CHAPTER II

LITERATURE REVIEW

2.1 HISTORY OF THE GENUS *Elaeis*

The genus *Elaeis* is derived from a Greek word elaion, which means oil. The oil palm is today grown commercially in South-east Asia, Equatorial America, Africa and South Pacific. The high and increasing yields of the oil palm have led to a rapidly expanding world industry, now based in the tropical areas of Asia, Africa and America (Rajanaidu & Jalani, 1994). Its origin is believed to have been in Africa, but the most productive parts of the industry at present are in Malaysia and Indonesia, which provide most of the oil entering international trade (Oil World, 2012). As a commercial crop, its history is rather short, dating back to 1807 on the West African coast where its cultivation was commenced. It came to the East via the island of Mauritius in 1848, to Indonesia where four seedlings were planted and from there to Singapore Botanical Gardens about 1870 and only then into Malaya in 1875 (Jalani, 2012). The Family Palmae or Arecaceaae, where the genus *Elaeis* belongs is considered as old as any other family of flowering plants with fossils discovered in Cretaceous rocks dating back some 120 million years (Purseglove, 1972). Many plant taxonomists believe it to be the first monocotyledon to have branched out from primitive dicotyledonous stock and as such, the progenitor of all monocotyledons.

The world production of oil palm products has always been impossible to assess accurately owing to the quantities of produce that are not recorded, because they are produced in groves, smallholder plots and farms, and used for the farmer's domestic

purposes or sold locally (Oil World, 2012). Estimates suggest that world-wide production rose from 2.2 million tonnes of palm oil and 1.2 million tonnes of kernels in 1972 to 21 million tonnes of oil, 6 million tonnes of kernels and 2.6 million tonnes of kernel oil in 2000. Most of this increase can be attributed to Malaysia and Indonesia, and to some smaller Asian producers. The production of palm oil has now overtaken that of other vegetable oils, apart from soybean oil. The oil extracted from the oil palm is of two types which are palm oil and palm kernel, of which 90 % is used for food purposes and 10 % for non-food purposes (Jalani, 2012).

2.1.1 Evolution of Genus *Elaeis*

The genus *Elaeis* which belongs to subfamily *Cocoideae* is believed to have originated either in Africa or America and is one of the 240 genera of the family *Arecaceae.* It cannot be viewed in isolation from the other genera because of high degree of homogeneity among the chromosomes of the palms (Dransfield et al., 2005). The first species of the genus, *E. guineensis*, known as the African oil palm was described by Jacquin in 1763. The second species is *E. oleifera* known as the South American oil palm initially described as *E. melanococca* and applied by Gaertner in 1788 to be a form of *E. guineensis*. This species is distinguished from *E. guineensis* by its trunk. Purseglove (1972) however doubted the classification as *E. oleifera* hybridizes easily with *E. guineensis* and the degree of divergence between the two species rather justifies separation on a specific level rather than generic level. The third species was previously known as *Barcella odora*, but was named *Elaeis odora* by Wessels. Boer (1965); it is not cultivated, and little is known about it.

However, molecular markers indicated that inclusion of *E. odora* within the genus *Elaeis* is justified (Barcelos *et al*., 2002): the genetic distance between *E. odora* and

the other o species of *Elaeis* was similar to the distance between the latter, and less than the distance from *Cocos nucifera*, another member of the Cocosoideae subfamily. The fourth species, *E. madagascariensis*, was described by Beccari (1941) but is still controversial. Some taxonomists believe that these species is a variant of *E. guineensis* which was introduced around the 10th century when Africa influence entered Madagascar (Purseglove, 1972). The species is distinguished from *E.guineensis* by its male flower in which the fused filaments of the staminal tube are shorter and the anthers erect, instead of spreading at anthesis while the fruits are smaller, rounded and surrounded by larger bracts. Uhl & Dransfield (1987) recognized only two species, namely *E. guineensis* and *E. oleifera*, the third species as belonging to *Barcella odora* (Trail) Trail ex Drude and the fourth species as a variant of *E. guineensis* to Madagascar as mentioned above. For the purpose of this research, *E.guineensis* will only be extensively focused on.

2.1.2 Morphology and Biology of *E. guineensis* (African Palm oil)

Elaeis guineensis is a large, pinnate-leaved palm having a solitary columnar stem with short internodes. There are short spines on the leaf petiole and within the fruit bunch. The separate upper and lower ranks of leaflets on the rachis give the palm a characteristic untidy appearance. The palm is normally monoecious with male or female, but sometimes mixed, inflorescences developing in the axils of the leaves. The fruits are borne on a large, compact bunch. The fruit pulp, which provides palm oil, surrounds a nut, the shell of which encloses the palm kernel.

The early descriptions of the oil palm are listed in Hartley (1988). The only one of more than historical interest is the botanical description by Jacquin (1763). He described the palm from material from Martinique (to where it must have been

introduced); his description was detailed, but he described the flowers as either female or *hermaphroditi steriles* and seemed unaware that flowers of the two sexes were in separate inflorescences. The production of male and female inflorescences was first recorded by Miller in his *Gardener's dictionary* (London, 1768).

Before the end of the eighteenth century Gaertner (*De fructibus et seminibus plantarum*, Stuttgart, 1788) gave a more detailed description of the flower parts, recording that the male and female flowers are on separate inflorescences. Most of the early attempts at classification of varieties were unsatisfactory, as they were based on very limited acquaintance with the palm, and no knowledge of the inheritance of the characters described. However, it is the first description by Preuss (1902) of the *lisombe* palm, a name used in Congo, Cameroon and Nigeria for the thin-shelled *tenera* fruit form and still employed in quite recent times.

Janssens (1927) and Smith (1935) provided the first simple classifications which, in their essentials, have stood the test of time. Although nothing was known of the inheritance of the characters described, Janssens recognised that the fruit forms *dura* and *tenera*, distinguished by the thickness of shell, could be found in fruit types of different external appearance. Thus, both the common fruit type *nigrescens* and the green-fruited *virescens* were divided by Janssens into three forms, *dura*, *tenera* and *pisifera*.

2.1.3 Infraspecific Classification of *E. guineensis*

The taxonomy of *E. guineensis* variant forms has not yet been clarified. Several authors have described them as cultivars, varieties, subspecies or races. Purseglove (1972) believed that cultivars in the real sense do not occur in the species. He

therefore produced an infraspecific classification based on fruit characters, *viz., dura* (DD), *tenera* (Dd) and *pisifera* (dd), without committing them to specific taxonomic categories. Other author such as Whitemore (1979) debated the various forms *idolatrica, tenera, deli dura* and *pisifera* as varieties. As categories such as subspecies, variety or form may be inappropriate in the infraspecific classification of the species, the term race is proposed. Race is defined as a permanent variety or a microspecies. This term may prove appropriate because, in the genetic sense, races of a species are capable of exchanging genes and this has been amply demonstrated by hybridization between the races of *E. guineensis*.

Brief Descriptions of the various common races of *E. guineensis* (Willy, 2010).

- Race *dura* (DD): endocarp 2mm-8mm thick, comprising 25 % - 55 % of fruit weight; medium mesocarp content of 35 % - 55 % by weight; kernel tends to be large comprising 7 % - 20 % of fruit weight. The *Deli dura* is an ecological variant of *dura*, differing from latter in only minor differences such as kernel content (4 % - 10 %), mesocarp content (up to 65 %) and as such larger fruit size. The *macrocarya* form with endocarp 6mm-8mm thickness is an extreme *dura* race.
- Race *tenera* (Dd): endocarp 0.5mm-3mm thick, comprising 1 % - 32 % of fruit weight; medium to high mesocarp content of 60 % - 95 %; fibre ring darker in colour. The dividing line between *dura* and *tenera* is sometimes rather distinct as shown by the presence of a fibre ring in the mesocarp of the latter which is diagnostic; kernel usually smaller, comprising 3 % - 15 % of fruit. The oil content is higher at 24 % - 32 %.

- Race *pisifera* (dd): no endocarp, with small pea-like kernel in fertile fruits, but predominantly sterile with fruits rotting prematurely. It has a much higher sex ratio than *dura* and *tenera* and a significantly higher number of spikelets; even when fertile, the fruit to bunch ratio is low; infertile palms show strong vegetative growth.

2.1.4 Ecology and Morphology of *E. guineensis*

In its natural habitat in West Africa, *E. guineensis* often occur in association with *Raphia* or, if alone, in fresh water swamps (Chevalier, 1943). It is absent from undistributed tropical rain forests, where there is competition from the forest flora, absence of light due to the forest canopy and excessive moisture. It cannot survive or regenerate in high secondary forests for the same reasons. It may tolerate temporary flooding, if the water is not stagnant. Under semi-wild conditions, *E. guineensis* groves owe their presence to alteration of the natural vegetation by man for their development. In southern Nigeria, with its high population pressure, oil palm reaches its highest development and has given rise to palm groves. In both wild and semi-wild conditions, the yields are very low compared to plantation conditions, where individual plants are properly spaced and adequately cared for (Corley &Tinker, 2003).

The biggest area of semi-wild *E. guineensis* groves in the lowland regions of Western and Central Africa lies between 10^{o} N and 10^{o}S, which have a marked dry season lasting up to four to six months. The genus is ecologically suited to such conditions, although physiologically, it lowers yields. The present knowledge indicates that under plantation conditions, *Elaeis* cultivation can be extended from its natural habitat of 10^{o}N and 10^{o}S of Equator to 23^{o}N and 23^{o}S, i.e. between the tropics of Cancer and

Capricon. This has been brought about by man's deliberate planting of oil palm as it is an important component of world's vegetable oil supply (Latiff, 2000; Willy 2010).

It is generally known that *Elaeis* thrives well between a mean minimum temperature of 20^{o}C - 23^{o}C and a mean maximum temperature of 28^{o}C - 32^{o}C which normally occur in tropical countries. If the temperature falls below this, particularly at night temperature of below 19^{o}C, bunch development is affected and the yield reduced (Latiff, 2002). There is also evidence to show that a temperature below 15^{o}C stops growth in young seedlings. In Malaysia, oil palm can be grown on a wide range of soils. It is found that adequate soil moisture is more important than nutrient supply, which can be artificially supplied. In Peninsular Malaysia, the best areas are the coastal alluvial clay; Sabah's the riverine and coastal alluviums, and the soils of volcanic origin.

Morphology

- The oil palm has a typical adventitious root system. When the seed germinates, the radicle emerges through the germ-pore. The radicle continues to grow and is joined when the seed becomes detached by the primary roots. The short-lived radicle is replaced by adventitious roots emerging from the periphery of the radicle-hypocotyl junction. Primary roots emerge from the base of the swollen part. The primary roots descend deeply from the base of the trunk and spreads horizontally at varying depths in soil as the palm develops. They are up to 6mm-10mm in diameter and up to 20m in length. The secondary roots are produced from the pericycles of the primary roots amounting to 2mm-4mm in diameter ascending to the surface of the soil (Purseglove, 1972; Corley &Tinker, 2003).

- A trunk of an oil palm is not visible until the palm is three years or so old when the apex reaches its full diameter inform of an inverted cone. The base of the trunk is about 60 cm in diameter and the trunk is about 40 cm in diameter. The rate of trunk extension is about 25 cm - 50 cm per year. All mature palms have a solitary columnar stem with persistent frond bases. The stem of *E. guineensis* is stout and despite the crowded steeply ascending frond petioles bases, and then ultimately smooth. Frond bases persistenly adhere to the stem for at least 12 years and then fall away except for a few near the crown. The stem measures between 22 cm to 75 cm in diameter (Latiff, 2002).
- The frond consists of leaflets, each with a lamina and midrib, a central rachis to which the leaflets are attached, a petiole and a frond sheath. Only remnants of the frond sheath are visible externally. In a developing frond, the sheath is tubular and completely encloses but as it extends after growth of the sheath has stopped, the fibrous sheath splits and breaks up, leaving a row of spines along both edges of the petiole which are bases of fibrous sheath. The leaflets are long, ranging from 55 cm to 65 cm and often 100 cm. it is narrow with a width from 2.5 cm to 4.0 cm and a midrib with a number of parallel veins on the lamina. The cuticle is thick and has a very high resistance to diffusion of water vapour. Stomata are situated on the abaxial surface of the leaflets only. The leaflets are arranged in two planes (Latiff, 2002; Willy, 2010).
- The oil palm produces either a male or female or at certain stages a hermaphrodite inflorescence in each of the frond axil during the mature stage. An average female inflorescence may have more than 100 spikelets with over 4000 floral buds. The spikelets develop acropetally on the inflorescence. They are about 30 cm long with 12 - 30 flowers on central spikelets and less than 12

flowers on the lower and upper spikelets. Averagely, there are 700 - 1200 flowers per spikelet. Therefore, the total number of flowers is about 126000 and the number of pollen grains has been estimated to be in excess of 900 million with weight of 40 g per inflorescence. The presence of empty frond axils is an indication of abortion. The hermaphrodite inflorescence in *E. guineensis* is similar to that of *E. oleifera*. It has about 200 spikelets of varying lengths from 5 cm – 15 cm (Willy, 2010).

- The oil palm gives the highest yield of oil per unit area of any crop and produces two distinct oils, palm oil and palm kernel oil, both of which are extremely important in world trade. The time from flowering to harvesting of ripe fruits is five to six months. The fruits are borne on spikelets which are spirally arranged to form a compact bunch. The fruits are drupes of about 10 kg – 90 kg in weight. The pericarp comprises of three layers; exocarp (external layer), mesocarp (outer layer) and endocarp (internal layer). The exocarp protects the fruit; the mesocarp contains the oil and the endocarp is a hard shell containing the kernel which contains the kernel oil and the endosperm (Latiff, 2002; Willy, 2010).
- The seed is a nut which remains after the soft oily mesocarp has been removed usually by retting. It consists of an endocarp and possibly two or three kernels. The nut size varies greatly and is dependent both on the thickness of shell and size of kernel. It may be 2 cm - 3 cm in length and average 4g in weight. Deli *dura* and African nuts are larger, weighing up to 12 g, and the African *tenera* nuts are usually 2 cm or less in length and average 2 g (Corley &Tinker, 2003).

2.2 PLANT GENETIC RESOURCES

Plant genetic resources include the variation in crop plant genetic material that is accessible for present and potential future exploitation (FAO, 1996). The variation contains diversity which can be found at the nucleotide sequences levels, genotypes and alleles, and is required for the improvement of new cultivated varieties along with contributing to the flexibility of existing varieties (Hammer et al., 2003). Plant genetic resources are significant basis of genetic diversity and contribute to food security. Plant breeding is founded on the exploitation of genetic diversity within and between crop species and varieties (FAO, 1996).

2.2.1 Genetic Diversity

Genetic diversity has been defined as the germplasm of plants, animals or other organisms containing useful characters of actual or potential values, especially when these characters provide the variation in genes and genotypes between and within species or populations (Cromwell et al., 1999). It is the diversity that enables species to adapt to changing environments and provides an insurance against unknown future needs or conditions, thereby contributing to stability of farming systems at the local, national and global level. Genetic diversity comprises the total genetic variation present in a population or species. It is the differences within and between species or varieties in genes, alleles and genotypes, caused by mutation and recombination. Genetic diversity is the basis of selection in crop plants, it is essential for the development of new varieties by using novel combinations and traits (Preston, 2011).

In the field, genetic diversity among and between individuals and varieties is essential for resistance to pests and diseases, as well as tolerance and adaptation to climatic

conditions and changing climate (Preston, 2011). According to Newbury and Ford-Lloyd (1997), maintenance of the range and magnitude of genetic diversity present within a taxon is a primary aim of plant conservation. Therefore, to facilitate the conservation of genetic variation for present and future use, and to establish a baseline, the diversity of plant genetic resources, such as those held at MPOB needs to be characterized.

2.2.2 Characterization of Genetic Diversity

The level of genetic diversity present in populations is influenced by many factors including life form, breeding system, seed dispersal and geographic range. The overall result of this is a higher level of genetic diversity within populations and lower level of differentiation between populations in outbreeding species; a lower level of genetic diversity within populations and higher level of genetic diversity between population differentiations in inbreeding species (Hamrick & Godt, 1996). Genetic patterns of diversity may be variable across time such as in species due to changes in agriculture, and such as in ex-situ collections (due to genetic drift, cross pollination) and selective effects during regeneration, which can result in changes in genetic diversity and allele frequencies (Negra & Tiranti, 2009).

Genetic diversity has been explored using, predominantly, three marker types; morphology, proteins and DNA based methods. Before the advent of modern genetic technology, morphological markers were the classical method for characterization or estimation of genetic diversity. These comprised a diversity of traits and measurements that were recorded at all stages of plant development. Using morphological markers confer many advantages and morphological studies are often the first step in species studies and plant genetic resource activities serving to inform

molecular studies as to where diversity may exist (Karp et al., 1997). Advantages comprise the low cost and level of technology required, and the relation of markers to traits of agronomic importance (Newbury & Ford-Lloyd, 1997).

However, there remain limits to the usefulness or applicability of morphological markers, many of which are met by molecular methods. These include: the limited number of informative characters available, some of which may show little variation over material. The quantitative nature of their inheritance (being jointly influenced by genetics and environmental conditions of growth, because of this, some traits cannot be reliably isolated); related to this is the effect of environment that may mask the genetic coordinate and therefore influence genetic diversity estimates based on morphology (Spooner et al., 2005).

In the past, determination of the genetic structure of germplasm collections has mainly been done using passport data (van Hintem, 2000) or multivariate statistical methods such as cluster analysis, principal component analysis, multidimensional scaling; usually based on agronomic data (Franco et al., 2001).

2.3 CHEMOMETRICS

The development of the discipline chemometrics is strongly related to the use of computers in chemistry. Chemometrics is the chemical discipline that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments, and to provide maximum chemical information by analyzing chemical data (Matthias, 2007). Chemometrics is the use of mathematical and statistical methods for selecting optimal experiments and extracting maximum amount of information when analyzing multivariate data. Chemometrics can also be

called multivariate analysis (Matthias, 2007). This is because the statistical method of classification is usually by multivariate methods which include Principal Component Analysis (PCA), Cluster analysis and Discriminate analysis (Oyelola, 2004). They are all under exploratory statistical analysis. Wold and Sjostrom (1998) have reviewed that one of the successful area in chemometrics was pattern recognition method which is applied in both industry and academia.

Chemical pattern recognition method often used in chemometrics to seek the pattern from the multivariate data. In addition, Gonzalez (2012) has discussed that pattern recognition was part of Artificial Intelligence field that focused on finding the similarities and variation from large data sets in which brought about natural classification and grouping. In most cases, the raw data from the chemistry research were in form of chromatographic or spectroscopic. Hence, the chemist has dealt with chemometrics analysis to extract the maximum of useful information, thus, highlighted the differences in the chromatographic or spectroscopic results of different variables. As a result, exploratory data analysis (EDA) gave simple earlier visual idea of the main relationship between the objects and the variables (Brereton, 2003).

2.3.1 Principal Components Analysis (PCA)

PCA is a widely used mathematical tool for high dimension data analysis; that provides a guideline for how to reduce a complex dataset to one of lower dimensionality, to reveal any hidden, simplified structures that may underlie it (Jolliffe, 2002). Although PCA is a powerful tool capable of reducing dimensions and revealing relationships among data items, it has been traditionally viewed as a “blackbox" approach that is difficult to grasp for many of its users because of its coordinate transformation from original data space into Eigen space. PCA is a method

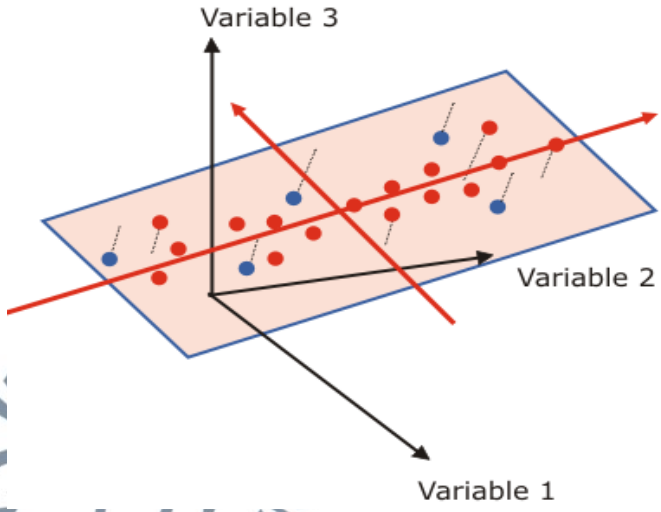
that projects a dataset to a new coordinate system by determining the eigenvectors and eigenvalues of a matrix. It involves a calculation of a covariance matrix of a dataset to minimize the redundancy and maximize the variance. Mathematically, PCA is defined as an orthogonal linear transformation and assumes all basis vectors are an orthonomal matrix (Jolliffe, 2002). PCA is concerned with finding the variances and coefficients of a dataset by finding the eigenvalues and eigenvectors.

PCA is one of the most useful statistical tools for screening multivariate data with significantly high correlations. Information from PCA may assist the plant breeder to identify limited traits for using in hybridization and selection programs (Doumbia et al., 2013). The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. PCA tool is used to interpret the variables that described the differences between samples, thus, identifies which variables contributed most to an observed difference or which variables were correlated. In short, PCA is the most basic workhorse of multivariate data analysis that acts as fundamental technique for other methods in “The Unscrambler” (CAMO, 2014).

Basically, the principle of PCA aims to find the direction in multidimensional space along which the distance between the data points is the largest thereby indicating the maximum variance in data set. In other words, it constructed as linear combinations of the initial variables that contributed most in making the samples different from each

other, thus minimizing the squared distance from the line to each point of data set. Data point was the representation of each sample in a data table, thus, the location was determined by the cell values of the corresponding row in the table. Thus, each variable act as coordinate axis in multidimensional space. PCs are computed iteratively where the first PC contained largest information or most explained variance while the subsequence PC contained less explained variance than previous one. Based on the geometric concept, a new set of coordinate axis were formed as shown in Figure 1 below where PCs were orthogonal to each other. Thus, the user can focus on the first few PCs in the interpretation which contained larger information than the subsequence ones (Esbensen *et al*., 2002; CAMO, 2014).

Figure 1: Score plots where PCs 1 and 2 are orthogonal to each other (CAMO, 2014)



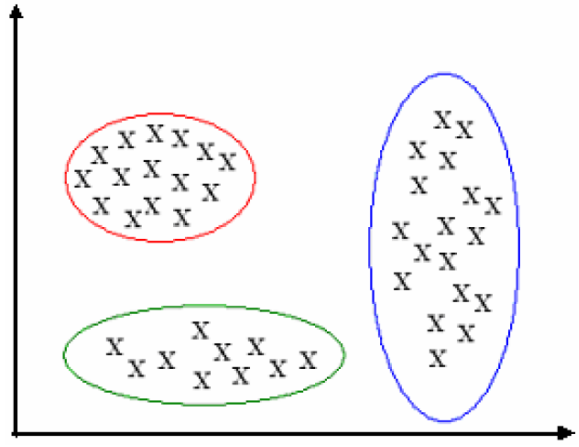
2.3.2 Clustering Analysis

Clustering deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to

other clusters (Manpreet et al., 2008). Thereby, clustering is the grouping of data items based on their similarity. Cluster analysis is classified as unsupervised pattern recognition due to the discovery of the data structures without providing any explanation as well as the pre-information about the data not being used in the classification (CAMO, 2014).

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion, which would be independent of the final aim of the clustering. Consequently, it is the user, which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection) (Manpreet et al., 2008).

Figure 2: Clustering pattern



Clustering can be in the form of distance-based clustering and conceptual clustering. Distance – based cluster is when two or more objects that belong to the same cluster

are “close” according to a given distance. On the other hand, conceptual clustering occurs when two or more objects that belong to the same cluster are defined by common concept. In other words, objects are grouped according to their fit to descriptive concepts not according to simple similarity measures.

2.3.2.1 Types of Clustering

The Unscrambler software offers two main groups of cluster analysis which are classified as non-hierarchical clustering (*k*-means and *k*-medians) and hierarchical cluster analysis (HCA).

K-means Clustering

The target of the clustering is to minimize the variability of the samples within the clusters along with maximizing the variability between clusters formed starting with *k* random clusters (StatSoft, 2013). The theory of *k*-means clustering algorithm provided by the Unscrambler starts with the randomized set of samples distribution between the defined *k* random clusters. Every defined clusters centroid is calculated using averages (*k*-means) method, then, the distance between each of the samples to its centroid is measured within the respective clusters. Consequently, the number of clusters *k* with the shortest distance between the samples to the centroid is selected by moving a set of samples to that particular *k* cluster. This nonhierarchical algorithm is repeated until there are no changes occurring and once the centroids have identified the natural clustering of points (Tan et al., 2006).

K- median clustering

In statistics and data mining, k-medians clustering is a cluster analysis algorithm. It is a variation of k-means clustering where instead of calculating the mean for each

cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the 1-norm distance metric, as opposed to the square of the 2-norm distance metric (which k-means does.). K median is used when one wishes to minimize the total 1-norm distance from each point to its nearest cluster center (Jain & Dubes, 1988; Bradley et al., 1997).

Hierarchical Clustering Analysis (HCA)

HCA used different linkage methods along with specific distance measure to create clusters. The chosen of both factors need to be appropriately selected based on the application domain as well as base on the real-world interpretation because they can affect the grouping results. Consequently, a dendrogram is created to portray the results of clusters arrangement produced yet formed based on the meet of triangle inequality of metric determined from the distance between samples (CAMO, 2014). Basically, HCA begins by treating each sample within its class, then, proceeded by combining the samples into another cluster based on their similarity until the clusters emerged to form one larger cluster (CAMO, 2014).

As discussed earlier, the distance measure between samples is important to be determined for the first part in relating the samples into the same group based on its closest distance. Hence, the Unscrambler offers several options for the distance measures between samples such as Euclidean distance, squared Euclidean distance, city-block distance and Chebyshev distance**.** Among these methods, Euclidean distance is the most usually chosen in measurement of the distance between two samples of the data sets that are suitably normalized by taken into consideration the different between two samples directly based on the magnitude of changes in the sample levels. The squared Euclidean distance intended to normalize the data by

measuring the similarity between clusters where some variable may control the distance between groups (Hoon *et al*., 2013 and CAMO, 2014).

2.4 PREVIOUS STUDIES

Principal component analysis and cluster analysis have been widely used in various studies for delineating the variability in large group of genotypes of many crop species, some of which are outlined below:

Shivani and Sreelakshmi (2014) assessed the genetic diversity in indigenous germplasm lines of safflower using multivariate tools. Principal component analysis revealed that 99% of the total diversity was explained on the basis of the first three principal components. The distribution of accessions within eight well defined clusters was visible as well with no apparent relationship with the geographical origin.

In evaluating the variation in seed vigour characters of West African rice (*Oryza sativa L*.) genotypes, seeds of 24 West African rice genotypes were selected based on seed vigour traits in the laboratory and field in two cropping seasons at the research farm of Federal University of Agriculture, Abeokuta, Nigeria. The analysis was subjected to principal component analysis and clustering analysis. The first three axes of the PCs across the two seasons captured 86.34% of the total variation. Cluster analysis classified these genotypes into four distinct groups based on germination and emergence percentages. Those characters identified by PCA could be included in the crop improvement programme for improved seed quality within West African low land rice germplasm (Adebisi et al., 2013).

Genetic diversity based on cluster and principal component analyses for yield and quality attributes in ginger was studied by Ravishanker et al. (2013). Through cluster

analysis, 25 genotypes were grouped into five main clusters while first six PCs with eigenvalues greater than one accounted for 76.19% of the total variability.

The comparative study of cowpea germplasms diversity from Ghana and Mali using morphological characteristics was carried out by Doumbia et al. (2013). Results from PCA indicated that first PC explained a total variation of 72.05% while UPGMA showed that at 0.97 level of similarity, almost all the 94 accessions studied were distinct from each other while at 0.95 levels, they were similar to each other.

Muhammad et al. (2013) utilized PCA and cluster analysis to study the genetic divergence in indigenous spinach. Results obtained shows that the cumulative contribution of the first three PCs reflected by PCA was 58.4 % while cluster analysis grouped the spinach genotypes into four major clusters regardless of collection origin.

In the research of WeldeMicheal et al. (2013), forty nine coffee germplasm (*Coffea arabica L.*) accessions were collected from Gomma Wereda in other to estimate information on genetic diversity. The experiment was conducted in simple lattice design with two replications during cropping season by superimposing on six years old coffee trees and grown under uniform coffee shade tree (*Sesbania sesban*) conditions. Data from 26 quantitative characters were recorded. The analysis of variance showed a significant variation among the accessions for all morphological traits. This shows the existence of variability among the tested materials. Cluster analysis grouped the 49 germplasm accessions into five clusters which make the accessions to be moderately divergent. The distances between these clusters are highly significant ($P<0.01$) which suggested a suitable hybridization program is possible. Principal component analysis showed a variation in the first two principal components which is explained by the larger share of the observed variation.

Ajmal et al. (2013) used multivariate analysis to study the genetic relationships among wheat germplasm comprising of 50 genotypes. The first three PCs with eigenvalues greater than one contributed to 70.59% of the variability amongst genotypes while cluster analysis set apart the 50 genotypes into 5 clusters based on Ward's method.

Principal component analysis was done by Lohani et al. (2012) to estimate the variability in germplasm of potato. The PCA of 54 potato genotypes based on correlation matrix of agronomic and quality traits showed that first 11 components explained 96.25% variation. Non-hierarchical Euclidean Cluster analysis based on PCA grouped the 54 potato genotypes into seven non-overlapping clusters.

Darvishzadeh et al. (2012) studied on the genetic diversity in population of Iranian Ajowan (*carum copticum L.*) based on agronomical and morphological characteristics. Ten populations were collected from different regions and evaluated in a completely randomized design with eight to ten replications. Among the characteristics studied, a high coefficient of variation was observed for the number of seeds (197.58), plant yield (57.56) and shoot dry matter (56.28). Cluster analysis using Ward's method classified ten populations of Ajowan into four groups. PCA using 18 agronomical and morphological characteristics indicated that the first two principal components with eigenvalue of more than one accounted for 74.5% of the total variance. Both PCA and cluster analysis result were consistent which displayed a considerable diversity of agronomical and morphological is useful in germplasm management.

In order to reveal the underlying association between agronomic traits of spring wheat in eastern region of Qinghai Tibet Plateau, Deyong (2011) used partial correlation analysis, PCA and cluster analysis to study the relationship between yield

grain and agronomic traits. PCA was able to indicate genotypes with high-yield potential and the varieties were grouped into 14 groups through cluster analysis.

Sanni et al. (2010) studied on the variability of 434 accessions of rice germplasm from Cote d'lvoire and evaluated 14 agro morphological traits in upland condition using experimental design and analyzed with multivariate methods. The result suggested that traits such as plant height, leaf length, tillering ability, number of days to heading and maturity, panicle length and grain size were the principal discriminatory characteristics. The result of PCA indicated that first three PCs explained about 58.41% of the total variation among the 14 characters studied. Seven cluster groups were obtained using the unweighted variable pair group method of the average linkage cluster analysis.

Multivariate statistical analysis was used by Lacis et al. (2010) to determine phonotypical variability and genetic diversity among 40 sour cherry accessions. Both cluster analysis and PCA showed an adequate grouping of accessions according to phenotypical data and known pedigree data. Nine components with eigenvalues larger than 1.0 were extracted which described 80% of the variability of the original traits while cluster analysis identified four main clusters.

Phenotypic diversity for quantitative and qualitative traits in a collection of pepper germplasm from different areas of Turkey was assessed by Bozokalfa et al. (2009). All the accessions were characterized for 67 agro-morphological traits and the morphological traits were subjected to PCA followed by hierarchical agglomerative clustering. The first six PC axes accounted for 54.29% of the variance among the 48 accessions and their lines while seven groups were created based on morphological and agronomic properties.

Peerasak et al. (2009) evaluated genetic variation in cultivated mungbean germplasm to assess the extent and pattern of their diversity. He evaluated 9 qualitative and 21 quantitative traits in 340 diverse cultivated mungbean accessions collected at the world vegetable center, Taiwan. The germplasm displayed a wide range of diversity for most of the traits evaluated. Cluster analysis grouped the germplasm into 5 major and 1 minor cluster. Principal component analysis revealed that the first three PCs explained 74.9% of the total variation. It is recommended that germplasm from west Asia be exploited more in cultivar programs. The results can help mungbean breeders choose the right combinations of parental genotypes carrying the desirable characters.

In the study of genetic variability of some quality traits in *Lathyrus spp*. germpalsm, 66 accessions representing eighteen species of the genus lathyrus were collected from different regions evaluated for variation of quality traits. Cluster analysis of coefficient of variance values for each accession identified the 66 accessions into 8 groups. High variability was attributed both genetic and environmental factors exhibited at both inter-specific and intra-specific levels. The most promising accession for breeding programs was *L. sativus* from Tunisia (Sammour et al., 2007).