

# Partitioning the Variation in Mammalian Substitution Rates

Nick G. C. Smith\* and Adam Eyre-Walker†

\*Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Sweden; and †Centre for the Study of Evolution & School of Biological Sciences, University of Sussex, Brighton, U.K.

We have used analysis of variance to partition the variation in synonymous and amino acid substitution rates between three effects (gene, lineage, and a gene-by-lineage interaction) in mammalian nuclear and mitochondrial genes. We find that gene effects are stronger for amino acid substitution rates than for synonymous substitution rates and that lineage effects are stronger for synonymous substitution rates than for amino acid substitution rates. Gene-by-lineage interactions, equivalent to overdispersion corrected for lineage effects, are found in amino acid substitutions but not in synonymous substitutions. The variance in the ratio of amino acid and synonymous substitution rates is dominated by gene effects, but there is also a significant gene-by-lineage interaction.

## Introduction

We know that molecular evolution does not proceed at a constant rate. Some genes and lineages consistently evolve faster or slower than others: for example, histone genes evolve very slowly in most taxa, and most genes evolve faster in rodents than in primates (Li 1997). But in addition to these gene and lineage effects, there is rate variation which is specific to certain genes in certain lineages; for example, the growth hormone gene has gone through a very rapid burst of evolution in ruminant artiodactyls and primates but not in other mammalian lineages (Wallis 1994). Such gene-by-lineage effects are equivalent to an overdispersion of the molecular clock when lineage effects have been removed (Gillespie 1989).

In mammals we have good evidence of all three types of evolutionary rate variation (Li 1997), but as yet we have no quantitative data on the relative contributions of gene, lineage, and gene-by-lineage effects. We have developed a method to solve this problem by using an analysis of variance (ANOVA) in which we partition the variance due to each effect and test whether the effects are statistically significant. The basic approach is as follows. We take orthologous genes from a series of lineages which either form a star phylogeny or form independent pairs of taxa (fig. 1); in the latter case we need to know the time of divergence for each pair of taxa to partition the variance correctly. In ANOVA we need an estimate of the error variance. This could potentially be estimated analytically from the formulae used to estimate the branch lengths; however, this is complex, so we have elected to estimate the error variance empirically by dividing each gene into its odd- and even-numbered codons. This yields two estimates of the substitution rate for each combination of lineage and gene. We calculate the branch lengths for each half of the gene and divide the branch length by the time of divergence if required (i.e., if we have independent pairs of lineages).

ANOVA partitions the variance according to an additive model, but models of molecular evolution are more naturally expressed in multiplicative terms; for ex-

ample, under a strict neutral model of molecular evolution, the rate of substitution is  $uf$ , where  $u$  is the nucleotide mutation rate and  $f$  the proportion of mutations which are neutral. We therefore took the logarithm of each branch length; this converts the model  $R_{ij} = RG_iL_jO_{ij}$ , where  $R$  is the geometric mean of the substitution rates across genes and lineages,  $G_i$  is the gene effect for the  $i$ th gene,  $L_j$  is the lineage effect for the  $j$ th lineage, and  $O_{ij}$  is the gene-by-lineage interaction for the  $i$ th gene in the  $j$ th lineage, to  $\log(R_{ij}) = \log(R) + \log(G_i) + \log(L_j) + \log(O_{ij})$ . We therefore have a two-way ANOVA with two observations for each combination of lineage and gene (fig. 1). The variance can be partitioned and tested for significance using a model II ANOVA.

We have applied our new method to nuclear and mitochondrial sequence data from mammals. For the nuclear data set we analyzed the variation in both amino acid and synonymous substitution rates using four taxa which form an approximate star phylogeny. But the ratio of synonymous to amino substitution rates in mitochondria is such that we had to analyze different data sets; for the amino acid substitution rate we analyzed seven taxa which form an approximate star phylogeny, and for the synonymous substitution rate we analyzed four independent pairs of taxa.

## Materials and Methods

### Mitochondrial Genes

For the mitochondrial data sets we analyzed the protein-coding genes from the complete mitochondrial genomes. The four pairs of taxa we used to analyze the variation in mitochondrial synonymous substitution rates were (with complete mitochondrial accession numbers in parentheses) pygmy chimp (D38116) and common chimp (D38113), horse (X79547) and donkey (X97337), grey seal (X72004) and harbor seal (X63726), and blue whale (X72204) and fin whale (X61145). For the analysis of variation in the rate of amino acid substitution, we used sequence data from seven mammalian species chosen such that they approximated a star phylogeny: cow (J01394), human (J01415), mouse (L07095), rabbit (AJ001588), African elephant (AJ224821), aardvark (Y18475), and nine-banded armadillo (Y11832). Alignments were taken from the AMntDB database (Lanave et al. 2000), and short genes were removed from consid-

Key words: molecular clock, overdispersion, substitution rate, ANOVA.

E-mail: a.c.eyre-walker@sussex.ac.uk.

*Mol. Biol. Evol.* 20(1):10–17. 2003

© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

eration, leaving nine genes: *atp6*, *co1*, *co2*, *co3*, *cytb*, *ndl*, *nd2*, *nd4*, and *nd5*.

### Nuclear Genes

We used the ACNUC (Gouy et al. 1984) and HOVERGEN (Duret, Mouchiroud, and Gouy 1994) databases to search for orthologous genes in the following four mammalian species: human, cow, mouse, and rabbit. Coding sequences were extracted using the EMBOSS package at HGMP (<http://www.hgmp.mrc.ac.uk/>), and DNA sequence alignment was performed on the basis of protein alignments using the default settings of the CLUSTALW (Thompson, Higgins, and Gibson 1994) multiple alignment program, as implemented in the DAMBE package (<http://web.hku.hk/~xxia/software/software.htm>). Because it is important to confirm orthology, we performed an additional check of the phylogeny using the PAML package (Yang 1997), which was used to test the three alternative four-species unrooted trees using a codon model of sequence evolution and the Shimodaira and Hasegawa (1999) test with multiple-comparison correction. For only one gene was the true species tree of [[human, cow], [mouse, rabbit]] rejected, leaving 29 genes. Accession numbers are available on request from the authors.

### Times of Divergence

To successfully partition the substitution rate between gene, lineage, and gene-by-lineage effects, we need to express the substitution rates in terms of substitution per unit time. For the nuclear genes and one of the mitochondrial DNA data sets, this is straightforward because the taxa in each case form an approximate star phylogeny; the divergences for each taxa are therefore over similar time scales. In contrast, in the case of the data set used to analyze mitochondrial synonymous substitution rates, we analyzed four independent pairs of taxa in which the taxa have been separated for different periods of time. We obtained the following times of divergence from Pesole et al. (1999), who inferred times of divergence from locally calibrated molecular clocks (chimps), biogeographic data (seals), and fossils (whales, horses-donkeys): chimps 2.5 Myr, seals 2.7 Myr, whales 5 Myr, horse-donkey 4 Myr.

### Substitution Rate Estimation

Rates of synonymous and amino acid substitution were estimated using the PAML package (Yang 1997). Amino acid substitution rates,  $K_a$ , were estimated from the amino acid sequences of the genes, using the Poisson model of amino acid substitution, with variation among sites modeled by a discretized gamma function (similar results were obtained assuming no variation in the rate of substitution across sites). Synonymous substitution rates,  $K_s$ , were estimated using a model of codon evolution which accounted for the transition-transversion rate bias, with codon usage bias modeled by the nucleotide frequencies at the three codon positions and with a different ratio of nonsynonymous to synonymous sub-

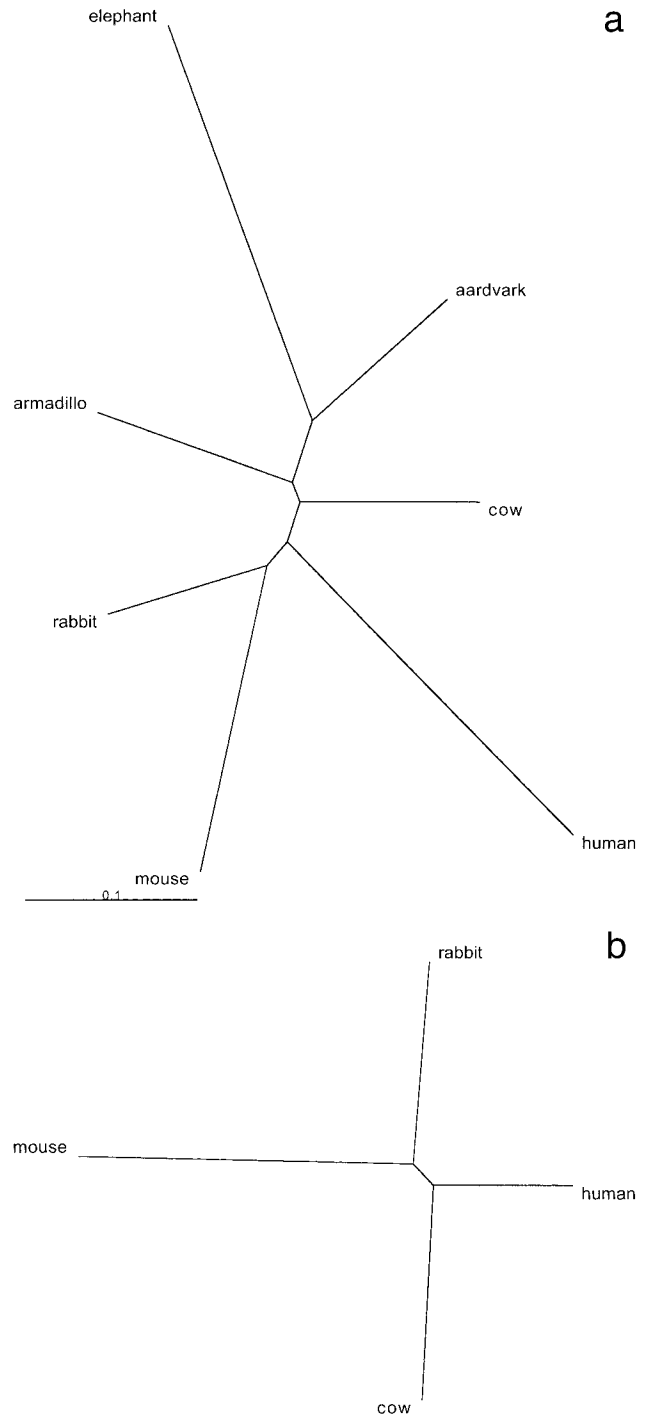


FIG. 1.—The phylogenies used for the mitochondrial  $K_a$  analysis (a) and the nuclear  $K_a$  and  $K_s$  analyses (b). The branch lengths are the averages across genes. In figure 1b the branch lengths are for  $K_s$ . A very similar tree, in terms of the proportion of the total tree length in the internal branch, is obtained using  $K_a$ . The likely root is between the cow and armadillo branches for the mitochondrial  $K_a$  tree and on the internal branch for the nuclear analyses.

stitution rates allowed down each lineage. To reduce the effect of changes at the amino acid level on synonymous site evolution, we restricted the analysis of  $K_s$  to codons for which the first two codon positions were invariant across all species.

Mitochondrial  $K_s$  rates were estimated by pairwise comparison between each of the four closely related species pairs, and mitochondrial  $K_a$  rates were estimated by analysis of the seven species alignments, with the tree assumed to be [[mouse, rabbit], [human, cow]], [[elephant, armadillo], [human, cow]], which is consistent with recent mammalian phylogenies (Madsen et al. 2001; Murphy et al. 2001). Nuclear  $K_s$  and  $K_a$  rates were estimated by analysis of the four species alignments, with the tree assumed to be ([human, cow], [mouse, rabbit]). Only the terminal branch lengths were included in the mitochondrial  $K_a$  analysis and the nuclear gene analyses (i.e., internal branch lengths were not used). We did this for two reasons: first, the deepest branch cannot be partitioned without the addition of an out-group taxon; second, the internal branches make distance estimates from different taxa nonindependent, for example, the divergence from mouse to the root of the mitochondrial  $K_a$  tree shares several internal branches with the divergence from rabbit to the root.

The sharing of lineages is not the only potential source of nonindependence; even the terminal branches are not independent of one another because they are all dependent on every sequence in the data set (excluding data sets in which we analyze pairs of independent taxa); for example, if we have three sequences for which the pairwise divergences are  $D_{12}$ ,  $D_{13}$ , and  $D_{23}$ , the individual branch lengths are calculated using all three pairwise comparisons, e.g.,  $(D_{12} + D_{13} - D_{23})/2$ . As Bulmer (1989) has shown, this nonindependence generates negative covariances between branches, which ultimately manifests itself as overdispersion. In our method this covariance is expected to be automatically partitioned into the error term; we examine this issue in our generic simulations below.

To check our assumption of a star phylogeny, we calculated the average distances along all branches in our two data sets. These are shown in figure 1; we only give the  $K_s$  tree for the nuclear genes because the  $K_a$  tree is very similar in terms of the proportion of the total tree length which is in the internal branch. As expected, it is evident that the internal branches in each tree are much shorter than the terminal branches. It should also be noted that the statistical significance of the gene effect and the gene-by-lineage interaction are independent of the star assumption; we can divide the divergence down any branch by an arbitrary constant with no effect on the probability values. Of course, the significance of the lineage effect is affected, as are the estimates of the variance components.

### The ANOVA of Substitution Rates

To test each effect (gene, lineage, and gene-by-lineage interaction) and to partition the variance of the substitution rates, we used a model II, two-way ANOVA with replication. We tested the interaction term by an  $F$ -test of the interaction mean sum-of-squares (MS) against the error MS. The gene and lineage effects were tested against the interaction MS; we did not attempt to combine the error and interaction MS. This did not affect

our results qualitatively. The variance was partitioned according to the following formulae:

$$V_{error} = MS_{error}$$

$$V_{interaction} = (MS_{interaction} - V_{error})/2$$

$$V_{gene} = (MS_{gene} - V_{error} - 2V_{interaction})/(2l)$$

$$V_{lineage} = (MS_{lineage} - V_{error} - 2V_{interaction})/(2g)$$

where  $g$  is the number of genes and  $l$  the number of lineages.

## Results

### Generic Simulations

Estimating the error associated with the estimate of the substitution rate empirically should overcome some of the problems which have been encountered in the analysis of overdispersion in the past, namely, the variance associated with correcting for multiple hits and the negative covariance between branch lengths. But there are two central assumptions in ANOVA—it is assumed that the error term is normally distributed and that the variance of the error term is homogeneous across the analysis (homoscedasticity). To investigate the first of these issues, we simulated DNA sequence evolution using PAML (Yang 1997) according to the Jukes-Cantor model under a number of different parameters. In each case we assumed that the sequences formed a star phylogeny and that each gene was 500 nucleotides in length. We varied the number of genes (10 or 30), the number of lineages (3, 4, or 10), and the level of divergence (0.1, 0.5, and 1.0 along each branch). We generated 1,000 simulated data sets for each parameter combination. We split each gene in half and performed an ANOVA on the logarithm of the distances, which were estimated using PAML assuming a star phylogeny and no molecular clock.

Table 1a shows the proportion of each set of the 1,000 simulated data sets in which the gene, lineage, or interaction effects were significant at 5%. Over all parameter combinations the level of type I error is as expected for the lineage effect but is highly conservative for the gene effect. The pattern for the interaction term is more complex. With only three lineages the level of type I error is unacceptably high over most parameter combinations. But the situation improves as the number of lineages increases, with the error being about double that expected with four lineages and the same as that expected with 10 lineages.

To investigate the problem of unequal variances, we repeated a subset of the simulations but with 10 genes of length 100, 200, . . . , 1,000 bp. The results are presented in table 1b. Interestingly, the lineage effect remains unaffected by the heterogeneity in variances, but the type I error increases for the gene and interaction effects. For the gene effect the level of type I error is at 5%, whereas the interaction effect rises to between 10% and 20%.

The heterogeneity in variance we have simulated is quite substantial, but the results suggest that caution

**Table 1**  
**Generic Simulation Results**

Lineages	Genes	Divergence	Gene (Proportion <0.05)	Lineage (Prop. <0.05)	Interaction (Prop. <0.05)
(a) Genes of equal length					
3	10	0.1	0.014	0.048	0.087
3	10	0.5	0.001	0.055	0.127
3	10	1	0.001	0.054	0.177
3	30	0.1	0.008	0.052	0.128
3	30	0.5	<0.001	0.052	0.269
3	30	1	<0.001	0.041	0.326
4	10	0.1	0.018	0.041	0.042
4	10	0.5	0.003	0.042	0.080
4	10	1	<0.001	0.055	0.101
4	30	0.1	0.005	0.046	0.044
4	30	0.5	<0.001	0.064	0.122
4	30	1	<0.001	0.044	0.211
10	10	0.1	0.054	0.041	0.044
10	10	0.5	0.005	0.060	0.051
10	10	1	0.001	0.074	0.048
10	30	0.1	0.054	0.038	0.044
10	30	0.5	0.001	0.073	0.040
10	30	1	<0.001	0.068	0.075
(b) Genes of unequal length					
3	10	0.1	0.04	0.045	0.13
3	10	1	0.03	0.03	0.17
4	10	0.1	0.05	0.039	0.13
4	10	1	0.04	0.038	0.16
10	10	0.1	0.15	0.04	0.10
10	10	1	0.01	0.039	0.14

should be taken in interpreting the mildly significant interaction effects. For this reason we have performed parametric simulations to check the significance of the interaction effects we have observed.

**Variation in Mitochondrial Substitution Rates**

The results of the ANOVA of amino acid substitution rates,  $K_a$ , in mitochondrial genes are given in table 2. All three effects are significant, with the gene effect being the strongest, followed by the lineage and gene-by-lineage interaction. Figure 2 shows the (geometric) mean amino acid substitution rates for each gene and lineage. The strong lineage effect seems to be largely due to the fast rates of evolution in human and elephant, and to a lesser extent, in mouse. In contrast, the gene effect is rather more evenly distributed. Figure 3 shows the substitution rate for each gene in each lineage, plotted on a log scale; if there were no gene-by-lineage interaction (or sampling error), the lines would be parallel. The lines are not parallel, but it is not easy to discern any obvious interactions, i.e., instances of a gene evolving particularly fast or slowly in a particular lineage.

Although we have determined that the mitochondrial  $K_a$  gene-by-lineage interaction is highly significant using our analytical ANOVA method, our test may be biased by the assumption that the expected variance of each observation is the same across the analysis (the assumption of homoscedasticity). This may not be the case for our data because there is variation in gene length, and gene and lineage distances. Therefore, we

**Table 2**  
**ANOVA Results**

Effect	df	MS	F-ratio	Probability	% Variance	% Non-error Variance
<b>Mitochondrial <math>K_a</math></b>						
Lineage.....	6	3.89	12.9	<10 <sup>-6</sup>	26.6	32.4
Gene.....	8	4.95	16.5	<10 <sup>-6</sup>	44.3	54.0
Gene by						
lineage.....	48	0.30	2.24	0.001	11.1	13.5
Error.....	63	0.13			13.4	
<b>Mitochondrial <math>K_s</math></b>						
Lineage.....	3	4.53	23.7	<10 <sup>-5</sup>	50.3	84.4
Gene.....	8	0.40	2.10	0.15	4.3	7.2
Gene by						
lineage.....	24	0.24	1.25	0.27	5.0	8.3
Error.....	36	0.19			40.4	
<b>Nuclear <math>K_a</math></b>						
Lineage.....	3	6.31	26.1	<10 <sup>-6</sup>	13.2	15.7
Gene.....	28	4.29	17.8	<10 <sup>-6</sup>	63.7	75.7
Gene by						
lineage.....	84	0.24	1.93	0.0005	7.3	8.7
Error.....	116	0.13			15.8	
<b>Nuclear <math>K_s</math></b>						
Lineage.....	3	8.06	36.8	<10 <sup>-6</sup>	36.5	60.0
Gene.....	28	0.65	2.95	0.00007	14.4	23.7
Gene by						
lineage.....	84	0.22	1.51	0.02	10.0	16.4
Error.....	116	0.15			39.2	
<b>Nuclear <math>K_a/K_s</math></b>						
Lineage.....	3	0.83	2.05	0.11	0.9	1.2
Gene.....	28	4.45	11.0	<10 <sup>-6</sup>	60.5	84.8
Gene by						
lineage.....	84	0.41	1.70	0.004	10.0	14.0
Error.....	116	0.24			28.6	

performed a simulation test of the significance of the mitochondrial  $K_a$  gene-by-lineage interaction. Amino acid sequence evolution was simulated using PAML according to the Poisson model with all amino acid frequencies equal. All other details of the simulation—phylogeny, gene lengths, and expected branch distances—were as observed for the mitochondrial  $K_a$  data. Expected branch distances were obtained by combining gene and lineage effects, assuming no gene-by-lineage interaction. We obtained our simulation  $P$  value for the

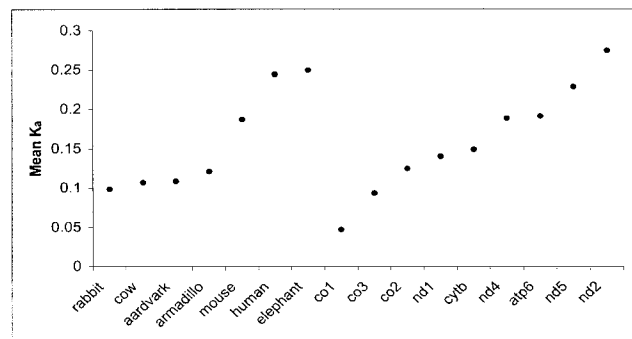


FIG. 2.—Mitochondrial  $K_a$  analysis: gene and lineage effects. The values plotted are the mean  $K_a$  values for each gene and lineage.

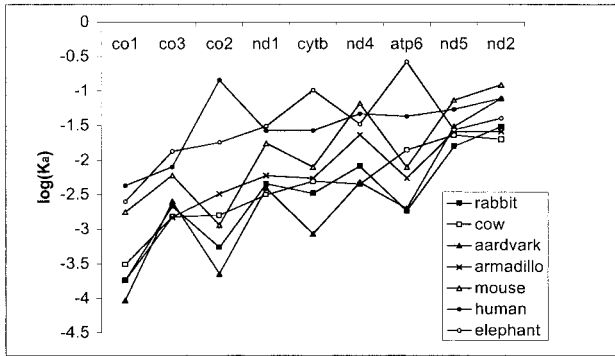


FIG. 3.—Mitochondrial  $K_a$  analysis: gene-by-lineage effects. The values plotted are the logarithm of the  $K_a$  value for each gene in each lineage.

interaction by applying our ANOVA method to the simulated data and recording for each of the 1,000 simulations whether the simulated gene-by-lineage interaction was stronger than the interaction observed for the real data. In only six cases out of 1,000 was the simulated interaction term greater than the observed value, thus supporting our result of a highly significant gene-by-lineage interaction for the mitochondrial  $K_a$  data.

In contrast to the  $K_a$  analysis, only the lineage effect is significant for the synonymous substitution rates in the mitochondrial data set, and this contributes around 85% of the nonerror variance. But the divergences in this data set are small, and this may restrict our power to detect gene and gene-by-lineage effects. Unfortunately, because of the fact that we needed a divergence date to partition the variance correctly, we were restricted in our choice of taxa. The mean substitution rate for each lineage and gene are plotted in figure 4. The rate of synonymous substitution in mitochondrial genes appears to be much higher in whales and horses than in seals and hominids (human-chimpanzee).

Variation in Nuclear Substitution Rates

There are several advantages in using mitochondrial genes for the study of variation in substitution rates. Mitochondria have been completely sequenced in many species, so we can be sure of the orthology, and we can be sure that location effects have not contributed to gene-by-lineage effects because the gene order is conserved in mammals. But the low  $K_a/K_s$  ratio in mito-

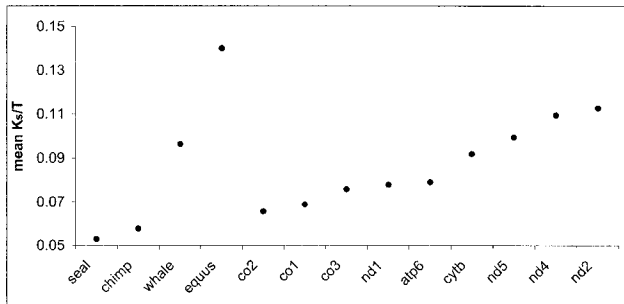


FIG. 4.—Mitochondrial  $K_s$  analysis: gene and lineage effects. The values plotted are the mean  $K_s$  values for each gene and lineage.

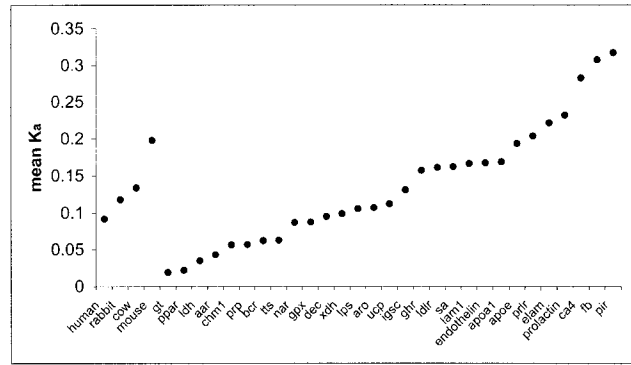


FIG. 5.—Nuclear  $K_a$  analysis: gene and lineage effects. The values plotted are the mean  $K_a$  values for each gene and lineage.

chondrial genes, and the need to have either a star phylogeny or divergence dates, means that we cannot compare  $K_a$  and  $K_s$  using a single set of genes and lineages. In contrast, nuclear genes have a ratio of nonsynonymous substitutions over synonymous substitutions, which is high enough to allow the analysis of both  $K_a$  and  $K_s$  for the same sequences using a star phylogeny. Furthermore, nuclear genes form the bulk of the coding complement in all organisms, so we need to partition the variance in their substitution rates if we are to understand the forces which affect most genes. But the orthology of nuclear genes requires careful validation. We identified a set of 29 nuclear genes from four mammalian species (see *Materials and Methods*).

The analysis of  $K_a$  in nuclear genes reveals highly significant lineage, gene, and gene-by-lineage effects (see table 2). Most of the nonerror variance in  $K_a$  is due to gene effects, with the remainder split between lineage and gene-by-lineage effects in the ratio 2:1. The mean substitution rates for each gene and lineage are plotted in figure 5. As with the mitochondrial  $K_a$  analysis, individual gene-by-lineage interactions are hard to identify; one possibility is the *dec* gene in mouse (fig. 6).

The analysis of nuclear  $K_s$  variation shows significant lineage, gene, and gene-by-lineage effects (see table 2). The nonerror variance in  $K_s$  is mostly due to lineage effects, with the remainder split fairly evenly between gene and gene-by-lineage effects. The plot of mean  $K_s$  values for each gene and lineage is shown in

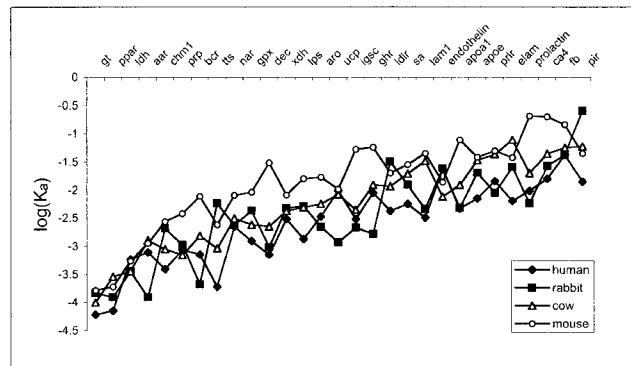


FIG. 6.—Nuclear  $K_a$  analysis: gene-by-lineage effects. The values plotted are the logarithm of the  $K_a$  value for each gene in each lineage.

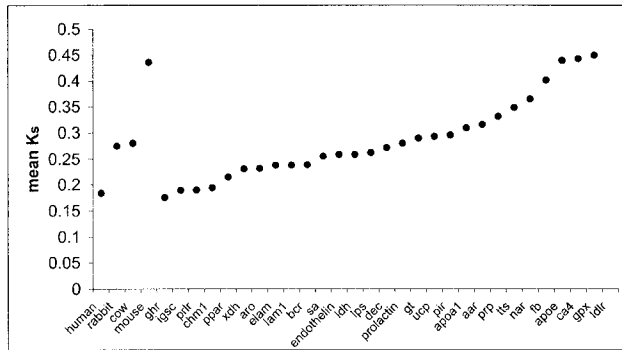


FIG. 7.—Nuclear  $K_s$  analysis: gene and lineage effects. The values plotted are the mean  $K_s$  values for each gene and lineage.

figure 7. As with the other analyses in which we found significant gene-by-lineage interactions, individual gene-by-lineage interactions are hard to identify for the  $K_s$  results, although *fb* in mouse and *ghr* in human could be candidates (fig. 8).

We tested the significance of the gene-by-lineage interaction in the nuclear  $K_a$  data in the same way as that described above for the mitochondrial  $K_a$  data. In only three cases out of 1,000 was the simulated interaction term greater than that observed in the nuclear  $K_a$  data, and so we concluded that our finding of a significant gene-by-lineage interaction in the nuclear  $K_a$  data is robust. We tested the significance of the gene-by-lineage interaction in the nuclear  $K_s$  data in a similar way, except for the fact that the DNA rather than amino acid evolution was simulated and analyzed according to the Jukes-Cantor model. In the case of the nuclear  $K_s$  data, our simulations showed that our method may have generated a false-positive gene-by-lineage interaction because in 103 cases out of 1,000 the simulated gene-by-lineage interaction was stronger than the observed interaction.

## Discussion

### Partitioning the Variance in Substitution Rates

We have partitioned the variance in amino acid and synonymous substitution rates between three effects: gene, lineage, and a gene-by-lineage interaction. For the amino acid substitution rates, we find evidence of all three effects; however, lineage effects are more pronounced for mitochondrial genes than for nuclear genes because they account for ~32% of the variance as opposed to ~16%. The level of gene-by-lineage interaction is quite similar in the two data sets at ~10%. There are strong lineage effects for the synonymous substitution rate in both mitochondrial and nuclear genes; however, although this is the sole significant effect for mitochondrial genes, there is also evidence of a gene effect for nuclear genes.

In all four ANOVAs the lineage effect is significant. For the nuclear genes, in which the comparison can be made, the lineage effect seems to be quite similar for the amino acid and synonymous substitution rates—humans have the lowest rate, followed by cow and rabbit, which have similar branch lengths, with mice show-

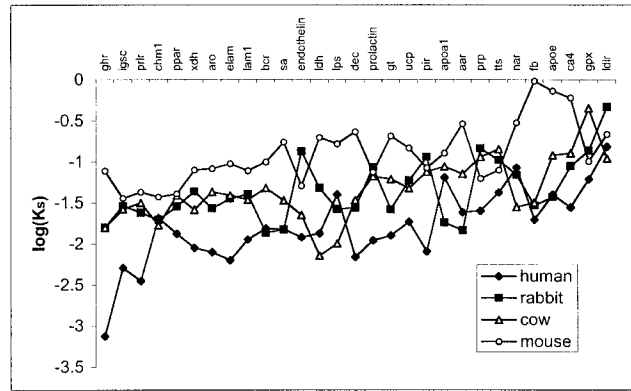


FIG. 8.—Nuclear  $K_s$  analysis: gene-by-lineage effects. The values plotted are the logarithm of the  $K_s$  value for each gene in each lineage.

ing the greatest rate. To investigate this further, we did an ANOVA of the logarithm of  $K_a/K_s$ —as expected, there was no lineage effect (see table 2), which suggests that the lineage effect is similar for the amino acid and synonymous substitution rates and that they probably have a common cause. The cause of the lineage effect is most probably the differences in the rate of mutation between lineages, due to variation in factors such as generation time and metabolic rate, but there are other possibilities. For example, if slightly deleterious mutations segregated at both synonymous and amino acid sites, then differences in effective population size would generate correlated differences in rate along lineages between  $K_a$  and  $K_s$ .

A gene effect for synonymous substitution rates has been clearly demonstrated only once before (Mouchiroud, Gautier, and Bernardi 1995); this was also the case in mammals. A gene effect is also implicit in the observation that the genes which are close together on mammalian chromosomes have similar rates of synonymous substitution (Matassi, Sharp, and Gautier 1999; Lercher, Williams, and Hurst 2001); if the synonymous substitution did not vary in a consistent manner between lineages, it would be difficult to detect a local similarity in rate. The significant gene effect might be due to this local similarity, although the effect is quite weak, particularly in rodents (Lercher, Williams, and Hurst 2001). The gene effect might also be due to CpG islands which are likely to reduce the rate of evolution through the stabilization of the CpG dinucleotide (Bulmer 1986; Sved and Bird 1990).

### Gene-by-Lineage Interactions and Overdispersion

The molecular clock is said to be overdispersed when the variation in the rate of evolution of a gene across lineages is greater than that expected under a Poisson process. But, overdispersion can arise through a variety of processes, including variation in the mutation rate between lineages. This led Gillespie (1989) to suggest that lineage effects should be removed before testing for overdispersion. Gene-by-lineage interactions are equivalent to Gillespie's (1989) definition of overdispersion: it is the variation in the rate of evolution

which cannot be attributed to gene or lineage effects. Overdispersion has been observed previously for both synonymous and amino acid substitution rates in mammalian nuclear genes (Gillespie 1989; Ohta 1995). But there was some doubt over the significance of the observed overdispersion because two factors were not accounted for: (1) the variance associated with the estimate of the substitution rate (Bulmer 1989), and (2) the covariance associated with estimating the rate for a non-independent set of lineages (Bulmer 1989). Simulations suggest that the rate of amino acid substitution is genuinely overdispersed in mammals (Nielsen 1997); however, a similar conclusion was not reached for the synonymous substitution rate. The level of overdispersion is usually measured by the index of dispersion; this is the variance in the “corrected” number of substitutions for a gene across lineages divided by the mean corrected number of substitutions, where the corrected number of substitutions is the number of substitutions for a gene in a lineage weighted by the relative rate of evolution in the lineage across genes. Gillespie (1989) found that the index of overdispersion was 6.95 for the rate of amino acid substitution and 4.64 for the rate of synonymous substitution across primates, artiodactyls, and rodents for a data set of 20 genes; similar results were obtained by Ohta (1995) on a larger data set, 5.60 for  $K_a$  and 5.89 for  $K_s$ . The values of the index of overdispersion for our data are comparable: 8.6 for  $K_a$  and 4.8 for  $K_s$  for nuclear genes, and 10.0 and 8.2 for mitochondrial genes. Our simulation results suggest that the overdispersion of synonymous substitution rates is not significant.

The cause of overdispersion in the rate of amino acid substitution has been one of the central arguments in the neutralist-selectionist debate (Gillespie 1986, 1989, 1991; Takahata 1987, 1988; Cutler 2000). Three neutral explanations for overdispersion have been suggested. First, overdispersion in the rate of amino acid substitution could be due to overdispersion in the mutation rate; if this were the case, then we would not expect an interaction for  $K_a/K_s$ . But there is a highly significant gene-by-lineage interaction for  $K_a/K_s$  (see table 2), and so it seems likely that there is overdispersion in the rate of amino acid substitution which is not caused by variation in the mutation rate.

Second, it has been suggested that overdispersion could be due to variation in the proportion of effectively neutral mutations caused by variation in the effective population size (Takahata 1987; Araki and Tachida 1997). If the strength of selection against a deleterious mutation is such that  $|N_e s| \ll 1$ , where  $N_e$  is the effective population size and  $s$  the strength of selection, then the mutation can spread by random genetic drift and become fixed in the population. But if  $|N_e s| \gg 1$ , the probability of fixation is essentially zero, and the mutation cannot become fixed. Therefore, variation in the effective population size can change the proportion of mutations which can become fixed and, hence, the rate of evolution. Single-locus models show that this process can lead to overdispersion under some parameter combinations (Takahata 1987; Araki and Tachida 1997). But it seems likely that much of this overdispersion effect will

manifest itself as a lineage effect in a multilocus system. If the effective population size decreases, the proportion of effectively neutral mutations will increase, but it will increase in all genes, generating an overall increase in the rate of evolution across all genes. Overdispersion will be present because the change in the proportion of effectively neutral mutations will be different for different genes, but the effect may be too subtle to detect.

Third, it has been suggested that overdispersion might arise through “fluctuating neutral space” (Takahata 1987; Iwasa 1993)—neutral substitutions change the proportion of mutations which are neutral. Under certain circumstances this model is expected to generate adaptive evolution for the following reason. The proportion of mutations which are neutral cannot increase indefinitely, so some neutral mutations will reduce the proportion of mutations which are neutral whereas others increase it (at equilibrium we expect the number of mutations becoming neutral to be equal to the number of mutations becoming nonneutral). Consider a neutral mutation at one site which changes another site, which was formerly neutral, to one that is subject to selection. On average there is a one in four chance that this second site will be occupied by the most advantageous allele, because there are four nucleotides. The first substitution will therefore create advantageous alleles at the second site, and adaptive evolution will ensue.

But it could be the case that the mutation at the first site will not be neutral unless the second site is occupied by its most advantageous allele. In this model, neutral space will fluctuate, but it will do so slowly. If the mutation at the first site only affects the selection at one other site, there is a one in four chance that the second site will be occupied by the most advantageous of the alleles and that the mutation at the first site will be neutral. If the first mutation affects the selection at two sites, the probability that the first mutation is neutral drops to 1 in 16; if it affects three sites, the probability is reduced further to 1 in 64. Hence, neutral mutations which significantly decrease the neutral space will be rare. This behavior may be exactly what is required to explain overdispersion because, as Gillespie (1993) and Cutler (2000) have shown, the change in the rate of evolution needs to be at a rate which is slower than the rate of substitution.

Gillespie (1986, 1989, 1991) has proposed that overdispersion is a consequence of bursts of adaptive evolution. This model certainly seems to be more consistent with some of the dramatic examples of overdispersion that are known. For example, there has been a very marked acceleration in the rate of amino acid substitution in the ancestral lineages leading to both the hominoids and colobine monkeys in the lysozyme gene; in each case,  $K_a > K_s$ , suggesting that the acceleration is due to adaptive evolution (Messier and Stewart 1997; Yang 1998). But there are potential problems with the hypothesis that overdispersion is caused by bursts of adaptive evolution. As Gillespie (1993) and Cutler (2000) have shown, the environment needs to change at a rate which is slower than the rate at which substitutions occur for overdispersion to be produced. Given

that a typical mammalian protein of 500 codons goes through one amino acid substitution every 2 Myr, the rate of environmental change needs to be very slow, and yet the habitat of most species changes much more rapidly. For example, a large proportion of the world was under ice during the last ice-age just 10,000 years ago. There are two possible solutions to this problem. First, the rate of evolution of a protein may not depend upon the external environment but on the substitutions which have occurred in the protein or in other related proteins. This is similar to the covarion model and might be termed a fluctuating adaptive space model. Second, adaptive evolution might be associated with the occupation of a vacant or new niche, and such niches may only become available at a much slower rate than the rate of environmental change.

### Acknowledgments

We are very grateful to Brandon Gaut, in whose laboratory these ideas were first formulated, to Holly Hilton and Spencer Muse for early discussions, and to the BBSRC (N.G.C.S and A.E.W.) and Royal Society (A.E.W) for funding.

### Literature Cited

- Araki, H., and H. Tachida. 1997. Bottleneck effect on evolutionary rate in the nearly neutral mutation model. *Genetics* **147**:907–914.
- Bulmer, M. 1986. Neighbouring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.* **3**:322–329.
- . 1989. Estimating the variability of substitution rates. *Genetics* **123**:615–619.
- Cutler, D. 2000. Understanding the overdispersed molecular clock. *Genetics* **154**:1403–1417.
- Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN, a database of homologous vertebrate genes. *Nucleic Acids Res.* **22**:2360–2365.
- Gillespie, J. H. 1986. Variability of evolutionary rates of DNA. *Genetics* **113**:1077–1091.
- . 1989. Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.* **6**:636–647.
- . 1991. *The causes of molecular evolution*. Oxford University Press, Oxford.
- . 1993. Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. *Genetics* **134**:971–981.
- Gouy, M., F. Milleret, C. Mugnier, M. Jacobzone, and C. Gautier. 1984. ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Res.* **12**:121–127.
- Iwasa, Y. 1993. Overdispersed molecular evolution in constant environments. *J. Theor. Biol.* **164**:373–393.
- Lanave, C., S. Liuni, F. Licciulli, and M. Attimonelli. 2000. Update of AMmtDB: a database of multi-aligned Metazoa mitochondrial DNA sequences status. *Nucleic Acids Res.* **28**:153–154.
- 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: implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18**:2032–2039.
- Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, Mass.
- Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610–614.
- Matassi, G., P. M. Sharp, and C. Gautier. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**:786–791.
- Messier, W., and C.-B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* **385**:151–154.
- Mouchiroud, D., C. Gautier, and G. Bernardi. 1995. Frequencies of synonymous substitutions are gene-specific and correlated with frequencies of non-synonymous substitutions. *J. Mol. Evol.* **40**:107–113.
- Murphy, W. J., E. Elzirik, W. E. Johnson, Y. P. Zhing, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**:614–618.
- Nielsen, R. 1997. Robustness of the estimator of the index of overdispersion for DNA sequences. *Mol. Phylogenet. Evol.* **7**:346–351.
- Ohta, T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**:56–63.
- Pesole, G., C. Gissi, A. De Chirico, and C. Saccone. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. *J. Mol. Evol.* **48**:427–434.
- Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**:1114–1116.
- Sved, J., and A. P. Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. USA* **87**:4692–4696.
- Takahata, N. 1987. On the overdispersed molecular clock. *Genetics* **116**:169–179.
- . 1988. More on the episodic clock. *Genetics* **118**:387–388.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW—improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- Wallis, M. 1994. Variable evolutionary rates in the molecular evolution of mammalian growth hormones. *J. Mol. Evol.* **38**:619–627.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555–556.
- . 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**:568–573.

Edward Holmes, Associate Editor

Accepted July 25, 2002