JSE Journal of Systematics and Evolution

Journal of Systematics and Evolution 49 (5): 386–395 (2011) doi: 10.1111/j.1759-6831.2011.00145.x

Research Article

Phylogeny and species delimitation of the C-genome diploid species in *Oryza*

[1,2]Li-Li ZANG [1]Xin-Hui ZOU [1]Fu-Min ZHANG [3,4]Ziheng YANG [1,2]Song GE*

[1](*State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences,* Beijing 100093, China)

[2](*Graduate University of the Chinese Academy of Sciences,* Beijing 100049, China)

[3](*Department of Biology,University College London,* London WC1E 6BT, United Kingdom)

[4](*Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences,* Beijing 100101, China)

Abstract The diploid *Oryza* species with C-genome type possesses abundant genes useful for rice improvement and provides parental donors of many tetraploid species with the C-genome (BBCC, CCDD). Despite extensive studies, the phylogenetic relationship among the C-genome species and the taxonomic status of some taxa remain controversial. In this study, we reconstructed the phylogeny of three diploid species with C-genome (*Oryza officinalis*, *O. rhizomatis*, and *O. eichingeri*) based on sequences of 68 nuclear single-copy genes. We obtained a fully resolved phylogenetic tree, clearly indicating the sister relationship of *O. officinalis* and *O. rhizomatis*, with *O. eichingeri* being the more divergent lineage. Incongruent phylogenies of the C-genome species found in previous studies might result from lineage sorting, introgression/hybridization and limited number of genetic markers used. We further applied a recently developed Bayesian species delimitation method to investigate the species status of the Sri Lankan and African *O. eichingeri*. Analyses of two datasets (68 genes with a single sample, and 10 genes with multiple samples) support the distinct species status of the Sri Lankan and African *O. eichingeri*. In addition, we evaluated the impact of the number of sampled individuals and loci on species delimitation. Our simulation suggests that sampling multiple individuals is critically important for species delimitation, particularly for closely related species.

Key words Bayesian species delimitation, *Oryza*, phylogeny, taxonomy.

The genus *Oryza* consists of the cultivated rice (*O. sativa* L.) and additional 22 wild species and is represented by six diploid (A-, B-, C-, E-, F-, G-genome) and four tetraploid (BC-, CD-, HJ-, HK-genome) groups (Ge et al., 1999). Rice and its relatives not only provide an enormous gene pool for genetic improvement of rice cultivars but also offer a good model system for studying many intriguing biological questions involving comparative and functional genomics, polyploid evolution, speciation and biogeography, as well as ecological adaptation and domestication (Ge et al., 1999; Wing et al., 2005; Sang & Ge, 2007; Ammiraju et al., 2010; Zheng & Ge, 2010).

In the rice genus, three diploid species with the C-genome type (*O. officinalis* Wall. ex Watt., *O. rhizomatis* Vaughan, and *O. eichingeri* Peter) comprise the core species of the *O. officinalis* complex (Vaughan, 1989). Geographically, *O. officinalis* is distributed widely in South and Southeast Asia as well as northern Australia and Papua New Guinea; *O. rhizomatis* has only been reported from Sri Lanka, and *O. eichingeri* is distributed both in Sri Lanka and West and East Africa (Vaughan et al., 2003; Zhang & Ge, 2007) (Fig. 1). These species not only possess valuable genes useful for rice improvement (Brar & Khush, 1997) but also provide parental donors to many tetraploids with the C-genome type (BBCC, CCDD) (Tateoka, 1965; Vaughan, 1989) and thus have been extensively investigated (see reviews in Nayar, 1973; Vaughan et al., 2003; Zhang & Ge, 2007). To date, a few studies have attempted to reveal the species relationship among the C-genome species but obtained inconsistent results (Shcherban et al., 2000a, 2000b; Bao & Ge, 2004; Bao et al., 2006; Bautista et al., 2006; Zhang & Ge, 2007; Wang et al., 2009). Based on restriction fragment length polymorphism and sequence analyses of the integrase coding domain of a gypsy-like retrotransposon, Shcherban et al. (2000a, 2000b) showed that *O. eichingeri* was the divergent lineage, whereas *O. officinalis* and *O. rhizomatis* were sister

Received: 3 May 2011 Accepted: 26 May 2011

* Author for correspondence. E-mail: gesong@ibcas.ac.cn; Tel.: 86-10-62836097; Fax: 86-10-62590843.

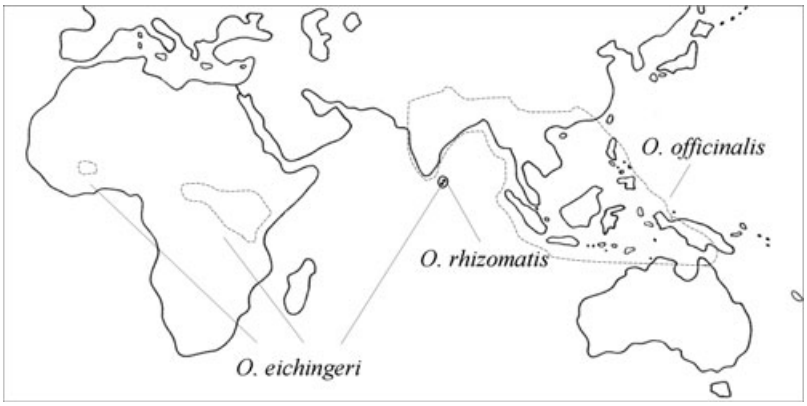

Fig. 1. Geographical distribution of *Oryza* species with the C-genome. The outlined areas are the distribution range of the three species.

groups. This result was confirmed by Bao & Ge (2004) and Wang et al. (2009) using sequences of multiple nuclear and chloroplast genes. However, other studies (Bao et al., 2006; Bautista et al., 2006; Zhang & Ge, 2007) found that *O. eichingeri* was genetically more similar to *O. rhizomatis* than to *O. officinalis*. Thus, the species tree of the diploid C-genome species remains elusive. With accumulation of sequence data, it has been a widespread practice to use multiple genes to resolve the conflicting gene trees at different hierarchical levels (Rannala & Yang, 2008; Zou & Ge, 2008). In our previous study on phylogenetic reconstruction of the diploid genomes in *Oryza* (Zou et al., 2008), we sequenced 142 single-copy genes and clarified the relationships among all diploid genome types of the rice genus, demonstrating the power of phylogenomics in the reconstruction of rapid diversification. In this study, we expanded the dataset used in Zou et al. (2008) by sequencing a set of genes from additional samples with the hope to resolve the species relationship among three C-genome diploid species.

The second goal of this study is to investigate the species status of two geographical races of *O. eichingeri*. This species is particularly interesting and its taxonomic status has been a focus of considerable debate because it is the only *Oryza* species that occurs in both Africa and Asia (Nayar, 1973; Vaughan, 1989; Vaughan et al., 2003; Zhang & Ge, 2007) (Fig. 1). Previous investigations suggested that *O. eichingeri* from two continents showed sufficient differentiation and should be treated as two species (Sharma & Shastry, 1965; Federici et al., 2002). Based on molecular population genetics study, however, Zhang & Ge (2007) indicated that a long-distance dispersal from West Africa to Sri Lanka was more likely to play a role in the disjunctive distribution of *O. eichingeri* and thus suggested that they should be treated as geographic races rather than distinct species. Traditionally a species is typically identified based on presence of fixed morphological characters (Wiens & Servedio, 2000). However, morphological characters are affected by many factors such as the environment and sampling and often lead to inaccurate species classification (Davis & Nixon, 1992; Padial et al., 2010). Despite many methods available (Godfray, 2002; Tautz et al., 2003; de Queiroz, 2007; Knapp et al., 2007; Wiens, 2007), species delimitation, particularly for recently diverged species, is still a challenge. Recently, Yang & Rannala (2010) developed a Bayesian model for using multilocus data to delimitate closely related or recently diverged species. This method has been evaluated with simulated datasets (Yang & Rannala, 2010; Zhang et al., 2011) and applied to several empirical datasets of rotifers, lizards, and humans (Yang & Rannala, 2010), forest geckos (Leache & Fujita, 2010), and butterfly (Zhang et al., 2011). Nevertheless, empirical studies on plant species are lacking and the utility of the method needs to be validated with more empirical datasets. The Asian and African populations of *O. eichingeri* provide an ideal example for an empirical test of the species delimitation method.

Here, we reconstruct the phylogenetic relationships of the C-genome diploid species using sequences from 68 nuclear single-copy genes. We obtained a highly resolved phylogenetic tree, clearly indicating the sister relationship of *O. officinalis* and *O. rhizomatis*. Then we used two types of datasets to apply the Bayesian species delimitation method by Yang & Rannala (2010) to investigate the species status of the Sri Lankan and African *O. eichingeri*. The results support the distinct species status of the Sri Lankan and African populations of *O. eichingeri*. In addition, the two datasets allow us to evaluate the impact of the number of sampled individuals and the number of loci on species delimitation by the Bayesian method.

Table 1 Species and geographic origins of the individuals sampled

Species	Accession†	Source
Oryza officinalis (O)	101412	India
	102460	Bangladesh
	105081	Myanmar
	7904, 104972‡	China
	81972	Thailand
	105080	Vietnam
	105093	Malaysia
	81796	Indonesia
	105100	Brunei
	105085	Philippines
	106519	Papua New Guinea
	106522	Papua New Guinea
O. eichingeri (S)	81803	Sri Lanka
	105407	Sri Lanka
	105413	Sri Lanka
	105415‡	Sri Lanka
O. eichingeri (U)	101425	Uganda
	105159§	Uganda
	105162	Uganda
	IP7	Cote d'Ivoire
O. rhizomatis (R)	103410‡	Sri Lanka
	103421	Sri Lanka
	105448	Sri Lanka
	105950	Sri Lanka
O. punctata	103903‡	Tanzania

†All accessions were obtained from leaf materials or seeds provided by the Genetic Resources Center of the International Rice Research Institute (IRRI) at Los Banos, Philippines, except for 7904, which was collected by the authors, and IP7, which was provided by Dr. G. Second (France). ‡Samples used for phylogenetic reconstruction. §Sample sequenced in this study.

1 Material and methods

1.1 Plant materials and sequencing

Zou et al. (2008) sequenced at least 62 genes from the representatives of all the diploid genomes including three C-genome diploid species. Of these, 58 were selected for use in the present study. Zhang & Ge (2007) also sequenced 10 nuclear single-copy genes from 12 individuals of *O. officinalis*, four individuals of *O. rhizomatis*, and four each of the Sri Lankan and Ugandan *O. eichingeri*. All 10 genes were used in this study. Therefore, we retrieved and analyzed the sequences of 68 genes from three C-genome species (*O. officinalis*, *O. rhizomatis*, *O. eichingeri* from Sri Lanka) and one B-genome species (*O. punctata* Kotschy ex Steud) (Table 1). In addition, we sequenced the 58 genes used in Zou et al. (2008) for one *O. eichingeri* sample from Uganda (Table 1) because this sample was not sequenced in Zou et al. (2008). DNA extraction, PCR amplifications, and purification of the products were carried out by conventional methods. Primers used for amplifying and sequencing these 58 genes of the Ugandan *O. eichingeri* were from Zou et al. (2008). When direct sequencing failed, we used the pGEM-T Easy Vector (Promega, Madison, WI, USA) to clone and sequence. Sequencing was carried out on an ABI automated sequencer (Applied Biosystems, Foster City, CA, USA). All the sequences obtained in this study have been deposited in the GenBank database (Accession Nos JN258623-JN258680).

In the species delimitation study, we used two types of datasets to run the reversible jump Markov Chain Monte Carlo (rjMCMC) algorithms (Yang & Rannala, 2010): (i) dataset A consists of sequences of 10 genes from the population samples of Zhang & Ge (2007); and (ii) dataset B consists of sequences of 68 genes from a single individual for each species/population.

1.2 Phylogenetic analyses

Phylogenetic trees were reconstructed using maximum likelihood (ML), maximum parsimony (MP) implemented in PAUP* version 4.0b10 (Swofford, 2003), and Bayesian inference (BI) performed with MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003). In ML and MP analyses, the branch-and-bound algorithm was used for tree searching and the tree reliability was assessed with a non-parametric bootstrap strategy (Felsenstein, 1985), with 1000 replicates for MP and 500 replicates for ML. In BI analyses, four independent MCMC runs were carried out, each starting with randomly choosing topologies for the four simultaneous chains, one cold and three incrementally heated chains. The four chains were run for at least one million generations, with samples taken at every 1000 generations, with the first 25% of samples discarded as burn-in. Models were selected by Modeltest 3.7 (Posada & Crandall, 1998). The homogeneity across gene fragments was tested using the incongruence length difference (ILD) test (Farris et al., 1994), as implemented in PAUP. We reconstructed phylogenies both for each of the 68 genes separately and for the concatenated data matrix. We also carried out a new Bayesian analysis (Liu & Pearl, 2007) which incorporated coalescent theory to account for lineage sorting during rapid speciation. This method, called Bayesian estimation of species trees (BEST), uses the gene tree distributions to reconstruct posterior distributions of a species tree (Edwards et al., 2007; Liu & Pearl, 2007). Four MCMC chains were run simultaneously for 10 000 000 generations with a burn-in period of 30 000. Each analysis was carried out at least twice with different starting seeds. In all phylogenetic reconstruction, one accession of the B-genome *O. punctata* (Table 1) was used as the outgroup because the B-genome was sister to the C-genome in the genus (Ge et al., 1999; Zou et al., 2008).

1.3 Bayesian species delimitation

Bayesian species delimitation was carried out using the program Bayesian phylogenetics and

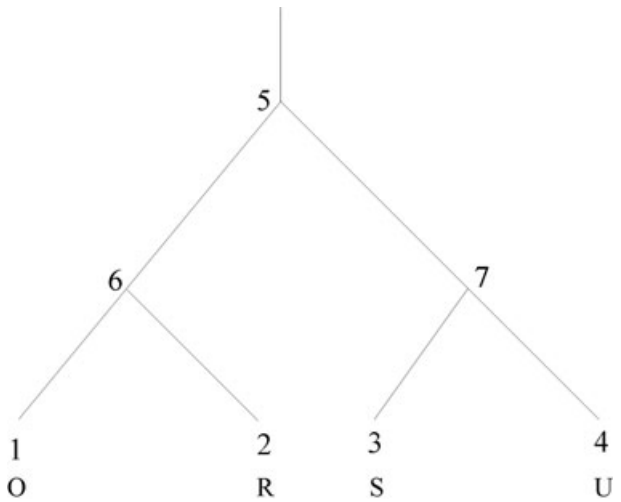

Fig. 2. Guide tree used in Bayesian species delimitation. There are seven species/populations of *Oryza*, including 1, 2, 3, and 4 that are extant, and 6 and 7 that are ancestors of 1 and 2, and 3 and 4, respectively. Species 5 is the ancestor of all four species/populations. A species-delimitation model is represented by flags (0 or 1) on the internal nodes 5, 6, and 7 on the guide tree, so that 000 means all three internal nodes are collapsed and that is only one species, whereas 111 means all three internal nodes exist and represent species divergence so that there are four distinct species in the model. This study mainly examines Pr(111), the posterior probability that there are four distinct species. Four species/populations were simplified as O (*O. officinalis*), R (*O. rhizomatis*), S (*O. eichingeri* from Sri Lanka), and U (*O. eichingeri* from Uganda).

phylogeography (BPP; Rannala & Yang, 2003; Yang & Rannala, 2010). The program requires a guide tree, as well as specification of prior distributions of ancestral effective population size (θ_0), and the root age (τ_0). It uses multiple-gene data to estimate the posterior probabilities for different species delimitation models that are compatible with the guide tree. Based on the nucleotide diversity of extant populations estimated in Zhang & Ge (2007), we used a gamma prior distribution G (1, 250) for population size parameter (θ_0), with mean $1/250 = 0.004$ nucleotide differences per site and a gamma prior G (1, 250) for the age of the root in the species tree again with the mean $1/250 = 0.004$ mutations per site. We also varied the priors to examine their impact (see below). Each analysis was run more than twice to confirm consistency between runs. Yang & Rannala (2010) indicated that, for a large number of loci, the rjMCMC may have some mixing problems, having difficulty moving between species-tree models. Thus, as suggested, we used the τ-threshold method that does not use rjMCMC (Yang & Rannala, 2010) on a fixed species tree. This method also involves the same parameters as BPP, and the posterior probability P is interpreted as the probability that two groups form a single species. Figure 2 shows the guide tree. When running BPP under that guide tree, the posterior distribution indicates that all the three internal nodes 5, 6, and 7 represent speciation events so that *O. officinalis* (O), *O. rhizomatis* (R), Sri Lankan *O. eichingeri* (S), and Ugandan *O. eichingeri* (U) are four distinct species.

To examine the impact of the number of genes and the sample size on the results, we drew random samples of different sizes from datasets A and B without replacement. For dataset A, we generated samples consisting of 2, 5, and 8 genes, with 10 replications in each case. For dataset B, we generated 10, 20, 30. . . 60 genes, with 100 replicates for each case.

1.4 Impact of guide trees and priors

In species delimitation, the guide phylogeny and prior distributions are the most important factors that affect the posterior probabilities (Leache & Fujita, 2010; Yang & Rannala, 2010; Zhang et al., 2011). In the study on forest geckos, Leache & Fujita (2010) suggested that a misspecified guide tree may lead to incorrect results, causing the method to split species. Here, we used two alternative guide trees to examine the impact of the guide tree.

We evaluated the effect of the priors by considering three different sets of priors. The first set of priors was based on $\theta_0 \sim G(1, 250)$ and $\tau_0 \sim G(1, 250)$, both with a prior mean 0.004 and variance 1.6×10^{-5}. The second set of priors assumed larger ancestral population sizes and deeper divergences: $\theta_0 \sim G(1, 25)$ and $\tau_0 \sim G(1, 25)$ with the prior mean 0.04 and variance 0.0016. The last set of priors assumed smaller ancestral population sizes and shallower divergence times: with $\theta_0 \sim G(1, 2500)$ and $\tau_0 \sim G(1, 2500)$ with the prior mean 0.0004 and variance 1.6×10^{-7}. Note that those priors have the same shape but the prior means are two orders of magnitude difference. The first set was applied to all the runs, and the other two were used in only a few analyses for comparison.

2 Results

2.1 Phylogeny inferred from 68 nuclear single-copy genes

The 68 nuclear genes are distributed throughout the 12 rice chromosomes. To examine the suitability of combining the genes, we applied the incongruence length difference test and found no significant incongruence ($P = 0.106$). After removing regions with ambiguous alignment, we concatenated the 68 genes into a data matrix of 59 390 bp, with exons accounting for 38.16% (22 663 bp). Of them, 2829 sites (4.76%) were

Table 2 Details of exon, intron, and third codon position datasets, as well as all sites combined, and results obtained using maximum likelihood (ML), maximum parsimony (ML), Bayesian inference (BI), and Bayesian estimation of species trees (BEST) analyses

Dataset	No. of sites (bp)	No. of variable sites (%)	No. of informative sites (%)	Branch support (ML/MP/BI/BEST)		
				((O,R)(S,U))	(O,R)	(S,U)
All sites	59390	2829	236	100/100/1.00/1.00	100/100/1.00/1.00	100/100/1.00/1.00
Intron	36727	2345	200	100/100/1.00/1.00	100/100/1.00/1.00	100/100/1.00/1.00
Exon	22663	484	36	100/100/1.00/1.00	100/100/1.00/1.00	100/100/1.00/1.00
Third codon	6442	367	27	100/100/1.00/1.00	100/100/1.00/1.00	100/100/1.00/1.00

O, *Oryza officinalis*; R, *O. rhizomatis*; S, *O. eichingeri* from Sri Lanka; U, *O. eichingeri* from Uganda. The BEST analysis applied according to Liu & Pearl, 2007.

variable including 236 (0.4%) parsimony-informative sites (Table 2). Phylogenetic reconstruction of the concatenated data using ML, MP, and BI yielded a single fully resolved tree with high bootstrap support (100%) or Bayesian posterior probability (PP = 1.0) for all internal branches (Fig. 3). It is evident that *O. rhizomatis* and *O. officinalis* are highly supported monophyletic groups and *O. eichingeri* is sister to the *O. rhizomatis*–*O. officinalis* clade. The Sri Lankan and Ugandan *O. eichingeri* populations form a highly supported monophyly, and show a level of divergence comparable to that between *O. rhizomatis* and *O. officinalis*, implying their high level of genetic differentiation.

Because different genome regions and different sites in the sequence are expected to be under different selective pressures and have different rates of evolution, we carried out further phylogenetic analyses based on the combined datasets of exon, intron, and the third codon positions, and obtained the same topology with 100% bootstrap support or 1.0 Bayesian posterior probabilities for all internal branches (Table 2). Using the BEST method that accounts for lineage sorting, we obtained the same species tree with posterior probability of 1.0, suggesting that lineage sorting had not biased the phylogenetic reconstructions in multilocus analyses.

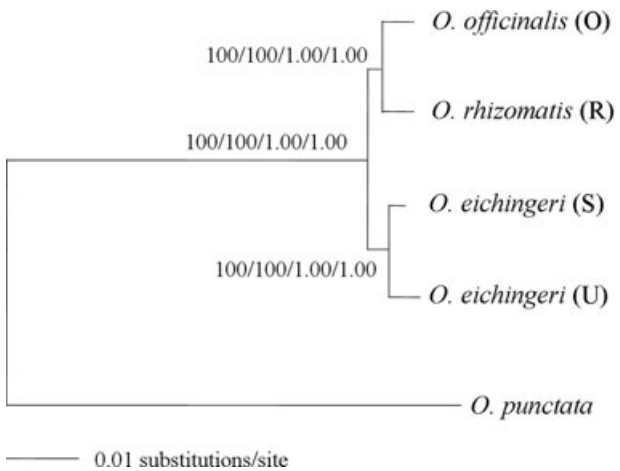

Fig. 3. Maximum likelihood tree generated from the concatenated sequences of 68 nuclear genes under the TVM+I model. The same topology was obtained from maximum parsimony, Bayesian inference, and the Bayesian Estimation of Species Trees method. Numbers above branches indicate bootstrap support of maximum likelihood and maximum parsimony, posterior probability of Bayesian inference and Bayesian Estimation of Species Trees, respectively.

2.2 Bayesian species delimitation

We applied the rjMCMC method to analyze the datasets A and B, using the prior $\theta_0 \sim G\,(1, 250)$ and $\tau_0 \sim G\,(1, 250)$. We focused on the two populations (the Sri Lankan and Ugandan) of the species *O. eichingeri* because they are distributed disjunctively on two continents and their species status has been a matter of debate. As shown in Table 3, analysis of dataset A supported the fully resolved tree model, with four distinct species for all four lineages. The posterior probability was 1.0 if all genes were used. However, analysis of dataset B generated the posterior tree probability Pr(111) = 0.209, i.e., Pr(110) = 0.791, with weak support for grouping S and U into one species. If we carried out the analysis using single locus of dataset A, tree 111 reaches high posterior probability (Pr ∼ 1) (Table S1). These results suggest that dataset A, with multiple individuals sampled from the same species/population, is more informative than dataset B, in which only one individual was sampled for each species/population.

To examine the correlation between the number of genes and the posterior probability, we drew random samples for each dataset and calculated the mean posterior probability of the species-tree models. As shown in Fig. 4, for dataset A, Pr(111) increased with the gene number, whereas in dataset B, the posterior probability fluctuated (see also Table S2).

To assess how many samples support four distinct species effectively, we examined two combinations of sampling loci and individuals: (i) sampling 2, 5, 8 loci, respectively for two individuals; and (ii) sampling 2, 5, 8 loci of four individuals for each of four species/populations. In each case, 10 replicates were generated. We found that Pr(111) varied from 0.892 to 1.0 in the case of four individuals. For the case of two individuals, however, Pr(111) was much lower and ranged between 0.391 and 0.592. These results suggest that it

Table 3 Bayesian species delimitation to investigate the species status of the Sri Lankan and African *Oryza eichingeri*. Results are based on datasets A, comprising 68 genes with a single sample, and B, comprising 10 genes with multiple samples

Dataset	Guide tree			Misspecified guide tree $\theta_0 = \tau_0 = 0.004$	τ-threshod method $\theta_0 = \tau_0 = 0.004$
	$\theta_0 = \tau_0 = 0.004$	$\theta_0 = \tau_0 = 0.04$	$\theta_0 = \tau_0 = 0.0004$		
A	1	1	1	1	0.9908
B	0.2089	0.1502	0.9451	0.9927	0.7436

The figures represent Pr(111), the posterior probability of tree 111, which assumes that the four populations, *O. officinalis*, *O. rhizomatis*, *O. eichingeri* from Sri Lanka, and *O. eichingeri* from Uganda, are four distinct species. The first column used the prior distribution $\theta_0 \sim G$ (1, 250) and $\tau_0 \sim G$ (1, 250) (a mean of 0.004); the second and third columns used the distribution $\theta_0 \sim G$ (1, 25) and $\tau_0 \sim G$ (1, 25), $\theta_0 \sim G$ (1, 2500) and $\tau_0 \sim G$ (1, 2500), respectively. Two control analyses under the misspecified guide tree and one species model used prior distribution $\theta_0 \sim G$ (1, 250) and $\tau_0 \sim G$ (1, 250).

is important to sample multiple individuals for accurate species delimitation.

2.3 Impact of prior distribution and guide tree

Previous studies suggested that when using rjMCMC, the prior distribution for θ_0 has a large impact on species delimitation (Leache & Fujita, 2010). Therefore, we applied two priors besides the original $\theta_0 \sim G$ (1, 250) and $\tau_0 \sim G$ (1, 250), varying those parameters by two orders of magnitude. Under a larger ancestral population size and deeper divergent time: $\theta_0 \sim G$ (1, 25) and $\tau_0 \sim G$ (1, 25), with the mean 0.04 and

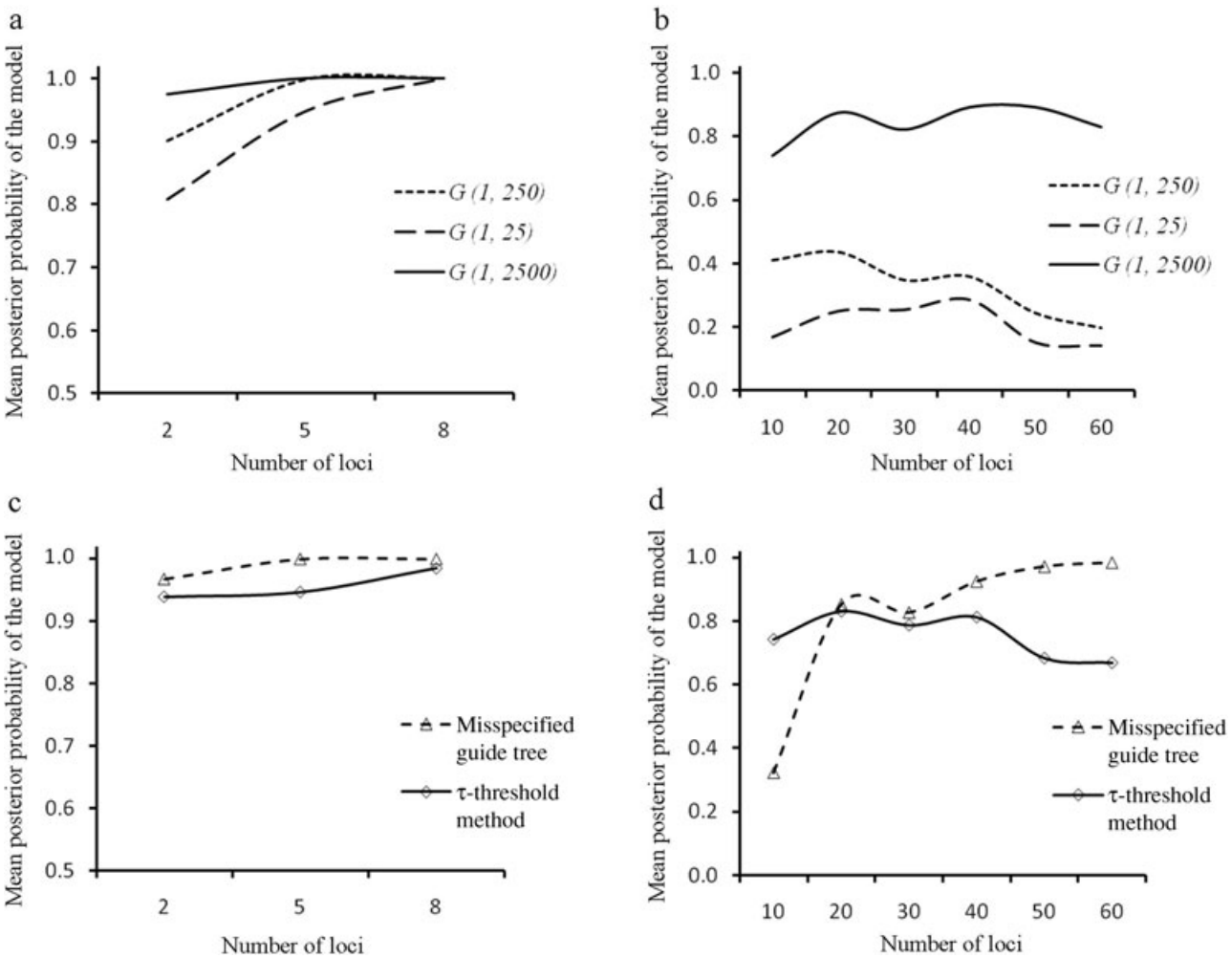

Fig. 4. Mean posterior probability Pr(111) across replicate datasets as a function of the number of randomly selected loci for datasets A and B under different conditions: three different priors, a misspecified guide tree, and τ-threshold method.

variance 0.0016, the posterior probability for the tree 111 decreased slightly for dataset A (Fig. 4: a) and decreased substantially for dataset B (Fig. 4: c). Another prior assumed smaller ancestral population size and shorter divergence time: $\theta_0 \sim G$ (1, 2500) and $\tau_0 \sim G$ (1, 2500) with the prior mean 0.0004 and variance 1.6×10^{-7}. In this case, Pr(111) increased slightly for dataset A (Fig. 4: a) but increased two-fold for dataset B (Fig. 4: c).

A misspecified guide tree could have a significant impact on Bayesian species delimitation (Leache & Fujita, 2010). We analyzed datasets A and B using a misspecified guide tree ((O,S),(R,U)) and found that both datasets strongly supported the distinct species status of O, R, S, and U, with the posterior probabilities close to 1.0 (Table 3). It is evident that under a misspecified guide tree, dataset A generated stable posterior probabilities with a different number of genes sampled (Fig. 4: b); the posterior probabilities for dataset B fluctuated as the number of sampled genes increased (Fig. 4: d). Use of other misspecified guide trees produced very similar results (data not shown).

We further analyzed the two datasets using the τ threshold method running the MCMC on the four-species tree. When using the original $\theta_0 \sim G$ (1, 250) and $\tau_0 \sim G$ (1, 250), the results showed that the posterior probabilities of tree 111 for datasets A and B were very high (0.9908 and 0.7436, respectively), supporting the Sri Lankan *O. eichingeri* (S) and Ugandan *O. eichingeri* (U) to be distinct species. Under the priors $\theta_0 \sim G$ (1, 25) and $\tau_0 \sim G$ (1, 25), the posterior probability of tree 111 was stable for dataset A (0.9887) but decreased slightly for dataset B (0.6734). Under the priors $\theta_0 \sim G$ (1, 2500) and $\tau_0 \sim G$ (1, 2500), we obtained similar posterior probability for dataset A (0.9943) and high posterior probability for dataset B (0.9928). Similarly, the posterior probability for dataset A was not sensitive to the number of genes sampled, but that for dataset B increased with the number of genes sampled (Fig. 4: b, d).

3 Discussion

3.1 Phylogenetic relationship of the C-genome diploid species in *Oryza*

The relationships among the three diploid species with the C-genome type have been investigated intensively using different genetic markers. Different results were produced, favoring either the relationship between *O. officinalis* and *O. rhizomatis* (Shcherban et al., 2000a, 2000b; Bao & Ge, 2004; Zou et al., 2008; Wang et al., 2009) or between *O. eichingeri* and *O. rhizomatis* (Bao et al., 2006; Bautista et al., 2006; Zhang & Ge, 2007). With a much larger dataset of 68 genes and the inclusion of additional samples of *O. eichingeri* from Africa, we obtained a fully resolved phylogeny of the C-genome diploid species. Analyses with different phylogenetic approaches strongly suggest that *O. officinalis* and *O. rhizomatis* are most closely related and *O. eichingeri* is the outgroup, which is further divided into two highly differentiated geographical races/populations occurring in Africa and Sri Lanka.

The incongruent phylogenies of the C-genome species in previous studies raise an interesting question regarding the species tree and gene tree conflicts. Several factors might explain the conflicts. First, lineage sorting arising from ancient polymorphism and rapid speciation may cause incongruent phylogenies across genes (Rannala & Yang, 2003). Using the sequences of 10 nuclear single-copy genes, Zhang & Ge (2007) found incongruent topologies among 10 gene trees regarding the position of three C-genome diploid species and estimated that two speciation events leading to the three species happened at such a short time interval (~0.63–0.68 million years) that the polymorphism in the ancestral population of these species could persist easily from the first divergence to the second. Rapid evolutionary radiations have been suggested to be the most plausible explanation for conflicting gene trees in many plant and animal species, particularly at lower taxonomic ranks (see reviews in Rannala & Yang, 2008; Zou et al., 2008).

The second reason for incongruent gene phylogenies among studies is hybridization/introgression between *O. eichingeri* and *O. rhizomatis*. These two species are sympatric in Sri Lanka, although their habitats are slightly different (Vaughan et al., 2003), and there is evidence that in Sri Lanka, the two species hybridize frequently (Bautista et al., 2006; Zhang & Ge, 2007). Finally, the limited number of genetic markers used in most of the previous studies might also have caused the inconsistent results, especially as the two speciation events occurred close in time (Rannala & Yang, 2008; Zou et al., 2008). The present study used a combined analysis of 68 genes and thus overcame the noises of ancient polymorphisms to obtain a fully resolved phylogenetic relationship.

3.2 Taxonomic status of the African and Sri Lankan *Oryza eichingeri*

Oryza eichingeri is the only wild *Oryza* species that is distributed in both Asia and Africa, and there has been a long-lasting debate regarding the taxonomic status of the Asian and African populations (Tateoka, 1965; Nayar, 1973; Vaughan, 1989; Vaughan et al., 2003; Zhang

& Ge, 2007). Based on the studies of gross morphology, Sharma & Shastry (1965) named the Sri Lanka *O. eichingeri* a new species, *O. collina*, but the nomenclature was later retracted by Vaughan et al. (2003). However, molecular data suggested high levels of genetic differentiation between the African and Sri Lankan populations of *O. eichingeri* (Shcherban et al., 2000a; Bao & Ge, 2004; Bao et al., 2006; Bautista et al., 2006; Zhang & Ge, 2007).

In the present study, we carried out species-delimitation analyses on two types of multilocus datasets using a recently developed Bayesian coalescent-based method (Yang & Rannala, 2010). This method has the advantage of accounting for species phylogenies and coalescent events in extant and extinct species and accommodating lineage sorting and uncertainties in the gene trees (Yang & Rannala, 2010; Zhang et al., 2011). Our results based on sequences of 10 genes for population samples (dataset A) strongly support the species status of the African and Sri Lankan populations. The support for two species in the analyses of the 68 single-sample genes (dataset B) was not as high, probably due to lack of information in the dataset. We appreciate that the species status of allopatric populations and geographical races is often controversial and arbitrary. Although this study supports the species status of the two geographical races of *O. eichingeri*, we leave it to future work to carry out detailed morphological investigation and molecular population studies, with extensive sampling across the entire geographical distribution.

3.3 Factors that affect posterior probability

Recent studies using BPP to analyze both simulated and empirical datasets indicated that the posterior probability of the species-delimitation model may be influenced by the prior distribution for θ and τ, by the guide tree, and by migrations between populations, mutation rate variation among loci, and other factors not studied here (Leache & Fujita, 2010; Yang & Rannala, 2010; Zhang et al., 2011). In this work, we tested two datasets with different priors, guide trees, and τ threshold model, and found that the results for dataset A were stable under all conditions. However, results for dataset B showed much fluctuation, suggesting that the dataset with one sequence from each species is more sensitive to the prerequisite or selection of parameters.

Based on a simulation study, Zhang et al. (2011) suggested that the correct species model could be inferred with 50 loci when only one sequence was sampled. They also found that the posterior probabilities increased as the number of genes increased. In our analysis, even with 60 loci, Pr(111) was still less than 0.5. There were two possible reasons: (i) the tentative species S and U are the same species without sufficient genetic divergence; or (ii) more genes are needed to make accurate species delimitation. Based on the results from dataset A, in conjunction with many previous studies, reason 1 can be excluded. Through the analyses of random samples from dataset B, we realized that Pr(111) ranged from 0.0 to 1.0 in every 100 replicates. It is reasonable to believe that if even more genes were used, the phenomenon might still exist. Here we suggest that for relatively closely related plant species, sampling multiple individuals from each population is critically important for species delimitation. Zhang et al. (2011) obtained high posterior probability with only one or two loci when five or 10 sequences were sampled from each species in their simulation study. In our real-data analysis, sampling four sequences from each species/population and including only five loci caused the speciation probabilities to reach 1.0.

Acknowledgements We thank Liang TANG and Xiao-Ming ZHENG for helpful suggestions on the manuscript, and Jie GUO and Zhe LI for methodology support. We also thank Bin AI, Shan-Shan LI, and other members of the Ge Laboratory (Institute of Botany, Chinese Academy of Sciences) for technical assistance. We are grateful to the International Rice Research Institute (Los Banos, Philippines) for providing seed samples. This work was supported by the National Natural Science Foundation of China (Grant Nos 30990240 and 30121003).

References

Ammiraju JSS, Song X, Luo M, Sisneros N, Angelova A, Kudrna D, Kim H, Yu Y, Goicoechea JL, Lorieux M, Kurata N, Brar D, Ware D, Jackson S, Wing RA. 2010. The *Oryza* BAC resource: A genus-wide and genome scale tool for exploring rice genome evolution and leveraging useful genetic diversity from wild relatives. Breeding Science 60: 536–543.

Bao Y, Ge S. 2004. Origin and phylogeny of *Oryza* species with the CD genome based on multiple-gene sequence data. Plant Systematics and Evolution 249(1–2): 55–66.

Bao Y, Zhou HF, Hong DY, Ge S. 2006. Genetic diversity and evolutionary relationships of *Oryza* species with the B- and C-genomes as revealed by SSR markers. Journal of Plant Biology 49: 339–347.

Bautista NS, Vaughan D, Jayasuriya AHM, Liyanage ASU, Kaga A, Tomooka N. 2006. Genetic diversity in AA and CC genome *Oryza* species in southern South Asia. Genetic Resources and Crop Evolution 53: 631–640.

Brar DS, Khush GS. 1997. Alien introgression in rice. Plant Molecular Biology 35: 35–47.

Davis JI, Nixon KC. 1992. Populations, genetic-variation, and the delimitation of phylogenetic species. Systematic Biology 41: 421–435.
De Queiroz K. 2007. Species concepts and species delimitation. Systematic Biology 56: 879–886.
Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences USA 104: 5936–5941.
Farris JS, Kallersjo M, Kluge AG, Bult C. 1994. Testing significance of incongruence. Cladistics 10: 315–319.
Federici MT, Shcherban AB, Capdevielle F, Francis M, Vaughan D. 2002. Analysis of genetic diversity in the *Oryza officinalis* complex. Electronic Journal of Biotechnology 5: 173–181.
Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.
Ge S, Sang T, Lu BR, Hong DY. 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proceedings of the National Academy of Sciences USA 96: 14400–14405.
Godfray HCJ. 2002. Challenges for taxonomy – The discipline will have to reinvent itself if it is to survive and flourish. Nature 417: 17–19.
Knapp S, Polaszek A, Watson M. 2007. Spreading the word. Nature 446: 261–262.
Leache AD, Fujita MK. 2010. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). Proceedings of the Royal Society B-Biological Sciences 277: 3071–3077.
Liu L, Pearl DK. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Systematic Biology 56: 504–514.
Nayar, NM. 1973. Origin and cytogenetics of rice. Advances in Genetics 17: 153–292.
Padial JM, Miralles A, De la Riva I, Vences M. 2010. The integrative future of taxonomy. Frontiers in Zoology 7: 16.
Posada D, Crandall KA. 1998. Modeltest: Testing the model of DNA substitution. Bioinformatics 14: 817–818.
Rannala B, Yang ZH. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.
Rannala B, Yang ZH. 2008. Phylogenetic inference using whole genomes. Annual Review of Genomics and Human Genetics 9: 217–231
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
Sang T, Ge S. 2007. Genetics and phylogenetics of rice domestication. Current Opinion in Genetics and Development 17: 533–538.
Sharma SD, Shastry SVS. 1965. Taxonomic studies in genus *Oryza* L. 6. A modified classification. Indian Journal of Genetics and Plant Breeding 25: 173–178.
Shcherban AB, Vaughan DA, Tomooka N. 2000a. Isolation of a new retrotransposon-like DNA sequence and its use in analysis of diversity within the *Oryza officinalis* complex. Genetica 108: 145–154.
Shcherban AB, Vaughan DA, Tomooka N, Kaga A. 2000b. Diversity in the integrase coding domain of a gypsy-like retrotransposon among wild relatives of rice in the *Oryza officinalis* complex. Genetica 110: 43–53.
Swofford DL. 2003. PAUP*: Phylogenetic analysis using parsimony (* and other methods), Version 4.0. Sunderland: Sinauer Associates.
Tateoka T. 1965. A taxonomy study of *Oryza eichingeri* and *O. punctata*. The Botanical Magazine Tokyo 78: 156–163.
Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP. 2003. A plea for DNA taxonomy. Trends in Ecology & Evolution 18(2): 70–74.
Vaughan DA. 1989. Genus *Oryza* L. Current status of taxonomy. IRRI Research Paper Series 138. Manila: International Rice Research Institute.
Vaughan DA, Morishima H, Kadowaki K. 2003. Diversity in the *Oryza* genus. Current Opinion in Plant Biology 6: 139–146.
Wang B, Ding Z, Liu W, Pan J, Li C, Ge S, Zhang D. 2009. Polyploid evolution in *Oryza officinalis* complex of the genus *Oryza*. BMC Evolutionary Biology 9: 250.
Wiens JJ. 2007. Species delimitation: New approaches for discovering diversity. Systematic Biology 56: 875–878.
Wiens JJ, Servedio MR. 2000. Species delimitation in systematics: inferring diagnostic differences between species. Proceedings of the Royal Society of London Series B-Biological Sciences 267: 631–636.
Wing RA, Ammiraju JS, Luo M, Kim H, Yu Y, Kudrna D, Goicoechea JL, Wang W, Nelson W, Rao K, Brar D, Mackill DJ, 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 Molecular Biology 59: 53–62.
Yang ZH, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences USA 107: 9264–9269.
Zhang LB, Ge S. 2007. Multilocus analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. Molecular Biology and Evolution 24: 769–783.
Zhang C, Zhang DX, Yang ZH. 2011. Evaluation of a Bayesian coalescent method of species delimitation. Systematic Biology (in press).
Zheng XM, Ge S. 2010. Ecological divergence in the presence of gene flow in two closely related *Oryza* species (*Oryza rufipogon* and *O. nivara*). Molecular Ecology 19: 2439–2454.
Zou XH, Ge S. 2008. Conflicting gene trees and phylogenomics. Journal of Systematics and Evolution 46: 785–807.
Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biology 9: R49.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Species delimitation analyses for each gene of dataset A using $\theta_0 \sim G$ (1, 250) and $\tau_0 \sim G$ (1, 250) as prior distribution.

Table S2. Posterior probabilities of Pr(111) for every 100 replications randomly selected from dataset B with different gene samples using $\theta_0 \sim G$ (1, 250) and $\tau_0 \sim G$ (1, 250) as prior distribution.