High-resolution species trees without concatenation

Scott V. Edwards*†, Liang Liu‡§, and Dennis K. Pearl‡

*Department of Organismic and Evolutionary Biology, and Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138; and ‡Department of Statistics, Ohio State University, Columbus, OH 43210-1247

Edited by Joseph Felsenstein, University of Washington, Seattle, WA, and approved February 13, 2007 (received for review August 15, 2006)

The vast majority of phylogenetic models focus on resolution of gene trees, despite the fact that phylogenies of species in which gene trees are embedded are of primary interest. We analyze a Bayesian model for estimating species trees that accounts for the stochastic variation expected for gene trees from multiple unlinked loci sampled from a single species history after a coalescent process. Application of the model to a 106-gene data set from yeast shows that the set of gene trees recovered by statistically acknowledging the shared but unknown species tree from which gene trees are sampled is much reduced compared with treating the history of each locus independently of an overarching species tree. The analysis also yields a concentrated posterior distribution of the yeast species tree whose mode is congruent with the concatenated gene tree but can do so with less than half the loci required by the concatenation method. Using simulations, we show that, with large numbers of loci, highly resolved species trees can be estimated under conditions in which concatenation of sequence data will positively mislead phylogeny, and when the proportion of gene trees matching the species tree is <10%. However, when gene tree/species tree congruence is high, species trees can be resolved with just two or three loci. These results make accessible an alternative paradigm for combining data in phylogenomics that focuses attention on the singularity of species histories and away from the idiosyncrasies and multiplicities of individual gene histories.

coalescent theory | importance sampling | molecular clock | yeast

Many biological disciplines have as their focus the phylogenetic relationships of species-species trees. With the advent of large-scale comparative genomic data sets and enhanced computational power, statistical methods such as maximum likelihood and Bayesian phylogenetic inference have provided sophisticated approaches to incorporation of heterogeneous models of sequence evolution in combined multilocus data sets (1–3). These methods have vastly increased the efficiency and statistical power that can be gleaned from DNA sequences and will greatly contribute to the ultimate goal of comprehending the full scope of Darwin's Tree of Life (4). The taxonomic units of the Tree of Life are species composed of large numbers of genes distributed across multiple independently segregating chromosomes and linkage groups. Thus, most phylogenetic studies in fact use methodologies that focus not on estimation of species trees per se but on estimation of gene trees, with the usual assumption being that the gene tree resolved by combining many genes is congruent with the species tree. This assumption will hold widely, except in cases when (*i*) horizontal gene transfer and other reticulate processes, such as interspecific gene flow, are common; (*ii*) gene duplication has caused gene lineage splits in the absence of splits in the history of species; and (*iii*) gene lineages fail to coalesce before divergence of species (looking backward in time). This paper addresses this latter problem, which occurs because of internodes in the species tree that are short when scaled by the effective population size of the relevant branches (reviewed in refs. 5 and 6), but the method also works in the absence of discordance between gene and species trees.

The current practice of concatenating sequences from genetically separate loci into a single supermatrix has its origins in debates in the 1990s about "total evidence" and does not permit heterogeneity among gene trees in phylogenetic analysis. Along with dense taxon sampling (7, 8), concatenation of sequences from multiple genes into supermatrices (9) is thought to maximize power to make inferences about species history (for recent examples see refs. 10–13). However, new analytical results suggest that for any species tree of five or more taxa, there exist branch lengths in the species tree (invariably short ones) for which gene trees that do not match the species tree are more common than gene trees matching the species tree, so-called anomalous gene trees (14). In such a situation, phylogenetic analysis of concatenated sequences can positively mislead inference of species relationships (15). Even when gene and species trees are topologically concordant, as occurs when species tree branch lengths are long, there is a need for phylogenetic methods that estimate species trees as distinct from gene trees, if only because species trees are a more realistic goal for systematics. Many recent models for estimating historical population parameters make reference to the species history in which gene histories are embedded (16–18), but phylogenetic inference itself still largely retains its focus on gene trees.

Some of these models, such as gene tree parsimony (19) or Takahata's method (20), estimate only the topology of the species tree. Supertree methods (21) have the advantage of being able to combine diverse sources of information but, unlike gene tree parsimony and methods based on population genetics, do not rely on any biological justification for explaining incongruences between genes or input trees. Population genetics models (21) jointly estimate the species tree topology and branch lengths and effective population sizes scaled by the mutation rate (μt and $\theta = 4\,\mathrm{N}\mu$, respectively), but overall, the few models making the distinction between gene and species trees either are not amenable to large scale multilocus analysis (22), lack a formal statistical framework, or are appropriate only for data sets for which some discordance exists (19, 20, 23, 24). Felsenstein (chapter 28 in ref. 2) outlined a likelihood extension of a model by Nielsen (23) that allows for sampling of more than one allele per species to estimating species trees, but this method has not yet been applied to real data.

Here we apply a Bayesian method whose explicit focus is the estimation of the distribution of species trees and that effectively deals with incongruences frequently observed in gene histories due to incomplete lineage sorting or deep coalescence (5, 6) as well as gene tree uncertainty. Many Bayesian phylogenetic analyses of multiple gene data sets assess congruence among

Author contributions: S.V.E. and L.L. contributed equally to this work; S.V.E., L.L., and D.K.P. designed research; L.L. performed research; S.V.E., L.L., and D.K.P. contributed new reagents/analytic tools; L.L. analyzed data; and S.V.E., L.L., and D.K.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: BEST, Bayesian estimation of species trees; MCMC, Markov Chain Monte Carlo.

†To whom correspondence should be addressed. E-mail: sedwards@fas.harvard.edu.

§Present address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138.

This article contains supporting information online at www.pnas.org/cgi/content/full/0607004104/DC1.

gene trees by analyzing them completely independently from one another, with each gene, for example, having a uniform prior on the topology. In this approach, which we call the independent model, topologies for gene trees are estimated without regard to information from other genes and, crucially, without any assumption that the topologies of different genes will be correlated because of shared species history (12). Particularly in Bayesian methods, where the prior can have an influence on posterior distributions of gene trees, such an approach could yield an unrealistically variable distribution of gene trees, because there is no common species tree constraining gene trees. Both the concatenation and independent model approaches are unrealistic to the extent they do not consider the partial correlation that must exist because of shared species history. The Bayesian hierarchical model we describe here improves on these options by using a joint model that uses the joint distribution of gene trees for many loci given a species tree as a prior (17). Such a prior has the advantage of allowing heterogeneity in gene trees as a means of estimating the posterior distribution of species tree topologies and branch lengths.

Results

Theory. Full details of the methods, which we call Bayesian Estimation of Species Trees (BEST), appear in supporting information (SI) *Materials and Methods* and in refs. 32 and 33. In step 1 (SI Fig. 4), our goal is to estimate the posterior distribution of gene trees (**G**) given the data (D), $f(\mathbf{G}|\mathrm{D})$. The prior on the gene trees, $f(\mathbf{G})$, comes from the distribution of gene trees given a species tree, considering all possible species trees under the coalescent model (17). Integrating across all possible species trees would be very slow; therefore, we restrict our attention to the species tree topology specified by the gene tree branch lengths for each random vector of gene trees, as follows. The specified species tree topology is an ultrametric tree whose nodes are as deep as possible while still being no deeper than the corresponding nodes of all gene trees in the sampled vector, a constraint consistent with the idea that gene divergences always occur before species divergence of the same taxa (25, 26). This constrained species tree topology is used to generate an approximate prior distribution on gene trees $K(\mathbf{G})$, varying only branch lengths of the species tree. The likelihood of the sequence data given the gene trees and substitution parameters μ, $f(\mathrm{D}|\mathrm{G},\mu)$, is given by the substitution model, which in our case is the general time-reversible model (GTR). Because we have not searched the entire space of possible species trees, the gene tree distribution from this first step is best considered an approximate posterior distribution of gene trees, $K(\mathbf{G}|\mathrm{D})$, under the coalescent model. The approximate posterior distribution $K(\mathbf{G}|\mathrm{D})$ is estimated by Markov Chain Monte Carlo (MCMC) using prior $K(\mathbf{G})$ and likelihood $f(\mathrm{D}|\mathrm{G},\mu)$ (*SI Materials and Methods*, Eq. **3**); this is achieved by incorporating this prior into the popular Bayesian phylogenetic analysis program, MrBayes (27).

With $K(\mathbf{G}|\mathrm{D})$ in hand, in step 2 we estimate $f(\mathrm{S},\theta|D)$, the posterior distribution of the species tree topology and branch lengths (S) and ancestral population sizes (θ) given the sequence data. For each gene tree vector $\mathbf{G_i}$ ($\mathrm{i} = 1$ to N samples, in our case ≈8,000) from step 1, we conduct a MCMC procedure that yields a sample of size x from $f(\mathrm{S},\theta|\ \mathbf{G_i})$, the posterior distribution for the species tree for gene tree vector $\mathbf{G_i}$. For each vector, we use a birth-death prior on the species tree and $f(\mathbf{G_i}|\mathrm{S})$, the probabilities of gene tree vectors given the species tree (SI Fig. 5), which again come from the coalescent model of Rannala and Yang (17). By combining the x samples from $f(\mathrm{S},\theta|\mathbf{G_i})$ associated with each vector, we estimate $f(\mathrm{S},\theta|D)$ with the resulting Nx sampled values. Step 3 is an importance sampling step to correct for the fact we used an approximate prior for the species tree in step 1. We correct for this by attaching weight $f(\mathbf{G_i})/K(\mathbf{G_i})$ to each species tree generated in step 2. This effectively aligns the ith sample from the approximate posterior distribution with what would have occurred if the true prior $f(\mathbf{G})$ had been used. The entire approach assumes free recombination between loci and lack of recombination within loci; it has been known for some time (28), and recent theory confirms for gene tree topologies (29), that even small amounts of recombination between loci will render the histories of linked gene trees nearly independent of one another, conditional on the species tree.

Gene Trees. Throughout our analysis, we used a gene-specific relative mutation parameter μ_i that allows for among-locus rate variation, which is essential to avoid having gene tree branch length variation attributed solely to coalescent effects and to prevent overestimation of ancestral θ (30). Because standard interpretation of gene trees in a coalescent framework requires a molecular clock, the first Bayesian hierarchical model we used (see *SI Materials and Methods*) estimated posterior probabilities of gene trees under this constraint for a recent data set consisting of 106 protein-coding regions sequenced from eight species of yeast (12). However, because estimating gene trees under a molecular clock can lead to errors, particularly with highly diverged sequences such as in the yeast data set, we also incorporated a simple correction whereby gene trees were first estimated without the constraints of a clock but in the posterior of step 1 were converted to ultrametric gene trees with the same total length as the unconstrained tree (see *SI Materials and Methods*). Although most coalescent programs tacitly assume a molecular clock on gene trees, we found that estimating gene trees without a clock was crucial to accurate estimation of the species tree.

The original study obtained 23 different topologies for the 106 genes when analyzed by parsimony or maximum likelihood (12). We found that consideration of 24 distinct topologies was sufficient to explain on average >95% and in most cases, ≈99% of the posterior distribution of gene trees for all analyses (SI Tables 1–4). The posterior distribution of gene trees under the independent model with a molecular clock was populated almost exclusively by 13 distinct topologies across all 106 genes, although most of this distribution was concentrated on six topologies (Fig. 1). In this analysis, the highest-probability tree for only 27 of 106 gene trees matched the concatenated tree published by Rokas *et al.* (ref. 12; topology 1, Fig. 1), whereas 38 genes yielded a maximum posterior probability gene tree in which *Saccharomyces kudriavzevii* and *Saccharomyces bayanus* form a clade (topology 2, Fig. 1). As expected, the posterior distribution of gene trees was noticeably more concentrated on a few (eight) trees under a joint model with a clock. However, we found that only 10 of the 106 genes in the data set were consistent with a molecular clock by a likelihood ratio test, suggesting that many of the gene trees estimated under a clock could be erroneous. We found that, without the constraints of a clock on gene trees, the effect of the joint model in concentrating the probability distribution of gene trees around a few topologies is even more evident (Figs. 1 and 2). Under these conditions, only three gene trees are plausible, many fewer than implied by the independent analyses used in studies of intergene phylogenetic signal (12). Moreover, under a joint model, the highest probability tree for 89 of the 106 genes matched the concatenated tree of the 106 genes made by Rokas *et al.* (12), and the next most common alternative was favored by only eight genes (Figs. 1 and 2). Comparison of the fit of the various models used using Bayes factors (3, 31) shows a clear superiority of the joint model under a relaxed clock compared with the three other models (see SI Fig. 6).

Species Tree. For the main analyses, we used a gamma (0, 120) prior on θ at each node to estimate the posterior distribution of the species tree topology and branch lengths for the yeast data set. We also tried gamma priors of (1, 10) and (1, 1,000) and found that these priors had little effect on the posterior distri-

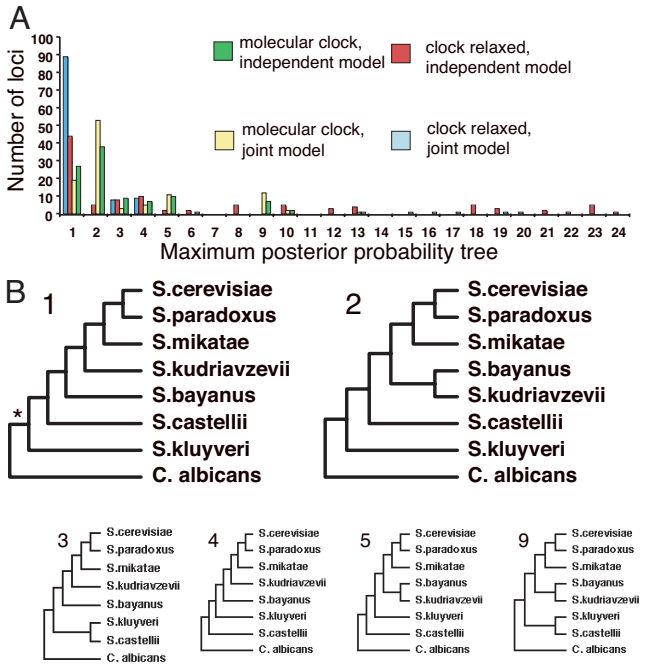

Fig. 1. The distribution of gene trees for the 106-gene yeast data set. (*A*) The number of genes (*y* axis) yielding each of 24 topologies according to the maximum posterior probability criterion (*x* axis) is shown for each of four analyses: independent (green) and joint (yellow) model with a molecular clock and independent (red) and joint (blue) model without a molecular clock. (*B*) The two most commonly encountered maximum posterior probability trees (for both species and genes) are shown below, with the next four most common shown in the bottom row (trees 3–5 and 9). The asterisk in tree 1 indicates the branch whose length differed drastically between BEST and MCMCcoal (30). The complete posterior distribution of gene trees for all four analyses is given in SI Tables 1–4.

bution of the species tree (see *SI Materials and Methods*). Here we focus primarily on the species tree topology, the primary focus of most phylogenetic analyses. The distribution of the prior species tree from which gene tree distributions were generated in step 1 contained only a single topology, topology 1 in Fig. 1. This result suggests there is considerable and consistent information in the gene tree vectors for this prior tree. Additionally, before importance sampling, the posterior distribution of the species tree also contained a single topology that was identical to the topology used in the prior, despite proposing eighty million vectors of gene trees and despite the fact that three topologies dominated the distribution of gene trees under the joint model without a clock (Fig. 2). So, from the point of view of the topology alone, the proposal scheme was very efficient and converged on a posterior modal species tree matching the concatenated gene tree of Rokas *et al.* (12) with probability 1.0 at each node. However, applying importance sampling to the posterior distribution the species tree topology, branch lengths, and ancestral population sizes culled nearly 98% of the distribution, leaving only ≈2% after importance sampling weights were applied. This result is likely a consequence of insufficiently varying the branch lengths in the species tree prior of step 1 and because the large size of the data set strongly skewed the distribution of weights among MCMC samples. Indeed, a challenging aspect of the importance sampling approach used here is that if the species trees sampled in step 1 comprise only a small fraction of the posterior distribution, the importance weights will tend to be very skewed and may misestimate the posterior density of parameters in these regions (32). Even so, the posterior distributions of species tree branch lengths and θ values are very similar before and after importance sampling, indicating that the importance sampling, although severe, did little to change these distributions. We have found that increased flexibility in searching the space of branch lengths, ancestral population sizes, and topologies in the species tree prior of step 1 increases the efficiency of the importance sampling substantially, a result that will be particularly significant when dealing with data sets containing large numbers of species. Other data sets yield more reasonable efficiencies of ≈20% (33). It is also significant that *S. kudriavzevii* and *S. bayanus* form a clade (topology 2, Fig. 2; see SI Fig. 6 and *SI Materials and Methods*) with high probability in the consensus species tree from the posterior distribution when gene trees were estimated under a molecular clock or when estimated under the independent prior without a clock. Particularly with highly diverged sequences as in the yeast data set, relaxing the molecular clock on gene trees is critical during species tree estimation.

By sampling increasing numbers of loci from the yeast data set at random, we found that the "correct" species tree could be estimated with high (>0.95) confidence with as few as eight genes, whereas the concatenation approach worked well only with at least 20 genes as reported by Rokas *et al.* (ref. 12; see *SI Materials and Methods*). We also compared our branch length and effective population size estimates (SI Table 5)with those delivered by a Bayesian approach that assumes rather than estimates a particular species tree (30). Although the parameter estimates differ substantially between the two approaches, in particular with regard to the length of what is broadly considered a long branch leading to *Candida albicans* (Fig. 1, tree 1), for a variety of reasons, we favor our estimates and suggest that the ability to estimate gene trees without a clock and to use complex substitution models with MrBayes) are positive features of our method (see SI Fig. 7).

Simulations. To explore the efficiency of the species tree approach, we simulated single coalescent gene lineages sampled from each tip of four- and eight-taxon species trees. We varied species branch lengths and effective population sizes at each node to produce gene trees with high and low probabilities of matching the species trees from which they were sampled. Under scenarios in which the proportion of gene trees matching the species tree was very high, we found that the correct species tree could be recovered with high probability with fewer than three genes (Fig. 3*A*). However, for species trees in which the probability of gene trees matching the species tree was low, we found that as many as 120 genes was required to accurately estimate the species tree in the case of eight species (Fig. 3 *B* and *C*). Remarkably, the BEST method was able to correctly reconstruct the species tree with high probability even when the proportion of gene trees matching the species tree was less than 10% (Fig. 3*C*). For a given sampling effort (e.g., 10 genes), the chance of correctly reconstructing the species tree increased as the proportion of gene trees matching the species tree increased (*SI Materials and Methods*). Overall, when the proportion of gene trees matching the species tree is low, the analysis suggests that increasing the number of independently segregating loci is crucial to achieving high confidence in the species tree. We also asked whether there existed situations in which Bayesian estimation of the species tree and concatenation of gene sequences would yield significantly different results and found that application of the joint model could estimate the correct species tree with high confidence in situations where concatenation under the appropriate model of nucleotide substitution using either Bayesian or maximum-likelihood methods yielded the wrong tree with high confidence (*SI Materials and Methods*).

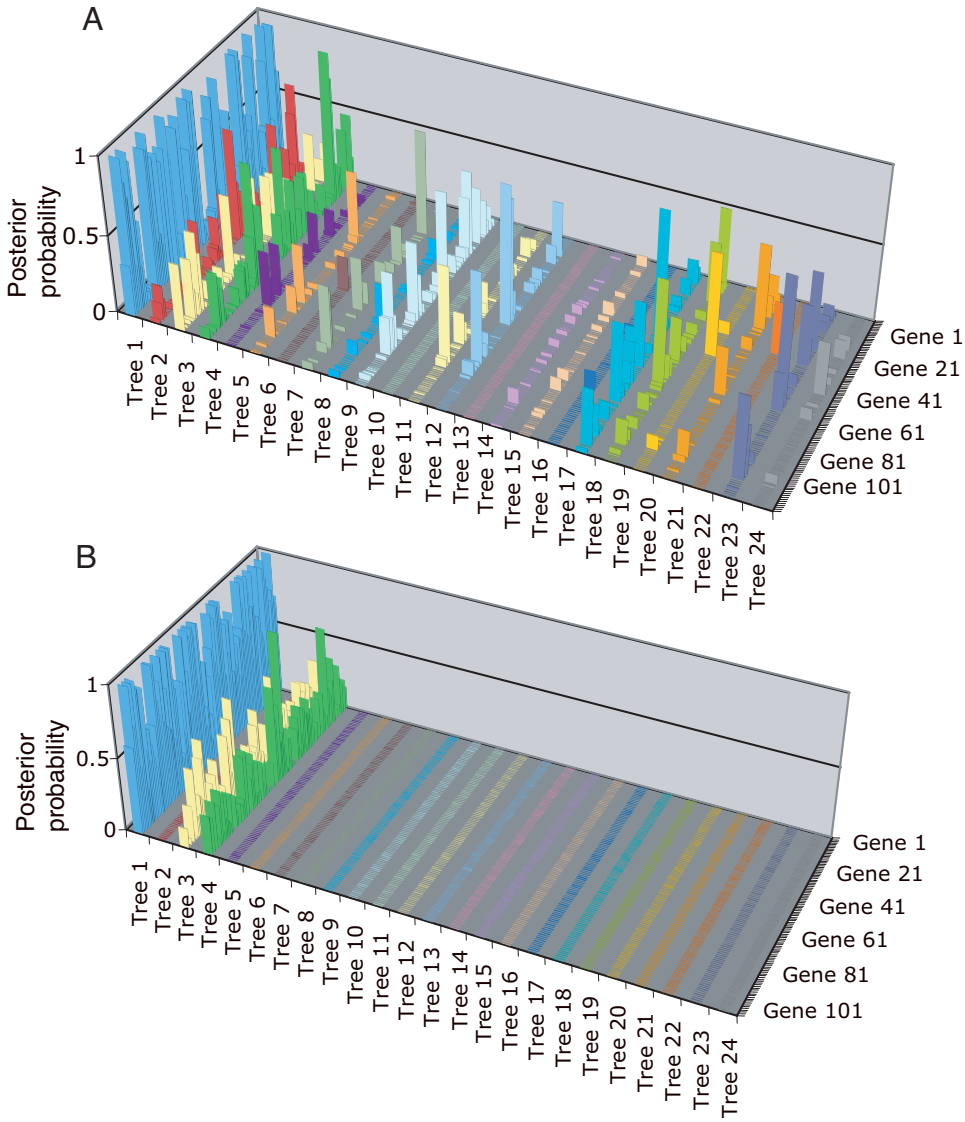

Fig. 2. Shifting phylogenetic landscapes for gene trees under different models. The complete posterior probability distributions for the independent (*A*) and joint (*B*) models without a molecular clock are shown.

Discussion

We have applied a multilocus Bayesian species tree method (BEST; ref. 33) to a prominent yeast phylogenomics data set and have shown that it is more efficient in estimating the species tree than concatenation is in estimating the gene tree. Increased taxon sampling has recently been shown to increase the congruence and accuracy of gene trees in the yeast data set (8), sometimes permitting accurate estimation with only two or three yeast genes. Yet the accuracy and efficiency of these two approaches cannot be compared directly, because they estimate different parameters; whereas concatenation estimates an average gene tree, BEST estimates the species tree containing the gene trees, as well as the individual gene trees themselves. The effect of increased taxon sampling on species tree estimation has not yet been explored, and presumably both taxon sampling and a focus on species trees rather than gene trees will have positive effects on phylogenetic analysis. Finally, we have shown through simulation how analysis focusing on the species tree can circumvent phylogenetic inconsistency in a situation where concatenation can positively mislead phylogeny estimation (14, 15).

Our analysis illustrates an important application of the concept of a joint prior in Bayesian phylogenetic analysis. This prior, whose distribution is specified by the approximate species tree in step 1 of the BEST method, is based on a logical assumption: that, barring evolutionary forces in addition to mutation and drift, such as gene flow or horizontal gene transfer, gene trees from independently segregating loci should be similar to one another. Such a prior can result in posterior gene tree distributions that are considerably more concentrated around a few topologies than when loci are analyzed independently of each other or of an overarching species tree (Figs. 1 and 2). Of course, this result also suggests that the signal in the yeast data set for many of the genes may be weak enough to be influenced by the prior. Our results suggest that other priors that incorporate the idea of a correlation among gene trees might also benefit Bayesian phylogenetic analysis. They also suggest there is much less need to explain incongruence among gene trees in the yeast data set (12) "or to increase multilocus congruence" by manipulation of substitution parameters, than recent critiques of the yeast analysis would imply (35, 36). Incongruence among gene trees has been viewed as a problem for MCMC analyses (37), but

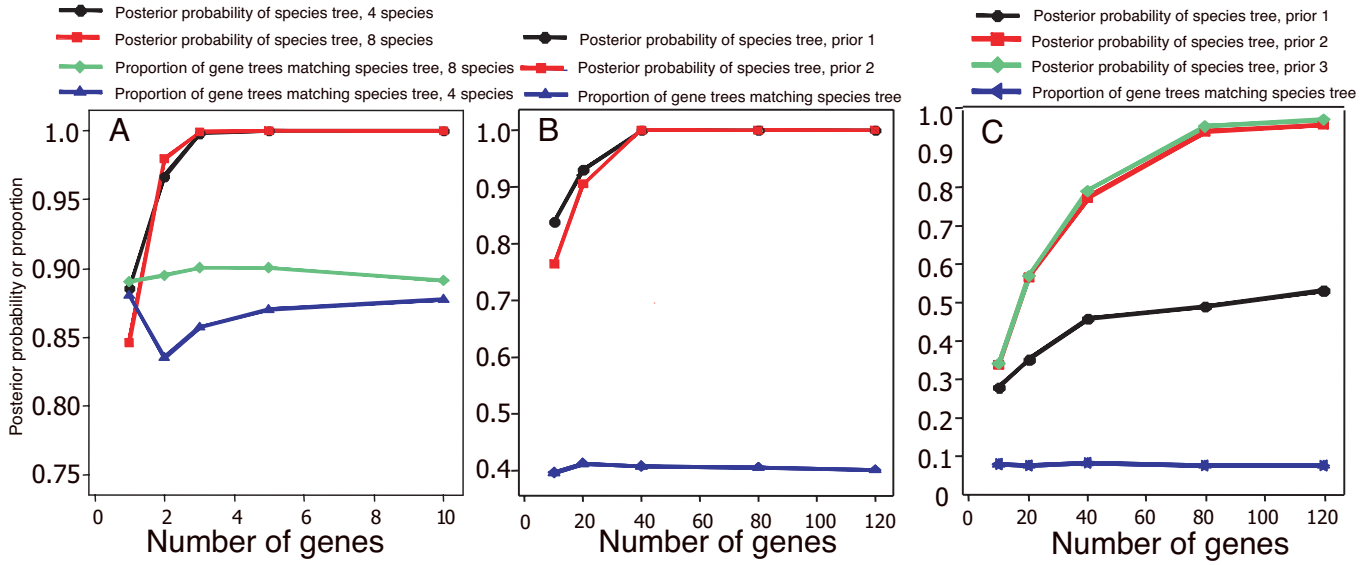

Fig. 3. Robustness and efficiency of the joint model for estimating species trees. (*A*) The number of genes required to resolve the correct species tree with four and eight species when the proportion of gene trees matching the species tree is high. Here this proportion varies between ≈83% and 90% (in blue and green, 100 gene trees per simulation) because the critical internodes in the species tree are relatively long on the scale of the effective population size (θ). The gamma-distributed prior on θ for each node was (1, 200), indicating a mean θ of 1/200 and variance of 1/40,000. A prior mean of 1/200 is consistent with what we know about θ in natural populations of yeast (45, 46). (*B*) The number of genes required to resolve the correct four-species tree when the proportion of gene trees matching the species tree (in blue) is low (≈40%). Prior 1 on θ is (1, 200), and prior 2 is (1, 1,000). (*C*) The number of genes required to resolve the correct eight-species tree when the proportion of gene trees matching the species tree (in blue) is low (<10%). Prior 1 on θ is (1, 100), prior 2 is (1, 500), and prior 3 is (1, 1,000).

the coalescent provides a mechanism in which multiple gene trees can be reconciled in a single species history. Importantly, unlike other methods for estimating species trees, the Bayesian method we have outlined incorporates error and uncertainty in gene trees through their posterior probability distributions.

Most of the discussions surrounding Bayesian posterior probability values, their fundamental difference from bootstrap proportions and their overcredibility when the evolutionary model, likelihood function, or prior distributions are inappropriate (38–41), have all been based on phylogenetic analysis of gene trees. Thus it remains to be explored whether posterior probabilities on species trees are subject to error to the same extent. Our analysis suggests that increasing the number of loci will increase the efficiency of estimating species phylogenies (Fig. 3), as it does for estimating simple population parameters, such as θ for a single population (42). However, the precise allocation of effort to number of loci, and the number and length of sampled sequences per species remain to be explored.

Inspection of the posterior distributions of parameters of the yeast species tree suggests that topologies, branch lengths, and ancestral population sizes will be estimated with decreasing certainty in that order (see *SI Materials and Methods*). Although we suspect the method will be most accurate in cases of recent speciation, our analysis of the yeast data set indicates that estimation of species trees on the time scale of tens of millions of years are possible. Clearly, initial estimation of gene trees without the constraints of a molecular clock is critical, especially with highly diverged sequences such as in the yeast data. Our simulations also illustrate how the BEST approach is applicable to instances in which gene tree/species tree concordance is high or complete, and that it need not be confined to data sets with high gene tree heterogeneity. Incorporation of multiple sequences sampled per species, as well as hybridization, gene flow, and lateral gene transfer into a more general model of phylogenetic inference, is an important goal for future work, one that could obviate the need to discard such genes from analysis simply because their topologies disagree with those of the majority (43). Species tree methods offer a more inclusive approach to systematics that directly bridges phylogenetics and population genetics and that could help clarify some of the most famous adaptive radiations, such as those for Darwin's finches and African cichlid fish, in which lack of reciprocal monophyly and gene tree discordances are the norm (6, 32, 44).

Materials and Methods

Theory, Data Sets, and BEST Analysis. We describe in full the Bayesian hierarchical model used by BEST elsewhere (32, 33) and in *SI Materials and Methods*. The programs used in this paper are available at www.stat.osu.edu/~dkp/BEST. A. Rokas kindly supplied a concatenated nexus file of the 106-gene yeast data set with character partitions for each gene. A modified MrBayes was run for 80 million cycles per analysis on computers in the Ohio Supercomputer Center to estimate posterior distributions. A GTR + Γ + I model was used for all analyses. The analyses using the joint prior required a prior on $\theta = 4N\mu$. Although the same prior distribution for θ is used for all nodes, both the proposed values for each cycle in the MCMC and the posterior distributions for θ are different for different nodes. For the analyses in Figs. 1 and 2, we used a gamma distribution prior for θ of (1, 200), i.e., with mean of 1/200 = 0.005 and variance $1/(200)^2$ = 0.000025, as well as priors of (1, 10) or (1, 1,000). Priors for θ in Fig. 3 are given in the legend and are based in part on known parameters for natural yeast populations (45, 46). In general, these various priors had moderate effect on the estimated posterior distributions of gene trees, but very little effect on the estimated posterior distribution of species trees. On the other hand, they did have a strong effect on the estimated ancestral population sizes (see *SI Materials and Methods*).

Bayes Factor Analysis. We used the harmonic means of the likelihoods to compute Bayes factors between the four models (independent and joint models, with and without a molecular

clock on gene trees) and the concatenation model. Bayes factors were computed after applying the importance sampling weights. The likelihood profiles for the four models are shown in *SI Materials and Methods*. The log Bayes factor favoring the joint model with relaxed clock over the next best model (independent model with a relaxed clock) was 130, indicating strong support. The independent model was actually favored over the joint model when a molecular clock was enforced. The concatenation model (with a relaxed clock) was the worst model (Bayes factor, 5,238). These conclusions remain true even with adjustments for the increased number of parameters in the more flexible models.

Analysis of BEST Efficiency. To determine the efficiency of the BEST method, we first randomly chose eight yeast genes and used the gene trees from their posterior distributions to build species trees as described above. We then repeated this 10 times to see how many samples of 10 can recover the species tree. For each of these 10 replicates, we found that every estimated species tree was the same as topology 1 with an average posterior probability of each node >0.95. We then repeated this for five genes instead of eight. We found that in one of the 10 estimated species trees, the topology was different from topology 1, although the other 9 correctly estimated topology 1 with high support >0.95. We therefore conservatively estimate that eight genes in the yeast data set are sufficient to estimate the correct species tree with high probability.

Simulations. We used the program MCMCcoal (30) to generate simulated gene genealogies on four- and eight-taxon species trees. In analysis 1, the proportion of gene trees matching the species tree was low (≈10–40%), and in analysis 2 the proportion was high (≈85–90%; see *SI Materials and Methods*). Gene trees and branch lengths generated using MCMCcoal were analyzed directly in BEST without simulation of DNA sequences to focus specifically on the ability of BEST to reconstruct species trees given gene tree heterogeneity but in the absence of gene tree error. To compare the consistency of concatenation vs. BEST analysis, we used MCMCcoal to simulate 30 gene trees on an 8-taxon species tree (species A–H) with branch lengths (see *SI Materials and Methods*). We then simulated 500 bp of DNA sequence on each of these gene trees using the Jukes–Cantor (JC) model of nucleotide substitution and built gene trees by concatenation using MrBayes and the JC model as well as by the joint model used in BEST. The BEST analysis of the data resulted in essentially a single species tree in the posterior distribution with the correct topology [(H, (G, (F, (E, (D, (C, (A, B)))))))] and receiving >98% of the posterior probability, despite the fact that 36 different gene trees had substantial posterior probability in the analysis. By contrast, analysis of the same DNA sequences by concatenation under the appropriate model of nucleotide substitution using either Bayesian or maximum-likelihood methods yielded the wrong tree with high confidence [(H,(G,((F,E),(D,(C,(B,A)))))))]. This convergence to the wrong tree under concatenation was seen in data sets of only 10 loci each of 500 bp and was virtually guaranteed provided that at least one internal branch in the species was short. The inconsistency of concatenation increased with increasing numbers of loci, whereas the BEST analysis continued to converge on the correct tree.

We thank A. Rokas (Broad Institute, Cambridge, MA), T. Collins (Florida International University, Miami, FL), and J. Gatesy (University of California, Riverside, CA) for providing data and discussing data analysis. L. Kubatko, N. Rosenberg, J. Degnan, E. Louis, G. Liti, and J. Fay provided helpful discussion and shared manuscripts throughout this project. We thank two anonymous reviewers and the Editor for numerous helpful comments. C. Moritz, J. McGuire, C. Moreau, C. Davis, J. Degnan, L. Kubatko, S. Smith, A. Rokas, and the Systematics Discussion Group in Organismic and Evolutionary Biology at Harvard University provided helpful comments on the manuscript. This work was supported by National Science Foundation Grants DEB-0315806 (to S.V.E.) and DMS-0112050 (to D.K.P.).

1. Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) *Syst Biol* 53:47–67.
2. Felsenstein J (2003) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
3. Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002) *Syst Biol* 51:673–688.
4. Cracraft J, Donoghue MJ, eds. (2004), in *Assembling The Tree of Life* (Oxford Univ Press, New York), pp 468–489.
5. Maddison WP (1997) *Syst Biol* 46:523–536.
6. Avise, JC (1994) *Molecular Markers, Natural History and Evolution* (Chapman and Hall, New York).
7. Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) *Syst Biol* 51:664–671.
8. Hedtke SM, Townsend TM, Hillis DM (2006) *Syst Biol* 55:522–529.
9. Driskell AC, Ane C, Burleigh JG, McMahon, M. M., O'Meara B, C. & Sanderson MJ (2004) *Science* 306:1172–1174.
10. Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) *Nature* 439:965–968.
11. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) *Science* 311:1283–1287.
12. Rokas A, Williams BL, King N, Carroll SB (2003) *Nature* 425:798–804.
13. Rokas A, Kruger D, Carroll SB (2005) *Science* 310:1933–1938.
14. Degnan JH, Rosenberg NA (2006) *PLoS Genet* 2:762–768.
15. Kubatko LS, Degnan JH (2007) *Syst Biol* 56:1–9.
16. Beerli P, Felsenstein J (2001) *Proc Natl Acad Sci USA* 98:4563–4568.
17. Rannala B, Yang Z (2003) *Genetics* 164:1645–1656.
18. Hey J, Nielsen R (2004) *Genetics* 167:747–760.
19. Slowinski J, Page RDM (1999) *Syst Biol* 48:814–825.
20. Takahata N (1989) *Genetics* 122:957–966.
21. Bininda-Emonds OR (2005) *Methods Enzymol* 395:745–757.
22. Wilson IJ, Weale ME, Balding DJ (2003) *J R Stat Soc A* 166:155–188.
23. Nielsen R (1998) *Theor Popul Biol* 53:143–151.
24. Maddison WP, Knowles LL (2006) *Syst Biol* 55:21–30.
25. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).
26. Edwards SV, Beerli P (2000) *Evolution (Lawrence, Kans)* 54:1839–1854.
27. Huelsenbeck JP, Ronquist F (2001) *Bioinformatics* 17:754–755.
28. Robertson A, Hill WG (1983) *Proc R Soc London Ser B* 219:253–264.
29. Slatkin M, Pollack JL (2006) *Genetics* 172:1979–1984.
30. Yang Z (2002) *Genetics* 162:1811–1823.
31. Kass RE, Raftery AE (1995) *J Am Stat Assoc* 90:773–795.
32. Liu L, Pearl DK (2006) in *Mathematical Biosciences Institute Technical Report #53* (Ohio State University, Columbus, OH), p 24.
33. Liu L, Pearl DK (2006) *Syst Biol*, in press.
34. Stephens M, Donnelly P (2000) *J R Stat Soc B* 62:605–635.
35. Gatesy J, Baker RH (2005) *Syst Biol* 54:483–492.
36. Collins TM, Fedrigo O, Naylor GJ (2005) *Syst Biol* 54:493–500.
37. Mossel E, Vigoda E (2005) *Science* 309:2207–2209.
38. Yang Z, Rannala B (2005) *Syst Biol* 54:455–470.
39. Lewis PO, Holder MT, Holsinger KE (2005) *Syst Biol* 54:241–253.
40. Misawa K, Nei M (2003) *J Mol Evol* 57 Suppl 1:S290–S296.
41. Suzuki Y, Glazko GV, Nei M (2002) *Proc Natl Acad Sci USA* 99:16138–16143.
42. Felsenstein J (2006) *Mol Biol Evol* 23:691–700.
43. Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) *Mol Biol Evol* 24:412–426.
44. Avise JC (2000) *Phylogeography: The History and Formation of Species* (Harvard Univ Press, Cambridge, MA).
45. Liti G, Barton DB, Louis EJ (2006) *Genetics* 174:839–850.
46. Fay JC, Benavides JA (2005) *Genetics* 170:1575–1587.