Theor Appl Genet (2011) 123:21–31
DOI 10.1007/s00122-011-1563-2

ORIGINAL PAPER

Genetic diversity and domestication history of African rice (*Oryza glaberrima*) as inferred from multiple gene sequences

Zhi-Ming Li · Xiao-Ming Zheng · Song Ge

Received: 29 April 2010 / Accepted: 26 February 2011 / Published online: 12 March 2011

Abstract Nucleotide variation in 14 unlinked nuclear genes was investigated in species-wide samples of African rice (*Oryza glaberrima*) and its wild progenitor (*O. barthii*). Average estimates of nucleotide diversity were extremely low in both species (θ_{sil} = 0.0007 for *O. glaberrima*; θ_{sil} = 0.0024 for *O. barthii*). About 70% less diversity was found in *O. glaberrima* than in its progenitor *O. barthii*. Coalescent simulation indicated that such dramatic reduction of nucleotide diversity in African rice could be explained mainly by a severe bottleneck during its domestication. The progenitor of African rice maintained also low genetic diversity, which may be attributed to small effective population size in *O. barthii*. Self-pollinating would be another factor leading to the unusually low diversity in both species. Genealogical analyses showed that all *O. glaberrima* accessions formed a strongly supported cluster with seven *O. barthii* individuals that were sampled exclusively from the proposed domestication centers of African rice. Population structure and principal component analyses found that the *O. glaberrima* group was homogeneous with no obvious genetic subdivision, in contrast to the heterogeneous *O. barthii* cluster. These findings support a single domestication origin of African rice in areas of the Upper Niger and Sahelian Rivers.

Communicated by G. Bryan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1563-2) contains supplementary material, which is available to authorized users.

Z.-M. Li · X.-M. Zheng · S. Ge (✉)
State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Xiangshan, Beijing 100093, China
e-mail: gesong@ibcas.ac.cn

Z.-M. Li · X.-M. Zheng · S. Ge
Graduate University, Chinese Academy of Sciences, Beijing 100039, China

Introduction

In the rice genus, two cultivated species, Asian rice (*Oryza sativa* L.) and African rice (*O. glaberrima* Steud.), were domesticated independently in Asia and Africa (Chang 1976; Second 1982). While Asian rice currently spreads worldwide as the world's largest food crop, African rice is grown primarily in tropical west Africa. With its distinct origin, African rice differs from its Asian counterpart in many qualitative and quantitative traits (Sarla and Swamy 2005; Vaughan et al. 2008). More importantly, African rice varieties have many unique and useful traits such as weed competitiveness, tolerance to various abiotic stresses (acidity, salinity and drought) and resistance to diseases/pests (Ghesquiere et al. 1997; Linares 2002; Sarla and Swamy 2005). These features have been combined with high yield of Asian rice by rice breeders to develop a series of popular hybrid varieties known as NERICA (New Rice for Africa) that are high-yielding, drought- and pest-resistant and adapted to the growing conditions of west Africa (Jones et al. 1997; Sarla and Swamy 2005).

For cultivated plants, a general consensus is that varying degrees of reduction of genetic diversity in crops relative to their wild progenitors occurred during the process of domestication because of dual bottlenecks imposed by domestication and breeding (Buckler et al. 2001; Tenaillon et al. 2004; Zeder et al. 2006). Although extensive analyses of reproductive barriers between Asian and African rice were made (Sarla and Swamy 2005), much less has been

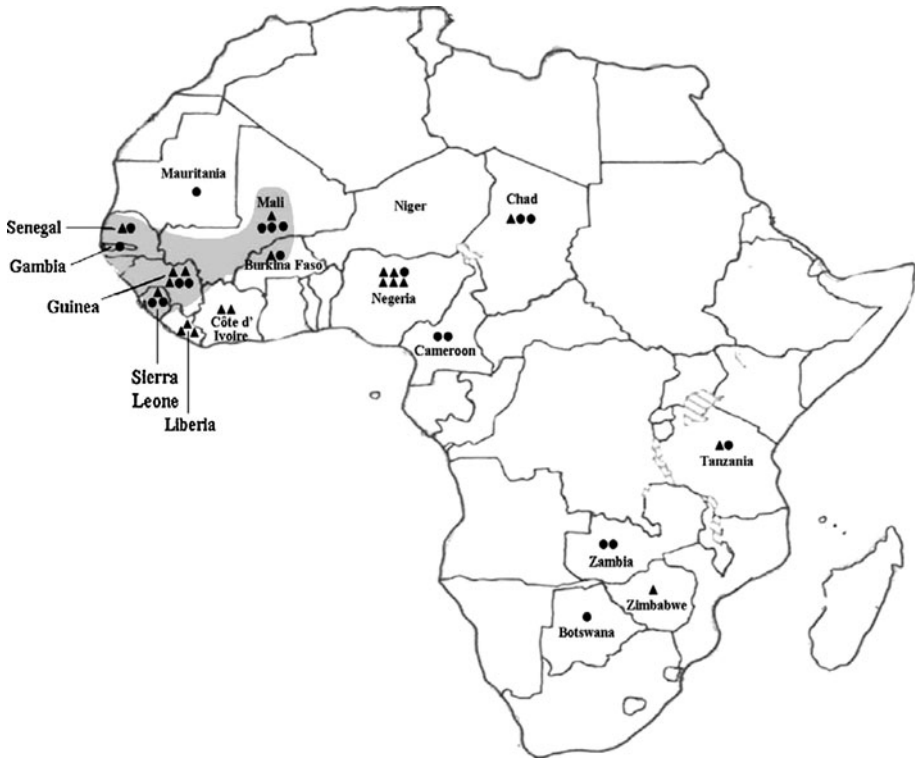

Fig. 1 Geographical origin of 20 *O. glaberrima* accessions and 20 *O. barthii* individuals sampled in the present study. The *shaded area* represents three domestication centers proposed by Portères (1970). *Triangles* and *circles* indicate the *O. glaberrima* accessions and *O. barthii* individuals sampled in this study, respectively

conducted on the population genetics and domestication history of African rice, in contrast to Asian rice that has received extensive studies using a variety of approaches (see reviews in Sweeney and McCouch 2007; Sang and Ge 2007; Vaughan et al. 2008). It is generally accepted that African rice originated from the Niger River Delta about 3,500 years ago, with *O. barthii* as its progenitor (Linares 2002; Sarla and Swamy 2005). Based on isozyme, RFLP, SSR and MITE markers, previous studies have found dramatic reduction of genetic diversity associated with the domestication of African rice and revealed substantially lower genetic diversity in African than in Asian rice (Second 1982; Wang et al. 1992; Ishii et al. 2001). However, these studies obtained inconsistent results regarding the level of genetic diversity present in African rice and the extent of genetic loss during its domestication, either because of different marker systems used or because of limited numbers of accessions/populations sampled. In particular, population genetics of *O. barthii*, the immediate ancestor of African rice, has not been the focus of research, although the value of wild species for improving crop yield has been widely recognized.

Another fundamental question regarding the domestication of a given crop plant is the number of domestication events and their geographic origin (Zeder et al. 2006; Burger et al. 2008). The domestication of African rice occurred much later than that of Asian rice, but certainly prior to the introduction of Asian rice to Africa (Sarla and Swamy 2005; Sweeney and McCouch 2007). Earlier suggestions of an Asian origin of African rice and its introduction into Africa have been refuted, but two hypotheses have been proposed with regard to the domestication centers of *O. glaberrima*. Portères (1970) suggested that African rice was first domesticated in the Inland Delta of the Upper Niger River and then spread to two secondary centers along the Sahelian Rivers (the Niger, Senegal, Gambia and their tributaries) (Sarla and Swamy 2005) (Fig. 1). An alternative hypothesis proposed that African rice was first selected by ancient hunting–gathering human populations at several different localities within the vast forest and savanna areas (Richards 1996). These hypotheses are mainly based on fragmented linguistic and archeobotanical evidence and have never been tested using genetic data.

Recent years have witnessed successful utilization of multilocus DNA sequences in addressing population genetics and evolutionary history of crop species and their wild relatives (Tenaillon et al. 2004; Wright and Gaut 2005; Liu and Burke 2006; Haudry et al. 2007; Zhu et al. 2007; Burger et al. 2008). In this study, we provide the first analysis of the levels and patterns of nucleotide variation in African rice (*O. glaberrima*) and its wild ancestor (*O. barthii*) based on the sequences of 14 unlinked nuclear loci. We first assessed species-wide level of nucleotide diversity of African rice and its progenitor. We were interested in assessing the relative levels of genetic diversity of African rice compared to that of Asian rice, which have been extensively investigated based on multilocus sequences (Tang et al. 2006; Caicedo et al. 2007; Zhu et al. 2007). Then, we assess how much

reduction of nucleotide variation occurs in African rice relative to its wild progenitor, and explore the severity of domestication bottleneck in African rice. Finally, we address the origin and domestication center of African rice, which has not been attempted using molecular population genetic data despite two hypotheses (Sarla and Swamy 2005). Given the fact that African rice is an important staple food for people across west Africa and carries numerous agronomically important alleles (Jones et al. 1997; Murray 2004; Sarla and Swamy 2005), knowledge of population genetics and domestication history of African rice is of great importance for the effective conservation and utilization of the cultivated and wild germplasm and to facilitate the genetic improvement and breeding of African rice.

Materials and methods

Plant materials

A total of 20 accessions of African rice (*O. glaberrima*) and 20 individuals of African wild rice (*O. barthii*) (Table 1) were used in this study after verification by growing them in greenhouse. Their geographic origin is shown in Fig. 1. The cultivated *O. glaberrima* accessions were collected from 11 countries and included all three cryptic subpopulations (Semon et al. 2005). The wild individuals were sampled to cover the entire geographical distribution of *O. barthii* and can be divided into two areas on the basis of geographical origin: ten from the putative domestication centers (Portères 1976; Linares 2002) and ten from areas outside the domestication centers (Fig. 1). One accession of Asian wild rice (*O. nivara*) was used as an outgroup, because this species was closely related to African rice (Wang et al. 1992; Zhu and Ge 2005). Seeds of these samples were kindly provided by the International Rice Genebank at the International Rice Research Institute (Manila, Philippines). To exclude the accessions/individuals that might originate from potential introgression/hybridization between the cultivated and wild species, we grew all samples in the greenhouse of the Institute of Botany (CAS). We identified three cultivated accessions and one wild individual that showed intermediate morphology and these samples were excluded from this study. The 40 samples used in this study were verified to be typical African rice and *O. barthii* (Table 1).

DNA extraction, amplification and sequencing

Total genomic DNA was extracted from fresh young leaves or gel-dried leaves, using the hexadecyltrimethylammonium

Table 1 Plant materials used in this study

Taxon	Accession[a]	Origin	Code[b]
Oryza glaberrima	86790	Liberia	gla_LIB2 (I)
	96775	Nigeria	gla_NIG2 (II)
	96837	Guinea	gla_GUI1 (I)
	100136	Guinea	gla_GUI2 (I)
	100854	Guinea Bissau	gla_GUI3 (I)
	100977	Zimbabwe	gla_ZIM (I)
	102193	Mali	gla_MAL (I)
	102210	Liberia	gla_LIB1 (III)
	102289	Liberia	gla_LIB3 (III)
	103344	Senegal	gla_SEN
	103474	Burkina Faso	gla_BUR
	103477	Côte d' Ivoire	gla_COT (I)
	103989	Sierra Leone	gla_SIE (I)
	104018	Tanzania	gla_TAN
	104034	Ivory Coast	gla_IVO
	104042	Chad	gla_CHA
	104525	Nigeria	gla_NIG3 (II)
	104527	Nigeria	gla_NIG4 (III)
	104558	Nigeria	gla_NIG1 (II)
	112564	Nigeria	gla_NIG5 (III)
Oryza barthii	86447	Guinea	bar_GUI2
	86525	Cameroon	bar_CAM2
	100122	Gambia	bar_GAM1
	100927	Sierra leone	bar_SIE2
	100933	Mali	bar_MAL2
	101252	Burkina Faso	bar_BUR
	101937	Senegal	bar_SEN
	103912	Tanzania	bar_TAN
	104078	Nigeria	bar_NIG
	104120	Chad	bar_CHA1
	104132	Cameroon	bar_CAM1
	104287	Mali	bar_MAL3
	105612	Zambia	bar_ ZAM1
	105613	Botswana	bar_ BOT
	106194	Guinea	bar_GUI1
	106205	Mali	bar_MAL1
	106234	Sierra Leone	bar_SIE1
	106291	Mauritania	bar_ MAU
	106295	Chad	bar_CHA2
	106489	Zambia	bar_ZAM2
Oryza nivara	103407	Sri Lanka	Niv_LKA1

[a] All accessions were obtained from seeds provided by the Genetic Resources Center of the International Rice Research Institute at Los Banos, Philippines

[b] Individual accession is abbreviated by the first three letters of the taxon name followed by the code of its origin of country. Roman numbers in parentheses are the subpopulation identity defined by Semon et al. (2005)

Table 2 Summary of genes surveyed and primer sequences used in the study

Locus	Chromosome location	Alignment length (bp)		Primer sequence (5′–3′)	Functional association
		Total	Noncoding		
Adh1	11	853	853	Adh1F5: TCCCGTGTTCCCTCGGATCTTC Adh1R3: GTCACACCCTCTCCA ACACTCT	Alcohol dehydrogenase I
CatA	2	601	601	CatAU: TCGTTATCATCAGCGTGGAC CatAL: CATGCCACTAACTAGCTCAG	Catalase gene
Cbp1	12	816	506	Cbp5F1: CATTTTACATAGCTGGGGAGTC Cbp9R1: CTGGGGTCACTTTCTTGATAG	Serine carboxypeptidase I
GBSSII	7	622	483	Gbs4F1: GTGACAGGACAGAAACAGTG Gbs7R1: TCCAGTGTGCCAGTCATTTG	Granule-bound starch synthase II
GPA1	5	1271	733	Gpa15F: GTGCTTTATGCAAGAGTACGG Gpa15R: CACAATGGCTAACTAGAGAAC	G protein α subunit
Ks1	4	880	460	Ks1F3: TTTCCTGGTATGCTTAGCCTTGC Ks1R3: GCGAAAAGCCATTGCACATGTTG	*ent*-kaurene synthase I
Lhs1	3	1056	895	Lhs1F3: AGAGGTACCGCAGCTGCAAC Lhs1R: CTCCTTCATGCTDAGTGGGC	MADS-box transcription factor 1
Os0053	2	736	587	Ganjin1F: AGAGAAAGCGGTGTCAGACG GanjinR: CACGTTCTCCTTGGCGAGCG	Putative aldehyde decarbonylase
NP39	9	489	239	NP39F: CGGCTATCCAATCCTTGAAG NP39R: AGTTCTGCCTCTAGTTCGTC	Copia-like retroelement pol polyprotein
NP122	8	664	390	NP122F: GTGGATTCTGGAGGGAACGC NP122R: GCATTGTGGGCAGCCATATT	Rubredoxin-type Fe(Cys)4 protein family protein
Os1977	3	592	582	Wanderer356U: GTGTTGTTGCTGCATATTTG Wanderer356L: CCTACGGTCATTCCAAAACC	Porin, eukaryotic type family protein
SSII1	10	917	844	OsS3F: AAGAGCAAAGAGGTGTTGGAG OsS3R: CATCGTCTGGAACTTCATCAAG	Soluble starch synthase II-1
TFIIAc-1	1	968	893	Seq1: AGGTGAAGAGCAAGGTTA P4: TCRCAGGCCACRATCTTCAC	Gamma subunit of transcription factor II A
Waxy	6	535	358	Waxy12F: GCTTCGAGCCCTGTGGACTC Waxy12R1: CACGTTCTCCTTGGCGAGCG	Granule-bound starch synthase
Total		11000	8424		

bromide method as previously described in Ge et al. (1999). The genes used in this study represent 14 unlinked nuclear loci across all 12 chromosomes in *O. sativa*. Ten of these genes had been previously used in another study on the nucleotide diversity of Asian cultivated and wild rice (Zhu et al. 2007). Four additional genes, *GPA1*, *NP39*, *NP122* and *Os1977*, were chosen to better represent all the rice chromosomes. Detailed information about the genomic location and putative gene functions, as well as the primer sequences of the amplified regions, is listed in Table 2 and the schematic diagrams of the 14 genes are provided in Fig. S1 (See Supplemental Material). Polymerase chain reaction (PCR) amplification was performed in a total volume of 25 µl using T-personal thermocycler (Biometra, Germany). The reaction mix contained 10–50 ng of genomic DNA, 0.2 µM of each primer, 200 µM of each deoxyribonucleotide triphosphate, 10 mM Tris–Cl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, and a mixture of 0.75 U Ex *Taq* DNA polymerase (TaKaRa, Japan). PCR reaction was carried out in a T-gradient 96 U themocycler (Biometra, Gottingen, German) as described in Zhu et al. (2007). To reduce recombinant molecule during PCR, we used long extension time during PCR reactions, including 3-min extension during each amplification cycle. Amplified products were separated by electrophoresis on 2.0% agarose gels stained with ethidium bromide using a 100-bp DNA ladder and gel purified with either a Pharmacia purification kit (Amersham Pharmacia Biotech, Piscataway, NJ, USA) or a Dingguo purification kit (Dingguo, Beijing, China). Sequencing was performed on ABI 3730 automated sequencer (Applied Biosystems, Foster City, CA, USA).

Because *O. glaberrima* and *O. barthii* are predominantly self-pollinating species (Vaughan 1994), sampling individuals was essentially equivalent to sampling gametes and the

purified PCR products could be sequenced directly. In one case where haplotypes could not be readily inferred because of heterozygosity (bar_BOT, accession number: 105613), PCR products were cloned into *pGEM* T-easy vectors (Promega, Madison, WI, USA) and six clones were sequenced. When singletons (polymorphisms that occurred in only one sequence relative to all the remainder sequences) were found, repeated PCR amplification and sequencing were conducted to exclude those singletons resulting from *Taq* polymerase errors, as described in Zhang and Ge (2007). All sequences have been deposited in the EMBL/GenBank Data Libraries under accession numbers. FJ810879–FJ811452.

Sequence analysis and neutrality test

Sequence data were assembled with the ContigExpress program (Informax Inc., North Bethesda, MD, USA). Sequences were aligned using ClustalX version 1.81 (Thompson et al. 1997) with additional manual refinements. Indels (insertions/deletions) were not included in the analysis because most of the software programs, such as DnaSP, do not deal with them. For each locus and taxon, we calculated the number of segregating sites (S) and two parameters of nucleotide diversity: π, the expected heterozygosity per nucleotide site (Nei 1987), and θ_w, an estimate of $4Ne\mu$, where Ne is the effective population size and μ is the mutation rate per nucleotide (Watterson 1975). Estimates of nucleotide diversity were based on total sequences and silent sites separately using the DnaSP version 4.10.8 (Rozas et al. 2003).

Several tests based on either the frequency spectrum of polymorphisms or the haplotype distribution (Tajima 1989; Fu and Li 1993) and on the relationship between intraspecific and interspecific diversity (Hudson et al. 1987) were performed to detect deviations from the neutral equilibrium model of evolution using the programs DnaSP and HKA (http://lifesci.rutgers.edu/~heylab/). Tajima's D (Tajima 1989) was based on the discrepancy between the mean pairwise difference (π) and Watterson's estimator (θ_w), while D^* and F^* of Fu and Li (Fu and Li 1993) relied on the difference between the number of polymorphic sites in external branches (polymorphisms unique to an extant sequence) and number of polymorphic site in internal phylogenetic branches (polymorphisms shared by extant sequences). These tests were calculated at all sites and at silent sites, respectively. The Hudson–Kreitman–Aguade (HKA) test (Hudson et al. 1987) was used to test for heterogeneity in the ratio of polymorphism to divergence among loci. This multilocus test can discriminate between selection force and population demography because selective force influences particular regions of the genome during speciation process. *Oryza nivara* sequences were used as outgroups for the HKA and Fu and Li's tests.

Coalescent simulations

To estimate the intensity of the bottlenecks associated with domestication of African rice, we conducted a coalescent simulation to model the impact of a bottleneck on sequence diversity using Hudson's ms program (Hudson 2002). A simple model of reduction in effective population size was used, assuming that an ancestral population experienced an instantaneous change in effective population size t generations ago. This model was previously described in domesticated wheat, in which the bottleneck intensity α was defined as the ratio of the ancestral population size (N_a) to cultivated population size (N_p) (Haudry et al. 2007). Higher value of α indicates more severe bottleneck during domestication. We used the *O. barthii* data to calibrate the simulation parameters for the ancestral population and *O. glaberrima* data as the observed data for cultivated species because African rice originated from *O. barthii* about 3,500 years ago (Linares 2002; Sarla and Swamy 2005). In addition to N_a and N_p, we had three more parameters for each locus: (1) $\tau = t/2\ N_p$, the time since the bottleneck, where $t = 3{,}500$ is the time when domestication started; (2) θ_{wild}, estimated by *O. barthii* data; (3) $4Nc$, the recombination parameter estimated based on *O. barthii* data using Hudson's LDhat program (http://www.stats.ox.ac.uk/~mcvean/LDhat/LDhat1.0).

We explored 20 values of α on a grid ranging from 1 (no reduction in effective population size) to 10.5 (severe reduction in effective population size), with 10,000 simulations for each locus. We estimated the bottleneck intensity by combined multilocus likelihood $L(\alpha)$ over loci. The intensity of the bottleneck was calculated to be the value of α that maximizes $L(\alpha)$ (Haudry et al. 2007).

Analyses of genealogy and population structure

The genealogical trees were constructed using the NJ (neighbor-joining), MP (maximum parsimony) and ML (maximum likelihood) methods. NJ and MP analyses were implemented in PAUP* version 4.0b10 (Swofford 2001), with NJ performed with Kimura's 2-parameter distances (Kimura 1980) and MP conducted using heuristic searches (tree bisection–reconnection branch swapping; MulTrees option in effect) with ten random additions of taxa. Topological confidence was assessed by bootstrap analysis with 1,000 replicates. ML analysis was conducted in PHYML program version 3.0 (Guindon and Gascuel 2003). The parameters used for ML tree construction were GTR model of nucleotide substitution, BIONJ as starting tree, non-parametric bootstrap analysis with 100 replicates, estimated proportion of invariable sites and estimated gamma shape parameter.

To further identify geographic division in genetic diversity of the species and origin of the cultivated accessions,

we conducted a Bayesian clustering analysis using the program Structure v. 2.2 with the "LinkageModel" (Falush et al. 2003). This analysis used multilocus genotypes to infer the fraction of an accession's genetic ancestry that belongs to a population, under a given number of populations (K). Since *O. glaberrima* and *O. barthii* are predominantly selfing, we treated the data as haploid and thus groupings were based on allele frequencies. The posterior probabilities were estimated using a Markov chain Monte Carlo method (MCMC). We used a genetic assignment algorithm to the data with $K = 2$–6 clusters, without using prior information on geographic origin of the samples. For comparison, we also performed a principal component analysis (PCA) using the R PACKAGE of P. Legendre. PCA analysis makes no assumptions about population structure and provides a suited method for description of the principal axes of genetic variation between populations (Patterson et al. 2006).

Results

Nucleotide diversity and neutrality test

Sequences of 14 unlinked loci were obtained from 40 accessions, representing African rice and its wild ancestor. The length of aligned sequence for each locus varied between 489 and 1,271 bp, with a total of 11,000 bp in length, including 8,424 bp of noncoding sequence (Table 2). A total of 29 indel polymorphisms, ranging from 1 to 5 bp, were found exclusively in introns and flanking regions and were excluded from further analyses. A total of 84 SNPs (an average of 1 SNP every 131 nucleotides) were found across all cultivated and wild accessions, with the number of SNP being 19 in *O. glaberrima* and 76 in *O. barthii*.

Standard statistics of sequence variation for each locus are summarized in Table 3, including the estimates of nucleotide diversity in different regions at individual loci. Different levels of nucleotide variation were found across loci, with no polymorphism detected at four loci (*Cbp1*, *GPA1*, *NP122* and *Waxy*) for *O. glaberrima* ($\pi_{sil} = 0$, $\theta_{sil} = 0$), and the highest level of polymorphism found at *CatA* for *O. barthii* ($\pi_{sil} = 0.0082$, $\theta_{sil} = 0.0066$). At the species level, the average estimate of polymorphism over 14 loci is substantially lower in the cultivated species ($\pi_{sil} = 0.0006$, $\theta_{sil} = 0.0007$) than in the wild species ($\pi_{sil} = 0.0025$, $\theta_{sil} = 0.0024$). It is obvious from Table 3 that nucleotide diversity in the wild species was about four times that of the cultivated rice at both entire region and silent sites, with much lower diversity found in the cultivated rice than its wild progenitor at almost all loci.

Neutrality tests including Tajima's D (Tajima 1989) and D^* and F^* of Fu and Li (1993) indicate no significant value at any locus (Table 3). The multilocus HKA test (Hudson et al. 1987) was further conducted to examine whether levels of polymorphism and divergence across loci would be correlated, as expected under the neutral model of molecular evolution. Using one sequence of *O. nivara* as an outgroup, we applied the HKA test across loci to each of the two taxa. The multilocus test was not significant for the wild species (*O. barthii*/*O. nivara*, $x^2 = 10.75$, $P = 0.632$), but rejected the null hypothesis of proportionality between polymorphism and divergence in *O. glaberrima*/*O. nivara* contrast ($x^2 = 25.26$, $P = 0.021$). We compared the contribution from each locus to the overall test statistics to ask whether or not the observed values of polymorphism and divergence were higher or lower than the expected. Result showed that locus *NP39* contributed greatly to the significant multilocus HKA statistics for cultivated rice contrast (*O. glaberrima*/*O. nivara*) (Fig. S2, See Supplemental Material). When this locus was excluded, the multilocus HKA test detected no rejection of null hypothesis (*O. glaberrima*/*O. nivara*, $x^2 = 14.42$, $P = 0.275$). Note that the nucleotide diversity of African rice ($\pi_{sil} = 0.0019$; $\theta_{sil} = 0.0029$) is substantially higher than that of *O. barthii* ($\pi_{sil} = 0.0004$; $\theta_{sil} = 0.0010$) at *NP39*, suggesting that selection might influence the *NP39* locus to some extent because balancing selection would elevate diversity (Wright and Gaut 2005).

Because the low level of nucleotide polymorphism in African rice could be explained by either genetic bottleneck or selection, we performed coalescent simulation and compared the simulation results with our observed data to assess the impact of a bottleneck on sequence diversity. Most loci produced similar likelihood peaks for the summary statistics except for *NP39* (Fig. S3, See Supplemental Material). This is consistent with the HKA test, suggesting that *NP39* might not evolve neutrally or might be subjected to strong selection pressures in African rice. Therefore, we excluded this locus from our simulation. The maximum likelihood estimate of the bottleneck intensity α for the domestication of African rice was 3.5 (Fig. S3, See Supplemental Material), indicating that substantial reduction of nucleotide polymorphism in African rice because of domestication bottleneck.

Genealogical analyses and population structure

Single gene trees were basically not powerful enough to obtain resolved trees because of low polymorphisms. The combined analyses using NJ, MP and ML methods obtained essentially the same topology except for slightly different statistical supports for some clades. As demonstrated in the ML tree (Fig. 2), all *O. glaberrima* accessions formed a monophyletic group with statistical supports over 93% for all three methods (named as the *O. glaberrima* or

Table 3 Summary of nucleotide polymorphisms and neutrality tests at 14 loci

Taxon	Locus	*S*	π_T	θ_T	π_{sil}	θ_{sil}	*D*	*D**	*F**	R_M
Oryza glaberrima	*Adh1*	1	0.0001	0.0003	0.0001	0.0003	−1.1644	−1.5396	−0.8790	0
	CatA	2	0.0011	0.0009	0.0011	0.0009	0.2822	0.8662	0.8131	0
	Cbp1	0	0.0000	0.0000	0.0000	0.0000	/	/	/	/
	*GBSS*II	1	0.0002	0.0005	0.0002	0.0006	−1.1644	−1.5396	−1.6477	0
	GPA1	0	0.0000	0.0000	0.0000	0.0000	/	/	/	/
	Ks1	1	0.0005	0.0003	0.0007	0.0005	0.7226	0.6495	0.7652	0
	Lhs1	3	0.0012	0.0008	0.0013	0.0009	1.1858	1.0065	1.2139	0
	Os0053	1	0.0001	0.0004	0.0002	0.0005	−1.1644	−1.5396	−1.6477	0
	NP39	3	0.0011	0.0017	0.0019	0.0029	−0.9752	−1.2550	−1.3552	0
	NP122	0	0.0000	0.0000	0.0000	0.0000	/	/	/	/
	Os1977	3	0.0021	0.0014	0.0021	0.0015	1.1526	1.0065	1.2034	0
	SSII1	1	0.0001	0.0003	0.0001	0.0003	−1.1643	−1.5396	−1.6477	0
	TFIIAc-1	3	0.0008	0.0009	0.0008	0.0009	−0.2604	−0.1243	−0.1857	1
	Waxy	0	0.0000	0.0000	0.0000	0.0000	/	/	/	/
	Average		0.0005	0.0005	0.0006	0.0007				0.07
Oryza barthii	*Adh1*	12	0.0063	0.0041	0.0063	0.0041	1.9569	1.0507	1.5225	1
	CatA	14	0.0082	0.0066	0.0082	0.0066	0.9190	0.7784	0.9509	1
	Cbp1	3	0.0011	0.0010	0.0015	0.0015	0.0887	−0.1243	−0.0759	0
	*GBSS*II	2	0.0013	0.0009	0.0016	0.0011	1.0483	0.8662	1.0502	0
	GPA1	4	0.0012	0.0009	0.0017	0.0010	1.0241	0.1745	0.4719	0
	Ks1	3	0.0007	0.0010	0.0009	0.0010	−0.7924	−1.2550	−1.2976	0
	Lhs1	6	0.0015	0.0016	0.0018	0.0018	−0.1322	−0.1542	−0.1709	0
	Os0053	12	0.0042	0.0046	0.0049	0.0054	−0.3333	−0.5907	−0.5987	2
	NP39	2	0.0004	0.0012	0.0004	0.0010	−1.5128	−2.0531	−2.1885	0
	NP122	2	0.0004	0.0009	0.0006	0.0012	−1.1407	−0.5935	−0.8503	0
	Os1977	7	0.0029	0.0034	0.0029	0.0034	−0.4638	−1.2035	−1.1490	0
	SSII1	1	0.0002	0.0003	0.0002	0.0003	−0.5916	0.6495	0.3673	0
	TFIIAc-1	3	0.0012	0.0009	0.0013	0.0009	1.0196	1.0065	1.1616	1
	Waxy	5	0.0018	0.0026	0.0024	0.0036	−1.0373	−1.2127	−1.3425	0
	Average		0.0022	0.0021	0.0025	0.0024				0.36

S number of segregating sites

π_T: average number of nucleotide differences per site between two sequences (Nei 1987) calculated on the total number of polymorphic sites

π_{sil}: average number of pairwise nucleotide differences per site calculated on the silent sites

θ_T: the Watterson estimator of θ per base pair (Watterson 1975) calculated on the total number of polymorphic sites

θ_{sil}: Watterson's estimator of θ per base pair calculated on the silent sites

D: Tajima's *D* (Tajima 1989)

D/F**: Fu & Li's *D**, Fu & Li's *F** (Fu and Li 1993)

R_M: estimates of the minimum number of recombination events (Hudson and Kaplan 1985)

African rice clade hereafter). Within this clade, no clear subdivision was detected, although 20 cultivated accessions represented three subpopulations of *O. glaberrima* as defined by Semon et al. (2005). Interestingly, seven *O. barthii* samples mixed with the *O. glaberrima* accessions that were sampled exclusively from domestication centers, with three from flood plain area (bar_MAL2, bar_MAL3, bar_BUR) and two from each of Mangrove brackish land (bar_SEN, bar_GAM1) and forest highland (bar_GUI1, bar_GUI2). These areas were previously considered the domestication centers of African rice (Portères 1976; Murray 2004). The remaining 13 accessions of *O. barthii* occurred in the branches outside the *O. glaberrima* clade and formed three well-supported subclades, implying subdivision of population structure of the wild species (Fig. 2).

Based on haplotype segments from the 14 loci, we conducted a STRUCTURE analysis to assign each cultivated accession and wild individual to a cluster of origin without

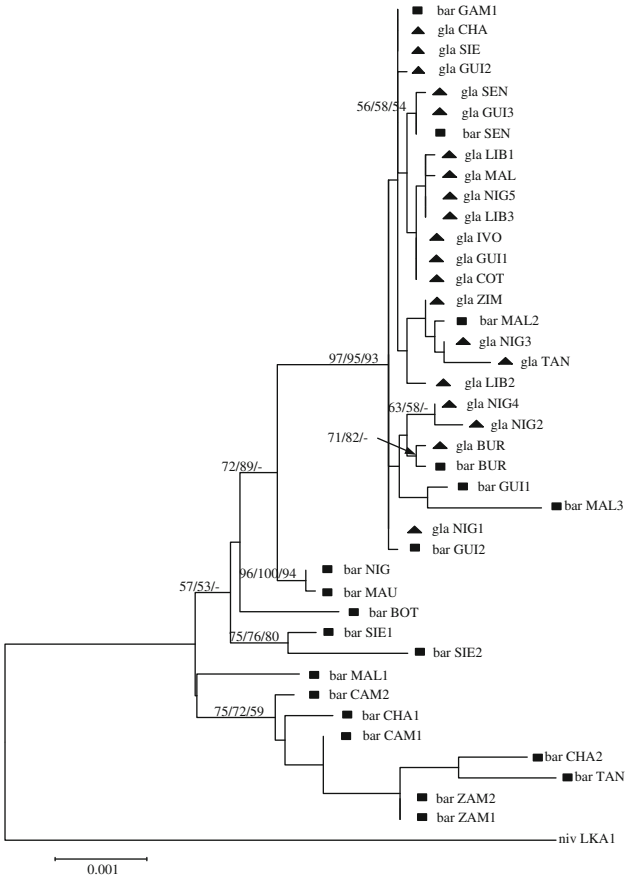

Fig. 2 Maximum likelihood phylogentic tree constructed with concatenated sequences of 14 nuclear loci. The cultivated accessions are indicted by *triangles* and the wild individuals by *squares*. Three numbers near branches are bootstrap percentages of ML and MP analyses as well as Bayesian posterior probabilities

prior information regarding the geographic origin of individual samples. Figure 3 shows the results by successively increasing K from 2 to 6, with the highest likelihood value at $K = 3$. To be conservative, we first examined the population structure at $K = 2$ and found a cultivar-wild gradient, with a clear difference between the wild individuals outside domestication centers and the cultivated accessions. At $K = 3$, where all *O. glaberrima* accessions are homogeneous, *O. barthii* individuals segregate into three groups. The first group consisting of six individuals has genetic affinity to *O. glaberrima* samples; the second group comprises seven individuals from both within and outside domestication centers; the third group includes seven samples from areas outside domestication centers. The three clusters remain unchanged when K increases from 3 to 6, suggestive of no further genetic population stratification of the sample under study. It should be noted, however, that *O. glaberrima* group remains homogeneous with no apparent genetic subdivision. In contrast, the *O. barthii* samples are obviously heterogeneous (Fig. 3), paralleling the observation in the genealogical analyses. Similarly, principal component analysis (PCA) did not detect apparent subdivision for the *O. glaberrima* group, although a clear separation is observed between the cultivated and wild accessions (Fig. 4). In contrast to homogeneity of *O. glaberrima* accessions, the wild populations could be divided clearly by two eigenvectors into subgroups, and one subgroup consisting of six individuals from the domestication center cluster with *O. glaberrima* samples (Fig. 4). The PCA analysis parallels the results of genealogical and STRUCTURE analyses.

Discussion

Extremely low level of genetic diversity in African rice and its progenitor

Population genetic studies on plant species have been increasingly conducted using multiple gene sequences and demonstrated a wide range of nucleotide diversity among species. As reviewed by Haudry et al. (2007) and Zhang and Ge (2007), about 20-fold difference in nucleotide diversity has been reported across species, with the π_{sil} values from 0.0012 (*Triticum turgidum* ssp. *dicoccum*, Haudry et al. 2007) to 0.0230 (*A. lyrata* ssp. *lyrata*, Ramos-Onsins et al. 2004). The present study showed that both cultivated and

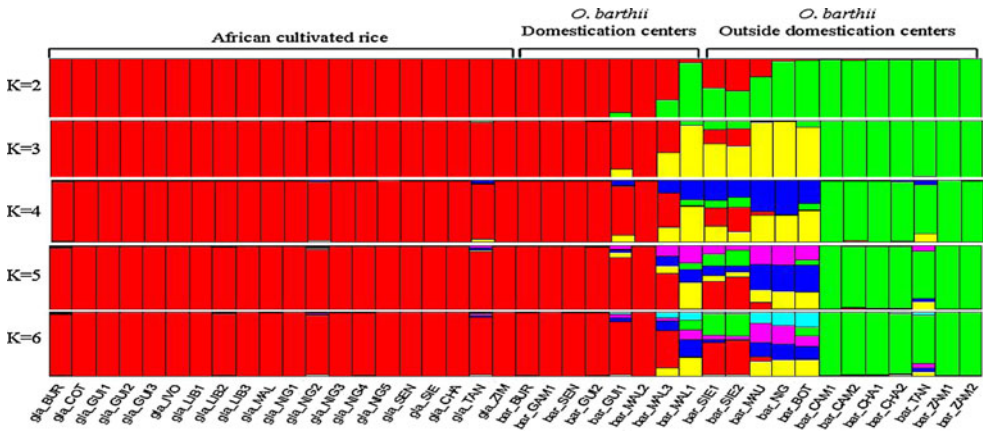

Fig. 3 Ancestries of 40 *O. glaberrima* and *O. barthii* samples inferred with STRUCTURE (Pritchard et al., 2000) at $K = 2$ to 6. Each sample is represented by a column partitioned into *colored segments.* the lengths of which correspond to its ancestry coefficients-inferred ancestral groups

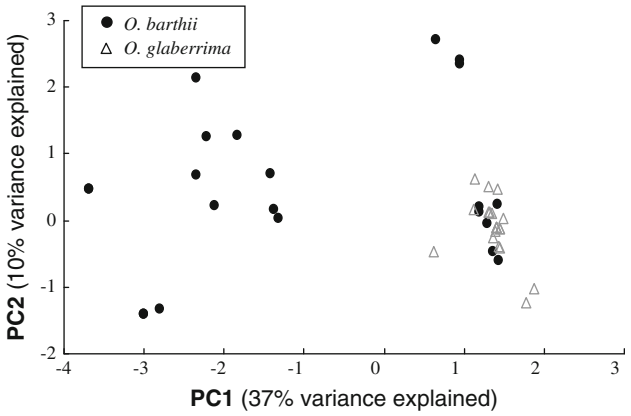

Fig. 4 Principal component analysis (PCA) of 40 *O. glaberrima* and *O. barthii* samples based on sequences of 14 nuclear loci. The first eigenvector (PC1) explained 37% of variation and the second (PC2) explained 10% of variation

wild African rice maintained extremely low levels of nucleotide diversity ($\pi_{sil} = 0.0006$ for *O. glaberrima*, $\pi_{sil} = 0.0025$ for *O. barthii*). This is noteworthy because the lowest nucleotide diversity found so far for cultivated plants is $\pi_{sil} = 0.0012$ in wheat (Haudry et al. 2007) and for wild species is $\pi_{sil} = 0.0028$ in wild soybean (Hyten et al. 2006). In addition, recent studies using sequences of multiple genes found that six other *Oryza* species have intermediate levels of nucleotide diversity, with π_{sil} values ranging from 0.0032 (*O. sativa*) to 0.0072 (*O. rufipogon*) (Tang et al. 2006; Caicedo et al. 2007; Zhang and Ge 2007; Zhu et al. 2007). Note that African rice has only one-sixth of the nucleotide diversity found in Asian rice and even the indica rice (*O. sativa* ssp. *indica*) has higher diversity ($\pi_{sil} = 0.0028$–0.0029, Tang et al. 2006; Zhu et al. 2007) than either African rice or *O. barthii*. A low level of genetic variation in the cultivated and wild African rice was previously reported using molecular markers, such as isozyme, RFLPs and SSRs (Second 1982; Wang et al. 1992; Ishii et al. 2001; Semon et al. 2005), but comparison among species was difficult due to the different marker systems and sampling strategies used in these studies.

It is widely appreciated that cultivated plants experience varying degrees of reduction of genetic diversity relative to their wild progenitors Such a loss of genetic diversity for major crops was previously estimated to be of the order 30–40% (Buckler et al. 2001) and appears to vary from 31 to 70% based on a recent report (Haudry et al. 2007). Dramatic reduction (over 50% loss) of nucleotide diversity has been demonstrated by recent studies on major cereal crops such as barley (*H. vulgare* ssp. *vulgare*, Kilian et al. 2006), Asian rice (*O. sativa*, Zhu et al. 2007) and wheat (*Triticum turgidum* ssp. *dicoccum*, Haudry et al. 2007). In comparison, genetic loss in African rice is much more severe after domestication, with 76% less diversity in the domesticated species than its wild progenitor, as measured by π_{sil}. An obvious explanation for the low genetic diversity of *O. glaberrima* would be a genetic bottleneck during its domestication from a small initial population of *O. barthii*. Based on our coalescent simulations, the extremely low nucleotide diversity in African rice can be explained by severe bottleneck during domestication, with high value of the bottleneck intensity ($\alpha = 3.5$) (Fig. S3, See Supplemental Material), which is consistent with its single origin in west Africa (see discussion below). It should be noted, nevertheless, that a domestication-selective sweep due to strong selection cannot be excluded entirely as a factor causing low nucleotide diversity in African rice, due to low statistical power in loci with little polymorphism.

The unusually low diversity in *O. glaberrima* might also be attributed to the low genetic diversity in *O. barthii*, the immediate ancestor of African rice, because *O. barthii* possessed the lowest nucleotide diversity for a wild plant species reported so far (see reviews in Haudry et al. 2007 and Zhang and Ge 2007). The low level of diversity in *O. barthii* may be explained by factors such as demography, mating system and selection. Selection seems unlikely at the loci studied here, because statistical tests did not find significant deviation from neutrality for almost all loci. It is possible that *O. barthii* maintained a small historic effective population size after its divergence from other *Oryza* species (Zhu and Ge 2005), because other wild *Oryza* species that have been investigated by multilocus sequences have apparently higher nucleotide diversity ($\pi_{sil} = 0.0037$ for *O. rhizomatis*; 0.0044 for *O. officinalis*; 0.0052 for *O. eichingeri*; 0.0063 for *O. nivara*; 0.0072 for *O. rufipogon*) (Tang et al. 2006; Caicedo et al. 2007; Zhang and Ge 2007; Zhu et al. 2007). Self-pollination would be another factor leading to low diversity of *O. barthii* because it reduces effective population size and effective recombination rates and thus decreases diversity (Charlesworth 2003). However, empirical studies indicate that the mating system mainly influences diversity at the population rather than species levels (Savolainen et al. 2000; Wright et al. 2003) and its effect needs investigation based on samples collected from disparate populations.

Alternately, as pointed out by previous studies (Chang 1976; Vaughan et al. 2008), African rice might experience a double evolutionary bottleneck, i.e., the first was associated with the divergence of its ancestor (*O. barthii*) from Asian *Oryza* species in history and the second happened during the domestication of *O. glaberrima* from the initial populations of *O. barthii* in Africa.

Domestication history and population structure of African rice

It is generally accepted that *O. glaberrima* was domesticated about 3,500 years ago, with the annual *O. barthii* as

its progenitor (Portères 1976; Linares 2002; Murray 2004; Sarla and Swamy 2005; Vaughan et al. 2008). However, the geographic origin of *O. glaberrima* remains largely elusive, despite two distinct hypotheses (Linares 2002). Portères (1970) hypothesized that African rice was first domesticated in the Inland Delta of the Upper Niger River, in what is today Mali, and subsequently spread to two secondary centers along Sahelian Rivers: one in the coast of Gambia and Guinea Bissau, and the other in the Guinea forest between Sierra Leone and the western Ivory Coast (Sarla and Swamy 2005) (Fig. 1). An alternative hypothesis proposed multiple origins of *O. glaberrima* at several different localities within the extensive forest and savanna areas (Richards 1996; Linares 2002).

Although phylogenetic tree based on multiple loci should be interpreted with caution because of recombination between loci and lineage sorting of closely related species, tree topologies provide a general picture of accession/individual relationship. The present genealogical analysis shows that all *O. glaberrima* accessions along with seven *O. barthii* individuals formed a cluster with strong statistical support (Fig. 2). In addition, all the seven *O. barthii* individuals in this cluster were collected from the domestication centers defined by Portères (1970), in which three individuals were sampled from the primary center (flood plain area in the Inland Delta of the Upper Niger River) and four from two secondary centers (Mangrove brackish land in the coast of Gambia and Guinea Bissau as well as the forest highland between Sierra Leone and the Western Ivory Coast). Such a pattern of variation is also evident in STRUCTURE and PCA analyses, consistent with the single origin of African rice (Portères 1976; Sarla and Swamy 2005). However, the precise place where *O. glaberrima* originated needs further investigation with extensive sampling.

A previous study revealed two major ecotypes in *O. glaberrima*: a floating photosensitive type grown in deep water and coastal mangrove lands, and an early erect ecotype cultivated in upland and rainfed lowlands (Ghesquiere et al. 1997). Using 93 nuclear microsatellite markers, Semon et al. (2005) studied a collection of 198 accessions of *O. glaberrima* and revealed three *O. glaberrima* subpopulations in addition to two admixed subpopulations between African and Asian rice based on the STRUCTURE analyses. They claimed that the three subpopulations were associated with ecological differentiation of floating and non-floating, and lowland and upland rice. It should be noted, however, that Semon et al. (2005) did not detect clear subdivision for *O. glaberrima* by both genealogical and PCA analyses, implying low level of genetic diversity and very weak if any genetic structure for the cultivated African rice. In our case, genealogical, STRUCTURE and PCA analyses all failed to detect subdivisions of *O. glaberrima* samples. In contrast, we detected significant subdivisions within the wild *O. barthii* (Fig. 4). These observations not only suggest the higher genetic diversity in *O. barthii* relative to *O. glaberrima*, but also imply very weak if any genetic structure for the cultivated African rice. Nevertheless, a better understanding of population genetic structure for both cultivated and wild African rice, as well as their underlying mechanism, needs further investigations utilizing extensively samples of the cultivated *O. glaberrima* accessions and the natural populations of the wild *O. barthii* with a larger number of molecular markers.

Acknowledgments We thank Qi-Hui Zhu, Lin-Bin Zhang, Xin-Hui Zou and other members of Ge's group for technical assistances. We are grateful to Jie Li (Xishuangbanna Tropical Botanical Garden, CAS) and the two anonymous reviewers for their valuable comments and suggestions. We also thank the International Rice Research Institute (Los Banos, Philippines) for providing leaf and seed samples. This work was supported by the National Basic Research Program of China (2007CB815704), National Natural Science Foundation of China (30990240, 30121003) and the grants from the Chinese Academy of Sciences.

References

Buckler ES, Thornsberry JM, Kresovich S (2001) Molecular diversity, structure and domestication of grasses. Genet Res 77:213–218

Burger JC, Chapman MA, Burke JM (2008) Molecular insights into the evolution of crop plants. Am J Bot 95:113–122

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL et al (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3:1745–1756

Chang TT (1976) The origin, evolution, cultivation, dissemination, and diversification of Asian and African rices. Euphytica 25:425–441

Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. Phil Trans R Soc Lond B 358:1051–1070

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure: extensions to linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Ge S, Oliveira GCX, Schaal BA, Gao LZ, Hong DY (1999) RAPD variation within and between natural populations of wild rice *Oryza rufipogon* from China and Brazil. Heredity 82:638–644

Ghesquiere A, Sequier J, Second G, Lorieux M (1997) First steps towards a rational use of African rice, *Oryza glaberrima*, in rice breeding through a 'contig line' concept. Euphytica 96:31–39

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C et al (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol 24:1506–1517

Hudson RR (2002) Generating samples under a Wright–Fisher nuetral model of genetic variation. Bioinformatics 18:337–338

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159

Hyten DL, Song QJ, Zhu YL, Choi IY, Nelson RL, Costa JM et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103:16666–16671

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

Jones MP, Dingkuhn M, Aluko GK, Semon M (1997) Inter-specific *O. sativa* L. *O. glaberrima* Steud: progenies in upland rice improvement. Euphytica 92:237–246

Kilian B, Ozkan H, Kohl J, von Haeseler A, Barale F, Deusch O et al (2006) Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. Mol Genet Genomics 276:230–241

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–134

Linares OF (2002) African rice (*Oryza glaberrima*): History and future potential. Proc Natl Acad Sci USA 99:16360–16365

Liu AZ, Burke JM (2006) Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics 173:321–330

Murray SS (2004) Searching for the origins of African rice domestication. Antiquity 78. (http://antiquity.ac.uk/projgall/murray/index.html)

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Patterson N, Price A, Reich D (2006) Population structure and eigenanlysis. PLoS Genet 2:e190

Portères R (ed) (1970) Primary cradles of agriculture in the African Continent. Cambridge University Press, Cambridge

Portères R (1976) The origin of African plant domestication. Mouton, The Hague

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguade M (2004) Multilocus analysis of variation and speciation in the closely related species *Arabidopsis helleri* and *A. lyrata*. Genetics 166:373–388

Richards P (1996) Redefining nature: ecology, culture and domestication. Berg, Oxford

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Sang T, Ge S (2007) Genetics and phylogenetics of rice domestication. Curr Opin Genet Dev 17:533–538

Sarla N, Swamy BPM (2005) *Oryza glaberrima*: a source for the improvement of *Oryza sativa*. Curr Sci 89:955–963

Savolainen O, Langley CH, Lazzaro BP, Freville H (2000) Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol Biol Evol 17:645–655

Second G (1982) Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. Jpn J Genet 57:25–57

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

Sweeney M, McCouch S (2007) The complex history of the domestication of rice. Ann Bot 100:951–957

Swofford D (2001) *PAUP: phylogenetic analysis using parsimony analysis (and other methods)*. Version 4.0b10. Sinauer Associated, Sunderland

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tang T, Lu J, Huang JZ, He JH, McCouch SR, Shen Y et al (2006) Genomic variation in rice: genesis of highly polymorphic linkage blocks during domestication. PLoS Genet 2:1824–1833

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol 21:1214–1225

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The Clustal X windons interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Vaughan DA (1994) The wild relatives of rice: a genetic resources handbook. International Rice Research Institute, Los Banos

Vaughan DA, Lu BR, Tomooka N (2008) The evolving story of rice evolution. Plant Sci 174:394–408

Wang ZY, Second G, Tanksley SD (1992) Polymorphism and phylogenetic relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. Theor Appl Genet 83:565–581

Watterson GA (1975) On the number of segregating sites in genetical models without recombinations. Theor Popul Biol 7:256–276

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519

Wright SI, Lauga B, Charlesworth D (2003) Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol Ecol 12:1247–1263

Zeder MA, Emshwiller E, Smith BD, Bradley DG (2006) Documenting domestication: the intersection of genetics and archaeology. Trends Genet 22:139–155

Zhang LB, Ge S (2007) Multilocus analysis of nucleotide variation and speciation in *Oryza officinalis* and its close relatives. Mol Biol Evol 24:769–783

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

Zhu QH, Zheng XM, Luo JC, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888