CHAPTER II

LITERATURE REVIEW

2.1 Origin, Taxonomy, Botany and Development of Oil Palm

2.1.1 Origin

The origin of the oil palm is believed to be in Africa, more precisely to be in the Gulf of Guinea. Wild and semi-wild palm groves are found in a coastal belt running from the Northernmost Senegal region through Sierra Leone, Liberia, Ivory Coast, Ghana, Togo, Benin, Nigeria, Cameroon, and The People's Republic of Congo up to the Southernmost Democratic Republic of Congo and Angola. The center of origin and diversity of the oil palm appears to be concentrated in the tropical forests of Nigeria, Cameroon, Congo and Angola (Verheye, 2010; Corley & Tinker, 2003).

The oil palm was introduced in South-east Asia in 1848, when four seedlings were planted in Buitenzorg (now Bogor) Botanic Garden in Java (Indonesia). Materials derived from this introduction were used to establish the first plantations in Indonesia and Malaysia (Ngando-Ebongue., et al., 2012; Rajanaidu, 1994). The center of major production of oil palm however, is in South East Asia with most production realized in Malaysia and Indonesia. Both countries contribute about 85 % of the world palm oil production (Rivas et al., 2012; Wahid et al., 2005).

2.1.2 Taxonomy

The oil palm is a tropical perennial monocotyledonous tree plant. The Genus name is from the Greek work “elaion” meaning oil and “elaia” meaning olive. The scientific classification of oil palm is as follows: (Ngando-Ebongue et al., 2012; Dransfield et al., 2005).

Kingdom:	Plantae
Division:	Spermatophyta
Subdivision:	Monocotyledonae
Order:	Arecales
Family:	Arecaceae
Subfamily:	Arecoideae
Tribe:	Cocoseae
Sub tribe:	Elaeidinae
Genus:	*Elaeis*
Species:	*guineensis* and *oleifera*

2.1.3 Botany

Oil palm is a diploid (Chromosome number of 2n=32) oleaginous tropical perennial crop (Dransfield et al., 2005). *Elaeis guineensis* is a feather palm with its stem topped by 35-60 pinnate fronds borne on a columnar stem. A 20-year oil palm would be about 12 m when the height is measured from the ground to base of frond 41. Oil palm

produces about 2-3 fronds per month in a regular sequence and each frond consists of a rachis and petiole. The petiole is about 150 cm long and 300-400 leaflets are attached to the rachis. Each leaflet is around 130 cm in length. Oil palm is monoecious and allogamous, producing separate male and female inflorescences on the same plant in alternate cycles. The duration of each floral cycle fluctuates greatly, largely influenced by genetic and environmental factors. Although monoecious, it is largely cross pollinated (Verheye, 2010; Corley & Tinker, 2003; Rajanaidu et al., 2000).

The inflorescence is a compound spike carried on a stout peduncle. The number of spikelets per inflorescence is nearly the same in both genders-about 100 to 110 spikelets but the number of male and female flowers per spikelet is highly variable, i.e. 700-1200 flowers in male as compared with 5-30 flowers in the female spikelets. The flowers are bisexual but in 'males', the stigmas are suppressed while in the 'female' flowers the stamens are undeveloped. A large quantity of pollen (ca. 30 g to 40 g) is produced by each male inflorescence.

It has been established that oil palm is mainly insect-pollinated especially by the imported weevil, *Elaeidobius kamerunicus* (Faust). At anthesis, a female flower is receptive for 36 hr to 48 hr. The flower within an inflorescence anthesize sporadically and there may be an interval of up to four weeks between the initial and the last flushes, most of the flowers anthesize within the first week (Verheye, 2010; Corley & Tinker, 2003; Rajanaidu et al., 2000; Rajanaidu, 1999)

2.1.4 Cultivars and Classification

According to Verheye (2010), cultivars of oil palm in the strict sense do not occur. As the oil palm is monoecious and cross pollinated, individual palms are usually

heterozygous, clonal material cannot be made. Classification of cultivars is based on fruit forms and fruit types.

2.1.4.1 Classification Based on Fruit Colour

The differences shown by fruit colour are caused by a monogenetic inheritance. The variation in fruit colour is an important indication of carotene content and some recessive plants do not produce carotene at all. The classification of fruits of *E. guineensis* on the basis of the colour of their exocarp is as follows: Nigrescens which is the most common are unripe fruits, deep violet to black at the apex and ivory coloured towards the base. They also have the highest carotenoid contents.

The Nigrescens are of two types: Rubo-nigrescens which are ripe fruits deep reddish orange with brown cap on the upper half. While the other, Rutilo-nigrescens are ripe fruits paler orange with black cap on the upper half. Virescens are unripe fruits green which ripen to a light reddish-orange with small greenish tip, anthocyanins little or absent.

The last type are the Albescens which are extremely rare, unripe fruits deep green, ripening to pale yellow or ivory with blackish or green cap in the upper half; little or no carotene present (Corley & Tinker, 2003; Latiff, 2000).

2.1.4.2 Classification Based on Fruit Forms

In addition to differentiation by their fruit pigmentation and characteristics, *dura*, *pisifera* and *tenera* are classified according to endocarp, or shell thickness and mesocarp content. Fruit of the dominant homozygote *dura* (Sh+ Sh+) have 2-8 mm thick shell (endocarp) comprising 25-55 % of weight of fruit, medium mesocarp content of 35-55 % of weight. The heterozygote *tenera* race (Sh+ Sh-) has 0.5-3 mm

thick endocarp comprising of 1-32 % weight of fruit, medium to high mesocarp content of 60-95 % and the recessive homozygote *pisifera* (Sh- Sh-) are female sterile and have no endocarp (shell less) and about 95 % mesocarp. The *tenera* is the result of hybridization between *dura* and *pisifera* and is of high commercial value (Ngando-Ebongue, 2012; Verheye, 2010; Corley & Tinker, 2003; Teoh, 2002).

2.1.5 Oil Palm Breeding and Development

The earliest record of palm into Far East Asia was four seedlings planted in a botanical garden 1848 in Java in the Dutch East Indies. Two of these were from Amsterdam botanical garden, the other from Bogor or Mauritius in Indian Ocean. The palm that sprang from these seedlings formed the entire source of planting materials in Malaysia (Okwuagwu, 2013).

2.1.5.1 Oil Palm Breeding in Malaysia

Oil palm is currently the most important plantation crop in Malaysia, and since 1971, Malaysia has been one of the world's largest producer and exporter of palm oil. Deli *dura*, is the breeding population used in most of the commercial programmes for hybrid production of oil palm plantations in Malaysia and elsewhere. The Deli *dura* provided the foundation for development of planting materials for oil palm industry in Malaysia and is widely used for seed production and genetic improvement programs in Malaysia and Indonesia (Kushairi, 2009; Teoh, 2002; Kushairi & Rajanaidu, 2000).

The *pisifera* palms are predominantly female sterile, where bunches do not develop beyond the stage of anthesis. They cannot be exploited for commercial planting, instead they are used as the male parents in *dura* x *pisifera* (D x P) crosses to generate the *tenera* hybrids. This was made possible after Beinaert and Vanderweyen (1941)

showed the monogenic inheritance if the shell gene. Since then, the *tenera* has been used as commercial material in industrial plantations throughout the world. This discovery was the cornerstone that revolutionized the oil palm industry into more productive D x P or *tenera* planting materials (Ngando-Ebongue, 2012; Kushairi & Rajanaidu, 2000).

According to Rajanaidu et al., (2000), the goal of palm breeding and selection methodology is focused mainly on producing quality materials for maximum mesocarp and kernel. As a result of breeding and selection, the Malaysian palm oil board (MPOB) had developed various breeding populations with special traits (PS series) using the elite *dura* and *pisifera*. To date MPOB has developed PS1 (dwarf palm), PS2 (high iodine value), PS3 (large kernel), PS4 (high carotene *E. oleifera*), PS5 (large fruit *dura*), PS6 (thin shell *tenera*), PS7 (high bunch index), PS8 (high vitamin E), PS9 (*Bactris gasipeas*-not an oil palm species), PS10 (long stalk), PS11 (high carotene *E. guineensis*), PS12 (high oleic acid) and PS13 (low lipase) (Kushairi, 2009; Teoh, 2002).

2.1.5.2 Oil Palm Breeding in Nigeria

When plantation had expanded in South East Asia, Nigeria was still depending on the wild groves. At the turn of the 20th century, Nigeria found it necessary to have research stations to develop plantation technology and by 1922, genetic selections were made on the wild groves. By 1939, Nigeria institute for oil palm research (NIFOR) was established, where all the genetic materials were assembled, evaluated and developed into a plantation crop (Okwuagwu, 2013). Oil palm breeding programme in Nigeria started in 1939 at the oil palm research station (OPRS) which later became WAIFOR, then NIFOR main station. The breeding programmme has a

relatively broader range of materials compared with other centres worldwide. NIFOR obtained materials from palm groves and through international exchange programmes.

Among notable germplasm prospection made by NIFOR was in 1960 and 1973. The one made in 1973 was in collaboration with the Malaysian Agricultural Research and Development Institute (MARDI) (Okwuagwu, 1986). The foundation materials of NIFOR breeding programme were based on mass selection of *tenera* and *dura* grove palms, where 12 palms at Calabar, 11 palms at Aba and 166 palms at Ufuma were selected. Current materials in breeding and seed production are based mainly on the selection and introgression of eight *dura* parents originating from Opobo, Delta, Calabar, Aba, Ufuma, Angola and Ecuador, as well as five Deli *dura* sources from Serdang, United plantations IRHO (ex Pobe and ex La Me) and Banting (Corley & Tinker, 2003; Kushairi & Rajanaidu, 2000).

2.2 GENETIC RESOURCES

Genetic resources can be defined as all materials that are available for improvement of a cultivated plant species (Haussmann et al., 2004). It could also be referred according to Gepts (2006) as the sum total of genes, gene combinations, or genotypes available for the genetic improvement of crop plants.

It has generally been recognized that the narrowness of gene pools has been a major obstacle to selection progress in oil palm (Arasu & Rajanaidu, 1975). Due to the narrowness of the oil palm gene pools, voices were raised by many organizations for the need to enrich genetic variability in oil palm from their natural environments (Rajanaidu, 1994). According to Corchard et al. (2009), the amount of genetic diversity and understanding of the genetic structure of natural oil palm collections in

relation to breeding populations are guides to the introduction of new materials in the base populations of breeding programmes. Rajanaidu and Ainul (2013), also stated that the knowledge of genetic variability and relation among oil palm germplasm is essential for selection of promising materials and also the presence of genetic variability in germplasm helps to increase the breeding efficiency, resulting in higher selection gain.

It has been stated that until a collection has been properly evaluated and its attributes become known to breeders, it has little practical use (Sapey et al., 2012). Strelchenko et al. (1998) also postulated that in order to optimize and accelerate breeding, it is essential to screen, evaluate and classify the genetic variability available in the germplasm. The author maintained that genetic variability is especially important for collecting, maintaining ex-situ and study plant genetic resources in germplasm program. Collection, conservation and evaluation of germplasm offer an opportunity of exploiting advantages of natural diversity for crop improvement (Thomas & Mathur, 1991).

2.2.1 Germplasm Collection

The role of germplasm of cultivated plants has been well recognized. Germplasm is the total gene pool of a species consisting of landraces, advanced breeding lines, popular cultivars, wild and weedy relatives. It forms the raw material for any crop improvement program (Upadhyaya et al., 2010). Vavilov (1951) was the first to recognize the importance of genetic diversity for crop improvement and organized extensive germplasm collections of various crops from their ‘centers of origin’ and distribution for conservation. Since then the germplasm collections of major crop plants continued to grow in number and size in the world.

According to Marshall and Brown (1975), the objective of germplasm collection is to gather maximum amount of genetic variability in the minimum of samples. The germplasm collection of any crop consists of diverse types of collections such as those derived from centres of diversity, areas of cultivation and breeding programmes (Thomas & Mathur, 1991). Okwuagwu et al. (2011) reported that the need for extensive germplasm collection to broaden the base of the oil palm breeding material and to safeguard crop vulnerability, inherent in growing of uniform and closely related cultivars over wide areas has become imperative.

In the case of the oil palm, due to its narrow genetic base, MPOB engaged in series of germplasm collection expeditions to gather oil palm breeding materials from different countries in various part of the world, viz: Nigeria, Cameroon, Zaire (Congo DR), Tanzania, Madagascar, Angola, Senegal, Gambia, Sierra Leone, Guinea and Ghana. The allied species *E. oleifera* was sampled in Honduras, Nicaragua, Costa Rica, Panama, Colombia, Brazil, Suriname and Ecuador (Kushairi, 2009; Hayati et al., 2004; Rajanaidu, 1994; Rajanaidu & Jalani, 1994). Two to thirty-two palms were collected and the number of sites visited in a country depended on the distribution and density of natural palm groves.

In 1973, MARDI and NIFOR in collaboration collected oil palm genetic materials at 45 sites with an average of 20 palms per site and 200 seeds per palm. A total of 919 (595 *duras* and 324 *teneras*) bunches were harvested (Corley & Tinker, 2003; Rajanaidu, 1994; Obasola et al., 1983). Some of the outstanding planting materials from the Nigerian collection have been used to initiate and develop an entirely new breeding programme (Rajanaidu, 1994).

2.2.2 Germplasm Evaluation

Germplasm evaluation, in the broad sense and in the context of genetic resources, is the description of the material in a collection. It covers the whole range of activities, starting from the receipt of samples by the curator and growing these for seed increase, characterization and preliminary evaluation, and also for further or detailed evaluation and documentation (Maji & Shaibu, 2012; Sapey et al., 2012; Thomas & Mathur, 1991).

After collection of germplasm, there is need for its systematic evaluation in order to know its various morphological, physiological and developmental characters and other special features such as stress tolerance, pest and disease resistance etc. (Thomas & Mathur, 1991). The evaluation involves assaying germplasm for agronomic traits which interact with the environment (Upadhyaya et al., 2010). Sapey et al. (2012) collected oil palm germplasm in the northern regions of Ghana for evaluation. Data on stem height, bunch weight, bunch length, bunch depth, bunch diameter, bunch width, bunch spine, bunch stalk weight, fruit length and width were recorded in-situ while mesocarp, kernel and shell to fruit ratios were computed.

Okwuagwu et al. (2011) collected fourteen oil palm accessions in ten locations from the highland areas of Afikpo in Eastern Nigeria. Data on stem height, bunch weight, bunch length, bunch width, bunch spine length, bunch stalk weight, fruit length and width were also evaluated and recorded in-situ while mesocarp, kernel and shell to fruit ratios were computed. Okoye and Okwuagwu (2008) have also evaluated the performance of 51 D x T inter-population progenies, derived from the second cycle modified reciprocal recurrent selection (RRS) breeding programme of NIFOR. Data on number of harvested bunches (BN), single bunch weight (SBN) and fresh fruit

bunch (FFB) yield was collected in order to estimate variability among the progenies and to identify the best performing progenies for introduction into the locality where they conducted the experiment.

For the MPOB-Nigerian germplasm, detailed data on yield and its components, bunch parameters, vegetative characters, physiological parameters and fatty acid composition have been collected (Rajanaidu et al., 2000). Rajanaidu and Rao (1988) as well as Rosenquist (1990) evaluated the performance of Nigerian genetic material in detail. Base on the performance of populations and families within populations, elite palms have been selected on traits for high yield and dwarfness, high iodine value and high kernel content (Rajanaidu et al., 2000).

2.2.3 Germplasm Utilization

The potential of a germplasm sample is largely unknown at the time of its collection and a number of desirable characters are identified whenever a diverse group of germplasm samples are evaluated and screened (Thomas & Mathur, 1991). The oil palm germplasm at the MPOB have been evaluated and is being utilized in a number of ways. About 3 % *teneras* in the Nigerian collection had oil yields comparable to current planting materials, as a result some of the extremely outstanding Nigerian *teneras* have been progeny tested with a range of Deli *dura* available in the industry and MPOB. The *pisiferas* in the Nigerian collection too are progeny tested to study their combining ability with Deli *duras* (Rajanaidu, et al., 2000).

The MPOB germplasm materials are also used in broadening the genetic base of the Deli *duras* and *teneras*. The outstanding *dura* and *pisifera* palms from the Nigerian collection have also been used to initiate an entirely new breeding programme, with

the objective of producing superior alternatives to the current Deli *duras* and modern *pisiferas*. MPOB also embarked on a programme to diversify the current oil palm breeding materials with the elite Nigerian palms, as a result of this, PORIM Series 1 (PS1) – high yielding dwarf palms, PORIM Series 2 (PS2) – high in yields and high iodine values and PORIM Series 3 (PS3) – high yielding Nigerian *dura* palms with high kernel content were all developed (Rajanaidu et al., 2000; Latiff, 2000; Rajanaidu, 1999).

2.2.4 Germplasm Conservation

Conservation involves all activities including collecting, maintenance, storage, management, protection and regeneration aimed at ensuring the continued existence, evolution and availability of genetic resources; in-situ and ex-situ (Gonzalez-Benito et al., 2004; Rao, 2004; Rao & Hodgkin, 2002; Koski et al., 1997). Conservation of genetic resources can be done in-situ or ex-situ. in-situ conservation is the conservation of genetic resources on site, in the natural and original population, or on the site where genetic resources of a particular population developed their distinctive properties (Gepts, 2006; Koski et al., 1997). ex-situ conservation on the other hand is the conservation method that entails removal of individual plants or propagating material from its site of natural occurrence, i.e., conservation "off-site" in gene banks as seed, tissue or pollen; in plantations or in other live collections (Gepts, 2006; Fao, 1993).

The oil palm germplasm are being preserved in-situ (natural palm groves in Africa), in vitro (cryopreservation technique) and field gene bank (ex-situ) (Rajanaidu et al., 2000). At present, the field gene bank of MPOB represents the largest ex-situ oil palm

germplasm in the world with about 59 625 palms covering 406 ha (1000 ac) (Kushairi, 2009; Rajanaidu et. al., 2000).

2.3 GERMPLASM CHARACTERIZATION AND EVALUATION

Characterization is the description of a character or its quality. The word "characterize" is also synonymous to "distinguish" that is, to mark as separate or different, or to separate into kinds, classes or categories. Characterization of genetic resources refers to the process by which accessions are identified or differentiated. This identification may in broad term refer to any difference in appearance or make up an accession. Characterization in gene banks and germplasm management stands for the description of characters that are usually heritable, easily seen by the eye and equally expressed in all environments (IPGRI/CIP, 2003).

Germplasm characterization according to Ferreira (2006) refers to the observation, measurement and documentation of heritable plant traits in a collection. The resulting data allows for identifying and classifying accessions, which is essential for collection management. Characterization of plant germplasm therefore aims at describing and understanding the genetic diversity of the accessions in question.

2.3.1 Genetic Diversity Measurement Tools

According to Aremu (2011), techniques for measuring or evaluating germplasm include use of multivariate data collected from morphological or agronomic traits which may effectively display discrete, continuous, binomial or ordinal variable; and use of multiple sets arising from morphological, biochemical and DNA-based data.

Standard characterization and evaluation of accessions may be routinely carried out using different methods, including traditional practices such as the use of descriptor

lists of morphological characters. They may also involve evaluation of agronomic performance under various environmental conditions. In contrast, genetic characterization refers to the description of attributes that follow a Mendelian inheritance or that involve specific DNA sequences. It involves the application of molecular markers and the identification of particular sequences through diverse genomic approaches (de Vicente et al., 2005).

According to Ferreira (2006), the current characterization of plant germplasm collections relies strongly on morphological descriptors as they are reliable, easy to study and relatively low cost to evaluate. However, in recent times, germplasm characterization based on molecular methods has experienced great development. The method which reveals sequence polymorphism known as "molecular markers" has fomented a revolution in the speed and quality of large-scale plant germplasm characterization.

Adequate characterization for agronomic and morphological traits is necessary to facilitate utilization of germplasm by breeders and in order to achieve this, germplasm accessions of all crops are characterized for morphological and agronomic traits in batches over the years (Upadhayaya et al., 2008). As stated Upadhayaya et al. (2008), the major objectives of germplasm characterization are:

i. Describe accessions, establish their diagnostic characteristics and identify duplicates;
ii. Classify groups of accessions using sound criteria;
iii. Identify accessions with desired agronomic traits and select entries for more precise evaluation;

iv. Develop interrelationships between or among traits and between geographic groups of cultivars; and

v. Estimate the extent of variation in the collection.

The accessibility of collections depends largely on the information available on them. Both characterization and evaluation data provide an effective source of information for genetic diversity studies. The results obtained can be used to help understand pattern of variation in crop species and to identify groups of accessions with high diversity or with shared characteristics (Rao & Hodgkin, 2002). Quite a number of researches have been carried out on oil palm using both morphological descriptors and molecular markers.

Noh et al. (2012) evaluated the performance of 11 oil palm AVROS *pisiferas* based on their 40 *dura* x *pisifera* (D x P) progenies tested on inland soils predominantly of Serdang series. Analysis of variance (ANOVA) showed low genetic variability among *pisifera* parents for most of the characters indicating uniformity of the *pisifera* population.

In order to find possible genetic differences and or similarities among commercial oil palm materials, Arias et al. (2012) assessed 17 microsatellite markers in nine crosses (D x P) of *E. guineensis* from various commercial companies in Malaysia, France, Costa Rica and Columbia. The obtained dendrogram shows formation of two groups based on their genetic similarity, group A with ~ 76 % similarity and group B with ~ 66 % genetic similarity.

Agyei-Dwarko et al. (2012) examined the phenotypic variability, heritability and correlation among growth parameters in Dura x Pisifera (D x P) oil palm seedlings at

the Oil Palm Research Institute Kade, between July 2008 and February 2010. All the characters studied were significantly and positively correlated with each other except leaf area ratio (LAR) in which the correlations were negative with seven other traits.

Bunch components of oil palm germplasm collected in the northern regions of Ghana were evaluated by Sapey et al. (2012). Data collected were analyzed using standard procedures; elementary statistics, (mean values, standard error, range and coefficient of variation). Results obtained revealed some variation, for both qualitative and quantitative traits assessed on the accessions.

Okwuagwu et al. (2008) conducted a study to assess the extent of genetic variability, broad sense heritability, and correlation between yield and yield components in three Deli *dura* x *tenera* (D x T) breeding populations. The combined analysis of variance for number of bunches (BN), average bunch weight (ABW) and fruit bunch yield (FFB) revealed significant genotypic differences.

Molecular markers and protein markers have also been used to investigate the genetic diversity in the oil palm germplasm. These include restriction fragment length polymorphisms (RFLPs) of the ribosomal DNA (Maizura et al., 2006); random amplified polymorphic DNA (RAPD) (Morelzsohn et al., 2002); microsatellite DNA markers (Noorhariza et al., 2010); RFLP using cDNA probes and amplified fragment length polymorphism (AFLP) (Barcelos et al., 2002); simple sequence repeats (SSRs) (Zulkifli et al., 2012; Noorhariza et al., 2012) and Isoenzyme markers (Hayati et al., 2004).

2.3.2 Grouping Techniques in Measuring Genetic Diversity

Genetic relationship among and within breeding materials can be identified and classified, using multivariate groupings (Aremu, 2011). The use of established multivariate statistical algorithm is important in classifying breeding materials from germplasm accessions, lines, and other races into distinct and variable groups depending on genotype performance. The widely used techniques irrespective of the data source (morphological, biochemical, and molecular marker data) are Cluster analysis, principal component analysis (PCA), principal coordinate analysis (PCOA), canonical correlation and multidimensional scaling (MDS) (Maji & Shaibu, 2012; Odewale, 2012; Aremu, 2011; Zafar et al., 2008).

2.4 MULTIVARIATE STATISTICAL ANALYSIS

Multivariate analysis is a tool to find patterns and relationships between several variables simultaneously (CAMO, 2011). Multivariate analysis comprises a set of techniques dedicated to the analysis of data sets with more than one variable (Abdi, 2003). According to Rencher (2002), multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or objects in one or more samples. The measurements are referred to as variables and the individuals or objects as units or observations.

Multivariate analysis is the analysis of observations on several correlated random variables for a number of individual. It becomes necessary when dealing with several variables simultaneously. A series of univariate statistical analysis carried out consecutively for each of the variable is in general not adequate as it ignores the correlation among the variables. Multivariate analysis in contrary provides a more

accurate depiction of the behavior of data that are highly correlated, and can indicate the relative importance of the characteristics involved and yield more meaningful information (CAMO, 2011).

Multivariate analysis is the branch of statistics that is concerned with the relationship among sets of dependent variables and the individuals which bear them. The objectives of scientific investigation for which multivariate methods or techniques most usually lend themselves include the following (Johnson & Wichern, 2007):

a. Data reduction or structural simplification: the phenomenon being studied is represented as simple as possible without sacrificing valuable information. This is to make interpretation easier.

b. Sorting and grouping: grouping of "similar" objects or variables are created based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required.

c. Investigation of the dependence among variables: the nature of relationship among variables of interest. Are all the variables mutually independent or are one or more variables dependent on the others? If so how?

d. Prediction: relationships between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observations on the other variables.

e. Hypothesis construction and testing: specific statistical hypotheses formulated in terms of the parameters of the multivariate population are tested. This may be done to validate assumptions or reinforce prior convictions.

2.4.1 Principal Component Analysis (PCA)

PCA is probably the most popular multivariate statistical technique and it is widely used by almost all scientific disciplines. It is also likely to be the oldest multivariate technique, as its origin can be traced back to Pearson (1901), but its modern instantiation was formalized by Hotelling (1933) who also coined the term "principal component" (Abdi & Williams, 2010). Jolliffe (2002) also stated that the most generally accepted earliest descriptions of PCA were given by Pearson (1901) and Hotelling (1933).

Principal component analysis is defined as a method of data reduction to elucidate the relationships between two or more characters and to divide the total variance of the original characters into a limited numbers of uncorrelated new variables (Wiley, 1981). This will appropriate visualization of the differences among the individuals and identify possible groups. The reduction is accomplished by linear transformation of the original variables into a new set of uncorrelated variables known as principal components (PCs). According to Jolliffe (2002), the central idea of 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 components retain most of the variation present in all the original variables.

Abdi (2003), stated that the goal of PCA is to decompose a data table with correlated measurements into a new set of uncorrelated (i.e., orthogonal) variables. These variables are called, depending upon the context, principal components, factors, eigenvectors, singular vectors, or loadings. Each unit is also assigned a set of scores

which correspond to its projection on the components. The results of the analysis are often presented with graphs plotting the projections of the units onto the components, and the loadings of the variables. The importance of each component is expressed by the variance (i.e., eigenvalue) of its projections or by the proportion of the variance explained.

PCA is also a way to sum up the original data with a large amount of variables into simpler data with only a few variables (Wise et al., 2006). Similar to clustering, PCA is often used to find patterns in data - to see if some samples differ from others, observed as clusters of samples located at a distance from other samples in plots called scores plots. PCA finds the directions with the largest variance, meaning the largest spread of the points, and fits a line into that direction in such a way that the minimum squared distance from the line to each point is minimized (Beebe et al., 1998). The fitted lines are known as the principal component (PC).

The first step in PCA is to calculate eigenvalues, which define the amount of total variation that is displayed on the PC axes. The first PC summarizes most of the variability present in the original data relative to all remaining PCs. The second PC explains most of the variability not summed up by the first PC and uncorrelated with the first, and so on (Jolliffe, 1986). Because PCs are orthogonal and independent of each other, each PC unveils different properties of the original data and may be interpreted independently. In this way, the total variation in the original data set may be broken down into components that are cumulative. The proportion of variation accounted for by each PC is expressed as eigenvalue divided by the sum of the eigenvalues. The eigenvector defines the relation of the PC axes to the original data axes (Mohammadi & Prassana, 2003).

PCA aims to describe as much of the data as possible with as little principal components as possible, to extract most of the interesting data from the vast amount of initial data that is mostly irrelevant. In other words, a PCA model that consists of only a few principal components is created to approximate and simplify the data. The PCs are fitted so that the first principal component is aligned with the direction with the most variance (the direction where the spread of the points is largest).The second PC is fitted so, that it goes to that direction orthogonal to PC1 that has the next largest variance in the points.

The third PC and all the rest of the PCs are always orthogonal to all of the preceding PCs and similarly go to the direction of the largest remaining variance. The PCs are in order of significance: PC1 describes the most variance and is therefore most important. Because there is random noise distributed evenly to the whole row space and the amount of variance explained by the first PC is largest, the signal-to-noise ratio of PC1 is highest of all PCs, decreasing for each following PC. Eventually, the later PCs have signal-to-noise so low that they mostly describe just noise and are therefore useless (Abdi & Williams, 2010; Jolliffe, 2002; Beebe et al., 1998).

PCA can be performed on two types of data matrices: (1) A variance-covariance matrix and (2) A correlation matrix. With characters of different scales, a correlation matrix by standardizing the original data is preferred. If the characters are of the same scale, a variance-covariance matrix can be used (Wiley, 1981). PCA can also be used to ascertain the optimum number of clusters in a study. In this scenario, the objective is to maximize the variation explained by the first PC of each cluster. It begins with all individuals in a single cluster and splits them until the second eigenvalue of all clusters is less than a level specified by the user (Thompson et al., 1998).

The most interesting information gained from PCA is the scores and loadings. As already stated, the distance of samples (points) from a particular PC is called the scores on the respective PC. The scores therefore, hold information on the samples. Samples with similar scores are close to each other in the row space and are therefore somehow similar. The goal of PCA as know is to summarize and visualize multivariate data and this visualization is given in the form scores plots. The loadings on the other hand hold information on the variables (Wise et al., 2006) and are a measure of alignment of the particular PC with the variable axes. Variables with high loadings on a particular PC are well described by that PC. The loadings plots make it possible to find out why some samples are being similar and forming clusters in the scores plots, and this is a great advantage of PCA over clustering.

2.4.2 Clustering Analysis

Cluster analysis is the name for a group of multivariate techniques whose primary aim is to group objects based on the characteristics they possess. Cluster analysis classifies objects so that each object is very similar to others with respect to some predetermined selection criterion (Hair et al., 1998). Cluster analysis has been referred to as Q analysis, typology construction, classification analysis, and numerical taxonomy. This variety of names is due in part to the usage of clustering methods in diverse disciplines as psychology, biology, sociology, economics, engineering and business. Although, the names differ across discipline but all have a common dimension, thus, the primary value of cluster analysis lies in the classification of data, as suggested by the "natural' groupings of the data themselves. Cluster analysis also differs from PCA in that cluster analysis groups objects while PCA is primarily aimed with grouping variables (Hair et al., 1998).

Clustering is the most important unsupervised data mining method, that is, the clusters are formed without any prior knowledge of what the objects involved are (Tan et al., 2005). It involves finding structure in a collection of unlabeled data so that inter cluster similarity is minimized and intra cluster similarity is maximized between the samples. 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 shows similarity between them and dissimilarity to objects belonging to other belonging to other clusters (Singh et al., 2008). Therefore, clustering is the grouping of data items based on their similarity (Fan & Li, 1998).

Cluster analysis as described by Brereton (2003), is a well-established approach that was developed primarily by biologists to determine similarities between organisms. It is a multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to recognize these entities into relatively homogenous groups of objects called clusters (Ben-Hur & Guyon, 2003). According to Ferreira and Hitchcock (2009), cluster analysis involves sorting data (or items) into natural grouping based on similarity as grouping of data helps to reveal information about the data such as outliers, dimensionality, or previously unnoticed interesting relationships.

Clustering algorithms may be classified as may be classified as exclusive, overlapping and probabilistic clustering. In exclusive clustering, data are grouped in an exclusive way so that if a certain datum belongs to a definite cluster, it could not be included in another cluster. On the contrary, overlapping clustering uses fuzzy sets to cluster data so that each point may belong to two or more clusters with different degree of membership. In this case, data will be associated to an appropriate membership value.

Hierarchical clustering algorithm is based on the union between the two nearest clusters and the last type of clustering uses a completely probabilistic approach (Singh et al., 2008).

Most commonly clustering algorithms, as stated by Ferreira and Hitchcock (2009), rely on the pairwise dissimilarities between the objects to guide the clustering. There are many ways of defining (dis) similarity among objects and the choice of dissimilarity measure depends largely on the data being worked on (discrete, continuous, binary, etc.) and when the items are clustered, the dissimilarity between any two items is indicated by some sort of distance (Siracli et al., 2013). Common measurements of distance between two multivariate data vectors include ordinary Euclidean distance, Manhattan or Mahalanobis distance (Svetlana et al., 2012).

An important class of clustering methods is hierarchical cluster analyses and there are two main types of hierarchical clustering methods; agglomerative and divisive or partitional method. An agglomerative hierarchical method begins with each object as its own cluster, and then successively merges the most similar clusters together until the entire set of data becomes one group. On the contrary, the divisive method starts with a single large cluster divides it into smaller clusters known as partitions (Ferreira & Hitchcok, 2009). Hierarchical cluster analysis requires the selection of a distance measure and a linkage method for linking two clusters. The most common distance measure used is the Euclidean distance which is the actual geometric distance between two points (Mohammadi & Prassana, 2003).

Various linkages can be used to determine which groups should be merged in agglomerative hierarchical clustering. Single linkage merges groups based on the minimum distance between two objects in two groups (Sneath, 1957). Complete

linkage merges groups based on the maximum distance between two objects in two groups (McQuilty, 1960; Sokal & Sneath, 1963). Average linkage merges groups based on the average distance of all the objects in one group to all the objects in the other groups (Sokal & Michener, 1958). Ward's method (1963) is another important hierarchical clustering method and it is similar to the linkage methods in that it begins with N Clusters, each containing one object and differs from it as it does not use cluster distance to group objects. Instead, the total within cluster sum of squares (SSE) is computed to determine the next two groups merged at each step of algorithm.

There are other clustering methods besides hierarchical methods, some of which include partitioning methods such as the well-known K-means (MacQueen, 1967) and K-medoids (Kaufman & Rousseuw, 1987). However, based on the survey of cluster analysis in scientific literature, Kettering (2006) indicated that hierarchical clustering was by far the most widely used form of clustering in practice.

2.5 REVIEW OF RESEARCH WORKS ON PRINCIPAL COMPONENT ANALYSIS AND CLUSTERING

Multivariate data analysis has been used extensively by several authors to decompose mixed data structure into its components. To reduce this relatively large number of variables to a simpler number of orthogonal factors, the concentrated data have been treated by multivariate statistical methods, i.e. principal component analysis and cluster analysis.

The genetic variability among the fruits of 22 date palm cultivars using six nutritional characters was studied by Hamza et al. (2014) to enable classify the available germplasm into distinct groups on the basis of their genetic diversity using their

nutritional characteristics from proximate composition. Multivariate analysis techniques were used to classify the cultivars and the results of PCA indicated that the contribution of the first three factors with eigenvalue greater than one accounted for 73.03 % of the total variation. Cluster analysis also placed the 22 date palm cultivars into three main clusters.

Jonah et al. (2014) applied multivariate analysis on the growth and yield of twelve cultivars of Bambara groundnut collected in Northern-Eastern Nigeria. Principal components axis 1 to 11 accounted for 99 % of the total variability observed among the cultivars of the Bambara groundnut evaluated.

Multivariate analysis was also applied on some metric traits of bread wheat (Ahmed et al., 2014). The nineteen genotypes of bread wheat evaluated where grouped into three clusters on the basis of average linkage while analysis for principal components observed that PCs with Eigenvalue greater than 1 accounted for 73.94 %. it was also found out that genotypes grouped into the three clusters, showed maximum inter cluster diversity and that the presence of significant genetic variability among tested genotypes shows excellent opportunity for improvement through wide hybridization by crossing genotypes of different clusters.

Okii et al. (2014) studied the morphological diversity of tropical common bean germplasm and PCA clustered the germplasm into three major groups and the first two PCs with eigenvalue greater than two accounted for 45.31 % of the total variability in the bean germplasm.

The genetic diversity for cotton leaf curl virus, fiber quality and some morphological traits in cotton was also assessed by Saeed et al. (2014). The data was subjected to

multivariate analysis and PCA showed first five PCs having eigenvalue greater than 1 explaining 71.3 % of the total variation while cluster analysis classified 100 accessions of cotton into four divergent groups.

Ren et al. (2014) also applied multivariate analysis to assess the genetic diversity and population structure of 196 peanut cultivars based on one hundred and forty-six highly polymorphic simple sequence repeat (SSR). The results from cluster analysis, principal component ordinate analysis (PCoA) and PCA based on the marker genotypes revealed five distinct clusters for the entire population that were related to their germplasm regions.

Multivariate analysis of phenotypic diversity of landraces of rice of West Bengal was carried out by Chakravorty et al. (2013). Characters were found to be interrelated among themselves and PCA was undertaken to have an idea about their independent impact. The first six components in the PCA analysis with eigenvalues greater than 1 contributed 75.9 % of the variability among genotypes evaluated for different agro-morphological traits. Clustering based on the first two principal components indicated 10 clusters.

Tahir et al. (2013) assessed the genetic divergence in sugarcane genotypes based on 14 parameters. Principal component analysis showed that there were two principal components accounting for 88 % of the total variation in the tested breeding material. Clustering analysis using Ward's method revealed that there were 3 clusters at a linkage of 4.5 and no corresponding of clustering with geographic location of the genotypes.

Tewodros (2013) assessed the nature and extent of diversity on one hundred accessions of taro (*Colocasia esculenta*) based on data collected on 17 qualitative and 13 quantitative traits. Both sets of data were subjected to multivariate analysis using PCA and clustering. First seven PCs with eigenvalue greater than one accounted for 64.9 % of variation while cluster analysis based on qualitative characters indicated the formation of seven clusters.

Odewale et al. (2012) employed multivariate analysis as a tool in the assessment of the physical properties of the date palm fruits in Nigeria. PCA divided the 17 out of 20 variables determined into four components which explained 97.03 % of the total variation. The dendrogram generated by the unweighted pair group method using arithmetic averages (UPGMA) cluster analysis grouped the ten new accessions into four distinct clusters indicating genetic diversity.

The morphometric characterization of coconut germplasm was also evaluated by Sarkar et al. (2012). To examine the relationship among the quantitative variables, PCA and clustering were applied. First three components contributed 66.73 % of the observed variation while the twenty entries of coconut were grouped into seven clusters.

Multivariate analysis was also used by Odewale et al. (2012) to study the pattern of genetic diversity and relationship among six exotic and four indigenous coconut cultivars and a hybrid. The result of the PCA indicated that the contribution of the first two factors with eigenvalue greater than unity accounted for 69.5 % of the total variation. The cultivars were also classified into two distinct groups.

PCA and clustering analysis were performed by Ahmad et al. (2012) to investigate the level of diversity among cotton germplasm based on phenotypic data of 16 characters, before and after year 1975 in Pakistan. As revealed through PCA, pre-1975 group showed 41 % genetic diversity, while post-1975 accounted for 37 % genetic diversity. 1st principal component being the most diverse component accounted for 1 - 53 % of degree of association in pre-1975 as compared to 1 - 46 % in post-1975 cultivars. Cluster analysis grouped all 39 accessions into five clusters on similarity basis.

Beiragi et al. (2012) applied multivariate analysis method to determine the importance of traits associated with maize. According to PCA, four PCs had eigenvalue greater than 1 which accounted for 67 % of the total variance in the data on early planting date and 73 % of the total variance in the data on delayed planting date. Based on cluster analysis, the 18 corn hybrids used were separated into four and five major groups with each having two or more subgroups on early and delayed sowing dates respectively.

To assess the genetic diversity for various rice traits and their association with yield, Ashfaq et al. (2012) evaluated forty rice genotypes on the basis of various morphological traits in a field experiment. Traits were studied through PCA and the combined variation among these traits was 67.7 % with four principal components exhibiting eigenvalue of more than one. The genotypes were also classified in to different clusters on the basis of the different traits with similar genotypes classified in to same cluster.

Sohrabi et al. (2012) stated that genetic diversity is prerequisite for any crop improvement as it helps in the development of superior recombinants. To this effect fifty Malaysian upland rice accessions were evaluated for 12 growth traits, yield and

yield components. According to UPGMA cluster analysis, all accessions were clustered into six groups while the first PCA indicated 76.7 % of the total variation.

Maji and Shaibu (2012) applied PCA for rice germplasm characterization and evaluation. Cluster analysis using Ward's method classified the 123 populations into seven distinct groupings. Principal components analysis resulted in the first two components with eigenvalue greater than 1 accounting for 78 % of the total variation. The results of the PCA were found to be closely in line with those of clustering.

Yusuf et al. (2011) used multivariate analysis to investigate the extent of genetic diversity in rapeseed (*Brassica campestris* L.) germplasm based on agro-morphological and seed quality traits. Cluster analysis based on fifteen agro-morphological and six seed quality traits, divided 114 accessions into six and five clusters during 2005 and 2006, respectively. Findings on PCA indicated that the first seven and five PCs with eigenvalues greater than 1 contributed 74.09 % and 66.08 % of the variability amongst accessions during 2005 and 2006, respectively. Nine important characters were also noted to have contributed positively to the first two PCs during both years.

According to Ahmadizadeh and Felenji (2011), PCA and clustering analysis in potato cultivars facilitate identifying desirable traits and their relationship with yield and reliable classification of genotypes. This was due to their research to investigate genetic diversity in potato using agro-morphological and yield components. PCA showed that, three components explained 80.1 % of the total variation among traits. In grouping based all the traits, 22 investigated cultivars were placed on two clusters using Ward's method.

Khavari et al. (2011) used multivariate analysis to characterize 34 maize hybrids. According to PCA, seven principal components had eigenvalues greater than 1 and accounted for 85.12 % of total variance in the data. Based on cluster analysis, the 34 new corn hybrids were separated into seven major groups with each having two or more subgroups.

Denton and Nwangburuka (2011) evaluated the genetic variability in eighteen accessions of *Solanum anguivi* collected from various diverse sources using principal component analysis and single linkage cluster analysis (SLCA) technique. The accessions were classified into five groups using PCA while the first three principal axes accounted for over 72 % of the total variation.

Iannucci et al. (2011) evaluated oat germplasm to determine the genetic diversity among 109 genotypes. All the accessions were characterized according to 13 bioagronomic traits. PCA showed that the first six axes accounted for 81 % of the variability while clustering entries identified nine groups based on their morphological and agronomic properties.

According to Evgenidis et al (2011), determination of germplasm diversity and genetic relationships among breeding materials is an invaluable aid in crop improvement strategies. In view of this, principal component and cluster analysis were employed as a tool in the assessment of tomato hybrids and cultivars based on morphophysiological data, yield and quality, stability of performance, heterosis and combining abilities. First two principal components accounted for 78.77 % of total variance two groups were formed from clustering analysis.

The genetic divergence of 40 lime accessions was carried out by Rahman and Mansur (2011) using D^2 and PCA. The genotypes under study fell into six clusters. First five axes obtained from the PCA accounted for 85.24 % of the total variation among 9 characters describing 40 lime accessions, while the first two accounted for 53.09 %.

Agro-morphological variation in the sesame germplasm was estimated by Seymus and Uzun (2010) using 21 morphologic and agronomic descriptors to characterize and identify genetic diversity. Multivariate analyses were performed in order to establish similarity and dissimilarity pattern. The first seven PC axes explained 69.9 % of the total variation while cluster analysis identified eight main clusters based on agro-morphological characters.

Makinde and Ariyo (2010) used multivariate analysis to evaluate the genetic divergence in twenty two genotypes of groundnut. The first three axes of the PCA captured 55 % of the total variance. The dendrogram obtained from the single linkage cluster analysis grouped accessions into eight clusters. All genotypes were distinct at 100 % level of similarity while at 25 %, they could not be discriminated.

The genetic diversity of sixty four new advanced breeding lines of groundnut was estimated by Kumar et al. (2010). The data recorded on fourteen characters were subjected to multivariate analysis to study the variability within the genotypes. The first three axes of PCA captured 59.52 % of the total variability. Ward's clustering method grouped the genotypes in to three different distinct clusters.

Zafar et al. (2008) evaluated soybean germplasm for some important morphological traits using multivariate analysis. Cluster analysis divided the 139 genotypes into five clusters based on the traits scored for. PCA showed that first three components

accounted for 69.77 % of the total variance. They concluded that principal component and clustering obtained from their study can make better choice for soybean breeders for selecting genotypes among large number of accessions.

The morphological and agronomic diversity of *Brassica napus* crops was carried out by Soengas et al. (2008). Thirty-five B. napus populations from different geographic origins and uses were evaluated and data was recorded on 17 morphological and agronomic traits. PCA and cluster analysis were performed to classify the population. Eight principal components accounted for 94 % of the total variability while populations were grouped into five clusters.

Idehen et al. (2007) employed PCA and SLCA to analyze the variation pattern in 20 accessions of "egusi" melon based on fourteen quantitative characters. The first three principal components accounted for 76.33 % and 78.70 % of the total variation in the early and late seasons respectively. SLCA summarized the relationship among the accessions at various levels of similarity into a dendrogram while the accessions were sorted into six distinct groups.

Ackerfield and Wen (2002) used SLCA for clustering, utilizing the unweighted pair-group arithmetic average Clustering (UPGMA) method and PCA to detect morphological variation among 105 specimens of *hedera* with each representing operational taxonomic unit (OUT). The taxa were separated into two major groups with no overlap and first three principal components accounted for 71.1 % of total variation.

Zizumbo-Villareal and Pinero (1998) statistically and numerically evaluated the pattern of morphological variation and diversity in forty-one populations of coconut,

using seventeen morphological traits. PCA and cluster analysis indicated four main groups of coconut populations. The first three principal components accounted for 89 % of the total variance.