Heterogeneity of the Transition/Transversion Ratio in
Drosophila and Hominidae Genomes

Vladimir B. Seplyarskiy,*1,2 Peter Kharchenko,3 Alexey S. Kondrashov,1,4 and Georgii A. Bazykin1,2

1Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia
2Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia
3Center for Biomedical Informatics, Harvard Medical School
4Department of Ecology and Evolutionary Biology and Life Sciences Institute, University of Michigan

*Corresponding author: E-mail: pamjat@mail.ru.
Associate editor: Koichiro Tamura

Abstract

Mutation rate varies between sites in the genome. Part of this variation can be explained by well-recognized short nucleotide contexts, but a large component of this variation remains cryptic. We used data on interspecies divergence and intraspecies polymorphism in Drosophila and Hominidae to analyze variation of the average rate of the 12 possible kinds of single-nucleotide mutations and in the transition/transversion ratio $\kappa$ at single-nucleotide resolution. Both the average mutation rate and $\kappa$ vary by a factor of $\sim 3$ between nucleotide sites. The characteristic scale of variation in $\kappa$ is up to at least $\sim 50$ nucleotides in Drosophila and $\sim 5$ nucleotides in Hominidae. Genome segments with locally elevated mutation rates possess lower values of $\kappa$; however, a substantial fraction of variation in $\kappa$ cannot be directly explained by the local mutation rates.

Key words: transition/transversion ratio, mutagenesis, SNP, D. melanogaster, H. sapiens.

Introduction

The mutation rate is known to vary along the genome at different scales. At a large scale, different chromosomes may mutate at rates that differ from each other by a factor of two or more (Castreres et al. 2004). At a medium scale, the mutation rate varies between chromosome segments characterized by different replication timings and chromatin accessibilities (Stamatoyannopoulos et al. 2009; Chen et al. 2010), as well as due to other factors. At a small scale, the mutation rate varies between individual nucleotide sites, due to influences both of simple contexts, such as CpG or TpT in mammals (Hwang and Green 2004; Arndt and Hwa 2005; Zhao et al. 2006; Hodgkinson and Eyre-Walker 2011), or more complex contexts spanning several neighboring nucleotides (Panchin et al. 2011), and of unknown “cryptic” factors (Hodgkinson et al. 2009; Johnson and Hellmann 2011).

Ignoring insertions, deletions, and complex mutations, the mutation process can be fully described by the rates of each of the 12 possible single-nucleotide substitutions. These rates can differ substantially. In particular, transversions are typically approximately two times rarer than transitions in D. melanogaster (Begun et al. 2007), and approximately four times rarer in humans (Nachman and Crowell 2000; Kondrashov 2003; Hodgkinson and Eyre-Walker 2010). It is also known that the relative rates of each of the 12 substitutions can vary along the genome (Hodgkinson and Eyre-Walker 2011), but this variation has not been studied in detail.

Here, we extend the analysis of Hodgkinson et al. (2009) and Johnson and Hellmann (2011), who studied small-scale variation of the average mutation rate, and investigate small-scale variation of the transition/transversion ratio. We detected a substantial heterogeneity of this ratio between individual sites in Drosophila and in Hominidae, with a significant positive correlation between adjacent sites. Most of the variation we observe is cryptic, in the sense that we cannot attribute it to any simple cause.

Materials and Methods

Data

Multiple complete genome alignment of D. simulans, D. yakuba, and D. erecta to D. melanogaster (dm3) was downloaded from UCSC Genome Browser (Fujita et al. 2011) (http://genome.ucsc.edu). Polymorphism data for 37 complete genomes of D. melanogaster, aligned to dm3 reference genome of D. melanogaster (Jordan et al. 2007, release 0.5), and for six complete genomes of D. simulans aligned to dm2 reference genome of D. melanogaster (Begun et al. 2007) were downloaded from DPGP (http://www.dpgp.org/). dm2 coordinates were converted to dm3 coordinates using lift-over (http://hgdownload.cse.ucsc.edu/liftOver/dm3/liftOver/). FlyBase genes (Tweedie et al. 2009, BDGP release 5) were used to map D. melanogaster introns and intergenic regions onto the four-species alignment.

Multiple complete genome alignment of Pongo pygmaeus, Gorilla gorilla, and Pan troglodytes to Homo sapiens (hg18) was obtained from UCSC Genome Browser (Fujita et al. 2011) (http://genome.ucsc.edu). Data on human variation were obtained from nine diploid human nuclear genotypes downloaded from Galaxy bioinformatics platform (Taylor et al. 2007; Schuster et al. 2010).
Levy et al. 2007

Four phylogenetic analyses (I–IV) used for studying the local variation in the overall mutation rate and in the transition/transversion ratio in Drosophila and Hominidae. Lines denote phylogenetic lineages, and solid triangles denote within-species polymorphism. Black color corresponds to the ancestral allele, and red and blue colors correspond to different derived alleles. In each analysis, the genome-wide estimate of the frequency of SNPs or replacements on a particular phylogenetic branch (red) is compared with the frequency of SNPs or replacements on the same phylogenetic branch (red) at sites, or in the vicinity of sites, where another SNP or replacement has occurred (blue). Dyak, Dere, Dsim, and Dmel stand for D. yakuba, D. erecta, D. simulans, and D. melanogaster; Ppyg, Ggor, Ptro, and Hsap stand for Pongo pygmaeus, Gorilla gorilla, Pan troglodytes, and Homo sapiens.

(http://usegalaxy.org) and the reference human genome, for a total of 19 haploid genomes. The following individual diploid genotypes were used: KB1 (454 method) (Schuster et al. 2010), ABT (SOLiD method) (Schuster et al. 2010), NA18507 (Bentley et al. 2008), NA19240 (Drmanac et al. 2010), Craig Venter (Levy et al. 2007), NA12891, NA12892, Chinese individual (Wang et al. 2008), and Korean individual (Ahn et al. 2009). UCSC hg18 Known Genes (Hsu et al. 2006) were used to map H. sapiens introns and intergenic regions onto the alignment.

Only constitutive introns, and constitutive segments of intergenic regions, were kept for analysis; coding regions were excluded to reduce the effect of natural selection on our inferences. Nucleotides masked by RepeatMasker, not aligned, containing gaps or non-ACGT characters, or (for hominds) contained in a CpG dinucleotide were not considered. Furthermore, to reduce the effect of sequencing errors, in all analyses involving polymorphism data, only nonsingletons, for example, nucleotides observed in more than one individual, were used. To remove the possible effect of complex mutational events on correlations between neighboring SNPs, we excluded those pairs of sites that contained a pair of derived SNPs in the same genotype from all analyses of neighboring SNPs within a single species.

Analysis
For each nucleotide site, allele replacements and ancestral and derived states of SNPs were inferred using maximum parsimony, as shown in figure 1. For each analysis, only sites that matched the phylogenetic pattern of black versus red in the corresponding panel of figure 1 were used as the background set, and those that matched the pattern of black versus red versus blue were used as the analyzed set, as detailed below. In particular, for a more reliable inference of the ancestral state in polymorphism data, we only considered such SNPs in which one of the alleles matched two or more agreeing outgroup species; this allele was considered ancestral.

Analysis of the Average Mutation Rate
Let us first consider the unweighted average mutation rate, defined as the arithmetic mean of the frequencies of the 12 possible kinds of single-nucleotide substitutions. We measured differences between the frequencies of substitutions at sites where another independent substitution was observed and at sites where no other substitutions were observed. Because substitutions of different kinds occur at different rates, the ratio of these two rates was calculated separately for each of the 12 kinds of mutations, and the average values of these ratios were used.

Specifically, for each pair of different nucleotides x and y, the frequency of a SNP or a replacement involving the ancestral nucleotide x and the single derived nucleotide y was obtained as

\[ f(x,y) = N_x(x,y)/N_a(x) \]

Here, \( x, y \in \{A, C, G, T\} \), \( x \neq y \); \( N_a(x) \) is the number of ancestral sites occupied by the nucleotide x, and \( N_x(x,y) \) is the number of sites at which an \( x > y \) SNP or replacement is observed, according to the pattern in figure 1 (black to red).

In the analyses II–IV, we similarly estimate the frequency of a SNP or replacement that occurred in another phylogenetic branch (black to blue):

\[ f_{ec}(x,y) = N_{ec}(x,y)/N_a(x) \]

where \( N_{ec}(x,y) \) is the number of sites at which an \( x > y \) SNP or replacement is observed, according to the pattern in figure 1 (black to blue). In analysis I, \( f_{ec}(x,y) = f(x,y) \).

Average mutation rate for the case of a single derived nucleotide in tables 1 and 2 were obtained as the arithmetic mean of \( f \) over the 12 possible mutation types:

\[ \bar{f_s} = \left( \sum_x \sum_y f(x,y)/12 \right) \]

For each pair of different derived nucleotides y and z (\( z \in \{A,C,G,T\}, z \neq x, z \neq y \)), the frequency of the observed SNPs
or replacements involving these two derived nucleotides was defined as

\[ f_m(x, y, z) = \frac{N_m(x, y, z)}{N_a(x)}, \]

where \( N_m(x, y, z) \) is the number of sites at which the \( x > y \) (blue) and \( x > z \) (red) SNPs or replacements are observed, according to the pattern in Figure 1. The average frequencies of mutation rates for the cases of two derived nucleotides in Tables 1 and 2 were defined as the arithmetic mean of \( f_m \) across the 24 possible patterns of double mutations:

\[ \bar{f}_m = \left( \frac{\sum_x \sum_y \sum_z f_m(x, y, z)}{24} \right). \]

In analysis I, we cannot distinguish between the two possible successions of mutations, and assume \( f_m(x, y, z) = f_m(x, z, y) \). The expected number of double derived alleles was calculated, for each ancestral nucleotide \( x \) and derived nucleotides \( y \) and \( z \), from the frequencies of single derived alleles

\[ e_m(x, y, z) = f_s(x, y) f_s(x, z) N_a(x). \]

Finally, the observed to expected ratio was calculated as the mean of such ratios for all types of double differences:

\[ r_m = \left( \frac{\sum_x \sum_y \sum_z N_m(x, y, z)}{e_m(x, y, z)} \right) / 24. \]

Table 1. Counts and Frequencies (in parentheses) of Patterns Corresponding to the Presence of One or Two Derived Nucleotides at the Same Site, and the Expectations for the Counts and Frequencies of the Presence of Two Derived Nucleotides, for Each of the Four Analyses Performed on Drosophila.

<table>
<thead>
<tr>
<th>Type of Noncoding Region</th>
<th>Single Derived Nucleotide</th>
<th>Two Derived Nucleotides at a Site, Observed</th>
<th>All Sites with a Single Derived Nucleotide (%)</th>
<th>Two Derived Nucleotides at a Site, Expected</th>
<th>Observed/Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triallelic SNPs (analysis I)</td>
<td>Introns 392,688 (6.91 × 10⁻³)</td>
<td>7,315 (1.24 × 10⁻⁴)</td>
<td>1.86</td>
<td>2,525 (4.12 × 10⁻⁵)</td>
<td>3.48</td>
</tr>
<tr>
<td>Coincident SNPs (analysis II)</td>
<td>Introns 523,130 (5.72 × 10⁻⁴)</td>
<td>9,725 (1.00 × 10⁻⁵)</td>
<td>1.86</td>
<td>3,268 (3.23 × 10⁻⁶)</td>
<td>3.59</td>
</tr>
<tr>
<td>Polymorphisms at sites of replacement (analysis III)</td>
<td>Introns 392,688 (6.91 × 10⁻³)</td>
<td>2,743 (4.59 × 10⁻⁴)</td>
<td>0.70</td>
<td>1,223 (2.07 × 10⁻⁵)</td>
<td>2.28</td>
</tr>
<tr>
<td>Replacement at sites of replacement (analysis IV)</td>
<td>Intergenic regions 523,130 (5.27 × 10⁻³)</td>
<td>3,486 (3.59 × 10⁻⁴)</td>
<td>0.67</td>
<td>1,502 (1.65 × 10⁻⁵)</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Table 2. Counts and Frequencies (in parentheses) of Patterns Corresponding to the Presence of One or Two Derived Nucleotides at a Site, and the Expectations for the Counts and Frequencies of the Presence of Two Derived Nucleotides, for Each of the Four Analyses Performed on Hominidae.

<table>
<thead>
<tr>
<th>Type of Noncoding Region</th>
<th>Single Derived Nucleotide</th>
<th>Two Derived Nucleotides at a Site, Observed</th>
<th>All Sites with a Single Derived Nucleotide (%)</th>
<th>Two Derived Nucleotides at a Site, Expected</th>
<th>Multiple Mutation Rate/Single-Mutation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triallelic SNP (analysis I)</td>
<td>Introns 981,033 (5.82 × 10⁻⁴)</td>
<td>819 (4.92 × 10⁻⁵)</td>
<td>0.083</td>
<td>442 (2.73 × 10⁻⁶)</td>
<td>2.55</td>
</tr>
<tr>
<td>Polymorphism at sites of replacement (analysis III)</td>
<td>Introns 1,321,296 (6.93 × 10⁻⁴)</td>
<td>1,113 (5.88 × 10⁻⁵)</td>
<td>0.084</td>
<td>703 (3.97 × 10⁻⁶)</td>
<td>2.21</td>
</tr>
<tr>
<td>Replacement at sites of replacement (analysis IV)</td>
<td>Intergenic regions 575,893 (3.84 × 10⁻⁴)</td>
<td>2,403 (8.13 × 10⁻⁵)</td>
<td>0.42</td>
<td>1,442 (5.03 × 10⁻⁶)</td>
<td>1.83</td>
</tr>
<tr>
<td>Triallelic SNP (analysis I)</td>
<td>Introns 720,188 (4.28 × 10⁻⁴)</td>
<td>3,139 (9.66 × 10⁻⁵)</td>
<td>0.44</td>
<td>1,854 (5.86 × 10⁻⁷)</td>
<td>1.85</td>
</tr>
<tr>
<td>Polymorphism at sites of replacement (analysis III)</td>
<td>Introns 1,582,163 (1.52 × 10⁻⁵)</td>
<td>8,159 (3.98 × 10⁻⁶)</td>
<td>0.52</td>
<td>5,591 (2.83 × 10⁻⁷)</td>
<td>1.76</td>
</tr>
<tr>
<td>Replacement at sites of replacement (analysis IV)</td>
<td>Intergenic regions 1,873,583 (1.67 × 10⁻⁵)</td>
<td>10,880 (4.90 × 10⁻⁶)</td>
<td>0.58</td>
<td>7,239 (3.41 × 10⁻⁷)</td>
<td>1.83</td>
</tr>
</tbody>
</table>

Note.—Frequencies have been calculated as arithmetic means for all kinds of nucleotide substitutions. For all analyses, the difference between the observed and the expected frequencies of double mutations was highly significant (\( p < 10^{-100} \), chi-square test).
To investigate the effects of the two immediately adjacent nucleotides (−1 and +1) on the mutation rate, we conducted a similar analysis, measuring \( N_m \) and \( e_m \) separately for each of the 16 combinations of the −1 and +1 nucleotides; the final value of \( r_m \) was then obtained by averaging over the \( 24 \times 4 \times 4 = 384 \) values. The effect of adjacent nucleotides was estimated in analysis IV (fig. 1), which provides the most data.

For analyses of SNPs or replacements at neighboring sites, additional filtering was applied: an alignment segment of length 80 was analyzed if the reconstructed ancestral state, for the given analysis, was a nucleotide (A, C, G, or T) in at least 70 of these sites. This was done to avoid the possible biases associated with higher SNP or replacement frequencies in regions with low local alignment quality. Calculations were done separately for each offset \( k \in \{-30 \ldots -1, 1 \ldots 30\} \). Here, we did not discriminate between the mutation types to simplify presentation. Frequency \( f_n(k) \) of SNPs or replacements (red) at a site located at offset \( k \) from site of a SNP or a replacement (blue), according to the pattern in figure 1, was obtained as

\[
f_n(k) = \left( \sum_x \sum_y N_n(x, y; k) / \sum_x \sum_y N_n(x, y) \right).
\]

Here, \( N_n(x, y; k) \) is the number of \( x > y \) SNPs or replacements, according to the pattern in figure 1 (black to red), observed at offset \( k \) nucleotides from some SNP or replacement that occurred according to the pattern in figure 1 (black to blue). The expected frequencies of SNPs or replacements were defined as their average frequency in a window of 200 nucleotides, centered at the site of a black to blue SNP or replacement; this approach was used to reduce the possible biases associated with large-scale variation in the mutation rates. Exact 95% confidence intervals for binomial proportions were calculated using the Clopper–Pearson method (Clopper and Pearson 1934).

**Analysis of the Transition–Transversion Ratio**

For a single-site analysis, we calculated the ratio of the number of transitions to the number of transversions separately for each of the two possible pairs of mutations for each ancestral nucleotide \( x \):

\[
\kappa(x, w : w \approx x) = N_t(x, y : y \sim x) / N_t(x, z : z \approx x).
\]

Here, \( w \in \{A, C, G, T\} \), \( w \neq x \), \( w \neq y \), \( w \neq z \), the sign \( \approx \) denotes a pair of nucleotides separated by a transition, and the sign \( \sim \) denotes a pair of nucleotides separated by a transition. The mean values of \( \kappa_{tv} \) were obtained as the arithmetic mean over all eight pairs of mutations:

\[
\bar{\kappa}_{tv} = \left( \sum_x \sum_{y \neq x} \kappa_{tv}(x, y) / 8 \right).
\]

The value \( \kappa_{tv} \) of the transition–transversion ratio conditional on another transversion at the same site was obtained as

\[
\kappa_{tv}(x, w : w \approx x) = N_m(x, w, y : y \sim x) / N_m(x, w, z : z \approx x).
\]

The mean values of \( \kappa_{tv} \) were obtained as

\[
\bar{\kappa}_{tv} = \left( \sum_x \sum_{y \neq x} \kappa_{tv}(x, y) / 8 \right).
\]

If the transition–transversion ratio at a site is independent on the presence of another transversion at this site, we expect \( \kappa_{tv}(x, y) = \kappa_{tv}(x) \). We calculated \( \kappa_{tv}(x, y) \) to \( \kappa_{tv}(x, y) \) ratio for each pair of mutations:

\[
r_s(x, y : y \approx x) = \kappa_{tv}(x, y : y \approx x) / \kappa_{tv}(x, y : y \approx x).
\]

Finally, the mean value of \( r_s \) was obtained as the arithmetic mean for all possible pairs of mutations:

\[
r_s = \left( \sum_{x, y \neq x} r(x, y) / 8 \right).
\]

To investigate the effects of the two immediately adjacent nucleotides (−1 and +1) on \( r_s \), we conducted a similar analysis, measuring \( r(x,y) \) separately for each of the 16 combinations of the −1 and +1 nucleotides; the final value of \( r_s \) was then obtained by averaging over the \( 8 \times 4 \times 4 = 64 \) values. The effect of adjacent nucleotides was estimated in analysis IV (fig. 1), which provides the most data.

For analysis of nearby sites, we did not discriminate between types of transitions or transversions. The transition/transversion ratio was obtained by counting all transitions and transversions that occurred according to the pattern in figure 1 (black to red), in both directions of some SNP or replacement that occurred according to the pattern in figure 1 (black to blue), in 30 nonoverlapping windows of 5 nucleotides each:

\[
k(p) = 2 \left( \sum x \sum_{y \neq x} \sum_{i = 0}^{4} N_t(x, y; 5p - \left( \frac{p}{|p|} \right)^i) / \sum x \sum_{y \neq x} \sum_{i = 0}^{4} N_t(x, y; 5p - \left( \frac{p}{|p|} \right)^i) \right).
\]

where \( p \in [-30 \ldots -1, 1 \ldots 30] \) is the number of the considered window, and multiplication by two is due to the fact that there are two possible transversions per each transition. The cases when the focal SNPs or replacements (blue) occurred by transitions, and the cases when they occurred by transversions, were considered separately. The expected transition/transversion ratio was obtained by counting all transitions and transversions that occurred according to the pattern in figure 1 (black to red) along all the genome. 95% confidence intervals on \( k(p) \) values were obtained by bootstrap.

In analyses of allele frequency spectra (AFS), only such nucleotide sites where used that at least 30 of 37 (in *D. melanogaster*) or 16 of 18 (for *H. sapiens*) genotypes carried a valid (A, C, G, or T) nucleotide. The frequency of the minor allele was measures among these 30 (16) genotypes; if between 31 and 37 (between 17 and 18) valid genotypes were present, random 30 (16) of them were chosen. Polymorphic sites were excluded in this analysis.

**Results**

**Using Polymorphism and Divergence Data to Study Variation of Mutation Rate**

First, we examine the local variation of the mutation rate in *Drosophila* and extend the earlier analyses (Hodgkinson et al. 1996).
et al. 2009; Hodgkinson and Eyre-Walker 2010; Johnson and Hellmann 2011) on the variation of the mutation rate in Hominiidae. With this aim, we used four analyses, considering sites with SNPs observed within a single species (analysis I, fig. 1A and A'), SNPs in different species (analysis II, fig. 1B and B'), SNPs in one species and replacement in another species (analysis III, fig. 1C and C'), and replacements in two different species (analysis IV, fig. 1D and D').

These four analyses are based on a common idea. Each time, we compare two values: first, the genome-wide density of derived alleles that segregate within a population or are fixed replacements between species (red in fig. 1); and second, the same value obtained at sites, or in the vicinity of sites, where another point mutation is known to have occurred, as revealed by the polymorphism in the same or a different species, or divergence of a different species (blue in fig. 1). The former value is used as a proxy for the genome-wide mutation rate; the latter value is used as a proxy for the mutation rate at a site, or in the vicinity of a site, with a potentially altered mutation rate. The comparison of the two values reveals the degree of correlation between the two mutations, and by inference, the variation in the mutation rate across the genome. As detailed below, all four analyses revealed that a mutation increases the probability of another mutation being observed at the same site and (in Drosophila) at nearby sites, confirming the earlier results (Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2010; Johnson and Hellmann 2011) obtained for Hominiidae and extending them for Drosophila.

There are two limitations to this approach. First, it is not suited to studying correlations between separate occurrences of the same substitution (e.g., A > T) at the same site (hereafter, we treat orthologous sites in different species as “the same” or “single” site). In analysis I, such occurrences will result in an SNP with two derived alleles identical by state (e.g., with ancestral allele A and two independently derived alleles T), which are essentially impossible to distinguish from an SNP with a single derived allele (with alleles A and T). In analysis II, the same biallelic polymorphism in D. melanogaster and D. simulans can appear due to incomplete lineage sorting in the ancestral population, rather than from two independent mutations (Begun et al. 2007). Even for human and chimpanzee, where shared polymorphisms are rare (Nachman and Crowell 2000; Hodgkinson et al. 2009), Hodgkinson et al. (2009) and Johnson and Hellmann (2011) observed a much stronger correlation between mutations in the two species in the same site when the derived alleles were the same, indicating that a high prevalence of parallel polymorphism is due to strong forces, mutational or otherwise, different from those responsible for other variation in the mutation rate. With analyses III and IV, two identical derived alleles at a site are also strongly overrepresented (Bazylkin et al. 2007; Johnson and Hellmann 2011), again suggesting the action of these strong forces. We reasoned that this effect should be studied separately, and therefore excluded patterns which may be due to repeated occurrences of the same substitution at a site from our analyses.

The second limitation appears if we try to apply analysis I to nearby sites. In this case, the same pattern could result from either non-independence of single-nucleotide substitutions at such sites or a complex mutational event affecting two sites simultaneously (Silva and Kondrashov 2002; Hodgkinson and Eyre-Walker 2010; McDonald et al. 2011; Schrider et al. 2011). Indeed, around 3% of nucleotide substitutions in both Drosophila and Hominiidae appear as parts of complex mutational events (Schrider et al. 2011). To deal with this limitation, we excluded those pairs of sites that contain a pair of derived SNPs in the same genotype. This same approach was taken for the case and the control and therefore should not bias the results.

**Variation in the Drosophila Mutation Rate**

Analysis I, when applied to a single site, deals with situations when, within a species, an ancestral and two different derived nucleotides coexist at the same site (a triallelic SNP). The frequency of triallelic SNPs in D. melanogaster is 3.5 times higher than that expected on the basis of the frequencies of biallelic SNPs and the assumption that the two mutations at a site are independent (table 1).

When applied to nearby sites, analysis I considers correlations in the occurrence of biallelic SNPs close to each other. The frequency of SNPs in D. melanogaster is substantially elevated within a DNA segment up to ~15 nucleotides surrounding the site that harbors another SNP (fig. 2A).

Analysis II, when applied to a single site, deals with situations when biallelic SNPs occur at the same site in two species. An SNP in D. simulans increases the probability of a coincident SNP in D. melanogaster by a factor of 2.3 (table 1).

For nearby sites, analysis II considers situations when a biallelic SNP in D. melanogaster occurs close to a site of a biallelic SNP in D. simulans. We observed a slight, although statistically significant, increase in the frequency of D. melanogaster SNPs at sites located up to ~15 nucleotides away from site of a D. simulans SNP (fig. 2B).

Analysis III, when applied to a single site, deals with situations when a biallelic SNP in a lineage of one species occurs at a site of replacement in another lineage. A replacement between D. yakuba and D. erecta increases the probability of a SNP at the same site in D. melanogaster by a factor of ~2 (table 1).

For nearby sites, analysis III considers situations when a SNP in D. melanogaster occurs close to the site of D. yakuba–D. erecta replacement. Frequency of SNPs is elevated at ~10 nucleotides surrounding the site of D. yakuba–D. erecta replacement (fig. 2C).

Analysis IV, when applied to a single site, considers co-occurring nucleotide replacements at the same site in lineages of different species. Nucleotide replacement between D. yakuba and D. erecta increases the probability of a nucleotide replacement at the same site between D. melanogaster and D. simulans by a factor of ~2 (table 1) similar to the pattern previously observed for synonymous sites (Bazylkin et al. 2007). Results were very similar if the effect of the two nearest adjacent nucleotides was
Fig. 2. Density of SNPs or nucleotide replacements as a function of distance from the site of SNP or nucleotide replacement in intergenic regions in Drosophila, shown for each of the four analyses in figure 1. Positive values on the horizontal axis correspond to positions 3’ of the considered nucleotide, and negative values correspond to positions 5’ of the considered nucleotide. Black solid lines show the density of SNPs (A–C) or of interspecies differences (D) near the sites that harbor other SNPs (A,B) or interspecies differences (C,D); gray dashed lines show the averages of the corresponding densities over a window of 100 nucleotides around the same sites. Error bars are 95% binomial confidence intervals; in panels A, C, and D, error bars are so small that they would be barely visible.

controlled for (supplementary table S1, Supplementary Material online).

For nearby sites, analysis IV considers situations when D. melanogaster–D. simulans replacement occurs close to a site of D. yakuba–D. erecta replacement. Fig. 2D shows that the frequency of nucleotide replacements within these two species pairs is correlated up to distances of ~20 nucleotides.

Variation in the Hominid Mutation Rate

In Hominidae, although the topology of the phylogenetic tree of the available species is different, the same four analyses can be performed (fig. 1). However, as no reliable genome-scale data set for the data on chimpanzee polymorphism exists, we were unable to perform analysis II requiring these data. The previous analyses of local variation of the mutation rate in hominids correspond to our analyses I (Hodgkinson and Eyre-Walker 2010) and II (Hodgkinson et al. 2009) (fig. 1A’ and B’). Analysis IV corresponds to the study by Johnson and Hellmann (2011), except that for assessing divergence they used orangutan, and we use a more closely related gorilla species.

For all analysis performed in Hominidae, we exclude CpG dinucleotides because of their hypermutability (Hwang and Green 2004). This removal did not lead to an artifactual decrease in the mutation rate at sites adjacent to mutation (Hodgkinson et al. 2009; Johnson and Hellmann 2011) because our approach guarantees that the two derived alleles occurred as independent events and therefore the second allele never originated at the background of the first one.

Analysis I, when applied to a single site, revealed that the frequency of triallelic SNPs in human is ~2.3 times higher than expected (table 2), consistent with the previous observations (Hodgkinson and Eyre-Walker 2010).

Analysis I for nearby sites revealed that in human, at sites immediately adjacent to a site harboring a biallelic SNP, the frequency of other biallelic SNPs is elevated by a factor of 1.6 (fig. 3A); however, this increase does not extend to the sites further away.

Analysis III, when applied to a single site, revealed that a nucleotide replacement at a site in gorilla lineage increases the probability of a SNP in human by a factor of ~1.3 (table 2).

For nearby sites, analysis III deals with situations when a SNP in human occurs close to a site of a replacement in gorilla lineage. Frequency of human SNPs is slightly elevated for the two sites immediately adjacent to the site of replacement in gorilla lineage, but not further away (fig. 3B).

Analysis IV, when applied to a single site, shows that a nucleotide replacement in gorilla lineage is associated with an increase of the probability of replacement in human lineage by a factor of 1.83 (table 2). As was the case in Drosophila, the results were very similar if the effect of the two nearest adjacent nucleotides was controlled for (supplementary table S2, Supplementary Material online).

For nearby sites, analysis IV considers situations when a replacement in human lineage occurred close to a site of replacement in gorilla lineage. Figure 3C shows that the occurrence of nucleotide replacements within these two species pairs is slightly correlated for the two immediately adjacent sites, but not beyond.

Using Polymorphism and Divergence Data to Study Variation in the Transition/Transversion Ratio

The overall mutation rate notwithstanding, the rates of mutations of different kinds can experience local heterogeneity. We investigate the heterogeneity of transition/transversion ratio \( \kappa \). We concentrate on the dependence of \( \kappa \) on the mutation events at the same or nearby sites, using the same logic as for the analysis of heterogeneity of the mutation rates: biased, compared with the genome average, values of \( \kappa \) for mutations at or near sites known to have also undergone a specific kind of a mutation (a transversion or transition) implies variation of \( \kappa \) between sites. The same four analyses outlined in figure 1 are used for this purpose. As their application to a single site is not intuitive, we describe it here in detail.

Consider a site with a known ancestral nucleotide (e.g., A). At such a site, there are always two possible transversions (A \( \rightarrow \) C and A \( \rightarrow \) T) and one possible transition (A \( \rightarrow \) G; as in the study of the overall mutation rate, we exclude cases of independent (parallel) occurrence of the same derived nucleotide). In analysis I, when the ancestral state is known, only two kinds of double mutation events are possible: 1) with the two derived nucleotides produced by two different transversions and 2) with one derived nucleotide
produced by a transversion and the other by a transition. Among the sites at which one of the derived alleles is a pre-defined transversion (e.g., A > C), the ratio of the rate of the transition (A > G) and the remaining transversion (A > T) $k_{tv}$ can be compared with $k_{tv}$, which is the ratio of the rates of the same two SNPs or replacements (A > G and A > T) at sites which harbor exactly one mutation; significant difference between these two values implies that the transition–transversion ratio is biased at sites of transversions. Conversely, $k_{tv}$ cannot be conditioned on the rate of transition. Therefore, data on two derived nucleotides co-occurring at a site allow us to study how $k_{tv}$ depends on the occurrence of a transversion, but not of a transition, at the same site (tables 3 and 4, supplementary tables S1–S5, Supplementary Material online). This limitation does not, however, apply when we study how $k$ depends on what happens at a different site; therefore, the dependence of $k$ on both transversions and transitions at nearby sites was studied. For analysis of neighboring sites, $k$ was defined, for a nucleotide site, as the rate of the only possible transition over the average rate of the two possible transversions.

### Variation of the Transition/Transversion Ratio in Drosophila

Data on $k_{tv}$ obtained by analysis I for sites that harbor a transversion in D. melanogaster are presented in table 3, separately for all the eight possible pairs of transitions and transversions. The occurrence of a transversion at a site is associated with a decreased transition/transversion ratio, so that $r_k$, that is, the ratio of $k_{tv}$ and $k_{tv}$ over all pairs of transitions and transversions, equals 0.76 ± 0.02 (hereafter, the means and the 95% confidence intervals obtained with bootstrapping are presented for point estimates).

Analysis I for nearby sites shows that $k$ is reduced near a site of a transversion ($P < 10^{-92}$ for the nearest five nucleotides, Fisher’s exact test) but not near a site of a transition ($P = 0.89$, Fisher’s exact test), compared with the genome average. The reduction in $k$ spans distances of up to ~30 nucleotides around the site of transversion (fig. 4A). A similar pattern was obtained when segments within 10 nucleotides from the nearest indel were excluded (supplementary fig 1, Supplementary Material online).

Analysis II, when applied to a single site, shows that at sites of a transversion SNP in D. simulans, the observed transition/transversion ratio for SNPs segregating in D. melanogaster is lower than the genome average, so that $r_k = 0.91 ± 0.06$ (supplementary table S3, Supplementary Material online). The results of the reciprocal analysis, when we compare the transition/transversion ratio in D. simulans with that at sites of a transversion in D. melanogaster, were similar (data not shown).

Analysis II for nearby sites revealed that $k$ is reduced at distances of ~100 nucleotides, compared with the genome average, both around the sites of a transversion ($P < 10^{-20}$ for the nearest 100 nucleotides, Fisher’s exact test) and around the sites of a transition ($P < 10^{-8}$, Fisher’s exact test); nevertheless, this reduction has a larger magnitude.

### Fig. 3.

Density of SNPs or nucleotide replacements as a function of distance from the site of SNP or nucleotide replacement in intergenic regions in Hominidae. Positive values on the horizontal axis correspond to positions 3’ of the considered nucleotide, whereas negative values correspond to positions 5’ of the considered nucleotide. Four panels correspond to the four analyses in figure 1. Black solid lines show the density of SNPs (A,B) or of interspecies differences (C) around sites that harbor SNPs (A) or interspecies differences (B,C); gray dashed lines show the corresponding densities within a window of 100 nucleotides around corresponding sites. Error bars are 95% binomial confidence intervals.
Table 3. Ratio of the Rates of Transitions and Transversions at Sites of Biallelic and Triallelic SNPs (analysis I) in Drosophila.

<table>
<thead>
<tr>
<th>Transition and Transversion Used to Calculate $\kappa_{tv}$</th>
<th>Additional Transversion Required in Calculation of $\kappa_{tv}$</th>
<th>Intron</th>
<th>Intergenic Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A &gt; G/A &gt; T</td>
<td>A &gt; C</td>
<td>1.02</td>
<td>0.78</td>
</tr>
<tr>
<td>A &gt; G/A &gt; C</td>
<td>A &gt; T</td>
<td>1.88</td>
<td>1.76</td>
</tr>
<tr>
<td>G &gt; A/G &gt; T</td>
<td>G &gt; C</td>
<td>1.84</td>
<td>1.45</td>
</tr>
<tr>
<td>G &gt; A/G &gt; C</td>
<td>G &gt; T</td>
<td>3.52</td>
<td>2.16</td>
</tr>
<tr>
<td>C &gt; T/C &gt; G</td>
<td>C &gt; A</td>
<td>3.50</td>
<td>2.17</td>
</tr>
<tr>
<td>C &gt; T/C &gt; A</td>
<td>C &gt; G</td>
<td>1.85</td>
<td>1.26</td>
</tr>
<tr>
<td>T &gt; C/T &gt; G</td>
<td>T &gt; A</td>
<td>1.85</td>
<td>1.50</td>
</tr>
<tr>
<td>T &gt; C/T &gt; A</td>
<td>T &gt; G</td>
<td>1.01</td>
<td>0.81</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>2.06</td>
<td>1.49</td>
</tr>
</tbody>
</table>

**Notes.** — $\kappa_{tv}$ transition/transversion ratio calculated from sites of biallelic SNPs; $\kappa_{tv}$, transition/transversion ratio calculated from sites of triallelic SNPs harboring an additional transversion.

* *, **, *** denote $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively, as determined by chi-square test for the differences between $\kappa_{tv}$ and $\kappa_{tv}$, in each of the eight comparisons.

Analysis IV for nearby sites revealed that $\kappa$ is reduced, compared with the genome average, both near the sites of a transversion ($P < 10^{-48}$ for the nearest five nucleotides, Fisher’s exact test) and a transition ($P < 10^{-14}$, Fisher’s exact test). As in analyses I–III, this reduction is larger around the sites of a transversion ($P < 10^{-6}$, Fisher’s exact test) (fig. 4D).

Analysis I for nearby sites shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.

Analysis IV, when applied to a single site, shows that $\kappa$ is reduced at distances of $\sim 100$ nucleotides both near the sites of transversions ($P < 10^{-50}$ for the nearest 100 nucleotides, Fisher’s exact test) and transitions ($P < 10^{-4}$, Fisher’s exact test). However, this reduction is much stronger near the sites of transversions ($P < 10^{-18}$, Fisher’s exact test) (fig. 4C). The lowest value of $\kappa$ is observed not for the sites immediately adjacent to the site of the mutation, but $20–30$ nucleotides away, possibly reflecting the presence of a short-scale ($\sim 10$ nucleotides) increase in $\kappa$ close to a site of interspecies replacement and an intermediate-scale effect that decreases $\kappa$ near such sites.
analyzed the genome-wide AFS separately for minor alleles that originated through a transition and a transversion. In *Drosophila*, the AFS of transversions was biased toward low frequencies, compared with transitions (fig. 6); the value of Tajima's *D* ([Tajima 1989](#)) for transversions (/C0 1.51 for introns and /C0 1.54 for intergenic regions) was lower than for transitions (/C0 0.80 for introns and /C0 0.84 for intergenic regions), suggesting genome-wide action of weak negative selection against transversions.

**Variation of the Transition/Transversion Ratio in Hominidae**

Analysis I, when applied to a single site, for human triallelic SNPs shows that a transversion decreases the observed transition/transversion ratio at the same site, so that /Rk=0.31±0.03 (table 3).

Analysis III, when applied to a single site, shows that a transversion that occurred in the gorilla lineage decreases the observed transition/transversion ratio for the SNPs segregating at the same position in human, so that /Rk=−0.91±0.10 (supplementary table S7, Supplementary Material online).

Analysis IV, when applied to a single site, shows that a transversion in the gorilla lineage decreases the observed transition/transversion ratio for replacements at the same site in the human lineage, so that /Rk=−0.58±0.03 (supplementary table S8, Supplementary Material online). Controlling for the context effects changed this value only marginally: /Rk=0.67±0.15 if the effect of the two nearest adjacent nucleotides was accounted for (supplementary table S9, Supplementary Material online).

For nearby sites, analysis I shows that /k is reduced, compared with the genome average, within the five nucleotides adjacent to the site of transversions (P = 8.1 × 10−3 for the nearest five nucleotides, Fisher's exact test) but not of transitions (P = 0.12); the difference between these two values was insignificant (P = 0.29) (fig. 7A), and we observed no

**Fig. 4.** Transition/transversion ratio /k as a function of distance from the site of transversion (magenta) or transition (green) in *Drosophila*. Positive values on the horizontal axis correspond to positions 3′ of the considered nucleotide, whereas negative values correspond to positions 5′ of the considered nucleotide. Four panels correspond to the four analyses in figure 1. Gray dashed line shows the genome average /k. Average values of /k for non-overlapping windows of length 5 are presented. Error bars are 95% confidence intervals obtained by nonparametric bootstrapping.

For nearby sites, analysis I shows that /k is reduced, compared with the genome average, within the five nucleotides adjacent to the site of transversions (P = 8.1 × 10−3 for the nearest five nucleotides, Fisher's exact test) but not of transitions (P = 0.12); the difference between these two values was insignificant (P = 0.29) (fig. 7A), and we observed no

**Fig. 5.** Transition/transversion ratio (/k) for different numbers of SNP in windows spanning different distances from the site of transversion or transition in *Drosophila* intergenic regions. The number of SNPs and the mean /k values for these SNPs were measured within a window centered at the site of transversion (black) or transition (gray). Each window had a total length of 60 nucleotides and spanned the distances 1–30 nucleotides, 31–60 nucleotides, or 61–90 nucleotides from the central site in both directions. The numbers on the horizontal axis correspond to the number of SNPs or interspecies differences observed in a given window. *, **, *** denote P < 0.05, P < 0.01, and P < 0.001, respectively, as determined by the Fisher’s exact test for the difference between /k near the site of a transition and /k near the site of a transversion.
differences between the observed- and the genome-average $k$ for higher distances. In analyses III and IV, no differences of $k$ near the sites of either transition or transversions from the genome average was observed (fig. 7B and C).

In human, the AFSs for minor transversions and transitions were indistinguishable (fig. 8), suggesting that their genome-wide occurrence is only marginally, if at all, affected by selection. The values of Tajima’s D (Tajima 1989) were also similar (transversions, $-0.46$ for introns, $-0.42$ for intergenic regions; transitions, $-0.43$ for introns, $-0.39$ for intergenic regions).

Discussion

Variation in the Mutation Rate

By applying four comparative analyses to Drosophila and Hominidae, we observed positive correlations of occurrences of SNPs and of interspecies differences at the same site, previously reported for comparison of human and chimpanzee SNPs (Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2010), for divergence at noncoding sites between human and orangutan (Johnson and Hellmann 2011) and for divergence at synonymous coding sites in Drosophila and in mammals (Bazykin et al. 2007). Furthermore, we observe correlations of occurrences of SNPs and of interspecies differences at nearby sites in Drosophila, at distances up to $\sim 20$ nucleotides. In line with previous findings (Silva and Kondrashov 2002; Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2010; Johnson and Hellmann 2011), no correlation at nearby sites was observed in hominids, except perhaps in the immediately adjacent nucleotides.

Our four analyses, although different from each other in the logic and in the data used (fig. 1), produced consistent results both in Drosophila (fig. 2) and in Hominidae (fig. 3), thus making our conclusions robust. In particular, sequencing or genome assembly errors could only affect the results of the analyses I and II. Also, in all the analyses dealing with SNPs, we ignored singleton SNPs, which were encountered

![Fig. 6. Minor AFS for polymorphic transversions (black) and transitions (gray) in D. melanogaster. Horizontal axis represents the number of genotypes (of 30) carrying the minor allele. A, introns; B, intergenic regions.](image)

![Fig. 7. Transition/transversion ratio ($\kappa$) as a function of distance from the site of transversion (magenta) or transition (green) in Hominidae. Positive values on the horizontal axis correspond to positions 3’ of the considered nucleotide, whereas negative values correspond to positions 5’ of the considered nucleotide. Four panels correspond to the four analyses in figure 1. Gray dashed line shows the genome average $\kappa$. Average values of $\kappa$ for nonoverlapping windows of length 5 are presented. Error bars are 95% confidence intervals obtained by nonparametric bootstrapping.](image)
only once in our sample and thus could possibly be the artifacts of such errors.

The observed correlations could conceivably emerge due to small-scale heterogeneity of selection, if sites under negative selection, as well as selectively neutral sites, are distributed in the genome nonuniformly. However, selection is unlikely to be the main cause of the observed effects, for two reasons. First, we excluded coding exons and UTRs, which are under the strongest selection, from all our analyses. Still, in Drosophila, sites under selection are common even in introns and intergenic regions (Andolfatto 2005). However, in Hominidae, which produced in the same-site analyses results similar to Drosophila, most of the sites in introns and in intergenic regions are effectively neutral (Shabalina et al. 2001). Second, both in Hominidae and in Drosophila, the magnitude of the correlations we observed was the largest in analysis I, which deals only with polymorphic variants, although selection is not as efficient in removing slightly deleterious alleles as in preventing their fixation (McDonald and Kreitman 1991; Charlesworth and Eyre-Walker 2008). Nevertheless, the correlation between substitutions at adjacent sites is observed at distances up to ~20 nucleotides in Drosophila, but not in hominids, consistently with stronger selection in non-coding regions in Drosophila.

Thus, we conclude that correlations we observed are mostly due to local heterogeneity in the mutation rate. However, we cannot relate this heterogeneity to any simple contexts. First, the CpG dinucleotide hypermutable in humans was excluded from analysis. Second, while the mutation rates depend on the surrounding nucleotides, heterogeneity is not reduced after this nonuniformity is controlled for (supplementary tables S1 and S2, Supplementary Material online). Therefore, we regard this heterogeneity as cryptic, following Hodgkinson et al. (2009).

Pairs of single-nucleotide differences originating as a single complex mutational event may contribute to the observed correlations in analysis I. To prevent this complication from affecting our results, we excluded from analysis I individual genotypes carrying two derived SNPs (there was a substantial excess of such genotypes; data not reported). Triallelic SNPs might also originate as a single-mutational event (Hodgkinson and Eyre-Walker 2010), leading us to interpret one complex event as two correlated mutations in analysis I. Remarkably, when applied to a single site, analysis I reveals a stronger correlation than other analyses (tables 1 and 2), consistently with the hypothesis of single-mutation origin of triallelic SNPs (Hodgkinson and Eyre-Walker 2010).

Perhaps, the patterns of the cryptic variation of the mutation rate are subject to evolution and, thus, become more and more different in more and more distant genomes. Indeed, there is practically no correlation between the replacements that occurred on the path connecting two species of lemurs, Microcebus murinus and Otolemur garnetti, and the density of SNPs in humans (data not shown).

Variation in the Transversion/Transition Ratio

Both in Drosophila and in Hominidae, the transition/transversion ratio is substantially reduced at sites of a transversion (tables 3 and 4; supplementary tables S3–S5, S7, and S8, Supplementary Material online). Excess of parallel SNPs among transversions has been previously observed in Hominidae (Hodgkinson and Eyre-Walker 2010). Our results on nonparallel SNPs and substitutions suggest that this pattern is mainly due to local variation in the transition/transversion ratio.

In Drosophila, \( \kappa \) is substantially reduced at ~100 nucleotides near the sites of transversions (fig. 4). In analyses II–IV, it was also reduced near the sites of transitions, but the magnitude of the reduction was always larger near the sites of transversions than near the sites of transitions. For nearby sites in Hominidae, we observed a statistically significant difference only in analysis I, where \( \kappa \) was reduced at distances of up to five nucleotides from the site of a mutation, and in this case, the difference in \( \kappa \) close to a site of a transversion and a transition was statistically insignificant (fig. 7).

For the same reasons as with the differences in the mutation rate, the observed differences in \( \kappa \) cannot be easily explained solely by selection. Conceivably, negative selection acting outside the coding regions, for example, constraining the evolution of regulatory sequences, could favor transitions over transversions, since the latter is a more radical nucleotide substitution from the structural viewpoint. Such selection should lead to an upward bias in \( \kappa \), compared with that inherent in the mutation pattern. In Drosophila, two observations suggest that weak genomewide selection against transversions actually does play a role. First, higher genome-average \( \kappa \) in interspecific replacements (analysis IV) than in polymorphism (analyses I–III) is consistent with selection, since divergence is affected by selection to a larger extent than polymorphism.

Fig. 8. Minor AFS for transversions (black) and transitions (gray) in H. sapiens. Horizontal axis represents the number of genotypes (of 16) carrying the minor allele. A, introns; B, intergenic regions.
Second, the alleles that originated through a transversion are observed in population at lower frequencies, compared with those that originated through a transition (fig. 6), also suggesting negative selection against transversions. Since selection is negatively correlated with local SNP density, this phenomenon could lead to a negative correlation between the local SNP density and \( k \), consistent with our observations in Drosophila. Furthermore, it could lead to reduced \( k \) within and, as long as the selective constraint in nearby sites is correlated, around the sites of transversions.

However, the reduction in \( k \) around the sites of transversions is the strongest in analysis I dealing with polymorphism, and not in analysis IV dealing with divergence (fig. 4), although selection should have the strongest effect in the latter. This suggests that while the correlation between \( k \) and the overall mutation rate can have a selective cause, selection is unlikely to solely explain positive correlation of \( k \) in neighboring sites. Furthermore, among all the same-site analyses, \( k_{\text{trans}}/k_{\text{SNP}} \) is the lowest, indicating the strongest nonuniformity, in analysis I (table 4) in Hominidae. However, in Hominidae, we see no sign of selection against transversions: the genome-wide \( k \) are indistinguishable in polymorphism (fig. 7A) and in divergence (fig. 7B and C), and no effect of selection on allele frequencies is observed (fig. 8). This is to be expected: the noncoding regions in Hominidae are under low selection pressure, compared with Drosophila (Jareborg et al. 1999). Therefore, at least in Hominidae, and partly in Drosophila, the observed effect is due to nonuniformity of the mutation patterns, rather than selection.

Similar to the differences in the mutation rates, differences in \( k \) reflect patterns in the mutation patterns, which mostly cannot be attributed to any simple contexts (supplementary tables S6 and S9, Supplementary Material online). Thus, local heterogeneity of the mutation process affects not only the overall mutation rate but also the relative rates of mutations of different kinds.

Heterogeneity of \( k \) could be due to reduced \( k \) close to polymorphic insertions or deletions (indels), as described by Tian et al. (2008). Therefore, we repeated analysis I excluding the sites at distances below 10 nucleotides from polymorphic indels in Drosophila. The pattern for \( k \) did not change (supplementary fig. 2, Supplementary Material online), indicating that the observed nonuniformity is not associated with the indels. Mutations affecting several sites simultaneously (complex mutational events) could contribute to the nonuniformity of \( k \), and indeed, a stronger nonuniformity of \( k \) was observed when all polymorphism was analyzed (data not shown). However, we control for this possibility by excluding alleles cosegregating within a single individual, and therefore, complex mutational events do not cause the observed results.

The observed nonuniformity of \( k \) along the genome could be explainable by the heterogeneity of the mutation rate or could be an independent phenomenon. In the former case, \( k \) is expected to be correlated with the local mutation rate, while in the latter case, \( k \) should be correlated with the presence of other transversions in the vicinity. We detected both these effects (fig. 5): \( k \) is correlated with the local mutation rate, but when this rate is controlled for, it is different for positions in the vicinity of transversions and of transversions. Unfortunately, relaxed negative selection also increase rate of polymorphism and therefore may provoke the correlation between polymorphism density and \( k \). But as discussed before, the main effect in analysis I, we addressed to variation in mutation rate and \( k \). Therefore, while a fraction of the variation in \( k \) can be explained by the local mutation rates or selection, it also has an independent cryptic component of strong variation with the characteristic length of \( \sim 100 \) nucleotides.

In total, our results indicate that not only the overall mutation rate but also the peculiarities of the mutation process are characterized by strong heterogeneity along the genome. The qualitative aspects of this heterogeneity appear to be conserved at least across Metazoans, and it may be an underappreciated contributor to the patterns of variation and evolution.

Supplementary Material
Supplementary figures 1 and 2 and tables S1–S9 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
This work was supported by grant 11.G34.31.0008 from the Russian Ministry of Science and Education and by the Molecular and Cellular Biology Program of the Russian Academy of Sciences. Research performed at Moscow State University (Moscow, 119991, Russia) and Institute of Information Transmission Problems (Moscow, 127994, Russia).

References


