Theor Appl Genet (2000) 101:264–268

J. Hu · J. Zhu · H.M. Xu

Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops

Received: 15 October 1999 / Accepted: 24 November 1999

Abstract A genetic model with genotype×environment (GE) interactions for controlling systematical errors in the field can be used for predicting genotypic values by an adjusted unbiased prediction (AUP) method. Mahalanobis distance, calculated based on the genotypic values, is then applied to measure the genetic distance among accessions. The unweighted pair-group average, Ward's and the complete linkage methods of hierarchical clustering combined with three sampling strategies are proposed to construct core collections in a procedure of stepwise clustering. A homogeneous test and *t*-tests are suggested for use in testing variances and means, respectively. The coincidence rate (*CR*%) for range and the variable rate (*VR*%) for the coefficient of variation are designed to evaluate the property of core collections. A worked example of constructing core collections in cotton with 21 traits was conducted. Random sampling can represent the genetic diversity structure of the initial collection. Preferred sampling can keep the accessions with special or valuable characteristics in the initial collection. Deviation sampling can retain the larger genetic variability of the initial collection. For better representation of the core collection, cluster methods should be combined with different sampling strategies. The core collections based on genotypic values retained larger genetic variability and had superior representatives than those based on phenotypic values.

Key words Core collection · Stepwise cluster · Sampling strategy · Genotypic value

Communicated by P.M.A. Tigerstedt

J. Hu · J. Zhu (✉) · H.M. Xu
Department of Agronomy, Zhejiang University, Hangzhou, 310029, China
e-mail: jzhu@zju.edu.cn
Tel.: +86-571-6971444, Fax: +86-571-6049815

Introduction

Since the concept of the core collection was proposed (Frankel and Brown 1984a, b; Brown 1989), investigations on developing core collections have increased and have been more extensively involved in different aspects for sampling strategy, core size, etc. (Erskine and Muehlbauer 1991; Perry et al. 1991; Zeuli and Qualset 1993; Hintum 1994, 1995; Basigalup et al. 1995; Diwan et al. 1995). Various data have been used to analyze the genetic diversity in crops, including morphological, agronomic and ecogeographical traits or molecular and biochemical markers. Each of these criteria has its advantages and disadvantages for measuring genetic diversity. Molecular markers such as RAPDs and RFLPs can reflect direct changes at the DNA sequence level. A disadvantage of molecular-, isozyme- and seed protein-markers is their bulk and the cumbersome work involved; it is obviously unrealistic to subject an entire collection, or even a large fraction of it, to molecular and biochemical analysis (Gepts 1995) unless simple methods are found. Phenotypic values have often been used to select collections (Holbrook et al. 1993; Diwan et al. 1994). Most traits of crop varieties are quantitative traits that are affected by environmental errors in the field and also by GE interaction. Therefore, genetic sorting based on phenotypic data can not correctly reflect the genetic diversity of the initial germplasm resources. The same phenotype can be achieved by different genotypes. Accessions with a similar phenotype may sometimes be evolutionarily unrelated (Singh et al. 1991). If genotypic values can be predicted based on phenotypic values, then genetic distance based on genotypic values among accessions can be measured more accurately. A core collection constructed by genotypic values will be more representative of the initial collection.

In the present study, genetic models with GE interaction for controlling systematical field errors were used to predict genotypic values by an adjusted unbiased prediction (AUP) method (Zhu 1993; Zhu and Weir 1996). Methods are proposed for constructing core collections

using stepwise clustering combined with three sampling strategies based on the genotypic values. Methods for evaluating representatives of core collections are also proposed. Cotton data with 21 traits were analyzed for constructing core collections. These analyses are used as demonstrations for evaluating sampling strategies and cluster methods, and also for comparing core collections based on phenotypic and genotypic values.

Models and Analysis methods

Genetic models

Field experiments are usually used to control variation due to cultural management, fertility trends or other environmental factors. In general, blocking is employed to control field variation by arranging plots in appropriate ways. When a large number of accessions is compared in a field trial to evaluate germplasm resources, genetic materials can be planted either in random or in an order based on rows and columns of field without blocks, and the same control can be planted, as a check, among genetic materials at certain intervals.

When genetic experiments are conducted for several environments, with at least two replicates per environment, the observed values can be expressed as:

$$Y_{hg(ij)}=\mu+E_h+R_{i(h)}+C_{j(h)}+G_{g(ij)}+GE_{hg(ij)}+e_{hg(ij)},$$

where μ is the population mean; E_h is the fixed effects of the hth environment; $R_{i(h)}$ is the fixed effect of the ith row in the hth environment; $C_{j(h)}$ is the fixed effect of the jth column in the hth environment; $G_{g(ij)}$ is the gth genotype effect in the ith row and the jth column within the hth environment, $G_{g(ij)}\sim(0, \sigma^2_G)$; $GE_{hg(ij)}$ is the interactive effect between the hth environment and the gth genotype, $GE_{hg(ij)}\sim(0, \sigma^2_{GE})$; $e_{hg(ij)}$ is the residual effect, $e_{hg(ij)}\sim(0, \sigma^2_e)$.

An adjusted unbiased prediction (AUP) method (Zhu 1993; Zhu and Weir 1996) can be used to predict genotypic values which can then be used in the calculation of genetic distances and in cluster analysis.

Construction and evaluation of core collection

Genetic distance calculation and cluster analysis

Mahalanobis distance (Mahalanobis 1936) calculated by a variance-covariance matrix can deal with correlations among traits and eliminate the scalar differences between traits. Therefore, the Mahalanobis distance is used in the present research.

The unweighted pair-group average method (Sokal and Michener 1958), Ward's method (Ward 1963) and the complete linkage method (Sorensen 1948) of hierarchical cluster analysis are used for grouping accessions. Finally, accessions are divided into hierarchical groups; each of the subgroups at the lowest level of dendrogram could have one or two accessions with most genetic similarity.

Stepwise clustering and sampling strategies

Based on the dendrogram of clusters, three sampling strategies combined with stepwise clustering are proposed to construct core collections.

(1) *Random sampling.* For this strategy, one accession from each subgroup with two accessions at the lowest level of sorting is randomly selected. If there is only one accession in a subgroup, it is directly sampled for the next cluster.

(2) *Preferred sampling.* By this strategy, accessions with maximum or minimum values of traits are preferred to select from each subgroup at the lowest level of sorting. Both accessions are selected if two accessions in a subgroup have maximum or minimum values of the traits. The other procedures are similar to the random sampling strategy.

(3) *Deviation sampling.* The degree of deviation of two accessions are compared in each subgroup at the lowest level of sorting; the accession with larger degree of deviation is selected for the next cluster analysis. If there is only one accession in a subgroup, it is also directly sampled for the next cluster. The degree of deviation of one accession can be determined by the formula:

$$s_i^2=\sum_{j=1}^{m}\frac{g_{ij}^2}{\sigma_j^2},\quad i=1,2,\ldots,n,\quad j=1,2,\ldots,m,$$

where σ_j^2 is the genotypic variance of the jth trait, and g_{ij} is the ith genotype value of the jth trait.

The genetic distance among all accessions selected on the basis of dendrograms from the first cluster analysis are calculated, then the second cluster analysis of the accessions is performed and accessions are selected based on a new dendrogram of clusters by one of the three sampling strategies, respectively. In the same way, the stepwise cluster analyses can be conducted until selected accessions are reduced to 20–30% of the initial collection (Crossa et al. 1995; Yonezawa et al. 1995), and then the construction of the core collections is completed.

Evaluation of the core collection

A homogeneity test (F-test) for variances and a t-test for means (α=0.05) can be performed to determine the difference of traits between core collections and the initial collection. Then the percentage of the significant difference between the core collections and the initial collection is calculated for the mean difference percentage (MD%) or the variance difference percentage (VD%) of traits.

The coincidence rate (CR%= $\frac{1}{m}\sum_{j=1}^{m}\frac{R_C}{R_I}\times100$) and the variable rate (VR%= $\frac{1}{m}\sum_{j=1}^{m}\frac{CV_C}{CV_I}\times100$) are designed to evaluate the properts of the core collection in terms of the initial collection, where R_C=range of the core collection, R_I=range of the initial collection, CV_C=coefficient of variation of the core collection, CV_I=coefficient of variation of the initial collection, m=number of traits.

The core collection is considered to be the representative of the initial collection under the following situations: (1) no more than 20% of the traits have different means (significant at α=0.05) between the core collection and the initial collection; and (2) the CR% retained by the core collection is no less than 80%.

Worked example

Constructing nine core collections

As a demonstration of constructing core collections, we analyzed 21 traits for 168 accessions of cotton germplasm. These traits were 11 agronomy traits (plant height, height of fruid branch, length of fruiting node, length of boll stalk, number of fruiting branch per plant, bolls per plant, incidence of infected plant, index of wilt disease, growth period, boll weight, and lint percentage), five fiber traits (length, uniformity, strength, elongation and micronaire) and five seed traits (seed length, seed width, ratio of length to width, seed index and kernel weight).

Nine core collections were developed by three cluster methods (the unweighted pair-group average method,

Table 1 Percentage of the trait differences between the core collections and the initial collection

Statistic	CoreC1S1	CoreC1S2	CoreC1S3	CoreC2S1	CoreC2S2	CoreC2S3	CoreC3S1	CoreC3S2	CoreC3S3
VD%[a]	9.5	33.3	57.1	4.8	33.3	42.9	4.8	57.1	42.9
MD%[b]	0	0	0	0	0	0	0	0	0
CR%[c]	82.8	100.0	93.3	90.7	100.0	95.3	88.7	100.0	94.1
VR%[d]	112.0	123.8	130.9	112.1	119.8	122.4	108.2	126.6	124.4

[a] Percentage of significant difference (α=0.05) between core collection and the initial collection for variance of traits
[b] Percentage of significant difference (α=0.05) between core collection and the initial collection for means of traits
[c] Coincidence rate
[d] Variable rate

C1; Ward's method, C2; and the complete linkage method, C3) combined with three sampling strategies (random sampling, S1; preferred sampling, S2; and deviation sampling, S3). The nine core collections were named CoreC1S1, CoreC1S2, CoreC1S3, CoreC2S1, CoreC2S2, CoreC2S3, CoreC3S1, CoreC3S2 and CoreC3S3, respectively.

There was no significant difference (α=0.05) for the means of all traits between each of the nine core collections and the initial collection. The *MD*% was 0% and the *CR*% was larger than 80% in the nine core collections. This indicated that each of the nine core collections was representative of the initial collection.

Evaluation of the cluster methods

Effects of the different cluster methods on the core collections were compared under the conditions of the same sampling strategy. There was no significant difference for all trait means between the initial collection and each of the core collections developed by the three cluster methods (Table 1).

CoreC1S1, CoreC2S1 and CoreC3S1 were developed by random sampling combined with the unweighted pair-group average method, Ward's method and the complete linkage method, respectively. As compared with Ward's method and the complete linkage method, the unweighted pair-group average method tended to give higher values of *VD*%, slightly higher or similar values of *VR*%, and lower values of *CR*% (Table 1). Ward's method and the complete linkage method gave very similar results for *VD*%. *CR*% and *VR*% in the core collection developed by Ward's method were larger than those developed by the complete linkage method. Therefore, considering all parameters, it can be concluded that Ward's method and the unweighted pair-group average method were slightly better than the complete linkage method when random sampling was used to develop the core collections.

Three core collections were developed by preferred sampling combined with the unweighted pair-group average method (CoreC1S2), Ward's method (CoreC2S2) and the complete linkage method (CoreC3S2), respectively. The complete linkage method gave a higher *VD*% and *VR*% as compared with the unweighted pair-group average method and Ward's method (Table 1). The three cluster methods had the same results for *CR*%=100%. The unweighted pair-group average method and Ward's method gave a very similar *VD*%. By using the unweighted pair-group average method, the constructed core collection tended to have a slightly larger *VR*% than that obtained by using Ward's method. It can be concluded that the complete linkage method was better than the other two cluster methods, and the unweighted pair-group average method was slightly better than Ward's method when preferred sampling was used to develop the core collections.

When a deviation sampling strategy was employed, three core collections were obtained by combining with the unweighted pair-group average method (CoreC1S3), Ward's method (CoreC2S3) and the complete linkage method (CoreC3S3), respectively. As compared with Ward's method and the complete linkage method, the unweighted pair-group average method gave a higher *VD*% and *VR*% (Table 1). The three cluster methods resulted in a similar *CR*%. There were similar *VD*% and *VR*% between the core collection developed by the complete linkage method and by Ward's method. For the *VD*% and *VR*% of core collections developed by deviation sampling, the higher values were better. It can be concluded that the unweighted pair-group average method was better than Ward's method and the complete linkage method; whereas the effect of Ward's method and the complete linkage method was similar when deviation sampling was used to develop the core collections.

Among the nine core collections studied, a better representation of the initial collection could be obtained for CoreC2S1 and CoreC1S1 by random sampling, for CoreC3S2 by preferred sampling, and for CoreC1S3 by deviation sampling.

Evaluation of sampling strategies

The effectiveness of the different sampling strategies for core collections was compared under the conditions of the same cluster method. There was no significant difference in trait means between the initial collection and the core collections developed by the three sampling strategies (Table 1).

Among the three core collections constructed by the unweighted pair-group average method combined with random sampling (CoreC1S1), preferred sampling (CoreC1S2) and deviation sampling (CoreC1S3), respec-

Table 2 Comparison between core collections based on genotypic and phenotypic values

Statistic	Phenotyic			Genotypie		
	PCoreC2S1	PCoreC3S2	PCoreC1S3	GcoreC2S1	GCoreC3S2	GCoreC1S3
VD%[a]	4.8	28.6	14.3	4.8	57.1	57.1
MD%[b]	0	0	9.5	0	0	0
CR%[c]	85.1	98.5	91.5	90.7	100.0	93.3
VR%[d]	105.8	119.3	112.4	112.1	126.6	130.9

[a] Percentage of significant difference (α=0.05) between core collection and the initial collection for variance of traits
[b] Percentage of significant difference (α=0.05) between core collection and the initial collection for means of traits
[c] Coincidence rate
[d] Variable rate

tively, CoreC1S1 had the lowest *VD*%, *CR*% and *VR*% (Table 1). When using the preferred sampling strategy, a value of 100% was obtained for *CR*%, while *VD*% and *VR*% tended to be higher than those with random sampling but lower than with deviation sampling. Deviation sampling gave the largest values of *VD*% and *VR*%, but an intermediate value of *CR*%, among the three sampling strategies (Table 1).

Three core collections were developed by Ward's method combined with random sampling (CoreC2S1), preferred sampling (CoreC2S2) and deviation sampling (CoreC2S3), respectively. These core collections had the same trends as those developed by the unweighted pair-group average method combined with the three sampling strategies, respectively (Table 1).

Among the three core collections developed by the complete linkage method combined with random sampling (CoreC3S1), preferred sampling (CoreC3S2) and deviation sampling (CoreC3S3), respectively, CoreC3S1 had also the lowest *VD*%, *CR*% and *VR*% (Table 1). Preferred sampling tended to give 100% *CR*%, a larger *VD*% and slightly larger *VR*% as compared with deviation sampling. Deviation sampling gave an intermediate value of *CR*% as compared with random sampling and preferred sampling, which was similar to the results in core collections developed by the unweighted pair-group average method and Ward's method combined with the three sampling strategies (Table 1).

Therefore, the properties of core collections constructed by the three sampling strategies were the same when using the unweighted pair-group average method and Ward's method, and were almost the same when using the complete linkage method. It was evident that the properties of the core collections developed by all three sampling strategies were stable and feasible.

Comparison between core collections based on genotypic and phenotypic values

CoreC2S1, CoreC3S2 and CoreC1S3 were chose to compare the difference of core collections based on genotypic and phenotypic values. Phenotypic core collections (PCoreC2S1, PCoreC3S2 and PCoreC1S3) were constructed in the same way as the genotypic core collections (GCoreC2S1, GCoreC3S2 and GCoreC1S3) except that genotypic values were replaced by phenotypic values.

MD% was smaller than 20% and *CR*% was larger than 80% in PCoreC2S1, PCoreC3S2 and PCoreC1S3; it was considered that the core collections were representative of the initial collection (Table 2).

There were the same *VD*% and *MD*% between PCoreC2S1 and GCoreC2S1 developed by random sampling. *CR*% and *VR*% in PCoreC2S1 were smaller than those in GCoreC2S1 (Table 2). Genotypic core collections developed by random sampling were considered to be better representative than those developed by phenotypic values.

PCoreC3S2 developed by preferred sampling had the same *MD*% but a smaller *VD*% and *VR*% as compared with GCoreC3S2 (Table 2). The core collections developed by preferred sampling based on genotypic values kept accessions with maximum or minimum values for traits (*CR*%=100%); however, the *CR*% obtained based on phenotypic values of PCoreC3S2 did not attain 100%. The reason was that the core collection with these phenotypic values did not ensure the retention of accessions with maximum genetic variation for the traits studied.

The *VD*%, *VR*% and *CR*% of PCoreS1S3 were smaller than those of GCoreC1S3. The *MD*% of PCoreC1S3 was markedly larger than that of GCoreC1S3.

The results showed that core collections based on genotypic values retained larger genetic variability for traits and had better genetic representation than the core collections based on phenotypic values, especially in the core collection developed by deviation sampling.

Discussion

Sampling strategies can affect the property of core collections. Three sampling strategies applied in this study have their own respective characteristics. Core collections developed by random sampling can determine the genetic diversity structure of initial genetic resources, because accessions are randomly sampled from each of the subgroups at the lowest level of sorting, with the smallest *VD*% and *VR*% among the core collections by the three sampling strategies. Random sampling can be

used if a core collection maintains the genetic diversity pattern of the initial collection. The core collection developed by deviation sampling can be considered as representative of the genetic variability of the initial collection, because the deviation sampling strategy selects accessions with a larger value of s_i^2. The variances and the coefficient of variation in the core collection should be larger. The results showed that the core collection constructed in this way had the highest *VD*% and *VR*% by all three sampling strategies. Deviation sampling can be used if a core collection retains a larger genetic variability of the initial collection. Core collections developed by preferred sampling can produce accessions with both maximum and minimum values of traits, and at the same time still retain the genetic variation structure of the initial collection. The core collection assembled by this way make *CR*%=100%. The core collection with a larger *VD*% and *VR*% is considered to provide a good representation of the genetic diversity of the initial collection. The preferred sampling strategy can be used for developing a core collection retaining accessions with special or valuable characteristics in the initial germplasm collection.

How to ascertain the suitable group numbers (i.e. how to ascertain threshold values of genetic distance for classification criteria or cutting point) after cluster analysis is still not fully resolved theoretically. Ascertaining threshold values is often affected by subjective factors. After the number of groups is determined, some accessions will be sampled from each group. However, up to now there is no suitable way of deciding the number of accessions selected from each group. Group size is not considered when selecting an equal number of accessions from each group. Sampling in proportion to the number of accessions in a group does not consider the genetic relationship among the groups. Methods of constructing core collection by stepwise clustering in the present study does not need to determine the threshold value (cut-off point) or to consider the group number and group size. One accession from each subgroup with two accessions of the most similar genetic variation is selected at the lowest level of sorting in each cluster.

The results of this study show that the unweighted pair-group average method, Ward's method, and the complete linkage method of hierarchical clustering should be combined with different sampling strategies for constructing a core collection. The representation of core collections developed by different sampling strategies and cluster methods is quite distinct.

Acknowledgements We are grateful to R.X. Li for the use of part of his data. This research was supported by the National Natural Science Foundation of China.

References

Basigalup DH, Barnes DK, Stucker RE (1995) Development of a core collection for perennial *Medicago* plant introductions. Crop Sci 35:1163–1168

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

Crossa J, Delacy IH, Taba S (1995) The use of multivariate methods in developing a core collection. In: Hodgkin T, Brown AHD, Hintum van ThJL, Morales EAV (eds) Core collections of plant genetic resources. John Wiley Sons, Chichester, UK, pp 77–92

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

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

Erskine W, Muehlbauer FJ (1991) Allozyme and morphological variability, outcrossing rate and core collection formation in lentil germplasm. Theor Appl Genet 83:119–125

Frankel OH, Brown AHD (1984a) Current plant genetic resources – a critical appraisal. In: Genetics: new frontiers (vol IV). Oxford and IBH Publishing, 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. Geoge Allen and Unwin, London, UK, pp 249–257

Gepts P (1995) Genetic markers and core collections. In: Hodgkin T, Brown AHD, Hintum van ThJL, Morales EAV (eds) Core collections of plant genetic resources. John Wiley and Sons, Chichester, UK, pp 127–146

Hintum van TJL (1994) Comparison of marker system and construction of a core collection in a pedigree of European spring barley. Theor Appl Genet 89:991–997

Hintum van TJL (1995) Hierarchical approaches to the analysis of genetic diversity in crop plants. In: Hodgkin T, Brown AHD, Hintum van ThJL, Morales EAV (eds) Core collections of plant genetic resources. John Wiley and Sons, Chichester, UK, pp 23–34

Holbrook CC, Anderson WF, Pittman RN (1993) Selection of a core collection from the US germplasm collection of peanut. Crop Sci 33:859–861

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

Perry MC, McIntosh 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

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Science Bull 38:1409–1438

Sorensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol Skrifter 5:1–34

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Yonezawa K, Nomura T, Morishima H (1995) Sampling strategies for use in stratified germplasm collections. In: Hodgkin T, Brown AHD, Hintum van ThJL, Morales EAV (eds) Core collections of plant genetic resources. John Wiley and Sons, Chichester, UK, pp 35–53

Zeuli PLS, Qualset CO (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theor Appl Genet 87:295–304

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

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