Theor Appl Genet (2007) 114:559–568
DOI 10.1007/s00122-006-0457-1

ORIGINAL PAPER

Genetic diversity analysis of traditional and improved Indonesian rice (*Oryza sativa* L.) germplasm using microsatellite markers

Michael J. Thomson · Endang M. Septiningsih · Fatimah Suwardjo · Tri J. Santoso · Tiur S. Silitonga · Susan R. McCouch

Received: 23 March 2006 / Accepted: 4 November 2006 / Published online: 29 November 2006

Abstract The archipelago of Indonesia has a long history of rice production across a broad range of rice-growing environments resulting in a diverse array of local Indonesian rice varieties. Although some have been incorporated into modern breeding programs, the vast majority of these landraces remain untapped. To better understand this rich source of genetic diversity we have characterized 330 rice accessions, including 246 Indonesian landraces and 63 Indonesian improved cultivars, using 30 fluorescently-labeled microsatellite markers. The landraces were selected across 21 provinces and include representatives of the classical subpopulations of *cere*, *bulu*, and *gundil* rices. A total of 394 alleles were detected at the 30 simple sequence repeat loci, with an average number of 13 alleles per locus across all accessions, and an average polymorphism information content value of 0.66. Genetic diversity analysis characterized the Indonesian landraces as 68% *indica* and 32% *tropical japonica*, with an *indica* gene diversity of 0.53 and a *tropical japonica* gene diversity of 0.56, and a F_{st} of 0.38 between the two groups. All of the improved varieties sampled were *indica*, and had an average gene diversity of 0.46. A set of high quality Indonesian varieties, including Rojolele, formed a separate cluster within the *tropical japonicas*. This germplasm presents a valuable source of diversity for future breeding and association mapping efforts.

Communicated by A. Paterson.

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-006-0457-1 and is accessible for authorized users.

M. J. Thomson · E. M. Septiningsih · F. Suwardjo ·
T. J. Santoso · T. S. Silitonga
Indonesian Center for Agricultural Biotechnology
and Genetic Resources Research and Development,
Jl. Tentara Pelajar 3A, Bogor 16111, Indonesia

S. R. McCouch (✉)
Department of Plant Breeding and Genetics,
Cornell University, 162 Emerson Hall,
Ithaca, NY 14853, USA
e-mail: srm4@cornell.edu

Present Address:
M. J. Thomson · E. M. Septiningsih
International Rice Research Institute,
Los Banos, Laguna, Philippines

Introduction

Knowledge of the genetic diversity and population structure of germplasm collections is an important foundation for crop improvement. Due to the importance of rice as a major world crop, the origin and diversity of *Oryza sativa* has attracted great interest. Current hypotheses point to a polyphyletic origin of cultivated rice, a result of at least two independent domestication events, likely from either perennial or annual types of *O. rufipogon* (Cheng et al. 2003; Second 1982; Vitte et al. 2004).Two sub-species of *O. sativa*, *indica* and *japonica*, have long been recognized and recent evidence suggests an ancient *indica*/*japonica* divide dating between 200,000 and 440,000 years ago based on nuclear genome sequence comparisons and between 86,000 and 200,000 years ago based on chloroplast sequences, which significantly pre-dates the domestication of rice (Ma and Bennetzen 2004; Tang et al. 2004; Vitte et al. 2004; Zhu and Ge 2005). This deep population structure is readily apparent in the rice landraces and improved varieties grown around the world. Representative samples of the global rice gene pool have been extensively studied using molecular markers, beginning with Glaszmann's pioneering study that identified six groups within *O. sativa* from 1,688 rice accessions using isozyme markers (Glaszmann 1987). A more recent analysis using microsatellite markers identified five major groups from a diverse sample of 234 rice accessions: *indica*, aus, *tropical japonica*, *temperate japonica* and aromatic (Garris et al. 2005).

While the overall population structure of global *O. sativa* germplasm has been well characterized, more detailed analyses of rice germplasm on a regional or country-specific basis have only just begun. One such example is Indonesia—a country with a wealth of biodiversity that is largely untapped. A long history of traditional rice production across numerous environments in Indonesia has led to a diverse array of local Indonesian rice varieties. Indonesian rice varieties have previously been classified into three traditional categories: *cere* (or *tjereh*), *bulu* and *gundil*. *Cere* rice has been characterized as having thin stems with many tillers, narrow leaves, and no awns, while *bulu* rices have thick stems with few tillers, broad leaves, and long awns, and *gundil* types are essentially the same as *bulu* but lacking awns (Takahashi 1997). Rice breeders have associated *cere* with *indica*, while *bulu* and *gundil* were once referred to as "*javanica*" (i.e. rice that was found on the island of Java, Indonesia) but the term *tropical japonica* is now preferred (Khush 1997). Several Indonesian varieties have been employed in international breeding programs, such as the variety Peta that was used as one of the parents of the green revolution variety IR8 (IRRI 1967). More recently, the breeding program at the International Rice Research Institute has made use of Indonesian *tropical japonica* landraces to develop new plant type (NPT) varieties that have fewer tillers, thick stems, and large panicles for more efficient grain production (Peng et al. 1999). Although a few varieties have been incorporated into modern breeding programs, the vast majority of traditional Indonesian germplasm remains uncharacterized and underutilized.

New strategies have recently been developed to make better use of plant germplasm collections for crop improvement, such as using advanced backcross QTL populations and introgression lines to identify and transfer beneficial alleles from exotic germplasm (Li et al. 2005; Tanksley and McCouch 1997). Another method for allele mining is to use association mapping to directly identify useful alleles in germplasm collections using linkage disequilibrium (LD) (Flint-Garcia et al. 2003). A two-tiered approach has been proposed for association mapping in the human genome in which an isolated population is used for an initial whole-genome scan, followed by high resolution mapping with more diverse populations where the haplotype blocks are smaller (Gabriel et al. 2002). Following this strategy, populations representing different subsets of Indonesian landraces may provide a valuable resource for future association mapping studies in rice. Before performing an association mapping study, it is essential to first define the population structure within the germplasm to avoid spurious associations (Flint-Garcia et al. 2005). As a first step toward defining the genetic diversity and population structure of Indonesian germplasm we have characterized 330 rice accessions, including 246 geographically-diverse Indonesian landraces and 63 Indonesian improved cultivars, using capillary electrophoresis with 30 fluorescently-labeled microsatellite markers. Previous studies have demonstrated the advantages of using fluorescently-labeled microsatellite markers for genetic diversity analysis in rice with multiple dyes allowing efficient genotyping using multiplex panels of markers (Blair et al. 2002; Coburn et al. 2002; Jain et al. 2004). The use of capillary electrophoresis provides additional benefits, including greater automation, ease of use, and precision in fragment sizing, which makes it the preferred technology for simple sequence repeat (SSR) genotyping in rice (Garris et al. 2005; Lu et al. 2004). The purpose of this study is to analyze the genetic diversity and population structure of Indonesian germplasm in comparison to global rice varieties, to compare the diversity and structure of Indonesian landraces with the improved varieties, and to evaluate the potential usefulness of Indonesian germplasm for association mapping.

Materials and methods

Plant material

The 330 rice accessions consisted of 309 Indonesian varieties (246 Indonesian landraces and 63 Indonesian improved cultivars), 18 international varieties, and three accessions of *O. rufipogon* (see Supplementary Table 1). The Indonesian landraces were selected from the rice germplasm collection at the Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (Bogor, Indonesia) to represent the broad geographic range of rice cultivation across 21 provinces of Indonesia (Fig. 1). Most of the landraces originated from one of four major Indonesian islands: 73 landraces were from Sumatra, 49 were from Kalimantan (on the island of Borneo), 49 were from Sulawesi, and 48 were from the island of Java (Fig. 1). The Indonesian improved varieties were requested from plant breeders from the Indonesian Institute of Rice Research (Sukamandi, Indonesia) and from the rice germplasm collection in Bogor to compare the genetic structure of Indonesian improved varieties with the traditional landraces. Eighteen international varieties were included in the study, most of which were selected from accessions previously clustered into population groups to serve as controls for the genetic diversity analysis, in addition to two previously studied Indonesian varieties, Trembese and Popot-165 (see Garris et al. 2005). These consisted of three *temperate japonica* (Nipponbare, Koshihikari, and Caloro), three *tropical japonica* (Moroberekan, Sinampaga Selection, and Maintmolotsy 1226), four *indica* (IR36, IR64, Tetep and Binulawan), three aus (Dhala Shaitta, DV85, and Kasalath), one aromatic (JC1), and one not previously defined (Cabacu). In addition, three improved varieties in the Indonesian breeding program were confirmed to be introductions from other countries, and have been labeled as international varieties: Kartuna (Philippines), Kalimutu (Kenya), and Gajah Mungkur (Kenya). Three *O. rufipogon* accessions originally from Indonesia were requested from the International Rice Research Institute (IRRI, Los Banos, Philippines) to compare the genetic relationship of wild Indonesian accessions to the landraces and improved varieties (IRGC 81802, 92605, and 105958).

SSR marker genotyping

Plants were grown in pots in a greenhouse in Bogor, Indonesia, and young leaf tissue was harvested from a single plant for each accession. Although there may be heterogeneity within landraces, a single plant was chosen to be able to detect heterozygosity within an individual, which would have been indistinguishable from heterogeneity within a landrace if a bulked sample had been used. Total genomic DNA was extracted after crushing in liquid nitrogen in microfuge tubes using a Tris/SDS extraction buffer (100 mM Tris–HCl pH 8, 50 mM EDTA pH 8, 500 mM NaCl, 1.25% SDS (w/v), 0.38 g sodium bisulfite per 100 ml of buffer) and chloroform extraction followed by ethanol precipitation. Agarose gel electrophoresis was used to estimate DNA concentration, and each sample was then diluted to approximately 5–10 ng/μl.

Thirty evenly-spaced SSR markers were selected as a subset of markers previously used for genetic diversity analysis of *O. sativa*, with a preference for those SSRs that followed a stepwise mutation pattern (Table 1; Garris et al. 2005). Multiplex panels were designed to run five markers in each panel based on previously published allele size ranges (Temnykh et al. 2000). For each SSR marker, the forward primer was labeled with one of three fluorescent labels (D2, D3, and D4) for running on the Beckman CEQ 8000 using WellRED dyes (Proligo, Boulder, CO, USA; IDT, Coralville, IA, USA). PCR reactions contained 0.75 U of FastStart Taq (Roche Applied Science, Indianapolis, IN, USA) with the following components in a 20 μl reaction volume: FastStart 10× PCR buffer with 2 mM $MgCl_2$, 10 mM dNTP mix, 5 μM forward and reverse primers, and 15–30 ng genomic DNA. Reactions were run for 35 cycles using a touchdown PCR program (touchdown annealing at 60–55°C). Each SSR marker was run in a separate PCR reaction and amplification products were pooled before loading. All six panels were optimized by adjusting the dilution of each of the five amplification products to equalize signal strength for each panel. Fragments were size separated by capillary electrophoresis using a Beckman CEQ 8000 Genetic Analyzer in the lab in Bogor, Indonesia. Allele sizes were obtained using Beckman's Fragment Analysis software, followed by a manual binning step for markers

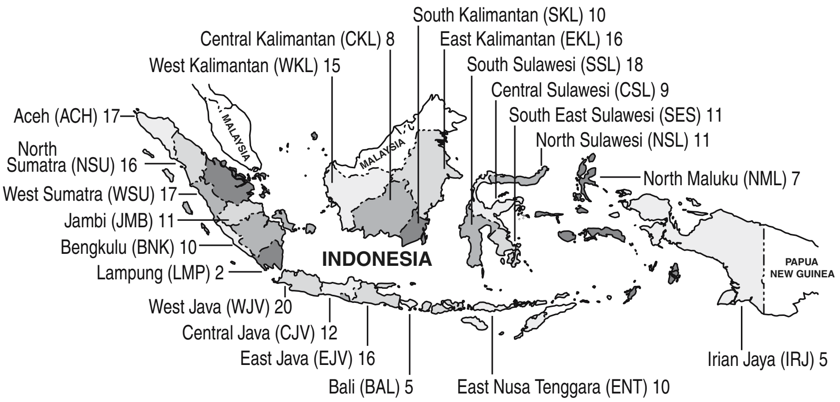

Fig. 1 Map of Indonesia, showing the 21 provinces represented by the 246 Indonesian landraces, an abbreviation of each province, and the number of landraces sampled from each province

Table 1 Data summary for 30 fluorescently-labeled microsatellite markers across 330 rice accessions

Panel	Label[a]	Marker	Chr.	Motif[b]	No. of alleles	Rare alleles[c]	Size range (bp)	Major allele[d]		PIC
								Size (bp)	Frequency (%)	
1	D2	RM433	8	(AG)13	10	7	206–236	226	69	0.45
1	D2	RM5	1	(GA)14	12	7	94–132	114	27	0.80
1	D3	RM55	3	(GA)17	12	8	153–243	229	60	0.55
1	D4	RM215	9	(CT)16	13	9	111–225	151	39	0.70
1	D4	RM514	3	(AC)12	13	8	234–274	250	40	0.71
2	D2	RM214	7	(CT)14	24	22	114–168	136	73	0.46
2	D3	RM11	7	(GA)17	12	6	122–148	140	27	0.82
2	D3	RM144	11	(ATT)11	19	15	220–283	220	36	0.78
2	D4	RM171	10	(GATG)5	6	2	321–345	341	69	0.45
2	D4	RM237	1	(CT)18	15	11	119–151	133	42	0.69
3	D2	RM133	6	(CT)8	7	4	189–233	229	50	0.46
3	D3	RM259	1	(CT)17	15	8	120–179	159	32	0.81
3	D3	RM287	11	(GA)21	14	9	95–157	115	28	0.80
3	D4	RM250	2	(CT)17	17	14	114–192	154	65	0.54
3	D4	RM507	5	(AAGA)7	5	3	216–304	256	71	0.34
4	D2	RM161	5	(AG)20	12	7	70–182	159	49	0.66
4	D3	RM124	4	(TC)10	5	1	267–275	269	37	0.67
4	D3	RM283	1	(GA)18	10	8	143–167	155	65	0.46
4	D4	RM162	6	(AC)20	17	12	205–245	211	40	0.76
4	D4	RM277	12	(GA)11	8	5	116–134	122	58	0.52
5	D2	RM431	1	(AG)16	10	6	241–263	253	47	0.66
5	D3	RM154	2	(GA)21	23	16	148–204	174	21	0.88
5	D3	RM484	10	(AT)9	4	2	292–300	296	72	0.34
5	D4	RM105	9	(CCT)6	6	1	98–137	122	32	0.72
5	D4	RM536	11	(CT)16	12	8	218–252	234	45	0.69
6	D2	RM125	7	(GCT)8	11	7	113–152	125	47	0.63
6	D2	RM19	12	(ATC)10	13	5	215–257	248	27	0.82
6	D3	RM541	6	(TC)16	24	19	104–200	184	23	0.87
6	D4	RM413	5	(AG)11	14	8	70–182	80	32	0.77
6	D4	RM474	10	(AT)13	31	26	228–306	230	26	0.88
		Mean			13	9			45	0.66

[a] Fluorescent label for Beckman CEQ: D2 = black, D3 = green, and D4 = blue

[b] Motif of the SSR and number of repeats as previously published (http://www.gramene.org)

[c] Rare alleles are defined as alleles with a frequency less than 5%

[d] Major allele is defined as the allele with the highest frequency

with an intermediate allele size. The average percent missing data across the 30 loci is 4%. The SSR genotype data for 330 accessions with the 30 SSR markers is available in Supplemental Table 1.

Data analysis

The summary statistics including the number of alleles per locus, major allele frequency, gene diversity, polymorphism information content (PIC) values, and classical F_{st} values were determined using PowerMarker version 3.23 (Liu and Muse 2005). For the unrooted phylogentic tree, genetic distance was calculated using the "C.S. Chord 1967" distance (Cavalli-Sforza and Edwards 1967) followed by phylogeny reconstruction using neighbor-joining as implemented in PowerMarker with the tree viewed using Treeview (Page 1996). The allele frequency data from PowerMarker was used to export the data in binary format (allele presence = "1" and allele absence = "0") for analysis with NTSYS-pc version 2.1 (Rohlf 1997). A similarity matrix was calculated with the Simqual subprogram using the Dice coefficient, followed by cluster analysis with the Sahn subprogram using the UPGMA clustering method as implemented in NTSYS-pc. The similarity matrix was also used for principal coordinate analysis (PCoA) with the DCenter, Eigen, Output, and MXPlot subprograms in NTSYS-pc. Bootstrapping of the UPGMA tree was performed using PowerMarker with 1,000 iterations followed by the Phylip Consense module with the majority rule setting (Felsenstein 1985). A model-based cluster analysis was then performed using the program Structure version 2.1 (Pritchard et al. 2000). The optimum number of populations (K) was selected after five independent runs of a burn-in of 10,000 iterations followed by 100,000 iterations using a model allowing for admixture and correlated allele frequencies and testing for $K = 2$ to $K = 8$ (Falush et al. 2003). The optimum value of $K = 2$ was then used to determine inferred ancestries. An analysis of multilocus associations was also performed using Brown and Feldman's (1981) method and Ohta's (1982) method and as implemented in PopGene version 1.32 (Yeh et al. 1997). For this analysis the data set was divided into two populations, representing the *indica* and *japonica* subspecies, using the previous cluster results. For Brown's analysis, the data was first converted to haploid data.

Results

Overall SSR diversity

A total of 394 alleles were detected at 30 microsatellite markers across 330 rice accessions (Supplemental Table 1). The number of alleles per locus ranged from 4 alleles (RM484) to 31 alleles (RM474), with an average of 13 alleles across the 30 loci (Table 1). The PIC values ranged from 0.34 (RM507 and RM484) to 0.88 (RM154 and RM474), with an average of 0.66. Rare alleles, defined as those alleles with a frequency less than 5%, were identified at all 30 loci, with an average of nine rare alleles per locus. The frequency of the most common allele at each locus ranged from 21% (RM154) to 73% (RM214). On average, 45% of the 330 rice accessions shared a common major allele at any given locus (Table 1).

Genetic distance-based analysis

The genetic distance-based results seen in the unrooted neighbor-joining tree revealed two major groups in the Indonesian germplasm (Fig. 2). By using predefined international accessions to assign identities to each group, the larger group corresponds to *indica*, while the smaller group corresponds to *tropical japonica*. The three *temperate japonica* controls were clustered within the larger *japonica* group, making it difficult to strictly define the boundary between the *temperate* and *tropical japonica* accessions. However, since none of the Indonesian accessions were placed directly within the *temperate japonica* sub-cluster, all were assumed to be *tropical japonica*. The three aus varieties clustered separately, but closer to the *indicas* than to the *japonicas*. Two Indonesian varieties (Ketan Putih and Jonoko) were found between the aus and the *indica* clusters. Furthermore, the single aromatic control (JC1) clustered closer to the *tropical japonica* group than to the *indicas*. Three Indonesian varieties (Jambu, Sasak Jalan and Ingsa Bondol) were near JC1 and distinct from the *tropical japonica* cluster. The three Indonesian *O. rufipogon* accessions fell between the *indica*/aus and the *japonica*/aromatic groups (Fig. 2).

The genetic similarity analysis using UPGMA clustering agreed with the neighbor-joining data, with the aus clustering near the *indica* group, and the aromatic JC1 near the *japonica* cluster (Supplementary Fig. 1). Again Sasak Jalan and Jambu appeared outside of the *japonica* cluster and near JC1, but a slight difference was seen in the placement of Ingsa Bondol between the *temperate* and *tropical japonicas*, instead of outside the *japonica* cluster as in the neighbor-joining tree. Using the UPGMA clustering to define each Indonesian accession into the *indica* and *tropical japonica* groups revealed that 68% of the Indonesian landraces were *indicas* and 32% were *tropical japonicas*. In contrast, all of the Indonesian improved varieties were *indicas*. A closer examination of the location of the Indonesian cultivars in the UPGMA cluster reveals a distinct concentration of improved cultivars in one portion of the *indica* group (in the same cluster as IR36 and IR64) while most of the Indonesian *indica* landraces are in a separate cluster, which consists of only 4% improved varieties (Supplementary Fig. 1). Likewise, a closer examination of the *tropical japonica* reveals a distinct cluster that contains several high quality Indonesian landraces (including Rojolele and Pandanwangi) and a variety from Brazil (Cabacu). This cluster is separate

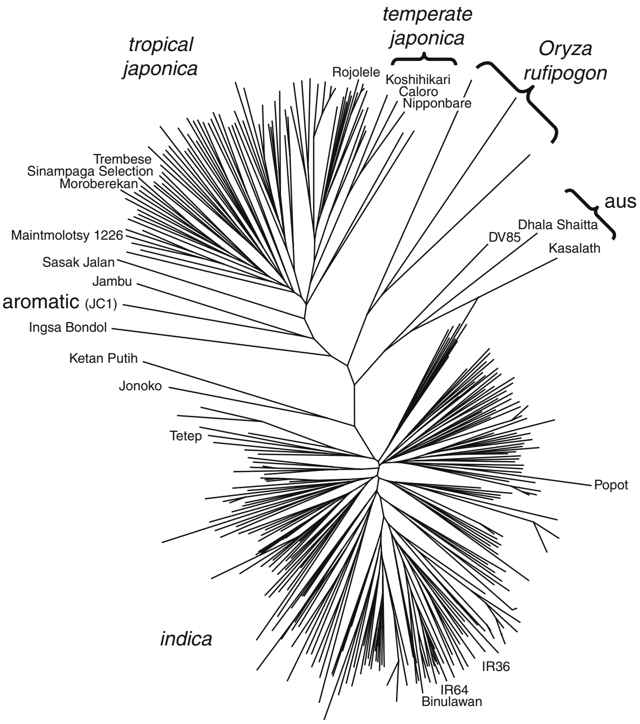

Fig. 2 An unrooted neighbor-joining tree showing the genetic relationships between the 330 rice accessions based on 30 microsatellite markers. The five major groups of *Oryza sativa* are labeled: *tropical japonica*, *temperate japonica*, aromatic, aus and *indica*, along with three accessions of *O. rufipogon*. Twelve varieties with known genetic relationships (Garris et al. 2005) are labeled, in addition to Caloro, Nipponbare, IR64, Tetep, Rojolele, and the five atypical Indonesian varieties (Sasak Jalan, Jambu, Ingsa Bondol, Ketan Putih, and Jonoko)

from the larger group of *tropical japonica*s, which includes the international varieties Moroberekan (Ivory Coast), Sinampaga Selection (Philippines), and Maintmolotsy 1226 (Madagascar) (Supplementary Fig. 1). A PCoA presented a similar picture with the accessions separating into *indica* and *japonica* groups, with the *O. rufipogon* accessions in the middle between the aus group and JC1 (Fig. 3). Again it appears that the Indonesian variety Ketan Putih is close to the aus group, and Jambu and Ingsa Bondol are near JC1. In this case, Jonoko and Popot appear to lie on the fringe of the *indica* group.

Population structure analysis

An analysis of population structure identified the highest log likelihood with the number of populations set at two ($K = 2$ using the program STRUCTURE). Since these two populations correspond to *indica* and *japonica*, it is likely that the other expected groups were not identified due to the small number of representative accessions: three *temperate japonica*, three aus, and one aromatic. Using $K = 2$, inferred ancestries for each accession were computed, revealing 84 *tropical japonica* accessions having greater than 95% shared ancestry, and 222 *indica* accessions having greater than 95% shared ancestry (Supplementary Fig. 2; Supplementary Table 1). The remaining 24 accessions were of mixed ancestry: 10 were predominantly *indica*s (sharing 73–92% *indica* ancestry) with potential admixture with *japonica*s, followed by Popot (67% *indica*), Jonoko (68% *indica*), and the three aus and Ketan Putih ranging from 47 to 65% *indica* ancestry (Supplementary Fig. 2). Interestingly, the three *O. rufipogon* accessions showed 64–83% *japonica* ancestry. JC1 showed 81% *japonica*, while three other Indonesian varieties (Ketan Siam, Ingsa Bondol, and Jambu) ranged from 71 to 82% *japonica* ancestry. Care should be taken in interpreting these results, however, since the $K = 2$ setting will force all accessions into just two populations, causing the more distantly related groups, such as the *O. rufipogon* and aus accessions, to appear to be admixtures between *indica* and *japonica*.

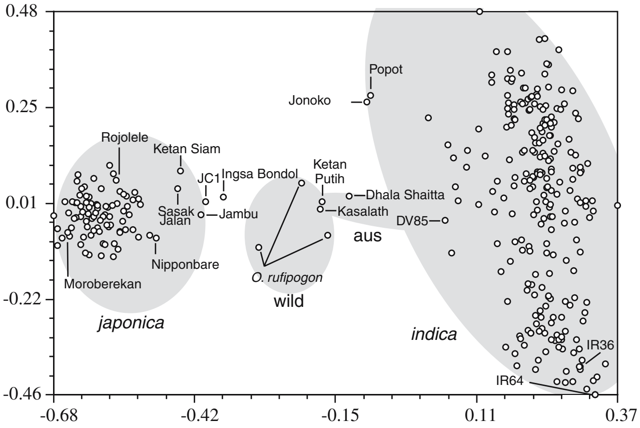

Fig. 3 Principal coordinate analysis with four groups outlined: *indica*, aus, wild, and *japonica*. Control varieties and atypical accessions are labeled

Population differentiation and diversity

When the 309 Indonesian accessions are analyzed by sub-group, the *japonica* group has slightly higher gene diversity at 0.56 compared with 0.54 for the *indica* group, although the larger *indica* group averaged 9.0 alleles per locus compared to 7.3 alleles per locus for the *japonicas* (Table 2). An estimate of the population differentiation between the Indonesian *indica* and *japonica* groups is represented by an F_{st} of 0.38. When limited to the 246 Indonesian landraces, the diversity and differentiation between the *indica* and *japonica* groups is almost the same as seen with the entire set of germplasm (Table 2). A comparison of the landraces versus the improved subgroups reveals a much lower amount of gene diversity in the improved subgroup (0.46) and only 4.9 alleles per locus, in comparison to the 0.69 gene diversity and 11.4 alleles per locus of the set of landraces (both *indica* and *tropical japonica*). This difference between the improved varieties and the landraces is still seen when the *japonica* accessions are removed from the landraces: the *indica* landraces contain a gene diversity level of 0.53 and 8.3 alleles per locus, which is greater than the *indica* improved varieties (Table 2). The population differentiation between the *indica* landraces and improved varieties is represented by an F_{st} value of 0.12. To test if the comparisons of gene diversity between groups were affected by the different sample sizes per group, the data was reanalyzed to compare populations of equal sizes by randomly selecting a subset of the larger population to equal the size of the smaller group for each comparison. Although the average numbers of alleles per locus changes depending on the population size, the average gene diversity and PIC values were more robust and gave similar results using either method (Supplementary Table 2).

A test of the population substructure at the multilocus level was performed using Brown and Feldman's (1981) method that compares single-locus versus two-locus contributions to the total variance in the number of heterozygous loci in two randomly chosen gametes within and among the subpopulations. When the *indica*

Table 2 SSR diversity and population differentiation across different sub-groups of Indonesian accessions using 30 SSR loci

Sub-groups	Sample size	Mean no. alleles/locus	Major allele frequency	Mean gene diversity	Mean PIC value	F_{st}
All Indonesian varieties						
(*indica* and *japonica*)	309	11.9	0.46	0.68	0.64	0.38
Indica	231	9.0	0.57	0.54	0.51	
japonica	78	7.3	0.56	0.56	0.53	
All Indonesian landraces						
(*indica* and *japonica*)	246	11.4	0.45	0.69	0.65	0.38
indica	168	8.3	0.59	0.53	0.49	
japonica	78	7.3	0.56	0.56	0.53	
All Indonesian varieties						
(Landraces and improved)	309	11.9	0.46	0.68	0.64	0.13
Landraces	246	11.4	0.45	0.69	0.65	
Improved	63	4.9	0.65	0.46	0.42	
All Indonesian *indica*						
(Landraces and improved)	231	9.0	0.57	0.54	0.51	0.12
Landraces	168	8.3	0.59	0.53	0.49	
Improved	63	4.9	0.65	0.46	0.42	

and *japonica* subpopulations were analyzed, the single-locus components account for 17% of the variance, while the multilocus associations accounted for 83% of the variance (Supplementary Table 3). Wahlund's covariance was very high, accounting for 88% of the variance of the two-locus components, while the average disequilibrium component accounted for just 11% of the variance. The high Wahlund's covariance and variance of disequilibrium and a low interaction component suggests founder effects or significant population subdivision (Brown and Feldman 1981). Similarly, when Ohta's (1982) method was used to compute the variance components of LD between the *indica* and *japonica* populations, the $D'^2_{IS} > D'^2_{ST}$ and $D^2_{ST} > D^2_{IS}$, indicating that limited migration and population subdivision explained the multilocus associations, rather than epistatic natural selection (Supplementary Table 4).

Discussion

The results from the genetic distance-based analysis of 309 Indonesian rice varieties with 30 microsatellite markers reveals a clear divide of Indonesian germplasm into *indica* and *tropical japonica* groups (Figs. 2, 3). Across all of the Indonesian varieties, 75% of the accessions are classified as *indicas* and 25% as *tropical japonicas*, while the subset of 246 Indonesian landraces were comprised of 68% *indica* and 32% *japonica*. These results are quite different from the 34% *indicas* and 65% *japonicas* found across 130 traditional Indonesian varieties previously analyzed with isozyme data (Glaszmann 1988). This is likely the result of different sampling strategies by the two studies, possibly due to Glaszmann's selective inclusion of *tropical japonica* ("*javanica*") type varieties. In comparison, a more recent study including 3,670 Indonesian varieties analyzed with 11 isozymes found 69% of the sampled Indonesian germplasm as *indica*, 28% as *japonica* and 3% intermediates, although no distinction was made between landraces and improved varieties (Khush et al. 2003). Since the isozyme study by Khush et al. (2003) sampled the largest number of Indonesian varieties to date, it may provide the most accurate estimate of the proportion of *indica* versus *japonica* rices in Indonesian germplasm.

In contrast to the significant proportion of *tropical japonica* accessions detected in the 246 Indonesian landraces in the current study, all of the Indonesian improved varieties sampled were *indicas*. This is likely a result of the Indonesian breeding program focusing on high-yielding irrigated rice varieties, which are largely *indica*, rather than upland varieties which tend to be *tropical japonica*. Notably, the recent effort to develop NPT varieties at IRRI evaluated a number of *bulu* varieties from Indonesia in an attempt to use *tropical japonica* germplasm to develop improved irrigated rice varieties (Virk et al. 2004). While these NPT varieties were very high yielding, they required backcrossing to *indica* germplasm for improved disease resistance and for *indica*-type grain quality (Virk et al. 2004). This effort has demonstrated the potential of using *tropical japonica* germplasm to improve irrigated rice varieties, although these NPT lines have not yet been widely used in Indonesia. In fact, our observation that Indonesian improved varieties clustered around IR64 in the dendrogram confirms the extensive use of modern high-yielding *indica* varieties from IRRI in the

Indonesian breeding program. This is highlighted by an analysis of the pedigrees of 15 Indonesian improved varieties, all of which had IR64 as a recent parent (data not shown).

A comparison of the traditional Indonesian rice categories to the SSR data revealed a strong correspondence of *cere* rice with *indica*, but less agreement between the *bulu* and *gundil* groups with tropical *japonica*: out of the 202 of the landraces with previously assigned classes, 156 were labeled *cere* and consisted of 81% *indica* and 19% tropical *japonica*, 37 were *bulu* and consisted of 43% *indica* and 57% *tropical japonica*, and 9 were classed as *gundil* and consisted of 56% *indica* and 44% *tropical japonica* (Supplementary Table 1). This weak correspondence may be due to the small sample size of the *bulu* and *gundil* groups included in the study, imprecision in the traditional methods of classifying Indonesian rice into the three groups, or a lack of association between the genetic variation at the relatively few loci controlling the primary morphological characters in comparison to the background genetic variation represented by the 30 presumably neutral SSR loci across the genome.

Strikingly, the population structure analysis showed very little admixture between the *indica* and *japonica* groups: 218 of the 231 Indonesian *indica* accessions shared over 95% *indica* ancestry, while 74 out of 78 of the Indonesian *tropical japonica* accessions shared over 95% *japonica* ancestry (Supplementary Fig. 2). This compares with a previously published analysis of global germplasm that identified 50% of the *indica* accessions over 95% shared ancestry (the rest were between 65 and 95%) and 45% of the *tropical japonicas* over 95% shared ancestry, with the remainder between 62 and 95% (Garris et al. 2005). The greater proportion of shared ancestry in the Indonesian germplasm may be a result of a narrower genetic pool, or it might reflect a statistical artifact due to the diversity of samples analyzed and the $K = 2$ population structure analysis versus the $K = 5$ analysis in Garris et al. (2005). Although it is difficult to date the introduction of *indica* and *tropical japonica* rice into Indonesia, it is likely that these two groups were introduced separately and have remained distinct despite several thousand years of sympatric cultivation in Indonesia. Possible explanations for the continued existence of these distinct genetic groups are adaptation to different environments and the partial sterility barrier that separates *indica* and *japonica* (Harushima et al. 2002; Oka 1988). In Indonesia irrigated rice is largely *indica*, while upland rice consists mostly of *tropical japonica* varieties, which are generally separate growing environments. However, in some instances *indica* and *japonica* varieties grow near each other, and in these cases the low fertility of intercrossed progeny undoubtedly acts to prevent large scale admixture between the *indica* and *tropical japonica* groups. A clear subdivision between the *indica* and *japonica* groups was also seen with significant multilocus associations using Brown and Feldman's (1981) method. While the recent study of Yu et al. (2003) also found significant multilocus effects between *indica* and *japonica*, the Indonesian data had a much higher Wahlund's covariance and lower average disquilibrium, suggesting that founder effects rather than selection may be more responsible for the multilocus associations (Brown and Feldman 1981). This difference is possibly due to the limited geographic origin of the Indonesian accessions compared to a more geographically diverse set of germplasm in Yu et al. (2003). Likewise, a greater effect from population subdivision rather than epistatic selection within the Indonesian germplasm was found when Ohta's (1982) analysis was performed.

Although all of the Indonesian varieties could be assigned as belonging to either the *indica* or *tropical japonica* groups, a few atypical accessions were detected. Five Indonesian varieties fell outside of the larger clusters: Ketan Putih (from Bengkulu) and Jonoko (Central Java) fell between the aus and *indica* groups, while the Jambu (West Sumatra) was closer to the aromatic JC1, and Ingsa Bondol (Bali) and Sasak Jalan (East Kalimantan) were related to the *japonica*s, but clustered separately from the main group. The unusual nature of these five accessions may be due to admixture with varieties recently introduced from outside of Indonesia.

It is also of interest to note that the three Indonesian *O. rufipogon* accessions showed a higher proportion of *japonica* ancestry than *indica* from the structure analysis (with a range of 64–83% *japonica* ancestry). This is in agreement to a recent report of a group of perennial *O. rufipogon* accessions (five from China and one from Indonesia) being more closely related to a cluster of *japonica O. sativa* accessions (Cheng et al. 2003). A more thorough analysis of *O. rufipogon* accessions from Indonesia is needed to compare with the accessions from China to further define the relationships between the different genetic groups within *O. rufipogon*. Considering that only 19 Indonesian *O. rufipogon* accessions are held in the IRRI germplasm collection (http://www.singer.grinfo.net/), this might also warrant additional collection of wild accessions from unexplored regions of Kalimantan and Irian Jaya (West Papua).

Abundant genetic diversity was detected across the Indonesian rice varieties, with an overall gene diversity of 0.68 and average of 11.9 alleles per locus. Surprisingly, the *tropical japonica* group had a gene diversity value of 0.56, which was slightly, but not significantly, higher than the *indica* group with a gene diversity value of 0.54. In comparison, a published survey of 234 global rice accessions (including a few Indonesian landraces) had an overall gene diversity of 0.70 and 11.8 alleles per locus, an *indica* diversity of 0.55 and 7.3 alleles per locus and a *tropical japonica* gene diversity of 0.47 and average of 6.1 alleles per locus (Garris et al. 2005). Other studies have shown higher diversity within *indica* germplasm than within the *tropical japonicas* (Glaszmann 1987; Li and Rutger 2000), although the opposite has also been found (Yu et al. 2003). In a recent study of rice germplasm in China, the *indica* group had a gene diversity of 0.68, while the *japonica* group was 0.57 (Gao et al. 2005). While the gene diversity of the Indonesian *Japonica* group (0.56) is similar to that found in China, the *indica* diversity was lower, possibly indicating a more narrow genetic base of *indica* rice in Indonesia compared with China.

While the *indica* and *tropical japonica* subsets of Indonesian germplasm both contained abundant diversity, a comparison of the improved varieties versus the landraces revealed much lower diversity in the set of 63 improved varieties sampled in this study. For example, the 168 Indonesian *indica* landraces had a gene diversity of 0.53 and an average of 8.3 alleles per locus, while the Indonesian *indica* improved varieties had a gene diversity of 0.46 and an average of 4.9 alleles per locus. The difference between the Indonesian landraces and the Indonesian improved varieties was even more pronounced than a comparison of mostly Chinese landraces and cultivars, where the *indica* landraces had a gene diversity of 0.60 and the *indica* improved varieties 0.55 (Yang et al. 1994). Furthermore, the Indonesian *indica* cultivars tended to cluster in one section of the larger *indica* group, and did not cover the complete range of *indica* diversity found in the landraces. This suggests that breeding programs can target these untapped clusters of landraces to introduce novel sources of genetic variation. One important example is a sub-group of the *tropical japonica*s that contains two popular high-quality landraces: Rojolele and Pandanwangi. These two varieties are prized in Indonesia by consumers due to their fragrant aroma and excellent taste resulting in a significant premium when sold on the market, but the desired characteristics of these two varieties have yet to be incorporated into improved Indonesian varieties. While these high quality *tropical japonica* varieties are currently grown in upland environments, it will be interesting to see if the desirable grain quality traits can be incorporated into high yielding irrigated varieties.

Given the differences seen between the *indica* and *japonica* groups as well as between Indonesian accessions versus global rice germplasm, the choice of population will be very important when planning association mapping experiments. For example, in any given set of germplasm, the *indica* and *japonica* accessions will need to be treated as separate populations to avoid false associations due to the underlying population structure, as seen in maize populations (Flint-Garcia et al. 2005). In addition, it is likely that different sets of rice germplasm will have varying amounts and distributions of LD, providing different levels of resolution for association mapping based on the size of the haplotype blocks in different regions of the genome. It is possible that isolated populations of Indonesian landraces that have experienced genetic bottlenecks would have more LD and larger haplotype blocks, providing a useful resource for an initial association mapping study, while more diverse populations may provide a higher mapping resolution once a candidate region is targeted. Detailed analyses of the amount of LD and the size of the haplotype blocks in different sets of rice germplasm and different genomic regions are needed before efficient association mapping experiments can be designed.

Acknowledgments This material is based upon work supported by the US National Science Foundation under grant no. 0301919 as an International Research Fellowship Program award to M.J.T. We also acknowledge the Indonesian Agency for Agricultural Research and Development for an RPTP 2004 grant to E.M.S. We would also like to thank Dr. Sugiono Moeljopawiro, Dr. M. Herman, and Dr. Sutrisno for hosting M.J.T. in Indonesia, Dr. Ida Hanarida Soemantri for providing access to the Indonesian rice germplasm collection, Matsuma for technical support and Lois Swales for help in formatting this manuscript.

References

Blair MW, Hedetale V, McCouch SR (2002) Fluorescent-labeled microsatellite panels useful for detecting allelic diversity in cultivated rice (*Oryza sativa* L.). Theor Appl Genet 105:449–457

Brown AHD, Feldman MW (1981) Population structure of multilocus associations. Proc Natl Acad Sci USA 78:5913–5916

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257

Cheng CY, Motohashi R, Tsuchimoto S, Fukuta Y, Ohtsubo H, Ohtsubo E (2003) Polyphyletic origin of cultivated rice: based on the interspersion pattern of SINEs. Mol Biol Evol 20:67–75

Coburn JR, Temnykh SV, Paul EM, McCouch SR (2002) Design and application of microsatellite marker panels for semiautomated genotyping of rice (*Oryza sativa* L.). Crop Sci 42:2092–2099

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

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Flint-Garcia SA, Thornsberry JM, Buckler ESt (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman M, Buckler E (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Gao LZ, Zhang CH, Chang LP, Jia JZ, Qiu ZE, Dong YS (2005) Microsatellite diversity within *Oryza sativa* with emphasis on indica–japonica divergence. Genet Res 85:1–14

Garris AJ, Tai TH, Coburn JR, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638

Glaszmann JC (1988) Geographic pattern of variation among asian native rice cultivars (*Oryza-Sativa*-L) based on 15 isozyme loci. Genome 30:782–792

Glaszmann JC (1987) Isozymes and classification of asian rice varieties. Theor Appl Genet 74:21–30

Harushima Y, Nakagahra M, Yano M, Sasaki T, Kurata N (2002) Diverse variation of reproductive barriers in three intraspecific rice crosses. Genetics 160:313–322

IRRI (1967) In: Annual report for 1966, International Rice Research Institute (IRRI), Manila, Philippines, p 59–82

Jain S, Jain R., McCouch S. (2004) Genetic analysis of Indian aromatic and quality rice (*Oryza sativa* L.) germplasm using panels of fluorescently-labeled microsatellite markers. Theor Appl Genet 109:965–977

Khush GS (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol Biol 35:25–34

Khush GS, Brar DS, Virk PS, Tang SX, Malik SS, Busto GA, Lee YT, McNally R, Trinh LN, Jiang Y, Shata MAM (2003) Classifying rice germplasm by isozyme polymorphism and origin of cultivated rice. In IRRI (ed) Discussion Paper No. 46. International Rice Research Institute, Los Banos

Li ZK, Fu BY, Gao YM, Xu JL, Ali J, Lafitte JR, Jiang YZ, Rey JD, Vijayakumar CHM, Maghirang R, Zheng TQ, Zhu LH (2005) Genome-wide introgression lines and their use in genetic and molecular dissection of complex phenotypes in rice (*Oryza sativa* L.). Plant Mol Biol 59:33–52

Li Z, Rutger JN (2000) Geographic distribution and multilocus organization of isozyme variation of rice (*Oryza sativa* L.). Theor Appl Genet 101:379–387

Liu K, Muse SV (2005) PowerMarker: Integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129

Lu H, Redus MA, Coburn JR, Rutger JN, McCouch SR, Tai TH (2004) Population structure and breeding patterns of 145 US rice cultivars based on SSR marker analysis. Crop Sci 45:66–76

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Ohta T (1982) Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci USA 79:1940–1944

Oka HI (1988) Functions and genetic base of reproductive barriers. In Oka HI, (ed) Origin of cultivated rice. Tokyo/Elsevier Science/Japan Scientific Societies Press, Amsterdam, pp 156–159, 181–209

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Mol Biol 12:357–358

Peng S, Cassman KG, Virmani SS, Sheehy J, Khush GS (1999) Yield potential trends of tropical rice since the release of IR8 and the challenge of increasing rice yield potential. Crop Sci 39:1552–1559

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rohlf F (1997) NTSYS-pc: numerical taxonomy and multivariate analysis system, 2.1 edn. Department of Ecology and Evolution, State University of NY, Stony Brook

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

Takahashi N (1997) Differentiation of ecotypes in cultivated rice. In Matsu T, Futsuhara Y, Kikuchi F, Yamaguchi H (eds) Science of the rice plant, vol 3. Genetics, Tokyo, pp 112–118

Tang JB, Xia HA, Cao ML, Zhang XQ, Zeng WY, Hu SN, Tong W, Wang J, Yu J, Yang HM, Zhu LH (2004) A comparison of rice chloroplast genomes. Plant Phys 135:412–420

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066

Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:697–712

Virk PS, Khush GS, Peng J (2004) Breeding to enhance yield potential of rice at IRRI: the ideotype approach. Int Rice Res Notes 29:5–9

Vitte C, Ishii T, Lamy F, Brar DS, Panaud O (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). Mol Genet Genomics 272:504–511

Yang GP, Maroof MAS, Xu CG, Zhang Q, Biyashev RM (1994) Comparative analysis of microsatellite DNA polymorphism in landraces and cultivars of rice. Mol Gen Genet 245:187–194

Yeh FC, Yang RC, Boyle TBJ, Ye ZH, Mao JX (1997) POPGENE, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Canada

Yu SB, Xu WJ, Vijayakumar CHM, Ali J, Fu BY, Xu JL, Jiang YZ, Marghirang R, Domingo J, Aquino C, Virmani SS, Li ZK (2003) Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. Theor Appl Genet 108:131–140

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