Strategies for Developing Core Collections of Safflower (*Carthamus tinctorius* L.) Germplasm – Part I. Sampling from Diversity Groups of Quantitative Morphological Descriptors

KK Suresh and R Balakrishnan
Department of Statistics, Bharathiar University, Coimbatore-641046, Tamil Nadu

Evaluation data collected from a germplasm collection of 3,250 safflower (*Carthamus tinctorius* L.) accessions was used to divide the whole collection into six major clusters based on multivariate cluster analysis. The clusters were further grouped into 30 diversity groups based on the geographical origin of the accessions and the plant types. Estimates of the phenotypic diversity of the core samples were obtained through simple random sampling and 5 stratified random sampling strategies. They were compared for varying sampling fractions ranging from 5% to 25% of the whole collection. In the core samples obtained through simple random sampling or stratified random sampling by frequency proportion method or diversity proportion method, the pooled diversity index based on 28 descriptors was close to the diversity of the whole collection. However, when the accessions from different diversity groups were allocated with equal frequency or in proportion to the logarithm of the number of accessions in each group or in proportion to the square root-proportion of the number of accessions in each group, the resultant core samples had higher levels of diversity than the whole collection. About 10-15% of the whole collection was found to be adequate as the core sample size.

Key Words: Core Sample, Information Measure, Kullback Divergence Statistics (KDS), Safflower, Shannon Diversity Index (SDI), Stratified Sampling

The potential use of plant genetic resources contained in large germplasm collections could be greatly enhanced by constituting sub-samples called core collections. The broad guideline for constituting a core sample as suggested by Frankel (1984) is that, it should include the maximum of genetic variation contained in the whole collection with minimum duplication. Obviously, the quality of the core sample is dependent upon good passport and evaluation data on the accessions that constitute the whole collection. Constituting a core sample of reasonable size and sufficient diversity by making use of available data on the accessions has been one of the most important issues of germplasm management. The basic issues that are of primary concern are (i) Core sample size, and (ii) Sampling strategies that could help in selecting accessions that reproduce the variation for several characters in the whole collection to the maximum possible extent. Brown (1989 a, b) addressed the issue of optimum core sample size and suggested that about 10 per cent of the whole collection would allow preservation of about 80% of alleles in a large collection. Methods for obtaining core samples using different sampling strategies had been suggested by several workers. They were mainly concerned with grouping of accessions into homogenous groups or clusters and selecting sub samples from each group to obtain a pooled core sample. The grouping approaches described were hierarchical (Hintum TJL Van, 1995; Peeters and Martenelli, 1989) or non-hierarchical cluster analysis methods using quantitative or a mixture of both quantitative and qualitative descriptors (Spagnoletti Zeuli and Qualset, 1993; Mahajan *et al.*, 1996; Harch *et al.*, 1996; Bisht *et al.*, 1998). Grouping of the accessions based on their geographical origin had also been suggested by several authors. The most common method of sampling is simple random sampling from each group to obtain a core sample of desired size. Several strategies had also been suggested for deciding appropriate sampling fraction from each group or strata. These methods included proportional allocation, log frequency allocation, square-root frequency proportion allocation etc. (Brown 1989b; Spagnoletti Zeuli and Qualset, 1993; Mahajan *et al.*, 1999; Balakrishnan and Suresh, 2000). As an alternative to simple random sampling, Noirot *et al.*, (1996) suggested that the accessions could be ranked on the basis of their relative contribution to the overall variance and a desired proportion of top ranked accessions could be selected from each group to constitute the core sample. This approach was found to be very useful not only in obtaining core samples of reasonable size, but also the diversity in terms of several qualitative descriptors were almost equal to that of the whole collection (Mahajan *et al.*, 1996; Balakrishnan *and Suresh* 2000).

In the present investigation, the diversity in a large germplasm collection of 3,250 safflower accessions has been studied for a set of 18 qualitative and 10 quantitative descriptors. These accessions were grouped into 6 major clusters based on the morphological and agronomic characters. These major clusters were further divided into 30 diversity groups based on the geographical origin and plant type of the accessions. Simple random sampling and stratified random sampling with 5 methods of group allocation in the core samples were used to compare the diversity of the core sample with that of the whole collection. An information measure was computed to study how far the resultant core sample deviated from the joint density distribution of the whole collection with respect to the set of descriptors. A measure of divergence of the core sample from the reserve collection or the non-core group was also evaluated under the various sampling strategies.

Materials and Methods

The data set on safflower (*C. tinctorius* L.) pertains to a collection of 3,250 accessions that were grown and evaluated at the Germplasm Management Unit of the All India Coordinated Research Project on Oilseeds at Solapur, Maharashtra, India (latitude: 17°14' N and longitude: 75°56' E). The majority of these accessions were from India and the remaining accessions from various other countries that were obtained from the world genebank through the courtesy of late Dr. PF Knowles and which were later assembled from different coordinating centres at the gemplasm management unit at Solapur. The material was raised during 1987-1990 in single rows of 6m length spaced at 45 cm between rows and 35 cm within rows in augmented block design under protective irrigation. Ghorpade *et al.*, (1991) had published a comprehensive catalogue of the passport and evaluation data of these accessions. For the present investigation, ten quantitative and 18 qualitative traits from this data set were considered for the purpose of constituting the core set from this germplasm collection.

The list of descriptors used for the present investigation is provided in Table 1. The number of accessions from different geographical regions varied widely, the majority of the accessions coming from India (2444) followed by United States of America (330). There were far lesser number of accessions from other countries and they are presented in Table 2.

Table 1. List of descriptors used for the study

1. Shape of lower stem leaf : Lanceolate narrow, lanceolate broad
2. Margin of lower stem leaf: Serrate, deeply serrate, entire, deeply lobed
3. Primary head shape before flowering: Conical, flattened, oval
4. Texture of upper leaves: Fleshy, normal, leathery
5. Shape of upper stem leaf: Lanceolate narrow, lanceolate broad, oblong
6. Margin of upper stem leaf: Entire, serrate, slightly serrate, few serrate, deeply lobed
7. Spines on upper stem leaf: Non-spiny, few spines, intermediate, many spines
8. Attitude of OIB* to head: Closed, open
9. OIB* cross section shape: Flat, grooved
10. Location of spines on OIB*: None, tip only, tip & few basal, tip & few apical, tip & all along margin
11. Number of spines on OIB*: None, intermediate, many
12. Length of spines on OIB*: None, short, intermediate, long
13. Bracts enclosing head: Complete, incomplete
14. Growth habit: Bushy, cone shaped, appressed, erect
15. Branch location on main stem: Primarily basal, upper 1/3, upper 2/3, base to apex
16. Pollen production: Sparse, intermediate, abundant
17. Pappus on the acheme: Present, absent, negligible
18. Hull thickness: Thin/ intermediate, thick
19. Days to 50% elongation: 20-23, 23-26, etc.,..... 38-41, > 41
20. Days to primary branch initiation: 30-35, 35-40, etc., ... 55-60, > 60
21. Days to first flowering: <= 55, 55-60, etc., ...95-100, >100
22. Days to 50% flowering: <= 65, 65-70, etc., ... 105-110, > 110
23. Days to physiological maturity: 70-80, 80-90, etc., ...140-150, > 150
24. Plant spread (cm): 10-20, 20-30, etc.,..... 80-90, > 90
25. Number of primary branches: <= 3, 3-6, etc.,..... 24-24, > 27
26. Number of capitula per plant: 0-10, 10-20, etc., ... 90-100, >100
27. Mean inter-node length (cm) : 1, 2, etc., ...,10
28. Main Capitula diameter (cm): 0.5-1.0, 1.0-1.5, etc., ...3.0-3.5, >3.5

* OIB: Outer Invocular Bracts

Cluster analysis

A random sample of 325 accessions (10%) from the whole collection was initially subjected to cluster analysis using Ward's method. For the purpose of cluster analysis, the attribute data for the qualitative descriptors were converted into numerical scores based on the method suggested by Balakrishnan and Sanghvi (1968). These numerical scores depend on the dispersion matrix of

Table 2. Region-wise breakup of accessions under major clusters of safflower germplasm collection

Geo. Region	Major clusters						
	1	2	3	4	5	6	Total
IND	795	1298	88	33	225	5	2444
USA	3	125	41	66	84	11	330
IRN	0	9	10	7	22	14	62
PAK	3	20	3	2	37	2	67
SUN	1	4	4	0	21	2	32
EUR	0	10	2	2	17	1	32
ESP	0	4	11	0	14	1	30
AFR	0	23	1	0	17	3	44
EGY	0	8	11	2	7	0	28
TUR	0	18	9	7	25	10	69
M.E.	0	9	6	4	7	7	33
AUS	0	12	1	2	3	0	18
Others	0	20	18	1	20	2	61
Total	802	1560	205	126	499	58	3250

IND: India and Bangladesh
USA: USA, Canada and Mexico
IRN: Iran and Iraq
PAK: Pakistan and Afghanistan
SUN: Russia, Germany and Poland
EUR: Switzerland, France, United Kingdom, Hungary and Italy
ESP: Spain and Portugal
AFR: Algeria, Ethiopia, Kenya, Libya, Morocco and Sudan
EGY: Egypt
TUR: Turkey
M.E: Israel, Syria, Lebanon and Jordan
AUS: Australia

the individual descriptors. The dispersion matrix was computed for each descriptor by pooling the frequency densities of the attribute values over 13 different geographical regions. The method is more objective in that it preserves the distance among various groups and avoids arbitrary scaling of the qualitative attributes into numerical values. The values for the 10 quantitative traits were standardized before subjecting them to cluster analysis. An initial set of 6 clusters was identified from this sample of 325 accessions. The remaining accessions were allocated to these clusters based on two methods, *viz.*, (i) The Euclidean distance method and (ii) the classification scheme of Boulton and Wallace (1970). For the Euclidean method, the cluster centres were computed and the Quick Cluster procedure of SPSS software package was used to allocate the remaining accessions into the initial clusters. For the Boulton-Wallace method, the joint density distributions of the descriptors under the 6 initial clusters were computed and any accession that had the highest probability of being nearest to any cluster was assigned to that cluster. An information measure (Wallace and Boulton, 1968) was also computed to assess the better of the two classification schemes. Based on this, the clustering scheme of Boulton and Wallace was considered for the data set of 3250 safflower accessions.

The number of accessions in each of the six main clusters and their distribution over different geographical regions is presented in Table 2. For further dividing these major clusters into diversity groups, geographical origin of these accessions (whether they belong to India, USA or others) and the plant growth type (bushy, cone shaped, erect or appressed) were taken into consideration. The diversity groups are presented in Table 3.

Diversity index

For computing the diversity index using the Shannon formula, the numerical descriptors were converted into appropriate class intervals and each class interval was treated as a descriptor state. For the i^{th} descriptor if P_{ij} is the proportion of its j^{th} attribute state in the population, then the population Shannon Diversity Index (SDI) for that descriptor can be computed using the formula:

$$SDI_i = -\sum_{d=1}^{s} P_{ij} * Log_e(P_{ij})$$

A pooled diversity index SDI across all the 28 descriptors was then computed by adding the SDI values for the individual descriptors. Similarly in case of computing the pooled diversity index for a core sample of a given size, the same formula was used by replacing the population proportion P_{ij} with the sample proportion P_{ij} for a given descriptor state.

Estimation of mean and variance of the pooled SDI through sampling

For obtaining core samples, 5 different sizes were considered. They were approximately fixed at 5, 10, 15, 20 and 25% of the whole collection and were respectively 150, 330, 480, 660 and 810. For drawing samples from each of the 30 diversity groups, simple random sampling and stratified random samples were considered. For drawing the samples through stratified random sampling, five methods of allocation were considered. They were:

1. Frequency proportion (FP method): In this method, the number of accessions drawn from any group was in proportion to the group size.
2. Square root -proportion (SQP method): In this method if P_i was the proportion of accessions from

i^{th} group in the whole collection, then q_i, the proportion allocated to that group in the core sample was computed as $q_i = \sqrt{p_i} / [-\sum_{i=1}^{s}(\sqrt{p_i})]$, *s*, being the number of groups.

3. Log frequency (LF method): In this method if F_i is the number of accessions in the i^{th} group in the whole collection then, the proportion q_i allocated to that group in the core sample was computed as: $q_i = \text{Log } F_i / [\sum_{i=1}^{s}(\text{Log } F_i)]$, *s* being the number of groups).
4. Equal frequency (EF method): In this method equal number of accessions were selected from each group for a given size of the core sample.
5. Diversity Proportion (DP method): In this case the number of accessions selected from each group depended on the proportion of the pooled diversity of that group i.e. the proportion of $(N_i * SDI_{st(I),})$ where N_i is the number of accessions in the i^{th} group and $SDI_{st(I)}$ is the pooled SDI of that group. The diversity index for each group is also provided in Table 3.

The pooled SDI being a complex parameter, to estimate its expected value and sampling variance, for the method of simple random sampling, 100 independent random samples of a given size were drawn without replacement from the given data set. In case of stratified random sampling the same procedure was followed and it was ensured that the number of accessions from each group were fixed as per the method of allocation. The sample pooled SDI was computed in each case and also the mean and variance of the SDI was computed over the 100 samples. Also the mean SDI for the individual descriptors for the core sample as well as for the unselected accessions (non-core group) was computed. Mahajan *et al.*, (1999) suggested an estimator with appropriate weights for the individual groups for estimating the pooled SDI when the stratified random sampling was used. The weights may have to be fixed such that this index approached the value of the population SDI. But in the present investigation our main interest was only in the diversity of the core sample *per se.* Hence, the resultant core sample was treated as one homogeneous mixture of accessions from different diversity groups and the properties of the core sample were studied in terms of the mean and variance of the pooled SDI. Also, the closeness of the joint density distribution of the descriptors in the core sample to that of the whole collection was evaluated by means of an information measure. This is explained below.

Table 3. Diversity groups based on geographical origin and plant types

Main cluster	Group ID	Description entries	No. of SDI	Group
1	01	From India - Bushy	253	14.76
	02	.....- Cone shaped	455	15.10
	03	.....- Erect/appressed	94	14.86
2	04	From India-Bushy	648	15.72
	05	.....- Cone shaped	378	19.94
	06	.....- Erect/appressed	272	19.21
	07	From USA – Bushy	69	18.16
	08	.....- Cone shaped	35	20.03
	09	..... – Erect/appressed	21	21.44
	10	Other regions-Bushy	36	21.07
	11	.....- Cone shaped	55	21.30
	12	.....- Erect/appressed	46	20.35
3	13	From India – Bushy/cone shaped	50	23.39
	14	.....- Erect/appressed	38	23.88
	15	From USA – Bushy/cone shaped	24	25.46
	16	.....- Erect/appressed	17	24.55
	17	Other regions – Bushy/cone shaped	22	23.62
	18	.....- Erect/appressed	54	26.34
4	19	Mostly Erect type...India	33	22.70
	20	.....USA	66	22.56
	21	.....Other regions	27	25.15
5	22	From India – Bushy	37	20.47
	23	.....- Cone shaped	99	21.63
	24	.....- Erect/appressed	89	23.92
	25	From USA – Bushy	12	19.28
	26	.....- Cone shaped	26	23.18
	27	.....- Erect/appressed	46	25.06
	28	Other regions – Bushy/cone shaped	33	24.50
	29	.....- Erect/appressed	157	25.80
6	30	Mostly erect/appressed type	58	25.40
		Pooled SDI for whole collection		26.14

Following Boulton and Wallace (1970), for each multi-state or qualitative descriptor '*d*' the probability of occurrence of descriptor state '*m*' of the attribute in the whole collection is estimated by

$$p[m,d] = \{n[m,d] + 1\}/\{n[d] + M[d]\} \quad (1)$$

where n[m,d] denotes the number of accessions having attribute state *m* of the descriptor *d*; n[d], the number of accessions having any known value of the attribute *d*; and M[d], the number of descriptor states of descriptor *d*. In general, if data on all attributes *d* are available

on all the accessions, n[d] would be equal to the number of accessions in the whole collection. Based on information theory concepts, the length of the information code that can optimally indicate the possession of descriptor state *m* of attribute *d* is equal to

$$c[m,d] = -\log_e p[m,d] = -\log_e \{n[m,d] + 1\}/\{n[d] + M[d]\} \quad (2)$$

The estimate (1) is slightly biased to prevent the divergence of (2) when n[m,d] = 0, and has the useful effect of allowing an accession to be taken into consideration without much error even though it has a descriptor state *m* not possessed by any existing member of that group. For each accession *s* in the core sample, a message length F[s] that is required to optimally encode all the *d* attributes of *s* using the joint density distribution of the whole collection, is then computed using the formula:

$$F[s] = \sum_{des} c[x[d,s],d], \quad (3)$$

where $\sum_{des}$ means summing over all discrete or qualitative attributes; and *x*[d,s] indicates any descriptor state *m* of attribute *d* possessed by the accession. The message length c[x,d] is obtained from (2) and hence the total message length that is attributed to each of the accessions in a core sample of a given size can be computed. The average message length can therefore be computed over the 100 repeated random samples that are drawn by the sampling procedure. The mean message length would depend on the size of the core sample. Any sampling scheme that gives the least mean message length for a given sample size would have the joint density distribution of the descriptors closest to that of the whole collection.

The deviation of the frequency patterns in the core sample from those of the whole collection was also tested. Since, the core samples formed a part of the whole collection, a direct comparison was not possible. Hence, the frequency patterns of the individual descriptors in the core sample as well as for the non-core group were compared by means of a chi-square test (Rao, 1973). Such a test implies the statistical significance of the differences between the diversity of the core sample and the whole collection. Such deviations in the diversity measure of the individual descriptors can be pooled over and can be evaluated by computing the statistic:

$KDS = \sum_{des} \{\sum_{j=1}^{s} (p_{1j} - p_{2j}) \log_e (p_{1j} / p_{2j})\}$, called Kullback Divergence Statistic (Goldstein and Dillon, 1978), where p_{1j} and p_{2j} represent the relative frequency of the j^{th} descriptor state of any descriptor in the core sample and the non-core group respectively and *s* being the number of descriptor states and $\sum_{des}$ indicating summation over all the descriptors. In order to obtain valid estimates of KDS in cases where the core sample or the non-core group may have zero frequency for a descriptor state, a slightly biased estimate of p_{ij} were used as shown in equation (1) which did not result in any appreciable error. The mean KDS values were computed for the core samples obtained by various schemes and their values compared. The magnitude of the mean KDS would indicate the divergence of the core sample from the non-core group and hence from the whole collection.

Results and Discussion

The number of accessions from different geographical regions that are grouped into six major clusters is presented in Table 2. Cluster 2 was the largest cluster with nearly 50% of the accessions, followed by Cluster 1. Cluster 6 was the smallest with 58 accessions. Out of the 802 accessions in Cluster 1, 795 were from India, which has the largest number of accessions in the whole collection. In Cluster 2, apart from a large number of accessions from India, there were a large proportion of accessions from USA also. Similarly, in Clusters 3, 4 and 5, the accessions from these two regions had a larger proportion of accessions. However, Cluster 6 did not show any predominant region and this cluster was taken as a single diversity group. In Table 3, the grouping of accessions into 30 diversity groups is shown based on their geographical origin and the plant growth habit.

The pooled diversity SDI for the whole collection was equal to 26.14 and the group-wise diversity index as shown in Table 3 indicated wide variation among the diversity groups. Even though groups-1,2 and 3 had fairly large number of accessions, the diversity in these groups was quite small as compared to that of the whole collection. This may be due to the fact these groups have mostly accessions from India, where a large number of duplicates were suspected. Diversity groups obtained from Cluster 2 had only moderate diversity and even in this case a majority of the accessions were from India. Cluster 3, which had smaller number of accessions from India, USA and other countries had a diversity index approaching that of the whole collection. Cluster 4 indicated moderate diversity from the groups containing accessions from India and USA, but showed a diversity approaching the population diversity for the group

consisting of accessions from other countries. Again in Cluster 5, groups consisting of accessions from USA and other countries had a diversity that was close to that of the whole collection. Cluster 6, consisting mostly of erect plant types had a diversity index close to that of the whole collection. In general, the groups consisting of accessions from India, the main source of origin of safflower, had relatively smaller diversity compared to the groups that consist of accessions from other countries. Balakrishnan and Suresh (2000) indicated that in a core collection of about 570 safflower accessions from this data set, nearly 45% of the accessions from India were probably duplicates in view of a very high similarity index among them. This could be due to the fact that the Indian accessions had been pooled from different coordinating centres within a narrow geographical area, mostly from the states of Maharashtra, Karnataka and Andhra Pradesh which are the predominant regions of safflower cultivation.

Using a pooled estimator for the SDI by combining the stratum diversity indices with the weights proportional to the group size, resulted in the estimate of the pooled diversity equal to 18.94, much smaller than the actual population diversity index of 26.14. Since the main interest of the present investigation was to assess the diversity of the core samples obtained through various stratification schemes, each of the core samples was treated as a homogeneous mixture of accessions from different diversity groups and the pooled diversity in the core samples were directly computed. The mean and variance of the pooled diversity index estimated using 100 random samples (of appropriate sizes) are presented in Table 4. Core samples constituted by simple random sampling, FP method and DP method had almost the same level of pooled diversity as that of the whole collection. In these methods, the variance of the pooled SDI reduced with increased sample size and about 10-15% of the sample could be considered optimum for these methods beyond which reduction in variance was not appreciable. In the case of SQP, LF and EF methods, the diversity indices of the core samples progressively increased with increasing sample size. Core samples obtained through the EF method had maximum diversity even at 15% of the sample size. All these three methods yielded higher pooled diversity for the core samples than the whole collection. This result would not have been obvious if we had differentiated the individual diversity groups in the resultant core sample and used an estimator with weights. In contrast to this Mahajan *et al.*, (1999) indicated in their results that the weighted-index was about 75-80% of the population SDI. Table 4 also indicated that by using SQP, LF or EF methods, the core sample size could be kept at 10% of the whole collection, beyond which there was not much appreciable reduction in the variance of the estimated pooled SDI.

The information measure presented as the average message length in Table 4 also indicated that simple random sampling, FP method and PD method (in that

Table 4. Diversity measures for the core samples drawn from the whole collection by simple random sampling and stratified random sampling

Sample size	Mean SDI*	Variance (SDI)	Mean Message Length	Mean KDS**
1. Simple random sampling				
150	25.72	0.3876	3917	0.855
330	26.00	0.2163	8637	0.402
480	26.00	0.1230	12535	0.290
660	26.04	0.0900	17239	0.221
810	26.10	0.0671	21195	0.193
2. Stratified - Frequency proportion method				
150	25.97	0.1336	3945	0.770
330	26.07	0.0509	8651	0.347
480	26.13	0.0376	12587	0.251
660	26.00	0.0289	17201	0.190
810	26.00	0.0153	21098	0.163
3. Stratified - Square root proportion method				
150	28.28	0.1068	4362	1.327
330	28.50	0.0581	9606	1.560
480	28.58	0.0228	13992	1.624
660	28.62	0.0211	19253	1.798
810	28.65	0.0142	23637	1.971
4. Stratified - Log frequency method				
150	29.22	0.1147	4589	3.179
330	29.57	0.0335	10175	3.186
480	29.60	0.0209	14783	3.361
660	29.60	0.0130	20296	3.755
810	29.57	0.0101	24860	4.168
5. Stratified - Diversity proportional method				
150	26.98	0.1218	4115	1.115
330	27.08	0.0420	9017	0.646
480	27.08	0.0393	13096	0.572
660	27.11	0.0237	18001	0.535
810	27.13	0.0192	22101	0.560
6. Stratified - Equal frequency method				
150	29.85	0.0885	4770	4.402
330	30.03	0.0260	10494	4.417
480	30.01	0.0223	15204	4.664

* - Pooled Shannon Diversity Index based on 28 descriptors
** - Kullback Divergence Statistic measuring the divergence of the core sample from the non-core group

order) had the least average message lengths for any given core sample size. This evidently indicated that the core samples obtained by these methods had their joint density distribution for the descriptors as close to that of the whole collection. Because of this the pooled diversity of the core samples obtained by these methods were nearly equal to that of the whole sample. The core samples obtained through EF, LF and SQP methods (in that order) had larger message lengths for any given core sample size. This indicated that the joint density distribution of the core samples obtained by these methods were farther away from that of the whole collection, resulting in a higher diversity index for the core samples.

The mean divergence of the core samples from the reserve collection (non-core group) similarly showed interesting trends. In case of simple random sampling, FP and PD methods, the KDS values were low indicating that the frequency patterns in the core samples were similar to those of the reserve collection and the divergence *decreased* with increasing sample size. However, in the case of SQP, LF and EF methods, not only the core samples were more divergent from the non-core group, but also, this divergence increased with increasing sample size. The EF method had the largest amount of divergence for any given sample size.

A more detailed analysis of the deviation of the core sample from the whole collection in terms of SDI for the individual descriptors is presented in Table 5. The results have been presented for the core sample size of 10% of the whole collection and the SDI values are presented in a standardized form so that its value has a range of 0-1. The results clearly indicated that the diversity levels of the core samples obtained by simple random sampling, stratified random sampling using frequency proportion and diversity proportion methods were at par with those of the whole collection for all the descriptors. Core samples obtained using the square root proportion, log frequency and equal frequency methods of stratified random sampling had significantly higher levels of diversity than the whole collection for many descriptors. Interestingly, core samples obtained through equal frequency method had maximum diversity for several descriptors.

The general conclusion from the present investigation was that sampling of the accessions through stratification based on the diversity groups was more efficient than simple random sampling. This was due to the fact that sampling variances of the pooled SDI estimated through the five stratification schemes were much smaller than that of the simple random sampling method. The core samples obtained by frequency proportion method of stratified random sampling had nearly the same level of diversity as that of simple random sampling but the sampling variance of the estimate of pooled SDI was much less than that of simple random sampling. The diversity proportion method which was meant to give due importance to diversity levels of the individual groups as well as their size resulted in core samples with marginal increase in the estimated pooled SDI and lesser sampling variance. The other three methods of stratified random sampling not only yielded core samples with higher levels of diversity but also had reduced sampling variances for the estimate of pooled SDI. The square root proportion and log frequency methods of allocation were mainly aimed at reducing undue weights given to larger groups that might contain higher levels of genetic redundancy. As the larger groups in the present study had relatively lesser diversity than the smaller groups, these two methods resulted in core samples with higher diversity than the whole collection. Of all the six methods considered in the present study, the equal frequency method of allocation resulted in core samples with maximum diversity. This was due to the fact that the smaller groups had larger levels of diversity. Other strategies like the diversity dependent strategy without regard to group size (Yonezawa *et al.*, 1995) and the GL strategy based on the proportion of the product of logarithm of group size and group diversity (Mahajan *et al.*, 1999) yielded similar results to those of the DP strategy and the LF strategy respectively. Hence these results are not reported here.

The present study indicated that by appropriate stratification of the accessions and group allocation in the core sampling methodology, it was possible to obtain a core sample that might contain higher level of diversity than the whole collection. This leads us to examine whether it would be possible to predict by how much the diversity level of the core sample is likely to exceed that of the whole collection. Obviously, it could be difficult to predict the expected improvement in diversity of the core sample based on any one of the above sampling strategies. The results could vary depending upon the diversity levels of individual descriptors in different diversity groups; the group sizes and the sampling

Table 5. Mean diversity indices for descriptors in the core samples obtained using different sampling strategies at 10% core sample size

Descriptor name	Whole collection	Simple random	FP method	DP method	SQP method	LF method	EF method
Shape of lower stem leaf	0.544	0.541	0.548	0.550	0.568	0.592	0.609
Margin of lower stem leaf	0.322	0.325	0.319	0.337	0.378	0.391	* 0.421
Primary head shape	0.376	0.376	0.368	0.435	**0.522	**0.593	**0.627
Texture of upper leaves	0.624	0.627	0.629	0.653	*0.694	**0.694	**0.732
Shape of upper stem leaf	0.395	0.393	0.402	0.453	* 0.480	**0.515	**0.535
Margin of upper stem leaf	0.535	0.532	0.541	0.577	**0.627	**0.668	**0.686
No. of spines on upper stem leaf	0.520	0.518	0.521	0.569	**0.630	**0.682	**0.715
Attitude of OIB to head	0.731	0.731	0.734	0.804	**0.890	**0.951	**0.971
OIB cross section shape	0.600	0.600	0.600	0.668	**0.775	**0.840	**0.884
Location of spines on OIB	0.248	0.242	0.241	0.297	**0.379	**0.443	**0.475
No. of spines on OIB	0.534	0.533	0.531	0.594	**0.698	**0.767	**0.810
Length of spines on OIB	0.607	0.602	0.612	0.663	**0.744	**0.800	**0.832
Bracts enclosing head	0.319	0.322	0.320	0.372	**0.480	**0.545	**0.597
Growth habit	0.893	0.892	0.896	0.914	**0.932	**0.944	**0.949
Branch location on main stem	0.809	0.806	0.805	0.835	*0.848	**0.858	**0.869
Pollen production	0.822	0.818	0.828	0.848	0.870	*0.896	**0.914
Pappus on the acheme	0.246	0.234	0.248	0.248	0.268	0.276	0.290
Hull thickness	0.638	0.636	0.646	0.651	0.695	0.716	*0.740
Days to 50% elongation	0.682	0.683	0.681	0.694	*0.712	**0.715	**0.706
Days to primary branch initiation	0.704	0.705	0.703	0.717	*0.729	**0.733	**0.727
Days to 1st flowering	0.809	0.811	0.811	0.825	**0.843	**0.849	**0.844
Days to 50% flowering	0.807	0.810	0.812	0.828	**0.850	**0.858	**0.854
Days to physiol. maturity	0.650	0.653	0.653	0.661	**0.676	**0.664	**0.648
Plant spread	0.704	0.703	0.701	0.718	0.721	0.734	0.740
No. of primary branches	0.723	0.724	0.721	0.723	0.732	0.733	0.743
No. of capitula/plant	0.684	0.684	0.684	0.697	0.714	0.726	0.746
Internode length	0.558	0.563	0.565	0.571	0.578	0.591	0.596
Main capitula diameter	0.557	0.554	0.559	0.574	0.598	*0.623	**0.634
No. of accessions	3250	330	330	330	330	330	330

*, ** - Mean proportions of accessions in the core sample with different attributes of the descriptor significantly different from those of the whole collection at 5% and 1% probability levels, respectively

method. The basic objective was to reproduce almost the same level of genetic diversity in the whole collection through a smaller collection called the core sample. But practical considerations suggest that a user of the germplasm collection might require a core sample with desired or pre-determined levels of diversity for one or more descriptors simultaneously, preferably with much higher levels of diversity than the whole collection itself. As an example, an attribute (say, spines on leaves) may be present in 10% of the accessions and absent in 90% of the accessions in the whole collection. This indicates a standardized SDI of 0.47 (range 0-1) for this descriptor. Can the user of the genetic resource obtain a core sample with say, the attribute being present in 70% of the accessions and absent in 30% of the accessions in the core sample (with a standardized SDI of 0.88)? This implies not only much higher level of the diversity in the core sample, but also an entirely different distribution pattern of the descriptor states in the core sample. If the diversity level of only one descriptor is pre-determined, then it is quite simple to draw such a sample. However, when the diversity levels of several descriptors that are both qualitative and quantitative are simultaneously pre-assigned for the core sample, none of the sampling strategies considered above serve this objective as these methods involve only random selection. Hence it requires an entirely different approach to be adopted for obtaining such a core sample. In Part II of this investigation a new method is proposed for obtaining a core sample with pre-determined distribution patterns for one or more descriptors simultaneously.

Acknowledgements

The authors are highly grateful to the Project Director, Directorate of Oilseeds Research, Hyderabad, for permission to use the data on safflower accessions published by the Institute.

Appendix I

Converting a qualitative attribute to numerical score (Refer Balakrishnan and Sanghvi, 1968 for more details)

Let p_{ijk} be the estimated proportions of the k^{th} descriptor state of the j^{th} character S_j in the i^{th} group or stratum ($i = 1,2.....q$; $k = 1,2.....s_j + 1$). The dispersion matrix of the estimates p_{ijk} is given by $\mathbf{A}_{ij}$, where

$A_{ijk} = p_{ijk} (1- p_{ijk})/n_{ij}$, $k = l$

$= -p_{ijk} * p_{ijl} /n_{ij}$, $k \neq l$,

$k, l = 1,2,.....s_j +1$ and

n_{ij} is the sample size for the j^{th} character from i^{th} group. Since $\sum p_{ijk} = 1$ ($k = 1$ to $s_j +1$); the rows and columns of $\mathbf{A}_{ij}$ add up to 0.

The proportions may be taken as the means of variables X_{jk} which take only the values 0 and 1, subject to the condition that one and only one of the X_{jk}'s can be unity on any individual.

The common dispersion matrix of the variable X_{jk} over the q groups is estimated by $\mathbf{C}_j$, where

$$C_{jkl} = \sum_{i=1}^{q} n_{jk}^2 A_{ijkl} / \sum_{i=1}^{q} n_{ij}$$

Having computed the common dispersion matrix, the procedure of obtaining the transformed variables in terms of the original variables is given by Rao (1952, chapter 9), The only requirement is that we have to ignore or delete any row and the corresponding column of $\mathbf{C}_j$ before the process. The procedure is illustrated with the following example.The pooled dispersion matrix of the descriptor *'Number of spines on upper stem leaf'* with the descriptor states: No spines, Few, Intermediate and Many (pooled over the 13 geographical regions shown in Table 2) was:

0.0376 -0.0102 -0.0263 -0.0010

-0.0102 0.1197 -0.1056 -0.0039

-0.0263 -0.1056 0.1518 -0.0205

-0.0010 -0.0039 -0.0205 0.0254

Eliminating the first row and column corresponding to 'No spines' and using the procedure of Rao (1952), the transformed variables for the descriptor states are:

No: Y(1) = 0 (because we have not considered this row and column).

Few: Y(2) = 2.8904 X(2), where X(2) = 1 when this attribute is present

Int: Y(3) = 3.6431 X(2) + 4.1296 X(3), where only X(3) =1 for this attribute; X(2) = 0

Many: Y(4) = 3.1549 X(2) + 3.2794 X(3) + 8.0325 X(4), where only X(4) = 1 for this attribute and X(2) = X(3) = 0

As one and only one attribute can take a value of 1 when it is present, in the above equations the numerical scores for the 4 qualitative attributes are: 0, 2.89, 4.13 and 8.03 respectively. Thus this scoring system is more objective than the arbitrary scaling of 0,1,2 and 3 for the above attributes as it takes into consideration the dispersion matrix of the attributes' frequency in the population.

We may arbitrarily choose any row and the corresponding column to be ignored. But by convention, we may omit the row/column corresponding to the lowest ranked attribute in the case of an ordered-multi state descriptor (as in the present example). In the case of binary attributes the scores are given as 1 and 0 as usual.

References

Balakrishnan R and KK Suresh (2000) Some strategies for obtaining core samples from germplasm collections using strata of geographical origins - a case study in safflower (*Carthamus tinctorius* L.). *Statistics and Applications* **2**: 49-64.

Balakrishnan V and LD Sanghvi (1968) Distance between populations on the basis of attribute data. *Biometrics.* **24**: 859-865.

Bisht IS, RK Mahajan, TR Loknathan and RC Agrawal (1998) Diversity in Indian sesame collection and stratification of germplasm accessions in different diversity groups. *Genet. Resour. Crop Evol.* **45**: 325-335.

Boulton DM and CS Wallace (1970) A program for numerical classification. *Computer J.* **13**: 63-69.

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

Brown AHD (1989b) The case of core collections. In: AHD Brown, OH Frankel, DR Marshall & JT Williams (eds). *The Use of Plant Genetic Resources.* Cambridge University Press, Cambridge, pp 136-156.

Frankel OH (1984) Genetic perspectiveness of germplasm conservation. In: WK Arber, K Llimensee, WJ Peacock and P Starlinger (eds.), *Genetic manipulation: Impact on Man and Society*, Cambridge Univ. Press, Cambridge. pp. 161-170.

Goldstein M and WR Dillon (1978) Discrete Discriminant Analysis. John Wiley & Sons.

Ghorpade DS, RC Patil, R Balakrishnan, KV Venkateswara Rao, VD Shende and KP Deolankar (1991) Safflower Genetic Resources-Evaluation and Analysis. Germplasm Management Unit & Project Coordinating Unit (M.P.A.U.), Solapur and Directorate of Oilseeds Research, Hyderabad, India.

Harch BD, KE Basford, IH DeLacy, PK Lawrence and A Cruickshank (1996) Mixed data types and the use of

pattern analysis on the Australian groundnut germplasm data. Genet. Resour. Crop Evol. **43**: 363-376.

Hintum TJL van (1995) Hierarchical approaches to the analysis of genetic diversity in crop plants. In: T. Hodgkin, AHD. Brown, TJL. van Hintum and EAV Morales (eds.), *Core Collections of Plant Genetic Resources.* John Wiley & Sons. pp. 23-34.

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

Mahajan RK, IS Bisht and PL Gautam (1999) Sampling strategies for developing Indian sesame core collection. *Indian J. Plant Genet. Resour.* **12**: 1-9.

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

Peeters JP and JA Martinelli (1989) Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theor. Appl. Genet.* **78**: 42-48.

Rao CR (1952) *Advanced Statistical Methods in Biometric Research.* John Wiley & Sons, New York.

Rao CR (1973) *Linear Statistical Inference and its Applications.* John Wiley & Sons, New York.

Spagnoletti Zeuli PL and CO Qualset (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theor. App. Genet.* **87**: 295-304.

Yonezawa K, T Nomura and H Morishima (1995) Sampling strategies for use in stratified germplasm collections. In: T Hodgkin, AHD Brown, TJL van Hintum and EAV Morales (eds.), Core Collections of Plant Genetic Resources, John Wiley & Sons, pp. 35-54.

Wallace CS and DM Boulton (1968) An information measure for classification. *Computer J.* **11**: 185-194.