Testing Multispecies Coalescent Simulators using Summary Statistics

Elizabeth S. Allman, Hector Baños, and John A. Rhodes

Abstract—As genomic-scale datasets motivate research on species tree inference, simulators of the multispecies coalescent (MSC) process have become essential for the testing and evaluation of new inference methods. However, the simulators themselves must be tested to ensure that they give valid samples. This work develops methods for checking whether a collection of gene trees is in accord with the MSC model on a given species tree. When applied to well-known simulators, we find that several give flawed samples. The tests presented are capable of validating both topological and metric properties of gene tree samples, and are implemented in a freely available R package `MSCsimtester` so that developers and users may easily apply them.

Index Terms—multispecies coalescent model, incomplete lineage sorting, MSC simulators

1 Introduction

Simulation software plays an important role in the development of phylogenetic methods, providing our only means of 1) verifying that inference software performs properly on large scale data, and 2) comparing the performance of different inference methods. Although simulation studies are often performed under conditions in which model fit is much better than is likely for empirical data, these studies are nonetheless essential to methodological progress.

With increasing attention to the analysis of large multilocus datasets for which incomplete lineage sorting (ILS) may have led to discordant gene trees, simulators of the *multispecies coalescent (MSC) model* of ILS have become one component of the simulation pipeline. Although there are many causes of gene trees discordance other than ILS, the MSC model is sometimes viewed as the null model — to be considered before further complications such as hybridization, lateral gene transfer, gene duplication and loss, and/or population structure are invoked [1]. Inference of a species tree under the MSC can now be performed in a variety of ways [2]–[7], and methods continue to be developed and refined.

In outline, a large-scale simulation study of species tree inference methods might begin with choices of one or more fixed species trees, with branch lengths in generations, and population sizes for each branch. The number of taxa on the trees may range from small (say eight) to quite large (thousands, e.g. [6]), with anywhere from hundreds to thousands of gene trees being simulated. Sequences are then simulated on each gene tree, forming the simulated data to be analyzed. While sequence simulation software has been well vetted over the many years of development of gene tree inference methods, the same is not true of simulators producing gene trees under the MSC, and it is this component of the pipeline on which we focus.

Unfortunately, testing an MSC simulator for correctness is not simple, as the theoretical distribution of gene trees the model produces is quite complex. For instance, under the MSC on any species tree, each possible gene tree topology has positive probability, with the full probability density having a quite complicated dependence on the species tree topology, edge lengths, and population sizes. Straightforward comparison of a sample of gene trees to the theoretical distribution is simply not a practical approach. This has led some of the most careful developers to validate their software by comparing its output to that of other simulators, rather than to theoretical predictions [8]. For other MSC software, we have been unable to find any information on testing. Due to the difficulty of forming a coherent understanding of a large sample of gene trees, even knowledgeable users may not be able to spot simulation flaws, and are left to trust that the software does what is claimed.

Here we introduce several testing tools, based on theoretical distributions of summary statistics that capture either metric or topological information on the species tree parameter. Our first test is based on the distribution of pairwise distances on gene trees as developed in [9], and the second on counts of rooted triples on gene trees generated under the MSC [10]. Implemented in a freely available R package, `MSCsimtester`, these tests can be applied to gene tree samples from any simulator to study whether its output is in accord with the MSC. Although examining such summary statistics cannot give an ironclad guarantee of correctness, we believe they are likely to uncover most problems.

We applied our `MSCsimtester` tools to output from five well-known MSC simulators on a species tree: `SimPhy` [8], `Phybase` [11], `Hybrid-Lambda` [12], `Mesquite` [13], and `DendroPy` [14]. See Section 4 for the criterion employed in making our choices. Our goal in this short note is not to check the accuracy of all existing MSC simulators, but to introduce novel methodology for testing them and future ones.

Our tests discovered (initially) that only two of the five simulators, `SimPhy` and `DendroPy`, behave as expected un-

- *E. Allman and J. Rhodes are with the University of Alaska Fairbanks, USA. Email: e.allman@alaska.edu, j.rhodes@alaska.edu*
- *H. Baños is with Dalhousie University, CA. Email: hbanos@dal.edu*

Manuscript received XXX; revised YYY.

der the MSC. In fact, Phybase also generates valid samples under the MSC, but its documentation was not sufficiently explicit on specifying its input, and our initial interpretation of the manual was incorrect. Our tools enabled us to catch and diagnose this user error, and the manual has since been updated after correspondence with the developer. A fourth simulator, Hybrid-Lambda [12] passes our topological tests but samples metric gene trees incorrectly. After this was uncovered in a preliminary version of this work, it has been documented on the Hybrid-Lambda website. The last, Mesquite [13], produces samples with neither gene tree topologies nor metric properties in accord with the MSC model. To the best of our knowledge, this has not been publicized. 5 While we notified the authors of these simulators of the problems in advance of this publication, the larger community should be aware of the need to interpret results of previous simulation work with them cautiously. We strongly suggest that users of other simulators, and developers of new ones, test them with the MSCsimtester tools, which are available at the Comprehensive R Archive Network (CRAN).

2 Background and Testing Methods

We give an informal description of the MSC model, to motivate and introduce the summary statistics upon which we focus.

Suppose first that 3 taxa are related by the species tree with topology $((a, b), c)$. Acknowledging that species are composed of populations, we depict this by a tree whose edges are 'pipes' as in Figure 1. The length of each pipe is elapsed time measured in generations, and the width of the pipe represents population size, which may vary over time and edge. When individual genes are sampled from the leaves of the species tree, they trace backwards in time within the species tree until they coalesce at a common ancestral individual. Coalescence is a random process, which informally can be described as individual genes lineages choosing their 'parent' uniformly at random from those existing in the previous generation, a panmictic viewpoint. Thus the only population detail of importance under the MSC is size. Importantly, there is a greater chance of coalescence when populations are small, but no requirement that lineages coalesce within any specified finite time.

Simplifying assumptions on population sizes are often made by modelers and programmers, such as that all populations at all times and on all edges throughout the tree are a constant N. More realistic is to at least allow different population sizes N_e for each edge (pipe) e of the species tree. While it would be highly desirable to be able to simulate gene trees under the MSC using arbitrary population size functions $N_e(t)$ varying with time, current simulators do not make this easy on a large tree. Nonetheless, our testing framework could accommodate that generality.

When gene trees are produced under the MSC they have two characteristics that can be viewed somewhat separately. One is the metric information, which is reflected in the distribution of pairwise distances between two fixed taxa across the gene trees. The second is topology, which is reflected in the distribution of rooted triple trees on three fixed taxa displayed on gene trees. In testing the performance of a simulator, it is essential that both metric and topological properties of a sample be examined.

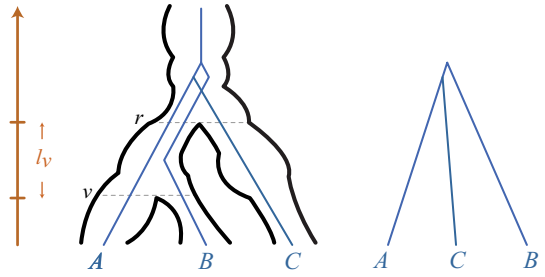

Fig. 1. (Left) A metric species tree $((a{:}\ell_a, b{:}\ell_b){:}\ell_v, c{:}\ell_c)$ drawn in black, with population sizes depicted by the widths of edges. Time is measured in generations before the present, and the vertices v and r are labeled, showing where the populations a and b merge, and further in the past where the populations ab and c merge. The vertex v is the *most recent common ancestor* of species a and b, $MRCA(a, b)$, and $r = MRCA(ab, c)$. With one lineage A, B, C sampled from each species, $A \in a, B \in b, C \in C$, a (blue) metric gene tree depicting ancestral lineages forms within the species tree. (Right) The same metric gene tree depicted more simply.

2.1 The distribution of pairwise distances on gene trees

Consider now the species tree $((a{:}\ell_a, b{:}\ell_b){:}\ell_v, c{:}\ell_c)$ shown in Figure 1, with root r and v the most recent common ancestor of a and b. For simplicity, fix constant population sizes N_v on the edge above v, and N_r above the root r. Tracking the lineages of two genes A and B sampled from a and b backwards in time, they cannot coalesce until they reach the population above v. On the edge (population) above v coalescence occurs by a Poisson process at a constant rate $1/N_v$. Thus the time T to coalescence above v is exponentially distributed for those times more recent than the root of the tree, $T \in (0, \ell_v]$. At the root r of the tree there is an instantaneous change in the coalescence rate, to the value $1/N_r$, and the coalescent process begins afresh. For the purpose of illustration, assuming that $N_v > N_r$, then coalescence above the root r occurs at a faster rate than below. Piecing together the 3 regions analyzed here (below v, between v and r, above r), the density for the pairwise distance between A and B on gene trees is a piecewise exponential, such as that shown in Figure 2 for $N_v = 2000$ and $N_r = 1000$.

Note that the discontinuities in the density in Figure 2 occur due to 1) the impossibility of coalescence below v, and 2) the discontinuity in population size at r. For larger trees, with more edges leading from the MRCA of a pair of taxa to the root, discontinuities in the density arise whenever there is a discontinuity in the population size. More generally, note that if the population sizes vary along edges with time, then the distribution need not be piecewise exponential, but is computable from the population size functions $N_e(t)$. Finally, there is no reason to require that the species tree be ultrametric, as edge lengths are in generations, and generation time may vary on different branches. The precise form of the pairwise distance density, and its derivation, is given in Section 3, following work in [9].

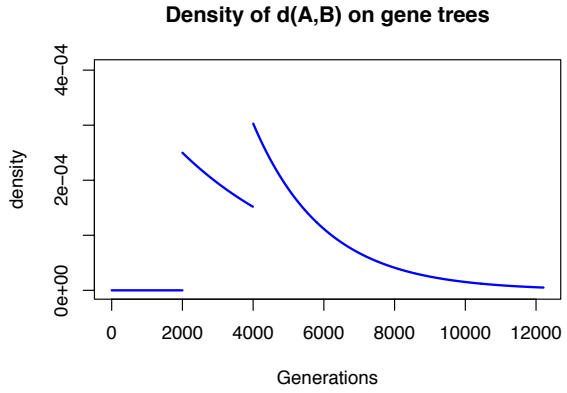

Fig. 2. The plot of the probability density function of $d(A, B)$ on gene trees for the species tree shown in Figure 1, with all edge lengths equal to 1000 generations, and population size parameters $N_v = 2000$ on the internal edge and $N_r = 1000$ for the population ancestral to the root.

2.2 The distribution of rooted triple topologies on gene trees

To test topological features of a sample of gene trees simulated under the MSC, we use the frequencies of rooted triples. For the species tree of Figure 1, again suppose the population sizes N_v and N_r are constant. Then the probability that lineages from a and b fail to coalesce in the edge of length ℓ_v between v and r is e^{-x}, where $x = \ell_v/N_v$. The quantity x here is the length of the edge in *coalescent units*, a unit convenient for addressing the confounding effects of population size and time. If the lineages fail to coalesce before the root, then lineages from a, b, c will all be present above r, and all three rooted gene topologies $((A, B), C)$, $((A, C), B)$, and $((B, C), A)$ are equally likely to form. This leads to the rooted triple probabilities

$$\begin{aligned} \mathbb{P}\left(((A,B),C)\right) &= 1 - \frac{2}{3}e^{-x}, \\ \mathbb{P}\left(((A,C),B)\right) &= \frac{1}{3}e^{-x}, \\ \mathbb{P}\left(((B,C),A)\right) &= \frac{1}{3}e^{-x} \end{aligned} \tag{1}$$

which were derived by [10]. More general formulae, accommodating larger trees and changing population sizes are derived in Section 3.

2.3 Metric and topological tests

To assess the accuracy of any MSC simulator, one first produces a large sample of gene trees from a fixed metric species tree with population size parameters. For testing using pairwise gene tree distance densities, after choosing some pair of taxa, a histogram of the pairwise distances between these taxa across the gene trees can be compared to the theoretical density. This comparison can be done visually, as major deviations from the theoretical predictions will be obvious. Additionally, one can perform a statistical test, such as that of Anderson and Darling [15], to compare the empirical distribution from the simulation to the theoretical one, giving a p-value to quantify the fit. Empirical cdfs for the Anderson-Darling test results can then be compared to the expected uniform cdf for the p-value distribution, $U(0, 1)$.

For verifying topological accuracy, one begins by tabulating the frequencies of the three rooted triple topologies displayed on the simulated gene trees for any (all) choice(s) of three taxa. The simulation can then be tested in two ways: First, one judges the fit of empirical frequencies to the expected ones using Pearson's chi-squared test to obtain a p-value. Second, one finds the maximum likelihood estimator of the internal branch length x from the tabulated frequencies, and compares this to the true value computed using Equations (1) or their analog for larger trees. The first test is a more formal test and has the advantage of being sensitive to imbalances between the counts for the two topologies incongruent with the species tree.

Both the metric and topological tests we propose require choosing a subset of two or three taxa. While one can apply the tests for all pairs and triples, we caution that in performing all such tests on the same data well known statistical issues arise in interpreting results of the multiple comparisons. Although methods such as the Holm-Bonferroni [16] could be applied to give conservative versions of such a family of tests, we do not believe such a rigorous approach is necessary given the goal of testing a simulator. Indeed, as our analyses of particular simulators have shown, choosing a pair of taxa connected by a many-edges path in the species tree, or a rooted triple of taxa whose internal edge and/or path from the most recent common ancestor to the root is composed of many edges in the species tree, quickly indicates problematic behavior.

3 Derivations of Summary Distributions

Here we derive formulas for the distributions of pairwise distances and rooted triples displayed on gene trees under the MSC, for a metric species tree with population sizes specified for each edge.

Let $(S, \{\ell_e\}, \{N_e\})$ be a metric species tree with population size functions, where each edge e has length ℓ_e and population size $N_e : [0, \ell_e) \to \mathbb{R}^{>0}$. Here $N_e(t)$ denotes the population size for a haploid organism t generations above the child node of e. There is also an 'above the root' population size function $N_r : [0, \infty)$. (For diploid taxa, the population sizes should be doubled). For technical reasons, we assume $1/N_e(t)$ is integrable on finite intervals.

3.1 Pairwise distance distribution

Let v be the most recent common ancestor of taxa a and b (that is, the node on S where A and B lineages enter the same population for the first time), and let P_a be the path in S from a to v, P_b be the path in S from b to v, and P_v be the path in S from v to the root r. Then $P_v = (e_1, e_2, ..., e_k)$, where v is incident to e_1 and r is incident to e_k. Finally, let $g_a = \sum_{e \in P_a} \ell_a$ and $g_b = \sum_{e \in P_b} \ell_b$ be the number of generations from a and b, respectively, to v. Then the distance $d(A, B)$ is a random variable

$$Y = g_a + g_b + 2X,$$

where X is the random variable giving the time to coalescence of two lineages at v. Let $c(x)$ be the probability density function for X.

To compute $c(x)$, let $N^* : [0, \infty) \to \mathbb{R}^{>0}$ be the piecewise 'union' of the N_e for $e \in P_v$ and N_r, which with $m_0 = 0$, $m_j = \sum_{i=1}^{j} \ell_{e_i}$ for $1 \leq j \leq k$, $m_{k+1} = \infty$, and N_{k+1} the population function ancestral to the root is given by

$$N^*(x) = N_{e_i}(x - m_{i-1})$$

for $x \in [m_{i-1}, m_i)$, $1 \leq i \leq k+1$. Since the coalescent process for two lineages in the same population of size $N^*(x)$ occurs with instantaneous rate $1/N^*(x)$, the probability density function is [9]

$$\begin{aligned} c(x) &= \frac{1}{N^*(x)} \exp\left(-\int_0^x \frac{1}{N^*(\tau)} d\tau\right) \\ &= \left(\prod_{j=1}^{i-1} \eta_j\right) \frac{\exp\left(-\int_0^{x-m_{i-1}} \frac{1}{N_{e_i}(\tau)} d\tau\right)}{N_{e_i}(x - m_{i-1})}, \end{aligned}$$

for $x \in [m_{i-1}, m_i)$, where

$$\eta_i = \exp\left(-\int_0^{\ell_i} \frac{1}{N_{e_i}(\tau)} d\tau\right)$$

is the probability that 2 lineages entering edge e_i fail to coalesce on it.

Since $X = \frac{Y - g_a - g_b}{2}$, setting $g_{ab} = g_a + g_b$ this shows the probability density function for Y is

$$f(y) = \begin{cases} 0 & \text{for } y \leq g_{ab}, \\ \left(\prod_{j=1}^{i-1} \eta_j\right) \frac{\exp\left(-\int_0^{\frac{y - g_{ab} - 2m_{i-1}}{2}} \frac{1}{N_{e_i}(\tau)} d\tau\right)}{2N_{e_i}(\frac{y - g_{ab} - 2m_{i-1}}{2})} & \text{for } \begin{cases} g_{ab} + 2m_{i-1} \leq y < g_{ab} + 2m_i, \\ 1 \leq i \leq k, \end{cases} \\ \left(\prod_{j=1}^{k} \eta_j\right) \frac{\exp\left(-\int_0^{\frac{y - g_{ab} - 2m_k}{2}} \frac{1}{N_r(\tau)} d\tau\right)}{2N_{e_r}(\frac{y - g_{ab} - 2m_k}{2})} & \text{for } g_{ab} + 2m_k \leq y. \end{cases}$$

In the special case that a population size function $N_e(t)$ is constant, this shows that the corresponding piece of f is a shifted, scaled, and possibly truncated exponential density.

3.2 Rooted triple frequencies

Suppose the rooted triple $((a, b), c)$ is displayed on S, and let $P = (e_1, e_2, \ldots e_i)$ denote the path from the most recent common ancestor of a, b on S to the most recent common ancestor of a, b, c. With the notation of the previous subsection, the probability that A and B lineages fail to coalesce within P is $\prod_{j=1}^{i} \eta_j$. Note that the gene triplets $((A, C), B)$ and $((B, C), A))$ can only form if the A, B lineages do not coalesce on P. Moreover, if A, B have not coalesced on P then by the exchangeability of lineages in the same populations under the MSC, the probability that any particular pair of A, B, C coalesce first is $1/3$. Thus

$$\mathbb{P}\left(((A, C), B)\right) = \mathbb{P}\left(((B, C), A)\right) = \frac{1}{3} \prod_{j=1}^{i} \eta_j.$$

Since the probabilities of the three possible topologies sum to 1,

$$\mathbb{P}\left(((A, B), C)\right) = 1 - \frac{2}{3} \prod_{j=1}^{i} \eta_j.$$

In the special case of constant population sizes $\eta_j = \exp(-\ell_{e_i}/N_{e_i})$. More generally, the length of e_i in coalescent units is $\int_0^{\ell_i} \frac{1}{N_{e_i}(\tau)} d\tau$, and $\prod_{j=1}^{i} \eta_i = \exp(-x)$ where x is the length of P in coalescent units.

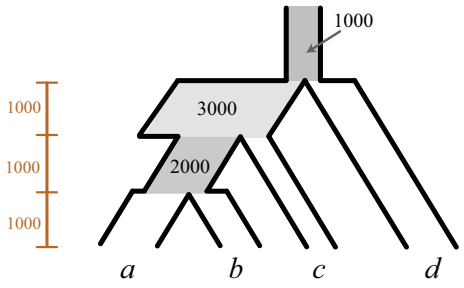

Fig. 3. 4-taxon metric species tree, with constant population sizes on each internal edge, for which data was simulated using the parameter values shown.

4 Simulations and Results

Using both metric and topological tests we investigated five popular simulators: `SimPhy` [8], `Phybase` [11], `Hybrid-Lambda` [12], `Mesquite` [13], amd `DendroPy` [14]. These simulators were chosen since they 1) allow input of a species tree, either in Newick notation or graphically, with edge lengths in generations, and 2) allow a population size N_e to be assigned independently to each edge e in the species tree. We consider these minimal requirements for a simulator appropriate for large scale simulations studies involving the MSC. All these software packages have many functionalities beyond MSC simulation, but our focus is restricted to the accuracy of MSC samples.

We also investigated the performance of the well-established and highly flexible `ms` [17] using a small species tree. However, for simulating from the MSC on a tree, `ms` requires careful conversion of a Newick-formatted tree for input, which violates our first criterion. This conversion, when performed manually, is prone to error, and therefore highly undesirable for many-taxon simulations. Although `PhyloNet` [18] provides a tool for inputting a species tree and then calling `ms` to generate a gene tree sample, its functionality is restricted to a constant population size on the entire species tree, violating our second criterion. On a manually produced small tree with varying population sizes, however, no aberrations were found in the gene tree samples `ms` produced (results not shown), as we expected.

We performed tests with a number of species trees, using a variety of choices of constant population sizes per

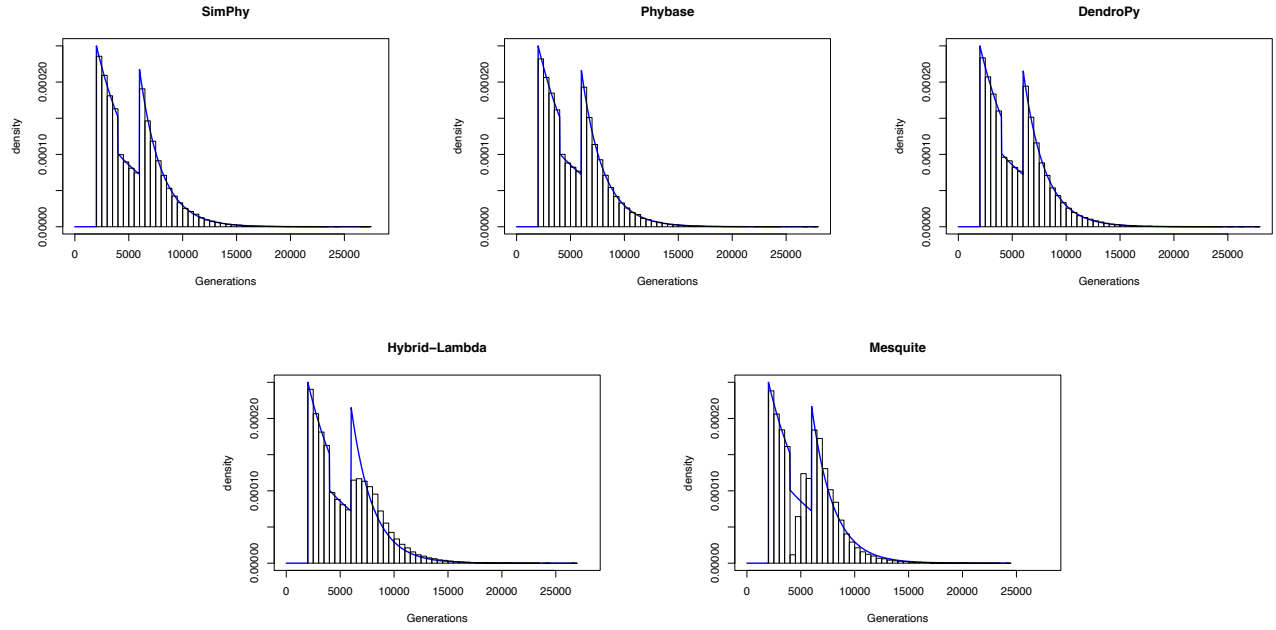

Fig. 4. Pairwise distance distributions for $d(A, B)$ from 100,000 simulated gene trees for the species tree and population sizes shown in Figure 3. Cumulative distribution functions are shown in the Supplement, Figure F2.

edge. Constant population sizes were used since none of the simulators implement time-varying population sizes. Here we show only representative examples of this work, using the species tree and populations depicted in Figure 3. Additional test results, on trees with up to 6 taxa, are shown in the supplementary materials, in Figures F2-F23 and Tables T1-T4. Samples of 100,000 gene trees were simulated with `Mesquite`, `Phybase`, `Simphy`, `DendroPy`, and of size 99,999 (its maximum) for `Hybrid-Lambda`.

For assessing the accuracy of metric features of the sample, the values of $d(A, B)$ on the gene trees were extracted, and a histogram was produced. These are shown in Figure 4, with the theoretical distribution superimposed on them. There is a good match for `SymPhy` and `DendroPy`, as was seen in all our simulations with these packages. When our initial `Phybase` simulation did not match expectations, we learned that species tree edge lengths should be supplied as μt, where t is in generations and μ is a mutation rate, while population sizes should be specified as $\theta = 4\mu N_d$ where N_d is diploid population size. Taking $\mu = 1$ and $\theta = 2N$ with N the haploid population size, `Phybase`'s sample matched expectations well. Both the `Hybrid-Lambda` and `Mesquite` simulations show pronounced deviations from the theoretical distribution. However, when the input tree is specified with branch lengths in coalescent units, `Hybrid-Lambda` correctly gives a sample of gene trees whose pairwise distances in coalescent units match the theory well (results not shown); the poor fit occurs only when species tree is entered in numbers of generations and the branch-specific population sizes are entered separately.

To quantify deviation of the sampled $d(A, B)$ density from the expected distribution, we apply the Anderson-Darling test. As is well known, even the small numerical errors arising from computer round-off in a simulation or its analysis may prevent the extremely close fit to theory that a large sample should exhibit. As a result, with very large samples such tests can produce misleadingly small p-values, leading to excessive rejection. To address this, we divided each of our samples into 100 subsamples of size 1000, computing a p-value for each. A good fit is then shown by a roughly uniform distribution of p-values for the subsamples, and an empirical cdf that matches that of a uniform distribution $U(0, 1)$. Figure 5 shows these p-value distributions and empirical cdf's for the samples from `SimPhy` and `Mesquite`, formally confirming the conclusions already described. Similar results are presented for pairs of taxa other than A, B, and for other species trees, in the Supplementary materials.

Topological features of samples were analyzed by tabulating counts of all rooted triple topologies across the sampled gene trees and then performing a chi-squared test, and by computing the MLE of the internal edge length on the species tree triple. Results are shown in Table 1. For the programs `Hybrid-Lambda`, `Phybase`, `SimPhy`, and `DendroPy` no p-values were extreme enough to suggest poor fit. However, the p-values for `Mesquite` strongly suggest poor model fit, with values extremely close to 0. Internal edge length estimates were also poorest for the `Mesquite` sample. Note that although `Hybrid-Lambda` had poor performance on our metric tests with units in generations, this procedure confirms it gave a good topological sample. This is consistent with the observation that a `Hybrid-Lambda` sample is accurate when species tree edge lengths are given in coalescent units. Indeed, assuming the `Hybrid-Lambda` algorithm is based on coalescent units, its errors may occur in conversion into numbers of generations.

5 Discussion

These results indicate that inadequate attention has been given previously to ensuring MSC simulators perform correctly. As novel methods are developed that can scale to ge-

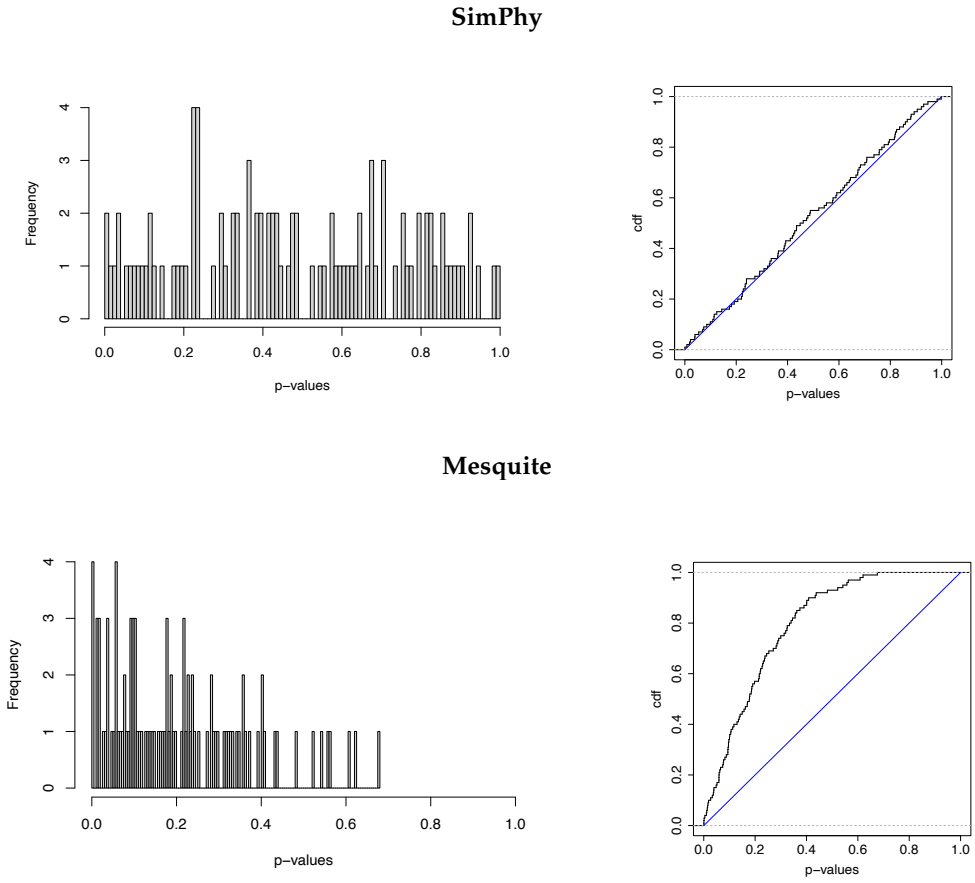

Fig. 5. Distributions and cdfs of p-values from Anderson-Darling tests comparing empirical and theoretical $d(A, B)$ densities for `SimPhy` (Top) and `Mesquite` (Bottom) on the species tree of Figure 3. The p-values were computed for 100 independent subsamples of size 1000 from the pairwise distance data underlying Figure 4. The pronounced divergence of the histogram from the expected density for `Mesquite` indicates poor fit. More rigorously, the deviation of Mesquite's empirical cdf from the blue cdf for a uniform distribution of p-values shows that Mesquite's sample is flawed.

nomic data or incorporate multilocus datasets of thousands of gene trees, it is imperative that such methodologies be tested on simulated datasets, both small and large, before practitioners can reliably trust analyses. The tests based on metric and topological summary statistics implemented in the R package `MSCsimtester` are a practical tool to uncover errors in simulators, and in user input to simulators. We recommend these tests be routinely used by developers of such simulators and, until software has been fully vetted, by anyone performing multispecies coalescent simulations.

Moreover, it is important that those currently conducting simulation studies, or interpreting the results of previous work (e.g., `Mesquite` simulations appear in [19], [20]), are aware of the problems these tests illuminate. The current version of `Hybrid-Lambda` is reliable only when metric species trees (and by extension networks) are specified in coalescent units, such as was done in [21], but caution is still warranted. A preliminary version of this paper showed that at least one earlier version of `Mesquite` produced erroneous samples, and we cannot verify that any early version performed correctly. New simulations in this work show problems remain in the current `Mesquite` version 3.7, and it should not be used for simulating under the MSC.

Note that our statistical methods consider distances for only one pair of taxa, or the rooted triple counts for only one choice of three taxa, at a time. When applied to many pairwise distances or rooted triple counts, even when conditioned on the species tree, these are not typically independent tests when applied to the same gene tree simulation. Although it would be desirable to test the full joint distribution as a step toward stronger testing, devising methods to do so is difficult. We believe the approach given here is adequate for revealing most simulator errors.

Finally, as the inference of species networks arising from hybridization or horizontal gene transfer receives more attention, simulators of the network multispecies coalescent (NMSC) model, such as `Hybrid-Lambda`, are likely to become more available. Enhancements of MSC simulators on a tree (e.g., permitting population sizes to vary with time on branch lengths) are also desirable. Developing testing methods for such software, that can be applied for general species networks and trees, will present new challenges. Understanding the distributions of summary statistics in such a setting is a considerably more complex task than that considered here.

6 Supplementary Material

Results of additional simulations are available in the supplementary material.

TABLE 1
Rooted triple topology counts from 100,000 gene trees sampled from the MSC on the species tree of Figure 3, and the p-values from the chi-squared tests. In the last column, the internal edge length of the species tree rooted triple in coalescent units with its Maximum Likelihood estimate from the simulated datasets. Small p-values and poor internal edge length estimates indicate poor model fit. Table entries are rounded.

	$((A,B),C)$	$((A,C),B)$	$((B,C),A)$	p-value	Int. edge
Expected	59564	20217	20217	-	0.500
Mesquite	56166	17424	26410	0	0.419
Hybrid-Lambda	59703	20227	20069	0.492	0.503
Phybase	59649	20122	20229	0.749	0.502
SimPhy	59800	20074	20126	0.306	0.506
DendroPy	59496	20049	20455	0.118	0.498

	$((A,B),D)$	$((A,D),B)$	$((B,D),A)$	p-value	Int. edge
Expected	71027	14487	14487	-	0.833
Mesquite	70229	14807	14964	1.26e-07	0.806
Hybrid-Lambda	71116	14462	14421	0.798	0.836
Phybase	71272	14335	14393	0.219	0.842
SimPhy	71061	14324	14615	0.225	0.835
DendroPy	71006	14584	14410	0.587	0.833

	$((A,C),D)$	$((A,D),C)$	$((C,D),A)$	p-value	Int. edge
Expected	52231	23884	23884	-	0.333
Mesquite	50256	24699.	25045	3.16e-35	0.293
Hybrid-Lambda	51997	24110	23892	0.203	0.328
Phybase	52367	23813	23820	0.691	0.336
SimPhy	52193	23832	23975.	0.784	0.333
DendroPy	52190	24028	23782	0.512	0.332

	$((B,C),D)$	$((B,D),C)$	$((C,D),B)$	p-value	Int. edge
Expected	52231	23884	23884	-	0.333
Mesquite	54905	22466	22629	4.57e-63	0.391
Hybrid-Lambda	51844	24131	24024	0.044	0.325
Phybase	52336	23826	23838	0.801	0.336
SimPhy	52116	24035	23849	0.534	0.331
DendroPy	52383	23921	23696	0.371	0.337

7 Acknowledgements

This work was supported, in part, by the National Institutes of Health [R01 GM117590], awarded under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological Mathematical Sciences, an NIGMS Institutional Development Award (IDeA) [2P20GM103395], and a National Science Foundation award [DMS 2051760]. H.B. was also partially supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation grant 735923LPI (DOI: https://doi.org/10.46714/735923LPI) awarded to Andrew J. Roger and Edward Susko.

References

[1] J. Degnan, "Modeling hybridization under the network multispecies coalescent," *Syst. Biol.*, vol. 67, no. 5, pp. 786–799, 2018. [Online]. Available: http://dx.doi.org/10.1093/sysbio/syy040

[2] L. Liu, "BEST: Bayesian estimation of species trees under the coalescent model." *Bioinformatics*, vol. 24, no. 21, pp. 2542–3, 2008.

[3] J. Heled and A. Drummond, "Bayesian inference of species trees from multilocus data," *Mol. Biol. and Evol.*, vol. 27, no. 3, pp. 570–580, 2010.

[4] J. Chifman and L. Kubatko, "Quartet inference from SNP data under the coalescent," *Bioinformatics*, vol. 30, no. 23, pp. 3317–3324, 2014.

[5] P. Vachaspati and T. Warnow, "ASTRID: Accurate species trees from internode distances," *BMC Genomics*, vol. 16, no. Suppl 10, p. S3, 2015.

[6] C. Zhang, M. Rabiee, E. Sayyari, and S. Mirarab, "ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees," *BMC Bioinformatics*, vol. 19, no. Suppl 6, p. 153, 2018.

[7] J. Rhodes, "Topological metrizations of trees, and new quartet methods of tree inference," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. to appear, 2019.

[8] D. Mallo, L. De Oliveira Martins, and D. Posada, "SimPhy: Phylogenomic simulation of gene, locus, and species trees," *Syst. Biol.*, vol. 65, no. 2, pp. 334–344, 2016. [Online]. Available: http://dx.doi.org/10.1093/sysbio/syv082

[9] E. Allman, C. Long, and J. Rhodes, "Species tree inference from genomic sequences using the log-det distance," *SIAM J. Appl. Algebra Geometry*, vol. 3, no. 1, pp. 1–30, 2019.

[10] P. Pamilo and M. Nei, "Relationships between gene trees and species trees." *Mol. Biol. and Evol.*, vol. 5, pp. 568–583, 1988.

[11] L. Liu and L. Yu, "Phybase: An R package for species tree analysis," *Bioinformatics*, vol. 26, no. 7, pp. 962–963, 2010. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btq062

[12] S. Zhu, J. Degnan, S. Goldstien, and B. Eldon, "Hybrid-Lambda: Simulation of multiple merger and Kingman gene genealogies in species networks and species trees," *BMC Bioinformatics*, vol. 16, no. 1, p. 292, Sep 2015. [Online]. Available: https://doi.org/10.1186/s12859-015-0721-y

[13] W. P. Maddison and D. Maddison, "Mesquite: A modular system for evolutionary analysis," 2021. [Online]. Available: http://www.mesquiteproject.org/

[14] J. Sukumaran and M. T. Holder, "DendroPy: A python library for phylogenetic computing," *Bioinformatics*, vol. 26, pp. 1569–1571, 2010.

[15] T. Anderson and D. Darling, "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes," *Ann. Math. Statist.*, vol. 23, no. 2, pp. 193–212, 1952.

[16] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
[17] R. Hudson, "Generating samples under a Wright-Fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, pp. 337–338, 2002.
[18] C. Than, D. Ruths, and L. Nakhleh, "PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary histories," *BMC Bioinformatics*, vol. 9, p. 322, 2008.
[19] W. P. Maddison and L. L. Knowles, "Inferring Phylogeny Despite Incomplete Lineage Sorting," *Systematic Biology*, vol. 55, no. 1, pp. 21–30, 02 2006. [Online]. Available: https://doi.org/10.1080/10635150500354928
[20] R. Yoshida, L. Zhang, and X. Zhang, "Tropical principal component analysis and its application to phylogenetics," *Bulletin of Mathematical Biology*, vol. 81, no. 2, pp. 568–597, 2019. [Online]. Available: https://doi.org/10.1007/s11538-018-0493-4
[21] A. G. Pereira and C. G. Schrago, "Incomplete lineage sorting impacts the inference of macroevolutionary regimes from molecular phylogenies when concatenation is employed: An analysis based on cetacea," *Ecology and evolution*, vol. 8, no. 14, pp. 6965–6971, 06 2018.