Received: 11 Jun 2022 | Revised: 14 May 2023 | Accepted: 29 May 2023

Electronic Journal of Plant Breeding

Research Note

Principal Component Analysis (PCA) and hierarchial clustering in tobacco (*Nicotiana tabacum* L.) for yield and yield attributing traits

B. P. Maruthi Prasad[1] , B. R. Patil*[1], D. Geeta[2] and P. S. Matiwade[3]

[1]Department of Genetics and Plant breeding, University of Agricultural Sciences, Dharwad – 580005.
[2]AINP (T), Agricultural Research Station, Nipani. UAS, Dharwad-580005
[3]Agronomy,Agricultural Research Station, Nipani. UAS, Dharwad-580005
***E-Mail:** patilbr@uasd.in

Abstract
Multivariate statistical analysis techniques like Principal Component Analysis (PCA) and heirarchial clustering were used to evaluate Genetic diversity among 246 genotypes of Tobacco for six major yield and yield-related traits. The hierarchial clustering indicated that all the genotypes were clustered into eight major groups. The cluster III had the maximum number of genotypes with highest intra cluster distance and cluster IV and VIII showed maximum inter cluster distance indicating that the characterized tobacco genotypes in these clusters has high potential for various breeding goals. Principal component analysis and genotype by trait biplot analysis showed that the first four components accounted for 94.75 per cent of the total variation, with principal component 1 (PC1) accounting for 55.96 per cent and PC2 for 20.97 per cent of the total variation. The high yielding genotypes with other yield attributes identified in this study would offer valuable genetic material for breeding elite tobacco varieties.

Keywords: Tobacco genotypes, PCA, Rotated component matrix, Eigen value.

India is the second-largest country in the production and export of Tobacco with a total production of 8,01,420 tonnes from 4,04,290 hectares with a productivity of 1982 kg per hectare. In Karnataka tobacco produced in an area of 90,000 ha having productivity of 922 kg per hectare with production of 83,000 tonnes (Anon., 2020). The tobacco grown in Nippani, Karnataka and surrounding regions is one of the world's finest tobacco which is rolled into beedis. Since the area under tobacco is declining, there is a need for increasing the production and productivity of the tobacco by identifying desirable genotypes. Further genetic base of the crop can be expanded by the introgression of lines with large genetic diversity.

To start crop improvement program, a detailed knowledge of the amount of genetic diversity present in the germplasm for various characters is necessary for selecting parents in hybridization, since various plants are expected to yield high hybrid vigour. The principal component analysis (PCA) is a well-known data reduction technique in modern data analysis because it is a simple and non-parametric methodology for extracting relevant information from data sets. It was initially proposed by Pearson (1901) and developed by Hotelling (1933). PCA provides a roadmap to reduce a complex data set to a lower dimension while retaining most of the variation in the data set (Abdi and Williams, 2010). PCA accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal (Ringer, 2008). By using a few components each sample can be represented by relatively few numbers instead of values for thousands of variables. It extracts

the low-dimensional set of features by taking a projection of irrelevant dimensions from a high-dimensional data set with a motive to capture as much information as possible (Anderson, 1972 and Morrison, 1982). The rotation procedure seeks to establish a simpler relationship between the factors and the components that is more likely to correspond to constructs of interest (Gorsuch, 1983). Varimax orthogonal rotation tries to *max*imize *variance of the squared loadings in each factor* in SS. Hence its name (*var*iance). As the result, each factor has only few variables with large loadings by the factor. Varimax directly "simplifies" columns of the loading matrix and by that it greatly facilitates the interpretability of factors. Varimax performs well mostly in combination with the *Kaiser's normalization* (equalizing communalities temporarily while rotating).

Hierarchical clustering plays a crucial role in plant breeding, enabling researchers to analyze and classify plant varieties based on their genetic similarities. This technique allows breeders to group plants into clusters based on their genetic profiles, helping to identify potential parents for crossbreeding and the selection of desirable traits. The objective of this research is to determine variation range among tobacco germplasm and to classify them into different clusters (Kalagare *et al.,* 2022).

The experiment was laid out in an Augmented block design with six blocks containing 242 genotypes and four checks (repeated two times in each block) *viz*., Vedaganga-1, A-119, Bhavyashree, and NBD 209 at a spacing of 1.0 × 0.75 m. The experimental material evaluated at Agricultural Research Station, Nipani, Karnataka, during *Rabi, 2020*. Nipani is close to the western ghats, with a mean annual rainfall of 730.1 mm. The observations were recorded for 6 characters *viz*., plant height, number of leaves, internode length, leaf length, leaf width, and leaf yield per hectare. The Principal component analysis (PCA) was carried out using the prcomp function of the *stats package, factoextra* (Kassambara and Mundt, 2020)*, factoMineR* (Le *et al*., 2008)*, ggplot2* (Wickham, 2016). Further Scree plot was used to access components or factors which explains most of the variability in the data and represents the values in descending order. The rotated component matrix is constructed using *psych* and *pals* packages of R studio.

Intra and inter-cluster distances: Tobacco is one of the important commercial crops in India and the knowledge on the genetic diversity of the crops is essentially important for future breeding. In the present study an attempt was made in this direction and 246 tobacco genotypes used in this study were clustered into 8 major clusters. Similar results were reported by Murphy *et al.* in 1987. Intra and inter-cluster distances are presented in **Table 1**. The values of intra cluster distances were lesser than inter cluster distance indicating that genotypes within cluster are less diverse among themselves. The nearest and farthest cluster from each cluster based on average cluster distance values are given in **Table 2**. The intra cluster distance ranged from 1.66 for cluster VII to 2.32 for cluster III. The highest inter cluster distance was observed between cluster IV and VIII with a distance of 6.83 indicating wider genetic diversity. The genotypes from these clusters (**Table 4**) can be used in future breeding programme to generate more variability. The lowest distance of 2.25 was observed between cluster I and VII indicating narrow genetic diversity (Hosseinzadeh *et al.,* 2015; Ann *et al.,* 1983; Wang *et al.,* 2014; Hemavathy *et al*., 2017).

Cluster means for different traits: The cluster mean for different traits are presented in **Table 3**. The genotypes in the cluster V recorded maximum mean value for internode length, leaf length and leaf width followed by cluster IV, while for traits plant height and number of leaves per plant VI exhibited a maximum mean value. The cluster IV exhibited maximum mean value for leaf yield per hectare. Cluster VII exhibited a lowest mean value for all the traits studied. The cluster V ranked first followed by cluster IV and VI, further these clusters also showed maximum mean values for the traits studied. Therefore, selecting desirable genotypes from these clusters will lead to maximum genetic gain.

Table 1. Average inter-cluster and intra cluster values in tobacco genotypes

Cluster	I	II	III	IV	V	VI	VII	VIII
I	**4.73**	5.34	6.79	5.65	4.74	5.84	4.87	7.01
II		**3.75**	8.28	6.99	5.91	5.22	4.00	4.48
III			**6.20**	6.82	6.76	8.13	8.20	9.92
IV				**3.94**	5.14	7.51	6.34	8.91
V					**3.08**	6.34	5.07	7.84
VI						**3.85**	5.60	6.29
VII							**3.90**	5.81
VIII								**3.43**

Table 2. The nearest and the distant clusters from each clusters

Cluster Number	Number of Genotypes	Nearest Cluster	Farthest Cluster
I	47	VII	VIII
II	33	VII	IV
III	49	I	VIII
IV	23	I	VIII
V	12	I	VIII
VI	45	II	IV
VII	22	I	VIII
VIII	15	II	IV

Table 3. Cluster mean for different quantitative and qualitative traits in tobacco genotypes

S. No.	Characters	Clusters															
		I		II		III		IV		V		VI		VII		VIII	
		Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank	Mean	Rank
1	Plant height (cm)	75.60	7	79.16	6	90.21	5	105.98	3	108.12	2	114.23	1	60.54	8	96.12	4
2	Number of leaves	16.97	2	11.65	7	11.79	6	12.88	4	12.81	5	19.57	1	11.58	8	16.75	3
3	Internode length	3.42	7	3.44	6	4.47	4	5.37	2	5.63	1	4.88	3	2.53	8	4.23	5
4	Leaf length	34.87	7	40.59	5	45.90	3	46.02	2	49.83	1	44.94	4	31.23	8	38.95	6
5	Leaf width	13.20	7	15.14	6	17.82	4	19.13	2	20.46	1	17.86	3	11.53	8	15.21	5
6	Yield per hectare	526.86	7	685.71	6	890.72	4	2036.00	1	1053.78	3	1299.77	2	408.88	8	890.28	5
			37		36		26		14		13		14		48		28
	Rank		6th		5th		3rd		2nd		1st		2nd		7th		4th

Distance between different pairs of genotypes: The distance between 246 tobacco genotypes in all possible combination is calculated using *dist* function of *stats* package by Euclidean method which computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix. The top 50 distance are selected from results and represented in **Table 5**. The results indicated that distance between Line 93-103-93 (88-47 x Sokh) and GT-4 was maximum (9.92). The genotypes represented in 'Positive value' column of **Table 5** *i.e.,* GT-4, KDH-959, NBD 325, Anand-2, S-20, NBD 324, NBD 209, ABD 95, NBD 324 and GT-5 are belong to cluster III and IV rather than cluster V which is having high mean value for all the traits. The reason may be that cluster III was having high intra cluster distance indicating maximum divergence within cluster. Further, these genotypes also showed positive PC value >1.0 (**Table 8**) indicating high values for the traits in these genotypes. The genotypes in 'Negative value' column of **Table 5** *i.e*., Line 93-103-93 (88-47 x Sokh), ABD 116, ABD 121, DWFC, F-7-124, Trabizonal, NPN 81, ABD 127, Bhagyalakshmi, A-428, 320-2-30-28-18-I, ABD 124 and 320-2-30-28-20-12 belong to cluster VIII which was having low mean values for all the traits and these genotypes also showed negative PC value >1.0 (**Table 9**) indicating low values for the traits in these genotypes.

The first principal component is a linear combination of original predictor variables which captures the maximum variance in the data set. It determines the direction of the highest variability in the data. The larger the variability captured in the first component, the larger is the information captured by the component. No other components can have variability higher than the first component. The tobacco genotypes included in the present investigation for PCA analysis comprised of 246 genotypes, which were studied for six yield attributing traits. All the characters were estimated based on principal component scores and presented in **Table 6**. Out of six, only four principal components (PCs) exhibited more than 0.5 eigen values and showed about 94.75 per cent of total variability. Scree plot explained the percentage of variance associated with each principal component obtained by drawing a graph between eigenvalues and

Table 4. Clustering pattern in tobacco genotypes

Clusters	Number of genotypes	Genotypes
I	47	Vedaganga-1, Bhavyashree, Kukumarthi, Kodani, V-54, Bankete A-1, Keliu-49, Pilliu-19, 103-9-101-28-31 (A-2 x Olor), ABD -7, ABD -15, ABD 30, TI-421, TI-55, NBD 259, 169-19-16 (88-47-Sokha), SB-154, 169-19-6 (88-47-Sokha), Bhavyashree, NBD-80-1, NBD-80-2, NBD 95, NBD 111, NBD 115, NBD 271, NBD 276, ABD 91, ABD 103, ArBD-4, ArBD-7, ArBD-8, ArBD-33, NyBD-3, NyBD-4, NyBD-59, ABD 118, ABD 138, ABD 151, ArBD-39, NBD 309, NBD 300, NBD 297, NBD 316, NBD 307, NBD 308, NBD 311, ABD 167.
II	33	A-119, BL 4-2, NBD 257, NBD 122, 35-19-39-24, NBD 236, ABD 60, ABD 61, ABD 67, ABD 70, ABD 73, ABD 77, ABD 84, ABD 90, ABD 125, ABD 128, ABD 130, ABD 152, ArBD-5, ArBD-9, ArBD-32, NyBD-56, G.M. Koyali, Line {34-30 x(A-119)2}103-6-1-40-22-34-26-35-22-25-2, Line 543-41-12-14 (RPK type), ABD 66, C-11, Sender patti special, HDBRG-LP-2, Thangam, NBD 289, NBD 315, ABD 163.
III	49	NBD 209, Keliu-20, Anand -23, Anand -119, GT-4, 103-9-101, Smyrna, Subhelav selection, KDH-959, NBD-239-4, NBD 209, RPK-1-2, NBD-48-1, 22-10-1 (11-47-Sokha), A-1-11-65, 169-2 (N&L), Jhakhari Rampur, AKBT-03-02, ABT-10, NBDS-57-1, NBD-57-1, NBD-71, NBD 85, NBD 136, NBD 154, NBD 155, ABD 54, ABD 104, ABD 146, NyBD-5, F-7-127, GT-5, NPN 64, NPN 75, NyBD 55, NBD 312, NBD 317, NBD 319, NBD 320, NBD 321, NBD 322, NBD 324, NBD 325, NBD 326, NBD 327, NBD 334, ABD 164, ABD 173, ABD 174.
IV	23	Anand-2, S-20, Gundsurti, Sanand local, Red Russion, V-58, Abirami, Jayalaxmi, 575-28-110, GT-5, GT-7, NBD 119, NBD 164, ABD 68, ABD 92, ABD 94, ABD 95, ABD 96, ABD 99, ABD 100, ABD 101, NBD 318, NBD 313.
V	12	783-51, 114-4 (RPK type), S-12, S-112, Dumbara, ABD 119, ABD 24, ABD 36, GT-9, NBD 43, NBD 159, NBD 314.
VI	45	NBD 260, NBD 239-4, NBD 261, K-20-Plule leaves, NBD 277, ABD 43, ABD 46, ABD 50, ABD 51, ABD 52, ABD 72, ABD 117, ABD 118, ABD 120, ABD 123, ABD 131, Pilliu-37, Line-169-119 (upper leaves long internode), B.S.P (Black Spangle parent), N.C.D. (Necrotic Crinkle Dwarf), Line 114-16 (Female parent of GT-4), Line 181-83-1 (S-20 xK-20), Line 132-2-2 {K-20 x Skh} x K-20, Line 543-37-38-24 (A-119 x Olar), Line 121-13-27-29 (108-15 x Olar), ABD 10, ABD 65, ABD 67, ABD-101 (GABT-11), Oriental, Bhagya, Jati Patti, 16-12-21-106-4-26, 320-2-80-25-84-10-I, Jati, Kunkumarthi, Vairam, NPN 63, NPN-65, NPN 66, NPN 73, NPN 30, NBD 290, NBD 302, NBD 323.
VII	22	ABD 62, ABD 69, ABD 71, ABD 78, ABD 79, ABD 87, ABD 102, ABD 107, ABD 109, ABD 110, ABD 111, ABD 112, ABD 113, ABD 115, ABD 132, Line-1-1, KL, Margadhan, ArBD-40, NBD 292, NBD 310, ABD 145.
VIII	15	ABD 116, ABD 121, ABD 124, ABD 127, Line 93-103-93 (88-47 x Sokh), Xanthi, Samsan, Trabizonal, Bhagyalakshmi, 320-2-30-28-18-I, 320-2-30-28-20-12, DWFC, F-7-124, NPN 81, A-428.

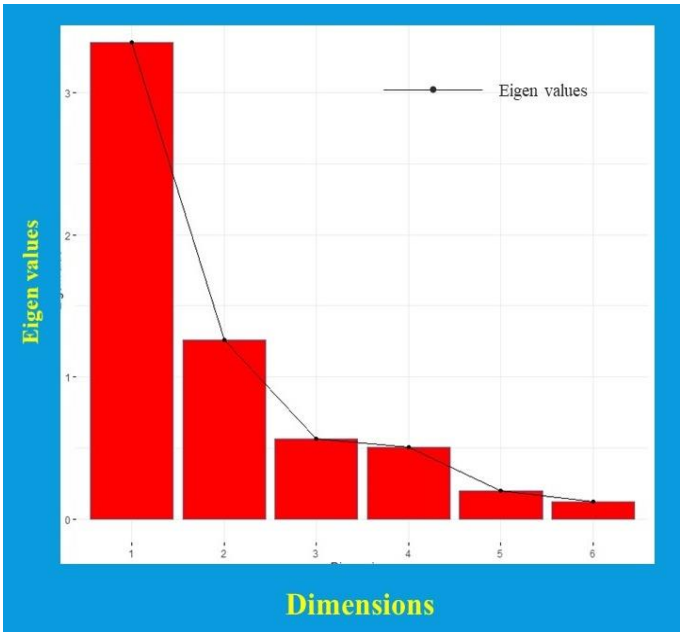

Fig. 1. Scree plot

Table 5. Top fifty distance between different pairs of genotypes

Distance between genotypes				Distance between genotypes			
Rank	Negative value	Positive value	Distance	Rank	Negative value	Positive value	Distance
1	Line 93-103-93 (88-47 x Sokh)	GT-4	9.92	26	Bhagyalakshmi	KDH-959	8.59
2	ABD 116	GT-4	9.89	27	ABD 121	S-20	8.56
3	ABD 121	GT-4	9.81	28	A-428	S-20	8.54
4	Line 93-103-93 (88-47 x Sokh)	KDH-959	9.77	29	A-428	KDH-959	8.54
5	DWFC	GT-4	9.44	30	Trabizonal	Anand-2	8.47
6	F-7-124	**GT-4**	9.42	31	Bhagyalakshmi	S-20	8.46
7	ABD 116	KDH-959	9.33	32	ABD 121	Anand-2	8.46
8	DWFC	KDH-959	9.22	33	320-2-30-28-18-I	GT-4	8.45
9	Line 93-103-93 (88-47 x Sokh)	**NBD 325**	9.18	34	DWFC	S-20	8.45
10	Samsan	GT-4	9.14	35	DWFC	Anand-2	8.45
11	ABD 121	KDH-959	9.12	36	A-428	GT-4	8.43
12	F-7-124	KDH-959	9.01	37	ABD 124	GT-4	8.42
13	Trabizonal	GT-4	8.99	38	320-2-30-28-20-12	GT-4	8.40
14	Line 93-103-93 (88-47 x Sokh)	Anand-2	8.91	39	DWFC	NBD 325	8.39
15	ABD 116	S-20	8.91	40	ABD 116	ABD 95	8.39
16	Line 93-103-93 (88-47 x Sokh)	S-20	8.87	41	320-2-30-28-18-I	KDH-959	8.39
17	Line 93-103-93 (88-47 x Sokh)	NBD 324	8.83	42	ABD 121	NBD 325	8.38
18	ABD 116	Anand-2	8.79	43	Bhagyalakshmi	GT-4	8.36
19	NPN 81	GT-4	8.76	44	ABD 116	NBD 324	8.36
20	Trabizonal	KDH-959	8.71	45	A-428	Anand-2	8.33
21	Trabizonal	S-20	8.70	46	Bhagyalakshmi	Anand-2	8.31
22	Line 93-103-93 (88-47 x Sokh)	NBD 209	8.67	47	Line 93-103-93 (88-47 x Sokh)	GT-5	8.29
23	ABD 127	GT-4	8.60	48	Thangam	GT-4	8.28
24	ABD 116	NBD 325	8.59	49	ABD 121	NBD 324	8.26
25	Samsan	KDH-959	8.59	50	Line 93-103-93 (88-47 x Sokh)	ABD 95	8.22

principal component number (**Fig. 1**). PC1 showed 55.96 per cent variability with eigen value 3.36 followed by PC2 with eigen value 1.26, which exhibited 20.97 per cent variability which then declined gradually in further PC's. From the graph, it is clear that the maximum variation was observed in PC1 which explains the maximum variance in comparison to other five PCs. So, selection of lines from this PC will be desirable. Thus, higher the explained variance, the higher will be the information contained in this components. These scores can be utilized to construct precise selection indices whose intensity can be decided by variability explained by each of the principal components. High PC scores for particular genotypes in a particular component denote high values for the variables in those particular genotypes. In 2012, Maji and Shaibu highlighted the significance of germplasm evaluation and characterization in the regular practices of plant breeders. They emphasized the usefulness of employing tools such as PCA (Principal Component Analysis), clustering, and multivariate statistical analysis. These analytical techniques proved valuable for estimating morphological diversity within and between collections of germplasm. Furthermore, these tools played a crucial role in assessing the potential breeding value of different varieties and identifying significant differences between germplasm types, as well as quantifying the extent of variation among various crop species. Kalivas *et al*. (2016) also reported that by understanding the genetic diversity and structure of tobacco, breeders can make decisions regarding the selection of parents for crossbreeding, the development of new cultivars, and the preservation of valuable genetic resources.

Rotation Method: Varimax with Kaiser Normalization: In rotated component matrix, the eigenvectors are converted to factor loadings, a term that can refer to either the factor pattern matrix or the factor structure matrix. The factor pattern matrix is the beta weights for generating the variables from the factor scores (for the case where both the variables and the factor scores are standardized). The factor structure matrix is the correlations between the variables and the factors. For an orthogonal rotation the factor pattern matrix and the factor structure matrix are identical.

Table 6. Eigenvalues, % variance and cumulative eigenvalues of Tobacco germplasm.

Principal components	eigenvalue	percentage of variance	cumulative percentage of variance
PC 1	3.36	55.96	55.96
PC 2	1.26	20.97	76.92
PC 3	0.57	9.43	86.35
PC 4	0.50	8.40	94.75
PC 5	0.20	3.27	98.02
PC 6	0.12	1.98	100.00

Extraction method: Principal Component Analysis

Table 7. Rotated component matrix

S. No.	Traits	Components					
		1	2	3	4	5	6
1	Plant height	0.383	0.32	0.277	0.355	0.74	0.017
2	Number of leaves	-0.077	0.983	0.09	-0.007	0.138	-0.001
3	Internode length	0.309	-0.011	0.205	0.905	0.207	0.012
4	Leaf length	0.924	-0.063	0.159	0.235	0.156	-0.194
5	Leaf width	0.867	-0.053	0.235	0.241	0.212	0.295
6	Leaf Yield per ha	0.228	0.108	0.935	0.191	0.159	0.013
SS loadings		1.905	1.095	1.088	1.082	0.704	0.125
Proportion Var		0.318	0.183	0.181	0.18	0.117	0.021
Cumulative Var		0.318	0.5	0.681	0.862	0.979	1

Most PCAs are conducted using the Varimax rotation (Kaiser, 1955), although others are available. In this approach, pairs of factors are rotated in the two-dimensional space formed by their two axes such as to maximize the sum of the variances of the squared loadings. The factors are systematically rotated in pairs until changes are negligible. This procedure has the effect that factor loadings tend to be as extreme as possible (either zero or high), a quality shared by other members of the Orthomax family of rotations (Gorsuch, 1983). It is important to recognize that rotated principal components are not principal components (the axes associated with the Eigen value decomposition) but are merely components. In this context, unrotated principal components are denoted as PC's, while rotated PCs are now onwards labeled as Rotated components (RC's) (Shukla, 2015).

A principal factor matrix after Varimax rotation with Kaiser Normalization for these six factors is given in **Table 7**. The values in the table for factor loadings indicate the contribution of each variable to the factors (PC's). The results clearly showed RC1 was having high proportion of variation of 0.318 with highest contribution from leaf length (0.924) and leaf width (0.867) followed by factor RC2 with high loading of number of leaves per plant (0.983), RC3 with high loading of trait Leaf yield per hectare (0.935), RC4 with high loading of internode length (0.905) and RC5 with high loading of plant height (0.74). whereas, RC6 showed least variability. The rotated component matrix results also showed positive loading for the trait plant height in all the RC's. Further it was observed that RC3 and RC5 showed positive loadings for all the traits . On the basis of PC scores genotypes were selected for utilization in genetic improvement programme in Tobacco. The genotypes contributing higher PC scores will be useful for the transferring the single traits in the genotype lacking that traits. It can be concluded that PC analysis highlights the characters with maximum variability. So, intensive selection procedures can be designed to bring about rapid improvement of yield and its attributes. Top twenty genotypes selected on the basis of PC scores and having positive values with sharing their presence in different PC's are presented in **Table 8**. The results showed that GT-4 showed its presence in PC1, PC2 and PC5 and NBD 325 showed its presence in PC2, PC4 and PC5 indicating that these are the promising genotypes with maximum loadings for different traits, while some genotypes showed presence in maximum two components. Further, top twenty genotypes selected on the basis of PC scores and having negative values with sharing there presence in different PC's are presented in **Table 9**. The results revealed that ABD-84 showed its presence in PC2, PC4 and PC6, B.S.P (Black Spangle parent) showed its presence in the PC3, PC4 and PC6, ABD 128 showed its presence in PC4, PC5 and PC6

Table 8. Rotated component scores of top twenty Tobacco genotypes having positive values >1.0 in each PCs

Genotype	RC1	Genotype	RC2	Genotype	RC3	Genotype	RC4	Genotype	RC5	Genotype	RC6
Gundsurti	**2.356**	Anand -23	**3.814**	KDH-959	**3.982**	NBD 325	**3.239**	GT-4	**2.608**	NBD 334	**3.694**
NBD 111	**2.281**	NBD 324	**2.879**	NBD 119	**3.279**	ABD 119	**3.113**	B.S.P (Black Spangle parent)	**2.521**	Sanand local	**3.108**
ABD 101	**2.205**	NBD 209	**2.786**	575-28-110	**3.255**	ABD 36	**2.751**	NyBD-56	**2.493**	169-19-6 (88-47-Sokha)	**2.577**
Bankete A-1	**2.071**	ABD 173	**2.362**	GT-7	**3.068**	114-4 (RPK type)	**2.585**	ABD 94	**2.450**	NBD 277	**2.396**
ArBD-8	**1.933**	GT-4	**2.311**	Jayalaxmi	**3.032**	ABD 24	**2.560**	NBD-48-1	**2.366**	Sender patti special	**2.320**
Red Russion	**1.763**	NBD 321	**2.262**	GT-5	**2.856**	783-51	**2.472**	Smyrna	**2.129**	NBD 159	**2.169**
S-20	**1.699**	NBD 325	**2.156**	Abirami	**2.562**	S-112	**2.406**	ABD 163	**2.091**	NBD-80-1	**1.921**
GT-4	**1.681**	ABD 174	**2.094**	ABD 138	**2.297**	Kodani	**2.245**	RPK-1-2	**1.853**	NBD 319	**1.846**
NBD 164	**1.616**	NPN 66	**1.935**	NBD 313	**2.223**	NBD 159	**2.212**	V-58	**1.850**	ABD 101	**1.832**
NBD 155	**1.546**	Keliu-20	**1.912**	ArBD-39	**2.060**	S-12	**2.172**	Anand -119	**1.736**	NPN 30	**1.815**
114-4 (RPK type)	1.540	NPN 75	1.856	NBD 209	1.885	GT-9	2.048	NBD 324	1.702	A-428	1.806
169-2 (N&L)	1.540	NBD-57-1	1.836	NyBD 55	1.830	Dumbara	1.927	ABD 109	1.686	NBD 289	1.790
ABD 112	1.527	Bhagyalakshmi	1.797	NBD 308	1.748	S-20	1.923	NBD 271	1.503	ABD 109	1.669
NBD 136	1.525	Line 121-13-27-29 (108-15 x Olar)	1.774	ABD 43	1.677	NBD 324	1.884	NBD 320	1.502	RPK-1-2	1.605
ABD 99	1.506	GT-5	1.726	Anand-2	1.669	NBD 314	1.794	NBD 325	**1.441**	Bhavyashree	1.511
Sender patti special	1.479	ABD 164	1.697	ArBD-7	1.571	Abirami	1.772	ABD 70	1.392	NBD 316	1.503
ABD 96	1.462	Line 181-83-1 (S-20 xK-20)	1.639	ArBD-32	1.495	Anand-2	1.771	NBD 327	1.379	Line 181-83-1 (S-20 xK-20)	1.491
ABD 95	1.461	NBD 319	1.614	NBD 334	1.482	V-54	1.668	F-7-127	1.379	TI-421	1.450
ABD 110	1.452	Line 114-16 (Female parent of GT-4)	1.601	NBD 154	1.437	Line 121-13-27-29 (108-15 x Olar)	1.646	ABD 118	1.372	NBD 318	1.450
ArBD-40	1.432	NBD 322	1.515	NBD 307	1.370	NBD 43	1.457	ABD 96	1.339	ABD 62	1.439

*Genotypes marked with same colour indicating presence of genotype in different component

indicating that these genotypes were having lowest loadings for different traits. Hence these genotypes (low loading) could be used in crossing progamme of tobacco with high loading genotypes like GT-4 and NBD 325 to achieve maximum genetic gain. From the principal component scores it is possible to identify the genotypes contributing highest variation for several characters and genotypes are ranked accordingly. Hence, for cluster analysis, PCA has added advantage for selection of genotypes in breeding programmes (Kalagare *et al.,* 2022).

In summary, the purposes of this work is to draw some inference about the available tobacco germplasm using multivariate statistical analysis techniques like heirarchial clustering and PCA. The results demonstrated that PCs 1 and 2 have contributed a variance of 55.96 percent and 20.97 per cent respectively, which can be used to categorize tobacco genotypes according to their loadings for different traits. Further results of heirarchial clustering divided the 246 genotypes into eight clusters indicating presence of significant variation among the tested genotypes. Further breeding programmes

Table 9. Rotated component scores of top ten Tobacco genotypes having negative values >1.0 in each PCs

Genotype	RC1	Genotype	RC2	Genotype	RC3	Genotype	RC4	Genotype	RC5	Genotype	RC6
NBD 277	-2.288	ABD 109	-2.717	S-112	-2.405	ABD 163	-3.216	Bankete A-1	-3.153	114-4 (RPK type)	-3.319
NPN 81	-2.214	NBD 292	-1.913	S-12	-2.158	Line 93-103-93 (88-47 x Sokh)	-2.883	Line 114-16 (Female parent of GT-4)	-2.408	NBD 271	-2.536
NBD 290	-2.098	ABD 163	-1.770	ABD 107	-1.718	A-428	-2.699	Bhagyalakshmi	-2.298	103-9-101	-2.504
ABD 121	-2.092	NyBD-56	-1.717	ABD 110	-1.673	Sender patti special	-2.444	ArBD-39	-2.267	NBD 323	-2.428
ABD 117	-2.013	Line 93-103-93 (88-47 x Sokh)	-1.708	320-2-80-25-84-10-I	-1.663	B.S.P (Black Spangle parent)	-2.281	Keliu-20	-2.168	ABD 104	-2.257
Pilliu-19	-2.000	ABD 119	-1.664	Vairam	-1.647	ArBD-40	-2.202	ArBD-33	-2.058	ABD 125	-2.254
Samsan	-1.981	ABD 107	-1.661	ABD 125	-1.615	Bhagyalakshmi	-2.141	114-4 (RPK type)	-2.050	NBD 209	-1.987
ABD 127	-1.916	ABD 30	-1.658	NBD 292	-1.598	DWFC	-1.795	NBD 323	-2.046	Anand -23	-1.916
ABD 116	-1.875	NBD 313	-1.623	ABD 109	-1.527	Trabizonal	-1.710	NyBD-4	-1.873	ABD 102	-1.905
ABD 124	-1.874	ABD 84	-1.508	ABD-101 (GABT-11)	-1.469	320-2-30-28-18-I	-1.662	ABD 116	-1.777	ABD 84	-1.718
320-2-30-28-20-12	-1.865	ABD 95	-1.486	Line 121-13-27-29 (108-15 x Olar)	-1.455	ABD 112	-1.634	ABD 121	-1.624	ABD 123	-1.686
NBD 260	-1.739	ArBD-9	-1.421	Line-1-1	-1.435	NyBD-56	-1.567	ABD 65	-1.537	ABD 124	-1.683
NPN 30	-1.679	F-7-124	-1.353	B.S.P (Black Spangle parent)	-1.399	ABD 130	-1.464	ABD 10	-1.506	NBD 276	-1.677
ABD 43	-1.618	ABD 79	-1.336	ArBD-40	-1.389	ABD 84	-1.375	ArBD-32	-1.481	ABD 146	-1.664
NPN 73	-1.534	ABD 151	-1.330	GT-5	-1.363	GT-5	-1.365	BL 4-2	-1.449	NBD 154	-1.579
Trabizonal	-1.523	ABD 61	-1.324	Bhagya	-1.353	Line 132-2-2 {K-20 x Skh} x K-20	-1.362	ABD 128	-1.442	ArBD-32	-1.576
NBD 257	-1.514	NBD 122	-1.291	ABD 115	-1.331	ABD 128	-1.343	Thangam	-1.400	ABD 128	-1.514
NBD 289	-1.494	DWFC	-1.290	Jati	-1.323	NyBD-4	-1.289	HDBRG-LP-2	-1.390	ArBD-9	-1.510
NBD 327	-1.491	NBD 289	-1.279	NPN 63	-1.308	ABD 164	-1.197	NBD 314	-1.381	B.S.P (Black Spangle parent)	-1.481
N.C.D. (Necrotic Crinkle Dwarf)	-1.483	GT-7	-1.273	169-2 (N&L)	-1.230	NBD 316	-1.182	NBD 43	-1.352	BL 4-2	-1.449

could take the advantage of those genetic materials and use them as a parental genotype in breeding programme to achieve more genetic advance for different yield attributes. The genoytpes with high loadings for various traits like GT-4, KDH-959, NBD 325, Anand-2, S-20, NBD 324, NBD 209, ABD 95, NBD 324 and GT-5 can be crossed to develop model plant which is superior to all the traits. These genotypes can also be crossed with genotypes with low loadings for various traits like Line 93-103-93 (88-47 x Sokh), ABD 116, ABD 121, DWFC, F-7-124, Trabizonal, NPN 81, ABD 127, Bhagyalakshmi, A-428, 320-2-30-28-18-I, ABD 124 and 320-2-30-28-20-12 to generate more variability.

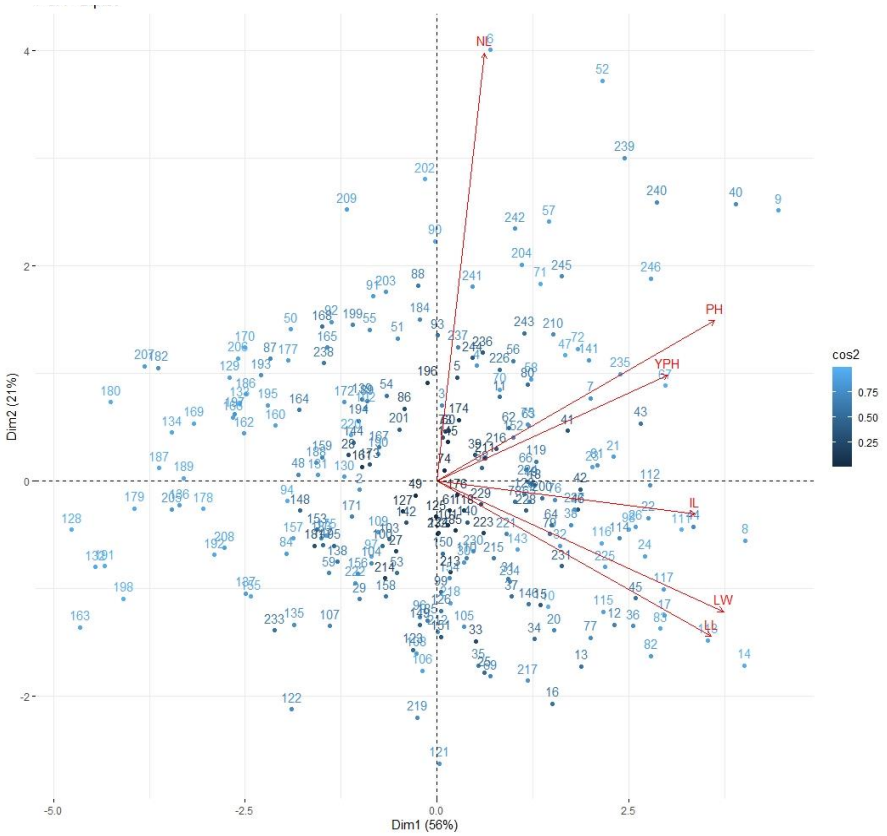

Fig. 2. PCA – Biplot for Tobacco genotypes

REFERENCES

Abdi, H. and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: Computational statistics*, **2**(4):433-459. [Cross Ref]

Anderson, T. W. 1972. An introduction to multivariate analysis. Wiley Eastem Pvt. Ltd, New Delhi.

Ann, D.J. and Kim, Y.D. 1983. Varietal classification on the basis of cluster analysis in burley tobacco of *N. tabacum* L. *Journal of the Korean Society of Tobacco Science*, **5**(2): 25-32.

Anonymous, 2020. Area, production, productivity of Tobacco in India and Karnataka. www.indiastat.com.

Gorsuch, R.L. 1983. Factor Analysis. *NJ Laurence Erlbaum Associates*. Hillsdale.

Hemavathy, A.T., Bapu, J.R. and Priyadharshini, C. 2017. Principal component analysis in pigeonpea (*Cajanus cajan* (L.) millsp.). *Electronic Journal of Plant Breeding*, **8**(4):1133-1139. [Cross Ref]

Hosseinzadeh, F.N., Shahadati, M.Z., Kiani, G., Salavati, M.R., Zamani, P., Mahdavi, A. and Alinejad, R. 2015. Investigation of genetic diversity among different oriental tobacco (*Nicotiana tabacum* L.) varieties using multivariate methods.126-134.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**(6):417. [Cross Ref]

Kaiser, H.F. 1955. An analytic rotational criterion for factor analysis. *Amer. Psychologist*. **10**:438.

Kalivas, A., Ganopoulos, I., Bosmali, I., Tsaliki, E., Osathanunkul, M., Xanthopoulou, A., Moysiadis, T., Avramidou, E., Grigoriadis, I., Zambounis, A., Tsaftaris, A., Nianiou-Obeidat, I. and Madesis, P. 2016. Genetic diversity and structure of tobacco in greece on the basis of morphological and microsatellite markers. *Crop Science*, **56**: 2652-2662. [Cross Ref]

Kalagare, V.S., Ganesan, N.M., Iyanar, K., Chitdeshwari, T. and Chandrasekhar, C.N. 2022. Multivariate analysis in parental lines and land races of pearl millet [*Pennisetum glaucum* (L.) R. Br.]. *Electronic Journal of Plant Breed*ing, **13**(1): 155-167. [Cross Ref]

Kassambara, A. and Mundt, F. 2020. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. https://CRAN.R-project.org/package=factoextra

Le, S., Josse, J. and Husson, F. 2008. Facto Mine R: An R Package for Multivariate Analysis. Journal of Statistical Software, 25, 18-21. http://www.jstatsoft.org/v25/i01. [Cross Ref]

Maji, A.T. and Shaibu, A. A. 2012. Application of principal component analysis for rice germplasm characterization and evaluation. *J. Plant Breed. Crop Sci*., **4**: 87-93. [Cross Ref]

Morrison, D.E. 1982. Multivariate Statistical Methods (2nd ed. 4th Print, 1978). McGraw Hill Kogakusta Ltd.

Murphy, J.P., Rufty, T. and Rodgers, D. 1987. A representation of the pedigree relationships among flue-cured tobacco cultvars. *Tob Sci*., **31**:70-75.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, **2**: 559. [Cross Ref]

Ringer, M. 2008. What is principal component analysis?. *Nature Biotechnology*, **26**(3): 303-304. [Cross Ref]

Shukla, A. K. 2015.Characterization and evaluation of field pea genotypes for yield and quality attributing traits. M.Sc. Thesis, Jawaharlal Nehru Krishi Vishwa Vidyalaya, Jabalpur.

Wang, Y., Zuo, A., Liang, R., Hu, Q., Li, X., Jin, B., Ye, W. and Zhang, J. 2014. Evaluation of principal components and cluster analysis on flue-cured tobacco of industrial grade. *Southwest China Journal of Agricultural Sciences*, **27**(4): 1733-1736.

Wickham, H. 2016. *GGPLOT2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.