Strategies for Developing Core Collections of Safflower (*Carthamus tinctorius* L.) Germplasm – Part II. Using an Information Measure for Obtaining a Core Sample with Pre-determined Diversity Levels for Several Descriptors Simultaneously

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

A new technique for establishing a core sample from a large germplasm collection based on an information measure is proposed. The procedure is viewed as a classification scheme having two groups, *viz.*, the core sample group and the non-core group with pre-assigned properties in terms of group sizes and the density distributions of several descriptors. That is, for a known or fixed core sample size, the proportion p[m,d,t] of each state *m* for each multi-state qualitative descriptor *d* is pre-assigned within each group *t*. Similarly, for each continuous or quantitative descriptor *d*, the mean μ[d,t] and standard deviation σ[d,t] are pre-assigned. The probability of finding an accession with a given set of attributes or values in the vicinity of either of the two groups is estimated from the density of the pre-assigned composite model for each group. Then, based on an information measure a group is identified that has the highest density in the neighbourhood of any accession in the measurement space and each accession in the collection is assigned to that group *i.e.* the core sample or the non-core group. The information measure that is used is the length of the message that can optimally convey the description of the measurements, which equals the negative value of the logarithm of the probability of obtaining the given set of measurements in the neighbourhood of any one of the two groups. An accession that has the least length with regard to any group is assigned that group. The method is illustrated with a data set of safflower germplasm described by Balakrishnan and Suresh (2000) and Part I of this article. A scheme by which the accessions are stratified as per their geographical origin is also considered. The diversity indices of the core samples constituted by using various options in the proposed method are compared with those of the whole collection as well as those obtained through simple or stratified random sampling.

Key Words: ***Carthamus tinctorius*** **L., Core Collection, Information Measure, Shannon Diversity Index (SDI)**

One of the major issues about which a germplasm manager must be concerned is the need to improve the accessibility of the collections of a genetic resource to the users. In practice, plant breeders who are the main users of the germplasm collections are interested in having fairly small number of genotypes which possess or likely to possess the characters needed in their breeding programmes. For this purpose simple random sampling of the accessions in the main collection may not be adequate as the users wish to identify unique genetic resources. Rather, the sample should be biased to assure the inclusion of accessions from poorly represented geographic regions/diversity groups or should provide a reasonable amount of diversity in morphological traits and variation in metric traits. In part I of this article, we considered an approach that used passport, characterization and other data of a safflower germplasm data set containing 3,250 accessions (Ghorpade *et al.*, 1991) to first identify diversity groups of accessions and then select random samples from each group. It was found that core samples with higher levels of diversity than the whole collection could be obtained by some of the stratified random sampling methods. Nevertheless it was pointed out that the properties of the resultant core sample might vary depending upon the diversity levels of the individual descriptors in different groups, the group sizes and the method of sampling. Therefore, a different approach is required if the user of the genetic resource wants to have a core collection of desired size with pre-determined frequency patterns for two or more descriptors, preferably with higher levels of diversity than the whole collection. The present investigation suggests such a procedure by which the user can fix the properties of the core sample in terms of diversity measures of several qualitative and quantitative descriptors simultaneously.

Materials and Methods

The safflower germplasm data set as detailed in Part I of this article was used for developing the procedure for obtaining a core sample with pre-determined levels of diversity for 15 descriptors as listed in Table 1. In order to pre-assign the diversity of the core sample in

terms of individual qualitative and quantitative descriptors the following computations were made.

Table 1. List of descriptors used for the study

1. Margin of lower stem leaves: Serrate, deeply serrate and deeply lobed
2. Texture of upper leaves: Fleshy, normal, leathery
3. Spines on upper stem leaf: Non-spiny, few spines, intermediate, many spines
4. Location of spines on OIB*: None, tip only, tip & few basal, tip & few apical, tip & all along margin
5. Number of spines on OIB*: None, intermediate, many
6. Length of spines on OIB*: None, short, intermediate, long
7. Bracts enclosing head: Complete, incomplete
8. Growth habit: Bushy, cone shaped, appressed, erect
9. Pollen production: Sparse, intermediate, abundant
10. Hull thickness: Thin/ intermediate, thick
11. Days to physiological maturity: 70-80, 80-90, etc., ...140-150, > 150
12. Number of primary branches: <= 3, 3-6, etc., ...24-24, > 27
13. Number of capitula per plant: 0-10, 10-20, etc., ... 90-100, >100
14. Mean inter-node length (cm) : 1, 2, etc., ...,10
15. Main Capitula diameter (cm): 0.5-1.0, 1.0-1.5, etc., ...3.0-3.5, >3.5

* OIB: Outer Invoclucre Bracts

(a) ***Qualitative Descriptors***

The relative frequencies of accessions in the whole collection with respect to the individual descriptors were computed. In order to have a higher level of diversity in the core sample, two types of frequency transformations were used, *viz.*, the square root-proportion and the log-frequency transformations. For the square root-proportion transformation, if p_i was the relative frequency of a descriptor state in the whole collection, the relative frequency q_i in the core sample for that descriptor state was computed as $q_i = \sqrt{p_i} / [\sum_{i=1}^{s} \sqrt{p_i}]$, where s is the number of descriptor states for a specific descriptor. For the log-frequency transformation, if F_i was the absolute frequency of a descriptor state in the whole collection, the relative frequency q_i in the core sample for that descriptor state was computed as $q_i = \log(F_i)/[\sum_{i=1}^{s} \log(F_i)]$. These two frequency transformations have a useful property that descriptor states with high relative frequency in the whole collection have reduced relative frequency in the core sample and those with low relative frequency in the whole collection have somewhat higher relative frequencies in the core sample. By these methods, the diversity level for a descriptor in the core sample as measured by Shannon Diversity Index (SDI) was improved substantially if the SDI was low in the whole collection for this descriptor. For a given pre-assigned core sample size, the relative frequencies of various attributes in the core sample and the non-core group could therefore be easily computed. It was ensured that in case the absolute frequency for a descriptor state in the core sample exceeded the corresponding frequency in the whole collection, the frequency for that descriptor state in the core sample was made equal to that of the whole collection. The excess number of accessions was added to the largest frequency class in the core group. This implied minor modifications in the relative frequencies for the descriptor states for both the groups. In Table 2 and Table 3, the computations have been illustrated for a sample of qualitative descriptors by fixing the core sample size as 420 using the square root-proportion and log frequency transformations respectively. It should be noted that once the q_i's are fixed (that are independent of the core sample size), then the corresponding relative frequencies in the non-core group will depend on the size of the core sample and the relative frequencies in the whole collection. The core sample size of 420 was fixed for this illustration based on the results of Balakrishnan and Suresh (2000) for the safflower data set. Similar computations for a core sample of 10% of the whole collection (325) were made but not illustrated here.

The Shannon Diversity Index (SDI) for each descriptor for the core sample and the non-core groups was computed by using the formula:

$$SDI = -\sum_{i=1}^{s} p_i \log_e (p_i),$$

where p_i were estimated as $(n_i + 1)/(n + s)$, n_i being the frequency of a particular descriptor state; n- the total number of observations and s- the number of descriptor states for a particular descriptor. The estimates of p_i are slightly biased to enable a descriptor state with zero frequency to be included in the computation of the SDI though it results in a negligible error. The SDI value was divided by its maximum possible value, i.e. by $\log_e(s)$, so that its value ranges from 0 to 1.

(b) ***Quantitative Descriptors***

The results of principal component method arrived at by Balakrishnan and Suresh (2000) for the safflower data were used for fixing the frequency patterns of the quantitative descriptors. Based on this study, a core sample of 420 accessions was recommended for the safflower germplasm after eliminating probable duplicates from the first stage core sample of nearly 570 accessions. This core sample had higher diversity for the quantitative

Table 2. Examples of fixing the descriptor state frequencies of the core sample as per the square-root proportion transformation

Descriptor & descriptor states	Freq. In whole collection*	Relative frequency in the whole collection	Relative frequency fixed for the core sample*	Freq. in non-core group**	Relative frequency in non-core group
Texture of upper leaves					
Fleshy	62	0.0191	0.0917	23	0.0081
Normal	2260	0.6954	0.5536	2028	0.7166
Leathery	928	0.2855	0.3547	779	0.2753
Standardized SDI		0.6245	0.8366		0.5776
Location of spines on OIB					
None	133	0.0409	0.1346	77	0.0272
Tip only	62	0.0191	0.0919	23	0.0081
Tip & few basal	88	0.0271	0.1095	42	0.0148
Tip & few apical	6	0.0018	0.0143	0	0.0000
Tip & all along the margin	2961	0.9111	0.6496	2688	0.9499
Standardized SDI		0.2517	0.6748		0.1589
Geographical origin					
India & Bangladesh	2444	0.7520	0.3570	2294	0.8106
USA, Canada & Mexico	330	0.1015	0.1312	275	0.0971
Iran & Iraq	62	0.0191	0.0569	38	0.0135
Pakistan & Afghanistan	67	0.0206	0.0591	42	0.0149
Russia, Germany & Poland	32	0.0098	0.0408	15	0.0052
Switzerland, France, U.K., Hungary, Italy	32	0.0098	0.0408	15	0.0052
Spain & Portugal	30	0.0092	0.0396	13	0.0047
Algeria, Ethiopia, Kenya, Libya, Morocco & Sudan	44	0.0135	0.0479	24	0.0084
Egypt	28	0.0086	0.0382	12	0.0042
Turkey	69	0.0212	0.0600	44	0.0155
Israel, Syria, Jordan & Lebanon	33	0.0102	0.0415	16	0.0055
Australia	18	0.0055	0.0306	5	0.0018
Other countries	61	0.0188	0.0564	37	0.0132
Standardized SDI		0.4161	0.8627		0.3271

* Size of the main collection = 3250; Size of core sample = 420
** Size of non-core group = 2830

descriptors than the whole collection and the same frequency pattern as found in the resultant core sample was used for fixing the diversity of the quantitative descriptors in the present study. A sample of frequency patterns for days to physiological maturity and number of primary branches/plant are presented in Table 3. For further analysis in this investigation, the quantitative descriptors were categorized as qualitative descriptors with appropriate class-intervals being treated as the descriptor states.

Two approaches were followed for constituting the core samples. In the first approach a combination of 10 qualitative and 5 quantitative descriptors with pre-assigned frequency distributions were considered (the detailed list as provided in Table 1) with a core sample of 420 accessions. In the second approach only the 10 qualitative descriptors were considered with pre-assigned frequency densities for the core sample and the core sample size was fixed at 10%, *i.e.* 325 accessions. Some of the morphological descriptors were selected based on their association with reaction of the germplasm material to *alternaria* leaf spot disease (Venkateswara Rao *et al.*, 1993) and safflower aphids (Balakrishnan *et al.*, 1994). Three of the descriptors *viz.*, margin of lower stem leaves, location of spines on OIB and bracts enclosing head were selected in view of low diversity for them in the whole collection. Growth habit and hull thickness had fairly high level of diversity in the whole collection. The quantitative traits were included based on their agronomic importance. Under both these approaches, two options were tried in fixing the frequency patterns of the qualitative descriptors, *viz.*, (a) the square

Table 3. Examples of fixing the descriptor state frequencies of the core sample as per the log-frequency transformation

Descriptor & descriptor states	Freq. in whole collection*	Relative frequency in the whole collection	Relative frequency fixed for the core sample*	Freq. in non-core group**	Relative frequency in non-core group
Texture of upper leaves					
Fleshy	62	0.0191	0.1476	0	0.0000
Normal	2260	0.6954	0.4867	2056	0.7264
Leathery	928	0.2855	0.3657	774	0.2736
Standardized SDI		0.6245	0.9123		0.5370
Location of spines on OIB					
None	133	0.0409	0.2101	45	0.0158
Tip only	62	0.0191	0.1476	0	0.0000
Tip & few basal	88	0.0271	0.1923	7	0.0026
Tip & few apical	6	0.0018	0.0143	0	0.0000
Tip & all along the margin	2961	0.9111	0.4357	2778	0.9816
Standardized SDI		0.2489	0.8437		0.0676
Geographical origin					
India & Bangladesh	2444	0.7520	0.1547	2379	0.8406
USA, Canada & Mexico	330	0.1015	0.1072	285	0.1007
Iran & Iraq	62	0.0191	0.0763	30	0.0106
Pakistan & Afghanistan	67	0.0206	0.0777	34	0.0121
Russia, Germany & Poland	32	0.0098	0.0640	5	0.0018
Switzerland, France, U.K., Hungary, Italy	32	0.0098	0.0640	5	0.0018
Spain & Portugal	30	0.0092	0.0629	4	0.0013
Algeria, Ethiopia, Kenya, Libya, Morocco & Sudan	44	0.0135	0.0699	15	0.0052
Egypt	28	0.0086	0.0616	2	0.0008
Turkey	69	0.0212	0.0782	36	0.0128
Israel, Syria, Jordan & Lebanon	33	0.0102	0.0646	6	0.0021
Australia	18	0.0055	0.0429	0	0.0000
Other countries	61	0.0188	0.0760	29	0.0103
Standardized SDI		0.4161	0.9811		0.2646

* Size of the main collection = 3250; Size of core sample = 420; ** Size of non-core group = 2830

root-proportion method and (b) the log-frequency method.

In order to group the accessions as a basis for stratified sampling, strata of geographical origin of the accessions were taken into consideration. For this, the passport data of the accessions was used. The stratification method was included in the two approaches described above by simply considering an additional variable, *viz.*, geographical origin by treating it as yet another qualitative descriptor with descriptor states being the codes for the source country of the accessions. Again, in order to allow for a better representation of accessions from geographical regions with a very small proportion of accessions in the whole collection, the square root-proportion method and the log-frequency methods were followed to pre-determine the frequency patterns of the geographical regions in the core sample. The pre-assigned frequencies for accessions in the core sample from different geographical regions are also presented in Tables 2 and 3.

In effect, eight schemes were considered for constituting the core samples that are referred to as Scheme-1 to Scheme-8 in further discussions. They are:

Schemes	Method of fixing frequency distribution of the core samples and the core sample size
Scheme-1	Frequency patterns of 10 qualitative descriptors as per square root-proportion method + pre-assigned frequency distribution of 5 quantitative descriptors; no stratification: pre-assigned core sample size = 420
Scheme-2	Same as Scheme-1 but with stratification of accessions as per their geographical origin with their allocation as per square root-proportion method
Scheme-3	Frequency patterns of 10 qualitative descriptors as per the square root-proportion method; no stratification: pre-assigned core sample size = 325 (10% of whole collection)
Scheme-4	Same as Scheme-3 but with stratification of accessions as per their geographical origin
Scheme-5	Frequency patterns of 10 qualitative descriptors as per log-frequency method + pre-assigned frequency distribution of 5 quantitative descriptors; no stratification : pre-assigned core sample size = 420
Scheme-6	Same as Scheme-5 but with stratification of accessions as per their geographical origin with their allocation as per log-frequency method
Scheme-7	Frequency patterns of only 10 qualitative descriptors fixed as per the log-frequency method; no stratification: pre-assigned core sample size = 325
Scheme-8	Same as Scheme-7 but with stratification of accessions as per their geographical origin with their allocation as per log-frequency method

Table 4. Examples of frequency profiles of some quantitative descriptors in the core sample and the non-core sample group

Descriptor	Core sample		Non-core group	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
No. primary branches/plant				
≤ 3	6	0.0143	12	0.0042
3 - 6	82	0.1952	545	0.1926
6 - 9	104	0.2476	996	0.3519
9 - 12	82	0.1952	626	0.2212
12 - 15	43	0.1024	375	0.1325
15 - 18	48	0.1143	222	0.0784
18 - 21	16	0.0381	44	0.0155
21 - 24	26	0.0619	7	0.0025
24 - 27	4	0.0095	1	0.0004
> 27	9	0.0215	2	0.0008
Total	420	1.0000	2830	1.0000
Standardized SDI		0.8546		0.6969
Days to physiological maturity				
70- 80	1	0.0024	0	0.0000
80- 90	3	0.0071	1	0.0004
90-100	5	0.0119	78	0.0276
100-110	11	0.0262	477	0.1685
110-120	61	0.1452	1325	0.4682
120-130	149	0.3548	717	0.2533
130-140	140	0.3333	219	0.0774
140-150	48	0.1143	12	0.0042
150-160	2	0.0048	1	0.0004
Total	420	1.0000	2830	1.0000
Standardized SDI		0.6934		0.6093

Having fixed the frequency patterns of the descriptors for the core sample group and the non-core group, the next step was to compute appropriate information measure. This information measure served as the criterion for the classification of individual accessions with a given set of attribute values into the core sample or the non-core group. The details of the theoretical framework and development of this information statistic has been explained by Wallace & Boulton (1968). An algorithm has been proposed by Boulton and Wallace (1970) that can easily be adapted to our present objective of allocating an accession to the core sample group or the non-core group. The computational aspects of this algorithm are explained below.

1. For each multi-state or qualitative descriptor '*d*' the probability of occurrence of descriptor state '*m*' of the attribute in group '*t*' (t =1 meaning the core sample group and t = 2 meaning the non-core group) is estimated by

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

where n[m,d,t] denotes the number of accessions in group *t* having attribute state *m* of the descriptor *d*; n[d,t], the number of accessions in group *t* 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,t] would be equal to the number of accessions in group *t*. This is fixed beforehand by assigning the core sample size. The length of the information code that can optimally indicate the possession of descriptor state *m* of attribute *d* is equal to

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

The estimate (1) is slightly biased to prevent the divergence of (2) when n[m,d,t] = 0, and has the useful effect of allowing an accession to be assigned to a group without much error even though it has a descriptor state *m* not possessed by any existing member of that group.

2. For each continuous or quantitative descriptor *d*, the distribution is assumed to be normal within each group *t*. Its mean is estimated by

$$\mu[d,t] = (\Sigma_t\, x[d,k])/n[d,t] \quad (3)$$

where Σ_t indicates summation over the entries in group *t* and *x*[d,k] indicates the value of accession *k* on attribute *d*. The standard deviation is estimated by

$$\sigma[d,t] = \{(\Sigma_i(x[d,k])^2 - n[d,t]\ (\mu[d,t])^2)/ (n[d,t]\ -1)\}^{0.5} \quad (4)$$

A distribution normalizing constant g[d,t] is estimated by

$$g[d,t] = \log_e\ (\sigma[d,t]/(K^*\ \varepsilon[d])) \quad (5)$$

where $K = 1/\sqrt{2\Pi}$. It is assumed that a measurement $x[d,k]$ of the attribute d of an accession k is quoted to a least count of $\varepsilon[d]$, i.e. to an accuracy of $\pm\varepsilon[d]$ and that the probability of getting such a measurement from the distribution (μ,σ) is approximately

$$(K^*\ \varepsilon[d]/\sigma)^*\ \exp(-(x[d,k]\ -\ \mu)^2/2\sigma^2)$$

3. For each accession k, a message length F[k,t] that is required to optimally encode all the d attributes of k using the density distribution of group $t(t = 1 \text{ or } t = 2)$, is then computed using the formula:

$$F[k,t] = I[t] + \Sigma_{des}\ c[x[d,k],d,t] + \Sigma_{qty}\ (g[d,t] + (x[d,k]\ -\ m[d,t])^2\ /\ 2(s[d,t])^2) \quad (6)$$

where Σ_{des} means summing over all discrete or qualitative attributes; Σ_{qty} means summing over all continuous or quantitative attributes and $x[d,k]$ indicates any descriptor state m of attribute d possessed by the accession k in the case of a qualitative descriptor and it indicates the numerical value in the case of a quantitative descriptor. It is also assumed that the variables under consideration are independent of each other. The first parameter I[t] in (6) is an estimate of the message length for the group label to which an accession belongs. This depends on the abundance of group t in terms of the number of accessions in that group and is estimated by

$$I[t] = \log_e(S/n[t]),$$

where S is the total number of accessions in the whole collection (=3250 for the present case) and n[t] is the pre-assigned size of group t. For the present study, n[1] is the core sample size (= 420 or 325) and n[2] is the non-core group size (=2830 or 2925). In the present study the quantitative traits with appropriate class-intervals were treated like qualitative descriptors and hence equations (3) to (5) and the third component of (6) were not used.

For each accession k in the whole collection, depending upon the values possessed by it on various descriptors, the value of F[k,t] is computed for each of the two groups *viz.*, the core sample group and the non-core group. If $F(k,1) \leq F(k,2)$ then the accession is assigned to the core sample, else it is assigned to the non-core group. The computations are illustrated as follows. Let us assume that a core sample of 420 accessions is required with frequency densities for 2 of the descriptors as shown in Table 2a and no stratification is involved. In Table 4 the values of c[m,d,t] are shown for these two sample descriptors along with the descriptor state frequencies in the core sample and non-core group. By using the values in Table 4, it can be seen that any accession possessing descriptor state 'Fleshy leaves' for the descriptor 'texture of upper leaves' and 'Tip and few basal' for the descriptor 'location of spines on OIB' has the values F[k,1] = I(1) + 2.3838 + 2.2019 = 6.6319 and F[k,2] = I(2) + 4.7302 + 4.1886 = 9.0572. Here, $I(1) = \log_e(3250/420) = 2.0462$ and $I(2) = \log_e(3250/2830) = 0.1384$ are the message lengths that correspond to the group abundance for the two groups, *viz.*, the core sample and the non-core group. Hence the accession is allocated to group t = 1 i.e. the core sample, for F(k,t) has a minimum value for t =1.

Let us assume that the selection scheme involves stratification of the accessions as per their geographical origin and the expected region-wise allocation in the core sample is as shown in Table 2a. We now take the message lengths that correspond to the stratum or group to which an accession belongs and add them to the above values and the accession is assigned to the core sample or the other group depending upon the new computed values. For example, if an accession from USA region possesses the above two attribute values, then F(k,1) = 6.6319 + 2.0454 = 8.6773 and F(k,2) = 9.0572 + 2.3322 = 11.3894, where 2.0454 and 2.3322 are the message lengths for the two groups that correspond to USA region (Table 4). Hence it is assigned to the core sample group as F(k,1) is still less than F(k,2). However, when more number of descriptors belonging to both qualitative and quantitative types are taken into consideration, the assignment to either of the two group might change depending upon the values of c[m,d,t] for each group. The above procedure is repeated for all the accessions in the whole collection and each accession is allocated to either of the two groups.

It should be pointed out that even though we specify the initial core sample size, the final core sample size might vary to some extent depending upon the actual allocation of the accessions to either group based on the above procedure. If the resultant core sample size is lesser than the pre-assigned size, this core sample could be treated as the final product as it is still likely to be close to the pre-assigned frequency patterns or diversity levels. However, if the resultant core sample

Table 5. Computation of message length c[m,d,t] for encoding the attribute values under the core sample and the non-core group – an illustration

Descriptor & descriptor states	Frequency in core sample*	Frequency in non-core group*	c[m,d,t] for core sample	c[m,d,t] for non-core group
Texture of upper leaves				
Fleshy	39	23	2.3838	4.7302
Normal	232	2028	0.5921	0.3343
Leathery	149	779	1.0367	1.2893
Location of spines on OIB				
None	56	77	1.9916	3.6060
Tip only	39	23	2.3632	4.7717
Tip & few basal	46	42	2.2019	4.1886
Tip & few apical	6	0	4.2603	7.2566
Tip & all along margin	273	2688	0.4390	0.0529
Geographical origin				
India & Bangladesh	150	2294	1.0535	0.2141
USA, Canada & Mexico	55	275	2.0454	2.3322
Iran & Iraq	24	38	2.8519	4.2891
Pakistan & Afghanistan	25	42	2.8126	4.1914
Russia, Germany & Poland	17	15	3.1804	5.1800
Switzerland, France, U.K., Hungary & Italy	17	15	3.1804	5.1800
Spain & Portugal	17	13	3.1804	5.3136
Algeria, Ethiopia, Kenya, Libya, Morocco & Sudan	20	24	3.0262	4.7337
Egypt	16	12	3.2375	5.3877
Turkey	25	44	2.8126	4.1460
Israel, Syria, Jordan & Lebanon	17	16	3.1804	5.1194
Australia	13	5	3.4317	6.1609
Other countries	24	37	2.8519	4.3150

Size of the whole collection = 3250; Size of core sample = 420; Size of non-core group = 2830

* Frequency patterns as illustrated in Table 2.

size far exceeds the expected size (as is the case in the present study), the accessions allocated to the core sample could be ranked in ascending order based on the F(k,1) values and the required number of accessions as per their ranking could be taken out of this core sample and the rest allocated to the non-core group or the reserve collection. Since F(k,1) is the message length that corresponds to an accession being in the vicinity of the core sample with pre-assigned joint density of the descriptors, the required number of accessions that have shorter message length F(k,1) could be selected to the core collection and the remaining ones with larger values of F(k,1) could be relocated to the reserve collection. In this method, if fairly large number of accessions are left out of the first stage core sample, it is quite likely that the frequency patterns in the reduced core sample might yield somewhat lesser diversity than the pre-assigned one for one or more descriptors. An alternate but a cumbersome method is to split the first stage core sample into 2 groups, one with the pre-assigned size and the other with the balance number of accessions and repeat the whole procedure again with pre-assigned higher levels of diversity to obtain a second stage core sample. This procedure after a few iterations would result in a core sample that is quite close to the pre-assigned diversity index for several descriptors simultaneously. However, this procedure might be attempted if suitable computer program is available to perform the computations.

The diversity of various descriptors as measured by Shannon Diversity Index (SDI) was computed for the core samples obtained by various schemes. The SDI values were divided by $\log_e(n)$ to make it range from 0 to 1, where n stands for the number of descriptor states. The deviation of the frequency patterns in the core samples obtained from different schemes from those of the whole collection was also tested. Since, the core samples form a part of the whole collection, a direct comparison was not possible. Hence, the frequency

patterns of the individual descriptors in the core sample were compared against those in the non-core group 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.

In order to compare the results of the new method with the simple random sampling method, accessions were drawn at random from the whole collection without replacement to obtain 100 independent core samples of sizes 325 and 420, respectively. The mean SDI was computed for each descriptor from these 100 samples. Also the same procedure was repeated for stratified random sampling with strata of 13 geographical regions with the proportions of allocation as listed in Table 2 (square root-proportion method) and Table 3 (log-frequency method) for the geographical regions. The SDI values for the descriptors in the core samples obtained through simple random sampling or stratified random sampling were compared with those obtained through the new procedure using a t-test (Poole, 1974).

Results and Discussion

Core samples were obtained based on the 8 schemes (as listed under the earlier section). Scheme 1 resulted in a first stage core sample of 510 accessions against the pre-assigned size of 420 accessions. Similarly scheme 2, 5 and 6 resulted in first stage core samples of size 542, 514 and 556 respectively against the pre-assigned size of 420. Hence, 420 accessions were retained in the core sample for each of these schemes based on the ascending order values of F(k,1). Schemes 3,4,7 and 8 resulted in first stage core samples of size 359, 408, 407, 408 respectively against the pre-assigned size of 325 accessions. Out of these first stage core samples only the required number of 325 accessions were retained as per the ordering of F(k,1) values.

The standardized SDI values for various descriptors under consideration are presented in Table 6 for the core samples obtained through schemes 1, 2, 5 and 6 (sample size = 420). For comparison, the SDI values for the whole collection and the mean diversity for core samples obtained by simple random sampling (SRS) and stratified random sampling with strata of geographical origins are also included in this table.

A perusal of Table 6 indicates that without stratification of the accessions, the simple random sampling (SRS) method yielded a core sample with mean SDI values of the descriptors that are almost equal to those of the whole collection. By using Scheme-1 with the frequency patterns of the 10 qualitative traits fixed as per the square root-proportion method, the SDI values of these descriptors except growth habit were significantly

Table 6. Diversity measure[$] for descriptors in the core samples obtained through the schemes 1, 2, 5 and 6

Descriptor	Whole collection	No Stratification			With Stratification			Pre-assigned SDI[$]		
		SRS#	Scheme-1	Scheme-5	SRS–	Scheme-2	SRS-LF#	Scheme-6	Scheme –1&2	Scheme –5&6
Margin of lower stem leaf	0.322	0.321	0.595**	0.740**	0.384	0.553**	0.410**	0.712	0.751	0.973
Texture of upper leaves	0.625	0.626	0.879**	0.882**	0.701	0.853**	0.727**	0.852	0.835	0.912
Spines on upper stem leaves	0.521	0.518	0.835**	0.862**	0.655	0.809**	0.703**	0.852	0.845	0.983
Location of spines on OIB	0.249	0.245	0.642**	0.859**	0.359	0.559**	0.403**	0.779	0.686	0.844
No. of spines on OIB	0.535	0.534	0.963**	0.965**	0.718	0.942**	0.791**	0.974	0.870	0.988
Length of spines on OIB	0.607	0.605	0.945**	0.944**	0.781	0.925**	0.834**	0.941	0.890	0.990
Bracts enclosing head	0.319	0.323	0.771**	0.884**	0.414	0.677**	0.450**	0.840	0.722	0.968
Growth habit	0.893	0.893	0.897	0.899	0.927	0.881	0.908	0.892	0.962	0.995
Pollen production	0.822	0.821	0.983**	0.965**	0.930	0.984*	0.967	0.974	0.954	0.995
Hull thickness	0.638	0.636	0.850**	0.807**	0.737	0.863*	0.762	0.850	0.889	0.990
Days to physiological maturity	0.650	0.648	0.566**	0.597*	0.671	0.525**	0.556**	0.759	0.676	0.676
No. of primary branches	0.723	0.726	0.804**	0.705	0.717	0.780**	0.660*	0.730	0.846	0.846
No. of capitula/plant	0.684	0.685	0.857**	0.742*	0.723	0.841**	0.705	0.734	0.887	0.887
Inter-node length	0.559	0.563	0.717**	0.607	0.622	0.693*	0.608	0.719	0.756	0.756
Main capitula diameter	0.556	0.555	0.642**	0.581	0.606	0.643	0.580	0.701	0.742	0.742
Geographical origin	0.416	0.418	0.805**	0.773**	0.863	0.878	0.981	0.844	0.863	0.981
No. of accessions	3250	420	420	420	420	420	420	420	–	–

$: Shannon Diversity Index in standardized form

#: SRS– Simple Random Sampling; SRS-SQP: Stratified Random Sampling with Square root-proportion allocation for geographical regions;

SRS-LF: Stratified Random Sampling with Log-Frequency allocation for geographical regions

*, **: SDI significantly different from the corresponding SRS values at 5% and 1% levels, respectively.

higher than those obtained through SRS or the whole collection. For the descriptors margin of lower stem leaves, location of spines on OIB and bracts enclosing head, the diversity level was quite low in the whole collection but Scheme 1 yielded a core sample with significantly higher diversity than the whole collection. The diversity levels for the quantitative descriptors were also significantly higher than those of SRS, except days to physiological maturity for which the diversity in the core sample was significantly lesser under Scheme 1. Nevertheless the diversity in the core sample by Scheme-1 was somewhat short of the pre-assigned levels for some of the descriptors as it was indicated earlier that this core sample was reduced to the required size from the first stage core collection of 510 accessions. The results for Scheme 5 indicated similar trends as for Scheme-1 and in addition the diversity levels for the descriptors were much higher in case of qualitative attributes. For both these schemes, the geographical diversity in the core sample was significantly higher than the whole collection or as that of core sample through SRS. The schemes 2 and 6 wherein stratification of the accessions was considered, the SDI values for the qualitative descriptors (except 'growth habit') were significantly higher than those obtained through stratified random sampling where the accessions in the strata were assigned as per the square root-proportion method or log-frequency method. The comparison of the SDI values of quantitative traits for these two schemes against stratified random sampling indicated that days to physiological maturity had significant lower level of diversity for both these schemes. In general, core samples obtained through schemes 5 and 6 (with stratification of accessions as per geographical origin) had lower levels of diversity for the descriptors than those obtained through schemes 1 and 2. But the stratification obviously increased the diversity level of the accessions from different geographical regions.

In Table 7, the diversity index for the descriptors is presented for core samples obtained through schemes 3, 4, 7 and 8 (core sample size = 325) and for the core samples obtained through simple or stratified random sampling. In these schemes only the distribution patterns for 10 qualitative traits were pre-assigned. The diversity levels for the quantitative traits have also been presented in this table for the sake of comparison. The trends were similar to those of schemes 1, 2, 5 and 6, except that the quantitative descriptors had the same level of diversity as that obtained under simple random sampling or stratified random samples. Also the diversity levels in the core samples obtained through the schemes 3, 4, 7 and 8 were higher for location of spines on IOB and bracts enclosing head which had low diversity in the whole collection. For these schemes also the realized diversity levels were short of the target in view of the fact that the first stage core samples were reduced to the required size. For the schemes which included stratification of accessions as their geographical origins (2, 6, 4 and 8), the SDI values were somewhat less than the schemes without stratification. The results also indicated that for the descriptors that had high level of diversity in the whole collection, not much improvement could be obtained in the core samples. But for the descriptors with low to moderate levels of diversity in the whole collection, the proposed method enhanced their diversity significantly in the core sample.

Review of earlier work so far in this direction indicates a general rule for establishing a core sample that is as follows. Make use of passport, characterization and other relevant information to first identify groups of accessions; then select a sample from each group either randomly or by purposive sampling. The sampling fraction from the different groups might however vary depending on the data. The method suggested by Noirot *et al.*, (1996) markedly deviated from this approach. They demonstrated that by using principal component analysis (PCA) on a set of quantitative descriptors, the sampling strategy could be changed by selecting accessions with the largest deviations from the centroid of the whole collection or the individual strata to constitute the core sample. Though the PCA method was not specific about qualitative variables, it resulted in core samples that had almost the same or slightly higher amount of diversity for several qualitative attributes as that in the whole collection (Mahajan *et al.*, 1996, Balakrishnan *et al.*, 2000 and Balakrishnan and Suresh, 2000). However, it would be a great advantage from the users' point of view if a core sample could be obtained with pre-assigned levels of diversity for several characteristics that are both multi-state and continuous for a fixed sample size. For this purpose we need to specify the density distribution of the core sample, which automatically fixes the density distribution of the remaining accessions. Then the problem becomes one of assigning an accession with a given set of attribute values to either of the two groups based on some objective criterion. Wallace & Boulton (1968) using the information theory concepts have shown that this problem is equivalent to a classification scheme that minimizes an information measure with the following properties for the groups:

Table 7. Diversity measure[$] for descriptors in the core samples obtained through the schemes 3,4, 7 and 8

Descriptor collection	No Stratification Whole	With Stratification		Pre-assigned SDI[$]					
		SRS[#]	Scheme-3	Scheme-7	SRS-SQP[#]	Scheme-4	SRS-LF#	Scheme-8	Schm-3&4	Schm-7&8
Margin of lower stem leaf	0.322	0.329	0.567**	0.626**	0.390	0.599**	0.408	0.627**	0.751	0.973
Texture of upper leaves	0.625	0.627	0.803**	0.908**	0.697	0.896**	0.732	0.911**	0.835	0.912
Spines on upper stem leaves	0.521	0.529	0.788**	0.780**	0.653	0.786**	0.706	0.776	0.845	0.983
Location of spines on OIB	0.249	0.249	0.873**	0.946**	0.357	0.767**	0.404	0.893**	0.686	0.844
No. of spines on OIB	0.535	0.532	0.843**	0.908**	0.722	0.951**	0.798	0.950**	0.870	0.988
Length of spines on OIB	0.607	0.609	0.809**	0.861**	0.780	*0.869*	0.840	0.892	0.890	0.990
Bracts enclosing head	0.319	0.322	0.981**	0.879**	0.424	0.876**	0.455	0.876**	0.722	0.968
Growth habit	0.893	0.892	0.902	0.887	0.929	0.853	0.908	0.859	0.962	0.995
Pollen production	0.822	0.815	0.904*	0.939**	0.933	0.957	0.964	0.951	0.954	0.995
Hull thickness	0.638	0.638	0.722	0.779*	0.723	0.805	0.756	0.805	0.889	0.990
Days to physiological maturity	0.650	0.654	0.659	0.670	0.675	0.649	0.643	0.651	0.676	0.676
No. of primary branches	0.723	0.724	0.699	0.720	0.717	0.694	0.704	0.709	0.846	0.846
No. of capitula/plant	0.684	0.686	0.684	0.707	0.723	0.693	0.733	0.707	0.887	0.887
Inter-node length	0.559	0.559	0.679	0.601	0.623	0.621	0.657	0.620	0.756	0.756
Main capitula diameter	0.556	0.560	0.640*	0.614	0.601	0.630	0.625	0.630	0.742	0.742
Geographical origin	0.416	0.415	0.713**	0.868**	0.863	0.868	0.981	0.849	0.863	0.981
No. of accessions	3250	325	325	325	325	325	325	325	-	-

$: Shannon Diversity Index in standardized form

#: SRS – Simple Random Sampling;
SRS-SQP: Stratified Random Sampling with Square root-proportion allocation for geographical regions;
SRS-LF: Stratified Random Sampling with Log-Frequency allocation for geographical regions

*, **: SDI significantly different from the corresponding SRS values at 5% and 1% levels, respectively.

a) A probability distribution function in measurement space is provided for each group (*viz.*, the core and non-core sample groups in the current context).
b) The parameters describing the probability distribution for a group are assigned values that are essentially maximum likelihood estimates based on the individuals assigned to that group.
c) Each individual is assigned to that group which has the highest density in the neighbourhood of the thing in the measurement space.

The information measure F(k,t) specified in equation (6) is an ideal criterion as it is derivable using the existing data for the accessions in the whole collection. It is based on comparisons between an individual on the one hand and a group on the other, and uses the well defined concept of probability that a member of a group of known distribution would be found to have certain measurement attributes. The properties of the core samples that were obtained in the present investigation using various schemes of pre-assigning the density distribution of the core sample clearly demonstrated that it would be possible to substantially increase the diversity of attributes which are low or moderate in the whole collection. Probably this would not have been possible by the other methods available so far. The frequency patterns of the final core sample size were not close to the pre-assigned ones, and the deviations that we observed could be mainly attributable to the inclusion of several qualitative and quantitative descriptors each having its own independent density distribution. The increased diversities of several attributes from the smaller core samples obtained from Schemes-3, 4, 7 and 8 clearly indicated the potentials of the proposed method as a new and powerful technique for constituting core samples.

The proposed method could be easily adapted to a variety of options with regard to grouping or stratification of the accessions. For the present study we have taken into consideration the geographical origins of the accessions as the grouping criterion. The core samples obtained using the stratification scheme had better diversity than the whole collection for the attribute under consideration, *viz.*, the geographical origin.

For pre-assigning the descriptor state frequencies for the core sample we have considered the square root-frequency proportion method and the log-frequency method as they were found to be useful frequency transformation methods. In fact, we could use a general class of frequency transformation of the form: $q_i = p_i^{\alpha}/(\sum_{j=1}^{s} p_j^{\alpha})$, s being the number of descriptor states and α is some constant greater than zero (Theil, 1972). The square root-frequency proportion method used in this investigation corresponds to a value of $\alpha = 0.5$.

Any value of a less than 1 will increase the SDI value in the core sample as compared to that of the whole collection, but SDI would be much higher if a is less than 0.5. The log-frequency transformation suggested by Brown (1989) is one such method but it generally reduces the frequency levels of predominant attributes rather sharply. A value of a > 0.5 towards 1 will correspondingly reduce the diversity in the core sample than that when a < 0.5. The proportional method of frequency transformation corresponds to $\alpha = 1$. But this method of pre-assigning the frequency patterns for all the descriptors in the core sample to the same level as that in the whole collection is just equivalent to simple random sampling. Also, it is not useful for obtaining a core sample using the proposed information measure as it would result in F(k,1) > F(k,2) for all accessions and no accessions could be allocated to the core sample. But it is possible to adopt different proportions of q_i for different descriptors by keeping different values of α such that $\sum q_i = 1$ for each descriptor. For example, we can pre-assign the sampling fraction from different geographical regions using log-frequency transformation and pre-assign the density distribution for a qualitative descriptor by fixing $\alpha = 0.3$ or 0.4. In the present study we have fixed the frequency patterns for the quantitative descriptors based on a different criterion.

It should be noted that if an optimum core sample size could be decided for the set of pre-assigned density distributions for various descriptors, then the large deviations in the first stage core samples from the pre-assigned ones could be reduced to a great extent. The frequency patterns in the core sample and the core sample sizes were considered mainly to illustrate the advantages of the new procedure. Further research is needed for optimizing the core sample size for a given set of frequency patterns or *vice versa.* The proposed method has one major advantage that any new accession that is added to the genetic resource may be assigned to the core sample group or the non-core sample group by computing the appropriate information statistic F(k,t) that depends on the measurements of the new accession on several descriptors. It is also possible to have more than one working core collection from the same genetic resource by using different pre-assigned frequency profiles for different sets of descriptors.

Conclusions

In the present investigation, a new approach for obtaining a core sample from any large germplasm collection has been proposed. The problem has been treated as one of objectively assigning an accession to the core sample group or the reserve collection which have pre-assigned sample sizes and density distributions for various attributes. For this, an information measure that depends on the probability of obtaining an accession with a given set of measurements or attributes in the vicinity of the presumed composite density of either of the two groups has been used. The new method also incorporates grouping of the accessions into strata based on some criterion like geographical origins. The results of the present investigation have clearly demonstrated the potential use of this methodology for obtaining core samples with desirable combination of distribution profiles of various characteristics.

References

Balakrishnan R, K Venkateswara Rao, RV Dathkhele and RC Patil (1994) Evaluation of germplasm collections of safflower (*Carthamus tinctorius* L.) in India for morphological characters and its association with reaction to aphid infestation. *J. Oilseeds Res.* **11**: 229-236.

Balakrishnan R, NV Nair and TV Sreenivasan (2000) A method for establishing a core collection of *Sachharum officinarum* L. germplasm based on quantitative-morphological data. *Genet. Resour. Crop Evol.* **47**: 1-9.

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.

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

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

Ghorpade DS, RC Patil, R Balakrishnan, KV Venkateswara Rao, V.D. 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.

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.

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.

Poole RW (1974) *An Introduction to Quantitative Ecology.* McGraw Hill, Tokyo.

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

Theil H (1972) *Statistical Decomposition Analysis.* North-Holland Publishing Co., London.

Venkateswara Rao K, R Balakrishnan, CD Deokar and RC Patil (1993) Evaluation of germplasm collections of safflower in India for morphological characters and its association with reaction to *Alternaria* leaf spot. *J.Oilseeds Res.* **10**: 282-287.

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