Probabilistic modelling improves relative dating from gene phylogenies

Moisès Bernabeu[1,2], Carmen Armero[3] and Toni Gabaldón[1,2,4,5]

1. Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell, 1-3, Barcelona, 08034, Spain
2. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, Barcelona, 08028, Spain
3. Department of Statistics and Operations Research, Universitat de València, Burjassot, 46100, Spain
4. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
5. Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona, Spain

*Corresponding author: toni.gabaldon@bsc.es

Abstract

Establishing the timing of past evolutionary events is a fundamental task in the reconstruction of the history of life. State-of-the-art molecular dating methods generally involve the reconstruction of a species tree from conserved, vertically evolving genes, and the assumption of a molecular clock calibrated with the fossil record. Although this approach is extremely useful, its use is limited to speciation events and does not account for genes following different evolutionary paths. Recently, an alternative methodology for the relative dating of evolutionary events has been proposed that considers the distribution of branch lengths across sets of gene trees. Here, we validate this methodology using a fossil-calibrated phylogeny and propose a model-based formalisation using a Bayesian framework. Our analyses revealed that the normalisation of the compared branch lengths with branch lengths of a shared reference clade results in narrower distributions, allowing the correct inference of the relative ordering of evolutionary events. Moreover, we show that distributions of normalised lengths can be modelled using gamma or lognormal distributions. Finally, we demonstrate that inference of the posterior distribution of the mode allows accurate relative age estimation, as assessed by a strong correlation with the molecular clock-dated tree. Overall, we provide a novel, model-based approach to infer relative ages from sets of gene phylogenies.

Keywords: Phylogenomics; Relative dating; Bayesian Inference; Gene trees; Branch length distribution

Introduction

Evolutionary biology aims at reconstructing the past history of living organisms. This process involves inferring a timeline, which minimally includes a relative ordering of events and, ideally, time estimates framed within the geological history of Earth. The fossil record, coupled with radiometric dating and stratigraphy, can provide relatively accurate estimates of the time at which different groups of organisms lived, but its application is mostly limited to macro-organisms containing fossilisable structures.

Molecular dating is a more recent dating approach that exploits the fact that homologous sequences accumulate differences with time to estimate how long ago they diverged. This approach assumes a Molecular clock that correlates evolutionary time to genetic sequence divergence (Zuckerkandl & Pauling, 1965). Importantly, the parameters of the molecular clock can be calibrated using dated fossils, thereby providing dates that can be placed along the geological timeline. The standard molecular dating approach starts by reconstructing the evolutionary relationships of extant species using their genetic information (Boussau & Daubin, 2010; Zuckerkandl & Pauling, 1965). For this, sets containing exclusively orthologous genes are selected (Kapli et al., 2020). Moreover, gene families showing phylogenetic signal saturation (Philippe et al., 2011) or evolving via horizontal gene transfer are discarded. The resulting species tree is calibrated by assigning dated fossils to certain ancestral nodes, this allows to introduce constraints to species divergence times and infer the molecular change rate, which in turn allows us to provide divergence time estimates to all speciation events included in the tree (Dos Reis et al., 2016).

This approach has been successful in dating the divergence of many macro-organismal lineages (Kumar et al., 2017). However, it presents several limitations. Firstly, the correct interpretation of the fossil record is crucial to obtaining an accurate dating, and different studies often provide conflicting results (Porter & Riedman, 2023). Secondly, the majority of the organisms across the Tree of Life lack a robust fossil record, or even known fossils.

Thirdly, the molecular dating approach is limited to reconstructed speciation events and cannot precisely date events outside the tree nodes.

To address some of these problems and aiming to provide insights into a broader set of evolutionary events, an alternative methodological framework using gene trees has been proposed that uses branch length ratios in gene trees to infer relative times (Pittis & Gabaldón, 2016a). This method, initially applied to investigate gene acquisition events in the lineage leading from the first eukaryotic common ancestor to the last common ancestor of extant eukaryotes, provides relative dating by using normalised branch lengths (Pittis & Gabaldón, 2016a; Susko et al., 2021; Vosseberg et al., 2021).

Although the branch length ratio method provides a new framework for analysing the relative timing of evolutionary events independently of the fossil record and is not just limited to speciation events, several caveats have been already discussed. Firstly, the presence of unsampled or extinct (ghost) lineages may confound to some extent the branch length ratio analysis conclusions, particularly when applied to gene transfers (Susko et al., 2021; Tricou et al., 2022). However, simulations show that inferences about relative timing of events are overall more likely to be correct, and that the extent of the risk and the potentially conflicting lineages can be assessed for a given backbone species phylogeny including the events of interest (Bernabeu et al., 2024). Secondly, the modelling implemented by (Pittis & Gabaldón, 2016a) to separate waves of gene acquisitions (a mixture of normal distributions) was criticised by showing that a lognormal distribution better fit the data (Martin et al., 2017). Although the conclusions of the paper were not dependent on this modelling (Pittis & Gabaldón, 2016b; Susko et al., 2021), the criticism underscored the lack of a thorough mathematical formalisation of the method.

Here, we developed a probabilistic framework for the branch length ratio method. To test our methodology, we use a well-established molecular clock-dated tree of mammal evolution as

a ground truth for relative dates that we inferred from genome-wide sets of gene phylogenies (Álvarez-Carretero et al., 2022). We found that both the gamma and lognormal distributions properly fit the empirical distributions of normalised branch lengths, which are mostly skewed. Moreover, the use of a Bayesian framework allows us to infer a posterior distribution for the modes, as the best proxy for event timing, and perform a statistically-sound assessment of their relative ordering.

Methods

Sequence data

We selected a taxonomically-balanced set including 24 out of the 72 species considered in a recently reconstructed dated phylogeny of mammals (Álvarez-Carretero et al., 2022), and downloaded their corresponding genomes and gene annotations from Ensembl v101 (Cunningham et al., 2022) as of September 2022 (Supplementary Table 1).

Phylome generation

We extracted the protein sequence of the longest isoform of each protein encoded in the selected genomes, and reconstructed a phylome (i.e., a complete collection of phylogenies from genes encoded in a genome of interest), using *Homo sapiens* as a seed, and using the PhylomeDB pipeline as implemented in phylomizer (https://github.com/Gabaldonlab/phylomizer) (Fuentes et al., 2022). In brief, for each *Homo sapiens* protein (seed), the pipeline runs a BLAST v2.13.0 (Altschul et al., 1990) against all 24 selected species' proteomes. We selected those hits with a coverage over the query sequence higher than 33% and an e-value lower than 1e-5. In addition, we limited the homologous gene set to the top 200 sequences. These sequences were aligned using MUSCLE v3.8.1551 (Edgar, 2004), MAFFT v7.407 (Katoh & Standley, 2013) and Kalign v2.04 (Lassmann & Sonnhammer, 2005) in forward and reverse orientation. Then, the six resulting

alignments were merged into a consensus alignment using M-Coffee v12.0 (Wallace et al., 2006). The consensus alignment was trimmed using trimAl v1.4.15 (Capella-Gutierrez et al., 2009) with a gap threshold of 0.1 and conserving a minimum of 30% of the positions of the original alignment. This trimmed alignment was used to reconstruct a phylogeny using IQ-TREE v1.6.9 (Nguyen et al., 2015), under the best-fitting model selected from a subset of the available ones (DCmut, JTTDCMut, LG, WAG, VT) using ModelFinder (Kalyaanamoorthy et al., 2017), and, the support was assessed using 1,000 ultra-fast bootstrap replicates.

Tree distance calculation

We implemented a custom script (https://github.com/Gabaldonlab/brlens) for calculating the phylogenetic distances of interest (Fig. 1). This script takes as input a gene tree, the reference species tree (the pruned dated tree from (Álvarez-Carretero et al., 2022)), and a table listing the ancestral events of interest and the species descending from it (the clades table). From the species tree, the script derives two types of information (Fig. 1a, 1): (i) a "species-to-age" dictionary, numbering all nodes ancestral to the seed sequence (from 1, the most recent to n, the most ancient) and indicating, for each other species in the tree the most recent common ancestor (MRCA) with respect to the seed species and (ii) a "first split" dictionary, defining for each considered event the two descendant clades (Fig. 1a, 1).

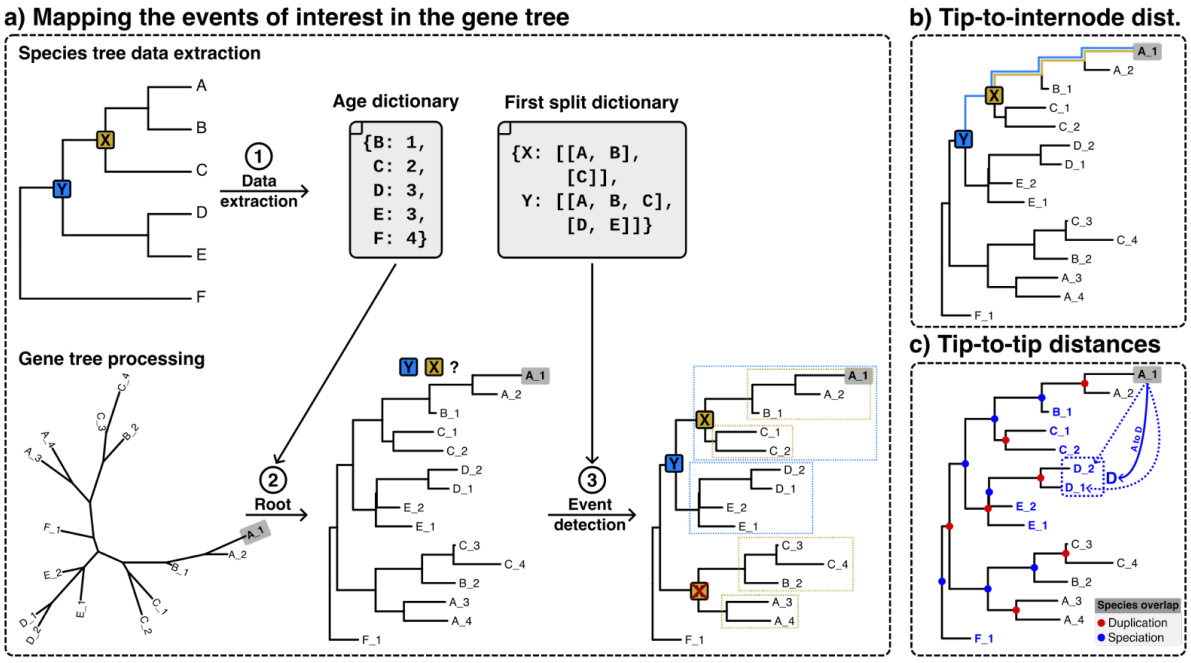


Figure 1. Schematic representation of the computational pipeline used for obtaining evolutionary distances. A1 (in bold and shadowed) represents the seed sequence. a) (1) Extraction of the age and first split dictionaries from the species tree, (2) gene tree rooting using the age dictionary, and (3) detection of the events' MRCA by assessing the presence of the first split, boxes in the tree mean the two monophyletic groups after the first split for each event in the tree, the X event below cannot be considered, for two reasons i) it does not contain the seed, and ii) despite all its descendants belonging to the X group, its first split does not match the species tree one. b) Tip-to-internode distance calculation, once the event's node is detected, the distance between the seed and the node is calculated by summing up the branch lengths in their connecting path. c) The tip-to-tip distance is calculated by obtaining the distance between the seed tip and all its orthologs (in blue) in each of the other species. In the example, the distance of the seed (A1) to the species D is calculated by retrieving the distances to all the co-orthologs belonging to the species D (D1, and D2, indicated with the blue dashed box).

Each gene tree is rooted at its oldest node using the species-to-age dictionary, as described in (Huerta-Cepas et al., 2007), (Fig. 1a, 2). Then, duplication and speciation nodes are inferred using the species overlap algorithm (Gabaldón, 2008), as implemented in the ETE3 Python package (Huerta-Cepas et al., 2016). In addition, the pipeline uses the clades table to label the tree leaves to indicate to which clade they belong (Fig. 1a, 3). We retrieved the subtrees of the events of interest, that is, those monophyletic clades whose MRCA is the event of interest. To this end, we designed an "MRCA function", which retrieves the largest monophyletic subtree containing all the sequences from the species belonging to the target clade that accomplish the following conditions: (i) any MRCA to tip distance is 0, (ii) the first split in the subtree is congruent with the species tree, allowing for missing species; and (iii) the subtree contains the seed sequence (Fig. 1a, 3). We calculated two types of distances, first, tip-to-tip distances, which are the distances from the seed sequence to all the tree tips that are orthologous to the seed. Second, the tip-to-internode distances, which are the distance between the seed and a speciation node corresponding to an internal node.

Normalisation of tree distances

To account for across-gene differences in evolutionary rates, we used a phylogenetic normalisation approach similar to that of (Pittis & Gabaldón, 2016a). Here, we used Primates as the reference clade for normalisation. We have chosen this group as it is close enough to the seed species (*H. sapiens*), and large enough to allow sampling a significant number of branch lengths. We thus normalised the raw distances of interest by dividing them by the median of the MRCA-to-tip distances of the Primates clade. We observed some large distances due to (principally) small normalising groups, which provided near to 0 normalising factors and then extremely large normalised distances. To solve this, we removed normalised distances greater than the 99th quantile for the tip-to-internode distances and the 90th quantile for the tip-to-tip distances (we used a more stringent quantile in tip-to-tip distances as we observed extremely long normalised distances caused by the effect of

some gene family expansions). All the tree functions and calculations were implemented using ETE3 (Huerta-Cepas et al., 2016).

Modelling tree distance distributions

Normalised evolutionary distances are positive real numbers that exhibit a right-skewed distribution. Therefore, we need a probability distribution with support for the positive reals to model their stochastic behaviour. The two probability distributions we choose as data-generating models for learning about these distances are the gamma and the lognormal distribution. Both are highly tuned to the shape exhibited by the data and have analytical expressions for their most important features, especially the mode, which will be the characteristic used for the relative dating. The continuous gamma distribution $Ga(\alpha, \beta)$ depends on two parameters, shape $\alpha > 0$ and rate $\beta > 0$. Its mean and variance are α/β and α/β^2, respectively. The mode is 0 if $\alpha < 1$ and $(\alpha - 1)/\beta$ otherwise. The continuous lognormal distribution $logN(\mu, \sigma^2)$ depends on two parameters, $-\infty < \mu < \infty$ and $\sigma > 0$. Its mean and variance are $\exp(\mu + \sigma^2/2)$ and $((\exp\{\sigma^2\} - 1)\exp\{2\mu + \sigma^2\})$, respectively. The mode is $\exp\{\mu - \sigma^2\}$.

Our methodological statistical framework is Bayesian Inference (BI), which we will use to infer the parameters of the two presented distributions. The three essential elements of a Bayesian statistical analysis are: first, a prior probabilistic distribution $\pi(\theta)$ for all the parameters of interest θ. In our case, $\theta = (\alpha, \beta)$ when dealing with the gamma distribution, and $\theta = (\mu, \sigma)$ in the case of the lognormal model. Second, the likelihood function $L(\theta)$ of the parameters for the observed data, which we will represent as D from now on. In our study, the data are the normalised distances. And finally, the posterior distribution for θ, $\pi(\theta|D)$, which combines the prior and the data information using Bayes' theorem as follows

$$\pi(\theta|D) \propto L(\theta)\pi(\theta).$$

We decided to give maximum prominence to the data and minimum to the prior distribution. To this end, we considered prior independence and selected wide and poorly informative uniform distributions, $U(0, 100)$, for each of the parameters.

The subsequent posterior distribution of the parameters of both the gamma and the lognormal distributions is not analytical. For this reason, we use Markov Chain Monte Carlo (MCMC) methods, in particular Gibbs sampling, to obtain an approximate sample of this distribution to allow us to make inference about the target parameters and the derived output quantities. MCMC was implemented via JAGS v4.3.0 (Plummer 2003). We ran three independent chains with 100,000 iterations each, removed 10% of the initial iterations which we considered a burn-in period, and used a thinning of ten iterations. Convergence was assessed using both graphic and numerical diagnostic tools. In particular, the Gelman-Rubin's statistic, which is the quotient of the variances of the chains within and between the independent runs, with values close to 1 indicating convergence, and the effective sample size (ESS) which accounts for the number of independent samples. The greater the ESS, the more samples are suitable to behave as the posterior. In addition, we plotted the traces and autocorrelation for all the parameters and chains.

To assess the robustness and sensitivity of this procedure, we collected random subsamples from the original tree sets for each considered event. For instance, the boreoeutherians' event node is present in a set of trees, B. We sampled random strict subsets of trees ($b_i \subset B$) and repeated the inference process for both models in these subsets. The size of the subsets was determined using a percentage of the original sample size of the trees, we got subsets from 10% ($b_{10\%}$) to 100% (B) including both 15% and 25% to gain insights in this range.

Results

Genome-wide distribution of phylogenetic distances

We first set out to investigate the shape of the distribution of phylogenetic distances obtained from a genome-wide collection of gene phylogenies (i.e., a phylome, Sicheritz-Pontén & Andersson, 2001), to assess their potential to perform relative dating of evolutionary events. For this, we reconstructed the human phylome in the context of 24 mammalian species, for which a recent highly resolved timed phylogeny is available ((Álvarez-Carretero et al., 2022), see Methods). This phylome includes 16,828 gene trees and is available for browsing or download at PhylomeDB with the PhylomeID 0593 (Fuentes et al., 2022). We first measured, for each gene tree, the tip-to-internode phylogenetic distance between the human seed gene and four events of interest: namely the origin of primates, boreoeutherians, placentals, and therians. All resulting tip-to-internode distances distributions were close to 0 and largely overlapped (Fig. 2a). As previously done by Pittis and Gabaldón (2016a), and to account for differences in evolutionary rates across gene families, we normalised the raw distances by dividing them by the median of the branch lengths observed in the primates clade (see Methods). This normalisation resulted in sharper distributions that are farther away from 0 and are more separated (Fig. 2b). This allows a better relative timing of the considered events based on the ordering of the peaks of these normalised distance distributions. This ordering agrees with the sorting based on the dated species tree. As expected from an erosion of the phylogenetic signal with time, distances to older events were associated with a higher dispersion, and with a lower number of gene trees containing that event (Supplementary Fig. 1d).

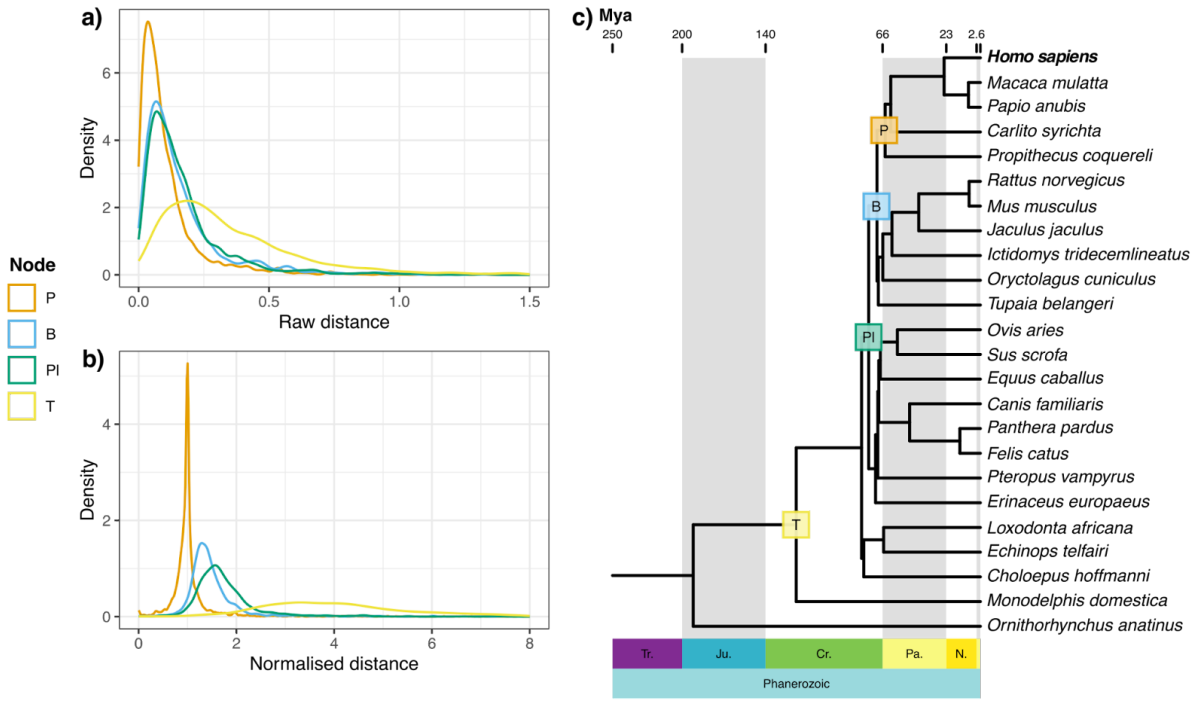


Figure 2. Distributions of the tip-to-internode distances to the events of interest. a) Distribution of tip-to-internode raw distances. b) Normalised tip-to-internode distributions for each event. c) Dated species tree from (Álvarez-Carretero et al., 2022), labelled nodes are those assessed in the distribution of a) and b). The legend shows the group that originated in the event, T: Theria, Pl: Placentalia, B: Boreoeutheria and P: Primates.

We used the same approach to calculate tip-to-tip distances between the seed human sequence of its tree and its orthologs in each of the other species (see Methods). In this case, raw distances (Supplementary Fig. 2) show values in a restricted range around 0 and 2, while normalised distances (Supplementary Fig. 3) had a wider range, from 0 to 15, and were more separated and easier to discriminate. Unexpectedly, we found that the distances to the closest species to the human seed (*Papio anubis* –PAPAN– and *Macaca mulatta* –MACMU) had bimodal distributions. Upon further investigation, we found that gene trees underlying the first and second peaks differed in the encoded functions: the first peak (including genes

having shorter normalised distances) contained mostly informational genes (DNA and RNA processing), whereas genes underlying the second peak (having longer normalised distances) are enriched in metabolic functions (Supplementary Methods 1.2 and Supplementary Figs. 4-7). From these analyses, we conclude that the normalisation of tip-to-internode and tip-to-tip phylogenetic distances from collections of gene phylogenies has the potential to infer a correct relative timing of evolutionary events, as proposed earlier (Pittis & Gabaldón, 2016b; Susko et al., 2021). We also note that rate differences may result in multimodal distributions after normalisation, particularly in recent events. For comparison, we explored alternative normalisation approaches, but concluded that they did not provide significant advantages over this one (see Supplementary Results).

Modelling the distribution of genome-wide phylogenetic distances

To infer the underlying probabilistic distribution of the branch lengths, we used gamma and lognormal models. We carried out Bayesian inference (BI) on the parameters of these distributions using MCMC sampling via JAGS (Plummer, 2003). We set non-informative prior distributions as the prior for each parameter of the gamma distribution, $\pi(\alpha) = \pi(\beta) = U(0, 100)$, as well as for each parameter of the lognormal distribution, $\pi(\mu) = \pi(\sigma) = U(0, 100)$. We ran three independent MCMC chains with 100,000 iterations each, removed 10% of the initial iterations which we considered as a burn-in period, and used a thinning of ten iterations. We further checked the convergence and autocorrelation using both numerical and graphical methods. We repeated the inference process in random subsamples of trees for each studied event.

The three independent chains for both models and parameters converged and provided enough posterior samples to infer the posterior distribution of the parameters. In the case of the gamma distribution, $\widehat{R}$ values are close to 1 meaning that the within and between chains variability is similar. Moreover, the ESS values range from ~5,800 to ~92,000, these high ESS

values allow us to treat the posterior samples as independent. Despite all the parameters converging, the autocorrelation decreases slowly (mean autocorrelation at lag 3 of 0.35); however, it reaches 0 in some lags. Regarding the lognormal model, all the $\widehat{R}$ values are close to 1, as in the gamma model. The ESS ranges from ~69,000 to ~97,000, providing more independent posterior samples than the gamma model. The lognormal model improves the autocorrelation, which reaches a mean value of 0.003 in just 3 lags (Supplementary tables 2-4). Both models accurately fitted the branch lengths.

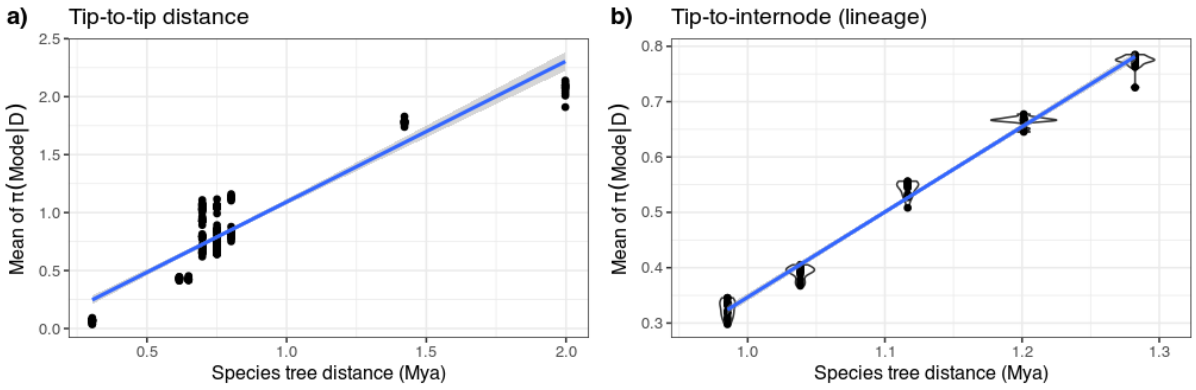
The mode of the inferred Gamma distribution of branch lengths as a proxy for relative timing

We next explored the potential of model-based inference for the relative timing of evolutionary events using sets of gene trees. Evolutionary events (such as the origin of a new clade) are usually assumed to be punctual. Then, the resulting branch length variability should result from analytical (alignment or phylogenetic inference errors) or biological factors (varying evolutionary rates not captured by the normalisation). Given the non-symmetrical nature of the distance distributions, we hypothesised that the mode of the distribution is the best proxy for the time point of interest, and tested this assumption by comparing the inferred modes with the corresponding distances in the dated tree (Álvarez-Carretero et al., 2022). The posterior distributions of the modes for each event (Fig. 4a) had very low dispersion, which means that different events can be easily distinguished.

To test whether the inferred modes were accurately retrieving temporal information from gene trees, we compared them with the molecular clock-dated species tree of our set of species (Álvarez-Carretero et al., 2022). We obtained the distances in Million years ago (My) from the *H. sapiens* tip to all the internal nodes in the path from the tip to the root in the dated species tree, and the same normalised distances for the phylome set of gene trees. Then we calculated the posterior distribution of the mode for each node. We also used the subsequent posterior distribution of the tip-to-tip distance and the corresponding distances in the species tree. The correlation between both dating methodologies was high (Fig. 3).

306 Regarding the tip-to-tip distances (Fig. 3a), there are several species with equal distances,
307 this is expected for the ultrametric property of the dated species tree, which means that the
308 distance to all members of a monophyletic sister clade will be the same. Despite this, the
309 inferred normalised distances agree with those in the dated tree. This correlation is even
310 higher when focusing on the distances within the lineage leading to humans (Fig. 3b). These
311 results indicate that the normalised distances obtained from collections of gene trees are a
312 good proxy of time, and that they allow a correct sorting of evolutionary events.

313



314

315 **Figure 3. Correlation between the mean of the posterior sample for the mode and the**
316 **species tree distance**. a) Tip-to-tip distance $R^2 = 0.91$ and b) tip-to-internode (human
317 lineage except the human speciation and the root) distance $R^2 = 0.99$.

318

319 ***A probabilistic framework for relative timing***

320 In BI, we can compare two parameters, for example θ_1 and θ_2, through the subtraction or
321 division of their posterior distributions. In the case of the subtraction, which we used here,
322 the posterior distribution would be $\pi(\theta_1 - \theta_2|D)$, where θ_i is the mode of the evolutionary
323 event i. Thus, posterior distributions for the subtraction of two events ($\pi(\theta_1 - \theta_2|D)$) around
324 zero, would indicate no relevant difference between θ_1 and θ_2, and the most likely hypothesis

is that they happened simultaneously. Conversely, posterior distributions for the subtraction far from 0 would refer to events that happened at different times.

We assessed whether this method accurately distinguished contiguous evolutionary events (Fig. 4b). All the comparisons between the timing of the events concluded that they occurred at different times with a probability of 1, although some of the compared clades occurred close in evolutionary absolute time (e.g., the origin of Placentalia and Boreoeutheria).

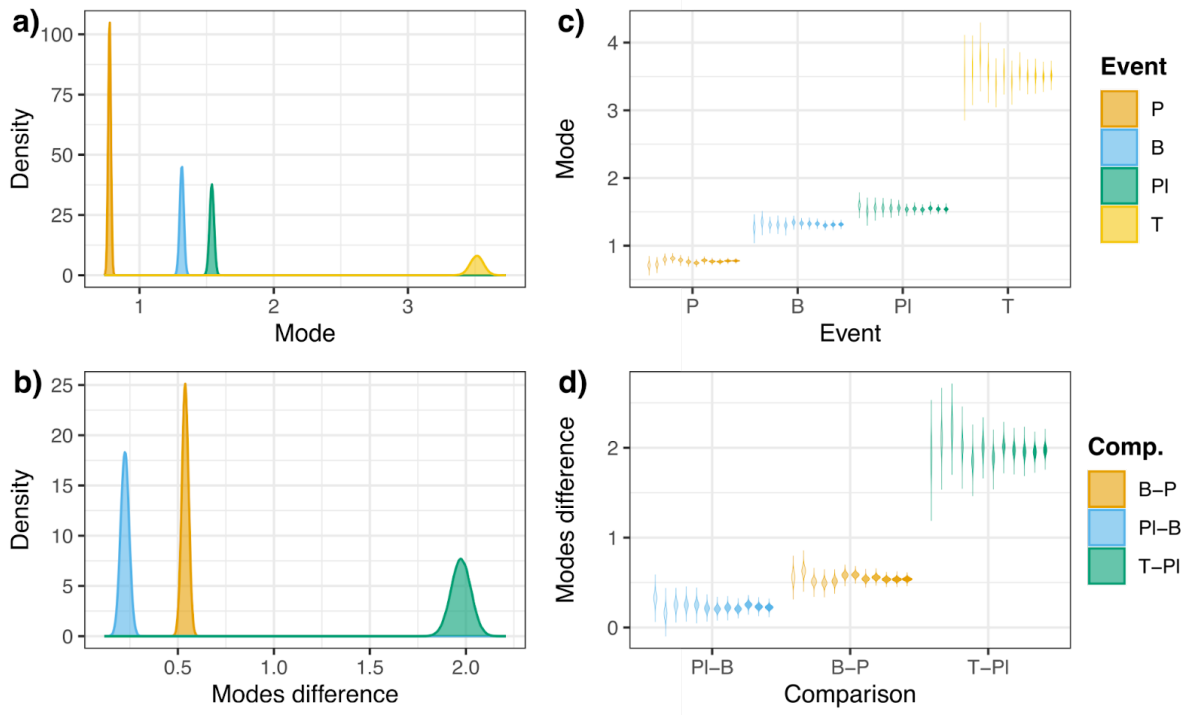


Figure 4. Posterior distributions of the events relative timing. a) Posterior distribution of the mode for each event. b) Posterior distributions of the difference between modes of adjacent events (Comp.: comparison groups). c) Posterior distributions, shown as violin plots, of the mode of each event for each subsampling set. d) Posterior distributions, shown as violin plots, of the difference between modes of adjacent events for each subsampling set. Subsampling ranges from 10% of the total number of trees to 100%, increasing by 10% each,

we also included 15% and 25% of the trees. The more opacity of the violin, the higher the percentage of trees used. T: Theria, Pl: Placentalia, B: Boreoeutheria and P: Primates.

To assess the robustness of the approach, we repeated the inference on random subsets of the trees (Fig. 4c). The inferred modes of the subsets and the full-dataset inferred were always similar, with deviations ranging from 0.69% in the Boreoeutheria node to 10.47% in the primates node for the smaller subset. In this study, we used a seed-based approach, in which homologs are inferred by searching the seed sequence in a given set of genomes. Because of this homology search strategy, the events closer to the seed sequence are over-represented with respect to the ancestral ones. As a result of this, the subsampling has a stronger effect on the variability of the deeper nodes, as seen in Fig. 4c. Despite the congruence with the inferred mode, the standard deviation of the posterior distributions for the mode ranges between 0.005 in the primates node (most recent event) using the complete tree set to 0.081 in therians (deepest event) using 10% of trees (Supplementary Figs. 9 and 10a). As expected, the standard deviation of the inference increased when we used fewer trees, as there is less information to infer the statistic (Supplementary Table 5). Importantly, however, the standard deviation values are small relative to the distance between modes and thus, such variation is unlikely to cause a shift in the relative timing (Fig. 4d). Furthermore, the standard deviation does not decrease constantly, there are some points in which the standard deviation of the posterior increases for a higher number of trees (25% for Theria and 20% for Primates subsample in Supplementary Fig. 10a).

This suggests that the nature of the gene trees in the subsample, and not just the number of trees, is important. We used the same subsampling strategy to assess whether, when using fewer trees, the method allows us to discriminate between events (Fig. 4d). The subtraction of posterior samples using less data showed that, despite using only 10% of the trees, the

studied events could be differentiated with probability 1. As we observed in the events' case, the variation between the standard deviation of the subsamples is around 70% (Supplementary Table 6). However, the farthest comparison (Theria origin against Placentalia) shows a slightly higher standard deviation (Supplementary Fig. 11a). The deviation from inferred differences between modes (i.e., using all the trees) is between 13% and 25% (Supplementary Table 6). Despite this, the measure of interest here is the area under the curve of the modes' comparison ($P(mode_1 - mode_1 > 0 \mid D)$) when it is greater than 0, for the closest event we found that using the 10% of the trees the probability that Placentalia clade originated before Boreoeutheria clade is ~1. Despite variations in the dispersion of the distributions, the differences between the inferred mode from different subsamples are low for both the events and the comparisons. This test supports the robustness of the inference method to provide probabilistic information about the sorting of evolutionary events.

Discussion

Here, we have formalised and tested the branch length ratio method for the relative timing of evolutionary events based on gene trees (Pittis & Gabaldón, 2016a; Susko et al., 2021). A key step in this method is the normalisation of a phylogenetic distance of interest by the median of the branch lengths in an evolutionarily consistent clade present in all the gene phylogenies. Our results show that, as compared to raw distances, normalised distances serve to better discriminate the timing of evolutionary events. Moreover, we show that these distances, inferred from the distributions of branch lengths in gene tree sets, show a high correlation with molecular clock dating of a species tree based on a concatenated set of single-copy orthologous genes. Finally, we show that our implementation of the approach in a Bayesian framework enables a probabilistic interpretation of the relative timing of events.

The branch length ratio method can use gene family trees including duplication events. In comparison, molecular dating of a species tree generally relies exclusively on single-copy genes. As a result, the branch length ratio method can exploit a larger amount of available data. We here show that the distribution of normalised distances in collections of gene trees exhibits variability but consistently has a mode that correlates well to the timing of the studied evolutionary event. We furthermore show that the use of Bayesian inference can account for uncertainty and allow precise estimations of the relative timing of compared events.

The branch lengths of a gene tree depend mainly on the divergence time among sequences and the evolutionary rate of the gene. Although the rate in all the tree branches is not necessarily conserved due to potential heterotachy, we here show that branch length normalisation and the modelling of the distributions of these normalised branch lengths across gene trees effectively provides accurate timing information. Regarding this normalisation, we have found that the median for the root-to-tip distances of a specific clade preserved across the gene trees is a reliable measure of the rate, as shown by the high observed correlation with molecular-clock-based absolute dating. Telford et al. (2014) had previously used the tree length (the sum of all branch lengths) divided by the number of leaves as a proxy of rate, but here we show that this measure is strongly affected by heterotachy. Moody et al. (2022) used the mean root-to-tip distance of the minimal ancestor deviation rooted gene tree. This measure, as the one proposed before, does not account for the asymmetry of the branch length distributions. Nevertheless, this measure is more accurate as it uses an evolutionary distance, from an event common in all the trees to the present although the event comprises the whole tree. Here we assumed that the set of paths from the primates' origin event to the present (i.e., primates MRCA to tip distances) is a good representation of the gene evolutionary rate, as Pittis and Gabaldón (2016a) previously did by using the paths from the Last Eukaryotic Common Ancestor (LECA) to the present. This

measure has been proven as a way for obtaining a relative-to-time measure by assuming branch lengths as a random variable with its intrinsic variability.

Recently, Moody et al. (2022) revealed that species trees obtained from concatenated conserved genes result in different estimations for the archaeal-bacterial branch length depending on verticality, functional and model shifts, such as using only ribosomal proteins or proteins with signs of HGT, among others. Moreover, Eme et al. (2023) have shown that some species are artifactually close in the species phylogeny when using ribosomal protein sets due to coevolution in similar environmental pressures. Using a functionally broader gene set, they retrieved the currently accepted topology. Thus, the use of larger gene sets, as enabled by the branch length ratio method is likely to alleviate functional and rate biases typical of reduced gene sets.

Here, we modelled the evolutionary distances as a gamma-distributed random variable, which explains the relative timing of an evolutionary event. Similarly to the molecular clock method, we assumed that the event was punctual (i.e., it occurred once in a time frame). Thus, our modelling was focused on retrieving a specific distribution value rather than the whole distribution Given that the normalised distributions were primarily asymmetrical we opted for the mode, rather than the mean of the distributions as a proxy for the time of the event.

As proposed by Martin et al. (2017), we tested the lognormal distribution. Although both lognormal and gamma distributions showed similar behaviours, we believe the latter to be more intuitive in the context of sequence evolution, as it is the one currently used in phylogenetics to model rate variation among sites. Nevertheless, any distribution with positive support and skewness could potentially model the relative dating (once its suitability is proved).

Thanks to Bayesian modelling, we can assign probabilities to comparisons of the timing of two or more events. This can be done in a robust way, even when using a reduced set of genes. Despite all these advantages, the branch length ratio method is sensitive to genes with a high degree of heterotachy. The presence of heterotachy between the target and the normalising lineages is expected to result in anomalously large or short normalised branch lengths due to the disparity in evolutionary rates. Nevertheless, these trees can be detected and removed. By using across-genome comprehensive collections of gene trees (phylomes) distances and a Bayesian framework to model them, our work shows that we can retrieve reliable inferences of relative timing for ancestral events that correlate with previously used dating methodologies and provides a new approach for comparing evolutionary events.

Acknowledgements

We thank members of the Gabaldón group for insightful discussions.

Funding

We acknowledge support from the Spanish Ministry of Science and Innovation for grants PID2021-126067NB-I00, CPP2021-008552, PCI2022-135066-2, and PDC2022-133266-I00, cofounded by ERDF “A way of making Europe”; from the Catalan Research Agency (AGAUR) SGR01551; from the European Union’s Horizon 2020 research and innovation programme (ERC-2016-724173); from the Gordon and Betty Moore Foundation (Grant GBMF9742); from the “La Caixa” foundation (Grant LCF/PR/HR21/00737), and from the Instituto de Salud Carlos III (IMPACT Grant IMP/00019 and CIBERINFEC CB21/13/00061- ISCIII-SGEFI/ERDF).

Data availability

The raw output data is stored in the Zenodo repository https://doi.org/10.5281/zenodo.8417362. The code used for the distances calculation and the posterior distributions calculation is available in the GitHub repository https://github.com/Gabaldonlab/brlens.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Álvarez-Carretero, S., Tamuri, A. U., Battini, M., Nascimento, F. F., Carlisle, E., Asher, R. J., Yang, Z., Donoghue, P. C. J., & dos Reis, M. (2022). A species-level timeline of mammal evolution integrating phylogenomic data. *Nature*, *602*(7896), 263–267. https://doi.org/10.1038/s41586-021-04341-1

Bernabeu, M., Manzano-Morales, S., & Gabaldón, T. (2024). On the impact of incomplete taxon sampling on the relative timing of gene transfer events. *PLOS Biology*, *22*(3), e3002460. https://doi.org/10.1371/journal.pbio.3002460

Boussau, B., & Daubin, V. (2010). Genomes as documents of evolutionary history. *Trends in Ecology & Evolution*, *25*(4), 224–232. https://doi.org/10.1016/j.tree.2009.09.007

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Dos Reis, M., Donoghue, P. C. J., & Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, *17*(2), 71–80. https://doi.org/10.1038/nrg.2015.8

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5). https://doi.org/10.1093/nar/gkh340

493 Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W.,
494 Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J.,
495 Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G.
496 (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes.
497 *Nature*. https://doi.org/10.1038/s41586-023-06186-2
498 Fuentes, D., Molina, M., Chorostecki, U., Capella-Gutiérrez, S., Marcet-Houben, M., & Gabaldón,
499 T. (2022). PhylomeDB V5: An expanding repository for genome-wide catalogues of
500 annotated gene phylogenies. *Nucleic Acids Research*, *50*(D1), D1062–D1068.
501 https://doi.org/10.1093/nar/gkab966
502 Gabaldón, T. (2008). Large-scale assignment of orthology: Back to phylogenetics? *Genome*
503 *Biology*, *9*(10), 235. https://doi.org/10.1186/gb-2008-9-10-235
504 Huerta-Cepas, J., Dopazo, H., Dopazo, J., & Gabaldón, T. (2007). The human phylome.
505 *Genome Biology*, *8*(6), R109. https://doi.org/10.1186/gb-2007-8-6-r109
506 Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and
507 Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6),
508 1635–1638. https://doi.org/10.1093/molbev/msw046
509 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017).
510 ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature*
511 *Methods*, *14*(6). https://doi.org/10.1038/nmeth.4285
512 Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age.
513 *Nature Reviews Genetics*, *21*(7), 428–444.
514 https://doi.org/10.1038/s41576-020-0233-0
515 Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
516 Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4).
517 https://doi.org/10.1093/molbev/mst010
518 Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for

519 Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7),
520 1812–1819. https://doi.org/10.1093/molbev/msx116
521 Lassmann, T., & Sonnhammer, E. L. (2005). Kalign – an accurate and fast multiple sequence
522 alignment algorithm. *BMC Bioinformatics*, *6*(1), 298.
523 https://doi.org/10.1186/1471-2105-6-298
524 Martin, W. F., Roettger, M., Ku, C., Garg, S. G., Nelson-Sathi, S., & Landan, G. (2017). Late
525 Mitochondrial Origin Is an Artifact. *Genome Biology and Evolution*, *9*(2), 373–379.
526 https://doi.org/10.1093/gbe/evx027
527 Moody, E. R., Mahendrarajah, T. A., Dombrowski, N., Clark, J. W., Petitjean, C., Offre, P.,
528 Szöllősi, G. J., Spang, A., & Williams, T. A. (2022). An estimate of the deepest
529 branches of the tree of life from ancient vertically evolving genes. *eLife*, *11*, e66695.
530 https://doi.org/10.7554/eLife.66695
531 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and
532 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
533 *Molecular Biology and Evolution*, *32*(1), 268–274.
534 https://doi.org/10.1093/molbev/msu300
535 Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., &
536 Baurain, D. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences
537 Are Not Enough. *PLoS Biology*, *9*(3), e1000602.
538 https://doi.org/10.1371/journal.pbio.1000602
539 Pittis, A. A., & Gabaldón, T. (2016a). Late acquisition of mitochondria by a host with
540 chimaeric prokaryotic ancestry. *Nature*, *531*(7592), 101–104.
541 https://doi.org/10.1038/nature16941
542 Pittis, A. A., & Gabaldón, T. (2016b). *On phylogenetic branch lengths distribution and the late*
543 *acquistion of mitochondria* [Preprint]. Biorxiv. https://doi.org/10.1101/064873
544 Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using

Gibbs Sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*.

Porter, S. M., & Riedman, L. A. (2023). Frameworks for Interpreting the Early Fossil Record of Eukaryotes. *Annual Review of Microbiology*, *77*(1), 173–191. https://doi.org/10.1146/annurev-micro-032421-113254

Sicheritz-Pontén, T., & Andersson, S. G. (2001). A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, *29*(2), 545–552. https://doi.org/10.1093/nar/29.2.545

Susko, E., Steel, M., & Roger, A. J. (2021). Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events. *Journal of Theoretical Biology*, *526*, 110788. https://doi.org/10.1016/j.jtbi.2021.110788

Telford, M. J., Lowe, C. J., Cameron, C. B., Ortega-Martinez, O., Aronowicz, J., Oliveri, P., & Copley, R. R. (2014). Phylogenomic analysis of echinoderm class relationships supports Asterozoa. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1786), 20140479. https://doi.org/10.1098/rspb.2014.0479

Tricou, T., Tannier, E., & de Vienne, D. M. (2022). Ghost lineages can invalidate or even reverse findings regarding gene flow. *PLOS Biology*, *20*(9), e3001776. https://doi.org/10.1371/journal.pbio.3001776

Vosseberg, J., van Hooff, J. J. E., Marcet-Houben, M., van Vlimmeren, A., van Wijk, L. M., Gabaldón, T., & Snel, B. (2021). Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nature Ecology & Evolution*, *5*(1), 92–100. https://doi.org/10.1038/s41559-020-01320-z

Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6), 1692–1699. https://doi.org/10.1093/nar/gkl091

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history.

571 *Journal of Theoretical Biology*, *8*(2), 357–366.

572 https://doi.org/10.1016/0022-5193(65)90083-4