Assessment on Evaluating Parameters of Rice Core Collections Constructed by Genotypic Values and Molecular Marker Information

WANG Jian-cheng 1, 2, HU Jin 1, ZHANG Cai-fang 1, ZHANG Sheng 1
(1 Seed Science Center, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China; 2 Crop Research Institute, Shandong Academy of Agricultural Sciences, Ji’nan 250100, China)

Abstract: Eleven evaluating parameters for rice core collection were assessed based on genotypic values and molecular marker information. Monte Carlo simulation combined with mixed linear model was used to eliminate the interference from environment in order to draw more reliable results. The coincidence rate of range (CR) was the optimal parameter. Mean Simpson index (M\(_S\)), mean Shannon-Weaver index of genetic diversity (M\(_D\)) and mean polymorphism information content (M\(_PIC\)) were important evaluating parameters. The variable rate of coefficient of variation (VR) could act as an important reference parameter for evaluating the variation degree of core collection. Percentage of polymorphic loci (p) could be used as a determination parameter for the size of core collection. Mean difference percentage (MD) was a determination parameter for the reliability judgment of core collection. The effective evaluating parameters for core collection selected in the research could be used as criteria for sampling percentage in different plant germplasm populations.

Key words: core collection; genotypic value; molecular marker information; Monte Carlo simulation; mixed linear model; evaluating parameter; rice

Germplasm collections are important for crop improvement and research. The existence of human being will depend on the quantity and diversity of germplasm collections to a great extent in the future. Many countries and organizations have founded hundreds of gene banks, and millions of crop resources have been preserved [1]. However, with continuous collecting of germplasm resources, the sizes of collections have been becoming larger and larger, which hindered the preservation, the evaluation, the research and the use of germplasm resources. For this reason, Frankel [2] initiated a ‘core collection’ concept. A core collection is defined as a representative sample of the entire collection with minimum repetitiveness and maximum genetic diversity of a crop species and its relatives. With the core collection, it is convenient to study and utilize germplasm resources, which has received the extensive attention all over the world. Rice is one of the most important crops in China. Abundant rice germplasm preservation contributes to advanced rice breeding in China. Up to 2000, there had been 75597 accessions of rice germplasm preserved in China [3]. Therefore, rice was the first crop used for core collection research in China [4]. However, the huge amount disturbed the utilization of rice germplasm.

One common approach for constructing a core collection is grouping the entire collection according to growing regions or ecotypes, then, selecting representative core accessions from each of the groups to form core subsets, and the entire core collection is constructed by combining all core subsets. The core accessions need to be analyzed on genetic diversity to ensure their representativeness [5]. For a core collection, the representativeness is the most important property. The representativeness evaluation is a significant step in the procedure of core collection construction. A series of evaluating parameters are required in representativeness evaluation for core accessions. Therefore, selecting a series of impactful evaluating parameters is an important aspect of core collection construction research. Only if the validities of evaluating parameters for core collection were confirmed, studies on sampling method and sampling...
percentage can have estimation standards. However, the research is scarce for those existing evaluating parameters.

The key to improve the representativeness of a core collection is the reliable stratification on the initial collection and scientific selection within groups. Only if the representativeness of core subsets from groups is improved, the representativeness of the entire core collection could be substantially improved. A core subset is a swatch of a core collection, and its evaluating parameters are almost the same as in core collection. The present research used rice as a material to assess the evaluating parameters of core collection at the core subset level. Monte Carlo simulation combined with mixed linear model was used in present research on evaluating parameters for core subset based on genotypic values and molecular marker information, which eliminated the interference from environment and achieved more reliable and effective evaluating parameters for core collection evaluation.

MATERIALS AND METHODS

Monte Carlo simulation for germplasm population

Based on the method of Bataillon et al. [6], Monte Carlo simulation was used to establish germplasm population (could be treated as group). Since the representativeness evaluation was only discussed at group level, the simulation process was predigested as follows: supposing that mutation, excursion, migration and selection did not exist, the initial gene pool was initialized with 600 correlative QTLs (quantitative trait loci). Each QTL was evaluated by the random number of normal distribution. Two hundred, 300 and 400 accessions were randomly selected from the initial gene pool. Each accession was self-pollinated for 100 generations to get homozygosity. Therefore, each homozygous accession could be treated as a traditional variety and all the accessions formed a heterogeneous-homozygous initial germplasm group [7] for core subsets construction.

Pleiotropism and multigenic effects were supposed to exist in the initial group. A certain number (10-50 in present research) of QTLs were randomly selected from each accession of the initial group (QTLs in one accession were corresponding to another) to form a quantitative trait. The genotypic value of one trait could be achieved by adding each QTL value of the trait [8]. Repeated random selecting for 30 times led to genotypic values of 30 quantitative traits in the group. It was supposed that molecular markers were completely linked with QTLs, 100 QTLs were randomly selected from each accession (QTLs in one accession were corresponding to another) to form the molecular marker information of the group. Therefore, there were correlative genotypic values of quantitative traits and molecular marker information in the group, which could be used for comprehensive assessment of the genetic diversity of the group. Before constructing core subsets, genotypic values of each trait were standardized (μ = 0, σ = 1; where μ is the population mean of the trait and σ is the standard deviation of the trait).

The evaluating parameters for core collection

Eleven evaluating parameters, including the ones suggested by the former researchers [9-10] and by the present study, were selected. Seven evaluating parameters for a quantitative trait: mean difference percentage (MD), variance difference percentage (VD), changeable rate of maximum (CRMAX), changeable rate of minimum (CRMIN), changeable rate of mean (CRMEA), coincidence rate of range (CR) and variable rate of coefficient of variation (VR). Four evaluating parameters for molecular marker information: percentage of polymorphic loci (p), mean Simpson index (MD), mean Shannon-Weaver index of genetic diversity (MI), and mean polymorphism information content (MPIC). These parameters are formulated as follows:

\[
MD = \left( \frac{S_t}{n} \right) \times 100 \%
\]

Where, \( S_t \) is the number of traits which have significant differences (α=0.05) of means between the initial germplasm group and core subset; \( n \) is total number of traits.

\[
VD = \left( \frac{S_V}{n} \right) \times 100 \%
\]

Where, \( S_V \) is the number of traits which have significant differences (α=0.05) of variances between
the initial germplasm group and core subset, \( n \) is total number of traits.

\[
CR_{\text{MAX}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Max}_{C(i)}}{\text{Max}_{I(i)}} \times 100;
\]

Where, \( \text{Max}_{C(i)} \) is the maximum value of the \( i \)th trait of core subset; \( \text{Max}_{I(i)} \) is the maximum value of the \( i \)th trait of the initial germplasm group; \( n \) is total number of traits.

\[
CR_{\text{MIN}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Min}_{C(i)}}{\text{Min}_{I(i)}} \times 100;
\]

Where, \( \text{Min}_{C(i)} \) is the minimum value of the \( i \)th trait of core subset; \( \text{Min}_{I(i)} \) is the minimum value of the \( i \)th trait of the initial germplasm group; \( n \) is total number of traits.

\[
CR_{\text{MEA}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{Mea}_{C(i)}}{\text{Mea}_{I(i)}} \times 100;
\]

Where, \( \text{Mea}_{C(i)} \) is the mean value of the \( i \)th trait of core subset; \( \text{Mea}_{I(i)} \) is the mean value of the \( i \)th trait of the initial germplasm group; \( n \) is total number of traits.

\[
CR = \frac{1}{n} \sum_{i=1}^{n} \frac{R_{C(i)}}{R_{I(i)}} \times 100;
\]

Where, \( R_{C(i)} \) is the range of the \( i \)th trait of core subset; \( R_{I(i)} \) is the range of the corresponding trait of the initial germplasm group; \( n \) is total number of traits.

\[
VR = \frac{1}{n} \sum_{i=1}^{n} \frac{CV_{C(i)}}{CV_{I(i)}} \times 100;
\]

Where, \( CV_{C(i)} \) is the coefficient of variation of the \( i \)th trait of core subset; \( CV_{I(i)} \) is the coefficient of variation of the corresponding trait of the initial germplasm group; \( n \) is total number of traits.

\[
p = \left( \frac{k}{n} \right) \times 100 \%
\]

Where, \( k \) is the number of polymorphic loci; \( n \) is the total number of molecular marker loci.

\[
M_D = \frac{1}{n} \sum_{i=1}^{n} \left( 1 - \sum_{j=1}^{m} \frac{p_{ij}^2}{1} \right);
\]

Where, \( P_{ij} \) is the frequency of the \( j \)th allele in the \( i \)th locus; \( m \) is the number of alleles in the \( i \)th locus; \( n \) is the total number of molecular marker loci.

\[
M_I = -\frac{1}{n} \sum_{i=1}^{n} \frac{p_{ij} \ln p_{ij}}{2}.
\]

Where, \( P_{ij} \) is the frequency of the \( j \)th allele in the \( i \)th locus; \( m \) is the number of alleles in the \( i \)th locus; \( n \) is the total number of molecular marker loci.

\[
M_{PC} = \frac{1}{n} \sum_{i=1}^{n} \left( 1 - \sum_{j=1}^{m} \frac{p_{ij}^2}{1} - \sum_{j=1}^{m} \frac{2 p_{ij}^2 p_{ij}^j}{2} \right);
\]

Where, \( p_{ki} \) is the frequency of the \( i \)th allele in the \( k \)th locus; \( p_{ji} \) is the frequency of the \( j \)th allele in the \( k \)th locus; \( m \) is the number of alleles in the \( i \)th locus; \( n \) is the total number of molecular marker loci.

All calculations on evaluating parameters were based on core accessions from non-standardized group.

**Constructing method of core subset**

The constructing method of stepwise clusters \([9]\) was modified and used in the present research. The process is: first, a precise sampling percentage of the core subset to the initial germplasm population is given. Next, the genetic distances between accessions are calculated and accessions are grouped by hierarchical cluster analysis based on the genetic distances. One accession from a subgroup with the least distance (this subgroup is unique in the whole dendrogram) is randomly removed and the other accession of the subgroup is sampled. Then, the genetic distances among the rest accessions are calculated again, and the sampling is performed by the same way. The stepwise samplings are performed until the percentage of the rest accessions reaches the given sampling percentage. By this way, the core subset is successfully constructed.

The method of modified stepwise clusters performs sampling based on the group with the least genetic distance, which could efficiently eliminate redundant accessions. Moreover, this method isn’t affected by cluster methods because subgroup with the least distance is unique in each procedure of cluster. Euclidean distance combining single cluster method \([11]\) was used in the present research for core subset construction.

**Core subset construction**

For three simulated germplasm populations (200, 300 and 400 accessions, respectively), 90 core subsets were constructed in each germplasm population by sampling 1% to 30% of accessions with three replications per percentage, and 270 core subsets were achieved from the three populations. In order to evaluate the validity of Monte Carlo simulation, a
worked example of rice was used. The rice population with 90 genotypes (marked from 1 to 90) was used in the present research. All genotypes came from the same growing region and served as a germplasm group to construct core subset. The observed data of eight agronomic quantitative traits with two replications and three years in this group were recorded. Information of 60 molecular markers was collected. Genotypic value of each trait was unbiasedly predicted by the mixed linear model approach\(^{[12]}\). Since the rice group was relatively small, 87 core subsets were constructed from 2% to 30% with three replications per percentage for contrasting to simulated data.

**Data processing**

For all the core subsets (357 total), all the eleven evaluating parameters were calculated and the changing trend of each evaluating parameter was figured. Variance analysis was used to test significance of difference for the same evaluating parameter at different sampling percentages. Tukey’s test (\(\alpha = 0.05\)) was used to perform multiple comparisons and letter-marking method was used to show the comparing results. Number of homogeneous populations of Tukey’s test (For example, according to alphabetical order, if the largest letter was ‘c’, the number of homogeneous populations were 3; if the largest letter was ‘f’, the number of homogeneous populations were 6) was used to assess the validity of each evaluating parameter. Larger number of homogeneous populations meant that more core subsets were distinguished, and the corresponding evaluating parameter was more valid.

The Monte Carlo simulation, core subset construction and calculation of evaluating parameters were conducted in MATLAB. The values of evaluating parameters were analyzed using ANOVA procedure with multiple comparisons method of Tukey in SAS software. The significance level was at \(\alpha=0.05\).

**RESULTS**

**Validity of the eleven evaluating parameters**

At sampling percentages ranging from 1% to 30% or 2% to 30% (Table 1), the number of homogeneous populations of \(CR\) was larger than that of other evaluating parameters whether in simulated germplasm group or in true rice germplasm group. \(MD\), \(VD\) and \(CR_{MEA}\) had smaller number of homogeneous populations (less than 3) than other evaluating parameters in all germplasm groups. In each germplasm group, the ranks of the number of homogeneous populations of \(MD\), \(MI\) and \(MPIC\) were

<table>
<thead>
<tr>
<th>Parameter</th>
<th>200 simulated accessions</th>
<th>300 simulated accessions</th>
<th>400 simulated accessions</th>
<th>90 rice accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of homogeneous</td>
<td>Rank (^{a})</td>
<td>No. of homogeneous</td>
<td>Rank (^{a})</td>
</tr>
<tr>
<td></td>
<td>populations</td>
<td></td>
<td>populations</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>VD</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>CR(_{MAX})</td>
<td>11</td>
<td>3</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>CR(_{MEAN})</td>
<td>17</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>CR(_{MEA})</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>CR</td>
<td>18</td>
<td>1</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>VR</td>
<td>1</td>
<td>7</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>(p)</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>(MD)</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>(MI)</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>(MPIC)</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

\(^{a}\) The order of the number of homogeneous populations for each evaluating parameter.
similar and the numbers of homogenous populations were relatively large. \( CR_{\text{MAX}} \) had large number of homogenous populations (not less than 8) in three simulated germplasm groups while had 5 homogenous populations in rice germplasm group. \( CR_{\text{MIN}} \) had the second large number of homogenous populations in the simulated germplasm group with 200 accessions and rice germplasm group while had the fourth one in the simulated germplasm group with 300 accessions. The number of homogenous populations of \( VR \) was quite large in simulated germplasm groups with 300 accessions (6 homogeneous populations) and 400 accessions (14 homogeneous populations) whereas it was very small in the simulated germplasm groups with 200 accessions (1 homogeneous populations) and rice germplasm group (2 homogeneous populations). \( p \) had 5 homogeneous populations in the rice germplasm group while had 3 ones in the three simulated germplasm groups respectively.

At sampling percentage of 10% to 30% (Table 2), \( CR \) ranked the first in the number of homogeneous populations in the simulated germplasm group with 200 accessions and rice germplasm group, and the second in the simulated germplasm groups with 300 accessions and 400 accessions. The ranks in the number of homogeneous populations of \( MD, VD, CR_{\text{MEA}} \) and \( p \) were lower than those in other evaluating parameters. The ranks in the number of homogeneous populations of \( MD, M_I, M_{PIC} \) were similar to those in Table 1 and \( MD \) had the largest number of homogeneous populations in the simulated germplasm group with 300 accessions. \( CR_{\text{MAX}} \) had large number of homogeneous populations but not stable among groups. \( CR_{\text{MIN}} \) had small number of homogeneous populations in the simulated germplasm group with 300 accessions (2 homogeneous populations) while had similar rank in number of homogeneous populations to \( CR_{\text{MAX}} \) in other three germplasm groups. \( VR \) had quite large number of homogeneous populations in the simulated germplasm groups with 300 accessions and 400 accessions while had just 1 homogeneous population in the simulated germplasm group with 200 accessions and rice germplasm group.

Synthesizing the results above, the order of the validity of the eleven evaluating parameters was: \( CR > M_D, M_I, M_{PIC} > CR_{\text{MAX}}, CR_{\text{MIN}} > VR > p > MD, VD, CR_{\text{MEA}} \).

**The stability of the eleven evaluating parameters**

Since \( MD, VD \) and \( CR_{\text{MEA}} \) had poor validity, their stability would not be discussed any more. \( CR \) was increased steadily with the increase of sampling percentage within germplasm group and the changing rate was almost same among germplasm groups (Fig. 1). \( M_I \) had similar changing rules to \( CR \) (Fig. 2). The changing trends of \( MD \) and \( M_{PIC} \) were almost the same.

| Table 2. Number of homogeneous populations of Tukey’s test (\( \alpha = 0.05 \)) of the eleven evaluating parameters in different germplasm populations at the sampling percentages of 10% to 30%. |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Parameter       | 200 simulated accessions | 300 simulated accessions | 400 simulated accessions | 90 rice accessions |
| No. of homogeneous populations | Rank* | No. of homogeneous populations | Rank* | No. of homogeneous populations | Rank* | No. of homogeneous populations | Rank* |
| \( MD \)         | 1 | 5 | 1 | 7 | 1 | 7 | 1 | 4 |
| \( VD \)         | 1 | 5 | 1 | 7 | 2 | 6 | 1 | 3 |
| \( CR_{\text{MAX}} \) | 8 | 3 | 5 | 4 | 8 | 3 | 5 | 2 |
| \( CR_{\text{MIN}} \) | 9 | 2 | 2 | 6 | 5 | 4 | 2 | 3 |
| \( CR_{\text{MEA}} \) | 1 | 5 | 3 | 5 | 3 | 5 | 1 | 4 |
| \( CR \)         | 11 | 1 | 8 | 2 | 8 | 2 | 6 | 1 |
| \( VR \)         | 1 | 5 | 5 | 4 | 12 | 1 | 1 | 4 |
| \( p \)          | 1 | 5 | 1 | 7 | 1 | 7 | 1 | 4 |
| \( M_D \)        | 6 | 4 | 10 | 1 | 8 | 3 | 2 | 3 |
| \( M_I \)        | 6 | 4 | 9 | 3 | 8 | 3 | 2 | 3 |
| \( M_{PIC} \)    | 6 | 4 | 9 | 3 | 8 | 3 | 2 | 3 |

* The order of the number of homogeneous populations for each evaluating parameter.
to $M_I$ in figure (Figures not shown), therefore, $M_I$ was discussed as the representative of $M_D$, $M_I$ and $M_PIC$. $CR$, $M_D$, $M_I$ and $M_PIC$ were considered to be stable among and within germplasm groups. $CR_{MAX}$ and $CR_{MIN}$ were increased with the increase in sampling percentage within germplasm groups, whereas the changing rate of $CR_{MAX}$ in the simulated germplasm group with 400 accessions was significant larger than those in the other three germplasm groups (Fig. 3), and $CR_{MIN}$ had different changing rates among germplasm
Therefore, CRMAX and CRMIN acted stably within germplasm group but unstably among germplasm groups. p reached 100% at low sampling percentage then kept stable within germplasm group and its different changing rates among germplasm groups just existed at low sampling percentage (Fig. 5). So p acted stably within germplasm group and had better stability than CRMAX and CRMIN among germplasm groups. With sampling percentage increasing, the value of VR waved largely (Fig. 6). Moreover, the changing rate of VR was significantly different among germplasm groups. Therefore, it acted unstably within and among germplasm groups. Synthesizing the results above, the order of the stability of the eight evaluating parameters was: CR, Ml, Ml, Ml, CRMIN, CRMAX, M, Ml, Ml, and p (Figs. 1 to 5). In each germplasm group, CRMAX and CRMIN had relative high changing rates while Ml just had high changing rate at low sampling percentage (Figs. 2 to 4). The value of p reached 100% when the sampling percentage was over 5% in the three simulated germplasm groups, and over 9% in the rice germplasm group (Fig. 5). This suggested that all the core subsets had almost the same p value when the sampling percentage was large.

Therefore, the order of sensibility of the seven evaluating parameters was: CR > CRMAX, CRMIN, Ml, Ml, Ml, Ml, p.

**DISCUSSION**

For the research of constructing core collection, phenotypic values are mainly used, and the representativeness evaluation are mainly based on phenotypic values \(^{[13-14]}\). Constructing core collection
based on genotypic values has just been reported in cotton \cite{9, 11} and rice \cite{15}. Most traits of germplasm materials are quantitative traits under the control of polygenes, which means that they are easily affected by environmental conditions and experimental errors. Moreover, effects of interaction between gene and environment (GE effects) exist in phenotypic values \cite{9}. Therefore, stratification or evaluation based on phenotypic values may be not accurate \cite{1}. In the present research, the core subset construction and evaluation was based on genotypic values, which eliminate the effects of experimental errors, environmental effects and GE effects to draw more accurate results. For the three simulated germplasm groups, Monte Carlo simulation was used to achieve genotypic values. For the rice germplasm group, mixed linear model approach was used to unbiasedly predict genotypic values. The action of the eleven evaluating parameters acted similarly in the simulated germplasm groups and rice germplasm group, which showed that the simulating method adopted in the present research was reliable. The core subset constructing method used in the present research was based on random selection. Therefore, although the number of accessions among core subsets had not much difference (1% sampling interval), core accessions among subsets had large difference. The small sampling interval had little interference to the final results.

An available core collection evaluating parameter should be able to distinguish different core collections effectively and act stably among different germplasm groups. Therefore, validity and stability are primary for evaluating parameters. Sensibility is a reference index because it implies the acute degree of the evaluating parameter’s value change when accessions changing within germplasm group. Synthesizing the orders of all evaluating parameters, \( CR \), which had the highest validity, stability and sensibility, could be the optimal evaluating parameter. \( M_D, M_I \) and \( M_{PIC} \) ranked the second in the validity, with slightly lower sensibility but higher stability than \( CR_{MAX} \) and \( CR_{MIN} \). Therefore, \( M_D, M_I \) and \( M_{PIC} \) are important evaluating parameters for core collection. \( CR_{MAX} \) and \( CR_{MIN} \), which had similar validity to \( M_D, M_I \) and \( M_{PIC} \), and high sensibility, also could be used in core collection evaluation. Essentially, \( CR \) synthesized the advantage of \( CR_{MAX} \) and \( CR_{MIN} \), which led it to be more appropriate for core collection evaluation. \( VR \) has lower validity than \( CR_{MAX} \) and \( CR_{MIN} \), and acted unstable among germplasm populations. However, the value of \( VR \) waved much with sampling percentage increasing. Therefore, \( VR \) could be as an important referential parameter for evaluating the variation degree of core collection. \( p \) had relatively high stability but low validity and sensibility. For a core collection with large size, \( p \) was easy to reach 100% when the number of molecular markers was large. Therefore, the application of \( p \) is restricted by germplasm population size and molecular marker number. However, it could act as a determination parameter for the size of core collection because it had high changing rate at a certain sampling percentage. The constructing method used in the present research behaved well in maintaining mean and variance of a population. Therefore, most core
subsets constructed in the present research had similar mean and variance to initial population, which led to low validities of $MD$, $VD$ and $CR_{MEA}$. However, much research has reported that $MD$ is an important parameter for determining whether a core collection is acceptable$^{[13, 16]}$. Generally, only if the $MD$ of a core collection was not over 20%, the core collection would be able to well represent the genetic diversity of the initial population$^{[9, 15]}$. Therefore, $MD$ could be used as a determination parameter for the reliability judgement of core collection.

The main aim of core collection research is to find out effective methods to conserve maximum genetic diversity by minimum accessions$^{[2, 17]}$. Basigalup et al$^{[18]}$ suggested that larger range and variance led to more representative core collection. Hu et al$^{[9]}$ treated $CR$ as an important evaluating parameter and applied in cotton core collection construction. The present research further proved the advantage of $CR$. Evaluating parameter based on population mean was found to be of low validity, which was accordant to the research results of Zhang et al$^{[10]}$. The information used in core collection construction is very limited at present. The genetic diversity of a collection can be reflected perfectly only if various kinds of traits have been evaluated compositively, such as morphology, agronomic traits, biochemical characters, molecular markers, and so on$^{[19]}$. However, Various kinds of traits of initial population usually include not only continuous data, but also discrete data. For the two kinds of data, different kinds of evaluating parameters are required. For quantitative traits, mean and variance can be used to classify continuous data and change them to discrete data$^{[20-21]}$.

It is an important work to confirm reasonable sampling percentage in core collection construction. Brown$^{[17]}$ suggested that core collection with 10% sampling percentage could represent 70% genetic diversity of the initial population when the number of the initial accessions was over 3000. Upadhyaya and Ortiz$^{[13]}$ constructed a chickpea core collection with 1956 core accessions from 16 991 initial accessions with 10% sampling percentage. Zewdie et al$^{[22]}$ also used 10% sampling percentage in sorghum core collection construction. Wang et al$^{[23]}$ constructed a Changjiang spring sowing soybean with just 8.58% sampling percentage (2 148 initial accessions). Xu et al$^{[11]}$ constructed a cotton core collection from 168 initial accessions with 30% sampling percentage. Generally, small sampling percentage is suitable for large size initial population, while large one is suitable for small size initial population. However, confirming reasonable sampling percentage should be based on characteristics of different germplasm populations. Although the germplasms are different in species, the evaluating parameters for core collection may be the same. A series of evaluating parameters selected by the present research is able to be applied in different germplasm populations, which can be used as criteria for sampling percentage. Therefore, solving the problem of selecting evaluating parameters leads to solving the problem of reasonable sampling percentage in great degree. For example, when $MD$ is 20% and $CR$ is 80%, the corresponding sampling percentage could be considered to be a reasonable percentage. If a more reliable core collection is required, the criterion mentioned above should be modified to be more strict and also, other evaluating parameters should be combined for evaluation.

ACKNOWLEDGEMENTS

This study was financially supported by the National Natural Science Foundation of China (Grant No. 30270759) and the Science and Technology Department of Zhejiang Province (Grant No. 2005C32001).

REFERENCES


