# ANALYSIS OF GENETIC DIVERSITY AND POPULATION STRUCTURE OF WILD LOQUAT (*UAPACA KIRKIANA* (Müell) Arg.)) USING DARTSEQ-GENERATED SINGLE NUCLEOTIDE POLYMORPHISMS

**JANE MAURINE GATI (BSc Biochemistry)**

**I56/CTY/PT/27122/2011**

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE (BIOTECHNOLOGY) IN THE SCHOOL OF PURE AND APPLIED SCIENCES OF KENYATTA UNIVERSITY**

**APRIL 2022**

## DECLARATION

I declare that this is my original work and has not been presented for degree award in any other university or other awards.

**Signature ……………………………** **Date…….………………...………….**

**Jane Maurine Gati**

**I56/CTY/PT/27122/2011**

**Department of Biochemistry, Microbiology and Biotechnology**

**Kenyatta University**

## SUPERVISORS

We confirm that the work reported in this thesis was carried out by the student under our supervision:

**Signature**……………………………… **Date**………………………………

**Prof. Steven M. Runo**

Associate Professor

Department of Biochemistry, Microbiology and Biotechnology

Kenyatta University

**Signature**……………………………… **Date**………………………………

**Dr. Alice Muchugi**

Genebank Manager

Genetic Resources Unit

World Agroforestry Centre

## DEDICATION

To my dear children Antony and Gianna, you give me the strength to carry on.

## ACKNOWLEDGEMENTS

Working at the International Centre For Research in Agroforestry (ICRAF) was instrumental towards my MSc thesis. During the 12 months that I was attached there as a research fellow, I learnt basic to advanced molecular biology techniques, including DNA extraction, gel electrophoresis, restriction enzyme digests and Polymerase Chain Reaction (PCR).

I thank Dr Alice Muchugi and the ICRAF capacity development unit for granting me this opportunity to learn at ICRAF. I am incredibly thankful for the provision of funding for the project and even affording me a stipend at the most needed time of my life.

I am grateful to Prof. Steven Runo, my university supervisor, for his patience, immense support, and mentorship throughout my thesis project.

Further on, I want to thank the ICRAF molecular biology laboratory research assistants Robert Kariba and Samuel Muthemba for their technical support and making this demanding time joyful and always efficient.

Completion of my coursework and proposal registration would not have been possible without the support of Dr Marc Ghislain (International Potato Centre). Thank you for funding my university tuition fee; I much appreciate it.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF APPENDICES

## ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| **AFLP** | Amplified Fragment Length Polymorphism |
| **ALP** | Amplified Length Polymorphism |
| **AMOVA** | Analysis of Molecular Variance |
| **BIC** | Bayesian Information Criterion |
| **CA** | Correspondence Analysis |
| **CoP** | Coefficient of Parentage |
| **CTAB** | Cetyl Trimethyl Ammonium Bromide |
| **DAPC** | Discriminant Analysis of Principal Components |
| **DArT** | Diversity Arrays Technology |
| **DArTseq** | Diversity array technology sequencing |
| **Dest** | Measure of population differentiation |
| **DRC** | Democratic Republic of Congo |
| **Dst** | Gene diversity among samples |
| **Dstp** | Corrected gene diversity among samples |
| **EM** | Expectation maximization |
| **FASTQ** | a text file with sequence data arising from a flow cell |
| **$F_{IS}$** | Inbreeding coefficient per overall loci (allele frequencies within populations) |
| **$F_{ST}$** | Fixation index (variance between populations) |
| **Fstp** | Corrected fixation index |
| **GBS** | Genotyping by Sequencing |
| **Hs** | Genetic diversity within populations |
| **Ht** | Overall gene diversity |
| **Htp** | Overall genetic diversity |

| | |
|---|---|
| **ICRAF** | International Centre for Research in Agroforestry |
| **IFTs** | Indigenous fruit trees |
| **ISSR** | Inter-simple sequence repeat |
| **iTOL** | interactive Tree of Life |
| **KDCompute** | An online platform and an application for analysis of sequence data |
| **MAF** | Minor Allele Frequency |
| **NGS** | Next-Generation-Sequencing |
| **Nipals** | Nonlinear Estimation by Iterative Partial Least Squares |
| **NJ** | Neighbor-joining |
| **PC** | Principle Components |
| **PCA** | Principal Component Analysis |
| **PCoA** | Principal Coordinate Analysis |
| **PCR** | Polymerase Chain Reaction |
| **PIC** | Polymorphism Information Content |
| **PPCA** | Probabilistic Principal Component Analysis |
| **QTL** | Quantitative Trait Loci |
| **RAPDs** | Random Amplified Polymorphic DNAs |
| **RFLP** | Restriction Fragment Length Polymorphism |
| **RGAP** | Resistance Gene Analogue Polymorphism |
| **SCARs** | Sequence Characterized Amplified Regions |
| **SCoT** | Start Codon Targeted |
| **SilicoDArT** | Dominant microarrays markers that are scored for presence or absence of one allele |
| **SMC** | Simple Matching Coefficient |

| | |
|---|---|
| **SNPs** | Single Nucleotide Polymorphisms |
| **SSRs** | Simple Sequence Repeats |
| **STRs** | Short Tandem Repeats |
| **STSs** | Sequence Tagged Sites |
| **SVD** | Singular Value Decomposition |

# ABSTRACT

*Uapaca kirkiana* (Müell) Arg, is a popular fruit tree that grows in the wild and is majorly found in the Miombo Woodland. It is popularly known as sugar plum or the wild loquat by the English name. It is a species of plant in the Euphorbiaceae family. *U. kirkiana* has been found to grow naturally south of the equator in Mozambique, Tanzania, Burundi, Zambia, Malawi, Zimbabwe, Burundi, Angola and Democratic Republic of Congo. There are 60 known species of the genus *Uapaca*. Increased consumption and utilization of *U. kirkiana* has led to high demand for the fruit and tree. Increased population and human activities have led to high pressure on land. As a result, forest reserves and national parks have been cleared to create space for the growing demand leading to loss of biodiversity. The domestication of *U. kirkiana* is a more significant step towards the management and conservation of biodiversity. Information on the amount as well as the distribution of genetic diversity is essential in effective management of germplasm resources. However, minimal molecular genetic evaluation on *U. kirkiana* has been carried out. The objectives of the research were to assess the genetic diversity and population genetic parameters, genetic relationships and population structure in *U. kirkiana* sampled from International Centre for Research in Agroforestry gene bank locations. Leaf material from 500 samples of *U. kirkiana* were collected, air-dried and well-preserved using silica gel then kept at –20 C till the extraction of DNA. The extraction of genomic DNA was done using the Cetyl Trimethyl Ammonium Bromide method with variations. Samples were then loaded onto 96 well plates and were sequenced at the Diversity Arrays Technology Pty. Ltd Australia. Data analysis was conducted through R, PHYLIP, and iTOL applications. The populations were divided into four groups by discriminant analysis of principal components and in the Neighbor joining analysis where cluster 1 had a total of 3 individuals, cluster 2 with 47, cluster 3 with 2 and cluster 4 with 289 individuals. However, the grouping pattern did not correspond to the geographical distribution of the plant. The overall genetic diversity was low with a value of *Ht*=0.1040. Analysis of molecular variance results indicated a high genetic density of 93.4% within samples and a lower genetic density of 1.3% between populations. Since the population was divided into four clusters, it would be economical to select a representative sample of each cluster to be preserved for germplasm conservation. The genetic diversity was low across the populations which may have been a result of the tree conservation strategy. The Germplasm conservation unit at International Centre for Research in Agroforestry may want to use populations that are genetically distant to increase diversity and enhance the long-term existence of the fruit tree. Genetic information obtained from this study will be beneficial in the domestication program and the genetic resources unit at the International Centre for Research in Agroforestry. Further analysis of *U. kirkiana* accessions for sex markers will lead to identification of the sex-specific markers at the molecular level and this information will be helpful in selection of the most desirable.

# CHAPTER ONE

# INTRODUCTION

## 1.1 Background information

*Uapaca kirkiana* is known as sugar plum or the wild loquat in its English name and is major fruit tree found in the Miombo woodlands. *U. kirkiana*, is a species of a plant in the Euphorbiaceae family. It is known by different scientific names including *U. homblei, U. goetzei, U. albida, U. banguelensis,* and *U. greenwayi*. *U. nitida, U. paludosa* (syn. *U. guineensis*) and *U. sansibarica* (syn. *U. macrocephala*) are close relatives of *U. kirkiana* (Mwase *et al*., 2010). *U. kirkiana* occurs naturally south of the equator in Mozambique, Tanzania, Burundi, Zambia, Democratic Republic of Congo, Burundi, Malawi, Angola, and Zimbabwe. There are about 60 species in the genus *Uapaca* and there is more diversity in the Zaire basin and the Southern region of Miombo woodlands (Ngulube *et al.*, 1995).

*U. kirkiana*, regarded as a fruit, is a crucial famine food in Tanzania used to make sweet beer, jam or sweet meat. The roots are used to treat indigestion, the flowers are utilised for honey production and the leaves are used as fodder. The wood is used to make charcoal and items such as spoons, furniture and timber. Therefore, the wild loquat has been identified to be a preferred fruit in the regions where it is found because of its role in nutrition, economic empowerment, and food security. It is easy to distinguish *U. kirkiana* from the related species because of its broad and feathery leaves as well as the rounded crown (Orwa *et al.*, 2009; Mwase *et al*., 2010).

Genotyping is one way through which difference in genetic makeup in an organism can be determined. Through single nucleotide polymorphisms (SNPs) genotyping, genetic variations between members of a species can be determined. Single nucleotide polymorphisms markers are plenty in the genome and bi-allelic. As such, SNPs provide the highest accuracy when compared to other molecular markers. Therefore, SNPs have been preferred for Quantitative Trait Loci (QTL) mapping and population diversity studies (Mammadov *et al.*, 2012). Genotyping is important in the identification of the genetic traits of economic importance and beneficial in genomic and marker-assisted selection. Knowledge of the genetic diversity as well as the genetic structure of a plant is essential for crop improvement (Chen and Sullivan 2003). To genotype a germplasm identified by a broad set of SNPs may prove to be costly. Therefore, next-generation sequencing techniques applied in genotyping use a fraction of a genome. That way, much of the effort that would have been concentrated on the large data set into finding polymorphic sites in a set of lines relevant in each study. A subgroup of SNP markers is selected depending on the study and location in the genome to create a basis for the analysis of all the selected SNPs at once (Sonah *et al.*, 2013).

Diversity array technology sequencing (DArTseq) is a genotyping system that utilizes Next-Generation-Sequencing (NGS) platform in the discovery of markers. DArTseq allows testing many samples at the same time and it helps in the analysis of samples whose sequences are unknown (Huttner *et al.*, 2005). Like Amplified Length Polymorphism (AFLP), DArTseq reduces the DNA complexity in a sample to get a genome representative. A typical DArTseq method consists of restriction enzyme digestion and adapter ligation, then PCR amplification and finally, detection through hybridisation. DArTseq can be used in genetic map construction, diversity analysis, Quantitative Trait Loci (QTL) analysis, cultivar identification and genome profiling (Appleby *et al.,* 2009). When compared to other genetic markers like Single Sequence

Repeats (SSRs), Diversity Arrays Technology (DArT) markers survey more loci per reaction, and are therefore more suitable in the analysis of “orphan crop” species in which molecular markers have not been developed or genetic information is unavailable (Huttner *et al*., 2005; Hurtado *et al*., 2008 ).

## 1.2 Problem statement and justification

Indigenous fruit trees (IFTs) make more than 75% of the Miombo woodlands. *U. kirkiana* fruits and products are most preferred by consumers (Akinnifesi *et al*., 2002). A study by Kalaba *et al*. (2009) recorded biodiversity loss of IFTs in the Miombo woodlands. There has been scarcity of *U. kirkiana* in the Miombo woodland as result of charcoal burning and land clearing. Increased population and human activities such as agriculture have led to increased pressure on land. As a result, forest reserves and national parks have been cleared to create space for the growing demand. According to Jinga *et al*. (2020) there will be a loss in *U. kirkiana* as a result of climate change between the year 2050 and 2070.

There has been increased consumption and utilisation of *Uapaca kirkiana,* notably among the low-income households (Mithöfer and Waibel, 2003). This demand can be met through the cultivation of the indigenous trees on farms. The International Centre for Research in Agroforestry (ICRAF) began a domestication program for *Uapaca kirkiana* in Southern Africa in order to conserve biodiversity, avoid losses due to deforestation and provide a source of income to the rural community (Mithöfer and Waibel, 2003).

The domestication of *U. kirkiana* is a more significant step towards the management and conservation of biodiversity. However, data on the amount and genetic diversity distribution is essential in the effective management of germplasm resources. Indeed, a comprehensive genetic structure of populations is crucial for a sustainable domestication strategy and such a genetic structure is not available now. Previous studies conducted on *U. kirkiana* to assess the genetic difference were based on AFLPs (Mwase *et al.,* 2007; Mwase *et al.,* 2010). Generally, minimal molecular genetic evaluation on *U. kirkiana* has been carried out (Lengkeek *et al.,* 2006). Therefore, understanding the genetic characteristics of *U. kirkiana* will help determine diversity and population structure, information that will be beneficial in the domestication program as well as to the ICRAF Genetic Resources Unit.

### 1.3 Null hypothesis

i) There is no genetic variability within *U. kirkiana* species under study

ii) There is no systematic organization of the genetic variability in the *U. kirkiana* species under study

### 1.4 Objectives

#### 1.4.1 General objective

To determine genetic differences in *Uapaca kirkiana* (Müell) Arg. based on SNPs generated through DArTseq.

#### 1.4.2 Specific objectives

i) To determine genetic diversity and population genetic parameters of *U. kirkiana* (Müell) Arg. from selected locations in Africa.

ii) To determine genetic relationships and population structure of *U. kirkiana* (Müell) Arg. from selected locations in Africa.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1 *Uapaca Kirkiana* (Müell) Arg and other related species

*U. kirkiana*, an indigenous tree from sub-Saharan Africa, grows in hot and dry zones of Tanzania, Mozambique, Zimbabwe, Malawi, Zambia, Burundi, southern Democratic Republic of Congo (DRC) and eastern Angola (Ngulube *et al.*, 1995). *Uapaca* genus is composed of 60 species, 49 of which are found in tropical Africa and the rest are restricted to Madagascar. *U. kirkiana* is classified into the family Euphorbiaceae, clan Phyllanthoideae, subclan Antidesmeae, and as the only representative of the subtribe Uapacinae (Ngulube *et al.*, 1996). The primary relatives of *U. kirkiana* are *Pterocarpus angolensis*, *Brachystegia spp*, *Pericopsis angolensi*, *Julbernardia*, *Parinari curatellifolia* and other *Uapaca* species. *U. kirkiana* can be identified from other *Uapaca* species by its distinctively wide, rugged leaves and adjusted crown. *U. kirkiana* is dioecious with distinct male and female trees, and unisexual inflorescences begin from axillary locations in the leaves or branchlets. The spatial dispersions of male and female trees in characteristic populaces are generally unreported (Ngulube *et al.*, 1996).

## 2.2 How population diversity is measured in plant species

Genetic characteristics and demographic features define the variability of a population (Luck *et al*., 2003). Richness, distribution, genetic diversity and size of a population are the descriptive features of population diversity. While population richness describes the number of populations in a locality, distribution explains how the populations are spread out over an area. Populations can be uniform, clumped or randomly distributed (Turchetto *et al*., 2016). The population size is determined by the number of individuals in a population. Genetic differentiation occurs within and among populations and is determined by the amount of genetic diversity. The greater the genetic diversity, the more adapted the population to the ecological changes (Jump *et al*., 2009).

Several studies have utilised the above characteristics to assess the extent of diversity within a group of individuals. In a study by Luo *et al.* (2019), assessment of genetic diversity was done based on allele properties including expected heterozygosity (*He*), polymorphism information content (PIC) and minor allele frequency (MAF). In another study conducted by Baloch *et al*. (2017), genetic diversity was determined by calculating the genetic distance among the landraces and then conducting a Neighbour Joining (NJ) tree analysis based on the genetic distance matrix. On the other hand, Mahboubi *et al*. (2020) assessed genetic variability based on PIC, genetic distance and analysis of clusters using the NJ dendrogram.

## 2.3 Molecular markers and their role in genetic diversity

Molecular markers, also called DNA markers are sequences of DNA in a genome that occur in different forms and can be identified by use of molecular techniques (Avinash *et al*., 2014). DNA markers are classified into hybridization-based markers and Polymerase Chain Reaction (PCR)-based markers. Restriction Fragment Length Polymorphism (RFLP) represents a hybridization-based molecular marker. PCR-based molecular markers consist of Amplified Length Polymorphism (ALP), Amplified Fragment length Polymorphism (AFLP), Random Amplified Polymorphic DNAs (RAPDs), Restriction Fragment Length Polymorphism (RFLP), Simple Sequence Repeats (SSRs), Sequence Tagged Sites (STSs), Sequence Characterized

Amplified Regions (SCARs), Start Codon Targeted (SCoT) and Microsatellites or Short Tandem Repeats (STRs) (Kordrostami and Rahimi, 2015).

Molecular markers are further classified into those that can demonstrate homozygosity or heterozygosity. Co-dominant markers show heterozygosity, while dominant markers show homozygosity. RAPDs and AFLPs are dominant markers while Inter-simple sequence repeats (ISSRs) SNPs, SSRs, STSs, RFLP and SCARs are codominant markers (Idrees and Irshad, 2014). Other studies have classified the PCR-based markers into those that are used for random genome profiling like AFLP, RAPD, ISSR, SCARs, Resistance Gene Analogue Polymorphism (RGAP), and those that target specific genome sites such as SNPs, GBS, DArT and microsatellites (Grover *et al*., 2016). The study according to Dar *et al*. (2019) agrees with this classification and in this study, there were three classes of molecular markers including hybridization-based markers, PCR-based and sequence-based markers. Most recent technologies such as genotyping by sequencing uncover the level of genetic variation in an extraordinary way without having to sequence the entire genome of a species (Porth and El-Kassaby, 2014).

Molecular markers have been used to provide information on the genetic characteristics of plants which has helped to determine the distribution and the amount of genetic variability within species and among populations. DNA markers have also been used in fingerprinting and in developping linkage maps in plants. In a study by Bakoumé *et al*. (2015), SSR markers were used to determine the genetic diversity of the world’s largest oil palm (*Elaeis guineensis* Jacq.). The genetic diversity was high with a mean allele per locus of 13.1 and 0.644 heterozygosity. From the analysis, the population was divided into three clusters. In another study by

Etminan *et al.* (2016), ISSR and SCoT markers were used to analyse genetic diversity in durum wheat genotypes. The ISSR and SCoT markers were highly polymorphic at 98.7% and 100% respectively. In addition, the genetic variations were high within populations with values of 90% for ISSR and 93% for SCoT markers. In a different study by Pidigam *et al*. (2019), RAPD markers were used to determine the genetic variation in yardlong bean (*Vigna unguiculate* (L.) Walp subsp. Sesquipedalis Verdc. The RAPD markers were found to have 100% polymorphism and the population was divided into five clusters.

## 2.4 Estimating the amount of genetic diversity in plants

Genetic diversity is made up of different traits that are inherited within same species. Genetic diversity is a result of different alleles in a gene of dissimilar individuals that lead to contrasting phenotypes (Ellegren and Galtier, 2016). Diversity is vital in plant survival and improvement of crops. It is because of diversity that plants can adapt to varied environments and withstand changing climates (Govindaraj *et al*., 2015). Several factors influence genetic diversity including evolutionary forces such as mutation, migration and genetic drift. These forces alter allelic frequency thereby affecting genetic diversity. Domestication of plant species and inbreeding reduce diversity whereas geneflow within populations, mutations and outbreeding increase diversity (Ingvarsson and Dahlberg, 2019). Morphological, cytological and biochemical markers have been used in the analysis of genetic diversity. However, as a result of introduction of genomic tools, molecular markers have taken precedence (El-Esawi, 2017).

To measure genetic diversity the Coefficient of Parentage (CoP) is determined where a value of one and zero represent relation and unrelation respectively. Determination of genetic distance between entities is another way of measuring diversity. When genotypes have closely related genes the genetic distance between them is reduced (Bhandari *et al.*, 2017). Allelic diversity which is determined by Polymorphism Information Content (PIC), percentage polymorphic loci, gene diversity (He) and average number of alleles per locus can also be used to determine the extent of genetic diversity. The diversity between and among populations is determined according to Nei (1978), using the equation $H_T = H_S + D_{ST}$ where $H_T$ is total observed diversity; $H_S$ is diversity within population; and $D_{ST}$ is diversity between population. F-statistics indices are used to determine the level of expected heterozygosity based on the expression $F_{IT} = 1\text{-}F_{IS} + 1\text{-}\ F_{ST}$, where variance of allele frequencies within populations is ($F_{IS}$), the allele frequencies variance between populations is ($F_{ST}$), and ($F_{IT}$) is the inbreeding coefficient compared to the whole population (Pagnotta, 2018). Multivariate statistics such as Principal Component Analysis (PCA), Principal Coordinate Analysis (PCoA), Correspondence Analysis (CA), factor analysis, cluster analysis and canonical analysis are some of the statistical analyses that can be used to determine genetic diversity in plant species (Bhandari *et al.*, 2017).

Various software packages most of which are based on multivariate statistics such as Structure (Pritchard *et al.*, 2000), GeneAlEx (Genetic Analysis in Excel), (Peakall and Smouse, 2006), Arlequin (Excoffier and Lischer, 2010) and Discriminant Analysis of Principal Components (DAPC) (Jombart *et al.*, 2010) have also been used to determine the allelic diversity in plant species using individual parameters. The choice of each will depend on the data type, objectives and reproducibility (Pagnotta, 2018).

## 2.5 Methods for determining population structure in plants

Population structure is the distribution of the total amount of genetic variations in a population. Genetic variations within and among populations as well as their spatial arrangement are considered when describing the structure of a population. Clustering methods categorise and define individuals based on genetic relativeness (Chakraborty, 1993). Clustering is achieved by descriptive analysis and assigning of individuals in a population to groups based on genetic distances and similarity indices (Rokach and Maimon, 2005).

The four significant categories of clustering methods that have been applied in population genetics studies include partitional, hierarchical, overlapping, and ordination methods (Milligan and Cooper, 1987). The sequential agglomerative hierarchical clustering method is amongst the popular clustering algorithms. In this method, clustering starts with an individual assigned as a separate group. As the clustering continues, two clusters are combined and the result is one cluster bearing all the data (Rasmussen, 1992; Jombart *et al*., 2010).

Neighbour Joining (NJ) tree method is a sequential agglomerative hierarchical clustering analysis. NJ tree is a distance-based evolutionary method where the distance matrix from individuals within a population are used to build a phylogenetic tree (Saitou and Nei, 1987). In several studies distance matrices have been used to develop NJ tree to assess phylogenetic diversity of different plant species. In a study according to Xiong *et al*. (2022) the NJ tree constructed from genetic distances showed a clear pattern of segregation with four clades and four subclades. In this study, there was consistency between the species relationship and all the other accessions. In another study by Yang *et al*. (2016), the NJ tree was used to analyse the genetic relationships in 37 watermelon (*Citrullus lanatus*) genotypes. In this analysis there were three clusters, and the genotypes in each of the clusters were consistent with the place of origin.

Partitioning methods, also known as non-hierarchical clustering techniques, produce a single data partition. For instance, K-means algorithm identifies the K-number of clusters and then assigns each observation to the nearest mean while optimising homogeneity measurements within groups and heterogeneity between clusters (Natingga, 2017). Overlapping methods allow clusters to overlap and ordination techniques select a proportion of data to represent in an entire dataset. Hierarchical and non-hierarchical methods yield separate clusters that are non-overlapping (Milligan and Cooper, 1987).

Discriminant analysis of principle components (DAPC) is a non-hierarchical clustering method based on K-means algorithm that classifies and defines clusters of individuals that are genetically related. Different studies have used DAPC method to determine population structure. In a study by Deperi *et al*. (2018), the population structure of a tetraploid potato panel was determined using DAPC. From the analysis, five clusters were identified within the population. In a different study according to Fatokun *et al*. (2018), DAPC method was used to identify and describe the population structure of the world cowpea (*Vigna unguiculata* (L.) Walp.) germplasm collection. The results indicated that there were three distinct clusters in the population. In a study to assess the genetic diversity and population structure in White Yam (*Dioscorea rotundata* Poir.), DAPC was used to validate model-based admixture analysis. The DAPC clustering identified four groups that were in agreement with the groups identified in the phylogenetic tree (Bhattacharjee *et al*., 2020).

## 2.6 Population diversity studies in *U. Kirkiana*

Previous studies on *U. kirkiana* have been conducted to assess the genetic variability based on AFLP. In a study according to Mwase *et al*., (2007), analysis of genetic diversity in *U. kirkiana* obtained from Malawi revealed that there were three clusters of subpopulations. The genetic diversity was moderate with a value of $G_{ST}$ =0.079. Analysis of Molecular Variance (AMOVA) results showed that there was a high genetic density of 92% within populations and a lower genetic density of 6.8% among populations (Mwase *et al*., 2007). In a different study by Mwase *et al*. (2010), morphological characteristics and AFLPs were used to study the genetic variation in *U. kirkiana* samples collected from Zambia, Zimbabwe, Malawi and Tanzania. The degree of differentiation ranged between $F_{ST}$ = 0.002 and $F_{ST}$ = 0.259. There was a high genetic diversity within the individuals with a value of *H* mean = 0.256. AMOVA results indicated that there was a high genetic density of 90.8% within populations and a lower genetic density of 9.2% among populations (Mwase *et al*., 2010).

## 2.7 Determination of genetic differences using DArTseq

Diversity Arrays Technology Pty. Ltd (Canberra, Australia) are the proprietary owners of the DArT system. DArTseq is a high throughput sequencing approach that was developed by the company. Through the multiple staged process, large samples can be analysed at the same time to yield high-quality data (Kilian *et al*., 2012). Through DArTseq two types of markers namely SilicoDArTs and SNPs are generated. In the scoring of data SNP markers are codominant and are represented by 0 for homozygous reference allele, 2 for the homozygous alternate allele, 1 for heterozygous allele and (-) for the missing value. SilicoDArT markers are dominant and are represented by (1) for the presence and (0) absence of a restriction fragment. Calls with non-zero counts but too low counts to score confidently as (1) are represented by (-). Therefore, SilicoDArT markers are considered the equivalent of the AFLP markers (Sánchez-Sevilla *et al*., 2015).

DArTseq has been used to explore the genetic diversity and population structure of various plant species in several studies. In a study by Yang *et al*. (2016) DArTseq was used to determine the genetic diversity and population structure of core watermelon (*Citrullus lanatus*). The genetic diversity in the watermelon genotypes ranged between 0.03 and 0.5. The population was grouped into three distinct clusters that were correlated with their point of origin (Yang *et al*., 2016). In a study by Seyedimoradi *et al*., (2020), SilicoDArT markers obtained from DArTseq were used to determine the genetic diversity and population structure of chickpea (*Cicer arietinum* L.). In this study, the chickpea genotypes were found to have high genetic diversity and the population was divided into four distinct clusters (Seyedimoradi *et al*., 2020). From the study conducted by Adu *et al*., (2021) ,SilicoDArT and SNP markers from DArTseq were used for analysis of population structure and genetic diversity in cassava (*Manihot esculenta* Crantz). The genetic diversity was moderate and the population was clustered into two subpopulations with a lot of admixture (Adu *et al*., 2021).

### 2.8 The working principle of DArTseq

The markers used in Diversity Arrays Technology sequencing (DArTseq) are polymorphic parts of the genomic DNA. The markers are recognized in a differential hybridisation platform that has been designed for this process. These markers possess two observable alleles which are either dominant or co-dominant (Huttner *et al.*, 2005).

In the analysis of samples using the DArT technology, a discovery array is formed from a subset of genomes representative of the relevant genome. The pool of genomes collectively referred to as metagenome must undergo a level of complexity reduction to reduce repetitive DNA, which otherwise would affect DArT sequences and are of no significance to polymorphism. The discovery array then identifies polymorphic DArT markers grouped into a genotyping arrangement. Individual clones from the genomic representations are amplified and spotted onto glass slides to attain a discovery array. There are labelled genomic presentations of individual genomes earlier included in the metagenome pool; these are hybridized to discovery array and polymorphic clones known as DArT markers. The DArT markers detected, as a result, are placed into a genotyping array for routine genotyping work (Huttner *et al.*, 2005). DArT software is used to determine the intensities of hybridization during sequencing. The level of genetic diversity within the metagenome pool determines the efficacy of identifying DArT markers with polymorphism. The diversity arrays detect polymorphism through variations in single base-pair that occur at the sites of restriction endonucleases as well as deletions and rearrangements occurring within DNA fragments (Wittenberg *et al.,* 2007). Though DArT DNA polymorphisms, deletions and insertions can be detected, DArTseq assays unselected populations of fragments for quantifiable differences in hybridisation signal among input genotype samples (Huttner *et al.,* 2005).

## 2.9 Imputation of missing values in DArTseq data

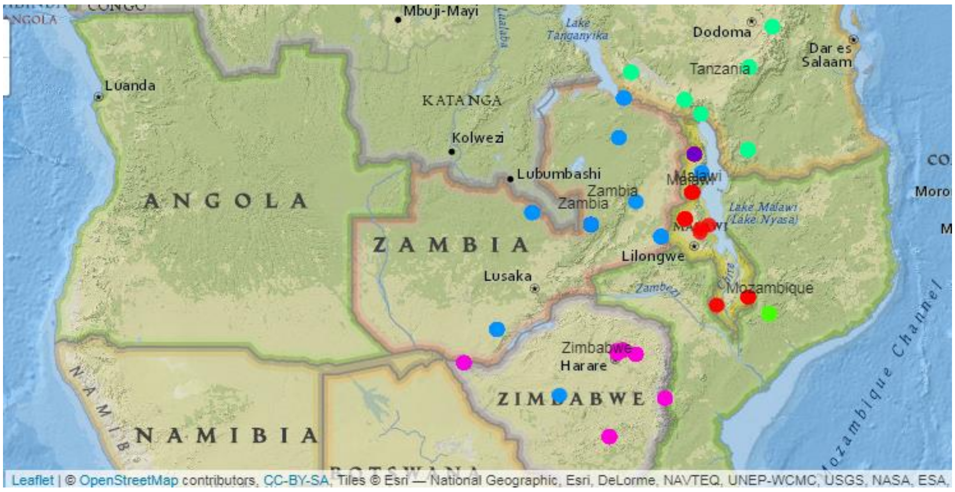
KDCompute is an online platform and an application developed by DArT PTY Ltd to analyse, impute and process sequence data without intensive computing power. Presence of missing data can lead to biasness and wrong conclusions in studies. As such, it is necessary to find a solution towards the issue of missing data (Hunter and Schmidt, 2004). Transcription, varying weather patterns, errors in measurement, damaged and dead plants are some of the causes of missing values. To compensate for the missing values, the missing data patterns must be established. These patterns help determine the imputation method (Negash, 2015). Expectation-Maximization (EM), Probabilistic Principal Component Analysis (PPCA), Singular Value Decomposition (SVD) and Nonlinear Estimation by Iterative Partial Least Squares (Nipals) are some of the imputation methods frequently used (John, *et al*., 2019). Small amounts of missing data are accommodated by Nipals but not more than 5% missing data. While SVD accepts high amount of missing data more than 10%, PPCA accommodates 10-15% of missing data (Stacklies *et al*., 2007). Expectation-Maximization imputation method uses an iterative algorithm to identify a parameter that utilizes the log likelihood when there are missing values (Dempster *et al*., 1977).

# CHAPTER THREE

# MATERIALS AND METHODS

## 3.1 Population samples and DNA extraction

Leaf material from 500 *Uapaca kirkiana* trees were randomly collected from International Centre For Research in Agroforestry (ICRAF) gene bank locations in Mozambique, Tanzania, Malawi, Zambia and Zimbabwe (Figure 3.1). From Mozambique leaves were obtained from 80 trees. From Tanzania, Malawi and Zimbabwe, leaves were collected from 100 trees in each country. From Zambia, leaves were plucked from 120 trees. The leaf samples were preserved in silica gel immediately after collection then shipped via courier to ICRAF headquarters in Nairobi for analysis. On arrival the collected leaves were sorted, some leaves had decomposed due to poor handling and storage and were discarded as they were no longer viable for DNA extraction. The remaining leaf samples (470) were then stored at -20°C till DNA extraction. The collected leaf sample size was determined by the amount of quality genomic DNA obtained for use in genotyping.



Leaflet | © OpenStreetMap contributors, CC-BY-SA, Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA,

**Figure 3.1:** Geographical distribution of 342 accessions *U. kirkiana* according to dartR results. Circles represent the sample location and colours indicate the country where the samples were obtained. Pink is for Zimbabwe, Blue is for Zambia, Red is for Malawi, Lime green is for Mozambique and Plain green is for Tanzania.

Total genomic DNA was extracted from 100mg of each leaf sample using the CTAB method according to Doyle and Doyle (1987). The leaves were obtained from old trees which tend to have high levels of polyphenols. Therefore, the CTAB protocol was modified to eliminate protein and secondary metabolites within sample and to obtain pure genomic DNA with high concentration (Porebski *et al.*,1997). Agarose gel electrophoresis (0.8%) was used to determine the approximate concentration and quality of the extracted DNA (Sambrook *et al.*, 1989). The purity of DNA was determined using Thermo Scientific™ NanoDrop 2000 spectrophotometer under the 260/280nm and 260/230 column. Any DNA that did not have the 260/280nm ratio ranging from 1.8 to 2.0 and 260/230 ration between 2.0 and 2.2 was discarded (Lucena-Aguilar *et al.*, 2016). The ability of the genomic DNA to amplify was determined using restriction enzymes digests after which the resulting fragments were analysed through PCR. This was necessary to eliminate contaminating nucleases and because DArT platform utilizes the same procedure in obtaining the restriction fragments. In the event that DNA could not be digested by restriction enzymes it was counted that the same would happen at DArT Pty Ltd during analysis. Thus, such DNA, degraded DNA, as well as those with short fragments, were eliminated. The final DNA concentration in nanograms per microliter (ng/µl) required for the analysis was measured using Thermo Scientific™ Qubit Fluorometer. Diversity arrays technology Pty Ltd requires at least 50 ng/µl of DNA for sequencing. Therefore, DNA concentration of less than 50 ng/µl was disregarded (Baloch *et al.*, 2017).

### 3.2 DNA normalisation, library preparation, and sequencing

DNA samples of 20 µl and a concentration of 50-100 ng/µl were loaded onto four 96 well plates and sent to Diversity Arrays Technology Pty. Ltd (Canberra, Australia) for analysis. The DArT sequencing steps are described in detail by Kilian *et al*. (2012) and DArT Pty Ltd at www.diversityarrays.com. Here is an outline of the process: the first step involved digestion of genomic DNA with a mixture of restriction enzymes to allow selection of a part of the genome depending on outlined criteria for instance size. As a result, polymorphic fragments that were relevant in the analysis of genetic diversity were selected. The polymorphic fragments were then used to create a library by cloning them into the *Escherichia coli* bacteria. The process was followed by polymerase chain-reaction (PCR) which amplified the generated libraries. Amplicon cleaning and evaluation through capillary electrophoresis was done and the resulting fragments were sequenced creating a FASTQ file with sequence reads of polymorphic fragments. Since there was reference no genome for *U. kirkiana*, the process was repeated to include different reads from the library. SilicoDArT and SNP markers were then identified based on different algorithms and the resulting data was heterozygous and homozygous (Kilian *et al*., 2012).

## 3.3 Data management and statistical analysis

### 3.3.1 Processing of raw data and SNP calling

Initial data processing was done at DArT Pty Ltd, Australia using the DArTsoft.v.7.4.7 to analyse images and score SilicoDArT and SNP markers. The data presented from DArT Pty Ltd, Australia, was in two formats: SNPs and SilicoDArT data sets in comma-separated values format. The data sets contained the parameters explaining the quality of the markers. The Polymorphic Information Content (PIC), call rate and reproducibility were used to explain allelic diversity. The sequences obtained from DArT Pty Ltd were filtered for insignificant markers and genotypes when generating the SNPs and SilicoDArTs. As a result, out of the 376 DNA samples that were sent to DArT Pty Ltd for DArTseq, only 342 were reported. A table for the list of samples that were analysed and reported by DArT Pty Ltd is shown (Table 3.1). Assigned identities were the sample names used during analysis of data. Provenance was the original collection point of *U. kirkiana* and laboratory identities were the names assigned to the samples for sequencing.

**Table 3.1:** Laboratory identity, assigned identity, provenance and country of origin of *U. kirkiana* from Miombo woodland used in this study

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **2** | 1 | Nyamukwarara | Zimbabwe |
| **4** | 2 | Mbala | Zambia |
| **7** | 3 | Lwilomelo | Zimbabwe |
| **10** | 4 | Choma | Zambia |
| **11** | 5 | Mapanzure | Zimbabwe |
| **16** | 6 | Musana | Zimbabwe |
| **17** | 7 | Nyamukwarara | Zimbabwe |
| **18** | 8 | Choma | Zambia |
| **19** | 9 | Lwilomelo | Zimbabwe |
| **22** | 10 | Mbala | Zambia |
| **23** | 11 | Choma | Zambia |
| **28** | 12 | Serenje | Zambia |
| **29** | 13 | Lwilomelo | Zimbabwe |
| **32** | 14 | Musana | Zimbabwe |
| **33** | 15 | Mapanzure | Zimbabwe |
| **34** | 16 | Litende | Malawi |
| **35** | 17 | Choma | Zambia |
| **36** | 18 | Lwilomelo | Zimbabwe |
| **39** | 19 | Mbala | Zambia |
| **40** | 20 | Musana | Zimbabwe |
| **41** | 21 | Murewa | Malawi |
| **42** | 22 | Domboshawa | Zimbabwe |
| **43** | 23 | Murewa | Zimbabwe |
| **44** | 24 | Mapanzure | Zimbabwe |
| **45** | 25 | Luwawa | Malawi |
| **46** | 26 | Luwawa | Malawi |
| **47** | 27 | Nyamukwarara | Zimbabwe |
| **48** | 28 | Choma | Zambia |
| **49** | 29 | Lwilomelo | Zimbabwe |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **51** | 30 | Lwilomelo | Zimbabwe |
| **53** | 31 | Litende | Malawi |
| **55** | 32 | Mapanzure | Zimbabwe |
| **56** | 33 | Musana | Zimbabwe |
| **59** | 34 | Luwawa | Malawi |
| **60** | 35 | Gombea | Tanzania |
| **61** | 36 | Musana | Zimbabwe |
| **63** | 37 | Musana | Zimbabwe |
| **64** | 38 | Mbala | Zambia |
| **65** | 39 | Litende | Malawi |
| **66** | 40 | Domboshawa | Zimbabwe |
| **67** | 41 | Mbala | Zambia |
| **68** | 42 | Lwilomelo | Zimbabwe |
| **69** | 43 | Choma | Zambia |
| **70** | 44 | Mbala | Zambia |
| **71** | 45 | Musana | Zimbabwe |
| **74** | 46 | Mbala | Zambia |
| **78** | 47 | Luwawa | Malawi |
| **79** | 48 | Lwilomelo | Zimbabwe |
| **80** | 49 | Luwawa | Zambia |
| **81** | 50 | Mbala | Zambia |
| **82** | 51 | Domboshawa | Zimbabwe |
| **83** | 52 | Lwilomelo | Zimbabwe |
| **84** | 53 | Litende | Malawi |
| **86** | 54 | Murewa | Malawi |
| **89** | 55 | Nyamukwarara | Zimbabwe |
| **91** | 56 | Choma | Zambia |
| **94** | 57 | Litende | Malawi |
| **95** | 58 | Luwawa | Malawi |
| **96** | 59 | Litende | Malawi |
| **97** | 60 | Gombela | Tanzania |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **98** | 61 | Lwilomelo | Zimbabwe |
| **100** | 62 | Litende | Malawi |
| **101** | 63 | Kitwe | Zambia |
| **102** | 64 | Iringa | Tanzania |
| **103** | 65 | Kasama | Zambia |
| **104** | 66 | Choma | Zambia |
| **105** | 67 | Gombela | Tanzania |
| **106** | 68 | Chipata | Zambia |
| **107** | 69 | Choma | Zambia |
| **108** | 70 | Chipata | Zambia |
| **109** | 71 | Iringa | Tanzania |
| **110** | 72 | Mbeya | Tanzania |
| **111** | 73 | Mbeya | Tanzania |
| **112** | 74 | Chipata | Zambia |
| **113** | 75 | Chipata | Zambia |
| **114** | 76 | Mbeya | Tanzania |
| **116** | 77 | MUAP | Malawi |
| **117** | 78 | Kasama | Zambia |
| **119** | 79 | Kasama | Zambia |
| **121** | 80 | Mpwapwa | Tanzania |
| **122** | 81 | Chipata | Zambia |
| **123** | 82 | Mbeya | Tanzania |
| **124** | 83 | Serenje | Zambia |
| **125** | 84 | Gombela | Tanzania |
| **126** | 85 | Kyela | Tanzania |
| **127** | 86 | Mbala | Zimbabwe |
| **128** | 87 | Gombela | Tanzania |
| **129** | 88 | Iringa | Tanzania |
| **130** | 89 | Iringa | Tanzania |
| **132** | 90 | Kitwe | Zambia |
| **134** | 91 | Gombela | Tanzania |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **136** | 92 | Kitwe | Zambia |
| **137** | 93 | Serenje | Zambia |
| **138** | 94 | Mpwapwa | Tanzania |
| **139** | 95 | Serenje | Zambia |
| **140** | 96 | Iringa | Tanzania |
| **142** | 97 | Kasama | Zambia |
| **144** | 98 | Kitwe | Zambia |
| **145** | 99 | Gombela | Tanzania |
| **146** | 100 | Zambia | Zambia |
| **147** | 101 | Sumbawanga | Tanzania |
| **148** | 102 | Kasama | Zambia |
| **149** | 103 | Mpwapwa | Tanzania |
| **150** | 104 | Mbala | Zambia |
| **151** | 105 | Iringa | Tanzania |
| **152** | 106 | Choma | Zambia |
| **153** | 107 | Kasama | Zambia |
| **154** | 108 | MUAP | Malawi |
| **155** | 109 | Mbala | Zambia |
| **156** | 110 | Kasama | Zambia |
| **158** | 111 | Litende | Malawi |
| **159** | 112 | Luwawa | Malawi |
| **161** | 113 | Luwawa | Malawi |
| **162** | 114 | Luwawa | Malawi |
| **163** | 115 | Litende | Malawi |
| **164** | 116 | Litende | Malawi |
| **166** | 117 | Litende | Malawi |
| **167** | 118 | Litende | Malawi |
| **168** | 119 | Kasungu-MUAP | Malawi |
| **169** | 120 | Luwawa | Malawi |
| **170** | 121 | Sumbawanga | Tanzania |
| **171** | 122 | Litende | Malawi |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **172** | 123 | Gombela | Tanzania |
| **174** | 124 | Mozambique | Mozambique |
| **175** | 125 | Kyela | Tanzania |
| **176** | 126 | Choma | Zambia |
| **177** | 127 | Serenje | Zambia |
| **178** | 128 | Choma | Zambia |
| **179** | 129 | Mbeya | Tanzania |
| **180** | 130 | Zambia | Zambia |
| **186** | 131 | Mozambique | Mozambique |
| **187** | 132 | Zambia | Zambia |
| **189** | 133 | Chipata | Zambia |
| **190** | 134 | Kitwe | Zambia |
| **192** | 135 | Kitwe | Zambia |
| **193** | 136 | Serenje | Zambia |
| **195** | 137 | Zambia | Zambia |
| **196** | 138 | Choma | Zambia |
| **197** | 139 | Kitwe | Zambia |
| **198** | 140 | Mozambique | Mozambique |
| **199** | 141 | Iringa | Tanzania |
| **200** | 142 | Sumbawanga | Tanzania |
| **202** | 143 | Kyela | Tanzania |
| **203** | 144 | Iringa | Tanzania |
| **204** | 145 | Mozambique | Mozambique |
| **205** | 146 | Mozambique | Mozambique |
| **206** | 147 | Serenje | Zambia |
| **207** | 148 | Mozambique | Mozambique |
| **210** | 149 | Mpwapwa | Tanzania |
| **212** | 150 | Mozambique | Mozambique |
| **213** | 151 | Mozambique | Mozambique |
| **103AA** | 152 | Kasama | Zambia |
| **104AA** | 153 | Choma | Zambia |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **105AA** | 154 | Gombela | Tanzania |
| **112AA** | 155 | Chipata | Zambia |
| **114AA** | 156 | Mbeya | Tanzania |
| **150AA** | 157 | Mbala | Zambia |
| **29AA** | 158 | Lwilomelo | Zimbabwe |
| **39AA** | 159 | Mbala | Zambia |
| **48AA** | 160 | Choma | Zambia |
| **51AA** | 161 | Lwilomelo | Zimbabwe |
| **55AA** | 162 | Mapanzure | Zimbabwe |
| **63AA** | 163 | Musana | Zimbabwe |
| **74AA** | 164 | Sumbawanga | Tanzania |
| **78AA** | 165 | Luwawa | Malawi |
| **82AA** | 166 | Domboshawa | Zimbabwe |
| **86AA** | 167 | Murewa | Zimbabwe |
| **89AA** | 168 | Nyamukwarara | Zimbabwe |
| **91AA** | 169 | Choma | Zambia |
| **95AA** | 170 | Luwawa | Malawi |
| **K107** | 171 | Kasungu | Malawi |
| **K117** | 172 | Phalombe | Malawi |
| **K13** | 173 | Luwawa | Malawi |
| **K136** | 174 | Mpwapwa | Tanzania |
| **K25** | 175 | Kasungu | Malawi |
| **K35** | 176 | Litende | Malawi |
| **K45** | 177 | Phalombe | Malawi |
| **K47** | 178 | Phalombe | Malawi |
| **K50** | 179 | Litende | Malawi |
| **K52** | 180 | Phalombe | Malawi |
| **K72** | 181 | Choma | Zambia |
| **K79** | 182 | Kasungu | Malawi |
| **K81** | 183 | Kasungu | Malawi |
| **K87** | 184 | Litende | Malawi |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **K93** | 185 | Kasungu | Malawi |
| **K98** | 186 | Serenje | Zambia |
| **M1** | 187 | Serenje | Zambia |
| **M10** | 188 | Choma | Zambia |
| **M101** | 189 | Mbeya | Tanzania |
| **M102** | 190 | Kasungu | Malawi |
| **M103** | 191 | Mbeya-Nyoka | Tanzania |
| **M104** | 192 | Mbeya-Nyoka | Tanzania |
| **M105** | 193 | Mbeya-Nyoka | Tanzania |
| **M106** | 194 | Gombela-Songea | Tanzania |
| **M108** | 195 | Sumbawanga | Tanzania |
| **M109** | 196 | Luwawa | Malawi |
| **M11** | 197 | Serenje | Zambia |
| **M111** | 198 | Murewa | Malawi |
| **M112** | 199 | Luwawa | Malawi |
| **M113** | 200 | Iringa | Tanzania |
| **M114** | 201 | Chipata | Zambia |
| **M115** | 202 | Mapanzure | Zimbabwe |
| **M116** | 203 | Gombela-Songea | Tanzania |
| **M117** | 204 | Mbeya-Nyoka | Tanzania |
| **M118** | 205 | Sumbawanga | Tanzania |
| **M12** | 206 | Mapanzure | Zimbabwe |
| **M120** | 207 | Kasungu | Malawi |
| **M13** | 208 | Murewa | Malawi |
| **M14** | 209 | Kasungu | Malawi |
| **M15** | 210 | Serenje | Zambia |
| **M16** | 211 | Mbeya | Tanzania |
| **M17** | 212 | Chipata | Zambia |
| **M18** | 213 | Kasungu | Malawi |
| **M19** | 214 | Kasungu | Malawi |
| **M2** | 215 | Mbeya-Kyela | Tanzania |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **M20** | 216 | Phalombe | Malawi |
| **M21** | 217 | Mbeya-Kyela | Tanzania |
| **M22** | 218 | Serenje | Zambia |
| **M23** | 219 | Choma | Zambia |
| **M24** | 220 | Murewa | Malawi |
| **M25** | 221 | Mapanzure | Zimbabwe |
| **M29** | 222 | Mapanzure | Zimbabwe |
| **M3** | 223 | Mapanzure | Zimbabwe |
| **M30** | 224 | Kyela | Tanzania |
| **M31** | 225 | Kasungu | Malawi |
| **M34** | 226 | Litende | Malawi |
| **M4** | 227 | Murewa | Malawi |
| **M5** | 228 | Serenje | Zambia |
| **M52** | 229 | Kasungu | Malawi |
| **M53** | 230 | Mapanzure | Zimbabwe |
| **M54** | 231 | Choma | Zambia |
| **M55** | 232 | Chipata | Zambia |
| **M56** | 233 | Kasungu | Malawi |
| **M58** | 234 | Murewa | Zimbabwe |
| **M59** | 235 | Luwawa | Malawi |
| **M6** | 236 | Nyamukwarara | Zimbabwe |
| **M60** | 237 | Mpwapwa | Tanzania |
| **M62** | 238 | Utete-Iringa | Tanzania |
| **M66** | 239 | Gombela-Songea | Tanzania |
| **M67** | 240 | Murewa | Zimbabwe |
| **M69** | 241 | Chipata | Zambia |
| **M7** | 242 | Murewa | Zimbabwe |
| **M70** | 243 | Sumbawanga | Tanzania |
| **M72** | 244 | Gombela-Songea | Tanzania |
| **M73** | 245 | Murewa | Zimbabwe |
| **M74** | 246 | Sumbawanga | Malawi |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **M75** | 247 | Choma | Zambia |
| **M77** | 248 | Mpwapwa | Tanzania |
| **M79** | 249 | Mpwapwa | Tanzania |
| **M8** | 250 | Nyamukwarara | Zimbabwe |
| **M80** | 252 | Gombela-Songea | Tanzania |
| **M82** | 252 | Utete-Iringa | Tanzania |
| **M83** | 253 | Mbeya-Kyela | Tanzania |
| **M85** | 254 | Mapanzure | Zimbabwe |
| **M86** | 255 | Serenje | Malawi |
| **M87** | 256 | Litende | Malawi |
| **M89** | 257 | Kasungu | Malawi |
| **M9** | 258 | Luwawa | Malawi |
| **M92** | 259 | Mbeya-Nyoka | Tanzania |
| **M92AA** | 260 | Mbeya-Nyoka | Tanzania |
| **M94** | 261 | Choma | Zambia |
| **M95** | 262 | Mbeya-Nyoka | Tanzania |
| **M97** | 263 | Litende | Malawi |
| **M99** | 264 | Choma | Zambia |
| **UK-M1** | 265 | Chipata | Zambia |
| **UK-M103** | 266 | Nyamukwarara | Zimbabwe |
| **UK-M104** | 267 | Nyamukwarara | Zimbabwe |
| **UK-M109** | 268 | Phalombe | Malawi |
| **UK-M110** | 269 | Phalombe | Malawi |
| **UK-M112** | 270 | Phalombe | Malawi |
| **UK-M117** | 271 | Serenje | Zambia |
| **UK-M118** | 272 | Serenje | Zambia |
| **UK-M119** | 273 | Serenje | Zambia |
| **UK-M124** | 274 | Sumbawanga | Tanzania |
| **UK-M125** | 275 | Sumbawanga | Tanzania |
| **UK-M126** | 276 | Sumbawanga | Tanzania |
| **UK-M127** | 277 | Sumbawanga | Tanzania |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **UK-M133** | 278 | Utete-Iringa | Tanzania |
| **UK-M136** | 279 | Utete-Iringa | Tanzania |
| **UK-M14** | 280 | Choma | Zambia |
| **UK-M2** | 281 | Chipata | Zambia |
| **UK-M20** | 282 | Gombela-Songea | Tanzania |
| **UK-M25** | 283 | Gombela-Songea | Tanzania |
| **UK-M26** | 284 | Gombela-Songea | Tanzania |
| **UK-M29** | 285 | Kasungu | Malawi |
| **UK-M3** | 286 | Chipata | Zambia |
| **UK-M31** | 287 | Kasungu | Malawi |
| **UK-M34** | 288 | Kasungu | Malawi |
| **UK-M37** | 289 | Kasungu | Malawi |
| **UK-M4** | 290 | Chipata | Zambia |
| **UK-M42** | 291 | Litende | Malawi |
| **UK-M44** | 292 | Litende | Malawi |
| **UK-M48** | 293 | Litende | Malawi |
| **UK-M51** | 294 | Luwawa | Malawi |
| **UK-M56** | 295 | Luwawa | Malawi |
| **UK-M58** | 296 | Mapanzure | Zimbabwe |
| **UK-M6** | 297 | Chipata | Zambia |
| **UK-M60** | 298 | Mapanzure | Zimbabwe |
| **UK-M61** | 299 | Mapanzure | Zimbabwe |
| **UK-M66** | 300 | Mapanzure | Zimbabwe |
| **UK-M7** | 301 | Chipata | Zambia |
| **UK-M71** | 302 | Kyela | Tanzania |
| **UK-M78** | 303 | Kyela | Tanzania |
| **UK-M8** | 304 | Chipata | Zambia |
| **UK-M81** | 305 | Mbeya-Nyoka | Tanzania |
| **UK-M84** | 306 | Mbeya-Nyoka | Tanzania |
| **UK-M85** | 307 | Mbeya-Nyoka | Tanzania |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **UK-M86** | 308 | Mpwapwa | Tanzania |
| **UK-M90** | 309 | Mpwapwa | Tanzania |
| **UK-M91** | 310 | Murewa | Zimbabwe |
| **UK-M92** | 311 | Murewa | Zimbabwe |
| **UK-M96** | 312 | Murewa | Zimbabwe |
| **UK-M97** | 313 | Murewa | Zimbabwe |
| **UK-Z1** | 314 | Chipata | Zambia |
| **UK-Z12** | 315 | Choma | Zambia |
| **UK-Z13** | 316 | Choma | Zambia |
| **UK-Z17** | 317 | Iringa | Tanzania |
| **UK-Z18** | 318 | Iringa | Tanzania |
| **UK-Z20** | 319 | Iringa | Tanzania |
| **UK-Z23** | 320 | Iringa | Tanzania |
| **UK-Z26** | 321 | Kasama | Zambia |
| **UK-Z27** | 322 | Kasungu | Malawi |
| **UK-Z36** | 323 | Kyela | Tanzania |
| **UK-Z37** | 324 | Kyela | Tanzania |
| **UK-Z38** | 325 | Litende | Malawi |
| **UK-Z42** | 326 | Luwawa | Malawi |
| **UK-Z43** | 327 | Luwawa | Malawi |
| **UK-Z49** | 328 | Mbala | Zambia |
| **UK-Z54** | 329 | Mpwapwa | Tanzania |
| **UK-Z55** | 330 | Mpwapwa | Tanzania |
| **UK-Z56** | 331 | Mpwapwa | Tanzania |
| **UK-Z60** | 332 | MUAP | Zambia |
| **UK-Z64** | 333 | Serenje | Zambia |
| **UK-Z66** | 334 | Serenje | Zambia |
| **UK-Z67** | 335 | Serenje | Zambia |
| **UK-Z68** | 336 | Serenje | Zambia |
| **UK-Z69** | 337 | Serenje | Zambia |
| **UK-Z74** | 338 | Zambia | Zambia |

| LAB ID | ASSIGNED ID | PROVENANCE | COUNTRY OF ORIGIN |
|---|---|---|---|
| **UK-Z77** | 339 | Zambia | Zambia |
| **UK-Z78** | 340 | Zambia | Zambia |
| **UK-Z79** | 341 | Zambia | Zambia |
| **UK-Z86** | 342 | Luwawa | Zambia |

Missing data was imputed on the KDCompute server at (https://kdcompute.igss-africa.org/kdcompute/login). The analysis was based on four imputation methods namely Expectation-Maximization (EM), Probabilistic Principal Component Analysis (PPCA), Singular Value Decomposition (SVD) and Nonlinear iterative partial least squares (Nipals). Each imputation method was run on the dataset with an additional 10% introduced missing values. The imputed introduced missing values were then compared to the original data set to calculate a Simple Matching Coefficient (SMC). The method with the highest SMC method was then used to impute the original data set.

### 3.3.2 Statistical analysis

DartR (Gruber *et al.*, 2018) R software package (R Core Team, 2017), was used to convert the dataset to distance matrices with 1000 bootstraps (Gruber *et al.*, 2018). Through the g2phylip() function and 1000 bootstrap replicate, a matrix of genetic distances between subpopulations was calculated to produce an input file for Phylip application (Felsenstein, 2005). Phylip application was then used to derive a neighbour-joining (NJ) dendrogram which was visualized on the interactive Tree of Life (iTOL) application (Letunic and Bork, 2019). The genetic diversity was calculated using the basic.stat() function in dartR package (Gruber *et al.*, 2018) and AMOVA was determined using poppr in R. Population structure was determined using discriminant analysis of principal components (DAPC) method (Jombart *et al*., 2010) in the *adegenet* package (Jombart, 2008) for R software (R Core Team, 2017). To identify the clusters in the dataset, find.clusters () function in DAPC was used (Jombart *et al*., 2010). A specific maximum value of K=40 groups which is equivalent to max.n.clust=40, was applied. While retaining the maximum number of all the Principle Components (PCs), a graph of Bayesian Information Criterion (BIC) values against cumulative values of K was plotted. Table.value() function was used to verify whether all the actual subpopulations were retrieved by the method. The results obtained from the discriminant analysis principle component were plotted using the scatterplot() function to include the retained eigenvalues principal component analysis. DAPC summary heatmaps of the first 50 individuals and all individuals in the dataset were then plotted. The results of the heatmap were drawn in a composite plot using the function compoplot(). The DAPC was cross-validated using xvalDapc() function.

# CHAPTER FOUR

# RESULTS

## 4.1 To determine genetic diversity and population genetic parameters of *U. kirkiana* (Müell) Arg. from selected locations in Africa

### 4.1.1 Evaluation of allelic diversity

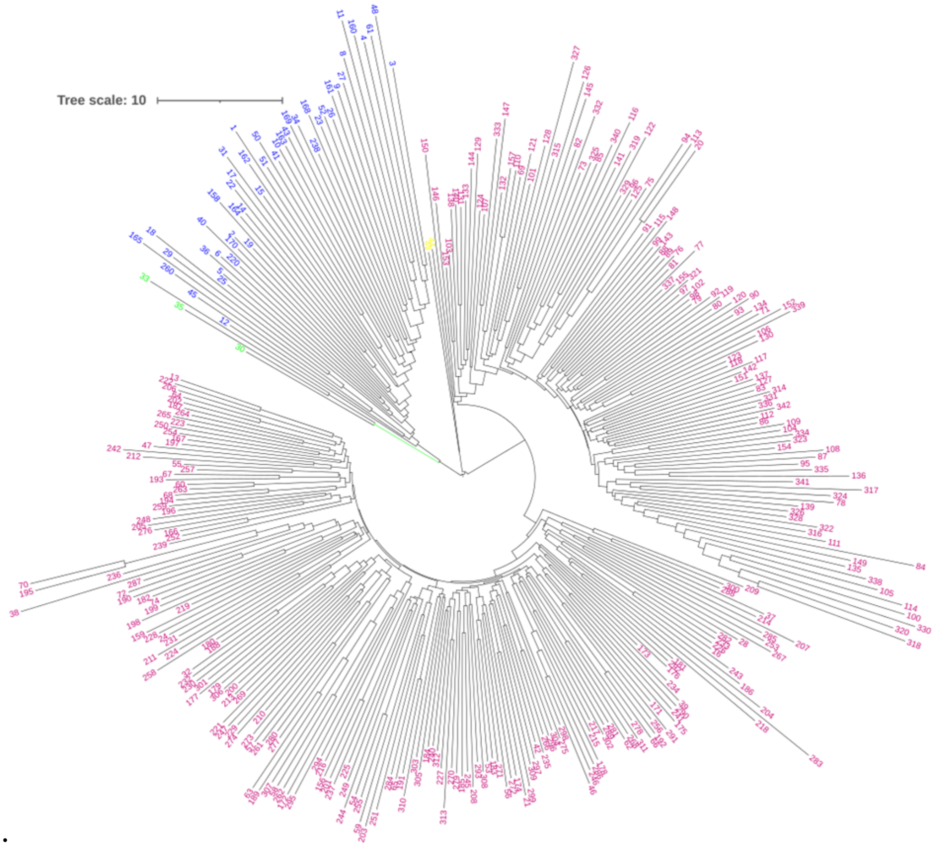
DArTseq generated codominant 28393 SNPs in 342 accessions of *Uapaca kirkiana* obtained from various locations in Malawi, Mozambique, Tanzania, Zambia, and Zimbabwe. The mean average call rate was 50%. The overall average of the polymorphism information content (PIC) of the reference and SNP allele was 0.1, with values ranging between 0.003 and 0.5. The average reproducibility rate, which is the proportion of technical replicate assay pairs for which the marker score was consistent, was 99.95%. Out of the four imputation algorithms EM, PPCA, SVD and Nipals; EM, PPCA and SVD had a simple matching coefficient (SMC) of 0.56, with Nipals having the lowest SMC of 0.53 (Table 4.1)

**Table 4.1:** Imputation report for compensating missing values in the SNP data for *U.kirkiana*

| Imputation Methods | Timings (minutes) | Scores |
|---|---|---|
| **EM** | 23.581 | 0.5637605 |
| **PPCA** | 4.160 | 0.5577847 |
| **SVD** | 15.050 | 0.5550635 |
| **Nipals** | 2.370 | 0.5299328 |

### 4.1.2 Assessment of phylogenetic relations

The neighbour-joining tree analysis classified the accessions into four clusters (Figure 4.1 and Figure 4.2). The first cluster was composed of a total of 3 individuals (subgroup 1) from Zimbabwe and Tanzania. Cluster two contained 47 individuals (subgroup 2) from Zimbabwe, Zambia, Malawi and Tanzania. Cluster three contained 2 individuals (subgroup 3) from Zambia. The fourth cluster consisted of 289 individuals (subgroup 4) from Zimbabwe, Malawi, Zambia, Tanzania and Mozambique (Figure 4.1 and Figure 4.2). During the construction of the NJ tree, one individual (7) an individual from Zimbabwe formed an outlier and therefore was deleted from the tree.



.

**Figure 4.1:** Neighbour joining tree created from 1000 bootstrap replicates for 341 *U. kirkiana* accessions. Based on the NJ analysis, there were four clusters that did not correspond to the geographical location of the plant. Cluster 1 (green colour) had 3 individuals, Cluster 2 (yellow colour) had 2 individuals, Cluster 3 (Fuchsia colour) had 289 individuals and cluster 4 had 47 individuals.

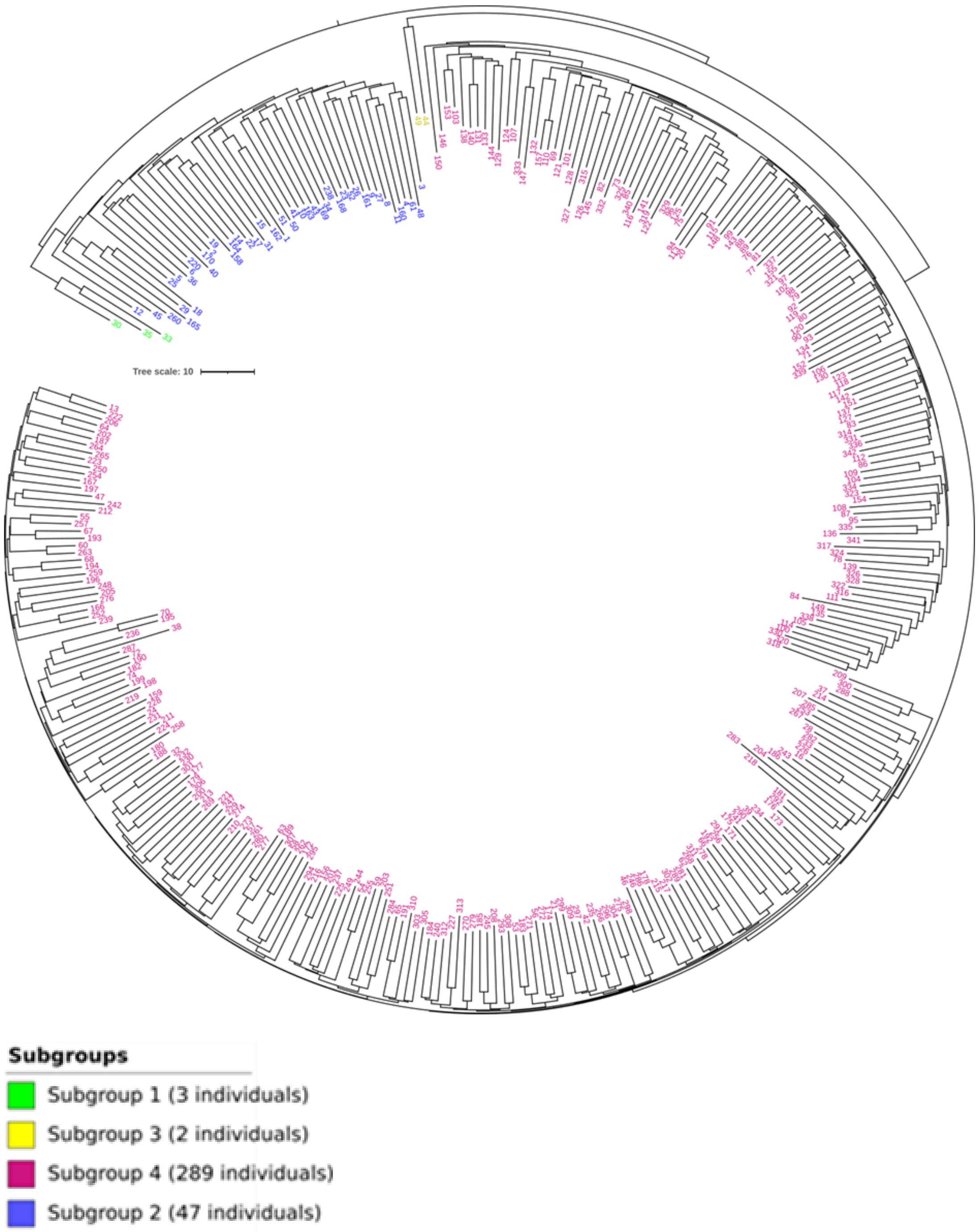

Tree scale: 10
Subgroups
🟩 Subgroup 1 (3 individuals)
🟨 Subgroup 3 (2 individuals)
🟥 Subgroup 4 (289 individuals)
🟦 Subgroup 2 (47 individuals)

**Figure 4.2:** Inverted Neighbour joining tree created from 1000 bootstrap replicates for 341 *U. kirkiana* accessions. Based on the NJ analysis, there were four clusters that did not correspond to the geographical location of the plant. Cluster 1 (green colour) had 3 individuals, Cluster 2 (yellow colour) had 2 individuals, Cluster 3 (Fuchsia colour) had 289 individuals and cluster 4 had 47 individuals.

### 4.1.3 Determination of population genetics parameters

The heterozygosity (Ho) within populations was 0.0849. The fixation index (Fst) was used to determine deviations from the Hardy Weinberg equilibrium. The overall Fst within the populations was 0.0551, and the corrected fixation index (Fstp) was equivalent to Fst. The overall gene diversity (Ht) was 0.1040, which was comparable to the genetic diversity within populations (Hs) but higher than the gene diversity among samples (Dst) that had a value of 0.0057. Even so, corrected overall genetic diversity (Htp) was the same as the overall gene diversity and the corrected gene diversity among samples (Dstp) was comparable to the gene diversity among samples. The inbreeding coefficient per overall loci (Fis) was 0.1364 whereas the measure of population differentiation (Dest) was 0.0065 (Table 4.2).

**Table 4.2:** Genetic diversity between and among populations of *Uapaca kirkiana*

| **Ho** | **Hs** | **Ht** | **Dst** | **Htp** | **Dstp** | **Fst** | **Fstp** | **Fis** | **Dest** |
|---|---|---|---|---|---|---|---|---|---|
| **0.0849** | 0.0983 | 0.1040 | 0.0057 | 0.1041 | 0.0059 | 0.0551 | 0.0565 | 0.1364 | 0.0065 |

Legend: Ho: heterozygosity within populations, Hs: genetic diversity within populations, Ht: overall gene diversity; Dst: gene diversity among samples; Htp: corrected overall gene diversity; Dstp: corrected gene diversity among samples; Fst: fixation index; Fstp: corrected fixation index; Fis: inbreeding coefficient per overall loci; Dest: a measure of population differentiation.

## 4.2 To determine genetic relationships and population structure of *U. kirkiana* (Müell) Arg. from selected locations in Africa

### 4.2.1 Determination of genetic relationships in populations

Analysis of molecular variance (AMOVA) is used to test whether there is significant population structure or not. In this study, there was significant diversity among samples (P>0.001). The distribution of genetic diversity was 1.3% between populations, 4.8% between samples within populations and 93.9% within samples. The population differentiation statistics was 0.05 between samples within populations, 0.06 within samples, and 0.01 overall. The analysis of molecular variance (AMOVA) results revealed a high genetic density within samples and a lower genetic density of between populations (Table 4.3).

**Table 4.3:** Analysis of molecular variance in populations of *U. kirkiana*. Estimation of *P*-value was based on 999 permutations. Legends: DF, degrees of freedom, SSD, sum of squared deviations; MSD, mean squared deviation.

| AMOVA | DF | SSD | MSD | Sigma | % | Statphi | P |
|---|---|---|---|---|---|---|---|
| **Between populations** | 4 | 2074.49 | 518.62 | 4.39 | 1.28 | | 0.001 |
| **Between samples within populations** | 91 | 32208.84 | 353.94 | 16.53 | 4.84 | 0.05 | 0.001 |
| **Within samples** | 96 | 30804.89 | 320.88 | 320.88 | 93.88 | 0.06 | 0.001 |
| **Total** | 191 | 65088.22 | 340.78 | 341.80 | 100.0 | 0.01 | |

### 4.2.2 Discriminant Analysis of Principal Components (DAPC)

From the find.clusters () function a graph of cumulative variance due to the PCA eigenvalues was generated. The graph was used to determine the number of principle components (PCs) to retain for use in the step of analysis (Figure 4.3).

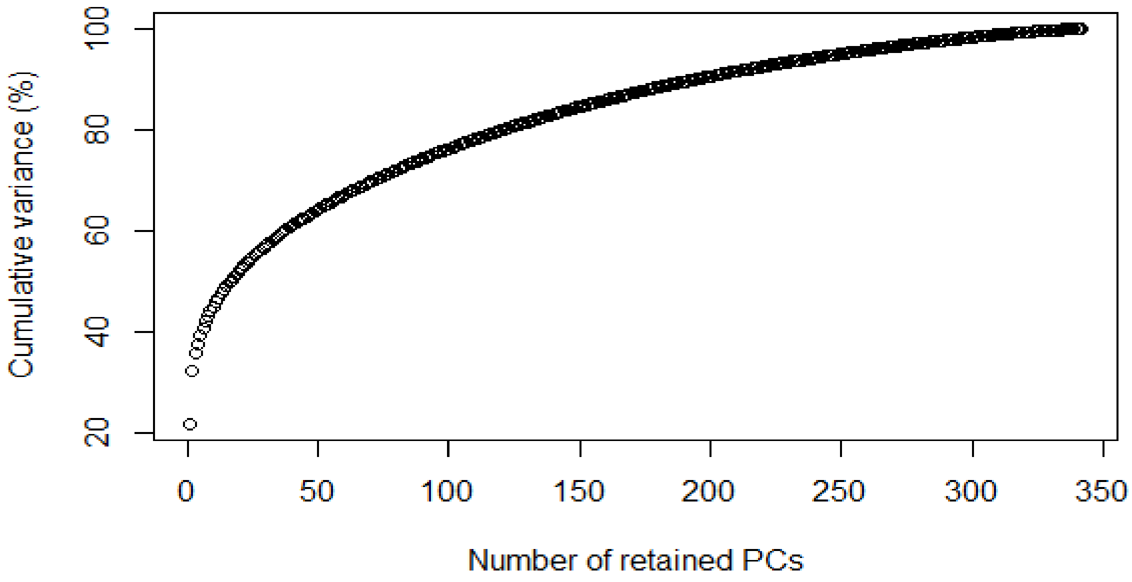


**Figure 4.3**: A graph of cumulative variance due to the PCA eigenvalues for DAPC

The number of PCs retained in the first step (325) were used to plot a graph of BIC against K. This graph was used to establish the values of K. From the figure (Figure 4.4), there was a decrease of BIC values up to K=4 which led to the narrowing down on the number of K values to use in the analysis as 4.

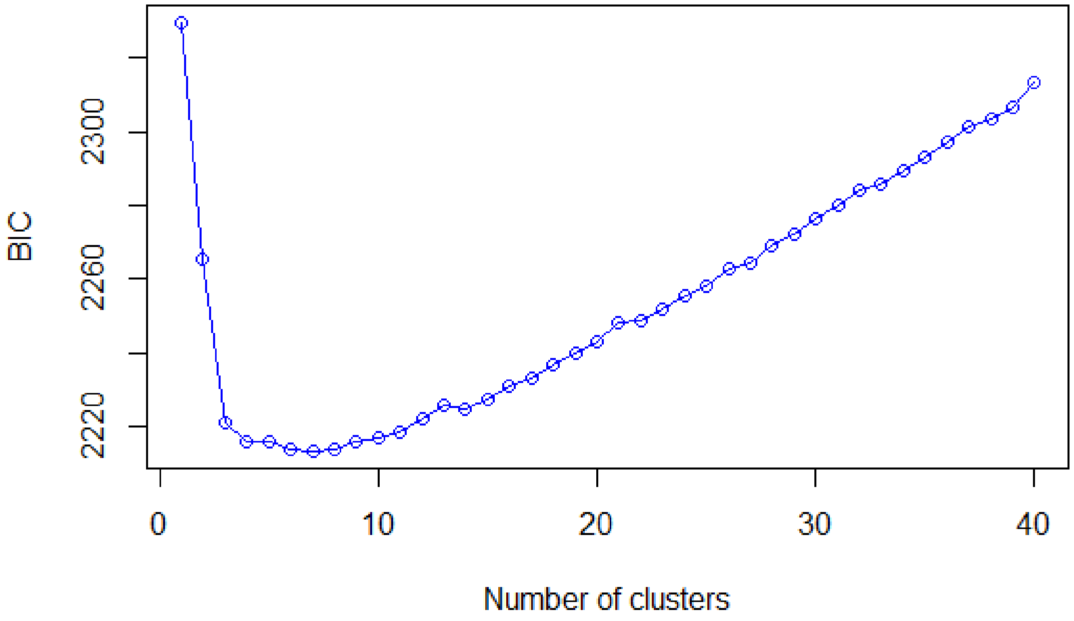


**Figure 4.4:** Value of BIC versus number of clusters for identifying K-values for DAPC

At the point of decrease of BIC values in the second graph, the curve formed an elbow suggesting that four clusters should be retained. The output from find.clusters() consisted of a list of statistics, including summary statistics (Kstat) of K=1 to K=40. The Kstat values ranged from 2214.635 to 2311.317. The number of clusters identified as well as the associated statistic were also included in the find.clusters() output. Consequently, K=4 was listed and the Kstat value was found to be 2215.827. Also, the assignment of all the individuals in the metapopulation to levels1 to 4 was shown. Each level had varying individual sizes: level 1 had a total of 173 individuals, level 2-52, level 3-102 and level 4-15 individuals. The actual groups in the subpopulations were all well identified by the method where the actual subpopulations were 341, and the inferred groups were 4. The output for find.clusters was as follows: Kstat, stat, population levels and size A graph similar to find.clusters() function was obtained from the dapc() function. Unlike K-means, where too many PCs were profitable, a few PCs were retained in dapc(). The number of PCs to retain were selected. A discriminant analysis eigenvalues plot was displayed showing three linear discriminants which were all retained (Figure 4.5).

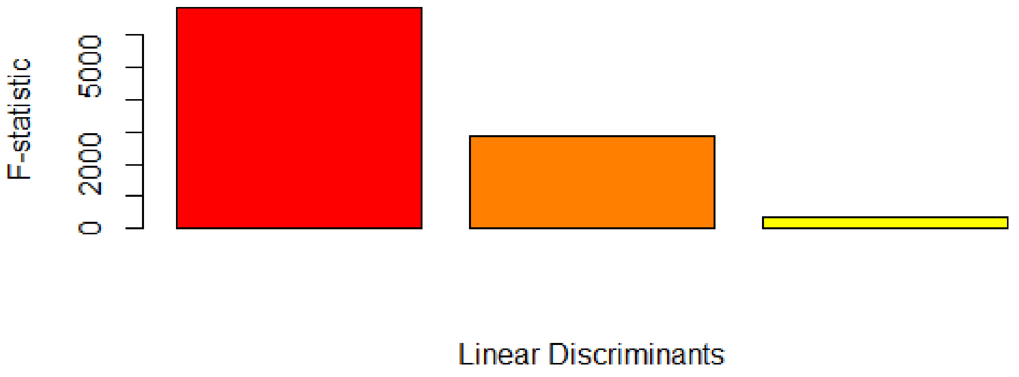


**Figure 4.5:** Discriminant analysis eigenvalues showing three linear discriminants that were retained and which accounted for 67.1% conserved variance

The first 60 PCs of PCA were used representing a proportion of conserved variance of 67.1%. Cross-validation of the DAPC confirmed these values (Figure 4.9).

From the DAPC scatter plot, the overall population was divided into four clusters. Individuals from cluster 4 were observed to spread along the y-axis. In the scatterplot, the graph of the PCA eigenvalues was retained (Figure 4.6).

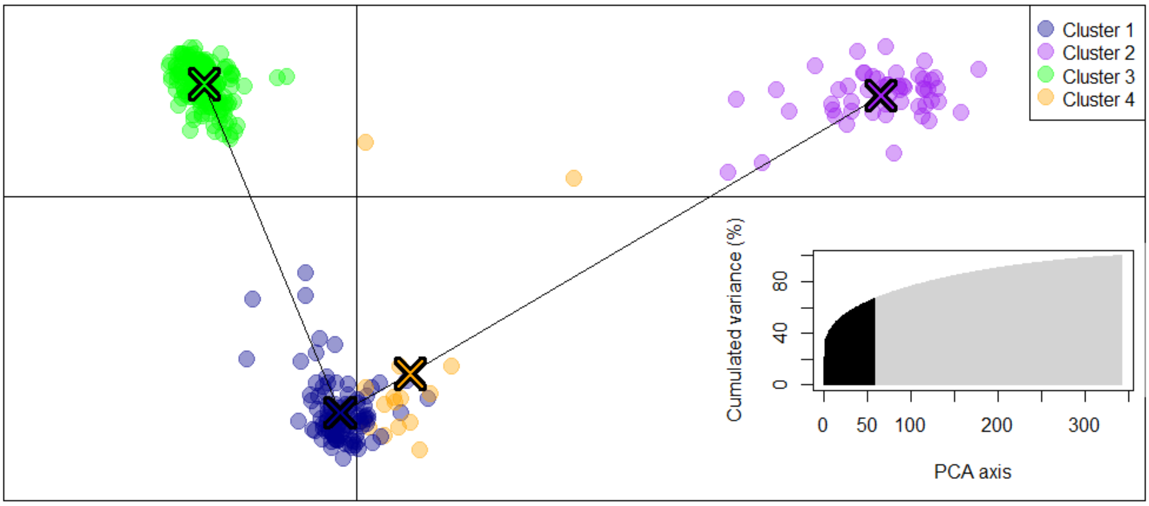


**Figure 4.6:** DAPC scatterplot. Crosses indicate the centre of each group; a minimum spanning tree indicates the actual closeness between subpopulations. Individuals are represented as coloured dots.

A summary of the DAPC showed the probabilities of the membership based on the discriminant functions that were retained. The summary of the DAPC was computed as summary statistics in R and displayed as summary (dapc) (Table 4.4).

**Table 4.4:** Summary statistics of DAPC outlining the group size per cluster of *U. kirkiana*

| **No. of dimensions** | **No of subpopulations** | **Assign per population** | | | | **Prior group size** | | | | **Post group size** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3** | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| | | 1 | 1 | 1 | 1 | 102 | 52 | 173 | 15 | 102 | 52 | 173 | 15 |

The summary (dapc) was composed of the number of dimensions, number of subpopulations, the overall number of populations, and population size before and after assignment. Whole numbers were observed after successful reassignment of individuals to their clusters as was indicated in the assign.per.pop slot. The clusters remained unchanged in size after reassignment. The summary (dapc) was used to generate a heatmap for the first individuals in the dataset. The initial clustering before DAPC was represented by the blue crosses which were consistent with the DAPC classification as the crosses were within the red rectangles (Figure 4.7).

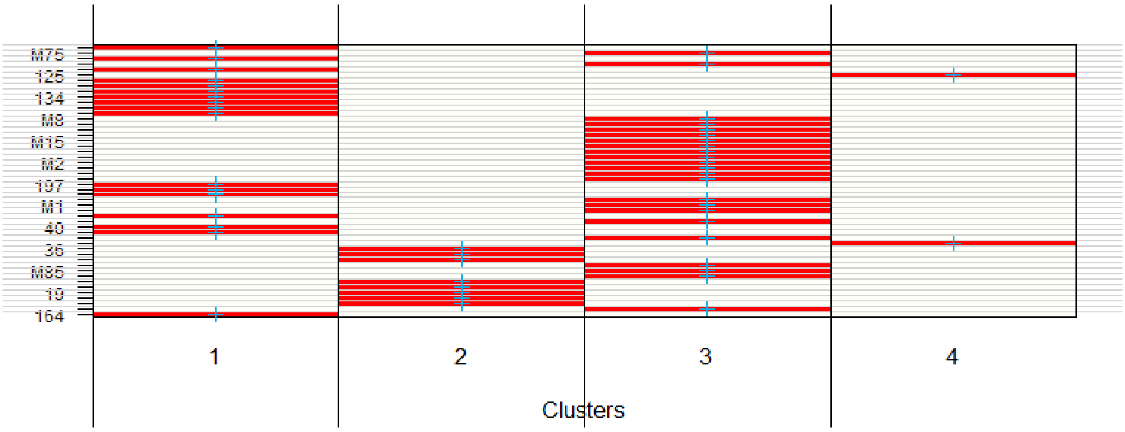


**Figure 4.7:** A DAPC summary heatmap of the first 50 individuals of *U. kirkiana* in the dataset

Heat colours are the membership probabilities; red is equivalent to one, and white is equal to zero. Blue crosses are the initial clusters that were provided to DAPC. The summary(dapc) was also used to plot a compoplot (Figure 4.8). The individuals in the compoplot had membership probabilities of 100% in a cluster. There were no individuals with less than 99% membership probability in a cluster.

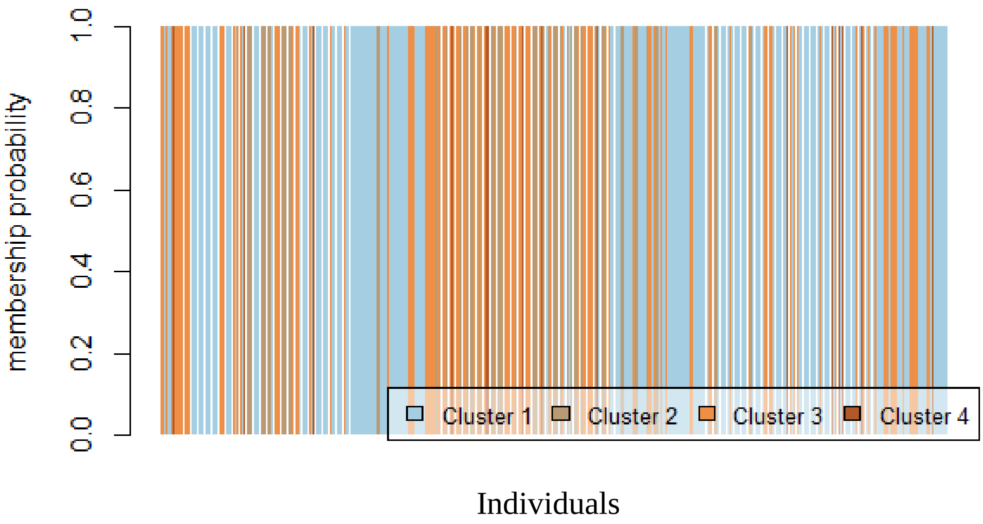


**Figure 4.8:** Composite plot of membership probability in each of the clusters identified in DAPC

### 4.2.3 Cross-validation of DAPC

When choosing the number of PCs for analysis, it is important to ensure that the suitable number of PCs are selected. The choice of the appropriate number of PCs to retain helps to include a greater source of variation in the data. DAPC cross-validation makes it possible to determine that the number of PCs retained is enough. From the DAPC cross-validation analysis, the number of PC's achieving the highest mean success were 60, and the number of PC's achieving the lowest mean squared error was 60 (Figure 4.9). Implying that retaining 60 PCs would account for the greatest source of variation in this study, which was the case.

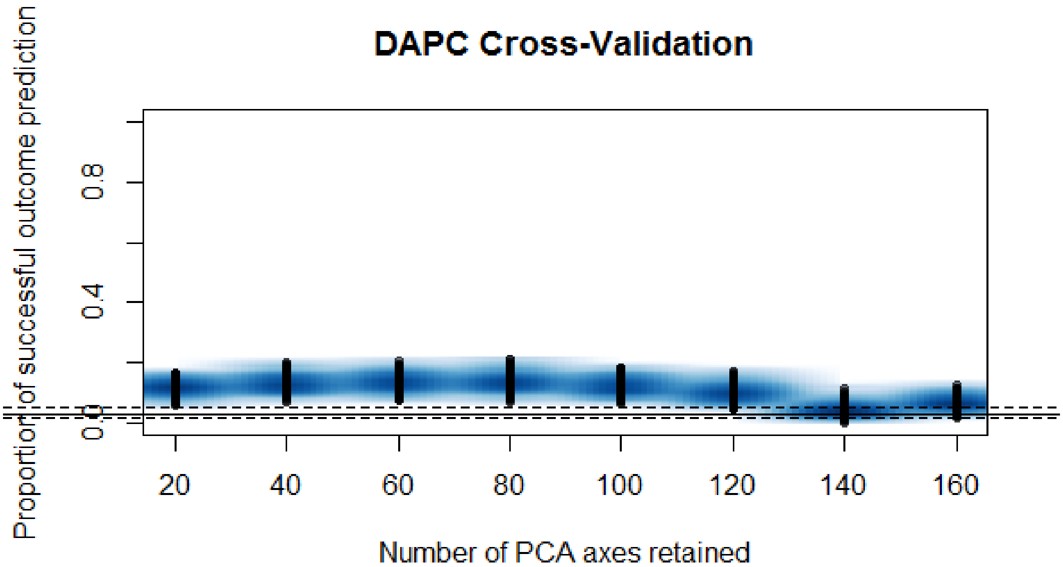


**Figure 4.9*:*** DAPC cross-validation confirming the true number principle components for achieving the highest mean

# CHAPTER FIVE

# DISCUSSION

## 5.1 Assessment of allelic diversity

DArT sequencing led to the discovery of 28393 SNPs from 342 individuals from *Uapaca Kirkiana.* The markers had an average reproducibility of 99.95% and an average call rate of 0.5%, which met the criteria for marker quality control, as illustrated by Kilian *et al.* (2012). DArT SNPs are filtered for insignificant markers and genotypes at DArT Pty. However, too much filtering would result in loss of significant markers (Gruber *et al*., 2019). An initial analysis was conducted on EM, PPCA, SVD and Nipals algorithms to determine the best method for imputing the original dataset. A comparison of the SMC from the four methods of imputation showed that EM was the highest-scoring method. As such, EM was the preferred method of imputation for the original dataset. The markers did not have chromosome information as this was the first time that *Uapaca kirkiana* was being sequenced. The SNPs were highly polymorphic with a PIC value ranging from 0.003 to 0.5 suggesting that the markers were genetically diverse (Avolio *et al*., 2012).

## 5.2 Assessment of genetic parameters and relations

SNPs provide accurate genomic data compared to other markers and thus have been used in genome level profile studies (Mammadov *et al*., 2012). Information derived from SNP data, including genetic diversity and population structure, is vital in crop improvement, germplasm conservation and crop management (Ríos, 2015). In this study, low levels of genetic diversity indices were observed. There was low differentiation among populations as was shown in AMOVA, low Dest and low Fst levels of 0.0065 and 0.0551, respectively implying that there is low genetic diversity and low population differentiation. The low genetic diversity may, in part, suggest that there is high gene flow within populations. There was a high genetic density of 82% within populations signifying high genetic differentiation. According to Pongratz *et al*. (2002), gene flow restrictions between populations lead to high genetic differentiation. Consequently, a restricted gene pool can lead to a decrease in diversity which is detrimental to the survival of the population (Giles *et al*., 1998).

Plant breeding has been shown to have an impact on genetic diversity by increasing crop uniformity in the field. In a study by Rauf *et al.* (2010), the domestication of plants led to losses in genetic diversity. *Uapaca kirkiana* one of the indigenous fruit trees that have been domesticated (Akinnifesi *et al*., 2002), could suggest that the loss of genetic diversity is due to breeding. Further, the inbreeding coefficient per overall was low levels of inbreeding within a population (Szczecińska *et al*., 2016), leading to a deduction that the observed low levels in Dest and Fst statistics could be largely due to the small population size evident in the domesticated trees (Furlan *et al*., 2012).

## 5.3 Determination of population structure

Population structure is a metapopulation resulting from individuals that are assembled into local populations depending on genetic differences shared (Woodruff, 2001). Parametric and nonparametric approaches infer the population structure and individuals' allocation to subpopulations (Alhusain and Hafez, 2018). DAPC is a non-hierarchical clustering technique used to classify and define individuals based on genetic relativeness (Jombart *et al*., 2010). Analysis of the population structure through DAPC yielded four clusters that were well defined based on their genetic make-up arising from the individuals. The results were comparable to the four-clustered dendrogram that was obtained using the Neighbour-joining (NJ) clustering analysis. However, the assignment of individuals between the Neighbour joining method and the DAPC differed. In the NJ dendrogram, the subpopulations were assigned into distinct groups. However, in the DAPC the populations were assigned to a cluster one individual at a time. Moreover, the cluster sizes in the two methods were not identical. The difference in the assignment of individuals in both methods is as a result of different algorithms that are used in the two methods of clustering. DAPC uses K-means clustering whereas NJ clustering uses the Ward method which is an agglomerative hierarchical clustering method (Rasmussen, 1992; Jombart *et al*., 2010). In comparison, K-means is a partitioning non-hierarchical method that identifies the K-number of clusters and then assigns each observation to the nearest mean, while optimising homogeneity measurements within groups and heterogeneity between clusters (Natingga, 2017).

The DAPC method for population structure analysis is dependent on PCA to reduce the dimension of data and linear discriminant analysis. The number of PCAs to be retained as well as the quality of the resulting DAPC are confirmed through cross-validation of the DAPC. The ideal number of PCs to retain in the DAPC are those that are linked to the lowest mean squared error (Jombart and Collins, 2015). From this study, the number of PCs associated with the means squared error was 60, which were the actual number of PCs selected from the onset of DAPC analysis. Therefore, the cross-validation results confirmed that the number of PCs retained in the analysis was optimum. From the DAPC heatmap and compoplot, individual membership probability in a cluster was 100% implying that the clusters were well defined and there was no admixture within the population.

# CHAPTER SIX

# CONCLUSIONS AND RECOMMENDATION

## 6.1 Conclusions

DArT sequencing of SNPs identified polymorphic markers and revealed diversity among the populations that were analysed. The *U. kirkiana* population was structured and composed of four clusters. As such, it would be economical to select a representative sample of each cluster to be preserved for germplasm conservation. There was a high genetic density within populations and a lower genetic density among populations. The genetic diversity was low across the populations, which may have been as a result of the tree conservation strategy.

## 6.2 Recommendation

- The germplasm conservation unit at ICRAF may want to use populations that are genetically distant to increase diversity and enhance the long-term survival of the fruit tree.
- Further analysis of *U. kirkiana* accessions for sex markers will lead to the identification of the sex-specific markers at the molecular level, and this information will be helpful in the selection of the most desirable varieties for conservation purposes.

## REFERENCES

Adhikari, S., Saha, S., Bandyopadhyay, T. K., and Ghosh, P. (2016). Marker assisted sex determination in dioecious crops: An advancement in molecular biology. In Gupta, N (Eds.). *Research Trends in Molecular Biology*, Chapter II, 35-70. Research Signpost, Kerala, India.

Adu, B.G., Akromah, R., Amoah, S., Nyadanu, D., Yeboah, A., Aboagye, L.M., Amoah, R.A. and Owusu, E.G. (2021). High-density DArT-based SilicoDArT and SNP markers for genetic diversity and population structure studies in cassava (*Manihot esculenta* Crantz). *PloS One, 16*(7), p.e0255290.

Akinnifesi, F. K., Kwesiga, F. R., Mhango, J., Mkonda, A., Chilanga, T., and Swai, R. (2002, August). Domesticating priority miombo indigenous fruit trees as a promising livelihood option for small-holder farmers in Southern Africa. In *XXVI International Horticultural Congress: Citrus and Other Subtropical and Tropical Fruit Crops: Issues, Advances and 632* (pp. 15-30).

Alhusain, L., and Hafez, A. M. (2018). Nonparametric approaches for population structure analysis. *Human Genomics*, *12*(1), 25.

Alstrom-Rapaport, C., Lascoux, M., Wang, Y. C., Roberts, G., and Tuskan, G. A. (1998). Identification of a RAPD marker linked to sex determination in the basket willow (*Salix viminalis* L.). *Journal of Heredity*, *89*(1), 44-49.

Appleby, N., Edwards, D., and Batley, J. (2009). New technologies for ultra-high throughput genotyping in plants. In *Plant Genomics* (pp. 19-39). Humana Press.

Avinash, M., Anurag, K. S., and Gaur, R. K. (2014). Molecular markers: Tool for genetic analysis. In Ashish S. V. and Anchal S (Eds.). *Animal Biotechnology, Models in Discovery and Translation.* Chapter 16, 289-305. Academic Press, Elsevier, Inc.

Avolio, M. L., Beaulieu, J. M., Lo, E. Y., and Smith, M. D. (2012). Measuring genetic diversity in ecological studies. *Plant Ecology*, *213*(7), 1105-1115.

Bakoumé, C., Wickneswari, R., Siju, S., Rajanaidu, N., Kushairi, A. and Billotte, N. (2015). Genetic diversity of the world's largest oil palm (*Elaeis guineensis* Jacq.) field genebank accessions using microsatellite markers. *Genetic Resources and Crop Evolution*, *62*(3), 349-360.

Baloch, F.S., Alsaleh, A., Shahid, M.Q., Çiftçi, V., E. Sáenz de Miera, L., Aasim, M., Nadeem, M.A., Aktaş, H., Özkan, H. and Hatipoğlu, R. (2017). A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. *Plos One, 12*(1), p.e0167821.

Bhandari, H. R., Bhanu, A. N., Srivastava, K., Singh, M. N. and Shreya, H. A. (2017). Assessment of genetic diversity in crop plants-an overview. *Advances in Plants Agriculture Resources, 7*(3), 279-286.

Bhattacharjee, R., Agre, P., Bauchet, G., De Koeyer, D., Lopez-Montes, A., Kumar, P.L., Abberton, M., Adebola, P., Asfaw, A. and Asiedu, R. (2020). Genotyping-by-sequencing to unlock genetic diversity and population structure in white yam (*Dioscorea rotundata* Poir.). *Agronomy, 10*(9), p.1437.

Chakraborty, R. (1993). Analysis of genetic structure of populations: meaning, methods, and implications. In *Human Population Genetics* (pp. 189-206). Springer, Boston, MA.

Chen, X., and Sullivan, P. F. (2003). Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *The Pharmacogenomics Journal, 3*(2), 77-96.

Chirwa, P. W., and Akinnifesi, F. K. (2008). Ecology and biology of *Uapaca kirkiana*, Strychnos cocculoides and Sclerocarya birrea in Southern Africa. In Akinnifesi F.K., R.R.B. Leakey, O.C. Ajayi, G. Sileshi, Z. Tchoundjeu, P. Matakala, F.R. Kwesiga (Eds.), *Indigenous Fruit Trees in the Tropics: Domestication, Utilization and Commercialization* (pp. 322-340). CAB International Publishing, Wallingford, UK.

Dar, A.A., Mahajan, R. and Sharma, S. (2019). Molecular markers for the characterization and conservation of plant genetic resources. *Indian J Agric Sci, 89*, pp.1755-1763.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

Deperi, S.I., Tagliotti, M.E., Bedogni, M.C., Manrique-Carpintero, N.C., Coombs, J., Zhang, R., Douches, D. and Huarte, M.A. (2018). Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. *PloS One, 13*(3), p.e0194398.

Diversity Arrays Technology. (2014). KDCompute Data Analysis. Retrieved October 22, 2020, from https://kdcompute.seqart.net/kdcompute/login

Doyle, J. J. and Doyle, J. L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue* (No. RESEARCH).

El-Esawi, M. A. (2017). Genetic diversity and evolution of Brassica genetic resources: from morphology to novel genomic technologies–a review. *Plant Genetic Resources, 15*(5), 388-399.

Ellegren, H., and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics, 17*(7), 422-433.

Etminan, A., Pour-Aboughadareh, A., Mohammadi, R., Ahmadi-Rad, A., Noori, A., Mahdavian, Z. and Moradi, Z. (2016). Applicability of start codon targeted (SCoT) and inter-simple sequence repeat (ISSR) markers for genetic diversity analysis in durum wheat genotypes. *Biotechnology and Biotechnological Equipment*, *30*(6), 1075-1081.

Excoffier, L. and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, *10*(3), 564-567.

Fatokun, C., Girma, G., Abberton, M., Gedil, M., Unachukwu, N., Oyatomi, O., Yusuf, M., Rabbi, I. and Boukar, O. (2018). Genetic diversity and population structure of a mini-core subset from the world cowpea (*Vigna unguiculata* (L.) Walp.) germplasm collection. *Scientific Reports*, *8*(1), pp.1-10.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. *Department of Genome Sciences, University of Washington, Seattle.*

Furlan, E., Stoklosa, J., Griffiths, J., Gust, N., Ellis, R., Huggins, R. M., and Weeks, A. R. (2012). Small population size and extremely low levels of genetic diversity in island populations of the platypus, *Ornithorhynchus anatinus*. *Ecology and Evolution*, *2*(4), 844-857.

Georges, A. and Gruber, B. (2019). SNP Analysis using dartR: Guide to Preparatory Analysis. Canberra: The Institute for Applied Ecology, University of Canberra.

Giles, B. E., Lundqvist, E., and Goudet, J. (1998). Restricted gene flow and subpopulation differentiation in *Silene dioica*. *Heredity*, *80*(6), 715-723.

Govindaraj, M., Vetriventhan, M. and Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genetics Research International*, *2015*, 431487.

Grover, A. and Sharma, P.C. (2016). Development and use of molecular markers: past and present. *Critical Reviews in Biotechnology*, *36*(2), pp.290-302.

Gruber, B., Unmack, P. J., Berry, O. F., and Georges, A. (2018). DartR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, *18*(3), 691-699.

Gruber, B., Unmack, P., Berry, O., and Georges, A. (2019). Introduction to dartR. *User Manual*.

Gupta, P. K., Rustgi, S., and Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity*, *101*(1), 5.

Hunter, J. E., and Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Hurtado, P., Olsen, K. M., Buitrago, C., Ospina, C., Marin, J., Duque, M., .and Adeleke, M. (2008). Comparison of simple sequence repeat (SSR) and diversity array technology (DArT) markers for assessing genetic diversity in cassava (Manihot esculenta Crantz). *Plant Genetic Resources*, *6*(3), 208-214.

Huttner, E., Wenzl, P., Akbari, M., Caig, V., Carling, J., Cayla, C., and Uszynski, G. (2005). Diversity arrays technology: a novel tool for harnessing the genetic potential of orphan crops. In *Discovery to Delivery: BioVision Alexandria 2004, Proceedings of the 2004 Conference of The World Biological Forum. CABI Publishing: UK* (pp. 145-155).

Idrees, M., and Irshad, M. (2014). Molecular markers in plants for analysis of genetic diversity: a review. *European Academic Research*, *2*(1), 1513-1540.

Ingvarsson, P. K. and Dahlberg, H. (2019). The effects of clonal forestry on genetic diversity in wild and domesticated stands of forest trees. *Scandinavian Journal of Forest Research*, *34*(5), 370-379.

Jinga, P., Palagi, J., Chong, J. P., and Bobo, E. D. (2020). Climate change reduces the natural range of African wild loquat (*Uapaca kirkiana* Müll. Arg., Phyllanthaceae) in south-central Africa. *Regional Environmental Change, 20*(3), 1-13.

John, C., Ekpenyong, E. J. and Nworu, C. C. (2019). Imputation of missing values in economic and financial time series data using five principal component analysis approaches. *CBN Journal of Applied Statistics (JAS), 10*(1), 3.

Jombart, T., and Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. *London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling*.

Jombart T, Devillard S and Balloux, F (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11: 94.

Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403-1405.

Jump, A. S., Marchant, R., and Peñuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends in plant science, 14*(1), 51-58.

Kalaba, F. K., Chirwa, P. W., and Prozesky, H. (2009). The contribution of indigenous fruit trees in sustaining rural livelihoods and conservation of natural resources. *Journal of Horticulture and Forestry, 1*(1), 1-6.

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., and Aschenbrenner-Kilian, M. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. In *Data production and analysis in population genomics* (pp. 67-89). Humana Press, Totowa, NJ.

Kordrostami, M., and Rahimi, M. (2015). Molecular markers in plants: concepts and applications. *Genet. 3rd Millenn, 13*, 4024-4031.

Lengkeek, A. G., Mwangi, A. M., Agufa, C. A., Ahenda, J. O., and Dawson, I. K. (2006). Comparing genetic diversity in agroforestry systems with natural forest: a case study of the important timber tree *Vitex fischeri* in central Kenya. *Agroforestry Systems*, *67*(3), 293-300.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research, 47*(W1), W256-W259.

Lucena-Aguilar, G., Sánchez-López, A. M., Barberán-Aceituno, C., Carrillo-Avila, J. A., López-Guerrero, J. A., and Aguilar-Quesada, R. (2016). DNA source selection for downstream applications based on DNA quality indicators analysis. *Biopreservation and Biobanking,14*(4), 264-270.

Luck, G. W., Daily, G. C., and Ehrlich, P. R. (2003). Population diversity and ecosystem services. *Trends in Ecology & Evolution*, *18*(7), 331-336.

Luo, Z., Brock, J., Dyer, J.M., Kutchan, T., Schachtman, D., Augustin, M., Ge, Y., Fahlgren, N. and Abdel-Haleem, H., (2019). Genetic diversity and population structure of a *Camelina sativa* spring panel. *Frontiers in Plant Science, 10,* 184.

Mahboubi, M., Mehrabi, R., Naji, A. M. and Talebi, R. (2020). Whole-genome diversity, population structure and linkage disequilibrium analysis of globally diverse wheat genotypes using genotyping-by-sequencing DArTseq platform. *3 Biotech*, *10*(2), 1-13.

Mammadov, J., Aggarwal, R., Buyyarapu, R., and Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International Journal of Plant Genomics, 2012*, 1-11.

Milligan, G. W., and Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied Psychological Measurement*, *11*(4), 329-354.

Mithöfer, D., and Waibel, H. (2003). Income and labour productivity of collection and use of indigenous fruit tree products in Zimbabwe. *Agroforestry Systems*, *59*(3), 295-305.

Mwase, W. F., Akinnifesi, F. K., Stedje, B., Kwapata, M. B., and Bjørnstad, Å. (2010). Genetic diversity within and among southern African provenances of *Uapaca kirkiana* Müell. Årg using morphological and AFLP markers. *New Forests*, *40*(3), 383-399.

Mwase, W. F., Erik-Lid, S., Bjørnstad, Å., Stedje, B., Kwapata, M. B., and Bokosi, J. M. (2007). Application of amplified fragment length polymorphism (AFLPs) for detection of sex–specific markers in dioecious *Uapaca kirkiana* Muell. Årg. *African Journal of Biotechnology, 6*(2), 137-142.

Natingga, D. (2017). *Data Science Algorithms in a Week*. Packt Publishing Ltd.

Negash, A. W. (2015). *Application of mixed model and spatial analysis methods in multi-environmental and agricultural field trials* (Doctoral dissertation).

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics, 89*(3), 583-590.

Ngulube, M. R., Hall, J. B., and Maghembe, J. A. (1998). Reproductive ecology of *Uapaca kirkiana* (Euphorbiaceae) in Malawi, southern Africa. *Journal of Tropical Ecology*, *14*(6), 743-760.

Ngulube, M. R., Hall, J. B., and Maghembe, J. A. (1996). A review of the silviculture and resource potential of a miombo fruit tree: *Uapaca kirkiana* (Euphorbiaceae). *Journal of Tropical Forest Science*, 8(3), 395-411.

Ngulube, M. R., Hall, J. B., and Maghembe, J. A. (1995). Ecology of a miombo fruit tree: *Uapaca kirkiana* (Euphorbiaceae). *Forest Ecology and Management*, *77*(1-3), 107-117.

Orwa, C., Mutua, A., Kindt, R., Jamnadass, R., and Anthony, S. (2009). Agroforestree Database: a tree reference and selection guide version 4.0. *World Agroforestry Centre, Kenya*, *15*.

Pagnotta, M. A. (2018). Comparison among methods and statistical software packages to analyze germplasm genetic diversity by means of codominant markers. *J-Multidisciplinary Scientific Journal*, *1*(1), 197-215.

Peakall, R. O. D. and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, *6*(1), 288-295.

Pidigam, S., Munnam, S.B., Nimmarajula, S., Gonela, N., Adimulam, S.S., Yadla, H., Bandari, L. and Amarapalli, G. (2019). Assessment of genetic diversity in yardlong bean (Vigna unguiculata (L.) Walp subsp. sesquipedalis Verdc.) germplasm from India using RAPD markers. *Genetic Resources and Crop Evolution*, *66*(6), pp.1231-1242.

Pongratz, N., Gerace, L., and Michiels, N. K. (2002). Genetic differentiation within and between populations of a hermaphroditic freshwater planarian. *Heredity*, *89*(1), 64-69.

Porebski, S., Bailey, L. G., and Baum, B. R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant molecular Biology Reporter*, *15*(1), 8-15.

Porth, I., and El-Kassaby, Y. A. (2014). Assessment of the genetic diversity in forest tree populations using molecular markers. *Diversity*, *6*(2), 283-295.

Pritchard, J. K., Wen, X. and Falush, D. (2010). Documentation for structure software: Version 2.3. *University of Chicago, Chicago, IL*.

R Core Team (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rasmussen, E. M. (1992). Clustering algorithms. *Information retrieval: data structures & algorithms*, *419*, 442.

Rauf, S., da Silva, J. T., Khan, A. A., and Naveed, A. (2010). Consequences of plant breeding on genetic diversity. *International Journal of Plant Breeding*, *4*(1), 1-21.

Ríos, R. O. (2015). *Plant breeding in the omics era*. Springer International Publishing, Switzerland

Rokach, L., and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer, Boston, MA.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406-425.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular cloning: a laboratory manual* (No. Ed. 2). Cold spring harbor laboratory press.

Sánchez-Sevilla, J. F., Horvath, A., Botella, M. A., Gaston, A., Folta, K., Kilian, A., and Amaya, I. (2015). Diversity Arrays Technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria ananassa*). *PLoS One*, *10*(12), e0144960.

Seyedimoradi, H., Talebi, R., Kanouni, H., Naji, A. M. and Karami, E. (2020). Genetic diversity and population structure analysis of chickpea (Cicer arietinum L.) advanced breeding lines using whole-genome DArTseq-generated SilicoDArT markers. *Brazilian Journal of Botany*, *43*(3), 541-549

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., and Belzile, F. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PloS one*, *8*(1), e54603.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., abd Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, *23*(9), 1164-1167.

Szczecińska, M., Sramko, G., Wołosz, K., and Sawicki, J. (2016). Genetic diversity and population structure of the rare and endangered plant species *Pulsatilla patens* (L.) Mill in East Central Europe. *PLoS One*, *11*(3), e0151730.

Turchetto, C., Segatto, A. L. A., Mäder, G., Rodrigues, D. M., Bonatto, S. L., and Freitas, L. B. (2016). High levels of genetic diversity and population structure in an endemic and rare species: implications for conservation. *AoB Plants*, *8*.

Wittenberg, A. H. (2007). *Genetic mapping using the Diversity Arrays Technology (DArT): application and validation using the whole-genome sequences of Arabidopsis thaliana and the fungal wheat pathogen Mycosphaerella graminicola*. Wageningen University and Research.

Woodruff, D. S. (2001). Populations, species, and conservation genetics. In: Levin S (ed). *Encyclopedia of Biodiversity*, 4th edition., 811–829. Academic, San Diego.

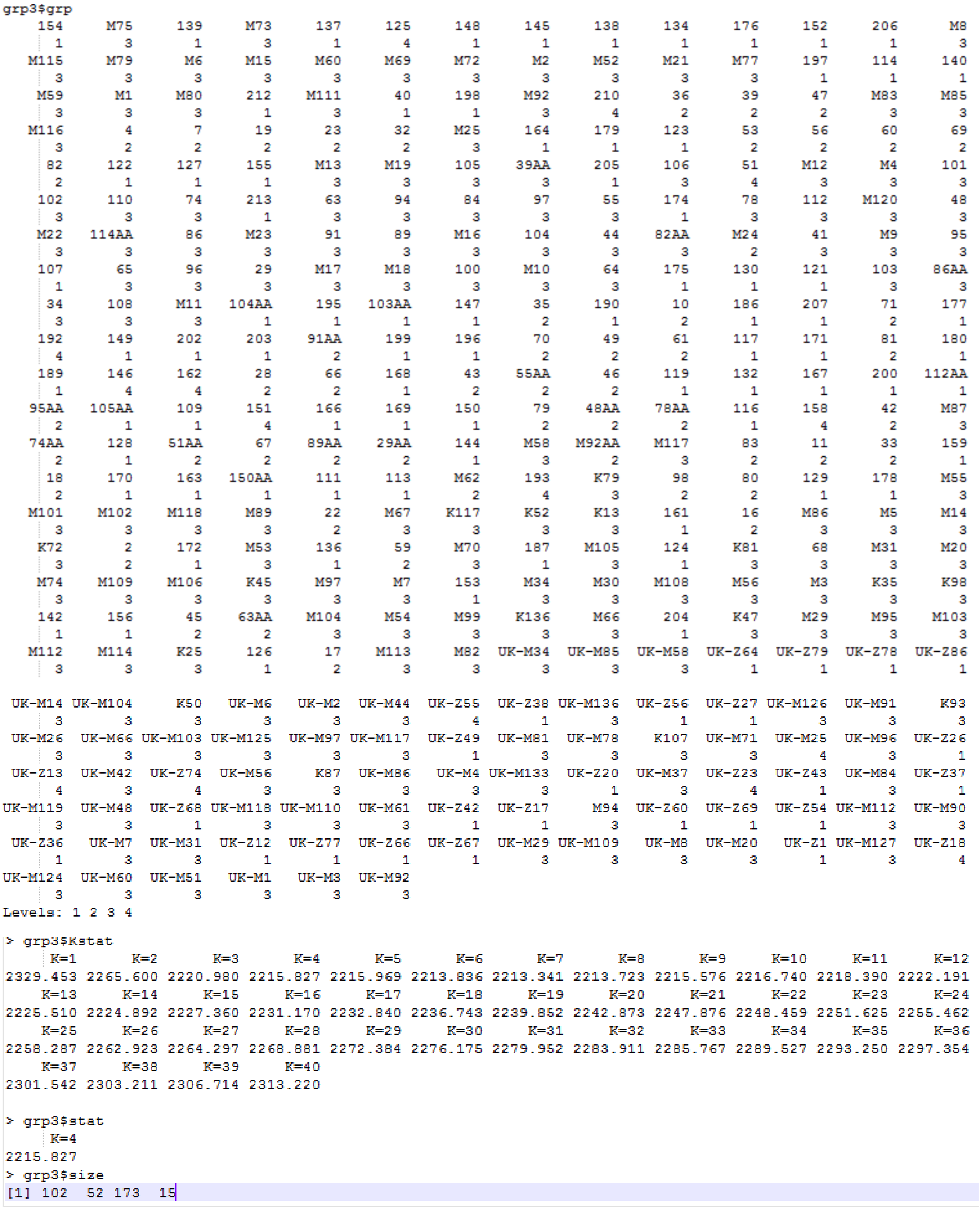
Xiong, H., Chen, Y., Gao, S. J., Pan, Y. B. and Shi, A. (2022). Population Structure and Genetic Diversity Analysis in Sugarcane (Saccharum spp. hybrids) and Six Related Saccharum Species. *Agronomy*, *12*(2), 412.

Yang, X., Ren, R., Ray, R., Xu, J., Li, P., Zhang, M., Liu, G., Yao, X. and Kilian, A. (2016). Genetic diversity and population structure of core watermelon (Citrullus lanatus) genotypes using DArTseq-based SNPs. *Plant Genetic Resources*, *14*(3), pp.226-233.

Zhang, J., Wang, X., Yao, J., Li, Q., Liu, F., Yotsukura, N., Krupnova, T.N. and Duan, D., 2017. Effect of domestication on the genetic diversity and structure of Saccharina japonica populations in China. *Scientific Reports*, *7*(1), 1-11.

# APPENDICES

**Appendix 1:** Find.clusters () output showing how individuals were assigned into clusters in DAPC. 1, 2, 3, and 4 values in the table indicate the clusters. The value above the cluster number is the identity of the individual.



```
grp3$grp
     154     M75     139     M73     137     125     148     145     138     134     176     152     206      M8
       1       3       1       3       1       4       1       1       1       1       1       1       1       3
    M115     M79      M6     M15     M60     M69     M72      M2     M52     M21     M77     197     114     140
       3       3       3       3       3       3       3       3       3       3       3       1       1       1
     M59      M1     M80     212    M111      40     198     M92     210      36      39      47     M83     M85
       3       3       3       1       3       1       1       3       4       2       2       2       3       3
    M116       4       7      19      23      32     M25     164     179     123      53      56      60      69
       3       2       2       2       2       2       3       1       1       1       2       2       2       2
      82     122     127     155     M13     M19     105    39AA     205     106      51     M12      M4     101
       2       1       1       1       3       3       3       3       1       3       4       3       3       3
     102     110      74     213      63      94      84      97      55     174      78     112    M120      48
       3       3       3       1       3       3       3       3       3       1       3       3       3       3
     M22   114AA      86     M23      91      89     M16     104      44    82AA     M24      41      M9      95
       3       3       3       3       3       3       3       3       3       3       2       3       3       3
     107      65      96      29     M17     M18     100     M10      64     175     130     121     103    86AA
       1       3       3       3       3       3       3       3       3       1       1       1       3       3
      34     108     M11   104AA     195   103AA     147      35     190      10     186     207      71     177
       3       3       3       1       1       1       1       2       1       2       1       1       2       1
     192     149     202     203    91AA     199     196      70      49      61     117     171      81     180
       4       1       1       1       2       1       1       2       2       2       1       1       2       1
     189     146     162      28      66     168      43    55AA      46     119     132     167     200   112AA
       1       4       4       2       2       1       2       2       2       1       1       1       1       1
    95AA   105AA     109     151     166     169     150      79    48AA    78AA     116     158      42     M87
       2       1       1       4       1       1       1       2       2       2       1       4       2       3
    74AA     128    51AA      67    89AA    29AA     144     M58   M92AA    M117      83      11      33     159
       2       1       2       2       2       2       1       3       2       3       2       2       2       1
      18     170     163   150AA     111     113     M62     193     K79      98      80     129     178     M55
       2       1       1       1       1       1       2       4       3       2       2       1       1       3
    M101    M102    M118     M89      22     M67    K117     K52     K13     161      16     M86      M5     M14
       3       3       3       3       2       3       3       3       3       1       2       3       3       3
     K72       2     172     M53     136      59     M70     187    M105     124     K81      68     M31     M20
       3       2       1       3       1       2       3       1       3       1       3       3       3       3
     M74    M109    M106     K45     M97      M7     153     M34     M30    M108     M56      M3     K35     K98
       3       3       3       3       3       3       1       3       3       3       3       3       3       3
     142     156      45    63AA    M104     M54     M99    K136     M66     204     K47     M29     M95    M103
       1       1       2       2       3       3       3       3       3       1       3       3       3       3
    M112    M114     K25     126      17    M113     M82  UK-M34  UK-M85  UK-M58  UK-Z64  UK-Z79  UK-Z78  UK-Z86
       3       3       3       1       2       3       3       3       3       3       1       1       1       1

  UK-M14 UK-M104     K50   UK-M6   UK-M2  UK-M44  UK-Z55  UK-Z38 UK-M136  UK-Z56  UK-Z27 UK-M126  UK-M91     K93
       3       3       3       3       3       3       4       1       3       1       1       3       3       3
  UK-M26  UK-M66 UK-M103 UK-M125  UK-M97 UK-M117  UK-Z49  UK-M81  UK-M78    K107  UK-M71  UK-M25  UK-M96  UK-Z26
       3       3       3       3       3       3       1       3       3       3       3       4       3       1
  UK-Z13  UK-M42  UK-Z74  UK-M56     K87  UK-M86   UK-M4 UK-M133  UK-Z20  UK-M37  UK-Z23  UK-Z43  UK-M84  UK-Z37
       4       3       4       3       3       3       3       3       1       3       4       1       3       1
 UK-M119  UK-M48  UK-Z68 UK-M118 UK-M110  UK-M61  UK-Z42  UK-Z17     M94  UK-Z60  UK-Z69  UK-Z54 UK-M112  UK-M90
       3       3       1       3       3       3       1       1       3       1       1       1       3       3
  UK-Z36   UK-M7  UK-M31  UK-Z12  UK-Z77  UK-Z66  UK-Z67  UK-M29 UK-M109   UK-M8  UK-M20   UK-Z1 UK-M127  UK-Z18
       1       3       3       1       1       1       1       3       3       3       3       1       3       4
 UK-M124  UK-M60  UK-M51   UK-M1   UK-M3  UK-M92
       3       3       3       3       3       3
Levels: 1 2 3 4

> grp3$Kstat
     K=1      K=2      K=3      K=4      K=5      K=6      K=7      K=8      K=9     K=10     K=11     K=12 
2329.453 2265.600 2220.980 2215.827 2215.969 2213.836 2213.341 2213.723 2215.576 2216.740 2218.390 2222.191 
    K=13     K=14     K=15     K=16     K=17     K=18     K=19     K=20     K=21     K=22     K=23     K=24 
2225.510 2224.892 2227.360 2231.170 2232.840 2236.743 2239.852 2242.873 2247.876 2248.459 2251.625 2255.462 
    K=25     K=26     K=27     K=28     K=29     K=30     K=31     K=32     K=33     K=34     K=35     K=36 
2258.287 2262.923 2264.297 2268.881 2272.384 2276.175 2279.952 2283.911 2285.767 2289.527 2293.250 2297.354 
    K=37     K=38     K=39     K=40 
2301.542 2303.211 2306.714 2313.220 

> grp3$stat
     K=4 
2215.827 
> grp3$size
[1] 102  52 173  15
```

**Appendix 2:** Distribution of *U.kirkiana* samples in the four subgroups from the NJ analysis. The subgroups were made up of individuals from different countries. The samples in subgroups did not correspond to the area of location

| **Group 1** | | |
|---|---|---|
| **Sample ID** | **Provenance** | **Country of Origin** |
| **30** | Lwilomelo | Zimbabwe |
| **33** | Musana | Zimbabwe |
| **35** | Gombea | Tanzania |

| **Group 2** | | |
|---|---|---|
| **Sample ID** | **Provenance** | **Country of Origin** |
| **44** | Mbala | Zambia |
| **49** | Lwilomelo | Zimbabwe |

| **Group 3** | | |
|---|---|---|
| **Sample ID** | **Provenance** | **Country of Origin** |
| **1** | Nyamukwarara | Zimbabwe |
| **2** | Mbala | Zambia |
| **3** | Lwilomelo | Zimbabwe |
| **4** | Choma | Zambia |
| **5** | Mapanzure | Zimbabwe |
| **6** | Musana | Zimbabwe |
| **8** | Choma | Zambia |
| **9** | Lwilomelo | Zimbabwe |
| **10** | Mbala | Zambia |
| **11** | Choma | Zambia |
| **12** | Serenje | Zambia |
| **14** | Musana | Zimbabwe |
| **15** | Mapanzure | Zimbabwe |
| **17** | Choma | Zambia |
| **18** | Lwilomelo | Zimbabwe |
| **19** | Mbala | Zambia |
| **22** | Domboshawa | Zimbabwe |
| **23** | Murewa | Zimbabwe |
| **25** | Luwawa | Malawi |
| **26** | Luwawa | Malawi |
| **27** | Nyamukwarara | Zimbabwe |
| **29** | Lwilomelo | Zimbabwe |
| **31** | Litende | Malawi |
| **34** | Luwawa | Malawi |
| **36** | Musana | Zimbabwe |
| **40** | Domboshawa | Zimbabwe |
| **41** | Mbala | Zambia |
| **43** | Choma | Zambia |
| **45** | Musana | Zimbabwe |
| **48** | Lwilomelo | Zimbabwe |
| **50** | Mbala | Zambia |
| **51** | Domboshawa | Zimbabwe |
| **52** | Lwilomelo | Zimbabwe |
| **61** | Lwilomelo | Zimbabwe |
| **158** | Lwilomelo | Zimbabwe |
| **160** | Mbala | Zambia |
| **161** | Lwilomelo | Zimbabwe |
| **162** | Mapanzure | Zimbabwe |
| **163** | Musana | Zimbabwe |
| **164** | Litende | Malawi |
| **165** | Luwawa | Malawi |
| **168** | Nyamukwarara | Zimbabwe |
| **169** | Choma | Zambia |
| **170** | Luwawa | Malawi |
| **220** | Murewa | Malawi |
| **238** | Utete-Iringa | Tanzania |
| **260** | Mbeya-Nyoka | Tanzania |

| Group 4<br>Sample ID | Provenance | Country of Origin |
|---|---|---|
| **13** | Lwilomelo | Zimbabwe |
| **16** | Litende | Malawi |
| **20** | Musana | Zimbabwe |
| **21** | Murewa | Malawi |
| **24** | Mapanzure | Zimbabwe |
| **28** | Choma | Zambia |
| **32** | Mapanzure | Zimbabwe |
| **34** | Luwawa | Malawi |
| **37** | Musana | Zimbabwe |
| **38** | Mbala | Zambia |
| **39** | Litende | Malawi |
| **42** | Lwilomelo | Zimbabwe |
| **46** | Mbala | Zambia |
| **47** | Luwawa | Malawi |
| **53** | Litende | Malawi |
| **54** | Murewa | Malawi |
| **55** | Nyamukwarara | Zimbabwe |
| **56** | Choma | Zambia |
| **57** | Litende | Malawi |
| **58** | Luwawa | Malawi |
| **59** | Litende | Malawi |
| **60** | Gombela | Tanzania |
| **62** | Litende | Malawi |
| **63** | Kitwe | Zambia |
| **64** | Iringa | Tanzania |
| **65** | Kasama | Zambia |
| **66** | Choma | Zambia |
| **67** | Gombela | Tanzania |
| **68** | Chipata | Zambia |
| **69** | Choma | Zambia |
| **70** | Chipata | Zambia |
| **71** | Iringa | Tanzania |
| **72** | Mbeya | Tanzania |
| **73** | Mbeya | Tanzania |
| **74** | Chipata | Zambia |
| **75** | Chipata | Zambia |
| **76** | Mbeya | Tanzania |
| **77** | MUAP | Malawi |
| **78** | Kasama | Zambia |
| **79** | Kasama | Zambia |

| Group 4<br>Sample ID | Provenance | Country of Origin |
|---|---|---|
| **80** | Mpwapwa | Tanzania |
| **81** | Chipata | Zambia |
| **82** | Mbeya | Tanzania |
| **83** | Serenje | Zambia |
| **84** | Gombela | Tanzania |
| **85** | Kyela | Tanzania |
| **86** | Mbala | Zimbabwe |
| **87** | Gombela | Tanzania |
| **88** | Iringa | Tanzania |
| **89** | Iringa | Tanzania |
| **90** | Kitwe | Zambia |
| **91** | Gombela | Tanzania |
| **92** | Kitwe | Zambia |
| **93** | Serenje | Zambia |
| **94** | Mpwapwa | Tanzania |
| **95** | Serenje | Zambia |
| **96** | Iringa | Tanzania |
| **97** | Kasama | Zambia |
| **98** | Kitwe | Zambia |
| **99** | Gombela | Tanzania |
| **100** | Zambia | Zambia |
| **101** | Sumbawanga | Tanzania |
| **102** | Kasama | Zambia |
| **103** | Mpwapwa | Tanzania |
| **104** | Mbala | Zambia |
| **106** | Choma | Zambia |
| **107** | Kasama | Zambia |
| **108** | MUAP | Malawi |
| **109** | Mbala | Zambia |
| **110** | Kasama | Zambia |
| **111** | Litende | Malawi |
| **112** | Luwawa | Malawi |
| **113** | Luwawa | Malawi |
| **114** | Luwawa | Malawi |
| **115** | Litende | Malawi |
| **116** | Litende | Malawi |
| **117** | Litende | Malawi |
| **118** | Litende | Malawi |
| **119** | Kasungu-MUAP | Malawi |

| Group 4<br>Sample ID | Provenance | Country of Origin |
|---|---|---|
| **120** | Luwawa | Malawi |
| **121** | Sumbawanga | Tanzania |
| **122** | Litende | Malawi |
| **123** | Gombela | Tanzania |
| **124** | Mozambique | Mozambique |
| **125** | Kyela | Tanzania |
| **126** | Choma | Zambia |
| **127** | Serenje | Zambia |
| **128** | Choma | Zambia |
| **129** | Mbeya | Tanzania |
| **130** | Zambia | Zambia |
| **131** | Mozambique | Mozambique |
| **132** | Zambia | Zambia |
| **133** | Chipata | Zambia |
| **134** | Kitwe | Zambia |
| **135** | Kitwe | Zambia |
| **136** | Serenje | Zambia |
| **137** | Zambia | Zambia |
| **138** | Choma | Zambia |
| **139** | Kitwe | Zambia |
| **140** | Mozambique | Mozambique |
| **141** | Iringa | Tanzania |
| **142** | Sumbawanga | Tanzania |
| **143** | Kyela | Tanzania |
| **144** | Iringa | Tanzania |
| **145** | Mozambique | Mozambique |
| **146** | Mozambique | Mozambique |
| **147** | Serenje | Zambia |
| **148** | Mozambique | Mozambique |
| **149** | Mpwapwa | Tanzania |
| **150** | Mozambique | Mozambique |
| **151** | Mozambique | Mozambique |
| **152** | Kasama | Zambia |
| **153** | Choma | Zambia |
| **154** | Gombela | Tanzania |
| **155** | Chipata | Zambia |
| **156** | Mbeya | Tanzania |
| **157** | Mbala | Zambia |
| **159** | Mbala | Zambia |
| **166** | Domboshawa | Zimbabwe |
| **167** | Murewa | Zimbabwe |
| **171** | Kasungu | Malawi |
| **172** | Phalombe | Malawi |
| **173** | Luwawa | Malawi |
| **174** | Mpwapwa | Tanzania |
| **175** | Kasungu | Malawi |
| **176** | Litende | Malawi |
| **177** | Phalombe | Malawi |
| **178** | Phalombe | Malawi |
| **179** | Litende | Malawi |
| **180** | Phalombe | Malawi |
| **181** | Choma | Zambia |
| **182** | Kasungu | Malawi |
| **183** | Kasungu | Malawi |
| **184** | Litende | Malawi |
| **185** | Kasungu | Malawi |
| **186** | Serenje | Zambia |
| **187** | Serenje | Zambia |
| **188** | Choma | Zambia |
| **189** | Mbeya | Tanzania |
| **190** | Kasungu | Malawi |
| **191** | Mbeya-Nyoka | Tanzania |
| **192** | Mbeya-Nyoka | Tanzania |
| **193** | Mbeya-Nyoka | Tanzania |
| **194** | Gombela-Songea | Tanzania |
| **195** | Sumbawanga | Tanzania |
| **196** | Luwawa | Malawi |
| **197** | Serenje | Zambia |
| **198** | Murewa | Malawi |
| **199** | Luwawa | Malawi |
| **200** | Iringa | Tanzania |
| **201** | Chipata | Zambia |
| **202** | Mapanzure | Zimbabwe |
| **203** | Gombela-Songea | Tanzania |
| **204** | Mbeya-Nyoka | Tanzania |
| **205** | Sumbawanga | Tanzania |
| **206** | Mapanzure | Zimbabwe |
| **207** | Kasungu | Malawi |

| Group 4<br>Sample ID | Provenance | Country of Origin |
|---|---|---|
| **208** | Murewa | Malawi |
| **209** | Kasungu | Malawi |
| **210** | Serenje | Zambia |
| **211** | Mbeya | Tanzania |
| **212** | Chipata | Zambia |
| **213** | Kasungu | Malawi |
| **214** | Kasungu | Malawi |
| **215** | Mbeya-Kyela | Tanzania |
| **216** | Phalombe | Malawi |
| **217** | Mbeya-Kyela | Tanzania |
| **218** | Serenje | Zambia |
| **219** | Choma | Zambia |
| **221** | Mapanzure | Zimbabwe |
| **222** | Mapanzure | Zimbabwe |
| **223** | Mapanzure | Zimbabwe |
| **224** | Kyela | Tanzania |
| **225** | Kasungu | Malawi |
| **226** | Litende | Malawi |
| **227** | Murewa | Malawi |
| **228** | Serenje | Zambia |
| **229** | Kasungu | Malawi |
| **230** | Mapanzure | Zimbabwe |
| **231** | Choma | Zambia |
| **232** | Chipata | Zambia |
| **233** | Kasungu | Malawi |
| **234** | Murewa | Zimbabwe |
| **235** | Luwawa | Malawi |
| **236** | Nyamukwarara | Zimbabwe |
| **237** | Mpwapwa | Tanzania |
| **239** | Gombela-Songea | Tanzania |
| **240** | Murewa | Zimbabwe |
| **241** | Chipata | Zambia |
| **242** | Murewa | Zimbabwe |
| **243** | Sumbawanga | Tanzania |
| **244** | Gombela-Songea | Tanzania |
| **245** | Murewa | Zimbabwe |
| **246** | Sumbawanga | Malawi |
| **247** | Choma | Zambia |

| Group 4<br>Sample ID | Provenance | Country of Origin |
|---|---|---|
| **248** | Mpwapwa | Tanzania |
| **249** | Mpwapwa | Tanzania |
| **250** | Nyamukwarara | Zimbabwe |
| **251** | Gombela-Songea | Tanzania |
| **252** | Utete-Iringa | Tanzania |
| **253** | Mbeya-Kyela | Tanzania |
| **254** | Mapanzure | Zimbabwe |
| **255** | Serenje | Malawi |
| **256** | Litende | Malawi |
| **257** | Kasungu | Malawi |
| **258** | Luwawa | Malawi |
| **259** | Mbeya-Nyoka | Tanzania |
| **261** | Choma | Zambia |
| **262** | Mbeya-Nyoka | Tanzania |
| **263** | Litende | Malawi |
| **264** | Choma | Zambia |
| **265** | Chipata | Zambia |
| **266** | Nyamukwarara | Zimbabwe |
| **267** | Nyamukwarara | Zimbabwe |
| **268** | Phalombe | Malawi |
| **269** | Phalombe | Malawi |
| **270** | Phalombe | Malawi |
| **271** | Serenje | Zambia |
| **272** | Serenje | Zambia |
| **273** | Serenje | Zambia |
| **274** | Sumbawanga | Tanzania |
| **275** | Sumbawanga | Tanzania |
| **276** | Sumbawanga | Tanzania |
| **277** | Sumbawanga | Tanzania |
| **278** | Utete-Iringa | Tanzania |
| **279** | Utete-Iringa | Tanzania |
| **280** | Choma | Zambia |
| **281** | Chipata | Zambia |
| **282** | Gombela-Songea | Tanzania |
| **283** | Gombela-Songea | Tanzania |
| **284** | Gombela-Songea | Tanzania |

| Group 4<br>Sample ID | Provenance | Country of Origin |
|---|---|---|
| **285** | Kasungu | Malawi |
| **286** | Chipata | Zambia |
| **287** | Kasungu | Malawi |
| **288** | Kasungu | Malawi |
| **289** | Kasungu | Malawi |
| **290** | Chipata | Zambia |
| **291** | Litende | Malawi |
| **292** | Litende | Malawi |
| **293** | Litende | Malawi |
| **294** | Luwawa | Malawi |
| **295** | Luwawa | Malawi |
| **296** | Mapanzure | Zimbabwe |
| **297** | Chipata | Zambia |
| **298** | Mapanzure | Zimbabwe |
| **299** | Mapanzure | Zimbabwe |
| **300** | Mapanzure | Zimbabwe |
| **301** | Chipata | Zambia |
| **302** | Kyela | Tanzania |
| **303** | Kyela | Tanzania |
| **304** | Chipata | Zambia |
| **305** | Mbeya-Nyoka | Tanzania |
| **306** | Mbeya-Nyoka | Tanzania |
| **307** | Mbeya-Nyoka | Tanzania |
| **308** | Mpwapwa | Tanzania |
| **309** | Mpwapwa | Tanzania |
| **310** | Murewa | Zimbabwe |
| **311** | Murewa | Zimbabwe |
| **312** | Murewa | Zimbabwe |
| **313** | Murewa | Zimbabwe |
| **314** | Chipata | Zambia |
| **315** | Choma | Zambia |
| **316** | Choma | Zambia |
| **317** | Iringa | Tanzania |
| **318** | Iringa | Tanzania |
| **319** | Iringa | Tanzania |
| **320** | Iringa | Tanzania |
| **321** | Kasama | Zambia |
| **322** | Kasungu | Malawi |
| **323** | Kyela | Tanzania |
| **324** | Kyela | Tanzania |
| **325** | Litende | Malawi |
| **326** | Luwawa | Malawi |
| **327** | Luwawa | Malawi |
| **328** | Mbala | Zambia |
| **329** | Mpwapwa | Tanzania |
| **330** | Mpwapwa | Tanzania |
| **331** | Mpwapwa | Tanzania |
| **332** | MUAP | Zambia |
| **333** | Serenje | Zambia |
| **334** | Serenje | Zambia |
| **335** | Serenje | Zambia |
| **336** | Serenje | Zambia |
| **337** | Serenje | Zambia |
| **338** | Zambia | Zambia |
| **339** | Zambia | Zambia |
| **340** | Zambia | Zambia |
| **341** | Zambia | Zambia |
| **342** | Luwawa | Zambia |