Phylogenetic relationships and genome divergence among the AA- genome species of the genus Oryza as revealed by 53 nuclear genes and 16 intergenic regions

Ting Zhu, Ping-Zhen Xu, Jiang-Peng Liu, Sheng Peng, Xin-Chun Mo, Li-Zhi Gao

Abstract

Rapid radiations have long been regarded as the most challenging issue for elucidating poorly resolved phylogenies in evolutionary biology. The eight diploid AA-genome species in the genus Oryza represent a typical example of a closely spaced series of recent speciation events in plants. However, questions regarding when and how they diversified have long been an issue of extensive interest but remain a mystery. Here, a data set comprising >60 kb of 53 singleton fragments and 16 intergenic regions is used to perform phylogenomic analyses of all eight AA-genome species plus four diploid Oryza species with BB-, CC-, EE- and GG-genomes. We fully reconstruct phylogenetic relationships of AA-genome species with confidence. Oryza meridionalis, native to Australia, is found to be the earliest divergent lineage around 2.93 mya, whereas O. punctata, a BB-genome species, serves as the best outgroup to distinguish their phylogenetic relationships. They separated from O. punctata approximately 9.11 mya during the Miocene epoch, and subsequently radiated to generate the entire AA-genome lineage diversity. The success in resolving the phylogeny of AA-genome species highlights the potential of phylogenomics to determine their divergence and evolutionary histories.

1. Introduction

Rice acts as the world’s staple food crop, providing daily caloric needs for more than half of the global population. Future increases in rice production call for genetic improvement of rice cultivars to feed the growing population. The rice genus Oryza L comprises approximately 21 wild and two cultivated species distributed in Asia, Africa, Australia and the Americas (Vaughan, 1989; Vaughan et al., 2003). These species have been classified into ten distinct genome types, represented by six diploids (AA, BB, CC, EE, FF, and GG) and four allotetraploid species (BBCC, CCDD, HHJJ, and HHKK) (Aggarwal et al., 1997; Ge et al., 1999, 2001). Since wild species of Oryza contain traits of agronomic interest, such as the resistance and tolerance to biotic and abiotic stresses (Chang, 1976; Sitch et al., 1989; Vaughan, 1989; Khush, 1997), they serve as a valuable reservoir for rice genetic improvement. Nevertheless, the majority of the Oryza germplasm has been actively utilized in rice improvement programs due to their sterility barriers (Brar and Khush, 1997; Piegu et al., 2006). Genetic resources of the AA-genome group, also called the Oryza sativa complex, have long been attractive for rice breeders. As the most recently diverged lineage in Oryza (Vaughan, 1989; Ge et al., 2001), sexual hybridization between cultivated rice and wild species with AA-genomes is easier than with other genome types. The resolution of phylogenetic relationships among the AA-genome species of Oryza will be essential in directing future efforts to explore beneficial alleles and thus greatly facilitate efficient germplasm conservation and utilization in rice breeding programs.

The Oryza sativa complex grows world-wide and consists of eight diploid species (Vaughan, 1989). Asian cultivated rice, O. sativa L, with two main subspecies, O. sativa ssp. indica and O. sativa ssp. japonica (Nayar, 1973; Oka, 1988; Morishima et al., 1992) is of Asian origin and is globally cultivated today. Of its two presumed wild progenitors, the perennial O. rufipogon is distributed throughout tropical Asia and Oceania, whereas the annual O. nivara occurs in tropical continental Asia. Another cultivated species in the genus, O. glaberrima, was parallel domesticated in West Africa where it is endemic (Vaughan, 1989). There are two additional wild species also endemic to Africa, O. barthii and O. longistaminata. The former is the annual wild progenitor of O. glaberrima, while the latter is a perennial, rhizomatous and partially self-incompatible grass species (Ghesquiere, 1987). The
Australian endemic *O. meridionalis* is a primarily inbred annual species, while the South American endemic *O. glumaepatulata* varies in its perennial period (Vaughan, 1989). Sterility barriers, which have been extensively investigated, vary between pairs of AA-genome species (Sitch, 1990; Naredo et al., 1998). The variability of sterility barriers among species is indicative of their incomplete speciation. Different extent of subsequent interspecies introgression apparently increases a risk to clearly trace the evolution and taxonomy of AA-genome species. Indeed, taxonomic difficulties still remain in the AA-genome group due to the lack of instability of clear morphological distinguishing characteristics (Vaughan, 1989; Juliano et al., 1998; Lu et al., 2000). To efficiently enhance germplasm exploration of these wild species, there have been continued efforts put to investigate the evolution and phylogeny of the AA-genome species over the past decades. A wealth of data from the nuclear, chloroplast and mitochondrial genomes (Ge et al., 1999; Guo and Ge, 2005; Zhu and Ge, 2005; Chuayjaeng and Volkert, 2006; Duan et al., 2007) has improved our understanding of the diversification and phylogeny of the AA-genome species. However, evolutionary relationships of these species are inconsistent from one study to another and have not been fully resolved. The basal split in the AA-genome group, for example, is particularly problematic. In agreement with previous investigations (Iwamoto et al., 1999), the *O. longistaminata* accessions might markedly diverge from all other AA-genome species (Cheng et al., 2002; Ren et al., 2003). Nevertheless, *O. meridionalis* was recognized as the earliest species of the AA-genome *Oryza* species with data from RFLPs (Wang et al., 1992; Doi et al., 2000), MITE-AFLPs (Park et al., 2003), and intron sequences of four nuclear genes (Zhu and Ge, 2005). In addition, phylogenetic positions of *O. longistaminata* and *O. glumaepatulata* in the AA-genome group are either still contradictory each other or show limited resolution (Zhu and Ge, 2005; Duan et al., 2007). To thoroughly elucidate their genome evolution, however, these studies lacked genome-scale sampling, such as phylogenetic breadth of marker genes with broad functional representation and randomly chosen intergenic genomic regions of chromosome dimension.

Increasing evidence has shown that the reliance on a single dataset might result in misleading phylogenetic inferences to analytical inconsistency and biological processes (Wendel and Doyle, 1998; Zou et al., 2008). The incongruence between phylogenies may become one of the most pervasive challenges when reconstructing molecular phylogenetics by using different datasets, such as individual genes. The current century has witnessed new phylogenetic reconstructions that use large quantities of genome-wide markers to elucidate previously controversial evolutionary relationships at all taxonomic levels (Bapteste et al., 2002; Rokas et al., 2003a; Hackett et al., 2008; Zou et al., 2008; Cranston et al., 2009). In effect, a gene tree does not necessarily reflect a species tree, even if the orthology of marker genes are clearly identified and employed. Thus, a large number of genetic markers, including unlinked loci with extensive functional representation as well as intergenic genomic regions, are needed to comprehensively track organismal history. Such a robust phylogeny will build a foundation for future insights into rice genome evolution.

In light of recent radiation and rapid diversification of the AA-genome species, questions regarding when and how these genomes diverged have long been an issue of extensive interest but still remain a mystery. The completion of the rice genomes (Yu et al., 2002; International Rice Genome Sequencing Project, 2005) has made the genus *Oryza* become an attractive model to study comparative and evolutionary genomics in the grass family (Shimamoto and Kyozuka, 2002; Wing et al., 2005; Bennetzen, 2007). By using the rice genome as a reference, increasing genome-wide comparative analyses have been performed among species in the genus *Oryza* (Ma and Bennetzen, 2004; Aamiraju et al., 2006, 2008, 2010; Piegu et al., 2006; Zhang et al., 2007; Huang et al., 2008; Lu et al., 2009; Roorkiwal et al., 2009; Chen et al., 2013). Since the complete genome sequences of all wild AA-genome species will not be fully sequenced in the foreseeable future, estimates of genomic divergence and divergence times among them from genome-wide sequences would be a worthwhile endeavor for the rice scientific community. A comparison of homologous sequences between pairs of species not only helps to trace their evolutionary history, but also provides important hints to investigate critical functional changes after their divergence.

Here, 53 single-copy nuclear genes representing diverse functional categories, together with 16 intergenic regions distributed throughout the genome, were sequenced for all eight AA-genome species as well as four diploid species with BB-, CC-, EE-, and GG-genomes in the genus *Oryza*. The study aims to reconstruct phylogenetic relationships among the AA-genome species in depth and identify their basal lineage. A concatenated dataset was also used to estimate genomic divergences and split times of AA-genome species in comparisons with closely related diploid species with BB-genome in the genus *Oryza*.

### 2. Materials and methods

#### 2.1. Plant materials

A total of nine accessions of eight AA-genome species were sampled, including *O. sativa* (one accession each for the *indica* and *japonica* subspecies, respectively), *O. glaberrima*, *O. rufipogon*, *O. nivara*, *O. glumaepatulata*, *O. meridionalis*, *O. longistaminata*, and *O. barthii*. To better elucidate their evolutionary relationships, we also sampled one accession each of the BB-genome species (*O. punctata*), the CC-genome species (*O. officinalis*), the EE-genome species (*O. australis*), and the GG-genome species (*O. granulata*). These four non-AA genome species, representing the majority of diploid genomes, are related to the AA-genome species in the genus (*Ge et al., 1999, 2001*).

Of them, ten accessions were provided by the International Rice Research Institute (Manila, Philippines), while three accessions were collected in Yunnan Province, China. Seeds were germinated on filter paper and the seedlings were transplanted into soil in the greenhouse. All the studied accessions with their scientific names, geographic origins and genome type were given in Table 1, and these materials are available upon request.

#### 2.2. Selection of single-copy genes and primer design

To ensure the chosen genes were single-copy genes in the genome, we downloaded the database of the RGAP rice genome (version 7.0, [http://www.rice.plantbiology.msu.edu/](http://www.rice.plantbiology.msu.edu/)) (Kawahara et al., 2013) for gene screening. We extracted the protein sequences with nr-KOME cDNA evidence and then performed extensive searches against the *japonica* genome sequences in all six reading frames using TBLASTN at an e-value cutoff of 10⁻⁷. To select applicable markers, we searched for target genes based on the following (1) they should be a singleton in both rice and *Sorghum bicolor* genomes at the node of grass post-duplication, according to Phytozone V8.0 ([http://www.phytozone.net/](http://www.phytozone.net/)), and thus kept away from the paralogy problem, (2) they should be unlinked, preferably on different chromosomes, and thus represent historically independent markers, and (3) they should have extensive representation of their functional diversity. Although there are two copies in both *S. bicolor* and *Zea mays*, *Brachypodium distachyon*, *Arabidopsis thaliana*, *Vitis vinifera* and *Glycine max* and thus was also included for further analyses.
The sequences of these putative genes were first downloaded from the RGAP rice genome (version 7.0, http://www.rice.plantbiology.msu.edu/). To further refine their annotations, similarity searches were performed by comparing against NCBI databases. A total of 53 single-copy genes were ultimately sampled. They all randomly distribute throughout the 12 rice chromosomes and represent at least two major categories of gene function relevant to either developmental or ecological adaptations. Phylogenetic performance of these genes has not yet been evaluated until now. Supplementary Table S1 summarizes properties of the whole genome set.

Among non-coding regions of these nuclear genes, introns evolve fast and may provide more informative sites. As a result, non-coding sequences were prevalently chosen as phylogenetically informative characters, especially at a lower taxonomic level (Mason-Gamer et al., 1998). Of the 53 surveyed single-copy gene fragments in this study, 11 introns longer than 500 bp were successfully amplified and included in the subsequent data analyses. The PCR primers used for the amplification (Supplementary Table S2) were designed by using Primer premier 5 using reference genomic sequences is usually neutral, they were ideally used to reconstruct the phylogeny (Lee and Wen, 2004; Johnson et al., 2008) to estimate the divergence between genomes. We randomly sampled a total of 60 intergenic regions throughout the whole rice genome and selected 42 regions after the removal of TE s using RepeatMasker (http://www.repeatmasker.org/). Of these 16 intergenic regions that were amplified and conserved across all the examined species were further analyzed. Primers for the intergenic regions were also designed based on reference sequences of the japonica genome. Since the evolution of intergenic genomic sequences is usually neutral, they were ideally used to examine the hypothesis of a molecular clock.

2.3. DNA isolation, amplification and sequencing

Total genomic DNA was isolated from fresh leaf tissues using the CTAB method (D oyle and Doyle, 1987). The quantity and quality of the extracted DNA were checked by electrophoresis on 1% agarose gel. PCR protocols were taken as follows: 2.5 μl buffer, 2.0 μl dNTPs, 0.15 μl Taq polymerase, 0.5 μl each of 10 μM solutions of the two primers, and 1 μl DNA template were included for each 25 μl reaction. Reaction mixtures were heated to 94°C for 5 min and then subjected to 35 cycles of 94°C for 30 s, 50–55°C for 1 min, and 72°C for 1–1.5 min, finally followed by 72°C for 10 min in the Biometra T1 thermocycler (Waltman, American). PCR products with expected size were purified by using the EZ Spin Column PCR product purification Kit (BBI), and then double-strand sequenced with two primers used for PCR amplifications at Beijing Genomics Institute, Shenzhen, China. The fragments which failed by direct sequencing were ligated into pMD18-T vector (Takara) and then cloned into Escherichia coli. Totally or partially, missing sequences existed for some gene fragments due to difficulties in amplifying as shown in Supplementary Table S2. All sequences reported in this study were deposited in the GenBank database under accession numbers KC609022-KC609255, KC610818-KC611172, KC626093-KC626136, KC679079-KC679237, and KC700663-KC700674. (Supplementary Tables S2 and S4).

2.4. Sequence alignment and analyses

The obtained gene sequences were annotated based on annotations of the RGAP rice genome (version 7.0). Individual sequences were aligned with Clustal W version 1.81 (Thompson et al., 1997) and then adjusted manually. Sequences of ambiguous alignment were excluded from our data analyses. We plotted the number of transitions and transversions vs. divergence to provide a visual display of substitution saturation with an asymptotic relationship using DAMBE 4.5.45 (Xia and Xie, 2001). Without including outgroups for any case, the GC content, parwise divergence, variable site frequency, and parsimony-informative site frequency were calculated by using SeqState ( Muller, 2005). Gaps in the alignments were coded as present/absent characters, following the simple gap coding method of Simmons and Ochoterena (Simmons and Ochoterena, 2000) as implemented by SeqState ( Muller, 2005). Gaps were treated as either missing data or new characters. We also performed an analysis with gap characters alone to assess phylogenetic information of indels in the dataset. Finally we used maximum likelihood methods implemented in HyPhy (Pond and Muse, 2005) to test the heterogeneity in the rate of synonymous substitution (dS) and nonsynonymous substitution (dN) for each gene under the hypothesis of a molecular clock.

2.5. Consensus network analyses

Consensus networks (Bandelt, 1995; Holland and Moulton, 2003) were built to visualize the extent of contradiction among phylogenetic trees, which involved combining splits from different trees into a potentially hyperdimensional graph (Holland et al., 2004). With the BB-, CC-, and EE- genome species as outgroups, respectively, consensus networks were individually constructed by optimal Bayesian inference (BI) trees, which were obtained from datasets of 53 single-copy genes and 16 intergenic genomic sequences using Splitstree 4.10 (Huson and Bryant, 2006). To examine the conflict among topologies of gene trees, we did not take branch lengths into consideration when
importing optimal BI trees as source data. Thus, edge lengths in the final network are proportional to the number of trees in which a particular split appears. To illustrate incongruity between individual gene trees, all splits occurred at a threshold value ranging from 0.05 to 0.25.

### 2.6. Selection of outgroup

The root of the phylogenetic trees of AA-genome species is unclear, and the choice of outgroup may have a significant impact on phylogenetic reconstruction around this node. We thus chose independent outgroups of the BB-, CC- and EE-genome species in the genus *Oryza* to individually root different datasets to test whether a given outgroup could be used alone to reflect the phylogeny of the tree, particularly to determine the basal lineage of the studied group. We took the approach used in multi-criteria decision-making which ensures that we are able to choose the adequate outgroup species and uncover effects of outgroup choice on the construction of phylogenetic trees. In our case, the two evolutionary characters, genetic distances and compositional qualities, were included. We also applied information from these preliminary tree searches as an additional signal of outgroup adequacy and used the consensus network to determine the proper outgroup.

### 2.7. Phylogenetic estimates of concatenated datasets

We first produced a large concatenated data matrix, including all single-copy genes and intergenic sequences from the studied species. The dataset was then divided into two groups, one with 16 intragenic genomic sequences and the other with 53 single-copy genes. We further separated the latter dataset into two subgroups comprising coding regions and non-coding sequences, respectively. Phylogenetic trees were reconstructed by neighbor joining (NJ), maximum parsimony (MP) and BI. NJ and MP were implemented with PAUP 4.0b10 (Swofford, 2003). Kimura 2-parameter distance estimate was used in the NJ method, whereas MP analyses were performed by heuristic search with MULPARS option and tree-bissection-reconnection (TBR) branch-swapping. Topological robustness was assessed by bootstrap analysis with 1000 replicates using simple taxon addition. Characters were equally weighted in all phylogenetic analyses. Appropriate nucleotide substitution models for each dataset were determined by Modeltest 3.7 (Posada and Crandall, 1998). The best models were chosen according to the Akaike Information Criterion (AIC) and then used for subsequent Bayesian analyses. BI was conducted with MrBayes 3.12 (Ronquist and Huelsenbeck, 2003) with MCMC estimation of posterior probability distributions. Four parallel MCMC were run for 3,000,000 generations with trees sampled every 100 generations. For all BI analyses, the first 5000 sampled trees from each run were discarded as burn-in to ensure that the chains reached stationary.

In addition, we attempted to construct phylogenetic trees by using the amino acid sequences of single-copy genes. ModelGenerator (Keane et al., 2006) was first used to determine the best-fitting evolutionary model of the proteins according to the Bayesian Information Criterion. Then, a discrete gamma JTT model of amino acid substitution was applied. The maximum likelihood method with a PhyML (Guindon et al., 2005) algorithm was employed to reconstruct phylogenetic trees of amino acid sequences with the BB-, CC- and EE-genome species as outgroups, respectively.

### 2.8. Tests of relative rates

Genomic divergences between pairs of *Oryza* species were calculated using the Jukes–Cantor method in MEGA4 program (Nei and Gojobori, 1986; Tamura et al., 2007). Divergence times between lineages were estimated by following the hypothesis of molecular clock with the formula \( T = K/2r \), where \( r \) refers to the absolute rate of substitutions/site/year and \( K \) is the estimated number of substitutions per site between homologous sequences. To test whether the relative rates between different lineages are heterogeneous, we performed Tajima’s relative rate test (Tajima, 1993) among lineages by using BB as outgroup. The congruence between different datasets was calculated using the partition homogeneity test, as implemented in PAUP version 4.0b10 (Swofford, 2003) using 1000 replicates. The resulting P-values were used to determine whether the datasets from different sources exhibited significant incongruence.

### 3. Results

#### 3.1. Comparisons of individual datasets

To reconstruct reliable phylogenetic relationships, it is necessary to make certain the orthology of the applied marker genes. In this study, we selected single-copy genes in both *O. sativa* and *S. bicolor*, which belong the BEP and PACCAD clades (Bouchenak-Khelladi et al., 2008), respectively. Single-copy genes detected within these two genomes indicated that they might not have experienced any duplication events after the split from their common ancestor, thus confirming them to be orthologous. After an extensive screen of the whole rice genome, we identified and sequenced 53 nuclear single-copy genes as well as 16 intragenic genomic regions distributed throughout 12 chromosomes for all eight AA-genome species and four diploid species with BB-, CC-, EE-, and GG-genomes as well. Of 53 single-copy genes represented diverse biological functions, for example, eight are relevant to ecological adaptations, while twenty-one are involved in developmental processes. These genes were first tested under the hypothesis of molecular clock by performing relative rate tests for all the datasets, and we found that nine genes had no rate heterogeneity at synonymous sites and eight at nonsynonymous sites. The length of aligned sequences for each locus varied from 485 bp to 2483 bp, and the mean GC content of these genes ranged from 30.6% to 67.4%. They appeared similar for the same gene across species, but varied greatly among genes. Our results indicate that the candidate genes were randomly sampled from the genome and thus adequate information becomes available at different taxonomic levels. Other detailed information of these single-copy genes are presented in Supplementary Table S5.

A sequence data matrix of 60,821 bp was first generated for a total of 69 sequences, which include 53 single-copy gene fragments and 16 intragenic genomic regions from accessions of nine AA-genome species. Here, the variable sites were 2876 bp (4.7%), including 762 bp (1.3%) informative sites, and the indels of this aligned matrix were 6485 bp. Next, we separately analyzed the single-copy genes and intragenic genomic regions. Genomic sequences of 53 single-copy genes resulted in a data matrix of 52,920 bp, with exons (17,326 bp) accounting for 32.7%. The concatenated data matrix contained 2640 (5.0%) variable sites and 735 (1.4%) parsimony-informative sites, and the indels after alignment were 5308 bp. Given different evolutionary behaviors between exons and non-coding regions, we further divided the dataset of single-copy genes into two subgroups. In addition to a larger incorporated dataset of 36,341 bp for introns than that of 17,326 bp for exons, introns apparently showed both more variable sites of 1664 bp (4.6%) and informative sites 596 bp (1.6%) than exons with 445 (2.5%) variable sites and 144 (0.8%) parsimony-informative sites. Moreover, the indels after alignment were 676 bp and 4461 bp for exons and non-coding regions,
respectively, suggesting that the majority of insertion and deletion mutations might be apt to occurring within non-coding regions. Overall, 16 intergenic regions were integrated into a dataset containing 8026 characters, with 318 bp (4.0%) variable sites and 65 bp (0.8%) parsimony-informative sites, and the indels after alignment were 1312 bp. The above-described information for all taxa is given in Table 2.

3.2. Phylogenetic analyses based on individual single-copy genes and intergenic regions

One of the most difficulties in phylogenetic reconstruction is widespread occurrence of the incongruence between alternative phylogenies generated from single data sets (Rokas et al., 2003b). Incongruence may occur at all taxonomic levels, from phylogenies of closely related species to relationships between major classes or phyla and higher taxonomic groups (Mason-Gamer and Kellogg, 1996; Giribet et al., 2001; Hwang et al., 2001; Kopp and True, 2002). To individually examine the extent of incongruence, we reconstructed phylogenetic trees of the 53 single-copy gene fragments and 16 intergenic regions using the BI method as an example. Here, BB-, CC-, and EE- genome species were independently used as outgroups. The obtained 160 BI trees were evidence for a substantial incongruence with more than 20 topologies, not considering those generated by using different outgroups (Supplementary Tables S6 and S7). Note that a number of sharply different scenarios were observed when these loci were analyzed separately. Some gene trees yielded a topology identical to the topology of the species tree with varied posterior probabilities, although others were able to better resolve some clades but not others. However, the remaining gene trees appeared in sharp contrast to the topology of the assumed species tree. Among the collected trees, we observed that conflicting clades broadly existed, which might be caused by different evolutionary histories of these genes or systematic errors. In addition, we obtained numerous gene trees with “bush-like” topologies, particularly among O. sativa. ssp. indica, O. sativa. ssp. japonica as well as their wild ancestors of O. nivara and O. rufipogon. Of them, posterior probabilities varied widely from one clade to another, and there were no branches strongly supported by most genes with enough high BI values. Such findings show that sufficient information was always absent to distinguish phylogenetic relationships of lower taxonomic groups in particular. It implies that a single or fewer marker genes from either single-copy gene fragments or intergenic regions could not alone actually resolve phylogenetic relationships of the AA- genome species, leaving some species unsettled. Our results thus demonstrate that extensive incongruence is ubiquitously present among gene phylogenies, and fewer gene fragments cannot result in the resolution of phylogeny trees, especially at lower taxonomic levels.

In order to investigate the extent of incongruence of phylogenetic reconstruction, we applied consensus networks to visualize the contradictions among candidate genes. Here, the consensus network of 63 small trees with BB- genome species as outgroup was constructed by means of media network (Fig. 1A). Then, we examined features of consensus networks by increasing the threshold from 0.05 to 0.25 (Fig. 1A). When the threshold reached 0.25, the intricacy of network cycles decreased and collapsed. Nevertheless, at a threshold of 0.05, the network appeared a structure consisting of high-dimensional hypercube, showing quite complicated phylogenetic patterns among gene trees and a variable extent of contradiction. At a threshold of 0.10, there was some uncertainty in the phylogeny with respect to O. glaberrima / O. barthii and O. glumaepatula. Such an uncertainty also existed in the placement of the root position of O. longistaminata, O. meridionalis and O. punctata; at a threshold of 0.15, we found that the ambiguity of O. glaberrima/O. barthii, O. glumaepatula was hidden. However, the uncertainty of O. longistaminata, O. meridionalis and O. punctata still held until the threshold increased to 0.25. At a threshold of 0.25, the tree ended up with an unique bifurcating tree structure, evidently indicating that the data were quite tree-like in nature. Common phylogenetic signals existed among large numbers of independent genes and/or intergenic genomic regions make them overcome the incongruence governed mainly by their different evolutionary histories. As shown in Fig. 2, the consensus network analyses of small trees similarly exhibited phylogenetic patterns when using CC- and EE- genome species as outgroups.

3.3. Impacts of outgroup selection on phylogenetic reconstruction

The uncertainty of the root position, either O. meridionalis or O. longistaminata, still needs to be further confirmed among the AA-genome species. To better identify the root position of the AA-genome species, it is necessary to identify the best outgroup and assess the impact of different outgroup choice on phylogenetic analyses. For all these gene trees generated by single-copy genes and intergenic regions, for example, there were only 25 gene or intergenic regions (36%) that reached the same tree topology when reconstructed with different outgroups, indicating that outgroup choice indeed affects the reconstruction of their phylogeny. We employed BB-, CC- and EE- genome species in the genus Oryza to determine an appropriate candidate outgroup. Given the application of different outgroups to our analyses, some loci exhibited the same topology in spite of different outgroups, while others made quite different topologies, which might be affected by the choice of outgroup (Supplementary Tables S6 and S7).

Table 2

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Description of sequence information for different datasetsa.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aligned length (range in bp)</td>
</tr>
<tr>
<td>All sites</td>
<td>60821 (56714–58021)</td>
</tr>
<tr>
<td>Single-copy genes</td>
<td>52920 (49446–51118)</td>
</tr>
<tr>
<td>Exons</td>
<td>17326 (16737–17259)</td>
</tr>
<tr>
<td>Introns</td>
<td>36341 (33480–34645)</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>8026 (6966–7609)</td>
</tr>
</tbody>
</table>

a The sequences of outgroups were not included in data analysis.
To assess the impact of outgroup choice on phylogenetic analyses, we selected 44 gene trees with BB-, CC- and EE- genome species as outgroups to build the consensus network (Fig. 1B). For the same dataset at the same threshold, stimulations of the consensus network were found to be consistent to each other. The generation of reticulating trees was different by reason of different rooting species. When comparing reticulations among three consensus networks, we found that the intricacy of network cycles was different from one outgroup to another. At a threshold of 0.1, the network presented branches appearing at a frequency of 10% or higher, and there was a three-dimensional cycle when EE- genome species served as outgroup; several extra two-dimensional ones were observed while CC- genome species was applied; and when we rooted the network by BB- genome species as outgroup, the intricacy of the network was much smaller in the presence of only a two-dimensional cycle at the junction of ingroup and outgroup. The most suitable outgroup, BB- genome species, resulted in smaller threshold values, and the intricate relationships between the networks were more stable in the same thresholds values (Figs. 1 and 2). The above-mentioned results indicate that intricate relationships between outgroup and ingroup might essentially depend on outgroup and that inadequate taxon sampling with improper choice of outgroup is apt to causing topological incongruence. It is likely that the root of any phylogenetic tree appears unstable, but an appropriate outgroup may largely account for the stability of phylogenetic trees. Since the BB- genome species, *O. punctata*,

Fig. 1. Consensus networks constructed from BI trees with different thresholds and outgroups. (A) Consensus networks from a collection of 63 optimal BI trees reconstructed by single-copy genes and intergenic regions with merely BB- genome species as outgroup. The threshold was 0.05 (I), 0.1 (II), 0.15 (III) and 0.25 (IV), respectively; (B) consensus networks from a collection of optimal BI trees of both single-copy genes and intergenic regions with BB- (I), CC- (II) and EE- (III) genome species as outgroup, respectively. The threshold of 0.1 was used for all the analyses.
Fig. 2. Consensus networks constructed from BI trees using CC and EE as outgroups with different thresholds. Consensus network of collection of 49 optimal BI tree of single gene and intergenic regions with (A) CC- and (B) EE- as outgroup. The threshold was 0.05 (I), 0.1 (II), 0.15 (III), 0.2 (IV) and 0.25 (V), respectively.

Fig. 3. Saturation analyses in the concatenated datasets. Saturation analyses of the concatenated datasets were performed for (A) single-copy genes, (B) intergenic regions, (C) introns, and (D) exons, respectively. The x-axis represents the corrected pairwise divergence estimated by maximum likelihood, while the y-axis is the observed divergence (uncorrected P-distance). Transitions are shown in blue, while transversions are in green. A best-fit second degree polynomial was applied to describe the trend.
By using BB- genome species as the outgroup, we analyzed both the whole integrated and the above-mentioned partitioned data sets using NJ, MP and BI. The obtained trees displayed almost the same topology with sufficiently high bootstrap values or BI (Fig. 4 and Supplementary Fig. S1). The resolved topology appears the same as the consensus network, which is essentially a kind of super tree. Our results indicate that all the studied *Oryza* species could be clearly distinguished from one another and the alternative topologies that occurred in small trees could be rejected. The concatenation of sufficient number of unlinked nuclear loci yields the species tree with remarkable bootstrap support. The Asian cultivated rice *O. sativa* and its presumed wild progenitors, *O. nivara* and *O. rufipogon*, formed a monophyletic group. Meanwhile, the African cultivated rice, *O. glaberrima*, is most closely related to the African wild species *O. barthii*, these two species formed a monophyletic group with 100% bootstrap or 1.00 PP. *O. glumaepatula* grown in America was sister to the Asian *O. sativa/O. rufipogon*/*O. nivara* and the African *O. barthii/O. glaberrima* groups, with 100% bootstrap or 1.00 PP. *O. longistaminata* was next sister to *O. glumaepatula*. Apparently, Australian *O. meridionalis* was the basal lineage of AA-genome species with strong bootstrap supports.

### 3.5. Genome divergence

Besides morphological characters, molecular tools offer an alternative source of information with which to estimate the genome divergence as a whole. In this study, we sampled all 53 single-copy genes together with 16 intergenic genomic regions to determine overall extent of genome divergence between rice and AA-genome *Oryza* species by using the *japonica* genome sequence as a reference (Table 4). We first analyzed the combined data matrix comprising of gene and intergenic region sequences to estimate the genome divergence at the DNA level. It is apparent that each pairwise JC distance was in agreement with their phylogenetic positions in the topology (Fig. 4). The two subspecies of Asian cultivated rice, *japonica* and *indica*, had an overall genome divergence of 0.43%. We found that the two presumed wild ancestral species *O. rufipogon* (0.44%) and *O. nivara* (0.48%) exhibited almost close divergence with *japonica*. Genome divergence of *O. glaberrima* and *O. barthii* with *japonica* were 0.64% and 0.59%, respectively, and further estimation showed that African cultivated rice *O. glaberrima* was fairly close to its wild progenitor, *O. barthii*, with a genome divergence of 0.51%. Genome divergence between *O. glumaepatula*, *O. longistaminata* and *japonica* was estimated to be 0.85% and 2.16%, respectively. Among AA-genome species, *O. meridionalis* showed the largest divergence of 3.0% with *japonica*. By and large, the average JC distance of AA-genome species was 1.1%. Additionally, genome divergence between BB-genome species of *O. punctata* and *japonica* was also obtained, approximated to be 6.4%.

Levels of genome divergence were further analyzed and compared by using data sets of intergenic regions, exons, introns, and the entire gene fragments, respectively (Table 4). The obtained JC distances for most species were in agreement with their phylogenetic positions in the topology. For the AA-genome species, the average JC distances in intergenic regions were twice as large as genic regions, and the intron regions exhibited relatively rapid divergence with a mean of 1.4%, almost three times faster than genic regions (0.4%). We further calculated levels of divergence with synonymous and nonsynonymous substitutions within coding regions of the 53 genes by the method of Nei-Gojobori using the Jukes–Cantor correction. Although relatively higher levels were observed with synonymous than nonsynonymous substitutions, we failed to detect significant difference between pairwise datasets (*P* < 0.001, Fisher’s exact test).

### Table 3
Tajima’s relative rate test for concatenated sequences using BB as an outgroup with estimates of the ratio of substitution rate between pairwise lineages.

<table>
<thead>
<tr>
<th>Genome</th>
<th>ind</th>
<th>ruf</th>
<th>niv</th>
<th>gla</th>
<th>bar</th>
<th>glu</th>
<th>lon</th>
</tr>
</thead>
<tbody>
<tr>
<td>ind</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ruf</td>
<td>0.48</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>niv</td>
<td>0.63</td>
<td>0.53</td>
<td>0.96</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gla</td>
<td>0.95</td>
<td>0.91</td>
<td>1.38</td>
<td>0.91</td>
<td>0.75</td>
<td>1.57</td>
<td></td>
</tr>
<tr>
<td>bar</td>
<td>0.76</td>
<td>0.74</td>
<td>0.90</td>
<td>0.79</td>
<td>0.91</td>
<td>1.16</td>
<td>0.77</td>
</tr>
<tr>
<td>glu</td>
<td>0.66</td>
<td>0.63</td>
<td>0.74</td>
<td>0.66</td>
<td>0.74</td>
<td>1.41</td>
<td>0.65</td>
</tr>
</tbody>
</table>

* Corresponds to the .05 significance level.
also obtained, approximated to be 6.58%. Among the AA-genome species, the lineages were homogeneous (Table 3). Our data showed that, although substitution rates differed from one fragment to another, the variation was not significant for all the sequenced fragments along most of the lineages. Although a significant difference was not detected between pairwise datasets of synonymous vs nonsynonymous substitutions, values of ti/tv for 11 genes were below 1 (Table 3). Thus, we calculated the values of dS, dN and dN/dS for each gene. 

### 3.6. Timing the divergence of AA-genome species

To time the divergence events leading to the major extant AA-lineages, we applied a molecular-clock approach to estimate their relative and absolute divergence times. In this study, Tajima’s relative test was first used to determine whether substitution rates are homogeneous (Table 3). Our data showed that, although substitution rates differed from one fragment to another, the variation was not significant for all the sequenced fragments along most of the lineages. Although a significant difference was not detected between pairwise datasets of synonymous vs nonsynonymous substitutions, values of ti/tv for 11 genes were below 1 (Table 3). Thus, we calculated the values of dS, dN and dN/dS for each gene (data not shown) and discarded five genes with either higher dN/dS values above 1 or very low values, respectively, which supposedly evolve under positive or negative selective pressures. To avoid one-fragment bias due to the variability in nucleotide substitution rates, we calculated an average rate of combined dataset of coding regions. In addition, molecular-clock was calibrated to date divergence times in this study. By using the fossil-calibrated average substitution rate of 5.9 × 10⁻⁹ mutations per synonymous site per year determined by the coding regions of Adh1 and Adh2 genes in grasses (Gaut et al., 1996), we obtained average substitutions of 6.02 × 10⁻⁹ mutations per synonymous site per year. As illustrated in Fig. 5 and Supplementary Table S9, divergence times among all the AA-genome species and BB-genome species as well were estimated with synonymous substitutions of these coding regions. In the present study, we found that the BB-genome species O. punctata shared a common ancestor and split from AA-genome species ~9.11 mya. Of the studied AA-genome species, O. meridionalis foremost diverged from other AA-genome members ~2.93 mya, and O. longistaminata and the other remaining species afterward split from O. meridionalis ~2.42 mya. Then, O. glumaputula split off ~1.12 mya. Our estimation showed that, African cultivated rice O. glaberrima with its wild progenitor O. barthi shared a common ancestor with Asian counterpart cultivated rice O. japonica. Furthermore, the AA-genome species split off one another ~0.86 mya. Here, indica and japonica genomes as well as their two presumed wild ancestral species O. rufipogon and O. nivara shared a common ancestor ~0.72 mya, whereas African O. barthi and O. glaberrima were split one another ~0.41 mya.

### Table 4

<table>
<thead>
<tr>
<th>Genes</th>
<th>Coding regions</th>
<th>Non-coding regions</th>
<th>Intergenic regions</th>
<th>Genes + Intergenic regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Synonymous</td>
<td>Nonsynonymous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of sites</td>
<td>47612</td>
<td>5421</td>
<td>5421</td>
<td>31880</td>
</tr>
<tr>
<td>jap-ind</td>
<td>0.0042</td>
<td>0.0071</td>
<td>0.0019</td>
<td>0.0050</td>
</tr>
<tr>
<td>jap-ruf</td>
<td>0.0044</td>
<td>0.0073</td>
<td>0.0015</td>
<td>0.0097</td>
</tr>
<tr>
<td>jap-niv</td>
<td>0.0042</td>
<td>0.0091</td>
<td>0.0014</td>
<td>0.0049</td>
</tr>
<tr>
<td>jap-gla</td>
<td>0.0066</td>
<td>0.0104</td>
<td>0.0016</td>
<td>0.0079</td>
</tr>
<tr>
<td>jap-bar</td>
<td>0.0060</td>
<td>0.0099</td>
<td>0.0015</td>
<td>0.0070</td>
</tr>
<tr>
<td>jap-glu</td>
<td>0.0037</td>
<td>0.0139</td>
<td>0.0033</td>
<td>0.0098</td>
</tr>
<tr>
<td>jap-lon</td>
<td>0.0223</td>
<td>0.0300</td>
<td>0.0085</td>
<td>0.0274</td>
</tr>
<tr>
<td>jap-mer</td>
<td>0.0336</td>
<td>0.0363</td>
<td>0.0111</td>
<td>0.0401</td>
</tr>
<tr>
<td>jap-pun</td>
<td>0.0661</td>
<td>0.1680</td>
<td>0.0341</td>
<td>0.0706</td>
</tr>
</tbody>
</table>

* The genomic divergence was estimated by Jukes-cantor distances in MEGA.

### Table 5

<table>
<thead>
<tr>
<th>Genes</th>
<th>Genome divergence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Synonymous</td>
</tr>
<tr>
<td>No. of sites</td>
<td>47612</td>
</tr>
<tr>
<td>jap-ind</td>
<td>0.0031</td>
</tr>
<tr>
<td>jap-ruf</td>
<td>0.0033</td>
</tr>
<tr>
<td>jap-niv</td>
<td>0.0029</td>
</tr>
<tr>
<td>jap-gla</td>
<td>0.0034</td>
</tr>
<tr>
<td>jap-bar</td>
<td>0.0032</td>
</tr>
<tr>
<td>jap-glu</td>
<td>0.0060</td>
</tr>
<tr>
<td>jap-lon</td>
<td>0.0145</td>
</tr>
<tr>
<td>jap-mer</td>
<td>0.0166</td>
</tr>
<tr>
<td>jap-pun</td>
<td>0.0658</td>
</tr>
</tbody>
</table>

* The divergence at amino acid level was calculated with JTT model in MEGA.
4. Discussion

4.1. Fully resolved phylogeny of the AA-genome species

Although a robust phylogeny among diploid species of the genus Oryza has been reconstructed (Ge et al., 1999; Zou et al., 2008), evolutionary relationships among AA-genome species are still contentious. Previous efforts have often resulted in conflicting results with different datasets to infer their relationships (Morishima et al., 1992; Wang et al., 1992; Ishii et al., 1996; Aggarwal et al., 1999; Iwamoto et al., 1999; Ge et al., 1999; Lu et al., 2000; Cheng et al., 2002; Park et al., 2003; Ren et al., 2003; Zhu and Ge, 2005; Duan et al., 2007), which were largely limited by the nature of molecular markers, number of loci and their representation in the genome. In this study, the alignment of ~60 kb represents 69 nuclear loci from the twelve rice chromosomes and has thoroughly resolved the phylogeny of AA-genome species with sufficient confidence, regardless of the applied methodologies and datasets of different sources. This study not only considered phylogenetic breadth of marker genes with diversified functions, particularly including both ecological adaptation-relevant and development-related gene fragments, some of which comprise sequences of long introns in length, but also included the sampling of 16 intergenic regions which usually are assumed to neutralize evolve. The application of these loci was evaluated and their potential values were confirmed in phylogenetic reconstruction, showing that they are able to broadly represent functional diversity and informative sites of chromosome dimension throughout the entire genome.

Our study has fully resolved evolutionary relationships of the AA-genome species and confirmed that the group is a distinct and monophyletic lineage in the genus Oryza, as proposed by former studies (Wang et al., 1992; Aggarwal et al., 1999; Ge et al., 1999; Zhu and Ge, 2005; Zou et al., 2008), although they occupy a globally extensive geographic range than the other Oryza species. Of the AA-genome species, O. meridionalis and O. longistaminata were previously regarded as two distantly related species (Morishima et al., 1992; Wang et al., 1992; Ishii et al., 1996; Lu et al., 2000; Cheng et al., 2002; Park et al., 2003; Zhu and Ge, 2005), but basal lineage of the group still remained unsolved. Some workers insisted that the African O. longistaminata was the earliest divergent species in the O. sativa complex (Iwamoto et al., 1999; Cheng et al., 2002; Ren et al., 2003). On the contrary, others proposed that O. meridionalis located at the root of the group (Wang et al., 1992; Ge et al., 1999; Park et al., 2003; Zhu and Ge, 2005), indicating that the AA-genome might have originated in Australia where this species nowadays grows. The present phylogenomic analyses all placed O. meridionalis as the basal lineage, demonstrating that it was the earliest divergent lineage that might separate from others at the early stage of the AA-genome species radiation. The three African species failed to form a monophyletic group, implying that geographic ranges may not necessarily account for different divergence times in their evolutionary histories.

As revealed by many previous investigations (Oka, 1988; Wang et al., 1992; Morishima et al., 1992; Park et al., 2003; Zhu and Ge, 2005), we confirmed that O. barthii is the wild progenitor of O. glaberrima. This study thus further verified that two cultivated rice, O. sativa and O. glaberrima, were parallel evolved from their wild progenitors, O. nivara/O. rufipogon and O. barthii, respectively, suggesting that two domestication events independently occurred in different geographic continents in the genus. The phylogenetic position of O. glumaepatula has long been conflicting from one study to another and remains unresolved (Zhu and Ge, 2005; Duan et al., 2007). In this study, we convincingly showed that O. glumaepatula grown in America was sister to the Asian O. sativa/O. rufipogon/O. nivara and the African O. barthii/O. glaberrima groups.

4.2. Phylogenetic reconstruction of AA-genome species

Although the reconstruction of gene trees seems essential for inferring evolutionary relationships of species, they may differ in several ways from the species tree. There is a need for caution, as previous workers have pointed out that, individual genes are not sufficient to estimate the correct genome phylogeny and a number of factors may have led to the incongruence between single-gene phylogenies (Kopp and True, 2002; Rokas et al., 2003a,b). These differences are usually caused by the complex evolutionary history, called as non-bifurcating, which violate the assumptions of simple sequence substitution models. The latter may include concerted evolution (Gao and Innan, 2004) as well as asymmetric sequence evolution (Conant and Wagner, 2003). In the present study, different genes resulted in alternative phylogenies, which might be caused by variable sites and rates of evolution (Ishii et al., 2001). It is intriguing to clarify which explanatory sources may have led to the incongruence. There are many factors, e.g., analytical and biological, which may influence the observed tree topology (Wendel and Doyle, 1998; Rokas et al., 2003a,b). Compared with ambiguous phylogenetic relationships of AA-genome species with merely four nuclear genes (Zhu and Ge, 2005), we obtained the same phylogeny of these species in spite of model specifications, outgroups choices and analytical methodologies. Hence, there is no reason to suggest that our finding might have experienced systematic bias after data combination. In addition, stochastic error (Zhu and Ge, 2005) may be avoided, at least largely reduced in this study through genome-scale sampling of source data including unlinked nuclear loci with diversified biological functions as well as intergenic genomic regions. In this study, topological convergence of trees was observed with NJ, MP and BI methods, as formerly confirmed (Bouchenak-Khelladi et al., 2008). Of them, BI methods could reduce the model specification error to a certain extent. Our analyses thus clearly indicate that, let alone the to-be-investigated causes of incongruence between single-gene phylogenies, the effect of concatenating single genes into one incorporated large data is able to overcome the pervasive incongruence between single trees and reach an authentic species tree. Accordingly, even though individual genes examined supported alternative trees, the concatenated data exclusively supported a single tree. Perhaps surprisingly, none of the known or predicted factors to cause phylogenetic error could systematically account for the observed incongruence (Sanderson and Shaffer, 2002; Rokas et al., 2003b), suggesting that there may be no good predictor of the phylogenetic informativeness of genes (Rokas et al., 2003b).

The stability of the junction between outgroup and ingroup has been a long-standing problem in phylogenetic analyses. The former studies told that appropriate outgroup is a fundamental prerequisite when the discrepancy between two conflicting phylogenetic hypotheses deeply relies on the root position (Bergsten, 2005; Kennedy et al., 2005; Morrison, 2010). With BB-, CC- and EE-genome species separately served as outgroup, consensus network analyses of the same set of genes showed that, as more informative characters were added, the effect of adding conflicting characters becomes minimized. Hence, the addition of more informative characters can lead to a point where outgroup choice plays a minimal role, while a proper choice of outgroup can largely decrease the influence to the outcome of tree topology.

4.3. Genome divergence of AA-genome species

The genomic divergence between rice and closely related AA-genome Oryza species has long been an issue of wide interest but unresolved. On basis of a large sample of genome-wide sequences, we obtained the first insights into the average nucleotide and
amino acid divergences of the given group of species. Comparing with an overall genome divergence of 0.43% between the two subspecies of Asian cultivated rice, *O. japonica* and *indica*, the previously reported observation (0.96–1.4%) (Ma and Bennetzen, 2004) might be overestimated. The values of pairwise discrepancy at either DNA or amino acid level were consistent with their phylogenic performance in the topology, suggesting that estimates should be reasonable and approximately represent genome divergence among them. Moreover, when separately using data sets of intergenic regions, exons, introns, and the entire gene fragments, the obtained levels of genome divergence were fairly in agreement with their phylogenetic position in the topology for most species. Attributable to the nature of neutral evolution, it is not surprising that the intron regions diverged relatively faster than genic regions, and the average JC distances in intergenic regions were almost as twice as genic regions in this study. The latter estimate is different from previous study which reported that intergenic regions evolved 4-fold faster than genic regions (Ma and Bennetzen, 2004). However, sequence divergences independently calculated by non-coding regions and exons unexpectedly revealed that mutation rates in synonymous sites of exons were mostly higher than non-coding regions. This seems somewhat consistent with the observation in primates, indicating that genetic distances in most of genic regions were higher than those in genomic sequences and thus explained that some of the genes may be located in regions with higher mutation rates (Subramanian and Kumar, 2003). In addition, repetitive elements have been demonstrated to play an important role in rice genome evolution and were estimated to compose a large portion of the rice genome (Gao et al., 2004) and wild related species (Piegu et al., 2006). However, this study did not contain such repetitive elements in genetic distance calculations. Prior studies estimated from large-scale alignments of genomic sequences between human and chimpanzee suggested a slightly higher average rate in repetitive elements than nonrepetitive sequences (Chen and Li, 2001). Therefore, overall estimates of genetic distance between the studied species call for taking into account both homologous repetitive and nonrepetitive sequences. Nevertheless, the comparisons in this study have hitherto provided an initial knowledge concerning their genetic or genomic divergence among rice and closely related AA- genome species and have important implications for understanding their evolutionary histories and critical functional changes which separate cultivated rice from other close relatives.

### 4.4. Recent rapid speciation of AA- genome species

As a recently radiated lineage with an extensive pantropical distribution, when the AA- genome species diverged and dispersed throughout Asia, Africa, America and Australia has long been an unsolved question (Nayar, 1973; Chang, 1976; Second, 1991; Ma and Bennetzen, 2004; Zhu and Ge, 2005). Previous attempts have been made to time evolutionary events in the genus *Oryza*, but these estimates largely differed from one to another (Chang, 1976; Barbier et al., 1991; Second, 1991; Ma and Bennetzen, 2004; Zhu and Ge, 2005). Ma and Bennetzen (Ma and Bennetzen, 2004) compared a total of 37 kb of sequence of the genomes of *indica*, *japonica* and African cultivated species, *O. glaberrima*, and suggested that the *indica* and *japonica* shared a common ancestor ∼0.44 mya and *O. glaberrima* diverged from a common ancestor with *indica and japonica* ∼0.64 mya. By means of intron sequences of four nuclear genes, Zhu and Ge (Zhu and Ge, 2005) roughly estimated a divergence of 2.0 mya for the AA- genome species. Based on the same method, they reported the Asian cultivated rice, *O. sativa*, diverged from the African cultivated rice, *O. glaberrima*, about 0.7 mya, while two subspecies of *O. sativa* (ssp. *indica* and ssp. *japonica*) separated approximately 0.4 mya. In order to obtain deep insights into the origin and evolutionary history of the extant AA- genome species in a phylogenomic perspective, we applied the molecular-clock approach to estimate their relative and absolute divergence times by combining a large dataset. To avoid the variability in rates of nucleotide substitutions and the bias of fewer gene fragments (Zhu and Ge, 2005), we analyzed both the pairwise synonymous and nonsynonymous average estimates of sequence divergence for coding regions, and dated approximate divergence times for diverse evolutionary events among the AA- genome species with BB- genome species *O. punctata* as outgroup. Our results obtained an earlier estimate of ∼0.86 mya when African cultivated rice *O. glaberrima* with its wild progenitor *O. barthi* shared a common ancestor with Asian counterpart cultivated rice as well as wild progenitors (Ma and Bennetzen, 2004; Zhu and Ge, 2005). In addition, we dated the divergence of *indica* and *japonica* as well as their two presumed wild ancestral species *O. rufipogon* and *O. nivara* to ∼0.72 mya, which is slightly larger than previous estimates (Ma and Bennetzen, 2004; Zhu and Ge, 2005) and much longer before the recognized domestication time of about 10,000 years ago between these two subspecies (Oka, 1988).

The biogeographic implications regarding when and how recent rapid speciation occurred across the eight AA- genome species have long been pursued by evolutionary biologists (Chang, 1976; Second, 1991; Barbier et al., 1991; Zhu and Ge, 2005). Despite limitations to the use of clock based on sequence data, such as substitution rate heterogeneity among lineages (Seelanan et al., 1997), estimating divergence events for the AA- genome species has roughly provided a comprehensive timeframe for understanding their evolutionary temps and biogeographic histories. Based on a molecular-clock estimate in this study, the eight AA- genome species diverged from their closely related BB- genome species, *O. punctata*, approximately 9.11 mya during the Miocene epoch. Provided that 2.93 mya could conservatively be the divergence time for the AA- genome group, it seems likely that the episode of rapid diversification gave rise to almost the whole diversity of the AA- genome lineages and radiated in the Early-Pleistocene. It is noticeable that the event has led to the divergence between African cultivated rice *O. glaberrima* with its wild progenitor *O. barthi* and Asian counterpart cultivated rice *O. sativa* as well as wild progenitors occurred as recently as approximately 0.86 mya during the Holocene. The finding is consistent to frequent migration after human origins and early agriculture together with the domestication and cultivation of major crops throughout the world to suit their needs during the Quaternary period. Since the global continents have been in their modern-day positions in the Pleistocene, the current distribution of the extant AA- genome *Oryza* species across four continents (Chang, 1976; Vaughan et al., 2003) but their fairly recent diversification is almost impossible to be explained by the previously proposed vicariance rising from the fracture and drift of the Gondwana supercontinent (Chang, 1976). Long-distance transeoceanic dispersal events were suggested for this group (Zhu and Ge, 2005) and other recently diverged lineages of the genus *Gossypium* (Seelanan et al., 1997). Before extraordinary efforts to search for further evidence, it would be appealing to investigate whether such rapid diversification was also resulted from adaptive ecological radiation rather than long-distance dispersals.

Rapid evolutionary radiations have long been regarded as the most challenging issue to elucidate the poorly resolved phylogenies in numerous organisms including aphids, black flies, bees, birds, turtles, mammals, and higher plants (Rokas and Carroll, 2006; Qiu et al., 2006; Nishiara et al., 2007; Rasmussen and Kellis, 2007; Whitfield and Lockhart, 2007; Zou et al., 2008). Indeed, it is such rapid radiations that phylogenetic relationships appeared an irresolvable polytomy for many cases (Fisher et al., 2001; Takezaki et al., 2004; Dopazo and Dopazo, 2005). Through
combining a sufficient amount of molecular data and appropriate applications of phylogenetic methods, however, increasing studies have shown that many assumed polytomies could be better resolved into sequential bifurcations (Walsh et al., 1999; Enard and Pääbo, 2004; Takezaki et al., 2004; Wortley et al., 2005). The latest genome-wide analyses by means of adequate sequence data, for example, have properly resolved soft polytomies, and clarified evolutionary relationships and genetic divergence among human, chimpanzee and gorilla, as a result of the rapid radiations with an estimated divergence as recent as a few Mya (Chen and Li, 2001; Enard and Pääbo, 2004; Patterson et al., 2006). The AA-genome Oryza species apparently represent a typical example of closely spaced series of fairly recent speciation events in flowering plants. Thanks to the unprecedented opportunity provided by the completed rice genome (Goff et al., 2002; Yu et al., 2002; International Rice Genome Sequencing Project, 2005), this study has fully determined evolutionary relationships with confidence through the utilization of unlinked intergenic genomic sequences and sufficiently diversified functional genes representing different evolutionary behaviors which may be of importance to assist phylogeny inference. Our results typify that phylogenetic reconstruction is powerful to better understand rapid speciation and genome divergence, which largely relies on improved suitable analytical methodology and particularly increasing amount of genome-wide sampling to mitigate the incongruence and better solve the mosaics of conflicting genealogies.

5. Conclusion
The success in resolving the phylogeny of AA-genome Oryza species in this study has undoubtedly proven particular values of phylogenomics to reconstruct evolutionary histories and elucidate rapid recent speciation events. This study also highlights the power of phylogenomics to evaluate levels of genomic divergence and determine genetic and evolutionary discrepancy between cultivated rice and closely related relatives in the genus. The completion of the rice genomes and the availability of the BAC libraries of Oryza species representing all rice genome types (Ammiraju et al., 2006) have attracted wide interest to perform the whole genome de novo sequencing of the genus Oryza and AA-genome species in particular through massively parallel next-generation sequencing technologies. Without doubt, a reliable resolution of phylogenetic relationships and preliminary understanding of genome divergence of AA-genome species has built a significant foundation for the future studies on evolutionary, comparative and functional genomics. Such a phylogenetic framework holds tremendous potential to accelerate studies of rice genome diversity, adaptation, evolution and domestication and, in return, will largely benefit genetic improvement of cultivated rice in search of beneficial agronomic genes from gene pool of wild AA-genome species.

Acknowledgments
We thank the International Rice Research Institute (Los Banos, Philippines) for providing rice materials. We are grateful for BR Holland who kindly offered the Python script. This work was supported by Talents Program of Yunnan Province (2008A0409), Hundreds Oversea Talents Program of Yunnan Province, a grant from Natural Science Foundation of Yunnan (2010CC011), the Hundreds Talents Program of the Chinese Academy of Sciences (CAS), a grant from the Chinese Academy of Science (KSCX2-YW-N-029), a grant from Chinese Department of Science and Technology (973 Program 2007CBB15703), and a startup grant from Kunming Institute of Botany, CAS to L.-Z. GAO.

Appendix A. Supplementary material
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2013.10.008.

References
Curr. Opin. Plant Biol. 6, 139–146.
Walsh, H., Kidd, M., Moum, T., Friesen, V., 1999. Polytomies and the power of 
Wang, Z., Second, G., Tanksley, S., 1992. Polymorphism and phylogenetic 
relationships among species in the genus Oryza as determined by analysis of 
history and molecular evolution. In: Soltis, D.E., Soltis, P.S., Doyle, J.J. (Eds.), 
265–296.
Wortley, A.H., Rudall, P.J., Harris, D.J., Scotland, R.W., 2005. How much data are 
needed to resolve a difficult phylogeny? Case study in Lamiales. Syst. Biol. 54, 
697–709.
biology and evolution. J. Hered. 92, 371–373
2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 
296, 79–92,
Zhang, S., Gu, Y.Q., Singh, J., Coleman-Derr, D., Brar, D.S., Jiang, N., Lemaux, P.G., 
2007. New insights into Oryza genome evolution: high gene colinearity and 
differential retrotransposon amplification. Plant Mol. Biol. 64, 589–600.
Zhu, Q., Ge, S., 2005. Phylogenetic relationships among A-genome species of the 
genus Oryza revealed by intron sequences of four nuclear genes. New Phytol. 
167, 249–265.
Analysis of 142 genes resolves the rapid diversification of the rice genus. 
Genome Biol. 9, R49.