Received: 28 Nov 2024 | Revised: 25 Jun 2025 | Accepted: 26 Jun 2025

Electronic Journal of Plant Breeding

Research Note

Inquisition on principal component and K- mean clustering analysis for yield and its contributing traits of bread wheat (*Triticum aestivum* L.)

Rishabh Singh, Vijay Kumar Yadav*, Prem Kumar, Shivanshu Shekhar, Vaibhav Chauhan and Amit Kumar

Department of Genetics and Plant Breeding,Chandra Shekhar Azad University of Agriculture & Technology, Kanpur (U.P.) India
***E-Mail:** vkyadu@gmail.com

Abstract
An experiment was conducted to assess genetic diversity *via* K cluster and principal component analysis (PCA) for yield and its contributing characters in 41 bread wheat genotypes at Crop Research Farm, Nawabganj, Chandra Shekhar Azad University of Agriculture & Technology, and Kanpur (U.P.) during *Rabi* season 2020-21. K- Clustering statistics was used to analyse the genetic divergence in the present group of material. All the 41 genotypes were classified into seven clusters. Cluster I comprised of ten genotypes, cluster II comprised of eight genotypes, cluster III of ten genotypes, clusters IV, V and VI consisted of two genotypes while cluster of VII had seven genotypes. Cluster-VII ranked first in days to maturity, productive tiller per plant, ear length and 1000 seed weight (gm). The maximum contribution towards the expressions of genetic divergence was exhibited by grain yield per plant (30.73%) followed by plant height (31.59%), chlorophyll content (13.41), number of spikelets per spike (6.95%), 1000 grains weight (5.73%), days to 50% flowering (4.63%), ear length (4.36%), days to maturity(1.83%) and number of grains per spike (0.61). Cluster I and VI are most desirable for breeding programme for creating the desired variability and the genotypes in these clusters deserve due consideration for improvement of traits like chlorophyll content, plant height and spike length. The outcomes of this study can contribute to the strategic development of wheat breeding initiatives aimed at exploiting heterosis.

Keywords: Wheat, PCA, genetic diversity, cluster analysis, grain yield.

Wheat (*Triticum aestivum* L.) is one of the most widely grown crops, feeding about 40% of the world's population (Mao *et al*., 2023). It has gained considerable importance worldwide due to its vast farmed area, high production, and its prominence in the worldwide food grain trade. Wheat is consumed in a number of forms, including bread, chapatti, porridge, flour, and suji. To improve wheat, we must tap into the genetic diversity in wheat germplasm to identify different genotypes for use in wheat breeding. Estimating the genetic relationships among the accessions is a viable strategy for parental selection in wheat hybridization programme since it can lead to promotion of genetic recombination in the progenies, thus, enhancing the yield potential. Principal component analysis (PCA) and cluster analysis are two complementary methods widely used to assess genetic diversity in wheat, each offering distinct insights into population structure (Sarfraz *et al*., 2021). The advantage of PCA over cluster analysis is that each germplasm line can be assigned to a single group (Mohammadi, 2002), whereas cluster analysis is an appropriate technique for examining ancestral relationships (Mellingers, 1972). The primary goal of this study was to identify the genetic diversity of wheat genotypes using statistical approaches such as PCA and K-cluster analysis.

The present experiment was conducted to analyze the genetic diversity present among the genotypes 41 bread wheat genotypes using PCA and K-means clustering based on observations of various quantitative traits. The experiment was conducted at Crop Research Farm, Nawabganj, Chandra Shekhar Azad University of Agriculture & Technology, and Kanpur (U.P.) during *rabi* 2019- 2020 season. Each genotype was sown in two rows of 5.0 m length with a spacing of 22.5×5 cm. The experiment was laid out in a Randomized Block Design (RCBD) with three replications. Recommended agro-techniques and plant protection techniques were adopted to cultivate a healthy crop. Observations on ten characters *viz*., days to 50% flowering, days to maturity, number of productive tillers per plants, plant height, ear length, number of spikelets per spike, number of grains per spikes, 1000-grain weight (gm), chlorophyll content and grain yield per plant were recorded in five plants of each genotype, chosen randomly. The replication wise mean data was used for statistical analysis. The statistical tool SAS (Statistical Analyses Software) was utilized for principal component and K cluster mean analysis (Hartigan and Wong,1979).

Principal component analysis determines the pattern of variation, whereas relationship analysis is used to determine the variation and estimate the relative contribution of individual attributes to total variability. It is a kind of multivariate analysis used to examine the importance of the largest contributor to total variance along each axis of differentiation (Khare, 2022). The PCA scores for 41 genotypes were computed and presented graphically in **Fig 1** and combined PCA and Factor Analysis (FA) analysis delivers a clear insight into how every agronomic trait contributes to general variability of different traits of wheat are mentioned in **Table 1 and 2.** Grain yield per plant had very high positive loading on both PC1 (0.508) and Factor 1 (0.512), indicating it as the highest contributing trait towards yield determination. Likewise, productive tillers per plant contributed positively (PCA: 0.352, FA: 0.463), showing its direct contribution in increasing yield. Grains per spike was characterized by high positive scores (PCA: 0.426, FA: 0.395), supporting its significance in yield improvement. Spikelets per spike was also a positive contributor to PC1 (0.426) and was found to have moderate to high engagement in FA (0.143 in F1, -0.578 in F4), indicating its contribution to spike architecture. Interestingly, plant height was negatively loaded in both analyses (PCA: -0.393, FA: -0.533), suggesting that lower plants could be more efficient in regards to yield. Days to maturity were highly positively loaded in PC2 (0.593) and Factor 2 (0.668), suggesting its correlation with late maturity and grain development. 1000-grain weight was also synchronized (PCA: 0.564, FA: 0.529), indicating that genotypes with greater grain weight fall under the category of maturity trait cluster. Chlorophyll content, relating to physiological efficiency, was found to be most prominent in PC3 (0.627) and FA Factor 3 (0.588), indicating its involvement in early vigor or stress tolerance. Days to 50% flowering indicated high loading in PC3 (0.560) and FA (0.744 in Factor 3), consistent with early process development. Ear length, while moderately expressed in PCA (PC4: -0.676) and FA (F4: -0.700), was negatively correlated, suggesting longer ears might not necessarily have a greater yield. The results were supported by the previous studies by Hailegiorgis *et al*. (2011), Mishra *et al*. (2015), Khan *et al.* (2015) and Abdelghany *et al.* (2023).

As indicated in **Table 3**., PCA I, PCA II, and PCA III combined explain 75% of the variation, each explaining around 25%. Genotypes with high values on PCA I, i.e., IC-534929 (20.650), EC-539267 (19.024), and EC-578142 (18.529), are located highly on the axis loaded with yield-related characteristics such as grain yield per plant, number of grains per spike, and productive tillers. This indicates that these genotypes have high yield potential. Conversely, genotypes such as EC-434545 (9.732) and IC-28755 (9.718) had low PCA I value, indicating low performance related to yield. PCA II discriminates genotypes with respect to factors such as days to maturity and grain weight. IEC-539267 (44.914), EC-11360 (44.296), and IC-406688 (44.509) have high scores for PCA II, indicating delayed maturity and potentially larger seeds. Genotypes such as EC-576889 (40.325) and IC-28755 (41.089) have lower scores, reflecting premature maturity. PCA III, which is indicative of physiological characteristics like chlorophyll content, brings to the fore genotypes such as IEC-539267 (61.171), EC-112558 (61.104), and IC-144903 (60.883) as physiologically efficient ones. Generally, the distribution of PCA scores allows easy visualization of genotypic diversity and facilitates the determination of potentially elite lines with favourable combinations of traits. Genotypes which have a score significantly along the axes can be used as ideal parents in breeding programs to enhance yield, duration of maturity, and physiological tolerance in wheat. Similar result were observed by Janmohammadi *et al.* (2014) Kumar *et al.* (2021), Elahi *et al*. (2021) and Šućur *et al.* (2024).

K- means Clustering via Principal component analysis: The K- cluster within sum of squares clustering for the 40 genotypes were computed and presented graphically in **Fig 2. and Table 4**. The K cluster mean analysis revealed that cluster I and cluster III had 10 genotypes, cluster II had 8 genotypes, cluster VII had 7 genotypes each, cluster IV, V and cluster VI had 2 genotypes each (**Table 5)**. The individual clusters showed supremacy for several attributes as indicated by the cluster mean value in **Table 5**. Cluster V showed the lowest K-cluster means for days to 50% flowering, ear length per plant and thousand grain weight. Similar reports were also reported by Baranwal *et al.* (2013), Marino and Alvino (2020) and Yasin *et al.* (2024) while working on wheat. Cluster VI has the lowest K-cluster mean in terms of days to maturity and productive tillers per plant and grain yield per plant. while Cluster IV had the lowest K-cluster mean in terms

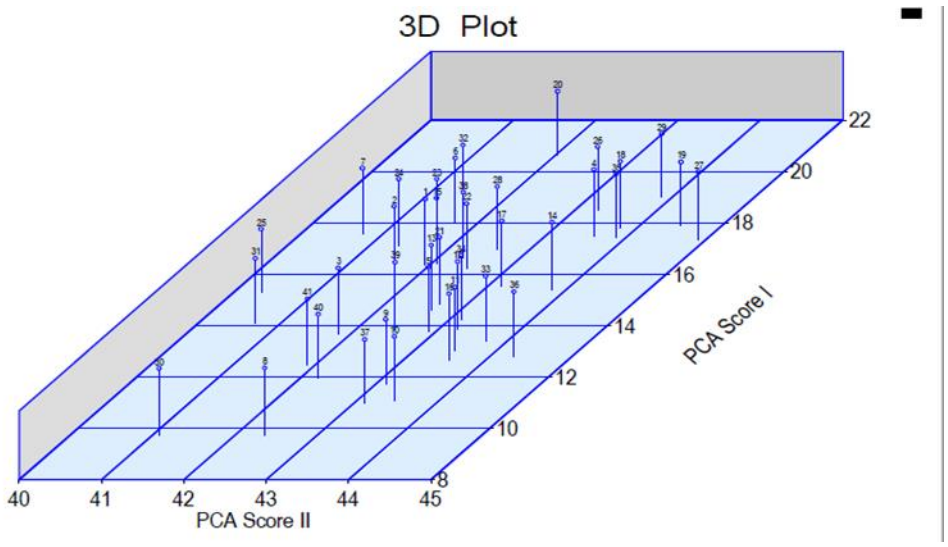

Fig.. 1. 3D plot of PCA Score I and II

Table 1. Canonical Root Analysis (PCA) for different characters

	Trait	Vectors	1 Vector	2 Vector	3 Vector	4Vector	5 Vector
		Variance Explained	**34.00716**	**14.34531**	**13.63747**	**10.25202**	**9.10799**
		Cumulative Variance Explained	**34.00716**	**48.35247**	**61.98995**	**72.24197**	**81.34996**
1	Day of 50% flowering		0.07682	0.24173	0.55995	0.47091	0.02119
2	Day to maturity		0.06038	0.59295	-0.31349	-0.20278	0.02164
3	Productive tiller per plant		0.35235	0.09306	-0.21093	0.22767	0.66625
4	Plant height		-0.39331	-0.10527	0.26858	-0.26468	0.25313
5	Ear lenght per plant		0.23532	0.23646	0.18476	-0.67552	0.19743
6	No of spikelets per spike		0.42600	-0.24649	0.18975	-0.35463	-0.21863
7	No of grain per spike		0.42582	-0.35462	-0.07390	0.02672	-0.33396
8	Weight of 1000 grains		0.13077	0.56355	0.02338	0.08009	-0.49175
9	Chlorophyll content		0.13348	0.09679	0.62707	-0.01342	0.10924
10	Grain yield per plant		0.50844	-0.02914	-0.04668	0.16075	0.19762

Table 2. Factor analysis for different characters

	Traits	1 Factor	2 Factor	3 Factor	4 Factor
1	Day of 50% FLOWERING	0.0810	0.0491	0.7436	0.1942
2	Day to maturity	0.0640	0.6682	-0.1999	-0.0638
3	Productive tiller per plant	0.4632	0.0995	-0.0219	0.0646
4	Plant height	-0.5326	-0.1176	0.0491	-0.0900
5	Ear length per plant	-0.1559	0.2954	-0.0071	-0.7001
6	No of spikelets per spike	0.1433	-0.2209	0.0320	-0.5776
7	No of grain per spike	0.3953	-0.3303	-0.0778	-0.2046
8	Weight of 1000 grains	0.1337	0.5288	0.2082	0.0280
9	Chlorophyll content	-0.1016	-0.0306	0.5879	-0.2523
10	Grain yield per plant	0.5117	-0.0410	0.0895	-0.1258

Table 3. Principal factor score of 41 genotypes in three principal components

	Genotypes	PCA I	PCA II	PCA III
		X Vector	Y Vector	Z Vector
1	IC-290161	16.354	41.950	59.706
2	EC-267020	16.103	41.660	58.767
3	EC-577448	13.660	41.856	59.807
4	EC-299060	17.502	43.590	59.796
5	EC-576591	13.796	42.898	57.219
6	IC-534929	17.981	41.735	58.863
7	IC-75208	17.570	40.755	59.273
8	EC-434545	9.732	42.365	60.750
9	EC-463396	11.730	43.122	58.164
10	CHUNGMAI	11.078	43.460	57.883
11	IC-128664	13.024	43.498	57.028
12	IC-144903	13.865	43.233	60.883
13	IC-145522	14.588	42.653	58.800
14	EC-112558	15.380	43.838	61.104
15	IC-35163	16.399	42.073	59.486
16	IC-554661	12.671	43.551	59.536
17	IC-535800	15.555	43.163	58.328
18	EC-578064	17.829	43.793	59.493
19	IC-406688	17.885	44.509	57.530
20	IC-534929	20.650	42.023	57.129
21	IC-145237	14.859	42.656	59.585
22	IC-335540	16.199	42.506	59.024
23	IC-252469	17.180	41.795	58.570
24	IC-28889	17.118	41.355	59.974
25	EC-576889	15.323	40.325	56.389
26	IC-533610	18.514	43.271	56.752
27	IEC-539267	17.348	44.914	61.171
28	IC-401927	16.964	42.608	56.790
29	EC-539267	19.024	43.868	56.725
30	IC-28755	9.718	41.089	60.578
31	EC-11071	14.067	40.704	58.992
32	EC-578142	18.529	41.633	58.028
33	EC-609338	13.394	43.741	58.725
34	IC-240801	14.226	43.146	58.279
35	EC-577738	17.416	43.888	57.607
36	EC-11360	12.801	44.296	58.350
37	IC-539313	11.004	43.125	56.876
38	IC-375938	16.630	42.313	59.389
39	PBW-373	13.832	42.486	60.941
40	K-1317	11.955	42.221	57.731
41	HD-2967	12.486	41.901	59.047

Table 4. K - clustering pattern for different Genotypes

Group	n	Within SS	Cluster Members
1	10	8.1994	EC-577448, EC-576591, IC-144903, IC-145522, EC-112558, IC-535800, IC-145237, EC-609338, IC-240801, PBW-373.
2	8	8.3394	EC-463396, CHUNGMAI, IC-128664, IC-554661, EC-11360, IC-539313, K-1317, HD-2967
3	10	6.2403	IC-290161, EC-267020, IC-534929, IC-75208, IC-35163, IC-335540, IC-252469, IC-28889, IC-401927, IC-375938
4	2	2.3258	IC-534929, EC-578142
5	2	0.8611	EC-576889, EC-11071
6	2	0.8139	EC-434545, IC-28755
7	7	4.1976	EC-299060, EC-578064, IC-406688, IC-533610, EC-539267, EC-539267, EC-577738

Table 5. The K-cluster means for seven clusters

Cluster	D50F	DM	PTPP	PH	ELP	NSPP	NGPS	TGW	CC	GYP
1 Cluster	74.157	117.690	8.10	97.050	8.836	16.070	55.027	40.362	45.476	15.386
2 Cluster	72.849	119.750	7.30	100.092	8.553	15.283	53.992	40.257	44.368	12.965
3 Cluster	73.357	116.453	8.87	93.493	8.713	17.653	59.350	39.629	45.569	18.240
4 Cluster	73.117	115.917	9.23	83.667	8.817	18.050	61.350	40.117	44.472	19.830
5 Cluster	71.033	116.950	8.23	94.683	8.043	16.800	58.483	36.403	45.348	16.585
6 Cluster	72.983	114.450	6.86	116.750	8.767	16.283	57.083	40.540	45.578	12.710
7 Cluster	74.005	121.463	9.38	88.862	9.463	16.776	55.842	40.977	45.318	18.859

D50F-days to 50% flowering, DM-days to maturity, PTPP-productive tillers per plant, PH- plant height, ELP-Ear length, NSPP- number of spike per plant, NGPS- number of grains per spike, TGW-1000 grain weight, CC- chlorophyll content, GYP- grain yield per plant.

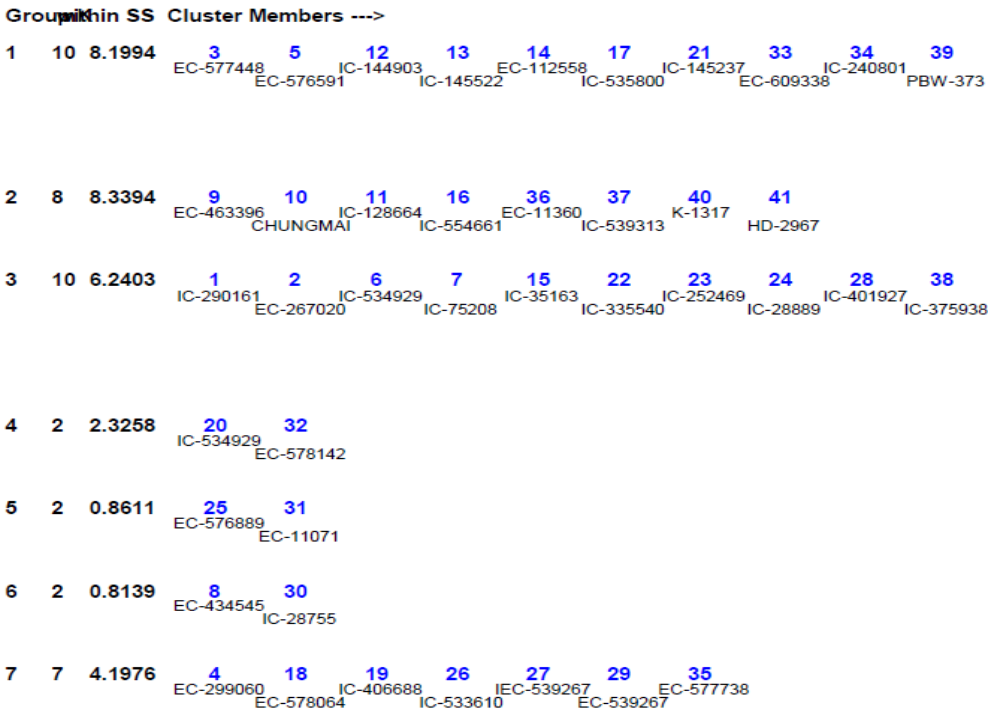

Fig.2. Groups within SS clustering members through K Clustering method

of plant height. Cluster II showed the lowest K-Cluster mean for number of spikelets per spike, number of grains per spike and chlorophyll content. The cluster mean value indicates how each unique cluster represented supremacy for different attributes. Cluster I recorded the highest K-cluster means for days to 50% flowering. Cluster IV showed the greatest K-cluster mean at number of spikelet per spike. Number of grains per spike and grain yield per plant. Cluster VI had the greatest K-cluster mean for plant height and chlorophyll content. Cluster VII had the greatest K- cluster mean for days to maturity, productive tillers per plant, ear length per plant and thousand grain weight. These results were in agreement with Khodadadi *et al.* (2014), Singh *et al.* (2014) Kumar *et al.* (2016), Sthapit *et al*. (2020), and Dervishi *el al*. (2022). However, for improving the grain yield in wheat, the diverse clusters, *i.e*. cluster IV and cluster VI may be used in breeding programs.

PCA, Factor Analysis, and K-means clustering reflected an integrated view of the genetic diversity and the trait contribution of wheat genotypes. Yield-conducive traits like grain yield per plant, productive tillers, and grains per spike were found to be the most contributory, while physiological and maturity traits were also highly contributory towards total variability. Genotypes such as IC-534929, EC-539267, and EC-578142 exhibited better performance, suggesting their potential towards yield enhancement. Clustering analysis also classified genotypes into separate groups, each expressing dominance for a given character. Cluster IV and Cluster VI were noteworthy groups with diametrically opposite characteristics and thus were good candidates for hybridization to realize heterosis and enhance yield potential. Combining multivariate analysis and clustering successfully delineates elite genotypes and clusters, facilitating strategic selection in wheat breeding programs.

REFERENCES

Abdelghany, M., Makhmer, K., Zayed, E., Salama, Y. and Amer, K. 2023. Genetic variability, principal components and cluster analysis of twenty-eight egyptian wheat genotypes. *Scientific Journal of Agricultural Sciences*, **5**(1): 107-118**.**

Baranwal, D.K., Mishra, V.K. and Singh, T., 2013. Genetic diversity based on cluster and principal component analyses for yield and its contributing characters in wheat (Triticum aestivum L.). *Madras Agric. J*, **100**(4-6), pp.320-323.

Dervishi, A., Rumano, M., Ruzi, P. and Çakalli, A. 2022. The genetic diversity and variation in crude protein content of wheat (*Triticum aestivum L.*) promising cultivars for breeding in Albania. *Agronomy Science*, **77**(3): 79-88. [Cross Ref]

Elahi, T., Pandey, S. and Shukla, R.S. 2021. Agro-morphological diversity in promising wheat genotypes grown under restricted irrigated condition. *Electronic Journal of Plant Breeding*, ***12***(3):643-651. [Cross Ref]

Hailegiorgis, D., Mesfin, M. and Genet, T. 2011. Genetic divergence analysis on some bread wheat genotypes grown in Ethiopia. *Journal of Central European Agriculture*, **12**(2): 344-352. [Cross Ref]

Hartigan, J.A. and Wong, M.A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, **28**(1): 100-108. [Cross Ref]

Janmohammadi, M., Movahedi, Z. and Sabaghnia, N. 2014. Multivariate statistical analysis of some traits of bread wheat for breeding under rainfed conditions. *Journal of Agricultural Sciences, Belgrade*, **59**(1):1-14. [Cross Ref]

Khan, M.A., Anjum, A., Bhat, M.A., Padder, B.A., Mir, Z.A. and Kamaluddin, M., 2015. Multivariate analysis for morphological diversity of bread wheat (*Triticum aestivum* L.) germplasm lines in Kashmir valley. *Journal of Science*, **5**(6): 372-376.

Khare, V., 2022. Multivariate analysis and role of direct–indirect effect for yield and its component traits in bread wheat (*Triticum aestivum* L.). *Electronic Journal of Plant Breeding*, **13**(2): 447-454.

Khodadadi, M., Dehghani, H. and Fotokian, M.H. 2014. Genetic diversity of wheat grain quality and determination the best clustering technique and data type for diversity assessment. *Genetika*, **46**(3): 763-774. [Cross Ref]

Kumar, J., Kumar, A., Singh, S.K., Singh, L., Kumar, A., Chaudhary, M., Kumar, S. and Singh, S.K., 2016. Principal component analysis for yield and its contributing traits in bread wheat (*Triticum aestivum*) genotypes under late sown condition. *Current Advances in Agricultural Sciences,* **8**(1): 55-57. [Cross Ref]

Kumar, S., Gupta, V., Yadav, S.S., Mamrutha, H.M., Singh, S.K., Chatrath, R. and Singh, G.P., 2021. Multivariate analysis in wheat germplasm captures variability for agro-morphological and physiological traits. *Indian Journal of Agricultural Sciences,* **91** (9): 1322–27. [Cross Ref]

Mao, H., Jiang, C., Tang, C., Nie, X., Du, L., Liu, Y., Cheng, P., Wu, Y., Liu, H., Kang, Z. and Wang, X. 2023. Wheat adaptation to environmental stresses under climate change: Molecular basis and genetic improvement. *Molecular Plant*, **16**(10): 1564-1589. [Cross Ref]

Marino, S. and Alvino, A. 2020. Agronomic traits analysis of ten winter wheat cultivars clustered by UAV-derived

vegetation indices. *Remote Sensing*, **12**(2): 249. [Cross Ref]

Mellingers, J.S. 1972. Measures of genetic similarity and genetic distance. VII. *Studies in Genetics*, 145-153.

Mishra, C.N., Tiwari, V., Satish-Kumar, S.K., Gupta, V., Kumar, A. and Sharma, I. 2015. Genetic diversity and genotype by trait analysis for agromorphological and physiological traits of wheat (*Triticum aestivum* L.). *SABRAO J. Breed. Genet*., **47** (1) 40-48.

Mohammadi, S.A. 2002. August. Statistical methods in genetics. In *Proceedings of the Sixth International Conference of Statistics. August* (26-28).

Sarfraz, Z., Shah, M.M., Iqbal, M.S., Nazir, M.F., Al-Ashkar, I., Rehmani, M.I.A., Shahid Iqbal, M., Ullah, N. and El Sabagh, A. 2021. Rendering multivariate statistical models for genetic diversity assessment in A-genome diploid wheat population. *Agronomy*, **11**(11): 2339. [Cross Ref]

Singh, G., Kulshreshtha, N., Singh, B.N., Setter, T.L., Singh, M.K., Saharan, M.S., Tyagi, B.S., Verma, A. and Sharma, I. 2014. Germplasm characterization, association and clustering for salinity and waterlogging tolerance in bread wheat (*Triticum aestivum*). *The Indian Journal of Agricultural Sciences*, **84**(9): 1102-10. [Cross Ref]

Sthapit, S.R., Marlowe, K., Covarrubias, D.C., Ruff, T.M., Eagle, J.D., McGinty, E.M., Hooker, M.A., Duong, N.B., Skinner, D.Z. and See, D.R., 2020. Genetic diversity in historical and modern wheat varieties of the US Pacific Northwest. *Crop Science*, **60**(6): 3175-3190. [Cross Ref]

Šućur, R., Mladenov, V., Banjac, B., Trkulja, D., Mikić, S., Šumaruna, M. and Börner, A., 2024. Phenotypic marker study of worldwide wheat germplasm. *Italian Journal of Agronomy*, **19**(1):100002. [Cross Ref]

Yasin, B.A., Shubhra, S., Abdu, M. and Shiferaw, G. 2024. Assessing genetic diversity in bread wheat (*Triticum aestivum* L.) using D2 statistical analysis. *Asian Journal of Dairy and Food Research*, 1-6. [Cross Ref]