Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2011, **2**, 427–436 doi: 10.1111/j.2041-210X.2011.00103.x

A simple polytomy resolver for dated phylogenies

Tyler S. Kuhn[1], Arne Ø. Mooers[2] and Gavin H. Thomas[3]*

[1]*Biological Sciences and* [2]*IRMACS, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6; and* [3]*School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK*

Summary

1. Unresolved nodes in phylogenetic trees (polytomies) have long been recognized for their influences on specific phylogenetic metrics such as topological imbalance measures, diversification rate analysis and measures of phylogenetic diversity. However, no rigorously tested, biologically appropriate method has been proposed for overcoming the effects of this phylogenetic uncertainty.

2. Here, we present a simple approach to polytomy resolution, using biologically relevant models of diversification. Using the powerful and highly customizable phylogenetic inference and analysis software BEAST and R, we present a semi-automated 'polytomy resolver' capable of providing a distribution of tree topologies and branch lengths under specified biological models.

3. Utilizing both simulated and empirical data sets, we explore the effects and characteristics of this approach on two widely used phylogenetic tree statistics, Pybus' gamma (γ) and Colless' normalized tree imbalance (I_c). Using simulated pure birth trees, we find no evidence of bias in either estimate using our resolver. Applying our approach to a recently published Cetacean phylogeny, we observed the expected small positive bias in γ and decrease in I_c.

4. We further test the effect of polytomy resolution on diversification rate analysis using the Cetacean phylogeny. We demonstrate that using a birth–death model to resolve the Cetacean tree with 20%, 40% and 60% of random nodes collapsed to polytomies gave qualitatively similar patterns regarding the tempo and mode of diversification as the same analyses on the original, fully resolved phylogeny.

5. Finally, we applied the birth–death polytomy resolution approach to a large (> 5000 tips), but unresolved, supertree of extant mammals. We report a distribution of fully resolved model-based trees, which should be useful for many future analysis of the mammalian supertree.

Key-words: birth–death model, gamma, imbalance, phylogenetics, polytomy, simulation, supertree

Introduction

In phylogenetic analysis, polytomous nodes (multifurcations rather than bifurcations) can be considered 'soft' (incomplete taxonomic resolution; Maddison 1989; DeSalle, Absher, & Amato 1994) or 'hard' (multiple simultaneous splitting events; Hoelzer & Meinick 1994a,b). Unlike 'hard' polytomies that reflect the true topology, the presence of 'soft' polytomies, which represent missing or ambiguous data, will influence results from many types of phylogenetic analysis. Although some phylogenetic methods (e.g. BiSSE, Maddison, Midford, & Otto 2007) have been adapted to allow for polytomies (FitzJohn, Maddison, & Otto 2009), most methods require complete, bifurcating trees, e.g. identifying changes in diversification rates through time (e.g. with LASER, Rabosky 2006), estimating phylogenetic diversity and isolation scores (e.g. the EDGE of Existence program, Isaac *et al.* 2007), and calculating tree shapes (most indices require fully resolved trees). The problem of resolving polytomies is particularly acute when missing tips are added to phylogenies based on taxonomic information, as is frequently the practice when constructing clade-wide supertrees (see, e.g. Angiosperms, Davies *et al.* 2004; Ruminants, Hernandez & Vrba 2005; Primates, Ranwez *et al.* 2007). In doing this, many nodes are formed with no age estimates, and many additional polytomies are created.

We suggest that there are two general approaches for appropriately dealing with polytomous nodes, both of which can be implemented within a Bayesian framework. The first involves the addition of missing taxa as empty sequences at the tree inference stage where the placement of missing species can be

*Correspondence author. E-mail: gavin.thomas@bristol.ac.uk
Correspondence site: http://www.respond2articles.com/MEE/

constrained using priors on topology. Topology priors might be derived from published phylogenies or taxonomic information. This has the advantage that the full suite of Bayesian phylogenetic tools (e.g. relaxed molecular clocks, molecular evolutionary parameters, tree priors) can be readily incorporated into the tree-building process along with the missing taxa. However, many previously published supertrees, with a high proportion of polytomies and representing a significant amount of research time, cannot be dealt with in this manner. We present a simple approach suitable for application to previously published trees (particularly supertrees) by modelling diversification at polytomies in dated phylogenetic trees. At present, there is no widely accepted model-based method of dealing with soft polytomies. Current methods either involve random resolution of the tree topology (either without specifying branch lengths or with branch lengths drawn from a specific distribution) or conversely by allowing analyses to work around the effects of polytomies rather than explicitly attempting to resolve them (see, e.g. FitzJohn, Maddison, & Otto 2009). Many uses of phylogenies require branch length distribution for all tips, so we do not consider topology-only approaches here.

Several methods of assigning branch lengths exist. First, Purvis (1995) introduced an approach suggested by Sean Nee in which unknown node ages are proportional to the log of the daughter clade divided by the log of the parent clade (the LnN approach). This has been applied to dating unknown nodes within published phylogenies (Purvis 1995; Bininda-Emonds *et al*. 2007; Fritz, Bininda- Emonds, & Purvis 2009), but we note that this approach cannot properly be applied to polytomies that contain resolution nested within polytomies (see discussion in model comparison section). Second, branch lengths can be distributed evenly between the known parent age and the known daughter age (the equal splits, or EQS, approach; see Webb, Ackerly, & Kembel 2008). Third, branch lengths can be randomly assigned to the paths created during polytomy resolution. To our knowledge, this has not been published (but see, Day, Cotton, & Barraclough 2008), and we refer to two variants of the random approach: RND and RND2. The EQS, RND and RND2 approaches may be viable alternative approaches to polytomy resolution, as they do not make reference to any particular *a priori* model. However, their behaviour, inherent biases and impacts on phylogenetic inference have yet to be studied.

We propose an alternative approach that uses the constant rate birth–death model to sample from topologies and branch length distributions at polytomies, referred to herein as the BD approach. Analyses can then be applied to the resulting *pseudo-posterior* distributions of trees. This approach leverages the power of the BEAST phylogenetic inference package (Drummond *et al*. 2002; Drummond & Rambaut 2007) to explore tree space. We demonstrate the efficacy of this model-based approach to polytomy resolution by comparing its behaviour to other previously used resolution approaches, as well as through its application to both simulated and published phylogenies. We explore how the polytomy-resolved phylogenies perform in commonly used tests of lineage diversification. We also show that the method can be applied to large trees by providing a distribution of fully resolved mammal supertrees generated from a recently updated mammalian supertree containing 5020 terminal taxa but >2500 unresolved nodes (Fritz, Bininda-Emonds, & Purvis 2009).

Resolving polytomies with a birth–death model

BEAST (Drummond & Rambaut 2007) implements Bayesian approaches to phylogenetic and phylogeographic analyses. Priors can be placed on, for example, the molecular evolutionary parameters, branch rates and tree topology. A useful (though not unique) property of BEAST is that it allows sampling from the prior only and the application of prior constraints to both the tree topology and branch lengths. Importantly, BEAST does not produce negative branch lengths, a common stumbling block for many polytomy resolution approaches. This, in addition to the prior-only sampling scheme and flexible XML input language, makes BEAST particularly well-suited as a general polytomy resolution tool. Specifically, it is possible to input a partially resolved tree, where the known resolved topology and node ages are constrained and allow the Bayesian Markov chain Monte Carlo search algorithm to permute the unresolved portions of the tree based on a specific biological model, such as (but not limited to) the constant rate birth–death model.

The polytomy resolution approach presented here is comprised of two separate stages: (1) production of an XML input file containing the topology constraints and (2) model-based tree permutations in BEAST. We provide two scripts (see supplementary materials) using the library APE (Paradis, Claude, & Strimmer 2004) for the R statistical language (R Development Core Team 2010) that define topology constraints in which the dichotomous portions of the user-input tree remain fixed, leaving the polytomies free to be permuted. The first (stand-alone) script writes a complete XML input file including topology constraints and a full set of BEAST input commands, including specification of a birth–death tree prior. The second script only defines topology constraints, allowing the user to adjust the model settings either by directly editing the XML or using the program BEAUTI (Drummond & Rambaut 2007). We encourage users to take advantage of the flexibility of the Bayesian framework to explore broad but appropriate prior distributions. The specific model used will of course depend on each researcher's interests and data set. In the example scenarios, we will utilize the stand-alone BD model R script. Within the present BD script, a uniform prior is employed for both the mean growth rate ($\lambda - \mu$) and relative death rate (μ/λ) parameters. The BEAST MCMC is then used to estimate these parameters based on the distribution of constrained nodes. In some situations, it may be appropriate or required to specify different prior distributions on these parameters, or even fix their values; however, here, we wish to provide a widely applicable preset approach and demonstrate its functionality.

We stress that by using a birth–death prior to resolve polytomies, our proposed method is necessarily biased towards favouring the birth–death model in analyses of diversification.

Because most known phylogenies do not conform to a constant rate birth–death process, most applications of our approach will therefore be biased. However, we demonstrate below that the bias is predictable and, in the context of diversification analyses, conservative because constant rate birth–death is the standard null model.

Testing the approach

PROOF OF CONCEPT

We resolved a single 10-tip polytomy to ensure that the estimated model parameters (e.g. the birth and death rates) were appropriately optimized and that the resulting tree distribution conformed to a birth–death model. We suggest that, assuming convergence and mixing of relevant tree statistics (see below), a posterior distribution of 10 000 trees will generally be adequate to explore tree space. We conducted preliminary analyses (not shown) to estimate the likely burnin period and found it to be short even for large trees. Consequently, we ran analyses for 11 111 000 iterations, sampling trees every 1000 iterations with a 10% burnin to yield posterior distributions of 10 000 trees. This is the default within the stand-alone R script but can readily be changed as appropriate in either the R script or the XML file generated by the script. For all BEAST output, we assessed mixing, convergence and that 10% burnin was appropriate by visual inspection of three statistics in TRACER v1.5 (Rambaut & Drummond 2009): net diversification rate ($\lambda - \mu$), relative extinction rate (μ/λ) and root age.

We consider these three statistics to be the most relevant for determining whether tree space has been adequately sampled because they refer directly to the tree-sampling prior or to the tree structure itself. With BEAST, a particular node age will be changed in 50% of the possible move types affecting that node (Drummond *et al.* 2002), and as a result, a trace of the node's age represents a conservative estimate of the number of changes made to it. We therefore use the root node as a standard marker for all the other nodes that are being sampled (i.e. nodes involved in polytomies) by incorporating a small amount of uncertainty in the prior on root age. Similar to standard analysis in BEAST, we regard a post-burnin estimated sample size (ESS) value >200 as evidence that stationarity has been reached. We note that because it is tightly constrained, the root age may not be useful as a means of assessing convergence between independent runs. For this 10-tip polytomy, ESS values calculated in TRACER v1.5 were between 7000 and 9500 for $\lambda - \mu$, μ/λ and root age (although we note that estimated $\lambda - \mu$ and μ/λ are not independent of one another).

COMPARISON AMONG RESOLUTION METHODS

To assess the relevance of our proposed polytomy resolution approach, we compare its behaviour to that of two previously used approaches (LnN and EQS) as well as two unpublished random resolution approaches (RND and RND2). These non-model-based methods involve a two-step process of random topology resolution followed by branch length estimation. The first step is identical in all four non-model-based approaches, whereas branch length inference differs. In contrast, tree topology and branch lengths are estimated simultaneously in the BD approach. All R-scripts used for method comparisons are available from the authors upon request.

Purvis (1995) developed a method to determine unknown node ages based on the theoretical relationship between clade size and node age distribution within either the pure birth or random birth–death process (Grafen 1989; Nee in Purvis 1995). Purvis proposed the following relationship:

$$T_D = T_A \cdot \frac{\log(N_D)}{\log(N_A)} \qquad \text{eqn 1}$$

where the age of the daughter node (T_D) is proportional to the size of the daughter clade (N_D) and the size of the ancestral clade (N_A) and the age of the ancestral node (T_A). We refer to this approach as LnN approach. Although this approach has been used for providing an age estimate for undated nodes within supertrees (Bininda-Emonds *et al.* 2007), it is prone to generating negative branch lengths where there is resolved tree topology nested within a polytomy. The standard response to this is simply to place the node age equidistant between the age of the mother and daughter node. A fuller discussion of this challenge with examples is included in the supplementary materials.

Unlike the LnN approach, both the EQS and RND approaches were developed for resolving polytomies with nested constraints. For each polytomous node, the total path length from the polytomy to a constrained daughter node along a newly resolved path is divided using a broken stick method. For the EQS approach, the total path length between polytomy and constrained daughter node is split equally, assigning the length of each branch, l_b, along this path using

$$l_b = \frac{1}{n} \cdot l_T \qquad \text{eqn 2}$$

where n is the number of edges (or sticks), and l_T is total path length between the polytomy and the constrained daughter node. For the RND approach, the total path length from polytomy to constrained daughter node is split into n sections of random length where the sum of all n random sections must equal the total path length. To avoid the negative branch length issues discussed for the LnN approach, the EQS and RND approaches must estimate the path lengths of the shortest polytomy to constrained daughter node path first. If no constrained daughter nodes exist, path lengths will then be estimated sequentially starting with the path with the most new nodes.

The RND2 approach was developed for testing the single polytomy scenarios and is not easily transferrable to a nested constraint polytomy resolution application. The RND2 approach estimates edge lengths sequentially through the tree, beginning with the first node up from the root, then following that path to a tip, before returning to the next node to tip path. At each edge, a random number is drawn from a uniform 0–1 distribution. This value represents the proportion of the remaining path length that is assigned to the current edge.

We test the behaviour and inherent biases in these four approaches, as well as the BD approach, using a simplest case scenario, a single polytomy (number of taxa = 10, 100, 500) with no internal constraints and a root age of 1. We simulate trees (10 000 trees for $N = 10$; 1000 trees for $N = 100$ and 500) under each method and compare the summary tree statistic, Pybus' gamma, γ (Pybus & Harvey 2000) using the program TreeStat v1.2 (Rambaut & Drummond 2008). Distributions for these parameters are shown using a modified violin plot (Hintze & Nelson 1998; Adler 2005).

For the smallest polytomy ($N = 10$), the EQS, LnN and RND2 methods perform poorly (Fig. 1a). At this tree size, both the EQS and LnN approaches produce negatively biased γ estimates. This behaviour for the LnN approach was also noted by Vos (2006). Conversely, the RND2 approach produces positively biased γ estimates. In all cases, these biases appear to be strongly size dependant (Fig. 1b,c). The RND approach performs better at all tree sizes but shows evidence of a size-dependant bias in γ. Most significantly, this γ bias shifts from slightly negative for $N = 10$, to positive for $N = 100$ and 500. This bias is problematic for application of the RND approach to a wide variety of size varying trees, in particular for very large supertrees with $N >> 1000$. In contrast with these four approaches, the BD approach, which does produce positively biased γ estimates, shows no evidence of size dependency. This small, size-independent bias appears to be a result of the birth–death tree prior implemented in BEAST. When the single polytomy scenario is run through BEAST using a Yule Process tree prior, rather than a birth–death tree prior, there is no observed bias in γ (see supplementary materials). It is worth noting that although a single polytomy represents the simplest scenario, it is the most challenging scenario for the BD approach. This is because there are no internal node constraints providing information for estimation of the $\lambda - \mu$ and μ/λ parameters. Because of the significant size-dependant biases in the LnN and RND2 approaches, we do not attempt to modify these approaches to nested constraint polytomy scenarios.

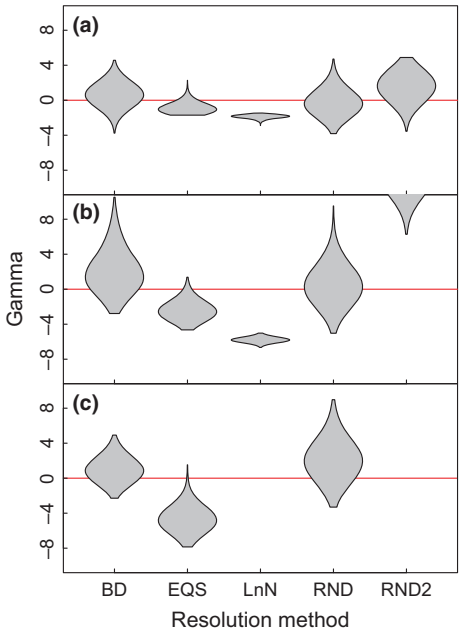

Fig. 1. Comparison between polytomy resolution methods for the simplest case scenario – a single polytomy. (a) A single 10-tip polytomy. (b) A single 100-tip polytomy. (c) A single 500-tip polytomy. Resolution methods shown are the BEAST birth–death method (BD), the Equal Splits method (EQS), the Log clade size method (Ln*N*), the Random brokenstick approach (RND) and the alternate random approach (RND2). For all methods, the expected Pybus' gamma, γ, value is shown in red [$E(\gamma) = 0{\cdot}0$]. Gamma values for the LnN and RND2 methods are outside of the plot area for the 500-tip polytomy (c). The BD approach is the only approach with a consistent, size-independent bias in γ. All other resolution methods show increased bias as the tree size increases.

SIMULATED BIRTH–DEATH TREES

To compare the behaviour of the BD, EQS and RND approaches on a more relevant scenario involving a tree with nested node constraints, we simulated two sets of 10 trees, one with 64 tips and one with 250 tips. For each data set, we randomly selected and collapsed 40% of the nodes back to polytomies. Both of these trees were simulated under a birth–death model where $\lambda = 0{\cdot}1$ and $\mu = 0{\cdot}0$ using the R package GEIGER (Harmon *et al.* 2008). Further simulations using varying λ and μ parameters are discussed below for the BD approach.

To assess the behaviour of the three polytomy approaches, we compared the original value and recovered the *pseudo-posterior* distribution for two summary tree statistics, Pybus' γ (Fig. 2, top) and Colless' normalized tree imbalance, I_c (Fig. 2, bottom; Colless 1982; Mooers & Heard 1997). Only results for the 250-tip tree are reported here; however, results from the 64-tip tree are included in the supplementary materials. For all trees, the BD method was the only approach able to reliably recover γ and I_c. The bias in I_c for EQS and RND is easy to explain: each polytomy is resolved under a Yule model of diversification, but there may be structure between it and the tips: the lineages emanating from any given polytomy may represent larger clades. After resolution, such trees will be less balanced than the Yule expectation.

Both the EQS and RND methods showed a strong negative bias in the recovered γ value. This was expected for the EQS approach, which was shown to have a strong negative bias in the single polytomy scenario. The negative bias observed in the RND approach is unexpected, and at odds with the observed positive bias in the single polytomy scenario. In this case, the bias appears to reflect the different behaviours of the RND approach when nested constraints exist. For the simple single polytomy scenarios, there was no 'shortest path' for the RND method to select first; thus, edge lengths were systematically assigned, beginning with the first node (in the cladogram) up from the root. Once edge lengths on this path were assigned; the next node-tip path was dealt with. However, in scenarios where there are nested constraints, and thus where a 'shortest path' does exist, the RND approach must assign edge lengths to that shortest path first to avoid negative branch lengths. It appears that as a result of this requirement, node ages are on

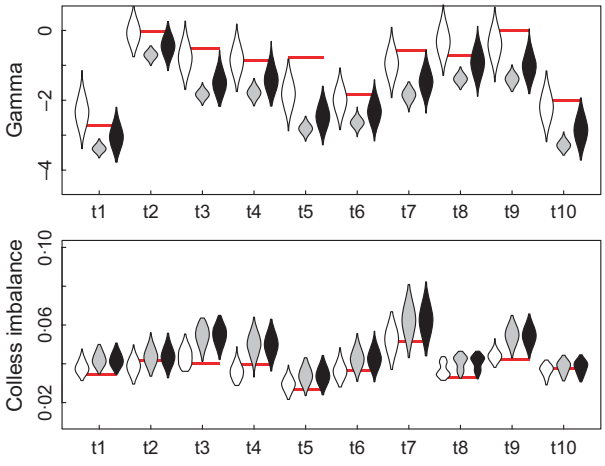

Fig. 2. Comparison of polytomy resolution methods on ten (t1, t2,…t10) independent 250-tip simulated pure birth trees where 40% of nodes have been collapsed. (White) BEAST birth–death method; (grey) Equal Splits (EQS) method; and (black) Random brokenstick (RND) approach. Initial Pybus' gamma, γ, and Colless' tree imbalance, I_c, values for the fully resolved starting tree are shown in red. The only approach where the 95% confidence interval consistently overlaps the initial value is the BD method. As expected, the EQS method significantly underestimates γ. Interestingly, the RND approach also underestimates γ, a result that is at odds with the results from the simple single polytomy scenario. With respect to I_c, both approaches implemented in R produce a noticeable increase in tree imbalance. The BD method appears much more capable of recovering the original I_c value.

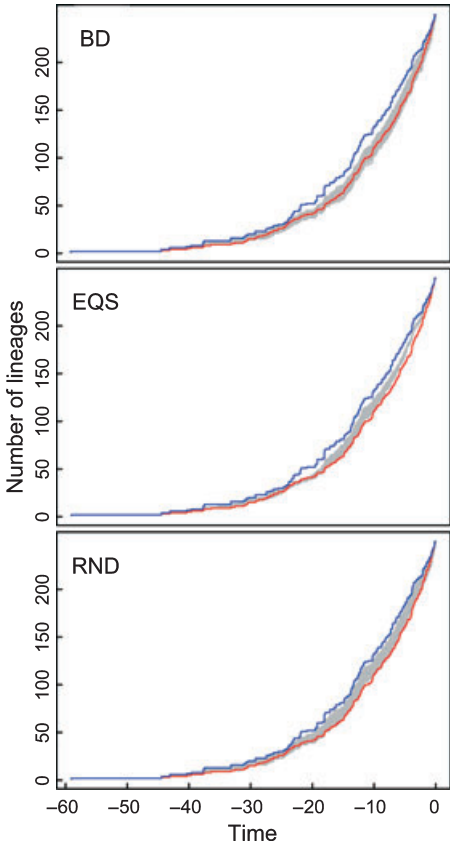

Fig. 3. Lineage through time comparison of the polytomy resolution approaches. The bottom red line in each graph depicts lineages through time for the original starting tree. The top blue line in each graph depicts lineages through time for the polytomized tree (tree 3 with 40% nodes collapsed is shown; results for other trees are similar), with the nodes all shifted towards the root. The grey lines (1000 tightly overlapping lines shown) depict lineages through time for the polytomy-resolved tree distributions. Both the EQS and RND methods show a distribution of trees with node heights intermediate between the polytomized tree in blue and the original tree, while the BD method produces a distribution centred about the original tree.

average shifted more towards the root, resulting in a negative γ bias.

We further illustrate the behaviour of these three methods by showing plots of lineages through time for one of the 10 250-tip trees (Fig. 3), comparing the original tree, the polytomized tree and the resolved distribution. Results from the other nine 250-tip trees and the ten 64-tip trees are not shown, but are consistent with the result shown in Fig. 3. The BD approach again results in a better fit of the *pseudo-posterior* distribution of branching times to the original tree (Fig. 3; top, red line). Both the EQS and RND methods do produce a noticeable shift of branching times towards the tips – i.e. towards the original tree. This is consistent with the observed negative γ bias (Fig. 2; top, grey and black distributions).

These comparisons suggest that although the EQS and RND approaches do not bias the diversification rate to any particular *a priori* model (e.g. the birth–death process), they do introduce size-dependent biases in both γ and I_c. Of particular concern regarding the RND approach is the inconsistency of the observed biases. For small single polytomy trees, the bias in γ is negative, but as the size of the tree increases, the bias becomes increasingly positive. Further complicating inference made from an RND resolved distribution, if nested constraints are present, the γ bias shifts to a negative bias. We consider this sufficient evidence to support the BD approach as the most consistently characterizable method for resolving polytomies.

We checked the behaviour of the BD approach on a suite of constant rate birth–death parameters (λ = 0·1, 0·2, 0·3 and μ = 0, 0·05, 0·09, 0·15, 0·25) by simulating multiple 64-tip 10-tree data sets using the R package GEIGER. Once simulated, for each of these 10-tree data sets, 40% of the nodes, chosen at random, were collapsed to polytomies (R script available upon request). A less conservative approach, with 20% of nodes collapsed, produced similar results but with tighter confidence limits. In addition to these 20% and 40% polytomized trees, we generated one data set of 60% polytomized trees – exceeding the amount of polytomies within the mammalian supertree. The results were not qualitatively different from those reported for the 40% polytomized trees (see supplementary materials, Fig. S5). For each of the 10-tree sets, we resolved the trees with the BD approach and recorded the estimated $\lambda - \mu$ and μ/λ parameters as recovered by BEAST (Fig. 4a,b) and the two summary tree statistics, Pybus' γ (Fig. 4c) and Colless' normalized tree imbalance, I_c (Fig. 4d).

As expected, both the λ and μ estimates from the *pseudo-posterior* distribution of resolved trees encompass the original values (λ = 0·1, μ = 0·0), and there was no bias in estimates

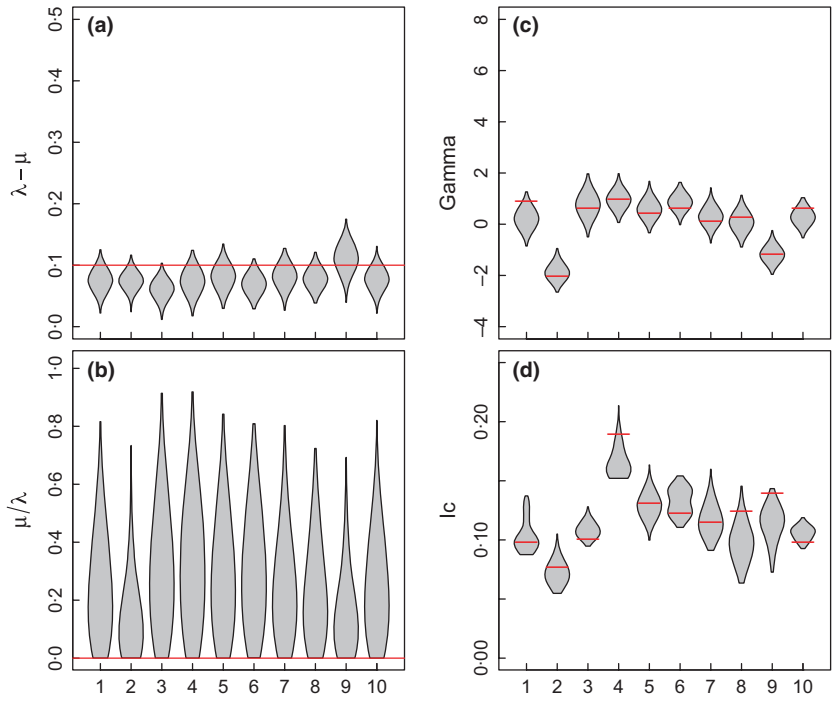

Fig. 4. Results from resolution of simulated 64-tip trees ($\lambda = 0{\cdot}1$, $\mu = 0$) with 40% of resolved nodes collapsed to polytomies. Horizontal red bars indicate the expected value for each tree. (a) estimated mean growth rates ($\lambda - \mu$) and (b) estimated relative death rates (μ/λ) produced from the 40% polytomized trees conform well with the expected values. The distribution of estimated μ/λ is right-skewed, as is common for analysis where the input μ is set to 0. (c) Values for Pybus' gamma, γ, statistic appear unbiased for resolutions of each of the 10 independent simulated trees. (d) The range of Colless' normalized tree Imbalance, I_c, for the polytomized trees consistently overlaps the score for all well-balanced starting trees indicating that the polytomization approach in BEAST is unbiased. The one exception, (tree 4), which by chance was noticeably more imbalanced than expected [$E(I_c) = 0{\cdot}1$ for $N = 64$], resulted in an expected and significantly more balanced distribution. See supplementary materials for further analysis at differing λ and μ.

of γ (pure birth simulation mean $\gamma = 0$; Fig. 4c). Similarly, I_c does not appear to change substantially between the original simulated tree (mean $I_c = 0{\cdot}1$ for $N = 64$) and resolved polytomous trees (Fig. 4d).

EMPIRICAL TEST – CETACEAN RADIATION

We tested whether our approach could have been used to capture a recently published diversification pattern. Steeman *et al.* (2009) utilized a near-complete phylogeny of cetaceans ($N = 87$) to explore competing hypotheses about the tempo of modern whale diversification. Their results supported a pulse of increased diversification related to periods of ocean restructuring, rather than an initial radiation of cetacean lineages.

We obtained the fully resolved cetacean phylogeny (provided by Dan Rabosky) and measured its shape with I_c and γ. We then produced three sets of 10 randomly 'polytomized' trees, with either 20%, 40% or 60%, of the internal nodes collapsed to polytomies, to mimic a partially resolved and dated supertree of the same group. We then resolved the polytomous nodes for each of these 30 trees using the BD approach. The birth and death rates for the original fully resolved tree were calculated with BEAST and compared to those from the BD resolved trees. The estimated net diversification rates ($\lambda - \mu$) from the BD resolved trees were similar to and unbiased from the original tree (original tree: $\lambda - \mu = 0{\cdot}0952$; mean from 20% trees $\lambda - \mu = 0{\cdot}0955$; mean from 40% trees $\lambda - \mu = 0{\cdot}0955$; mean from 60% trees $\lambda - \mu = 0{\cdot}0954$). However, estimates of the relative diversification rate (μ/λ) increased with size of polytomy (original tree: $\mu/\lambda = 0{\cdot}1400$; median from 20% trees $\mu/\lambda = 0{\cdot}1519$; median from 40% trees $\mu/\lambda = 0{\cdot}1611$; median from 60% trees $\mu/\lambda = 0{\cdot}2129$).

Comparison of the γ and I_c metrics between the original, fully resolved cetacean tree and the 30 resolved polytomy trees is presented in Fig. 5. Unlike the simulated birth–death trees discussed earlier, there is a noticeable, if non-significant, increase in the γ value, indicating a shift in internal nodes towards the tips relative to the original tree, and the resolved cetacean trees are biased to be more balanced than the true input phylogeny. Both of these patterns are to be expected, because the expected γ for a Yule tree is 0 (vs. the observed value on the original tree of $-0{\cdot}623$), and real-world trees are known to be more imbalanced (here $I_c = 0{\cdot}164$) than expected under a null model of speciation/diversification [here, $E(I_c) = 0{\cdot}08$ for $N = 87$; Mooers & Heard 1997].

We then reanalysed the 30 resolved polytomy tree distributions using the methods described by Steeman *et al.* (2009), using the R library LASER (Rabosky 2006) and additional code kindly provided by Dan Rabosky. For these analyses, we

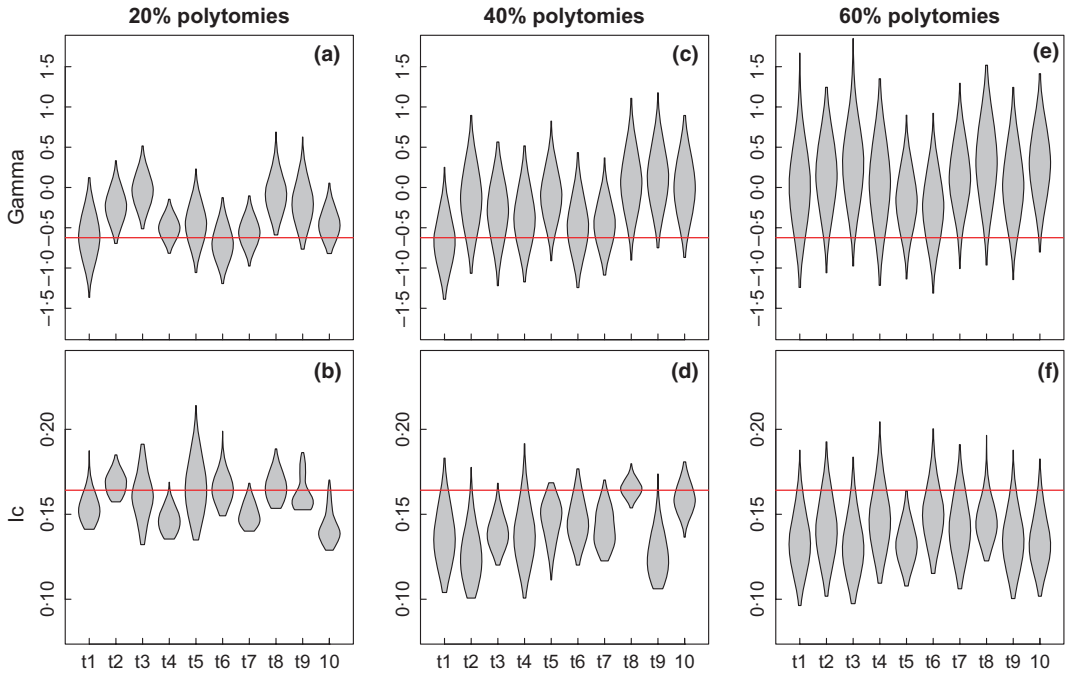

Fig. 5. Results from resolution of the cetacean phylogeny with artificially generated polytomies (at 20%, 40% and 60% of total). The results for Pybus' gamma, γ, (a, c, e) from the pseudo-posterior distribution of resolved trees appear again to suffer from a slight positive bias, although in almost all cases, the distribution of estimates overlaps the value of the fully resolved molecular phylogeny. Colless' normalized tree imbalance, I_c (b, d, f) appears to be shifted towards more balanced trees in almost all randomly polytomized tree distributions. Both of these observations are consistent with expectations from theoretical models [$E(I_c) = 0{\cdot}08$ for $N = 87$ and $E(\gamma) = 0$], where the imbalance of empirical trees is known to be more imbalanced than expected (here $I_c = 0{\cdot}164$), and the γ statistic for the fully resolved cetacean tree is lower than expected ($\gamma = -0{\cdot}623$).

combined and resampled the 10 sets of trees from the 20%, 40% and 60% resolved distributions respectively to produce three distributions of 10 000 trees. Steeman *et al.* (2009) compared the fit of seven models describing the diversification of cetaceans. These included two constant rates models (pure birth and constant rate birth–death), two models of diversity dependence (linear diversity dependence and exponential diversity dependence) and three rate shift models based on the timings of major periods of ocean restructuring. The latter three models were used to test for rate shifts at 35–31 Ma, 13–4 Ma or both periods combined. Steeman *et al.* (2009) rejected the constant rates models and favoured the combined ocean restructuring model.

Our results are consistent with those from the original Cetacean tree (Table 1). The ocean restructuring model is, on average, the best fitting model and is favoured in 99·2%, 75·5% and 58·9% of polytomy-resolved trees from the 20%, 40% and 60% polytomy tree distributions, respectively. The major change with decreasing topology resolution (from 20% polytomy through to 60% polytomy trees) is in the number of times that a constant rates model cannot be rejected. With 20% polytomies, the pure birth model is favoured in only 0·5% of trees, while it is acceptable in 20% of trees in analyses with 40% polytomies and 32% of trees with 60% polytomies. Again, this is not surprising because the polytomies are resolved using the constant rates model. Nonetheless, this suggests that resolving polytomies using constant rate birth–death models does not mask strong patterns in the data, even with a high proportion of polytomies. Where the favoured model departs from the true best model, it does so conservatively. We consider this departure conservative because the standard null hypothesis for diversification analysis is that of a constant rates pure birth or birth–death model, and the bias inherent in our approach will diminish the chances of rejecting the null hypothesis. Moreover, we note that the parameter estimates for both background and elevated diversification rates for the original tree fall well within the 95% sampling intervals of the polytomy-resolved trees. We caution that it is not appropriate to use results from the BD approach to determine whether diversification of a particular phylogeny follows a constant rates model. The BD approach may only be used for testing whether alternate diversification models better fit the data. Although this distinction is subtle, it is necessary to acknowledge the limitations of the approach. In this manner, acceptance of an alternate diversification model will be conservative.

MAMMALIAN SUPERTREE

One of the main applications of phylogenetic analysis to phylogenies with unresolved polytomous nodes is in analyses of taxonomically completed supertrees. It is important that a polytomy resolution approach is capable of dealing with the compounding issues of such large data sets. In this section, we applied our 'polytomy resolver' to the recently published

Table 1. Maximum likelihood analysis of diversification rates in complete and polytomy-resolved cetacean phylogenies. Number of parameters (k), log-likelihood (Log$L$), P (from likelihood ratio statistic for each model against the pure birth model) and Akaike Information Criterion (AIC) values are based on fitting models to the original cetacean tree. Background and elevated diversification rates (lineages/million years) are based on the original cetacean tree with 95% sampling intervals from 10 000 trees with 40% polytomies resolved in parentheses. The Best model columns are the number of times each model is the favoured best AIC model from 10 000 trees with 20%, 40% and 60% polytomies, respectively

Model	k	Log$L$	P	AIC	Best model 20%	Best model 40%	Best model 60%	Background	Elevated
Pure birth	1	22·527		−43·053	49	2008	3233	0·104 (0·100–0·112)	
Birth–death (constant rate)	2	22·527		−41·053	0	1	15	0·104 (0·095–0·110)	
Density dependent, linear	2	22·385	0·595 (0·086–0·996)	−40·770	0	2	0		
Density dependent, exponential	2	22·590	0·722 (0·543–0·993)	−41·180	0	0	0		
Ocean restructuring	2	25·465	0·015 (0·003–0·592)	−46·930	9916	7552	5895	0·081 (0·075–0·102)	0·137 (0·113–0·148)
35–31 Ma only	2	23·114	0·278 (0·091–0·838)	−42·229	35	45	740	0·102 (0·097–0·111)	0·207 (0·080–0·267)
13–4 Ma only	2	24·753	0·035 (0·009–0·795)	−45·507	0	395	117	0·085 (0·079–0·104)	0·134 (0·109–0·146)

mammalian supertree (Bininda-Emonds *et al.* 2007; Davies *et al.* 2008; as updated by Fritz, Bininda- Emonds, & Purvis 2009). This large supertree (N = 5020 tips), which represents the most complete summary of phylogenetic relationships for all mammalian taxa, is only 50% resolved. Although the possibility remains that some of these nodes may represent hard polytomies, the majority result from insufficient information.

Owing to the complexity of this analysis, with 2503 bifurcating node constraints, several important considerations needed to be addressed. The first was related to the sheer volume of information. The input file required to code 2503 constraints required more than 155 000 lines of XML code. For this reason, we recommend using the stand-alone input file generator, as some versions of the BEAUti user interface are not capable of accepting the 5020 mammalian taxa. Similarly, the tree log file output from BEAST that contained the entire 10 000 trees exceeds 3·8 gigabytes in size and cannot be opened in most graphical text editors. In addition to these logistic issues, the complexity of this analysis meant we needed to decrease the sampling frequency of the BEAST analysis. Previously, we sampled at a frequency of once per 1000 iterations, but test runs of the full supertree showed this was insufficient for this data set, and so we recorded samples every 2000 iterations. We divided the analysis into seven independent runs, each lasting between 2·5 and 5 million iterations (several analyses were cut short because of power failures). For all analyses, careful examination of parameter estimates using Tracer v1.5 indicated a burnin period of *c.* 500 000 steps was required to achieve stationarity. In all independent runs, the parameter estimates converged to similar values. When a sufficiently independent sample was suspected, we made use of the standard diagnostics available to confirm stationarity. We required ESS values for the parameters of interest, mean growth rate, relative death rate and root age (with a small amount of uncertainty incorporated into this constraint) be well above the accepted value of 200. In this case, ESS values of the final compilation ranged from 730 to 1250.

From this final compilation, we report two distributions of fully resolved 5020 tip trees; 10 000 trees, representing the full data set, and a much smaller resampled set of 100 trees. Both distributions are available in the online supplementary materials; note the full 10 000 trees are contained in an 800+ megabyte zip file. Pybus' γ and I_c values for both the 100 and 10 000 tree distributions are shown in Fig. S14 (supplementary materials). As there are no estimates of the 'true' γ and I_c values, we report only the γ and I_c distributions (mean γ 4·92 [3·85,5·96] and mean I_c 5·55 E−3 [5·35 E−3,5·77 E−3]). In addition, we include a plot of the lineages through time for the full 100 tree distribution and the unresolved mammal tree (Fig. S15; supplementary materials). As expected, the branch length distribution of the resolved mammalian supertree is markedly different from that of the unresolved tree, with branching times shifted noticeable towards the tips (Fig. S15).

Conclusion

The Bayesian polytomy resolution approach presented here has several important benefits over previous approaches. Rather than designing metrics that ignore polytomies, or approaches that address only terminal polytomies, this approach allows for inference based on a biologically relevant model-based simulation of branch lengths for all nodes within a polytomous tree, including very large supertrees. This makes it possible to utilize previously developed metrics that require fully resolved trees, without modification and with minimal violation of assumptions.

We document the behaviour and biases of several published and unpublished polytomy resolution approaches –

approaches that estimate a distribution of node ages or branch lengths. The birth–death approach (the BD approach) generally recovered the expected starting values for diversification parameters (mean growth rate and relative death rate) as well as two phylogenetic shape metrics (Pybus' γ and Colless' tree imbalance). There was a small positive bias in Pybus' γ; however, this bias and the behaviour of the BD approach showed no size dependency. Biases, notably related to phylogeny size, were detected in all other non-model-based approaches (EQS, RND) making inferences on diversification from such approaches inappropriate.

Trees based on our method are necessarily biased towards constant rate birth–death models but are generally conservative because this is the typical null model for diversification rate analyses. Even with this known and predictable bias, our analyses of cetacean diversification, with 60% of nodes in the cetacean phylogeny resolved using a birth–death model, still recovered the same best model (Ocean Restructuring) as that obtained from the fully resolved phylogeny. However, there will inevitably be a loss of power, as the proportion of polytomies increases such that constant rates models are more likely to be favoured. We typically find that polytomy-resolved trees return slightly higher (more positive) values of γ than the 'correct' fully resolved tree. Consequently, while false inference of slow-downs will be rare using our approach, some instances of real slowdowns may be missed.

Although the implementation of our approach here is based on a birth–death model, it can in principle be extended to other tree priors such as rate heterogeneous birth–death models. A particularly interesting extension would be to develop priors that build in information from species traits. For example, body size has frequently been shown to have a strong phylogenetic signal and may be informative about the possible relationships between species in unresolved parts of the phylogeny. To our knowledge, such a model has not yet been implemented in BEAST.

Acknowledgements

We thank Dan Rabosky for providing the cetacean phylogeny and R-scripts necessary to replicate their analysis, Rakesh Parhar for help with tree generation and measurement, and Karen Magnuson-Ford for scripting the RND approach we tested. We thank the SFU evolution group (FAB*) for discussion and comments. This work was supported by funding from the Canadian Natural Sciences and Engineering Research Council (NSERC Canada) and by a Natural Environment Research Council (NERC, UK) Postdoctoral Research Fellowship (grant number NE/G012938/1).

References

Adler, D. (2005) *vioplot: Violin plot. R package version 0.2.* http://cran.r-project.org/web/packages/vioplot/index.html (accessed 25 October 2010).

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. & Purvis, A. (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Colless, D.H. (1982) Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, **31**, 100–104.

Davies, T.J., Barraclough, T.G., Chase, M.W., Soltis, P.S., Soltis, D.E. & Savolainen, V. (2004) Darwin's abominable mystery: insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences*, **107**, 1904–1909.

Davies, T.J., Fritz, S.A., Grenyer, R., Orme, C.D.L., Bielby, J., Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., Gittleman, J.L. & Mace, G.M. (2008) Phylogenetic trees and the future of mammalian biodiversity. *Proceedings of the National Academy of Sciences*, **105**, 11556.

Day, J.J., Cotton, J.A. & Barraclough, T.G. (2008) Tempo and mode of diversification of Lake Tanganyika Cichlid Fishes. *PLoS ONE*, **3**, e1730.

DeSalle, R., Absher, R. & Amato, G. (1994) Speciation and phylogenetic resolution. *Trends in Ecology and Evolution*, **9**, 297–298.

Drummond, A.J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Drummond, A.J., Nicholls, G., Rodrigo, A. & Solomon, W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307.

FitzJohn, R., Maddison, W. & Otto, S. (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, **58**, 595–611.

Fritz, S.A., Bininda- Emonds, O.R.P. & Purvis, A. (2009) Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology letters*, **12**, 538–549.

Grafen, A. (1989) The phylogenetic regression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **326**, 119–157.

Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. & Challenger, W. (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

Hernandez, F.M. & Vrba, E.S. (2005) A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. *Biological Reviews*, **80**, 269–302.

Hintze, J.L. & Nelson, R.D. (1998) Violin plots: a box plot-density trace synergism. *The American Statistician*, **52**, 181–184.

Hoelzer, G. & Meinick, D. (1994a) Patterns of speciation and limits to phylogenetic resolution. *Trends in Ecology & Evolution*, **9**, 104–107.

Hoelzer, G. & Meinick, D. (1994b) Reply from G.A. Hoelzer and D.J. Melnick. *Trends in Ecology and Evolution*, **9**, 298–299.

Isaac, N.J.B., Turvey, S.T., Collen, B., Waterman, C. & Baillie, J.E.M. (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PLoS ONE*, **2**, e296.

Maddison, W. (1989) Reconstructing character evolution on polytomous cladograms. *Cladistics*, **5**, 365–377.

Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

Mooers, A.Ø. & Heard, S. (1997) Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology*, **72**, 31–54.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Purvis, A. (1995) A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **348**, 405–421.

Pybus, O.G. & Harvey, P. (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B*, **267**, 2267–2272.

R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.r-project.org (accessed 25 October 2010).

Rabosky, D. (2006) LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary Bioinformatics*, **2**, 247–250.

Rambaut, A. & Drummond, A.J. (2008) TreeStat v1.2: tree statistic calculation tool. http://tree.bio.ed.ac.uk/software/treestat/ (accessed 5 June 2010).

Rambaut, A. & Drummond, A.J. (2009) Tracer v1.5: an MCMC trace analysis tool. http://beast.bio.ed.ac.uk/ (accessed 1 December 2009).

Ranwez, V., Berry, V., Criscuolo, A., Fabre, P.H., Guillemot, S., Scornavacca, C. & Douzery, E.J.P. (2007) PhySIC: a veto supertree method with desirable properties. *Systematic Biology*, **56**, 798–817.

Steeman, M., Hebsgaard, M., Fordyce, R.E., Ho, S., Rabosky, D., Nielsen, R., Rahbek, C., Glenner, H., Sorensen, M. & Willerslev, E. (2009) Radiation of extant cetaceans driven by restructuring of the oceans. *Systematic Biology*, **58**, 573–585.

Vos, R.A. (2006) A new dated supertree of the primates. In: *Inferring Large Phylogenies: The Big Tree Problem*, pp. 94–164. PhD thesis. Simon Fraser University, Burnaby, Canada.

Webb, C.O., Ackerly, D.D. & Kembel, S.W. (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, **24**, 2098–2100.

Received 13 August 2010; accepted 3 February 2011
Handling Editor: Emmanuel Paradis

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Fig. S1. Inferred gamma bias for the simplest single polytomy scenarios using a Yule tree prior in BEAST rather than Birth-death tree prior. Two sizes of polytomies (100 tips and 500 tips) are shown. In both cases, there is no evidence of a bias in the *pseudo-posterior* distribution of gamma estimates [$E(\gamma) = 0.0$]. This is in contrast with the slight positive γ bias shown for the BD model (Fig. 1). Researchers wishing to use the Yule prior can do so by running the included 'PolytomyResolverConstraints' R script, and combining the output XML tags with a user generated BEAUti XML file where the tree prior is set to a Yule prior.

Figs S2–S4. Comparison between BD, EQS and RND approaches for simulated 64-tip trees. Three different sets of 10 trees were simulated ($\lambda = 0.1$, $\mu = 0.0, 0.05, 0.09$). None of these polytomy resolution approaches appear affected by the different birth and death rates used to simulate trees. Similar to Fig. 2 from the main text, the EQS approach has a strong negative γ bias (grey). Comparison of results for these smaller trees with the 250 tip tree results shown in Fig. 2, demonstrates the size-dependent bias observed in the RND approach (black). In Figs S2–S4, the RND approach better recovers the true γ values (red lines) than does the RND approach in the 250-tip trees (Fig. 2). The BD approach does not demonstrate any strong bias, with the 95% confidence interval overlapping the best estimate in all the 10-tree datasets. As expected there is no bias in the imbalance estimate (I_c) for any of these three 10-tree datasets (panel B).

Fig. S5. Parameters estimated from the *pseudo-posterior* distribution of trees resolved using the BD approach. Sixty percent of internal nodes chosen at random from the starting trees, the same 10-tree dataset presented in the main content (Fig. 4), were collapsed to polytomies. Similar to the 40% polytomized trees presented in the main content, the parameter estimates (grey) for the BD resolved tree distributions encompass the initial values (red bars) for all four parameters. No biases are apparent in γ or I_c. The mean growth rate ($\lambda - \mu$) does appear to be consistently underestimated. This is likely related to the challenges of estimating a relative death rate (μ/λ).

Figs. S6–S13. Parameter estimates from *pseudo posterior* distribution of 40% polytomized 64-tip simulated trees resolved using the BD approach. For Figs S6–S12, the BD approach was able to recover the original value for all four parameters (mean growth rate, $\lambda - \mu$; relative death rate, μ/λ; Pybus' gamma, γ; and Colless' tree imbalance, I_c). In Fig. S13, where the birth and death rates were high, the BD approach was not able to reliably estimate the birth and death rate parameters. However, there is again no observable bias in γ or I_c.

Fig. S14. Pybus' gamma (γ) and Colless' tree imbalance (I_c) for the mammalian supertree resolved using the BEAST birth-death (BD) method. Two pseudo-posterior tree distributions are shown, the complete set of 10 000 trees, and a subsampled set of 100 trees. Although the γ and I_c distrubtions are less smooth, there does not appear to be a difference between the full set and the subsample. The true γ and I_c values for the mammalian supertree are not known.

Fig. S15. A lineage through time plot showing both the original unresolved mammalian supertree (blue line), and the 100-tree distribution of resolved trees (grey lines, 100 tightly overlapping lines shown). It is clear from this figure that the polytomy resolution approach has a noticeable effect on the node ages, shifting nodes towards the tips as more and more polytomies are resolved. The true distribution of lineages through time is not known.

Data S1. *Pseudo-posterior* distribution of resolved mammalian supertrees. Mammalian supertree resolution was done using the stand-alone Polytomy Resolver script. This approach resolved all polytomies under a constant rates birth-death model. Both the complete distribution of 10,000 trees and a resampled set of 100 trees are available. Log files are available from the authors upon request.

Data S2. Polytomy Resolver scripts. See supplementary materials for detailed instructions on running the "PolytomyResolver.R" standalone R-script or the "PolytomyResolverConstraints.R" customizable R-script.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.