RESEARCH ARTICLE

SABRAO Journal of Breeding and Genetics
55 (6) 1963-1971, 2023
http://doi.org/10.54910/sabrao2023.55.6.10
http://sabraojournal.org/
pISSN 1029-7073; eISSN 2224-8978

PCA-BASED MULTIVARIATE APPROACH FOR SEGMENTATION OF VARIANCE IN INDIAN MUSTARD (*BRASSICA JUNCEA* [L] CZERN & COSS)

S. GHOSH[1], H. AVINASHE[1*], N. DUBEY[1], G.P. SHARADHI[1], K. DANALAKOTI[1], S. SACHAN[2], and S. CHOUDHARY[3]

[1]Department of Genetics and Plant Breeding, Lovely Professional University, Phagwara, Punjab, India
[2]Department of Agricultural Economics and Extension, Lovely Professional University, Phagwara, Punjab, India
[3]Department of Plant Pathology, School of Agriculture, Lovely Professional University, Phagwara, Punjab, India
*Corresponding author's email: havinashe@gmail.com
Email addresses of co-authors: sreejayaghosh873@gmail.com, nidhi.19843@lpu.co.in, sharadhigp@gmail.com, keerthanakeerthi1652@gmail.com, sharad.19461@lpu.co.in, sneha.28385@lpu.co.in

SUMMARY

The presented study assessed 27 Indian mustard (*Brassica juncea* [L.] Czern & Coss) genotypes for 13 quantitative and one biochemical trait by a PCA-dependent multivariate analysis, which split the total divergence into 14 accountable principal components among them. The first five PCs, which showed an Eigenvalue of more than one, contributed significantly to about 76.35% of the total divergence. Interpretation of the PCA-Biplot declared that the genotypes, namely, two and 11, mostly subsidized the overall variance, which is about 20% and 17%, respectively. The study of the bar plot of contribution percentage presented relevance with the PCA-biplot study, which again indicated the significance of genotypes two and 11 in total variance. The biplot analysis revealed that traits, NSB and SYP, contribute appreciably to the variation of the genotypes placed in the second coordinate, showing a detrimental interference with PC 1 and positive interference with PC 2. Likewise, in the third coordinate, traits, such as LMS, influenced the variance of the genotypes of that coordinate. The percentage contribution study for features in the first and second PCs revealed that characteristics, such as, LMS, PH, NSMS, BYP, NSS, and SYP participated prominently to accelerate the total variance. This research work can be a groundwork for further crop improvement featuring the studied materials.

Keywords: Indian mustard, diversity, multivariate analysis PCA, biplot analysis, oil content

Key findings: The study summarized genotypes, BR-23, Parbati mustard, and PDZ-1, as sponsoring maximum to the entire variance presented, and traits, such as, LMS, PH, NSMS, BYP, NSS, and SYP, predominantly attributed to the variation represented by these genotypes.

Communicating Editor: Dr. Quaid Hussain

Manuscript received: May 19, 2023; Accepted: September 4, 2023.

Citation: Ghosh S, Avinashe H, Dubey N, Sharadhi GP, Danalakoti K, Sachan S, Choudhary S (2023). PCA-based multivariate approach for segmentation of variance in Indian mustard (*Brassica juncea* [L] Czern & Coss). *SABRAO J. Breed. Genet.* 55(6): 1963-1971. http://doi.org/10.54910/sabrao2023.55.6.10

INTRODUCTION

Brassica juncea (L.) Czern & Coss, commonly known as Indian mustard, belonging to the family Brassicaceae (Cruciferae), is one of the most significant oilseed brassica crops grown in India and globally. More or less a hundred species have come under the genus *Brassicaceae,* among which the rapeseed-mustard group has exclusive cultivation as oilseed crops worldwide (Singh and Singh, 2018). Indian mustard is an amphidiploid (AABB) believed to have evolved as a consequence of interspecific cross or crosses between plants of *B. rapa (AA)* by *B. nigra (BB)* (Nagaheru, 1935; Olsson, 1960). It falls under the category of one of the earliest domesticated crops, as various primeval literatures affirmed (Allchin, 1969). Mustard seeds contain around 28%–36% protein with 38%–45% oil, which is rich in linoleic acid and oleic acids and a potent source of omega-3 and omega-6 fatty acids, rendering it a healthier option for edible oil (Mustafa *et al*., 2022; Das *et al*., 2022). Mustard use is not just an oilseed but rather an essential condiment in culinary practices. India, Canada, Pakistan, Hungary, Nepal, Great Britain, and the United States are chief contributors to the total production of mustard seeds. In India, 86.72% of the area under mustard cultivation is mainly from the states, namely, Madhya Pradesh, Rajasthan, Uttar Pradesh, Haryana, Assam, and West Bengal, accounting for 89.53% of the total production (DRMR, 2022).

India produces only around 10.6 million tons of edible oil annually while spending a foreign exchange surpass of USD 18.99 billion in the fiscal year ending March 31, 2022, to import 14–14.5 million tons of edible oil, making it the seventh-highest importer of edible oil (DFPD, 2021). Crop improvement in mustard can be a step toward making India self-sufficient in oil production. Improving the mustard crop is challenging due to the complicated nature of yield and yield-contributing characteristics and their inheritance. There is evidence of additive and non-additive gene activity in the succession of mustard characteristics (Singh *et al*., 2016; Verma *et al*., 2016; Manjunath *et al*., 2017). Genetic divergence research aids in developing cultivars with a higher yield, broader adaptability, desirable traits, and pest and disease resistance through the selection of superior genotypes and parents for a crossing program (Ram *et al*., 2015; Govindaraj *et al*., 2015). Thus, the first step in crop improvement is to understand the variation present in the existing material. In that regard, the current study proceeded with a set of n genotypes of Indian mustard to evaluate the extent of genetic diversity through the multivariate principal component analysis.

A multivariate-based analysis has immense potential in justifying variances among different genotypes and designating the contribution of distinct variables; yet, more studies need to focus on this aspect in plant breeding and genetics. PCA uses different approaches, such as, dimension reduction, identification of potential variables, and description of component variations. PCA groups the genotypes and traits in such a way as to accelerate the selection of superior genotypes and valuable characteristics based on their contribution to total variation, converting the vast phenotypic data into a convenient size without neglecting much information (Hamman, 1972). It is essential to delineate their exact contribution and the quality of the contribution to understanding the divergence apparent among the genotypes. The idea about the supportive traits is that they account for the performance and variance present in the genotypes. Accounting for the immense opportunities of PCA in plant breeding studies, the current manuscript has offered various aspects of it in understanding the divergence present in 27 Indian mustard genotypes. Likewise, Yadav *et al*. (2022) evaluated the principal component in 18 genotypes of Indian mustard to find out the magnitude of influence toward diversity.

Table 1. Genotypes used in the study.

No.	Genotypes	No.	Genotypes
1	Bhagirathi	15	Pusa Jai Kisan
2	BR-23	16	Pusa Karishma
3	Durga	17	Pusa Mustard-24
4	Durga Mani	18	Pusa Mustard-27
5	Gujarat Mustard-1	19	Pusa Mustard-28
6	Gujarat Mustard-2	20	RH-119
7	JD-6	21	RH-30
8	KBS-3	22	RH-701
9	Kranti	23	RNG-73
10	NRCHB-1	24	Rohini
11	Parbati Mustard	25	SMR-9
12	PDZ-1	26	Urvashi
13	Pusa Bold	27	Vaibhav
14	Pusa Jagannath		

MATERIALS AND METHODS

Location of the experiment

Evaluation of the experimental material had a trial conducted at the area (31.25° N and 75.707° E) of the Agriculture Research Farm under the Faculty of Genetics and Plant Breeding, School of Agriculture, Lovely Professional University, Phagwara, Punjab. The duration of cropping was around five months, from mid-October until March of the following year.

Experimental design and material

The experimental material consisted of 27 genotypes of Indian mustard of ICAR - DRMR (Table 1). The materials, laid out in a randomized block design, had three replications and involved four rows per genotype, with a row measurement of 60 cm and 15 cm plant spacing. All the recommended practices remained during the cropping period.

Recorded observations

The observations' documentation ensued on randomly selected five competitive plants of each genotype from each replication. Noting evaluations for 14 total traits occurred. Traits, such as, plant height (PH), length of the main shoot (LMS), the number of primary branches (NPB), the number of secondary branches (NSB), the number of siliqua on the main shoot (NSMS), the length of siliqua on the main shoot (LSMS), number of seeds siliqua^{-1} (NSS), 1000-seed weight (TAW), biological yield plant^{-1} (BYP), seed yield plant^{-1} (SYP), and harvest index (HI) incurred measuring on an individual plant basis, while attributes, days to 50% flowering (DFF) and days to maturity (DM) achieved scrutiny on a block basis. Working out oil content (OC) employed the help of a Soxhlet apparatus using n-hexane as a solvent.

Statistical evaluation

The principal component analysis (PCA) execution helped estimate the contribution and quality accountable for the dominance of the variance observed in the study (Ingebriston and Lyon, 1985). The raw data's first calibration was to match unitary variables with their analysis based on their co-variances. The software used for the data analysis was Agricolae, R package version 1.3-5, and Metan R package (Olivoto and Lúcio, 2020; de Mendiburu, 2021).

RESULTS AND DISCUSSION

Analysis of principal components

Principal component analysis is a method of compression and binarization of a vast data set in a manageable and unambiguous manner.

PCA is a conventional multivariate method of reduction in the number of explanatory variables, which solves the problem of collinearity in the spread of data (Bair *et al*., 2006). PCA lays out the data variables in a few linear combinations to epitomize the data (Maitra and Yan, 2008). A PCA generally runs with the help of a covariance matrix of different pairs of characteristics. As an Eigenvector-based analysis, Eigenvalues or Eigenvectors continue calculating from the covariances. Based on these, the Eigenvalues' data set acquires characterization into specific orthologous clusters of variables called principal components. The first few components are accountable for most of the variability present in the samples.

The study calculated a principal component analysis of 14 quantitative traits of 27 Indian mustard varieties. It divided the total variance into 14 contributing principal components. Figure 1 depicts the principal components in the form of corrplots for their easy visualization. It is evident from the illustration that PC 1 is the highest contributing component of the total variance, and the intensity of contribution decreases in the following principal components. PC 10 still showed some amount of visibility in terms of involvement. Afterward, going further, the participation of PC 11 to PC 14 was quite negligible. The percentage involvement of the first 10 PCs sustained elaboration in the Scree Plot in advance.

Scree plot analysis

Delineating the contribution of these PCs in the scree plot appears in Figure 2. The first principal component explains the highest variance of 29%, followed by PC 2 to PC 10, which explains 15%, 13%, 10.2%, 9.2%, 5.9%, 5.3%, 5.2%, 2.9%, and 2.6% of the variance, correspondingly. The 10 PCs altogether accounted for 98.2% of the total available variance. The first four principal components subsidized around 67.2% of the complete variance. In Table 2, all the Eigenvalues, variations, and cumulative variances are available. As a rule of Gutten's lowest limit principle, if exempting Eigenvalues of less than one, then divergences of PC1 to PC5 only need consideration. The first five principal components yielded Eigenvalues of 4.06, 2.09, 1.82, 1.42, and 1.28, respectively, and subsidized a total cumulative variance of 76.34%. The first two PCs contributed about 44% of the total variance. Thus, the first two PCs underwent further use to exemplify PCA-Biplot analysis and the contribution and quality of representation of genotypes and variables to the total variation.

Thakral *et al*. (2015) found a similar trend in the analysis of 60 Indian mustard genotypes where out of 16 PCs, selecting the first 11 PCs was according to their Eigenvalues. These PCs together contributed to about 75% of the total variance present in the study, wherein the first PC alone explained 13.19% of the total variance.

Pankaj *et al*. (2017) observed a parallel result while working on 43 genotypes of Indian mustard, where primitive nine PCs selection was due to their showing an Eigenvalue more than one. Together, these PCs subsidized around 77% of the total variability and the first PC contributed to 16.65% of the entire variance and the contribution trail decreased progressively in the later PCs.

Godara *et al.* (2022) also found comparable results while working on 310 genotypes of Indian mustard. In total, 11 PCs formed had four PCs taken into consideration based on Eigenvalues, which accounted for 65% of the complete variance. A 24.16% of the variation has a contribution from the first PC alone. Then, similarly, the involvement faded gradually in further PCs.

Analysis of PCA-Biplot

The biplot constitution took the first two PCs into account (PC 1 and PC 2) and plotted along the *x-axis* and *y-axis,* respectively, as depicted in the biplot in four coordinates (Figure 3). The first coordinate (C-1) contained four genotypes, viz. 2, 11, 18, and 27 (Table 1), and plotted in the positive direction of both PC 1 and PC 2. The trait TWA correlates with these genotypes. Among these four genotypes, 2 and 11 are outlying at the edge, showing maximum divergence.

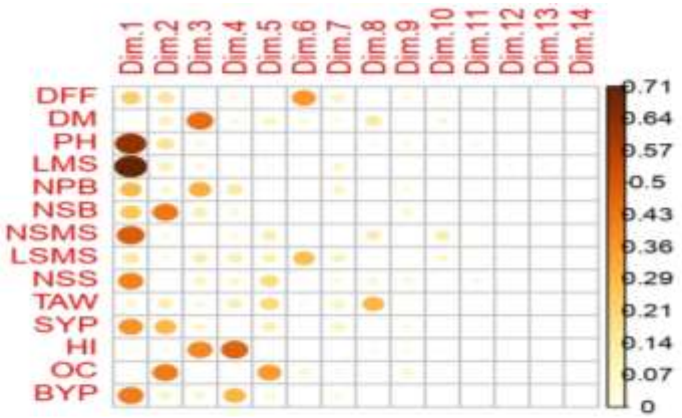

Figure 1. Corrplot of the contribution of each trait in each Principal Component.

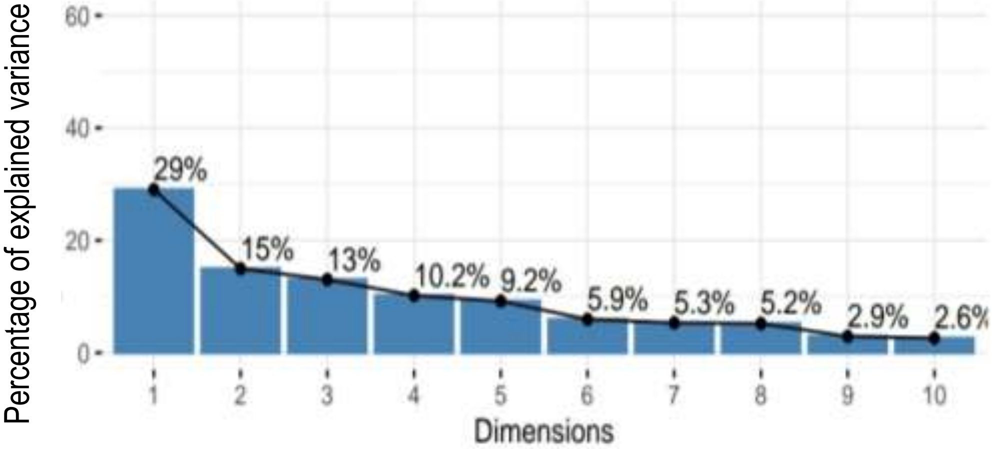

Figure 2. Scree plot on variability explained by each component of 27 mustard genotypes.

Table 2. Eigenvalues, % variance, and cumulative variance.

PCs	Eigenvalue	Variance percent	Cumulative Variance percent
PC 1	4.062087745	29.01491247	29.01491
PC 2	2.093835762	14.95596973	43.97088
PC 3	1.822594064	13.01852903	56.98941
PC 4	1.425287428	10.18062449	67.17004
PC 5	1.285141789	9.17958421	76.34962
PC 6	0.830883554	5.93488253	82.28450
PC 7	0.746433344	5.33166674	87.61617
PC 8	0.727666485	5.19761775	92.81379
PC 9	0.407560820	2.91114872	95.72494
PC 10	0.364552675	2.60394768	98.32888
PC 11	0.125638905	0.89742075	99.22630
PC 12	0.061464645	0.43903318	99.66534
PC 13	0.044857705	0.32041218	99.98575
PC 14	0.001995078	0.01425056	100.00000

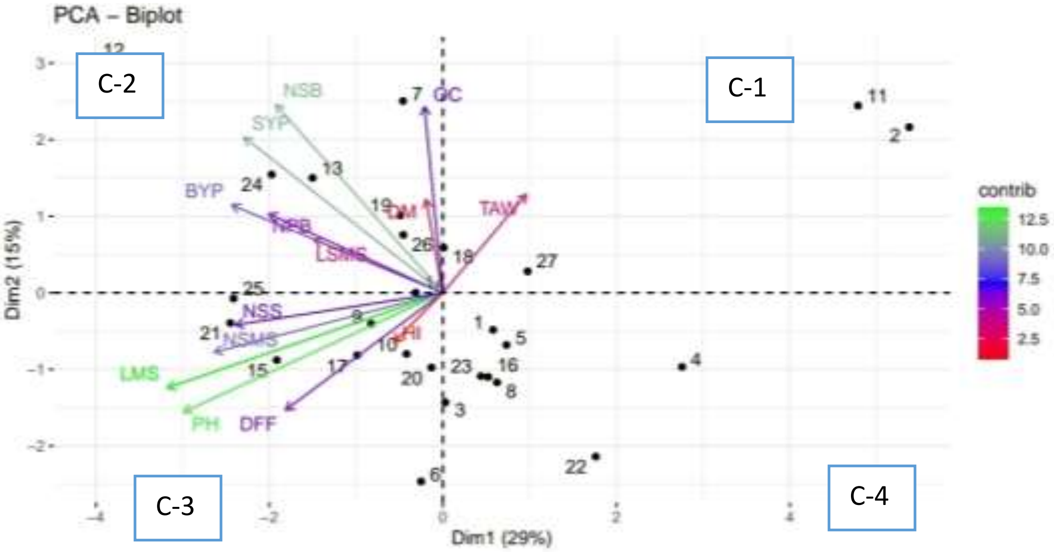

Figure 3. PCA-Biplot for 27 Indian mustard genotypes with 14 quantitative traits along the first two Principal Components.

Seven genotypes, namely, 1, 7, 12, 13, 19, 24, and 26 (Table 1), plotted in the second coordinate, depicted a negative value toward PC 1 but a positive value toward PC 2. The associative variables are NSB, SYP, BYP, NPB, LSMS, DM, and OC. Variables, viz., NSB and SYP, are outlying maximum from the center, indicating their consistent contribution toward the variation in these genotypes. Plotting of genotype 12 moves away from the center, showing maximum divergence among the genotypes of these coordinates.

Eight genotypes, viz., 6, 9, 10, 15, 17, 20, 21, and 25, have fallen under the third coordinate (C-3), and they have exhibited negative values for the first two PCs (Table 1). A set of variables, such as, LMS, PH, DFF, NSMS, NSS, and HI, is the accompanying variability of these genotypes. Among all variables, LMS scattered the most from the axis, thus contributing utmost to the divergence of coordinated genotypes.

Seven genotypes, i.e., 1, 3, 4, 5, 8, 22, and 23, showed positivity toward PC 1 while negativity toward PC 2 plotting in a location in the fourth coordinate (Table 1). Genotypes 22 and 4, located away from the center, contributed to the total variance. Similar results were evident in the research work of Saleem *et al*. (2017), where the traits DFF, DM, PH, NSMS, NPB, NSB, SYB, and LMS explained negative contribution in the first PC. In addition, traits, namely, LSMS, SYP, TAW, and DM, contributed positively, and DFF, PH, and LMS contributed negatively in the second PC of the mentioned study.

Similarly, various traits have contributed distinctly to different PCs and partitioned variance in those dimensions in the research work of Godara *et al*. (2022). It indicated that attributes, such as, TAW, LMS, NSS, and LSMS contributed negatively, while others, such as, NPB, PH, SYP, DM, OC, and NSB, showed positive relations in the first PC. Further, the qualities, i.e., LSMS, NSB, NSS, SYP, and NPB, provided negative contributions, while TAW and PH characteristics showed positive contributions toward the variation in the second PC.

Contribution of variables along with their quality toward variance

In the bar plot of the contribution percentage of variables, six traits exceeded the expected average threshold, namely, LMS, PH, NSMS, BYP, NSS, and SYP (Figure 4.a). The variable LMS presented the highest contribution to the total variance, followed by PH, NSMS, BYP, NSS, SYP, NPB, NSB, DFF, LSMS, TAW, HI, OC,

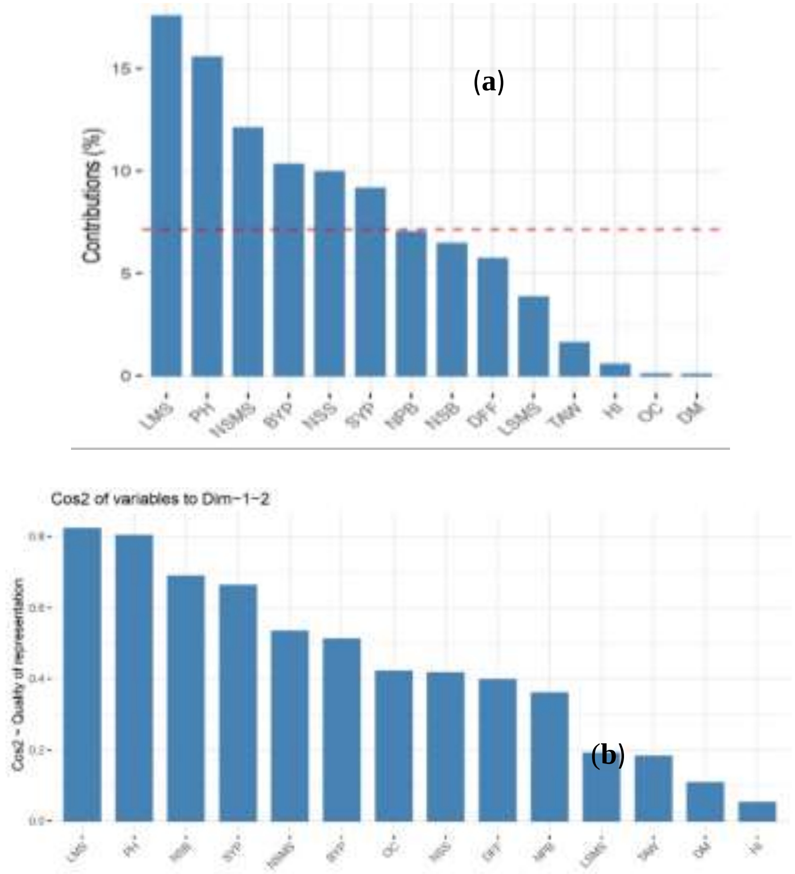

Figure 4. (a) Bar plot of the contribution percentage (red-dotted line indicates the expected average contribution) and (b) Bar plot of the quality of representation of 14 variables to Principal Components 1-2.

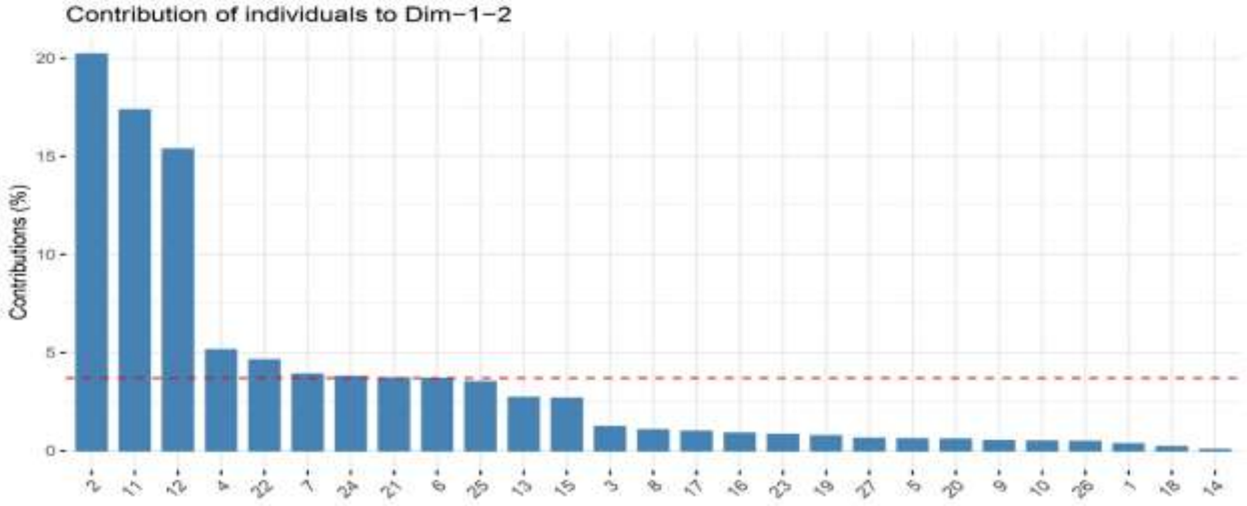

Figure 5. Bar plot of the contribution percentage of 27 Indian mustard genotypes to Principal Components 1-2 percentage (red-dotted line indicates the expected average contribution).

and DM (Figure 4.a). Even though the trait LMS provided the highest quality of contribution, followed by PH, other variables indicated a different trend in the case of quality of representation (Figure 4.b).

Contribution of genotypes toward variance

Similarly, in the bar plot of the contribution percentage for genotypes, genotype 2 contributed the maximum toward the entire divergence, followed by genotypes 11, 12, 4, 22, 7, 24, 21, 6, 25, 13, 15, 3, 8, 17,16, 23, 19, 27, 5, 20, 9, 10, 26, 1, 18, and 14 (Table 1, Figure 5). Six genotypes were exposed to surpass the average contribution toward variance. Among them, three genotypes, namely, 2, 11, and 12, exhibited 15%–20% variation each, and the other three genotypes, i.e., 4, 22, and 7 contributed about 4%–5% individually of the total variance. In addition, the genotypes, such as, 24, 21, and 8 displayed average contributions to the divergence.

CONCLUSIONS

The investigation divided the total variability into 14 PCs, out of which five PCs contributed significantly to the total variance**.** PCA-biplot helped to visualize the assortment of different genotypes in terms of their diversity, which revealed genotypes 2 and 11 subsidizing most of the overall divergence. The bar plot of the contribution percentage also supported the previous statement and showed that genotypes 2, 11, 12, 4, 7, and 22 exceeded the expected average cut-off of contribution. A discovery implied that six variables, LMS, PH, NSMS, BYP, NSS, and SYP, surpassed the average contribution range toward the entire diversity. It might be reliable to utilize the variation among the genotypes for further selection or in a hybridization program.

REFERENCES

Allchin FR (1969). Early cultivated plants in India and Pakistan. In: P.J. Ucko and G.W. Dimbleby (eds.). The Domestication and Exploitation of Plants and Animals. Duckworth and Co., London.

Bair E, Trevor H, Paul D, Robert T (2006). Prediction by supervised principal components. *J. Am. Stat. Assoc*. 101(473): 119-137(19).

Das G, Tantengco OAG, Tundis R, Robles JAH, Loizzo MR, Shin HS, Patra JK (2022). Glucosinolates and Omega-3 fatty acids from mustard seeds: Phytochemistry and pharmacology. *Plants* 11(17): 2290.

de Mendiburu F (2021). Agricolae: Statistical Procedures for Agricultural Research. *R package version* 1.3-5, https://CRAN.R-project.org/package=agricolae.

DFPD (2021). Department of Food and Public Distribution, Oil Division. Importance of Edible Oils in the Country's Economy. Available: https://dfpd.gov.in/oil-division.htm.

Godara P, Kumar S, Kumar D (2022). Evaluation of genetic variation in Indian mustard (*Brassica juncea* L. Czern and Coss) using multivariate techniques. *J. Agric. Res. Technol*. *47*(3): 344.

Govindaraj M, Vetriventhan M, Srinivasan M (2015). Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. *Genet. Res. Int*. 2015: 431487. https://doi.org/10.1155/2015/431487.

Hamman HK (1972). Utilization of multivariate statistics for discriminating among flue-cured tobacco varieties. Technical Bulletin Number 212, North Carolina Agricultural Experimental Station (Raleigh, NC).

ICAR-DRMR (2022). ICAR-Directorate of Rapeseed-Mustard Research. Available: https://www.drmr.res.in/about_rmcrop.php.

Ingebriston SE, Lyon RJP (1985). Principal components analysis of multi-temporal image pairs: *Int. J. Remote Sens.* 6(5): 687-696.

Maitra S, Yan J (2008). Principle component analysis and partial least squares: Two-dimension reduction techniques for regression. *Casualty Actuarial Soc. 79*: 79-90.

Manjunath H, Phogat D, Kumari P, Singh D (2017). Genetic analysis of seed yield and yield attributes in Indian mustard (*Brassica juncea* L. Czern and Coss.). *Electr. J. Plant Breed*. 8: 182-186. 10.5958/0975-928X.2017.00026.6.

Mustafa H, Mahmood T, Bashir H, Hasan E, Din A, Habib S, Altaf M, Qamar R, Ghias M, Bashir M, Anwar M (2022). Genetic and physiological aspects of silique shattering in rapeseed and mustard. *SABRAO J. Breed. Genet*. *54*(2): 210-220.

Nagaheru U (1935). Genome analysis in Brassica with special reference to the experimental formation of *B. Napus* and peculiar mode of fertilization. *Jap. J. Bot.* 7: 389-452.

Olivoto T, Lúcio AD (2020). Metan: An R package for multi-environment trial analysis. *Methods Ecol. Evol. 11*(6): 783-789. doi:10.1111/ 2041-210X.13384. https://doi.org/10.1111/2041-210X.13384.

Olsson G (1960). Species crosses within the genus *Brassica*. II. Artificial *Brassica juncea* Coss. *Hereditas.* 46: 171-222.

Pankaj R, Avtar R, Kumari N, Jattan M, Rani B (2017). Multivariate analysis in Indian mustard genotypes for morphological and quality traits. *Electr. J. Plant Breed. 8*(2): 450-458.

Ram KB, Singh VV, Singh BK, Priyamedha, Kumar A, Singh D (2015). Comparative tolerance and sensitive response of Indian Mustard (*Brassica juncea* L. Czern and coss) genotypes to high temperature stress. *SABRAO J. Breed. Genet*. 47(3): 315-325.

Saleem N, Jan SA, Atif MJ, Khurshid H, Khan SA, Abdullah M, Jahanzaib M, Ahmed H, Ullah SF, Iqbal A, Naqi S, Ilyas M, Ali N, Rabbani MA (2017). Multivariate-based variability within diverse Indian mustard (*Brassica juncea* l.) genotypes. *Open J. Genet.* 7: 69-83. https://doi.org/10.4236/ojgen.2017.72007.

Singh M, Singh VV (2018). Physiological approaches for breeding drought-tolerant brassica genotypes. *SABRAO J. Breed. Genet.* 50: 360-372.

Singh VV, Gurjar N, Ambawat S, Yadav S, Singh BK, Ram B, Meena ML, Singh BR, Singh S, Singh D (2016). Morphological and molecular diversity among full-sib progenies of Indian mustard (*Brassica juncea* L.). *SABRAO J. Breed. Genet.* 48(2): 180-188.

Thakral NK, Avtar R, Singh A (2015). Evaluation and classification of Indian mustard (*Brassica juncea* L.) genotypes using principal component analysis. *J. Oilseed Brassica*. 6(1): 167-174.

Verma S, Singh VV, Meena MI, Rathore SS, Ram B, Singh S, Garg P, Singh BR, Gurjar N, Ambawat S, Singh D (2016). Genetic analysis of morphological and physiological traits in Indian mustard (*Brassica juncea* L.). *SABRAO J. Breed. Genet.* 48(4): 391–401.

Yadav D, Singh L, Jafri SKF, Singh A, Singh V (2022). Study on genetic variability and principal component analysis in Indian mustard [*Brassica juncea* (L.) Czern and Coss]. *Pharma Innov.* 11(4): 2166-2169.