CHAPTER IV

PRINCIPAL COMPONENT ANALYSIS OF THE MPOB-NIGERIAN OIL PALM GERMPLASM

4.1 INTRODUCTION

Present day plant breeding entails data collection on a large number of plant characteristics or traits. Some of these traits may be correlated to each other in such a way that the increase in one, constantly leads to the increase or decrease in another. When such traits are found out, as it is done in oil palm breeding, the breeder becomes confronted with the challenge of keying out a few that the limited resources can be concentrated upon and be certain of not losing out information on the others (Oboh & Fakorede, 1997).

Oil palm, being a cross pollinated crop has individual palm having varying in performance from each other and hence, assessment of individual palm is of utmost important. Evaluation of palms is therefore needed for various purposes, like for selecting of superior palms (Mandal, 2008). It is has been known, that accurate evaluation in oil palm is usually hindered by cumbersome data accumulated over the years (Oboh & Fakorede, 1990) and as such, some of these variables may not be adequate for germplasm evaluation, characterization and management. Both characterization and evaluation data provide an effective source of information for genetic diversity and help in understanding patterns of variation in crop species (Rao & Hodgkin, 2002).

Statisticians have helped developed statistical methods generally known as multivariate statistical techniques to effectively handle the situation in which a large amount of accumulated data makes it difficult for breeders to take decision. These grounded multivariate statistical algorithms have proved important in classifying germplasm, ordering variability for large number of accessions or analyzing genetic relationships among traits in any breeding materials (Zafar et al., 2008).

One of these techniques is principal component analysis (PCA). This technique is a descriptive procedure for analyzing relationships that may exist in a set of multivariate dataset. It is designed to reduce the number of variables that need to be considered to a small number of indices (called the principal components) that are linear combination of the original variables (Ekezie, 2013). Ezekie (2013) identified three major principal components account for 99.48 % of total variation in the oil palm progeny studied using PCA.

In view of this, the objectives of this chapter were to use PCA to describe the pattern of variation, as well as identify the variables contributing to the total variation in the MPOB-Nigerian oil palm germplasm.

4.2 MATERIALS AND METHODS

4.2.1 Plant Breeding Material

The Malaysian Agricultural Research and Development (MARDI) and the Nigerian Institute for Oil Palm Research (NIFOR) in 1973 jointly made collections of oil palm genetic materials in Nigeria to broaden the genetic base of the oil palm. The Nigerian oil palm germplasm was collected at 45 sites (45 populations) (Table 1, Figure 1), with an average of 20 palms per site and 200 seeds per palm. A total of 919 (595 *duras* and 324 *teneras*) open-pollinated bunches were harvested during this collection (Rajanaidu & Rao, 1987). The prospected oil palm breeding materials were planted using eight cubic lattice design on September, 1975 and identified as Trial 0.149 (Rajanaidu & Rao, 1987) at the MPOB Station at Kluang, Malaysia. Kluang is characterized by Rengam soil series and a rainfall of about 2500 mm/yr. However, only the *dura* Nigerian oil palm breeding materials were selected for this research.

4.2.2 Data Collection

Data collections were carried out for bunch yield, bunch quality components, morphological traits, physiological traits and fatty acid composition. The procedures for these traits are presented in **APPENDIX A – D.**

4.2.3 Standardization

Prior to performing PCA on the dataset, data for traits were standardized to a mean of zero and a standard deviation of one to annihilate the biasness due to dissimilar magnitude of the units of measurement (Jolliffe, 2002). The average of the quantitative was standardized using the formula (Microsoft Office Excel, 2010);

$$z = \frac{x - \mu}{\sigma} \qquad (2)$$

Where;

x = value to standardize

μ = arithmetic mean

σ = standard deviation of the distribution

4.2.4 Principal Component Analysis (PCA)

PCA was performed with the aid of "THE UNSCRAMBLER®X" software (CAMO software version 10.1). PCA was performed separately based on all the 43 variables; yield, yield components and bunch components; vegetative and physiological traits; as well as fatty acid traits of the MPOB-Nigerian germplasm materials.

4.2.5 Theory of PCA

PCA is a powerful technique for pattern recognition that attempts to explain the variance of a large set of inter-correlated variables and transforming into a smaller set of independent (uncorrelated) variables (principal components) (Svetlana et al., 2012). PCA analyzes a data table representing observations described by several dependent variables which are, in general, inter-correlated. PCA also represents the pattern of similarity of the observations and the variables by displaying them as points in maps (Abdi & Williams, 2010).

4.2.6 Prerequisite Notions and Notations

Matrices are denoted in upper case bold, vectors are denoted in lower case bold, and elements are denoted in lower case italic. Matrices, vectors and elements from the same matrix all use the same letter (e.g., **A, a**, *a*). The transpose operation is denoted by the superscript ^{T}. The identity matrix is denoted **I**. The data table to be analyzed by PCA comprises *I* observations by *J* variables and it is represented by the $I \times J$ matrix **X**, whose generic element is $x_{i,j}$. The matrix **X** has rank *L* where $L \leq \min \{I, J\}$. In general the data table will be preprocessed before the analysis. The columns of **X** will be centered so that the mean of each column is equal to 0 (i.e., $\mathbf{X}^{T}1=0$, where 0 is a *J* by 1 vector of zeros and 1 is an *I* by vector of ones).

The inertia of a column is defined as the sum of the squared elements of this column as is computed as:

$$\gamma_i^2 = \sum_i^J x_{i,j}^2. \tag{3}$$

The sum of all the γ_i^2 is denoted Γ and it is called the *inertia* of the data table or *total inertia*. The center of gravity of the rows also called centroid, denoted *g*, is the vector of the means of each column of **X**. When **X** is centered, its center of gravity is equal to the $I \times J$ row of vector 0^T. The distance of the *i*-th observation to g is equal to:

$$d_{i,g}^2 = \sum_i^J (x_{i,j} - g_i). \tag{4}$$

When the data are centered, Equation 4 is reduced to:

$$d_{i,g}^2 = \sum_i^J x_{i,j.}^2 \tag{5}$$

4.2.7 Eigenvalues

After PCA, the size of each component can be measured and often called an eigenvalue. The earlier (and more significant) the components are, the larger their size and the higher the degree of correlation among the variables in the data, the fewer components to capture common information. Eigenvalue of a PC is defined as the sum of squares of the scores and given as;

$$g_a = \sum_{i=1}^{I} t_{ia}^2, \tag{6}$$

where g_a is the a[th] eigenvalue (Brereton, 2003). Eigenvalues are presented in percentages. Successive eigenvalues correspond to smaller percentage. The cumulative percentage eigenvalue is often used to determine (approximately) what proportion of the data has been modeled using PCA and is given by $\sum_{a}^{A} = 1g_a$. The closer the value to 100 %, the more faithful is the PC model (Breteton, 2003; Vyas & Kymaranayake, 2006).

4.2.8 Scores and Loadings

PCA results in a transformation of the original data matrix into scores and loadings. The scores have as many rows as the original data matrix and the loadings have as many columns as the original data matrix. Each scores matrix consists of a series of column vectors, and each loadings matrix a series of row vectors. The scores matrices and loadings matrices are composed of several such vectors, one for each principal component.

Correlation of principal components and original variables is given by loadings and loadings reflect the relative importance of the variable components and not the importance of the variable itself. Variable loadings can be classified as "strong", "moderate", and "weak", corresponding to absolute loading values of greater than 0.75, 0.75-0.50, and 0.50-0.30, respectively (Jolliffe, 2002). Each successive component is characterized by a pair of eigenvectors. After PCA, the original variables are reduced to a number of significant principal components (Brereton, 2003).

4.2.9 Optimum Number of PCs Extracted

The eigenvalue-one criterion also known as the Kaiser criterion (Kaiser, 1960) was used to retain the optimum number of PCs needed and it is one of the most commonly used criteria for solving the number-of-components problem. With this approach, component with an eigenvalue greater than 1.00 were retained and interpreted while component with less than 1.00 eigenvalue were discarded.

4.3 RESULTS

4.3.1 PCA of All Variables

As presented in Table 9, the result of PCA on all the traits studied indicated eight components underlying the characters of the MPOB-Nigerian oil palm germplasm. Only PCs with eigenvalue greater than one were selected for this analysis. The cumulative variance for each component in which eight principal components obtained, accounting for 90.53 % of the total variance is illustrated graphically in Figure 2a.

PC1 explained 42.97 % of the total variance and had strong positive loadings on fresh fruit bunch (FFB), average bunch weight (ABW), mean fruit weight (MFW), mesocarp to fruit (MF), oil yield (OY), total economic product (TEP), petiole cross section (PCS), rachis length (RL), leaflet length (LL), leaflet width (LW), no of leaflets (LN), leaf area (LA), leaf area index (LAI), bunch dry mass (BDM), total dry mass (TDM), radiation conversion efficiency (*e*) and fractional interception of radiation (*f*); moderate positive loadings on mean nut weight (MNW), oil to dry mass (ODM), oil to bunch (OB), kernel yield (KY) and vegetative dry mass (VDM); strong negative loadings on kernel to fruit (KF) and kernel to bunch (K/B); moderate negative loadings on shell to fruit (SF) and frond production (FP). First principal component represent mostly yield, bunch, vegetative and physiological parameters.

PC2 accounted for 15.12 % of the total variance and had moderate positive loadings on Shell to fruit (SF), trunk diameter (DIAM), vegetative dry mass (VDM), iodine value (IV) and C18:1. PC2 had moderate negative loadings on no of bunches (BNO), parthenocarpic to bunch (PB), bunch index (BI) and net assimilation ratio (NAR).

PC3 explained 10.12 % of the total variance and had moderate positive loadings on BI and stearic acid (C18:0); moderate negative loadings on frond production, trunk height (HT) and myristic acid (C14:0).

PC4 accounted for 8.57 % of the total variance and had moderate positive loadings on iodine value (IV) and oleic acid (C18:1); moderate negative loadings on palmitic acid (C16:0).

PC5 explained 4.49 % of the total variance with moderate negative loadings on fruit to bunch (F/B).

PC6 had 3.91 % of the total variance with moderate negative loadings on NAR while PC7 and PC8 on the other hand accounted for 2.69 % and 2.56 % of the total variance respectively, with moderate positive loadings of linoleic acid (C18:2) on PC7 and negative loadings on PC8.

Figure 2b shows the correlation loadings plot for each variable for the displayed components. The plot contains two ellipses, the inner and outer ellipses. The variables in the inner circle indicate 50 % of explained variance while those in the outer ellipse indicate 100 % variance and have high contribution to PC1.

Figure 2c shows the projection of the variables on principal component 1 and 2. From the graph, parameters in O and P had high contribution on PC1 and both groups of variables are anti-correlated. Parameters in Q and R had high contribution to PC2. Some of the variables in P are strongly correlated to each other as can be seen in FFB and BDM, likewise TEP and OY.

Table 9: Variables correlation loading matrix and total variance explained for all traits

Parameter	Principal component							
	1	2	3	4	5	6	7	8
FFB	.908*	-.299	.194	-.038	.012	-.178	.042	-.018
BNO	-.412	-.573**	.214	-.106	.481	-.316	.117	.119
ABW	.885*	-177	-.042	.029	-.319	.101	-.095	-.089
MFW	.834*	.074	-.348	.064	-.053	-.209	.242	.110
MNW	.664**	.336	-.401	-.097	-.038	-.301	.274	.108
PB	.015	-.708**	.215	.196	.078	.202	-.381	.004
MF	.792*	-.443	.027	-.324	-.069	.136	.046	.041
KF	-.814*	.010	.103	-.365	-.250	-.007	-.260	-.150
SF	-.610**	.617**	-.101	-.231	.251	-.186	.096	.035
ODM	.603**	-.344	-.462	.324	.172	.180	.044	-.032
OWM	.264	-.430	-.484	.298	.328	.278	-.276	-.209
FB	.220	.024	-.459	-.029	-.519**	-.151	.103	-.132
OB	.749**	-.470	-.182	.345	-.064	.151	-.012	-.037
KB	-.785*	.073	.024	-.397	-.331	-.044	-.216	-.173
OY	.880*	-.434	.041	.144	-.022	-.070	.017	-.033
KY	.505**	-.248	.304	-.494	-.357	-.286	-.159	-.220
TEP	.878*	-.433	.085	.053	-.076	-.108	-.010	-.065
FP	-.590**	.134	-.568**	.387	.260	.010	.041	-.010
PCS	.899*	.349	-.057	-.066	-.096	-.035	-.110	-.006
RL	.927*	.168	.187	-.195	-.018	.062	.033	.026
LL	.942*	.198	-.050	-.039	.020	.075	.066	-.014
LW	.775*	.439	.117	-.232	.079	.162	-.059	.056
LN	.853*	.158	.115	-.229	.049	.079	-.234	-.005
HT	-.002	.373	-.740**	.330	-.068	.097	-.100	-.222
LA	.919*	.314	.025	-.162	.057	.126	-.052	.014
LAI	.919*	.314	.025	-.162	.057	.126	-.052	.014
DIAM	.419	.633**	-.029	-.204	.301	-.139	-.134	.181
LAR	.360	-.226	.445	-.318	.351	.435	.212	.103
BDM	.909*	-.306	.207	-.081	.026	-.157	.032	-.009
VDM	.689**	.532**	-.420	.143	.035	-.029	-.118	-.034
TDM	.980*	.025	-.045	.044	.035	-.127	-.032	-.022
BI	.424	-.676**	.532**	-.146	-.021	-.135	.133	.010
E	.820*	-.251	-.152	.238	.025	-.362	-.033	-.052
F	.945*	.258	.067	-.137	.037	.096	-.021	.009
NAR	-.332	-.620**	-.175	.417	-.039	-.500**	.036	-.062
IV	-.091	.607**	.377	.514**	-.033	.040	.157	-.179
C120	.112	-.339	-.390	-.227	-.479	.271	.148	.426
C140	-.139	-.415	-.587**	-.476	-.181	.166	.200	.228
C160	.090	-.443	.466	-.585**	.333	-.121	-.173	-.141
C161	.202	.296	-.137	-.498	.261	-.228	.098	-.307
C180	.025	.190	.649**	.478	-.199	.104	.079	-.131
C181	-.011	.583**	.407	.541**	-.046	-.197	-.021	.172
C182	-.091	-.111	.034	-.194	.028	.310	.538**	-.592**
Eigenvalue	18.48	6.50	4.39	3.68	1.93	1.68	1.16	1.10
Variance (%)	42.97	15.12	10.12	8.57	4.49	3.91	2.69	2.56
Cumulative (%)	42.97	52.08	68.30	76.87	81.36	85.28	87.96	90.52

* Variables that have loadings greater than0.75, **Variables that have loadings 0.75 – 0.50

Figure 2: Principal component analysis plots (a) Cumulative variance; (b) Correlation loadings plot of all variables along PC1 and PC2; (c) loading plot of all variables on PC1 and PC2; (d) PCA scores plot for the MPOB-Nigerian oil palm populations; (e) Bi-plot of 44 MPOB-Nigerian oil palm populations and 43 variables

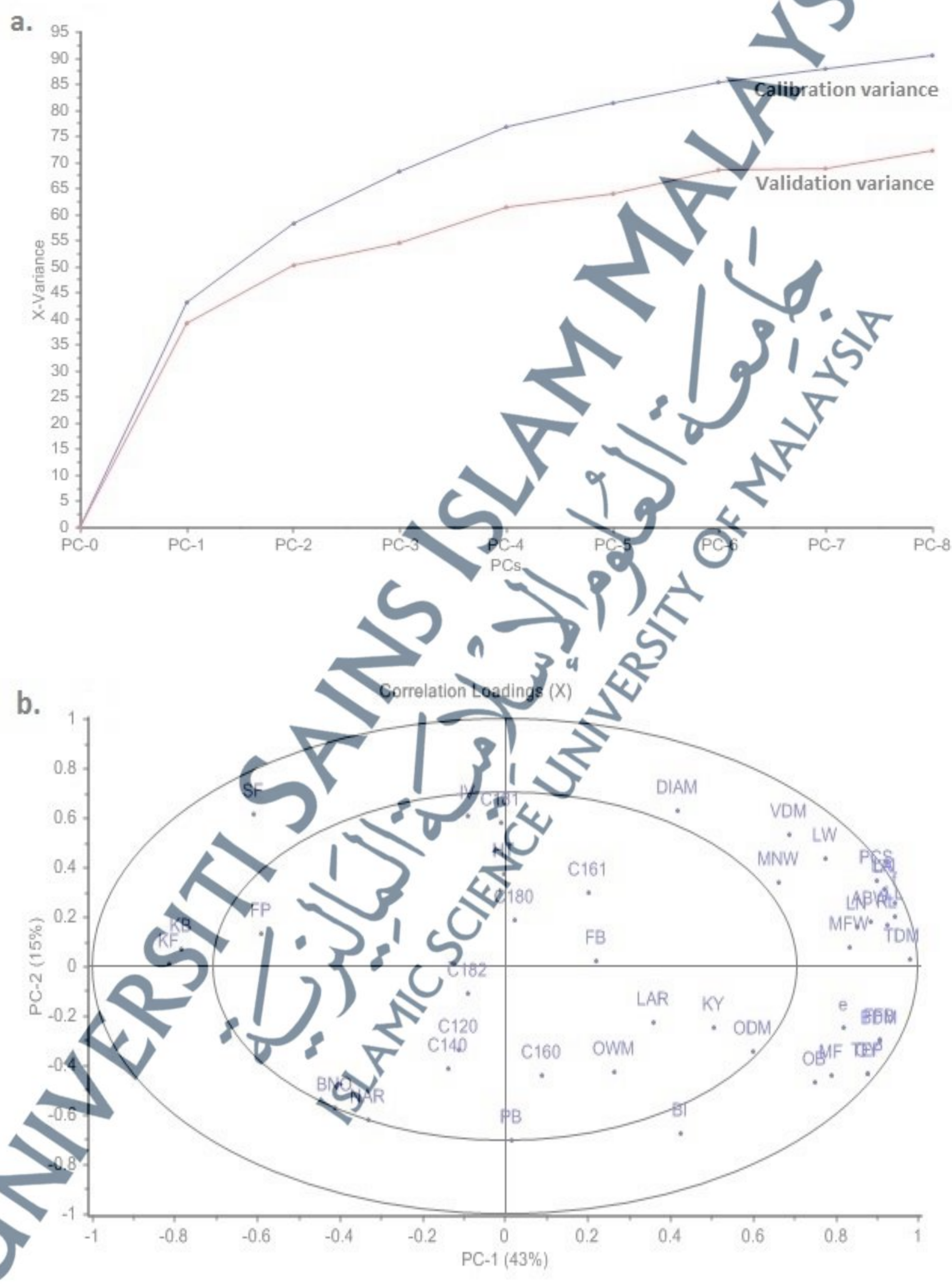
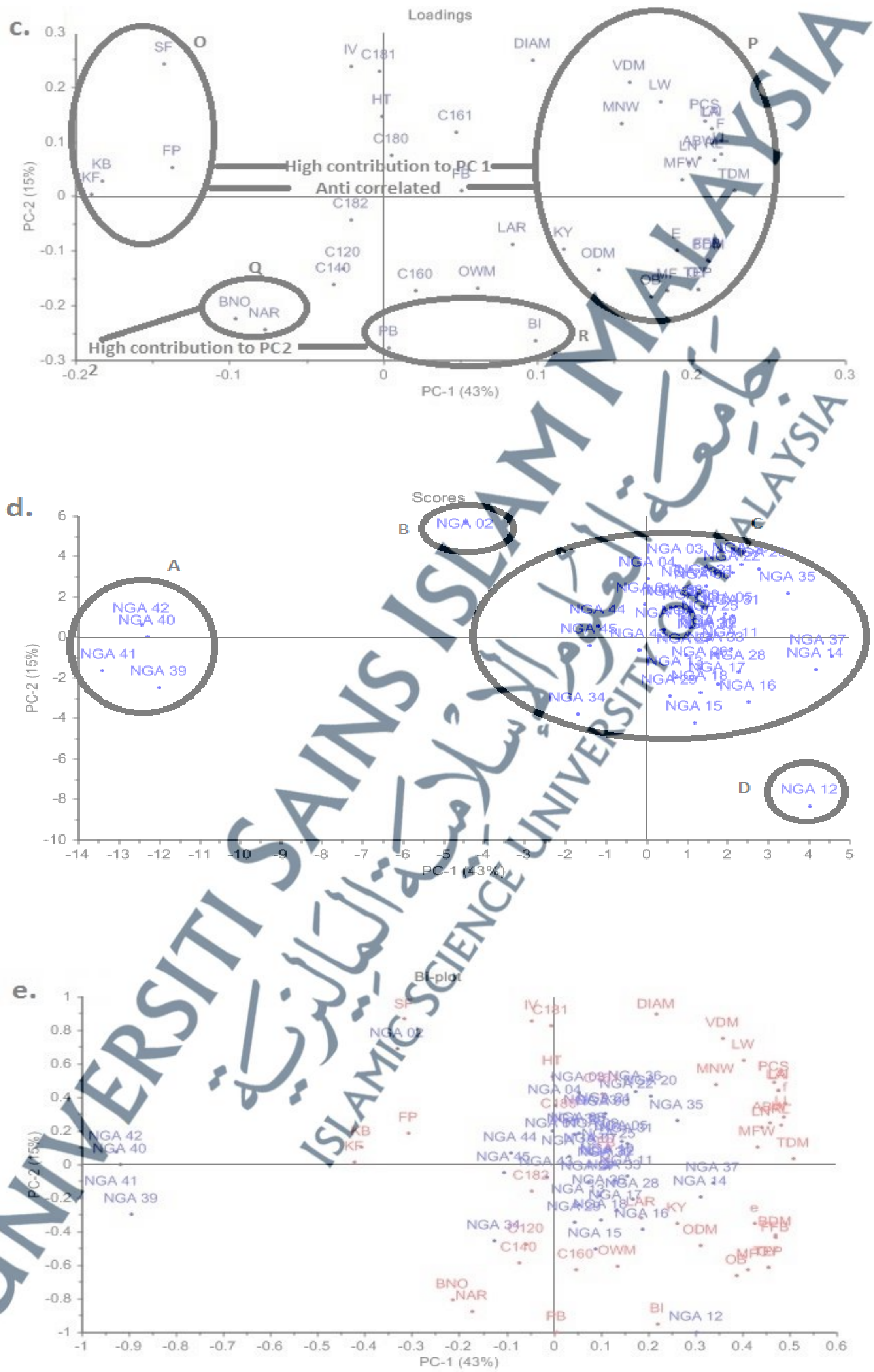


Figure 2 contd.



The projection of the MPOB-Nigerian oil palm germplasm on PC1 and PC2 is shown graphically in Figure 2d. Interestingly, from the scores plot it can be seen that the populations in A show distinctness from others. In B and D, only one population can be seen. C has the highest number of populations and the oil palm populations are very close to one another. Group B and D are also anti-correlated as they appear at opposite end to each other.

The projection of scores plot and loadings plot is displayed simultaneously in Figure 2e. It can be seen from the graph that NGA 12 can best be described by high bunch index (BI), NGA 02 by high shell-to-fruit content (S/F). Also, these two populations are different as both are anti-correlated. NGA 34 is also best described by high C12:0 and C14:0. Furthermore, C18:2 can be seen close to the centre of the plot and as such it has no influence in the describing the populations.

4.3.2 PCA of Yield and Bunch Quality Components

Figure 3a shows the cumulative variance for each component and four components with eigenvalue of greater than1 were obtained, summing 87.33 % of the total variance in the yield and bunch quality components traits of the MPOB-Nigerian oil palm germplasm. As presented in Table 10, PC1 accounted for 54.01 % of the total variance and had strong positive contributions on FFB, ABW, MFW, MF, ODM, OB, OY and TEP; moderate positive loadings on MNW and strong negative loadings on KF, SF and KB. PC2 explained 16.17 % of the total variance with moderate positive loadings on MNW, strong negative loadings on PB and moderate negative loadings on BNO. PC3 explained 10.5 % of the total variance with strong positive loadings on KY while PC4 accounted for 7.15 % of the total variance with moderate positive loadings on FB and moderate negative loadings on BNO.

Figure 3b shows the correlation loadings plot for each variable for the displayed components. The variables in the inner circle indicate 50 % of the explained variance while those in the outer circle indicate 100 % variance and they have high contribution to PC1. The projection of yield and bunch quality traits on PC1 and PC2 is shown graphically in Figure 3c. Variables in R and S contributed to the PC1 and are anti-correlated to each other. As seen in the Figure 3c, TEP and OY in S are strongly correlated as well as OY and MF. These variables are overlapped on one another.

Figure 3d shows the distribution of the oil palm germplasm on PC1 and PC2 based on their yield and bunch quality traits. Populations in V and X are quite separated from the other populations and are at the opposite end to each other, i.e. they are anti-correlated. Populations in U are also separated from others and the members are close to each other. Group W has the highest number of populations; some members of the group are strongly similar to one another while a few members are separated. The plot of loadings and scores simultaneously is displayed in Figure 3e.

4.3.3 PCA of Vegetative and Physiological Traits

As displayed in Figure 4a, only three PCs with eigenvalue greater than one explained 90.79 % of the total variance for the vegetative and physiological traits of the MPOB-Nigerian oil palm germplasm. As indicated in Table 11, PC1 contributed 61.94 % of the total variance and had moderate positive loadings on FP and NAR; strong negative loadings on PCS, RL, LL, LW, LN, LA, LAI, BDM, TDM and *e*; moderate negative loadings on DIAM, VDM and *f*. PC2 explained 17.89 % of the total variance with strong positive loadings on BI; moderate positive loadings on LAR; strong negative loadings on HT and moderate negative loadings on FP and VDM.

Table 10: Variables correlation loading matrix and total variance explained for yield and bunch components

Parameters	Principal components			
	1	2	3	4
FFB	0.891*	0.030	0.330	-0.286
BNO	-0.287	0.638**	0.047	-0.501**
ABW	0.791*	-0.403	0.190	0.175
MFW	0.814*	0.463	-0.104	-0.153
MNW	0.550**	0.699**	-0.101	-0.229
PB	0.213	-0.835*	0.119	0.145
MF	0.939*	-0.227	0.000	0.068
KF	-0.846*	-0.15	0.350	0.259
SF	-0.798*	0.411	-0.216	-0.256
ODM	0.779*	-0.146	-0.448	0.128
OWM	0.457	-0.459	-0.491	0.247
FB	0.288	-0.459	0.077	0.605**
OB	0.933*	-0.223	-0.116	0.220
KB	-0.826*	-0.015	0.361	0.331
OY	0.967*	-0.154	0.160	-0.081
KY	0.404	0.044	0.858*	-0.050
TEP	0.941*	-0.133	0.281	-0.081
Eigenvalue	9.18	2.75	1.78	1.22
Variance (%)	54.01	16.17	10.5	7.15
Cumulative (%)	54.01	70.18	80.68	87.83

* Variables that have loadings greater than0.75; ** Variables that have loadings 0.75 – 0.50

Figure 3: Principal component analysis plots (a) Cumulative variance (b) Correlation loadings plot of yield and bunch quality traits along PC1 and PC2; (c) Loading plot of yield and bunch quality traits on PC1 and PC2; (d) PCA scores plot of the MPOB-Nigerian oil palm populations; (e) Bi-plot of 44 MPOB-Nigerian oil palm populations and 17 yield and bunch quality traits.

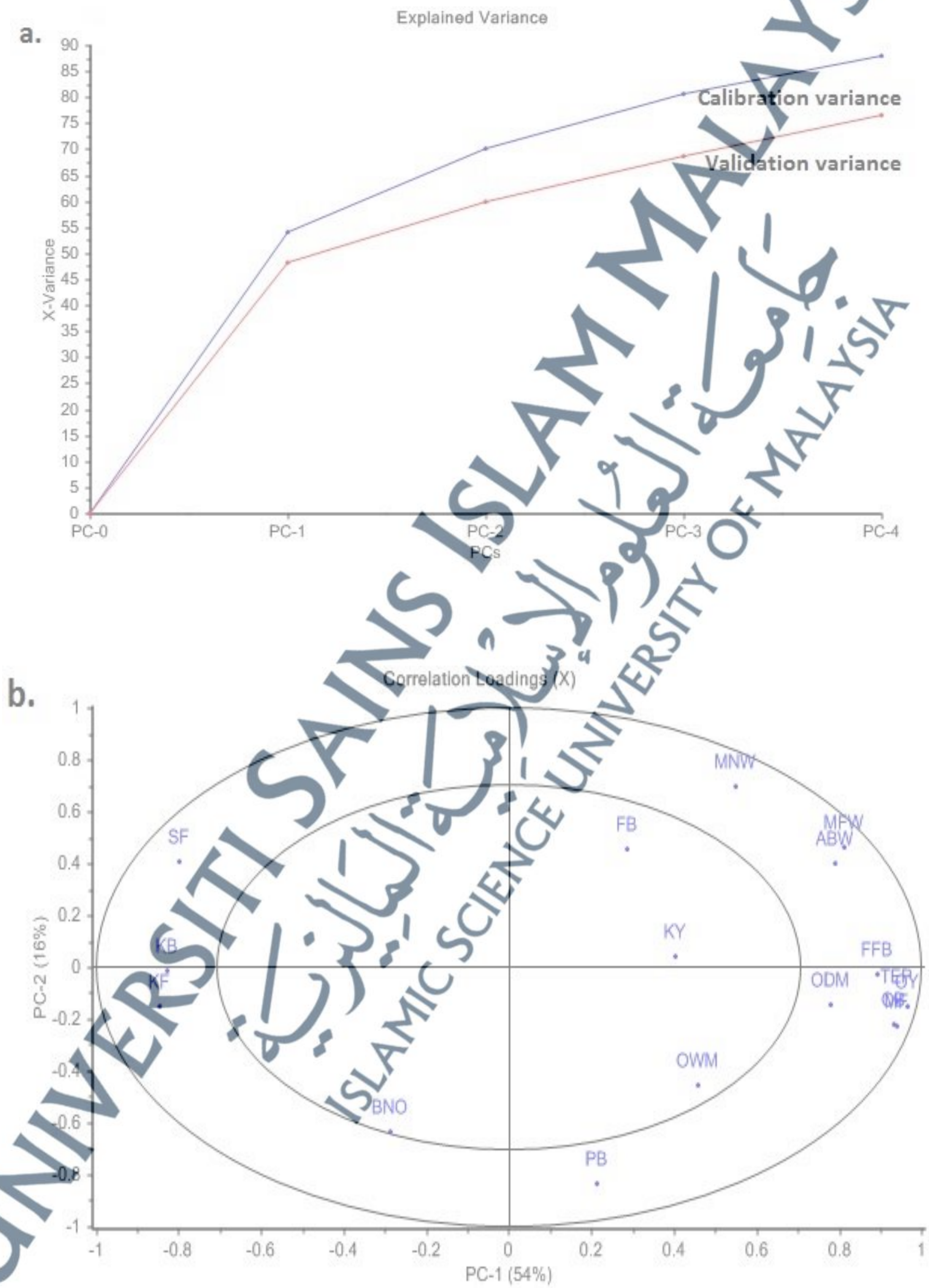
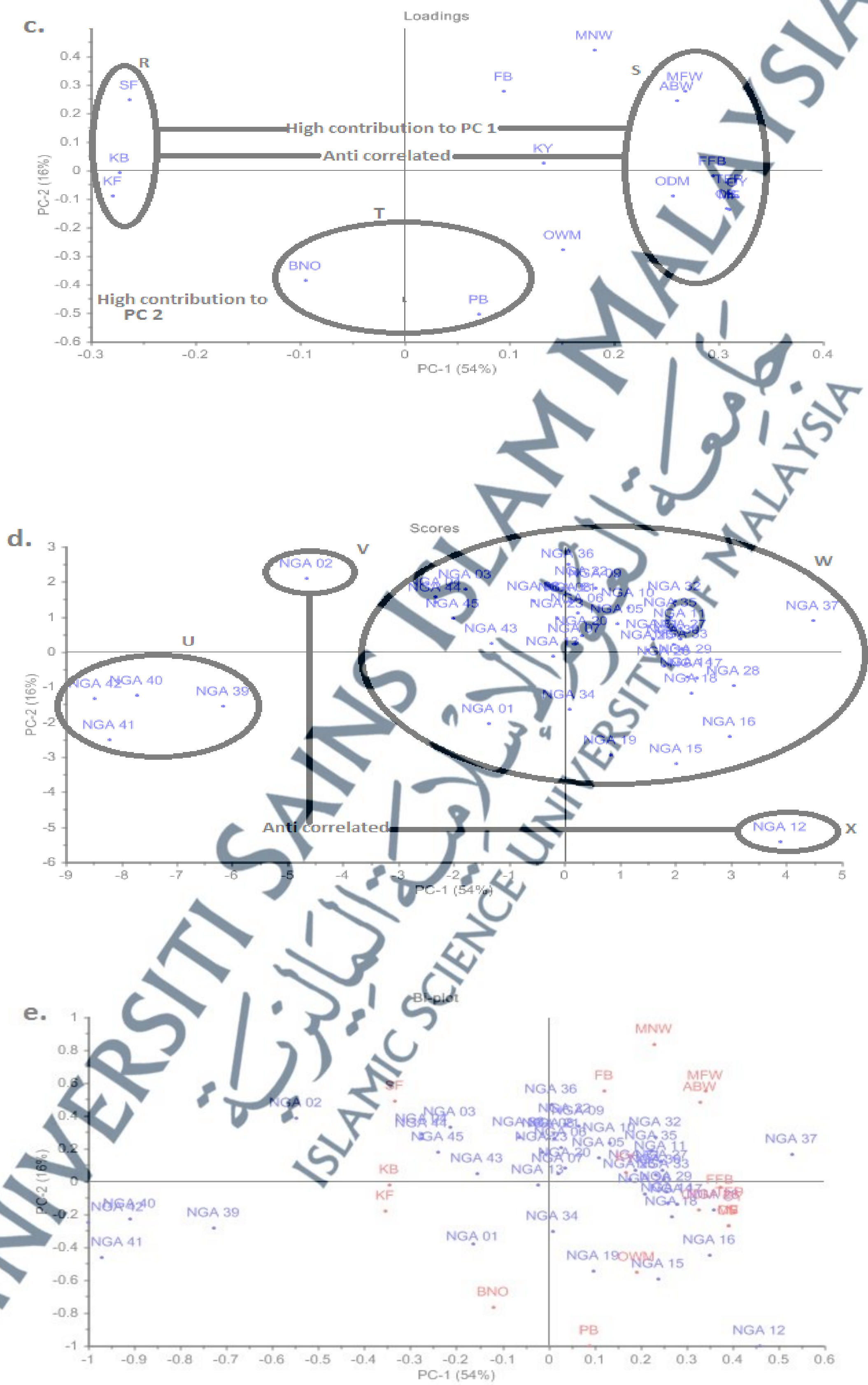


Figure 3 Contd.



The last component, PC3 explained 10.96 % of the total variance and had strong positive loadings on NAR and moderate positive loadings on *f*.

Figure 4b shows the correlation loadings plot for each variable for the displayed components. The plot contains two ellipses and most of the variables are on the outer circle. These variables indicate 100 % variance while those in the inner ellipse indicate 50 % variance. The Figure 4c shows how the variables contribute to each PC (axes) and their correlations. Parameters in K and M have high contribution to PC1 and both are anti-correlated. Those in K exhibit positive correlations to PC1 while parameters in M show positive correlations. Furthermore, parameters in L and N contributed high to PC2 and are also anti-correlated. Parameters in L are positively correlated to PC2 while those in N are negatively correlated to PC2.

The scores plot (Figure 4d) shows the dispersion of the MPOB-Nigerian oil palm accessions on PC1 and PC2. The oil palm populations are presented in three groups on the score plot. It is visible from the graph that group V is well separated from the other two groups. Some populations in T are very close to another while some are well spread out. Populations in U are clustered together as overlapping of populations can be seen and members in this group are also clustered around populations in T. The scatter plot of scores for PC1 and PC2 with loadings displayed on the same plot is shown in Figure 4e and from the graph it can be seen that samples are close to variables that best describes them.

4.3.4 PCA of Oil Quality Traits

Three PCs with eigenvalue greater than one accounted for 81.89 % of the total variance as shown in Figure 5a. PC1 as shown in Table 12 accounted for most of the variation, contributing 50.38 % of the total variance and had strong positive loadings on IV, C18:0, C18:1; strong negative loadings on C14:0, C16:0 and moderate negative loadings on C12:0. PC2 explained 17.82 % of the total variance with C12:0 having moderate positive loadings on it and C16:0 exhibiting strong negative loadings on it. PC3 explained 13.69 % of the total variance with strong positive loadings on C18:2.

Figure 5b shows the correlation loadings plot for each variable for the displayed components. From the plot, only C182 is in the inner ellipse and it indicates 50 % variance while other variables indicate 100 % variance and contribute high to the PC1. The loading of the oil quality traits on PC 1 and 2 is shown in Figure 5c. C12:0, C14:0 and C16:0 are negatively correlated on PC1 while on PC2, C18:2 tend towards zero and do not describe the PC.

The sample score plot for PC 1 and PC 2 is shown in Figure 5d and from the graph, two groups can be observed, the group on the right corner of the graph (H) and the one at the left corner of the graph (G). H comprises of just populations NGA 02 and NGA 19 while group G comprises of the rest of the populations. In G, some populations are also well separated out, while some are clustered together.

The scatter plot for scores and loadings simultaneously is shown in Figure 5e. It can be seen from the graph that NGA 02 and NGA 19 have high C18:0, C18:1 and IV content; NGA 13 has high C16:0; NGA 34 best described by C12:0; NGA 38 and NGA 39 by C14:0 while other populations are close to the centre of the graph which is described by C18:2.

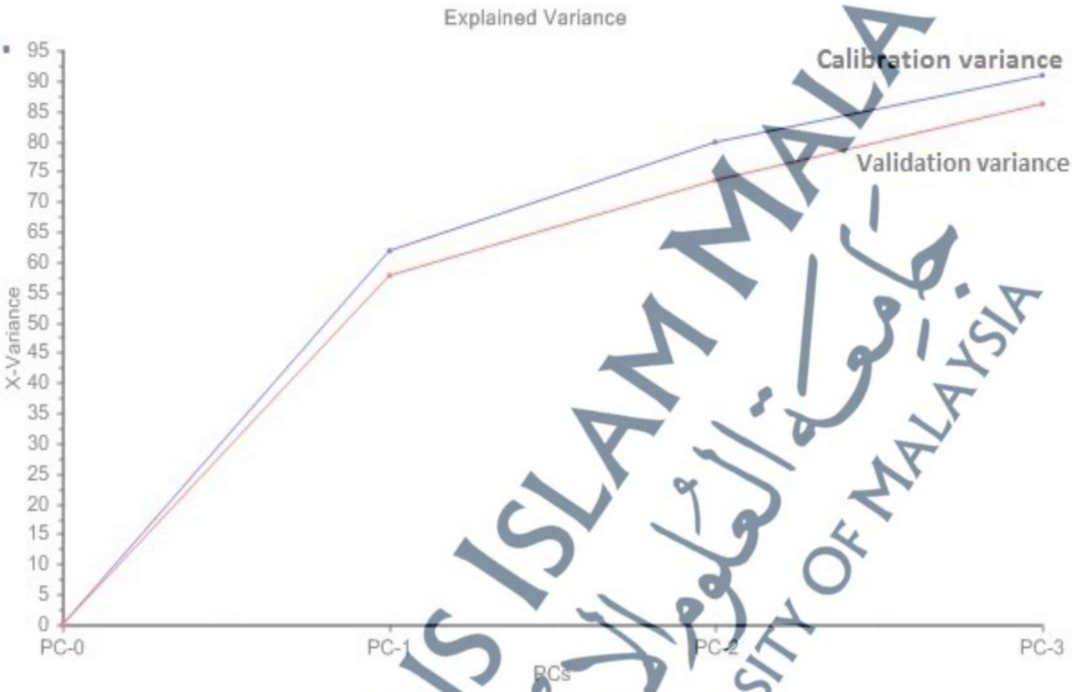
Table 11: Variables correlation loading matrix and total variance explained for vegetative and physiological traits

Parameter	Principal component		
	1	2	3
FP	.631**	-.597**	.167
PCS	-.947*	-.213	.101
RL	-.972*	.123	-.048
LL	-.948*	-.082	.098
LW	-.894*	-.141	-.291
LN	-.917*	.047	-.061
HT	.025	-.843*	.277
LA	-.987*	-.112	-.087
LAI	-.987*	-.112	-.087
DIAM	-.581**	-.426	-.291
LAR	-.394	.567**	-.477
BDM	-.823*	.440	.346
VDM	-.731**	-.650**	.192
TDM	-.937*	-.018	.340
BI	-.311	.925*	.188
F	-.681**	.095	.712**
E	-.993*	-.040	-.045
NAR	.545**	.231	.783*
Eigenvalue	11.15	3.22	1.97
Variance (%)	61.94	17.89	10.96
Cumulative (%)	61.94	79.83	90.79

* Variables that have loadings greater than0.75; ** Variables that have loadings 0.75 – 0.50

Figure 4: Principal component analysis plots (a) Cumulative variance; (b) Correlation loadings plot of vegetative and physiological traits along PC1 and PC2; (c) loading plot of vegetative and physiological traits on PC1 and PC2; (d) PCA scores plot of the MPOB-Nigerian oil palm populations; (e) Bi-plot of 44 MPOB-Nigerian oil palm populations and 18 morpho-physiological traits

a.



b.

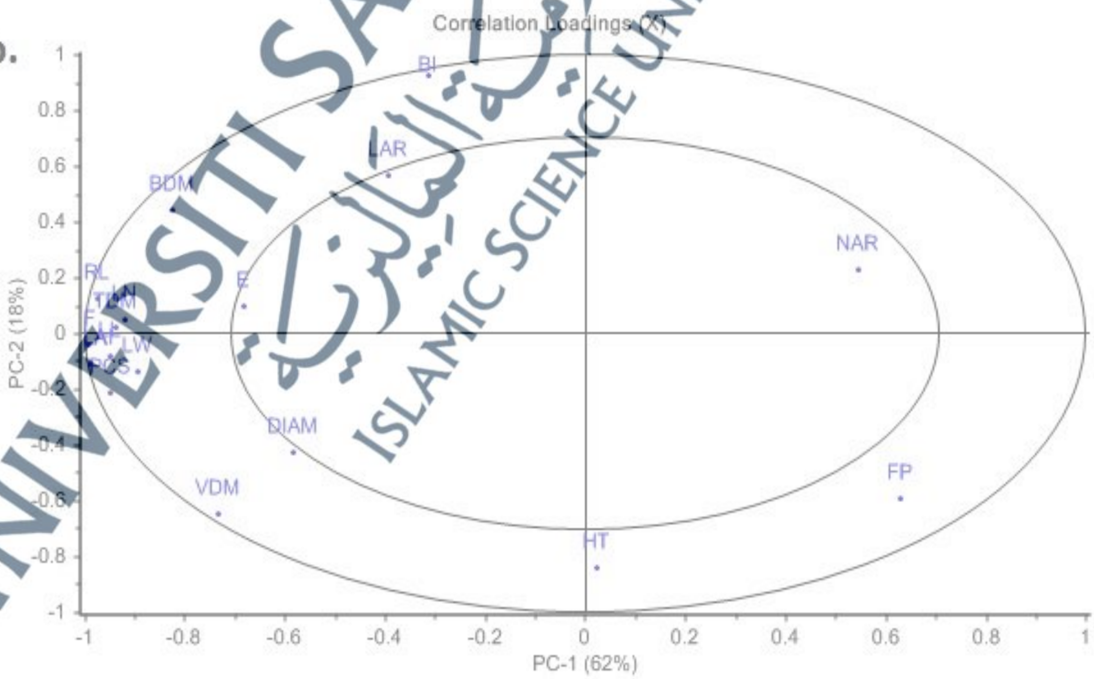


Figure 4 Contd.

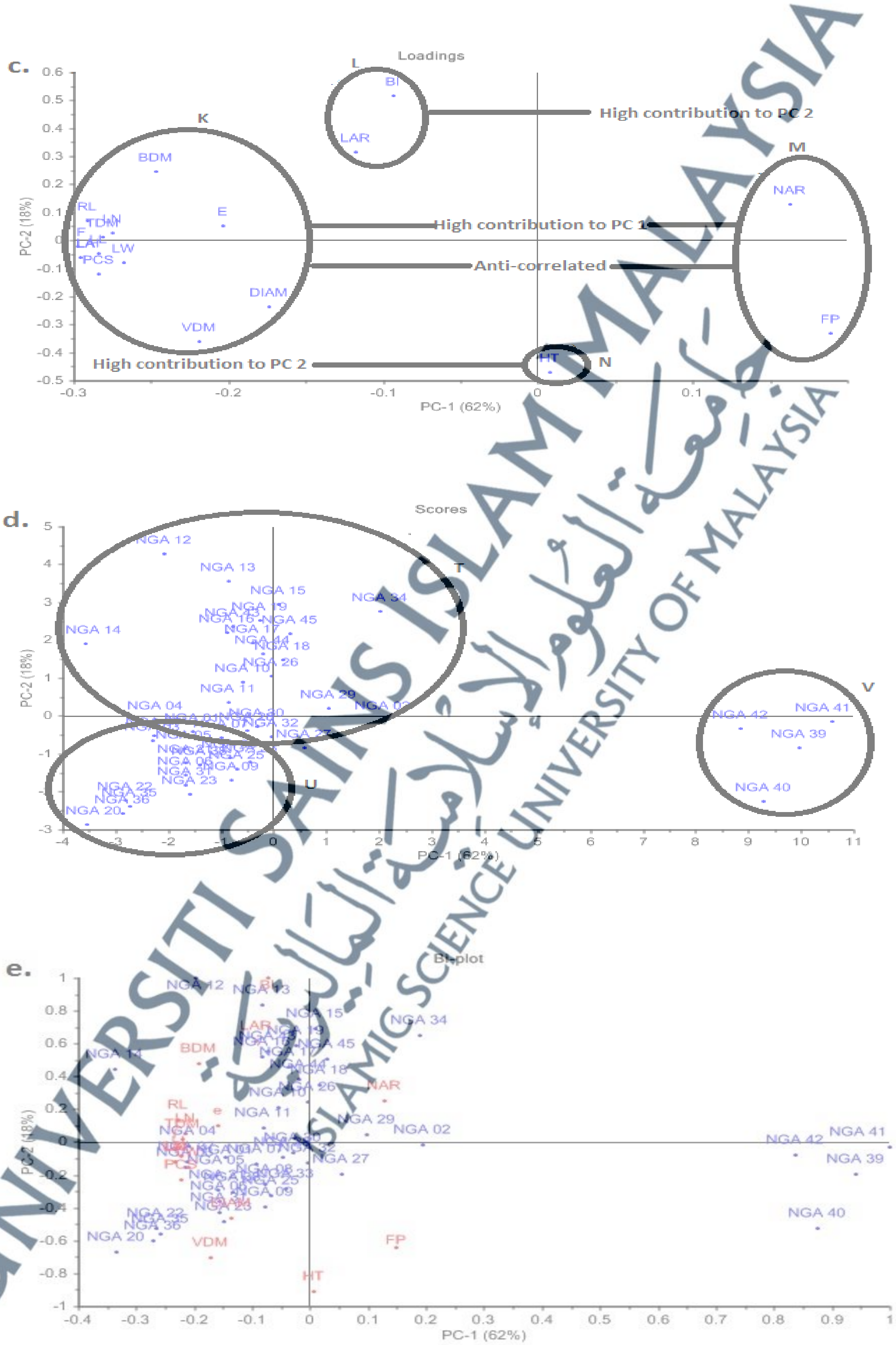


Table 12: Variables correlation loading matrix and total variance explained for fatty acid traits

Parameter	Principal component		
	1	2	3
IV	.855*	-.008	.206
C120	-.577**	.698**	-.042
C140	-.879*	.326	.061
C160	-.857*	-.428	-.070
C161	-.201	-.770*	.133
C180	.789*	.233	.207
C181	.879*	-.004	-.219
C182	-.155	.035	.966*
Eigenvalue	4.03	1.43	1.10
Variance (%)	50.38	17.82	13.69
Cumulative (%)	50.38	68.21	81.89

* Variables that have loadings greater than0.75; ** Variables that have loadings 0.75 – 0.50

Figure 5: Principal component analysis plots (a) Cumulative variance; (b) Correlation loadings plot of fatty acid traits along PC1 and PC2; (c) loading plot of fatty acid traits on PC1 and PC2; (d) PCA scores plot of the MPOB-Nigerian oil palm populations; (e) Bi-plot of 44 MPOB-Nigerian oil palm populations and 8 fatty acid traits

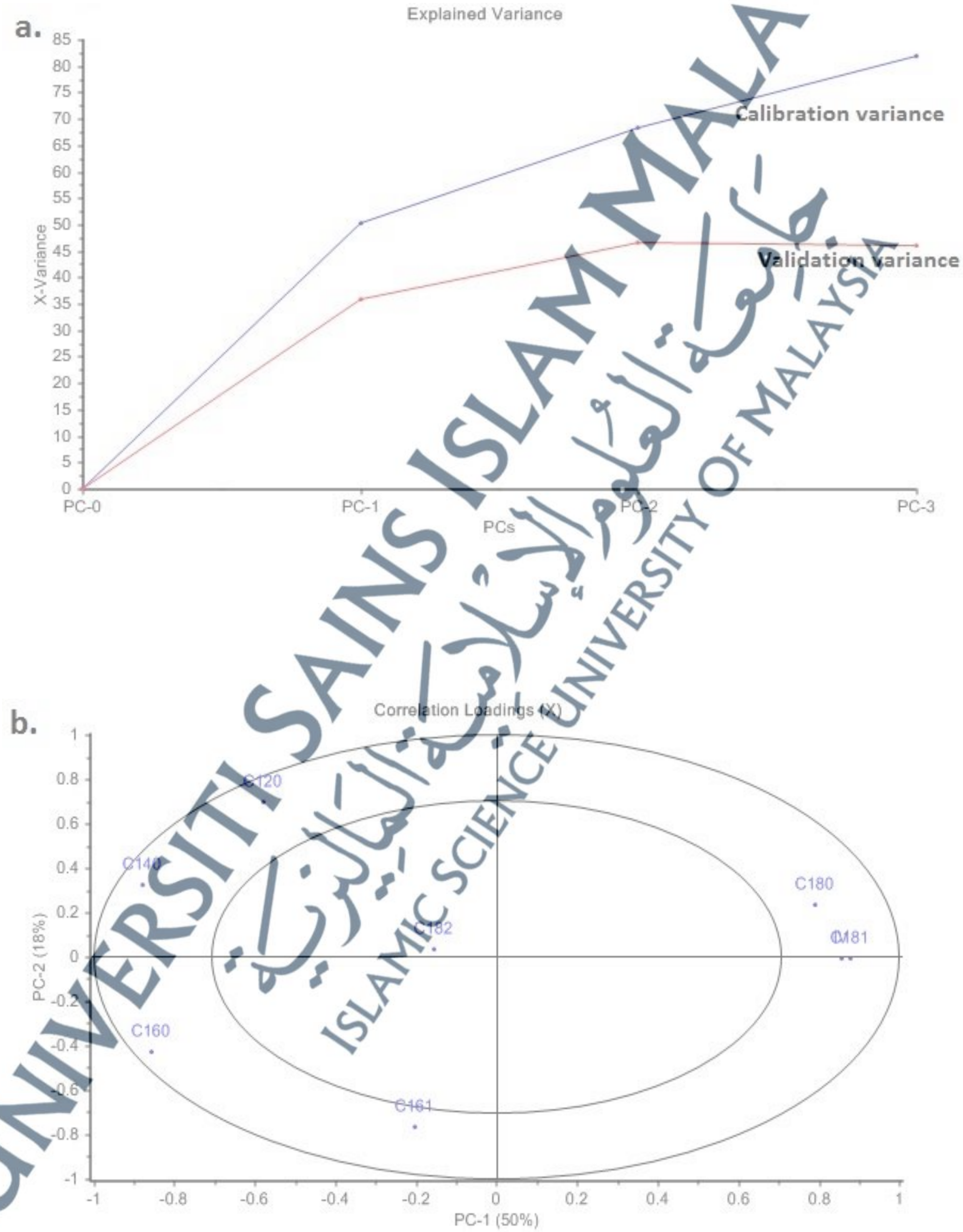
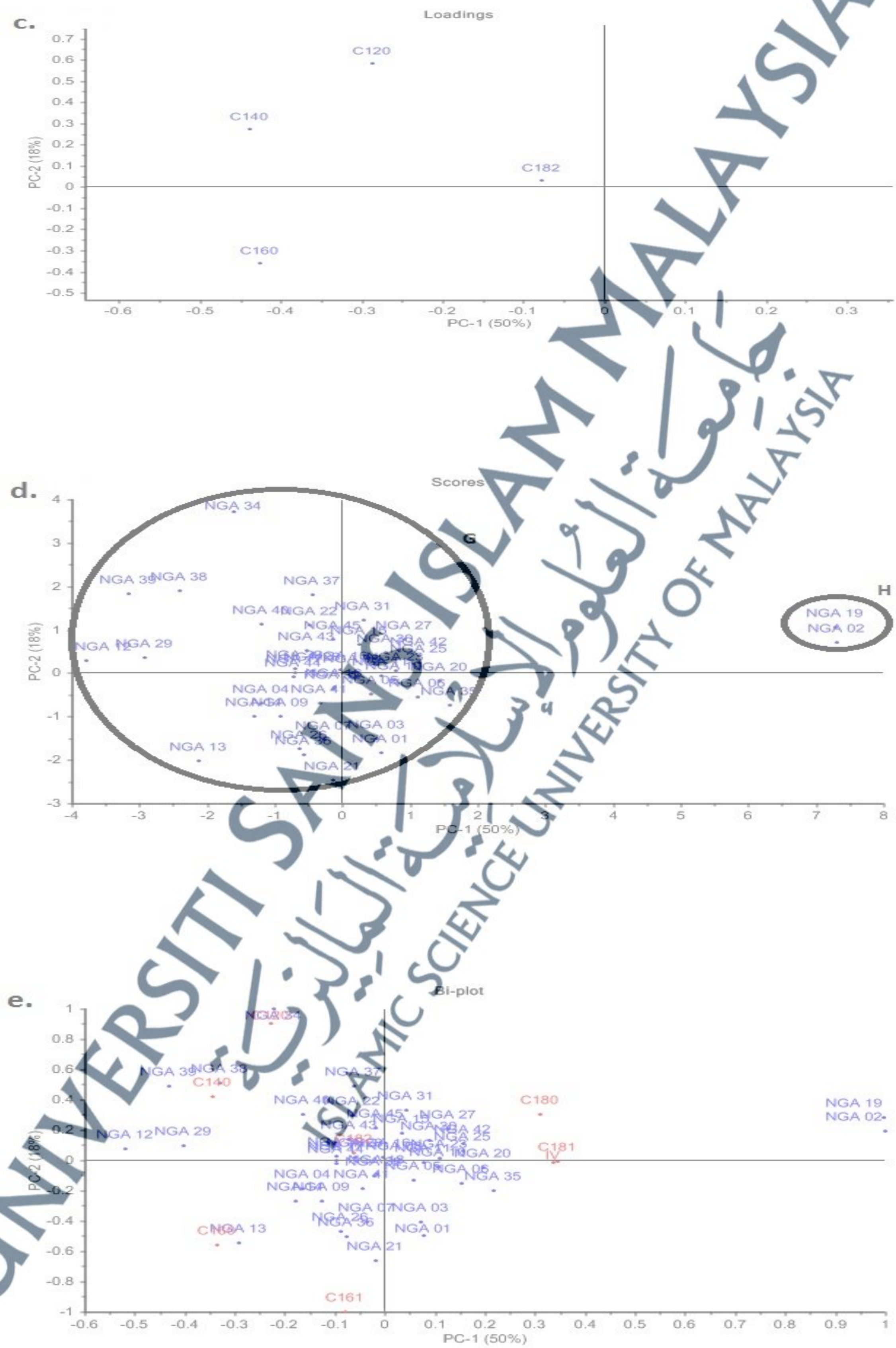


Figure 5 Contd.



4.4 DISCUSSION

4.4.1 Number of Principal Components Extracted

According to Iezzoni and Pritts (1991), the eigenvalue can be used as a criterion to determine how many PCs should be retained. They considered PCs with eigenvalue more than 1 as inherently more informative than any single original variable alone. The number of principal components taken into account was based on the eigenvalue one criterion. The rationale for this is as stated by Kaiser (1960) is that each observed variable contributes one unit of variance to the total variance in the data set, and as such any component that displays an eigenvalue greater than one is accounting for a significant amount of variance and is worthy of being retained. On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variance. For the purpose of this research, components with eigenvalues less than one were not retained and were viewed as trivial, as the purpose of principal component analysis is to reduce a number of observed variables into a relatively smaller number of components (Wiley, 1981).

The explained variance graph, which is often measured as percentage of the total variance in the data is a measurement of the proportion in the data accounted for by the individual PC (Ahmadizadeh & Felenji, 2011). It gives an indication of how much of the variation in the data is described by the different components (CAMO, 2011). The calibration variance indicates the number of PC that was selected for this research based on the eigenvalue one criterion and is based on fitting the calibration data to the model while the validation variance is the proof of the optimum number of principal components needed to describe the PCA model and is computed by testing the model on data that were not used to build the PCA model (CAMO, 2011). Since both the calibration variance and validation variance computed in this study do not differ significantly, it indicates that data are truly representative of the model.

4.4.2 Loadings

The two dimensional scatter plots of X-loadings describe the data structure in terms of variable contributions and correlations. Every variable analyzed has a loading on each PC and this reflects how much the individual variable contributes to that PC, and how well the PC takes into accounts the variation in the variable (Wise et al., 2006). According to Iezzoni and Pritts (1991), the magnitude of the loadings indicates the relative contribution of the individual variable to each PC based on the interrelationships among the variables, and the biological meaning is determined by eigenvectors (weight) and the PC scores.

Variables on the component 1 vs. component 2 represent the largest variations in the data set. PC1 is generally better correlated with the variables than PC2; this is expected as PCs are extracted successively, each one accounting for as much of the remaining variance as possible (CAMO, 2011; Beebe et al., 1998).

4.4.3 Scores Plot

The scores plot is a two dimensional scatter plot (or map) of scores for two specified components (PCs) for PCA. The scores plot shows how well the data is distributed and gives information in the samples (Wise et al., 2006). The scores plot (PC1 vs PC2) was used in this study because the two components summarize more variation in the data than any other pair of components. The closer the samples are in the scores plot, the more similar they are with respect to the components concerned (i.e. they have close values for the corresponding variables). On the other hand, samples for which scores differ greatly are quite different from each other with respect to the variables. The score describes the major features of the sample, relative to the variables with high loadings on the PC (CAMO, 2011; Iezzoni and Pritts 1991).

The Bi-plot is also a two dimensional scatter plot or map of scores for two specified components (PCs) with loadings displayed on the same plot. It enables one to interpret sample propertries and variable relationship simultaneously (CAMO, 2011). As stated by Firincioglu et al. (2009), biplot diagrams can be used to select genotypes that might have favourable combinations of traits for use in a breeding project. For instance, if the aim was to increase seed yield, genotypes that fall in the right quadrants are generally above average for this specific trait. Also, Ahmadizadeh and Felenji (2011) used biplot in comparing different genotypes. In the present study, these data have revealed different associations between traits in the MPOB-Nigerian oil palm germplasm. This indicates that through selection, there is possibility of getting desirable genotypes that can combine high yield with worthy traits (Iannucci, 2011).

4.4.4 PCA of All variables

PCA was applied to the data set of 43 different traits after standardization (the mean of the values for each variable was from each variable value and the result was divided by the standard deviation of the values for each variable). The rationale for this is because after standardization each parameter contributes equally to the data set variance and carries equal weight in principal component calculation (Kumar & Singh, 2006). The result from the PCA identified eight axes with eigenvalue greater than one to have accounted for 90.52 % of the total variation for all the 43 traits studied in the MPOB-Nigerian oil palm germplasm accessions. Okoye and Okwuagwu (2012) reported 89.74 % cumulative of the total multivariate variation of yield performance and parental stock in classifying twenty-four genotypes of oil palm. Zizumbo-Villarreal and Pinero (1998) reported 80 % variation in coconut germplasm in Mexico. Oboh and Fakorede (1997) also reported 90 % total variation from factor analysis on oil palm progeny mean data.

From the Table 9, variables that have significant positive as well as negative impact on the PCs can be said to be the most contributing characters to the variation, especially those on PC1 are the likely source of variation in the oil palm germplasm studied and can be used in the selection of diverse genotypes (Ashfaq et al., 2012). It can be assumed that the characters that delineate these oil palm populations include combination of traits and not a single trait. These findings supports the statement of Oboh and Fakorede (1997) that, within the limits imposed by the environment and genetic composition of the material evaluated; the oil palm breeders not measure more than one or two traits from the first three PCs to obtain information in all other traits in the dataset. The major contributing characters in the oil palm germplasm can be said to be the traits with the highest loading (either positive or negative) on the extracted PC and according to Ahmad et al. (2014), it is customary to choose one variable from the groups of variables loaded significantly on each PC. Hence, for the eight PCs extracted, the best choice of variable would be TDM, PB, HT, C16:0, FB, NAR and C18:2 respectively.

From the loadings and scores plots (Fig. 2b and 2c), PCA did a good job on showing the relationship between the variables and how the variables have contributed to the scores plot; as the loadings cannot be interpreted without scores, and vice versa (CAMO, 2011). Variables in P and O (Fig. 2b) are anti-correlated and this indicates that when the proportion of the traits in P increases, those in O will decrease and vice versa. Thus, populations that have high scores in S/F, K/B, K/F and FP will be low in oil yield and oil yield is very essential in oil palm breeding. Variables that are in the same PC and are very close to each other signify that there are positive correlations between them, and the increase in one will lead to the increase in the other. Jonah et al. (2014) reported similar result in growth and yield of groundnut. The loading plot showed a significant association between FFB/BDM, O/B, M/F, TEP/OY, ODM and *e*; PCS, LAI, *f*, LL, RL, ABW, LN, MFW, TDM and also between BNO and NAR. These associations can also be validated in the correlation studies (Table 8). Jonah et al. (2014) also found that results from correlation studies supported that of PCA.

Okoye et al. (2007) found relatively large effect of BN, FFB, M/F and O/B on oil yield and also positive associations among OY, OWM, M/B, M/F, O/B, F/B and ODM as revealed by GT biplot. This evidence suggests that selection of palms based on these traits will result in high oil yield. The oil palm accessions can be grouped with respect to their scores for each variable (Fig. 2c). Populations in A are very distant from the other populations and they include oil palm populations 39, 40, 41 and 42. These populations are best described as having high proportion of the variables S/F, K/F, K/B and FP and as such they are low yielding population - they exhibit high negative loadings to characters that contribute to most of the variation. Rajanaidu and Rao (1987) reported that these same populations to be of low yield because they were collected from dry areas (Northern Nigeria) and as such received less rainfall (less than 1692mm per year). Furthermore, in the same scores plot, population 02 and 12 are also separated from the rest and also at opposite angle to each other. This is so because population 12 has a high leverage and is best described as being the highest oil yielding population and population 02 is poorly described for variables population 12 has high scores for (FFB, BDM, TEP, OY, OB, M/F); likewise population 12 is poorly described by variables in which population 02 has high scores for. This is why these two populations are anti-correlated to each other. Population 02 is best described by high IV content. Rajanaidu and Rao (1987) as well found NGA 12 to have the highest yielding FFB. Other populations in C also have higher than average for the variables in P and Populations 43, 44 and 45 are average populations in terms of yield as they are found close to the center of the plot (the center is the average of the data set). Populations 14, 35, and 37 are also lie far from the center of the plot and they are also well described by PC1, i.e. they also exert high influence on the PC1. These results conform to the studies of Ithnin et al. (1996), who studied the genetic variability of the Nigerian oil palm germplasm using restriction fragment length polymorphism (RFLP). It was reported by the researchers that populations 12, 35, 36 and 37 showed a higher level of polymorphic than those from dry and intermediate ecotypes. Populations 12, 14, 35 and 37 belong to the same ecotype and where collected from Southern Nigeria with rainfall between 2570 – 4000 mm per annum. This may indicate why they differ from others and populations 43, 44 and 45; also from the same ecotype belonging to an intermediate area with annual rainfall of 1620 – 1750 mm, which also suggests why they are of average samples.

4.4.5 PCA of Yield, Yield Components and Bunch Components

The first four PCs accounted for 87.83 % of the total variation which well described the PCA model. From Table 10, it can be said that variables that contributed mostly to variation are loaded on PC1 and they include those with positive impacts- FFB, ABW, MFW, MNW, M/F, O/B, OY and TEP; and those with negative impacts- S/F, K/F and K/B. On PC 2 BNO (positive) and PB (negative) accounted for most of the variation, KY positively had high influence on PC3 while in PC4, BNO (negative) and FB (positive) accounted for most variation. This signifies that all traits for yield, yield component and bunch traits are source of variation in oil palm except for oil to mesocarp weight (OWM). This finding has been ascertained by many other researchers that characters which load high positively or negatively contributed more to the diversity and are the characters which differentiate the accessions (Ahmad et al., 2014; Hamza et al., 2014; Denton & Nwangburuka, 2011; Iannucci et al., 2011).

In Figure 3b, variables in R and S showed high contribution to PC1 and this indicates that these variables capture most of the variation in terms of the yield, yield component and bunch quality traits. Also both groups of variables are anti-correlated and this signifies populations with high proportion of S/F, K/B and K/F will have low proportions of the other variables. This criterion can be used for selecting high yielding accessions for breeding purpose. Ahmadizadeh and Felenji (2011) found similar pattern in diversity among potato cultivars. Also, BNO and PB showed high contribution to PC 2 with BNO negative impact to the variation. This suggests that a palm with large number of bunches may have high proportion of parthenocarpic fruits. This result can also be validated in the correlation studies where significant correlation exists between BNO and PB.

Figure 3c showed how the samples are distributed in respect to their scores for the variables, and from the graph, populations in U; NGA 39, NGA 40, NGA 41 and NGA 42 are best described by variables in R. These populations have high percentage of S/F, K/B and K/F and they have low proportion of other traits. Population with the highest scores for variables in PC is NGA 37 as it can be seen separate from other populations in W. Other populations that well describe this PC include; NGA 12, NGA 28 and NGA 16. These populations can be selected for good bunch quality traits and oil yield. Also NGA 12 and NGA 02 are anti-correlated; they are far apart because NGA 12 has high positive scores for traits that are well loaded on PC1 while NGA 02 is negatively correlated to this PC. NGA 12 also has high scores for BNO and PB, as well as NGA 15, 19 and 01. Also kernel yield (KY), NGA 43, 44 and 45 are well described by it and can be selected for breeding high kernel yielding palms. These results follow similar pattern as that of Svetlana et al. (2012) and Hossain (2011).

4.4.6 PCA of Vegetative and Physiological Traits

Three PCs with eigenvalue greater than one captured 90.79 % of the total variation and the calibrated variance and validation variance perfectly fit which indicates that the PCA is a good model. Kjær et al. (2004) also reported 57.4 % total variance from two PCs with eigenvalue greater than one of 10 quantitative morphological variables in Sago Palm, a member of the oil palm family. All the vegetative and physiological traits had significant influence on PCs as those traits not captured on the first PC were captured on the subsequent PCs. Oboh and Fakorede (1997) also found vegetative characteristics accounting for the largest mean and total variation on progeny mean and individual palm basis in their research. This indicates that one or two traits can be used to obtain information on all other traits and these traits can be improved on and used as criteria in selecting oil palms for improvement or breeding programme.

Furthermore, FP and NAR imparted high positive significant on PC1 while PCS, RL, LL, LW, LN, LA, LAI, DIAM, BDM, VDM, TDM, *f* and *e* imparted high negative significant. This signifies that populations with high positive scores for FP and NAR will score low for the other traits and vice versa. Similar trend was reported by Soengas et al. (2008) in morphologic and agronomic diversity of *Brassica napus* crops. From Figure 4b, NAR and FP are anti-correlated to other variables on the left quadrant of the graph. In K, the strong relationship between TDM, LL, LN, RL, LAI, LW, F and PCS suggests that if one of the variables is increased, others will also be increased. This result is validated in the correlation studies in Table 9. Bi-plot diagrams can be used to select genotypes that might have favourable combinations of traits for use in a breeding project (Firincioglu et al. (2009). From the biplot in Figure 4e high bunch index (BI) is well described by population 12 and NGA 13 while LAR by NGA 15, 43, 16, 45, 34, 17 and NGA 44. NGA 14 is best choice for BDM while for NAR, NGA 29 and NGA 02 is the best choice. Tewodros (2013) found similar trend in the morpho-agronomical characterization of taro and also Ahmad et al. (2012) reported similar result for genetic diversity in germplasm and cultivars of upland cotton.

Also for HT and FP which loaded negatively on PC2, populations with more positive scores on the on the PC have lower than average values for these variables, while populations with negative scores, have higher than average values for these traits (Rahman & Munsur, 2009). Populations that have high negative loadings for height include NGA 20, 22, 35, 36, 39, 40, 41 and 42. Therefore, oil palm population with high positive loadings for height can be exploited for a long time as short trunk height is a significant trait in oil palm breeding (Sapey et al., 2012). Palms that well describe short trunk height include NGA 12, 13, 15, 34 and NGA 45. These palms also have high BI, LAR and BDM. This signifies that HT is negatively correlated to BI. This is validated in the correlation studies where HT had negative significant correlation with BDM and BI but its negative correlation with LAR was not significant. It can also be assumed that the taller an oil palm tree, the lesser its bunch index and bunch dry matter. FP and HT are also close together in the Figure 4b and as such there are similar. FP also was anti-correlated with BDM and VDM but there was no significant correlation between FP and BDM but significantly correlated with VDM. Kumar et al. (2010) reported similar trend for genetic diversity in groundnut, also Ackerfield (2002) in morphometric analysis of ivy genus.

4.4.7 PCA of Oil Quality Traits

The PCA of the oil quality traits resulted in three PCs which accounted for 81.89 % of the total variation. This is validated in the work of Arasu et al. (1987) where it is stated that the phenotypic variation for most fatty acid traits is considerably larger in these Nigerian germplasm materials. The validation variance curve in Figure 5 indicated that two PCs are enough to describe the variation in the fatty acid composition of the oil palm germplasm. However, due to the eigenvalue one criterion used in this study, three PCs with eigenvalue greater than one were retained. Chakravorty et al. (2013) observed similar trend in phenotypic diversity of landraces of rice.

In Table 11, IV, C18:0 and C18:1 loaded positively on the first PC and this indicates that populations with high positive scores for these traits have high content for them. Populations with high content for IV, C18:0 and C18:1 can be seen clearly in group H (Figure 5c)-NGA 02 and NGA 19. This is the most probable reason why these two populations are marked out from the rest of the populations which have negative scores for the traits. Khan et al. (2009) found oleic acid (C18:0) as one of the traits accounting for variation in safflower.

Furthermore on PC2, C12:0 was positively correlated and from the score plot of fatty acid traits (Figure 5c), population 34 well describes this fatty acid and this is why NGA 34 is also quite separated from other populations in G. Also, C14:0 is well described by NGA 39, 38 and 37. The last PC was significantly loaded by C18:2 and NGA 34 also scored high for this trait. It is of notice that variation for fatty acid traits is quite limited as compared to other traits, but, nevertheless this can be an important source of selection for palms that have high content for good oil quality. Khan et al., (2009) reported similar pattern in safflower.

Despite the fact that only few PCs were generated from all the characters of the MPOB-Nigerian oil palm germplasm, PCA was able to account for high variation of the total variations encountered in the oil palm germplasm. Thus, the capacity of the PCA in data reduction without loss of information was ascertained (Wiley, 1981).

4.5 CONCLUSION

In short, Principal Components Analysis (PCA) has proved useful in a number of ways. Firstly, it has helped in the reduction of the basic dimensions of the variability in the measured dataset to the smallest number of meaningful dimensions. Secondly, it has examined the correlations between the variables, the elimination of variables which contributed relatively little information, and lastly the examination of the grouping of individuals on two dimensional scatter plots. With respect to the strong inter-correlation between the variables, a little number of variables may be adequate to distinguish between the oil palm accessions. In this study, 43 variables were used and only eight (8) components were extracted based on these variables. The implication of this analysis suggests that the time and labour in measuring a large number of variables may be preserved by measuring a few of the variables loaded significantly on the extracted PCs and generalizing the information as the totality of the other variables. The variables that were loaded on the first PC and explained the most variation include; FFB, ABW, MFW, MNW, MF, ODM, OB, OY, KY, TEP, FP, PCS, RL, LL, LW, LN, LA, LAI, BDM, VDM, TDM, *e*, and *f*. These variables can be used as criteria for improvement of oil palm accessions for high oil yield. Furthermore, the oil palm populations with high scores for these variables can be selected singly or in combination as palm of superior yield quality.