Genetic Diversity of Tropical Hybrid Rice Germplasm Measured by Molecular Markers

HE Zhi-zhou\textsuperscript{1,2}, XIE Fang-ming\textsuperscript{2}, CHEN Li-yun\textsuperscript{1}, Madonna Angelita DELA PAZ\textsuperscript{2}

(\textsuperscript{1}Rice Research Institute, Hunan Agricultural University, Changsha 410128, China; \textsuperscript{2}International Rice Research Institute, Metro Manila, DAPO Box 7777, Philippines)

Abstract: Investigation of genetic diversity and relationships among breeding lines is of great importance to facilitate parent selection in hybrid rice breeding programs. In this study, we characterized 168 hybrid rice parents from International Rice Research Institute with 207 simple sequence repeat (SSR) and 353 single nucleotide polymorphism (SNP) markers. A total of 1 267 SSR and 706 SNP alleles were detected with the averages of 6.1 (SSR) and 2.0 (SNP) alleles per locus respectively across all lines. Based on the genetic distances estimated from the SSR and SNP markers separately and combined, the unrooted neighbor-joining cluster and STRUCTURE analyses consistently separated the 168 hybrid rice parents into two major groups: B-line and R-line, which is consistent with known parent pedigree information. The genetic distance matrices derived from the SSR and SNP genotyping were highly correlated ($r = 0.81$, $P < 0.001$), indicating that both of the SSR and SNP markers have distinguishable power to detect polymorphism and are appropriate for genetic diversity analysis among tropical hybrid rice parents. A subset of 60 SSR markers were also chosen by the Core Hunter with 368 alleles, and the cluster analysis based on the total and subset of SSR markers highly corresponded at $r = 0.91$ ($P < 0.001$), suggesting that fewer SSR markers can be used to classify and evaluate genetic diversity among parental lines.

Key words: tropical hybrid rice; genetic diversity; simple sequence repeat marker; single nucleotide polymorphism marker

Germplasm diversity and genetic relationships are essential information in crop improvement programs. Current tropical hybrid rice breeding programs rely mainly on the use of cytoplasmic male sterility (CMS) system, which includes two major parental groups: B lines (male sterility maintainer lines) and R lines (male sterility restorer lines). To fully exploit heterosis among the B and R lines, detailed and comprehensive information of genetic diversity and structure among parental lines is required, which is particularly important for parental selection in designing crossing programs (Hallauer et al, 1988).

Traditionally, the coefficient of parentage (COP), calculated from the pedigree records (Malécot, 1948), is used by breeders as an indirect measure to assess genetic similarity among breeding germplasms. Apart from pedigree information, diversity analysis could also be based on morphological, physiological or biochemical markers. Since 1990’s, several types of molecular markers are being increasingly utilized for investigating germplasm diversity and genetic relationships (Hamrick and Godt, 1997; Westman and Kresovich, 1997; Melchinger, 1999). Among these DNA markers, simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers are currently widely employed for rice genetic diversity analysis (Chen et al, 1997; Garris et al, 2005; McNally et al, 2009; Chen et al, 2011). SSR markers are codominant, multiallelic and highly polymorphic whereas SNP markers are more abundant, which can be done at high throughput and automated for genotyping (Ching et al, 2002; Duran et al, 2009; Edwards and Batley, 2010). To evaluate the performance and relationship between these two widely used molecular markers in genetic diversity analysis, several comparative studies in recent years have been conducted. Van Inghelandt et al (2010) examined 1 537 commercial maize inbred lines with 359 SSR and 8 244 SNP markers, which revealed the same results regarding the diversity and the structure of heterotic pools. Genetic distances were also observed to be highly correlated between the two marker systems and indicated that the genetic distance estimates could be more precise based on SNP markers than on SSR markers. Hamblin et al (2007), Jones et al (2007) and Nelson et al (2011) proved that SNP markers could provide increased marker data quality whereas SSR markers were more powerful to assign inbreds to sub-populations, and genetic distances were highly correlated for more closely related germplasm, but much weaker for more distantly related germplasm, and suggested that with increasing SNP markers, an equivalent polymorphism and discriminating power might be obtained.

The successful hybrid maize breeding in the past century had a major contribution to extensive studies in the relationship of genetic diversity of maize inbred
lines and heterotic groups. In rice, the overall genetic diversity of representative samples of the global rice germplasm has been well investigated (Yu et al, 2003; Garris et al, 2005; Caicedo et al, 2007). With the development of hybrid rice breeding, studies in the genetic diversity and heterotic groups have also attracted an increasing attention. Li et al (1999) evaluated 33 parental lines with 20 SSR markers and classified the parental lines into three groups. Xu et al (2002) investigated 80 tropical hybrid parental lines based on their pedigrees and SSR genotyping and found that the genetic diversity among the R lines was greater than that among the B lines. In the evaluation of genetic diversity of 100 Chinese parental lines of hybrid rice based on COP data, Wang and Lu (2006) found a potentially high degree of diversity among the hybrid rice germplasm wherein the cluster analysis identified 10 clusters which clearly separated the B lines, R lines and other parental lines to different groups. Additionally, Duan et al (2002) observed a low genetic diversity among 35 restorer lines genotyped with 25 SSR markers, which indicated a vulnerable genetic background and a limited utilization in rice heterosis breeding. To accelerate the development of hybrid rice in the tropics, detailed analysis of genetic diversity and the relationships among parental lines should be well characterized. To our knowledge, the comparative analysis on genetic diversity and structure of tropical hybrid rice parents based on SSR and SNP markers has not been reported.

In this study, the genetic diversity and structure among 168 tropical hybrid rice parental lines as genetic donors in hybrid breeding programs were analyzed using SSR, SNP markers and pedigree information. Also, the efficiency of using a subset of SSR markers to reflect the genetic diversity among the potential parental lines was also explored.

**MATERIALS AND METHODS**

**Rice materials**

One hundred and sixty-eight rice inbred lines, representing the diversity available among the current and historical germplasms in hybrid rice breeding programs at International Rice Research Institute (IRRI) were chosen for SSR and SNP genotyping. These lines have been historically used in the breeding programs in the tropical Asian countries and many are the parents of commercial hybrids. Of the 168 inbred lines, 103 are maintainer (B) lines and 65 are restorer (R) lines. The pedigree information of these lines was retrieved from IRRI International Crop Information System (ICIS) database.

**SSR genotyping**

Two leaves from five 21-day old healthy plants per line were bulked, lyophilized, and placed in 50 mL Falcon tubes and sent to Pioneer Hi-Bred Ltd. in Canada for DNA extraction and SSR genotyping with the protocol of Smith et al (2010). For DNA extraction, the CTAB method was conducted and the DNA content was adjusted to 3 ng/µL as polymerase chain reaction (PCR) templates. DNA profile of each inbred line was done using 207 fluorescent dyed-labeled forward SSR markers synthesized by Applied Biosystems. PCR was performed in 384-well PCR plates, with each reaction containing 9 ng of genomic DNA, 0.1 µmol/L of each primer, 2.5 mmol/L MgCl2, 0.5 mmol/L of each dNTP, 1 × PCR buffer, and 0.2 U of Taq Gold DNA polymerase. Reactions were carried out in a BIORAD Tetrad2 Peltier Thermal Cycler (Hercules, CA) with 10 min at 95 °C, followed by 35 cycles of 30 s at 94 °C, 30 s at 60 °C, and 30 s at 72 °C, with a final extension of 72 °C for 5 min. The PCR products were multiplexed into groups of six wherein 4 µL of each PCR product was added to 220 µL of deionized water to make a total volume of 244 µL. The samples (2 µL of each multiplexed sample and 8.04 µL of loading buffer) loaded in a 96-well plate were genotyped on an ABI 3730 DNA Analyzer (Applied Biosystems) for capillary electrophoresis. PCR fragment sizes were determined using the GeneMapper v4.7 software (Applied Biosystems), followed by manual allele binning.

**SNP genotyping**

SNP genotyping was conducted at Molecular Marker Applications Laboratory (MMAL) at IRRI using the GoldenGate Genotyping Protocol of the Veracode Technology. The 384-plex indica × indica oligonucleotide pool assay (OPA), a subset from the 44K-SNP Affymetrix array was used as markers (McCouch et al, 2010). Scan results were generated from Illumina’s BeadXpress Reader, where the allele calls generated from Genome-Studio software were corrected using the ALCHEMY software (http://www.illumina.com/systems/beadxpress.ilmn; Wright et al, 2010). SNP markers that were monomorphic or with more than 20% missing data were considered non-informative, and were removed from further analysis. A total of 353 SNP markers were retained for genetic diversity and structure analysis.

**Statistical analysis**

COP values between all pairs of parental lines were determined with fully expanded genealogical information extracted from the IRRI ICIS database. Pedigree distances were calculated as 1-COP value.
All analyses described below were performed for SSR data as well as for SNP data. The summary statistics include the number of alleles per locus, major allele frequency, gene diversity and polymorphism information content (PIC) values. The Fst values were determined using PowerMarker version 3.25 (Liu and Muse, 2005). Genetic distances were calculated using the C. S. Chord distance since it has been shown by analysis of simulations to generate correct tree topologies regardless of the microsatellite mutation model (Cavalli-Sforza and Edwards, 1967; Takezaki and Nei, 1996). The unrooted phylogenetic tree was based on the neighbor-joining method implemented in the PowerMarker with the tree viewed using MEGA version 5.05 (Tamura et al, 2007).

The model-based program STRUCTURE (Pritchard et al, 2000) was used to infer population structure and to assign individuals to populations. Models with a putative number of sub-populations (K) from 1 to 10 with admixture and correlated allele frequencies were considered (Falush et al, 2003). Five independent runs of a burn-in of 10 000, run length of 100 000 iterations for each K were implemented. To determine the K value, both LnP(D) value and Evanno’s ΔK were used (Evanno et al, 2005). LnP(D) is the log likelihood of the observed genotype distribution in K clusters and can be output by STRUCTURE simulation. Evanno’s ΔK takes consideration of the variance of LnP(D) among repeated runs and can usually indicate the ideal K. The optimum value of K was then used to determine inferred ancestries. An individual was assigned to a specific population if it had more than 0.8 membership in that population, while inbreds with membership probabilities less than 0.8 were assigned to an admixed group. Using the software GenAlEx 6.41, Mantel test was performed to assess the relationships between pairs of genetic matrices generated using SSR, SNP and pedigrees. A value of 0.61–0.80 suggests that there is a strong correlation whereas a value of 0.81–0.99 suggests a very strong correlation between the two matrices being compared. The analysis of molecular variance (AMOVA) was subjected to estimate molecular diversity at each hierarchical level among and within the groups for SSR markers wherein the significance of PhiPT value was tested with 9 999 permutations.

The program Core Hunter, an algorithm for sampling genetic resources based on multiple genetic measures (Thachuk et al, 2009), was employed to determine a subset of SSR markers from the original 207 polymorphic SSR markers. Mantel test was used to compute the cophenetic correlation (r) to test the goodness of fit of the cluster analysis to the similarity matrix generated by using the total set of markers (207 SSR markers) and the subset of markers selected by the Core Hunter.

**RESULTS**

**Overall genetic diversity revealed by SSR and SNP markers**

The complete set of 207 SSR markers detected a total of 1267 alleles across the 168 hybrid rice parental lines whereas the subset of 60 SSR markers chosen by Core Hunter detected a total of 368 alleles. The allele numbers per locus for the 207 SSR markers ranged from 2 to 28 with an average of 6.12 (Table 1), while the PIC values ranged from 0.05 to 0.87 with an average of 0.45. Gene diversity ranged from 0.05 to 0.88 with an average of 0.50, indicating a moderate level of diversity for these hybrid rice parents. Major allele frequency, defined as the frequency of the most common allele at each locus, ranged from 22% to 98%. On average, 62% of the 168 parental lines shared a common major allele at any given loci.

For the 353 SNP markers, as expected, a total of 706 alleles with an average of 2.00 alleles per locus were observed and the average PIC value was 0.23, and the average gene diversity was 0.29, nearly as half as that of SSR markers. When the SSR and SNP data were combined for analysis, the polymorphism results showed a medium value between SSR and SNP markers. The overall fixation index Fst were 0.16, 0.15 and 0.15 for the SSR, SNP and SSR-SNP combined markers, respectively.

**Genetic distance (GD) based on molecular markers and pedigrees**

The GD matrices for the total set (207) and subset (60) of SSR, 353 SNP, and the combined (207 SSR and 353 SNP) molecular markers were calculated using the same method (C.S. Chord distance), while the pedigree distance was calculated as 1-COP value (Table 2). The average GD between pairs of all the parental lines was 0.45 based on the SSR markers and 0.26 based on the SNP markers, whereas the GD calculated from the pedigree records was 0.85, which was much higher than the molecular GD estimates.

There was a highly significant correlation (r = 0.91, P < 0.001) between the total set (207) and subset (60) of SSR markers. The Mantel test correlation (Table 3) between GD estimates based on the total set (207) SSR and 353 SNP markers was also highly significant at r = 0.81 (P < 0.001). The GD estimates based on the SSR-SNP combined markers were significantly corrected with the total set (207) SSR GD at r = 0.94 (P < 0.001), and with the SNP GD at r = 0.96 (P < 0.001). Moreover,
the correlation coefficient between GD estimates based on the molecular markers and the GD based on the pedigree was only moderately significant at \( r = 0.66 \) for SSR, and at \( r = 0.60 \) for SNP (\( P < 0.001 \)).

### Genetic relationships and clustering

The neighbor-joining trees (Fig. 1) based on the GDs of the SSR (207), SNP (353) and SSR-SNP combined (560) markers consensually revealed two major groups for the 168 hybrid rice parents. The majority of the B and R lines were separated into their respective clusters, which were in accordance with the parental pedigree information. The STRUCTURE program implements a model-based clustering method for inferring relationships between populations and assigning individuals to each population. When we ran the STRUCTURE simulation using all the 168 lines based on SSR, SNP and SSR-SNP combined markers, the Ln\( P(D) \) value increased with \( K \) from 1 to 10, but showed an evident increase at \( K = 2 \), suggesting two distinctly divergent populations. There was also a sharp peak of Evanno’s \( \Delta K \) at \( K = 2 \). These statistical parameters confirmed that the 168 hybrid rice parents were most probably grouped as two major populations (Fig. 2), which was also consistent with the clustering based on genetic distances. Using \( K = 2 \), based on the membership probability threshold of 0.8, 83% of the lines were classified into one of the two groups for SSR markers, 68.5% for SNP markers, and 77.4% for SSR-SNP combined markers. AMOVA indicated that most of the genetic variances were within groups, accounted for 84.2%, 84.9% and 84.6% (\( P < 0.001 \)) of the total variances revealed by SSR, SNP and SSR-SNP combined markers while the remaining 15% was due to the differences among groups.

### Genetic diversity and differentiation within groups

The genetic diversity indices based on molecular markers and pedigree information were calculated for each group separately (Table 1). For SSR markers, the average number of alleles per locus ranged from 1 to 14 with an average of 5.09 for the R lines. For SNP

### Table 1. Average (minimum–maximum) of polymorphism for all lines and by groups.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All lines (168)</th>
<th>B line (103)</th>
<th>R line (65)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of alleles per locus</td>
<td>6.12 (2–28)</td>
<td>4.52 (1–14)</td>
<td>5.09 (2–20)</td>
</tr>
<tr>
<td>PIC value</td>
<td>0.45 (0.05–0.87)</td>
<td>0.40 (0.00–0.81)</td>
<td>0.42 (0.03–0.84)</td>
</tr>
<tr>
<td>Gene diversity</td>
<td>0.50 (0.05–0.88)</td>
<td>0.44 (0.00–0.83)</td>
<td>0.47 (0.03–0.85)</td>
</tr>
<tr>
<td>Major allele frequency</td>
<td>0.62 (0.22–0.98)</td>
<td>0.66 (0.24–1.00)</td>
<td>0.65 (0.24–0.98)</td>
</tr>
<tr>
<td>Fst</td>
<td>0.16</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>SNP marker (( n = 353 ))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of alleles per locus</td>
<td>2.00 (2–2)</td>
<td>1.97 (1–2)</td>
<td>1.99 (1–2)</td>
</tr>
<tr>
<td>PIC value</td>
<td>0.23 (0.01–0.38)</td>
<td>0.21 (0.00–0.38)</td>
<td>0.23 (0.00–0.38)</td>
</tr>
<tr>
<td>Gene diversity</td>
<td>0.29 (0.01–0.50)</td>
<td>0.25 (0.00–0.50)</td>
<td>0.28 (0.00–0.50)</td>
</tr>
<tr>
<td>Major allele frequency</td>
<td>0.79 (0.50–1.00)</td>
<td>0.82 (0.50–1.00)</td>
<td>0.80 (0.50–1.00)</td>
</tr>
<tr>
<td>Fst</td>
<td>0.15</td>
<td>0.18</td>
<td>0.09</td>
</tr>
<tr>
<td>SSR and SNP markers combined (( n = 560 ))</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of alleles per locus</td>
<td>3.52 (2–28)</td>
<td>2.91 (1–14)</td>
<td>3.14 (1–20)</td>
</tr>
<tr>
<td>PIC value</td>
<td>0.31 (0.01–0.87)</td>
<td>0.28 (0.00–0.81)</td>
<td>0.30 (0.00–0.84)</td>
</tr>
<tr>
<td>Gene diversity</td>
<td>0.37 (0.01–0.88)</td>
<td>0.32 (0.00–0.83)</td>
<td>0.35 (0.00–0.85)</td>
</tr>
<tr>
<td>Major allele frequency</td>
<td>0.73 (0.22–1.00)</td>
<td>0.76 (0.24–1.00)</td>
<td>0.75 (0.24–1.00)</td>
</tr>
<tr>
<td>Fst</td>
<td>0.15</td>
<td>0.18</td>
<td>0.11</td>
</tr>
</tbody>
</table>

### Table 2. Pairwise genetic distance calculated from molecular markers and pedigrees.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All lines</th>
<th>B line</th>
<th>R line</th>
<th>B line vs R line</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of pairs</td>
<td>14 028</td>
<td>5 253</td>
<td>2 080</td>
<td>6 695</td>
</tr>
<tr>
<td>SSR</td>
<td>0.45 (0.00–0.76)</td>
<td>0.40 (0.00–0.68)</td>
<td>0.43 (0.01–0.76)</td>
<td>0.49 (0.26–0.74)</td>
</tr>
<tr>
<td>SNP</td>
<td>0.26 (0.00–0.50)</td>
<td>0.23 (0.00–0.43)</td>
<td>0.26 (0.00–0.50)</td>
<td>0.29 (0.12–0.50)</td>
</tr>
<tr>
<td>SSR-SNP combined</td>
<td>0.33 (0.00–0.58)</td>
<td>0.29 (0.00–0.51)</td>
<td>0.32 (0.00–0.58)</td>
<td>0.36 (0.23–0.57)</td>
</tr>
<tr>
<td>Pedigree</td>
<td>0.85 (0.00–1.00)</td>
<td>0.78 (0.00–1.00)</td>
<td>0.83 (0.09–1.00)</td>
<td>0.91 (0.45–1.00)</td>
</tr>
</tbody>
</table>

### Table 3. Correlation coefficients of Mantel test between genetic distance estimates.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All lines</th>
<th>B line</th>
<th>R line</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR vs SNP</td>
<td>0.81</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>SSR vs SSR-SNP combined</td>
<td>0.94</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>SNP vs SSR-SNP combined</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>SSR vs pedigree</td>
<td>0.66</td>
<td>0.72</td>
<td>0.36</td>
</tr>
<tr>
<td>SNP vs pedigree</td>
<td>0.60</td>
<td>0.66</td>
<td>0.27</td>
</tr>
<tr>
<td>SSR-SNP combined vs pedigree</td>
<td>0.66</td>
<td>0.72</td>
<td>0.33</td>
</tr>
</tbody>
</table>
markers, the number of alleles per locus ranged from 1 to 2 with an average of 1.97 for the B lines, and 1.99 for the R lines. The combined markers revealed 2.91 alleles per locus for the B lines, and 3.14 alleles per locus for the R lines. All of these indices indicated that the R lines had slightly higher gene diversity and PIC values than the B lines. The genetic distance calculated from molecular markers and pedigree information also showed that GDs within the R lines were greater than those within the B lines. The GD between the B and R lines was higher than that within the B and R lines (Table 2).

We performed independent clustering within the B and R lines, and found that each of the two major groups (B and R groups) could be further subdivided into several sub-groups. Based on the combined molecular markers, the 103 B lines could be further divided into three sub-groups. Sub-group B3 consisted of 14 entries that were derived from IR68897B, IR62829, IR46830B and IR54752B, and sub-group B2 was composed of 36 entries, most of which were

**Fig. 1.** Neighbor-joining trees and STRUCTURE plot when $K = 2$ for the 168 hybrid rice parents based on C. S. Chord genetic distances calculated from 207 SSR (A), 353 SNP (B) and 560 SSR-SNP combined markers (C).

The red and green lines represent B and R lines, respectively. The y-axis indicates the probability of population membership.

**Fig. 2.** Average LnP(D) and $\Delta K$ over five repeats of STRUCTURE simulations.

LnP(D) with $K = 1$–10, $\Delta K$ with $K = 2$–9 for the 560 combined markers.
directly derived from the crosses involved with IR58025B as shown in their pedigrees. Sub-group B1 included 53 entries that were derived from diverse germplasm sources.

The 65 R lines could also be divided into three subgroups. Sub-group R1 contained only one R line, sub-group R2 was composed of 22 entries, and sub-group R3 included 42 entries. Both of the sub-groups R2 and R3 were derived from diverse parental lines.

**DISCUSSION**

We investigated 168 hybrid rice parents with 207 SSR markers and 353 SNP markers, and the analysis revealed an abundant allelic variation with an average of 6.12 alleles per locus over 207 SSR loci, which exceeded the previously reported values of 2.6 (Duan et al, 2002), 4.2 (Xu et al, 2002) and 4.56 (Wang and Lu, 2006) alleles per locus. The higher number of alleles per locus observed in the present study can be attributed to the larger number of parental lines surveyed and the inclusion of more SSR markers than in the previous studies.

When comparing the diversity results generated by SSR and SNP markers, we found that the number of alleles per locus, gene diversity and PIC values based on SSR markers were higher than those computed from SNP markers, indicating that SSR markers are more polymorphic than SNP markers and the result is consistent with previous studies in other crops (Van Inghelandt et al, 2010; Varshney et al, 2010). The average number of alleles per SNP locus was 2.00 due to the fact that SNP markers are usually biallelic and the maximum gene diversity observable with biallelic markers is 0.5, whereas for multi-allelic markers such as SSR markers, the maximum can be 1. This can explain the lower polymorphism observed in SNP markers than in SSR markers on an individual marker basis (Vignal et al, 2002). The different mutational properties of these two classes of markers result in differences in heterozygosities and allele frequencies. Differences in the major allele frequency may also account for the differences in information content of SSR and SNP markers. SNP markers (0.79) have a higher major allele frequency than SSR markers (0.62). Despite the differences in the average of polymorphism parameters calculated from SSR and SNP markers, we observed the same trends for SSR, SNP and SSR-SNP combined markers in these parameters for all the studied lines as well as for the B and R lines separately. These results imply that both markers are appropriate to examine genetic diversity in hybrid rice parental lines.

Our results showed that the genetic distance estimates were considerably lower for the molecular markers than for the pedigrees. Within the molecular markers, the SNP-GD was much lower than the SSR-GD. This finding is also in accordance with the previous studies in maize (Hamblin et al, 2007; Jones et al, 2007; Van Inghelandt et al, 2010). Theoretically, the estimates of genetic distance based on pedigree data will somewhat differ from those based on molecular marker data due to the different assumptions of the calculations. The difference of genetic distance estimates between molecular marker systems may be explained by the different number of alleles per locus, which resulted in different average allele frequencies. Despite the differences in the GD estimates calculated from molecular markers and pedigrees, we found significant correlations between these estimates (Table 3). Estimates of GD based on SSR and SNP markers were also well correlated, indicating that both markers have similar distinguishing power within the studied hybrid parental lines. Highly significant correlations between molecular markers and pedigree-based distances were well documented in previous studies wherein a usually high correlation coefficient can be found with detailed and accurate pedigree records (Bernardo et al, 2000; Liu et al, 2003; Van Inghelandt et al, 2010). Generally, molecular markers reflect the actual level of genetic variation existing among genotypes at the DNA level and therefore provide a more accurate estimate than pedigree information. Although marker-based distance is more reliable for genetic relationship evaluation, pedigree record is cost-effective and valuable for breeders, therefore, we suggest the integration of molecular data and pedigree information will greatly facilitate accurate diversity analysis to improve breeding efficiency.

It was also shown in this study that though the program Core Hunter was originally developed for sampling genetic resources, it was also a powerful tool in choosing a subset of SSR markers. For the objective of choosing a core set of SSR markers, the number of effective alleles which is correlated with the expected proportion of heterozygous loci and a measure of the average number of alleles within a locus, was optimized wherein, a subset of 60 markers from a total of 207 SSR markers was able to discriminate the B lines from the R lines.

There are broadly two types of clustering methods, one is the distance-based method and the other is the model-based method. In this study, the genetic distance-based neighbor-joining trees and model-based population structure analysis for SSR, SNP and combined markers accordingly revealed two clear
The B-line group and the R-line group are the two major groups for the hybrid parents, which were in agreement with the parental grouping information. The B-line group and the R-line group are the two major parental groups in three-line hybrid rice system. In the studied hybrid rice parents, we detected a low genetic diversity within the B lines compared with the R lines. It may be due to that the B lines were developed from limited germplasm sources during the earlier stage of tropical hybrid rice breeding program at IRRI. As shown in the pedigree information, a lot of B lines were derived from the crosses with an elite B-line (IR 58025B), which is the most popular hybrid rice parent used in the tropical Asia countries. The tendency to choose a few elite parental lines for developing new breeding lines resulted in the abundance of homonyms and synonyms in germplasm collections and also raising the attention for narrow genetic diversity in hybrid rice breeding programs.

Another significant finding was that the cluster analysis of this study clearly revealed two major parental groups, but showed a certain degree of parental mixture genetically. There are some B lines grouped into the R-line cluster, and vice-versa. These admixed parents are likely to be the result of shared ancestry with the other groups during the process of new B- and R-line breeding. Wang and Lu (2006) suggested that differential selection and mutation effect could also be looked into as possible factors for such admixture. However, by using molecular markers, it will be possible to avoid from the crosses between parental groups and selection with admixed ancestry based on population structure assignment, which will lead to the development of heterotic pools in hybrid rice and an improved degree of heterosis.

It is important to assign parental lines to subgroups within the two major parental pools where criteria for shaping sub-groups can be arbitrary or natural (Laurentin, 2009). In the clustering analysis, three sub-groups within each of the two major groups were identified, which suggests that selection of parental lines from different sub-groups might be an effective way for making hybrid combinations combined with field performance of hybrid. It has been observed that lines within a group or sub-group have a low level of genetic dissimilarity and that crosses between genetically divergent lines selected from different groups or sub-groups generally produce better-performing hybrids than the closely related parents (Tracy and Chandler, 2006; Lu et al, 2009).

Several studies have compared SSR and SNP markers for genetic relationship and grouping analysis in maize. Hamblin et al (2007) found 69 SSR markers performed better at clustering germplasm into populations than 847 SNP markers did. Van Inghelandt et al (2010) also found that SSR markers gave higher group membership probabilities than SNP markers in population structure analysis. In the present study, we found more mixed lines in the SNP neighbor-joining tree than in SSR and in combined markers. For $K = 2$, the STRUCTURE analysis revealed 32.5% lines unassigned to its corresponding group based on SNP markers, whereas 17.3% and 22.6% unassigned lines for SSR and combined markers. According to Laval et al (2002), $(K−1)$ times more SNP markers should be needed to achieve the same clustering power as a set of SSR markers with $K$ alleles. Yu et al (2009) proposed the ratio of 1:10 for SSR:SNP to provide robust kinship analysis. Van Inghelandt et al (2010) suggested that 7 to 11 times more SNP markers than SSR markers should be used for analyzing genetic relationships. Nelson et al (2011) found that, where SNP markers are selected to maintain high gene diversity and genome coverage, only 2–3 times more SNP markers are needed to reveal associations among lines compared with SSR markers.

**CONCLUSIONS**

The results of this study indicate that both of the SSR and SNP markers are capable of discriminating the two major groups of the 168 hybrid rice parents which were in general supported by the pedigree data from ICIS. Though SSR markers provide more polymorphism information than SNP markers, SNP markers have a number of technical advantages. Given the trend towards increased use of SNP markers, by increasing SNP numbers and selecting SNP markers with high gene diversity and even genome coverage, it might improve the distinguishing power of SNP markers in rice genetic diversity analysis. The results also indicate that the combined data from SSR and SNP markers can provide more detailed diversity information and verify the independent analysis of the markers.

**ACKNOWLEDGEMENTS**

The authors thank T. LOMAS and his team from the Pioneer Hi-Bred International, Inc. for conducting the SSR genotyping. We are grateful to R. VAIRAMANI from the Pioneer Hi-Bred International, Inc., for his critical reading and editing of the manuscript. This research was supported by the IRRI-Pioneer Scientific Knowledge Exchange Program (SKEP) and the China Scholarship Council (CSC) (Grant No. 2009325011).
REFERENCES


