Estimating diversification rates on incompletely-sampled phylogenies: theoretical concerns and practical solutions

Supplemental Information
Jonathan Chang, Daniel L. Rabosky, Michael E. Alfaro

Contents

The taxonomic and sampling fraction equations
Comparing analyses with rate heterogeneity using PASTIS and TACT
Comparing the running time of PASTIS and TACT
Comparing stochastic polytomy resolution, the sampling fraction method, and the taxonomic method
Detecting rate heterogeneity using the sampling fraction method
References

The taxonomic and sampling fraction equations

In this section we work through the taxonomic and sampling fraction methods of computing likelihoods under simplified conditions. Although what we present here has been thoroughly described and characterized elsewhere (Bailey 1964; Raup 1985; Nee et al. 1994; Foote et al. 1999; Maddison et al. 2007; FitzJohn et al. 2009; Harmon 2018), we work through these cases for completeness. We use the notation from Harmon (2018) throughout.

Given some starting number of species $N_0$ at the start of the birth-death process $t = 0$, the probability of observing $N$ species at the end of the process is:

$$P(N|N > 0) = \sum_{j=1}^{\text{min}(N_0,N)} \binom{N_0}{j} \binom{N-1}{j-1} \alpha^{N_0-j} \beta^{N-j} \frac{\lambda e^{(\lambda - \mu) t} - \mu}{\lambda e^{(\lambda - \mu) t} - \mu}$$

where $\alpha$ and $\beta$ are defined as:

$$\alpha = \frac{\mu e^{(\lambda - \mu) t} - 1}{\lambda e^{(\lambda - \mu) t} - \mu}$$
$$\beta = \alpha \left( \frac{\lambda}{\mu} \right)$$
Simplifying this for the case where \( N_0 = 1 \), we obtain Raup (1985) equation A17:

\[
P(N|N_0 = 1) = (1 - \alpha)(1 - \beta)\beta^{N-1}
\]

Substituting \( \mu = 0 \) to assume the pure-birth condition gives:

\[
\beta = \frac{e^{\lambda t} - 1}{\lambda e^{\lambda t}}
\]

The simplified expression therefore becomes:

\[
P(N|N_0 = 1, \mu = 0) = (1 - \beta)\beta^{N-1}
\]

Note that in the notation of Raup (1985), the number of species at the start of the process \( N_0 \) is written as \( a \). We use \( N_0 \) here to avoid confusion with the common parameterization \( \epsilon = \frac{a}{\lambda} \).

For the case where the number of extant species at the start of the process \( N_0 = 2 \), the expanded expression becomes:

\[
P(N|N_0 = 2, \mu = 0) = 2\alpha\beta(1 - \alpha)(1 - \beta) + [(1 - \alpha)(1 - \beta)]^2
\]

The equations for the sampling fraction method are commonly defined as differential equations introduced by Maddison et al. (2007). These two equations are \( D_N(t) \), the probability that a lineage at past time \( t \) will be presently observed as a clade of size \( N \), and \( E(t) \), the probability that a lineage at past time \( t \) will not be observed at the present. We also need to specify the initial conditions \( D_0 \) and \( E_0 \). Because we are assuming the pure-birth condition \( \mu = 0 \), we only need to consider \( D_N(t) \); \( E(t) = 0 \) because species that exist at the start the process will never go extinct.

\[
D_N(t) = \frac{D_0(\lambda - \mu)^2e^{-(\lambda - \mu)t}}{(\lambda - \lambda E_0 + e^{-(\lambda - \mu)t}(\lambda E_0 - \mu))^2}
\]

As we are assuming \( \mu = 0 \), we can simplify this to:

\[
D_N(t|\mu = 0) = \frac{D_0\lambda^2e^{-\lambda t}}{(\lambda - \lambda E_0 + e^{-\lambda t}(\lambda E_0))^2}
\]

When sampling is not complete, the starting conditions are \( D_0 = 1 - f \) and \( E_0 = f \), where \( f \) is the sampling fraction (FitzJohn et al. 2009).

When the number of extant species at the start of the process \( N_0 = 2 \), the computation using the sampling fraction method uses the same equations as above, but squared (since there are now two tips) and multiplied by \( \lambda \), the probability of observing the initial speciation event.

**Comparing analyses with rate heterogeneity using PASTIS and TACT**

To compare how PASTIS (Thomas et al. 2013) and TACT deal with potential diversification rate heterogeneity, we simulate a phylogeny containing a single rate shift. First, we use TreeSim (Stadler 2011) to generate a pure-birth phylogeny with \( N = 100 \) tips generated over \( t = 100 \) Ma with a speciation rate \( \lambda = 0.039 \). From this tree, we
identify a clade that we will simulate a shift to an increased diversification rate with \( N = 500 \) additional tips on the rate-shifted clade.

We use TACT on the original \( N = 100 \) tip phylogeny, and specify a taxonomy file that will generate \( N = 500 \) tips in the shift clade. We deviate from the default settings of TACT, and set the minimum crown capture probability to 0 (normally 0.8). This forbids TACT from considering the stem as a valid location to place unsampled taxa, which matches the behavior of PASTIS.

As we are only examining the impact of rate heterogeneity, we use this TACT phylogeny as input to PASTIS, to ensure that the same topology is used, but allow the branching times (and therefore diversification rate) to vary from the TACT analysis. We modify the default PASTIS MrBayes template file in several ways:

- Set the extinction prior to a flat \( \text{Beta}(1, 1) \) distribution (the default fixes the extinction rate to 0). This matches the behavior of TACT which permits the relative extinction rate to be anywhere from 0 to 1.
- Time-calibrate the phylogeny by placing a fixed prior on the root and shift clade.
- We set the clock rate prior to a \( \text{Lognormal}(-7, 0.6) \) prior. As this is a time-calibrated analysis the clock prior is required.

\[
\begin{align*}
\lambda &= 0.87 \\
\mu &= 0.27 \\
\gamma &= 2.33 \\
\lambda &= 0.44 \\
\mu &= 0 \\
\gamma &= -5.13
\end{align*}
\]

Figure S1: Phylogenies and lineage-through-time plots for TACT and PASTIS. Phylogenies for TACT (a) and PASTIS (b) have the rate shift node marked with an arrow, and show the maximum-likelihood estimate of the speciation rate \( \lambda \), extinction rate \( \mu \), and \( \gamma \) statistic for the rate-shifted clade. (c) Lineage-through-time plot comparing TACT (dashed line) and PASTIS (dotted line) against a Yule pure-birth expectation (solid line). This is a reproduction of Figure 3 from the main text, included here for convenient reference.

We omitted sequence data to draw directly from the prior distribution. We ran MrBayes v3.2.7a for 100,000 generations with a sampling frequency of 1,000 for a total of 100 samples, and summarized the resultant phylogenies into a single consensus phylogeny.

The stochastic polytomy resolution phylogenies for TACT and PASTIS are shown in Fig. S1 (a, b). For the shift clade, we computed the maximum likelihood estimate of the speciation and extinction rate, and also the \( \gamma \) statistic, on both phylogenies. The \( \gamma \) statistic is similar to a one-tailed t-test and will detect a significant (\( \alpha = 0.05 \)) deceleration in diversification rate when \( \gamma < -1.645 \) (Pybus and Harvey 2000). The TACT phylogeny fails to reject the null hypothesis (\( \gamma = 2.33, p = 0.99 \)), but the PASTIS phylogeny strongly rejects the null (\( \gamma = -5.13, p < 0.001 \)), supporting decelerating diversification rates.
We also plotted the number of lineages through time (LTT) for each of the PASTIS and TACT phylogenies (Fig. S1 c). We compare these empirical LTT plots and to the LTT plot expected under a constant rate pure birth process. The TACT LTT plot tracks closely to the Yule pure birth expectation line. However, consistent with our results from the $\gamma$ statistic test, the PASTIS LTT plot suggests a pattern of more lineages appearing early in the shift clade’s history than would be expected under a Yule process.

For a more direct comparison of the estimates from PASTIS and TACT, we plot a rank-order correlation of the DR statistic, a fast approximation of the tip-specific speciation rate $\lambda$ (Jetz et al. 2012; Title and Rabosky 2019), and the node branching times for the rate-shifted clade (Fig. S2). In both comparisons, PASTIS deviates substantially from the 1:1 correlation line, as PASTIS tends to have earlier branching times and therefore slower DR estimates compared to TACT.

Overall, this simulation suggests that PASTIS is less able to accommodate an increase in diversification rate in the shift clade, and is instead using the slower tree-wide diversification rate. When rates are heterogenous on phylogenies, such as in this example, PASTIS may produce phylogenies that incorrectly suggest an “early burst” of speciation followed by a slowdown in diversification rate.

**Comparing the running time of PASTIS and TACT**

We used the `fishtree` package (Rabosky et al. 2018; Chang et al. 2019) to download phylogenies of three ray-finned fish groups: Pleuronectiformes, Carangaria, and Percomorphaceae. We did not use the “missing clades” feature of PASTIS for simplicity. We used the `omit_sequences` option to only sample under the prior, and modified the template to generate a single sample for comparison with TACT. We used the standard `time` command to compare
how long TACT and the PASTIS-formatted MrBayes file (Ronquist and Huelsenbeck 2003; Thomas et al. 2013) took to run. All comparisons were run on a 2017 iMac 27” with a 3.4 GHz Intel Core i5 processor.

Table S1: Comparison of TACT and MrBayes (PASTIS) wall-clock times to generate a single complete phylogeny for three groups of ray-finned fishes.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sampled species</th>
<th>Total species</th>
<th>Constraints</th>
<th>TACT</th>
<th>PASTIS</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleuronectiformes</td>
<td>220</td>
<td>780</td>
<td>149</td>
<td>8.75 s</td>
<td>12.53 s</td>
<td>1.43x</td>
</tr>
<tr>
<td>Carangaria</td>
<td>415</td>
<td>1057</td>
<td>224</td>
<td>16.452 s</td>
<td>52.132 s</td>
<td>3.17x</td>
</tr>
<tr>
<td>Percomorphaceae</td>
<td>6863</td>
<td>17301</td>
<td>3221</td>
<td>8.69 m over 7 d</td>
<td>over 1000x</td>
<td></td>
</tr>
</tbody>
</table>

To generate a single sample, in all cases TACT ranged from slightly faster (1.4x) to substantially faster (over 1000x; Tbl. S1). MrBayes, which PASTIS relies on to generate its phylogenies, did not begin sampling even after a week of runtime for the percomorph phylogeny, and we stopped the analysis as we did not believe it would not begin sampling in a reasonable amount of time.

We speculated that MrBayes performed poorly due to the large number of constraints imposed on its topology. To test the scaling behavior of its constraint algorithm, we removed all but one constraint from the percomorph Nexus file, and found it took about 52 minutes to generate a single complete phylogeny with a single topological constraint. A future version of MrBayes could, however, potentially improve its scaling behavior with respect to the number of topological constraints.

Table S2: Comparison of TACT and PASTIS (MrBayes) times to generate 100 complete phylogenies for three groups of ray-finned fishes. The PASTIS + MrBayes analysis was set to run its MCMC for 1 million generations, sampling every 10,000 generations. The TACT times are assuming serial execution with a single compute core. For fully parallel execution, such as on a compute cluster, the TACT timings from Tbl. S1 can be used. We did not attempt to run PASTIS for the 100 sample test as the single sample test (Tbl. S1) did not finish initializing even after a week of computation.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>TACT</th>
<th>PASTIS</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleuronectiformes</td>
<td>14.59 m</td>
<td>5.86 m</td>
<td>0.40x</td>
</tr>
<tr>
<td>Carangaria</td>
<td>27.41 m</td>
<td>14.60 m</td>
<td>0.53x</td>
</tr>
<tr>
<td>Percomorphaceae</td>
<td>14.49 h</td>
<td>did not attempt</td>
<td>—</td>
</tr>
</tbody>
</table>

To generate a distribution of complete, stochastically resolved phylogenies, PASTIS and MrBayes have the advantage of needing to set up its analysis only once (such as loading the tree and setting priors and model parameters), and thereafter drawing samples from a posterior distribution; partial likelihood calculations for analysis can be cached in a single long analysis, further speeding up runtime. TACT has no facility for reducing this setup time, so a serial analysis using TACT will have longer runtimes than PASTIS for small phylogenies (Tbl. S2). However, we note that TACT is trivially parallel: if one had access to a high-performance computing cluster, it would be straightforward to simultaneously generate 100 (or more) complete phylogenies in the same time it would take to generate a single complete phylogeny.
Comparing stochastic polytomy resolution, the sampling fraction method, and the taxonomic method

As a visual comparison of stochastic polytomy resolution, the sampling fraction method, and the taxonomic method, we generated a single 500-tip phylogeny with the parameters $\lambda = 0.04$ and $\mu = 0.001$ using the sim.bd.taxa function from TreeSim (Stadler 2011). This phylogeny was subsampled at various sampling fractions $f$ to create a set of phylogenies, consisting of 100 incomplete replicate trees. We then added tips that were dropped in the previous step using stochastic polytomy resolution to form our second set of phylogenies, consisting of 500 complete replicate trees, and computed the averaged likelihood surface for all of these phylogenies using stochastic polytomy resolution, the sampling fraction method, and the taxonomic method. This is an extended version of the same procedure described in the main text, corresponding to Figure 3 of the main text.

The sampling fraction method (Fig. S3 b–g) has a broad and flat likelihood surface as sampling decreases, while the surface for stochastic polytomy resolution (Fig. S3 i–n) remains relatively steep, similar to the complete phylogeny (Fig. S3 a). The taxonomic method’s (Fig. S3 h) likelihood surface does not vary with respect to sampling, but it has an extremely broad likelihood surface, but may be preferable to the sampling fraction method at very low sampling ($f < 0.01$, Fig. S3 g).

To test how the power to detect rate heterogeneity is sensitive to various tree parameters, such as speciation rate, extinction rate, and tree size, we used the R package TreeSim (Stadler 2011) to simulate 100 phylogenies under all combinations of the following parameters: net diversification $r \in \{0.05, 0.1, 0.25\}$, relative extinction $\epsilon \in \{0, 0.1, 0.5, 0.9\}$, and number of tips $N \in \{100, 500, 1000\}$. We only present the results for the 1,000 tip trees; the full set of simulations is available in the Dryad online supplement.

To test the sampling fraction method, we use these simulated “complete” trees and randomly drop tips under four different sampling regimes $f \in \{0.75, 0.5, 0.25, 0.05\}$ with 100 replicates for each condition. To compare against stochastic polytomy resolution, we re-add the dropped tips 100 times for each incomplete tree. This procedure generated 14,400 phylogenies with 1,440,000 simulated distributions of waiting times under stochastic polytomy resolution. We then compute the maximum likelihood estimate for the net diversification rate $r$ for both the incompletely sampled phylogeny, and the phylogeny where missing tips were re-added via stochastic polytomy resolution. As expected, the estimates of the net diversification rate for both methods are extremely tightly correlated (Pearson’s $r > 0.95$ for all combinations of simulating parameters; Fig. S4, S5, S6), suggesting that inference based on stochastic polytomy resolution does not result in biased rates relative to the sampling fraction method.

To determine the range of plausible support for alternative rate configurations, we first compute the maximum likelihood estimate for the rate parameters $r$ and $\epsilon$. We then search the likelihood surface of $r$ along the half-open intervals $(0, r_{MLE})$ and $[r_{MLE}, 4)$ to detect the point where the log-likelihood decreases by $m = 3$. This $m$-unit support measure, first introduced by Edwards (1972), is equivalent to an Akaike information criterion $\Delta AIC = 2$ with a difference in $k = 2$ parameters, in this case the location of a proposed rate shift and its new speciation parameters. We only examine the net diversification rate here, as estimating extinction rates can be difficult on molecular phylogenies (Rabosky 2009). Moreover, computing the covered area in 2-dimensional parameter space enclosed by the $m = 3$ boundary would be computationally infeasible for our over 1 million simulations, as this is equivalent to numerically identifying all roots in the complex plane.
Figure S3: Likelihood surfaces for a simulated 500-tip phylogeny. (a) Surfaces for the complete phylogeny at sampling fraction $f = 1$, (h) under the taxonomic method, (b–g) using the incomplete sampling method $f \in \{0.9, 0.75, 0.5, 0.25, 0.1, 0.01\}$, and (i–n) using complete stochastically resolved phylogenies are provided. In all panels, the thick black line corresponds to the $\Delta\text{AIC} = 2$ boundary that encloses the maximum likelihood estimate of the speciation and extinction rates. The contour lines of equal likelihood are also identical across all panels. This is an expanded version of Figure 3 in the main text.
Figure S4: Correlation plot of the estimates of the net diversification rate \( r \) for stochastic polytomy resolution compared to the sampling fraction method, where the generating net diversification rate \( r = 0.05 \) for an \( N = 1000 \) tip phylogeny. Columns in the panel grid are different values of the relative extinction rate \( \epsilon \), while rows in the panel grid are different sampling fractions \( f \). The red dashed line is the 1:1 correlation.
Figure S5: Correlation plot of the estimates of the net diversification rate $r$ for stochastic polytomy resolution compared to the sampling fraction method, where the generating net diversification rate $r = 0.1$ for an $N = 1000$ tip phylogeny. Columns in the panel grid are different values of the relative extinction rate $\epsilon$, while rows in the panel grid are different sampling fractions $f$. The red dashed line is the 1:1 correlation.
Figure S6: Correlation plot of the estimates of the net diversification rate $r$ for stochastic polytomy resolution compared to the sampling fraction method, where the generating net diversification rate $r = 0.25$ for an $N = 1000$ tip phylogeny. Columns in the panel grid are different values of the relative extinction rate $\epsilon$, while rows in the panel grid are different sampling fractions $f$. The red dashed line is the 1:1 correlation.
Figure S7: Comparison of the plausible support ranges of the net diversification rate $r$ for phylogenies simulated with different values of $r$ (top, middle, and bottom panels) using stochastic polytomy resolution (solid black line) and the sampling fraction method (dashed gray line). The endpoints of these ranges conceptually represent the difference in rates needed to support an alternate rate configuration using a $\Delta \text{AIC} = 2$ condition with a difference in $k = 2$ parameters.
We compared the range of plausible net diversification rate \( r \) under the sampling fraction method using the incomplete phylogenies, and under stochastic polytomy resolution using the simulated distributions of waiting times (Fig. S7). In all cases, the plausible range for \( r \) using stochastic polytomy resolution (solid line) was narrower than the plausible range using the sampling fraction method (dashed line), suggesting increased power to detect potential rate heterogeneity when using stochastic polytomy resolution.

These analyses were performed on various computing clusters (see Acknowledgements in main text) and took approximately 2 CPU-years of compute time.

**Detecting rate heterogeneity using the sampling fraction method**

To demonstrate that the broader range of supported rate estimates (Fig. S7) can fail to detect rate heterogeneity, we create a simple test case using a popular method for detecting rate shifts on phylogenies, Bayesian Analysis of Macroevolutionary Mixtures (BAMM; Rabosky 2014). We stress that this is not an issue specific to BAMM, but is shared across all methods that estimate diversification rates on incomplete phylogenies. BAMM, as one of the more popular methods, serves as a useful method to demonstrate the downstream impacts of inference on phylogenies obtained using stochastic polytomy resolution versus ones analyzed with the sampling fraction method. These issues we identify could apply equally to other methods such as BiSSE and RPANDA (Morlon et al. 2016).

We generated two topologies for testing: the first, for the sampling fraction method, is a rooted three tip phylogeny, where one of the tips has a sampling fraction specified to represent a total diversity of \( N = 1000 \) species. The second, for stochastic polytomy resolution, is the same phylogeny, but where one of the tips has been replaced with a simulated \( N = 1000 \) tip phylogeny, for a total of 1,002 tips observed in the tree. In all cases, the root age of the phylogeny is 20 Ma, and the crown age of the diverse clade is 10 Ma. For both analyzed phylogenies, BAMM should place a rate-shift event for a higher speciation rate on the diverse clade, as it is fairly obvious that the diverse clade should have a much faster speciation rate relative to the rest of the tree.

To analyze these phylogenies, we ran BAMM with all default parameters, except for:

- `expectedNumberOfShifts = 0.1` to permit Bayesian odds ratio testing
- `lambdaIsTimeVariablePrior = 0` disables time-variable BAMM, matching the generating model
- `muInit0 = 0` sets the initial extinction parameter \( \mu \) to 0
- `updateRateMu0 = 0` forbids changes to the extinction parameter from its initial state, ensuring we only estimate pure-birth models

To aid with convergence we also set the following parameters:

- `initialNumberOfEvents = 50` places a large number of rate shift events on the starting tree
- `numberOfChains = 4` enables Metropolis coupling

For the full stochastic polytomy resolution phylogeny, we set `useGlobalSamplingProbability = 1` and `globalSamplingFraction = 1.0`. For the three tip sampling fraction phylogeny, we instead generate a sampling probability file that specifies the percentage of sampled diversity that each tip represents. Specifically: the “diverse” tip \( f = 0.001 \), the “depauperate” tip \( f = 1 \), the diverse + depauperate clade \( f = 0.001998002 \), and the entire tree (diverse + depauperate + outgroup) \( f = 0.002997003 \).
We ran Bamm for each tree for 1 million MCMC generations, sampling every 1,000 generations, for a total of 1,000 posterior samples, and discarded the first 10% of samples as burnin. We used the coda R package (Plummer et al. 2006) to diagnose convergence, ensured the number of effective sample sizes was greater than 200, and examined the trace plot to check that chains appeared to reach stationarity.

![Figure S8: BAMM phylorate plots for a complete phylogeny, representative of an analysis using stochastic polytomy resolution (left panel) and an incomplete phylogeny, representative of an analysis using the sampling fraction method (right panel). The best shift configuration obtained using Bamm placed a rate shift at the base of the clade consisting of \( N = 1000 \) species; however, the same Bamm analysis using the sampling fraction method did not support a rate shift on the exemplar branch representing the same amount of diversity. The color scale for speciation rate \( \lambda \) is identical in both phylorate plots.]

For the full phylogeny representing the stochastic polytomy resolution method, the 95% credible set of rate configurations all included one or more rate shift to a higher speciation rate regime. The best shift configuration placed a single rate shift at the root of the diverse clade on the phylogeny (Fig. S8, left phylogeny). In contrast, the three tip phylogeny representing the sampling fraction method did not support any rate shifts in the 95% credible shift set (Fig. S8, right phylogeny), despite both analyses representing the exact same pattern of diversification. We conclude that the sampling fraction method has reduced power to detect rate heterogeneity even when that heterogeneity is fairly extreme.

**References**


