Research article

Exploration of Phenotypic Dissimilarity for Drought Tolerance in Maize Inbred Line Collection

Pattama Hannok*, Phongsit Kaewunta, Walailak Khunyota, Anittaya Kaewnut, Phongsathorn Pachimkanthong and Chutiwat Tangthavonkarn

Division of Agronomy, Faculty of Agricultural Production, Maejo University, Chiang Mai, Thailand

Received: 21 October 2021, Revised: 7 March 2022, Accepted: 18 April 2022

DOI: 10.55003/cast.2022.01.23.001

Abstract

Cluster analysis is a type of exploratory analysis that is used for classifying unknown individuals into groups of members that share certain similarity. Grouping breeding materials into different clusters based on their performances under a given condition allows plant breeders to select breeding lines more efficiently. This study aimed to group breeding materials based on their performance under water stress conditions using cluster analysis. The experiment was conducted in RCB design with 3 blocks. Fifty maize inbred lines were grown and exposed to water stress conditions. Twenty-four phenotypic traits were collected and some of them were subjected to cluster analysis. The Partitioning Around Medoid algorithm was used to cluster 50 inbred lines from chosen phenotypes. Cluster validation was then carried out in a subsequent experiment by testing the statistically significant differences between chosen inbred lines and tolerant and susceptible clusters. According to the analysis, 4 clusters with different numbers of inbred lines were obtained. Lines in cluster no. 3 showed the most tolerance compared with the other clusters. In contrast, individuals in cluster no. 1 were the most susceptible. The results of cluster validation in another year also supported this cluster result. Therefore, we concluded that the clustering method was an efficient way to differentiate tolerant and susceptible inbred lines. Furthermore, the results suggested that plant height, tassel size, spikelet density, days to silking, and anthesis-silking interval were secondary traits that could be used in selection for drought tolerance in maize.

Keywords

cluster analysis;

partitioning around medoid;

ordination analysis;

inbred line collection;

drought tolerance

*Corresponding author: Tel.: (+66) 53873630 Fax: (+66) 53498168
E-mail: pattama_h@mju.ac.th

1. Introduction

Developing drought tolerant maize varieties is one method to mitigate yield losses caused by water stress-inducing phenomena such as rain delay and unexpected low amount of rainfall during the crop seasons. One important plant breeding process is the identification and grouping of breeding materials into different groups based on either similarity of performance or genetics. This assists plant breeders to efficiently choose breeding lines. However, grouping breeding materials using molecular markers may not provide information about plant adaptivity, especially in the case of breeding drought tolerant plants since drought tolerance is an adaptive trait that involves response to specific environment conditions and magnitude of abiotic stress. It is worthy of note that plant developmental and growth stage, percentage of field capacity and stress duration are the major factors that cause different degrees of damage in maize plants [1, 2].

Cluster analysis is one type of exploratory analysis and is based on a dimensionality reduction technique that reduces the multidimensional space of a dataset into two-dimensional (2D) spaces. It is used for classifying unknown individuals into groups based on their similarities. Cluster analysis usually uses unsupervised machine learning algorithms. This means that no historical record or background information is required for this method. The role of cluster analysis is to split set of unknown individuals into numbers of group (*k*). One of the popular algorithms for clustering a biological dataset is Partitioning Around Medoid (PAM) [3]. The advantage of PAM over other clustering algorithms is that PAM algorithm is appropriate for any dataset with high correlation such as biological dataset and no minimum number of individuals is required, except *k* has to be lower than the numbers of individuals [3]. Machine learning with PAM algorithm finds the medoids within each *k* cluster instead of centroids, which are commonly used in other algorithms. Medoid is less sensitive to outliers that may naturally exist in a biological dataset. A method is a central point of each *k* group [4]. With PAM algorithm, dissimilarity distance among individuals can be estimated from either Pearson, Spearman or Kendall correlation coefficients and this allows the space between individuals to be specified. Individuals with high similarity can be grouped into the same cluster whereas individuals from different clusters show high values of dissimilarity distance. As described, it seems that this method can be helpful for plant breeders who need to investigate whether variability is existing in their collections. Cluster analysis was used for the purpose of plant accession classification [5-7] but became less popular after the beginning of genomics era. However, it has been rising in popularity again with advances in computational science and bioinformatics. In the present study, in order to gain the hidden information from collected phenotypic traits that plants explicitly possess under the environment of interest, this study used phenotypic traits which were typically recorded under water stress in order to group our breeding materials into clusters based on their adaptiveness to water stress.

2. Materials and Methods

2.1 Experimental design

Fifty maize inbred lines obtained from National Corn and Sorghum Research Center, and Nakhon Sawan Field Crop Research Center, Thailand were investigated under water stress conditions. The experiment was conducted in a controllable environment using a randomized complete block design with 3 blocks. Each block was carried out on different planting dates during 2019-2020. Fifty maize inbred lines were planted in 2:1 ratio of soil and organic matter as a planting material, and they were contained in 450 planting bags (50 lines*3 blocks*3 plants/plot). All maize plants were grown following standard practices until harvesting.

In order to investigate the performance of each maize inbred line with some of their yield components, only a moderate degree of water stress was determined (growth stage before flowering time, low level of soil moisture and short period of water stress) for testing their performance of these 50 inbred lines. To do so, once maize plants were at the V_6 growth stage, water-withdrawal was started and continued until low soil moisture scale was observed. Next, a limited amount of water was given to the maize plants for 3 consecutive days. During this period, the majority of the maize plants were at the V_8 growth stage. Subsequently, watering was resumed and continued until harvesting.

2.2 Trait phenotyping and their correlation

Twenty-four phenotypic traits were collected after maize plants had experienced water stress during V_6-V_8 growth stage. Those 24 traits were plant height at V_T stage (PH), leaf rolling in the midst of water stress period (LR), tassel size (TS), spikelet density (SPD), 50% days to silking (SD), 50% days to anthesis (AD), anthesis and silking interval (ASI), normalized difference vegetation index (NDVI) at V_7, V_9, V_{10} growth stage (represented as V7, V9 and V10 respectively), leaf greenness (SPAD), chlorophyll content (CHL), percentage of leaf greenness at 30, 40 and 50 days after flowering (shown as p30, p40 and p50, respectively), number of top ears (nTop), number of total ears (nAll), total fresh weight of unhusked ears (FW) and husked ears (FWhusked), fresh weight of unhusked top ears (FWT) and husked top ears (FWThusked), kernel number of top ears (nKerT), kernel weight of top ears (KWt) and hundred-kernel weight of top ears (hunKWt).

Pearson correlation coefficients among all 24 traits were estimated via R version 3.6.3. In this study, the yield components, e.g., FW, FWhusked, FWT and FWThusked were considered as the primary traits. Therefore, only secondary traits that showed statistically significant relationships with these 4 primary traits were chosen and subjected to cluster analysis.

2.3 Cluster analysis and validation

The Partitioning Around Medoid (PAM) method was implementing for clustering all 50 inbred lines from the chosen phenotypes. Data standardization was first carried out to scale those chosen phenotypic traits. Subsequently, a correlation-based dissimilarity matrix was built from package 'factoextra' [8] in the R statistical program, version 3.6.3 [9] and *k* clusters were determined using 30 indices, which was run in a package 'NbClust' [10]. After that, all 3 components, e.g., the dissimilarity matrix, *k* cluster and chosen phenotypic traits in a scaled unit were used to cluster 50 inbred lines with package 'ggplot2' [11] and 'cluster' [12]. Moreover, each cluster result was checked for quality by estimating Average Silhouette width. In addition, the least square means for the chosen phenotypic traits of each resulting cluster were statistically compared and observed for intra-variation within each cluster.

Then, only eight inbred lines from tolerant and susceptible clusters were randomly chosen for validating clusters under water stress in the year 2021. Those lines were Ki58, Nei462013, Nei492006, Nei492024, Nei452006, Ki59, Kei1509 and Kei1421, in which the first 4 inbred lines were from tolerant clusters whereas the latter 4 lines were from susceptible ones. These 8 inbred lines were grown and treated by the same procedure as described above. At the harvest, only 4 yield components were collected and analyzed in test of a significance and *post-hoc* analysis.

3. Results and Discussion

3.1 Phenotypic correlation

Pearson correlation coefficients among 24 phenotypic traits were computed and tested for a significance of 0.05 and the results are shown in Table 1. The result revealed that PH, V9 and V10 showed significant positive correlation (r = 0.29 to 0.59, P<0.05) with the 4 main traits of yield components whereas SD and ASI had a negative relationship with those traits (r = -0.30 to -0.46, P<0.05). Only traits with significant correlation in both positive and negative coefficients were further used to develop a correlation-based dissimilarity matrix for cluster analysis. Therefore, 9 out of 24 phenotypic traits, e.g., plant height at V_T stage (PH), 50% days to silking (SD), anthesis and silking interval (ASI), normalized difference vegetation index at V_9 (V9), normalized difference vegetation index at V_{10} (V10), total fresh weight of unhusked ears (FW), total fresh weight of husked ears (FWhusked), fresh weight of unhusked top ears (FWT) and fresh weight of husked top ears (FWThusked) were chosen for that purpose.

3.2 Cluster analysis

All chosen 9 traits as described above were scaled before building a dissimilarity matrix among 50 maize inbred lines and illustrated via a heat map (Figure 1). As seen in Figure 1, different shades of colors between pairs of inbred line represented different degrees of dissimilarity in which maximum degree was 8 (dark blue) and the minimum was 0 (red). Thirty indices that were available in function ‘nbclust()’ in package ‘factoextra’ were used for *k* determination with the majority rule. It was found that 16 out of 30 indices suggested *k* = 4 for the cluster analysis. D-index was one of the 16 indices that suggested *k* = 4 as illustrated in Figure 2. Many clustering algorithms have been reported for indicating the numbers of clusters. It was difficult to decide the best algorithm for each dataset. Once Package ‘Nbclust’ provided 30 indices for this purpose, and it helped us to decide the appropriate number. According to the majority rule, *k* = 4 was therefore chosen for the further steps. In general, all indices for *k* determination share the same goal which is to minimize intra-variation within a cluster and maximize inter-variation between clusters [13, 14].

After reducing the dimensions of this dataset into 2 dimensions that could be used to investigate the hidden groups among 50 inbred lines, two-dimensional scores plot were obtained from the PAM algorithm. Under water stress conditions, the total phenotypic variation among 50 inbred lines from 9 correlated traits explained in this 2D plot was accounted for 67.7% (Figure 3). According to the 2D plot, different numbers of inbred lines were automatically classified in each cluster based on their dissimilarities. Cluster no. 3 consisted of 17 inbred lines was followed by cluster no. 4 (16 lines) whereas only 8 inbred lines were grouped in cluster 1 (the smallest cluster) and the rest were in cluster no. 2. Apparently, cluster no. 3 lay further away from the plot origin. This suggested that its 16 inbred lines probably performed differently from the inbred lines in other clusters.

The main advantage of clustering inbred lines based on their phenotypic performance is that the process makes it possible to observe phenotypic plasticity [15] in a specific condition of environments, especially abiotic stresses. The effect of QTL-by-environment interaction can change the magnitude of genetic force and influence plasticity [16]. Therefore, using cluster analysis based on integrated multi-phenotypic traits in this study may reflect the true performance of plants better than using molecular markers.

In order to designate the comparative degree of drought tolerance for these 4 clusters, the least square (LS) mean of all 24 traits was used to statistically compare clusters at a significance

Table 1 List of maize inbred lines in each cluster with relative tolerance degree

Cluster	Comparative degree	Inbred lines
1	Susceptible	DTMA192 Ki53 Nei452008(TF1) Nei542010 Kei1421 Kei1509 Ki59 Nei452006
2	Intermediate	DTMA193 Nei412019 Nei452013 Nei452029 Nei502007 Nei532003 Nei532005 Nei582016 Nei9008
3	Tolerant	DTMA202 Kei1420 Kei1508 Kei1608 Kei1715 Ki11 Ki47 Ki48 Nei452004 Nei452015(TF3) Nei452026 Nei582046 Nei9202(T) Nei462013 Nei492006 Nei492024 Ki58
4	Intermediate	Kei1519 Kei1533 Kei1615 Kei1618 Kei1719 Ki3 Ki45 Ki60 Nei422006 Nei422007 Nei452009 Nei492022 Nei502002 Nei512008 Nei512013 Nei541017

Note: underlined inbred lines were chosen for cluster validation

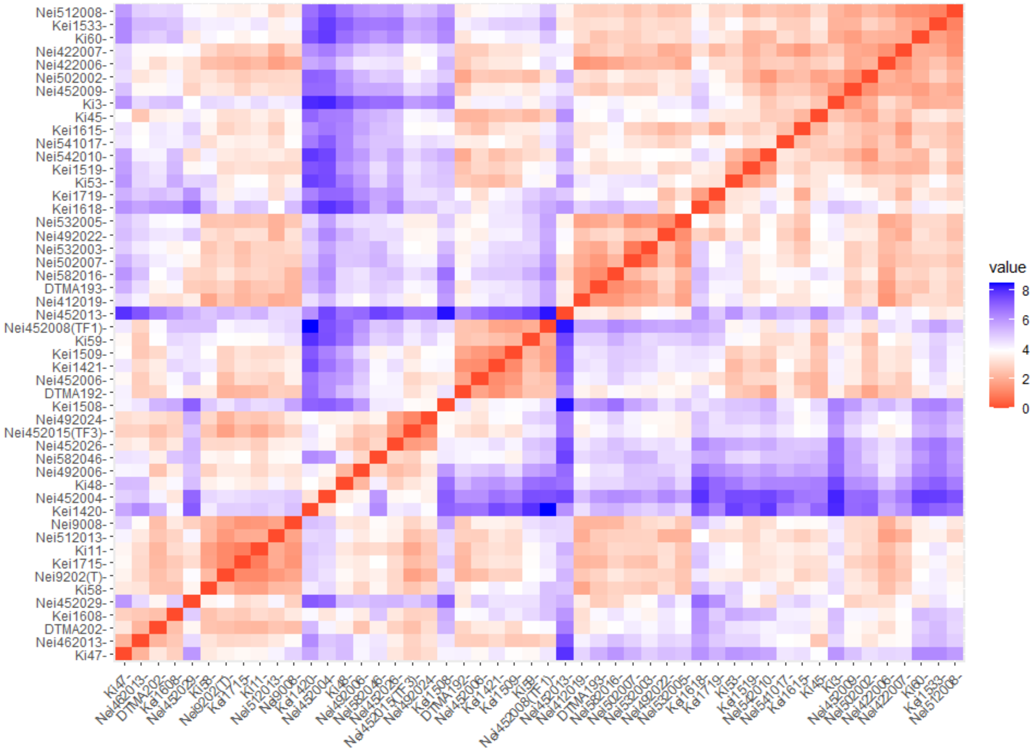

Figure 1. A heat map for MDS-obtained dissimilarity matrix among 50 inbred lines corresponding to 9 chosen traits, e.g., plant height at V_T stage, 50% days to silking, anthesis and silking interval, normalized difference vegetation index at V_9 and V_{10}, total fresh weight of unhusked ears, total fresh weight of husked ears, fresh weight of unhusked top ears and fresh weight of husked top ears

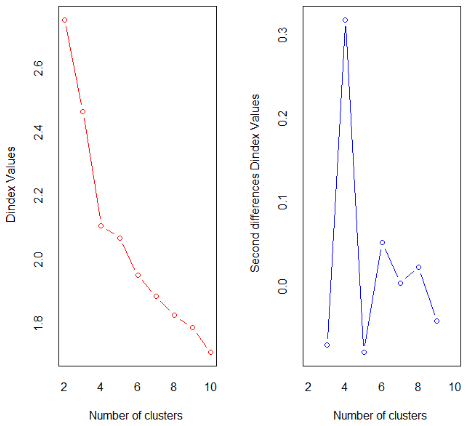

Figure 2. Knee plots obtained from D-index

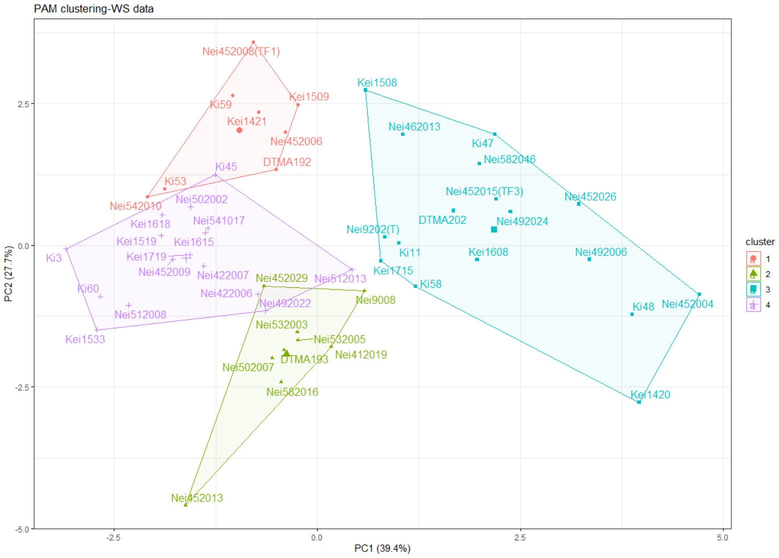

Figure 3. 2D score plots for 50 maize inbred lines under water stress conditions

level of 0.05. According to the LS mean comparison (data not shown), only 11 out of 24 traits, including primary and secondary traits were found to have statistically significant differences among the 4 clusters. Those traits were PH, TS, SPD, SD, ASI, FW, FWThusked, FWT, FWhusked, KWt and hunKWt. The first 5 phenotypic data were secondary traits and the latter ones were primary traits. Only 2 secondary traits, TS and ASI, were found in the index list of the International Maize and Wheat Improvement Center (CIMMYT) for drought maize breeding program [17]. Lower numbers of tassel size (TS) and anthesis-silking interval (ASI) were desirable for drought tolerance. Strong genetic correlation between ASI and grain yield was previously reported for maize [18]. Therefore, ASI is always used in the indirect selection for drought tolerance in maize. In addition to its role in phenotypic selection, ASI is one of the agronomic traits that is often included and used for predicting genomic estimating breeding value (GEBV) in the genomic selection approach [19-

21]. Furthermore, TS and SPD reflect the responses of male reproductive organs under drought stress. Recently, differential gene expression in young tassel samples was monitored and reported [22]. They found that genes regulating the metabolic pathways of carbohydrate and lipid metabolisms in tassel tissue were down-regulated under water stress and caused male infertility and immature tassels. For our study, we hereby reported that PH, TS, SPD, SD and ASI were effective secondary traits for use in selection for drought tolerance.

The comparative tolerance degree was designated as Tolerant, Intermediate and Susceptible for each cluster in this study (Table 1). Subsequently, the 8 inbred lines underlined in Table 1 of tolerant and susceptible comparative degree were randomly chosen and tested in year 2021. Similar work was done with maize roots, where PAM clustering was used with multiple phenotypes including aerenchyma content, cortical cells, cortical cell files, hydraulic conductance, metaxylem vessels and so on. The results showed 5 clusters of maize root phenotypes with contrasting performance under water stress. Some of those phenotypic traits have been suggested as breeding ideotypes [23].

3.3 Cluster validation

Four yield components, FW, FWThusked, KWt and hunKWt, were collected from 4 inbred lines of tolerant cluster (cluster no. 3) and susceptible cluster (cluster no. 1), each which had been grown under water stress conditions in another year as previous described. Figure 4 shows the strong significant differences between tolerant and susceptible clusters (P<0.01) for all pairs of phenotypic traits. This bar plot clearly indicates that the clustering method with $k = 4$ we used in this present study was an effective and helpful way to differentiate inbred lines based on their adaptive performances into different groups. This method enabled us to select the superior parental lines for further uses.

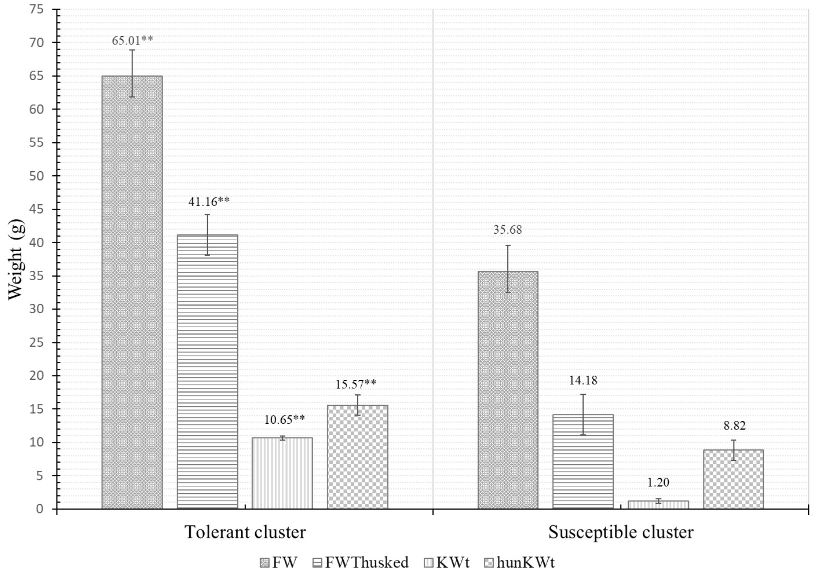

Figure 4. Mean comparison with standard error between tolerant and susceptible clusters for field weight (FW), husked top ear (FWThusked), kernel weight (KWt) and hundred-kernel weight (hunKWt). Means with asterisks show statistically significant difference at 1% level of each trait between 2 clusters.

4. Conclusions

In order to differentiate material collection from performance under the environment of interest, cluster analysis is another tool that can be used to explore the hidden patterns and enables the grouping of individual lines based on their dissimilarity. Grouping individuals using multivariate data and the subsequent display of their distances on a 2D score plot is quite helpful for breeding work when the history record is unknown. In this current study, we proved that the clustering method was effective to differentiate tolerant and susceptible inbred lines. Furthermore, we then suggested that plant height (PH), tassel size (TS), spikelet density (SPD), days to silking (SD), and anthesis-silking interval (ASI) were effective as a secondary trait for use in selection for drought tolerance. According to the information we obtained in this study, 17 potential inbred lines from tolerant clusters were listed and are considered as good candidate for future works. The work has enabled us to select superior inbred lines for developing biparental populations between tolerant and susceptible lines. In addition, the other inbred lines designated as the intermediate cluster should also be further investigated at the molecular level as some of them may prove to be of use.

5. Acknowledgements

We thank Dr. Sansern Jampathong (a former scientist of National Corn and Sorghum Research Center, Faculty of Agriculture, Kasetsart University) and Mr. Suriphat Thaitad (a senior scientist of Nakorn Sawan Field Crops Research Center, Field and Renewable Energy Crops, Department of Agriculture) for providing us plant materials for this research. This work was sponsored by Agricultural Research Development Agency (Public Organization), PRP6105022140, Thailand.

References

[1] Aslam, M., Maqbool, M.A. and Cengiz, R., 2015. *Drought Stress in Maize* (*Zea mays* L.)*: Effects, Resistance Mechanism, Global Achievement and Biological Strategies for Improvement.* New York: Springer.

[2] Sah, R.P., Chakraborty, M., Prasad, K., Pandit, M., Tudu, V.K., Chakravarty, M.K., Narayan, S.C., Rana, M. and Moharana, D., 2020. Impact of water deficit stress in maize: Phenology and yield components. *Scientific Reports,* 10, 2944, DOI:10.1038/s41598-020-59689-7.

[3] Kaufman, L. and Rousseeuw, P.J., 1990*. Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons.

[4] Van der Laan, M., Pollard, K. and Bryan, J., 2003. A new partitioning around medoids algorithm, *Journal of Statistical Computation and Simulation*, 73(8), 575-584.

[5] Broich, S.L. and Palmer, R.G., 1980. A cluster analysis of wild and domesticated soybean phenotypes. *Euphytica,* 29, 23-32.

[6] Peeters, J.P. and Martinelli, J.A., 1989. Hierarchical cluster analysis as a tool to manage variation in germplasm collections. *Theoretical and Applied Genetics,* 78, 42-48.

[7] Kouamé, C.N. and Quesenberry, K.H., 1993. Cluster analysis of a world collection of red clover germplasm. *Genetic Resoures and Crop Evolution,* 40, 39-47.

[8] Kassambara, A. and Mundt, F., 2020. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. *R Package Version 1.0.7*. [online] Available at: https://CRAN.R-project.org/package=factoextra.

[9] R Core Team, 2021. *R: A Language and Environment for Statistical Computing*. [online] Available at: https://www.R-project.org/.

[10] Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A., 2014. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6), 1-36.

[11] Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

[12] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K., 2021. *Cluster: Cluster Analysis Basics and Extensions*. *R Package Version 2.1.2*. [online] Available at: https://CRAN.R-project.org/package=cluster.

[13] Bezdek, J.C. and Pal, N.R., 1998. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3), 301-315.

[14] Ray, S., and Turi, R.H., 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques,* Calcutta, India, December 27-29, 1999, pp. 137-143.

[15] Schlichting, C.D., 1986. The evolution of phenotypic plasticity in plants. *Annual Review of Ecology and Systematics*, 17(1), 667-693.

[16] Marais, D.L.D., Hernandez, K.M. and Juenger, T.E., 2013. Genotype-by-environment interaction and plasticity: Exploring genomic responses of plants to the abiotic environment. *Annual Review of Ecology, Evolution, and Systematics*, 44(1), 5-29, .

[17] Bänziger, M., Edmeades, G.O., Beck, D. and Bellon, M., 2000. *Breeding for Drought and Nitrogen Stress Tolerance in Maize: From Theory to Practice*. Mexico: CIMMYT.

[18] Edmeades, G.O., Bolaños, J., Hernàndez, M. and Bello, S., 1993. Causes for silk delay in a lowland tropical maize population. *Crop Science*, 33, 1029-1035, DOI: 10.2135/cropsci1993.0011183X003300050031x.

[19] Shikha, M., Kanika, A., Rao, A.R., Mallikarjuna, M.G., Gupta, H.S. and Nepolean, T., 2017. Genomic selection for drought tolerance using Genome-Wide SNPs in maize. *Frontier in Plant Science*, 8(550), DOI:10.3389/fpls.2017.00550.

[20] Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E.A., Babu, R., Zhang, X. and Trachsel, S., 2018. Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Frontier in Plant Science,* 9(366), DOI:10.3389/fpls.2018.00366.

[21] Yuan, Y., Cairns, J.E., Babu, R., Gowda, M., Makumbi, D., Magorokosho, C., Zhang, A., Liu, Y., Wang, N., Hao, Z., San, Vicente F., Olsen, M.S., Prasanna, B.M., Lu, Y. and Zhang, X., 2019. Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in Maize. *Frontier in Plant Science,* 9(1919), DOI:10.3389/fpls.2018.01919.

[22] Wang, N., Li, L., Gao, W.W., Wu, Y.O., Yong, H.J., Weng, J.F., Li, M.S., Zhang, D.G., Hao, Z.F. and Li, X.H., 2018. Transcriptomes of early developing tassels under drought stress reveal differential expression of genes related to drought tolerance in maize. *Journal of Integrative Agriculture*, 17(6), 1276-1288, DOI: 10.1016/S2095-3119(17)61777-5.

[23] Klein, S.P., Schneider, H.M., Perkins, A.C., Brown, K.M. and Lynch, J.P., 2020. Multiple integrated root phenotypes are associated with improved drought tolerance. *Plant Physiology*, 183(3), 1011-1025, DOI: 10.1104/pp.20.00211.