

Original Research Article

Application of principal component analysis to access selected rice germplasm phenotypic diversity

ABSTRACT

Without variability, it is not possible to conduct a plant breeding program. Germplasm is hence the critical first step in initiating a breeding program. Diversity in *O. glaberrima* accessions is enormous. It needs to be organized and characterized in order to facilitate its use by plant breeders. Two principal components (PRIN1) and (PRIN2) accounted for most of the variability observed in characters studied/measured. PRIN 1 accounted for 56% of the phenotypic and morphological variation. The PRIN1 was loaded on plant height, number of panicle, biomass wet weight, panicle wet weight and grain yield. PRIN 2 accounted for 23% of the variation. PRIN 2 was loaded on biomass wet weight, biomass dry weight, panicle wet weight, panicle dry weight, and harvest index. The test for univariate statistics described individual variable to explore pattern of response to variation showed strong statistical significance ($P < 0.001$) on phenotypic differences in all the variables that were measured. These traits are the most important characters contributing to the overall variability. The dendrogram produced grouping that defined five distinct clusters and minimum genetic distance between clusters varies from 0 to 2.0. All the selected *Oryza glaberrima* accessions and the *Oryza sativa* were distributed across the five clusters respectively.

KEY WORDS: Principal Component Analysis, germplasm evaluation, genetic diversity, multivariate.

INTRODUCTION

A huge number of variables are regularly measured by plant breeders of which some may not be of sufficient discriminatory power for germplasm evaluation and characterization of traits (Shaibu and Ogburia, 2002). In such a situation, principal component analysis (PCA) may offer a useful insight in data set to eliminate redundancy. In any crop, germplasm resource not only serves as a valuable source of useful genes but also provides scope for building up a basic population of wide genetic variability. Bringing improvement over existing crop varieties is a continuous process in plant breeding. To achieve this objective, the breeder has to identify diverse parents having high genetic variability for combining desirable characters. Therefore, knowledge of sound genetic diversity is essential for undertaking any recombination breeding program. In earlier days, geographical diversity was considered as a measure of genetic diversity but recently it is observed that genetic materials from same eco geographic origin also possess diverse genetic makeup and it is not uncommon that the genetic materials of different eco geographic origin possess similar genetic architecture due to free and easy material transfer facilities. The usefulness of multivariate analysis for the study of morphologically complex individual and for measuring the degree of divergence

Comment [ar1]: Does not make any sense

Comment [ar2]: Traits and characters are synonyms

Comment [ar3]: Improve the sentence

Comment [ar4]: ?

Comment [ar5]: ?

between biological populations has been shown in different fields of research. Multivariate statistical techniques which simultaneously [analyse analyses](#) multiple measurements on each individual under investigation are widely used in an analysis of genetic diversity irrespective of whether it is morphological, biochemical or molecular marker-based and subsequently, classification of germplasm collections. Among the multivariate techniques, principal component analysis (PCA) and cluster analysis have been shown to be very useful in selecting genotypes for breeding program that meet the objective of a plant breeder (Mohammadi and Prasanna, 2003). PCA has been reported to reveal patterns and eliminate redundancy in data sets (Adams, 1995) while morphological and physiological variations routinely occur in crop species. Cluster analysis is commonly used to study genetic diversity and for forming principal subclass for grouping accessions with similar characteristic into one homogenous category. Clustering is also used to summarize information on relationships between objects by grouping similar units so that the relationship may be easily understood for scientific knowledge and advancement.

Maji and Shaibu (2012) described that multivariate analysis has been used frequently for genetic diversity analysis in many crops such as barley (*Hordeum vulgare* L.), [S](#)sorghum (*Sorghum bicolor* L. Moench), wheat (*Triticum* spp.), peanut (*Arachis hypogaea* L.) (Upadhyaya *et al.*, 2009), [and](#) vineyard peach (*Prunus persica* L. Batsch) and [in](#) rice (Gana *et al.*, 2013). [Multivariate analysis has been used frequently for genetic diversity analysis in many crops such as finger millet \(Dagnachew *et al.*, 2012b\) and rice \(Gana *et al.*, 2013\).](#) The

main objective of any plant scientist is to identify an optimum number of plant characters which are sufficient to elucidate the maximum variability in field crop. This study was undertaken to run a classification analysis on selected *Oryza glaberrima* accessions and Asian rice varieties by means of descriptive statistics and to understand the association of various characters, PCA and cluster analysis which would enable breeders to classify the available germplasm into distinct groups on the basis of their genetic diversity. However, the results obtained would be valuable in developing an effective breeding.

MATERIALS AND METHODS

Plant materials, data collection and experimental design

The study was conducted to characterize 30 rice germplasm accessions consisted of 22 selected *Oryza glaberrima* and 8 *Oryza sativa* lines obtained from Africa Rice gene bank were evaluated under lowland conditions at Africa Rice lowland breeding station, at the International Institute of Tropical Agriculture, Ibadan, Nigeria. IITA is located at (Latitude 7° 3 'N and Longitude 3° 45 'E in the forest-savannah agro ecology). These lines were

Comment [ar6]: Sentence repetition

Comment [ar7]: Make no sense

evaluated in replicated trials in a factorial lattice design during the dry season of 2014. The 21 day old seedlings were transplanted into a well levelled puddled field of 1.2m² plot at spacing of 20cm between rows and hills consisting two row plots of 3m length. Ten morphological characters were measured including days to flowering, plant height, number of tiller, number of panicle, biomass wet weight, biomass dry weight, panicle wet weight, panicle dry weight, harvest index and grain yield.

DATA ANALYSIS

Univariate statistics, standard deviation and probability level for each one of the 10 characters were calculated. Clustering of genotypes into similar groups was performed using Ward's hierarchical algorithm based on squared Euclidean distances. For the ten agro-morphological characters, the data were standardized to have a mean of zero and a variance of one prior to squared Euclidean distance calculation. The pseudo F statistic and the pseudo T² statistic (Jobson, 1992) were examined to establish the numbers of clusters using SAS version 9.2 software (SAS Institute, 1996, Cary, NC). In order to identify the patterns of morphological variation, (PCA) was conducted.

RESULTS AND DISCUSSIONS

Of the ten phenotypic characters measured, ten principal component axes were revealed from PCA analysis. Two principal components (PRIN1) and (PRIN2) accounted for most of the variability observed in characters measured (Table 1). PRIN1 accounted for 56% of the phenotypic and morphological variation. The PRIN was loaded on plant height, number of panicle, biomass wet weight, panicle wet weight and grain yield. PRIN2 accounted for 23% of the variation. PRIN 2 was loaded on biomass wet weight, biomass dry weight, panicle wet weight, panicle dry weight, and harvest index. The most important descriptions were those associated with grain yield based on the analysis of natural variability in phenotypic characters measured in rice (Table 1). The scree plot graphs showed the eigenvalue against the component number (Figure 2). The values in the two columns of the table 1 showed two principal component axes. From the third component on, it showed that the line is almost flat (Figure 1), meaning that each successive component is accounting for smaller and smaller amounts of the total variance. In general, we are interested in keeping only those principal components whose eigenvalues are greater than 1. Components with an eigenvalue of less than 1 account for less variance than the original variable (which had a variance of 1), and so are of little use. Hence, the point of principal components analysis is to redistribute the variance in the correlation matrix (using the method of eigenvalue decomposition) to

Comment [ar8]: Compare and discuss the results with the previous research papers

Comment [ar9]: Explain and discuss

Comment [ar10]: Compare and discuss the results with the previous research papers

Comment [ar11]: Add reference

redistribute the variance to first components extracted (Table 1). The test for univariate statistics described individual variable to explore pattern of response to variations showed strong statistical significantly ($P < 0.001$) on phenotypic differences in all the variables that were measured (Table 2). This indicates that the quantitative traits may be modified variously by the environmental conditions and each contributing such a small amount of phenotype that their individual effects cannot be detected by Mendelian methods but by only statistical methods. High variability observed for most of the characters indicated the possibility of improvement of these characters by direct selection. These traits should be given prominence in rice improvement. Days to flowering, number of tillers, biomass dry weight and harvest index negatively contributed to the first component axes (PRIN1), and only days to flowering negatively contributed to the total variation in these second principal component axes (PRIN2). The evaluation of germplasm frequently comprises measurement of phenotypic characters or agronomic traits of interests, such as resistance to pests and diseases and tolerance to physiological stresses that are influenced by environment which they grow. The data measuring different characters are most often required by plant breeders. According to Ng and Padulosi (1992) and (Peters and Williams, 1984), breeders continually working and looking for variation in germplasm for use in crop improvement. Therefore, continuous evaluation of germplasm should be done to broaden the genetic base of the species and identification of additional genes of interest or alternative sources that control a particular attribute or trait for use in crop improvement. As demonstrated in this study, genetic variability in crop species should be exploited further to identify traits of interest so as to develop new rice varieties with high stability to resist or tolerate adverse environments and biotic conditions demonstrated by Gana, (2006). Awopetu and Gana (1997) studied the genetic relationship of some rice varieties and observed that origin, habitat and breeding background contributed to variation in the rice population. Therefore, this study is conducted to explain variation among the selected *Oryza glaberrima* accessions *Oryza sativa* varieties and identify traits that contribute to variability in this population and for their possible exploitation in breeding programs. However, consistent effort is therefore required to evaluate rice germplasm collection and breeding rice adaptable to harsh African environment high grain yield.

Comment [ar12]: Statistically significant

Comment [ar13]: Compare and discuss the results with the previous research papers

Comment [ar14]: Revise the sentence

Comment [ar15]: No need

138 Table1:Eigenvector(“Weight”)andEigenvalue(“Load”)ofthecorrelationmatrixandtheir
139 contributiontototalvariationinrice.

Descriptor variables	PRIN 1	PRIN 2
Days to flowering (days)	-0.39	-0.01
Plant height (cm)	0.23	0.01
Number of tiller (#)	-0.41	0.04
Number of panicle (m ⁻²)	0.40	0.03
Biomass wet weight (gm ⁻²)	0.28	0.27
Biomass dry weight (gm ⁻²)	-0.28	0.43
Panicle wet weight (gm ⁻²)	0.18	0.53
Panicle dry weight (gm ⁻²)	0.01	0.60
harvest index (m ⁻²)	-0.35	0.30
Grain yield (gm ⁻²)	0.39	0.11
Eigen value	5.65	2.31
Proportion	0.56	0.23

140

141 Table2:Univariateteststatisticsofsomephenotypiccharactersidentifiedbyprincipal
142 componentanalysisclusters.

143

Variable	Total standard deviation	Pooled standard deviation	Between standard deviation	R-Square	F Value	Pr> F
Days toflowering(days)	347.8	124.3	370.4	0.99	1743.2	<.0001
Plantheight(cm)	14.1	11.9	8.5	0.32	10.1	<.0001
Number oftillers(#)	44.2	8.4	46.3	0.96	598.4	<.0001
Number ofpanicle(#)	121.9	28.9	126.4	0.94	373.0	<.0001
Biomass wetweight(gm ⁻²)	114.8	91.6	76.5	0.39	13.7	<.0001
Biomass dryweight(gm ⁻²)	188.8	139.0	139.6	0.48	19.8	<.0001
Panicle wetweight(gm ⁻²)	60.89	52.8	34.4	0.28	8.35	<.0001
Panicle dryweight(gm ⁻²)	57.3	54.4	22.9	0.14	3.5	<.0001
Harvestindex(%)	63.3	35.9	56.0	0.69	47.6	<.0001
Grain yield(tha ⁻¹)	195.3	79.2	191.1	0.84	114.1	<.0001

144

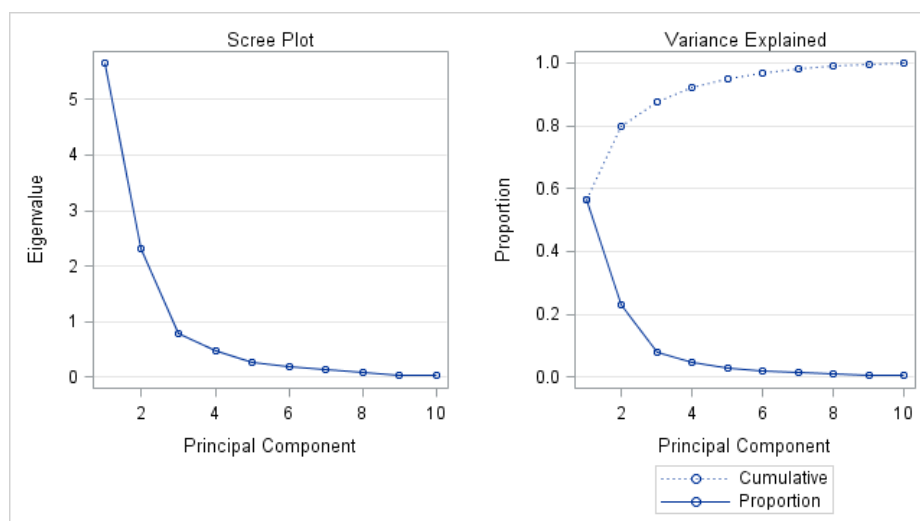


Figure 1 Scree plot and variance explained by the principal component analysis

Diversity analysis

Similarity indices and pattern of relationships among the *O. glaberrima* and *O. sativa* from clusters and principal component analysis are useful in evaluating the potential breeding value of both the checks and *O. glaberrima* lines. The checks-FARO52, TOX4004, NERICA-L25, NERICA-U4, IR77298-14.1-2-B-10, IR74371-1-1, CG 14 and Apower were distributed across three clusters. The dendrogram produced grouping that defined five distinct clusters and minimum genetic distance between clusters varies from 0 to 2.0. Cluster 1 composed of 14 *Oryza glaberrima* and one *O. sativa* check (NERICA L-25) diverged 0.1, 0.3, 0.8, 1 and 2.0. Cluster 2 grouping comprised of 3 *O. glaberrima* lines (TOG6520, TOG6311 and A2-123) and 3 *O. sativa* checks namely; TOX4004, IR74371-54-1-1, IR77298-14-1-2-B-10 including *O. glaberrima* check (CG 14) diverged at 0.5 and 1.0 level. RAM 133, TOMYANGOB and FARO52 were grouped into cluster 3 at 0.6 and 0.9 levels while cluster 4 comprised of 3 checks and one *O. glaberrima* line – TOG 5534 while cluster 5 comprised of TOG 5396 which stands out from other clusters. The standalone of one of the selected *O. glaberrima* accession may have some underlying unique features not found in other groups (Figure 2). This work has demonstrated the level of similarity between and within the selected lines. Genetic diversity association generated by unweighted pair-group method of mathematical averages (UPGMA) clustering of amplified fragment length polymorphism (AFLP) method have been reported in *Musa* species (Shaibu, 2012). However, this work on

phenotypic diversity and similarity will help create segregation in breeding population that may be useful in diversity study and crop improvement.

CONCLUSION

Variation in plant phenotypic traits among rice accessions is a common phenomenon. Without variability, it is not possible to conduct a plant breeding program. Germplasm is hence the critical first step in initiating a breeding program. Diversity in *O. glaberrima* accessions is enormous. It needs to be organized and characterized in order to facilitate its use by plant breeders. There is need to understand how this tremendous variation originates and the manner in which it is organized or classified. We therefore need to know the sources to which breeders may go to find *O. glaberrima* germplasm with enormous variation to initiate their breeding programs, and the rationale for selecting materials for breeding and improving rice adaptable to African environment.

Comment [ar16]: Write down the conclusion not the introduction

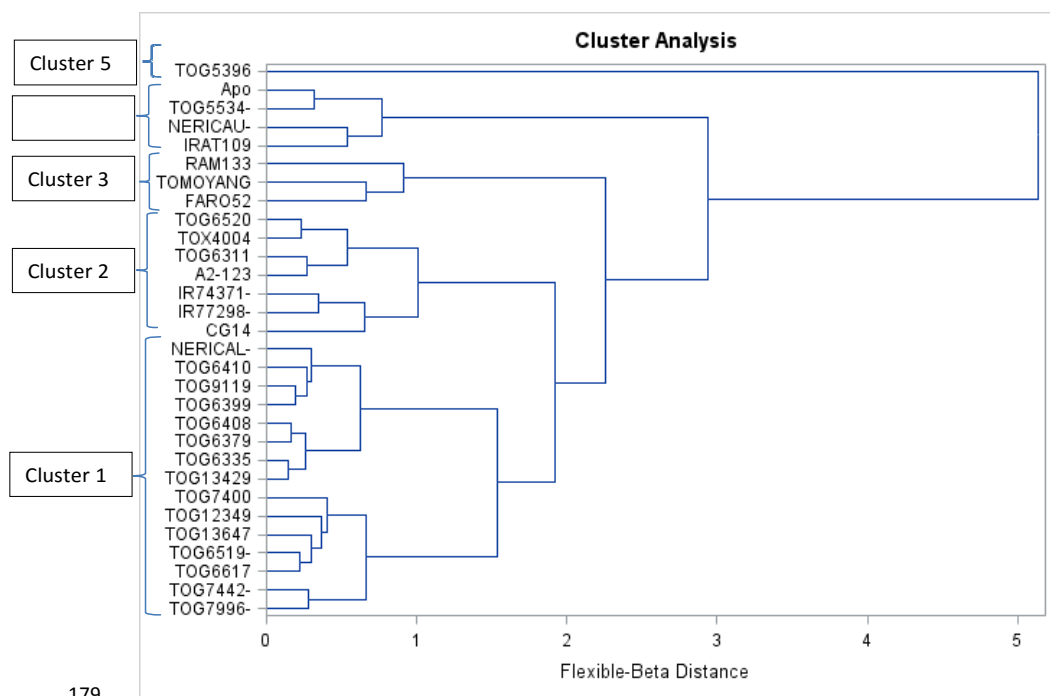


Figure 2. Clustering pattern and dendrogram of selected *O. glaberrima* and *O. sativa* lines.

REFERENCES

- 183 1. Shaibu, A. A. and Ogburn M. N. (2002). Structural organization and morphological
184 diversity of somaclonal variation in triploid plantain (Musa Sp. AAB group)
185 landraces. The Nigerian Journal of Research and Production vol. 1, 31-37
- 186 2. Adams MW (1995). An estimate of homogeneity in crop plants with special reference
187 to genetic vulnerability in dry season. *Phaeolus vulgaris. Euphytica*, 26: 665-679.
188
- 189 3. Mohammadi SA, Prasanna BM (2003). Analysis of genetic diversity in crop plants-
190 salient statistical tools and considerations. *Crop Sci.* 43, 1235-1248.
191
- 192 4. Maji, A. T. and Shaibu, A. A. (2012). Application of principal component analysis for
193 rice germplasm characterization and evaluation. *J. Plant Breeding and Crop Sci.* Vol.
194 4, 87-93.
195
- 196 5. Upadhyaya HD, Gowda CLL, Pundir RPS, Reddy VG, Singh S (2006). Development
197 of core subset of finger millet germplasm using geographical origin and data on 14
198 quantitative traits. *Gen. Res. Crop Evo.* 53: 679-685.
199
- 200 6. Gana AS, Shaba SZ, Tsado EK (2013). Principal component analysis of
201 morphological traits in thirty-nine accessions of rice (*Oryza sativa* L.) grown in a
202 rainfed lowland ecology of Nigeria. *J. Plant Breed. Crop Sci.* Vol. 5, 120-126.
203
- 204 7. Dagnachew Lule, Kassahun Tesfaye, Masresha Fetene and Santie De Villiers
205 (2012b). Multivariate Analysis for Quantitative Traits in Finger Millet (*Eleusine*
206 *coracana* subsp. *coracana*) Population Collected from Eastern and Southeastern
207 Africa: Detection for Patterns of Genetic Diversity. *Int. J. Agric. Res.* 7(6): 303-314.
208
- 209 8. Jobson JD (1992). Applied multivariate data analysis. Volume II: Categorical and
210 Multivariate Methods. New York: Springer Verlag, USA.
211
- 212 9. SAS Institute (1996), SAS/STAT software: Users guide, version 9.2. SAS Inst., Cary,
213 NC.
214
- 215 10. NgNQ, Padulosi S (1992). Constraints in the accessibility and use of germplasm
216 collection. In Biotechnology enhances research on tropical crops in Africa edited by
217 Thottappilly, L. M. Monti D. R. Mohan Raj. A. W. Moore IITA Ibadan.
218
- 219 11. Peters JP, Williams JT (1984). Towards better use of gene bank with special reference
220 to information. FAO/IBPGR Plant genetic resources Newsletter 60: 22-31.
221
- 222 12. Gana AS (2006). Variability studies of the response of rice varieties to biotic and
223 abiotic stresses. Unpublished Ph.D Thesis, University of Ilorin.
224
- 225 13. Awopetu JA, Gana, AS (1997). A Numerical Analysis of Genetic relationship within
226 rice accessions. Nig. J. Gene X 11: 1-8.

Comment [ar17]: References are lacking