Systematic Botany (2010), 35(3): pp. 512–523

Phylogenomic Analysis of BAC-end Sequence Libraries in *Oryza* (Poaceae)

Karen A. Cranston,[1,4,8,9] Bonnie Hurwitz,[1,2,9] Michael J. Sanderson,[1,3] Doreen Ware,[2,5] Rod A. Wing,[3,6] and Lincoln Stein[2,7]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson Arizona 85721, U. S. A.
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, U. S. A.
[3]BIO5 Institute, University of Arizona, Tucson Arizona 85721, U. S. A.
[4]Biodiversity Synthesis Center, Field Museum of Natural History, 1400 S. Lake Shore Drive, Chicago Illinois 60605, U. S. A.
[5]Robert W. Holley Center for Agriculture and Health, United States Department of Agriculture – Agricultural Research Service
[6]Arizona Genomics Institute and Department of Plant Sciences, University of Arizona, Tucson Arizona 85721, U. S. A.
[7]Ontario Institute for Cancer Research, Ontario M5G 0A3, Canada.
[8]Author for correspondence (kcranston@fieldmuseum.org)
[9]these authors contributed equally

Communicating Editor: Fernando Zuloaga

***Abstract*—**Analyses of genome scale data sets are beginning to clarify the phylogenetic relationships of species with complex evolutionary histories. Broad sampling across many genes allows for both large concatenated data sets to improve genome-scale phylogenetic resolution and also for independent analysis of gene trees and detection of phylogenetic incongruence. Recent sequencing projects in *Oryza sativa* and its wild relatives have positioned rice as a model system for such "phylogenomic" studies. We describe the assembly of a phylogenomic data set from 800,000 bacterial artificial chromosome (BAC) end sequences, producing an alignment of 2.4 million nucleotides for 10 diploid species of *Oryza*. A supermatrix approach confirms the broad outline of previous phylogenetic studies, although the nonphylogenetic signal and high levels of missing data must be handled carefully. Phylogenetic analysis of 12 chromosomes and nearly 2,000 genes finds strikingly high levels of incongruence across different genomic scales, a result that is likely to apply to other low-level phylogenies in plants. We conclude that there is great potential for phylogenetic inference using data from next-generation sequencing protocols but that attention to methodological issues arising inevitably in these data sets is critical.

***Keywords*—**BAC-end sequencing, gene trees, missing data, *Oryza*, Phylogenomics, rice.

Application of genome-scale data sets to phylogenetic problems, or phylogenomics, has become increasingly feasible in many taxa (Rokas et al. 2003; Savard et al. 2006; Dunn et al. 2008; Nishihara et al. 2007; Pollard et al. 2006). Plant studies have mainly focused on the plastid genome (Moore et al. 2006; Jansen et al. 2005; Cronn et al. 2008) and to a lesser extent EST libraries of nuclear genomes (Lee et al. 2002; Sanderson and McMahon 2007), but next generation sequencing technologies promise to expose the plant nuclear genome to much more comprehensive phylogenomic investigation. Early phylogenomics studies tackled troublesome nodes in the tree of life through analysis of a large number of concatenated nucleotides. The major advantage of genome-scale data, however, may be the availability of many genes that can be analyzed independently to produce a collection of potentially incongruent gene trees. This incongruence may indicate that different genes have different evolutionary histories (Pamilo and Nei 1988; Hudson 1992; Doyle 1992; Maddison 1997). Studying the phylogenetic distribution of gene trees allows for insight into the evolutionary processes that underlie different regions of the genome.

While phylogenomic studies in yeast (Rokas et al. 2003) or *Drosophila* (Pollard et al. 2006) have the luxury of fully sequenced genomes across closely related taxa, analyses of nonmodel species instead usually rely on samples of large numbers of loci across the genome, either from traditional directed gene-by-gene sequencing or, increasingly, from data mining or high-throughput sequencing approaches short of generating complete genomes. In the latter case, the availability of a full annotated genome for a model organism can greatly reduce the time and cost required for assembly of sequences from closely-related species. The creation of bacterial artificial chromosome (BAC) libraries and subsequent sequencing of either full clones or end sequences is one approach to generation of this type of genomic sequence data. Given the size and complexity of many plant genomes (Soltis et al. 2003) and the resulting cost and bioinformatics challenge of whole-genome sequencing, it is likely that these types of sequencing strategies will continue to be important in plant biology for some time to come.

BAC-end sequences per se have rarely been used in phylogenetics, but the structure of the data share many properties with other data sets assembled from genomic libraries, such as ESTs, and with aspects of database mining, that have been more heavily exploited. Unlike data in conventional phylogenetic analyses, which begin largely with a set of homologous sequences given a priori (e.g. via primer design), data extracted from genomic libraries are initially unstructured, and homologies have to be identified by algorithmic procedures, leading to often relatively heterogeneous sequence clusters that are then aligned. Some protocols along these lines have been developed for analysis of EST libraries (Lee et al. 2002; Sanderson and McMahon 2007; Dunn et al. 2008; Kullberg et al. 2008), or mining subsets of GenBank or genome databases (Driskell et al. 2004; Philippe et al. 2005; McMahon and Sanderson 2006; Ciccarelli et al. 2006; Robbertse et al. 2006). A typical consequence of all such procedures is the assembly of alignments with a significant fraction of data missing owing to lack of representation of one or more taxa in that cluster, either through failure in sampling leading to missing data in the source database, or failure in detection during homology searches. Previous work has suggested that the fragmentation caused by missing data may or may not interfere with accurate tree reconstruction, depending on the quantity and patterns of missing data (Wiens 1998; Sanderson et al. 2007). Moreover, missing data is modeled differently across phylogenetic inference procedures and scaling data upward could conceivably exacerbate this issue.

In this paper we exploit deep BAC-end sequence libraries from the *Oryza* Map Alignment Project (OMAP) (Wing

et al. 2005; Kim et al. 2008). Rice (*Oryza sativa* L.) is well known as the world's most important food crop (Vaughan et al. 2003). Draft genomes of both the indica (*O. sativa* L. ssp. *indica*) and japonica (*O. sativa* L. ssp. *japonica*) cultivars (Goff et al. 2002; Yu et al. 2002), as well as the finished japonica genome (IRGSP 2005) are now available. In addition, the modest size of the genome, availability of wild and cultivated species, and the presence of both diploid and polyploid species make *Oryza* an ideal candidate for both empirical and methodological studies. To further this aim, OMAP has used the finished *Oryza sativa* subsp. *japonica* genome as a reference to construct deep BAC libraries and physical maps of twelve wild and one cultivated *Oryza* species. The volume of data generated by OMAP is large, amounting at present to over 1.5 million sequences for potential phylogenetic analysis (see http://www.omap.org).

Historical patterns of gene flow, demography, selection, and diversification in rice have been inferred using molecular sequence data at many different scales (Caicedo et al. 2007; Zhang and Ge 2007; Zhu et al. 2007), from a few loci to complete genomes, and from a few exemplar species to much more intense population-level samples. However, confounding processes of hybridization, introgression, polyploidization/diploidization (Wang et al. 2005), selective sweeps (Caicedo et al. 2007; Olsen et al. 2006), linkage disequilibrium (Garris et al. 2003; Semon et al. 2005; Mather et al. 2007), and domestication (Olsen et al. 2006; Zhu et al. 2007) have altered the signal left in the genomes of *Oryza* species, raising obstacles to robust inferences. Even the most basic phylogenetic question of the species tree for *Oryza* remains unresolved in a few important places, such as the nearest relatives of *O. sativa* (Ishii et al. 2001; Ren et al. 2003; Zhu and Ge 2005; Duan et al. 2007). Zou et al. (2008) exploited the availability of the complete genome sequence for *O. sativa* to undertake a phylogenomic analysis in *Oryza*. They designed primers to amplify single copy loci for 62 genes across 11 species in *Oryza* and the outgroup *Leersia*, plus 80 additional genes for a subset of six of these *Oryza* species. They recovered a well-supported species tree confirming the monophyly of the cytogenetically defined genome groups, strong support for much of the deeper phylogeny within *Oryza*, and probable evidence of lineage sorting at some nodes in the tree.

However, Zou et al. (2008) considered only three species in the AA genome group, and for these only 62 genes were sampled from their larger analysis. This closely knit genome group (or species complex (Tateoka 1962)), includes the two domesticated rice species, Asian *O. sativa* and African *O. glaberrima* Steud., and six wild species, distributed in tropical or subtropical Asia, Australia, Africa, and Latin America (Lu et al. 2000; Vaughan et al. 2003). This is a group with a difficult taxonomic history, uncertain species boundaries, and incomplete crossability barriers between some taxa (Lu et al. 2000). The Zou et al. (2008) study found extensive gene tree conflict at deeper nodes within *Oryza*, and it seems probable that a broader sampling of AA species would uncover this within the AA species as well. The OMAP data permits us to extend the taxon sampling to five AA genome species and to a much larger sample of loci across the genome.

The AA genome species are those that are most relevant to unlocking the complex domestication history of cultivated rice. Genetic variation within and between species, studied using both phylogenetic and population genetic techniques, is addressing questions about the origins of the different cultivars of domesticated rice, including the number of domestications (Sang and Ge 2007; Vaughan et al. 2008). A well-supported species level phylogeny will give a broad evolutionary context to future work into rice domestication.

In this paper we investigate whether the initial pool of ~820,000 BAC-end sequences, analyzed with attention to the unique problems entailed by the low coverage of the genomes, can shed light on the recalcitrant problem of the closest relatives of *O. sativa*. Despite widespread incongruence between different genomic regions, our genome-scale approach confirms previous results about the relationships between these species.

Materials and Methods

***Species Included in the Study*—**In addition to the genome reference sequence from *O. sativa*, we used OMAP BAC-end sequences (BES) from the cultivated African species *O. glaberrima*, three African species, *O. barthii* A. Chev., *O. punctata* Kotschy ex Steud. and *O. brachyantha* A. Chev. & Roehr., one Australian species, *O. australiensis* Domin and four Asian species, *O. rufipogon* Griff., *O. nivara* Sharma & Shastry, *O. officinalis* Wall. ex G. Watt, and *O. granulata* Nees & Arn. ex G. Watt. These species are sampled from the genome groups AA (*O. sativa*, *O. rufipogon*, *O. nivara*, *O. glaberrima* and *O. barthii*), BB (*O. punctata*), CC (*O. officinalis*), EE (*O. australiensis*)), FF (*O. brachyantha*), and GG (*O. granulata*). See (Wing et al. 2005) for details of the OMAP project, including source of plant accessions and construction and alignment of BAC based physical maps. *Oryza* also contains ~9 polyploid species, all evidently derived from taxa outside of the AA genome group. Although BAC libraries have been constructed for four of these species, we restricted our attention to phylogenetic inference in the diploid taxa because inferring reticulate evolutionary histories adds an additional layer of complexity in this already difficult problem (Huber and Moulton 2006).

***Sequencing and Genomics Pipeline*—**A starting pool of BAC-end sequences (BES) for nine *Oryza* species was obtained from the GenBank GSS Division using the following query: "OR__Ba or OR_BBa or OG_BBa or OP__Ba or OO__Ba or OA_CBa or OB__Ba or OG_ABa". In addition, sequences from *O. barthii* not yet submitted to Genbank were obtained directly from the OMAP group. To identify BES in coding regions of the genome, we aligned BES from the wild rice species 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). The alignments were then postprocessed to find the best hit. Since a BES could be split across multiple regions of a gene, multiple high scoring pairs for the gene with the best match were retained. Overlapping BES from the same species were aligned using ClustalW (Thompson et al. 1994) and default parameters, to create a consensus exon sequence. We then performed an all-vs-all BLAST search of the consensus exon sequences for each of the wild rice species and the CDS sequences from *O. sativa* to identify exon sequences that were not single copy. If an exon was found to have multiple hits to its own genome or that of another species, we removed it from the analysis to avoid problems with paralagous sequences in the phylogenetic inference. The remaining exons were filtered to ensure 70% coverage between sequences, and assembled into clusters of orthologs via a single linkage clustering algorithm, keeping only clusters with four or more sequences. Four sequences is the minimum necessary to produce an informative unrooted phylogenetic tree (though not always sufficient).

Orthologous exons were then aligned 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. If less than four species remained after the alignment had been filtered the alignment was removed. 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 (Castresana 2000), which is designed to identify regions of poor alignment quality, did highlight some large blocks for removal, but further investigation revealed that the basis for removal was generally the 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.

We applied one final filter to generate a set of alignments for gene tree construction. In this step, we restricted the set to alignments that contained *O. punctata* as the outgroup, *O. sativa* and at least two of the other four AA-genome species in the ingroup (to ensure a minimum of four species overall).

Phylogenetic Analysis—Phylogenies were reconstructed at two different taxonomic scales, first with data for all 10 species of *Oryza* (henceforth the "ORYZA" data set); second with just the five AA genome species and its probable closest outgroup, *O. punctata* (henceforth the "AA" data set). Using a supermatrix approach (de Queiroz and Gatesy 2007), we first inferred the larger phylogeny based on a concatenated alignment using Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Markov Chain Monte Carlo (BMCMC) methods. Maximum parsimony analyses were done using PAUP* v. 4.0b10 (Swofford 2003); heuristic search with default parameters; ML using Garli (Zwickl 2006) for analyses and PAUP for consensus tree construction from Garli bootstrap result; BMCMC using Mr. Bayes v. 3.1.2 (Huelsenbeck and Ronquist 2001); nruns = 3, ngen = 3,000,000, samplefreq = 1,000. For likelihood-based analyses, we employed a GTR + Γ model of sequence evolution, using empirical base frequencies and an estimated alpha parameter for the gamma distribution. All bootstrap reanalyses (Felsenstein 1985) used 500 replicates.

Using the AA alignment, phylogenetic analyses were undertaken at several different "genomic" scales: based on the supermatrix consisting of the entire concatenated alignment (the "AA_all" data set), supermatrices constructed from each chromosome (the "AA_chrom" data set), and finally from individual genes (the latter sometimes consisting of only parts of genes: henceforth labeled the "AA_genes" data set). We did not do the chromosome and gene level analyses due to increased amounts of missing data in the 10-species ORYZA alignment.

As for the concatenated alignments, we used MP and ML reconstruction. The model of evolution was GTR + Γ, except for the gene level analysis, where we used HKY + Γ to prevent overparameterization with these much shorter alignments (where short length is coupled with low sequence divergence). We used 10 categories for the discrete gamma distribution and a proportion of invariant sites equal to zero. We also attempted these analyses in RAxML (Stamatakis et al. 2005). RAxML does not implement a proportion of invariant sites and uses a completely different implementation of rate heterogeneity, the CAT model (Stamatakis 2006). This implementation forgoes the usual modeling of rates as drawn from a gamma distribution and bins the rates into a number of categories (by default, 25) and assigns each site to one of the rate categories. Unfortunately, for an alignment with only 6 species, there is not enough information in each site to accurately estimate the per-column rate (A. Stamatakis, pers. comm.). The more standard GTR + Γ model in RAxML is limited to 4 gamma-distributed rate categories. For this reason, results from RAxML are not included here, but we mention this because it is an obvious candidate when choosing software for large scale phylogenetics using ML. Trees have been deposited in Dryad (http://hdl.handle.net/10255/dryad.1611).

Hypothesis Testing—We examined both the sequence data and the output trees using a variety of diagnostic and descriptive methods, including testing for bias of GC content, partition homogeneity, strength of support for various topologies, effect of missing data and incongruence between trees produced from different genome regions. We describe details of each of these procedures below.

We tested for bias in GC content across the species in the AA_all alignment and in each of the AA_chrom alignments. We used a χ^2 test, implemented as the "basefreq" command in PAUP*, to test the observed frequency of each base against the expected value under a hypothesis of equal frequencies.

We tested for significant incongruence between the chromosomes using the partition homogeneity, or incongruence length difference (ILD), (Farris et al. 1995) test applied to the AA_chrom alignments. We defined twelve partitions based on the chromosome boundaries in *O. sativa* and performed the ILD test using the "HomPart" command in PAUP*, with a parsimony-based branch and bound search, and 1,000 replicates.

To test whether some trees were supported significantly better than other trees by a given alignment we used an SH test (Shimodaira and Hasegawa 1999). Given the small number of taxa, there are a countable number of possible topologies, and we can calculate the likelihood for each possible tree and test against the maximum likelihood topology. The SH test corrects for multiple comparisons and also allows for a posteriori selection of topologies for comparison. There are 105 possible topologies for the five AA genome species with *O. punctata* as the root. This is the first set of hypotheses tested. The second set includes only the 15 topologies that contain the clade with *O. glaberrima* and *O. barthii*, which had 100% support in all analyses of the concatenated alignments. One disadvantage of the SH test is that it can be overly conservative, including more trees in the final confidence set with larger sets of input trees (Strimmer and Rambaut 2002). The second set of trees allowed all of the possible hypotheses for the relationships of *O. sativa* while restricting to a smaller number of total input trees.

We also tested the effect of missing data on the phylogenetic results. In both the AA and ORYZA alignments, the amount of missing sites across species is not consistent, and missing sites tend to appear in large blocks, resulting from a lack of BES for that exon, rather than from small insertions or deletions. We used the AA_all alignment to determine whether the amount or distribution of missing data was affecting the phylogenetic analysis. To test the effect of the missing data, we then used a simulation approach. First, we removed all sites that contain any amount of missing data, producing a 'gapless' alignment. We then generated an alignment of equal length as the original alignment by sampling site patterns,, with replacement, from the gapless alignment. Then, we added the same number of missing characters as in the original alignment, but scattered randomly across the species (see Fig. 1). We compared the resulting topology and bootstrap scores / posterior probabilities from ML and BMCMC estimation of the original alignment, gapless alignment, simulated full-length alignment and the alignment with simulated missing data.

Finally, we used some nonstandard methods to describe and compare phylogenies built from chromosome and gene-level alignments. Consensus tree methods (reviewed by Bryant (2003)) can extract the common signals between the trees, but do not convey the details about incongruence between the input trees. Bipartition scores (bootstrap proportions or posterior probabilities) describe the statistical confidence associated with the two sets of taxa separated by an edge in the phylogenetic tree (a clade on a rooted tree). We used bipartition scores to describe incongruence between the chromosome trees, which have complete taxon sampling. However, in the case of the gene trees, these scores must be interpreted with caution due to the different taxa sampled in the different gene trees. For example, suppose (A,(B,(C,(D,E)))) is the correct tree. The clade (C,D) might be strongly supported on all trees that are missing taxon E, but this should not be taken to contradict the clade (D,E) on the fully sampled tree. For this reason, in the gene tree summary, we used three-species rooted triplets so that we can clearly label the resolution of the three species, as well as indicate cases in which a gene tree is not informative for a given set of species.

Results

Sequence Alignments—We obtained 820,247 BAC-end sequences (BES) for eight *Oryza* species from the GenBank GSS database and one *Oryza* species directly from OMAP researchers. The mean read length of these sequences is 655 bp and the minimum and maximum length is 101 and 1,012, respectively. Running BLAST searches against these sequences against the IRGSP V3 gene models yielded 44.5% coverage of the total CDS sequence from *O. sativa* (44,492,676 bp), representing 27,751 of the 37,544 total genes in rice. Creating a consensus sequence of two or more overlapping BES gene hits from within a single species generated a pool of exons across the nine species having 14.6% exon coverage, representing 23,192 total genes. The all-against-all BLAST procedure aimed at excluding duplications reduced the pool of exon sequences to orthologs having 9.7% exon coverage from 13,198 genes of *O. sativa*. Finally, after removing any clusters with fewer than four sequences, we were left with 5.5% exon coverage in 9,481 genes with orthologous sequence suitable for multiple sequence alignment and phylogenetic inference.

The full concatenated alignment of exonic sequence from 9,481 genes across the 10 diploid species is 2.45 million nucleotides long. See Table 1 for details of the chromosome-by-chromosome and full concatenated alignments over the 10 species and 6 species taxon sets. The AA_all alignment, with *O. punctata* (BB) as the outgroup, contains a total of 1.21 million sites. This is two orders of magnitude larger than the recent phylogenetic studies on the AA genome: four genes, 2,750 sites (Duan et al. 2007), four genes, 2,315 sites (Zhu and

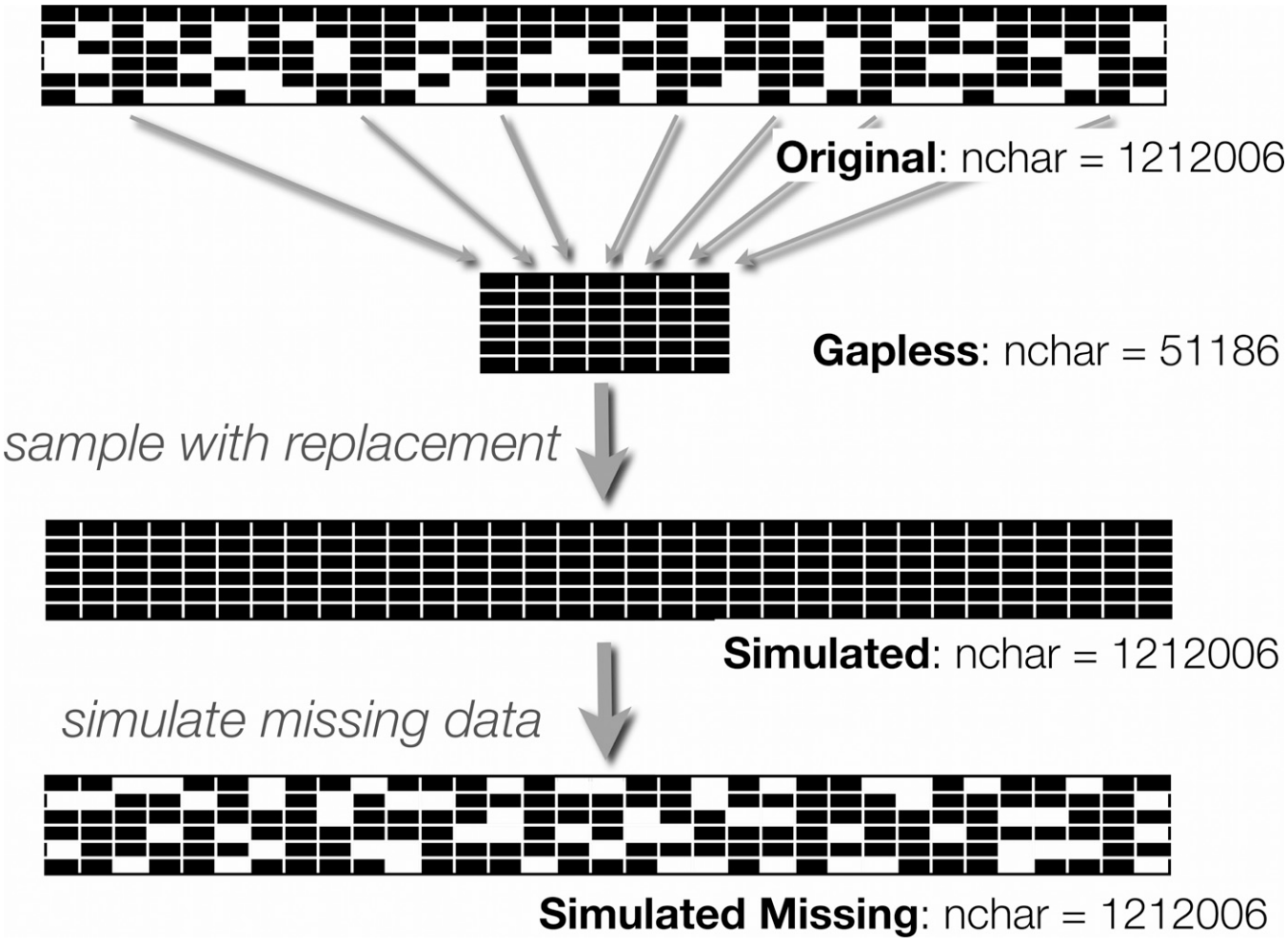

FIG. 1. Simulation of data sets to test the effect of missing data. In the original alignment, missing data is distributed unevenly among the species. In the final simulated alignment, missing sites are distributed randomly among the species.

Ge 2005) and one order of magnitude larger than the recent phylogenomic analysis of Zou et al. (2008), which included 142 genes and 124,000 sites.

However, these alignments have a large amount of missing data. Many genes are missing for many species, and the distribution of missing data is highly skewed across the species. For the model organism *O. sativa*, with a full genome sequence, there was essentially no missing data, but the sequences for the remaining members of the AA genome group contained between 40 and 60% missing data, while the non-AA genome species had up to 72% missing data (see Table 2). Overall, 54% of the nucleotides in the ORYZA alignment and 31% of the nucleotides in the AA alignment are missing.

TABLE 1. Summary statistics for alignments used in this paper. "Genes" refers to the number of genes represented by exons in the alignment, not to the presence of full gene sequences. Chromosome sizes are taken from the finished genome sequence of *O. sativa* (IRGSP 2005).

		ORYZA alignment		AA alignment	
Chromosome	Size (MB)	Genes	Nucleotides	Genes	Nucleotides
1	45.05	1,369	350,899	689	170,534
2	36.78	1,241	310,898	627	159,730
3	37.37	1,345	346,286	664	166,171
4	36.15	906	241,307	446	123,537
5	30.00	732	190,502	345	93,344
6	31.60	762	200,118	382	102,750
7	30.28	724	179,486	370	85,084
8	28.57	555	146,697	278	73,543
9	30.53	491	129,537	253	62,086
10	23.96	488	127,103	231	60,288
11	30.76	411	120,011	204	61,080
12	27.77	457	116,711	220	53,859
Total	388.82	9,481	2,459,555	4,718	1,212,006

TABLE 2. Species used in this study, along with genome type and sequence information in the ORYZA alignment with 10 species and AA alignment with six species. The two cultivated species are highlighted in bold.

		ORYZA alignment		AA alignment	
Species	Genome	Nucleotides	% missing	Nucleotides	% missing
O. sativa	**AA**	**2,438,526**	**0.9**	**1,202,073**	**0.8**
O. rufipogon	AA	1,114,050	54.7	765,348	36.9
O nivara	AA	1,409,694	42.7	904,409	25.4
O. glaberrima	**AA**	**1,071,443**	**56.4**	**760,162**	**37.3**
O. barthii	AA	1284,292	47.8	869,757	28.2
O. punctata	BB	892,781	63.7	546,078	54.9
O. officinalis	CC	942,897	61.7	-	-
O. australiensis	EE	688,229	72.0	-	-
O. brachyantha	FF	709,288	71.2	-	-
O. granulata	GG	672,568	72.7	-	-

The initial massive alignments for the concatenated analyses pare down to a much smaller amount of data at the level of gene tree analyses when we require that the alignment contain at least two of the AA-genome species in addition to *O. sativa* and *O. punctata*. The full alignments for the AA-genome species contain nearly 5,000 genes. The filtering of these alignments for gene tree construction left 1,720 genes with a mean alignment length of 350 nucleotides. All six species are present in 307 of these alignments, five species in 546 alignments and four in the final 867 alignments. Even with this data reduction from 9,481–1,720 genes, the number of genes in this analysis far exceeds any previous phylogenetic study in *Oryza*.

Supermatrix Analysis—Relationships reconstructed from the larger ORYZA alignment are generally consistent and well-supported across reconstruction methods (Fig. 2). The AA genome group of species is unequivocally supported, and within the AA-genome group, support for the monophyly of the Asian species *O. rufipogon* and *O. nivara* is 100% in all trees, as is monophyly of the African species *O. barthii* and *O. glaberrima*. However, the position of *O. sativa* is not consistently well supported. In the MP analyses, *O. sativa* is the sister group to the other Asian species with 100% bootstrap support, a relationship also found in the BMCMC trees, albeit with only 68% posterior probability. The ML analysis places *O. sativa* as the sister group of the other AA genome species, again lower support values (72% bootstrap). In the AA_all supermatrix, the position of *O. sativa* is also uncertain, with MP analysis placing *O. sativa* as sister to the Asian species, and ML / BMCMC placing it as sister to all the other AA genome species, but again with low support (51% bootstrap, 70% posterior probability). See Fig. 3.

The topology resulting from the AA_all alignment is sensitive to the method of modeling rate heterogeneity in likelihood-based analyses, in particular to the number of categories in the discretization of gamma-distributed rates. As we increase the number of rate categories for the gamma distribution (and therefore better estimate the continuous distribution), the likelihood increases and the maximum likelihood topology switches from the expected tree to the tree where *O. sativa* is sister to the other AA genome species (Fig. 4). When the data sets are analyzed with a simple HKY model (Hasegawa et al. 1985) without gamma-distributed rates, both alignments yield topologies with *O. sativa* as sister to *O. rufipogon* and *O. nivara*, with high bootstrap proportions and posterior probabilities (trees not shown).

The SH test based on the complete set of all 105 possible trees returns a set of seven trees that cannot be rejected as statistically different from each other. The smaller input list of trees that include the African (*O. glaberrima*, *O. barthii*) clade shortens the list of nonrejected trees to only three, the two trees seen in Figs. 2 and 3 as well as the topology where

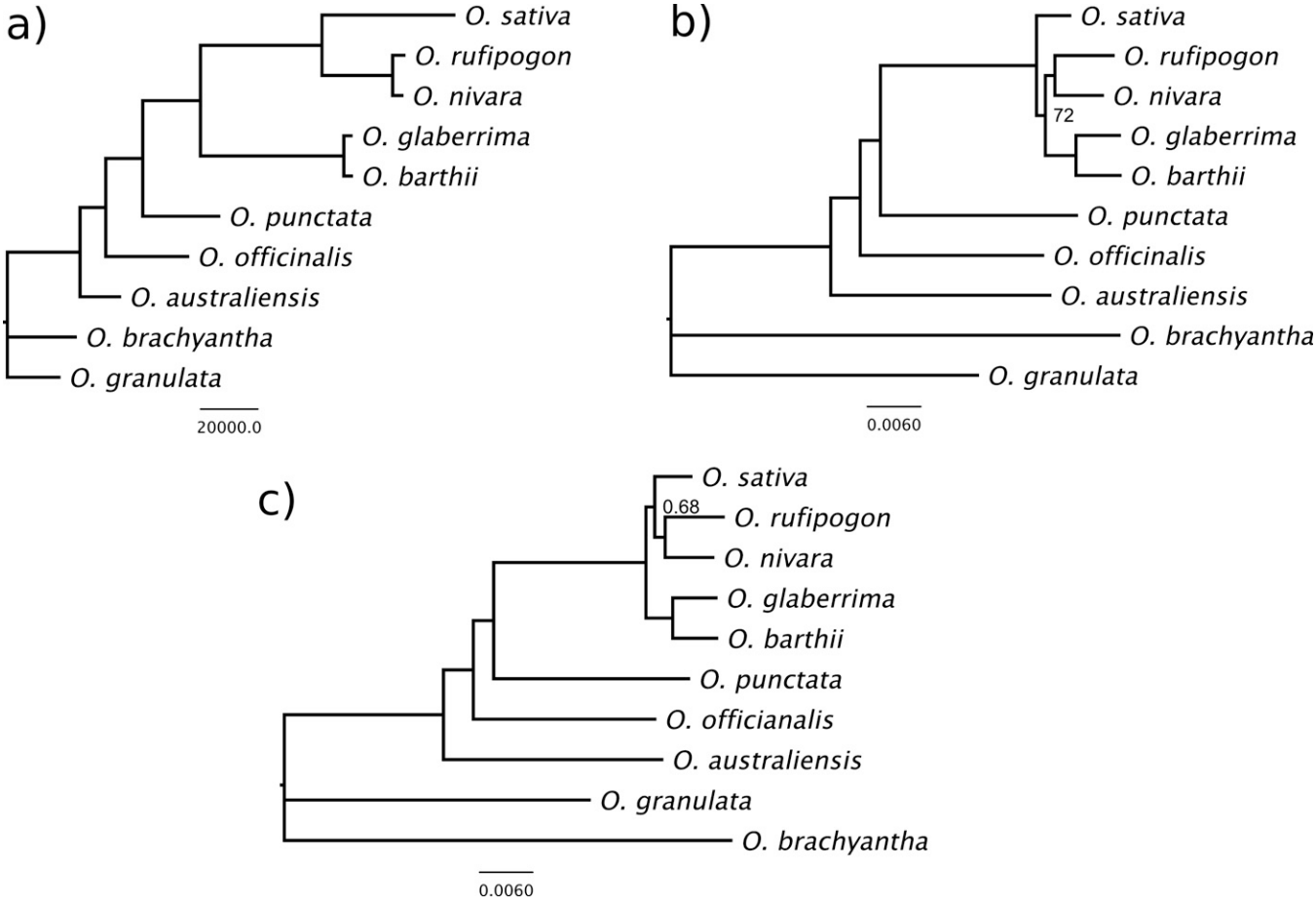

FIG. 2. Unrooted majority-rule consensus phylogenies for all 10 *Oryza* species. a) Maximum Parsimony (MP) analysis; b) Maximum Likelihood (ML) analysis; c) Bayesian Markov Chain Monte Carlo (BMCMC) analyses. Node labels are bootstrap proportions for MP and ML and posterior probabilities for BMCMC. Unlabeled nodes have 100% bootstrap proportion/1.0 posterior probability. Note differences in placement of *O. sativa* between the phylogenies produced using different methods.

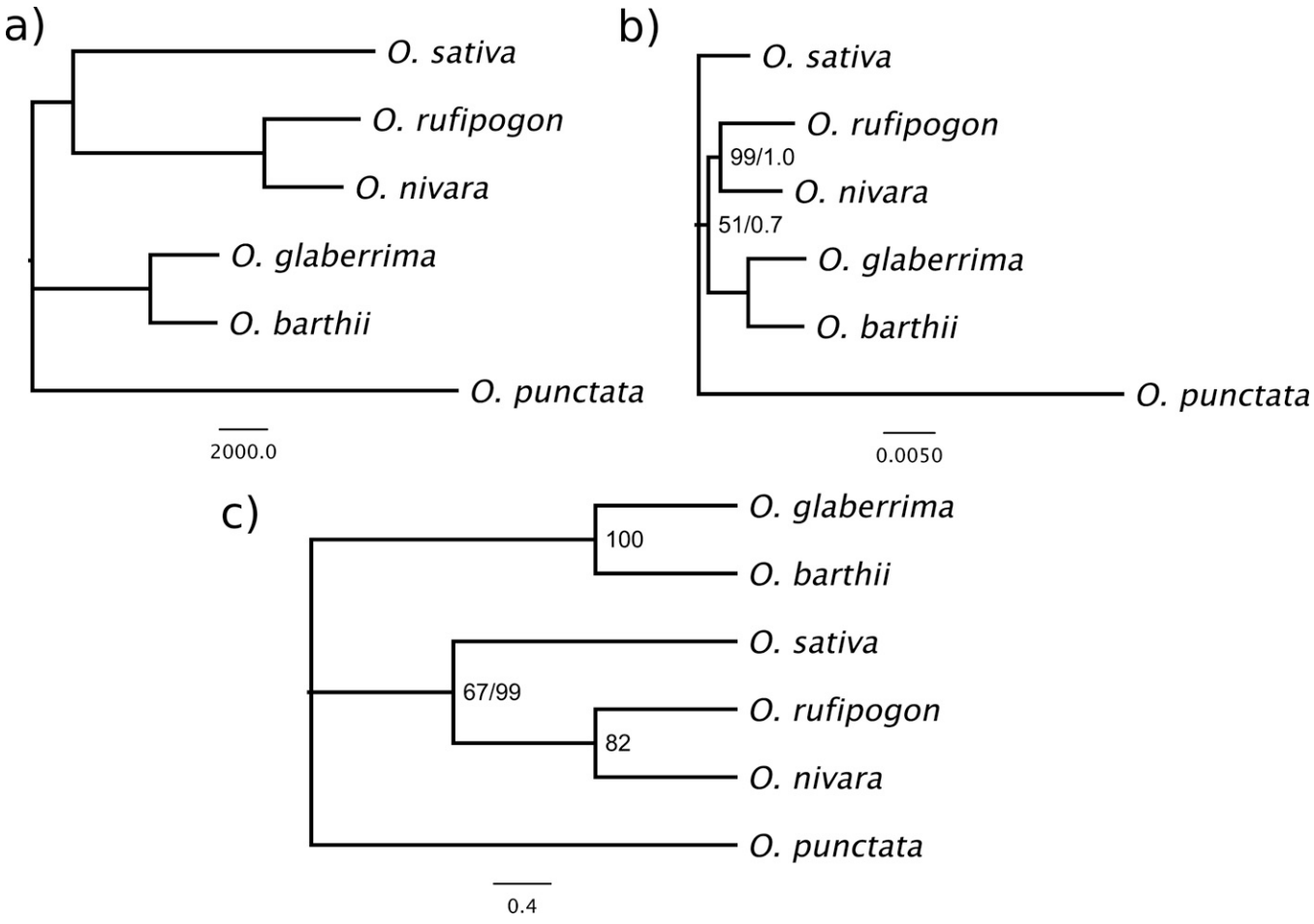

FIG. 3. Unrooted majority-rule consensus phylogenies for the AA-genome *Oryza* species. a) Maximum Parsimony (MP) analysis of all sites; b) Maximum Likelihood (ML) and Bayesian Markov Chain Monte Carlo (BMCMC) analyses of all sites; c) ML and BMCMC analyses of alignment with gaps removed. Node labels are bootstrap proportions for MP, and bootstrap proportion/posterior probabilities for ML/BMCMC. Unlabeled nodes have 100% bootstrap proportion / 1.0 posterior probability. Branch lengths in b) and c) are from BMCMC analyses.

O. sativa is sister to the African species. When increasing the number of categories in the gamma distribution, as described above, the set of topologies does not change, but the topology labeled as `best' differs (as we see with the directly-calculated likelihood values in Fig. 4).

The large length of these alignments (2.45 million and 1.1 million nucleotides for the ORYZA and AA alignments, respectively) posed computational challenges to phylogenetic inference programs. We were able to use PAUP*, RAxML, Garli, and MrBayes for analysis. However, Garli and MrBayes required approximately five and three GB RAM, respectively, which is more memory than would be available on many desktop computers. We could not open the alignment files in PhyML (Guindon and Gascuel 2003) and were unable to perform Shimodaira's AU test of topologies in CONSEL (Shimodaira and Hasegawa 2001). Computational speed was less of an issue than amount of memory required, which is not unanticipated given the small number of species. For programs that would import the data matrix, the time to read the file and initialize the data for likelihood calculations was nearly as long, or longer, than the time required to infer the most likely tree.

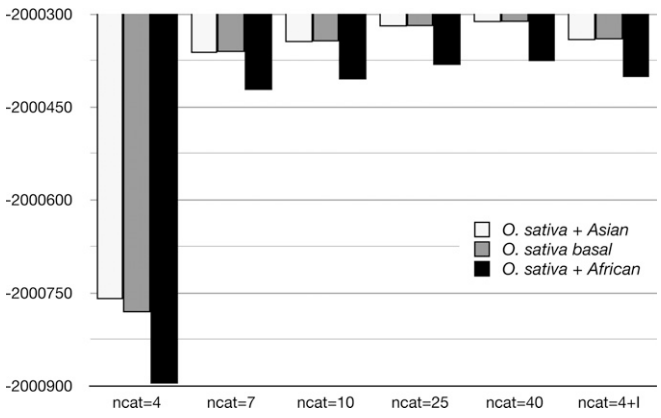

FIG. 4. Effect of number of categories for the discrete Gamma distribution on the log likelihood for three topologies of the AA-genome species.

Systematic Bias?—One possible explanation for the differences between the MP and ML results in the AA_all alignment is biased GC content across the species (Phillips et al. 2004). While the differences in base frequency across the AA species are statistically significant (p = 0.046), the absolute differences are small, and the significance is due to the statistical power enabled with 1.2 million nucleotides. For example, the range of GC content across the AA genome species was between 0.484 in *O. rufipogon* and *O. glaberrima* to a maximum value of 0.487 in *O. barthii*, a difference of only 0.003. At

the level of individual chromosomes, five of the six shortest chromosomes show no significant differences in nucleotide frequency. Only the longest chromosome alignments show significant differences.

The results of the missing data simulations indicate that the distribution of missing data is important in this data set. The gapless alignment, without any missing data, contains 51,168 nucleotides. The phylogeny reconstructed from this alignment does not differ in topology across the reconstruction methods, with *O. sativa* as sister to the other Asian AA genome species, *O. rufipogon* and *O. nivara* with 63% bootstrap support in ML and 100% posterior probability in BMCMC (Fig. 3). The full-length simulated alignment increases the support values to 100% for both ML and BMCMC. Finally, when we add the same amount of missing data as the original alignment, but distributed randomly across the species, the same topology is recovered with 99% bootstrap support and 98% posterior probability for the clade containing *O. sativa*, *O. rufipogon*, and *O. nivara*.

Chromosome Trees—Six different bootstrap (majority rule consensus) trees were recovered from ML analysis of the twelve AA_chrom alignments. These include the two strongly supported trees from the AA_all concatenated analyses, the topology grouping *O. sativa* with the other Asian species and the topology where *O. sativa* is sister to all the other AA species.

Inspection of bipartition support values suggests that incongruence between chromosomes is significant. Across the twelve consensus trees, there are six different bipartitions with greater than 95% bootstrap support, and only one, the grouping of the African species *O. glaberrima* and *O. barthii*, is found with high support in all chromosomes. Figure 5 illustrates the varying support for three common bipartitions across the chromosomes. We confirmed the differing signal between the chromosomes using a partition homogeneity test ($p = 0.001$). When using MP reconstruction, there are eight different topologies and only three statistically significant partitions, none of which are present in all of the chromosomes. For only four of the chromosomes (3, 6, 8 and 10) did MP and ML return the same topology.

Given the results seen with the AA_all alignment, we investigated whether missing data was causing some of this incongruence. When all missing data was removed from the chromosome alignments, there is still widespread incongruence between chromosomes and between reconstruction methods. In the ML analysis, removal of missing data changes the consensus tree topology for 10 of the 12 chromosomes (noting that some of these changes are a result of differences in amount of resolution) and increases the amount of incongruence between the chromosomes, such that only chromosomes 2 and 6 share the same topology. While missing data seem to affect the phylogenetic reconstruction of the chromosome-level alignments, as seen with the concatenated sequences, the presence of missing data is not solely responsible for the phylogenetic differences between the chromosomes.

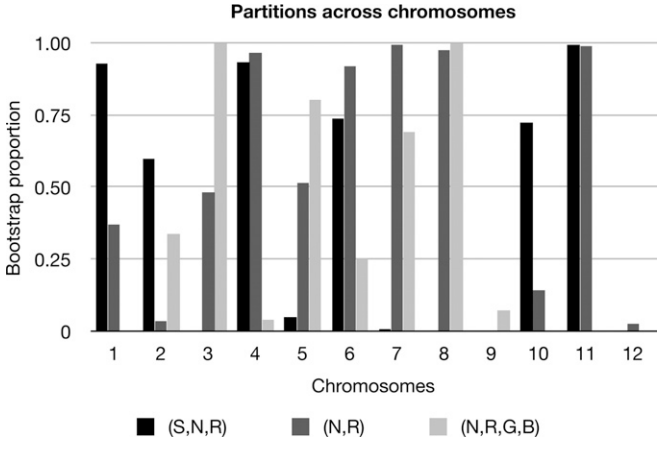

FIG. 5. Variability in bootstrap support for three common bipartitions across the 12 chromosomes. Species are S = *O. sativa*; N = *O. nivara*; R = *O. rufipogon*; G = *O. glaberrima*; B = *O. barthii*.

'Gene' Trees—Summarizing the maximum likelihood analysis of the 1,720 `genes' and 500 bootstrap replicates presents a challenge due to the nonoverlapping taxon sets and sheer number of resulting phylogenies. The list of bootstrap (majority-rule consensus) trees contains 235 unique topologies, with varying levels of resolution, across the alignments of four, five, and six species. There are 389 completely unresolved topologies, indicating alignments with no information about species relationships. This was not unanticipated, given the shorter length of these alignments and low level of sequence divergence between the species. Looking specifically at the six species alignments, we obtained 118 different topologies. The number of possible rooted multifurcating topologies for the five ingroup species is 236 (Felsenstein 2004). Our result, which represents half of the possible topologies, indicates the great extent of gene tree incongruence.

We summarized the signals in the trees using triplets of species rather than bipartition scores. Results for four different triplets are given in Fig. 6. We note that for all triplets, all three possible relationships are supported by at least 14% of the gene trees. The triplets that include *O. sativa*, one Asian species and one African species illustrate varying relationships across the shortest internal branch in the species phylogeny, separating the Asian and African species. These are the two pie charts at the top of Fig. 6. Under a hypothesis of lineage sorting with three species (Pamilo and Nei 1988), the gene tree matching the species tree is expected to be the most prevalent, with the two alternate hypotheses at lower, but equal, frequency. This is what we see for these triplets. Support for the different relationships in the triplet comprising the three Asian species is more equally distributed, perhaps indicating the presence of factors other than lineage sorting in these closely related species. Finally, for the triplet including *O. sativa* with the two African species, we see most support for the grouping of the two African species but the alternate resolutions exist in more than 10% of the trees. The African clade was supported at 100% by all supermatrix analysis, but we do see some contradictions of this relationship in the gene trees. Nearly one-quarter of the total gene trees were completely unresolved, highlighting cases where our 'gene fragments' were too short to be informative for species this closely related.

The gene trees are well-distributed across the genome, and there does not appear to be any significant relationship between genome location and topology, however (results not shown). While this may be taken as evidence to support the hypothesis that incomplete lineage sorting is the primary cause of the incongruence (Pollard et al. 2006), we caution that the sampling density of the genes is rather sparse. Over 450 Mb of genome sequence in *O. sativa*, our 1,720 genes constitute an average distance between genes of approximately

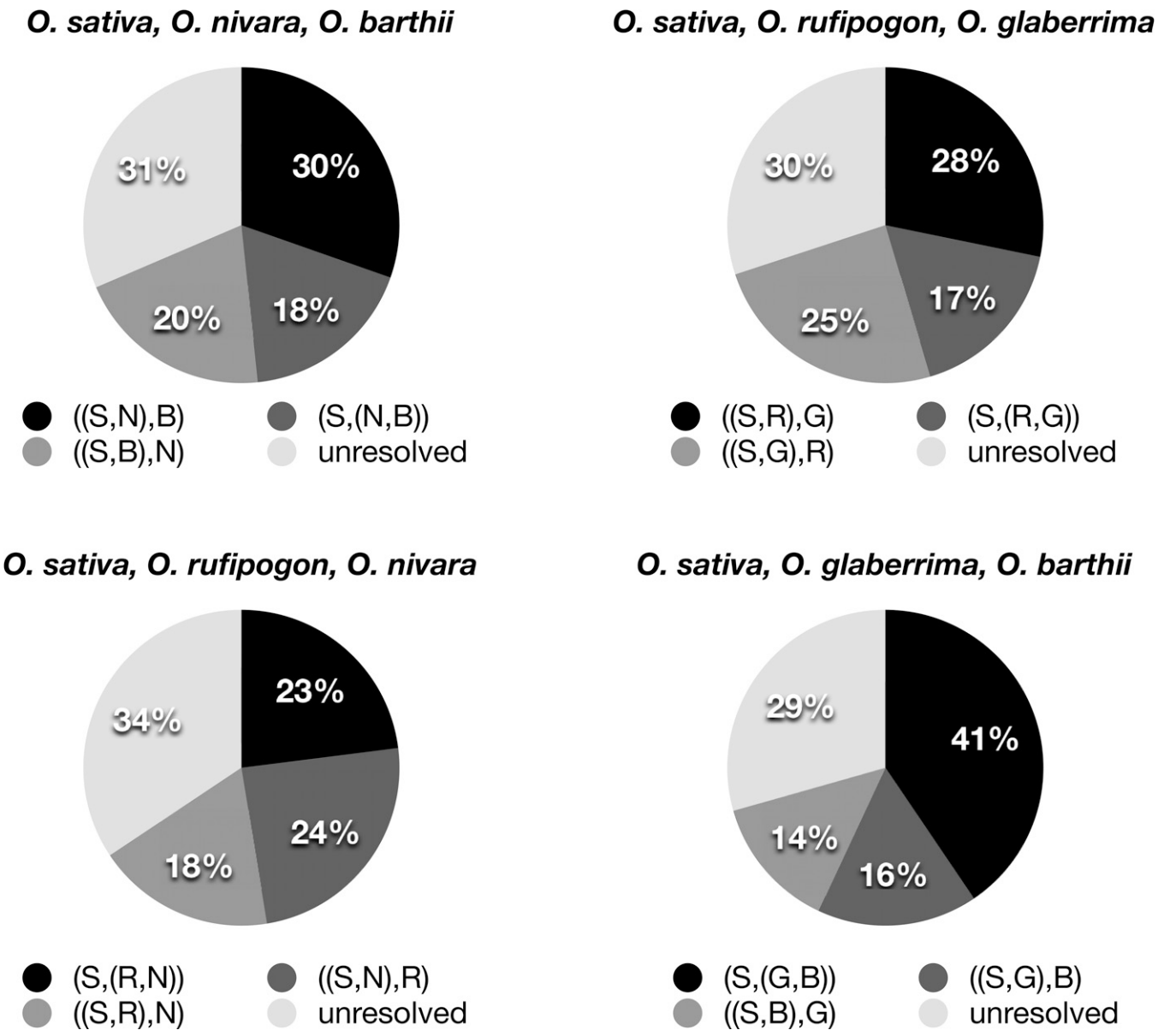

FIG. 6. Support for various triplets within the gene trees. In all charts, the lightest gray segment is the unresolved ('star') phylogeny and the black segment represents the triplet relationship where *O. sativa* is more closely related to an Asian species than to an African species. Species are S = *O. sativa*; N = *O. nivara*; R = *O. rufipogon*; G = *O. glaberrima*; B = *O. barthii*.

250 kb. This may not be sufficiently dense to detect blocks of incongruence gene trees that may instead be due to hybridization/introgression.

Discussion

A BAC-end Oryza alignment—We have demonstrated the utility of high-throughput sequences for phylogenetic analysis. Despite low overall coverage, the assembly of BES described here has provided the largest phylogenetic data set yet analyzed for *Oryza*. Our concatenated ORYZA alignment of ten species contains 2.45 million sites across 9,481 genes, and there are 1.2 million sites in the alignment of AA genome species. In comparison to other published phylogenomic analyses, this data set is similar in size to the one megabase eukaryotic alignment (Nishihara et al. 2007) and surpassed only by complete-genome analyses of 9,405 genes in *Drosophila*) (Pollard et al. 2006) and more than 122 million nucleotides in primates (Ebersberger et al. 2007).

As expected, the alignments constructed from BES are more sparse than those from more traditional phylogenetic sequencing protocols. Many genes are missing from many species. Even so, we have assembled 307 genes (82,000 sites) with sequence over each of the six AA genome species. This gene set is comparable to the recent analysis of 142 rice genes (Zou et al. 2008), which included a dense sampling alignment of 52 genes and 52,000 sites.

Phylogenetic Analysis—Elucidating the phylogeny of *Oryza* is a challenge even with data of this scale. The phylogenies produced from our large concatenated data sets support the monophyly of the AA genome species. This result is consistent with previous work based on both single and multiple

gene analyses (Ge et al. 1999, 2002; Zou et al. 2008). The topology within the AA genome clade is more uncertain. Previous studies show generally good support for distinct clades containing the Asian and African species, although this is complicated by poor resolution within these clades, low support values, lack of monophyly among multiple accessions of the same species and incongruence between phylogenies built using different genes (Zhu and Ge 2005; Duan et al. 2007; Zou et al. 2008).

In our analyses, MP analysis consistently places *O. sativa* as sister to *O. rufipogon* and *O. nivara*, the other two Asian AA genome species, with 100% bootstrap support. Initial likelihood-based analysis conflict on the placement of *O. sativa*, also supporting topologies that place *O. sativa* as the sister to the other four AA-genome species (Bayesian analysis of both large alignments and maximum likelihood analysis of the six-species alignment). The support values for the placement of *O. sativa* are weak in all likelihood-based analyses. The weak support is confirmed by results from SH tests of topology, which indicate that the maximum likelihood topology is not statistically superior to the second and third-best trees, which differ with respect to their placement of *O. sativa*.

The differences between the phylogenies inferred by ML and MP seem to be due to a combination of errors in modeling rate heterogeneity and treatment of missing data. The low level of sequence divergence translates to high heterogeneity of estimated rates across sites. Using a fine-grained discretization of the gamma distribution (by increasing the number of categories) gives improved likelihood scores, alters the relative order of likelihood scores for the two best topologies and reduces the differences in likelihood between these two topologies. It is unlikely that the differences between methodological treatments are due to artifacts such as base composition differences (Phillips et al. 2004) or long branch attraction (Felsenstein 1978). Composition differences were slight, and given the relatively recent time scale of diversification in *Oryza*, rate heterogeneity between lineages is unlikely to be sufficient to cause statistical inconsistency.

In general, our BAC-end data is consistent with previous studies that indicate that the evolution of the AA genome species is more complex than it is for *Oryza* as a whole. The strongest phylogenetic signal places *O. sativa* as the sister group to the Asian *O. rufipogon* and *O. nivara,* but alternate resolutions of these species do find support in some analyses and in some genomic regions.

***Missing Data*—**The use of BAC-end sequences induced a high level of missing data in the alignments and this data was unevenly distributed across the species. The results of phylogenetic inference were affected by the presence of missing data, specifically the location of the model organism *O. sativa* on the phylogeny and support for its placement. When we included all sites, *O. sativa* was weakly supported as sister to the other four AA genome species. Removing all sites with missing data produced a phylogeny where *O. sativa* was weakly supported as sister to only the other two Asian species (*O. rufipogon* and *O. nivara*). Simulating a random pattern of missing data in the alignment of 1.2 million nucleotides increases support for the three-species Asian clade to greater than 95% in both ML and Bayesian reconstructions. Based on these results, we hypothesize that obtaining a fully resolved and highly supported species tree may require nearly complete alignments of the length we have assembled here.

Previous studies on the effect of missing data on likelihood-based analyses have argued that the percent of missing data is not important, as long as the total amount of data is large (Wiens 2006; Philippe et al. 2004). In our alignment, it is the pattern of missing data across the species that seems to be affecting the phylogenetic reconstruction. This is despite having a large total number of sites. We have complete genes for *O. sativa* (i.e. no missing data), approximately 40–55% missing data for the other AA genome species and > 50% missing data for species of the other genome types. Although this uneven distribution of incomplete sites has been present in other studies, the effect may be different among closely related species than over a tree of greater taxonomic breadth (for example, the eukaryotic tree of Philippe et al. (2004)). The missing data in the AA-genome species without a whole genome sequence may be causing them to cluster in the phylogeny, especially in the presence of high rate heterogeneity (Lemmon et al. 2009). We note that missing data is less of a concern in the gene-level analysis, because when whole fragments were missing, we eliminate that species from the alignment and subsequent inference of the gene tree.

***Phylogenetic Incongruence*—**Discordance across the *Oryza* genome becomes more evident when we construct phylogenies on the scale of whole chromosomes or individual genes. The chromosome-level trees display statistically significant differences between chromosomes and between reconstruction methods. This is despite the fact that each chromosome is itself a large alignment on the scale of the phylogenomic data sets for yeast (127,000 sites; (Rokas et al. 2003)) or insects (101,000 sites; (Savard et al. 2006). Elimination of missing data, which reduced conflict in the trees from the concatenated alignment, did not reduce the chromosome-level incongruence, leading us to conclude that we are seeing differences in phylogenetic signal rather than systematic bias.

Gene trees displayed even more severe incongruence, supporting nearly half of the total number of possible trees in the six-species alignments. Across the shortest branch in the phylogeny, support for triplets of species roughly follows the 2:1:1 distribution suggested in cases of incomplete lineage sorting (Pamilo and Nei 1988). Within the three-species Asian clade of *O. sativa*, *O. nivara,* and *O. rufipogon*, the different resolutions were nearly equal, perhaps indicating the presence of other biological factors such as hybridization / introgression. Triplets describing the relationship of *O. sativa* with the two African species also showed some incongruence, with nearly one-fifth of the triplets pairing *O. sativa* with one of the African species. Overall, nearly one-quarter of the gene trees were completely unresolved, highlighting cases where our gene fragments were too short to be informative for species this closely related.

There is growing interest in phylogenetic analysis of independent gene trees (rather than simple concatentation) and subsequent use of gene tree discordance in the inference of a species tree. While the groundbreaking Rokas et al. (2003) phylogenomic data set for yeast species display little statistically significant incongruence (Jeffroy et al. 2006), *Oryza* is very different due to much more recent speciation and subsequent short branch lengths. Gene tree discordance can be due to a number of issues, including incomplete lineage sorting, gene duplication (if some genes are paralogs rather than orthologs) or introgression/hybridization. A challenge for low-level phylogenomic analyses of the type described here for *Oryza* data is to apply new methods developed to use

gene tree information, including some methods that explicitly model the process of lineage sorting as causing incongruence (Ané et al. 2007; Liu and Pearl 2007; Edwards et al. 2007; Mossel and Roch 2007). The surprisingly extensive fraction of the *Oryza* genome suggesting alternate phylogenetic relationships merits further study to determine its implications for the history of introgression, incomplete lineage sorting, and other processes in this species complex. Incomplete lineage sorting has been singled out as an important factor leading to incongruence at deeper nodes outside of the AA genome group (Zou et al. 2008). Within the AA genome, we have begun to explore some of the technical challenges and the biological results from methods that reconcile gene and species trees across the rice genome (Cranston et al. 2009).

BAC-end Phylogenomics—Traditionally, generation of data for phylogenetic analysis has been done through a priori selection and sequencing of complete genes for each species. This study instead examines the potential of high-throughput BAC-end sequencing for phylogenetic reconstruction. We have found both advantages and disadvantages of using these data for phylogenomics. On the positive side, large-scale sequencing projects provide a huge amount of data, creating alignments of a size rarely seen in phylogenetic studies. The sheer number of nucleotides available allows for the use of nuclear exons, even combined with a low level of sequence divergence. The existence of fully sequenced and annotated genomes for model organisms such as *O. sativa* permits cost-effective low-coverage sequencing of related species when there are insufficient resources to sequence full de novo genomes (as compared to phylogenomic analyses with yeast or *Drosophila*), where full genomes are available for a large number of closely related species). Having a model organism as an 'anchor' species also allows for simpler detection of orthologs and leveraging of information about genome location. The massive amount of genomic data allows for stringent filters for criteria such as sequence homology and overlap or presence of particular species, while still resulting in hundreds or thousands of genes appropriate for analysis. However, next-generation sequencing technologies, such as this BAC-end sequencing protocol, are often low coverage, producing a very large matrix of species by genes, but with many missing cells. These 'gene' level alignments are DNA fragments rather than full genes, and some fragments can be too short to be informative on their own. These issues of missing data are, of course, balanced by the large amount of total data.

As next-generation sequencing technologies develop and sequencing costs decrease, large-scale sequencing efforts for other genera are likely to produce data similar to our BAC-end sequence matrix for *Oryza*. These data will be characterized by a large number of sequences with genome-wide distribution, DNA fragments instead of full genes, sparse data matrices with high levels of missing data and a focus on coding sequences. This is the first study to focus on the use of BES for phylogenomics and highlights the potential for alignments built from high-throughput sequences to both identify incongruence and resolve troublesome nodes in the plant tree of life.

Acknowledgments. We thank Alexei Stamatakis for discussions about RaxML, and Derrick Zwickl for dicussions about Garli. We thank the OMAP team for insights on the construction and analysis of this data set. BH, LS, and DW gratefully acknowledge support from the National Science Foundation (NSF) Division of Biological Infrastructure grant #0321678 and DW acknowledges support from USDA-ARS. This work was also supported by NSF grants to MJS and NSF grants DBI-0321678 (to RAW and LS) and the Bud Antle Endowed Chair (to RAW).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412–426.

Bryant, D. 2003. A classification of consensus methods for phylogenetics, Pages 163–184 in *BioConsensus* (M. Janowitz, F. Lapointe, F. McMorris, B. Mirkin, and F. Roberts, eds.). DIMACS. Providence, Rhode Island: American Mathematical Society.

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

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540–552.

Ciccarelli, F., T. Doerks, C. von Mering, C. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283–1287.

Cranston, K. A., B. Hurwitz, D. Ware, L. Stein, and R. A. Wing. 2009. Species trees from highly incongruent gene trees in rice. *Systematic Biology* 58: 489–500.

Cronn, R., A. Liston, M. Parks, D. S. Gernandt, R. Shen, and T. Mockler. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36: e122.

de Queiroz, A. and J. Gatesy. 2007. The supermatrix approach to systematics. *Trends in Ecology & Evolution* 22: 34–41.

Doyle, J. J. 1992. Gene trees and species trees - molecular systematics as one-character taxonomy. *Systematic Botany* 17: 144–163.

Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, B. C. O'Meara, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. *Science* 306: 1172–1174.

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

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452: 745–749.

Ebersberger, I., P. Galgoczy, S. Taudien, S. Taenzer, M. Platzer, and A. von Haeseler. 2007. Mapping human genetic ancestry. *Molecular Biology and Evolution* 24: 2266–2276.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences USA* 104: 5936–5941.

Farris, J., M. Källersjö, A. Kluge, and C. Bult. 1995. Constructing a significance test for incongruence. *Systematic Biology* 44: 570–572.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401–410.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791.

Felsenstein, J. 2004. *Inferring phylogenies*. Sunderland: Sinauer Associates.

Garris, A. J., S. R. McCouch, and S. Kresovich. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). *Genetics* 165: 759–769.

Ge, S., A. Li, B. R. Lu, S. Z. Zhang, and D. Y. Hong. 2002. A phylogeny of the rice tribe *Oryzeae* (Poaceae) based on *matK* sequence data. *American Journal of Botany* 89: 1967–1972.

Ge, S., T. Sang, B. R. Lu, and D. Y. Hong. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proceedings of the National Academy of Sciences USA* 96: 14400–14405.

Goff, S., D. Ricke, T. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. M. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong,

T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W.-L. Sun, L. Chen, B. Cooper, S. Park, T. C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* 296: 92–100.

Guindon, S. and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22: 160–174.

Huber, K. T. and V. Moulton. 2006. Phylogenetic networks from multi-labelled trees. *Journal of Mathematical Biology* 52: 613–632.

Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles. *Genetics* 131: 509–513.

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

IRGSP. 2005. The map-based sequence of the rice genome. *Nature* 436: 793–800.

Ishii, T., Y. Xu, and S. McCouch. 2001. Nuclear- and chloroplast-microsatellite variation in A-genome species of rice. *Genome* 44: 658–666.

Jansen, R. K., L. A. Raubeson, J. L. Boore, C. W. DePamphilis, T. W. Chumley, R. C. Haberle, S. K. Wyman, A. J. Alverson, R. Peery, S. J. Herman, H. M. Fourcade, J. V. Kuehl, J. R. McNeal, J. Leebens-Mack, and L. Cui. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology* 395: 348–384.

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

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

Kullberg, M., B. M. Hallstrom, U. Arnason, and A. Janke. 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. *Zoologica Scripta* 37: 115–127.

Lee, Y., R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang, and J. Quackenbush. 2002. Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA). *Genome Research* 12: 493–502.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. M. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology* 58: 130–145.

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

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

Maddison, W. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Mather, K. A., A. L. Caicedo, N. R. Polato, K. M. Olsen, S. McCouch, and M. D. Purugganan. 2007. The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* 177: 2223–2232.

McMahon, M. M. and M. J. Sanderson. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 Papilionoid legumes. *Systematic Biology* 55: 818–836.

Moore, M. J., A. Dhingra, P. S. Soltis, R. Shaw, W. G. Farmerie, K. M. Folta, and D. E. Soltis. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology* 6: 17.

Mossel, E. and S. Roch. 2007. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. http://arxiv.org/pdf/0710.0262.

Nishihara, H., N. Okada, and M. Hasegawa. 2007. Rooting the eutherian tree: The power and pitfalls of phylogenomics. *Genome Biology* 8: R199.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302: 205–217.

Olsen, K. M., A. L. Caicedo, N. Polato, A. McClung, S. McCouch, and M. D. Purugganan. 2006. Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics* 173: 975–983.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* 5: 568–583.

Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005. Phylogenomics. *Annual Review of Ecology Evolution and Systematics* 36: 541–562.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. H. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Molecular Biology and Evolution* 21: 1740–1752.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* 21: 1455–1458.

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

Ren, F., B. Lu, S. Li, J. Huang, and Y. Zhu. 2003. A comparative study of genetic relationships among the AA-genome *Oryza* species using RAPD and SSR markers. *Theoretical and Applied Genetics* 108: 113–120.

Robbertse, B., J. B. Reeves, C. L. Schoch, and J. W. Spatafora. 2006. A phylogenomic analysis of the Ascomycota. *Fungal Genetics and Biology* 43: 715–725.

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

Sanderson, M. J., C. Ané, O. Eulenstein, D. Fernandez-Baca, J. Kim, M. M. McMahon, and R. Piaggio-Talice. 2007. Fragmentation of large data sets in phylogenetic analysis. in *Reconstructing evolution: new mathematical and computational advances*, eds. O. Gascuel and M. Steel. Oxford: Oxford University Press.

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

Sang, T. and S. Ge. 2007. The puzzle of rice domestication. *Journal of Integrative Plant Biology* 49: 760–768.

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

Semon, M., R. Nielsen, M. P. Jones, and S. R. McCouch. 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.

Shimodaira, H. and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* 16: 1114–1116.

Shimodaira, H. and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17: 1246–1247.

Soltis, D. E., P. S. Soltis, M. D. Bennett, and I. J. Leitch. 2003. Evolution of genome size in the angiosperms. *American Journal of Botany* 90: 1596–1603.

Stamatakis, A. 2006. Phylogenetic models of rate heterogeneity: A high performance computing perspective. *in Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium*. Rhodes, Greece: IEEE.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456–463.

Strimmer, K. and A. Rambaut. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society B: Biological Sciences* 269: 137–142.

Swofford, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). version 4. beta 10. Sunderland: Sinauer Associates.

Tateoka, T. 1962. Taxonomic studies of *Oryza* I. *O. latifolia* complex. *The Botanical Magazine, Tokyo* 75: 418–427.

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

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

Vaughan, D. A., H. Morishima, and K. Kadowaki. 2003. Diversity in the *Oryza* genus. *Current Opinion in Plant Biology* 6: 139–146.

Wang, X. Y., X. L. Shi, B. L. Hao, S. Ge, and J. C. Luo. 2005. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *The New Phytologist* 165: 937–946.

Wiens, J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology* 47: 625–640.

Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics* 39: 34–42.

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

Yu, J., S. Hu, J. Wang, G. K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, X. Huang, W. Li, J. Li, Z. Liu, L. Li, J. Liu, Q. Qi, J. Liu, L. Li, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Zhang, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Ren, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, J. Wang, W. Zhao, P. Li, W. Chen, X. Wang, Y. Zhang, J. Hu, J. Wang, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, G. Li, S. Liu, M. Tao, J. Wang, L. Zhu, L. Yuan, and H. Yang. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92.

Zhang, L. B. and S. Ge. 2007. Multilocus analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. *Molecular Biology and Evolution* 24: 769–783.

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

Zhu, Q. H., X. M. Zheng, J. C. Luo, B. S. Gaut, and S. Ge. 2007. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: Severe bottleneck during domestication of rice. *Molecular Biology and Evolution* 24: 875–888.

Zou, X.-H., F.-M. Zhang, J.-G. Zhang, L.-L. Zang, L. Tang, J. Wang, T. Sang, and S. Ge. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biology* 9: R49.

Zwickl, D. J. 2006. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Ph. D. Thesis. Austin, Texas: The University of Texas at Austin.