Int J Adv Life Sci Res. Volume 4(3) 32- 43
https://doi.org/10.31632/ijalsr.2021.v04i03.005

International Journal of Advancement in Life Sciences Research

Online ISSN: 2581-4877

journal homepage http://ijalsr.org

Original Article

Visual Clustering Analysis of some traditional Mango (*Mangifera indica* L.) varieties of Murshidabad District, West Bengal using Clust Vis web tool

Mitu De[1], Subhasree Dutta[2], Susanta Ray[3], and Santi Ranjan Dey*[2]

[1]*Department of Botany, Gurudas College, Kolkata 700054, India*
[2]*Department of Zoology, Rammohan College, Kolkata 700009, West Bengal, India*
[3]*Goalpara Tarun Sangha Junior High School, Khagra, Murshidabad, India*

**Correspondence E-mail : srdey1@rediffmail.com*

Abstract

A clustergram or a heatmap is one of several techniques that directly visualize data without the need for dimensionality reduction. Heatmap is a representation of data in the form of a map or diagram in which data values are represented as colours. Cluster heatmaps have high data density, allowing them to compact large amounts of information into a small space. "ClustVis", is a web tool for visualizing clustering of multivariate data using Principal Component Analysis and Heatmap. Using this web tool, genetic relationships among the traditional mango (*Mangifera indica* L.) varieties can be visualized. In this investigation ten (10) indigenous mango varieties were selected. These were elite varieties of Murshidabad *viz.* Anaras, Bhabani, Champa, Dilpasand, Kalabati, Kohinoor, Kohitoor, Molamjam. The morphological and biological characters were analyzed using this tool. Analysis and assessment of the current status of mango genetic resources will be important for ascertaining the relationship among traditional varieties. This data may be used for appropriate conservation and sustainable utilization measures. This information may also be needed to carry out breeding programs to develop improved cultivars for sustainable livelihoods of local communities.

Keywords: *ClustVis; Heatmap; traditional mango variety; Murshidabad.*

Introduction

Clustering is a popular unsupervised learning method. It is used by analysts during exploratory data analysis (EDA), which depends on the discovery of patterned relations and structures among data instances and attributes (Hastie *et al,* 2005). Data scientists need tools that facilitate iterative, rapid exploration of the space of data clusterings. Data visualization is a central tool for the initial analysis of biological data, and dimensionality reduction techniques, such as principal component analysis (PCA) are commonly employed to project high dimensional data onto two or three dimensions so it can be visualized. A clustergram, or a heatmap is one of several techniques that directly visualizes data without the need for dimensionality reduction (Eisen *et al,* 1998). So Heatmaps are excellent for data visualization.

Mango has got various local names viz. In Murshidabad district a single variety "Himsagar" is also known as "Shadulla" among some people and "Khirsapati" in Northern Murshidabad. So for proper identification with statistical analysis of morphological and biological characters are very much important.

Cluster analysis is usually performed by using similarity and dissimilarity data. In this way proper identification/classification of varieties can be performed.

Heatmaps: Tools for Visual Clustering Analysis

There are several tools for Visual Clustering Analysis. For examples of some tools are ClusterSculptor (Nam *et al,* 2007) and Cluster Sculptor (Bruneau *et al,* 2015). These tools enable users to supervise clustering processes. ClustVis is a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. This web server is freely available at http://biit.cs.ut.ee/clustvis/. Lately ClustVis has been used for the whole-unique-sequence-based analysis (Ionov and Rogovskyy, 2020).

Heat maps make it easy to visualize complex data and understand it at a glance. The practice we now call heat maps is thought to have originated in the 19th century, where manual gray-scale shading was used to depict data patterns in matrices and tables. There are variations of heatmap viz. web heat maps and tree maps (Zhao *et al,* 2014). The term heatmap was first trademarked in the early 1990s, when software designer Cormac Kinney created a tool to graphically display real-time financial market information. Nowadays, heatmaps can still be created by hand, using Excel spreadsheets, or with special software and web tools. So a heatmap is a graphical representation of data that uses a system of color-coding to represent different values. Heatmaps are used in various forms of analytics but are most commonly used to show user behaviour on specific webpages or webpage templates. The primary purpose of Heat Maps is to better visualize the volume of locations/events within a dataset and assist in directing viewers towards areas on data visualizations that matter most.

Heat maps in Life Sciences

Biology heat maps are typically used in molecular biology to represent the level of expression of many genes across a number of comparable samples (e.g. cells in different states, samples from different patients) as they are obtained from DNA microarrays. In heat maps the data is displayed in a grid where each row represents a gene and each column represents a sample. The colour and intensity of the boxes is used to represent changes (not absolute values) of gene expression. Clustergrams or heatmaps are easy to interpret and are widely used to visualize biological data in print publications. Hierarchically clustered heatmaps can also be used to visualize biological networks by displaying network connections in a symmetric adjacency matrix (Pavlopoulos *et al,* 2008).

Heatmap clustering

A heatmap (or heat map) is another way to visualize hierarchical clustering. It's also called a false colored image, where data values are transformed to color scale.The columns/rows of the data matrix are re-ordered according to the hierarchical clustering result, putting similar observations close to each other. Cluster heatmaps have high data density, allowing them to compact large amounts of information into a small space (Wilkinson and Friendly, 2009). Cluster heatmaps are commonly used in biology and related fields to reveal hierarchical clusters in data matrices. These types of visualization technique have high data density and reveal clusters better than unordered heatmaps alone.

Heatmap and hierarchical clustering for morphological parameters

Cluster heatmaps have been used in various studies viz. studying interactions between environmental variables and microbial communities (Wang *et al,* 2012); unsupervised pattern discovery in human chromatin structure through genomic segmentation (Hoffman *et al*, 2012). Heat maps of species abundance clustering are being published. The influence of heavy metals on total soil bacterial population and its diversity pattern from 10 km distance of a Zinc smelter in Feng County, Qinling Mountain, China was reported by Feng Shen and Co workers in 2016 (Shen *et al,* 2016).

From literature it was found that Heatmap and hierarchical clustering for morphological and physiological parameters under well-watered and drought stress conditions in 49 switchgrass (*Panicum virgatum* L.) genotypes after 30 days of treatment. Heatmaps and Clustering analysis of switchgrass genotypes showed two main groups (Liu *et al,* 2015). In 2020 Guijarro-Real and co workers worked on the Morphological Diversity and Bioactive Compounds in Wall Rocket (*Diplotaxis erucoides* (L.) DC.) and used clustering tools (Guijarro-Real *et al*. 2020). Toubiana and co workers in 2020 worked on the morphological and metabolic profiling of a tropical-adapted potato using heatmaps (Toubiana *et al,* 2020). The ability of maize populations/landraces to tolerate drastically extreme environments over the past four centuries in Algeria leads to characterize these genetic resources for germplasm management as well as the identification of the best landraces useful for genetic improvement with Heatmaps (Aci et al, 2018).

Cluster Analysis of Multivariate Data

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. Similar to a contour plot, a heat map is a two-way display of a data matrix in which the individual cells are displayed as colored rectangles. The color of a cell is proportional to its position along a color gradient. Usually, the columns (variables) of the matrix are shown as the columns of the heat map and the rows of the matrix are shown as the rows of the heat map, as in the example below. The order of the rows is determined by performing hierarchical cluster analyses of the rows. This tends to position similar rows together on the plot. The order of the columns is determined similarly. Usually, a clustered heat map is made on variables that have similar scales, such as scores on tests in the example below. If the variables have different scales, the data matrix must first be scaled using a standardization transformation such as z-scores or proportion of the range (Metsalu and Vilo, 2015).

Intra-specific diversity among traditional mango varieties of Murshidabad district

Mango has rich intra-specific diversity and there are about 1600 cultivars in the world. India has the richest collection of mango cultivars with mango being grown in all the states of the country. Different varieties have unique taste, flavour, texture and size. Custodian farmers have been maintaining, promoting and adapting a number of indigenous mango varieties on their farms. These custodian farmers protect, nurture and propagate the rich mango legacy of India. Detailed and well documented information about the available genetic material together with a broad, well maintained varietal diversity are essential for breeding efforts. This should also include local varieties (Subedi *et al.,* 2005), which may have a low market, but high breeding value.

Mango Diversity

Local mango genetic resources have been found to be community assets for fulfilling the nutritional and other local requirement, such as fuel-wood and shade. Maintenance of local mango orchards is a relatively cost-effective strategy since many of the cultivars are adapted to marginal conditions. The International Plant Genetic Resources Institute descriptors allow for the use of visual assessment tools of morphological traits to characterize mango germplasm (IPGRI (2006). Morphological variations of mango varieties have been studied extensively worldwide (Uddin *et al.,* 2006; Galvez – Lopez *et al*., 2010; Majumder *et al,* 2011). In 2009 Njuguna *et al*., (2009) suggested about the absence of fruits, traits such as the color of young leaf, leaf margin type, leaf fragrance strength, tree height and stem circumference can be used to distinguish between cultivars particularly between the traditional and exotic varieties. Each of the clusters generated by the dendogram possessed varieties that can be used as parents in breeding efforts. Moreover the original habitats of local mango have been rapidly changing in response to biotic, economic and other pressures in recent years (Subedi *et al.* 2004).

Mango diversity in West Bengal

Murshidabad district of West Bengal is famous for their traditional mango varieties. More than

200 mango varieties were recorded during the time of the royals of Murshidabad. After independence with rapid industrialization this germplasm is under threat (Mukherjee, 1953). In West Bengal proper documentation of the mango varieties is at a nascent stage. Usually for assessment, the quality characters of Mango are very much important (Dash and Hota, 1997; Majumder, 2013), because the quality characters are parameters of selection of proper plants for propagation. Till date the marketable quality characters are only considered when assessing the traditional varieties. In spite of many valuable morphological traits, genetic diversity conserved in local mango cultivars and in its exotic germplasm has not been assessed fully in Murshidabad district either by using morphological characters or DNA-based genetic markers. This recent documentation of the mango diversity from these areas show that the traditional mango varieties which are in general are low yielding are being replaced by new high yielding hybrids.

Materials and Method

Study Area

Murshidabad is a district of West Bengal in eastern India. Situated on the left bank of the river Ganges, the district is very fertile. Covering an area of 5,341 km² (2,062 sq mi) (Fig. 1).

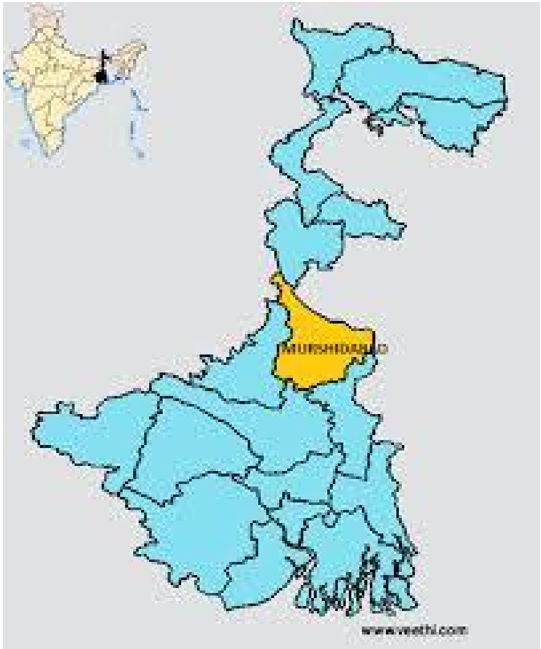


Fig.1. Map of West Bengal showing the location of Murshidabad district

The district comprises two distinct regions separated by the Bhagirathi River. To the west lies the Rarh,a high, undulating continuation of the Chota Nagpur plateau. The eastern portion, the Bagri, is a fertile, low-lying alluvial tract, part of the Ganges Delta. The district is drained by the Bhagirathi and Jalangi rivers and their tributaries. Bhagirathi is a branch of the Ganges, and flows southwards from Farakka barrage where it originates from the Ganges. It flows southwards through the district and divides it into more or less equal halves (Figure 2).

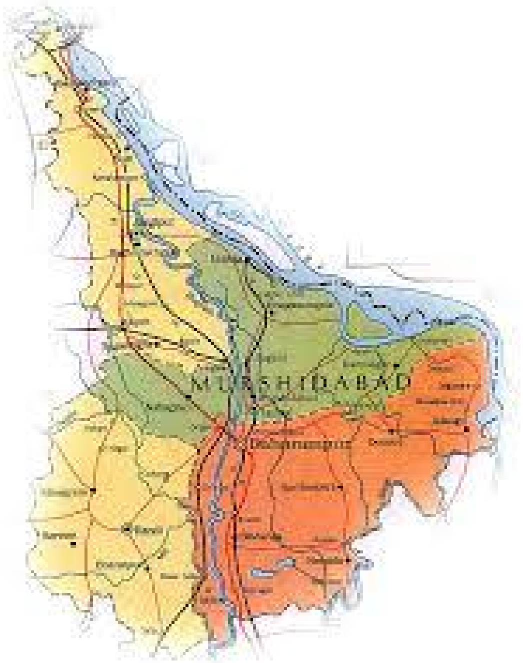


Fig.2. Map of Murshidabad district, Green Region is Lalbag or Murshidabad Subdivision

Methods

The ClustVis web tool allows users to upload their own data and easily create Principal Component Analysis (PCA) plots and heatmaps (Metsalu and Vilo, 2015). As an output, users can download PCA plot and heatmap in one of the preferred file formats.

Measuring Euclidean distance for Dendogram Analysis

In general, if you have p variables X1,X2, . . . ,Xp measured on a sample of n subjects, the observed data for subject i can be denoted by xi1, xi2, . . . , xip and the observed data for subject j by xj1, xj2, . . . , xjp. The Euclidean distance between these two subjects is given by

$$d_{ij} = \sqrt{\sum (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

When using a measure such as the Euclidean distance, the scale of measurement of the variables under consideration is an issue, as changing the scale will obviously effect the distance between subjects (e.g. a difference of 10cm could being a difference of 100mm). In addition, if one variable has a much wider range than others then this variable will tend to dominate. To get around this problem each variable can be standardised (converted to z-scores). However, this in itself presents a problem as it tends to reduce the variability (distance) between clusters. This happens because if a particular variable separates observations well then, by definition, it will have a large variance (as the between cluster variability will be high). If this variable is standardised then the separation between clusters will become less (Sokal, R and Michener, C. 1958).

UPGMA is a text-book algorithm that belongs to the family of agglomerative clustering methods that share the following common bottom-up scheme (cf. e.g. [4, p.162]). They take as input a dissimilarity D on a set X, i.e. a real-valued, symmetric map on X × X which vanishes on the diagonal, and build a collection of clusters or subsets of X which correspond to a rooted tree with leaf-set X. To do this, at each step two clusters with the minimum inter-cluster dissimilarity are combined to create a new cluster, starting with the collection of clusters consisting of singleton subsets of X, and finishing when the cluster X is obtained. Different formulations of the inter-cluster dissimilarity, which specifies the dissimilarity of sets as a function of the dissimilarities observed on the members of the sets, lead to different heuristic criteria of the agglomerative methods. UPGMA, as the name average linkage analysis suggests, uses the mean dissimilarity across all pairs of elements that are contained within the two clusters. UPGMA is minimized over all possible pairs of clusters. Since the arithmetic mean is used it is often more stable than linkage methods in which only a subset of the elements within the clusters are used (e.g. the single-linkage method) (Wilkinson and Friendly,2008).

Materials: Mango varieties

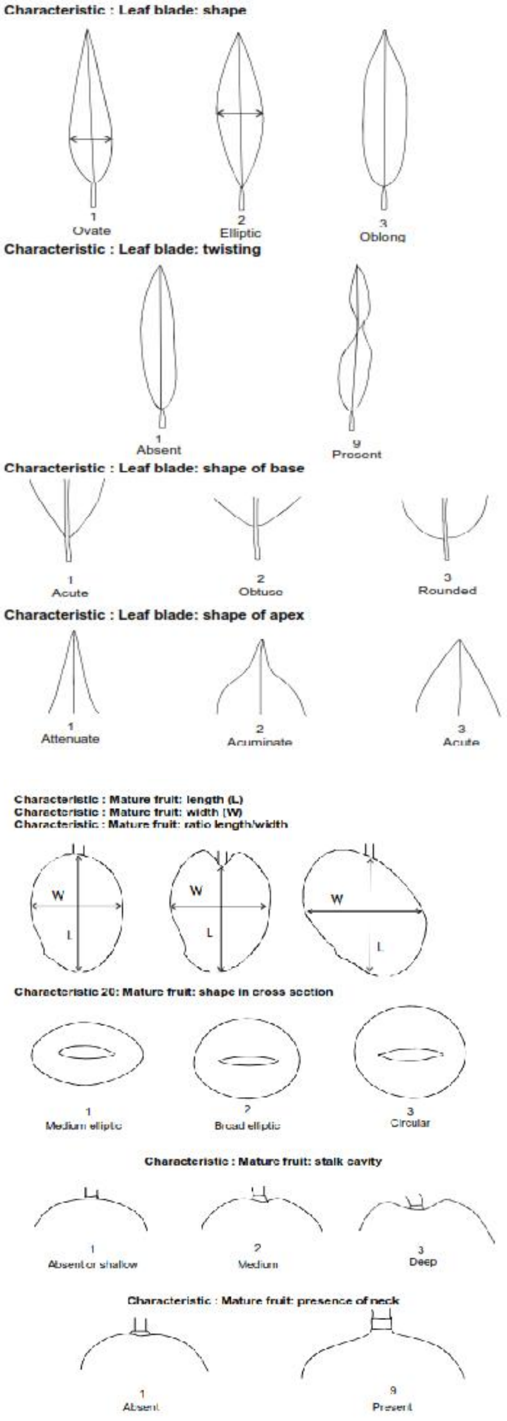
30 (Thirty) traditional Mango samples, in different names were spotted from old orchards of Lalbag or Murshidabad Subdivision of Murshidabad District. Initial characterization of these 30 mango samples were done as per PPV& FR. Data available on Scale of Mango varieties for two seasons (http://www.plantauthority.gov.in). A total of 50 (fifty) different fruit characters and plant morphology were considered. After characterization, an Excel sheet for the varieties against the characters (Figure 3) as per PPV&FR scale was prepared. Cluster analysis was performed for identification of varieties using CLUSTVIS (https://www.biit.cs.ut.ee/clustvis_large/). It was found that these 30 samples merged into 10 different lines of mango varieties. From the morphological characters 10 different varies, genetic distance analysis using UPGMA was performed. Determination of genetic relatedness among the varieties was found from UPGMA analysis (http://genomes.urv.cat).

Measured Characters (PPV& FR Scale)

The characters and corresponding scales for the description are given as Fig. 3. The description is given below.

Fruit Character:

Mature fruit length, width, ratio of length/width, shape of cross section, colour of the skin, density of lenticels. Colour contrast between lenticels and skin is weak, size of the lenticels, roughness of the surface, presence of cavity at stalk, depth of cavity at the stalk. Short length neck is present/absent, shape of ventral shoulder, shape of dorsal shoulder upward/downwards, presence of groove in the ventral shoulder is absent/present, and bulging on the ventral shoulder is present/absent. Sinus is absent/absent; depth of sinus. Bulging at the proximal stylar scar is absent/present or weak, point at the stylar scar is also absent/present. Diameter of the stalk attachment. Predominant colour of skin of the ripe fruit is yellow/ green/red/orange, speckling of skin is absent/present, thickness of skin is very thin/thick, adherence of the flesh to the skin is weak/strong or minimal/maximal, main colour of the flesh, firmness of the flesh, juiciness and a coarse texture of flesh. The amount of fiber attached to the stone of ripe fruit is low/high; amount of fibre attached to the skin is also low/high. Stone surface is ridged, seed kernel is oblong in view, and seed is mono embryonic/polyembyronic in nature. The variety has a tendency of late/early fruit maturity.


Characteristic : Leaf blade: shape
1 Ovate
2 Elliptic
3 Oblong
Characteristic : Leaf blade: twisting
1 Absent
9 Present
Characteristic : Leaf blade: shape of base
1 Acute
2 Obtuse
3 Rounded
Characteristic : Leaf blade: shape of apex
1 Attenuate
2 Acuminate
3 Acute
Characteristic : Mature fruit: length (L)
Characteristic : Mature fruit: width (W)
Characteristic : Mature fruit: ratio length/width
W
L
W
L
W
L
Characteristic 20: Mature fruit: shape in cross section
1 Medium elliptic
2 Broad elliptic
3 Circular
Characteristic : Mature fruit: stalk cavity
1 Absent or shallow
2 Medium
3 Deep
Characteristic : Mature fruit: presence of neck
1 Absent
9 Present

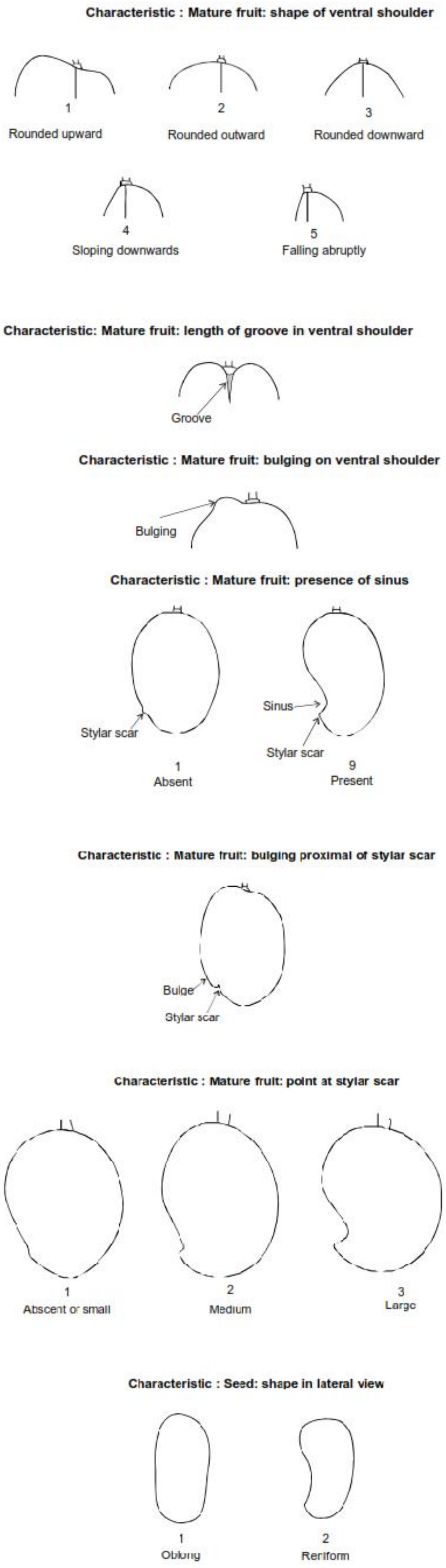

Characteristic : Mature fruit: shape of ventral shoulder
1 Rounded upward
2 Rounded outward
3 Rounded downward
4 Sloping downwards
5 Falling abruptly
Characteristic: Mature fruit: length of groove in ventral shoulder
Groove
Characteristic : Mature fruit: bulging on ventral shoulder
Bulging
Characteristic : Mature fruit: presence of sinus
Stylar scar
1 Absent
Sinus
Stylar scar
9 Present
Characteristic : Mature fruit: bulging proximal of stylar scar
Bulge
Stylar scar
Characteristic : Mature fruit: point at stylar scar
1 Absent or small
2 Medium
3 Large
Characteristic : Seed: shape in lateral view
1 Oblong
2 Reniform

Fig. 3 Explanations of some characteristics of Mango

Plant Characters:

This variety of Mango is known for presence/absence of anthocyanin coloration with negligible intensity for the young leaf. The leaf blade is generally medium/short/long in length, with a narrow width/broad; length/width ratio of the leaf blade is medium/low/high, colour. Twisting is present/absent in the leaf blade, shape of base is normal/acute, shape of apex is acuminated/pointed/blunt, and petiole length is Long/short (>3.0 cm). Time of flowering in this variety is medium/late/early. Inflorescence length is long/medium/short, diameter is also long/short/medium, ratio of diameter/length of inflorescence is long/short/medium, anthocyanin coloration is absent or weak.

Results

From the cluster analysis of PPV&FR data we have identified 10 different varieties that are landraces of Murshidabad. These varieties are Anaras, Bhabani, Champa, Dilpasand, Kalabati, Kohinoor, Kohitoor, Molamjam. The detailed morphology of 10 (ten) varieties were used for Cluster analysis. The description of just one sample variety is given below:

ANARAS (Endemic Variety of Murshidabad)

Mature fruit is medium in length (10 cm), with a medium width, ratio of length/width is small, shape of cross section is broad elliptic, color of the skin is green and yellow, and density of lenticels is sparse. The color contrast between lenticels and skin is weak, with negligible size of the lenticels. The corkiness due to the presence of the lenticels is absent, shallow cavity at stalk is present. Neck is absent. Shape of ventral shoulder is rounded upwards, shape of dorsal shoulder is rounded outwards. Groove in the ventral shoulder is absent, and bulging on the ventral shoulder is also absent. Sinus is absent; depth of sinus is negligible. Bulging at the proximal stylar scar is absent, point at the stylar scar is absent. Diameter of the stalk attachment is medium. Predominant color of skin of the ripe fruit is yellow green, speckling of skin is absent, thickness of skin is medium, adherence of the flesh to the skin is medium. Main color of the flesh of the ripe fruit is light yellow. Firmness of the flesh of the ripe fruit is soft, with medium juiciness, and medium texture of flesh.

Anaras mango variety is found only in Lalbag subdivision of Murshidabad.

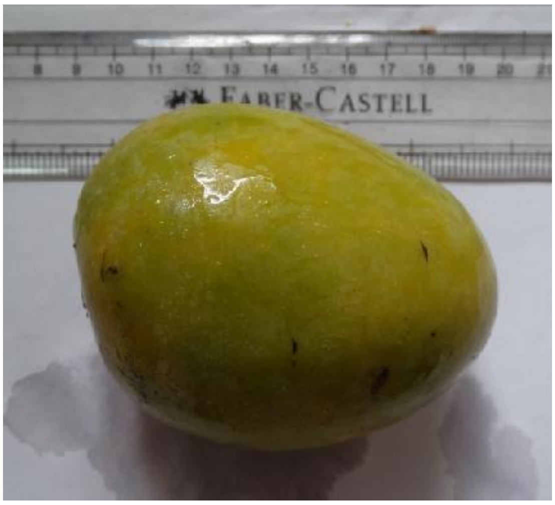


Fig.4. Mature Fruit of Anaras Variety

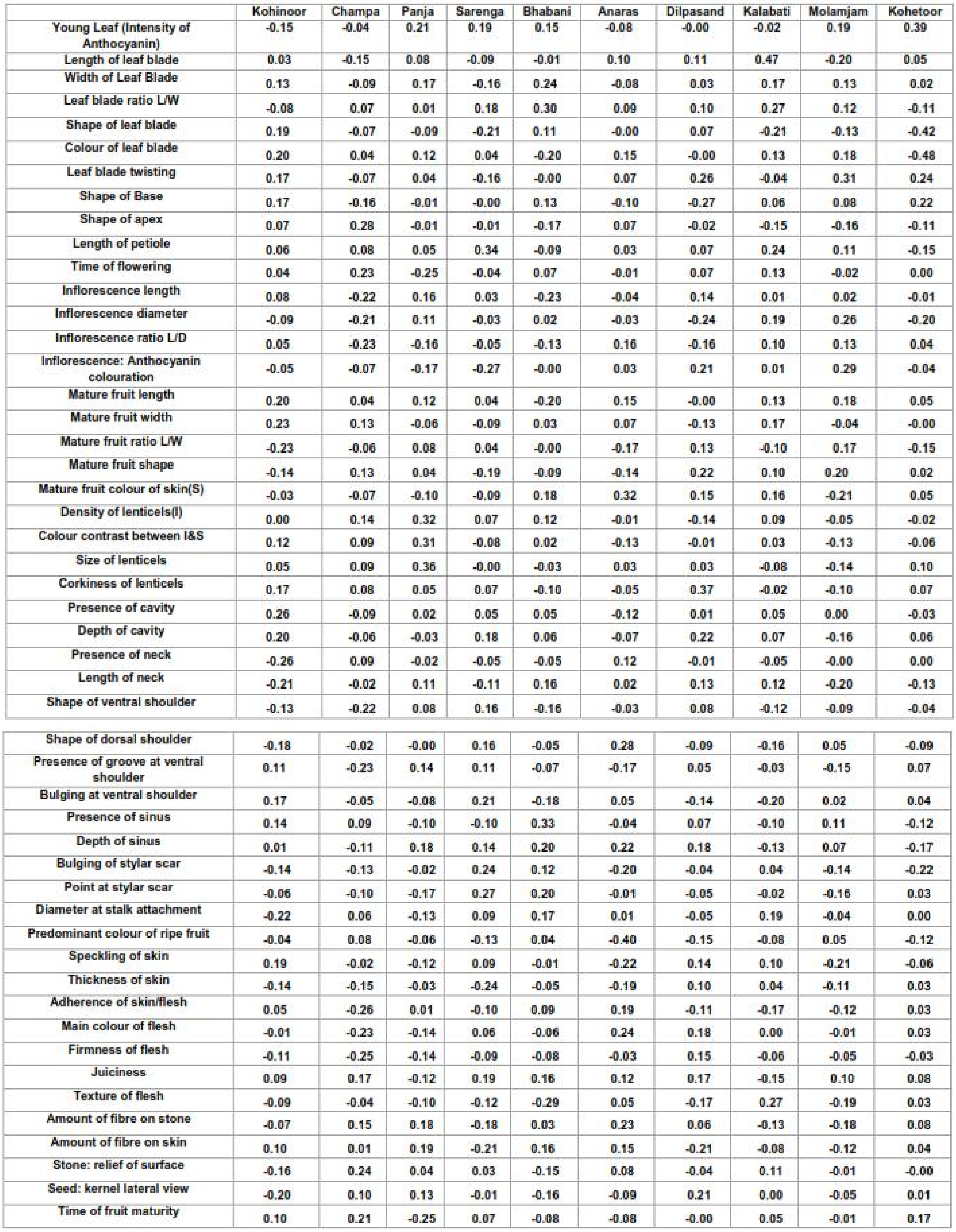
Components Contribution of Individual varieties as distance from axis and Characters for Identification / Measurements:

In Table 1 the data on the Components Contribution of Individual varieties as distance from axis and Characters for Identification/Measurements are given.

Analysis of Principal Components

From statistical analysis we observed 10 Principal Components are important for Varietal identification. The individual and cumulative distance of those PC are shown. (Table 2).

Table 1: Contribution in PC of each variety against each morphological character (As per PPV&FR)



	Kohinoor	Champa	Panja	Sarenga	Bhabani	Anaras	Dilpasand	Kalabati	Molamjam	Kohetoor
Young Leaf (Intensity of Anthocyanin)	-0.15	-0.04	0.21	0.19	0.15	-0.08	-0.00	-0.02	0.19	0.39
Length of leaf blade	0.03	-0.15	0.08	-0.09	-0.01	0.10	0.11	0.47	-0.20	0.05
Width of Leaf Blade	0.13	-0.09	0.17	-0.16	0.24	-0.08	0.03	0.17	0.13	0.02
Leaf blade ratio L/W	-0.08	0.07	0.01	0.18	0.30	0.09	0.10	0.27	0.12	-0.11
Shape of leaf blade	0.19	-0.07	-0.09	-0.21	0.11	-0.00	0.07	-0.21	-0.13	-0.42
Colour of leaf blade	0.20	0.04	0.12	0.04	-0.20	0.15	-0.00	0.13	0.18	-0.48
Leaf blade twisting	0.17	-0.07	0.04	-0.16	-0.00	0.07	0.26	-0.04	0.31	0.24
Shape of Base	0.17	-0.16	-0.01	-0.00	0.13	-0.10	-0.27	0.06	0.08	0.22
Shape of apex	0.07	0.28	-0.01	-0.01	-0.17	0.07	-0.02	-0.15	-0.16	-0.11
Length of petiole	0.06	0.08	0.05	0.34	-0.09	0.03	0.07	0.24	0.11	-0.15
Time of flowering	0.04	0.23	-0.25	-0.04	0.07	-0.01	0.07	0.13	-0.02	0.00
Inflorescence length	0.08	-0.22	0.16	0.03	-0.23	-0.04	0.14	0.01	0.02	-0.01
Inflorescence diameter	-0.09	-0.21	0.11	-0.03	0.02	-0.03	-0.24	0.19	0.26	-0.20
Inflorescence ratio L/D	0.05	-0.23	-0.16	-0.05	-0.13	0.16	-0.16	0.10	0.13	0.04
Inflorescence: Anthocyanin colouration	-0.05	-0.07	-0.17	-0.27	-0.00	0.03	0.21	0.01	0.29	-0.04
Mature fruit length	0.20	0.04	0.12	0.04	-0.20	0.15	-0.00	0.13	0.18	0.05
Mature fruit width	0.23	0.13	-0.06	-0.09	0.03	0.07	-0.13	0.17	-0.04	-0.00
Mature fruit ratio L/W	-0.23	-0.06	0.08	0.04	-0.00	-0.17	0.13	-0.10	0.17	-0.15
Mature fruit shape	-0.14	0.13	0.04	-0.19	-0.09	-0.14	0.22	0.10	0.20	0.02
Mature fruit colour of skin(S)	-0.03	-0.07	-0.10	-0.09	0.18	0.32	0.15	0.16	-0.21	0.05
Density of lenticels(l)	0.00	0.14	0.32	0.07	0.12	-0.01	-0.14	0.09	-0.05	-0.02
Colour contrast between l&S	0.12	0.09	0.31	-0.08	0.02	-0.13	-0.01	0.03	-0.13	-0.06
Size of lenticels	0.05	0.09	0.36	-0.00	-0.03	0.03	0.03	-0.08	-0.14	0.10
Corkiness of lenticels	0.17	0.08	0.05	0.07	-0.10	-0.05	0.37	-0.02	-0.10	0.07
Presence of cavity	0.26	-0.09	0.02	0.05	0.05	-0.12	0.01	0.05	0.00	-0.03
Depth of cavity	0.20	-0.06	-0.03	0.18	0.06	-0.07	0.22	0.07	-0.16	0.06
Presence of neck	-0.26	0.09	-0.02	-0.05	-0.05	0.12	-0.01	-0.05	-0.00	0.00
Length of neck	-0.21	-0.02	0.11	-0.11	0.16	0.02	0.13	0.12	-0.20	-0.13
Shape of ventral shoulder	-0.13	-0.22	0.08	0.16	-0.16	-0.03	0.08	-0.12	-0.09	-0.04
Shape of dorsal shoulder	-0.18	-0.02	-0.00	0.16	-0.05	0.28	-0.09	-0.16	0.05	-0.09
Presence of groove at ventral shoulder	0.11	-0.23	0.14	0.11	-0.07	-0.17	0.05	-0.03	-0.15	0.07
Bulging at ventral shoulder	0.17	-0.05	-0.08	0.21	-0.18	0.05	-0.14	-0.20	0.02	0.04
Presence of sinus	0.14	0.09	-0.10	-0.10	0.33	-0.04	0.07	-0.10	0.11	-0.12
Depth of sinus	0.01	-0.11	0.18	0.14	0.20	0.22	0.18	-0.13	0.07	-0.17
Bulging of stylar scar	-0.14	-0.13	-0.02	0.24	0.12	-0.20	-0.04	0.04	-0.14	-0.22
Point at stylar scar	-0.06	-0.10	-0.17	0.27	0.20	-0.01	-0.05	-0.02	-0.16	0.03
Diameter at stalk attachment	-0.22	0.06	-0.13	0.09	0.17	0.01	-0.05	0.19	-0.04	0.00
Predominant colour of ripe fruit	-0.04	0.08	-0.06	-0.13	0.04	-0.40	-0.15	-0.08	0.05	-0.12
Speckling of skin	0.19	-0.02	-0.12	0.09	-0.01	-0.22	0.14	0.10	-0.21	-0.06
Thickness of skin	-0.14	-0.15	-0.03	-0.24	-0.05	-0.19	0.10	0.04	-0.11	0.03
Adherence of skin/flesh	0.05	-0.26	0.01	-0.10	0.09	0.19	-0.11	-0.17	-0.12	0.03
Main colour of flesh	-0.01	-0.23	-0.14	0.06	-0.06	0.24	0.18	0.00	-0.01	0.03
Firmness of flesh	-0.11	-0.25	-0.14	-0.09	-0.08	-0.03	0.15	-0.06	-0.05	-0.03
Juiciness	0.09	0.17	-0.12	0.19	0.16	0.12	0.17	-0.15	0.10	0.08
Texture of flesh	-0.09	-0.04	-0.10	-0.12	-0.29	0.05	-0.17	0.27	-0.19	0.03
Amount of fibre on stone	-0.07	0.15	0.18	-0.18	0.03	0.23	0.06	-0.13	-0.18	0.08
Amount of fibre on skin	0.10	0.01	0.19	-0.21	0.16	0.15	-0.21	-0.08	-0.12	0.04
Stone: relief of surface	-0.16	0.24	0.04	0.03	-0.15	0.08	-0.04	0.11	-0.01	-0.00
Seed: kernel lateral view	-0.20	0.10	0.13	-0.01	-0.16	-0.09	0.21	0.00	-0.05	0.01
Time of fruit maturity	0.10	0.21	-0.25	0.07	-0.08	-0.08	-0.00	0.05	-0.01	0.17

Table 2. Variance explained by Principal Component (10 Basic Component)

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10
Individual	0.25	0.17	0.13	0.11	0.09	0.09	0.07	0.05	0.04	0.00
Cumulative	0.25	0.42	0.55	0.66	0.75	0.84	0.91	0.96	1.00	1.00

From statistical analysis we observed 10 Principal Components (PC) are important for Varietal identification. The individual and cumulative distance of those PC are shown below (Fig 5).

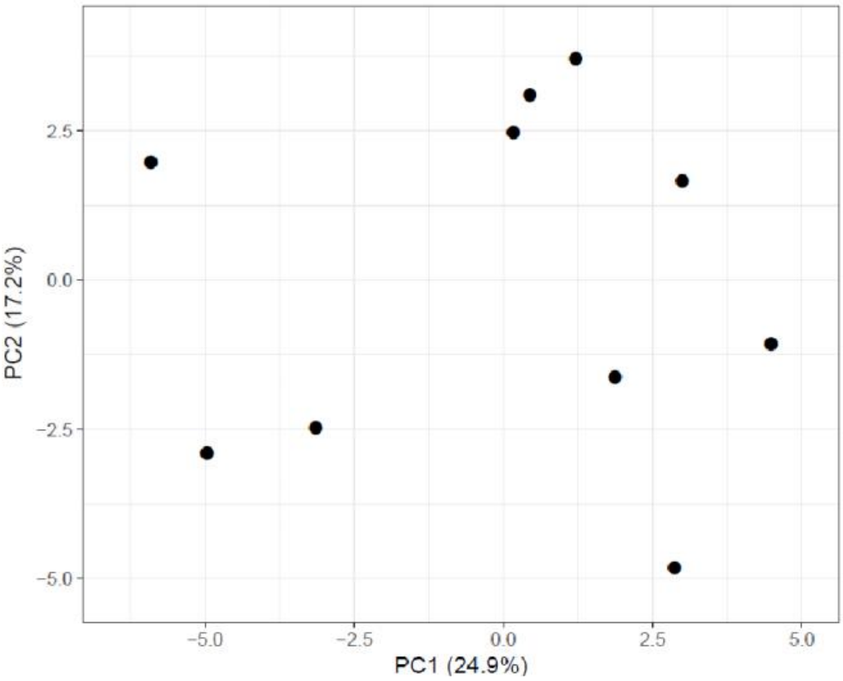


Figure 5. Cluster Analysis of the Varieties showing their relatedness

The Cluster Heat Map: The morphological characters were used to generate this Map.

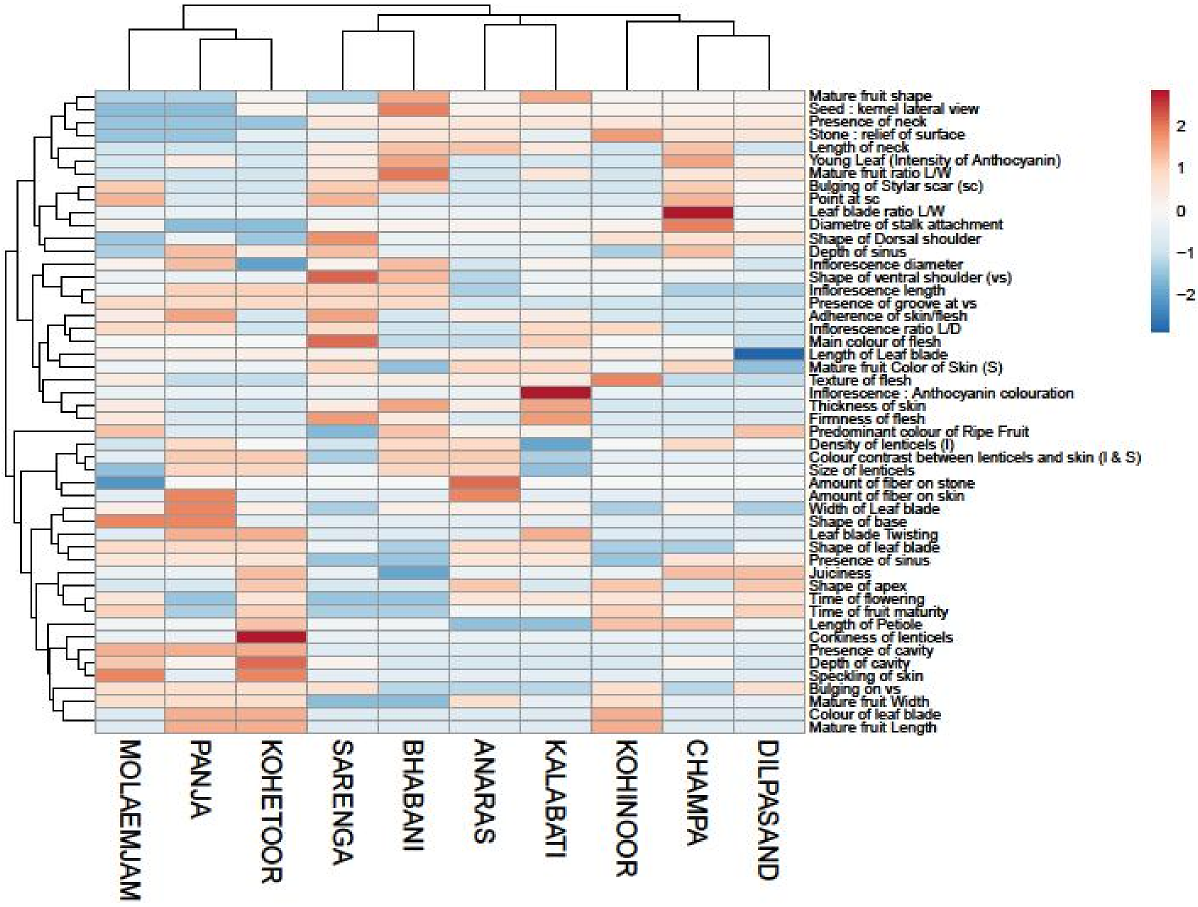


Fig 6. The Cluster Heat Map showing Similarity and Dissimilarity distribution of characters among 10 selected varieties/Germplasms. The Red are the similarity areas/characters and blue are dissimilarity areas/characters. The intensity of colour is representation of degree of similarity and dissimilarity.

The UPGMA Dendogram: From the morphological data this dendrogram was constructed

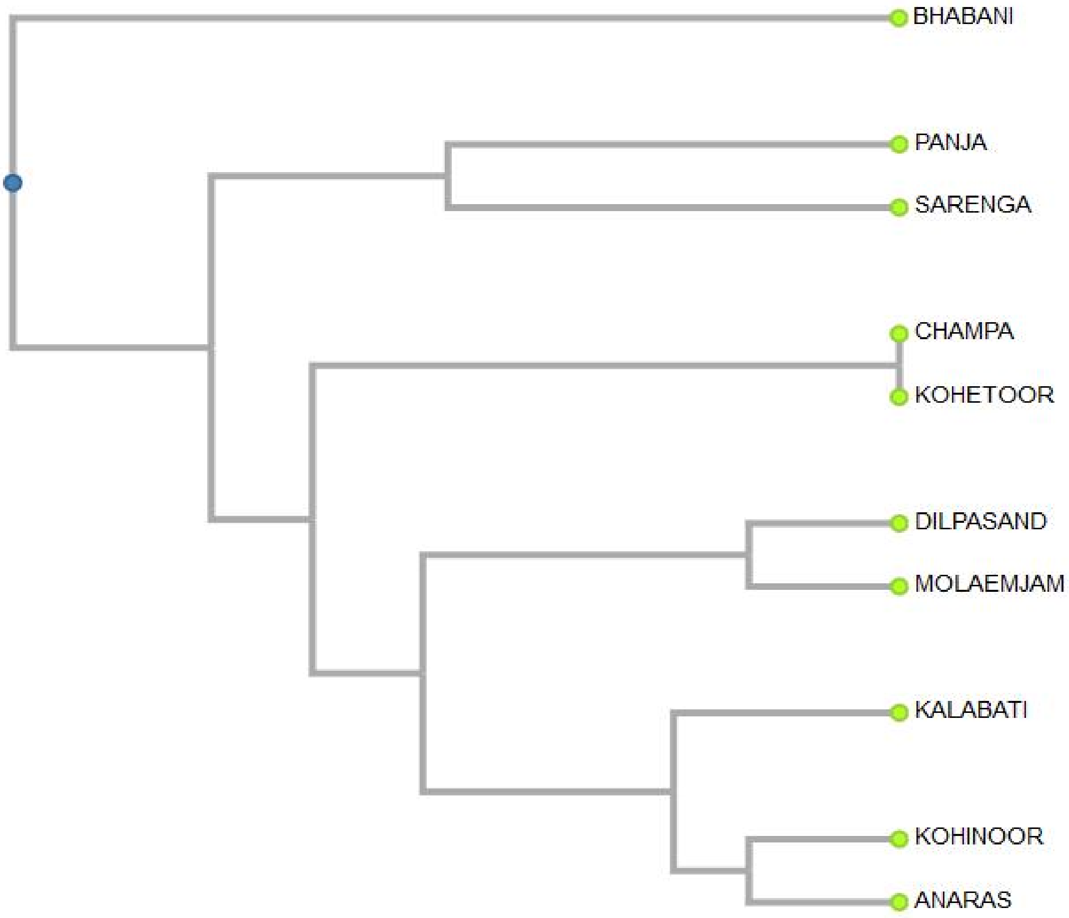


Fig. 7. The UPGMA Dendrogram drawn from CLUSTER analysis based on morphological characters of Mango varieties.

Discussion

PCA analysis:

Principal Component analysis has been used to measure the character contribution of variation among the different Genotypes of mango. From Table 1 & 2 it was found that the varieties are highly diversified in nature. In the analysis it was found that there were 9 major component determining diversities among these germplasms. From PC 1 to PC 9 they show an eigenvalue greater than 1.0 (5.91, 3.70,4.04, 2.92, 1.52, 2.40, 3.71, 1.97 and 1.68 respectively). If we consider the X-Y axis PC 1 and PC 2 showed 24.9% and 17.2% variability among the varieties.

Cluster and Heatmap analysis:

Cluster and Heatmap (Figure 5 and 6) analysis according to similarity- dissimilarity metrics showed that these varieties are highly diversified in nature. A single cluster of 3 varieties are seen rest of varieties are highly distant from each other.

UPGMA Analysis

From Cluster analysis, we have identified 10 unique varieties from 30 characterized sample. From UPGMA (Fig.7) we found among these 10 varieties Bhabani is unique and isolated from the ancestral stock earliest due to farmers selection. The varieties Panja and Sarenga, Kohitoor and Champa, Dilapasand and Molamjam, Kohinoor and Anaras are very close whereas Kalabati are distinct from all these varieties.

The characterization of the germplasm available for target traits is an essential step for selection and breeding programs (Guijarro-Real, 2020). It also is a provision to protect the interests of public sector breeding institutions and the farmers.

Conclusion

The Government of India enacted "The Protection of Plant Varieties and Farmers' Rights (PPV &FR) Act, 2001" adopting *sui generis* system. It emphasizes farmers' Rights as positive rights capturing the spirit of FAO International Treaty on Plant Genetic Resources for Food and Agriculture and UN Convention on Biological Diversity. At present the PPV&FR Authority has notified 156 crop species and has opened the registration for these crops. Indian legislation is not only in conformity with International Union for the Protection of New Varieties of Plants (UPOV), 1978, but also have sufficient. The studies mango genotypes showed considerable variability for most of the traits that could be explored for future improvement. Analysis of cluster based on morphology suggests the genotypes could be grouped altogether. This grouping is useful for breeders. Both the analysis of Principal Component and cluster are useful tools for providing information on variability and identification of proper varieties. The characterization of the germplasm available for target traits will aid in the selection of parents in future breeding programs. This data may be used for appropriate conservation and sustainable utilization measures. This information may also be needed to carry out breeding programs to develop improved cultivars for sustainable livelihoods of local communities.

Acknowledgment:
The authors would also take this opportunity to thank the mango orchard owners of Murshidabad who gave their germplasm for the characterisation. This work has been carried out with the financial assistance of a Research Project of the West Bengal Biodiversity Board (WBBB) to the first author as principal investigator.

Conflicts of Interest

The authors declare no conflict of interest.

References

Aci, M. M., Lupini, A., Mauceri, A., Morsli, A., Khelifi, L., & Sunseri, F. (2018). Genetic variation and structure of maize populations from Saoura and Gourara oasis in Algerian Sahara. *BMC genetics*, *19*(1), 1-10.

Bruneau, P.,Pinheiro, P. Broeksema, B. Otjacques, B. (2015). Cluster sculptor, an interactive visual clustering system. *Neurocomputing,* 150: 627–644,

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998).Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95, 14863–14868.

Dash, R. C. and Hota, B. N. 1977. Some indigenous mango varieties of Orissa. *Orissa J. Hort.,* 5: 35-52.

De, Mitu, Das, Jhilik, Saha, Meghna, Pal, Ankush and Dey, Santi Ranjan. (2014). Documentation and Characterization of the Indigenous Mango (Mangifera indica L.) varieties of Malda and Murshidabad districts of West Bengal. *J. Environ. & Sociobiol.* (ISSN 0973-0834): 11(2), Dec: 187-198pp.

Galvez-Lopez D., Salvador-Figueroa M., Adriano-Anaya L. and Mayek-Perez N. (2010). Morphological Characterization of Native Mangoes from Chiapas, Mexico. *Subtropical Plant Science,* 62: pp 18-26.

Guijarro-Real, Carla, Jaime Prohens, Adrián Rodríguez-Burruezo, and Ana Fita. (2020). "Morphological Diversity and Bioactive Compounds in Wall Rocket (*Diplotaxis erucoides* (L.) DC.)" *Agronomy* 10, no. 2: 306.

Hastie, T., Tibshirani, R., Friedman, J. And Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*.; 9(5):473–6.

Ionov, Yurij; Rogovskyy, Artem S. (2020): Principal Component Analysis plot (A) and heatmap (B) generated, via the ClustVis, by the whole-unique-sequence-based analysis (WUSA). PLOS ONE. https://doi.org/10.1371/journal.pone.0226378.g004

International Plant Genetic Resourcces Institute IPGRI (2006). Descriptors for mango (*Mangifera indica* L). International Plant Genetic Resources Institute, Rome, Italy.

Liu, Yiming & Zhang, Xunzhong & Tran, Hong & Shan, Liang & Kim, Jeongwoon & Childs, Kevin &

Ervin, Erik & Frazier, Taylor & Zhao, Bingyu. (2015). Assessment of drought tolerance of 49 switchgrass (Panicum virgatum) genotypes using physiological and morphological parameters. *Biotechnology for Biofuels*, *8*(1), 1-18.

Majumder, D. N. 2013. Genetic Diversity in Mango (*Mangifera Indica* L.) through multivariate analysis. *Bangladesh J. Agril. Res*., 38(2): 343-353.

Majumder, D. N. (2013). Genetic Diversity in Mango (*Mangifera Indica* L.) through multivariate analysis. *Bangladesh J. Agril. Res*., 38(2): 343-353.

Majumder Dan, Hassan L, Rahim MA Kabir MA (2011). Studies on physio-morphology, floral biology and fruit characteristics of mango. *J Bangladesh Agri Univ.* 9 (2):187-199.

Metsalu, T and Vilo,J. (2015). ClustVis: a web tool for visualising clustering of multivariate data using PCA and Hetmap. *Nucleic Acid Research*. 43: Web Server issue.

Mukherjee, S.K. 1953. The mango-its botany cultivation, uses and future improvements, especially as observed in India. Econ. Bot. 7: 130–162.

Nam, E. J., Han, Y., Mueller, K., Zelenyuk, A. and Imre, D. (2007). Cluster Sculptor: A visual analytics tool for high-dimensional data. In Proc. IEEE VAST'07.

Njuguna J.K., Wepukhulu S.B. and Wanjala S. (2009). Mango cultivar evaluation programme in Kenya. *Acta Horticulturae*, 820: pp 133-135.

Pavlopoulos, G. A., Wegener, A.-L. & Schneider, R. (2008). A survey of visualization tools for biological network analysis. *Biodata mining* 1, 12.

Shen, Feng & Li, Yanxia & Zhang, Min & Awasthi, Mukesh & Ali, Amjad & li, Ronghua & Wang, Quan & Zhang, Zengqiang. (2016). Atmospheric Deposition-Carried Zn and Cd from a Zinc Smelter and Their Effects on Soil Microflora as Revealed by 16S rDNA. *Scientific Reports*. 6. 39148.

Sokal, R and Michener, C. 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 1409–1438..

Subedi A., Bajracharya J., Joshi B.K., Regmi H.N., Gupta S.R. and Hari B.C.K. 2005. Characterization and genetic diversity of mango (Mangifera indica L.) in Nepal. In: Sthapit BR, Upadhyay MP, Shrestha PK, Jarvis DI (eds.). On-farm conservation of agricultural biodiversity in Nepal. Volume I. Assessing the amount and distribution of genetic diversity on-farm. Proceedings of the Second National Workshop, 25- 27 Aug 2004, Nagarkot, Nepal. International Plant Genetic Resources Institute, Rome, Italy, pp. 167-175.

Toubiana D, Cabrera R, Salas E, et al. (2020). Morphological and metabolic profiling of a tropical-adapted potato association panel subjected to water recovery treatment reveals new insights into plant vigor. *The Plant Journal : for Cell and Molecular Biology.* Sep;103(6): 2193-2210.

Uddin MZ, Rahim MA, Barman JC Wadud MA. (2006). A study on the physical characteristics of some mango germplasm grown in Mymensingh condition. *Int J Sustain Crop Prod.* 1(2): 33-38

Wilkinson L, Friendly M. (2009). The history of the cluster heat map. The American Statistician; 63(2):179–84.

Zhao, Shilin & Guo, Yan & Sheng, Quanhu & Shyr, Yu. (2014). Advanced Heat Map and Clustering Analysis Using Heatmap3. BioMed research international. 986048. 10.1155/2014/9860