Indian Journal of Agricultural Sciences **85** (12): 1560–4, December 2015/Article

Identification of a suitable clustering method and allocation strategy for core set development in salt stress tolerant rice (*Oryza sativa*) germplasm

SOUMYA RANJAN BARDHAN[1], A R RAO[2], PRABINA KUMAR MEHER[3], SUDEEP MARWAHA[4] and S D WAHI[5]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 20 January 2015; Accepted: 10 August 2015

ABSTRACT

Preserving genetic diversity in repository of germplasm is essential for crop breeding programs. However, maintenance and protection of all the germplasms in gene bank is difficult due to its voluminous size. Hence the development of core set with minimum number of germplasm representing maximum genetic diversity of the population has become an alternative. From the available clustering methods and allocation strategies, identifying a suitable combination is essential for the development of species-specific core set. In the present study, data on 219 salt stress tolerant rice (*Oryza sativa* L.) germplasm accessions with 14 phenotypic traits and 2915 genome wide Single Nucleotide Polymorphisms (SNPs) is considered to identify a suitable combination of clustering method and allocation strategy for core set development. Eight different combinations consisting of two clustering methods, viz. Ward's and UPGMA along with four different allocation strategies, viz. L, D, LD and NY allocation with three level of sampling intensities (20%, 25% and 30%) have been tried. Based on the study carried out during 2013-14 at Indian Agricultural Statistics Research Institute, New Delhi, it is concluded that the Ward's clustering method with NY allocation, irrespective of sampling intensity, is suitable for developing core set with maximum diversity.

Key words: Core set, Genetic diversity, Rice germplasm, Salinity stress, SNP genotyping

Rapid climate change has resulted in various abiotic stresses such as high and low temperature, excess and deficient water, salinity, etc., which are responsible for loss in crop production and productivity (Bansal *et al.* 2014). The strategy of integrating molecular breeding and genetic engineering through utilization of plant genetic resources is gaining momentum for developing climate-resilient cultivars (Varshney *et al.* 2011). Characterization and conservation of diversified plant genetic resources are the prerequisite for generation of genomics resources, which can be used by the breeders to develop stress tolerant cultivars. However, the presence of high degree of redundancy in different *in situ* as well as *ex situ* collections is a major problem for the management of plant genetic resources (Bansal *et al.* 2014). Moreover, the sequencing of each line in a large germplasm collection is almost impossible. Therefore, there is a need to develop a core set (Frankel and Brown 1984), which represents the entire crop genetic diversity (Glaszmann *et al.* 2010). This can be used as a valuable resource for future reanalysis and integrative genomic analysis (Hawkins *et al.* 2010).

[1]Senior Research Fellow (e mail: bardhan143@gmail.com), [2]Principal Scientist (e mail: arrao@iasri.res.in), [3]Scientist (e mail: pkmeher@iasri.res.in), [4]Senior Scientist (e mail: sudeep@iasri.res.in), [5]Principal Scientist (e mail: sdwahi@iasri.res.in)

Several gene banks possess a large number of salinity stress tolerance rice germplasms. Several approaches have been proposed for the development of a core set by using either quantitative (phenotypic traits) or qualitative trait (SNP markers) data or mixture of both (Sarkar *et al.* 2011, 2012 and 2014). 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), which are very large in number. However, all the SNPs may not be associated with each trait under consideration and by including all the SNPs in the model may increase complexity of analysis. Further, several clustering and sampling strategies have been suggested for the core set identification. However, it is required to choose an appropriate combination of clustering and sampling methods to develop a well diversified core set and hence, the present study is taken up to develop a core set with maximum genetic diversity by using relevant and reduced number of SNPs.

MATERIALS AND METHODS

The quantitative and qualitative trait data on 219 lines

of salt stress tolerant rice germplasm was considered for the present study (Sarkar *et al.* 2014). The dataset contains observations on 14 phenotypic characters, viz. days to 50% flowering, number of tillers per plant, plant height (cm), 100 seed weight (gm), number of panicle per tiller, panicles length (cm), number of filled grains, number of unfilled grains, grain length (cm), grain weight (g), biomass/plant (g), yield per plant (g), Na content of leaf per dry weight (mmol/g) and K content of leaf per dry weight (mmol/g). In addition to this, the dataset also contains 2915 SNP genotyping information in the form of 0, 1 and 2, which indicate the frequency of major allele present in each genotype for a given SNP locus.

Initially, SNPs with high variable importance and associated with each phenotypic trait were identified using Random Forest (RF; Breiman 2001) and Least Absolute Shrinkage Selection Operator (LASSO; Tibshirani 1996) methods. Under RF model, the SNPs having high variable importance, i.e. higher value of mean decrease in accuracy were identified as important SNPs for each trait. Similarly under LASSO model, the SNPs having non-zero regression coefficients were identified. Based on both LASSO and RF unique SNPs were identified from the selected SNPs of all 14 traits. The information on these selected SNPs of all the germplasm along with the phenotypic traits were then used as mixture data for core set development. For implementing RF and LASSO, "randomForest" and "glmnet" Packages of R-software were used respectively.

The identification of core set involves two-step procedure in which the accessions were first classified into homogeneous clusters and then a fraction of accessions from each cluster were selected by using an appropriate allocation strategy. Two most widely used clustering methods, i.e. Unweighted Pair Group Method with Arithmetic Mean (UPGMA, Sokal and Michener 1958, Murtagh 1984, D'haeseleer 2005) and Ward's clustering method (Ward 1963) were used. Four different allocation strategies, viz. L-allocation (Brown 1989), D-allocation, NY-allocation (Neyman 1934) and LD-allocation were used to draw samples from different clusters identified by the two clustering methods.

The diversity of core set was compared with the diversity of population (original) data set by computing the percentage mean difference (MD %), variance difference (VD %), coincidence rate (CR %) and variable rate (VR %) using quantitative variables (Hu *et al*. 2000). The different measures are as follows:

$$MD(\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{|Me - Mc|}{Mc} \times 100;$$

$$VD(\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{|Ve - Vc|}{Vc} \times 100;$$

$$CR(\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{Rc}{Re} \times 100;$$

$$VR(\%) = \frac{1}{m}\sum_{j=1}^{m}\frac{CVc}{CV\,e} \times 100;\ (j=1,2,\ldots,m \text{ traits})$$

where Me, Ve, Re and CVe and Mc, Vc, Rc and CVc stands for mean, variance, range and coefficient of variation for the entire dataset (population) and core set respectively. Nei (1978) genetic diversity for marker data was used to measure genetic diversity in both core and population germplasm.

RESULTS AND DISCUSSION

Parameter optimization

Several combinations of *ntree* (*i.e.,* 500, 1000, 1500 and 2000) and *mtry* (*i.e.* 54, 583, 875, 1458 and 2915) are opted to fit the RF model for idenfying important SNPs associated with each phenotypic trait. An optimum combination of parameters in RF for each trait is chosen on the basis of lowest Out Of Bag (OOB) error rate presented in Table 1, whereas no such parameter optimization is required for fitting LASSO model. It can be seen that in most of the traits, OOB error rate is stable around 1500 trees with *mtry* value as 1458. Also, it is observed that the number of trees required to obtain lowest OOB error rate in case of grain weight and grain length is lower compared to other traits. Further, it can be noticed that in case of Na-content, the number of trees required is largest and the OOB error is lowest as compared to the other traits.

SNPs selection

All the genome wide SNPs may not have association with the traits considered under study. Hence, only the important SNPs associated with the traits are filtered out by using RF and LASSO. Under RF, for each trait, the SNPs are ranked based on decreased order of variable importance and based on which top 10% (=290) SNPs are selected. This resulted in a total of 4060 (14*290) markers,

Table 1 Optimum value of *mtry* and *ntree* along with lowest Out Of Bag (OOB) error rate for all the 14 quatitative traits under RF methodology.

Trait name	mtry	ntree	OOB error
50% flowering	1458	1500	47.95
Plant height	1458	1500	55.25
Number of tiller per plant	1458	1500	55.25
Number of panicle per tiller	1458	1500	49.32
Panicle length	1458	1500	44.75
Number of filled grains	1458	1500	47.95
Number of unfilled grains	1458	1500	45.21
Grain length	1458	1000	52.05
Grain weight	54	1000	44.75
100 seed weight	1458	1500	47.95
Biomass per plant	1458	1500	51.14
Yield per plant	1458	1500	48.4
Na- content of leaf per dry weight	1458	2000	22.37
K- content of leaf per dry weight	1458	1500	43.84

with redundant SNPs, to be associated with all the traits. Similarly under LASSO, for each trait, SNPs with non-zero regression coefficients are selected and a total of 2382 SNPs (with redundancy) are selected for all the 14 traits. Finally, the SNPs selected under RF and LASSO method are pooled. From the selected SNP markers a total of 1362 unique markers are obtained, which constitutes the qualitative dataset and 14 phenotypic traits constitutes the quantitative dataset which in turn leads to mixture data. So the mixture data set consists of 1362 SNPs marker and 14 quantitative traits for 219 accessions of salt stress tolerant rice germplasm were subjected to further analysis. It is observed that in case of grain weight and grain length less number of trees in RF (*ntree*) is required (Table 1) than other phenotypic traits to get the lowest OOB error rate and this may be due the fact that the decision trees grown in the random forest are highly correlated. Further, it is seen that OOB error rate is minimum in case of trait "Na content" and is presumably that the SNPs are highly associated with this phenotypic trait.

Cluster analysis, sampling intensities and allocation strategy

Kumar *et al*. (2015) reported the formation of three major groups in the data set considered in the present study. However, a maximum of six groups can be safely formed in the considered data set. Hence the cluster analysis was carried out with six clusters using UPGMA and Ward's clustering methods. Six clusters having 56, 54, 37, 25, 28, 19 germplasm accessions respectively are obtained under Ward's clustering method, whereas the clusters sizes are 193, 12, 9, 2, 2 and 1 under UPGMA method (Table 2). From the identified clusters under both Ward's and UPGMA methods, sampling intensities of 20%, 25% and 30% are applied to draw samples under each allocation methods for the development of core sets. The proportion of observations drawn from each cluster under different allocation strategies and sampling intensities are given in Table 2. It is observed that the number of observations drawn from each clusters to develop core set under different allocation methods in Ward's clustering are lower than the respective cluster sizes, whereas the number of observations that are drawn under D, L and LD allocations in UPGMA method are higher (except NY) than the sizes of the clusters C4, C5 and C6 (Table 2). Further, it is observed that size of sample in case of NY allocation is less than or equal to the size of cluster, irrespective of the sampling intensities and clustering methods. This may be a limitation with L, D and LD allocations and presumably because of (i) smaller clusters size, (ii) large variation in cluster sizes, (iii) high mean Gower's distance (Gower 1971) in small cluster size, (iv) smoothing of the effect of cluster size by log transformation, (v) non-consideration of true cluster size as weight. Therefore NY allocation method may be used as a better alternative to the other considered allocation methods. Besides, from the cluster analysis it is also observed that sample size is much smaller than the cluster size irrespective of allocation methods and sampling intensities in Ward's clustering and therefore Ward's method may be used with some advantage as compared to the UPGMA for clustering.

Diversity analysis of core set

Values of MD%, VD%, VR% and CR% computed under different clustering methods, allocation methods and sampling intensities are reported in Table 3. Similarly, the diversity index meant for qualitative traits namely Nei's average expected heterozygosity is computed both for corset and population and given in Table 5. It can be seen that the value of MD% and VD% under Ward's method is less than the UPGMA method. On the other hand, UPGMA is better than Ward's clustering while compared using VR% and CR% measure. Further, it is observed that the representativeness of the core set for population is better under NY allocation than other allocations irrespective of the sampling intensities and clustering techniques used

Table 2 Number of samples to be drawn from each cluster, under Ward's and UPGMA clustering techniques, by different allocation methods and sampling intensities.

	Ward's Clustering						UPGMA Clustering						Sampling
Cluster name	C1	C2	C3	C4	C5	C6	C1	C2	C3	C4	C5	C6	intensities
Cluster size	56	54	37	25	28	19	193	12	9	2	2	1	
D	7	7	8	7	7	7	11	12	10	4	4	1	20%
	9	9	10	8	9	9	14	15	12	6	4	3	25%
	11	11	12	10	11	11	17	18	15	8	6	4	30%
L	8	8	8	7	7	6	20	9	8	3	3	2	20%
	10	10	9	8	9	8	25	12	10	5	3	0	25%
	13	12	11	10	10	9	30	14	13	5	4	0	30%
NY	11	11	8	5	6	4	39	3	2	1	0	0	20%
	14	14	10	6	7	5	48	3	2	1	0	0	25%
	17	16	12	7	9	6	58	4	2	1	0	0	30%
LD	8	8	8	5	7	6	20	10	10	3	3	0	20%
	11	11	10	8	9	8	25	13	13	4	4	0	25%
	13	13	12	9	11	9	30	16	16	5	4	0	30%

Table 3 Diversity indices for core sets under different allocation methods with different sampling intensities using both the clustering methods

AM	SI	UPGMA clustering				Ward's clustering			
		MD%	VD%	VR%	CR%	MD%	VD%	VR%	CR%
D	20%	8.433	34.281	120.165	89.071	3.850	25.336	93.594	68.389
	25%	8.734	31.195	118.094	88.194	4.708	22.420	96.217	77.243
	30%	7.933	33.061	121.236	89.493	3.700	23.850	104.238	81.318
L	20%	6.759	31.440	120.209	90.754	3.766	23.850	104.238	81.318
	25%	6.505	30.517	118.285	93.988	4.227	19.842	98.906	82.313
	30%	5.263	28.041	117.891	95.065	2.981	24.200	97.904	71.428
LD	20%	7.637	29.247	119.060	93.203	4.321	22.827	107.238	84.755
	25%	5.787	28.145	115.205	88.699	1.953	13.532	100.006	79.870
	30%	4.980	26.497	113.405	92.057	4.084	19.247	102.150	85.150
NY	20%	2.724	18.825	101.824	76.719	2.664	20.479	96.877	68.565
	25%	1.967	12.864	100.218	79.550	3.363	14.924	104.973	89.917
	30%	3.322	7.688	99.929	85.374	2.444	11.402	100.815	87.404

AM- Allocation methods; SI- Sampling intensities

Table 4 Nei's index for core sets and entire population under different allocation methods, sampling intensities and clustering methods

Sampling	Allocation methods											
	D			L			LD			NY		
	20%	25%	30%	20%	25%	30%	20%	25%	30%	20%	25%	30%
Ward's	0.22	0.22	0.21	0.20	0.21	0.22	0.22	0.21	0.21	0.22	0.22	0.22
UPGMA	0.21	0.22	0.22	0.22	0.21	0.22	0.22	0.22	0.22	0.22	0.22	0.21
Population	0.222											

while measuring diversity using MD% and VD% measures. Similarly, LD allocation is better than other allocations as far as VR% and CR% measures are concerned. Hence, it can be inferred that the representativeness of the core set in NY allocation along with Ward's method is higher than the other combinations of allocation and clustering techniques while compared using MD% and VD% measures. Similarly, the combination of LD allocation and UPGMA is better to represent the population diversity in the core using VR% and CR% as diversity measures. This argument is also supported by Nei's diversity index that the NY allocation along with Ward's method and LD allocation along with UPGMA are better in representing the population diversity in the core set (Table 4).

The main aim in developing core set is to preserve maximum diversity of the population with lesser number of germplasm as compared to population. In other words, more is the diversity in the core set better is the representativeness. In this study, a dataset on 14 phenotypic traits and 1362 selected SNPs corresponding to 219 of salt stress tolerant rice germplasm genotypes are used for developing a core set. Different combinations of 2 clustering techniques, 4 allocation methods with 3 levels of sampling intensities are used. After analyzing the results from cluster analysis, diversity measures using phenotypic traits and Nei's diversity index, it can be concluded that Ward's clustering with NY allocation is suitable for the development of diversified core set using both phenotypic and genome wide SNP genotypic data of salt stress tolerance rice germplasm. At the same time it will also be premature to say that the UPGMA clustering under LD allocation will always be inferior. Also, it is suggested to use RF and LASSO to identify genome wide SNPs associated with traits under consideration.

ACKNOWLEDGEMENT

First author acknowledges the receipt of Junior Research Fellowship from PG School, IARI during his M Sc programme.

REFERENCES

Bansal K C, Lenka S K and Mondal T K. 2014. Genomic resources for breeding crops with enhanced abiotic stress tolerance. *Plant Breeding* **133**: 1–11.

Breiman L. 2001. Random Forests. *Machine Learning* **45** (1): 5–32.

Brown A H D. 1989. Core collection: A practical approach to genetic resources management. *Genome* **31**: 818–24.

D'haeseleer P. 2005. How does gene expression clustering work? *National Biotechnology* **23**: 1 499–1 501.

Frankel O H and Brown A H D. 1984. Plant genetic resources today: a critical appraisal. (*In*) *Crop Genetic Resources: Conservation and Evaluation*, pp 249–57 Holden J H W and

Williams J T (Eds). George Allen and Unwin, London

Glaszmann J C, Kilian B, Upadhyaya H D and Varshney R K. 2010. Accessing genetic diversity for crop improvement. *Current Opinion in Plant Biology* **13**: 167–73.

Gower J C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* **27**: 857–74.

Hawkins R D, Hon G C and Ren B. 2010. Next-generation genomics: an integrative approach. *Nature Review Genetics* **11**: 476–86.

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. *Theoretical and Applied Genetics* **101**: 264–8.

Kumar V, Singh A, Mithra S V A, Krishnamurthy S L, Parida S K, Jain S, Tiwari K K, Kumar P, Rao A R, Sharma S K, Khurana J P, Singh N K and Mohapatra T. 2015 Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Research* (In Press).

Murtagh F. 1984. Complexities of Hierarchic Clustering Algorithms: the state of the art. *Computational Statistics Quarterly* **1**: 101–13.

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–90.

Neyman J. 1934. On the two different aspects on the representative method: The method of stratified sampling and the method of purposive selection. *Journal of Royal Statistical Society* **97**: 558–606.

Sarkar R K, Rao A R, Wahi S D and Bhat K V. 2011. A comparative performance of clustering procedures for mixture of qualitative and quantitative data- an application to black gram. *Plant Genetic Resources: Characterization and Utilization* **9**(4): 523–7.

Sarkar R K, Rao A R, Wahi S D and Bhat K V. 2012. Performance of clustering procedures for grouping germplasms based on mixture data with missing observations. *Indian Journal of Agricultural Sciences* **82**(12): 1 055–8.

Sarkar R K, Meher P K, Wahi S D, Mohapatra T and Rao A R. 2014. An approach to the development of a core set of germplasm using a mixture of qualitative and quantitative data. *Plant Genetic Resources: Characterization and Utilization, pp* 1–8, doi:10.1017/S1479262114000732.

Sharma R, Rao V P, Upadhyaya H D, Reddy V G and Thakur R P. 2010. Resistance to grain mold and downy mildew in a mini-core collection of sorghum germplasm. *Plant Disease* **94**(4): 439–44.

Sokal R and Michener C D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **28**: 1 409–38.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society* **58**: 267–88.

Varshney R K, Bansal K C, Aggarwal P K, Datta S K and Craufurd P Q. 2011. Agricultural biotechnology for crop improvement in a variable climate: hope or hype? *Trends in Plant Science* **16**: 363–71.

Ward J H. 1963. Hierarchical Grouping to optimize an objective function. *Journal of American Statistical Association* **58**(301): 236–44.

Wen W, Franco J, Chavez-Tovar V H, Yan J and Taba S. 2012. Genetic characterization of a core set of a tropical maize race Tuxpeno for further use in maize improvement. *PLoS ONE* **7**(3): e32626. doi:10.1371/journal.pone.0032626.

Yu J Z, Kohel R J, Fang D D, Cho J, Van Deynze A, Ulloa M, Hoffman S M, Pepper A E, Stelly D M, Jenkins J N, Saha S, Kumpatla S P, Shah M R, Hugie W V and Percy R G. 2012. A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. *Genes Genomes Genetics* **2**: 43–58.