Article (Methods)

Factors influencing the accuracy in dating single gene trees

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

1. École Normale Supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'École Normale Supérieure (IBENS), F-75005 Paris, France

Corresponding author: H. Roest Crollius, hrc@bio.ens.psl.eu

Abstract (250 words max)

Molecular dating is a cornerstone of evolutionary biology, yet it is by far not an exact science. The inference of precise dates using gene sequences is difficult, in part because of the stochastic process of DNA mutation, selective forces that alter substitution rates and many unknown parameters linked to population genetics in ancestral lineages. Dating species divergence is one important challenge in this field, which is usually performed by concatenating extant sequences sampled within a genome as representative of a lineage, and computing distances between these lineages. However, concatenates precludes the dating of events specific to a gene family, such as gene duplication. During evolutionary time, individual gene sequences record different signatures of base substitutions and at rates that may deviate

substantially from the average rate. No formal study exists that quantifies which parameters influence this deviation. Here we designed a strategy to date events within a gene family, and we test the influence of more than 30 parameters on dating accuracy. We developed this approach on approximately 5,000 primate gene families comprising 12 genomes that display no gene loss nor gene duplications. We then test its relevance in the complete set of primate gene families to date gene duplications. Our result are compared to previous fossil and molecular dating approaches, and provide a practical set of guidelines for accurate molecular dating at the single gene family level.

Introduction

Since Zuckerkandl and Pauling (1962; 1965), Margoliash (1963), Doolittle and Blombäck (1964) formulated that biological sequences contain a signature of passed substitutions, the field of molecular phylogenetics has made great progress in reconstructing evolutionary history. It has greatly benefited from recent advances in sequencing genetic material from an increasing diversity of organisms, and many computational method have been developed to date divergences between species or between sequences. One major prior requirement to do so is to ascertain a process of substitution accumulation, i.e., a molecular clock. First envisioned as roughly constant, the exceptions became the rule and most dating softwares nowadays implement clock rate variation. Indeed it has been observed that rates of substitution per million year vary across taxa (Wu and Li 1985; Britten 1986; Pagel et al. 2006). The mechanisms at the origin of the variations are varied and may include population size, impact of selection at the molecular level, generation time, efficiency of the DNA repair

system, metabolism, etc. For some of these reasons, rates also vary across the genome (Wolfe et al. 1989). However it is generally agreed that by selecting multiple genes, these variations are attenuated, their mean rate is approximately constant and the molecular clock can still be used (Kumar and Hedges 1998). Dating events that took place in the history of species that share a relatively recent common ancestor also make this simplification more accurate.

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) were developed to assess these rate variations. Nowadays, most models accommodate a non constant molecular clock providing that calibrations are specified. How to relax the clock constancy is however subject to debate: autocorrelated rates assume a progressive change between connected branches while uncorrelated clock rates just follow an overall distribution such as the Gamma distribution. Different empirical studies support either the uncorrelated rates model (Drummond et al. 2006; Rannala and Yang 2007; Heath et al. 2012) or the autocorrelated clock rates (Thorne et al. 1998; 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). Algorithmic implementations include widely used softwares 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).

Molecular dating is currently mostly used to date speciation times, i.e., the divergence between two lineages, in which case a large number (typically hundreds) of genes sampled from extant genomes descending from the speciation of interest are carefully selected and concatenated. In order to confidently apply molecular dating using fossil-calibrated

substitution models, it is necessary to characterize the uncertainty of the dating method. A growing body of literature directly addresses the precision of molecular dating (Pulquério and Nichols 2007; Burbrink and Pyron 2008; dos Reis and Yang 2013; dos Reis et al. 2015; Zhu et al. 2015; Kumar and Hedges 2016; Warnock et al. 2017; Tao et al. 2019), which remains a controversial topic where several limitations have been underscored, notably the uncertainty in calibration points themselves and the difficulty in characterizing the type of clock relaxation (dos Reis et al. 2015). However molecular dating has naturally expanded to other subjects such as virus lineages at the epidemiological timescale (To et al. 2016; Volz and Frost 2017; Bletsa et al. 2019), transposable elements (Boissinot et al. 2000; Ovchinnikov et al. 2002; Khan et al. 2006), micro-RNAs (Sempere et al. 2006), and might be of interest for cancer phylogenetics (Lecca et al. 2015; Schwartz and Schäffer 2017). As has repeatedly been noted, the amount of sequence data and observable substitutions is crucial for accurate dating (Bromham et al. 2000; Lanfear et al. 2010; Smith et al. 2018; Bletsa et al. 2019; Bromham 2019) which 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 on 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 which creates a disproportionate amount of noise relative to the weak signal. Here we analyse the impact of these parameters on thousands of single protein gene families in primates. We first develop a procedure for estimating the dating confidence after benchmarking several strategies for dating speciation nodes. We take advantage of the fact that thousands of gene trees contain the same speciation nodes to measure the dispersion of these estimates as a measure of reliability of the dating method. In a subsequent analysis,

we regress the individual gene tree errors against gene tree characteristics, to identify the correlates of a greater dating accuracy.

Results

Dating procedures

We tested seven dating procedures, which all follow the same general steps (methods): first compute and process the coding sequence alignment and second, date the nodes by fitting a molecular clock.

For the first step, we tested four strategies to compute and process the alignment: (i) *original*, where the alignment is from Ensembl Compara; (ii) *cleaned*, where the Ensembl Compara alignment is filtered to remove poorly aligned sequence segments with the HmmCleaner program (Di Franco et al. 2019); (iii) *FSA*, where the sequences were aligned using the FSA program, a probabilistic and rather conservative method (Bradley et al. 2009); (iv) *FSA+cleaned*, where the *cleaning* is applied on the FSA alignment.

For the second step, we either applied the Mean-Path-Length (MPL) algorithm (Britton et al. 2002; 2007) or a relaxed clock model with Beast 2 (Bouckaert et al. 2019). MPL first requires precomputed synonymous branch lengths with Codeml from the PAML 4 package (Yang 2007), which we generated by testing two alternative models: 1) the free-ratio branch model (here called branch model), or 2) a fixed branch rate but a variable site rate (gamma distributed, here called site model). MPL then converts these branch lengths into absolute node dates using a non parametric smoothing process: the path lengths from a node to all its descendent leaves is averaged to deliver an ultrametric tree. This average distance to the

present is then rescaled in absolute age using point calibrations. In contrast, Beast is entirely parametric and takes into account tree topology uncertainty, calibration uncertainty, and a model of rate variation.

The first and second step of the dating procedure were tested in seven combinations as follows: (1) *original+siteMPL*, (2) *original+branchMPL*, (3) *cleaned+branchMPL*, (4) *FSA+branchMPL, (5) FSA+cleaned+branchMPL*, (6) *FSA+Beast*, (7) *FSA+cleaned+Beast*.

As input, Ensembl Compara gene trees and coding sequences were pruned to only retain *Simiiformes* sequences and two outgroup sequences. *Simiiformes* is the primate group sister to *Tarsiidae* and is further divided into *Catarrhini* (Old World monkeys and apes) and *Platyrrhini* (New World monkeys) (fig. 1). To ensure that all trees used for this benchmark display the same topology, we eliminated trees with gene duplications and/or gene losses, resulting in 5,204 gene trees. As calibration, we used *Simiiformes*, the common ancestral node dated at 43.15 Mya in TimeTree (Kumar et al. 2017) to jointly estimate the ages of the twelve nodes using each of the above seven procedures (supplementary material S1, S2).

Choice of the best dating procedure

We computed four summary statistics for each combination of pipeline and speciation in order to characterise the dispersion of age estimates: the mean absolute deviation from the median (MAD; fig 1), the interval that contains 95% of the dates for a given speciation (supplementary material S3 and S4), the skew (supplementary material S5) and kurtosis (supplementary material S6) of each distribution. The mean absolute deviation from the median (MAD) is a good compromise between sensitivity (it accounts for all data points) and robustness to outliers (by using deviation from the median). We therefore propose to use the MAD to estimate parameters influencing dating error. In the following, for the sake of

simplicity we use the word “error” in place of MAD or of deviation.

In order to rank the accuracy of the seven methods, we averaged errors over all ancestral nodes (fig. 1). This shows that the *FSA+cleaned+branchMPL* approach performs best, with an average error of 2.7 My. Consistently, this method generates the smallest error for 5 out of 12 ancestral nodes, and overall it generates lower MAD values for all ancestors relative to other methods. The *FSA+cleaned+Beast* method also generates minimal errors for 5 ancestors, but interestingly can also generate very high values for others, such as for *Catarrhini* and *Hominoidea*. The only difference between these two approaches is the dating method, suggesting that Beast performs less consistently across varying conditions than the MPL method. Overall, the other statistics confirmed these results, with the 95% interval and the distribution skew also ranking the *FSA+cleaned+branchMPL* first, while it comes second with the distribution of kurtosis statistic. Overall, while more sophisticated than MPL, the Beast method results in similar error ranges. The skew and kurtosis values however show dramatic differences between the *FSA+Beast* and other approaches (supplementary material S5 and S6). This suggests that a subset of trees are not handled well with this approach, generating dates that are extremely far from the median. Of note, to avoid negative branch lengths in the MPL procedure, if a computed date exceeds the range imposed by the two flanking nodes of the node of interest, then we set the date to be equal to the nearest flanking node. In Beast this is not required because branch lengths are designed to be positive, but highly heterogeneous trees may generate the extreme skew and kurtosis values observed. The variability of Beast performances and its higher skew and kurtosis statistics could be caused by a higher sensitivity to reduced numbers of available sites in the alignments, especially in the more extreme *FSA+cleaned* version. This dispersion of the estimates may also reflect the uncertainty of the input more faithfully than the MPL smoothing procedure.

The comparison of MAD values across the seven dating strategies for the 12 ancestral nodes (fig. 1) also reveals the striking influence of the alignment method. For example, the MPL dating algorithm was used with several combinations of alignment procedures, showing that the *original* alignments consistently result in a higher error, while the error is equivalently low for the *cleaned*, *FSA* and *FSA+cleaned* procedures. Interestingly, among the 12 ancestral nodes, *Catarrhini* and to a lesser extent *Hominoidea* depart from this trend. *Catarrhini* appears much less sensitive to the alignment stringency than other ancestors, but much more to the dating model, with maximum error values with the Beast method.

The benchmark presented here also confronted two substitution models within the MPL method, showing that given the same input alignments (*original*) the branch model (1) consistently performs better than the site model (2).

Among the possible dating algorithms, we also tested the "chronos" function from the R package "APE" (Paradis 2013) but optimisation issues made interpretation of results too ambiguous for their application in this study (supplementary material S7, S8).

In conclusion, from the error values (fig. 1), we decided that the *FSA+cleaned+branchMPL* procedure was the most accurate overall and use it for the rest of our analyses. It also advantageously allows to easily set more calibration nodes, in order to estimate a single speciation age one at a time, by calibrating all other nodes in the tree.

Magnitude of the dispersion

The 95% intervals provide an overview of the spread of the distributions. We were surprised to note that, even in the pipelines with most frequent minimal intervals (procedures 5 and 7, supplementary material S3), the intervals overlap substantially between ancestors

(supplementary material S4). Most critically, they represent an important fraction of the total age interval between the present and the *Simiiformes* calibration point at 43 Mya. For example, *Hominoidea* is dated close to the midpoint between the present and the calibration point (20 Mya), and the smallest 95% interval, 17 My, represents 40% of this range (procedure 5 in supplementary material S3).

In addition to the above measures of dispersion (MAD and quantile ranges), we computed the deviation from two sets of reference dates to effectively measure an error. The first comes from TimeTree (Kumar et al. 2017), a database of published molecular time estimates providing consensus ages, while a second source of reference dates comes from a fully parametric estimation (dos Reis et al. 2018) (see supplementary material S4).

Comparisons with actual reference ages show a clear trend: our median ages are consistently younger than both our reference sources, especially compared to the parametric estimation (supplementary material S4). More strikingly the 95% confidence intervals of the references do not always contain our own median age. Notably, we observe this trend with both the MPL and the Beast methods, despite the latter relying on a relaxed clock model with continuous calibration distributions. Although this shift is disconcerting, it appears to be consistent across speciations. We note that the two reference sources, TimeTree and the parametric estimates, are also discordant with a majority of disjoint 95% confidence intervals. The systematic tendency towards more recent ages could be due to multiple factors, such as a too simplistic model in our procedure that does not account for complex single genes processes including different species-wise rates or hybridizations. Of note, incomplete lineage sorting would increase the estimated ages and cannot explain this tendency. A second possible cause may be that our TimeTree calibration for Simiiforms at 43 Mya is too recent.

For example, if the calibration was was set 20% further in the past at 52 My, then all 12 computed MAD estimates would fit in the 95% interval of TimeTree dates for these nodes. However, a deficit in site and substitutions seems the most likely explanation, because the age underestimation is stronger when alignments are subject to the HMMcleaner pre-treatment which removes sites from the alignments (e.g. *FSA+branchMPL* compared to *FSA+cleaned+branchMPL*). Although such removed sites should only be those misaligned, the program must also be sometimes pruning substitutions that are specific to few species.

Because of this lack of consensus and to further dissect the factors influencing the precision of date estimates, we used our own median ages as reference values, with the added advantage of removing errors due to lineage specific rates and only considering gene-to-gene variability.

Desirable gene tree features for accurate dating

We next sought to determine the characteristics of gene families and trees that correlate with the accuracy of the dating estimates provided by the *FSA+cleaned+branchMPL* pipeline. To estimate the error per tree, we averaged the deviation from the median age over tree nodes. We regressed it against more than 60 features related to alignment quality and substitution rate estimations, as well as measures generated by the programs in the pipeline, such as the likelihood of the gene tree and the proportion of nucleotides removed by HmmCleaner. With a Lasso regression against the gene tree deviation, we selected features with an absolute coefficient above 0.01 and refitted the regression by Ordinary Least Squares (OLS).

In a first analysis on the same 5,204 trees as above, we noticed that a few parameters show a dominant effect (high regression coefficients) on the dating error. Upon closer examination, these parameters all described anomalous or obviously aberrant gene family characteristics, such as consecutive null branch lengths (in dS or in dN) or non-overlapping sequences in the

same alignment (supplementary material S9). We removed 3,768 trees with such anomalous characteristics, resulting in a final set of 1,436 trees for a regression aimed at detecting more biologically relevant influences on the dating error.

From the magnitude of coefficients in the multiple regression (fig. 2), the main feature that increases the error is the heterogeneity of dS rate per My (p-value = 0), followed by the alignment length (p-value = 0). This can be explained by the loss of statistical power that coincides with the use of individual coding sequences. Indeed, the decile with lowest accuracy is associated with highest dS heterogeneity and shortest sequences (table 1). The GC-content of the alignment also plays a significant role in increasing the accuracy (p-value = 0.003) but within a modest GC% range. The adjusted R^2 of the OLS fit is 0.47.

	Error (My)	dS rate heterogeneity (10^{-3} subst/codon/My)	Alignment length (nucleotides)	Mean GC content (%)
10% predicted lowest accuracy	3.74 ± 1.42	1.7 ±1.1	1289 ± 702	52.0 ± 8.5
10% predicted highest accuracy	1.28 ± 0.39	0.46 ± 0.26	7031 ± 3866	55.7 ± 7.4

Table 1: Characteristics of extreme gene families according to the fitted predicted error

Comparing features between ancestors

After having established the main factors that influence the average dating accuracy over the complete tree, we must consider that the node positions in a tree correspond to very different situations. Dating errors may therefore be due to different factors depending on the specific nodes. An obvious difference between nodes is their reference age, such as between *Catarrhini* (ancestor of Old World monkeys and apes), estimated at 29.4 ± 1.9 Mya in TimeTree, and the much more recent *Pan* and *Macaca* genera. Moreover our species

phylogeny contains sister clades with variable numbers of descendants where the two subtrees can be very balanced, as for the descendants of *Catarrhini*, or very imbalanced as for the subtrees descending from *Hominoidea* (fig. 1). In particular, subtrees with a high density of speciation node suffer from a 'node density artefact', where more substitutions are inferred because more branches are available (Hugall and Lee 2007). Dating nodes separately also makes it possible to measure deviations for each node independently from the factors affecting other parts of the tree (i.e., beyond the two nearest calibration points). We executed the dating pipeline *FSA+cleaned+branchMPL* for each speciation, with the calibration scheme adapted such that all nodes except the node of interest are set to the reference age from TimeTree. We applied the same regression procedure as above, except for the dS rate and its heterogeneity, which were measured separately for the branches directly connected to the node of interest, and for the branches located elsewhere in the tree. We termed the two measures "local" and "non-local" heterogeneity respectively. Of note, local and non-local dS rates are correlated, which would confuse the interpretation of their relative influence on dating errors. To avoid this, we considered the non-local dS rates on the one hand, and the part of the local dS rates that is not explained by the non-local rates on the other hand (relative local dS rate, see methods).

The twelve regressions produced slightly different results in terms of which features had the most impact on the dating error, so we selected the top five features of each regression for comparison (fig. 3, supplementary material S10). In contrast to the regression on the whole tree, results for individual nodes show a major role for non-local dS rates on improving the dating accuracy, and to a lesser extent for the relative local dS rates. Our results are consistent with the expectation that a large dS rate coincides with more statistical power, thus decreasing the error, but it is somewhat unexpected that this effect was not observed

when errors were computed on the whole tree (fig. 2). On the other hand, as for the regression on the whole tree, the dS rate heterogeneity has a major influence on dating errors, especially for the local dS rate heterogeneity, underlining that the constancy of the molecular clock (low dS rate heterogeneity) is a key factor to gain higher precision. Finally, in line with the global regression, the alignment length consistently decreases the dating error of individual speciation nodes.

We sorted the speciation nodes by age, with *Catarrhini* the oldest and *Pan* the most recent, but interestingly there is no consistent sign of an age effect on the error. This is somewhat unexpected because older nodes rely on more distant sequence comparisons, and therefore may suffer from more noise in the alignment. The speciation-specific regressions (fig. 3) show a lower R^2 (ranging from 0.103 for *Papionini* to 0.317 for *Pan*, supplementary material S11*)* compared to the regression computed on the entire tree (fig 2, R^2=0.47), most likely because there is less signal in the local error to be explained by averaged gene tree feature.

Predicting average errors in new trees

As a practical application, it may be of interest to exploit the relationships between tree features and dating errors established above, in order to predict the errors one would we be subject to when dating nodes in a new (unknown) tree. We applied this reasoning to the 19,331 primate trees that contain duplication nodes and/or gene losses (testing set) using regression coefficients computed on the 5,204 gene trees that have no duplication and no loss (training set).

Results show that two thirds (66%) of the trees in the testing set display predicted errors that overlap 90% of the training set error distribution (fig. 4a, light green vs light blue). The remaining fraction (34%) of the trees in the testing set display a much larger predicted error,

up to 20 My different from the reference date. This prediction of a higher error is caused by this subset of the tested gene trees having a high dS rate heterogeneity (fig. 4b), and smaller alignments (fig. 4c), both of which exert the strongest influence on the dating error (supplementary material S12).

In summary, among the 14,155 new trees used in the prediction, 4,796 are predicted to have an error higher than that of most training trees (3.22 My threshold, corresponding to the 10th decile of the training set; vertical dotted line in fig. 4a). For practical purposes, we can consider that this 10th decile corresponds to trees with intolerable errors on date estimates. Using this guideline, we observe that trees of the testing set with dS rate heterogeneities larger than 2.23×10^{-3} substitutions/codon/My (fig. 4c) or with alignment length longer than 822 bp (fig. 4b) capture more than 50% of trees with such intolerable errors. As a rough guideline, we propose that these thresholds be used to reject trees unsuitable for dating evolutionary events in single gene trees.

Discussion

Molecular dating is a preliminary step for many evolutionary analyses, such as correlations with paleo environments or climates, comparisons of different clade evolutionary histories, analyses of diversification dynamics, etc. However, the power of dating methods depends on the amount of data available (Bromham et al. 2000; Lanfear et al. 2010; Smith et al. 2018; Bletsa et al. 2019; Bromham 2019) and this is the rationale for concatenating alignments when estimating speciation ages. Nonetheless individual gene tree histories are also of interest for evolutionists, in particular with regards to gene duplications. For instance Aguileta et al. (2006) revisited the textbook example of the globin family and re-estimated rates and divergence times of the paralogs. Here, we provide a detailed investigation of molecular dating behaviour when applied to single gene families, where the number of sites

is therefore much more limited than typically found in concatenations of gene sequences. Aiming towards such gene-scale analyses, we therefore estimated empirically the confidence of single genes dating, using primates families of strictly orthologous genes. We show the range of confidence displayed in individual gene trees. Our results are empirical, that is, we used a genomic database of gene trees to estimate a sample variation in dating errors. As such, our approach is distinct from and complementary to statistical confidence interval estimation, such as bootstrapping.

Magnitude and causes of the average age deviations

First, our empirical 95% intervals of age has values as high as 18.2 My at *Catarrhini*, representing a large time span compared to the 43 Mya of the *Simiiformes* calibration. Although this span is lower for more recent nodes, it may seem that such large uncertainty intervals are prohibitive. However, when considering the predicted dating error averaged across the 12 speciation nodes of primate trees, 90% are less 3.22 My per node. Depending on the question being addressed by the dating approach, and depending on the age of the target nodes, this suggest that errors made when dating from single gene families may in some contexts be negligible for the interpretation of results.

The 5,235 selected gene trees represent a biased fraction of the approximately 22,000 gene trees rooted at *Simiiformes.* They likely representing a better subset for dating, because the absence of gene duplication and loss should be associated with a more stable evolution in terms of constancy of rate and selective pressure. Within this curated fraction, we detected variation in characteristics influencing the dating accuracy: as expected, more statistical power, provided by larger dS rates and longer alignments lower the dating error while an increase in rate variation between local branches increases the error. This confirms previous observations that the number of sites and substitutions is a limiting factor with regard to

statistical power, and also that the clock rate does not appear to be constant across branches, even after accounting for alignment length. Furthermore, the clock rate is not constant across genes of the same genome either. We find that this rate variation affects synonymous substitutions, raising the question of the causal mechanisms: are synonymous substitutions neutral enough for dating? How large is the impact of the stochasticity of the substitution process, and how much is it amplified by uncertainty in inference? This rate unreliability is consistent with current knowledge in the field, and pragmatic approaches have been proposed to solve this heterogeneity, such as the "gene-shopping" approach from Smith et al. (2018), wherein trees with apparently non constant rate are discarded prior to the analysis. Our regression of the dating errors allowed us to pinpoint those non-desirable characteristics, and could be further used to find possible biological reasons behind them.

In our pipeline selection, we promptly discarded the *siteMPL* strategy, clearly less accurate than the *branchMPL*. This finding contrasts with observations that fitting clock rates site-wise is better than branch-wise for dating (Smith et al. 2018). Fitting both models simultaneously should naturally lead to improved results and could be done using codon model implementations such as in HyPhy (Kosakovsky Pond et al. 2019).

Our analysis was a first attempt at quantifying the most prevalent factors influencing the dating accuracy of individual gene trees. In the course of this study, a number of insights suggest which recent developments could lead to improvements:

- using a fully probabilistic gene-species tree reconciliation such as implemented in GeneRax (Morel et al. 2020);
- for recent timescales, considering incomplete lineage sorting (Scally et al. 2012) and individual gene histories more generally: speciation at the genomic scale is thought to happen through "islands of differentiation", i.e., some genomic regions would differentiate early,

while other remain under gene flow (Nosil et al. 2009, Wolf and Ellegren 2017 for a review);
- using models of substitution rate variation, both branch-wise and site-wise, and treating neutral sites adequately. Here we use a codon model to account for dS and dN but it might be a sufficient approximation to distinguish the 3rd codon position. Accurate nucleotide models would be especially useful because codon models require more sites, a limiting factor here. Regarding this site neutrality question, the distribution of fitness effects of non-synonymous substitution remains under active research (e.g. Moutinho et al. 2019, Rousselle et al. 2019), and holds potential implications for the molecular clock;
- using a mixture of relaxed clock models as in Lartillot et al. (2016) or Bletsa et al. (2019);
- considering the uncertainty in independent calibrations, ideally from fossil dates using a speciation birth-death prior as described in Yang and Rannala (2006) and dos Reis (2016), and implemented in the MCMCTree program (Yang 2007). Indeed, they and others have observed that the uncertainty in fossil calibrations cannot be avoided and propagates in the estimates proportionally with age.

Elaborate inference cannot circumvent data limitations

On all regressions analyses, adjusted R^2 were relatively low (0.47 for the global regression, down to 0.103 for the *Papionini* regression), indicating a low explained variance. To investigate the missing variance, we should incorporate additional biological characteristics, but it is likely that intrinsic stochastic properties of the mutational and evolutionary processes account for most of this uncertainty.

These limitations inherent to the data are not solved even by an advanced method such as Beast and its statistically elaborate procedure. Indeed, we did not manage to obtain significantly less dispersed estimates, although we do advocate the use of such programs for the many benefits it brings such as obtaining credibility intervals from the posterior

distribution. Beast produced clearly more recent ages, a trend that is difficult to evaluate given our uncertainty in the reference ages. Of note, it appears computationally intractable to date individual nodes with Beast when setting too many calibration points perhaps because of prior distribution truncation.

Despite the expectation that a more complete model would improve accuracy, which was not satisfied by our results with Beast, it should nonetheless be highlighted that the most important advantage of Bayesian methods is the ability to properly render the uncertainty caused by insufficient data. Strategies for doing so are straightforward in the Bayesian framework where one commonly fits the posterior probability without data. One extra requirement is to rigorously transmit this uncertainty between analysis steps. Achieving this is more easily done when the inference proceeds through all steps at once, such as jointly estimating tree topology, substitutions and clock rates, given that model parameters remain identifiable. In our experience it appears challenging to run such computations at the scale of comparative genomics. First, MCMC implies that long running times and complex models might be numerically intractable and second, MCMC requires a thorough and experimented human inspection to validate the inference output. In the face of this, a database like Ensembl Compara Vertebrata contains 197 species and 51,267 gene families as of release 98, and is expected to significantly grow in the near future. Therefore future work on faster non-parametric algorithms should be a promising path forward.

Our study leads to practical guidelines for accurate and confident phylogenetic molecular dating of single gene trees. They 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 correlations with other gene trees, as recent developments suggest (Duchêne et al. 2019).

Our results should help identify pitfalls in dating analysis, at different steps. As proof of concept, we predicted expected errors for trees containing duplications and losses, a benchmark that provides easily comparable values of uncertainty. For example, a new dating algorithm could be tested on the same data and compared with our present results. Alternatively, it would be interesting to test different genome wide sets of gene families sampled from taxa bearing different ranges of characteristics concerning the alignment and substitution rates.

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 is still mysterious, both at the lineage scale and at the genome scale. Future developments answering this question will be methodologically challenging but shall shed light on important evolutionary processes.

Material and Methods

Source and number of trees

Species phylogenetic tree

The tree of the 99 species that are in Ensembl Compara 93 (July 2018; Zerbino et al. 2018) and based on the NCBI taxonomy was edited to resolve polytomies using multiple sources, including a dated phylogeny from TimeTree (data retrieved Jan. 2019; Kumar et al. 2017) and reference literature (Perelman et al. 2011; Mittermeier et al. 2013; Upham et al. 2019).

We used reference ages from both TimeTree and the estimation of (dos Reis et al. 2018), calibration strategy A, auto-correlated clock, where their *Anthropoidea* node is synonym of

Simiiformes.

Gene trees and alignments

The complete forest of 23,904 reconciled phylogenetic gene trees (NHX format) and the corresponding multiple sequence alignments of the coding domain of the longest transcribed isoforms (fasta format) were downloaded from Ensembl Compara 93 (July 2018; Zerbino et al. 2018) via the Perl API.

Data processing

We first removed species with low genome assembly quality, as quantified by the N50 (size of scaffold, in number of genes, such that that 50% of genes are contained in larger scaffolds) and the K70 (number of the longest scaffolds which contain more than 70% of the genes). Species with N50<30 genes and K70>400 scaffolds were removed, including 3 primate species out of 24 (*Rhinopithecus bieti*, *Rhinopithecus roxellana*, *Carlito syrichta)* leaving 21 primate species for this study including 18 *Simiiformes*.

Our species tree topology was then applied on the downloaded gene trees from Ensembl, requiring that some nodes be edited (automatically; Peres and Roest Crollius 2015). In the original gene tree forest from Ensembl, 729 subtrees supported by aberrant branch lengths (above 10,000 nucleotide substitutions per site) were removed, and 6962 annotated gene splits were merged. From the original alignments and trees, which might be rooted up to the *Opisthokonta* ancestor of fungi and metazoans, subtrees rooted at *Simiiformes* were extracted, while retaining at least two nearest outgroup sequences.

In the first control analysis, we only used gene trees without duplications and losses within *Simiiformes*, so that all tree topologies are exactly the same, for a total of 5,235 trees.

Dating pipeline

Realigning and cleaning

In the *FSA* and *FSA+cleaned* procedures (numbered 4 to 7), the *Simiiformes* protein sequences were realigned using FSA (Fast Statistical Alignment, Bradley et al. 2009) with default parameters, then back-translated to the corresponding nucleotide alignment.

HmmCleaner (Di Franco et al. 2019) finds segments of sequence that appear inconsistent with the other aligned sequences, interpreting them as sequencing errors. In the *cleaned* procedures, we replace these segments by gaps. It was applied on the 5,235 amino-acid alignments that do not have duplications or losses. Out of these, 31 alignments caused the program to fail, resulting in the 5,204 subtrees under study here.

Computing synonymous substitution rates

Most of our dating procedures make use of branch lengths measured in synonymous substitution numbers (dS), computed with codeml from the PAML software (Yang 2007). A typical parameter file is provided in supplementary material S13, but most importantly we fit:

1) A “free-ratio” branch model (parameter "model=1"), whereby the ω parameter (instantaneous dN/dS ratio) is independent across branches. In this model, a single ω is fitted for all sites in the alignment;

2) A discretised Gamma distributed site model with 10 rate categories (parameters "model=0", "NSsites=5", "ncatG=10") on ω, in which case each site category of ω is constant across branches.

Sites with missing data are treated as ambiguous (parameter "cleandata=0").

Mean-Path-Length algorithm to obtain ages from substitution numbers

We developed a custom Python implementation of the algorithm described in Britton et al (2002; 2007). Briefly, the MPL computes the depth of a node by averaging the path lengths (sum of all successive branches leading to the present). We obtained the node age in million years by multiplying the mean path length with the ratio of the root calibration by the root mean-path-length. As a result, the rescaled tree is ultrametric, i.e., all leaves are at the same distance from the root. Internal node calibrations can be set as well, in which case the MPL formula needs to be adjusted, and we implemented it differently than Britton et al. (2007) (supplementary material S14). Hard calibrations are used, i.e., they are point values and determine strict boundaries. When the mean-path-length ratio results in negative branch lengths, we adjusted the node position in order to convert it to a null branch length.

Penalised-likelihood algorithm for dating

We used the chronos function from the R package “APE” 5.3 (Paradis 2013) with varying parametrisations:

- a constant rate clock fitted by maximum likelihood (“model=discrete, nb.rate.cat=1”);
- correlated clock fitted by penalised likelihood (“model="correlated"”, default) with various penalty weights (lambda=1, 100 and 10000);
- a relaxed clock fitted by penalised likelihood (“model="relaxed"”) with the same lambdas;
- a Mean-Path-Length fit (function “chronoMPL”) was also performed.

The actual function used was modified to allow all possible calibration schemes during initialisation. Optimisation issues that occurred on our data are described in supplementary material S8.

Bayesian node dating with Beast 2

Using the alignments as input for Beast 2.6.1 (Bouckaert et al. 2019) we ran the following model: a GTR+Γ nucleotide substitution model with two site partitions corresponding to the codon positions {1,2} and 3. We used a Yule Birth-Death model for the branching pattern, and a relaxed uncorrelated lognormal clock rate was fitted on branches, with estimated mean and standard deviation. Calibrations for *Primates* and *Simiiformes* were taken from TimeTree and specified as gamma distributions manually tuned to match TimeTree point estimate and 95% interval: location 70.8 My, α=4.6 and β = 0.656 for *Primates*, and location 40.9 My, α=4.6 and β = 0.656 for *Simiiformes*. The MCMC was run 50,000,000 iterations in order for all ESS (effective sample size) to be above 200. The resulting dates were annotated on the imposed input tree topology from Ensembl using TreeAnnotator from Beast 2.

Quantifying dispersion

The mean absolute deviation from the median (MAD) is a robust estimate of deviation: 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. As such it appeared to us as a good compromise for evaluating the dating procedures. Additionally it corresponds to the metric used in the Brown-Forsythe test of variance equality, a convenient property when testing the significance of the dispersion difference between procedures.

Regression

All statistical analyses on the output dates were performed in a Jupyter notebook environment using Python 3.5 and additional scientific computing packages including Numpy, Pandas, Statsmodels, Scikit-learn, Biopython and Ete3 (supplementary material

S16).

Collecting features of the gene families

We measured about 60 characteristics of the gene families as explanatory variables in the regression. They include:

- Tree features:

  - number of edited nodes when imposing the species tree;

  - minimum and mean bootstrap values of nodes, from Ensembl;

- 22 alignment features:

  - global statistics such as alignment length and proportion of split sequences;

  - 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.

- substitution rate features:

  - proportion of synonymous sites in the alignment;

  - the ratio of transitions over transversions (kappa);

  - the tree length in dS, dN and t (total codon rate);

  - median and standard deviation of ω (dN/dS);

- features from program executions:

  - log-likelihood, computation time and convergence warnings from codeml;

  - proportion of sequences cleaned by HmmCleaner

Computing mean rate and rate heterogeneity per tree

The mean rate is computed as the ratio of the total branch length in substitutions per site (d)

divided by the total branch length in million years (l), corresponding to the maximum-likelihood rate estimate from a Poisson model:

$$\bar{r}_{tree} = \frac{\sum_{branch} d_{branch}}{\sum_{branch} l_{branch}} \quad (1)$$

The rate heterogeneity was taken as the weighted deviation of the branch rate, where each branch is weighted by its length in million year. This weighting was motivated by the rationale that the longest branches are expected to give more stable rate estimates and by a slightly better R^2 in our regressions.

$$\bar{\sigma}^2_{tree}(r) = \frac{1}{\sum_{branch} l_{branch}} \times \sum_{branch} l_{branch} \left(\frac{d_{branch}}{l_{branch}} - \bar{r}\right)^2 \quad (2)$$

Transforming 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 approximately normally distributed explanatory variables. For this, a semi-automated procedure was set up, where a set a transform functions were selected if they minimize the skew. Possible transforms were 'none', 'logarithm base 10', and 'square root', with additional increment or sign modification if needed. All discretisations (in all case binary-encoded) were manually enforced, in cases where a large proportion of values were constant, or where distributions were clearly bimodal. Transforms for each variable can be found in supplementary material S15.

Reducing multicollinearity

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

- first we obtained the covariance 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. We thus decorrelated many instances of metrics where the standard deviation is correlated with the mean (supplementary material S15).
- 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). We iteratively removed features with the highest multicollinearity impact on the matrix, until a multicollinearity condition number less than 20 was reached.

Removing trees with outlier binary values

Twelve binary features showed up with high coefficients in the regressions. These include the presence of consecutive branch of length zero, and the presence of split sequences, that we can consider as obvious drivers of errors, and therefore remove them to better analyse the remaining features. We individually tested their effect on our response variable (the age estimate deviation) using one-tailed t-tests and a Benjamini-Hochberg correction for multiple testing (supplementary material S9). We then removed the 3768 gene trees having such anomalies, in order to fit a regression on mostly continuous variables.

Final feature selection with Lasso and OLS refitting

Lasso (Tibshirani 1996) is a regression fitting algorithm where feature selection is performed simultaneously, 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. However Lasso coefficients are (by design) biased, 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. Statsmodels 0.10.1 implementations were used.

Predicting the dating accuracy on sets of trees

The same regression as above was run on the 5,204 gene trees (training set) that have no duplications or loss, producing coefficients and feature transforms that were then applied on a testing dataset composed of the other 19,331 *Simiiformes* gene trees.

To compute the prediction, all regression variables (i.e., gene tree features) were collected for the testing set, in a similar manner than for the training. However, duplication nodes do not have reference dates, so the heterogeneity of the dS rate must be approximated for the gene trees with duplications. This approximation must only be done between two consecutive speciations for which we have reference dates; the subtree inbetween these consecutive speciations, which contains one or more duplications, is replaced by an unresolved subtree, in a way that retains the same path lengths from the parent speciation node to the descendent speciation nodes. This allowed us to compute the dS rate and dS rate heterogeneity for the whole tree, with equations (1) and (2) respectively. As this reduces the heterogeneity around duplication nodes, the overall tree heterogeneity is underestimated, and so will the predicted dating error. Therefore the resulting predicted error must be interpreted as a lower boundary. To reduce multicollinearity between features, we applied the same decorrelation procedure as in the training regression.

Source code available at

https://github.com/DyogenIBENS/Phylorgs (version 0.0.0); the Python code from the Jupyter notebooks with source and intermediate data is archived at Zenodo.org (doi:10.5281/zenodo.3992663).

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

References

Aguileta G, Bielawski JP, Yang Z. 2006. Evolutionary rate variation among vertebrate β globin genes: Implications for dating gene family duplication events. *Gene* 380:21–29.

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

Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History. *Mol. Biol. Evol.* 17:915–928.

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.

Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K. 2007. Estimating Divergence Times in Large Phylogenetic Trees. Anderson F, editor. *Syst. Biol.* 56:741–752.

Britton T, Oxelman B, Vinnersten A, Bremer K. 2002. Phylogenetic dating with confidence

intervals using mean path lengths. *Mol. Phylogenet. Evol.* 24:58–65.

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. 2019. Linking Branch Lengths Across Sets of Loci Provides the Highest Statistical Support for Phylogenetic Inference. *Mol. Biol. Evol.* [Internet]. Available from: https://academic-oup-com.insb.bib.cnrs.fr/mbe/advance-article/doi/10.1093/molbev/msz291/5673393

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

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

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

Hugall AF, Lee MSY. 2007. The Likelihood Node Density Effect and Consequences for Evolutionary Studies of Molecular Rates. *Evolution* 61:2293–2307.

Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res.* 16:78–87.

Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, et al. 2019. HyPhy 2.5 - a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.*

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.

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

Lecca P, Casiraghi N, Demichelis F. 2015. Defining order and timing of mutations during cancer progression: the TO-DAG probabilistic graphical model. *Front. Genet.* 6:309.

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

Mittermeier RA, Rylands AB, Wilson DE eds. 2013. Handbook of the mammals of the world: 3. Primates. In: Handbook of the mammals of the world. Vol. 3. Lynx Ediciones.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss. *bioRxiv*:779066.

Moutinho AF, Bataillon T, Dutheil JY. 2019. Variation of the adaptive substitution rate between species and within genomes. *Evol. Ecol.* [Internet]. Available from: https://doi.org/10.1007/s10682-019-10026-z

Nosil P, Funk D J, Ortiz-Barrientos D. 2009. Divergent selection and heterogeneous genomic divergence. *Mol. Ecol.* 18:375–402.

Ovchinnikov I, Rubin A, Swergold GD. 2002. Tracing the LINEs of human evolution. *Proc. Natl. Acad. Sci.* 99:10522–10527.

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

Paradis E. 2013. Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Mol. Phylogenet. Evol.* 67:436–444.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A Molecular Phylogeny of Living Primates. *PLOS Genet.* 7:e1001342.

Peres A, Roest Crollius H. 2015. Improving duplicated nodes position in vertebrate gene trees. *BMC Bioinformatics* 16:A9.

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.

dos Reis M. 2016. Notes on the birth–death prior with fossil calibrations for Bayesian estimation of species divergence times. *Philos. Trans. R. Soc. B Biol. Sci.* 371:20150128.

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. 2013. The unbearable uncertainty of Bayesian divergence time estimation. *J. Syst. Evol.* 51:30–43.

Rousselle M, Simion P, Tilak MK, Figuet E, Nabholz B, Galtier N. 2019. Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *bioRxiv*:643619.

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

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169–175.

Schwartz R, Schäffer AA. 2017. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* 18:213–229.

Sempere LF, Cole CN, Mcpeek MA, Peterson KJ. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J. Exp. Zoolog. B Mol. Dev. Evol.* 306B:575–588.

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.

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.

Upham NS, Esselstyn JA, Jetz W. 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biol.* 17:e3000494.

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

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.

Wolf JBW, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. *Nat. Rev. Genet.* 18:87–100.

Wolfe KH, Sharp PM, Li W-H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337:283–285.

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.

Yang Z, Rannala B. 2006. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol. Biol. Evol.* 23:212–226.

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.

Figure legends

Figure 1. Mean absolute deviation from the median (MAD) of age estimates, for each method (columns) and each speciation (rows), in million years (color gradient scaled by row). Minimum values for each row are hashed in yellow.

Figure 2. Regression coefficients of the transformed normalized variables against the response variable "mean absolute age deviation". Coefficients are ranked by absolute value (except the intercept), indicating their relative importance in the age deviation. Error bars represent the 95% confidence interval. Ordinary Least Squares regression was fitted on the variables selected by Lasso, and absolute coefficient higher than 0.01 were retained (** p-values < 0.01).

Figure 3. Regression coefficients for eight features with the most impact on the dating error, for each ancestral node ranked by decreasing ages from left to right on the x-axis. Error bars represent the 95% confidence interval.

Figure 4. a) Predicted error for trees with duplications and losses (green, n=2774), based on the regression coefficients for the training trees (blue, n=1436). b) Distributions of the alignments length for the training and testing sets. The X axes is in centred and normalized units.

Figure 1

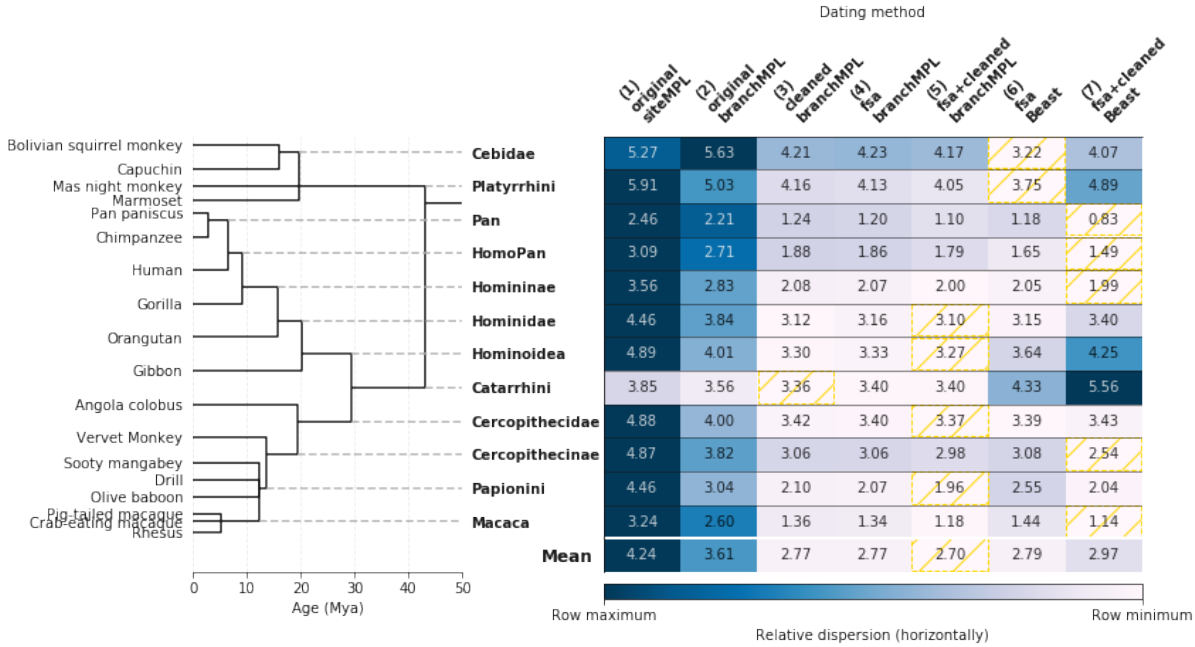

Figure 2

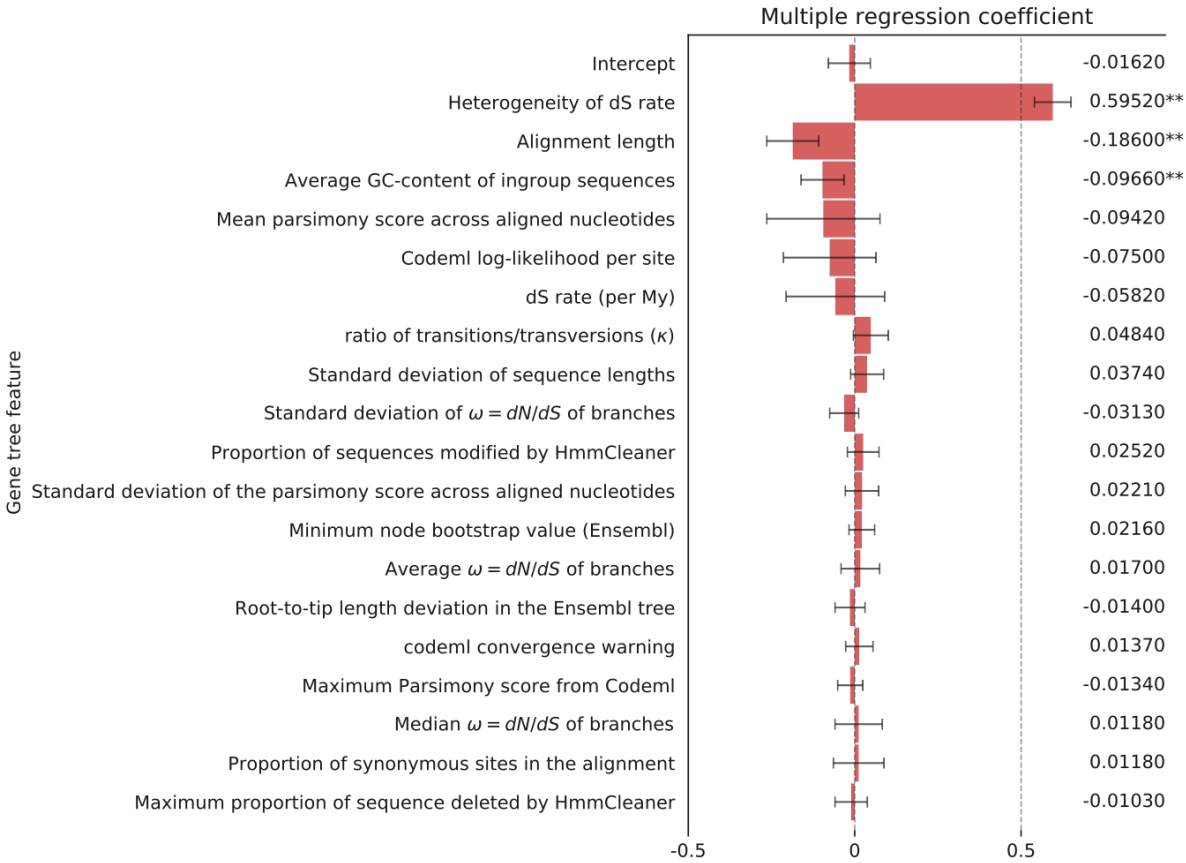

Figure 3

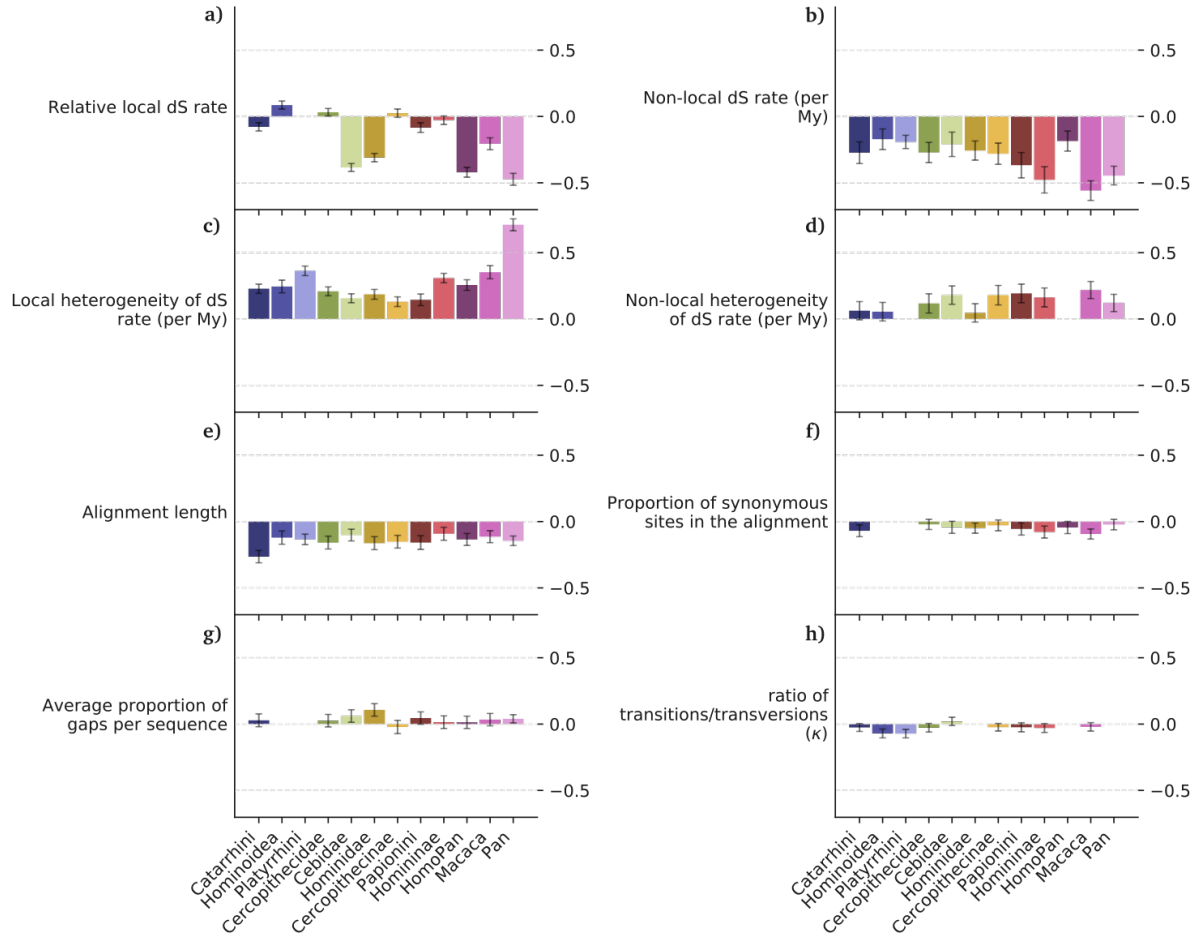
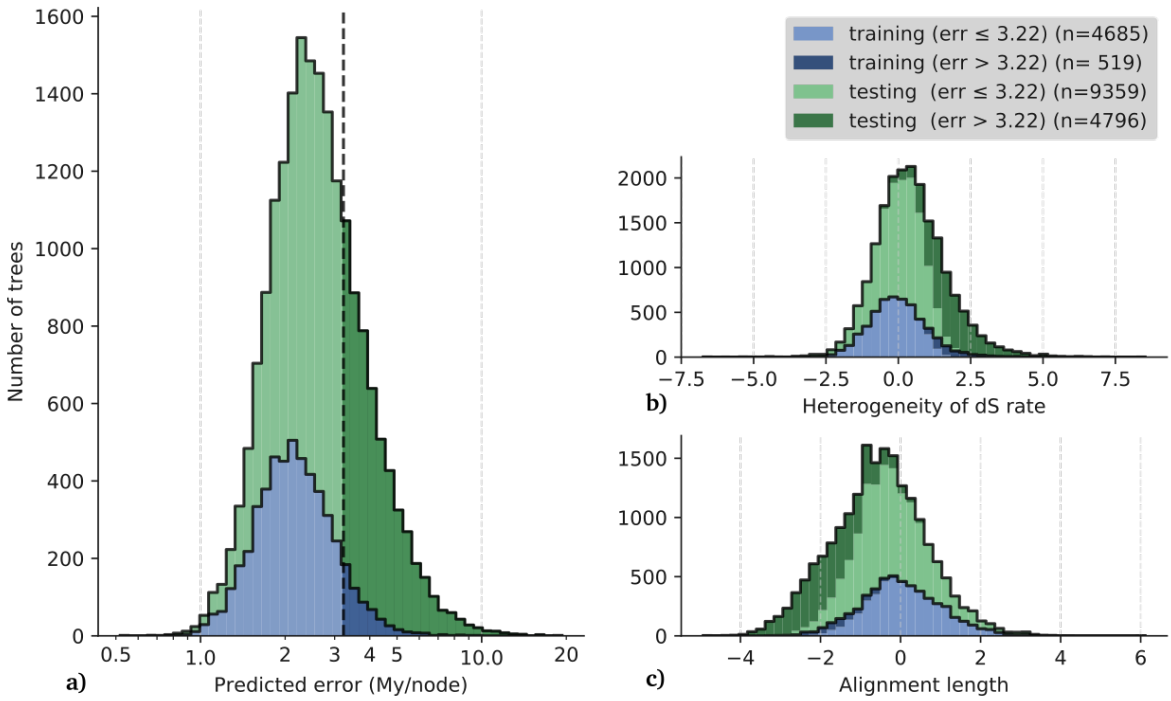
Figure 4