Molecular Phylogenetics and Evolution 57 (2010) 888–898

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Rapid and accurate species tree estimation for phylogeographic investigations using replicated subsampling

Sarah Hird [a], Laura Kubatko [b], Bryan Carstens [a,*]

[a] Department of Biological Sciences, 202 Life Sciences Building, Louisiana State University, Baton Rouge, LA 70803, USA
[b] Departments of Statistics and Evolution, Ecology, and Organismal Biology, Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 1 June 2010
Revised 5 August 2010
Accepted 10 August 2010
Available online 19 August 2010

Keywords:
Subsampling
Coalescent
Species tree estimation
Phylogeography

ABSTRACT

We describe a method for estimating species trees that relies on replicated subsampling of large data matrices. One application of this method is phylogeographic research, which has long depended on large datasets that sample intensively from the geographic range of the focal species; these datasets allow systematicists to identify cryptic diversity and understand how contemporary and historical landscape forces influence genetic diversity. However, analyzing any large dataset can be computationally difficult, particularly when newly developed methods for species tree estimation are used. Here we explore the use of replicated subsampling, a potential solution to the problem posed by large datasets, with both a simulation study and an empirical analysis. In the simulations, we sample different numbers of alleles and loci, estimate species trees using STEM, and compare the estimated to the actual species tree. Our results indicate that subsampling three alleles per species for eight loci nearly always results in an accurate species tree topology, even in cases where the species tree was characterized by extremely rapid divergence. Even more modest subsampling effort, for example one allele per species and two loci, was more likely than not (>50%) to identify the correct species tree topology, indicating that in nearly all cases, computing the majority-rule consensus tree from replicated subsampling provides a good estimate of topology. These results were supported by estimating the correct species tree topology and reasonable branch lengths for an empirical 10-locus great ape dataset.

1. Introduction

Phylogeographic studies require large sample sizes (Avise, 2000) and sequence data from hundreds (Wares and Cunningham, 2001; Zamudio and Savage, 2003) or even thousands (Bernatchez, 2001) of individuals have frequently been collected. This level of sampling is necessary to meet the aims of phylogeographic research, including the identification of population genetic structure and the estimation of the relationships among these populations. Initially, phylogeographic investigations were conducted using single-locus sequence data; in this case, an estimate of the gene tree was used as the primary tool for inferring both population structure and the relationships among populations. For a variety of reasons (Brumfield et al., 2003), phylogeographic investigations have adopted multilocus datasets, and these investigations have continued to rely on large sample sizes: datasets of 6–10 genes for 50–150 individuals are common (Dolman and Moritz, 2006; Geraldes et al., 2008), and some investigations sample hundreds of individuals (Garrick et al., 2008; Peters et al., 2008) or an appreciably greater number of loci (Lee and Edwards, 2008; Moeller and Tiffin, 2008). However, these large multilocus datasets can complicate analyses, particularly in the case of recently developed approaches for species tree estimation. While this work is discussed in the context of phylogeography largely due to the research interests of some of the authors, it should be noted that the results described below should be broadly applicable to phylogenetic research at a variety of levels of divergence.

Phylogeographic research has benefited from the development of phylogenetic methods that directly estimate the species phylogeny, either from previously estimated gene trees (Maddison and Knowles, 2006; Carstens and Knowles, 2007; Kubatko et al., 2009) or concurrently with estimation of gene trees (Edwards et al., 2007; Liu and Pearl, 2007; Ané et al., 2007; Heled and Drummond, 2010). While investigations that utilize these approaches have been smaller in terms of sample size than many single-locus phylogeographic investigations, they are still on the order of 6–10 genes for 5–10 individuals per lineage (Knowles and Carstens, 2007; Belfiore et al., 2008; Carling and Brumfield, 2008). With the advent of next-generation sequencing, future phylogeographic datasets will ideally include hundreds of loci and thousands of individuals.

Several characteristics of phylogeographic data are problematic for species tree estimation. First, since phylogenetic methods for

* Corresponding author. Fax: +1 225 578 2597.
E-mail address: carstens@lsu.edu (B. Carstens).

1055-7903/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2010.08.006

estimating the species tree require data from a single non-recombining locus, the resulting data often are limited in the number of variable sites that they contain. For example, investigations cited above utilize data that are slightly over 1 kb in length that contain an average of 40 variable sites. Since the accuracy of gene tree estimates is correlated to the number of variable sites (Hillis, 1995), these relatively low levels of sequence variation may be problematic in that they often result in gene trees that have relatively low levels of nodal support, particularly for nodes near the tips of the tree. Some approaches to species tree estimation address this shortcoming by simultaneously assessing the uncertainty in the gene tree and species tree posterior distributions using MCMC (e.g., Bayesian Estimation of Species Trees (BEST) (Liu and Pearl, 2007)). Second, it can be difficult to collect complete data matrices for multilocus phylogeographic investigations, and these missing data can complicate the estimation of species phylogeny. Third, the computational load imposed by such large datasets can be significant – in cases where many individuals are used, calculating a rigorous gene tree can take greater than a month, in some cases much longer.

Here we consider one method for estimating species phylogenies given some of the difficulties inherent to phylogeographic data. We explore the use of replicated subsampling; our approach involves the sampling of a small number of alleles from a data set that contains some large number of alleles (where an allele is a sampled variant at a given locus and a locus is a set of non-recombining contiguous base pairs, as are typically sampled in phylogenetic studies), the estimation of gene trees from these subsamples, and the subsequent estimation of the species trees using STEM (Kubatko et al., 2009). STEM computes the maximum-likelihood species phylogeny given an input consisting of a set of gene trees. Several characteristics of the coalescent process suggest that replicated subsampling is an approach worth exploring. For instance, samples collected by phylogeographic researchers consist of n sampled individuals, containing from 1 to many alleles at a given genetic locus, all related via descent from some common ancestral allele. At a given locus, the time until any two alleles share a common ancestor is a function of the effective population size, and overall the process is exponentially distributed such that most of the coalescent events occur in the recent past, but the waiting time on the last and deepest few coalescent events can be long. When this process is generalized as a gene tree, the tip branches are likely to be short and the deeper internodes long (Nordborg, 2000). Mutations provide an incomplete record of the coalescent process; for empirical systems, mutations are more likely to occur when the waiting time between coalescent events is long than when it is short (e.g., on the longer branches of the gene trees). Replicated subsampling will draw alleles that provide reasonably good estimates of these longer branches of the genealogies, even at the expense of information pertaining to the recent coalescent events. Since the deeper coalescent events within a population are expected to provide information regarding the relationships among populations (Hudson, 1991), a sample of the alleles by a given phylogeographic investigation should contain valuable information about the species tree.

It is unclear as to the degree of subsampling that will be required to produce good estimates of the longest branches of the gene trees, and thus of the species tree. Theory predicts that the probability of sampling the deepest coalescent event within a population is approximated by $n - 1/n + 1$, where n is the number of alleles (Saunders et al., 1984); this suggests a fairly large number of alleles would be required to sample the deepest nodes of a gene tree. However, this theory was developed for sampling within a single population and thus absent a phylogenetic framework; we intend to elucidate how sampling alleles from related lineages will influence the information provided by the samples.

2. Methods

2.1. Genealogical data simulation

Our approach proceeds by randomly subsampling n alleles per species from a large dataset for between two and thirty loci. It then estimates one genealogy per locus, and uses these genealogies to estimate the species phylogeny with STEM. We conduct this procedure 100 times and construct a consensus tree in order to evaluate the accuracy of the species tree estimates. Computationally, it is less difficult to estimate 100 small (e.g., five taxa) gene trees than a single large (e.g., 500 taxa) gene tree due to the reduction in gene tree space. Our first subsampling simulation (referred to henceforth as the "Genealogical Data – 30 Loci" simulation) has six distinct steps (Fig. 1):

Simulation Steps

1. Simulate a 5-taxa species tree topology under a Yule (pure birth) process using Mesquite (Maddison and Maddison, 2004), total tree depth 5N generations (where N is the effective population size). A pure birth process will tend to produce species trees with shorter internodes than a birth–death process; the use of the pure birth process here produces trees that should be more challenging to estimate.
2. Simulate gene genealogies onto the species tree using the program *ms* (Hudson, 2002). For each species tree, we simulated a 30-locus dataset with 100 alleles per species – thus, these genealogies contain 500 tips using a per locus $\theta = 4N_e\mu$ of 10.0.

Analysis Steps

3. Randomly prune (i.e., remove) alleles from each genealogy until only one allele per species remains: the 500-tip genealogy is reduced to a 5-tip genealogy for each of the 30 loci. This was accomplished using a Perl script and the program PAUP* (Swofford, 2002).
4. Use the 30 pruned genealogies as input dataset for STEM. The output (the maximum-likelihood species tree and its likelihood) are saved.
5. Remove two loci and estimate the species phylogeny using STEM. Repeat until only two loci are left.
6. Return to step (3) 99 times.
7. Begin the Analysis Steps again; prune each locus to three alleles per species intraspecific sampling effort.
8. Begin the Analysis Steps again; prune each locus to five alleles per species intraspecific sampling effort.
9. Begin the Analysis Steps again; prune each locus to 10 alleles per species intraspecific sampling effort.

Hereafter, a "treatment" refers to a given intraspecific subsampling effort and number of loci (e.g., three alleles per species and two loci is one treatment). We repeated the entire method for 40 independent species tree topologies. The first 20 topologies were simulated at a total tree depth of 5N generations (where N is the effective sample size) and the second 20 topologies were simulated at a total tree depth of 10N generations. Both of these tree depths represent extremely rapid diversification, with an average waiting time for cladogenesis between 1N and 2N (for 5N trees) and between 2N and 4N (for 10N trees) and produce data with significant levels of polymorphism shared among populations.

We evaluated the performance of STEM in two ways. First, within each treatment, we computed the symmetric difference distance (SDD; Robinson and Foulds, 1981) between each STEM replicate and the actual species tree topology. SDD is a pairwise comparison and returns the number of nodes that are present in one tree and not the other (a value of 0 corresponds to identical

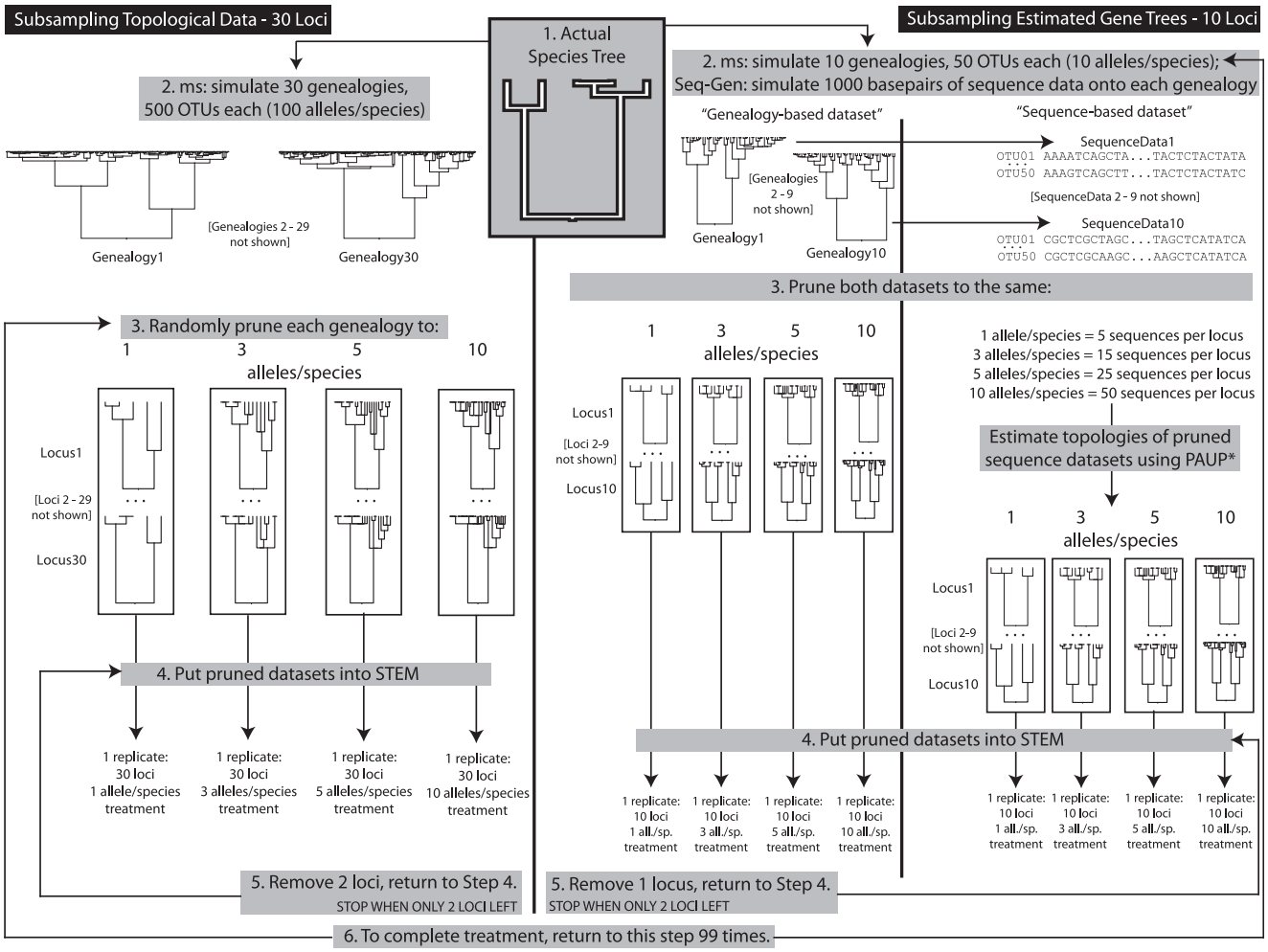

Fig. 1. Schematic of simulations conducted on each of the 40 species trees. Subsampling Topological Data – 30 loci on left and subsampling estimated gene trees – 10 loci on right. The right side is further subdivided into the paired "Genealogy-based dataset" and "Sequence-based dataset". Gray boxes and numbers correspond to steps outlined in the methods text. Thick vertical bars separate the datasets.

topologies). The percentage of replicates that had an SDD of 0 was used as a metric for accuracy of topology estimates. Second, we computed Kuhner–Felsenstein (KF; Kuhner and Felsenstein, 1994) distances between each replicate and the actual species tree. This is a metric for similarity in branch lengths and topology; larger scores correspond to less accurate species tree estimates. We computed the KF distance between each replicate and the actual species tree using the TreeDist package in Phylip (Felsenstein, 2005) then calculated the mean and variance for each treatment. We scaled all KF distances by the total depth of the species tree (e.g., either 5N or 10N) so that we could compare performance across tree depths. Finally, a consensus tree was found for each treatment using SumTrees (Sukumaran, 2008); we considered the consensus tree to be correct if the SDD between the consensus tree and the actual species tree was 0.

2.2. Estimated gene trees simulation

The above analyses were conducted using the actual simulated genealogies and thus represent the optimal scenario. Empirical datasets are limited in their ability to estimate the genealogy by the mutational process, as such STEM (in addition to other approaches that estimate the species tree from gene trees) should be less accurate when estimated gene trees are used as input. To assess the effect of using estimated genealogies, we simulated sequence data onto the genealogies, which allowed us to control for the stochastic selection of randomly choosing alleles by systematically removing the same alleles from all datasets. We used the following approach to assess the accuracy of replicated subsampling when estimated gene trees are used as input (henceforth referred to as the "Estimated Gene Trees – 10 Loci" simulation):

Simulation Steps

1. Use the same 40 species tree topologies generated by Mesquite for the Genealogical Data – 30 Loci simulation (five taxa, both 5N and 10N tree depth).
2. Simulate genealogies for 10 loci onto a species tree. Each genealogy has 10 alleles per species; this is the genealogy-based dataset. Simulate 1000 basepairs of sequence data onto each of the 10 genealogies in the genealogy-based dataset. Each individual sequence corresponds to one allele on the genealogy. This was done with the program Seq-Gen (Rambaut and Grassly, 1997); we used an HKY model of sequence evolution, a transition–transversion ratio of 0.5, equal nucleotide probabilities and scaled branches by 0.003 (in order to end up with 30–80 variable sites per dataset). We refer to these data as the sequence-based dataset.

Analysis Steps

1. Prune each locus to the one allele per species. Estimate genealogies from the pruned sequence datasets. In PAUP*, we conducted a likelihood search using SPR branch swapping, and saved trees with the molecular clock enforced and midpoint rooting.
2. Use the 10 pruned genealogies as an input dataset for STEM. The output (the maximum-likelihood species tree and its likelihood) are saved.
3. Remove one locus and analyze with STEM again. Repeat until only two loci are left.
4. Begin the Analysis Steps again; prune each locus to three alleles per species intraspecific sampling effort. Conduct steps (4) and (5).
5. Begin the Analysis Steps again; prune each locus to five alleles per species intraspecific sampling effort. Conduct steps (4) and (5).
6. Begin the Analysis Steps again; do not prune the loci. Conduct steps 4 and 5.
7. Return to step (1) 99 times.

We repeated the entire method for all 40 species tree topologies. Performance was assessed with the topological metric (SDD), the scaled branch length metric (KF/N distances) and consensus tree methods discussed above. Portions of this research were conducted with high performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu).

2.3. Empirical dataset

We used 10 non-coding loci (Fischer et al., 2006) for six hominid species: gorillas (*Gorilla gorilla* [$n = 18$]), humans (*Homo sapiens* [$n = 2$]), bonobos (*Pan paniscus* [$n = 9$]), chimpanzees (*Pan troglodytes* [$n = 30$], Sumatran orangutans (*Pongo abelii* [$n = 6$]) and Borneo orangutans (*Pongo pygmaeus* [$n = 10$]). For two of the loci, there were only 16 gorilla samples and for a third locus, there were only nine Bornean orangutans. We subsampled three alleles per species for all 10 loci and performed 100 replicates. The replicates were then used as input for STEM ($\theta = 0.00211$, averaged across taxa from Fischer et al., 2006). We then calculated a 50% majority-rule consensus tree.

3. Results

Simulated species trees had a wide range of divergence dates for the speciation events, and their shortest internodes ranged from 0.024N to 2.88N. These levels of divergence are expected to produce a large number of shared alleles across lineages, and a qualitative examination of the simulated data (not shown) supported this expectation.

3.1. Subsampling Genealogical Data – 30 Loci

As intraspecific sampling increased, estimates of the species tree improved, both with regard to topology and branch lengths (Fig. 2). For example, with one sampled allele per species and two loci (total tree depth 5N), 53.2% of the replicates had the correct topology

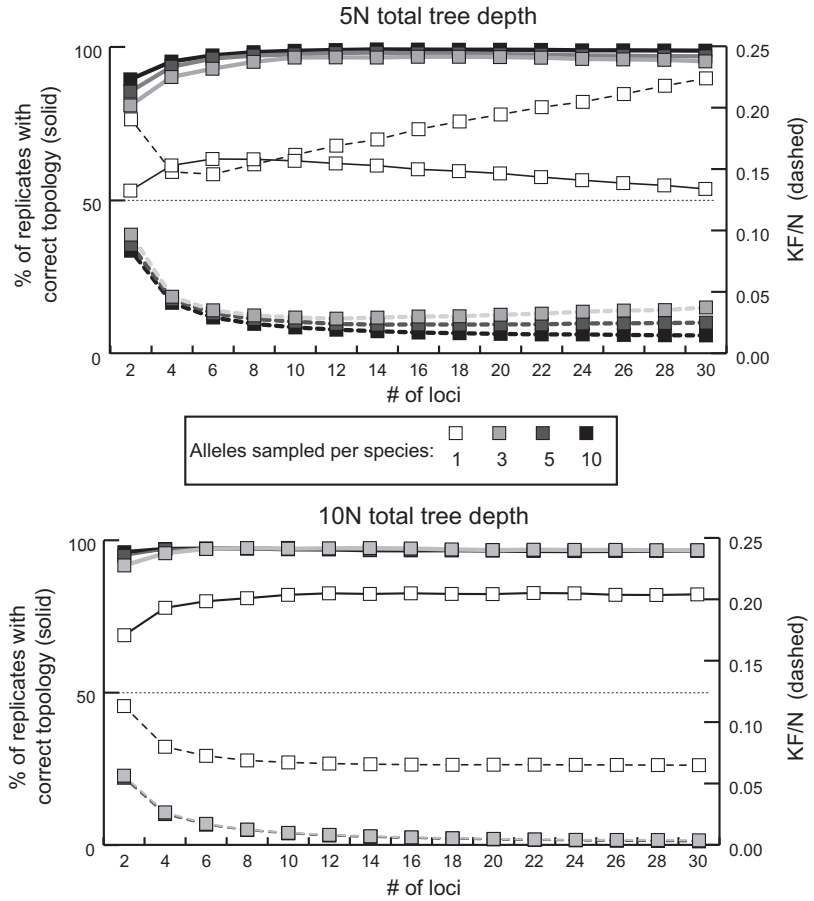

Fig. 2. Results from the Genealogical Data – 30 Loci simulation; the percentage of replicates with the correct topology (solid lines) and KF/N distances (dashed lines).

(Table 1), and as alleles were added this percentage improved substantially (three alleles = 80.95%, five alleles = 85.45%, 10 alleles = 89.45%). However, as loci are added, the improvement that results from adding alleles diminishes such that there is little beyond three alleles when more than six loci are used (Fig. 2). Thus, these results suggest that replicated subsampling of three alleles and six loci will almost always accurately estimate the topology of the species tree. Additionally, as the species tree depth increases from 5N to 10N, these results generally improve (Fig. 2). Similar to the estimates of topology, the branch length estimates usually improve as data are added, and KF/N distances consistently approach zero as both loci and intraspecific sampling increase (Table 1, Fig. 2, Supplementary Table S2). The exceptions are the 5N, one allele treatments, where accuracy never exceeds 65% and branch length estimates worsen as data are added.

The accuracy of species tree estimates also depends on the shape of the species tree; some topologies resulted in genealogies that produced poor estimates of the species phylogeny, while others did not (Fig. 3). Performance is related to the length of the internodes of the species tree: the treatments with lowest percentage of correctly estimated topologies tended to be on species trees with shortest internodes ($R^2 = 0.45125$, Fig. 4). These results are consistent with another recent exploration of STEM's performance (McCormack et al., 2009).

3.2. Subsampling estimated gene trees – 10 loci

The results using the estimated gene trees (sequence-based dataset) mirrored those from the actual genealogies (genealogy-based dataset), although overall topological accuracy and goodness of branch length estimates were worse (Fig. 5). For example, the percentage of replicates estimating the correct topology in the three sampled alleles per species and five loci treatment (total tree depth of 5N), decreases from 88.5% to 58.4% when estimated gene trees are used as input (Supplementary Table S3). However, in almost all cases, the majority of the replicates correctly estimated the topology even in cases where performance was appreciably lower. This suggests that the use of a majority-rule consensus tree in conjunction with replicated subsampling may provide a good estimate of the species tree topology, even when species divergence is extremely recent. The one species tree from our simulations where this is not true (tree35; Fig. 3) is an example of the anomalous gene tree zone (Degnan and Rosenberg, 2006), where the most commonly sampled genealogy does not reflect the branching order of the species tree.

When estimated gene trees were used as input, the most noticeable decrease in performance involved branch lengths, as measured by the KF/N distances (Fig. 5, Supplementary Table S4). KF/N distances are appreciably higher when estimated gene trees are used. In this case, it is clear that users of STEM will benefit by gathering data from additional loci: going from two loci to six loci decreased KF/N by approximately 0.15. However, the differences in branch length may not be dramatic for species trees of certain shapes – in some cases the increase from estimated to actual was <0.001 KF/N units (Supplementary Table S4), effectively the same tree (Fig. 6).

It requires an average of 2.15 s for estimating 100 replicates of the one allele per species gene trees, 363.65 s (6.06 min) for three alleles per species, 4274.2 s (71.2 min) for five alleles per species and 87,178 s (24.21 h) for 10 alleles per species. STEM computes the analytical solution to all these datasets in <1 s.

3.3. Empirical dataset

After subsampling three alleles per species for all 10 loci of the great ape dataset, we calculated a 50% majority-rule consensus tree of the 100 replicates (Fig. 7). The consensus tree fully recovered the accepted great ape topology with 100% consensus among replicates. The total tree depth was 18.64N generations (Fig. 7).

4. Discussion

Phylogeographic investigations require methods that work well with large numbers of individuals, and that can handle large amounts of data in an efficient manner. Our subsampling method reduces the computational load of large datasets by repeatedly pruning it to a small number alleles and analyzing the smaller dataset. Since the estimation of gene trees is the most computationally intensive step in an analysis that uses STEM to estimate species phylogeny, replicated subsampling is likely to require less time and produce equally good results.

4.1. Subsampling Genealogical Data – 30 Loci

The results of our first simulation show that as data are added, results improve. In most cases, the topological correctness exceeds 95% and in all cases it exceeds 50%. Perhaps the most striking result is the poor performance of subsampling one allele per species, especially when compared to the sampling of three alleles per species. This may be explained by the fact that no intraspecific coalescent information exists when only one allele per species is sampled. The percentage of replicates with the correct topology plateaus before reaching 100% in both the 5N and 10N trees. It is worth noting, however, that all the 10N treatments reached 100% correctness to the exclusion of one species tree (tree35) with an extremely difficult topology that contains an internode of 0.024N generations. Finally, these results suggest that as an empirical guide, subsampling three alleles per species for six loci will result in the correct topology in >90% of replicates and result in KF/N values <0.05. In cases where six loci are not available, four loci may be adequate, as the four loci treatments performed almost as well as six loci.

4.2. Subsampling estimated gene trees – 10 loci

The difference between simulated gene trees and simulated sequence data is the error associated with the estimation of genealogies from sequence data. This error is attributable to the inherent variance in the mutational process, as well as some degree of phylogenetic estimation error. In combination, this error can be substantial so we analyzed its effect on the proposed subsampling method. This simulation paired the same alleles across the genealogies known without error and the sequence data simulated from the genealogies thus comparing the performance of the exact same alleles. In this way, decreases in performance were entirely attributable to estimation error. Unsurprisingly, the actual genealogies performed better than the estimated gene trees (by 8.35–39.9%). As either loci or alleles were added, the difference between the performances of the paired datasets grew. So although adding both loci and alleles produces an improvement in the accuracy of the species tree estimation, the actual genealogies had larger gains than the estimated gene trees (data not shown).

4.3. Application of method

Perhaps the most obvious application of this method involves subsampling a dataset many times and computing a consensus tree of the replicates. This application performed well in our simulations (Fig. 8). Across all replicates of the "Genealogical Data – 30 Loci" simulation, 90.83% of the treatments produced a consensus tree that matched the topology of the actual species tree. Failure

Table 1
Average percentage of replicates with correct topology (RF%) and KF/N distance (KF/N) and the associated variances (RFvar and KFvar, respectively) for the various treatments.

		Loci[b]														
Alleles[a]		2	4	6	8	10	12	14	16	18	20	22	24	26	28	30
A. 5N (total tree depth)																
1	RF%	53.20	61.40	63.50	63.40	62.90	62.05	61.35	60.15	59.55	58.80	57.60	56.60	55.65	54.90	53.75
	RFvar	1077.33	1065.52	1086.89	1156.99	1181.57	1249.63	1308.87	1368.24	1421.73	1487.43	1586.15	1647.31	1693.29	1771.25	1829.67
	KF/N	0.191	0.148	0.146	0.154	0.162	0.169	0.174	0.182	0.189	0.194	0.200	0.205	0.211	0.218	0.224
	KFvar	0.099	0.251	0.327	0.377	0.408	0.437	0.455	0.459	0.467	0.469	0.471	0.479	0.483	0.498	0.501
3	RF%	80.95	90.20	92.90	95.15	96.55	96.60	96.55	96.80	96.80	96.70	96.50	96.10	95.95	95.80	95.30
	RFvar	482.79	193.85	108.31	52.77	23.52	19.83	22.37	21.22	21.43	22.96	29.00	34.52	36.58	38.69	49.48
	KF/N	0.097	0.046	0.035	0.031	0.029	0.028	0.029	0.030	0.030	0.031	0.032	0.034	0.035	0.035	0.037
	KFvar	0.008	0.010	0.012	0.014	0.016	0.018	0.022	0.023	0.025	0.027	0.030	0.034	0.036	0.038	0.042
5	RF%	85.45	93.55	96.20	97.20	97.60	98.05	98.15	97.90	97.90	97.65	97.50	97.25	97.15	97.00	96.95
	RFvar	378.05	112.58	41.64	20.91	13.41	9.21	9.92	11.78	12.62	15.92	17.21	19.88	20.98	23.89	25.00
	KF/N	0.088	0.044	0.032	0.028	0.026	0.024	0.023	0.023	0.023	0.023	0.024	0.024	0.024	0.025	0.025
	KFvar	0.009	0.010	0.012	0.013	0.014	0.015	0.016	0.018	0.019	0.022	0.023	0.025	0.025	0.026	0.027
10	RF%	89.45	95.25	97.25	98.30	98.75	99.00	99.20	99.15	99.10	99.10	99.05	98.90	98.90	98.85	98.80
	RFvar	317.31	80.20	29.36	12.75	6.72	5.16	3.12	3.82	4.62	3.99	4.05	6.73	6.73	7.82	7.75
	KF/N	0.083	0.041	0.029	0.024	0.021	0.019	0.018	0.017	0.016	0.016	0.015	0.015	0.015	0.015	0.015
	KFvar	0.010	0.010	0.012	0.013	0.013	0.013	0.014	0.014	0.015	0.016	0.017	0.018	0.018	0.018	0.018
B. 10N (total tree depth)																
1	RF%	69.05	77.60	80.55	81.55	82.30	82.25	81.90	82.40	82.25	82.15	82.65	82.45	82.05	82.15	82.25
	RFvar	711.71	732.34	862.84	888.95	901.88	931.62	986.99	1049.83	1101.31	1137.71	1150.12	1162.04	1222.73	1256.58	1267.88
	KF/N	0.113	0.080	0.073	0.069	0.068	0.067	0.066	0.066	0.066	0.065	0.065	0.065	0.065	0.065	0.065
	KFvar	0.653	1.119	1.405	1.558	1.652	1.719	1.769	1.815	1.847	1.872	1.893	1.901	1.912	1.928	1.935
3	RF%	91.80	95.15	96.55	96.95	97.00	96.80	96.90	96.60	96.55	96.55	96.70	96.70	96.50	96.65	96.75
	RFvar	232.41	148.51	100.98	114.89	145.57	134.98	124.79	140.22	173.79	204.80	192.20	198.45	198.45	211.25	211.25
	KF/N	0.057	0.027	0.017	0.012	0.010	0.008	0.007	0.006	0.005	0.005	0.004	0.004	0.004	0.004	0.004
	KFvar	0.010	0.003	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002
5	RF%	94.75	96.40	97.15	97.35	97.10	97.00	97.15	97.00	96.80	96.50	96.50	96.65	96.60	96.50	96.50
	RFvar	196.20	124.84	140.22	151.25	125.00	156.80	204.80	198.45	217.80	217.80	217.80	231.20	245.00	245.00	245.00
	KF/N	0.054	0.025	0.017	0.012	0.010	0.008	0.007	0.006	0.005	0.005	0.004	0.004	0.004	0.003	0.003
	KFvar	0.009	0.003	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
10	RF%	95.90	96.85	97.65	98.00	97.75	97.40	97.40	97.20	97.20	97.10	97.10	96.80	96.90	96.60	96.65
	RFvar	151.78	129.71	145.57	140.45	174.05	180.00	211.25	217.80	211.25	224.45	245.00	252.05	245.00	238.05	224.45
	KF/N	0.054	0.026	0.017	0.012	0.009	0.008	0.007	0.006	0.005	0.005	0.004	0.004	0.003	0.003	0.003
	KFvar	0.007	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

[a] Level of intraspecific sampling, or number of alleles sampled per species.
[b] Number of loci used to calculate the species tree with STEM.

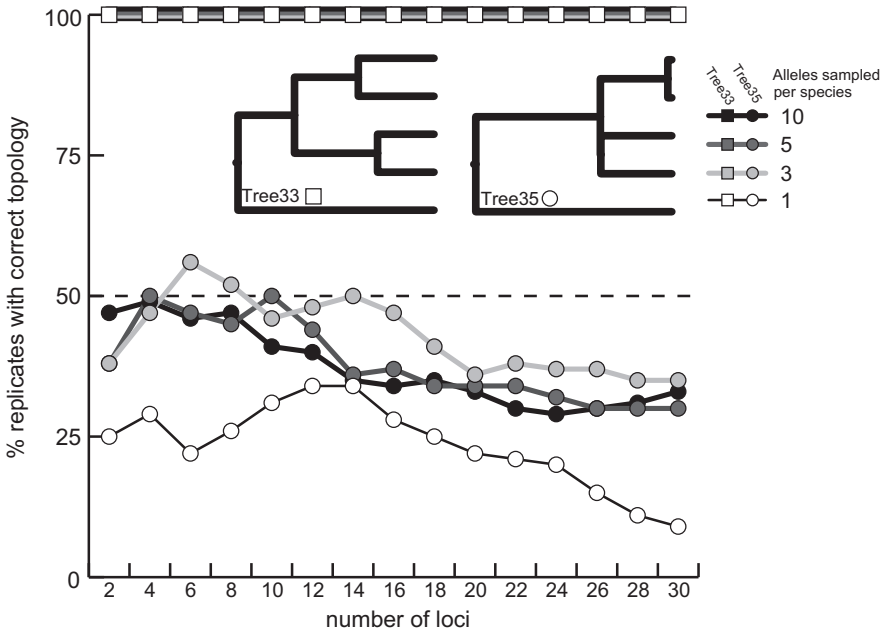

Fig. 3. Comparison of performance of STEM on a tree with relatively long (tree33) and short (tree35) internodes from Genealogical Data – 30 Loci simulation.

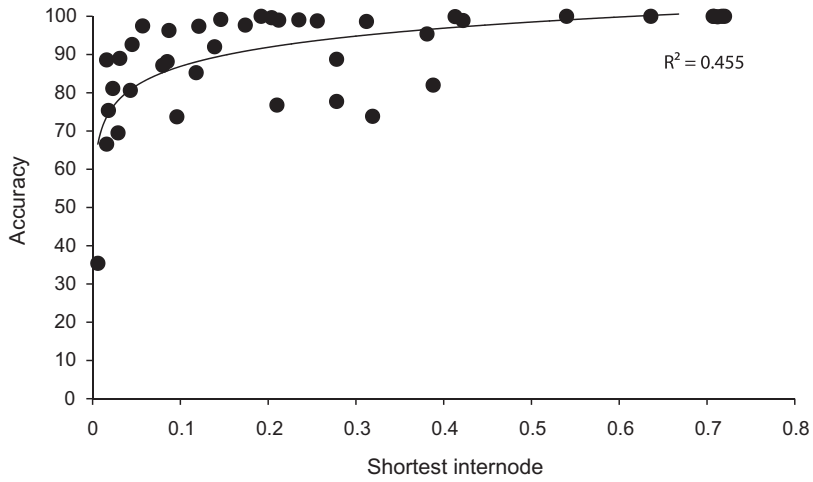

Fig. 4. Relationship between the shortest internode on the species tree and the percentage of replicates with the correct topology (R^2 of logarithmic trendline equals 0.4559).

of the consensus tree to recover the correct topology was correlated with intraspecific sampling. When the one allele/species treatments were excluded, the number of correct consensus trees rose to 97.6% and if the anomalous tree (tree35) was excluded, the number rises further to 99.9%.

We have shown our method works well for simulated data. In order to use this approach with confidence on species phylogenies without a known species tree, we analyzed an empirical dataset that has generally accepted phylogenetic relationships. The great ape dataset included 10 loci for six hominid taxa: gorillas, humans, bonobos, chimpanzees, Sumatran orangutans and Borneo orangutans (Fischer et al., 2006). This dataset was part of a larger dataset that was used to show that the genetic distances between subspecies of chimpanzees and bonobos were as much as the genetic variation within humans. There was no phylogenetic analysis done with these loci in Fischer et al. (2006), however the phylogenetic relationships between the great apes is generally accepted to be: ((Gorilla,(Human,(Chimp, Bonobo))),(Orangutan)). We subsampled three alleles per species for 10 loci and performed 100 replicates. The replicates were then used as input for STEM ($\theta = 0.00211$). We then calculated a 50% majority-rule consensus tree (Fig. 7). The consensus tree fully recovered the accepted great ape topology with 100% consensus among replicates. The estimates of branch lengths we similar to published estimates of divergence time between great ape taxa. The branch separating bonobos and chimpanzees on our STEM tree is 2.7N generations long. If we assume an average generation time of 15 years (Won and Hey, 2005), the branch becomes 41.52N years and if we assume an average effective population size of 15,200 (average of two chimpanzee subspecies Ne and the bonobo Ne estimates from Won and Hey, 2005), their divergence time is approximately 631,100 years ago. Won and Hey (2005) estimate this divergence to be 0.86–0.89 million years ago. Given our very broad assumptions, we believe this is a reasonable performance of our method in estimating branch length. Additionally, we reduced the dataset to both five loci and two loci, and the results were the same: recovery of the accepted

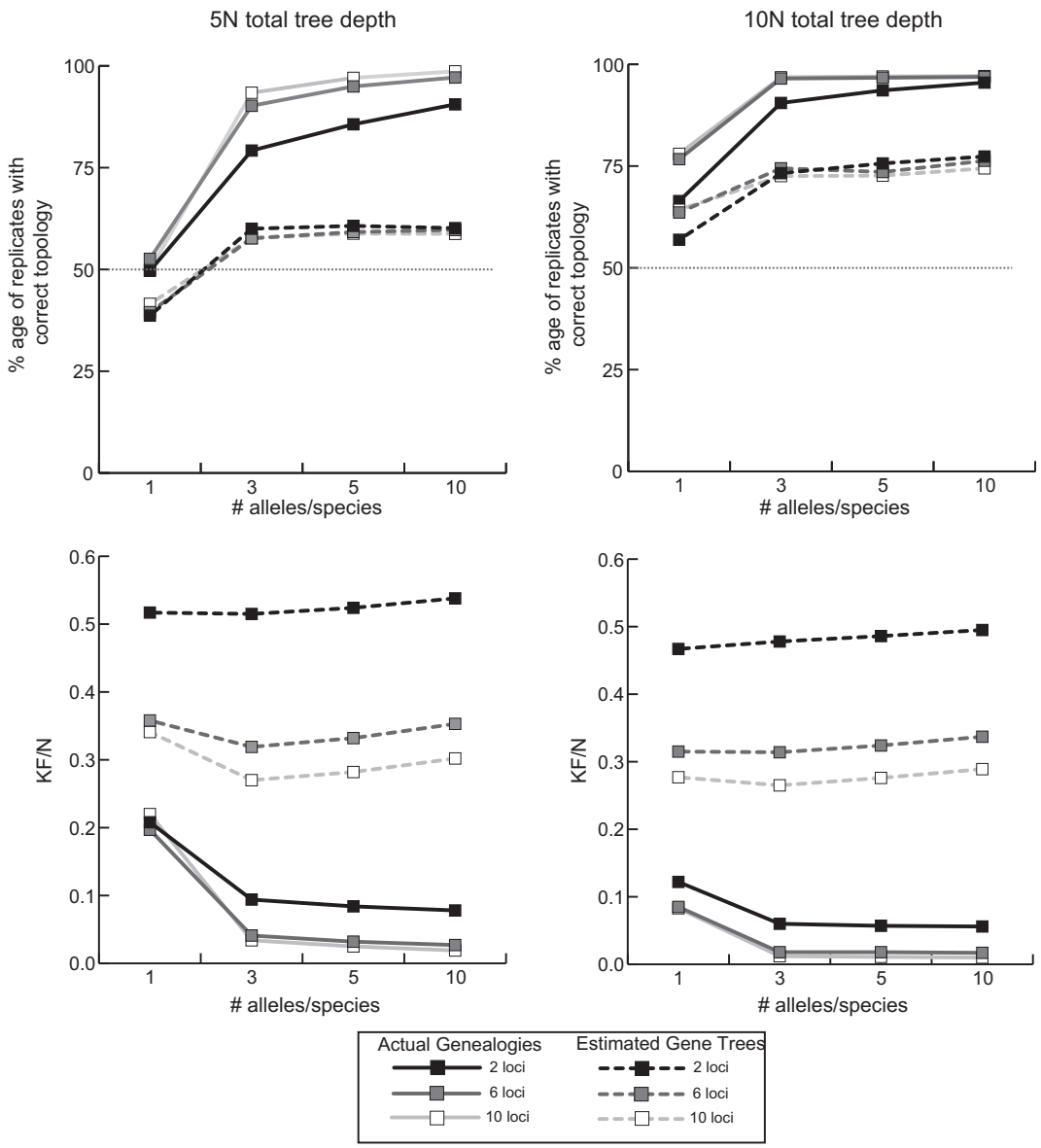

Fig. 5. Results of estimated gene trees – 10 loci simulation; solid lines indicate performance from actual genealogies known without error; dashed lines indicate performance of estimated gene trees.

topology with branch lengths similar to those shown in Fig. 7. An additional benefit of this method is a record of the proportion of times that a particular node occurs across replicates this proportion could be utilized as a measure of nodal support.

4.4. Performance of method

Overall, subsampling alleles provides a good estimate of the underlying species tree across a range of data. As more information is added, the percentage of replicates with the correct species topology increases and the probability that the consensus tree is correct increases. The branch lengths approach the correct values as more data are added as well. Consequently, there is a trade-off between the amount of data and the speed of the analysis. Given the profile of the accuracy curves for our simulations, we propose that a reasonable trade-off between correctness and computational speed occurs using eight loci and three alleles per species, which produced the smallest dataset that was >95% accurate at both 5N and 10N (Table 1). If eight loci are unattainable, six loci with five alleles per species or four loci with 10 alleles per species also have >95% accuracy at both 5N and 10N. For our simulation of gene trees estimated from sequence data, a decrease in accuracy was found, but 50% majority-rule consensus trees should provide a good estimate of the species tree, as all treatments had greater than 50% accuracy. Finally, our method provides computational efficiency. We used GARLI (Zwickl, 2006) to calculate a 500 taxa tree similar to the sequence datasets in our analysis, which it finished in 3 h. Our replicated subsampling method takes between 6 min (for three alleles per species) to 24 h (for 10 alleles per species), calculating gene trees with PAUP* under maximum-likelihood. This represents an increase in efficiency, as we not only have an estimate of the species tree, we also have support for the nodes of the trees as well. The support for the bipartitions are similar to the support assigned with a statistical jackknife, which also relies

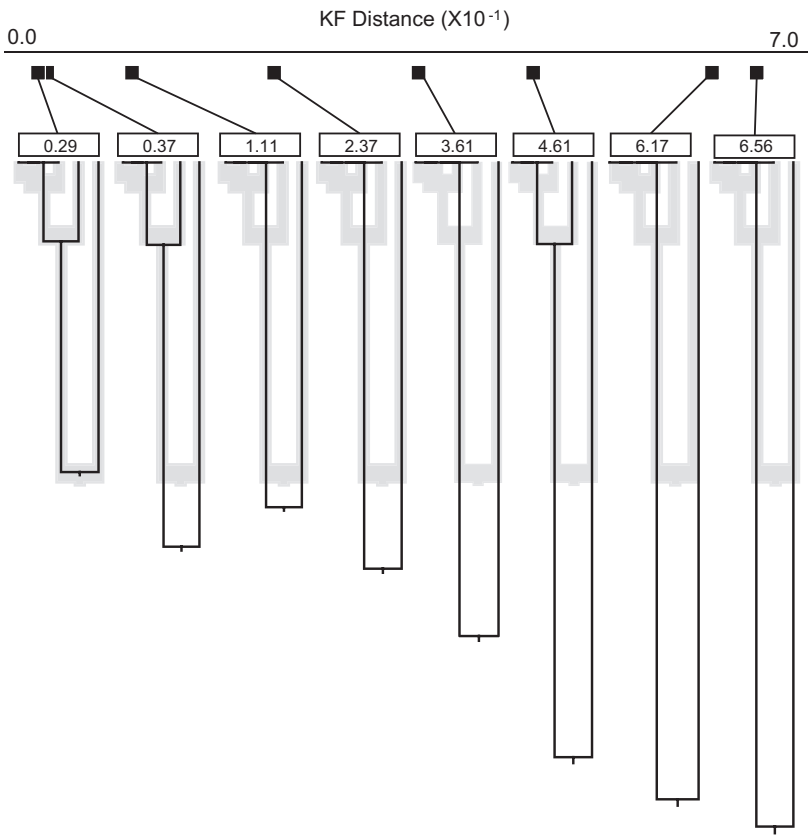

Fig. 6. Visualization of differences in KF distance. The gray phylogeny is the actual species tree. The black trees (drawn on the same scale) show how similar the topologies are for several KF distance values.

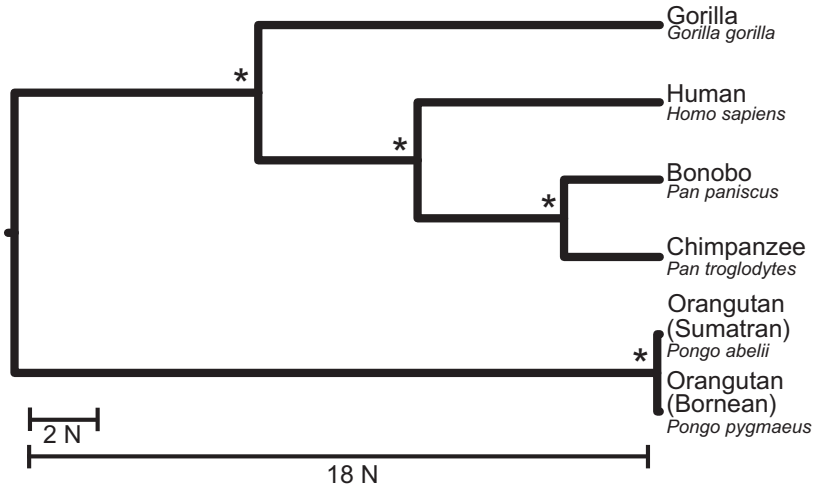

Fig. 7. Results from the empirical great ape dataset – consensus tree of 100 STEM replicates; *Indicates 100% consensus.

on subsampling taxa (as compared to a bootstrap, which subsamples sites, with replacement).

4.5. Application to empirical systems

The approach described above is likely to be broadly applicable to phylogenetic and phylogeographic investigations into empirical systems, but there are several important caveats. First, thorough sampling is critical; all analyses have been conducted by subsampling a small number of alleles at random from a larger set, and this scheme is not equivalent to sampling one or two individuals from a widespread species. Second, while STEM assumes that the OTUs represent independent evolutionary lineages, the actual species boundaries in empirical systems may be unknown, which could produce inaccurate estimates of species phylogeny. For systems such as these, we advocate a joint

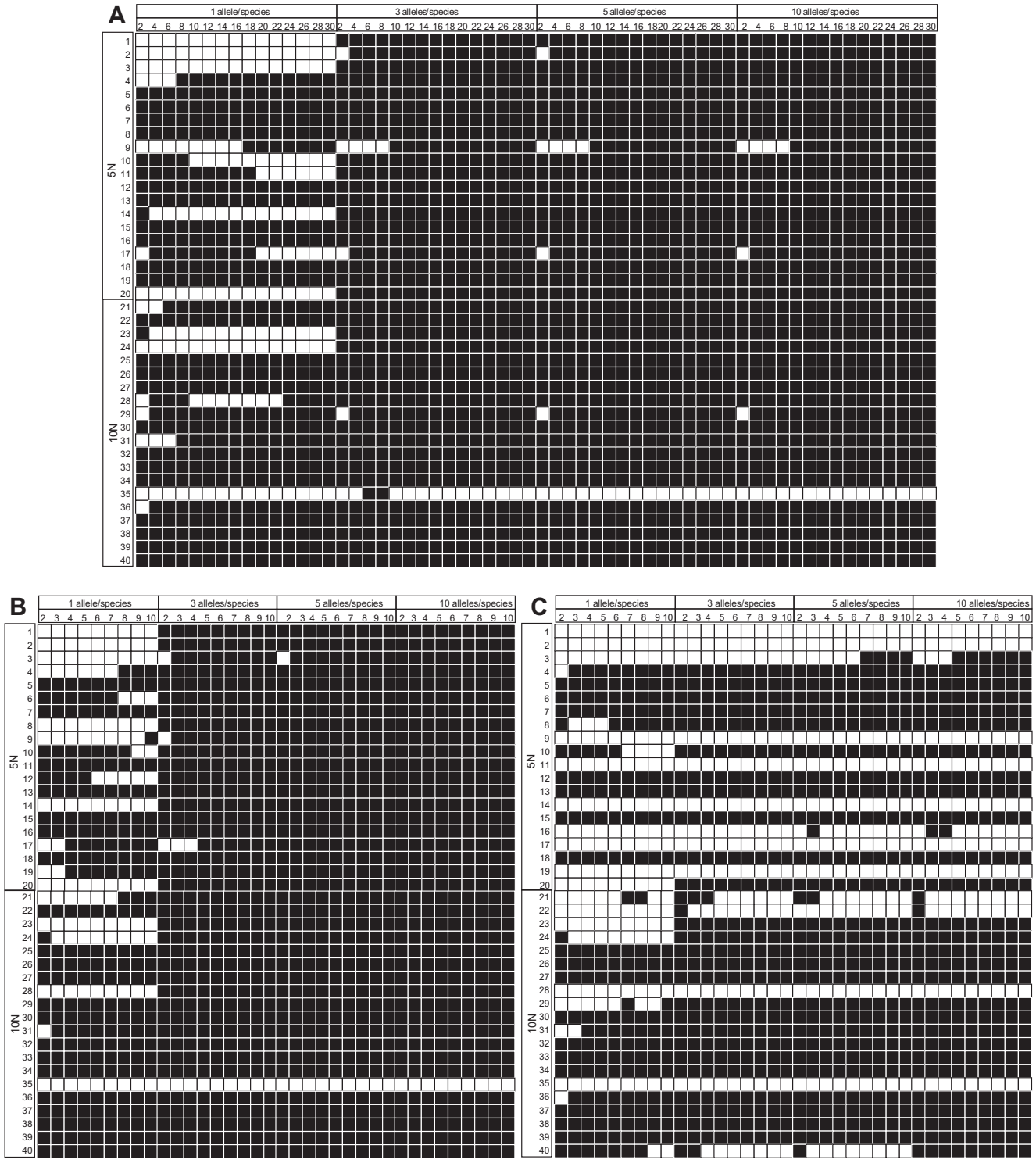

Fig. 8. Consensus tree accuracy for the Genealogical Data – 30 Loci (A) and the genealogy-based dataset (B) and sequence-based dataset (C) from the estimated gene trees – 10 loci simulations. Black boxes indicate which treatments produced a consensus tree from the 100 replicates that matched the actual species topology. White boxes indicate a consensus tree different from the actual species tree.

estimation of species trees and lineage membership (e.g., O'Meara, 2010; Carstens and Dewey, 2010). Finally, no simulation study can capture the complexities of a particular empirical system, and for this reason we advocate the use of power analyses that are tailored to the empirical system of interest. To facilitate these analyses, we have made all the scripts used in this research available at: http://www.lsu.edu/faculty/carstens/archives/HKC.2010.zip.

5. Conclusions

Our analyses suggest that replicated subsampling is a useful approach to species tree estimation. For large phylogeographic datasets, which are likely to consist of multilocus allelic data from hundreds of individuals, this approach will correctly estimate species topology even in difficult cases where lineage divergence is rapid and recent. In the vast majority of cases, it will also provide good estimates of branch lengths.

Acknowledgments

The authors wish to acknowledge that funding for this project was provided by the National Science Foundation Grant NSF (DEB-0956069). Portions of this research were conducted with high performance computational resources provided by Louisiana State University (http://www.hpc.lsu.edu). We thank N. Reid, T. Pelletier, J. McVay, A. Espindola, Y.E. Tsai, D. Ence. J. Charboneau and M. Koopman for thoughtful discussions and critique of the manuscript. We also thank A.E. Saitou and two anonymous reviews for comments and suggestions that improved this manuscript.

References

Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Belfiore, N.M., Liu, L., Moritz, C., 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). Syst. Biol. 57, 294–310.

Bernatchez, L., 2001. The evolutionary history of brown trout (*Salmo trutta* L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation. Evolution 55, 351–379.

Brumfield, R.T., Beerli, P., Nickerson, D.A., Edwards, S.V., 2003. The utility of single nucleotide polymorphism in inferences of population history. Trends Ecol. Evol. 18, 249–256.

Carling, M.D., Brumfield, R.T., 2008. Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in *Passerina* Buntings. Genetics 178, 363–377.

Carstens, B.C., Dewey, T.A., 2010. Species delimitation using a combined coalescent and information-theoretic approach: An example from North American Myotis bats. Syst. Biol. 59, 400–414.

Carstens, B.C., Knowles, L.L., 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst. Biol. 56, 400–411.

Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68.

Dolman, G., Moritz, C., 2006. A multilocus perspective on refugial isolation and divergence in rainforest skinks (*Carlia*). Evolution 60, 573–582.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104, 5936–5941.

Felsenstein, J., 2005. PHYLIP (Phylogenetic Inference Package) Version 3.06. Distributed by the author. Department of Genome Sciences, University of Washington.

Fischer, A., Pollack, J., Thalmann, O., Nickel, B., Pääbo, S., 2006. Demographic history and genetic differentiation in apes. Curr. Biol. 16, 1133–1138.

Garrick, R.G., Rowell, D.M., Simmons, C.S., Hillis, D.M., Sunnucks, P., 2008. Fine-scale phylogeographic congruence despite demographic incongruence in two low-mobility saproxylic springtails. Evolution 62, 1103–1118.

Geraldes, A., Basset, P., Gibson, B., Smith, K.L., Harr, B., Yu, H.-T., Bulatova, Y.Z., Nachman, N.W., 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol. Ecol. 17, 5349–5363.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Hillis, D.M., 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44, 3–16.

Hudson, R.R., 1991. Gene genealogies and the coalescent process. In: Futuyma, D., Antonovics, J. (Eds.), Oxford Surveys in Evolutionary Biology. Oxford University Press, New York, pp. 1–44.

Hudson, R.R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

Knowles, L.L., Carstens, B.C., 2007. Estimating a geographically explicit model of population divergence. Evolution 61, 477–493.

Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.

Lee, J.Y., Edwards, S.V., 2008. Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed Fairy Wren (*Malurus melanocephalus*). Evolution 62, 3117–3134.

Liu, L., Pearl, D.K., 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56, 504–514.

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.

Maddison, W.P., Maddison, D.R., 2004. Mesquite: A Modular System for Evolutionary Analysis. Version 1.01. Available at <http://mesquiteproject.org>.

McCormack, J.E., Huang, H., Knowles, L.L., 2009. Breaking them down and building them up: evaluating the accuracy of maximum-likelihood estimates of species trees. Syst. Biol. 58, 501–508.

Moeller, D.A., Tiffin, P., 2008. Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. Evolution 62, 3069–3081.

Nordborg, M., 2000. Coalescent theory. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics, third ed. John Wiley and Sons, Ltd., West Sussex, England, pp. 843–872.

O'Meara, B.C., 2010. New heuristic methods for joint species delimitation and species tree estimation. Syst. Biol. 59, 59–73.

Peters, J.L., Zhuravlev, Y.N., Fefelov, I., Humphries, E.M., Omland, K.E., 2008. Multilocus phylogeography of a Holarctic duck: colonization of North America from Eurasia by Gadwall (*Anas strepera*). Evolution 62, 1469–1483.

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Robinson, D.R., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Saunders, I.W., Tavare, S., Watterson, G.A., 1984. On the genealogy of nested sub-samples from a haploid population. Adv. Appl. Prob. 16, 471–491.

Sukumaran, J., 2008. SumTrees: Summarization of Split Support on Phylogenetic Trees. Part of the: DendroPy Phylogenetic Computation Library. Available at <http://sourceforge.net/projects/dendropy>.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (and other methods). Version 4. Sinnauer Associates, Sunderland, MA.

Wares, J.P., Cunningham, C.W., 2001. Phylogeography and historical ecology of the North Atlantic intertidal. Evolution 55, 2455–2469.

Won, Y., Hey, J., 2005. Divergence population genetics of chimpanzees. Mol. Biol. Evol. 22, 297–307.

Zamudio, K.R., Savage, W.K., 2003. Historical isolation, range expansion, and secondary contact of two highly divergent mitochondrial lineages in spotted salamanders (*Ambystoma maculatum*). Evolution 57, 1631–1652.

Zwickl, D.J., 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion. Ph.D. Dissertation. The University of Texas at Austin. Available at: <http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html>.