Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 67(1) 2013 71-78

Establishment of Castor Core Collection Utilizing Self–Organizing Mapping (SOM) Networks

C. Sarada and K. Anjani

Directorate of Oilseeds Research, Rajendranagar, Hyderabad, Andhra Pradesh

Received 30 November 2011; Revised 2 February 2012; Accepted 26 February 2012

SUMMARY

A core collection can be defined as a representative sample of entire germplasm collection with minimum repetitiveness and maximum genetic diversity of a crop species and its relatives. The success of development of a most representative core collection mainly depends on non-overlapping grouping of whole collection. In the present study, a promising method viz., Self Organizing Mapping (SOM) network clustering technique was applied, which was first time attempted in establishment of core collection in a crop species. An attempt was made to compare SOM with clustering methods *viz*., Ward's and K-means clustering to understand the superiority of SOM over these two methods in forming castor core representative of whole collection. Forty experimental cores were constructed using these clustering methods as well two clustering algorithms (single and two stage) and two allocation methods, viz., proportional and logarithmic methods. Three sample sizes representing 10 per cent, 15 per cent and 20 per cent of total collection were drawn, and a fourth sample size of 524 accession based on progresss was made. Thus formed experimental cores were evaluated based on the four parameters viz., mean difference percentage (MD), variance difference percentage (VD), coincidence rate percentage (CR) and variable rate percentage (VR). The results indicated that SOM method performed better as compared to Ward's and K-means clustering methods conserving maximum diversity existing in the whole germplasm collection.

Keywords: Castor, Core collection, K-means clustering, Self Organizing Mapping networks, Ward's clustering method.

1. INTRODUCTION

Developement of a core collection which identifies a small number of germplasm accessions to represent the gentic diversity present in a large collection of germplasm is a very important research area in crop improvement. The concept of core collection was introduced by Frankel and Brown (1984) and Brown (1989) to minimize the cost of germplasm conservation whilst ensuring maximum genetic diversity. A core collection can be defined as a representative sample of entire germplasm collection with minimum repetitiveness and maximum genetic diversity of a crop species and its relatives.

Core collections were developed in crops to improve conservation of genetic resources and to stimulate effective utilization of genetic resources in crop improvement. Germplasm cores were developed on the basis of passport data, morphological, agronomic or quantitative traits in crops like chick pea (Upadhyaya 2001) Groundnut (Upadhyaya *et al*. 2003), Island cotton (Xu *et al*. 2006). Besides morphological and phenotypic data, biochemical and molecular markers data were employed for assessing genetic diversity (Vollman 2005, Escribano *et al.* 2008, Zhao 2010). The procedure to form a core was mostly based on a combination of practical experience, passport data and botanical variety.The criteria used for assessing the degree of diversity in a core were most often based on mean, range, variance, Shannon index etc. Evaluation of a primary core and subsets formed based on quantitative traits provided new information, which had direct

Corresponding author : C. Sarada
E-mail address : saradac@yahoo.com

implications on genetic conservation of a crop species and in deciding the strategies for future collecting as well as for restructuring of core subsets (Jana and Addala 1999).

For formation of a core in castor (*Ricinus communis* L.) crop, where no prior attempts were made to develop a core, more accurate studies in relation to sampling strategies, hierarchical relationships, genetic structure and gene pools are needed.

The major objective of the present study was to identify ideal methodology for construction of balanced castor core based on quantitative characters representing maximum genetic diversity existing in the total germplasm collection.

The success of development of a most representative core collection mainly depends on the reliable non-overlapping grouping of whole collection. Thus the efforts should be focused on creation of distinct groups in such a way that the accessions should be homogenous within the groups and heterogeneous between groups.

Among multivariate data analysis techniques, clustering is a set of methods capable of creating groups (clusters or classes) of objects based on their degree of association. Clustering techniques are purely empirical classification methods relying upon inductive approaches, thus influenced by the nature of the input data to be analyzed. Errors during cluster analysis contribute to ineffective decisions. (Chen *et al.* 1995)

Choosing an appropriate clustering method for a given data set is a difficult task. The most common classification strategies applied in the development of core collection are hierarchical methods like centroid, UPGMA, Ward's minimum variance method and its modifications (Taba *et al.* 1998, Upadhyaya *et al.* 2001a, Franco *et al*. 2005, 2006; Weihai 2008) and non hierarchical methods like K-means (Grenier 2001a,b; Amalraj *et al*. 2006). Each method has its own advantages and disadvantages. All these classical clustering techniques perform better and give good approximation for well characterised data sets with compact, isolated clusters. But in real situation, the concerned empirical data such as germplasm accessions data, does not possess the ideal conditions distinct cluster structures required for cluster definition. Presence of outliers also may lead to inappropriate clustering. In such situation outliers are to be removed to carry out the analysis. Self Organizing Mapping (SOM) networks is a neural network clustering method which can take care of outlisers and other imperfections (Mangiameli *et al*. 1996).

Several studies presented comparisons between the clustering performance of SOM and classical clustering algorithms (Chen *et al*. 1995, Mangiameli *et al*. 1996, Waller *et al*. 1998). Mangiameli *et al*. (1996) compared performance of SOM with seven hierarchical clustering methods using 252 data sets having various levels of imperfections such as data dispersion, outliers, irrelevant variables, and non uniform cluster densities. They revealed that SOM was superior in accuracy and robustness as compared to seven clustering methods. However, this method has not been used so far in development of germplasm core in any crop. Therefore, in the present study also a comparison of SOM and conventional clustering techniques was made to develop a germplasm core in castor (*Ricinus communis L*).

2. MATERIALS AND METHODS

2.1 Materials

The Directorate of Oilseeds Research, Hyderabad, India collected castor germplasm accessions from all over the country through explorations and introduction from 36 countries. A total of 3003 accessions were considered for the study. Data on 14 quantitative traits viz., plant height, number of nodes, total length of primary spike, length of primary spike covered by capsules, total number of spikes/plant, days to flowering, days to 50% maturity, 100-seed weight, seed yield/plant at 120 days after planting (DAP), seed yield/plant at 150DAP, seed yield/plant at 180 DAP, seed yield/plant at 210DAP, total seed yield/plant and oil content (%), were considered for establishment of castor core collection. The preliminary analysis revealed great variability for all the 14 quantitative traits.

2.2 Methodology

The two broad steps involved in the construction of core were

1. Grouping/clustering of the accessions.
2. Selecting the accessions with in a cluster with a predetermined sampling strategy.

2.2.1 Grouping /Clustering

As the genebank data is highly diverse and consisting of rare types which depart from the ideal conditions of traditional clustering techniques, SOM was considered for clustering whole germplasm accessions. Apart from this Ward's method and K-means clustering methods were considered for comparison. Earlier studies on clustering performance suggested that, in order to maximise the efficiency of cluster analysis, a two-stage approach should be contemplated (Punj and Stewart 1993, Balakrishnan *et al*. 1996). In the present study also two–stage clustering algorithm was attempted by combining the two methods in a sequential order, in order to improve classification reliability and validity (Punj and Stewart 1993).

Self –Organising mapping networks (SOM)

SOM (Kohonen 2001) is a specific family of neural networks using unsupervised training, where no target output is provided and the network evolves until stabilisation. It is an effective modelling tool for the visualisation of high dimensional data. Non-linear statistical relationships between high dimensional data are converted into simple geometric relationships of their image points on a low dimensional display, usually a two dimensional grid of nodes.

A general architecture of SOM (Fig. 1) consists of a set of input nodes, output nodes and weight parameters. Each input node is fully connected to every output node via a variable connection. A weight parameter is associated with each of these connections. The weights between the input nodes and output nodes are iteratively changed during the learning phase until a termination criterion is satisfied. For each input vector, there is one associated winner node on the output map.

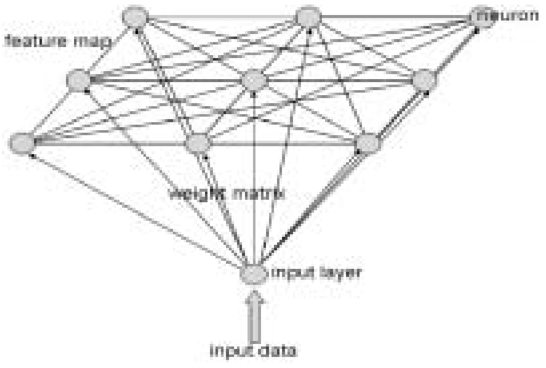

Fig. 1. General SOM architecture

Batch- SOM algorithm

In the present study, Batch Self–Organising map (Batch-SOM) was considered for clustering. In this procedure each map node was mapped to a weighted average of the fixed data points, based on the current winner assignment. This important learning rule is named "Batch map"; it is significantly faster in terms of computation time. The algorithm is briefly explained here as give by Naidoo (2007).

Let S be a fundamental set of some items $x(i)$ and $d[x(i), x(j)]$ be some distance measure between $x(i)$, $x(j) \in S$. The set median m over S shall minimize the expression

$$D = \Sigma_{x(i)\in S} d[x(i), m] \qquad (1)$$

The reason for naming the median is that it is relatively easy to show that the usual median of real numbers is defined by equation (1) whenever the distance measure satisfied the following

$$d[x(i), x(j)] = [x(i) - x(j)] \qquad (2)$$

In the above case it was assumed that m belongs to the fundamental set S however it is possible to find a hypothetical item m such that D attains its absolute minimum value. In contrast to the set median the term generalized median is used to denote the value of m that gives the absolute minimum value for D as it was shown earlier that the convergence limits during the learning process were

$$m_i(t + 1) = m_i(t) + h_{ci}(t)\,[x(t) - m_i(t)] \quad (3)$$

It is now useful in understanding what the convergence limits m_i^* represent. Assume that the convergence to some ordered state is true m then the expected values of $m_i(t + 1)$ and $m_i(t)$ must be equal.

In other words in the stationery state

$$E\{h_{ci}(x - m_i^*)\} = 0 \text{ for all value of } i$$

In the simplest case $h_{ci}(t)$ was defined as follows:

$$h_{ci} = \begin{cases} 1 \text{ if } i \text{ belongs to some topological} \\ \qquad\qquad\qquad \text{neighbourhood set } N_c \\ 0 \;\; \text{otherwise} \end{cases}$$

The convergence limit m_i^* can be defined as follows

$$m_i^* = \frac{\int_{V_i} xp(x)\, d(x)}{\int_{V_i} p(x)\, d(x)} \quad (4)$$

where V_i is the set of those values in the integrands that are able to update vector m_i, in other words the winner node c for each $x \in V_i$ must belong to neighbourhood set N_i of cell i.

The iterative process in which a number of samples of x is first classified into the respective V_i regions and the updating of the m_i^* is made iteratively as defined by equation 4 can be expressed in the following steps.

Step 1: For the initial reference vectors take the first K training samples, where K is the number of reference vectors.

Step 2: For each map unit i, collect a list of copies of those training samples x whose nearest reference vector belongs to unit i.

Step 3: Take for each new reference vector the mean over the union of the lists in N_i.

Step 4: Repeat step 2 and step 3 until convergence or the maximum iterations.

If a general neighbourhood function h_{ji} is used and $\overline{x}_j$ is the mean of the $x(t)$ in the Voronoi set V_j, then it shall be weighted by the number n_j of sample V_j and the neighbourhood function.

The following equation is obtained:

$$m_i^* = \frac{\Sigma_j n_j h_{ji} \overline{x}_j}{\Sigma_j n_j h_{ji}} \quad (5)$$

where the sum over j is taken for all units of the self organizing map, or if h_{ji} is truncated over the neighborhood set N_j in which it is defined.

For case in which no weighting in the neighbourhood is used equation (5) becomes

$$m_i^* = \frac{\sum_{j \in N_j} n_j \overline{x}_j}{\sum_{j \in N_j} n_j \overline{x}_j} \quad (6)$$

It should be noticed that algorithm contains no learning rate parameter: therefore it has no convergence problems and yields stable asymptotic values for m_i compared to other Kohonen SOM algorithms.

2.2.2 Sampling Strategies

The purpose of sampling accessions from groups is to recover most of the diversity existing in genebank, while maximising variance and the distance between accessions in a sample. A large core collection would face the same management problems as that of the whole collection. Similarly, a core collection that is too small to contain a significant fraction of the base collections' diversity would not serve the purpose of conserving maximum diversity of a base collection in a core set (Balakrishnan 2007).

A sampling strategy involves defining a sampling intensity, a sampling method, and an allocation method. (Thompson 2002, Franco *et al*. 2005). The sampling intensity defines the sample size as a proportion of a population size. A core collection is always substantially smaller than the base or whole collection and in practice most of the core collections developed so far were between 5 per cent to 20 per cent of the size of the base collection (Brown 1989, Brown and Spillane 1999, Van Hintum *et al*. 2000, Franco *et al*. 2005). Brown (1989 b) suggested that it should be no more than 10% of the base collection and always less than 2000 entries. Balakrishnan *et al*. (2000) proposed a regression model to explain the progress curve of the cumulative Genralised Sum of Squares (GSS) per cent (diversity) accounted for by the selected accessions. The shape of the rate-of-progress-curve indicates that an optimum core size could be fixed until 50% of the total GSS or diversity was covered (for detailed procedure Balakrishnan *et al.* (2000) may be referred).

In the present study, we considered 10 per cent, 15 per cent, 20 per cent sampling sizes along with the sampling size of 524 obtained by constructing progress-curve based on methodology given by Balakrishnan *et al*. (2000). Fig. 2 illustrates the shape of the rate of progress curve indicating that an optimum core size can be fixed at 524 for the castor germplasm accessions. The most commonly used sampling method viz., simple random sampling without replacement (SSRWOR) was considered for the study.

An allocation method determines the number of accessions to be selected from each cluster. Brown (1989) described three allocation methods: constant (C) where the number of accessions allocated is the same for all the groups; proportional (P) where the number of accessions allocated is proportional to the number of accessions in each group; and logarithmic (L) where the number of accessions allocated is proportional to the logarithm of the number of accessions in each group. Other methods have been proposed more recently such as the D allocation strategy (Franco *et al.* 2005), where Gower's distances are used between accessions within each cluster, stepwise clustering (Hu *et al.* 2000), where successive dendrograms are constructed and some accessions are selected in each step or the M-strategy (Schoen and Brown 1993), where the number of observed alleles at each marker locus is maximized. In the present study two allocation methods viz., logarithmic and proportional allocation methods were considered.

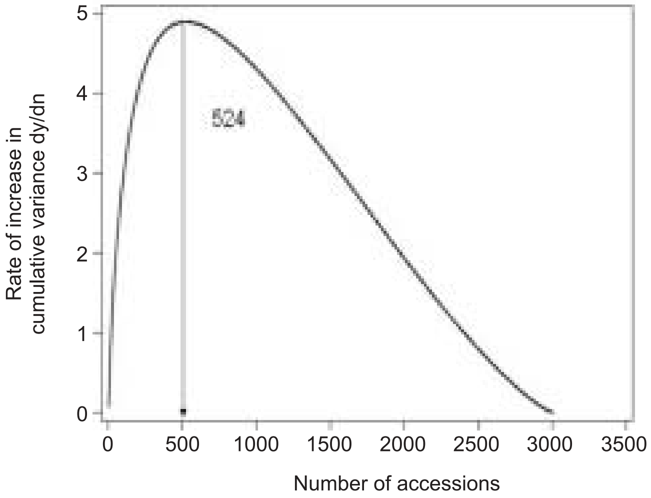

Fig. 2. Determination of number of accessions based on progress curve

2.2.3 Evaluation of the Core collection

A homogeneity test (F test) for variances and a t test of means ($\alpha = 0.05$) were performed to determine the significant differences in traits between core collection and whole collection. Based on the calculated results of t test, F test, range and coefficient of variation, the following four important evaluating parameters were considered.

Mean difference percentage (MD) = $(S_t/n)\times100$, Where S_t is the number of traits which have significant difference ($\alpha = 0.05$) of means between whole and core collection; n is the total number of traits.

Variance difference percentage (VD) = $(S_F/n)\times100$ where S_F is the number of traits which have significant difference ($\alpha = 0.05$) of variances between whole and core collection; n is the total number of traits.

Coincidence rate of range percentage (CR) $= \frac{1}{n}\sum_{i=1}^{n} R_{c(i)}/R_{e(i)} \times 100$, where, $R_{c(i)}$ is the range of the ith trait of core subset; $R_{e(i)}$ is the range of the corresponding trait of entire germplasm collection; n is total number of traits.

Variable rate of coefficient of variation percentage (VR) $= \frac{1}{n}\sum_{j=1}^{m} CV_{c(i)}/CV_{ce(i)} \times 100$ where, $CV_{c(i)}$ is the coefficient of variation of the i^{th} trait of core subset; $CV_{ce(i)}$ is the coefficient of variation of the corresponding trait of the entire germplasm collection; n is total number of traits.

Hu *et al.* (2000) treated the CR and MD as the most important evaluating parameters. Wang *et al.* (2007) assessed the evaluating parameters of rice core collection using Monte Carlo simulations combined with mixed linear models and concluded that the CR has the highest validity, stability and sensibility could be the optimal evaluating parameter. MD can be used as determination parameter for reliability judgment of core collection. VR parameter as an important referential parameter for evaluating the degree of variation in core collection. The core collection can be considered to represent the genetic diversity of the initial collection if MD ≤ 20 per cent and CR is ≥ 80 per cent and the corresponding sampling percentage could be a considered to be a sampling percentage. Moreover, in the same sampling percentage, smaller MD leads to more representative core collections, and core collection with larger CR, VD and VR is considered to provide good representative of the genetic diversity of the whole collection (Hu *et al.* 2000, Wang *et al.* 2007). SAS 9.2 and SAS/Entrprise Miner 6.0 statistical software was used in the present study.

3. RESULTS AND DISCUSSION

3.1 Results

From the whole collection of 3003 accession 40 experimental cores were developed. Of these, 20 experimental cores generated with logarithmic allocation method, were rejected as they were having more than 20 per cent MD. The results of 20 cores, generated using proportional allocation method were presented in Table 1. MD of all the experimental cores were zero indicating no significance difference between the means of quantitative traits of experimental cores and that of whole germplasm collection.

Table 1. Comparison of experimental cores

Experimental core	MD	VD	CR	VR
C1A1S1	0	7.1	72.8	97.0
C1A1S2	0	7.1	73.9	97.7
C1A1S3	0	14.3	81.3	101.2
C1A1S4	0	7.1	78.7	98.0
C1A2S1	0	21.4	73.5	99.7
C1A2S2	0	14.3	74.8	97.2
C1A2S3	0	21.4	81.9	99.1
C1A2S4	0	28.6	86.4	103.8
C2A1S1	0	7.1	79.2	102.1
C2A1S2	0	7.1	77.8	100.0
C2A1S3	0	7.1	81.3	100.2
C2A1S4	0	7.1	77.5	101.6
C2A2S1	0	28.6	75.7	100.0
C2A2S2	0	14.3	71.8	96.4
C2A2S3	0	21.4	78.6	98.9
C2A2S4	0	28.6	83.6	101.7
C3S1	0	14.3	72.8	99.2
C3S2	0	7.1	77.9	99.7
C3S3	0	42.9	89.6	103.8
C3S4	0	21.4	79.8	101.5

C1 = Ward's C2 = K-means C3= Batch-SOM

A1 = Single-stage A2 = Two-stage

S1 = 301 S2 = 450 S3= 601 S4 = 524

It was observed that in Ward's clustering method (Single- Stage) for the four sample sizes, the MD was zero, VD and VR had increased with the sample sizes, but the CR was less than 80 per cent, except in case of experimental core C1A1S3 where CR was 81.3 per cent, hence this core was considered.

Ward's clustering method with two-stage algorithm yielded higher VD and VR with respect to corresoponding sample sizes of single-stage algorithm and CR more than 80 per cent for C1A2S3 and C1A2S4. However, CR (86.4 percent), VD (28.57) and VR (103.8) were higher for experimental core C1A2S4 as compared to C1A2S3.

For the K-means clustering method increase of VD and VR with the sample size was not observed. For single- stage alogorithm VD was same for all sample sizes. VR was higher as compared to Ward's method only C2A1S3 could satisfy the desired condition more than 80 per cent CR.

Similarly, the K-means clustering method with two stage algorithm C2A2S4 with CR of 83.6 per cent could be chosen as other sample sizes could not satisfy the condition for the CR.

Irrespective of the clustering method two-stage clustering algorithm performed better with respect to extracting higher variance differences as compared to single–stage. And it was observed that sample size 524 (optimum core size for conserving the maximum diversity) was found to be having higher CR, VD and VR for both the clustering methods.

The evaluating parameters of the experimental cores developed by Batch-SOM method indicated that CR, and VR increased with sample sizes but no such

Table 2. Selected Experimental cores from different clustering methods.

Experimental core	MD	VD	CR	VR
C1A1S3	0	14.3	81.3	101.2
C1A2S4	0	28.6	86.4	103.8
C2A1S3	0	7.1	81.3	100.2
C2A2S4	0	28.6	83.6	101.7
C3S3	0	42.9	89.6	103.8

pattern was observed with VD. Even though VD and VR were higher for all the four experimental cores. C3S3 alone could satisfy the condition of CR more than 80 per cent. Thus the experimental core C3S3 with 89.6 per cent has been selected from this method.

A comparison of all the experimental cores selected from each clustering methods presented in Table 2 indicated that among the selected cores from different clustering methods C3S3 has the highest CR (89.6 per cent) and VD (42.9 per cent) and VR (103.8 per cent) followed by C1A2S4 which was on par with C3S3 with respect to VR but less in CR and VD. Thus the results indicated that Batch–SOM method performed better as compared to Ward's and K-means clustering methods in terms of extracting the diversity present in the whole germplasm collection. Thus the experimental core developed by Batch-SOM with a sample size of 601 accessions with proportional allocation methods was selected as a preliminary core. For the better management and conservation of diversity, size of the core should contain a mangeable number of accessions. So, as size of the preliminary core is difficult to manage, thus, a core of 165 accessions was developed utilizing the same methodology and evaluating parameters were calculated with respect to preliminary core and whole collection (Table 3). The result indicated that MD was zero implying that there was no significant differences between the means of the final core developed and of whole collection as well as of preliminary core. Higher VD and VR signifies that a very high degree of variation was conserved in final core with respect to whole germplasm collection. Thus the final core comprising of 165 accession was considered as a valid core for conserving maximum diversity of the whole collection.

Table 3. Comparison of final core with preliminary core and whole collection.

Evaluation Parameter	Preliminary core (per cent)	Whole collection (per cent)
MD	0	0
VD	78.6	92.9
CR	112.6	98.1
VR	115.9	120.3

3.2 Discussion

Results proved the effectiveness of SOM method of clustering over the classical methods in extracting maximum diversity of the whole collection. The CR, VD and VR per cent were highest in SOM in the present study. Thus, efficient algorithms like Self Organizing mapping Neural Networks can be used for the development of core subsets from the whole germplasm collection to establish an effective core conserving the maximum diversity exist in the whole germplasm collection. However, there are limitations of SOM method such as it is more efficient for grouping large datasets than the smaller datasets, the initial number of cluster is to be specified and no theoretical frame work for selection of optimum number of parameters is defined. The training algorithm not optimizes the error function (Mangiameli *et al*. 1996).

REFERENCES

Amalraj, V.A., Balakrishnan, R., Jebdadhas, A. Williuam and Balasundaram, N. (2006). Constituting a core collection of *Saccharum spontaneum* L.- comparison of three stratified random sampling procedures *Genet. Resour. Crop Evol.*, **53,** 1563-1572.

Balakrishnan, P.V.S., Cooper, M.C., Jacob, V.S. and Lewis, P.A. (1996). Comparative performance of the FSCL neural net and K-means algorithm for market segmentation. *Euro. J. Op. Res.*, **93,** 346-357.

Balakrishnan, R., Nair, N.V. and Sreenivasan, T.V. (2000). A method for establishing a core collection of Saccharum offcinarum L. germplasm based on quantitative-morphological data. *Genet. Resour. Crop Evol.,* **47,** 1-9.

Balakrishnan, R. (2007). Statistical methods for constituting core subsets of germplasm collections. *Stat. Appl.*, **5(1 and 2)**, 37-45.

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

Brown, A.H.D. and Spillane, C. (1999). Implementing core collections-principles, procedures, progress, problems and promise. Pp. 1-9 in *Core collections for today and tomorrow* (R.C. Johnson and T. Hodgkin, eds.), International Plant Genetic Resources Institute, Rome, Italy.

Chen, S.K., Mangiameli, P. and West, D. (1995). The comparative ability of self-organizing neural networks to define cluster structure. *Omega, Int. J. Manage. Sci*., **23,** 271-279.

Chung, H.-K., Kim, K.-W., Chung, J.-W., Lee, J.-R., Lee, S.-Y., Dixit, A., Kang, H.-K., Zhao, W., McNally, K.L., Hamilton, R.S., Gwag, J.-G. and Park, Y.-J. (2009), Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. *J. Integrative Plant Biol.*, **51,** 1116-1125.

Escribano, P., Viruel, M.A. and Hormaza, J.I. (2008). Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya (*Annona cherimola*, Annonaceae), an underutilised subtropical fruit tree species. *Ann. Appl. Biol.,* **153(1)**, 25-32.

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

Franco, J., Crossa, J., Warburton, M.L. and Taba, S. (2006). Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. *Crop Sci.*, **46,** 854-864.

Frankel, O.H. and Brown, A.H.D. (1984). Current plant genetic resources — a critical appraisal. In: *Genetics: New Frontiers*, Vol. IV, Oxford and IBH Publ. Co., New Delhi, India.

Grenier, Hamon, C.P. and Bramel-Cox, P.J. (2001a). Core collection of sorghum: I. Stratification based on eco-geographical data. *Crop Sci.*, **41,** 234-240.

Grenier, Hamon, C.P. and Bramel-Cox, P.J. (2001b). Core collection of sorghum: II. comparison of three random sampling strategies *Crop Sci.*, **41,** 241-246.

Hu, J., Zhu, J. and Xu, H.M. (2000). Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theo. Appl. Genet.,* **101,** 264-268.

Jana, S. and Addala, K.R.V. (1998). Evolving issues in genetic resources conservation Pp. 1-12 in Triticee III (A.A. Jaradat, ed.). Science Publishers Inc., Enfield, New Hampshire, USA. Johnson R.C. and T. Hodgkin 1999. *Core Collections for Today and Tomorrow*. International Plant Genetic Resources Institute, Rome, Italy.

Kohonen, Teuvo (2001). *Self Orgnizing Maps* 3 Ed. Springer Series in Information Sciences, **30**, Springer, Berlin, Heidelberg, New York.

Mangiameli, P., Shaw, K., Chen and David, West (1996). A comparison of SOM neural network and hierarchical clustering methods. *European J. Oper. Res.*, **93,** 402-417.

Naidoo, A.G.V. (2007). A multidimensional measure of poverty in South Africa. Unpublished Thesis. © University of Pretoria.

Punj, G. and Stewart, D.W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *J. Market. Res.,* **20,** 134-148.

Schoen, D.J. and Brown, A.H.D. (1993). Conservation of allelic richness in wild crop relatives is aided by assessment of genetic-markers. *Proceedings of the National Academy of Sciences USA,* **90,** 10623-10627.

Taba, S., Diaz, J., Franco, J. and Crossa, J. (1998). Evaluation of carribean maize accessions to develop a core subset. *Crop Sci*., **38**., 1378-1386.

Upadhyaya, H.D., Bramel, P.J. and Sube Singh (2001). Development of a chickpea core subset using geographic distribution and quantitative traits. *Crop Sci.*, **41,** 206-210.

Upadhyaya, H.D., Ortiz, R., Bramel, P.J. and Sube Singh (2003). Development of a groundnut core collection using taxonomical, geographical, and morphological descriptors. *Genet. Resour. Crop Evol*., 139-148.

van Hintum, T., Brown, A.H.D., Spillane, C. and Hodgkin, T. (2000). Core collections of plant genetic resources. *IPGRI Technical Bulletin* No. 3. International Plant Genetic Resources Institute, Rome, Italy.

Vollmann, J., Grausgruber, Stift, Dryzhyruk and Lelley (2005). Genetic diversity in camelina germplasm as revealed by seed quality characteristics and RAPD polymorphism. **124(5)**, 446-453.

Waller, N.G. and Kaiser, H.A. (1998). A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika*, **63,** 5-22.

Wang, J.C., Hu, J., Zhang, C.F. and Zhang, S. (2007). Assessment on evaluating parameters of rice core collections constructed by genotypic values and molecular marker information. *Rice Sci.*, **14(2)**, 101-110.

Weihai, M., Yi Jinxin and Darasinh Sihachakr (2008). Development of core subset for the collection of Chinese cultivated eggplants using morphological based passport data. *Plant Genet. Resour. : Characterization utilization,* **6**, 33-40.

Xu, H.M., Mei, Y.J., Hu., J., Zhu, J. and Gong, P. (2006). Sampling a core collection of island cotton (*Gossypium barbadense* L.) based on the genotypic values of fiber traits. *Genet. Resour. Crop Evol.,* **53(3)**, 515-521.

Zhao, Weiguo, Cho, Gyu-Taek, Ma, Kyung-Ho Chung, Jong-Wook, Gwag, Jae-Gyun, Park, Yong-Jin (2010). Development of allele –mining set in rice using a heuristic alogroithm and SSr genotype data with least redundancy of the post–genomic era. *Molecular Breeding,* **26**(4), 639-651.