Theor Appl Genet (2004) 108:1172–1176
DOI 10.1007/s00122-003-1536-1

ORIGINAL PAPER

C. T. Li · C. H. Shi · J. G. Wu · H. M. Xu ·
H. Z. Zhang · Y. L. Ren

Methods of developing core collections based on the predicted genotypic value of rice (*Oryza sativa* L.)

Received: 16 May 2003 / Accepted: 10 November 2003 / Published online: 28 January 2004

Abstract The selection of an appropriate sampling strategy and a clustering method is important in the construction of core collections based on predicted genotypic values in order to retain the greatest degree of genetic diversity of the initial collection. In this study, methods of developing rice core collections were evaluated based on the predicted genotypic values for 992 rice varieties with 13 quantitative traits. The genotypic values of the traits were predicted by the adjusted unbiased prediction (AUP) method. Based on the predicted genotypic values, Mahalanobis distances were calculated and employed to measure the genetic similarities among the rice varieties. Six hierarchical clustering methods, including the single linkage, median linkage, centroid, unweighted pair-group average, weighted pair-group average and flexible-beta methods, were combined with random, preferred and deviation sampling to develop 18 core collections of rice germplasm. The results show that the deviation sampling strategy in combination with the unweighted pair-group average method of hierarchical clustering retains the greatest degree of genetic diversities of the initial collection. The core collections sampled using predicted genotypic values had more genetic diversity than those based on phenotypic values.

Introduction

With the rapid increase in the number of accessions contained in crop germplasm collections, redundant resources have become an obstacle to the effective maintenance and utilization of these collections. To solve this problem, Frankel (1984) proposed the concept of the core collection. The design of the core collection should minimize repetitiveness within the collection and should represent the genetic diversity of a crop species and its relatives. The core collection could serve as a working collection that could be extensively examined, and the accessions excluded from the collection would be retained as the reserve collection (Frankel and Brown 1984a, 1984b; Brown 1989a, 1989b). Thus, The establishment of core collections is a helpful means by which to make better use of plant germplasm and to assist in the management of the entire collection.

Dozens of core collections have been successfully developed by various criteria and sampling strategies. Some examples are those for perennial *Glycine* spp., peanut (*Arachis hypogae* L.), annual *Medicago* spp. and sugarcane (*Saccharum spontaneum* L.) (Brown et al. 1987; Holbrook et al. 1993; Diwan et al. 1995; Tai et al. 2001). The kinds of data that have been used to analyze the genetic diversity of crops include morphological, agronomic and ecogeographical traits as well as molecular and biochemical markers (Perry et al. 1991; Joe and Orlando, 1996; Lerceteau et al. 1997; Hokanson et al. 1998; Ortiz et al. 1998; Huaman et al. 1999; Parsons et al. 1999; Chavarriaga-Aguirre et al. 1999; Marita et al. 2000; Upadhyaya and Ortiz 2001; Chandra et al. 2002). To date, most core collections have been constructed using phenotypic values. Most traits of crop varieties are quantitative traits under the control of multiple genes that are easily influenced by environmental conditions (e.g. weather, soil, cultural and field management). Hence, genetic grouping based on phenotypic data may not reflect the true genetic diversity of the initial germplasm resources depending on the degree of environmental effects. The core collections constructed solely on the basis of phenotypic data might not perfectly represent the genetic diversities of the original crop collections (Tanksley and McCouch 1997).

In the investigation reported here, various methods were proposed to construct 18 core collections of rice by three sampling strategies and hierarchical clustering with

Communicated by D.J. Mackill

C. T. Li · C. H. Shi (✉) · J. G. Wu · H. M. Xu · H. Z. Zhang · Y. L. Ren
Department of Agronomy,
College of Agriculture and Biotechnology,
Zheijang University,
310029 Hangzhou, China
e-mail: chhshi@zju.edu.cn

six amalgamation rules based on the predicted genotypic values of 13 traits. Sampling strategies and cluster methods were evaluated to find the best method for determining core collections of rice. These methods were also compared to determine the effect of using phenotypic values instead of predicted genotypic values on the design of core collections.

Materials and methods

Materials

The seeds of 992 varieties of rice germplasm were sown on May 25, 2001, and 27-day-old seedlings were individually transplanted at 20×20-cm spacing at the experiment farm of Zhejiang University. Twenty-four seedlings of each variety were planted in plots with two replications. Plant samples for each variety were derived at maturity from eight plants in the middle of the plot. The phenotypic values of 13 traits were studied: days to heading, plant height (centimeters), number of panicles per plant, flag leaf length (centimeters), flag leaf width (centimeters), ratio of the flag leaf length to width, panicle length (centimeters), number of seeds per panicle, number of fertile seeds per panicle, seed setting rate (percentage), grain weight (milligrams), panicle weight per plant (grams) and yield per plant (grams). The mean values from these plant samples served as the original data for the 13 traits.

Statistical models

In the single environment experiment with two replications, the observed values could be expressed as$y_{ij} = \mu + G_i + e_{ij}$, where y_{ij} is the jth replication's observed values of the ith variety; μ is the population mean; G_i is the genotypic effect of the ith variety, which was predicted by the adjusted unbiased prediction (AUP) method with phenotypic values (Zhu 1993; Zhu and Weir 1996), $G_i \sim \left(0, \sigma_G^2\right)$, i=1, 2, 3, ..., g; e_{ij} is the residual effect, $e_{ij} \sim \left(0, \sigma_e^2\right)$. On the basis of the predicted genotypic values, the genetic distances among different varieties could be calculated and used for classifying the accessions.

Cluster analysis and sampling strategies

Mahalanobis distance (Mahalanobis 1936) can eliminate the scalar difference between traits and account for correlations among traits (Hu et al. 2000). It was, therefore, applied to calculate the genetic distance among the different accessions. The single linkage, median linkage, centroid, unweighted pair-group average, weighted pair-group average and flexiable-beta methods were used for grouping the accessions. Based on the clustering in the dendrograms, random, preferred and deviation samplings were employed to construct core collections (Hu et al. 2000), and the sample size was 15% (Perry et al. 1991).

Evaluation of the core collection

A homogeneity test (F-test) for variance and a t-test for means (α=0.05) were applied to test for significant differences between the means of each trait for the core collection and the initial collection. The following parameters were used to evaluate the properties of the core collections in terms of the initial collection: VD% [percentage of significant difference (α=0.05) between core collection and the initial collection for variances of traits], MD% [percentage of significant difference (α=0.05) between core collection and the initial collection for means of traits], CR% (the coincidence rate, $CR\% = \frac{1}{m}\sum_{j=1}^{m}\frac{R_C}{R_I} \times 100$) and VR% (the variable rate, $VR\% = \frac{1}{m}\sum_{j=1}^{m}\frac{CV_C}{CV_I} \times 100$), where R_C is the range of each trait for the core collection, R_I is the range for initial collection, CV_C is the coefficient of variation of trait for the core collection, CV_I is the coefficient of variation of trait for the initial collection and m is the number of traits.

Core collections were considered to well represent the genetic diversity of the initial collection if the two following criteria were met: (1) no more than 20% of the traits had different means (significant at α=0.05) between the core collection and the initial collection and (2) CR% was retained by the core collection in no less than 80% of the traits (Hu et al. 2000).

Results

Phenotypic means and predicted genotypic values for 13 traits of rice

The phenotypic means, predicted genotypic values and their ranges for 13 traits of rice are listed in Table 1. The results show that there was a large variation among materials studied. The genotypic values predicted by the AUP method with phenotypic means also showed abundant genetic diversity among these rice varieties, but the ranges were smaller. Consequently, the initial germplasm collections of rice could be used to construct core collections, and the genetic distances among different

Table 1 Phenotypic means and predicted genotypic values for 13 traits in rice (*SE* standard error)

Traits	Phenotypic means		Predicted genotypic value	
	Mean ± SE	Range	Mean ± SE	Range
Days to heading time	72.09±10.78	57.00~116.00	72.08±10.49	57.41~114.81
Plant height (cm)	86.16±12.78	54.25~153.11	86.16±12.42	55.17~151.19
Number of panicles per plant	7.39±1.56	3.23~20.00	7.39±1.27	4.00~17.67
Flag leaf length (cm)	29.89±5.60	18.02~54.98	29.89±5.16	18.95~53.00
Flag leaf width (cm)	1.51±0.21	0.79~2.31	1.51±0.20	0.83~2.26
Flag leaf length/width ratio	20.08±3.83	11.57~37.02	20.08±3.48	12.35~35.47
Panicle length (cm)	20.28±2.11	11.78~27.76	20.28±1.97	12.35~27.26
Number of seeds per panicle	104.59±29.42	44.82~245.15	104.59±27.63	48.44~236.62
Number of fertile seeds per panicle	68.92±21.06	17.11~156.62	68.92±18.80	22.69~147.19
Seed setting rate (%)	65.99±9.97	13.38~91.07	65.99±7.83	24.65~85.70
Grain weight (mg)	21.65±3.25	11.57~44.73	21.65±3.16	11.84~44.10
Panicle weight per plant (g)	12.52±3.03	3.48~30.50	12.52±2.42	5.29~26.90
Yield per plant (g)	10.49±2.74	2.33~25.92	10.49±2.16	4.05~22.66

Table 2 Comparison of the percentages for the differences between the core collection and the initial collection in rice

Core collections	VD%[a]	MD%[b]	CR%[c]	VR%[d]	Number of entries
GCoreC1S1	100	38.5	96.0	134.4	133
GCoreC2S1	46.2	0	88.3	112.2	119
GCoreC3S1	23.1	7.7	77.8	108.8	145
GCoreC4S1	69.2	0	86.2	112.8	129
GCoreC5S1	30.8	0	79.6	112.7	112
GCoreC6S1	76.9	0	87.4	117.0	114
GCoreC1S2	100	46.2	100	133.6	146
GCoreC2S2	84.6	0	100	120.0	134
GCoreC3S2	84.6	7.7	100	120.1	125
GCoreC4S2	92.3	0	100	120.2	137
GCoreC5S2	84.6	15.4	100	120.5	119
GCoreC6S2	84.6	0	100	119.7	120
GCoreC1S3	100	38.5	98.1	140.1	137
GCoreC2S3	100	0	93.1	127.4	130
GCoreC3S3	100	15.4	95.3	131.1	130
GCoreC4S3	100	0	94.7	131.5	137
GCoreC5S3	100	15.4	92.2	129.5	133
GCoreC6S3	100	7.7	94.2	131.1	123

[a] VD%, Percentage of significant difference (α=0.05) between core collection and the initial collection for variance of traits

[b] MD%, Percentage of significant difference (α=0.05) between core collection and the initial collection for means of traits

[c] CR%, Coincidence rate

[d] VR%, Variable rate

varieties could be calculated for classifying the accessions based on the predicted genotypic values.

Construction of 18 core collections

Eighteen core collections were constructed by six clustering methods, including the single linkage method (C1), the median linkage method (C2), the centroid method (C3), the unweighted pair-group average method (C4), the weighted pair-group average method (C5) and the flexible-beta method (C6), combined with random sampling (S1), preferred sampling (S2) and deviation sampling (S3) of a 15% sample size. The 18 core collections were named using these descriptors (Table 2). Each of these core collections was established based on the genotypic values predicted by the AUP method from the mean phenotypic values.

The results show that there was no significant difference (α=0.05) in the means of all traits between 15 core collections and the initial collections and that there were 16 core collections with CR% larger than 80%. When we considered all those that met both conditions, 13 core collections remained, and these 13 are useful in determining appropriate rules for building core collections.

Evaluation of the core collection with 13 traits

The effects of different cluster methods under the condition of the same sampling strategy are listed in Table 2. When the random sampling strategy was employed for the six core collections, three cores did not satisfy the established rules: the MD% of GCoreC1S1 had significant differences (MD% >20%) compared with the initial collections, while GCoreC3S1 and GCoreC5S1 had small CR% (less than 80%). When the remaining three core collections were compared, GCoreC6S1 was found to have the highest VD% and VR% and the same or similar MD% and CR%. Consequently, the flexible-beta method (C6) was the best clustering strategy when random sampling was employed.

Six core collections were developed using the preferred sampling strategy combined with six cluster methods. Each of these six core collections had a CR% of 100%. The MD% of GCoreC1S2 showed significant differences (MD% >20%) compared with the initial collections, so GCoreC1S2 was excluded. With respect to the remaining five core collections, GCoreC4S2 had the highest VD% and VR% and the lowest MD%. Therefore, when the sampling strategy was preferred sampling, the best way of clustering was the unweighted pair-group average method (C4).

The last set of six core collections was constructed by combining the deviation sampling strategy with the different clustering methods. GCoreC1S3 did not show the diversity of the initial collection since its MD% was more than 20% (38.5%). Among the other five core collections, GCoreC4S3 had the highest VD%, CR% and VR% and the lowest MD%. Consequently, the unweighted pair-group average method (C4) would be the best clustering method if deviation sampling was used to develop the core collections.

The six hierarchical cluster methods applied in this study have their own characteristics when combined with these sampling strategies. The single linkage method (C1) always had the highest MD% regardless of the sampling strategy it was combined with. The MD% of the weighted pair-group average method (C5) was 15.4% when combined with preferred sampling and deviation sampling,

Table 3 Comparison between core collections based on genotypic and phenotypic values of 13 traits in rice

Statistical parameter[a]	PcoreC6S1	GCoreC6S1	PcoreC4S2	GCoreC4S2	PCoreC4S3	GCoreC4S3
VD%	69.2	76.9	84.6	92.3	100	100
MD%	0	0	0	0	0	0
CR%	83.5	87.4	100	100	93.8	94.7
VR%	115.4	117.0	119.2	120.2	130.7	131.5
Number of entries	106	114	135	137	130	137

[a] VD%, Percentage of significant difference (α=0.05) between core collection and the initial collection for variance of traits; MD%, percentage of significant difference (α=0.05) between core collection and the initial collection for means of traits; CR%, coincidence rate; VR%, variable rate

while its CR% would be less than 80% when combined with random sampling. As to the centroid method (C3), its MD% was markedly larger than those of the other three clustering methods. It can then be concluded that these three cluster methods (C1, C3 and C5) were not suitable to construct core collections of rice. The unweighted pair-group average method (C4) always had the lowest MD% and relatively high VD%, CR% and VR% when combined with all three sampling strategies. Hence, it would be the best clustering method to construct a core collection of rice because of this stability.

The results also suggest that better representations of the initial collection could be obtained for GCoreC6S1 by random sampling, for GCoreC4S2 by preferred sampling and for GCoreC4S3 by deviation sampling. If we consider the parameters of the 18 core collections overall, the GCoreC4S3 had the highest VD% and VR%, a relatively high CR%, and the lowest MD%; therefore, it would be the optimal choice on the basis of the 18 core collections. The deviation sampling strategy combined with the unweighted pair-group average method of hierarchical clustering could retain the greatest degree of genetic diversity of the initial rice collection.

Comparison between core collections based on genotypic and phenotypic values of 13 traits in rice

The best clustering methods and sampling strategy combinations were selected to compare differences between core collections based on genotypic values and those based on phenotypic values. The core collections based on phenotypes (PcoreC6S1, PcoreC4S2 and PcoreC4S3) were constructed using the same approaches and sample size as the core collections derived from genotypes (GCoreC6S1, GCoreC4S2 and GCoreC4S3) (Table 3).

The results show that in PcoreC6S1, PcoreC4S2 and PcoreC4S3 MD% were all smaller than 20% and CR% were larger than 80%. The core collections based on phenotypic values also retained the genetic diversity of the initial collections. However, the VD%, CR% and VR% of PcoreC6S1, PcoreC4S2 and PcoreC4S3 were less than or equal to those of GCoreC6S1, GCoreC4S2 and GCoreC4S3. Therefore, with respect to rice germplasm, the core collections based on predicted genotypic values better represented the genetic diversity than the core collections based on phenotypic values.

Discussion

Many core collections were successfully developed after Frankel proposed the theory of the core collection in 1984. The criteria by which core collections are established have intrinsic advantages and disadvantages for evaluating the genetic diversity of an entire collection. While molecular markers have been used to assess genetic diversity at the DNA level in crop collections (Lerceteau et al. 1997), studies at this level applied to entire collections using molecular or other biochemical analysis would be laborious and costly. Morphological, agronomic and ecogeographical data based on the phenotypic values have generally been used to develop core collections (Holbrook et al. 1993; Diwan et al. 1994; Huaman et al. 1999). Because of the environmental effects, the same phenotype might be achieved by different genotypes (Singh et al. 1991), and some accessions having different genotypes might be excluded from the core collections. Therefore, the core collections based on the phenotypic values are not the perfect representation of the initial collection. It is necessary to employ suitable genetic models that can be used to reduce the environment effects and experimental errors and allow the prediction of genotypic values of the crop traits. Core collections based on the predicted genotypic values would be better than those based on phenotypic values because the environmental effects have been reduced.

In crop germplasm collections, the genetic resources are usually not equably collected. Some varieties are usually overrepresented while others are deficient, and this leads to an imbalance in the genetic diversity. Hence, a reasonable sampling strategy is required that would reduce these imbalances without bias. In general, the deviation sampling strategy could select accessions with a larger value of standard bias degree for traits $\left(s_i^2\right)$, and the variances and the coefficient of variation in the core collection should be larger. The core collections constructed by this sampling strategy always had the highest VD% and VR% regardless of the clustering method it was combined with. The core collections that were constructed with deviation sampling strategy were able to retain a larger genetic variability of the initial collection.

As a result, the deviation sampling strategy, combined with the unweighted pair-group average method of hierarchical clustering, might be the best way to retain the genetic diversity of the initial collections. This sampling strategy and the clustering method proposed in this experiment based on the predicted genotypic values may be used for other crops to construct core collection(s).

Acknowledgements The project was supported by the Science and Technology Office of Zhejiang Province (grant no. 011102471) and the 151 Foundation for the Talents of Zhejiang Province.

References

Brown AHD (1989a) Core collection: a practical approach to genetic resources management. Genome 31:818–824

Brown AHD (1989b) The case for core collections. In: Brown AHD, Frankel OH, Marshall RD, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 136–156

Brown AHD, Grace JP, Speer SS (1987) Designation of a core collection of perennial *Glycine*. Soybean Genet Newsl 14:59–67

Chandra S, Huaman Z, Hari Krishna S, Ortiz R (2002) Optimal sampling strategy and core collection size of Andean tetraploid potato based on isozyme data—a simulation study. Theor Appl Genet 104:1325–1334

Chavarriaga-Aguirre P, Maya MM, Tohme J, Duque MC, Iglesias C, Bonierbale M W, Kresovich S, Kochert G (1999) Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. Mol Breed 5:263–273

Diwan N, Bauchan GR, McIntosh MSA (1994) Core collection for the United States annual *Medicago* germplasm collection. Crop Sci 34:279–285

Diwan N, Mclntosh MS, Bauchan GR (1995) Methods of developing a core collection of annual medicago species. Theor Appl Genet 90:755–761

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber WK, Llimensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170

Frankel OH, Brown AHD (1984a) Current plant genetic resources—a critical appraisal. In: Genetics: new frontiers, vol 4. Oxford and IBH Publ, New Delhi, India, pp 1–11

Frankel OH, Brown AHD (1984b) Plant genetic resources today: a critical appraisal. In: Hoden HW, Williams JT (eds) Crop genetic resources: conservation and evaluation. George Allen and Urwin, London, pp 249–257

Hokanson SC, Szewc-McFadden AK, Lamboy WF, McFerson JR (1998) Microsatellite (SSR) markers reveal genetic identities, genetic diversity and relationships in a *Malus* × *domestica* borkh. core subset collection. Theor Appl Genet 97:671–683

Holbrook CC, Anderson W F, Pittman R N (1993) Selection of a core collection from the U.S. germplasm collection of peanuts. Crop Sci 33:859–861

Hu J, Zhu J, Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theor Appl Genet 101:264–268

Huaman Z, Aguilar C, Ortiz R (1999) Selecting a Peruvian sweetpotato core collection on the basis of morphological, ecogeographical, and disease and pest reaction data. Theor Appl Genet 98:840–844

Joe T, Orlando GD (1996) AFLP analysis of gene pools of a wild bean core collection. Crop Sci 36:1375–1384

Lerceteau E, Robert T, Petiard V, Crouzillat D (1997) Evaluation of the extent of genetic variability among *Theobroma cacao* accessions using RAPD and RFLP markers. Theor Appl Genet 95:10–19

Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55

Marita JM, Rodriguez JM, Nienhuis J (2000) Development of an algorithm identifying maximally diverse core collections. Genetic Resour Crop Evol 47: 515–526

Ortiz R, Ruiz-Tapia EN, Mujica-Sanchez A (1998) Sampling strategy for a core collection of *Peruvian quinoa* germplasm. Theor Appl Genet 96:475–483

Parsons BJ, Newbury HJ, Jackson MT, Ford-Lloyd BV (1999) The genetic structure and conservation of aus, aman and boro rices from Bangladesh. Genet Resour Crop Evol 46:587–598

Perry MC, Mclntosh MS, Stoner AK (1991) Geographical patterns of variation in the USDA soybean germplasm collection: II. allozyme frequencies. Crop Sci 31:1356–1360

Singh SP, Nodari R, Gepts P (1991) Genetic diversity in cultivated common bean. I. Allozymes. Crop Sci 31:19–23

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

Tai PYP, Miller JD (2001) A core collection for *Saccharum spontaneum* L. from the world collection of sugarcane. Crop Sci 41:879–885

Upadhyaya HD, Ortiz R (2001) A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theor Appl Genet 102:1292–1298

Zhu J (1993) Methods of prediction genotype value and heterosis for offspring of hybrids. J Biomath 8: 32–44

Zhu J, Weir BS (1996) Diallel analysis for sex-linked and maternal effects. Theor Appl Genet 92:1–9