GENETIC CHARACTERIZATION OF SELECTED *Moringa oleifera* PROVENANCES FROM THE KENYAN COASTAL REGION

Sarah Kwamboka Ondieki

A thesis submitted in partial fulfillment of the requirements for the award of the Degree of Master of Science in Molecular biology of Masinde Muliro University of Science and Technology.

November, 2024

DECLARATION

This thesis is my own original research prepared with no other than indicated sources and has not been presented elsewhere for any other award.

Signature……. Date……

Sarah Kwamboka Ondieki

Reg: SBB/G/01-53140/2018

CERTIFICATION

We the undersigned certify that we have read and hereby recommend for acceptance of Masinde Muliro University of Science and Technology a thesis entitled; **'***Genetic Characterization of Selected Moringa Oleifera Provenances from The Kenyan Coastal Region'*.

Signature…… Date……

Dr. Jackson Cheruiyot Korir

School of Natural Sciences

Masinde Muliro University of Science and Technology

Signature…… Date……

Dr. Carol Hunja

School of Science and Computing

South Eastern Kenya University

COPYRIGHT

This thesis is copyright materials protected under the Berne Convection, the copyright Act 1999 and other international and national enactments in that behalf, on intellectual property. It may not be reproduced by any means in full or in part except for short extracts in fair dealing so for research or private study, critical scholarly review or discourse with acknowledgment, with written permission of the Director of Post-Graduate Studies on behalf of both the author and Masinde Muliro University of Science and Technology.

DEDICATION

This work is dedicated to my parents Dr. Peterson Ondieki and Mrs. Hellen Ondieki, my loving husband Geoffrey Nyambuti, my beloved sons Leeroy and Owen and my lovely daughter Briella.

ACKNOWLEDGEMENT

Foremost, my utmost gratitude goes to God for granting me good health and life. Secondly, I sincerely acknowledge my supervisors, Dr. Jackson C. Korir of the Department of Biological Sciences, Masinde Muliro University of Science and Technology (MMUST) and Dr. Carol Hunja of the Department of Life Sciences, South Eastern Kenya University (SEKU) for guidance and mentorship. Thank you for your unwavering support towards the actualization of this thesis.

My sincere appreciation goes to the National Research Fund (NRF) for funding this research and SEQART Africa, based at International Livestock and Research Institute (ILRI), Kenya for providing the laboratory and equipment for conducting this study.

To all my dearest friends who have continually encouraged and held my hand, provided some needed insight on my work and gave me a listening ear continually, I appreciate you from the bottom of my heart. To my loving husband, Dr. Geoffrey who ensured I never lacked the fees and persistently reminded me to keep focus, I am grateful beyond measure for your support, thank you.

To my father, Dr. Peterson Ondieki, who has been my mentor and a great support throughout my education and a shoulder to lean on thank you. Your encouragement and insight is well appreciated. I also give my thanks to my mother, Hellen, my brothers and my sister Dr. Moraa for your support through my academic journey.

ABSTRACT

Globally, moringa oleifera is a major economic crop. Traditional medicine has used it to treat diabetes, inflammation, cancer, bacterial, viral, and fungal infections, joint pain, and cardiovascular health. Moringa has been used for food, cattle feed, and water purification. Cross-pollinated moringa has great species variation. Rapid growth and adaptation to varied climates have led to widespread production of M. oleifera in Coastal and Eastern Kenya. Identifying useful genotypes and improving drumstick tree cultivars requires studying its genetic diversity. Genetic diversity patterns must be understood to improve breeding strategies. Marker-assisted selection for desirable cultivars will boost breeding efficiency. A variety of molecular markers, including AFLP, SSR, RAPD, ISSR, and RFLP, have been used to characterize the Moringa plant. This study used SNPs, a cutting-edge molecular marker. This study meticulously characterized 164 genotypes of Moringa oleifera sourced from 17 provenances in Coastal Kenya through the application of genome sequencing, employing the genotyping by sequencing (GBS) technique. The analysis also discerned polymorphisms (SNP Calling) within the chosen genotypes and conducted investigations into population genetics. DNA was extracted from homogenized and crushed Moringa leaves. After reducing genomic complexity, library preparation includes barcoding, electrophoresis, pooling, and sequencing library preparation, followed by sequencing. Sequencing and SNP mining used Illumina Hiseq 2500 next-generation sequencing. The process of binning was implemented to eliminate the noise. *In silico* assembly of sequence contigs was undertaken. BLAST facilitated the alignment of the sequences with the Moringa reference genome. The analysis of data was conducted utilizing the DArT R and KDCompute platforms. The genetic characterization was accomplished through the implementation of cluster analysis, principal coordinate analysis (PCoA), three-dimensional plotting, and the construction of a phylogenetic tree. The computation of molecular variance analysis (AMOVA) and principal component analysis was conducted utilizing the genetic distance matrix. The SNP calling process was executed using the SNP caller algorithm provided by Illumina within the CASAVA software framework. A total of 20,921 SNPs were identified, exhibiting an average call rate of 0.82. The mean polymorphism content (PIC) for the SNPs was calculated to be 0.24, while the reproducibility rate was determined to be 0.98. The cluster analysis clearly indicated that the genotypes were organized into four distinct clades. The similarity coefficient derived from hierarchical clustering suggested that the genotypes exhibited reduced variability. A 3D plot was generated interactively utilizing DArT R, specifically through the Adegenet package. A phenetic tree was constructed employing a neighbor-joining methodology utilizing DArT R. In the analysis of population genetics, the F statistic (Fst) was calculated using the StAMPP package in DArT R. The results indicated that Gede and Samburu displayed the lowest heterozygosity/correlation, recorded at 0.0003, while Pwani University and Samburu demonstrated the highest gene correlation, reaching a value of 0.37. The Euclidean metric served as a means of quantifying distance, revealing that the mean distance among the populations was 33.024. The analysis of molecular variance (AMOVA) indicated a modest variation of 2.55% within the population and a slightly higher variation of 2.73% among the populations. The notable resemblance among the genotypes may be ascribed to the Moringa plants across different provenances sharing a common lineage. Considering the prevalence of SNPs and their role as a source of allele variations, this research may facilitate the identification of associations between gene allelic forms and phenotypes, thereby allowing for the connection of alleles to advantageous traits such as rapid growth and increased seed production.

.

TABLE OF CONTENTS

PLAGIARISM STATEMENT..... ii

DECLARATION..... iii

CERTIFICATION..... iii

COPYRIGHT..... iv

DEDICATION..... v

ACKNOWLEDGEMENT..... vi

ABSTRACT..... vii

LIST OF FIGURES..... xii

LIST OF TABLES..... xiii

ACRONYMS AND ABBREVIATION..... xiv

CHAPTER ONE..... 1

INTRODUCTION..... 1

1.0 Background of the study..... 1

1.1 Statement of the problem..... 4

1.2 Justification of The Study..... 4

1.3 Significance of the study..... 5

1.4 Objectives of The Study..... 6

1.4.1 General Objective..... 6

1.4.2 Specific Objectives..... 6

1.4.3 Research Questions..... 6

CHAPTER TWO..... 7

LITERATURE REVIEW..... 7

2.1 Introduction..... 7

2.2 Utilization of *M. oleifera*..... 8

2.2.1 Human Consumption of *M. oleifera*..... 8

2.2.2 Industrial uses of *M. oleifera* oil..... 9

2.2.3 Water Purification..... 10

2.2.4 *M. oleifera* as a Forage Plant..... 11

2.3 Genetic Diversity of *Moringa oleifera*..... 12

2.4 SNPs as a Molecular Marker..... 14

2.5 Importance of SNP in Crop Improvement..... 18

2.6 Genotyping by Sequencing..... 20

2.7 The Utilization of GBS in Plant Breeding..... 22

2.8 *In silico* Technologies..... 25

CHAPTER THREE..... 29

MATERIALS AND METHODS..... 29

3.1 Study Area..... 29

3.2 Experimental Design..... 30

3.3 Sampling Method..... 31

3.4 Genotyping by Sequencing (GBS)..... 32

3.4.1 DNA Isolation..... 32

3.4.2 Agarose Gel Electrophoresis..... 33

3.4.3 Library Preparation..... 33

3.5 Data Analysis..... 34

CHAPTER FOUR..... 37

4.1 Characterization of *Moringa oleifera* Provenances..... 37

4.1.1 Simple Cluster Analysis..... 37

4.1.2 Hierarchical Cluster Analysis..... 41

4.1.3 Principal Coordinate Analysis Results..... 46

4.1.4 3D Plot Result..... 49

4.2 To Identify Polymophisms in the Selected Genotypes..... 50

4.3. Population Genetics..... 55

4.3.1 Bar plot on the Number of Individuals per Population..... 55

4.3.2 F Statistics..... 55

4.3.3 Population Euclidean Distances..... 58

4.3.4 Analysis of Molecular Variance (AMOVA)..... 58

4.3.5 Neighbor-Joining Tree..... 59

CHAPTER FIVE..... 61

DISCUSSION..... 61

5.1 Characterization of *Moringa oleifera*.....61

5.1.1 Clustering Analysis.....61

5.1.2 Hierarchical Clustering.....62

5.1.3 Principal Coordinate Analysis.....66

5.1.4 3D Plot.....67

5.2 To Identify Polymorphisms in the Selected Varieties.....67

5.3 Population Genetics Analysis/ Study.....70

5.3.1 Population Euclidean Distances.....70

5.3.2 F Statistics.....71

5.3.3 Neighbor Joining Tree.....71

5.3.4 Analysis of Molecular Variance (AMOVA).....72

CHAPTER SIX.....74

CONCLUSSION AND RECOMMENDATION.....74

6.1 Conclusion.....74

6.2 Recommendation.....75

REFERENCES.....76

APPENDIX.....88

Appendix I. Publication.....88

Appendix II.Single Nucleotide Polymorphisms of Selected Moringa.....89

Appendix III. Sequencing and BLAST.....91

LIST OF FIGURES

Figure 2.1: Essential uses of various M. oleifera parts 9

Figure 2. 2: Schematic overview of steps in genotyping by sequencing (GBS) library construction, sequencing, and analysis 27

Figure 2.3: In silico analysis of SNPs 34

Figure 3. 1: Location of research site in SEKU, Kitui County, Kenya 35

Figure 3. 2: : Location of Seed collection areas (provenances) 36

Figure 4. 1: Phylogenetic tree showing diversity and variance between M. oleifera genotype 47

Figure 4. 2: Hierarchical clustering dendrogram with au and bp value 53

Figure 4.3: Principal Coordinate Analysis plot 55

Figure 4.4: Three dimensional plot 57

Figure 4.5: Barplot showing populations/provenances 63

Figure 4. 6: Phenetic tree for the population using Neighbor-joining approach from DArT 68

LIST OF TABLES

Table 3. 1: No of Trees sampled per provenance.....37

Table 4. 1: A summary of the genotypes obtained in Fig. 1 by simple cluster analysis.....46

Table 4. 2: A summary of genotypes obtained by Hierarchical clustering dendrogram (Figure 2) with similarity coefficients indicated.....52

Table 4. 3: Principal Coordinate Analysis (PCoA) showing the number of genotypes per coordinates summarized from Figure 9.....56

Table 4. 4: DArT SNP marker quality and diversity.....59

Table 4. 5: Shows the SNP position.....60

Table 4. 6: Extrapolation showing part of the DArT SNPs.....61

Table 4. 7: Silico DArT data.....61

Table 4. 8: Fst table showing the distances in centiMorgans.....64

Table 4. 9: Population genetic distances using Euclidean as a distance measure.....65

ACRONYMS AND ABBREVIATION

AFLP - Amplified Fragment Length Polymorphism

PCR - Polymerase Chain Reactions

SNP - Single Nucleotide Polymorphisms

WGS - Whole Genome Sequencing

NGS - Next Generation Sequencing

RAPD - Random Amplified Polymorphic DNA

ISSR - Inter Simple Sequence Repeats

SSR - Simple Sequence Repeat

SRAP - Sequence Related Amplified Polymorphism

LD - Linkage disequilibrium

eSNP - electronic SNP

EST - Expressed Sequence Tag

WGAS- Whole Genome Association Studies

QTL - Quantitative Trait Loci

GBS - Genotyping by Sequencing

RRS - Reduced Representation Sequence

DnaSP- DNA Sequence Polymorphism Platform

MAS - Marker-assisted Selection

GS - Genome Sequencing

GWAS- Genome Wide Association Studies

PAV - Presence Absence Variation

AMOVA -Analysis of Molecular Variance

CHAPTER ONE

INTRODUCTION

1.0 Background of the study

Moringa oleifera is a member of the *Moringa*ceae, a monogenetic family of shrubs and trees (Ramachandran et al., 1980); (Stephenson & Fahey, 2004). *M. oleifera* is the most commonly cultivated *M. oleifera* plant, with a diploid chromosome (2n = 28). It's used as a medicinal plant, human food, animal fodder, antimicrobial agent, and biofuel resource (Gopalakrishnan et al., 2016). Its resistance to abiotic stresses makes it an ideal tree for arid-zone cultivation (Sciences, 2015). *M. oleifera* is a nutritionally and medicinally important *Moringa* species that is grown in over 60 countries around the world (Leone et al., 2015). From India, *M. oleifera* was introduced to a number of tropical and subtropical countries (Nouman et al., 2014). Further research into the conservation and enhancement of these important species was suggested in these reviews.

Drumstick (common name) is an essential food source, containing proteins, vitamins, minerals, and antioxidant properties that help protect cells from damage (Nautiyal and Venkataraman 1987). Pods are a good source of calcium, phosphorus, and vitamin C in terms of nutrition(Fuglie, 2005). It is also utilized in feeding programs to lower malnutrition in a number of third-world nations. Numerous compounds present in ***M. oleifera*** have been utilized for centuries in Siddha therapy due to their hypotensive and anticancer effects (Soller and Beckmann 1983). According to Jonathan and Weden (1990), fresh or dried leaves, blossoms, and immature pods are all utilized extensively in cooking and are highly valued in the modern diet. *M. oleifera* produces a very nutritious edible oil that is low in

polyunsaturated fatty acids and high in oleic acid (N73%). The crop is still underutilized despite its worth (Pandey et al., 2011).

The drumstick is a crop characterized by extensive cross-pollination, resulting in a diverse genetic composition derived from multiple species. Achieving genetic enhancement and the development of superior cultivars necessitates a comprehensive understanding of the genetic diversity present among advanced breeding lines. The integration of morphological research with biomolecular and molecular analysis offers a comprehensive insight into the variability present in advanced breeding lines. In various crops, markers including RFLPs, RAPDs, AFLPs, SSRs, and ISSRs (Powell et al. 1996) are frequently employed to assess plant genetic diversity at the genomic level. Single nucleotide polymorphisms represent the latest advancements in molecular markers and are therefore preferred. This arises from the characteristic that SNP assays do not necessitate the size separation of DNA, can be automated in an assay-plate format or on microchips, and are more readily identifiable in the majority of single-copy regions of the genome compared to SSRs (Rafalski, 2002). Classification, conservation, and cultivar growth necessitate a comprehensive grasp of genetic variation. Saini et al. (2013) employed molecular markers, including RAPD, ISSR, and the Cytochrome P450 marker system, to assess genetic variation among common breeding lines of *M. oleifera* L. The application of morphological and cytological techniques for assessing genetic diversity in breeding lines is not merely labor-intensive; it also carries the potential for errors across multiple dimensions (Monica et al. 2010). Conventional methods of crop breeding aimed at various traits could gain from a more profound comprehension of genetic diversity through the application of specific markers (Bretting and lead into Widrelechner 1995). In a breeding program for plants, it is crucial to assess the

uniqueness and distinctness of the phenotype, alongside identifying the genetic constitution of the genotype (Davierwala et al. 2000). Given their role as the foundation of genetic material for the enhancement of elite varieties, it is essential to assess their genetic diversity (Makkar & Becker, 1997). An expedited breeding program would benefit from the implementation of marker-assisted germplasm selection to enhance the desirable traits.

The detection of polymorphism at the genomic level reveals the presence of variation and polymorphism, facilitating the precise identification of genetic diversity within a population (Wu et al., 2010); (S.-T. He et al., 2021). Consequently, the application of genetic markers for the enhancement of crops has gained notable traction in recent years (GarciaI et al. 2004). Marker-assisted selection (MAS) refers to the application of molecular markers for the indirect selection of desirable traits in crops. MAS has been extensively employed in plant breeding to enhance crop yield, quality, and resilience against biotic and abiotic stresses, serving as a crucial and effective methodology. Genotyping-by-sequencing (GBS) represents an innovative approach in marker-assisted selection that has the potential to enhance the efficiency of plant breeding and the advancement of crop improvement (J. He et al., 2014).

Ensuring a robust level of genetic diversity stands as a paramount concern for the effective enhancement of *M. oleifera*, given the species' considerable multipurpose value yet its lack of genetic characterization. Due to the complexities of environmental interactions, epistasis, and the largely enigmatic genetic regulation of traits, the morphological markers available for individual identification are limited and do not consistently reflect genetic relationships (Perez et al., 1998). Nonetheless, there exists a paucity of information regarding the genetic characterization of the plant. Consequently, it is essential to conduct a genetic

characterization of the *M. oleifera* plant to pinpoint genes responsible for advantageous traits, including accelerated growth and enhanced yields. In Kenya, the available information regarding the genetic fingerprinting and molecular analysis of *M. oleifera* is quite scarce. Moreover, the entire genome of the plant remains unsequenced. Consequently, this study seeks to undertake a molecular characterization of chosen *M. oleifera* provenances from Coastal Kenya, alongside a functional analysis that connects these polymorphisms to a range of genes associated with desirable cultivar traits.

1.1 Statement of the problem

Moringa oleifera is a very essential plant as it can be utilized as food for humans, fodder for animals, for water purification and is of medical significance. *M. oleifera* as a commercial tree has a range of medicinal, nutritional, and pharmacological properties, which have recently been reviewed (Gopalakrishnan et al., 2016). As the population in Kenya is increasing, arable land becomes scarce and this leads to inadequate food production which may lead to malnutrition of children and lack of enough fodder for animals. Thus, genetic assessment for desirable traits is prerequisite for crop improvement Programme. Since *M. oleifera* is a perennial tree crop that is predominantly cross pollinated in nature, it needs a reliable and fast method for cultivar selection for its crop improvement program (kumar Ganesan et al., 2014). No priori sequencing has been done for the *M. oleifera* plant in Kenya. This study seeks to carry out diversity studies and population genetic analysis for selected *M. oleifera* genotypes from Coastal Kenya. Functional analysis could also be done. Single nucleotide polymorphisms are a virtually limitless source of polymorphic markers for use in high-resolution genetic mapping of traits and association studies based on candidate genes or potentially entire genomes (Rafalski, 2002).

1.2 Justification of The Study

In this research, Single nucleotide polymorphisms were used as molecular markers because SNP assays do not require size separation of DNA, can be automated in an assay-plate format or on microchips, and are easier to locate in most single-copy regions of the genome than SSRs (Rafalski, 2002). Classification, conservation, and cultivar growth all require a thorough understanding of genetic variation. The use of morphological and cytological methods to determine genetic diversity in breeding lines is not only time consuming, but it can also result in errors on several levels (Mahmood et al., 2010).

In a plant, breeding program, evaluating the uniqueness and distinctness of the phenotype, as well as identifying the genotype's genetic constitution, is essential (Davierwala et al. 2000). Since they are the source of genetic material used to enhance elite varieties, it is vital to evaluate their genetic diversity (Makkar & Becker, 1997)**.** A faster-breeding program would profit from the use of marker-assisted germplasm selection for the desirable traits. Polymorphism detection at the genome level indicates the existence of variation and polymorphism, allowing for the effective identification of genetic diversity within a population (Wu et al., 2010). As a result, in recent years, the use of genetic markers for crop improvement has become increasingly prevalent (Rafalski, 2002).

The swift progress in next-generation sequencing (NGS) and the accompanying bioinformatics computing capabilities has led to reduced costs, driven by innovations in sequencing technologies. The recent progress has accelerated whole genome association studies (WGAS) and facilitated the identification of a significant increase in new SNPs across both model and non-model plants. Single nucleotide polymorphisms (SNPs) represent the most prevalent marker technique employed across various plant species following the

genomic era. This method boasts numerous advantages, such as stability, user-friendliness, low mutation rates, and the capacity for high-throughput genotyping (Morgil et al., 2020).

1.3 Significance of the study

This study was carried out to ascertain the genetic diversity of the selected *Moringa oleifera* varieties for future improvement of its breeding cultivars. This is essential as the identification of the genotype diversity within a population would achieve this. Genetic diversity is important for a healthy population by maintaining different varieties of genes that might improve the Moringa plant as it is a very important plant with several uses. Genetic diversity within and between the studied provenances will allow the selection of superior genotypes, which can then be utilized for development of genetic stock for hybridization programs.

1.4 Objectives of The Study

1.4.1 General Objective

To characterize selected *Moringa oleifera* provenances from Coastal Kenya using molecular analysis.

1.4.2 Specific Objectives

i. To characterize the selected *Moringa oleifera* provenances by genotyping by sequencing.
ii. To identify polymorphisms in the selected varieties.
iii. To undertake population genetic studies on the selected *M. oleifera* varieties.

1.4.3 Research Questions

i. What is the genetic diversity of the selected *Moringa oleifera* provenances by genotype sequencing?

ii. What are polymorphisms of the selected varieties?

iii. What are population genetic studies of the selected *M. oleifera* varieties?

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction.

Moringa oleifera is a unique monogenetic shrub and tree that has its origins in the northwest region of India, specifically in the cities of Agra and Oudh, situated south of the Himalayan Mountains. Despite the designation "Shigon" for *M. oleifera* being recorded in the first century A.D. within the "Shushruta Sanhita," evidence exists indicating that the tree was cultivated in Ancient India (Foidl et al., 2001). Given that the seeds possessed edible oil, they were utilized by the Indians for medicinal purposes (Rayan, 2019). The general populace likely recognized its significance as a source of fodder or as a vegetable (Ramachandran et al., 1980). This particular plant has the capacity to thrive at elevations reaching up to 1,000 meters above sea level in its natural habitat. It flourishes on rocky outcroppings; however, it is more prevalent in pastureland and river basins. It has now disseminated throughout the Middle East and nearly the entire tropical belt (Makkar & Becker, 1997). In the early years of the twentieth century, *M. oleifera* made its introduction from India to Eastern Africa. During the 1920s, it was initially introduced as a living barrier and decorative flora. As an uncultivated species, it is esteemed for its resilience to drought and disease. The elevated protein levels found in the leaves, twigs, and roots, alongside the substantial protein and oil concentrations in the seeds, the extensive array of distinctive polypeptides present in the seeds capable of binding to various moieties, the presence of growth factors within the leaves, and the plant's significant sugar and starch content exemplify these characteristics (Foidl et al., 2001).

2.2 Utilization of *M. oleifera*

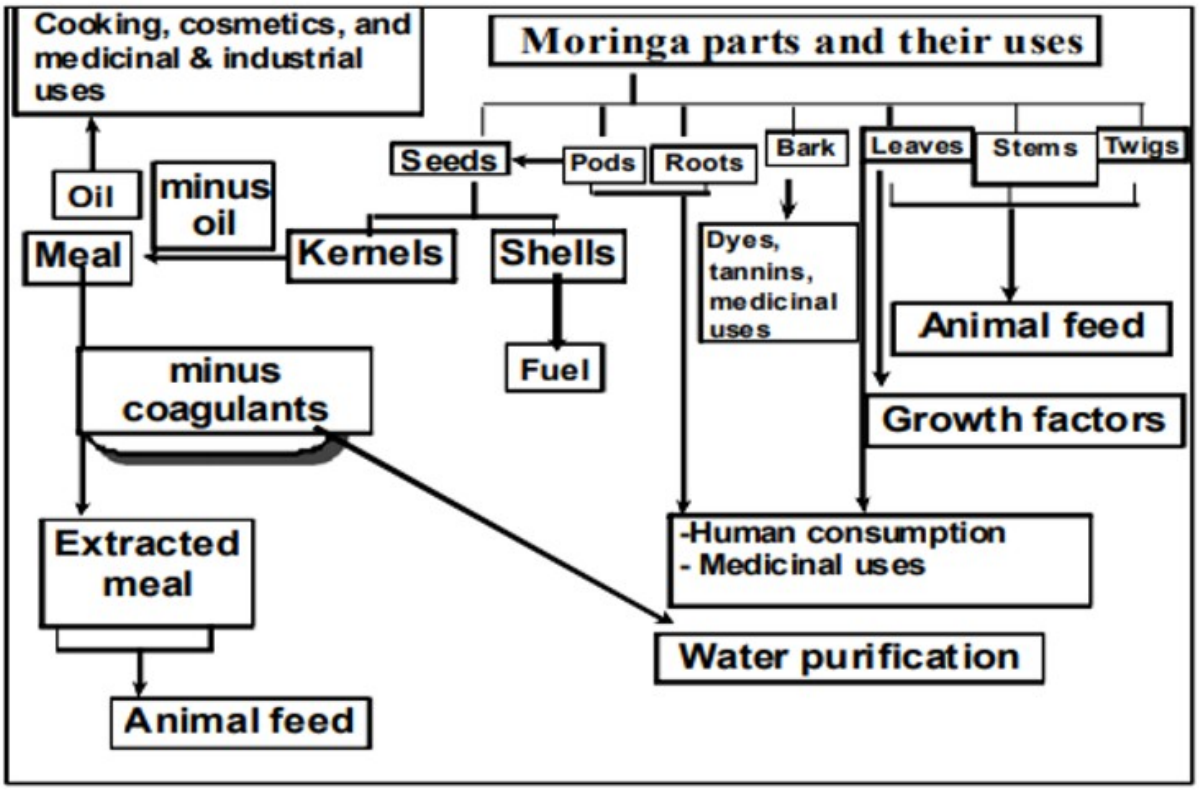


Figure 2. 1: Essential uses of various M. oleifera parts

(Anwar et al., 2007).

2.2.1 Human Consumption of *M. oleifera*

The tender foliage is rich in nutrients and can therefore be prepared and consumed similarly to spinach, or incorporated into soups and salads (Mahmood et al., 2010).They are rich in provitamin A, vitamins B and C, minerals—particularly iron—and the sulphur-containing amino acids methionine and cysteine (Gopalakrishnan et al., 2016). The amino acid composition of the leaf protein exhibits a remarkable balance. *M. oleifera* is replete with a

significant array of essential nutrients and antinutrients throughout its entirety. The leaves of *M. oleifera* are rich in essential minerals, including calcium, potassium, zinc, magnesium, iron, and copper. Vitamins including beta-carotene, folic acid, pyridoxine, and nicotinic acid, as well as vitamins C, D, and E, are all present (Smarandache & Christianto, 2019). Among the various compounds present, one can find anti-cancerous agents including glucosinolates, isothiocyanates, glycoside compounds, and glycerol-1-9-octadecanoate. Additionally, a range of phytochemicals is represented, encompassing tannins, sterols, terpenoids, flavonoids, saponins, anthraquinones, alkaloids, and reducing sugars. The leaves of *M. oleifera* are characterized by their low caloric content, making them a suitable addition to the dietary regimen of individuals dealing with obesity (Gopalakrishnan et al., 2016). Immature green pods offer a delightful culinary experience when subjected to boiling, resembling the enjoyment derived from consuming green beans. When the pods can be swiftly divided without the presence of any discernible fiber strands, they are deemed suitable for human consumption. These contain a significant amount of free leucine. It is advisable to boil the seeds for a brief period prior to consumption in order to eliminate the delicate transparent hull and to properly drain the water. The powdered seeds serve as a means to enhance the flavor profile of sauces. The roots of young plants may be desiccated and pulverized to create a piquant seasoning base reminiscent of horseradish. The *M. oleifera* tree is thus commonly known as the "Horseradish Tree." (Foidl et al., 2001). One may also create a flavorful hot sauce from the roots by simmering them in vinegar. Once delicately blanched, the flowers may be consumed raw, serving as a delightful enhancement to salads. The resin extracted from the trunk of the tree can additionally serve as a thickening agent for sauces (Devarai, 2005).

2.2.2 Industrial uses of *M. oleifera* oil.

The de-hulled seed (kernel) exhibits an oil content of 42 percent, characterized by a vibrant yellow hue (Adegbe et al., 2016). Due to its minimal tendency to degrade and lose its desirable qualities, it serves as an effective lubricant for precision instruments such as timepieces. (Ramachandran et al., 1980). It may also serve as a culinary oil for vegetables (Foidl et al., 2001). The oil's capacity to absorb and retain volatile compounds renders it advantageous in the perfume industry for the stabilization of fragrances (Anwar & Bhanger, 2003). The concentration of free fatty acids varies between 0.5 and 3%. *M. oleifera* seed oil comprises 13 percent saturated fatty acids and 82 percent unsaturated fatty acids. Its concentration of oleic acid is markedly elevated at 70%, in contrast to other vegetable oils that typically contain approximately 40% oleic acid (Adegbe et al., 2016).

2.2.3 Water Purification.

Moringa oleifera presents a promising natural and environmentally friendly alternative to chlorine for water disinfection, alongside its notable health benefits and coagulant characteristics (Faurie et al., 2018). The seeds of *M. oleifera* possess an oil content ranging from 30 to 42 percent, while the press cake generated as a by-product of the oil extraction process is characterized by its substantial protein content. Roughly 1% of these proteins exhibit activity as cationic polyelectrolytes, with molecular weights spanning from seven to 17 K Dalton. Consequently, this protein could serve as a non-toxic organic polypeptide for the sedimentation of mineral particles and organics in the purification of drinking water, the cleaning of vegetable oil, and the sedimentation of fibers within the juice and beer industries.

Consequently, it functions as a significant coagulant, creating natural connections among the colloid particles (Zaku et al., n.d.). As noted by Camacho et al. (2017) and Baptista et al. (2015), *M. oleifera* may be employed alongside conventional water treatment techniques or independently, contingent upon the intended application of the product and the characteristics of the water source. They also found that it can serve to remove inorganic coagulants and possesses the potential for renewability. The water-soluble, positively charged proteins found in *M. oleifera* function as an effective coagulant in the treatment of water and wastewater. Crushed seeds have historically served as indicators of murky waters in households across Sudan and Kenya. Furthermore, they have been successfully employed in both pilot and full-scale wastewater treatment operations in Malawi, demonstrating their relevance in the interaction of flocculation and filtration methods for waters with relatively low turbidity (Faurie et al., 2018). Conversely, industrial coagulants like alumina may pose certain hazards. Their appropriate application requires the availability of skilled personnel, a resource that is predominantly absent in many developing nations (Anwar et al., 2007). Moreover, these industrial coagulants incur substantial expenses and represent a considerable burden on the hard currency reserves of developing nations. The characteristics of natural polypeptides sourced from *M. oleifera* seeds have been acknowledged in China for centuries (Bhatia et al., 2007).

The British colonization of India facilitated the effective dissemination of this knowledge globally. It has demonstrated notable efficacy in Egypt and Sudan for the purification of Nile water intended for consumption (Foidl et al., 2001). The desiccated seeds are meticulously ground into a fine powder following the removal of their wings. The powder is combined

with water, stirred for approximately five minutes, and subsequently filtered through a woven cloth to obtain pure water after nearly an hour. In contrast, impurities can be effectively coagulated by immersing a cloth containing the seed powder in water for an extended period, such as overnight. The seeds are subsequently extricated from the fabric, and the filtered liquid is poured off, leaving the coagulated matter settled at the bottom. Colloids can be extracted with an efficiency of up to 99 percent. In the case of moderately polluted water, a single seed suffices for each litre; conversely, for water that is extremely contaminated, two seeds are required (Mahmood et al., 2010).

2.2.4 *M. oleifera* as a Forage Plant

The *M. oleifera* tree, with its remarkable nutritional attributes, serves as an effective fresh forage source for cattle. The leaves are rich in protein, carotene, iron, and ascorbic acid, whereas the pod contains a significant amount of the amino acid lysine (Zaku et al., n.d.). Experiments utilizing *M. oleifera* as a dietary supplement for the purpose of enhancing cattle fattening. Feeding tests were carried out with a cohort of 24 animals. Throughout the day, various animals engaged in grazing upon gamba, a pasture characterized by the presence of certain leguminous plants. Throughout the night, 12 of the animals, organized into 3 groups of 4, were provided ad libitum with freshly cut pasture, while another 12 were given ad libitum access to chopped 35-day-old *M. oleifera*. The *M. oleifera* group exhibited a significantly greater weight gain compared to the group that was sustained on pasture; (Dogra et al., 1975). A notable benefit of *M. oleifera* lies in its superior fresh material productivity per unit area when juxtaposed with other forage crops. *M. oleifera* represents a notably economical and efficient source of forage for cattle (Mendieta-Araica et al., 2011).

The incorporation of *M. oleifera* leaves into cattle feed led to a notable 32 percent increase in daily weight gain. The diets of milk cows were enhanced with an addition of 15 to 17 kilograms of fresh *M. oleifera* leaves daily, resulting in a notable improvement in dairy production by 43%. The enhancement in milk yield reached an impressive 58 percent following the supplementation of 2 kg of dry matter in the feed. The augmentation in milk production reached 65 percent subsequent to the supplementation of 3 kg of dry matter per day (Fahey, 2005). Reflect on the potential advancements that could arise from enhancing milk production in developing countries through this approach (Foidl et al., 2001). It possesses the capacity to alleviate considerable discomfort for those experiencing protein deficiency (Mahmood et al., 2010).

The *M. oleifera* leaves exhibit a significant crude protein content, with extracted and unextracted values recorded at 43.5 and 25.1 percent, respectively, culminating in an actual protein content exceeding 95 percent. Approximately 95% of the total crude protein was found to be located in the rumen or post-rumen, with a significant fraction exhibiting resistance to degradation in the rumen while remaining available for processing in the post-rumen phase. The results suggest that *M. oleifera* leaves, whether extracted or unextracted, serve as appropriate protein supplements for high-yielding dairy cows. The rates of NDF digestion, which indicate the speed at which feed or fiber is fermented in the rumen, were found to be elevated in both the extracted and unextracted M. This suggests that the fiber quality of these leaves is indeed exceptional (Nouman et al., 2014).

2.3 Genetic Diversity of *Moringa oleifera*

M. oleifera is a tree that undergoes cross-pollination, resulting in considerable variation within the genus. The examination of this variability is highly beneficial for the selection and improvement of genotypes. A significant number of studies employed diverse molecular methodologies to explore the genetic diversity of *M. oleifera* (Kumar Ganesan et al., 2014); (Saini et al., 2013).

A multitude of studies examining the genetic diversity of the *M. oleifera* plant have been conducted, including those by da Silva et al. (2010) focusing on specimens from Brazil, Shahzad et al. (2013) analyzing samples from Pakistan, Ojuederie et al. (2013) investigating varieties from Nigeria, Rajalakshmi et al. (2019), and Saini et al. (2013) exploring the genetic makeup of *M. oleifera* in India, as well as Kleden et al. (2017) studying the plant in Indonesia.

An understanding of the genetic diversity of *M. oleifera* plants is essential for enhancing growth, breeding, and industrial propagation. Kumar Ganesan and colleagues (2014) established the essential morphological characteristics of *M. oleifera* plants for the purpose of investigating genetic diversity. Molecular methods can be employed to analyze plant genetic variation, providing a contrast to morphological studies. RAPD, ISSR, AFLP, SSR, and SRAP represent a selection of the most commonly employed molecular markers in the exploration of genetic diversity within plant species. A swift plant-breeding initiative would benefit from the application of marker-assisted germplasm selection to enhance desirable traits. The identification of polymorphism at the genomic level reveals substantial evidence of variation and diversity, facilitating the straightforward detection of genetic diversity within a population (Wu et al., 2010). Consequently, the application of genetic markers for

the enhancement of crops has gained significant traction in recent years (GarciaI et al. 2004). The most prevalent PCR-based molecular markers include restriction fragment length polymorphism (RFLP) as noted by Thomas et al. (2000), simple sequence repeats (SSR) highlighted by Popoola et al. (2017), amplified fragment length polymorphism (AFLP) discussed by Muluvi et al. (2004), as well as microsatellites and mini satellites examined by Saini et al. (2013). Additionally, inter-simple sequence repeats (ISSR) have been explored by Rajalakshmi et al. (2019) and Robiansyah et al. (2015), alongside RAPD as referenced by CHARAN et al. (n.d.), Kumar et al. (2017), Kleden et al. (2017), and Ojuederie et al. (2013).

Understanding the genetic diversity present in advanced breeding lines may subsequently facilitate improved varietal selection and crossbreeding (Kumar et al., 2017). Single nucleotide polymorphisms (SNPs) represent the latest advancements in molecular markers, allowing for the detection of genetic variations in plants with greater efficiency and prominence (Morgil et al., 2020).

2.4 SNPs as a Molecular Marker

Comprehending the mechanisms by which phenotypic variations among organisms or individual members of the same species are inscribed within their genomes represents a significant challenge in contemporary molecular biology. Single nucleotide polymorphisms (SNPs) are regarded as the fundamental drivers of genetic diversity. These factors represent over 80% of all recognized polymorphisms and are therefore considered the principal contributors to phenotypic variation within a species. SNPs possess the capacity to serve as valuable biological markers (Shastry, 2009).

Single nucleotide polymorphisms, commonly referred to as SNPs, appear to represent the most prevalent variations in DNA sequences. While the vast majority of SNPs are situated in noncoding genomic areas, most SNPs found within coding regions do not exert harmful effects on gene products. Single nucleotide polymorphisms play a crucial role in the exploration of genetic variations across diverse species, encompassing both model organisms and economically significant flora and fauna. SNP genotyping has emerged as a promising area of research in recent years, with the establishment of several innovative techniques (Brookes, 1999). Given the intricate relationships between marker SNPs and functional SNPs, their significance as genetic markers is evident, facilitating the identification of functional SNPs. There exists considerable potential for identifying associations between gene allelic forms and phenotypes, especially in the context of multifactorial diseases. Recent investigations into the scope of linkage disequilibrium (LD) within human populations have revealed that extensive regions of elevated LD extend over significantly larger distances than previously anticipated (Shastry, 2009). This finding suggests that the examination of SNP haplotypes, rather than focusing on individual SNPs, represents a significantly more effective approach for associating alleles with traits. The advancements in SNP discovery and assay development, alongside the application of haplotype diversity for association mapping, have reached notable milestones. A multitude of plant species, including maize, exhibit significant polymorphism, facilitating the identification of SNPs. When inbred lines are accessible, it becomes feasible to directly ascertain haplotypes. The thorough examination of the correlation between genetic and physical distance becomes straightforward when employing populations that are optimal for high-resolution genetic mapping. Single nucleotide polymorphisms can be identified

through various methodologies. The re-sequencing of PCR amplicons, whether preceded by pre-screening or not; the discovery of electronic SNPs (eSNPs) within shotgun genomic libraries; and the identification of eSNPs in expressed sequence tag (EST) libraries serve as illustrative examples of these methodologies. The simplest approach to discern SNP polymorphisms involves sequencing DNA segments, which are amplified through PCR, from a diverse group of individuals (Khan et al., 2012).

SNP genotyping typically involves the generation of a distinct DNA product from a designated area of the genome, followed by either direct methods (such as sequencing) or indirect approaches for SNP identification. Consequently, a polymerase chain reaction (PCR) amplification phase of the specific genome fragment is utilized in the vast majority of the methodologies. There exist alternative methodologies, such as isothermal smart amplification and techniques utilizing advanced next-generation whole genome sequencing technologies (Chan, 2009).

Single nucleotide polymorphisms alter gene function by creating a defect in cellular signaling, potentially elucidating the phenotypic transition (A. J. Bokes, 1999). Single nucleotide polymorphisms (SNPs) within gene coding regions can be classified as synonymous, meaning they do not result in changes to the amino acid sequence, or nonsynonymous, which indicates that they do induce alterations in the amino acid structure. Given that the primary sequence of the protein remains intact, synonymous SNPs have traditionally been regarded as lacking significance. In humans, it is believed that SNPs

account for approximately 90% of the variance observed in sequences. Single nucleotide polymorphisms (SNPs) may exist within the genome, manifesting in both coding regions and non-coding sequences. Intergenic sequences, 5′- or 3′-untranslated regions, intronic regions, and associated noncoding regions such as promoters and transcription factor binding sites represent various loci where SNPs may be located beyond coding regions (Ryan Hunt et. al 2009).

The advent of next-generation sequencing technologies has led to remarkable progress in the field of genomics. This text delineates the examination of studies that have adeptly employed next-generation sequencing techniques for extensive SNP identification, alongside the stringent standards for detecting single nucleotide polymorphisms (SNPs) through these methodologies (Chan, 2009).

Advanced sequencing techniques employ highly parallel approaches to simultaneously sequence hundreds of thousands to millions of reads. The four commercialized conventional techniques available in next-generation sequencing include parallel pyrosequencing in picoliter reactors, Illumina sequencing, SOLiD or polony sequencing, and single molecule sequencing. ChIP-seq, transcript sequencing, microRNA discovery, whole genome resequencing, and single nucleotide polymorphism (SNP) discovery represent merely a selection of the most recent applications for these methodologies. The utilization of these technologies for SNP discovery appears to be particularly demanding, as it requires a significant degree of precision (Chan, 2009). In contrast to ChIPseq or transcript sequencing

methodologies, which rely on the identification of read sequences, the process of SNP discovery requires the precise detection of individual base pairs and their variations. Given the high frequency of SNPs, occurring every few hundred to thousand base pairs, any errors in sequencing can result in imprecise SNP detection (Rafalski, 2002). Various methodologies have been devised for SNP genotyping. The approach selected for a particular genotyping endeavor is influenced by the quantity of SNPs (n) and the DNA samples (m) that require genotyping. The Affymetrix SNP GeneChip and Illumina GoldenGate BeadChips assays are well-suited for extensive genome-wide or large-scale projects characterized by a high sample size and a relatively small number of markers. The TaqMan assay stands out as the preferred technology for projects necessitating a small number of SNPs alongside a substantial population, owing to its high throughput capabilities and remarkable accuracy, precision, time efficiency, and cost-effectiveness (Shen et al., 2009).

Recent advancements in sequencing technology have facilitated the identification of single nucleotide polymorphisms (SNPs) and insertions/deletions, which are fundamental to numerous allele variations. The considerable degree of intraspecific nucleotide diversity, coupled with the presence of numerous gene and expressed sequence tag (EST) sequences, facilitates the relatively straightforward discovery of SNPs in various crop species, including corn and soybean (Shastry, 2009). The direct readout of SNP haplotypes is indeed achievable for these species. The evaluation based on haplotypes offers a more comprehensive perspective compared to assessments focused solely on individual SNPs, and it plays a more significant role in analyzing phenotypic associations. A limited number of elite crop germplasm may have recently encountered constraints, resulting in an increase in linkage

disequilibrium (LD) and facilitating the association of SNP haplotypes at candidate gene loci with phenotypes. In instances where sufficient linkage disequilibrium exists, comprehensive genome scans may elucidate genomic regions correlated with fascinating phenotypic traits (Rafalski, 2002). The increasing allure of SNP and indel markers for high-throughput applications, including marker-assisted breeding, EST mapping, and the integration of genetic and physical maps, can be attributed to recent technological advancements. These polymorphisms may serve as fundamental genetic markers that can be located near nearly any gene (Chan, 2009). The application of SNPs in elucidating the interactions between allelic variants of a gene and their corresponding phenotypes holds significant potential. The advancements in SNP discovery and assay development, alongside the application of haplotype diversity for association mapping, have reached noteworthy milestones. Plant genomes offer a range of unique advantages for researchers employing SNP-based technologies. A multitude of plant species, including maize, exhibit a significant level of polymorphism, thereby facilitating the identification of SNPs. Single nucleotide polymorphisms and *indels* represent a nearly boundless reservoir of polymorphic markers, which can be employed in the high-resolution genetic mapping of traits and in association studies that focus on candidate genes or entire genomes (Khan et al., 2012). A deeper understanding of the distribution of linkage disequilibrium and recombination frequencies across plant chromosomes, especially in crop species, is essential. Significant insights can be gleaned from contemporary advancements in the genomics of both humans and animals within this domain. The imminent realization of automated high-density mapping is a testament to the swift progress in technological innovations. Particularly concerning quantitative traits, the challenges of population growth and accurate phenotype

determination, as opposed to marker assays, continue to present significant limitations (Rafalski, 2002).

2.5 Importance of SNP in Crop Improvement

Single nucleotide polymorphisms generate genetic diversity within individuals of the same species and may manifest at varying frequencies throughout the genome in distinct species. Single nucleotide polymorphisms (SNPs) can instigate variations in phenotypic traits among individuals, influencing characteristics such as the color of plants or fruits, their size, the timing of ripening and flowering, as well as factors related to crop quality, grain yield, and resilience against both abiotic and biotic stresses (Morgil et al., 2020). Single nucleotide polymorphisms can lead to modifications in the amino acid sequence within a gene's exon; however, they may also remain silent in their effects. Such occurrences may also transpire in noncoding regions. Single nucleotide polymorphisms (SNPs) can influence the activity of gene promoters, thereby facilitating the synthesis of functional proteins via the process of transcription. Consequently, the identification of functional SNPs within genes and the assessment of their effects on phenotype will assist product developers in gaining a deeper comprehension of the ramifications for gene function (Shastry, 2009). Plant breeding occurs through two primary methodologies: conventional breeding and marker-assisted breeding. Plant breeding serves as the cornerstone for enhancing scientific productivity across the domains of food, feed, and industry, and this role is poised to persist into the future. The conventional breeding process, which entails hybridization among diverse parental lines followed by an extended period of selection (spanning 5 to 15 years) grounded in phenotypic criteria to realize a refined outcome, is a significant factor contributing to its growing

prominence (Mammadov et al., 2012). The swift progress in sequencing technology, including SNP genotyping and genome sequencing, has led to innovative and effective approaches for delineating intricate characteristics and subsequently identifying the genes responsible for them. Initially employed in the realm of human genetics, these techniques are progressively finding application in the field of plant genetics and the enhancement of product growth. The capacity to create experimental populations from germplasm collections and homozygous individuals in plants within a constrained timeframe through the application of these innovative technologies represents a noteworthy advantage. The primary advantages of SNP markers lie in their adaptability and rapidity, alongside the straightforward manner in which they facilitate data management. The presence of only two alleles at each locus facilitates the integration of data from diverse cohorts, thereby enabling the construction of extensive databases comprising this information. Upon thorough assessment of the quality of these results, this will also guarantee consistent allele detection across different genotyping platforms.

The amalgamation of the fusion sequence and SNPs provides an extensive viewpoint on the complete SNP catalog for each species, supported by a high-quality reference genome (Chan, 2009). Single nucleotide polymorphisms, being the most prevalent type of DNA variation, can demonstrate distinctiveness at a genome-wide locus, thus reflecting selective pressures on SNP variants at the target locus and offering significant marker sets for particular germplasm pools. The development of SNP markers has become a conventional approach, particularly in scenarios where a reference genome is accessible, owing to the emergence of technologies that enhance the verification and identification of SNPs. The

application of whole-genome sequencing (GWAS) and de novo sequencing for previously uncharacterized genomes has experienced a significant rise in prominence (Morgil et al., 2020). Single nucleotide polymorphisms (SNPs) play a crucial role in plant genetics, enabling the classification of cis-regulatory variability within a species via allele-specific analysis. They are instrumental in identifying genes associated with complex genetic traits by elucidating their distribution, thereby enhancing our understanding of the species' adaptation in a given environment. This scenario provides an opportunity for researchers to explore the implications of varying conditions, particularly through the application of RNA-seq technology to pinpoint SNPs at the transcriptome level (Mammadov et al., 2012). SNP discovery, whether de novo or reference-based, is frequently undertaken across diverse species, including plants that possess minimal genetic information. The advent of NGS empowers researchers to identify all SNPs while simultaneously gaining insights into genomic location and genotyping within a singular phase. The utilization of RNA-seq data for SNP analysis presents numerous advantages. A multitude of SNPs can be discerned, and the expression levels of countless functional genes exhibiting sequence variations can be simultaneously observed. Furthermore, genotypes can serve as a basis for estimating phenotypes by pinpointing discrepancies in coding regions associated with the biological and agronomic characteristics of plants. It encompasses information pertinent to investigations including gene characterization, measurement of gene expression, and the examination of posttranslational processes. Moreover, recent advancements in technology have facilitated de novo SNP scanning, even in situations where a reference genome sequence for a particular plant variety is lacking. Consequently, the desired product may be produced utilizing solely information from data repositories prior to the commencement of

fieldwork, thereby facilitating the efficient cultivation of a crop. The applications of SNP technology are extensive, serving the needs of evolutionary and molecular geneticists, plant breeders, and various industries, thereby enhancing our comprehension and advancement of crop species (Morgil et al., 2020).

2.6 Genotyping by Sequencing

GBS exemplifies a novel methodology for the identification and genotyping of SNPs within crop genomes and populations, utilizing next-generation sequencing protocols. The GBS method, performed on a single lane of flow cells, involves the digestion of genomic DNA with restriction enzymes, followed by the ligation of a barcode adapter, PCR amplification, and the subsequent sequencing of the amplified DNA pool. To perform a comprehensive analysis and interpretation of GBS datasets, the establishment of bioinformatics pipelines is crucial. GBS has demonstrated its efficacy in advancing genome-wide association research (GWAS), investigating genomic diversity, performing genetic linkage analysis, identifying molecular markers, and supporting genomic selection across a diverse range of plant breeding programs, establishing itself as a leading approach for marker-assisted selection and a resource-efficient technique. Genotyping-by-sequencing exemplifies a novel utilization of next-generation sequencing methodologies focused on the identification and genotyping of single nucleotide polymorphisms to improve crop varieties. Owing to its cost-effective characteristics, GBS offers a compelling strategy for enriching mapping and breeding populations with a comprehensive assortment of SNP markers. Progress in sequencing chemistries and base-calling software will allow NGS technologies to attain higher sequencing throughputs per run, thereby enabling deeper multiplexing while ensuring

a stable average sequencing depth for each sample. GBS has developed into a financially sound alternative to other whole genome genotyping platforms, as the volume and quality of sequence data produced per run continually improve, enabling even higher plexing and lower costs per sample. It is expected that high-density SNP markers obtained from NGS will be widely utilized in GWAS, MAS, and GS. Plant breeders will have the ability to sequence large crop genomes and create highly detailed genetic linkage maps based on breeding populations. In the future, plant breeders may have the capacity to perform marker-assisted selection or genomic selection on innovative germplasm or species through genotyping-by-sequencing, all without the requirement of existing molecular resources. GBS will play a crucial role in the field of plant genetics and breeding, as sequence-based genotyping has become available for a wide range of genomic studies (J. He et al., 2014).

This methodology necessitates no antecedent sequencing of plant genomes, facilitates direct genotyping of plants with intricate genomes devoid of SNP identification, and yields high-density, cost-effective genotype data. GBS serves as an exceptionally effective method for molecular characterization due to its advantageous attributes (Fu et al., 2014). Mapping of quantitative trait loci (QTL) and conducting association analyses necessitate the availability of genotypic information (Beissinger et al., 2013).

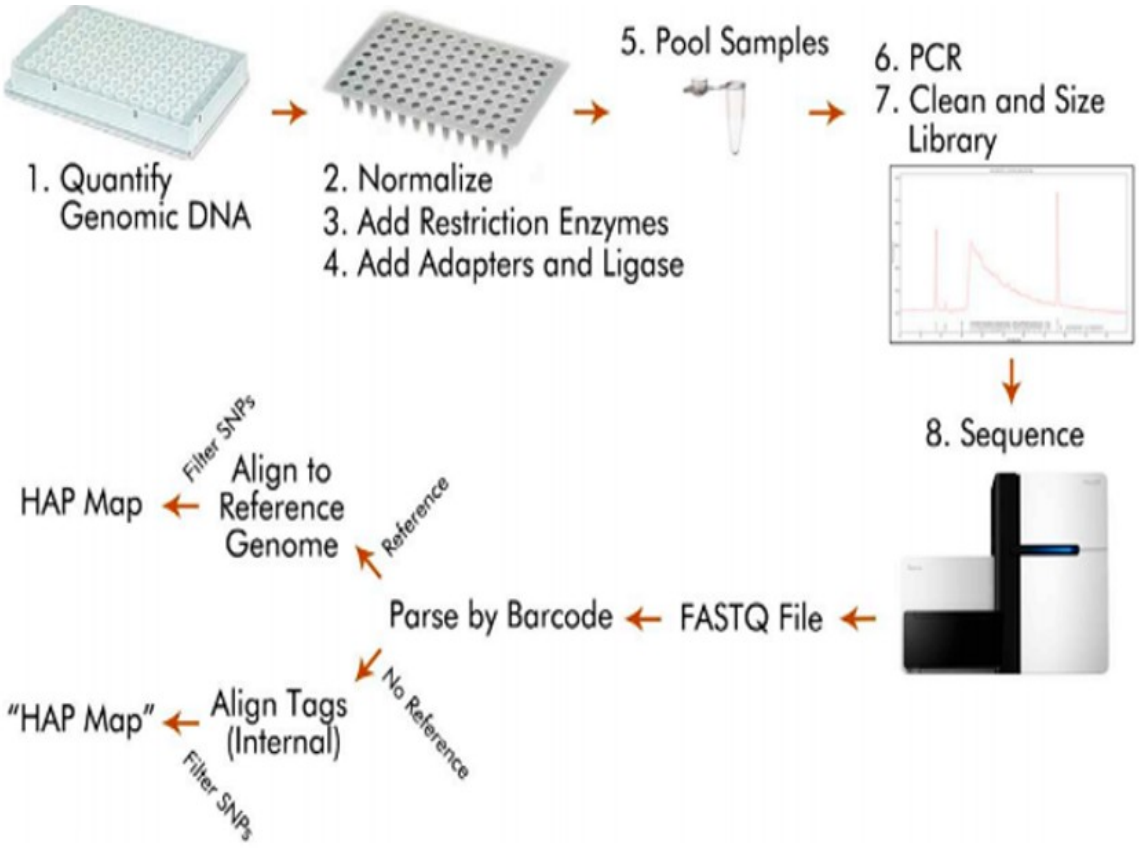


Figure 2. 2: Schematic overview of steps in genotyping by sequencing (GBS) library construction, sequencing, and analysis

(J. A. Poland & Rife, 2012)

2.7 The Utilization of GBS in Plant Breeding

The application of GBS for the identification of high-density SNP markers and the generation of genetic linkage maps presents significant potential and numerous applications within the realm of plant breeding (J. He et al., 2014). The spatial relationships among markers situated along the chromosomes, as evidenced by their recombination frequency, are illustrated on genetic linkage maps. In breeding programs, such maps prove advantageous as they facilitate the simplification of QTL and association analyses (Scheben et al., 2017). The domain of plant genomics and the methodology of applied plant breeding

stand to undergo significant transformation through the advent of next-generation sequencing applications. The integration of phenotype and genotype, coupled with the application of this understanding to formulate phenotypic predictions and select superior plant varieties, constitutes a primary objective of functional genomics within agricultural species (Glaubitz et al., 2014). Innovative methodologies for sequence-based genotyping have been developed to harness the capabilities of NGS for the advancement of plant breeding and genomics. GBS represents a viable technological approach that employs enzyme-mediated complexity reduction, focusing specifically on a limited segment of the genome through the use of restriction endonucleases. This method incorporates DNA barcoded adapters to generate multiplex libraries of samples suitable for next-generation sequencing. This methodology is referred to as the reduced restriction sequence (RRS) (Wang et al., 2020). This method has demonstrated efficacy across numerous species, capable of producing tens of thousands to hundreds of thousands of molecular markers (Elshire et al., 2011). GBS has been employed to augment breeding initiatives for a range of significant crops, including chickpea (Verma et al., 2015), maize (Su et al., 2017; Wang et al., 2020; Elshire et al., 2011), rice (Furuta et al., 2017; Spindel et al., 2013), Napier grass (Muktar et al., 2019), wheat (J. Poland et al., 2012), soybean (Sonah et al., 2013), and barley (Abed et al., 2019). The integration of GBS techniques with phenotypic data establishes a robust framework for the efficient mapping and identification of genes associated with agronomic traits, facilitating their subsequent introgression into crop germplasm (J. He et al., 2014). The versatility of GBS in accommodating various species, populations, and research objectives renders it an exemplary instrument for investigations in plant genetics.

Identifying DNA sequence variations and evaluating markers across a broad spectrum of materials are the two paramount elements of genotyping germplasm. This has traditionally involved a two-phase methodology, consisting of the identification of markers followed by the development of assays and subsequent genotyping. The simultaneous execution of marker discovery and genotyping represents a notable benefit of sequence-based genotyping methodologies (J. A. Poland & Rife, 2012).

This facilitates the identification of innovative germplasm collections or potentially new species without the prerequisite of locating and characterizing polymorphisms in advance. An additional significant characteristic of GBS datasets is the continual evolution of the raw data. With the progression of bioinformatics tools, the advent of reference genomes, and the expansion of sequence data, the raw sequences derived from GBS can undergo reanalysis, uncovering novel insights such as new polymorphisms and annotated genes. Each of these components contributes to enhancing the intrinsic worth of the unrefined dataset. The detection of single nucleotide polymorphism (SNP) and presence–absence variation (PAV) across diverse populations, both with and without reference genomes, represents one of the pioneering and extensively utilized applications of next-generation sequencing (C. Kim et al., 2016; K. D. Kim et al., 2020; De Donato et al., 2013; Deschamps et al., 2012).

The primary objective of the GBS approach is to concurrently identify polymorphisms and collect genotypic data across the entire population of interest, rather than merely detecting polymorphisms and subsequently applying them to a predetermined assay. GBS serves as an

exceptionally rapid and adaptable platform suitable for a diverse array of species and germplasm collections. Its efficacy in genomic selection (GS) within plant breeding programs is underscored by this integrated one-step methodology (J. A. Poland & Rife, 2012). GBS proves to be particularly advantageous for novel germplasm and uncharacterized species, as it facilitates straightforward de novo discovery and the application of new molecular polymorphisms. The mitigation of ascertainment bias linked to marker discovery in panels that diverge from the target population represents, in numerous ways, the paramount advantage of sequence-based genotyping methodologies. This presents a distinct advantage for association studies, given that varying allele frequencies considerably influence the power and precision of the research (Larsen et al., 2018). With the introduction of novel germplasm into the breeding pool, one can uncover informative polymorphisms that serve valuable functions in breeding endeavors. GBS has demonstrated its utility as a method for examining genetic diversity (Fu et al., 2014); (Spindel et al., 2013); (Peterson et al., 2014). GS represents a distinctive methodology that amalgamates molecular markers with phenotypic and pedigree information to enhance the precision of breeding concerning genotypic parameters, in contrast to traditional MAS (Heffner et al., 2009). Both theoretical and practical inquiries into GS indicate its capacity to accelerate the advancement of novel crop varieties. GS, within the framework of the GBS strategy, possesses the potential to substantially augment conventional crop enhancement. It serves as an essential element for advancing genomics-assisted breeding into commercially viable crops characterized by large, intricate genomes (J. A. Poland & Rife, 2012). It offers comprehensive genome-wide scanning and multiplexed sequencing systems at an economical price point. In order to simplify the process, the GBS methodology utilizes

restriction enzymes alongside DNA barcoded adapters (Wickland et al., 2017). Regardless of the presence of reference genome sequences, GBS is capable of conducting SNP discovery and genotyping concurrently (Spindel et al., 2013). Consequently, GBS can be employed to investigate plant breeding and genetics through various methodologies, including linkage maps, genome-wide association studies, genomic selection, and studies of genomic diversity (Muktar et al., 2019). GBS serves as an exemplary method for research that spans from single-gene markers to comprehensive whole-genome analysis, owing to its distinctive attributes (Chung et al., 2017).

2.8 *In silico* Technologies

Bioinformatics research is perpetually advancing through the creation of both online and stand-alone tools, innovative software, and sophisticated algorithms designed for the analysis of SNPs (Morgil et al., 2020). The newly developed open-source bioinformatics software has significantly accelerated SNP detection while simultaneously lowering costs. The crucial aspect lies in the selection of software, sequencing platform, file requirements, algorithmic foundations, operating systems, and the organism of interest, all of which significantly influence the decision regarding the appropriate bioinformatics platform or pipeline to employ. A plethora of databases and resources currently exists to elucidate SNPs across various plant species. The approaches employed for SNP mining will exhibit considerable similarity for both databases obtained and those derived from high-throughput sequencing. Next-generation sequencing technologies, such as Illumina GA/Solexa, SOLiDTM, and Oxford Nanopore high-throughput sequencing, produce substantial volumes of sequence data, consequently leading to the discovery of numerous novel SNPs. The

preferred method may differ depending on the source data and the approach taken. There exist distinct analytical procedures pertinent to the two categories of sequence data: reference sequence data, which is derived from species possessing an accessible reference sequence, and de novo sequence data. In both scenarios, one can identify three principal steps:

i. Group the sequence reads according to their sequence similarities, and confirm identity of reads covering the same part of the genome or having the same transcript origin
ii. Align the reads
iii. Scan for sequence variants
iv. Should a reference sequence be accessible, the initial task involves identifying an appropriate homology search tool to align the new sequence reads with the reference. A variety of tools exist for both global and local mapping, including BLAST and SSAHA, specifically designed for whole-genome data analysis. Should the short reads originate from Illumina, there exist specialized tools like SOAP and MAQ that have been meticulously developed for such purposes. The UniGene set is meticulously crafted for the purpose of mapping transcript data. The subsequent phase involves the selection of a multiple or pairwise alignment tool for the purpose of aligning the mapped reads to the reference sequence. Software tools such as CAP3 and Phrap have been widely employed for this objective. In the context of de novo sequence data, an extra procedure known as clustering is required to organize the short sequence reads into coherent groups. TeraClu, TGICL, and d2cluster are the

predominant tools employed to fragment the input data and subsequently assemble them into distinct contigs. Subsequent to the clustering phase, all nucleotides from distinct reads at the same position on the gene are aligned in a comparable manner utilizing CAP3 and Phrap.

v. The final step is the SNP calling or validation.

In the context of analyzing fragments derived from a trace file or a base quality score, it is noteworthy that tools such as PolyBayes, PolyPhred, novoSNP, and SNPdetector are widely recognized in the field. In the context of de novo SNP mining, one may utilize tools such as AutoSNP, QualitySNP, and MAVIANT. A variety of SNP mining tools and databases tailored for plant research exist, including dbSNP, ESTree DB, POLYMORPH, SNiPlay, AutoSNPdb, and IRIS, among others.

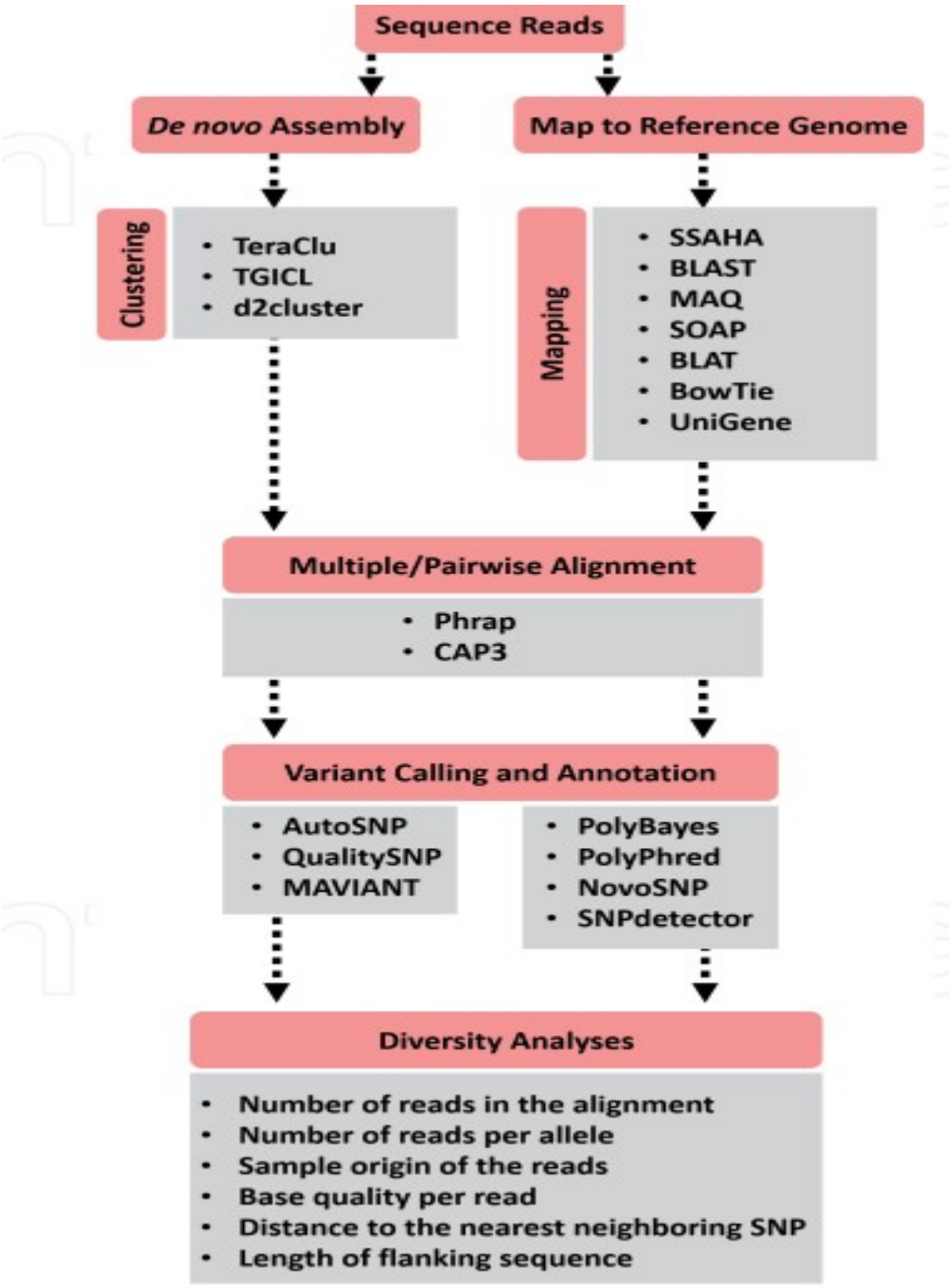


Figure 2. 3: In silico analysis of SNPs

(Morgil et al., 2020).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Study Area

The *M. oleifera* provenance trial was carried out in South Eastern Kenya University (SEKU) located 15 km off Kwa Vonza Market, along the Kitui-Machakos main road, Yatta/Kwa Vonza location, Lower Yatta Sub-County, Kitui County. The site lies at latitudes and longitudes of 037.755460 E 01.313580S respectively. The site elevation is 1173m a.s.l. as evident in figure 4 (Festus et al., 2022).

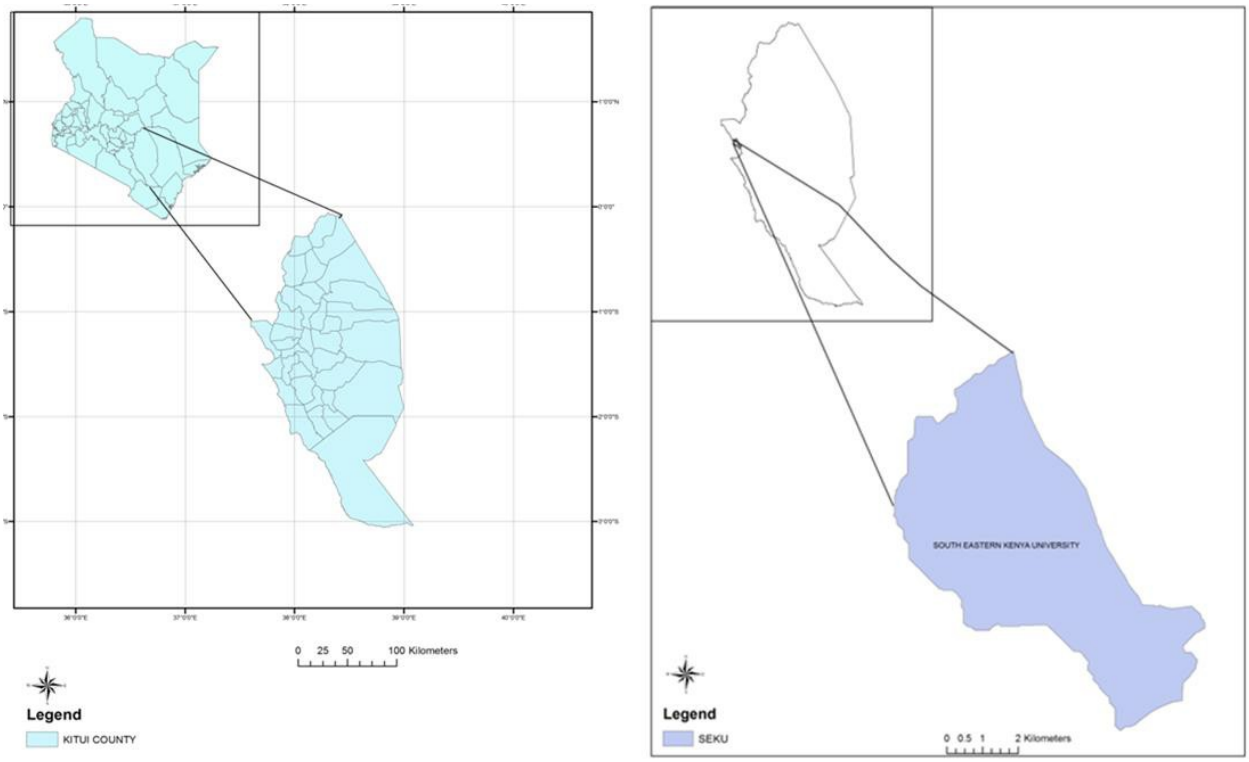


Figure 3. 1: Location of research site in SEKU, Kitui County, Kenya

(Festus et al., 2022)

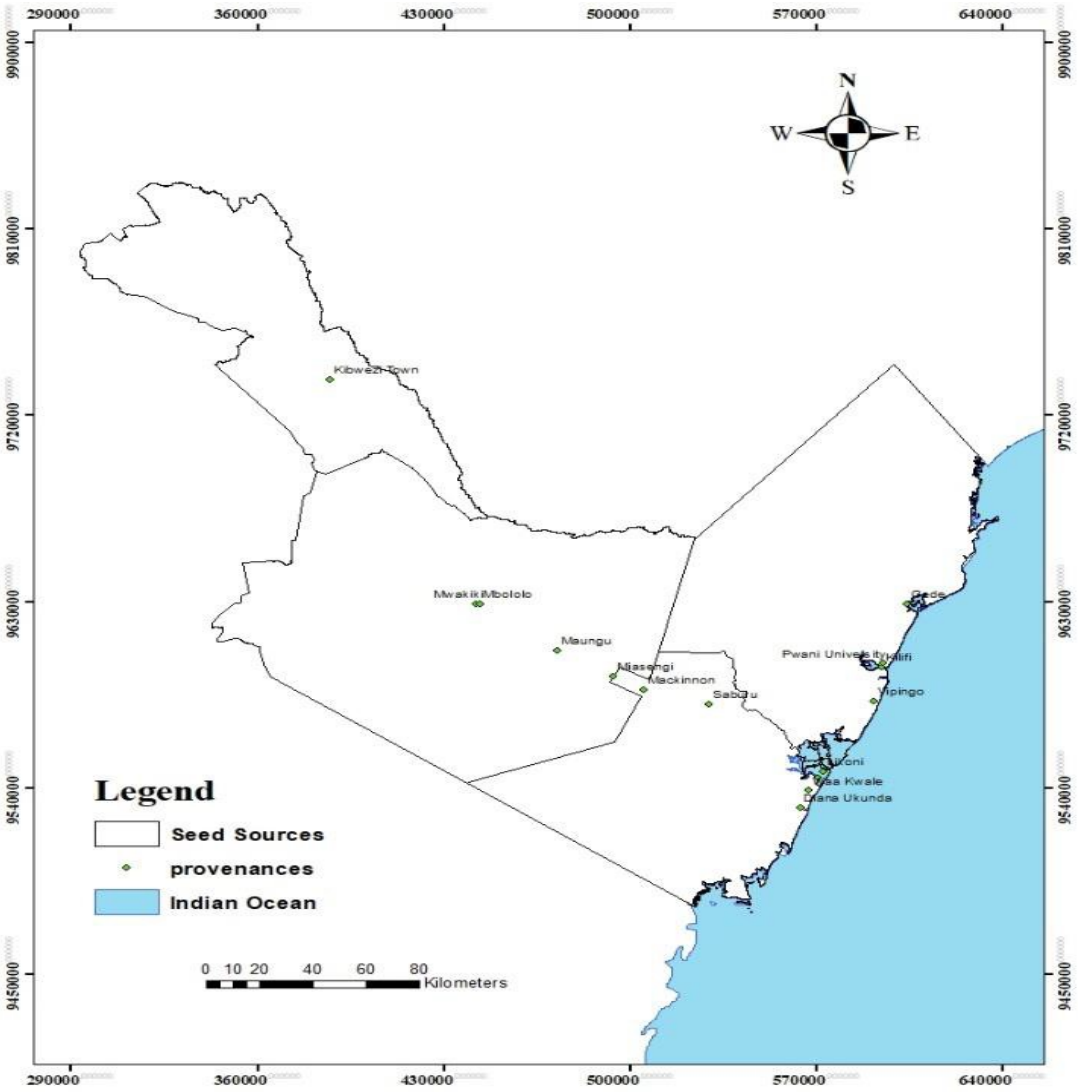


Figure 3. 2: : Location of Seed collection areas (provenances)

(Festus et al., 2022)

3.2 Experimental Design

The original introduction sites in Kenya's Coast and some areas of its Eastern region were used to select provenances (fig. 5). 164 most productive plants were selected (9 to 10 plants per provenance as seen in table 1) using a randomized complete block design (RCBD) (Festus et al., 2022).

Table 3. 1: No of Trees sampled per provenance

	Provenances	No. of trees sampled
1	Mwakiki	10
2	Pwani University	10
3	Shika Adabu	10
4	Miasenyi	10
5	Diani-Ukunda	10
6	Maungu	10
7	India	10
8	Kilifi	10
9	Likoni	10
10	Kibwezi Town	10
11	Mbololo	10
12	Gede	9
13	Arabuko Sokoke	9
14	Mackinon	9
15	Waa Kwale	9
16	Vipingo	9
17	Samburu	9
	Total	**164**

3.3 Sampling Method

Approximately 10 *M. oleifera* leaves per sampled tree were selected and placed in a microtube. The samples were freeze dried (lypholization) overnight using a lypholizer. The dried *M. oleifera* samples were then grinded using a Geno Grinder at 1300 rpm for 3 minutes. The resulting ground *M. oleifera* samples were utilized in isolation of DNA.

3.4 Genotyping by Sequencing (GBS)

3.4.1 DNA Isolation

Genomic DNA isolation was performed using NucleoMag® 96 Plant according to manufacturer's protocol. The reversible adsorption of nucleic acids to paramagnetic beads under suitable buffer conditions is the basis of the kit.

A plant DNA isolation summary by DArT was utilized. 175µl MC1 was added to each *M. oleifera* tissue sample. It was spinned down to collect liquid, incubated for 1 hr at 62˚c then spinned down for 10 minutes at 3000 RPM. 100µl lysate was taken into a deep well plate and 7.5µl beads and 100µl MC2 added to each sample. The samples were mixed on the shaker for 5 minutes. The plate was put on the magnetic separator and the supernatant discarded. The plate was then removed from the magnetic separator and 150µl of MC3 added. Mixing was done on the shaker for 5 minutes, then the plate was put on the magnetic separator and the supernatant discarded. The plate was removed from the separator and 150 µl of MC4 was added. The samples were then mixed on the shaker for 5 minutes, the plate was put on the magnetic separator and the supernatant discarded. The plate was removed from the separator and 150 µ l of 80% Ethanol added. After mixing the samples on the shaker for 5 minutes, the plate was put on the magnetic separator and the supernatant discarded. The plate was left on the magnetic separator and 150 µl MC5 was added. Incubation was done on separator for 60 seconds and the supernatant was discarded ensuring there was no carry over. The samples were removed from the magnetic separator and 30 µl of MC6 was added. Mixing on the shaker was done for 5 minutes, then the plate was put on the magnetic separator and lastly the eluted DNA was removed for PCR.

3.4.2 Agarose Gel Electrophoresis

Agarose gel electrophoresis of the DNA is utilized to show the integrity of the DNA. Quantification of the isolated DNA was done using 0.8% Agarose gel to visualize bands.

3.4.3 Library Preparation

By integrating two restriction enzymes (Pst1-CTGCAG-, Mse I) along with adapters marked by barcodes for sample identification, a genomic representation of the samples was generated for operation within a single lane on the Illumina HiSeq2500 instrument. The DNA barcoded adapters were incorporated at the termini of the DNA fragments of interest, accompanied by ligase to facilitate ligation with primers for subsequent amplification. The PST 1 enzyme served as a barcode adapter, whereas MSE 1 functioned as a common adapter. The site for MSE was successfully amplified through PCR techniques. The assessment of quality control for the GBS library was conducted through the application of 1.2% Agarose gel electrophoresis. Equimolar quantities of the amplified fragments were combined utilizing Tecan, a device employed for the pooling of substantial samples. The pooled samples were subsequently utilized in the c-Bot bridging PCR (Illumina), a fully automated process for clonal cluster generation, followed by sequencing with the Illumina HiSeq2500 system.

The sequencing libraries were prepared in accordance with the optimized protocols established by SEQART Africa. The Illumina Hiseq 2500 was employed for the sequencing of DNA libraries. The analytical program, which has been developed and patented by DArT Pvt. Ltd. in Australia, was employed to produce two distinct categories of data: scores for "presence/absence" markers, referred to as SilicoDArTs, and SNP markers, as noted by

Spinoso-Castillo et al. in 2020. Markers of superior quality were discerned for this specific inquiry through the filtration of FASTQ data according to established quality parameters.

3.5 Data Analysis

The genotypic characterisation was conducted with the NGS method DArTseq™. DArTseq integrates next-generation sequencing platforms with methyl filtration-based DArT complexity reduction methodologies (Sohail et al., 2012; Kilian et al., 2012). The data from 164 genotypes was analyzed using KDCompute, a platform offered by Seqart Africa for diverse studies, and the Diversity Arrays Technology (DArT) R, a R statistics package created by DArT. Distinct and clearly defined bands were evaluated for the presence (1) and absence (0) of single nucleotide polymorphisms (SNPs) utilizing Gel Analyzer 2010a to create a binary data matrix (Spinoso-Castillo et al., 2020) A genomic representation of the samples was produced through the digestion and ligation of genomic DNA using a combination of two restriction enzymes (Pst1-CTGCAG- and Mse I) and adapters tagged with barcodes to distinguish each sample for processing in a single lane on the Illumina HiSeq2500 instrument. The data were subsequently examined based on the matrix. The polymorphism information content (PIC) was computed using silico DArT. Analysis of molecular variance (AMOVA) and principal component analysis (PCA) were conducted using the genetic distance matrix. AMOVA was employed to assess the degree of genetic variation both among and within the population, as well as the level of population genetic differentiation.

Package DArT R devoloped by (Gruber et al., 2018)**,** which is an R stats package was utilised in this analysis for:

- loading DArT TM SNP and SilicoDArT data and Intelligent interpretation and input of DArT comma-delimited files to a compact genlight form of the R {adegenet} package.
- Filtering loci and individuals on criteria drawn from the DArT locus metadata (such as repAvg, Avg-PIC) or on computed statistics such as call rate.
- Relabelling individuals and recoding populations into new aggregations, and deleting selected individuals or populations.
- Visualization using Principal Coordinates Analysis (PCoA) and Neighbour-joining trees (Gruber et al., 2019)**.**

Diversity studies were conducted employing both simple and hierarchical clustering methodologies. The premise underlying cluster analysis posits that the data exhibits discontinuity. It illustrates the correlation of genetic resemblance among genotypes, grounded in phenotypic traits or evolutionary connections. It functions to differentiate between various groups and categorize related lines or genomes into a singular classification. The methodology employed is the Unweighted Paired Group Method with Arithmetic Mean (UPGMA). Alongside basic clustering, hierarchical clustering was conducted to derive the similarity coefficient, which proved crucial in assessing the degree of variability among the genotypes. A hierarchical cluster dendrogram was developed, plotting height against distance with two p values. The figure on the left (red) represents an approximately unbiased p-value (Au), calculated through multiscale bootstrap resampling, while the figure on the right (green) denotes a bootstrap probability p-value (Bp), derived from standard bootstrap

resampling.

Furthermore, principal coordinate analysis (PCoA) serves as a method for reducing the dimensionality of extensive datasets, thereby enhancing interpretability while minimizing the loss of information. This is accomplished through the generation of new uncorrelated variables that progressively maximize variance. Principal coordinate analysis (PCoA) serves as a method for data reduction, specifically designed to handle quantitative data types. Multi-correlated data are transformed through PCoA into a new set of uncorrelated variables for subsequent analysis. This new collection of variables is constituted by linear combinations of the original variables. The basis lies in the progression of eigenvalues and the orthogonally independent eigenvectors (principal components) organized in a descending sequence according to the magnitude of variance. The components generate scatter plots of data that exhibit the ideal attributes for examining inherent variability and correlation (Bhandari et al., 2017).

Ultimately, a distance matrix-based approach was employed to construct a phylogenetic tree.

Cluster analysis was conducted employing KDCompute, utilizing the unweighted pair-group method with arithmetic average (UPGMA) grounded in Nei's genetic distance matrix. Clustering analysis was conducted using the KDCompute software, which employs genotypic marker data to perform both simple and hierarchical clustering based on the Euclidean distance between samples. The analysis of hierarchical clustering was conducted employing the pvclust function in R (Suzuki & Shimodaira, 2015), which leverages multiscale bootstrap resampling techniques for each hierarchical cluster. The KDCompute software employs the R package 'pvclust' to generate a clustering dendrogram accompanied

by P-Values through bootstrap methods, as well as a compilation of clusters derived from the genotypic marker data. A dendrogram was created to illustrate the similarity index among the 164 genotypes and their corresponding clustering pattern. The DArT R methodology, employing the prcomp function from the R stats package, was applied to execute the PCoA of the genotypic data derived from DArT mapping. A graphical representation of PC1 against PC2, highlighting the two most significant principal components, was created alongside a pair-wise grid layout of the most relevant principal components, featuring a Bi-plot and the raw data from the PCA analysis. The Adegenet package of DArT R was employed with five factors. A three-dimensional representation was also apparent from DArT R.

SNP calling was done by Illumina's SNP caller aligorithm in the CASAVA software. For the population genetics analysis, F statistic (Fst) utilising the functions StAMPP package in DArT R was performed. Euclidean was used as a measure of distance. A phenetic tree was constructed using a neighbor-joining approach using DArT R. Analysis of molecular variance (AMOVA) was utilized to carry out population genetic analysis using KDCompute Software (silicoDArT) and DArT R.

CHAPTER FOUR

RESULTS

4.1 Characterization of *Moringa oleifera* Provenances

Through a combination of cluster analysis, PCoA and 3D Plot, the genotypes were characterized. Clustering generated varied phylogenetic trees arising from both simple and hierarchical clustering. Through PCA a plot was generated with two Principal Coordinates of 71% and 15 %.

4.1.1 Simple Cluster Analysis

The *M. oleifera* SNP profile was clustered into 4 clades as shown in figure 6 and table 2. Clade III had the highest number of genotypes, which amounted to 58. Clade II consisted of 48 genotypes. Clade IV had 38 genotypes whereas clade I had the least genotypes which were 20. In Clade I, all 9 genotypes from Mackinon were clustered here whereas India had 6 genotypes, Samburu and Gede gad 2 genotypes and Arabuko Sokoke had only 1 genotype. The 9 genotypes from Mackinon sub-clustered together. All the genotypes in clade one amounted to 20. Clade II consisted of 48 genotypes with Shika Adabu having all of its genotypes clustered here. Arabuko Sokoke had 7 genotypes, Gede and Kibwezi Town had 6 genotypes, Waa Kwale had 5 genotypes, Samburu had 4 genotypes, India had 3 genotypes, Kilifi, Mbololo and Diani-Ukunda had 2 genotypes with Vipingo having 1 genotype.

The highest number of genotypes were evident in Clade III and they amounted to 58. Mwakiki, Miasenyi and Likoni had their 10 genotypes clustered in this clade. 7 genotypes were from Mbololo, 5 genotypes from Vipingo, 4 genotypes from Waa Kwale, 3 genotypes

from Diani-Ukunda, 2 genotypes from Pwani University, Kilifi and Maungu and 1 genotype from Gede, Samburu and India.

Lastly, Clade IV exhibited 38 genotypes with Pwani University and Maungu having 8 of its genotypes clustered here. Kilifi had 6 genotypes, 5 genotypes were from Diani-Ukunda, 4 genotypes from Kibwezi Town, 3 genotypes from Vipingo, 2 genotypes from Samburu whereas Arabuko Sokoke and Mbololo had only 1 genotype (Fig. 6)

Table 4. 1: A summary of the genotypes obtained in Fig. 1 by simple cluster analysis

Clade	I	II	III	IV
Provenance	Mackinon – 9 India – 6 Samburu – 2 Gede – 1 Arabuko Sokoke -1	Shika Adabu – 10 Arabuko Sokoke – 6 Gede – 6 Kibwezi – 6 Waa Kwale – 5 Samburu – 4 India – 3 Kilifi – 2 Mbololo - 2 Diani-Ukunda – 2 Vipingo - 1	Mwakiki – 10 Miasenyi – 10 Likoni – 10 Mbololo – 7 Vipingo – 5 Waa Kwale – 4 Diani-Ukunda-3 Pwani University – 2 Kilifi – 2 Maungu – 2 Gede – 1 Samburu – 1 India - 1	Pwani University – 8 Kilifi – 6 Diani-Ukunda - 5 Kibwezi Town -4 Vipingo – 3 Samburu – 2 Arabuko Sokoke -1 Mbololo - 1
Total	**20**	**47**	**58**	**38**

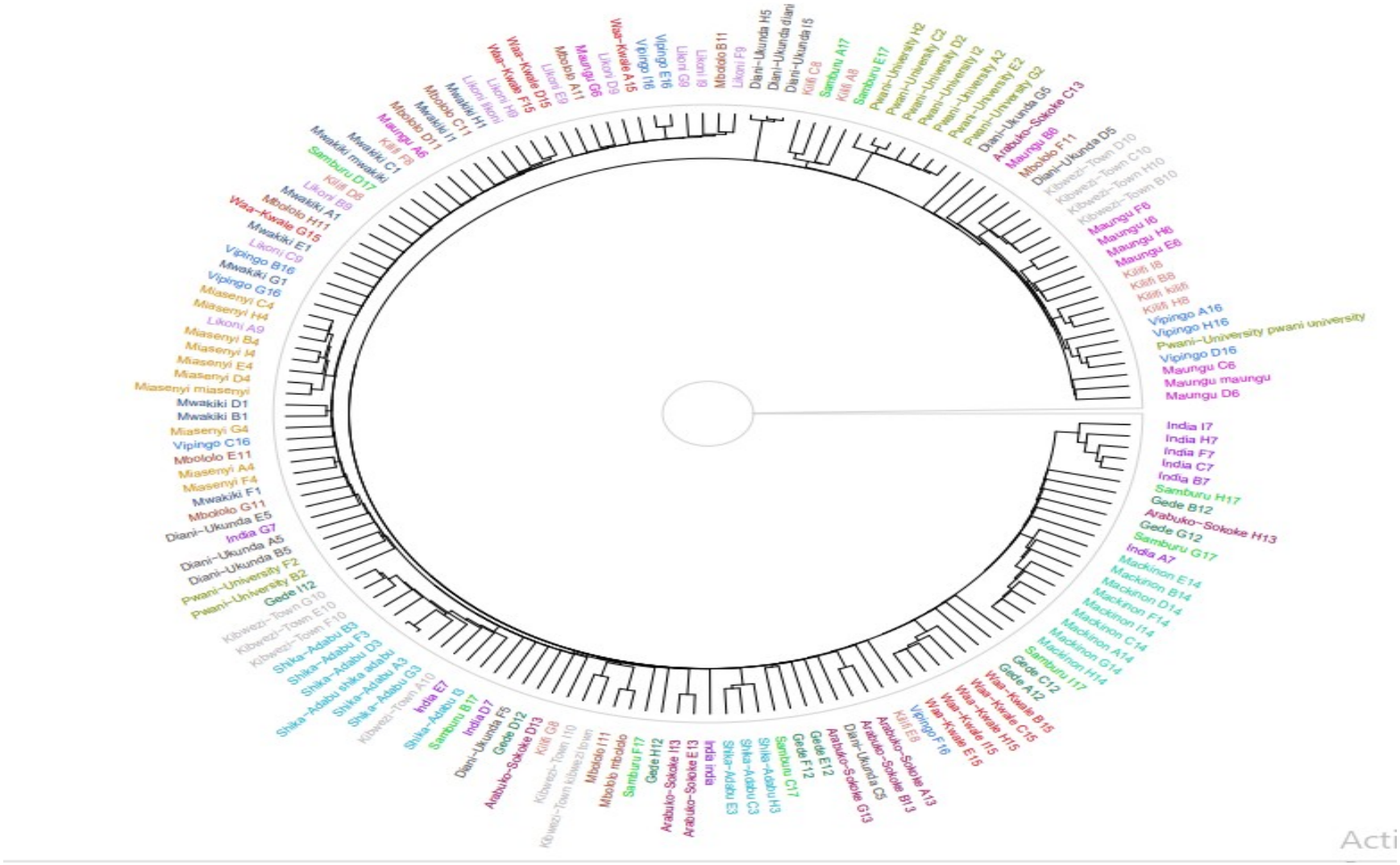


Figure 4. 1: Phylogenetic tree showing diversity and variance between M. oleifera genotype

4.1.2 Hierarchical Cluster Analysis

The hierarchical cluster dendrogram (Fig. 8 and table 3) exhibited four clades based on the SNP data which corresponded with simple clustering analysis. However, similarity coefficients were obtained in hierarchical clustering which are essential in determining the level of diversity in between the genotypes. The confidence level for the Au p-value (similarity coefficient) showed a similarity of 94% for Clade I, 96% for Clade II and 97% for Clade III and IV. Hierarchical cluster analysis showed varied levels of genetic variation among the genotypes from the 17 provenances. The similarity coefficient ranged from 63% to 100%, with most being above 90% which indicates a high level of similarity between the studied genotypes.

Clade I consisted of 11 genotypes with Miasenyi having all of its 10 genotypes clustered here. The other 1 genotype was from Likoni. Clade II had 54 *Moringa* genotypes with Mwakiki having all its 10 genotypes clustered in this clade. This was closely followed by Likoni provenance, which had 9 genotypes. Mbololo exhibited 7 genotypes, Shika Adabu, and Vipingo had 5 genotypes, Kibwezi Town had 4 genotypes, Diani-Ukunda, Maungu and Waa Kwale had 3 genotypes, Kilifi had 2 genotypes while India, Samburu and Arabuko Sokoke exhibited 1 genotype each. Clade II exhibited 7 sub clades. 5 genotypes of Shika Adabu provenance clustered in the 1st sub clade with 4 genotypes of Kibwezi Town and 1 genotype of India provenance indicating that these 12 genotypes had least variability. They exhibited a similarity coefficient of 94%. The 12 genotypes also clustered near Clade I, which indicates their close relationship. The 2nd sub clade had 3 genotypes whereby Diani-Ukunda, Mwakiki and Samburu had 1 genotype each with a similarity coeffient of 79%. 2 genotypes from Likoni, 2 from Mbololo and 1 genotype from Kilifi clustered in the 3rd sub

clade with a similarity coefficient of 79%. The 4th sub clade consisted of 14 genotypes which had a similarity coefficient of 90%. Six genotypes clustered together whereby Mbololo had 2 genotypes whereas Waa Kwale, Likoni, Maungu and Kilifi had 1 genotype each, with a similarity coefficient of 87%. The other 8 genotypes clustered together with Likoni having 4 genotypes, Vipingo 2 genotypes, whereas Waa Kwale and Arabuko Sokoke had 1 genotype each with a similarity coefficient of 90%. Sub clade 5 had 19 genotypes which had a similarity coefficient of 97%. 5 of the genotypes clustered together whereby Mbololo had 3, Samburu and Maungu had 1 genotype each. The 5 genotypes had a similarity coefficient of 87%. The other 14 genotypes also clustered together with Mwakiki having 9 genotypes. Vipingo had 2 genotypes, Maungu, Waa Kwale and Likoni provenances had 1 genotype each. It is thus evident that the 9 Mwakiki genotypes were closely related and had less variability with the other 5 genotypes from various provenances as discussed above because they had a similarity coefficient of 82%. The 4th sub clade had a similarity coefficient of 83%. Clade I and II had a similarity coefficient of 94%.

In Clade III, there were 94 *M. oleifera* genotypes with Mackinon and Gede provenances having all of their 9 genotypes clustered in this clade. India and Pwani University also exhibited 9 of their 10 genotypes in Clade III. Arabuko Sokoke and Samburu had 8 genotypes, Waa Kwale and Diani-Ukunda had 7 of its genotypes, Kilifi and Maungu had 6 genotypes, Shika Adabu had 5 genotypes, Kibwezi Town and Vipingo had 4 genotypes and lastly Mbololo had 2 genotypes. Clade IV had the least genotypes amounting to 5 whereby Kibwezi Town had 2 genotypes, while Pwani University, Mbololo and Shika Adabu had 1 genotype each.

From the hierarchical dendrogram, it is evident that Clade III had five sub clades, which had a similarity coefficient of 91%. In the 1st sub clade, there were only 3 genotypes whereby Samburu provenance had 2 and Kilifi had with a similarity coefficient of 97%. The 2nd sub clade had 12 genotypes with a similarity coefficient of 79%. 6 genotypes clustered together with Maungu having 5 and Diani-Ukunda 1 exhibiting a similarity coefficient of 83%. The other 6 genotypes, which included 4 genotypes from Kibwezi Town and 1 from Maungu and Vipingo, also clustered together with a similarity coefficient of 72%. Sub clade 3 had the highest number of genotypes amounting to 37 with a similarity coefficient of 79%. 15 of the genotypes clustered together whereby Pwani University had 7 genotypes, Maungu and Kilifi had 3 genotypes each while Arabuko Sokoke had 2 genotypes. The similarity coefficient of these genotypes was 63%. The other 22 genotypes clustered together and had a similarity coefficient of 77%. 2 genotypes from Shika Adabu and 1 genotype from Kilifi had a similarity coefficient of 100%. All 9 genotypes of Mackinon clustered together with a 91% similarity coefficient. The remaining 10 genotypes had a similarity coefficient of 75%. 4 genotypes from Waa Kwale and 1 from Vipingo had a similarity coefficient of 90% while 3 genotypes from Vipingo, 1 from Mbololo and Diani-Ukunda clustered together with a similarity coefficient of 93%. The 4th sub clade had 29 genotypes with a similarity coefficient of 92%. 2 genotypes of Pwani University clustered together with a similarity coefficient of 100%. 6 genotypes from Samburu, 7 from Gede, 3 genotypes from Diani-Ukunda and 1 genotype from Shika Adabu clustered together with a similarity coefficient of 91%. 3 genotypes from Arabuko Sokoke, 2 genotypes from Diani-Ukunda, Gede and India and 1 genotype from Kilifi clustered together and had a similarity coefficient of 90%. Lastly, the 5th sub clade had 14 genotypes exhibiting a 96% similarity coefficient. It was evident that

India had 7 genotypes, Arabuko Sokoke had 4, Mbololo had 2 whereas Kilifi had 1 genotype clustered together.

Clade IV had the least number of genotypes amounting to 5. The similarity coefficient was 97% implying less genetic variability among the genotypes. 2 genotypes from Kibwezi Town clustered together and had a similarity coefficient of 100% while 1 genotype from Mbololo, Shika Adabu and Pwani University clustered together with a similarity coefficient of 91%

Table 4. 2: A summary of genotypes obtained by Hierarchical clustering dendrogram (Figure 2) with similarity coefficients indicated.

Clade	Genotypes per Provenance	Sub-clades	Similarity Coefficient (%)	
I	Miasenyi - 10	Miasenyi - 2	100	94
	Likoni - 1	Miasenyi - 8	94	
		Likoni - 1		
II	Mwakiki - 10	Shika Adabu - 5	94	96
	Likoni - 9	Kibwezi Town - 4		
	Mbololo - 7	India - 1		
	Shika Adabu - 5	Diani- Ukunda - 1	79	
	Vipingo - 5	Mwakiki - 1		
	Kibwezi Town - 4	Samburu - 1		
	Diani-Ukunda - 3	Likoni - 2	79	
	Maungu - 3	Mbololo - 2		
	Waa Kwale - 3	Kilifi - 1		
	Kilifi - 2	Mbololo - 2	90	
	India - 1	Waa Kwale - 2		
	Samburu - 1	Likoni - 5		
	Arabuko Sokoke - 1	Maungu - 1		
		Kilifi - 1		
		Arabuko Sokoke - 1		
		Vipingo - 2		
		Mbololo - 3	97	
		Samburu - 1		
		Maungu - 2		
		Mwakiki - 9		
		Vipingo - 2		
		Waa Kwale - 1		
		Likoni - 1		
IV	Kibwezi Town - 2	Kibwezi Town - 2	100	97
	Mbololo - 1	Mbololo - 1	91	
	Shika Adabu - 1	Shika Adabu - 1		
	Pwani University - 1	Pwani University - 1		

Clade	Genotypes	Sub-clades	Similarity Coefficient (%)	
III	Mackinon - 9	Samburu - 2	97	97
	Gede - 9	Kilifi - 1		
	India - 10	Maungu - 6	79	
	Pwani University - 10	Diani-Ukunda - 1		
	Samburu - 8	Kibwezi Town - 4		
	Arabuko Sokoke - 7	Vipingo - 1		
	Waa Kwale - 7	Pwani University - 7	79	
	Diani-Ukunda - 7	Maungu - 3		
	Kilifi - 6	Kilifi - 4		
	Maungu - 6	Arabuko Sokoke - 2		
	Shika Adabu - 5	Shika Adabu - 2		
	Kibwezi Town - 4	Diani-Ukunda - 1		
	Vipingo - 4	Mackinon - 9		
	Mbololo - 2	Waa Kwale - 4		
		Vipingo - 4		
		Mbololo - 1		
		Pwani University - 2	92	
		Samburu - 6		
		Gede - 7		
		Diani-Ukunda - 5		
		Shika Adabu - 1		
		Arabuko Sokoke - 3		
		Kilifi - 1		
		Gede - 2		
		India - 2		
		India - 7	96	
		Arabuko Sokoke - 4		
		Mbololo - 2		
		Kilifi - 1		

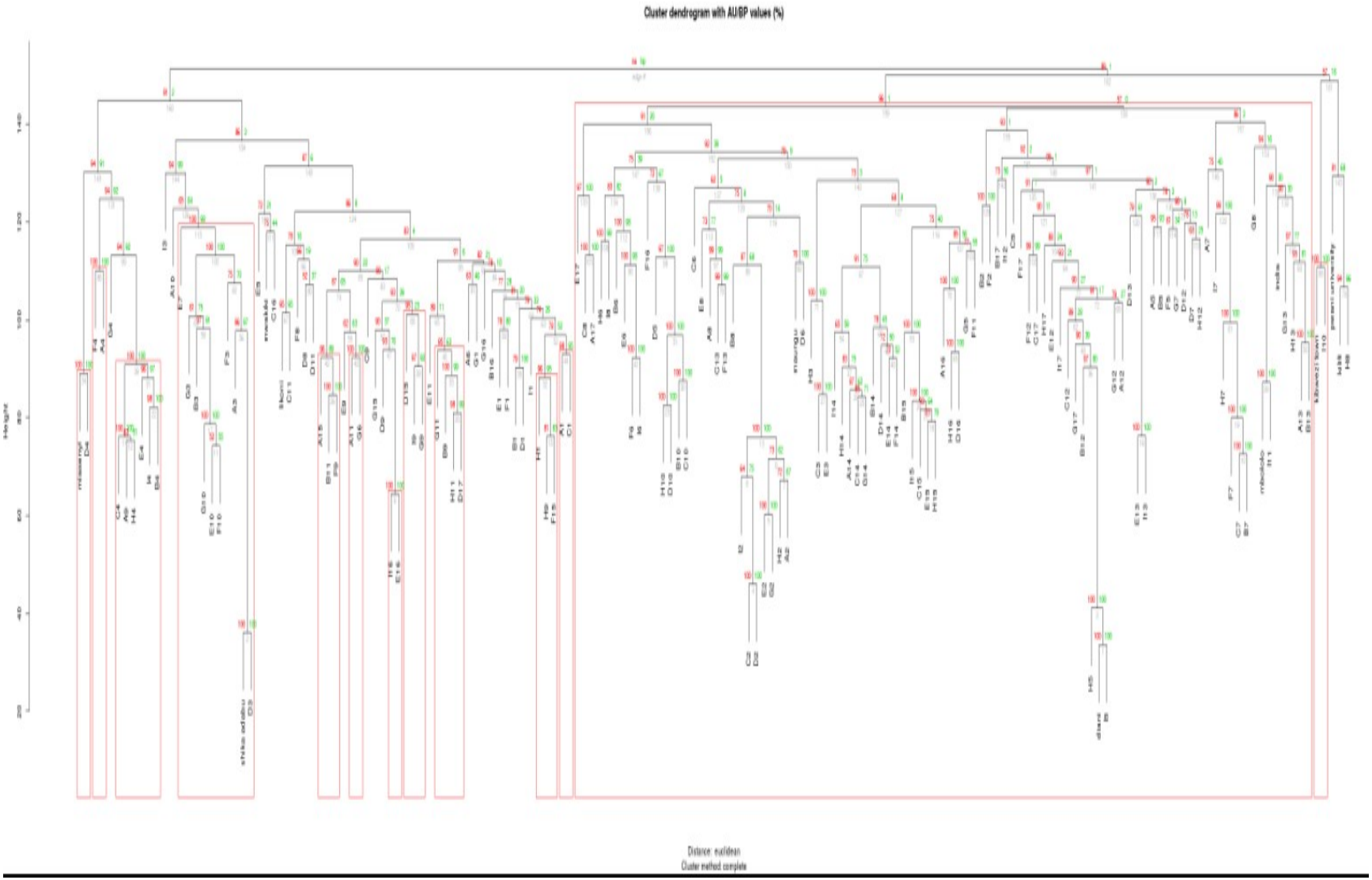


Figure 4. 2: Hierarchical clustering dendrogram with au and bp value

4.1.3 Principal Coordinate Analysis Results

The *M. oleifera* SNP profile classified the 164 genotypes into four major groups. This grouping pattern corresponds with that of clustering analysis as shown by the hierarchical dendrogram. The genetic similarity of individuals and populations were visualized by a principal component analysis (PCoA) plot, which represented the samples based on the eigenvalues in the two-dimensional graph with respect to the first two principal coordinates as shown in figure 8.

Principle component (PC) 1 amounted to 71% variance and the 2nd Principle component had 15% variance as evident in the PCoA plot. According to the PCoA score plot, the 1st coordinate which lies in the positive of PC2 and negative of PC1 had 31 genotypes. 7 genotypes were from Mackinon, Pwani University had 6 genotypes, India and Waa Kwale had 4 genotypes each, Diani-Ukunda had 3, Gede had 2 genotypes whereas Kilifi, Mbololo, Arabuko Sokoke, Vipingo and Samburu had 1 genotype each.

50 genotypes lie in the negative of PC1 and negative of PC2. These 8 genotypes from Maungu, 6 genotypes from Arabuko Sokoke and Kilifi, 5 genotypes from Kibwezi Town and Samburu, 3 genotypes from Shika Adabu, India, Gede and Vipingo, 2 genotypes from Pwani University, Diani-Ukunda and Mackinon and 1 genotype from Mbololo and Waa Kwale. These genotypes had a high genetic similarity. From the PCA plot, there were 37 genotypes in the positive of PC1 and negative of PC2 whereby 7 genotypes were from Shika Adabu, Diani-Ukunda and Kibwezi Town had 5 genotypes, Miasenyi and Gede had 4 genotypes, India had 3 genotypes, Arabuko Sokoke, Mwakiki and Samburu had 2 genotypes each whereas Pwani University, Kilifi and Mbololo had 1 genotype each.

Lastly, 46 genotypes were in the positive of PC1 and positive of PC2 with Likoni having all its 10 genotypes here. Mwakiki had 8 genotypes, Mbololo had 7 genotypes, Miasenyi had 6 genotypes, 5 genotypes belonged to Vipingo, 4 genotypes were from Waa Kwale, Kilifi and Maungu had 2 genotypes each whereas Pwani University and Samburu had 1 genotype each.

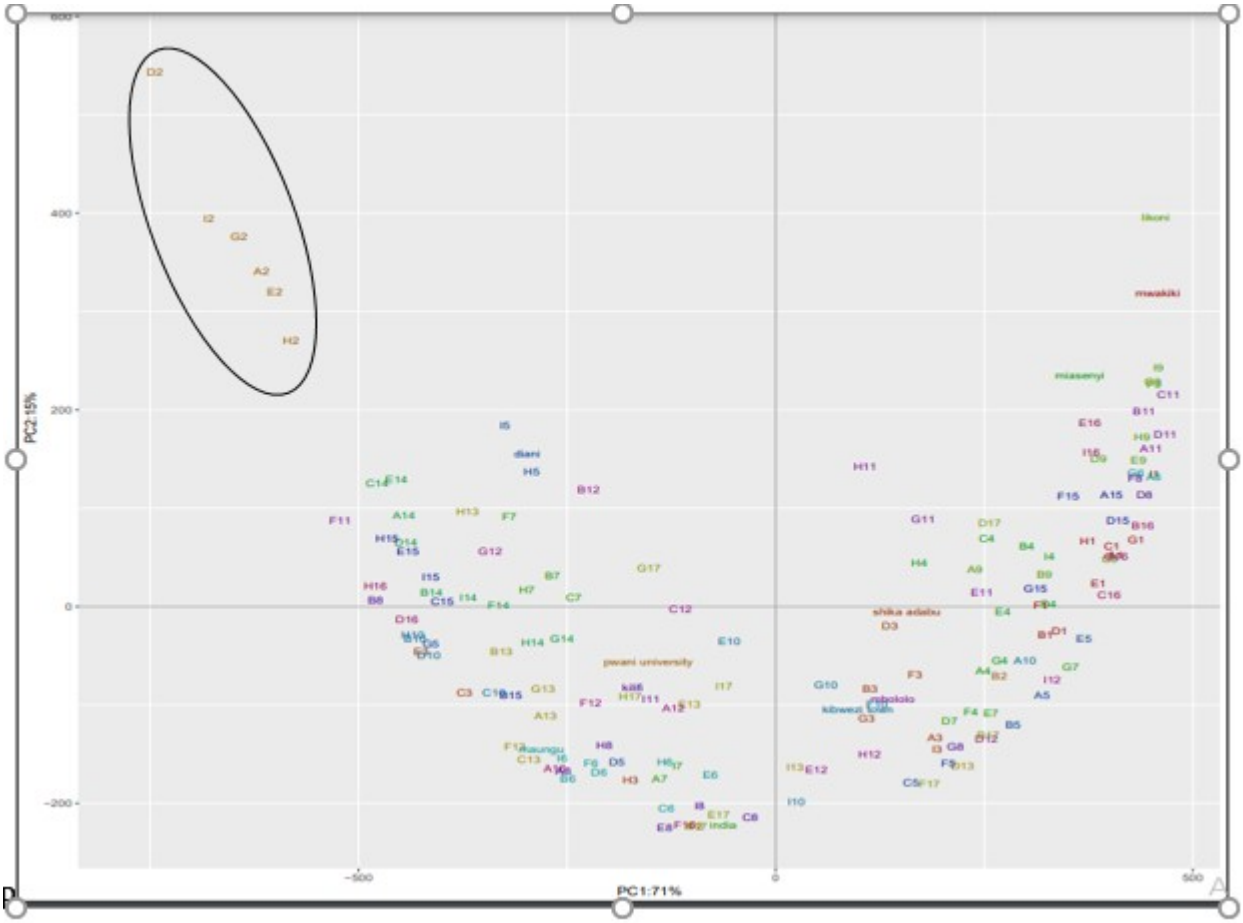


Figure 4. 3: Principal Coordinate Analysis plot

Where 1 – Mwakiki, 2 – Pwani University, 3 – Shika Adabu, 4 – Miasenyi, 5 – Diani-Ukunda, 6 – Maungu, 7 – India, 8 – Kilifi, 9 – Likoni, 10 – Kibwezi Town, 11 – Mbololo, 12 – Gede, 13 – Arabuko Sokoke, 14 – Mackinon, 15 – Waa Kwale, 16 – Vipingo and 17 – Samburu.

Table 4. 3: Principal Coordinate Analysis (PCoA) showing the number of genotypes per coordinates summarized from Figure 9.

Coordinates	I:PC1-ve,PC2+ve	II: PC-ve,PC2-ve	III: PC +ve,PC2-ve	IV:PC1+ve,PC2+ve
Genotypes	Mackinon - 7	Maungu - 8	Shika Adabu - 7	Likoni - 10
	Pwani University -6	Arabuko Sokoke - 6	Diani-Ukunda - 5	Mwakiki - 8
	India - 4	Kilifin- 6	Kibwezi Town - 5	Mbololo - 7
	Waa Kwale - 4	Kibwezi Town-5	Miasenyi - 4	Miasenyi - 6
	Diani-Ukunda - 3	Samburu - 5	Gede - 4	Vipingo - 5
	Gede - 2	Shika Adabu - 3	India - 3	Waa Kwale - 4
	Kilifi - 1	India - 3	Arabuko Sokoke - 2	Kilifi - 2
	Mbololo - 1	Gede - 3	Mwakiki - 2	Maungu - 2
	Arabuko Sokoke - 1	Vipingo - 3	Samburu - 2	Pwani University - 1
	Vipingo - 1	Pwani University - 2	Pwani University - 1	Samburu -1
	Samburu - 1	Diani-Ukunda- 2	Kilifi - 1	
		Mackinon - 2	Mbololo - 1	
		Mbololo - 1		
		Waa Kwale - 1		
Total	**31**	**50**	**37**	**46**

4.1.4 3D Plot Result

An interactive 3D plot was produced using DArT R (Adegenet package) due to the fact that 3 dimensions were indicated by the scree plot (Gruber et al., 2019)

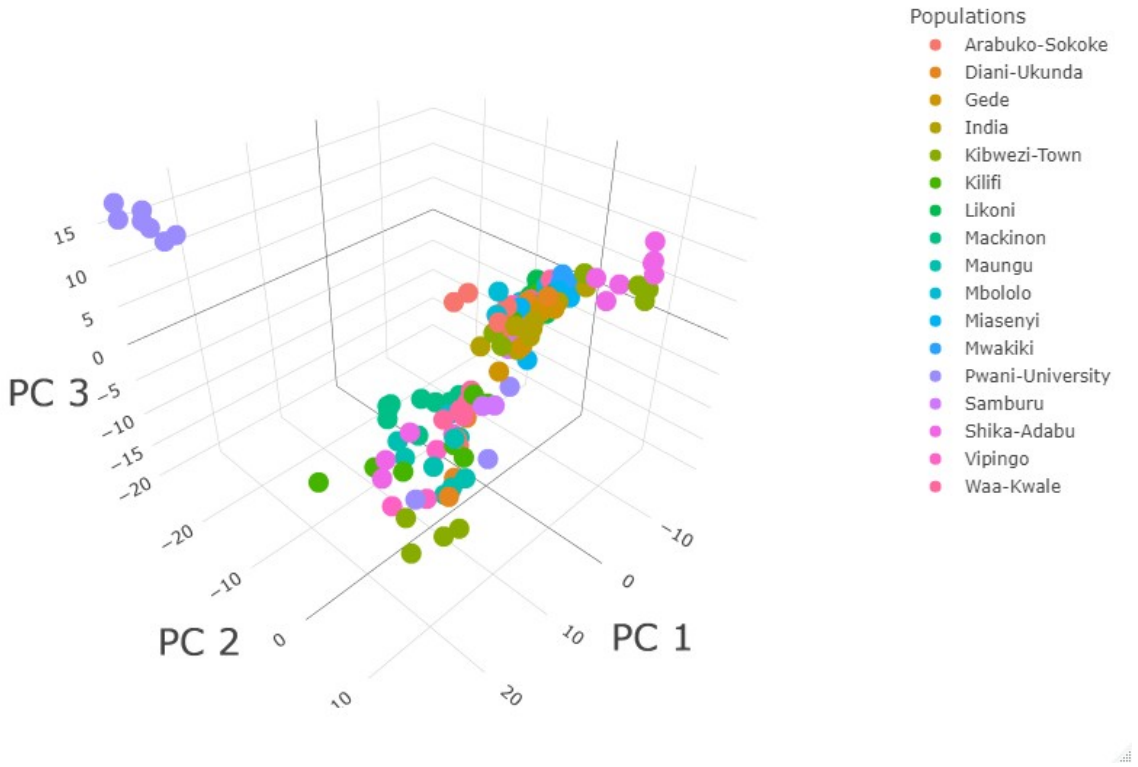


Figure 4. 4: Three dimensional plot

4.2 To Identify Polymophisms in the Selected Genotypes

A comprehensive analysis yielded a total of 37,853 markers (SilicoDArTs; table 8 and DArTSNPs; table 7), which were meticulously scored utilizing DArTsoft v.7.4.7 as binary data (1 indicating presence, 0 indicating absence, for a marker within the genomic DNA of each accession). The DArTsoft v.7.4.7 software was utilized to compute various quality parameters for DArTseq and SNP markers, including call rate, polymorphic information content (PIC), and reproducibility. There existed 16,932 SilicoDArTs and 20,921 DArTSNPs markers.

The heterozygosity of the samples (FreqHet) varied from 0 to 0.698, with an average heterozygosity calculated at 0.11. The range of samples exhibiting homozygosity for the SNP allele (FreqHomSnp) varied from 0 to 0.99, with a mean value of 0.2, as presented in table 7. The proportion of samples exhibiting a genotype score of “1” in the SNP allele row (OneRatioSnp) varied from 0.006 to 1, with a mean value of 0.3. The average count of the reference, defined as the total of the tag read counts across all samples divided by the number of samples exhibiting non-zero tag read counts for the reference allele (AvgCountRef), varied from 2.57 to 261.87, with a mean of 24.96.

The mean count of the SNPs, calculated as the total tag read count across all samples divided by the number of samples exhibiting non-zero tag read counts for the SNP allele (AvgCountSnp), varied from 2.5 to 181.97, with an overall average of 16.77. The ratio of technical replicate assay pairs exhibiting consistency in marker scores (RepAvg) was found to be 1. The fraction of samples exhibiting a genotype score of "1" (One Ratio) was 1. The

polymorphism information content (PIC) for the SNP allele row (PICSnp) varied from 0 to 0.5, with an average of 0.24. In contrast, the average PIC for the Reference and SNP allele rows ranged from 0.006 to 0.5, yielding an average of 0.2 (table 5).

Table 4. 4: DArT SNP marker quality and diversity

Marker quality and diversity	Mean	Min	Max
Call rate	0.82	0.396	1
Reproducibility	0.98	0.006	1
One ratio reference allele	0.8	0.006	1
SNP allele frequency	0.3	0.006	1
Heterozygosity	0.11	0	0.698
PIC for SNP	0.24	0	0.5

Table 4. 5: Shows the SNP position

AlleleID	AlleleSequence	TrimmedSequence	SNP	SnpPosition
100081123	TGCAGAACAGCTACTCTTCCCTTGATATCCATCTTGGGCAGTCATATTCTGTCTTGGTCACAGCTAACC	TGCAGAACAGCTACTCTTCCCTTGATATCCATCTTGGGCAGTCATATTCTGTCTTGGTCACAGCTAACC	33:T>C	33
100081125	TGCAGATAATTCGTCGTGCGAATGAGAATGGGGAAGATCCATTACATCTAAGCAACAGATTTTGTGAAG	TGCAGATAATTCGTCGTGCGAATGAGAATGGGGAAGATCCATTACATCTAAGCAACAGATTTTGTGAAG	19:G>A	19
100081139	TGCAGCTCTGATTCGTAAAGTTAGAGGCACATCATTCTTCTCTAGAGTTGCTTCCTGAGTTTCTGAAAG	TGCAGCTCTGATTCGTAAAGTTAGAGGCACATCATTCTTCTCTAGAGTTGCTTCCTGAGTTTCTGAAAG	20:T>C	20
100081142	TGCAGCTTGAATTTGAAAGCAAGGGCTTCAGGTTTATGTCACCTGCATATGTTTCTGGGACTTCTTGTA	TGCAGCTTGAATTTGAAAGCAAGGGCTTCAGGTTTATGTCACCTGCATATGTTTCTGGGACTTCTTGTA	26:T>C	26
100081143	TGCAGCTTTCGTTTTTGTAAAGTCAGCTGAACCACTAGGGACACATTTTCCATGTGAAAAAAACAACTC	TGCAGCTTTCGTTTTTGTAAAGTCAGCTGAACCACTAGGGACACATTTTCCATGTGAAAAAAACAACTC	23:C>T	23
100081157	TGCAGTAACATCAAACCAAGTATCTTGGTTGGTGGCTTAGCTGCATATATGGTCAGAGAAGACATTGGA	TGCAGTAACATCAAACCAAGTATCTTGGTTGGTGGCTTAGCTGCATATATGGTCAGAGAAGACATTGGA	26:G>A	26
100081166	TGCAGTTATTATGGAAACCAGGGAGGTGGAGCACTTCCTAAAGATACTTCTGTAGCTGTCTGCACAATA	TGCAGTTATTATGGAAACCAGGGAGGTGGAGCACTTCCTAAAGATACTTCTGTAGCTGTCTGCACAATA	59:C>T	59
100081174	TGCAGTTTTGAACCAAGTGTCAGCATCAGTGAATTGCAGACAACCTACATTTTAGATTATAGACCATAA	TGCAGTTTTGAACCAAGTGTCAGCATCAGTGAATTGCAGACAACCTACATTTTAGATTATAGACCATAA	66:T>G	66
100081180	TGCAGAAAAGAAAAGCAGTTGCTGGTCATTCGAAGGATAACCCTGAACATGCAGTAAAACTGTTTCTTC	TGCAGAAAAGAAAAGCAGTTGCTGGTCATTCGAAGGATAACCCTGAACATGCAGTAAAACTGTTTCTTC	31:G>C	31
100081190	TGCAGAATAAGTTGCTGCTAACAGCAATGCAACAAATGCTGGAGGGAGTAAGTGGGGCTCAAAGGAAGA	TGCAGAATAAGTTGCTGCTAACAGCAATGCAACAAATGCTGGAGGGAGTAAGTGGGGCTCAAAGGAAGA	16:G>A	16
100081191	TGCAGAATGAAACAAATACAAATGATGTTATTGCAGTTAGAAAAAGATAAATTTTTTTCCGAGAAATAG	TGCAGAATGAAACAAATACAAATGATGTTATTGCAGTTAGAAAAAGATAAATTTTTTTCCGAGAAATAG	35:G>A	35
100081193	TGCAGAATGAGTAAAAGGGAAGAAAAAGTGGGGGGAATTTGGGAAAGGAGGGAAATTGAGGAGAATAAA	TGCAGAATGAGTAAAAGGGAAGAAAAAGTGGGGGGAATTTGGGAAAGGAGGGAAATTGAGGAGAATAAA	66:A>T	66
100081198	TGCAGAGAGGAGTGTAATTTCTGGAGACAATCAAGGTGGGAAGTTCTGTTTCAATTCTGGCAGCATCTA	TGCAGAGAGGAGTGTAATTTCTGGAGACAATCAAGGTGGGAAGTTCTGTTTCAATTCTGGCAGCATCTA	45:C>T	45
100081202	TGCAGAGGCTCAGATGATCATGGGAAATTGGCAACAATATACAAGTCAGAGAAGAGAGATCAGATGGTA	TGCAGAGGCTCAGATGATCATGGGAAATTGGCAACAATATACAAGTCAGAGAAGAGAGATCAGATGGTA	54:A>G	54
100081203	TGCAGAGGTGACGGTTGGTAGTAAGAACCGCATGTATGAGCGGCCTGACGCATCTGGAGATCTGCTTGT	TGCAGAGGTGACGGTTGGTAGTAAGAACCGCATGTATGAGCGGCCTGACGCATCTGGAGATCTGCTTGT	43:C>T	43
100081208	TGCAGATACACTAACTGTAAAAATGAGTGTCAAGTAAAGGTAAGTGCAAATTGCCTATGCCCCTTTACA	TGCAGATACACTAACTGTAAAAATGAGTGTCAAGTAAAGGTAAGTGCAAATTGCCTATGCCCCTTTACA	50:T>G	50
100081217	TGCAGCAACAGCAGCAACTCTCCCAGCTGCAGCAACAACTGCCACAGCAGCAGCAAATGGTGGGGTCAG	TGCAGCAACAGCAGCAACTCTCCCAGCTGCAGCAACAACTGCCACAGCAGCAGCAAATGGTGGGGTCAG	40:G>A	40
100081223	TGCAGCAGGATCTAAAATTGTCTTTGATTCAGTTTCCGCATGAAGGAGTAAGCAGAGAAGATCAACCTA	TGCAGCAGGATCTAAAATTGTCTTTGATTCAGTTTCCGCATGAAGGAGTAAGCAGAGAAGATCAACCTA	68:A>T	68
100081235	TGCAGCCTCTTTCTGCCAAACTCCGTCTTCCAGCACTTCTCCTGGTTCTCTCCTTCACCATAGCCATGG	TGCAGCCTCTTTCTGCCAAACTCCGTCTTCCAGCACTTCTCCTGGTTCTCTCCTTCACCATAGCCATGG	20:C>T	20
100081239	TGCAGCGTCAGGAATAGATGACTAATAAAAACGTCTTGAAGACACTATTTGACTCCGAAAAACCTGACG	TGCAGCGTCAGGAATAGATGACTAATAAAAACGTCTTGAAGACACTATTTGACTCCGAAAAACCTGACG	31:C>T	31
100081246	TGCAGCTGGTGATCTTGAAGCATGCACTCAATGTGAAAGGAAATGTACGTGGAATTTCTGCTTAGTCTC	TGCAGCTGGTGATCTTGAAGCATGCACTCAATGTGAAAGGAAATGTACGTGGAATTTCTGCTTAGTCTC	54:T>A	54
100081253	TGCAGCTTTAGCAGCCTGCACCCCAGCTAGTGCATCTTCAATAACAATACACTGAGAGAAATAGACAAA	TGCAGCTTTAGCAGCCTGCACCCCAGCTAGTGCATCTTCAATAACAATACACTGAGAGAAATAGACAAA	57:G>A	57
100081263	TGCAGGCTGATAACTATGAAAAGGTCTTCTCCCACATGAAACCTAACAGCATTCTTGGGCTTTCTCATG	TGCAGGCTGATAACTATGAAAAGGTCTTCTCCCACATGAAACCTAACAGCATTCTTGGGCTTTCTCATG	66:A>G	66

Table 4. 6: Extrapolation showing part of the DArT SNPs

AlleleID	AlleleSeq	TrimmedS	Chrom_G(	ChromPos	AlnCnt_G(	AlnEvalue	SNP	SnpPositi(	CallRate	OneRatioRef	OneRatioSnp	FreqHomRef	FreqHomSnp	FreqHets	PICRef	PICSnp	AvgPIC	AvgCount	AvgCount
100081123	TGCAGAA	TGCAGAA	JAJFZO01(	1081492	1	2.57E-29	33:T>C	33	0.932927	1	0.006536	0.993464	0	0.006536	0	0.012986	0.006493	10.81053	4
100081125	TGCAGAT/	TGCAGAT/	JAJFZO01(	1134086	1	2.57E-29	19:G>A	19	0.987805	0.919753	0.358025	0.641975	0.080247	0.277778	0.147615	0.459686	0.30365	32.05376	23.88732
100081139	TGCAGCT(	TGCAGCT(	JAJFZO01(	8071886	1	2.57E-29	20:T>C	20	0.908537	0.993289	0.006711	0.993289	0.006711	0	0.013333	0.013333	0.013333	5.97753	3
100081142	TGCAGCT	TGCAGCT	JAJFZO01(	1380373	1	2.57E-29	26:T>C	26	0.957317	0.993631	0.025478	0.974522	0.006369	0.019108	0.012658	0.049657	0.031157	8.20635	5
100081143	TGCAGCT	TGCAGCT	JAJFZO01(	4803571	1	2.57E-29	23:C>T	23	0.987805	1	0.018519	0.981481	0	0.018519	0	0.036351	0.018176	14.75622	7
100081157	TGCAGTA/	TGCAGTA/	JAJFZO01(	1287261	1	2.57E-29	26:G>A	26	0.97561	0.9875	0.1125	0.8875	0.0125	0.1	0.024687	0.199688	0.112187	23.70918	13
100081166	TGCAGTT/	TGCAGTT/	JAJFZO01(	12952433	1	2.57E-29	59:C>T	59	0.987805	1	0.006173	0.993827	0	0.006173	0	0.012269	0.006135	14.60891	14
100081174	TGCAGTT	TGCAGTT	JAJFZO01(	9636080	1	2.57E-29	66:T>G	66	0.926829	1	0.006579	0.993421	0	0.006579	0	0.013071	0.006536	5.97354	3
100081180	TGCAGAA	TGCAGAA	JAJFZO01(	1587065	1	2.57E-29	31:G>C	31	0.963415	1	0.012658	0.987342	0	0.012658	0	0.024996	0.012498	19.62694	10.5
100081190	TGCAGAA	TGCAGAA	JAJFZO01(	5471128	1	2.57E-29	16:G>A	16	0.95122	1	0.00641	0.99359	0	0.00641	0	0.012738	0.006369	10.77604	3
100081191	TGCAGAA	TGCAGAA	JAJFZO01(	3841676	1	4.29E-27	35:G>A	35	0.823171	1	0.022222	0.977778	0	0.022222	0	0.043457	0.021728	23.40964	8.33333
100081193	TGCAGAA	TGCAGAATGAGTAAA		0	0	999	66:A>T	66	0.920732	0.993377	0.006623	0.993377	0.006623	0	0.013157	0.013157	0.013157	29.06011	6
100081198	TGCAGAG	TGCAGAG	JAJFZO01(	3413525	1	2.57E-29	45:C>T	45	0.945122	0.96129	0.109677	0.890323	0.03871	0.070968	0.074422	0.195297	0.13486	11.21667	4.63636
100081202	TGCAGAG	TGCAGAG	JAJFZO01(	2564438	1	2.57E-29	54:A>G	54	0.737805	1	0.008264	0.991736	0	0.008264	0	0.016392	0.008196	10.44595	6
100081203	TGCAGAG	TGCAGAG	JAJFZO01(	6652954	1	2.57E-29	43:C>T	43	0.890244	0.931507	0.157534	0.842466	0.068493	0.089041	0.127604	0.265434	0.196519	19.8	18.07143
100081208	TGCAGAT/	TGCAGAT/	JAJFZO01(	4479332	1	1.19E-27	50:T>G	50	0.957317	1	0.012739	0.987261	0	0.012739	0	0.025153	0.012577	23.35385	8
100081217	TGCAGCA/	TGCAGCAA	CAGCAG(	0	0	999	40:G>A	40	0.792683	0.923077	0.223077	0.776923	0.076923	0.146154	0.142012	0.346627	0.24432	10.23129	6.85714
100081223	TGCAGCA(	TGCAGCA(	JAJFZO01(	4203361	1	2.57E-29	68:A>T	68	0.646341	0.990566	0.009434	0.990566	0.009434	0	0.01869	0.01869	0.01869	10.96774	10
100081235	TGCAGCC	TGCAGCC	JAJFZO01(	9486987	1	2.57E-29	20:C>T	20	0.689024	0.955752	0.061947	0.938053	0.044248	0.017699	0.08458	0.116219	0.100399	9.02308	7.42857
100081239	TGCAGCG	TGCAGCG	JAJFZO01(	1403201	1	1.19E-27	31:C>T	31	0.95122	0.961538	0.121795	0.878205	0.038462	0.083333	0.073964	0.213922	0.143943	14.22162	9.55
100081246	TGCAGCT(	TGCAGCT(	JAJFZO01(	142616	1	2.57E-29	54:T>A	54	0.707317	0.991379	0.025862	0.974138	0.008621	0.017241	0.017093	0.050386	0.03374	11.85507	6.33333
100081253	TGCAGCT	TGCAGCT	JAJFZO01(	496515	1	2.57E-29	57:G>A	57	0.835366	0.963504	0.043796	0.956204	0.036496	0.007299	0.070329	0.083755	0.077042	4.46951	3.14286
100081263	TGCAGGC	TGCAGGC	JAJFZO01(	4947313	1	2.57E-29	66:A>G	66	0.884146	1	0.006897	0.993103	0	0.006897	0	0.013698	0.006849	7.22807	4
100081264	TGCAGGC	TGCAGGC	JAJFZO01(	3849388	1	2.57E-29	65:C>T	65	0.987805	0.993827	0.006173	0.993827	0.006173	0	0.012269	0.012269	0.012269	10.90955	8
100081270	TGCAGGG	TGCAGGG	JAJFZO01(	8530295	1	1.19E-27	55:T>G	55	0.878049	0.013889	0.986111	0.013889	0.986111	0	0.027392	0.027392	0.027392	5	3.41916
100081274	TGCAGGT	TGCAGGTTATGCTAC(		0	0	999	65:T>A	65	0.640244	0.914286	0.161905	0.838095	0.085714	0.07619	0.156735	0.271383	0.214059	24.48276	14.19048
100081283	TGCAGTA(	TGCAGTACATACAGA		0	0	999	54:T>A	54	0.603659	0.838384	0.262626	0.737374	0.161616	0.10101	0.270993	0.387307	0.32915	7.20952	7.07143

Table 4. 7: SilicoDArT data

CloneID	AlleleSeq	TrimmedSeq	Chrom_G	ChromPos_	AlnCnt_G	AlnEvalue_	CallRate	OneRatio	PIC	AvgReadDepth	StDevReadDepth	Qpmr	Reproducibility
1E+08	TGCAGCA	TGCAGCAAA	JAJFZO01	7182074	1	2.00E-25	0.987805	0.5	0.5	18.7767	12.18142	1.34569	1
1E+08	TGCAGGC	TGCAGGCATC	JAJFZO01	7234507	1	2.32E-18	0.95122	0.5	0.5	91	50.20822	1.88081	1
1E+08	TGCAGGT	TGCAGGTGA	JAJFZO01	2234104	1	2.28E-10	0.926829	0.5	0.5	43.76106	24.93328	1.82621	0.96875
1E+08	TGCAGGT	TGCAGGTTTGGTTTTTGG		0	0	999	0.97561	0.5	0.5	16.35577	11.29663	1.33205	1
1E+08	TGCAGAA	TGCAGAAAACCCTTTTCA		0	0	999	0.926829	0.5	0.5	52.76577	31.15204	1.68763	1
1E+08	TGCAGTT	TGCAGTTCCA	JAJFZO01	1745915	1	2.57E-29	0.97561	0.5	0.5	12.48485	8.21844	1.36118	1
1E+08	TGCAGCC	TGCAGCCAAT	JAJFZO01	3364433	1	2.57E-29	0.97561	0.5	0.5	55.0297	37.0123	1.41039	0.972222
1E+08	TGCAGTT	TGCAGTTCACATTATGTA		0	0	999	0.987805	0.5	0.5	17.20192	11.0178	1.31335	1
1E+08	TGCAGTT	TGCAGTTTGA	JAJFZO01	3545927	1	1.19E-27	0.97561	0.5	0.5	31.69524	19.18885	1.63721	0.967742
1E+08	TGCAGTG	TGCAGTGAA	JAJFZO01	728694	1	1.21E-17	0.981707	0.496894	0.499981	14.27778	9.7978	1.3863	0.971429
1E+08	TGCAGAA	TGCAGAAGC	JAJFZO01	1189910	1	3.32E-28	0.981707	0.496894	0.499981	23.06	14.7308	1.39655	1
1E+08	TGCAGTG	TGCAGTGAC	JAJFZO01	1927057	1	3.27E-22	0.981707	0.503106	0.499981	76.65094	45.20249	1.51127	1
1E+08	TGCAGAT	TGCAGATGATGACGAAT		0	0	999	0.981707	0.503106	0.499981	47.79381	32.58389	1.30261	1
1E+08	TGCAGTT	TGCAGTTGTA	JAJFZO01	3339549	1	2.57E-29	0.981707	0.496894	0.499981	47.09524	32.57885	1.42585	0.972222
1E+08	TGCAGCC	TGCAGCCATCTTCCACCT		0	0	999	0.969512	0.496855	0.49998	25.63462	15.48762	1.62878	1
1E+08	TGCAGTT	TGCAGTTGAG	JAJFZO01	3953646	1	2.57E-29	0.969512	0.496855	0.49998	18.52679	12.48997	1.43282	1
1E+08	TGCAGAA	TGCAGAAAG	JAJFZO01	1059215	1	5.55E-26	0.957317	0.503185	0.49998	28.26852	16.4354	1.80234	1
1E+08	TGCAGGC	TGCAGGCAA	JAJFZO01	5178365	1	1.19E-27	0.957317	0.496815	0.49998	33.43396	20.50173	1.63188	1
1E+08	TGCAGTG	TGCAGTGCAC	JAJFZO01	5647666	1	5.55E-26	0.969512	0.503145	0.49998	46.29126	27.30896	1.56535	1
1E+08	TGCAGTA	TGCAGTACTG	JAJFZO01	336292	1	1.01E-16	0.969512	0.496855	0.49998	121.25	64.25313	2.02043	0.967742
1E+08	TGCAGTA	TGCAGTAAGA	JAJFZO01	414810	1	5.55E-26	0.957317	0.496815	0.49998	42.83929	25.23056	1.46251	1
1E+08	TGCAGTT	TGCAGTTCTA	JAJFZO01	904378	1	1.20E-22	0.932927	0.496732	0.499979	7.92523	5.24128	1.48091	0.971429
1E+08	TGCAGCC	TGCAGCCCAA	JAJFZO01	5589654	1	9.29E-24	0.945122	0.496774	0.499979	56.2973	32.72368	1.69628	1
1E+08	TGCAGAT	TGCAGATTGA	JAJFZO01	1184248	1	3.98E-21	0.932927	0.496732	0.499979	62.34906	35.35428	1.78743	0.96875
1E+08	TGCAGAA	TGCAGAAATC	JAJFZO01	8307354	1	2.58E-24	0.932927	0.503268	0.499979	82.93	41.43604	2.06084	1
1E+08	TGCAGCG	TGCAGCGGAC	JAJFZO01	986015	1	5.55E-26	0.945122	0.496774	0.499979	43.26923	25.2637	1.85436	1
1E+08	TGCAGTG	TGCAGTGGTG	JAJFZO01	3256843	1	2.57E-29	0.945122	0.503226	0.499979	29.83654	17.29891	1.63022	0.970588
1E+08	TGCAGTG	TGCAGTGCAA	JAJFZO01	6440321	1	2.57E-29	0.920732	0.496689	0.499978	11.32743	7.61048	1.66402	0.967742

4.3. Population Genetics

4.3.1 Bar plot on the Number of Individuals per Population

Diani-Ukunda, India, Kibwezi Town, Kilifi, Likoni, Maungu, Mbololo, Miasenyi, Mwakiki, Pwani University and Shika Adabu provenances had 10 genotypes each whereas Arabuko Sokoke, Gede, Mackinon, Samburu, Vipingo and Waa Kwale had provenances had 9 genotypes as shown in fig. 10.

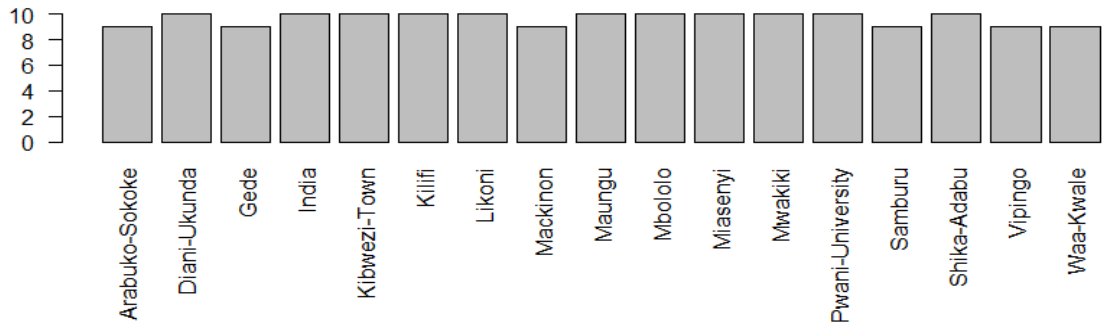


Figure 4. 5: Bar plot showing populations/ provenances

4.3.2 F Statistics

For Fst, the functions StAMPP package were used as they enable parallel computing hence faster. Fst nboots = 1, percentage was 95 and nclusters = 1. The genetic distances were in centiMorgans (cM)

Table 4. 8: Fst table showing the distances in centiMorgans

	Arabuko-Sokoke	Diani-Ukunda	Gede	India	Kibwezi-Town	Kilifi	Likoni	Mackinon	Maungu	Mbololo	Miasenyi	Mwakiki	Pwani-University	Shika-Adabu	Samburu	Vipingo	Waa-Kwale
Arabuko-Sokoke	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Diani-Ukunda	0.08839573	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Gede	0.08350471	0.05436086	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
India	0.09123628	0.10682290	0.0831651876	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Kibwezi-Town	0.09667251	0.10330296	0.1358056494	0.13164604	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Kilifi	0.05980152	0.04185168	0.0809777092	0.09218634	0.06229153	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Likoni	0.23878713	0.16166061	0.1723103958	0.19987526	0.23928502	0.16178895	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Mackinon	0.27382156	0.24313803	0.2851214155	0.28918113	0.25815672	0.21890244	0.40816105	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Maungu	0.08556645	0.06754122	0.1125005719	0.12296296	0.08263517	0.01545695	0.18599915	0.2446184	NA	NA	NA	NA	NA	NA	NA	NA	NA
Mbololo	0.12637178	0.08676262	0.0908647419	0.11953297	0.15208738	0.08184686	0.02263307	0.3232576	0.10975701	NA	NA	NA	NA	NA	NA	NA	NA
Miasenyi	0.22036031	0.17774202	0.1937571500	0.20481694	0.23564218	0.17685166	0.19439592	0.4105880	0.20447853	0.17187817	NA	NA	NA	NA	NA	NA	NA
Mwakiki	0.22037932	0.13525650	0.1419753162	0.18225293	0.21543672	0.15565820	0.09504356	0.4012120	0.18731091	0.07154605	0.1881584	NA	NA	NA	NA	NA	NA
Pwani-University	0.19971295	0.22030874	0.2620468680	0.24810477	0.21604694	0.14690566	0.37524804	0.3183340	0.19077273	0.28724940	0.3614970	0.3717897	NA	NA	NA	NA	NA
Shika-Adabu	0.11499504	0.11537405	0.1323495056	0.13495842	0.06140674	0.09905247	0.22643222	0.2960261	0.12646844	0.14104922	0.2338500	0.1948355	0.2572970	NA	NA	NA	NA
Samburu	0.06033521	0.04156700	0.0003597574	0.06309578	0.10437766	0.02906750	0.13491462	0.2425008	0.07033899	0.06255422	0.1729000	0.1281419	0.2160161	0.1113208	NA	NA	NA
Vipingo	0.08982599	0.04312460	0.0648701125	0.10214033	0.09899893	0.02812665	0.09184752	0.2474152	0.05503756	0.03094450	0.1659348	0.1012815	0.2093105	0.1174521	0.03704638	NA	NA
Waa-Kwale	0.11575017	0.09217460	0.0829989068	0.12887380	0.14181266	0.08259242	0.14365688	0.2877884	0.10662040	0.07174563	0.2027366	0.1505751	0.2641509	0.1388378	0.06071000	0.05671634	NA

Table 4. 9: Population genetic distances using Euclidean as a distance measure

```
  Distance measure: euclidean
    No. of populations = 17
    Average no. of individuals per population = 9.647059
    No. of loci = 20921
    Miniumum Distance:  19.32
    Maximum Distance:  48.45
    Average Distance:  33.024

Completed: gl.dist.pop
                 Arabuko-Sokoke Diani-Ukunda     Gede    India Kibwezi-Town   Kilifi   Likoni Mackinon   Maungu  Mbololo
Diani-Ukunda           31.22686
Gede                   33.60015      28.17896
India                  31.70141      30.38961 30.71674
Kibwezi-Town           30.58207      28.96980 33.57191 31.67714
Kilifi                 27.95676      23.67949 29.62305 28.59167      24.74401
Likoni                 38.27676      31.25721 32.96569 34.65025      37.02305 31.45459
Mackinon               44.88385      40.78608 45.22790 45.02551      40.99392 39.02065 47.81913
Maungu                 29.98550      26.02887 31.98760 31.13192      26.46278 20.04084 32.81997 40.22005
Mbololo                33.50289      28.30668 30.45371 31.23737      32.84962 27.40933 19.32103 45.07161 29.52481
Miasenyi               37.11695      32.64182 34.71259 35.31060      36.91502 32.88203 30.36716 48.45087 34.47238 31.28294
Mwakiki                37.76208      29.93681 31.63940 33.98191      35.71957 31.66914 23.32104 47.81128 33.60243 24.16349
Pwani-University       37.65138      37.59373 42.18504 40.48078      36.77890 31.72450 46.01462 43.07642 34.80793 41.99274
Samburu                30.90221      26.11825 25.45931 28.17443      30.49295 23.96683 30.11192 42.13376 27.73024 27.67423
Shika-Adabu            31.72473      29.76132 32.88128 31.63799      24.06153 27.99242 35.38716 43.29907 30.11396 31.49249
Vipingo                32.16829      25.28038 29.84218 30.68687      29.10405 22.95677 25.90581 41.36523 25.36065 24.02726
Waa-Kwale              35.33271      30.69813 32.56732 34.02535      33.65759 29.34749 30.16808 44.18125 30.99582 28.37661
                    Miasenyi  Mwakiki Pwani-University  Samburu Shika-Adabu  Vipingo
Diani-Ukunda
Gede
India
Kibwezi-Town
Kilifi
Likoni
Mackinon
Maungu
Mbololo
Miasenyi
Mwakiki             30.39267
Pwani-University    45.22906 46.18995
Samburu             33.20795 30.52589          38.62910
Shika-Adabu         36.20826 33.64458          39.53715 30.79956
Vipingo             31.88543 27.56765          36.96960 26.59748     30.34971
Waa-Kwale           34.77817 31.58159          41.52538 29.99404     32.83590 28.69758
```

4.3.3 Population Euclidean Distances

The genetic distance between the 17 populations of *M. oleifera* genotypes were calculated and computed using DArT R package, based on Euclidean as a measure of distance. It was evident that the Euclidean distances ranged from 19.32 and 48.45 with an average distance of 33.024. The least distance was that of Mbololo and Likoni while the highest distance was between Miasenyi and Mackinon.

4.3.4 Analysis of Molecular Variance (AMOVA)

Analysis of molecular variance (AMOVA) was utilized to carry out population genetic analysis using KDCompute Software (silicoDArT) and DArT R. The standard deviation and variance for individual alleles were calculated. The sum of the tag read counts for all samples, divided by the number of samples with non-zero tag read counts (AvgReadDepth) had a range of 255.26 to 5.05 with an average of 39.8. The standard deviation of the number of tag reads for all samples with non-zero tag read counts (StDevReadDepth) ranged from 0 to 124.14 exhibiting an average of 18. 96. The average of the normalized non-zero tag read counts divided by the standard deviation of the normalized non-zero tag read counts which is the variance (Qpmr), exhibited a range of 1.25 and 2262.8 with an average Qpmr of 2.58. Reproducibility ranged from 0.95 to 1 and had an average reproducibility of 0.99. From these data, it is evident that the average of the sum of the tag read counts for all the SNPs called was 39.8, the standard deviation of the SNPs was 18.96 whereas the variance of the SNPs was 2.58. From DArT R, the table below summarizes the values of analysis of molecular variance.

4.3.5 Neighbor-Joining Tree

It constructs a phenetic tree using a neighbor-joining approach or UPGMA using DArT R and thus enables the visualization of the populations' genetic similarity. Euclidean distances were utilized. A phylogenetic tree with 17 tips and 15 internal nodes was resultant. It is evident from the phenetic tree (figure 11) that there were 4 clusters of the *M. oleifera* provenances. The 1st cluster had 4 populations. Pwani University and Mackinon populations were sub grouped together, indicating their close relationship. Maungu and Kilifi were also sub clustered together. In the 2nd cluster were Shika Adabu and Kibwezi Town whereas in cluster 3 only Arabuko Sokoke was evident. Lastly, cluster 4 had 10 populations clustered together. Diani-Ukunda, India, Samburu, Gede, Waa Kwale, Vipingo, Miasenyi, Mwakiki, Mbololo and Likoni provenances were all clustered here. This shows their close genetic similarity.

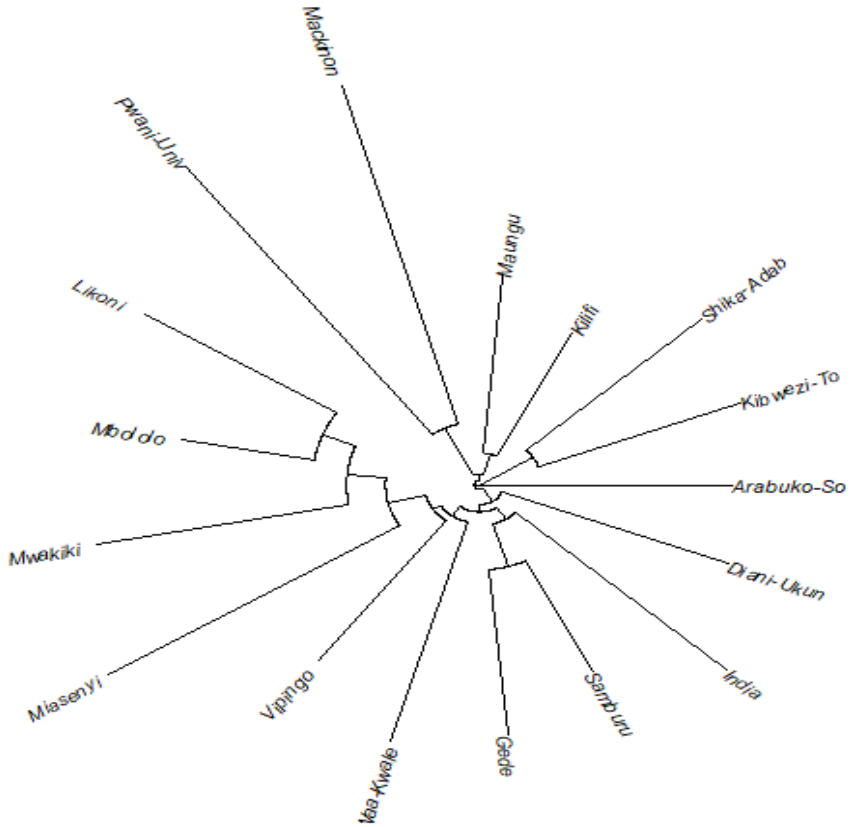


Figure 4. 6: Phenetic tree for the population using Neighbor-joining approach from DArT

CHAPTER FIVE

DISCUSSION

5.1 Characterization of *Moringa oleifera*

5.1.1 Clustering Analysis

In the cluster analysis, the accessions were not clustered according to their geographical relatedness of collections as evident in table 2 and figure 6, where four clades were notable. There were instances where all genotypes clustered in one clade, notably all 9 genotypes from Mackinon clustered in Clade I, all 10 genotypes from Shika Adabu clustered in Clade II as well as all 10 genotypes from Likoni, Miasenyi and Mwakiki clustering in Clade III. Even though all the said genotypes were clustered in one clade, they sub-clustered with other genotypes from different provenances. For instance, 7genotypes from Miasenyi sub-clustered with 1 genotype from Likoni. 8 genotypes from Pwani University clustered in Clade IV with 7 of them sub-clustering together. 6 genotypes from India clustered in Clade I but only 5 sub-clustered together. Another notable sub-clade had 15 genotypes whereby 3 Shika Adabu were clustered alone whereas the other 7 genotypes were clustered together with 1 genotype from India and 4 from Kibwezi Town.

(Ravi et al., 2020) explained that *M. oleifera* is an out-crossed species expected to maintain variation within populations than between populations.

It is notable from simple clustering dendrogram (from DArT R) that there were 163 samples as opposed to the 164 samples that were analyzed. This could be due to absolute similarity

from sequence duplicity or close homology of two of the genotypes which could have been algorithmically collapsed into one sample.

5.1.2 Hierarchical Clustering

As evident in table 3, all Miasenyi genotypes clustered in Clade I, 2 of the genotypes sub-clustered together with a similarity coefficient of 100% while the other 8 genotypes sub-clustered with 1 Likoni genotypes exhibiting a similarity coefficient of 94%. This can be attributed to a high similarity between the genotypes. The other 9 Likoni genotypes were clustered in Clade II, where they further sub-clustered with genotypes from varied provenances in the 3rd, 4th and 5th sub-clusters and exhibited different similarity indexes. Mwakiki had all 10 genotypes in Clade II whereby 9 clustered with 2 Vipingo genotypes and 1 genotype from Maungu, Waa Kwale and Likoni with a similarity coefficient of 82% while one genotype clustered with one genotype from Diani-Ukundu and Samburu with a similarity coefficient of 79% in the 2nd sub-clade. Notably, all 9 genotypes of Mackinon were clustered in Clade III where they sub-clustered in the 3rd sub-clade with a similarity coefficient of 91%. All 10 Pwani University genotypes clustered in Clade III but they further sub-clustered with different genotypes. 7 of the genotypes sub-clustered together with a similarity coefficient of 100%, indicating that they were 100% similar.

(Hassani et al., 2020) utilized start codon targeted (SCoT) markers. The 10 plant samples of *M. oleifera* were divided into two main distinct clusters based on genetic similarity and intraspecific differentiation, with roughly 60% genetic similarities between them. Thus, from our study it is evident that SNPs showed a higher similarity coefficient(63% to 100%) which was similar with the clustering analysis with a wide range of similarity coefficients

that revealed a high level of genetic variation among the genotypes from the two countries whereby the similarity coefficient ranged from 38% to 89% (Rufai et al., 2013). Ten *M. oleifera* accessions from different Indonesian islands were successfully identified genetically through molecular analysis utilizing SRAP molecular markers. The obtained polymorphism percentage of 81.40% suggests a comparatively large genetic diversity of *M. oleifera* (Ridwan et al., 2021). Using RAPD markers, (A. V. C. da Silva et al., 2012) examined the genetic diversity of 17 accessions of *M. oleifera* from the Germplasm Bank and found that it ranged from 0% to 52%. According to the study, genetic diversity values of 55% or below were determined to be low. The UPGMA clustering produced three clusters with an average similarity between the accessions of 48%, ranging from 27% to 69%. The genetic diversity of provenances of *M. oleifera* from Costal regions of Tanzania was studied using RAPD primers, and the results of the cluster analysis revealed five clusters with similarity ranging from 54% to 96% according to (Mgendi et al., 2010). As per (Ravi et al., 2020), an excellent predictor of *M. oleifera* genetic diversity is the number of polymorphic loci or the percentage of polymorphisms or similarity. The percentage of polymorphisms in *M. oleifera* genetic diversity reported from multiple earlier research was lower than this finding, such as 48.68% in *M. oleifera* from India (Saini et al., 2013) and 62% in *M. oleifera* from the United States using RAPD markers (A. V. C. da Silva et al., 2012), 66.5% in *M. oleifera* from India, Malawi and Kenya (Muluvi et al., 2004)using AFLP markers, 55.9% in *M. oleifera* from Timor Island, Indonesia (Kleden et al., 2017), 40.00% using cytochrome p450 in *M. oleifera* from India and 48.57% using ISSR markers (Saini et al., 2013).

There have also been reports of polymorphisms with comparable or even slightly higher genetic similarity percentages, such as those by (Ojuederie et al., 2013) in their study on the

M. oleifera plant from Nigeria, which used RAPD markers and got 81.50%, (kumar Ganesan et al., 2014) in their study of the *M. oleifera* from India, using SSR markers to get 82.86%, and (Ravi et al., 2020) who studied *M. oleifera* from India and had a similarity coefficient of 88.25% using cytochrome p450.

The *M. oleifera* genotypes from most of the populations were seen clustering together in the clades III and IV. This demonstrates that there is no clear correlation between the genetic distance as indicated by the dendrogram and the geographic origin. According to this result, the genetic diversity within and between these two groups cannot be attributed to their geographic distance. This observation also suggests that it is not possible to attribute the demographic divergence seen in these populations to isolation brought on by distance.

Comparing and contrasting the two clustering techniques revealed some discrepancies. From the study it is evident from UPGMA clustering analyses there were 4 clades which differed in the number of genotypes in each clade per cluster analysis. There were several noticeable sub clades too from the hierarchical clustering analysis. This was also reported by (A. V. C. da Silva et al., 2012). Some authors (Costa et al., 2011; Dos Santos et al., 2011) have employed regrouping as an alternative to make the results more insightful. The employment of multiple grouping techniques prevents erroneous conclusions due to variations in group classification, optimization, and sorting. As a result, this choice is made when allocating materials among a specific collection of genotypes (Arriel et al., 2006). According to (Shahzad et al., 2013), M. oleifera accessions from 9 distinct countries (India, Senegal, Tanzania, Mozambique, Zimbabwe, Florida, Belize, Mexico, and Haiti) shared a high degree of genetic similarity with one accession from Pakistan (Punjab accession) and exhibited a

low genetic diversity with each other. In light of this, they made the assumption that the accretions from the nine various countries came from Punjab and were transferred by humans to the other nine countries. This is evident in our study as the 164 genotypes exhibited a low genetic diversity and can be attributed to the Moringa plant in Coastal Kenya having come from a common ancestry, that is, India.

The absence of geographic clustering suggests that the genetic differences between individuals from various geographic locations are not very significant. This might be brought on by the movement of seeds, cuttings, and/or high rates of gene flow between the nearby populations (kumar Ganesan et al., 2014).

Start codon targeted (SCoT) markers were used by (Hassani et al., 2020). The 10 plant samples of *M. oleifera* were divided into two main distinct clusters based on genetic similarity and intraspecific differentiation, with roughly 0.60 genetic similarity between them. Thus, from our study it is evident that SNPs showed a higher similarity coefficient which was quite similar with the clustering analysis with a wide range of similarity coefficients that revealed a high level of genetic variation among the genotypes from the two countries whereby the similarity coefficient ranged from 0.38 to 0.89 (Rufai et al., 2013)**.**

Miasenyi had all of its genotypes clustered in clade I, which can be attributed to less variability between the genotypes. The *M. oleifera* genotypes from most of the populations were seen clustering together in the clades III and IV. This demonstrates that there is no clear correlation between the genetic distance as indicated by the dendrogram and the geographic origin. According to this result, the genetic diversity within and between these two groups cannot be attributed to their geographic distance. This observation also suggests that it is not

possible to attribute the demographic divergence seen in these populations to isolation brought on by distance.

Interestingly, the within and between counties analysis of *M. oleifera* rendered high similarity coefficients that cannot be attributed to their geographic distance. This observation also suggests that it is not possible to attribute the demographic divergence seen in these populations to isolation by distance. The absence of geographic clustering suggests that the genetic differences between individuals from various geographic locations are not very significant. This might be brought on by the movement of seeds, cuttings, and/or high rates of gene flow between the nearby populations (kumar Ganesan et al., 2014.

5.1.3 Principal Coordinate Analysis

The PCoA frequently determines the eigenvalues and eigen-vectors of a matrix that represents all of the distances between the data points, measured using the Euclidean or Gower distance. It creates a 2 or 3 dimensional scatter map of the samples with the least amount of distortion possible, with the distances between the samples in the plot reflecting their genetic distances (Bhandari et al., 2017).The *M. oleifera* SNP profile classified the 164 genotypes of into four major groups (Fig. 9). This grouping pattern corresponds with that of clustering analysis as shown by the hierarchical dendrogram.

The genetic similarity of individuals and populations were visualized by a principal coordinate analysis (PCoA) plot, which represented the samples based on the eigenvalues in the two-dimensional graph with respect to the first two principal coordinates. Principle component (PC) 1 amounted to 71% variance and the 2nd Principle component had 15% variance as evident in the PCA plot.

In PCoA, all accessions were designated with various colors based on their different population numbers to indicate their region specificity (Fig. 9). The intermixing of color across the coordinates further supports the UPGMA tree (Fig. 9), which shows that there was no location-specific grouping of the genotypes. (kumar Ganesan et al., 2014) also found out that there was no geographic specific grouping in their study and concluded that lack of clustering in accordance to geographical areas was indicative of individuals from different geographical areas lacking a significant difference genetically. It is thus evident that there was a high genetic similarity between the genotypes studied.

Evidently, seven genotypes from Pwani University consistently clustered together and separately from the rest of the genotypes across all the analyses. These genotypes seem to be unique requiring further analysis. It is possible that they represent independently sourced germplasm in comparison to the other provenances that indicate a single seed source.

5.1.4 3D Plot

From figure 9, it is evident that 7 genotypes of Pwani University clustered together and further from the rest of the genotypes. This is also evident from figure 7 and 8 which show that 7 genotypes of Pwani university clustered together in Clade IV. The rest of the genotypes clustered together and can thus, be deduced that grouping of the genotypes was not specific to the location of origin.

It's noteworthy that the genotypes from Pwani University clustered separately. It is remarkable that regardless of the technique employed, these genotypes consistently grouped uniquely and independently. This may be as a result of the seeds' distinctive origin. Nevertheless, the *M. oleifera* plants in the different provenances may share a common ancestor, that may account for the great resemblance between the genotypes.

5.2 To Identify Polymorphisms in the Selected Varieties.

As evident in Table 6, the allele with allele ID 100081123|F|0-33:T>C-33:T>C whose TGCAGAACAGCTACTCTTCCCTTGATATCCATCTTGGGCAGTCATATTCTGTCT TGGTCACAGCTAACC had a mutation or base pair change at position 33. The nucleotide change was from T to C hence the SNP was at position 33.

Another allele with Allele ID 100081125|F|0-19:G>A-19:G>A with the sequence TGCAGATAATTCGTCGTGCGAATGAGAATGGGGAAGATCCATTACATCTAAGC AACAGATTTTGTGAAG had a single nucleotide polymorphism at position 19 from G to A.

The call rate for the 20,921 SNPs ranged from 0.396 to 1 with an average call rate of 0.82. The proportion of samples for which the genotype score is "1" was averagely 0.30 with a minimum score of 0.006 and a maximum score of 1. The SNP exhibited a reproducibility of 0.98. There was an evidently high reproducibility and consistent marker score.

PIC for the SNPs was averagely 0.24 with a maximum PIC being 0.5. PIC is divided into three categories with regard to the polymorphism value, high (PIC value higher than 0.5), medium (value between 0.25 and 0.5) and low (lower than 0.25) (Vaiman et al., 1994; Xie et al., 2010)**.** In a study conducted by (Rajalakshmi et al., 2019), (Genetic Diversity, Population Structure and Correlation Study in *M. oleifera* Lam. Using ISSR and SRAP) Markers, fifteen inter-simple sequence repeat (ISSR) primers produced the average PIC value of 0.53 while the genetic variability study using sequence-related amplified polymorphism (SRAP) primer combinations exhibited an average PIC of 0.44. (Muslihatin et al., 2022) used ISSR markers

to study the genetic diversity of *M. oleifera* where the average PIC value in the results of their study was 0.46 (table 5).

(kumar Ganesan et al., 2014) carried out Genetic diversity and population structure study of drumstick by utilizing morphological and SSR markers whereby the PIC value for SSR primers exhibited an average of 0.15. For the characterization of *M. oleifera*, (Saini et al., 2013) utilized RAPD, ISSR, and cytochrome P450 makers and reported high PIC values of 0.72, 0.81, and 0.68 for RAPD, ISSR, and cytochrome P450, respectively. It is evident that the average PIC for the SNP markers was low. (kumar Ganesan et al., 2014) postulated that the reason for lower PIC could be due to the fact that SSR markers generated on average 1.84 bands per primer followed by minimum number of unique bands generated in the genotypes studied in comparison with (Muluvi et al., 1999) who reported 59 bands per primer with AFLP marker. In our study, this could also be attributed to SNP markers being changes on single nucleotide bases. (Sousa et al., 2017) found a mean PIC value of 0.35 with 11,187 SNP markers while studying *C. Arabica* and concluded that the low PIC value evidences the narrow genetic base of *C. Arabica*. Since PIC measures a marker's capacity to detect polymorphism, it can be assumed that the SNPs' low PIC average was brought on by the marker's specificity and selectivity. The heterozygosity of the loci ranged from 0 to 0.698 and exhibited an average heterozygosity of 0.11. This heterozygosity average was similar to that of (kumar Ganesan et al., 2014) average heterozygosity across 20 loci was 0.10.

The sum of the tag read counts for all samples, divided by the number of samples with non-zero tag read counts (AvgReadDepth) had a range of 255.26 to 5.05 with an average of 39.8. The standard deviation of the number of tag reads for all samples with non-zero tag read

counts (StDevReadDepth) ranged from 0 to 124.14 exhibiting an average of 18. 96. The average of the normalized non-zero tag read counts divided by the standard deviation of the normalized non-zero tag read counts which is the variance (Qpmr), exhibited a range of 1.25 and 2262.8 with an average Qpmr of 2.58. Reproducibility ranged from 0.95 to 1 and had an average reproducibility of 0.99.

M. oleifera being a cross-pollinated tree, high heterogeneity in many character forms was observed resulting in vast genetic diversity. *M. oleifera* seems to be a very polymorphic plant since there were several single nucleotide polymorphisms observed. With an average call rate of 95%, 116,190 SilicoDArT markers were called on the 105 Napier grass accessions. The greatest anticipated heterozygosity (He) and polymorphism information content (PIC) values for biallelic markers were both 0.5. The expected heterozygosity (He) varied from 0 to 0.5, and the PIC ranged from 0 to 0.38. He and PIC values were generally 0.24 and 0.19, respectively. With an average call rate of 87%, a total of 85,452 SNP markers were called on the accessions. According to (Muktar et al., 2019), the PIC values ranged from 0 to 0.38 with an average of 0.11 and the He values ranged from 0 to 0.5 with an average of 0.13.

5.3 Population Genetics Analysis/ Study

In-depth knowledge and understanding of the gene flow pattern and population genetic structure in *M. oleifera* through molecular genetic diversity and population genetic structure of worldwide collections is of immense promise. Documentation of genetic diversity and conservation of germplasm is a necessity to develop elite varieties.

5.3.1 Population Euclidean Distances

The genetic distance between the 17 populations of *M. oleifera* genotypes were calculated and computed using DArT R package, based on Euclidean as a measure of distance (table 10). It was evident that the Euclidean distances ranged from 19.32 and 48.45 with an average distance of 33.024. The least distance was that of Mbololo and Likoni while the highest distance was between Miasenyi and Mackinon. The genetic distances exhibited between the populations was relatively low and could thus be concluded that there was least variability in the populations. This could be attributed to the spread of seeds as most of the provenances are geographically close and the source of planting material arising from a common ancestry. According to (Mgendi et al., 2010; Muluvi et al., 1999),the core collection of *M. oleifera* populations extant in Africa consists of a relatively small number of accessions imported from India, which has led to low levels of genetic diversity.

5.3.2 F Statistics

F statistics is used to test variances of populations (Gruber et al., 2018). Samburu and Gede exhibited the least variance 0.0003 whereas Pwani University and Mwakiki showed the highest variance with 0.37. A moderate Fst value would indicate minimal significant genetic divergence between the populations of *M. oleifera*, according to (Rufai et al., 2013). This study exhibited low Fst values between the populations signaling very low genetic divergence (table 9).

5.3.3 Neighbor Joining Tree

The 1st cluster had 4 populations (figure 12). Pwani University and Mackinon populations were sub grouped together, indicating their close relationship. Maungu and Kilifi were also sub clustered together. In the 2nd cluster was Shika Adabu and Kibwezi Town whereas in

cluster 3 only Arabuko Sokoke was evident. Lastly, cluster 4 had 10 populations clustered together. Diani-Ukunda, India, Samburu, Gede, Waa Kwale, Vipingo, Miasenyi, Mwakiki, Mbololo and Likoni provenances were all clustered here. This shows their close genetic similarity (Fig. 12).

Mackinon and Pwani University cluster together as evident in figure 12 yet they are geographically distant as shown in figure 5. This observation is also true for Maungu and Kilifi, Samburu and Gede ad also Mbololo and Likoni. Hence, the conclusion that no geographic isolation was evident and could be resultant from the accessions having a common ancestry. This could also be explained from the perspective that no evolution has taken place to guarantee change in genetic makeup of *M. oleifera* since it was introduced less than a century ago by the Indians.

In the groups analyzed, no obvious geographic isolation was observed. (Rufai et al., 2013) and (Mgendi et al., 2010) reported a similar finding based on RAPD markers. This further demonstrates that population genetic divergence cannot be explained solely by geographic distance (Verma et al., 2015). Similar findings in other plant species have been reported by numerous studies (Singhet al., 2014; (Bronzini de Caraffa et al., 2002), but some of these studies were able to distinguish genotypes depending on geographic origin (Desplanque et al., 1999); (Muluvi et al., 1999) (Li & Nelson, 2002).

5.3.4 Analysis of Molecular Variance (AMOVA)

AMOVA (analysis of molecular variance) was carried out to detect population differentiation utilizing molecular markers. Analysis of molecular variance (AMOVA) was utilized to carry out population genetic analysis using KDCompute Software (silicoDArT).

The standard deviation and variance for individual alleles were calculated. The sum of the tag read counts for all samples, divided by the number of samples with non-zero tag read counts (AvgReadDepth) had a range of 255.26 to 5.05 with an average of 39.8. The standard deviation of the number of tag reads for all samples with non-zero tag read counts (StDevReadDepth) ranged from 0 to 124.14 exhibiting an average of 18. 96. The average of the normalized non-zero tag read counts divided by the standard deviation of the normalized non-zero tag read counts which is the variance (Qpmr), exhibited a range of 1.25 and 2262.8 with an average Qpmr of 2.58. Reproducibility ranged from 0.95 to 1 and had an average reproducibility of 0.99. From these data, it is evident that the average of the sum of the tag read counts for all the SNPs called was 39.8, the standard deviation of the SNPs was 18.96 whereas the variance of the SNPs was 2.58.

The molecular variance analysis (AMOVA) described a 2.55%, variation within the population and 2.73% among the populations. In this study the higher variation was among population. AMOVA was carried out for SSR data and 2% variation was found between regions; 3% of variation was found among the populations and the maximum variation was observed within populations (95%). The majority of variation within populations is anticipated to be maintained by woody, perennial, out-crossed species (Hamrick et al., 1989). In contrast to our finding that the greatest variation was seen between populations, (Muluvi et al., 1999) studied four pairs of AFLP primer combinations and found significant variation between regions and populations. However, (kumar Ganesan et al., 2014; Rufai et al., 2013), reported 95% of the total genetic variation within the population. The high similarity between the genotypes could be attributed to the *M. oleifera* plants in the various

provenances having the same ancestry. It could also be resultant from the samples having been introduced from different sources.

CHAPTER SIX

CONCLUSSION AND RECOMMENDATION

6.1 Conclusion

This research has provided a comprehensive genetic characterization of *Moringa oleifera*, shedding light on its genomic diversity and genetic population analysis. Through molecular markers analysis, utilizing SNPs and genotyping, essential insights into the genetic diversity of this invaluable plant have been unearthed.

It was evident from this study that there was less genetic diversity. Hierarchical clustering dendrogram showed a high similarity coefficient between the genotypes and thus less genetic divergence. Domestication of plants can cause reduction in genetic diversity. *M. oleifera* in Kenyan Coast originated from India and has since been domesticated.

It was also evident that the genotypes did not cluster according to their geographical location. Cluster analysis and population study showed that, no geographical isolation exists between genotypes from the Coastal region of Kenya.

This could imply that spread of planting materials has taken place in the form of cuttings, seed and/or high rates of gene flow between the adjacent populations. This could also be explained from the perspective that no evolution has taken place to guarantee change in genetic makeup of *M. oleifera* since it was introduced less than a century ago by the Indians.

This study contributes significantly to the body of knowledge on *M. oleifera* genetics, paving the way for future research endeavors and practical applications aimed at harnessing the full potential of this remarkable plant. The management, conservation, association mapping, and marker-assisted selection of the drumstick for future improvement will all benefit significantly from this study.

6.2 Recommendation

This research may help identify relationships between gene allelic forms and phenotypes, enabling the alleles to be linked to desired features like rapid growth and high seed output. This is because SNPs are highly frequent and involved in allele variations.

More studies could be carried out using different molecular markers to ascertain the genetic divergence of the drumstick tree in Kenya.

Continued research into functional genomics of *M. oleifera* will deepen our understanding of its biological mechanisms and facilitate development of innovative biotechnological applications.

REFERENCES

Abed, A., Légaré, G., Pomerleau, S., St-Cyr, J., Boyle, B., & Belzile, F. J. (2019). Genotyping-by-sequencing on the ion torrent platform in barley. In *Barley* (pp. 233–252). Springer.

Adegbe, A. A., Larayetan, R. A., & Omojuwa, T. J. (2016). Proximate analysis, physicochemical properties and chemical constituents characterization of Moringa oleifera (Moringaceae) seed oil using GC-MS analysis. *American Journal of Chemistry*, *6*(2), 23–28.

Anwar, F., & Bhanger, M. I. (2003). Analytical characterization of Moringa oleifera seed oil grown in temperate regions of Pakistan. *Journal of Agricultural and Food Chemistry*, *51*(22), 6558–6563.

Anwar, F., Latif, S., Ashraf, M., & Gilani, A. H. (2007). Moringa oleifera: A food plant with multiple medicinal uses. *Phytotherapy Research: An International Journal Devoted to Pharmacological and Toxicological Evaluation of Natural Product Derivatives*, *21*(1), 17–25.

Baptista, A. T. A., Coldebella, P. F., Cardines, P. H. F., Gomes, R. G., Vieira, M. F., Bergamasco, R., & Vieira, A. M. S. (2015). Coagulation–flocculation process with ultrafiltered saline extract of Moringa oleifera for the treatment of surface water. *Chemical Engineering Journal*, *276*, 166–173.

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., Vaillancourt, B., Buell, C. R., Kaeppler, S. M., & de Leon, N. (2013). Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics*, *193*(4), 1073–1081.

Bhandari, H. R., Bhanu, A. N., Srivastava, K., Singh, M. N., & Shreya, H. A. (2017). Assessment of genetic diversity in crop plants-an overview. *Adv. Plants Agric. Res*, *7*(3), 00255.

Bhatia, S., Othman, Z., & Ahmad, A. L. (2007). Coagulation–flocculation process for POME treatment using Moringa oleifera seeds extract: Optimization studies. *Chemical Engineering Journal*, *133*(1–3), 205–212.

Bronzini de Caraffa, V., Giannettini, J., Gambotti, C., & Maury, J. (2002). Genetic relationships between cultivated and wild olives of Corsica and Sardinia using RAPD markers. *Euphytica*, *123*(2), 263–271.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177–186. https://doi.org/10.1016/S0378-1119(99)00219-X

Camacho, F. P., Sousa, V. S., Bergamasco, R., & Teixeira, M. R. (2017). The use of Moringa oleifera as a natural coagulant in surface water treatment. *Chemical Engineering Journal*, *313*, 226–237.

Chan, E. Y. (2009). Next-generation sequencing methods: Impact of sequencing accuracy on SNP discovery. In *Single Nucleotide Polymorphisms* (pp. 95–111). Springer.

CHARAN, A. I., SRIVASTAVA, N. K., SINGH, S., CHARAN, A. A., & SHARAN, A. (n.d.). *MOLECULAR CHARECTERIZATION OF Moringa oleifera WITH RAPD MARKERS*.

Chung, Y. S., Choi, S. C., Jun, T.-H., & Kim, C. (2017). Genotyping-by-sequencing: A promising tool for plant genetics research and breeding. *Horticulture, Environment, and Biotechnology*, *58*(5), 425–431.

da Silva, J. P., Serra, T. M., Gossmann, M., Wolf, C. R., Meneghetti, M. R., & Meneghetti, S. M. (2010). Moringa oleifera oil: Studies of characterization and biodiesel production. *Biomass and Bioenergy*, *34*(10), 1527–1530.

De Donato, M., Peters, S. O., Mitchell, S. E., Hussain, T., & Imumorin, I. G. (2013). Genotyping-by-sequencing (GBS): A novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PloS One*, *8*(5), e62137.

Deschamps, S., Llaca, V., & May, G. D. (2012). Genotyping-by-sequencing in plants. *Biology*, *1*(3), 460–483.

Desplanque, B., Boudry, P., Broomberg, K., Saumitou-Laprade, P., Cuguen, J., & Van Dijk, H. (1999). Genetic diversity and gene flow between wild, cultivated and weedy forms of Beta vulgaris L.(Chenopodiaceae), assessed by RFLP and microsatellite markers. *Theoretical and Applied Genetics*, *98*(8), 1194–1201.

Devarai, L. (2005). Moringa Oleifera: A review on nutritive importance. *Asian J Plant Sci*, *4*, 417–421.

Dogra, P. D., Singh, B. P., & Tandon, S. (1975). Vitamin C content in Moringa pod vegetable. *Current Science*. https://agris.fao.org/agris-search/search.do?recordID=US201303093087

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379.

Fahey, J. W. (2005). Moringa oleifera: A review of the medical evidence for its nutritional, therapeutic, and prophylactic properties. Part 1. *Trees for Life Journal*, *1*(5), 1–15.

Faurie, D. G., Boussougou, S.-K., & Mosesane, M. J. (2018). *The Study of the Sewerage Disinfection Properties of Moringa Oleifera Leaf Powder*. 9.

Festus, M., Danson, K., Josphert, K., & Kennedy, M. (2022). Growth and fruiting of selected provenances of Moringa oleifera Lam. In South Eastern region of Kenya. *Magna Scientia Advanced Research and Reviews*, *5*(2), 008–018.

Foidl, N., Makkar, H. P. S., & Becker, K. (2001). The potential of Moringa oleifera for agricultural and industrial uses. *What Development Potential for Moringa Products, 20*.

Fu, Y.-B., Cheng, B., & Peterson, G. W. (2014). Genetic diversity analysis of yellow mustard (Sinapis alba L.) germplasm based on genotyping by sequencing. *Genetic Resources and Crop Evolution*, *61*(3), 579–594.

Fuglie, L. J. (2005). The Moringa Tree: A local solution to malnutrition. *Church World Service in Senegal*, *5*, 75–83.

Furuta, T., Ashikari, M., Jena, K. K., Doi, K., & Reuscher, S. (2017). Adapting genotyping-by-sequencing for rice F2 populations. *G3: Genes, Genomes, Genetics*, *7*(3), 881–893.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PloS One*, *9*(2), e90346.

Gopalakrishnan, L., Doriya, K., & Kumar, D. S. (2016). Moringa oleifera: A review on nutritive importance and its medicinal application. *Food Science and Human Wellness*, *5*(2), 49–56. https://doi.org/10.1016/j.fshw.2016.04.001

Gruber, B., Unmack, P., Berry, O., & Georges, A. (2019). *Introduction to dartR*. CRAN.

Gruber, B., Unmack, P. J., Berry, O. F., & 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.

Hamrick, J. L., Blanton, H. M., & Hamrick, K. J. (1989). Genetic structure of geographically marginal populations of ponderosa pine. *American Journal of Botany*, *76*(11), 1559–1568.

Hassani, S. M. R., Talebi, R., Pourdad, S. S., Naji, A. M., & Fayaz, F. (2020). In-depth genome diversity, population structure and linkage disequilibrium analysis of worldwide diverse safflower (Carthamus tinctorius L.) accessions using NGS data generated by DArTseq technology. *Molecular Biology Reports*, *47*(3), 2123–2135.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, *5*, 484.

He, S.-T., Yang, J., Wei, J., Wu, J.-C., Zheng, Y.-X., Zhang, Y.-P., & Peng, X.-M. (2021). Genetic diversity of eleven Moringa oleifera Lam. Germplasm introduced to Yunnan, southwest China and their backgrounds during worldwide cultivation. *Genetic Resources and Crop Evolution*, 1–12.

Heffner, E. L., Sorrells, M. E., & Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1–12.

Khan, M. A., Han, Y., Zhao, Y. F., & Korban, S. S. (2012). A high-throughput apple SNP genotyping platform using the GoldenGate™ assay. *Gene*, *494*(2), 196–201.

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., & Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science*, *242*, 14–22.

Kim, K. D., Kang, Y., & Kim, C. (2020). Application of Genomic Big Data in Plant Breeding: Past, Present, and Future. *Plants*, *9*(11), 1454.

Kleden, M. M., Soetanto, H., Kusmartono, K., & Kuswanto, K. (2017). Genetic diversity evaluation of Moringa oleifera, lam from east flores regency using marker random amplified polymorphic DNA (RAPD) and its relationship to chemical composition and in vitro gas production. *AGRIVITA, Journal of Agricultural Science*, *39*(2), 219–231.

kumar Ganesan, S., Singh, R., Choudhury, D. R., Bharadwaj, J., Gupta, V., & Singode, A. (2014). Genetic diversity and population structure study of drumstick (Moringa oleifera Lam.) using morphological and SSR markers. *Industrial Crops and Products*, *60*, 316–325.

Kumar, P., Dolkar, R., Manjunatha, G., & Pallavi, H. M. (2017). Molecular fingerprinting and assessment of genetic variations among advanced breeding lines of Moringa oleifera L. by using seed protein, RAPD and Cytochrome P450 based markers. *South African Journal of Botany*, *111*, 60–67.

Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., Migicovsky, Z., Myles, S., & Toldam-Andersen, T. B. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PLoS One*, *13*(8), e0201889.

Leone, A., Spada, A., Battezzati, A., Schiraldi, A., Aristil, J., & Bertoli, S. (2015). Cultivation, Genetic, Ethnopharmacology, Phytochemistry and Pharmacology of Moringa oleifera Leaves: An Overview. *International Journal of Molecular Sciences*, *16*(6), Article 6. https://doi.org/10.3390/ijms160612791

Li, Z., & Nelson, R. L. (2002). RAPD marker diversity among cultivated and wild soybean accessions from four Chinese provinces. *Crop Science*, *42*(5), 1737–1744.

Mahmood, K. T., Mugal, T., & Haq, I. U. (2010). Moringa oleifera: A natural gift-A review. *Journal of Pharmaceutical Sciences and Research*, *2*(11), 775.

Makkar, H. P. S., & Becker, K. (1997). Nutrients and antiquality factors in different morphological parts of the Moringa oleifera tree. *The Journal of Agricultural Science*, *128*(3), 311–322.

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP Markers and Their Impact on Plant Breeding. *International Journal of Plant Genomics*, *2012*, e728398. https://doi.org/10.1155/2012/728398

Mendieta-Araica, B., Spörndly, E., Reyes-Sánchez, N., & Spörndly, R. (2011). Feeding Moringa oleifera fresh or ensiled to dairy cows—Effects on milk yield and milk flavor. *Tropical Animal Health and Production*, *43*(5), 1039–1047.

Mgendi, M. G., Manoko, M. L., & Nyomora, A. (2010). *Genetic diversity between cultivated and non-cultivated Moringa oleifera Lam. Provenances assessed by RAPD markers*.

Morgil, H., Gercek, Y. C., & Tulum, I. (2020). Single nucleotide polymorphisms (SNPs) in plant genetics and breeding. In *The Recent Topics in Genetic Polymorphisms*. IntechOpen.

Muktar, M. S., Teshome, A., Hanson, J., Negawo, A. T., Habte, E., Entfellner, J.-B. D., Lee, K.-W., & Jones, C. S. (2019). Genotyping by sequencing provides new insights into the diversity of Napier grass (Cenchrus purpureus) and reveals variation in genome-wide LD patterns between collections. *Scientific Reports*, *9*(1), 1–15.

Muluvi, G. M., Sprent, J. I., Odee, D., & Powell, W. (2004). Estimates of outcrossing rates in Moringa oleifera using Amplified fragment length polymorphism (AFLP). *African Journal of Biotechnology*, *3*(2), Article 2. https://doi.org/10.4314/ajb.v3i2.14932

Muluvi, G. M., Sprent, J. I., Soranzo, N., Provan, J., Odee, D., Folkard, G., McNicol, J. W., & Powell, W. (1999). Amplified fragment length polymorphism (AFLP) analysis of genetic variation in Moringa oleifera Lam. *Molecular Ecology*, *8*(3), 463–470.

Muslihatin, W., Saputro, T. B., Latifah, N. I., & Himayani, C. E. S. (2022). The Genetic Diversity of Moringa Oleifera on Poteran Island-Madura Based on Petiole Colors Using ISSR (Inter Simple Sequence Repeat) Method. *7th International Conference on Biological Science (ICBS 2021)*, 39–46.

Nouman, W., Basra, S. M. A., Siddiqui, M. T., Yasmeen, A., Gull, T., & Alcayde, M. A. C. (2014). Potential of Moringa oleifera L. as livestock fodder crop: A review. *Turkish Journal of Agriculture and Forestry*, *38*(1), 1–14.

Ojuederie, O. B., Igwe, D. O., Okuofu, S. I., & Faloye, B. (2013). Assessment of genetic diversity in some Moringa oleifera Lam. Landraces from Western Nigeria using RAPD markers. *The African Journal of Plant Science Biotechnology*, *7*(1), 15–20.

Peterson, G. W., Dong, Y., Horbach, C., & Fu, Y.-B. (2014). Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity*, *6*(4), 665–680.

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, *5*(3).

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., & Sorrells, M. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, *5*(3).

Popoola, J. O., Bello, O. A., OLUGBUYIRO, J., & Obembe, O. O. (2017). Simple sequence repeats (SSR) analysis of genetic intraspecific relationships of Moringa oleifera populations from Nigeria. *Sci. Int.(Lahore)*, *29*(3), 645–657.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, *5*(2), 94–100.

Rajalakshmi, R., Rajalakshmi, S., & Parida, A. (2019). Genetic diversity, population structure and correlation study in Moringa oleifera Lam. Using ISSR and SRAP markers. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences*, *89*(4), 1361–1371.

Ramachandran, C., Peter, K. V., & Gopalakrishnan, P. K. (1980). Drumstick (Moringa oleifera): A multipurpose Indian vegetable. *Economic Botany*, 276–283.

Ravi, R. D., Siril, E. A., & Nair, B. R. (2020). The efficiency of Cytochrome P450 gene-based markers in accessing genetic variability of drumstick (Moringa oleifera Lam.) accessions. *Molecular Biology Reports*, *47*(4), 2929–2939.

Rayan, W. A. (2019). Evaluating the efficiency of DNA Barcode rbcL for detection of genetic relationships between four Moringa spp. Genotypes. *THE EGYPTIAN JOURNAL OF EXPERIMENTAL BIOLOGY (Botany)*, *15*(2), 333–340.

Robiansyah, I., Ramadan, A., Al-kordy, M. A., Ghushash, A. S., & Hajar, A. S. (2015). Characterization of a new biotype Moringa of Saudi Arabia using RAPD and ISSR markers. *Buletin Kebun Raya*, *18*(2), 99–110.

Rufai, S., Hanafi, M. M., Rafii, M. Y., Ahmad, S., Arolu, I. W., & Ferdous, J. (2013). Genetic dissection of new genotypes of drumstick tree (Moringa oleifera Lam.) using random amplified polymorphic DNA marker. *BioMed Research International*, *2013*.

Saini, R. K., Saad, K. R., Ravishankar, G. A., Giridhar, P., & Shetty, N. P. (2013). Genetic diversity of commercially grown Moringa oleifera Lam. Cultivars from India by RAPD, ISSR and cytochrome P 450-based markers. *Plant Systematics and Evolution*, *299*(7), 1205–1213.

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161.

Shahzad, U., Khan, M. A., Jaskani, M. J., Khan, I. A., & Korban, S. S. (2013). Genetic diversity and population structure of Moringa oleifera. *Conservation Genetics*, *14*(6), 1161–1172. https://doi.org/10.1007/s10592-013-0503-x

Shastry, B. S. (2009). SNPs: Impact on gene function and phenotype. *Single Nucleotide Polymorphisms*, 3–22.

Shen, G.-Q., Abdullah, K. G., & Wang, Q. K. (2009). The TaqMan method for SNP genotyping. In *Single nucleotide polymorphisms* (pp. 293–306). Springer.

Smarandache, F., & Christianto, V. (2019). On the Efficacy of Moringa Oleifera as Anticancer Treatment: A Literature Survey. *BAOJ Cancer Research & Therapy*, *5*(2). https://digitalrepository.unm.edu/math_fsp/324

Sohail, Q., Shehzad, T., Kilian, A., Eltayeb, A. E., Tanaka, H., & Tsujimoto, H. (2012). Development of diversity array technology (DArT) markers for assessment of population structure and diversity in Aegilops tauschii. *Breeding Science*, *62*(1), 38–45.

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., Normandeau, É., Laroche, J., Larose, S., & Jean, M. (2013). An improved genotyping by sequencing

(GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PloS One*, *8*(1), e54603.

Sousa, T. V., Caixeta, E. T., Alkimim, E. R., de Oliveira, A. C. B., Pereira, A. A., Sakiyama, N. S., de Resende Júnior, M. F. R., & Zambolim, L. (2017). Population structure and genetic diversity of coffee progenies derived from Catuaí and Híbrido de Timor revealed by genome-wide SNP marker. *Tree Genetics & Genomes*, *13*(6), 1–16.

Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., Lorieux, M., Ahmadi, N., & McCouch, S. (2013). Bridging the genotyping gap: Using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theoretical and Applied Genetics*, *126*(11), 2699–2716.

Spinoso-Castillo, J. L., Escamilla-Prado, E., Aguilar-Rincón, V. H., Ramos, V. M., de los Santos, G. G., Pérez-Rodríguez, P., & Corona-Torres, T. (2020). Genetic diversity of coffee (Coffea spp.) in Mexico evaluated by using DArTseq and SNP markers. *Genetic Resources and Crop Evolution*, *67*(7), 1795–1806.

Stephenson, K. K., & Fahey, J. W. (2004). Development of tissue culture methods for the rescue and propagation of endangered Moringa spp. Germplasm. *Economic Botany*, *58*(1), S116–S124.

Su, C., Wang, W., Gong, S., Zuo, J., Li, S., & Xu, S. (2017). High density linkage map construction and mapping of yield trait QTLs in maize (Zea mays) using the genotyping-by-sequencing (GBS) technology. *Frontiers in Plant Science*, *8*, 706.

Suzuki, R., & Shimodaira, H. (2015). pvclust: Hierarchical clustering with P-values via multiscale bootstrap resampling. *R Package Version*, *2*.

Vaiman, D., Mercier, D., Moazami-Goudarzi, K., Eggen, A., Ciampolini, R., Lépingle, A., Velmala, R., Kaukinen, J., Varvio, S.-L., & Martin, P. (1994). A set of 99 cattle microsatellites: Characterization, synteny mapping, and polymorphism. *Mammalian Genome*, *5*(5), 288–297.

Verma, S., Gupta, S., Bandhiwal, N., Kumar, T., Bharadwaj, C., & Bhatia, S. (2015). High-density linkage map construction and mapping of seed trait QTLs in chickpea (Cicer arietinum L.) using Genotyping-by-Sequencing (GBS). *Scientific Reports*, *5*(1), 1–14.

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., & Lu, Y. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Scientific RepoRtS*, *10*.

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, *18*(1), 1–12.

Wu, J.-C., Yang, J., Gu, Z.-J., & Zhang, Y.-P. (2010). Isolation and Characterization of Twenty Polymorphic Microsatellite Loci for Moringa oleifera (Moringaceae). *HortScience*, *45*(4), 690–692. https://doi.org/10.21273/HORTSCI.45.4.690

Xie, W.-G., Zhang, X.-Q., Cai, H.-W., Liu, W., & Peng, Y. (2010). Genetic diversity analysis and transferability of cereal EST-SSR markers to orchardgrass (Dactylis glomerata L.). *Biochemical Systematics and Ecology*, *38*(4), 740–749.

Zaku, S. G., Emmanuel, S., Tukur, A. A., & Kabir, A. (n.d.). *African Journal of Food Science Review*.

APPENDIX

Appendix I. Publication

Magna Scientia Advanced Research and Reviews

eISSN: 2582-9394

Cross Ref DOI: 10.30574/msarr

Journal homepage: https://magnascientiapub.com/journals/msarr/



(RESEARCH ARTICLE)

Check for updates

Genetic diversity of selected *Moringa oleifera* lam. Provenances from the coastal region of Kenya

Sarah Kwamboka Ondieki [1,*], Carol W. Hunja [2], Geoffrey M. Muluvi [2], Jackson C. Korir [1], Festus Mutiso [3], Johnson U. Kitheka [3], Danson Kioko [3], Josphert Kimatu [2], James K. Ndufa [4] and Kennedy Mutati [5]

[1] *School of Natural and Applied Sciences, Department of Biological Sciences, Masinde Muliro University of Science and Technology, P.O Box 190-50100 Kakamega, Kenya.*
[2] *School of Science and Computing, South Eastern Kenya University, P. O Box 170-90200 Kitui, Kenya.*
[3] *School of Agriculture, Environment, Water and Natural Resources, South Eastern Kenya University, P.O Box 170-90200 Kitui, Kenya.*
[4] *Kenya Forestry Research Institute, P.O. Box 20412-00200 Nairobi, Kenya.*
[5] *Department of Water & Agriculture, County Government of Kitui, P.O. Box 33-90200 Kitui, Kenya.*

Magna Scientia Advanced Research and Reviews, 2023, 08(02), 118–128

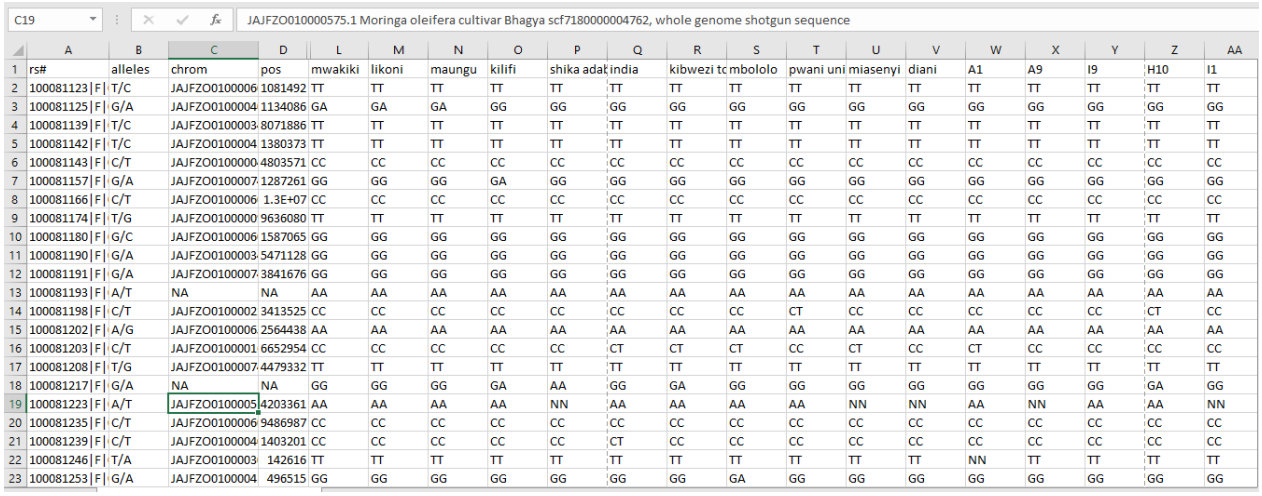
Publication history: Received on 13 June 2023; revised on 22 July 2023; accepted on 25 July 2023

Article DOI: https://doi.org/10.30574/msarr.2023.8.2.0113

Abstract

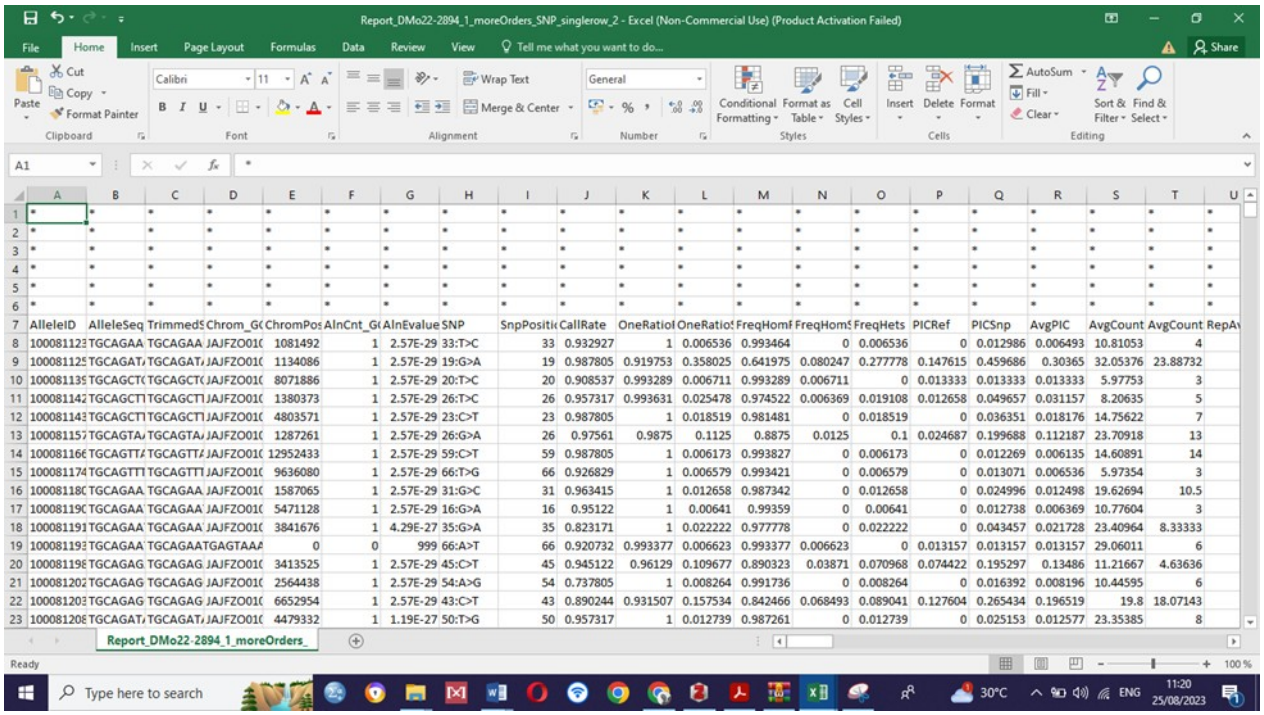
Moringa oleifera Lam (syn. *M. Pterygosperma Gaertn*) is one of the most economically important crops widely cultivated globally. Traditionally, it has been used as a remedy for conditions such as diabetes, inflammation, cancer, bacterial,

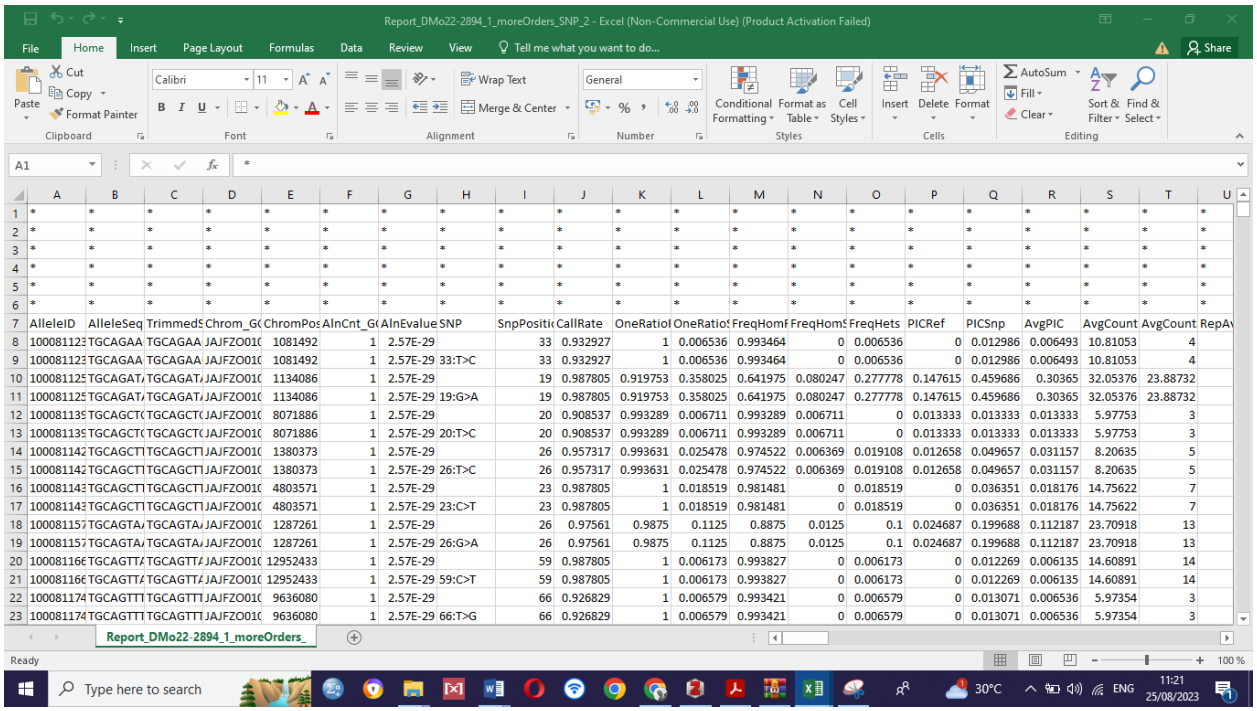
Appendix II.Single Nucleotide Polymorphisms of Selected Moringa.

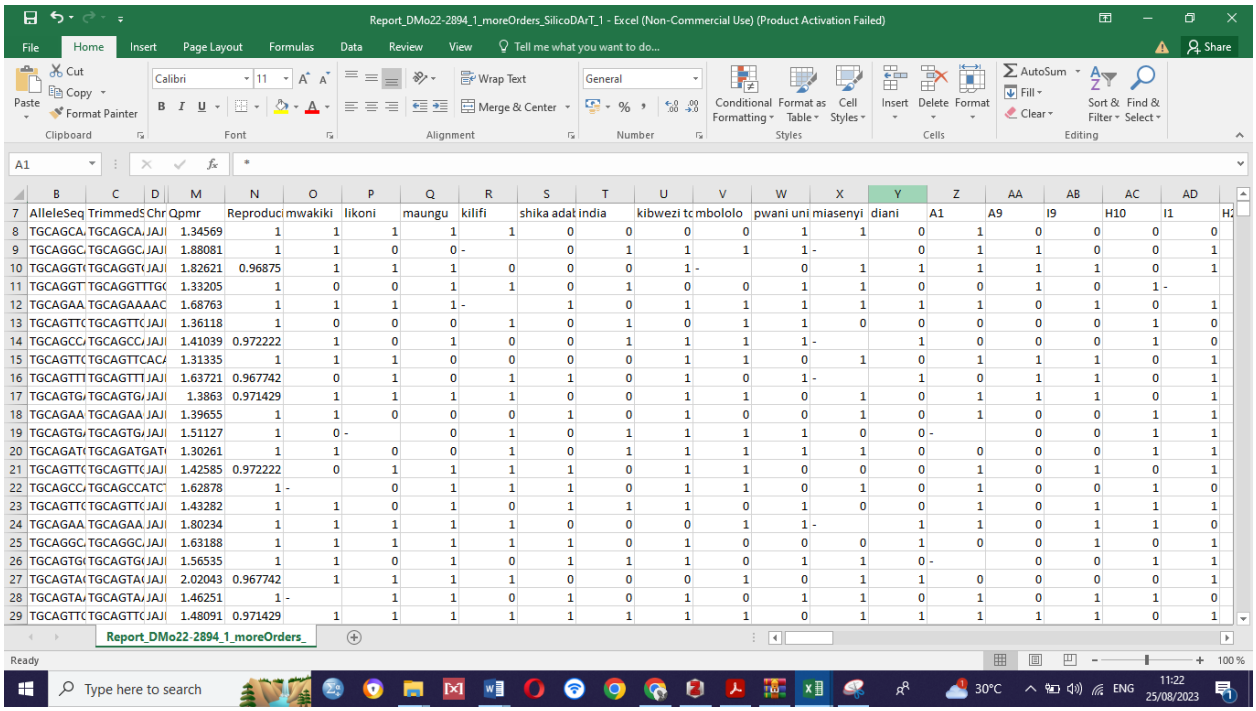


C19 | JAJFZO010000575.1 Moringa oleifera cultivar Bhagya scf7180000004762, whole genome shotgun sequence

	A	B	C	D	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	rs#	alleles	chrom	pos	mwakiki	likoni	maungu	kilifi	shika adak	india	kibwezi tc	mbololo	pwani uni	miasenyi	diani	A1	A9	I9	H10	I1
2	100081123 | F |	T/C	JAJFZO0100006	1081492	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
3	100081125 | F |	G/A	JAJFZO0100004	1134086	GA	GA	GA	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
4	100081139 | F |	T/C	JAJFZO0100003	8071886	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
5	100081142 | F |	T/C	JAJFZO0100004	1380373	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
6	100081143 | F |	C/T	JAJFZO0100000	4803571	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
7	100081157 | F |	G/A	JAJFZO0100007	1287261	GG	GG	GG	GA	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
8	100081166 | F |	C/T	JAJFZO0100006	1.3E+07	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
9	100081174 | F |	T/G	JAJFZO0100000	9636080	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
10	100081180 | F |	G/C	JAJFZO0100006	1587065	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
11	100081190 | F |	G/A	JAJFZO0100003	5471128	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
12	100081191 | F |	G/A	JAJFZO0100007	3841676	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
13	100081193 | F |	A/T	NA	NA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
14	100081198 | F |	C/T	JAJFZO0100002	3413525	CC	CC	CC	CC	CC	CC	CC	CC	CT	CC	CC	CC	CC	CC	CT	CC
15	100081202 | F |	A/G	JAJFZO0100006	2564438	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
16	100081203 | F |	C/T	JAJFZO0100001	6652954	CC	CC	CC	CC	CC	CT	CT	CT	CC	CT	CC	CT	CC	CC	CC	CC
17	100081208 | F |	T/G	JAJFZO0100007	4479332	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
18	100081217 | F |	G/A	NA	NA	GG	GG	GG	GA	AA	GG	GA	GG	GG	GG	GG	GG	GG	GG	GA	GG
19	100081223 | F |	A/T	JAJFZO0100005	4203361	AA	AA	AA	AA	NN	AA	AA	AA	AA	NN	NN	AA	NN	AA	AA	NN
20	100081235 | F |	C/T	JAJFZO0100006	9486987	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
21	100081239 | F |	C/T	JAJFZO0100004	1403201	CC	CC	CC	CC	CC	CT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
22	100081246 | F |	T/A	JAJFZO0100003	142616	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	NN	TT	TT	TT	TT
23	100081253 | F |	G/A	JAJFZO0100004	496515	GG	GG	GG	GG	GG	GG	GG	GA	GG	GG	GG	GG	GG	GG	GG	GG


Report_DMo22-2894_1_moreOrders_SNP_singlerow_2 - Excel (Non-Commercial Use) (Product Activation Failed)
Report_DMo22-2894_1_moreOrders_
Ready
Type here to search
30°C
ENG
11:20
25/08/2023


Report_DMo22-2894_1_moreOrders_SNP_2 - Excel (Non-Commercial Use) (Product Activation Failed)
Report_DMo22-2894_1_moreOrders_
Ready
Type here to search
30°C
ENG
11:21
25/08/2023


Report_DMo22-2894_1_moreOrders_SilicoDArT_1 - Excel (Non-Commercial Use) (Product Activation Failed)
File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...
Share
Cut Copy Format Painter Paste
Calibri 11
Wrap Text
Merge & Center
General
Conditional Formatting
Format as Table
Cell Styles
Insert Delete Format
AutoSum Fill Clear
Sort & Filter
Find & Select
Clipboard Font Alignment Number Styles Cells Editing
A1
AlleleSeq TrimmedS Chr Qpmr Reproduci mwakiki likoni maungu kilifi shika adat india kibwezi tc mbololo pwani uni miasenyi diani A1 A9 I9 H10 I1
Report_DMo22-2894_1_moreOrders_
Ready
100 %
Type here to search
30°C
ENG
11:22
25/08/2023

Appendix III. Sequencing and BLAST



SEQART AFRICA
Affordably High Quality Genotyping

SEQART AFRICA
ILRI Campus, Old Naivasha Road P.O Box 30709 Nairobi, K
www.seqart.net

Metadata info for order(s): DMo21-2867 DMo22-2894

Order number: DMo21-2867
Client name [email]: Dr. Carol Hunja[chunja@seku.ac.ke]
Order date and time: 2021-07-05 12:24:09
Product ordered: Moringa DArTseq (1.0)
Sequenced on: HiSeq at: 2021-08-06 13:01:25

Order number: DMo22-2894
Client name [email]: Dr. Carol Hunja[chunja@seku.ac.ke]
Order date and time: 2022-03-01 13:02:59
Product ordered: Moringa DArTseq (1.0)
Sequenced on: HiSeq at: 2022-03-31 15:44:04

Blast alignment parameters:

E-value: 5e-10
Min sequnence % identity: 90

Aligned to the following blast databases:
Model reference: GCA_021397835.1

Data files in the report
NB:There was a contamination of "Foreign" DNA in the tissues.
File name: Report_DMo22-2894_1_moreOrders_SilicoDArT_1.csv
File type: SilicoDArT
File description: SilicoDArT Format: SilicoDArTs are scored in a binary fashion, representing genetically "

Metadata columns in the file:

AlleleSequence	The sequence of the tag which is present in samples with genotype score "1"
AvgReadDepth	The sum of the tag read counts for all samples, divided by the number of sam
CallRate	The proportion of samples for which the genotype call is either "1" or "0", rat
CloneID	Unique identifier of the sequence tag

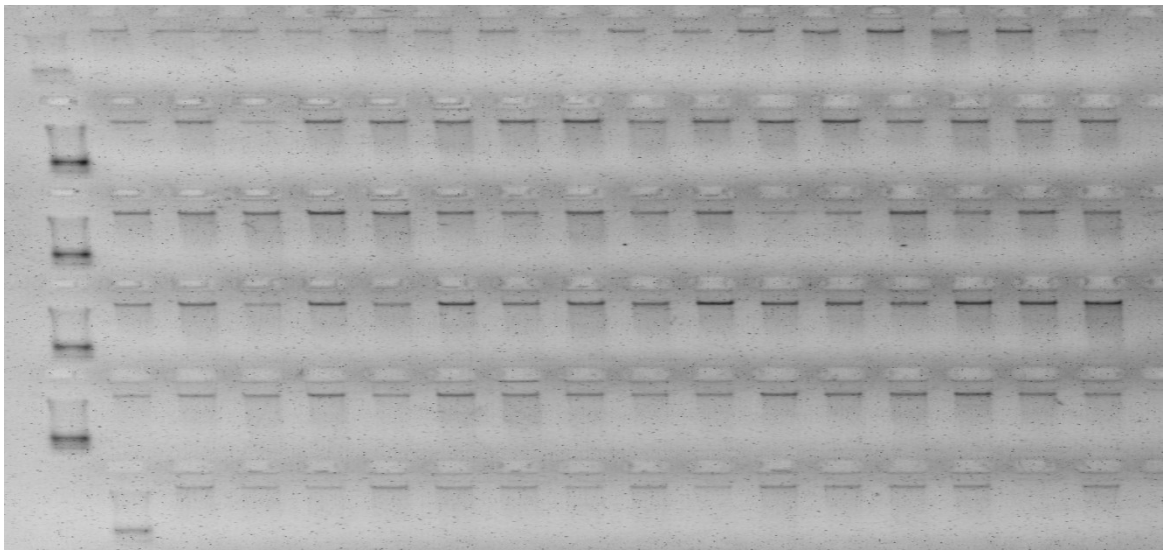
AlnCnt_	Total count of aligning markers / tags with selection criteria described below
AlnEvalue_	E value of the best alignment to an existing model genome
ChromPos_	Position(s) on contig(s) with the best alignment of marker / tag to an existing
Chrom_	Contig(s) with the best alignment of marker / tag to an existing model genom

Header rows:

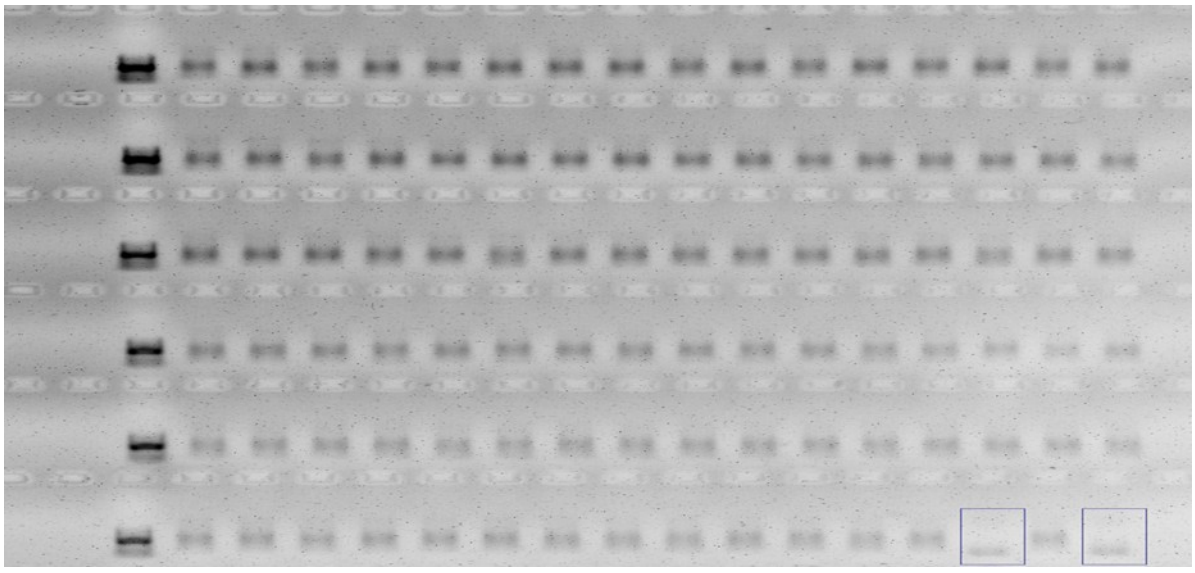
1	Order number where sample belongs to - important for multi-orders reports
2	DArT plate barcode
3	client plate barcode
4	well row position
5	well column position
6	sample comments
7	genotype name

List of not reported samples [OrderNumber PlateID Row Col SampleName]

DMo21-2867 1 D 2 blank
DMo21-2867 1 E 2 blank
DMo21-2867 1 F 2 blank
DMo21-2867 1 G 2 blank
DMo21-2867 1 H 2 blank
DMo21-2867 1 A 3 blank
DMo21-2867 1 B 3 blank
DMo21-2867 1 C 3 blank
DMo21-2867 1 D 3 blank
DMo21-2867 1 E 3 blank
DMo21-2867 1 F 3 blank
DMo21-2867 1 G 3 blank
DMo21-2867 1 H 3 blank



a. Image of 0.8% agarose gel electrophoresis



b. Image of 0.8% agarose gel electrophoresis