Prevalence of Multinucleotide Replacements in Evolution of Primates and *Drosophila*

Nadezhda V. Terekhanova,[1,2] Georgii A. Bazykin,[1,2] Alexey Neverov,[1,2] Alexey S. Kondrashov,[1,3] and Vladimir B. Seplyarskiy*,[1,2]

[1]Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

[2]Sector for Molecular Evolution, Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia

[3]Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan

***Corresponding author:** E-mail: pamjat@mail.ru.

Associate editor: Asger Hobolth

Abstract

Evolution of sequences mostly involves independent changes at different sites. However, substitutions at neighboring sites may co-occur as multinucleotide replacement events (MNRs). Here, we compare noncoding sequences of several species of primates, and of three species of *Drosophila* fruit flies, in a phylogenetic analysis of the replacements that occurred between species at nearby nucleotide sites. Both in primates and in *Drosophila*, the frequency of single-nucleotide replacements is substantially elevated within 10 nucleotides from other replacements that occurred on the same lineage but not on another lineage. The data imply that dinucleotide replacements (DNRs) affecting sites at distances of up to 10 nucleotides from each other are responsible for 2.3% of single-nucleotide replacements in primate genomes and for 5.6% in *Drosophila* genomes. Among these DNRs, 26% and 69%, respectively, are in fact parts of replacements of three or more trinucleotide replacements (TNRs). The plurality of MNRs affect nearby nucleotides, so that at least six times as many DNRs affect two adjacent nucleotide sites than sites 10 nucleotides apart. Still, approximately 60% of DNRs, and approximately 90% of TNRs, span distances more than two (or three) nucleotides. MNRs make a major contribution to the observed clustering of substitutions: In the human–chimpanzee comparison, DNRs are responsible for 50% of cases when two nearby replacements are observed on the human lineage, and TNRs are responsible for 83% of cases when three replacements at three immediately adjacent sites are observed on the human lineage. The prevalence of MNRs matches that is observed in data on de novo mutations and is also observed in the regions with the lowest sequence conservation, suggesting that MNRs mainly have mutational origin; however, epistatic selection and/or gene conversion may also play a role.

Key words: **multinucleotide replacements, complex mutations, mutagenesis, *D. melanogaster*, *H. sapiens*.**

Introduction

Evolution at all levels, including that of DNA sequences, primarily proceeds through small steps. Simple evolutionary models do not include complex events, assuming that different sites of diverging sequences accumulate nucleotide substitutions independently (e.g., Yang and Bielawski 2000). However, several factors can lead to violations of this assumption.

First, changes at different sites can be correlated due to complex mutations. A complex mutation consists of multiple changes in the DNA sequence, which appear not too far from each other as parts of a single event; each change may be a single-nucleotide mutation (multinucleotide mutations, MNMs), or more complex changes may be involved. As with other types of mutations, the frequency of occurrence of complex mutations can be most reliably estimated from data on de novo mutations, because other factors of evolution have had least chance to affect the fate of these mutations; the second best option is the mutation-accumulation experiments where selection has been relaxed (Kondrashov 2008). Analyses of de novo mutations (Kondrashov 2003; Chen et al. 2009) and of mutation-accumulation experiments (Keightley et al. 2009) reveal that complex mutations occur with a nonnegligible frequency. Patterns in within-population variation (Hodgkinson and Eyre-Walker 2010; Schrider et al. 2011) and in interspecies divergence (Averof et al. 2000; Smith et al. 2003) also demonstrate a substantial frequency of occurrence of correlated changes. Although other, nonmutational, explanations are possible for complex events observed in polymorphism and divergence data (see later), in each instance, arguments have been put forward that these observations are primarily due to complex mutational events (Averof et al. 2000; Smith et al. 2003; Hodgkinson and Eyre-Walker 2010; Schrider et al. 2011).

The differences in contamination by non-MNMs may be one reason why the estimates of the rates of MNMs vary widely. Early estimates based on the frequencies of transitions between the two serine codon families in metazoans, and of double-nucleotide substitutions in a pseudogene of primates, put the rate of MNMs at approximately 0.1 per site per billion years (Averof et al. 2000). In primates, the fraction of mutations that are involved in a MNM was thus estimated as approximately 2% (Averof et al. 2000); this estimate dropped to 0.3% after the local heterogeneity of the mutation rates was taken into account (Smith et al. 2003). The rate of double mutations was estimated as approximately 10^{-11} per site per generation from data on human de novo nonsense mutations (Kondrashov 2003). The frequency of MNMs among the de novo mutations in mouse was estimated at 0.2–1% (Wang et al. 2007). More recently, the frequencies of single nucleotide polymorphisms (SNPs) in human put this estimate at 0.9% for MNMs hitting the adjacent sites (Hodgkinson and Eyre-Walker 2010) or at 3% for MNMs spanning distances of up to 20 nucleotides (Schrider et al. 2011). To the best of our knowledge, the only estimate for the rate of the MNMs in *Drosophila* is based on the observation of four individual dinucleotide mutations among 174 mutations observed in a mutation-accumulation experiment (Keightley et al. 2009). Most analyses do not distinguish between dinucleotide, trinucleotide, and higher order co-occurring mutations (but see Wang et al. 2007).

Second, nonindependent evolution at different sites on one lineage can be caused by selection. A strong correlation of nonsynonymous substitutions, and a weaker effect in synonymous substitutions, was observed in evolution of rodent (Bazykin et al. 2004), HIV-1 (Bazykin et al. 2006), and *Drosophila* (Callahan et al. 2011; Bazykin and Kondrashov 2012) proteins; a similar effect was also observed for coding SNPs of *Ciona savignyi* (Donmez et al. 2009). The contrast between the nonsynonymous and synonymous substitutions or SNPs suggests that this effect is caused by selection. Runs of adjacent substitutions in coding sequences were also described by Stoletzki and Eyre-Walker (2011).

Third, evolution at nearby sites can be correlated due to nonallelic gene conversion (Chen et al. 2007). When a segment of DNA is converted by a paralogous sequence carrying a number of nucleotide differences, the resulting pattern may be reminiscent of multiple simultaneous substitutions (Kidd et al. 2010, fig. S3). A typical conversion tract spans DNA segments of more than 300 nucleotides (Chen et al. 2007; Mancera et al. 2008; Paigen and Petkov 2010).

Multinucleotide replacement events (MNRs) can be accounted for in evolutionary models. There are a lot of approaches, based on HMM (Yang 1995), maximum likelihood (ML; Whelan and Goldman 2004), Bayesian graphical models (Poon et al. 2008), and so on. In ML modeling of evolution of protein-coding sequence, including MNRs significantly improves the fit of the model (Whelan and Goldman 2004; Kosiol et al. 2007; Miyazawa 2011). The model implemented in PAML (Yang 2007) put the estimates for the rates of dinucleotide replacements (DNRs) and trinucleotide replacements (TNR) for aligned protein domains in Pandit database (Whelan et al. 2006) at 21% and 3.5%, respectively (Kosiol et al. 2007).

Here, we estimate the frequency of MNRs from data on divergence of noncoding (intronic and intergenic) sequences of primates and *Drosophila*. We show that the frequency of single-nucleotide substitutions is increased by the presence of a nearby substitution on the same lineage to a larger extent than by the presence of a nearby substitution on another lineage, implying a substantial frequency of MNRs.

Results

To estimate the fraction α of single-nucleotide replacements that occurred as part of an MNR, we use a phylogenetic approach (fig. 1). In each analysis, we consider a trio of species, including the "focal" species (*Homo sapiens* or *Drosophila melanogaster*), a "proxy" sister species, and an outgroup. Using the outgroup to infer the ancestral state, we then consider the replacements that have occurred either on the focal or on the proxy lineage since their divergence from their last common ancestor. The logic of the analysis is as follows. Replacements on different lineages may occur only as independent events; conversely, multiple nearby replacements on the same lineage may occur either as independent events or as MNRs. Therefore, to calculate the rate of DNRs $\alpha_d(k)$, we compare the conditional frequencies of substitutions on the focal lineage, given another substitution that has occurred at distance k nucleotides on the focal lineage ($s_d(k)$) or on the proxy lineage ($d_d(k)$):

$$\alpha_d(k) = s_d(k) - d_d(k).$$

A similar approach can be used to estimate the frequency of TNR events, on the basis of the difference in the frequencies of substitutions pairs when one more substitution has occurred nearby on the same lineage or on a different lineage. The frequency of TNRs $\alpha_t(k,l)$ can be calculated as the difference between the prevalence of these two patterns, after the DNRs are controlled for (see Materials and Methods). Three substitutions on the same lineage may occur due to five scenarios: as three independent events; involving three different sets of DNRs; or involving a TNR (fig. 1*E*). Therefore, we calculate the rate of TNRs $\alpha_t(k,l)$ as the difference between the conditional frequencies at which two substitutions at distances k and $k+l$ from the first substitution are observed, given the first substitution that has occurred on the same (focal) phylogenetic lineage ($s_t(k,l)$) or on the other (proxy) lineage ($d_t(k,l)$). We also need to subtract the probabilities that the first and the second mutations come as a DNR and that the first and the third mutations come as a DNR (see Materials and Methods for a more detailed explanation). Therefore,

$$\alpha_t(k,l) = s_t(k,l) - d_t(k,l) - \alpha_d(k)d_d(k+l) - \alpha_d(k+l)d_d(k).$$

This approach controls for the nonuniformity of the replacement rates along the genome, as long as this nonuniformity is conserved between the focal and the proxy species. It allows us to infer the rates of DNRs and TNRs for different distances between sites k and l; it can also be easily extended for more complex events, that is, those involving more than three simultaneous nucleotide substitutions or other types of mutations.

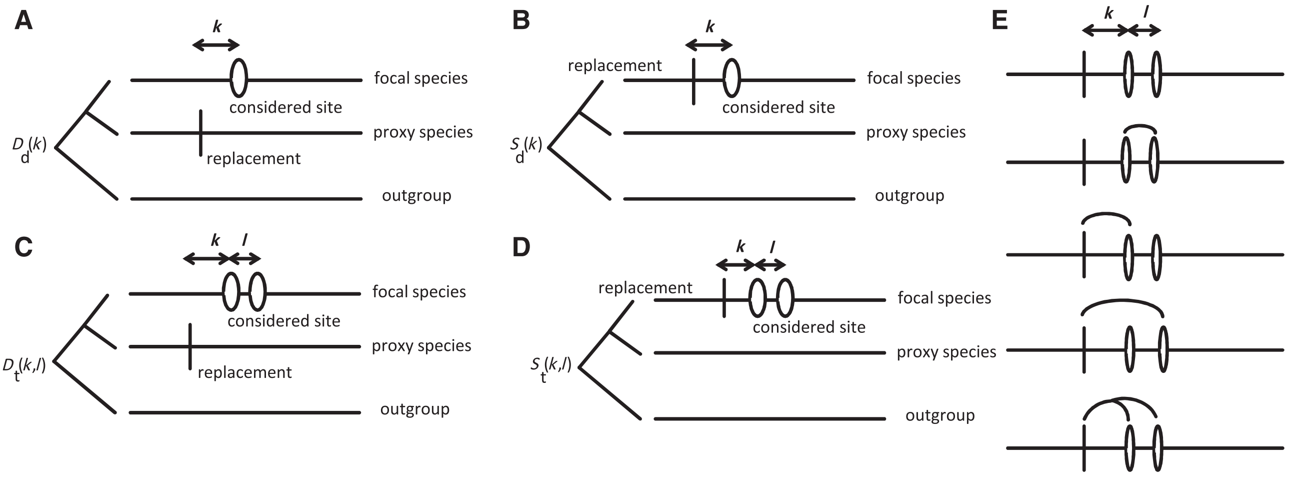

Fig. 1. Inferring the frequencies of DNRs (*A*,*B*) and of TNRs (*C*–*E*). The schematic phylogenetic tree at the left represents, from top to bottom, the two sister species, focal and proxy, and the outgroup; the adjacent horizontal lines represent multiple sequence alignments for the corresponding species. (*A*,*B*) The frequencies of substitutions on the focal lineage are measured within the set of sites (vertical ovals), such that another substitution (vertical line) is observed at distance *k* from them on the proxy (*A*) or on the focal (*B*) lineage ($d_d(k)$ and $s_d(k)$, respectively). (*C*,*D*) The frequencies of pairs of substitutions on the focal lineage are measured within the set of sites (pairs of vertical ovals at distances *l* from each other), such that another substitution (vertical line) is observed at distance *k* from them on the proxy (*A*) or the focal (*B*) lineage ($d_t(k,l)$ and $s_t(k,l)$, respectively). *E*, five scenarios that can give rise to three adjacent substitutions on the same lineage, from top to bottom: three distinct mutations; two distinct mutations (an arc connects positions involved in an MNR), one being a DNR involving the second and the third, the first and the second, or the first and the third of the three nucleotides; and one TNR.

Primates

In primates, we study four trios of species. All trios use the lineage of *H. sapiens* as the focal lineage but consider a range of species at increasing phylogenetic distances from *H. sapiens* as the proxy and the outgroup species.

Even for the shortest considered phylogenetic distances (human–chimpanzee pair), the observation of a substitution in the proxy species does not bias the frequency of a substitution at a nearby site on the *H. sapiens* lineage (fig. 2, red lines). The sole exception is the immediately adjacent sites ($k = 1$), where this frequency is underestimated due to exclusion of a subset of pairs of substitutions giving rise to a CpG dinucleotide. This result, consistently with previous findings (Hodgkinson et al. 2009; Johnson and Hellmann 2011; Seplyarskiy et al. 2012), shows that there is little, if any, short-scale variation in the mutation rates along the genomes of primates, perhaps with the exception of the immediately adjacent positions.

In contrast, a substitution that occurs on the *H. sapiens* lineage radically increases the probability of another substitution on the same lineage nearby (fig. 2, blue lines). Therefore, a substantial fraction of substitutions within a lineage at nearby sites comes as a single-MNR event. These results suggest that DNRs spanning distances of up to 10 nucleotides are responsible for at least $\sum_{1\leq k\leq 10}\alpha_d(k) = 0.023$ of all observed single-nucleotide substitutions. $\alpha_d(k)$ decreases rapidly with k, so that $\alpha_d(1)$ exceeds $\alpha_d(10)$ by at least a factor of approximately 6 (fig. 2, green lines); among all DNRs at distances up to 10 nucleotides, 36% occur at immediately adjacent sites.

A substantial fraction of substitutions also come as part of a TNR. Figure 3 presents the results for the case when two of the three mutations involved in a TNR occurred at adjacent sites ($l = 1$), and figure 4*A*, all values of $\alpha_t(k,l)$ for $k \leq 10$ and $l \leq 10$ (see also supplementary table S1, Supplementary Material online). TNRs spanning distances of up to 10 nucleotides are responsible for at least $\sum_{1\leq k+l\leq 10}\alpha_t(k,l) \approx 0.0060$ of all observed single-nucleotide substitutions or for 26% of DNRs spanning such distances. Similar to DNRs, the rate of TNRs decreases rapidly with the distance spanned, so that $\alpha_t(1,1)$ exceeds $\alpha_t(10,1)$ by a factor of approximately 7 (fig. 3, green lines). Among TNRs spanning distances up to 10 nucleotides, 7.8% affect the three immediately adjacent nucleotides (fig. 4*A*).

Between the four analyzed phylogenies, the length of the considered focal (*H. sapiens*) lineage differs by a factor of 3.8. Correspondingly, the number of pairs of nearby nucleotide sites each carrying a substitution on the human lineage $D_d(k)$ (which scales with lineage length quadratically) differs by a factor of approximately 10, and the fraction of sites, among sites with a substitution on the human lineage, that also carry another substitution at a nearby site $d_d(k)$ (which scales with lineage length linearly) differs by a factor of approximately 4. Nevertheless, we get similar estimates of $\alpha_d(k)$ from different phylogenies, implying that our estimates of the frequencies of MNRs are robust (fig. 5*A*).

Although the fraction of MNRs among all replacements is constant and independent of the phylogeny, their contribution to the genome-level patterns is disproportionally high for very closely related species. Indeed, in the human–chimp comparison, 50% of pairs of substitutions at adjacent sites were due to DNRs (fig. 5*B*) and 83% of triplets of substitutions at adjacent sites were due to TNRs. This fraction is reduced as species accumulate more differences: In the human–macaque comparison, DNRs and TNRs are responsible for only 22%

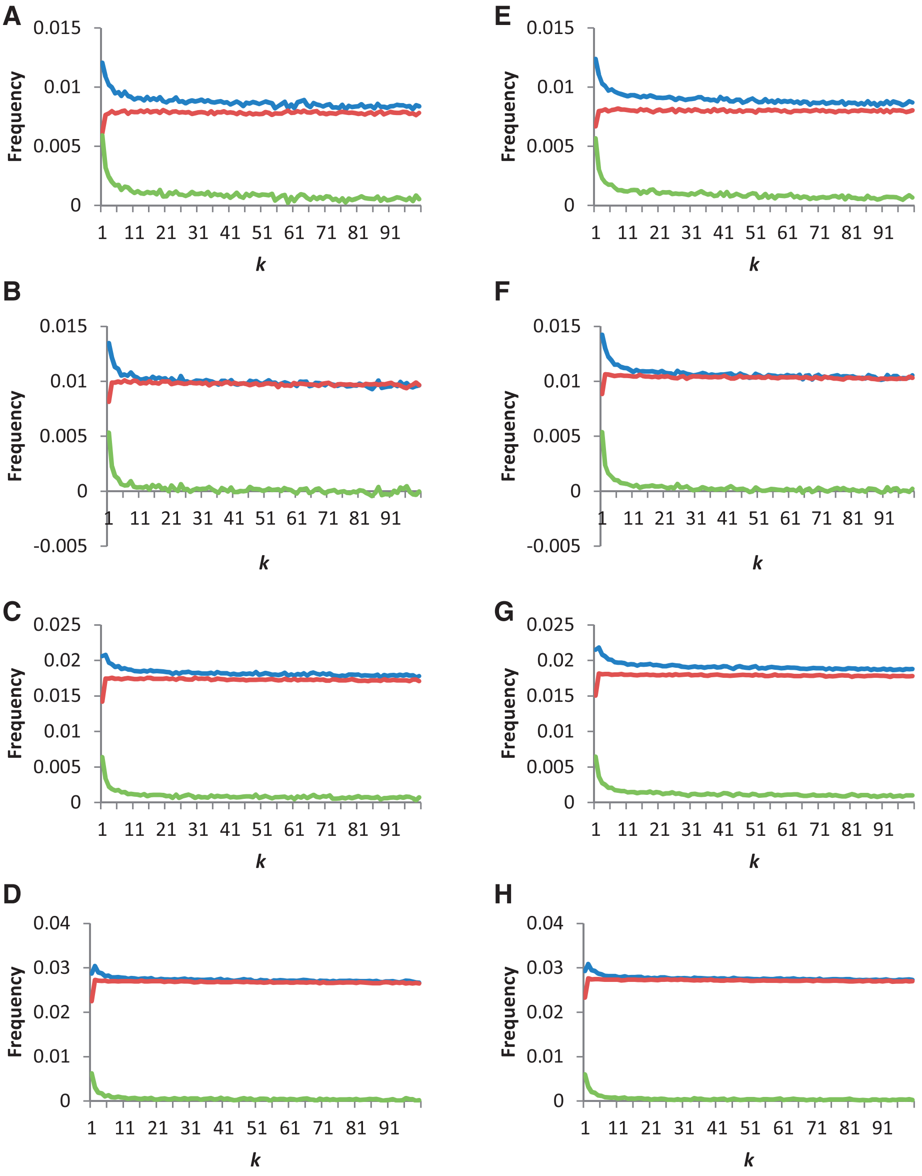

FIG. 2. Frequencies of DNRs in primates for different distances between sites. $d_d(k)$ (red), $s_d(k)$ (blue), and $\alpha_d(k)$ (green) are shown for distances $1 \le k \le 100$ between the sites (horizontal axis). Left column (*A*–*D*), introns; right column (E–*H*), intergenic regions. (*A*,*E*) *Homo sapiens* and *Pan troglodytes* (*Gorilla gorilla* as outgroup), (*B*,*F*) *H. sapiens* and *G. gorilla* (*Pongo pygmaeus* as outgroup), (*C*,*G*) *H. sapiens* and *P. pygmaeus* (*Macaca mulatta* as outgroup), and (*D*,*H*) *H. sapiens* and *M. mulatta* (*Callithrix jacchus* as outgroup). CpG dinucleotides are excluded, which leads to underestimation of $d_d(1)$ and $s_d(1)$ (see Materials and Methods).

(44%) of such pairs (fig. 5*B*). This is intuitive: In closely related species, multiple substitutions at adjacent sites were unlikely to have accumulated by chance and are usually due to an MNR.

Drosophila

In contrast to primates, in *Drosophila*, even a substitution that occurs on the proxy lineage substantially increases the frequency of nearby substitutions on the *D. melanogaster* lineage (figs. 6 and 7, red lines). This effect spans distances of tens of nucleotides and is consistent with pervasive short-scale heterogeneity of the substitution rate in *Drosophila* (Seplyarskiy et al. 2012).

Nevertheless, a substitution that occurs on the *D. melanogaster* lineage increases the frequency of another substitution on the *D. melanogaster* lineage nearby more radically than the substitution on the proxy lineage (figs. 6 and 7, blue lines), implying a high frequency of MNRs (figs. 6 and 7, green lines). The contributions of DNRs and TNRs spanning distances of up to 10 nucleotides are $\sum_{1 \le k \le 10} \alpha_d(k) \approx 0.056$ and $\sum_{1 \le k+l \le 10} \alpha_t(k,l) \approx 0.039$ (fig. 4*B* and supplementary table S2, Supplementary Material online), respectively; DNRs and TNRs affecting two or three immediately adjacent nucleotides are responsible for $\alpha_d(1) \approx 0.02$ and $\alpha_t(1,1) \approx 0.004$ of all single-nucleotide substitutions, respectively. Therefore, DNRs and TNRs contribute, respectively, approximately 2 times and approximately 6 times more to the substitution rate in *Drosophila* than in primates. DNRs are responsible for 34% of pairs of substitutions at adjacent sites, and TNRs are responsible

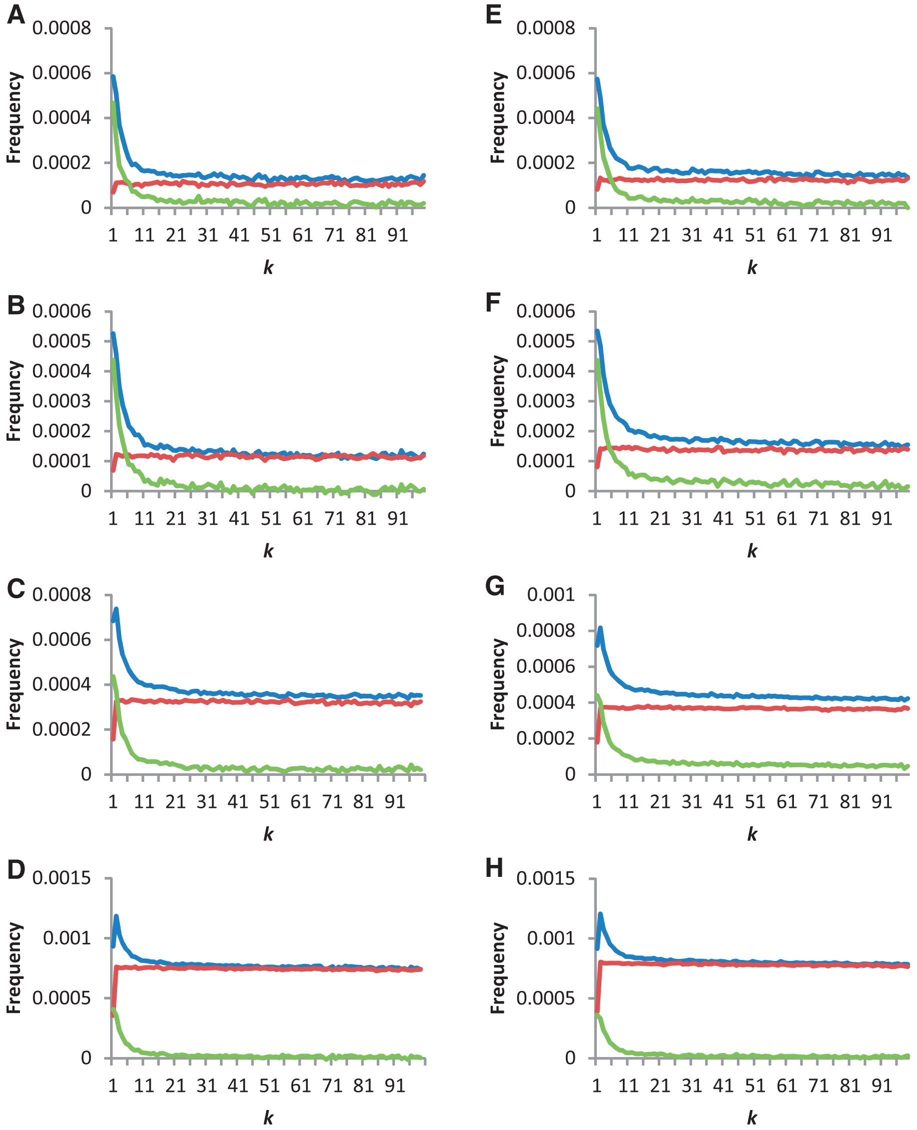

Fig. 3. Frequencies of TNRs, such that two of the three replacements affect adjacent sites ($l = 1$), for different distances k from the third site in primates. $d_t(k,l)$ (red), $s_t(k,l)$ (blue), and $\alpha_t(k,l)$ (green) are shown for distances $1 \leq k \leq 100$ (horizontal axis). The panels correspond to the panels in figure 2.

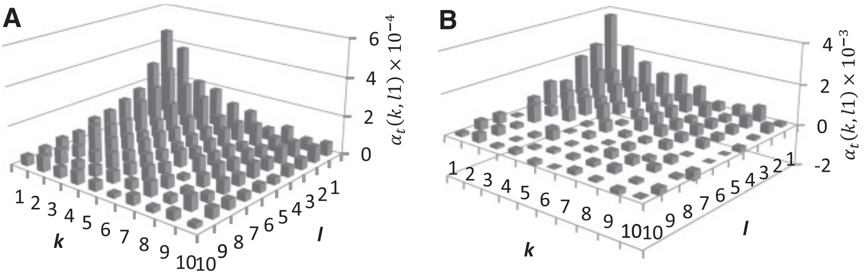

Fig. 4. Frequencies of TNRs. $\alpha_t(k,l)$ for all values of k and l between 1 and 10. (*A*) Human–chimpanzee comparison and (*B*) *D. melanogaster*–*D. simulans* comparison. Only intronic sites are shown.

for 43% of triplets of substitutions at adjacent sites. As in primates, both $\alpha_d(k)$ and $\alpha_t(k,l)$ decrease rapidly with k over a distance of approximately 10 nucleotides. Slightly higher estimates of $\alpha_d(k)$ are obtained using the ML-based inference of the ancestral states (supplementary fig. S1, Supplementary Material online). This difference is due to inclusion of nucleotide sites with multiple replacements in the ML analysis; when such less reliable sites are excluded, identical estimates are obtained using the two methods.

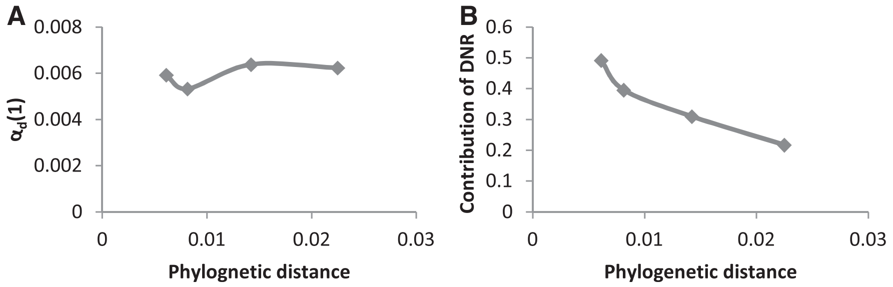

Fig. 5. Dependence of the contribution of DNRs on the length of the phylogenetic lineage of the focal species in introns. (*A*) $\alpha_d(1)$ as the function of $d_d(1)$ and (*B*) $\alpha_d(1)/s_d(1)$ as the function of $d_d(1)$. At each panel, the four points, from left to right, correspond to the substitutions on the *Homo sapiens* lineage after its divergence from *Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*, and *Macaca mulatta*, respectively.

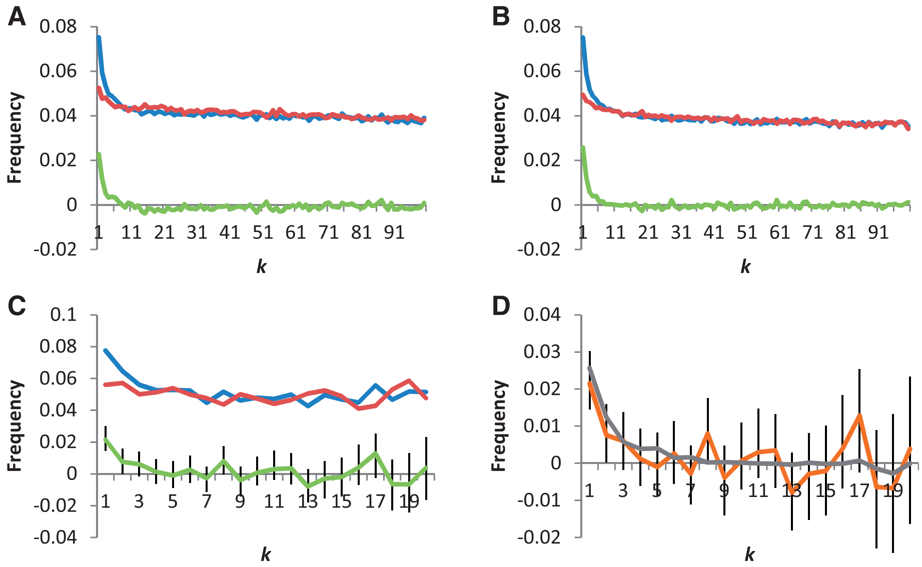

Fig. 6. Frequencies of DNRs in *D. melanogaster–D. simulans* comparison for different distances between sites. (A–C) $d_d(k)$ (red), $s_d(k)$ (blue), and $\alpha_d(k)$ (green) are shown for distances $1 \leq k \leq 100$ between the sites (horizontal axis). (*A*) Introns; (*B*) intergenic regions; and (*C*) positions 8–30 in introns with lengths up to 120 nucleotides. (*D*) $\alpha_d(k)$ for all intronic sites (gray) and positions 8–30 in introns with lengths up to 120 nucleotides (orange) plotted together. The differences between the two curves are insignificant for all *k*. Error bars for $\alpha_d(k)$ correspond to 95% confidence intervals obtained by 1,000 bootstrap simulations; for data from all intronic sites and intergenic regions, they are not shown because they would be barely visible.

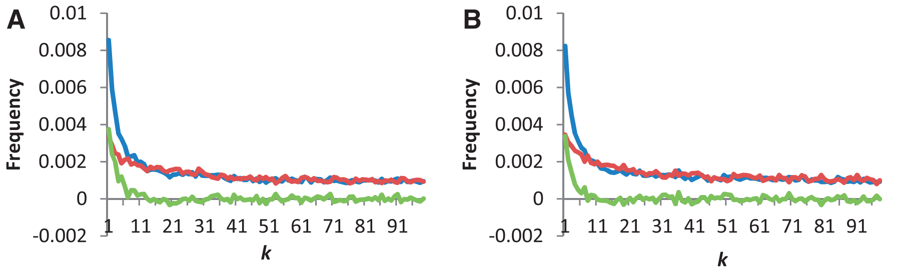

Fig. 7. Frequencies of TNRs in *D. melanogaster–D. simulans* comparison, such that two of the three replacements affect adjacent sites ($l = 1$), for different distances *k* from the third site in drosophilids. $d_t(k,l)$ (red), $s_t(k,l)$ (blue), and $\alpha_t(k,l)$ (green) are shown for distances $1 \leq k \leq 100$ between sites (horizontal axis). (*A*) Introns and (*B*) intergenic regions.

Discussion

Our results imply that at least 2.3% of single-nucleotide replacements observed in evolution of the human lineage, and at least 5.6% of those observed on the *D. melanogaster* lineage, occurred as part of MNR events spanning multiple nucleotides at distances up to 10 from each other. These estimates control for the nonuniformity of replacement rates along the genomes, as long as this nonuniformity is preserved in both evolving lineages. Such nonuniformity may occur due to variation in the mutation rates, which is known to be high both in primates (Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2011; Johnson and Hellmann 2011; Seplyarskiy et al. 2012) and, in particular, in *Drosophila* (Stamatoyannopoulos et al. 2009; Seplyarskiy et al. 2012; Weber et al. 2012) or to differences in the strength of negative selection affecting noncoding regions. For example, if the local mutation rates are autocorrelated at distances up to approximately 10 nucleotides, $d_d(k)$ and $s_d(k)$ for $k < 10$ will be elevated equally (because any such heterogeneity will cause mutations to be clumped along the sequence, no matter on what lineage they occur), but $\alpha_d(k)$ will not be affected.

Still, the observation of MNRs can come from a number of sources: nonuniformity of mutation rates and/or selection that varies between the two analyzed lineages; genome sequencing, assembly, or alignments errors; mutations simultaneously affecting several nucleotides; epistatic interactions among replacements; and/or nonallelic gene conversion.

Local patterns of mutation and selection may change in the course of evolution. If the sister species are so distant that homologous genome segments have different local replacement rates (e.g., because their local mutation rates diverged), α may provide a biased estimate for the frequency of MNRs; in this case, for example, substitutions may be clustered on the same lineage due to differences in the mutation rates between lineages rather than due to MNRs. To estimate this effect, we used as proxy species four different primate species located at different phylogenetic distances from the focal species (there is no range of suitable *Drosophila* species) and compared the obtained $\alpha_d(k)$. The results were similar (fig. 5), suggesting that the contribution of the change in the local mutation or selection rates to our estimates of the rate of MNRs is minor.

Assembly and/or alignment errors may lead to spurious clustering of nucleotide replacements, and sequencing errors may have the same effect if they are clustered in the genome (Löytynoja and Goldman 2008; Wong et al. 2008; Mallick et al. 2009; Schneider et al. 2009). With our approach, such double errors in the genomes of proxy species do not affect $\alpha_d(k)$ or $\alpha_t(k)$. Only multiple errors at nearby sites in the genomes of *H. sapiens* or *D. melanogaster* can inflate $\alpha_d(k)$ or $\alpha_t(k)$; however, these genomes have high-quality sequences, and errors in them are infrequent. Nevertheless, we did two additional tests to estimate their contribution. First, we reasoned that errors are unlikely to coincide in two species with independently assembled genomes. However, when only nucleotides matching between human and chimpanzee are used in the human–orangutan comparison, the values of $\alpha_d(k)$ remain very similar to those when obtained using all nucleotides (supplementary fig. S2, Supplementary Material online). Second, to minimize the effect of alignment errors, we only considered nucleotides further than 20 nucleotides from alignment gaps; again, the results were similar (supplementary figs. S3 and S4, Supplementary Material online), implying that the contribution of alignment errors is also minor.

MNRs may occur due to epistatic selection. Under positive epistasis, that is, when the first nucleotide replacement facilitates the second, replacements can co-occur on the same lineage (Bazykin et al. 2004). This is the leading cause of co-occurrence of nonsynonymous replacements within coding regions (Bazykin et al. 2004, 2006; Callahan et al. 2011; Bazykin and Kondrashov 2012). However, epistasis is impossible in the absence of selective constraint. Indeed, if the ancestral and the two-substitution variants have the same fitness, but the intermediate variant is deleterious, the rate of evolution decreases as the selection against the intermediate variant increases (Kimura 1985). Conversely, if selection against the intermediate variant is weak, we expect a near-neutral rate of evolution, but little clumping, and no clumping at all if the intermediate variant is neutral. Clumping due to epistasis can also occur when the double mutant has a higher fitness than the ancestral variant and the two-substitution variant as a whole is positively selected. In this case, the overall rate of evolution (and, correspondingly, degree of conservation) can be higher, lower, or the same as in the neutral regions, depending on the parameters (fig. 1 in Lynch and Abegg 2010). Still, data suggest that this mode of selection is more likely in regions where the overall conservation is high. Indeed, clumping of nonsynonymous substitutions caused by positive selection is stronger in conservative regions of proteins (Bazykin and Kondrashov 2012). More generally, positive selection affects a larger fraction of substitutions in conservative regions of proteins (Bazykin and Kondrashov 2012) and of noncoding regions (Cai et al. 2009, Halligan et al. 2011). Therefore, excluding the conservative sites should restrict the influence of epistatic selection.

The estimates for the fraction of intronic or intergenic sites in primates that are selectively constrained based on divergence and polymorphism patterns range from 2.8% to 11% (Waterston et al. 2002; Dermitzakis et al. 2005; Asthana et al. 2007; Ponting and Hardison 2011; Ward and Kellis 2012). Selection in noncoding regions of *Drosophila* is more prevalent and constrains the evolution of more than 50% of intronic and intergenic sites (Andolfatto 2005). The higher fraction of genome under selection in *Drosophila*, compared with human, seems consistent with a higher observed rate of MNRs in the former if MNRs are caused by selection. However, to minimize the contribution of selection in our analyses, we excluded all the nucleotide sites that show signs of conservation between species (see Materials and Methods). In doing so, we excluded the fraction of the genome (∼40% in humans and ∼60% in *Drosophila*) that is larger than the currently estimated fraction of the genome under selective constraint. Choosing an even more radical threshold (excluding ∼70% of the genome in *Drosophila*) leads to very similar results (supplementary fig. S5, Supplementary Material online).

To further study the possible contribution of selection in *Drosophila* to MNRs, we compared the rate of MNRs at positions 8–30 of short introns (fig. 6*C*) and the rest of the intronic and intergenic sites (fig. 6*A* and *B*). Positions 8–30 of short introns are the class of sites in *Drosophila* that is least subject to both positive or negative selection (Parsch et al. 2010). Because of an approximately 10× smaller sample size, the confidence intervals for the estimates of the rate of MNRs at this class of sites were large; still, we observe a significantly nonzero $\alpha_d(k)$ for $k < 3$, and no difference between the $\alpha_d(k)$ for short and all introns for all values of k (fig. 6*D*). In summary, the data seem to suggest that the contribution of selection to MNRs in nonconstrained regions is minor.

Conceivably, MNRs may also be caused by nonallelic gene conversion. In this process, a segment of DNA may be converted in one of the lineages by a similar genomic region. If the conversion tract spans more than one nucleotide difference, this will lead to nearby correlated changes in the genome sequence and may inflate $\alpha_d(k)$. However, our results show that the majority of MNRs span distances of no more than 10 nucleotides, whereas conversion tracts are much longer (Hilliker et al. 1994; Jeffreys and May 2004; Mancera et al. 2008; Wang et al. 2012). To estimate the magnitude of any potential contribution of gene conversion, we calculated $\alpha_d(k)$ for the 50% of the genome that lacked close paralogous sequences and was thus unlikely to have experienced nonallelic gene conversion. The results of this analysis (supplementary figs. S6 and S7, Supplementary Material online) were similar to those obtained using the entire genome. Therefore, it is unlikely that gene conversion is a major contributor to the MNRs.

Finally, MNRs can be caused by MNMs. Although direct sequencing experiments are the best way to measure the mutation rates, such data sets so far are of limited scope; few MNMs have been observed in them, and therefore, the resulting estimates have a large variance. Still, our 2.3% estimate for the rate of DNRs in primates falls near to 2–3% obtained by 1000 Genomes Project Consortium et al. (2010); and our 5.6% estimate for this rate in *Drosophila* is consistent with four dinucleotide mutations among 174 de novo mutations (2.3%) observed by Keightley et al. (2009). These numbers are also concordant with indirect measurements of MNM rates in humans based on polymorphism frequencies (2.0%; Schrider et al. 2011).

In summary, our results support a major role for MNRs in metazoans. A large fraction of the MNRs involves more than two nucleotides. Moreover, although the MNRs spanning immediately adjacent nucleotides are numerically the most prevalent, they comprise only approximately 30% of DNRs, and only approximately 10% of TNRs; the remaining MNRs span larger distances. Although a high prevalence of MNRs appears to be universal (Schrider et al. 2011), we show that their frequency differs between species, so that the fraction of MNRs among all replacements in *Drosophila* exceeds that in primates by a factor of approximately 2 for DNRs and by a factor of approximately 7 for TNRs. Such a high fraction of MNRs should be taken into account in evolutionary models, especially when estimation of epistatic selection is the goal, because MNRs may facilitate adaptation (Kimura 1985).

Materials and Methods

Data

Multiple complete genome alignment of *Callithrix jacchus*, *Macaca mulatta*, *Pongo pygmaeus*, *Gorilla gorilla*, and *Pan troglodytes* to *H. sapiens* (hg18) (Fujita et al. 2011) was obtained from UCSC Genome Browser (http://genome.ucsc.edu). Data on human variation in nine diploid nuclear genotypes and the reference human genome, for a total of 19 haploid genotypes, were obtained as described in Seplyarskiy et al. (2012).

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 162 complete genomes of *D. melanogaster* aligned to dm3 reference genome of *D. melanogaster* (Mackay et al. 2012) were downloaded from http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/, and data for six complete genomes of *D. simulans* (Begun et al. 2007) aligned to dm2 reference genome of *D. melanogaster* were downloaded from DPGP (http://www.dpgp.org/). dm2 co-ordinates were converted to dm3 co-ordinates using liftover (http://hgdownload.cse.ucsc.edu/goldenPath/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.

The sets of analyzed nucleotide sites were formed as follows. Nucleotides masked by RepeatMasker, not aligned, or containing gaps or non-ACGT characters were not considered. To reduce the effect of natural selection on our inferences, we excluded all coding regions, 5′- and 3′-UTRs, all alternatively spliced regions, and nucleotide sites with high interspecies conservation: phastCons (Siepel et al. 2005) scores >0.1 in *Drosophila* and PhyloP (Pollard et al. 2010) scores >0.1 in primates. (For supplementary fig. S5, Supplementary Material online, phastCons scores >0.05 in *Drosophila* were excluded.) Including polymorphic sites may bias the estimates of divergence when the compared species are closely related (Keightley and Eyre-Walker 2012); therefore, we also excluded sites polymorphic in human in the analyses of primates, or in *D. melanogaster* or in *D. simulans* in the analyses of *Drosophila*.

To infer the lineage-specific nucleotide substitutions, two approaches were used. First, we used a maximum parsimony-based approach. To this end, in each comparison, we compared trios of species including two sister species and an outgroup. In primates, we considered several trios, with increasing phylogenetic distance between the sister species: *H. sapiens* and *Pan troglodytes* (*G. gorilla* as outgroup), *H. sapiens* and *G. gorilla* (*P. pygmaeus* as outgroup), *H. sapiens* and *P. pygmaeus* (*M. mulatta* as outgroup), and *H. sapiens* and *M. mulatta* (*C. jacchus* as outgroup). In *Drosophila*, we compared *D. melanogaster* and *D. simulans*, using matched positions of *D. yakuba* and *D. erecta* to infer the ancestral state; positions that differed between *D. yakuba* and *D. erecta* were excluded. We then assumed that a substitution has occurred on one of the sister lineages if the nucleotides on the other sister lineage and in the outgroup coincided and differed from the nucleotide on the considered lineage.

Second, in *Drosophila*, where the inference of the ancestral state is less reliable due to higher phylogenetic distances between species, we used an ML-based approach. To this end, we used the baseml program of PAML (Yang 2007).

Analysis

To estimate the fraction α of single-nucleotide substitutions that occurred as part of an MNR, we compared the lineage-specific substitutions in the genomes of two sister species, using the genome of *H. sapiens* or *D. melanogaster* as the "focal" genome, and its sister genome as the "proxy" genome. The focal genomes have a higher quality sequence and annotation; therefore, in each analysis, we used them to measure the substitution rates.

We estimated the fraction of single-nucleotide differences that arose due to a DNR as follows. For each distance, between the two analyzed nucleotide sites $k \in (1..100)$, we calculated two values: 1) $d_{\mathrm{d}}(k)$, the fraction of sites carrying a single-nucleotide substitution on the focal lineage, among sites at distance k from a site with a substitution on the proxy lineage:

$$d_{\mathrm{d}}(k) = \frac{D_{\mathrm{d}}(k)}{P},$$

and 2) $s_{\mathrm{d}}(k)$, the fraction of sites carrying a single-nucleotide substitution on the focal lineage, among sites at distance k from a site with another substitution on the focal lineage:

$$s_{\mathrm{d}}(k) = \frac{S_{\mathrm{d}}(k)}{F}.$$

Here, $D_{\mathrm{d}}(k)$ is the number of pairs of nucleotide sites with co-ordinates $(i,\ i+k)$, such that one substitution occurred at the first site on the proxy linage and another at the second site on the focal lineage; $S_{\mathrm{d}}(k)$ is the number of pairs of nucleotide sites with co-ordinates $(i,\ i+k)$, such that one substitution occurred at each of the two sites on the focal lineage (fig. 1*A* and *B*); and P and F are the overall numbers of single-nucleotide substitutions on the proxy and the focal lineage, respectively.

Because two substitutions that occur on different lineages can only arise due to two distinct replacement events, $D_{\mathrm{d}}(k) = x_{\mathrm{fp}}(k)P$, where $x_{\mathrm{fp}}(k)$ is the probability of a substitution on the focal lineage at a site at distance k from a substitution on the proxy lineage, and

$$d_{\mathrm{d}}(k) = x_{\mathrm{fp}}(k).$$

In contrast, two substitutions on the same lineage can occur due to either two distinct replacement events or a single DNR. Therefore, $S_{\mathrm{d}}(k) = x_{\mathrm{ff}}(k)F + \alpha_{\mathrm{d}}(k)F$, where $x_{\mathrm{ff}}(k)$ is the probability of a substitution on the focal lineage at a site at distance k from another substitution on the focal lineage, and $\alpha_{\mathrm{d}}(k)$ is the probability that a single-nucleotide substitution at position i in the focal genome originated as part of a DNR also involving a substitution at position $i+k$. Therefore,

$$s_{\mathrm{d}}(k) = x_{\mathrm{ff}}(k) + \alpha_{\mathrm{d}}(k).$$

If the distribution of the replacement rates along the genome is preserved between the two lineages, $x_{\mathrm{fp}}(k) = x_{\mathrm{ff}}(k)$, and

$$\alpha_{\mathrm{d}}(k) = s_{\mathrm{d}}(k) - d_{\mathrm{d}}(k).$$

To estimate the contribution of TNRs, we calculate the frequencies of cases when two substitutions occurred at sites separated by l nucleotides in the focal genome, within the set of sites at distance k from the nearest of them in the proxy genome $d_{\mathrm{t}}(k,l)$ or in the focal genome $s_{\mathrm{t}}(k,l)$:

$$d_{\mathrm{t}}(k,l) = \frac{D_{\mathrm{t}}(k,l)}{P}$$

$$s_{\mathrm{t}}(k,l) = \frac{S_{\mathrm{t}}(k,l)}{F},$$

where $D_{\mathrm{t}}(k,l)$ is the number of triplets of nucleotide sites with co-ordinates $(i,\ i+k,\ i+k+l)$, such that one substitution occurred at the first site on the proxy lineage and two more in the second and the third sites on the focal lineage; and $S_{\mathrm{t}}(k,l)$ is the number of triplets of nucleotide sites with co-ordinates $(i,\ i+k,\ i+k+l)$, such that a substitution occurred at each of these three sites on the focal lineage (fig. 1*C* and *D*). The first pattern can arise due to either three distinct replacement events or two replacement events, one of them being a DNR:

$$d_{\mathrm{t}}(k,l) = x_{\mathrm{fp}}(k)\,x_{\mathrm{fp}}(k+l)P + x_{\mathrm{fp}}(k)\alpha_{\mathrm{d}}(l)P,$$

The second pattern can arise under five different scenarios (fig. 1*E*):

$$\begin{aligned} S_{\mathrm{t}}(k,l) = {} & x_{\mathrm{ff}}(k)x_{\mathrm{ff}}(k+l)F + x_{\mathrm{ff}}(k)\alpha_{\mathrm{d}}(l)F + \alpha_{\mathrm{d}}(k)x_{\mathrm{ff}}(k+l)F \\ & + x_{\mathrm{ff}}(k)\alpha_{\mathrm{d}}(k+l)F + \alpha_{\mathrm{t}}(k,l)F, \end{aligned}$$

where $\alpha_{\mathrm{t}}(k,l)$ is the probability that a single-nucleotide substitution at position i in the focal genome originated as part of a TNR also involving substitutions at positions $i+k$ and $i+k+l$, and the five components of the sum correspond to the five scenarios in figure 1*E*. Therefore,

$$\alpha_{\mathrm{t}}(k,l) = s_{\mathrm{t}}(k,l) - d_{\mathrm{t}}(k,l) - \alpha_{\mathrm{d}}(k)d_{\mathrm{d}}(k{+}l) - \alpha_{\mathrm{d}}(k{+}l)d_{\mathrm{d}}(k).$$

CpG dinucleotides are hypermutable in vertebrates; this can lead to correlations between consecutive mutations at adjacent sites if the first mutation creates a CpG (Bird 1980). To avoid the contribution of this phenomenon to measurements of MNR rates, in all analyses involving primates, we excluded all sites involved in a CpG dinucleotide in the outgroup species. For analyses involving pairs of adjacent sites, we also excluded pairs of substitutions in the same lineage that possibly involved a CpG intermediate, that is, N_1pG → CpN_2 or CpN_1 →N_2pG, where N_1 and N_2 correspond to any nucleotide and pairs of substitutions on different lineages, such that the ancestral variant was N_1pG or CpN_2 and one of the derived variants was a CpG dinucleotide. This latter filter leads to underestimation of $d_{\mathrm{d}}(1)$, $s_{\mathrm{d}}(1)$, $d_{\mathrm{t}}(1,l)$, and $s_{\mathrm{t}}(1,l)$ but should not bias $\alpha_{\mathrm{d}}(k)$ or $\alpha_{\mathrm{t}}(k,l)$.

To estimate the possible contribution of nonallelic gene conversion to MNRs, we took the following approach. We reasoned that genomic regions that have a close paralog elsewhere in the genome are more likely to undergo nonallelic gene conversion. To identify these regions, we ran basic local alignment search tool (BLAST) (Altschul et al. 1990) of the entire *H. sapiens* (*D. melanogaster*) genome against itself. Among the resulting BLAST hits, we removed all those with co-ordinates overlapping the query sequence. We then sorted the remaining BLAST hits by their *e* values and considered the number of BLAST hits from the top of this list that cumulatively covered 50% of the genome. The 50% of the genome that was covered by these BLAST hits and the remaining 50% were then considered as prone and not prone to nonallelic gene conversion, respectively.

Supplementary Material

Supplementary figures S1–S7 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Ministry of Education and Science of the Russian Federation (grant nos 11.G34.31.0008 and 8814), by the Russian Foundation for Basic Research (grant no. 12-04-33202 mol_a_ved), and by the Molecular and Cellular Biology Program of the Russian Academy of Sciences. The research was performed at Moscow State University (Moscow, 119991, Russia) and Institute of Information Transmission Problems (Moscow, 127994, Russia).

References

Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol.* 215(3):403–410.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437:1149–1152.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci U S A.* 104:12410–12415.

Averof M, Rokas A, Wolfe KH, Sharp PM. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287:1283–1286.

Bazykin GA, Dushoff J, Levin SA, Kondrashov AS. 2006. Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. *Proc Natl Acad Sci U S A.* 103:19396–19401.

Bazykin GA, Kondrashov FA. 2012. Major role of positive selection in the evolution of conservative segments of *Drosophila* proteins. *Proc Biol Sci.* 2791742:3409–3417.

Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS. 2004. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature* 429:558–562.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5:e310.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* 8:1499–1504.

Callahan B, Neher RA, Bachtrog D, Andolfatto P, Shraiman BI. 2011. Correlated evolution of nearby residues in Drosophilid proteins. *PLoS Genet.* 7:e1001315.

Chen J-M, Copper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nature* 8:762–775.

Chen J-M, Férec C, Cooper DN. 2009. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. *Hum Mutat.* 30:1435–1448.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* 5:e1000336.

Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat Rev Genet.* 6:151–157.

Donmez N, Bazykin GA, Brudno M, Kondrashov AS. 2009. Polymorphism due to multiple amino acid substitutions at a codon site within *Ciona savignyi*. *Genetics* 181:685–690.

Fujita PA, Rhead B, Zweig AS, et al. (27 co-authors). 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39: D876–D882.

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. *Mol Biol Evol.* 28(9):2651–2660.

Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A. 1994. Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. *Genetics* 137(4):1019–1026.

Hodgkinson A, Eyre-Walker A. 2010. Human triallelic sites: evidence for a new mutational mechanism? *Genetics* 184:233–241.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet.* 12:756–766.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. *PLoS Biol.* 7:e1000027.

Jeffreys AJ, May CA. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet.* 36(2):151–156.

Johnson PLF, Hellmann I. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol Evol.* 3: 842–850.

Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J Mol Evol.* 74:61–68.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* 19: 1195–1201.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 143(5):837–847.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. *J Genet.* 64:7–19.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat.* 21:12–27.

Kondrashov AS. 2008. Another step toward quantifying spontaneous mutation. *Proc Natl Acad Sci.* 27:9133–9134.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol.* 24(7):1464–1479.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320(5883):1632–1635.

Lynch M, Abegg A. 2010. The rate of establishment of complex adaptations. *Mol Biol Evol.* 27(6):1404–1414.

Mackay TF, Richards S, Stone EA, et al. (52 co-authors). 2012. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 482: 173–178.

Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* 19: 922–933.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479–485.

Miyazawa S. 2011. Selective constraints on amino acids estimated by a mechanistic codon substitution model with multiple nucleotide changes. *PLoS One* 6(3):e17244.

Paigen K, Petkov P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet.* 11(3):221–233.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol Biol Evol.* 27:1226–1234.

Pollard KS, Hubisz MJ, Rosenboom K, Siepel A. 2010. Detection of nonneutral substitution rates on Mammalian phylogenies. *Genome Res.* 20:110–121.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res.*11:1769–1776.

Poon AF, Lewis FI, Frost SD, Kosakovsky Pond SL. 2008. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* 24(17):1949–1950.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol.* 1:114–118.

Schrider DR, Hourmozdi JN, Hahn MW. 2011. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol.* 21:1051–1054.

Seplyarskiy VB, Kharchenko P, Kondrashov AS, Bazykin GA. 2012. Heterogeneity of the transition/transversion ratio in *Drosophila* and Hominidae genomes. *Mol Biol Evol.* 29(8):1943–1955.

Siepel A, Bejerano G, Pedersen, et al. (13 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034–1050.

Smith NGC, Webster MT, Ellegren H. 2003. A low rate of simultaneous double-nucleotide mutations in primates. *Mol Biol Evol.* 20:47–53.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet.* 41(4):393–395.

Stoletzki N, Eyre-Walker A. 2011. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. *Mol Biol Evol.* 28:1371–1380.

Tweedie S, Ashburner M, Falls K, et al. (43 co-authors). 2009. FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* 37:D555–D559.

Wang J, Fan HC, Behr B, Quake SR. 2012. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150(2):402–412.

Wang J, Gonzalez KD, Scaringe WA, Tsai K, Liu N, Gu D, Li W, Hill KA, Sommer SS. 2007. Evidence for mutation showers. *Proc Natl Acad Sci U S A.* 104:8403–8408.

Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675–1678.

Waterston RH, Lindblad-Toh K, Birney E, et al. (223 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Weber CC, Pink CJ, Hurst LD. 2012. Late-replicating domains have higher divergence and diversity in *Drosophila melanogaster*. *Mol Biol Evol.* 29(2):873–882.

Whelan S, de Bakker P, Quevillon E, Rodriguez N, Goldman N. 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Res.* 34:D327–D331.

Whelan S, Goldman N. 2004. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics* 167(4):2027–2043.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473–476.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139(2):993–1005.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* 15(12):496–503.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.