IDENTIFICATION OF MICROSATELLITE MARKERS FOR FINGER MILLET (*ELEUSINE CORACANA)* BY ANALYSIS OF ROCHE 454 GS-FLX TITANIUM SEQUENCE DATA

BY

GIMODE DAVIS MUSIA (B.Sc)

Reg No. I56/23200/2010

A Thesis Submitted in Partial Fulfillment of the Requirements for the Award of the Degree of Master of Science (Biotechnology) in the School of Pure and Applied Sciences of Kenyatta University

December, 2013

DECLARATION

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

Gimode Davis Musia

Signature _____ Date _____

We confirm the work reported in this thesis was carried out by the candidate under our supervision

Dr. Alice Muchugi
Department of Biochemistry and Biotechnology
Kenyatta University
P.O Box 43844-00100
Nairobi

Signature _____ Date _____

Dr. Santie De Villiers
Pwani University
P.O Box 195-80108
Kilifi

Signature _____ Date _____

DEDICATION

To my father, who I look up to with great admiration

ACKNOWLEDGEMENT

I would like to express my depth of gratitude to the Almighty God, from whom I receive every perfect gift including the gift of life. For fortitude, providence and the ability to reach this far.

To my supervisor, Dr. Santie De Villiers. For your patience and for being an ever available source of inspiration. Despite my ineptitude, you have been a guide and a light in the labyrinth that is the scientific jungle. From you I have gleaned and learnt much in the course of this project. They say that, the mind is not a vacuum to be filled but a fire to be lit. You have been the spark that has ignited my passion and desire to soar even higher in the scientific realms.

I am grateful to the late Prof. Jesse Machuka for being more than just a supervisor. You were a mentor, from whom I learnt a lot, not just in science, but on matters of life as a whole. Many thanks to Dr. Alice Muchugi for diligently helping me to go through the thesis and for never tiring of meeting me even outside office hours. I appreciate the support I received from the staff and fellow students at the International Crops Research Institute for the Semi-Arid Tropics. Special thanks to Mr. Vincent Njung'e and Miss. Annis Saiyiori. For providing an excellent working environment and for taking your time to offer technical guidance.

A million thanks to my parents, Dr. Edwin and Mrs. Jescah Gimode. For ever being there for me and for walking with me through the journey of life. Not only are you my parents, but you are also my mentors in the world of academia. Because of you, I am challenged to exceed my limitation and enlarge the boundaries of my expectations. To my brother Kevin O'bbede and my sister Sharon Kadagaya, I could never ask for better siblings.

I am indebted to my wife Winnie Riziki, for your encouragement and for being a buttress of support to me. I appreciate the immense support of my friends. Worthy of special mention are Joyce Nthiga and Paddy Nthiga, for the constant reminder that the word 'impossible' should not feature in one's dictionary. To everyone else who has assisted me in any way, great or small, may God bless you abundantly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

AFLP	Amplified fragment length polymorphism
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
Bp	Base pairs
CAPS	Cleaved amplified polymorphic sequences
CCD	Charge coupled device
cpSSR	Chloroplast simple sequence repeat
CTAB	Cetyl trimethyl-ammonium bromide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DTT	Dithiothreitol
EDTA	Ethylene diamine tetra-acetic acid
emPCR	Emulsion polymerase chain reaction
ESTs	Expressed Sequence Tags
EtOH	Ethanol
FAO	Food and Agricultural Organization
GUI	Graphical user interface
HCl	Hydrochloric acid
ICECP	ICRISAT *Eleusine coracana* Primer
ICRISAT	International Crop Research Institute For The Semi-Arid Tropics
ISSR	Inter simple sequence repeats
KCl	Potassium chloride

MABC	Marker assisted back crossing
MAS	Marker assisted selection
MB	Megabase
MFS	Major facilitator superfamily
$MgCl_2$	Magnesium chloride
MIRA	Mimicking intelligent read assembly
MISA	Microsatellite
mRNA	Messenger ribonucleic acid
mtSSR	Mitochondrial simple sequence repeat
NaCl	Sodium chloride
NaOAc	Sodium acetate
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
PCR	Polymerase chain reaction
PIC	polymorphic information content
QC	Quality control
QTL	Quantitative trait loci
RAPD	Random amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SCAR	Sequence characterized amplified regions
SciRoKo	SSR classification and identification by Robert Kofler
SNP	Singe nucleotide polymorphism

SSLP	simple sequence length polymorphisms
SSR	Simple sequence repeat
STR	Short tandem repeats
TE	Tris ethylene diamine tetra-acetic acid
UTR	Untranslated region
UV	Ultra violet
w/v	weight per volume

ABSTRACT

Finger millet is an important cereal cultivated in Eastern Africa as well as Southern India. It is a staple crop that is characterized by ability to thrive on a variety of environmental conditions, excellent grain storage quality and ability to withstand significant levels of salinity. Scientific research aimed at improving this important cereal has been negligible and it is regarded as one of the orphaned crop species. The aim of this project was to isolate microsatellite markers from finger millet by analyzing data provided following 454 GS-FLX Titanium sequencing. This is a next generation sequencing platform that confers the potential of isolating a greater number of microsatellites at a much lower cost than the conventional Sanger sequencing platform. These markers are hyper variable, and exhibit wide genomic distribution, co-dominant inheritance, reproducibility, multi-allelic nature and chromosome specific location. The 2 genotypes of finger millet studied were obtained from ICRISAT, Nairobi. These were KNE 755 and KNE 796. *In silico* tools were used to mine for the microsatellites from sequence data obtained after enrichment and 454 GS-FLX sequencing of finger millet genomic DNA. These tools included NGSQC, for quality screening, MIRA for sequence assembly, SciRoKo for SSR mining and BatchPrimer3 for primer design. *In vitro* analysis was undertaken to evaluate the markers supplied by Ecogenics. This involved PCR analysis followed by allele calling using the Gene Mapper software package. The allelic data generated was subjected to statistical analysis using PowerMarker to reveal polymorphism. The *in silico* study resulted in the identification of 92 primers that were unique from the published as well as the Ecogenics supplied primers. From the *in vitro* study, 49 markers were polymorphic and the average polymorphic information content (PIC) was 0.4153. These markers are a significant addition to the existing 82 SSRs. They are also valuable tools that will be useful for conducting further genomic studies in finger millet, including MAS, fingerprinting studies as well as assaying genetic diversity. This study has demonstrated the use of NGS to rapidly and cost-effectively generate genomic sequences containing SSR motifs. It points to the possibility of using cutting edge technology to advance research in underutilized crops such as finger millet by researchers in Kenya. It also shows that it is possible to develop abundant genomic resources for hitherto understudied crops which have great significance to the development of third world countries in Africa and Asia.

CHAPTER ONE

INTRODUCTION

1.1 Background

Eleusine coracana, finger millet, is a tetraploid crop (2n=4x=36) belonging to the subfamily *Chloridoideae* (Srinivasachary *et al*., 2007). It is an important cereal in Eastern Africa and is also widely cultivated in Southern India (Dida *et al*., 2008). In the countries where it is grown, it is commonly referred to as Wimbi (Swahili), Bulo (Uganda), Tellebun (Sudan) and Ragi (India) (Reddy *et al*., 2011). Its value stems from its versatility as a staple food, and its excellent grain storage quality attributed to its high polyphenol content and small grain size that deters storage pests (Chetan and Malleshi, 2007). The fact that it thrives on a variety of environmental conditions, particularly in marginal areas, makes it an ideal crop during famine. Finger millet seed can lie dormant for weeks in times of drought, germinating at the onset of rains and can be harvested in just forty five days (Styslinger, 2011). It can also withstand significant levels of salinity and is not plagued by many serious diseases, except finger millet blast disease (Barbeu, 1993; National Research Council, 1996; Dida *et al*., 2007).

Because of its high nutritional value, it is an important food security crop (Dida *et al*., 2007). Its calcium content is higher than that of other major cereals. It is also rich in iron as well as the amino acids leucine, tryptophan, phenylalanine and methionine. Consequently, it is a major preventive agent against malnutrition (Barbeu, 1993; National Research Council, 1996). After grain harvest, the finger millet stover (the leaves and stalks that remain after the grain is harvested) is an important source of

fodder for livestock. This contributes in making finger millet of significant importance in food and feed security in areas that rely on crop and livestock farming (Krishnappa *et al*., 2009).

Production statistics show that 94% of global millet output is from developing countries mainly in Asia and Africa. In these countries, 95% of millet produced is consumed as food with a large number of households aiming to produce just enough to feed them. However, many often are unable to meet this goal (FAO and ICRISAT, 1996). The first of the Millennium development goals is to eradicate extreme poverty and hunger by 2015 (http://www.undp.org/mdg/basics.shtml). Not only is it critical to halve hunger by 2015, it is also vital to address the challenge of hidden hunger. This is due to lack of micronutrients such as vitamins and minerals whose impact on the body is profound relative to the amount needed (IPGRI *et al*., 2005). Finger millet, which is high in energy, rich in micronutrients and essential amino acids can play a role in achieving this goal.

Although finger millet has been grown as a subsistence crop for many decades, there has been very little scientific intervention aimed at improving and developing this important cereal (Kumari and Pande, 2010). This may be due to the fact that millets in general are of little economic importance to developed countries in comparison to major cereals such as maize, wheat and rice (Kothari *et al*., 2005). Finger millet is regarded as one of the orphaned crop species, which are crops with underexploited potential for

contributing to food security, health (nutrition/medicinal), income generation and environmental services (Jaenicke and Hoschle-Zeledon, 2006).

A number of characteristics can be attributed to orphaned species in general, such as: (i) great potential for improving income, food security and combating micronutrient deficiency, (ii) they are adapted to specific agro-ecological niches and marginal lands, (iii) they may be highly nutritious with multiple uses and (iv) they receive little attention in terms of research aimed at improving them. This may partly be attributed to the fact that their importance is largely local or regional in scale (Dawson *et al*., 2009).

Research done on finger millet includes attempts to improve it through hybridization of various varieties, in order to create useful variability (Krishnappa *et al*., 2009). Expressed sequence tags (ESTs) linked to drought stress and salt tolerance have also been identified which can aid in enhanced production in challenging environments (Dawson *et al*., 2009). There has been an attempt to produce blast resistant transgenic finger millet by introduction of a synthetic anti-fungal *pin* gene (Latha *et al*., 2005). In addition, quite significant research has been done with the aim of developing regeneration protocols through tissue culture (Eapan and George 1989; Poddar *et al*., 1997; Kumar *et al*., 2001). As a first step towards mapping traits of agronomic importance, a skeleton genetic map has been constructed by Dida *et al*. (2007).

1.2 Statement of the problem

Finger millet is a crop of immense importance in areas where it is grown. Properties such as its excellent grain storage quality, high nutritional value and ability to thrive in

marginal areas make it an important food security crop. However, it still remains an orphaned crop with little research having been undertaken to improve it, as compared to the major cereals such as maize, wheat and rice. According to Damme *et al*. (2010), no molecular markers associated with agronomically important traits have been developed for the crop. In addition, only 82 microsatellite markers have been published for finger millet (Dida *et al*., 2007). These were developed through the traditional Sanger sequencing method and none of them have been linked to agronomically important traits.

1.3 Justification

The aim of this study was to isolate additional microsatellite markers for finger millet by analyzing next generation sequencing (NGS) data from short sequence repeat (SSR) enriched finger millet DNA libraries. Isolation of microsatellites by traditional Sanger sequencing has been hindered by the fact that the method is difficult, time consuming and costly, yielding only a small number of SSRs (Malausa *et al*., 2011). The use of NGS in this study was aimed at overcoming these drawbacks. This is in view of the efficiency and ever decreasing cost of NGS approaches (Castoe *et al*., 2010). Thus, this study sought to generate more microsatellite markers to supplement the few published SSRs. These markers will be useful for finger millet improvement by facilitating activities such as genome wide screens for variation, fingerprinting, analysis of genetic diversity and genotyping among many other applications (Rudd *et al*., 2005; Dawson, 2009; Anithakumari *et al*., 2010).

1.4 Hypothesis

Analysis of 454 GS-FLX Titanium sequence data is not an ideal method of developing additional markers for finger millet.

1.5 Objectives of the study

1.5.1 General objective

To develop microsatellite markers for finger millet.

1.5.2 Specific objectives

i. To develop novel SSRs from 454 GS-FLX Titanium sequence data obtained from microsatellite enriched finger millet genomic DNA.

ii. To evaluate polymorphism of the new SSRs among ten finger millet varieties.

CHAPTER TWO

LITERATURE REVIEW

2.1 Genetic markers

The discovery of genetic markers opened a new frontier in crop improvement as they enabled polymorphisms to be detected at the genomic level. Adoption of these markers revolutionized plant biotechnology with techniques constantly being developed to assess genetic variation with greater precision, speed and cost effectiveness (Kumar *et al*., 2009). Genetic markers can be described as DNA sequences that are easily detected and their inheritance monitored (Kumar, 2009). Examples of common molecular markers include; Restriction Fragment Length Polymorphisms (RFLPs), Random Amplified Polymorphic DNA (RAPDs), Cleaved Amplified Polymorphic Sequences (CAPS), Sequence Characterized Amplified Regions (SCARs), Simple Sequence Repeats (SSRs), Inter Simple Sequence Repeats (ISSRs) and Single Nucleotide Polymorphisms (SNPs) (Maheswaran, 2004).

Use of molecular markers allows indirect selection of traits of interest at the seedling stage. This results in saving time, resources and energy that would have been expended in raising large numbers of individuals in segregating populations for several generations. Additionally, indirect selection ensures the absence of confounding effects as a result of the environment as well as facilitating the pyramiding of economically important genes such as for drought resistance. These are tasks that are difficult to achieve by conventional breeding (Gupta and Varshney, 2000).

2.2 Application of molecular markers in finger millet research

Various researchers have employed the use of several molecular markers in studying different aspects of finger millet. RAPD markers have been used most frequently in studies on finger millet. This is probably due to their simplicity and applicability (Bardakci, 2000). However, these markers are not reproducible and have problems with data scoring. RFLPs and to a lesser extent, SSRs, have also been used in studying finger millet. Dida *et al*. (2007) employed RFLP, AFLP, EST and SSR markers to generate a skeleton genetic map of finger millet that covered the nine homoeologous chromosomes of the crop. Parani *et al*. (2001) used PCR-RFLP together with particular restriction enzyme combinations to generate species specific markers for finger millet.

Gupta *et al*. (2010) investigated genetic relatedness between three finger millet varieties with variable seed coat colours, i.e. brown, white and golden using RAPD and ISSR markers. The study revealed maximum similarity between the genotypes with white coloured seed coat and the one with golden coloured seed coat. Das and Misra (2010) studied diversity among fifteen finger millet genotypes using RAPDs in which they reported polymorphism ranging from 50% to 100% for the primers they used. Kumar and Pande (2010) analyzed genetic diversity in eleven germplasms using RAPDs and reported 61.9% polymorphism.

Fakrudin *et al*. (2007) conducted diversity studies using RAPDs in twelve accessions from different geographical regions and reported 85.82% polymorphism across the whole set of samples. They observed fewer polymorphisms for Indian accessions

compared to the more diverse African accessions. Babu *et al*. (2007) used 32 finger millet genotypes to assess diversity using RAPDs and established that RAPDs were useful in discriminating between finger millet genotypes, reporting polymorphism of between 44.5 and 100%. Panwar *et al*. (2010) did a comparative evaluation of genetic diversity using RAPDs, SSRs and cytochrome P450. The level of polymorphism they reported for each of the markers was 49.43, 50.2 and 58.7% for RAPDs, SSRs and cytochrome P450, respectively in the 52 genotypes they studied.

Sinha and Pande (2010) did a finger printing study of finger millet using microsatellites. They hailed the success of microsatellites in uncovering variation in finger millet, which is a self-pollinated plant and thus shows little or no polymorphism with other markers. Dida *et al*. (2008) used 45 SSR markers to evaluate genotypic variation among 79 finger millet accessions across their range of cultivation in Africa and Asia. This so far has been the largest study conducted in terms of number of markers and number of accessions studied. The study predicted origins of breeding material unknown prior to the study. The study also substantiated the theory that finger millet was domesticated in Africa and then introduced to India. It is worth noting that all the finger millet SSRs used for the various studies were obtained from among the 82 SSRs published by Dida *et al*. (2007).

2.3 Microsatellites

Microsatellites, also referred to as short tandem repeats (STRs), simple sequence repeats (SSRs) or simple sequence length polymorphisms (SSLPs), are tandem repeated motifs

of 1-6 bases. They are found in all prokaryotic and eukaryotic genomes and are present in both coding and non-coding regions (Zane *et al*., 2002). In eukaryotes, these simple sequences are highly repeated and thus offer a nearly unlimited and accessible supply of polymorphisms (Tautz, 1989). They can be analysed by PCR owing to the fact that the sequences that flank specific SSRs are conserved within a particular species, across species, within a genus and at times across related genera and these conserved areas can be used to design primers for each SSR that is unique (Varshney *et al*., 2002; Parida *et al*., 2009).

Microsatellites are characterized by high degree of length polymorphism (Zane *et al*., 2002). This is attributable to the different number of repeats in the microsatellite regions, which makes it possible to easily and reproducibly detect them by PCR (Kalia *et al*., 2011). These markers are hyper variable, and exhibit wide genomic distribution, co-dominant inheritance, reproducibility, multi-allelic nature and chromosome specific location. These are attributes that make them ideal for plant breeding and genetics (Parida *et al*., 2009). Majority of microsatellites are found in non-coding DNA, either in intergenic regions or in introns, and these comprise the microsatellites that are mostly used as markers (Ellegren, 2004). Their abundance in the genome coupled with the possibility of associating them with many phenotypes makes them powerful for applications in diverse areas in plant genetics (Victoria *et al*., 2011).

2.4 Classification of microsatellites

Microsatellites have been classified in a number of ways depending on their size, location in the genome and type of repeat motif. Repeats of one, two, three, four, five and six nucleotides comprise what are classified as microsatellites. Longer repeat units form minisatellites and in extreme cases, satellite DNA (Ellegren, 2004). The number of nucleotides per repeat unit yields the classification of microsatellites as mononucleotides $(A)_n$, dinucleotide $(CA)_n$, trinucleotide $(CGT)_n$, tetranucleotide $(CAGA)_n$, pentanucleotide $(AAATT)_n$ or hexanucleotide $(CTTTAA)_n$ repeats (Kalia *et al*., 2011).

Microsatellites have been further described by Wang and co-workers (2009) as follows; simple perfect repeats, which are tandem arrays of a single repeat sequence such as $[AGG]_n$, simple imperfect arrays, which consist of one or more repeat units of different lengths such as $[AAC]_n[ACT][AAC]_{n+1}$, compound perfect arrays, which are two or more different repeat motifs of the same length such as, $[AGG]_n$ $[AATC]_n$ and compound imperfect motifs, which are interrupted by one or more repeats of different lengths such as $[GGAT]_n[ACT][GTAA]_{n+1}$. According to their location in the genome, microsatellites have been classified as nuclear (nuSSR), mitochondrial (mtSSR) or chloroplast (cpSSR).

2.5 Origin of microsatellite polymorphisms in the genome

The molecular mechanisms underlying the development of microsatellite variation are still not yet completely understood (Anmarkrud *et al*., 2008). Microsatellite polymorphisms are mainly as a result of variability in length polymorphism rather than

in the primary sequence. This is in contrast to unique DNA (Ellegren, 2004). Replication slippage is the primary mechanism reputed to underlie the change in length in microsatellite DNA (Levinson and Gutman, 1987).

Following the commencement of replication of a repeat region, the two strands may dissociate. After realignment occurs, the nascent strand may be out of register, resulting in the nascent strand having a different length from the template strand. An increase in repeat length comes about if a loop is introduced on the nascent strand. On the other hand, if a loop is introduced on the template strand, the consequence is a decrease in repeat length. This is illustrated in Figure 1.

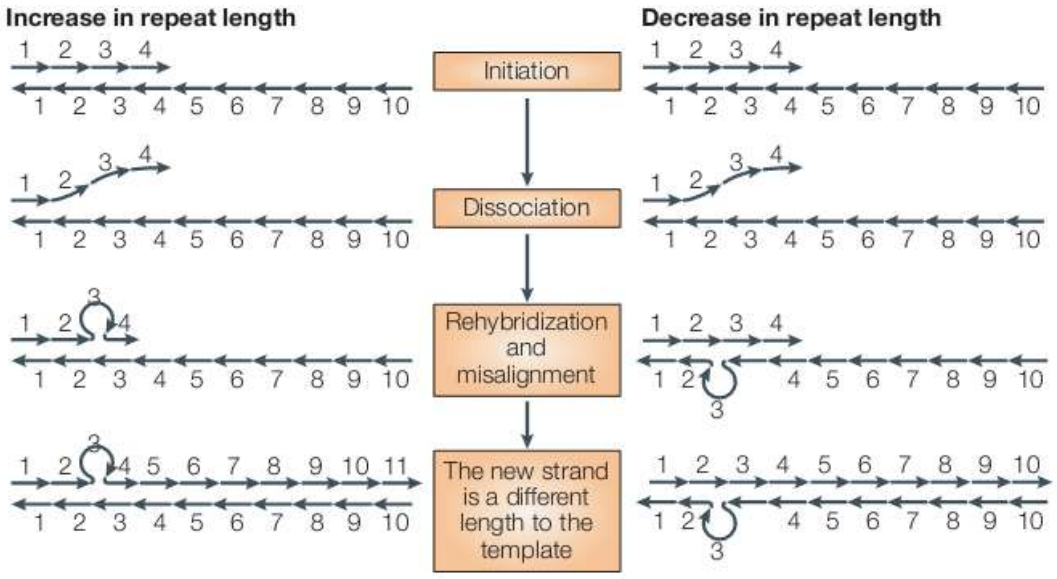


Figure 1: Following the commencement of replication of a repeat region, the two strands might dissociate. Alignment of the nascent strand out of register leads to a different length from the template strand. A loop introduced on the nascent strand results in an increase in repeat length. A loop that is formed in the template strand leads to a decrease in repeat length (Ellegren, 2004).

Ordinarily, the mutation events would be corrected by mismatch repair as well as exonucleolytic proof reading. However, if this fails to occur, variation in the microsatellite may result. Therefore, microsatellite instability can be seen as a balance between the generation of replication anomalies via slip-strand mispairing and the correction of some of these errors through exonucleolytic proof reading and mismatch repair (Strand *et al*., 1993; Li *et al*., 2002; Ellegren, 2004). The enzymatic activity that results in replication slippage is that of DNA polymerase. The process involves the DNA polymerase momentary halting and dissociating from the DNA. When dissociation occurs, only the end portion of the nascent strand separates from the template and later anneals to another repeat unit (Hile and Eckert, 2004).

A second model that has been cited to explain the rise of microsatellite polymorphism is unequal recombination (Chistiakov, 2005). This is also referred to as recombination by unequal cross over or gene conversion and is also responsible for expansion and contraction of repeat length (Richard and Paques, 2000). The mechanism of recombination has been particularly studied in yeast and humans, where trinucleotide repeat expansions have been reported to be responsible for neurological diseases. Richard and Paques (2000) state that the occurrence of these expansions during the lifetime is not precisely known. They suggest that it might be meiotic pre- or post-zygotic. However, it is worth noting that there seems to be no literature siting studies of this mechanism in plants

2.6 Genomic distribution of microsatellites

Microsatellites are largely located in non-coding DNA. However, they can also be found in transcribed regions, albeit in lower proportions (Kalia *et al*., 2011). The lower frequency of SSRs in coding regions can be attributed to selection against frame shift mutations in coding regions (Li *et al*., 2002).The location of a microsatellite in the genome determines its effect and can thus have an impact on all aspects of genetic function. For instance, a microsatellite in the coding region can affect the expression of a protein (Kalia *et al*., 2011). The expansion of SSRs in the 3'-UTR can cause transcription slippage, resulting in expansion of mRNA which can disrupt splicing and other cellular functions. Microsatellites in the introns can act in regulating gene transcription and translation in addition to acting as a co-regulator for gene expression in conjunction with SSRs located in the 5'-UTR. The effects of microsatellite expansion and contraction can ultimately lead to change in phenotype (Li *et al*., 2004).

2.7 Targeting microsatellites

To target microsatellites, 18-24 bp primers are designed that are specific to either side that flanks the repeat element. This is possible because the flanking regions of microsatellites are usually conserved. These microsatellite loci are easily amplified by PCR and the difference in repeat length among individuals in a population can be analyzed by gel electrophoresis (Wang *et al*., 2009). Additionally, the PCR products can be analyzed by automated sequencers to facilitate the detection of microsatellites. This is made possible by incorporating a fluorophore into the PCR product. The fluorophore absorbs a specified wavelength of light emitted by a diode laser. The light emitted by the

fluorophore is captured by a detector which makes a digital record of the light excitation in the form of a band on a gel or chromatogram (Wang *et al*., 2009).

When ordering the primers, the M13 labeling strategy can be specified. This is a cost effective strategy of utilizing a fluorescent labeled primer that universally works with each SSR primer set as described by Schuelke (2000). To accomplish this, 3 primers are required: a forward primer, specific to the target template DNA with M13 (-19bp) tail at its 5' end, a reverse primer specific to the template sequence and a universal fluorescent labeled M13 (-19bp) primer. Typically, the amount of forward primer used is less than half the amount of reverse primer. When setting up the PCR, conditions are selected that allow the forward primer with the M13 (-19bp) sequence to be incorporated into the PCR product during the first cycles. Once the forward primer is exhausted, the universal M13 (-19bp) primer anneals, enabling the universal fluorescent labeled M13 (-19bp) primer to become the forward primer. In this process, the fluorescent dye is incorporated into the PCR product. This is illustrated in the Figure 2.

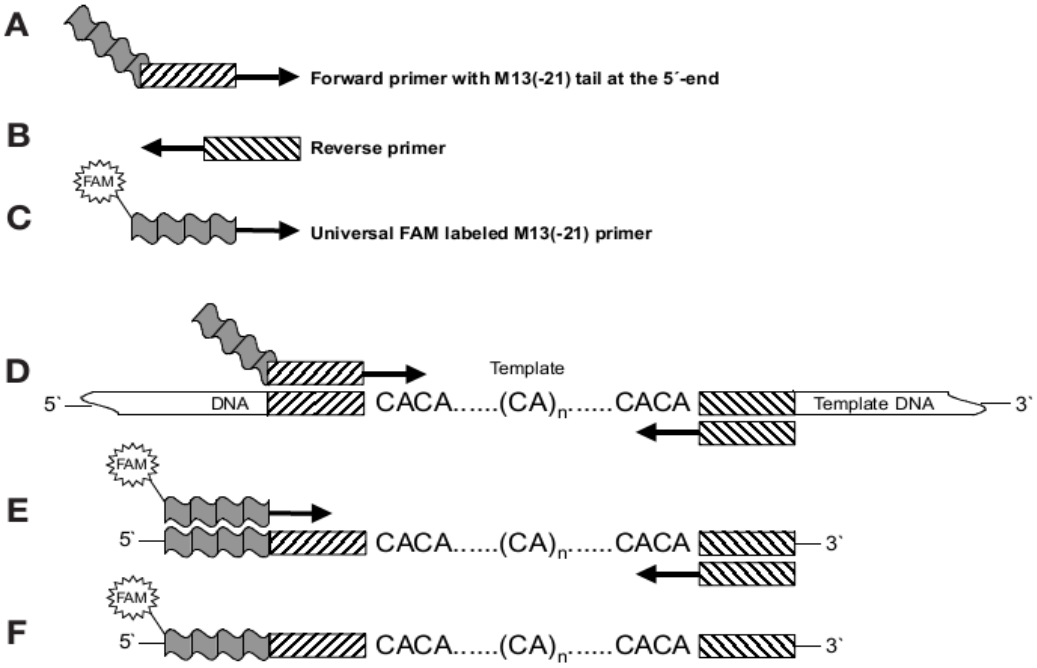


Figure 2: A and B; the crossed boxes represent the microsatellite-specific primers. C; the gray wavy box represents the universal M13(-21) sequence, and the star the fluorescent FAM label. D; In the first PCR cycles, the forward primer with the M13(-21) tail is incorporated into the PCR products. E and F; the products are amplified with the FAM-labeled universal M13(-21) primer being the forward primer resulting in the final PCR product being fluorescent labeled (Schuelke, 2000).

By using four different fluorescent dyes, that is VIC, FAM, PET and NED, it is possible to co-load up to four sets of markers together thus drastically reducing the cost of separation and detection of microsatellites by capillary electrophoresis.

2.8 Applications of microsatellites in plant research

There are diverse applications of microsatellite markers in plant research. Their hypervariable nature and extensive genome coverage makes them the marker of choice for many applications in plant research (Kalia *et al*., 2011). These include genome

mapping, cultivar identification, MAS, genetic diversity studies, phylogenetic relationships and population studies (Wang *et al*., 2009).

2.8.1 Genome mapping

According to Wang *et al*. (2009), genome mapping comprises genetic mapping, comparative mapping, physical mapping and association mapping. More than 80 genome maps have been constructed so far for various plant species using microsatellites. These include rice (Varshney *et al*., 2005), wheat (Yu *et al*., 2004; Breseghello and Sorrells, 2006a and Hiebert *et al*., 2007), barley (Ramsey *et al*., 2000; Marcel *et al*., 2007 and Varshney *et al*., 2007), cotton (Guo *et al*., 2007; Yu *et al*., 2007 and Yu *et al*., 2011), ryegrass (Studer *et al*., 2010), white clover (Zhang *et al*., 2007), potato (Bjorn *et al*., 2008) and spinach (Khattak *et al*., 2006).

Wang *et al*. (2009), describes comparative mapping as the alignment of chromosomal fragments of various related species based on genetic mapping of common DNA markers. It can facilitate the identification of major gene syntenies, chromosome re-arrangements and micro-syntenies (i.e., conserved gene order in the chromosomes of different species) between species. Major and micro-syntenies can be useful in developing markers for specific chromosomal regions which can further facilitate MAS. Physical maps of genomes are able to provide the actual physical distance between markers or genes in base pairs. They enable the assembling of genome DNA sequences and positional cloning (Wang *et al*., 2009).

Microsatellite markers can also be used for association mapping. This is a technique that relies on linkage disequilibrium to study the relationship between a phenotypic variation and genetic polymorphism. Association mapping can be used to implement MAS in plant breeding programs, especially in conjunction with microsatellite markers that have been clearly linked to particular phenotypes of interest (Breseghello and Sorrells, 2006b; Wang *et al*., 2009).

2.8.2 Marker assisted selection

Marker-assisted selection (MAS) entails the use of genetic markers to follow genomic regions that encode specific characteristics of a plant (Barr, 2009). The efficiency of plant breeding programs can be extensively improved through SSR-MAS. In the selection process, two types of SSRs can be used; those closely linked to a locus of a trait, that is, flanking SSRs, or those developed within the target region itself. SSRs within the target gene are definitely more efficient; however they require more effort to develop (Wang *et al*., 2009). Through MAS, it is possible to bypass conventional phenotypic selection of plants in the field, thus greatly increasing efficiency and precision of breeding programs (Kalia *et al*., 2011).

An example of how microsatellites have been successfully used in MAS is in pea breeding, where researchers were able to select for powdery mildew resistance with 98.4% success rate being achieved (Ek *et al*., 2005). SSRs have also been used in MAS for *Fusarium* head blight resistance in wheat. This is a disease that has been extensively studied because of its potential to cause massive losses in grain yield, reduction in

baking and seed quality and contamination with mycotoxins (Buerstmayr *et al*., 2009). Another example of MAS is the use of SSRs for conversion of normal maize lines into quality protein maize. These are rich in lysine and tryptophan (Danson *et al*., 2006; Brumlop and Finckh, 2011). In combination with marker assisted back crossing (MABC), MAS has been extensively used to introgress submergence tolerance (sub 1) QTL in rice. This has enabled the development of rice varieties that are tolerant to submergence while maintaining characteristics that are preferred by farmers and consumers (Iftekharuddaula *et al*., 2011).

2.8.3 Genetic diversity

Genetic diversity refers to any variation in nucleotides, genes, chromosomes or whole genomes. Microsatellites provide a powerful marker system that can be used to deduce genetic diversity within and between species. Information from genetic diversity studies can be instrumental in choosing parental lines for breeding programs as well as for the classification of plant germplasm accessions (Wang *et al*., 2009). Diversity analysis information is also important for crop conservation, development of new varieties and taxonomic studies (Kalia *et al*., 2011).

2.8.4 Population studies

The genetic structure of natural populations can be affected by events such as seed dispersal, pollen flow and plant introduction and domestication by humans. Using microsatellites, it is possible to elucidate the population structure within and among natural populations and to identify potential progenitors (Wang *et al*., 2009). This is

exemplified in the study by Spooner *et al*. (2007), where they used 50 microsatellites to genotype 750 potato accessions. The results confirmed the reclassification of cultivated potato into 4 species, namely *S. tuberosum, S. ajanhuiri, S. juzepczuki* and *S. curtilobum.*

Organelle specific microsatellite; that is cpSSR and mtSSR exhibit uniparental inheritance, conserved gene order, lack of heteroplasmy (i.e., lack of a mixture of organelle genomes) and lack of recombination of organelle genomes. As such, they can be used to determine structure and variation within natural populations, as well as phylogenetic relationships (Kalia *et al*., 2011).

2.9 Challenges of microsatellite isolation

Despite their great utility and applicability in plant genetics, microsatellites have not been widely exploited in most species, particularly the minor crops such as finger millet. This is attributed to the fact that they need to be isolated *de novo* from most species being examined for the first time (Kalia *et al*., 2011). This is because microsatellites are usually found in non-coding regions where the nucleotide substitution is much higher than in coding regions, thus making it more difficult to design universal primers for microsatellites that match conserved sequences (Zane *et al*., 2002). Additionally, large scale isolation of microsatellites is technically challenging because of their relatively low frequency in plant genomes. Also, the palindromic nature of A-T dinucleotides, which are the most common SSRs in plants, presents challenges in isolating them from libraries (Powell *et al*., 1996).

2.10 Strategies of microsatellite isolation

The classical method of microsatellite isolation has involved the use of primers already developed for other species (Csencsics *et al*., 2010). Alternatively, it has been necessary to create enriched microsatellite libraries, a task that involves cloning, hybridization to detect positive clones, plasmid isolation and Sanger sequencing (Castoe *et al*., 2009). This approach was used to generate the few (82) published SSRs for finger millet (Dida *et al*., 2007).

Large scale Sanger sequencing is often very expensive and extremely labor intensive, involving sub cloning of the DNA into vectors and amplification in hosts. The methodology involves reading of the sequences from DNA fragments of different lengths. These are generated by a DNA polymerase that ceases incorporating nucleotides whenever it encounters a labeled terminator (Brautigam and Gowik, 2010). After sequencing, primers are designed for use in locus specific PCR analysis and identification of polymorphism (Powell *et al*., 1996). Research has led to improvement over the standard isolation methods and development of alternative procedures that are less time consuming but result in significant increase in microsatellite yield (Zane *et al*., 2000; Kalia *et al*., 2011).

It is possible to significantly lower the cost of microsatellite isolation by incorporating next-generation sequencing (NGS) in place of the hitherto conventional Sanger sequencing. This is in view of the ever-decreasing cost of NGS approaches. Additionally, NGS based approaches circumvent the cloning step, making it possible to

sequence more fragments than those limited by successful cloning into plasmid vectors during library construction in Sanger sequencing (Castoe *et al*., 2009). follow

2.11 Next generation sequencing (NGS)

NGS refers to newer methods of sequencing strategies that offer the ability to produce large numbers of sequences at reduced cost and time as compared to Sanger sequencing (Csencsics *et al*., 2010; Metzker, 2010). The reduction in cost and labour potentially enables researchers to develop larger number of microsatellites for use in studying non-model plants (Csencsics *et al*., 2010). Unlike in the case of Sanger sequencing, for which the nucleic acids have to be sub cloned into vectors and amplified in hosts, all NGS platforms avoid this step and the DNA is directly sequenced (Brautigam and Gowik, 2010).

Commercially available NGS platforms include the Genome Sequencer FLX and Titanium from 454 life sciences/Roche, the Illumina Genome Analyzer from Solexa and SOLiD (Sequencing by Oligo Ligation and Detection) from Applied Biosystems. All these are regarded as second-generation NGS platforms because they require an emulsion PCR (emPCR) amplification step prior to sequencing (Brautigam and Gowik, 2010). Third generation NGS platforms feature the latest development in sequencing technology. Their distinctive feature is their ability to sequence single DNA molecules without a prior amplification step. This is referred to as single molecule sequencing. Sequencers under this category include; Heliscope single molecule sequencer, Single

molecule real time (SMRT) sequencer, Nanopore DNA sequencer and the Ion torrent sequencer (Pareek *et al*., 2011).

This study used the Roche 454 GS-FLX Titanium platform. It is based on sequencing by synthesis and is able to sequence up to 0.1 gigabases (10^9) per run with a mean read length of between 300-400 nucleotides, which is sufficient for isolating microsatellites together with enough flanking sequence for primer design (Allentoft *et al*., 2009). Despite the read length being shorter than in Sanger sequencing, the number of copies per read (sequencing depth) is high enough to provide confidence in the sequence results and the cost of producing the sequences is greatly diminished ($90/MB) when compared to traditional sequencing ($1000/MB) (Wall *et al*., 2009).

2.12 Roche/454 sequencing

The Roche/454 sequencing platform was the first commercially available NGS platform (Mardis, 2008). 454 sequencing utilizes the technique of sequencing by synthesis. This differs from the classical chain termination Sanger sequencing in that, it depends on the detection of the release of a pyrophosphate whenever a nucleotide is incorporated; hence, it is also referred to as pyrosequencing (Shekhar *et al*., 2011).

In pyrosequencing, prior to sequencing, fragments of 300-800bp are attached to two different adaptors at either end. These fragments are captured on micro beads on which emPCR is performed to create copies of each fragment. The micro beads are then loaded into microwells on a PicoTitre plate where the pyrosequencing reaction takes place.

During the process of pyrosequencing, a dNTP is incorporated into a growing DNA strand by DNA polymerase. This results in the α phosphate of the dNTP becoming part of the phospho-diester backbone and the release of the β and γ phosphates in the form of pyrophosphate (ppi). ATP sulfurylase converts the ppi into ATP in the presence of adenosine phosphosulphate. Luciferase, a firefly enzyme utilizes the ATP to convert luciferin to oxyluciferin resulting in the emission of light that can be detected by a luminometer connected to a charge coupled device (CCD) camera. The amount of light produced is proportional to the number of nucleotides incorporated (Mardis, 2008; Turner, 2011; Zalapa *et al*., 2012). The general overview of 454 sequencing is illustrated in Figure 3.

The use of NGS to identify microsatellites is relatively new and only a few studies have reported this approach comprehensively (Csencsics *et al*., 2010). In most of the studies, 454 is the dominant platform used for SSR isolation. This is attributed to the fact that the read length of between 350-600 bp per read is sufficiently long to allow detection of SSRs directly from raw reads (Zalapa *et al*., 2012)

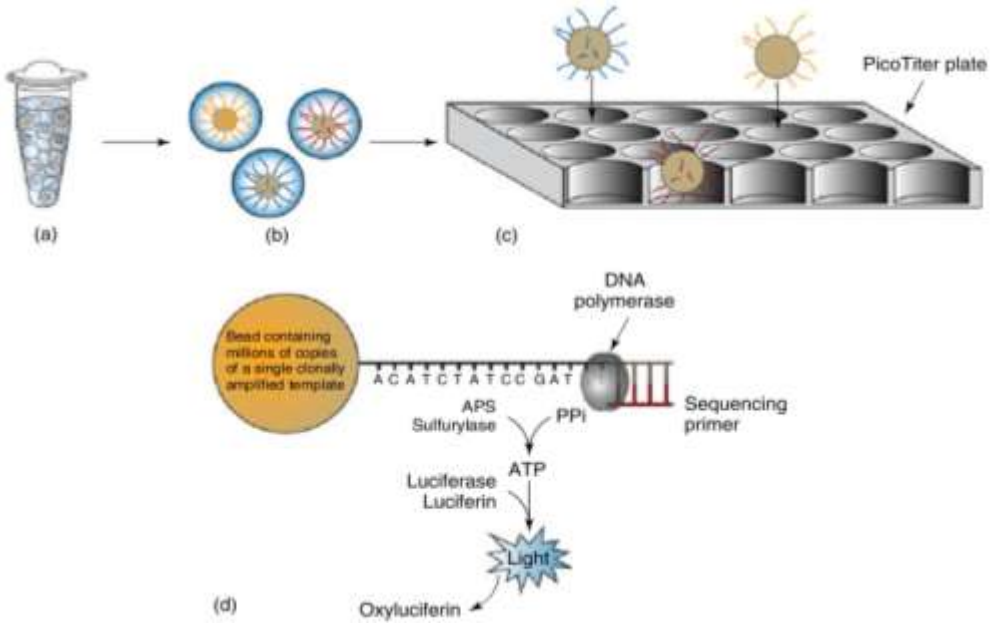


Figure 3: in (a), beads, templates, and amplification reagents are emulsified, creating aqueous compartments that contain single template molecules. In (b) template molecules are amplified onto beads through emPCR. In (c), individual beads are added to wells on a PicoTiter Plate. In (d) bead-bound templates are sequenced by pyrosequencing (Turner, 2011).

2.13 Methods of Microsatellite isolation using NGS

Using NGS platforms such as Roche 454 GS-FLX Titanium as is the case in this study, it is feasible to detect microsatellites by shotgun sequencing, whereby candidate microsatellites are identified from a set of randomly sampled shot gun reads as done in the studies by Abdelkrim *et al*. (2009), Allentoft *et al*. (2009) and Castoe *et al*. (2010). Alternatively, an enrichment step for specific microsatellites can be incorporated prior to sequencing (Malausa *et al*., 2011).

Castoe *et al*. (2010) points out drawbacks of using the enrichment based approach for isolating microsatellites. Since enrichment requires *a priori* choices about the types of microsatellite loci to target, there are biases as to which microsatellites are identified. This limits the diversity of the microsatellites identified to only a small subset of all possible motifs. Additionally, uninformed *a priori* choices about what motifs to target may limit the success of obtaining ample numbers of useful microsatellite loci. However, after carrying out a comparative study of both approaches, Malausa *et al*. (2011) point out some advantages of using enrichment. In their study, enrichment resulted in increased number of microsatellite loci isolated while reducing the proportion of unwanted motifs such as AT motifs which present difficulty during amplification. Based on their findings, enrichment improved isolation efficiency by close to 300%.

Secondly they suggest that enrichment increases the number of multiple reads obtained for a particular microsatellite locus, thus it is possible to design primers targeting non polymorphic sequences that flank microsatellite motifs. This diminishes the chances of designing markers with a high percentage of null alleles owing to mismatches between primers and polymorphic nucleotides in flanking regions that can occur in some individuals or populations. In light of this, our choice of the enrichment option found merit (Malausa *et al*., 2011).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Library enrichment and sequencing

Library enrichment and 454 sequencing were carried out at Ecogenics GmbH, Switzerland. Two finger millet genotypes; KNE 755 which is susceptible to finger millet blast and KNE 796 which is tolerant to finger millet blast (H. Ojulong, ICRISAT, pers comm.), were used. Leaves of each genotype were sampled 2-3 weeks after planting. These were dried on silica gel, packaged and sent to Ecogenics GmbH for genomic DNA extraction, library enrichment for two dinucleotide motifs, namely, CT and GT, and 454 GS-FLX Titanium sequencing.

3.2 Sequence assembly and SSR mining

The sequencing service provider supplied raw 454 GS-FLX Titanium sequences as well as possible primer sequences. From the raw sequences, high quality nucleotides (contained in the .fna data files) were selected using the quality scores (.qual data files). The selection was done using NGS QC (Patel and Jain, 2012). Sequence length of 80bp and a phred score of 20 were regarded as the threshold for high quality sequences. In the absence of a reference genome, *de novo* assembly was performed using MIRA (Chevreux *et al*., 2004). The output from MIRA as well as the remaining single reads, were fed into SciRoKo 3.4 (Kofler *et al*., 2007) for SSR mining and extraction of flanking sequences.

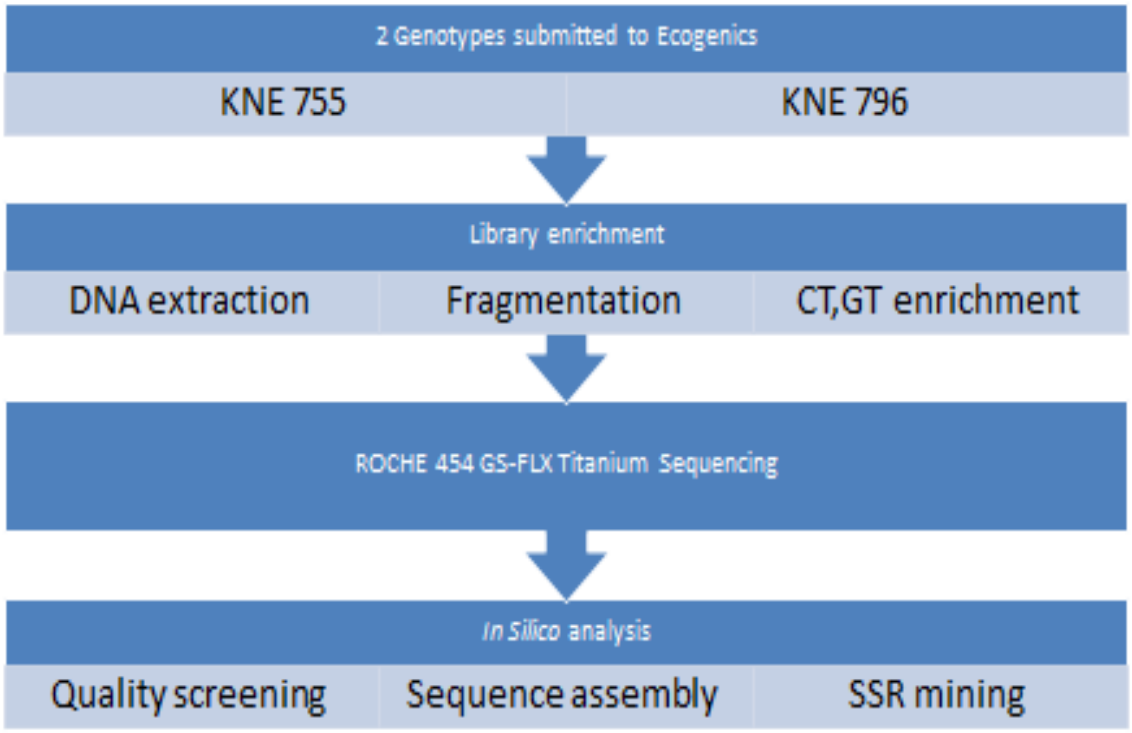


Figure 4: Flowchart showing the process of *in silico* analysis of the sequences supplied by Ecogenics.

3.3 Primer design and analysis

Primers for the microsatellites identified were designed using BatchPrimer3 (You *et al*., 2008), a primer design tool based on Primer 3 (Rozen and Skaletsky 2000) that can accept input sequences in batches of up to 500 sequences at a time. After eliminating repeated primers, the designed primers were analysed using PrimerAnalyser (Kalendar *et al*., 2011). This tool was used to filter out primers with less than 40% GC content and to eliminate those that could potentially form primer dimers. Additionally, the tool was used to measure primer efficiency, with a selection being made for primers with greater than 75% efficiency. The selected primers as well as the Ecogenics primers were subjected to *in silico* PCR using CLC Genomics Workbench v 5 (www.clcbio.com). In

the absence of a reference genome, the primers were tested against the sequences provided by the sequencing service provider.

3.4 Comparison of primers

Microsoft Excel and ExamDiff v 1.9, which is a tool for visual file comparison, were used to compare the designed primers with those supplied by the sequencing service provider. A further comparison of all the new primer sequences was made against the published microsatellite primers (Dida *et al*., 2007).

3.5 Primer evaluation

3.5.1 DNA extraction

Ten different finger millet genotypes, mostly from East Africa were selected from the ICRISAT mini-core collection that is broadly representative of the global finger millet diversity. Details of these genotypes that were used for primer evaluation are summarised in Table 1.

Table 1: Details of the genotypes used for validating the primers

Sample No.	**Acc/local name**	**Country**	**Region**
253	GBK-044047A	Kenya	Laikipia
124	GBK-000414A	Kenya	South Nyanza
169	GBK-011135A	Kenya	Machakos
386	Sansamula	Tanzania	Sumbawanga
408	Namakonta	Tanzania	Sumbawanga
5	Ebega	Uganda	Serere
77	Bulo	Uganda	Hoima
33	Emorumoru(rock)	Uganda	Amuria
271	IE2572*		
281	IE2957*		

*From ICRISAT Minicore collection

DNA was extracted from leaves of 10 different individuals from each genotype of 2-3 week old seedlings according to Mace *et al*. (2003) but omitting the phenol:chloroform extraction step. Each sample was cut into ≤0.5cm pieces and put into a 2 ml eppendorf micro centrifuge tube. 450µl of preheated (65ºC) extraction buffer (100mM Tris-HCl [pH 8], 20mM EDTA, 2-3% w/v CTAB, 1.4M NaCl) together with β-Mercapto-ethanol (0.03-3% v/v) was added to each of the samples and the latter ground using steel balls and a Tissue Lyser II (Qiagen®, Hilden, Germany*)*. The homogenate underwent solvent

extraction using 450µl of chloroform: isoamylalcohol (24:1) and the crude DNA pellet for each sample was precipitated using isopropanol (0.7 vol, -20ºC). 3µl of ribonuclease-A (10mg/ml) was used to remove RNA. A second solvent extraction was done and the final DNA pellet precipitated using 315 µl of ethanol-acetate solution (30ml EtOH, 1.5ml 3M NaOAc [pH 5.2]). The pellets were cleaned using 70% ethanol and re-suspended in low salt TE buffer (10mM Tris, 0.1mM EDTA [pH 8]). Extracted DNA was visualised on a 0.8% (w/v) agarose gel and quantified spectrophotometrically using a Nanodrop® 1000 (Thermo Scientific, Florida, USA), followed by dilution to 10 ng/µl in TE buffer (10mM Tris, 0.1mM EDTA pH 8.0). The DNA was stored at 4°C.

3.5.2 PCR reaction

Due to cost constraints and since these primers were received along with the sequencing results from the provider, PCR reactions were carried out using primer pairs that were synthesized for 101 of the 178 recommended SSRs that were supplied by Ecogenics.

The primers for KNE 755 and KNE 796 were combined and collectively named ICECP (ICRISAT *Eleusine coracana* Primer). All forward primers contained a 19 bp M13-tag (5'- CACGACGTTGTAAAACGAC - 3') on the 5' end that were fluorescently labelled to allow detection of amplification products (Schuelke, 2000). PCR amplification was performed in 10 µl 384 well microtitre plates (AB Gene, USA) and each reaction comprised of 1 x PCR buffer (20 mM Tris-HCl, pH 7.6; 100 mM KCl; 0.1 mM EDTA; 1 mM DTT; 0.5% (w/v) Triton X-100; 50% (v/v) glycerol), 2 mM $MgCl_2$, 0.16 mM dNTPs, 0.16 µM fluorescent labeled M13-forward primer, 0.04 µM forward primer, 0.2 µM reverse primer, 0.2 units of Taq DNA polymerase (SibEnzyme Ltd, Novosibirsk,

Russia) and 30ng of template DNA. PCR reactions were performed on a GeneAmp 9700® thermocycler (Applied Biosystems, California, USA) with initial denaturation of 94°C for 5 minutes, followed by 35 cycles of 94°C for 30 seconds, 59°C for 1 minute and 72°C for 2 minutes, followed by final elongation at 72°C for 20 minutes. Amplification was confirmed by running 3 µl of the products on a 2% (w/v) agarose gel stained with GelRed® (Biotium, California, USA) and visualized under UV light.

Amplification products (2.5 µl–3.5 µl of each) were co-loaded in sets of 3 to 4 markers together with the internal size standard, GeneScan™ –500 LIZ® (Applied Biosystems, California, USA) and Hi-Di™ Formamide (Applied Biosystems, California, USA) and separated by capillary electrophoresis using an ABI Prism® 3730 Genetic analyzer (Applied Biosystems, California, USA). Allele calling was performed with Gene Mapper 4.0 (Applied Biosystems, California, USA).

3.5.3 BLAST search

All the sequences were subjected to a BLASTX search against gene ontology databases using the tool Blast2GO (Conesa *et al*., 2005). This was done in order to determine if there were sequences that could be linked to both an SSR and a functional gene.

3.6 Data analysis

Allelic data generated by Gene Mapper 4.0 (Applied Biosystems, California, USA) was subjected to statistical analysis with PowerMarker V3.25 (Liu and Muse, 2005) to reveal the level of polymorphism of the markers among the selected individuals. The software

was able to calculate the polymorphic information content (PIC) for each marker. This describes the usefulness of a marker in detecting polymorphism within a population and depends on the number of detectable alleles and their frequency distribution (Botstein *et al*., 1980)

CHAPTER FOUR

RESULTS

4.1 Output following library enrichment and sequencing

Ecogenics GmbH, supplied data files that included; fasta sequences (.fna file), quality scores (.qual file) and possible SSR primers. Table 2 gives a summary of the data obtained after enrichment and sequencing.

Table 2: Sequence data supplied by Ecogenics

Genotype	**No. of sequences**	**Max seq length**	**Min seqlenth**	**Average seq length**	**N50 seq length**	**No. of SSR primers**
KNE 755	6106	479	23	125	144	45
KNE 796	5357	517	25	140	159	56

4.2 Quality assessment of sequences

The NGS QC tool, which is a command line software, required that the paths of both the .fna (i.e., the fasta sequences) and .qual (i.e., the quality scores) files be specified as input. The QC parameters that were specified, required discarding of sequences that were less than 80bp long as well as those with a phred score (i.e., scores for evaluating the quality of individual bases in the sequences) of less than 20. For KNE 755, sequences of low quality were 27, while sequences shorter than 80bp were 1344. For KNE 796, low quality sequences were 24 while sequences shorter than 80bp were 884.

Thus the selected high quality sequences were 4735 and 4448 for KNE 755 and KNE 796 respectively. These results are summarized in Table 3.

Table 3: Quality assessment results for the sequences.

Genotype	**Original no. of seq**	**Low quality seq**	**Seq shorter than 80bp**	**High quality seq**
KNE 755	6106	27	1344	4735 (77.55%)
KNE 796	5357	24	884	4448 (83.03%)

4.3 Sequence assembly

Following assembly by MIRA, 629 reads from KNE 755 were assembled into 46 contigs, with the largest contig being 816 bp. For KNE 796, 523 reads were assembled into 41 contigs with the largest contig being 552 bp (Table 4).

Table 4: Results obtained after assembling of the sequences.

Genotype	**No. of assembled reads**	**No. of contigs**	**Largest contig (bp)**	**N50 contig size**	**N95 contig size**
KNE 755	629	46	816	292	131
KNE 796	523	41	552	314	151

4.4 SSR mining and extraction of flanking sequences

SSR mining by SciRoKo was performed in MISA mode. After mining, sequences containing SSRs with sufficient flanking sequence for designing primers of 18-25bp were selected as presented in Table 5.

Table 5: Results obtained after mining of SSRs.

Genotype	**No. of SSRs**	**SSRs with flanking sequence**	**(% of total)**
KNE 755 contigs	55	55	100
KNE 796 contigs	26	26	100
KNE 755 single reads	1224	767	62.66
KNE 796 single reads	1230	790	64.22

4.5 Primer design and analysis

Primer design using BatchPrimer3 yielded a total of 283 primer pairs for KNE 755 and 274 primer pairs for KNE 796. These included 32 and 9 primer pairs for the contigs of KNE 755 and KNE 796, respectively. Out of these, 113 KNE 755 primers had less than 40% GC content, 86 had dimers while 42 had efficiency of less than 75%. For KNE 796, 94 had less than 40% GC content, 82 had dimers and 48 had efficiency of less than 75%.

Thus, the selected good primers were 42 and 50 for KNE 755 and KNE 796 respectively (Table 6). *In silico* PCR results showed that 29 and 40 KNE 755 and KNE 796 primers supplied by Ecogenics could amplify DNA, translating to 64.4% and 71.4% of the primers respectively. For the new primers, 33 and 43 KNE 755 and KNE 796 primers could potentially amplify DNA. These were 78.6% and 86.0% of the primers respectively (Table 8).

Table 6: Primers selected after analysis

Genotype	**Primer sets identified from BP3***	**< 40% GC contents**	**Dimers**	**< 75% efficiency**	**Good primers**	**% Good primers**
KNE 755	283	113	86	42	42	15
KNE 796	274	94	82	48	50	18

*BP3 – BatchPrimer3

From assembled contig sequences, only 3 and 1 pair of good primers were obtained for KNE 755 and KNE 796, respectively. Table 7 shows a comparison of the numbers of the new primer pairs with those from the sequencing service provider.

Table 7: Comparison of the numbers of new primers and the Ecogenics primers identified.

Source	**KNE 755**	**KNE 796**	**Total**
New primers	42	50	92
Ecogenics primers	45	56	101

Table 8: Amplification potential of the primers.

Genotype	**Ecogenics amplifying primers**	**Total**	**Percent**	**New amplifying primers**	**Total**	**Percent**
KNE 755	29	45	64.4	33	42	78.6
KNE 796	40	56	71.4	43	50	86.0

4.6 DNA extraction

The concentration of extracted DNA ranged from 29.4 ng/µl to 106.8 ng/µl. The optical density ratio (260nm/280nm), which is a measure of purity of the DNA, ranged from 1.75 to 1.94, where pure DNA should have a ratio of 1.8.

4.7 PCR and capillary electrophoresis

Figure 11 shows a 2% (w/v) agarose gel image captured following electrophoretic seperation of the PCR products. Some of the markers could clearly amplify genomic regions. These were revealed by clear sharp bands, for example ICECP 9 (marked by a red ellipse) amplified a 200bp fragment (200bp indicated by the blue ellipse on the ladder). On the other hand, smears were indicative of nonspecific amplification while very low molecular weight bands revealed presence of primer dimers and/or failure of amplification. ICECP 1 was an example of a smear (indicated by the green ellipse), while ICECP 8 was an example of a low molecular weight band (indicated by a yellow ellipse). Despite the fact that not all the samples amplified during PCR, all were loaded onto the sequencer as it would have been too time-consuming to select the small number of samples that did not work from the 96 well plate in which the PCR was performed.

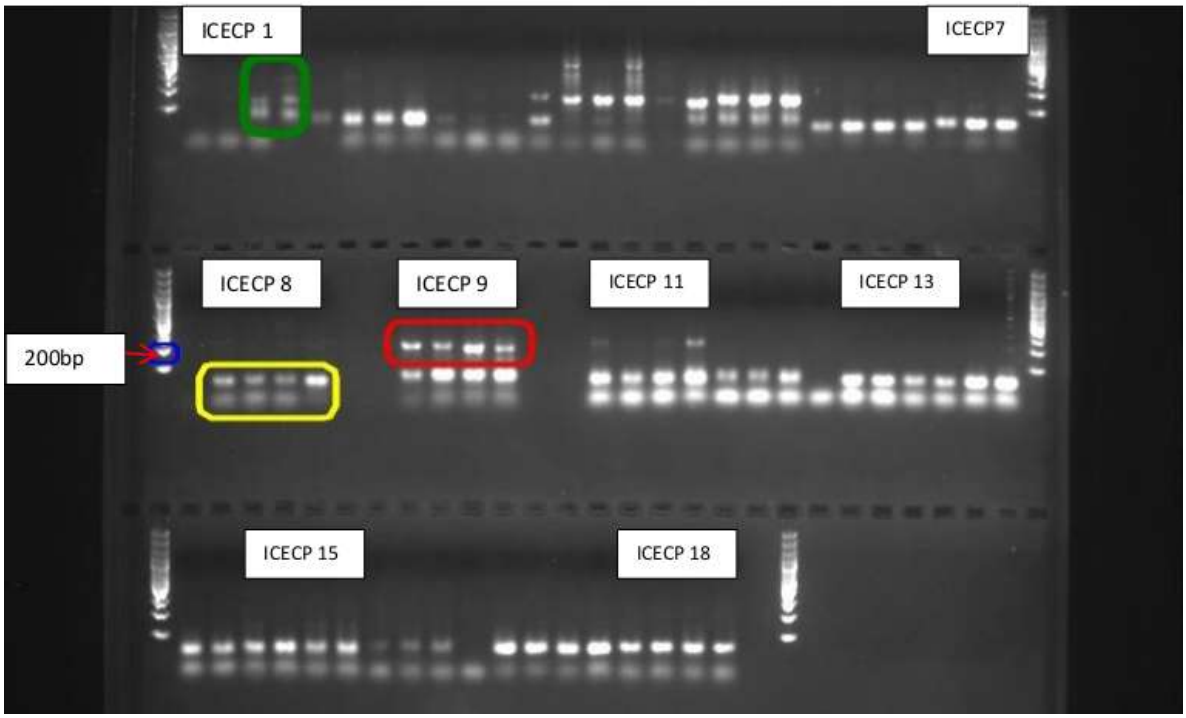


Figure 5: An agarose gel image of the amplification products of the markers ICECP 1- ICECP 9 and ICECP 11- ICECP 18. The markers were loaded in groups of 4, with ICECP 1 represented by the first four wells at the top left corner of the gel.

Separation of the amplicons was carried out by capillary electrophoresis. The raw data generated was analyzed using Gene Mapper 4.0 (Applied Biosystems, California, USA). This was to facilitate allele calling and identification of the various characteristics of the markers as illustrated in Figure 12, Figure 13, Figure 14 and Figure 15. Some of the markers were mono-allelic, i.e., amplified only one allele. Di-allelic and multi-allelic markers amplified two alleles and multiple alleles respectively. Gene Mapper analysis also enabled the discernment of markers that were monomorphic and polymorphic. Monomorphic markers exhibited no variations within the genotypes being studied while polymorphic markers showed variations. The best markers were both mono-allelic and polymorphic as depicted in Figure 13. The genotype that amplified best (most efficiently

and most frequently) was 77 with 52.5% of the markers amplifying. The genotypes with the least number of amplifying markers were; 169, 271 and 281. 44.5% of the markers amplified in each of these genotypes (Table 10).

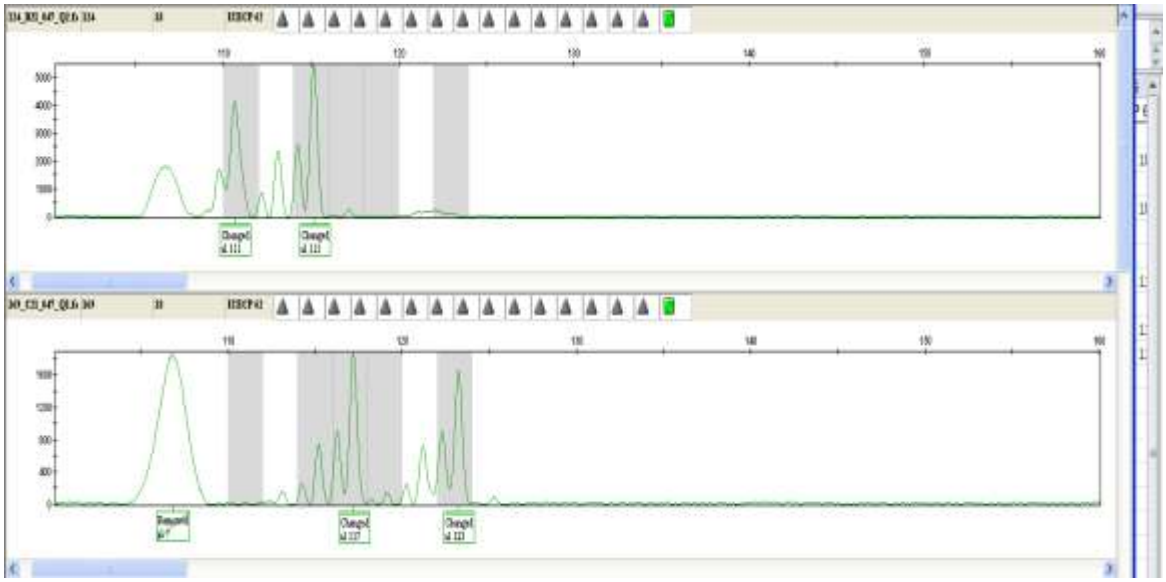


Figure 6: Gene Mapper screen shot showing an example of a di-allelic marker

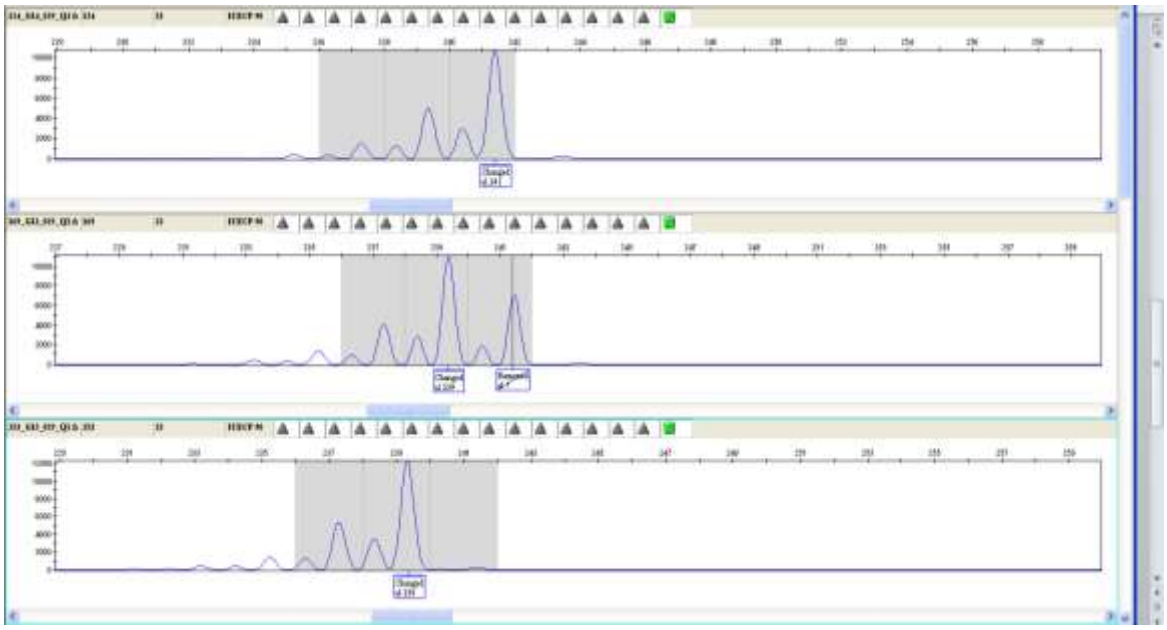


Figure 7: Gene Mapper screen shot of the best type of marker, a mono-allelic, polymorphic marker

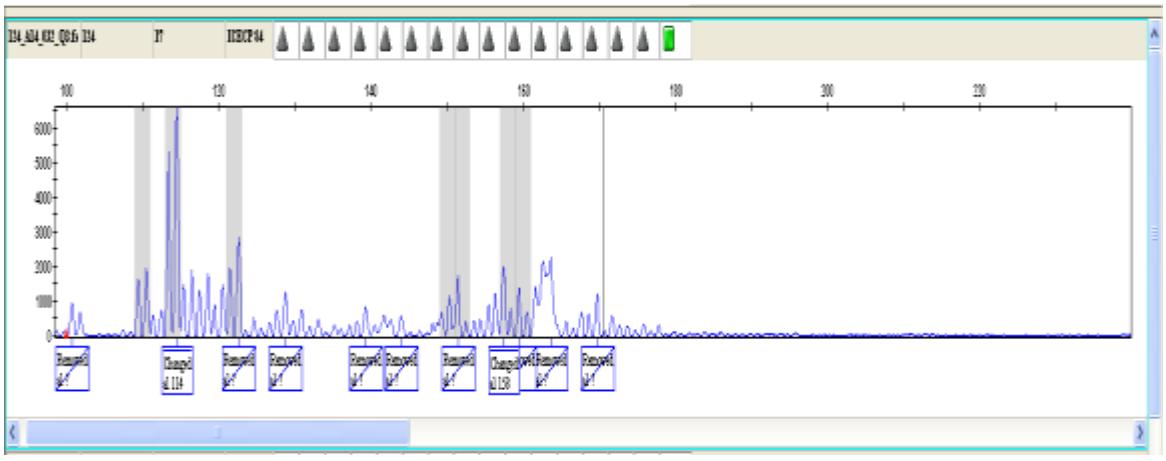


Figure 8: Gene Mapper screen shot of a marker that was difficult to score and that were mostly avoided for further analyses.

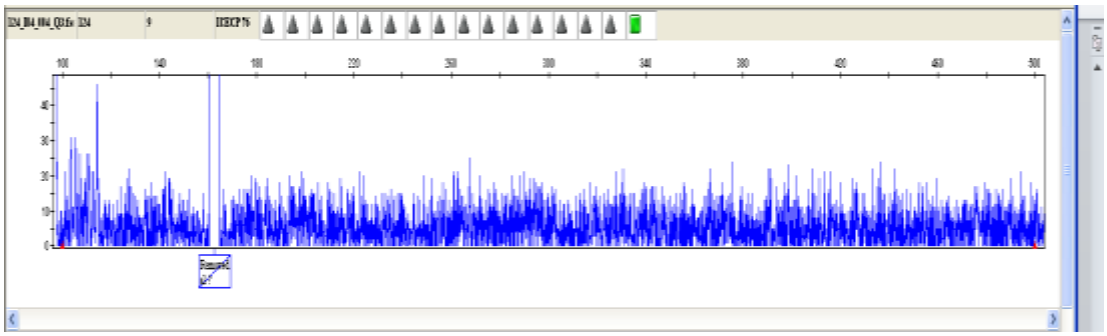


Figure 9: Gene Mapper screen shot showing an example of a marker that did not work and therefore did not display a discernible amplification product.

Table 9: Number of markers that amplified each genotype

Genotype	Markers that worked (from a total of 101)	% of total
5	48	47.5
33	49	48.5
77	53	52.5
124	48	47.5
169	45	44.5
253	49	48.5
271	45	44.5
281	45	44.5
386	51	50.5
408	51	50.5

Statistical analysis using PowerMarker V3.25 (Liu and Muse, 2005) showed that 49 primers were polymorphic, 10 were monomorphic, while 42 did not work.

Table 10: Power marker output showing the 49 markers that were polymorphic

Marker	Major.Allele	SampleSize	No. of obs.	AlleleNo	Availability	GeneDiversity	Heterozygosity	PIC
ICECP 54	0.3000	10.0000	10.0000	6.0000	1.0000	0.8000	0.0000	0.7716
ICECP 47	0.2857	10.0000	7.0000	5.0000	0.7000	0.7755	0.0000	0.7397
ICECP 89	0.3500	10.0000	10.0000	5.0000	1.0000	0.7200	0.7000	0.6722
ICECP 50	0.4500	10.0000	10.0000	4.0000	1.0000	0.6750	0.1000	0.6191
ICECP 58	0.5000	10.0000	10.0000	4.0000	1.0000	0.6600	0.0000	0.6102
ICECP 84	0.4500	10.0000	10.0000	4.0000	1.0000	0.6650	1.0000	0.6035
ICECP 5	0.5000	10.0000	10.0000	4.0000	1.0000	0.6350	0.8000	0.5729
ICECP 96	0.4444	10.0000	9.0000	3.0000	0.9000	0.6420	0.0000	0.5676
ICECP 3	0.5000	10.0000	6.0000	3.0000	0.6000	0.6111	0.0000	0.5355
ICECP 95	0.6000	10.0000	10.0000	4.0000	1.0000	0.5800	0.0000	0.5350
ICECP 4	0.5500	10.0000	10.0000	3.0000	1.0000	0.5950	0.9000	0.5280
ICECP 68	0.6000	10.0000	5.0000	3.0000	0.5000	0.5600	0.0000	0.4992
ICECP 73	0.5000	10.0000	10.0000	3.0000	1.0000	0.5800	1.0000	0.4918
ICECP 53	0.6000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 63	0.6000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 64	0.6000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 90	0.6000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 61	0.7000	10.0000	10.0000	4.0000	1.0000	0.4800	0.0000	0.4500
ICECP 62	0.7000	10.0000	10.0000	4.0000	1.0000	0.4800	0.0000	0.4500
ICECP 37	0.7000	10.0000	10.0000	3.0000	1.0000	0.4600	0.0000	0.4102
ICECP 69	0.7000	10.0000	10.0000	3.0000	1.0000	0.4600	0.0000	0.4102
ICECP 66	0.7500	10.0000	10.0000	4.0000	1.0000	0.4150	0.2000	0.3894
ICECP 11	0.5000	10.0000	2.0000	2.0000	0.2000	0.5000	0.0000	0.3750
ICECP 67	0.5000	10.0000	2.0000	2.0000	0.2000	0.5000	0.0000	0.3750
ICECP 70	0.5000	10.0000	10.0000	2.0000	1.0000	0.5000	0.0000	0.3750
ICECP 71	0.5000	10.0000	10.0000	2.0000	1.0000	0.5000	0.0000	0.3750
ICECP 97	0.5000	10.0000	2.0000	2.0000	0.2000	0.5000	0.0000	0.3750
ICECP 46	0.5714	10.0000	7.0000	2.0000	0.7000	0.4898	0.0000	0.3698
ICECP 40	0.6000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 85	0.6000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 98	0.6000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 99	0.6000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 42	0.6250	10.0000	8.0000	2.0000	0.8000	0.4688	0.0000	0.3589
ICECP 44	0.6667	10.0000	9.0000	2.0000	0.9000	0.4444	0.0000	0.3457
ICECP 59	0.6667	10.0000	9.0000	2.0000	0.9000	0.4444	0.0000	0.3457
ICECP 82	0.6667	10.0000	3.0000	2.0000	0.3000	0.4444	0.0000	0.3457
ICECP 92	0.6667	10.0000	3.0000	2.0000	0.3000	0.4444	0.0000	0.3457
ICECP 93	0.7778	10.0000	9.0000	3.0000	0.9000	0.3704	0.0000	0.3402
ICECP 56	0.7000	10.0000	10.0000	2.0000	1.0000	0.4200	0.0000	0.3318
ICECP 52	0.7143	10.0000	7.0000	2.0000	0.7000	0.4082	0.0000	0.3249
ICECP 72	0.8000	10.0000	10.0000	3.0000	1.0000	0.3400	0.0000	0.3142
ICECP 80	0.8000	10.0000	10.0000	3.0000	1.0000	0.3400	0.0000	0.3142
ICECP 101	0.8000	10.0000	10.0000	3.0000	1.0000	0.3400	0.0000	0.3142
ICECP 1	0.7500	10.0000	8.0000	2.0000	0.8000	0.3750	0.0000	0.3047
ICECP 43	0.8000	10.0000	10.0000	2.0000	1.0000	0.3200	0.0000	0.2688
ICECP 48	0.8889	10.0000	9.0000	2.0000	0.9000	0.1975	0.0000	0.1780
ICECP 57	0.9000	10.0000	10.0000	2.0000	1.0000	0.1800	0.0000	0.1638
ICECP 81	0.9000	10.0000	10.0000	2.0000	1.0000	0.1800	0.0000	0.1638
ICECP 91	0.9000	10.0000	10.0000	2.0000	1.0000	0.1800	0.0000	0.1638
Mean	0.6219	10.0000	8.6735	2.8367	0.8673	0.4849	0.0959	0.4153

The results of the *in vitro* PCR were compared with the results of the *in silico* PCR. 53 out of 101 markers had similar results and this is shown in Table 12.

Table 11: Comparison of the output of the *in vitro* vs *in silico* PCR.

Name	CLC output	*in vitro*
ICECP 1	Y	Y
ICECP 2	Y	N
ICECP 3	Y	Y
ICECP 4	Y	Y
ICECP 5	Y	Y
ICECP 6	Y	N
ICECP 7	Y	N
ICECP 8	Y	N
ICECP 9	Y	Y
ICECP 10	N	N
ICECP 11	Y	Y
ICECP 12	Y	N
ICECP 13	Y	N
ICECP 14	Y	N
ICECP 15	Y	N
ICECP 16	Y	N
ICECP 17	Y	N
ICECP 18	Y	N
ICECP 19	N	N
ICECP 20	N	N
ICECP 21	Y	N
ICECP 22	N	N
ICECP 23	Y	Y
ICECP 24	Y	Y
ICECP 25	Y	N
ICECP 26	Y	N
ICECP 27	Y	N
ICECP 28	N	N
ICECP 29	N	N
ICECP 30	Y	N
ICECP 31	Y	Y
ICECP 32	N	N
ICECP 33	N	N
ICECP 34	Y	N
ICECP 35	Y	Y
ICECP 36	N	Y
ICECP 37	N	Y
ICECP 38	Y	N
ICECP 39	N	N
ICECP 40	N	Y
ICECP 41	N	Y
ICECP 42	N	Y
ICECP 43	N	Y
ICECP 44	N	Y
ICECP 45	Y	Y
ICECP 46	Y	Y
ICECP 47	N	Y
ICECP 48	Y	Y
ICECP 49	Y	Y
ICECP 50	Y	Y
ICECP 51	N	N
ICECP 52	N	Y
ICECP 53	Y	Y
ICECP 54	Y	Y
ICECP 55	Y	N
ICECP 56	Y	Y
ICECP 57	Y	Y
ICECP 58	Y	Y
ICECP 59	Y	Y
ICECP 60	N	N
ICECP 61	N	Y
ICECP 62	N	Y
ICECP 63	Y	Y
ICECP 64	N	Y
ICECP 65	Y	N
ICECP 66	Y	Y
ICECP 67	Y	Y
ICECP 68	N	Y
ICECP 69	Y	Y
ICECP 70	N	Y
ICECP 71	N	Y
ICECP 72	N	Y
ICECP 73	N	Y
ICECP 74	Y	N
ICECP 75	Y	N
ICECP 76	Y	N
ICECP 77	Y	N
ICECP 78	Y	N
ICECP 79	Y	N
ICECP 80	N	Y
ICECP 81	Y	Y
ICECP 82	Y	Y
ICECP 83	Y	Y
ICECP 84	Y	Y
ICECP 85	N	Y
ICECP 86	N	N
ICECP 87	N	N
ICECP 88	Y	N
ICECP 89	Y	Y
ICECP 90	Y	Y
ICECP 91	Y	Y
ICECP 92	Y	Y
ICECP 93	Y	Y
ICECP 94	N	Y
ICECP 95	Y	Y
ICECP 96	Y	Y
ICECP 97	Y	Y
ICECP 98	Y	Y
ICECP 99	Y	Y
ICECP 100	Y	N
ICECP 101	Y	Y

NB: Red = difference

Black = similar

4.8 BLAST search

Following a BLASTX search on the NCBI site, out of all 101 sequences identified in this study that contained SSRs and that amplified *in vitro*, only one was linked to a functional gene. This was ICECP 36, which was linked to the Major Facilitator Superfamily Protein (MFS), with the query yielding a total bit score of 240 and E value of 2e-05. The bit score measures the similarity of the query sequence to the sequence in the database. The E value measures the reliability of the score. It describes the probability due to chance, of there being another alignment with a similarity greater than the given score (http://blast.ncbi.nlm.nih.gov).

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.1 Discussion

Use of NGS in this study has enabled the production of large numbers of sequences at reduced cost and time. Consequently, it has been possible to develop a large number of microsatellites for use in studying finger millet. This has been achieved without having to sub clone the DNA into vectors and amplifying them in hosts, as it would have been necessary if Sanger sequencing was employed.

The technique has resulted in the production of markers that are highly polymorphic and can be easily and reproducibly detected by PCR (Kalia *et al*., 2011). The markers are also hyper variable, and exhibit wide genomic distribution, co-dominant inheritance, reproducibility, multi-allelic nature and chromosome specific location. These are attributes that make them ideal for future breeding and genetic studies in finger millet.

Despite the fact that the use of NGS enables the generation of huge amounts of DNA sequence reads (Brautigam and Gowik, 2010), the data supplied by Ecogenics GmbH was surprisingly, less than expected. Typically, a 454 sequencing run produces circa 1,000,000 reads with a typical length of 400-500 bases (Turner, 2011). Several factors may have contributed to the diminished output. First, Ecogenics GmbH required that leaf material be sent to them as starting material. This meant that the leaf material had to be sampled, dried, packaged then sent to Ecogenics by courier. They performed the DNA extraction, enrichment and sequencing. It is probable that the handling of leaf material prior to DNA extraction may have diminished their quality as good sources for

DNA extraction. Thus the DNA obtained may not have been of optimal quality or sufficient amounts for optimal sequencing results.

Secondly, the ICRISAT laboratory routinely performs DNA extractions for finger millet. As such, there is an optimized protocol that ensures sufficient yields for routine genotyping experiments. However, it was observed that when compared to other plants such as sorghum, finger millet often yielded less DNA. Consequently, a researcher has to be extra diligent to succeed in obtaining sufficient quantities of DNA from finger millet samples. It is possible that Ecogenics was not able to optimize their DNA extraction protocol to suit finger millet. This probably resulted in low yields of DNA following extraction.

Thirdly, prior to sequencing, Ecogenics performed an enrichment step to select for the presence of GT and CT dinucleotides in the DNA fragments submitted to sequencing. The purpose of enrichment was to increase microsatellite detection efficiency and reduce genome complexity (Zalapa *et al*., 2012). Additionally, including an enrichment step enabled the exclusion of unwanted motifs such as AT repeats that could be problematic during amplification (Malausa *et al*., 2011). However, it has been observed previously that incorporating an enrichment step may potentially reduce the capacity to obtain ample numbers of microsatellite loci (Castoe *et al*., 2010). Zalapa *et al*. (2012) pointed out that studies which include an enrichment step prior to sequencing tended to report very low numbers of sequences with SSRs. Therefore, the enrichment step in this study may have contributed in reducing the number of reads obtained after sequencing.

Phred scores typically range from 4 to 60 with higher scores denoting higher sequence quality. Phred scores are logarithmically linked to error probabilities. Thus a score of 10 means the accuracy of base call is 90%. A score of 20, 30 or 40 indicates an accuracy of 99%, 99.9% and 99.99% respectively (Ewing *et al*., 1998). Thus, a score of 20 was considered adequate for this study.For KNE 755, 77.55% of the sequences met the quality criteria while for KNE 796, 83.03% of the sequences were above the minimum sequence quality threshold (Table 3). This indicated that the quality of the DNA was good. This further lends credence to the assumption that the low number of reads may have resulted from low quantity of DNA rather than low quality.

After sequence assembly, very few contigs were formed. This was due to the small number of available raw sequences. Long sequence lengths that result from sequence assembly into contigs are desirable for SSR isolation and primer design. If the number of sequences was large enough, the assembly step would have compensated for the shorter than expected read length of the raw sequences, which could not be achieved efficiently in this study.

SciRoKo was the tool of choice for SSR mining. Detection was done in MISA mode, which allowed for isolation of di- to hexanucleotides. The software not only isolated the microsatellites, but it was also instructed to extract flanking sequences for primer design from sequences with microsatellite loci. After mining of SSRs and extraction of flanking regions, all the sequences that assembled into contigs had flanking regions, as expected. Assembling sequences into contigs was useful because it enabled the joining of single

reads, some of which had the microsatellite loci near the end of the sequence. This made it possible to design primers for sequences that would otherwise have been discarded. This was confirmed by the fact that 37.34% and 35.78% of the KNE 755 and KNE 796 sequences which did not assemble into contigs, had to be discarded due to lack of flanking regions (Table 5).

Primer3 is a primer design software routinely used for generating primers. Unfortunately the standard web based tool does not facilitate the design of primers in batch. To mitigate this challenge, it is possible to download the Primer3_core standalone software. The drawback of this option is that it is complex to set up and run, and a researcher needs to be conversant with working on the command line of a computer (Rozen and Skaletsky, 2000). BatchPrimer3 simplifies the process of designing primers from many sequences. It is a program based on Primer3, but with extended capability for batch detection of primers as well as exporting of results into a Microsoft Excel spread sheet that is easy to work with. With good internet connection, there is no need to download any software. Up to 500 sequences can be processed simultaneously via the graphical user interphase (GUI), thus there is no need for specialized computer skills. Additionally, the sequences are sent to a remote server for primer design, circumventing the need for high end computation infrastructure (You *et al*., 2008).

SSRs with flanking sequences were subjected to primer design using BatchPrimer3. The results showed that 58% of KNE 755 contigs and 35% of KNE 796 contigs contained valid primer sequences. For the single reads, 33% of KNE 755 and 34% of KNE 796

sequences had primers successfully designed for them (Table 6). One of the reasons that could account for the low numbers of successfully synthesized primers was that the flanking regions of the SSR containing sequences could have been too short to enable the design of good primers.

Primers with a GC content of 40-60% ensure stable binding of the primer to the template because the G-C bonds have increased melting temperature than A-T bonds. It is essential to avoid primers that can potentially form dimers as such primer dimers can potentially inhibit amplification of template DNA by competing for PCR reagents (Vallone and Butler, 2004). Thus, it was necessary to further process the primers designed to eliminate dimers and primers with less than 40% GC content. The tool PrimerAnalyser was used (Kalendar *et al*., 2011). The tool further calculated the potential efficiency of the primers and those with greater than 75% efficiency were selected. The results indicated that only 15% of the KNE 755 primers and 18% of the KNE 796 could meet the criteria for potentially good primers (Table 6). This showed that *in silico* tools such as PrimerAnalyser had practical benefit, since they enabled selection of primers with the highest likelihood of working, prior to *in vitro* analysis. This results in huge savings on cost and labor of validating primers.

The primers that were finally selected had an average potential PCR efficiency of 83.2% for KNE 755 and 82.7% for KNE 796. In terms of classes of microsatellites, 71% of KNE 755 primers were dinucleotides, 14% were trinucleotides, 4% were tetranucleotides and 9% were pentanucleotides. For KNE 796 80% were dinucleotides,

8% were trinucleotides, 4% were tetranucleotides, 6% were pentanucleotides and 2% were hexanucleotides (Appendix I & II). The higher percentages of dinucleotides is consistent with the results obtained by other researchers. In their study, Cavagnaro *et al*. (2010) observed that dinucleotides had more repeat units than any other SSR type as well as the highest cumulative sequence length. Zhu *et al*. (2012) reported similar observations and noted that the SSR frequency decreased sharply as repeat number increased, leading to an abundance of dinucleotides over other motifs. The enrichment step could also be a contributing factor to the greater abundance of dinucleotides.

The new primers were unique from the published primers as well as from those supplied by the sequencing service provider. This was unexpected considering that the new primers and the Ecogenics primers were derived from the same sequences. Since Ecogenics did not provide any information regarding the SSR detection method they employed, the lack of similarity between the Ecogenics primers and the new primers was probably as a result of the use of different microsatellite detection platforms and SSR selection criteria as well as use of different primer design tools. The 92 new primers were slightly less than the number supplied by the sequence service provider. However, combining the two sets of primers provided a larger number of primers for finger millet than was initially available (Table 7).

The Ecogenics and the new primers were subjected to basic *in silico* PCR to determine which ones could potentially amplify genomic regions. In the absence of a reference genome, the raw sequences were fed into CLC Genomics workbench v5 (www.clcbio.com) as template for analysis of the primers. From the Ecogenics primers,

64.4% and 71.4% of the KNE 755 and KNE 796 primers amplified *in silico.* From the new primer set, 78.6% and 86.0% of the KNE 755 and the KNE 796 amplified (Table 8). Thus the new set of primers displayed a higher percentage amplification efficiency when compared to the Ecogenics primers. This could be attributed to the fact that the new primers were screened to select the most efficient primers.

Cost effective SSR detection can be achieved by using polyacrylamide gel electrophoresis followed by imaging using radioactive isotopes, silver staining or ethidium bromide (Wang *et al*., 2009). Alternatively, agarose gel electrophoresis can be employed to check that the PCR amplification was successful; unfortunately, it cannot resolve alleles that differ by small increments of the repeat element (Wang *et al*., 2009). Automated capillary systems are available that can resolve differences of up to a single base pair. An example of this system is the ABI 3730 (Applied Biosystems) which was used for this study. In such cases agarose electrophoresis is done prior to fragment analysis so that only markers that amplify suitable PCR products are submitted to the expensive step of fragment analysis on ABI 3730.

After fragment analysis the data was subjected to allele calling using Gene Mapper software. The results showed that of the markers that worked, most worked well for the majority of varieties tested and are therefore robust markers with good potential for future use with finger millet.

Out of the polymorphic markers, 32 amplified in all genotypes, 12 amplified in less than 100% but greater than 50% of the genotypes, while 5 amplified in less than 50% of the

genotypes. The 49 polymorphic markers showed varying degrees of polymorphism ranging from 0.16 to 0.77 with an average PIC of 0.4153. PIC values normally range from 0 to 1. Values close to 0 indicate absence of variation, while values close to 1 reveal maximum polymorphism. Markers with PIC values above 0.5 are generally regarded to be highly polymorphic. In this study, 11 markers had a PIC greater than 0.5 and were thus considered to be highly polymorphic (Table 11 and Appendix VI).

A comparison was made between the *in vitro* PCR results and the *in silico* PCR results. Of the 101 markers tested, 53 had results that were consistent both *in silico* and *in vitro.* Thus, 52.5% of the markers showed the same result in both approaches. This shows that, *in silico* PCR does not necessarily predict the performance of the primers *in vitro,* since there are other PCR parameters such as amplification program as well as salt and primer concentration in the PCR mix which cannot be easily simulated (Bellemain *et al*., 2010). Thus, an *in vitro* step should be performed to give full guarantee on the characteristics of a primer pair (Table 12).

Only one primer pair was linked to a functional protein. This is not surprising since microsatellites are found in transcribed regions in much lower proportions (Kalia *et al*., 2011). This is due to selection against frame shift mutations in coding regions (Liu *et al*., 2002).

The primer ICECP 36 was linked to the major facilitator superfamily (MFS) protein. This is one of the two largest families of membrane transporters known. It consists of 74 families, each of which is usually concerned with the transport of a certain type of

substrate (Reddy *et al*., 2012). They translocate substrates against their electrochemical gradient by coupling the movement of an ion or a second solute down its gradient (Kaback *et al*. 2001). This is the first report of an SSR marker linked to a specific protein in finger millet.

5.2 Conclusion

The *in silico* study developed 92 new primers in addition to the 101 primers supplied by the sequence service provider. The latter were validated *in vitro* in this study and revealed 49 polymorphic and hence useful primers that can be added to the existing 82 SSRs available to date (Dida *et al.*, 2007). The new markers are valuable tools that will be useful for conducting studies that involve characterizing and fingerprinting cultivated varieties of finger millet, assaying genetic diversity and MAS.

This study demonstrates the use of NGS to rapidly and cost-effectively generate genomic sequences containing SSR motifs. It points to the possibility of using cutting edge technology to advance research in orphan crops such as finger millet by researchers in Kenya. It also shows that it is possible to develop abundant genomic resources for hitherto understudied crops which have great significance to the development of third world countries in Africa and Asia.

5.3 Recommendations

This study has resulted in the development of markers that are an important genomic resource for finger millet. I recommend that the 49 polymorphic markers be used to further advance genetic studies in finger millet. The markers can be effectively utilized for studies involving genome mapping, cultivar identification, MAS, genetic diversity studies, phylogenetic relationships and population studies.

Additionally, subject to availability of funds, all the remaining new markers identified in this study, should be evaluated *in vitro* to confirm their ability to reliably amplify polymorphic SSRs in finger millet.

REFERENCES

Abdelkrim, J., Robertson, B.C., Stanton, J.L. and Gemmell, N.J. (2009). Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, 46:185-192.

Allentoft, M.E., Schuster, S.C., Holdaway, R.N., Hale, M.L., McLay, E., Oskam, C., Thomas, M., Gilbert, P., Spencer, P., Willerslev, E. and Bunce, M. (2009). Identification of microsatellites from an extinct moa species using high throughput (454) sequence data. *BioTechniques,* 46 (3):195-200.

Anithakumari, A.M., Tang, J., Eck, H.J., Visser, R.G.F., Leunissen, J.A.M., Vosman, B. and Linden, C.G. (2010). A pipeline for high throughput detection and mapping of SNPs from EST databases. *Molecular Breeding*, 26:65-75.

Anmarkrud, J.A., Kleven, O., Bachmann, L. and Lifjeld, J.T. (2008). Microsatellite evolution: Mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus *HrU10. BioMed Central Evolutionary Biology*, 8:138 [online] DOI: 10.1186/1471-2148-8-138.

Babu, B., Senthil, N., Gomez, S., Biji, K., Rajendraprasad, N., Kumar, S. and Babu, R. (2007). Assessment of genetic diversity among finger millet *Eleusine coracana* L Gaertn accessions using molecular markers. *Genetic Resources and Crop Evolution*, 54:399–404.

Barbeu, W.E. and Hilu, K.W. (1993). Protein, calcium, iron and amino acid content of selected wild and domesticated cultivars of finger millet. *Plant Foods for Human Nutrition*, 43:97-104.

Bardakci, F. (2000). Random Amplified Polymorphic DNA (RAPD) Markers. *Turkish Journal of Biology*, 25:185-196.

Barr, A.R. (2009). Marker assisted selection in theory and practice. In: Ceccarelli, S., Guimarães, E.P. and Weltzien, E. eds. *Plant breeding and farmer participation.* Rome: Food and Agriculture Organization of the United Nations. 479-517.

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P. and Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BioMed Central Microbiology*, 10:189 [online], DOI: 10.1186/1471-2180-10-189.

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

Boutros, P. C. and Okey, A. B. (2004). PUNS: Transcriptomic- and genomic-in silico PCR for enhanced primer design. *Bioinformatics,* 20(15): 2399–2400.

Brautigam, A. and Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology,* 12: 831–841.

Breseghello, F. and Sorrells, M.E. (2006b). Association mapping of kernel size and milling quality in wheat (*Triticum aes-tivum* L.) cultivars. *Genetics* 172:1165–1177.

Breseghello, F. and Sorrells, M.E. (2006a). Association Analysis as a Strategy for Improvement of Quantitative Traits in Plants. *Crop Sci*ence, 46: 1323–1330.

Brumlop, S. and Finckh, M.R. (2011). *Applications and potentials of marker assisted selection (MAS) in plant breeding*. Bonn: Bundesamt für Naturschutz (BfN).

Buerstmayr, H., Ban. T. and Anderson, J.A. (2009). QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: a review. *Plant Breeding*, 128:1-26.

Castoe, T.A., Poole, A.W., Gu, W., Jason De Koning, S.P., Daza, J.M., Smith, E.N. and Pollock, D.D. (2010). Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources,* 10:341-347.

Cavagnaro, P.F., Senalik, D.A., Yang, L., Simon, P.W., Harkins, T.T., Kodira, C.D., Huang, S. and Weng, Y. (2010). Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus L.*). *Biomed Central Genomics*, 11:569–586.

Chethan, S. and Malleshi, N.G. (2007). Finger millet polyphenols: Characterization and their nutraceutical potential. *American Journal of Food Technology*, 2: 618-629.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Muller, W. E., Wetter, T. and Suhai, S. (2004). Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research,* 1(6):1147-1159.

Chistiakov, D.A., Hellemans, B. and Volckaert, F.A.M. (2005). Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture,* 255:1-29.

Conesa, A., Götz, S., García-Gómez., J.M., Terol1, J., Talón, M. and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* 21(18):3674-3676.

Csencsics, D., Brodbeck, S. and Holderegger, R. (2010). Cost –effective, species-specific microsatellite development for the endangered dwarf bulrush (*Typha minima*) using next-generation sequencing technology. *Journal of Heredity*, 101(6):789-793.

Damme, V., Gomez-Paniagua, H. and Carmen de Vicente, M. (2010). The GCP molecular marker toolkit, an instrument for use in breeding food security crops. *Molecular Breeding* [online], DOI 10.1007/s11032-010-9512-3.

Danson. J.W., Mbogori, M., Kimani, M., Lagat, M., Kuria, A. and Diallo, A. (2006). Marker assisted introgression of opaque2 gene into herbicide resistant elite maize inbred lines. *African Journal of Biotechnology* , 5 (24):2417-2422.

Das, S. and Misra, R.C. (2010). Assessment of genetic diversity among finger millet genotypes using RAPD markers. *Indian Journal of Agricultural Resources,* 44 (2):112 – 118.

Dawson, I.K., Hedley, P.E., Guarino, L. and Jaenicke, H. (2009). Does biotechnology have a role in the promotion of underutilised crops?. *Food Policy,* 34:319–328.

D'hoop, B.B., Paulo, M.J., Mank, R.A., van Eck, H.J. and van Eeuwijk, F.A. (2008). Association mapping of quality traits in potato (*Solanum tuberosum* L.). *Euphytica,* 161 (1-2):47-60.

Dida, M., Srinivasachary., Ramakrishnan, S., Bennetzen, J.L., Gale, M.D. and Devos, K.M. (2007). The genetic map of finger millet, *Eleusine coracana*. *Theoretical and Applied Genetics*, 114:321—332.

Dida, M.M., Wanyera, N., Melanie, L., Dunn, H., Jeffrey, L., Bennetzen, J.L. and Devos, K.M. (2008). Population structure and diversity in finger millet (*Eleusine coracan*) germplasm. *Tropical Plant Biology,* 1:131-141.

Eapan, S. and George, L. (1989). High frequency plant regeneration through somatic embryogenesis in finger millet (*Eleusine coracana* Gaertn). *Plant Science*, 61(1):127-130.

Ek, M., Eklund, M., Von Post, R., Dayteg, C., Henriksson, T., Weibull, P., Ceplitis, A., Isaac, P. and Tuvesson, S. (2005). Microsatellite markers for powdery mildew resistance in pea (*Pisum sativum* L.). *Hereditas*, 142:86-91.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 5:435-445.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Resources*, 8 (3):175–185.

Fakrudin, B., Kulkarni, R. S., Shashidhar, H. E. and Hittalmani, S. (2007). Genetic diversity assessment of finger millet, *Eleusine coracana*, germplasm through RAPD analysis. *Plant Genetic Resources Newsletter*, 138, 52–54.

FAO and ICRISAT. (1996). The World Sorghum Economies: Facts, Trends and Outlook. FAO, Rome, Italy and ICRISAT, Andhra Pradesh, India.

FAO. (2010). The state of food insecurity in the world: addressing food insecurity in protracted crises. Rome, Italy.

Guo, W., Cai, C., Wang, C., Han, Z., Song, X., Wang, K., Niu, X., Wang, C., Lu, K., Shi, B. and Zhang, T. (2007). A Microsatellite-Based, Gene-Rich Linkage Map Reveals Genome Structure, Function and Evolution in Gossypium. *Genetics*, 176 (1):527-541.

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, R., Verma, K., Joshi, D.C., Yadav, D. and Singh, M. (2010). Assessment of Genetic Relatedness among Three Varieties of Finger Millet with Variable Seed Coat Color Using RAPD and ISSR Markers. *Genetic Engineering and Biotechnology Journal*, 2:1-9.

Hiebert, C.W., Thomas, J.B., Somers, D.J., McCallum, B.D. and Fox, S.L. (2007). Microsatellite mapping of adult-plant leaf rust resistance gene Lr22a in wheat. *Theoretical and Applied Genetics*, 115:877–884.

Hile, S. E. and Eckert, K. A. (2004). Positive correlation between DNA polymerase α-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *Journal of Molecular Biology*, 335:745–759.

Iftekharuddaula, K.M., Newaz, M.A., Salam, M.A., Ahmed, H.U., Mahbub, M.A.A., Septiningsih, E.M., Collard, B.C.Y., Sanchez, D.L., Pamplona, A.M. and Mackill, D.J. (2011). Rapid and high-precision marker assisted backcrossing to introgress the SUB1 QTL into BR11, the rainfed lowland rice mega variety of Bangladesh. *Euphytica*, 178 (1):83-97.

IPGRI, GFU, MSSRF. (2005). Meeting the millennium development goals with agricultural biodiversity. International Plant Genetic Resources Institute (Bioversity International), Rome, Italy, the Global Facilitation Unit for Underutilized Species, Rome, Italy and the MS Swaminathan Research Foundation, Chennai, India.

Jaenicke, H., Höschle-Zeledon, I. eds. (2006). Strategic Framework for Underutilised Plant Species Research and Development with Special Reference to Asia and the Pacific, and to Sub-Saharan Africa. International Centre for Underutilised Crops, Colombo, Sri Lanka and the Global Facilitation Unit for Underutilized Species, Rome, Italy. (http://www.icuc-iwmi.org/Publications/Strategy%20Documents/summary%20for%20strategic%20framework.htm).

Kaback, H.R., Sahin-Toth, M. and Weinglass, A.B. (2001). The kamikaze ap-proach to membrane transport. *Nature Reviews. Molecular and Cell Biology* ,2:610–620.

Kalendar, R., Lee, D. and Schulman, A.H. (2011). Java web tools for PCR, *in silico* PCR, and oligonucleotide assembly and analysis. *Genomics*, 98(2):137-144.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011). Microsatellite markers: an overview of the recent progress in plants. *Euphytica*, 177:309–334.

Khattak, J.Z.K., Torp, A.M. and Andersen, S.B. (2006). A genetic linkage map of *Spinacia oleracea* and localization of a sex determination locus. *Euphytica*, 148:311–318.

Kofler, R., SchlÖtterer, C. and Lelley, T. (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Genome analysis*, 23(13):1683–1685.

Kothari, S.L., Kumar, S., Vishnoi, R.K., Kothari, O. and Watanabe, K.N. (2005). Application of biotechnology for improvement of millet crops: Review of progress and future prospects. *Plant Biotechnology*, 22(2):81-88.

Krishnappa, M., Ramesh, S., Chandraprakash, J., Gowda, J., Bharathi and Doss, D.D. (2009). Genetic analysis of economic traits in finger millet. *Journal of SAT Agricultural Research*, 7:1-5.

Kumar, S., Agarwal, K. and Kothari, S. L. (2001). *In vitro* induction and enlargement of apical domes and formation of multiple shoots in finger millet, *Eleusine coracana* (L.) Gaertn and crowfoot grass, *Eleusine indica* (L.) Gaertn. *Current Science*, 81(11):1482-1485.

Kumar, P., Gupta, V.K., Misra, A.K., Modi, D.R. and Pandey, B.K. (2009). Potential of Molecular Markers in Plant Biotechnology. *Plant Omics Journal*,2(4):141-162.

Kumari, K. and Pande, A. (2010). Study of genetic diversity in finger millet (*Eleusine coracana* L. Gaertn) using RAPD markers. *African Journal of Biotechnology*, 9(29):4542-4549.

Latha, A.M., Rao, K.V. and Reddy, V.D. (2005). Production of transgenic plants resistant to leaf blast disease in finger millet (Eleusine coracana (L.) zaertn.). *Plant Science*, 169:657-667.

Levinson, G. and Gutman, G. A. (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. *Nucleic Acids Research*, 15:5323–5338.

Li, Y., Korol, A.B., Fahima, T., Beiles, A. and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, 11:2453-2465.

Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution*, 21:991–1007.

Liu, K. and Muse, S.V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, 21(9):2128-2129.

Mace, E.S., Buhariwalla, H.K. and Crouch, J.H. (2003). A High-Throughput DNA Extraction Protocol for Tropical Molecular Breeding Programs. *Plant Molecular Biology Reporter*, 21:459a–459h.

Maheswaran, M. (2004). Molecular markers: history, features and applications. *Advanced Biotechnology*, 17-24.

Malausa, T., Gilles, A., Meglécz, E., Blanquart, H., Duthoy, S., Costedoat, C., Dubut, V., Pech, N., Castagnone-sereno, P., Délye, C., Feau, N., Frey, p., Gauthier, P., Guillemaud, T., Hazard, L., Le Corre, V., Lung-Escarmant, B., Malé, P.G., Ferreira, S. and Martin, J. (2011). High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources*, 11:638-644.

Marcel, T.C., Varshney, R.K., Barbieri, M., Jafary, H., de Kock, M.J.D., Graner, A. and Niks, R.E. (2007). A high-density consensus map of barley to compare the distribution of QTLs for partial resistance to Puccinia hordei and of defence gene homologues. *Theoretical and Applied Genetics*, 114 (3):487-500.

Mardis, E.R. (2008). Next generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9:387-402.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11:31-46.

National Research Council. (1996). Finger Millet. In: National Academy of Sciences. *Lost Crops of Africa.* Washington: National Academies Press. [online], 1,39-58. Available from http://www.nap.edu/catalog/2305.html [accessed 10th June 2011].

Parani, M., Rajesh, K., Lakshmi, M., Parducci, L., Szmidt, A. E. and Parida, A. (2001). Species identification in seven small millet species using polymerase chain reaction - restriction fragment length polymorphism of *trnS-psbC* gene region. *Genome*, 44:495–499.

Pareek, C.S., Smoczynski, R. and Tretyn, A.J. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52:413 – 435.

Panwar, P., Nath, M., Kumar, V., Yadav. and Kumar, A. (2010) Comparative evaluation of genetic diversity using RAPD, SSR and cytochrome P450 gene based markers with respect to calcium content in finger millet (*Eleusine coracana* L. Gaertn.). *Journal of Genetics*, 89.

Parida, S.K., Kalia., S.K., Kaul, S., Dalal, V., Hemaprabha, G., Selvi, A., Pandit, A., Singh, A., Gaikwad, K., Sharma, T.R., Srivastava, P.S., Singh, N.K. and Mohapatra, T. (2009). Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. *Theoretical Applied Genetics*, 118:327–338.

Patel, R.K. and Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE* 7(2) [online], DOI:10.1371/journal.pone.0030619.

Poddar, K., Vishnoi, R.K. and Kothari, S.L. (1997). Plant regeneration from embryogenic callus of finger millet *Eleusine coracana* (L.) Gaertn. on higher concentrations of NH_4NO_3 as a replacement of NAA in the medium. *Plant Science*, 129 (1):101-106.

Powell, W., Machray, G.C. and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science*, 1(7):215-222.

Ramsay, L., Macaulay, M., degli Ivanissevich, S., MacLean, K., Cardle, L., Fuller, J., Edwards, K.J., Tuvesson, S., Morgante, M., Massari, A., Maestri, E., Marmiroli, N., Sjakste, T., Ganal, M., Powell, W. and Waugh, R. (2000). A simple sequence repeat-based linkage map of barley. *Genetics*, 156:1997–2005.

Reddy, I.N.B., Reddy, D.S., Narasu, M.L. and Sivaramakrishnan, S. (2011). Characterization of disease resistance gene homologues isolated from finger millet (*Eleusine coracana* L. Gaertn). *Molecular Breeding*, 27:315-328.

Reddy, V. S., Shlykov, M. A., Castillo, R., Sun, E. I. and Saier, M. H. (2012). The major facilitator superfamily (MFS) revisited. *FEBS Journal*, 279:2022–2035.

Richard, G.F. and Paques, F. (2000). Mini- and microsatellite expansions: the recombination connection. *EMBO Reports*, 1:122-126.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, 132:365-86.

Rudd, S., Schoof, H. and Klaus, M. (2005). Plant Markers-A database of predicted molecular markers from plants. *Nucleic Acids Research*, 33:D628–D632.

Schuelke, M. (2000). An economic method for the fluorescent labelling of PCR fragments. A poor man's approach to genotyping for research and high-throughput diagnostics. *Nature Biotechnology*, 18:233-234.

Shekhar, C.P., Smoczynski, R. and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4):413–435.

Sinha, A. and Pande, A. (2010). Finger Printing of *Eleusine coracana* l. (gaertn) using Microsatellite Markers. *Bioresearch Bulletin*, 2:51-58.

Spooner, D.M., Nunez, J., Trujillo, G., Herrera., Mdel, R., Guzman, F. and Ghislain, M. (2007). Extensive simple sequence repeat genotyping of potato landraces supports a major re-evaluation of their gene pool structure and classification. *Proceedings of the National Academy of Sciences USA*, 104, 19398-19403.

Srinivasachary., Dida, M.M., Gale, M.D. and Devos, K.M. (2007). Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theoretical and Applied Genetics*, 115(4):489-499.

Strand, M., Prolla, T. A., Liskay, R. M. and Petes, T. D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, 365:274–276.

Studer, B., Kölliker, R., Muylle, H., Asp, T., Frei, U., Roldán-Ruiz, I., Barre, P., Tomaszewski, C., Meally, H., Barth, S., Skøt, L., Armstead, I.P., Dolstra, O. and Lübberstedt, T. (2010). EST-derived SSR markers used as anchor loci for the construction of a consensus linkage map in ryegrass (*Lolium* spp.). *BioMed Central Plant Biology*, 10:177 [online], DOI:10.1186/1471-2229-10-177

Styslinger, M. (2011). Finger millet: A once and future staple. *Nourishing the Planet*. http://blogs.worldwatch.org/nourishingtheplanet/finger-millet-a-once-and-future-staple [accessed: 12th Feb, 2012]

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, 17(16):6463-6471.

Turner, D.J. (2011). Next-generation DNA sequencing technologies. *Encyclopedia of Analytical Chemistry* [online], DOI: 10.1002/9780470027318.9209.

Vallone, P.M. and Butler, J.M. (2004). AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques*, 37:226-231.

Varshney, R.K., Sigmund, R., Borner, A., Korzun, V., Stein, N., Sorrells, M.E., Langridge, P. and Graner, A. (2005). Interspecific transferability and comparative mapping of barley EST–SSR markers in wheat rye and rice. *Plant Science*, 168:195–202.

Varshney, R.K., Thiel, T., Stein, N., Langridge, P. and Graner, A. (2002). *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cellular and Molecular Biology Letters*, 7:537 – 546.

Varshney, R.K., Mahendar,T., Aggarwal, R.K. Börner, A. (2007). Genic molecular markers in plants: Development and applications. In Varshney, R.K. and Tuberosa, R. eds. *Genomics-Assisted Crop Improvement:Volume 1: Genomics Approaches and Platforms*, The Netherlands, Springer:13-29.

Varshney, R.K., Marcel, T.C., Ramsay, L., Russell, J., Röder, M.S., Stein, N., Waugh, R., Langridge, P., Niks, R.E. and Graner, A. (2007). A high density barley microsatellite consensus map with 775 SSR loci. *Theoretical and Applied Genetics*, 114 (6):1091-1103.

Victoria, F.C., da Maia, L.C. and de Oliveira, A.C. (2011). *In silico* comparative analysis of SSR markers in plants. *Biomed Central Plant Biology*,11:15 [online] DOI:10.1186/1471-2229-11-15.

Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., Barakat, A., Wolcott, E., Liang, H., Landherr, L., Tomsho, L.P., Hu, Y., Carlson, J.E., Ma, H., Schuster, S.C., Soltis, D.E., Soltis, P.S., Altman, N. and DePamphilis, C.W. (2009). Comparison of next generation sequencing technologies for transcriptome characterization. *Biomed Central Genomics*, 10:347 [online], DOI:10.1186/1471-2164-10-347.

Wang, M.L., Barkley, N.A. and Jenkins, T.M. (2009). Microsatellite Markers in Plants and Insects. Part I: Applications of Biotechnology. *Genes Genomes and Genomics*, 3(1).

You, F.M., Huo, N., Gu, Y.Q., Luo, M., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J. and Anderson, O.D. (2008). BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *Biomed Central Bioinformatics*, 9:253 [online] DOI:10.1186/1471-2105-9-253.

Yu, J.K., Dake, T.M., Singh, S., Benscher, D., Li, W., Gill, B. and Sorrells, M.E. (2004). Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome*, 47:805–818.

Yu, J., Yu, S., Lu, C., Wang, W., Fan, S., Song, M., Lin, Z., Zhang, X. and Zhang, J. (2007), High-density Linkage Map of Cultivated Allotetraploid Cotton Based on SSR, TRAP, SRAP and AFLP Markers. *Journal of Integrative Plant Biology*, 49:716–724.

Yu, Y., Yuan, D., Liang, S., Li, X., Wang, X., Lin, Z. and Zhang, X. (2011). Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC

1 population between *gossypium hirsutum* and *G. barbadense.* *BioMed Central Genomics*, 12:15 [online], DOI:10.1186/1471-2164-12-15.

Zalapa, J.E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., McCown, B., Harbut, R. and Simon, P. (2012). Using next -generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany*, 99(1):1-16.

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

Zhang, Y., Sledge, M.K. and Bouton, J.H. (2007). Genome mapping of white clover (*Trifolium repens* L.) and comparative analysis within the *Trifolieae* using cross-species SSR markers. *Theoretical and Applied Genetics*, 114 (8):1367-1378.

Zhu, H., Senalik, D., McCown, B.H., Zeldin, E.L., Speers, J., Hyman, J., Bassil, N., Hummer, K., Simon, P.W. and Zalapa, J.E (2012). Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (*Vaccinium macrocarpon Ait*.). *Theoretical and Applied Genetics*, 124:87-96.

APPENDICES

Appendix I: New KNE 755 primers identified in this study

Seq ID	Orientation	tm	GC%	Seq	Motif	Motif Len	SSR	SSRLen	Primer Efficiency
KNE755_000175	FORWARD	54.88	42.86	CTTCTAGCGTGAGTTTGTGTT	GT	2	GTGTGTGTGTGT	12	86
KNE755_000175	REVERSE	54.87	47.62	CACGACTGTTAGAACCTCATC					84
KNE755_000181	FORWARD	54.6	47.62	CCTAGGTTGGGTAAGGTAAAC	GGGTT	5	GGGTTGGGTTGGGTT	15	79
KNE755_000181	REVERSE	56.26	50	TTCCGTTAGGAAGGAACG					84
KNE755_000460	FORWARD	55.79	61.11	GGTTGGCCAGAGAGAGAG	GTAC	4	GTACGTACGTACGTAC	16	79
KNE755_000460	REVERSE	56.34	55.56	TCTCGTCTCGTCGTCGTA					86
KNE755_000634	FORWARD	55.42	61.11	CTCCTCTGGTCACCTTCC	CA	2	CACACACACACACACACACA	28	90
KNE755_000634	REVERSE	55.53	50	GAGGGAGAAAATCCGTTG					76
KNE755_000796	FORWARD	54.6	42.86	ATCATACCAGCACTAAAGCAC	GA	2	GAGAGAGAGAGAGAGAGAC	20	85
KNE755_000796	REVERSE	55.68	52.63	GCTCGGGCTCGTATAACTA					77
KNE755_000820	FORWARD	55.75	42.86	CCCACTTGGTCATTCTTCTAT	TC	2	TCTCTCTCTCTC	12	75
KNE755_000820	REVERSE	53.93	42.86	GAGATGAGAAAGTTGATGGAG					77
KNE755_000830	FORWARD	54.27	47.37	TGATGTGATTGAGGAGGAG	CAG	3	CAGCAGCAGCAGCAG	15	81
KNE755_000830	REVERSE	55.27	47.62	CCCAGATCCTTAGTGTAATCC					79
KNE755_000891	FORWARD	54.35	47.62	CCCATACTCACACACACTACA	AG	2	AGAGAGAGAGAGAGAGAGA	24	79
KNE755_000891	REVERSE	54.78	47.62	TATAGCTTCCTCTTCCCTAGC					78
KNE755_000985	FORWARD	54.07	42.86	GATGAGATAGGTCTTCCCTTT	AC	2	ACACACACACACACACACAC	22	79
KNE755_000985	REVERSE	54.98	52.63	AGGTCTCCACATTCTCCAG					85
KNE755_000999	FORWARD	54.94	55.56	GTCACCAAGCTTCACCTG	CAA	3	CAACAACAACAACAACAACA	33	90
KNE755_000999	REVERSE	55.3	42.11	CGTTGTTGGTGTTGTTGTT					80
KNE755_001006	FORWARD	54.06	42.86	TATATCCTCTGTTGCACCTCT	GT	2	GTGTGTGTGTGT	12	81
KNE755_001006	REVERSE	54.97	47.62	AGGGAGATATGTGAGAGGAAG					78
KNE755_001119	FORWARD	54.81	42.86	GTGGTGACATTGTGGTTAGAT	CA	2	CACACACACACACACA	16	82
KNE755_001119	REVERSE	55	42.86	CCTGATTTATCCTGTGTGTGT					86
KNE755_001195	FORWARD	56.8	61.11	GCTCCGACATCCTCTCTG	TC	2	TCTCTCTCTCTCTC	14	83
KNE755_001195	REVERSE	55.75	47.37	ATGAGATAGCGGGAAGTCA					91
KNE755_001562	FORWARD	54.97	47.62	CCAAGTCTCTAGTTGTTGGTG	GT	2	GTGTGTGTGTGTGTGTGTGTG	30	84
KNE755_001562	REVERSE	55.75	52.63	CTCATCTCTATGGGCGTTC					88
KNE755_001683	FORWARD	55.35	42.86	TCAGTCTTGCTGTTCCTCTTA	CT	2	CTCTCTCTCTCTCTCTCTCT	22	79
KNE755_001683	REVERSE	54.93	55.56	TATGGCGCAGGAGTAGAG					85
KNE755_001707	FORWARD	57.04	61.11	TCCTCCTCTGACCAGGTG	TC	2	TCTCTCTCTCTCTCTCTC	18	88
KNE755_001707	REVERSE	54.56	45.45	TCACATACACACACACAGAGAG					75
KNE755_001884	FORWARD	55.88	52.63	GATGGAAGCAGAGATGAGC	GA	2	GAGAGAGAGAGA	12	79
KNE755_001884	REVERSE	55.11	55.56	CACCTGAGCGGATACAAG					87
KNE755_001973	FORWARD	54.99	42.86	TTTAAACTACCCTACCCCAAC	GGGTT	5	GGGTTGGGTTGGGTT	15	86
KNE755_001973	REVERSE	56.26	50	TTCCGTTAGGAAGGAACG					84
KNE755_002147	FORWARD	62.04	55	ATTGAGAGTGGCACGGGAGT	AG	2	AGAGAGAGAGAGAG	14	90
KNE755_002147	REVERSE	55.97	47.62	CCTCACTCTCTTGGTCAGAAT					84
KNE755_002155	FORWARD	54.88	47.62	ACTGTACATCTCCACAACCAC	GGGTT	5	GGGTTGGGTTGGGTT	15	88
KNE755_002155	REVERSE	56.26	50	TTCCGTTAGGAAGGAACG					84
KNE755_002386	FORWARD	54.3	40.91	TTATACTCTAAGGGGTTGGTTC	AC	2	ACACACACACACACACAC	18	88
KNE755_002386	REVERSE	54.08	47.37	ATGACTGGAGCATGTCATC					89
KNE755_002536	FORWARD	55.68	42.86	AGTGATACCGCTTCGTTCTAT	TC	2	TCTCTCTCTCTCTCTC	16	81
KNE755_002536	REVERSE	54.48	61.11	CAGAGACTAGCGGAGACG					91
KNE755_002600	FORWARD	57.56	47.37	TAAGCTCCGGCGAGATTAT	AGC	3	AGCAGCAGCAGCAGC	15	89
KNE755_002600	REVERSE	55.71	42.86	GTCCATCTTTTCCATCTCAGT					76
KNE755_002602	FORWARD	55.07	52.63	CTCTGAAGGGTGTGAGGTT	AG	2	AGAGAGAGAGAGAGAGAGA	24	93
KNE755_002602	REVERSE	59.63	66.67	CTCCCGTCCTCTGCTCTG					80
KNE755_003575	FORWARD	55.49	45	CAGACCTTTACATTGCGTCT	AC	2	ACACACACACAC	12	85
KNE755_003575	REVERSE	54.68	42.86	ATTCAGGTGTGTGTGTGTGTA					78
KNE755_003577	FORWARD	55.22	42.86	AAGCAAAATGGTACTCTCTCC	CT	2	CTCTCTCTCTCT	12	83
KNE755_003577	REVERSE	55.17	42.86	TGCTTGTTACGGTGTTACTCT					87
KNE755_004173	FORWARD	54.82	47.37	AAGGTAGGTTCGTTCCGTA	AACC	4	AACCAACCAACC	12	90
KNE755_004173	REVERSE	55.5	52.38	GGTACCCGTAACCGTAGTTAG					81
KNE755_004275	FORWARD	55.36	42.86	CTACAAATTGACCCACGAGTA	GA	2	GAGAGAGAGAGAGAGA	16	85
KNE755_004275	REVERSE	54.52	52.38	GTAACGAGTGCTGAGAGTAGC					89
KNE755_004449	FORWARD	54.66	55	CTCTCACGTCTCGTCTCTCT	TC	2	TCTCTCTCTCTC	12	83
KNE755_004449	REVERSE	55.13	40.91	CCTACATGAGAATAACCGTCTT					83
KNE755_004703	FORWARD	55.26	42.86	TTCTCTGGACATGAAGCTCTA	CT	2	CTCTCTCTCTCT	12	83
KNE755_004703	REVERSE	54.28	55	CTCATGCTAGAGAGGGAGAG					81
KNE755_004777	FORWARD	52.78	50	TAGAAAGACGGGAAGGAG	AG	2	AGAGAGAGAGAGAGAGAGA	52	81
KNE755_004777	REVERSE	52.9	55.56	ACCTACCGTCGTCGTACT					84
KNE755_004777	FORWARD	56.01	52.63	CTCGTTGGGCATAAGAGAG	AGT	3	AGTAGTAGTAGT	12	86
KNE755_004777	REVERSE	58.72	66.67	GACCGACCACCACCTACC					89
KNE755_004794	FORWARD	55.05	42.86	AACTCTGAAGTTAAGCGTCCT	GA	2	GAGAGAGAGAGAGAGAGAC	22	87
KNE755_004794	REVERSE	55.68	52.63	GCTCGGGCTCGTATAACTA					77
KNE755_004822	FORWARD	55.45	52.38	GAGATACGTGAGAGGAAGGAC	AC	2	ACACACACACACAC	14	79
KNE755_004822	REVERSE	55.91	47.62	GATATCCTCTGTTGCACCTCT					85
KNE755_004904	FORWARD	54.97	42.86	ACGGAGAGAGTATCTTTTGCT	AG	2	AGAGAGAGAGAGAGAGAGA	20	80
KNE755_004904	REVERSE	55.1	42.86	TTCTCTCCACGAGTAAGATCA					82
KNE755_005272	FORWARD	55.93	55.56	CTCGCATGGTATGTACGG	TTC	3	TTCTTCTTCTTCTTC	15	81
KNE755_005272	REVERSE	53.93	42.86	GGATGATATAGAGCTTGATGG					84
KNE755_005355	FORWARD	54.18	55.56	TCTCTCGTCTCGTCTCGT	GTGTG	5	GTGTGGTGTGGTGTGGTGTG	20	81
KNE755_005355	REVERSE	54.63	55	CACACTACACCACAGACCAC					88
KNE755_005404	FORWARD	55.18	50	CAGCACATCCTTCCATGT	GT	2	GTGTGTGTGTGTGTGTGTGT	20	85
KNE755_005404	REVERSE	55.18	47.37	ATTTGCCTCACACACACAC					82
KNE755_005640	FORWARD	58.67	61.11	GCACCCTTGGCACCTAGT	CA	2	CACACACACACACA	14	86
KNE755_005640	REVERSE	55.07	42.11	TCTGATGAGCGAAATGTGT					88

Primers from contigs									
Seq ID	**Orientation**	**tm**	**GC%**	**Seq**	**Motif**	**Motif Len**	**SSR**	**SSRLen**	**Primer Efficiency**
755_rep_c14	FORWARD	55.68	52.63	GCTCGGGCTCGTATAACTA	TC	2	tctctctctctctctctctc	22	77
755_rep_c14	REVERSE	54.6	42.86	ATCATACCAGCACTAAAGCAC					85
755_c31	FORWARD	55.01	52.63	GAGCGGTTGGATAAGAGAG	CAA	3	caacaacaacaacaa	15	79
755_c31	REVERSE	55.02	42.86	CTTATACACCGTTGCTTCATC					81
755_rep_c42	FORWARD	54.86	47.62	GAGAGAGGGAGTATCTTTTGC	GA	2	gagagagagagaga	14	76
755_rep_c42	REVERSE	54.98	42.86	CTGGAGTATGATCGAAACAAG					78

Appendix II: New KNE 796 primers identified in this study

Seq ID	Orientation	tm	GC%	Seq	Motif	Motif Len	SSR	SSRLen	Primer Efficiency
KNE796_000303	FORWARD	54.93	42.86	TGTGTGTCACCATCTCTAACA	TG	2	TGTGTGTGTGTGTGTG	16	82
KNE796_000303	REVERSE	55.03	42.86	TGTACACACAACACACACACA					75
KNE796_000374	FORWARD	56.33	42.86	TACAAAACGCGAGAGAGTACA	GA	2	GAGAGAGAGAGAGAGAGAGAGA	24	78
KNE796_000374	REVERSE	56.25	60	GTACTCCCTCCCTACCTCCT					82
KNE796_000397	FORWARD	54.34	47.62	CATCTGTCCACAACTCTCTCT	TG	2	TGTGTGTGTGTGTG	14	81
KNE796_000397	REVERSE	56	47.62	GTGGAGGAGGATGAACATTAC					76
KNE796_000519	FORWARD	60.51	55.56	GCAAGCGTGGCTTCATTC	AG	2	AGAGAGAGAGAGAGAGAGAG	22	88
KNE796_000519	REVERSE	52.16	42.86	GTTCTATCAGAACACATGGAG					84
KNE796_000731	FORWARD	55.97	47.37	CTCTTAGCGCACGAAAACT	GA	2	GAGAGAGAGAGAGA	14	93
KNE796_000731	REVERSE	55.47	42.86	TCTGAGAAGAAGAAGCAGACA					77
KNE796_000733	FORWARD	56.24	52.63	CCCTGACCTCAACATCATC	AC	2	ACACACACACACACACACACACACA	38	82
KNE796_000733	REVERSE	50.41	42.86	TCTATTAGTCGTCTGAGTGTG					88
KNE796_000738	FORWARD	55.31	42.86	CATGACTTACTCCCTCAACAA	TC	2	TCTCTCTCTCTCTCTC	16	85
KNE796_000738	REVERSE	55.1	42.86	TCATCGGAGTACAAGAAGAGA					77
KNE796_000772	FORWARD	57.14	66.67	CTCCTCCTCCTCCTCCAC	CT	2	CTCTCTCTCTCT	12	76
KNE796_000772	REVERSE	53.65	52.63	TAGAGAAGAGGTGGTGACG					82
KNE796_000945	FORWARD	54.96	42.86	GTAGTTGTTGTTCGTTCGTTC	GTCG	4	GTCGGTCGGTCGGTCG	16	79
KNE796_000945	REVERSE	61.35	66.67	GTCGTCGTCGTCGGTACG					86
KNE796_000948	FORWARD	59.32	55.56	GATGGTGGTTGCCAGGTT	CA	2	CACACACACACACACACACA	22	94
KNE796_000948	REVERSE	55.74	42.86	AATGAGACGAGGGTCATTAAC					82
KNE796_001160	FORWARD	60.05	61.11	ACACGCACACCCACACAC	CA	2	CACACACACACACACACACACA	24	82
KNE796_001160	REVERSE	54.24	40.91	AGGCTCATATGTAATCCTAACC					79
KNE796_001225	FORWARD	56.44	47.37	TCGTTTCTTACGGATCGAG	GT	2	GTGTGTGTGTGTGTGTGTGTGTGTG	64	87
KNE796_001225	REVERSE	56.28	61.11	CTCGTGTCGTCGTCGTAG					85
KNE796_001237	FORWARD	55.67	42.11	CTTTGTTTATGACGCACCA	CA	2	CACACACACACACA	16	87
KNE796_001237	REVERSE	56	52.38	CTCTGTGTGTGTGTGAGTGTG					84
KNE796_001380	FORWARD	55.24	42.86	TTTTAGTTTCCGACCTAGTCC	CGGGT	5	CGGGTCGGGTCGGGT	15	80
KNE796_001380	REVERSE	55.09	50	TCCCGTTTAGGAAGGAAC					80
KNE796_001400	FORWARD	54.8	42.86	CAGTTGACGACTAGCAAATCT	AC	2	ACACACACACACACACACACAC	24	77
KNE796_001400	REVERSE	53.73	45	TATCATACGGTCTCTTGTCG					87
KNE796_001532	FORWARD	54.75	52.38	AGTACTAGGATGGGTGACCTC	GA	2	GAGAGAGAGAGAGA	14	90
KNE796_001532	REVERSE	55.32	50	CAAGCTCAGGCTCGTATAAC					84
KNE796_001724	FORWARD	54.97	55	TCTACCGCTGAGAGAGAGAG	AG	2	AGAGAGAGAGAGAG	14	78
KNE796_001724	REVERSE	54.45	42.86	CCCACATGATGACCTATTTAC					77
KNE796_001740	FORWARD	56	52.63	CAGATCCACCCACCATATC	AG	2	AGAGAGAGAGAGAGAGAGAGAG	26	86
KNE796_001740	REVERSE	54.49	52.38	CTCTCTCCTACCCTAGCTGAT					84
KNE796_001754	FORWARD	55.26	52.63	CTATGGCTAACACGGACCT	CT	2	CTCTCTCTCTCT	12	90
KNE796_001754	REVERSE	55.18	42.86	CAGGATCTTTCCCCTTTACTA					76
KNE796_002067	FORWARD	58.56	55.56	TCTCCTTCCCCTTTGTGG	CCT	3	CCTCCTCCTCCTCCTCCT	18	81
KNE796_002067	REVERSE	54.88	52.63	ATCAGAGGAGGATGTGGAG					85
KNE796_002130	FORWARD	55.97	47.62	AAACAGACTCAGGGATCAGAG	GTGAT	5	GTGATGTGATGTGAT	15	86
KNE796_002130	REVERSE	55.7	52.63	GTATGTCCACCACGAACCT					94
KNE796_002236	FORWARD	55.23	47.62	GACAGCAGTTGTACCATCACT	AC	2	ACACACACACACACACACACACA	26	82
KNE796_002236	REVERSE	55.25	42.86	GTTGACGCAAACTGCTTATAC					83
KNE796_002861	FORWARD	56.23	55.56	CAGCAAAGGTGCTGACAC	CCT	3	CCTCCTCCTCCTCCT	15	93
KNE796_002861	REVERSE	55.15	55.56	AAAGGAGGAGGAGGAGGT					75
KNE796_002884	FORWARD	54.95	47.37	GCATACATCAACGTGAAGG	CT	2	CTCTCTCTCTCTCTCTCTCTCTCT	28	86
KNE796_002884	REVERSE	57	50	CACCACACACACACACACAT					75
KNE796_002932	FORWARD	54.41	42.86	TAGCTTTCTGTAGCTCTTTGG	TG	2	TGTGTGTGTGTGTGTGTG	18	77
KNE796_002932	REVERSE	55.59	52.38	GACACAGACACACACACACAC					80
KNE796_003172	FORWARD	55.99	45	AAGAGCGATGAGATGAGTGA	GGTGT	5	GGTGTGGTGTGGTGTGGTGT	20	78
KNE796_003172	REVERSE	53.45	47.62	ATCTCTTGTCCTCACCTACTG					78
KNE796_003195	FORWARD	55.12	42.86	ACACACACACACACAACAAAC	CA	2	CACACACACACACA	14	75
KNE796_003195	REVERSE	54.36	42.86	CTCGTAGCCACAAACTAATTC					77
KNE796_003350	FORWARD	56.32	47.62	GTGCCCTAGGTTTTGTGTAAC	AC	2	ACACACACACACACAC	16	82
KNE796_003350	REVERSE	55.59	47.62	TACTAGCACCTTTGGAGTGTG					87
KNE796_003421	FORWARD	54.33	52.38	CCTGCTCTCTCAGCTCTAGTA	CT	2	CTCTCTCTCTCTCTCTCTCT	20	81
KNE796_003421	REVERSE	55.79	47.62	GTTCTTGTGTGTGTGTGTGTG					77
KNE796_003429	FORWARD	55.13	42.86	GAAAGTACAATGGTGGTGAGA	GT	2	GTGTGTGTGTGTGTGTGT	18	79
KNE796_003429	REVERSE	55.29	52.38	CGACTCTCTCTCTGTAACGTG					82
KNE796_003479	FORWARD	57.16	61.11	CTGTACGGACCCCATGTC	AC	2	ACACACACACACACACACAC	20	94
KNE796_003479	REVERSE	55.01	47.62	GAGAGGACAGGGTAGCATATT					80
KNE796_003500	FORWARD	54.87	47.62	CATCGTTCTCCTACAACTGTC	CA	2	CACACACACACACACACA	18	83
KNE796_003500	REVERSE	55.12	50	TCTACCGTCTCTTCGACCTA					84
KNE796_003563	FORWARD	55.67	42.86	CAACTCCATTAACTCAGCGTA	TC	2	TCTCTCTCTCTCTCTC	16	82
KNE796_003563	REVERSE	54.76	52.38	CACTGTAAGACCACTCACTCC					90
KNE796_003563	FORWARD	54.37	52.38	CAGCGTACTCATCTCTCTCTC	GT	2	GTGTGTGTGTGTGTGTGTGTGTGTG	30	79
KNE796_003563	REVERSE	54.76	52.38	CACTGTAAGACCACTCACTCC					90
KNE796_003626	FORWARD	54.66	55.56	CACCCTGAGTTGGATCTG	GT	2	GTGTGTGTGTGTGTGTGTGTGTGTG	44	84
KNE796_003626	REVERSE	60.3	60	GTAACGTACCGACCGACCAC					89
KNE796_003883	FORWARD	54.81	45	ATGCTCCAAAGTCTCAGTGT	GT	2	GTGTGTGTGTGTGTGTGTGTGTGTG	48	85
KNE796_003883	REVERSE	55.39	42.86	AAGATAGATGGAACCAGGAGA					83
KNE796_003937	FORWARD	55.34	55.56	TCTATCGCCTCCCTCTTC	CCG	3	CCGCCGCCGCCGCCG	15	84
KNE796_003937	REVERSE	55.92	47.62	GGTTAGTTACTCGGTGTCGTT					89
KNE796_004004	FORWARD	55.5	50	CGTCTCTAACAAGGGTCGTA	GT	2	GTGTGTGTGTGTGTGT	16	91
KNE796_004004	REVERSE	50.69	52.63	ACTCTACGAGTTCCTCCTC					83
KNE796_004013	FORWARD	56.1	57.89	CTAGACTCGTGCCAAGGAG	TG	2	TGTGTGTGTGTGTGTG	16	96
KNE796_004013	REVERSE	55.1	47.62	TTCTTCCTCCTACTGGATCTC					77

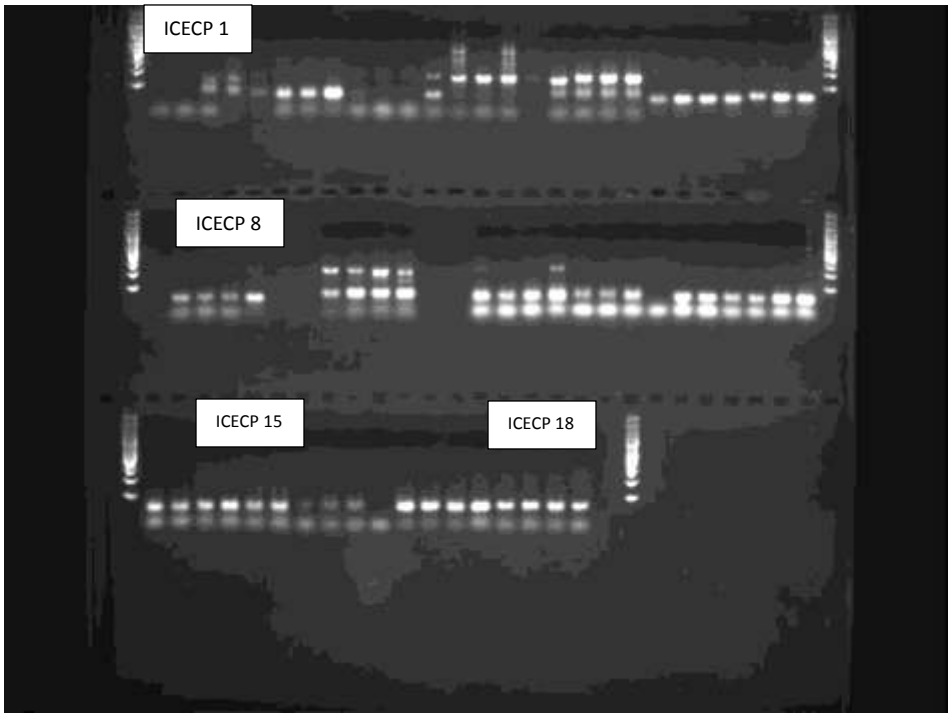
Seq ID	Orientation	tm	GC%	Seq	Motif	Motif Len	SSR	SSRLen	Primer Efficiency
KNE796_004033	FORWARD	54.88	50	GTAAGGTACTACGGTACGAAGG	AGTGTG	6	AGTGTGAGTGTGAGTGTG	18	88
KNE796_004033	REVERSE	54.92	47.62	CTCACTGCTACACTCACACAA					83
KNE796_004363	FORWARD	55.38	42.86	TAGATGCAAGATGTGAGTTCC	TG	2	TGTGTGTGTGTGTGTGTGTG	22	79
KNE796_004363	REVERSE	54.99	42.86	CCTTTCTTGGTGTGTGTAAAG					78
KNE796_004388	FORWARD	55.63	47.37	CTTCATGGGAGAAACTTGG	GA	2	GAGAGAGAGAGAGAGAGAGAGAGA	38	84
KNE796_004388	REVERSE	58.35	61.11	CTTGCCGCCTCTCTCTCT					79
KNE796_004661	FORWARD	55.51	42.86	GTGTGTGTGCATGAGTTGTTA	GT	2	GTGTGTGTGTGTGT	14	75
KNE796_004661	REVERSE	55.1	45	AAGGATTGACACAACCTGAG					89
KNE796_004870	FORWARD	59.05	61.11	CACCACACACGCACACAC	AC	2	ACACACACACACACACACACACACA	30	80
KNE796_004870	REVERSE	50.56	47.37	TGTGATGAGATAGGAGAGG					75
KNE796_004884	FORWARD	55.79	42.86	TTCTTGTGTGTGTGTGTGT	GA	2	GAGAGAGAGAGAGAGAGA	18	75
KNE796_004884	REVERSE	54.33	52.38	CCTGCTCTCTCAGCTCTAGTA					81
KNE796_004931	FORWARD	55.29	50	GAGATGGCGGTTGTGATA	GA	2	GAGAGAGAGAGAGAGAGAGAGAGA	28	87
KNE796_004931	REVERSE	54.98	52.63	CACCACTCCTCTCATCCTT					79
KNE796_004944	FORWARD	54.58	42.86	TCCTCTGGTGTGTATTTAAGC	TC	2	TCTCTCTCTCTCTCTCTCTCTC	24	79
KNE796_004944	REVERSE	54.47	50	CGACCGACCTACGTTATAGT					79
KNE796_004944	FORWARD	54.9	42.86	TCAATTAGCAGGTCAGTAAGC	GGTC	4	GGTCGGTCGGTCGGTCGGTC	20	78
KNE796_004944	REVERSE	56.09	61.11	CTCCGTGTGTCGTCGTAG					87
KNE796_004957	FORWARD	54.71	47.37	AGGAAGGACACAAGCTCAT	GA	2	GAGAGAGAGAGAGA	14	87
KNE796_004957	REVERSE	62.88	66.67	GGCGCTAACACCCCCTTC					87
Primers from contigs									
Seq ID	**Orientation**	**tm**	**GC%**	**Seq**	**Motif**	**Motif Len**	**SSR**	**SSRLen**	**Primer Efficiency**
796_rep_c33	FORWARD	55.02	42.86	CTTATACACCGTTGCTTCATC	TTG	3	ttgttgttgttgttgttg	18	81
796_rep_c33	REVERSE	55.06	52.63	AGGTTGTCGAACTGGAGAC					87

Appendix III: Primers from Ecogenics

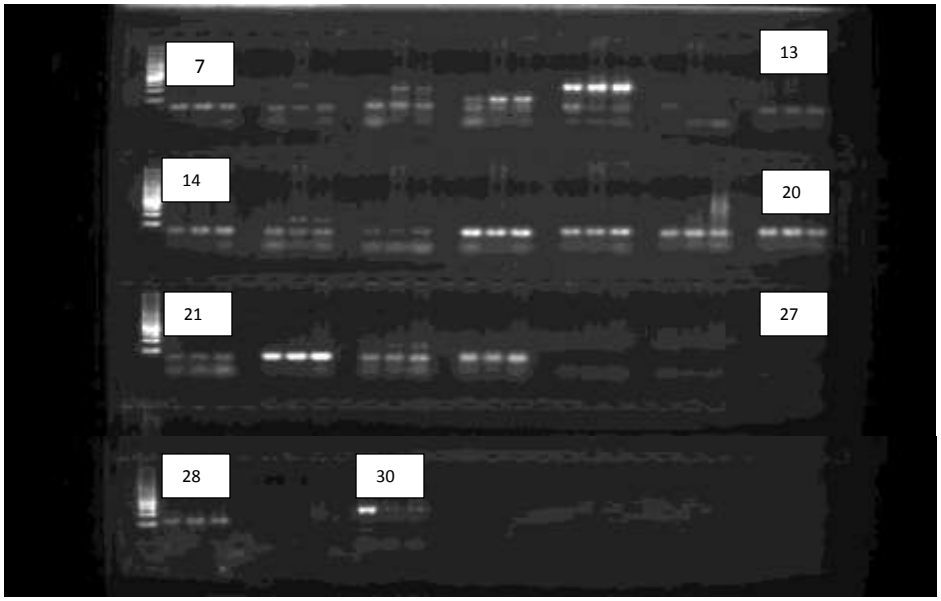
Name	Repeat motif	Repeat length	Amplicon size	Forward Primer Sequence	Reverse Primer Sequence
ICECP 1	(TG)	14	115	TCGTGCCCTTTCTCTCTCTC	TGCAATCTGATCAACAGCCG
ICECP 2	(CA)	22	104	TTGACTTCCAGTACCCCGAG	GTTAGTGGAAAGGACCAAGCAG
ICECP 3	(CA)	16	137	GCACATTGTTCCTCTTCCCTC	TGGTACATCTCAACATGGTTTCTC
ICECP 4	(TC)	11	119	TCATGGCTGACCTCCCTTTC	TGCATCCCTTTTTGCTCTACTC
ICECP 5	(CA)	14	123	CTCCTCTGGTCACCTTCCTC	CGAGGGAGAAAATCCGTTGG
ICECP 6	(TC)	13	120	TTCTCTCCCACGTCGTTGTC	ATGTCGAAGGAGGAGCGTC
ICECP 7	(TG)	13	124	GGTCGTGAGCTTGTCCATTC	AGAGAACCCAGAGGGAGTAG
ICECP 8	(AG)	12	217	AAACCCACCCACCCATACTC	TCTATCTGCTGCCTGCTCTG
ICECP 9	(AC)	11	177	TTTGGATCAGCCGTTTCGC	AGGTCTCCACATTCTCCAGC
ICECP 10	(AC)	11	87	TGTAACAAGATGGCCGGAGG	ATCTACCGCGCAAGTACCAG
ICECP 11	(AG)	19	219	CCAGAACCAAATCCGTCACC	ACCGAGCCCACTCAGTTATG
ICECP 12	(GT)	12	105	TGCTTTACTGGCTCGATCTC	GGAGGAGGACGACACAAAAC
ICECP 13	(AG)	14	145	GACCTTCTACTTGTTGTCTCTCC	TATGGCATGGAGACACTGGG
ICECP 14	(GA)	13	127	AACCCGTGGCAGTGTTGC	AACCCTTCCCACTGTTCCTC
ICECP 15	(GT)	15	111	GCTCCAAGTCTCTAGTTGTTGG	TCTCATCTCTATGGGCGTTCG
ICECP 16	(CT)	11	155	GCCTTCTTCCTCACAGCATC	CAGGAGTAGAGCATGGACGG
ICECP 17	(AC)	11	192	CATATGTAATAGCCAAACCCCC	GGGTTCCCGTTACCCTTACC
ICECP 18	(AG)	23	155	GCGGATCATGTTGGGCATAAG	CTCGTGTCGTCGTCGTAGAG
ICECP 19	(AG)	13	140	TGCTTGTTTGGATCTTGCTC	CGACCACCACTCTCTCTCTC
ICECP 20	(GA)	29	155	TGCTTGTTTGGATCTTGCTC	GTAGAGACGACCGACCGAC
ICECP 21	(CA)	17	102	CATGTGCACCCATGTGAACG	TGGCTGTGTCTGTGAGTGC
ICECP 22	(AG)	14	84	GTTAATCGTGCTGGTCCTCG	ACCCCAACGATGTACTACCC
ICECP 23	(CA)	14	115	TACCTTACTTATCCCGCGCC	CTGGACCTGGACGGATCG
ICECP 24	(CT)	12	240	TCAATTTTGGGCTCGCCTTC	ACAGAAGATTGAACAGTGCGG
ICECP 25	(AC)	13	115	TTCGATACGGGGCGAAAATG	CTGCTGCTGCTGCTCATC
ICECP 26	(GAT)	11	119	AGCAAGAGCAAGAGAAACAAG	GCCAAGTCCTCAGTCAAGTC
ICECP 27	(AG)	12	126	GTGAGGTTTGGTTGCGAGTG	TCGACCTCCCGTCCTCTG
ICECP 28	(CA)	12	98	AGACGACTTCACGCAGCTC	GACACGAGCCGAAATGGAAC
ICECP 29	(AG)	24	112	TACTGCTCTTGTAGCAGGGG	TAGTACGACGACGACCGACC
ICECP 30	(AG)	13	221	ACAGTAGGAAATAAGTAATTGCATGAG	TGGCATGAACGTGTTGACAG
ICECP 31	(AC)	11	131	CAATTGCCCCAAACTTGAATTG	AGGTACCCTGTGAAGTCATCC
ICECP 32	(TG)	11	88	CCCAAGCGTCTCTCTCTCTC	TGTTCATACCTTACCAAAATAAATC
ICECP 33	(AC)	12	82	ATCGGACATCCTTTTCTGGC	ACTGTGTGTGTTTGTCGTG
ICECP 34	(GTA)	11	165	ACACGGGAAAACGGGAAAAC	AGGAAGGAAGTATAGGGTAGGG
ICECP 35	(AG)	26	134	TAGAAAGACGGGAAGGAGCC	CGACCGACCACCACCTAC
ICECP 36	(GA)	11	91	GAAGTCCTCGTGTTGCATCC	CAAGCTCGGGCTCGTATAAC
ICECP 37	(TC)	12	93	CGACACGGTTTCCATTGGTC	GCAAAGAGGTAGCGAAGCTG
ICECP 38	(AC)	20	142	TGCATAGATTTGTGTTCTTTCTTGC	GGTCCGCATGTTAGTTACGTC
ICECP 39	(TG)	11	96	TCCGAGAAACTTTGTGATCGG	GACTCCCCCTATCCACGTC
ICECP 40	(AGG)	8	129	CAAGTCACCGTTGTCGCAG	ACTACTGCTTGCTCCTCGTC
ICECP 41	(AAG)	8	131	CAAGTCACCGTTGTCGCAG	ACACTACTGCTTGCTCCTCG
ICECP 42	(TG)	13	86	AGCTAGCCAGCCATTTGTCC	GCAAACGAACACAACAGTAACG
ICECP 43	(AG)	12	80	TCTTTTGCTTGTTTGGATCTTACTC	CTCGCTCTCTCTGTCTCTCTC
ICECP 44	(TG)	13	83	AGCTAGCCAGCCATTTGTCC	CAAACGACACAACAGTAACGC
ICECP 45	(AG)	18	137	TCGAATTATTGGCGCACAGG	TCTGTGAGACGACCACGAC
ICECP 46	(AC)	12	121	CCTCGTGTTGACGTGCAG	ACCACCGTGACCTTCCTTC
ICECP 47	(CA)	21	84	AATCACAGCAACCAGCAATC	TCGAGCTGTGTGTGAGTGAG
ICECP 48	(AC)	11	117	ACAGGAGCAGGTACAGATCG	TATCCAAACACAGCGTCAGC
ICECP 49	(GA)	12	163	AAACGCGAGAGAGTACAGGG	CTCGTGTCGTCGTCGTAGAG
ICECP 50	(AG)	17	119	CATCATTCTGCTTCCCCTGC	CGCGATGGATGATTGGATTG
ICECP 51	(AG)	11	86	TGCAAGCGTGGCTTCATTC	GGTCCGTCGTGTTCCTACTG
ICECP 52	(CA)	21	93	GGTGACCAGGATCATACCCC	GCTTTTACTTGAAGGCCCATTTG
ICECP 53	(CA)	12	114	TCTCAGTGGTATTTTGGTCCC	AGGTTGGTGGATCTGGATTTTG
ICECP 54	(AG)	20	106	CGCAAGCCGACAACAAAATG	AGAAGAAGCAGACACGACAG
ICECP 55	(AG)	11	190	CGTGTCGGGGGAGATAGAG	TTAAAACCCCCGGTAACCCC
ICECP 56	(TG)	11	120	ATCTCGTTGCATTCCGGTTG	TCAAGCCCTTATGCCCCC
ICECP 57	(CA)	11	118	TGATGGTGGTTGCCAGGTTC	CCAAGTGGTGTCGAGGAAAG
ICECP 58	(AC)	20	105	CGAATTCAGCTAGCGTGCC	GCTGAACCTTGTGCGGTG
ICECP 59	(CT)	11	110	GAGTCGCAATAGCTGAAGGC	TCAACGACCGGACGAAGAC
ICECP 60	(AC)	12	82	TAATTTGGGCTCGACCTTGC	CTTGTCGCGCGCATCATATC
ICECP 61	(GA)	11	88	CGCCGTGCTCACATCAGG	CCCTGCTTGTCAGTTCCTTC
ICECP 62	(TG)	11	96	ACAGACCTCCTCTCTCTCCC	ACCTGGATTGCTGATGGAGC
ICECP 63	(AG)	13	119	ACCAGATCCACCCACCATATC	ATCCGTCCCCTCTCTCCTAC
ICECP 64	(AC)	17	96	AGAAAACGGGAAAGATCCCAG	TTTCTCTGGCACCAGCAATC
ICECP 65	(GA)	15	151	ACACTAGAAAGAGAGAGAGAGTG	GTAGTAGACGACCGACCGAC
ICECP 66	(AC)	13	100	AGACAGCAGTTGTACCATCAC	CATGTTACGGAGAGGGGTCG
ICECP 67	(AG)	12	160	TGGAGAGAGAGTATCGTTTTGC	TACGTACGGAACGGAACGG
ICECP 68	(AC)	15	83	TGCATGTCATTAAAATGTATGTGTG	CTCAAAGCACCTCACAAGGC
ICECP 69	(CA)	12	120	TCACACGCACACCCACAC	AGGCTCATATGTAATCCTAACCTA
ICECP 70	(TC)	11	96	TCAAGCTCGGGCTCGTATAG	AACTTGAAACCCGACGTTGC
ICECP 71	(AG)	12	98	CTGAACGAAGGCCTGTTTCC	TGTGTTGATGTTGGGTGTGC
ICECP 72	(TG)	13	98	TCCAATTCTGTCCCATCCCC	CGAACCACGTTGCTCACATC
ICECP 73	(AC)	11	93	TGTTTCTGTGAGCTATCTTTGGC	AAGACCCTGGTACGCATCTC
ICECP 74	(GA)	11	191	ATGCTCTTTCTCACGGAGCG	CGTGCGTTTGTCAAGTAGGG
ICECP 75	(CA)	13	118	CAGCATCCTCCAAAAGAGGC	GATCCATGTTAGCGTGCGTG
ICECP 76	(GT)	15	250	CCCATGACTACCGACAACCG	TAAATAGTCCGCCTCCCGTC
ICECP 77	(AC)	14	104	CCGTCTCCCAATCCTACCTG	GGTGGTGGTTTCAGGCTTTC
ICECP 78	(GT)	22	112	CCACCCTGAGTTGGATCTGG	TAACGTACCGACCGACCAC
ICECP 79	(CA)	13	111	CGGAACCAAGAGACAATCGC	ATACCCAAGCCACAACATGC

Name	Repeat motif	Repeat length	Amplicon size	Forward Primer Sequence	Reverse Primer Sequence
ICECP 80	(GA)	13	97	ACGAGAAAATGGAATCGCGG	CTTCTTGCTTGTGTCCTGCC
ICECP 81	(TG)	18	128	TTCTAGCAGTGTGTGTGTGC	AATGAAGCAGTGGGAGGGTC
ICECP 82	(GT)	24	119	AGGGGATGCTCCAAAGTCTC	GGAACCAGGAGACAATACGC
ICECP 83	(CT)	13	149	TCGCTCTCTCTCTCTGTGTG	ATTTGTCTGTCTCGTATCTTACTAC
ICECP 84	(TC)	12	103	TTCATGTGATGCAAGCCACC	CAGGCAAGATGCTGTTCTGG
ICECP 85	(TC)	14	96	CAGCAGCATCTATTTTCCATTGAC	GAAGAGAGGGAGCTTCGCC
ICECP 86	(TG)	11	80	TGTGAGTTCCTCTCTCTCTCTC	CCTAAGCAGGTTGCGTCTTG
ICECP 87	(AC)	12	80	ATCACTAAAGACCATAGCCAACC	CGTCATTCCTCTAGCGTGTG
ICECP 88	(GA)	19	158	GCTTCATGGGAGAAACTTGGG	CTTGCCGCCTCTCTCTCTG
ICECP 89	(GA)	11	128	CCACCAGAATCCAATGGCAC	TCGACTTTGTTTGCATGCTG
ICECP 90	(TG)	17	220	ATTCATCGACTCCCCAGTCC	AGCATGGACGAAGCGAAATC
ICECP 91	(GT)	11	127	AGATGAAAATGACTCGGTCTTGAG	CTTCTCAGTCCTTCACCCCC
ICECP 92	(AG)	26	237	CCCGTTTCCACCATCACAAC	GTACGACGACGACCGACC
ICECP 93	(AC)	12	167	AAGGAAGGAGAGGGCTCCAC	AGGGCCCACAGATAAACCTC
ICECP 94	(GT)	12	128	TGCTGGAGATCGCTGAAAAG	CGACCTTGCTTGCAAAAACC
ICECP 95	(GA)	14	101	GATGGCGGTTGTGATATACGG	GTCACCACTCCTCTCATCC
ICECP 96	(TC)	12	134	AGCAGGTCAGTAAGCTAGGC	CTCCGTGTGTCGTCGTAGAG
ICECP 97	(AG)	12	122	GTTACTTGGTAAACCGCCCC	AGTGTAGGAGTCAAAACAAAGC
ICECP 98	(AC)	12	191	CGAGTGAGTGTTCGTGTGTG	TGCATGAAATAGATGGGCCG
ICECP 99	(TC)	11(TG)13	125	ATCGACCTTTCCCTTCCTCC	TACTACAAGGGAGTTGGGCG
ICECP 100	(CT)	14	83	CACTCTCTCTCGCTCTCTCAC	CCGACCACGCACCGATTG
ICECP 101	(AAC)	16	182	AGGTTGTCGAACTGGAGACC	TAGCTGACCTATCGACGTGC

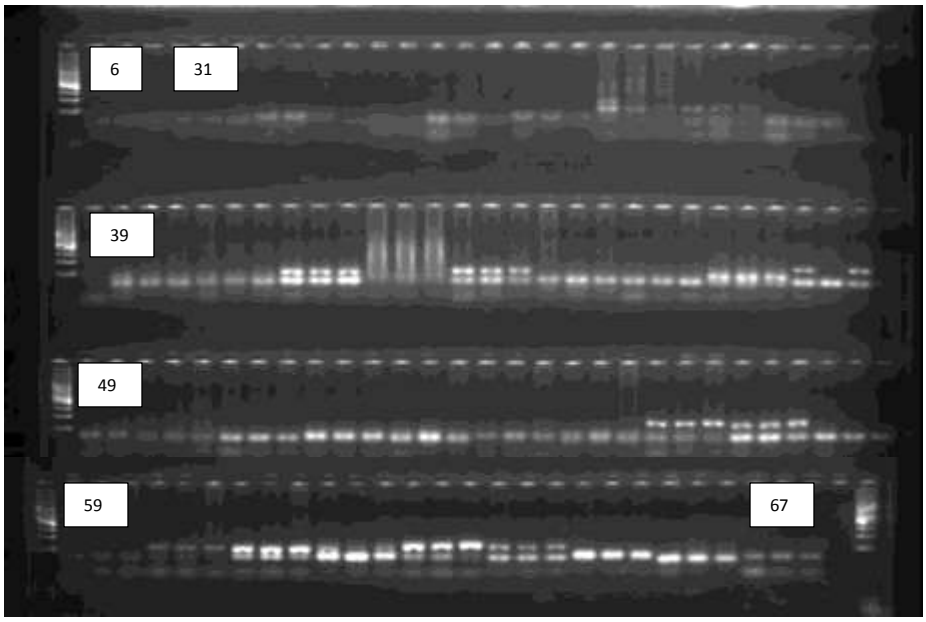
Appendix IV: Gel images



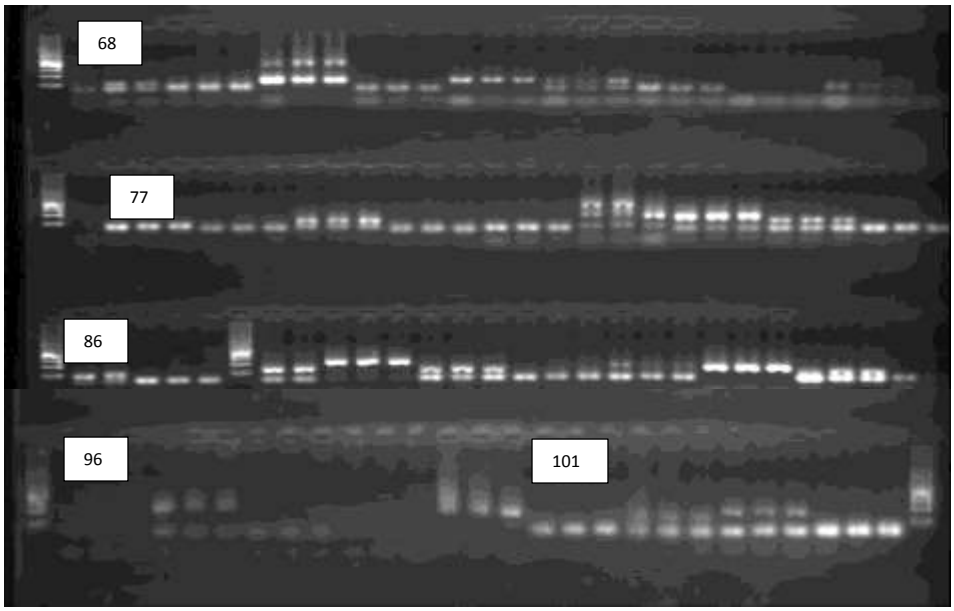
ICECP 1-ICECP 18. The markers were loaded in groups of 4, with ICECP 1 represented by the first four wells at the top left corner of the gel.



ICECP 7-ICECP 30. The markers were loaded in groups of 3, with ICECP 7 represented by the first three wells at the top left corner of the gel.



ICECP 6 and ICECP 31-ICECP 67. The markers were loaded in groups of 3, with ICECP 6 represented by the first three wells at the top left corner of the gel and ICECP 31 by the next three wells.



ICECP 68-ICECP 101. The markers were loaded in groups of 3, with ICECP 68 represented by the first three wells at the top left corner of the gel.

Appendix V: PCR results for the markers

Polymorphic	Monomorphic	Didn't work	Difficult to score,	worked<100%>50%	worked<50%
ICECP 1	ICECP 9	ICECP 2	ICECP 83	ICECP 1 (8/10)	ICECP 9 (2/10)
ICECP 3	ICECP 23	ICECP 6		ICECP 3 (6/10)	ICECP 11 (2/10)
ICECP 4	ICECP 24	ICECP 7		ICECP 42 (8/10)	ICECP 23 (1/10)
ICECP 5	ICECP 31	ICECP 8		ICECP 44 (8/10)	ICECP 24 (1/10)
ICECP 11	ICECP 35	ICECP 10		ICECP 45 (8/10)	ICECP 35 (1/10)
ICECP 37	ICECP 36	ICECP 12		ICECP 46 (7/10)	ICECP 67 (2/10)
ICECP 40	ICECP 41	ICECP 13		ICECP 47 (7/10)	ICECP 82 (3/10)
ICECP 42	ICECP 45	ICECP 14		ICECP 48 (9/10)	ICECP 92 (3/10)
ICECP 43	ICECP 49	ICECP 15		ICECP 49 (5/10)	ICECP 97 (2/10)
ICECP 44	ICECP 94	ICECP 16		ICECP 52 (7/10)	
ICECP 46		ICECP 17		ICECP 68 (5/10)	
ICECP 47		ICECP 18		ICECP 93 (9/10)	
ICECP 48		ICECP 19		ICECP 96 (9/10)	
ICECP 50		ICECP 20		ICEIP 1 (7/10)	
ICECP 52		ICECP 21		ICEIP 2 (9/10)	
ICECP 53		ICECP 22		ICEIP 3 (9/10)	
ICECP 54		ICECP 25			
ICECP 56		ICECP 26			
ICECP 57		ICECP 27			
ICECP 58		ICECP 28			
ICECP 59		ICECP 29			
ICECP 61		ICECP 30			
ICECP 62		ICECP 32			
ICECP 63		ICECP 33			
ICECP 64		ICECP 34			
ICECP 66		ICECP 38			
ICECP 67		ICECP 39			
ICECP 68		ICECP 51			
ICECP 69		ICECP 55			
ICECP 70		ICECP 60			
ICECP 71		ICECP 65			
ICECP 72		ICECP 74			
ICECP 73		ICECP 75			
ICECP 80		ICECP 76			
ICECP 81		ICECP 77			
ICECP 82		ICECP 78			
ICECP 84		ICECP 79			
ICECP 85		ICECP 86			
ICECP 89		ICECP 87			
ICECP 90		ICECP 88			
ICECP 91		ICECP 100			
ICECP 92					
ICECP 93					
ICECP 95					

ICECP 96					
ICECP 97					
ICECP 98					
ICECP 99					
ICECP 101					

Appendix VI: Summary statistics for PowerMarker output

Marker	Maj Allele Freq	GenotypeNo	SampleSiz	No. of obs.	AlleleNo	Availability	GeneDiversity	Heterozygc	PIC
ICECP 2	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 6	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 7	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 8	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 10	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 12	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 13	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 14	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 15	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 16	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 17	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 18	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 19	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 20	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 21	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 22	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 25	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 26	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 27	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 28	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 29	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 30	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 32	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 33	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 34	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 38	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 39	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 51	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 55	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 60	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 65	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 74	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 75	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 76	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 77	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 78	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 79	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 83	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 86	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 87	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 88	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 100	0.0000	0.0000	10.0000	0.0000	0.0000	0.0000	1.0000	NaN	1.0000
ICECP 54	0.3000	6.0000	10.0000	10.0000	6.0000	1.0000	0.8000	0.0000	0.7716
ICECP 47	0.2857	5.0000	10.0000	7.0000	5.0000	0.7000	0.7755	0.0000	0.7397
ICECP 89	0.3500	6.0000	10.0000	10.0000	5.0000	1.0000	0.7200	0.7000	0.6722
ICECP 50	0.4500	5.0000	10.0000	10.0000	4.0000	1.0000	0.6750	0.1000	0.6191
ICECP 58	0.5000	4.0000	10.0000	10.0000	4.0000	1.0000	0.6600	0.0000	0.6102
ICECP 84	0.4500	3.0000	10.0000	10.0000	4.0000	1.0000	0.6650	1.0000	0.6035

Marker	Maj Allele Freq	GenotypeNo	SampleSiz	No. of obs.	AlleleNo	Availability	GeneDiversity	Heterozygc	PIC
ICECP 5	0.5000	5.0000	10.0000	10.0000	4.0000	1.0000	0.6350	0.8000	0.5729
ICECP 96	0.4444	3.0000	10.0000	9.0000	3.0000	0.9000	0.6420	0.0000	0.5676
ICECP 3	0.5000	3.0000	10.0000	6.0000	3.0000	0.6000	0.6111	0.0000	0.5355
ICECP 95	0.6000	4.0000	10.0000	10.0000	4.0000	1.0000	0.5800	0.0000	0.5350
ICECP 4	0.5500	3.0000	10.0000	10.0000	3.0000	1.0000	0.5950	0.9000	0.5280
ICECP 68	0.6000	3.0000	10.0000	5.0000	3.0000	0.5000	0.5600	0.0000	0.4992
ICECP 73	0.5000	2.0000	10.0000	10.0000	3.0000	1.0000	0.5800	1.0000	0.4918
ICECP 53	0.6000	3.0000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 63	0.6000	3.0000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 64	0.6000	3.0000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 90	0.6000	3.0000	10.0000	10.0000	3.0000	1.0000	0.5400	0.0000	0.4662
ICECP 61	0.7000	4.0000	10.0000	10.0000	4.0000	1.0000	0.4800	0.0000	0.4500
ICECP 62	0.7000	4.0000	10.0000	10.0000	4.0000	1.0000	0.4800	0.0000	0.4500
ICECP 37	0.7000	3.0000	10.0000	10.0000	3.0000	1.0000	0.4600	0.0000	0.4102
ICECP 69	0.7000	3.0000	10.0000	10.0000	3.0000	1.0000	0.4600	0.0000	0.4102
ICECP 66	0.7500	4.0000	10.0000	10.0000	4.0000	1.0000	0.4150	0.2000	0.3894
ICECP 11	0.5000	2.0000	10.0000	2.0000	2.0000	0.2000	0.5000	0.0000	0.3750
ICECP 67	0.5000	2.0000	10.0000	2.0000	2.0000	0.2000	0.5000	0.0000	0.3750
ICECP 70	0.5000	2.0000	10.0000	10.0000	2.0000	1.0000	0.5000	0.0000	0.3750
ICECP 71	0.5000	2.0000	10.0000	10.0000	2.0000	1.0000	0.5000	0.0000	0.3750
ICECP 97	0.5000	2.0000	10.0000	2.0000	2.0000	0.2000	0.5000	0.0000	0.3750
ICECP 46	0.5714	2.0000	10.0000	7.0000	2.0000	0.7000	0.4898	0.0000	0.3698
ICECP 40	0.6000	2.0000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 85	0.6000	2.0000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 98	0.6000	2.0000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 99	0.6000	2.0000	10.0000	10.0000	2.0000	1.0000	0.4800	0.0000	0.3648
ICECP 42	0.6250	2.0000	10.0000	8.0000	2.0000	0.8000	0.4688	0.0000	0.3589
ICECP 44	0.6667	2.0000	10.0000	9.0000	2.0000	0.9000	0.4444	0.0000	0.3457
ICECP 59	0.6667	2.0000	10.0000	9.0000	2.0000	0.9000	0.4444	0.0000	0.3457
ICECP 82	0.6667	2.0000	10.0000	3.0000	2.0000	0.3000	0.4444	0.0000	0.3457
ICECP 92	0.6667	2.0000	10.0000	3.0000	2.0000	0.3000	0.4444	0.0000	0.3457
ICECP 93	0.7778	3.0000	10.0000	9.0000	3.0000	0.9000	0.3704	0.0000	0.3402
ICECP 56	0.7000	2.0000	10.0000	10.0000	2.0000	1.0000	0.4200	0.0000	0.3318
ICECP 52	0.7143	2.0000	10.0000	7.0000	2.0000	0.7000	0.4082	0.0000	0.3249
ICECP 72	0.8000	3.0000	10.0000	10.0000	3.0000	1.0000	0.3400	0.0000	0.3142
ICECP 80	0.8000	3.0000	10.0000	10.0000	3.0000	1.0000	0.3400	0.0000	0.3142
ICECP 101	0.8000	3.0000	10.0000	10.0000	3.0000	1.0000	0.3400	0.0000	0.3142
ICECP 1	0.7500	2.0000	10.0000	8.0000	2.0000	0.8000	0.3750	0.0000	0.3047
ICECP 43	0.8000	2.0000	10.0000	10.0000	2.0000	1.0000	0.3200	0.0000	0.2688
ICECP 48	0.8889	2.0000	10.0000	9.0000	2.0000	0.9000	0.1975	0.0000	0.1780
ICECP 57	0.9000	2.0000	10.0000	10.0000	2.0000	1.0000	0.1800	0.0000	0.1638
ICECP 81	0.9000	2.0000	10.0000	10.0000	2.0000	1.0000	0.1800	0.0000	0.1638
ICECP 91	0.9000	2.0000	10.0000	10.0000	2.0000	1.0000	0.1800	0.0000	0.1638
ICECP 9	1.0000	1.0000	10.0000	2.0000	1.0000	0.2000	0.0000	0.0000	0.0000
ICECP 23	1.0000	1.0000	10.0000	1.0000	1.0000	0.1000	0.0000	0.0000	0.0000
ICECP 24	1.0000	1.0000	10.0000	1.0000	1.0000	0.1000	0.0000	0.0000	0.0000
ICECP 31	1.0000	1.0000	10.0000	10.0000	1.0000	1.0000	0.0000	0.0000	0.0000
ICECP 35	1.0000	1.0000	10.0000	1.0000	1.0000	0.1000	0.0000	0.0000	0.0000
ICECP 36	1.0000	1.0000	10.0000	10.0000	1.0000	1.0000	0.0000	0.0000	0.0000

Marker	Maj Allele Freq	GenotypeNo	SampleSiz	No. of obs.	AlleleNo	Availability	GeneDiversity	Heterozygc	PIC
ICECP 41	1.0000	1.0000	10.0000	10.0000	1.0000	1.0000	0.0000	0.0000	0.0000
ICECP 45	1.0000	1.0000	10.0000	9.0000	1.0000	0.9000	0.0000	0.0000	0.0000
ICECP 49	1.0000	1.0000	10.0000	5.0000	1.0000	0.5000	0.0000	0.0000	0.0000
ICECP 94	1.0000	1.0000	10.0000	10.0000	1.0000	1.0000	0.0000	0.0000	0.0000
Mean	0.4046	1.5048	10.0000	4.8476	1.4952	0.4848	0.6489	NaN	0.6144