BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 16 2013, pages 1970–1979
doi:10.1093/bioinformatics/btt340

Phylogenetics

Advance Access publication June 12, 2013

Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency

Guy Baele* and Philippe Lemey

Department of Microbiology and Immunology, Rega Institute, KU Leuven, 3000 Leuven, Belgium

Associate Editor: David Posada

ABSTRACT

Motivation: The advent of new sequencing technologies has led to increasing amounts of data being available to perform phylogenetic analyses, with genomic data giving rise to the field of phylogenomics. High-performance computing is becoming an indispensable research tool to fit complex evolutionary models, which take into account specific genomic properties, to large datasets. Here, we perform an extensive Bayesian phylogenetic model selection study, comparing codon and nucleotide substitution models, including codon position partitioning for nucleotide data as well gene-specific substitution models for both data types. For the best fitting partitioned models, we also compare independent partitioning with standard diffuse prior specification to conditional partitioning via hierarchical prior specification. To compare the different models, we use state-of-the-art marginal likelihood estimation techniques, including path sampling and stepping-stone sampling.

Results: We show that a full codon model best describes the features of a whole mitochondrial genome dataset, consisting of 12 protein-coding genes, but only when each gene is allowed to evolve under a separate codon model. However, when using hierarchical prior specification for the partition-specific parameters instead of independent diffuse priors, codon position partitioned nucleotide models can still outperform standard codon models. We demonstrate the feasibility of fitting such a combination of complex models using the BEAGLE library for BEAST in combination with recent graphics cards. We argue that development and use of such models needs to be accompanied by state-of-the-art marginal likelihood estimators because the more traditional and computationally less demanding estimators do not offer adequate accuracy.

Contact: guy.baele@rega.kuleuven.be

Received on January 24, 2013; revised on June 3, 2013; accepted on June 10, 2013

1 INTRODUCTION

The startling advances in sequencing technology have led to a dramatic increase in the scale and ambition of phylogenetic analyses. As more and more complete genomes are sequenced, phylogenetics is entering a new era—that of phylogenomics, which uses phylogenetic principles to extract information from genomic data (Eisen and Fraser, 2003). Until recently, molecular phylogenies based on a single or few orthologous genes often yielded contradictory results (Jeffroy *et al.*, 2006). One branch of the expanding field of phylogenomics aims to reconstruct the evolutionary history of organisms on the basis of their genomes, as opposed to performing single-gene studies, which have long dominated the field (Delsuc *et al.*, 2005). By expanding the number of characters that can be used in phylogenetic reconstruction from a few thousand to tens of thousands, access to genomic data could potentially alleviate sampling problems that hampered previous phylogenetic analyses.

Phylogenomic analyses involve estimating the underlying evolutionary history of sequences either as an intermediate goal or as an end point. Statistical phylogenetics provides a framework for estimating historical patterns, inferring intrinsic parameters of evolutionary processes, and testing hypotheses under the auspices of the neutral theory of molecular evolution (Kumar *et al.*, 2012). In contrast to the few genes that were previously available, the complete genomes of many species that are now accessible for inferring evolutionary relationships constitute large quantities of data that lead to reduced estimation errors associated with site sampling, to very high power in the rejection of simple evolutionary hypotheses and to high confidence in estimated phylogenetic patterns.

Traditionally, phylogenetic analyses based on many genes combined data into a contiguous block, a practice that is still commonly used today. Under this concatenated model, all genes are not only assumed to evolve under the same tree but also to evolve at the same rate. Although multigene datasets have the advantage of providing greater resolution—with more information it is likely to find trees that more accurately reflect evolutionary history—it may prove challenging to account for the heterogeneous nature of the data. Different genes undergo different selective pressures, and the degree of site rate heterogeneity may vary from gene to gene (Bevan *et al.*, 2007). Likelihood calculations on trees have been shown to clearly benefit from accommodating different evolutionary pressures (as observed in different codon positions, or different genes) in heterogeneous data (Bull *et al.*, 1993; Nylander *et al.*, 2004; Yang, 1996a, b).

Selection is a key evolutionary process in shaping genetic diversity and a major focus of phylogenomics investigations (Kumar *et al.*, 2012). Using statistical methods in evolutionary genetics, researchers frequently evaluate the strength of selection operating on genes or even individual codons in the entire phylogeny or in a subset of branches. Codon substitution models have been particularly useful for this purpose because they allow estimating the ratio of non-synonymous and synonymous substitution rates (dN/dS) in a phylogenetic framework. Two different versions of codon substitution models were simultaneously introduced that both allowed estimating a single dN/dS ratio across all sites and branches (Goldman and Yang, 1994;

*To whom correspondence should be addressed.

Muse and Gaut, 1994). Various extensions have since been proposed, such as codon models that model variation in dN/dS among sites complemented with empirical Bayes approaches to identify the sites under specific selection regimes (Nielsen and Yang, 1998).

Codon substitution models can to some extent be approximated by partitioning nucleotide models according to codon positions, which accommodates differences in evolutionary dynamics at the three codon positions. For example, Yang (1996a, b) takes into account the nucleotide frequency bias, the substitution rate bias and the difference in the extent of rate variation among the three codon positions and shows that incorporating these features can yield drastically different divergence time estimates compared with models not incorporating this complexity. Although full codon models approximate biological reality more closely, codon position partitioned nucleotide models have the advantage to be far more computationally efficient. While Yang (1996a, b) only used the HKY evolutionary model (Hasegawa *et al.*, 1985) in his analyses, Shapiro *et al.* (2006) also included the GTR model (Tavaré, 1986) in a comprehensive model evaluation study (see also Baele *et al.*, 2011). Shapiro *et al.* (2006) showed that codon position partitioned nucleotide models are biologically motivated, computationally practical alternatives to codon models for the analysis of protein-coding sequences. It is therefore not surprising that such approaches are also being exploited for detecting positively selected sites (Lemey *et al.*, 2012).

The large state space of full codon substitution models renders these models computationally expensive compared with standard nucleotide substitution models, which explains why they are frequently fit to trees to scrutinize selection processes but generally not used to reconstruct phylogenies. However, as computational power has increased, phylogenetic inference using codon-based models is becoming more and more realistic. Here, we examine whether the use of full codon models is feasible in the phylogenomics era by fitting several codon models to a full genome mitochondrial dataset in a Bayesian framework. We use state-of-the-art model-selection approaches to assess whether increased biological realism, as obtained by modelling gene-specific properties, goes hand in hand with increased model performance. Finally, we provide estimates of computation time required on both multi-core CPU systems and a system equipped with one of the latest graphics cards available.

2 METHODS

2.1 Data

To examine the differences in model fit between nucleotide and codon models and the performance gains that graphical processing units (GPUs) afford in statistical phylogenetics, we fit a range of evolutionary models (see next section) to the mitochondrial genomes from 62 extant carnivores and a pangolin outgroup (Suchard and Rambaut, 2009). This genomic sequence alignment contains 10 869 nt columns that code for 12 mitochondrial proteins and when translated into a 60-state vertebrate mitochondrial codon model, yields a total of 3623 alignment columns, of which 3601 site patterns are unique. Conducting a Bayesian phylogenetic analysis on such an extensive dataset using a codon substitution model would take up an unpractically large amount of time, but the computational hindrance has to a large extent been removed by recent developments that exploit a special codebase in CUDA to perform the calculations on graphics cards (Suchard and Rambaut, 2009). The analyses reported there could at that time not be run in double precision (i.e. increased precision for floating point numbers to minimize rounding errors) on a single graphics card due to memory restrictions. With the advent of new graphics cards and the release of the BEAGLE library (Ayres *et al.*, 2012), we here revisit codon substitution model estimates for the carnivores dataset.

2.2 Evolutionary models

Goldman and Yang (1994) and Muse and Gaut (1994) developed the first codon-based evolutionary models (GY and MG, respectively), i.e. models that have codons as their states, incorporating biologically meaningful parameters such as transition/transversion bias, variability of a gene and amino acid differences. While nucleotide models have 4 states and amino acid models have 20 states, a full vertebrate mitochondrial codon model has 60 states (ignoring the four nonsense or stop codons). Both the GY and MG codon models assume that mutations occur independently at the three codon positions and therefore only consider substitutions that involve a single-nucleotide substitution. As in nucleotide models, codons are also assumed to evolve independently from one another. Although more realistic codon models have been proposed, e.g. for mammalian genes (Rubinstein *et al.*, 2011), we restrict ourselves to the standard GY codon substitution model implementation in BEAST (Drummond *et al.*, 2012) and allow for substitution rate heterogeneity among codons using a discrete gamma distribution (i.e. each codon is allowed to evolve at a different substitution rate) (Yang, 1996a, b).

2.3 Model selection

We use state-of-the-art marginal likelihood estimation techniques to compare different models using Bayes factors. In particular, we focus on recent implementations of path sampling (PS) and stepping-stone sampling (SS) in BEAST (Baele *et al.*, 2012), which have been previously introduced in Bayesian phylogenetics by Lartillot and Philippe (2006) and Xie *et al.* (2011), respectively. Although these approaches are computationally demanding, they considerably outperform the still widely used harmonic mean estimator (HME). The HME has been shown to systematically overestimate the marginal likelihood, even for simple (Gaussian) cases, while PS and SS estimate the marginal likelihood with much lower error (Lartillot and Philippe, 2006; Xie *et al.*, 2011). A recent study comparing the performance of HME, PS and SS as marginal likelihood estimators for coalescent models and molecular clock models in BEAST (Baele *et al.*, 2012) has also demonstrated that PS and SS yield reliable model selection whereas the HME performs poorly. Further evaluations of these estimators were carried out in the context of a comparison between model selection and model averaging (Baele *et al.*, 2013a), providing additional evidence against the use of the HME and sHME. Here, we include the HME and sHME for the sake of comparison and it serves to illustrate that they may lead to flawed conclusions.

The HME only requires samples from the posterior and can therefore be calculated from an MCMC sample that is obtained by a standard Bayesian phylogenetic analyses under a particular model. If one collects n samples from the posterior, the HME is estimated as follows

$$p(Y|M) = \frac{n}{\sum_{i=1}^{n} \frac{1}{p(Y|\theta_i, M)}}, \quad (1)$$

with $p(Y|M)$ the marginal likelihood and $p(Y|\theta_i, M)$ the likelihood (with M the model under evaluation). The sHME is based on a mixture of the prior and the posterior, but in practice it only uses samples from the

posterior (Newton and Raftery, 1994). If one collects n samples from the posterior, the sHME is estimated as follows

$$p(Y|M) = \frac{\delta_n/(1-\delta) + \sum_{i=1}^{n} p(Y|\theta_i, M)/\left\{\delta p(Y|M) + (1-\delta)p(Y|\theta_i, M)\right\}}{\delta_n/(1-\delta)p(Y|M) + \sum_{i=1}^{n}\left\{\delta p(Y|M) + (1-\delta)p(Y|\theta_i, M)\right\}^{-1}}. \quad (2)$$

BEAST uses a δ value of 0.01 and a simple, readily converging iterative scheme to calculate the sHME.

PS and SS rely on drawing MCMC samples from a series of distributions, each of which is a power posterior differing only in its power, along the path going from the prior to the unnormalized posterior defined by the model M. Both Lartillot and Philippe (2006) and Xie *et al.* (2011) define this path to be:

$$q_\beta(\theta) = p(Y|\theta, M)^\beta p(\theta|M), \quad (3)$$

where $p(Y|\theta, M)$ is again the likelihood function and $p(\theta|M)$ the prior. This formulation ensures that the power posterior is equivalent to the posterior distribution when $\beta = 1.0$ and equivalent to the prior distribution when $\beta = 0.0$.

Our implementation of PS in BEAST (Baele *et al.*, 2012, 2013a) involves adapting the original PS assumptions, i.e. spreading the different values of β evenly between 0.0 to 1.0 and only collecting one sample from each power posterior (before β is updated). Xie *et al.* (2011) found that the efficiency of PS can be considerably improved by choosing β values according to evenly spaced quantiles of a Beta(α, 1.0) distribution rather than spacing β values evenly from 0.0 to 1.0. As suggested by Xie *et al.* (2011), we use a value of $\alpha = 0.3$, which results in half of the β values evaluated being <0.1. We have also chosen to use multiple samples per β, leading to the following expression for the (log) marginal likelihood, assuming $K+1$ path steps, which yield a collection of samples $(\beta_k, \theta_k)_{k=0\ldots K}$, with $\beta_0 = 0$ and $\beta_K = 1$

$$\ln p(Y|M) = \sum_{k=0}^{K-1}(\beta_{k+1} - \beta_k)\left(\frac{\sum_{i=1}^{n} \ln p(Y|\theta_{k,i}, M)}{2n} + \frac{\sum_{i=1}^{n} \ln p(Y|\theta_{k+1,i}, M)}{2n}\right). \quad (4)$$

For SS, we follow Xie *et al.* (2011) in calculating the marginal likelihood using n samples from a series of $K+1$ power posteriors as follows

$$p(Y|M) = \prod_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n} p(Y|\theta_i, M)^{\beta_k - \beta_{k-1}}. \quad (5)$$

To perform the marginal likelihood estimations, we use the BEAST software package (Drummond *et al.*, 2012) to draw inference under the standard nucleotide and the codon partition models as well as estimating the marginal likelihoods for those models using the different Monte Carlo estimators presented earlier. We use BEAST in combination with BEAGLE (Ayres *et al.*, 2012), an application programming interface (API) and library that enables one to exploit the massive parallelism that modern-day graphics cards (GPUs) have to offer, for the computationally demanding codon models. We compare the performance of two (re)scaling schemes that BEAGLE offers when calculating (the partial) likelihoods to help avoid roundoff (Suchard and Rambaut, 2009): 'delayed' (which postpones scaling until the first underflow or overflow) and 'always' (which yields a better precision in each iteration at an increased computational cost). The posterior-based estimators were run for 10 million iterations, discarding the first million as burn-in (2 million for the codon-based models). PS and SS were run for an initial burn-in of 2.5 million iterations, after which 64 power posteriors were run for 275 thousand iterations each, discarding the first 25 thousand iterations as the burn-in.

2.4 Prior specifications

In recent work, we have shown the importance of using proper priors (probability distributions that integrate to 1) when performing Bayesian model selection and by extension, when performing Markov chain Monte Carlo analyses (Baele *et al.*, 2013a). The frequently used constant function, often inaccurately called a uniform distribution, over an infinite interval is an example of an improper prior; the use of such priors may lead to a posterior distribution that does not exist. Further, recent model-selection approaches such as PS and SS explicitly sample from the prior distribution when calculating the marginal likelihood. It is important to stress that an improper prior distribution frequently leads to an infinite marginal likelihood (even if the estimation method returns a non-infinite value), which in turn implies that the Bayes factor is not well-defined (Friel and Petitt, 2008), making inference based on improper priors highly suspect. Despite this importance, attributing little attention to proper prior specification has unfortunately been common practice when calculating Bayes factors in phylogenetics, which can inadvertently affect model comparison conclusions.

We use the following priors in our analyses: a birth–death process (Gernhard, 2008) was used as a tree-prior with a diffuse normally distributed prior on the log growth rate and a uniform prior (between 0.0 and 1.0) on the relative death rate; a diffuse normally distributed prior on the log transition/transversion parameter of the HKY model and of the GY94 codon model; a diffuse normally distributed prior on the log coefficient bearing on the non-synonymous/synonymous rate ratio; diffuse gamma distributed priors on the relative rate parameters of the general time-reversible (GTR) model (Tavaré, 1986); an exponential prior on (each of) the rate heterogeneity parameter(s) (Yang, 1996a, b); and an exponential prior on the standard deviation of the lognormal distribution specifying the rates of an uncorrelated relaxed clock model (UCLD). As an alternative to the diffuse priors for substitution model parameters in gene-partitioned models, we also explore hierarchical phylogenetic modelling (HPM) approaches (Suchard *et al.*, 2003). HPMs use hierarchical prior distributions on the gene-specific parameters that are in turn characterized by unknown estimable hyperparameters, which ensures that within-partition parameters vary around an unknown common mean for each gene.

2.5 Hardware/Software

Nucleotide substitution models and their codon position partitioned versions were fitted on a 40-core Intel Xeon(R) E7- 4870 2.40 GHz system, using a single BEAGLE instance (Ayres *et al.*, 2012). Given the relatively low dimensions of these models and the lower number of unique site patterns (compared with those of higher-dimension models, such as codon models), the proposed system provides reasonable computing performance, which will not be further discussed here. The use of codon models, considered to be more realistic models of sequence evolution in coding genes, however, requires considerable computational effort to evaluate the likelihood of phylogenetic histories, the number of which increases drastically with the number of sequences in the dataset. Parallel calculation of the finite-time transition probabilities is therefore key to speeding up such analyses, preferably on adequate hardware.

We perform the codon-based analyses on a state-of-the-art desktop PC sporting a 3.4 GHz Intel Core i7 CPU and 16 GB of 1.6 GHz DDR3 RAM, equipped with an NVIDIA GTX 590 consisting of two GPUs and hence with a total 1024 CUDA cores and 3 Gb of GDDR5 RAM. At the time this study was initiated, this graphics card was the fastest consumer-oriented card available with satisfactory double precision performance for scientific computing. More recent releases include the

NVIDIA GTX 680 (1 GPU) and GTX 690 (2 GPUs) graphics cards, which have taken a performance hit in their scientific computing capabilities at the expense of a newly released range of NVIDIA workstation cards, which have more graphics memory and higher double precision performance and should hence be more suited for scientific computing.

3 RESULTS

We first compared the standard HKY and GTR nucleotide models, assuming varying rates across sites [modelled using a discrete gamma distribution with four rate categories (Yang, 1996a, b)], with the HKY-based and GTR-based codon position partitioned nucleotide models analysed in the work of Shapiro *et al*. (2006) and the codon model of Goldman and Yang (1994), also accommodating varying substitution rates across codons. For each of the models, the marginal likelihoods were calculated assuming both a strict clock and a uncorrelated relaxed clock with an underlying lognormal distribution (UCLD) (Drummond *et al*., 2006). The results of this model comparison are listed in Table 1. We consider the common codon partition (CP) models where all three codon positions are considered separately (denoted CP_{123}) and where the first and second codon position are grouped together (denoted CP_{112}), as the third codon positions generally evolve much faster than the first and the second codon positions.

None of the marginal likelihood estimators in Table 1 selects the GY94 codon model as the best-fitting mode, indicating that increased biological realism offered by explicitly model substitution in codon space is not reflected in an increased model fit. The model that seems to most adequately capture the substitution complexity in the data is a full fledged codon position partitioned nucleotide model, which assumes a separate GTR model for each codon position, as well as different rate heterogeneity patterns and nucleotide frequency compositions across codon positions (the partitioning also specifies different relative rates for the codon positions). The various estimators consistently select that model as the best model, with apparently only small differences in the ranking of the models. We note, however, that the HME and sHME yield drastically different estimates for the log marginal likelihood compared with PS and SS, which supports earlier claims that such posterior-based estimators overestimate the marginal likelihood (Lartillot and Philippe, 2006; Xie *et al*., 2011). The GY94 codon model ranks among the various codon position partitioned nucleotide models, albeit in different parts of the overall ranking, depending on the estimator used. Standard nucleotide models, with or without varying rates across lineages, appear to be too simplistic for this dataset, as established by all the model-selection approaches. For each of the models tested, a relaxed molecular clock (with underlying lognormal distribution; UCLD) is consistently shown to outperform a strict clock.

Next, we introduce gene-specific partitioning for the models tested in Table 1. To this purpose, we assume one model instance for each of the 12 genes present in the dataset, resulting for example in 12 GY94 codon models being estimated and 36 GTR models being estimated for the most parameter-rich codon position partitioned models, along with 36 gamma distributions being estimated and 36 sets of empirical frequencies being used. For each of these models, we have calculated marginal likelihoods assuming both a strict and a relaxed clock (Table 2).

Table 1. Model comparison between nucleotide models, their codon position (CP) partitioned versions and codon models. CP models are allowed different rates relative to each other, resulting in models that are labelled CP_{112} and CP_{123} with the subscript indicating the rate category for each CP (Shapiro *et al*. 2006)

Model	Clock	Resource	HME	sHME	PS	SS
Nucleotide models						
$HKY+\Gamma$	SC	(CPU − 64 bit)	−194777.15	−194764.43	−195148.46	−195151.20
$HKY+\Gamma$	RC	(CPU − 64 bit)	−194542.78	−194535.03	−194979.87	−194980.25
$GTR+\Gamma$	SC	(CPU − 64 bit)	−194505.35	−194497.86	−194898.23	−194903.69
$GTR+\Gamma$	RC	(CPU − 64 bit)	−194271.22	−194263.58	−194731.50	−194730.51
Codon position partitioned nucleotide models						
$HKY_{112}+CP_{112}+\Gamma_{112}$	SC	(CPU − 64 bit)	−188176.33	−188169.32	−188566.99	−188569.42
$HKY_{112}+CP_{112}+\Gamma_{112}$	RC	(CPU − 64 bit)	−188023.10	−188011.56	−188460.51	−188462.16
$HKY_{123}+CP_{123}+\Gamma_{123}$	SC	(CPU − 64 bit)	−187825.40	−187816.71	−188223.94	−188227.21
$HKY_{123}+CP_{123}+\Gamma_{123}$	RC	(CPU − 64 bit)	−187666.27	−187658.34	−188121.53	−188127.63
$GTR_{112}+CP_{112}+\Gamma_{112}$	SC	(CPU − 64 bit)	−187992.12	−187984.52	−188418.74	−188422.25
$GTR_{112}+CP_{112}+\Gamma_{112}$	RC	(CPU − 64 bit)	−187839.60	−187829.99	−188317.03	−188321.83
$GTR_{123}+CP_{123}+\Gamma_{123}$	SC	(CPU − 64 bit)	−187376.73	−187369.36	−187826.43	−187833.38
$GTR_{123}+CP_{123}+\Gamma_{123}$	RC	(CPU − 64 bit)	−187220.85*	−187214.11*	−187724.26*	−187728.51*
Codon models						
$GY94+\Gamma$	SC	(GPU − 64 bit)	−188014.32	−188007.59	−188404.52	−188404.97
$GY94+\Gamma$	RC	(GPU − 64 bit)	−187860.70	−187849.43	−188293.14	−188296.01

Note: Posterior-based model selection approaches (i.e. HME and sHME) were run for 10 million iterations (of which 1 million serve as burn-in; 2 million for the codon models), while PS and SS were run for 64 path steps/ratios with 250 000 iterations (+25 000 iterations burn-in) per path step/ratio after an initial 1 million iterations that serve as burn-in (2 million for the codon models). Rescaling in BEAGLE was set to default (delayed).

Table 2. Model comparison between gene-specific codon models and gene-specific codon position partitioned models (with 12 genes present in the alignment) and gene-specific nucleotide models

Model	Clock	Resource	HME	sHME	PS	SS
Gene-specific nucleotide models with gene-specific rate variation (no relative rates)						
$12 \times (HKY + \Gamma)$	SC	(CPU − 64 bit)	−194396.68	−194386.04	−194834.57	−194833.18
$12 \times (HKY + \Gamma)$	RC	(CPU − 64 bit)	−194172.15	−194163.49	−194682.65	−194684.08
$12 \times (GTR + \Gamma)$	SC	(CPU − 64 bit)	−194117.13	−194103.29	−194709.01	−194717.70
$12 \times (GTR + \Gamma)$	RC	(CPU − 64 bit)	−193894.85	−193877.87	−194555.70	−194556.55
Gene-specific codon position partitioned models: among codon position rate heterogeneity, homogeneous rates among genes						
$12 \times (HKY_{112} + CP_{112} + \Gamma_{112})$	SC	(CPU − 64 bit)	−187271.76	−187258.79	−187867.90	−187861.02
$12 \times (HKY_{112} + CP_{112} + \Gamma_{112})$	RC	(CPU − 64 bit)	−187116.61	−187106.30	−187765.45	−187769.85
$12 \times (HKY_{123} + CP_{123} + \Gamma_{123})$	SC	(CPU − 64 bit)	−186701.68	−186688.82	−187353.09	−187360.19
$12 \times (HKY_{123} + CP_{123} + \Gamma_{123})$	RC	(CPU − 64 bit)	−186544.57	−186530.02	−187254.74	−187260.86
$12 \times (GTR_{112} + CP_{112} + \Gamma_{112})$	SC	(CPU − 64 bit)	−186986.84	−186973.81	−187807.31	−187820.66
$12 \times (GTR_{112} + CP_{112} + \Gamma_{112})$	RC	(CPU − 64 bit)	−186839.53	−186824.79	−187714.81	−187722.89
$12 \times (GTR_{123} + CP_{123} + \Gamma_{123})$	SC	(CPU − 64 bit)	−186179.71	−186167.95	−187116.56	−187131.50
$12 \times (GTR_{123} + CP_{123} + \Gamma_{123})$	RC	(CPU − 64 bit)	−186031.91	−186014.53	−187007.28	−187020.53
Gene-specific codon position partitioned models: among codon position and among gene rate heterogeneity						
$12 \times (HKY_{112} + CP_{112} + \Gamma_{112})$	SC	(CPU − 64 bit)	−187276.91	−187252.66	−187883.78	−187892.05
$12 \times (HKY_{112} + CP_{112} + \Gamma_{112})$	RC	(CPU − 64 bit)	−187111.03	−187098.52	−187776.69	−187776.81
$12 \times (HKY_{123} + CP_{123} + \Gamma_{123})$	SC	(CPU − 64 bit)	−186871.26	−186861.97	−187382.98	−187385.31
$12 \times (HKY_{123} + CP_{123} + \Gamma_{123})$	RC	(CPU − 64 bit)	−186727.09	−186704.22	−187322.07	−187321.20
$12 \times (GTR_{112} + CP_{112} + \Gamma_{112})$	SC	(CPU − 64 bit)	−186977.48	−186962.17	−187816.57	−187827.14
$12 \times (GTR_{112} + CP_{112} + \Gamma_{112})$	RC	(CPU − 64 bit)	−186841.06	−186811.89	−187720.80	−187734.61
$12 \times (GTR_{123} + CP_{123} + \Gamma_{123})$	SC	(CPU − 64 bit)	−186162.58	−186146.43	−187114.60	−187118.11
$12 \times (GTR_{123} + CP_{123} + \Gamma_{123})$	RC	(CPU − 64 bit)	−186004.06*	−185990.45*	−187024.19	−187038.32
Gene-specific codon models (no relative rates)						
$12 \times (GY94 + \Gamma)$	SC	(GPU − 64 bit)	−186465.43	−186452.31	−187006.16	−187010.41
$12 \times (GY94 + \Gamma)$	RC	(GPU − 64 bit)	−186310.12	−186300.91	−186917.05	−186919.59
Gene-specific codon models (relative rates)						
$12 \times (GY94 + \Gamma)$	SC	(GPU − 64 bit)	−186358.25	−186350.38	−186938.58	−186940.48
$12 \times (GY94 + \Gamma)$	RC	(GPU − 64 bit)	−186217.09	−186205.22	−186851.45*	−186855.58*

Note: Posterior-based model-selection approaches (HME and sHME) were run for 10 million iterations (of which 1 million serve as burn-in; 2 million for the codon models), while PS and SS were run for 64 path steps/ratios with 250 000 iterations (+25 000 iterations burn-in) per path step/ratio after an initial 1 million iterations of burn-in (2 million for codon models). Rescaling in BEAGLE was set to default (delayed).

The gene-specific partitioning tested in Table 2 yields a different model ranking compared with Table 1. Whereas a gene-independent codon model was outperformed by its codon position partitioned nucleotide approximation, this is no longer the case when gene-specific evolutionary patterns are taken into account. The gene-specific codon model that allows for different (relative) rates between the different genes, with a different gamma distribution to model varying rates across sites within each gene, achieves the best performance as assessed by both PS and SS. However, the posterior-based model-selection approaches (HME and sHME) fail to detect this, and also other subtle differences can be observed for these estimators when comparing various versions of the gene-specific codon position partitioned nucleotide models in Table 2. The strong tendency to overestimate marginal likelihoods by the HME is also reflected in Table 2.

Comparing Table 1 and Table 2 reveals that each of the models analysed benefits from taking into account gene-specific properties of the dataset studied, i.e. that it is composed of 12 protein-coding genes (see Figure 1 for gene-specific estimates of the key parameters of the codon model). One gene that stands out from all the others is the ATP8 gene, exhibiting a dN/dS ratio that is between 3 to 15 times higher than that of the other genes and the lowest transition/transversion ratio of the genes considered. Further, the ATP8 gene is the only gene with a rate heterogeneity parameter lower than 1, implying that most sites have very low substitution rates (Yang, 1996a, b), but few sites also have high rates, whereas all the other genes have on average more intermediate rates across sites. The COX1, COX2, COX3, CYTB and ND1 genes have clearly lower dN/dS ratios, and the CYTB, ND1 and ND3 genes have a much higher transition/transversion ratio. Finally, allowing for different relative rates between the different genes shows that certain genes (such as ATP6 and CYTB) may evolve twice as fast as some of the other genes (e.g. ATP8, COX1 and COX2).

Gene partitioning allows capturing variation in the substitution process but considerably increases the number of parameters that need to be estimated, in particular for the GTR-based codon position partitioned nucleotide substitution models. Because standard diffuse priors offer little protection against

over-parameterization, and prior specification impacts marginal likelihood estimates, we also explore HPM approaches allowing to share information between different genes through hierarchical prior specification (Suchard *et al.*, 2003). We put hierarchical priors on the gene-specific κ and ω parameters of the GY94 codon model, on the α parameters and the five free evolutionary parameters of the GTR model. Log marginal likelihoods, shown in Table 3, were calculated using the same settings as in Tables 1 and 2.

The HPM approach offers marked improvements of the marginal likelihoods, and as expected, this is more pronounced for the parameter-rich codon position partitioned nucleotide models (about 300 log units higher) compared with the codon model (about 100 log units higher). As a consequence, a GTR-based nucleotide model partitioned according to gene and codon position now yields the highest marginal likelihood. Similar to the partitioning with independent prior specification, we can notice discrepancies in model selection outcome between the HME and sHME on the one hand, and PS and SS on the other (Table 3).

In the past decades, codon models have largely been avoided for phylogenetic reconstruction because they are computationally prohibitive. Recent development of dedicated APIs and high-performance computing libraries have, however, made it possible to harness the large numbers of computing cores available in graphics cards, which is particularly useful in the case of datasets with many unique site patterns (Suchard and Rambaut, 2009; Ayres *et al.*, 2012), such as the one analysed here. At the time of the introduction of these techniques in statistical phylogenetics, the dataset analysed here could not be analysed in double precision on what were state-of-the-art graphics cards at that time because they were limited in the amount of memory. Hence, to analyse this dataset using codon models, three graphics cards were combined in the analysis of this dataset (Suchard and Rambaut, 2009), amounting to a considerable cost

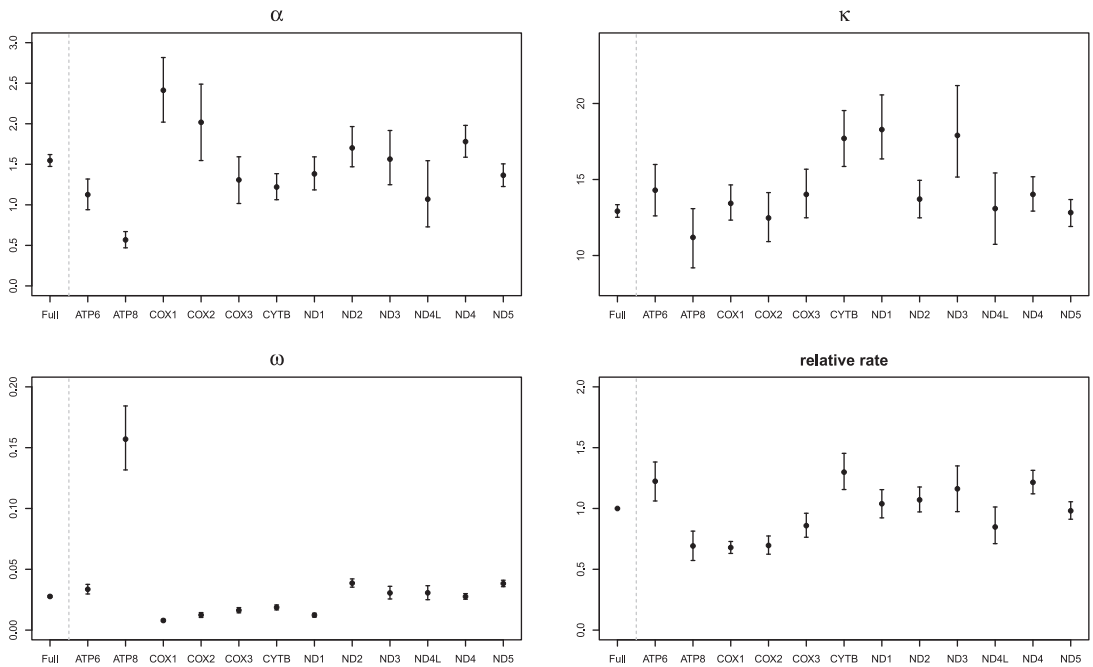

Fig. 1. Gene-specific estimates of the rate heterogeneity parameter (α), the transition/transversion ratio (κ), synonymous-non synonymous substitution bias (ω) and relative rate of evolution (μ). The estimate for the corresponding parameter in a model that does not account for the gene-specific properties is shown to the left of the vertical dotted line

Table 3. Hierarchical phylogenetic models for the best fitting codon position partitioned nucleotide models and codon model. Posterior-based model-selection approaches (HME and sHME) were run for 10 million iterations (of which 1 million serve as burn-in; 2 million for the codon models), while PS and SS were run for 64 path steps/ratios with 250 000 iterations (+25 000 iterations burn-in) per path step/ratio after an initial 1 million iterations of burn-in (2 million for codon models)

Model	Clock	Relative Rates	Resource	HME	sHME	PS	SS
$12 \times (GTR_{123} + CP_{123} + \Gamma_{123})$	RC	No	(CPU − 64 bit)	−186021.35*	−186003.13*	−186728.49	−186741.56
$12 \times (GTR_{123} + CP_{123} + \Gamma_{123})$	RC	Yes	(CPU − 64 bit)	−186041.67	−186020.56	−186710.23*	−186718.40*
$12 \times (GY94 + \Gamma)$	RC	Yes	(GPU − 64 bit)	−186214.76	−186202.85	−186749.45	−186758.76

Note: Rescaling in BEAGLE was set to default (delayed).

for this particular hardware set-up but yielding a speedup over CPU by a factor of 52 in double precision. Recently, the BEAGLE library (Ayres *et al.*, 2012) became available for use with BEAST, reporting a speedup factor of 37 for a codon model on a dataset consisting of 15 taxa and 6264 codon site patterns of which 6080 were unique, using a GTX 580 graphics card in double precision.

In Table 4, we compare the performance of a GTX 590 graphics card and a 40-core Xeon(R) E7- 4870 2.40 GHz system using a auto-sizing thread pool, two different BEAGLE rescaling schemes and different numbers of BEAGLE instances, both for a single codon model and a codon model for each gene. We demonstrate that, for a single codon model and using a delayed rescaling scheme (the default in BEAGLE) with a single-core instance, a state-of-the-art graphics card offers a speed increase with a factor over 30 compared with a single-core instance on a modern CPU running BEAST (i.e. without BEAGLE). Because the number of BEAGLE instances divides the likelihood calculations into two or more parts, thereby allowing each core to calculate part of the likelihoods corresponding to unique site patterns, multi-core or multi-GPU systems may benefit from using two or more BEAGLE instances. The speedup further increases to a factor >60 when two such BEAGLE instances are used, which allows for both GPUs on the graphics card to be used and half of the site patterns being calculated on each GPU. The speed increase provided by a multi-core CPU system levels off at a factor of 12, with the maximum performance being reached when using approximately 32 cores, illustrating that a costly multi-core CPU architecture cannot achieve the same degree of speed ups as a graphics card. No further increases in performance by using additional cores could be obtained. Although we cannot exclude communication latency as a possible cause for this observation, the most likely explanation is the considerable difference in memory bandwidth between both systems, with the GTX 590 sporting a theoretical total memory bandwidth of 327.7 Gb/s, much higher than the performance of a multi-core CPU system such as ours, yielding a typical memory bandwidth of 67 Gb/s [benchmarked using a multithreaded version of *stream*; Vladimirov (2012)].

The situation is different when attempting to fit independent codon models to each of the 12 genes and each with its own gamma distribution to model site heterogeneity. The partitioning strategy implies a change in the meaning of BEAGLE instances; the specification of 12 partitions translates into 12 likelihoods that need to be evaluated, which are naturally distributed over multiple cores as 12 instances. The specification of more BEAGLE instances partitions the likelihood calculations even more by splitting further each partition (e.g. two instances would yield 24 sets of likelihood). This implies that there is little benefit in using instances on the GTX 590 graphics card, as each of the GPUs already handles six likelihood sets. However, this approach may still profit from a 40-core CPU system, where initial increases in the number of BEAGLE instances yield the highest speedups, but there is a diminishing

Table 4. Comparison of speed estimates on different CPUs/GPUs/operating systems (OS), all run in double precision

Gene partitions	Rescaling	CPU	GPU	# instances	OS	Hours/million	Factor
No	–	Xeon(R) E7– 4870 2.40 GHz	–	–	Ubuntu 12.04	344.17	–
No	Delayed	–	GTX 590	1	Windows 7	10.53	32.68
No	Delayed	–	GTX 590	2	Windows 7	5.69	60.49
No	Delayed	–	GTX 590	4	Windows 7	6.06	56.79
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	1	Ubuntu 12.04	323.54	1.06
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	2	Ubuntu 12.04	159.80	2.15
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	4	Ubuntu 12.04	92.97	3.70
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	8	Ubuntu 12.04	48.75	7.06
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	16	Ubuntu 12.04	33.46	10.29
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	24	Ubuntu 12.04	28.30	12.16
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	32	Ubuntu 12.04	27.15	12.68
No	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	40	Ubuntu 12.04	37.82	9.10
No	Always	–	GTX 590	1	Windows 7	13.77	24.99
No	Always	–	GTX 590	2	Windows 7	7.79	44.18
No	Always	–	GTX 590	4	Windows 7	7.87	43.73
Yes	–	Xeon(R) E7– 4870 2.40 GHz	–	–	Ubuntu 12.04	192.55	–
Yes	Delayed	–	GTX 590	1	Windows 7	5.70	33.78
Yes	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	1	Ubuntu 12.04	171.95	1.12
Yes	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	2	Ubuntu 12.04	123.64	1.56
Yes	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	4	Ubuntu 12.04	94.25	2.04
Yes	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	8	Ubuntu 12.04	78.13	2.46
Yes	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	16	Ubuntu 12.04	69.63	2.77
Yes	Delayed	Xeon(R) E7– 4870 2.40 GHz	–	24	Ubuntu 12.04	72.73	2.65
Yes	Always	–	GTX 590	1	Windows 7	7.70	25.01

Note: Processors available are a 4-core Intel Core i7 running at 3.40 Ghz and a 40-core Intel Xeon(R) CPU E7- 4870 setup running at 2.40 GHz (release: Q2 2011). The GPU available is an nVidia GeForce GTX 590 (release: Q1 2011) which consists of 2 GPU units, each consisting of 512 cores and each equipped with 1.5 Gb of memory. The timing estimates are obtained for the GY94 codon model with a relaxed molecular clock (UCLD) and they are based on 100.000 MCMC iterations.

return in speed up for further instances as the number of likelihoods considerable outweighs the number of cores available. Simultaneously estimating 12 codon models proves daunting even on a multi-core CPU system, whereas on a GPU there is no performance difference compared with estimating a single codon model (although the MCMC chain will have to run longer to get adequate ESS values when estimating multiple codon models).

4 DISCUSSION

The Bayesian phylogenetic model comparison we present here consistently shows that partitioning by gene yields an increased model fit. Using standard diffuse priors, a separate codon model for each gene accompanied with gene-specific among-codon rate variation and gene-specific relative substitution rates offers the best performance, followed by codon partition models and trailed by standard nucleotide models. However, when substituting the independent diffuse priors by a hierarchical prior specification over the gene-specific parameters, a more parameter-rich GTR-based nucleotide substitution model partitioned according to gene and codon position emerges again as the best fitting model. These results can only be uncovered using recent model-selection approaches, such as path sampling and stepping-stone sampling.

By demonstrating an increased model fit for gene partitioning, we corroborate the results of earlier studies, e.g. by Nylander *et al.* (2004), who combined morphology and nucleotide data from four genes in a study on model heterogeneity across data partitions. Through Bayes factor comparisons the authors showed a dramatic increase in model fit when extending two-partition models (one partition for the morphology data and one joint partition for the four genes) to five-partition equivalents (one partition for each of the four genes), emphasizing the importance of accommodating across-partition heterogeneity. The authors also showed that within-partition rate variation was by far the most important model component (i.e. much more than across-partition heterogeneity), but that the difference in fit between substitution models was only pronounced when comparing JC (Jukes and Cantor, 1969) and GTR (Tavaré, 1986). It is important to note that the Bayes factor comparisons reported in (Nylander *et al.*, 2004) are calculated using the HME (Newton and Raftery, 1994). Because convincing evidence has been presented for the poor performance of the HME in recent years (Lartillot and Philippe, 2006; Xie *et al.*, 2011; Baele *et al.*, 2012, 2013a,b), we advocate for caution when interpreting such results and encourage the use of PS and SS (over HME and sHME).

Consistent with the increased model fit for the gene-partitioned models, we observed considerable variation in evolutionary parameters across the 12 mitochondrial genes. We examined whether this parameter variation observations could be associated with the asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Faith and Pollock (2003) and Reyes *et al.* (1998) obtained the same relative gene order with respect to the duration of the single-strand state of the parental H strand: COX1 < COX2 < ATP8 < ATP6 < COX3 < ND3 < ND4L < ND4 < ND1 < ND5 < ND2 < CYTB. When comparing the relative orders of α, κ, ω and the relative rate in Figure 1 against this relative gene order, only ranking according to the relative rate results in a similar order, with the exception of ATP6 and the ND genes: COX1 < COX2 < ATP8 < COX3 < ND4L < ND5 < ND1 < ND2 < ND3 < ND4 < ATP6 < CYTB.

Our comparison of independent diffuse priors and hierarchical priors for the gene partitioned models illustrates the impact of prior specification on marginal likelihood estimation and model selection. We have previously highlighted that the outcome of Bayesian model selection is dependent on prior choices (Baele *et al.*, 2013a). Through hierarchical prior specification, HPMs offer a middle ground between the extreme scenarios of independently fitting different models across genes and fitting a single model to all genes (Suchard *et al.*, 2003). While accommodating parameter variation among genes will be appropriate in most cases, there is less information available in each gene to inform the gene-specific parameters. HPMs allow borrowing of strength of information from one partition by another, providing more precise gene-specific parameter estimates, and resulting in further model improvements in our comparisons. BEAST supports a generic implementation of hierarchical prior specification and HPM approaches can therefore be applied to different problems, such as HIV within-host evolution for different (groups of) patients (Edo-Matas *et al.*, 2011) and phylogeographic problems (Cybis *et al.*, 2013).

By adopting hierarchical prior specification across gene-specific parameters in GTR nucleotide substitution models with both gene and codon position partitions, we notice a better model fit compared with gene partitioning with a standard codon substitution models. However, we are essentially comparing the most complex parametrization among conventional nucleotide substitution models with the simplest codon substitution model, and many assumptions can still be relaxed to make codon models more realistic. To illustrate this point, we observe a marginal likelihood improvement of about 300 log units or more between the HKY-based and GTR-based codon position partitioned nucleotide models. So, we also expect model fit improvements for a codon model that would consider different substitution rates for the different types of nucleotide substitutions within a codon, instead of merely distinguishing between transitions and transversions as is done in the GY94 model. Such a GTR-type of model applied at the nucleotide level, but with the constraint that the nucleotide sequence must encode some full-length amino acid sequence, is the rationale of the codon substitution models in the style of the codon model of Muse and Gaut (1994). The codon model of Muse and Gaut (1994) may offer a more realistic parameterization than the GY94 model, which has no natural mechanistic interpretation at the nucleotide level (Rodrigue *et al.*, 2008), resulting in a possible increase in model fit over the GY94 model. More importantly, codon models can model varying selective pressure, but a gene-specific dN/dS is a very coarse approximation of this variation and many realistic codon codons now accommodate among-site variation in dN/dS (Nielsen and Yang, 1998), in dN and dS separately (Kosakovsky Pond and Muse, 2005), and/or among lineage variation in dN/dS (Yang, 1998). Further research is needed to implement such models in BEAST (Drummond *et al.*, 2012) and to assess their model fit.

For each of the models tested in this manuscript, be it with or without gene partitioning, a relaxed molecular clock (with

underlying lognormal distribution; UCLD) is shown to outperform a strict clock, using all of the estimators. Given that mammalian datasets of mitochondrial DNA exhibit a wide variation in substitution rate across lineages, additional clock models, such as autocorrelated [see e.g. Drummond *et al.* (2006)] or random local clocks (Drummond and Suchard, 2010) should ideally be included in our model comparison. For example, Nabholz *et al.* (2008) have shown that the distribution of estimated mitochondrial substitution rates across species shows a very large variance, with the rates spanning two orders of magnitude. The authors also show that the family taxonomic level explains 75% of this variance, while the order taxonomic level explains 21%, indicating that entire orders could all have (for example) low substitution rates, which may be appropriately modeled using autocorrelated relaxed clock models (Thorne *et al.*, 1998).

Finally, we have reported massive increases in computation speed using the BEAGLE library for BEAST in combination with the latest graphics cards. However, the GTX 590 we used here is essentially designed to offer tremendous single precision performance, as required for visualization purposes in the gaming community. Hence, its double precision performance is not keeping the same development pace, which will also be the case for future cards from the same series. Double precision performance has recently increased with the advent of a new line of nVidia Tesla K20 graphics cards, designed for scientific computing. Further research will be needed to determine to what extent these new cards can improve efficiency in the field of phylogenetics.

ACKNOWLEDGEMENTS

We thank the associate editor and two anonymous reviewers for their helpful comments. We acknowledge the support of the National Evolutionary Synthesis Center (NESCent) through a working group (Software for Bayesian Evolutionary Analysis).

Funding: The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864.

Conflict of Interest: none declared.

REFERENCES

Ayres,D.L. *et al.* (2012) BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.*, **61**, 170–173.

Baele,G. *et al.* (2011) Context-dependent codon partition models provide significant increases in model fit in atpB and rbcL protein-coding genes. *BMC Evol. Biol.*, **11**, 145.

Baele,G. *et al.* (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.*, **29**, 2157–2167.

Baele,G. *et al.* (2013a) Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol. Biol. Evol.*, **30**, 239–243.

Baele,G. *et al.* (2013b) Make the most of your samples: bayes factor estimators for high-dimensional models of sequence evolution. *BMC Bioinformatics*, **14**, 85.

Bevan,R.B. *et al.* (2007) Accounting for gene rate heterogeneity in phylogenetic inference. *Syst. Biol.*, **56**, 194–205.

Bull,J.J. *et al.* (1993) Partitioning and combining data in phylogenetic analysis. *Syst. Biol.*, **42**, 384–397.

Cybis,G.B. *et al.* (2013) Graph hierarchies for phylogeography. *Phil. Trans. R. Soc. B, Biol. Sci.*, **368**, 20120206.

Drummond,A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.

Drummond,A.J. and Suchard,M.A. (2010) Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, **8**, 114.

Drummond,A.J. *et al.* (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.

Delsuc,F. *et al.* (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet*, **6**, 361–375.

Edo-Matas,D. *et al.* (2011) Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models. *Mol. Biol. Evol.*, **28**, 1605–1616.

Eisen,J.A. and Fraser,C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science*, **300**, 1706–1707.

Faith,J.J. and Pollock,D.D. (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics*, **165**, 735–745.

Friel,N. and Petitt,A.N. (2008) Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. B.*, **70**, 589–607.

Gernhard,T. (2008) The conditioned reconstructed process. *J. Theor. Biol.*, **253**, 769–778.

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

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

Jeffroy,O. *et al.* (2006) Phylogenomics: the beginning of incongruence? *Trends Genet.*, **22**, 225–231.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In: Munro,H.M. (ed.) *Mammalian protein metabolism*. Academic Press, New York, pp. 21–132.

Kosakovsky Pond,S.L. and Muse,S. (2005) Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.*, **22**, 2375–2385.

Kumar,S. *et al.* (2012) Statistics and truth in phylogenomics. *Mol. Biol. Evol.*, **29**, 457–472.

Lartillot,N. and Philippe,H. (2006) Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, **55**, 195–207.

Lemey,P. *et al.* (2012) A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, **28**, 3248–3256.

Muse,S.V. and Gaut,B.S. (1994) A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, **11**, 715–724.

Nabholz,B. *et al.* (2008) Strong variations of mitochondrial mutation rate across mammals - the longevity hypothesis. *Mol. Biol. Evol.*, **25**, 120–130.

Newton,M.A. and Raftery,A.E. (1994) Approximating Bayesian inference with the weigthed likelihood bootstrap. *J. R. Stat. Soc. B*, **56**, 3–48.

Nielsen,R. and Yang,Z.H. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

Nylander,J.A. *et al.* (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.*, **53**, 47–67.

Reyes,A. *et al.* (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.*, **15**, 957–966.

Rodrigue,N. *et al.* (2008) Bayesian comparisons of codon substitution models. *Genetics*, **180**, 1579–1591.

Rubinstein,N.D. *et al.* (2011) Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. *Mol. Biol. Evol.*, **28**, 3297–3308.

Shapiro,B. *et al.* (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.*, **23**, 7–9.

Suchard,M.A. and Rambaut,A. (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatic*, **25**, 1370–1376.

Suchard,M.A. *et al.* (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.*, **52**, 649–664.

Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Waterman,M.S. (ed.) *Some Mathematical Questions in Biology: DNA Sequence Analysis*. American Mathematical Society, Providence (RI), pp. 57–86.

Thorne,J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.

Vladimirov,A. (2012) *Terabyte RAM servers: memory bandwidth benchmark and how to boost RAM bandwidth by 20% with a single command*. Technical report.
Xie,W. *et al*. (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, **60**, 150–160.
Yang,Z. (1996a) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, **11**, 367–372.
Yang,Z. (1996b) Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.*, **42**, 587–596.
Yang,Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.*, **15**, 568–573.