Factors influencing the accuracy and precision in dating single gene trees

Guillaume Louvel[1,2], Hugues Roest Crollius[2]

1. Écologie Systématique Évolution (ESE), Université Paris-Saclay, CNRS, AgroParisTech, 91190 Gif-Sur-Yvette, France
2. École Normale Supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École Normale Supérieure (IBENS), F-75005 Paris, France

Corresponding author: Guillaume Louvel, guillaume.louvel@normalesup.org

Abstract

Molecular dating is the inference of divergence time from genetic sequences. Knowing the time of appearance of a taxon sets the evolutionary context by connecting it with past ecosystems and species. Knowing the time of appearance of gene lineages through duplications or transfers would provide a context to understand adaptation at the genomic level. However, molecular clock inference faces uncertainty due to the variability of the rate of substitution between species, between genes and between sites within genes. When dating speciations, per-lineage rate variability can be informed by fossil calibrations, and gene-specific rates can be either averaged out or modeled by concatenating multiple genes. By contrast when dating gene-specific events, fossil calibrations only inform about speciation nodes and concatenation does not apply to divergences other than speciations.

This study aims at benchmarking the accuracy of molecular dating applied to single gene trees, and identify how it is affected by gene tree characteristics. We analyze 5205 alignments of genes from 21 Primates in which no duplication or loss is observed. We also simulated alignments based on characteristics from Primates, under a relaxed clock model, to analyze the dating accuracy. Divergence times were estimated with the bayesian program Beast2.

From the empirical dataset, we find that the precision of dating decreases with shorter alignments, high rate heterogeneity between branches and low average rate, features that underlie the amount of dating information in alignments, hence statistical power. The best precision is associated with core biological functions such as ATP binding, cellular organization and anatomical development, probably because such constrained genes evolve more constantly. We then investigated the accuracy of dating on simulated alignments, by controlling the three above parameters separately. It confirmed the factors of precision, but also revealed biases when branch rates are highly heterogeneous. This corroborates findings that relaxed clock inferences are mainly driven by the tree prior when calibrations are lacking and rate heterogeneity is high. Our study finally provides a general scale of parameters that influence the dating precision and accuracy, which can be generalized to other taxa. To tackle molecular dating events like deep coalescence, horizontal gene transfers and gene duplications, future models should overcome the lack of power due to limited information from single genes.

Keywords: molecular clock, molecular dating, uncertainty, gene tree, primates, phylogenetics, phylogenomics

Introduction

The phylogenetic tree is the most prevalent concept for representing evolutionary relationships, but a species tree and a gene tree do not necessarily tell the same story: while nodes in a species tree represent speciations, meaning the divergence of sister species from their parent, nodes in a gene tree represent the divergence of genes which do not result directly from speciation, but from the coalescence of alleles or from locus duplication and horizontal transfer. Therefore reconstructing a species tree generally requires sequences assumed to be orthologous and congruent with the species history; subsequent inferences such as dating taxon divergences usually lump together many genes to increase statistical power, for example by concatenating orthologs. However, the specificities of gene evolution are not only a nuisance that hinders species tree inference, they are also the raw material of morphological evolution, adaptation and species diversification. In this regard, focusing on single genes instead of concatenations can reveal gene-specific evolutionary processes. In particular, genes have heterogeneous substitution dynamics due to functional specificities or location in the genome, or as a consequence of events of duplication and transfer. Estimating the rate and timing of sequence evolution is the subject of molecular clock dating and we evaluate here its accuracy and precision when applied to single genes.

The molecular clock is a set of methods to date the divergence between DNA or protein sequences by linking absolute time to the number of observed substitutions (Zuckerkandl and Pauling 1962; Margoliash 1963; Doolittle and Blombäck 1964; Zuckerkandl and Pauling 1965). The model considers a sequence that evolves by substitution over a given lapse of time t. It states an approximately proportional relationship between time t and the number of substitutions N, by modelling N as a Poisson distributed random variable of expectation $r \times t$, where r is the rate of substitutions per unit of time. Thus an average of $r \times t$ substitutions occur in the duration t. In the simplest model the rate is assumed to be constant over the entire phylogenetic tree (Langley and Fitch 1974). The number of substitutions can be obtained from the branch lengths inferred with classic phylogenetic tree inference programs (PhyML, Iqtree, RaxML, Phylobayes, MrBayes, PAUP, etc), but to infer the rate, times must be known, and vice-versa. Fossils typically provide independent dating and are used to assign ages to some of the speciation nodes, allowing the rate to be derived and all non-calibrated nodes to be dated.

However while rate constancy is a convenient simplification, it has been observed that rates of substitution per unit of time vary across taxa (Wu and Li 1985; Britten 1986; Pagel et al. 2006). The causes of variations are diverse and may include population size, impact of selection at the molecular level, generation time, efficiency of the DNA repair system, metabolism, etc (Gillespie 1991). For some of these reasons, rates also vary across the genome (Wolfe et al. 1989). At the scale of a single sequence, the heterogeneity of the rate across branches does not necessarily follow the same pattern between sites, meaning that different sites accelerate or decelerate in an independent manner. This is called heterotachy (literally “different rates” Philippe et al. 2003). Some authors (Kumar and Hedges 1998) have

argued that by selecting many genes these variations are attenuated, their mean rate is approximately constant and the molecular clock can still be used, especially in recent clades. Elsewhere it is generally accepted that such simple models can generate overconfident but wrong results (Revell et al. 2005; Lozano-Fernandez et al. 2017).

Initially, tests of constancy such as the relative rate tests (e.g. in Sarich and Wilson 1966; Fitch and Langley 1976; Gu and Li 1992; Tajima 1993) or the likelihood-ratio test of the clock (Felsenstein 1981) were developed to assess these rate variations. Nowadays, most models accommodate a non constant molecular clock providing that enough calibrations are specified. How to relax the clock constancy is not settled: autocorrelated rates assume a progressive change between connected branches (Thorne et al. 1998) while uncorrelated clock rates just follow an overall distribution (Drummond et al. 2006). In the uncorrelated case, the branch-wise rates are most commonly modeled as a lognormal, gamma or exponential, but alternatively the "white noise" model has the interesting property that the variance of the rate is smaller on longer branches (Lepage et al. 2007). Different empirical studies support either the uncorrelated rates model (Drummond et al. 2006; Rannala and Yang 2007; Linder et al. 2011; Heath et al. 2012) or the autocorrelated clock rates (Thorne et al. 1998; Lepage et al. 2007; dos Reis et al. 2018; Smith et al. 2018), while a mixture of the two models over different timescales has also been proposed (Lartillot et al. 2016; Bletsa et al. 2019). However it is uncertain that autocorrelation can actually be detected (Ho et al. 2015; Tao et al. 2019) and the modalities of the rate relaxation are likely specific to each taxonomic group, timescale studied and calibration scheme. Algorithmic implementations include widely used software such as Beast 2 (Bouckaert et al. 2019) and MCMCTree (Rannala and Yang 2007) which exploit the Bayesian framework through Markov Chain Monte Carlo (MCMC) sampling, while some faster alternatives take advantage of least squares strategies, like LSD (To et al. 2016) and treedater (Volz and Frost 2017). Accelerating probabilistic algorithm has also been made possible by using approximate likelihoods (Thorne et al. 1998; Guindon et al. 2010; dos Reis and Yang 2011).

In contrast to the early ages of molecular biology in the nineteen-sixties where analyses could only be performed on a single protein at a time such as haemoglobins, cytochrome c or fibrinopeptides (Margoliash 1963; Doolittle and Blombäck 1964; Zuckerkandl and Pauling 1965), modern high-throughput sequencing gives access to hundreds or thousands of genes per genome, providing a more general insight. The modern procedure to date species divergences therefore uses many loci whose sequences are concatenated. This concatenation approach aims to increase the amount of information, thus the precision (Bromham et al. 2000; Lanfear et al. 2010; Smith et al. 2018; Bletsa et al. 2019). Even in a favorable situation with many genes, the accuracy of molecular dating has been questioned, because of flawed procedures such as reusing inferred dates as secondary calibrations without propagating uncertainty (Graur and Martin 2004; Schenk 2016), but also because of inherent limitations of the problem (Pulquério and Nichols 2007; Burbrink and Pyron 2008; dos Reis and Yang 2013; Zhu et al. 2015; Kumar and Hedges 2016; Warnock et al. 2017); the most critical limitations are the difficulty of characterizing the type of clock relaxation and the uncertainty in calibration points themselves (dos Reis et al. 2015).

If accurate dating requires sufficient sequence information, this means that inferring dates in virus, transposable element or single gene evolutionary histories is even more difficult than with concatenations of hundreds of genes. Indeed, on such short sequences, low statistical power amplifies the following known limitations of dating by molecular clock: methodological errors in tree topologies and substitution rates, rate variation of multiple biological origins and inherent stochasticity in the substitution process, all of which create a disproportionate amount of noise.

Here we characterize the level of accuracy in dating nodes from single gene trees, and what characteristics of the gene tree are related to a higher accuracy. We first evaluate a dataset of primate genes under a cross-validation procedure where we date speciation events in each tree and compare each individual tree estimate with the median age over all trees. Using the median age as a point of reference produces a measure of the precision of dating. In these primate gene trees we find characteristics associated with a higher precision (as measured by the deviation from the median), mainly the length of the sequence alignment, the heterogeneity of the rate between branches and the mean rate of the tree. In addition to measuring precision we also compare with reference ages from TimeTree, and observe a bias towards younger ages. Because several explanations for such bias cannot be disentangled from the empirical dataset we simulated gene trees to measure how the precision and accuracy of dating varies depending on alignment length, rate heterogeneity between branches and mean rate of substitution.

Results

Single gene estimates show a high dispersion and a bias towards younger ages

We perform a benchmark on dating speciation nodes from single gene trees with an empirical dataset of trees that include genes from 21 primate species obtained from Ensembl version 93. We focus on the Simiiformes clade as ingroup, which contains all Primates but the Lemuriformes that we use as outgroup. Speciation ages for all nodes have been estimated many times independently based on fossil-calibrated dating on concatenation of genes, and resulting consensus ages can be obtained from TimeTree (Kumar et al. 2017). Because our aim is to replicate the uncertainty in dating nodes that lack fossil calibrations, we do not set any calibration except for the Simiiformes ancestor at 43.2 My (C.I [41.0, 45.7]). We then quantify the uncertainty relative to the surrounding interval of 43.2 My. The choice of this root calibration and its associated uncertainty is arbitrary because all trees are then compared by this yardstick. Likewise, the choice of Primates for the source trees is arbitrary; the specific selection of species does not matter, what matters is that we collect natural replicates of the same tree. We selected 5205 gene trees that do not display any duplication or loss in this tree of 21 primate species, so that each gene tree shares exactly the same topology. These were dated with Beast 2, providing 5205 age estimates at each internal node (fig. 1).

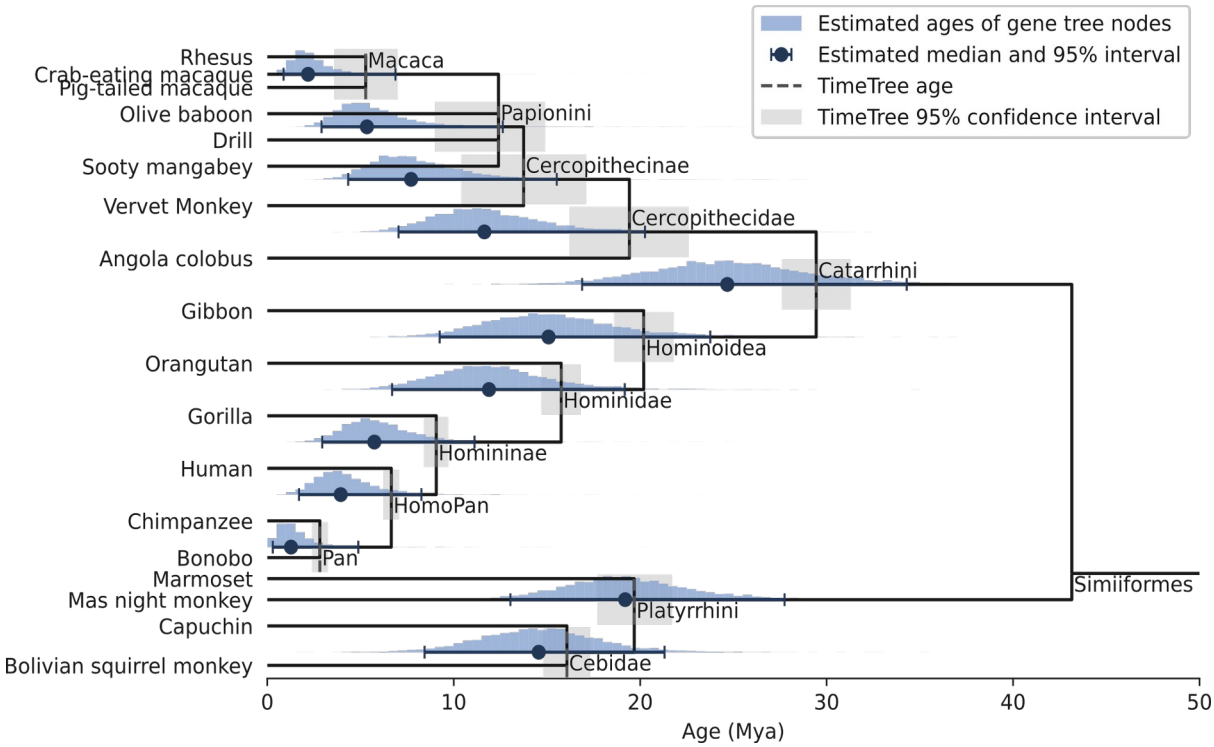

Figure 1: Distribution of speciation ages estimated on primate gene trees independently (5205 trees). Each histogram is rescaled to display the same height, but their total count is each equal to 5205. Dating was performed with Beast under the nucleotide substitution model HKY and a relaxed clock rate (see Methods).

By comparison with the reference age estimates from TimeTree, we obtain clearly younger estimates on all nodes, sometimes even outside of the confidence interval from the reference (median, fig. 1). The highest shift of 10 My affects Cercopithecidae. This suggests that either our estimates are biased, or less plausibly that the reference ages themselves are inaccurate. In our dating procedure, several simplifications may bias ages. To start with, calibrating only one node is insufficient, but this is precisely the purpose of our analysis, since we study gene trees for which nodes lack calibrations. With one calibration but variable branch rate, rate and time are unknown because they are conflated into the branch length. In fact, molecular clock "dating" is as much about estimating rates using calibrations than estimating dates from rates, this shift in focus being due to the high variability of molecular rates. Without many calibrations, variation in rates cannot be faithfully infered, and in turn dates remain uncertain. Accordingly, our empirical dataset is expected to display heterogeneous branch rates of different origins. First there are species-specific rate variations, in which all genes experience the same trend in a specific lineage (also called lineage effects, Ho 2014). For example the average branch length of Cercopithecidae is higher than their Hominoidea sister (supp. info. S1). This is consistent with the known generation times in these clades: based on Pacifici et al. (2013) the averages are 4035 days for Cercopithecidae versus 6132 for Hominoidea. In addition to species trends, each gene tree may experience

particular variations of branch rates, therefore producing dispersed ages when compared (called residual effects, i.e. variation remaining after gene and lineage effects, Ho 2014). Finally, heterotachy also likely occurs within genes, between sites. However, these gene and site heterotachies should just cause dispersion but not loss of accuracy, because we can expect that the average of gene ages should fall on the correct value (after correction for lineage specific rates). Another simplification of our inference is to consider instantaneous segregation of genes at speciations. In reality, speciation takes up to several million years and segregated genes can correspond to older allelic divergence. This phenomenon of deep coalescence should cause speciations to appear older than they are, and not younger as is the case in our results. Conversely, shallow coalescence caused by introgression between recently diverged species would produce younger age, but we deem implausible that introgression would be so pervasive among genes so as to strongly bias ages. In summary, variation in lineage specific rates is the most plausible explanation for our bias towards younger ages, unless the inference obtained from Beast is not statistically consistent.

Aside from comparisons with the reference ages, we quantified the dispersion of the estimates at each node with the following metrics: the mean absolute deviation from the median (MAD) of the node age and the 95% inter-quantile range (IQR95); the averages of these metrics over the 12 speciations yield 3 My and 16 My, respectively. These dispersions are larger for deeper nodes, such as Catarrhini with a MAD of 5.6 My and an IQR95 of 32.4 My. Divided by the Simiiformes calibration of 43.2 My, they represent 13 % and 75 % of this interval, respectively. In other words, the confidence interval for the age of Catarrhini is barely shorter than the age of Simiiformes.

In the following, we break down the causes leading to the dispersion of our estimates by measuring gene features and searching for a relationship with the dating precision.

The dispersion of age estimates is associated with low statistical signal and branch rate variation

We computed 56 features from each gene tree related to alignment quality and substitution rate estimations (supp. info. S2). We regressed them against the absolute deviation from the median age averaged across all speciation nodes, chosen to describe the dating imprecision in a single gene tree. In order to select a restricted set of independent variables we used a Lasso regression with a regularization parameter ‘alpha’ of 0.02, which retained 10 parameters further fitted by Ordinary Least Squares (adjusted $R^2 = 0.288$, fig. 2). Variables were normalized so that coefficients can be compared and they were centered. Apparent outliers, such as trees with extreme rates, were discarded from the regression (see Methods - Removing trees with outlier values), which was therefore based on 5170 trees.

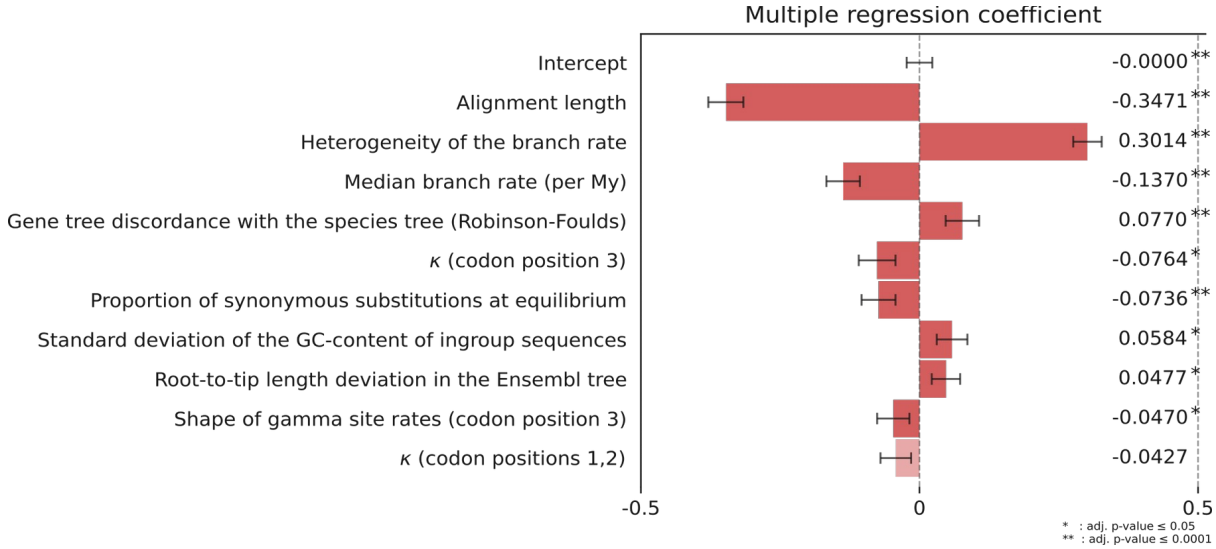

Figure 2: Coefficients of regression of the gene tree features against the age dispersion. Features were selected with a Lasso regression with parameter alpha=0.02. Adjusted $R^2 = 0.288$. Error bars indicate the 95% confidence intervals.

By order of coefficient size, the length of the sequence alignment comes first (coefficient 0.35), associating short alignments with high dispersion. It is followed by the rate heterogeneity across branches which is positively associated with the dispersion (0.30). Next comes the mean rate of substitution (-0.14). We hypothesize two distinct causes for these correlations: first, the strength of the statistical signal is linked to the alignment length and to the mean rate, because both influence the number of observable substitutions, thus the amount of information available to infer ages. Second, some model parameters are in practice difficult to infer individually. In particular, a variable branch rate cannot be estimated separately when there are no internal calibrations, meaning that it could take arbitrary values as long as the product of rate by time equals the branch length in substitutions. Consequently, it would take values dictated by the tree prior (the Birth-Death branch process) that cannot possibly recover the exact Simiiformes speciation dates (dos Reis and Yang 2013). Here the rate heterogeneity is the standard deviation of the rate across branches, as inferred by Beast. This indicates that while Beast manages to detect a high rate heterogeneity in a given gene tree, the ages it infers in such a tree are quite distinct from the median ages, and probably far from their true age. The fourth largest association is a measure of the incongruence between the species tree and the gene tree reestimated using IQtree. This means that lack of support for the species topology is associated with low dating precision. This could be attributable to either actual incongruence being masked by the reconciliation step in our dataset, to low signal from short sequences or to such departure from a strict clock that the true topology cannot be recovered.

The next five coefficients have p-values below 0.05, and correspond to characteristics of the substitution process or the alignment, for which we provide the following interpretations. The kappa parameter (for codon position 3), or ratio of transitions over transversions, is

negatively correlated. It may relate to the degree of saturation: saturation brings this ratio closer to one, and it should increase dating uncertainty, although we think that very few gene trees show saturation in our Primates dataset. Next is the proportion of synonymous substitutions at equilibrium, negatively correlated. This value that we measured with 'codeml' from PAML depends on the alignment and represents the amount of substitutions that would be synonymous if the ratio dN/dS was equal to one. This result indicates that gene sequences that offer more opportunity for synonymous changes favor precise dating. For the standard deviation of the GC-content between sequences of the alignment, the correlation is positive. This could have two causes that are not mutually exclusive, one methodological, one biological: in the substitution model, sequence composition is assumed homogeneous. Only few phylogenetic models can actually model composition changes over the tree (Foster 2004) while this is known to highly impede phylogenetic inferences. Under a purely biological explanation so even if we fitted the appropriate model, genes subject to composition shifts could be evolving heterogeneously in other aspects, in particular with heterogeneous rates. The next variable is the standard deviation of root-to-tip path lengths, an approximation of the branch rate heterogeneity that can easily be computed on any phylogram. It is, as expected negatively correlated but weakly, probably because our more precise measure of across-branch rate heterogeneity already captures most of the effect. In (Smith et al. 2018) it is used as a criterion for selecting trees which are suitable for dating, but here we see that its predictive power is limited. Last to be significant, the shape of the site-wise gamma distribution of rates is a parameter in which a lower value means a more skewed gamma distribution. Thus, as indicated by the negative coefficient (fig. 2), rates that are heterogeneous site-wise (not only branch-wise) appear to also increase dating uncertainty.

To situate these features in their real scale, we summarized them in the 10% of trees with highest or lowest predicted dispersion (table 1). Additionally, we searched for gene functions overrepresented in these two extreme subsets compared to the full set of 5170 genes, based on human Gene Ontology (GO) annotations (table 2, supp. info. S3). For the most dispersed trees, no enriched function was detected. On the other hand, well dated trees (low dispersion) appear to have constrained functions. Negative selection on this set is further evidenced by 67 "Human Phenotype Ontology" overrepresented terms related to various developmental abnormalities (nervous system, eye, ear morphology, movement, see supp. info. S3). These functions are a subset of the functions retrieved when using the 10% longest genes (supp. info. S4), indicating that the length is the main driver of the functions found in the set of genes that have the highest dating precision.

Subset of gene families	Observed dispersion (My)	Alignment length (nucl)	Rate heterogeneity (10^{-4} subst/nucl/My)	Mean rate (10^{-3} subst/nucl/My)
10% highest predicted dispersion	3.45 ± 1.6	1008 ± 506	11.3 ± 9.76	2.16 ± 1.25
10% lowest predicted dispersion	1.41 ± 0.576	5255 ± 3247	4.69 ± 3.94	2.75 ± 1.13

Table 1: Characteristics of gene families whose dispersion is predicted to be the lowest/highest (mean ± standard deviation)

	GO category	GO term IDs	GO term names
10% lowest predicted dispersion	Molecular Function	GO:0005524	ATP binding
		GO:0032559	Adenyl ribonucleotide binding
		GO:0030554	Adenyl nucleotide binding
	Biological Process	GO:0071840	Cellular component organization or biogenesis
		GO:0016043	Cellular component organization
		GO:0007010	Cytoskeleton organization
		GO:0006996	Organelle organization
	Cellular Compartment	GO:0005604	Basement membrane
		GO:0043228	Non-membrane-bounded organelle
		GO:0043232	intracellular non-membrane-bounded organelle

Table 2: Overrepresented Gene Ontology terms in gene families with lowest/highest predicted dispersion, using g:Profiler with default parameters: significance threshold 0.05 and multiple testing correction g:SCS.

In this regression of gene tree characteristics, the explained variable (deviation from the median) is a measure of dispersion, not of accuracy. However we can only measure accuracy if we have access to the "true" age. Besides, the distribution of ages (fig. 1) suggests a bias in our estimates that empirical data alone is powerless to explain. We therefore performed a simulation to confirm the features that impact dating accuracy.

Simulating gene tree variability impact on dating

Simulations allow to measure two facets of the dating quality: first the precision or how fine-grained the estimates are; second and more importantly, the accuracy or how close the estimates are from the true value on average. We focused on the top 3 variables obtained from the regression: alignment length, across-branch rate heterogeneity and mean rate of substitution. Our three variables of interest were set to representative values spanning their observed range in the real trees (supp. info. S5). For each set of parameters, we first simulated trees according to a relaxed clock model, then simulated codon sequence evolution along this tree (see Methods - Simulating alignments). Dates were then reconstructed from

simulated data with the same method as for the empirical dataset. We expected the three variables to reproduce the same effect on dispersion as with the above empirical dataset. However we expected the accuracy to be unaffected because the model we use to simulate alignments and trees is very similar to the model used to reconstruct dates: Beast dating was thus performed without discrepancy between its inference model and the process that generated the data, and in this condition we expect it to be statistically consistent, i.e. converge to the true value as the amount of data increases.

We ran 500 simulations for each parametrization, by fixing two parameters and varying the third, around a central point defined by length=3000 nucleotides, mean rate=0.004 substitutions/codon/My and branch rate heterogeneity=¼ (defined as the standard deviation divided by the mean of the branch rates, abbreviated σ/μ).

The dispersion, as seen by the interquantile range covering 95% of the data (IQR95) (fig. 3) varies with each variable in the same direction as observed in the regression of primate gene trees: it increases for shorter alignments (3b), higher rate heterogeneity (3c) and lower evolutionary rate (3d). In almost all speciation nodes and parameter values, we recover an unbiased median age, falling accurately on the age from the underlying species tree. However, we find shifts for the highest rate heterogeneity between branches (3d, σ/μ = 1). Cebidae appears younger while Catarrhini appears older than in reality, an effect that we find when sampling from the prior (supp. info. S6). This shows that in presence of very high across-branch rate variation and uninformative calibrations, the prior on the time tree (Birth-Death) strongly influences the ages.

Discussion

Factors of precision

From an empirical dataset of 5205 gene trees in 21 Primates, we identified characteristics related to the dispersion of age estimates. More statistical power, provided by higher rates or longer alignments, lowers the dating dispersion while an increase in rate variation between branches increases the dispersion. This confirms previous observations that the number of sites and substitutions is a limiting factor with regard to statistical power (Bromham et al. 2000; Lanfear et al. 2010; Smith et al. 2018; Bletsa et al. 2019; Bromham 2019), and also that the clock rate does not appear to be constant across branches. In our analysis the dispersion also correlates with characteristics of the substitution process (kappa, GC content), indicating either that the dating model is too simple, or that multiple kinds of heterogeneities (composition, rates) co-occur in some trees. The 5205 selected gene trees only contain strict orthologs and as such represent a biased fraction of the 24,614 gene trees extracted from Simiiformes: this selection likely represents a more reliable subset for dating, because the absence of gene duplication and loss may be associated with a more stable evolution in terms of constancy of rate and selective pressure.

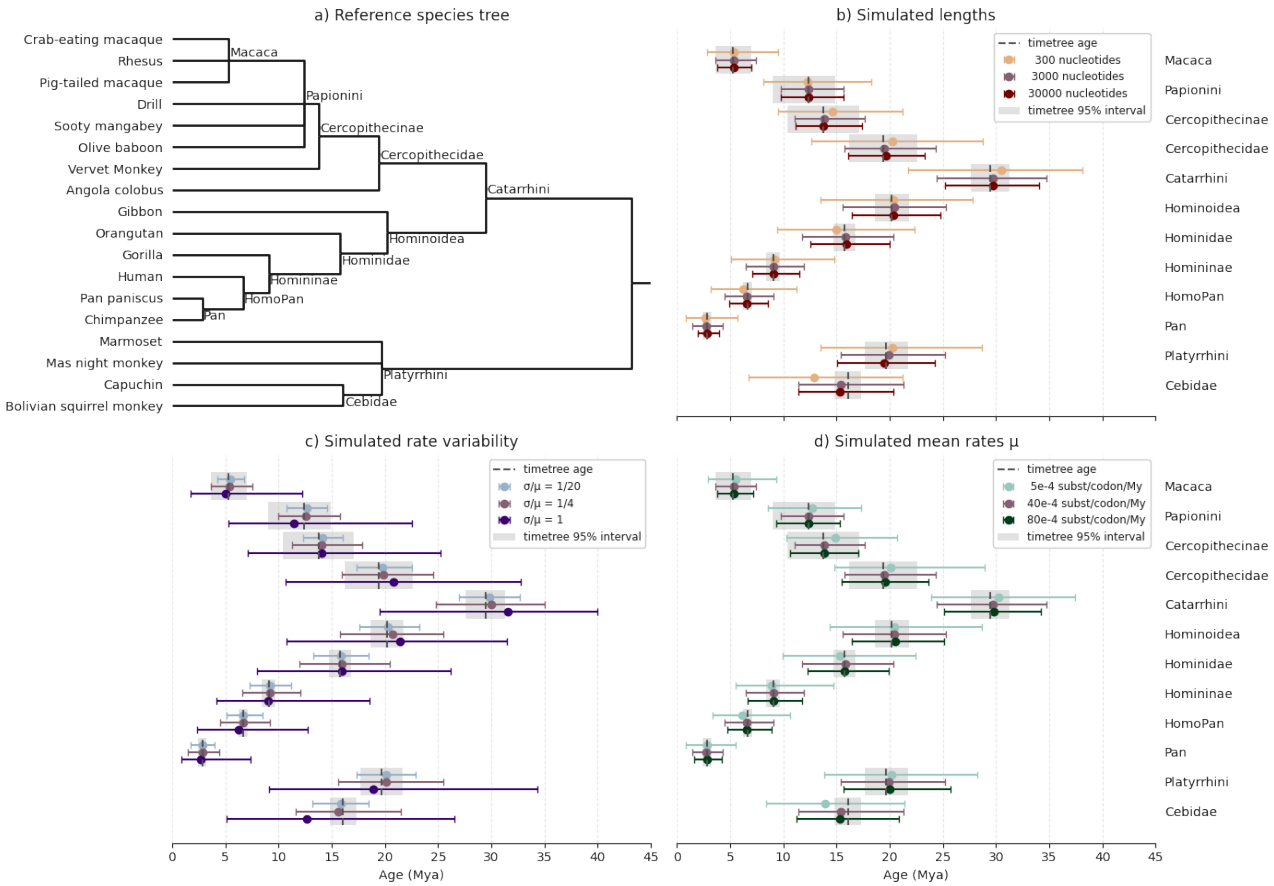

Figure 3: Speciation dates on simulated alignments, under a relaxed clock model and a fixed species tree of Simiiformes, under various parameter values. Central dots represent the median date, and the interval represent the range [0.025, 0.975] of the distribution. a) Dated species tree from TimeTree; b) the number of sites in the alignments, in number of nucleotides; c) the diffusion parameter, defining the degree of relaxation of the clock rate, such that the ratio σ/μ (standard deviation of the rates over the mean rate) equals one twentieth, one quarter, and one; d) The mean rate of substitution, in number of substitutions per site per million years. When not specified, length is 3000, rate is 40e-4 and σ/μ is ¼.

In the regression based on empirical gene trees, the adjusted R^2 is low (0.29), indicating a low fraction of explained variance. To investigate the missing variance, we should incorporate additional biological characteristics of the sequences, but as we show in simulations, stochastic properties of the mutational and evolutionary processes already account for a large uncertainty. Furthermore in our empirical dataset, incomplete lineage sorting (ILS), introgression or reciprocal paralog loss may also obfuscate the relationship between molecular change and time, but we have not used models able to identify these processes. Among other simplifications we used, our dating method employs a nucleotide substitution model which cannot distinguish neutral from non-neutral substitutions, the latter being usually more clock-like in absolute time, whereas neutral substitutions tend to show a generation time effect (Ho 2014). A codon model would be able to separate synonymous from non-synonymous substitutions and test this hypothesis, but being more complex (using

61 character frequencies instead of 4) it also requires more data, which as we have seen is the limiting factor at the scale of single gene trees.

Limiting the bias when dating single gene trees

In the empirical estimates, we observed a shift towards ages younger than the TimeTree reference. Then we simulated sequences without species-specific rates and that bias was not reproduced. This is unsurprising because we used the same model for inference as for simulating the data, and thus were expecting statistical consistency. However in the simulated case with very high rate heterogeneity, the ages are biased by the time tree prior (Birth-Death model). In the real primate trees it is likely that the bias in instead driven by species-specific rates: this creates systematic trends as opposed to the random heterogeneity in the simulations. In such circumstances the model with a relaxed clock requires calibrations on internal nodes to properly infer branch rates and times (Duchêne et al. 2014). Here, the choice of a single calibration was made precisely with the aim to measure the dating accuracy on uncalibrated nodes, such as those occuring in gene trees if we consider events other than speciations. Some studies use the relative ordering of events from different taxa (Lutzoni et al. 2018) or from different gene families (Pittis and Gabaldón 2016; Vosseberg et al. 2021), but across-branch rate heterogeneity is a major hurdle to doing so and conditions for applicability need to be rigorously verified, in particular the absence of systematic bias (Susko et al. 2021).

How to reduce uncertainty

Stochastic uncertainty is the amount of information that is lost assuming that the model is correct, as opposed to *inductive uncertainty* due to inadequate modelling (Holland 2013). Inadequate modelling generally produces systematic errors. Our simulations show what level of stochastic uncertainty to expect in gene trees with realistic rate characteristics, and they confirm the drivers of uncertainty identified in empirical gene trees. We highlight the requirement for a high number of sites and substitutions, which is critical for statistical power. This problem can probably only be circumvented by incorporating information beyond single genes, such as genomic context and linked model parameters between multiple gene trees, as recent developments suggest (Duchêne et al. 2020). Our study also reinforces the notion that fossil calibrations are quantitatively more important to accurate dating than sequence data alone. These solutions are unfortunately harder to design for gene trees with duplications or transfers. Some pragmatic approaches that have been proposed to handle gene specific heterogeneity are the removal of loci with apparently non constant rate prior to the analysis (Jarvis et al. 2014), a “gene-shopping” approach (Smith et al. 2018), but discarding data might be unsatisfactory for genome scale analyses that look for the most general picture. Our results may be extended to a similar gene filtering approach as it pinpoints gene features that are correlated with the dating uncertainty, although its predictive power would need to be improved.

Regarding *inductive uncertainty*, richer models avoid overconfident but erroneous results. Also, when the inference of tree topology, substitutions and clock rates is performed jointly by an integrative method such as Beast, it guarantees that uncertainty is properly cumulated at all steps. However in our experience it appears challenging to run a complex parametric inference such as Beast at the scale of comparative genomics. Such integrative methods might not be numerically tractable because they generally require MCMC which implies prohibitive running times and in addition, MCMC requires a thorough and experimented human inspection to validate the inference output. In the face of this, numerous projects aiming at sequencing broad segments of biodiversity on Earth are emerging, under the auspices of the Earth Biogenome Project (EBP, Lewin et al. 2018; Lewin et al. 2022). Extremely large gene families, composed of tens of thousands of genes, will become common place. Therefore the development of faster non-parametric algorithms is still relevant, alongside improvements in the computational footprint of probabilistic algorithms.

Research on the molecular clock enters an exciting time, with huge amounts of data and increasingly sophisticated methods to dissect the hidden mechanisms of sequence evolution. How the rate of molecular evolution varies still holds mysteries, both at the scale of lineages and at the scale of genomes. Future developments answering this question will be methodologically challenging but shall shed light on important evolutionary processes.

Methods

Source and number of trees

The species tree and gene trees were obtained from Ensembl Compara 93 (July 2018; Zerbino et al. 2018) Metazoa dataset: 99 species descending from Opisthokonta (the last common ancestor of fungi and metazoan) and 23,904 reconciled gene trees. Low quality genomes, aberrant gene branch lengths, split genes were removed (supp. info. S7, Species and gene trees preprocessing).

Nodes from the species tree were assigned an age from TimeTree (data retrieved Jan. 2019; Kumar et al. 2017). The corresponding sequences of the coding domain of the longest transcribed isoforms (fasta format) were downloaded as fasta multiple sequence alignments via the Ensembl Perl API.

Focusing on the 21 primate species (18 Simiiformes and 3 Lemuriformes), we extracted 24,614 subtrees at their Simiiformes node plus two of the outgroup sequences with shortest branch lengths.

From those we selected the 5235 gene trees without duplications and losses within Simiiformes, so that they share a single tree topology.

Alignment and cleaning

We built codon alignments: protein sequences were aligned using FSA version 1.15.9 (Fast Statistical Alignment, Bradley et al. 2009) with default parameters, then back-translated to the corresponding nucleotide (codon) alignment.

HmmCleaner (version 0.180750, Di Franco et al. 2019) finds segments of sequence that appear inconsistent with the other aligned sequences, interpreting them as sequencing errors. We apply it and replace the detected segments by gaps. It was applied to the 5235 alignments that do not have duplications or losses, on the amino-acid data which were then backtranslated to codons. Out of these, 30 alignments caused HmmCleaner to fail, resulting in the 5205 subtrees under study here.

Bayesian node dating with Beast 2

Using the alignments as input for Beast 2.6.3 (Bouckaert et al. 2019) we ran the following model: a HKY nucleotide substitution model with two site partitions corresponding to the codon positions {1,2} and 3. The Birth-Death model" is set as the tree prior (Gernhard 2008), and a relaxed uncorrelated lognormal clock rate is fitted on branches, with mean and standard deviation being estimated. All 12 Simiiformes clades (see species tree in fig. 1) as well as Primates and Lemuriformes were constrained to be monophyletic (note that Macaca, Papionini and Platyrrhini are polytomic). Only Primates and Simiiformes were calibrated precisely, the others having uninformative priors. The calibrations were specified as gamma distributions manually tuned to match TimeTree point estimate and 95% interval: offset 70.8 My, shape 4.6 and scale 0.656 for Primates, and offset 40.9 My, shape 4.0 and scale 0.575 for Simiiformes (therefore the mean prior age for Primates is $70.8 + 4.6 \times 0.656 = 73.8$ My, and for Simiiformes it is $40.9 + 4.0 \times 0.575 = 43.2$ My). BEASTGen (v1.0.1) was used to automatically generate the parameter file for each alignment from a common template. The MCMC was run in one chain of 20,000,000 iterations, after a pre-burnin of 1,000,000. The 390 trees which had an ESS (effective sample size) under 200 for any variable were resumed and extended with 20,000,000 additional iterations, leading to 13 trees with one ESS under 200 that we discarded. The resulting mean dates were annotated on the species tree using TreeAnnotator from Beast 2.

Global age dispersion from all primate trees

Given that all speciation nodes in the primate species tree are replicated in our set of 5205 gene trees, we measured the dispersion of estimated ages for each speciation. We used two metrics of dispersion, the MAD and the IQR95. The mean absolute deviation from the median (MAD) is a robust estimate of dispersion: it is less influenced by skewed distributions than when using the mean as center, and less influenced by outliers than the standard deviation. It is also still sensitive enough compared to median deviation from the median or interquartile range. The IQR95, or interquantile range covering 95% of the data (between percentiles 2.5 and 97.5) is more robust but less sensitive.

Regression

Statistical analyses on the output dates were performed with Python 3.8 and additional scientific computing packages including Numpy, Pandas, Statsmodels, Scikit-learn, Biopython and Ete3 (supp. info. S8).

Quantifying the average deviation of ages for a single gene tree

As the dependent variable to regress, we used the average deviation of ages in each gene tree: for one gene tree, the absolute deviation from the median speciation age is obtained at each node, then the deviations of the 12 nodes are averaged.

Retrieving mean rate and rate heterogeneity per tree

The clock model that we fit in Beast is unlinked between codon positions {1,2} and position {3}, meaning that the rate parameters are infered separately for each of these site partitions. Beast outputs different types of rates: First, it outputs the sampled parameters of the uncorrelated log-normal clock model (its mean and standard deviation). Additionally, it computes a posteriori the mean and variance of the branch rate. The latter estimation differs in that it is not a parameter of the model, but a statistic that is computed at the end of each iteration on the proposed tree. We monitor both estimates because they yield quite different values, although being correlated (supp. info. S9). For each sampled tree, it then produces the mean and variance. We use the latter rate statistics as variables in the regression. Since there are as many values as sampled trees of the MCMC chain, we take the median over the chain as final value.

As we estimated separately the mean rates $m_{1,2}$ and m_3 for the corresponding codon positions, we finally combined them to obtain the total rate of substitutions per codon:

$$m = 2 \times m_{1,2} + m_3$$

This rate by codon equals three times the average rate by nucleotide, but we chose the codon metric for consistency with the simulation parameters based on a codon model in INDELible.

Similarly as measure of across-branch rate heterogeneity, we summed the rate standard deviations based on the returned variances $v_{1,2}$ and v_3 from Beast:

$$s = 2 \times \sqrt{v_{1,2}} + \sqrt{v_3}$$

Collecting features of the gene families

We measured 56 characteristics of the gene families as explanatory variables in the regression. They include (full list in supp. info. S2):

- Tree features:
  - mean bootstrap values of nodes, from Ensembl;
  - the root-to-tip deviation, or standard deviation of the path lengths from the root to the leaves, from the Ensembl trees.

- whether the gene tree was a posteriori edited to fit the species tree topology (See supp. info. S7, Species and gene trees preprocessing).

- alignment features:
  - global statistics such as alignment length and proportion of pairs of sequences that do not share sites (indicative of "split genes");
  - statistics over sequences like the mean frequency of gaps, the mean percentage of ambiguous nucleotides, the mean GC-content, the mean CpG content;
  - statistics over sites, like the column entropy and the parsimony score;
  - proportions of sequences cleaned by HmmCleaner;

- substitution process features estimated with the program 'codeml' from PAML 4.9e (Yang 2007):
  - proportion of synonymous substitutions at equilibrium, based on measured codon frequencies and the fitted substitution matrix (using names from 'codeml' output, it is S/(N+S));
  - mean and standard deviation of ω (dN/dS) under the free-ratio model;

- substitution process features estimated during the Beast dating inference, separately for codon positions {1,2} and {3}:
  - the ratio of transitions over transversions (kappa);
  - the gamma shape for site variable rates;
  - the mean clock rate;
  - the standard deviation of the clock rate;

- MCMC related statistics of the Beast runs:
  - the number of iterations the chain was run (either 20 or 40 millions);
  - whether any parameter had an ESS below 200.

Transforming regressed features

We chose the appropriate transformation of features in order to stay close to the assumptions of a linear modelling framework with ordinary least squares, that is explanatory variables with low skew. For this, a semi-automated procedure was set up, where the transform that minimizes the skew was selected from the predefined set made of 'no transform', 'logarithm base 10', and 'square root', with additional increment or sign modification if needed by the function domain of validity. Some variables were binary encoded in cases where a large proportion of values was constant, or where distributions were clearly bimodal. Transforms for each variable can be found in supp. info. S10. Finally, all variables were normalized (divided by their standard deviation) and centered before the regressions.

Reducing multicollinearity

Multicollinearity being known to impede ordinary least squares fitting, we reduce it prior to the regression with two strategies:

- first we obtained covariances from a Factor Analysis (analogous to Principal Component Analysis, but accounting for discrete and ordinal data). From the results, we removed some features from clusters of heavily correlated features. We also "decorrelated" pairs of features by dividing one by the other or by computing the residues of the simple regression between the two linked variables. For instance we decorrelated standard deviation metrics when they correlate with the mean of the same variable (supp. info. S10).
- Afterwards we checked the multicollinearity condition number, which is the square root of the highest eigenvalue divided by the smallest eigenvalue of X^TX (X being the design matrix). Since the multicollinearity condition number was already less than the usual cutoff of 20, it was not needed to remove additional features.

Removing trees with outlier values

To avoid a misleading impact of outliers on the linear regression, trees with excessive rates computed by Beast were excluded. The cutoffs were chosen based on the histograms, which display very long tails. As mentionned above, the 13 trees with insufficient ESS (less than 200) were discarded as well. Finally, alignments including pairs of sequences which do not share any common site, i.e. unaligned pairs, were also excluded, as this can lead to meaningless branch lengths. In total these filters removed 35 outlier gene trees (supp. info. S11).

Final feature selection with Lasso and OLS refitting

Lasso (Tibshirani 1996) is a regression fitting algorithm which simultaneously performs feature selection, thus dealing with highly dimensional design matrices. We therefore apply it in a first pass to discard coefficients with an absolute value inferior to 0.01, using a penalty value (parameter alpha) of 0.02. However Lasso coefficients are biased (by design), so that p-values and R^2 cannot be readily computed. For this reason we subsequently refit by ordinary least squares (OLS) using the selected features, with a covariance set to MacKinnon and White's (1985) heteroscedasticity robust type 1. Finally, the p-values were subjected to a Bonferroni correction by multiplying them by the number of features used in the Lasso step. Statsmodels 0.13.2 implementations were used. Detailed regression output statistics are in supp. info. S12.

Gene functional annotation overrepresentation

From the 5170 gene trees without duplication or loss retained in the regression, human genes were used as the reference set of functions. Overrepresentation of the 517 human genes from the most dispersed and least dispersed trees was done with the online tool g:Profiler version e111_eg58_p18_30541362 (Raudvere et al. 2019) using default parameters (significance threshold 0.05 and multiple testing correction by g:SCS).

As specified in the table 1, the set of most (or least) dispersed trees is obtained from the fitted regression line. It is therefore a predicted dispersion, distinct from the observed dispersion displayed in column 1.

Simulating alignments

Alignments were generated by simulating the evolution of a sequence along the above tree of Primates, using INDELible v1.03 (Fletcher and Yang 2009). The three tested parameters were the mean rate, the rate heterogeneity across branches and the alignment length, with 500 simulation replicates for each set of parameter values. These parameters were chosen to represent the extent of variation in the real data of 5205 trees (as measured to be used in the regression, see below and supp. info. S5). To generate the variable branch rates, we applied an independent log-normal relaxed clock simulation with the 'simclock' R package; the tested mean rates were 5×10^{-4}, 40×10^{-4} and 80×10^{-4} substitutions.site^{-1}.My^{-1}. To obtain the desired rate heterogeneities, we tuned the "diffusion" parameter s^2 so that the standard deviation of the branch rates is comparable with empirical estimates, as per the relation between standard deviation σ, mean μ and diffusion s^2 of a log-normal distribution:

$$s^2 = \log\left(1+\frac{\sigma^2}{\mu^2}\right)$$

The chosen diffusion values (s^2) of 0.6931, 0.0606 and 0.0025 correspond to a standard deviation of the branch rate equal to 1, 1/4 and 1/20 of the mean rate (σ/μ), respectively. For the alignment length, we set the root sequence lengths to 300, 3000 or 30,000 nucleotides in INDELible. The evolutionary model includes insertions and deletions, which were parametrized to occur at a rate of 8.4×10^{-7} and 16.8×10^{-7} My^{-1} respectively (corresponding to 1/14 and 2/14 of the median substitution rate 1.177 codon^{-1}.My^{-1}, a ratio taken from Fletcher and Yang 2009), with lengths following a Lavalette distribution with parameters (2, 300). The sequences evolve according to a codon model with *kappa* = 4 (transitions/transversions) based on estimates from the Primates gene trees (supp. info. S13), *omega* = 0.175 (dN/dS) and codon frequencies taken from the concatenated alignments of Simiiformes genes. No across-site rate heterogeneity was modeled. The random seed was set to 9342.

We then dated Primates speciations from these simulated sequences with Beast 2, as for the real Primates dataset above, except that the tree prior is "Calibrated Yule" for panels b and d of fig. 3. This tree prior was updated to Birth-Death in panel c because it was showing a bias towards younger ages. The number of trees with any ESS under 200 is given in supp. info. S14.

Code and data availability

The Python statistical analysis, the source data and the intermediate data are archived at Zenodo.org with the DOI: 10.5281/zenodo.8235767). The core phylogenetics/bioinformatics library is also available at https://github.com/DyogenIBENS/Phylorgs (version 0.1.0).

Acknowledgment

We wish to thank Alexandra Louis for the management of data and software resources used in this work, and Pierre Vincens for computing support.

Fundings

This work was supported by a grant from Fondation pour la Recherche Médicale to G.L. (FRM FDT201904008392).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

References

Bletsa M, Suchard MA, Ji X, Gryseels S, Vrancken B, Baele G, Worobey M, Lemey P. 2019. Divergence dating using mixed effects clock modelling: An application to HIV-1. *Virus Evol.* [Internet] 5. Available from: https://academic.oup.com/ve/article/5/2/vez036/5561482

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, Maio ND, et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* 15:e1006650.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast Statistical Alignment. *PLOS Comput. Biol.* 5:e1000392.

Britten RJ. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* 231:1393–1398.

Bromham L. 2019. Six Impossible Things before Breakfast: Assumptions, Models, and Belief in Molecular Dating. *Trends Ecol. Evol.* 34:474–486.

Bromham L, Penny D, Rambaut A, Hendy MD. 2000. The Power of Relative Rates Tests Depends on the Data. *J. Mol. Evol.* 50:296–301.

Burbrink FT, Pyron RA. 2008. The Taming of the Skew: Estimating Proper Confidence Intervals for Divergence Dates. *Syst. Biol.* 57:317–328.

Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* 19:21.

Doolittle RF, Blombäck B. 1964. Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals: Evolutionary Implications. *Nature* 202:147–152.

dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. 2015. Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Curr. Biol.* 25:2939–2950.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Duchêne DA, Tong KJ, Foster CSP, Duchêne S, Lanfear R, Ho SYW. 2020. Linking Branch Lengths Across Sets of Loci Provides the Highest Statistical Support for Phylogenetic Inference. *Mol. Biol. Evol.* 37:1202–1210.

Duchêne S, Lanfear R, Ho SYW. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Mol. Phylogenet. Evol.* 78:277–289.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368–376.

Fitch WM, Langley CH. 1976. Protein evolution and the molecular clock. *Fed. Proc.* 35:2092–2097.

Fletcher W, Yang Z. 2009. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Mol. Biol. Evol.* 26:1879–1888.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–495.

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

Gillespie JH. 1991. The Causes of Molecular Evolution. Oxford University Press Available from: https://academic.oup.com/book/53090

Graur D, Martin W. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* 20:80–86.

Gu X, Li W-H. 1992. Higher rates of amino acid substitution in rodents than in humans. *Mol. Phylogenet. Evol.* 1:211–214.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* 59:307–321.

Heath TA, Holder MT, Huelsenbeck JP. 2012. A Dirichlet Process Prior for Estimating Lineage-Specific Substitution Rates. *Mol. Biol. Evol.* 29:939–955.

Ho SYW. 2014. The changing face of the molecular evolutionary clock. *Trends Ecol. Evol.* 29:496–503.

Ho SYW, Duchêne S, Duchêne D. 2015. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol. Ecol. Resour.* 15:688–696.

Holland BR. 2013. The Rise of Statistical Phylogenetics. *Aust. N. Z. J. Stat.* 55:205–220.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. *Nature* 392:917–920.

Kumar S, Hedges SB. 2016. Advances in Time Estimation Methods for Molecular Data. *Mol. Biol. Evol.* 33:863–869.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* 34:1812–1819.

Lanfear R, Welch JJ, Bromham L. 2010. Watching the clock: Studying variation in rates of molecular evolution between species. *Trends Ecol. Evol.* 25:495–503.

Langley CH, Fitch WM. 1974. An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* 3:161–177.

Lartillot N, Phillips MJ, Ronquist F. 2016. A mixed relaxed clock model. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* 24:2669–2680.

Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. 2022. The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci.* 119:e2115635118.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* 115:4325–4333.

Linder M, Britton T, Sennblad B. 2011. Evaluation of Bayesian models of substitution rate evolution--parental guidance versus mutual independence. *Syst. Biol.* 60:329–342.

Lozano-Fernandez J, dos Reis M, Donoghue PCJ, Pisani D. 2017. RelTime Rates Collapse to a Strict Clock When Estimating the Timeline of Animal Diversification. *Genome Biol. Evol.* 9:1320–1328.

Lutzoni F, Nowak MD, Alfaro ME, Reeb V, Miadlikowska J, Krug M, Arnold AE, Lewis LA, Swofford DL, Hibbett D, et al. 2018. Contemporaneous radiations of fungi and plants linked to symbiosis. *Nat. Commun.* 9:5451.

Margoliash E. 1963. Primary Structure and Evolution of Cytochrome C. *Proc. Natl. Acad. Sci. U. S. A.* 50:672–679.

Pacifici M, Santini L, Di Marco M, Baisero D, Francucci L, Grottolo Marasini G, Visconti P, Rondinini C. 2013. Generation length for mammals. *Nat. Conserv.* 5:89–94.

Pagel M, Venditti C, Meade A. 2006. Large Punctuational Contribution of Speciation to Evolutionary Divergence at the Molecular Level. *Science* 314:119–121.

Philippe H, Casane D, Gribaldo S, Lopez P, Meunier J. 2003. Heterotachy and functional shift in protein evolution. *IUBMB Life* 55:257–265.

Pittis AA, Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* 531:101–104.

Pulquério MJF, Nichols RA. 2007. Dates from the molecular clock: how wrong can we be? *Trends Ecol. Evol.* 22:180–184.

Rannala B, Yang Z. 2007. Inferring Speciation Times under an Episodic Molecular Clock. *Syst. Biol.* 56:453–466.

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47:W191–W198.

dos Reis M, Gunnell GF, Barba-Montoya J, Wilkins A, Yang Z, Yoder AD. 2018. Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and Calibration Strategies on Divergence Time Estimation: Primates as a Test Case. *Syst. Biol.* 67:594–615.

dos Reis M, Yang Z. 2011. Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times. *Mol. Biol. Evol.* 28:2161–2172.

dos Reis M, Yang Z. 2013. The unbearable uncertainty of Bayesian divergence time estimation. *J. Syst. Evol.* 51:30–43.

Revell LJ, Harmon LJ, Glor RE. 2005. Under-parameterized Model of Sequence Evolution Leads to Bias in the Estimation of Diversification Rates from Molecular Phylogenies.Linder P, editor. *Syst. Biol.* 54:973–983.

Sarich VM, Wilson AC. 1966. Quantitative Immunochemistry and the Evolution of Primate Albumins: Micro-Complement Fixation. *Science* 154:1563–1566.

Schenk JJ. 2016. Consequences of Secondary Calibrations on Divergence Time Estimates. *PloS One* 11:e0148228.

Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era.Escriva H, editor. *PLOS ONE* 13:e0197433.

Susko E, Steel M, Roger AJ. 2021. Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events. *J. Theor. Biol.* 526:110788.

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135:599–607.

Tao Q, Tamura K, Mello B, Kumar S. 2019. Reliable Confidence Intervals for RelTime Estimates of Evolutionary Divergence Times. *Mol. Biol. Evol.* [Internet]. Available from: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msz236/5602325

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15:1647–1657.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58:267–288.

To T-H, Jung M, Lycett S, Gascuel O. 2016. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst. Biol.* 65:82–97.

Volz EM, Frost SDW. 2017. Scalable relaxed clock phylogenetic dating. *Virus Evol.* [Internet] 3. Available from: http://academic.oup.com/ve/article/doi/10.1093/vex025/4100592/Scalable-relaxed-clock-phylogenetic-dating

Vosseberg J, van Hooff JJE, Marcet-Houben M, van Vlimmeren A, van Wijk LM, Gabaldón T, Snel B. 2021. Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat. Ecol. Evol.* 5:92–100.

Warnock RCM, Yang Z, Donoghue PCJ. 2017. Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution. *Proc. R. Soc. B Biol. Sci.* 284:20170227.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci.* 82:1741–1745.

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

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res.* 46:D754–D761.

Zhu T, Dos Reis M, Yang Z. 2015. Characterization of the Uncertainty of Divergence Time Estimation under Relaxed Molecular Clock Models Using Multiple Loci. *Syst. Biol.* 64:267–280.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution, and genetic heterogeneity. *Horiz. Biochem.*:189–225.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357–366.

610