Large scale variation in the rate of *de novo* mutation in humans and its relationship to divergence and diversity.

Thomas Smith
Adam Eyre-Walker

School of Life Sciences
University of Sussex
Brighton
BN1 9QG

Correspondence : a.c.eyre-walker@sussex.ac.uk

Abstract

The rate of divergence between species, and the level diversity within a species, are expected to depend on the rate of mutation. Here we investigate the relationship between divergence, diversity and the rate of mutation across the human genome at scales of 1MB and 100KB using >40,000 *de novo* mutations (DNM). We show that there is significant variation in the rate of DNM across the human genome, but that the variation is modest. Different types of mutation show similar levels of variation and appear to vary in concert. Regressing the rate of DNM against a range of genomic features suggests that nucleosome occupancy is the most important correlate, but that GC content, recombination rate, replication time and various histone methylation signals also correlate significantly. In total the model explains ~75% of the explainable variance. As expected the rate of divergence between species and the level of diversity within humans is correlated to the rate of DNM. However, the correlations are weaker than if all the variation in divergence was due to variation in the mutation rate. We provide evidence that this is due the effect of biased gene conversion on the probability that a mutation will become fixed. Finally, we show that the correlation between divergence and DNM density declines as increasingly divergent species are considered. Our results have important implications for the use of divergence and diversity data to study variation in the mutation rate.

Introduction

Until recently, the distribution of germ-line mutations across the genome was largely studied using patterns of nucleotide substitution between species in putatively neutral sequences (see [1] for review of this literature), since under neutrality the rate of substitution should be equal to the mutation rate. However, the sequencing of hundreds of individuals and their parents has led to the discovery of thousands of *de novo* mutations (DNMs) in humans [2-6]; it is therefore possible to start analysing the pattern of DNMs directly rather than inferring their patterns from substitutions. Initial analyses have shown that the rate of DNM increases with paternal age [4], a result that was never-the-less inferred by Haldane some 70 years ago [7], varies across the genome [5] and is correlated to a number of factors, including the time of replication [3], the rate of recombination [3], GC content [5] and DNA hypersensitivity [5].

Here we use a collection of over 40,000 DNMs to address a range of questions pertaining to the large-scale distribution of DNMs. First, we investigate whether there is variation in the mutation rate at a large-scale that cannot be explained in terms of variation at smaller scales. We quantify this variation and investigate to what extent the variation is correlated between different types of mutation, and to what extent it is correlated to a range of genomic variables.

Second we investigate the relationship between variation in the divergence between species and the rate of DNM across the human genome. Although, divergence between species has been used to study the pattern of DNM, it has recently been shown that the rate of mutation, at a per MB scale, is not as strongly correlated to the rate of nucleotide substitution between species as it could be if all the variation in divergence between 1MB blocks was due to variation in the mutation rate [3]. Instead, the rate of divergence appears to correlate to the rate of recombination as well. This might be due to one, or a combination, of several factors. First, recombination might affect the probability that a mutation becomes fixed by the process of biased gene

conversion (BGC) (review by [8]). Second, recombination can affect the probability that a mutation will be fixed by natural selection; in regions of high recombination deleterious mutations are less likely to be fixed, whereas advantageous mutations are more likely. Third, low levels of recombination can increase the effects of genetic hitch-hiking and background selection, both of which can reduce the diversity in the human-chimp ancestor, and the time to coalescence and the divergence between species. And fourth, the correlation of divergence to both recombination and DNM density might simply be due to limitations in multiple regression; spurious associations can arise if multiple regression is performed on two correlated variables that are not known without error. For example, it might be that divergence only depends on the mutation rate, but that the mutation rate is partially dependent on the rate of recombination. In a multiple regression, divergence might come out as being correlated to both DNM density and the recombination rate, because we do not know the mutation rate without error, since we only have limited number of DNMs. It could even be the case that divergence only correlates to recombination rate, because it is a less error prone measure of the mutation rate than DNM density. Here, we introduce a test that can resolve between some of these explanations.

Third, we examine the relationship between diversity within humans and the large-scale variation in the mutation rate. The role of the mutation rate variation in determining the level of genetic diversity across the genome has long been a subject of debate. It was noted many years ago that diversity varies across the human genome at a large scale and that this variation is correlated to the rate of recombination [9-11]. Because the rate of substitution between species is also correlated to the rate of recombination, Hellmann et al. [9, 10] inferred that the correlation between diversity and recombination was at least in part due to a mutagenic effect of recombination. This is consistent with the results of Francioli et al. [3] who have recently shown that the rate of DNM is correlated to the rate of recombination. However, no investigation has recently been made as to whether this explains all the variation in diversity.

Results

De novo mutations

To investigate large scale patterns of *de novo* mutations in humans we compiled data from four studies which between them had discovered 43,433 DNMs: 26,939 mutations from Wong et al. [6], 11016 mutations from Francioli et al. [3], 4931 mutations from Kong et al. [4] and 547 mutations from Michaelson et al. [5]. We divided the mutations up into 9 categories reflecting the fact that CpG dinucleotides have higher mutation rates than non-CpG sites, and the fact that we cannot differentiate which strand the mutation had occurred on: CpG C->T (a C to T or G to A mutation at a CpG site), CpG C->A, CpG C->G and for non-CpG sites C->T, T->C, C->A, T->G, C<->G and T<->A mutations. The proportion of mutations in each category in each of the datasets is shown in figure 1.

We find that the pattern of mutation differs significantly between the 4 studies (Chi-square test of independence on the number of mutations in each of the 9 categories, $p < 0.0001$). This appears to be largely due to the relative frequency of C->T transitions in both the CpG and non-CpG context. In the data from Wong et al. [6] and Michaelson et al. [5] the frequency of C->T transitions at CpG sites is ~13% whereas it is ~17% in the other two studies, a discrepancy which has been noted before between the studies of Michaelson et al. and Kong et al. [12]. For non-CpG sites the frequency of C->T transitions is ~24% in all studies except that of Wong et al. in which it is 26%. It is not clear whether these patterns reflect differences in the mutation rate between different cohorts of individuals, possibly because of age [3, 4, 6] or race [13] or whether the differences are due to methodological problems associated with detecting DNMs. Since the differences are relatively small and it is not clear whether one study represents a more representative sample than the other, we combined the data.

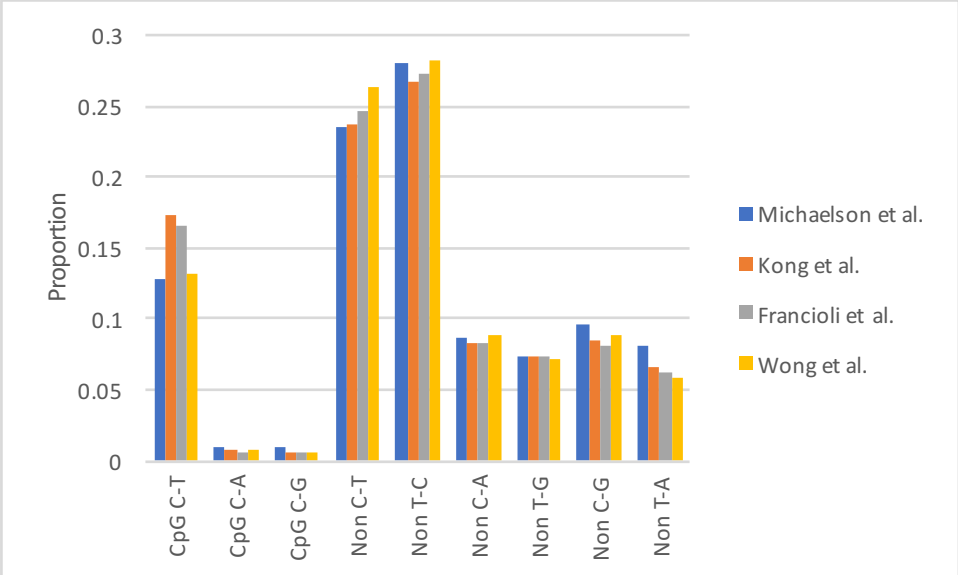


Figure 1. The proportion of DNMs in each of the mutational types in the four datasets.

Distribution of rates

To investigate whether there is large scale variation in the mutation rate we divided the genome into non-overlapping windows of 10KB, 100KB, 1MB and 10MB and fit a gamma distribution to the number of mutations per region, taking into account the sampling error associated with the low number of mutations per region. The coefficient of variation (CV) of the fitted gamma distributions are 0.41, 0.29, 0.21 and 0.17 for 10KB, 100KB, 1MB and 10MB respectively. If all the variation at the larger scales is explainable by variation at a smaller scale, then the CV at scale *x* should be equal to the CV at some finer scale, *y*, divided by the square-root of *x/y*; for example, the CV at the 100KB scale given the variation at the 10KB scale should be $0.41/\sqrt{10} =$ 0.13, which is considerably smaller than the observed CV, suggesting that there is more variation at the 100KB scale than expected. This demonstrates that there is large scale variation in the mutation rate. To characterise this further, we plotted CV against the scale on a log scale; the relationship is approximately linear but the slope (-0.13) is considerably less than the expected slope of -0.5 if all variation at larger scales was due to variation at smaller scales (Figure 2). For the rest of this analysis we concentrate on this

large-scale variation in the mutation rate and consider it at two scales of 100KB and 1MB.

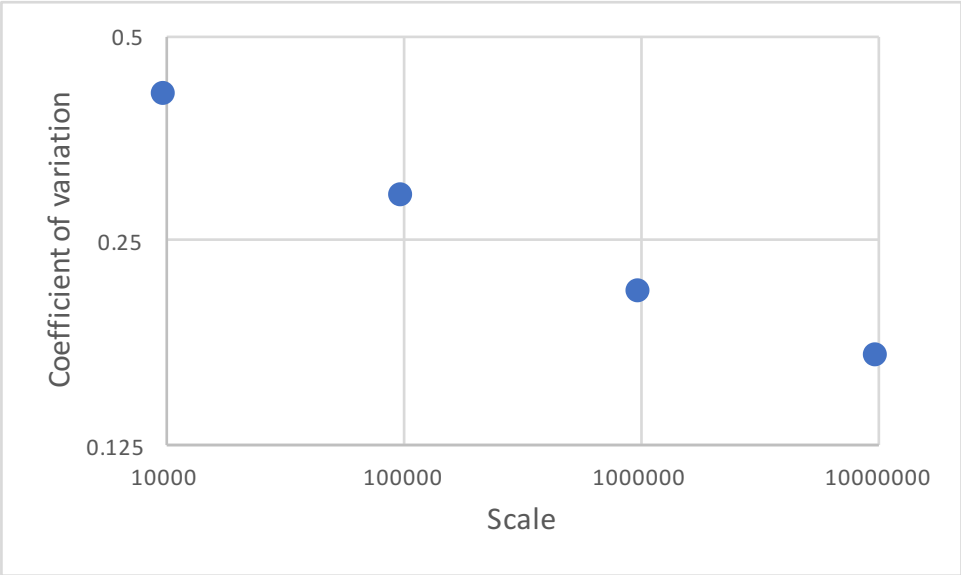


Figure 2. The coefficient of variation of the gamma distribution fitted to the number of DNMs per block, versus the size of the blocks.

The level of variation at both the 100KB and 1MB scales is significant (i.e. the lower 95% confidence interval of the CV is not zero) when all mutations are considered together (Table 1), however the level of variation is quite modest (Figure 3). A gamma distribution with a coefficient of variation of 0.21, as we find for the MB data, is a distribution in which 90% of regions have a rate of mutation that is within 35% the mean (i.e. rates within the range of 0.65 to 1.35); at the 100KB level, roughly 90% of regions have mutation rates that are within 47% the mean (Table 1) (Figure 3).

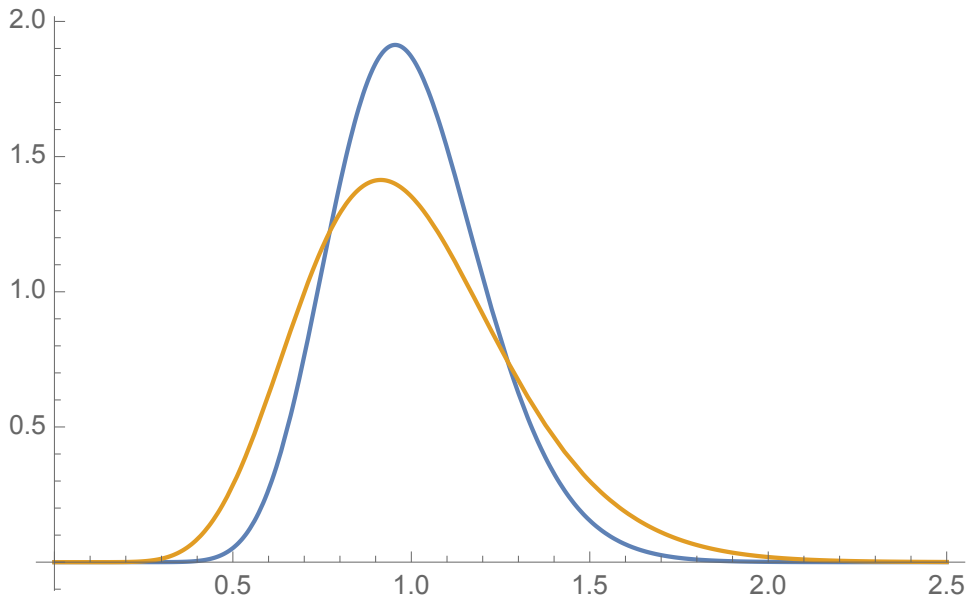


Figure 3. The distribution of the mutation rates, relative to the mean, inferred from the distribution of DNMs at the 100KB (blue line) and 1MB (orange line) scales.

We also find significant variation for CpG transitions and non-CpG transitions and transversions (Table 1). However, we do not find significant variation for either CpG transversions (the lower confidence interval for the coefficient of variation is zero), or when we split the data into most individual mutational types; this is probably because we have too little data.

Mutation type	100KB	1MB
All	0.29 (0.27, 0.32)	0.21 (0.20, 0.23)
CpG	0.41 (0.32, 0.49)	0.28 (0.23, 0.32)
nonCpG	0.30 (0.28, 0.32)	0.21 (0.19, 0.22)
CpG transitions	0.41 (0.30, 0.50)	0.27 (0.21, 0.32)
CpG transversions	0.24 (0.0, 0.27)	0.47 (0, 0.73)
nonCpG transitions	0.29 (0.22, 0.30)	0.19 (0.17, 0.21)
nonCpG transversions	0.34 (0.28, 0.39)	0.24 (0.20, 0.27)

Table 1. The coefficient of variation for a gamma distribution fitted to the density of DNMs, and the 95% confidence intervals of the coefficient of variation.

Given that there is variation in all mutational types, for which we have enough data, it is of interest to investigate whether the amount of variation differs between the mutational types. To investigate this, we ran a series of likelihood ratio tests in which fit separate and common distributions to the different mutational types. We found significantly more variation at non-CpG sites than CpG sites at both scales and more variation for non-CpG transversions than transitions at the 100KB scale (it is almost significant at the 1MB scale as well) ($p<0.05$) (Table S1). Never-the-less, although significant, the differences in terms of the coefficient of variation are quite modest (Table 1).

Correlations between mutational types

Given that there is variation in the mutation rate at the 1MB and 100KB levels and that this variation is quite similar for different mutational types, it would seem likely that the rate of mutation for the different mutational types are correlated. We find that this is indeed the case. At the 1MB scale we find significant correlations between the rates of CpG and non-CpG mutations ($r = 0.17$, $p < 0.001$), CpG transitions and transversions ($r = 0.050$, $p = 0.012$), and non-CpG transitions and transversions ($r = 0.25$, $p<0.001$). In all cases these

correlations are about as strong as you would expect given the high level of sampling error; i.e. if we simulate data using a gamma distribution fit to both mutational categories, we find the mean correlation from 100 simulations are 0.18, 0.036 and 0.20 for three comparisons respectively, with 31%, 74% and 98% of the simulated correlations being smaller than those observed. A very similar pattern is apparent at the 100KB; the observed and expected correlations between CpG and non-CpG mutations, CpG transitions and transversions, and non-CpG transitions and transversions are 0.035 (expected = 0.045), 0.013 (0.0086) and 0.061 (0.050) respectively, with 7%, 76% and 94% of simulated correlations being smaller than the observed.

Correlations with genomic variables

To try and understand why there is large scale variation in the mutation rate, we compiled a number of genomic variables which have previously been shown to correlate to the rate of germline or somatic DNM, or divergence between species: recombination rate, GC content, replication time, nucleosome occupancy, transcription level, DNA hypersensitivity and several histone methylation and acetylation marks [3, 5, 14, 15].

The results from individual regressions and a multiple regression are broadly concordant, as are the results at the two different scales; we find that DNM density is positively correlated to DNA hypersensitivity, H3K27 acetylation, H3K4 methylation 1, nucleosome occupancy and recombination rate, and negatively correlated to H3K4 methylation 3, H3K9 methylation 3 and replication time (indicating lower mutation rates in early replicating DNA) (Table 2). The correlation for GC content changes from positive when regressed against DNM density by itself to negative in the multiple regression, and the correlations with H3K27 methylation 3 is positive at the 100KB scale but non-significantly negative at the 1MB scale. The biggest effect, as judged by the standardized slope is for nucleosome occupancy followed by GC content, recombination rate and replication time, which are similar in their level of correlation in the multiple regression (Table 2).

	100KB		1MB	
Factor	Individual regression slope	Multiple regression slope	Individual regression slope	Multiple regression slope
DNAse hypersensitivity	0.025***	0.027	0.026	
GC content	0.053***	-0.083	0.098***	-0.14
H3K27 acetylation	0.030***	0.039	0.036	
H3K27 methylation 3	0.0069***		-0.0056	-0.034
H3K4 methylation 1	0.052***	0.042	0.083***	0.12
H3K4 methylation 3	-0.014***	-0.039	-0.052*	-0.089
H3K9 methylation 3	-0.011***	-0.036	-0.086***	
Nucleosome occupancy	0.077**	0.12	0.15***	0.29
Recombination rate	0.071***	0.048	0.20***	0.12
Replication time	-0.020***	-0.081	-0.029	-0.13
RNA seq.	-0.0072***		-0.041*	

Table 2. The standardized slope from regressing DNM rate against individual features and all significant features in a multiple regression. The standardized slope is obtained by subtracting the mean from each feature and dividing it by its standard deviation; this makes the slopes from different features comparable. Note that a negative slope for replication time indicates that the mutation rate is higher for later replicating regions. Features were selected in the multiple regression model using backwards stepwise regression and AIC. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Overall the regression models only explain 7.0% and 1.4% of the variance in DNM density, at the 1MB and 100KB scales respectively, however since there are very few DNMs per MB or 100KB there is considerable sampling error. To estimate how much of the variance is potentially explainable we used the multiple regression model to predict the mutation rate for each MB or 100KB region and used these expectations to simulate the observed number of DNMs; we then calculated the coefficient of determination without refitting the regression model. The average coefficient of determinations from the simulations were 9.5% and 1.7% suggesting that the model explains 74% and 82% of the explainable variance at the 1MB and 100KB scales respectively.

This method can potentially over-estimate the explainable variance if sampling error affects the parameter estimates of the model substantially; i.e. if the fitted model was different for two sets of simulated data generated from the same model. To investigate whether this was the case we re-ran the simulation but refit the regression model each time, but only using the factors included in the model used to simulate the data. In this case the average coefficients of determination are 9.6% and 1.8% for the 1MB and 100KB scales, very similar to the coefficients when not refitting the model, suggesting that the model does not vary greatly between different simulated datasets. This also suggests that our regression model has substantial predictive power.

Correlation with divergence

The rate of divergence between species is expected to depend, at least in part, on the rate of mutation. To investigate whether variation in the rate of substitution is correlated to variation in the rate of mutation we calculated the divergence between humans and chimpanzees. There are however at least three different sets of human-chimpanzee alignments: pairwise alignments between human and chimpanzee (PW)[16] found on the University of California Santa Cruz (UCSC) Genome Browser, the human-chimp alignment

from the multiple alignment of 46 mammals (MZ)[17] from the same location, and the human-chimp alignment from the Ensembl Enredo, Pecan and Ortheus primate multiple alignment (EPO) [18]. We find that the correlation depends upon the human-chimpanzee alignments used and the amount of each block (either 1MB or 100KB) covered by aligned bases (Figure 4). The correlation is significantly negative if we include all windows for the UCSC PW and MZ alignments at the 1MB scale (similar results are obtained at 100KB), but becomes more positive as we restrict the analysis to windows with more aligned bases. In contrast the correlations are always positive when using the EPO alignments, and the strength of this correlation does not change once we get above 200,000 aligned bases per 1MB. Further analysis suggests there are some problems with the PW and MZ alignments because divergence per Mb window is inversely correlated to mean alignment length ($r = -0.31$, $p < 0.0001$) for the PW alignments and positively correlated ($r = 0.57$, $p < 0.0001$) for the MZ alignments (Figure S1). The EPO alignment method shows no such bias and we consider these alignments to be the best of those available. Therefore, we use the EPO alignments for the rest of this analysis.

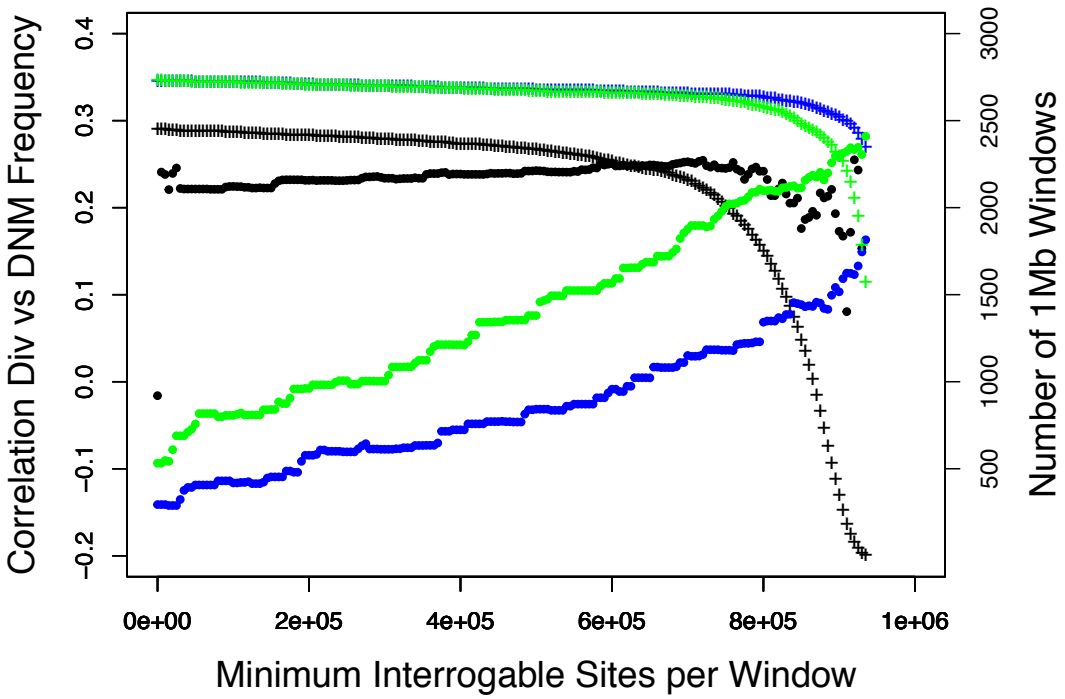


Figure 4. The correlation between divergence from human to chimpanzee and the density of DNMs in humans (dots) as a function of the number of aligned sites per window for three sets of alignments: UCSC pairwise alignments (blue), UCSC multi-way aligments (green) and EPO multi-species alignments alignments (black). Also shown as crosses are the number of 1MB blocks.

The EPO alignments allow us to consider lineage specific changes using parsimony to reconstruct ancestral states. As expected the divergence along the human lineage is correlated to the rate of DNMs (0.24 at the 1MB scale, 0.064 at the 100KB scale). However, the correlation between the rate of DNMs and divergence is not expected to be perfect even if variation in the mutation rate is the only factor affecting the rate of substitution between species; this is because we have relatively few DNMs and hence our estimate of the density of DNMs is subject to a large amount of sampling error. To investigate how strong the correlation could be, we follow the procedure

suggested by Francioli et al. [3]; we assume that variation in the mutation rate is the only factor affecting the variation in the substitution rate across the genome between species and that we know the substitution rate without error (this is an approximation, but the sampling error associated with the substitution rate is small relative to the sampling error associated with DNM density because we have so many substitutions). We generate the observed number DNMs according to the rates of substitution, and then consider the correlation between these simulated DNM densities and the observed substitution rates. We repeated this procedure 100 times to generate a distribution of expected correlations. Performing this simulation, we find that we would expect the correlation between divergence and DNM density to be 0.50 at the 1MB level and 0.24 at 100KB level, if variation in the mutation rate explained all the variation in the substitution rate, considerably greater than the observed values of 0.24 and 0.064 respectively. In none of the simulations was the simulated correlation as low as the observed correlation. Similar patterns hold for almost all mutational types; the level of divergence is positively correlated to the density of DNMs, often significantly so, but the observed correlations are substantially lower than the simulated correlations (Table S2).

There are several potential explanations for why the correlation is weaker than it could be; the pattern of mutation might have changed, or there might be other factors that affect divergence. Francioli et al. [3] showed that including recombination in a regression model between divergence and DNM density significantly improved the coefficient of determination of the model; a result we confirm here; the coefficient of determination when recombination is included in a regression of divergence versus DNM density increases from 0.058 to 0.18, and from 0.0041 to 0.048 for the 1MB and 100KB datasets respectively.

As detailed in the introduction there are at least three explanations for why recombination might be correlated to the rate of divergence independent of its effect on the rate of DNM: (i) biased gene conversion, (ii) recombination affecting the efficiency of selection, (iii) recombination affecting the depth of

the genealogy in the human-chimpanzee ancestor and (iv) problems with regressing against correlated variables that are subject to sampling error. We can potentially differentiate between these four explanations by comparing the slope of the regression between the rate of substitution and the recombination rate, and the rate DNM and the recombination rate. If recombination affects the substitution rate, independent of its effects on DNM mutations, because of biased gene conversion (BGC), then we expect the slope between divergence and recombination rate to be greater than the slope between DNM density and recombination rate for Weak->Strong (W->S), smaller for S->W, and unaffected for S<->S and W<->W changes. The reason is as follows; BGC increases the probability that a W->S mutation will get fixed but decreases the probability that a S->W mutation will get fixed. This means that regions of the genome with high rates of recombination will tend to have higher substitution rates of W->S mutations than regions with low rates of recombination hence increasing the slope of the relationship between divergence and recombination rate. The opposite is true for S->W mutations, and S<->S and W<->W mutations should be unaffected by BGC. If selection is the reason that divergence is correlated to recombination independently of its effects of the mutation rate, then we expect all the slopes associated with substitutions to be less than those associated with DNMs. The reason is as follows; if a proportion of mutations are slightly deleterious then those will have a greater chance of being fixed in regions of low recombination than high recombination. If the effect of recombination on the substitution rate is due to variation in the coalescence time in the human-chimp ancestor, then we expect all the slopes associated with substitution to be greater than those associated with DNMs; this because the average time to coalescence is expected to be shorter in regions of low recombination than in regions of high recombination. Finally, if the effect is due to problems with multiple regression then we might expect all the slopes to become shallower. Since the DNM density and divergences are on different scales we divided each by their mean to normalise them and hence make the slopes comparable.

The results of our test are consistent with the BGC hypothesis; the slope of divergence versus RR is greater than the slope for DNM density versus RR

for all W->S mutations and less for all S->W mutations (Figure 5); these differences are significant for most of the comparisons at the 1MB and 100KB scales (Table 3)(significance was assessed by bootstrapping the data by block and then recalculating the slopes). There are no significant differences between the slopes for W<->W and S<->S mutations except at the 100KB scale for non-CpG S<->S. Never-the-less it is worth noting that the DNM slope is consistently greater than the divergence slope at both spatial scales suggesting that there might be some effect of recombination affecting the efficiency of selection. Unfortunately, the obvious analysis, of regressing the W<->W and S<->S substitution rate against DNM density and recombination rate is inconclusive; rather than observe a negative correlation as we might expect under the efficiency of selection model we observe that the substitution rate is significantly positively correlated to RR. However, this might be simply due to RR and DNM density being positively correlated but DNM density being an error prone measure of the mutation rate (i.e. within the multiple regression we cannot hold the mutation rate constant).

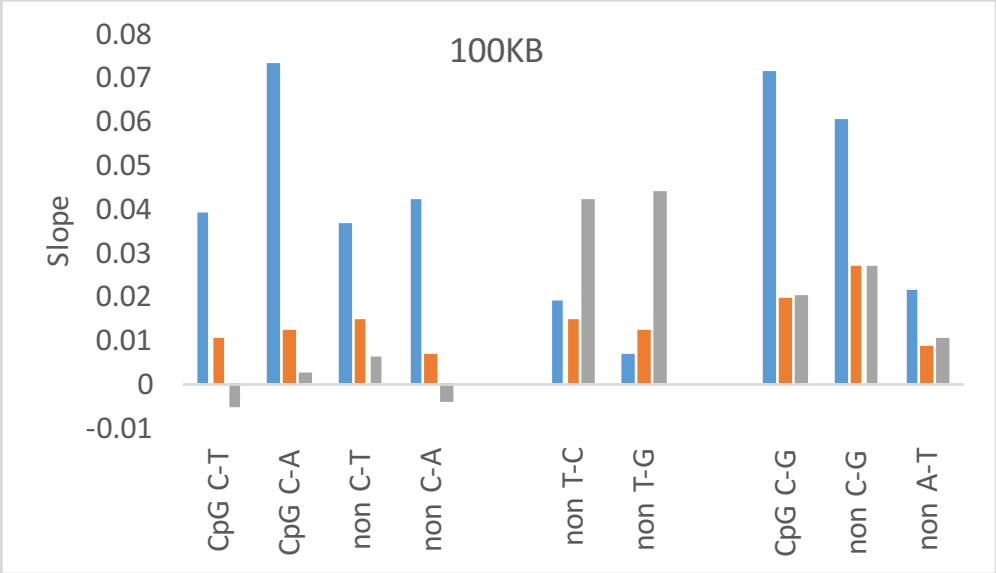


Figure 5. The slope between normalised DNM density and recombination rate (RR) (blue), normalised SNP density and RR (red), and normalised substitution density and RR (green). In each case the values were normalised by dividing the values by the mean.

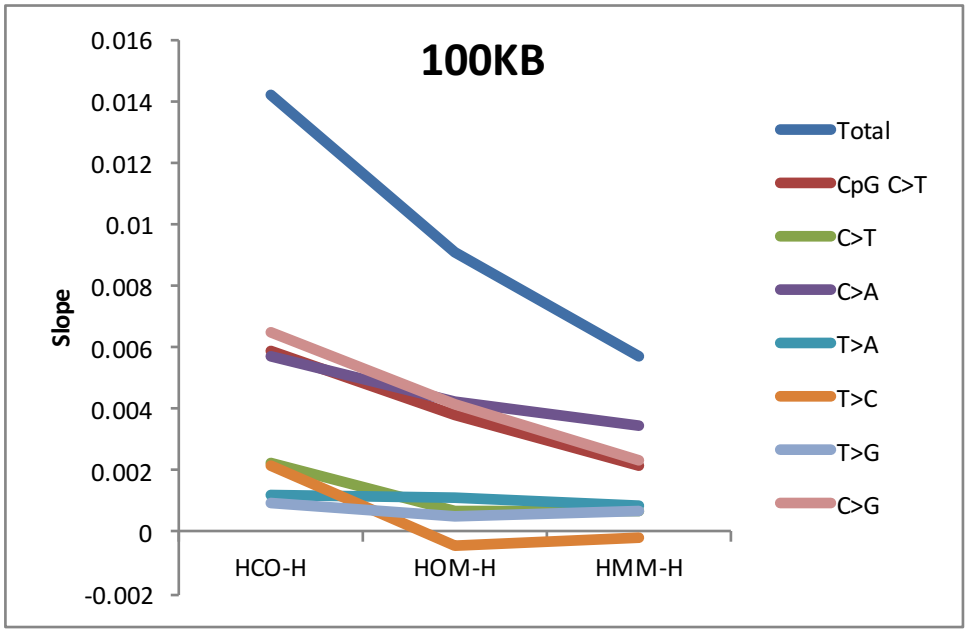
	100KB		1MB	
	Proportion of bootstraps diversity slope > DNM slope	Proportion of bootstraps divergence slope > DNM slope	Proportion of bootstraps diversity slope > DNM slope	Proportion of bootstraps divergence slope > DNM slope
CpG C->T	0.0045	0	0.085	0.002
CpG C->A	0.11	0.074	0.096	0.065
non-CpG C->T	0.0016	0	0.0029	0
non-CpG C->A	0.0010	0	0.078	0.0064
non-CpG T->C	0.28	0.9997	0.54	1
non-CpG T->G	0.67	0.9982	0.80	0.9999
CpG C->G	0.13	0.13	0.24	0.29
non-CpG C<->G	0.0041	0.0043	0.31	0.26
non-CpG A<->T	0.19	0.22	0.17	0.28

Table 3. Proportion of bootstrap replicates in which the slope of the normalised diversity versus recombination rate, or normalised divergence versus recombination rate, is greater than the slope of the normalised DNM density and recombination rate. 10,000 bootstrap replicates were performed in each case.

Other species

Divergence between species, usually humans and macaques, is often used to control for mutation rate variation in various analyses. But how does the

correlation between divergence and the DNM rate in humans change as the species being compared get further apart? To investigate this, we compiled data from a variety of primate species – human/chimpanzee/orang-utan (HCO) considering the divergence along the human and chimp lineages, human/orang-utan/macaque (HOM) considering the divergence along the human and orang-utan lineages, and human/macaque/marmoset (HMM) considering the divergence along the human and macaque lineages. This yields two series of divergences of increasing evolutionary divergence: the human lineage from HCO, HOM and HMM, and chimp from HCO, orang-utan from HOM and macaque from HMM. All divergences were normalised by dividing by their mean. For both series we see a clear tendency for the slope of the regression between divergence and DNM rate to decrease as a function of evolutionary divergence at both the 100kb and 1MB scales (Figure 6 for the human lineage, Figure S2 for the chimpanzee, orang-utan ad marmoset lineages). If we calculate the correlation coefficient between the slope and the evolutionary stratum, assigning 1, 2 and 3 to the strata (e.g. 1 for chimp, 2 for orangutan and 3 for macaque), we find that the correlations are negative for all mutational types, for both sets of evolutionary divergence and scales (binomial test of positive versus negative for both 100kb and 1MB using both the human and other lineage $p < 0.01$) (Figure 6).



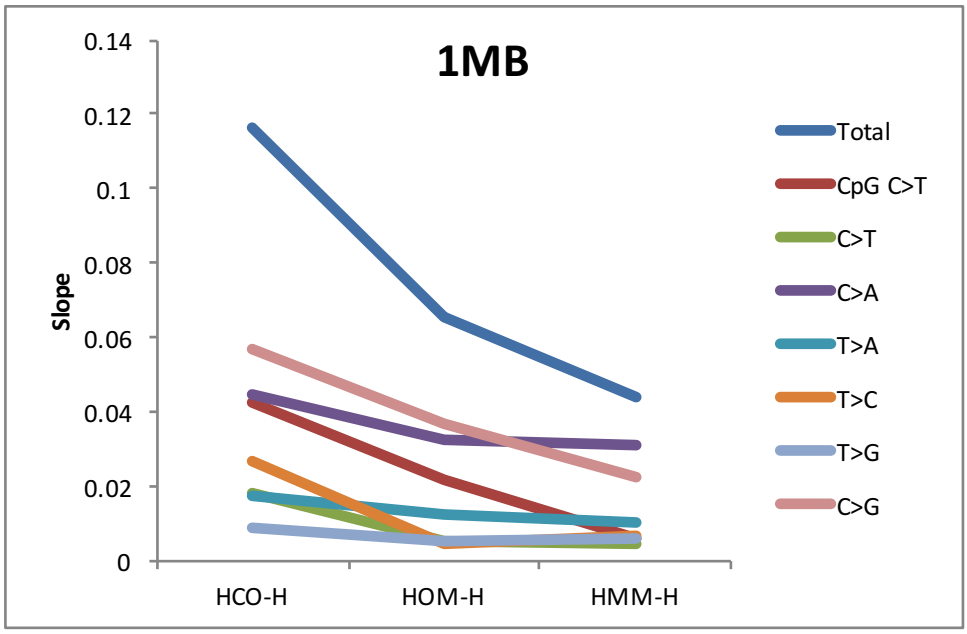


Figure 6. The slope of the linear regression between divergence and DNM rate for 100kb (top panel) and 1MB (bottom panel). HCO-H is the human divergence since humans split from chimpanzee, from a comparison of human, chimpanzees and orang-utans; HOM-H is the human divergence since humans split from orang-utans, using human, orang-utan and macaque; HMM-H is the human divergence since humans split from macaques using human, macaque and marmoset.

Correlation with diversity

Just as we expect there to be correlation between divergence and DNM rate, so we might expect there to be correlation between DNA sequence diversity within the human species and the rate of DNM. To investigate this, we compiled the number of SNPs in 1MB and 100kb blocks from the 1000 genome project [19, 20]. There is a positive correlation between SNP density and DNM rate at both the 1MB ($r = 0.36$, $p<0.001$) and 100KB scales ($r = 0.13$, $p<0.001$). This positive correlation is observed for all mutational types, however in some cases the correlations are not significant (Table S3).

Using a similar strategy to that used in the analysis of divergence we calculated the correlation we would expect if all the variation in diversity was due to variation in the mutation rate by assuming that the level of diversity was known without error, and hence was a perfect measure of the mutation rate (we have on average 31,000 SNPs per MB, so there is little sampling error associated with the SNPs). We then simulated the observed number of DNMs according to these inferred mutation rates. The expected correlations are 0.41 and 0.17 at the 1MB and 100KB scales; these are significantly greater than the observed correlation ($p<0.01$ in both cases) but the difference is less dramatic than the difference for divergence. However, this is deceptive because for most mutational types the observed correlation is considerably smaller than the expected correlation; on average the observed correlation is ~45% the expected correlation when each mutational type is considered separately (Table S3), fairly similar to the average effect seen for divergence (Table S2).

The fact that the correlation between diversity and DNM density is not as strong as it could be, could be caused by BGC. To investigate this, we repeated our BGC test used in the analysis of the divergence data – i.e. we compared the slope of the relationship between diversity and recombination rate to the slope of the regression between DNM density and recombination rate (as before the variables were normalised by dividing by the mean). As

expected we find the slope of the regression between diversity and RR to be greater than the slope between DNM density and RR for all W->S mutations and less than for all S->W, except non-CpG T->C mutations (Figure 5). In almost all cases the slope of diversity versus RR is between the slope of divergence versus RR and DNM versus RR, as expected, since BGC is expected to have smaller effects on diversity than divergence. The effects are often significant at the 1MB scale but not significant at the 100KB scale (Table 3).

As with divergence we observe that the slope associated with mutational types not affected by BGC is lower for diversity than DNM, which is consistent with selection being more efficient against deleterious mutations in regions of the genome with higher RR. However, the differences in slope are not significant except for non-CpG S<->S changes.

Discussion

We have considered the large-scale distribution of DNMs along the human genome and the relationship between the rate of DNM, divergence between species, and diversity within a species. We find evidence that there is large scale variation in the mutation at the 100KB and 1MB; this is variation that cannot be explained by variation at smaller scales. However, the variation in the mutation rate is quite modest; at the MB scale 90% of regions have a mutation rate that is within ± 35% of the mean, at the 100KB scale this increases to ± 47% of the mean. It seems likely that there will be more variation at smaller scales but how this will scale up remains to be investigated.

Although we do not have enough DNM data to consider each mutational type individually, it is evident that the rates of CpG and non-CpG mutation vary across the genome as do the rates of non-CpG transitions and transversions at both the 1MB and 100KB scales; we do not have enough data to determine what is happening with CpG transversions. The rate of mutation of the different mutational types are about as strongly correlated to each other as

they could be, suggesting that they vary in concert and are likely to be influenced by similar factors.

We confirm that replication time, recombination rate and GC content are all independently correlated to the rate of DNM, but we also show that nucleosome occupancy and two histone marks are correlated to DNM density. The strongest effect we find comes from nucleosome occupancy. Although, nucleosome occupancy has previously been investigated, no significant effects were detected [5]. However, Michaelson et al. [5] only considered a small number of DNMs at a scale of single nucleotides. Overall the regression model explains 73% and 82% of the explainable variance at the 1MB and 100KB scales respectively, which is quite remarkable since the factors, such as replication time, are not being measured in the relevant cells, the male and female germ-line.

Some caution should be exercised in interpreting the results of the multiple regression because most of the variables in the model are subject to experimental error. This means that some variables might be included in the model when they should not be. For example, let us imagine that factor X affects the mutation rate whereas factor Y does not; however, X and Y are mildly correlated. If we can measure X and Y without error then a multiple regression should show that the rate of DNM is correlated to X but not Y. However, if we cannot measure X without error then we may find that the rate of DNM correlates to both X and Y.

As expected the rate of divergence between species is correlated to the rate of DNM, however, the strength and even the sign of the correlation depends on the alignments being used. The correlations between divergence and DNM density are actually negative if no filtering is applied to the UCSC alignments, and there is a negative correlation between divergence and alignment length for the pairiwise alignments from the UCSC genome browser, and a positive correlation for the multi-species alignment. It is clear that there are problems with these alignments and results obtained using these alignments should be treated with caution.

As Francioli et al. [3] showed the correlation between divergence and DNM density is worse than it would be if variation in the mutation rate was the only factor affecting divergence. We show that this is also true for diversity within humans. Francioli et al. [3] showed that although the rate of DNM is correlated to the rate of recombination, divergence is correlated to the rate of recombination independently of this effect. We have shown that the reason recombination affects divergence and diversity independently of its effects on the rate of mutation is likely to be due to effect of biased gene conversion since the slope of the relationship between divergence and recombination rate is smaller than the slope for DNM rate and recombination for S->W changes, but greater for W->S changes; W<->W and S<->S changes are unaffected.

Although, biased gene conversion appears to affect the relationship between both divergence and diversity, and the rate of mutation, this is clearly not the only factor, since the correlation between divergence, diversity and DNM density for mutations, which are unaffected by biased gene conversion, is worse than it could be if all the variation in divergence and diversity for these mutational types was caused by variation in the rate of mutation; the difference between the expected and observed correlation is generally significant at both scales (Tables S2 and S3). The fact that the relationship between divergence or diversity and DNM density is not as strong as it could, could be due to a number of reasons. First the mutation rate might be evolving through time. In this case, we might expect the ratio of the observed and expected correlations, for W<->W and S<->S mutations, to be smaller for divergence than diversity and yet they are remarkably similar (the average ratio for divergence = 0.52, for diversity = 0.56, Tables S2, S3). Second, there might be variation in the effective population size across the genome; this would generate variation in diversity that is not associated with the mutation rate, and potentially variation in the divergence through variation in coalescence time in the human-chimp ancestor. Here one would expect the effect on the correlation between divergence and DNM density to be smaller than the effect for diversity, since variation in the effective population size will only affect the overall divergence to a small extent. Third, variation in effective

population size across the genome could generate variation in the effciciency of selection. But again we would expect the effect to be different for divergence and diversity. The reason why the correlation between divergence, diversity is less than perfect for W<->W and S<->S mutations remains unclear.

We also show that the relationship between divergence and DNM rate gets weaker (the slopes get shallower) as more and more divergent species are considered. This might be due to two factors. First we might expect the mutation rate of a region to evolve through time eroding the relationship between divergence and the current mutation rate. Second, the relationship might get weaker because we are underestimating the divergence as species get more divergent. This might tend to affect the most divergent blocks the most. However, we see no obvious effect of this; the mutation that should be most affected is CpG transitions and the decay in the slope (between divergence and DNM rate) is no faster than for other mutational types (Figure 6).

These results are consistent with those of Terekhanova et al. [21] who showed that the substitution rate for W<->W and S<->S along the human lineage was correlated to that of other primates at the 1MB scale, but that the strength of this correlation declined as more divergent species were considered. They showed that a fraction of this correlation was due to variation in the substitution rate that was not correlated to genomic features in humans; possibly the 25% of the variance that we find is unexplainable by genomic features.

Divergence between species has often been used to control for mutation rate variation in humans (for example [22-24]). This is clearly not satisfactory given that divergence is more strongly correlated to the rate of recombination than the rate of DNM, and the relationship between divergence and the rate of DNMs decreases as evolutionary divergence increases. However although we have too few DNMs to construct a mutation rate map directly, our regression model for predicting the mutation rate from genomic features is sufficiently

good to yield a reasonable prediction of the mutation rate, at least down to 100KB scale. Never-the-less it should be appreciated that there may be much more variation in the mutation rate at finer scales and that it may be necessary to control for this variation in some analyses.

It has been known for sometime that diversity across the human genome is correlated to the rate of recombination[9-11] and there has been much debate about whether this is due to mutagenic effects of recombination or the effect of recombination on processes such as genetic hitch-hiking and background selection. Divergence between humans and other primates is correlated to the rate of recombination, which was initially interpreted as being due to a mutagenic effect of recombination [9, 11] but subsequently it has been interpreted as evidence of BGC [25]. Both of these hypotheses appear to be correct – the rate of DNM is correlated to the rate of recombination ([3]; results above), but recombination also affects which mutations become fixed through BGC.

Materials and methods

DNM data

Details of DNM mutations were downloaded from the supplementary tables of the respective papers: 26,939 mutations from Wong et al. [6], 11016 mutations from Francioli et al. [3], 4931 mutations from Kong et al. [4] and 547 mutations from Michaelson et al. [5]. These were all mapped to hg19/GRCh37. Only autosomal DNMs were used. From these DNMs we constructed a series of datasets. In the first we considered all DNMs; in subsequent analyses we only considered DNMs that mapped to regions of the genome for which we had all genomic variables, such as replication time data, or which mapped to regions for which we had divergence data.

Alignments.

Three sets of alignments were used in this analysis, all based on human genome build hg19/GRCh37: (i) the University of California Santa Cruz (UCSC) pairwise (PW) alignments [16] for human-chimpanzee (hg19-panTro4

downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsPanTro4/) (ii) the UCSC MultiZ (MZ) 46-way alignments [17] downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/multiz46way/ and (iii) Ensembl Enredo, Pecan, Ortheus (EPO) 6 primate multiple alignment, release 74, [18] downloaded from ftp://ftp.ensembl.org/pub/release-74/emf/ensembl-compara/epo_6_primate/. We found that the EPO alignments were the most reliable – see main text – and they were used for the majority of the analyses.

Filtering of EPO alignments and construction of main data set.

In analyses involving the divergence betwee species we only considered DNMs that mapped to sequences that were alignable between species, in 100KB and 1MB blocks in which at least half the sequence was alignable between the species. This left us with 35,401 DNMs for the human/chimpanzee/orang-utan comparison, 31,185 DNMs for the human/orang-utan/macaque comparison and 23,534 DNMs for the human/macaque/marmoset comparison.

Selection and filtering of SNPs.

All SNPs from the 1000 genomes project phase 3 [20] were downloaded from hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/. After removing all multi-allelic SNPs and, structural variants and indels we were left with 77,818,368 autosomal SNPs. After filtering out windows which had less than 50% of nucleotides aligning between human-chimpanzee-orangutan and no recombination rate scores we were left with 71,917,321 SNPs.

Genomic features.

Male specific standardised recombination rate data [26] was downloded from http://www.decode.com/additional/male.rmap, which provides recombination rates in 10KB steps. For each 100KB and 1Mb window the recombination rate was calculated as the mean of these scores with a score assigned to the window in which the position of its first base resided. For replication time data we downloaded the Encode Repli-seq wavelet smoothed signal data [27, 28], provided in 1kb steps, for the GM12878, HeLa, HUVEC, K562, MCF-7 and

HepG2 cell lines from the UCSC ftp site ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/. We computed the mean replication time for all autosomes for 100kb and 1Mb windows across all 6 cell lines. Replication times were assigned to windows based upon their start coordinates. GC content was calculated directly from the human genome (hg19/GCRh37) for 100kb and 1Mb windows. Nucleosome occupancy for the GM12878 cell line was used [29] downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhNsome/. Nucleosome occupancy scores are provided at high, but variable resolution, with scores spanning 1 to 27,362 bases. Mean nucleosome occupancy was calculated per 100kb and 1Mb window, accounting for this variation.

Statistical analysis.

The R stats package, R version 3.3.1, was used for all correlations and regression analyses of observed variables. Simulations to derive expected variables and comparisons to observed variables were done using Mathematica version 10.

To estimate the mutation rate distribution we use the method of [30]. In brief we assume that the mutation rate in each block is $\alpha\bar{u}$ where $\bar{u}$ is the average mutation rate per site and α is the rate above or below this mean. α is assumed to be gamma distributed. The number of mutations per block is assumed to be Poisson distributed with a mean $\alpha\bar{u}l$ where *l* is the length of the block. This means that the number of mutations per block is a negative binomial. We fit the distribution using maximum likelihood using the *NMaximize* function in *Mathematica*.

We investigated the correlation between different types of mutation across blocks by fitting a single distribution to both types of mutation; i.e. by finding the distribution which when fitted to both distributions of mutations across sites, maximizes the likelihood. We then used this distribution to simulate data; we drew a random variate for each block from the distribution assigning

this as the rate for that block. We then generated two Poisson variates with the appropriate means such that the total number of DNMs for each type of mutation was expected to equal the total number of DNMs of those types. A similar procedure was used to test the fit of the regression model.

Literature cited

1. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics. 2011;12(11):756-66. Epub 2011/10/05. doi: 10.1038/nrg3098. PubMed PMID: 21969038.
2. Conrad DF, Keebler JE, Depristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011;43(7):712-4. Epub 2011/06/15. doi: ng.862 [pii] 10.1038/ng.862. PubMed PMID: 21666693.
3. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. Nature genetics. 2015;47(7):822-6. Epub 2015/05/20. doi: 10.1038/ng.3292. PubMed PMID: 25985141; PubMed Central PMCID: PMC4485564.
4. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471-5. doi: Doi 10.1038/Nature11396. PubMed PMID: ISI:000307761600028.
5. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151(7):1431-42. Epub 2012/12/25. doi: 10.1016/j.cell.2012.11.019. PubMed PMID: 23260136.
6. Wong WS, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, et al. New observations on maternal age effect on germline de novo mutations. Nat Commun. 2016;7:10486. doi: 10.1038/ncomms10486. PubMed PMID: 26781218; PubMed Central PMCID: PMCPMC4735694.
7. Haldane JB. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. Annals of eugenics. 1947;13(4):262-71. PubMed PMID: 20249869.
8. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285-311. Epub 2009/07/28. doi: 10.1146/annurev-genom-082908-150001. PubMed PMID: 19630562.
9. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet. 2003;72(6):1527-35. Epub 2003/05/13. doi: S0002-9297(07)60451-0 [pii] 10.1086/375657. PubMed PMID: 12740762; PubMed Central PMCID: PMC1180312.
10. Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, Ptak SE. Why do human diversity levels vary at a megabase scale? Genome Res. 2005;15(9):1222-31. Epub 2005/09/06. doi: 15/9/1222 [pii]

10.1101/gr.3461105. PubMed PMID: 16140990; PubMed Central PMCID: PMC1199536.
11. Lercher MJ, Hurst LD. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 2002;18(7):337-40. Epub 2002/07/20. doi: S0168952502026690 [pii]. PubMed PMID: 12127766.
12. Eyre-Walker A, Eyre-Walker YC. How Much of the Variation in the Mutation Rate Along the Human Genome Can Be Explained? G3-Genes Genom Genet. 2014;4(9):1667-70. doi: Doi 10.1534/G3.114.012849. PubMed PMID: ISI:000342570600011.
13. Harris K. Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci U S A. 2015;112(11):3439-44. doi: 10.1073/pnas.1418652112. PubMed PMID: 25733855; PubMed Central PMCID: PMCPMC4371947.
14. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488(7412):504-7. doi: 10.1038/nature11273. PubMed PMID: 22820252.
15. Tyekucheva S, Makova KD, Karro JE, Hardison RC, Miller W, Chiaromonte F. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol. 2008;9(4):R76. Epub 2008/05/02. doi: gb-2008-9-4-r76 [pii] 10.1186/gb-2008-9-4-r76. PubMed PMID: 18447906; PubMed Central PMCID: PMC2643947.
16. Chiaromonte F, Yap VB, Miller W. Scoring pairwise genomic sequence alignments. Pac Symp Biocomput. 2002:115-26. PubMed PMID: 11928468.
17. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004;14(4):708-15. doi: 10.1101/gr.1933104. PubMed PMID: 15060014; PubMed Central PMCID: PMCPMC383317.
18. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18(11):1814-28. doi: 10.1101/gr.076554.108. PubMed PMID: 18849524; PubMed Central PMCID: PMCPMC2577869.
19. 1000_Genomes_Project_Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73. Epub 2010/10/29. doi: nature09534 [pii]
10.1038/nature09534. PubMed PMID: 20981092; PubMed Central PMCID: PMC3042601.
20. 1000_Genomes_Project_Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65. doi: Doi 10.1038/Nature11632. PubMed PMID: ISI:000310434500030.
21. Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. Evolution of local mutation rate and its determinants. Mol Biol Evol. 2017;in press.
22. Burgess R, Yang Z. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. 2008;25(9):1979-94. Epub 2008/07/08. doi: msn148 [pii]
10.1093/molbev/msn148. PubMed PMID: 18603620.
23. Gossmann TI, Woolfit M, Eyre-Walker A. Quantifying the variation in the effective population size within a genome. Genetics. 2011;189(4):1389-402.

Epub 2011/09/29. doi: 10.1534/genetics.111.132654. PubMed PMID: 21954163; PubMed Central PMCID: PMC3241429.
24. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5(5):e1000471. Epub 2009/05/09. doi: 10.1371/journal.pgen.1000471. PubMed PMID: 19424416; PubMed Central PMCID: PMC2669884.
25. Duret L, Arndt PF. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 2008;4(5):e1000071. Epub 2008/05/10. doi: 10.1371/journal.pgen.1000071. PubMed PMID: 18464896; PubMed Central PMCID: PMC2346554.
26. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002;31(3):241-7. PubMed PMID: 12053178.
27. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 2007;17(6):917-27. doi: 10.1101/gr.6081407. PubMed PMID: 17568007; PubMed Central PMCID: PMCPMC1891350.
28. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(1):139-44. Epub 2009/12/08. doi: 10.1073/pnas.0912402107. PubMed PMID: 19966280; PubMed Central PMCID: PMC2806781.
29. ENCODE-Project-Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799-816.
30. Hodgkinson A, Chen Y, Eyre-Walker A. The large scale distribution of somatic mutations in cancer genomes. Human Mutation. 2012;33:136-43.

Supplementary Tables

Model	# of parameters	Log likelihood
100 kb		
Combined distribution for CpG and non-CpG	3	-57399.40
Separate distributions for CpG and non-CpG	4	**-57397.10**
Combined distribution for CpG transitions and transversions	3	-18612.00
Separate distributions for CpG transitions and transversions	4	-18612.00
Combined distribution for non-CpG transitions and transversions	3	-58207.50
Separate distributions for non-CpG transitions and transversions	4	-58205.70
1 MB		
Combined distribution for CpG and non-CpG	3	-13179.50
Separate distributions for CpG and non-CpG	4	**-13176.03**
Combined distribution for CpG transitions and transversions	3	-6465.54
Separate distributions for CpG transitions and transversions	4	-6464.82
Combined distribution for non-CpG transitions and transversions	3	-13404.10
Separate distributions for non-CpG transitions and transversions	4	**-13401.79**

Table S1. Likelihood values for fitting combined and separate distributions to categories of mutations. Each pair of lines represents a likelihood ratio test; bold figures denote a significant result.

	100kb	100kb	1MB	1MB
Mutation	Obs. Correlation	Exp. correlation	Obs. correlation	Exp. correlation
All	0.064***	0.24**	0.24***	0.50**
CpG C-T	0.055***	0.11	0.20***	0.22
CpG C-A	0.021**	NA	0.053*	0.084
CpG C-G	0.011	NA	0.033	0.087**
non C-T	0.016*	0.14	0.063**	0.28**
non C-A	0.047***	0.12	0.16***	0.28**
non T-C	0.015*	0.14	0.080***	0.32**
non T-G	0.0099	0.096**	0.040	0.22**
non C-G	0.059***	0.11	0.23***	0.25
non T-A	0.013	0.091**	0.094***	0.18**

Table S2. The observed and expected correlations between the density of DNMs and substitutions at the 100kb and 1MB scales; the expected correlation is the mean correlation from 100 simulations assuming that all the variation in the substitution rate is due to variation in the mutation rate (and assuming the pattern of mutation has not changed along the human lineage). We are not able to simulate data for CpG transversions due to the fact that some regions have no substitutions of this type. Indicated is whether the observed correlation is greater than zero and whether the expected correlation is significantly greater than the observed. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	100kb	100kb	1MB	1MB
Mutation	Obs. Correlation	Exp. correlation	Obs. correlation	Exp. correlation
All	0.13***	0.17	0.36***	0.41**
CpG C-T	0.074***	0.11**	0.16***	0.24**
CpG C-A	0.011	0.034**	0.019	0.073*
CpG C-G	0.019**	0.032*	0.065**	0.063
non C-T	0.032***	0.12**	0.082***	0.29**
non C-A	0.055***	0.10**	0.13***	0.25**
non T-C	0.013	0.12**	0.025	0.28**
non T-G	0.011	0.070**	0.026	0.18**
non C-G	0.085***	0.12**	0.25***	0.32**
non T-A	0.027***	0.073**	0.069***	0.18**

Table S3. The observed and expected correlations between the density of DNMs and the density of SNPs at the 100kb and 1MB scales; the expected correlation is the mean correlation from 100 simulations assuming that all the variation in the density of SNPs is due to variation in the mutation rate. Indicated is whether the observed correlation is greater than zero and whether the expected correlation is significantly greater than the observed. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

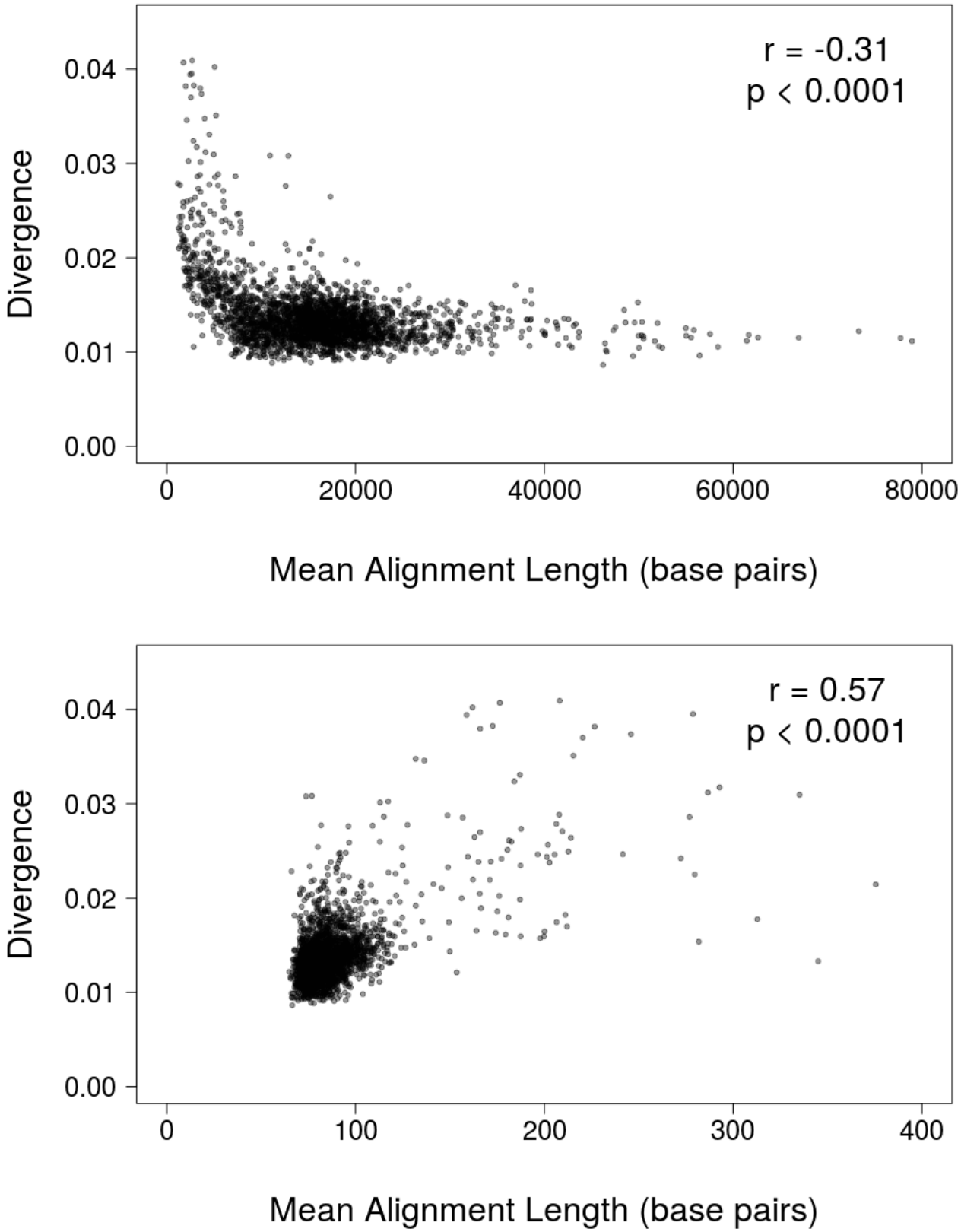
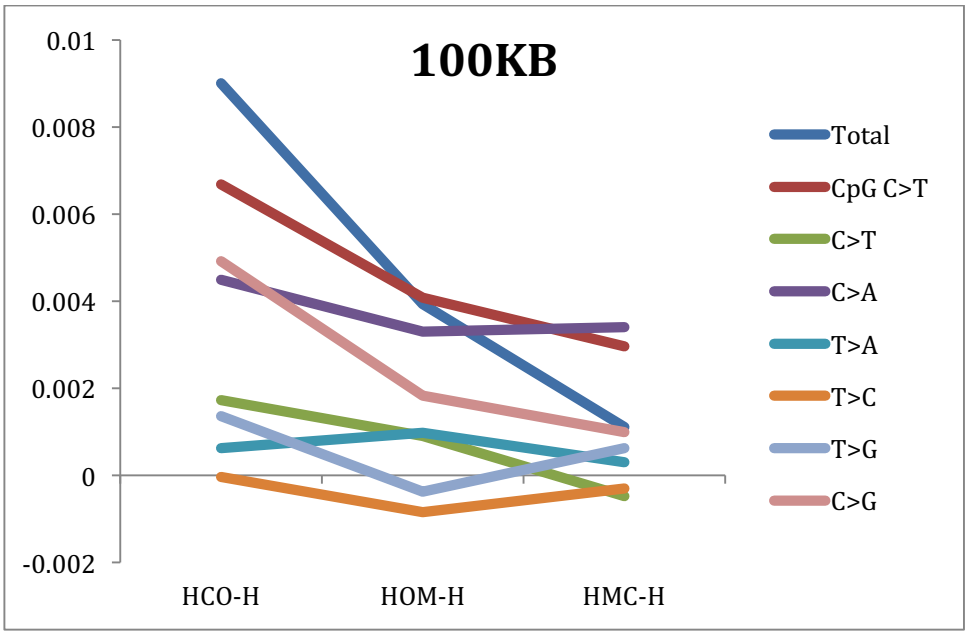


Figure S1. Divergence (number of substitutions per bas pair) as a function of alignment length in the UCSD pairwise alignments (top panel) and the UCSD multiz alignments (bottom panel).



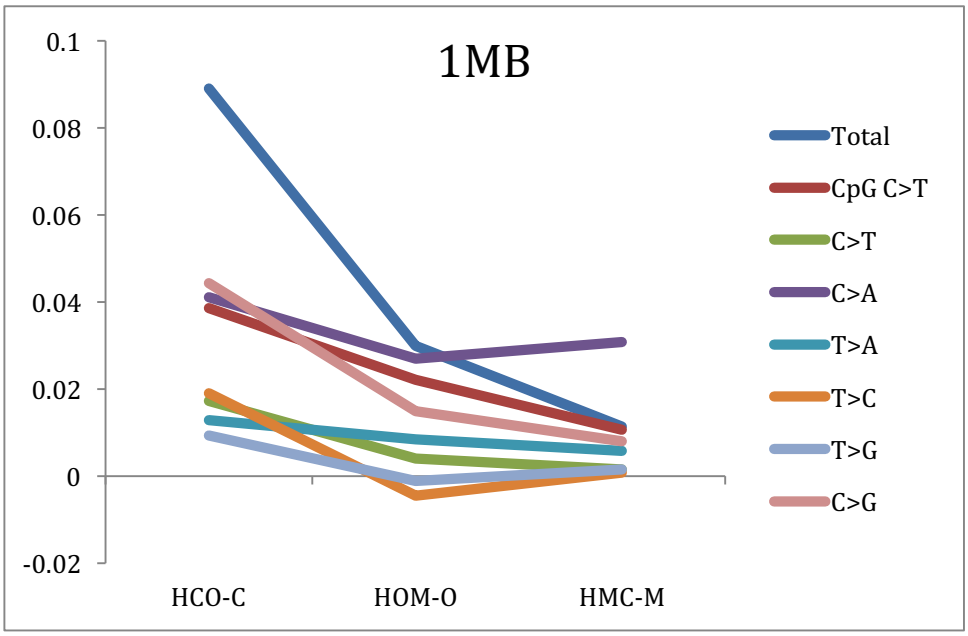


Figure S2. The slope of the linear regression between divergence and DNM rate for 100kb (top panel) and 1MB (bottom panel). HCO-C is the chimpanzee divergence since humans split from chimpanzee, from a comparison of human, chimpanzees and orang-utans; HOM-O is the orang-utan divergence since humans split from orang-utans, using human, orang-utan and macaque; HMC-C is the *Callithrix* divergence since humans split from macaques using human, macaque and *Callithrix*.