© NIAB 2003
ISSN 1479-2621

Plant Genetic Resources 1(2-3); 151–156
DOI: 10.1079/PGR200322

A strategy for selecting diverse accessions using Principal Component Analysis from a large germplasm collection of soybean

R. L. Sapra* and S. K. Lal

Pulse Research Lab, Genetics Division, Indian Agriculture Research Institute, Pusa, New Delhi 110012, India

Received 25 June 2003; Accepted 8 December 2003

Abstract

We suggest a diversity-dependent strategy, based on Principle Component Analysis, for selecting distinct accessions/parents for breeding from a soybean germplasm collection comprising of 463 lines, characterized and evaluated for 10 qualitative and eight quantitative traits. A sample size of six accessions included all the three states, namely low, medium and high of the individual quantitative traits, while a sample of 16–19 accessions included all the 60–64 distinct states of qualitative as well as quantitative traits. Under certain assumptions, the paper also develops an expression for estimating the size of a target population for capturing maximum variability in a sample three accessions.

Keywords: breeding; diversity; germplasm; Principal Component Analysis; sampling; soybean; inertia

Introduction

One of the major issues that plant breeders face is the need to look for distinct and unique variability with the aim of increasing the level of expression of economic traits. For carrying out the breeding programme, the number of accessions to be assessed needs to be minimized. However, at the same time it is preferable to retain maximum variability in order to generate large numbers of combinations. In any breeding programme, 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 estivum* 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 × 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. This procedure is efficient when the groups are well defined and non-overlapping, and within-group variability is low. However, when the number of lines is large, it is often difficult to satisfy these criteria. Various strategies have been proposed to select a number of entries from each group, particularly in

* Corresponding author. E-mail: saprarl@rediffmail.com

https://doi.org/10.1079/PGR200322 Published online by Cambridge University Press

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. The simplified method which we suggest in this paper uses the Principal Component Score Strategy (PCSS) suggested by Hamon and Noirot (1996), Noirot *et al.* (1996), Mahajan *et al.* (1996), Umesh *et al.* (1999) and others, based on quantitative traits. We have calculated an inertia score on the basis of the quantitative traits. This inertia is further used for selecting the quantitative as well as qualitative traits. This approach ensures maximum diversity in the selected accessions while keeping the sample size as minimum as possible. The paper also discusses theoretically the issue of minimum sample size and estimates, under certain assumptions, the required target population size for including variability ranging from low to high.

Materials and methods

Data on 463 lines representing Indian as well as exotic variability were subjected to Principal Component Analysis (PCA) based on quantitative traits to calculate an inertia score (defined below). The crop was grown in 5 m rows during kharif (crop season beginning with the arrival of monsoon in July) 2000 in an augmented design (Federer, 1956) with five checks. Observations were made for the eight quantitative and 10 qualitative traits listed in Table 1.

An observation for each trait represented the mean of 10 randomly selected plants. Three independent analyses were undertaken to validate and verify the results, namely (i) the entire set of 463 lines, (ii) a random sample of population of 231 of the 463 lines and (iii) the remaining 232 lines. In each case the first three principle components with an eigenvalue of 1 or more were extracted. The Relative Contribution (RC) for each accession was determined as follows:

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

$$P_i = \sum_{j=1} Y_{ij}^2$$

The relative contribution for the ith accession is given by $RC_i = P_i/Nk$.

The data for all the samples were arranged in descending order of relative contribution in order to select the entries (accessions) in the following manner. The quantitative data were converted into code data in three states: low (1), medium (2), high (3). Inclusion of entries was based on the score values beginning from the first entry

Table 1. List of traits used for morphological characterization

Character	States	Code
1. Early plant vigour	Poor	1
	Good	2
	Very good	3
2. Days to 50% flowering		Quantitative
3. Flower colour	White	1
	Light purple	2
	Purple	3
4. Leaf shape	Broad	1
	Intermediate	2
	Narrow	3
5. Leaflet colour	White	1
	Light green	2
	Green	3
	Dark green	4
6. Number of leaflets	Low	1
	Medium	2
	High	3
7. Pubescence colour	Grey	1
	Light tawny	2
	Tawny	3
8. Plant height		Quantitative
9. Number of primary branches		Quantitative
10. Number of pods per plant		Quantitative
11. Days to 80% maturity		Quantitative
12. Seed coat colour	Yellowish white	1
	Yellow	2
	Green	3
	Buff	4
	Reddish brown	5
	Grey	6
	Imperfect black	7
	Black	8
13. Hilum colour	Yellow	1
	Buff	2
	Brown	3
	Green	4
	Grey	5
	Imperfect black	6
	Black	7
14. Seed coat surface lustre	Shiny	3
	Intermediate	5
	Dull	7
	Heavy bloom	9
15. Number of seeds per pod		Quantitative
16. Seed shattering	Absent	0
	Present	1
17. Weight of 100 seeds		Quantitative
18. Seed yield per plant		Quantitative

https://doi.org/10.1079/PGR200322 Published online by Cambridge University Press

having the highest score. If a particular entry did not add a state, it was ignored and the next entry was considered for inclusion. The procedure was continued until all the states for each quantitative trait were included in the sample.

After selecting the entries for the quantitative traits, the states of the qualitative traits were checked. If the sample included all the qualitative states, the procedure was terminated; otherwise the same procedure was followed as above for quantitative traits until all the states were included.

Minimum sample size

The minimum sample size depends on whether the sample is to be used for breeding or to be used for conserving the variability by developing a core set. In the former case, the sample size needs to be smaller and is primarily based on quantitative traits, whereas in the latter case, it will be larger and should ideally represent the entire spectrum of variability containing at least a copy of all the alleles present. For example, if a breeder carries out a crossing programme based on 10 morphological traits that can be categorized in three states (i.e. low, medium and high), then to capture all three states of each trait a minimum sample size will be three: one accession with states 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 (first state of each trait), the second 2, 2, 2, 2, 2, 2, 2, 2, 2, 2 and the third 3, 3, 3, 3, 3, 3, 3, 3, 3, 3. Clearly there is a large number of possible combinations of triplets including all the states of the individual traits. The probability of obtaining such a triplet of accessions can be predicted as follows:

Let d be the number of descriptors each having three states.

Total number of possible genotypic combinations $(G) = 3^d$. From these G genotypes, one can get $(G)^3$ triplets. However, the desired number of triplets (triplets containing all the three states of each trait) will be: $3^{d-1}2^{d-1}$. Thus, the probability $p0$ of obtaining a desired triplet is

$$3^{d-1}2^{d-1}/(G)^3 = 2^{d-1}/3^{2d+1}. \quad (1)$$

For 10 descriptors, $p_0 = 2^9/3^{21}$, so the chance of selecting at random such a triplet is remote. Of course, where traits (descriptors) are associated with one another because of genetic linkage and/or pleiotropy, the sample size will be reduced.

Target population size

To estimate the minimum size population to obtain at least one desired triplet of accessions with a probability p, we have followed the approach of Sapra *et al.* (1998, 2003). Assuming that the occurrence of all the genotypic combinations (G) is equally likely, and letting N be the population size, then the total number of triplets (T) will be ${}^{N}C_3$, i.e. $N(N-1)(N-2)/6$.

Since the probability of not obtaining a desired triplet in T triplets is $(1-p_0)^T$, the probability p of obtaining at least one desired triplet is $1-(1-p_0)^T$, which reduces to the relationship $T = \log(1-p)/\log(1-p_0)$. Thus $N(N-1)(N-2)/6 = \log(1-p)/\log(1-p_0)$, and if N is large, then

$$N \approx \sqrt[3]{6\{\log(1-p)/\log(1-p_0)\}}. \quad (2)$$

Results and discussion

Expression (1) indicates a very low probability ($2^{d-1}/3^{2d+1}$) of finding a desired triplet of accessions and is heavily dependent on the number of descriptors. Therefore, selecting desirable accessions through random sampling is likely to be unfruitful. Our calculations suggest that the probability for eight descriptors is almost 20 times higher than that for 10 descriptors. However, based on expression (2) and Table 2, the estimated minimum target population size for the desired triplet of accessions for 10 descriptors is almost three times that of eight descriptors with a very high degree of certainty (0.9999) under certain assumptions. The estimated target population size (380) for eight descriptors was less than the available population size (463). Therefore, an attempt was made to search for the desired triplet by writing a computer program to generate all the desired genotypic combinations. However, this attempt was not successful, indicating that the desired triplet(s) was missing in the population under consideration, and suggesting that more accessions were required to be sampled to include all the three states of the eight quantitative traits. We did not proceed further in searching more than three accessions, as there may be several such sets of accessions, and choosing a particular set might again be problematical. Alternatively, we have resorted to the PCA technique, which can make full use of the quantitative data.

Table 2. Target population size for two levels of descriptors (8 and 10) at three levels of probability (0.95, 0.99, 0.9999) for getting at least one desired triplet

	Estimated target population size	
Probability	Eight descriptors	Ten descriptors
0.95	262	702
0.99	302	810
0.9999	380	1020

https://doi.org/10.1079/PGR200322 Published online by Cambridge University Press

Table 3. Principal component, eigenvalue and the cumulative % of variance for the three populations

	Entire population (463)		Population I (231)		Population II (232)	
Component	Eigenvalue	Cumulative variance (%)	Eigenvalue	Cumulative variance (%)	Eigenvalue	Cumulative variance (%)
1	3.513	43.917	3.477	43.466	3.606	45.071
2	1.140	58.168	1.294	59.641	1.253	60.735
3	1.118	72.143	1.059	72.878	0.980	72.988

Table 4. Mean and coefficient of variation (CV%) for the three populations

	Entire population (463)		Population I (231)		Population II (232)	
Descriptor	Mean	CV%	Mean	CV%	Mean	CV%
DAY_FLW	45.22	10.77	45.16	11.20	45.28	10.33
DAY_MAT	100.83	4.57	100.82	4.57	100.85	4.59
PLT_HT	61.72	32.97	61.68	34.43	61.77	31.52
POD_PLT	93.67	47.44	94.43	46.03	92.92	48.92
PRI_BRN	6.04	30.09	6.04	30.63	6.04	29.57
SED_POD	2.08	10.38	2.06	11.16	2.10	9.51
SED_WT	7.72	22.62	7.85	22.80	7.60	22.38
SED_YLD	12.33	43.00	12.71	44.14	11.97	41.47

The first principal component accounted for more than 43%, and the three major components accounted for more than 72% of the variation for all the three populations (Table 3). The mean and coefficient of variation for the eight quantitative traits are shown in Table 4. To follow trends in the variation, the data were arranged in decreasing order of relative contribution. For each population, groups of 30 accessions were formed in a sequential manner and average coefficients of variation were determined (Table 5). In general, a declining trend (except in a few cases) in average coefficient of variation was observed as the group number increased among all the three populations (Fig. 1). The top 30 entries showed the highest average coefficient of variation, whereas the bottom 30 entries showed the lowest average coefficient of variation among all the populations (Table 5). Thus the bottom entries are less variable than the top entries.

Table 5. Average coefficient of variation for the three populations arranged in descending order of relative contribution for different groups of sizes 30 each

Group no.	Entire population (463)	Population I (231)	Population II (232)
1	40.39	37.98	36.23
2	34.09	30.83	24.90
3	28.09	26.75	26.68
4	29.33	21.49	23.47
5	26.70	21.54	22.21
6	25.75	17.9	19.82
7	21.52	17.57	17.68
8	23.60		
9	23.42		
10	19.38		
11	17.65		
12	17.79		
13	16.29		
14	15.63		
15	13.98		

This study indicates that as few as six accessions (among the nine top accessions arranged in descending order of relative contribution or inertia) included all the 24 quantitative states in all the three populations. With an additional set of 10–13 accessions, selected according to the procedure described above, all the states (60–64) were included in the sample. Thus, the sample forms a minimal yet diverse set. Since the selection procedure is based on the inertia or relative contribution, the method ensures a high degree of quantitative diversity in the sample. As all the states have been included in the sample, the procedure should give a pooled diversity

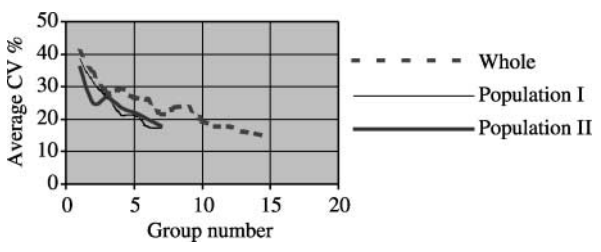


Fig. 1. Relationship between average coefficient of variation (CV %) and various groups.

https://doi.org/10.1079/PGR200322 Published online by Cambridge University Press

Table 6. Sample size, pooled Shannon Diversity Index (SDI, pooled over 10 qualitative traits) and number of qualitative states included in the sample for the three populations

	Entire population (463)		Population I (231)		Population II (232)	
Sample size	Pooled SDI	No. of states	Pooled SDI	No. of states	Pooled SDI	No. of states
15			3.146	26	2.951	25
20			3.335	28	3.149	27
25			3.342	30	3.359	29
30	3.321	31	3.481	31	3.369	29
35	3.357	32	3.503	31	3.364	30
40	3.374	32	3.423	31	3.484	30
45	3.444	32	3.491	32	3.433	30
50	3.455	32	3.545	33	3.468	30
55	3.633	32	3.551	33	3.450	30
60	3.567	32	3.520	35	3.509	30
65	3.546	32	3.569	35	3.243	31
70	3.548	32	3.579	35	3.371	31
All cases	3.625	40	3.644	40	3.563	40

index higher than that of the whole population. Sample size is affected by the amount of diversity (number of distinct states/alleles) present in the population and not the population size. Only those accessions which make a significant relative contribution in terms of variability are included in the sample.

The criterion 'pooled diversity index' has been used by other workers for inclusion of entries in a sample (e.g. Bisht *et al.*, 1999, for the development of an Indian sesame germplasm core set). However, as Table 6 suggests, the pooled diversity index may not give a reliable picture of the diversity. For example a pooled Shannon Diversity Index for a sample size of 55 is 3.633, which is higher than that of the whole collection (3.625). Moreover, this sample would include only 32 out of the possible 40 descriptor states.

Conclusion

This paper proposes a strategy to select accessions for use in a breeding programme. The idea is to form three classes per trait and then select accessions so that the two-way classification of classes × traits is as complete as possible with a minimum number of accessions. A forward-selection approach using inertia score based on quantitative traits was followed for maximum variability (qualitative and quantitative) in the sample. The sample selected through the suggested procedure is primarily diversity-dependent and depends on the number of traits, their distinct states/alleles and is independent of the population size. However, any correlation between traits will affect sample size. A sample size of six accessions included all the three states, namely low, medium and high of the individual quantitative traits, while a sample of 16–19 accessions included all the 60–64 distinct states of qualitative as well as quantitative traits. The chance of selecting a desired triplet of accessions through random strategy is very remote. Under certain assumptions, we have also developed an expression for estimating the size of a target population for capturing maximum variability in a sample of three accessions. We believe this selective approach, based on inertia score, will ensure higher diversity with minimum sample size, particularly suited for the purpose of plant breeding.

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, Mahajan RK, Loknathan TR, Gautam PL, Mathur PN and Hodgkin T (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 Worall WD (1987) Electrophoretic variation among and within strains of 'Kharkof' Wheat maintained at 11 locations. *Euphytica* 36: 815–822.

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

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

Hamon S and Noirot M (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, pp. 8–12.

https://doi.org/10.1079/PGR200322 Published online by Cambridge University Press

Lal SK and Rana VKS (2000) Genetic enhancement of soybean. *Indian Journal of Genetics* 60: 483–494.

Mahajan RK, Bisht IS, Agrawal RC and Rana RS (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, Hamon S and Anthony F (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, Narain P and Chauhan SVS (1998) A general model for sample size determination for collecting germplasm. *Journal of Bioscience* 23: 647–652.

Sapra RL, Narain P, Chauhan SVS, Lal SK and Singh BB (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. In: Hodgkin T, Brown AHD, van Hintum ThJL and Morales EAV (eds) *Core Collections of Plant Genetic Resources*. Chichester: John Wiley, pp. 55–76.

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

Umesh S, Sapra RL, Chauhan SVS and Narain P (1999) Studies in world tomato (*Lycopersicon esculentum* Mill.) germplasm collection: methodology for establishing a representative core set using characterization data. *Indian Journal of Plant Genetic Resources* 12: 290–301.

https://doi.org/10.1079/PGR200322 Published online by Cambridge University Press