ISSN 1009 - 3095 *Journal of Zhejiang University(SCIENCE)* V.2, No.1, P.89 - 93, Jan. - Mar., 2001
http://www.chinainfo.gov.cn/periodical; http://www.zju.edu.cn;
http://lib.zju.edu.cn/English; http://www.zjupress.com

STUDIES ON PROPERTY OF SAMPLE SIZE AND DIFFERENT TRAITS FOR CORE COLLECTIONS BASED ON GENOTYPIC VALUES OF COTTON

HU Jin(胡 晋), ZHU Jun(朱 军), XU Hai-ming(徐海明)

(*Department of Agronomy, Zhejiang University, Hangzhou 310029, China*)

Received Mar.28, 2000; revision accepted May 25, 2000

Abstract: Studies were conducted on specific core collections constructed on the basis of different traits and sample size by the method of stepwise cluster with three sampling strategies based on genotypic values of cotton. A total of 21 traits (11 agronomy traits, 5 fiber traits and 5 seed traits) were used to construct main core collections. Specific core collections, as representative of the initial collection, were constructed by agronomy, fiber or seed trait, respectively. As compared with the main core collection, specific core collections tended to have similar property for maintaining genetic diversity of agronomy, seed or fiber traits. Core collections developed by about sample size of 17% ($P_2 = 0.17$) and 24% ($P_1 = 0.24$) with three sampling strategies could be quite representative of the initial collection.

Key words: specific core collection, stepwise cluster, sampling strategy, genotypic values
Document code: A **CLC number:** S32

INTRODUCTION

The potential user of core collection falls into three main groups: plant breeders who wish to find and utilize germplasm; germplasm specialists who wish to study germplasm or genetic variation; curators who require assistance with germplasm management (Mackay, 1995). Germplasm specialists and curators are concerned mainly with management issues, while breeders pay much attention primarily to germplasm use. Constructing a core collection for a kind of crop or initial collections can reduce the size of a gene bank to make it more manageable and usable. One core collection is probably enough for germplasm management, but probably not for the demand of plant breeders, who request a set of accessions containing a characteristic not previously described or used in developing a core collection. Therefore, besides one main core collection based on all traits in the data base, specific core collections can be constructed based on different kinds of traits in order to seek accessions for a breeding plan. For instance, if the fiber quality of cotton needs to be improved, breeders can look for useful genetic materials from a core collection constructed mainly by fiber characteristics. Constructing only one core collection might be difficult to meet the demand for management and identification of accessions (Chang, 1991).

How many accessions can be sampled into one core collection is determined by the sample size of the core collection. There are different opinions on the size of core collection. Brown (1989) suggested that the core collection should contain 5% to 10% of the germplasm collection. Diwan et al. (1994) developed a core collection for annual Medicago with 17% of initial collection, the 5% and 10% sample size of core collection were judged insufficient to represent the initial collection (Diwan et al., 1995). Zeuli and Qualset (1993) thought that an evaluator or breeder could process about 500 accessions easily, thus they chose 16% as sample size of core collection to represent 3038 original accessions. The size of peanut core collection was 11.2% of initial collection (Holbrook et al., 1993). Cassava core collection contained 630 accessions to represent 5169 initial accessions; sample size of core collection was about 12.2% (Wheatley et al., 1993). Sample size of 20-

* Project(No 39470377) supported by NSFC

30% from initial accessions was suitable for core collection in the study of Yonezawa et al. (1995). Crossa et al. (1995) stopped sampling when the size of core collection was 27% of original collection in developing a core collection. However, most researches on size of core collection were made based on phenotypic values.

In this study, genotypic values were used to construct core collections using stepwise cluster combined with three sampling strategies (Hu et al., 2000). Comparison was conducted for specific core collections of different kinds of traits and for sample size of core collections based on genotypic values.

MATERIALS AND METHODS

Plant materials

A two-year data set of 21 traits for 168 accessions of upland cotton germplasm was used as a working example. Eleven agronomy traits (plant height, height of fruit 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) were considered.

Genetic models and predicted genotypic values

When experiment is conducted to evaluate germplasm resources by using a large number of accessions, genetic materials can be planted along with the check in plots arranged by rows and columns of field. A genetic model with genotype (environment (GE) interactions for controlling systematical errors in the field can be used for analyzing variance components (Hu, 1999; Hu et al., 2000). 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 calculation of genetic distances and cluster analysis.

Construction and evaluation of core collection

Mahalanobis distance calculated based on genotypic values is applied to measure genetic distance among accessions (Mahalanobis, 1936; Hu et al., 2000).

Three sampling strategies (random sampling, S1; preferred sampling, S2; and deviation sampling, S3) and three cluster methods (the unweighted pair-group average method, C1; Ward's method (Ward, 1963), C2; and the complete linkage method, C3) were combined, respectively, according to the best combinations between sampling strategies and cluster methods in a previous study (Hu et al., 2000). Three main core collections using a total of 21 traits (T) and nine specific core collections using 11 agronomy traits (A), five fiber traits (F), five seed traits (S) were constructed, respectively according to Hu et al. (2000) methods. The main core collections were named as TCoreC2S1, TCoreC3S2, TCoreC1S3, the specific ones as ACoreC2S1, FCoreC2S1, SCoreC2S1, ACoreC3S2, FCoreC3S2, SCoreC3S2, ACoreC1S3, FCoreC1S3, SCoreC1S3, respectively. Moreover, core collections based on 21 traits with different sample size were constructed by the methods of CoreC2S1, CoreC3S2, and CoreC1S3 (Hu et al., 2000). Core collections were developed until selected accessions were reduced to an average of 24% ($\bar{P}_1 = 0.24$) and average of 17% ($\bar{P}_2 = 0.17$) of the initial collection, respectively.

Homogeneous test (F-test) for variances and t-test for means ($\alpha = 0.05$) were used to compare the differences between core collections and the initial collection. Then the percentage of the significant difference between the core collection and the initial collection was calculated for the mean difference percentage ($MD\%$) or the variance difference percentage ($VD\%$) of traits (Hu et al., 2000). The coincidence rate for range ($CR\%$) and the variable rate for coefficient of variation ($VR\%$) were used to evaluate the properties of the core collection in terms of the initial collection (Hu et al., 2000).

The core collection was considered to be representative of the initial collection under the following conditions: (1) $MD\%$ was no more than 20% (significant at $\alpha = 0.05$); and (2) $CR\%$ of the core collection was no less than 80%.

RESULTS AND DISCUSSION

Comparison of specific core collections of different traits with main core collection

Three specific core collections were constructed by Ward's method (Ward, 1963) and random sampling strategy based on agronomy (ACoreC2S1), fiber (FCoreC2S1) and seed (SCoreC2S1) traits, respectively. There was no significant difference (*MD%* = 0%) for means between the three specific core collections and the initial collection. As compared with the initial collection, the *CR%* was larger than 80% in the three specific core collections. It was indicated that the three specific core collections developed by agronomy, fiber and seed traits, respectively could be representative of the initial collection (Table 1).

Table 1 Percentage of trait difference of specific core collections and main core collection with average 26% sample sizes by Ward's method and random sampling strategy for different traits

Statistics	ACoreC2S1	FCoreC2S1	SCoreC2S1
VD% [a]	9.1 (9.1[e])	20.0 (0)	0 (0)
MD% [b]	0 (0)	0 (0)	0 (0)
CR% [c]	89.8 (90.2)	83.6 (90.4)	90.3 (92.2)
VR% [d]	110.9 (108.7)	114.1 (115.1)	109.7(116.6)

[a]Percentage of significant difference ($\alpha = 0.05$) between core collection and the initial collection for variance of traits; [b] Percentage of significant difference ($\alpha = 0.05$) between core collection and the initial collection for means of traits; [c]Coincidence rate; [d]Variable rate; [e]Data in parenthesis were values of agronomy, fiber or seed trait in TCoreC2S1 based on a total 21 traits.

Comparison of *VD%*, *MD%*, *CR%* and *VR%* of the three specific core collections with the values in parenthesis (Table 1) for agronomy, fiber and seed traits in the main collection based on a total of 21 traits (TCoreC2S1), respectively, showed that: there was the same zero *MD%* in the three specific core collections; *VD%*, *CR%* and *VR%* were similar in the specific core collection of agronomy trait (ACoreC2S1); *VD%* was increased; *CR%* was decreased; *VR%* was almost the same in FCoreC2S1; *VD%* was the same; *CR%* was approximately the same; *VR%* was slightly decreased in SCoreC2S1.

Three specific core collections constructed by the preferred sampling strategy and complete linkage method were based on agronomy (ACoreC3S2), fiber (FCoreC3S2) and seed (SCoreC3S2) traits, respectively. *MD%* was 0%, *CR%* was 100% (larger than 80%), in all three specific core collections. Therefore, each of the three specific core collections developed by different traits was representative of the initial collection (Table 2).

Table 2 Percentage of trait difference between specific core collections and main core collection with average 26% sample sizes by the complete linkage method and preferred sampling strategy for different traits

Statistics	ACoreC3S2	FCoreC3S2	SCoreC3S2
VD% [a]	9.1 (27.3[e])	100.0 (100.0)	0 (80.0)
MD% [b]	0 (0)	0 (0)	0 (0)
CR% [c]	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
VR% [d]	115.8 (120.5)	134.6 (139.2)	118.6 (127.8)

[a], [b], [c], [d]are the same as in Table 1, [e]data in parenthesis were values of agronomy, fiber or seed trait in TCoreC3S2 based on a total of 21 traits.

Comparison of the statistics of the three specific core collections with the values in parenthesis (Table 2) of agronomy, fiber and seed traits in the main core collection (TCoreC3S2), respectively, showed that: the value of *MD%* and *CR%* were the same; *VD%* was decreased; *VR%* was similar to that in ACoreC3S2; *VD%* was the same; *VR%* was similar to that in

FCoreC3S2; *VD%* and *VD%* were decreased in SCoreC3S2.

When deviation sampling strategy was combined with the unweighted pair-group average method, three specific core collections were constructed based on agronomy (ACoreC1S3), fiber (FCoreC1S3) and seed (SCoreC1S3) traits, respectively. The *MD%* was 0% and *CR%* was larger than 80% in the three specific core collections. It was found that each of the three specific core collections developed by different traits could be representative of the initial collection (Table 3). Note sentence is betten coithout the correetion in redink.

Comparison of the statistics of the three specific core collections with the values in parenthesis (Table 3) of agronomy, fiber and seed traits in the main core collection (TcoreC1S3), respectively, showed that: they had the same value of *MD%*; *CR%* was slightly enlarged in the three specific core collections; *VD%* was increased and % was similar in ACoreC1S3; *VD%* and *VR%* were decreased in FCoreC1S3; *VD%* was greatly increased and *VR%* was slightly decreased, in SCoreC1S3.

Table 3 Percentage of trait differences between specific core collections and main core collection with average 27% sample sizes by the unweighted pair-group average method and deviation sampling strategy for different traits

Statistics	ACoreC3S2	FCoreC3S2	SCoreC3S2
VD% [a]	45.5 (36.4[e])	80.0 (100.0)	100.0 (60.0)
MD% [b]	0 (0)	0 (0)	0 (0)
CR% [c]	94.0 (93.7)	93.5 (90.0)	97.0 (95.6)
VR% [d]	123.6 (125.9)	127.9 (137.9)	128.4 (135.5)

[a, b, c, d] are the same as in Table 1, [e]data in parenthesis were values of agronomy, fiber or seed trait in TCoreC1S3 based on a total of 21 traits.

It can be considered that specific core collections tend to have similar property for maintaining genetic diversity of agronomy, seed or fiber traits, as compared with main core collection, especially by random sampling and deviation sampling. Since specific core collections in the present study could be representative of the initial collection, genotypes containing desired attributes could then be conveniently searched for within the specific core collection. However, specific core collection cannot replace but only complement main core collection. Since core entries are only marked in the database, and are not related to accession quantity in a gene bank, more than one core collection will not be too difficult to develop.

Comparison of core collections constructed on the basis of different sample sizes

In the six core collections, the values of *MD%* were 0% and those of *CR%* were larger than 80% (Table 4). This showed that each of the six core collections developed on the basis of different sample size could still be representative of the initial collection.

Table 4 Percentage of trait differences between core collections constructed on the basis of two sample sizes

Statistics	$\bar{P}_1 = 0.24$			$\bar{P}_2 = 0.17$		
	CoreC2S1	CoreC3S2	CoreC1S3	CoreC2S1	CoreC3S2	CoreC1S3
VD% [a]	4.8	57.1	57.1	14.3	66.7	81.0
MD% [b]	0	0	0	0	0	0
CR% [c]	90.7	100.0	93.3	87.5	100.0	92.0
VR% [d]	112.1	126.6	130.9	115.7	132.5	142.5

[a, b, c, d] are the same as in Table 1.

As compared with core collections constructed on the basis of $\bar{P}_1 = 0.24$, the small core collection ($\bar{P}_2 = 0.17$) tended to have higher values of *VD%*, the same values of MD%, and similar values of *CR%* and *VR%* when based on random sampling (CoreC2S1); the small core collection tended to have higher values of *VD%*, the same values of *MD%* and *CR%*, slightly higher values of *VR%* when constructed on the basis of the preferred sampling strategy (CoreC3S2); the small core collection had higher values of *VD%* and *VR%*, similar values of *MD%* and *CR%* when constructed based on deviation sampling strategy (CoreC1S3).

Study on sample size of core collections based on genotypic values has not been presented up to now. In the present study, core collections with two sample sizes constructed by stepwise cluster approaches based on genotypic values could represent initial genetic diversity. The appropriate sample size of core collections should be determined according to such factors as species, crops, number of initial accessions, ability of keeping germplasm, and easy manipulation.

In future study, molecular markers can be used in core collection research (Hokanson et al., 1998; Divaret et al., 1999) because molecular markers such as RAPDs and RFLPs can reflect direct changes at the DNA sequence level. However, molecular markers are not genes, core collections constructed by the molecular marker method are difficult to match with those based on analyzing the traits of accessions. Perhaps information from molecular markers combined with genotypic values is better for developing a core collection.

References

Brown, A. H. D., 1989. The case for core collections. *In*: The Use of Plant Genetic Resources. Brown, A. H. D., Frankel, O. H., Marshall, D. R. et al. (eds.), Cambridge University Press, Cambridge, p.136 - 156.

Chang, T. T., 1991. Guidelines on developing core collections of rice cultigens. *In*: Rice Germplasm Collecting, Preservation, Use: Proceedings of third international workshop on rice germplasm. Pollard, L., (ed.), IRRI, Philippines, p.105 - 107.

Crossa, J., Delacy, I. H. and Taba, S., 1995. The use of multivariate methods in developing a core collection. *In*: Core Collections of Plant Genetic Resources. Hodgkin, T., Brown, A. H. D., Hintum, van Th. J. L., et al. (eds.), John Wiley & Sons, Chichester, UK, p.77 - 92.

Divaret I., Margalé, E. and Thomas, G., 1999. RAPD markers on seed bulks efficiently assess the genetic diversity of a *Brassica oleracea* L. collection. *Theor. Appl. Genet.*, **98**: 1029 - 1035.

Diwan, N., Bauchan, G. R. and McIntosh, M. S. A., 1994. Core collection for the United States annual *Medicago* germplasm collection. *Crop. Sci.*, **34**: 279 - 285.

Diwan, N., McIntosh, M. S. and Bauchan, G. R., 1995. Methods of developing a core collection of annual *Medicago* species. *Theor. Appl. Genet.*, **90**: 755 - 761.

Hokanson, S. C., Szewc-McFadden, A. K., Lamboy, W. F. et al., 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, C. C., Anderson, W. F. and Pittman, R. N., 1993. Selection of a core collection from the U.S. germplasm collection of peanut. *Crop. Sci.*, **33**: 859 - 861.

Hu, J., 1999. Studies on Methods of Developing Core Collections by Stepwise Clusters in Crops. Ph. D. Dissertation, Zhejiang University, Hangzhou, China.

Hu, J., Zhu, J. and Xu, H. M., 2000. Methods of constructing core collections by stepwise cluster with three sampling strategies based on genotypic values of crops. *Theor. Appl. Genet.*, (in printing).

Mackay, M. C., 1995. One core collection or many? *In*: Core Collections of Plant Genetic Resources. Hodgkin, T., Brown, A. H. D., Hintum, van Th. J. L., et al. (eds.), John Wiley & Sons, Chichester, UK, p.199 - 210.

Mahalanobis, P. C., 1936. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, **2**: 49 - 55.

Ward, J. H., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**: 236 - 244.

Wheatley, C. C., Orrego, A. J. I., Sanchez, T. et al. 1993. Quality evaluation of the cassava core collection at CIAT. *In*: First International Scientific Meeting Cassava Biotechnology Network: proceedings. Roca, W. M. and Thro, A. M. (eds.), Cali, Colombia: CIAT, p.255 - 264.

Yonezawa, K., Nomura, T. and Morishima, H., 1995. Sampling strategies for use in stratified germplasm collections. *In*: Core Collections of Plant Genetic Resources. Hodgkin, T., Brown, A. H. D., Hintum, van Th. J. L., et al. (eds.), John Wiley & Sons, Chichester, UK, p.35 - 53.

Zeuli, P. L. S. and Qualset, C. O., 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. *Journal of Biomath.*, **8**: 32 - 44.

Zhu, J. and Weir, B. S., 1996. Diallel analysis for sex-linked and maternal effects. *Theor. Appl. Genet.*, **92**: 1 - 9.