Likelihood Analysis of Asymmetrical Mutation Bias Gradients in Vertebrate Mitochondrial Genomes

Jeremiah J. Faith[1] **and David D. Pollock**[2]

Department of Biological Sciences and Biological Computation and Visualization Center,
Louisiana State University, Baton Rouge, Louisiana 70803

Manuscript received August 2, 2002
Accepted for publication June 17, 2003

ABSTRACT

Protein-coding genes in mitochondrial genomes have varying degrees of asymmetric skew in base frequencies at the third codon position. The variation in skew among genes appears to be caused by varying durations of time that the heavy strand spends in the mutagenic single-strand state during replication (D_{ssH}). The primary data used to study skew have been the gene-by-gene base frequencies in individual taxa, which provide little information on exactly what kinds of mutations are responsible for the base frequency skew. To assess the contribution of individual mutation components to the ancestral vertebrate substitution pattern, here we analyze a large data set of complete vertebrate mitochondrial genomes in a phylogeny-based likelihood context. This also allows us to evaluate the change in skew continuously along the mitochondrial genome and to directly estimate relative substitution rates. Our results indicate that different types of mutation respond differently to the D_{ssH} gradient. A primary role for hydrolytic deamination of cytosines in creating variance in skew among genes was not supported, but rather linearly increasing rates of mutation from adenine to hypoxanthine with D_{ssH} appear to drive regional differences in skew. Substitutions due to hydrolytic deamination of cytosines, although common, appear to quickly saturate, possibly due to stabilization by the mitochondrial DNA single-strand-binding protein. These results should form the basis of more realistic models of DNA and protein evolution in mitochondria.

VERTEBRATE mitochondrial DNA (mtDNA) has been a prevalent model system for genomic biodiversity and molecular evolution research largely due to the genome's manageable size of ~17 kb, its high copy number, and a variety of protein and structural RNA-encoding genes that have proven useful for phylogenetic inference (Wolstenholme 1992; Pollock *et al.* 2000). Mutations in these genes are important in disease and the aging process (Linnane *et al.* 1989; Lindahl 1993; Wallace 1999).

The directions of transcription for the 37 genes of the circular mitochondrial genome are unevenly distributed between the heavy (H) strand and the light (L) strand (so named for their different buoyancy densities in CsCl gradients). The light strand of mtDNA codes for 8 tRNAs and 1 protein, while the heavy strand codes for 14 tRNAs, 2 rRNAs, and 12 proteins.

There is strong heterogeneity in the rate of substitution in vertebrate mitochondria, both between taxonomic groups and among sites. Because of the importance of substitution rates in evolutionary analysis, there is considerable interest in the underlying transcription and replication processes that give rise to differences in mitochondrial mutation processes. Much work has gone into deciphering the mechanism of mtDNA replication, especially in the mouse and human (Shadel and Clayton 1997; Clayton 2000), and in those organisms it is now known to proceed by an asymmetrical mechanism with the two origins of replication well separated along the genome (Figure 1). Replication begins first at the origin of H-strand replication (O_H), located in the predominately triple-strand D-loop region of the mitochondrial genome. A γ-polymerase extends an RNA primer to produce a nascent H strand, while the displaced parental H strand is then coated with single-strand-mtDNA-binding proteins (mtSSBs). About two-thirds of the way around the genome, the replication fork reaches the origin of L-strand replication (O_L) and a secondary structure is formed that binds a second polymerase molecule to begin synthesis of a new L strand. The parental H strand becomes double stranded as the L strand is synthesized, but for much of the genome considerable time passes before this occurs. Thus, there is a gradient along the genome of duration of time spent in the single-strand state (D_{ssH}; Tanaka and Ozawa 1994; Reyes *et al.* 1998), with the genes closest to O_L in the direction of light-strand replication spending the shortest time in the single-strand state, and the genes closest to O_H spending the longest time (Figure 1). Genes between and behind the two origins also have moderately high D_{ssH}. The entire process of mtDNA replication requires ~2 hr to complete, with

[1] *Present address:* Cold Spring Harbor Laboratory, 1 Bungtown Rd., P.O. Box 100, Cold Spring Harbor, NY 11724.

[2] *Corresponding author:* Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803. E-mail: dpollock@lsu.edu

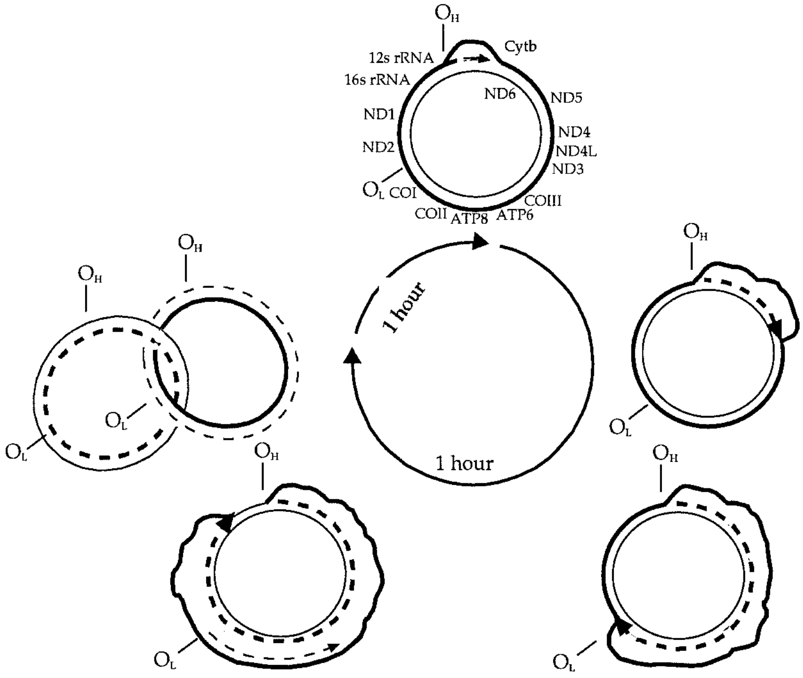

FIGURE 1.—Genome replication in mammalian mtDNA. Replication of the heavy strand (thick line) begins first at O_H, while replication of the light strand (thin line) does not begin until the heavy-strand replication fork passes O_L. Placement of the RNA and protein-coding genes are labeled to show that CytB spends the longest time in the single heavy-strand state (squiggly lines), while COI spends the shortest time. ND1 and ND2 are immediately behind the origin of replication for both strands and also spend a relatively long time in the single-strand state. The entire replication process takes 2 hr, but only ~1 hr is spent synthesizing DNA.

~1 hr spent on DNA synthesis (see SHADEL and CLAYTON 1997; CLAYTON 2000 for review).

It has recently been inferred that the strand asymmetry and gradients of exposure to mutagenesis in the single-strand state during replication lead to gradients of asymmetry in the nucleotide substitution process and thereby to gradients of asymmetry in base composition between the strands. In 1991, it was noticed that T/C ratios at third codon positions in vertebrate mitochondria vary along the genome (LIMAIEM and HENAUT 1984b; DELORME and HENAUT 1991). The strength of this skew appeared to correlate with the position of the gene within its transcript, and similar results were found in *Drosophila yakuba*, despite a different gene arrangement. These findings appeared to be consistent with observations in phages and animal viruses (LIMAIEM and HENAUT 1984a). In the same year, ASAKAWA *et al.* (1991) observed that T and G are found preferentially on H-strand-encoded genes. The bias is clearly not gene specific, since ND1 and ND2, located on the H strand in starfish and the L strand in sea urchins, have TG skews in each organism that match the bias in their coding strands (ASAKAWA *et al.* 1991).

Combining these observations, it was determined that asymmetric compositional bias (skew) in third codon positions and other sites depends on which strand the gene is coded on and on how long the gene spends in the single-strand state during replication (JERMIIN *et al.* 1994, 1995; TANAKA and OZAWA 1994; REYES *et al.* 1998). This conclusion is not completely straightforward because both selection and mutation can affect the substitution process that determines base frequencies. Third codon positions are usually studied to reduce bias caused by the effects of selection on the amino acid sequence, but many third codon positions are not completely free to vary—for example, some sites have only two possible codons to code for the same amino acid. JERMIIN *et al.* (1995) tried to avoid these confounding effects by looking at intergenic spacer regions, but this greatly reduced the number of sites, both because the amount of intergenic region in the genome is limited and because these regions are also subject to insertions and deletions, which reduce the number of sites with useful information. TANAKA and OZAWA (1994) solved the problem by looking at only fourfold redundant sites that can freely exchange between all nucleotides without altering the amino acid and therefore have little or no selective pressures. Even with fourfold redundant sites, there is the possibility of substitution bias arising due to selection for codon usage. If codon bias exists, there is no reason to believe that it would not be consistent across the genome and therefore would not bias the results but might add some noise to the data. This is particularly true in the mitochondrion, since all of the protein-coding genes on the light strand are transcribed simultaneously on a single RNA strand, and gene expression is therefore independent of gene location (CLAYTON 2000).

REYES *et al.* (1998), building on work by TANAKA and OZAWA (1994), presented a detailed hypothesis for variation of mutational pressures in mammalian mtDNA. They concluded that nucleotide skew was correlated with the duration of time spent in the single-strand state, for which TANAKA and OZAWA (1994) had earlier developed a simple measure (see METHODS). This hypothesis was derived from similar models of evolution in bacteria (see FRANCINO and OCHMAN 1997 for a review), studies on mtDNA replication mechanisms, and research on the relative rates of different types of single-strand and double-strand mutations (see Lindahl 1993 for a review). Analysis of base frequency gradients cannot by itself pinpoint the specific nature of the underlying mutations responsible, but an increase in transitions was considered most consistent with compositional analyses. On the basis of mutation studies (Sancar and Sancar 1988; Frederico *et al.* 1990; Lindahl 1993; Ames *et al.* 1995), an increased rate of hydrolytic deamination on the single-strand heavy strand was hypothesized to be the most plausible factor creating biases in single-strand mutation rates, with cytosine-to-uracil mutations occurring two times more often than adenine-to-hypoxanthine mutations (Reyes *et al.* 1998). The cytosine-to-uracil mutations may also be induced by oxygen free radicals (Tanaka and Ozawa 1994). These two mutation pressures are consistent with the observed gradient of bias toward adenine and cytosine on the complementary light strand. It has been shown that under some conditions, an alternate replication mechanism that does not involve a long duration in the single-strand state can occur, and it is unknown which mechanism might predominate in germ cells (Holt *et al.* 2000). The observation of a substitution bias that correlates with D_{ssH} does not imply that the alternative mechanism is not in effect in germ cells, but rather only that the classical mechanism happens often enough to create the asymmetrical bias.

Previous analyses have considered the effect of asymmetric mutation bias primarily in terms of base frequencies, but as Reyes *et al.* (1998) hypothesize, it is probable that only a few specific mutations occur noticeably more often in the single-strand state, leading to changes in specific types of substitution. To analyze rates of different substitution types in mitochondrial genomes directly and test the role of deamination in creating a gradient of skewed base frequencies, we carried out likelihood-based analyses on 42 complete vertebrate mitochondrial genomes. To minimize the effects of rate variation across taxa, the genomes analyzed had the same gene order, including control regions, and excluded taxonomic groups with fast evolutionary rates. Our data set should thus represent the ancestral vertebrate condition, but has predominant representation of the mammals, and thus has substantial overlap with the Reyes *et al.* (1998) data set. Detailed analysis was limited to third codon positions in sites with the same conserved redundancy class throughout the entire genome set. This set of sites will have had the weakest selection pressures acting upon it and avoids biases in frequency distributions that may be caused by including sites that have only recently joined a new redundancy class through substitution at the first or second codon positions. These sites are assumed to be effectively neutral to directly link the substitution and mutation processes. We considered twofold redundant sites in addition to fourfold redundant sites to allow separate analysis of transitions in the absence of the potentially confounding effect of transversion substitutions. By incorporating the phylogenetic tree and additional taxa into our likelihood analyses, we were able to improve accuracy, evaluate gradients of substitution rates rather than simply base frequencies, and test hypotheses using likelihood ratios.

METHODS

Sequence alignment, manipulation, and pruning: Sequences were obtained for all 118 complete vertebrate mitochondrial genomes that had been submitted to GenBank and revised by the National Center for Biotechnology Information RefSeq program (Pruitt *et al.* 2000) at the time our analyses began. Mitochondrial genes were parsed from these genomes according to their annotations and placed into a MySQL relational database. Alignments for each homologous gene set were created automatically using Perl scripts and ClustalW v1.8 (Thompson *et al.* 1994). For protein-coding genes, the amino acid sequences were aligned first, and nucleotide alignments were derived via computational comparisons of the nucleotide sequence with the aligned amino acid sequence. There are a few instances of known frameshifts, and for these the noncoding basepair(s) were removed automatically prior to nucleotide alignment. Predicted amino acid sequences and annotated amino acid sequences were compared as a check on these procedures. Final alignments were examined by eye to check for obvious anomalies and alignment of gapped regions (presumably corresponding to regions of structural malleability).

Since gene order is hypothesized to be critical for mutation bias, gene orders of all 118 vertebrate genomes in the database were scanned to determine the largest subset of taxa with identical gene order (including order, duplication, and presence/absence of the origin of light-strand replication and the D-loop). Primates and the European hedgehog are known to have faster rates of evolution than other vertebrates (Gissi *et al.* 2000), and most were therefore removed from the data set [additional fast-evolving taxa were removed in an additional test data set (see below), but results were qualitatively unchanged]. Other known fast-evolving taxonomic groups had already been removed due to gene rearrangements, and by visual inspection the remaining taxa do not appear to be evolving dramatically faster or slower than the cluster that contains them (Figure 2). The remaining 42 organisms served as the experimental data set (Table 1). This gene arrangement is considered "ancestral" because it includes taxa as divergent as mammals, sharks, and bony fish and also because alternative arrangements tend to be in small clusters and limited to specific vertebrate subgroups.

From the complete set of alignments, four data sets were created for analysis. For phylogenetic tree reconstruction, the ribosomal RNAs (rRNAs) and protein-

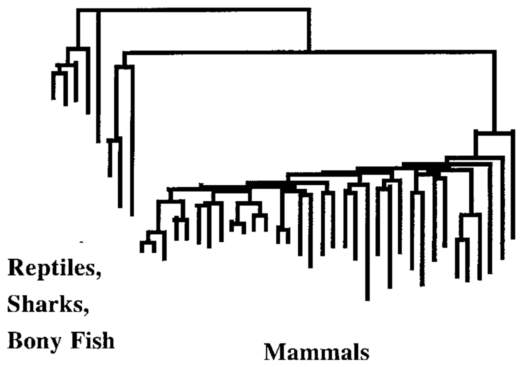

Figure 2.—Phylogenetic tree used in analysis. Visual analysis shows that taxa with exceptionally long branch lengths have been excluded. An expanded and fully labeled version of this tree is available as supplementary data (http://www.genetics.org/supplemental/).

coding regions were concatenated. This data set was 15,054 bp long. Although sites in these genes are probably evolving differently, and the average model is not necessarily accurate for each site, a reasonable joint model is not available and the benefits of a large data set would appear to outweigh concerns over model variations, particularly since the purpose of this analysis is discovery of new model complexity, not improved phylogenetic reconstruction. From the same set of 12 concatenated proteins, third codon positions were parsed out for conserved fourfold redundant sites (4×) and the two types of conserved twofold (2×) redundant sites, coding for either purines (2× *R*) or pyrimidines (2× *Y*). Redundancy at a specific codon in a single sequence is easily defined as the number of codons that code for the same amino acid. For example, a site coding for alanine would be fourfold redundant with the vertebrate mitochondrial genetic code, since there are four possible codons (all with the same nucleotides at the first and second codon positions) for this amino acid. We defined conserved redundancy in the alignment as those sites for which all codons in each sequence were in the same redundancy class. For example, the conserved fourfold redundant sites included those sites with any of the amino acids threonine, alanine, arginine, valine, glycine, or proline, but no other amino acids or gaps. This measure of redundancy class is preferable to that measured in a single genome because the site has probably been in the redundancy class over the entire evolutionary tree, whereas for a single genome a site may have recently changed its redundancy class, and therefore its base frequencies may not reflect the equilibrium values for that class. Sites coding for serine and leucine were considered sixfold redundant sites and not included, and methionine at the first amino acid position was also excluded. The 4×, 2× *R*, and 2× *Y* data sets were 583, 221, and 445 bp, respectively. Third codon positions from the different redundancy classes do not have variable selective constraints for protein function. The 4× sites are free to vary among all nucleotides, while the 2× sites are restricted to exchange between only two nucleotides [*i.e.*, purine to purine (2× *R*) or pyrimidine to pyrimidine (2× *Y*)]. The individual protein-coding regions were also sometimes considered separately for comparison to earlier results, but in these cases ATP8, ND3, and ND4L were discarded, as they contained too few sites to accurately estimate model parameters (see results for further consideration on tradeoffs among number of model parameters, amount of data, and comparative inference). ND6 was excluded from all redun-

TABLE 1

RefSeq accession numbers for genomes used in the analysis

Raja radiata (NC_000893)
Squalus acanthias (NC_002012)
Scyliorhinus canicula (NC_001950)
Mustelus manazo (NC_000890)
Ornithorhynchus anatinus (NC_000891)
Volemys kikuchii (NC_003041)
Rattus norvegicus (NC_001665)
Mus musculus (NC_001569)
Echinops telfairi (NC_002631)
Loxodonta africana (NC_000934)
Orycteropus afer (NC_002078)
Dasypus novemcinctus (NC_001821)
Ochotona collaris (NC_003033)
Sciurus vulgaris (NC_002369)
Oryctolagus cuniculus (NC_001913)
Pongo pygmaeus (NC_002083)
Echinosorex gymnura (NC_002808)
Talpa europaea (NC_002391)
Thryonomys swinderianus (NC_002658)
Cavia porcellus (NC_000884)
Chalinolobus tuberculatus (NC_002626)
Myoxus glis (NC_001892)
Soriculus fumidus (NC_003040)
Artibeus jamaicensis (NC_002009)
Pteropus scapulatus (NC_002619)
Pteropus dasymallus (NC_002612)
Ceratotherium simum (NC_001808)
Rhinoceros unicornis (NC_001779)
Equus asinus (NC_001788)
Equus caballus (NC_001640)
Felis catus (NC_001700)
Canis familiaris (NC_002008)
Halichoerus grypus (NC_001602)
Ovis aries (NC_001941)
Bos taurus (NC_001567)
Physeter catodon (NC_002503)
Balaenoptera musculus (NC_001601)
Balaenoptera physalus (NC_001321)
Dogania subplana (NC_002780)
Chelonia mydas (NC_000886)
Eumeces egregius (NC_000888)
Paralichthys olivaceus (NC_002386)

dancy class analyses since it is the only protein coded on the L strand.

Phylogenetic tree reconstruction: A neighbor-joining tree (NJ_{Tree1}) was calculated with PAUP* (Swofford 2000) from a 15,054-bp concatenated alignment using analytical distances calculated under the general time-reversible (GTR) model and empirical base frequencies (Lanave *et al.* 1984). NJ_{Tree1} was then used to obtain maximum-likelihood estimates for parameters in the GTR + Γ (general time reversible with rates gamma distributed across sites) model. A final neighbor-joining tree (NJ_{Tree2}) was calculated using the new parameters and maximum-likelihood distance estimates. It has been shown that substitution model parameters are relatively insensitive to errors in the phylogenetic tree estimate as long as the tree is approximately right (Yang *et al.* 1994; Sullivan *et al.* 1996), so we made no further efforts to revise the tree (Figure 2). All analyses were repeated on a similar tree of 40 taxa with the two longest individual branches in the mammals (*Loxodonta africana* and *Pongo pygmaeus*) removed. In addition to the taxon removal, this tree had some minor branch rearrangements, but results were qualitatively the same as the 42-taxon tree (data not shown). A more detailed version of these trees and all sequence alignments are available as supplementary data on the genetics web site (http://www.genetics.org/supplemental/) and at www.biology.lsu.edu/webfac/dpollock/asymmetry.html.

Calculation of statistics and parameters: Base frequencies and substitution rates were obtained with PAML v3.1 (Yang 1997) using the fixed topology and branch lengths of NJ_{Tree2} under the GTR model with empirical base frequencies. A nonreversible unrestricted model was also tried, but gave ambiguous results, probably due to overparameterization of the model. Although a reversible rather than a nonreversible model of substitution is almost universally used for this reason, this does not mean that the reversible model is a better reflection of biological reality. Complementary substitution rates, for example, are constrained to have an inverse relationship, so differences among regions should be interpreted as an increase in one substitution rate relative to the other. Substitution rates, the probabilities of substitution from each nucleotide i to nucleotide j, were calculated as $s_{ij} = \lambda_{ij}\pi_j$, where $\lambda_{ij} = \lambda_{ji}$ is the symmetric rate parameter governing exchange between the nucleotides, and π_j is the equilibrium frequency of the nucleotide mutated to. The HKY model (Hasegawa *et al.* 1985) was also run to obtain κ, the ratio of the transition rate parameter (α), and the transversion rate parameter (β).

For each of the redundancy class data sets, GC and AT skews were calculated using the formula of Perna and Kocher (1995),

$$\text{GC skew} = \frac{f_G - f_C}{f_G + f_C}$$

$$\text{AT skew} = \frac{f_A - f_T}{f_A + f_T}, \qquad (1)$$

where f_A, f_C, f_G, and f_T are the empirical frequencies of each nucleotide. In the absence of strand bias, f_G should equal f_C, and f_A should equal f_T due to base-pairing rules and the fact that the same process are by definition occurring on both strands (Sueoka 1962; Lobry 1996).

For a single gene, D_{ssH} (the duration of the single-strand state of the parental H strand) was calculated as in Tanaka and Ozawa (1994) and Reyes *et al.* (1998), using $D_{ssH} = (L - 2(\bar{x} - O_L))/L$ for ND1 and ND2, the two genes behind the origins of replication, and $D_{ssH} = 2((O_L - \bar{x})/L)$ for the remaining genes (Figure 1), where L is the length of the genome, O_L is the position of the light-strand origin of replication, and $\bar{x}$ is the average position of the gene. This calculation assumes that the movement of the replication forks is constant along the genome and equal for both strands. Position numbers began at the start of tRNA-Phe (the first position after the D-loop) and increased in the direction of H-strand synthesis during replication. For gene alignments, D_{ssH} was calculated for each gene and each site as the average of all organisms in the alignment. We obtained the same relative gene order with respect to D_{ssH} as Reyes *et al.* (1998): COX1 < COX2 < ATP8 < ATP6 < COX3 < ND3 < ND4L < ND4 < ND1 < ND5 < ND2 < CYTB.

To evaluate the effect of D_{ssH} continuously across the genome, the most likely model for both low and high D_{ssH} values was calculated, and comparative support for these two models was evaluated at each 4× site. The 4× sites with the 70 lowest and 70 highest D_{ssH} values were used to obtain parameters for the two extreme models. Maximum-likelihood estimates of model parameters were obtained using the fixed topology and branch lengths of NJ_{Tree2} under the GTR model with empirical base frequencies and using PAUP* (Swofford 2000) rather than PAML v3.1 (Yang 1997) as a matter of computational convenience. These parameter estimates were then fixed and for each model a maximum-likelihood (ML) analysis was run on the entire 4× data set. Likelihood values were calculated for each site under both models, and relative support for the two models at each site was then measured using Δ ln L, the difference in log-likelihood values for the two models. The difference in log-likelihood for the 70 lowest and 70 highest D_{ssH} sites evaluated separately and jointly was also calculated. The joint analysis provides the likelihood under the assumption that these two regions are evolving under the same model, and twice this Δ ln L statistic has an expected distribution of χ^2 with 9 d.f. A large Δ ln L indicates that the null hypothesis of a single model for both regions should be rejected, *i.e.*, that these two regions are probably evolving with different substitution rates.

It is of interest to know how much of the site-specific

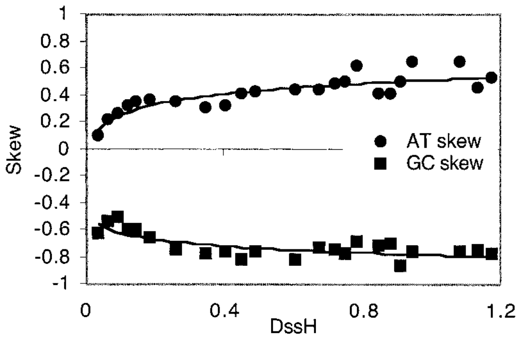

Figure 3.—Plot of AT and GC skews on D_{ssH} for 25- to 26-bp partitions of the 4× redundant sites. The logarithmic regression lines are also shown.

support derives from rate parameters and how much derives from base frequencies. To address this, two further sets of models were obtained: the first used the rate parameters from the previously calculated high and low D_{ssH} sets, but used base frequencies obtained from an ML analysis of the entire 4× data set (6 d.f. between independent and joint models); the second, conversely, used base frequencies from the high and low sets, but rate parameters from the entire data set (3 d.f. difference between independent and joint models). Site-specific support was evaluated using Δ ln L, as before.

RESULTS

Skew and DssH: In agreement with the results of Reyes *et al.* (1998), a regression of AT skew on D_{ssH} for third codon positions at 4× redundant sites from each protein-coding gene was significant ($R^2 = 0.748$; $P = 0.003$), but we found that the relationship between GC skew and D_{ssH} was not significant ($R^2 = 0.279$; $P = 0.144$). Since each gene is of variable length and contains variable numbers of 4× redundant sites, we also divided the 4× redundant sites into 23 partitions of ~25 sites each. For these partitions, a linear regression for AT skew is also significant ($y = 0.295x + 0.242$, $R^2 = 0.687$; $P < 0.001$), as is the GC skew ($y = 0.156x - 0.631$, $R^2 = 0.373$; $P = 0.001$). These data points may be distributed on a curved rather than a straight line, however, and logarithmic curves (Figure 3) give better fits in both cases [for the AT curve, $y = 0.113 \ln(x) + 0.513$, $R^2 = 0.750$, $P < 0.001$; for the GC curve, $y = -0.069 \ln(x) - 0.782$, $R^2 = 0.581$, $P < 0.001$)]. Although the GC skew appears to fall early on with increasing D_{ssH} in both the gene-by-gene graph and the equal-partition graph, the first 25 4× sites already have a GC skew of −0.623, even though these sites spend relatively little time in the single-strand state. The last 25 4× sites have a GC skew of −0.766, which is close to −1.0, the largest possible negative skew.

TABLE 2

Number of sites in data subsets

	No. of sites		
Gene	2× *R*	2× *Y*	4×
COI	51	113	147
COII	25	40	34
ATP6	10	17	34
COIII	25	53	55
ND4	24	37	60
ND1	26	34	56
ND5	37	51	75
ND2	19	11	31
CYTB	29	66	74
Mean	27.3	46.9	62.9
SD	11.5	30.2	35.6

Substitution rates: Evaluation of changes in substitution rates can establish which types of substitution are most prevalent in the single-strand state. Since it has been hypothesized that transitions are most important, it is useful to analyze the 2× redundant sites in addition to the 4× sites, since transversions at the 4× sites can potentially confound analysis of transition rates. Linear regression of substitution rates on D_{ssH} for third codon positions in each gene (Table 2) were performed for 4×, 2× *R*, and 2× *Y* sites (see Table 3 for slopes, R^2, and rates). For the 4× sites, the TC/CT (Figure 4) and AT/TA regressions are most significant, while nearly all regressions involving G are not (*e.g.*, AG/GA, Figure 4). In a strikingly similar pattern, the regressions of substitution rates for 2× *Y* sites were significant, while for the 2× *R* sites they were not (Figure 5). Notably, the slopes of increase in pyrimidine transitions are extremely similar for both 4× and 2× *Y* sites. A regression of the transition/transversion ratio, $\kappa = \alpha/\beta$, on D_{ssH} shows a small but significant increasing trend with D_{ssH} ($y = 1.94x + 5.36$, $R^2 = 0.545$, $P = 0.023$; Figure 6).

It is worth noting here that there are explicit tradeoffs in this analysis among the number of parameters being estimated, the size of the data sets, and the accuracy of parameter estimates. When estimating skew, which is dependent only on the four equilibrium frequencies (three independent parameters), we were able to obtain accurate estimates for 23 4× data sets of only 25 sites each. For estimating substitution rates at the 4× sites, there are another six parameters for the reversible model, and the gene average of 63 sites (Table 2) is more appropriate. Fewer sites are necessary (and fewer are available) for the 2× sites, since only one free equilibrium frequency parameter and one rate parameter are in the 2× models. With a greater number of densely sampled taxa, we might be able to accurately sample smaller gene regions, and if the samples were more closely related there might be more sites in the conserved redundancy classes. It is clearly critical that the topology and branch

TABLE 3

Linear regression of substitution rates on D_{ssH} for 4×, 2× *R*, and 2× *Y* sites

	Slope		R^2		Significance	
s_{ij}	4×	2×	4×	2×	4×	2×
s_{AG}	−0.0247	−0.0286	0.106	0.368	0.392	0.083
s_{GA}	6.13E-02	1.424	0.058	0.270	0.533	0.152
s_{CT}	−0.211	−0.210	0.793	0.810	0.001***	0.001***
s_{TC}	0.269	0.438*	0.724	0.765	0.004***	0.002***
s_{AC}	0.116	—	0.581	—	0.017*	—
s_{CA}	0.145	—	0.322	—	0.111	—
s_{AT}	−0.120	—	0.674	—	0.007**	—
s_{TA}	0.192	—	0.603	—	0.014*	—
s_{CG}	−0.0067	—	0.029	—	0.664	—
s_{GC}	0.0897	—	0.064	—	0.513	—
s_{GT}	0.101	—	0.130	—	0.340	—
s_{TG}	0.0484	—	0.278	—	0.144	—

*$P < 0.05$; **$P < 0.01$; ***$P < 0.005$.

lengths in these analyses are obtained from the entire genome data set and not reestimated for these small data sets, since the addition of so many more free parameters would greatly reduce the precision of substitution rate estimates.

Site-specific analyses: A regression of relative site-specific support for the two most extreme models (derived from the 70 lowest and highest D_{ssH} values) was extremely significant, despite a great deal of noise in the plot ($y = -1.62x + 0.335$, $R^2 = 0.115$, $P < 0.001$; Figure 7). The slope was still extremely significant for a regression on the middle points, with the first 70 and last 70 points removed ($y = -1.63x + 0.261$, $R^2 = 0.073$, $P < 0.001$). Separate analyses of models with divergent base frequencies only, or divergent rate parameters only, showed that most of the significance was in the equilibrium frequencies ($y = -1.38x + 0.312$, $R^2 = 0.098$, $P < 0.001$) rather than in the rate parameters ($y = -0.122x + 0.057$, $R^2 = 0.023$, $P < 0.001$; Figure 8). Twice the difference in log-likelihood scores for the smallest and largest 70 sites evaluated both separately and together was 69.03. With only nine free parameters different between these two nested models, the probability that these two sets of sites evolve under the same model is <0.001.

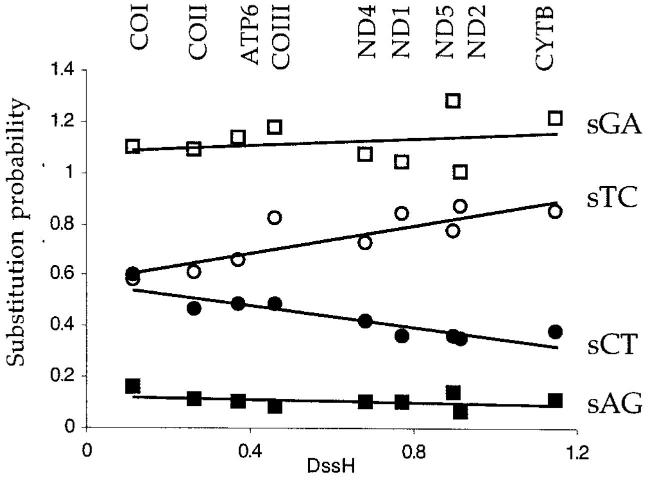

FIGURE 4.—Plot of gene-specific AG, GA, CT, and TC substitution rates on D_{ssH} for 4× redundant sites. These are the substitution rates for the light strand, but single-strand mutations occur on the complementary heavy strand. The linear regression lines are also shown.

DISCUSSION

Our analysis indicates that for vertebrates with the "ancestral" gene order and replication process, mutation patterns on the heavy strand during mtDNA replication lead to a strong and consistent increasing gradient of T → C substitutions on the light strand (Figure 9). This is consistent with the logarithmic increase of the AT skew with time in the single-strand state (D_{ssH}), and a similar linear increase in estimated rates of T → C substitutions is seen for both 4× and 2× *Y* redundant sites. The conserved 2× *Y* sites provide particularly strong evidence, as there is no possibility that this relationship is confounded by variable or unequal transversion rates. In COI, the gene with the smallest D_{ssH} value, there is no asymmetric bias in T → C over C → T substitutions for either the 4× or the 2× *Y* sites, indicating that the development of such a bias in genes with higher D_{ssH} values is entirely attributable to time spent in the single-strand state. The linear increase in κ, the ratio of transition to transversion rate parameters, is also consistent with this interpretation, although it does not exclude a possible gradient in any of the transversion substitutions. The significant linear regression of relative site-specific support for the extreme high and

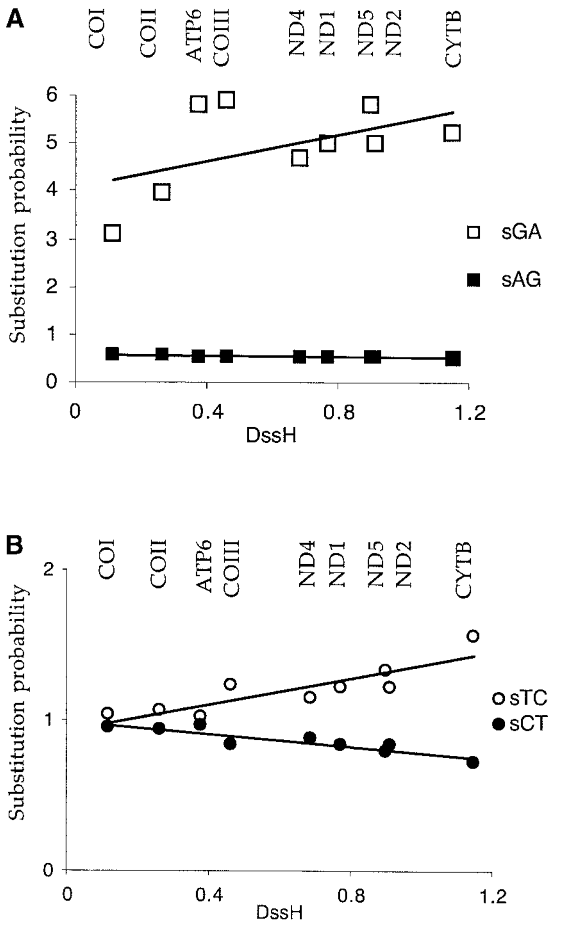

FIGURE 5.—Plot of gene-specific substitution rates on D_{ssH} for 2× redundant sites. AG and GA substitution rates are shown for 2× *R* sites (A), while TC and CT substitution rates are shown for 2× *Y* sites (B). The linear regression lines are also shown.

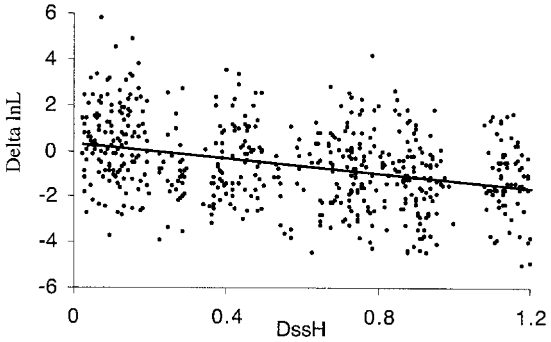

FIGURE 7.—Plot of contrasting site-specific support in 4× redundant sites for extreme models of evolution *vs.* D_{ssH}. Relative support for each model is measured as Δln *L*, the difference in the log-likelihood of the data at the site for each model. The linear regression line is also shown.

low D_{ssH} value models also lends support to a continuous gradient of mixed substitution rates with changing D_{ssH}.

It has been hypothesized that C → U mutations on the heavy strand are the most common mutation in the single-strand state, which should lead to an even stronger gradient of G → A than T → C substitutions on the light strand (REYES *et al.* 1998). However, we find little evidence for a linear gradient of G → A substitutions in either the 4× or the 2× *R* sites. In both data sets we instead see a strong but nearly constant bias in favor of G → A substitutions over A → G substitutions. This bias is larger than the bias of T → C to C → T substitutions in 4× sites at all points in the genome.

Given the empirical evidence for up to 200-fold higher rates of cytosine deamination in the single-strand *vs.* double-strand states (SANCAR and SANCAR 1988; FREDERICO *et al.* 1990; IMPELLIZZERI *et al.* 1991), we interpret these data as indicating that excess cytosine deaminations occur even for the smallest D_{ssH} values and that the rate of substitution of these mutations quickly reaches a plateau (Figure 9). It is plausible on the basis of these data that D_{ssH} values of zero already have an asymmetric substitution process, which could be due to the short period of time spent in the single-strand state after the passing of the heavy-strand replication fork,

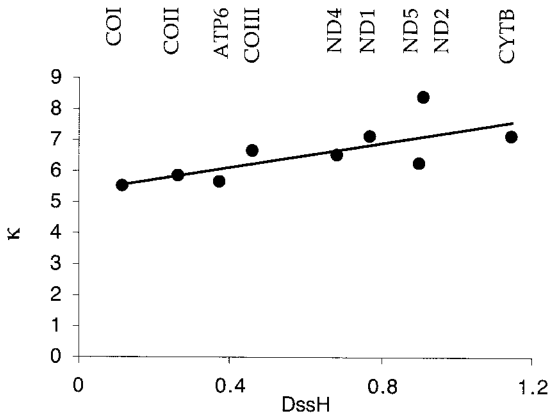

FIGURE 6.—Plot of gene-specific transition-to-transversion ratios on D_{ssH} for 4× redundant sites. The linear regression line is also shown.

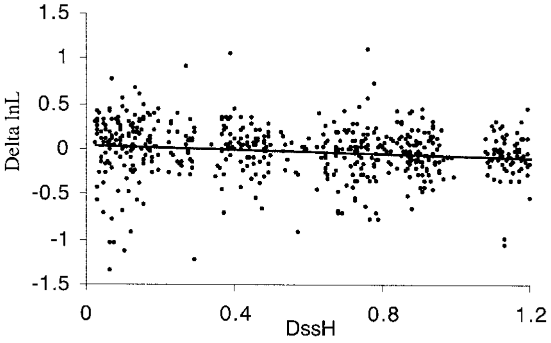

FIGURE 8.—Plot of contrasting site-specific support on D_{ssH} in 4× redundant sites for two models of evolution with divergent rate parameters but the same base frequencies. Relative support for each model is measured as Δln *L*, the difference in the log-likelihood of the data at the site for each model. The linear regression line is also shown.

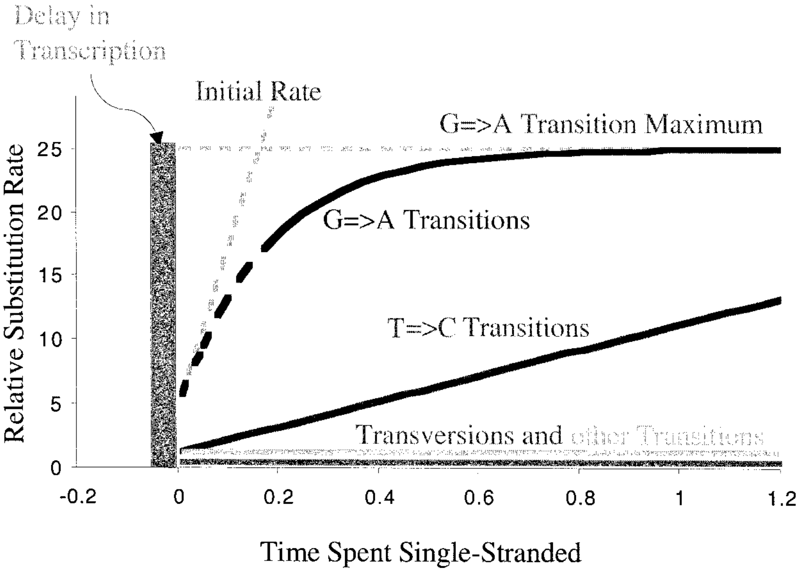

FIGURE 9.—Diagram summarizing the interpretation of relative substitution rates *vs.* time spent single stranded. The time spent single stranded is measured in units corresponding to the time required for polymerase gamma to traverse the length of the genome, and the scale for relative substitution rates is arbitrary. For consistency with the text, the types of transition refer to the light strand (coding direction), although the asymmetric mutations occur on the single-strand heavy strand. The transversion rates are depicted as a horizontal line at the bottom since they are much slower than transitions, and there is no clear evidence that they change along the genome. The T → C transitions are depicted as increasing linearly and intersecting the *y*-axis near 1.0, indicating that an excess of this type of transition does not occur prior to initiation of light-strand replication. The G → A transitions are depicted as starting out with a linear increase, but rapidly saturating as a theoretical maximum rate is approached. The ratio of the initial G → A and T → C rates of increase is a reflection of the relative mutation propensity of the complementary nucleotides on the single-strand heavy strand, and both of these rates should be proportional to the inverse of the rate of polymerization by gamma polymerase. The maximum rate of G → A transitions is hypothesized to be reached according to the time required to protect the single-strand DNA with mtSSB protein. The *y*-intercept for both increasing transition types should be affected by both the rate of polymerization and the time required to initiate light-strand replication after the passing of the heavy-strand replication fork. The other transition rates (A → G and C → T) are assumed to be constant and depicted at slightly higher level than the transversions.

but before initiation of light-strand replication (Figure 9). The presence of asymmetric substitutions at zero $D_{\rm ssH}$ values would otherwise have to be explained by some other mechanism, such as codon bias or time spent in the single-strand state during transcription (see REYES *et al.* 1998 for a review of other possibilities). In transcription, however, both strands enter the single-strand state, and deamination should occur more often in the nontranscribed strand (FRANCINO and OCHMAN 1997; REYES *et al.* 1998). This is not in agreement with our data, since C → T substitution rate biases are not detectable in the nontranscribed strand. An analysis of conserved 4× redundant sites in ND6, the only gene transcribed on the light strand, gave base frequencies of 0.1701 for A, 0.0527 for C, 0.2942 for G, and 0.4830 for T on the light strand, meaning that AT skew on the heavy strand was 0.479, while the GC skew was −0.69. ND6 has a $D_{\rm ssH}$ of 1.04, intermediate between ND5 and CYTB, and its AT skew is compatible with this $D_{\rm ssH}$ in the absence of any effect of transcription. The GC skew is slightly less negative than the average of heavy-strand-encoded genes with similar $D_{\rm ssH}$ values (−0.758), which may be due to sampling error or may indicate a small effect due to transcription, but if so the change is in a direction opposite to expectation. Thus, both transcription and codon bias due to selection play only a minor role, if any, in producing the gradients of asymmetric bias at redundant sites.

A plateau in the rate of light-strand G → A substitutions combined with a linear increase in light-strand T → C substitutions can be explained by either a plateau in the heavy-strand C → U mutation rate or a selective constraint against loss of too many G's in the light strand. A plateau in the C → U mutation rate could be plausibly associated with the coating of the single strand with mtSSBs [which occurs in ~5 min (SANCAR and SANCAR 1988)], such that the initial rapid increase is due entirely to the time spent single stranded prior to completion of the coating. If so, adenine-to-hypoxanthine (hX) deaminations in the single-strand state, leading to T → C substitutions on the light strand, apparently continue unabated in the face of this protection against further C → U deaminations. If selection plays a role, the constraint against loss of G's could be either a general selective pressure on the frequency of G's or a specific pressure against loss of G at specific sites. In the latter case, the G-selected sites would evolve more slowly than unrestricted sites, and therefore the slowest sites should be more heavily biased toward G. We tested this by running a GTRγ model with 100 rate categories and fixed $\mathrm{NJ}_{\rm Tree2}$ on the 4× sites, selecting the 20 sites with the lowest posterior rate probability and evaluating their base composition. The percentage of G content at these sites was 3.9 as opposed to 4.8% G content for all 4× sites combined. Thus, there is no support for site-specific constraints on selection for G in the 4× sites. A linear regression of posterior rate probabilities was also calculated against $D_{\rm ssH}$, but no significant trend was found (slope = −0.11, R^2 = 0.002, P = 0.284). There is a great deal of noise in this analysis, but this result is compatible with the hypothesis that the G → A substitutions have saturated early on (Figure 9).

It is worth noting that in the skew plots and in the κ plots, and to a lesser extent in the substitution rate plots, points attributable to ND1 and ND2 appear slightly misplaced. These deviations from the rest of the genes could be corrected by a shift in their D_{ssH} to slightly higher values. An important assumption in the calculation of D_{ssH} is that the length of time in the single-strand state is well predicted by the length of the genome that the two replication forks must traverse and thus that the two replication forks travel at equal and constant speeds. A misplacement of ND1 and ND2 with these calculations might be caused by either a slower rate of light-strand replication or a slowdown of light-strand replication as it crosses secondary structure in the rRNA-coding and control regions.

It appears from this analysis that variation in different kinds of DNA substitutions along the genome can be directly linked to functional aspects of the replication system in vertebrate mitochondria. Although we attempted to determine the ancestral vertebrate substitution process in this study, it is exciting to consider that variation in substitution patterns during the course of evolution might be interpretable in terms of changes in the function of particular proteins, rather than uninterpretable alterations in the average nucleotide frequencies that result. We removed the taxa with the most potential to have changed from the ancestral evolutionary pattern, but there is still no guarantee that these patterns do not change within our data set. With this model, differences in the slope and intercept of the T → C substitution patterns (on the light strand), along with the initial rate of increase and saturation level of the G → A substitutions (Figure 9), can be linked to changes in the rate of replication, the number of replications per generation in the germ line, the efficiency of light-strand replication initiation, the functionality of protection offered by the mtSSB protein, and the location of the origins of replication.

Having definitive knowledge of the evolutionary behavior at individual sites can allow large improvements in phylogenetic inference (Pollock and Bruno 2000), so a correct mechanistic model based on these patterns that allows for the continuous nature of substitution rate differences along the genome is likely to enhance phylogenetic reconstruction dramatically. When a model based on this analysis becomes available, it should be possible to evaluate gradients in individual substitution rates along the entire genome simultaneously, leading to greater accuracy and fewer parameters overall. While this manuscript was in review, a model was published that allows for a symmetric substitution process, with asymmetric deviations to be experimentally added one at a time (Bielawski and Gold 2002). Such a model, combined with continuous gradients along the genome, may allow accurate estimates of parameters with much smaller data sets than those used here and should be an important component of codon-based models of protein evolution.

While the vertebrate mitochondrial genomes are the only taxonomic group with a large set of densely sampled complete genomes with the same gene order, it would be of great interest if new genomes that allow comparative analysis of changes in the mitochondrial replication process become available. Comparison of densely sampled data sets from more closely related organisms will allow evaluation of the degree to which such changes occur. On the basis of current rates of mitochondrial genome sequencing (the number of available complete vertebrate mitochondrial genomes has approximately doubled since this study was initiated), useful data sets are likely to appear in the near future, although current sampling strategies tend to emphasize sampling divergent lineages. We hope that this study will help motivate a focus on higher-density sampling of closely related organisms for the express purpose of more accurately analyzing substitution processes across the genome.

We thank Marly Dows and Carla Haslauer for assistance with data preparation and manipulation. We thank Frank Burbrink, Jim McGuire, Mohamed Noor, and two anonymous reviewers for constructive comments on the manuscript. D.D.P. and J.J.F. were supported by a research competitiveness subprogram grant from the State of Louisiana [LEQSF (2001-04)-RD-A-08], and D.D.P. was also supported by the State of Louisiana's Millennium Research Program (Biological Computation and Visualization Center) and by National Institutes of Health grant GM-65612.

LITERATURE CITED

Ames, B. N., M. K. Shigenaga and T. M. Hagen, 1995 Mitochondrial decay in aging. Biochim. Biophys. Acta **1271:** 165–170.

Asakawa, S., Y. Kumazawa, T. Araki, H. Himeno, K. Miura *et al.*, 1991 Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. J. Mol. Evol. **32:** 511–520.

Bielawski, J. P., and J. R. Gold, 2002 Mutation patterns of mitochondrial H- and L-strand DNA in closely related Cyprinid fishes. Genetics **161:** 1589–1597.

Clayton, D. A., 2000 Transcription and replication of mitochondrial DNA. Hum. Reprod. **15:** 11–17.

Delorme, M. O., and A. Henaut, 1991 Codon usage is imposed by the gene location in the transcription unit. Curr. Genet. **20:** 353–358.

Francino, M. P., and H. Ochman, 1997 Strand asymmetries in DNA evolution. Trends Genet. **13:** 240–245.

Frederico, L. A., T. A. Kunkel and B. R. Shaw, 1990 A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29:** 2532–2537.

Gissi, C., A. Reyes, G. Pesole and C. Saccone, 2000 Lineage-specific evolutionary rate in mammalian mtDNA. Mol. Biol. Evol. **17:** 1022–1031.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Holt, I. J., H. E. Lorimer and H. T. Jacobs, 2000 Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. Cell **100:** 515–524.

Impellizzeri, K. J., B. Anderson and P. M. Burgers, 1991 The spectrum of spontaneous mutations in a Saccharomyces cerevisiae uracil-DNA-glycosylase mutant limits the function of this enzyme to cytosine deamination repair. J. Bacteriol. **173:** 6807–6810.

Jermiin, L. S., D. Graur, R. M. Lowe and R. H. Crozier, 1994 Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. J. Mol. Evol. **39:** 160–173.

Jermiin, L. S., D. Graur and R. H. Crozier, 1995 Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. Mol. Biol. Evol. **12:** 558–563.

Lanave, C., G. Preparata, C. Saccone and G. Serio, 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20:** 86–93.

Limaiem, J., and A. Henaut, 1984a Demonstration of a sudden change in the use of codons in the vicinity of transcription termination. Crit. Rev. Acad. Sci. III **299:** 275–280.

Limaiem, J., and A. Henaut, 1984b Fluctuation of the incidence of the 4 bases along the mitochondrial genome of mammals using correspondence factorial analysis. Crit. Rev. Acad. Sci. III **298:** 279–286.

Lindahl, T., 1993 Instability and decay of the primary structure of DNA. Nature **362:** 709–715.

Linnane, A. W., S. Marzuki, T. Ozawa and M. Tanaka, 1989 Mitochondrial DNA mutations as an important contributor to ageing and degenerative diseases. Lancet **1:** 642–645.

Lobry, J. R., 1996 Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:** 660–665.

Perna, N. T., and T. D. Kocher, 1995 Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J. Mol. Evol. **41:** 353–358.

Pollock, D. D., and W. J. Bruno, 2000 Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. **17:** 1854–1858.

Pollock, D. D., J. A. Eisen, N. A. Doggett and M. P. Cummings, 2000 A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. Mol. Biol. Evol. **17:** 1776–1788.

Pruitt, K. D., K. S. Katz, H. Sicotte and D. R. Maglott, 2000 Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. **16:** 44–47.

Reyes, A., C. Gissi, G. Pesole and C. Saccone, 1998 Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. **15:** 957–966.

Sancar, A., and G. B. Sancar, 1988 DNA repair enzymes. Annu. Rev. Biochem. **57:** 29–67.

Shadel, G. S., and D. A. Clayton, 1997 Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. **66:** 409–435.

Sueoka, N., 1962 On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48:** 582–592.

Sullivan, J., K. E. Holsinger and C. Simon, 1996 The effect of topology on estimates of among-site rate variation. J. Mol. Evol. **42:** 308–312.

Swofford, D. L., 2000 *Phylogenetic Analysis Using Parsimony (*and Other Methods).* Sinauer Associates, Sunderland, MA.

Tanaka, M., and T. Ozawa, 1994 Strand asymmetry in human mitochondrial DNA mutations. Genomics **22:** 327–335.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Wallace, D. C., 1999 Mitochondrial diseases in man and mouse. Science **283:** 1482–1488.

Wolstenholme, D. R., 1992 Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. **141:** 173–216.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

Yang, Z., N. Goldman and A. Friday, 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11:** 316–324.

Communicating editor: J. Hein