corresponding impacts on the mutation rates, although still higher than the corresponding impacts at non-synonymous sites. This implies that some selection acts on synonymous transitions within CpG context, and that this selection is weaker than the corresponding selection at non-synonymous sites. Several analyses revealed weak selection favoring Cs and Gs at synonymous sites [25,20].

**Discussion**

Our results show that negative selection is stronger within CpG contexts than in less mutable sites at identical codon positions. We can see that the per nucleotide site rate of transitions, accepted in the course of human-chimpanzee divergence, is on average 7.1 times higher within CpG contexts than outside CpG contexts (Table 3). A comparison of this figure with the impact of CpG on the corresponding mutation rate (Table 1) suggest that a transition that occurred within CpG context gets fixed in the course of human-chimpanzee divergence with a probability of $7.1/14.5 = 0.49$ of the probability of fixation of a transition that occurred outside CpG context. Thus, nucleotides within CpG context are protected by a stronger selection.

In the case of SNPs, we observed a similar but weaker effect. On average, non-synonymous SNPs caused by transitions are 11.2 times more common within CpG context than outside of it. Thus, a non-synonymous transition mutation that occurred within CpG context is observed as a SNP with a chance that constitutes only $11.2/14.5 = 0.77$ of the chance of observing a transition that caused the same amino acid replacement but occurred outside CpG context.

In other words, in the case of transitions, CpG context increases the level of intrahuman diversity and in particular the rate of non-synonymous divergence less than proportionally to its impact on the mutation rate. This demonstrates that negative selection at non-synonymous sites within CpG context is stronger than at sites outside it. This seemingly counterintuitive pattern probably has a simple evolutionary explanation: nucleotide sites that are not under strong negative selection will eventually lose most of their hypermutable

---

**Table 3. Average impacts of CpG context for different types of sites using orangutan as outgroup.**

<table>
<thead>
<tr>
<th>CpG Impact</th>
<th>Type</th>
<th>Divergence</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-synonymous</td>
<td>transition</td>
<td>7.1</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>transversion</td>
<td>2.5</td>
<td>2.4</td>
</tr>
<tr>
<td>synonymous</td>
<td>transition</td>
<td>8.6</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td>transversion</td>
<td>2.1</td>
<td>2.3</td>
</tr>
<tr>
<td>non-coding</td>
<td>transition</td>
<td>12.2</td>
<td>13.7</td>
</tr>
<tr>
<td></td>
<td>transversion</td>
<td>2.4</td>
<td>2.0</td>
</tr>
</tbody>
</table>

---

**Figure 1. CpG impact on transitions in amino acid changes.** The effect on human-chimpanzee divergence is shown in blue; the effect on non-synonymous SNPs in human in red. doi:10.1371/journal.pgen.1000281.g001
CpG contexts. Thus, hypermutable contexts must be disproportionally common at sites under strong negative selection.

It is not surprising that a stronger negative selection within CpG contexts affects the rates of evolution more than it affects intraspecies diversity. Indeed, a substantial fraction of SNPs that segregate within a population are nevertheless subject to negative selection that is strong enough to prevent their fixation [22]. The large difference between the impacts of CpG context on polymorphism and divergence suggests that the observed effect is mostly due to nucleotide sites under weak selection, which affects divergence more than polymorphism. Such sites are abundant in human protein coding genes [9–11,29].

Predictably, the impacts of CpG context at mostly selectively neutral noncoding sites do not differ substantially from its impacts on the mutation rate. In contrast, coding synonymous sites within CpG contexts evolve slower and are less diverse within humans than what would be expected on the basis of the mutation rates alone. This is not surprising because the impact of CpG context must be sensitive to even weak selection [25,28]. Indeed, CpG contexts are greatly underrepresented at purely neutral sites, but even a rather weak selection is expected to increase their prevalence substantially, as long as the coefficient of selection is of the order of the reciprocal of the effective population size or higher [22]. CpG contexts are much more common within synonymous sites than within non-coding sites [25].

CpG context exerts a much weaker influence on the rate of transversions than on the rate of transitions (see Table 1). Thus, it is not surprising that the effects, which we can easily observe in the case of transitions, are not visible in the case of transversions. More data are needed to determine if these effects, however weak, are still present in the case of transversions.

Our estimates of the impact of CpG context on divergence (Tables 2 and 3) are probably too low due to substitutions in the outgroup lineage. However, these estimates depend only slightly on whether orangutan or macaque is used as an outgroup, although in the second case the prevalence of multiple substitutions at a site should be much higher. Also, the estimates computed from only human and chimpanzee genomes assuming equilibrium of the CpG context are only slightly higher than the estimate obtained using an outgroup. Further, the estimate of the impact of CpG context on human-chimpanzee divergence due to transitions at non-synonymous sites is much lower than the corresponding estimate for non-coding sites computed using the same outgroup (Table 3). This indicates that the low impact of CpG contexts not just an artifact of the assumption of parsimony. Even under the impossible assumption that every site that is located within CpG context in either human or chimpanzee sequence was also located within CpG context in their last common ancestor, the resulting estimate of the impact of this context on the rate of divergence equals 12 and is still lower than CpG impact on raw mutation rate.

The analysis of intrahuman diversity relies on the chimpanzee sequence for determining the identity of ancestral alleles. Misidentification of ancestral alleles would result in an underestimation of the impact of CpG context because ancestral CpGs would preferentially evolve in the chimpanzee lineage. To evaluate a possible extent of this bias we repeated the analysis using major and minor alleles instead of inferred ancestral and derived alleles. The resulting estimate of the impact of CpG context on non-synonymous transitions is 11.5, which is only slightly higher than 11.2 (Table 2).

Negative selection can also be detected in polymorphism data independently of intraspecies nucleotide diversity through changes in the distribution of allele frequencies, because such selection causes an excess of low-frequency alleles. In particular, minor allele frequencies of non-synonymous SNPs that affect slowly evolving (conserved) protein sites are reduced [30,31]. The excess of rare alleles was not statistically significant in the two datasets of human SNPs used in this study. The effect of weak negative selection on allele frequency distribution is expected to be much smaller than on divergence and data on rare SNPs in protein coding regions are sparse. Thus, the analysis of allele frequency distribution may lack statistical power.

Our analysis suggests that mutation rates can be used in computational methods to predict which amino acid replacements are deleterious [32]; a replacement that occurred at a highly mutable site is more likely to be deleterious. Currently, prediction methods rely on the properties of an encoded amino acid sequence, its conservation between species, and the properties of the corresponding protein. Our analysis suggests that taking the DNA-level features of an amino acid replacement into account will increase the accuracy of prediction of its effect on protein function.

Materials and Methods

To determine the impact of CpG context on mutation rates we constructed a human-chimpanzee-orangutan alignment for a ~1 Mb piece of orangutan genomic sequence (gi:119380173), and analyzed it assuming parsimony. To study the impact of CpG context on the rate of evolution, we constructed human-chimpanzee-orangutan and human-chimpanzee-macaque alignments of coding regions of individual genes by finding the orthologous macaque gene for each UCSC human-chimpanzee pair with the by-directional best BLAST hits approach [33]. We also repeated the analysis on just two sequences assuming equilibrium CpG content (data not shown). This analysis resulted in similar estimates.

For the analysis of intrahuman diversity we used a comprehensive and systematic Applera dataset [27]. Chimpanzee nucleotides corresponding to human SNP positions were identified using the SNP UCSC genome track [34]. Applera set is gene centric. Therefore, for the analysis of non-coding diversity, we used randomly ascertained SNPs from the Perlegen set [35]. We also verified that coding SNPs from the Perlegen dataset produced estimates highly similar to those based on the Applera dataset. We analyzed each population separately and excluded SNPs, which were fixed in the population and could not be mapped to chimpanzee nucleotides (~4.6%).

Statistical analysis was carried out using R statistical package v2.7.0 [36]. p-Values for individual amino acid residue contingency tables were computed by Monte Carlo simulations with the number of replicates B = 10^6. To obtain combined p-values we used Stouffer's z-scores [37] and Fisher's sum of logs of p [38] methods. Cochran-Mantel-Haenszel test of conditional independence [39] was utilized to ensure there was no three-way interaction with the amino acid residue type.

Author Contributions

References