Syst. Biol. 68(1):145–156, 2019

DOI:10.1093/sysbio/syy047
Advance Access publication June 25, 2018

Optimal Rates for Phylogenetic Inference and Experimental Design in the Era of Genome-Scale Data Sets

ALEX DORNBURG[1,*], ZHUO SU[2], AND JEFFREY P. TOWNSEND[2,3,4]

[1]*North Carolina Museum of Natural Sciences, Raleigh, 1671 Goldstar Drive, NC 27601, USA;*
[2]*Department of Ecology and Evolutionary Biology, Yale University, New Haven, 165 Prospect Street, CT 06525, USA;*
[3]*Department of Biostatistics, Yale University, New Haven, 60 College Street, CT 06510, USA; and*
[4]*Program in Computational Biology and Bioinformatics, Yale University, New Haven, 300 George Street, CT 06511, USA*
[*]*Correspondence to be sent to: North Carolina Museum of Natural Sciences, Raleigh, NC 27601, USA;*
E-mail: alex.dornburg@naturalsciences.org.

Received 13 February 2018; reviews returned 4 June 2018; accepted 13 June 2018
Associate Editor: Rachel Mueller

Abstract.—With the rise of genome-scale data sets, there has been a call for increased data scrutiny and careful selection of loci that are appropriate to use in an attempt to resolve a phylogenetic problem. Such loci should maximize phylogenetic information content while minimizing the risk of homoplasy. Theory posits the existence of characters that evolve at an optimum rate, and efforts to determine optimal rates of inference have been a cornerstone of phylogenetic experimental design for over two decades. However, both theoretical and empirical investigations of optimal rates have varied dramatically in their conclusions: spanning no relationship to a tight relationship between the rate of change and phylogenetic utility. Herein, we synthesize these apparently contradictory views, demonstrating both empirical and theoretical conditions under which each is correct. We find that optimal rates of characters—not genes—are generally robust to most experimental design decisions. Moreover, consideration of site rate heterogeneity within a given locus is critical to accurate predictions of utility. Factors such as taxon sampling or the targeted number of characters providing support for a topology are additionally critical to the predictions of phylogenetic utility based on the rate of character change. Further, optimality of rates and predictions of phylogenetic utility are not equivalent, demonstrating the need for further development of comprehensive theory of phylogenetic experimental design. [Divergence time; GC bias; homoplasy; incongruence; information content; internode length; optimal rates; phylogenetic informativeness; phylogenetic theory; phylogenetic utility; phylogenomics; signal and noise; subtending branch length; state space; taxon and character sampling.]

The rapid proliferation of sequencing technology over the past decade has enabled historically unparalleled progress in our efforts to reconstruct of the Tree of Life (Near et al. 2012b; Wiens et al. 2012; Prum et al. 2015; Ren et al. 2016). Correspondingly, phylogenetic trees have assumed a central role in studies that span topics from conservation (Purvis et al. 20005; Dornburg et al. 2017a; Pollock et al. 2017) to the modeling of cancer and infectious diseases (Zhao et al. 2016; Reiter et al. 2017). However, despite data sets spanning thousands if not millions of characters, some phylogenetic problems continue to defy resolution. Faced with a deluge of sequence data, the question of how to select data most appropriate for a given phylogenetic problem has become a major topic of interest (Salichos and Rokas 2013; Pisani et al. 2015; Shen et al. 2017). This question is not new to phylogenomics and has been a driving question in the theory of phylogenetic experimental design for over two decades (Graybeal 1993; Xia et al. 2003).

It is well established that the interaction between time and the rate of character evolution can be used to predict the probability of convergence in states that do not reflect shared evolutionary history (Graybeal 1993; Xia et al. 2003). Given the expected relationship between time, character evolution, and inference, theory predicts the existence of an optimal rate of character evolution that maximizes the generation of phylogenetic information, while minimizing the accumulation of homoplasious characters (Goldman 1998). Efforts to determine optimal rates for phylogenetic inference continue to the present day (Klopfstein et al. 2017; Steel and Leuenberger 2017), as does the development of approaches that assess whether a set of characters are evolving at rates appropriate for resolving a given phylogenetic problem (Susko and Roger 2012; Townsend et al. 2012; Su and Townsend 2015). However, several studies have failed to recover a relationship between the mean rate of change and the phylogenetic utility of a locus (Aguileta et al. 2008). Further, others have argued that focusing on rate alone ignores other interacting terms such as tree structure, complex patterns of character evolution, or taxon sampling (Heath et al. 2008a; Townsend et al. 2012; Su and Townsend 2015).

In an effort to reconcile apparently contradictory views of phylogenetic experimental design, we demonstrate that all of these views are essentially correct. Determination of the optimal rates for phylogenetic inference depends on each of these factors, which exhibit a range of dependency and influence.

CHARACTERS AS UNITS FOR OPTIMAL RATES

The evolutionary rate relevant to phylogenetic utility is the rate of evolution of the homologous character—not the rate of the gene or locus

A core criterion of locus selection for accurate inference has been the evolutionary rate of a given gene or

locus (Blouin et al. 1998; Moreira and Philippe 2000; Xia et al. 2003; Nosenko et al. 2013). However, an investigation of fungal genomes found no relationship between the rate at which a locus was evolving and phylogenetic utility (Aguileta et al. 2008). Is the assumption of the predictive utility of rates wrong? We argue that it is not. Instead, we view the assumption that genes or loci can be characterized by a single evolutionary rate as, perhaps, the greatest barrier to understanding of the impact of rates of molecular evolution on phylogenetic utility.

It has become standard practice to incorporate rate variation across sites into molecular evolutionary models used for phylogenetic inference (Yang 1996; Sullivan et al. 1999). However, this practice has not yet taken root in phylogenetic experimental design, and loci are still often contrasted by pairwise distances or mean rates (Aguileta et al. 2008; Makowsky et al. 2010; Lanier et al. 2014). While determining or comparing mean rates is certainly useful for certain empirical problems (Braasch et al. 2016), characterizing loci by a single rate can oversimplify and, in the worst cases, wholly mislead phylogenetic experimental design. There may be some loci in which all sites evolve at similar rates in a clock-like manner, in which case there will likely be a strong correlation between the mean rate of the locus and its utility for a phylogenetic problem. In contrast, numerous loci are comprised of sites that are highly heterogeneous in their substitution rates. For these loci, attempting to predict phylogenetic utility using the mean rate across all sites can mask a large number of suboptimal sites—in the worst case scenarios, reflecting a rate of change that is not actually occurring to any individual character the data set. As such, failure to account for intra-locus site rate variation can underlie broad findings in which there is no correlation between the mean rate of evolution of the locus and its informativeness (Aguileta et al. 2008).

The utility of a character for phylogenetic inference can be quantified by comparing the expected phylogenetic signal supporting a correct resolution (R) of a quartet to the amount of noise expected to accumulate in support of the most supported incorrect quartet (Townsend et al. 2012; Su et al. 2014; Su and Townsend 2015). By calculating the point at which this difference is largest (by the old calculus trick of setting the partial derivative of the function of utility to zero), an optimal rate is defined (Appendix). Using this framework, we can explore the impact of among-site rate variation on inference. Consider a hypothetical phylogenetic quartet and a set of characters evolving close to the optimal rate. This data set would be predicted to be of high experimental design utility under both a mean rate criterion and the framework of Su and Townsend (2015). However, as the variance in rates begins to increase, there is a clear loss in utility predicted by the equations of Su and Townsend (Fig. 1).

The issues that arise from failing to account for site-rate variation in phylogenetic experimental design should be intuitive and reminiscent of classical phylogenetic problems such as the estimation of divergence times using a strict molecular clock. Both among-site rate variation within loci as well as shifts in life history that correlate to changes in molecular rates between taxa are commonly observed across the tree of life (Bromham et al. 2015; Dornburg et al. 2012; Berv and Field 2017; Minias et al. 2017; Gan et al. 2018). As such, we argue that use of mean rate alone can drive erroneous expectations of phylogenetic utility and promote poor experimental design.

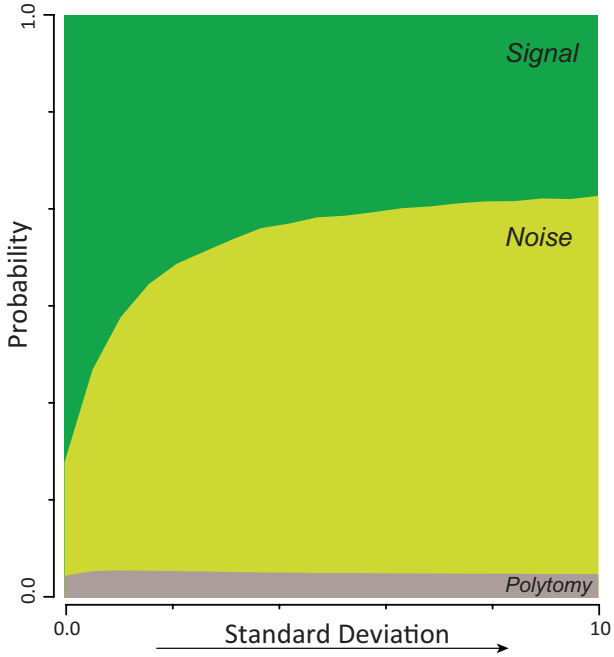

FIGURE 1. The effect of increasing variance of site rates on phylogenetic utility: probability of noise compared with signal increases as the standard deviation of a unimodal distribution of rates of evolution increases (where the mean of the distribution is set at the optimal rate $\hat{\lambda}$).

Optimal Rates, Divergence Time, and Internode Length

The utility of a character evolving at a rate depends not only on the depth of the phylogenetic inference, but also on the length of the internode to be resolved.

It has long been recognized that the time scale of a given phylogenetic problem will impact which sites are of utility for topological resolution (Graybeal 1993; Xia et al. 2003; Mueller 2006). In addition to the rate of character evolution and the passage of time, it has also become increasingly clear that the time between divergence events, or internode distances, can strongly impact the utility of a set of characters for inference. As demonstrated by Townsend et al. (2012) and Susko and Roger (2012), rates of evolution that are highly informative for some internodes within a timeslice of a tree might not contribute to resolution or—in the worst case—might positively mislead inference. The complexity of calculating optimal rates for specific internode distances was recently highlighted

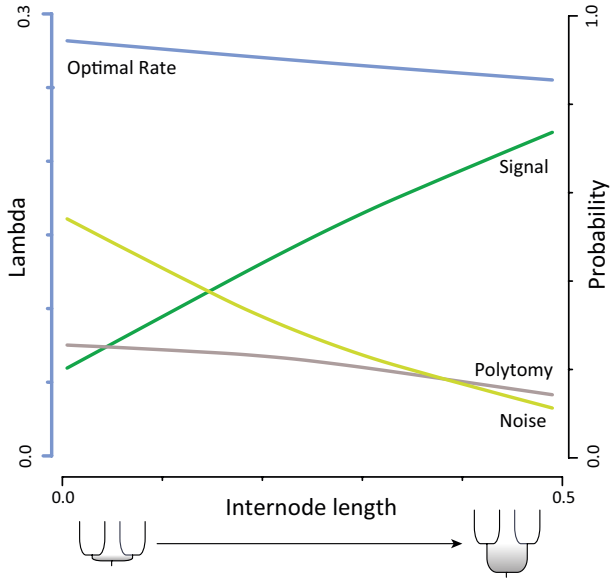

FIGURE 2. Optimal rate of change $\hat{\lambda}$ for a character for different quartet internode lengths t_0, based on a quartet tree with four subtending branches of length $T = 1$ under a Jukes–Cantor model.

by Steel and Leuenberger (2017), who examined optimal rates given an asymptotically short internode as in Townsend (2007). If we are to advance phylogenetic experimental design, a generalized evaluation of optimal rates under different combinations of time, internode, and rate is warranted.

For every short, deep internode length t_0 with four subtending branches of length T, we can again calculate the value at which R (the amount of resolution, equal to the phylogenetic signal supporting a correct topology minus the amount of noise in support of the most supported incorrect topology; *c.f.* Townsend et al. (2012); Su et al. (2014); Su and Townsend (2015) is largest, defining how the optimal rate $\hat{\lambda}$ changes (Appendix, Eq. A6). Comparing the effect on the optimal rate $\hat{\lambda}$ to the effect on probabilities of signal, noise, and polytomy (based on a character evolving at rate $\hat{\lambda}$) reveals a stark contrast. The length of the internode has enormous impact on the probability of resolving the internode; however, it has very little impact on the optimal rate of a character to resolve that internode (Fig. 2). The reason for this contrast can be explained by consideration of the fundamental drivers of signal and noise in phylogenetic inference. Probability of resolution is dramatically impacted by the length of the internode, because the probability that a character yields a shared derived state decreases monotonically (and approximately proportionally) with internode length t_0 as t_0 values approach zero. If the internode is very long, there is a high chance a character will change along its duration. If the internode is very short, there is a low chance a character will change along its duration.

Conversely, optimal rate is relatively unaffected by internode length. If T is long, slower rates will be optimal so that important changes during the internode t_0 are not obscured by subsequent changes. If T is short, faster rates will be optimal to ensure a change occurs during the internode t_0. For most persistent phylogenetic problems, $t_0 \ll T$. Consequently, the main issue that modulates optimal rate is therefore the length of T (Fig. 2). This mismatch between optimal rates and predictions of information content reveals that quantification of the optimality of rates does not provide a sufficient criterion for prediction of phylogenetic performance. The significant impact of t_0 on the probability of resolution clarifies that nevertheless, neglecting internode length is a potential pitfall of phylogenomic experimental design and inference.

Optimal Rates and Subtending Branch Lengths

The utility of a character also depends on the relative rates and times of evolution of subtending lengths of the internode to be resolved.

Expanding from the consideration of internodes, the subtending branching times of the quartet are also important to consider (Fig. 3). Varying the lengths of subtending branches will impact the degree to which rates are informative for resolving a specific node (Su and Townsend 2015). Consider assigning an internode length of $t_0 = 0.1$ and a length of $T_3 = T_4 = 1$, then $T_2 = T_3 + t_0 = 1.1$ (T_2). Examining a range of values for T_1 between 1.1 to 10 and solving at each value for the optimal rate $\hat{\lambda}$ for each value by calculating the value of λ which R is maximized (Appendix) provides a depiction of how increasing the length of a single subtending branch can increase optimal rates (Fig. 4).

Optimal Rates and Character States

The utility of a character evolving at a rate depends on the number of states the character can adopt.

Investigations of optimal rates have largely focused on time and tree structure, while simplifying model assumptions of character evolution to quantify rates (Shpak and Churchill 2000; Townsend 2007; Steel and Leuenberger 2017). These simplifications often involve conditioning character evolution on equal probabilities of change between character states (s). However, most empirical data sets will violate this assumption. Heterogenous rates of change between characters such as asymmetric transitions between transitions or transversions, or rarely reversed morphological states will drive a mismatch between the quantified rates of character evolution under simplified assumptions and those exhibited by the data (Yang 1996; Phillips et al. 2004; Leaché et al. 2015). Su et al. (2014) incorporated explicit specification of substitution models into locus scrutiny, providing a framework from which we can generalize both how the substitution model and the expansion or contraction of character state space (e.g., nucleotide versus amino acid substitution matrix) in itself impact quantification of optimal rates similar to the approach above.

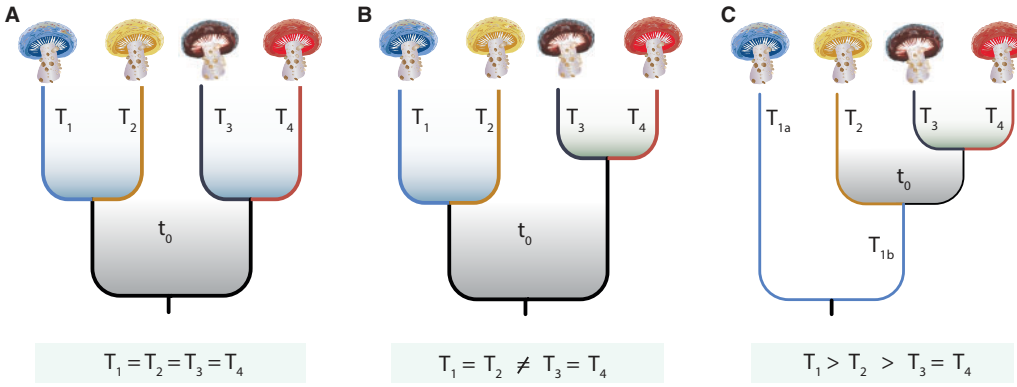

FIGURE 3. Rooted phylogenetic quartets, in which the topology and/or the shared evolutionary history between taxa is varied by changing the length of the quartet internode t_0 and the lengths of the subtending branches $T_1 - T_4$. A) A balanced topology is illustrated with equal values for subtending branches (T). In contrast, B) an unbalanced topology is illustrated with unequal times of divergence, comparing $T_1 = T_2$ with $T_3 = T_4$, and C) a pectinate topology is illustrated in which $T_1 > T_2 > T_3 = T_4$. The differences in shared history among scenarios result in different optimal rates.

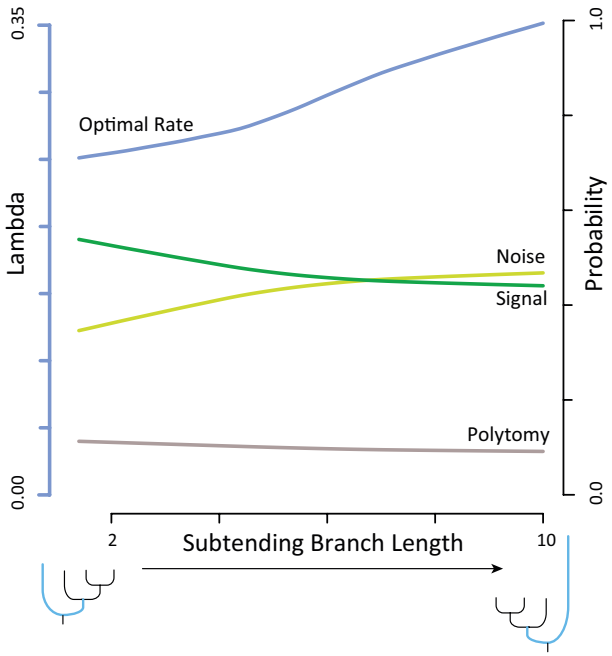

FIGURE 4. Optimal rate $\hat{\lambda}$ for a single character to resolve a pectinate quartet tree. The internode length (t_0) is 0.1 and the three subtending branches are of lengths $T_3 = T_4 = 1$, and $T_2 = 1.1$ (as in Fig. 3C), varying the divergence of the deepest-branching lineage, T_1 (light branch in the graphics along the x-axis).

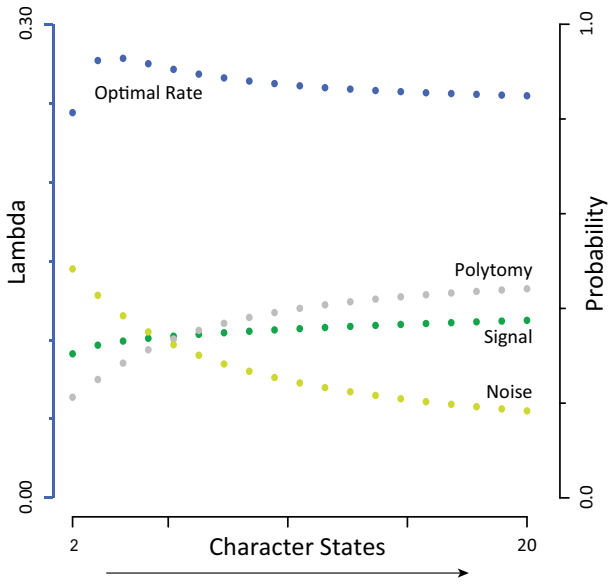

FIGURE 5. Optimal rate of change $\hat{\lambda}$ for a range of character state spaces from 2 to 20 as well as signal, noise, and polytomy probabilities, for resolution of a quartet with an internode $t_0 = 0.1$ and four subtending branches $T_1 = T_2 = T_3 = T_4 = T = 1$.

By expressing R as a function of λ and s (Appendix), we can expand the above framework and numerically solve for $\hat{\lambda}$ at any given value of s. The solution demonstrates that the character state space *per se* has little impact on optimal rate estimates (Fig. 5). However, there is a mismatch between the effect of s on optimal rates versus the effect of s on resolution. Increasing the state space lowers probabilities of incorrect resolution due to homoplasy. In contrast, evaluating $\hat{\lambda}$ across a range of substitution models demonstrates little effect of substitution model on either optimal rate estimates or predicted probabilities of resolution (Fig. 6).

Optimal Rates and Taxon Sampling

The utility of a character evolving at a rate depends not only on the depth of the phylogenetic inference and the length of the internode to be resolved, but also on the degree of ingroup (and outgroup) taxon sampling.

The question of the effect of addition of taxa on phylogenetic inference has long been a topic of

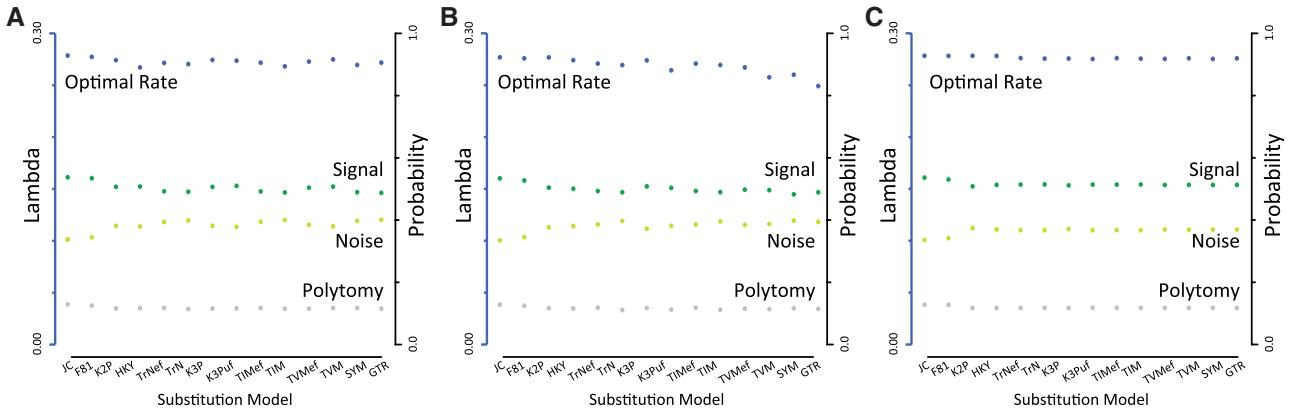

FIGURE 6. Optimal rate $\hat{\lambda}$ for a character to resolve a quartet with an internode $t_0 = 0.1$ and four subtending branches $T_1 = T_2 = T_3 = T_4 = T = 1$, as a function of applicable reversible models of molecular evolution. The models used in these analyses are based on the parameter values of models estimated for the (A) actin (*ACT1*) marker of 29 taxa of the yeast genus Candida and allied teleomorph genera (from Su et al. 2014; see also Su and Townsend 2015); (B) 12S sequences of 47 species of lemurs (from Federman et al. 2016); and (C) the glycosyltransferase (*glyt*) marker for 83 species of Antarctic teleost fishes (from Near et al. 2012a; see also Dornburg et al. 2017a).

conversation in the phylogenetic community (Pollock et al. 2002; Zwickl and Hillis 2002a; Hedtke et al. 2006a; Heath et al. 2008b; Crawley and Hilu 2011). For tractability, investigations of optimal rates for inference often ignore this subject by focusing on a defined topology with a set number of leaves, or even restricting attention to a simple phylogenetic quartet. However, additional information arising as a consequence of taxa sampled that surround the quartet results in a gap in our ability to accurately predict utility. Recently Townsend and Lopez-Giraldez (2010) partially bridged this gap, demonstrating changes in optimal rate that occur as a consequence of adding an additional taxon unit.

Townsend and Lopez-Giraldez (2010) considered placing an additional lineage into a quartet. For this lineage placed onto the quartet to yield novel resolution of the quartet, three things must happen. If (1), one or more changes in character state occurred on the internode, (2) no changes occur along four subtending lineages, and (3) the sister lineage to the ingroup formed by the addition of the new taxon underwent one or more changes in character state since diverging along an independent evolutionary pathway [Time from root of tree (T) – new internode ($\bar{t}$)], then the expectation of the informativeness of sampling a character of a new taxon is

$$e^{-sT\lambda}\left(1-e^{-t_0\lambda}\right)\left(1-1-e^{-(T-\bar{t})\lambda}\right), \qquad (1)$$

where s represents the character state space and t_0 represents the internode length; *c.f.* Equation 3 in Townsend and Lopez-Giraldez (2010). Townsend and Lopez-Giraldez (2010) illustrated the effect of taxon sampling compared to character sampling, demonstrating that the proximity to the deep internode of the additional taxon has a marked effect on the overall utility of additional taxon sampling—whereas it has a modest effect on the optimal rate: in their setting, Townsend and Lopez-Giraldez (2010) found the optimal rate ($\hat{\lambda}_t$) for a novel ingroup lineage to be $\hat{\lambda}_t \approx \frac{1}{2.2T}$. This rate is faster than the optimal rate for addition of a character to a lineage present in an extant quartet, $1/4T$ (Townsend 2007; Fischer and Steel 2009), consistent with both the simulation-based expectation that higher rates should become more advantageous as taxon sampling is increased (Hillis and Cannatella 1998; Heath et al. 2008a) and with theoretical analyses of several asymptotic scenarios (Townsend and Leuenberger 2011). Adding taxa provides additional information on the historical state of characters (Heath et al. 2008a), so it should be no surprise that taxon sampling strategies will impact calculations of optimal rates. Although the above framework has yet to be expanded to the complex taxon sampling strategies observed across empirical data sets, it is clear that adding one or more taxa that diverged close to an internode of interest will alter the predicted utility of different markers.

Optimal Rates and Character Sampling

Optimal rates of evolution vary depending on the desired level of confidence that we demand to support a phylogenetic hypothesis.

Up to now, we have defined the "optimal rate" to be the rate at which a single character is most likely to provide a pattern supportive of the correct resolution. We have also defined the desired "resolution" to be just a single character of support over either other resolution of a quartet. However, defining the utility of a data set to be gathered does not depend only on parameters of the tree and the molecular evolutionary model or the number of characters we are sampling. We must also specify the amount of node support necessary to proclaim a phylogenetic problem "resolved". Oft-used metrics of phylogenetic support, such as bootstrap support values, are one way to quantify support, but bootstrap proportions and Bayesian Posterior Probabilities have a narrow scope and can strongly support the incorrect topology. As few as four additional characters

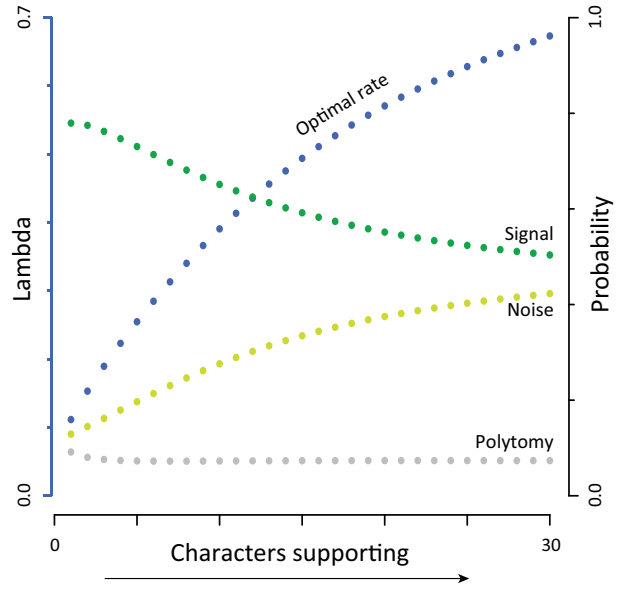

FIGURE 7. Optimal rate of change $\hat{\lambda}$ to resolve a quartet with an internode $t_0 = 0.1$ and four subtending branches $T_1 = T_2 = T_3 = T_4 = 1$, with the goal to achieve a specified level of resolution with maximum probability (defined as the number of characters supporting the true quartet above and beyond the support for the next most supported quartet) from 2 to 30.

supporting a given tree can provide 95% bootstrap proportion for a given topological inference (Felsenstein 1985). Bayesian Posteriors are generally viewed as less conservative, even more rapidly maxing out their scope for conveying level of support (Alfaro 2003). As we gather data sets spanning thousands, if not millions of characters, and evaluation of data is based more on the degree of support in light of potential confounders rather than the statistical significance differentiating a result from a polytomy, it is essential to also quantify and reflect on the relationship between optimal rates and the targeted level of evidence sufficient to consider a node "strongly resolved".

Townsend et al. (2012) quantified the probability of resolution based on one character supporting the correct topology. By expanding this formulation to include a specified number of supporting characters, we can express R as a function of λ and s (Appendix) and numerically solve for $\hat{\lambda}$ for any desired number of characters supporting the quartet tree. The solution demonstrates that targeting more characters that support the correct quartet resolution will increase optimal rate estimates (Fig. 7). While the solution that faster rates are necessary to obtain more characters is potentially intuitive, faster rates present an inevitable problem for inference. Increasing the target differential of supporting characters (and therefore, the optimal rate) also has the effect of lowering predicted probabilities of resolution, because instances of convergence in character state are expected to increase (Fig. 7). As we march eagerly into the use of genome-scale data, this predicted relationship between optimality and homoplasy raises a general question of experimental design: when evaluating large, complex topologies that are assemblages of many distinct hypotheses, how many characters should we target to gather sufficient evidence? As cases of strongly supported topological incongruence become increasingly pervasive (Brown et al. 2018; Reddy et al. 2017; Springer and Gatesy 2018; Ilves et al. 2018), consideration of this axis of experimental design represents an important and exciting frontier.

Optimal Rates and Phylogenetic Experimental Design

Studies of optimal rate have been a conceptual keystone of phylogenetic experimental design, potentially empowering cost effective sequencing, maximizing productivity of staff, and enabling accurate inference (Townsend and Leuenberger 2011; Klopfstein et al. 2017). Far from being intractable, we have demonstrated that optimal rates for phylogenetic inference are complex but calculable under a range of conditions that include varying tree structure, character states, and underlying molecular models. While all of these factors do have an impact on quantification of optimal rates, we have demonstrated that their effects on optimal rate values are usually modest. For example, acceleration of the rate of evolution of a subtending branch by as much as an order of magnitude can markedly affect the utility of a character to help resolve a phylogenetic problem (Su and Townsend 2015), but only produces a minor change in the optimal rate to solve that problem (Fig. 4). Moreover, the difference in optimal rate between different complexities of reversible models of molecular evolution, ranging from a Jukes-Cantor to a general time reversible model, is virtually negligible (Fig. 6). The modest sensitivity of optimal rate to other parameters of the evolutionary process demonstrated here suggests that, generally, optimal rates reside within a known window across a wide range of conditions for a temporally specific phylogenetic problem, and supports the practice of predicting utility of a set of characters on the basis of their evolutionary rate.

Although this apparent robustness of optimal rate is encouraging, there are several caveats to consider with empirical data. First, it is important to consider that optimal rates by our definition do depend critically on what specific phylogenetic problem an investigator is attempting to resolve. Empirical studies—including our own—typically demand that their data to resolve nodes across a wide range of time scales (Near et al. 2013; Federman et al. 2016; Cantalapiedra et al. 2017; Forrestel et al. 2017). This desire for utility across historical time creates an essential conflict of experimental design. Optimal rates for resolving one node can be suboptimal, uninformative, or even positively misleading for resolving another. Thoughtful contemplation of this fundamental apothegm of phylogenetic experimental design is of high utility when screening genomic scale data for loci of predicted utility for resolution of

some of the most recalcitrant nodes across the Tree of Life.

Taxon sampling will also alter predictions of what rates are optimal. While a reasonable expectation that additional taxon sampling will allow for faster rates to be of utility exists, how complex taxon sampling strategies impact optimal rate estimates is not as yet entirely clear. In principle, effective taxon sampling should enable investigators to harness faster-evolving markers with less fear of misleading homoplasy (Zwickl and Hillis 2002b; Hedtke et al. 2006b; Wilson 2011; Crawley and Hilu 2012; Hilu et al. 2014). Intuitively, adding more taxa provides a concave, sublinear advantage of breaking up long branches, improving ancestral information, thereby generally increasinging accuracy. However, adding more taxa will lead to a convex, nonlinear increase in the complexity of fully resolving all branches of a phylogenetic tree topology, demanding resolution of more hypothetical ancestral relationships from the same data. Further, addition of taxa can increase the probability of introducing new rate heterogeneities and biases, thereby adding long branches and potential model violations to an erstwhile tractable phylogenetic problem (Reddy et al. 2017). While we lack a theoretical framework that evaluates phylogenetic utility explicitly for highly sampled, complex trees, it is clear that phylogenetic information content is not evenly distributed between taxa, within loci, or among characters within loci (Townsend and Lopez-Giraldez 2010; Bordewich et al. 2017). This variation of phylogenetic information content creates one of the biggest challenges for utilizing phylogenetic experimental design: translating the expectations of optimal rates back to empirical data sets.

The distribution of site rates within empirical data sets often extends far beyond the boundaries of optimal rate estimates for a given phylogenetic problem. As data sets containing hundreds if not thousands of loci are becoming increasingly common (Jarvis et al. 2014; Eytan et al. 2015; Prum et al. 2015), how then do we apply the expectations of optimal rates and inference to the selection of loci for inference? Given the ubiquity of site-rate heterogeneity within loci, the practice of considering the mean rate as a criterion for a gene being optimal for inference (Fig. 1) is not appropriate. In the worst case scenario, assigning a single rate to a complex locus can drive a mismatch between empirical results and presumed theoretical expectations between optimal rates and phylogenetic utility (c.f. Aguileta et al. 2008; Moeller and Townsend 2011, 2013). Fortunately over the past several years early approaches that incorporate site rate heterogeneity (Townsend 2007) have become increasingly sophisticated (Susko and Roger 2012; Townsend et al. 2012; Su et al. 2014; Su and Townsend 2015). By assessing how rates at each individual site contribute to the overall predicted utility of a locus, these approaches have laid a foundation that facilitates a much needed theoretical expansion of phylogenetic experimental design. However, additional challenges remain.

It is our view that a solitary focus on optimal rates estimated under specific conditions offers diminishing returns for empirical phylogeneticists. Indeed, multiple authors have pointed out (consistent with theory) that near-optimal rates are almost as informative as optimal rates (Yang 1998; Klopfstein et al. 2017). Instead, we believe that continuing to build a practical theory of phylogenetic experimental design is critical if we are to meet the challenge of providing a more robust predictive framework for empirical studies. For example, it has been well established that rates of molecular evolution can also change along branches. Changes in life history can drive extreme changes in rate between lineages (Smith and Donoghue 2008; Dornburg et al. 2012; Lanfear et al. 2013), while clade-wide changes in diversification dynamics drive rate increases or decreases across an entire topology within a geologic time-slice (Steiper and Seiffert 2012; Berv and Field 2017). However, phylogenetic experimental design rarely incorporates either phenomenon. Integration of the predicted relationship between molecular rates and life history into considerations of phylogenetic utility is an avenue of tremendous potential. Likewise, studies of experimental design largely consider site rates to evolve under a Poisson process. However, empirical data sets often do not. Compositional biases have been repeatedly highlighted as problematic in empirical data sets (Cox et al. 2014; Li et al. 2014; Dornburg et al. 2017b; Reddy et al. 2017). Although the impact of non-randomly evolving compositional patterns on phylogenetic experimental design methods has not been rigorously evaluated, it is reasonable to intuit that high levels of bias could drive convergences in site patterns that appear to be evolving at a rate near "optimal". A theory of bias in phylogenetic experimental design represents another highly useful area of development for the phylogenomics community.

Conclusion

Over the past several decades, we have made tremendous strides towards developing a conceptual understanding of the relationship between the rate of evolution and our ability to resolve a given phylogenetic problem. It is well known that the factors of a phylogenetic problem highlighted here can markedly affect accurate inference. However, despite this complexity, these factors often result in very minor changes in optimal rates of change for phylogenetic resolution. This inequivalency of effect on expected information and optimality is yet one more example of the utility of the further development of theory of phylogenetic experimental design. The effort devoted should be analogous to the longstanding effort within the phylogenetic community devoted to developing inference methods. Just as phylogenetic inference methods have grown to accommodate complex evolutionary patterns, so too must the theory of phylogenetic experimental design if it is to be comprehensive. Given that lineage- and/or locus-specific patterns of rate heterogeneity as well as

compositional biases are common features of genomic data sets, developing theory that aids in overcoming these challenges to empirical phylogenomics should provide exciting and highly useful avenues of research. Such advances are critical if we are to achieve consistency of inference across the Tree of Life and generate a robust understanding of the processes that maintain biodiversity.

Funding

The authors thank the Notsew Orm Sands Foundation for support to JPT for this research.

Acknowledgments

The authors thank Art Bogan and Katerina Zapfe for valuable comments on an early version of this manuscript.

Appendix

To demonstrate the impact that site rate variation can have on inference, a metric that facilitates quantification of the expected level of support a given character contributes towards resolution of the true tree is needed. Such a framework was recently developed (Townsend et al. 2012; Su and Townsend 2015) by deriving a series of equations that calculate the predicted probability of a character exhibiting a synapomorphic pattern of character states at the terminal tips of a phylogenetic quartet that is consistent with the true quartet topology (denoted "y" in Equation 4 of Townsend et al. 2012 and in Equation 3 in Su and Townsend 2015), and the probabilities of this character exhibiting a homoplasious pattern of character states consistent with either of the two possible incorrect quartet topologies (denoted x_1 and x_2 in Equations 5 and 6 in Townsend et al. 2012, and Su and Townsend 2015). This theory is built upon the well-established expectations of nucleotide substitution under a GTR model. To calculate the average rate of substitution, λ, for a given character, we must consider the instantaneous rate (q_{ij}) of changes in nucleotide from i to j, where $j \neq i$ and $i =$ A, G, C, or T in the instantaneous rate matrix (*c.f.* Q(λ); Whelan et al. 2001) and calculate

$$\lambda = \sum_i \sum_{j \neq i} \pi_i q_{ij}, \tag{A1}$$

with π_i ($i =$ A, G, C, T) representing the frequencies of each nucleotide state at equilibrium (also see Townsend et al. 2012; Su et al. 2014; Su and Townsend 2015). From here, we can generate a substitution probability matrix $P(\lambda, t)$ that describes the probability of a nucleotide state change in a finite time (t) through the following equation

$$P(\lambda, t) = e^{Q(\lambda)t}, \tag{A2}$$

Via eigendecomposition of Equation A2, Su and Townsend (2015) developed a theoretical framework that predicts the probability of a change in character state pattern (e.g., yielding synapomorphies or homoplasy) for a phylogenetic quartet with four uneven subtending branches. The length (in time) of the internode of this quartet tree is represented by t_0; the lengths of the four subtending branches are denoted as T_1, T_2, T_3, and T_4. The rate of substitution of a molecular character under investigation can be specified as λ along the internode and along each of the four subtending branches. The assumption of homogeneity of the average substitution rate can be relaxed by allowing characters to evolve at different rates in the internode and the four subtending branches (*c.f.* Fig. 1 in Su and Townsend 2015).

Based on this framework, an optimal value of λ for the character to correctly resolve the quartet tree ($\hat{\lambda}$) can be calculated as the value of λ at which the level of resolution of an internode $R = y - \mathrm{Max}(x_1, x_2)$ is maximized, where y is the probability of the single character supporting the correct quartet subtree, and x_1 and x_2 are the probabilities of the single character supporting each of the two incorrect quartet subtrees. Because of symmetry, support for either incorrect tree is equal (this symmetry to applies to all subsequent figures). Because R is uniformly concave, this optimum can be obtained by the old calculus trick of setting the partial derivative of R with respect to λ equal to zero and solving (analytically or numerically) for its zero value. Provided the second derivative is negative, this procedure yields an optimal value of λ, $\hat{\lambda}$, at which R is maximized. Obtaining $\hat{\lambda}$ is eased by calculating the value of λ at which $y - x_1$ is maximized: by symmetry, the rate that maximizes $R = y - \mathrm{Max}[x_1, x_2]$ will also maximize $\bar{R} = y - x_1$.

To evaluate the scenario presented in Figure 3A, we can use Equation 3 of Su and Townsend (2015),

$$\begin{aligned} &y(\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4; t_0, T_1, T_2, T_3, T_4) \\ &= \sum_M \sum_N \sum_{C_1 = C_2} \sum_{C_3 = C_4 \neq C_1} \pi_M P_{MN}(\lambda_0, t_0) P_{MC}(\lambda_1, T_1) \\ &\quad \times P_{MC_2}(\lambda_2, T_2) P_{NC_3}(\lambda_3, T_3) P_{NC_4}(\lambda_4, T_4), \end{aligned} \tag{A3}$$

where M and N denote the ancestral states of the nucleotide character at the two ends of the internode, whose length in time is represented by t_0; and C_1, C_2, C_3, and C_4 represent the nucleotide character's states at the terminal tips of the four subtending branches (T_1, T_2, T_3, and T_4, respectively); and $P_{ij}(\lambda, t)$ provides the probability that the character with average substitution rate λ will change from nucleotide i to nucleotide j ($j \neq i$) after time t. For Figure 3A, the lengths of the four subtending branches are all equal to 1, so we can assign $T_1 = T_2 = T_3 = T_4 = 1$, and assign the instantaneous rate and base frequency parameters to be consistent with the Jukes–Cantor (JC) model of nucleotide substitution ($\pi_i = 0.25$, and $r_{ij} = 1$, where $j \neq i$, $i =$ A, G, C, or T, and r_{ij} represents the relative rate parameter in the

instantaneous rate matrix $\mathbf{Q}(\lambda)$, yielding

$$y=\frac{3}{64}+\frac{27}{64}e^{\frac{-16\lambda}{3}}+\frac{9}{32}e^{\frac{-8\lambda}{3}}-\frac{3}{16}e^{-\frac{16\lambda}{3}-\frac{4t_0\lambda}{3}}-\frac{3}{8}e^{-4\lambda-\frac{4t_0\lambda}{3}}-\frac{3}{16}e^{-\frac{8\lambda}{3}-\frac{4t_0\lambda}{3}}. \quad (A4)$$

Evaluating Equations 4 and 5 in Su and Townsend (2015), symmetry of the quartet tree dictates that

$$x_1=x_2=\frac{3}{64}+\frac{3}{64}e^{\frac{-16\lambda}{3}}-\frac{3}{32}e^{\frac{-8\lambda}{3}}+\frac{3}{16}e^{-\frac{16\lambda}{3}-\frac{4t_0\lambda}{3}}-\frac{3}{8}e^{-4\lambda-\frac{4t_0\lambda}{3}}+\frac{3}{16}e^{-\frac{8\lambda}{3}-\frac{4t_0\lambda}{3}}, \quad (A5)$$

where x_1 and x_2 can be calculated similarly to Equation A3. The character evolving at a rate that maximizes $R=y-Max[x_1,x_2]$ is predicted to provide the greatest support. Invoking symmetry, the rate that maximizes $R=y-Max[x_1,x_2]$ will also maximize $\bar{R}=y-x_1$, therefore,

$$\bar{R}=\frac{3}{8}e^{\frac{-16\lambda}{3}}+\frac{3}{8}e^{\frac{-8\lambda}{3}}-\frac{3}{8}e^{-\frac{16\lambda}{3}-\frac{4t_0\lambda}{3}}-\frac{3}{8}e^{-\frac{8\lambda}{3}-\frac{4t_0\lambda}{3}}, \quad (A6)$$

and

$$\frac{\partial\bar{R}}{\partial\lambda}=-2e^{\frac{-16\lambda}{3}}-e^{\frac{-8\lambda}{3}}-\frac{3}{8}e^{-\frac{16\lambda}{3}-\frac{4t_0\lambda}{3}}\left(-\frac{16}{3}-\frac{4t_0}{3}\right)-\frac{3}{8}e^{-\frac{8\lambda}{3}-\frac{4t_0\lambda}{3}}\left(-\frac{8}{3}-\frac{4t_0}{3}\right). \quad (A7)$$

Solving for the value of λ at which Equation A7 equals zero is analytically challenging. However, a numerical solution for $\frac{\partial\bar{R}}{\partial\lambda}=0$ can be readily obtained (and therefore, the solution of $\frac{\partial R}{\partial\lambda}=0$, and therefore, $\hat{\lambda}$, the optimal value of λ). To generate Figure 2, we evaluated Equation A7 above at a range of values of t_0, then solved for $\hat{\lambda}$ numerically at each value.

In the scenario presented in Figure 4, the internode $t_0=0.1$, and we fix the length of three of the four subtending branches such that $T_3=T_4=1$ and $T_2=T_3+t_0=1.1$. We then calculate $\hat{\lambda}$ depending on T_1, which can be obtained using Equation A3—as it was in Equations A4–A6—by substituting the quartet tree branch lengths and JC model instantaneous rate and base frequency parameters into Equations 3–5 in (Su and Townsend, 2015). In this case,

$$y=0.047-\frac{3e^{-4.27\lambda}}{32}-\frac{3e^{-2.93\lambda}}{32}+0.14e^{\frac{-8\lambda}{3}}-\frac{3}{16}e^{-4.27\lambda-\frac{4T_1\lambda}{3}}+0.42e^{-4.13\lambda-\frac{4T_1\lambda}{3}}-\frac{3}{16}e^{-2.93\lambda-\frac{4T_1\lambda}{3}}-0.094e^{-2.8\lambda-\frac{4T_1\lambda}{3}}+0.047e^{-1.47\lambda-\frac{4T_1\lambda}{3}} \quad (A8)$$

Using symmetry ($T_3=T_4$),

$$x_1=x_2=0.047-\frac{3e^{-4.27\lambda}}{32}+\frac{3e^{-2.93\lambda}}{32}-0.047e^{\frac{-8\lambda}{3}}+\frac{3}{16}e^{-4.27\lambda-\frac{4T_1\lambda}{3}}+0.047e^{-4.13\lambda-\frac{4T_1\lambda}{3}}-\frac{3}{16}e^{-2.93\lambda-\frac{4T_1\lambda}{3}}-0.094e^{-2.8\lambda-\frac{4T_1\lambda}{3}}+0.047e^{-1.47\lambda-\frac{4T_1\lambda}{3}}. \quad (A9)$$

In this scenario, defining

$$\bar{R}=y-x_1=-\frac{3e^{-2.93\lambda}}{16}+0.1875e^{\frac{-8\lambda}{3}}-\frac{3}{8}e^{-4.27\lambda-\frac{4T_1\lambda}{3}}+0.375e^{-4.13\lambda-\frac{4T_1\lambda}{3}}, \quad (A10)$$

and differentiating

$$\frac{\partial\bar{R}}{\partial\lambda}=0.55e^{-2.93\lambda}-0.5e^{\frac{-8\lambda}{3}}-\frac{3}{8}e^{-4.27\lambda-\frac{4T_1\lambda}{3}}\left(-4.27-\frac{4T_1}{3}\right)+0.375e^{-4.13\lambda-\frac{4T_1\lambda}{3}}\left(-4.13-\frac{4T_1}{3}\right). \quad (A11)$$

a numerical solution for $\frac{\partial\bar{R}}{\partial\lambda}=0$ can be readily obtained.

The scenario presented in Figure 5 corresponds to a quartet tree with four even subtending branches as described by Townsend et al. (2012) in the first iteration of phylogenetic signal and noise analysis. Townsend et al. (2012) derived the expressions for y, x_1, and x_2 for a molecular character evolving at an average substitution rate of λ for resolving a quartet tree with an internode t_0 and four subtending branches of equal length T. The character follows the Poisson model of molecular evolution and the number of character states in the Poisson model is s. By substituting $t_0=0.1$ and $T=1$ into Equation 4 in Townsend et al. (2012),

$$y=\frac{-1+s}{s^3}+\frac{e^{-\frac{3.1s\lambda}{-1+s}}\left(-8+12s-4s^2\right)}{s^3}+\frac{e^{-\frac{2s\lambda}{-1+s}}\left(2+e^{-\frac{0.1s\lambda}{-1+s}}(4-4s)-4s+2s^2\right)}{s^3}+\frac{e^{-\frac{4s\lambda}{-1+s}}(-1+3s-3s^2+s^3+e^{-\frac{0.1s\lambda}{-1+s}}(4-8s+5s^2-s^3))}{s^3}. \quad (A12)$$

Evaluating Equations 5 and 6 in Townsend et al. (2012) and using the symmetry between x_1 and x_2 in this scenario,

$$x_1=x_2=\frac{-1+s}{s^3}+\frac{e^{-\frac{3.1s\lambda}{-1+s}}(-8+12s-4s^2)}{s^3}+\frac{e^{-\frac{2s\lambda}{-1+s}}(2-2s+e^{-\frac{0.1s\lambda}{-1+s}}(4-6s+2s^2))}{s^3}$$

$$+\frac{e^{-\frac{4s\lambda}{-1+s}}(-1+s+e^{-\frac{0.1s\lambda}{-1+s}}(4-6s+2s^2))}{s^3}. \quad \text{(A13)}$$

In this scenario,

$$\begin{aligned}\bar{R}=y-x_1=&\frac{e^{-\frac{2s\lambda}{-1+s}}(2+e^{-\frac{0.1s\lambda}{-1+s}}(4-4s)-4s+2s^2)}{s^3}\\&-\frac{e^{-\frac{2s\lambda}{-1+s}}(2-2s+e^{-\frac{0.1s\lambda}{-1+s}}(4-6s+2s^2))}{s^3}\\&-\frac{e^{-\frac{4s\lambda}{-1+s}}(-1+s+e^{-\frac{0.1s\lambda}{-1+s}}(4-6s+2s^2))}{s^3}\\&+\frac{e^{-\frac{4s\lambda}{-1+s}}(-1+3s-3s^2+s^3+e^{-\frac{0.1s\lambda}{-1+s}}(4-8s+5s^2-s^3))}{s^3},\end{aligned} \quad \text{(A14)}$$

and

$$\begin{aligned}\frac{\partial\bar{R}}{\partial\lambda}=&-\frac{0.1e^{-\frac{2.1s\lambda}{-1+s}}(4-4s)}{(-1+s)s^2}+\frac{0.1e^{-\frac{4.1s\lambda}{-1+s}}(4-6s+2s^2)}{(-1+s)s^2}\\&+\frac{0.1e^{-\frac{2.1s\lambda}{-1+s}}(4-6s+2s^2)}{(-1+s)s^2}\\&-\frac{2e^{-\frac{2s\lambda}{-1+s}}(2+e^{-\frac{0.1s\lambda}{-1+s}}(4-4s)-4s+2s^2)}{(-1+s)s^2}\\&-\frac{0.1e^{-\frac{4.1s\lambda}{-1+s}}(4-8s+5s^2-s^3)}{(-1+s)s^2}\\&+\frac{2e^{-\frac{2s\lambda}{-1+s}}(2-2s+e^{-\frac{0.1s\lambda}{-1+s}}(4-6s+2s^2))}{(-1+s)s^2}\\&+\frac{4e^{-\frac{4s\lambda}{-1+s}}(-1+s+e^{-\frac{0.1s\lambda}{-1+s}}(4-6s+2s^2))}{(-1+s)s^2}\\&-\frac{4e^{-\frac{4s\lambda}{-1+s}}(-1+3s-3s^2+s^3+e^{-\frac{0.1s\lambda}{-1+s}}(4-8s+5s^2-s^3))}{(-1+s)s^2}.\end{aligned} \quad \text{(A15)}$$

The value of $\hat{\lambda}$ at any given value of s can then be solved numerically.

Lastly, in the various scenarios presented in Figure 6 where a character evolves according to different models of nucleotide substitution and the quartet tree has an internode $t_0=0.1$ and four equal subtending branches $T_1=T_2=T_3=T_4=1$, the value of $\hat{\lambda}$ can be computed in the same way as it was for Figures 1 and 4. For each model of nucleotide substitution considered, Equations 3–5 can be evaluated by setting $T_1=T_2=T_3=T_4=1$, $t_0=0.1$, and the instantaneous rate and base frequency parameters as those estimated for that particular model for the ACT1 marker in the analysis by Su et al. (2014). For example, for the F81 model (Felsenstein 1981), the estimated model parameters are $\pi_T=0.32$, $\pi_C=0.29$, $\pi_A=0.23$, and $\pi_G=0.16$, and $rTC=rTA=rTG=rCA=rCG=rAG=1$. By substituting those model parameter values and the tree branch length values into Equations 3–5 in Su and Townsend (2015), formulae for y, x_1, and x_2 can be obtained. In this case,

$$\begin{aligned}\frac{\partial\bar{R}}{\partial\lambda}=&1.96264e^{-5.57823\lambda}-1.91478e^{-5.44218\lambda}\\&-1.41610\times10^{-17}e^{-4.08163\lambda}+0.66748e^{-2.85714\lambda}\\&+0.42726e^{-2.85714\lambda}-.63570e^{-2.72109\lambda}\\&-.40691e^{-2.72109\lambda}+0.00660e^{-1.49660\lambda}\\&-.00660e^{-1.49660\lambda}-.00600e^{-1.36054\lambda}\\&+0.00600e^{-1.36054\lambda}+4.13029\times10^{-19}e^{-0.13605\lambda}.\end{aligned} \quad \text{(A16)}$$

The value of $\hat{\lambda}$ for the model can then be computed by setting $\frac{\partial\bar{R}}{\partial\lambda}=0$ and solving numerically. The same approach can be used to calculate $\hat{\lambda}$ for the other models of nucleotide substitution.

References

Aguileta G., Marthey S., Chiapello H., Lebrun M.-H., Rodolphe F., Fournier E., Gendrault-Jacquemard A., Giraud T. 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst. Biol. 57:613–627.

Alfaro M.E. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Berv J.S., Field D.J. 2017. Genomic signature of an Avian Lilliput effect across the K-Pg Extinction. Syst. Biol. 63:1–13.

Blouin M.S., Yowell C.A., Courtney C.H., Dame J.B. 1998. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol. Biol. Evol. 15:1719–1727.

Bordewich M., Deutschmann I.M., Fischer M., Kasbohm E., Semple C., Steel M. 2017. On the information content of discrete phylogenetic characters. J. Math. Biol. 1–18.

Braasch I., Gehrke A.R., Smith J.J., Kawasaki K., Manousaki T., Pasquier J., Amores A., Desvignes T., Batzel P., Catchen J., Berlin A.M., Campbell M.S., Barrell D., Martin K.J., Mulley J.F., Ravi V., Lee A.P., Nakamura T., Chalopin D., Fan S., Wcisel D., Cañestro C., Sydes J., Beaudry F.E.G., Sun Y., Hertel J., Beam M.J., Fasold M., Ishiyama M., Johnson J., Kehr S., Lara M., Letaw J.H., Litman G.W., Litman R.T., Mikami M., Ota T., Saha N.R., Williams L., Stadler P.F., Wang H., Taylor J.S., Fontenot Q., Ferrara A., Searle S.M.J., Aken B., Yandell M., Schneider I., Yoder J.A., Volff J.-N., Meyer A., Amemiya C.T., Venkatesh B., Holland P.W.H., Guiguen Y., Bobe J., Shubin N.H., Di Palma F., Alföldi J., Lindblad-Toh K., Postlethwait J.H. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. 48:427–437.

Bromham L., Hua X., Lanfear R., Cowman P.F. 2015. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. Am. Nat. 185:507–524.

Brown J.W., Wang N., Smith S.A. 2017. The development of scientific consensus: Analyzing conflict and concordance among avian phylogenies. Mol. Phylogenet. Evol. 116:69–77.

Cantalapiedra J.L., Prado J.L., Hernández Fernández M., Alberdi M.T. 2017. Decoupled ecomorphological evolution and diversification in Neogene-Quaternary horses. Science. 355:627–630.

Cox C.J., Li B., Foster P.G., Embley T.M., Civán P. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst. Biol. 63:272–279.

Crawley S.S., Hilu K.W. 2011. Impact of missing data, gene choice, and taxon sampling on phylogenetic reconstruction: the Caryophyllales (angiosperms). Plant Syst. Evol. 298:297–312.

Crawley S.S., Hilu K.W. 2012. Caryophyllales: Evaluating phylogenetic signal in trnK intron versus matK. J. Syst. Evol. 50:387–410.
Dornburg A., Brandley M.C., McGowen M.R., Near T.J. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol. Biol. Evol. 29:721–736.
Dornburg A., Federman S., Lamb A.D., Jones C.D., Near T.J. 2017a. Cradles and museums of Antarctic teleost biodiversity. Nat. Ecol. Evol. 1:1379–1384.
Dornburg A., Townsend J.P., Brooks W., Spriggs E., Eytan R.I., Moore J.A., Wainwright P.C., Lemmon A., Lemmon E.M., Near T.J. 2017b. New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. Mol. Phylogenet. Evol. 110:27–38.
Eytan R.I., Evans B.R., Dornburg A., Lemmon A.R., Lemmon E.M., Wainwright P.C., Near T.J. 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. BMC Evol. Biol. 15:113.
Federman S., Dornburg A., Daly D.C., Downie A., Perry G.H., Yoder A.D., Sargis E.J., Richard A.F., Donoghue M.J., Baden A.L. 2016. Implications of lemuriform extinctions for the Malagasy flora. Proc. Natl. Acad. Sci. U.S.A. 113:5041–5046.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783.
Fischer M., Steel M. 2009. Sequence length bounds for resolving a deep phylogenetic divergence. J. Theor. Biol. 256:247–252.
Forrestel E.J., Donoghue M.J., Edwards E.J., Jetz W., du Toit J.C.O., Smith M.D. 2017. Different clades and traits yield similar grassland functional responses. Proc. Natl. Acad. Sci. U.S.A. 114:705–710.
Gan H.M., Tan M.H., Lee Y.P., Schultz M.B., Horwitz P., Burnham Q., Austin C.M. 2018. More evolution underground: accelerated mitochondrial substitution rate in Australian burrowing freshwater crayfishes (Decapoda: Parastacidae). Mol. Phylogenet. Evol. 118:88–98.
Goldman N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. R. Soc. B Biol. Sci. 265:1779–1786.
Graybeal A. 1993. The phylogenetic utility of cytochrome b: lessons from bufonid frogs. Mol. Phylogenet. Evol. 2:256–269.
Heath T.A., Zwickl D.J., Kim J., Hillis D.M. 2008a. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst. Biol. 57:160–166.
Heath T.A., Zwickl D.J., Kim J., Hillis D.M. 2008b. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. Syst. Biol. 57:160–166.
Hedtke S.M., Townsend T.M., Hillis D.M., Collins T. 2006a. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522–529.
Hedtke S.M., Townsend T.M., Hillis D.M., Collins T. 2006b. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522–529.
Hillis D.M., Cannatella D. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3–8.
Hilu K.W., Black C.M., Oza D. 2014. Impact of gene molecular evolution on phylogenetic reconstruction: a case study in the rosids (Superorder Rosanae, Angiosperms). PLoS One. 9:e99725.
Ilves K.L., Torti D., López-Fernández H. 2018. Exon-based phylogenomics strengthens the phylogeny of Neotropical cichlids and identifies remaining conflicting clades (Cichliformes: Cichlidae: Cichlinae). Mol. Phylogenet. Evol. 118:232–243.
Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Vargas Velazquez A.M., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320–1331.
Klopfstein S., Massingham T., Goldman N. 2017. More on the best evolutionary rate for phylogenetic analysis. Syst. Biol. 66:769–785.
Lanfear R., Ho S.Y.W., Jonathan Davies T., Moles A.T., Aarssen L., Swenson N.G., Warman L., Zanne A.E., Allen A.P. 2013. Taller plants have lower rates of molecular evolution. Nat. Commun. 4:1879.
Lanier H.C., Huang H., Lacey Knowles L. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylogenet. Evol. 70:112–119.
Leaché A.D., Banbury B.L., Felsenstein J., de Oca A.N.-M., Stamatakis A. 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Syst. Biol. 64:1032–1047.
Li B., Lopes J.S., Foster P.G., Martin Embley T., Cox C.J. 2014. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Mol. Biol. Evol. 31:1697–1709.
Makowsky R., Cox C.L., Roelke C.E., Chippindale P.T. 2010. Analyzing the relationship between sequence divergence and nodal support using Bayesian phylogenetic analyses. Mol. Phylogenet. Evol. 57:485–494.
Minias P., Whittingham L.A., Dunn P.O. 2017. Coloniality and migration are related to selection on MHC genes in birds. Evolution. 71:432–441.
Moeller A.H., Townsend J.P. 2011. Phylogenetic informativeness profiling of 12 genes for 28 vertebrate taxa without divergence dates. Mol. Phylogenet. Evol. 60:271–272.
Moeller A.H., Townsend J.P. 2013. Response to: The relative utility of sequence divergence and phylogenetic informativeness profiling in phylogenetic study design. Mol. Phylogenet. Evol. 66:436.
Moreira D., Philippe H. 2000. Molecular phylogeny: pitfalls and progress. Int. Microbiol. 3:9–16.
Mueller R.L. 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst. Biol. 55:289–300.
Near T.J., Dornburg A., Eytan R.I., Keck B.P., Smith W.L., Kuhn K.L., Moore J.A., Price S.A., Burbrink F.T., Friedman M., Wainwright P.C. 2013. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc. Natl. Acad. Sci. U.S.A. 110:12738–12743.
Near T.J., Dornburg A., Kuhn K.L., Eastman J.T., Pennington J.N., Patarnello T., Zane L., Fernández D.A., Jones C.D. 2012a. Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic fishes. Proc. Natl. Acad. Sci. U.S.A. 109:3434–3439.
Near T.J., Eytan R.I., Dornburg A., Kuhn K.L., Moore J.A., Davis M.P., Wainwright P.C., Friedman M., Smith W.L. 2012b. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. U.S.A. 109:13698–13703.
Nosenko T., Schreiber F., Adamska M., Adamski M., Eitel M., Hammel J., Maldonado M., Müller W.E.G., Nickel M., Schierwater B., Vacelet J., Wiens M., Wörheide G. 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. 67:223–233.
Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.
Pisani D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N., Wörheide G. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. U.S.A. 112:15402–15407.
Pollock D.D., Zwickl D.J., McGuire J.A., Hillis D.M., Crandall K. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51:664–671.
Pollock L.J., Thuiller W., Jetz W. 2017. Large conservation gains possible for global biodiversity facets. Nature. 546:141–144.
Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R. 2015. A comprehensive phylogeny of birds

(Aves) using targeted next-generation DNA sequencing. Nature. 526:569–573.
Purvis A., CarDillo M., Grenyer R., Collen B. 2005. Correlates of extinction risk: phylogeny, biology, threat and scale. In: Purvis A, Gittleman JL, Brooks T, editors. Phylogeny and Conservation. Cambridge UK: Cambridge University Press, p. 295–316.
Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.-L., Harshman J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 2017. Why do phylogenomic data sets yield conflicting trees? data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66:857–879.
Reiter J.G., Makohon-Moore A.P., Gerold J.M., Bozic I., Chatterjee K., Iacobuzio-Donahue C.A., Vogelstein B., Nowak M.A. 2017. Reconstructing metastatic seeding patterns of human cancers. Nat. Commun. 8:14114.
Ren R., Sun Y., Zhao Y., Geiser D., Ma H., Zhou X. 2016. Phylogenetic resolution of deep eukaryotic and fungal relationships using highly conserved low-copy nuclear genes. Genome Biol. Evol. 8:2683–2701.
Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.
Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. 1:126.
Shpak M., Churchill G.A. 2000. The information content of a character under a Markov model of evolution. Mol. Phylogenet. Evol. 17:231–243.
Smith S.A., Donoghue M.J. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science. 322:86–89.
Springer M.S., Gatesy J. 2018. On the importance of homology in the age of phylogenomics. Syst. Biodivers. 16(3):210–228.
Steel M., Leuenberger C. 2017. The optimal rate for resolving a near-polytomy in a phylogeny. J. Theor. Biol. 420:174–179.
Steiper M.E., Seiffert E.R. 2012. Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. Proc. Natl. Acad. Sci. U.S.A. 109:6006–6011.
Sullivan J., Swofford D.L., Naylor G. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16:1347–1356.
Susko E., Roger A.J. 2012. The probability of correctly resolving a split as an experimental design criterion in phylogenetics. Syst. Biol. 61:811–821.
Su Z., Townsend J.P. 2015. Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: analytical predictions of long-branch effects. BMC Evol. Biol. 15:86.
Su Z., Wang Z., López-Giráldez F., Townsend J.P. 2014. The impact of incorporating molecular evolutionary model into predictions of phylogenetic signal and noise. Front. Ecol. Evol. 2:11.
Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.
Townsend J.P., Leuenberger C. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60:358–365.
Townsend J.P., Lopez-Giraldez F. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59:446–457.
Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835–849.
Whelan S., Liò P., Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 17:262–272.
Wiens J.J., Hutter C.R., Mulcahy D.G., Noonan B.P., Townsend T.M., Sites J.W., Reeder T.W. 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. 8:1043–1046.
Wilson J.J. 2011. Assessing the value of DNA barcodes for molecular phylogenetics: effect of increased taxon sampling in lepidoptera. PLoS One. 6:e24769.
Xia X., Xie Z., Salemi M., Chen L., Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26:1–7.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.
Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125–133.
Zhao Z.-M., Zhao B., Bai Y., Iamarino A., Gaffney S.G., Schlessinger J., Lifton R.P., Rimm D.L., Townsend J.P. 2016. Early and multiple origins of metastatic lineages within primary tumors. Proc. Natl. Acad. Sci. U.S.A. 113:2140–2145.
Zwickl D.J., Hillis D.M. 2002a. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.
Zwickl D.J., Hillis D.M. 2002b. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588–598.