Published in *Molecular Biology and Evolution* (2007) 24(2). Copyright 2007, Oxford University Press. Used by permission. DOI:10.1093/molbev/msl170.

Bayesian Estimation of Concordance among Gene Trees

Cécile Ané[1,*], Bret Larget[1], David A. Baum[1], Stacey D. Smith[2,ǂ] and Antonis Rokas[3]

[1]University of Wisconsin; [2]Duke University; [3]The Broad Institute of MIT and Harvard
[*]ane@wisc.edu; [ǂ]affiliation in 2012: University of Nebraska-Lincoln, ssmith19@unl.edu

Keywords: Gene genealogy; Concordance; Bayesian phylogenetics; Total evidence; Consensus methods; Dirichlet process.

Manuscript accepted November 3, 2006; originally available online November 9, 2006.

Multigene sequence data have great potential for elucidating important and interesting evolutionary processes, but statistical methods for extracting information from such data remain limited. Although various biological processes may cause different genes to have different genealogical histories (and hence different tree topologies), we also may expect that the number of distinct topologies among a set of genes is relatively small compared with the number of possible topologies. Therefore evidence about the tree topology for one gene should influence our inferences of the tree topology on a different gene, but to what extent? In this paper, we present a new approach for modeling and estimating concordance among a set of gene trees given aligned molecular sequence data. Our approach introduces a one-parameter probability distribution to describe the prior distribution of concordance among gene trees. We describe a novel two-stage Markov chain Monte Carlo (MCMC) method that first obtains independent Bayesian posterior probability distributions for individual genes using standard methods. These posterior distributions are then used as input for a second MCMC procedure that estimates a posterior distribution of gene-to-tree maps (GTMs). The posterior distribution of GTMs can then be summarized to provide revised posterior probability distributions for each gene (taking account of concordance) and to allow estimation of the proportion of the sampled genes for which any given clade is true (the sample-wide concordance factor). Further, under the assumption that the sampled genes are drawn randomly from a genome of known size, we show how one can obtain an estimate, with credibility intervals, on the proportion of the entire genome for which a clade is true (the genome-wide concordance factor). We demonstrate the method on a set of 106 genes from 8 yeast species.

Introduction

The growth of multigene and even genome-wide data sets for phylogenetic analysis has simultaneously presented systematists with greater power to estimate evolutionary history and enormous challenges for extracting, analyzing, and summarizing phylogenetic signal. Exploration of these large data sets has made it abundantly clear that different genes sampled from the same set of taxa can produce markedly different phylogenies (Kellogg et al. 1996; Baker and DeSalle 1997; Baldauf et al. 2000; Giribet et al. 2001; Cronn et al. 2003; Pollard et al. 2006). Although some of these differences may be due to incorrect estimation of gene genealogies, incongruent gene trees can also be due to the existence of different evolutionary histories for different parts of the genome. Ideally, systematists would like to make inferences about the diversity of gene genealogies within the genome and then find ways to summarize their findings. Here we develop a Bayesian approach for making such genome-wide inferences while properly taking into account uncertainty in the estimates of the phylogeny for each sampled gene.

Existing approaches to synthesizing phylogenetic signal from different loci have historically taken two general forms, total evidence and consensus methods. The former approach advocates analysis of concatenated data sets, despite the potential for different evolutionary histories (Kluge 1989; Barrett et al. 1991). The justification is that, if most of the genome has been transmitted vertically, it should share a common tree, equivalent with the “species” tree. Genes that have minority histories are taken as instances of phylogenetic noise, which can be swamped with sufficient sampling (Barrett et al. 1991; de Queiroz 1993). Although combining data from multiple genes can result in strongly supported phylogenetic estimates, the implicit assumption of a single divergent history may undermine interpretation of measures of support on the combined tree (e.g., Lewis et al. 2005; Mossel and Vigoda 2005) and precludes investigation of potentially interesting biological processes such as incomplete lineage sorting, hybridization, and lateral gene transfer that underlie discordant histories (Wendel and Doyle 1998).

In contrast to the total evidence approach, consensus methods typically entail identifying the best estimate of the phylogeny of each gene and creating a consensus of these separate point estimates. The recently developed consensus network approach, implemented in the program SplitsTree (Holland et al. 2004, 2006; Huson and Bryant 2006), extends previous consensus methods by allowing the user to pinpoint nodes with conflict and visualize the frequency of alternative resolutions. Although consensus methods can retain a diversity of potential topologies for a given gene, they lack an objective way to incorporate uncertainty in individual-gene tree estimates. Additionally, because each gene is analyzed independently, information on the trees supported by one gene do not influence the choice of gene trees for another gene. This seems undesirable because we often have a strong prior belief that genes from the same organisms are more likely than chance to share the same genealogical history (as pointed out by Penny et al. [1982]). Suchard et al. (2003) and Suchard (2005) proposed Bayesian methods for the simultaneous estimation of the species tree and multiple gene trees. These methods consider topology and branch length, making them computationally intractable for reasonable numbers of taxa. We here sought to develop an alternate method that considers tree topology only, which increases the potential for analyzing data from larger numbers of taxa.

The work described here was initially motivated by a particular challenge that could not be met by existing analytical methods: estimating the proportion of the genome for which a given clade is true, the clade's concordance factor (Baum 2007). It is generally agreed that one of the underlying objectives of phylogenetic research is to estimate the dominant tree for a set of sampled taxa. One way (of several) to understand the concept of a "dominant tree" is as a tree composed of those clades that are true for a plurality of the genome, that is, clades whose concordance factors exceed that of any contradictory clades. Such a primary concordance tree is something that a systematist might wish to estimate and use as a basis for taxonomy (Baum 2007). However, existing methods are not designed to estimate concordance factors, and thus, even when we have sequence data for multiple genes from the same organisms/taxa, we cannot directly estimate concordance factors or primary concordance trees. We, therefore, hoped to develop a method that could estimate concordance factors while taking account of uncertainty in the gene trees for the sampled genes and the fact that only a finite number of genes have been sampled from the genome.

The estimation of concordance factors and concordance trees could be achieved if we had a statistically valid method to estimate the complete distribution of evolutionary histories within a multigene data set. From such a distribution, these and many other evolutionary parameters of interest could be extracted. In this paper, we describe such a method. In contrast to the total evidence approach, we did not want to assume *a priori* that all genes share a common evolutionary history. Likewise, we hoped to improve upon existing consensus methods by properly accounting for uncertainty in individual-gene tree estimates and by allowing genealogical information from one gene to influence our estimates of another gene's genealogy.

We used a Bayesian approach because this provides a formal framework for combining prior beliefs about genealogical concordance with evidence contained in aligned sequence data from individual genes. We employ a novel two-stage Bayesian Markov chain Monte Carlo (MCMC) approach where we first calculate the posterior distribution of trees from single-gene analyses and then use these results along with a prior distribution on gene tree concordance for a second-stage MCMC to estimate the joint probability distribution of the gene-to-tree map (GTM) (described more formally below). Further, we develop methods for estimating the proportion of the genome for which any given clade is true (the genome-wide concordance factor) based on the sequence data for a set of genes randomly sampled from the genome. The posterior distribution of GTMs has many other potential uses besides the estimation of concordance factors. For example, it becomes possible to infer the true tree for a given gene conditional on genealogical information from other genes, to identify particular genes with outlier gene tree topologies, or to estimate the proportion of the genome that was transferred during an introgression event.We demonstrate Bayesian concordance analysis using an 8-taxon, 106-gene data set (Rokas et al. 2003) so as to demonstrate the method's potential for the statistical analysis of genealogical concordance at a genome-wide scale.

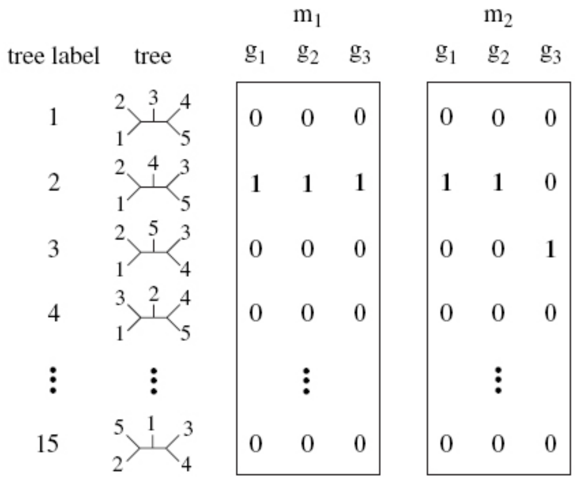
Methods

GTMs

Consider a collection of aligned molecular sequences with one sequence from each of the several loci for a matching set of individuals. Provided that the alignments are correct, every site at every locus has a single history. We will assume, in addition, that all sites within a single locus share the same evolutionary history. In the remainder of this paper, we will refer to each locus with assumed common history as a "gene." In the case of a coding sequence consisting of several exons separated by large introns where we might suspect the possibility of recombination so that separate exons might have different histories, we could treat each exon as a different gene.

Under the assumption that each gene is associated with a single tree, there exists a true mapping from genes to trees. This map, which we will call a "gene-to-tree map" or GTM, can be represented in multiple ways. One way to represent the map is with a table where columns represent genes and rows represent trees. Each column contains a single one in the row corresponding to its true tree, and all other entries in the column are zeros. Figure 1 shows two GTMs. The left one is a case of complete concordance because all genes share the same tree, whereas the right GTM shows some discordance among genes. For multigene data sets, a GTM will be a very large, sparse matrix. For instance, if there are 8 taxa and 100 genes, there are 10,395 possible tree topologies, and the GTM would be a 10,395 by 100 matrix filled with zeros except for a single one in each column.

It may be helpful to note that a similar matrix representation can be used to visualize a GTM and to summarize individual-gene tree posterior probabilities. In the example from Figure 1, we could imagine filling each column of the matrix with the posterior probability of each tree topology as estimated from an MCMC sample in a Bayesian analysis.The matrix would be similarly constrained with each column summing to 1, but, although small, in theory, there will be positive values throughout the column.



		m1			m2		
tree label	tree	g_1	g_2	g_3	g_1	g_2	g_3
1	2 3 4 1 5	0	0	0	0	0	0
2	2 4 3 1 5	1	1	1	1	1	0
3	2 5 3 1 4	0	0	0	0	0	1
4	3 2 4 1 5	0	0	0	0	0	0
⋮	⋮		⋮			⋮	
15	5 1 3 2 4	0	0	0	0	0	0

Figure 1—Examples of GTMs. Left box (m_1): GTM showing complete concordance among gene trees because all genes are mapped to tree 2.Right box (m_2): GTM showing some discordance. The first and second genes share the same tree (tree 2), but the third gene has a different tree (tree 3).

A second and more concise representation of a GTM is as a vector of tree labels whose length is the number of genes where the *i*th element is the label corresponding to the tree mapped to by the *i*th gene. In Figure 1, the first GTM is represented as $m_1 = (2, 2, 2)$ and the second as $m_2 = (2, 2, 3)$. With this representation, it is easy to see that if there are G genes and T trees, then there are a total of T^G possible GTMs. In the small example from Figure 1 with five taxa and three genes, there are 15 possible tree topologies and $15^3 = 3{,}375$ possible GTMs. The GTM m^1 is one of the 15 in which all genes share the same tree topology, whereas m^2 shows some discordance.

Bayesian Estimation

Formally, the problem we address is the estimation of a GTM M showing the mapping between a set of G genes and T trees, each with the n leaves on the basis of data $X = (X_1,\ldots, X_G)$, where X_i is the aligned sequence data from the *i*th gene. In a Bayesian framework, we wish to compute the posterior probability $P\{M \mid X\}$ given a prior probability distribution $P\{M\}$ and a likelihood model $P\{X \mid M\}$. Under assumptions we will specify as we progress, we can construct the posterior distribution on M from the individual-gene posterior distributions and a prior distribution on the amount of concordance.

Model of Evolution

Consider first the likelihood model $P\{X \mid M\}$. Conditional on the GTM M, we assume that the G genes evolve independently, each according to its own evolutionary model. Thus, each gene may have its own branch lengths, base/amino acid/codon frequencies, transition/transversion ratio or synonymous/nonsynonymous ratio, shape parameter for the distribution of rates across sites, etc. (Liò and Goldman 1998; Whelan et al. 2001; Huelsenbeck 2002). We make the assumption that the parameters for gene *i* depend on the GTM M through the true tree for gene *i* only and are thus independent of the parameters for all other genes. Under these assumptions, we show in Appendix that the likelihood of GTMs is proportional to the product over the genes of the individual-gene posterior probabilities normalized by the prior probabilities of the topologies:

$$P\{X|M\} \propto \prod_{i=1}^{G} \frac{P\{T_i|X_i\}}{P\{T_i\}}, \tag{1}$$

where T_i is the tree assigned by M to gene *i* and the posterior probability $P\{T_i \mid X_i\}$ is obtained by analyzing gene *i* individually with the desired prior distribution on the gene's evolutionary parameters (including branch lengths). However, any distribution for the tree topology can be used in individual analyses and does not correspond to any assumption of our model. In fact, the likelihood of the GTM M on left-hand side of Equation (1) is defined independently of any prior distribution on GTMs, and so independently of any prior distribution on trees. In Equation (1), if we use a uniform prior distribution on topologies for each individual-gene analysis, the likelihood of the GTM M is proportional to the product of the corresponding individual-gene posterior probabilities:

$$P\{X|M\} \propto \prod_{i=1}^{G} P\{T_i|X_i\}. \tag{2}$$

Therefore, evolutionary parameters can be integrated out in individual-gene analyses, a uniform distribution on tree topologies serving as a tool, and the posterior probability distribution for the GTM is proportional to the product of its prior distribution and of its likelihood:

$$P\{M|X\} \propto P\{M\} \prod_{i=1}^{G} \mathrm{P}\{T_i|X_i\}. \tag{3}$$

Implications of Assumptions

The independence assumptions we make result in flexible prior distributions because they allow different genes to have different base frequencies or different branch lengths, even though they might share the same tree topology. The sequence data enters the likelihood of the GTM only through the individual-gene posterior distributions. This has the benefit that it suffices to analyze each gene separately and simply retain each individual posterior distribution for later use. The cost of this approach is that we do not benefit from the possibility of pooling information among genes about evolutionary parameters other than the tree topology. Note also that using uniform distributions on the tree topology in individual analyses is not necessary in general: if nonuniform priors were used for individual genes, the likelihood of a GTM would still be proportional to the ratio of the products of individual-gene posterior and prior distributions as in Equation (1), with no modification of our prior distribution on GTMs.

Prior Distribution on GTMs

A GTM determines which genes share the same tree and which genes do not. Thus, it determines clusters of genes, such that all genes in the same cluster are fully concordant and share the same tree. Our Bayesian framework requires a prior distribution for this clustering. The Dirichlet process prior probability distribution (Ferguson 1973; Antoniak 1974) is widely used as a prior distribution in Bayesian statistics for clustering problems. Our use of this distribution is distinct from previous uses in phylogenetics (Huelsenbeck et al. 2004; Liang L-J, Weiss RE, personal communication) because here the parameter defining the clusters, tree topology, is not a continuous variable. In our case, the value associated with each cluster is one of the T possible tree topologies. The Dirichlet process can be understood as a generalization of a Polyá urn scheme for assigning trees to genes by sequentially drawing topology-labeled cards from an urn, one for each gene, where the contents of the urn change as the sampling progresses.

We have two types of cards available: regular cards that have weight 1 and show a picture of one of the possible tree topologies, and a special "joker" with weight $\alpha > 0$. We have an endless supply of regular cards for each tree, but only one joker. In addition, we have a second well-shuffled infinitely large deck of regular tree cards where the proportion of cards with tree i is f_i (with these frequencies f_i summing to 1). We begin with the joker as the only card in the urn. We sequentially sample by choosing a card from the urn where the probability of drawing each card is proportional to its weight (one for each tree card and α for the joker). When the joker is drawn from the urn, we draw a tree card at random from the second deck and return it to the urn with the joker. If a regular card is drawn instead, we return it along with another identical card to the urn. In both cases, the next gene is mapped to the tree shown on the new regular card added to the urn. This continues until all genes are mapped to a tree, that is, until the urn contains as many tree cards as the number of genes being considered. After sampling c cards, the total weight of cards in the urn is α + c. If tree i is represented by r regular cards already in the urn, then it has probability $(r + \alpha f_i)/(c + \alpha)$ of being selected for the next gene: $r/(c + \alpha)$ is the chance that one of the r cards is chosen, and $\alpha/(c + \alpha)$ fi is the chance of picking the joker and then drawing tree i from the second deck. The Dirichlet process prior probability distribution for GTMs is determined by the parameter α and by the prior probabilities f_i that tree i is correct. All genes share this common prior distribution on tree topologies. The uniform distribution is the special case where $f_i = 1/T$. In this special case, the Dirichlet process has a single parameter, α.

Although the previous description of the urn sampling scheme depends on a specific ordering of the genes, it turns out that the probability of any GTM is independent of this ordering. For example, the GTM m_1 = (2, 2, 3, 2, 1) has probability

$$f_2 \frac{1 + \alpha f_2}{1 + \alpha} \; \frac{\alpha f_3}{2 + \alpha} \; \frac{2 + \alpha f_2}{3 + \alpha} \; \frac{\alpha f_1}{4 + \alpha},$$

whereas the probability of GTM m_2 = (1, 2, 2, 2, 3) with the same trees in a different order,

$$f_1 \frac{\alpha f_2}{1 + \alpha} \; \frac{1 + \alpha f_2}{2 + \alpha} \; \frac{2 + \alpha f_2}{3 + \alpha} \; \frac{\alpha f_3}{4 + \alpha},$$

is equal to the previous expression because the numerator is simply the product of the same factors in a different order and the denominator is identical.

To simplify the general expression of the probability of a GTM, we introduce the notation

$$A_n(x) = \prod_{i=0}^{n-1} (i + x), \tag{4}$$

so that $A_1(x) = x$, $A_2(x) = x(x + 1)$, and so on. For instance, the probability of GTM m_1 (and of m_2) is

$$\frac{A_1(\alpha f_1) A_3(\alpha f_2) A_1(\alpha f_3)}{A_5(\alpha)}.$$

More generally, if a GTM m maps the G genes to k distinct trees with $g(i) > 0$ genes mapped to tree $t(i)$, respectively, for $i = 1, \ldots, k$ so that $\sum_{i=1}^{k} g(i) = G$, the probability of m is

$$P\{m\} = \frac{\prod_{i=1}^{k} A_{g(i)}(\alpha f_{t(i)})}{A_G(\alpha)}. \tag{5}$$

In the special case of a uniform distribution on T trees, this is

$$P\{m\} = \frac{\prod_{i=1}^{k} A_{g(i)}(\alpha/T)}{A_G(\alpha)}. \tag{6}$$

The larger the value of α, the more probable it becomes *a priori* that there are a greater number of distinct gene trees. This should be clear intuitively from the urn sampling description of the process as α alone modulates the rate at which the joker is selected and new topologies can only arise when a joker is selected.

Continuum between Total Evidence and Consensus Approaches

The Dirichlet process prior distribution we describe here represents a compromise between two limiting cases.The limiting case of α = 0 corresponds to the total evidence approach that insists *a priori* that there is but a single tree for all genes. In the card/urn description, once the first card is added to the urn, the joker is never sampled again, and all genes share the same topology as the first gene. The other limiting case is α = ∞ which corresponds to a prior assumption of independence between gene trees, which is equivalent to a uniform distribution on all GTMs (when assuming a uniform distribution on the tree topology for each gene). In this case, the joker is selected every time, resulting in a draw from the second deck. As a result, knowledge of the true tree for one gene is uninformative about the true tree for any other gene, which is similar to the implicit assumption made in typical consensus approaches. In the latter case, with α = ∞, the posterior distribution on GTMs is proportional to the likelihood.

Prior Distribution on Number of Distinct Gene Trees

In the case of a uniform distribution on the T trees, the Dirichlet distribution is governed by α only. The expression for the prior probability distribution for the number k of distinct gene trees is

$$P\{k|\alpha, T, G\} = \frac{A_k(T - k + 1)}{A_G(\alpha)} \sum_{i=k}^{G} a(G, i) S(i, k) \left(\frac{\alpha}{T}\right)^i, \tag{7}$$

where α(G, i) is the absolute value of a Stirling number of the first kind and $S(i, k)$ is a Stirling number of the second kind. The Appendix contains details of the derivation of this expression and a description of how to compute Stirling numbers.

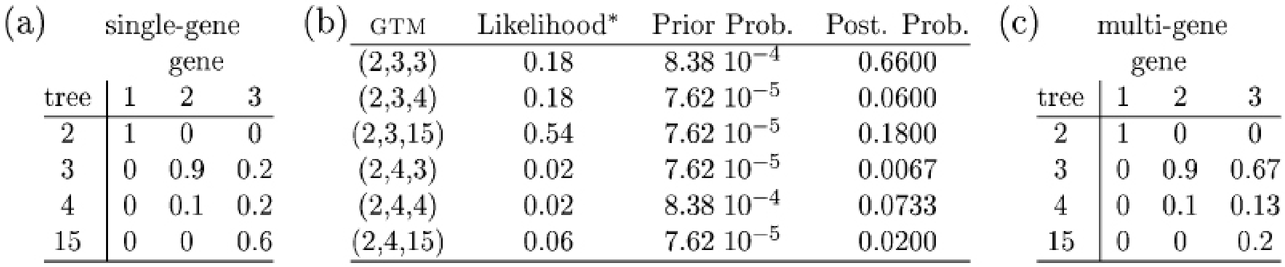
The probability that all gene trees are identical is found by substituting k = 1 in Equation (7) which simplifies to $TA_G(\alpha/T)/A_G(\alpha)$. This expression can be useful when considering an appropriate value for the prior parameter α. In addition, the choice of a can be guided by the prior probability that two randomly sampled genes share the same tree, which is $(1 + \alpha/T)/(1 + \alpha)$ ~ $1/(1 + \alpha)$ when the ratio α/T is negligible.

Numerical Example

The following small numerical example with three genes and five taxa will help to illustrate these ideas. We assume a uniform prior distribution on the T = 15 distinct tree topologies and set α = 1.5 (so that α/T = 0.1). This value of α corresponds to a prior probability of $(1 + a/T)/(1 + a) = 0.44$ that two randomly chosen genes share the same tree.

Equation (7) implies that the probabilities of there being one, two, and three distinct gene trees are 0.264, 0.528, and 0.208, respectively. Also, Equation (3) implies that the prior probability of each specific GTM with a single shared tree is the same as P{(1, 1, 1)} = 0.0176. Any specific GTM with two distinct gene trees such as (1, 2, 1) has prior probability 8.38 × 10^{-4}, and finally, each specific GTM with three distinct gene trees such as (1, 2, 3) has prior probability 7.62 × 10^{-5}.

Now suppose that the single-gene posterior probability distributions for the three genes are concentrated collectively on four of the 15 possible tree topologies as shown in Figure 2a. Figure 2b shows the corresponding likelihood, prior probability, and posterior probability for each GTM for which the posterior probability is positive. The likelihood of each GTM is simply the product of the single-gene posterior probabilities from Figure 2a. Posterior probabilities are proportional to the product of the likelihood and the prior probability and



(a)

	single-gene		
	gene		
tree	1	2	3
2	1	0	0
3	0	0.9	0.2
4	0	0.1	0.2
15	0	0	0.6

(b)

GTM	Likelihood*	Prior Prob.	Post. Prob.
(2,3,3)	0.18	$8.38\ 10^{-4}$	0.6600
(2,3,4)	0.18	$7.62\ 10^{-5}$	0.0600
(2,3,15)	0.54	$7.62\ 10^{-5}$	0.1800
(2,4,3)	0.02	$7.62\ 10^{-5}$	0.0067
(2,4,4)	0.02	$8.38\ 10^{-4}$	0.0733
(2,4,15)	0.06	$7.62\ 10^{-5}$	0.0200

(c)

	multi-gene		
	gene		
tree	1	2	3
2	1	0	0
3	0	0.9	0.67
4	0	0.1	0.13
15	0	0	0.2

Figure 2—(a) Example of the single-gene posterior distribution for three genes, with one distribution in each column. (b) Posterior probabilities of GTMs after concordance analysis of the single-gene distributions shown in (a). All posterior probability is concentrated on 6 of the 15^3 = 3,375 GTMs. Likelihood is proportional to the product of single-gene posterior probabilities as in Equation (2). This product is reported here. The posterior probabilities are proportional to the product of the prior probability and likelihood and sum to 1. (c) Posterior distribution for each gene conditional on the data from all three genes assuming α = 1.5 to account for the expected concordance, derived from (b).

such that the sum over GTMs of their posterior probabilities is 1. In this example, the GTM with the highest likelihood does not have the highest posterior probability. This reflects the choice of α = 1.5, which gives a rather small prior probability to GTMs with three distinct trees.

The posterior probability that a gene is mapped to a specific tree topology, taking concordance into account, is found by summing the posterior probabilities over all corresponding GTMs. These posterior distributions adjusted for concordance with α = 1.5 are shown in Figure 2c. Notice that these distributions are unchanged for genes 1 and 2. In the former case, there was no uncertainty, so additional information from other genes made no difference. Because gene 3 gave equal probability to trees 3 and 4, the probability distribution for gene 2, which only includes trees 3 and 4, was unchanged. In contrast, information from genes 1 and 2 significantly alter the posterior probability distribution for the trees assigned to gene 3. This is because the high single-gene posterior probability on tree 3 from gene 2 "pulls" the posterior distribution for gene 3.

Second-Stage MCMC

Once individual-gene analyses have been performed, posterior probabilities of GTMs are to be calculated according to Equation (3). Direct calculation of posterior probabilities was carried out in our running example, but this is not tractable in realistic problems as soon as many GTMs have nonzero posterior probability. To compute posterior distributions of GTMs from real data, we introduce a novel MCMC approach. The input given to this MCMC is a table summarizing the individual-gene analyses similar to Figure 2a, with each column containing the posterior distribution on trees from a single gene. The state space of our Markov chain is the set of possible GTMs. The basic update procedure we use is to cycle through the genes having each gene in turn propose a topology according to its single-gene posterior distribution and either accept or reject the proposal by the Metropolis-Hastings criterion. This proposal, commonly called "importance sampling," treats the individual-gene posterior distribution as its importance density. The acceptance probability is the minimum of 1 and the product of the prior, likelihood, and proposal ratios. This basic update procedure changes only the tree for a single gene, so all factors but one in the likelihood from Equation (2) will cancel. If the update for gene i changes tree T_i to tree T_i^*, the likelihood ratio is $P\{T_i^* \mid X_i\}/P\{T_i \mid X_i\}$. As proposals are from single-gene distributions, the proposal ratio is $P\{T_i \mid X_i\}/P\{T_i^* \mid X_i\}$, which cancels the likelihood ratio exactly. Hence, the acceptance probability for this proposal is simply the prior ratio. If $g(T_i)$ and $g(T_i^*)$ are the sizes of the clusters containing trees T_i and T_i^* before the proposal, respectively, then the prior ratio is

$$\frac{A_{g(T_i^*)+1}(\alpha/T)}{A_{g(T_i^*)}(\alpha/T)}\ \frac{A_{g(T_i)-1}(\alpha/T)}{A_{g(T_i)}(\alpha/T)}, \qquad (8)$$

which with simplification leads to the acceptance probability

$$\min\left\{1, \frac{g(T_i^*)+\alpha/T}{g(T_i)-1+\alpha/T}\right\}. \qquad (9)$$

Note that the acceptance probability depends on a through the ratio α/T only. This basic update procedure can mix slowly. For example, to give all genes in a cluster a new tree, it would be necessary to propose and accept the changes one at a time, which can be problematic, especially because during the transition an extra cluster of trees is maintained. To speed mixing, we implemented two different strategies. The first was to use Metropolis-coupled MCMC (MCMCMC) where we run one "cold chain" whose stationary distribution is the desired one along with several "heated" chains (Geyer 1991). Rather than raising the likelihood to different powers as in MrBayes (Huelsenbeck and Ronquist 2001), we instead heat the chains by using different prior distributions where the jth heated chain uses Dirichlet process prior parameter $\alpha = c^j\alpha_0$ for some constant $c > 1$, where α_0 is the parameter for the cold chain. Larger values of α speed mixing because GTMs with more clusters have larger probability. In the limit as α is infinite, a single cycle through the genes is sufficient to reach the stationary phase of MCMC.

The second update we used proposes a new tree for all genes in a cluster. We begin by selecting a cluster at random. Next, we determine the set of possible trees where all genes in the cluster have positive single-gene posterior probabilities and no genes outside the cluster currently have this tree mapped in the current GTM. This set contains the current tree and possibly others. We select a new tree from this set with probability proportional to the product of the single-gene posteriors of genes in the cluster for the tree. As in the basic update, the likelihood and proposal ratios cancel. Here, the update does not change the number or sizes of clusters, so the prior ratio is also 1, and the proposal is always accepted.

Concordance Factor

In order to summarize and visualize the posterior distribution on GTMs, we consider concordance among genes with respect to clades of taxa. We define the "sample-wide concordance factor" of the clade c to be the proportion of genes in the sample whose true tree contains clade c, and denote this value as $\rho_s(c)$ where the s refers to sample. As the true GTM is unknown, we will typically estimate the sample-wide concordance factor with its posterior mean and with a credibility interval, although the modal value can also be useful.

For example, with respect to the previous numerical example and the trees in Figure 1, the clade c = {12|345} appears in tree 2 and tree 3 but not in tree 4 or tree 15. In Figure 2b, the first listed GTM is (2, 3, 3), meaning that all trees contain clade c in this case. If these GTMs were correct, the concordance factor would be 3/3 = 1. Each of the next three GTMs in the table have exactly two of three trees with c and correspond to concordance factor of 2/3, whereas the last two GTMs correspond to a concordance factor of 1/3. In this example, no GTMs correspond to a concordance factor of 0 for clade c. The Bayesian posterior distribution for the sample concordance factor is obtained by summing the posterior probabilities of the GTMs corresponding to each possible value. In this case, the complete posterior distribution for $\rho_s(c)$ is as follows: 1/3 with probability 0.0933, 2/3 with probability 0.2467, and 1 with probability 0.66. The mean of this distribution is $\hat{\rho}_s(c)$ = 0.856. The smallest interval that contains 95% of the posterior probability is [1/3, 1]. The modal value is 1, implying that all three true trees have clade c, but the probability of this estimate is only 0.66.

Primary Concordance Tree

In principle, we can find the posterior distribution of the concordance factor for each possible clade and then rank these by their posterior means. A useful summary of these results is the "primary concordance tree" (Baum 2007). This tree is built from the clades with highest rank until no more clades can be added to the tree. All clades with a posterior mean concordance factor above 50% are necessarily compatible and appear in the primary concordance tree. The reason is that the sum of the concordance factors of two contradictory clades cannot exceed 1 given any particular GTM. Therefore, if the true concordance factor of a clade is more than 50%, then the true concordance factor of any contradictory clade is below 50%. The same holds true for posterior mean concordance factors. When concordance factors can be estimated with high certainty, the primary concordance tree shows the history shared by a plurality of the sampled genes. More generally, a concordance network can be built from all clades with posterior mean concordance factor greater than some critical value. The critical value will be typically chosen so that the network is not too busy to provide useful graphical information. It will always contain the primary concordance tree.

In our running numerical example, each 5-taxon tree contains exactly two nontrivial splits. Table 1 shows the mean posterior concordance factors for the 6 clades with positive posterior probability estimates. The two clades with the highest posterior probabilities are {12|345} and {125|34}, so the primary concordance tree is tree 3. It is important not to interpret the mean concordance factor incorrectly. The value 0.856 is not an estimate of the proportion of the three sampled genes that have clade {12|345}—this true proportion must be an integer multiple of 1/3. The value of 0.856 represents a summary of the distribution of the probabilities over the integer multiples of 1/3 where most of the probability is either at 2/3 or 1. Likewise, if the concordance tree were annotated with concordance factors on nodes, these should not be confused with measures of support (bootstrap, clade credibilities), which generally provide information as to the confidence or probability that the clade exists given the assumption of a single shared tree for all the data.

Table 1
Mean Posterior Concordance Factor $\hat{\rho}_s$ for the Distribution in Figure 2b

Clade	{12|345}	{123|45}	{124|35}	{125|34}	{13|245}	{134|25}
Trees with clade	2,3	4	2	3,15	4	15
$\hat{\rho}_s$	0.856	0.053	0.333	0.589	0.053	0.067

Genome-Wide Concordance Factors

We define the "genome-wide" concordance factor $\rho(c)$ for a clade c as the proportion of genes in the genome for which clade c is in the true tree. If we consider the sampled genes to be a random sample from all genes (the total number N of genes in the genome being known) and if we assume that gene trees for all N genes in the genome follow a Dirichlet process with parameter α, then we can obtain the posterior distribution of $\rho(c)$ given the sequence data of only the G sampled genes. It is shown in Appendix that the posterior distribution of the genome-wide concordance factor $\rho(c)$ can be easily calculated from the posterior distribution of the sample-wide concordance factor $\rho_s(c)$, through a closed Formula (20). We give below some explanation of the relationship between the posterior distribution of $\rho_s(c)$ and the posterior distribution of $\rho(c)$. Consider first the case when the sequence data are abundant and give a very well supported gene tree for each gene, such that the concordance analysis gives most support to a single GTM. In this case, there is very high support for a single number j of gene trees that have clade c among the G sampled genes, and $\rho_s(c)$ is inferred to be j/G with very high posterior probability. Uncertainty remains as to how many of the N - G unsampled genes have clade c in their true tree. Under our assumptions, the posterior probability of this number depends on the sequence data through the number j only, and its shape is roughly the shape of a binomial distribution centered near j/G and with variance inversely proportional to the number G of sampled genes.

Now consider the general case when the number of sampled genes having clade c in their true tree is not estimated with high credibility. Instead, for each j between 0 and G, we know the posterior probability q_j that j of the sampled genes have clade c. Then the posterior distribution of the genome-wide concordance factor of clade c depends on the sequence data only through the q_j's. Each value j contributes a distri-

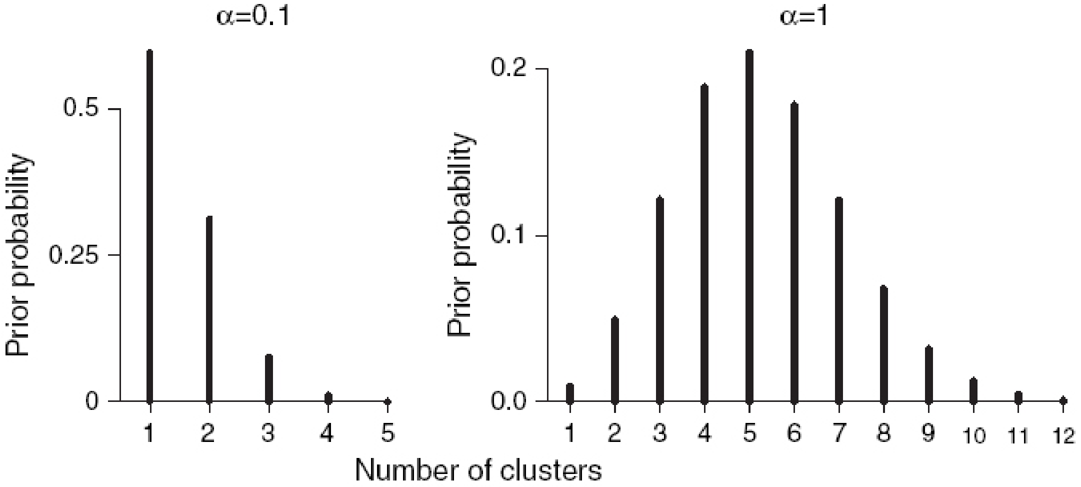


Figure 3—Prior distribution of the number of distinct gene trees as defined by the Dirichlet process with parameter α = 0.1 (left) and α = 1 (right) on 8-taxon trees and 106 genes.

bution as described above (centered near j/G), but this distribution is weighted by the posterior probability qj. In other words, each bar in the histogram of the posterior distribution of $\rho_s(c)$ contributes a small histogram, and these small histograms are added together to make the posterior distribution of $\rho(c)$. As a result, the posterior distribution of $\rho(c)$ will resemble that of $\rho_s(c)$, but it will be more dispersed, depending on the number of sampled genes G.

Multigene Yeast Data Set

Rokas et al. (2003) report apparent discordance among 106 genes sampled from 8 yeast species, even though a total evidence approach returned a tree with 100% bootstrap support (see also Phillips et al. 2004; Gatesy and Baker 2005; Burleigh et al. 2006). We reanalyzed this data set using our concordance approach. For each individual gene, we carried out an analysis using MrBayes. In each case, we used the general time reversible model with invariant sites and gamma distributed rates, employing MCMCMC with one cold and three heated chains, with default parameters for the prior distribution. We used the default values for tuning parameters, running each chain for 550,000 updates, discarding the first 50,000 as burn-in, and subsampling every 10th tree resulting in 50,000 sampled trees per gene. The combined sample included 479 distinct tree topologies. These separate samples were summarized in a 479 × 10^6 table where each column represents a single-gene posterior distribution.

In our approach, we allow different genes to have different topologies but allow the information from all other genes to improve the estimated tree topology of each particular gene. The parameter α modulates the strength of our prior beliefs about the number of distinct gene trees. We considered first a prior distribution with α = 0.1 for which the prior probability of no discordance (one tree shared by all 106 genes) is about 60%. To examine the effects of this prior assumption, we also considered a prior distribution with α = 1, which predicts considerably more discordance. These distributions have 1.51 and 5.24 expected distinct gene trees, respectively, among 106 genes and are displayed in Figure 3. An interactive Web site http://www.stat.wisc.edu/~larget/bucky.html plots the distribution of the number of distinct gene trees for a selected number of genes, number of taxa, and choice of α.

The software that implements the second-stage MCMC is named BUCKy (Larget 2006). For each value of α, we ran two independent sets of MCMCMC runs. Each run used one cold chain and 7 heated chains, set burn-in for 100,000 cycles, and retained an additional 1,000,000 cycles for analysis. Genes were updated one at a time (basic update). On a Dell machine with 1 Gb memory and a 2.8-GHz processor, each concordance analysis run required approximately 24 min. When the second update, wherein clusters of genes are reassigned to new trees, was added, the algorithm ran considerably slower (about 6 h) but found similar results. We examined the output from both runs with the same α for numerical similarity to check convergence.

For assessing the uncertainty in the genome-wide concordance factor estimate, we used a total number of N = 6,000 genes in the yeast genome (Goffeau et al. 1996). Additionally, we used the value N = 60,000 in order to assess the robustness of our method to a change in the value of N. In addition to determining the posterior distribution of the sample-wide and genome-wide concordance factors of the major

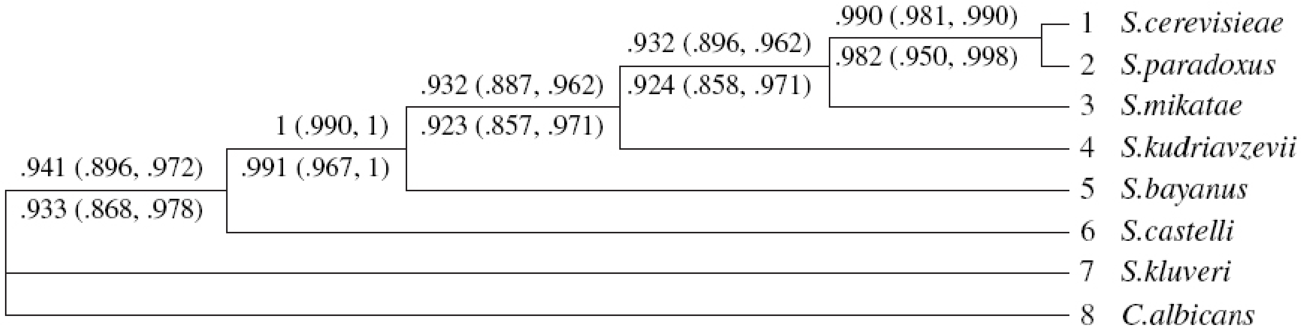


Figure 4—Total evidence tree of the yeast data set, from Rokas et al. (2003). It is also the primary concordance tree. Numbers are posterior mean concordance factors and their 95% credibility intervals, obtained with α = 1. Above edges, numbers refer to sample-wide concordance factors (qs), and below edges, numbers refer to genome-wide concordance factors.

clades, we determined the posterior distribution of the proportion of genes in the sample having the total evidence tree as their true tree. The posterior distribution of the genome-wide proportion of genes with this total evidence tree was calculated in a similar way that genome-wide concordance factors were inferred from sample-wide concordance factors (see Appendix).

Because GTMs contain information about which genes share the same tree and which genes do not, for each pair of genes we determined the posterior probability that the two genes share the same tree. A dissimilarity measure was obtained by subtracting this posterior probability from 1. Classical multidimensional scaling (Cox TF and Cox MAA 2001) was then used to display the dissimilarities among all 106 genes. This method places the genes on a two- dimensional space so that the pairwise distances between genes best approximates the pairwise dissimilarities.

Results

Yeast Data

Individual-Gene Analyses

The tree topology for the total evidence analysis in Rokas et al. (2003) is displayed in Figure 4. A Bayesian analysis on the concatenated data set has nearly 100% posterior probability for this tree as well. The average single-gene posterior probability of this tree across the 106 genes is 0.38. Equivalent averages for all other tree topologies are below 0.10, so this topology is the best single topology to explain all the data. Nonetheless, single-gene analyses reveal much potential discordance. In our analysis of the 106 genes, the tree topology in Figure 4 is the most probable tree in only 44 genes, has posterior probability less than 0.10 in 39 genes, has posterior probability less than 0.01 in 16 genes, and has estimated posterior probability of 0 in 3 of the 106 genes (*YGL192W*, *YJR068W*, and *YLR253W*). Although the posterior distribution for the first two of these genes overlaps with the distributions of many of the other genes, the single-gene distribution from *YLR253W* has positive probability on a set of tree topologies that has no overlap with the set of trees positively supported by any of the other 105 genes. The posterior distribution of the gene tree for *YLR253W* has probability greater than 99% for the clade with *Saccharomyces cerevisieae* and *Saccharomyces kudriavzevii*, whereas no other single-gene tree posterior distribution has measurable support for this clade. The collection of most probable trees among the 106 genes includes 18 separate trees. Most genes have considerable uncertainty, however, and the most probable tree is often not very strongly supported. Figure 5 shows the topologies with highest posterior probabilities for two typical genes and for the three genes that do not have support for the total evidence tree.

Second-Stage Concordance

Analysis

The individual-gene posterior probabilities were subject to a second round of MCMC to estimate the posterior distribution over GTMs. The posterior probability distribution was concentrated on GTMs with three distinct gene trees for both α = 0.1 and α = 1, with probabilities 0.997 and 0.98, respectively. In both cases, the probability of more than five distinct gene trees was negligible, and the probability of a single or two distinct gene trees was estimated as 0.

All the clades in the total evidence tree had mean sample-wide and genome-wide concordance factors that exceeded 0.92 (Figure 4) for both values of α. Figure 4 shows the posterior mean concordance factor of the dominant clades and their 95% credibility intervals. Although the 95% credibility intervals around these concordance factors are narrow,

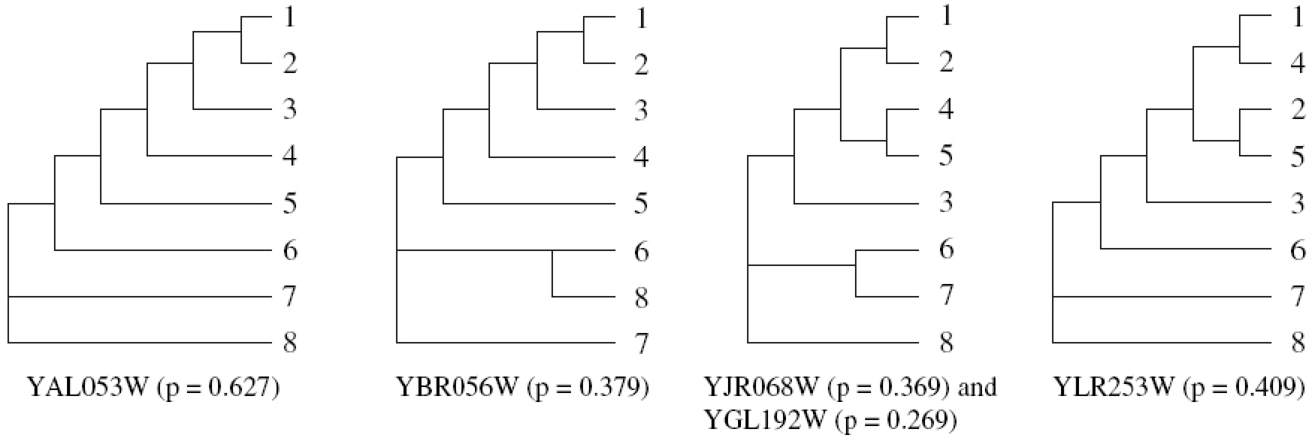


Figure 5—Trees with highest posterior probability *p* from individual-gene analyses. Species labels are as in Figure 4.

the genome-wide concordance factors were subject to greater uncertainty, as expected. For example, we infer that the clade {12345 | 678} (translation of numbers into taxon names as in Figure 4), uniting *S. cerevisieae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus*, is shared by all 106 genes (concordance factor of 1) and clade {12 | 345678}, uniting *S. cerevisieae* and *S. paradoxus*, is shared by all genes except for *YLR253W* with extremely high probability (concordance factor of 105/106). However, the genome-wide concordance factors for the two aforementioned clades have 95% credibility intervals spanning (0.97, 1) and (0.95, 1), respectively. The posterior distribution of the sample-wide and genome-wide concordance factors of clade {1234 | 5678} (with $\alpha = 1.0$) are shown in Figure 6. Figure 6 also shows distributions as might been obtained had there been only 26 genes (about a fourth of 106). To simulate such a reduced sample, the sample-wide concordance factor was subsampled while retaining the shape of the posterior distribution and then the concordance distribution was extrapolated to the whole genome assuming that 26 of 6,000 genes had been sampled. This artificial example shows that the posterior distribution of the genome-wide concordance factor is expected to become broader and is shifted toward low values when a lower number of genes are sampled. The shift of the mean genome-wide concordance factor toward low values reflects the fact that even with $\alpha = 1.0$, the prior probability of a given clade (especially a large clade) occurring in many gene trees is low. The total number of genes N in the genome has little effect on the analysis when $N \gg 1$: the posterior distributions of the genome-wide concordance factor (see Figure 6) obtained with $N = 6{,}000$ and $N = 60{,}000$ were not distinguishable when rescaled as density functions to be comparable to each other.

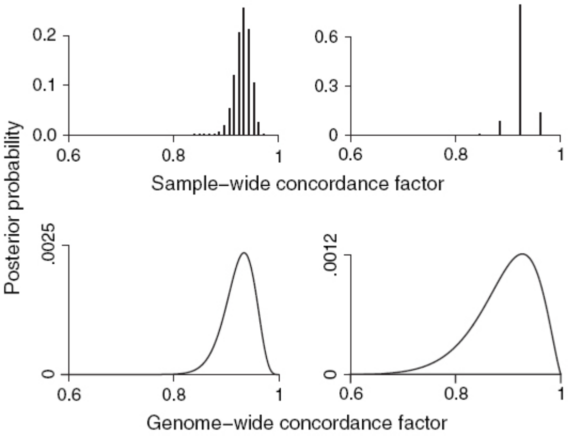


Figure 6—Left: posterior distribution of the sample-wide and genome-wide concordance factors of clade {1234 | 5678} (see Figure 4), obtained with $\alpha = 1$. The posterior distribution of the genome-wide concordance factor was obtained assuming a total number of $N = 6{,}000$. Right: expected posterior distribution of the sample-wide concordance factor (top) for clade {1234 | 5678}, as might be obtained from 26 genes only. The genome-wide concordance factor has a more dispersed posterior distribution (bottom), reflecting the lower number of sampled genes.

Table 2 shows both the single-gene and concordance-based posterior probabilities for each clade in the total evidence tree topology, as well as a few other selected clades, for four of the 106 genes. Genes *YAL053W* and *YBR056W* had a single-gene posterior probability for the total evidence tree topology of 63% and 31%, respectively, largely due to uncertainty in the relationships among *S. castelli*, *S. kluveri*, and *Candida albicans* (see Figure 5 as well). After the concordance analysis, these two genes have a posterior probability near 1 for the total evidence topology, as shown in Table 2. On the other hand, the concordance analysis pulls the posterior distribution from gene *YGL192W*, which initially gives no support to the total evidence tree, onto the tree that receives the highest support from the individual analysis of this gene (see Figure 4). In contrast, the single-gene posterior distribution

Table 2
Posterior Probability that a Given Clade Is True for a Given Gene from Individual-Gene and Concordance Analyses

	YAL053W		*YBR056W*		*YGL192W*		*YLR253W*	
Clade	sing.	conc.	sing.	conc.	sing.	conc.	sing.	conc.
{12 | 345678}	1.000	1.000	1.000	1.000	0.609	1.000	0.000	0.000
{123}45678}	0.998	1.000	0.843	0.998	0.004	0.000	0.000	0.000
{1234 | 5678}	0.982	1.000	0.929	0.998	0.000	0.000	0.002	0.002
{12345 | 678}	1.000	1.000	1.000	1.000	0.508	1.000	1.000	1.000
{123456 | 78}	0.637	1.000	0.374	0.997	0.001	0.000	0.999	0.999
{1245 | 3678}	0.002	0.000	0.065	0.002	0.330	1.000	0.410	0.410
{14 | 235678}	0.000	0.000	0.000	0.000	0.000	0.000	0.996	0.997

Note—sing., single; conc., concordance.

for the topology of gene *YLR253W* is identical to its concordance-based distribution because there was no measurable overlap in its single-gene distribution with those from any other gene.

After the concordance analysis, 100 genes give most support to the total evidence topology (as compared with 44 genes without accounting for concordance). Based on the GTMs sampled in the second-stage MCMC, the number of genes mapped to the total evidence tree is at least 80 (75% of the genes) with 100% posterior probability, and most probably 99 (93% of the genes), for both values of α. The extension to the genome-wide proportion showed that there is over 0.98 posterior probability that at least 85% of the genes in the genome have the total evidence tree as their true tree. Therefore, under our model, there is very strong evidence that the primary concordance tree shown in Figure 4 reflects the actual history of the majority of the genes in the yeast genome.

Pairwise Gene Dissimilarities

From the concordance analysis, we estimated the posterior probability that two specific genes do not share the same tree, which we used as a measure of dissimilarity between genes. Figure 7 shows an ordination of genes using classical multidimensional scaling with two dimensions, which explains 92.7% of the variation in gene-gene similarity (see Cox TF and Cox MAA [2001]). In this data set, for both values of α we considered, most sampled GTMs contained three distinct tree topologies, one of which is the total evidence tree topology, the second is the topology shown for genes *YGL192W* and *YJR068W* in Figure 5, and the third is one of the several tree topologies supported by *YLR253W*. Representation of gene-gene similarity in Figure 7 reflects this highly supported structure with three clusters of genes. The horizontal axis separates the genes based on how often they belong to the total evidence tree cluster, that is, based on their single-gene posterior probability for the total evidence tree. Most genes cluster in the top-left corner and support the total evidence tree, whereas genes on the far right (like *YJR068W* and *YGL192W*) do not support the total evidence tree and most likely belong to the second cluster. As expected, gene *YLR253W* appears as an outlier because it never shares the same tree with any other gene. It forms the third cluster by itself.

Table 3
Absolute Values of Stirling Numbers of the First Kind $a(n, k)$ (left) and Second Kind $S(n, k)$ (right)

		k		
n	0	1	2	3
1	0	1	0	0
2	0	1	1	0
3	0	2	3	1

		k	
n	1	2	3
1	1	0	0
2	1	1	0
3	1	3	1

Discussion

Bayesian Concordance Analysis

The method of analysis we have described provides biologists an alternative to either analyzing all data together in a concatenated matrix or analyzing each data set separately without sharing any information. As a Bayesian method, it requires that the biologist specifies prior expectations on the amount of concordance that is expected. Concordance is here modeled using a Dirichlet process prior which allows one to summarize expected concordance using a single parameter, α, that describes the probability of randomly selected genes having the same true gene tree and yields a prior probabilitydistribution on the number of distinct gene clusters (where a cluster is a set of genes mapped to the same tree). Setting α close to 0 embodies the biological assumption that the sampled taxa have not been subject to phenomena such as introgression, lateral gene transfer, or incomplete lineage sorting, in which case analysis of a single concatenated data set would be called for. At the other extreme, setting α close to infinity, implying independence among genes and pervasive genealogical discordance, might be appropriate when all the sampled individuals were drawn from a single panmictic population (and sampled genes are in linkage equilibrium). In this latter case, information from one locus has no informational value in selecting the correct genealogy for another

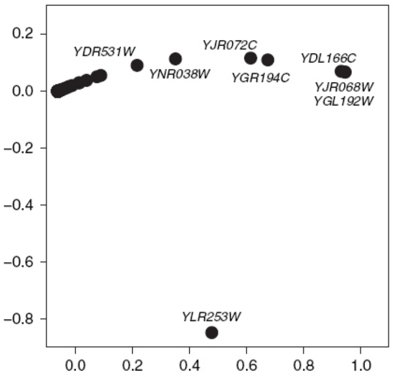


Figure 7—Two-dimensional representation of the 106 yeast genes. Similarity between pairs of genes is measured as the posterior probability that the two genes share the same tree (α = 1). Most genes cluster in the top-left corner and are not labeled.

locus, similar to standard consensus methods. Hence, Bayesian concordance analysis, as we call our method, allows one to select a point on the continuum from total evidence to consensus assumptions that is judged most plausible for the data at hand. Given that in almost all biologically reasonable cases a prior assignment of α that is greater than 0 and less than infinity would seem desirable, Bayesian concordance analysis would seem preferable to either a total evidence or consensus procedure for many multigene data sets.

The concern that might arise, common to other branches of Bayesian phylogenetics, is how a value of α should be selected *a priori*. The practitioner should use as much prior information as possible to assess the probability of a given amount of concordance. Information might come from previously analyzed data and such ancillary factors as the taxonomic scale of the study, population size (hence the chances of incomplete lineage sorting), and the propensity for hybridization or lateral gene transfer in the study organisms. Through the use of the interactive Web site provided (http://www.stat.wisc.edu/~larget/bucky.html), a biologist can identify an upper and lower bound for their prior expectations of concordance and compare the results to determine whether the choice of prior has an effect on the posterior distribution, as summarized, for example, in concordance factors. Although we find it valuable to be forced to articulate our prior beliefs of the degree of discordance, it is easy enough to imagine modifying the method such that one only selected a noncommittal prior distribution on a and then used the second-stage MCMC to obtain a posterior probability distribution of this parameter.

Missing Data

So far we have assumed that the sequence data of every sampled gene is available from every sampled taxon. However, it is frequent in multigene data sets that some gene sequences are missing for some taxa. Such missing data represents a technical issue, but the Bayesian concordance analysis can still be carried out on such data sets. Suppose, for instance, that gene *i* has sequence data X_i on a taxon set L_i, but that L_i misses some taxa from the full taxon set *L*. If L_i is not the complete list of all taxa, then it means gene *i* is missing for some taxa. In order to carry out the second stage of the concordance analysis, we need to know the posterior distribution of trees with all taxa in *L* from gene *i*. One naive way to proceed is to artificially add sequences of missing data to X_i and run MrBayes on this new data set. Although we have used such an approach successfully for small data sets (not shown), MrBayes (or any program) tends to mix very slowly in the presence of missing data. Also, considering the huge number of tree topologies that will have equal likelihood due to differing only in the position of a taxon with missing data, it is not clear if the typical number of sampled trees retained during first-stage MCMC will accurately sample the posterior distribution of tree topologies. Another way to proceed is to run the individual analysis of gene *i* on the original data, which yields the posterior distribution of trees with taxa in L_i given the data on gene *i*. The posterior distribution on full trees (with all taxa from *L*) can then be obtained analytically from the posterior distribution on smaller trees with

$$P\{\tau|X_i\} = \frac{P\{\tau_i|X_i\}}{K_i},$$

where τ is any tree on the full taxon set and τ_i is its subtree restricted to taxa from L_i only. Indeed, all trees containing the same subtree on L_i have the same likelihood given the data X_i, and when a uniform prior is used on tree topologies, all these trees have the same posterior probability. Each fixed tree on the smaller taxon set L_i with m_i taxa is the sub-tree of $K_i = (2m = 5)!!/(2m_i = 5)!!$ larger trees on the taxon set *L* (with *m* taxa), which determines the proportionality constant 1/K_i. Missing sequences make the posterior distribution on full trees very dispersed, which will likely increase the computational burden with high values of α.

Computation

Our approach involves single-gene analyses being run independently, with summaries of these runs then being input into the second-stage concordance analysis. This two-step procedure is computationally efficient because the initial analyses can be run in parallel. Also, despite the fact that the second MCMC is exploring the immense GTM state space, the acceptance ratio is easily calculated and mixing is quite efficient with the use of a diversity of update procedures. Nonetheless there are two obvious disadvantages of this two-step procedure. First, it requires one to assume that parameters other than tree topology are independent among genes, despite the fact the we might also expect other parameters such as relative branch lengths, to show some correlation among genes. Second, because the first-stage analysis retains only a finite sample from the individual-genes' posterior distributions, numerical errors will arise and will be propagated into the second stage of concordance analysis. This is a particular issue for tree topologies supported by some genes where the posterior probability is measured as 0 in other genes. The true posterior probability for every tree topology and every gene should be positive, not 0. In most cases, we can assume that a tree whose posterior probability is low enough never to occur in an MCMC sample is not the true tree for that gene. However, given an assumption of high concordance (i.e., low α compared with the total number *T* of topologies), a tree with a very low posterior probability from independent analysis of a single gene could nonetheless be a plausible tree for that gene if that tree has high posterior support from other genes. More work needs to be done to determine the real impact of inaccuracies in numerical summaries of single-gene posterior distributions. Also, we see some hope for the development of methods for estimating the posterior probability of specific tree topologies that are measured as 0 in standard single-gene Bayesian analyses.

An alternative strategy that could overcome the two problems with the two-step procedure should be mentioned: a one-step MCMC that simultaneously considers the probability of each tree for each gene, given both the sequence data (under standard priors for Bayesian phylogenetics) and the trees assigned to all other sampled genes (under the Dirichlet process prior with a specified value of α). Programs such as MrBayes already allow multiple genes (partitions) to be analyzed simultaneously while either assuming *a priori* that parameters are "linked" (the parameter is shared between the partitions) or "unlinked" (fully independent). A single-step Bayesian concordance MCMC analysis would not be forced to assume that parameters (tree topology, branch lengths, etc.) are completely linked or completely independent but could allow one to select a level of linkage according to one's prior expectations of concordance. However, our intuition is that such a single-step procedure would be computationally prohibitive for all but the smallest data sets.

Comparison with Recent Methods Accounting for Different Gene Trees

After the review of the manuscript, we became aware of a very similar approach used by L.-J. Liang and R.E. Weiss (personal communication). They propose a method similar to ours, where genes are first analyzed individually and then a second-stage analysis places a mixed Dirichlet process prior on the parameter of interest (ti/tv ratio, for instance) and uses the individual-gene samples of this parameter for importance sampling and MCMC. As with our method, L.-J. Liang and R.E. Weiss (personal communication) do not assume that all genes have tracked the same topology. However, whereas we make inferences on the tree topology, they treat topology as a nuisance variable and instead make inference on one or several continuous evolutionary parameters. Liu and Pearl (2006) have also recently proposed an interesting method for estimating multiple gene trees, under the assumption that discordance among gene trees is due to lineage sorting. By enforcing a molecular clock on branch lengths, their method has the advantage of sharing branch length information among genes.

Pros and Cons of the Dirichlet Process Prior

The Dirichlet process prior has some desirable features for modeling concordance. In particular, this process can be described with a single parameter α and is mathematically well understood and tractable. Furthermore, it is relatively easy to elicit meaningful priors from biologists based on prior beliefs as to the number of distinct gene trees. However, the Dirichlet process prior has a few disadvantages that should be noted. First, it implies a particular distribution of cluster sizes in which the number of genes in the largest cluster will be considerably larger than the number of genes in the second largest cluster. For example, with 100 genes and a low value of α compared with the total number of trees (T), the largest cluster contains 86.7 genes on average when there are two clusters. However, it is easy to imagine biological situations, for example, cases where a hybrid speciation event is thought to have occurred, in which one expects the largest two clusters to include a similar number of genes. Likewise, the Dirichlet process implies that even among minor clusters, equality in cluster size is not expected. With 100 genes and a low value of α, for example, when there are three clusters, the largest one has mean size of 76.5 genes, the second largest cluster has mean size of 18.8 genes, and the smallest cluster contains only 4.7 genes on average. This is at odds with expectations of lineage sorting on a resolved divergent population history, for which one expects the two minor histories to be represented at equal frequency in the genome (Pamilo and Nei 1988). We plan to explore different models for concordance that tend to predict more balanced cluster sizes.

A second, perhaps larger, problem with the Dirichlet process prior is its assumption that tree topologies are exchangeable. A GTM consisting of several distinct trees that are topologically similar would be just as likely *a priori* as a GTM with the same numbers of distinct trees but where the topologies are much more dissimilar. This arises because when a joker is drawn during the Dirichlet process, a card is drawn from a side deck that contains all trees at uniform frequency, rather than a set of trees whose distribution is shaped by the identity of the trees that are already in the urn. As a result, when individual-gene posterior distributions share a positive posterior probability for the same tree, the model does not allow that they might each have tracked different, but similar trees, with the overlap in their posteriors being due to "bleeding" of the posterior from one tree to other similar trees. This means that the Dirichlet process prior will tend to favor strong concordance among gene trees when gene trees are actually different but topologically similar. As a result, it may tend to overestimate concordance. To deal with this, we plan to explore a model that places a prior distribution on the pairwise distance between the trees of two randomly selected genes and also a strategy for using coalescent theory to simultaneously estimate the GTM and an underlying divergent population tree. Thus, although the use of a Dirichlet process prior does not solve all problems in the analysis of multigene data sets, it is a valuable step toward a rigorous statistical analysis of concordance and provides a leaping off point for future extensions to yield improvements in computational efficiency and to allow the more realistic incorporation of such biological phenomena as coalescence and genetic linkage.

Acknowledgments

D.A.B. and B.L. are supported in part by National Institutes of Health (NIH) grant R01 GM068950 and B.L. is supported in part by NIH grant R01 GM069801 and by National Science Foundation grant DEB 0445453.

Appendix

Likelihood Model

Consider first the likelihood model $P\{X\,|\,M\}$. Conditional on the GTM M, we assume that the G genes evolve independently, each according to its own evolutionary model of substitution with parameters $\theta = (\theta_1, \ldots, \theta_G)$. The conditional probability of the data $X = (X_1, \ldots, X_G)$ given M is then

$$P\{X\,|\,M\} = \int_\theta P\{X\,|\,M, \theta\}\, d\pi(\theta), \tag{10}$$

which depends on the assumption that M and θ are independent *a priori*. If we assume further that the gene-specific evolutionary parameters are mutually independent and that the GTM M assigns tree T_i to gene i, the likelihood simplifies to

$$\prod_{i=1}^{G} \int_{\theta_i} P\{X_i|T_i, \theta_i\}\, d\pi_i(\theta_i) = \prod_{i=1}^{G} P\{X_i|T_i\}, \tag{11}$$

because $\pi(\theta) = \prod_{i=1}^{G} \pi_i(\theta_i)$ and T_i and θ_i are independent for each i. By applying Bayes' theorem to each factor and using any prior distribution on tree topologies (not related to any assumption in our model), the previous expression becomes

$$\prod_{i=1}^{G} \frac{P\{T_i|X_i\}P\{X_i\}}{P\{T_i\}}, \tag{12}$$

so that the likelihood of the map M is proportional to the product of the individual-gene ratios of posterior to prior probabilities of the tree topology. Further, if we use the uniform prior distribution on topologies for each individual gene, we conclude that the likelihood of the map M is proportional to the product of the corresponding individual-gene posterior probabilities.

$$P\{X \mid M\} \propto \prod_{i=1}^{G} P\{T_i \mid X_i\}. \qquad (13)$$

GTM Prior Distribution

This expression depends in part on the coefficients of the polynomial $A_n(x)$ which when expanded is seen to be an nth degree polynomial in x with the form

$$A_n(x) = \sum_{k=1}^{n} a(n,k)x^k \qquad (14)$$

The coefficients $a(n, k)$ are integers and are absolute values of so-called Stirling numbers of the first kind which arise in many combinatorial settings. These numbers can be computed recursively with a(1, 1) = 1, $a(1, k) = 0$ for $k > 1$, $a(n, 0) = 0$ for $n > 0$, and $a(n, k) = a(n - 1, k - 1) + (n - 1)a(n - 1, k)$.

The typical Dirichlet process prior on a continuous parameter space generates a new distinct value each time the joker is drawn. The prior probability of k distinct sampled values in α sample of size n is then

$$\frac{a^k a(n,k)}{A_n(a)} \qquad (15)$$

(Antoniak 1974). This expression requires modification in the discrete case. We find this expression in the special case of a uniform distribution.

A GTM can contain k distinct gene trees if the joker is sampled i times for any $i \geq k$. If J is the number of times the joker is drawn and if D is the number of distinct gene trees, then

$$P\{D = k\} = \sum_{i=k}^{G} P\ \{J = i\}\ P\ \{D = k \mid J = i\}, \qquad (16)$$

where $P\{J = i\}$ is found from Equation (15). To find an expression for $P\{D = k\} \mid J = i\}$; we begin by noting that every sequence of i trees drawn from the uniform distribution has probability (1/T)i: we need to simply count the number of sequences with k distinct trees. First there are $S(i, k)$ ways to specify which of the i draws of the joker correspond to each of the k distinct trees where $S(i, k)$ is a Stirling number of the second kind. (In general, $S(n, k)$ counts the number of ways to partition n objects into k nonempty groups.) Given this specified partition, there are $T(T - 1)\ldots(T - k + 1) = A_k(T - k + 1)$ ways to choose the specific set of k distinct trees in the order selected. Putting this together, we find that $P\{D = k \mid J = i\} = S(i, k)A_k(T - k + 1) = T_i$, so that the prior probability distribution for k distinct gene trees among G sampled genes is

$$P\{k|\alpha, T, G\} = \frac{A_k(T-k+1)}{A_G(\alpha)} \sum_{i=k}^{G} a(G,i)S(i,k)\left(\frac{\alpha}{T}\right)^i. \qquad (17)$$

Notice that if T is quite large as will be the case for even moderate numbers of taxa and if a is much smaller than T which will be the case when we expect few distinct trees *a priori*, then the preceding sum is dominated by the first term $(\alpha^k a(G, k)/A_G(a)) \times A_k(T - k + 1)/T^k$ which is nearly identical to Equation (15) as the last factor is very nearly 1. On the other hand, if T is small, this probability vanishes when $k > T$ as $A_k(T - k + 1)$ would contain a factor of 0. Equation (7) is numerically tractable because Stirling numbers of the second kind can be computed by a recursive relationship similar to that for Stirling numbers of the first kind: $S(1, 1) = 1$, $S(1, k) = 0$ for $k > 0$, $S(n, 1) = 1$ for $n \geq 1$, and $S(n, k) = \mathrm{S}(n - 1, k - 1) + kS(n - 1, k)$.

Genome-Wide Concordance Factor

Consider a clade c partitioning the n taxa into two groups of m and $n - m$ taxa. Consider first the prior distribution of the number of trees with clade c among N genes. Let p_c be the proportion of tree topologies having clade c. This proportion depends on the clade's sizes only, and $p_c = \mathrm{UB}(m + 1)\mathrm{UB}(n - m + 1)/\mathrm{UB}(n)$ where $\mathrm{UB}(k) = (2k - 5)!!$ is the number of unrooted bifurcating trees with k taxa (and $\mathrm{UB}(k + 1)$ is the number of rooted trees with k taxa). Under the Dirichlet process, each time the joker is selected, the next gene tree has probability p_c of having the clade. If instead a regular tree card is selected, the probability of selecting a tree with clade c only depends on how many trees currently in the urn have the clade. Therefore, if we summarize the gene trees to a sequence of zeros and ones, assigning 1 to a gene when its tree has the clade and 0 when its tree does not have the clade, then this 0/1 process is a Dirichlet process with parameters α (weight of the joker) and frequencies $f_1 = p_c$ and $f_0 = 1 - p_c$. There are $\binom{N}{k}$ sequences of zeros and ones with exactly k ones, each having the same prior probability. This probability is obtained by applying Equation (5) to the new 0/1 Dirichlet process. It follows that the prior probability of sampling exactly k trees with clade c among N gene trees is

$$\binom{N}{k} \frac{A_k(\alpha p_c)A_{N-k}(\alpha(1-p_c))}{A_N(\alpha)}. \qquad (18)$$

Because the first gene tree (and so any gene tree) has probability p_c of having the clade, the mean of this distribution is Np_c. However, this distribution is more dispersed than the binomial distribution with same sample size N and same mean Np_c. (The binomial distribution has variance $Np_c(1 - p_c)$, whereas the prior distribution of the number of genes with clade c has a variance $(N - \alpha)/(1 + \alpha)$ times bigger.) Now suppose that we have a genome of N genes for which we have sequence data from G genes. Let q_j be the posterior probability that exactly j of the G sampled genes have clade c. Consider first the case when we know the true sample-wide GTM (the GTM of the G sampled genes). In particular, we know the exact number j of gene trees among the G genes that have clade c, so that $q_j = 1$ and all other q_j' are 0. In this case then, sampling an additional $N - G$ genes and mapping the new genes to state 1 when the gene tree has clade c and 0 otherwise is again like a Dirichlet process, no matter what the sample-wide GTM actually is. For this process, the joker has weight $\alpha + G$—the total weight of all the cards in the urn after sampling the first G genes and frequencies are $f_1 = (\alpha p_c - j)/(\alpha + G)$ and $f_0 = (\alpha(1 - p_c) + G - j)/\ (\alpha + G)$ (the probability of getting or not getting the clade for the first newly sampled gene). The posterior probability that k genes among the N genes in the genome have clade c is then the probability that $k - j$ of the nonsampled $N - G$ genes have the clade, which is then

$$\binom{N-G}{k-j}\frac{A_{k-j}(\alpha p_c+j)A_{N-G-k+j}(\alpha(1-p_c)+G-j)}{A_{N-G}(\alpha+G)} \tag{19}$$

by applying Equation (5) to the new 0/1 Dirichlet process. We now turn to the general case, when the sample-wide GTM M_s is not known with certainty. Because the sequence data X from the G sampled genes and the N - G non-sampled gene trees are independent given the sample-wide GTM, it follows that the posterior distribution of the genome-wide GTM $M = (M_s, M_{ns})$ is

$$P\{(M_s, M_{ns}) \mid X, \alpha\} = P\{M_{ns} \mid M_s, \alpha\} \cdot P\{M_s \mid X, \alpha\}$$

Furthermore, the posterior distribution of the number of trees in M_{ns} with clade c depends on the sample-wide GTM M_s only through the number of gene trees in M_s with clade c, so that the posterior probability that k genes in the genome have clade c is $\sum_{j=0}^{G} P\{M$ has k trees with $c \mid M_s$ has j trees with $c \mid X, \alpha\}$ $P\{M_s$ has j trees with $c \mid X, \alpha\}$. In the sum, the first term is given by Equation (19), whereas the second term is just q_j, so that the posterior probability that k genes among the N genes in the genome have clade c is

$$\sum_{j=0}^{G} q_j \binom{N-G}{k-j}\frac{A_{k-j}(\alpha p_c+j)A_{N-G-k+j}(\alpha(1-p_c)+G-j)}{A_{N-G}(\alpha+G)}. \tag{20}$$

The same derivations can be made with any feature looked for in gene trees. The feature considered above was the property of having the clade c. A similar formula is obtained with, for example, the property of being equal to a fixed tree τ, such as the total evidence tree. Equation (20) still relates the posterior distribution of the sample-wide proportion of genes with tree τ to the posterior distribution of the genome-wide proportion of genes with tree τ. The prior probability p_c of the clade just needs to be replaced by $1/T$, the prior probability of the tree τ. We implemented these calculations in the R programming language (Chambers 1998) and made our R functions available at http://www.stat.wisc.edu/~larget/bucky.html.

Literature Cited

Antoniak C. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann Stat* 2: 1,152-1,174.

Baker RH, DeSalle R. 1997. Multiple sources of character information and the phylogeny of Hawaiian Drosophilids. *Syst Biol* 46: 654-673.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290: 972-977.

Barrett M, Donoghue MJ, Sober E. 1991. Against consensus. *Syst Zool* 40: 486-493.

Baum DA. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* 56: 417-426.

Burleigh JG, Driskell AC, Sanderson MJ. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Syst Biol* 55: 426-440.

Chambers JM. 1998. *Programming with Data*. New York: Springer.

Cox TF, Cox MAA. 2001. *Multidimensional Scaling*. Boca Raton, Fla.: Chapman and Hall/CRC.

Cronn R, Small RL, Haselkorn T, Wendel JF. 2003. Cryptic repeated genomic recombination during speciation in *Gossypium gossypioides*. *Evolution* 57: 2,475-2,489.

de Queiroz A. 1993. For consensus (sometimes). *Syst Biol* 42: 368-372.

Ferguson TS. 1973. A Bayesian analysis of some nonparametric problems. *Ann Stat* 1: 209-230.

Gatesy J, Baker RH. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? *Syst Biol* 54: 483-492.

Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. *In*: Keramidas EM, editors. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Fairfax Station, Va.: Interface Foundation, p. 156-163.

Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* 413: 157-161.

Goffeau A, Barrell BG, Bussey H, et al. (16 co-authors). 1996. Life with 6,000 genes. *Science* 274: 546-567.

Holland BR, Huber KT, Moulton V, Lockhart PJ. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol Biol Evol* 21: 1,459-1,461.

Holland BR, Jermiin LS, Moulton V. 2006. Improved consensus network techniques for genome-scale phylogeny. *Mol Biol Evol* 23: 848-855.

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. *Mol Biol Evol* 19: 698-707.

Huelsenbeck JP, Larget B, Alfaro M. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol* 21: 1,123-1,133.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754-755.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23: 254-267.

Kellogg EA, Appels R, Mason-Gamer RJ. 1996. When genes tell different stories: the diploid genera of Triticeae (Gramineae). *Syst Bot* 21: 321-347.

Kluge AG. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst Zool* 38: 7-25.

Larget B. 2006. *Bayesian Untangling of Concordance Knots* (*BUCKy*), version 1.1 [cited November 27, 2006, from the Internet]. Department of Statistics, University of Wisconsin. Available from: http://www.stat.wisc.edu/~larget/bucky.html.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. *Syst Biol* 54: 241-253.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. *Genome Res* 8: 1,233-1,244.

Liu L, Pearl DK. 2006. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Mathematical Biosciences Institute Tech. Report 53 [cited November 27, 2006 from the Internet]. Available from: http://mbi.osu.edu/publications/pub2006.html.

Mossel E, Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309: 2,207-2,209.

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

Penny D, Foulds LR, Hendy MD. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* 297: 197-200.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol* 21: 1,455-1,458.

Pollard D, Iyer VN, Moses AM, Eisen MB. 2006. Whole genome phylogeny of the *Drosophila melanogaster* species subgroup: widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet* 2(10): e173.

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

Suchard MA. 2005. Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics* 170: 419-431.

Suchard MA, Kitchen CMR, Sinsheimer JS, Weiss RE. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst Biol* 52: 649-664.

Wendel JF, Doyle JJ. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. *In*: Soltis DE, Soltis PS, Doyle JJ, editors. *Molecular Systematics of Plants II*. Boston: Kluwer Academic, p. 265-296.

Whelan S, Liò P, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet* 17: 262-272.