Syst. Biol. 58(5):489–500, 2009
© The Author(s) 2009. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org
DOI:10.1093/sysbio/syp054
Advance Access publication on September 21, 2009

Species Trees from Highly Incongruent Gene Trees in Rice

KAREN A. CRANSTON[1,2,*], BONNIE HURWITZ[1,3], DOREEN WARE[3,4], LINCOLN STEIN[3,5], AND
ROD A. WING[6,7,8]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA;*
[2]*Biodiversity Synthesis Center, Field Museum of Natural History, Chicago, IL 60605, USA;*
[3]*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA;*
[4]*Robert W. Holley Center for Agriculture and Health, United States Department of Agriculture-Agricultural Research Service, Ithaca, NY 14853, USA;*
[5]*Ontario Institute for Cancer Research, Ontario, Canada M5G 0A3;*
[6]*Arizona Genomics Institute and* [7]*Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA; and*
[8]*BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA;*
**Correspondence to be sent to: Biodiversity Synthesis Center, Field Museum of Natural History, 1400 South Lakeshore Drive, Chicago, IL 60605, USA;*
E-mail: kcranston@fieldmuseum.org.

Abstract.—Several methods have recently been developed to infer multilocus phylogenies by incorporating information from topological incongruence of the individual genes. In this study, we investigate 2 such methods, Bayesian concordance analysis and Bayesian estimation of species trees. Our test data are a collection of genes from cultivated rice (genus *Oryza*) and the most closely related wild species, generated using a high-throughput sequencing protocol and bioinformatics pipeline. Trees inferred from independent genes display levels of topological incongruence that far exceed that seen in previous data sets analyzed with these species tree methods. We identify differences in phylogenetic results between inference methods that incorporate gene tree incongruence. Finally, we discuss the challenges of scaling these analyses for data sets with thousands of gene trees and extensive levels of missing data. [Bayesian MCMC; gene tree incongruence; multilocus analysis; phylogenetic inference; rice.]

A species tree is the underlying framework used in many studies of comparative and evolutionary biology. In the absence of full genome sequences, trees built from 1 or more genes have been used as a proxy for the species trees. However, the knowledge that trees inferred from different genes can conflict with each other has fueled much debate on how and when data sets from different sources can be combined in a phylogenetic analysis (Doyle 1992; Bull et al. 1993; de Queiroz et al. 1995; Wiens 1998). Given the increasing availability of genome-scale data, phylogenetic analysis can now span hundreds or thousands of genes (e.g., Rokas et al. 2003; Pollard et al. 2006; Ebersberger et al. 2007). Studies at this scale have reinforced the finding that gene tree incongruence is commonplace in many taxa and widespread throughout the genome.

Gene trees may differ in their evolutionary history for a number of biological reasons. One is coalescent stochasticity, which can result in the persistence of lineages beyond the speciation boundary (Pamilo and Nei 1988; Takahata 1989). In this case, the order of lineage coalescence events (the gene tree) can differ from speciation events in the species tree. A second reason for incongruence is movement of genes between species, whether by hybridization and introgression (Rieseberg et al. 2000) or horizontal gene transfer (Doolittle 1999). A third is failure to correctly infer orthology due to gene duplication and subsequent gene loss or incomplete sampling of gene copies, so that paralogous gene copies are inferred to be orthologs (Page 1994; Page and Charleston 1997). The case of incomplete lineage sorting has been shown to be particularly challenging, as some combinations of tree shape and branch lengths in a species trees can cause the most biologically likely gene tree to differ from the species tree (Degnan and Rosenberg 2006).

The fact that trees built from different genes can vary has been recognized for some time. The first method for constructing species tree from incongruent gene trees was published nearly 30 years ago (gene tree parsimony; Goodman et al. 1979). Inference of species trees from gene trees then received relatively little attention except for important developments in the gene tree parsimony approach reconciling gene duplications and losses (Page and Charleston 1997; Page 1998). Reasons for this may include a dearth of data sets with large numbers of genes and also lack of computational power to implement complex high-dimensional models of gene duplication and coalescent stochasticity. The focus shifted to the development of sophisticated models of nucleotide substitution and application of these models to likelihood and Bayesian inference of individual data sets. Also important was the hope that concatenation of many genes would solve the problem; that with enough sequence data, a predominant signal would emerge and this signal would be equal to the species tree (the "supermatrix" approach; de Queiroz and Gatesy 2007). The landmark phylogenomics paper of (Rokas et al., 2003) described the inference of a fully resolved and perfectly supported phylogeny of yeast using a supermatrix approach, despite incongruence between the 106 individual genes trees. This was followed by many studies of the yeast alignment and other data sets, showing that the simple concatenation of genome-scale data could mislead the inference of species phylogenies due to the effects of so called "nonphylogenetic" signal (Phillips et al. 2004; Jeffroy et al. 2006) or presence of incomplete lineage sorting (Kubatko and Degnan 2007).

The increase in availability of genome-scale data for phylogenomics, combined with better methods and models for inferring gene trees, has triggered renewed interest in gene tree incongruence. Rather than approaching species tree inference as a problem of separating signal (a species tree) from noise (gene tree variability), the availability of many genes across many species permits the use of information contained in the distribution of gene trees as part of the inference procedure. Following this approach, promising methods for inferring species trees using the incongruence in gene trees have recently been developed (Arvestad et al. 2003; Maddison and Knowles 2006; Ané et al. 2007; Edwards et al. 2007; Liu and Pearl 2007).

In this study, we focus on 2 such methods, Bayesian concordance analysis (BCA), as implemented in the software BUCKy (Bayesian Untangling of Concorance Knots) (Larget 2006; Ané et al. 2007) and Bayesian estimation of species trees, implemented in the software BEST (Liu and Pearl 2007; Liu 2008). BCA infers concordance factors (CFs), defined as the proportion of genes, or of the whole genome, that agree with a given bipartition. The calculation of CFs is a 2-stage process. The first stage is independent Bayesian analyses of the gene trees to produce a posterior density of phylogenies for each gene. The second stage infers a mapping of genes to tree topologies using both the independently inferred posterior samples and a prior distribution that describes the expected discordance between the genes. The result from this second stage is a set of sample-wide and genome-wide CFs for all bipartitions in the gene trees as well as an overall concordance tree. The model does not assume any specific biological process underlying the gene tree incongruence.

In contrast, the BEST method estimates a species tree by assuming that all incongruence between the gene trees is due to coalescent stochasticity. A coalescent-based prior describes the relationship between the genes trees and the common underlying species tree. Rather than inferring gene trees independently and then inferring a species tree, the method infers the gene trees and species tree jointly, starting from the gene alignments. Models and model parameters can vary between genes similar to a partitioned Bayesian analysis (Nylander et al. 2004). Using a full coalescent approach means that BEST estimates not only a species tree topology but also branch lengths in terms of divergence time and population size.

Both these methods use a Bayesian Markov chain Monte Carlo (BMCMC) framework, inferring a species-level tree over a distribution of gene trees and model parameters. This framework incorporates uncertainty in the gene and species tree reconstruction, which allows for more information about gene tree incongruence than simple point estimates of the gene trees. The inference of species trees from gene trees adds many additional parameters when compared with independent gene tree analysis or analysis of a concatenated data set, which may subsequently increase the length of analysis and sophistication of algorithms required for convergence of the underlying MCMC chains.

We are most interested in the application of these and comparable species tree methods to large phylogenomic data sets expected to play an increasingly important role in phylogenetic research as low-cost genome-scale sequencing technologies come online. One such example is the deep bacterial artificial chromosome (BACt)-end sequence (BES) libraries of the *Oryza* Map Alignment Project (OMAP) (Wing et al. 2005; Kim et al. 2008). OMAP has constructed BES libraries and physical maps of 12 wild and 1 cultivated *Oryza* species. These sequences were obtained from a high-throughput and low-coverage sequencing protocol designed to develop genome resources for species closely related to a fully sequenced model organism, in this case cultivated rice, *Oryza sativa*. The volume of data generated by OMAP is quite large, amounting at present to more than 1.5 million sequences for potential phylogenetic analysis (see http://www.omap.org).

Our test data set is a collection of 1700 genes in rice (*Oryza*), curated from approximately 450,000 BESs. The key features are a large number of genes over a relatively small number of species, with a matrix that is incomplete—not all genes are present in all species. This type of phylogenomic data set shares properties with those compiled using expressed sequence tag libraries, which are already having an impact on phylogenetic research programs (Sanderson and McMahon 2007; Hartmann and Vision 2008; Kullberg et al. 2008).

Biological factors also make rice an interesting test case. Phylogenetic analysis between closely related species such as rice or *Drosophila* (Pollard et al. 2006) lie at the intersection of phylogenetic and population genetic theory, where the effects of coalescent stochasticity result in very high levels of gene tree incongruence (Pamilo and Nei 1988; Maddison and Knowles 2006). The oft-analyzed yeast species (Rokas et al. 2003) represent a fairly deep divergence (at least 250 million years; Douzery et al. 2004), whereas estimates for divergence of *Oryza* are 8–14 million years for the whole genus and 2 million years for the AA-genome clade studied here (Vaughan et al. 2008). The species in this study are all members of the closely related AA-genome group in *Oryza* and their history includes 2 domestication events as well as evidence of hybridization/introgression (Sang and Ge 2007). Gene flow in rice has been detected in both wild and domesticated species (Semon et al. 2005; Caicedo et al. 2007; Zhou et al. 2008). The presence of gene flow may complicate the phylogenetic analysis, causing gene tree incongruence in some regions of the genome that cannot be explained by coalescent-based models.

A key question for analysis of these types of sequences is whether the BMCMC methods for species tree inference can scale to hundreds or thousands of gene trees across a range of phylogenetic depths. The largest data set yet tested with these species tree methods is yeast, with 106 loci in 8 species (Ané et al. 2007; Edwards et al. 2007). Phylogenetic analysis at lower taxonomic levels, where incongruence due to coalescent stochasticity is expected to be greater, have had far fewer loci, for example, 30 loci in 4 species of finches (Liu and Pearl 2007), 4 loci in 19 species of macaques (Liu and Pearl 2007), 7 loci in 8 species of gophers (Belfiore et al. 2008), 3 loci in 19 species (1–4 individuals per species) of sawflies (Linnen and Farrell 2008), and 5 loci in 5 species (20 individuals) of Mannakins (Brumfield et al. 2008). In this analysis, we investigate the phylogenetic results from these methods as well as their computational efficiency using a large number of loci from a set of very closely related species.

METHODS

Species Included in the Study

This study includes 6 diploid species of rice, genus *Oryza*. We used the cultivated rice species *O. sativa* ssp. *japonica* and *Oryza glaberrima* as well as the wild species *Oryza rufipogon*, *Oryza nivara*, and *Oryza barthii* and, as an outgroup, the wild species *Oryza punctata*. See Table 1 for genome types and geographic location.

Curation of Gene Alignments

We obtained a starting pool of BESs for 5 *Oryza* species from the GenBank GSS Division and sequences from *O. barthii* not yet submitted to Genbank directly from the OMAP group. We then aligned these BES to the coding sequences (CDS) from the *O. sativa* IRGSP V3 gene models using *blastall* (Altschul et al. 1990) (options: -p BLASTn, -w 7, -e 1e-10). We aligned overlapping BES from the same species using ClustalW (Thompson et al. 1994), with default parameters, to create a consensus exon sequence. To remove paralogous sequences, we performed an all-versus-all BLAST search of the consensus sequences for each of the wild rice species and the CDS sequences from *O. sativa*. If an exon had multiple hits to its own genome or that of another species, we removed that exon from the analysis. We then filtered these remaining single-copy exons to ensure 70% coverage between sequences and assembled them into clusters of orthologs via a single linkage clustering algorithm (where inclusion of a gene into a cluster requires only a single link via BLAST to any other gene), keeping only clusters with sequences from all 5 AA-genome species plus the outgroup.

We then aligned orthologous exons with the gene sequence from *O. sativa* using T-Coffee (Notredame et al. 2000); options: matrix = BLOSUM, ktuple = 2, tg_mode = 0, gapopen = −10. A small percentage of these alignments were found to be misaligned and to contain large gaps in the alignment. We filtered the data a second time to keep only the sequences in the alignment that had both 70% identity with the *O. sativa* gene sequence and 70% coverage. Visual inspection of a randomly selected subset of genes indicated that the alignments were of high quality, which was expected given the low sequence divergence. We also examined the alignments using GBlocks (Castresana 2000), which is designed to identify regions of poor alignment quality.

TABLE 1. The diploid rice species used in this study, along with geographic location and genome type. The 2 species in bold are the cultivated rice species

Species	Genome type	Geographic location
Oryza sativa	**AA**	**Southeast Asia**
Oryza rufipogon	AA	Southeast Asia
Oryza nivara	AA	Southeast Asia
Oryza glaberrima	**AA**	**West Africa**
Oryza barthii	AA	Africa
Oryza punctata	BB	Africa

Model Selection

For all analysis described, we employed a HKY+Γ model of sequence evolution, using empirical base frequencies and an estimated alpha parameter for a Γ distribution with 10 categories. Although there was more than enough data in the concatenated alignment to use a more highly parameterized model, we chose HKY based on results from ModelTest (Posada and Crandall 1998) for a random sample of the gene-level alignments as well as results from phylogenetic analysis of the larger alignment. This kept the model consistent across the analysis of the large concatenated alignment and the independent genes as well as the joint inference of gene and species trees and also helped to reduce the number of parameters for the complex coalescent-based approach.

Concatenated Analysis

For analysis of the full data set, we concatenated the final set of genes with data for all 6 species into a single alignment. We then inferred a species tree using a BMCMC approach as implemented in MrBayes v. 3.1.2 (Huelsenbeck and Ronquist 2001; nruns = 3, ngen = 5,000,000, samplefreq = 500). We repeated the analysis twice, once with a single partition across the genes and once splitting the alignments into a separate partition for each gene, allowing for different values of the model parameters across the genes. To ensure that the analyses had converged to the stationary distribution, we examined log likelihood plots for stationarity, ensured that potential scale reduction factor (PSRF) values for tree length, transition–transversion ratio (kappa), and alpha parameter for the rates-across-sites model were less than the generally accepted value of 1.20. We also calculated effective sample sizes for these parameters using Tracer (Rambaut and Drummond 2005) and then finally examined the mean standard deviation of the partition probabilities across the chains.

Before beginning the analyses, we recompiled MrBayes with the following changes to the file *mb.h* to increase allowable number of character sets (the upper limit before these changes was 150):

```
#define MAX_NUM_DIVS 310
#define MAX_NUM_CHARSETS 310
```

```
#define MAX_NUM_TREES 310
#define MAX_NUM_DIV_BITS 310.
```

Independent Bayesian Analysis

As a first stage for the BCA analysis, and to quantitate the overall level of incongruence between the genes, we inferred a posterior distribution for each gene tree independently using MrBayes. These analyses consisted of 3 separate MCMC chains per gene alignment, each of 2 million iterations. We used the same measures of convergence as for the concatenated analysis, although we limited graphical examination of log likelihood plots and effective sample sizes to a random sample of the total number of genes.

We summarized the independent gene tree analyses using the posterior distributions and point estimates for the trees. We also used the genealogical sorting index, *gsi* (Cummings et al. 2008), to summarize the incongruence between independently inferred gene trees. The *gsi* index provides a measure of the relative degree of exclusive ancestry of a group of species. It can be calculated for a single tree or a collection of trees (the ensemble *gsi*, or *egsi*) and statistical significance assessed using a permutation test. Using the implementation in R (R Development Core Team 2009), we calculated the *egsi* across the set of maximum a posteriori (MAP) gene trees for the Asian group (*O. sativa*, *O. nivara*, *O. rufipogon*) and the African group (*O. glaberrima, O. barthii*), with 5000 replications for the permutation test.

Bayesian Concordance Analysis

We performed BCA using the program BUCKy (Ané et al. 2007). Running BUCKy is a 2-step process. First, we summarized the MrBayes results for the independently inferred gene trees using the *mbsum* command included with BUCKy, combining the 3 MCMC chains into a single output file after removing 500 samples from each chain as burn-in. We then performed the BCA using the following BUCKy command: *bucky -n 1,000,000 -a 0.1*. The single adjustable parameter in BUCKy is α, which controls the a priori level of gene tree incongruence. We tested wide range of this parameter using α values of 0.01, 0.1, 0.5, 1.0, 5.0, 10.0, and 100.0. A larger value of α corresponds to greater gene tree incongruence—the probability that 2 genes share the same topology is approximately $1/(1 + \alpha)$. We also tested robustness of the method to inclusion of less informative genes using both all the genes with 6 species and also a smaller subset of "informative" genes (those that did produce a flat posterior distribution of topologies and the star topology as the majority rile consensus tree).

BEST Analysis

Finally, we jointly inferred posterior densities of the gene trees and species tree using BEST 2.2 (Edwards et al. 2007; Liu and Pearl 2007). Version 2.2 did not require the changes to the file *mb.h* as described for MrBayes, above, but we did need to compile as 32 bit (-m32 flag with the GCC compiler) to run BEST on a 64-bit version on Linux. Due to memory allocation errors in our initial tests with all 6-species genes and concerns about slow convergence, we performed the analyses with the smaller set of informative genes described above for BCA. We ran 3 independent analyses of 1.6 billion iterations, 2 MCMCMC chains and sampling every 50,000 iterations.

We also examined the performance of BEST with subsets of the total number of genes. We randomly selected sets of 10, 20, 30, and 40 genes, with 10 replicates each, for 40 subsampled data sets in total. Finally, we concatenated the longest gene alignments (the genes with more than 1000 nucleotides each). For all subsampled analysis, model parameters were the same as the 162-gene data set, but we ran the chains for 200 million iterations, sampling every 10,000.

Computer Hardware

For the concatenated and independent Bayesian analyses as well as BEST analyses, we used a Linux computer cluster comprised 12 servers and 96 cores with processor speeds ranging from 2.33 to 3.16 GHz and 8–16 GB of memory per server. We performed the *gsi* and second phase of the BCA analyses on a MacPro quad-core desktop machine with CPU speed of 2.66 GHz and 3 GB memory.

Results

Curation of Gene Tree Alignments

We started with 310,538 BESs over all 6 *Oryza* species. The mean read length of these sequences is 644 bp and the minimum and maximum length is 101 and 1012, respectively. BLASTing these sequences against the IRGSP V3 gene models yielded 31.9% coverage of the total CDS sequence from *O. sativa* (44,492,676 bp), representing 22,544 of the 37,544 total genes in rice. Creating a consensus sequence of 2 or more overlapping BES gene hits from within a single species generated a pool of exons across the 6 species having 8.9% exon coverage to *O. rufipogon*, 12% to *O. nivara*, 8.7% to *O. glaberrima*, and 7.8% to *O. punctata* representing 18 094 total genes. The all-against-all BLAST procedure aimed at excluding duplications reduced the pool of exon sequences to orthologs representing 7306 genes of *O. sativa*. Finally, after removing any clusters with fewer than 4 sequences, we were left with 1720 genes. All 6 species are present in 307 of these alignment, 5 species in 546 alignments, and 4 in the final 867 alignments. The alignments range in length from 99 to 2076 with a mean length of 350 nucleotides. There were 26 genes with more than 1000 nucleotides. The concatenated alignment of the 307 six-species genes contains 136,684 nucleotides, and the smaller set of 162 informative genes contained 87,619 nucleotides (see results from independent Bayesian analysis, below).

Visual inspection of a randomly selected subset of genes indicated that the alignments were of high quality, which was expected given the low sequence divergence. Examination of the concatenated alignments using GBlocks did highlight some large blocks for removal, but further investigation revealed that the basis for removal was generally a high amount of missing data in a given region rather than poor alignment per se. Therefore, we used the T-Coffee alignments without additional editing for subsequent phylogenetic analysis. (See supplementary material online for Nexus files containing these sequence data, Supplementary Material Online).

Concatenated Bayesian Analysis

The BMCMC analyses of the large concatenated alignment converged very quickly to a stationary distribution. For both the partitioned and the nonpartitioned analyses, the estimated burn-in based on log likelihood plots was 1000 samples (100,000 generations) per chain, leaving 9000 samples per chain (18,000 total) for inference. PSRF values of all model parameters were less than 1.20. Effective sample sizes of all parameters were at least 300, and most were greater than 1000.

The majority rule consensus tree topology, shown in the top left of Figure 1, was the same from both the partitioned and the nonpartitioned analysis and was also the MAP tree for both analyses. The posterior probability of this topology was 1.00 in the partitioned analysis and 0.914 in the nonpartitioned analysis. The only clade with posterior probability less than 1.00 was the (*O. sativa, O. nivara, O. rufipogon*) group in the nonpartitioned analysis, which had a probability of 0.914.

Independent Bayesian Analysis

The independent analyses converged very quickly. Examination of the log likelihood plots for a sample of 30 genes indicated that the chains had reached stationarity within the first few thousand iterations (data not shown). Based on these tests, we removed 500 samples (50,000 MCMC iterations) as burn-in for all genes before summarizing trees or model parameters. All PSRF values less than 1.20 for tree length, alpha and kappa for all genes. The mean standard deviation of the partition probabilities was less that 0.01 in all runs but 3. The 3 genes with larger values all had very little probability on any 1 tree, and the posterior sample of trees contained all possible topologies. The low phylogenetic signal may have been a cause of the slow convergence (although other genes that returned a similar result

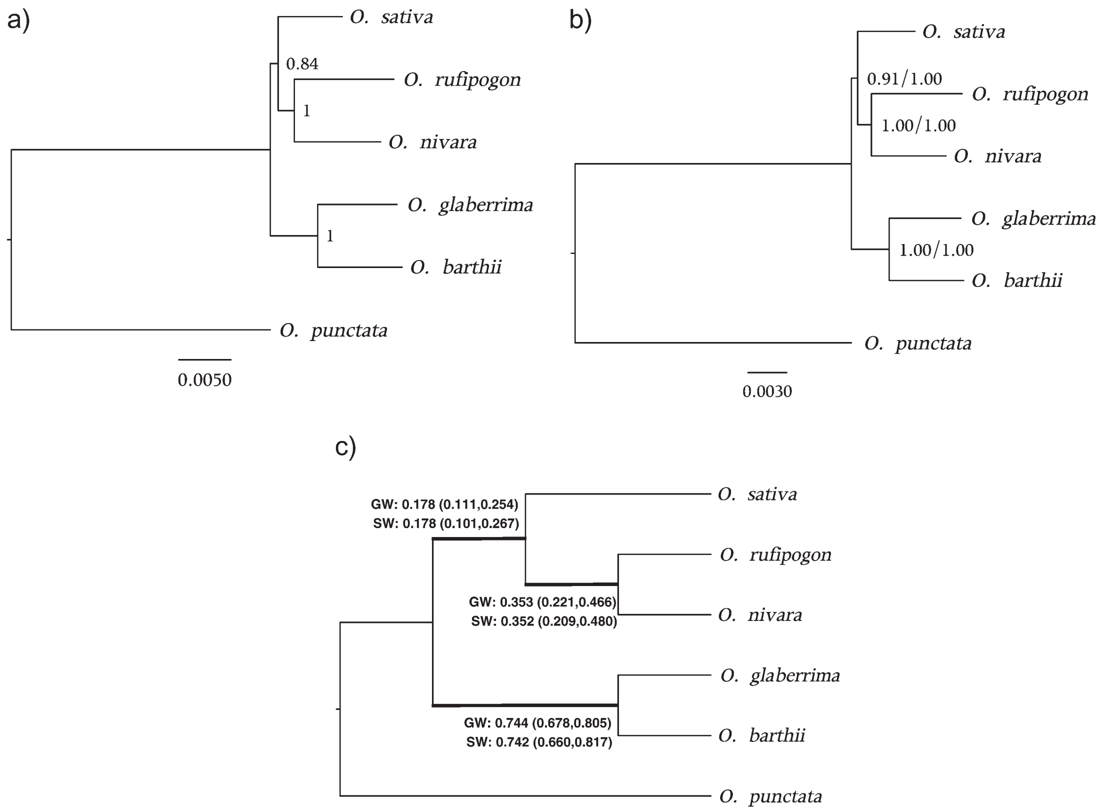

FIGURE 1. Species trees from the different analysis: a) MAP tree from Bayesian analysis of the concatenated data set. Node labels are posterior probabilities from analysis of the data set with a single partition (first number) and gene-by-gene partitioning (second number). Branch lengths are based on the nonpartitioned analysis and the tree is rooted using midpoint rooting for display purposes. b) MAP tree from BEST analysis of a smaller data set consisting of the longest gene alignments. Node labels are posterior probabilities. c) Primary concordance tree (cladogram) from the BCA of all genes, showing genome-wide (GW) and sample-wide (SW) CFs along with the 95% HPD interval for the highlighted branches.

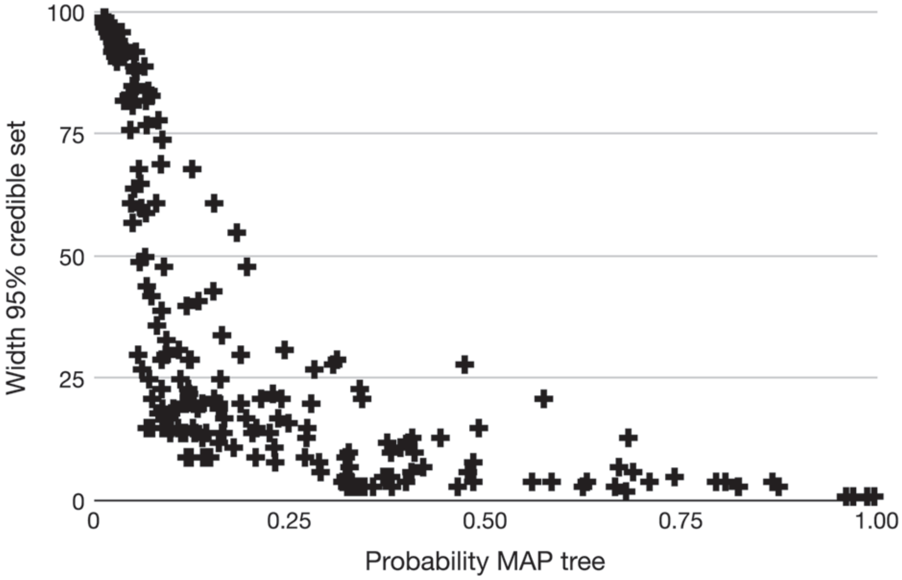

FIGURE 2. Plot of probability of the MAP tree against the width of the credible set in the independent Bayesian analyses. The width of the credible set is the number of topologies that comprise 95% of the posterior density. The posterior densities of gene tree range from very narrow (most or all probability on a single tree) through very wide (nearly equal posterior probability on each of the possible topologies). Each point represents 1 independent gene alignment.

with respect to the distribution of trees displayed faster convergence).

Overall, the width of the 95% highest posterior density (HPD) intervals (the number of trees in the posterior sample that comprise 95% of the total probability) ranged from 1, indicating a single topology with posterior probability greater than 0.95, to 99. Given 105 possible rooted topologies for the 5 ingroup species, an HPD of 99 means that nearly all possible gene tree topologies are represented in the posterior sample. See Figure 2 for a graphical representation of the posterior samples of trees. The topology returned in the concatenated Bayesian analysis is the MAP tree in only 8 of the 307 genes, with a mean posterior probability of 0.812. We contrast this result to the yeast analysis, where 44 of the 106 genes return the supermatrix topology as the most probable tree (Ané et al. 2007).

Using the *egsi* index, we estimated the relative degree of exclusive ancestry for the Asian species group (*O. sativa, O. nivara, O. rufipogon*) to be 0.300 (*P* value of 0.103). For the African species group (*O. glaberrima, O. barthii*), the *egsi* value was 0.581 with a *P* value of 0.071. The *egsi* is scaled between 0 and 1, with a value of 1 indicating a monophyletic group across all the input trees. The results for the rice gene trees indicate that neither the African nor the Asian group is supported across all the gene trees, and the index for the Asian group is not greater than what might be expected from a randomly generated set of gene trees.

Bayesian Concordance Analysis

The primary concordance tree, estimated using BCA and the independent posterior distributions of genes trees, is shown in Figure 1. It is the same topology as the concatenated Bayesian analysis. Note that the CFs in the BCA tree are much more conservative than the posterior probabilities in the topology estimated from the concatenated alignment (top right in Fig. 1). There was good agreement between the calculated sample-wide CFs and the extrapolated genome-wide CFs, which is not unexpected given that our data set is a large random sample of genes distributed across the whole rice genome. To determine the significance of the CFs in the primary concordance tree, we examined the 95% HPD intervals for clades that conflicted with those in the primary concordance topology (see Table 2). The (*O. barthii, O. glaberrima*) clade is the only CF in the primary concordance topology where the 95% HPD interval does not overlap with a conflicting clade.

The BCA method is robust to the prior probability on gene tree incongruence (the α parameter in BUCKy). See Figure 3 for a comparison of the alpha value on the 3 clades in the primary concordance topology. All the HPD intervals for a given clade overlap for all α values. We see the greatest effect of the prior distribution on the clade with the highest posterior probability in the concatenated analysis: (*O. glaberrima, O. barthii*). In this case, CFs decrease with decreasing α. In contrast, there is no effect on the clade including (*O. sativa, O. nivara, O. rufipogon*), which had lower probability in the concatenated analysis.

The BCA method is also robust to the inclusion of all 307 six-species gene trees, which includes alignments with very little phylogenetic signal. The difference in mean CFs between the data set with 162 alignments and the data set with all 307 alignments is less than 0.06 for all clades, and the difference in HPD interval width is

TABLE 2. Checking for overlapping 95% HPD intervals for sample-wide CFs of conflicting clades. Each group of 2 clades represents the primary concordance tree clade (in bold) and the conflicting clade with the highest CF

Clade	Mean CF	95% HPD interval	Overlaps?
(*Oryza glaberrima, Oryza barthii*)	0.707	(0.648, 0.759)	
(*Oryza nivara, O. glaberrima*)	0.071	(0.043, 0.111)	No
(*Oryza rufipogon, O. nivara*)	0.353	(0.235, 0.451)	
(*O. rufipogon, Oryza sativa*)	0.194	(0.105, 0.315)	Yes
(*O. sativa, O. rufipogon, O. nivara*)	0.202	(0.136, 0.272)	
(*O. sativa, O. nivara, O. barthii, O. glaberrima*)	0.238	(0.170, 0.302)	Yes

less than 0.05. Therefore, the inclusion of the uninformative alignments does not significantly change either the mean CF or its estimated level of uncertainty.

BEST Analysis

The data set of 162 genes did not converge in 1.6 billion iterations. Parameter values (log likelihood and the parameter *LnJointGenePr* for the species tree and tree lengths for a random sample of gene trees) differed greatly across 3 independent runs and all were unstable, continuing to increase up to the end of the analysis (see Fig. 4). For this reason, we do not report any phylogenetic results from analysis of the full data set with BEST. We were successful with some of the BEST analysis of the smaller data sets, although 2 of the 20-gene data sets, 2 of the 3-gene, and 5 of the 40-gene data sets did not converge after 200 million iterations. For those that did reach stationarity, we removed between 2000 and 8000 of the 20,000 total samples as burn-in.

For the set of longest genes (26 genes with more than 1000 nucleotides), the MAP tree was the same topology returned by the concatenated analysis and by BCA analysis (see Fig. 1). Probability of this topology was 0.509, which means that the MAP tree is equivalent to the majority rule consensus tree. The 95% credible set contained 12 trees, and the second and third most probable trees were those with alternate resolutions of the 3 Asian species.

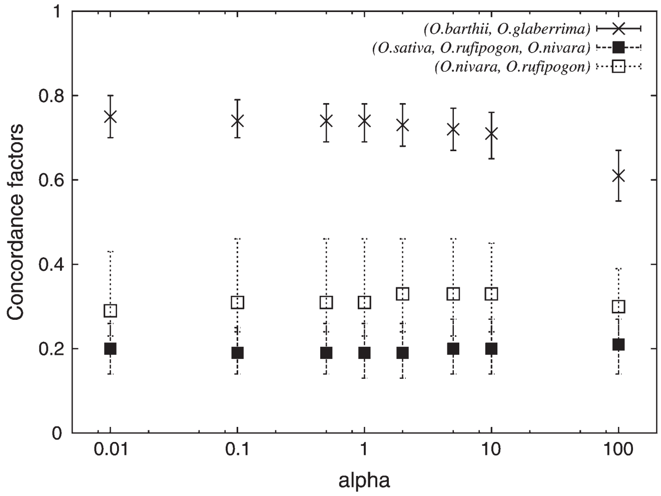

FIGURE 3. BCA: effect of changing the prior on gene tree incongruence (α) on the CFs for the 3 highlighted clades in the primary concordance tree shown in Figure 1. Error bars are the width of the 95% HPD interval.

This same topology was also the most frequent MAP tree in the subsampled genes (11 of 31 sets of genes). The probabilities of these MAP trees ranged from a high of 0.833 to a low of 0.217, with a mean posterior probability of 0.44 over all 32 of the subsampled data sets. However, the remaining analyses support alternate topologies. One of the other striking results is the number of maximally asymmetric topologies sampled by BEST, a tree shape that did not appear in concatenated Bayesian analysis or in BCA analysis. For example, the MAP tree is a maximally asymmetric topology in 10 of the subsampled gene analyses. See Figure 5 for results over all analyses. The use of more genes (or having more nucleotides in total) did not increase the frequency with which we observed any particular topology, meaning that conflicting phylogenetic signal was not correlated with fewer numbers of genes or nucleotides. Overall, most analyses support the African clade (*O. glaberrima, O. barthii*), with only 3 analyses giving less than 0.95% posterior probability to this clade in the species tree. Overall, subsampling 10, 20, 30, and 40 sets of rice genes does not give a stable result for the species tree under the BEST model.

Computational Efficiency

The various phylogenetic inference methods vary considerably in terms of running time. Concatenated Bayesian analysis required 5–8 h for 5,000,000 generations and 2 MCMC chains in MrBayes. Independent Bayesian analysis, which is also step 1 of the BCA method, required approximately 1 h per gene tree (2 million generations; 3 Metropolis-coupled MCMC chains in MrBayes), and this step was trivially parallelized on our computer cluster so that total elapsed time for all genes was less than 2 h. The second step of BCA analysis required less than 5 min for calculation of CFs across the posterior distributions of gene trees. In contrast, 1.6 billion iterations of BEST analysis for the concatenated 162 gene data set required nearly 2 months for each MCMC chain, for an analysis that did not converge over that period of time. Therefore, joint inference of gene trees and species trees greatly increases the computational burden over concatenated, independent or BCAs. The increased analysis time is, of course, due to the increased complexity of the BEST procedure, and the tradeoff is the that this method more explicitly models gene tree incongruence, providing estimates of other parameters in addition to topology.

DISCUSSION

We are seeing a move to phylogenetic analysis based on the information contained in gene tree incongruence

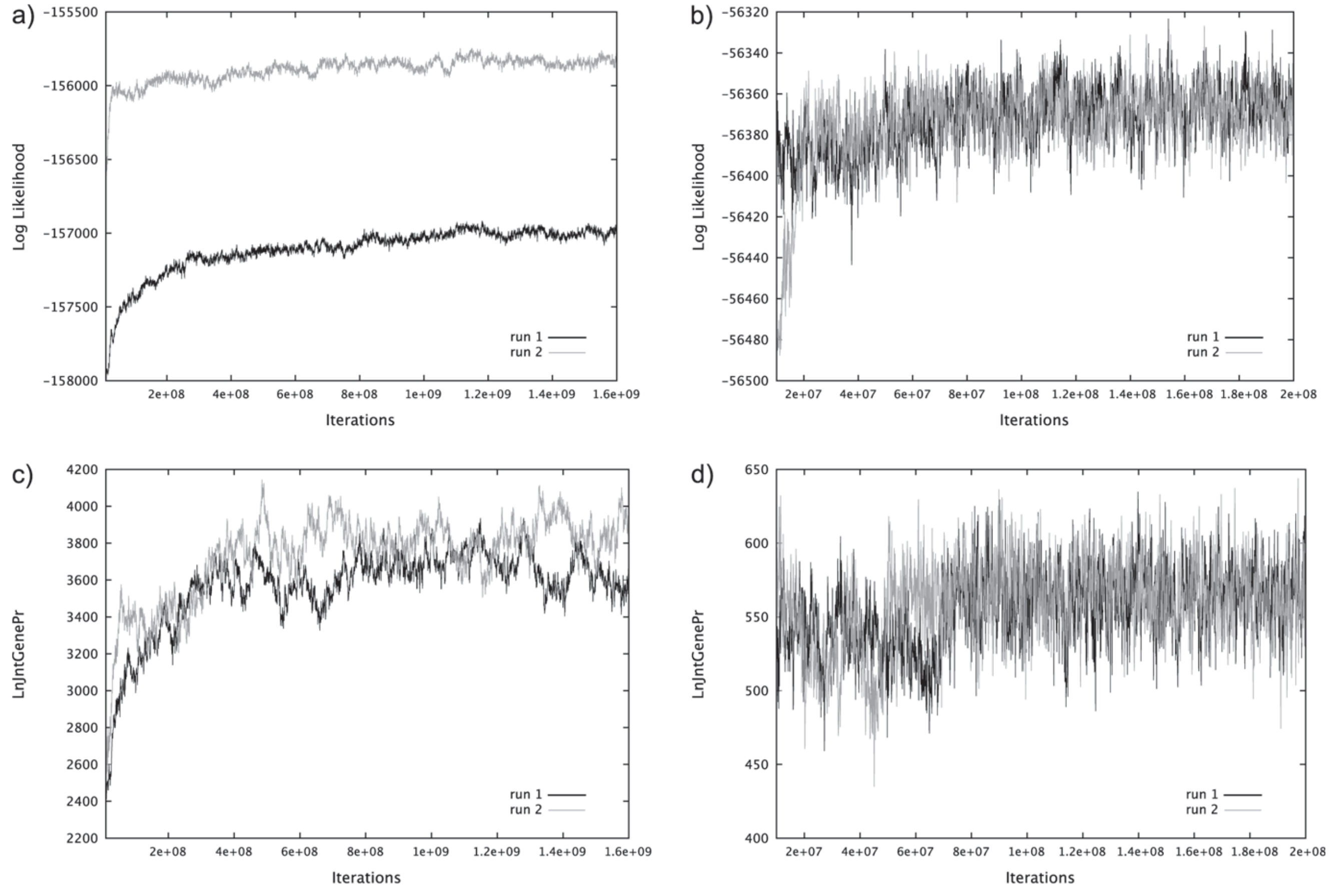

FIGURE 4. MCMC convergence in BEST: a) log likelihood, 162 genes, 1.6×10^9 iterations; b) log likelihood, 26 longest genes, 2×10^8 iterations; c) LnJointGenePr, 162 genes, 1.6×10^9 iterations; d) LnJointGenePr, 26 longest genes, 2×10^8 iterations. To show more detail in the latter part of the chains, the plots do not include the first 10,000,000 iterations. Note the differences in scale of the axes between the plots on the left and those on the right.

due to the availability of many genes for many species and simultaneous development of novel inference methods. We have examined the performance of 2 of these methods, BCA and BEST, using the largest number of genes that has been applied to either method. These data from rice also provide an example of species tree inference for closely related species with a high level of gene tree incongruence.

In terms of rice phylogeny, the analyses described here generally agree with the current hypothesis about the relationship between the *Oryza* species (Zhu and Ge 2005; Duan et al. 2007; Zou et al. 2008). The 2 African species, *O. barthii* and *O. glaberrima*, form a clade in nearly all the species trees, and the Asian species, *O. sativa*, *O. rufipogon*, and *O. nivara*, are supported by some, but not all, of the inference methods. Taking into account the incongruence between gene trees does not drastically change our overall view of rice phylogeny, but it does give a more varied picture of the support across the tree. The African clade is much more strongly supported across the analyses than the Asian clade, and some of the analyses do not support grouping of the 3 Asian species over other phylogenetic resolutions. In general, concatenation gives nearly perfect support for all clades across the *Oryza* phylogeny, whereas methods that explicitly consider gene tree incongruence give support values that are lower and more consistent with what we know about the biology of these species.

Although each of the methods used in this study is an example of Bayesian inference of and therefore incorporates uncertainty and uses posterior probabilities in the output, the underlying models and definition of the output differs greatly. MrBayes and BEST both output posterior probabilities of whole topologies and of bipartitions in the tree. These are the probabilities that tree or clade is true under the specified model, but that model differs greatly between the 2 programs. Concatenated analysis in MrBayes assumes a single shared topology underlying all genes, whereas BEST uses joint inference of a species tree and gene trees based on a coalescent explanation for incongruence between gene trees. In contrast, BCA infers a completely different measure—CFs, or the proportion of the genome that supports a bipartition—calculated using the posterior probabilities of gene-to-tree maps. It is important to consider these differences between the models and

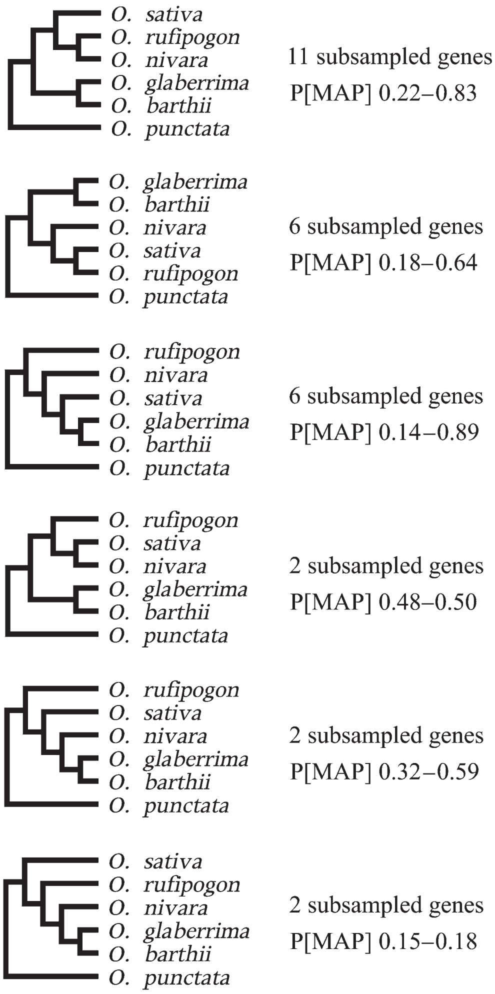

FIGURE 5. Most frequent MAP trees and the range of probability values for the analysis of subsampled sets of genes in BEST. The 2 African species, *Oryza glaberrima* and *Oryza barthii*, form a clade in each one, whereas relationships between the Asian species are less consistent across the trees.

output measures when we discuss results across inference methods.

Bayesian Concordance Analysis

The BCA worked well for the rice data, providing useful estimates for the proportion of the genome that supports various bipartitions for the *Oryza* species. The method returns the maximum concordance topology, which represents the collection of nonconflicting clades with the highest CFs. This topology is the same as the tree inferred by supermatrix approaches based on Bayesian, maximum likelihood and maximum parsimony frameworks (data not shown). CFs on clades in the species-level tree are much lower than the posterior probabilities or bootstrap proportions seen in traditional analyses, illustrating the difference between these measures. Only 1 partition in the tree, that of the African species *O. glaberrima* and *O. barthii*, returns a CF that is significantly higher than the estimate for any contradicting clade. These results are not surprising, given the extent of gene tree incongruence, and we feel that the CFs give an useful perspective for genome-wide phylogenetic analyses. The method is robust to changes in the parameter of the prior distribution and to inclusion of less phylogenetically informative genes. As implemented in the software BUCKy, concordance analysis is extremely fast, even when factoring in Bayesian inference of the independent gene trees as a first step.

For this type of data set, with very high gene tree incongruence, the CFs and the primary concordance tree from BCA give a more intuitive estimate of bipartition support across a species trees than does a supermatrix approach. Using the latter method, the posterior probabilities estimated for partitions on the species tree do not reflect the level of incongruence between the gene trees due to the assumption of a single underlying species tree. Although posterior probabilities have been shown to be an accurate measure of support with data simulated on a single topology (Huelsenbeck and Rannala 2004), there are known problem with the traditional BMCMC methods when the data actually describe a mixture of topologies (Mossel and Vigoda 2005). One disadvantage of the BCA method is that each gene is treated equally, even if there is great heterogeneity in number of nucleotides across genes. Use of the posterior sample of trees as input, rather than a point estimate of each gene tree, in advantageous in cases where the strength of the phylogenetic signal varies across genes but does not directly address gene length. It is also unknown how BCA analysis might be affected by the anomaly zone—combinations of topology and branch lengths in the species tree that cause an anomalous gene tree to be most likely. We know that maximum likelihood or maximum parsimony inference of concatenated sequences can be inconsistent under these conditions (Kubatko and Degnan 2007). If the independent Bayesian analysis of gene trees also suffers from this difficulty, then the BCA analysis of the posterior samples would also be influenced, although the effect remains to be tested.

Bayesian Estimation of Species Trees

Analysis of this rice data with the coalescent-based method BEST posed some methodological challenges. The BMCMC analysis of the 162 gene data was far from stationarity after nearly 2 months of analysis time. Does this mean that BEST is not immediately ready for analysis of this scale? We suggest that it is not the scale of the data that is the problem. The yeast data set is similar in size, with 106 genes over 8 species, when compared with our collection of 162 genes over 6 species. In the analysis of the yeast genes, 80 million MCMC iterations was sufficient (Edwards et al. 2007), whereas our analysis of the rice genes had not yet reached stationarity after

1.6 billion iterations. The depth of the phylogeny and the resulting extent of gene tree incongruence differs greatly between the rice and yeast data, and the increased complexity, combined with lower levels of sequence divergence, is likely a contributing factor to the slow convergence. Future work with rice, or other similar low-level phylogenetic analyses, will include investigating ways to ensure convergence. A simplistic approach is to simply run longer MCMC chains, but we may be able to increase the rate of convergence by thoroughly exploring Metropolis-coupled MCMC parameters, using different proposal mechanisms, especially for trees (Lakner et al. 2008), or perhaps inferring starting parameters for the individual genes before beginning the joint analysis.

A second reason for the difficulty with the coalescent approach is that regions of the rice genome almost certainly violate the assumption that all incongruence is due to coalescent stochasticity. Given the overlapping ranges and interspecies fertility in these species (Lu et al. 2000), we expect both incomplete lineage sorting and hybridization/introgression to play a significant role in the evolution of rice. BEST models all incongruence between gene trees as being due to coalescent stochasticity and assumes that gene flow is not present. In the face of gene flow, increasing the rate of MCMC convergence will not solve the underlying problem of incorrectly fitting a coalescent model. Some studies of coalescent-based species tree methods have found them to be relatively robust in the presence of gene flow (Eckert and Carstens 2008), but the BEST method has not yet been thoroughly tested in this way.

The full data set may contain regions of the rice genome where phylogenetic incongruence cannot be explained using coalescent-based analyses. Our experiments with smaller subsamples of genes were more successful in terms of MCMC convergence, although when examining the phylogenetic results, we see different MAP topologies from different subsamples of genes. The set of longest genes returns the same topology as we see in the concatenated and BUCKy analyses, as do many, but not all, of the other subsampled gene sets. The presence of other topologies may indicate that our subsamples do not contain enough information to accurately infer the species tree, that some of these random samples describe a different evolutionary history, or that the presence of gene flow in some data sets is affecting the inference of the species tree. We contrast this result to the yeast data set, where 8 randomly sampled genes were sufficient to recover the same topology inferred using the full data set of 106 genes (Edwards et al. 2007). Population genetic studies in rice have started to quantify gene flow in *Oryza* (Semon et al. 2005; Caicedo et al. 2007; Zhou et al. 2008). Additional types of population genetic and phylogenetic analyses with larger number of genes and/or multiple individuals per species will be able to precisely identify signatures of gene flow within the rice genome and allow us to explore these hypotheses about the results from BEST as compared with BCA or traditional single-tree Bayesian analysis.

Inferring Species Trees from Gene Trees with Genome-Scale Data

This study also highlights the differences between traditional gene-by-gene sequencing strategies, which produce full sequences from a collection of genes selected a priori, and high-throughput sequencing approaches. The latter method provides a large set of randomly sampled genes across the genome, a selection process that is ideal for these phylogenetic methods that infer species trees from gene trees. However, the sequencing strategy also means that many of the "genes" may be only partial gene fragments, as is the case with this rice data. We have compiled a large number of sequences with great variation in sequence length and high levels of missing data. Concatenated, this results in a huge number of nucleotides for supermatrix approaches, but the structure of the data is more limiting for a gene-by-gene method of analysis.

In addition to the sequences analyzed in this study, our bioinformatics pipeline identified an additional 1500 gene tree alignments that contain 4 and 5 species. We are able to analyze these genes with supermatrix techniques or by independently inferring trees on a gene-by-gene basis. Currently, neither BEST or BUCKy can utilize alignments or gene trees with different numbers of species. A future challenge is to incorporate alignments with missing data into the analyses that combine gene trees to infer species tree. The authors of BUCKy discuss the issue of missing data (Ané et al. 2007), suggesting that one possible solution is to add the missing species in all possible places on the input trees that lack those species in the posterior sample. This would likely increase analysis time, but only for the second phase of the BCA analysis, and given the speed of the method, this may not be a critical issue. Another option would be to add missing taxa to the alignments, so that some taxa only contain missing ('?') characters for some genes. This would almost certainly decrease the rate of MCMC convergence, adding analysis time to both phases of the BCA analysis. The issue of missing data in BEST is more complex. Adding missing taxa to the input alignments would certainly increase the difficulty of an already challenging MCMC convergence problem. It is also unknown how the coalescent prior that relates species trees and gene trees could be modified to allow for missing data. Ideally, we would like to infer each gene tree using only the available data but allow each gene tree to influence the species tree, even if the species tree contains a larger number of taxa. In general, it becomes difficult to interpret support for a particular clade under any joint method of analysis of species and gene trees. For example, high support for the clade (A,C) in a given gene tree does not negate the relationship ((A,B),C) in the species tree if B is not present in the alignment for the gene.

To maximize the information content in a data set such as the one we have presented for rice, novel methods for species tree inference must be able to deal with large amounts of missing data, particularly the case where gene alignments are missing data for an entire species. The whole genomes of *Drosophila* provide a complete set of genes across a dozen species, whereas other large phylogenomic data sets such as those for insects (Savard et al. 2006) or mammals (Philippe et al. 2004) have large number of loci over large number of species, but the data sets are much more sparse, with many loci missing for many species. In yeast, a small number of genes was sufficient to resolve the species tree, depending on the method used (Rokas et al. 2003; Edwards et al. 2007). More closely related species, with higher levels of gene tree incongruence, will likely require a larger number of genes. It is unlikely that accurate and precise inference of species trees from gene trees will require thousands of genes, but this necessitates some means of selecting genes for inclusion in the analysis. Other types of inference, such as quantifying gene tree incongruence, or differentiating between incomplete lineage sorting and hybridization in the genome, may be best done by maximizing the number of available genes.

Large phylogenomic data sets have highlighted that gene tree incongruence is widespread across genomes and across species in the Tree of Life. The BCA and BEST methods tested here are a very promising start to inferring species trees using the differences between the evolutionary histories of genes. In addition to providing an estimate of the species tree and various types of support for bipartitions across the taxa, these methods will be invaluable for identifying the specific evolutionary processes underlying the history of related species.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

B.H., L.S. and D.W. gratefully acknowledge support from the National Science Foundation (NSF) Division of Biological Sciences grant 0321678 and D.W. acknowledges support from United States Department of Agriculture - Agricultural Research Service. K.A.C. was supported by the NSF grant 0733365 to Michael Sanderson. This work was also supported by NSF grant DBI-0321678 to R.A.W. and L.S. and by the Bud Antle Endowed Chair (to R.A.W.).

Acknowledgements

First, we thank Scott Edwards and L. Lacey Knowles for organizing the Species Trees from Gene Trees symposium at the 2008 Evolution meeting. We also thank David Baum and Cecile Ané for discussions about BUCKy. This manuscript was greatly improved through comments by Scott Edwards, Jack Sullivan, Liang Liu and an anonymous reviewer.

References

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

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

Arvestad L., Berglund A.-C., Lagergren J., Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics. 19:i7–i15.

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

Brumfield R.T., Liu L., Lum D.E., Edwards S.V. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, Manacus) from multilocus sequence data. Syst. Biol. 57:719–731.

Bull J.J., Huelsenbeck J.P., Cunningham C.W., Swofford D.L., Waddell P.J. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42:384–397.

Caicedo A.L., Williamson S.H., Hernandez R.D., Boyko A., Fledel-Alon A., York T.L., Polato N.R., Olsen K.M., Nielsen R., McCouch S.R., Bustamante C.D., Purugganan M.D. 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 3:1745–1756.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540–552.

Cummings M.P., Neel M.C., Shaw K.L., Otto S. 2008. A genealogical approach to quantifying lineage divergence. Evolution. 62: 2411–2422.

de Queiroz A., Donoghue M.J., Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26: 657–681.

de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22:34–41.

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

Doolittle W.F. 1999. Lateral genomics. Trends Cell. Biol. 9:M5–M8.

Douzery E.J., Snell E.A., Bapteste E., Delsuc F., Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl. Acad. Sci. USA. 101:15386–15391.

Doyle J.J. 1992. Gene trees and species trees—molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Duan S., Lu B., Li Z., Tong J., Kong J., Yao W., Li S., Zhu Y. 2007. Phylogenetic analysis of AA-genome *Oryza* species (Poaceae) based on chloroplast, mitochondrial, and nuclear DNA sequences. Biochem. Genet. 45:113–129.

Ebersberger I., Galgoczy P., Taudien S., Taenzer S., Platzer M., von Haeseler A. 2007. Mapping human genetic ancestry. Mol. Biol. Evol. 24:2266–2276.

Eckert A.J., Carstens B.C. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49:832–842.

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

Goodman M., Czelusniak J., Moore G.W., Romeroherrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.

Hartmann S., Vision T.J. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8:95.

Huelsenbeck J.P., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

Huelsenbeck J.P., Ronquist F. 2001. MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Kim H., Hurwitz B., Yu Y., Collura K., Gill N., SanMiguel P., Mullikin J.C., Maher C., Nelson W., Wissotski M., Braidotti M., Kudrna D., Goicoechea J.L., Stein L., Ware D., Jackson S.A., Soderlund C., Wing R.A. 2008. Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. Genome Biol. 9: R45.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56: 17–24.

Kullberg M., Hallstrom B.M., Arnason U., Janke A. 2008. Phylogenetic analysis of 1.5 Mbp and platypus EST data refute the Marsupionta hypothesis and unequivocally support Monotremata as sister group to Marsupialia/Placentalia. Zool. Scr. 37:115–127.

Lakner C., van der Mark P., Huelsenbeck J.P., Larget B., Ronquist F. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol. 57:86–103.

Larget B. 2006. Bayesian untangling of Concordance Knots (BUCKy), version 1.1. Available from: http://www.stat.wisc.edu/ larget/ bucky.html.

Linnen C.R., Farrell B.D. 2008. Comparison of methods for species-tree inference in the sawfly genus Neodiprion (Hymenoptera: Diprionidae). Syst. Biol. 57:876–890.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542–2543.

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

Lu B.R., Naredo M., Juliano A., Jackson M.T. 2000. Preliminary studies on taxonomy and biosystematics of the AA-genome *Oryza* species (Poaceae). In: Jacobs S., Everett J., editors. Grasses: systematics and evolution. Melbourne (Australia): CSIRO. p. 51–58.

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

Mossel E., Vigoda E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science. 309:2207–2209.

Notredame C., Higgins D.G., Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Nylander J.A.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J.L. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Page R.D.M. 1994. Maps between trees and cladistic-analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43:58–77.

Page R.D.M. 1998. Genetree: comparing gene and species phylogenies using reconciled trees. Bioinformatics. 14:819–820.

Page R.D.M., Charleston M.A. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree species tree problem. Mol. Phylogenet. Evol. 7:231–240.

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

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:1634–1647.

Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.

R Development Core Team. 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rambaut A., Drummond A. 2005. Tracer version 1.3. Available from: http://beast.bio.ed.ac.uk/Tracer.

Rieseberg L.H., Baird S.J., Gardner K.A. 2000. Hybridization, introgression, and linkage evolution. Plant Mol. Biol. 42:205–224.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Sanderson M.J., McMahon M.M. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7(Suppl 1):S3.

Sang T., Ge S. 2007. The puzzle of rice domestication. J. Integr. Plant Biol. 49:760–768.

Savard J., Tautz D., Richards S., Weinstock G.M., Gibbs R.A., Werren J.H., Tettelin H., Lercher M.J. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. 16:1334–1338.

Semon M., Nielsen R., Jones M.P., McCouch S.R. 2005. The population structure of African cultivated rice *Oryza glaberrima* (Steud.): evidence for elevated levels of linkage disequilibrium caused by admixture with *O. sativa* and ecological adaptation. Genetics. 169:1639–1647.

Takahata N. 1989. Gene genealogy in 3 related populations—consistency probability between gene and population trees. Genetics. 122:957–966.

Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Vaughan D., Lu B., Tomooka N. 2008. The evolving story of rice evolution. Plant Sci. 174:394–408.

Wiens J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.

Wing R., Ammiraju J., Luo M., Kim H., Yu Y., Kudrna D., Goicoechea J., Wang W., Nelson W., Rao K., Brar D., Mackill D., Han B., Soderlund C., Stein L., SanMiguel P., Jackson S. 2005. The Oryza Map Alignment Project: the golden path to unlocking the genetic potential of wild rice species. Plant Mol. Biol. 59:53–62.

Zhou H.-F., Zheng X.-M., Wei R.-X., Second G., Vaughan D.A., Ge S. 2008. Contrasting population genetic structure and gene flow between *Oryza rufipogon* and *Oryza nivara*. Theor. Appl. Genet. 117:1181–1189.

Zhu Q.H., Ge S. 2005. Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. New Phytol. 167:249–265.

Zou X.-H., Zhang F.-M., Zhang J.-G., Zang L.-L., Tang L., Wang J., Sang T., Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biol. 9:R45.

Received 8 September 2008; reviews returned 20 November 2008; accepted 9 August 2009

Guest Associate Editor: Scott Edwards