

International Journal of Advanced Biochemistry Research



ISSN Print: 2617-4693
ISSN Online: 2617-4707
NAAS Rating (2026): 5.29
IJABR 2026; SP-10(8): 1734-1744
www.biochemjournal.com
Received: 28-06-2026
Accepted: 01-08-2026

Deepthi ST
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Dushyanthakumar BM
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Gireesh C
ICAR- National Institute of Seed
Science and Technology,
Regional Station, Bengaluru,
Karnataka, India

Vishwasgowda C
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Karthik N
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Supreetha K
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Vikas JB
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Sanjana UN
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Hemanth MJ
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Corresponding Author:
Deepthi ST
Department of Genetics and
Plant Breeding, College of
Agriculture, KSNUAHS,
Shivamogga, Karnataka, India

Multivariate analysis reveals key yield-contributing traits and phenotypic diversity among rice (*Oryza Sativa* L.) accessions from the subset of 3K rice genome panel

Deepthi ST, Dushyanthakumar BM, Gireesh C, Vishwasgowda C, Karthik N, Supreetha K, Vikas JB, Sanjana UN and Hemanth MJ

DOI: <https://www.doi.org/10.33545/26174693.2026.v10.i8Ss.9663>

Abstract

Rice (*Oryza sativa* L.) is a major staple crop and improving grain yield is a primary objective of rice breeding. Identifying key yield-contributing traits and exploiting genetically diverse germplasm are essential for accelerating genetic gain. In this study, 208 rice accessions representing a subset of the 3K Rice Genome Panel were evaluated under field conditions for twelve agronomic and yield-related traits. Pearson's correlation, path coefficient analysis, principal component analysis (PCA) and K-means clustering were employed to assess trait associations, phenotypic variation and genetic diversity. Grain yield per plant showed significant positive associations with productive tillers, filled spikelets per panicle, spikelet fertility and thousand-seed weight. Path analysis revealed high positive direct effects of productive tillers, filled spikelets per panicle and thousand-seed weight on grain yield, while other traits contributed indirectly through these components. The first five principal components accounted for 75.34% of total phenotypic variation, with the first two explaining 44.55%. Spikelet number, filled spikelets and spikelet fertility contributed mainly to PC1, whereas productive and total tillers contributed to PC2. K-means clustering grouped the accessions into five distinct phenotypic clusters, confirming substantial diversity. Overall, productive tillers, spikelet number, filled spikelets, spikelet fertility and thousand-seed weight were identified as key traits for yield improvement and strategic parent selection.

Keywords: Rice, 3k rice genome panel, correlation analysis, path coefficient analysis, principal component analysis, k-means clustering

1. Introduction

Rice (*Oryza sativa* L.) is the staple food for more than half of the global population and contributes nearly 20% of the dietary requirements consumed worldwide. Asia accounts for approximately 90% of global rice production and consumption, making rice a cornerstone of food security and rural livelihoods. Increasing demand resulting from rapid population growth, coupled with shrinking arable land, climate change, erratic rainfall, water scarcity and emerging biotic stresses, has intensified the need for the development of high-yielding and climate-resilient rice cultivars (Khush, 2005; Muthayya *et al.*, 2014; FAO, 2023)^[18, 25]. Grain yield is a complex quantitative trait controlled by numerous genes and is strongly influenced by environmental conditions and genotype × environment interactions. Unlike simply inherited traits, grain yield is determined by the cumulative contribution of several component characters including flowering duration, plant height, tillering ability, panicle architecture, spikelet production, grain filling and grain weight. Because these traits interact genetically and physiologically, direct selection based solely on grain yield is often inefficient, particularly during the early generations of breeding programmes (Yoshida, 1981; Ashikari *et al.*, 2005; Alam *et al.*, 2024)^[2, 3, 39]. Understanding the relationships among yield-contributing traits is therefore essential for developing effective selection strategies. Correlation analysis provides valuable information regarding the magnitude and direction of relationships among agronomic traits and enables breeders to identify characters that may be improved simultaneously through indirect selection. However, correlation coefficients alone cannot distinguish whether the observed association results from a direct influence or an

indirect effect mediated through other traits. Path coefficient analysis overcomes this limitation by partitioning correlation coefficients into direct and indirect effects, thereby identifying traits that genuinely influence grain yield.

The availability of genetically diverse germplasm forms the foundation of sustainable crop improvement. The 3K Rice Genome Project, jointly developed through international collaboration, represents one of the largest publicly available collections of rice genetic diversity (Li *et al.*, 2014; Wang *et al.*, 2018) [34]. This panel contains 3024 rice accessions representing the major rice subpopulations, including indica, japonica, aus, aromatic, tropical japonica, temperate japonica and admixed groups collected from diverse agroecological regions. The extensive genomic and phenotypic variation present within this resource offers unprecedented opportunities for identifying favourable alleles associated with yield, stress tolerance, grain quality and adaptation (Zhao *et al.*, 2011; Xu *et al.*, 2017) [37, 40].

Modern breeding programmes increasingly rely on multivariate statistical approaches to evaluate large germplasm collections. Principal component analysis (PCA) is among the most widely used techniques for reducing multidimensional datasets into a few orthogonal components while preserving most of the original information. PCA facilitates the identification of major sources of variation and visualizes relationships among traits and genotypes (Hotelling, 1933; Abdi and Williams, 2010) [1, 11]. Similarly, K-means clustering classifies accessions into phenotypically similar groups, assisting breeders in selecting genetically divergent parents for hybridization and maximizing heterosis (Caldo *et al.*, 1996; Fu, 2015) [5].

Although numerous studies have independently employed correlation analysis, path coefficient analysis, or multivariate statistical techniques in rice, comprehensive investigations integrating these methods using accessions from the 3K Rice Genome Panel under Indian agro-climatic conditions remain limited. Integrating these complementary statistical approaches provides a comprehensive understanding of yield determination and genetic diversity, thereby facilitating the identification of reliable selection criteria and superior parental combinations.

Therefore, the present investigation was undertaken to evaluate rice accessions representing the subset of 3K Rice Genome Panel with the following objectives: (i) to assess phenotypic variability for agronomic and yield-related traits; (ii) to determine the association among traits through Pearson's correlation analysis; (iii) to estimate the direct and indirect effects of component traits on grain yield using path coefficient analysis; (iv) to identify the principal sources of phenotypic variation through principal component analysis; (v) to classify the germplasm into distinct phenotypic groups using K-means clustering and (vi) to identify promising parental lines and key selection criteria for future rice breeding programmes.

2. Materials and Methods

2.1 Plant materials and experimental site

The experimental material comprised 208 rice (*Oryza sativa* L.) accessions, including 203 genetically diverse accessions representing a subset of the 3K Rice Genome Panel and five widely cultivated check varieties (Brown gora, Sabitha, Sahbhagidhan, IR 64 and DRR Dhan 44). The experiment was conducted during the Kharif 2024 season at the ICAR-

National Institute of Seed Science and Technology, Regional Station, Bengaluru. All recommended agronomic and plant protection measures were followed uniformly throughout the experimental period to ensure optimal crop growth and minimize environmental variation among the evaluated accessions.

2.2 Experimental design and phenotypic evaluation

The experiment was laid out in an Augmented Randomized Complete Block Design (Augmented RCBD), which is well suited for the evaluation of large germplasm collections where seed availability is limited. Five standard check varieties were replicated across experimental blocks, while each test accession was grown in a single plot. Observations were recorded from five randomly selected competitive plants for each accession following the Standard Evaluation System (SES) for Rice developed by the International Rice Research Institute (IRRI, 2014). Mean values of the five sampled plants were used for subsequent statistical analyses. Twelve agronomic and yield-related traits were recorded, including days to 50% flowering (DFF), plant height (PH), number of tillers per plant (NTP), number of productive tillers per plant (NPT), panicle length (PL), flag leaf length (FLL), flag leaf width (FLW), number of spikelets per panicle (NSP), number of filled spikelets per panicle (NFSP), spikelet fertility (SF), thousand seed weight (TSW) and grain yield per plant (GY). These traits were selected because of their direct or indirect influence on grain yield and their importance in rice breeding and genetic improvement programmes.

2.3 Statistical analysis

Descriptive statistics, including the mean, range, standard deviation and coefficient of variation, were computed for all measured traits to assess the extent of phenotypic variability among the evaluated accessions. The observed variation was used to determine the breeding potential of the germplasm and to identify traits exhibiting substantial genetic diversity. Pearson's correlation coefficients were estimated to determine the magnitude and direction of associations among grain yield and its component traits (Pearson, 1901) [27]. Correlation analysis was performed to identify characters that are closely associated with grain yield and could therefore serve as effective indirect selection criteria. The significance of correlation coefficients was tested at the 5% and 1% probability levels.

To distinguish the direct and indirect contributions of individual traits toward grain yield, path coefficient analysis was carried out by considering grain yield per plant as the dependent variable. The analysis followed the procedure described by Dewey and Lu (1959) [6], based on the path coefficient methodology originally proposed by Wright (1921) [35]. This approach partitions the observed correlation coefficients into direct and indirect effects, thereby facilitating the identification of traits exerting genuine causal influence on grain yield. Traits exhibiting both high positive direct effects and significant positive correlations with grain yield were considered reliable selection criteria for improving yield.

Principal component analysis (PCA) was performed using standardized trait values to summarize the multidimensional variation present among the rice accessions and to identify the major traits contributing to phenotypic diversity. Eigenvalues, the percentage of variance explained,

cumulative variance and variable loadings were used to interpret the principal components. The relationships among traits and genotypes were visualized using scree plots, variable contribution plots, correlation circles and PCA biplots.

To further characterize the phenotypic diversity within the germplasm, K-means clustering was employed to classify the accessions into distinct groups based on standardized trait values. The optimum number of clusters was determined using the elbow method by minimizing within-cluster variation while maximizing between-cluster separation. The resulting clusters were visualized through PCA-based scatter plots to illustrate the distribution and divergence of the evaluated accessions.

All statistical analyses were performed using R software (version 4.5.1). Principal component analysis and cluster analysis were conducted using the FactoMineR and factoextra packages (Le and Husson, 2008; Kassambara and Mundt, 2020) [17], while correlation and path coefficient analyses were performed using appropriate R packages for quantitative genetic analysis.

3. Results and Discussion

3.1 Phenotypic variability

Substantial phenotypic variation was observed among the evaluated accessions for all measured traits, indicating the presence of broad genetic diversity within the study material. The wide range recorded for flowering time, plant height, tillering ability, panicle architecture, spikelet production, grain filling, seed weight and grain yield demonstrates the quantitative inheritance of these characters and highlights the availability of exploitable genetic variation for crop improvement (Table 1).

Days to 50% flowering varied from 78.17 to 110.93 days with a mean of 93.75 ± 8.07 days, while plant height ranged from 89.22 to 141.39 cm, averaging 115.77 ± 10.28 cm. The mean number of total tillers and productive tillers per plant was 17.89 and 15.22, respectively, reflecting relatively low variation ($CV < 5\%$), indicating stable expression of these traits across the evaluated germplasm.

Among the yield components, number of filled spikelets per panicle (21.65%), number of spikelets per panicle (18.19%), flag leaf width (17.13%) and flag leaf length (17.07%) exhibited comparatively higher coefficients of variation, suggesting greater opportunity for genetic improvement through selection. Conversely, spikelet fertility showed the lowest coefficient of variation (3.26%), indicating relatively stable reproductive performance across accessions.

The observed variability demonstrates the richness of the evaluated subset of the 3K Rice Genome Panel and confirms its suitability for identifying superior donor genotypes possessing favourable alleles for grain yield improvement. The wide phenotypic variation observed in the present study is consistent with previous investigations on diverse rice germplasm, which reported considerable diversity for flowering behaviour, plant architecture, panicle traits, grain filling and yield components. Such variability forms the foundation for effective parent selection and sustained genetic gain in rice breeding programmes.

Similar observations have been reported in diverse rice germplasm collections by Li *et al.* (2014), Wang *et al.* (2018) [34], Kumar *et al.* (2024) [19], Shajan *et al.* (2024) [30] and Sagar *et al.* (2025) [28], emphasizing that broad

phenotypic variability forms the foundation of effective parent selection and long-term genetic gain.

3.2 Correlation analysis

The Pearson's correlation coefficients among the twelve agronomic and yield-related traits are presented in Table 2. Correlation analysis is an important biometrical tool in plant breeding because it reveals the magnitude and direction of relationships among traits, thereby facilitating indirect selection for complex quantitative characters such as grain yield. Since grain yield is governed by numerous interrelated components, understanding these associations enables breeders to identify traits that can be simultaneously improved through selection.

The results revealed considerable variation in the magnitude and direction of trait associations, indicating the complex nature of yield determination in rice. Grain yield per plant exhibited significant positive correlations with several reproductive and yield component traits, while a few vegetative traits showed weak or negative associations.

Among all traits, number of spikelets per panicle ($r = 0.202$, $P \leq 0.01$) showed one of the strongest positive correlations with grain yield, followed closely by number of filled spikelets per panicle ($r = 0.200$, $P \leq 0.01$). These results indicate that increased sink capacity contributes substantially to grain production. Accessions producing larger panicles with a greater number of spikelets generally accumulated higher grain yield because of increased grain-bearing potential. Similarly, spikelet fertility ($r = 0.182$, $P \leq 0.01$) was positively associated with grain yield, demonstrating that efficient grain filling is an important determinant of productivity. High spikelet fertility ensures that a greater proportion of spikelets develop into mature grains, thereby increasing final yield. Thousand-seed weight ($r = 0.152$, $P \leq 0.01$) also exhibited a significant positive association with grain yield, suggesting that heavier grains contribute directly to enhanced productivity. Improvement of grain weight is therefore expected to complement gains achieved through increased spikelet production.

Plant height showed a weak but significant positive association with grain yield ($r = 0.111$, $P \leq 0.05$), indicating that moderately taller plants may possess improved biomass production and assimilate availability for grain filling. However, excessive plant height is generally undesirable because of increased lodging susceptibility. In contrast, days to 50% flowering exhibited a significant negative correlation with grain yield ($r = -0.194$, $P \leq 0.01$). This finding suggests that earlier flowering genotypes tended to produce relatively higher grain yield under the experimental conditions, possibly because of better synchronization with favourable environmental conditions during reproductive development. Flag leaf length also showed a significant negative association with grain yield ($r = -0.171$, $P \leq 0.01$). Although the flag leaf is considered an important photosynthetic organ, increased leaf length alone may not necessarily translate into greater grain production, indicating that assimilate partitioning efficiency is more important than leaf size.

The correlation matrix further revealed extremely strong positive associations among the reproductive traits. The number of spikelets per panicle and the number of filled spikelets per panicle were almost perfectly correlated ($r = 1.000$), reflecting the close biological relationship between these two traits. Likewise, spikelet fertility showed strong

positive correlations with both spikelet number ($r = 0.878$) and filled spikelets ($r = 0.878$), indicating coordinated regulation of reproductive development and grain filling.

Overall, the correlation analysis demonstrated that grain yield was primarily associated with reproductive sink traits rather than vegetative growth traits. The positive relationships among spikelet number, filled spikelets, spikelet fertility and grain weight indicate that these characters should be considered collectively during selection for yield improvement.

The observed relationships agree with the findings of Yoshida (1981) [39], Ashikari *et al.* (2005) [3], Xing and Zhang (2010) [36], Huang *et al.* (2013), Wang *et al.* (2018) [34], Bharamappanavara *et al.* (2023) [4] and Hani and Thouseem (2024) [10], who similarly reported that spikelet production, grain filling efficiency and grain weight are the principal contributors to grain yield in rice. These studies collectively emphasize that simultaneous improvement of these component traits can substantially enhance yield potential.

However, correlation analysis measures only the degree of association among traits and does not distinguish between direct and indirect effects. Therefore, path coefficient analysis was undertaken to partition these relationships and identify the traits exerting genuine causal effects on grain yield.

3.3 Path coefficient analysis

Path coefficient analysis was performed to partition the observed correlation coefficients into direct and indirect effects, thereby identifying the traits that exert a genuine influence on grain yield per plant. While correlation analysis measures the overall association between two traits, it does not distinguish whether this relationship is due to a direct causal effect or is mediated through other correlated characters. Therefore, path coefficient analysis provides a more reliable basis for identifying effective selection criteria in crop improvement programmes (Table 3).

The results indicated that the contribution of individual agronomic traits to grain yield varied considerably. Several traits exhibited substantial direct effects on grain yield, whereas others contributed mainly through indirect effects mediated by closely associated characters.

Among the evaluated traits, number of productive tillers per plant (NPP) exhibited one of the highest positive direct effects on grain yield. This result indicates that productive tillers are a primary determinant of grain production because each productive tiller contributes directly to panicle formation and ultimately to grain yield. The high direct effect of productive tillers agrees with its positive association with grain yield observed in the correlation analysis and emphasizes its importance as a reliable selection criterion in breeding programmes. Similarly, number of filled spikelets per panicle (NFSP) exerted a strong positive direct influence on grain yield. Filled spikelets represent the effective sink capacity of a plant and directly determine the number of harvested grains. The strong direct effect observed in the present investigation indicates that improving grain filling efficiency is likely to produce substantial gains in grain yield.

Thousand-seed weight (TSW) also exhibited an appreciable positive direct effect on grain yield. Grain weight reflects grain size and density, both of which contribute directly to total grain production. Selection for increased thousand-seed

weight therefore provides an effective means of improving grain yield, particularly when combined with enhanced spikelet production and grain filling. Although number of spikelets per panicle (NSP) showed a strong positive correlation with grain yield, its direct effect was comparatively lower because a considerable proportion of its influence was transmitted indirectly through the number of filled spikelets per panicle. This indicates that producing a larger number of spikelets alone is insufficient for improving yield unless efficient grain filling is achieved. Thus, the effectiveness of increased spikelet production depends largely on the plant's ability to convert spikelets into fully developed grains. Spikelet fertility (SF) also contributed positively to grain yield primarily through indirect effects mediated by filled spikelets. Since spikelet fertility determines the proportion of spikelets that develop into mature grains, improvement in fertility indirectly enhances grain yield by increasing grain filling efficiency. Plant height exhibited only a small positive direct effect on grain yield despite showing significant positive correlations with several reproductive traits. This suggests that the contribution of plant height to yield is largely indirect and mediated through improvements in panicle architecture and reproductive development. Consequently, plant height alone may not serve as a reliable selection criterion unless accompanied by favourable reproductive characteristics. Panicle length similarly exerted a relatively low direct effect on grain yield. Although longer panicles generally accommodate more spikelets, the ultimate contribution of panicle length depends on spikelet fertility and grain filling rather than panicle size itself. Flag leaf length and flag leaf width exhibited negligible direct effects on grain yield. These traits may influence yield indirectly through enhanced photosynthetic capacity and assimilate production; however, their independent contribution to grain production appears to be relatively limited under the present experimental conditions. Days to 50% flowering showed a small negative direct effect on grain yield, indicating that delayed flowering may reduce productivity under the environmental conditions prevailing during the study. Early flowering genotypes may have benefited from better synchronization with favourable climatic conditions during grain filling, thereby achieving comparatively higher yields.

Overall, the present results demonstrate that productive tillers per plant, filled spikelets per panicle, spikelet fertility and thousand-seed weight constitute the principal determinants of grain yield among the evaluated rice accessions. Simultaneous improvement of these characters is therefore expected to maximize genetic gain and enhance the efficiency of indirect selection in rice breeding programmes.

These findings are in close agreement with those reported by Dewey and Lu (1959) [6], Ashikari *et al.* (2005) [3], Xing and Zhang (2010) [36], Huang *et al.* (2013), Wang *et al.* (2018) [34], Kumar *et al.* (2024) [19], Salunkhe *et al.* (2024) [29] and Yashwanth *et al.* (2025) [38], who also identified productive tillers, grain filling, spikelet fertility and grain weight as the most influential determinants of grain yield in rice.

3.4 Principal component analysis (PCA)

Principal component analysis (PCA) was employed to summarize the multidimensional phenotypic variation present among the 208 rice accessions and to identify the

major agronomic traits responsible for the observed diversity. PCA transforms correlated variables into a smaller number of orthogonal principal components while retaining most of the original variation. This approach facilitates the identification of key yield-related traits and provides a graphical representation of the relationships among traits and genotypes.

The eigenvalue analysis revealed that the first five principal components possessed eigenvalues greater than one, satisfying the Kaiser criterion and together explaining 75.34% of the total phenotypic variation observed among the evaluated accessions (Table 4). This indicates that these five components adequately summarize the multidimensional variation present in the germplasm.

The first principal component (PC1) accounted for 28.35% of the total variation with an eigenvalue of 3.40, indicating that this component captured the largest proportion of phenotypic diversity. The second principal component (PC2) explained 16.20% of the variation, while the third, fourth and fifth principal components accounted for 12.20%, 10.18% and 8.40%, respectively. Together, PC1 and PC2 explained 44.55% of the total phenotypic variation, demonstrating that a substantial proportion of the observed diversity can be effectively visualized using a two-dimensional PCA biplot.

Beyond the fifth principal component, eigenvalues declined sharply and each additional component explained less than 8% of the total variation, suggesting diminishing biological significance. The concentration of variation within the first few principal components indicates that the phenotypic diversity of the evaluated rice accessions is primarily governed by a limited number of agronomic traits, making PCA an effective approach for summarizing the overall variability of the germplasm.

Principal component loadings

Principal component loadings indicate the contribution of individual traits to each principal component. Traits with larger absolute loading values contribute more strongly to the variation represented by a particular component.

The first principal component (PC1) was predominantly influenced by number of spikelets per panicle (0.900), number of filled spikelets per panicle (0.900) and spikelet fertility (0.846), followed by plant height (0.577). These results demonstrate that reproductive sink traits constitute the major source of phenotypic variation among the evaluated rice accessions (Table 5). The high positive loadings of spikelet number and filled spikelets indicate that variation in reproductive capacity contributes substantially to differences in grain yield potential. Since these traits are directly involved in grain production, they represent important targets for indirect selection (Figure 1).

The second principal component (PC2) was characterized mainly by number of productive tillers per plant (0.860) and total tillers per plant (0.859). These findings indicate that tillering ability represents the second most important source of variation within the germplasm.

Plant height and panicle length exhibited moderate positive loadings in PC1, reflecting their contribution to plant architecture and reproductive development. Conversely, days to 50% flowering showed moderate positive loading in PC3, indicating that flowering behaviour contributes independently to phenotypic diversity rather than directly influencing the first two principal components. Flag leaf

length and flag leaf width contributed primarily to later principal components, suggesting that although these traits exhibit measurable variation, their influence on overall diversity is less pronounced compared with reproductive and tillering traits.

The predominance of reproductive sink traits in PC1 and tillering traits in PC2 agrees closely with the findings of the correlation and path coefficient analyses, reinforcing the conclusion that these characters represent the principal determinants of grain yield in the evaluated germplasm. The high $\cos^2$ values of reproductive traits indicate that these variables explain a substantial proportion of the overall phenotypic variation within the evaluated germplasm. In contrast, flag leaf dimensions and flowering time showed comparatively lower contributions to the first two principal components, indicating that these traits account for variation represented by later components (Figure 2).

The agreement among the eigenvalue analysis, loading matrix, scree plot, correlation circle and $\cos^2$ plot demonstrates the robustness of the PCA results and confirms that reproductive sink traits and tillering ability are the dominant contributors to phenotypic diversity (Figure 3). The PCA results obtained in the present investigation are in agreement with those reported by Jolliffe (2002) ^[15], Johnson and Wichern (2007) ^[14], Sinha and Mishra (2013) ^[32], Nachimuthu *et al.* (2014) ^[26], Wang *et al.* (2018) ^[34], Singh *et al.* (2021) ^[31], Bharamappanavara *et al.* (2023) ^[4] and Lakshmeesha *et al.* (2024) ^[20], who similarly identified spikelet number, grain filling, tillering ability and grain weight as the major sources of variation in diverse rice germplasm.

3.5 K-means cluster analysis

K-means clustering was employed to classify the 208 rice accessions into phenotypically distinct groups based on twelve agronomic and yield-related traits. Cluster analysis is an effective multivariate technique that groups genetically similar accessions while maximizing variation among clusters. Such classification provides valuable information for identifying genetically diverse parents that can be utilized in hybridization programmes to exploit heterosis and generate transgressive segregants.

The optimum number of clusters was determined using the elbow method, which minimizes within-cluster variation while maximizing between-cluster divergence. The PCA score plot demonstrated that the evaluated rice accessions were distributed into five distinct phenotypic clusters, indicating the existence of substantial diversity within the evaluated subset of the 3K Rice Genome Panel (Figure 4).

The clusters were clearly distinguishable in the PCA space, although slight overlap among neighbouring clusters suggested gradual phenotypic transitions among certain accessions. Such overlap is expected in quantitative traits because of continuous variation and the influence of polygenic inheritance. The formation of five well-defined clusters indicates that the evaluated germplasm possesses broad phenotypic diversity for flowering behaviour, plant architecture, tillering ability, panicle characteristics, grain filling and grain yield. Accessions within the same cluster exhibited relatively similar phenotypic performance, whereas those belonging to different clusters were considerably divergent.

Among the evaluated traits, reproductive characters such as number of spikelets per panicle, number of filled spikelets

per panicle, spikelet fertility and thousand-seed weight contributed substantially to the observed clustering pattern. These findings are consistent with the principal component analysis, where these same traits contributed most strongly to the first principal component. Similarly, total tillers per plant and productive tillers per plant largely influenced the separation of accessions along the second principal component, emphasizing the importance of tillering ability in determining phenotypic diversity.

The considerable dispersion of accessions across the PCA space suggests that the evaluated germplasm represents multiple genetic backgrounds originating from diverse geographical regions. Such diversity is highly desirable for breeding programmes because crosses involving genetically distant parents generally produce greater heterosis and increase the probability of recovering superior recombinant lines in segregating generations.

The clustering results therefore provide a valuable framework for parent selection. Crosses involving accessions belonging to different clusters are expected to generate wider genetic recombination and greater

phenotypic variation than crosses between closely related genotypes. The observed clustering pattern also demonstrates the usefulness of multivariate statistical methods in characterizing large germplasm collections. By reducing complex phenotypic information into a few distinct groups, K-means clustering facilitates efficient germplasm management and assists breeders in selecting parents possessing complementary agronomic characteristics.

The present findings agree with previous reports by MacQueen (1967) ^[23], Mohammadi and Prasanna (2003) ^[24], Wang *et al.* (2018), Hair *et al.* (2019), Bharamappanavara *et al.* (2023) ^[4] and Karoda *et al.* (2024) ^[16], who similarly reported substantial phenotypic diversity among rice germplasm using multivariate clustering approaches.

The close agreement between K-means clustering and principal component analysis further validates the robustness of the present investigation. Both approaches consistently demonstrated that reproductive sink traits and tillering ability represent the major determinants of phenotypic diversity among the evaluated rice accessions.

Table 1: Descriptive statistics of twelve agronomic and yield-related traits

Trait	Mean	SD	Minimum	Maximum	CV (%)
DFF	93.75	8.07	78.17	110.93	8.61
PH	115.77	10.28	89.22	141.39	8.88
NTP	17.89	0.64	16.55	19.32	3.6
NPP	15.22	0.64	13.77	16.58	4.24
PL	22.15	0.92	19.95	24.32	4.15
FLL	26.18	4.47	15.59	37.29	17.07
FLW	1.69	0.29	1.22	2.3	17.13
NSP	99.05	18.01	58.23	143.84	18.19
NFSP	83.42	18.06	42.74	128.63	21.65
SF	84.46	2.75	77.81	90.72	3.26
TSW	22.71	2.68	15.98	28.25	11.82
GYP	17.27	1.21	14.51	20.31	6.98

Abbreviations: DFF = Days to 50% flowering; PH = Plant height (cm); NTP = Number of total tillers per plant; NPP = Number of productive tillers per plant; PL = Panicle length (cm); FLL = Flag leaf length (cm); FLW = Flag leaf width (cm); NSP = Number of spikelets per panicle; NFSP = Number of filled spikelets per panicle; SF = Spikelet fertility (%); TSW = Thousand-seed weight (g); GYP = Grain yield per plant (g).

Table 2: Pearson's correlation coefficients among twelve agronomic and yield-related traits

Trait	DFF	PH	NTP	NPP	PL	FLL	FLW	NSP	NFSP	SF	TSW	GYP
DFF	1.000**	-0.059	0.154**	0.151**	-0.008	0.199**	-0.104*	-0.115*	-0.115*	-0.028	0.150**	-0.194**
PH	-0.059	1.000**	-0.260**	-0.270**	0.447**	0.193**	-0.082	0.320**	0.321**	0.334**	0.250**	0.111*
NTP	0.154**	-0.260**	1.000**	0.993**	-0.146**	0.015	-0.101*	-0.151**	-0.152**	-0.104*	-0.056	-0.042
NPP	0.151**	-0.270**	0.993**	1.000**	-0.159**	0.003	-0.096	-0.158**	-0.159**	-0.115*	-0.082	0.023
PL	-0.008	0.447**	-0.146**	-0.159**	1.000**	0.081	-0.018	0.244**	0.244**	0.207**	0.032	0.054
FLL	0.199**	0.193**	0.015	0.003	0.081	1.000**	0.021	0.051	0.051	0.048	0.036	-0.171**
FLW	-0.104*	-0.082	-0.101*	-0.096	-0.018	0.021	1.000**	0.024	0.024	-0.001	-0.089	-0.019
NSP	-0.115*	0.320**	-0.151**	-0.158**	0.244**	0.051	0.024	1.000**	1.000**	0.878**	0.056	0.202**
NFSP	-0.115*	0.321**	-0.152**	-0.159**	0.244**	0.051	0.024	1.000**	1.000**	0.878**	0.057	0.200**
SF	-0.028	0.334**	-0.104*	-0.115*	0.207**	0.048	-0.001	0.878**	0.878**	1.000**	0.082	0.182**
TSW	0.150**	0.250**	-0.056	-0.082	0.032	0.036	-0.089	0.056	0.057	0.082	1.000**	0.152**
GYP	-0.194**	0.111*	-0.042	0.023	0.054	-0.171**	-0.019	0.202**	0.200**	0.182**	0.152**	1.000**

Abbreviations: DFF = Days to 50% flowering; PH = Plant height (cm); NTP = Number of total tillers per plant; NPP = Number of productive tillers per plant; PL = Panicle length (cm); FLL = Flag leaf length (cm); FLW = Flag leaf width (cm); NSP = Number of spikelets per panicle; NFSP = Number of filled spikelets per panicle; SF = Spikelet fertility (%); TSW = Thousand-seed weight (g); GYP = Grain yield per plant (g).

Table 3: Direct and indirect effects of agronomic and yield-related traits on grain yield per plant

	DFF	PH	NTP	NPP	PL	FLL	FLW	NSP	NFSP	SF	TSW
DFF	-0.215	0.000	-0.713	0.707	-0.001	-0.020	0.004	-2.589	2.590	-0.006	0.047
PH	0.013	0.001	1.205	-1.267	0.037	-0.019	0.003	7.188	-7.195	0.066	0.079
NTP	-0.033	0.000	-4.635	4.662	-0.012	-0.001	0.004	-3.393	3.405	-0.020	-0.018
NPP	-0.032	0.000	-4.606	4.697	-0.013	0.000	0.004	-3.539	3.562	-0.023	-0.026
PL	0.002	0.000	0.679	-0.748	0.083	-0.008	0.001	5.479	-5.485	0.041	0.010
FLL	-0.043	0.000	-0.068	0.013	0.007	-0.098	-0.001	1.145	-1.147	0.009	0.011
FLW	0.022	0.000	0.473	-0.445	-0.001	-0.002	-0.041	0.543	-0.539	0.000	-0.028
NSP	0.025	0.000	0.701	-0.740	0.020	-0.005	-0.001	22.454	-22.443	0.173	0.018
NFSP	0.025	0.000	0.703	-0.745	0.020	-0.005	-0.001	22.454	-22.443	0.173	0.018
SF	0.006	0.000	0.480	-0.541	0.017	-0.005	0.000	19.703	-19.702	0.197	0.026
TSW	-0.032	0.000	0.256	-0.384	0.003	-0.004	0.004	1.253	-1.277	0.016	0.317

Abbreviations: DFF = Days to 50% flowering; PH = Plant height (cm); NTP = Number of total tillers per plant; NPP = Number of productive tillers per plant; PL = Panicle length (cm); FLL = Flag leaf length (cm); FLW = Flag leaf width (cm); NSP = Number of spikelets per panicle; NFSP = Number of filled spikelets per panicle; SF = Spikelet fertility (%); TSW = Thousand-seed weight (g); GYP = Grain yield per plant (g).

Table 4: Eigenvalues, percentage of explained variance and cumulative variance of principal components obtained from principal component analysis of twelve agronomic traits

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.4023	28.3528	28.3528
comp 2	1.9442	16.2018	44.5546
comp 3	1.4644	12.2029	56.7576
comp 4	1.2220	10.1831	66.9407
comp 5	1.0083	8.4026	75.3433
comp 6	0.9270	7.7253	83.0687
comp 7	0.7634	6.3616	89.4303
comp 8	0.6644	5.5366	94.9669
comp 9	0.4485	3.7375	98.7044
comp 10	0.1515	1.2628	99.9672
comp 11	0.0039	0.0328	100.0000
comp 12	0.0000	0.0000	100.0000

Table 5: Principal component loadings of twelve agronomic and yield-related traits contributing to the first ten principal components

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
DFF	-0.175	0.194	0.623	-0.134	-0.411	-0.024	0.420	0.394	0.151	0.034
PH	0.577	-0.151	0.435	0.240	0.339	0.078	-0.081	-0.169	0.489	0.033
NTP	-0.446	0.859	0.049	0.046	0.177	0.090	0.015	-0.125	0.018	0.001
NPP	-0.452	0.860	0.015	0.069	0.182	0.103	0.001	-0.061	0.036	0.012
PL	0.432	-0.087	0.335	0.092	0.652	-0.017	0.386	0.105	-0.314	-0.023
FLL	0.082	0.046	0.615	-0.401	0.063	0.238	-0.581	0.170	-0.157	-0.014
FLW	0.031	-0.168	-0.291	-0.437	0.059	0.790	0.246	-0.039	0.077	-0.001
NSP	0.900	0.332	-0.126	-0.148	-0.112	-0.051	0.004	-0.025	-0.053	0.153
NFSP	0.900	0.331	-0.125	-0.148	-0.112	-0.052	0.004	-0.027	-0.054	0.152
SF	0.846	0.363	-0.059	-0.124	-0.155	-0.054	0.037	-0.019	0.048	-0.319
TSW	0.169	-0.029	0.402	0.573	-0.420	0.360	0.047	-0.354	-0.217	-0.002
GYP	0.274	0.122	-0.336	0.625	0.034	0.288	-0.171	0.541	0.022	-0.002

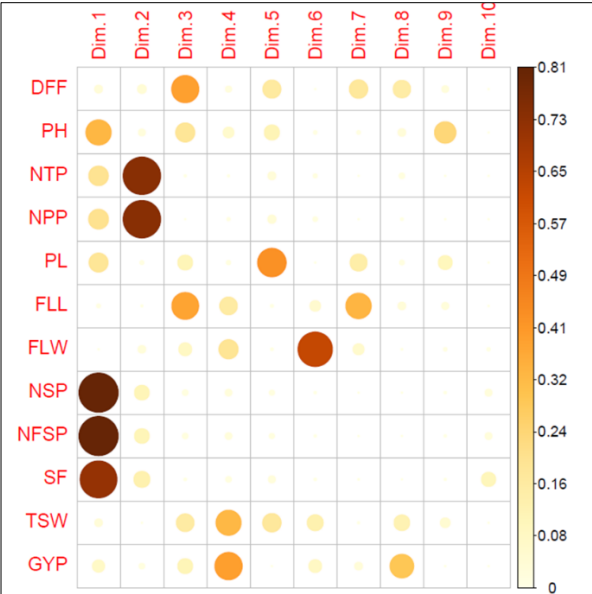


Fig 1: Correlation circle showing the contribution and association of twelve agronomic and yield-related traits to the first two principal components (PC1 and PC2)

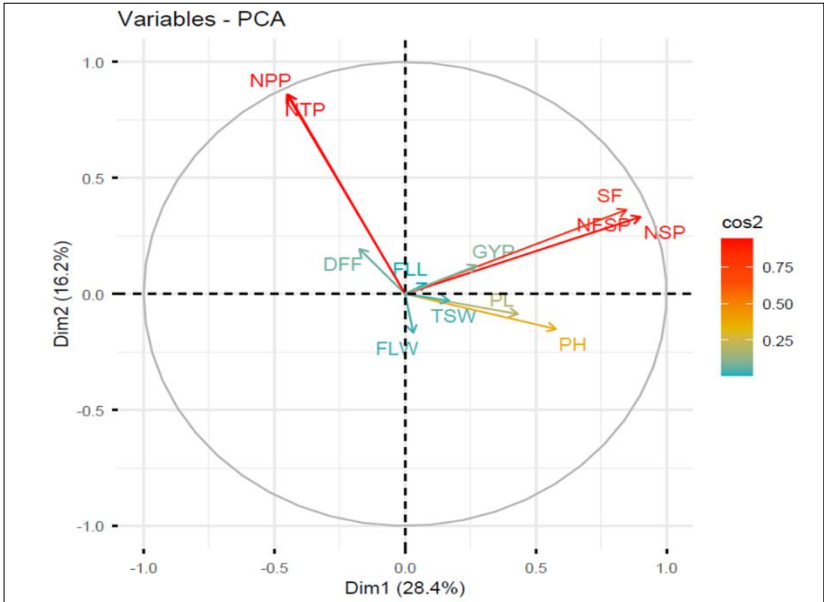


Fig 2: Contribution ($\cos^2$) of twelve agronomic and yield-related traits across the first ten principal components

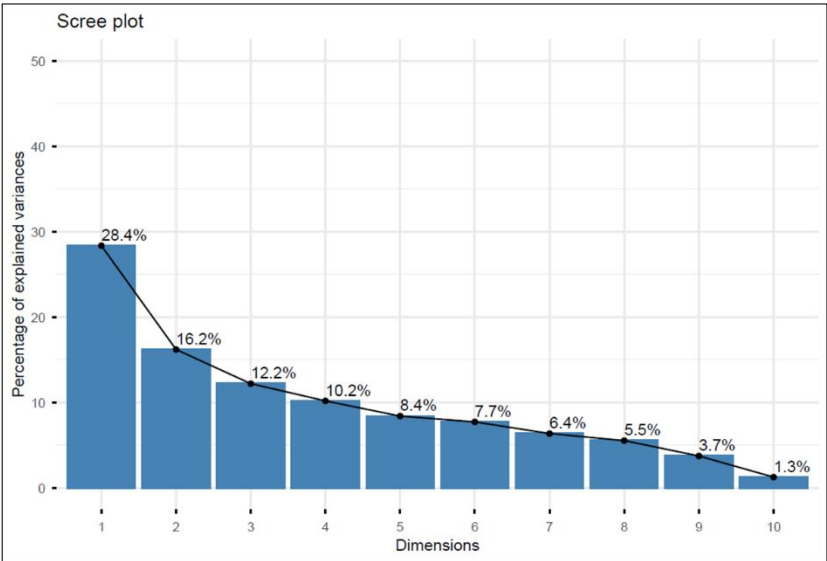


Fig 3: Scree plot showing the percentage of variance explained by each principal component

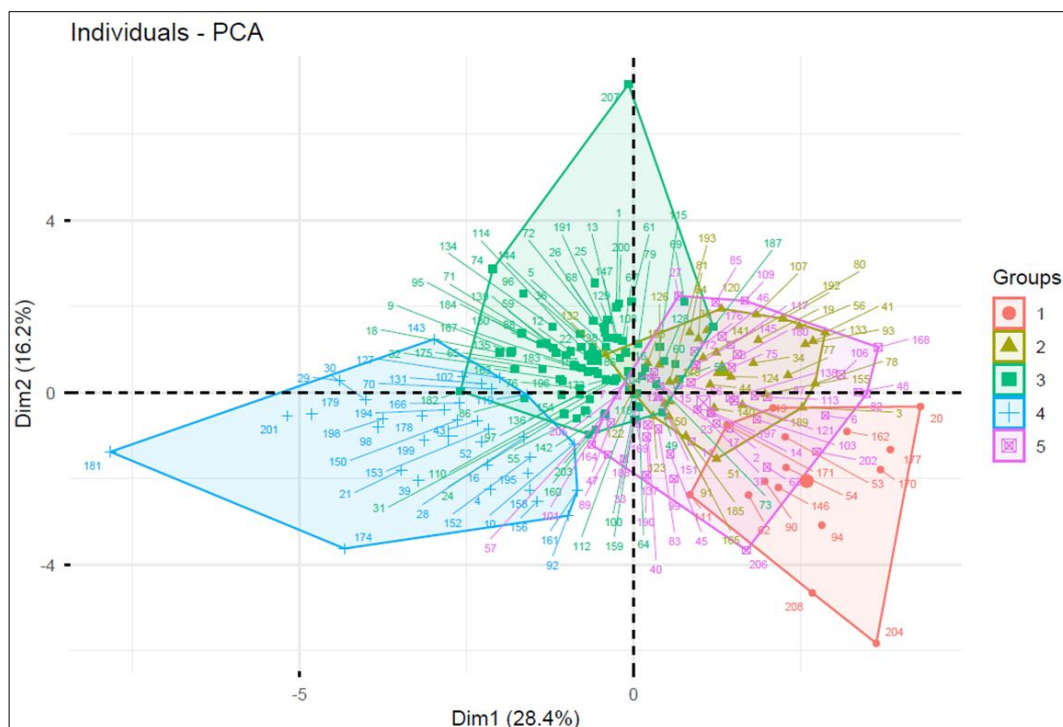


Fig 4: Principal component analysis (PCA) score plot showing the distribution of 208 rice accessions classified into five phenotypic clusters using K-means clustering

4. Conclusion

The present investigation revealed substantial phenotypic diversity among 208 rice (*Oryza sativa* L.) accessions representing a subset of the 3K Rice Genome Panel for twelve agronomic and yield-related traits. The broad variation observed for flowering time, plant height, tillering ability, panicle architecture, spikelet production, grain filling, grain weight and grain yield demonstrates the immense breeding potential of this diverse germplasm.

Correlation analysis identified number of spikelets per panicle, number of filled spikelets per panicle, spikelet fertility and thousand-seed weight as traits positively associated with grain yield, indicating that improvement of these characters would enhance productivity. The significant negative association between days to 50% flowering and grain yield further suggests that early-flowering accessions may possess an adaptive advantage under the prevailing environmental conditions.

Path coefficient analysis confirmed that productive tillers per plant, number of filled spikelets per panicle and thousand-seed weight exerted the greatest positive direct effects on grain yield. These traits therefore represent the most reliable selection criteria for improving grain yield through indirect selection. The close agreement between correlation and path analyses strengthens the confidence in their use as breeding targets.

Principal component analysis demonstrated that the first five principal components explained 75.34% of the total phenotypic variation, with the first two principal components accounting for 44.55% of the observed diversity. Reproductive sink traits, including spikelet number, filled spikelets and spikelet fertility, contributed predominantly to the first principal component, whereas tillering traits contributed mainly to the second principal component. These findings indicate that reproductive capacity and tillering ability are the principal sources of phenotypic variation within the evaluated germplasm.

K-means clustering classified the rice accessions into five distinct phenotypic groups, confirming the existence of substantial diversity within the study population. The clustering pattern provides a valuable basis for identifying genetically divergent parental combinations that may maximize heterosis and generate superior recombinants in future breeding programmes.

The results of the present investigation provide a strong scientific basis for parent selection, indirect selection strategies and the efficient utilization of the 3K Rice Genome Panel in rice improvement programmes aimed at developing high-yielding and genetically diverse cultivars.

Acknowledgements

The authors gratefully acknowledge ICAR-National Institute of Seed Science and Technology, Regional Station, Bengaluru and KSNUAHS, Shivamogga for providing research facilities.

References

1. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2(4):433-459. doi:10.1002/wics.101.
2. Alam M, Lou G, Abbas W, Osti R, Ahmad A, Bista S, *et al.* Improving rice grain quality through ecotype breeding for enhancing food and nutritional security in Asia-Pacific Region. Rice. 2024;17:47. doi:10.1186/s12284-024-00725-9.
3. Ashikari M, Sakakibara H, Lin S, Yamamoto T, Takashi T, Nishimura A, *et al.* Cytokinin oxidase regulates rice grain production. Science. 2005;309(5735):741-745.
4. Bharamappanavara M, Madhyavenkatapura AS, Appaiah MC, Patil BS, Vijjeswarapu A, Senguttuvel P,

- et al.* Multivariate analysis for assessment of genetic diversity in rice (*Oryza sativa* L.). *Electron J Plant Breed.* 2023;14:1001-1012.
5. Caldo RA, Sebastian LS, Hernandez JE. Morphology-based genetic diversity analysis of ancestral lines of Philippine rice cultivars. *Philipp J Crop Sci.* 1996;21:86-92.
 6. Dewey DR, Lu KH. A correlation and path-coefficient analysis of components of crested wheatgrass seed production. *Agron J.* 1959;51:515-518.
 7. Food and Agriculture Organization. FAOSTAT statistical database. Rome: Food and Agriculture Organization; 2023.
 8. Fu YB. Understanding crop genetic diversity under modern plant breeding. *Theor Appl Genet.* 2015;128:2131-2142.
 9. Hair JF, Black WC, Babin BJ, Anderson RE. Multivariate data analysis. 8th ed. Cengage Learning; 2019.
 10. Hani Babu, Thouseem N. Genetic variability, correlation and path analysis of yield and its attributing traits in rice (*Oryza sativa* L.). *Int J Plant Soil Sci.* 2024;36(8):1029-1043. doi:10.9734/IJPSS/2024/v36i84936.
 11. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol.* 1933; 24:417-441.
 12. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature.* 2012;490(7421):497-501. doi:10.1038/nature11532.
 13. International Rice Research Institute. Standard evaluation system for rice. 5th ed. Los Baños: International Rice Research Institute; 2014.
 14. Johnson RA, Wichern DW. Applied multivariate statistical analysis. 6th ed. Pearson Education; 2007.
 15. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002.
 16. Karoda K, Singh Y, Singh SK, Prajapati PK, Kumar V. Principal component analysis and genetic divergence studies for yield and quality related attributes of hybrid rice. *Biotechnol J Int.* 2024;28(5):26-33. doi:10.9734/bji/2024/v28i5738.
 17. Kassambara A, Mundt F. factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.7; 2020.
 18. Khush GS. What it will take to feed 5.0 billion rice consumers in 2030. *Plant Mol Biol.* 2005; 59:1-6.
 19. Kumar KP, Pushpam R, Manonmani S, Raveendran M, Santhiya S, Senthil A. Enhancing stress resilience in rice (*Oryza sativa* L.) through profiling early-stage morpho-physiological and molecular responses to multiple abiotic stress tolerance. *Front Plant Sci.* 2024; 15:1342441. doi:10.3389/fpls.2024.1342441.
 20. Lakshmeesha R, Mahesh HB, Pattanashetti BM, Harinikumar KM, Anil VS. Principal component analysis of yield and its attributing traits in recombinant inbred lines of rice under submerged condition (*Oryza sativa* L.). *Arch Curr Res Int.* 2024;24(4):39-44. doi:10.9734/acri/2024/v24i4658.
 21. Le S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. *J Stat Softw.* 2008;25(1):1-18.
 22. Li JY, Wang J, Zeigler RS. The 3,000 rice genomes project: New opportunities and challenges for future rice research. *GigaScience.* 2014;3:8. doi:10.1186/2047-217X-3-8.
 23. MacQueen J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1. Berkeley: University of California Press; 1967. p. 281-297.
 24. Mohammadi SA, Prasanna BM. Analysis of genetic diversity in crop plants salient statistical tools and considerations. *Crop Sci.* 2003;43:1235-1248.
 25. Muthayya S, Sugimoto JD, Montgomery S, Maberly GF. An overview of global rice production, supply, trade and consumption. *Ann N Y Acad Sci.* 2014; 1324:7-14.
 26. Nachimuthu VV, Robin S, Sudhakar D, Rajeswari S, Raveendran M, Subramanian KS, *et al.* Genotypic variation for micronutrient content in traditional and improved rice lines and its role in biofortification programme. *Indian J Sci Technol.* 2014;7(9):1414-1425. doi:10.17485/ijst/2014/v7i9.15.
 27. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag.* 1901; 2:559-572.
 28. Sagar SPP, Diwan JR, Gireesh C, Tembhurne BV, Mahantashivayogayya K, Suma TC, *et al.* Multivariate analysis of early seedling vigour and its component traits in a subset of rice (*Oryza sativa* L.) 3K panel. *Plant Arch.* 2025;25(1):2060-2068.
 29. Salunkhe H, Kumar A, Janeja H, Krishna B, Talekar N, Mehendi S, *et al.* Genetic variability, correlation and path-coefficient analysis for yield and yield attributing traits in aerobic rice (*Oryza sativa* L.). *Electron J Plant Breed.* 2024;15(1):226-232. doi:10.37992/2024.1501.026.
 30. Shajan KT, Lal GM, Bara BM. Evaluation of genetic variability, correlation and path analysis in yield and yield contributing characters in rice (*Oryza sativa* L.) germplasm. *Ann Agric Res.* 2024;45(3):213-218.
 31. Singh R, Kumar P, Sharma V. Principal component and cluster analyses for evaluating genetic diversity in rice. *Electron J Plant Breed.* 2021; 12:654-664.
 32. Sinha S, Mishra PK. Principal component analysis of rice germplasm for yield and quality traits. *Oryza.* 2013; 50:242-247.
 33. The 3,000 Rice Genomes Project. The 3,000 rice genomes project. *GigaScience.* 2014;3:7. doi:10.1186/2047-217X-3-7.
 34. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, *et al.* Genomic variation in 3,010 diverse accessions

- of Asian cultivated rice. *Nature*. 2018;557(7703):43-49. doi:10.1038/s41586-018-0063-9.
35. Wright S. Correlation and causation. *J Agric Res*. 1921;20:557-585.
36. Xing Y, Zhang Q. Genetic and molecular bases of rice yield. *Annu Rev Plant Biol*. 2010;61:421-442.
37. Xu Y, Li P, Yang Z, Xu C. Genetic mapping and genomic selection for crop improvement. *Crop J*. 2017;5:101-115.
38. Yashwanth JV, Shanthi P, Shekar MR, Kumari PL, Vemireddy LR. Evaluating correlation and path coefficient analysis in advanced breeding lines of rice (*Oryza sativa* L.) for improved productivity. *Plant Arch*. 2025;25(1):653-662.
39. Yoshida S. Fundamentals of rice crop science. Los Baños: International Rice Research Institute; 1981.
40. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, *et al*. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun*. 2011;2:467.