Multilocus estimation of divergence times and ancestral effective population sizes of *Oryza* species and implications for the rapid diversification of the genus

Xin-Hui Zou[1], Ziheng Yang[2,3], Jeff J. Doyle[4] and Song Ge[1]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China; [2]Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101, China; [3]Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London, WC1E 6BT, UK; [4]Department of Plant Biology, Cornell University, 412 Mann Library Building, Ithaca, NY 14853, USA

Author for correspondence:
Song Ge
Tel: +86 10 62836097
Email: gesong@ibcas.ac.cn

Received: *2 January 2013*
Accepted: *8 February 2013*

New Phytologist (2013) **198:** 1155–1164
doi: 10.1111/nph.12230

Key words: divergence time, *Oryza*, population size, rapid diversification, relaxed clock.

Summary

- Despite substantial investigations into *Oryza* phylogeny and evolution, reliable estimates of the divergence times and ancestral effective population sizes of major lineages in *Oryza* are challenging.
- We sampled sequences of 106 single-copy nuclear genes from all six diploid genomes of *Oryza* to investigate the divergence times through extensive relaxed molecular clock analyses and estimated the ancestral effective population sizes using maximum likelihood and Bayesian methods.
- We estimated that *Oryza* originated in the middle Miocene (*c*. 13–15 million years ago; Ma) and obtained an explicit time frame for two rapid diversifications in this genus. The first diversification involving the extant F-/G-genomes and possibly the extinct H-/J-/K-genomes occurred in the middle Miocene immediately after (within < 1 Myr) the origin of *Oryza*. The second giving rise to the A-/B-/C-genomes happened *c*. 5–6 Ma. We found that ancestral effective population sizes were much larger than those of extant species in *Oryza*.
- We suggest that the climate fluctuations during the period from the middle Miocene to Pliocene may have contributed to the two rapid diversifications of *Oryza* species. Such information helps better understand the evolutionary history of *Oryza* and provides further insights into the pattern and mechanism of diversification in plants in general.

Introduction

Rice has played an essential role for our species, as it provides the staple food for over half of the human population. Increasing the rice yield and improving its quality is an important task in our future efforts to meet the increasing demands on food supply from population growth and economic development worldwide (Khush, 2001). Understanding the evolutionary history of cultivated rice and its wild relatives will help achieve these goals by facilitating the utilization of useful genes in wild rice (Ge *et al*., 1999; Wing *et al*., 2005; Sang & Ge, 2007). Furthermore, rice and its relatives have become an excellent model system for various evolutionary and functionary studies owing to their relatively small genome sizes, dense genetic maps and extensive genome colinearity with other cereal species, as well as the completion of genome sequencing of two rice subspecies (Wing *et al*., 2005; Ammiraju *et al*., 2008). In order to take full advantage of the ideal system for comparative and functional studies, we need better understanding of its fundamental background including the evolutionary history and dynamics of rice and its relatives.

The genus *Oryza* consists of two cultivated and *c*. 22 wild rice species, and is represented by 10 distinct genome types, including six diploid (A, B, C, E, F and G) and four allotetraploid (BC, CD, HJ and HK) genome types (Aggarwal *et al*., 1997; Ge *et al*., 1999). Through extensive phylogenetic analyses, the phylogenetic relationships among different genome types and species in *Oryza* have been well established (Ge *et al*., 1999; Zou *et al*., 2008). However, a few important evolutionary parameters have not been studied extensively or remain controversial, including the divergence times among lineages and ancestral effective population sizes. To date, several studies have used different methods to date the divergence times of the *Oryza* species, with different estimates obtained for some lineages (Second, 1985; Guo & Ge, 2005; Ammiraju *et al*., 2008; Lu *et al*., 2009; Sanyal *et al*., 2010; Tang *et al*., 2010). Estimates of ancestral effective population sizes are almost lacking except for Zhang & Ge (2007) in which the ancestral effective population sizes of C-genome *Oryza* species were attempted. Recently, Ai *et al*. (2012) estimated the effective population sizes for 10 extant diploid

Oryza species, but no ancestral effective population size was obtained.

Divergence time is one of the key factors required to interpret the patterns of speciation and rates of adaptive radiation, and it is also necessary for estimating rates of genetic and morphological change and for understanding biogeographic history (Kumar & Hedges, 1998; Arbogast *et al.*, 2002). Absolute divergence times allow evolutionary events to be placed in the appropriate context of global climate changes and geographical events, thereby suggesting possible mechanisms of lineage divergence and speciation (Tiffney & Manchester, 2001; Stromberg, 2005; Prasad *et al.*, 2011). Effective population size is a central parameter in models of population genetics, conservation and human evolution, and plays important roles in plant and animal breeding (Rannala & Yang, 2003; Charlesworth, 2009). It helps answer numerous evolutionary questions by determining the equilibrium level of neutral or weakly selected variability in populations and evaluating the effectiveness of selection relative to genetic drift (Yang, 2002; Charlesworth, 2009). Estimates of the ancestral effective population size around the time of speciation can give very useful insights into the historical demographic process and speciation (Chen & Li, 2001; Broughton & Harrison, 2003; Rannala & Yang, 2003; Zhang & Ge, 2007; Zhou *et al.*, 2007; Yang, 2010).

In this study, we sampled all six diploid genomes of the genus *Oryza* and selected 106 single-copy nuclear genes across all 12 rice chromosomes to estimate the divergence times and ancestral population sizes of major lineages in this genus. With these parameters available, we were able to estimate the time frame of species diversification in *Oryza* and explored potential climatic factors that underlie the diversification of the *Oryza* genus. Such information may help better understand the evolutionary history of *Oryza* and provides further insights into the pattern and mechanism of speciation and diversification in plants in general.

Materials and Methods

Species and loci sampled

We sampled all six diploid genomes of *Oryza*, each represented by one species: *O. rufipogon* Griff. (A-genome), *O. punctata* Kotechy ex Steud. (B-genome), *O. officinalis* Wall. ex Watt (C-genome), *O. australiensis* Domin. (E-genome), *O. brachyantha* Chev. et Roehr. (F-genome), *O. granulata* Nees et Arn. ex Watt (G-genome). We also included *Leersia tisserantti* (A. Chev.) Launert, a species from the most closely related genus to *Oryza* as outgroup (Tang *et al.*, 2010). In our previous phylogenetic study, we sequenced DNA fragments of 142 nuclear genes that are distributed throughout the 12 chromosomes in rice for all these seven species (Zou *et al.*, 2008). Here, we selected 106 genes in which there were no missing data for all seven species from the 142 gene dataset. Because neutral sites are commonly required for molecular dating and estimating ancestral effective population size (Kimura, 1983), we chose to use synonymous sites and intron sequences (silent sites) from the 106 genes in our analyses. All the sequences were aligned using T-Coffee (Notredame *et al.*, 2000) and manually refined. After removing ambiguously aligned regions, the sequenced regions of the 106 genes ranged from 264 to 1675 bp in length, with total length of 88 526 bp. Variable sites ranged from 7.02% to 31.77%, and informative sites ranged from 1.0% to 14.39%. Details of the 106 genes are provided in Supporting Information Table S1.

Estimation of divergence time

We used two Bayesian Markov chain Monte Carlo (MCMC) programs, MULTIDIVTIME (Thorne *et al.*, 1998; Thorne & Kishino, 2002) and MCMCTREE (Yang & Rannala, 2006; Rannala & Yang, 2007), to estimate the divergence time. These two relaxed-clock methods can account for the rate heterogeneity among lineages, incorporate multiple loci into one analysis and deal with the heterogeneous rates among loci appropriately. Specifically, MULTIDIVTIME uses a rate-drift model to implement auto-correlated rates among adjacent lineages to relax the molecular clock, while MCMCTREE uses two strategies to model the change of evolutionary rate among lineages, that is, independent rates (clock2) and auto-correlated rates (clock3; Rannala & Yang, 2007). Node ages are constrained by hard bounds in MULTIDIVTIME, in which a zero probability of true divergence time being outside the minimum and maximum age constraint is imposed. In MCMCTREE, soft bounds are used so that the minimum and maximum age constraint may be violated with a small probability (2.5%).

Up to now, only one macrofossil record belonging to *Oryza*, the fossil of spikelets, has been reported from an excavation of Miocene age in Germany and was identified as *O. exasperata* (A. Braun) Heer. (Heer, 1855). It appears to resemble the extant *O. granulata*, based on its morphology (Heer, 1855; Tang *et al.*, 2010). Although this fossil suggests that *Oryza* had already differentiated in the Miocene (5–23 Ma), its precise date is unclear. Thus, we used the 5 Ma as the conservative minimum age constraint for the crown node of the genus *Oryza*. Several previous studies have attempted to date the origin of *Oryza*, with different age estimates ranging from *c.* 8 to *c.* 15 Ma (Guo & Ge, 2005; Ammiraju *et al.*, 2008; Lu *et al.*, 2009; Sanyal *et al.*, 2010; Tang *et al.*, 2010). We thus used 15 Ma as the maximum age constraint for the MCMCTREE and MULTIDIVTIME to calibrate the crown node of *Oryza*. Our previous study based on 142 nuclear genes obtained a robust phylogenetic tree of diploid *Oryza* species (Zou *et al.*, 2008), which was used as the input tree in this study.

First, the Bayesian relaxed-clock method implemented in MCMCTREE program in PAML v4.4 (Yang, 2007) was used for dating. The HKY85 + Γ model was used with different transition/transversion rate ration parameters (κ) and different shape parameters (α) among loci. A gamma prior G(6, 2) was assigned for κ, and G(1, 1) for α (The gamma distribution G(α, β), with shape parameter α and scale parameter β, has mean α/β and variance α/β^2). A calibration point at the basal position of *Oryza* was set as B(0.05, 0.15), with 100 million yr (Myr) as one time unit. The overall substitution rate (rgene) was assigned by a gamma distribution with prior G(3, 5), that is, mean 0.6 and standard deviation (SD) 0.35, based on the synonymous substitution rate from nuclear genes in grasses (i.e. 6×10^{-9} substitutions per site per year; Gaut *et al.*, 1996; Gaut, 1998; White & Doebley, 1999). Parameter σ^2 was used to specify variability of rates across branches and was set to G(2, 1). We varied the setting for σ^2 and calibration bounds to examine the effect of these priors on posterior estimates. A total of 40 000 generations was sampled every five steps after discarding the initial 20 000 samples as burn-in.

Secondly, we used MULTIDIVTIME to estimate the divergence time. The calibration point for the basal node of *Oryza* was set as B(0.5, 1.5), with 10 Myr as one time unit. The mean of the prior on the rate (rtrate) was set as 0.06 (i.e. 6×10^{-9} substitutions per site per year), with standard deviation (rtratesd) as 0.03, similar to MCMCTREE. Brownmean and brownsd set the mean and SD of the prior distribution for the rate variance across branches. We use the suggested strategy (Thorne & Kishino, 2002) to have rttm × brownmean to be 1 and to have brownsd equal to brownmean. Also, we varied the setting for calibration bounds and adjusted the parameter brownmean to examine the effect of these priors on posterior estimates. Model parameters of F84 + Γ were first estimated for each loci using baseml in PAML v4.4. Then, maximum likelihoods (ML) of the branch lengths for the ingroup rooted tree and their variance–covariance matrix were estimated by estbranches in MULTIDIVTIME program for each gene separately, after transforming the baseml output files by paml2modelinf program. Lastly, Bayesian MCMC analysis was conducted to approximate the posterior distributions of divergence time using multidivtime. We analysed a total of 100 000 generations with sample frequency as 100, after discarding the initial 500 000 samples as burn-in. Each analysis was conducted at least twice to ensure consistency between different runs. Additionally, convergence of the MCMC algorithm was assessed by Tracer v1.4 (Rambaut & Drummond, 2007).

Estimation of ancestral effective population size

The effective population size of modern species can be estimated by the observed polymorphism in populations, whereas that of the extinct ancestral species is much harder to determine. Takahata (1986) suggested a method for estimating the ancestral population size of two related species from multiple loci, based on the fact that the coalescent time in the ancestral population fluctuates over loci, which is in proportion to the effective size of ancestral population. This method has been expanded further to the two-species ML method (Takahata *et al.*, 1995; Takahata & Satta, 1997) and the Bayesian MCMC method (Yang, 2002) that we used here. Both methods are based on the coalescent model and extract information from the variation of sequence divergence among loci. The Bayesian method also makes use of information from the conflicting gene tree topologies as well. They all assume rate constancy among lineages, no intragenic recombination and no gene flow. Therefore, we first performed the likelihood-based relative rate test (RRT) to test the rate constancy among lineages (i.e. the clock hypothesis) using the program HyPhy (Pond *et al.*, 2005). In this case, multiple pairwise tests were conducted for all of species pairs for a given locus, using *L. tisserantti* as an outgroup. Considering the nonindependence of pairwise tests, we used Bonferroni corrections in order to reduce the probability of a type-I error (Rice, 1989). The clock hypothesis was rejected for any given gene if any of the pairwise tests for that gene were significant (Posada, 2001). Loci that evolved neutrally were then used for estimation.

The two-species ML method uses pairs of orthologous loci from two species, and the nucleotide divergence of orthologous sequences from a pair of species consists of two parts: the divergence that arose before and after the speciation. When multiple loci are considered, the divergence after speciation is constant across all loci, while the divergence before speciation differs among loci according to the exponential distribution with mean and SD all equal to $2N\mu$ (N is the effective population size of the ancestral species, μ is the rate of substitution per site per generation). Let θ and τ equal $4N\mu$ and $2t\mu$, where t represents the divergence time of two species; we could determine the value of θ and τ by maximizing the following log-likelihood function (Takahata & Satta, 1997):

$$L(\theta,\tau) = \sum_{i=1}^{m} \left[-n_i\tau - \log_e(1+n_i\theta) + \log_e \sum_{d=0}^{k_i} \frac{(n_i\tau)^d}{d!} \left\{ \frac{n_i\theta}{1+n_i\theta} \right\}^{k_i-d} \right] \qquad \text{Eqn 1}$$

(n_i and k_i, the total number of sites and the number of different sites at the ith locus, respectively; m, number of loci). The calculation is implemented in a program written by the present authors in the R language (R Development Core Team, 2008).

The Bayesian MCMC method was implemented in the MCMCcoal program (Yang, 2002; Rannala & Yang, 2003). The JC69 substitution model was used to correct for multiple hits at the same site. Unlike the two-species ML method that only analyses paired species, this method analyses multiple species and loci simultaneously. The species tree topology is assumed to be known and fixed in the analysis. The priors for θ and τ are approximated by independent gamma distributions and must be specified before the analysis. The mean of gamma priors for θ was assigned varying from 0.005 to 0.05, and the priors for population divergence time were assigned based on previous dating results (Ammiraju *et al.*, 2008; Lu *et al.*, 2009; Tang *et al.*, 2010). We initially conducted analyses assuming a constant rate for each locus. To examine whether the variation of evolutionary rates among loci has an effect on the posterior estimates, we performed analyses by incorporating the relative rate for each gene. When running the Bayesian MCMC algorithm, we sampled a total of 1 000 000 generations with sample frequency of 5, after discarding the initial 100 000 samples as burn-in. The same analysis was performed at least twice to check for convergence.

In order to investigate potential intragenic recombination, the sequence alignment was examined using the Recombination Detection Program (RDP; Martin *et al.*, 2010). Six automated recombination detection methods (RDP, GENECONV, Chimaera, MaxChi, BootScan and SiScan) were implemented and default settings were used. Only potential recombination signals detected by two or more of the above six methods were considered significant.

Results

Divergence time estimates of six diploid genome types based on 106 loci are shown in Table 1 and Fig. 1. Time estimates obtained from MCMCTREE and MULTIDIVTIME are very similar. The origin of the rice genus is dated *c.* 13–15 Ma, with 95% CI at (14.35, 16.57), (14.49, 16.64) and (12.12, 14.75) for clock2 and clock3 of MCMCTREE, and MULTIDIVTIME, respectively. The two programs dated the divergence of F-genome at *c.* 15 and *c.* 13 Ma, respectively. For the divergence of other genomes, the two programs gave largely consistent results: E-genome branched out from the ancestral lineage of the A-, B-, and C-genomes at *c.* 7 Ma, and C-genome diverged at *c.* 6 Ma, whereas the A- and B-genomes separated at *c.* 5.5 Ma.

Table 1 Divergence time estimation by the programs MCMCTREE and MULTIDIVTIME

	MCMCTREE			MULTIDIVTIME	
Node	Prior	Posterior (clock2)	Posterior (clock3)	Prior	Posterior
I	10.00 (5.00, 15.00)	15.30 (14.35, 16.57)	15.40 (14.49, 16.64)	9.63 (5.25, 14.53)	13.46 (12.12, 14.75)
II	8.01 (3.04, 13.84)	15.17 (14.20, 16.43)	15.35 (14.44, 16.59)	7.91 (3.22, 13.49)	12.80 (11.52, 14.04)
III	6.02 (1.58, 12.04)	7.50 (6.87, 8.23)	7.36 (6.83, 8.04)	6.09 (1.77, 11.93)	6.79 (6.00, 7.58)
IV	4.03 (0.58, 9.80)	6.13 (5.57, 6.76)	5.87 (5.44, 6.41)	4.21 (0.56, 9.88)	5.79 (5.10, 6.49)
V	2.04 (0.05, 6.99)	5.61 (5.08, 6.21)	5.31 (4.89, 5.81)	2.21 (0.06, 7.30)	5.49 (4.83, 6.16)

The priors of the node ages specified by the two programs and the posterior means and their 95% confidence intervals (in parentheses) are given in Myr. Nodes are numbered as in Fig. 1. The priors on node age were assessed by running Markov chain Monte Carlo (MCMC) without data.

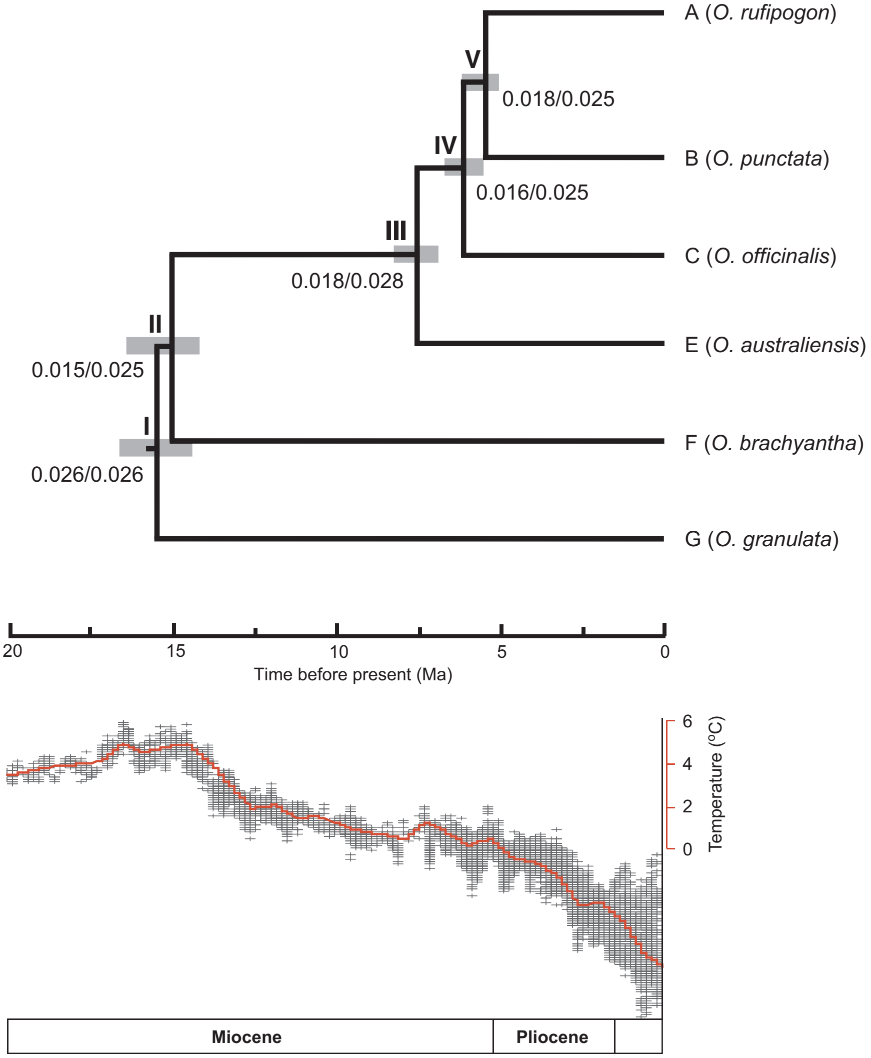

Fig. 1 The chronogram of the genus *Oryza* based on 106 loci, showing the results of Bayesian MCMCTREE analysis under the clock2 model with calibration of (5, 15) Myr. The capital letters represent the genome types, followed by the sampled species in parentheses. The ancestral nodes are indicated as I, II, III, IV and V. The branch length of the tree represents the posterior means of date estimates with the grey bars illustrating 95% Confidence Intervals (CIs). The details of the dates are shown in Table 1. The numbers besides the nodes indicate the θ values of ancestral populations estimated by two-species ML and Bayesian MCMC methods, respectively. Below the chronogram is the figure showing temperature changes since the Miocene (modified from Zachos *et al.*, 2001).

It is noted that considerable overlaps of the estimates in 95% CI were found between the divergence times of the F-genome and the G-genome (node I and II), and between the divergence times of the C-genome and the A- and B-genomes (node IV and V). This implies two episodes of rapid speciation that gave rise to the G- and F-genomes and the A-, B-, C-genomes. To explore whether the overlap was caused by the prior setting we used in the MCMC analyses, we examined the effect of priors for fossil calibration and rate heterogeneity across branches using two sets of parameters. In the first set, we narrowed and widened the width of the age constraints for node I by using different calibrations, that is, (10, 15), (5, 15) and (5, 20), respectively (in the form of minimum and maximum age in parenthesis, with time unit as Myr). In the second set, we changed the parameter settings that control rate variation across branches, that is, parameter σ^2 in MCMCTREE by using different gamma priors as G(1, 10), G(2, 1) and G(5, 1), respectively, and changed the brownmean parameter in MUTIDIVTIME by using different values equal to 1, 3, 5 and 10, respectively. Results show that different calibrations have only a slight effect on node age, and the variation in posterior mean times obtained by MCMCTREE was < 3.36 Myr, nearly the same results obtained by MULTIDIVTIME with a difference of < 0.26 Myr (Fig. S1b,d). Compared with the calibration, the prior for the rate heterogeneity across branches had slightly larger effects on posterior estimates, with the differences in posterior means < 6.09 Myr in MCMCTREE and < 0.88 Myr in MULTIDIVTIME (Fig. S1a,c). When we draw attention to the two time spans between nodes I and II and between nodes IV and V, the CIs always overlapped (Fig. S1).

As stated by Yang & Rannala (2006), for a specified set of fossil calibrations, the error of posterior time estimates cannot be reduced to zero by increasing the number of sites in the sequence. Yang & Rannala (2006) predicted that when the sequence data approach infinity, the posterior means and the 95% CIs for all node ages will lie on a straight line. We therefore examined whether or not adding more sequence data would improve the date estimates by plotting the posterior means of divergence time against the width of corresponding 95% CIs for each node, following Yang & Rannala (2006). As shown in Fig. S2, a nearly perfect linear relationship exists between the posterior means and the 95% CI bounds, regardless of the programs used ($r^2 = 0.84$–0.95). The high correlation coefficients of the plot suggest that our sequence data are highly informative, and it seems unlikely that the precision of time estimates might be improved by adding more sequences.

Our likelihood-based relative rate test (RRT) showed that among the 106 gene dataset, 66 genes did not reject the clock hypothesis and thus were included in our estimation of ancestral effective population size (Table S1). For the two-species ML method, we used several classes of paired species in which all pairs share the common ancestral population following the method of Satta *et al.* (2004). For node I, we used five pairs of species: A- and G-genomes, B- and G-genomes, C- and G-genomes, E- and G-genomes, F- and G-genomes; for node II, we used four pairs of species: A- and F-genomes, B- and F-genomes, C- and F-genomes, E- and F-genomes. Similarly, for nodes III, IV, and V, we used three, two, and one pairs of species, respectively. We obtained fairly large results for all ancestral polymorphism estimates (θ value), with all of them ≥ 0.013, and the estimates within each class were similar (Table 2).

Because variation of evolutionary rates among loci may influence the estimate and ignoring this rate variation might inflate the coalescent variation among loci, leading to the overestimation of the ancestral effective population size (Yang, 1997), we calculated the relative rate of each gene following Yang (2002); that is, for a given locus, averaging the JC69 distance from the A-, B-, C-, E-, and F-genomes to the G-genome, and then dividing by the mean of the value across all 66 loci. Although we found minimal rate of variation among the 66 loci (Fig. S3), we evaluated the effect of among-locus rate variation in the Bayesian MCMC method, in which the relative rate for each locus could be taken into consideration. Analyses with or without incorporating the relative rates of each locus did not obtain significantly different estimates ($P = 0.89$, Wilcoxon signed-ranks test), and thus we only show results that account for rate heterogeneity (Table 2).

Given that little prior information is available for θ, we performed a series of analyses using different priors for θ in the Bayesian MCMC method, with the mean being 0.005, 0.01, 0.02, 0.03 and 0.05, respectively, based on estimates of ancestral polymorphisms of *Oryza* C-genome species and of other genera

Table 2 Estimates of θ values of ancestral populations in *Oryza* using the two-species maximum likelihood (ML) and Bayesian Markov chain Monte Carlo (MCMC) methods based on 66 neutrally evolving loci

Node	Method	Paired taxa	θ	95% CI
I	Two-species ML	A-G	0.030	0.018, 0.042
		B-G	0.024	0.016, 0.032
		C-G	0.024	0.016, 0.032
		E-G	0.026	0.016, 0.036
		F-G	0.028	0.018, 0.038
		Average	0.026	
	Bayesian MCMC		0.026	0.020, 0.034
II	Two-species ML	A-F	0.017	0.009, 0.025
		B-F	0.016	0.008, 0.024
		C-F	0.013	0.005, 0.021
		E-F	0.015	0.007, 0.023
		Average	0.015	
	Bayesian MCMC		0.025	0.008, 0.051
III	Two-species ML	A-E	0.017	0.011, 0.023
		B-E	0.017	0.011, 0.023
		C-E	0.021	0.013, 0.029
		Average	0.018	
	Bayesian MCMC		0.028	0.022, 0.035
IV	Two-species ML	A-C	0.017	0.011, 0.023
		B-C	0.014	0.008, 0.020
		Average	0.016	
	Bayesian MCMC		0.025	0.015, 0.039
V	Two-species ML	A-B	0.018	0.012, 0.024
	Bayesian MCMC		0.025	0.008, 0.050

Species pairs and node specifications are labeled as those in Fig. 1. 95% CI (Confidence Intervals) for the two-species ML method are calculated as θ ± 2 SE.

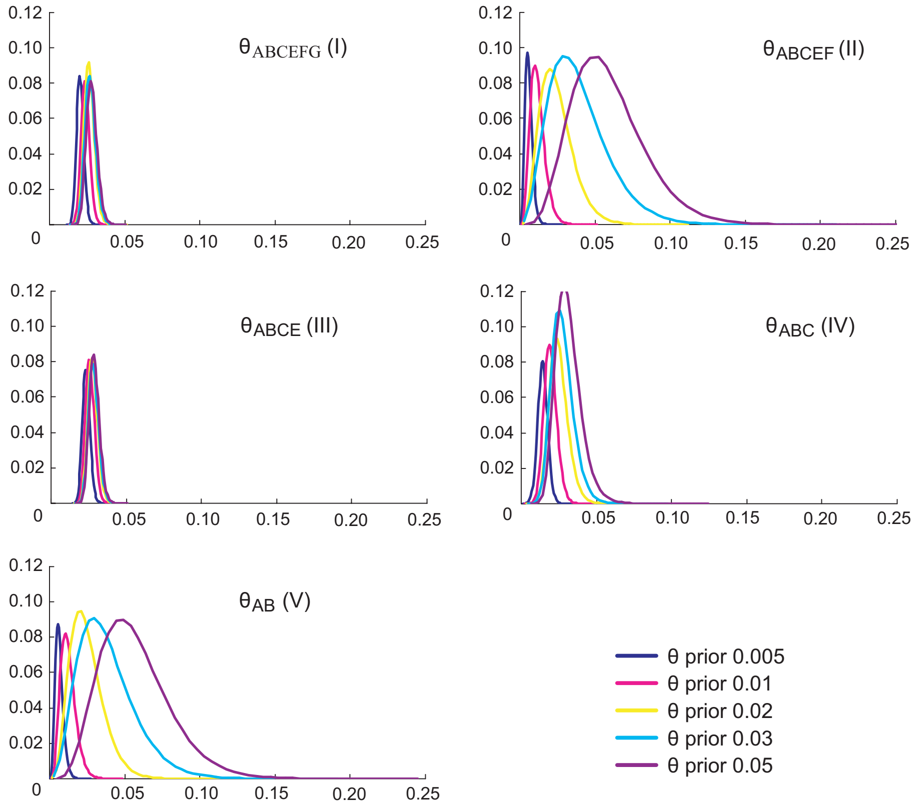

Fig. 2 The posterior distributions of θ values for the five internal nodes for the genus *Oryza* in Fig. 1. Different priors for θ were used in the analyses, with the prior means 0.005, 0.01, 0.02, 0.03 and 0.05, respectively. The ordinate and abscissa represent the frequency and midvalue of θ values, respectively.

(Rannala & Yang, 2003; Zhang & Ge, 2007; Zhou *et al.*, 2007; Yang, 2010). The posterior distributions of the θ values for the five internal nodes were plotted in Fig. 2 and the details are provided in Table S2. As shown in Fig. 2, the posterior θ estimates obtained based on different priors were very similar for the ancestral populations of all genomes (node I) and the A-/B-/C-/E-genomes (node III). For the ancestral population of A-/B-/C-genomes (node IV), the posterior estimates were slightly influenced by the priors, but the posterior distributions overlapped substantially. For the ancestral population of A-/B-/C-/E-/F-genomes and that of A-/B-genomes (nodes II and V), the mean and distribution of the posteriors were dependent on the priors. Although the priors for population divergence time (τ) were assigned based on previous dating results (Ammiraju *et al.*, 2008; Lu *et al.*, 2009; Tang *et al.*, 2010), we also varied the τ priors four-fold to examine the effect on posterior estimates. These analyses showed that the τ priors only had a negligibly small effect on the estimates (results not shown).

Taken together, all estimates for the ancestral effective population size by the two methods were largely consistent, except that the estimates for node III were slightly lower with ML than that with the Bayesian method (Table 2). Our estimates suggest large θ values for the ancestral population during the evolutionary history of *Oryza*. Taking the generation time for the ancestral species as 1 yr and the evolutionary rate as 6×10^{-9} substitutions per site per year (Gaut, 1998), we estimated the ancestral effective population size to be over 540 000 throughout the evolutionary history.

Discussion

Divergence time in *Oryza*

Despite increasing interest in the rice genus and the flood of sequence data in recent years, a reliable estimate of the time frame of evolution in this genus and its relatives has been difficult, mainly owing to the absence of grass macrofossils. Previous studies have attempted to date the origin of the major lineages of *Oryza* under the clock assumption and have obtained different age estimates of the origin of the genus, ranging from *c.* 8 to *c.* 15 Ma (Second, 1985; Guo & Ge, 2005; Ammiraju *et al.*, 2008; Lu *et al.*, 2009; Sanyal *et al.*, 2010). The inconsistent estimates are partly due to the different methods used and different species sampling, and partly because relatively few genes or loci were sampled. Recently, based on sequences of 20 chloroplast gene fragments, Tang *et al.* (2010) have reconstructed the phylogeny of the rice tribe and dated the major lineages of Oryzeae using relaxed-clock approaches. Nevertheless, Tang *et al.* (2010) used exclusively maternal-inherited chloroplast genes in their phylogenetic reconstruction, which may potentially have led to a fallacious cpDNA-based tree rather than the actual organismal phylogeny because of factors such as gene transfer and chloroplast capture (Soltis & Kuzoff, 1995).

Here, using two Bayesian relaxed clock methods that can incorporate different sources of information and adequately account for the uncertainty of fossil calibrations (Inoue *et al.*, 2010), we obtain an explicit time frame for this important group based on 106 nuclear genes which distribute randomly across all 12 rice chromosomes. Our results indicated that the rice genus originated *c.* 13–15 Ma, followed by two consecutive and rapid diversifications (Table 1, Fig. 1). The first diversification, which involved the extant F-/G-genomes and possibly the extinct H-/J-/K-genomes, occurred in the middle Miocene immediately after – within < 1 Myr – the origin of *Oryza*. The second one, which gave rise to the A-/B-/C-genomes, happened around – within < 0.5 Myr – the Miocene/Pliocene boundary. The MCMCcoal program, which accommodates ancestral polymorphisms and incomplete lineage sorting, also gave similar divergence time estimates (τ) when we used it to estimate ancestral effective population sizes.

As many other studies have indicated, fossil calibrations have great impact on dating results and relaxing the molecular clock assumption is also important when different lineages evolve with heterogeneous rates (Rannala & Yang, 2007). Therefore, we explored the impact of these two factors on posterior time estimations. We changed the age constraints and the parameters that specify the amount of rate variation across branches, that is, σ^2 in MCMCTREE and brownmean in MULTIDIVTIME. Although results showed small effects of these parameters on posterior estimates, the 95% CIs overlapped considerably in all analyses when we examined the two short intervals between speciation events that we mentioned above (Fig. S1). This fact demonstrates that these two short time spans were robust to the prior settings and imply two episodes of rapid diversification.

Large ancestral effective population size of *Oryza*

A reliable estimate of effective population size not only provides important parameters for studies in population genetics and evolution, but also facilitates numerous questions to be addressed involving biological conservation, as well as plant and animal breeding (Yang, 2002; Rannala & Yang, 2003; Zhang & Ge, 2007; Charlesworth, 2009). Previous studies of effective population size have mainly focused on the extant species of *Oryza* (Zhu & Ge, 2005; Zhang & Ge, 2007; Zhou *et al.*, 2008; Ai *et al.*, 2012), except that Zhang & Ge (2007) investigated nucleotide diversity in three species of C-genome and found that the ancestral effective population sizes were *c.* 2–10-fold larger than those of the extant species. In this study, we have obtained largely congruent estimates of the ancestral effective population size of all diploid genomes (Table 2). Compared with the nucleotide diversity of extant *Oryza* species – which ranged from 0.0011 for *O. punctata* to 0.0095 for *O. rufipogon* (Zhu & Ge, 2005; Ai *et al.*, 2012) – our estimates of ancestral polymorphisms (0.013–0.03) were much larger, indicating that the ancestral effective populations sizes of the rice genus were of the order of 10^5 throughout its evolutionary history.

These estimates were largely insensitive to prior settings in the Bayesian MCMC method. Prior settings for τ had nearly no effect on the estimates of θ values of the ancestral populations. When we changed the prior settings for θ, we found that θ values of the ancestral population of A-/B-/C-genomes, A-/B-/C-/E-genomes, and A-/B-/C-/E-/F-/G-genomes were very similar, over the 10-fold range of priors tested (Fig. 2 and Table S2). The insensitivity to priors implies a strong signal and sufficient information in the data. The fact that the posterior θ of the ancestral population of A-/B-genomes and that of A-/B-/C-/E-/F-genomes were sensitive to the prior settings suggested that the information for these two ancestral populations is much lower than the others, probably due to rapid diversification of these taxa.

Much larger θ estimates for ancestral species, relative to extant, have also been obtained for several other organisms, such as *Drosophila* (Machado *et al.*, 2002), field crickets (Broughton & Harrison, 2003), finches (Jennings & Edwards, 2005), cichlid fishes (Won *et al.*, 2005) and mangroves (Zhou *et al.*, 2007). The population size of the common ancestor of humans and chimpanzees was estimated to be *c.* 5–10 times larger than that of the extant human population (Chen & Li, 2001; Wall, 2003; Satta *et al.*, 2004; Burgess & Yang, 2008). Because large estimates for ancestral θ could be caused by methodological artifacts, including violation of the assumption of no intragenic recombination and no gene flow (Yang, 2010), we used the RDP program (Martin *et al.*, 2010) to investigate potential recombinant sequences in our data and detected recombination signals for 26 loci among them (Table S1). We thus re-analysed the data by excluding sequences showing evidence of recombination until no recombination signals were found. Very similar results for the reduced dataset (Table S3) suggest that recombination signals have no effect on our estimates. Moreover, intragenic recombination would reduce the coalescent variance among loci, leading to underestimates of ancestral effective population sizes (Takahata & Satta, 1997; Wall, 2003). Therefore, the violation of the assumption of no intragenic recombination would not overturn our conclusion of much larger θ estimates for the ancestral species of *Oryza*. Gene flow seems unlikely to cause inflated estimates in our case because hybridization or introgression between species with different genome types is difficult if not impossible, although gene flow has been documented between species with the same genome type (Vaughan *et al.*, 2003). Another possibility that the ancestral populations might be highly structured at the time of speciation cannot be ruled out entirely, because the restricted migration/gene flow between subpopulations would greatly increase the ancestral effective population size (Zhou *et al.*, 2007; Charlesworth, 2009).

Implications for the origin and diversification of *Oryza*

Rapid diversification, or radiation, is the rise of diversity of species within a lineage in a short evolutionary time span. This has happened throughout evolutionary history and has contributed greatly to the enormous diversity of life on earth (Arbogast *et al.*, 2002; Davies *et al.*, 2004; Rokas & Carroll, 2006). For example, consistent with the findings of the rapid rise and early diversification of flowering plants by Charles Darwin, Davies *et al.* (2004) found at least 10 significant rate accelerations of diversification within angiosperms. Similarly, studies of *Oryza* and its relatives have revealed numerous episodes of rapid diversification in their evolutionary history. In the subtribe Zizaniinae, for instance, Tang *et al.* (2010) identified a rapid diversification at *c.* 21.2 Ma, giving rise to eight genera within a short time interval. Within specific genome groups of *Oryza*, species radiations were also determined, including the A-genome species that started to radiate in the mid-Pleistocene (*c.* 2 Ma; Zhu & Ge, 2005), and the C-genome group that was diversified into three species within a short time span (Zhang & Ge, 2007). In particular, using sequences of 142 nuclear genes, Zou *et al.* (2008) clearly revealed two rapid diversification events that gave rise to almost all of the genome diversity of this genus. However, the inference of rapid diversifications by Zou *et al.* (2008) was mainly deduced from short branches in the gene trees, without an explicit time frame. The present study has provided strong evidence that the two rapid diversifications happened roughly at 13–15 and 5–6 Ma, respectively.

An important question arising from these studies is what factors underlie the multiple punctuated rapid diversifications in the *Oryza* genus. It is well recognized that evolution of organisms is profoundly influenced by the climate changes and tectonic events and that radiations are often associated with significant geological events (Richardson *et al.*, 2001; Stromberg, 2005). Evidence shows that there was a long period of cooling and rapid expansion of Antarctic continental ice-sheets from the early Eocene until a new warm phase began in the later Oligocene and peaked in the late middle Miocene (17–15 Ma; Zachos *et al.*, 2001). During the warm period, the extent of the global ice-sheet was reduced and remained low, and the trend toward higher temperatures is thought to have led to the expansions of many evergreen and thermophilic taxa (Tiffney & Manchester, 2001; Zachos *et al.*, 2001). It is likely that ancestral populations of *Oryza* expanded to Eurasia during this warm phase, provided that the rice genus originated in the Asia–Australia region (Vaughan *et al.*, 2005). This speculation is consistent with the finding that the macrofossil of *O. exasperata* (which resembles the extant *O. granulata* of the G-genome) was found in an excavation of Miocene age in Germany (Heer, 1855). After the warm peak in the late middle Miocene, cooling returned and ice-sheets on Antarctica were reestablished by 10 Ma (Zachos *et al.*, 2001). As a result, the tropical area of the Asia–Australia region became more scattered over two continents and thousands of islands, leading to highly localized biological diversity. At the same time, the vegetation of this area would have been significantly affected by a decrease in sea level, resulting in expansion of seasonal forest and savannah (Heaney, 1991). Thus, we assume that this warming and cooling process from early to middle Miocene could have accelerated the rapid diversifications of the *Oryza* lineage by giving rise to the F-, G-, and H-/J-/K-genomes (Fig. 1). This speculation corroborates the observation that the most basal lineages on the *Oryza* tree (e.g. *O. granulata*, *O. longiglumis*, *O. ridleyi*, *O. schlechteri*) existed mainly in forest and shaded habitats in this area (Quade *et al.*, 1994; Vaughan *et al.*, 2003, 2005).

The second rapid diversification of *Oryza* coincides with a period marked by extensive cooling and greater aridity around the Miocene/Pliocene boundary (Fig. 1). It has been well established that a shift from closed habitats to open habitats occurred during the late Miocene (Janis, 1993; Cerling *et al.*, 1997), associated with a global expansion of plants using C_4 photosynthesis in tropical and subtropical areas at the expense of C_3 plants (Cerling *et al.*, 1997; Edwards *et al.*, 2010). Meanwhile, significant faunal turnover was observed in many parts of the world, such as Pakistan, North America, South America, Europe and Africa (Cerling *et al.*, 1997). These environmental changes and accompanying new ecological niches presumably played key roles in promoting the second rapid diversification of the rice genus (5–6 Ma). Our speculation is consistent with the observation that the *Oryza* species of recent lineages, including all of the A- and B-genome species, usually exist in open or seasonally dry habitats (Vaughan *et al.*, 2003). Similar bursts of species diversification have been observed in many other organisms around the same time period, including mammals (Janis, 1993; Cerling *et al.*, 1997; Krause *et al.*, 2008), birds (Roy *et al.*, 2001; Fuchs *et al.*, 2007; Voelker *et al.*, 2009), leaf beetles (McKenna & Farrell, 2006), dinoflagellates (Lajeunesse, 2005) and red algae (Lindstrom *et al.*, 1996).

Comparisons of effective population size of ancestors with those of their descendant species provide valuable information on historical population dynamics such as population declines or expansions (Machado *et al.*, 2002; Broughton & Harrison, 2003; Won *et al.*, 2005; Zhang & Ge, 2007). In a study on *Drosophila* species, Machado *et al.* (2002) found that the estimated population size of the common ancestor of *D. pseudoobscure* and *D. persimilis* was significantly larger than that of either descendant species, suggesting that these two species might have experienced population contraction since their time of divergence. In our case, the nucleotide diversity of the extant F- and G-genome species (0.0033 and 0.0032; Ai *et al.*, 2012), which arose from the first rapid diversification, is only one-eighth of that of the ancestral diversity estimate for the genus (0.026, Fig. 1). Similarly for the second rapid diversification, the nucleotide diversity of extant A-, B- and C-genomes (0.0011–0.0073; Ai *et al.*, 2012) is less than one half of the value of their ancestral lineage (0.016–0.025, Fig. 1). The significantly larger population size of the ancestral species relative to extant species in *Oryza* is consistent with the argument that climate fluctuation contributed to the two rapid diversifications of *Oryza* species. When a large ancestral species or population became subdivided, the resulting daughter species would occupy part of the ancestral species range (Broughton & Harrison, 2003). Many lines of evidence based on the paleobotanical and paleofaunal data and isotopic records indicated that grasses underwent a dramatic expansion and became ecologically dominant during the Miocene, with global drying and cooling (Jacobs *et al.*, 1999; Stromberg, 2005). It would be interesting to investigate the population dynamics and speciation process of other groups of grasses that diversified during the same period. Such information may provide further insights into the understanding of patterns and mechanisms of plant species diversification associated with global climate changes.

Acknowledgements

We thank J. Guo for assistance with the R programming and A. N. Egan for her thoughtful comments on the intragenic recombination detection. We also thank Y-F. Wang, F-M. Zhang, B. Ai and L. Tang for their helpful comments on this manuscript. This work was supported by the National Natural Science Foundation of China (30990243).

References

Aggarwal RK, Brar DS, Khush GS. 1997. Two new genomes in the *Oryza* complex identified on the basis of molecular divergence analysis using total genomic DNA hybridization. *Molecular and General Genetics* **254**: 1–12.

Ai B, Wang ZS, Ge S. 2012. Genome size is not correlated with effective population size in the *Oryza*. *Evolution* **66**: 3302–3310.

Ammiraju JSS, Lu F, Sanyal A, Yu Y, Song X, Jiang N, Pontaroli AC, Rambo T, Currie J, Collura K *et al.* 2008. Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell* **20**: 3191–3209.

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics* **33**: 707–740.

Broughton RE, Harrison RG. 2003. Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. *Genetics* **163**: 1389–1401.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution* **25**: 1979–1994.

Cerling TE, Harris JM, MacFadden BJ, Leakey MG, Quade J, Eisenmann V, Ehleringer JR. 1997. Global vegetation change through the Miocene/Pliocene boundary. *Nature* **389**: 153–158.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**: 195–205.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics* **68**: 444–456.

Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. 2004. Darwin's abominable mystery: insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences, USA* **101**: 1904–1909.

Edwards EJ, Osborne CP, Stromberg CAE, Smith SA, Bond WJ, Christin PA, Cousins AB, Duvall MR, Fox DL, Freckleton RP *et al.* 2010. The origins of C_4 grasslands: integrating evolutionary and ecosystem science. *Science* **328**: 587–591.

Fuchs J, Ohlson JI, Ericson PGP, Pasquet E. 2007. Synchronous intercontinental splits between assemblages of woodpeckers suggested by molecular data. *Zoologica Scripta* **36**: 11–25.

Gaut BS. 1998. Molecular clocks and nucleotide substitution rates in higher plants. In: Hecht MK, MacIntyre RJ, Clegg MT, eds. *Evolutionary biology*. New York, NY, USA: Plenum Press, 93–120.

Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences, USA* **93**: 10 274–10 279.

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**: 14 400–14 405.

Guo YL, Ge S. 2005. Molecular phylogeny of Oryzeae (Poaceae) based on DNA sequences from chloroplast, mitochondrial, and nuclear genomes. *American Journal of Botany* **92**: 1548–1558.

Heaney LR. 1991. A synopsis of climatic and vegetational change in Southeast-Asia. *Climatic Change* **19**: 53–61.

Heer O. 1855. *Flora tertiaria helvetiae*. Paris, France: Winterthur.

Inoue J, Donoghue PCJ, Yang ZH. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Systematic Biology* **59**: 74–89.

Jacobs BF, Kingston JD, Jacobs LL. 1999. The origin of grass-dominated ecosystems. *Annals of the Missouri Botanical Garden. Missouri Botanical Garden* **86**: 590–643.

Janis CM. 1993. Tertiary mammal evolution in the context of changing climates, vegetation, and tectonic events. *Annual Review of Ecology and Systematics* **24**: 467–500.

Jennings WB, Edwards SV. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* **59**: 2033–2047.

Khush GS. 2001. Green revolution: the way forward. *Nature Reviews Genetics* **2**: 815–822.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge, UK: Cambridge University Press.

Krause J, Unger T, Nocon A, Malaspinas AS, Kolokotronis SO, Stiller M, Soibelzon L, Spriggs H, Dear PH, Briggs AW *et al.* 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evolutionary Biology* **8**: 220.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. *Nature* **392**: 917–920.

Lajeunesse TC. 2005. "Species" radiations of symbiotic dinoflagellates in the Atlantic and Indo-Pacific since the Miocene–Pliocene transition. *Molecular Biology and Evolution* **22**: 570–581.

Lindstrom SC, Olsen JL, Stam WT. 1996. Recent radiation of the Palmariaceae (rhodophyta). *Journal of Phycology* **32**: 457–468.

Lu F, Ammiraju JSS, Sanyal A, Zhang SL, Song RT, Chen JF, Li GS, Sui Y, Song X, Cheng ZK *et al.* 2009. Comparative sequence analysis of *MONOCULM1*-orthologous regions in 14 *Oryza* genomes. *Proceedings of the National Academy of Sciences, USA* **106**: 2071–2076.

Machado CA, Kliman RM, Markert JA, Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Molecular Biology and Evolution* **19**: 472–488.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**: 2462–2463.

McKenna DD, Farrell BD. 2006. Tropical forests are both evolutionary cradles and museums of leaf beetle diversity. *Proceedings of the National Academy of Sciences, USA* **103**: 10 947–10 951.

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

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

Posada D. 2001. Unveiling the molecular clock in the presence of recombination. *Molecular Biology and Evolution* **18**: 1976–1978.

Prasad V, Stromberg CA, Leache AD, Samant B, Patnaik R, Tang L, Mohabey DM, Ge S, Sahni A. 2011. Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. *Nature Communications* **2**: 480.

Quade J, Solounias N, Cerling TE. 1994. Stable isotopic evidence from paleosol carbonates and fossil teeth in Greece for forest or woodlands over the past 11 Ma. *Palaeogeography, Palaeoclimatology, Palaeoecology* **108**: 41–53.

R Development Core Team. 2008. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org.

Rambaut A, Drummond AJ. 2007. *Tracer v1.4*. [WWW document] URL http://beast.bio.ed.ac.uk/Tracer.

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 Z. 2007. Inferring speciation times under an episodic molecular clock. *Systematic Biology* **56**: 453–466.

Rice WR. 1989. Analyzing tables of statistical tests. *Evolution* **43**: 223–225.

Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM. 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science* **293**: 2242–2245.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. *PLoS Biology* **4**: e352.

Roy MS, Sponer R, Fjeldsa J. 2001. Molecular systematics and evolutionary history of akalats (Genus *Sheppardia*): a pre-Pleistocene radiation in a group of African forest birds. *Molecular Phylogenetics and Evolution* **18**: 74–83.

Sang T, Ge S. 2007. Genetics and phylogenetics of rice domestication. *Current Opinion in Genetics & Development* **17**: 533–538.

Sanyal A, Ammiraju JS, Lu F, Yu Y, Rambo T, Currie J, Kollura K, Kim HR, Chen J, Ma J *et al.* 2010. Orthologous comparisons of the *Hd1* region across genera reveal *Hd1* gene lability within diploid *Oryza* species and disruptions to microsynteny in sorghum. *Molecular Biology and Evolution* **27**: 2487–2506.

Satta Y, Hickerson M, Watanabe H, O'hUigin C, Klein J. 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. *Journal of Molecular Evolution* **59**: 478–487.

Second G. 1985. Evolutionary relationships in the sativa group of *Oryza* based on isozyme data. *Genetics, Selection, Evolution* **17**: 89–114.

Soltis DE, Kuzoff RK. 1995. Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). *Evolution* **49**: 727–742.

Stromberg CAE. 2005. Decoupled taxonomic radiation and ecological expansion of open-habitat grasses in the Cenozoic of North America. *Proceedings of the National Academy of Sciences, USA* **102**: 11 980–11 984.

Takahata N. 1986. An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. *Genetical Research* **48**: 187–190.

Takahata N, Satta Y. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proceedings of the National Academy of Sciences, USA* **94**: 4811–4815.

Takahata N, Satta Y, Klein J. 1995. Divergence time and population-size in the lineage leading to modern humans. *Theoretical Population Biology* **48**: 198–221.

Tang L, Zou XH, Achoundong G, Potgieter C, Second G, Zhang DY, Ge S. 2010. Phylogeny and biogeography of the rice tribe (Oryzeae): evidence from combined analysis of 20 chloroplast fragments. *Molecular Phylogenetics and Evolution* **54**: 266–277.

Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**: 689–702.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**: 1647–1657.

Tiffney BH, Manchester SR. 2001. The use of geological and paleontological evidence in evaluating plant phylogeographic hypotheses in the Northern Hemisphere tertiary. *International Journal of Plant Sciences* **162**: S3–S17.

Vaughan DA, Kadowaki K, Kaga A, Tomooka N. 2005. On the phylogeny and biogeography of the genus *Oryza. Breeding Science* **55**: 113–122.

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

Voelker G, Rohwer S, Outlaw DC, Bowie RCK. 2009. Repeated trans-Atlantic dispersal catalysed a global songbird radiation. *Global Ecology and Biogeography* **18**: 41–49.

Wall JD. 2003. Estimating ancestral population sizes and divergence times. *Genetics* **163**: 395–404.

White SE, Doebley JF. 1999. The molecular evolution of *terminal ear1*, a regulatory gene in the genus Zea. *Genetics* **153**: 1455–1462.

Wing RA, Ammiraju JS, Luo M, Kim H, Yu Y, Kudrna D, Goicoechea JL, Wang W, Nelson W, Rao K *et al.* 2005. The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species. *Plant Molecular Biology* **59**: 53–62.

Won YJ, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. *Proceedings of the National Academy of Sciences, USA* **102**(Suppl 1): 6581–6586.

Yang Z. 1997. On the estimation of ancestral population sizes of modern humans. *Genetical Research* **69**: 111–116.

Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* **162**: 1811–1823.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.

Yang Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biology and Evolution* **2**: 200–211.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* **23**: 212–226.

Zachos J, Pagani M, Sloan L, Thomas E, Billups K. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**: 686–693.

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.

Zhou HF, Zheng XM, Wei RX, Second G, Vaughan DA, Ge S. 2008. Contrasting population genetic structure and gene flow between *Oryza rufipogon* and *Oryza nivara. TAG. Theoretical and Applied Genetics.* **117**: 1181–1189.

Zhou RC, Zeng K, Wu W, Chen XS, Yang ZH, Shi SH, Wu CI. 2007. Population genetics of speciation in nonmodel organisms: I. Ancestral polymorphism in mangroves. *Molecular Biology and Evolution* **24**: 2746–2754.

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

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.

Fig. S1 Effect of the priors for fossil calibration and rate heterogeneity across branches on posterior estimates of divergence time.

Fig. S2 The posterior means of divergence time were plotted against the width of corresponding 95% Confidence Intervals (CIs) for each node.

Fig. S3 Relative rate or among-locus rate variation of 66 loci.

Table S1 Information on 106 loci used in the present study

Table S2 Posterior distributions of θ values for ancestral populations based on 66 loci using different prior means of θ of 0.005, 0.01, 0.02, 0.03 and 0.05

Table S3 Posterior distributions of θ values for ancestral populations based on the reduced dataset without recombination signal using different priors mean of θ from 0.005, 0.01, 0.02, 0.03 to 0.05

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.