Bayesian Inference of Species Trees from Multilocus Data

Joseph Heled*[,1] and Alexei J. Drummond[1,2,3]

[1]Department of Computer Science, University of Auckland, New Zealand
[2]Bioinformatics Institute, University of Auckland, New Zealand
[3]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, New Zealand

***Corresponding author:** E-mail: jheled@gmail.com.

Associate editor: Dr. Jeffrey Throne

Abstract

Until recently, it has been common practice for a phylogenetic analysis to use a single gene sequence from a single individual organism as a proxy for an entire species. With technological advances, it is now becoming more common to collect data sets containing multiple gene loci and multiple individuals per species. These data sets often reveal the need to directly model intraspecies polymorphism and incomplete lineage sorting in phylogenetic estimation procedures.

For a single species, coalescent theory is widely used in contemporary population genetics to model intraspecific gene trees. Here, we present a Bayesian Markov chain Monte Carlo method for the multispecies coalescent. Our method coestimates multiple gene trees embedded in a shared species tree along with the effective population size of both extant and ancestral species. The inference is made possible by multilocus data from multiple individuals per species.

Using a multiindividual data set and a series of simulations of rapid species radiations, we demonstrate the efficacy of our new method. These simulations give some insight into the behavior of the method as a function of sampled individuals, sampled loci, and sequence length. Finally, we compare our new method to both an existing method (BEST 2.2) with similar goals and the supermatrix (concatenation) method. We demonstrate that both BEST and our method have much better estimation accuracy for species tree topology than concatenation, and our method outperforms BEST in divergence time and population size estimation.

Key words: multispecies coalescent, species trees, gene trees, molecular systematics, Bayesian inference, censored coalescent.

Research article

Introduction

Despite huge advances both in evolutionary theory and in sequencing technology, estimating the "Tree of Life" even for a small subset of species can be challenging. Better mathematical models and more data improve our ability to infer a single gene phylogeny, but a gene history may be different from the species phylogeny. The potential for a discrepancy between the gene tree and the species tree has been known for decades and is especially problematic for closely related species or species with large population sizes. Building a species tree requires combining information from multiple genes; all gene phylogenies need to be "embedded" inside the species history while not violating the species tree constraints: The time of a common ancestor of a gene cannot be more recent than the time of divergence of the respective species.

This simple yet useful view assumes no significant gene flow between species such as horizontal gene transfer, reassortment, or introgression.

Early theoretical work included the analytical derivation of the probabilities for different gene tree topologies relating four individuals from two different species and showed that when the two populations diverged only recently an incorrect tree is not the exception but a common occurrence (Tajima 1983). Analytical results were also known for three individuals from three species (Nei 1987).

By the late 1980s, the discrepancy between species trees and gene trees was considered common knowledge, and Pamilo and Nei (1988) suggested that combining information from several independent loci was better than adding more samples. Pamilo and Nei also mentioned that a short branch in a species tree makes it likely that a gene tree has a different topology irrespective of the rest of the tree.

There are many potential sources of discrepancy between gene trees and species trees, including horizontal transfer, lineage sorting, and gene duplication/extinction. Early approaches to species tree estimation in the face of multiple gene trees included a parsimony-based method for constructing a species trees topology from gene trees (Maddison 1997). Many of the sources of inconsistency between gene trees and species trees have since been subject to further research, with the focus being on the development of statistical inference procedures. In this paper, we will term models that emphasize incomplete lineage sorting as the main source of inconsistency between gene trees and species trees, "multispecies coalescent models."

Recent research into multigene phylogenetics demonstrates that the common approach of concatenating sequences from multiple genes generates the wrong kind of average (Degnan and Rosenberg 2006) and can lead to poor estimation of the species tree (Kubatko 2007). Although

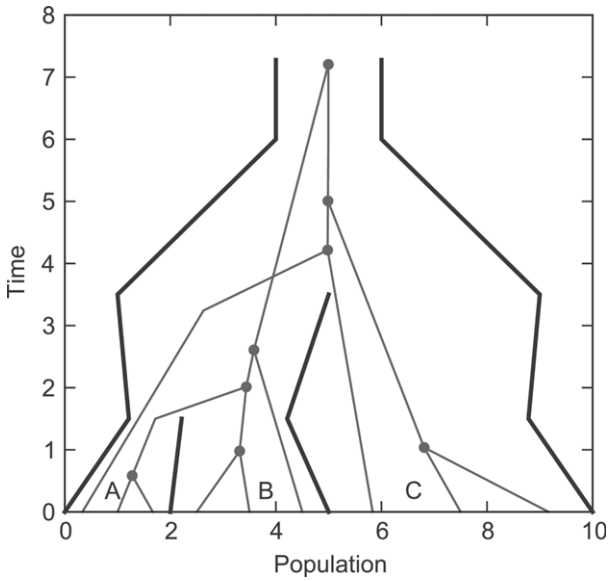

FIG. 1. Species tree visualization. One locus for three individuals from each of the three species giving a total of nine samples. Current population size ($t = 0$) of A is 2 and at time 1.5 (where it split from B) the population size is 1.

this common practice of concatenation can result in a well supported but incorrect tree, it is still a widely used method (Rokas et al. 2003; Wu and Eisen 2008), largely because of a lack of alternatives.

It has also been shown that the straightforward procedure of using the estimated gene tree topology that occurs most often among set of loci can be asymptotically guaranteed to produce the wrong estimate of the species tree in the so-called anomaly zone (Degnan and Rosenberg 2006). Two recent studies examined the performance of various methods in this problematic region of species trees (Huang and Knowles 2009; Liu and Edwards 2009).

A number of researchers have taken advantage of the multispecies coalescent model to develop methods that reconcile a set of gene trees with a shared species tree (Wilson and Balding 1998; Rannala and Yang 2003; Wilson et al. 2003; Liu and Pearl 2007; Liu et al. 2008). The multispecies coalescent assumes that each gene tree represents the relationships between orthologous genes from a small sample of individuals from multiple species and that there is no horizontal gene transfer or admixture between individuals from different species. A number of Bayesian approaches to inference have been developed in this context. The software package BATWING (Wilson and Balding 1998; Wilson et al. 2003) was developed to estimate a species tree from a single gene tree, including the times of speciation, population sizes, and growth rates. The MCMCcoal (Rannala and Yang 2003) software package estimates ancestral population sizes and divergence times on a known species tree based on a strict molecular clock and multiple gene trees. Finally, BEST provides estimates of the species tree topology, divergence times, and ancestral population sizes from a set of gene trees via an importance sampling method (Liu and Pearl 2007; Liu et al. 2008).

Multilocus species tree inference and the multispecies coalescent continue to be the subject of intense interest (Degnan and Rosenberg 2009; Knowles 2009; Liu et al. 2009). STEM (Kubatko et al. 2009; McCormack et al. 2009) is a new method that infers species trees from user supplied gene trees, population sizes, and gene tree rates. STAR and STEAC (Liu et al. 2009) are methods that use coalescence summary statistics as the basis of inference. In addition to the development of these new methods, the performance of consensus methods of species tree estimation has also been investigated (Degnan et al. 2009).

The Species Tree

Coalescent theory explicitly links the effective population size with the ancestral history of a small sample of genes from a population. In the context of Bayesian phylogenetic analysis, the coalescent acts as a prior distribution for gene trees. In its basic form, it is restricted to analyzing genes of individuals from the same species but it can be extended in a natural way to serve as a prior when building a multiple species phylogeny.

Please bear in mind that "species" above and in the rest of the text is not necessarily the same as a taxonomic rank, but designates any group of individuals that, after some "divergence" time, have no history of breeding with individuals outside that group. A species tree defines barriers for gene flow, and so the term is a catch all for taxonomic rank, subspecies, or any diverging "population structure."

A species tree specifies ancestral relationships (tree topology), the times ancestral species separated into two species (divergence times), and the population size history for each species. Each species (extant or ancestral) is represented by one branch of the species tree.

Gene trees are "embedded" inside a species tree by following the stochastic coalescent process back in time from the present within each branch, a process known as a multispecies coalescent or the "censored coalescent" (Rannala and Yang 2003). A species tree can be visualized by setting the *y* axis proportional to time and the intervals on the *x* axis proportional to population size as shown in figure 1 (Wilson et al. 2003).

Multiple samples per species are necessary for a complete estimation. Even two samples per species are sufficient, given enough loci. A single sample means no coalescent events for that extant species and so no information to estimate population size. This may in turn have a detrimental effect on inferring speciation times and perhaps even species topology.

In this paper, we describe a full Bayesian framework for species tree estimation. We have attempted to combine the best aspects of previous methods to provide joint inference of a species tree topology, divergence times, population sizes, and gene trees from multiple genes sampled from multiple individuals across a set of closely related species. We have achieved this by extending BEAST, a large existing open source software package for Bayesian phylogenetic

inference (Drummond and Rambaut 2007). The new method is named *BEAST (pronounced "star beast").

Methods

Given data D, we define the posterior distribution of the complete species tree S as follows:

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) \mathrm{d}G. \quad (1)$$

The data $D = d_1, d_2, \ldots, d_n$ is composed of n alignments, one per locus. $G = (G_1 \times G_2 \times \cdots \times G_n)$ is the space of all gene trees over the respective alignments, where $g_i \in G_i$ is one specific gene tree.

The term $P(d_i|g_i)$ is the "Felsenstein" likelihood of the ith sequence alignment given a gene tree (Felsenstein 1981), $P(g_i|S)$ is the multispecies coalescent, and $P(S)$ is a prior distribution on the space of species trees. The model assumes no recombination within loci and free recombination between loci.

In the context of species tree inference, gene trees act as "nuisance parameters" and are integrated out in the posterior. Direct evaluation of this integral is not possible, but it can be approximated using Markov chain Monte Carlo (MCMC) and a large amount of computation.

Computing the Multispecies Coalescent

The multispecies coalescent likelihood of a gene tree g embedded in species tree S is computed by combining the likelihood over all branches b of S,

$$P(g|S) = \prod_{b \in S} P(L_b(g)|N_b(t)). \quad (2)$$

Here, $L_b(g) = \{l, (t_0, t_1, \ldots, t_k, t_{k+1})\}$ is the lineage history of g over b and $N_b(t), t_0 \leqslant t \leqslant t_{k+1}$ is the effective population size function over b. t_0 and t_{k+1} are, respectively (moving back in time), the start and the end times of b and l is the number of lineages at time t_0 in the ancestral species represented by b. $l - k$ lineages remain at time t_k and this is unchanged at time t_{k+1}.

For example, the lineages of species C in figure 1 decrease from three at time 0–2, and the lineages of the (A,B) ancestor decrease from four to two.

The likelihood over a single branch is adapted from the equations given in Griffiths and Tavare (1994) and Heled and Drummond (2008) to account for no coalescent event during the last interval (t_k to t_{k+1}),

$$P(L_b(g)|N_b) = \prod_{i=0}^{k-1} \frac{1}{N_b(t_{i+1})} \prod_{i=0}^{k} \exp\left(- \int_{t_i}^{t_{i+1}} \frac{\binom{l-i}{2}}{N_b(t)} \mathrm{d}t \right). \quad (3)$$

The exact form of the demographic function N_b is left unspecified in equation (3). Most models in the literature assume a constant population size over the lifetime of the species, that is, $N_b(t) = c_b$. In addition to this simple model, we offer a model where population size changes linearly over the branch with one additional continuity constraint: The sum of population size of the two new species at the time of split is equal to the population size of the ancestral species. The example in figure 1 illustrates this continuous model.

The Species Tree Prior

The prior on the complete species tree is composed of the prior on the tree divergence times, $f_{\mathrm{BD}}(S)$, and the a prior on population sizes, $P_N(S)$:

$$P(S) = f_{\mathrm{BD}}(S) P_N(S). \quad (4)$$

The species topology prior is uniform on ranked labeled trees. For the divergence times, we use the reconstructed birth–death process (Gernhard 2008), parameterized by lineage birth and death rates λ and μ:

$$f_{\mathrm{BD}}(S) = n_s! \lambda^{n_s - 1} (\lambda - \mu) \frac{\mathrm{e}^{-(\lambda-\mu)x_1}}{\lambda - \mu \mathrm{e}^{-(\lambda-\mu)x_1}} \times \prod_{i=1}^{n_s-1} \frac{(\lambda-\mu)^2 \mathrm{e}^{-(\lambda-\mu)x_i}}{\lambda - \mu \mathrm{e}^{-(\lambda-\mu)x_i}}, \quad (5)$$

where n_s is the number of species and $x_1, x_2, \ldots, x_{n_s-1}$ are the divergence times of the species tree, x_1 being the root height.

Typically, there is no a priori knowledge regarding the birth/death rates, so this is in fact a hyperprior where both hyperparameters are estimated and a noninformative prior is used for both. In our implementation, the parameters are $r = \lambda - \mu$ and $a = \frac{\mu}{\lambda}$, and the priors used are uniform on $[0, 1]$ on a and $f(x) = 1/x$ for r.

The prior for population sizes depends on the model used. For constant population per branch, the population size is assumed to be a sample from a $\Gamma(2, \psi)$ distribution—a gamma distribution with a mean 2ψ and a shape of 2:

$$P_N(S) = \left(\prod_{b \in S} \frac{1}{\psi^2} N_b \mathrm{e}^{-N_b/\psi} \right) P_\psi(\psi). \quad (6)$$

Unless we have some specific knowledge about population size, we again use the noninformative prior $P_\psi(x) = 1/x$ for hyperparameter ψ.

In the continuous linear model, we have n_s population sizes at the tips of the species tree, and two per each of the $(n_s - 1)$ internal nodes, expressing the staring population size of each of the descendant species. The prior for the population sizes at the internal nodes are as above, but for the ones at the tips, they are assumed to come from a $\Gamma(4, \psi)$ distribution. Because $X_1, X_2 \sim \Gamma(2, \psi)$ implies $X_1 + X_2 \sim \Gamma(4, \psi)$, this corresponds to having the same prior on all final (most recent) population sizes of both extant and ancestral species.

Results

Examining Rapid Radiation Using Simulated Data

The first set of simulations explored a region of species tree space with seven extant species. A multipart data set was constructed as follows.

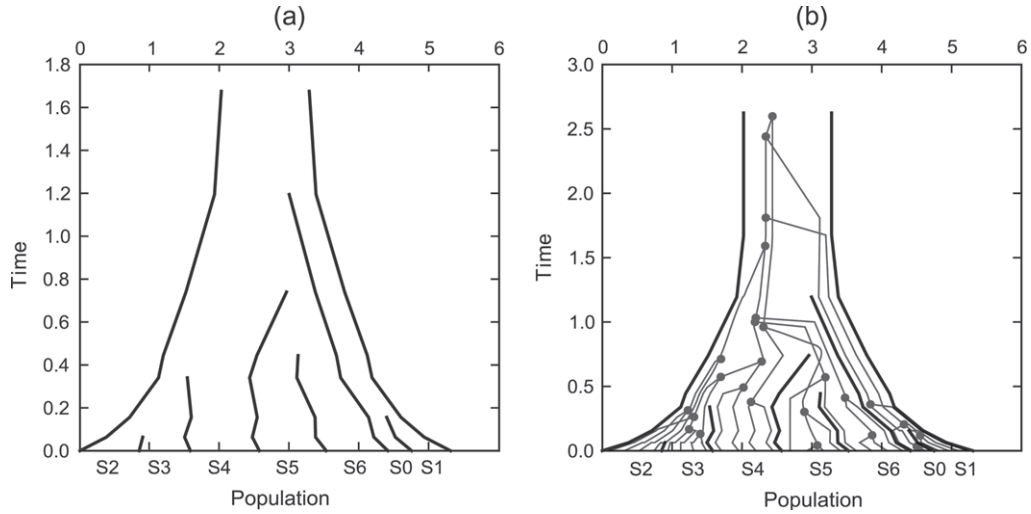

FIG. 2. (*a*) A simulated species tree from a birth–death process with continuous population sizes. (*b*) A single gene tree embedded inside the same species tree.

One hundred species trees were simulated using a birth–death process with $\lambda = 1$ and $\mu = 0.2$ (Gernhard 2008). For each species tree, a linear population size function was randomly assigned to each branch. First, each extant species population size at t_0 was set to $\rho \times z$, where z was a random Gaussian variate, $z \sim N(1.1, 0.4)$ and $\rho = \lambda - \mu = 0.8$. Second, because the continuity constraint determines the population size at the tip-ward end of each ancestral species, we only needed the population size at the root-ward end of each ancestral branch to be specified. This value was drawn from a log-normal distribution with a mean of $p_0 \times f$ in real space and variance $\nu = 0.4$, where p_0 is the population size at the tip-ward end of the branch. The reduction factor f controls the amount of population growth from the root to the tips. With $f = 1$, the total population size, across all species, would stay roughly the same, that is, the sum of the population size of two descendant population would on average be equal to the population size of the ancestral species. For our purposes, f was set to 0.7, generating trees with moderate expansion of total population across all species. Note that this population size simulation does not exactly match the prior we employ for inference.

One simulated tree is shown in figure 2*a*. A set of gene trees "embedded" inside the species tree were simulated, each with four individuals sampled per species. This was repeated for each species tree with 1, 2, 3, 4, 8, 16, and 32 independent loci. A single locus from one run is shown in figure 2*b*.

Finally, sequences (1,600 bp long) were simulated for each gene tree using the Jukes–Cantor model under a strict clock and a substitution rate of $r_\mu = 0.005$. Interpreting the simulated speciation times as millions of years, this rate corresponds to a real-life substitution rate of 5×10^{-9} per site per year. The overall scenario is of a rapid species radiation starting between 0.49 and 4.9 (mean 1.9) million years ago.

Seven Species Rapid Radiation Results

It is not immediately obvious how to summarize 700 runs of species tree estimation. It is typical to focus on a point estimate of the species tree topology and associated clades probabilities, but this is unsatisfactory here for several reasons. First, a Bayesian method generates a set of trees drawn from the posterior, not a single tree. Second, looking only at the topology ignores estimation of speciation times and population sizes. Third, two trees with different topologies may in fact be very similar when divergence times are considered. Finally, using only one bit of information per test would require running many replicates to obtain an accurate measure of the performance of the method.

We have attempted to define some performance measures that make the most of the simulations carried out. Figure 3 shows the averages of several measures computed from the posterior distribution of species trees; each graph point was obtained by averaging results from 100 runs, where each run was produced by the Bayesian analysis of a single simulated data set. Figure 3*a* plots two error measures, whereas figure 3*b* shows the mean number of species tree topologies in the 95% credible interval.

The first error measure is the "Normalized Rooted Branch Score." This is the rooted equivalent of the "branch score" metric suggested by Kuhner and Felsenstein (1994) and is defined in Appendix. A tree metric provides a direct way of measuring the distance of the estimate from the true species tree. Normalizing by the tree length makes it possible to meaningfully average this distance across runs with different simulated species trees. The second measure is the "Normalized Rooted Tree Score." This is an extension of the branch score that incorporates the population sizes. The definition is similar but uses the length of the branch in coalescent units, that is, the integral of the inverse of population size over the branch (see Appendix).

Although the tree score measures overall performance, some specific point estimates such as speciation times are

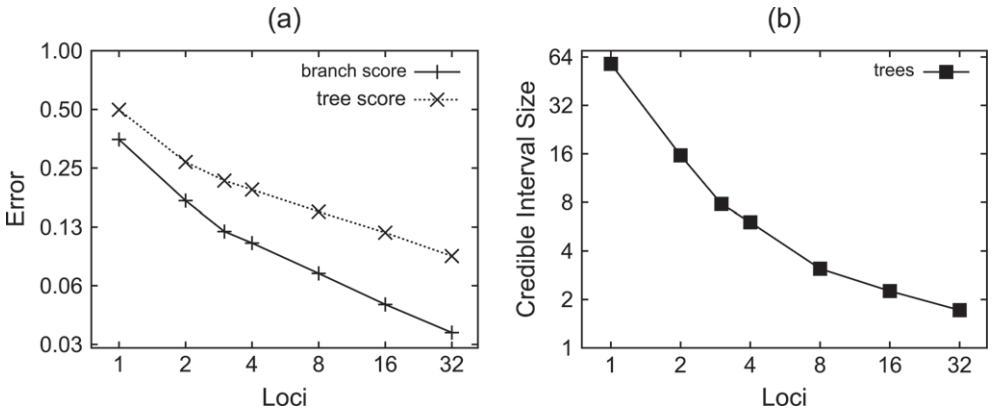

FIG. 3. (*a*) Species tree estimation error and (*b*) 95% credible interval size as a function of the number of loci. The number of individuals sampled per species is four for all experiments. Each graph point is obtained by averaging the error measure (described in the main text) over 100 analyses of simulated data sets. The "branch score" is a measure of the distance in tree space of the estimated species tree to the true tree, incorporating both topology and divergence times. The "tree score" is a measure of the distance between the estimated species tree and the true species tree incorporating information about the population size as well. For details of the tree metrics used, see main text.

of interest as well. However, note that evaluating the errors in point estimates of speciation times and population sizes can be done only for clades that appear in the true species tree. Figure 4*a* summarizes the estimation errors in speciation times and population sizes for all runs.

Errors are computed as the absolute value of the difference between the posterior median ν_i and the true value ν, normalized by the true value:

$$\text{Err} = \frac{\sum_{i=0}^{N} \left| \frac{\nu_i - \nu}{\nu} \right|}{N}. \quad (7)$$

The credible interval is the 95% highest posterior density (HPD) interval calculated from the posterior samples. Figure 4*b* shows the credible interval size of speciation time and population size point estimates, calculated as the credible interval range ($h_i - l_i$), normalized by the true value ν,

$$\text{HPD size} = \frac{\sum_{i=0}^{N} \frac{h_i - l_i}{\nu}}{N}. \quad (8)$$

Another statistic of interest is the frequentist coverage; that is, the percentage of estimates where the true value falls inside the credible interval. Coverage statistics are harder to estimate accurately given the relatively small number of runs, but in our analyses, they ranged between 89% and 98%.

Recent speciation events were harder to estimate (in terms of relative error) than older ones and so contributed more to the overall error. Figure 5 illustrates this by plotting the relative error as a function of split time for the two loci data set.

Seven Species Rapid Radiation—Effect of Sequence Length

Simulation may use any arbitrary sequence length, but real-world sequences are subject to practical limits such as time, cost, primer suitability, and haplotype block sizes. Figure 6 shows the results of analyzing the four loci data set where sequence length starts at 6,400 bp and is halved five times down to 200 bp. The vertical line shows the error for an "infinite" sequence, which is obtained by repeating the analysis using the simulated gene trees directly and no sequence data. Please note that long but recombination free sequences are not likely for species with large populations. The parameter ranges of simulated data were chosen to

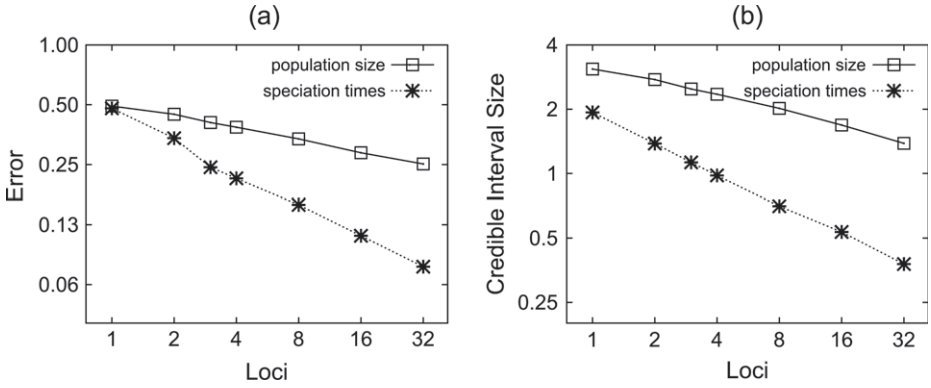

FIG. 4. (*a*) Relative error and (*b*) credible interval size for both population size and speciation time point estimates. The number of individuals sampled per species is four for all experiments. Each graph point is obtained by averaging over 100 analyses of simulated data sets (see main text for details).

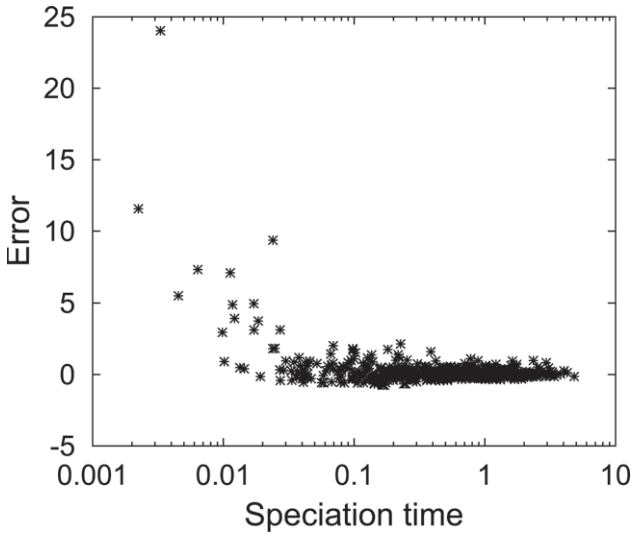

FIG. 5. Speciation time estimate error as a function of speciation time. Data are taken from the main 100 runs with two loci, four individuals per species, and the sequence length of 1,600 bp.

illustrate trends and may extend beyond real-life conditions. In our simulations, increasing the sequence length from 200 to 1,600 roughly halved the error associated with speciation time estimates (from 0.41 to 0.21). In comparison, increasing the sequence length from 200 to 800 roughly halved the average number of trees in the 95% credible set (from 18.5 down to 8.2).

Estimation of Relative Substitution Rates for Different Loci

So far sequence data were generated assuming that all loci evolve at the same rate according to a strict molecular clock, not something that can be justified for most real data sets. In this section, we examine the effect of allowing each gene to have a separate substitution rate. In general, it is not possible to estimate the absolute molecular clock rate for contemporaneous data sets, but it is possible to estimate relative rates. In our case, this means fixing the substitution rate of the one locus and estimating the rates of the other loci relative to the reference locus.

For simplicity, we reuse the setting of the four loci seven species rapid radiation. We set the substitution rate of the first locus to $r_\mu = 5 \times 10^{-3}$ and set the rates of the other loci to $r_\mu \times z$, where z is a random number distributed uniformly in $[0.2, 2]$.

Sequence data were regenerated under this new scenario and analyzed in the same way as described for the previous simulations. The analysis was carried out twice: Once under the correct model that incorporated variation in substitution rate across loci, and once assuming a single substitution rate across all loci, in order to measure the effect of this form of model misspecification.

When estimating the relative rates across loci, the point estimates of the rates had a mean error of 1.21 and an HPD size of 1.99, both values computed as explained for the other point estimates. In addition, the relative ordering of rates from each run was compared with their true ordering by computing the permutation distance: In 61 cases, the distance was 0, 30 had a distance 1, and 7 distance of 2, giving a mean permutation distance of 0.5 over the 100 runs. This suggests that our method is successfully able to discriminate between fast and slow loci in this set of simulations.

The performance of species tree estimation under rate variation across loci is shown in table 1. The two new sets of analyses are summarized in rows 2 and 3, and the first row shows the results of the equal rates simulations for comparison.

The results show that incorrectly assuming the rates across loci are equal had only a small impact on the estimation of species tree topology. However, ignoring rate variation did adversely affect the estimation of species divergence times, as shown by the drop in coverage percentage: Only 67% (compared with 93%) of the true speciation times fell inside the associated 95% credible interval.

Effect of Number of Sampled Individuals per Species

Next we considered the effect of varying the number of individuals sampled from each species. Gene trees for 32 individuals per species and four loci were simulated for each of the 100 species trees, then 16 individuals were removed to

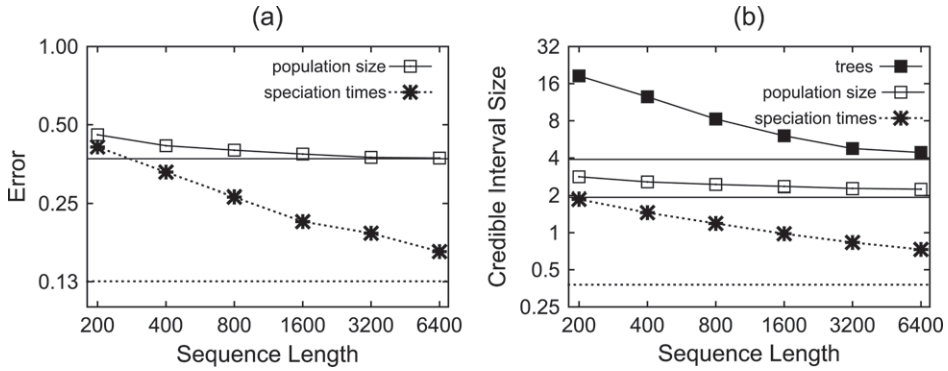

FIG. 6. (*a*) Relative error and (*b*) credible interval sizes as a function of sequence length. The number of individuals sampled per species is four for all experiments. The number of independent loci are four for all experiments. Each graph point is obtained by averaging over 100 analyses of simulated data sets. The horizontal line represents the theoretical maximum when sequence length approaches infinity and is calculated by using the gene trees directly without error.

Table 1. Summary of Seven Taxa, Four Loci Species Tree Estimation Where Genes Evolve at Different Rates. The Final Row Represents the Case Where the Model is Misspecified: The Truth is That Each Gene Has a Different Rate, But the Method Assumes That All Genes Have the Same Substitution Rate.

Data/model	Topology inside 95%	Mean 95% size	Normalized branch score	Normalized tree score	Speciation time inside 95%	Speciation time error/CI size	Population size error/CI size
Equal rates/equal rates	94	7.86	0.10	0.19	93	1.36/10.0	2.2/162
Rates vary/rates vary	95	8.08	0.12	0.19	93	1.43/12.0	2.2/189
Rates vary/equal rates	92	10.24	0.16	0.20	67	1.57/12.6	2.5/189

CI, credible interval.

leave 16, then halved again to 8, and so on down to 2 individuals per species. To reduce the considerable computational cost involved, the analysis was carried out using the gene trees directly, that is, without sequence data. The results are shown in figure 7.

Increasing the number of individuals sampled from each species up to 32 results in a marked improvement in all aspects of the species tree estimation, including population sizes. This result seems to be different to what is typically found when considering sampling schemes for a population. For a single population, the number of independent loci is the major factor and the return from additional individuals diminishes quite quickly. Our somewhat unexpected result may be specific to rapid radiations, where population sizes are comparable in size to the species lifetime. Under those conditions, additional sampled individuals result in more lineages crossing the species boundary (looking backward in time), adding more power to the estimate of divergence times and population sizes.

Comparison with BEST

In 2007, Liu and Pearl (2007) and Liu et al. (2008) introduced BEST, a Bayesian method whose goals are similar to *BEAST. Both software packages estimate species tree topology, divergence times, and population sizes from gene trees under a multispecies coalescent model. Both extend known and well-established software packages MrBayes (Huelsenbeck and Ronquist 2001) and BEAST (Drummond and Rambaut 2007)—providing users the convenience of specifying powerful and well-tested models for gene tree estimation. There are various modeling differences: BEST requires an outgroup, population size is assumed constant over the branch, and the species tree prior is uniform. However, there is also a major difference in the computational strategies employed—BEST estimates each gene tree individually, then infers the species tree in two additional stages using importance sampling. In contrast, *BEAST coestimates the species tree and all gene trees in one Bayesian MCMC analysis. We know that the main factor in estimating population size is the number of independent loci, and so believe that estimation will be better when using information from all gene trees simultaneously.

To compare the two methods, 100 four-loci species trees have been generated. The process described in the beginning of the Methods section was modified as follows to conform with BEST requirements: Trees with eight species were simulated but only those with a 7/1 split at the root were retained with the lone species forming an outgroup. Each species had four individuals except the outgroup which had only one, and finally, the population size was set to be constant along the branch with a value of the mean of the sizes at the start and at the end of the branch.

Table 2 summarizes the 100 runs, the error measures from BEST are substantially larger than the ones from *BEAST, and the coverage percentages for point estimates are much lower.

Comparison to Concatenation

A final experiment involving the simulated data sets compared the performance of *BEAST with the standard supermatrix method (concatenation). For each of 100 data sets (four loci, 1,600 bp), we selected a single random individual per species and concatenated the four alignments into a single supermatrix of size 7 species × 6,400 sites. We then used

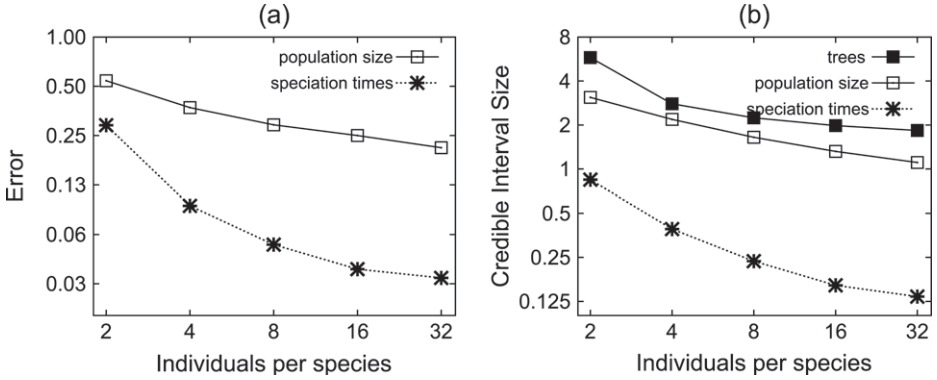

Fig. 7. (*a*) Relative error and (*b*) credible interval sizes, as a function of number of individuals sample from each species. Each graph point is obtained by averaging over 100 analyses of simulated data sets. The analysis used the true gene trees to reduce the computational cost.

Table 2. Comparison of *BEAST, BEST, and "Supermatrix" Performance in Estimating Species Trees.

	Topology inside 95%	Mean 95% size	Normalized branch score	Normalized tree score	Speciation time inside 95%	Speciation time error/CI size	Population size inside 95%	Population size error/CI size
*BEAST	97	11.78	0.10	0.20	96	0.41/1.49	98	0.33/1.11
BEST	88	12.88	0.58	0.64	56	1.32/2.06	56	0.59/2.15
Supermatrix	9	1.4	0.77	NA	0.7	21.11/5.27	NA	NA

NA, not available; CI, credible interval.

BEAST to estimate the species tree for each of these 100 concatenated data sets using a yule tree prior. Only 9 of the 100 replicates contained the true species tree topology in the 95% credible set of tree topologies. This is compared with 98 for *BEAST and 88 for BEST. The coverage percentage for speciation times was close to 0. The full set of measures can be found in table 2. It is clear that both multispecies coalescent methods are far superior to the supermatrix method.

Pocket Gophers

Belfiore et al. (2008) investigated the rapid species radiation in *Thomomys*, a genus of pocket gophers spread over a wide range including Texas, the Dakotas, Baja California, and the southern edge of Canada. The authors collected data from 28 individuals belonging to seven Thomomys species and two outgroups from the "sister tribe." Seven noncoding nuclear sequences were sequenced from each individual (alignment length 471 to 819 bp), though this was not possible for some of the outgroups.

The 28 individuals were distributed across the species as follows: two from each of 5 species, 3 northern pocket gophers (*Thomomys talpoides*), and 12 Botta's pocket gophers (*Thomomys bottae*). This uneven distribution of individuals among species is not optimal, but it does meet the minimum of two individuals from each of the species of interest.

For the analysis, we used a general time reversible substitution model and strict molecular clock with a separate mutation rate for each locus as described in the simulations section. The highest posterior tree is shown in figure 8*a*. There are four strongly supported clades, but the outgroup is not where we expect it to be.

Experts are certain that the *Orthogeomys heterodus* is a proper outgroup (Belfiore NM, personal communications), and we can incorporate this knowledge into our Bayesian model as part of the prior. Figure 8*b* shows the result on an almost identical run, where the proper relation of the outgroup is a priori enforced. The divergence times on the two trees are virtually identical except for the outgroup.

Given the proper tools, the reason for this discrepancy is easy to spot. Figure 9 shows the median species tree and embedded gene trees for the first run. The tendency to place the outgroup incorrectly appears to be caused by just one gene out of the six (TBO29 in green), which is closer to the (*T. bottae*, *Thomomys townsendi*, *Thomomys umbrinus*) clade. It is a good reminder of the fact that species tree are not the result of a consensus process—but of a "reverse auction"—where the lowest bidder sets the limit. Here, the other genes "do not care" as their shared ancestry takes place as expected in the common ancestor.

Belfiore et al. made a considerable effort to collect and sequence an outgroup, but it is only a requirement for MrBayes/BEST. *BEAST does not require any outgroup as it follows the BEAST mantra: "There are no unrooted phylogenies—only phylogenies whose root position is uncertain." *BEAST uses either a strict or a relaxed molecular clock in order to estimate the roots of the individual gene trees, which in turn combine, through the multispecies coalescent, to estimate the root of the species tree.

Discussion

A natural consequence of taking a model-based approach to statistical phylogenetics is the continuous refinement of the model to better reflect our understanding of the biological reality of evolving populations. Tempering this refinement is a need to control the number of parameters in the model. This balance necessitates a focus on modeling only

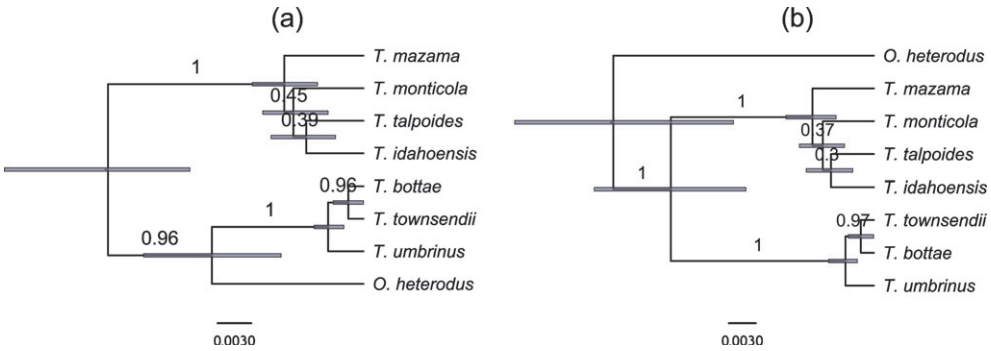

Fig. 8. Phylogeny for seven groups of western pocket gophers (Geomyidae, *Thomomys*). The analysis is based on seven noncoding nuclear genes from 28 individuals. Clade posterior probability is indicated on the branch. (*a*) Analysis with no monophyly constraints and (*b*) analysis with ingroup monophyly enforced.

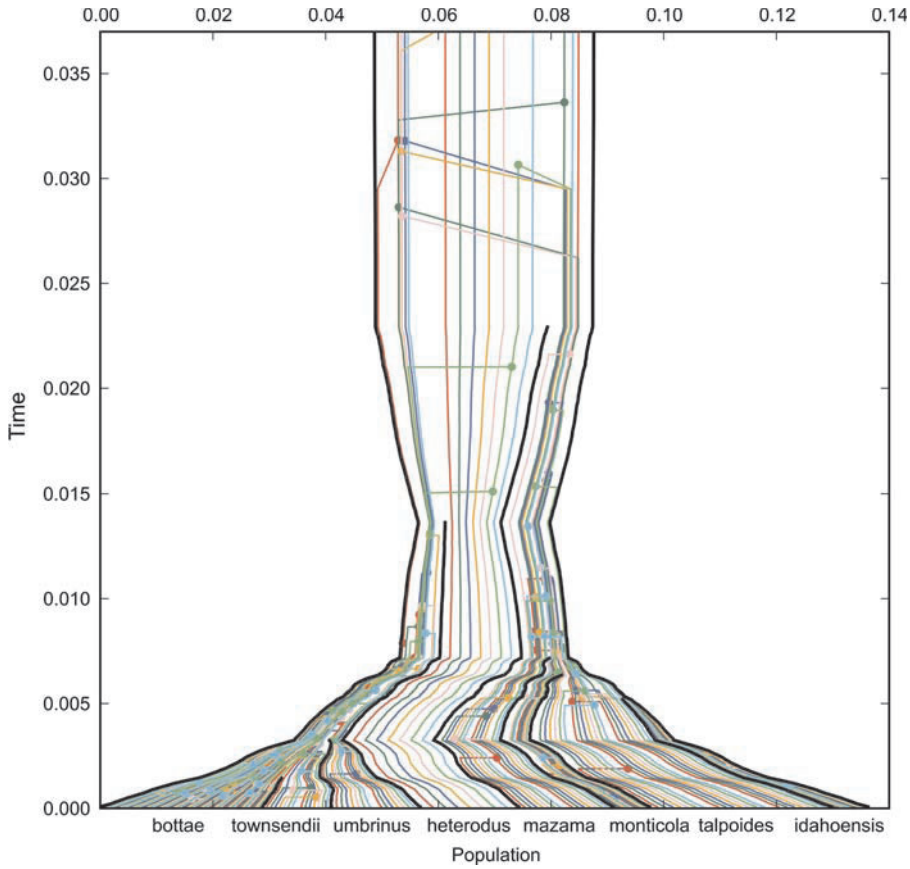

Fig. 9. Western pocket gophers (Geomyidae, *Thomomys*) species tree with embedded gene trees, each in a different color. The species tree was generated using median estimates for the divergence times and population sizes. Note that this representation is for the purpose of visual inspection only, and any inferences should be made directly from the posterior data.

the essential details of the evolutionary process. It has been known since Mendel's experiments (Orel 1996) that "unlinked" genetic loci segregate independently of each other within a species or isolated population, but it was not until a few decades ago that the implications of independent segregation were fully appreciated in the context of gene trees. Although the model we have described here is not new (Maddison 1997; Rannala and Yang 2003) (apart from small details such population size function and its prior), the contribution we have made is to describe and implement for practical use a full Bayesian inference of the species tree under the multispecies coalescent model. Our implementation is based on a popular existing software package for Bayesian phylogenetics, and as a result it can exploit existing models for gene trees such as relaxed molecular clocks (Drummond et al. 2006) and previously implemented priors for species tree such as the reconstructed birth–death prior (Gernhard 2008).

There is no doubt that our proposed model still represents a very idealized view of the genetic relationships of multiple loci between individuals from closely related species. However, we believe that despite its obvious simplifications, it represents a major improvement on the standard approaches to multigene phylogenetics. Besides the ability to model incomplete lineage sorting and ancestral polymorphism, our implementation also provides a very natural way to include multiindividual data and missing data into a phylogenetic analysis.

There are two obvious ways in which our model is deficient. The first, and arguably most important, is that it lacks any modeling of recombination within a locus. For nuclear loci in eukaryotes, recombination rates may often be comparable to mutation rates. The impact of this on species tree estimation under the multispecies coalescent has not been assessed; however, it is known that the presence of recombination can have a large biasing effect on estimated population sizes, if no recombination is assumed in the model (Schierup and Hein 2000).

The second deficiency of our model is, arguably, our treatment of speciation events. It has been suggested that incorporating a substantial period of limited gene flow as a transition between a single ancestral species and two descendant species is more realistic (Hey and Nielsen 2004). This approach has been taken for coalescent inference of sister species in the IM/IMα software packages (Hey and Nielsen 2007). Although we think that this is a good research direction and expect multispecies versions of isolation with migration models to be popular, we remain uncertain about how much power there will be to perform inference under such models, without making very strong prior assumptions

about the length of the transition period and associated migration rates.

Conclusions

A transition from single gene to multigene analyses in molecular systematics is well underway—and the extension of this transition into the fields of molecular ecology and phylogeography has revealed the need to more accurately model the relationships among gene trees at different loci. Although it is well established that "more loci are better" for estimating population sizes in a single population (Pluzhnikov and Donnelly 1996; Felsenstein 2006), the optimal sequencing strategy for phylogenetic questions is not yet established (but see Maddison et al. 2006). Our results lend weight to the growing notion that sequencing multiple independent loci from a small representative sample of individuals can, for many questions, yield better results than sampling a large numbers of individuals at just one genetic locus. However, it also seems clear that additional individuals per species provide a significant contribution to the accurate estimation of species tree divergence times and topologies, at least under the rapid radiation scenario studied here.

In conclusion, we agree with a recent suggestion that the multispecies coalescent represents a step toward the unification of molecular systematics and phylogeography (Edwards and Rausher 2009)—however, we can also see a number of natural further steps that need to be taken to address common situations facing researchers in molecular ecology and phylogeography. For example, it is often the case that the exact number of species, and the assignments of individuals to species or subspecies, is uncertain in recently radiated groups (Meyer 1993; Das et al. 2004; Glor et al. 2004; Belfiore et al. 2008; Leache 2009).

Using morphological and geographical data to define pairwise probabilities of species identity, we can envisage extensions of the presented method that not only sample the species tree topology, divergence times, and population sizes but also estimate the total number of species and consequently the assignments of individuals to species. This would provide similar capabilities to population structure inference packages such as STRUCTURE (Pritchard et al. 2000) and STRUCTURAMA (Huelsenbeck and Andolfatto 2007) but would simultaneously provide Bayesian inference about the relationships between the different populations/species. We hope that this general line of research will eventually lead to a model-based synthesis of the fields of population genetics, phylogenetics, and phylogeography.

Acknowledgments

We would like to thank David Bryant for invaluable discussions and Natalie Belfiore for helpful comments on the analysis of pocket gophers. J.H. was supported by Marsden grant UOA0502.

Appendix

Normalized Rooted Branch Score

To measure the distance between two rooted trees, we use the rooted branch score defined as

$$|T_1, T_2| = \sqrt{\sum_{c \in T_1 \cup T_2} (B(T_1, c) - B(T_2, c))^2}, \quad (9)$$

where $B(T, c)$ is the length of the branch connecting clade c to the tree if it is present in T, 0 otherwise.

The posterior estimate is the mean of this distance to the target over all posterior trees. This is normalized by tree length, so the score units can be interpreted as a percent. The normalization also allows us to meaningfully average scores from different runs.

$$\frac{1}{|T|} \frac{\sum_{i=0}^{N} |T, T_i|}{N}. \quad (10)$$

Normalized Rooted Tree Score

The distance between two rooted species trees is defined as:

$$|T_1, T_2| = \sqrt{\sum_{c \in T_1 \cup T_2} (D(T_1, c) - D(T_2, c))^2}, \quad (11)$$

where $D(T, c) = \int_{b_0}^{b_1} \frac{1}{N_b(t)} dt$, and b is the branch connecting clade c to the tree if it is present in T, 0 otherwise. The score is normalized by the "tree area," which is the total tree length in coalescent units.

If all populations are constant and equal to 1, this reduces to the branch score. Note that unlike the branch score, this is not a true metric because branch length and population size are confounded.

References

Belfiore NM, Liu L, Moritz C. 2008. Multilocus phylogenetics of a rapid radiation in the genus Thomomys (Rodentia: Geomyidae). *Syst Biol*. 57(2):294.

Das A, Mohanty S, Stephan W. 2004. Inferring the population structure and demography of Drosophila ananassae from multilocus data. *Genetics* 168(4):1975–1985.

Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. 2009. Properties of consensus methods for inferring species trees from gene trees. *Syst Biol*. 58(1):35.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet*. 2(5):e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*. 24(6):332–340.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 4(5):e88.

Drummond AJ, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 7:214.

Edwards SV, Rausher M. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63(1):1–19.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.

Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol Biol Evol*. 23(3):691–700.

Gernhard T. 2008. The conditioned reconstructed process. *J Theor Biol*. 253(4):769–778.

Glor RE, Gifford ME, Larson A, Losos JB, Schettino LR, Lara ARC, Jackman TR. 2004. Partial island submergence and speciation in an adaptive radiation: a multilocus analysis of the Cuban green anoles. *Proc R Soc Lond Ser B Biol Sci*. 271(1554):2257–2265.

Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*. 344: 403–410.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evol Biol*. 8(1):289.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics* 167(2):747–760.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci USA*. 104(8):2785.

Huang H, Knowles LL. 2009. What is the danger of the anomaly zone for empirical phylogenetics? *Syst Biol*. 58:527–536.

Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. *Genetics* 175(4):1787.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.

Knowles LL. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Syst Biol*. 58:463–467.

Kubatko LS. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*. 56(1):17–24.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25(7):971.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*. 11(3):459–468.

Leache AD. 2009. Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (Sceloporus). *Syst Biol*. 58:547–559.

Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. *Syst Biol*. 58:452–460.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*. 56(3):504–514.

Liu L, Pearl DK, Brumfield RT, Edwards SV. 2008. Estimating species trees using multiple-allele dna sequence data. *Evolution* 62(8):2080–2091.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol*. 53(1):320–328.

Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst Biol*. 58: 468–477.

Maddison WP. 1997. Gene trees in species trees. *Syst Biol*. 46(3): 523–536.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*. 55(1):21.

McCormack JE, Huang H, Knowles LL. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst Biol*. 58:501–508.

Meyer A. 1993. Phylogenetic relationships and evolutionary processes in East African cichlid fishes. *Trends Ecol Evol*. 8:279–284.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Orel V. 1996. Gregor Mendel: the first geneticist. New York: Oxford University Press.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol*. 5(5):568–583.

Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144(3): 1247–1262.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945–959.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164(4):1645–1656.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425(6960):798–804.

Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156(2):879–891.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460.

Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. *Genetics* 150(1):499–510.

Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J R Stat Soc Ser A (Statistics in Society)* 166(2): 155–188.

Wu M, Eisen J. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*. 9(10):R151.