



Deciphering Genetic Diversity and Scaling Exalted Genotypes of Soybean Using PCA

Ponnavada Abhishek ^{a*}, M.K. Shrivastava ^a,
Pawan Kumar Amrate ^a, Shivakumar M ^a, Yogendra Singh ^a,
Shivani Jawarkar ^a, Ravleen Kaur Badwal ^a
and Shrichand Patel ^a

^a Department of Genetics and Plant Breeding, Jawaharlal Nehru Krishi Vishwa Vidyalaya, Jabalpur, Madhya Pradesh-482004, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: <https://doi.org/10.9734/ijpss/2025/v37i105785>

Open Peer Review History:

This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: <https://pr.sdiarticle5.com/review-history/145201>

Original Research Article

Received: 06/08/2025

Accepted: 09/10/2025

Published: 16/10/2025

ABSTRACT

Soybean (*Glycine max*) is a nutritionally rich legume valued for its high protein and oil content, soil-enriching nitrogen fixation, and diverse food and feed uses. Its narrow genetic base necessitates studies on genetic diversity to enhance breeding efficiency and adaptability. Techniques like principal component analysis (PCA) and cluster analysis are crucial for classifying genotypes, identifying superior lines, and improving yield potential. This study aimed to assess the genetic diversity among 44 advanced soybean genotypes using Tocher's clustering, and Principal Component Analysis (PCA) during the *Kharif* season 2024 at JNKVV, Jabalpur. Tocher's method

*Corresponding author: E-mail: abhishekpinnavad@gmail.com, abhishekpinnavada@gmail.com;

grouped the genotypes into seven distinct clusters, with Cluster III showing the highest intra-cluster divergence and the maximum inter-cluster distance recorded between Cluster VI and Cluster III. PCA revealed that the first three principal components accounted for 76.58% of total variance, with PC1 alone contributing 51.26%, heavily influenced by key yield traits such as seed yield per plant, number of pods and seeds per plant, and biological yield. Genotypes 146, 154, 151, and 109 emerged as superior candidates for yield improvement based on high PC1 scores.

Keywords: Cluster analysis; principal component analysis.

1. INTRODUCTION

“Soybean (*Glycine max* (L.) Merr), commonly referred to as the "golden bean," is the most widely cultivated leguminous crop globally. It plays a vital role among grain legumes due to its rich protein and oil content. Containing approximately 38 to 42 percent high-quality protein and 18 to 20 percent oil, soybean serves as a potential solution to current and future global demands for edible oil and protein. As a complete protein source, it provides all nine essential amino acids” (Michelfelder, 2009). “Additionally, being a leguminous plant, it contributes to soil fertility by fixing atmospheric nitrogen adding about 65 to 100 kg N/ha/year which enhances the productivity of subsequent non-legume crops” (Fujitha et al., 1992).

The domesticated soybean ($2n = 40$) is classified under the order Fabales, within the Fabaceae (or Leguminosae) family, sub family Papilionoideae, tribe Phaseoleae and subtribe Glycininae. It belongs to the genus *Glycine*, which is divided into two subgenera: *Glycine* and *Soja*. The modern cultivated soybean was derived from its wild progenitor, *Glycine soja*.

It's significant nutritional value in preventing diabetes and heart disease are widely recognized (Dinesh et al., 2013). Regarding health promoting food, soybean-based products like soymilk, tofu, soy sauce, and natto are considered traditional dietary staples. Studies suggested consumption of tofu was linked to a lower risk of breast cancer, while soymilk reduces bone loss through isoflavones. Soy sauce, a fermented soybean product, is used in cooking, and known for its antioxidant properties (Wang & Komatsu, 2006). Soybean meal serves as supplement for fish and poultry diets (Kim et al., 2021).

Assessing genetic diversity in advanced developing lines supports the effective organization and classification of genotypes. In breeding programs, selecting offspring from

crosses involving genetically diverse parents based on genetic divergence analysis increases the likelihood of identifying transgressive segregants in later generations and achieving heterotic effects in early ones. Researchers continue to apply genetic diversity insights in soybean to select suitable parents for hybridization, aiming to develop desirable offspring (Jain & Ramgiry, 2000; Shrivastava et al., 2001; Kayande & Patil, 2009). The challenges posed by a narrowed genetic base in soybean reflected in reduced variability and limited adaptability highlight the critical need and relevance of ongoing research in this area (Rani et al., 2023).

Estimation of genetic diversity will be balanced and accurate by paying sufficient attention to the use of several data sets in data studies utilizing clustering algorithms and other multivariate techniques. Despite advances in agriculture, soybean yield per unit area remains relatively stagnant compared to major crops like rice, wheat, and maize, indicating that soybean has yet to experience a true green revolution through breeding efforts (Liu et al., 2020).

“Among multivariate analysis techniques, principal component analysis (PCA) serves as an effective method for reducing a large set of correlated variables into a smaller set of uncorrelated variables. Each principal component is formulated as a linear combination of the original variables and remains independent from the others. The objective of PCA is to retain the maximum amount of information by capturing the greatest possible variation present in the original dataset, based on the estimated ranking” (Cruz, 2006).

Cluster analysis, on the other hand, enables the grouping of individuals based on their similarities, aiming to maximize homogeneity within each cluster while ensuring clear heterogeneity between clusters. This method is effectively used for genotype classification and selection, serving as a complementary approach to principal

component analysis (PCA) (Mohammadi & Prasanna, 2003).

2. METHODS AND MATERIALS

The experiment was conducted during the *Kharif* season 2024 under All India Coordinated Research Project on Soybean, Seed Breeding Farm, Department of Genetics and Plant Breeding, College of Agriculture, JNKVV, Jabalpur (M.P.). The research site had homogeneous topography, medium-black soil with a pH of 7.5 and was free of stagnant water with recommended package and practices.

2.1 Climate and Weather Conditions

The city of Jabalpur (Latitude: 23.90 °N, Longitude: 79.58 °E, Elevation: 411.78 m above mean sea level). During the *Kharif* season, which spans from late June to early November, the region receives the bulk of its annual rainfall, along with high humidity and moderate temperatures. In 2024, the monsoon season followed its typical pattern, with almost 90% of the annual rainfall occurring between late June and late September. The total recorded rainfall during the season was approximately 1,420 mm, with the heaviest precipitation occurring in the weeks of July 29 – August 4 (217.2 mm) and August 19 – 25 (184.8 mm). The average

maximum temperature during the *Kharif* season was around 31.5°C, while the average minimum temperature hovered near 24.3°C. There were total of 48 rainy days recorded during the period, with the highest frequency of rainy days (6 days/week) observed in late July.

2.2 Experiment Details

The research material consists of total (42+2 Checks) genotypes; the list of genotypes along with their pedigree is mentioned below (Table 1). The experiment was carried out in three replications using Randomized Complete Block Design (RCBD) of plot size 0.45 m x 1.5 m (1 row/plot) with plant to plant spacing 7.0 cm

2.3 Observations Recorded

Observations were taken on three randomly chosen competitive plants from each advance breeding line in each replication to evaluate yield and yield-contributing traits. For quantitative trait observations, calculated the mean of the highest, middle, and lowest values for each trait, such as the number of primary branches per plant, the number of pods per plant, the number of seeds per plant, the 100 seed weight (g), the biological yield per plant (g), seed yield per plant (g) and harvest index (%) was recorded for this investigation

Table 1. List of pedigree with their code numbers

S. No.	Code no.	Pedigree
1	1	EC 457254 x JS 95-60
2	7	EC 457254 x JS 20-34
3	12	EC 457254 x JS 20-34
4	14	EC 457254 x JS 20-34
5	15	EC 457254 x JS 20-34
6	16	EC 457254 x JS 20-34
7	18	EC 457254 x JS 20-34
8	19	EC 457254 x JS 20-34
9	21	EC 538828 x NRC 155
10	22	EC 538828 x NRC 155
11	24	EC 538828 x NRC 155
12	25	EC 538828 x NRC 155
13	30	EC 538828 x NRC 155
14	32	EC 538828 x EC 457254
15	35	EC 457254 x JS 95-60
16	36	EC 457254 x JS 95-60
17	38	EC 457254 x JS 95-60
18	40	EC 457254 x JS 95-60
19	42	NRC 128 x JS 20-34
20	43	NRC 128 x JS 20-34
21	45	NRC 128 x JS 20-34
22	57	EC 457254 x JS 20-34

S. No.	Code no.	Pedigree
23	58	EC 457254 x JS 20-34
24	63	EC 538828 x JS 95-60
25	64	EC 538828 x JS 95-60
26	77	EC 538828 x NRC 128
27	94	CAT 47 x JS 20-69
28	95	EC 39072 x JS 95-60
29	96	EC 538828 x EC 457254
30	101	EC 457254 x JS 95-60
31	107	EC 457254 x JS 95-60
32	108	EC 457254 x JS 95-60
33	109	EC 457254 x JS 95-60
34	119	EC 457254 x JS 95-60
35	124	NRC 128 x JS 95-60
36	128	NRC 128 x JS 95-60
37	146	NRC 128 x JS 95-60
38	149	NRC 128 x JS 95-60
39	150	NRC 128 x JS 95-60
40	151	NRC 128 x JS 95-60
41	152	NRC 128 x JS 95-60
42	154	NRC 128 x JS 95-60
43	Check 1	JS 20-34
44	Check 2	JS 20-98

2.4 Methodologies Used to Genetic Divergence

To assess genetic divergence, Tocher's method was applied; utilizing transformed character means to convert correlated variables into standard, uncorrelated ones.

The Principal Component analysis (PCA) was done using R (R Core Team, 2019) and following main R packages: FactoMineR (Le *et al.*, 2008), FactoInvestigate (Thureau *et al.*, 2020), factoextra (Kassambara *et al.*, 2020), corrplot (Wei *et al.*, 2017), energy (Rizzo *et al.*, 2019), DDoutlier (Madsen *et al.*, 2018).

3. RESULTS AND DISCUSSION

3.1 Grouping of Genotypes into Different Clusters

Tocher's approach was used to categorise 44 genotypes into seven clusters based on D^2 values mentioned in Table 2. Cluster I to cluster IV was polygenotypic in which Cluster I having twelve genotypes, Cluster II with fifteen genotypes, cluster III with twelve genotypes, cluster IV with two genotypes, remaining three of seven clusters were monogenic of which each cluster has one genotype.

Similar findings were made by Manav and Arora (2018) who classified 45 soybean genotypes into

eight clusters, Patil *et al.* (2011) formed six clusters from thirty-six genotypes by taking eleven characters of soybean, Dubey (2018) studied 50 genotypes; in this, ten clusters were created based on the genotypes, Upadhyay *et al.* (2022) divided fifty genotypes into five clusters.

3.2 Intra and Inter Cluster Divergence D^2 Values

Cluster III showed maximum intra cluster D^2 value (8.72) followed by cluster II (6.97), cluster (6.85), cluster IV (3.48), while other clusters were mono-genotypic with no intra cluster divergence.

The highest inter cluster divergence was observed between genotypes of cluster VI and III (16.76), followed by cluster VI and V (15.61), cluster IV and III (14.90), cluster III and V (14.46), cluster I and V (14.43), cluster III and II (14.18), cluster V and IV (13.13), cluster VII and I (12.72), cluster VII and IV (12.28), cluster VII and VI, VI and IV (12.13), cluster VII and I (12.08), cluster II and I, VII and III (11.46), cluster IV and I (11.25), cluster IV and II (10.94), cluster III and I (10.82), cluster VII and II (10.21), cluster VII and V (9.98), cluster V and II (9.78) and cluster VII and II (9.55).

These findings are similar to Nayak and *et al.* (2023) showed highest intra cluster distance for cluster III, Kumar *et al.*, (2014) recorded highest number of genotypes in cluster II.

Table 2. Distribution of Soybean genotype with code number in different clusters

Cluster	No. of genotypes	Code no.
I	12	18, 152, 42, 154, 57, 101, 108, 149, 32, 1, 128, 77
II	15	19, 38, 16, 12, 107, 22, 45, 7, 43, 95, 155, 63, 94, 64, 156
III	12	25, 40, 146, 15, 24, 35, 96, 58, 151, 109, 150, 119
IV	2	30, 36
V	1	124
VI	1	14
VII	1	21

Table 3. Intra and Inter-cluster divergence (D^2) for seed yield and associated traits

Cluster number	Cluster matrix						
	I	II	III	IV	V	VI	VII
I	6.85	11.46	10.82	11.25	14.43	12.08	12.72
II		6.97	14.18	10.94	9.78	9.55	10.21
III			8.72	14.9	14.46	16.76	11.46
IV				3.48	13.13	12.13	12.28
V					0	15.61	9.98
VI						0	12.13
VII							0

3.3 Cluster Mean Values

The cluster mean values of different traits were given in Table 4. Cluster III has highest cluster mean primarily contributed by number of seeds per plant, on basis of cluster mean, Cluster III and VII exhibited yield contributing traits which includes number of pods per plants, number of seeds per plant, biological yield per plant. Early maturity genotypes are fall in the clusters of IV, early days to flowering is under the cluster VI. Cluster IV has lowest total cluster mean.

3.4 Principal Component Analysis

The principal Component Analysis (PCA) of ten traits across 44 soybean genotypes revealed critical insights into the underlying structure of trait variability:

3.5 Dimensional Reduction and Variance Explanation:

Only the first three principal components (PCs) had Eigen values greater than 1, collectively accounting for 76.58% of the total variability. Notably, PC1 alone explained 51.26%, highlighting its dominant contribution.

PC1 was strongly associated with seed yield per plant, number of seeds per plant, number of pods per plant, biological yield per plant and plant height. These traits are key yield contributors, making PC1 the most critical for yield-focused breeding.

PC2 correlated mainly with harvest index, suggesting its relevance for identifying genotypes with improved resource allocation.

PC3 was influenced by days to maturity and 100-seed weight indicating its importance for genotypes with larger seeds and efficient biomass use.

The scree plot depicting the relationship between Eigen values and principal components revealed that PC1 accounted for 51.25% of the total variation, followed by PC2 with 14.19% and PC3 with 11.12%. Collectively, the first three principal components explained 76.58% of the total variance. Among them, PC1 contributed the most to variation, as indicated by the steep initial slope of the curve, followed by PC2 and PC3. In contrast, PC4 through PC10 exhibited minimal variation individually. Therefore, focusing on genotypes associated with the traits represented by PC1 to PC3 would be beneficial for selection purpose.

3.6 Genotypic Performance

Genotypes of these code numbers 146, 154, 151, 109, and 150 scored high value on PC1, indicating superior yield potential and related traits. Genotypes with codes (64, 1, 38, and 151) ranked high on PC2, suggesting maturity-related advantages. Genotypes represented by the following code 43, 1, 45, 7, and 64 led PC3, reflecting favorable values for seed weight and maturity.

Table 4. Cluster means value for ten traits in soybean

	Days to 50% flowering	Days to maturity	Plant height (cm)	No. of primary branches	No. of pods per plant	No. of seeds per plant	100 seed wt (g)	Biological yield per plant (g)	Harvest Index (%)	Seed yield per plant (g)
Group. 1	45.75	97.44	45.17	3.03	52.81	87.31	8.48	22.66	33.66	7.58
Group. 2	37.69	99.22	43.53	3.38	43.53	70.89	9.99	21.4	34.51	7.21
Group. 3	47.69	95.75	50.41	4.81	72	150	10.57	38.01	40.95	15.36
Group. 4	41.33	90.5	41.06	3.5	10.17	21.17	9.65	13.25	21.53	2.85
Group. 5	38.67	99.67	48.44	3.33	36.33	52	14	16.01	43.62	6.93
Group. 6	36.67	95.67	42.22	3.33	55	79.67	6.81	14.05	43.99	6.07
Group. 7	39.67	92.67	42.44	4.67	71.67	117.33	11.74	26.78	48.57	13.04

Table 5. Eigen values, percentage of total variation and cumulative percentage for corresponding ten traits of soybean

Principal components (PC)	Eigen value	Variability (%)	Cumulative %
PC1	5.125539	51.25539	51.25539
PC2	1.419954	14.19954	65.45492
PC3	1.112653	11.12653	76.58145
PC4	0.753909	7.539095	84.12055
PC5	0.590448	5.904485	90.02503
PC6	0.486197	4.861971	94.887
PC7	0.326394	3.263945	98.15095
PC8	0.14336	1.433605	99.58455
PC9	0.035826	0.358258	99.94281
PC10	0.005719	0.05719	100

Table 6. Principal components for ten yield contributing traits of soybean lines

Traits	PC1	PC2	PC3
Days to 50% flowering	0.2960	-0.2050	-0.4550
Days to maturity	-0.1160	0.4590	0.4980
Plant height (cm)	0.3070	0.1810	-0.0010
No of primary branches per plant	0.2230	-0.4710	0.2590
No. of pods per plant	0.3960	0.2090	-0.0370
No. of seeds per plant	0.4210	0.1710	-0.0360
100 seed weight (g)	0.0860	-0.5130	0.6040
Biological yield per plant (g)	0.3880	-0.1480	-0.0340
Harvest index (%)	0.2800	0.3650	0.3210
Seed yield per plant (g)	0.4300	0.0200	0.0820

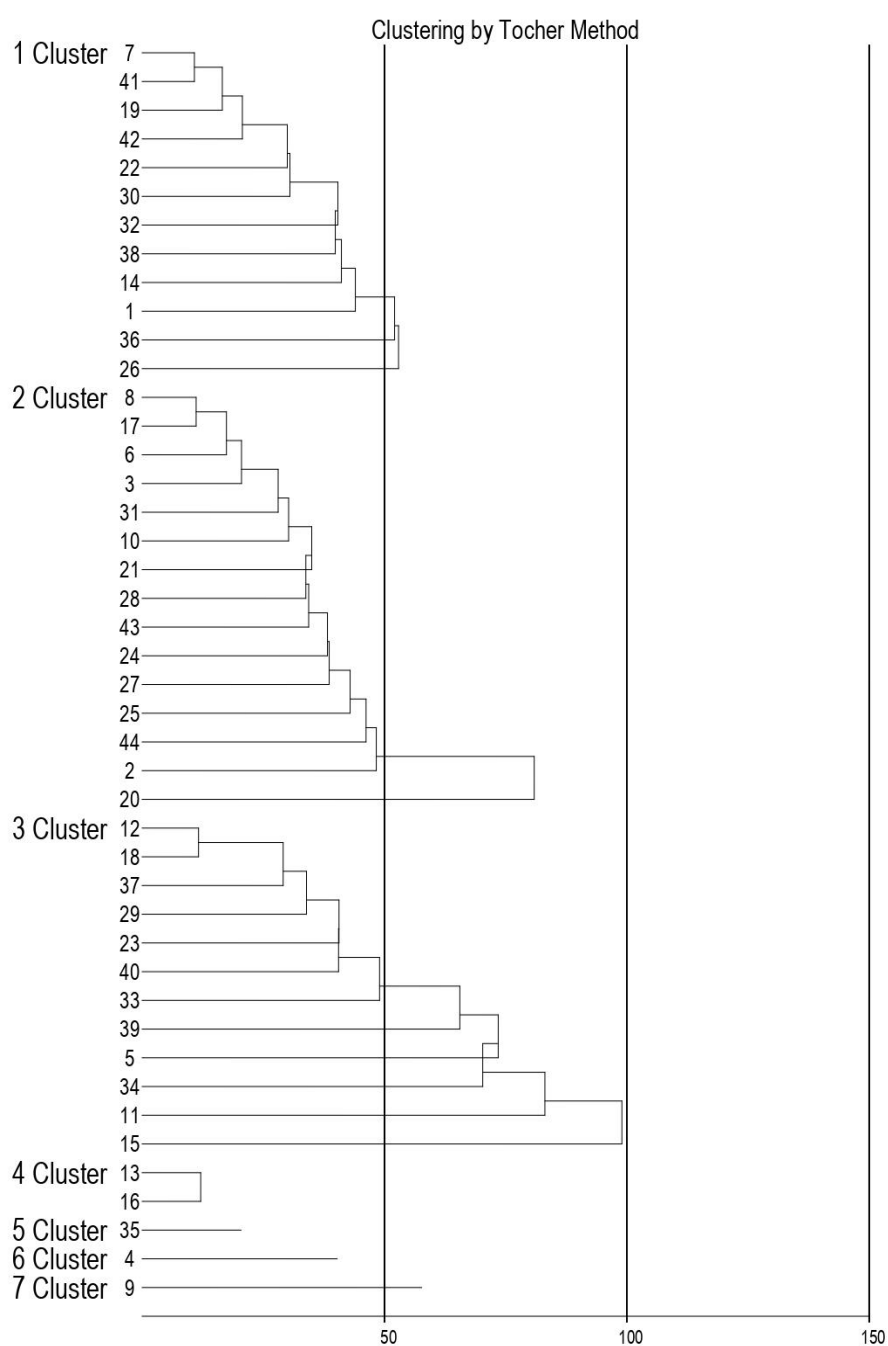


Fig. 1. Dendrogram showing cluster analysis

Table 7. Interpretation of rotated component matrix for the traits having values > 0.3 in each PCs

	PC 1	PC 2	PC 3
Characters	Seed yield per plant (g)	Harvest index (%)	Days to maturity
	No. of seeds per plant		100 seed weight (g)
	No. of pods per plant		
	Biological yield per plant (g)		
	Plant height (cm)		

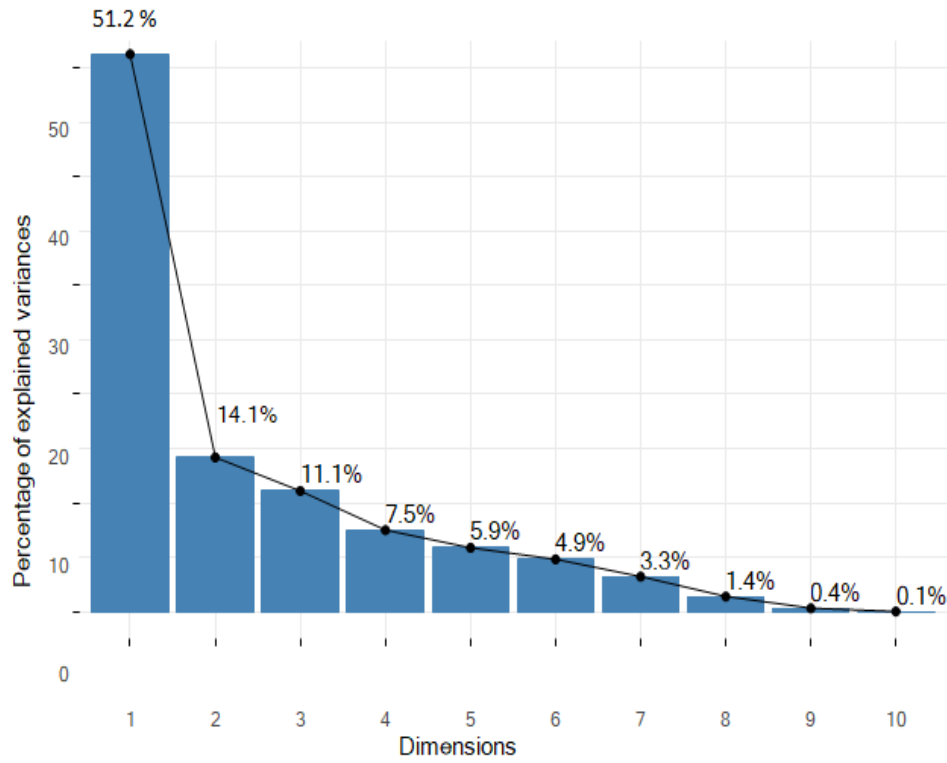


Fig. 2. Scree Plot laid between Eigen value and Cumulative variability of 10 PCs

Table 8. PC scores of soybean lines

Code no	PC1	PC2	PC3
77	-2.1349	0.5783	0.4590
96	-3.6704	1.1554	-0.5599
24	-1.4338	0.7293	-1.9583
15	-1.3671	-1.5072	0.5828
154	4.8000	-0.8476	-0.8418
14	-2.0071	0.1618	-1.5894
7	-0.0882	-0.8110	1.2863
19	-1.0327	0.7488	-1.0353
119	1.5246	0.8815	-0.9890
95	-0.3822	-0.9752	-0.9017
109	4.0351	-2.4804	-0.0846
42	1.5422	1.1997	-0.1176
64	-2.9137	2.3070	1.2377
18	-0.6928	0.4177	0.9069
146	5.0019	1.6453	0.6790
1	-3.7318	2.2071	2.0407
63	-1.4634	-0.0349	-0.7489
107	1.1090	0.4508	-0.4356
124	-0.2370	-1.0948	1.0046
40	-0.0536	-2.0603	-0.1045
12	-1.1924	-1.8674	-0.5941
58	0.0270	-1.3192	0.6622
152	2.0172	1.3770	0.6352
22	-2.5016	-1.3153	0.0809

Code no	PC1	PC2	PC3
25	-2.9264	-0.4728	-0.5748
57	0.4061	-1.2258	1.1081
36	0.5387	-0.5174	-0.8122
21	-1.3590	-0.5572	-0.8881
150	3.4463	0.3069	-0.8665
128	-1.7768	-0.8102	0.6793
32	-0.9760	-0.5567	-0.4466
94	1.0652	-0.8598	0.0926
151	4.1799	2.0213	0.6437
108	3.1662	-1.0945	0.4652
30	-1.1889	0.3862	-2.6647
16	-1.8394	0.2714	1.0578
149	2.3253	1.5031	0.8209
101	-0.9890	0.0930	1.1782
38	0.8457	2.1724	-1.4236
35	3.3068	0.5448	0.1020
45	0.6586	-0.7013	1.3639
43	-0.9674	-0.5143	2.0964
JS 20-34	-2.4530	0.2558	-0.2357
JS 20-98	-0.6168	0.2087	-1.3104

These results are consistent with the observations reported by Verma *et al.* (2021), where the first principal component (PC1) accounted for the highest proportion of variance, marked by an Eigen value of 2.17. In their analysis, the number of seeds per plant contributed the most to PC1, explaining approximately 64% of its variation. Similarly, El-Hashash (2016) highlighted that the first two principal components, each with Eigen values exceeding one, captured a substantial portion of variability in the dataset. Vianna *et al.* (2013) reported that four components with Eigen values above one accounted for 71.6% of the total variability, leading to the identification of seventeen superior genotypes, mainly associated with yield-related traits.

3.7 Breeding Implications

The concentration of major yield-related traits in PC1 emphasizes the potential of selecting genotypes with high PC1 scores for yield improvement programs. Meanwhile, PCs 2 and 3 provide supplementary selection dimensions for maturity and seed quality traits.

4. CONCLUSION

Using Mahalanobis D^2 statistics, cluster means and Tocher's clustering method revealed substantial genetic variability for the ten yield-related traits examined. The existence of considerable inter-cluster distances, particularly between Cluster VI and Cluster III, suggests that

hybridization between genotypes from these clusters may result in high levels of heterosis and transgressive segregation. The identified clusters and the genotypes within them offer promising genetic combinations for hybridization aimed at enhancing yield potential and its contributing traits in future breeding efforts.

PC1, which explained the largest portion of variation (51.25%), was primarily influenced by key agronomic traits such as seed yield per plant, number of seeds and pods per plant, biological yield, and plant height, making it the most critical component for yield-related selection. Genotypes such as 146, 154, 151, and 109 showed high PC1 scores and may serve as promising candidates for yield improvement. PC2 and PC3 contributed an additional 14.20% and 11.13% of the variation, respectively, with traits like days to maturity, harvest index, and 100-seed weight showing strong associations.

5. SUGGESTIONS FOR FURTHER WORK

Based on the D^2 analysis lines from divergent clusters *Viz.* Code no.(25, 40, 146, 15, 24, 35, 96, 58, 151, 109, 150, and 119) and (124, 36, 14, 30) should be prioritized for strategic crosses to create populations with wider recombination potential. Such crosses can help develop transgressive segregants for seed yield and component traits.

By combining all statistical analysis like PCA, D^2 analysis and disease screening genotypes with

code no. (146, 42, 45, 35, 96, 151, 109, 150, 124, 152, and 64) subjected to marker-assisted selection can significantly accelerate the identification and development of high-performing breeding lines, particularly for complex traits such as yield and disease resistance.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of this manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

- Cruz, C. D. (2006). *Programa Genes: análise multivariada e simulação*. Viçosa: UFV.
- Dinesh, K., Agarwal, Billore, S. D., Sharma, A. N., Dupare, B. U., & Srivastava, S. K. (2013). Soybean: Introduction, improvement, and utilization in India problems and prospects. *Agricultural Research*, 2(4), 293–300.
- Dubey, N., Harshal, A., Avinash, & Shrivastava, A. N. (2018). Evaluation of genetic diversity among soybean (*Glycine max* (L.)) genotypes using multivariate analysis. *Plant Archives*, 18(1), 908–912.
- El-Hashash, E. F. (2016). Genetic diversity of soybean yield based on cluster and principal component analyses. *Journal of Advances in Biology & Biotechnology*, 10(3), 1–9.
- Fujitha, K., Ofosus, Bady, K. G., & Ogatas. (1992). Biological nitrogen fixation in mixed legume–cereals cropping system. *Plant and Soil*, 141(1-2), 155–175.
- Jain, P. K., & Ramgiry, S. R. (2000). Genetic diversity in Indian and abroad collection of soybean. *Advanced Plant Science*, 13(2), 487–490.
- Kassambara, A., & Mundt, F. (2020). *Factoextra: Extract and visualize the results of multivariate data analyses*.
- Kayande, N. V., & Patil, S. P. (2009). Genetic divergence in soybean (*Glycine max* (L.) Merrill). *International Journal of Plant Sciences*, 4(1), 218–220.
- Kim, B. G., Lee, E. J., Oh, H. S., Choi, S. S., Han, Y. J., Lee, B. D., Lee, J. H., & Kim, Y. W. (2021). Study of enzyme-hydrolyzed soybean replacing fish meal and/or chicken meal on the growth of channel catfish (*Ictalurus punctatus*). *Fish and Shellfish Immunology*, 119, 13–23.
- Kumar, A., Pandey, A., Aochen, C., & Pattanayak, A. (2014). Evaluation of genetic diversity and interrelationships of agromorphological characters in soybean (*Glycine max*) genotypes. *Proceedings of the National Academy of Sciences*, 85(2), 397–405.
- Le, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.
- Liu, S. L., Zhang, M., Feng, F., & Tian, Z. X. (2020a). Toward a “green revolution” for soybean. *Molecular Plant*, 13, 688–697.
- Madsen, J. H. (2018). *DDoutlier: Distance & density-based outlier detection*.
- Manav, & Arora, R. N. (2018). Assessment of genetic diversity for yield and seedling traits in soybean (*Glycine max* L. Merrill). *Electronic Journal of Plant Breeding*, 9(1), 355–360.
- Michelfelder, A. J. (2009). Soy: A complete source of protein. *American Family Physician*, 79(2), 43–47.
- Mohammadi, S. A., & Prasanna, B. M. (2023). Analysis of genetic diversity in crop plants salient statistical tools and considerations. *Crop Science*, 43(4), 1235–1248.
- Nayak, S., Meena, B. L., Sharma, S. C., Tak, Y., Meghwal, H. P., Kumari, S., & Himanshi. (2023). Genetic diversity studies in soybean (*Glycine max* (L.) Merrill). *The Pharma Innovation Journal*, 12(11), 2315–2318.
- Patil, V. V., Girase, V. S., Ghule, K. S., Ghume, O. G., & Shaniware, Y. A. (2024). Correlation and path analysis studies in soybean (*Glycine max* (L.) Merrill). *International Journal of Advanced Biochemistry Research*, 8(12), 281–288.
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rani, R., Raza, G., Tung, M. H., Rizwan, M., Ashfaq, H., Shimelis, H., et al. (2023). Genetic diversity and population structure analysis in cultivated soybean (*Glycine max* [L.] Merr.) using SSR and EST-SSR

- markers. *PLOS ONE*, 18, e0286099. <https://doi.org/10.1371/journal.pone.0286099>
- Rizzo, M., & Szekely, G. (2019). *Energy: E-Statistics: Multivariate inference via the energy of data*.
- Shrivastava, M. K., Shukla, R. S., & Singh, C. B. (2001). Genetic divergence in soybean. *JNKVV Research Journal*, 34(1-2), 25–28.
- Thuleau, S., & Husson, F. (2020). *FactoInvestigate: Automatic description of factorial analysis*.
- Upadhyay, P., Shrivastava, M. K., Sharma, S., Thakur, S., & Anand, J. K. (2022). Morphological characterization of exotic lines of soybean (*Glycine max* (L.) Merrill) for developing ideotype. *Biological Forum – An International Journal*, 14(2), 546–551.
- Verma, V., Shrivastava, M. K., Mehra, S., Amrate, P. K., & Yadav, R. B. (2021). Estimation of genetic parameters for yield associated traits and principal component in advance breeding lines of soybean (*Glycine max* (L.) Merrill). *International Journal of Current Microbiology and Applied Sciences*, 10(1), 2704–2710.
- Vianna, V. F., Uneda-Trevisoli, S. H., Desiderio, J. A., Santiago, S. D., Charnai, K., Junior, J. A. F., Ferraudo, A. S., & Mauro, A. O. (2013). The multivariate approach and influence of characters in selecting superior soybean genotypes. *African Journal of Agricultural Research*, 8(30), 4162–4169.
- Wang, X., & Komatsu, S. (2006). Improvement of soybean products through the response mechanism analysis using proteomic technique. *Advances in Food and Nutrition Research*, 82, 117–148.
- Wei, T., & Simko, V. (2017). *R Package “Corrplot”: Visualization of a correlation matrix*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright (2025): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:

The peer review history for this paper can be accessed here:
<https://pr.sdiarticle5.com/review-history/145201>