Optimization and characterization of novel EST-SSR markers for cassava (*Manihot esculenta*).

Titus Mureithi Kathurima

A thesis submitted in partial fulfillment for the Degree of Master of Science in Biochemistry in the Jomo Kenyatta University of Agriculture and Technology

2009

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature…… Date…….

Titus Mureithi Kathurima

This thesis has been submitted for examination with our approval as university supervisors.

Signature…… Date…….

Dr. Esther Magiri.

JKUAT, Kenya.

Signature…… Date…….....

Dr. Morag Ferguson

International Institute of Tropical Agriculture, Kenya.

DEDICATION

This work I dedicated to my wife Purity Kagwiria and sons Elvis Munene and Evans Mutugi for their relentless support and understanding through out the trying moments.

ACKNOWLEDGEMENT

First and foremost may I take this opportunity to thank the almighty God for having been merciful and kind throughout the period of this project. May I take the earliest opportunity to thank my supervisors Dr Esther Magiri and Dr Morag Ferguson. Dr Magiri the university supervisor was there from the first step of developing of the proposal up to the end of thesis finalization. Thank you for your valuable corrections and co operation and personal commitment you showed throughout the learning period. I will always cherish each and every advice given during this project. Secondly my I extend my whole heartedly gratitude to my other supervisor Dr Morag Ferguson (IITA) who allowed me into her project and supported me with technical support and her valuable time. You ably gave your guidance to my project and all related aspects timely and promptly. There other attribute that I have learnt besides the project work which I consider to have added value to me as an aspiring scientist. Thank you for your time and guidance. I would like also to thank each and every colleague that I worked with from lab four. Inoster Nzuki for your technical training of carrying out PCR, gel analysis, genotyping with ABI machine and analyzing the data. Finally but not least the ILRI members whom I interacted with during this period, God bless you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ABI	Applied Biosystems
AFLP	Amplified Fragment Length Polymorphism
AIDS	Acquired immunodeficiency syndrome
BLAST	Basic Local Alignment Search Tool
CIAT	Centro Internacional de Agricultural Tropical
cDNA	Complementary deoxyribonucleic acid
DNA	Deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
EST	Expressed sequence tag
F	Forward
HIV	Human immunodeficiency virus
IITA	International Institute of Tropical Agriculture
MAS	Mass assisted selection
mM	Milimolar
NCBI	National Center for Biotechnology Information
PCR	Polymerase chain reaction
PIC	Polymorphic Information Content
QTL	Quantitative trait loci
R	Reverse
RAPD	Random Amplified Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism

RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SSR	Simple Sequence Repeats
Tm	Melting temperature
Ta	Annealing temperature
UPGMA	Unpaired Group Method of Arithmetic averages
USA	United States of America

ABSTRACT

Microsatellites, or simple sequence repeats (SSRs) are very useful molecular markers for a number of plant species. They are commonly used in cultivar identification, plant variety protection, as anchor markers in genetic mapping, and in marker-assisted breeding. Their utility is due to their abundance, hyper variability, and suitability for high-throughput analysis. Early development of SSRs was hampered by the high cost of library screening and clone sequencing. Currently, large public SSR datasets exist for many crop species, but the numbers of publicly available, mapped SSRs for cassava are relatively low. A database mining approach to identify SSR-containing expressed sequence tags (EST) in the IITA/Craig Venter Institute database was utilized. The overall aim of this study was to optimize and characterize in terms of polymorphism new EST-SSR primers that may be useful for diversity assessments and genetic linkage mapping in cassava. Seventy primer pairs were synthesized and used to amplify SSRs from diverse cassava DNA genotypes. This study identified 33 (63%) useful SSRs markers which were polymorphic in a set of 24 cassava genotypes from South America and Africa together with four parents of a mapping population from drought tolerant and four parents of a cassava brown streak disease (CBSD) mapping population. The polymorphic information content (PIC) values ranges were from 0.0 to 0.7381 and average allele frequency of 2.5. The high proportion of (63%) polymorphic EST-SSRs obtained in this work validates the use of transcribed sequences as a source of markers. These markers will be useful for genetic mapping, diversity assessments and genomic research.

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background

1.1.1 Origin and distribution of cassava

Cassava (*Manihot esculenta* Crantz,) is a perennial root crop in the family *Euphorbiaceae* (Purseglove, 1968). The cassava plant is believed to have arisen through hybridization events of wild *Euphorbiaceae* species (Nassar, 2002). Cassava was domesticated sometime between 5000 – 7000 years BC in the Amazon region (Gibbons, 1990). The crop is widely spread in the tropics (Jos, 1969) although it is originally considered to be a

native to countries of the New World, particularly Brazil and Mexico (Nassar, 2002). Since then, the crop has spread rapidly in most tropical countries in South America, Africa and Asia (Gibbons, 1990). The Portuguese introduced cassava into Africa in the 16th century from Brazil (Jones, 1959). Introduction of the crop into Africa commenced first in the West African coast and later to East Africa through Madagascar and Zanzibar (Jennings, 1976). Cassava was reported to be grown in Reunion and Zanzibar between 1736 and 1799 (Purseglove, 1968). Cassava was grown on the East African interior initially starting along the coastal region (Purseglove, 1968). Later, the crop spread further inland and more widely in the region around the late part of the 18th or 19th century (Purseglove, 1968). Cassava may have been introduced into the Lake Victoria basin along trade routes (Jameson and Thomas, 1970). The crop became established in Uganda during the 19th century and its value as a food security crop was soon realized.

In fact, between 1963 and 1964 about 175,000 ha were being grown in Uganda (Jameson and Thomas, 1970). Since many cultivars are drought resistant, cassava can survive even during the dry season when the soil moisture is low (Puttchacharoen, *et al*.,1998). The taxonomical classification is as follows (Allen, 2002).

Kingdom:	Plantae
Division:	Magnoliophyta
Class:	Magnoliopsida
Order:	Malpighiales
Family:	Euphorbiaceae
Subfamily:	Crotonoideae
Tribe:	Manihoteae
Genus:	Manihot
Species:	*Esculenta*

1.1.2 Ecological Conditions

Cassava is usually cultivated in areas considered marginal for other crops with soils of low fertility and annual rainfall of less than 600 mm as in the semi arid tropics (De Tafur *et al.,* 1997) to more than 1000 mm in the sub-humid and humid tropics (Pellet *et al.*, 1997). The uncertain rainfall, infertile soils and weak market infrastructure in Africa makes cultivation of cassava advantageous over rice, maize and other grains as a food stable (FAO, 2002). Whereas cassava can be grown under a wide range of ecological conditions, other crops has narrow ecological adaptation in Africa. Cassava is drought tolerant and this makes it a suitable food crop during periods of drought and or famine (Federal Agriculture Coordinating unit, Nigeria 1986). Besides, cassava has substantial leaf production whose leaves tend to drop to form organic matter in the soil thus

recycling soil nutrients (Howeler, 1998). Cassava requires little or no fertilizer and yet will maintain a steady production trend over a fairly long period of time in a continuous farming system although fertilizer application can increase yield (Asher, *et al*., 1980) Howeler, 1998). With its ability to suppress weeds particularly the improved varieties which develop many branches early enough to form a canopy shading weeds from solar radiation, cassava as a crop is a friend of the small scale farmer whose ability to weed is drastically reduced. Weeding is however vitally important in the growing period, before a canopy is established, if high yields are to be realized (Islam, *et al*., 1980). Unlike other crops, such as yam, maize, banana and plantain, cowpea or sorghum and millet which are eco-regionally specific, cassava cuts across all ecological zones.

1.1.3 Importance of cassava

Cassava is a very popular crop in most areas because of the many socio-economic uses, where it can be grown (FAO, 2002). More than 600 million people depend on cassava in Africa, Asia, and Latin America (FAO, 2002, Scott e*t al*., 2000). Cassava has gained such importance because of the following reasons. First, it has starchy roots, which are a very valuable source of energy (Purseglove, 1968). The roots can be boiled and processed in different ways for human consumption. For example, the roots have been processed into

fermented starches, dried as chips, or used as meal or pellets for animal feed (Hernan *et al*., 2004). Cassava provides a cheap source of dietary carbohydrate energy of 720.1 x 10- 12 kJ day–1, which makes it fourth in rank after rice, sugarcane and maize (Baguma, 2004). The root of cassava also contains other important nutrients. For example, cassava roots are known to comprise 2-3 % (dry weight basis) proteins. Reports elsewhere

suggest that protein content in the roots could actually be considerably higher (6 - 8 %) in some landraces, particularly from Central America (Hernan *et al.,* 2004). Nonetheless, this protein amount is little compared to what has been recorded in other crops. It is also possible to enhance protein content in roots using crop improvement techniques (Hernan *et al*., 2004). Similarly, considerable amounts of carotenes have been observed in yellow cassava roots (Iglesias *et al*., 1997). Leaves from cassava have been found to be very nutritious which are often used as both animal and human food (Hernan *et al*., 2004; Jaramillo *et al*., 2006). Cassava also has a huge industrial potential. For example, in South East Asia, cassava is being exploited for ethanol for automobile fuel (Baguma, 2004). Hence, the Food and Agriculture Organization (FAO) of United Nations (UN) has identified cassava as a crop that will spur rural industrial development and raise incomes for producers, processors, and traders and contribute to food security of its producing and consuming households. Similarly, in Africa, the New Partnership for Africa's Development (NEPAD) has identified cassava as 'A Poverty Fighter in Africa' and launched a Pan African Cassava Initiative that seeks to tap the enormous potential of the crop for food security and income generation (NEPAD, 2004).

1.1.4 Molecular markers and genomic mapping of cassava

A molecular marker can be defined as a specific piece of DNA with a known position on the genome (Launaud and Vincent 1997). Simple sequence repeats (SSR) molecular markers are useful for a variety of applications in plant genetics and breeding because of their reproducibility, multi-allelic nature, co-dominant inheritance, relative abundance and good genome coverage (Powell *et al.*, 1996). Markers have been useful for genetic and sequence-based physical maps in plant species, and simultaneously have provided

breeders and geneticists with an efficient tool to link phenotypic and genotypic variation (Mba *et al*., 2000). For instance there are an estimated 800 SSR markers available for cassava that covers 80% of the cassava genome (Mba *et al*., 2000). Additional markers are required to increase the density of markers on the genetic linkage map of cassava. This will enable scientist to map traits of interest in cassava more accurately. The current cassava genetic map was constructed from segregation of RFLP, SSR, RAPD and isoenzyme markers (Fregene *et al* 1997). Many of the markers on the map are RFLP markers which do not lend themselves easily to large scale high throughput marker assisted analysis of plant population (Mba *et al*., 2000).

There is an initiative to improve the cassava map through identification of SSR derived from expressed sequence tags (EST-SSRs) and single nucleotide polymorphisms. The distribution of EST-SSRs in the genetic maps mirrors the distribution of genes along the genetic map as ESTs are derived from expressed genes (Boventius and Weller, 1994; Suarez *et al.*, 2000).

1.1.5 Expressed sequence tags-simple sequence repeats (EST-SSR)

The expressed sequence tags (ESTs) are obtained from transcribed region of a gene. Each gene must be converted or transcribed into messenger RNA (mRNA) that serves as a template for protein synthesis (MacIntosh *et al.,* 2001). The mRNA guides the synthesis of a protein through a process called translation. The problem, however, is that mRNA is very unstable outside of a cell; therefore, scientists use an enzyme called reverse transcriptase to convert mRNA to complementary DNA (cDNA) (Leigh *et al.*, 2003). cDNA production is the reverse of the usual process of transcription in cells because the procedure uses mRNA as a template rather than DNA. cDNA is a much

more stable compound and it represents only expressed DNA sequence because it is generated from mRNA that is transcribed from the exons by excising (splicing) introns (Eujayl *et al.*, 2004). Once cDNA representing an expressed gene has been isolated, scientists can then sequence a few hundred nucleotides from either the 5' or 3' end to create 5' expressed sequence tags (5' ESTs) and 3' ESTs, respectively (Jongeneel, 2000). A 5' EST is obtained from the portion of a transcript (exons) that usually codes for a protein (Boguski, 1993). The expressed regions tend to be conserved across species and do not change much within a gene family (Elias *et al*, 2000).

EST sequenced data and full length cDNA clones can be downloaded from GenBank and scanned for identification of SSRs, using appropriate software (Boguski *et al*., 1993). Such SSRs are typically referred to as EST-SSRs or genic Microsatellites. Subsequently, locus specific primers flanking EST or genic SSRs can be designed to amplify the microsatellite present in the EST (Suarez *et al*, 2002). Thus, the generation of EST- SSR markers in this way is relatively easy and inexpensive because they are a by-product of the sequence data from genes or ESTs that are publicly available (Gupta *et al.*, 2003). However, the generation of EST-SSR markers is largely limited to those species or close relatives for which there is a sufficiently large number of ESTs available. EST-SSRs have some intrinsic advantages over genomic SSRs in that, they are present in expressed regions of the genome (Rungis *et al.*, 2004). The usefulness of these EST-SSRs also lies in their expected transferability because the primers are designed from the more conserved coding regions of the genome as opposed to non-genic regions however, some regions are more conserved and polymorphism tends to be

more limited (Pashley *et al.*, 2006). They are also reported to produce cleaner results for scoring as there are fewer null alleles (Leigh *et al.*, 2003; Rungis *et al.*, 2004) and fewer stutter bands (Leigh *et al.*, 2003;Woodhead *et al*., 2005; Eujayl *et al.*, 2004; Pashley *et al.*, 2006). Due to the advantages of EST-SSR markers over genomic SSR markers and the public availability of large a EST of sequence data, genic SSRs have been identified, developed and used in a variety of studies, for several plant species (Eujayl *et al*, 2002).

1.2 Problem Statement

The yield of local cassava landraces rarely exceeds 7-9 t/ha in Africa (FAO, 2002). Similarly only about 25 t/ha, compared to potential yields of over 80 t/ha, can be obtained when elite varieties are grown (FAO, 2002). Attaining optimal yields in cassava is constrained by a number of social, abiotic and biotic factors. Among these constraints are cassava diseases which are considered to be the most important constraint to production (Otim-Nape, 1993). Disease can cause up to 100 % yield loss (Jennings, 1994). Conventional breeding efforts are employed to bring the epidemics under control. Unfortunately the conventional procedures used, often takes very long, laborious, costly and of low precision.

Cassava germplasm is often characterized by morphological descriptors markers. Morphological descriptors are reliable, easy to study and relatively low cost to evaluate however, they have some limitations as they are influenced by environmental conditions. In traditional farming systems, the concept of a variety can encompass very diverse genetic entities. Traditional naming and classification systems are often based on traits

that are perceived subjectively. In so doing it is not uncommon to find confusion between varieties or use of different names for the same cultivar (Elias *et al*., 2001).

Besides the *de novo* development of SSRs is a costly and time-consuming endeavor (Zane *et al.*, 2002), and these problems are often compounded by a paucity of resources in taxa that lack clear economic importance. One possible solution to these sorts of problems would be to exploit publicly available genomic resources for the development of gene-based SSR markers (Smith and Waterman 1981). In fact, the rapid and inexpensive development of SSRs from expressed sequence tag (EST) databases has been shown to be a feasible option for obtaining high-quality nuclear markers (Gupta *et al.*, 2003; Bhat *et al.*, 2005). Recently, new SSRs have been identified within EST sequences and flanking primers have been designed. The aim of this study was to optimize the amplification of the new EST-SSR primers and study polymorphism in diverse array of cassava germplasm.

1.3 Justification

Africa is a continent in crisis; it is racked with hunger, poverty, and the HIV/AIDS pandemic Canadian Coalition to End Global Poverty (CCIC) (2004). Africa is also the region with the fastest population growth, the most fragile natural resources base and the weakest set of agricultural research and extension institutions World Bank (2000). African people are bearing the brunt of world food insecurity, malnutrition and poverty and its population is expected to increase to 1.2 billion by 2020, and its urban population will likely grow at an even faster rate (FAO, 2002). With urbanization and higher incomes, the composition and characteristics of food demand will be significantly

altered, domestic food production and food imports will have to increase to meet Africa's growing demand World Bank (2000).

With genetic improvement through efficient plant breeding and improvements in agronomy cassava can enhance its role as a powerful poverty fighter (NEPAD, 2004). Determining the genetic relationship and the genetic variability or diversity among varieties is important for efficient conservation of genetic resources (Lefebvre *et al*., 2001). Characterizing and defining the genetic entities or varieties is crucial for plant breeding efforts to develop new high yielding varieties that meet the demands of farmers and consumers (Ayad *et al*., 1997). Molecular marker technologies have found applications in the determination of the genetic basis of phenotypic expression and the manipulation of phenotypic variation in plants. These have been mostly through the use of markers in understanding heterosis; prediction of hybrid performance; identification and mapping of Quantitative trait loci (QTLs) and in marker assisted selection (MAS) (Tanksley and Nelson 1996). SSR markers have been useful for integrating the genetic, physical and sequence-based map in plant species and simultaneously have provided an efficient tool to link phenotypic and genotypic variation (Gupta *et al.*, 2000).

An initiative toward improving the cassava map involving the generation of expressed sequence tags and identification of SSRs contained within them is in progress. EST-SSR markers often have known or 'putative' functions and they may represent functional genes (Anderson and Lubbersted, 2003). The putative functions for a significant number of EST-SSR markers has been reported in barley (Thiel *et al.*, 2003), wheat (Yu *et al.*, 2004), and cotton (Han *et al.*, 2004). EST-SSR markers can contribute to 'direct allele

selection', if they are shown to be completely associated or even responsible for a targeted trait (Sorrells and Wilson, 1997). For example, recently, a Dof homolog (*DAG1)* gene that showed a strong effect on seed germination in *Arabidopsis* (Papi *et al.*,2002) has been mapped on chromosome 1B of wheat by using wheat EST-SSR primers (Gao *et al.*, 2004). Similarly, Yu *et al.* 2004 identified two EST-SSR markers linked to the photoperiod response gene (*ppd*) in wheat. The successful strategies to identity functional genes reported in other plant species can also be applied to cassava by developing EST-SSR markers.

1.4 Hypothesis

Newly identified EST-SSR markers can detect genetic variation among cassava varieties grown in different geographical zones.

1.5 General objective

To optimize and characterize new EST-SSR primers that may be useful for diversity assessments and genetic linkage mapping in cassava.

1.6 Specific objectives

1. To use bioinformatics tools to select newly developed EST-SSR primers.
2. To optimize PCR amplification conditions of new EST-SSR primers.
3. To determine polymorphism of EST-SSR loci in a diverse array of cassava germplasm.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Origin of cassava

The centre of origin and diversity of cassava is in Brazil (Olsen and Schaal, 1999). It was grown in Peru four thousand years ago and in Mexico some two thousand years ago (Okigbo, 1980). Cassava is a vital staple for about 500 million people world wide (FAO, 2002). It is ranked high among the top 10 most significant food crops produced in developing countries (Scott e*t al*., 2000).

2.2 Germplasm characterization

Germplasm characterization refers to the observation, measurement and documentation of heritable plant traits in a collection (Jones *et al*., 1997). The resulting data allows the identification and classification of accessions, building a catalog of descriptors with embedded biological information that is essential to germplasm collection management or to direct use in agriculture (Mohan *et al*., 1997). Germplasm characterization aims at the description and understanding of the genetic diversity of the organism under study. Today, germplasm characterization has been developed based mostly on morphological descriptors and molecular marker technology (Jones *et al*., 1997). Morphological descriptors are easy to study and relatively low cost to evaluate. However, they have some limitations as they are influenced by environmental conditions. Due to these limitations, molecular markers are now rapidly being adopted by crop improvement researchers globally as an effective tool for basic studies addressing biological components in agricultural production system (Jones *et al*., 1997; Mohan *et al*., 1997).

Molecular markers offer specific advantages in assessment of genetic diversity and in trait-specific crop improvement. They are also applied in breeding programs for mapping and tagging of the gene blocks associated with economically important traits (quantitative trait loci) (QTL) (Mohan *et al*., 1997). Morphological traits cannot, however, be replaced by any of the molecular techniques, the results of the molecular or biochemical studies should be considered as complementary to morphological characterization.

2.3 Genetic diversity studies

2.3.1 Variability and Conservation

Genetic diversity is the variation in the genetic composition within or among individuals' populations or species. Assessment of genetic variability is important for conservation and utilization of genetic resources (Lefebvre *et al*., 2001). It is important to conserve genetic diversity for utilization by the users (plant breeders, agronomists, and farmers) (Ayad *et al*., 1997). The success of any genetic conservation program is dependent on understanding the amount and distribution of genetic diversity present in the gene pool (Zhang *et al.,* 2000) and it creates a basis for improvement of agricultural production. Data from genetic diversity studies may also be used in developing conservation strategies within a gene bank. One such application is the development of a core collection. (Zhang *et al.*, 1998). Core collections are usually 10% of the total collection (Lanaud and Vincent, 1997) and attempt to maximize diversity. Genetic diversity is measured either on morphological, biochemical or DNA profiling (Lefebvre *et al*., 2001).

2.4 DNA-based Molecular Markers

The differences that distinguish one plant from another are encoded in the plant genetic material, the deoxyribonucleic acid (DNA). DNA is packaged in chromosome pairs (strands of genetic material), one coming from each parent. The genes, which control a plants characteristic are located on specific segments of each chromosome. All of the genes carried by a single gamete are known as the genome (King and Stansfield, 1997). To help identify specific genes located on a particular chromosome, most scientist use genetic markers. A genetic marker can be defined as a specific piece of DNA with a known position on the genome. Molecular marker technologies have found applications in the determination of the genetic basis of phenotypic expression and the manipulation of phenotypic variation in plants. These have been mostly through the use of markers in understanding heterosis; prediction of hybrid performance; identification and mapping of QTL; and in MAS (Stuber *et al.,* 1999). The following DNA based molecular markers are discussed below.

2.4.1 Restriction Fragment Length Polymorphism (RFLP)

Restriction Fragment Length Polymorphism (RFLP) is defined as the variation(s) in the length of DNA fragments produced by a specific restriction endonuclease from genomic DNAs of two or more individuals of a species. RFLP was first developed in the early 1980s for use in human genetic applications and was later applied to plants for determining genetic relationships (Botstein *et al*., 1980). RFLPs are detected by hybridizing labeled DNA probes to southern blots containing DNA digested with

restriction enzymes. Southern hybridization analysis of RFLP requires skill and is time consuming. RFLP markers have been successfully used to assess genetic diversity in cassava (Debener *et al.*, 1990).

2.4.2 Random Amplified Polymorphic DNA (RAPD)

The RAPD technique described by Williams *et al* (1990) makes use of arbitrary short oligomers (usually 10-mer) which anneal to random homologous target sites within the genome allowing the generation of a product (Tautz, 1989). They are an attractive choice for determining genetic relationships because they are simple to detect and they don't require DNA sequence information or synthesis of specific primers (Saghai-Maroof *et al.,* 1994). The technique uses short oligodeoxynucleotide primers of arbitrary nucleotide sequence (amplifiers) and PCR procedures. RAPD generates dominant markers. These sequences are both abundant and highly polymorphic in plants (Tautz, 1989; Saghai-Maroof *et al.,* 1994). Polymorphism is based on the disruption or displacement of homologous target sites between individuals which results in the loss of a product (Tautz, 1989). However, since the fragments are amplified from very short, random DNA sequences which are used to prime the PCR, there is some uncertainty about the homology of fragments from different genotypes and about the genome origin of the fragments. RAPD markers have been successfully used to assess genetic diversity in sweet potato (Gichuki *et al*., 2003) and cassava (Marmey *et al*., 1994) as well as other species. They have a problem of reproducibility therefore RAPDs are rarely used these days because of the challenge highlighted.

2.4.3 Amplified Fragment Length Polymorphism (AFLP)

The AFLP technique is a PCR-based fingerprinting technique developed mainly to overcome the disadvantages of other PCR based techniques such as sensitivity to reaction conditions, DNA quality and PCR temperature profiles that limit their application (Vos *et al*, 1995). The steps involved in AFLP are restriction digestion/ligation, pre-selective amplification, selective amplification and visualization (Robinson and Harris, 1999). The technique has become an attractive tool because it generates a high number of polymorphic products from previously uncharacterized genomes (Powell *et al*., 1996), and is reproducible across laboratories (Jones *et al.,* 1997). AFLP markers have mainly been used for phylogeny analysis, cultivar/accession identification and genetic variation studies (Robinson and Harris, 1999). For instance, this has been used to assess genetic diversity of maize inbred lines (Vuylsteke *et al*., 2000) and in *Moringa Oleifera* Lam (Muluvi *et al*., 1999). AFLP has previously been used successfully in genetic diversity studies in sweet potato (Zhang *et al.,* 1998). They have also been used to assess variation of East African sweet potato cultivars (Gichuki *et al*, 2003). From these studies AFLP were found to be more informative and reproducible than other molecular markers like RAPD.

2.4.4 Single Nucleotide Polymorphism (SNP)

A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide differs between two homologous chromosomes. SNP may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genesn (Rafalski, 2002). SNPs within a coding sequence will not necessarily

change amino acid sequence of the protein that is produced due to degeneracy of the genetic code (Syvanen, 2001). A single nucleotide polymorphism in which both forms lead to the same polypeptide sequence is termed *synonymous* (silent mutation). SNPs that are not in protein coding regions may still have consequences for gene splicing, transcription factors binding, or the sequence of non-coding RNA. SNPs are currently used as markers in genetic analysis and in breeding programs (Gupta *et al.*, 2000). They are used in detection of alleles associated with genetic diseases in humans and identification of individuals (Nikiforov *et al*., 1994). They are invaluable as a tool for genome mapping, offering the potential for high density genetic maps, which can be used to develop haplotyping system for genes or regions of interest (Rafalski, 2002). The low mutation rate of SNPs also makes them excellent markers for studying complex genetic traits and as tool for the understanding of genome evolution (Syvanen, 2001).

2.4.5 Simple Sequence Repeats (SSR) or Microsatellites

Simple Sequence Repeats (SSRs) are hyper-variable tandem repeats of DNA motifs 2-5 bases long, common in eukaryotic and prokaryotic genomes (Zhu *et al*., 2000). They are widely distributed in higher plants (Morgante and Olivieri, 1993; Wang *et al*, 1994). Although they are ubiquitous (Kijas *et al.,* 1995), retrieval of SSRs has not been easy in plants because of their relatively low abundance compared with animal genomes (Zhu *et al*., 2000). The variation comes from differences in number of repeat units originating from error in copying of DNA during replication by DNA polymerase (Robinson and Harris, 1999). The repeated sequence is often simple, consisting of two, three or four nucleotides (di-, tri-, and tetranucleotide repeats, respectively) (Wang *et al.*, 1994). One

example of a microsatellite is a dinucleotide repeat (CA)n, where n is the variable between alleles (Tautz *et al*., 1986). SSRs are highly reproducible, polymorphic and co-dominant molecular markers (Lelley *et al*., 2000). Co-dominancy allows the identification of heterozygotes. These markers are therefore more informative than RAPD and AFLP markers (Robinson and Harris, 1999). The SSR technique requires prior PCR optimization and polymorphism screening. SSR markers have been useful for integrating the genetic, physical and sequence-based map in plant species and simultaneously have provided an efficient tool to link phenotypic and genotypic variation (Gupta and Varshney, 2000). They have been identified in many plant genomes including those of soybean (Akkaya *et al.,* 1992; Morgante and Olivieri, 1993); barley (Saghai-Maroof *et al.,* 1994) and in cassava, in addition SSR markers have been used to search for duplicates in the CIAT core collection (Chavarriaga-Aguirre *et al.*, 1999) and to analyze variation in natural population of putative progenitors of cassava (Olsen and Schaal, 2001).

2.5 Factors affecting optimization of PCR condition

The success of PCR depends on various factors namely: primer concentration which should be between 0.1 mM and 0.5 mM. For most applications 0.2 mM produces satisfactory results (Innis and Gelfand, 1990). Too high primer concentrations increase the chance of mispriming, which results in nonspecific PCR products and limiting primer concentrations can result in lower levels of amplification than possible. The concentration of DNA template depends on the source, normally a concentration of 100-250 ng is used for mammalian genomic DNA and 50 ng for cassava per 10 µl reactions

(Innis and Gelfand, 1990) as well as the concentration of dNTPs (dATP, dCTP, dGTP and dTTP) should be approximately 2mM, otherwise the optimum concentration will vary, and a too high concentration inhibits the PCR reaction. Magnesium chloride concentration is critical and should be optimized since DNA polymerase requires magnesium for its activity thus increasing magnesium increases the reagent activity and reaction kinetics, which results in a more efficient or quick reaction. Magnesium is usually supplied to PCR amplification in the form of magnesium chloride. A series of PCR experiments should be carried out with Mg^{2+} concentrations varying from 1.5 to 3.0 mM in 0.5 mM steps (Innis and Gelfand, 1990). High concentrations of chelating agents (such as EDTA) and negatively charged ionic groups (such as phosphates) should be avoided. Some suppliers of DNA polymerases have included NH_4^+ ions to their buffers. It has been shown that the presence of NH_4^+ ions results in a high specificity of the primer-template binding over a broad temperature range (Innis and Gelfand, 1990) The PCR reaction conditions and reaction times depend on the type of DNA polymerase used. A standard polymerase buffer works well for a wide range of templates and primers but may not be optimal for any particular combination. The number of cycles necessary to obtain a sufficient amount of PCR product largely depends on the reaction efficiency. In a typical PCR, the maximum amount of product is approximately 10^{12} copies of the template. Starting from one copy, the most efficient PCR would reach this level in 40 cycles. Depending on the nature of the DNA template, starting with many more copies and the rule of thumb is carry out PCR with 25 cycles for plasmid DNA and 30-35 cycles for genomic DNA.

Annealing temperature**.** Primer length and sequence are of critical importance in designing the parameters of a successful amplification: the melting temperature of nucleic acid duplex increases both with its length, and with increasing (G+C) content: a simple formula for calculation of the Tm is

$Tm = 4(G + C) + 2(A + T)\ ^{o}C.$

Thus, the annealing temperature chosen for a PCR depends directly on length and composition of the primer(s). One should aim at using an annealing temperature (Ta) of $5^{o}C$ below the lowest Tm of their pair of primers to be used (Innis and Gelfand, 1990). A more rigorous treatment of Ta is given by Rychlik *et al*. (1990): they maintain that if the Ta is increased by $1^{o}C$ every other cycle, specificity of amplification and yield of products <1kb in length are increased. One consequence of having too low a Ta is that one or both primers will anneal to sequences other than the true target. A consequence of too high a Ta is that too little product will be made, as the likelihood of primer annealing is reduced; another and important consideration is that a pair of primers with very different Ta's may never give appreciable yields of a unique product, and may also result in inadvertent "asymmetric" or single-strand amplification of the most efficiently primed product strand. Annealing temperature is one of the most important parameters that need adjustment in the PCR reaction. Moreover, the flexibility of this parameter allows optimization of the reaction in the presence of variable amounts of other ingredients (Innis and Gelfand, 1990).

2.6 Detection methods for fragments generated using molecular markers.

The use of fluorescent labelled microsatellite primers and laser detection (e.g. automated sequencer) in genotyping procedures significantly improves the throughput and automatisation (Wenz *et al.*, 1998). The use of microsatellites, however, can be costly due to the high price of the fluorescent labels, which must be carried by one of the primers in the primer pair. The cost of labelled microsatellite assay increases substantially if a large number of loci have to be tested. In order to overcome this financial difficulty, Schuelke (2000) introduced a novel procedure in which three primers are used for the amplification of a defined microsatellite locus a sequence-specific forward primer with M13 tail at its 5' end, a sequence-specific reverse primer and the universal fluorescent-labelled M13 primer. Other strategies that use universal extension and a complimentary fluorescent tail have been developed and are cost effective. Agarose and denaturing (DNA or protein is denatured by urea or SDS)/non-denaturing polyacrylamide gel electrophoresis are the two common methods for manual separation of fragments generated using molecular marker systems. Separated fragments are visualized either by ethidium bromide or silver staining of the gels. However, accurate sizing is difficult with both agarose and polyacrylamide gels and these matrices do not allow resolution to within a single base pair unit. Moreover, the mobility is also affected by sequence composition so that repeat unit confounds migration of complementary strands in a gel based system. For example CA strands move faster on denaturing polyacrylamide gels than GT strands and this can result in stutter bands which can complicate analysis of fragments. (Saitoh *et al*., 1998). In population analysis using microsatellites, accurate sizing of the microsatellite allele is crucial and ideally the

detection system needs to be able to differentiate a one base pair difference. This is made possible with the use of semi-automated capillary based systems such as the ABI Prism 377 instrument (Applied Biosystems, Warrington, UK). The advent of fluorescent-based capillary detection systems such as ABI 310, 3100, 3130 and 3730 which work on capillaries filled with polymers and picks samples directly from the plate has provided much greater resolution. This has overcome the earlier limitations of conventional gel based systems such as exposure to toxic chemicals, speed and sizing of alleles. The ABI 3730 is a fluorescent based capillary detection system that uses polymer as the separation matrix. This facilitates the accurate sizing of the microsatellite allele to within ± 0.3 base pairs (Buhariwalla and Crouch 2004). Co-loading and multiplexed based on dye label; fragment size and fluorescence reduce the unit cost of high throughput genotyping. The peaks are sized and the alleles called using Genotyper and GeneMapper^{TM} softwares. This system has the advantages of automated filling of capillaries, automated sample loading and rapid electrophoresis (Buhariwalla and Crouch, 2004). High throughput genotyping with several markers is now possible because the systems offer the possibility of PCR multiplexing and post-PCR co-loading since markers are labelled with different dyes and hence can be detected simultaneously (Elnifro *et al*., 2000).

2.7 Application of bioinformatics in sequence alignment.

In bioinformatics, Basic Local Alignment Search, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences (Casey, 2005). A BLAST

search enables a researcher to compare a query sequence with a database of sequences, and identify sequences that resemble the query sequences above a certain threshold. To run BLAST requires an input query sequence (also called the target sequence) and a sequence database. BLAST finds sequences from the queries that are identical to sequences in the database according to a defined threshold. In typical usage, the query sequence is much smaller than the database, e.g., the query may be one thousand nucleotides while the database is several billion nucleotides. Input and output, complies with the FASTA format. The BLAST algorithm can be conceptually divided into three stages (Casey, 2005).

- In the first stage, BLAST searches for exact matches of a small fixed length W between the query and sequences in the database. For example, given the sequences AGTTAC and ACTTAG and a word length W = 3, BLAST would identify the matching substring TTA that is common to both sequences. By default, W = 11 for nucleic seed (the nucleotide that initiates an alignment).

- In the second stage, BLAST tries to extend the match in both directions, starting at the seed. The ungapped alignment process extends the initial seed match of length W in each direction in an attempt to boost the alignment score. Insertions and deletions are not considered during this stage. For our example, the un-gapped alignment between the sequences AGTTAC and ACTTAG centred around the common word TTA would be:

..AGTTAC..

| || | ..

ACTTAG.

If a high-scoring un-gapped alignment is found, the database sequence is passed on to the third stage.

- In the third stage, BLAST performs a gapped alignment between the query sequence and the database sequence using a variation of the Smith and Waterman (1981) algorithm. Alignments that are above specified threshold are then displayed to the user.

2.8 Polymorphic Information Content (PIC)

Polymorphic information content is a measure of the informativeness of a genetic marker in any species (Botstein *et al.* 1980). The PIC of a genetic marker is estimated by:-

$$PIC=1-\sum_{i=1}^{n}p_i^2-2\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}p_i^2p_j^2\right]$$

where *p*i is the frequency of the *i*th allele and *n* is the number of alleles (Botstein *et al*. 1980; Ott 1992). It was developed as a measure of the utility of a co-dominant genetic marker for ascertaining the allele transmitted by an affected parent carrying a dominant allele (the affected parent is assumed to be heterozygous) (Ott 1992). Theoretically, PIC values range from O to 1. At a PIC of O, the marker has only one allele. At a PIC of 1, the marker would have an infinite number of alleles. A PIC value of greater than 0.7 is considered to be highly informative, whereas a value of 0.44 is considered to be moderately informative. A gene or marker with only two alleles has a maximum PIC of 0.375. Clearly markers with greater numbers of alleles and more even frequencies tend

to have higher PIC values and thus are more informative than those with a large number of rare alleles. The PIC value of each SSR marker is a measure of marker diversity.

2.9 Heterozygosity (H)

Heterozygosity (H) is used as a measure of allelic diversity in a population. The diversity of individuals increases as *H* increases. The heterozygosity of an individual is estimated by

$$H = 1 - \sum_{i=1}^{k} Pi^2$$

where *p*i is the frequency of the *i*th allele and *k* is the number of alleles (Nei 1987; Ott 1992). Heterozygosity has different meanings in inbred and outbred populations and is a function of the individuals and populations sampled. When individuals are sampled from genetically narrow or genetically isolated populations, you would expect to find fewer alleles and a higher frequency of monomorphic loci than when individuals are sampled from genetically diverse populations. *H* tends to increase as the number of alleles increases; however, *H* depends on the frequencies of different alleles. Regardless of the number of alleles at a locus, *H* is maximum when allele frequencies are equal and for *k* equally frequent alleles (Guo and Elston. 1999). The limit (*H* = 1.0) is virtually impossible to reach in practice; however, multi-allelic markers with alleles well distributed across populations can have heterozygosities in the 0.70 to 0.90 range On the other hand, a multi-allelic locus with several rare alleles may not be very polymorphic or informative. If a genetic marker has five equally frequent rare alleles and one common allele and the allele frequencies are 0.02 for the rare alleles and 0.9 for the common a

CHAPTER THREE

3.0 MATERIAL AND METHODS

3.1 Germplasm used in the study

In this study 32 cassava genotypes listed in (Table 1) were used. Of these 24 accessions were from South America obtained from Centro Internacional de Agricultural Tropical (CIAT) germplasm collection and eight from Africa obtained from IITA gene bank. Another eight genotypes extracted in IITA laboratory were parents of trait specific mapping populations, four (A-D) genotypes from drought tolerant mapping population, and four (E-H) genotypes from Tanzania for cassava brown streak virus. Seventy EST-SSRs markers/primers (Thereafter designated as ESSR1 to ESSR70) listed in Appendix 1 were screened for polymorphism. This research was carried out in IITA laboratory hosted in International Livestock Research Institute (ILRI) Kenya.

Table 1: The cassava genotypes that were used for optimization and polymorphism screen

ID CODE	GENOTYPE	COUNTRY	REGION
CIAT6	BRA206	BRASIL	South America
CIAT56	COL2459	COLOMBIA	South America
CIAT80	GUA59	GUATEMALA	South America
CIAT819	BRA200	BRASIL	South America
CIAT834	BRA436	BRASIL	South America
CIAT857	BRA785	BRASIL	South America
CIAT1226	BRA1016	BRASIL	South America
CIAT1212	BRA842	BRASIL	South America
CIAT1303	COL233	COLOMBIA	South America
CIAT391	ARG12	ARGENTINA	South America
CIAT567	PAR23	PARAGUAY	South America
CIAT543	CR19	COSTARICA	South America
CIAT694	COL2638	COLOMBIA	South America
CIAT759	TAI1	THAILAND	Asia
CIAT989	GUA43	GUATEMALA	South America
CIAT1135	USA7	USA	North America
CIAT1370	BRA1001	BRASIL	South America
IITA235	TME290	NIGERIA	South America
IITA270	TME396	TOGO	Africa
IITA372	TME539	UGANDA	Africa
IITA750	TME1368	GHANA	Africa
IITA766	TME1389	GHANA	Africa
IITA190	TME230	TOGO	Africa
IITA394	TME589	LIBERIA	Africa
CIATA	BRA255	BRASIL	South America
CIATB	VEN77	VENEZUELA	South America
CIATC	COL1734	COLOMBIA	South America
CIATD	COL1468	COLOMBIA	South America
IITAE	Albert (6215)	TANZANIA	Africa
IITAF	Namikonga (6217)	TANZANIA	Africa
IITAG	Kibalia (6218)	TANZANIA	Africa
IITAH	Kalolo (6219)	TANZANIA	Africa

A-D drought tolerant mapping population, **E-F** cassava brown streak virus mapping population

3.2 Identification of EST-SSR primers using Bioinformatics as a tool.

3.2.1 Source of sequences used to design EST-SSR primers

There was construction of cDNA from two normalized libraries from drought stressed and susceptible cassava by IITA in collaboration with Craig Venter Institute. The libraries comprised cDNAs from drought stressed cassava lines Sauti, Gomani, Mbundumali, TME 1 and Mkondezi which showed varying responses to drought stress from typical susceptible to rapid leaf loss and maintenance of stem meristem for rapid re-growth to stay green (Generation Challenge Programme, 2006). RNA was extracted from pools of root, leaf and stem meristem tissue using Invitrogen's concert RNA reagent following manufacturer's protocol. The pools of tissue comprised equal weights of tissue from each treatment and cultivar. cDNA synthesis and normalization was performed by Evrogen, (Moscow, Russia) following their own protocols (Generation Challenge Programme, 2006). TIGR then constructed cDNA libraries and performed the sequencing and alignments. Five hundred and three contigs were obtained. Further searches from NCBI and gene bank revealed 1657 SSR containing EST sequences. This rendered a total of 2160 SSR containing ESTs and out of these Craig Venter Institute designed 422 possible pairs of primers (Generation Challenge Programme, 2006).

3.2.2 Identification of sequence for designing of primer

From the database of 2160 SSR containing ESTs, TIGR were able to design 422 candidate primer pairs. Comparison between the database and already existing primers were done to eliminate duplicate SSRs. A BLASTN tool was used where forward and reverse primers of 817 existing SSRs were blasted against 2160 IITA/ Craig Venter

Institute. EST containing SSR at high stringency (0.0001) and low stringency (0.001). If a sequence matched both forward and reverse primers, and the motif was the same, then it was confirmed as an exact match (Appendix 2 and 3). The exact matches that resulted from the BLAST were compared with the 422 primer sequences designed by Craig Venter Institute and the similar ones were excluded. EST-SSRs of di- or tri-nucleotide repeat motifs that did not exist in the primer data base were used for primer synthesis. PCR amplification of 70 of these primers was optimized and used for polymorphism screening across a diverse array of 32 cassava genotypes.

3.3 Reconstitution of the primers

The 70 EST-SSR primers pairs were received from the manufacturer (Bioneer® Buckinghum, UK) in lyophilised form. They were reconstituted by dissolving them in 1mM tris EDTA after centrifuging them at 10,000 rpm for 15 minutes to make a concentration of 100 pmoles/µl by addition of low salt TE buffer to a volume computed depending on the manufacturer packed quantity, then the primer were left to dissolve overnight in –20^{o}C. The primers were then diluted with double distilled water to a 1.0 pmole/µl that made a working solution by diluting 1ul of stock solution to 99µl of double distilled water.

3.4 DNA extraction

Genomic DNA was extracted from young cassava leaf following the modified method of Dellaporta (Dellaporta *et al.,* 1983). Leaf tissue (0.5 g - 0.2 g) was ground to fine powder in liquid nitrogen using a mortar and pestle. The powdered tissue was

transferred to a frozen 1.5 ml eppendorf tube containing 800 µl of extraction buffer (100mM Tris-HCl, 50mM EDTA, 500mM NaCl, 1% PVP, 700 µl of B-mercaptoethanol in one litre of extraction buffer) at 65^{o}C. Subsequently, 50 µl of 20% SDS was added and put on a water bath at 65^{o}C with intermitted mixing. After incubation for 20 minutes, the tubes were cooled to room temperature and 250 µl of ice-cold 5M potassium acetate added with gentle mixing by inverting the tubes after which they were incubated on ice for 20 minutes. The solution was centrifuged at 12,000 rpm for ten minutes in eppendorf table top centrifuge. The supernatant was transferred to a fresh 1.5 ml eppendorf tube and one volume of ice-cold isopropanol added with gentle mixing followed by incubation at -80^{o}C for one hour and centrifugation at 12,000 rpm. The supernatant was poured off and last drops of isopropanol removed by placing the tubes facedown on paper towels. The pellet was re-suspended in 500 µl of T.E buffer (50mM Tris-HCL /10mM EDTA pH 9.0). The pellet was re-extracted and precipitated by addition of one volume of ice-cold isopropanol followed by incubation for 20 minutes and centrifugation at 12,000 rpm for ten minutes. 100 µl of 10mM Tris-HCL, 1mM EDTA containing 10 mg/ml RNase was added and stored at 4^{o}C overnight to dissolve the pellet. The DNA was transferred to fresh eppendorf tube and stored at –20^{o}C.

3.4.1 Determination of DNA Quality

Five microlitre from each DNA sample was mixed with 6 X gel loading dye (0.25 % bromophenol blue, 0.25 % xylene cyanol and 30 % glycerol) then electrophoresis was done through a 0.8% agarose gel containing ethidium bromide in a 1x TBE buffer (0.045

M Tris-Borate and 0.001 M EDTA) at 100 V for 1 hour. It was important to know the quality and quantity of extracted DNA as PCR reactions are optimized for specific DNA concentration.This method utilizes ultraviolet florescence emitted by ethidium bromide molecules intercalated into the DNA. The amount of florescence was proportional to the total mass of DNA. The quality of DNA in the sample could be determined by check the smearing on the gel. The quantity of DNA in the sample could be estimated by comparing the florescent yield of the sample with that of a series of known DNA mass standards.

3.4.2 Determination of DNA Quantity

A spectrophotometer was used for estimating the purity of DNA extracted and samples obtained from CIAT and IITA gene banks. The concentrations of the genomic DNA were determined by mixing 5 µl of DNA with 495 µl of double distilled water in a microfuge tube. Quantification was done by adding the diluted DNA sample into a 10 mm ultraviolet silica cuvette after a blank, containing double distilled water only, had been set and then loaded into a GeneQuant ® spectrophotometer (Biochrom, Cambridge, UK). Absorbance was calculated automatically by the GeneQuant ® spectrophotometer (Biochrom, Cambridge, UK) and the DNA concentration printout generated. The Optical Density (OD) was taken at 260 nm and 280 nm and the ratio OD 260: OD 280 was calculated. Standardization of the DNA was performed by dilution of the stock DNA that was more than 50 ng/µl by adding double distilled water accordingly.

3.5 Preliminary optimization of PCR conditions

The assay to optimize the PCR conditions for each primer was conducted over the following range: template DNA total 50 ng to 15 ng, $MgCl_2$ 1 mM to 2 mM, dNTPs 0.1 mM, to 2.0 mM, primer 0.4 pmol to 1.6 pmol. Constant amount of *Taq* DNA polymerase (0.375U) and a constant amount of 1x *Taq* buffer was used. The total reaction volume was 10 µl. An empirically devised strategy based on modification of Taguchi method (Taguchi, 1987) was used to test the components of PCR together. The optimization trials consisted of four reaction components each at three concentration levels as shown in Table 2 and 3. The stocks used for the initial optimization consisted of a constant amount of *Taq* polymerase at 0.5U per 10 µl reaction volume; primer at 0.8 pmol/µl F/R; $MgCl_2$ at 25 mM; dNTPs at 25 mM; and template DNA at 50ng/µl. Based on this strategy, the initial optimization of the four major components in a PCR: concentrations of primer, template DNA, Mg^{2+} and dNTPs were empirically determined for all the primers used in this study then the PCR products electrophoresis was done through a 2.0 % agarose gel containing ethidium bromide in a 1x TBE buffer (0.045 M Tris-Borate and 0.001 M EDTA) at 100 V for 1 hour.

Table 2: Concentration levels for the components in a PCR.

Component	A	B	C
Primer F and R	0.4pmol	0.8pmol	1.2 pmol
DNA	5ng	10ng	15ng
Mg	1.0mM	1.5mM	2.0mM
dNTP	0.1mM	0.15mM	0.2mM

A, B and C are varying PCR components concentrations.

Table 3: Orthogonal array to test four components each at three different concentrations as shown in Table 2.

Reactions	Primer	DNA	Mg2+	dNTP
1	A	A	A	A
2	A	B	B	B
3	A	C	C	C
4	B	A	B	C
5	B	B	C	C
6	B	C	B	C
7	C	A	C	B
8	C	B	A	A
9	C	C	B	C
10	B	B	B	B
11	C	C	C	C
12	C	A	B	B

1-12 represents different PCR conditions testing varying primer, DNA, Mg2+ and dNTP concentrations.

3.6 Determination of optimal annealing temperature using gradient thermocycler

The optimal annealing temperature for each primer pair was determined by gradient PCR. This was done by running a PCR with one DNA sample. The PCR condition were:-50 ng DNA; 0.08 pmole/µl (F and R) primer; 0.375U *Taq* polymerase; 1.5 mM $MgCl_2$; 0.2 mM dNTPs and 1x PCR buffer (50 mM KCl; 10 mM Tris-HCl (pH 8.3)). The reaction was done in volume of 10 µl and consequently the reaction was run in a gradient thermocycler that subjected the DNA sample to different annealing temperatures ranging between 55.8^{o}C to 68.1^{o}C. A basic PCR profile consisting of 35 cycles was used in the initial optimizations. The PCR products were electrophoresed on a 2.0 % agarose gel containing ethidium bromide in a 1x TBE buffer (0.045 M Tris-Borate and 0.001 M EDTA) at 100 V for 1 hour. To confirm the annealing temperature

chosen, the PCR was repeated using four different DNA samples (6215, 6217, 6218 and 6219 as indicated in Table 1 in page 26.) amplified under the same PCR conditions.

3.7 Magnesium concentration and fluorescent tail optimization

Using a chosen 62^{o}C optimum annealing temperature each primer was amplified across one DNA sample over a range of magnesium concentrations (1.0 mM, 1.5 mM, 2.0 mM, 2.5 mM and 3.0 mM), 50ng of genomic DNA; 0.08 pmole/µl (F and R) primer; 0.375U *Taq* polymerase; 0.2 mM dNTPs; 1x PCR buffer (50 mM KCl; 10 mM Tris-HCl (pH 8.3)) and 0.125, 0.150 and 0.175 pmole/µl fluorescent tail. The reaction mix was done in total volume of 10 µl. The PCR temperature profile consisted of five minutes initial template denaturation step at 95°C, followed by 30 cycles of 30 seconds at 95°C, 2 minute at primer annealing temperature and 30 seconds at 72°C. This was followed by a final primer extension of 72°C for 30 minutes. A negative control was added consisting of all PCR components apart from DNA. The PCR products and a blank were electrophoresed on a 2.0 % agarose gel containing ethidium bromide in a 1 x TBE buffer (0.045 M Tris-Borate and 0.001 M EDTA) at 100 V for one hour. PCR products were then stored at 4°C prior to capillary electrophoresis fragment analysis.

3.8 Microsatellite genotyping

Once amplified, the PCR products were genotyped using ABI 3730 GeneMapper® software which utilizes an internal size standard, GeneScan-500Liz® which is orange in color, for sizing the DNA fragments. The master mix was prepared by adding 12 µl of

Liz standard to 1 ml of HiDi™ formamide and mixed. This mixture was enough for 96 reactions. 1 µl of the PCR products were aliquoted then pipetted into individual wells of a microplate to which 9 µl of the standard/formamide mix was added and gently mixed. The PCR products were denatured at 95 ^{O}C for five minutes and immediately chilled on ice to prevent re-association of the DNA. The denatured PCR fragments were separated using an ABI 3730 (Applied Biosystems, Warrington, UK) automated capillary DNA sequencer.

3.9 Post-PCR co-loading optimization for high throughput genotyping

All the successful primer amplified products were analyzed by loading in sets of four primers together to achieve high throughput genotyping. First the loading volume had to be optimized. The co-loading sets volumes varied from one microlitre to three microlitres of pooled products from four different markers bearing different fluorophore, according to their strength of amplification then one microlitre was then mixed with 9.0µl of capillary electrophoresis cocktail (12µl GeneScan-500Liz® size standard, 1ml Hidi™ formamide). Each of the DNA fragments was labeled with a proprietary fluorophore, which results in a single peak when run under denaturing conditions. The use of an internal lane size standard enabled automated data analysis using the Applied Biosystems fluorescence-based DNA electrophoresis systems and was essential for achieving high precision in sizing DNA fragments by electrophoresis. The formamide prevents the re-naturation of single stranded DNA.

3.10 Polymorphism screen of the EST – SSR markers

Following successful annealing and magnesium optimization of all primer pairs, 32 DNA genotypes Table 1 on page 27 were amplified across using the optimized PCR conditions Table 8 on page 48. The PCR conditions were done in a total volume of 10 µl. The PCR thermocycler profile consisted of 5 minutes initial template denaturation step at 95°C, followed by 30 cycles of 30seconds at 95°C, 2 minute at primer annealing temperature and 30 seconds at 72°C, this was followed by a final primer extension of 72°C for 30 minutes. The PCR products were electrophoresed on a 2.0 % agarose gel containing ethidium bromide in a 1x TBE buffer (0.045 M Tris-Borate and 0.001 M EDTA) at 100 V for 1 hour. PCR products were then stored at 4°C prior to capillary electrophoresis fragment analysis.The amplified PCR products were analysed using ABI 3730 genotype sequencer (Applied Biosystems, Warrington, UK).

3.11 DNA fragment analysis

The DNA fragments were analyzed using GeneMapper^{TM} software Version 3.7 (Applied Biosystems, Warrington, UK). Panels containing name, color and size range of fragment were prepared for each microsatellite locus. The panels were subsequently imported into the software for analysis. The third order least squares method was used for allele size calling. This method was selected because it uses regression analysis to build a best-fit cubic function curve. Genotypic data generated was then exported to Microsoft Excel^{TM} for subsequent editing. Samples which Relative Florescent Unit (RFU) was below 1500 they were re-amplified and re-genotyped. Power marker version 3.0 was

used to analyze the data. Polymorphic Information Content (PIC) for each primer were determined as described by Weir (1996)

$$PIC=1-\sum_{i=1}^{n} p_i^2 - 2\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i^2 p_j^2\right]$$

where *Pi* was the frequency of the i^{th} allele.

PIC values give the information that each marker impacts to the study, which is the measure of the usefulness of each marker in distinguishing one individual from another. PIC values are affected by the number and frequency of alleles in the population under study. The resulting polymorphic markers were used to describe the genetic diversity of the 32 cassava genotypes. Genetic diversity was measured in terms of number of alleles per locus and Nei's unbiased estimate of gene diversity (H) also known as expected Heterozygosity (He) (Nei, 1987). The difference in value between expected Heterozygosity (He) and observed heterozygosity (direct count) (Ho) provides an indication of the deviations from random mating system in relation to Hardy-Weinberg (H-W) equilibrium (Nei, 1987). Distance matrices were calculated on the basis of Euclidean distance among population and geographic origins. The distance matrices were subjected to UPGMA (Unpaired Group Method of Arithmetic averages) and clustering was done resulting in a construction of dendrogram using Darwin software. Bootstrap resampling (n = 1000) was performed to test the robustness of the dendrogram topology.

CHAPTER FOUR

4.0 RESULTS

4.1 BLAST results

The Eight hundred and seventeen (817) known primers were blasted against 2160 sequences containing EST-SSRs (Appendix 2 and 3). For the most stringent BLAST, forward primers generated 229 matches, while least stringent there were 246 matches. For the most stringent BLAST, reverse primers generated 252 matches, while least stringent there were 263 matches. Consequently one hundred and eighty six exact matches of TIGR EST-SSR sequences with known F and R primers were found. When the one hundred and eighty six exact matches were compared with the 422 sequences from TIGR candidate for primer designing, 55 of these perfectly matched. The remainder of 367 primer pairs flanking were unknown SSRs. A total of 177 primers were of single nucleotide repeat that were excluded from the study. The other 190 remaining primers pairs were dinucleotide repeats and higher repeats and they were distributed as shown in the Table 4 below. For the purpose of this study a total of seventy primers were selected for optimisation, of which 45 (24.0%), were dinucleotide repeats, 24 (13.0%) trinucleotide and 1 (0.53%) tetra nucleotide repeats. No hexanucleotides were selected for this study.

Table 4: Total number and distribution of EST-SSR primers after bioinformatics screening.

Repeat type	No.	Percentage (%)
Dinucleotide	49	26.0
Trinucleotide	133	70.0
Tetranucleotide	02	01.0
Hexanucleotide	06	03.0
Total	190	100.0

4.2 Determination of DNA quality.

The quality of genomic DNA preparation was analyzed using agarose gel electrophoresis (Figure 1). This was a critical step since high quality genomic DNA was required for PCR during genotyping. Analysis of the quality of genomic DNA was based on the electrophoretic mobility of the DNA samples on the gel. Electrophoresis in agarose gel separates out DNA fragments according to size (molecular weight) and charge. Largest fragments move much slower than short fragments. The concentration of the sample DNA loaded was estimated visually by comparison of the band intensity with a standard DNA of known concentrations e.g. Lambda DNA. The slow uniform electrophoretic mobility of the DNA indicates high molecular weight DNA (long fragments) of intact genomic DNA. Impurities in genomic DNA preparations such as carbohydrates, proteins and polyphenols interfere with the electrophoretic mobility of the DNA due to their interaction with the DNA and made it difficult for DNA to move out of the wells. A smearing on the gel indicates small fragments of degraded DNA.

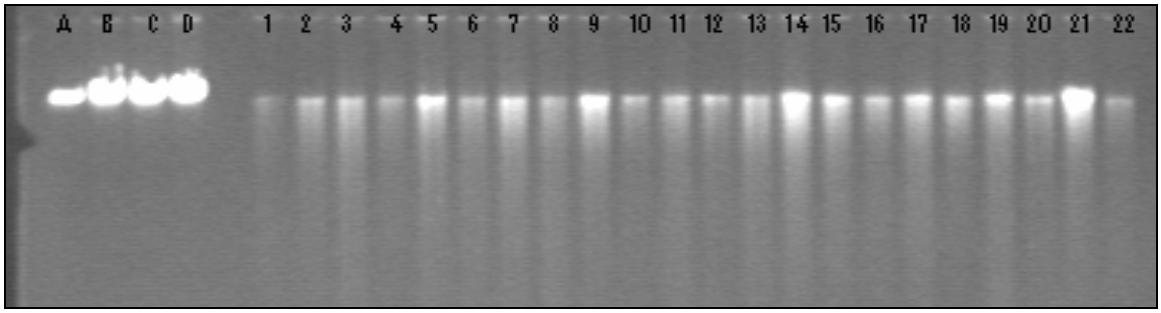

Figure 1: 0.8% Agarose gel showing the quality of South American cassava accessions. **A-D**; DNA concentration **A** = 50 ng **B**= 100 ng **C**= 150 ng **D**= 200 ng. lane one *1.BRA206 2.COL2459 3.GUA59 4.CUB1 5.BRA200 6.BRA436 7.BRA785 8.BRA990 9.BRA1016 10.BRA842 11.COL233 12.ARG12 13.PAR2314. MEX55 15.CR19 16.PER458 17.COL2638 18.TAI1 19.GUA.43 20.USA7 21.BRA1001 22.COL297*

4.3 Quality check and quantity estimation of DNA using agarose gel.

Agarose gel electrophoresis analysis enables quick and easy quantification and quality check of DNA. As little as 10 ng DNA can be detected by agarose gel electrophoresis with ethidium bromide staining. DNA samples are run on 0.8 % TBE agarose gel alongside 20 ng/µl lambda DNA digested with Hind III and EcoR1. The amount of sample DNA loaded was estimated by comparison of the band intensity with the lambda standard visually. The samples contained between 20 ng to 250 ng DNA as estimated by visual comparison with the standard marker. (Tables 5 and 6)

Table 5: Estimation of DNA concentration by agarose electrophoresis for African DNA samples

ID Code	Genotype	Country	Sample conc. (ng/µl)
IITA235	TME290	NIGERIA	100
IITA270	TME396	TOGO	150
IITA372	TME539	UGANDA	200
IITA750	TME1368	GHANA	250
IITA766	TME1389	GHANA	250
IITA190	TME230	TOGO	250
IITA394	TME589	LIBERIA	250
IITA E	Albert (6215)	TANZANIA	50
IITA F	Namikonga (6217)	TANZANIA	100
IITA G	Kibalia (6218)	TANZANIA	50
IITA H	Kalolo (6219)	TANZANIA	50

Table 6: Estimation of DNA concentration by agarose gel for South American DNA samples

ID Code	Genotype	Country	Sample conc. (ng/µl)
CIAT80	GUA59	GUATEMALA	20
CIAT62	CUB1	CUBA	30
CIAT819	BRA200	BRASIL	50
CIAT834	BRA436	BRASIL	20
CIAT857	BRA785	BRASIL	50
CIAT880	BRA990	BRASIL	20
CIAT1226	BRA1016	BRASIL	50
CIAT1212	BRA842	BRASIL	20
CIAT1303	COL233	COLOMBIA	40
CIAT391	ARG12	ARGENTINA	30
CIAT567	PAR23	PARAGUAY	30
CIAT443	MEX55	MEXICO	60
CIAT543	CR19	COSTARICA	50
CIAT584	PER458	PERU	40
CIAT694	COL2638	COLOMBIA	50
CIAT759	TAI1	THAILAND	40
CIAT989	GUA43	GUATEMALA	50
CIAT1135	USA7	USA	40
CIAT1370	BRA1001	BRASIL	100
CIAT1089	COL297	COLOMBIA	30
CIATA	BRA255	BRASIL	50
CIATB	VEN77	VENEZUELA	100
CIATC	COL1734	COLOMBIA	100
CIATD	COL1468	COLOMBIA	50

4.4 Spectrophotometric determination of DNA concentration.

The DNA samples were obtained from lab extraction, CIAT germplasm and IITA gene bank collection. They were quantified on the nanodrop by reading OD 260/280. All the DNA samples were of good quality as their OD ratios reading were ranging from 1.4-2.0 (Table 7). The OD 260 : OD 280 generally ranged between 1.4 and 2.0 indicating the DNA samples were reasonably free from contaminants such polyphenolics,

polysaccharides and proteins.The readings for estimating DNA concentration from the gel positively correlated with ones from spectrophotometer indicating this could also be used as an alternative for DNA quantification.

Table 7: Spectrophotometer reading for diverse cassava genotypes.

ID CODE	GENOTYPE	COUNTRY	Sample conc ng/µl	260/280
CIAT6	BRA206	BRASIL	36.19	1.63
CIAT56	COL2459	COLOMBIA	58.97	1.6
CIAT80	GUA59	GUATEMALA	25.14	1.38
CIAT819	BRA200	BRASIL	44.63	1.73
CIAT834	BRA436	BRASIL	15.54	1.35
CIAT857	BRA785	BRASIL	60.85	1.7
CIAT1226	BRA1016	BRASIL	78.33	1.86
CIAT1212	BRA842	BRASIL	55.63	1.75
CIAT1303	COL233	COLOMBIA	39.81	1.59
CIAT391	ARG12	ARGENTINA	26.24	1.55
CIAT567	PAR23	PARAGUAY	32.49	1.75
CIAT543	CR19	COSTARICA	79.9	1.82
CIAT694	COL2638	COLOMBIA	79.3	1.82
CIAT759	TAI1	THAILAND	44.9	1.72
CIAT989	GUA43	GUATEMALA	35.66	1.7
CIAT1135	USA7	USA	36.38	1.53
CIAT1370	BRA1001	BRASIL	94.54	1.86
IITA235	TME290	NIGERIA	97.42	1.78
IITA270	TME396	TOGO	293.66	1.84
IITA372	TME539	UGANDA	333.19	1.78
IITA750	TME1368	GHANA	415.23	1.85
IITA766	TME1389	GHANA	212.4	1.85
IITA190	TME230	TOGO	322.54	1.85
IITA394	TME589	LIBERIA	330.21	1.85
CIATA	BRA255	BRASIL	17.34	1.49
CIATB	VEN77	VENEZUELA	98.34	2.03
CIATC	COL1734	COLOMBIA	111.11	1.7
CIATD	COL1468	COLOMBIA	37.11	1.67
IITAE	Albert (6215)	TANZANIA	65.49	2.01
IITA F	Namikonga (6217)	TANZANIA	93.97	2.01
IITAG	Kibalia (6218)	TANZANIA	56.31	2.03
IITAH	Kalolo (6219)	TANZANIA	46.77	1.94

A-D drought tolerant mapping population, **E-F** cassava brown streak virus mapping population

4.5 Preliminary amplification of PCR conditions involving four components.

From the preliminary reactions the combination of 0.8 pmole/µl primer, 5 ng, 10 ng and 15 ng DNA, 1.5 mM and 2.0 mM $MgCl_2$ and 0.15 mM and 0.2 mM dNTPs showed appreciable PCR amplification at 65^{o}C annealing temperature.

4.6 Determination of annealing temperature.

The primer annealing temperatures were determined by running the primers on a gradient temperature using one DNA sample and later four DNA samples. This approach ensured that primers were subjected across different temperatures which resulted in different product intensities when viewed on 2.0 % agarose gel. It was observed that the primers amplified across a wide range of temperatures 55.8 ^{o}C to 68.1 ^{o}C (Figure 2). The optimum annealing temperature was 62^{o}C except for two markers ESSR 5 and ESSR 9 that amplified better at 57 ^{o}C (Figure 3 and 4). Testing annealing temperatures were advantageous during the initial optimization steps, and ultimately reduced the number of temperature profiles to be tested during the optimization of new primer sets.

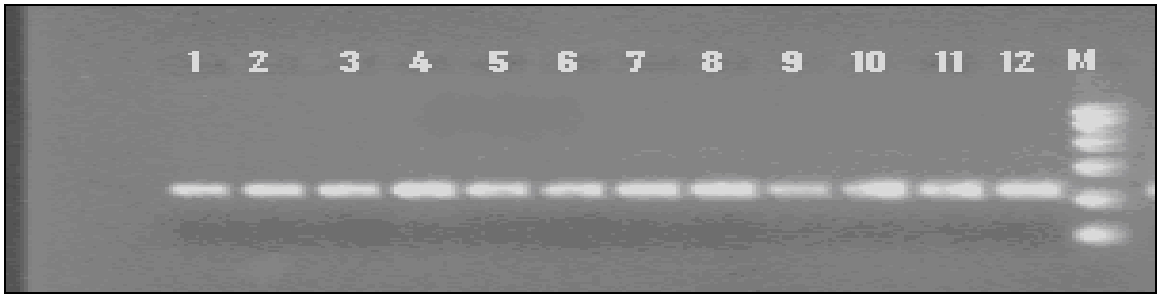

Figure 2: 2.0 % agarose gel showing gradient PCR testing annealing temperature for ESSR 7. **M**=Standard marker (Promega) 100 base pair ladder, lane one. **1**. 55.8 ^{o}C, **2**. 56.2 ^{o}C, **3**. 57.4 ^{o}C, **4**. 58.8 ^{o}C, **5.** 60.0 ^{o}C, **6**. 61.3 ^{o}C **7**.62.6 ^{o}C **8.** 64.1 ^{o}C **9**. 65.4 ^{o}C **10**. 66.8 ^{o}C **11**.67.6 ^{o}C **12**. 68.1 ^{o}C

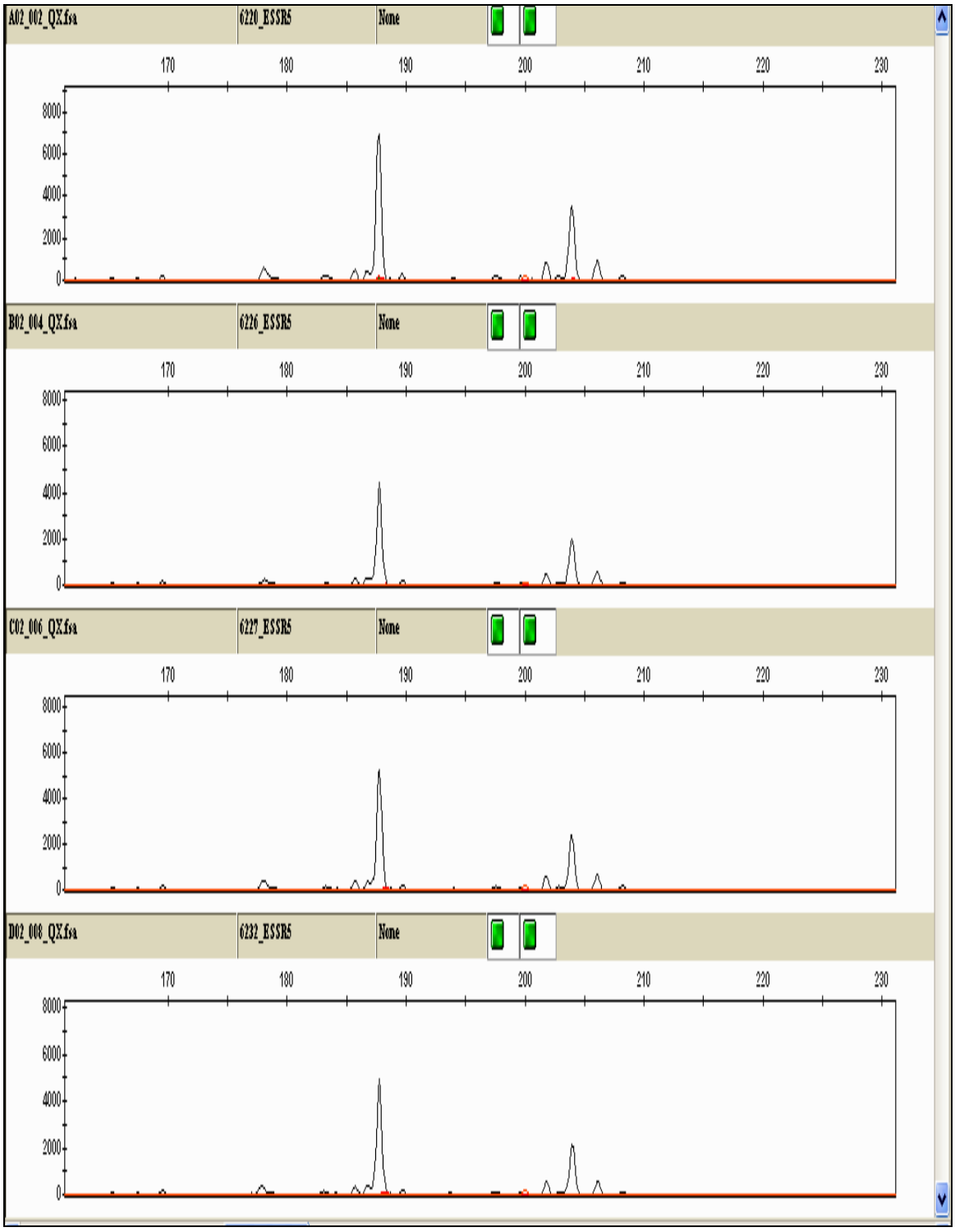

Figure 3: Electropherogram showing high amplification of marker 5 at $57^{O}C$

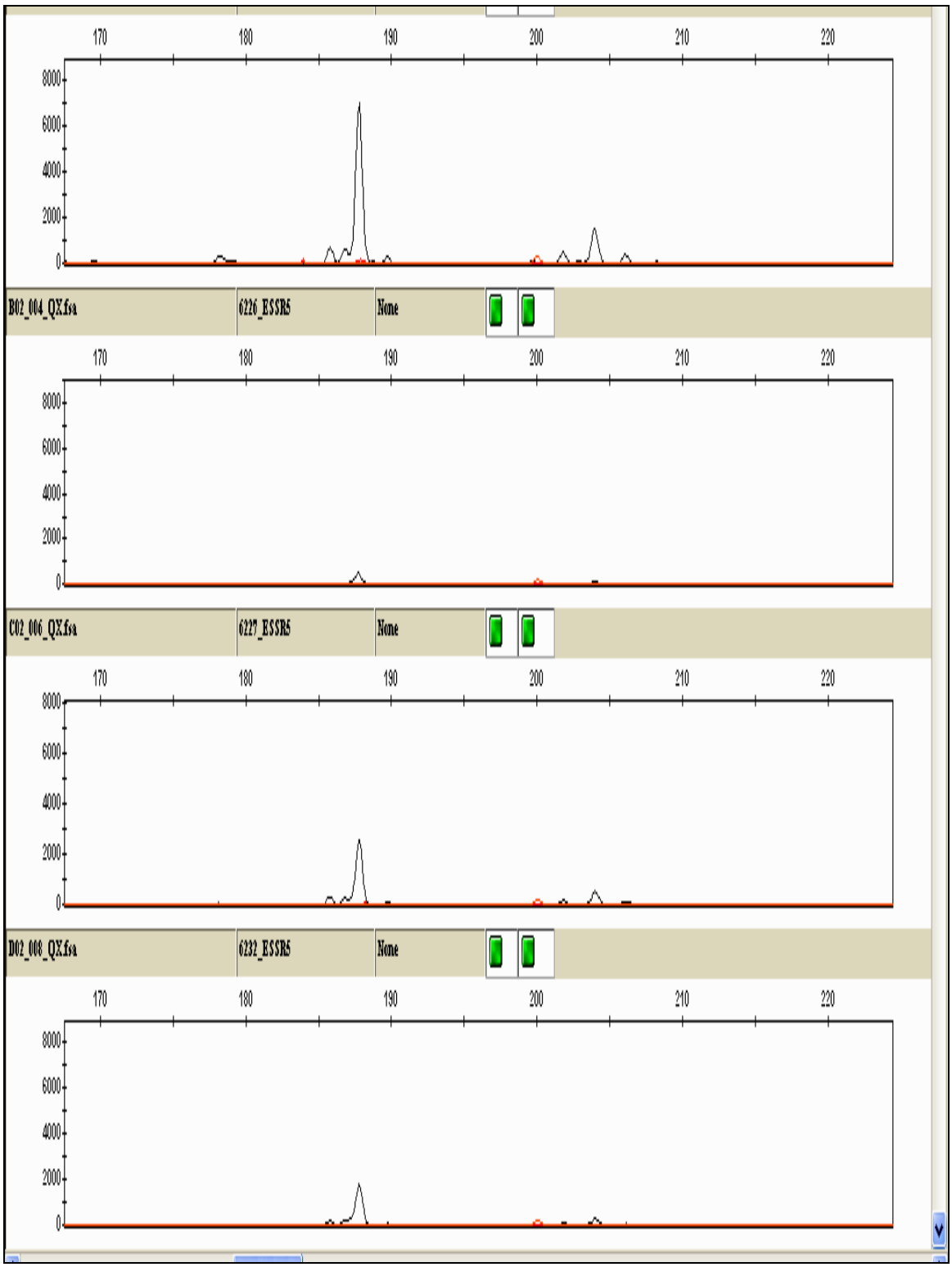

Figure 4: Electropherogram showing low amplification of marker 5 at 62^{O} C

4.7 Magnesium optimization using four different DNA samples

All the markers showed a very good amplification at the chosen temperature of 62 ^{o}C as represented in Figure 5. However, it was noted that ESSR 6, ESSR 10, ESSR 12, ESSR 18, ESSR 28, ESSR 49, ESSR 54, ESSR 55, ESSR 63, ESSR 67 and ESSR 68 failed to amplify while marker ESSR 16, ESSR 31, ESSR 48, ESSR 52, ESSR 59, ESSR64 and ESSR 69 amplified but the product size was greater than expected product size. Magnesium optimization was done to eliminate the A+ peaks that make it difficult to call alleles during data analyses. From the results of the peak it was clear that magnesium has an effect on the peaks (Figure 6). The magnesium concentrations that amplified were 1.0 mM, 1.5 mM, 2.0 mM, 2.5 mM and 3.0 mM Table 8 on page 48. The fluorescent tail concentration appeared to have no effect on the peaks thereby the concentration of 0.175 pmole/µl was the optimum and hence preferred for subsequent PCR reactions.

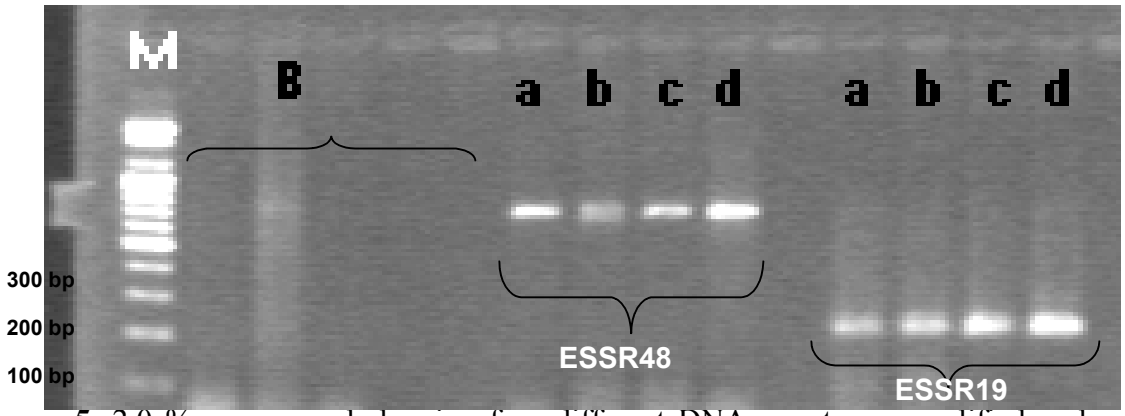

Figure 5: 2.0 % agarose gel showing four different DNA genotypes amplified under 2.0mM magnesium chloride concentration at 62^{o}C annealing temperature **a**:6215, **b:**6217.**, c:**6218 and **d:**6219. **M** = marker 100 base pair ladder (Promega), **B:** blank.

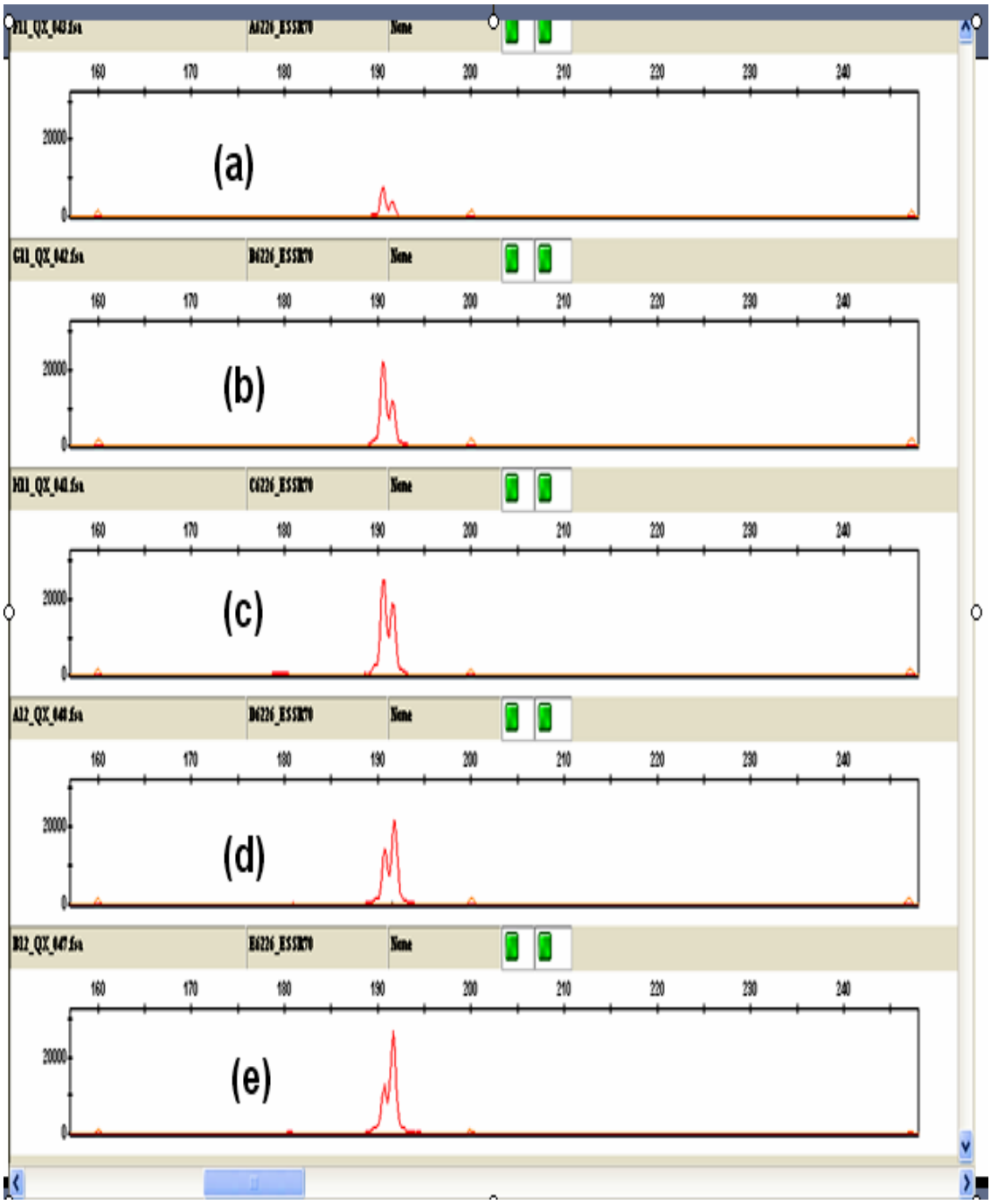

Figure 6: Electropherogram showing effect of varying magnesium chloride concentration on ESSR70, 3.0mM was the optimum concentration (a) 1.0 mM, (b) 1.5mM, (c) 2.0 mM, 2.5 mM, (e) 3.0 mM. .

4.8 Optimum PCR conditions

The following condition proved to be optimum after thorough screening; 0.175 Pmole/µl dye, 50ng Total DNA, 0.08 Pmole/µl F/R primer,1.0x buffer, 0.375U *Taq,* 0.2mM dNTPs, and 62^{o}C annealing temperature but ESSR5 and ESSR9 57^{o}C was the optimum while magnesium chloride had varied concentrations as shown in Table 8.

Table 8: The concentrations of optimized PCR conditions for all the 52 EST-SSR primers in 10 µl reaction volume.

SNO.	MARKER	$MgCl_2$ mM
1	IITA ESSR01	2.0
2	IITA ESSR02	2.0
3	IITA ESSR03	1.0
4	IITA ESSR04	2.0
5	IITA ESSR05	1.0
6	IITA ESSR07	2.0
7	IITA ESSR08	2.0
8	IITA ESSR09	2.0
9	IITA ESSR11	2.0
10	IITA ESSR13	1.0
11	IITA ESSR14	1.0
12	IITA ESSR15	2.0
13	IITA ESSR17	3.0
14	IITA ESSR19	2.0
15	IITA ESSR20	1.0
16	IITA ESSR21	1.0
17	IITA ESSR22	2.0
18	IITA ESSR23	2.0
19	IITA ESSR24	2.0
20	IITA ESSR25	2.0
21	IITA ESSR26	3.0
22	IITA ESSR27	3.0
23	IITA ESSR29	2.0
24	IITA ESSR30	3.0
25	IITA ESSR32	3.0
26	IITA ESSR33	2.0
27	IITA ESSR34	1.5

Table 8 cont…

28	IITA ESSR35	2.0
29	IITA ESSR36	2.0
30	IITA ESSR37	1.5
31	IITA ESSR38	1.0
32	IITA ESSR39	2.0
33	IITA ESSR40	1.0
34	IITA ESSR41	1.0
35	IITA ESSR42	2.0
36	IITA ESSR43	1.0
37	IITA ESSR44	3.0
38	IITA ESSR45	1.0
39	IITA ESSR46	2.5
40	IITA ESSR47	1.0
41	IITA ESSR50	2.0
42	IITA ESSR51	1.0
43	IITA ESSR53	1.0
44	IITA ESSR56	2.0
45	IITA ESSR57	2.0
46	IITA ESSR58	1.5
47	IITA ESSR60	2.0
48	IITA ESSR61	1.0
49	IITA ESSR62	2.0
50	IITA ESSR65	2.0
51	IITA ESSR66	2.0
52	IITA ESSR70	3.0

4.9 Post PCR Co-loading for high through-put genotyping

DNA marker screening of many genotypes with several markers required an effective strategy which should be cost effective and time saving. In this study the first strategy was to optimize PCR conditions for each primer and the second strategy was to establish a co-loading set whereby several markers were genotyped and analyzed simultaneously on the automated capillary electrophoresis system ABI-3730 (Applied Biosystems, Warrington, UK). To achieve high through-put genotyping, the optimum PCR conditions (Table 8) and the post PCR co-loading conditions (Tables 9) were established. Figure 7 shows an electropherogram obtained from ABI 3370 of optimized co-loading conditions using four different dyes (PET-red, VIC-green, FAM-blue and NED- yellow). The loading volume was optimised to avoid overloading that result in peaks with pink colour when analysed in the ABI machine.

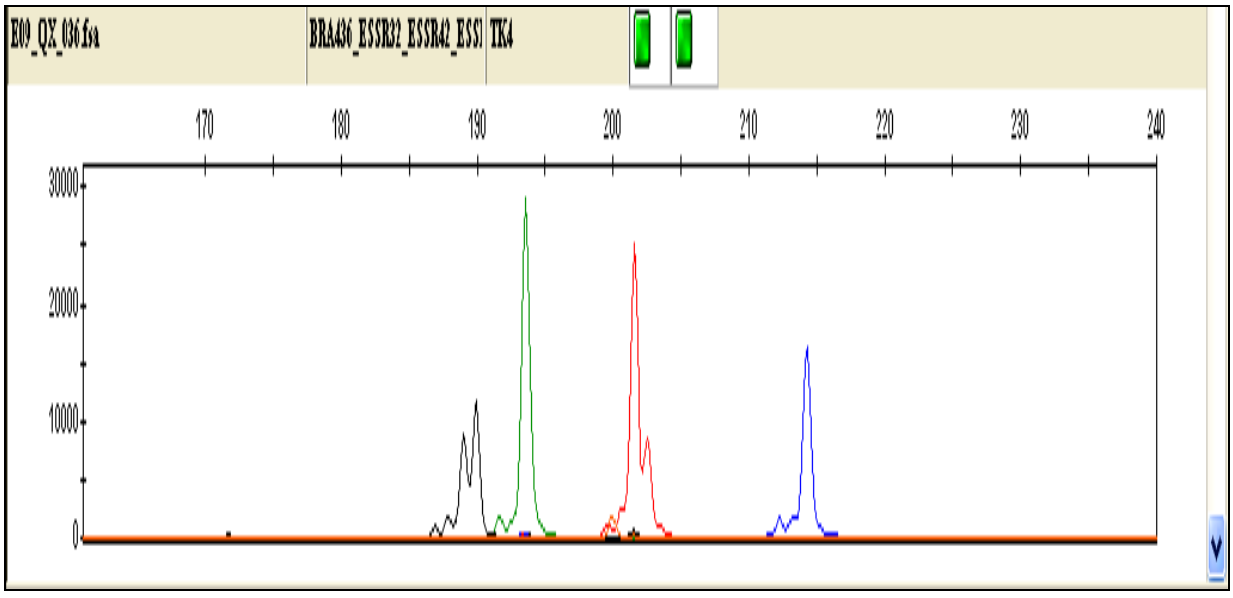

Figure 7: Electropherogram showing optimised co-loading set volume for high through put analysis **NED - yellow VIC-green PET-red and FAM-blue**

Table 9: (a-n): Optimized co-loading sets volumes for high through-put genotyping.

(a)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR60	178	2.0	PET	1.0 µl
ESSR25	184	2.0	VIC	1.5 µl
ESSR15	189	2.0	FAM	1.2 µl
ESSR39	211	2.0	NED	1.0 µl

(b)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR24	174	2.0	PET	1.5 µl
ESSR19	185	2.0	VIC	1.5 µl
ESSR29	205	2.0	FAM	1.2 µl
ESSR23	216	2.0	NED	1.2 µl

(c)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR32	202	2.0	PET	1.2 µl
ESSR42	188	2.0	VIC	1.2 µl
ESSR33	214	2.0	FAM	1.2 µl
ESSR30	194	3.0	NED	1.0 µl

(d)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR8	174	2.0	PET	2.5 µl
ESSR7	181	2.0	VIC	1.3 µl
ESSR1	188	2.0	FAM	1.3 µl
ESSR35	211	2.0	NED	1.0 µl

(e)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR21	178	1.0	PET	1.5 µl
ESSR2	186	2.0	VIC	1.2 µl
ESSR20	199	1.0	FAM	1.7 µl
ESSR57	211	2.0	NED	1.0 µl

(f)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR66	181	2.0	PET	1.0 µl
ESSR11	188	2.0	VIC	1.5 µl
ESSR65	216	2.0	FAM	1.0 µl
ESSR3	225	1.0	NED	1.0 µl

(g)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR40	177	1.0	PET	1.0 µl
ESSR47	183	1.0	VIC	1.0 µl
ESSR53	217	1.0	NED	1.0 µl

(h)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR17	179	3.0	PET	1.7 µl
ESSR26	193	3.0	VIC	3.5 µl
ESSR27	203	3.0	FAM	2.5 µl
ESSR4	188	2.0	NED	2.0 µl

(i)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR70	191	3.0	PET	1.0 µl
ESSR41	201	1.0	VIC	1.0 µl
ESSR44	205	2.0	FAM	1.5 µl
ESSR61	220	1.0	NED	1.0 µl

(j)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR37	176	1.5	PET	1.0 µl
ESSR9	178	2.0	VIC	2.0 µl
ESSR56	182	1.0	FAM	1.0 µl
ESSR22	205	1.0	NED	1.5 µl

(k)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR5	189	1.0	PET	3.0 µl
ESSR58	189	1.5	VIC	2.0 µl
ESSR14	202	1.0	FAM	2.5 µl
ESSR38	197	1.0	NED	1.5 µl

(l)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR13	184	1.0	PET	1.3 µl
ESSR34	207	2.0	VIC	1.3 µl
ESSR36	174	2.0	FAM	1.0 µl

(m)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR62	207	1.5	PET	1.0 µl
ESSR50	184	1.0	VIC	1.3 µl

(n)

PRIMER	OBSERVED PRODUCT SIZE	Mg Concentration mM	LABEL	LOADING VOL.
ESSR43	191	1.0	PET	1.0 µl
ESSR46	198	2.5	NED	1.0 µl
ESSR51	197	1.0	FAM	1.0 µl
ESSR45	176	1.0	VIC	1.2 µl

Label = PET-Red, NED-Yellow, FAM-Blue and VIC-Green

4.10 Polymorphism screening and allele profile interpretation

The polymorphism screening resulted in 19 markers detecting only one allele hence monomorphic marker (Figure 8) while others detected two or more alleles (Figure 9) hence resulted in 33 polymorphic markers. One of the markers ESSR62 (Figure 10) showed two very close loci which could easily confuse during allele calling. All the

markers amplified under the optimized conditions as recorded in Table 8 on page 48 and optimized high through-put co-loading volumes Table 9 on pages 52 to 54.

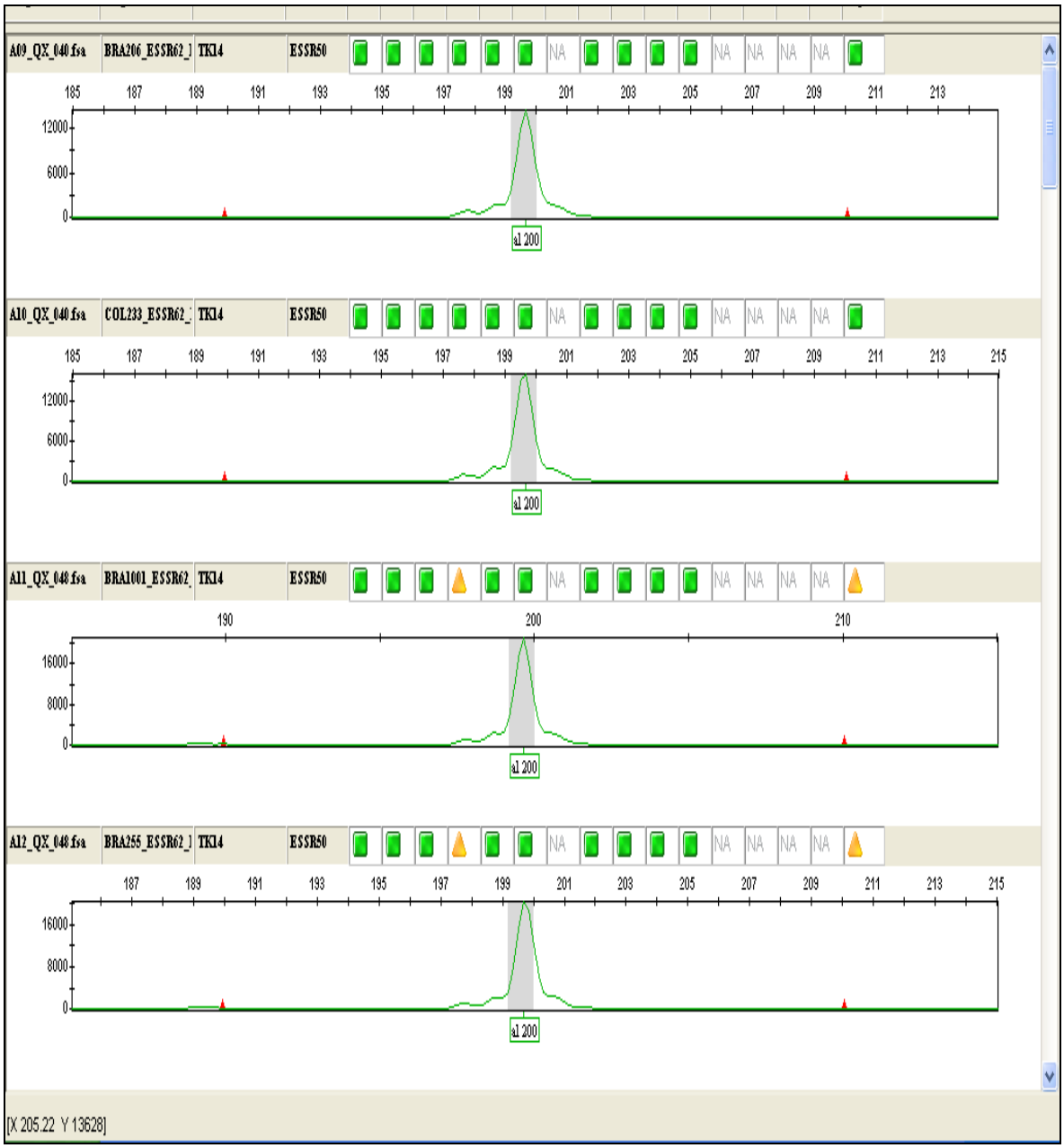

Figure 8: Electropherogram showing ESSR50 a monomorphic primer with single allele. Not a good marker for diversity studies.

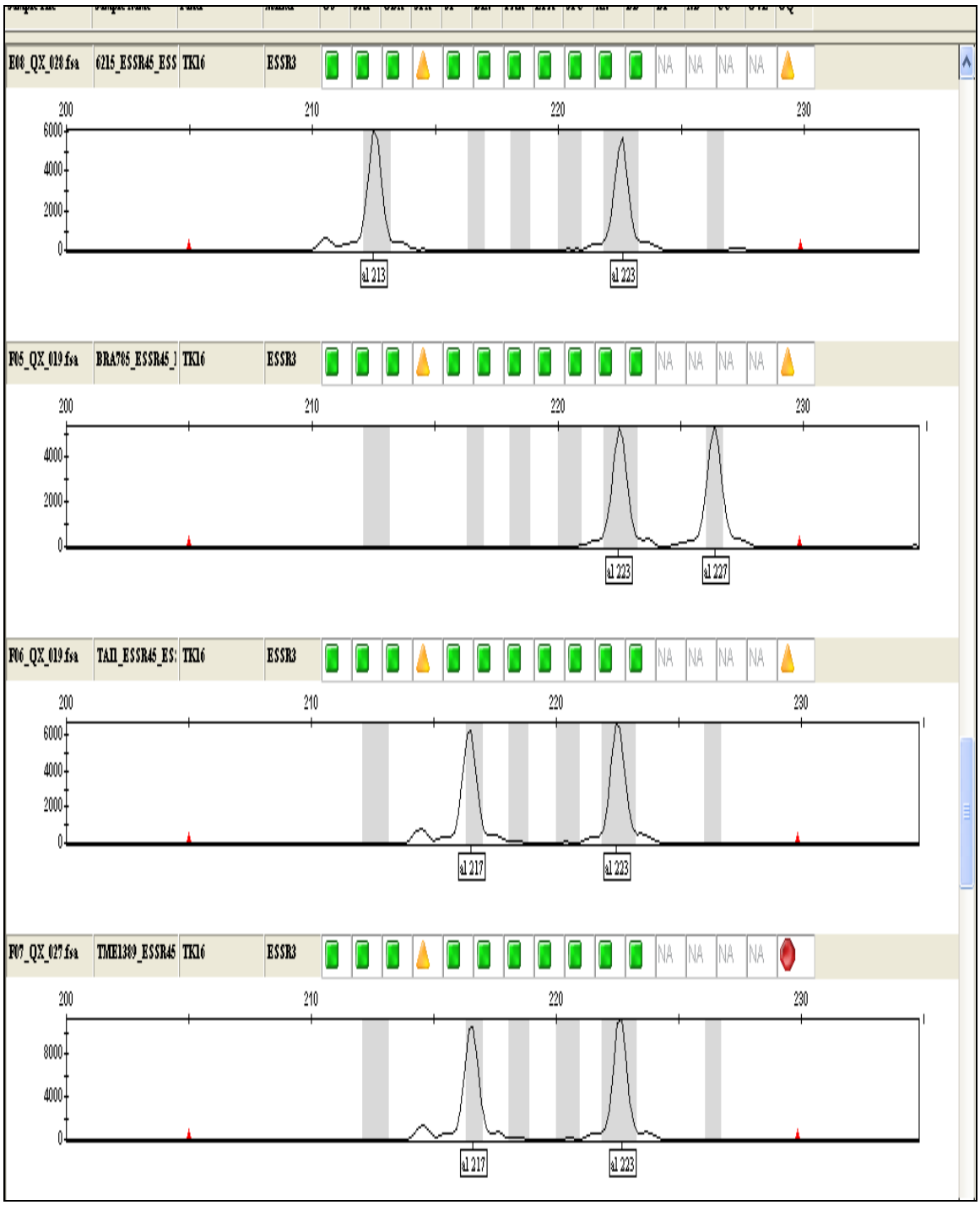

Figure 9: Electropherogram showing ESSR3 a Polymorphic primer with six alleles. Good marker for diversity studies.

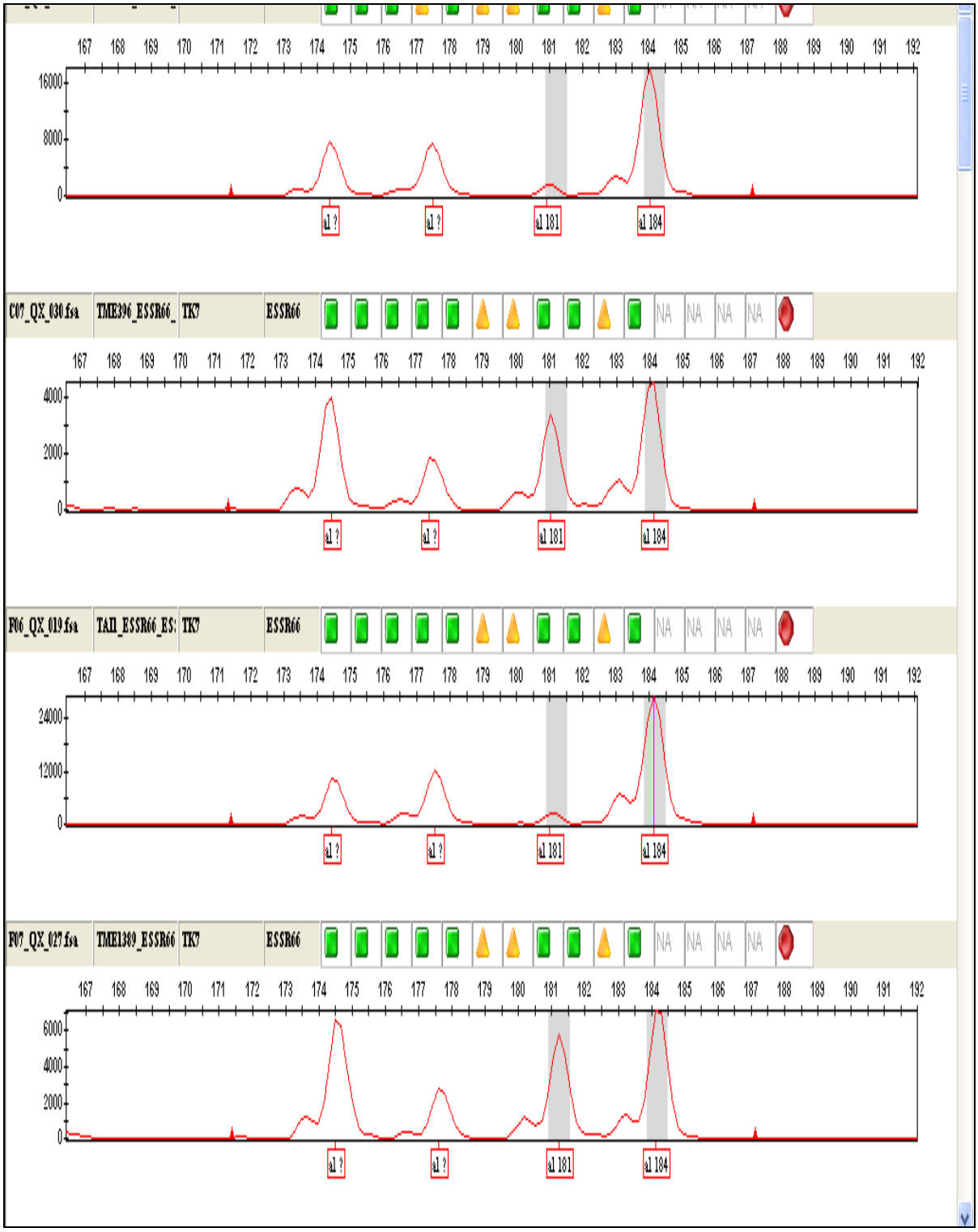

Figure 10: Electropherogram showing ESSR 66 with two close loci. Not a good marker for diversity studies

4.11 Allele profile

A total of 110 alleles were detected with the 33 EST-SSR polymorphic markers that were screened with the 32 DNA samples (Table 10). The number of alleles per marker ranged from 2 to 8 with the mean of 3.33. ESSR15 detected the most number of alleles eight, ESSR5 and ESSR53 seven alleles each, ESSR3 and ESSR8 six alleles each, ESSR26 five alleles, ESSR23 and ESSR37 four alleles each, ESSR7, ESSR11, ESSR19, ESSR25, ESSR4, ESSR13, ESSR14, ESSR30, ESSR34, ESSR35, ESSR40 and ESSR36 three alleles each and ESSR1, ESSR21, ESSR17, ESSR38, ESSR39, ESSR43, ESSR44, ESSR51, ESSR66, ESSR70 and ESSR24 two alleles each. A total of 19 EST SSR markers were monomorphic, they detected only one allele (Table 11). Fifteen markers showed polymorphism for drought resistance genotype parents and CSBD mapping population. ESSR15, ESSR30, ESSR33, ESSR5, ESSR53, ESSR19, ESSR23, ESSR26, ESSR21, ESSR25, ESSR35, ESSR39, ESSR40, ESSR7 and ESSR8 These markers have a potential for linkage mapping for a particular trait.

Table 10 Polymorphic markers: Number of alleles, Allele size, repeat type and polymorphic information content (PIC) of each Polymorphic EST-SSR markers.

SNO.	MARKER	REPEAT TYPE	Allele No.	Observed alleles size (base pairs)	PIC
1	IITA ESSR01	(AG)11	2	186,190	0.0712
2	IITA ESSR03	(CT)11	6	213,217,219,221,223,227	0.4387
3	IITA ESSR04	(CT)8	3	182,185,187	0.3099
4	IITA ESSR05	(TA)8	7	190,193,197,199,205,207,209	0.6228
5	IITA ESSR07	(CT)11	3	181,183,187	0.4678
6	IITA ESSR08	(CT)13	6	166,168,172,176,178,182	0.7381
7	IITA ESSR11	(CT)8	3	181,183,187	0.3416
8	IITA ESSR13	(TC)8	3	178,184,188	0.3994
9	IITA ESSR14	(AT)12	3	195,199,201	0.2048
10	IITA ESSR15	(TC)8	8	180,182,186,188,194,202,208,210	0.6050
11	IITA ESSR17	(TAT)6	2	179,182	0.2278
12	IITA ESSR19	(TTA)8	3	184,190,193	0.2713
13	IITA ESSR21	(AAT)6	2	181,184	0.3705
14	IITA ESSR23	(AT)6	4	212,214,216,222	0.2617
15	IITA ESSR24	(TA)6	2	174,176	0.0322
16	IITA ESSR25	(ATA)6	3	181,184,187	0.2019
17	IITA ESSR26	(ATA)10	5	184,190,193,202,205	0.6095
18	IITA ESSR27	(AT)7	3	196,204,206	0.2777
19	IITA ESSR30	(TG)6	3	186,190,192	0.4191
20	IITA ESSR33	(CT)7	2	214,216	0.3233
21	IITA ESSR34	(TA)7	3	206,208,212	0.3093
22	IITA ESSR35	(TA)7	3	208,210,212	0.4874
23	IITA ESSR36	(CT)7	3	161,173,175	0.4843
24	IITA ESSR37	(AT)6	4	177,179,181,183	0.6539
25	IITA ESSR38	(AT)6	2	193,196	0.3600
26	IITA ESSR39	(AG)6	2	186,188	0.3188
27	IITA ESSR40	(TC)6	3	176,178,180	0.5611
28	IITA ESSR43	(CT)6	2	189,191	0.0303
29	IITA ESSR44	(AT)6	2	204,206	0.3262
30	IITA ESSR51	(AT)6	2	194,196	0.3546
31	IITA ESSR53	(TC)7	7	213,217,219,221,223,227,237	0.5299
32	IITA ESSR66	(TCT)10	2	181,184	0.3457
33	IITA ESSR70	(CCA)7	2	189,195	0.3553
	Mean		**3.33**		**0.3731**

Table 11: Monomorphic primers-19 (37%)

SNO.	MARKER	REPEAT TYPE	Allele No	Observed Allele size (base pair)
1	IITA ESSR02	(AC)12	1	186
2	IITA ESSR09	(TC)9	1	176
3	IITA ESSR20	(ATT)5	1	197
4	IITA ESSR22	(TTAT)5	1	203
5	IITA ESSR29	(TC)6	1	203
6	IITA ESSR32	(AG)6	1	202
7	IITA ESSR41	(GA)6	1	200
8	IITA ESSR42	(TA)7	1	193
9	IITA ESSR45	(TC)6	1	177
10	IITA ESSR46	(AG)7	1	197
11	IITA ESSR47	(GA)6	1	181
12	IITA ESSR50	(AG)6	1	200
13	IITA ESSR56	(GCC)5	1	183
14	IITA ESSR57	(AAC)5	1	212
15	IITA ESSR58	(TGA)5	1	188
16	IITA ESSR60	(AGC)5	1	175
17	IITA ESSR61	(GCA)5	1	188
18	IITA ESSR62	(AGA)6	1	198
19	IITA ESSR65	(CTC)5	1	215

4.12 Allele frequency and Polymorphic information content (PIC)

The allele frequencies for a particular locus have influence on the PIC value. Monomorphic markers detected one allele with a frequency of 100% hence the PIC was zero. The polymorphic marker with generally evenly distributed alleles had higher PIC values. Figure 11 and 12 show the relationship between allele number and PIC value which clearly indicate that markers with more numbers of allele had higher PIC value. The effect of allele frequency can be demonstrated by marker ESSR24 (PIC 0.0322) and ESSR33 (PIC 0.3233) both with two alleles but of different frequency. PIC value is influenced by the number and frequency of alleles and is a valuable indicator of marker

polymorphism. A high PIC value indicates that the marker is highly informative and can distinguish closely related genotype. The most informative SSR markers were ESSR8, ESSR5, ESSR15, ESSR26 and ESSR37 with high PIC values of 0.7381, 0.6228, 0.6050, 0.66095 and 0.6539 respectively. The polymorphic markers indicated there was slight positive correlation tendency between the number of repeat units and number of alleles as shown in Figure 13.

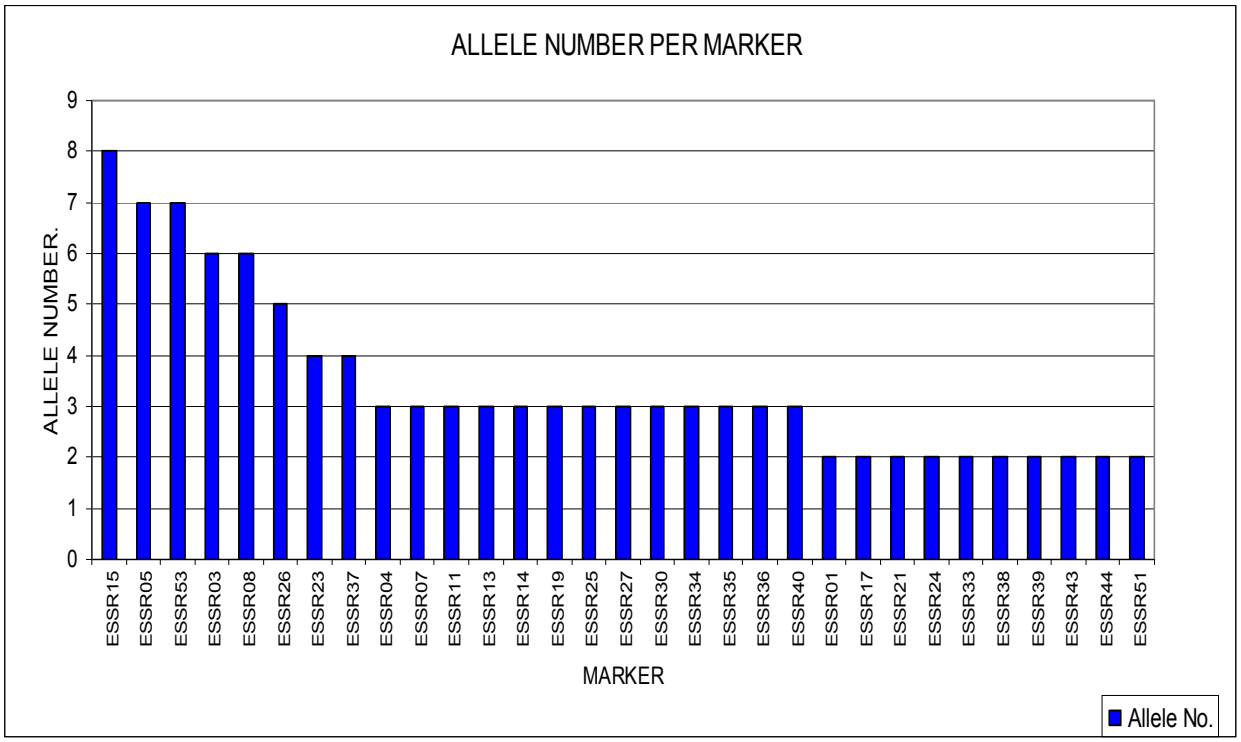

Figure 11: A graph showing allele number for each of the polymorphic primers.

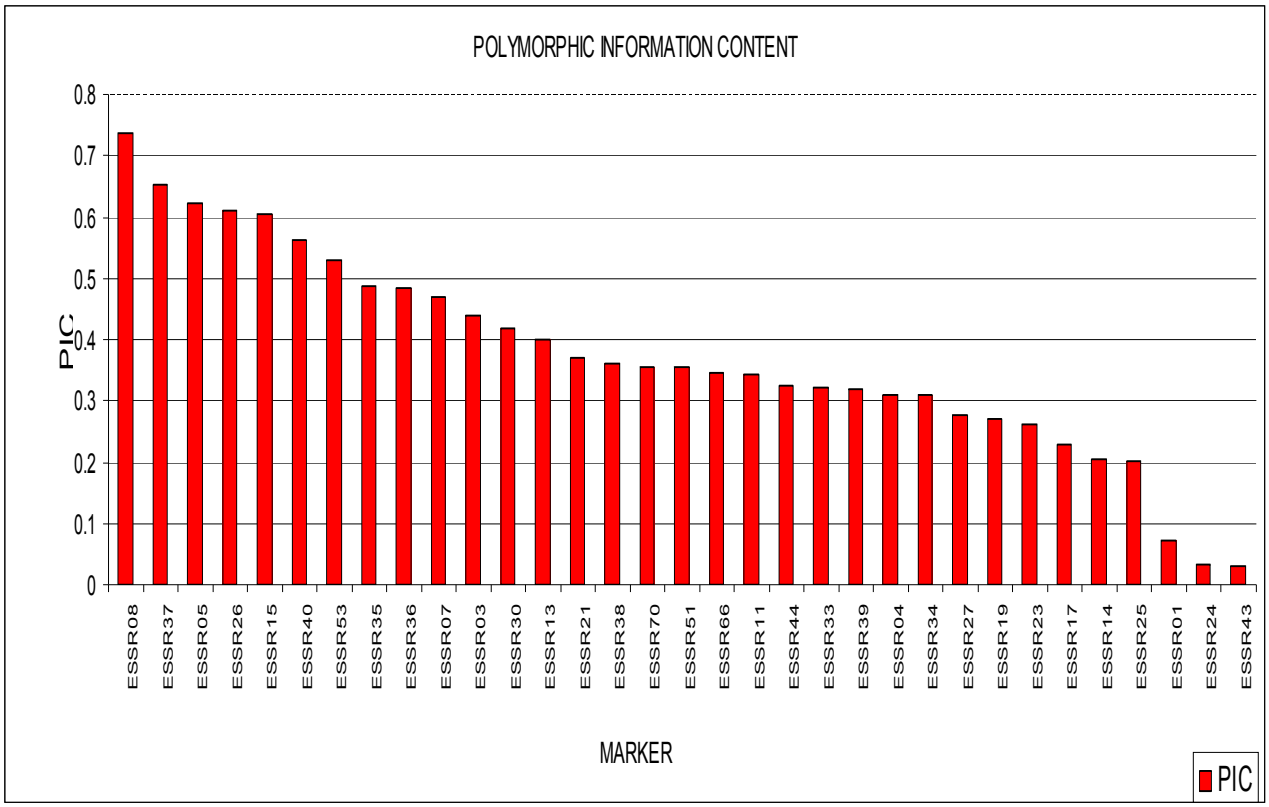

Figure 12: A graph showing polymorphic information content of the polymorphic primers.

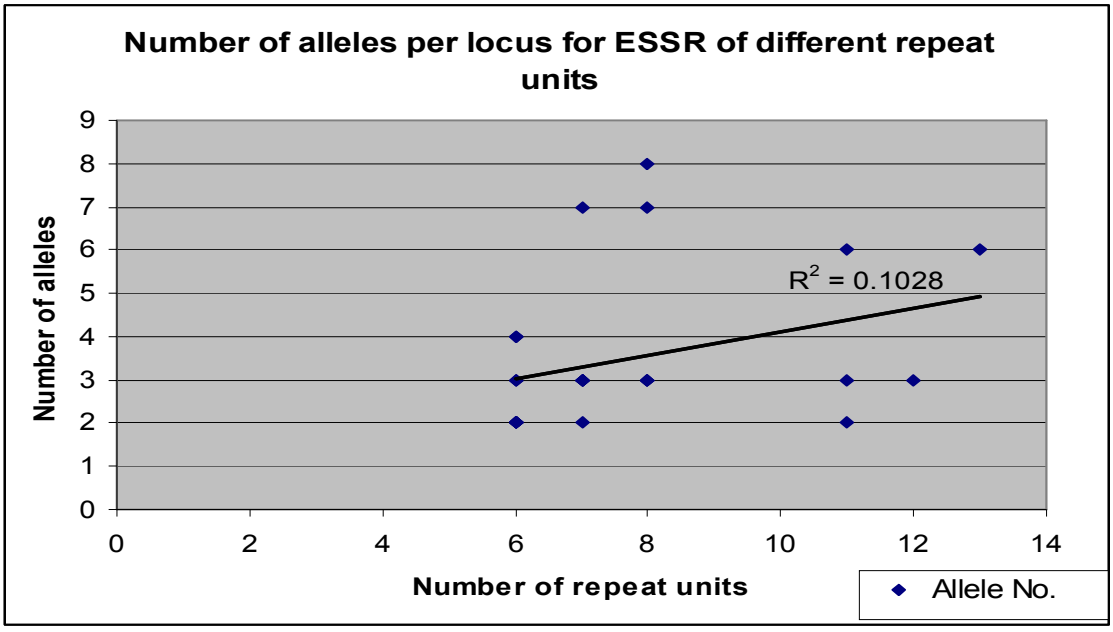

Figure 13: Number of alleles per locus for EST-SSR of different repeats units

4.13 Primer characterization

Of the 70 markers to be optimised 52, (74%) amplified successfully and gave the correct product size. Another 7 (10%) markers amplified but the product size was beyond the expected size and another 11 (16%) markers failed to amplify. Of the successful 52 markers, 19 (37%) were monomorphic while 33 (63%) markers detected more than one allele and hence were polymorphic. 78.8% of the polymorphic markers were dinucleotide repeats and the other 21.2% trinucleotide markers (Table 12). Tetranucleotide one amplified and it was monomorphic.

Table 12: Summary distribution of the seventy primers after PCR amplification.

Repeat	di	%	tri	%	tetra	%	Total	%
Successful	36	51	15	21	1	1	52	74
Not expected size	3	4	4	6	0	0	7	10
Failed	6	9	5	7	0	0	11	16
Total	45	64	24	34	1	1	70	100

Di- dinucleotide (two nucleotide repeats), Tri- trinucleotide(three nucleotide repeats), tetra- tetranucleotide(four nucleotide repeats).

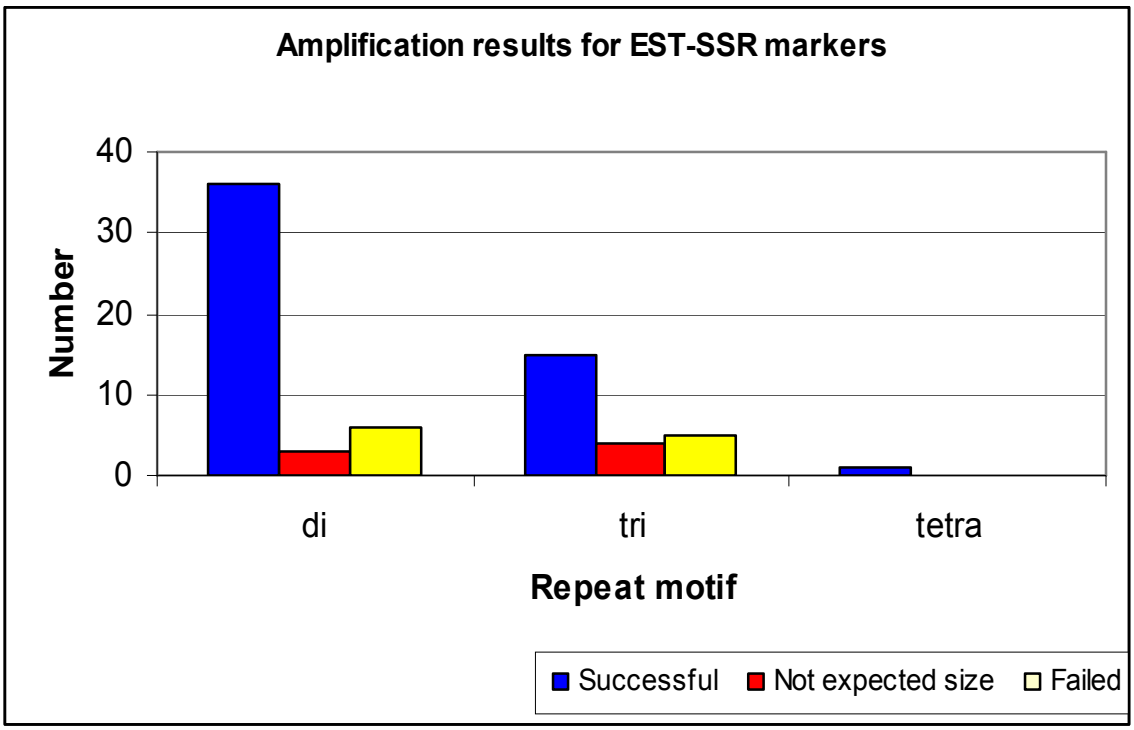

Figure 14: A graph showing distribution of the seventy primers after PCR amplification. Di- dinucleotide (two nucleotide repeats), Tri- trinucleotide (three nucleotide repeats), tetra- tetranucleotide (four nucleotide repeats).

4.14 Analysis for the optimized markers.

Figures 15 and 16 represents proportions in percentages of successfully amplified dinucleotide and trinucleotide motif repeats. On carefully checking the repeat type of the successful markers 10 (28%) dinucleotides were monomorphic and 26 (72%) were

polymorphic while 5 (53%) of trinucleotide were monomorphic and 7 (47%) were polymorphic. The only tetranucleotide present was monomorphic as shown in Table13.

Table 13 Polymorphic and monomorphic markers.

Repeat	di	%	tri	%	tetra	%	Total	%
Monomorphic	10	28	8	53	1	100	19	37
Polymorphic	26	72	7	47	0	0	33	63
Total	36	100	15	100	1	100	52	100

Di- dinucleotide, Tri- trinucleotide, tetra- tetranucleotide.

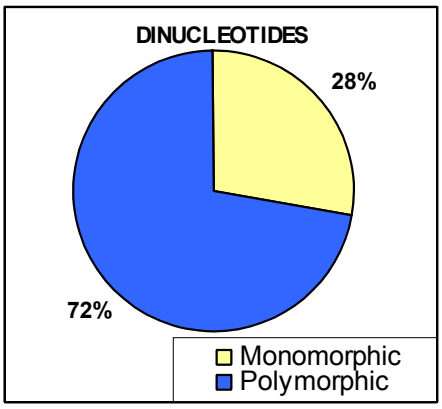

Figure 15: Dinucleotide distribution of 52 markers

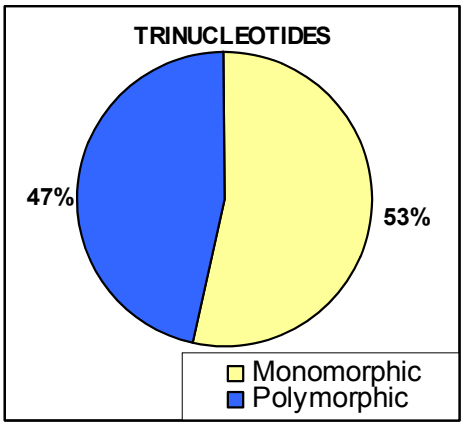

Figure 16: Trinucleotide distribution of 52 markers

4.14.1 Motif analysis for dinucleotides

The repeat type in dinucleotide with highest polymorphism was CT [12(46%)] followed by TA [11(42%) Figure 17. The rest were AG [2(8%)] and TG [1(4%)] Figure 17. A total of 10 dinucleotides were monomorphic with repeat type of AG [5(50%)], CT [3(30%)] AC [1(10%)] and AT [1(10%)] Figure 18. A total of six dinucleotides failed to amplify with CT [3(50%)], AT [29(33%)] and AG [1(17%)]. Finally 3 dinucleotides repeat type failed to amplify with CT [2(67%)] and AG [1(33%)], of the 20 CT repeats primers 60% were polymorphic, 15% were monomorphic, 15% failed to amplify and

10% amplified but the product size was not the expected. Of the 14 TA repeat primers 79% were polymorphic, 14% failed to amplify and 7% were monomorphic. Of the 9 AG repeat primers 59% were monomorphic, 22% were polymorphic, 11% failed to amplify while the other 11% amplified but the product size was not the expected.

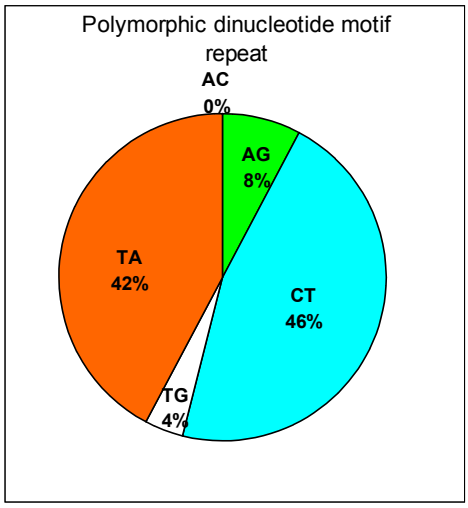

Figure 17: Polymorphic dinucleotide motif repeats.

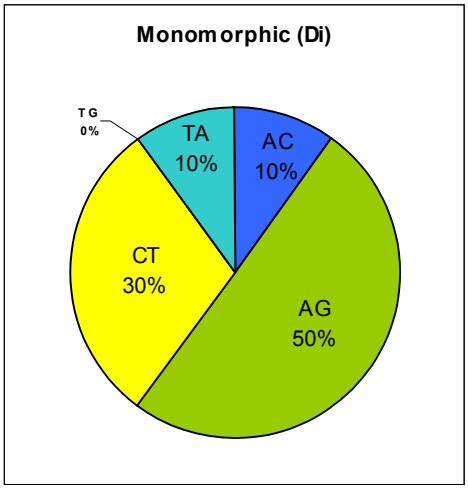

Figure 18: Monomorphic dinucleotide motif repeats

4.14.2 Motif analysis for trinucleotides.

A total of 7 trinucleotides repeats markers were polymorphic with ATA [5(72%)], CCA [1(14%)] and TCT [1(14%)]. A total of 8 trinucleotides repeats markers were monomorphic with AGA [2(25%)], ACC [1(12.5%)], GCC [1(12.5%)], CTC [1(12.5%)], AGC [1(12.5%)], ATT [1(12.5%)] and GCA [1(12.5%)].

4.14.3 Primers with unexpected product size

A total of three dinucleotide repeat primers amplified but the product size was not the expected size, one of AG repeat, two of TC repeats while four trinucleotides repeat primers amplified but the product size was not the expected. AGC, TCT, TGA and TGC each with one (Table 14)

Table 14 Markers of unexpected product size.

SNO.	MARKER	REPEAT TYPE	EXPECTED SIZE	OBSERVED SIZE
1	IITA ESSR16	(AGC)8	192	380
2	IITA ESSR31	(TC)6	177	320
3	IITA ESSR48	(TC)6	197	284
4	IITA ESSR52	(AG)6	220	746
5	IITA ESSR59	(TCT)7	216	270
6	IITA ESSR64	(TGC)5	212	414
7	IITA ESSR69	(TGA)5	210	318

4.14.4 Primers that failed to amplify

A total of six dinucleotide repeat primers failed to amplify one of AG, two of AT repeat and three of CT repeats while five trinucleotides repeats primers failed to amplify AGA, AGC, ATT, CAT and CCT each with one (Table 15). The failure of amplification could be attributed to presence of intron(s).

Table 15: Markers that failed to amplify

SNO.	MARKER	REPEAT TYPE	EXPECTED SIZE
1	IITA ESSR6	(CT)8	152
2	IITA ESSRI0	(CT)11	189
3	IITA ESSR12	(TA)8	154
4	IITA ESSR18	(ATT)5	193
5	IITA ESSR28	(CT)6	192
6	IITA ESSR49	(AG)7	186
7	IITA ESSR54	(AT)6	199
8	IITA ESSR55	(AGC)6	174
9	IITA ESSR63	(CAT)5	175
10	IITA ESSR67	(AGA)5	183
11	IITA ESSR68	(CTT)7	222

4.15 Heterozygosity

Nei's unbiased estimate of gene diversity (*He*) is independent of sample size and takes into account the number of alleles and their frequencies and thus it is a good measure of genetic diversity. The levels of observed heterozygosity (*Ho*) and the Nei's unbiased estimate of gene diversity are shown in Table 16. The value of unbiased heterozygosity (*He)* varied from 0.0306 to 0.7730 with average of 0.4292, while direct count heterozygosity was from 0.000 to 0.7778 with average of 0.4061. The smaller the value the least the diversity while the larger value indicates the genotypes are highly diverse.

Table 16 Number of allele, Nei's unbiased estimate of gene diversity (*He*) and observed Heterozygosity (*Ho*) per locus.

Locus	Number of Allele	Nei's unbiased estimate of gene diversity (He)	Observed Heterozygosity (Ho)
ESSR01	2.0	0.0740	0.0000
ESSR03	6.0	0.4697	0.5172
ESSR04	3.0	0.3605	0.4091
ESSR05	7.0	0.6570	0.6207
ESSR07	3.0	0.5576	0.4800
ESSR08	6.0	0.7730	0.7778
ESSR11	3.0	0.4150	0.5000
ESSR13	3.0	0.5137	0.3333
ESSR14	3.0	0.2182	0.2414
ESSR15	8.0	0.6362	0.5938
ESSR17	2.0	0.2622	0.3103
ESSR19	3.0	0.3012	0.2903
ESSR21	2.0	0.4911	0.5333
ESSR23	4.0	0.2872	0.2667
ESSR24	2.0	0.0328	0.0333
ESSR25	3.0	0.2128	0.2333
ESSR26	5.0	0.6632	0.5455
ESSR27	3.0	0.3173	0.1538
ESSR30	3.0	0.4709	0.4138
ESSR33	2.0	0.4055	0.5652
ESSR34	3.0	0.3657	0.3438
ESSR35	3.0	0.5512	0.5417
ESSR36	3.0	0.5559	0.4138
ESSR37	4.0	0.7054	0.6429
ESSR38	2.0	0.4709	0.6207
ESSR39	2.0	0.3980	0.2903
ESSR40	3.0	0.6326	0.2069
ESSR43	2.0	0.0308	0.0313
ESSR44	2.0	0.4105	0.3462
ESSR51	2.0	0.4608	0.4000
ESSR53	7.0	0.5571	0.6250
ESSR66	2.0	0.4444	0.5333
ESSR70	2.0	0.4620	0.5862
Mean	3.3	0.4292	0.4061
Total	110.0		

4.16 Genetic diversity

A total number of 94 alleles were detected from 21 samples in South American region and 83 from 11 samples in African region Table 17. Among countries in South America Brazil had the highest number of alleles 89 while others Colombia (69), Guatemala (58), USA (44), Venezuela (42), Costa Rica (40), Thailand (35), Paraguay (35), and Argentina (31). In Africa Tanzania had the highest number of alleles with (65), Togo (62), Ghana (54), Nigeria (49), Liberia (49) and Uganda (42). The country with the highest number of alleles indicates greater diversity among its genotypes while the one with least number of alleles indicates low levels of diversity among its genotypes.

Table 17: Number of alleles per locus in South America and Africa

Marker	AFRICA	SOUTH AMERICA
ESSR01	1	2
ESSR03	3	3
ESSR04	2	3
ESSR05	5	6
ESSR07	2	3
ESSR08	5	5
ESSR11	2	3
ESSR13	3	2
ESSR14	2	3
ESSR15	5	6
ESSR17	2	2
ESSR19	2	3
ESSR21	2	2
ESSR23	2	4
ESSR24	2	1
ESSR25	2	3
ESSR26	4	5
ESSR27	3	2
ESSR30	3	3
ESSR33	2	2
ESSR34	1	3
ESSR35	3	3
ESSR36	3	3
ESSR37	3	4
ESSR38	2	2
ESSR39	2	2
ESSR40	2	3
ESSR43	1	2
ESSR44	2	2
ESSR51	2	2
ESSR53	4	7
ESSR66	2	2
ESSR70	2	2
Mean	2.5	2.8
No. of individuals	11	21
Total	83	94

4.17 Genetic distances and phylogenetic analysis

The genetic distances were computed according to Euclidian distance computation using Powermarker V3.0 software (Appendix 4). The 32 genotypes were used in cluster analysis for cassava, a dendrogram of all the genotypes generally showed some evidence of structuring or grouping together of genotypes according to their geographic origins (Figure 19). The lowest computed distance was between genotype TME1368 from Ghana and 6219 from Tanzania with a value of 0.1951 and the highest was between genotype COL233 from Colombia and 6218 from Tanzania with a value of 0.8649. The thirty two genotypes grouped into clusters that indicate their geographical origin with those of South America and Africa grouping together in distinct clusters while a few from South America region clustered together with Africa genotypes.

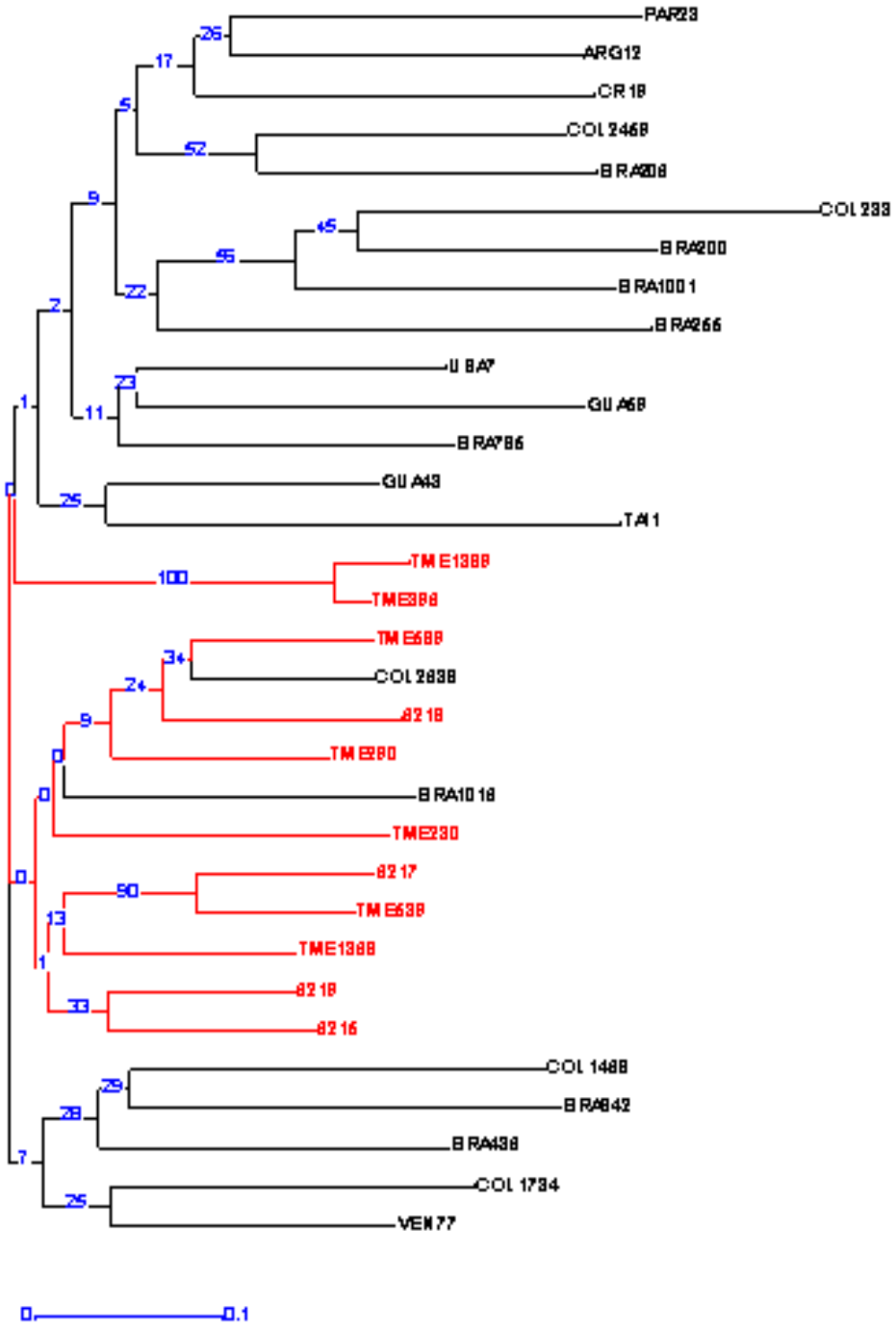

Figure 19: UPGMA tree for South America (black) and Africa genotypes (red).

CHAPTER FIVE

5.0 DISCUSSION

5.1 PCR optimisation

Optimization of PCR is a critical precursor for the accurate and robust mass screening with markers, particularly to post-PCR co-load markers to reduce the unit cost of high-throughput genotyping (Beaulieu *et al*, 2001). Optimization is often a balance between producing as much product as possible and under producing nonspecific, background amplifications. Since PCR is sensitive to a number of parameters including magnesium, template DNA, primer concentration, and annealing temperature during amplification reaction conditions needed to be optimized to avoid nonspecific amplification products such as primer-dimers or fragments of heterogeneous size (Blanchard *et al*., 1993). The annealing temperature of 62^{o}C falls within the range of 55 to 72°C generally that yields best results (Innis and Gelfand, 1990). Similarly, conditions to detect products on the ABI-3730 need to be optimized to improve the efficiency of microsatellite analysis. The DNA marker screening of many genotypes with several markers required an effective strategy which should be cost effective and time saving. The first strategy was to optimize PCR conditions for each primer and the second strategy was to establish a co-loading set whereby several markers were genotyped and analyzed simultaneously on the automated capillary electrophoresis system (ABI-3730). The volume of PCR products co-loaded depended on the efficiency of amplification, product size and the dye label. This study provides an optimum marker conditions for cassava assessment. These conditions have proved to be optimum and reproducible during the genotyping of 32

DNA samples. Thus this study has provided a protocol for future genotyping work involving the polymorphic EST-SSR in cassava studies

5.2 Magnesium chloride concentration

Magnesium chloride is an essential co-factor for the DNA polymerase in PCR (Innis and Gelfand, 1990, Blanchard *et al*., 1993) and its concentration must be optimized for every primer template system. Many components of the reaction bind magnesium ion, including primers, template, PCR products and dNTPs (Innis and Gelfand, 1990). The Mg ions binds tightly to the phosphate sugar backbone of nucleotides and nucleic acids, and variation in the MgCl2 concentration has strong effects on nucleic acid interactions. Variations in MgCl2 concentration below 4 mM improves performance of PCR by affecting specificity (higher concentration lower the specificity, lower concentrations raise specificity) (Blanchard *et al*., 1993). It is necessary for free magnesium ion to serve as an enzyme co-factor in PCR, the total magnesium ion concentration must exceed the total dNTP concentration (Saiki, 1989). Typically, to start the optimization process, 1.5 mM magnesium chloride is added to PCR in the presence of 0.8 mM total dNTPs. This leaves about 0.7 mM free magnesium for the DNA polymerase. In general, magnesium ion should be varied in a concentration series from 1.0–3.0 mM in 0.5 mM steps. The optimum magnesium chloride ($MgCl_2$) concentration in the reaction mix were 2.0 mM on all primers, however the A+ peak were observed on others that necessitated varying magnesium concentrations. Lower $MgCl_2$ concentration (1.0 mM) yielded low visible bands but with no A+ peaks and higher magnesium concentration (2.5 and 3.0 mM) yielded adequate but somewhat less amplification product. The importance of optimum magnesium concentration for PCR is well recognized (Innis and Gelfand, 1990).

5.3 Primer characterization

During this study 52 markers were screened across twenty four diverse genotype from South America and Africa including eight parents for drought resistance and CSBD genotypes. Out of these 33 (63%) primers were identified as polymorphic and 19 (37%) were monomorphic. The average PIC value for all 33 polymorphic SSR marker loci was 0.3731, ranging from 0.0303 to 0.7381. The markers that had high number of alleles generally had high PIC values indicating that some markers are more useful for differentiating between closely related genotypes than others. These results agree with the suggestions made by Buchanan *et al*. (1994) that loci with a large number of different alleles may have high PIC values, but if one or two alleles dominate, then the PIC value may still be relatively small. The effect of allele frequency was demonstrated by marker ESSR24 (PIC 0.0322) and ESSR33 (PIC 0.3233) both with two alleles but of different frequency. PIC value is influenced by the number and frequency of alleles and is a valuable indicator of marker polymorphism (Varshney *et al.,* 2005). A high PIC value indicates that the marker is highly informative and can distinguish closely related genotype. The most informative SSR markers were ESSR8, ESSR5, ESSR15, ESSR26 and ESSR37 with high PIC values of 0.7381, 0.6228, 0.6050, 0.66095 and 0.6539 respectively. Of the five most polymorphic markers four were dinucleotide of [TA (2)] and [TC (2)] motif repeat and one trinucleotide of (ATA) repeat. This agrees with other studies that the dinucleotide repeats of (AT/TA) repeat types are the most predominant, followed by (GA/CT)n/AG/TC) repeats (Powell *et al*.1996; Morgante and Olivieri 1993). Considering the SSR classes and motifs, the dinucleotide SSRs showed higher allele numbers (average 3.5 per locus) and PIC values (average of 0.371 per marker).

Based on this study the relationship of repeat unit length with number of alleles indicated, there was slight positive correlation tendency between the number of repeat units and number of alleles. It has been reported earlier in other studies that the degree of polymorphism increases with the total length of the repeat (Akkaya *et al.,* 1992; Hüttel *et al.,* 1999 Becher *et al.,* .2000; Ferguson *et al.,* 2004). Some other studies showed no relationship or weak correlation between SSR polymorphism and repeat unit length (Love *et al.,* 1990; Gupta and Varshney, 2000; He *et al.,* 2003). From this study TA (21%) repeats and TC (23%) were polymorphic, while trinucleotide of ATA (1.9%) repeats showed more polymorphism among other repeats that had fairly equal percentages. In general 72% of dinucleotide repeats showed polymorphism against 47% of trinucleotide.

5.4 Genetic Relationships

Polymorphic markers from this study were used to quantify the genetic relationship or relatedness of cassava genotype among the 32 DNA genotypes, a dendrogram was constructed by the UPGMA Method on the basis of Euclidian genetic distances. The phylogenetic analysis showed that in general genotypes from Africa were more closely related to each other than those from South America resulting in some clustering according to continent of origin. The clustering together of genotypes from different regions suggests that there has been genetic flow between the regions over time or that cassava has been introduced to Africa relatively recently and that divergence caused by mutation and selection has not had time to accumulate. Other sources of genetic differentiation could be selection for adaptation to agro-ecologies particularly disease, found in Africa, mutation and even bias in sampling. The clustering of EST-SSR

markers agrees with the studies conducted by (Mba *et al.,* 2001) for global cassava germplasm diversity using genomic SSR markers that confirmed genetic diversity structured by region, with Africa clustering from the rest of the world. The mixed cluster of genotypes could also be as result of close relatedness of the African genotypes which are improved cultivars by IITA from South American genotypes**.**

5.5 CONCLUSION

The newly identified polymorphic markers are novel as they don't match after BLAST with already existing markers in the database that are currently in use by researchers. The markers used were found to be informative in the cassava germplasm studied. ESSR 15 was the most informative SSR marker as it was observed to have the highest number of alleles and PIC value. In this study, the markers were able to cluster the cassava genotypes from Africa and those from the rest of the world. Through this study there has been increase of the number of EST-SSR markers that reveal polymorphism in cassava and the conditions have proved to be optimum and reproducible during the genotyping of 32 DNA samples. Thus this study provides protocol for future genotyping work involving cassava EST-SSR markers.

5.6 RECOMMENDATIONS

This study has presented a protocol for future genotyping work involving the 33 cassava EST markers. Following recommendations may be made:

(i) A genetic linkage mapping should be developed for the following markers that showed polymorphism for drought resistance and cassava brown streak disease genotype parents (ESSR15, ESSR30, ESSR33, ESSR5, ESSR53, ESSR19, ESSR23, ESSR26, ESSR21, ESSR25, ESSR35, ESSR39, ESSR40, ESSR7 and ESSR8)

(ii) The usefulness of these EST-SSRs lies in their application in genetic mapping, diversity assessments, marker assisted selection and genomic research and I highly recommend researchers to include them in their research activities.

REFERENCES

Akkaya, M.S., Bhagwat, A.A. and Cregan, P.B. (1992) Length polymorphism of simple sequence repeat DNA in soybean *Genetics*132**:**1131-1139.

Allen, A.C. (2002). The origins and taxonomy of cassava. In: Hillocks R.J.,Tresh J.M., Belloti A.C (Eds), Cassava: Biology, Production and Utilisation.CABI Publishing Oxon, UK and New York, USA, pp. 1-16.

Anderson, J.R., and Lubberstedt, T. (2003). Functional markers in plants, *Trends Plant Science* 8:554–560.

Asher, C.J Edwards, D.G and Howeler R.H. (1980) Nutritional disorders of cassava (*Manihot esculenta cruntz*). University of Queenland. St lucia, Queenland, Australia.

Ayad, W.G., Hodgkin, A., Jaradat and V.R. Rao, editors. (1997). Molecular genetic techniques for plant genetic resources. Report of and IPGRI workshop, 9-11 October 1995, Rome, Italy. International Plant Genetic Resources Institute, Rome Italy. Extraordinarily polymorphic microsatellites DNA in barley: Species diversity, chromosomal locations and population dynamics. *Proc Natl. Acad. Sci. (USA)* 91:5466-5470.*genetic. Plant J.* 3: 175-182.*genetic Plant J.* 3: 175-182.

Baguma, Y. (2004). Regulation of starch synthesis in cassava. Doctoral thesis. Swedish

Beaulieu, M., Larson, G.P., Geller, L., Flanagan, S.D and Krontiris, T.G., (2001). PCR candidate region mismatch scanning: adoption to quantitative, high-throughput genetyping. *Nucleic Acid Research* 29:1114-1124.

Becher, S.A., Steinmetz, K., Weising, K., Boury, S., Peltier, D., Renou, J.P., Kahl, G. and Wolff, K. (2000). Microsatellites for cultivar identification in Pelargonium. *Theoretical and Application Genetics* 101**:** 643-651.

Bhat, P.R., Krishnakumar, V., Hendre, P.S, Rajendrakumar, P., Varshney, R.K., and Aggarwal, R.K. (2005). Identification and characterization of expressed sequence tags-derived simple sequence repeats, markers from robusta coffee variety 'C ×R' (an interspecific hybrid of *Coffea canephora* × *Coffea congensis). Molecular Ecolology* Notes 5: 80–83.

Blanchard, M.M., P. Tailon-Miller, P. Nowotny and V. Nowotny, 1993. PCR buffer optimization with a uniform temperature regimen to facilitate automation. *PCR Methods Applications.*, 2: 234–40.

Boguski, M.S., Lowe, T.M.J. and Tolstoshev, C.M. (1993). dbEST-database for expressed sequence tags. *Nature Genetics* 4: 332–333.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980). Construction of genetic linkage map in man using restriction fragment length polymorphisms. *American Journal Human Genetics* 32: 314-331.

Boventius, H. and Weller, J. (1994). Mapping analysis of dairy cattle quantative trait loci maximum likelihood methodology using milk protein genes as genetic markers. *Genetics* 137:267-280.

Buchanan, F.C., Adams, L.J., Littlejohn, R.P., Maddox, J.F. and Crawford, A.M. (1994). Determination of evolutionary relationships among sheep breeds using microsatellites. *Genomics* 22: 397-403.

Buhariwalla, H.K. and Crouch, J.H. (2004). Optimization of marker screening protocol to assess the degree and distribution of genetic diversity in landraces of pigeonpea. *In*: Bramel, P. J. (2004). *Assessing the Risk of losses of Biodiversity in Traditional Cropping Systems: A case study of Pigeonpea in Andra Pradesh.* International Crops Research Institute for the Semi-Arid tropics Pantacheru, 502 324, Andhra Pradesh, India. Pg 168.

Canadian Coalition to End Global Poverty (CCIC) (2004). Recent Trends in Canadian Aid to Sub-Saharan Africa.http://www.ccic.ca/e/docs/003_acf_2004 10_subsaharan_africa_aid_trends.pdf).

Casey, R.M. (2005). BLAST Sequences Aid in Genomics and Proteomics. *Business Intelligence Network*. http://www.b-eye-network.com/view/1730.

Chavarrianga-Aguirre, P., Maya, M.M., Tohme, J., Duque, M.C., Iglesius, C. and Bonierbale, M.W. (1999). Microsatellite, isoenzymes and redundancy in the cassava core collection and to assess the usefulness of DNA markers to maintain germplasm collections. *Molecular Breeding* 5:263-273.

De Tafur, S.M., El-Sharkawy, M.A. and Caller, F. (1997). Photosynthesis and yield perfomance of cassava in seasonally dry and semi arid environments. *Photosynthesis.* 33: 229-257.

Debener, T., Salamini, F. and Gebhardt, C. (1990). Phylogeny of wild and cultivated *Solanum* species based on nuclear restriction fragment length polymorphisms (RFLPs). *Theoretical and Applied Genetics* 79**:** 360-368.

Dellaporta, S.L., Wood, J. and Hicks, J.B. (1983). A plant DNA mini preparation:Version II. *Plant Molecular Biology* 1:19-21

Elias, M., McKey, D., Panaud, O., Anstett, M.-C., Robert, T. (2001). Traditional management of cassava morphological and genetic diversity by the Makushi Amerindians (Guyana, South America): Perspectives for on farm conservation of crop genetic resources. Euphytica, 120, 143-157.

Elias, M., Panaud, O. and Robert (2000). Assessment of genetic variability in a traditional cassava (*Manihot esculenta Crantz*) farming system, using AFLP markers. *Heredity* 85:219-230.

Elnifro, E. M., Ashshi, A. M., Cooper, R. J. and Klapper, P. E. (2000) Multiplex PCR: optimization and application in diagnostic virology. *Clinical Microbiology*. Reviews.13, 559–570.

Eujayl, I., Sorrells M.E., Baum M., Wolters P., Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics* 104:339–407.

Eujayl, I., Sledge M.K, Wang, L., May, G.D., Chekhovskiy, K., Zwonitzer, J.C and Medicago M.A. (2004). *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theoretical and Applied Genetics* 108: 414–422.

FAO (2002). Partnership formed to improve cassava, staple food of 600 million people.http://www.fao.org/english/newsroom/news/2002/10541-en.html Cited on 17/10/2007.

Federal Agriculture Coordinating Unit (FACU) (1986). Niger State Agricultural Development Project II, Phase I 1988-1991. Kaduna, Nigeria: Federal Agriculture Coordinating Unit.

Ferguson, M.E., Burow, M.D., Schultz, S.R, Bramel, P.J., Paterson, A.H., Kresovich, S.and Mitchell, S. (2004). Microsatellite identification and characterization in peanut (*A.hypogaea L.*)**.** *Theoretical and Applied Genetics* 108**:**1064- 1070.

Fregene, M., Angel, F., Gomez, R., Rodriguez, F., Chavarrianga, P., Roca,W.,Tohme, J. and Bonierbale M. (1997). A molecular genetic map of cassava *(Manihot esculenta Crantz*). *Theoretical and Applied Genetics* 95:431-441.

Gao, L.F., Jing, R.L., Huo., N.X., Li., Y. and Li., X.P. (2004), One hundred and one new microsatellite loci derived from ESTs (EST-SSRs) in bread wheat, *Theoretical Applied Genetics* 108 pp. 1392–1400.

Generation Challege Programme, (2006). Generation challege programme competitive and commissioned research project mid-year reports.pp 147-148. http://www.generationcp.org/arm/ARM06/General/2006_midyear_reports.pdf

Gibbons, A. (1990). New view of early Amazonia. Science 248. 1488 – 1490.

Gichuki, S.T., Berenyi, M., Zhang, D., Hermann, M., Schmidt, J., Glössl, J. and Burg, K. (2003). Genetic diversity in sweet potato [*Ipomoea batats (L.) Lam*.] in relationship to geographic sources as assessed with RAPD markers. *Genetic Resources and Crop Evolution* 50: 429-437.

Guo, X., Elston, R.C. (1999) Linkage information content of polymorphic genetic markers. *Human Heredity* 49:112–118.

Gupta, P.K. and Varshney, R.K. (2000). The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113:163-185.

Gupta, P.K., Varshney, R.K, Sharma,P.C. and Ramesh, B. (2000). Molecular markers and their application in wheat breeding. *Plant Breeding* 118:369-390.

Gupta, P.K., Rustgi, S., Sharma, S., Singh, R., Kumar, N., Balyan, H.S.(2003). Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetic Genome* 270: 315–323.

Han, Z.G., Guo, W.Z., Song, X.L., Zhang, T.Z. (2004). Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. *Molecular Genetics and Genomics* 272, 308–327.

He, G., Meng, R., Newman, M., Gao, G.M., Pittman R.N. and Prakash, C.S. (2003). Microsetellites as DNA markers in cultivated peanut (*Arachis hypogaea L.*) *BMC Plant Biology* 3**:**3.

Hernan Ceballos, C. A. Iglesias, J. C. Perez, and A.G.O.Dixon. (2004). Cassava breeding: opportunities and challanges. *Plant molecular biology* 56: 506 – 516, 2004.

Howeler R.H. (1998) Cassava agronomy research in Asia –an overview 1993-1996. in Howeler R.H (ed) cassva Breeding agronomy and farmer participatory research in Asia. Proceedings 5th Regional workshop held in Danzhou, Hainan Vhaina. Nov3- 8 pp 355-375.

Hüttel, B., Winter, P., Weising, K., Choumane, W., Weigand, F. and Kahl, G. (1999).Sequence-tagged microsatellite-site markers for chickpea (*Cicer arietinum L.*)**.** *Genome* 42**:**210-217.

Iglesias, C.A., Mayer, J., Chávez, A.L., and Calle, F. (1997). Genetic potential and stability of carotene content in cassava roots. *Euphytica* 94: 367 – 373.

Innis, M.A and Gelfand D.H (1990). Optimization of PCR's. In: Innis, M.A, Gelfend, D.H, Sninsky, J.J. and White, T.J (eds), PCR Protocols: A Guide to Methods and Applications, pp 3-12. Academic Press, New York.

Islam, A.K.M.S., Edwards, D.G and Asher, C.J (1980) PH optima for crop growth: results of flowing culture experiment with six species. *Plant and soil* 54(3), 339-357.

Jameson, J.D and Thomas, D.G, (1970). Cassava. In: Jameson, D.G (ed). Agriculture in Uganda. Oxford University press, Oxford, pp. 247 – 251.

Jaramillo, G, N., Morante, J. C., Pérez, F., Calle, H., Ceballos, B., Arias and Bellotti. A. C. (2005). Diallel Analysis in Cassava Adapted to the Midaltitude Valleys Environment Crop Science 45:1058-1063.

Jennings, D. L. (1994). Breeding for resistance to African Cassava mosaic geminivirus in East Africa. Tropical Science Journal 1994. Volume 34. 110-122

Jennings, D.L. (1976). Cassava *Manihot esculenta* (Euphorbiacae). In: Simmons, N. (ed) *Evolution of crop plants*. Longman, London, pp. 81-84.

Jones, C.J., Edwards, K.J., Castiglione, S., Winfield, M.O., Sala, F. van de Wiel, C. Bredmeijer, G.Vosman, B., Matthes, M., Daly, A., Brettschneider, R., Bettini, P., Buiatti, M., Maestri, E., Malcevschi, A., Marmiroli, N., Aert, R., Volkaert, G., Rueda, J., Linacero, R., Vazquez, A. and Karp A. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding* 3: 381-390.

Jongeneel, C.V. (2000). Searching the expressed sequence tag (EST) databases: Panning for genes. *Briefings in Bioinformatics* 1:76-92.

Jos, J.S. (1969). Cytological aspects of cassava. In: cassava Production Technologies, Hrish, N, Nair R.G (eds), pp 10-14. Central Tuber Research Institute,Trivandrum

Kijas, J.M.H., Fowler, J.C.S. and Thomas, M.R. (1995). An evaluation of sequence tagged microsatellite site markers for genetic analysis within *Citrus* and related species, *Genome* 38:349-355.

King, R.C., Stansfield, W.D. (1997). A dictionary of genetics. 4th ed., Oxford University press New York-Oxford, pp188.

Launaud, C. and Vincent, L. (1997). Molecular techniques for increased use of genetic resources. In; Ayad, W.G., T. Hodgkin, A. Jaradat and V.R. Rao, (eds). 1997. Molecular genetic techniques for plant genetic resources. Report of an IPGRI workshop, 9-11 October 1995, Rome, Italy. International Plant Genetic Resources Institute, Rome, Italy.

Lefebvre, V., Goffnet, B., Chauvet, J. C., Caromel, B., Signoret, P. Brand, R. and Palloix, A. (2001). Evaluation of genetic distances between pepper inbred lines for cultivar protection purposes:comparison of AFLP, RAPD and phenotypic data. *Theoretical and Applied Genetics* 102: 741 -750.

Leigh, F., Lea, V., Law, J., Wolters, P., Powell, W., Donini, P. (2003). Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. *Euphytica* 133: 359–366.

Lelley,T., Stachel, M., Grausfruber, H. and Vollmann, J. (2000). Analysis of relationships between Aegilops tuschii and D genome of wheat utilizing microsatellites. *Genome* 43: 661-668.

Love, J., Knight, A., Mc Aleer, M., Todd, J. (1990). Towards construction of a high-resolution map of the mouse genome using PCR analyzed microsatellites. *Nucleic Acids Research* 18**:** 4123-4130.

MacIntosh G.C., Wilkerson C, Green, P.J. (2001). Identification and analysis of analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. *Plant Physiology*. 127(3): 765-776.

Marmey, P., Beeching J. R., Hamon S. & Charrier A. (1994). Evaluation of cassava (*Manihot esculenta* Crantz) germplasm using RAPD markers. *Euphytica* 74, 203-209.

Mba, R.E.C., P. Stephenson, K. Edwards, S. Melzer, J. Nkumbira, U. Gullberg, K. Apel, M. Gale, J. Tohme and M. Fregene, (2001). Simple sequence repeat (SSR) markers survey of the cassava (*Manihot esculenta* Crantz) genome: towards an SSR-based molecular genetic map of cassava. *Theoretical and Applied Genetics* 102: 21–31.

Mba, R.E.C., Stephenson, P., Edwards. K., Melzer, Mkumbira, J., Gullberg, U., Apel, K., gale, M., Tohme, J. and Fregene, M. (2000). Simple sequence repeat (SSR) markers survey of the cassava (*Manihot esculenta crantz*) genome toward a SSR-based molecular genetic map of cassava. *Theoretical and Applied Genetics* 102: 21-31.

Mohan, M., Nair, S., Bhagwat, A., Krishna T.G., Yano, M., Bhatia, C.R., Sasaki T. (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding* 3: 87-103.

Morgante, M. and Olivieri A.M. (1993). PCR amplified microsatelites as markersin plant genetics. *Plant Journal* 3:175-182.

Muluvi, G. M., Sprent, J. I., Soranzo, N., Provan, J., Odee, D., Folkard, G., McNicol, J. W. and Powell, W. (1999). Amplified Fragment Length Polymorphism (AFLP) analysis of genetic variation in *Moringa Oleifera* Lam. *Molecular Ecology* 8: 463-470.

Nassar, N.M.A (2002). Cassava, *Manihoti esculenta* Crantz, genetic resources: origin of

Nei, M., (1987). Molecular evolutionary genetics. Columbia university press, New York.

NEPAD Pan African Cassava Initiative. (2004). Cassava a powerful poverty fighter. Project implementation proposal. 60pp.

Nikiforov, T.T., Rendle, R.B., Goelat, P., Rogers, Y.H., Anderson, S., Trainor, G.L. and Knapp, M.R. (1994). Genetic bit analysis a solid phase method for typing single nucleotide polymorphisms. *Nucleic acid Research* 22: 4167-4175.

Okigbo, B.N. (1980). Nutritional implication of projects giving high priority to the production of stample of low nutritive quality. The case of cassava in the humid tropics of West Africa. Food and nutrition bulletin, United Nations University, Tokyo, 2(4):1-10.

Olsen, K. and Schaal, B. (1999). Evidence of the origin of cassava: Phylogeography of *Manihot esculenta*. *Proceeding of National Academy of Science* USA 96: 5586-5591.

Olsen, K.M., and B.A. Schaal. (2001). Microsatellite variation in cassava (*Manihot esculenta,* Euphorbiacea) and its wild relatives: further evidence for a southern

Amazonian origin of domestication. *American Journal of Botany* 88(1): 131–142.

Otim-Nape, G.W. (1993). Epidemiology of the African cassava mosaic geminivirus disease in Uganda. Ph.D Thesis, University of reading, U.K.

Otto J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* 51: 283-290.

Papi, M., Sabatini, S., Altamura, M.M., Hennig, L., Schafer, E., Costantino, P., Vittorioso P.(2002).Inactivation of the Phloem-specific Dof Zinc Finger gene DAG1 affects response to light andintegrity of the Testa of Arabidopsis seeds. *Plant Physiology*, 128, 411-417.

Pashley, C.H., Ellis, J.R., McCauley, D.E. and Burke, J.M. (2006). EST databases as a source for molecular markers: lessons from Helianthus. *Heredity* 97: 381–388.

Pellet, D. and El-Sharkawy, M.A. (1997). Cassava varietal response to fertilization, growth, dynamics and implications for cropping sustainability. *Experimental Agriculture* 33: 353-365.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingy, S. and Rafalski, A. (1996). The Comparison of RFLP, RAPD, AFLP, and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding* 2:225-238.

Powell,W., Machray, G., Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends Plant Science* 1: 215–222.

Purseglove, J.W. (1968). *Tropical Crops*: Dicotyledons. Longman, London.

Puttchacharoen, S., Howeler R.H., Jantawat, S. and Vichukit, V. (1998) Nutrientuptake and soil erosion losses in cassava and six other crops in a psamment in easternThailand. *Field crops Research* 57, 113-126.

Rafalski, A. (2002). Applications of single nucleotide polymorphism in crop genetics. Current. Opinion. *Plant Biology* 5: 94-100.

Robinson, J.P. and Harris, S.A. (1999). Amplified Fragment Length Polymorphisms and Microsatellites: Aphylogenetic perspective. In: Which DNA Marker for Which Purpose? Final Compendium of the Research Project Development, Optimization for molecular tools for assessment of biodiversity in forest trees in European Union DGXII Biotechnology FW IV Research Programme Molecular Tools for Biodiversity. Gillet, E. M. (ed.) 1999. URL. http://webdoc. Sub. Gwdg.de/ebok/y/1999/which marker/index. Htm.

Rungis, D., Berube, Y., Zhang, J., Ralph, S., Ritland, C.E., Ellis, B.E., Douglas, C., Bohlmann, J., Ritland (2004). Robust simple sequence repeats markers for spruce (*Picea* spp.) from expressed sequence tags. *Theoretical and Applied Genetics* 109: 1283–1294.

Rychlik, W., Spencer, W.J. and Rhoads, R.E. (1990). Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucleic Acids Research* 21:6409-6412.

Saghai-Maroof, M.A., Bujasher, R.M., Yang, G.P., Zhang, Q. and Allard, R.W. (1994). Extraordinary polymorphic microsatellite DNA in barley: species diversity,

chromosomal locations and population dynamics. *Proceeding of National Academy of Science USA* 91:5466-5470.

Saiki, R.K., (1989). The design and optimization of the PCR. *In:* Erlich, H.A. (Eds.), *PCR Technology - Principles and Applications for DNA Amplification,* pp: 7–16. Stockton Press, New York

Saitoh, H., Ueda, S., Kurusaki, and Kiuchi, M. (1998). The different mobility of complementary strands depends on the proportion AC/GT. *Forensic Science International* 2:81-90.

Schuelke, M. (2000). An economic method for the fluorescent labelling of PCR fragments.*Nature Biotechnology* 18: 233-234.

Scott, G. J., Best, R., Rosegrant, M. and Bokanga, M. (2000). Roots and Tubers in the global food system: A vision statement to the year 2020 (including Annex). A Co-publication of International Potato Center (CIP), Centro internacional de Agricultura Tropical (CIAT), International Food Policy Research Institute (IFPRI), International Institute of Tropical Agriculture (IITA), and International Plant Genetic Resources Institute (IPGRI). Printed in Lima, Peru: International Potato Center.sequence repeat DNA in soybean. *Genetics* 132:1131-1139

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* 147: 195–197.

Sorrells, M.E. and Wilson, W.A. (1997). Direct classification and selection of superior alleles for crop improvement. *Crop Science* 37: 691–697.

Stuber, C.W., Polacco,M.,Senior,M.L.(1999). Synergy of empirical breeding, Marker-assisted selection, and genomics to increase crop yield potential. *Crop Science* 39: 1571-1583.

Suarez, M.C., Bernal, A., Gutierrez, J., Tohme, J. and Fregene, M. (2002). Development of expressed sequence tags (ESTs) from polymorphic transcription derived fragments (TDFs) in cassava (Manihot esculenta crantz). *Genome* 43:62-67.

Syvanen, A.C. (2001). Genotyping single nucleotide polymorphism. *Nature Revolution Genetics* 2:930-942.

Taguchi, G. and Konishi, S. (1987). Orthogonal Arrays and Linear Graphs , American Supplier Institute Inc., Dearborn, MI.

Tanksley, S.D. and Nelson, J.C. (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theoretical Applied Genetics* 92:191–203.

Tautz, D. (1989). Hypervariability of Simple Sequences as a general source of polymorphic DNA markers. *Nucleic Acids Resources.* 17:3286-3292.

Tautz, D., Trick, M. and Dover, G.A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322: 652-656.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. (2003) Exploiting EST databases for the development of cDNA derived microsatellite markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics* 106, 411–422.

Varshney, R.K., Andreas, G. and Sorrells, M.E. (2005). Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice, *Plant Science.* 168 pp. 195–202.

Varshney, R.K., Graner, A., Sorrells, M.E. (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotechnology* 23(1)**:**48-55.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Vande Lee, T., Hordes, M., Frijters, A., Pot, J., Peleman, J., Kuper, M. and Zabeau, M. (1995). AFLP, A new technique for DNA fingerprinting. *Nucleic Acids Research* 23: 4407-4414.

Vuylsteke, M., Mank, R., Brugmans, B., Stam, P. and Kuiper, M. (2000). Further characterization of AFLP data as a tool in genetic diversity assessments among maize (*Zea mays* L.) inbred lines. *Molecular Breeding* 6: 265-276.

Wang, Z., Weber, J. L., Zhong, G. and Tanksley, S. D. (1994). Survey of plant short tandem DNA repeats. *Theoretical Applied Genetics* 88:1-6.

Weir, B.S. (1996). Genetic data analysis II. Methods for discrete population genetic data. Sinauer Associates, Inc. Sunderland, Massachusetts. 445pp

Wenz, H.M., Robertson, J.M., Menchen, S., Oaks, F., Demorest, D.M., Scheibler, D.,Rosenblum, B.B., Wike, C., Gilbert, D.A and Efcavitch, J.W. (1998). High-precision genotyping by denaturing capillary electrophoresis. *Genome Research* 83: 69-80.

Williams, J. G.K., Kubelik, A.R., Livak, K. J., Rafalski, J. A. and Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Resources* 18:7213-7218.

Woodhead, M., Russell, J., Squirrell, J., Hollingsworth, PM., Mackenzie, K., Gibby M. and W. Powell. (2005). Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. *Molecular Ecology* 14: 1681–1695.

World Bank (2000). Can Africa Claim the 21st Century? Washington, DC: World Bank publication.

Yu., J.K. Rota., M.L. Kantety., R.V. and Sorrells, J.K. (2004) EST-derived SSR markers for comparative mapping in wheat and rice, *Molecular Genetics* 271 pp. 742–751

Zane, L., Bargelloni, L., Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Molecular Ecology* 11: 1–16.

Zhang, D. P., Cervantes, J. C., Huaman, Z., Carey, E. E. and Ghislain, M.(2000). Assessing genetic diversity of sweet potato [(*Ipomoea batatas* (L.) Lam.] cultivars from Tropical America using *AFLP. Genetic Resources Crop Evolution.* 47: 659 – 665.

Zhang, D.P., Ghislain, M., Huamán, Z., Golmirzaie, A. and Hijimans, R. (1998). RAPD Variation in Sweetpotato [*Ipomoea batatas* (L.) Lam] cultivars from South America and Papua New Guinea. *Genetic Resources Crop Evolution* 45: 271 – 277

Zhu, Y., Strassmann, J.E. and Queller, D.C. (2000). Insertions, substitutions and the origin of microsatellites. *Genetical Research* 76: 227-236.

LIST OF APPENDICES

Appendix 1 The seventy primers for optimization

MARKER	Left primer sequence	Left Tm	Right primer sequence	Right Tm	Motif Repeats	Product size
IITA ESSR1	TCTGCTCAGCTGCCCAGCCA	70.238	TCGCAGCAAACCTCTCCCCA	68.594	(AG)11	162
IITA ESSR2	TGGAAATGCTGAAAGTGAACGCTTGA	69.644	AAAACACCAGCAAAATTGCACAGGAC	68.036	(AC)12	162
IITA ESSR3	GCAACAGGTGCCCGATGTGTAGC	70.019	CAGCGGCTGCTCCCATTCCT	69.18	(CT)11	194
IITA ESSR4	TCTCTCACAGGTCGCCCAACACA	69.581	GGTCACGTCAAGTACCTGTCAAGGCA	69.41	(CT)8	163
IITA ESSR5	TGCCACAACGCCTGTGTAGAATCG	70.264	ACCCAATGGAGCCGTAACAAATTCA	68.215	(TA)8	162
IITA ESSR6	TCCACCATTTCATTCATCAAGGCCA	70.254	GGAACGATTTTCTCAACCAAAATGCGA	70.239	(CT)8	152
IITA ESSR7	AGCACTCTAATCATGCAACTCCTTCGG	68.431	GCTCAATCAGGTGCCACAGCG	68.616	(CT)11	156
IITA ESSR8	TTCTGCCGAGCACGAATATTACCCC	69.649	TCGACTTGTTTTCAAGTGCATCCCA	68.776	(CT)13	152
IITA ESSR9	CTCTAGCCTGGAGCTCGTGACGACATT	70.516	TCCCAATGTAACCAGCACCACCG	70.03	(TC)9	152
IITA ESSRI0	AGCCACCACACACACCAAACGC	69.696	TCCAGACGCTGCATTTGCCA	68.438	(CT)11	189
IITA ESSR11	GAGGAGGTTTGGGACCCTCCCTG	70.142	GGAGGGTGGCTGTGAATCCCG	70.273	(CT)8	157
IITA ESSR12	TGTCAATACACTGTCAGACACGTTCGC	68.422	GCATCGTGACTTTTCTTGATAGGCCAG	68.675	(TA)8	154
IITA ESSR13	TCAGAATTCGAGCTGAGAGTGTTGAGG	68.076	CCCTTCTCTGAGGCCAGTCCCA	69.172	(TC)8	158
IITA ESSR14	ATGGGGTTCTCACAGTGACGGTTCC	70.3	TGCTCAGAGAATCCCAAAGGCACA	69.25	(AT)12	175
IITA ESSR15	GTCAGCCGTCATCCGGCCAT	69.945	GCTTTCTCTTCAAGCCAAAAGCGTCC	69.797	(TC)8	162
IITA ESSR16	TTGCCAGCATTGATACTGCACAAGC	69.252	GGCACCTGGGGACCTGTAATCAGTC	69.789	(AGC)8	178
IITA ESSR17	CTATTGGATGTGGGCTGGCGCT	69.471	CCACTCGCATGCTCCTCAAGCA	70.183	(TAT)6	154
IITA ESSR18	CACCGGATCCCACGTGCAAGA	70.809	CAAGGGTGACGTCCACTAAATCGACA	69.055	(ATT)5	193
IITA ESSR19	ACGGTAGTGCCCTTGAGGTTGGG	69.347	CAAATGGACAACATCAACGATCACAGG	69.244	(TTA)8	165
IITA ESSR20	TGTCAATTTGGGTCCAATTGCAACAGT	69.92	AGCAAGCACATGCCATTCTTTTCTTTC	68.236	(ATT)5	175
IITA ESSR21	TACAGGATTGACGTTGCTGTTGCATGT	69.408	CACAAATGGTGAAGACACAGAAAACGC	68.565	(AAT)6	157
IITA ESSR22	TTGGAATGCACTGAAACTCATTTGGGA	70.084	TGCATAATGAGGTCAAATGTTTGGGG	68.453	(TTAT)5	179
IITA ESSR23	GGTTGATGGGAATGTTGTTTGGCTC	68.486	TGGAGGGGAAGGAGAGATTTTTCAGA	68.104	(AT)6	191
IITA ESSR24	GGGACGCGTGAATTCTTGCTTTTG	69.601	GGTTTCTGAGAGAAAGCATGCGCAGA	70.545	(TA)6	150
IITA ESSR25	TCCTCGTCTTCAAAACCCACAAGGC	70.209	GGTCAGGCAAAGCAATTGGGC	68.083	(ATA)6	158

Appendix 1 cont…						
IITA ESSR26	GCGTGGAAGCAAGGCAATACTGAAT	68.249	TTTCTGCTGCTTTCGAAGCTCTCTGTT	68.484	(ATA)10	171
IITA ESSR27	AGTTGCTGGGTCCTGCGTTTAAGG	68.407	TTCTGGACGTCCTCTTCAGAGCCA	68.696	(AT)7	180
IITA ESSR28	TTCCTATTGACCGACATCCCTCTCCC	69.956	GAGCGAGCGAGACCGAGCGA	70.878	(CT)6	168
IITA ESSR29	TTCGCGTCTTCAATCCGTAGCCA	69.854	GCCGGTGTGAGTCGCGAGAA	69.322	(TC)6	153
IITA ESSR30	CGTGTTGTGCATCTGGGCCG	70.389	CCTTCGAAGTACAACCAAAGCCATGA	68.088	(TG)6	159
IITA ESSR31	CGGCCGCTGCATCAGAGCTT	70.327	TGCCTCTTGGCGGGGGTCTT	70.436	(TC)6	153
IITA ESSR32	GGGGAAATCACAAACTCCAAGCCA	69.129	TGGATCATCGGAGACCCCTCG	68.903	(AG)6	179
IITA ESSR33	CCGCAAGCAACGGCCAAGA	69.964	GGAATATCAACGGTGATGCCGGA	68.974	(CT)7	192
IITA ESSR34	TCACAGGCTGGAGTTTATGAAGGCG	69.329	ACTGCAGCCGCTCCTCCCAA	69.849	(TA)7	181
IITA ESSR35	GGGTCCTGAGCCACCTGCATC	68.638	TGCTCCCGGTAACCAGTGGTGG	70.424	(TA)7	189
IITA ESSR36	ACGATGTTTGTCTTTGAGGATTGGTGG	68.884	TGAGACAACACAGGTGGATTGCAGC	69.703	(CT)7	151
IITA ESSR37	TGGAGGCAGGGCCTTCTTTGC	69.941	TGCATCCCAAGCAAGAGAGGAGAA	68.464	(AT)6	150
IITA ESSR38	TGATCATAAAGCTGGAGCAGAGGCTGA	69.958	AAACTCATGCCCCTCGTGAAAACAA	68.594	(AT)6	170
IITA ESSR39	GTTGCAGCAAAGCTTGCTATCCAATCA	69.841	GGAGGCTCCACTCCCCACTGA	68.288	(AG)6	164
IITA ESSR40	GGGAGTACCTCGAGTACAACGAAGCAA	68.32	ATCGCATGCCTCTGCGTGGA	69.677	(TC)6	150
IITA ESSR41	TCCGCGAAAACAATTTGGCACA	69.551	TCCAATTCCATTTTCATCACCAGCA	68.406	(GA)6	176
IITA ESSR42	TCATTCTTTCCCTGTTTTGCCTTCG	68.184	GCCTTCGTCAGGCAAGGAGCA	69.256	(TA)7	172
IITA ESSR43	TTTGCTCACCAGCACCAGCGA	69.659	TCACGAGCTGACACGTTGCCG	70.254	(CT)6	165
IITA ESSR44	TTCCTCGTTAACGCTGGCCTTGTG	70.139	GAGAAAACGCAATTCCGAGCCAA	68.407	(AT)6	179
IITA ESSR45	GTCTCAGTCCCTGCCAGACCCG	69.768	TCGCCTTCCTCTTCTTTCTGTGTCCA	70.106	(TC)6	152
IITA ESSR46	TCCGTCAACTCTCTCACTCTGCGTTG	70.007	GCCTTGGTTCTAAGAGGGTGGGC	68.131	(AG)7	172
IITA ESSR47	TTCGCTTCTTGACATCTTCCGCC	68.756	TGCAGAGTCCATGGTTTGGCGA	70.282	(GA)6	159
IITA ESSR48	TCTCCGCCCTTCCCCCATCT	69.45	CACGGAAAGCTTGGTGTTTTTGGC	69.479	(TC)6	173
IITA ESSR49	CTGGCACAAAGTGCAGTTGGAGTTG	69.063	GCTGTGAAAATGAACTGCATGCCAC	68.847	(AG)7	162
IITA ESSR50	ACGCCAACTAGCCTCTGATTTCTCACA	68.966	GGCCAAAATCTTTGCAACGTGGT	68.14	(AG)6	178
IITA ESSR51	AGATGGAGAGGCAATGCTGGGC	68.996	CACTTGTTCCTGTGCTTAACCCACCTT	68.127	(AT)6	172
IITA ESSR52	ATGGGTGTCCTTGTGCCTACTGGA	68.001	CCCCAATTGCAGCAAGGCGT	69.621	(AG)6	196
IITA ESSR53	TCCCACTTCCCAGTCAACGCC	68.862	TCGCCTATGCCGACGGAGGA	70.435	(TC)7	178
IITA ESSR54	TGATCATAAAGCTGGAGCAAAGGCTG	68.371	TGGTAAAACTCATGCCCCTCGTGA	68.704	(AT)6	175
IITA ESSR55	AGGGTTGGAGGCTGAGCTGGC	69.327	TGGTGTAAGCGGCTCACCATTCTC	68.558	(AGC)6	150
IITA ESSR56	TTGGGTGACGATGACGCCGA	70.449	GGAGGTACCGGCTTGAAGGGGA	69.498	(GCC)5	161
IITA ESSR57	TCCATGAGCAGTGAAGGAGCTTCAAGT	69.304	GGCACATCATCTTCTCCAATCATAGCC	68.604	(AAC)5	188

Appendix 1 cont…						
IITA ESSR58	TGGATCTGATGAGGAAGGGGATCA	68.151	GGATGATCACCATCTTGCAAGCCTAA	68.175	(TGA)5	162
IITA ESSR59	GGAATGGTTGAAACGGGAAAGCC	68.571	CCAGTGATGATTGGGCTTCATGGTC	69.689	(TCT)7	192
IITA ESSR60	CGTTTCCCTTGCGCTCTCCG	69.484	CCCACTAAGCTGATTGGTTGCTGCT	68.511	(AGC)5	151
IITA ESSR61	CAGCAGCAGCAACAACATCCGC	70.456	CAAGCAGCCCTGCAATCCTCTTTC	69.214	(GCA)5	200
IITA ESSR62	AGCATGAGCGCATGTCTGTGAGC	69.293	TGGGTCGACACCAAATCTACCATTCA	69.263	(AGA)6	173
IITA ESSR63	CCACCAACATCCTCATCATGGAAGAC	68.723	AAGGTCATGATGAACGACTGGAGCA	68.206	(CAT)5	151
IITA ESSR64	GCTACGGGGGATTACACGACCTTTG	69.211	TGCACCACTCGCTCGTTCACC	69.195	(TGC)5	188
IITA ESSR65	GATGGAGCCGCTGACCTCCG	69.8	TGATCGCCGCTTCGACGACTT	69.545	(CTC)5	193
IITA ESSR66	CGTCTCTCCGGTGACGTTGTCG	69.559	GACGGAGCAAATTATCATCATCGAACC	68.29	(TCT)10	157
IITA ESSR67	TATGATCCAGCGCCCAGCGG	70.716	GGTGCATCTGGCGGAATGTCAA	69.568	(AGA)5	159
IITA ESSR68	CGCCGCCCTCGCTTAGCC	70.62	GCTGGAGGGTATGCTGCAGTGG	68.399	(CTT)7	198
IITA ESSR69	TGGAGGCTGTAATGGCTTGCTGG	69.5	TGGGGACAAGAGGACCAAATCCC	69.339	(TGA)5	186
IITA ESSR70	TTGACAGGCCCGCAGCTGGT	70.732	TGGTCTTCAGTCAGGGGAACAGGA	68.411	(CCA)7	172

Appendix 2 Sample of blast results.

BLASTN 1.6.1-Paracel [2005-04-25]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= SSRY1
(20 letters)

Database: cassava_ssr
2160 sequences; 1,474,476 total letters

Searching.....done

***** No hits found ******

Database: cassava_ssr
Posted date: Aug 29, 2007 2:58 PM
Number of letters in database: 1,474,476
Number of sequences in database: 2160

Lambda	K	H
1.37	0.711	1.31

Gapped

Lambda	K	H
1.37	0.711	1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 5
Number of Sequences: 2160
Number of extensions: 5
Number of successful extensions: 5
Number of sequences better than 1.0e-03: 0
length of query: 20
length of database: 1,474,476
effective HSP length: 11
effective length of query: 9
effective length of database: 1,450,716
effective search space: 13056444
effective search space used: 13056444
T: 0

A: 40
X1: 6 (11.9 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 17 (34.2 bits)
BLASTN 1.6.1-Paracel [2005-04-25]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= SSRY2
(20 letters)

Database: cassava_ssr
2160 sequences; 1,474,476 total letters

Searching.....done

***** No hits found ******

Database: cassava_ssr
Posted date: Aug 29, 2007 2:58 PM
Number of letters in database: 1,474,476
Number of sequences in database: 2160

Lambda	K	H
1.37	0.711	1.31

Gapped

Lambda	K	H
1.37	0.711	1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 7
Number of Sequences: 2160
Number of extensions: 7
Number of successful extensions: 7
Number of sequences better than 1.0e-03: 0
length of query: 20
length of database: 1,474,476
effective HSP length: 11
effective length of query: 9

effective length of database: 1,450,716
effective search space: 13056444
effective search space used: 13056444
T: 0
A: 40
X1: 6 (11.9 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 17 (34.2 bits)
BLASTN 1.6.1-Paracel [2005-04-25]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= SSRY3
(20 letters)

Database: cassava_ssr
2160 sequences; 1,474,476 total letters

Searching.....done

***** No hits found ******

BLASTN 1.6.1-Paracel [2005-04-25]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= SSRY1
(20 letters)

Database: cassava_ssr
2160 sequences; 1,474,476 total letters

Searching.....done

***** No hits found ******

Database: cassava_ssr

Posted date: Aug 29, 2007 2:58 PM
Number of letters in database: 1,474,476
Number of sequences in database: 2160

Lambda	K	H
1.37	0.711	1.31

Gapped

Lambda	K	H
1.37	0.711	1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 3
Number of Sequences: 2160
Number of extensions: 3
Number of successful extensions: 3
Number of sequences better than 1.0e-03: 0
length of query: 20
length of database: 1,474,476
effective HSP length: 11
effective length of query: 9
effective length of database: 1,450,716

Appendix 3 Excel sheet results for blast results (forward 0.0001 matches)

Known SSR 817	**SSR containing EST 2160**	**% match**	**No. of bases that match**	**Start base**	**Finish base**	**E-value**	E-VALUE
SSRY43	me512003460ssr1	100	23	345	367	2.10E-07	46.09
SSRY103	DV441775ssr1	100	21	105	125	2.70E-06	42.12
SSRY106	DV441775ssr1	100	20	111	130	1.10E-05	40.14
SSRY197	TA1363ssr2	100	22	270	291	7.70E-07	44.1
SSRY197	TA1363ssr1	100	22	270	291	7.70E-07	44.1
SSRY203	DV441775ssr1	100	20	130	149	1.10E-05	40.14
SSRY204	TA3249ssr1	100	20	718	737	1.10E-05	40.14
SSRY206	me512001909ssr1	100	20	58	77	1.10E-05	40.14
SSRY215	TA207ssr1	100	20	59	78	1.10E-05	40.14
SSRY216	TA1339ssr1	100	20	290	309	1.10E-05	40.14
SSRY217	TA1538ssr1	100	20	629	648	1.10E-05	40.14
SSRY233	TA393ssr1	100	20	206	225	1.10E-05	40.14
SSRY239	me512000753ssr1	100	20	361	380	1.10E-05	40.14
SSRY239	DV445170ssr1	100	20	271	290	1.10E-05	40.14
SSRY246	DV446397ssr1	100	20	419	438	1.10E-05	40.14
SSRY250	TA5492ssr1	100	20	327	346	1.10E-05	40.14
SSRY252	DV454501ssr1	100	21	153	173	2.70E-06	42.12
SSRY252	CK650997ssr1	100	21	190	210	2.70E-06	42.12
SSRY257	DV448119ssr1	100	20	160	179	1.10E-05	40.14
SSRY273	TA3249ssr1	100	20	239	258	1.10E-05	40.14
SSRY278	TA554ssr1	100	20	1235	1254	1.10E-05	40.14
SSRY292	DV456190ssr1	100	26	99	124	4.40E-09	52.03
SSRY293	CK651736ssr1	100	19	1	19	3.80E-05	38.16

Appendix 4 Euclidean distance for South American and African genotypes

OTU	6215	6217	6218	6219	ARG12	BRA1001	BRA1016	BRA200
6215	0.0000	0.4406	0.3132	0.2652	0.4635	0.6030	0.4602	0.6178
6217	0.4406	0.0000	0.3817	0.3185	0.4930	0.6166	0.6187	0.7305
6218	0.3132	0.3817	0.0000	0.2593	0.4964	0.5678	0.5283	0.7124
6219	0.2652	0.3185	0.2593	0.0000	0.3607	0.5289	0.5246	0.5992
ARG12	0.4635	0.4930	0.4964	0.3607	0.0000	0.4961	0.5226	0.5297
BRA1001	0.6030	0.6166	0.5678	0.5289	0.4961	0.0000	0.4944	0.5621
BRA1016	0.4602	0.6187	0.5283	0.5246	0.5226	0.4944	0.0000	0.6145
BRA200	0.6178	0.7305	0.7124	0.5992	0.5297	0.5621	0.6145	0.0000
BRA206	0.6403	0.5219	0.5788	0.3934	0.4092	0.5594	0.5094	0.6403
BRA255	0.5935	0.5953	0.4657	0.5406	0.7227	0.4123	0.4206	0.5935
BRA436	0.6482	0.5347	0.5027	0.3572	0.4377	0.4556	0.3788	0.6482
BRA785	0.4750	0.4746	0.5883	0.4546	0.6149	0.4925	0.4650	0.4750
BRA842	0.5613	0.5097	0.5015	0.3390	0.3994	0.5292	0.6182	0.5613
COL1468	0.4148	0.4298	0.4304	0.4194	0.2913	0.5001	0.6604	0.4148
COL1734	0.5292	0.5097	0.5084	0.4120	0.6269	0.5848	0.4192	0.5292
COL233	0.4821	0.7863	0.8649	0.7137	0.4738	0.4970	0.4961	0.4821
COL2459	0.5464	0.4167	0.5452	0.4046	0.5378	0.3614	0.4419	0.5464
COL2638	0.6931	0.4673	0.3193	0.4045	0.6108	0.5168	0.3650	0.6931
CR19	0.5452	0.6350	0.5168	0.4050	0.5144	0.5452	0.6215	0.5452
GUA43	0.5238	0.4080	0.4273	0.3227	0.3830	0.5781	0.5869	0.5238
GUA59	0.6619	0.5289	0.5581	0.4574	0.5613	0.6273	0.5500	0.6619
PAR23	0.6976	0.4836	0.5641	0.4340	0.3889	0.5265	0.4304	0.6976
TAI1	0.4624	0.5366	0.6202	0.4940	0.5107	0.5454	0.4750	0.4624
TME1368	0.5621	0.3064	0.2963	0.1951	0.2992	0.4949	0.4633	0.5621
TME1389	0.5094	0.5447	0.5269	0.4277	0.5144	0.5048	0.5254	0.5094
TME230	0.5837	0.4401	0.4633	0.3731	0.5597	0.5950	0.5145	0.5837
TME290	0.6788	0.4171	0.3421	0.4010	0.5131	0.5020	0.4231	0.6788
TME396	0.4983	0.5331	0.5157	0.4452	0.4930	0.5192	0.5368	0.4983
TME539	0.6599	0.2210	0.3806	0.3421	0.3830	0.4714	0.4790	0.6599
TME589	0.6800	0.4484	0.2676	0.4452	0.6230	0.5236	0.4290	0.6800
USA7	0.5743	0.4933	0.3771	0.3640	0.4262	0.4816	0.6197	0.5743
VEN77	0.5493	0.4440	0.4324	0.4334	0.5841	0.5953	0.4887	0.5493

COL233	COL1734	COL1468	BRA842	BRA785	BRA436	BRA255	BRA206	Appendix 4 cont…
0.7134	0.4037	0.4808	0.4382	0.4145	0.3836	0.4944	0.5359	**6215**
0.7863	0.5097	0.4298	0.5097	0.4746	0.5347	0.5953	0.5219	**6217**
0.8649	0.5084	0.4304	0.5015	0.5883	0.5027	0.4657	0.5788	**6218**
0.7137	0.4120	0.4194	0.3390	0.4546	0.3572	0.5406	0.3934	**6219**
0.4738	0.6269	0.2913	0.3994	0.6149	0.4377	0.7227	0.4092	**ARG12**
0.4970	0.5848	0.5001	0.5292	0.4925	0.4556	0.4123	0.5594	**BRA10**
0.4961	0.4192	0.6604	0.6182	0.4650	0.3788	0.4206	0.5094	**BRA10**
0.4821	0.5292	0.4148	0.5613	0.4750	0.6482	0.5935	0.6403	**BRA20**
0.5855	0.5062	0.4782	0.5902	0.5298	0.3536	0.7456	0.0000	**BRA20**
0.6217	0.4427	0.4596	0.5527	0.5046	0.6456	0.0000	0.7456	**BRA25**
0.5893	0.3536	0.5224	0.5094	0.4623	0.0000	0.6456	0.3536	**BRA43**
0.6174	0.4431	0.4870	0.6374	0.0000	0.4623	0.5046	0.5298	**BRA78**
0.6274	0.4124	0.5189	0.0000	0.6374	0.5094	0.5527	0.5902	**BRA84**
0.5876	0.5144	0.0000	0.5189	0.4870	0.5224	0.4596	0.4782	**COL14**
0.4625	0.0000	0.5144	0.4124	0.4431	0.3536	0.4427	0.5062	**COL17**
0.0000	0.4625	0.5876	0.6274	0.6174	0.5893	0.6217	0.5855	**COL23**
0.4738	0.5821	0.5467	0.5821	0.4556	0.4612	0.6342	0.4262	**COL24**
0.6734	0.3775	0.5440	0.5910	0.4478	0.4848	0.4206	0.6653	**COL26**
0.5209	0.4612	0.5100	0.7048	0.5519	0.5598	0.6561	0.5452	**CR19**
0.6217	0.5883	0.4851	0.4245	0.4731	0.5556	0.6236	0.5765	**GUA43**
0.6872	0.5103	0.6080	0.7030	0.4961	0.5547	0.5880	0.5984	**GUA59**
0.6737	0.5880	0.6218	0.5383	0.4854	0.3794	0.7367	0.5614	**PAR23**
0.7071	0.6342	0.5111	0.7134	0.4739	0.5534	0.6787	0.5724	**TAI1**
0.6217	0.4450	0.3607	0.3536	0.4293	0.3807	0.5224	0.5366	**TME13**
0.6800	0.4808	0.4541	0.5057	0.4190	0.4190	0.6766	0.5995	**TME13**
0.6528	0.4464	0.4714	0.3794	0.3334	0.5168	0.4450	0.5340	**TME23**
0.6899	0.4299	0.4243	0.4906	0.3836	0.4425	0.3858	0.5538	**TME29**
0.6491	0.4976	0.4643	0.5206	0.4389	0.4389	0.6777	0.6075	**TME39**
0.5971	0.5359	0.4091	0.4836	0.4983	0.4595	0.5439	0.5097	**TME53**
0.7542	0.4561	0.5015	0.5168	0.4258	0.5347	0.4906	0.6183	**TME58**
0.7630	0.5788	0.3496	0.5267	0.4250	0.6136	0.5389	0.6090	**USA7**
0.6554	0.2832	0.4714	0.5091	0.4382	0.5097	0.5222	0.5864	**VEN77**

TME1368	TAI1	PAR23	GUA59	GUA43	CR19	COL2638	COL2459	Appendix 4 cont....
0.3064	0.5678	0.4635	0.4823	0.4802	0.4915	0.3579	0.5864	**6215**
0.3064	0.5366	0.4836	0.5289	0.4080	0.6350	0.4673	0.4167	**6217**
0.2963	0.6202	0.5641	0.5581	0.4273	0.5168	0.3193	0.5452	**6218**
0.1951	0.4940	0.4340	0.4574	0.3227	0.4050	0.4045	0.4046	**6219**
0.2992	0.5107	0.3889	0.5613	0.3830	0.5144	0.6108	0.5378	**ARG12**
0.4949	0.5454	0.5265	0.6273	0.5781	0.5452	0.5168	0.3614	**BRA10**
0.4633	0.4750	0.4304	0.5500	0.5869	0.6215	0.3650	0.4419	**BRA10**
0.5621	0.4624	0.6976	0.6619	0.5238	0.5452	0.6931	0.5464	**BRA20**
0.5366	0.5724	0.5614	0.5984	0.5765	0.5452	0.6653	0.4262	**BRA20**
0.5224	0.6787	0.7367	0.5880	0.6236	0.6561	0.4206	0.6342	**BRA25**
0.3807	0.5534	0.3794	0.5547	0.5556	0.5598	0.4848	0.4612	**BRA43**
0.4293	0.4739	0.4854	0.4961	0.4731	0.5519	0.4478	0.4556	**BRA78**
0.3536	0.7134	0.5383	0.7030	0.4245	0.7048	0.5910	0.5821	**BRA84**
0.3607	0.5111	0.6218	0.6080	0.4851	0.5100	0.5440	0.5467	**COL14**
0.4450	0.6342	0.5880	0.5103	0.5883	0.4612	0.3775	0.5821	**COL17**
0.6217	0.7071	0.6737	0.6872	0.6217	0.5209	0.6734	0.4738	**COL23**
0.3536	0.3349	0.4501	0.6256	0.3162	0.6456	0.5303	0.0000	**COL24**
0.3836	0.5711	0.5224	0.5366	0.5848	0.5808	0.0000	0.5303	**COL26**
0.4733	0.5699	0.5950	0.3669	0.6178	0.0000	0.5808	0.6456	**CR19**
0.2926	0.4351	0.3830	0.5294	0.0000	0.6178	0.5848	0.3162	**GUA43**
0.4750	0.5666	0.5467	0.0000	0.5294	0.3669	0.5366	0.6256	**GUA59**
0.3449	0.5002	0.0000	0.5467	0.3830	0.5950	0.5224	0.4501	**PAR23**
0.4139	0.0000	0.5002	0.5666	0.4351	0.5699	0.5711	0.3349	**TAI1**
0.0000	0.4139	0.3449	0.4750	0.2926	0.4733	0.3836	0.3536	**TME13**
0.3720	0.5057	0.4837	0.5650	0.5295	0.5621	0.4996	0.4851	**TME13**
0.4080	0.5735	0.4612	0.5133	0.3817	0.5870	0.3472	0.5062	**TME23**
0.3237	0.6109	0.4663	0.5168	0.4684	0.6603	0.3697	0.5147	**TME29**
0.3708	0.5134	0.4930	0.5493	0.5406	0.5673	0.5126	0.4940	**TME39**
0.2682	0.4243	0.4304	0.6037	0.4212	0.7071	0.4273	0.2652	**TME53**
0.4525	0.6925	0.4958	0.5883	0.5185	0.5913	0.2431	0.4940	**TME58**
0.3705	0.5267	0.6501	0.4464	0.4795	0.4935	0.4996	0.5282	**USA7**
0.4731	0.6697	0.5354	0.5276	0.5358	0.5167	0.3536	0.5677	**VEN77**

Appendix 4 cont…	TME1389	TME230	TME290	TME396	TME539	TME589	USA7	VEN77
6215	0.4045	0.3878	0.3643	0.4228	0.4673	0.3956	0.4415	0.4729
6217	0.5447	0.4401	0.4171	0.5331	0.2210	0.4484	0.4933	0.4440
6218	0.5269	0.4633	0.3421	0.5157	0.3806	0.2676	0.3771	0.4324
6219	0.4277	0.3731	0.4010	0.4452	0.3421	0.4452	0.3640	0.4334
ARG12	0.5144	0.5597	0.5131	0.4930	0.3830	0.6230	0.4262	0.5841
BRA10	0.5048	0.5950	0.5020	0.5192	0.4714	0.5236	0.4816	0.5953
BRA10	0.5254	0.5145	0.4231	0.5368	0.4790	0.4290	0.6197	0.4887
BRA20	0.5094	0.5837	0.6788	0.4983	0.6599	0.6800	0.5743	0.5493
BRA20	0.5995	0.5340	0.5538	0.6075	0.5097	0.6183	0.6090	0.5864
BRA25	0.6766	0.4450	0.3858	0.6777	0.5439	0.4906	0.5389	0.5222
BRA43	0.4190	0.5168	0.4425	0.4389	0.4595	0.5347	0.6136	0.5097
BRA78	0.4190	0.3334	0.3836	0.4389	0.4983	0.4258	0.4250	0.4382
BRA84	0.5057	0.3794	0.4906	0.5206	0.4836	0.5168	0.5267	0.5091
COL14	0.4541	0.4714	0.4243	0.4643	0.4091	0.5015	0.3496	0.4714
COL17	0.4808	0.4464	0.4299	0.4976	0.5359	0.4561	0.5788	0.2832
COL23	0.6800	0.6528	0.6899	0.6491	0.5971	0.7542	0.7630	0.6554
COL24	0.4851	0.5062	0.5147	0.4940	0.2652	0.4940	0.5282	0.5677
COL26	0.4996	0.3472	0.3697	0.5126	0.4273	0.2431	0.4996	0.3536
CR19	0.5621	0.5870	0.6603	0.5673	0.7071	0.5913	0.4935	0.5167
GUA43	0.5295	0.3817	0.4684	0.5406	0.4212	0.5185	0.4795	0.5358
GUA59	0.5650	0.5133	0.5168	0.5493	0.6037	0.5883	0.4464	0.5276
PAR23	0.4837	0.4612	0.4663	0.4930	0.4304	0.4958	0.6501	0.5354
TAI1	0.5057	0.5735	0.6109	0.5134	0.4243	0.6925	0.5267	0.6697
TME13	0.3720	0.4080	0.3237	0.3708	0.2682	0.4525	0.3705	0.4731
TME13	0.0000	0.4633	0.4367	0.0000	0.4984	0.4762	0.4583	0.5055
TME23	0.4633	0.0000	0.3421	0.4790	0.4984	0.3878	0.4182	0.3901
TME29	0.4367	0.3421	0.0000	0.4317	0.3859	0.2885	0.4179	0.4501
TME39	0.0000	0.4790	0.4317	0.0000	0.5115	0.5059	0.4513	0.5352
TME53	0.4984	0.4984	0.3859	0.5115	0.0000	0.4684	0.4681	0.5122
TME58	0.4762	0.3878	0.2885	0.5059	0.4684	0.0000	0.4823	0.2737
USA7	0.4583	0.4182	0.4179	0.4513	0.4681	0.4823	0.0000	0.5525
VEN77	0.5055	0.3901	0.4501	0.5352	0.5122	0.2737	0.5525	0.0000

Appendix 5 Quantities for optimised PCR conditions for all the 33 polymorphic EST-SSR primers in 10 µl reaction volume.

MARKER	DYE 1.0pmole µl	DNA 50ng µl	PRIMER<F> 1.0pmole µl	PRIMER<R> 1.0pmole µl	BUFFER 10X µl	Taq 50U µl	dNTPS 2.5mM µl	$MgCl_2$ 25mM µl	Distilled Water µl	Annealing temp ^0C
IITA ESSR01	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR03	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR04	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR05	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	57
IITA ESSR07	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR08	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR11	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR13	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR14	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR15	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR17	0.175	1.0	0.8	0.8	1.0	0.075	0.8	1.2	4.55	62
IITA ESSR19	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR21	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR23	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR24	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR25	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR26	0.175	1.0	0.8	0.8	1.0	0.075	0.8	1.2	4.35	62
IITA ESSR27	0.175	1.0	0.8	0.8	1.0	0.075	0.8	1.2	4.35	62
IITA ESSR30	0.175	1.0	0.8	0.8	1.0	0.075	0.8	1.2	4.55	62
IITA ESSR33	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR34	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.6	4.75	62
IITA ESSR35	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR36	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR37	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.6	4.75	62

Appendix 5 cont…	DYE 1.0pmole µl	DNA 50ng µl	PRIMER<F> 1.0pmole µl	PRIMER<R> 1.0pmole µl	BUFFER 10X µl	Taq 50U µl	dNTPS 2.5mM µl	$MgCl_2$ 25mM µl	Distilled Water µl	Annealing temp ^0C
IITA ESSR38	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR39	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR40	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR43	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.55	62
IITA ESSR44	0.175	1.0	0.8	0.8	1.0	0.075	0.8	1.2	4.35	62
IITA ESSR51	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR53	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.4	4.95	62
IITA ESSR66	0.175	1.0	0.8	0.8	1.0	0.075	0.8	0.8	4.55	62
IITA ESSR70	0.175	1.0	0.8	0.8	1.0	0.075	0.8	1.2	4.35	62