A Low Rate of Simultaneous Double-Nucleotide Mutations in Primates

Nick G. C. Smith, Matthew T. Webster, and Hans Ellegren
Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

The occurrence of double-nucleotide (doublet) mutations is contrary to the normal assumption that point mutations affect single nucleotides. Here we develop a new method for estimating the doublet mutation rate and apply it to more than a megabase of human-chimpanzee-baboon genomic DNA alignments and more than a million human single-nucleotide polymorphisms. The new method accounts for the effect of regional variation in evolutionary rates, which may be a confounding factor in previous estimates of the doublet mutation rate. Furthermore we determine sequence context effects by using sequence comparisons over a variety of lineage lengths. This approach yields a new estimate of the doublet mutation rate of 0.3% of the singleton rate, indicating that doublet mutations are far rarer than previously thought. Our results suggest that doublet mutations are unlikely to have caused the correlation between synonymous and nonsynonymous substitution rates in mammals, and also show that regional variation and sequence context effects play an important role in primate DNA sequence evolution.

Introduction

Although point mutations are usually assumed to involve single nucleotides, both direct observations of mutational events in the laboratory (Nakazawa et al. 1994) and comparative analyses of DNA sequences (Wolfe and Sharp 1993) have suggested that simultaneous adjacent double-nucleotide mutation events (hereafter termed doublet mutations) do occur in nature. Recently, Averof et al. (2000) used sequence comparisons of coding and noncoding DNA from diverse organisms to reveal a high frequency of doublet mutations, about 2% of the single-nucleotide mutation rate. We have attempted to develop an improved estimate of the doublet mutation rate, focusing on mutation processes in primate noncoding DNA sequences. Noncoding sequences are expected to provide a more reliable indication of underlying mutation processes than coding sequences, because they are less likely to be affected by selection.

The importance of quantifying the doublet mutation rate becomes evident if we consider the consequences of incorrectly assuming that all point mutations occur at single nucleotides. Fundamentally, the pattern of neutral evolution is determined by the pattern of mutation, so unless we understand mutational processes we may falsely reject neutrality. For example, the classic test of the neutral theory using the index of dispersion of substitutions (see Gillespie 1991) will be biased by doublet mutations. The neutral prediction, that the index of dispersion is one, is based on the Poisson model assumptions that mutations are random, independent, and single. Thus doublet mutations will generate overdispersion of the molecular clock: if all mutations are doublets then, relative to singleton mutations, the mean number of substitutions will be doubled but the variance in the number of substitutions will be quadrupled; hence the index of dispersion will be two rather than one. More generally, the explicit models of molecular evolution required for substitution rate estimation and phylogenetic inference usually involve single nucleotide changes (e.g., Li 1997). If doublet mutations are common then such models, and the results they generate, may be biased.

Finally, estimating the level of doublet mutations should help resolve the debate concerning the causes of the correlation between synonymous and nonsynonymous substitution rates in mammals (Wolfe and Sharp 1993; Smith and Hurst 1999). Doublets provide a mutational (i.e., neutral) explanation for why nonsynonymous changes in protein-coding genes (those changes in DNA sequence which affect the amino acid sequence) covary with synonymous changes (those changes in DNA sequence which do not affect the amino acid sequence owing to the degeneracy of the genetic code). For example, changes at the second codon position are always nonsynonymous while changes at the third codon position are mostly synonymous: thus a doublet mutation can simultaneously generate both a synonymous and a nonsynonymous mutation. Note, however, that there is also a methodological debate over whether there really is a synonymous-nonsynonymous correlation in mammals (Bielawski, Dunn, and Yang 2000).

Materials and Methods

Substitution Data

Primate genomic alignments were built in a number of stages. Chimpanzee and baboon sequence data orthologous to regions of human chromosome 7, generated by the NIH Intramural Sequencing Center (NISC) Comparative Sequencing Initiative (see http://www.nisc.nih.gov/), were retrieved using National Center for Biotechnology Information (NCBI) Entrez (http://www.ncbi.nlm.nih.gov/). Those sequences reported as “working draft sequence” were broken up into their constituent unordered pieces. Regions of the human genome orthologous to the chimpanzee and baboon sequences were identified by BLAST searches (Altschul et al. 1997) against the human genome. These BLAST searches provided positional information for the removal of overlapping sequences. Human-chimpanzee-baboon alignments were generated using the default values of ClustalW (Thompson, Higgins, and Gibson 1994) as implemented at the Human Genome Mapping Project (HGMP) (http://www.hgmp.mrc.ac.uk/). The resulting alignments were checked to remove poorly aligned
regions, thereby ensuring high-quality genomic alignments, insensitive to changes in alignment parameters (see Smith, Webster, and Ellegren 2002). In all, 43 human-chimpanzee-baboon alignments with a total ungapped and unmasked length of 1.7Mb were obtained. Details of the sequences in our alignments are available on request.

The positions of alignments in human contigs were determined by BLAST searches against the human genome, and comparison to the contig annotation files available at NCBI allowed the masking of coding regions within alignments. Repetitive sequence elements were masked with RepeatMasker (A.F.A. Smit and P. Green, unpublished). Repetitive elements were masked because they are known to evolve at higher rates than nonrepetitive sequence (Chen and Li 2001), and so may generate strong regional variation and sequence context effects. Microsatellites were masked using the program Sputnik (Abajian, unpublished, http://abajian.net/sputnik/) because microsatellites cause alignment problems and show unusual substitution patterns.

Lineage-specific substitutions were classified using parsimony (e.g., if the human-chimp-baboon sequences are A-C-C, then a C to A change is inferred down the human lineage). For both pairwise and lineage-specific substitution we ignored both singletons which may be due to the hypermutability of methylated CG dinucleotides (CG to TG and CG to CA). To nullify the effect of CG mutation on doublet estimates, all potential CG-mediated tandems (CH to TG and DG to CA, where H is A/C/T and D is A/G/T) were removed, as were the corresponding near-neighbors. Tandems are pairs of adjacent differences, whereas near-neighbors are pairs of differences separated by one nucleotide. The corresponding near-neighbors had to be removed because not all putative CG tandems are generated by CG hypermutation.

Given that there are several alternative strategies for counting different types of substitutions, here we present our algorithms in greater detail. Slight alterations to the algorithms made little difference to estimates of the doublet mutation rate (results not shown). All alignments were analyzed separately after masking for genes and repeats. The following classes of sites were defined for both the human-chimp and human-baboon pairwise comparisons and the human and chimp lineages: substitutions (s), conserved sites (c) and masked sites (m). In addition, some sites were undefined (u) in the human and chimp lineages (i.e., parsimony uninformative). Then potential CpGs were masked (s → m) as described above (note that CpG masking differs between the pairwise comparisons and lineages because in the latter case the direction of substitution is known). Then types of substitutions were counted. The total number of substitutions was found by counting all the “s”-sites; tandems were found by counting all the adjacent pairs of “s”-sites (including overlapping pairs, e.g., “sss” contains two doublets), and near-neighbors were identified by counting all pairs of “s”-sites separated by one site (which could belong to any class, and again overlapping pairs were allowed, e.g., “scsms” contains two near-neighbors). Finally, the effective length of the alignment, required for calculating the Averof et al. (2000) measure of the doublet mutation rate, was found by counting all the “s”-sites and “c”-sites.

### Polymorphism Data

We downloaded the tenth release of The SNP Consortium (TSC) database, consisting of 1,255,326 mapped single-nucleotide polymorphisms (SNP’s) from http://snps.cshl.org/index.html. Tandems and near-neighbors could be identified because the polymorphism data files gave the base positions of each SNP along the chromosomes as well as positions within human contigs. All SNP’s identified as coding by BLAST searches of surrounding sequence against the January 2001 version of the human mRNA RefSeq database (Pruitt et al. 2000) were removed. Dawson et al. (2001) found no evidence for polymorphism tandems to occur preferentially in repetitive sequences, and so the polymorphism data was not masked for repeat sequences. Potential CG-mediated singletons, tandems, and near-neighbors were also removed as for substitutions. In the case of our polymorphism data, for which both the direction of mutation and the linkage patterns were unknown, the complementary removal of near-neighbors is particularly important because many of the putative CG tandems (identified as Y(R/K/S) and (M/S/Y/R where Y is C/T, R is A/G, K is G/T, S is C/G, and M is A/C) will not have been caused by CG hypermutation. Runs of three or more adjacent polymorphisms were also masked as these might be considered likely to result from sequencing errors; this procedure made no qualitative difference to our results (data not shown). Tandem polymorphisms were counted as all instances of pairs of polymorphisms one base pair apart; near-neighbor polymorphisms were counted as all instances of pairs of polymorphisms two base pairs apart.

### Results

#### Estimating the Doublet Mutation Rate

Averof et al. (2000) estimated the relative rate of doublet mutations, here termed $D_{oc}$, by comparing the observed number of tandem substitutions ($T_0$) with the expected number of tandem substitutions ($T_e$), calculating the expected number of tandem substitutions by assuming that a substitution was equally likely at every site:

$$D_{oc} = (T_0 - T_e)/(S_0 - 2(T_0 - T_e)).$$

$S_0$ is the observed number of substitutions, and so $S_0$ minus twice the inferred number of doublets gives the number of singletons. The terms doublet and tandem refer to adjacent bases, but doublets refer to the (inferred) mutation process, whereas tandems refer to differences between compared sequences. Thus we may observe many tandem differences, but we cannot infer doublet mutations unless we can show that the observed number of tandem differences exceeds the expected number.

The assumption of no rate variation between sites does not seem generally applicable (Yang 1996), however, and so it is worth considering the effect of rate variation between sites. We define two types of rate variation: sequence context and regional variation. By sequence context we refer to the dependence of mutation rates on the identities of nearby bases (Templeton et al. 2000; Zavolan...
and Kepler 2001), and by regional variation we refer to all larger scale variation in mutation rates.

Regional variation is suggested by $K_S$ variation in mammals, between genes throughout the genome (Wolfe, Sharp, and Li 1989; Matassi, Sharp, and Gautier 1999; Lercher, Williams, and Hurst 2001) as well as within genes (Tsunoyama, Bellgard, and Gojobori 2001), and by variation in noncoding substitution rates in primates (Smith, Webster, and Ellegren 2002). If there is regional variation in mutation rates, then it is easy to see that the expected number of tandems will be underestimated. For instance, consider 10 kb compared between two samples with a mean distance of 10%; if there is no variation in substitution rates, we expect 100 tandems (length of sequence times the square of the distance). Now imagine that there is rate variation with 9 kb at 8% and 1 kb at 28%, in which case although the mean distance is the same the expected number of tandems rises to 136. So, unless the regional variation is accounted for, 36 doublet mutations will be falsely inferred.

Sequence context occurs on a much finer scale than regional variation, but it can generate tandems in a similar way. For example, it has been shown that, with an adjacent 5’ G, the mutagenic guanine product dG-AF induces mutations at a much higher frequency with an adjacent 3’ C than with an adjacent 3’ G (Shibutani et al. 2001). Thus if a GGG trinucleotide undergoes a primary mutation to GGC, then the chance of a secondary mutation affecting the middle nucleotide is greatly increased. Thus sequence context can make the generation of a tandem more likely than expected if mutations were independent.

Here we present a simple method to address the problem of regional variation. Instead of comparing the observed number of tandems to the number of tandems expected under a given model, we compare the observed number of tandems ($T_o$), pairs of differences one base apart, with the observed number of what we refer to as near-neighbors ($NN_o$), pairs of differences two bases apart, e.g., as in AGGTT compared to ATGAT. Just as many additional near-neighbors will be generated by regional variation as tandems, so the difference gives the number of doublet mutations inferred from the data. The doublet mutation rate, $D_m$, is then quantified relative to the number of singleton mutations as follows:

$$D_m = (T_o - NN_o)/(S_o - 2(T_o - NN_o)). \tag{2}$$

The underlying assumption of our new method is that there are only two sorts of mutations, singletons and doublet; in other words there are no simultaneous near-neighbor mutations.

The formula for $D_m$ also corrects for sequence context effects to some extent, because sequence context effects will generate near-neighbors as well as tandems. The correction for sequence context effects, however, will certainly not be perfect because such effects appear to be much stronger at a distance of one base than at two bases (Krawczak, Ball, and Cooper 1998). The strongest known case of sequence context, the hypermutability of CG dinucleotides, can be explicitly accounted for by ignoring potentially affected differences (see Materials and Methods). Such an approach is not possible if weak sequence context effects are common, and so our approach for dealing with sequence context effects is to consider sequence differences along lineages of different lengths. Short lineages provide few primary mutations to affect secondary mutations, and sequence context effects should thus be weak.

Simulations

We performed some simulations to confirm that both the $D_{oe}$ and $D_m$ methods are robust to two notable features of DNA sequence evolution: (i) variation in the rates at which different types of DNA mutations occur and (ii) rate variation between sites (assuming that such rate variation is random with respect to genomic position; i.e., there is no systematic regional variation). Intuitive reasoning suggests that neither factor should affect the methods because they do not affect the relative positions at which mutations appear. We performed simulations of DNA sequence evolution using the evolve program in the PAML package (Yang 1997). In all cases, we generated three sequences of 1 million base pairs each according to a tree with distances similar to those in the human-chimpanzee-baboon tree—[(human:0.005, chimpanzee:0.005):0.04, baboon:0.045]) —and we analyzed the simulated sequences using the same methods employed for the human-chimpanzee-baboon alignments (see Materials and Methods; for simplicity we did not apply CpG masking to the simulated data). Four sets of simulated data were generated using different DNA substitution models: (1) JC69 model (Jukes and Cantor 1969), (2) HKY85 model (Hasegawa, Kishino, and Yano 1985) with equal base frequencies and a transition/transversion ratio of 5, (3) as in (2), but with unequal base frequencies $C = G = 0.4$ and $A = T = 0.1$, and (4) as in (3) but with rate variation between sites corresponding to a “discretized” gamma distribution (Yang 1994) with shape parameter 0.5 and eight rate categories.

The results of the simulations are shown in table 1, in which each simulation data set is analyzed in three comparisons of different lineage lengths corresponding to those performed on the human-chimpanzee-baboon data described below. In none of the 12 sets of analyses is there a significant difference between the observed and expected numbers of tandems ($T_o$ and $T_e$) or between the observed numbers of tandems and near-neighbors ($T_o$ and $NN_o$). These results suggest that neither the $D_{oe}$ nor the $D_m$ method is likely to generate biases through a failure to provide an explicit model of DNA evolution. The simulated “human-baboon” comparisons also suggest that the use of parsimony, or more specifically the failure to account for multiple substitutions at the same site, does not bias the $D_{oe}$ or $D_m$ methods. As one moves from simulation (1) to simulation (4), the failure to account for multiple hits leads to increasing underestimation of substitution rates; thus $S_{oe}$, $T_o$, and $NN_o$ all decrease. However, this bias does not seem to affect the difference between $T_o$ and $T_e$ or that between $T_o$ and $NN_o$. Therefore, although both the $D_{oe}$ and $D_m$ methods are somewhat ad hoc, the simulations indicate that they are not biased by multiple hits, nonregional rate variation between sites, or...
variation in the rates of different types of DNA mutations. Thus, although it may ultimately be desirable to address the issue of doublet mutations through explicit DNA models combined with likelihood or Bayesian approaches, the development of such methods does not appear to be a pressing concern.

Substitutions

Nucleotide substitutions were identified in the non-coding nonrepetitive regions of a large set of human-chimpanzee-baboon alignments (see Materials and Methods). With the three species alignments, a number of different comparisons were possible, each considering substitutions down lineages of different lengths. The human-baboon pairwise comparison represents the longest lineage, with a total of 50 million years of evolution (twice the time to the common ancestor 25 million years ago [Purvis 1995]). The human-chimpanzee pairwise comparison represents about 10 million years of evolution, given species divergence 5 million years ago (Chen and Li 2001). Finally, the human and chimpanzee lineage-specific substitutions both represent roughly 5 million years of evolution (i.e., time since species divergence), and they can be combined because substitutions in the two lineages are independent when parsimony is used to infer substitutions.

For all three substitution comparisons, both $D_{oe}$ and $D_{in}$ were calculated, and table 2 shows that in each case $D_{in}$ is much lower than $D_{oe}$ (human-baboon pairwise $D_{in} = 0.95\%$ and $D_{oe} = 1.74\%$, human-chimpanzee pairwise $D_{in} = 0.53\%$ and $D_{oe} = 1.14\%$, human-baboon pairwise $D_{in} = 0.35\%$ and $D_{oe} = 0.80\%$). Because $D_{oe}$ assumes no rate heterogeneity, the difference between $D_{in}$ and $D_{oe}$ indicates mainly the effect of regional variation within the 43 alignments ranging in length from 12 kb to 107 kb (but not variation between alignments because expected numbers of tandems were calculated separately for each alignment). The magnitude of the difference shows the necessity of accounting for regional variation in substitution rates.

It is also clear from table 2 that both estimates of the doublet mutation rate decrease as the length of the lineages decreases, with a greater than twofold difference in $D_{in}$ and $D_{oe}$ between the human-baboon pairwise substitutions and the human and chimpanzee lineage-specific substitutions. The effect of lineage length is consistent with sequence context effects because longer lineages provide more scope for secondary nonindependent mutations. Lineage length effects are also apparent in noncoding primate data reported by Averof et al. (2000), because most of their doublet mutations were inferred to be in the longest lineage (see their table 2 and figure 3).

Polymorphisms

Given the effect of lineage length on estimates of the doublet mutation rate, the obvious step after considering substitutions among primates is to turn to human polymorphisms. Not only is the human genome the subject of intensive efforts to determine polymorphisms (The International SNP Map Working Group 2001), but also humans have a lower effective population size than other primates (Kaessmann et al. 2001), and so offer the shortest possible lineage length among the primates, and thus the weakest sequence context effects. Because the average age, measured in generations, of neutral polymorphisms is four times the effective population size (Kimura 1983), estimates of a human effective population size of 10,000 (Jorde, Watkins, and Bamshad 2001) and a generation time of 25 years (Keightley and Eyre-Walker 2000) indicate that human polymorphisms have an average age of roughly 1 million years (possibly lower if ancestral generation times and effective population sizes were lower). Thus the lineage length of human polymorphisms is roughly five times shorter than the lineage length of human and chimpanzee lineage-specific substitutions (see table 2).

An excess of tandem SNP's relative to the number expected assuming no polymorphism level heterogeneity has previously been observed on human chromosome 22 (Dawson et al. 2001). Applying the $D_{oe}$ method to the singleton and doublet data given by Dawson et al. generates a doublet mutation rate of 0.70%. This value, however, is likely to be a serious overestimate because CG hypermutation mediated tandems (see Materials and Methods) were not removed from their analysis (indeed Dawson et al. (2001) found TG $\leftrightarrow$ CA tandems to be the most common), and because there is extreme regional variation in levels of polymorphism (see Dawson et al., Figure 1).

The estimation of the doublet mutation rate $D_{oe}$ using TSC SNP data, 0.27%, appears reasonable, although the results should be viewed with caution because the sampling of polymorphisms over the human genome is

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Numbers of Different Types of Substitutions in Simulated Sequence Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation 1</td>
</tr>
<tr>
<td>Human-baboon pairwise</td>
<td>85,121, 7,199, 7,270, 7,222</td>
</tr>
<tr>
<td>Human-chimp pairwise</td>
<td>9,917, 94, 99, 103</td>
</tr>
<tr>
<td>Human andchimp lineage-specific</td>
<td>9,362, 40, 9,578, 43</td>
</tr>
<tr>
<td>(So, T o, T e, N No)</td>
<td>48, 43, 51, 48</td>
</tr>
</tbody>
</table>

Note.—The numbers of all substitutions (So), observed tandems (To), expected tandems (Te), and observed near-neighbors (N No) in data sets simulated under four different evolutionary models (see Results for details). The expected tree lengths for the simulations were based on the observed values for the human-chimpanzee-baboon data set, thus allowing comparisons down lineages of different lengths as in table 2.
clearly patchy. The doublet mutation rate will only be underestimated if the sampling process is biased toward looking for polymorphisms at single nonadjacent sites, a finding which seems unlikely. Just as with the substitution comparisons, $D_{tn}$ is much less than $D_{oe}$. The large difference between the two doublet mutation rate estimates and the 25-fold excess of observed versus expected tandems probably reflect the fact that both polymorphism levels (see Figure 2b and supplementary material of The International SNP Map Working Group [2001]) and sampling intensity vary greatly across human chromosomes.

**Discussion**

The comparison of the $D_{tn}$ estimates for the polymorphism and substitution data should allow the determination of sequence context effects. When lineages are very short, then sequence context effects should be minimal, but it is not clear how to quantify this except by comparison of different lineage lengths. If one considers lineages of progressively shorter length, then when the $D_{oe}$ estimates approach an asymptotic minimum, sequence context effects may be inferred to be minimal at the shortest lineage lengths (but note that very short lineages contain relatively little information and thus yield imprecise estimates of the doublet mutation rate). The fact that the $D_{tn}$ estimate for human polymorphisms is only slightly less than that for the human and chimpanzee lineage-specific substitutions suggests that both estimates are inflated only marginally by sequence context effects. So our best estimate of the doublet mutation rate relative to the singleton mutation rate is 0.3%, the average of the polymorphism and human and chimpanzee lineage-specific substitution values. This value is over six times lower than the value of 2% obtained by Averof et al. (2000).

It is desirable to consider the various possible types of doublet mutations to see if the results of mutagenesis studies match with the results of sequence comparisons. For example, it is interesting to see whether CC to TT mutations, which have been found in the p53 gene as the result of UV exposure (Nakazawa et al. 1994), are revealed by our sequence comparisons. The most appropriate data for this purpose are the human and chimpanzee lineage-specific substitutions, because the use of the baboon as an outgroup allows the direction of substitution to be inferred.

Because mutations affecting both strands need to be considered, we compare tandems and near-neighbors for GG to AA changes as well as CC to TT changes (GNG to ANA and CNC to TNT for near-neighbors). There are six tandems and two near-neighbors in our sequence comparisons, and this excess of tandems is consistent with mutagenesis studies, although the difference is not significant because of the small numbers of differences. The lack of data in the present study precludes a full analysis of all possible types of doublets, so the generation of additional SNP data should help such studies considerably. Linkage information will allow a more discriminating test of doublet mutations: apart from rare recombination events doublet SNP’s should be in perfect linkage disequilibrium. Outgroup sequences, such as those from chimpanzee, will allow the determination of the direction of mutation in polymorphism studies, and thus the more accurate analysis of mutation processes.

What is the significance of our finding of a doublet mutation rate of 0.3% as opposed to the estimate of 2% obtained by Averof et al. (2000)? Does this discrepancy in doublet mutation rates result from the differences between the $D_{oe}$ and $D_{tn}$ methods or perhaps from the differences between the data sets? We can discount the latter possibility in two ways. First, we reanalyzed the data set of Averof et al. (2000). Their alignment of primate pseudo eta globin sequences, masked for CpG mutations, was kindly provided by M. Averof. We used the program baseml in the PAML package (Yang 1997) to reconstruct the ancestral sequences and we analyzed the changes down the lineage leading to rhesus monkey (this is the longest lineage, with the most data, and it provides the strongest signal of doublet mutations). Using this method, slightly different from that employed by Averof et al. (2000), we identified 232 substitutions, 17 tandems, and 14 near-neighbors and calculated the expected number of tandems

### Table 2

**Numbers of Different Types of Substitutions and Polymorphisms in Primate Sequence Data**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Time (MY)</th>
<th>$S_{o}$</th>
<th>$T_{o}$</th>
<th>$N_{No}$</th>
<th>$D_{tn}$</th>
<th>$T_{e}$</th>
<th>$D_{oe}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-huboon pairwise substitutions</td>
<td>50</td>
<td>46,808</td>
<td>2,773</td>
<td>2,337</td>
<td>0.95%</td>
<td>1,986</td>
<td>1.74%</td>
</tr>
<tr>
<td>Human-chimp pairwise substitutions</td>
<td>10</td>
<td>8,792</td>
<td>170</td>
<td>124</td>
<td>0.53%</td>
<td>72.3</td>
<td>1.14%</td>
</tr>
<tr>
<td>Human and chimp lineage-specific substitutions</td>
<td>5</td>
<td>9,613</td>
<td>123</td>
<td>90</td>
<td>0.35%</td>
<td>47.5</td>
<td>0.80%</td>
</tr>
<tr>
<td>Human polymorphisms</td>
<td>1</td>
<td>898,383</td>
<td>6,581</td>
<td>4,187</td>
<td>0.27%</td>
<td>295</td>
<td>0.71%*</td>
</tr>
</tbody>
</table>

*Note.*—Estimates of the doublet mutation rate in primates: the numbers of all substitutions ($S_{o}$), observed tandems ($T_{o}$), expected tandems ($T_{e}$), and observed near-neighbors ($N_{No}$) used to generate two different estimates of the doublet mutation rate, $D_{tn}$ and $D_{oe}$, calculated for a variety of different sequence comparisons ranging in lineage length from 1 million to 50 million years. $D_{oe}$ compares tandems and near-neighbors, whereas $D_{tn}$ compares the observed and expected numbers of tandems (see eqs. 1 and 2).

* * *

$a$ Our polymorphism estimate of $D_{oe}$ is derived from the reported density of one SNP per 3050 bp for the TSC data (The International SNP Map Working Group 2001).
as 8.29. Thus the $D_{oe}$ method yields a doublet mutation rate of 3.8%, whereas the $D_{in}$ method gives 1.3%, suggesting that the two methods give different results in both data sets. Second, we can partition our human and chimp lineage-specific results according to alignment-specific substitution rates. This partitioning enables a test of the idea that our human-chimpanzee-baboon data set may contain selectively constrained regions in which doublet mutations may be particularly constrained (unlike the pseudogene analyzed by Averof et al. (2000), which is almost certainly free of constraint). If so, we would predict the doublet mutation rate to be lower in regions with low substitution rates. However, this prediction does not hold true ($D_{in}$ is 0.58% in low substitution rate alignments and 0.18% in high substitution rate alignments, varying as expected given that the doublet mutation rate is conditioned on the singleton mutation rate).

Given that we can be confident that the $D_{oe}$ and $D_{in}$ methods yield different results when applied to real sequence data, an important conclusion of this study is the strength of regional variation and sequence context, effects we consider responsible for most of the excess of observed tandems relative to the expectation with no rate heterogeneity. The results of our study of primate genomic sequences are thus consistent with regional variation in synonymous and noncoding substitution rates across mammalian nuclear genomes (Lercher, Williams, and Hurst 2001; Tsunoyama, Belligard, and Gojobori 2001; Smith, Webster, and Ellegren 2002), sequence context in chloroplasts (Morton, Oberholzer, and Clegg 1997), sequence context in mitochondria (Howell and Smejkal 2000), and the utility of the auto-discrete-gamma model to describe sequence evolution in primate mitochondria (Yang 1995).

The direct effect of a low doublet mutation rate is obviously a diminished belief in the importance of doublet mutations. Such a low rate suggests that doublet mutations cannot fully explain the strong correlation between synonymous and nonsynonymous substitution rates, $K_S$ and $K_A$, observed in mammals (Wolfe and Sharp 1993). In the comparison presented by Wolfe and Sharp (1993), of 363 genes in mouse and rat, there were on average 28 nonsynonymous and 56 synonymous differences per gene after correction for multiple hits. A doublet mutation rate of 0.3% means that the expected number of doublets is only 0.25: thus over three quarters of genes will not be affected by doublet mutation. The effect of doublets on the $K_A$–$K_S$ correlation can be investigated by simulating substitutions according to Poisson distributions based on the data of Wolfe and Sharp (1993), assuming that all genes have independent synonymous and nonsynonymous singleton substitution rates based on the mean numbers of substitutions and numbers of sites (hence the expected $K_A$–$K_S$ correlation coefficient in the absence of doublet mutations is zero). If all doublet mutations are considered to generate one synonymous and one nonsynonymous mutation (a conservative assumption for our purposes), the expected increase in the $K_A$–$K_S$ correlation coefficient arising from a doublet mutation rate of 0.25 is just 0.005. Applying the Averof et al. (2000) doublet mutation rate of 2% means 1.7 doublets on average, which increases the expected $K_A$–$K_S$ correlation confidence by 0.042. Thus it seems highly unlikely that doublet mutations are solely responsible for the observed increase in the $K_A$–$K_S$ correlation coefficient above the neutral expectation (Ohta 1995), given that this has been quantified as an increase of 0.141 in the mouse-rat study of Smith and Hurst (1999).

**Acknowledgments**

H.E. is a Royal Swedish Academy of Sciences Research Fellow supported by a grant from the Knut and Alice Wallenberg foundation. This study was supported by the Swedish Research Council. Thanks to M. Averof for providing alignments.

**Literature Cited**


Lercher, M. J., E. J. B. Williams, and L. D. Hurst. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implica-


Kenneth Wolfe, Associate Editor

Accepted September 3, 2002