Bayesian Estimation of Nonsynonymous/Synonymous Rate Ratios for Pairwise Sequence Comparisons

Konstantinos Angelis,[1] Mario dos Reis,[1] and Ziheng Yang*[,1]
[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom
***Corresponding author:** E-mail: z.yang@ucl.ac.uk.
Associate editor: Xun Gu

Abstract

The nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) is an important measure of the mode and strength of natural selection acting on nonsynonymous mutations in protein-coding genes. The simplest such analysis is the estimation of the d_N/d_S ratio using two sequences. Both heuristic counting methods and the maximum-likelihood (ML) method based on a codon substitution model are widely used for such analysis. However, these methods do not have nice statistical properties, as the estimates can be zero or infinity in some data sets, so that their means and variances are infinite. In large genome-scale comparisons, such extreme estimates (either 0 or ∞) of ω and sequence distance (t) are common. Here, we implement a Bayesian method to estimate ω and t in pairwise sequence comparisons. Using a combination of computer simulation and real data analysis, we show that the Bayesian estimates have better statistical properties than the ML estimates, because the prior on ω and t shrinks the posterior of those parameters away from extreme values. We also calculate the posterior probability for $\omega > 1$ as a Bayesian alternative to the likelihood ratio test. The new method is computationally efficient and may be useful for genome-scale comparisons of protein-coding gene sequences.

Key words: **nonsynonymous/synonymous rate ratio, evolutionary distance, Bayesian estimation, pairwise comparisons, protein-coding sequences.**

Article

Introduction

The nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) is an important measure of the mode and strength of natural selection acting on protein-coding genes (Kimura 1977). A number of methods have been developed to estimate ω from pairwise sequence alignments, ranging from heuristic counting methods (Li et al. 1985; Nei and Gojobori 1986; Yang and Nielsen 2000) to maximum-likelihood (ML) methods based on an explicit Markov model of codon evolution (Goldman and Yang 1994). ML estimates (MLEs) of ω for thousands of genes are routinely calculated as descriptive statistics in genomic comparisons (Nielsen et al. 2005; Ge et al. 2008; Walters and Harrison 2010; Buschiazzo et al. 2012; Gladieux et al. 2013; Wang and Chen 2013). Although the ML method for pairwise comparisons produces reasonable estimates of ω and sequence distance (t) for most data sets, it suffers from a few problems when the data sets are extreme. For example, the MLE of ω ($\hat{\omega}$) is 0 when the two compared sequences have only synonymous differences and ∞ when they have only nonsynonymous differences. Similarly, when the sequences are identical, the MLE $\hat{t}$ is 0 and $\hat{\omega}$ is not unique. When the sequences are very divergent $\hat{t}$ may be ∞.

Because of these infinite or undefined estimates, neither $\hat{\omega}$ nor $\hat{t}$ have finite means or variances. Extreme values of $\hat{\omega}$ and $\hat{t}$ are commonly encountered in genome-level comparisons of thousands of genes, and those extreme estimates cause difficulties with the calculation of summary statistics (such as mean $\hat{\omega}$ and $\hat{t}$ across all genes in the genome). An estimation method that always produces finite and reasonable estimates for ω and t is thus desirable. Here, we develop a Bayesian method to calculate the posterior means of ω and t between two sequences, denoted $\tilde{\omega}$ and $\tilde{t}$. Using computer simulation, we show that the posterior means of ω and t are well behaved and have better Frequentist properties than the MLEs. We then use ML and the new Bayesian method to estimate ω and t from pairwise gene alignments for the genomes of four mammals (human, chimpanzee, mouse, and rat) and three bacterial strains (*Escherichia coli* O157:H7, *E. coli* K-12, and *Salmonella typhimurium* LT2). We show that extreme MLEs of ω and t are common in these data sets, and that the Bayesian method produces finite, well-behaved estimates. The new Bayesian method is computationally efficient and is implemented in the CODEML program of the PAML package (Yang 2007).

New Bayesian Approach to Estimate ω and t

Here, we summarize the main features of the new Bayesian approach. The joint posterior distribution of t and ω given the data x (the pairwise sequence alignment) is

$$f(t, \omega \mid x) = \frac{1}{C} f(x \mid t, \omega) f(t, \omega), \quad (1)$$

where $f(x \mid t, \omega)$ is the likelihood or the probability of observing the data x given t and ω, $f(t, \omega)$ is the prior and $C = \int\int f(x \mid t, \omega) f(t, \omega) \mathrm{d}t\mathrm{d}\omega$ is the normalizing constant. The posterior is proportional to the product of the likelihood and the prior. If the model involves the transition/transversion rate

© The Author 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Open Access

ratio (κ), its MLE ($\hat{\kappa}$) is used. If the model involves nucleotide or codon frequency parameters, they are estimated using the observed frequencies. When the data are informative, the likelihood dominates the posterior. When the data are uninformative, the prior may have a strong influence on the posterior. Here, we use two independent gamma distributions to construct the joint prior of t and ω:

$$f(t, \omega) = G(t \mid 1.1, 1.1) \times G(\omega \mid 1.1, 2.2), \quad (2)$$

where the gamma density $G(x \mid \alpha, \beta)$ has mean α/β and variance α/β^2. Here, the prior means of t and ω are 1 and 0.5, respectively, and the shape parameter $\alpha = 1.1$ indicates that the priors are quite diffuse. This joint prior has a mode away from (0,0) and the prior density decays to 0 as either t or ω approaches ∞, thus penalizing extreme values. The likelihood is calculated from a pairwise sequence alignment using a codon substitution model (Yang and Nielsen 1998). As point estimates of ω and t we use their posterior means

$$\tilde{\omega} = \mathrm{E}(\omega \mid x) = \frac{1}{C}\int_0^\infty\int_0^\infty \omega f(x \mid t, \omega, \hat{\kappa}) f(t, \omega)\mathrm{d}t\mathrm{d}\omega, \quad (3)$$

$$\tilde{t} = \mathrm{E}(t \mid x) = \frac{1}{C}\int_0^\infty\int_0^\infty t f(x \mid t, \omega, \hat{\kappa}) f(t, \omega)\mathrm{d}t\mathrm{d}\omega. \quad (4)$$

The posterior variances and covariance of ω and t can be similarly defined and can be calculated using standard numerical techniques. We use Gaussian quadrature to calculate all integrals numerically. We use similar techniques to calculate $P(\omega > 1 \mid x)$, the posterior probability that $\omega > 1$, which may be compared with the likelihood ratio test (LRT) of the null hypothesis H_0: $\omega = 1$ (see Methods and Materials).

We consider five different scenarios in which the numerical calculations of the integrals may differ. We simulated five data sets to represent those five scenarios, each consisting of 2 sequences of 100 codons, with different numbers of synonymous (S_d) and nonsynonymous (N_d) differences. The posterior and likelihood surfaces for the five cases are shown in figure 1.

Case I: ($S_d > 0$, $N_d > 0$). This is the most common case, with both synonymous and nonsynonymous differences observed. The data are quite informative about ω and t and the posterior distribution resembles the likelihood (fig. 1*A*′ and *A*). In our example data set, we have $S = 73.7$, $N = 226.3$, $S_d = 18.5$, $N_d = 6.5$, where S and N are the numbers of synonymous and nonsynonymous sites. The MLEs are $\hat{t} = 0.30$ and $\hat{\omega} = 0.11$ whereas the posterior means are $\tilde{t} = 0.31$ and $\tilde{\omega} = 0.13$.

Case II: ($S_d = N_d = 0$). In this case, the two sequences are identical. The likelihood is maximized when $t = 0$ and when $t = 0$, ω has no effect on the likelihood, so the MLE of ω is not unique (fig. 1*B*). In our example, $S = 73.3$, $N = 226.7$, $S_d = N_d = 0$. The posterior has a single mode and the posterior means are $\tilde{t} = 0.011$ and $\tilde{\omega} = 0.496$ (fig. 1*B*′). Note that the posterior mean of ω is almost equal to the prior mean, since the data are uninformative about ω. Also, the posterior mean is markedly different from the posterior mode, because the posterior distribution is highly skewed.

Case III: ($S_d > 0$, $N_d = 0$). Only synonymous differences are observed. In our example, $S = 74.4$, $N = 225.6$, $S_d = 24$ and $N_d = 0$. Then, we have $\hat{t} = 0.306$ and $\hat{\omega} = 0$ (fig. 1*C*). The posterior for ω has a mode away from 0 and $\tilde{t} = 0.316$ and $\tilde{\omega} = 0.014$ (fig. 1*C*′).

Case IV: ($S_d \gg 0$, $N_d \gg 0$). Only nonsynonymous differences are observed. In our example, $S = 73.2$, $N = 226.8$, $S_d = 0$, $N_d = 40$. The MLEs are $\hat{t} = 0.48$ and $\hat{\omega} = \infty$ (fig. 1*D*). The posterior has a well-defined mode and thus $\tilde{t} = 0.47$ and $\tilde{\omega} = 3.1$ (fig. 1*D*′).

Case V: ($S_d \gg 0$, $N_d \gg 0$). The two sequences are so divergent that they look like random sequences ($S = 75.9$, $N = 224.1$, $S_d = 75$, $N_d = 175$). Here, the likelihood increases with the increase of both t and ω, with the MLEs at $\hat{t} = \infty$ and $\hat{\omega} = \infty$ (fig. 1*E*). In the Bayesian analysis, the prior penalizes large values and thus the posterior means are $\tilde{t} = 10.31$ and $\tilde{\omega} = 0.72$ (fig. 1*E*′). Note that the posterior mean of ω is close to the prior mean, since the data of two nearly random sequences are uninformative about ω.

These five cases illustrate how the prior influences the posterior depending on whether the data are informative or not. The posterior means of t and ω are finite for all five cases, whereas the MLEs are not. We note that because the MLEs of t and ω may be infinite, their mean square errors (MSEs) are ∞ as well. The MSEs of the posterior means are in contrast always well defined. In this sense, the posterior mean estimates have better Frequentist properties than the MLEs. In the next section, we study the statistical properties of the Bayesian estimates of t and ω using simulated and real data, in comparison with the MLEs. We calculate the MSEs of the MLEs by excluding the infinite estimates.

Results

Analysis of Simulated Data

To examine the statistical properties of the posterior estimates of t and ω, we conducted a computer simulation. The program EVOLVER from the PAML package (Yang 2007) was used to generate pairwise sequence alignments of length $L_c = 500$ codons. We used $t = 0.1$, 0.5, 1, and 5 and $\omega = 0.01$, 0.1, 0.5, and 2 (16 combinations) with transition/transversion rate ratio $\kappa = 2$ and equal codon frequencies (1/61) to generate the data sets. The number of replicates was 10,000. The simulated data sets were analyzed using both ML and the new Bayesian method using the CODEML program (Yang 2007). The same prior (eq. 2) was used for all data sets. Equal codon frequencies are assumed in the model (Fequal model).

Figures 2 and 3 show the histograms (smoothed densities) of posterior mean estimates and MLEs of t and ω. As we see in figure 2, ML and Bayesian results are nearly identical for all combinations of $\omega = 0.1$ and 0.5 and $t = 0.5$ and 1. However, for $\omega = 0.01$, Bayesian estimates of ω are shifted to the right (too large) for all t values, as the prior for ω has a mean of 0.5 and affects the posterior estimates. For $\omega = 2$, posterior estimates of ω are shifted to the left (too small) due to the prior.

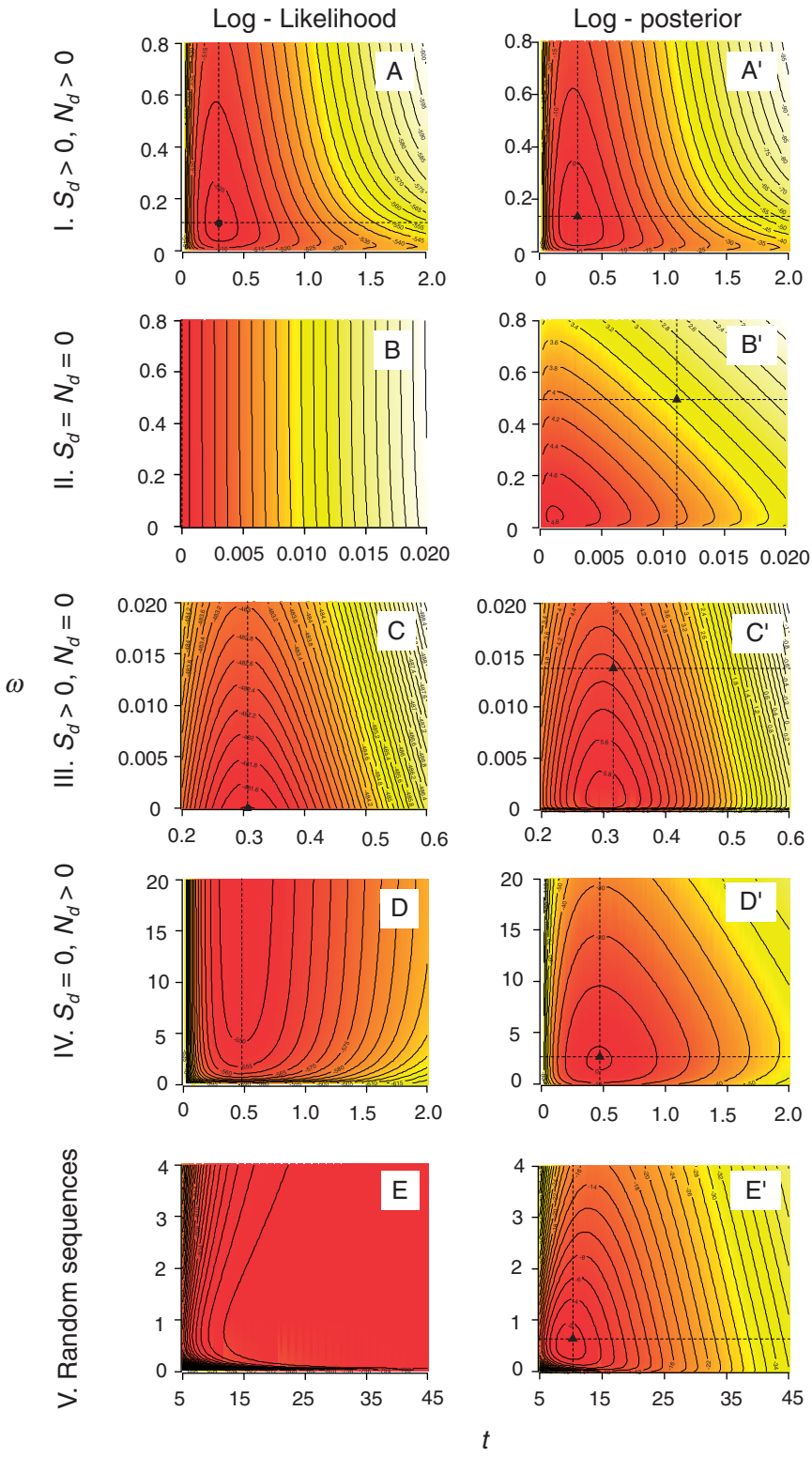

Fig. 1. Contour plots of log-likelihood (A–E) and log-posterior (A'–E') densities for ω and t for five synthetic pairwise sequence alignments of 100 codons. The dashed lines indicate the MLE. Five cases are analyzed: I. normal sequences (A and A'), II. identical sequences (B and B'), III. sequences with only synonymous changes (C and C'), IV. with only nonsynonymous changes (D and D'), V. random sequences (E and E').

Generally, both methods behave best (estimates are more concentrated around the true value) for intermediate distances ($t = 0.5$ and 1), because sequences of moderate divergences are the most informative. The estimates of t show similar patterns (fig. 3). Although for $t = 0.5$ and 1 the Bayesian estimates are almost identical to the MLEs, for $t = 0.1$ Bayesian results are slightly shifted to the right (too large) and for $t = 5$ they are shifted to the left (too small).

The means of the Bayesian and ML estimates, the square root of the MSE ($\sqrt{\mathbf{MSE}}$), and the 2.5% and 97.5% percentiles

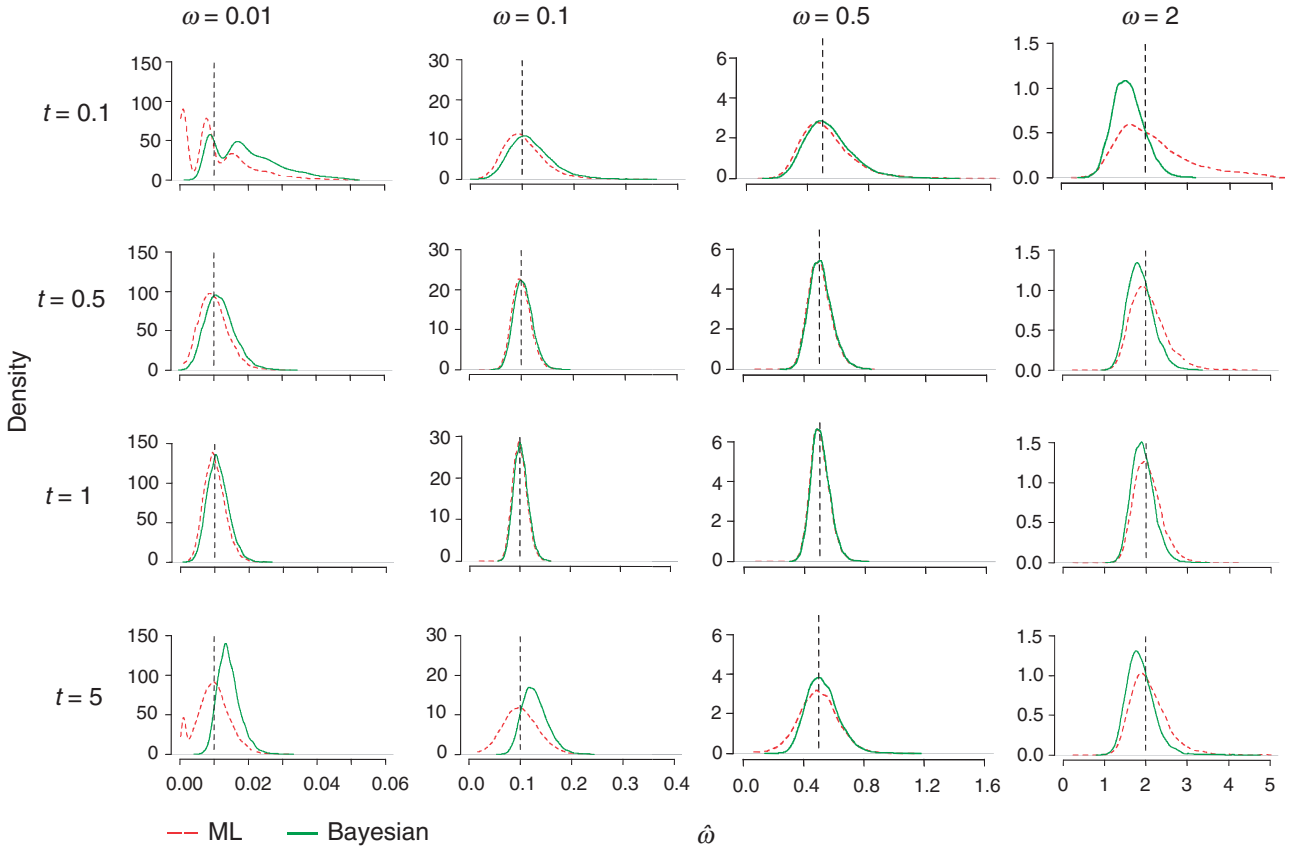

FIG. 2. Kernel density (smoothed histogram) of MLEs (dashed red) and Bayesian posterior means (solid green) for ω in simulated data sets. The true values of ω and t are shown on the top and left of the plots, respectively. The sequence length is 500 codons. The number of replicates is 10,000. The vertical dashed lines correspond to the true values of ω. Independent gamma priors are used $\omega \sim G(1.1, 2.2)$, $t \sim G(1.1, 1.1)$ (eq. 2).

of estimates from the 10,000 simulations are presented in tables 1 and 2. Those for the ML method are calculated after the infinite estimates are removed. We see that for highly similar ($t = 0.1$) and highly divergent ($t = 5$) sequences, the prior has a noticeable impact. For example, when $t = 0.1$ the mean of Bayesian estimates of ω is 0.02 when the true $\omega = 0.01$ and is 1.591 when the true $\omega = 2.0$. The mean MLEs are in comparison closer to the true values than the means of Bayesian estimates. However, the means for the MLEs are calculated after data sets in which $\hat{\omega} = \infty$ are excluded, whereas those same data sets are included in the calculation of the Bayesian estimates. Similar patterns are observed concerning estimates of t. Moreover, for small and intermediate ω and t, ML and Bayesian methods have similar MSE, but for large ω and t, the Bayesian has smaller MSE indicating that in those cases Bayesian estimates are preferable to the MLEs.

We also considered a test of positive selection, indicated by $\omega > 1$. For ML, a LRT is used to compare H_0: $\omega = 1$ against H_1: $\omega > 1$, at the 5% significance level. In the Bayesian framework, we require the posterior probability to exceed the threshold $P(\omega > 1 \mid x) > 0.95$. For the true $\omega = 0.01, 0.1, 0.5$, no data sets showed significant positive selection by either method. When the true $\omega = 2$ and $t = 0.5, 1, 5$, both methods correctly detect positive selection in almost 100% of the replicate data sets, so that the power of detecting positive selection is high in both methods but with the LRT having more power (table 1). When $\omega = 2$ and $t = 0.1$, positive selection is detected in 35% and 61% of data sets by the Bayesian and ML methods, respectively. In this case, given the short sequence distance, the prior has quite some impact on the ability of the Bayesian method to detect selection. In particular, the prior mean ($\omega = 0.5$) is smaller than the true value ($\omega = 2$), so that $\tilde{\omega}$ is shrunk away from 1.

Analysis of Real Data

We applied both ML and Bayesian methods to estimate ω and t for pairwise alignments of protein-coding genes from four mammalian genomes (human, chimpanzee, mouse, and rat) and from three bacterial genomes (*E. coli* O157:H7, *E. coli* K-12, and *S. typhimurium* LT2). In all analyses, the codon frequencies were estimated by using the observed codon frequencies in the genes (the F61 model).

Analysis of the Mammalian Data Set

We conducted three sets of pairwise comparisons: human versus chimpanzee, human versus mouse, and mouse versus rat. Figure 4 shows the distributions (smoothed histograms) of posterior means and the MLEs of t and ω in those comparisons. In the human–chimpanzee comparison, the Bayesian ω estimates are slightly shifted to the right compared with the MLEs for low ω values and shifted to the left for high ω values. The mean, median, and 25% and 75%

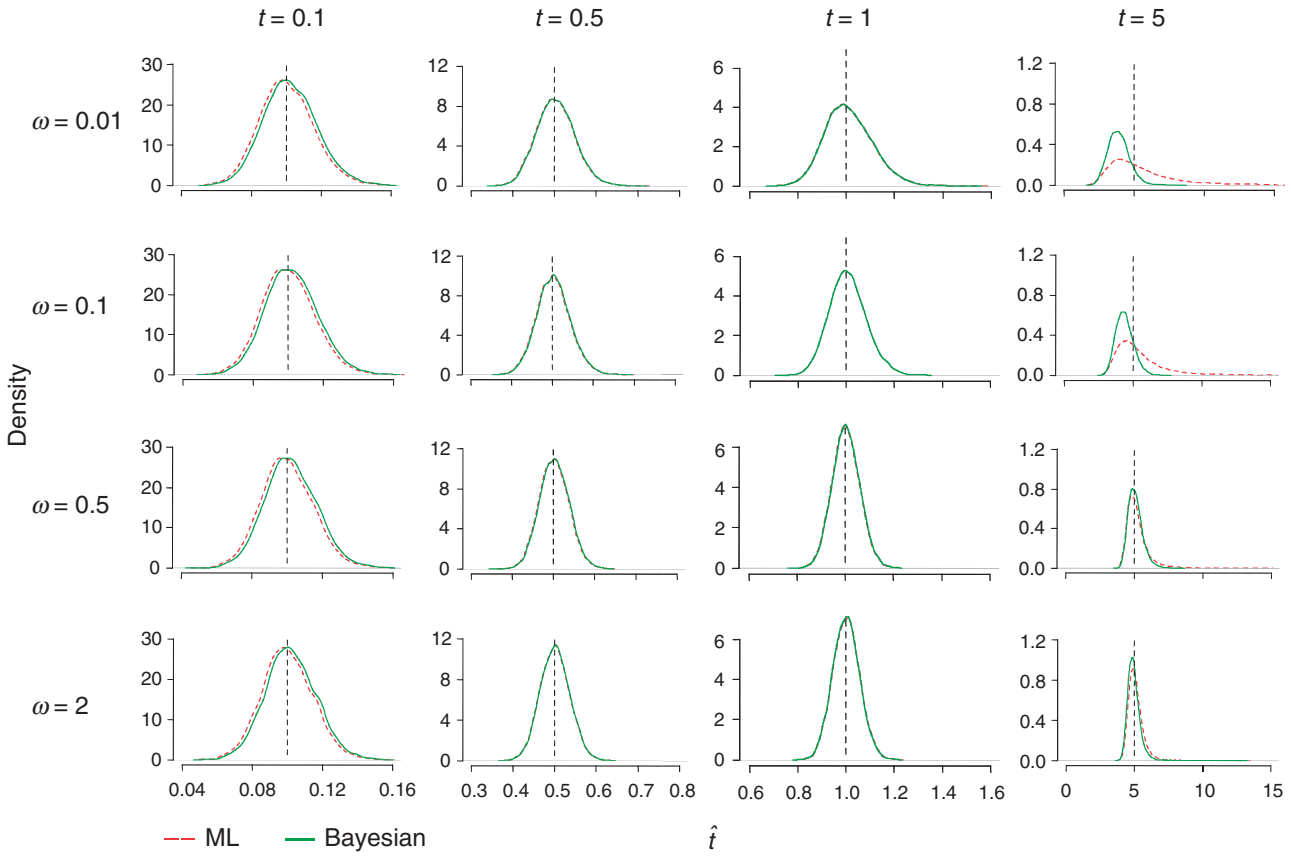

Fig. 3. Kernel density (smoothed histogram) of MLEs (dashed red) and Bayesian posterior means (solid green) for t in simulated data sets. Details as in figure 2.

Table 1. Summary Statistics of Bayesian (top, underlined) and ML (bottom) Estimates of ω from 10,000 Simulated Data Sets.

	$\omega = 0.01$					$\omega = 0.1$				$\omega = 0.5$					$\omega = 2$					
	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_0	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_∞	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_∞	P_+
$t = 0.1$	0.020	0.014	0.007	0.044	0	0.118	0.045	0.052	0.214	0.543	0.160	0.301	0.904	0	1.591	0.546	0.966	2.359	0	35.1
	0.011	0.009	0	0.033	2861	0.103	0.039	0.041	0.194	0.528	0.172	0.278	0.936	0	2.365	1.484	1.015	5.626	3	60.7
$t = 0.5$	0.012	0.005	0.005	0.021	0	0.104	0.018	0.072	0.141	0.511	0.076	0.379	0.677	0	1.878	0.329	1.360	2.543	0	98.3
	0.010	0.004	0.003	0.019	15	0.101	0.018	0.069	0.138	0.506	0.076	0.374	0.674	0	2.064	0.424	1.409	3.031	0	98.9
$t = 1$	0.011	0.003	0.006	0.018	0	0.102	0.014	0.076	0.132	0.506	0.062	0.397	0.637	0	1.922	0.278	1.466	2.497	0	99.9
	0.010	0.003	0.005	0.017	0	0.100	0.014	0.075	0.130	0.503	0.062	0.393	0.635	0	2.038	0.326	1.508	2.764	0	100
$t = 5$	0.014	0.005	0.009	0.022	0	0.129	0.038	0.089	0.183	0.526	0.109	0.348	0.755	0	1.876	0.374	1.331	2.642	0	97.4
	0.010	0.005	0	0.019	370	0.101	0.034	0.037	0.171	0.515	0.981	0.226	0.762	44	2.120	1.398	1.400	3.228	0	98.6

Note.—The Fequal model is used for codon frequencies. Results for ML have been calculated after removing infinite estimates. For $\omega = 0.1$, there were no data sets with 0 or infinite estimates. N_0 is the number of replicates with $\hat{\omega} = 0$, whereas N_∞ is the number of replicates with $\hat{\omega} = \infty$. P_+ is the proportion of replicates with significant evidence for positive selection indicated by $P(\omega > 1 \mid x) > 0.95$ in the Bayesian method or by a significant LRT at the 5% level (one-sided with critical value 2.71) in the likelihood method.

percentiles of the Bayesian estimates are 0.369, 0.320, and (0.180, 0.500) whereas those of the MLEs are 0.307, 0.193, and (0.062, 0.411) (table 3). The human and chimpanzee genes are very similar and the patterns are similar to those observed in computer simulation for low t values. Moreover, there are 377 and 2,507 gene alignments in which $\hat{t} = 0$ and $\hat{\omega} = 0$, respectively, as well as 2 and 423 alignments where $\hat{t} = \infty$ and $\hat{\omega} = \infty$, respectively. The Bayesian method does not produce any such extreme estimates. The number of genes in which the ω estimate is >1 is 1,121 for ML and 299 for the Bayesian method (table 4). The discrepancy is the result of two effects, a short evolutionary distance and a short sequence length, both indicating a lack of information and greater influence from the prior. Genes with $\hat{\omega} > 1$ tend to be small (median sequence length 313 codons, compared with 454 codons for all genes). For example, one gene among

Table 2. Summary Statistics of Bayesian (top, underlined) and ML (bottom) Estimates of t from 10,000 Simulated Data Sets.

	$t = 0.1$				$t = 0.5$				$t = 1$				$t = 5$				
	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	Mean	$\sqrt{\text{MSE}}$	2.5%	97.5%	N_∞
$\omega = 0.01$	0.102	0.015	0.074	0.134	0.504	0.045	0.421	0.596	1.013	0.100	0.837	1.223	3.910	1.322	2.600	5.506	0
	0.100	0.015	0.072	0.132	0.503	0.045	0.419	0.595	1.011	0.100	0.836	1.222	7.572	8.922	2.676	43.744	244
$\omega = 0.1$	0.102	0.015	0.075	0.133	0.503	0.041	0.427	0.587	1.007	0.077	0.865	1.171	4.406	0.869	3.317	5.795	0
	0.100	0.015	0.073	0.131	0.502	0.041	0.425	0.585	1.006	0.077	0.864	1.170	5.629	2.700	3.373	11.506	24
$\omega = 0.5$	0.102	0.015	0.075	0.132	0.503	0.036	0.436	0.574	1.004	0.057	0.895	1.118	5.158	1.469	4.249	6.368	0
	0.100	0.015	0.073	0.130	0.501	0.036	0.434	0.572	1.002	0.057	0.894	1.116	5.440	2.601	4.228	7.979	43
$\omega = 2$	0.102	0.015	0.075	0.131	0.501	0.035	0.434	0.571	1.001	0.056	0.895	1.112	4.988	0.737	4.274	6.035	0
	0.100	0.014	0.073	0.129	0.500	0.035	0.433	0.571	1.002	0.056	0.895	1.114	5.119	0.726	4.323	6.401	3

NOTE.—The Fequal model is used for codon frequencies. Results for ML have been calculated after removing the infinite estimates. For $t = 0.1$, 0.5, and 1, there were no data sets with 0 or infinite estimate. N_∞ is the number of replicates with $\hat{\omega} = \infty$.

those 1,121 with $\hat{\omega} > 1$ has $\hat{\omega} = 1.22$ (95% confidence interval—CI 0.37–4.01) and posterior mean $\tilde{\omega} = 0.93$ (95% credibility interval—CI 0.36–2.43). This gene has a length of 262 codons and has a small evolutionary distance with $\hat{t} = 0.043$ (95% CI 0.024–0.077) and $\tilde{t} = 0.047$ (95% CI 0.027–0.082), so that the prior has an impact. Another gene has $\hat{\omega} = 1.27$ (95% CI 0.75–2.16) and $\tilde{\omega} = 1.13$ (95% CI 0.60–2.13). This gene is 257 codons in length and the ML and Bayesian distance estimates are 0.17 (95% CI 0.13–0.24) and 0.18 (95% CI 0.13–0.24), respectively. The second gene has a similar length to the first but because the sequence distance is greater, the prior is much less important. In a third gene, of length 1,019 codons, the MLEs are $\hat{t} = 0.041$ (95% CI 0.030–0.056) and $\hat{\omega} = 1.27$ (95% CI 0.77–2.07), compared with the Bayesian estimates $\tilde{t} = 0.042$ (95% CI 0.031–0.057) and $\tilde{\omega} = 1.13$ (95% CI 0.59–2.14). In this case, the effect of the prior is unimportant, because the gene is long.

Among the 1,121 genes with $\hat{\omega} > 1$ only 78 have statistically significantly evidence of positive selection, based on the LRT ($\alpha = 5\%$) (table 4). All the 78 genes have the posterior mean $\tilde{\omega} > 1$. Moreover, out of them, three showed strong evidence of positive selection in the Bayesian analysis, with $P(\omega > 1 \mid x) > 0.95$ (table 4). The difference (78 vs. 3 genes) in the number of genes with $\omega > 1$ between the ML and the Bayesian method is consistent with the general expectation that the LRT tends to reject the null more readily than the Bayesian analysis. It is also consistent with the results observed in the computer simulations for $t = 0.1$ and $\omega = 2$. We note that the three genes significant in the Bayesian analysis have fairly large sequence divergences, with $\hat{t} \approx 0.1$, whereas the other 75 genes (for which the LRT is significant but the Bayesian evidence is not strong) have highly similar sequences, with $\hat{t} < 0.07$ (with median 0.021).

In the human–mouse comparison, the ML and Bayesian estimates are very similar. The sequence divergence is intermediate, the data are informative, and the prior does not have a noticeable impact. There are very few cases where the MLEs are extreme (0 or ∞). Also, the number of genes showing ω estimates >1 are nearly the same between the two methods (7 vs. 6) and the same two genes show significant evidence for positive selection by both methods. The mouse–rat comparison shows similar patterns to the human–mouse comparison: in both cases, the sequences are moderately divergent and the data are informative.

To examine the sensitivity of posterior estimates of t and ω to the prior, we reanalyzed the human–chimpanzee and human–mouse alignments using two alternative priors: AP1 and AP2. The first alternative prior (AP1) is $t \sim G(2, 2)$ and $\omega \sim G(2, 4)$. This has the same means as the default prior of equation (2) but the prior is more informative because of the larger shape parameter (2 vs. 1.1). In the second alternative prior (AP2), we used 2 for the shape parameter, but chose the rate parameter such that the prior mean roughly matches the median of the MLEs for all genes (table 3). Thus, for the human–chimpanzee comparison, AP2 is $t \sim G(2, 100)$, with the prior mean 0.02 (while the median of MLEs of t is 0.016), and $\omega \sim G(2, 10)$, with the prior mean 0.2 (while the median of MLEs of ω is 0.193). For the human–mouse comparison, AP2 specifies $t \sim G(2, 3)$, with the prior mean 0.67 (while the median of the MLEs is 0.686) and $\omega \sim G(2, 20)$, with the prior mean 0.1 (the median of the MLEs is 0.089). While it is in general not advisable to use the data to specify the prior, we note that in specific comparisons, some prior information may be available. For example, between the human and the chimpanzee, the distance t is very likely to be smaller than 0.1.

Posterior estimates of ω and t from the analysis using the default and alternative priors are illustrated in figures 5 and 6. In the human–chimpanzee comparison, the impact of the prior is apparent. The Bayesian ω estimates using the AP1 are higher than those using the default prior for low ω values ($\omega < 0.5$) and lower for high ω values ($\omega > 0.5$) (fig. 5*A*). With a more informative prior (shape parameter 2), the posterior means are closer to the prior mean 0.5. For the human–mouse comparison estimates under AP1 are close to those under the default prior (fig. 5*B*). The Bayesian estimates of t are less affected by the change in the prior in both comparisons and the estimates are approximately the same for the majority of the genes (fig. 6*A* and *B*). Prior AP2 has a more significant effect. In both comparisons, the Bayesian estimates of ω are smaller than those obtained using the default prior for almost all genes (fig. 5*C* and *D*). The priors are more informative (with shape parameter $\alpha = 2$) and have

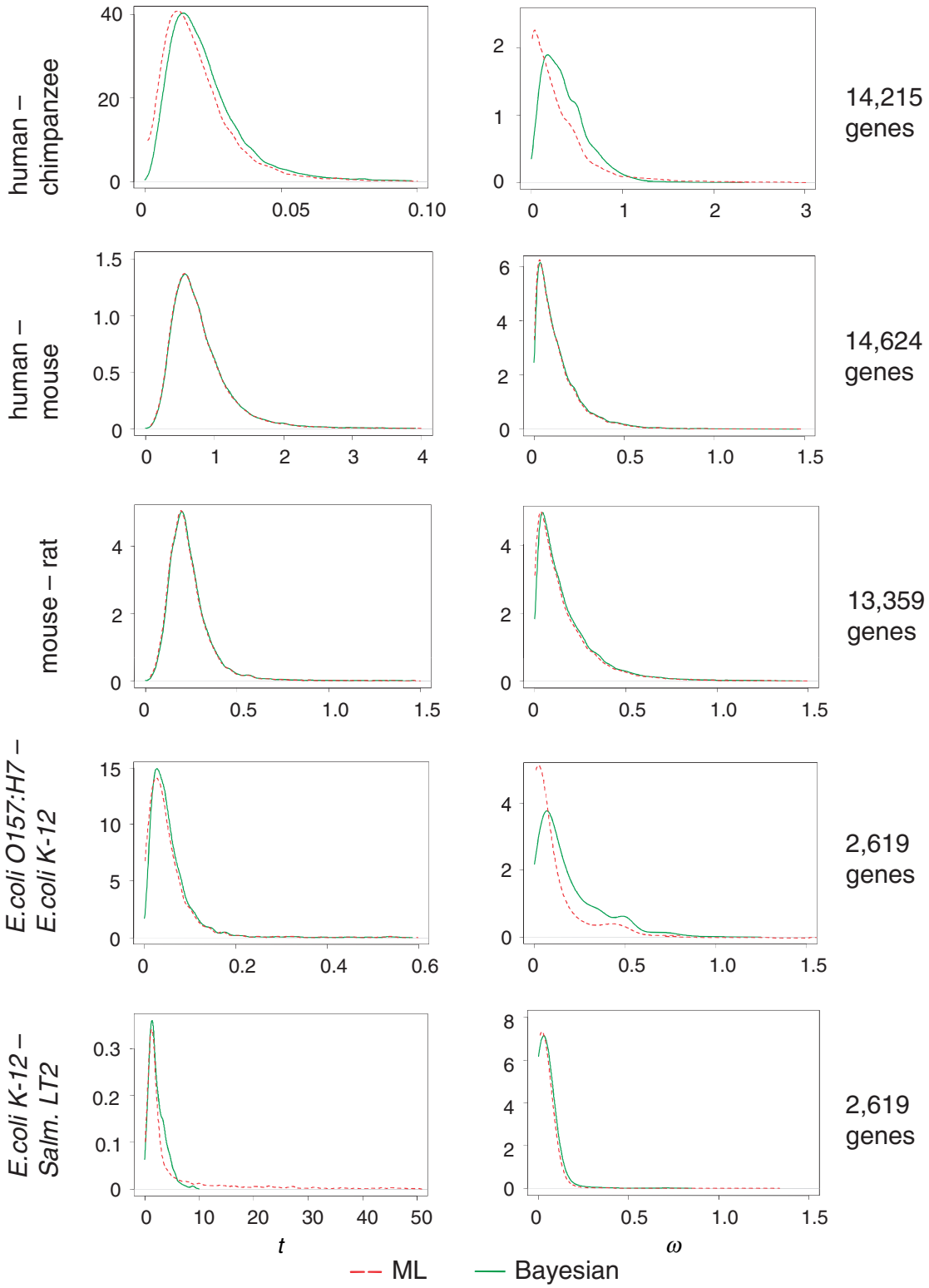

Fig. 4. Distributions (smoothed histograms) of Bayesian and ML estimates of t and ω from mammalian and bacterial pairwise gene comparisons. Numbers of genes analyzed in each comparison are shown in the right part of the figure.

lower means (0.2 and 0.1 for the human–chimpanzee and human–mouse comparisons, respectively, instead of 0.5) and thus affect posterior estimates more than the default prior. The effect is more apparent in the human–chimpanzee comparison because of the smaller sequence distances. Posterior estimates of t are less affected by the change in the prior (fig. 6*C* and *D*). In summary, the prior affects posterior estimates of ω when the genes are not informative about ω and does not affect significantly the posterior estimates of t.

Table 3. Descriptive Statistics of Bayesian (top, underlined) and ML (bottom) Estimates of t and ω from Pairwise Comparisons of Protein-Coding Genes from Mammalian Species and Bacterial Strains.

		ω							t						
		Mean	SD	Quartiles			N_0	N_∞	Mean	SD	Quartiles			N_0	N_∞
	No. of Genes			25%	50%	75%					25%	50%	75%		
Human–chimpanzee	14,215	0.369	0.246	0.180	0.320	0.500	0	0	0.025	0.072	0.013	0.019	0.028	0	0
		0.307	0.418	0.062	0.193	0.411	2507	423	0.022	0.042	0.010	0.016	0.025	377	2
Human–mouse	14,624	0.130	0.125	0.044	0.093	0.176	0	0	0.812	0.574	0.503	0.691	0.958	0	0
		0.126	0.157	0.040	0.089	0.170	221	0	0.849	1.252	0.499	0.686	0.952	0	30
Mouse–rat	13,359	0.168	0.168	0.055	0.118	0.228	0	0	0.242	0.179	0.163	0.215	0.281	0	0
		0.159	0.180	0.046	0.108	0.215	509	0	0.238	0.232	0.161	0.212	0.278	0	3
Escherichia coli K-12–*E.coli* O157	2,619	0.179	0.170	0.055	0.116	0.252	0	0	0.080	0.354	0.026	0.043	0.068	0	0
		0.099	0.174	0.001	0.034	0.110	912	31	0.073	0.527	0.020	0.038	0.064	121	6
E. coli K-12–*Salmonella typhimurium* LT2	2,619	0.037	0.042	0.016	0.025	0.042	0	0	2.261	1.546	1.153	1.836	3.129	0	0
		0.025	0.042	0.006	0.018	0.032	164	0	5.052	8.481	1.087	1.748	4.066	0	217

NOTE.—The F61 model is used for codon frequencies. Results for ML have been calculated after removing the infinite estimates. N_0 is the number of genes with the MLE $\hat{\omega}$ or $\hat{t}=0$, whereas N_∞ is the number of genes with the MLE $\hat{\omega}$ or $\hat{t}=\infty$.

Table 4. The Numbers of Genes with ω Estimate Greater or Less than 1 Using the Bayesian and ML Methods.

Data			Bayesian		
			$\tilde{\omega}<1$	$\tilde{\omega}>1$	N_L
Human–chimpanzee	ML	$\hat{\omega}<1$	13,094	0	78
		$\hat{\omega}>1$	822	299	
		N_B		3	
Human–mouse		$\hat{\omega}<1$	14,617	0	2
		$\hat{\omega}>1$	1	6	
		N_B		2	
Mouse–rat		$\hat{\omega}<1$	13,313	0	5
		$\hat{\omega}>1$	10	36	
		N_B		2	
Escherichia coli K-12–*E. Coli* O157		$\hat{\omega}<1$	2,574	0	0
		$\hat{\omega}>1$	43	2	
		N_B		0	
E. coli K-12–*Salmonella typhimurium* LT2		$\hat{\omega}<1$	2,617	0	0
		$\hat{\omega}>1$	2	0	
		N_B		0	

NOTE.—N_L is the number of genes with statistically significant $\hat{\omega}>1$ based on the LRT at the 5% level (one-sided with critical value 2.71) in the likelihood method, whereas N_B is the number of genes with $P(\omega>1 \mid x)>0.95$ in the Bayesian analysis.

Analysis of the Bacterial Data Set

We conduct two pairwise comparisons: *E. coli* K-12 versus *E. coli* O157:H7 and *E. coli* K-12 versus *S. typhimurium* LT2. Note that the two strains of *E. coli* have the same evolutionary distance from the *Salmonella*.

The sequences from the two *E. coli* strains are very similar, and the prior has an impact on Bayesian estimates, similar to the comparison of the human and chimpanzee genes. The mean, median, and 25% and 75% percentiles of the Bayesian ω estimates are 0.179, 0.116, and (0.055, 0.252) while the corresponding results for the MLEs are 0.099, 0.034, and (0.001, 0.110). The two methods are thus very different in analysis of those genes. Also, the MLE $\hat{\omega}=0$ in 912 genes and $\hat{\omega}=\infty$ in 31 genes. None of the genes with $\hat{\omega}>1$ is statistically significant at the $\alpha=5\%$ significance level according to the LRT and none has $P(\omega>1 \mid x)>0.95$ (table 4). The gene sequences from the *E. coli* K-12 and *Salmonella* are quite divergent. In most genes, the two methods produced similar estimates (fig. 4). However, some genes are very divergent with the MLE $\hat{t}=\infty$ in 217 genes.

Discussion

We suggest that if possible one should conduct joint comparative analysis of multiple protein-coding gene sequences on a phylogeny, instead of pairwise comparisons. In particular, a number of LRTs have been developed to detect positive selection that affects particular evolutionary lineages on the phylogeny or individual sites in the protein (see, e.g., Yang [2006a] and Cannarozzi and Schneider [2012], for reviews). To apply such tests of positive selection, it is essential to use multiple sequences, as a pair of sequences hardly contains enough information for the tests to have any power (e.g., Yang 2006b). Some proteins may evolve in an episodic manner and thus adaptive episodes may not be detected in pairwise comparisons, especially when the sequences are distantly related (Messier and Stewart 1997). In a pairwise comparison, positive selection is detected only if the ω averaged over all sites in the protein and over the whole evolutionary history connecting the two sequences is >1. This seems to be an extremely stringent criterion. Analysis of multiple sequences on a phylogeny allows one to detect episodic positive selection that affects a particular branch (Yang 1998).

Nevertheless, we note that pairwise sequence comparisons are widely used, especially in comparative genomics, sometimes to provide summary statistics of the data and sometimes because of lack of a third genome. The ML method has been used to estimate ω and t in pairwise comparisons of genes (e.g., Nielsen et al. 2005; Ge et al. 2008; Walters and Harrison 2010; Buschiazzo et al. 2012;

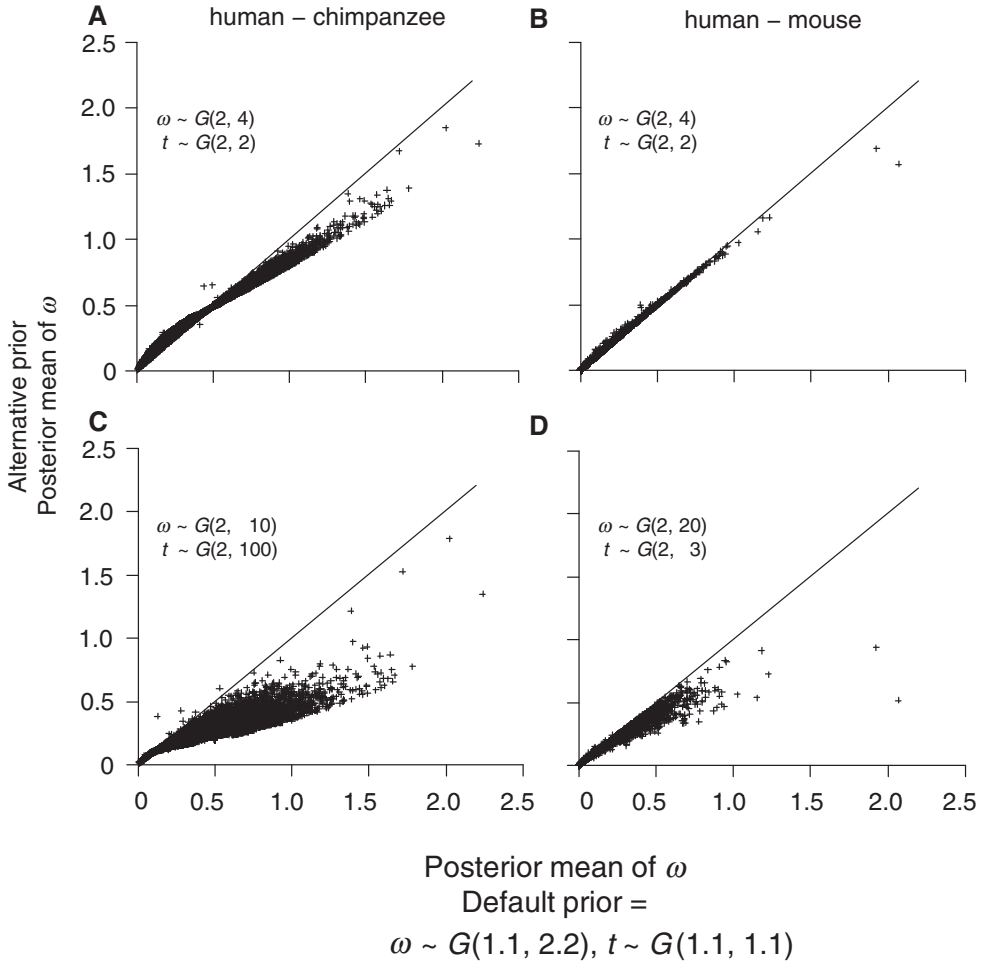

Fig. 5. Bayesian estimates of ω for the human–chimpanzee (A and C) and human–mouse (B and D) comparisons using two alternative priors plotted against estimates using the default prior (eq. 2). The alternative priors are: (A and B) $\omega \sim G(2, 4)$, $t \sim G(2, 2)$; (C) $\omega \sim G(2, 10)$, $t \sim G(2, 100)$; and (D) $\omega \sim G(2, 20)$, $t \sim G(2, 3)$.

Gladieux et al. 2013; Wang and Chen 2013). Counting methods are also used due to their simplicity (Garcia-Gil et al. 2003; Schenekar et al. 2011; Graves et al. 2013), even though they were found not to perform as well as ML in computer simulations (Yang and Nielsen 2000). Both counting and ML methods sometimes return 0 or ∞ as estimates, so that neither the expectation nor the variance of the estimates is finite. The infinity estimates of ω appear to be particularly confusing to many users of the methods. To avoid such extreme estimates, some authors (e.g., Novaes et al. 2008; Bajgain et al. 2011; Pellino et al. 2013) added a small arbitrary number (pseudocounts) to the numbers of synonymous and nonsynonymous substitutions before calculating ω. Other authors excluded genes with $d_S = 0$ from their analysis (e.g., Wang and Chen 2013). The Bayesian method implemented here may provide a better procedure than such ad hoc treatments. It always returns finite estimates of ω and t as the prior penalizes extreme values. Our computer simulation suggests that the Bayesian estimates of ω have nice statistical properties, with similar or smaller MSEs compared with the MLEs. The posterior means are close to the MLEs when the data are informative, that is, when the sequences are long and the sequence divergence is intermediate, but the differences can be large when the sequences are short and are either too similar or too divergent. Nearly identical sequences contain little information while extremely divergent sequences contain too much noise concerning ω. In both cases, the data are not informative and the prior has an impact on posterior estimates of ω. However, as sequence length increases the effect of the prior decreases irrespective of the true values of ω and t. Our Bayesian method is used for the analysis of only two sequences. A Bayesian method for the analysis of multiple sequences in a phylogeny requires calculation of high-dimensional integrals and is not pursued here.

We emphasize that MLEs $\hat{\omega} = \infty$ should not be taken as evidence for positive selection ($\omega > 1$) because the extreme estimate may well be due to chance effects when the numbers of changes are small. Instead, positive selection can be claimed only if the LRT is significant in the ML framework or when $P(\omega > 1 \mid x) > 0.95$ in the Bayesian analysis.

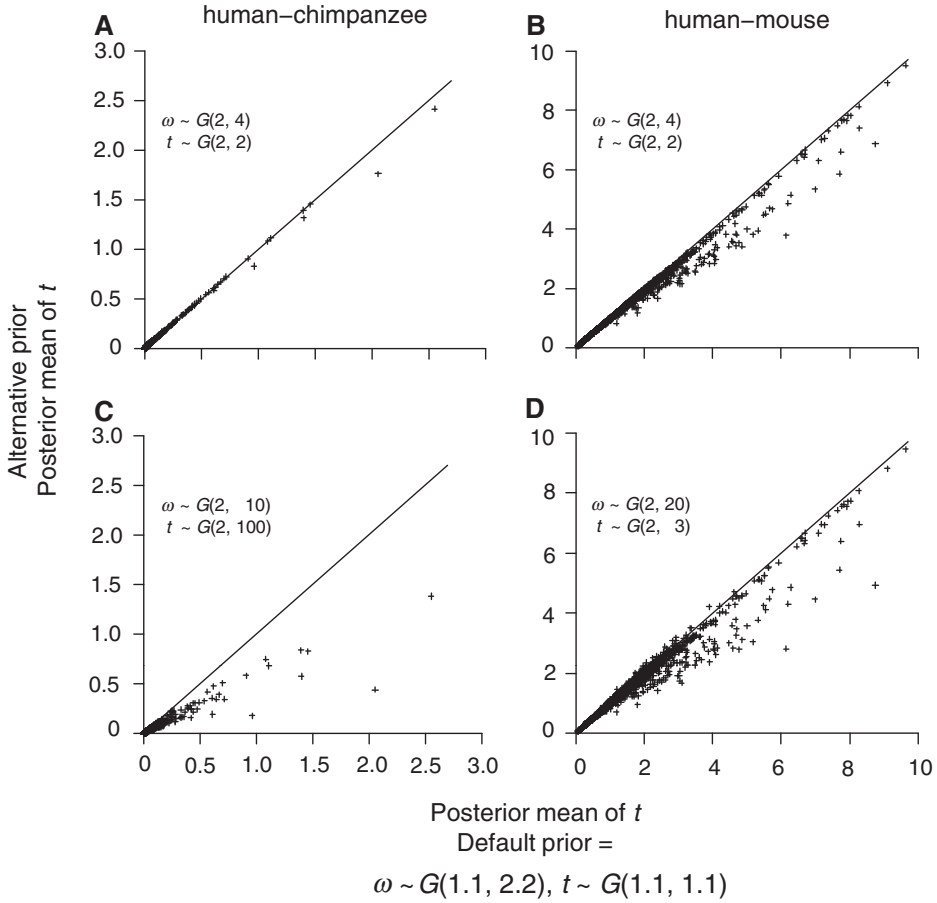

Fig. 6. Bayesian estimates of t for the human–chimpanzee (A and C) and human–mouse (B and D) comparisons using different gamma priors. The alternative priors are as in figure 5.

Program Availability

The Bayesian method of this article is implemented in the CODEML program in the PAML package. The program allows the user to specify gamma priors for t and ω. Although the Bayesian method is computationally more intensive than ML, it remains fast enough for large-scale screening. It takes 1–2 s to analyze one pair of sequences on a modern PC.

Methods and Materials

Theory

We use a simplified version of the model of Goldman and Yang (1994) to model the evolution of codon sequences (Yang and Nielsen 1998). The model accounts for the genetic code structure, the transition/transversion rate ratio, the codon frequencies as well as the d_N/d_S rate ratio ω. The instantaneous substitution rate from codon i to codon j ($i \neq j$) is given by

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition,} \end{cases} \quad (5)$$

where π_j is the equilibrium frequency of codon j. Stop codons are not considered (they are assumed not to occur within protein-coding genes). Therefore, the substitution rate matrix $Q = \{q_{ij}\}$ is of size 61×61 for the standard genetic code. The rate matrix is scaled so that the average rate of codon substitution equals $-\sum_{i=1}^{61} \pi_i q_{ii} = 1$, and thus time is measured by the expected number of nucleotide substitutions per codon site. We use standard theory to calculate the transition probability matrix over time t as $P(t) = \exp(Qt)$. The likelihood function on a pairwise sequence alignment x is

$$f(x \mid t, \kappa, \omega) = \prod_{h=1}^{L_c} \pi_i P_{ij}(t), \quad (6)$$

where i and j are the observed codons in the two sequences at site h and L_c is the number of codons.

The joint posterior distribution of ω and t is given by equation (1). If κ is a parameter in the model we replace it with its MLE ($\hat{\kappa}$). If the two sequences are identical so that $\hat{\kappa}$ is not unique, we fix it at 2. Besides the posterior means of ω and t given in equations (3) and (4), we also calculate the posterior variances and covariance

$$\mathrm{Var}(\omega \mid x) = E(\omega^2 \mid x) - [E(\omega \mid x)]^2, \quad (7)$$

$$\mathrm{Var}(t \mid x) = E(t^2 \mid x) - [E(t \mid x)]^2, \quad (8)$$

$$\mathrm{Cov}(\omega, t \mid x) = E(\omega t \mid x) - E(\omega \mid x)E(t \mid x). \quad (9)$$

Thus, six double integrals need to be computed, one for the normalizing constant C, and five for the different expectations in equations (3), (4), and (7)–(9).

Consider the calculation of the normalizing constant C. All other integrals are calculated in the same way. We write $g(t, \omega) = f(x \mid t, \omega)\, f(t, \omega)$. To avoid overflows and underflows, we set $h(t, \omega) = \exp\{\log[g(t, \omega)] - l_{\max}\}$, where $l_{\max}$ is the maximum of $g(t, \omega)$, a constant chosen for scaling. The normalizing constant can then be written as

$$C = \exp(l_{\max}) \int_0^\infty \int_0^\infty h(t, \omega)\mathrm{d}t\mathrm{d}\omega. \quad (10)$$

We use the Gaussian quadrature method to calculate all integrals numerically, which uses Legendre polynomials to approximate any continuous integrand function $f(x, y)$:

$$\int_{-1}^{1} \int_{-1}^{1} f(x, y)\mathrm{d}x\mathrm{d}y \approx \sum_{i,j=1}^{n} w_i w_j f(x_i, y_j). \quad (11)$$

The weights w_i and w_j and the points x_i and y_j at which the integrand is evaluated are predetermined given the total number of points n. In our case, the limits of the integrals are 0 and ∞ and we have to use a transformation to map the $(0, \infty)$ limits to $(-1, 1)$. A much more serious problem is that the integrand g may be spiky (i.e., it is highly concentrated in a very small interval) and the approximation will be very poor if the sampled points miss the spike in the integrand. The rationale behind our transformation is to find a probability density function (PDF) that has a similar shape to the integrand $g(t, \omega)$ and then we use its cumulative distribution function (CDF) to transform the integrand. Note that if the chosen PDF matches the posterior exactly, the new integrand will become perfectly flat after the transformation. The logistic distribution is used for that purpose.

Let $x_1 = \log t \sim \mathrm{Logistic}(\mu_1, \sigma_1)$ and $x_2 = \log\omega \sim \mathrm{Logistic}(\mu_2, \sigma_2)$. For any random variable $x \sim \mathrm{Logistic}(\mu, \sigma)$ the CDF is $F_L(x) = \frac{1}{1 + e^{-(x-\mu)/\sigma}}$. Thus, for equation (10), we use the following transformation (change of variables):

$$z_1 = 2F_L(x_1) - 1 \;\Rightarrow\; t = \exp\left\{\mu_1 + \sigma_1 \log\frac{1 + z_1}{1 - z_1}\right\}, \quad (12)$$

$$z_2 = 2F_L(x_2) - 1 \;\Rightarrow\; \omega = \exp\left\{\mu_2 + \sigma_2 \log\frac{1 + z_2}{1 - z_2}\right\}. \quad (13)$$

Thus, equation (10) becomes

$$\begin{aligned} C &= \exp(l_{\max}) \int_{-1}^{1} \int_{-1}^{1} r(z_1, z_2)\mathrm{d}z_1\mathrm{d}z_2 \\ &\approx \exp(l_{\max}) \sum_{i,j=1}^{n} w_i w_j r(z_{1_i}, z_{2_j}), \end{aligned} \quad (14)$$

where $r(z_1, z_2) = h(t, \omega)\frac{2t\sigma_1}{1-z_1^2}\frac{2\omega\sigma_2}{1-z_2^2}$ and t and ω are given by equations (12) and (13), respectively. We transform all other integrals in equations (3), (4), and (7)–(9) in the same way. Thus, we have

$$\begin{aligned} \mathrm{E}(\omega \mid x) &\approx \tfrac{1}{A} \sum_{i,j=1}^{n} w_i w_j \omega_i r(z_{1_i}, z_{2_j}), \\ \mathrm{E}(t \mid x) &\approx \tfrac{1}{A} \sum_{i,j=1}^{n} w_i w_j t_j r(z_{1_i}, z_{2_j}), \\ \mathrm{E}(\omega^2 \mid x) &\approx \tfrac{1}{A} \sum_{i,j=1}^{n} w_i w_j \omega_i^2 r(z_{1_i}, z_{2_j}), \\ \mathrm{E}(t^2 \mid x) &\approx \tfrac{1}{A} \sum_{i,j=1}^{n} w_i w_j t_j^2 r(z_{1_i}, z_{2_j}), \\ \mathrm{E}(t\omega \mid x) &\approx \frac{1}{A} \sum_{i,j=1}^{n} w_i w_j \omega_i t_j r(z_{1_i}, z_{2_j}), \end{aligned} \quad (15)$$

where $A = C\exp(-l_{\max})$. Notice that the exponential term $\exp(l_{\max})$ cancels out during calculations.

Our Bayesian calculation is performed after the MLEs are obtained. Thus if both $\hat{t}$ and $\hat{\omega}$ are finite, away from 0 and the observed p_S and p_N (proportion of synonymous differences per synonymous site and proportion of nonsynonymous differences per nonsynonymous site, respectively) are <0.74, we set $\mu_1 = \log\hat{t}$, $\mu_2 = \log\hat{\omega}$, $\sigma_1 = \left(\frac{1}{\hat{t}}\right)\sqrt{\hat{V}(\hat{t})}$, and $\sigma_2 = \left(\frac{1}{\hat{\omega}}\right)\sqrt{\hat{V}(\hat{\omega})}$. The variances $\mathrm{V}(\hat{t})$ and $\mathrm{V}(\hat{\omega})$ are estimated using the Nei and Gojobori (1986) method. Because the Nei and Gojobori method uses the Jukes and Cantor (1969) nucleotide substitution model (JC69) to correct for multiple hits, the use of 0.74 as an upper limit for the p_S and p_N guarantees an adequate estimation of $\mathrm{V}(\hat{t})$ and $\mathrm{V}(\hat{\omega})$.

In all other cases, we find numerically the point $(\bar{t}, \bar{\omega})$ that maximizes $\log\{g(t, \omega)\}$. We calculate the Hessian matrix at this point using the second-order difference method and use the inverse of the Hessian to estimate the variances $\mathrm{V}(\bar{t})$ and $\mathrm{V}(\bar{\omega})$. Then, we set $\mu_1 = \log\bar{t}$, $\mu_2 = \log\bar{\omega}$, $\sigma_1 = \left(\frac{1}{\bar{t}}\right)\sqrt{\hat{V}(\bar{t})}$, and $\sigma_2 = \left(\frac{1}{\bar{\omega}}\right)\sqrt{\hat{V}(\bar{\omega})}$. Notice that because of our choice of the prior, $\log(g)$ always has a mode and thus the optimization algorithm returns a point away from $(0, 0)$.

We use the same number of points n for both parameters ω and t in the Gaussian quadrature. With $n = 32$, each sum in equation (15) requires $32 \times 32 = 1{,}024$ evaluations of the $r(z_1, z_2)$ function. Tests suggest that using 32 points achieves high accuracy. The use of more points increases the computational time radically since evaluation of $r(z_1, z_2)$ requires evaluation of the likelihood which is computationally expensive. Moreover, we use the same techniques described above to calculate the posterior probability $P(\omega > 1 \mid x) = \frac{1}{C}\int_0^\infty \int_1^\infty f(x \mid t, \omega, \hat{\kappa})\, f(t, \omega)\mathrm{d}\omega\mathrm{d}t$, as a Bayesian equivalent of the LRT for positive selection indicated by $\omega > 1$.

Real Data Analysis

Both the new Bayesian method of this article and the ML method of Goldman and Yang (1994) were applied to compare protein-coding genes from mammalian species and bacterial strains. The mammalian data set is a subset of the data analyzed by dos Reis et al. (2012). There are 14,218 genes from the human and chimpanzee, with the sequence length ranging from 39 to 8,797 codons; 14,631 genes from the human and mouse with the sequence length from 13 to 8,787 codons; and 13,371 genes from the mouse and rat with the sequence length from 14 to 7,798 codons. The protein-coding sequences from the genomes of *E. coli* O157:H7, *E. coli* K-12, and *S. typhimurium* LT2 were downloaded from GenBank (accession numbers: U_00096, NC_002655, and NC_003197). Orthologous genes among the three genomes were identified by using the program BLAT (Kent 2002) to extract the best reciprocal hits. Only orthologs present in all three genomes are used. This bacterial data set consists of 2,631 genes from each strain, with the sequence length ranging from 20 to 1,485 codons. Codons involving alignment gaps and ambiguity nucleotides were removed prior to analyses. Moreover, genes with sequence length of 50 codons or less were excluded from the analysis. The number of genes analyzed in each comparison is reported in table 3 and figure 4.

Acknowledgments

The authors thank two anonymous referees for constructive comments. K.A. is supported by a UCL Impact studentship. Z.Y. is supported by a grant BB/J009709/1 from the Biotechnological and Biological Sciences Research Council (BBSRC) and a Royal Society-Wolfson Merit Award.

References

Bajgain P, Richardson BA, Price JC, Cronn RC, Udall JA. 2011. Transcriptome characterization and polymorphism detection between subspecies of big sagebrush (*Artemisia tridentata*). *BMC Genomics* 12:370.

Buschiazzo E, Ritland C, Bohlmann J, Ritland K. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol.* 12:8.

Cannarozzi GM, Schneider A. 2012. Codon evolution: mechanisms and models. Oxford: Oxford University Press.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc Biol Sci.* 279:3491–3500.

Garcia-Gil MR, Mikkonen M, Savolainen O. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Mol Ecol.* 12:1195–1206.

Ge G, Cowen L, Feng X, Widmer G. 2008. Protein coding gene nucleotide substitution pattern in the apicomplexan protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Comp Funct Genomics.* 2008: 879023.

Gladieux P, Devier B, Aguileta G, Cruaud C, Giraud T. 2013. Purifying selection after episodes of recurrent adaptive diversification in fungal pathogens. *Infect Genet Evol.* 17:123–131.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11: 725–736.

Graves CJ, Ros VI, Stevenson B, Sniegowski PD, Brisson D. 2013. Natural selection promotes antigenic evolvability. *PLoS Pathog.* 9:e1003766.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12: 656–664; Article published online before March 2002.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267: 275–276.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* 2:150–174.

Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151–154.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418–426.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3:e170.

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.

Pellino M, Hojsgaard D, Schmutzer T, Scholz U, Horandl E, Vogel H, Sharbel TF. 2013. Asexual genome evolution in the apomictic *Ranunculus auricomus* complex: examining the effects of hybridization and mutation accumulation. *Mol Ecol.* 22:5908–5921.

Schenekar T, Winkler KA, Troyer JL, Weiss S. 2011. Isolation and characterization of the CYP2D6 gene in Felidae with comparison to other mammals. *J Mol Evol.* 72:222–231.

Walters JR, Harrison RG. 2010. Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in *Heliconius butterflies*. *Mol Biol Evol.* 27:2000–2013.

Wang TC, Chen FC. 2013. The evolutionary landscape of the *Mycobacterium tuberculosis* genome. *Gene* 518:187–193.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 15: 568–573.

Yang Z. 2006a. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2006b. On the varied pattern of evolution of 2 fungal genomes: a critique of Hughes and Friedman. *Mol Biol Evol.* 23: 2279–2282.

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

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol.* 46:409–418.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.