Selecting Accessions in Soybean Collection with High Diversity

RL Sapra*, SK Lal, Akshay Talukdar and KP Singh
Division of Genetics, Indian Agricultural Research Institute, New Delhi-110 012

The paper suggests a diversity efficient and computationally convenient procedure for selecting distinct accessions for breeding as well as core-set formation purpose. Soybean germplasm data comprising of 270 accessions, evaluated for seven important quantitative traits were used for the study. Thirty entries scoring the highest inertia values in individual clusters, when selected, resulted in highest pooled Shannon Diversity Index as well as coefficient of variability for individual characters.

Key words: Soybean, Diversity, Principal component

Plant breeders look for distinct and unique variability out of the large number of accessions available with them with an aim of increasing the level of expression of economic traits in the crossed combinations. Their primary concern is how to select the desired variability to carry out the breeding programme preferably with low number of accessions keeping in view the time and resource constraints. Thus, choice of parents for hybridization is one of the most critical factors in deciding the outcome. This is important because it determines the kind of variability expected in segregating generations, thereby setting the limits for improvement. The choice of parents largely depends upon the objectives. Various methods of selecting parents in *Triticum aestivum* have been discussed by Bhatt (1973). These methods are based on either ecogeographic diversity (Cox and Worall, 1987), character compensation (Grafius, 1965), early generation testing (Thurling and Ratinam, 1987) or measures of combining ability. These methods generally consider few parents. Many breeding programmes in the major crop species have reached a stage where yield improvement is slow or has almost reached a plateau. In this situation, a different strategy has to be adopted to increase the pace of improvement. A narrow genetic base among released cultivars and the practice of using elite line x elite line crosses has been implicated in slowing the rate of genetic advance for yield (Lal and Rana, 2000). There are two potential ways of overcoming the yield barrier. One is to look for traits/genes, which can enhance the adaptability of the cultivar in a specific production system/agro-climatic region. The other way is to use genetically diverse parents. Clustering germplasm into various groups using hierarchical or non-hierarchical algorithms based on multivariate statistical techniques, and sampling from within discrete groups is a common method for maximizing diversity. What and how many entries to be selected from each group, particularly when the population is large, is again cumbersome procedure. Various strategies have been proposed to select a number of entries from each group, particularly in relation to developing core sets (Brown, 1989; Schoen and Brown, 1995). In an empirical analysis, these authors ranked the various strategies according to the highest to the lowest expected allele retention and found the M Maximization) strategy to be superior. Principal Component Score Strategy (PCSS) suggested by Hamon and Noirot (1996), Noirot *et al.* (1996), Mahajan *et al.*, 1996), Srivastava *et al.*, (1999), Sapra and Lal (2003) and others, based on quantitative traits for selecting the accessions. Sapra and Lal (2003) discussed the issue of minimum sample size and estimated, under certain assumptions, the required target population size for including variability ranging from low to high. Inertia score has been used to select accessions to maximize the variability in the sample. The suggested approach slightly differs from above as it makes use of both i.e. the inertia score as well as is the grouping done based on *K-Means Clustering* for ensuring high and representative variability in the sample.

Materials and Methods

Data on 270 lines representing Indian as well as exotic variability. The crop was grown in 5 m rows during *kharif* (crop season beginning with the arrival of *monsoon* in July) 2004 in an augmented design (Federer, 1956) with five checks. Data was recorded on seven quantitative and four qualitative traits. However, the quantitative data was subjected to Principal Component Analysis (PCA) to calculate inertia scores (defined below). First three Principal Components with an Eigen value of one or more

* *Email:* saprarl@gmail.com

were extracted. The Relative Contribution (RC) for each accession was determined as follows:

Let Y_{ij} be the score through PCA for i^{th} (i = 1, 2 . . . N) accession and j^{th} (j = 1, 2 . . . k) principal component. The inertia of the i^{th} accession is defined as

$P_i = \Sigma_j Y_{ij}^2$ The relative contribution for the i^{th} accession is given by $RC_i = P_i /Nk$. Accessions were arranged in descending order of inertia.

The data was also subjected to classification using *K-Means Clustering* after applying the z- transformation. The number of clusters (k) was taken equal to the number that is approximately equal to the 10% of the collection size, a number normally required for developing germplasm core set for retaining at least 70% or more allelic diversity in the sample. In our case 30 clusters were formed. The quantitative data for all the characters was converted into qualitative one by dividing the individual character range into five equal intervals. The pooled Shannon Diversity Index (pooled over the characters) was chosen as the measure of diversity and for comparison purpose.

Proposed Selection Procedure

(i) Select a single entry from each cluster whose inertia score is the highest in the cluster.

(ii) Count the frequencies of each character state and calculate the individual SDI for all the characters. Pooled SDI is the sum of individual SDIs.

Results and Discussions

The first three principal components could explain 71% of the total variation, whereas the first two components around 56%. Graph 1 and 2 give the spread in a two dimensional space of all the 270 data points as well as the points related to 30 accessions selected from 30 clusters and top 30 points in terms of inertia score respectively. If we examine the top 30 points in Graph 2, it is clear that these points are distantly located from the centre and hence play significant role in the divergence and that is why the top entries have higher values of diversity index as compared to that of the whole collection (Table 2). The points, which are located near the centre of the graph, have low inertia score and contribute less to the diversity. The pooled diversity decreases from top to bottom i.e., from 9.070 to 5.732 (Table 2). However, the diversity of individual characters does not show a declining trend for all the characters when we move from top to bottom (Table 2). For example the middle 30 accessions show lower values of SDI for plant height and pods/plant as compared to those of bottom accessions. The basic objective here is to look for those points, which can give higher diversity than that of the top 30 entries.

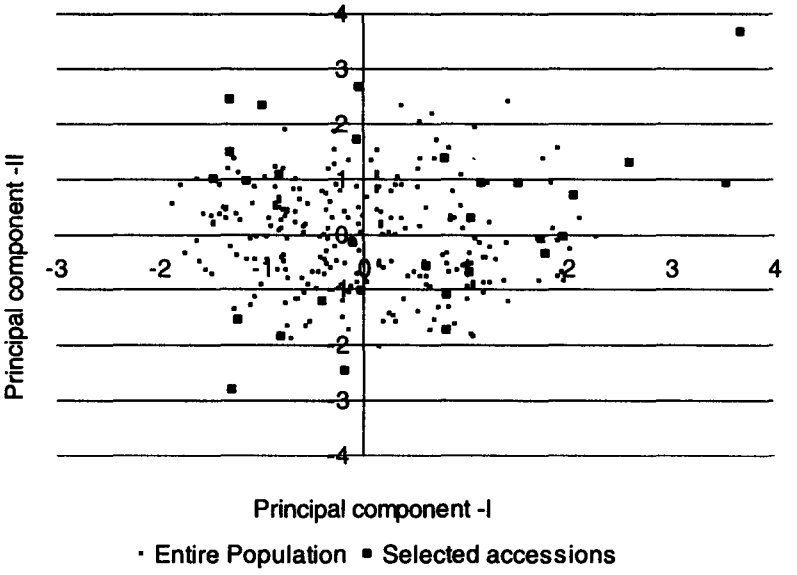
Graph 1 indicates the 30 points having the highest inertia and located in the 30 disjoint-clusters. On comparing the two graphs we find that nearly 60% of the points are common. Some of the points have come closer to the centre. *K-Means Clustering* divides the 270 objects into 30 clusters such that some metric relative to the centroids of the clusters is minimized. Thus, the selected points represent the homogenous clusters, are bound to give better representation of the variability in the sample. If we compare the values of SDIs between top 30 entries and clusters points, the values are higher for all the characters except for plant height. The sample of cluster points resulted in the highest pooled SDI (9.463). Table 3 gives a comparative view of the variation in terms of range and coefficient of variation (%) for the entire collection and cluster points. The range for individual characters is either equal to that of entire collection or almost approaching it. The CV (%) values are higher for all the traits for cluster-selected-points. Thus, selection based on inertia score and clustering could be an appropriate choice if the objective is to maximize the diversity for a given sample size. This procedure is diversity efficient, computationally convenient and

Table 1. Eigen value and variation

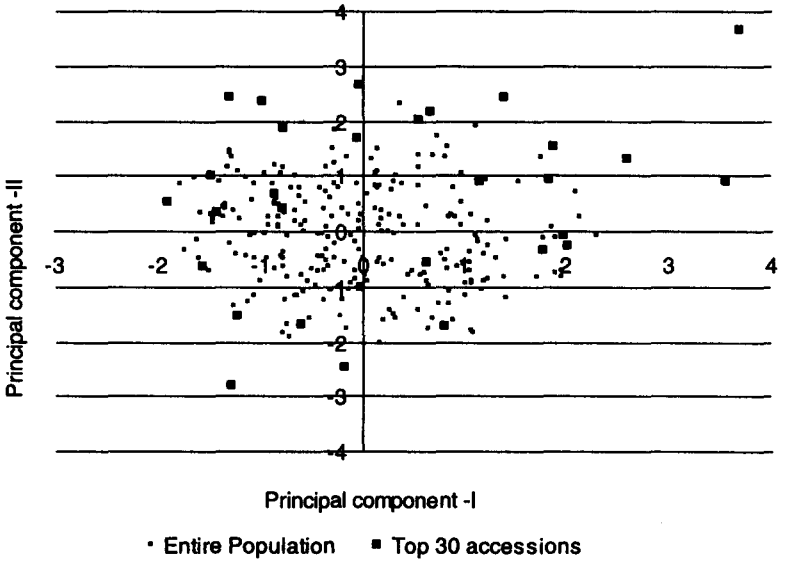
Principal Component	Eigen value	Variance (%)	Cumulative %
I	2.55	31.872	31.87
II	1.917	23.96	55.83
III	1.214	15.171	71

Table 2. Shannon Diversity Index for various samples, and entire collection for individual traits

	Days to 50% flowering	Days to maturity	Plant height	Number of branches	Number of pods/plant	Number of seeds	Yield per plant	Pooled SDI
Population	1.377	1.316	1.480	1.224	0.897	1.156	0.994	8.444
Top 30	1.448	1.320	1.431	1.328	1.111	1.200	1.232	9.070
Middle 30	0.976	1.239	1.067	1.012	0.543	1.198	0.730	6.766
Bottom 30	0.887	0.822	1.198	0.722	0.703	0.778	0.623	5.732
Clusters points	1.407	1.458	1.367	1.477	1.152	1.206	1.395	9.463



Graph 1: *Spread of entire population and accessions from 30 clusters in a two dimensional space*



Graph 2: *Sperad of entire population and top 30 accessions in a two dimensional space*

straight forward. The procedure can be easily adopted for large collections, particularly, where the information on geographical diversity is lacking or unavailable. Selection of entries from the non-hierarchical groups is not optional but logical as the entry with the highest inertia is included in the sample.

References

Bhatt GM (1973) Comparison of various methods of selecting parents for hybridization aimed at yield improvement in self-pollinated crops. *Australian Journal of Agriculture Research* **24**: 457-464.

Bisht IS, RK Mahajan, TR Loknathan, PL Gautam, PN Mathur and T Hodgkin (1999) Assessment of genetic diversity, stratification of germplasm accessions in diversity groups and sampling strategies for establishing a core collection of Indian sesame (*Sesamum indicum* L.). *Plant Genetic Resources Newsletter* **119**: 35-46.

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

Cox TS and WD Worall (1987) Electrophoretic variation among and within strains of *'Kharif'* Wheat maintained at 11 locations. *Euphytica* **36**: 815-822.

Federer TW (1956) Augmented (or Hoonuiaku) designs. *The Hawaiian Planter's Record* **IV**: 191-202.

Table 3. Showing variation in quantitative traits for the entire collection and cluster selected points (bold)

	Minimum	Maximum	Mean	CV(%)
Days to 50%	33	55	44.10	9.41
Flowering	35	55	43.67	12.36
Days to maturity	87	115	100.23	5.54
	87	115	99.63	8.30
Plant Height (cm)	18.67	97	61.01	27.60
	20.00	90	56.95	37.57
Number of Branches	0.67	9	4.13	33.13
	0.67	9	4.81	43.13
Number of Pods/plant	6.5	171.5	41.69	45.49
	6.5	171.5	51.17	63.43
Seeds/pod	1.2	2.73	1.99	12.47
	1.4	2.53	2.01	14.84
Yield/plant	0.42	22.43	4.92	72.95
	0.42	22.43	6.66	87.05

Grafius JE (1965) A Geometry of Plant Breeding. *Michigan State University Research Bulletin.* East Lansing: Michigan State University, 59 p.

Hamon S and M Noirot (1996) Some proposed procedure for obtaining a core collection using quantitative plant characterization data. Paper presented at the International Workshop on Okra, NBPGR, New Delhi, India, October 8-12.

Lal SK and VKS Rana (2000) Genetic enhancement of soybean. *Indian J. Genet* **60**: 483-494.

Mahajan RK, IS Bisht, RC Agrawal and RS Rana (1996) Studies on a South Asian okra collection: methodology for establishing a representative core set using characterization data. *Genetic Resources and Crop Evolution* **43**: 249-255.

Noirot M, S Hamon and F Anthony (1996) The principal component scoring: a new method of constituting a core collection using quantitative data. *Genetic Resources and Crop Evolution* **43**: 1-6.

Sapra RL, P Narain and SVS Chauhan (1998) A general model for sample size determination for collecting germplasm. *Journal of Bioscience* **23**: 647-652.

Sapra RL, P Narain, SVS Chauhan, SK Lal and BB Singh (2003) Sample size for collecting germplasms - a polyploid model with mixed mating system. *Journal of Bioscience* **28**: 155-161.

Schoen DJ and Brown AHD (1995) Maximizing genetic diversity in core collections of wild relatives of crop species. Hodgkin T, Brown AHD, van Hintum Th JL and Morales EAV (eds) Core Collections of Plant Genetic Resources. Chichester: John Wiley 55-76.

Srivastava, Umesh, Sapra RL, Chauhan SVS and Narain P (1999) Studies world tomato (*Lycopersicon esculentum* Mill.) germplasm collection: Methodology for establishing a representative core set using characterization data. *Indian J Plant Genet Resour.* **12**: 290-301.

Thurling N and M Ratinam (1987) Evaluation of parents selection methods for yield improvement of cowpea (*Vigna unguiculata* L. Walp.). *Euphytica* **36:** 913-926.