ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2015) **13**(2); 96–103
doi:10.1017/S1479262114000732

An approach to the development of a core set of germplasm using a mixture of qualitative and quantitative data

Rupam Kumar Sarkar[1], Prabina Kumar Meher[1], S. D. Wahi[1], T. Mohapatra[2] and A. R. Rao[1]*

[1]*Indian Agricultural Statistics Research Institute, New Delhi 110012, India and*
[2]*Central Rice Research Institute, Cuttack, Odisha 753006, India*

Received 13 February 2014; Accepted 31 May 2014 – First published online 26 June 2014

Abstract

Development of a representative and well-diversified core with minimum duplicate accessions and maximum diversity from a larger population of germplasm is highly essential for breeders involved in crop improvement programmes. Most of the existing methodologies for the identification of a core set are either based on qualitative or quantitative data. In this study, an approach to the identification of a core set of germplasm based on the response from a mixture of qualitative (single nucleotide polymorphism genotyping) and quantitative data was proposed. For this purpose, six different combined distance measures, three for quantitative data and two for qualitative data, were proposed and evaluated. The combined distance matrices were used as inputs to seven different clustering procedures for classifying the population of germplasm into homogeneous groups. Subsequently, an optimum number of clusters based on all clustering methodologies using different combined distance measures were identified on a consensus basis. Average cluster robustness values across all the identified optimum number of clusters under each clustering methodology were calculated. Overall, three different allocation methods were applied to sample the accessions that were selected from the clusters identified under each clustering methodology, with the highest average cluster robustness value being used to formulate a core set. Furthermore, an index was proposed for the evaluation of diversity in the core set. The results reveal that the combined distance measure A_1B_2 – the distance based on the average of the range-standardized absolute difference for quantitative data with the rescaled distance based on the average absolute difference for qualitative data – from which three clusters that were identified by using the *k*-means clustering algorithm along with the proportional allocation method was suitable for the identification of a core set from a collection of rice germplasm.

Keywords: consensus clustering; core set; germplasm; mixture data; robustness; single nucleotide polymorphisms

Introduction

A vast collection of crop-related global germplasm includes traditional landraces, modern cultivars and wild cultivars. However, only a fraction of these germplasm collections could be protected and maintained in gene banks. Frankel and Brown (1984) introduced the concept of a core collection as a subset of a larger germplasm collection that represents genetic and phenotypic diversity. Existing methodologies for the development of a core set are either based on qualitative or quantitative data. Occasionally, transformations are applied on quantitative data to make them qualitative or vice

* Corresponding author. E-mail: arrao@iasri.res.in; rao.cshl.work@gmail.com

versa to avoid the difficulty of handling mixture data (Kim *et al.*, 2007).

Various clustering methodologies are applied to obtain homogeneous strata either based on qualitative or quantitative data. However, the results have been shown to be highly dependent on clustering methodologies, and mostly heuristic methods (Kim *et al.*, 2007) have been followed to determine homogeneous strata in a germplasm collection. Various factors that need to be addressed for the development of a core set include the size of the core set, the formation of homogeneous strata and the sampling strategy (van Hintum and Th, 1999).

In the past, many studies have described the development of core sets from a large collection of germplasm, namely the development of a core set from the United States Department of Agriculture (USDA) rice germplasm (Yan *et al.*, 2007) and a rice mini-core from the USDA core collection (Agrama *et al.*, 2009) using PowerCore software (RDA-Genebank Information Center; http://www.genebank.go.kr/eng/PowerCore/PowerCore_Software.zip) (Kim *et al.*, 2007). Gangopadhyay *et al.* (2010) used the principal component score strategy to develop a core set of brinjal germplasm. Sharma *et al.* (2010) evaluated a sorghum mini-core from a core collection of landrace accessions to identify the sources of grain mold and downy mildew resistance. Yu *et al.* (2012) developed a core set of cotton germplasm with a genome-wide coverage of marker data. Wen *et al.* (2012) investigated how the tropical maize race *Tuxpeno* could be exploited in future maize improvement using genome-wide single nucleotide polymorphisms (SNPs). Gibert and Cortes (1997) presented the properties and details of distance matrices obtained by weighting qualitative and quantitative variables for cluster analysis. However, weighting of distance matrices from different sources is problematic because the objective choices of weighting parameters are often difficult. Crossa and Franco (2004) reviewed genomic classification techniques as well as statistical models based on mixed distribution models. Doring *et al.* (2004) proposed a fuzzy clustering procedure for mixture data. Sarkar *et al.* (2011) compared the performance of different clustering procedures based on mixture data. However, the identification of the optimum number of clusters with full utilization of mixture data for the development of a core set remains a challenge.

To our knowledge, most of the existing methodologies for the development of a core set are either based on qualitative or quantitative traits. Moreover, identification of a suitable distance measure, clustering methodology, number of clusters, allocation strategy and evaluation criteria for the development of a core set based on the mixture data of germplasm is yet to be fully explored. Therefore, in the present study, a systematic approach was proposed for the development of a core set of germplasm using mixture data. The approach thus developed is illustrated on rice germplasm having both quantitative and qualitative SNP genotyping data.

Materials and methods

The identification of a core collection is a two-step procedure in which the accessions are initially classified into homogeneous strata and then a fraction of accessions from each stratum are selected for core collection by using an appropriate sampling or allocation strategy. To enable clustering techniques to handle a mixture of qualitative and quantitative data, first, distances are calculated separately for qualitative and quantitative data using relevant measures. Then, these distance matrices are directly combined and used as inputs for cluster analysis.

In this study, a dataset comprising 219 salt-tolerant rice germplasm accessions having 14 agronomic/phenotypic characteristics and 2915 genome-wide SNPs (coded as 0 and 2 for dominant and recessive homozygotes, respectively, and 1 for a heterozygote for each individual) was considered.

Distance measures

The following three distance measures were considered to determine the distances between the accessions based on quantitative data:

(1) Distance based on the average of the range-standardized absolute difference:

$$A_1 = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{r_k},$$

where x_{ik} and x_{jk} are the ith and jth accessions of the kth quantitative variable; r_k is the range of the kth variable; and p is the total number of quantitative variables (Gower, 1971).

(2) Distance based on Pearson's correlation:

$$A_2 = (1 - r_{ij}^2),$$

where r_{ij} is the product moment correlation (similarity) between ith and jth accessions, thus dissimilarity = 1 − similarity.

(3) Rescaled distance based on the standardized score:

$$A_3 = \sum_{k=1}^{p}\frac{\left[\frac{x_{ik}-x_{jk}}{\sigma_k}\right]^2}{\max\left(d_{ij}^{*}\right)},$$

where σ_k is the standard deviation of the kth variable and max (d_{ij}^{*}) is the maximum of the distances between two accessions in the entire dataset.

https://doi.org/10.1017/S1479262114000732 Published online by Cambridge University Press

The following two different distance measures were considered to determine the distance between the accessions based on qualitative data:

(1) Distance based on the average mismatch:

$$B_1 = \frac{1}{m}\sum_{k=1}^{m} d_k,$$

where $d_k = 0$, if $y_{ik} = y_{jk}$, else $d_k = 1$ (Gower, 1971).

(2) Rescaled distance based on the average absolute difference:

$$B_2 = \frac{3}{2} \times \frac{\frac{1}{m}\sum_{k=1}^{m} |y_{ik} - y_{jk}|}{1 + \frac{1}{m}\sum_{k=1}^{m} |y_{ik} - y_{jk}|},$$

where y_{ik} and y_{jk} are the ith and jth accessions of the kth qualitative variable and m is the total number of qualitative variables. The distance B_2 is a modified measure of Munneke *et al.* (2005). The modification is done so that the value of B_2 lies in the range [0, 1].

The range of elements in the three quantitative distance matrices (A_1–A_3) and two qualitative distance matrices (B_1 and B_2) lies between 0 and 1. Thus, the various combined distance matrices for mixture data are computed by summing up the distance matrices corresponding to the qualitative and quantitative data, which are defined as follow:

$$A_1B_1 = ((a_{1ij}) + (b_{1ij})),$$

$$A_2B_1 = ((a_{2ij}) + (b_{1ij})),$$

$$A_3B_1 = ((a_{3ij}) + (b_{1ij})),$$

$$A_1B_2 = ((a_{1ij}) + (b_{2ij})),$$

$$A_2B_2 = ((a_{2ij}) + (b_{2ij})),$$

$$A_3B_2 = ((a_{3ij}) + (b_{2ij})),$$

where (a_{1ij}), (a_{2ij}), (a_{3ij}), (b_{1ij}), (b_{2ij}) and (b_{3ij}) represents the ijth elements of matrices A_1, A_2, A_3, B_1, B_2 and B_3, respectively. These qualitative, quantitative and combined distance matrices are used as inputs for clustering analysis. In this study, seven (five hierarchical and two partitioned) different clustering procedures, namely single linkage, complete linkage, unweighted pair-group method with arithmetic mean, weighted average, Ward's method, *k*-means and partitioning around the medoids, were considered to find the optimum number of homogeneous clusters. Here, the approach of Monti *et al.* (2003) was followed for assessing the stability of clusters by bootstrapping, and 1000 bootstrap samples were drawn from the distance matrices for each set of the cluster number $k = \{2, \ldots, 10\}$. A consensus clustering result was obtained by taking the ratio of the number of times any two observations are found together in the same cluster to the total number of times that are selected together in the bootstrap samples. As each clustering procedure exhibits different cluster memberships of individuals, the consensus clustering results obtained from different clustering procedures are merged together to obtain a merged-consensus clustering result (Simpson *et al.*, 2010). The merged-consensus clustering result is obtained by taking the average of the consensus clustering results for a particular cluster number k. Due to the absence of any *a priori* information on the clustering pattern, equal weights are given to each consensus clustering result, i.e. equal importance is given to each of the clustering procedure.

Optimum number of clusters

Mostly, *a priori* information is used for the determination of the number of clusters to classify the accessions, but in the absence of such information, it is beneficial to identify the optimum number of clusters. Moreover, identifying the optimal number of clusters is one of the most challenging issues and essential for effective and efficient clustering (Everitt, 1979). The optimal number of clusters (k) is estimated as the value of k at which the change in the area under cumulative density function (CDF) (ΔK) calculated across a range of possible values of k is largest. Let us suppose that M indicates a merged-consensus clustering result of order $N \times N$. Then, an empirical CDF, defined over the range [0, 1], is given by:

$$\mathrm{CDF}(c) = \frac{\sum_{i<j} 1\{M(i,j) \leq c\}}{\frac{N(N-1)}{2}},$$

where $1\{\ldots\}$ denotes an indicator function, $M(i, j)$, with (i, j) being the entry of the merged-consensus matrix M. The area under the CDF corresponding to M is computed using the formula:

$$\mathrm{AUC} = \sum_{i=2}^{m} [x_i - x_{i-1}]\mathrm{CDF}(x_i),$$

where $\{x_1, x_2, \ldots, x_m\}$ is the ordered set of entries of the merged-consensus matrix M, with $m = N(N-1)/2$ (Monti *et al.*, 2003).

Cluster robustness

After determining the optimal number of clusters, the best-fitted clustering pattern of germplasm is determined

https://doi.org/10.1017/S1479262114000732 Published online by Cambridge University Press

based on cluster robustness. The robustness of clusters under any clustering procedure is calculated by taking the average of the merged-consensus result of those individuals falling in the same group using the formula (Simpson *et al.*, 2010):

$$m(k) = \frac{1}{\frac{N_k(N_k-1)}{2}} \sum_{i<j(\in I_k)} M(i,j).$$

The average cluster robustness value is calculated across the k clusters using the clustering algorithm to choose the one that is best fitted to the data.

Allocation methods

The second and final stage for the development of a core set is to select the accessions from homogeneous groups based on a suitable sampling or allocation strategy. van Hintum *et al.* (2000) and Hu *et al.* (2000) used different sampling strategies, namely proportional allocation (P strategy), log frequency allocation (L strategy), constant allocation (C strategy) and simple random sampling (R strategy) for the identification of a core set. During this stage, the accessions from the identified robust clusters are sampled by using the following three different allocation methods:

(1) Proportional allocation

$$n_i = \left[n \times \frac{N_i}{\sum_{i=1}^{g} N_i} \right]$$

(2) Log-proportional allocation

$$n_i = \left[n \times \frac{\log(N_i)}{\sum_{i=1}^{g} \log(N_i)} \right],$$

where n_i is the number of accessions selected for the core set from the ith cluster; N_i is the number of accessions in the ith cluster; n is the size of the core set; g is the total number of clusters and the parentheses '[ ]' represent the nearest integer function.

(3) Random allocation of single entry (RASE). Here, no optimal number of clusters is determined and the accessions are grouped into the number of clusters equals to the size of the core set. A single entry is then selected from each of the cluster to construct the core set.

Evaluation of a core set

For quantitative data, the efficiency of methodologies for the identification of a core set is evaluated by using different indices, namely mean difference (MD), variance difference (VD), variable rate (VR) and coincidence rate (CR) (Hu *et al.*, 2000). For qualitative data, the aforementioned methodologies are evaluated using the index average polymorphic information content difference (APICD), which is given by:

$$\text{APCID}\,(\%) = \frac{|\overline{P_e} - \overline{P_c}|}{\overline{P_c}} \times 100,$$

where $\overline{P_c}$ and $\overline{P_e}$ are the average polymorphic information content of the core set and the entire set, respectively.

A combined evaluation index (CEI) was proposed by combining the above-mentioned five indices to evaluate the diversity of the core set based on mixture data. The CEI is given by:

$$\text{CEI} = (w_1 M_1 + w_2 M_2)/(w_1 + w_2),$$

where $M_1 = [(100 - \text{MD}) + (100 - \text{VD}) + (100 - \text{VR}_t) + \text{CR}]/4$, with $\text{VR}_t = |100 - \text{VR}|$, $M_2 = (100 - \text{APICD})$; and $w_1 = (N_{quant}/N_T)$ and $w_2 = (N_{qual}/N_T)$ N_{quant}, N_{qual} and N_T are the number of quantitative, qualitative and total number of variables, respectively, with $w_1 + w_2 = 1$ and $N_T = N_{quant} + N_{qual}$. The CEI represents the percentage of

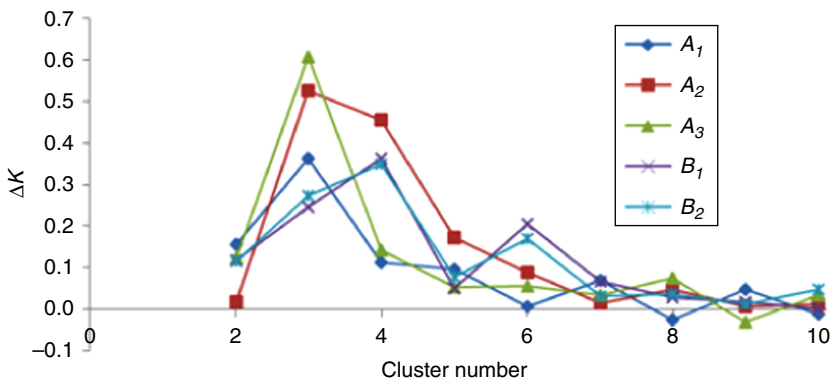

Fig. 1. Graphical representation of ΔK against the cluster numbers (k) for qualitative and quantitative data separately using the corresponding distance measures.

https://doi.org/10.1017/S1479262114000732 Published online by Cambridge University Press

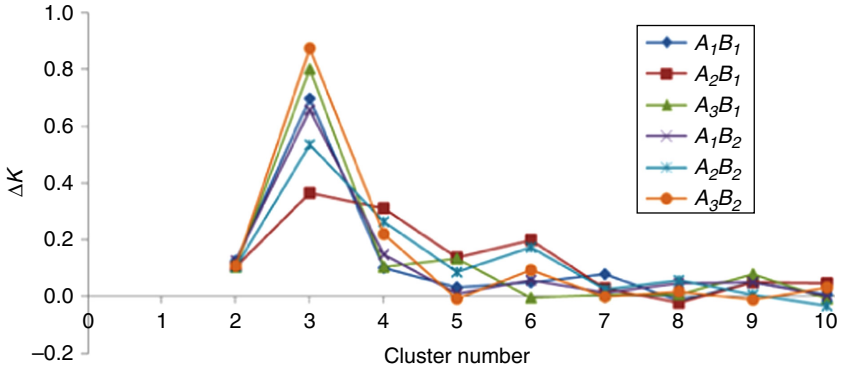

Fig. 2. Graphical representation of ΔK against the cluster numbers (k) for mixture data using combined distance measures.

resemblance between the core set and the entire set. The value of the CEI ranges between 0 and 100. Moreover, the value of 100 corresponds to the best representativeness of the entire population. The difference between the CEI under the proportional allocation, log-proportional allocation and RASE methods for all the distance measures is tested by using a large sample z-test.

All the required coding is done in R software. For the consensus and merged-consensus clustering results, the 'clusterCons' package was used (Simpson, 2010), and to sample the accessions for the core set, the 'ccChooser' package (Studnicki and Debski, 2012) in R software was used.

Results

The consensus clustering methodology was applied on the qualitative and quantitative data separately using the three distance measures for quantitative data (A_1–A_3) and two distance measures for qualitative data (B_1 and B_2). In Fig. 1, the values of ΔK were plotted against the cluster number (k). As shown in Fig. 1, the data consisted of three and four groups based on the quantitative and qualitative distance measures, respectively. Thus, the use of qualitative or quantitative data alone may result in widely different core sets. Moreover, dropping or transforming (qualitative to quantitative and vice versa) any kind of the variables from the analysis may result in a potential loss of information.

The plot of ΔK values against the number of clusters formed by the combined qualitative and quantitative data is shown in Fig. 2. It can be observed that for all the combined distance measures, the peak value of ΔK for the number of clusters was found to be equal to 3. So, the problem of choosing the number of clusters while using qualitative or quantitative data alone, in the case of disagreement, can be resolved by considering the combined distance measures based on mixture data.

The average cluster robustness values for the different clustering methodologies and combined distance measures are given in Table 1. For a given combined distance measure, the clustering methodology with the highest average cluster robustness value was then chosen for adopting the sampling strategy for the development of a core set. It was found that the k-means clustering algorithm was suitable for grouping germplasm based on the combined distance measures A_1B_1, A_3B_1 and A_1B_2, whereas the complete linkage clustering algorithm was suitable for grouping germplasm based on the combined distance measures A_2B_1, A_2B_2 and A_3B_3.

Table 1. Average cluster robustness values of different clustering methodologies using different combined distance measures

	Clustering algorithm						
Distance measures	Single	Complete	Average	Weighted	Ward's method	k-means	PAM
A_1B_1	0.220	0.742	0.505	0.479	0.742	0.744	0.732
A_2B_1	0.258	0.860	0.850	0.848	0.824	0.845	0.844
A_3B_1	0.223	0.765	0.743	0.762	0.760	0.773	0.765
A_1B_2	0.474	0.708	0.453	0.698	0.712	0.727	0.715
A_2B_2	0.256	0.864	0.838	0.854	0.838	0.850	0.823
A_3B_2	0.244	0.848	0.244	0.820	0.823	0.828	0.792

PAM, partitioning around the medoids.

https://doi.org/10.1017/S1479262114000732 Published online by Cambridge University Press

Table 2. Mean values (with their standard errors in parentheses) of the combined evaluation index (CEI) under the proportional allocation (CEI-P), log-proportional allocation (CEI-LP) and random allocation of single entry (CEI-R) methods and their pairwise absolute difference (AbsD)

	A_1	A_2	A_3	B_1	B_2	A_1B_1	A_2B_1	A_3B_1	A_1B_2	A_2B_2	A_3B_2
CEI-P	93.47 (5.9)	94.84 (4.5)	93.83 (5.0)	98.05 (1.8)	97.95 (1.3)	96.83 (2.1)	96.53 (2.8)	96.95 (2.3)	97.81 (1.6)	95.74 (2.7)	97.05 (2.6)
CEI-LP	93.81 (4.8)	93.68 (5.5)	94.48 (4.5)	89.65 (2.5)	88.66 (2.9)	91.11 (2.8)	93.07 (4.1)	93.21 (4.1)	95.84 (2.3)	93.38 (4.4)	92.58 (4.1)
CEI-R	93.44 (5.3)	94.30 (5.5)	94.99 (4.1)	86.84 (0.3)	89.00 (0.3)	85.61 (0.5)	93.87 (1.0)	89.16 (0.9)	88.33 (0.6)	90.27 (0.4)	95.37 (1.6)
AbsD (P–LP)	0.34	1.16*	0.65	8.74*	9.29*	5.72*	3.46*	3.74*	1.97*	2.36*	4.47*
AbsD (P–R)	0.03	0.54	1.16*	11.21*	8.95*	11.22*	2.66*	7.79*	9.48*	5.47*	1.67*
AbsD (LP–R)	0.37	0.62	0.51	2.81*	0.34*	5.5*	0.80*	4.05*	7.51*	3.11*	2.79*

* $P < 0.05$.

To sample the accessions, three different allocation methods, namely proportional allocation, log-proportional allocation and RASE, were adopted. For the first two allocation methods, accessions were selected from the three clusters identified under each combined distance measure to develop a core set with 20% of germplasm from the entire collection. In contrast, the random sampling of a single entry from each of the 44 clusters (approximately 20% of the total number of germplasm) was done to develop a core set by ignoring the optimal number of clusters using the RASE method. To evaluate the efficiency of the procedures to identify the core set, 500 independent core collections were simulated under each sampling strategy.

The mean values of the CEI, over 500 independent simulation runs, under the proportional, log-proportional allocation and RASE methods are presented in Table 2. The absolute differences in CEI values between the proportional, log-proportional allocation and RASE methods were statistically tested and are given in Table 2. From Table 2, it is evident that the differences in CEI values between the proportional and log-proportional methods and between the proportional and RASE methods were significantly higher and hence the proportional allocation method was best among the three allocation methods for the identification of a diverse core set irrespective of the distance measures used. In addition, the differences in CEI values between the combined distance measures and the qualitative/quantitative distance measures under the proportional, log-proportional and RASE methods are given in Table 3. For the proportional allocation method, the value of the CEI was highest for A_1B_2 among the combined distance measures. However, the CEI values of the distance measure A_1B_2 were significantly different from those of A_1 to A_3, and at the same time they were not significantly different from the CEI values of B_1 and B_2 (Table 3). In contrast, for the log-proportional allocation method, the CEI value of A_1B_2 was highest among the CEI values of all the distance measures (Table 2) and significantly different from the rest (Table 3). Moreover, a core set was constructed through heuristic methods using PowerCore (RDA-Genebank Information Center; http://www.genebank.go.kr/eng/PowerCore/PowerCore_Software.zip). The CEI value of the core set constructed by PowerCore was found to be 89.19, which was the lowest among all the combined distance measures under the proportional and log-proportional allocation methods.

Discussion

The use of qualitative and quantitative data separately to classify germplasm collections may result in different numbers of groups and, hence, different grouping patterns

https://doi.org/10.1017/S1479262114000732 Published online by Cambridge University Press

Table 3. Differences in combined evaluation index (CEI) values between the qualitative/quantitative distance measures and the combined distance measures under the proportional and log-proportional allocation methods

	Proportional allocation						Log-proportional allocation					
	A_1B_1	A_2B_1	A_3B_1	A_1B_2	A_2B_2	A_3B_2	A_1B_1	A_2B_1	A_3B_1	A_1B_2	A_2B_2	A_3B_2
A_1	3.36*	3.06*	3.48*	4.34*	2.27*	3.58*	2.70	0.74	0.60	2.03*	0.43	1.23
A_2	1.99*	1.69*	2.11*	2.97*	0.90*	2.21*	2.57	0.61	0.47	2.16*	0.30	1.10
A_3	3.00*	2.70*	3.12*	3.98*	1.91*	3.22*	3.37	1.41	1.27	1.36*	1.10	1.90
B_1	1.56*	1.86*	1.44	0.58	2.65*	1.34*	1.46*	3.42*	3.56*	6.19*	3.73*	2.93*
B_2	1.12*	1.42*	1.00*	0.14	2.21*	0.90*	2.45*	4.41*	4.55*	7.18*	4.72*	3.92*

* $P < 0.05$.

under each clustering methodology. Therefore, it is difficult to generalize the clustering patterns obtained from the analysis of qualitative and quantitative data separately. Moreover, dropping or transforming of any type of data that is generated by spending time and money may lead to loss of information. So, combining qualitative and quantitative information by distance indices is a beneficial way to handle such data. In the present study, six different combined distance measures were proposed and evaluated. While developing combined measures, care has been taken to combine the distance matrices corresponding to both qualitative and quantitative data. Prior to combining the qualitative and quantitative distance matrices, the elements of each of these matrices are set in a uniform scale, i.e. ranging between 0 and 1. Occasionally, the core set is identified based on the degree of correspondence between the clustering patterns obtained from the analysis of qualitative and quantitative data separately. In addition, the classification depends on the clustering methodology used in grouping the data. Therefore, it is also important to combine qualitative and quantitative data in the early stage of the analysis to draw valid inferences. Moreover, many clustering algorithms are used over time to generate a core set by applying a suitable allocation strategy. Odong *et al.* (2011) advocated the use of traditional clustering approaches over model-based clustering approaches to develop a core set, particularly for simple sequence repeat marker data. However, developing a core set based on phenotypic and SNP genotyping data, together by adopting a suitable procedure involving an appropriate combined distance measure, clustering methodologies, number of clusters, allocation method and evaluation strategy, is rarely known. Hence, the present study was undertaken to find an end-to-end solution for the identification of a core set.

In the present study, the consensus and merged-consensus clustering results were used to identify the optimum number of groups. Although there was a disagreement between the optimum number of clusters based on the quantitative and qualitative data (i.e. three and four), the number of clusters for mixture data was found to be 3. With regard to the selection of the best-fitted clustering algorithm, the *k*-means clustering algorithm gave the highest average cluster robustness values for the combined distance measures A_1B_1, A_3B_1 and A_1B_2. In contrast, the complete linkage clustering algorithm gave the highest average cluster robustness value for the combined distance measures A_2B_1, A_2B_2 and A_3B_2 (Table 1).

Odong *et al.* (2013) reviewed different criteria, under different circumstances, for the evaluation of a core set. Frequently, it may be difficult to comment on the diversity of a core collection based on the individual index. To avoid such confusion, a combined measure may help in drawing valid conclusions. Thus, a CEI involving MD, VD, VR, CR and APICD is used to evaluate the diversity in a core set. A comparison among the CEI values under all the combined distance measures indicates the superiority of the proportional allocation method over the log-proportional and RASE methods. Classifying germplasm to the lowest level, i.e. by taking the number of clusters equal to the size of the core set, and followed by the application of the RASE method for selecting accessions does not provide any gain over the proportional allocation and log-proportional allocation methods for all the combined distance measures, barring few exceptions (Table 2). Moreover, clustering germplasm with the number of clusters equal to the size of the core set leads to a violation of natural grouping in the clustering methodology. In addition, this will lead to bias in the selection of germplasm, as the probability of a germplasm being chosen from a smaller cluster is higher than that from a larger cluster. Hence, the identification of the optimum number of clusters based on sound statistical techniques followed by the selection of accessions based on the proportional allocation method is advisable for developing a diverse core set. Furthermore, Table 3 reveals that even though, in a majority of the cases, the combined distance measures performed better over the individual measures, under the proportional allocation method, there were few cases such as A_1B_2 versus B_2, A_2B_2 versus A_2 and A_3B_2 versus B_2 where the combined distance measures did not perform over the individual distance measures. Similarly, under the log-proportional allocation method, the combined

https://doi.org/10.1017/S1479262114000732 Published online by Cambridge University Press

distance measures A_2B_1 and A_2B_2 did not outperform the individual measure A_2. Hence, it cannot be concluded that combined distance measures will always perform better than individual distance measures. However, in the present study, the combined distance measure A_1B_2 performed best among the rest of the combined measures. Furthermore, the efficiency of the approach is established by a comparison with PowerCore (RDA-Genebank Information Center; http://www.genebank.go.kr/eng/PowerCore/PowerCore_Software.zip). This indicates the advantage of using the proposed approach for mixed data. Hence, the combined measure A_1B_2 using the *k*-means clustering algorithm along with the proportional allocation method to sample accessions can be preferred for the identification of a core set from a collection of rice germplasm.

Acknowledgements

The authors wish to express their gratitude to referees and editor for important comments and suggestions, which improved the paper substantially. Mr Sarkar acknowledge the receipt of fellowship from PG School, IARI, New Delhi during his Ph.D. study. Also, the authors wish to acknowledge World Bank Funded – National Agricultural Innovation Project (NAIP), ICAR Grants NAIP/Comp-4/C4/C-30033/2008-09.

References

Agrama HA, Yan WG, Lee F, Fjellstrom R, Chen M-H, Jia M and McClung A (2009) Genetic assessment of a mini-core subset developed from the USDA rice genebank. *Crop Science* 49: 1336–1346.

Crossa J and Franco J (2004) Statistical methods for classifying genotypes. *Euphytica* 137: 19–37.

Doring C, Borgelt C and Kruse R (2004) Fuzzy clustering of quantitative and qualitative data. In *Proceedings of the 2004 NAFIPS*. Banff, Alberta, Canada, pp. 84–89.

Everitt BS (1979) Unresolved problems in cluster analysis. *Biometrics* 35: 169–181.

Frankel OH and Brown AHD (1984) Plant genetic resources today: a critical appraisal. In: Holden JHW and Williams JT (eds) *Crop Genetic Resources: Conservation and Evaluation*. London: George Allen & Unwin Ltd, pp. 249–257.

Gangopadhyay KK, Mahajan RK, Kumar G, Yadav SK, Meena BL, Pandey C, Bisht IS, Mishra SK, Sivaraj N, Gambhir R, Sharma SK and Dhillon BS (2010) Development of a core set in brinjal (*Solanum melongena* L.). *Crop Science* 50: 755–762.

Gibert K and Cortes U (1997) Weighting quantitative and qualitative variables in clustering methods. *Mathware & Soft Computing* 4: 251–266.

Gower JC (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27: 857–874.

Hu J, Zhu J and Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theoretical and Applied Genetics* 101: 264–268.

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG and Park YJ (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23: 515–526.

Monti S, Tamayo P, Mesirov J and Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 52: 91–118.

Munneke B, Schlauch KA, Simonsen KL, Beavis WD and Doerge RW (2005) Adding confidence to gene expression clustering. *Genetics* 170: 2003–2011.

Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL and van Eeuwijk FA (2011) Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? *Theoretical and Applied Genetics* 123: 195–205.

Odong TL, Jansen J, van Eeuwijk FA and van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theoretical and Applied Genetics* 126: 289–305.

Sarkar RK, Rao AR, Wahi SD and Bhat KV (2011) A comparative performance of clustering procedures for mixture of qualitative and quantitative data – an application to black gram. *Plant Genetic Resources: Characterisation and Utilization* 9: 523–527.

Sharma R, Rao VP, Upadhyaya HD, Reddy VG and Thakur RP (2010) Resistance to grain mold and downy mildew in a mini-core collection of sorghum germplasm. *Plant Disease* 94: 439–444.

Simpson TI (2010) clusterCons: Calculate the consensus clustering result from re-sampled clustering experiments with the option of using multiple algorithms and parameter, R package version 3.0.2. http://cran.r-project.org/src/contrib/Archive/clusterCons/

Simpson TI, Armstrong JD and Jarman AP (2010) Merged consensus clustering to assess and improve class discovery with microarray data. *BMC Bioinformatics* 11: 590.

Studnicki M and Debski K (2012) ccChooser: Developing a core collections, R package version 3.0.2. http://cran.r-project.org/package=ccChooser

van Hintum T and Th JL (1999) The Core Selector, a system to generate representative selections of germplasm accessions. *Plant Genetic Resources Newsletter* 118: 64–67.

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

Wen W, Franco J, Chavez-Tovar VH, Yan J and Taba S (2012) Genetic characterization of a core set of a tropical maize race Tuxpeño for further use in maize improvement. *PLoS ONE* 7: e32626.

Yan W, Rutger JN, Bryant RJ, Bockelman HE, Fjellstrom RG, Thomas MC, Tai H and McClung AM (2007) Development and evaluation of a core subset of the USDA rice germplasm collection. *Crop Science* 47: 869–876.

Yu JZ, Kohel RJ, Fang DD, Cho J, Van Deynze A, Ulloa M, Hoffman SM, Pepper AE, Stelly DM, Jenkins JN, Saha S, Kumpatla SP, Shah MR, Hugie WV and Percy RG (2012) A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. *Genes Genomes Genetics* 2: 43–58.