Bioinformatics, 36(26), 2020, i884–i894
doi: 10.1093/bioinformatics/btaa820
ECCB2020

Data

Using a GTR+Γ substitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated

Jose Barba-Montoya[1,2], Qiqing Tao[1,2] and Sudhir Kumar[1,2,3,*]

[1]Institute for Genomics and Evolutionary Medicine, [2]Department of Biology, Temple University, Philadelphia, PA 19122, USA and [3]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia

*To whom correspondence should be addressed.

Abstract

Motivation: As the number and diversity of species and genes grow in contemporary datasets, two common assumptions made in all molecular dating methods, namely the time-reversibility and stationarity of the substitution process, become untenable. No software tools for molecular dating allow researchers to relax these two assumptions in their data analyses. Frequently the same General Time Reversible (GTR) model across lineages along with a gamma (+Γ) distributed rates across sites is used in relaxed clock analyses, which assumes time-reversibility and stationarity of the substitution process. Many reports have quantified the impact of violations of these underlying assumptions on molecular phylogeny, but none have systematically analyzed their impact on divergence time estimates.

Results: We quantified the bias on time estimates that resulted from using the GTR + Γ model for the analysis of computer-simulated nucleotide sequence alignments that were evolved with non-stationary (NS) and non-reversible (NR) substitution models. We tested Bayesian and RelTime approaches that do not require a molecular clock for estimating divergence times. Divergence times obtained using a GTR + Γ model differed only slightly (∼3% on average) from the expected times for NR datasets, but the difference was larger for NS datasets (∼10% on average). The use of only a few calibrations reduced these biases considerably (∼5%). Confidence and credibility intervals from GTR + Γ analysis usually contained correct times. Therefore, the bias introduced by the use of the GTR + Γ model to analyze datasets, in which the time-reversibility and stationarity assumptions are violated, is likely not large and can be reduced by applying multiple calibrations.

Availability and implementation: All datasets are deposited in Figshare: https://doi.org/10.6084/m9.figshare.12594638.

Contact: s.kumar@temple.edu

1 Introduction

Biological evolution at the molecular level is inherently complex. Nucleotide and amino acid substitution patterns vary from species to species, locus by locus and over time (Arenas, 2015; Nei and Kumar, 2000; Yang, 2014). Considerable attention has been paid to developing substitution models that better reflect the process of molecular evolution, resulting in increasingly complex, realistic evolutionary models for phylogenomic studies (Arenas, 2015; Yang, 2014). Markov models thoroughly describe the substitution processes that embrace the presence of biased base/amino acid compositions, differences in transition/transversion rates, non-uniformity of evolutionary rates among sites and differences in substitution patterns among genomic regions (Arenas, 2015; Tao *et al.*, 2020).

Widely used substitution models in molecular phylogenetics assume time-reversibility and stationarity of the substitution processes over the whole phylogenetic tree (Galtier and Gouy, 1998; Jayaswal *et al.*, 2011; Yang, 2014). The time-reversibility assumption requires that the instantaneous rate of change from base *i* to base *j* is equal to that of base *j* to *i* (Nei and Kumar, 2000). For large datasets, this assumption is expected to be frequently violated, and an unrestricted model is usually a better fit (Yang, 1994, 2014). Although this complexity is well appreciated in molecular evolutionary research, including phylogenetics and systematics, a vast majority of researchers employ a General Time Reversible (GTR) class of substitution models (Fig. 1).

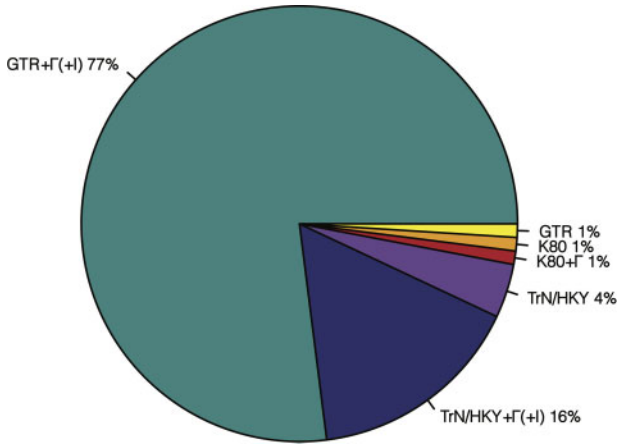

Fig. 1. A survey of substitution models selected in 141 research articles that published timetrees in year 2015–2017. More than 130 studies (>98%) used models that have more free parameters than the K80 model. All studies assumed stationarity and time-reversiblity of evolutionary processes, with the GTR + Γ and GTR + Γ+I being the most preferred models. K80, HKY, TrN and GTR represent Kimura-2-parameter (Kimura, 1980), Hasegawa–Kishino–Yano (Hasegawa *et al.*, 1985), Tamura–Nei (Tamura and Nei, 1993) and GTR model (Tavaré, 1986), respectively. Model + Γ(+I) means that either a gamma distribution for incorporating rate variation across sites is used, or a proportion of sites are assumed to be invariant across sequences, or both are used along with the corresponding substitution model

The GTR model provides for different rates for all the transitions and transversional substitutions as well as the unequal frequency of bases. In addition to time-reversibility, the use of a GTR model in phylogenetic methods, as implemented in most of the software packages, automatically assumes that the substitution process does not change over time; i.e. it is stationary. This translates into assuming that the rates of different types of base substitutions are the same across evolutionary lineages and over time. Violation of this stationarity assumption is evident from differences in base composition across sequences (e.g. Galtier and Gouy, 1998; Kumar and Gadagkar, 2001; Rosenberg and Kumar, 2003; Tamura and Kumar, 2002). Studies of empirical data have shown that unequal base compositions can mislead methods of phylogenetic inference by grouping sequences according to the base compositions instead of their phylogenetic relationships (Galen *et al.*, 2018; Galtier and Gouy, 1995; Lockhart *et al.*, 1994; Rosenberg and Kumar, 2003). Therefore, many models and methods that avoid the stationarity assumption have been developed (Blanquart and Lartillot, 2006, 2008; Foster, 2004; Galtier and Gouy, 1998; Tamura and Kumar, 2002; Yang, 1994).

Therefore, we expect that the use of the GTR model for substitution rates and phylogenetic inference would cause bias because it is an oversimplification of the correct model. This bias is known to impact phylogenetic inference (Galen *et al.*, 2018; Huang *et al.*, 2010; Philippe *et al.*, 2017; Singh *et al.*, 2009), but there is little known about its impact on the estimates of molecular divergence times. In an analysis of bias caused by the use of simple substitution models from the GTR class of models, Tao *et al.* (2020) reported that the complexity of the substitution model has a rather limited biasing effect in empirical data analysis. They found that the actual bias in the time estimates (TEs) became very small when even a few clock calibrations are applied. The main focus of this study is to examine if this trend holds when we consider datasets that violate the underlying assumptions of time-reversibility and stationarity made in all the current molecular dating analyses.

In the following, we quantified the bias that results from using the GTR substitution model, along with a provision to account for the rate variation among sites by using a Gamma distribution (GTR + Γ), to analyze computer-simulated nucleotide sequence alignments that were evolved without reversibility or stationarity of substitution process. Datasets were simulated for phylogenies in which evolutionary rates varied extensively among lineages with or without autocorrelation among lineages (Rannala and Yang, 2007; Tao *et al.*, 2019).

We applied Bayesian and RelTime relaxed clock methods (Rannala and Yang, 2007; Tamura *et al.*, 2012) for divergence time estimation. We report that the bias of TEs caused by the use of a GTR + Γ model with assumptions of stationarity and time-reversibility to analyze datasets that violate these assumptions. Results are presented for analyses using only a root calibration, as well as those in which multiple internal calibrations were assumed to be known.

2 Materials and methods

2.1 Data simulation

We conducted computer simulations to generate nucleotide sequence alignments in which the substitutional process was reversible (GTR), non-reversible (NR), or non-stationary (NS). All our analyses were conducted by using a model timetree derived from the bony-vertebrate clade in the Timetree of Life (Hedges and Kumar, 2009), from which we randomly sampled 100 taxa (Figs 2 and 3).

We simulated 50 model trees in which the evolutionary rates among branches were autocorrelated (AR datasets) and another 50 in which the rates varied independently (IR datasets); see Tamura *et al.* (2012) for more details. We used INDELible (Fletcher and Yang, 2009) to generate 150 alignments using 50 AR phylogenies such that three datasets were produced from each phylogeny. For one set of data, the nucleotide substitution followed a GTR model with stationarity and reversibility (GTR–AR dataset). In the second, the substitution process was not time reversible (NR–AR). And, in the third, the substitution process was NS (NS–AR). Similarly, 150 alignments were produced by using IR phylogenies, which resulted in 50 GTR–IR, 50 NR–IR and 50 NS–IR datasets (Fig. 3).

The GTR alignments were simulated under the GTR + Γ ($\alpha = 1.0$) model with 1000 base pairs (bp), a base composition of $\pi_T = 0.3$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.2$ and substitution rate parameters $a = 0.2$, $b = 0.4$, $c = 0.6$, $d = 0.8$, $e = 1.2$, and $f = 1$. The NR sequences were simulated under the unrestricted model, with 4000 bp, and substitution rate parameters T→C = 0.1, T→A = 0.2, T→G = 0.3, C→T = 0.4, C→A = 0.5, C→G = 0.6, A→T = 0.1, A→C = 0.2, A→G = 0.3, G→T = 0.4, G→C = 0.5 and G→A = 1.

The NS alignments were simulated under the three different GTR + Γ ($\alpha = 1.0$) models (mGTR1, mGTR2, and mGTR3), with different base composition and rate matrix for different parts of the phylogeny (Fig. 2). Alignments were 4000 bp long. For mGTR1, we used a base composition of $\pi_T = 0.3$, $\pi_C = 0.2$, $\pi_A = 0.3$, and $\pi_G = 0.2$, and substitution rate parameters $a = 0.2$, $b = 0.4$, $c = 0.6$, $d = 0.8$, $e = 1.2$, and $f = 1$; for mGTR2, we used a base composition of $\pi_T = 0.05$, $\pi_C = 0.45$, $\pi_A = 0.05$, and $\pi_G = 0.45$, and substitution rate parameters $a = 0.1$, $b = 0.2$, $c = 0.3$, $d = 0.4$, $e = 0.6$, and $f = 1$; for mGTR3, we used a base composition of $\pi_T = 0.45$, $\pi_C = 0.05$, $\pi_A = 0.45$, and $\pi_G = 0.05$, and substitution rate parameters $a = 0.15$, $b = 0.3$, $c = 0.45$, $d = 0.5$, $e = 0.75$, and $f = 1$.

2.2 Estimation of divergence times

2.2.1 Bayesian approach

All Bayesian analyses (300 datasets of 100 sequences each) were carried out with the program MCMCTree (Yang, 2007). The correct topology of the 100 taxa tree assumed (Fig. 2) to avoid confounding phylogeny inference errors with divergence time estimation bias. We used the AR model to analyze 150 AR datasets and the IR model for the 150 IR datasets; this was done to avoid confounding the bias due to the misspecification of the branch rates model with the bias due to the violation of stationarity and time-reversibility assumptions.

For all the analyses, we assigned to the overall rate (μ) a gamma hyperprior G(1, 1) with mean 1/1 = 1 substitutions per site per time unit (100 MY) or 10^{-8} substitutions per site per year. To the rate drift parameter (σ^2), we assigned another a gamma hyperprior G(1,

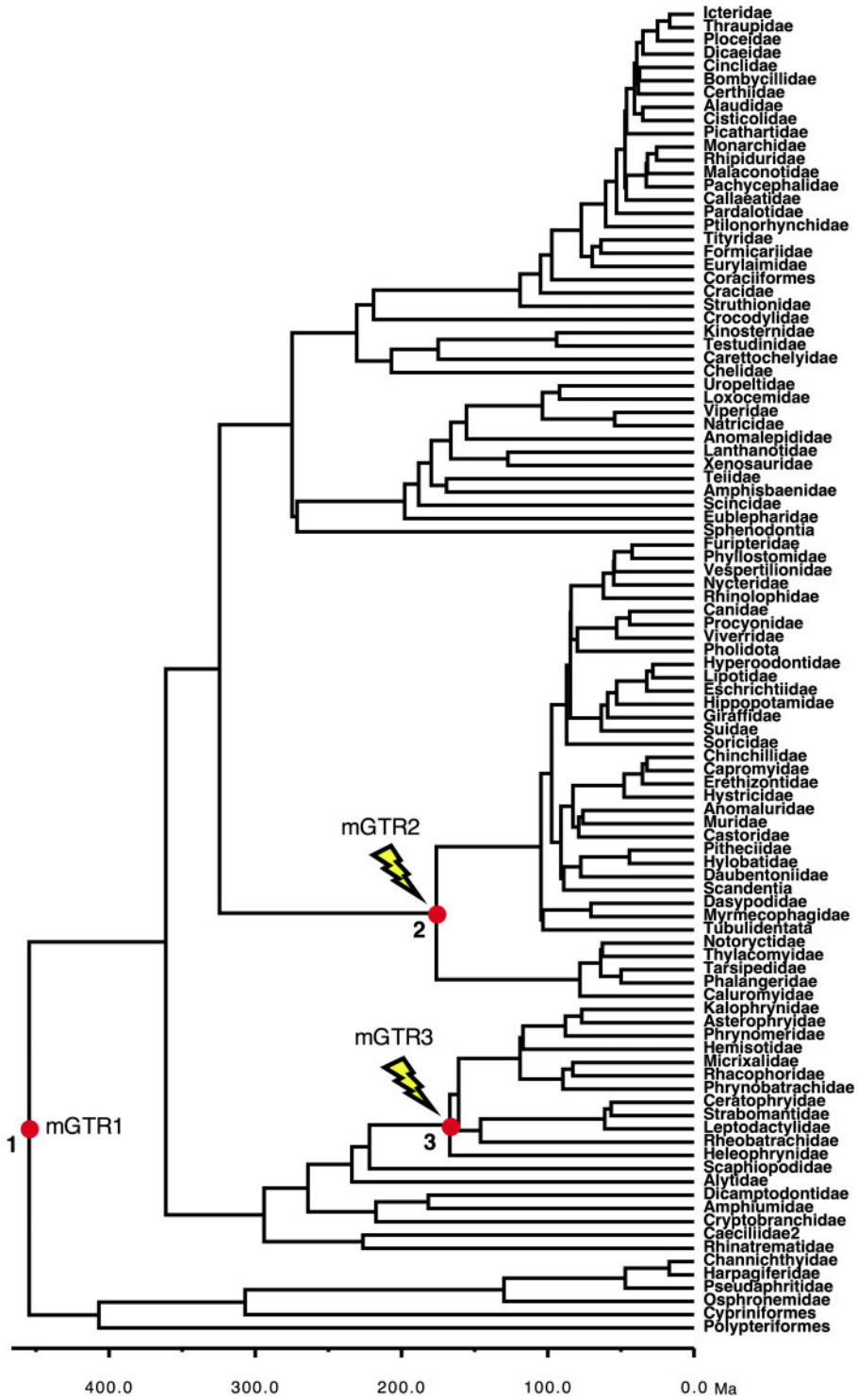

Fig. 2. Phylogeny of 100 taxa showing calibrated nodes. The tree has been scaled to time on the basis of TEs from the Timetree of Life (Hedges and Kumar, 2009). Calibrations are represented for three nodes (red dots). We used a uniform distribution U(min, max) for the three calibrations: (1) root calibration U(444.6, 464.6); (2) Calibration-2 U(166.2, 186.2); (3) Calibration-3 U(157, 177). For the NS alignments, a NS process was added by changing the base composition and rate matrix for two lineages, starting at the ascending branches of node 2 (mGTR2) and node 3 (mGTR3)

1) with mean 1, allowing large rate variation like the ones simulated here (Tamura *et al.*, 2012). The sequence likelihood was calculated under the GTR substitution model with a Γ distribution of site rates (five categories; Yang, 1994). The approximate likelihood method (dos Reis and Yang, 2011; Thorne *et al.*, 1998) was used in maximum likelihood (ML) estimation of branch lengths and the Hessian matrix. The parameters of the birth–death sampling process were fixed at $\lambda = 2, \mu = 2$, and $\rho = 0.6$ (Yang and Rannala, 2006). For each analysis, two runs were performed, each consisting of 5×10^6 iterations after a burn-in of 5×10^4 iterations and sampling every 200, resulting in a total of 5×10^4 samples from the two runs. We checked for convergence by comparing the posterior mean estimates between runs and by plotting the time series traces of the samples. We used two different calibration strategies to investigate the impact

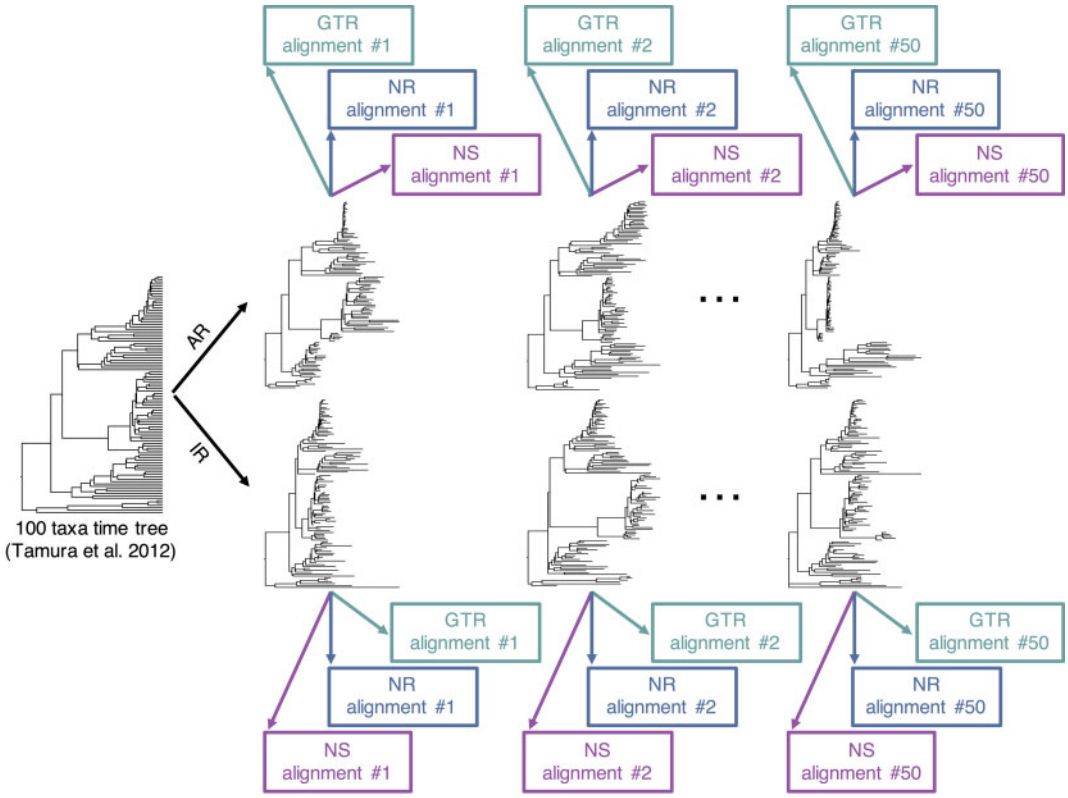

Fig. 3. A flowchart showing an overview of the simulation procedure used to generate datasets. We generated 150 alignments of 100 taxa from 50 phylogenies simulated using the AR model (50 GTR, 50 NR, and 50 NS) and 150 alignments from 50 phylogenies simulated using the IR model (50 GTR, 50 NR, and 50 NS)

of calibrations on time inference: a uniform root calibration only, and a uniform root calibration with another two additional uniform internal calibrations—three uniform calibrations (Fig. 2).

2.2.2 RelTime analysis under relative rate framework

All RelTime analyses were carried out in MEGA-CC software that was prototyped in MEGA X (Kumar *et al.*, 2012, 2018). For ensuring a direct comparison, we first used the ML method to estimate branch lengths under the GTR + Γ model with the correct topology by using RAxML (Stamatakis, 2014). Then, each phylogeny with branch length was used to infer a timetree by applying the RelTime method. Timetrees were computed using two different calibration strategies applying uniform constraints: a root calibration only and three calibrations (Fig. 2).

2.3 Measurements of accuracies

All comparisons of TEs from simulated data and correct times involved normalized values that were obtained by dividing given node time by the sum of node times in the tree. This procedure avoids normalization biases that may be caused by using any one node as an anchor. The percent difference in TEs (ΔTE) is the difference between the estimated time and the true time divided by the true time and multiplied by 100. For comparison of model-match and -mismatch cases, ΔTE is the difference between the estimated and the GTR data TE divided by the GTR data TE and multiplied by 100.

2.4 Measurements of coverage probability

We calculated the coverage probability of each node for each dataset. The coverage probability of a node was the proportion of datasets in which the Bayesian highest posterior density intervals (HPDs) or RelTime confidence intervals (CIs) of that node contained the true time. This was done for all datasets with only one root calibration and with three calibrations. True times were normalized to the sum of true times, and lower and upper bounds of HPDs (or CIs) were normalized to the sum of estimated node times.

2.5 Measurements of branch length linearity

Because the same phylogeny was used to simulate GTR, NR and NS datasets, their inferred branch lengths were directly comparable. For each phylogeny simulated under AR or IR rate scenario, we compared the branch lengths inferred using the GTR + Γ model in RAxML for GTR and NR data and GTR and NS data. The coefficient of determination of linear regression through the origin (R^2) was used to determine the linearity of inferred branch lengths. A higher R^2 value indicated a better linear relationship.

3 Results and discussion

We first present results from analyses without applying internal calibrations, which is essential to learn about the intrinsic time structure in the data because calibrations generally impose strong constraints on node ages. We then show results from analyses using a few internal calibrations, which allow us to examine whether the use of multiple calibrations reduce the bias in times obtained by using the GTR + Γ model.

3.1 Impact of violating the time-reversibility assumption

The use of the GTR + Γ model is expected to cause bias in estimating divergence times for the NR datasets because the analysis assumed the time-reversibility of the substitution process (model-mismatch). This bias is explored by comparing the TEs inferred by using the GTR + Γ model for GTR and for NR datasets simulated using the same phylogeny (50 replicates with autocorrelated and 50 with independent branch rates).

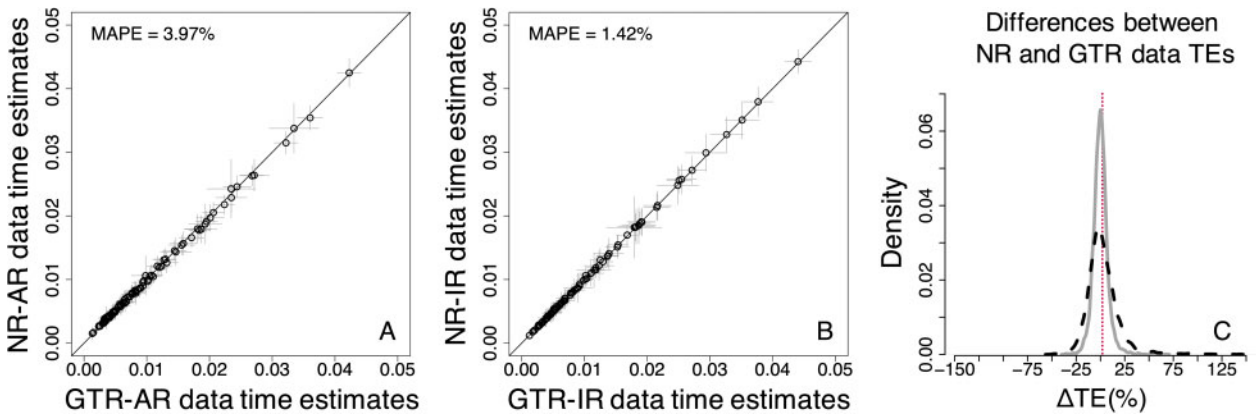

Fig. 4. Comparison of Bayesian TEs obtained by using the GTR + Γ model for analyzing GTR (model-match) and NR (model-mismatch) datasets simulated (**A**) with rate autocorrelation, AR, and (**B**) without rate autocorrelation, IR. Each data point represents the average of normalized times from 50 simulations (±1 SD—gray line). The MAPE is shown in the upper left corner of these panels. The black 1:1 line shows the trend if the estimates were equal. (**C**) Distributions of the normalized differences between GTR and NR data TEs for AR (black-dashed curve) and IR (gray curve) branch rates. For visual clarity, the distribution in (C) was truncated, removing a few outliers

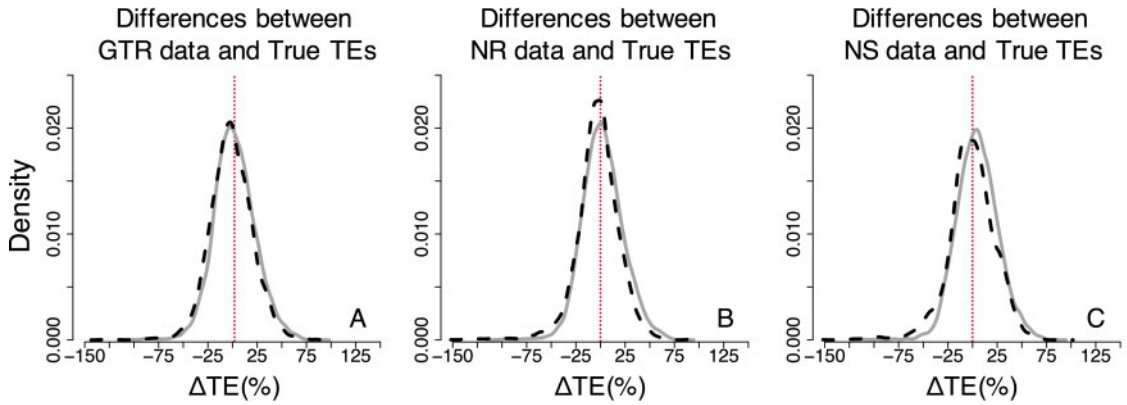

Fig. 5. Distributions of the normalized differences between estimated and true node times for GTR, NR, and NS datasets—Bayesian approach (root calibration only). Comparisons of AR (black-dashed curve) and IR (gray curve) performance for (**A**) GTR, (**B**) NR and (**C**) NS datasets. For visual clarity, the distribution in (A–C) was truncated, removing a few outliers

In Bayesian analyses, these estimates showed a high similarity (Fig. 4A and B). The mean absolute percent error (MAPE) was 3.97% when rates were autocorrelated and 1.42% when rates were independent among branches, i.e. the bias is surprisingly small for AR as well as IR datasets. The differences between TEs estimated from GTR and NR datasets (ΔTEs) showed a slightly higher dispersion for AR datasets as compared with the IR datasets (Fig. 4C).

Figure 5 compares the dispersions of ΔTEs between the true times and the times obtained when the models matched (Fig. 5A) and the times obtained when there was a model-mismatch for the NR datasets (Fig. 5B). Both of these comparisons show very similar shapes and trends for AR as well as IR datasets. Therefore, the violation of the assumption of the time-reversibility of the Markov process does not seem to have a strong biasing impact in the Bayesian analyses.

Similar results were observed for RelTime estimates. Again, the bias was small (Fig. 6), with MAPE equal to 4.74% when rates were autocorrelated and 1.44% when rates were independent. Moreover, ΔTEs between TEs for NR and GTR datasets showed a low and similar dispersion for both IR and AR datasets (Fig. 6C). However, the RelTime estimates showed greater noise (SDs) as compared with the Bayesian node times. It is not surprising because we assumed correct priors (e.g. tree prior and evolutionary rate model) in Bayesian analyses. In contrast, the RelTime method does not require such prior knowledge in estimating divergence times.

Tao *et al.* (2020) showed that one reason for the robustness of relaxed clock methods to model mis-specification was that the estimates of branch lengths under simple and complex models are often linearly related. In the relative rate framework underlying the RelTime method, divergence times are a function of the ratio of linear combinations of branch lengths. So, we examined the relationship of inferred branch lengths for GTR and NR phylogenies in which a GTR model was used for the inference. We found an excellent linear relationship for an AR and an IR dataset (Fig. 7A and B, $R^2 = 0.97$ and 0.98 for AR and IR dataset, respectively), which is similar to that reported in Tao *et al.* (2020). The pattern of linearity of branch lengths was observed across the majority of AR and IR datasets (Fig. 7C). These results indicated similar relative branch lengths were produced when the assumed model matched or did not match the actual evolutionary process, and therefore, similar divergence TEs. This linear relationship provides a fundamental reason for the results seen for RelTime (Fig. 6) and Bayesian (Figs 4 and 5) methods.

3.2 Impact of the violation of the non-stationarity assumption

Next, we explored the bias of TEs caused by the use of the GTR + Γ model to analyze data in which the substitution process was not stationary over time or among lineages (NS datasets). The results of Bayesian analyses showed that the bias on TEs caused by the use of the GTR + Γ model is again small (Fig. 8A and B). MAPE between TEs estimated from NS and GTR data was 9.92% when branch rates were autocorrelated and 7.38% when branch rates were

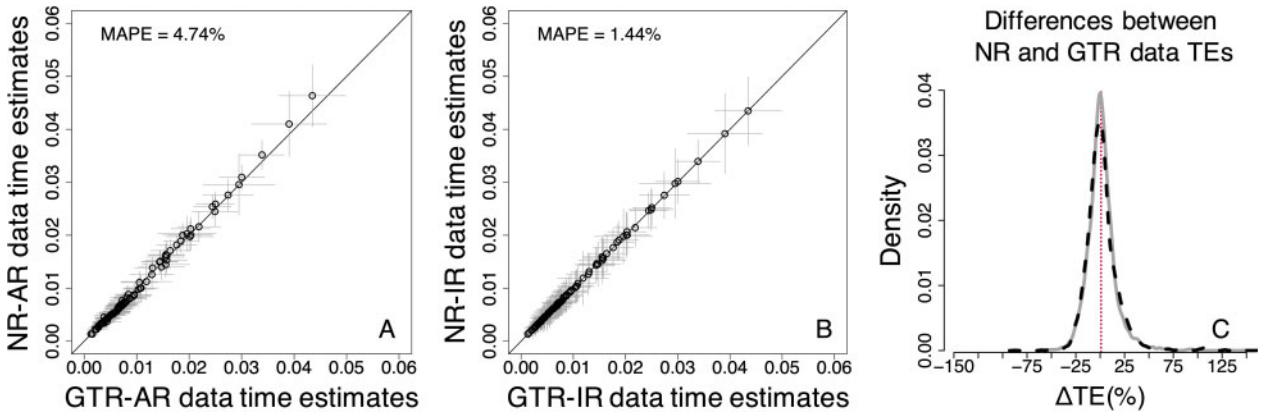

Fig. 6. Comparison of RelTime estimates obtained by using the GTR + Γ model for GTR (model-match) and NR (model-mismatch) datasets simulated (**A**) with rate autocorrelation, AR, and (**B**) without rate autocorrelation, IR. Each data point represents the average of normalized times from 50 simulations (±1 SD—gray line). The MAPE is shown in the upper left corner of these panels. The black 1:1 line shows the trend if the estimates were equal. (**C**) Distributions of the normalized differences between GTR and NR TEs for AR (black-dashed curve) and IR (gray curve) datasets. For visual clarity, distribution in (C) was truncated, removing a few outliers

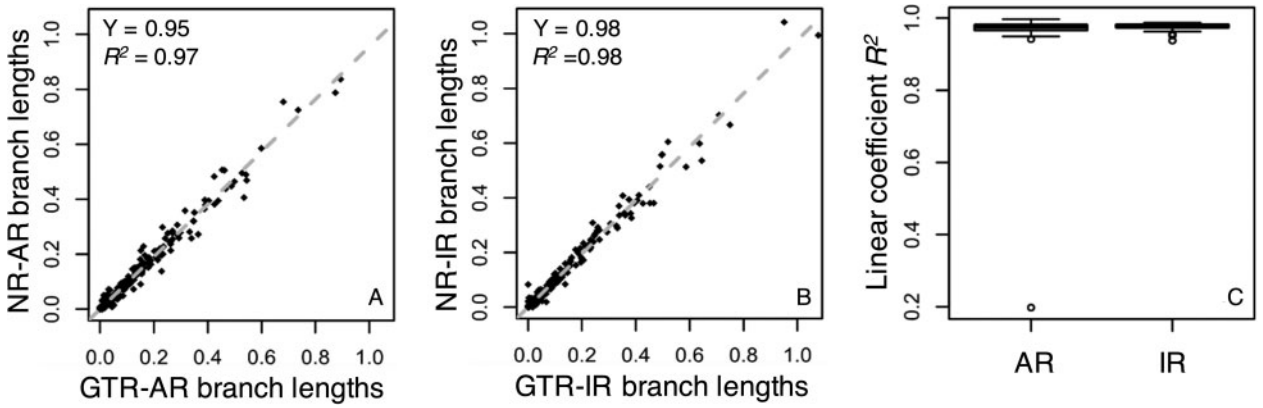

Fig. 7. Branch length comparisons for GTR and NR datasets. Branch lengths were inferred by using the GTR + Γ model for (**A**) an AR dataset and (**B**) an IR dataset simulated under the GTR model (x-axis, model-match case) and the NR model (y-axis, model-mismatch case). They all show good linear relationships. The gray-dashed line is the best-fit linear regression through the origin. The slope (Y) and coefficient of determination (R^2) are shown. (**C**) The dispersion of the linear trends of branch lengths. Boxes show the variation of the coefficient of determination of the linear regression (through the origin, R^2) between branch lengths inferred using the GTR + Γ model for 50 GTR and 50 NR datasets simulated under AR and IR scenarios

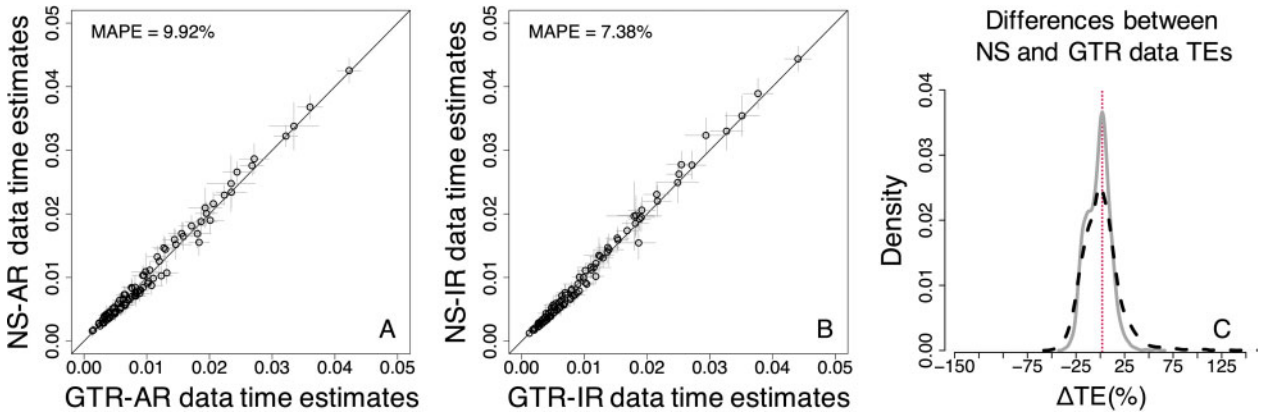

Fig. 8. Comparison of Bayesian TEs obtained by using the GTR + Γ model for GTR (model-match) and NS (model-mismatch) datasets simulated (**A**) with rate autocorrelation, AR, and (**B**) without rate autocorrelation, IR. Each data point represents the average of normalized times from 50 simulations (±1 SD—gray line). The MAPE is shown in the upper left corner of these panels. The black 1:1 line shows the trend if the estimates were equal. (**C**) Distributions of the normalized differences between GTR and NS data TEs for AR (black-dashed curve) and IR (gray curve) datasets. For visual clarity, the distribution in (C) was truncated, removing a few outliers

independent. The dispersion of node TEs was higher for AR datasets than IR datasets (Fig. 8C). However, overall, the bias is greater for NS datasets than the NR datasets (compare Figs 4 and 8), with consistent problems observed for some TEs. In particular, TEs from the lineages in which the base composition and rate matrix changed were misestimated. Furthermore, in comparison to NR datasets (Fig. 4C), we found that ΔTEs had a considerably larger dispersion for NS data (Fig. 8C).

As for the accuracy of Bayesian times, we found that the distributions of ΔTEs between the estimated and true TEs showed slightly

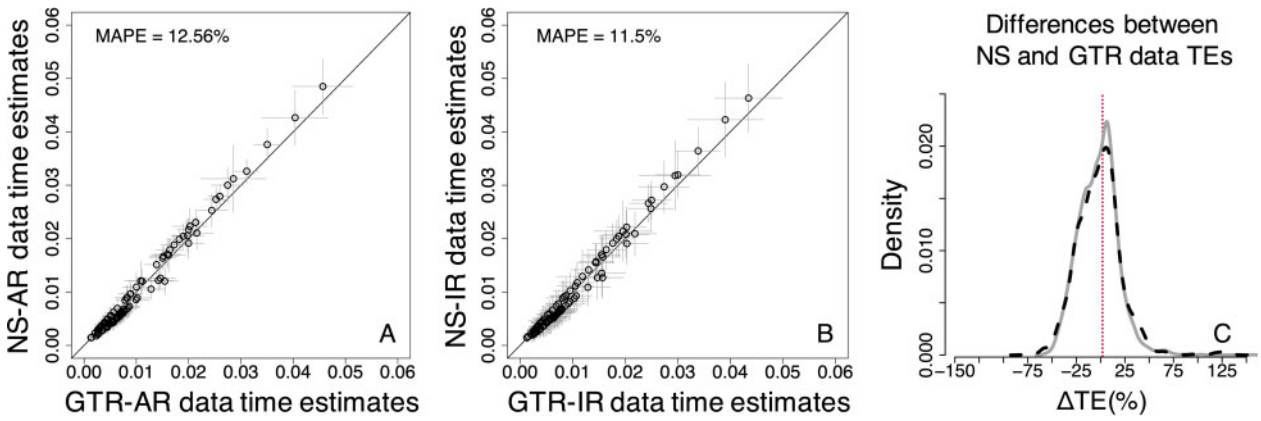

Fig. 9. Comparison of TEs obtained by using the GTR + Γ model for GTR and NS datasets—RelTime approach (root calibration only). (**A**) AR datasets. (**B**) IR datasets. Each data point represents the average of normalized times from 50 simulations (±1 SD—gray line). The MAPE is shown in the upper left portion of each plot. The black line represents equality between estimates. (**C**) Distributions of the normalized differences between GTR and NS TEs for AR (black-dashed curve) and IR (gray curve) datasets. For visual clarity, the distribution in (C) was truncated, removing a few outliers

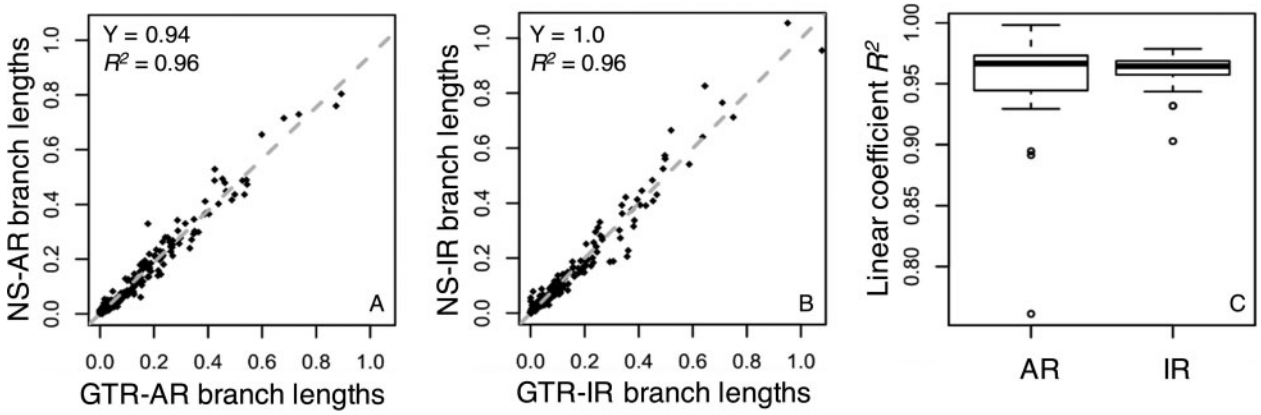

Fig. 10. Branch lengths comparisons between GTR and NS data. Branch lengths inferred using the GTR + Γ model for (**A**) an AR dataset and (**B**) an IR dataset simulated under the GTR model (x-axis, model case) and the NS model (y-axis, model case) show a good linear relationship. The gray-dashed line is the best-fit linear regression through the origin. The slope (Y) and coefficient of determination (R^2) are shown. (**C**) The dispersion of the linear trends of branch lengths. Boxes show the variation of the coefficient of determination of the linear regression (through the origin, R^2) between branch lengths inferred using the GTR + Γ model for 50 GTR and 50 NS datasets simulated under AR and IR scenarios

larger dispersion for NS than for NR and GTR datasets under AR and IR models (Fig. 5A versus C). Therefore, the violation of the assumption of stationary of the substitution process is likely to have a somewhat considerable biasing impact in the Bayesian time inference.

The results of RelTime analyses showed a pattern that resembled the Bayesian methods, as similar TEs were obtained when the GTR + Γ model was used on GTR datasets (model-match) and when the GTR + Γ model was used on NS datasets (model-mismatch, Fig. 9). However, the overall bias on TEs caused by the use of the GTR + Γ model was slightly larger for RelTime. The MAPE of TEs inferred from NS data was, on average, 12.56% for AR datasets (Fig. 9A) and 11.5% for IR datasets (Fig. 9B). In comparison to NR data (Fig. 5C), the ΔTEs between TEs estimated from NS data and GTR data displayed a larger dispersion (Fig. 9C).

We also examined the relationship of inferred branch lengths for GTR and NS phylogenies, in which the assumption of stationarity was violated, and a GTR + Γ model was used for branch lengths inference. There was a very high correlation (Fig. 10A and B, R^2 = 0.96 and 0.96 for AR and IR dataset, respectively). However, the linear trend was weaker as compared with that for the NR data and displayed higher variation (compare Figs 7 and 10). This slightly weaker linear relationship appears to be the reason for the results that higher biased estimates were produced by the use of the GTR + Γ model for NS data than NR data. This is particularly interesting, because figures 8A, 9A and 10A show very similar trends, indicating that the bias in estimating relative branch lengths is recapitulated in the TEs produced by both Bayesian and RelTime approaches.

3.3 Improvements offered by the use of calibrations

We also estimated divergence times using two internal calibrations shown in Figure 2 to examine whether the use of multiple calibrations constrained the TEs and reduced the possible error caused by the use of GTR + Γ. As expected, the biased estimates improved when we use calibrations strategically positioned on the nodes that experienced a change in the substitution model and base composition in the phylogeny.

In comparison to analyses with only the root calibration, the MAPE for the NS data was reduced to 8.08% and 6.33% for AR and IR datasets, respectively. For the NR case, the MAPE was reduced to 3.65% for AR datasets, and it remained almost identical (1.46%) for IR datasets (Fig. 11A, B, D and E). Furthermore, ΔTEs from NR and NS data showed a higher correspondence to those from GTR data (Fig. 11C and F) than in the analyses using a root calibration only (Figs 4C and 8C).

The accuracy of Bayesian TEs remained similar when internal calibrations were used (Fig. 12A–C), although ΔTEs displayed slightly lower dispersion. Overall, ΔTEs showed a high correspondence to those in the analyses using a root calibration only (Fig. 5A–

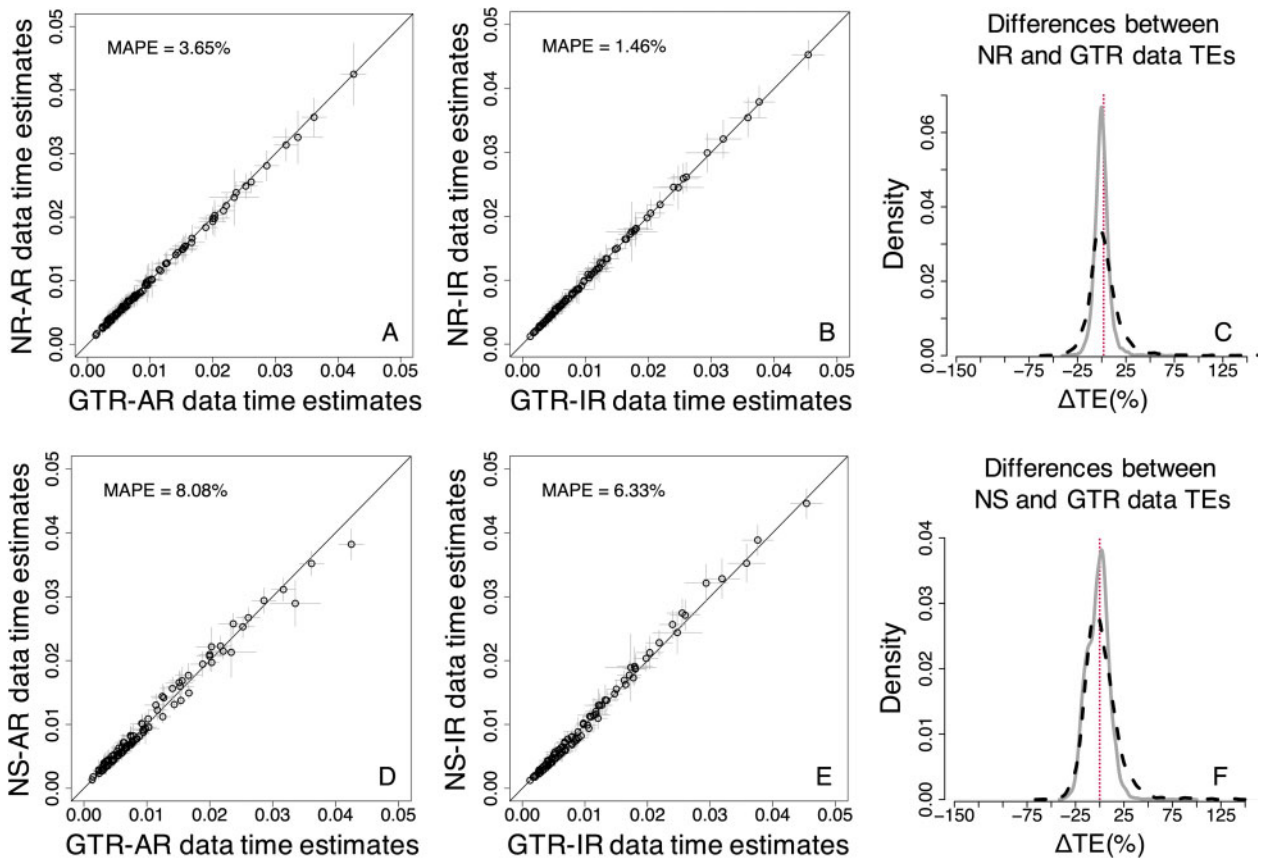

Fig. 11. Comparison of TEs obtained by using the GTR + Γ model for GTR, NR, and NS datasets—Bayesian approach (three calibrations). (**A–C**) NR datasets. (**D–F**) NS datasets. (A, B, D, and E) Each data point represents the average of normalized times from 50 simulations (±1 SD—gray line), generated using. The MAPE is shown in the upper left portion of each plot. The black line represents equality between estimates. Distributions of the normalized differences (C) between GTR and NR TEs, and (F) between GTR and NS TEs for AR (black-dashed curve) and IR (gray curve) datasets. For visual clarity, the distribution in (C) and (F) was truncated, removing a few outliers

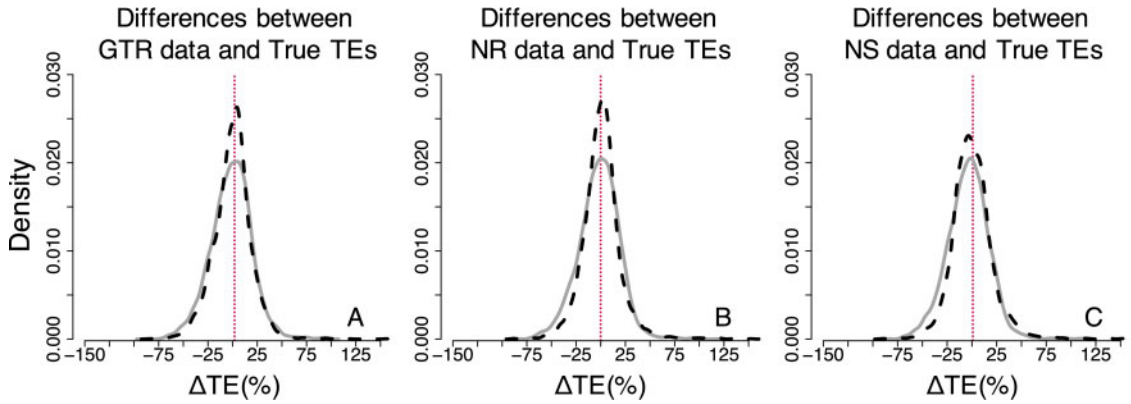

Fig. 12. Distributions of the normalized differences between estimated and true times on nodes for GTR, NR, and NS datasets—Bayesian approach (three calibrations). Comparisons of AR (black-dashed curve) and IR (gray curve) performance for (**A**) GTR, (**B**) NR, and (**C**) NS datasets. For visual clarity, the distribution in (A–C) was truncated, removing a few outliers

C), excluding NS-AR data, which displayed a significantly reduced dispersion.

3.4 Effect of calibrations in the RelTime estimates

When compared with the Bayesian method, we found that the bias on TEs was considerably reduced when RelTime was used, particularly the bias on NR data TEs (Fig. 13A, B, D, and E). When internal calibrations were applied, RelTime TEs inferred from GTR data showed higher similarity to those obtained by using NR and NS data (Fig. 13). The MAPE for NR data was reduced to 1.41% when rates were autocorrelated, and to 1.21% when rates were independent, the MAPE in the NS case was even further reduced to 5.0% and 5.2% for AR and IR data, respectively. Furthermore, ΔTEs from NR and NS data showed a high correspondence to those from GTR data (Fig. 13C and F). These results indicate that the use of internal calibrations can correct bias caused by the use of the GTR + Γ model in analyses of sequence alignments evolved under NS and NR substitution processes. The accuracy of RelTime TEs became higher when internal calibrations were used, as RelTime ΔTEs displayed a

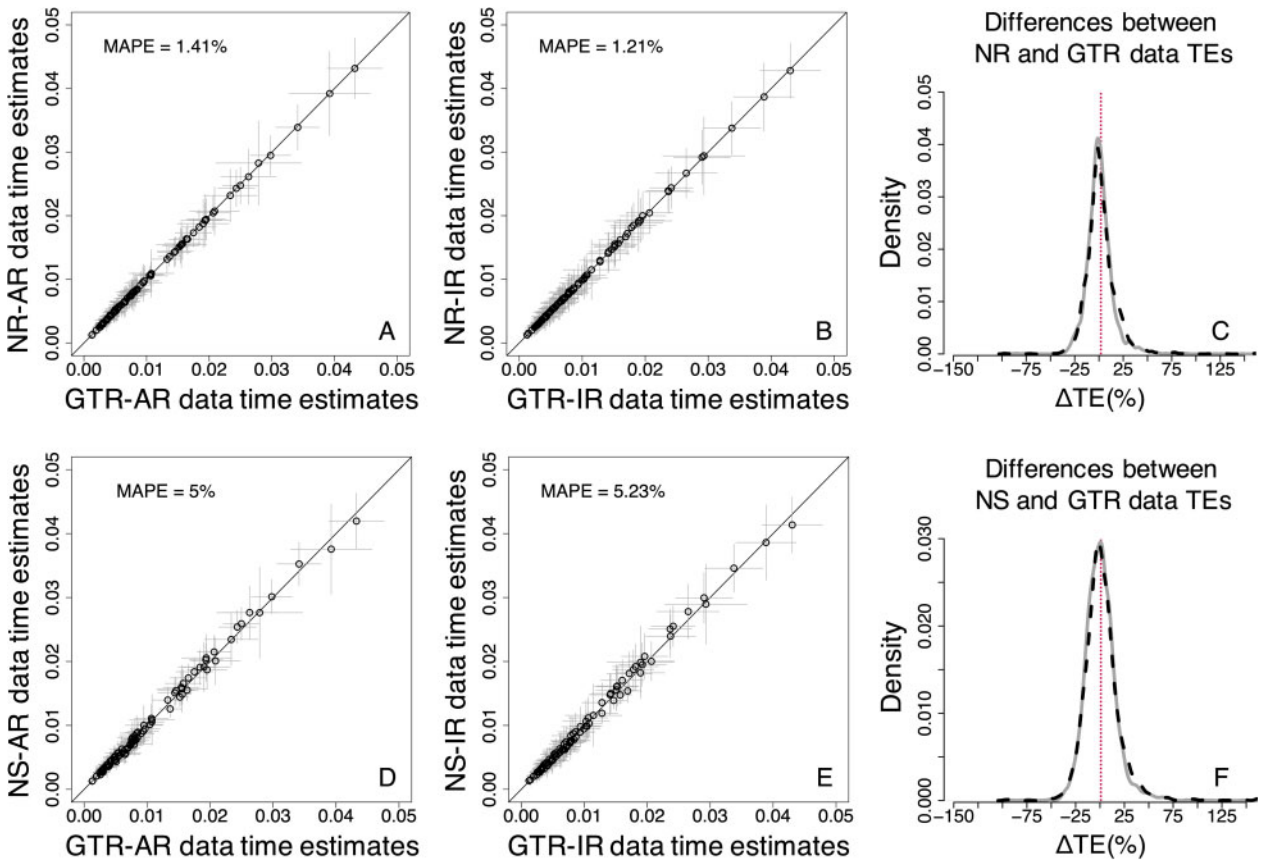

Fig. 13. Comparison of TEs obtained by using the GTR + Γ model for GTR, NR and NS datasets—RelTime approach (three calibrations). (**A–C**) NR datasets. (**D–F**) NS datasets. (A, B, D, and E) Each data point represents the average of normalized times from 50 simulations (±1 SD—gray line). The MAPE is shown in the upper left portion of each plot. The black line represents equality between estimates. Distributions of the normalized differences (C) between GTR and NR TEs, and (F) between GTR and NS TEs for AR (black-dashed curve) and IR (gray curve) datasets. For visual clarity, the distribution in (C) and (F) was truncated, removing a few outliers

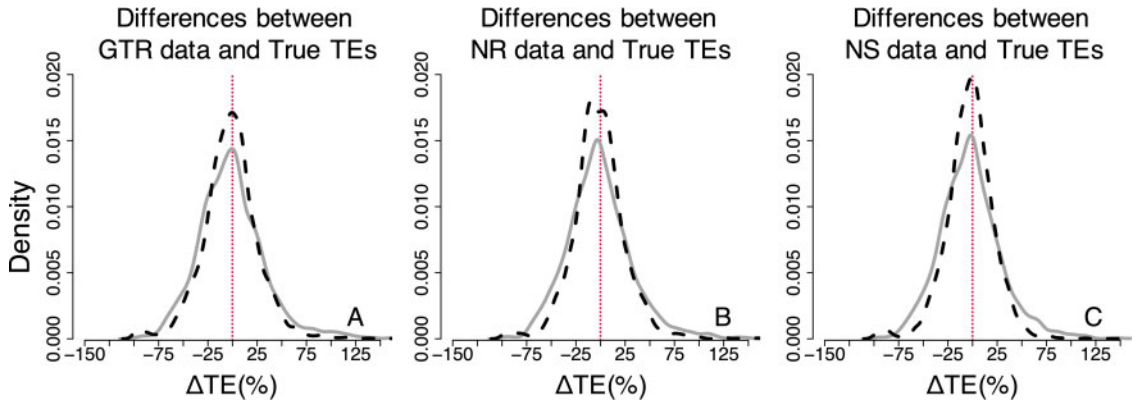

Fig. 14. Distributions of the normalized differences between estimated and true times on nodes for GTR, NR, and NS datasets—RelTime approach (three calibrations). Comparisons of AR (black-dashed curve) and IR (gray curve) performance for (**A**) GTR, (**B**) NR, and (**C**) NS datasets. For visual clarity, the distribution in (A–C) was truncated, removing a few outliers

reduced dispersion, particularly for NS data (Fig. 14). Although the bias caused by the use of the GTR + Γ model to analyze NR and NS data was reduced, RelTime TEs still displayed a larger dispersion.

3.5 Evaluation of coverage probability in Bayesian and RelTime approaches

Next, we estimated coverage probabilities that quantified how often the actual node times were contained in the 95% HPDs for Bayesian analyses or 95% CIs for RelTime analyses. We found that Bayesian HPDs obtained using the GTR + Γ model contained the actual node times for the majority of nodes for GTR data (Fig. 15A, median coverage probability = 0.96). This was expected because the used substitution model matched the actual evolutionary process. When the data were simulated under NR and NS models, we also found that HPDs obtained using the GTR + Γ model often included the true times (Fig. 15A, median coverage probability = 0.94 and 0.92 for NR and NS datasets, respectively). More interestingly,

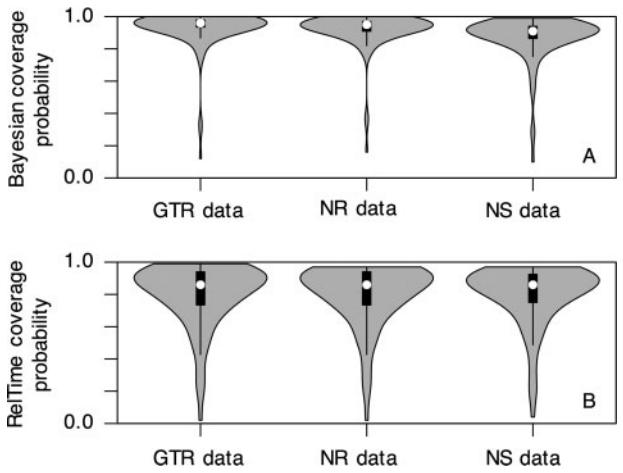

Fig. 15. Distributions of coverage probabilities of all the nodes for GTR, NR, and NS datasets. Coverage probability of (**A**) Bayesian HPDs and (**B**) RelTime CIs for each scenario is calculated using results of 50 IR and 50 AR simulated datasets obtained using the GTR + Γ model and three calibrations. White dot represents the median value

distributions of coverage probabilities across all the nodes for NR and NS datasets (model-mismatch cases) were very similar to the one for GTR datasets (model-match case, Fig. 15A).

Similar distributions of coverage probabilities between matched and mismatched cases were also observed in RelTime analyses (Fig. 15B). These results indicate that the use of the GTR + Γ model on datasets that evolved under much more complex processes is unlikely to impact the estimation of HPDs or CIs, resulting in a consistent conclusion for biological hypothesis testing. However, the overall coverage probability of RelTime CIs was slightly lower than the Bayesian HPDs. It may be because we assumed that correct priors (e.g. tree prior and evolutionary rate model) were known and used them in the Bayesian method, which maximized its performance. In contrast, the RelTime method does not use any such prior knowledge in estimating divergence times.

4 Conclusion

In this study, we have analyzed the bias on divergence time estimation caused by the use of the GTR + Γ model to analyze sequence alignments that violate the basic assumptions in phylogenetic analyses: time-reversibility and stationarity of substitution processes. Our results reveal that violating the time-reversibility assumption may only have a limited effect on the accuracy of divergence TEs. In contrast, the use of sequences with considerable variation in base compositions among sequences, in which the assumption of model stationarity is violated, has a greater effect on the accuracy and precision of divergence TEs.

Fortunately, we may expect an improvement of accuracy and precision if we use reliable calibrations that are strategically positioned on the phylogeny. Comparable trends were observed for node TEs between RelTime and Bayesian analyses, although overall RelTime estimates showed a larger dispersion and higher error.

Our results are mostly consistent with the conclusion of Tao *et al.* (2020) that show that the complexity of the substitution model has only a modest impact on divergence TEs. The primary reason for the good performance of the GTR + Γ for analyzing sequences that evolved under NS and NR processes is the high linearity between the branch lengths produced by the mismatched model with those from the correct model. The similar relative branch lengths can be transformed to similar divergence times estimates because the divergence times are a function of the ratio of linear combinations of branch lengths. The present results show that using the GTR + Γ model to analyze sequence alignments, whose basic assumptions are violated, may be sufficient in a majority of time inference tasks. Nevertheless, accounting for time-irreversibility and non-stationarity is still an important aspect of the determination of substitution rates and other phylogenetic inference.

Acknowledgements

We thank Dr. Marcos Caraballo for technical support.

Funding

This work was supported by grants from NSF [DBI 1356548], National Institutes of Health [R01GM126567-03] and National Aeronautics and Space Administration [NASA, NNX16AJ30G] to S.K.

Conflict of Interest: none declared.

References

Arenas,M. (2015) Trends in substitution models of molecular evolution. *Front. Genet.*, **6**, 319.

Blanquart,S. and Lartillot,N. (2006) A Bayesian compound stochastic process for modeling non-stationary and non-homogeneous sequence evolution. *Mol. Biol. Evol.*, **23**, 2058–2071.

Blanquart,S. and Lartillot,N. (2008) A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.*, **25**, 842–858.

dos Reis,M. and Yang,Z. (2011) Approximate likelihood calculation on a phylogeny for Bayesian Estimation of Divergence Times. *Mol. Biol. Evol.*, **28**, 2161–2172.

Fletcher,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**, 1879–1888.

Foster,P.G. (2004) Modeling compositional heterogeneity. *Syst. Biol.*, **53**, 485–495.

Galen,S.C. *et al.* (2018) The polyphyly of *Plasmodium*: comprehensive phylogenetic analyses of the malaria parasites (Order Haemosporida) reveal widespread taxonomic conflict. *R. Soc. Open Sci.*, **5**, 171780.

Galtier,N. and Gouy,M. (1998) Inferring pattern and process: maximum-likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, **15**, 871–879.

Galtier,N. and Gouy,M. (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. USA*, **92**, 11317–11321.

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

Hedges,S.B. and Kumar, S. *et al.* (eds) (2009) Discovering the timetree of life. In: *The Timetree of Life*, Oxford University Press, New York, pp. 3–18.

Huang,H. *et al.* (2010) Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.*, **59**, 573–583.

Jayaswal,V. *et al.* (2011) Two stationary non-homogeneous Markov models of nucleotide sequence evolution. *Syst. Biol.*, **60**, 74–86.

Kimura,M. (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Kumar,S. *et al.* (2012) MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*, **28**, 2685–2686.

Kumar,S. *et al.* (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.*, **35**, 1547–1549.

Kumar,S. and Gadagkar,S.R. (2001) Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics*, **158**, 1321–1327.

Lockhart,P.J. *et al.* (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, **11**, 605–612.

Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford.

Philippe,H. *et al.* (2017) Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.*, **2017**, 1–25.

Rannala,B. and Yang,Z.(2007) Inferring speciation times under an episodic molecular clock. *Syst. Biol.*, **56**, 453–466.

Rosenberg,M.S. and Kumar S. (2003) Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. *Mol. Biol. Evol.*, **20**, 610–621.

Singh,N.D. *et al.* (2009) Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. *Mol. Biol. Evol.*, **26**, 1591–1605.

Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Tamura,K. *et al.* (2012) Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. USA*, **109**, 19333–19338.

Tamura,K. and Kumar,S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.*, **19**, 1727–1736.

Tamura,K. and Nei,M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, **10**, 512–526.

Tao,Q. *et al.* (2019) A Machine Learning Method for Detecting Autocorrelation of Evolutionary Rates in Large Phylogenies. Molecular Biology and Evolution, **36**, 811–824.

Tao,Q. et al. (2020) Relative efficiencies of simple and complex substitution models in estimating divergence times in phylogenomics. *Evol. Biol.*, **37**, 1819–1831.

Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Am. Math. Soc. Lect. Math. Life Sci.*, **17**, 57–86.

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

Yang,Z. (1994) Estimating the Pattern of Nucleotide Substitution. *J. Mol. Evol.*, **39**, 105–111.

Yang,Z. (2014) *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford.

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

Yang,Z. and 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.