Bioinformatics, 31(13), 2015, 2061–2065
doi: 10.1093/bioinformatics/btv121
Advance Access Publication Date: 27 February 2015
Discovery Note

OXFORD

Phylogenetics

Mammalian genome evolution is governed by multiple pacemakers

Sebastián Duchêne and Simon Y. W. Ho*

School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on October 14, 2014; revised on February 17, 2015; accepted on February 20, 2015

Abstract

Genomic evolution is shaped by a dynamic combination of mutation, selection and genetic drift. These processes lead to evolutionary rate variation across loci and among lineages. In turn, interactions between these two forms of rate variation can produce residual effects, whereby the pattern of among-lineage rate heterogeneity varies across loci. The nature of rate variation is encapsulated in the pacemaker models of genome evolution, which differ in the degree of importance assigned to residual effects: none (Universal Pacemaker), some (Multiple Pacemaker) or total (Degenerate Multiple Pacemaker). Here we use a phylogenetic method to partition the rate variation across loci, allowing comparison of these pacemaker models. Our analysis of 431 genes from 29 mammalian taxa reveals that rate variation across these genes can be explained by 13 pacemakers, consistent with the Multiple Pacemaker model. We find no evidence that these pacemakers correspond to gene function. Our results have important consequences for understanding the factors driving genomic evolution and for molecular-clock analyses.

Availability and implementation: ClockstaR-G is freely available for download from github (https://github.com/sebastianduchene/clockstarg).

Contact: simon.ho@sydney.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The evolutionary process leaves genomic signatures that can be analyzed using phylogenetic methods. For example, methods based on molecular clocks can be used to estimate evolutionary rates and timescales (Zuckerkandl and Pauling, 1962). However, these methods must account for the complexities of molecular evolution at the genomic scale, which is governed by interactions between mutation, selection and drift. These factors lead to different patterns of rate variation among lineages and across loci (Gaut *et al.*, 2011; Ho and Duchêne, 2014)

Factors that cause rate disparities across loci are known as gene effects. Gene effects can be partly attributed to differences in the strength and direction of selection between genes, which lead to disparities in the proportion of sites that are free to vary (Dickerson, 1971). Evolutionary rates can also vary substantially among lineages, owing to differences in generation time, DNA repair mechanisms and other traits (Bromham, 2009). These forms of rate variation are known as lineage effects (Gillespie, 1989). Lineage and gene effects can co-occur, leading to a pattern of among-lineage rate variation that is sustained across genes, but with differences among the absolute gene-specific rates.

A more complex pattern arises from the interaction between gene and lineage effects, known as residual effects (Gillespie, 1989). Under this scenario, the pattern of among-lineage rate variation can differ across loci. This can occur if selective constraints vary across loci in a lineage-specific manner (Muse and Gaut, 1997). For example, a set of genes might be under positive selection in some lineages, but under purifying selection in other lineages (Gaut *et al.*, 2011). Changes in population size can also cause residual effects because they alter the proportion of sites that are effectively neutral (Takahata, 1987).

Partitioning rate variation into gene, lineage and residual effects is difficult (Smith and Eyre-Walker, 2003), but the influence of these

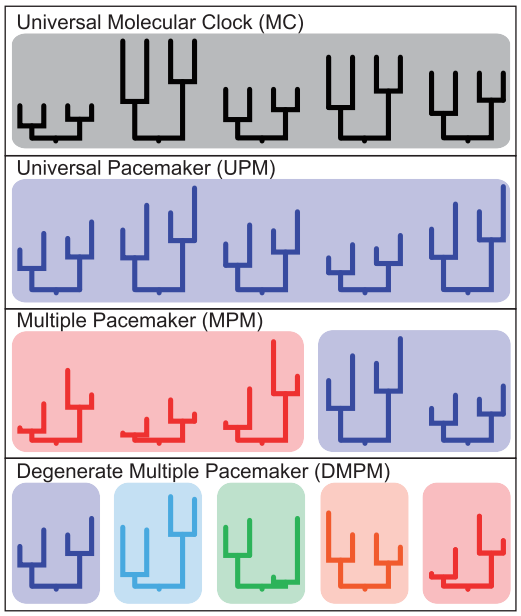

Fig. 1. Pacemaker models of genome evolution. Illustration of four models of genome evolution using gene trees. In the Universal Molecular Clock (MC) model, the evolutionary rate is constant among lineages but varies across genes. In the Universal Pacemaker (UPM) model, the evolutionary rate varies across genes and among lineages. However, all genes share the same pattern of among-lineage rate variation. In the Multiple Pacemaker (MPM) model, groups of genes share the same pattern of among-lineage rate variation. In the Degenerate Multiple Pacemaker (DMPM) model, each gene has a distinct pattern of among-lineage rate variation

three factors at the genomic scale can be described using the pacemaker models of molecular evolution (Fig. 1; Ho 2014; Snir *et al.* 2012). The Universal Molecular Clock (MC) is the simplest model of genomic evolution (Zuckerkandl and Pauling, 1962). It posits a constant rate of evolution among lineages, but allows the rate to vary across loci. In this model, rate variation is governed by gene effects only. The Universal Pacemaker (UPM) model assumes that both lineage and gene effects are present, but that residual effects are negligible. Therefore, a single pattern of among-lineage rate variation is shared across all loci. The Multiple Pacemaker (MPM) model suggests that there is a limited number of pacemakers, such that there are groups of genes with the same pattern of among-lineage rate variation, but with different absolute rates. This implies that residual effects within pacemakers are very small compared with those between pacemakers. In the Degenerate Multiple Pacemaker (DMPM) model, residual effects are pervasive and there is a distinct pattern of among-lineage rate variation for each gene. This extreme case of gene-specific among-lineage rate variation is also known as 'erratic' evolution (Ayala *et al.*, 1996; Rodríguez-Trelles *et al.*, 2001).

Support for the UPM model has come from studies of various organisms, including archaea, bacteria, plants, fungi and *Drosophila* species (Snir *et al.*, 2012, 2014; Wolf *et al.*, 2013). In these analyses, the UPM model was preferred over the MC, the MPM, and DMPM models. These studies compared models using goodness-of-fit statistics, but these are limited because they do not allow testing of all of the possible scenarios under the MPM model. The MPM hypothesis is much more complex than the UPM or the DMPM because it comprises a large family of models: the n loci in a data set can be assigned to one of k pacemakers, where k can take any value from 2 to $n-1$. Given the large number of ways in which the loci can be assigned to pacemakers, testing the goodness-of-fit of every scenario under the MPM model is computationally intractable. This problem can be overcome by using clustering algorithms (Duchêne and Ho, 2014; Snir, 2014).

Here, we analyze a publicly available data set of 431 genes from 29 mammalian taxa to test the different pacemaker models. We tested the UPM model using a clustering algorithm implemented in a new version of our program, ClockstaR (Duchêne *et al.*, 2014), which we have modified so that it can be used to analyze genome-scale data sets. We find that the patterns of evolutionary rate variation in the mammalian data set can be explained by 13 pacemakers, supporting the MPM model. Our results suggest that there is a degree of stability in genome evolution. However, we find no evidence of association between pacemakers and gene function.

2 Methods

We compiled a publicly available data set comprising 431 nuclear protein-coding genes from 33 mammalian taxa (Song *et al.*, 2012). We removed four taxa (horse, New World bat, Old World bat and tree shrew) because their phylogenetic placement in most gene trees did not match that in the species tree, leaving 29 taxa. This is a crucial requirement of our method, which assumes the same topology across all gene trees. We did not analyze the data sets from previous studies of the pacemaker models (Snir *et al.*, 2012, 2014; Wolf *et al.*, 2013) because they do not meet this assumption. Moreover, they contain large proportions of missing data, which can mislead the estimates of tree distances that form an important component of our analysis.

To evaluate the MPM model, we estimated gene-trees under maximum-likelihood implemented in phangorn v1.99 (Schliep, 2011). For each gene we selected the nucleotide substitution model according to the Bayesian information criterion. We optimized the branch lengths by fixing the tree topology to the species tree inferred for these data (Song *et al.*, 2012). We analyzed the gene trees using a modified version of the ClockstaR algorithm (Duchêne *et al.*, 2014), ClockstaR-G.

The algorithm in ClockstaR-G involves scaling the length of each gene tree to 1. This step controls for the possibility that genes differ in their absolute evolutionary rates, and allows us to focus on the proportional differences among branch lengths (i.e. the pattern of among-lineage rate variation). The scaled branch lengths are used as individual dimensions in Euclidean space to calculate the distance between trees. This is possible because all genes are assumed to share the same topology, such that we always compare the same branches among genes. In this respect, trees with similar patterns of among-lineage rate variation are expected to have small pairwise distances between them. A pacemaker is a cluster of trees in this space. For example, under the MC and the UPM models (Fig. 1), there is a single cluster of gene trees with pairwise distances of 0. In the MPM model, there is more than one cluster of gene trees, and the gene-tree distances within each cluster are necessarily smaller than those between clusters. In the DMPM model, the gene trees are randomly distributed across Euclidean space, such that it is not possible to identify discrete clusters. Clustering algorithms can be used to identify the pacemakers. ClockstaR-G uses the Clustering for Large Applications (CLARA) algorithm, which is efficient for large data sets (Kaufman and Rousseeuw, 2005).

The optimal number of clusters can be identified via the Gap statistic (Tibshirani *et al.*, 2001). We use CLARA to assign genes to each of the k clusters and we calculate the average silhouette width

($\langle S_k \rangle$), a measure of clustering fit (Rousseeuw, 1987). We then draw a number of random samples equal to the number genes from a multidimensional uniform distribution with the same dimensions and range of values as those of the gene trees. We cluster these samples using CLARA and we calculate $\langle S_k \rangle$ for every value of k. We replicate this procedure 500 times. The Gap statistic is computed by subtracting the $\langle S_k \rangle$ estimated for the original data from $\langle S_k \rangle$ estimated for randomly drawn samples, such that there is a range of Gap_k values for every k. The optimal k is that with the highest average Gap_k, ($\langle \text{Gap}_k \rangle$) and for which the 95% confidence interval does not overlap with that of Gap_{k-1} (Tibshirani *et al.*, 2001). $\langle S_k \rangle$ cannot be computed for $k = 1$ or for $k = N$. In these cases, the 95% confidence intervals of successive values of k are expected to overlap consistently. To select $k = 2$, Gap_2 should not overlap with Gap_3. ClockstaR-G allows the computation of $\langle S_k \rangle$ across different CPUs. The program is written in the R programming language v3.0.1 (R Core Team, 2008) and is freely available from an online repository (github.com/sebastianduchene/clockstarg).

We calculated the isolation and mean dissimilarity for the each of the pacemakers under the optimal k. The isolation is used to determine whether a cluster is well separated from other clusters, with low values indicating a high degree of separation. The mean dissimilarity measures the average cohesiveness of a cluster, with low values indicating that the points within a cluster are in close proximity. According to these statistics, a pacemaker is well defined if it has low values for the isolation and mean dissimilarity.

In ClockstaR-G, the UPM and the MC models are indistinguishable because the tree distance used here is expected to be 0 in both cases. We tested the MC model by concatenating the genes and analyzing the data in baseml, part of the PAML package v4.8 (Yang, 2007). The substitution model was chosen according to the Bayesian information criterion. We used a likelihood-ratio test to compare the fit of a free-rates model, in which there are no constraints on branch lengths, against a strict-clock model.

We used a simulation approach to test the sensitivity of our clustering method to detect different numbers of pacemakers. This involved simulating k trees, each with branch lengths drawn from a different lognormal distribution with a mean of 1 and a standard deviation of 30% of the mean. In our method of generating trees, the branch lengths in a tree are independently and identically distributed, as is commonly assumed in phylogenetic analyses (e.g. Ronquist *et al.*, 2012). We simulated the evolution of 431 nucleotide alignments with the same sequence length and substitution model parameters as those of the mammalian genes. The trees were rescaled to match the individual gene-tree lengths. We simulated a similar number of gene alignments along each of the k trees. For example, for $k = 10$ there were nine pacemakers with 43 genes each and one pacemaker with 44. We chose values for k of 1, 2, 5, 10, 13, 100 and 431. To analyze these data, we used the same method as that for the mammalian genes. We replicated our simulations five times for each k. This method allows us to assess the expected results under the UPM ($k = 1$), and the DMPM ($k = 431$).

Given that, in the MPM model, all genes are assigned to a limited number of pacemakers, the question arises as to whether these pacemakers are associated with any features of the genes. For example, genes with similar features might be subject to the same selective constraints across a range of taxa, leading to similar patterns of among-lineage rate variation. To answer this question, we analyzed the correspondence between pacemakers and several gene features: sequence length, nucleotide composition, protein family, C–G content and gene-tree length. We included these features in a classification tree and in a random forest with the pacemaker assignment as the target variable, as implemented in the R packages Tree v1.0 (Ripley, 2014) and RandomForest v4.6 (Liaw and Wiener, 2002).

The classification-tree algorithm uses one gene feature at a time as a binary classifier, known as a decision node. The features are added recursively to the tree until all the genes are classified into pacemaker. Notably, a particular feature can be used at multiple nodes in the tree. To avoid over-fitting, the least informative nodes are pruned by using a k-fold cross-validation (Hastie *et al.*, 2009). The features can be ranked in terms of their importance by counting the number of times that they are used in the tree. The random forest is an ensemble learning method that uses a large number of classification trees (Breiman, 2001). In this algorithm, feature importance is determined by removing one feature at a time, and computing the increase in misclassification error. We chose these methods over other classification algorithms, such as logistic regression, because they do not make parametric assumptions and because the inclusion of categorical predictor variables is straightforward.

3 Results and discussion

The MC model provided a poor fit to the concatenated alignment of mammalian genes. The log-likelihood of the unconstrained model (–8 400 383) was much higher than that of the strict-clock model (–8 463 250), such that the MC model was strongly rejected ($P < 0.01$).

Our test of the MPM model suggested that the optimal number of pacemakers for this data set is 13, with $\langle \text{Gap}_{13} \rangle = 0.14$, compared

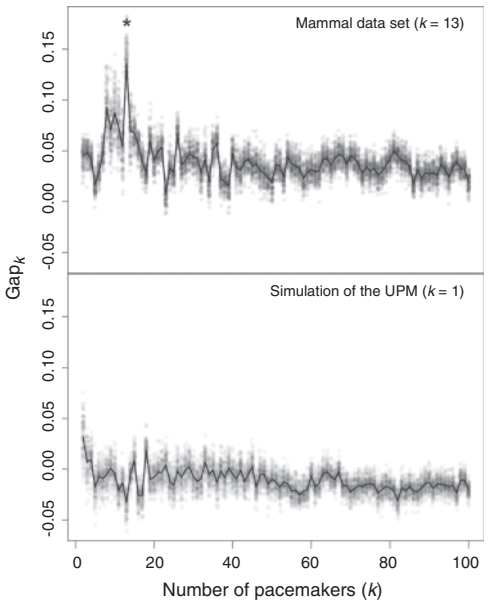

Fig. 2. Statistical fit of different numbers of pacemakers to 431 genes from 29 mammal taxa and one replicate of 431 genes simulated under the Universal Pacemaker (UPM) model. Statistical fit is measured by the Gap statistic, Gap_k. Results are shown for pacemakers (k) ranging from 2 to 100 and values of Gap_k for k>100 are shown elsewhere (Supplementary Fig. S2). The solid line represents the average Gap_k, and the gray points represent the values estimated from 500 bootstrap replicates for each number of pacemakers. Large values of Gap_k indicate high statistical fit. To select the optimal number of pacemakers, we consider the value of k with the highest average Gap_k, and for which the 95% confidence interval does not overlap with those of Gap_{k-1} and Gap_{k+1} For the mammal data set, the asterisk (*) denotes the optimal number of pacemakers. In the UPM simulations, the values for Gap_k overlap among all successive values of k, so it is not possible to identify an optimal number of pacemakers

with $\langle Gap_{12} \rangle = 0.053$ and $\langle Gap_{14} \rangle = 0.07$. The 95% confidence interval of Gap_{13} was 0.12–0.16, which did not overlap with those of Gap_{12} or Gap_{14}, which were 0.02–0.09 and 0.04–0.10, respectively (Fig. 2). For the data simulated under the UPM ($k = 1$), the 95% confidence intervals between all successive values of k overlapped (Fig. 2). This indicated low support for all numbers of pacemakers, such that the default value of $k = 1$ was selected. The different trends in the Gap_k between the mammalian data and the synthetic data generated according to the UPM model provide strong evidence in support of the MPM model. Our simulations of a range of values of k demonstrated that our method almost always identified the correct number of pacemakers, at least under the idealized conditions represented by simulated data. The exceptions were two simulations with $k = 100$, for which the method detected 101 and 97 pacemakers (Supplementary Fig. S1).

There was high variation in the isolation and mean dissimilarity of the pacemakers. However, some pacemakers had large numbers of genes with low mean dissimilarities, suggesting that a large proportion of genes had patterns of among-lineage rate variation that were very similar to each other. For example, pacemaker PM1 included 257 genes with a mean dissimilarity of 2.98, which is lower than that of most of the other pacemakers (Table 1). The most isolated pacemakers, such as PM10 through PM13, had small numbers of genes. This probably occurs because some genes have distinct patterns of among-lineage rate variation, such that they are equidistant to two or more clusters and the algorithm assigns them uniquely to a pacemaker.

We did not find a clear association between the gene features and the pacemaker assignments. There were high misclassification errors for the classification tree (38%) and for the random forest (40%), indicating that only approximately 60% of the genes could be accurately assigned to a pacemaker on the basis of gene features. Therefore, the pacemakers governed clusters of genes with different functions and molecular characteristics, such as nucleotide composition.

The only gene feature with some explanatory power was gene-tree length, which was the most frequent variable in the classification tree and was the most important variable in the random forest. This variable allowed us to distinguish between PM1 and PM2, which are the largest pacemakers and which have a mean difference in tree length of 0.25 substitutions/site (Table 1). This is perhaps because long trees are a consequence of high rates of evolution, which imply weak selective constraints. In these cases, lineage effects would be the dominant source of rate variation. In contrast, genes that yield short trees might be subject to stronger selective constraints, such that they are governed by residual effects. Consequently, there would be fewer pacemakers for genes with long trees than for genes with short trees. This result stands in contrast with those of Du *et al.* (2013), who did not find any relationship between gene-specific evolutionary rates and the degree of among-lineage rate heterogeneity. Further data from a range of taxonomic groups will enable this hypothesis to be tested more comprehensively.

Table 1. Features of genes assigned to 13 pacemakers in the mammalian genome, ordered by number of genes

Pacemaker	Number of genes	Mean tree length (subs/site)	Mean isolation	Mean dissimilarity
PM1	257	1.63	1.77	2.98
PM2	60	1.38	1.79	3.03
PM3	31	1.67	1.77	3.45
PM4	22	1.83	1.33	3.51
PM5	22	1.52	1.26	3.37
PM6	14	1.77	1.37	3.37
PM7	5	1.47	1.07	3.91
PM8	5	1.37	0.82	2.69
PM9	5	1.30	1.03	3.50
PM10	4	1.40	1.02	3.21
PM11	3	1.47	1.05	3.33
PM12	2	1.83	0.72	1.78
PM13	1	0.94	0	0

Wolf *et al.* (2013) proposed that there is stability in among-lineage rate variation throughout the genome. They showed that the UPM model explains a large proportion of the gene-specific rate variation among bacterial and archaeal lineages. Snir *et al.* (2014) also found support for the UPM model in *Drosohpila* and yeast genomes. Our analyses of the mammalian data set favor the MPM model, but the number of pacemakers is small compared with the total number of genes (13 pacemakers for 431 genes). This lends support to the hypothesis of stable genome evolution because genome-wide rate variation can be clustered into a small number of pacemakers.

Our method has appeared to perform well in analyses of data from simulations and mammals, but might be less reliable when the level of within-cluster variation is high compared with that between clusters. Further investigation of its performance under a wider range of simulation scenarios, including its sensitivity to potential confounding factors, will be valuable. In addition, improved genome annotation and understanding of gene-specific functions will enable elucidation of the factors that determine the number and influence of genomic pacemakers. This opens the way for improving schemes for assigning multiple relaxed-clock models in phylogenetic analyses of evolutionary timescales (dos Reis *et al.*, 2012; Duchêne and Ho, 2014; Ho and Duchêne, 2014). Furthermore, analyzing the genomes of a wider range of organisms will help to determine whether stability is a ubiquitous property of genome evolution. Our method provides a potential means of answering such questions about patterns of evolutionary rate variation on a genomic scale.

Funding

SD was supported by a Francisco José de Caldas Scholarship from the Colombian government and by a Sydney World Scholars Award from the University of Sydney. SYWH was supported by a Queen Elizabeth II Fellowship from the Australian Research Council (grant number DP110100383).

Conflict of Interest: none declared.

References

Ayala,F.J. *et al.* (1996) Molecular clock or erratic evolution? A tale of two genes. *Proc. Natl. Acad. Sci. USA*, **93**, 11729–11734.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Bromham,L. (2009) Why do species vary in their rate of molecular evolution? *Biol. Lett.*, **5**, 401–404.

Dickerson,R.E. (1971) The structure of cytochrome *c* and the rates of molecular evolution. *J. Mol. Evol.*, **1**, 26–45.

Du,X. *et al.* (2013) Why does a protein's evolutionary rate vary over time? *Genome Biol. Evol.*, **5**, 494–503.

Duchêne,S. *et al.* (2014) ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics*, **30**, 1017–1019.

Duchêne,S. and Ho,S.Y.W. (2014) Using multiple relaxed-clock models to estimate evolutionary timescales from DNA sequence data. *Mol. Phylogenet. Evol.*, **77**, 65–70.

Gaut,B. *et al.* (2011) The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. Syst.*, **42**, 245–266.

Gillespie,J.H. (1989) Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.*, **6**, 636–647.

Hastie,T. *et al.* (2009) *The elements of statistical learning*. Springer, New York.

Ho,S.Y.W. (2014) The changing face of the molecular evolutionary clock. *Trends Ecol. Evol.*, **29**, 496–503.

Ho,S.Y.W. and Duchêne,S. (2014) Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.*, **23**, 5947–5975.

Kaufman,L. and Rousseeuw,P.J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis*. 1st ed. Wiley, Hoboken, NJ.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.

Muse,S. V. and Gaut,B.S. (1997) Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics*, **146**, 393–399.

Dos Reis,M. *et al.* (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. London B*, **279**, 3491–3500.

Ripley,B. (2014) Tree: classification and regression trees. R package version 1.0-35.

Rodríguez-Trelles,F. *et al.* (2001) Erratic overdispersion of three molecular clocks: GPDH, SOD, and XDH. *Proc. Natl. Acad. Sci. USA*, **98**, 11405–11410.

Ronquist,F. *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Schliep,K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

Smith,N.G.C. and Eyre-Walker,A. (2003) Partitioning the variation in mammalian substitution rates. *Mol. Biol. Evol.*, **20**, 10–17.

Snir,S. (2014) Pacemaker partition identification. In: Darling,A. and Stoye,J. (eds.) *Algorithms in Bioinformatics*. Springer, New York, pp. 281–295.

Snir,S. *et al.* (2012) Universal pacemaker of genome evolution. *PLOS Comput. Biol.*, **8**, e1002785.

Snir,S. *et al.* (2014) Universal pacemaker of genome evolution in animals and fungi and variation of evolutionary rates in diverse organisms. *Genome Biol. Evol.*, **6**, 1268–1278.

Song,S. *et al.* (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. USA*, **109**, 14942–14947.

Takahata,N. (1987) On the overdispersed molecular clock. *Genetics*, **116**, 169–179.

Tibshirani,R. *et al.* (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, **63**, 411–423.

Wolf,Y.I. *et al.* (2013) Stability along with extreme variability in core genome evolution. *Genome Biol. Evol.*, **5**, 1393–1402.

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

Zuckerkandl,E. and Pauling,L. (1962) Molecular disease, evolution and genetic heterogeneity. In: Kasha,M. and Pullman,B. (eds.) *Horizons in Biochemistry*. Academic press, New York, pp. 189–225.