Multivariate characterization of mutant ginger genotypes using cluster and principal component analyses for yield and trait improvement

ABUA, M. N.[1], OBOK E.E[1], EDUGBO, R. E.[2*], USEN, J. M.[1] AND IWO, G. A.[1]
[1] Department of Crop Science, University of Calabar, PMB 1115 Calabar, Cross River State, Nigeria.
[2]Department of Applied Biology, Federal University of Technology, Babura, Jigawa State, Nigeria
e-mail: richmondedugbo@gmail.com

Abstract

Understanding phenotypic diversity is essential for effective ginger (*Zingiber officinale* Roscoe) improvement. This study evaluated 15 mutant ginger lines at the Cross River Basin Development Authority, Calabar, Nigeria during the 2023 cropping season using a Randomized Complete Block Design (RCBD) with three replications. Phenotypic variation was analysed using Principal Component Analysis (PCA) and hierarchical clustering. The first four principal components (PCs) explained 93.64% of total variability, with PC1 contributing 72.87%, PC2 13.37%, PC3 6.59%, and PC4 1.14%. Rhizome length showed the highest positive loading on PC1 (0.9912), while establishment count (−0.8189) and plant height (−0.4558) dominated PC2. The genotypes were grouped into four clusters, indicating substantial genetic diversity. Cluster 2 (six genotypes) exhibited higher establishment counts (up to 91%), more rhizome fingers (8–12 per plant), longer rhizomes (15.3–18.6 cm), and higher yields (up to 32.7 t ha^{-1}), making them ideal candidates for yield-focused breeding programmes. Cluster 3 (two genotypes) showed high emergence rates (>90%), taller plants (up to 102 cm), and broad leaves (5.2–6.1 cm), suggesting potential for improving vegetative vigour. The integration of PCA and clustering provides a robust framework for elite genotype selection and targeted trait improvement, offering a strategic pathway to develop high-yielding and regionally adaptable ginger cultivars.

Keywords: Ginger, Hierarchical clustering, Phenotypic diversity, Principal Component Analysis (PCA)

Introduction

Univariate analyses, while useful, are often inadequate for plant breeding studies because they fail to account for the interrelationships among multiple traits (Fikdalski, *et al.*, 2007). In contrast, multivariate analyses provide a more comprehensive understanding of complex datasets by simultaneously evaluating several variables and their correlations (Grobe, 2005). These techniques are particularly valuable in plant breeding programmes, where they enhance selection efficiency and improve decision-making (Beebe *et al.*, 1998).

Multivariate approaches enable the concurrent assessment of multiple traits, facilitating a better interpretation of the genetic and phenotypic variation within germplasm collections (Santos, 2010). Consequently, they have been widely applied in genetic diversity studies across diverse crops, including barley (*Hordeum vulgare* L.) (Cross, 1992), sorghum (*Sorghum bicolor* L. Moench) (Ayana and Bekele, 1999), wheat (*Triticum* spp.) (Hailu *et al.*, 2006), peanut (*Arachis hypogaea* L.) (Upadhyaya *et al.*, 2009), rice (Bharadwaj *et al.*, 2001), and ginger (*Zingiber officinale* Roscoe) (Abua *et al.*, 2021).

Among the commonly employed multivariate techniques, Principal Component Analysis (PCA) and cluster analysis are particularly useful for genotype selection aligned with breeding objectives (Mohammadi and Prasanna, 2003). Cluster analysis classifies genotypes into groups based on trait similarity, seeking to maximize homogeneity within clusters while enhancing heterogeneity among them (Rencher, 2002; Ralmawati *et al.*, 2019). This approach facilitates the detection of hidden patterns, relationships, and key discriminating traits within datasets (Pellegrini, 2017; Pan, 2003). Similarly, PCA is a powerful tool for identifying major sources of variation within complex data, enabling plant breeders to pinpoint key polygenic traits that contribute significantly to diversity (Afuape *et al.*, 2011; Beheshtizadeh *et al.*, 2013). Beyond simplifying data structure, PCA assists in selecting promising parental lines for genetic improvement and often serves as a basis for visualizing clusters derived from hierarchical clustering analyses (Wang *et al.*, 2018).

Ginger (*Zingiber officinale* Roscoe) has gained considerable economic importance due to its high market value, particularly in Nigeria, where its ginger is highly sought after for its superior oleoresin content (Abua *et al.*, 2024). The increasing demand has driven expansion of its cultivation into new agro-ecological zones, underscoring the need to evaluate and identify superior genotypes adapted to these diverse environments. Applying robust multivariate tools such as PCA and cluster analysis is essential for selecting elite genotypes with desirable agronomic traits for incorporation into breeding programmes. Therefore, this study aimed to classify, identify, and select superior ginger genotypes based on trait similarities using PCA and cluster analysis, thereby supporting targeted genetic improvement and the development of regionally adapted cultivars.

Materials and Methods

Study Site, Experimental materials and Experimental Design

The field experiment was conducted during the 2023 cropping season (March–December) at the Cross River State River Basin Development Authority, Nigeria. Fifteen ginger (*Zingiber officinale* Roscoe) mutant lines — UG2-9-01, UG1-5-49, UG1-5-38, UG1-5-52, UG1-5-18, UG2-2-35, UG1-5-04, UG1-5-22, UG1-5-35, UG2-11-03, UG1-11-07, UG1-5-31, UG1-7-24, UG1-5-48, and

UG1-13-02 — were obtained from the National Root Crops Research Institute (NRCRI), Umudike, Abia State, Nigeria. The experiment was laid out in a Randomized Complete Block Design (RCBD) with three replications. The experimental field measured 37.5 m × 9 m (total area: 337.5 m²), and each plot accommodated 12 plants. Ginger setts were planted at an inter- and intra-row spacing of 40 cm, corresponding to a plant population density of 62,500 plants ha⁻¹, with a total of 1,232 plants established across the entire field.

Data Collection

Data were recorded on key agronomic and yield-related traits, including: emergence percentage, establishment count, plant height, number of leaves per plant, leaf length, leaf width, number of tillers per plant, number of rhizome fingers per plant , rhizome length and rhizome yield.

Data Analyses

The collected data were subjected to hierarchical cluster analysis and Principal Component Analysis (PCA) to assess genetic relationships and identify traits contributing most to variation among genotypes. All statistical analyses were performed using the Plant Breeding Tools (PBTools) statistical software.

Results

Principal Component Analysis (PCA)

The results of the Principal Component Analysis (PCA) are summarized in Table 1. The first four principal components (PCs) accounted for 93.64% of the total variation among the ginger mutant lines. PC1 alone accounted for the highest proportion (72.87%), followed by PC2 (13.37%), PC3 (6.59%), and PC4 (1.14%). Trait loading revealed that rhizome length contributed most strongly to PC1 (loading = 0.9912), indicating its central role in explaining overall variability among the genotypes. In PC2, establishment count (loading = −0.8189) and plant height (loading = −0.4558) were the major contributors, while number of rhizome fingers had the highest loading in PC3 (0.9645). For PC4, plant height showed the strongest positive contribution (0.7840). Most traits had positive loadings on PC1, except for establishment count (−0.0041), number of leaves (−0.0166), and rhizome yield (−0.0176), which exhibited weak negative loading. Conversely, nearly all traits loaded negatively on PC2, except number of rhizome fingers (0.1025). The eigenvalues for PC1, PC2, PC3, and PC4 were 285.73, 52.48, 25.85, and 12.30, respectively, confirming the predominance of the first two components in explaining trait variation.

Cluster Analysis

Hierarchical cluster analysis grouped the 15 ginger mutant lines into four distinct clusters based on trait similarities (Table 2). Cluster 1 comprised five genotypes: UG2-9-01, UG1-5-52, UG1-5-18, UG1-5-31, and UG1-7-24. Cluster 2 included six genotypes: UG1-5-49, UG1-5-35, UG1-5-04, UG2-11-03, UG1-11-07, and UG1-5-48. Cluster 3 contained two genotypes: UG1-5-38 and UG1-5-22. Cluster 4 consisted of two genotypes: UG2-5-35 and UG1-13-02.

In the dendrogram (figure 2), clusters were visually separated, with Cluster 1 highlighted by green boundaries, Cluster 2 by red, Cluster 3 positioned on the extreme right, and Cluster 4 located between Clusters 1 and 2.

The cluster means and standard deviations of the evaluated traits are presented in Table 3. Distinct phenotypic profiles were observed across clusters: Cluster 1 (five genotypes) characterized by shorter plants, fewer and smaller leaves, shorter rhizomes, and low yield. Cluster 2 (six genotypes) exhibited higher establishment counts, more rhizome fingers, longer rhizomes, and high yield potential. Cluster 3 (two genotypes) displayed high emergence rates, taller plants, broader leaves, longer rhizomes, and high yields. Cluster 4 (two genotypes) characterized by lower emergence, greater leaf number, longer leaves, more tillers, fewer rhizomes, and low yield.

Discussion

The proportion of variance explained by principal components (PCs) is a critical determinant for selecting the number of components to retain in multivariate analyses. According to Rencher (2002), PCs retained for interpretation should collectively account for at least 70% of the total variation. In this study, the first four PCs explained 93.64% of the total variability among the evaluated ginger mutant lines, indicating that these components captured nearly all relevant phenotypic diversity. Consequently, the traits contributing most strongly to these components are highly informative for genotype differentiation and should receive priority in ginger breeding programmes.

The analysis revealed that rhizome length, establishment count, plant height, and number of rhizome fingers were the major contributors to the overall variation, suggesting that these traits are critical discriminators among the genotypes

Table 1: Principal component analysis of ginger mutant lines

Traits	PC1	PC2	PC3	PC4
Emergence	0.00089	-0.1253	0.1358	-0.0932
Establishment	-0.0041	-0.8189	0.1474	-0.5132
Plant height (cm)	0.0059	-0.4558	-0.0458	0.7840
Number of leaves	-0.0166	-0.2232	-0.0532	0.1961
Leaf length (cm)	0.0165	-0.1822	-0.0367	0.1292
Leaf width (cm)	0.0036	-0.0042	0.0129	0.0428
Number of tillers	0.0032	-0.0875	-0.0361	0.1258
Number of rhizome fingers	0.1282	0.1025	0.9645	0.1336
Rhizome length (cm)	0.9912	-0.0144	-0.1239	-0.0200
Rhizome yield ($t\ ha^{-1}$)	-0.0176	-0.0675	0.0806	0.1493
Standard deviation	16.9035	7.2400	5.0839	3.5077
Proportion of variance (%)	72.87	13.037	6.59	1.14
Cumulative proportion (%)	72.87	86.24	92.83	95.97
Eigen values	285.727.6	54.79	25.8465	12.3037

Table 2: Composition of cluster membership

Clusters	**Number of genotypes**	**Cluster Membership**
1	5	UG2-9-01, UG1-5-52, UG1-5-18, UG1-5-31, UG1-7-24
2	6	UG1-5-49, UG1-5-35, UG1-5-04, UG2-11-03, UG1-11-07, UG1-5-48
3	2	UG1-5-38, UG1-5-22
4	2	UG2-5-35, UG1-13-02

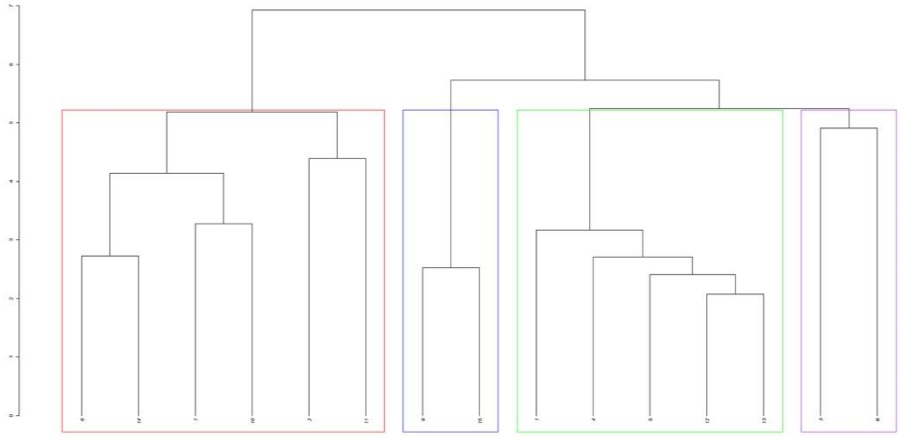

Figure 1: Dendrogram showing the relationship between genotypes

Table 3: Cluster Means and Standard Deviation values of ginger traits

Variable	Cluster	Min	Max	Mean	Std Dev
Emergence	1	33.33	46.17	40.99	5.98
Emergence	2	30.73	51.27	40.17	8.98
Emergence	3	51.27	51.30	51.28	0.02
Emergence	4	30.37	35.90	33.13	3.91
Establishment	1	58.97	79.47	65.93	8.42
Establishment	2	58.97	87.17	73.48	11.85
Establishment	3	58.97	69.23	64.10	7.25
Establishment	4	56.40	67.93	62.17	8.15
Plant height	1	14.03	17.53	16.39	1.40
Plant height	2	15.60	20.33	17.49	1.80
Plant height	3	17.00	43.40	30.20	18.67
Plant height	4	18.67	20.67	19.67	1.41
No. of leaves	1	8.33	9.37	8.94	0.56
No. of leaves	2	9.00	10.70	10.01	0.64
No. of leaves	3	9.33	9.87	9.60	0.38
No. of leaves	4	11.27	12.00	11.63	0.52
Leaf length	1	11.20	12.33	11.75	0.47
Leaf length	2	9.20	13.20	12.00	1.42
Leaf length	3	12.33	13.27	12.80	0.66
Leaf length	4	14.27	15.00	14.63	0.52
Leaf width	1	1.20	1.93	1.54	0.29
Leaf width	2	1.33	2.00	1.69	0.24
Leaf width	3	1.93	3.00	2.46	0.76
Leaf width	4	1.97	2.00	1.98	0.02
No. of tillers	1	2.67	3.27	2.92	0.26
No. of tillers	2	2.60	4.67	3.38	0.73
No. of tillers	3	2.67	3.00	2.83	0.23
No. of tillers	4	3.93	4.33	4.13	0.28
No. of rhizome finger	1	3.28	5.40	4.78	0.87
No. of rhizome finger	2	5.20	10.00	7.62	2.09
No. of rhizome finger	3	4.63	4.87	4.75	0.17
No. of rhizome finger	4	2.97	4.67	3.82	1.20
Rhizome length	1	6.67	9.87	8.40	1.40
Rhizome length	2	9.80	12.33	11.27	1.09
Rhizome length	3	9.97	10.17	10.07	0.14
Rhizome length	4	9.00	10.27	9.63	0.90
Rhizome yield	1	0.77	2.00	1.32	0.53
Rhizome yield	2	1.40	4.33	2.51	0.99
Rhizome yield	3	1.47	2.53	2.00	0.75
Rhizome yield	4	0.48	1.82	1.15	0.95

Therefore, should be considered in ginger breeding programmes. PC1, which explained the largest proportion of variation (72.87%), was strongly associated with rhizome length, indicating its potential as a key selection criterion for yield improvement. Similarly, PC2 was predominantly influenced by establishment count and plant height, while PC3 and PC4 were mainly explained by the number of rhizome fingers and plant height, respectively.

Eigenvector loadings indicated both positive and negative contributions of traits to each component, highlighting the presence of contrasting relationships among traits. For instance, traits with high positive loadings within a PC contribute synergistically to genotype differentiation, whereas those with negative loadings oppose these trends. These findings are consistent with the reports of Clifford and Stephenson (1975) and Guei *et al.* (2005), who emphasized that the first three principal components generally capture the most informative patterns of variation and are therefore highly relevant for breeding decisions.

While PCA identified the traits contributing most to overall diversity, hierarchical cluster analysis provided complementary insights by grouping genotypes with similar phenotypic characteristics. The 15 ginger mutant lines were classified into four distinct clusters, revealing substantial genetic diversity within the evaluated germplasm. Cluster 2 contained six genotypes characterized by higher establishment counts, more rhizome fingers, longer rhizomes, and high rhizome yield. These genotypes represent promising candidates for yield-focused breeding programmes. Cluster 3, comprising two genotypes, exhibited high emergence rates, taller plants, broader leaves, and longer rhizomes, making them potential donors of vigorous growth and biomass-related traits. Cluster 1 included five genotypes with shorter plants, fewer and smaller leaves, and lower rhizome yields, which could still serve as useful parental lines for introducing compact plant architecture. Cluster 4, with two genotypes, was distinguished by lower emergence, longer leaves, more tillers, and fewer rhizomes. These genotypes may be valuable for breeding vegetative growth traits rather than yield.

This integrative approach, combining PCA and cluster analysis, allows breeders to simultaneously identify the most discriminating traits and select elite genotypes tailored to specific breeding objectives. Similar findings have been reported by Vishnu *et al.* (2014) and RaviKumar *et al.* (2015) who highlighted the power of multivariate techniques in distinguishing genetically diverse germplasm and supporting efficient parent selection.

The identification of high-yielding genotypes such as those in Clusters 2 (UG1-5-49, UG1-5-35, UG1-5-04, UG2-11-03, UG1-11-07, UG1-5-48) and 3 (UG1-5-38, UG1-5-22) alongside key discriminating traits such as rhizome length, number of rhizome fingers, and establishment count would provide a strategic framework for ginger crop improvement. Incorporating these superior genotypes into breeding pipelines could accelerate the development of high-yielding, regionally adapted, and market-preferred cultivars. Moreover, genotypes from contrasting clusters offer an opportunity for heterotic hybridization, maximizing variability and potentially enhancing genetic gains. Overall, this study demonstrates the complementary value of PCA and hierarchical clustering in evaluating genetic diversity, prioritizing traits for selection, and identifying elite genotypes for breeding. Such integrative multivariate approaches are critical for designing efficient, targeted, and impactful ginger breeding strategies capable of meeting increasing global demand.

Conclusion

This study successfully applied Principal Component Analysis (PCA) and hierarchical cluster analysis to evaluate phenotypic diversity among 15 ginger (*Zingiber officinale* Roscoe) mutant lines. The first four principal components explained 93.64 % of the total variation, with rhizome length, establishment count, plant height, and number of rhizome fingers emerging as the most discriminating traits. Hierarchical clustering grouped the genotypes into four distinct clusters, revealing considerable genetic variability within the evaluated germplasm. Genotypes within Clusters 2 and 3 exhibited superior agronomic performance, characterized by higher emergence rates, greater establishment counts, longer rhizomes, more rhizome fingers, and higher yields, making them promising candidates for yield improvement programmes. Conversely, genotypes in Clusters 1 and 4, though associated with lower yields, possess unique vegetative growth traits that may be exploited for specific breeding objectives such as plant architecture modification and adaptability enhancement. The integration of PCA and clustering approaches provides a robust framework for elite genotype selection, trait prioritization, and strategic parent choice in ginger breeding. These findings contribute valuable insights for designing targeted crop improvement programmes aimed at developing high-yielding, regionally adaptable, and market-preferred ginger cultivars. Future studies should incorporate molecular diversity analyses alongside phenotypic evaluations to further refine selection strategies and enhance breeding efficiency.

Acknowledgements:

The authors are grateful to Dr Ayuk of Cross River Basin Development Authority for providing the piece of land used for this research

Conflict of interest statement:

The authors declare that there is no conflict of interest regarding the publication of this article.

Authors contributions:

All the authors worked together in designing the experiment, data collection result and discussion and reviewed the article.

References

Abua, M. N., Obok, E. E., Usen, M. J. and Iwo, G. A.(2024) Comparative evaluation of the growth and yield indices in some mutant lines of ginger (*Zingiber officinale.* Rosc.). *African Journal of Agriculture, Technology and Environment,* 13(1): 12-17

Abua, M.N., Ittah, M.A. and Iwo, G.A, (2021) Application of AMMI biplot model to evaluate some ginger (*Zingiber officinale* Rosc.) genotypes for adaptation and stability. *Journal of Agricultural* and *Crop Research*, 9 (5):121-133.

Afuape, S.O., Okocha P. I. and Njoku D. (2011). Multivariate assessment of the agromorphological variability and yield components among sweet potato (*Ipomoea batatas* (L.) Lam) landraces. *African journal of plant science* 5(2): 123-132.

Ayana A, Bekele E (1999). Multivariate analysis of morphological variation in sorghum (Sorghum bicolor L. Moench) germplasm from Ethiopia and Eritrea Genetic Resource. *Crop Evolution* 46: 273-284

Beebe, K. R., Pell, R. J., Seasholt, M. B. (1998). Chemometrics: A practical guide. New York, NY: John Wiley and Sons

Beheshtizadeh, H., A. Rezaie, A. Rezaie, and Ghandi. A. (2013). Principal component analysis and determination of the selection criteria in bread wheat (*Triticum aestivum* L.) genotypes. *International Journal of Agriculture and Crop Sciences* 5(18): 2024-2027.

Bharadwaj Ch, Tara SC, Subramanyam D (2001). Evaluation of different classificatory analysis methods in some rice (*Oryza sativa L.*) collections. *Indian Journal of Agricultural Science*. 71(2): 123-125.

Clifford, H.T. and Stephenson, W. (1975). An Introduction to Numerical Classification. Academic Press, London. p. 229

Cross, R. J. (1992). A proposed revision of the IBPGR barley descriptor list. Theory. Application *Genetics*. 84: 501-507

Fidalski, J., Tormena, C. A., and Scapim, C. A. (2007). Espacialização vertical e horizontal dos indicadores de qualidade para um latossolo vermelho cultivado com citros. *Revista Brasileira de Ciência do Solo*, 31(1), 9-19

Grobe, J. R. (2005). Aplicações da estatística multivariada na análise de resultados em experimentos com solos e animais (Master's thesis, Universidade Federal do Paraná, Curitiba

Guei, R.G.; Sanni, K.A. and Fawole, A.F.J. (2005). Genetic diversity of rice (O. sativa L.). *Agronomy Africa*, 5: 17-28

Hailu F, Merker A, Singh H, Belay G, Johansson E (2006). Multivariate analysis of diversity of tetraploid wheat germplasm from Ethiopia. Genet. Resource. *Crop Evolution*, 54: 83-97.

Mohammadi SA. And Prasanna BM. (2003). Analysis of genetic diversity in crop plants-salient statistical tools and considerations. *Crop Science*, 43:1235-1248.

Pan, W. (2003*).* *"*Cluster Analysis" *Journal of machine learning research:* pp7-14.

Pellegrini, M (2017). "Cluster analysis and subgrouping to investigate inter-individual variability to non-invasive brain stimulation: a systematic review." *Reviews in the Neurosciences* 29, 6, 675-697.

Ralmawati, W and Putri A .Y (2019). Fuzzy subtractive clustering *Journal of Mathematical Science and Statistics* pp 1-4

Ravikumar, B.N., Naga Kumari, Venkata Ramana R., Girija Rani, P.V. Satyanarayana, N.Chamundeswari, K. Mohan Vishnuvardhan, Y. Suryanarayana,M. Bharathalakshmi and Vishnuvardhan Reddy A. (2015). Principal component analysis and character association for yield components in rice (Oryza sativa L.) cultivars suitable for irrigated ecosystem. Current *Biotica*. 9(1):25-35.

Rencher, A. C., and Christensen, W. F. (2002). *Methods of Multivariate Analysis*. John Wiley and Sons, Inc.: Toronto, ON, Canada. https://doi.org/10.1002/0471271357

Santos, E. F. N. (2010). Formação de grupos produtivos em vacas leiteiras por meio de componentes principais. Revista brasileira biométrica, v. 28, n. 3, p. 15-22.

Upadhyaya HD, Reddy LJ, Dwivedi SL, Gowda CLL, Singh S (2009). Phenotypic diversity in cold-tolerant peanut (*Arachis hypogaea* L.) germplasm. Euphytica, 165: 279-291.

Vishnu, V.N., Robin S., Sudhakar, D., Raveendran, M., Rajeswari, S. and Manonmani, S. (2014). Evaluation of Rice Genetic Diversity and Variability in a Population Panel by Principal Component Analysis. *Indian Journal of Science and Technology,* 7(10) 1555–1562.

Wang S, Wang J and Han Y. (2018). Principal component analysis of early alcohol, drug and tobacco use with major depressive disorder in US adults. *Journal of Psychiatric Research,* 100: 113-120.