GMR
The Original

GENETIC DIVERGENCE AND PRINCIPAL COMPONENT ANALYSIS IN F_4 SEGREGATING LINES OF RICE (*ORYZA SATIVA* L.) USING MAHALANOBIS D^2 STATISTICS

Tanistha Singh Roy[1], Vishal Singh[2*], Karthik Chittibomma[3], Rubleen Kaur[4], Vivek Singh[5], Abhishek Tiwari[6], Birina Kaushiki[7], Shabnam Kumari[8], Saurabh Raj Pandey[9] and Puneet Kumar[10]

[1]Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh, India.
[2*]Research Scholar, Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, Uttar Pradesh, India.
[3]Ph.D. Scholar, Department of Genetics and Plant Breeding, ICAR-Indian Institute of Rice Research, Rajendranagar, Hyderabad, India.
[4]Ph.D. Scholar, Department of Plant Pathology, CCSHAU, Hisar, Haryana, India.
[5]Assistant Professor, Department of Genetics and Plant Breeding, College of Agriculture, RLBCAU, Jhansi, Uttar Pradesh, India.
[6]Assistant Professor, Department of Soil Science and Agricultural Chemistry, SAAST, C.S.J.M. University, Kanpur, Uttar Pradesh, India.
[7]Research Scholar, Department of Sericulture, Assam Agricultural University, Jorhat, Assam, India.
[8]Assistant Professor, Department of Agronomy, Teerthanker Mahaveer College of Agriculture Sciences, TMU, Moradabad, Uttar Pradesh, India.
[9]SRF, ICAR-Agricultural Technology Application Research Institute, Zone-III, Kanpur, Uttar Pradesh, India.
[10]Teaching Associate, Department of Entomology, Chandra Shekhar Azad University of Agriculture and Technology, Kanpur, Uttar Pradesh, India.
***Corresponding author:** Vishal Singh, vishal19@bhu.ac.in

ABSTRACT
Systematic characterization of genetic divergence within breeding populations is fundamental to plant genetic resources management, as it enables breeders to identify genetically diverse parental combinations with the potential to generate broad genetic variability and desirable transgressive segregants. In the present study, genetic divergence among forty F_4 rice (*Oryza sativa* L.) lines derived from the cross Kalarata × Pusa 44 was assessed using Mahalanobis D^2 statistics and principal component analysis (PCA). The lines were evaluated for twelve quantitative traits in a randomized complete block design with three replications during the Kharif season of 2025 at Banaras Hindu University, Varanasi, India. Tocher's method grouped the forty genotypes into thirteen non-overlapping clusters, with Clusters II and IV accommodating the maximum number of genotypes (seven each), while nine clusters comprised only two genotypes each, indicating substantial genetic divergence among a subset of the lines. The maximum inter-cluster distance was recorded between Clusters X and XI (D^2 = 611.96), followed by Clusters VIII and X (539.86), whereas the minimum distance occurred between Clusters II and III (27.59). Cluster I exhibited the highest mean values for grain yield per plant, filled grains per panicle, test weight and spikelet fertility, identifying it as an agronomically promising group. Filled grains per panicle contributed the largest proportion of total genetic divergence (56.28%), followed by plant height (18.08%) and days to 50% flowering (17.18%). PCA provided complementary evidence of trait variation, with the first three principal components accounting for 78.15% of the total variation (PC1 = 56.34%, PC2 = 12.98% and PC3 = 8.83%). Yield-related reproductive traits contributed predominantly to PC1, whereas morphological and phenological traits were more strongly associated with PC2. The substantial inter-cluster distances involving Clusters X, VIII and XI, together with the favourable trait means observed in Cluster I, highlight these groups as potential genetic resources for selecting diverse parents for future rice improvement programmes.

KEYWORDS: *Oryza sativa* · F_4 population · Genetic divergence · Mahalanobis D^2 · Tocher's clustering · Principal component analysis · Plant genetic resources

INTRODUCTION
Rice (*Oryza sativa* L.) ranks among the most crucial cereal crops grown globally, significantly contributing to worldwide food production. Ongoing genetic enhancement of this crop is crucial to satisfy the food needs of a growing global population, despite the increasing limitations on land, water, and climate resources (Muthayya et al., 2014; Zeigler and Barclay, 2008). Global rice production has continued to increase, surpassing 540 million tonnes on a milled basis, with India contributing nearly 152 million tonnes from approximately 52 million hectares during the 2025/26 season (USDA, 2025). Sustaining further genetic gains depends critically on the effective conservation and utilization of genetically diverse plant genetic resources (Salgotra and Chauhan, 2023; Bairwa et al., 2024). Genetic diversity provides the foundation for developing improved cultivars, as crosses between genetically diverse parents may generate greater variability and increase the opportunities for recovering desirable recombinants in subsequent generations (Falconer and Mackay, 1996; Khush, 1997).

Selection of parents for hybridization based solely on geographic origin or phenotypic dissimilarity is often unreliable, as morphological similarity does not necessarily reflect underlying genetic relationships (Murty and Arunachalam, 1966). Mahalanobis D^2 statistics (Mahalanobis, 1936) provide an objective multivariate measure of genetic divergence by accounting for correlations among the traits under consideration. This approach has been extensively used for characterizing genetic diversity in crop germplasm, commonly in combination with Tocher's clustering method (Rao, 1952). In rice, D^2 statistics have been widely employed to classify genotypes into distinct clusters and identify potential parental combinations for hybridization and population improvement (Nivedha et al., 2024; Mondal et al., 2024). Crosses between genotypes belonging to widely separated clusters may offer opportunities to exploit genetic diversity and recover desirable segregants, although their breeding value also depends on parental performance and the inheritance of target traits (Arunachalam, 1981; Rao, 1952; Nivedha et al., 2024). Principal component analysis (PCA), based on the eigen-decomposition of the trait correlation matrix, complements D^2 analysis by identifying a smaller number of principal components that account for a substantial proportion of the total variation and by highlighting the traits contributing most strongly to this variation (Hotelling, 1933; Jolliffe, 2002; Greenacre et al., 2022). Thus, the combined application of D^2 statistics and PCA provides complementary information on the extent, structure and major sources of phenotypic variation within breeding populations.

Recent multivariate studies of rice genetic resources have demonstrated the value of combining D^2-based clustering with PCA. Grain-number, panicle-related and morphological traits have frequently been identified as major contributors to genetic divergence, while genotypes from widely separated clusters have been suggested as potential parents for hybridization and as valuable accessions for germplasm conservation (Bhajipale et al., 2025; Gayathri et al., 2026; Islam et al., 2025; Sinha et al., 2019; Thakare et al., 2025). Similar approaches, sometimes supplemented by SSR marker analysis, have been used to characterize genetic relationships and identify divergent rice accessions for breeding and conservation purposes (Debsharma et al., 2023; Mohanavel et al., 2021; Mondal et al., 2024; Nivedha et al., 2024; Prakash et al., 2024; Yadav et al., 2024). Under salt-affected conditions, D^2- or Euclidean-distance-based analyses have grouped rice genotypes into distinct clusters, with the number of clusters and patterns of divergence varying according to the genetic material and environmental conditions evaluated (Arunkumar and Narayanan, 2023; Maurya et al., 2025; Shrivastav et al., 2025; Sharma et al., 2026). PCA combined with cluster analysis has also been used to compare patterns of genetic variation under sodic and normal soil conditions (Akilan et al., 2023). A recent evaluation of Bangladeshi Aman rice landraces published in *Scientific Reports* similarly identified grain-number and panicle-related traits as important contributors to phenotypic divergence, demonstrating the relevance of these traits across different germplasm sets and geographic regions (Islam et al., 2025). Despite these findings, the number of clusters, magnitude of inter-cluster distances and relative contribution of individual traits vary considerably among populations, generations and trait combinations. Consequently, the characterization of individual breeding populations remains important, particularly at early segregating generations, when the identification of divergent lines can help guide subsequent breeding decisions and the selection of potential parents for future hybridization programmes (Jain and Brar, 2009). Cluster-based characterization has also been applied to segregating rice populations evaluated under stress conditions, including F_3 populations subjected to salt stress (Laghari et al., 2025).

The present study characterizes forty F_4 lines derived from the cross between Kalarata, a pigmented black-pericarp landrace valued as a potential donor of grain-quality traits (Muthazhagu Kuppuraj et al., 2025), and Pusa 44, a high-yielding semi-dwarf indica variety widely cultivated in northern India (IARI, 2022). The study aimed to: (i) classify the forty F_4 lines into genetically distinct clusters using Mahalanobis D^2 statistics and Tocher's method; (ii) estimate inter- and intra-cluster distances and evaluate cluster mean performance for grain yield and its component traits; (iii) determine the relative contribution of individual traits to total genetic divergence; and (iv) examine the pattern of trait variation through PCA. The findings are expected to provide information on the genetic diversity and agronomic potential of the evaluated F_4 lines and help identify promising genetic resources for future rice improvement programmes.

MATERIALS AND METHODS

Plant material, site and experimental design

Forty F_4 rice lines derived from the cross Kalarata × Pusa 44 were evaluated at the Agricultural Research Farm, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India (25.14° N, 82.59° E; 100.51 m above mean sea level), during the Kharif season of 2025. The experiment was laid out in a randomized complete block design (RCBD) with three replications. Transplanting was carried out on 17 July 2025, and harvesting was completed by the first week of December 2025. Each replication comprised forty rows, with each genotype represented by ten plants. Individual plots measured 4 m × 2 m, with row-to-row and plant-to-plant spacing of 20 cm and 15 cm, respectively. All experimental plots were maintained under standard agronomic practices recommended for irrigated rice cultivation in the Eastern Gangetic Plains (IRRI, 2013; Singh et al., 2018).

TRAITS RECORDED

Recording of Observations

Twelve quantitative traits were recorded for the evaluated genotypes on a plant or plot basis, as appropriate: days to 50% flowering (DF), days to maturity (DM), plant height (PH, cm), panicle length (PL, cm), number of effective tillers per plant (ET), grain yield per plant (GYP, g), panicle weight (PW, g), number of filled grains per panicle (FG), number of unfilled grains per panicle (UFG), test weight (TW, g), grain length (GL, mm) and spikelet fertility (SF, %).

Mahalanobis D² analysis and clustering

Genetic divergence among the forty F_4 rice genotypes was estimated using Mahalanobis generalized distance (D^2) statistics (Mahalanobis, 1936), based on genotype means for the twelve quantitative traits. The analysis accounted for correlations among traits using the pooled variance–covariance matrix, following the procedure described by Rao (1952). The generalized squared distance between any two genotypes, *i* and *j*, was calculated as:[D^2_{ij}=(X_i-X_j)^T S^{-1}(X_i-X_j)]

where, (X_i) and (X_j) represent the vectors of trait means for genotypes *i* and *j*, respectively, and (S^{-1}) denotes the inverse of the pooled variance–covariance matrix.

The forty genotypes were subsequently grouped into clusters using Tocher's method (Rao, 1952). Under this method, genotypes were grouped sequentially according to their average generalized distances, with the objective of maintaining greater homogeneity within clusters than between clusters. Inter-cluster and intra-cluster distances were estimated as the average generalized distances among genotype pairs belonging to different clusters and within the same cluster, respectively. Cluster means were calculated for each of the twelve quantitative traits to assess the relative performance of the clusters.

The relative contribution of individual traits to total genetic divergence was estimated using the ranking procedure described by Rao (1952). For each pairwise comparison of genotypes, the trait contributing the maximum to the generalized distance was identified. The contribution of each trait was then expressed as the percentage of total pairwise comparisons in which it ranked first.

Principal component analysis

Principal component analysis (PCA) was performed following the methods described by Hotelling (1933), Jolliffe (2002) and Greenacre et al. (2022), using standardized genotype means (mean-centred and scaled to unit variance) for the twelve quantitative traits to identify the major axes of variation among the forty F_4 rice lines. Eigenvalues, percentage of variance explained and cumulative percentage of variance explained by successive principal components were calculated from the correlation matrix. Trait loadings and their contributions to individual principal components were examined to identify the traits responsible for the observed variation and to provide complementary information on the patterns of genetic divergence revealed by Mahalanobis D^2 analysis. A biplot of the first two principal components was constructed to display the positions of the lines in relation to the trait loadings.

Graphical representation of clusters

The inter- and intra-cluster D^2 matrix was displayed as a heat map. Hierarchical relationships among the clusters were visualized with an average-linkage (UPGMA; Sokal and Michener 1958) dendrogram computed from the inter-cluster D^2 matrix. To visualize the relative positions of the clusters, principal coordinate analysis (classical multidimensional scaling; Gower 1966) was applied to the inter-cluster Mahalanobis distances ($\sqrt{D^2}$). Cluster means were standardized across the thirteen clusters (z-scores) for the heat map of trait means, and the population mean of each trait was calculated as the mean of the cluster means weighted by the number of lines in each cluster. All graphics were prepared in Python using the SciPy and Matplotlib libraries.

RESULTS

Distribution of genotypes into clusters

Mahalanobis D^2 analysis followed by Tocher's clustering grouped the forty F_4 rice lines into thirteen distinct, non-overlapping clusters (Table 1), indicating substantial genetic diversity within the population. Clusters II and IV were the largest, each accommodating seven genotypes, followed by Clusters I and III, with four genotypes each. The remaining nine clusters (V–XIII) were smaller, each comprising only two genotypes. The distribution of genotypes among these clusters indicates considerable variation in the genetic relationships within the population. The genetic distinctness and breeding potential of these smaller clusters can be further assessed based on their inter-cluster distances and cluster mean performance, which may help identify promising parental combinations and potential sources of desirable segregants in subsequent breeding generations.

Table 1 Distribution of 40 F_4 rice lines into thirteen clusters based on Mahalanobis D^2 statistics

Cluster	No. of genotypes	Genotypes / lines
I	4	L19, L13, L9, L16
II	7	L38, L34, L11, L17, L25, L32,

Cluster	No. of genotypes	Genotypes / lines
		L14
III	4	L36, L31, L20, L40
IV	7	L37, L33, L12, L18, L26, L28, L15
V	2	L29, L23
VI	2	L8, L6
VII	2	L35, L10
VIII	2	L24, L22
IX	2	L7, L5
X	2	L27, L3
XI	2	L30, L4
XII	2	L21, L2
XIII	2	L39, L1

Inter-cluster and intra-cluster distances

Inter-cluster distances among the thirteen clusters varied widely (Table 2; Fig. 1), indicating a broad range of genetic divergence within the population. The maximum inter-cluster distance was observed between Clusters X and XI ($D^2 = 611.96$), followed by Clusters VIII and X (539.86), Clusters VII and X (470.34) and Clusters I and XI (452.14), indicating a high degree of genetic dissimilarity between the genotypes belonging to these cluster pairs. These widely separated clusters may represent potential sources of diverse parents for future hybridization programmes aimed at broadening genetic variability and increasing the possibility of recovering desirable segregants. In contrast, the minimum inter-cluster distance was recorded between Clusters II and III (27.59), followed by Clusters II and V (29.79), indicating comparatively greater genetic similarity between these groups. Intra-cluster distances (diagonal values) were lower than the smallest inter-cluster distance of the respective cluster for Clusters I–X, suggesting greater genetic similarity among genotypes within these clusters than between clusters. In contrast, the intra-cluster distances of Clusters XI (78.07), XII (124.29) and XIII (346.83) exceeded their smallest inter-cluster distances (41.44, 56.08 and 149.21, respectively), indicating greater heterogeneity between the two genotypes within each of these clusters. These results indicate that Tocher's clustering effectively differentiated the genotypes into groups based on their multivariate divergence.

Table 2 Inter-cluster (off-diagonal) and intra-cluster (diagonal) Mahalanobis D^2 distances among thirteen clusters of F_4 rice lines

Cluster	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
I	15.65	56.81	109.71	230.22	44.38	200.15	360.53	366.29	182.44	134.30	452.14	93.62	345.15
II		18.43	27.59	87.39	29.79	84.17	175.93	184.67	78.15	171.95	251.26	83.90	211.06
III			12.26	43.30	51.85	54.88	110.38	127.46	62.03	200.23	174.66	113.83	167.86
IV				18.75	138.31	39.16	39.68	69.91	65.18	341.53	82.24	195.28	152.52
V					14.16	128.66	241.26	275.70	129.74	115.20	337.75	56.08	273.05
VI						17.84	91.93	151.49	107.25	329.20	127.87	207.56	185.92
VII							21.70	54.22	97.25	470.34	41.44	281.96	173.36
VIII								31.12	92.68	539.86	76.87	314.91	207.14
IX									49.00	308.90	164.52	160.13	149.21
X										71.48	611.96	135.82	397.22
XI											78.07	391.96	246.52
XII												124.29	323.21
XIII													346.83

Diagonal values represent intra-cluster distance; all other values are inter-cluster D^2 distances

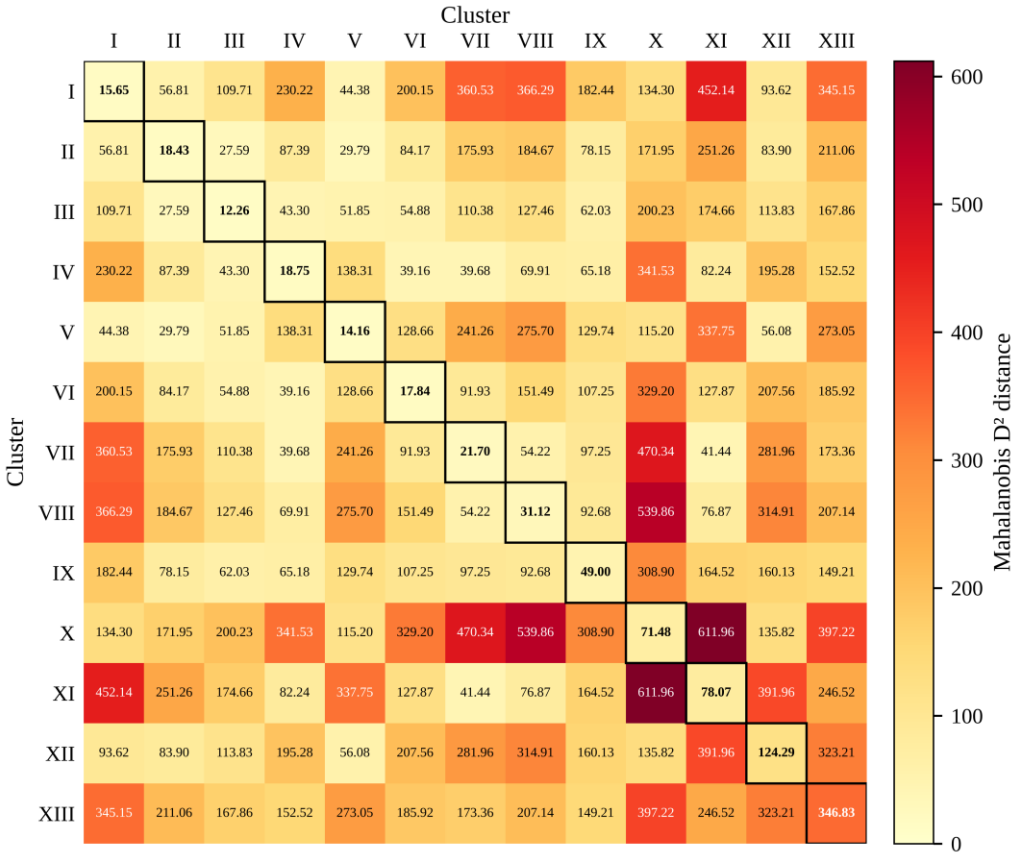

Fig. 1 Heat map of inter- and intra-cluster Mahalanobis D² distances among the thirteen clusters. Values are given in each cell; outlined diagonal cells are intra-cluster distances. Darker cells indicate greater genetic distance.

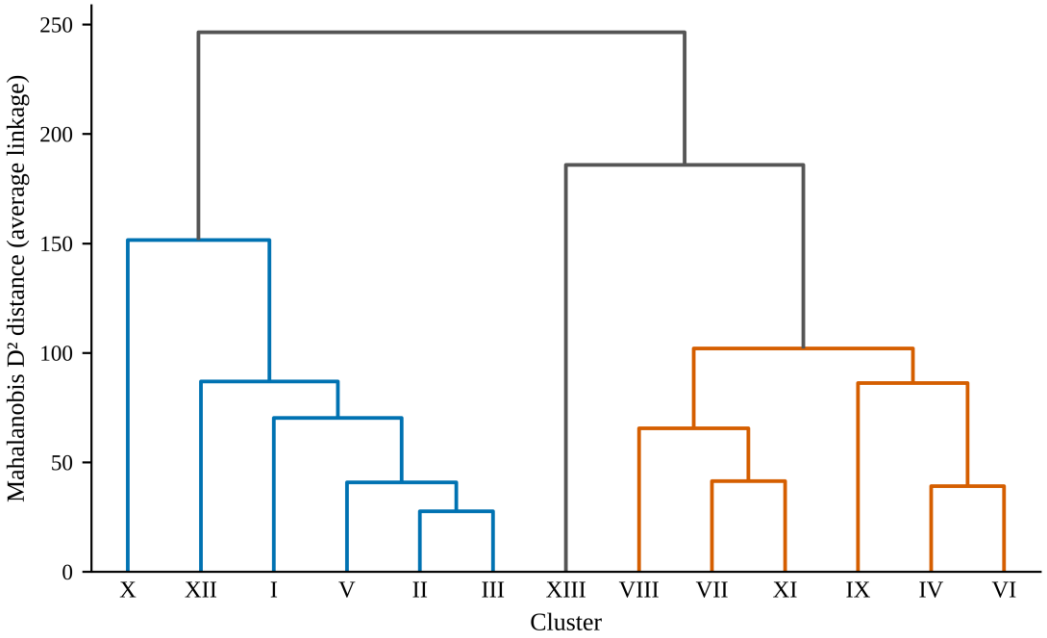

Fig. 2 Dendrogram showing the hierarchical relationship among the thirteen Tocher's clusters, derived from the Mahalanobis D² distance matrix in Table 2.

The average-linkage dendrogram (Fig. 2) arranged the thirteen clusters into two main groups: Clusters I, II, III, V, X and XII formed one group, and Clusters IV, VI, VII, VIII, IX, XI and XIII formed the other, the two groups joining at a linkage distance of approximately 246. Within the first group, Cluster X was the last to join (at approximately 152), and within the second group Cluster XIII joined last (at approximately 186), indicating that these two clusters were the most distinct members of their respective groups. The principal coordinate analysis of the inter-cluster distances (Fig. 3) showed a consistent pattern. The first two coordinates accounted for 73.5% of the total variation in the distance matrix (60.1% and 13.4%, respectively). Cluster X was

positioned at the negative extreme of the first coordinate, whereas Clusters XI, VIII and VII lay at the positive extreme, and Cluster XIII was separated from all other clusters along the second coordinate. Cluster I, which had the highest mean grain yield, lay on the negative side of the first coordinate close to Clusters XII and V, and was far from Cluster XI. The intra-cluster D^2 values were lowest in Clusters III (12.26), V (14.16) and I (15.65) and highest in Cluster XIII (346.83) (Fig. 4a); the pairs of clusters with the largest and smallest inter-cluster distances are ranked in Fig. 4b.

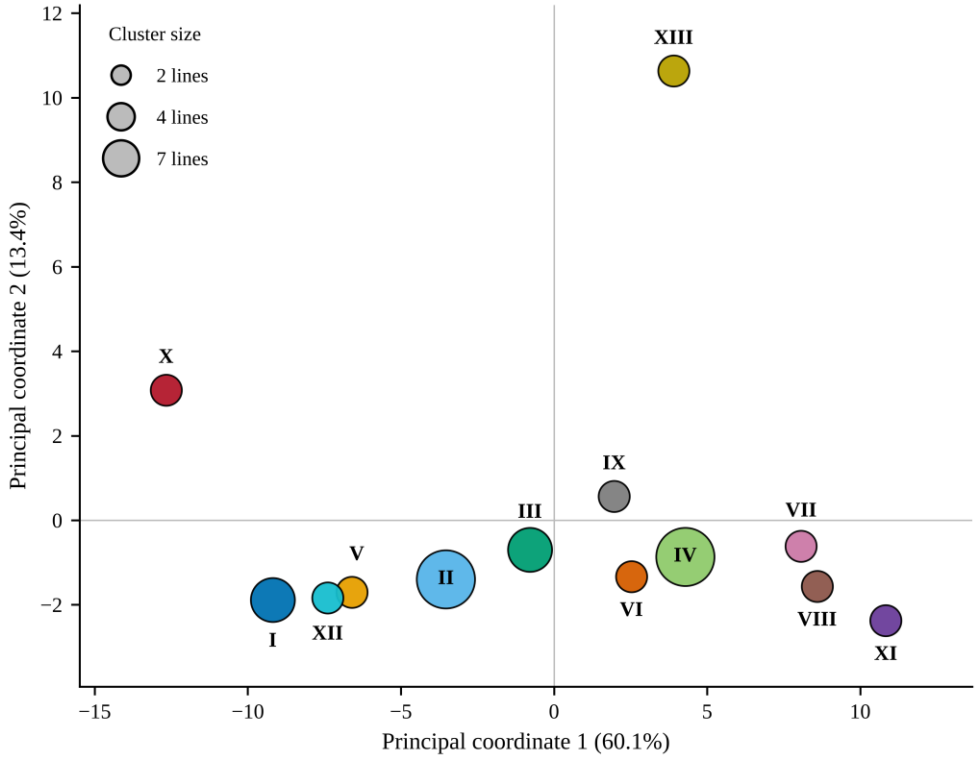

Fig. 3 Principal coordinate analysis (PCoA) of the thirteen Tocher's clusters based on the inter-cluster Mahalanobis distances ($\sqrt{D^2}$) in Table 2. The first two coordinates account for 60.1% and 13.4% of the total variation in the distance matrix. Circle area is proportional to the number of F_4 lines in the cluster.

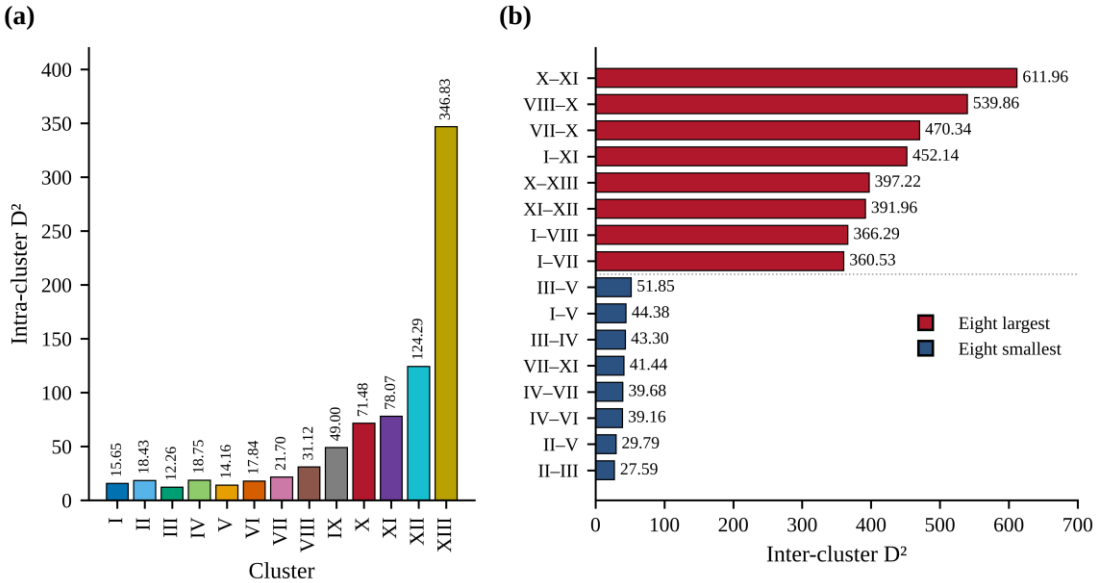

Fig. 4 (a) Intra-cluster D^2 values of the thirteen clusters and (b) the eight largest and eight smallest inter-cluster D^2 values.

Cluster mean performance

Cluster means for the twelve quantitative traits revealed considerable variation among the thirteen clusters with respect to flowering duration, maturity, plant height, yield components, grain characteristics and spikelet fertility (Table 3). Cluster I exhibited the highest mean values for grain yield per plant (24.18 g), filled grains per panicle (82.44), test weight (24.52 g) and spikelet fertility (79.22%; Fig. 6), along with an above-average plant height (132.17 cm). These favourable trait values identify Cluster I as an agronomically promising group

within the evaluated population and a potential source of superior lines for further selection. Cluster IX recorded the highest mean number of effective tillers per plant (13.05), whereas Cluster X exhibited the greatest mean plant height (142.32 cm), and Cluster XII recorded the greatest mean panicle length (27.25 cm). Cluster VIII showed the longest mean duration to 50% flowering (86.00 days). In contrast, Clusters XI and VII recorded the lowest mean values for grain yield per plant (12.97 g and 14.58 g, respectively), filled grains per panicle (42.27 and 45.03, respectively) and spikelet fertility (55.91% and 57.48%, respectively). These clusters showed comparatively lower performance for the major yield-related traits, although their genetic divergence may make them useful sources of variation for specific breeding objectives.

Table 3 Cluster means for twelve quantitative traits in 40 F_4 rice lines grouped by Mahalanobis D^2 analysis

Cluster	DF	DM	PH	PL	ET	GYP	PW	FG	UFG	TW	GL	SF
I	82.92	129.17	132.17	26.79	12.40	24.18	2.36	82.44	21.57	24.52	10.69	79.22
II	81.38	125.00	126.32	25.94	11.33	20.17	2.08	71.30	25.67	23.45	10.50	73.50
III	77.67	121.33	124.47	25.15	10.28	18.48	1.91	66.19	26.42	23.01	10.77	71.22
IV	73.00	119.48	118.33	25.25	9.98	16.15	1.86	55.51	28.87	22.12	10.09	65.67
V	81.67	127.33	134.53	25.52	11.13	21.78	1.98	71.40	25.47	23.57	10.49	73.69
VI	64.17	125.00	116.88	25.70	11.07	16.47	2.09	61.62	26.07	22.83	10.20	70.25
VII	71.83	116.00	116.47	24.05	9.45	14.58	1.82	45.03	33.40	22.17	9.34	57.48
VIII	86.00	116.00	113.30	25.43	9.23	15.27	1.81	48.68	32.60	21.98	10.17	59.76
IX	83.17	116.83	119.72	25.07	13.05	21.13	2.07	59.17	27.90	23.55	9.95	67.74
X	71.83	118.00	142.32	26.83	11.68	19.40	1.99	76.55	26.82	22.03	10.75	74.00
XI	70.00	118.67	112.67	23.45	9.32	12.97	1.94	42.27	33.15	22.23	9.70	55.91
XII	84.67	125.33	138.08	27.25	12.23	19.52	1.98	63.67	26.90	22.43	9.89	69.35
XIII	66.67	108.50	113.00	25.40	11.15	21.47	1.97	63.48	31.92	23.03	9.74	66.50

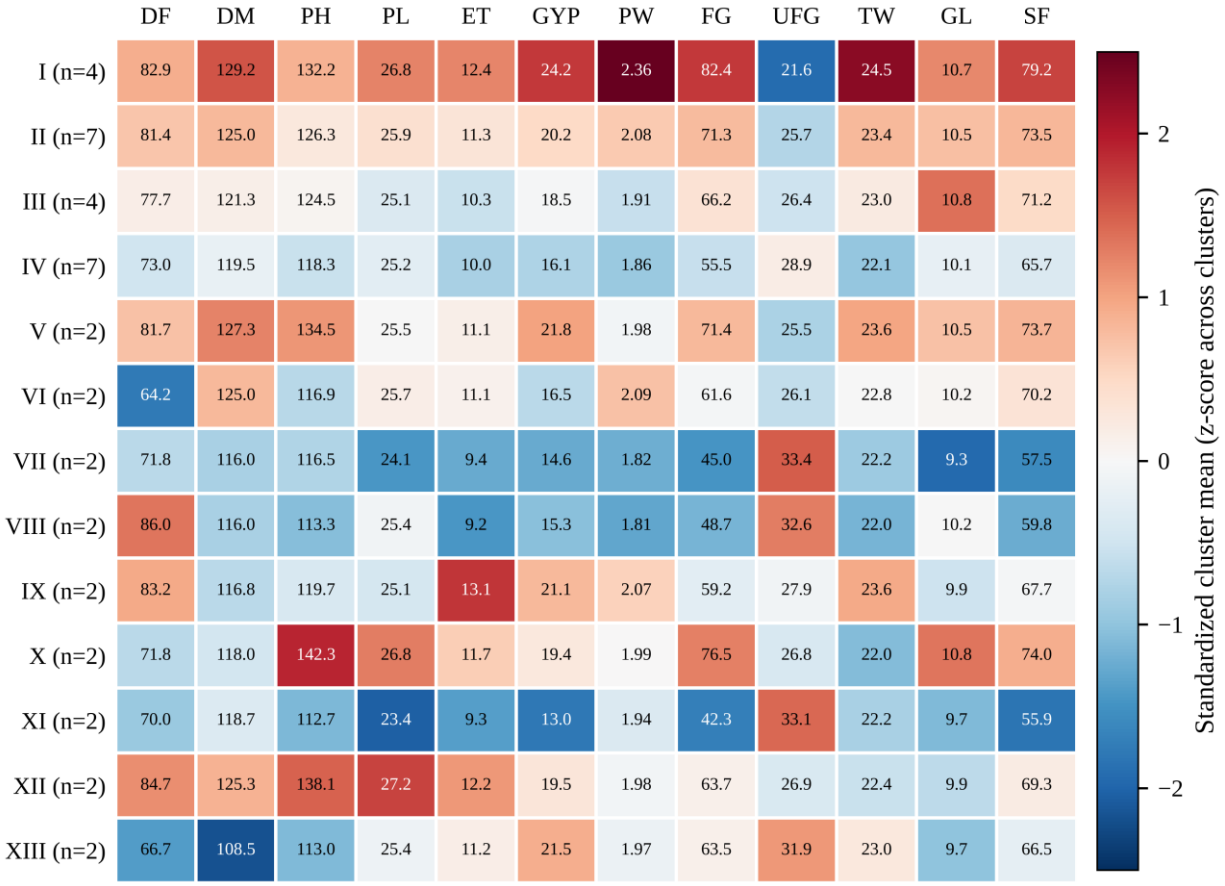

Fig. 5 Heat map of cluster means for the twelve quantitative traits (Table 3). Colours show cluster means standardized across the thirteen clusters (z-score; red, above and blue, below the mean of cluster means); cell labels give the original cluster means. DF, days to 50% flowering; DM, days to maturity; PH, plant height

(cm); PL, panicle length (cm); ET, effective tillers per plant; GYP, grain yield per plant (g); PW, panicle weight (g); FG, filled grains per panicle; UFG, unfilled grains per panicle; TW, test weight (g); GL, grain length (mm); SF, spikelet fertility (%).

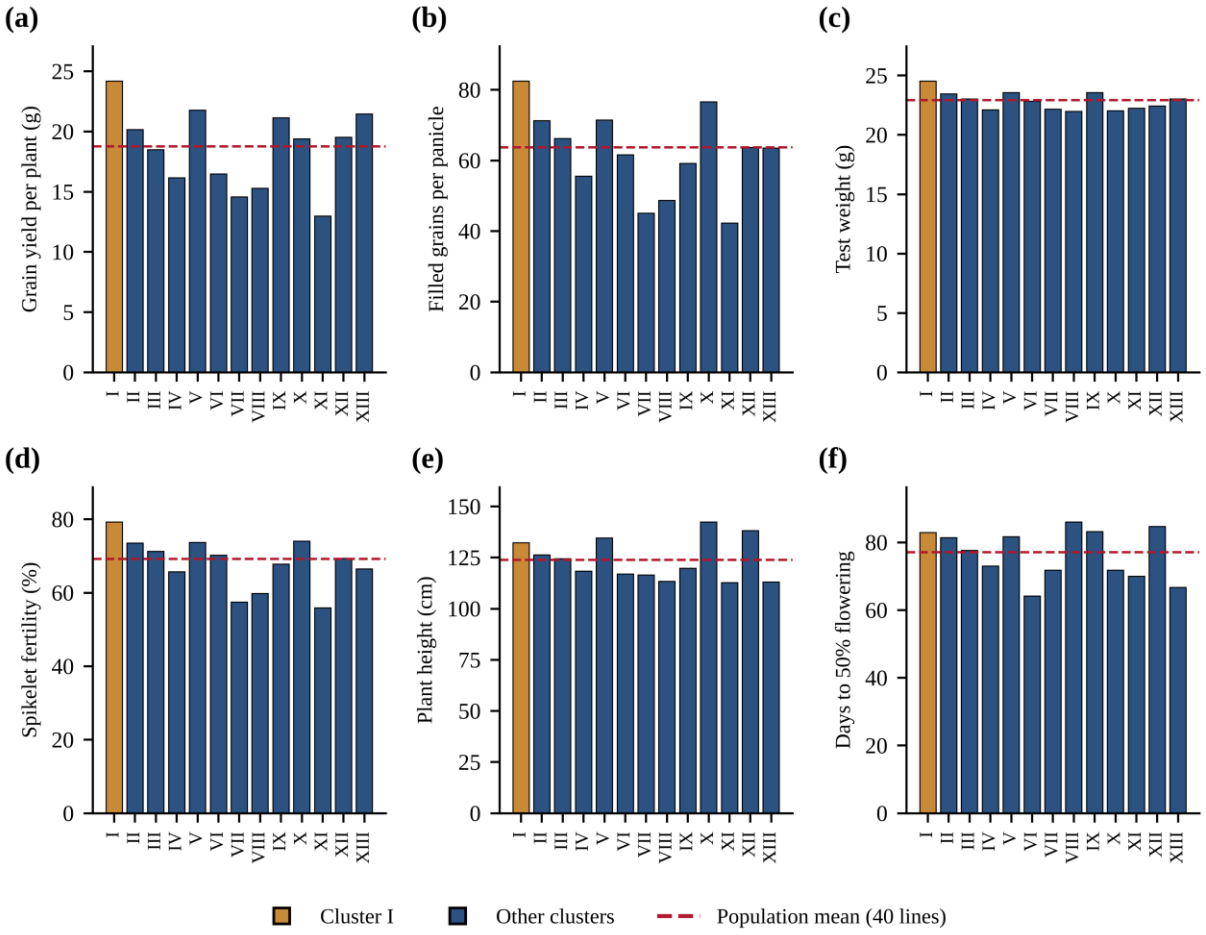

Fig. 6 Cluster means for (a) grain yield per plant, (b) filled grains per panicle, (c) test weight, (d) spikelet fertility, (e) plant height and (f) days to 50% flowering. Cluster I is highlighted; the dashed line is the population mean of the 40 lines (cluster means weighted by the number of lines per cluster).

Contribution of traits towards genetic divergence

The relative contributions of individual traits to total genetic divergence varied considerably (Table 4; Fig. 7). The number of filled grains per panicle was the largest contributor, accounting for 56.28% of the total divergence, followed by plant height (18.08%) and days to 50% flowering (17.18%). Days to maturity contributed 4.49%, whereas the number of unfilled grains per panicle (1.92%) and spikelet fertility (1.67%) made relatively minor contributions. Panicle length (0.26%) and grain yield per plant (0.13%) showed very low contributions. Effective tillers per plant, panicle weight, test weight and grain length had negligible contributions to the observed genetic divergence. These findings suggest that differences among the genotypes were primarily associated with variation in filled grains per panicle, plant height and flowering time, while the remaining traits contributed relatively little to the overall divergence under the conditions of the present study.

Table 4 Percentage contribution of individual traits towards total genetic divergence among 40 F_4 rice lines

Rank	Trait	Contribution (%)
1	Filled grains per panicle (FG)	56.28
2	Plant height (PH)	18.08
3	Days to 50% flowering (DF)	17.18
4	Days to maturity (DM)	4.49
5	Unfilled grains per panicle (UFG)	1.92
6	Spikelet fertility (SF)	1.67

Rank	Trait	Contribution (%)
7	Panicle length (PL)	0.26
8	Grain yield per plant (GYP)	0.13

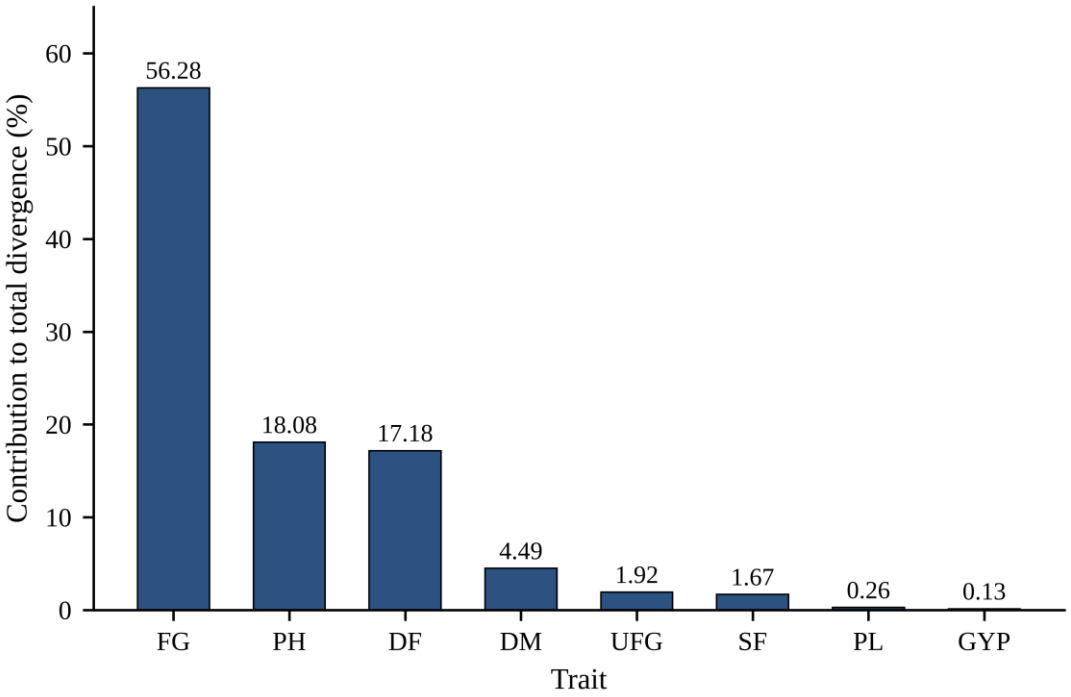

Fig. 7 Percentage contribution of individual traits towards total genetic divergence among the 40 F_4 rice lines. DF, days to 50% flowering; DM, days to maturity; PH, plant height; PL, panicle length; GYP, grain yield per plant; FG, filled grains per panicle; UFG, unfilled grains per panicle; SF, spikelet fertility.

Principal component analysis

Principal component analysis (PCA) provided complementary evidence of the patterns of genetic variation identified through Mahalanobis D^2 analysis. The first principal component (PC1) had the highest eigenvalue (6.760) and accounted for 56.34% of the total variation. The second principal component (PC2) had an eigenvalue of 1.558 and explained 12.98% of the variation, while the third principal component (PC3) had an eigenvalue of 1.060 and contributed 8.83%. Collectively, the first three principal components accounted for 78.15% of the total variation among the genotypes (Table 5; Fig. 8). The subsequent components (PC4–PC12) each contributed less than 6% to the total variation, with the cumulative variance reaching 100% across all twelve principal components.

Table 5 Eigenvalues, percentage variation and cumulative variance explained by twelve principal components

Component	Eigenvalue	Variation explained (%)	Cumulative variation (%)
PC1	6.760	56.34	56.34
PC2	1.558	12.98	69.32
PC3	1.060	8.83	78.15
PC4	0.711	5.93	84.08
PC5	0.575	4.79	88.87
PC6	0.452	3.77	92.64
PC7	0.351	2.92	95.57
PC8	0.212	1.76	97.33
PC9	0.150	1.25	98.58
PC10	0.124	1.03	99.62
PC11	0.044	0.36	99.98

Component	Eigenvalue	Variation explained (%)	Cumulative variation (%)
PC12	0.002	0.02	100.00

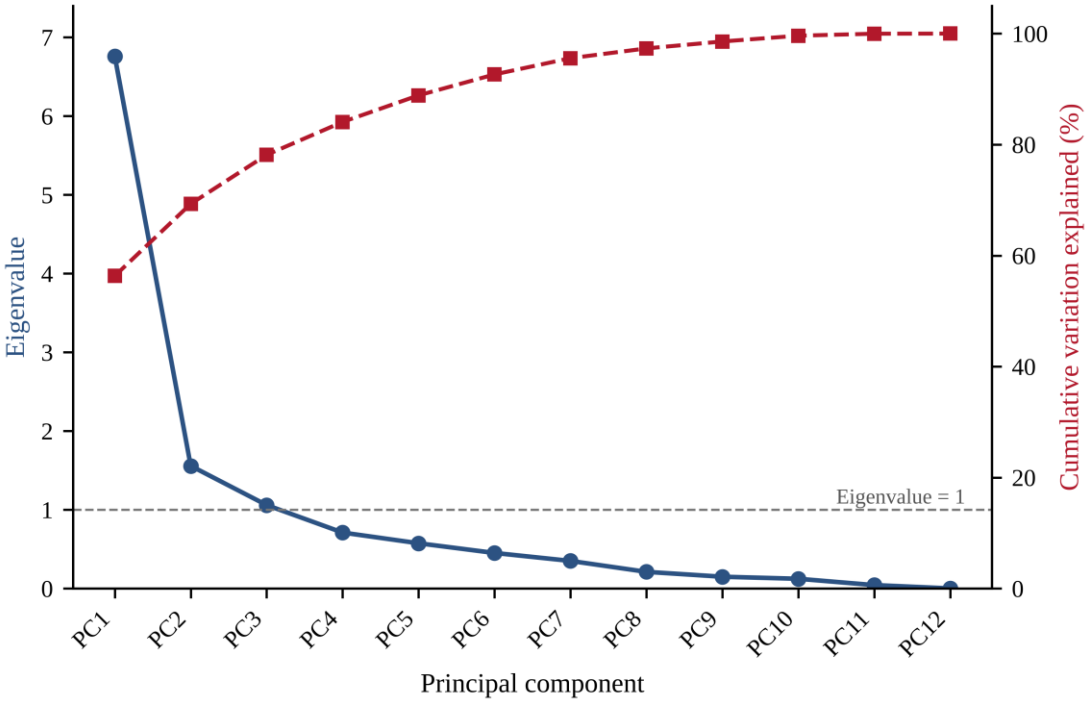

Fig. 8 Scree plot of eigenvalues (left axis) and cumulative variation explained (right axis) by the twelve principal components. The dashed horizontal line marks an eigenvalue of 1.

Examination of trait contributions to the first two principal components (Table 6; Fig. 9) revealed distinct patterns of trait association. In PC1, filled grains per panicle (13.39%), spikelet fertility (12.80%), grain yield per plant (11.37%), panicle weight (9.93%), unfilled grains per panicle (9.67%), effective tillers per plant (9.46%) and test weight (8.59%) were the major contributors. In contrast, days to 50% flowering (2.46%) and grain length (4.24%) showed relatively lower contributions to PC1. For PC2, grain length (20.80%) and plant height (20.60%) were the predominant contributors, followed by days to 50% flowering (12.13%), days to maturity (11.19%) and test weight (10.95%). Filled grains per panicle (0.01%) and spikelet fertility (0.40%) made negligible contributions to this component. Overall, PC1 was primarily associated with yield and reproductive traits, whereas PC2 was more strongly associated with morphological and phenological traits. The prominence of filled grains per panicle in PC1 is consistent with its major contribution to genetic divergence in the Mahalanobis D^2 analysis, while plant height and flowering time also emerged as important contributors to overall divergence. Thus, PCA provided complementary insights into the trait associations underlying genetic variation among the evaluated genotypes.

Table 6 Percentage contribution of individual traits to the first two principal components

Trait	PC1 (%)	PC2 (%)
Filled grains per panicle (FG)	13.39	0.01
Spikelet fertility (SF)	12.80	0.40
Grain yield per plant (GYP)	11.37	5.80
Panicle weight (PW)	9.93	6.78
Unfilled grains per panicle (UFG)	9.67	2.12
Effective tillers (ET)	9.46	4.21
Test weight (TW)	8.59	10.95
Panicle length (PL)	7.03	5.01
Plant height (PH)	6.11	20.60
Days to maturity (DM)	4.94	11.19

Trait	PC1 (%)	PC2 (%)
Grain length (GL)	4.24	20.80
Days to 50% flowering (DF)	2.46	12.13

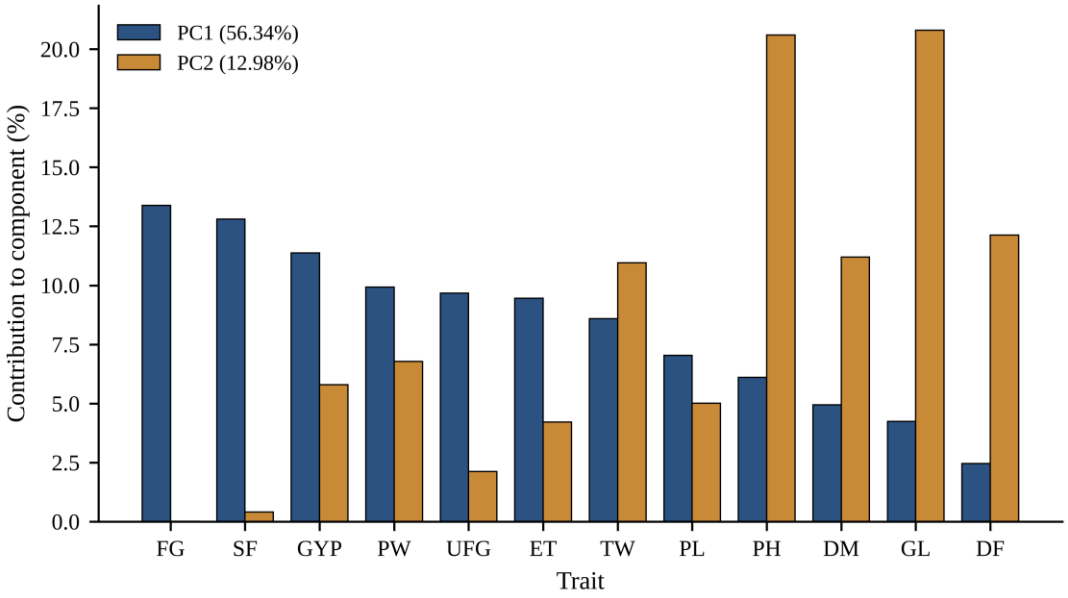

Fig. 9 Percentage contribution of individual traits to the first two principal components, PC1 and PC2.

The PCA biplot of PC1 and PC2, which together explained 69.32% of the total variation, showed that the forty F_4 lines were dispersed across all four quadrants (Fig. 10), reflecting wide variation among the lines. Lines such as L16, L19, L13, L9, L21 and L11 lay on the positive side of PC1, in the direction of the trait vectors for filled grains per panicle, spikelet fertility, grain yield per plant, panicle weight, effective tillers and test weight. In contrast, lines L39, L30, L10, L35, L24 and L4 lay on the negative side of PC1, opposite to these yield-related vectors and toward unfilled grains per panicle. Lines L3, L2, L22, L38, L36 and L23 occupied the upper part of PC2, in the direction of plant height, grain length and the phenological traits.

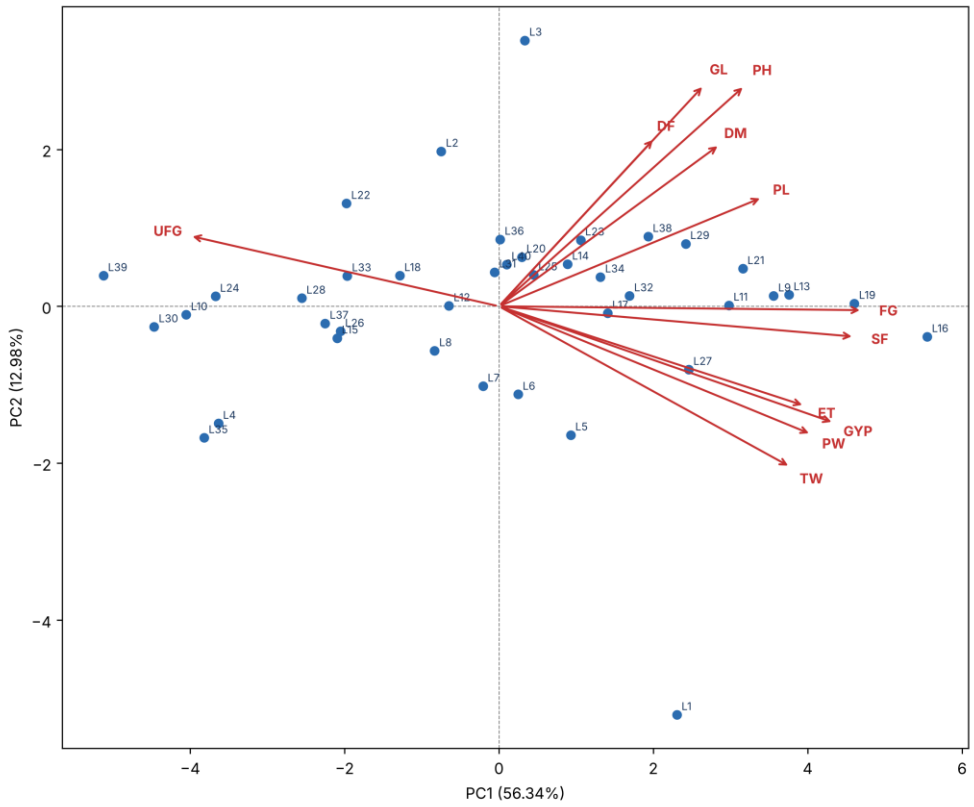

Fig. 10 PCA biplot of the forty F_4 rice lines based on PC1 (56.34%) and PC2 (12.98%). Points represent lines

(L1–L40) and arrows represent trait loadings. DF, days to 50% flowering; DM, days to maturity; PH, plant height; PL, panicle length; ET, effective tillers per plant; GYP, grain yield per plant; PW, panicle weight; FG, filled grains per panicle; UFG, unfilled grains per panicle; TW, test weight; GL, grain length; SF, spikelet fertility.

DISCUSSION

The classification of the forty F_4 lines into thirteen distinct clusters, several of which comprised only two genotypes, revealed considerable genetic diversity within the Kalarata × Pusa 44-derived population. This finding highlights the variability present among the segregating lines and is consistent with the expectation that crosses involving genetically contrasting parents can generate diverse progeny (Khush, 1997). The occurrence of multiple small clusters is comparable to observations from other Mahalanobis D^2-based studies of rice breeding populations and germplasm collections. The number of clusters and their genotype composition may vary depending on the genetic background of the material, the traits evaluated and the generation under study (Bhajipale et al., 2025; Sinha et al., 2019; Thakare et al., 2025).

The thirteen clusters identified in the present study are comparable in number to the ten clusters reported among 66 rice parental lines (Nivedha et al., 2024) and the seven clusters identified among 20 genotypes (Yadav et al., 2024). Similarly, studies involving salt-affected rice germplasm have reported eight to eleven clusters (Arunkumar and Narayanan, 2023; Maurya et al., 2025; Shrivastav et al., 2025; Sharma et al., 2026). The formation of small clusters containing two genotypes has also been reported by Arunkumar and Narayanan (2023), in which five of the eight clusters comprised two genotypes each. These findings demonstrate that cluster number and composition are strongly influenced by the genetic material and traits included in the analysis. The conservation and further characterization of genetically distinct landraces and breeding lines can contribute to the effective utilization of genetic resources in crop improvement programmes (Salgotra and Chauhan, 2023; Bairwa et al., 2024).

The inter-cluster distances ranged widely from 27.59 (Cluster II–III) to 611.96 (Cluster X–XI), while intra-cluster distances were lower than the smallest inter-cluster distance for Clusters I–X but not for Clusters XI, XII and XIII. This pattern indicates greater genetic similarity among genotypes within most clusters and substantial genetic differentiation between certain clusters. Such clustering patterns are consistent with the general principles underlying Tocher's method (Rao, 1952) and have also been reported in recent studies employing Mahalanobis D^2 analysis in rice (Gayathri et al., 2026; Thakare et al., 2025).

Genetically divergent parents are often considered useful candidates for hybridization because their crosses may generate greater genetic variability and provide opportunities for selecting desirable transgressive segregants (Arunachalam, 1981). From the standpoint of genetic divergence, the combinations involving Clusters X and XI (611.96), VIII and X (539.86), VII and X (470.34), and I and XI (452.14) represent potentially useful candidates for further evaluation. However, parental selection should not be based on genetic distance alone. Considering the comparatively favourable cluster means for grain yield and its components in Cluster I, crosses between Cluster I and genetically distant clusters, particularly those showing complementary trait characteristics, may offer opportunities to combine desirable agronomic attributes with broader genetic variability. Similar approaches have been recommended in studies of salt-affected rice germplasm, which emphasize the importance of considering both inter-cluster distance and cluster mean performance when selecting parents for hybridization (Akilan et al., 2023; Shrivastav et al., 2025). Arunkumar and Narayanan (2023) also highlighted the importance of combining high inter-cluster distances with favourable cluster means for yield-related traits during parental selection.

The number of filled grains per panicle emerged as the predominant contributor to genetic divergence (56.28%), followed by plant height (18.08%) and days to 50% flowering (17.18%). This contribution pattern differed from those reported in other rice divergence studies, in which grain yield, biological yield, spikelets per panicle or test weight were identified as the major contributors (Bhajipale et al., 2025; Sinha et al., 2019). Such differences may reflect variations in the parental background, segregating generation and set of traits evaluated in individual studies, indicating that the relative contribution of traits to genetic divergence is population- and study-specific rather than universally consistent.

Studies involving salt-affected rice germplasm further illustrate this variation. Spikelets per panicle was reported as the largest contributor to genetic divergence by Maurya et al. (2025) and Shrivastav et al. (2025), whereas grain length was identified as the leading contributor by Arunkumar and Narayanan (2023). Notably, grain yield per plant contributed very little to genetic divergence in Shrivastav et al. (2025), which is comparable to the negligible contribution observed in the present study (0.13%). Overall, the findings indicate that a relatively small number of reproductive and morphological traits accounted for most of the observed genetic divergence in the evaluated population. The results also highlight that grain yield per plant does not necessarily emerge as a major contributor to genetic divergence, even though it remains an important selection criterion in rice improvement programmes. Therefore, the selection of traits for genetic divergence analysis should consider the biological relevance of individual traits and the objectives of the breeding programme rather than assuming that yield-related traits will invariably dominate the divergence pattern (Gayathri et al., 2026).

The results of principal component analysis provided complementary information on the patterns of genetic

variation identified through Mahalanobis D^2 analysis. In PCA, filled grains per panicle, spikelet fertility and other yield-related traits were prominent contributors to PC1, whereas plant height, grain length and phenological traits contributed more strongly to PC2. Although these patterns were not identical to the trait contribution ranking obtained from D^2 analysis, the prominence of filled grains per panicle and plant height in both approaches highlights their importance in explaining genetic variation within the evaluated population. The use of D^2 clustering in combination with PCA has also been recommended in rice genetic diversity studies to provide complementary information on genetic relationships and trait associations (Gayathri et al., 2026; Sinha et al., 2019).

Similar applications have been reported in other rice populations. Among 66 rice parental lines, four principal components with eigenvalues greater than one accounted for 71.28% of the total variation, while the D^2 clusters were also associated with yield and grain-related traits (Nivedha et al., 2024). Biplot analysis of 48 aromatic rice genotypes has likewise been used to identify genotypes showing divergence for yield-attributing traits (Mondal et al., 2024). Combined PCA and D^2 analyses have also been employed in rice grown under sodic conditions, where the first five principal components explained 62.06% of the total variation among 89 genotypes (Sharma et al., 2026). Similarly, PCA combined with cluster analysis has been used to compare genotype grouping patterns under sodic and normal soil conditions (Akilan et al., 2023).

In the present study, the first three principal components collectively explained 78.15% of the total variation, indicating that a substantial proportion of the observed variability was captured by a relatively small number of components. This suggests that the twelve evaluated traits exhibited patterns of association that allowed the overall variation to be summarized through fewer principal components. PCA therefore provides a useful basis for understanding trait relationships and identifying major sources of variation in the population. Further evaluation of trait combinations across additional populations and environments would be necessary to determine whether a reduced set of traits could be used reliably for the cost-effective characterization of larger rice genetic resource collections (Greenacre et al., 2022; Nivedha et al., 2024).

CONCLUSION

Mahalanobis D^2 analysis grouped the forty F_4 rice lines derived from the cross Kalarata × Pusa 44 into thirteen distinct clusters, indicating considerable genetic diversity within the population. Filled grains per panicle, plant height and days to 50% flowering were identified as the major contributors to total genetic divergence. The maximum inter-cluster distance was observed between Cluster X and Cluster XI, followed by Cluster VIII and Cluster X, indicating substantial genetic differentiation among these groups. These clusters may serve as potential sources of genetic diversity for future hybridization and selection programmes. Principal component analysis provided complementary information on the patterns of genetic variation, with the first three principal components collectively explaining 78.15% of the total variation. The prominence of filled grains per panicle and plant height in both analytical approaches further highlighted their importance in characterizing the evaluated population. Considering both genetic divergence and agronomic performance, crosses involving agronomically superior genotypes from Cluster I and genetically distant genotypes from Clusters X, XI and VIII may be explored to generate broader genetic variability and identify desirable segregants. Such combinations could provide useful breeding material for the development of improved rice cultivars with enhanced yield potential and a broader genetic base.

REFERENCES

1. Akilan M, Jeyaprakash P, Shanmuganathan M, Meena S, Rajanbabu V, Vanniarajan C (2023) Comparative patterns of principal component and cluster analysis under sodicity and normal soil conditions in rice (*Oryza sativa* L.). Electron J Plant Breed 14(3):893–901. https://doi.org/10.37992/2023.1403.101
2. Arunachalam V (1981) Genetic distance in plant breeding. Indian J Genet Plant Breed 41(2):226–236
3. Arunkumar P, Narayanan R (2023) Estimation of genetic diversity for grain yield and its component traits in rice (*Oryza sativa* L.) under coastal saline condition. Mysore J Agric Sci 57(3):88–96
4. Bairwa RK, Yadav MC, Tiwari S, Mahendran A, Ellur RK, Krishnan SG (2024) Aromatic rice (*Oryza sativa* L.) landraces of Indo-Gangetic Plains of India harbor rich genetic diversity. Crop Sci 64(3):1711–1728. https://doi.org/10.1002/csc2.21211
5. Bhajipale VH, Patil SA, Meshram MP, Shende PV, Madke VS (2025) Genetic divergence analysis in advanced breeding lines of rice (*Oryza sativa* L.). Int J Adv Biochem Res 9(7):1692–1696
6. Debsharma SK, Syed MA, Ali MH, Maniruzzaman S, Roy PR, Brestic M, Gaber A, Hossain A (2023) Harnessing on genetic variability and diversity of rice (*Oryza sativa* L.) genotypes based on quantitative and qualitative traits for desirable crossing materials. Genes 14(1):10. https://doi.org/10.3390/genes14010010
7. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow
8. Gayathri D, Pandey AP, Sri MU, Akhilesh G, Gaur RS, Singh B (2026) Multivariate genetic diversity, morphological characterization, correlation and path coefficient analysis of rice (*Oryza sativa* L.) germplasm. Int J Plant Soil Sci 38(8):543–553
9. Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53(3–4):325–338
10. Greenacre M, Groenen PJF, Hastie T, d'Enza AI, Markos A, Tuzhilina E (2022) Principal component

analysis. Nat Rev Methods Primers 2:100. https://doi.org/10.1038/s43586-022-00184-w
11. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6):417–441
12. IARI (2022) Pusa 44: variety information. Indian Agricultural Research Institute, New Delhi
13. IRRI (2013) Standard evaluation system for rice, 5th edn. International Rice Research Institute, Los Baños
14. Islam MZ, Chakrabarty T, Akter N, Khalequzzaman M, Prince MFRK, Pittendrigh BR, Tomita M, Ali MP (2025) Genetic variability, correlation and path coefficient analysis of phenotypic traits and genetic diversity of Aman rice landraces (*Oryza sativa* L.). Sci Rep 15:18606. https://doi.org/10.1038/s41598-025-03547-x
15. Jain SM, Brar DS (eds) (2009) Molecular techniques in crop improvement, 2nd edn. Springer, Dordrecht
16. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
17. Khush GS (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol Biol 35(1–2):25–34
18. Laghari AA, Ahmad A, Memon S, Musavi SAM, Ali A, Kumar A, Guo Q (2025) Genetic diversity in F3 segregating populations of rice (*Oryza sativa* L.) genotypes under salt stress. Front Plant Sci 16:1568859. https://doi.org/10.3389/fpls.2025.1568859
19. Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2(1):49–55
20. Maurya RK, Dwivedi DK, Singh SP, Khan NA, Dixit S, Kumar A, Kumar A, Yadav HK, Kushwaha S (2025) Genetic diversity and trait contribution in rice genotypes using Mahalanobis D^2 statistic. Bioscan 20(2):194–198. https://doi.org/10.63001/tbs.2025.v20.i02.pp194-198
21. Mohanavel V, Yesudhas AS, Sharma A, Ramasamy A, Muthu Arjuna Samy P, Subramanian M, Muthusamy R (2021) Haplotype and diversity analysis of indigenous rice for salinity tolerance in early-stage seedling using simple sequence repeat markers. Biotechnol Rep 31:e00666. https://doi.org/10.1016/j.btre.2021.e00666
22. Mondal S, Pradhan P, Das B, Kumar D, Paramanik B, Yonzone R, Barman R, Saha D, Karforma J, Basak A, Dey P, Seleiman MF (2024) Genetic characterization and diversity analysis of indigenous aromatic rice. Heliyon 10(10):e31232. https://doi.org/10.1016/j.heliyon.2024.e31232
23. Murty BR, Arunachalam V (1966) The nature of genetic divergence in relation to breeding system in crop plants. Indian J Genet Plant Breed 26A:188–198
24. Muthayya S, Sugimoto JD, Montgomery S, Maberly GF (2014) An overview of global rice production, supply, trade, and consumption. Ann N Y Acad Sci 1324(1):7–14
25. Muthazhagu Kuppuraj M, Sangeetha S, Vanniarajan C, Maheswaran M (2025) Maternal effect on the inheritance of pericarp colour and grain dimension in rice (*Oryza sativa* L.). J Genet 104:5
26. Nivedha R, Manonmani S, Kalaimagal T, Raveendran M, Kavitha S (2024) Assessing the genetic diversity of parents for developing hybrids through morphological and molecular markers in rice (*Oryza sativa* L.). Rice 17(1):1–16. https://doi.org/10.1186/s12284-024-00691-2
27. Prakash S, Reddy SS, Chaudhary S, Vimal S, Kumar A (2024) Multivariate analysis in rice (*Oryza sativa* L.) germplasms for yield attributing traits. Plant Sci Today 11(1):64–75. https://doi.org/10.14719/pst.2231
28. Rao CR (1952) Advanced statistical methods in biometric research. John Wiley & Sons, New York
29. Salgotra RK, Chauhan BS (2023) Genetic diversity, conservation, and utilization of plant genetic resources. Genes 14(1):174. https://doi.org/10.3390/genes14010174
30. Sharma P, Gupta AK, Srivastava A, Saini A, Raj P, Srivastava S, Tyagi A, Bharti U, Vimal SC (2026) Genetic variability, heritability sand multivariate analysis of 89 rice (*Oryza sativa* L.) genotypes under sodic soil. Plant Sci Today 13(3). https://doi.org/10.14719/pst.12078
31. Shrivastav SP, Verma OP, Lal K, Singh V, Srivastava K (2025) Genetic divergence analysis in rice (*Oryza sativa* L.) germplasms under sodic soil. Indian J Agric Res 59(2):206–210. https://doi.org/10.18805/IJARe.A-5976
32. Singh RP, Singh KN, Singh AK, Singh VK (2018) Rice-based production systems of the Eastern Gangetic Plains: production constraints and management strategies. Indian J Agron 63(4):427–436
33. Sinha S, Bharati UC, Kumar V, Malik R, Sinha B (2019) Genetic diversity analysis of rice (*Oryza sativa* L.) germplasm through morphological markers. J AgriSearch 6(3):152–157
34. Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438
35. Thakare VV, Lal GM, Bara BM (2025) Assessment of genetic divergence in rice (*Oryza sativa* L.) germplasm using D^2 analysis. Arch Curr Res Int 25(1):267–282
36. USDA (2025) Grain: world markets and trade. United States Department of Agriculture, Foreign Agricultural Service, Washington, DC
37. Yadav SPS, Bhandari S, Ghimire NP, Mehata DK, Majhi SK, Bhattarai S, Shrestha S, Yadav B, Chaudhary P, Bhujel S (2024) Genetic variability, character association, path coefficient, and diversity analysis of rice (*Oryza sativa* L.) genotypes based on agro-morphological traits. Int J Agron 2024:9946332. https://doi.org/10.1155/2024/9946332
38. Zeigler RS, Barclay A (2008) The relevance of rice. Rice 1(1):3–9