Statistics and Applications
Volume 6, Nos.1 & 2, 2008 (New Series), pp. 175-187

Optimum sample size for clustering wild wheat populations using AFLP markers

M. Singh[1,2] and K. Chabane[1]
[1]*International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria*
[2]*Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Canada*

Abstract

Genetic markers such as amplified fragment length polymorphisms (AFLP) are used to fingerprint the genetic makeup of plant material. Cluster analysis is used to identify genetic similarities between populations. With a view to classify wild wheat populations on the basis of AFLP marker data, this study presents an empirical method for estimating the optimum number of plants, and illustrates the procedure on a total of eight populations of *Triticum dicoccoides* and *T. araraticum*. Applications of a resampling with replacement method and fitting of an exponential function of sample size to the least squares scaling of stress (LSS) derived from the similarity matrix were made. An expression for the optimum sample size was then derived from an exponential relationship. When applied on the presence/absence data of 203 polymorphic AFLP markers on the population accessions, the method showed that 4 - 5 plants are required to estimate the asymptotic value of the LSS measure within a difference of 5 - 10%.

Key words: Amplified fragment length polymorphism (AFLP); Bootstrapping method; Clustering methods; Wild wheat.

1 Introduction

Conservation of genetic resources is essential for continuous plant breeding efforts leading to plants with desired agronomic traits. Sampling strategies for conservation of genetic resources to preserve a diversity of genes or accessions of the species have been discussed in the literature (Brown 1989, Brown and Marshall 1995, Franco et al.

2005). Molecular studies for genetic polymorphisms have been carried out for a number of different applications, such as anthropological and evolutionary studies, genetic identification in plants and conservation of germplasm and application to crop improvement. Genetic markers, such as amplified fragment length polymorphism (AFLP) have been found to be very useful and economic in capturing the genetic structure of plant materials. Wild wheat populations can be very diverse in genetic makeup. Identifying similarities between populations will help in obtaining a more convenient preservation strategy of the genetic material with no substantial reduction of the diversity. Evaluations of genetic diversity have been carried out using morphological and molecular traits. A review of statistical methods applied in plant biodiversity studies can be found in Mohammadi and Prasanna (2003). Genetic similarities of the genetic entities can be done using a clustering method. In order to obtain clustering of genetic populations or accessions, one approach is to use morphological information but they are subject to the environmental variation, while the other approach would be to use genetic markers such as AFLP marker. Generating information has a cost, for example, the cost would be proportional to the number of plants assayed and the number of markers used. A question often asked is how many plants/accessions (at the minimum) are needed to provide nearly the same clusters as prevailing in the population of a given species. Determination of sample size to estimate the genetic diversity has been addressed by several researchers including (Nei, 1987; Barrett et al., 1998; Hegde, et al. 2002; Dreisigaker et al. 2004; Singh et al., 2006). Another aspect of diversity study is to examine similarity between the populations, for example by using clustering methods (Weir, 1996; Pritchard et al., 2000; Hegde et al., 2002). However, limited effort has been made for determining optimum sample size required for clustering of genetic materials based on genetic data.

Using the AFLP markers data, this study evaluates the effect of sample size on the clustering of accessions, for a chosen similarity definition and a clustering method. For a given clustering method, the clusters depend on the similarity matrix. It is expected that the similarity matrix based on the optimum sample size will be close to the similarity matrix if all the plants in the populations were used. We can use a measure of fit, called the least square scaling of the stress between the similarity matrix based on a sample and population, as

the standardized sum of squares of difference between the elements of the two matrices (LSS). If the same set of markers and similarity-measure are used for the sample and population similarities, so no further transformation would be necessary. Another variant of the measure of fit could be the standardized sum of squares of difference between the square of the elements of the two matrices (LSSS). These measures were introduced in psychometric applications and has been described in Cox and Cox (1994) and Gower and Hand (1996).

Sample size determination for the situations where the variance of estimate of diversity is expressible as sample size can be done explicitly as in Nei (1987). However, where there is no expression of variance available for genetic diversity, one may use empirical relationship with the sample size (Singh et al., 2006). In the present case of clustering the populations, since there is no explicit expression available for LSS or LSSS in terms of sample size, we can pursue the approach of Fairfield Smith (1938) for the uniformity field trials and its recent applications by Girma (2000) and Fagrouda and Van Meirvenne (2002). Fairfield Smith (1938) computed variance of means on a per-unit basis for various plot sizes from the same trial, and fitted a linear relationship of logarithm of the variance on logarithm of the sample size, equivalent to the variance being expressed as the power of the plot size. In the uniformity field trials, Smith (1938) generated various plot sizes by merging the neighboring plots of unit size; we followed the resampling approach to generate samples of various sizes. There are a number of resampling approaches available in literature, e.g. interpenetrating sub-sampling (Mahalanobis, 1944, 1946; Ghosh 2005), Jackknife method (Quenouille, 1949) and bootstrap method (Efron, 1979) which are used to estimate variance of a given statistic of interest. In the present case, our interest lies in fitting a relationship between LSS/LSSS, a quantity measuring squared deviation, and the sample size developed when the resampling was carried using simple random sampling with replacement (Efron, 1979). No study of the theoretical properties of the distribution of LSS has been aimed at.

The objective of this study was to 1) consider a similarity matrix for classification and a measure of fit or stress, to represent the closeness between the sample similarity matrix and the population similarity matrix, 2) generate stress measures with varying sample sizes using bootstrap methods, 3) identify a functional relation with help of the visual display, called an empirical relationship, and 4) use

the relationship to estimate optimum sample size, and demonstrate the application on the AFLP marker data to cluster 8 populations of wild wheat.

2 Material and methods

2.1 AFLP analysis and data

This study used data on the presence/absence of 203 polymorphic AFLP markers on wild relatives of wheat. Total number of markers used was 277. We used a total of 8 populations of *T. dicoccoides* and *T. araraticum* given in Table 1. Details of the experimental methods are given in Sasanuma et al. (2002).

Table 1. Populations, origin of collection and number of plants/accessions used for this study

Species	Populations	Country	Plants
T. dicoccoides	D1	Iran	9
	D1	Iran	9
	D2	Iraq	8
	D3	Jordan	10
	D4	Syria	9
	D5	Palestine	10
T. araraticum	A1	Turkey	7
	A2	Syria	9
	A3	Iran	9

2.2 Statistical Methods

Consider a set of m populations where a number of accessions ($\boldsymbol{n}_i, i = 1 \ldots m$) have been genotyped using AFLP markers on L loci. The binary data, $\boldsymbol{x}_{ijl}$, is the presence (1)/absence (0) of the AFLP marker band at the l-th locus on the j-th accession of the i-th population ($l = 1 \ldots L, j = 1 \ldots \boldsymbol{n}_i$). Clustering methods can be found in standard texts (Hand 1981, Jain and Dubes 1988) and statistical software (Payne et al. 2007). In order to cluster the populations based on data from all the accessions, we obtained a matrix of means $\left(\bar{x}_{il} = \sum_{j=1}^{j=n_i} x_{ijl}/n_i\right)$.

We used Euclidian distance between the populations and clustering was carried out using average linkage function (also called UPGMA method) which is most frequently used in clustering genetic materials. Let the similarity matrix based on the full data be denoted by a $m \times m$ matrix $S_P = (\psi_{ii'})$.

To generate similarity matrices for varying sample sizes, we follow bootstrap resampling method of Efron (1979). For a given sample size k, say, take a simple random sample with replacement k accessions out of the $\boldsymbol{n}_i$ samples available on a given i-th population $(i = 1 \ldots m)$. Form a matrix $\left(\bar{y}_{il} = \sum_{j=1}^{j=k} y_{ijl}/k \right)$ where y is the sample version of the x, the presence/absence of the l-th marker. Using the values of $(\bar{y}_{il}; i = 1 \ldots m, l = 1 \ldots L)$ we obtain a sample similarity matrix, $SB_{kr} = (\psi_{krii'})(m \times m)$, which with $r = 1$ will be called the first bootstrap (re-sampled) matrix. The process of resampling from each population can be repeated, say N times, to generate N (bootstrap) samples of matrices $SB_{kr}(r = 1, 2, \ldots N)$, each of size k. Then, one can obtain a bootstrap matrix $S\bar{B}_k \left(= \sum_{r=1}^{r=N} SB_{kr}/N \right) = \left(\sum_{r=1}^{r=N} \psi_{krii'}/N \right) = \left(\sum_{r=1}^{r=N} \bar{\psi}_{kii'} \right) =$ as an average value of the similarities. Actually, one would use $S\bar{B}_k$ to cluster the populations based on the marker information on k accessions. Since the cluster formation depends on a similarity matrix and a grouping method, therefore, for a given grouping method, the closeness between sample of a given size and the population can be defined as difference or a function of the difference between the matrices $S\bar{B}_k$ and S_P. We have used two methods of closeness, based on the concept of stress and squared-stress in psychology as a measure of fit based on least squares scaling. For a given sample size k, the two measures are:

Least-squares- scaling (LSS) stress:

$$LSS_k = \sum_{i<i'=1}^{m} \left(\bar{\psi}_{kii'} - \psi_{kii'} \right)^2 / \sum_{i<i'=1}^{m} \bar{\psi}^2_{kii'}$$

Least- squares- squared-scaling (LSSS) stress:

$$LSSS_k = \sum_{i<i'=1}^{m} \left(\bar{\psi}^2_{kii'} - \psi^2_{kii'} \right)^2 / \sum_{i<i'=1}^{m} \bar{\psi}^4_{kii'}$$

Similarly, the above process can be repeated from other sample sizes $k = 2, 3, \ldots A$ graphical display could exhibit an association between the sample size and the closeness of the sample to the population similarity as measured by the two stresses (k, LSS_k) or $(k, LSSS_k)$. The optimum k will be the one with the lowest value of $LSS_k/LSSS_k$, with a preference to a smaller k.

2.3 Estimation of the sample size

Following the approach of Singh et al. (2006) for determining sample size for diversity estimation, the association between a given sample size k and the LSS, ω_k say, was modeled using an exponential relationship:

$$\omega_k = A + BR^k + \eta_k$$

where $\eta_k s$ are assumed independent and identically distributed random errors with mean zero and constant variance. Estimation of the parameters A, B and R of the above model was carried out using ordinary least squares method. Generally R will be $< 1, \omega_k$ will decrease to A (if $B > 0$) as k increases to infinity. It would be sufficient to determine k when ω_k is close to A, say with a relative difference of ε (for which values such as 0.05 or 0.10 may be reasonable). In this case, we get the following equation:

$$\omega_k = A(1 + \varepsilon) = A + BR^k.$$

This gives the optimum k,

$$k_0 = \mathrm{In}(\varepsilon A/B)/\mathrm{In}(R).$$

The estimate of k_o were be obtained by replacing A, B and R by their least-squared estimates.

3 Application/results

Of the 277 AFLP markers, 203 were found polymorphic when 71 accessions/plants from a total of 8 populations of species *T. dicoccoides* and *T. araraticum* were assayed. Using Euclidian distance between the means of presence of marker bands to derive similarity between

the populations, similarity matrix from all the accession records and the bootstrapped matrices for various sizes up to the maximum size available for a given population were developed (Table 2 in Appendix). The clusters are given for various sizes in Fig 1. in Appendix. Using these matrices, the standardized stress values and squared stress values were obtained. The exponential functions of the sample size were fitted to the data on stress measures. The observed and fitted relations are shown in Fig. 2 in Appendix. The estimate of the parameters of the fitted relationships and estimates of optimum sample size for various degree of closeness ($\varepsilon = 0.05 - 0.20$) of the stress levels to its asymptotic limit are given in Table 3 in Appendix.

It may be noted that the exponential curves fit very well to the two stress measures, LSS and LSSS.

4 Discussion

In the approach followed, it would be pertinent to mention that the ω_k values generated by resampling to fit the exponential relationship would not be independent as the marker data used in obtaining ω_ks for larger k values are likely to contain the data used for smaller sample sizes. Thus, this method shares a similar practical limitation with Smith (1938) and Girma (2000) method. Further, we do not find any alternative approach to compare with.

Molecular studies for genetic polymorphisms are being carried out for a number of different applications, such as genetic disorders in different populations, genomics, and genetic identification of ethnic groups for forensic and legal applications, genetic identification of plants for commercial applications and conservation of germplasm. This paper addresses the question of the optimum size of the sample required to cluster the populations with similar genetic structure as detected by the AFLP markers.

This study had a limitation on the sample size (a maximum of 10) and there is a need to see what happens in case of larger number of accessions per populations to validate the results. Further the AFLP markers, being dominant markers, restrict exhibiting only two alleles per locus. In order to cover genetic structure comprising larger number of multiple alleles, we need to extend this approach of sample size determination from data on codominant markers and also on the

other species populations.

Acknowledgements : The authors wish to thank Professor M. S. Srivastava, University of Toronto, Dr. Rajender Parsad, ICARDA, Dr M. Baum, ICARDA, and a referee for providing constructive suggestions on an earlier version of the manuscript.

References

Barrett, B. A., Kidwell, K.K., and Fox, P. N. (1998).Comparison of AFLP and pedigree-based genetic diversity assessment methods using wheat cultivars from the Pacific Northwest. *Crop Sci.* **38**, 1271-1278.

Brown, A.H.D. (1989). Core collections: a practical approach to genetic resources management. *Genome* **31**, 818-824.

Brown, A.H.D. and Marshall, D.R. (1995). A basic sampling strategy: theory and practice. In: Guarino, L., V. Ramanatha Rao and R. Reid (eds.), Collecting Plant Genetic Diversity: Technical Guidelines, 76-91, CAB International, Wallingford.

Cox, T. F. and Cox, M. A. A. (1994). *Multidimensional Scaling.* Chapman & Hall, London.

Dreisigacker, S., Zhang, P., Warburton, M. L., Van Ginkel, M., Hoisington, D., Bohn, M., and Melchinger, A. E. (2004). SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different megaenvironments. *Crop Sci.* **44**, 381-388.

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Ann. Statist.* **7**, 126.

Fagrouda, M. and Van Meirvenne, M. (2002). Accounting for soil spatial autocorrelation in the design of experimental trials. *Soil Sci. Soc. Amer. Jour.* **66**, 1134-1142

Fairfield Smith, H (1938). An empirical law describing heterogeneity in the yield of agricultural crops. *J. Agric. Sci.* **28**, 1-23.

Franco J., Crossa, J., Taba, S. and Shands, H. (2005). A sampling strategy for conserving genetic diversity when forming core subsets. *Crop Sci.* **45**, 1035-1044.

Girma, T., Tarekegne, A. and Tanner, D.G. (2000). Estimation of optimum plot dimensions and replication number for wheat experimentation in Ethiopia. *Afric. Crop Sci. Jour.* **8(1)**, 11-23.

Ghosh, S. (2005). *Interpenetrating Samples. Standard article in Encyclopedia of Biostatistics*. John Wiley & Sons, New York.

Gower, J. C., and Hand, D. J. (1996). *Biplots*. Chapman & Hall, London.

Hand, D.J. (1981). *Discrimination and Classification*. John Wiley and Sons, New York.

Hegde, S. G., Valkoun, J., and Waines, J. G. (2002). Genetic diversity in wild and weedy aegilops, amblyopyrum, and secale species: A Preliminary Survey. *Crop Sci.* **42**, 608-614.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs.

Mahalanobis, P.C. (1944). On large-scale sample surveys. *Phil. Trans. of the Royal Soc.* London, Series **B 231**, 329-451.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute, *Jour. Royal Statist. Soc.* **109**, 325-378.

Mohammadi, S.A. and Prasanna, B. M. (2003). Analysis of genetic diversity in crop plants: salient statistical tools and considerations. *Crop Sci.* **43**, 1235-1248.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Payne, R.W., Harding, S.A., Murray, D.A., Soutar, D.M., Baird, D.B., Welham, S.J., Kane, A.F., Gilmour, A.R., Thompson, R., Webster, R., & Tunnicliffe Wilson, G. (2007). The Guide

to GenStat Release 10, Part 2: Statistics. VSN International, Hemel Hempstead.

Pritchard J.K., Stephens M., Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945959.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Jour Royal Statist. Soc.* **Series B 11**, 18-44.

Sasanuma, T., Chabane, K., Endo T. R. and Valkoun, J. (2002). Genetic diversity of wheat wild relatives in the Near east detected by AFLP. *Euphytica* **127** 81-93.

Singh M, Chabane K, Valkoun J, and Blake T. (2006). Optimum sample size for estimating gene diversity in wild wheat using AFLP markers. *Genet. Resour. Crop Evol* **53(1)**, 23-33.

Weir, B. S. (1996). *Genetic Data Analysis*. Sinauer Associate Inc., Sunderland.

M. Singh
Concordia University
1455 De Maisonneuve Blvd. West
Montreal, QC, H3G 1M8, Canada
Email: smurari@mathstat.concordia.ca

K. Chabane
International Center for Agricultural Research in the Dry Areas (ICARDA)
PO Box 5466, Aleppo, Syria

Appendix

Table 2 Similarly matrices using Euclidian distances on allelic frequencey computed from all accessions (full data) and frok bootstraps at different sample sizes

	D1	D2	D3	D4	D5	A1	A2	A3	D1	D2	D3	D4	D5	A1	A2	A3
D1	100					Full data										
D2	78	100														
D3	75	79	100													
D4	70	70	70	100												
D5	66	75	73	64	100											
A1	65	67	66	69	69	100										
A2	69	70	69	70	72	76	100									
A3	69	70	70	73	69	75	78	100								
D1	100					$k=1$			100					$k=2$		
D2	75	100							77	100						
D3	72	73	100						74	76	100					
D4	71	68	69	100					71	70	70	100				
D5	68	70	71	64	100				70	73	70	64	100			
A1	69	71	67	68	67	100			69	69	68	70	70	100		
A2	70	69	70	70	71	71	100		73	71	71	70	72	74	100	
A3	71	72	71	72	67	69	72	100	73	72	72	72	69	73	76	100
D1	100					$k=3$			100					$k=4$		
D2	78	100							78	100						
D3	74	76	100						75	77	100					
D4	71	70	70	100					72	71	70	100				
D5	68	73	71	65	100				70	74	71	64	100			
A1	68	69	66	68	67	100			70	70	68	69	70	100		
A2	71	71	71	70	72	73	100		73	72	71	70	72	75	100	
A3	72	73	72	73	69	72	75	100	74	74	73	73	70	74	77	100
D1	100					$k=5$			100					$k=6$		
D2	79	100							79	100						
D3	75	77	100						76	78	100					
D4	72	71	70	100					72	71	70	100				
D5	69	74	72	65	100				69	74	71	65	100			
A1	68	70	67	69	68	100			69	69	67	69	69	100		
A2	72	72	71	70	72	74	100		72	72	71	70	72	75	100	
A3	72	73	72	73	70	73	76	100	73	73	72	74	70	74	77	100
D1	100					$k=7$			100					$k=8$		
D2	79	100							79	100						
D3	75	78	100						76	78	100					
D4	72	71	70	100					72	71	70	100				
D5	69	75	72	65	100				69	75	72	65	100			
A1	69	70	67	68	69	100			69	70	67	69	68	100		
A2	72	73	71	70	72	75	100		72	72	71	70	72	75	100	
A3	72	74	72	73	70	74	77	100	73	74	73	74	70	74	77	100
D1	100					$k=9$			100					$k=10$		
D2	79	100							79	100						
D3	76	78	100						76	78	100					
D4	72	72	70	100					72	72	71	100				
D5	69	75	72	66	100				69	75	72	66	100			
A1	69	70	67	69	69	100			69	70	67	68	69	100		
A2	73	73	71	71	73	75	100		72	73	71	71	73	75	100	
A3	73	74	72	74	70	74	78	100	72	74	72	74	71	74	78	100

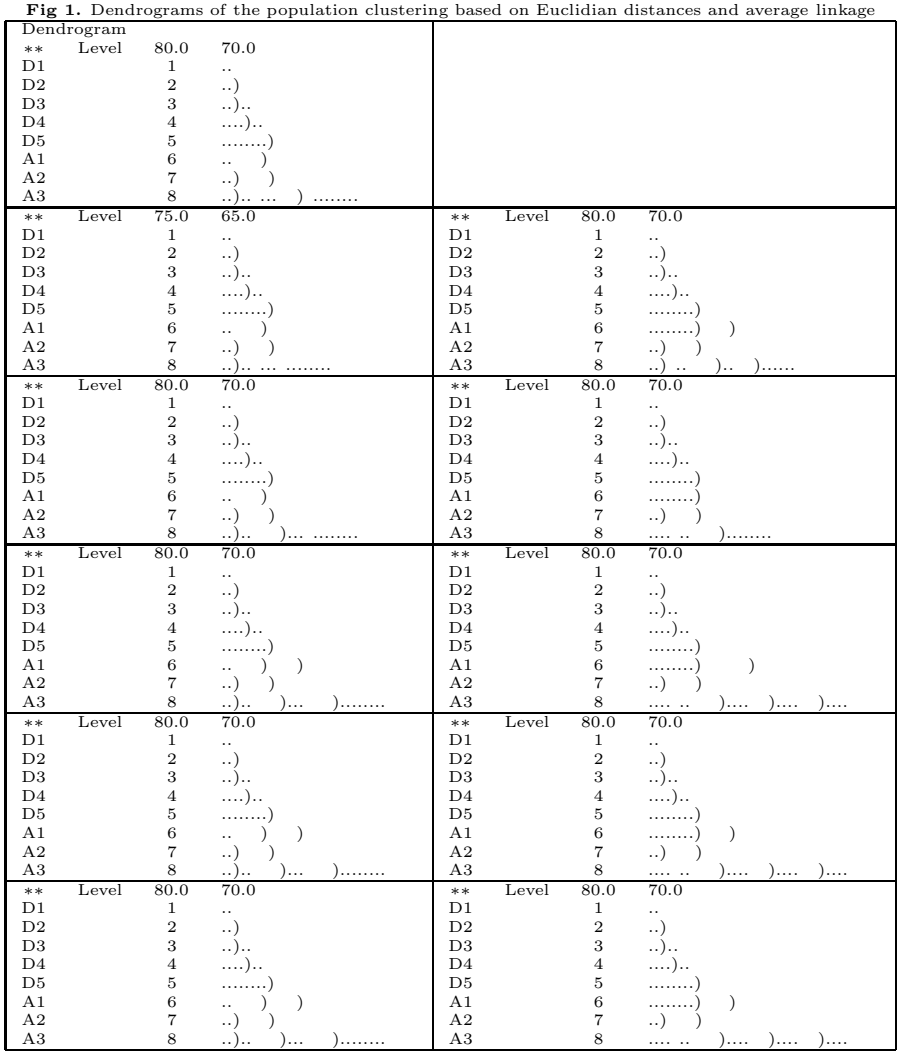
D1 to D5 : *T. dicoccoides* populations. A1 to A3 :*T. araraticum* populations.

Table 3 Estimated values of exponential relationship parameters and optimum values of sample size

		Estimated sample size		
		ε		
	Fitted function	0.05	0.10	0.20
LSS	$\omega_k = (0.0399 \pm 0.00125) + (0.0638 \pm 0.0112)(0.4891 \pm 0.071)^k$ $\bar{R}^2 = 94\%$	4.9	3.9	2.9
LSSS	$\omega_k = (0.4772 \pm 0.0580) + (0.2123 \pm 0.0309)(0.11337 \pm 0.00323)^k$ $\bar{R}^2 = 96\%$	4.9	4.0	3.0

LSS = Least-square-scaling stress measure. LSSS = Least-squares-squared-scaling stress measure. ω_k = LSS or LSSS for sample size k. $\bar{R}^2$ = percent variance accounted for . ε = relative difference in the asymptotic values of the stress measure.

Fig 1. Dendrograms of the population clustering based on Euclidian distances and average linkage



D1 to D5 : *T. dicoccoides* populations. A1 to A3 :*T. araraticum* populations.

Fig. 2 The least- squares-scaling (LSS) stress and least-squares-squared-scaling (LSSS) stress relationships with sample size