MOLECULAR MARKER DEVELOPMENT FOR CASSAVA MOSAIC DISEASE RESISTANCE USING BIOINFORMATICS TOOLS

by

AMBU VIJAYAN

(2010-09-105)

THESIS

Submitted in partial fulfilment of the

requirement for the degree of

MASTER OF SCIENCE (INTEGRATED) INBIOTECHNOLOGY

Faculty of Agriculture

Kerala Agricultural University, Thrissur

B.Sc.-M.Sc. (INTEGRATED) BIOTECHNOLOGY
DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM-695 522

KERALA, INDIA

2015

DECLARATION

I hereby declare that this thesis entitled **"Molecular marker development for cassava mosaic disease resistance using bioinformatics tool** "is a bonafide record of research work done by me during the course of research and that the thesis has not previously formed the basis for the award of any degree, diploma, associate ship, fellowship or other similar title, of any other university or society.

Vellayani
Date: 25.08.2015

AMBU VIJAYAN
(2010-09-105)

Produce tuber reduce hunger

भाकृ अनुप-केन्द्रीय कन्द फसल अनुसंधान संस्थान
(भारतीय कृषि अनुसंधान परिषद)
श्रीकार्यम, तिरुवनन्तपुरम -६९५०१७, केरल,भारत

ICAR-CENTRAL TUBER CROPS RESEARCH INSTITUTE
(Indian Council of Agricultural Research)
Sreekariyam, Thiruvananthapuram-695 017, Kerala, India

ISO 9001:2008

Dr. J. Sreekumar
Senior Scientist

CERTIFICATE

Certified that this thesis entitled **"Molecular marker development for cassava mosaic disease resistance using bioinformatics tool"** is a record of research work done independently by Mr. Ambu Vijayan (2010-09-105) under my guidance and supervision and that it has not previously formed the basis for the award of any degree, diploma, fellowship or associateship to him.

Place : CTCRI, Trivandurm
Date : 11.01.2016

Dr. J. Sreekumar
(Chairman of the Advisory Committee)
Senior Scientist (Agrl. Statistics)

Phone : 91-0471-2598551 to 2598554
Director (Per.) : 91-0471-2598431
(Res.) : 91-0471 - 2441957
Admn.Officer : 91-0471-2598193

RIGHT TO INFORMATION

Fax : 91-0471-2590063
E-Mail : ctcritvm@yahoo.com
ctcritvm@gmail.com
Web : http://www.ctcri.org

CERTIFICATE

We, the undersigned members of the advisory committee of Mr. Ambu Vijayan, a candidate for the degree of Master of Science (Integrated) in Biotechnology, agree that the thesis entitled "**Molecular marker development for cassava mosaic disease resistance using bioinformatics tools**" may be submitted by Mr. Ambu Vijayan, in partial fulfillment of the requirement for the degree.

11.01.16

Dr. J. Sreekumar
(Chairman, Advisory Committee)
Senior Scientist
Section of Extension and Social Sciences
Central Tuber Crops Research Institute
Sreekariyam, Thiruvananthapuram-695 017

11/1/16

Dr. B. R. Reghunath
(Member, Advisory Committee)
Professor and Head
Department of Plant Biotechnology
College of Agriculture, Vellayani

11/01/16

Dr. M. N. Sheela
Principal Scientist and Head
Division of Crop Improvement
Central Tuber Crops Research Institute
Sreekariyam
Thiruvananthapuram-695 017

11/1/16

Dr. C. Mohan
(Member, Advisory Committee)
Principal Scientist,
Division of Crop Improvement
Central Tuber Crops Research Institute
Sreekariyam

ACKNOWLEDGEMENT

It is with my heartfelt feelings, I wish to express my deep sense of gratitude and sincere thanks to Dr. J. Sreekmar, Senior Scientist ICAR-Central Tuber Crops Research Institute, and his valuable guidance, sincere help, patience, support and encouragement throughout the course of study.

My sincere gratitude to Dr. C. Mohan, Principal Scientist, ICAR-CTCRI. I express my indebtedness to sir for his guidance, incessant inspiration, untiring attention and patience towards my wet lab experiments while devoting his precious time in the midst of his busy schedule.

I express my sincere gratitude to Dr. S. K. Chakrabarti for giving me an opportunity to work at ICAR-CTCRI for my thesis work.

I am thankful to the members of the advisory committee, Dr. B. R. Reghunath, Professor and Head, Department of Plant Biotechnology, College of Agriculture, Vellayani, Dr. M. N. Sheela Principal Scientist and Head, Division of Crop Improvement, Central Tuber Crops Research Institute for their valuable counseling and constructive suggestions that were much helpful throughout my research progress.

I am indebted to all my teachers for giving me the confidence to walk ahead. I am thankful to Lekha mam, Sony mam, Swapna mam, Deepa mam and Geetha mam for their critical professional advice and direction throughout my research.

I take immense pleasure to express my thanks to Ms. Vidya. P and Aswathy G. H. Nair for providing me with space and being there for me always, rendering whatever help i needed during my wet lab experiments and for providing me advice and unwavering encouragement throughout my research work.

I express my heartfelt thanks to Dr. L. Rajamony, Dean, COA, Vellayani and also Dr. Sverup John, former Dean, COA, Vellayani for all the facilities provided.

I thankfully remember my best friends particularly Molu T. Lalson, Rajesh Krishnan R. K and my sisters Karthika Venugopal and Aarya Satheesan who have rendered their helping hands at all times and support during my research work.

With a grateful heart I thank the help, moral support and encouragement rendered by Abhijith Syam, Jayakrishnan K., Bharat Mohan and Anoop A.

I wish to express my deep gratitude to all non-teaching staff members of ICAR-CTCRI, for their timely help.

My special thanks to Amritha M. S and Dhanya O. G for helping me in completing this research work.

I thank all my beloved folks Sreekuttan, Edwin, Roshan, Sudheer, Aswathy, Rajalekshmi, Jayalekshmi and my seniors Sudheep, Achuth, Anoop, Adil, Athul, Nandu, Arun, Abeesh, Anjana Pamitha and leen.

At last I owe this achievement to my parents, K. Vijayakumar and Sreekala S. who always stood along my side and I will never forget the timely help, mental support, kindness and affection extended by Archa and Adarsh without them this work would have never seen light.

Finally, I take this opportunity to thank my brothers Mr. Mahendran, Mr. Vishnu and Mr. Jayakrishnan for their encouragement and moral support.

I acknowledge the favour of numerous persons who, though not been individually mentioned here, who have all directly or indirectly contributed to this thesis work.

AMBU VIJAYAN

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PLATES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND SYMBOLS USED

CMD	Cassava mosaic disease
SSR	Simple sequence repeats
SNP	Single nucleotide polymorphism
DNA	Deoxyribose nucleic acid
EST	Expressed sequence tag
NCBI	National Center for Biotechnology Information
cSNP	coding Single nucleotide polymorphism
ncSNP	Non coding Single nucleotide polymorphism
ORF	Open reading frames
CMV	Cassava mosaic virus
CMD	Cassava mosaic disease
EMBL	European Molecular Biology Laboratory
MAS	Marker assisted selection
%	Per cent
mM	millimolar
µl	Micro litre

@	At the rate of
°C	Degree Celsius
bp	Base pair
et al.	And other co workers
Fig.	Figure
g	Gram
g-1	Per gram
mg	Milli gram
ml	Millilitre
sec	Seconds
min	Minutes
ng	Nanogram

INTRODUCTION

1. INTRODUCTION

Cassava, (*Manihot esculenta* Crantz) (2n = 36) Family: Euphorbiaceae, which originated in Latin America is one of the most important food crops with a worldwide production of 270,293,801 tonnes. About 146,824,969 tonnes is produced in Africa, 90,372,457 in Asia where India had a production of 8,139,430 tonnes (FAOSTAT 2014). Various traits of the crop such as drought tolerance, heat tolerance and less requirement for agricultural fertilizers makes it an attractive crop. Cassava has monoecious flowering nature and so self- pollination in cassava is mainly prevented by protogyny (Alves, 2002) thus rendering the crop highly heterozygous.

Cassava is an essential staple food for over 700 million people all over the tropical and sub-tropical regions of the world. It can be grown all year round and provides food in periods of scarcity. The high starch content (20-40%) makes cassava a desirable energy source both for human consumption and industrial biofuel applications (Schmitz & Kavallari, 2009). Cassava is one of the most commonly used raw materials for the production of starch. High purity, low production costs, distinctive characteristics like clear viscous paste nature has made many industries adopt cassava starch as an alternative to more traditional sources like potato and maize. It is known for its drought tolerance and stable productivity even when cultivated in soils of low fertility.

Cassava originated in America and was transferred by the Portuguese to the rest of the world, particularly to the African continent, in the sixteenth century. In Africa and Latin America, cassava has been cultivated and produced for many centuries by small farmers, which has resulted in a large number of local cassava varieties. The genome of cassava is approximately 770 *Mb* (Awoleye et al., 1994), and the draft genome sequence of cassava was created through the whole genome shotgun strategy.

The cassava genome is predicted to contain 30,666 genes (Prochnik *et al.*, 2012). However, the function of most of the genes remains unclear. Cassava mosaic

disease (CMD) is the single most important disease affecting cassava cultivation. Economic losses due to CMD is estimated at US$ 1.5 billion annually in Africa. CMD is caused by gemini viruses of the genus Begomovirus (Family Geminiviridae) transmitted by a vector, white fly [*Bemisia tabaci, (Gennadius)*].

Expressed sequence tags (ESTs), which are short (300–500 bp) single read sequences from random cDNA clones, have a wide range of applications including the use as gene cloning reservoirs, evaluation of expression of tissue-specific gene, molecular markers for map based cloning and genomic sequence annotation. EST data have also led to a better understanding of both the existence and expression patterns of alternative transcripts and of coordinated gene expression. EST data represents a potentially significant resource for the detection of single nucleotide polymorphism (SNPs) in plants (Batley *et al.*, 2003). One tool which permits the disentanglement of the complications of gene expression is the analysis of ESTs. This strategy has advanced into an economical and capable gene discovery methodology. (Ohlrogge & Benning, 2000). About 74316793 million ESTs are available at the EST database of National Center for Biotechnology Information (NCBI).

Single nucleotide polymorphisms (SNPs) are markers of choice for high-density genetic mapping due to their sheer abundance in the genome (Rafalski, 2002). SNPs are known to occur at a rate of one per 100–500 bp in plant genomes, depending on the species. The advancements in sequencing ability along with the saving in sequencing cost allow for effective genome-wide discovery of SNPs. Organisms which have large genomes such as cassava, transcriptome sequencing (RNA-seq) offers an efficient way to restrict the sequencing to the expressed portion of the genome while it still identifies a large amount of genetic variation (Chepelev *et al.*, 2009). Considerable improvement on genomic resources for cassava, greatly achieved through the sequencing of the cassava genome (Prochnik *et al.*, 2012), greatly facilitates the characterization of variability within a crop by high throughput re-sequencing. RNA-seq has been successfully applied to large-

scale SNP discovery and EST- derived SNP development in various plant species (Paritosh *et al.*, 2013; Ferguson *et al.*, 2012).

Single sequence repeats (SSRs), also known as microsatellites are one of the most common and multipurpose marker type used in plant genetic mapping studies because of its advantageous features such as high abundance rate, specificity of locus, codominant inheritance, high information rate about polymorphism, and reproducibility (Varshney *et al.*, 2005). According to the origins of SSRs, they are divided into two categories: genomic SSRs or genic SSRs (EST-SSRs). Genic SSRs can be developed by screening the collection of clustered ESTs in publicly available databases and also they are derived from the expressed regions of the genome, so that they have increased potentials for tagging and mapping of genes and quantitative trait loci (QTLs).

The present study was undertaken to computationally develop SNPs and SSRs for cassava mosaic disease resistance and to understand the effectiveness of molecular markers in cassava in biotic stress response (cassava mosaic virus). SNP and SSR development tools were also evaluated to understand their performance.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

Cassava (*Manihot esculenta* Crantz, Euphorbiaceae) is grown for its starch-containing tubers, which feed over 500 million people worldwide and is the third most important food crop after cereals and grain legumes (Puonti-Kaerlas., 2001). In India it is grown in an area of 2.4×10^5 hectares both for direct consumption and starch grain (sago)-producing industries, mainly in the southern states of India. The major constraint for cassava production in Africa and the Indian subcontinent was the cassava mosaic disease (CMD) caused by gemini viruses in the genus Begomovirus (family Geminiviridae) (Hong *et al.,* 1993). The genomes of most gemini viruses are bipartite, termed DNA A and DNA B, the former encodes functions associated with viral replication and encapsidated and the latter encoding movement functions (Harrison & Robinson, 1999).

2.1 Cassava Mosaic Virus

Gemini viruses are serious plant pathogens, infecting a wide range of important crop plants in tropical, subtropical and, to a lesser extent, in temperate regions. The family Geminiviridae is divided into four genera on the basis of genome organization and biological properties. All have circular single-stranded DNA encapsidated in twinned (geminate) particles of approximately 20×30 nm (Bottcher *et al.*, 2004; Zhang *et al.*, 2001). The genus Begomovirus comprises more than 100 members which are transmitted by the whitefly *Bemisia tabaci* (Genn.) to dicotyledonous host plants. While most possess a bipartite genome, some begomoviruses from the old World have a monopartite genome (Rothenstein *et al.*, 2006).

CMD is caused by a number of begomoviruses representing distinct species, such as African cassava mosaic virus (ACMV), East African cassava mosaic virus (EACMV), East African cassava mosaic Cameroon virus, East African cassava mosaic Zanzibar virus and South African cassava mosaic virus (Berrie *et al.*, 2001). The causative agent of CMD in India is believed to be Indian cassava mosaic virus, ICMV (Hong *et al.*, 1993). Complete nucleotide sequencing of two cloned ICMV

DNAs, one from the southern state of Kerala (Hong *et al.*, 1993) and another from the central state of Maharashtra (Saunders *et al.*, 2002), showed that they were highly similar to each other, indicating them to be isolates of the same virus. In contrast, another distinct cassava infecting geminivirus (CIG) was reported from Sri Lanka, named Sri Lankan cassava mosaic virus (SLCMV), which has much lower sequence homology to ICMV (Saunders *et al.*, 2002). SLCMV had properties of a monopartite begomovirus, which reportedly captured the DNA B of ICMV following a recombination event (Saunders *et al.*, 2002). Using Polymerase Chain Reaction (PCR) analysis to specifically amplify parts of ICMV and SLCMV DNA A. In addition, by PCR-restriction fragment length polymorphism (PCR- RFLP) analysis, the presence of several novel forms of the above viral DNAs, whose partial sequence analyses indicate that they have probably arisen by accumulating random point mutations (Patil *et al.*, 2005).

2.2 Expressed Sequence Tags

Expressed sequence tag (EST) databases have become particularly attractive resources for such in silico mining, as was demonstrated in citrus (Chen *et al.*, 2006), coffee (Aggarwal *et al.*, 2007; Poncet *et al.*, 2006), sugarcane (Pinto *et al.*, 2004), sunflower. (Heesacker *et al.*, 2008; Pashley *et al.*, 2006) and particularly in the cereals (Kantety *et al.*, 2002; Thiel *et al.*, 2003; Yu *et al.*, 2004).

Several cassava genes encoding putative enzymes that may be involved in starch biosynthesis, such as soluble starch synthase (SSIII), starch phosphorylase and 1,4-a-glucan branching enzyme, have also been identified through EST sequencing (Lopez *et al.*, 2004). Processes of starch formation and storage in cassava roots are far more complex than starch biosynthesis since starch synthesized in the leaves has to be translocated to roots for maturation and storage (Alves & Setter, 2004). Starch accumulation in roots has been observed to occur as early as 25–40 DAP in some cassava cultivars (El-Sharkawy, 2004), suggesting that starch formation in cassava roots is very complex and a continuing process closely related to cassava growth and development. The dehydration-stress library uncovered numerous ESTs with recognized roles in drought-responses, including

those that encode late embryogenesis abundant proteins (LEA) thought to confer osmoprotective functions during water stress, transcription factors, heat-shock proteins as well as proteins involved in signal transduction and oxidative stress. The unigene clusters were screened for short tandem repeats for further development as microsatellite markers (Lokko *et al.*, 2007).

Since gene expression profiles of key growth and developmental stages of cassava becomes a prerequisite. large-scale analysis technologies such as genome-wide EST sequencing have been applied to identify related genes and/or pathways (Anderson *et al.*, 2004; Lokko *et al.*, 2007; Lopez *et al.*, 2004; Reilly *et al.*, 2007; Sakurai *et al.*, 2007). All these studies have indicated that very complex molecular networks govern cassava starch formation, and the molecular mechanisms controlling cassava growth and development are still poorly understood (Li *et al.*, 2010b).

In EST sequencing, sequencing redundancy usually reflects the average EST counts per gene which is generally calculated as the number of redundant observations divided by the sample size. Considering sequence 'validity' and assembly of contigs by splicing with two or more reads, one contig and one singleton was assigned to represent one EST per gene, respectively. The sequence redundancy was calculated according to the following equation: the sequence redundancy (%) = (total 'valid' sequences – singletons - contigs)/total 'valid' sequences 9100. (Li *et al.*, 2010a).

2.3 EST Databases

ESTs constitute an important tool for a better understanding of plant genome structure, gene expression and function. The development of an EST collection also provides an additional resource for the identification of new molecular markers and thus increases the density of gene markers on the genetic map (Lopez *et al.*, 2005). Accumulation of nucleotide sequence information from various organisms, including cassava, has been promoted as an effective method for gene discovery in recent decades (Mochida & Shinozaki, 2010). The development of several full-

length cDNA and expressed sequence tag (EST) collections has led to functional genomics studies in several plant species (Kikuchi *et al.*, 2003; Nanjo *et al.*, 2007; Seki *et al.*, 2002; Soderlund *et al.*, 2009; Taji *et al.*, 2008; Umezawa *et al.*, 2008); moreover, full-length cDNAs have been utilized to develop comprehensive transgenic lines of arabidopsis and rice (Ichikawa *et al.*, 2006; Sakurai *et al.*, 2011). Large-scale cassava cDNA collection projects have been conducted by various cassava research groups (Anderson *et al.*, 2004; Lokko *et al.*, 2007; Lopez *et al.*, 2004; Sakurai *et al.*, 2007), information resources from which have been used in transcriptomics research (An *et al.*, 2012; Sojikul *et al.*, 2010; Utsumi *et al.*, 2012). The cassava draft genome sequence is now publicly available, and the initial assembly spans 419.5 Mb, covering 54% of the estimated cassava genome size (770 Mb). At present, 30,666 protein-coding loci have been predicted from this genome sequence and 3,485 alternative splice forms are supported by ESTs (Prochnik *et al.*, 2012).

The expressed sequence tags (ESTs; which are partial sequences [200–800 bp] of expressed genes randomly picked from a cDNA library) databases are currently the fastest growing and largest portion of these publicly available DNA sequence databases (Cooke *et al.*, 1996; Ohlrogge & Benning, 2000). These databases are important for identifying expressed genes that are further used for developing DNA microarrays (Richmond & Somerville, 2000). The cDNA microarray technology (Schena *et al.*, 1995), depends on the availability of ESTs. This technology has been widely received (Duggan *et al.*, 1999) and used in plants to identify specific gene functions (Aharoni *et al.*, 2000); (Gutierrez *et al.*, 2002), evaluate transcript profiles induced by various physiological or environmental conditions (Reymond *et al.*, 2000); (van Hal *et al.*, 2000); (Lee *et al.*, 2002); (Oztur *et al.*, 2002); (Potokina *et al.*, 2002); (Zhu *et al.*, 2003a), and evaluate transcript profiles between genetically modified and control species (van Hal *et al.*, 2000). Although these are just a few examples that have materialized from genomics initiatives, continued genome sequencing projects for many important crops are still underway and are expected to provide future benefits (Anderson *et al.*, 2004).

Expressed sequence tags (ESTs) today represent a powerful and efficient tool for rapid identification of the genes that are preferentially expressed in certain tissues or cell types (Adams *et al.*, 1991) and are reported to be helpful for post transcriptomic large scale functional genomics particularly to gain new insights into reproductive molecular biology (Cerda, 2009).

2.4 Cassava Genome

The cassava genome (2n=36) (De Carvalho & Guerra, 2002) is highly heterozygous because of its outcrossing nature and broad tropical distribution (Fregene *et al.*, 2003; Siqueira *et al.*, 2010). Conventional breeding and marker-assisted selection have so far proved ineffective in achieving its potential regarding desirable traits, such as high-quality starch, storage root yield, avoidance to postharvest biological deterioration and resistance to diseases (Okogbenin & Fregene, 2003; Rabbi *et al.*, 2012). For instance, cassava storage root yield is approximately 13.6t ha^{-1} globally, which is two- to four fold below its potential productivity. The lack of a reference genome sequence and other genomic and transcriptomic resources has limited progress in basic biological research and breeding in cassava. Therefore, the draft genome sequence of a partial inbred cassava line, AM560, has been generated and publicly released relatively recently (Prochnik *et al.*, 2012) (*http://www.phytozome.net/cassava.php/*). The sequence integrated 26- and 0.9-fold coverage of Roche 454 and Sanger reads, resulting in 530-Mb assembled scaffolds (including 410-Mb of contigs with no gaps), that cover approximately 70% of the cassava genome.

Cassava, *Manihot esculenta* Crantz subsp. esculenta (Euphorbiaceae) is an ancient crop species; starch grains or radiocarbon-dated macroscopic remains are in the archeological record from 1800–7500 BP (Ugent *et al.* 1986; Dickau *et al.*, 2007). Molecular evidence based on the haplotypes of the single-copy nuclear gene glyceraldehyde 3-phosphate dehydrogenase and genetic variation in five microsatellite loci strongly support the view that cultivated cassava is most likely derived from wild populations (*M. esculenta* subsp. flabellifolia), particularly from

the populations occurring along the southern border of the Amazon basin (Olsen & Schaal, 2001; Olsen & Schaal, 1999).

2.5 Genetic Variability and Diversity in Cassava

Although cassava originated in South America and was exported to Africa and Asia, its population structure is poorly understood relative to better studied crops such as maize and rice. An understanding of genetic variation allows the development of robust systems of markers for mapping and breeding, including the characterization of germplasm that might provide useful alleles (Edwards & Batley, 2010). Initial marker development in cassava has relied upon simple sequence repeats (SSRs), such as microsatellite sequences (Raji *et al.*, 2009; Roa *et al.*, 2000), as well as ~2,000 SNPs identified in expressed sequence tags (Ferguson *et al.*, 2012; Tangphatsornruang *et al.*, 2008). Known SSR and SNP markers, however, are sparsely distributed across the cassava genome and may not be ideal for either fine-mapping or inexpensive large scale assays.

2.6 Single Nucleotide Polymorphism

A SNP is a mutation with a single DNA base substitution or minor allele frequency (MAF) observed with a frequency of at least 1% in a given population (Mah & Chia, 2007; Riva & Kohane, 2002). A nonsynonymous SNP (nsSNP) is a single base change in the coding region of a gene, which results in an amino acid substitution (AAS) in the corresponding protein product. nsSNPs results in actual changes in primary amino acid sequences, the function of the protein products might be altered. SNP prediction tools can either be sequence (Miller & Kumar, 2001) or structure based, because most disease-causing SNPs affect the protein stability and structure-based rules that have been established to distinguish functionally significant SNPs from those that are functionally neutral (Sunyaev *et al.*, 2000; Wang & Moult, 2001). There are many publicly available bioinformatics tools that provide systematic means of predicting the functional significance.

Sequence variations could be either SNPs or small insertion/deletions (indels) in genomic DNA of individuals of the same species or closely related

species (Brookes, 1999; Useche *et al.*, 2001). SNPs are one of the most commonly used genetic markers for studying complex genetic traits and genome evolution because of their abundance and slow mutation rate within the genome. In addition, SNPs in coding sequences are used to directly study the genetics of expressed genes and to map functional traits (Grivet *et al.*, 2003). In particular, non-synonymous SNPs (nsSNPs) are more attractive because they change the amino acid, possibly affecting protein function (Garg *et al.*, 1999a).

Sequence and structure based methodologies are the most common approaches used in SNP prediction tools. The advantage of using the sequence-based approach alone for prediction is that results for a large number of substitutions can be generated (Ng & Henikoff, 2003), as structural information tends to be less available. Sequence- based predictions can be more encompassing than structure-based ones as they can include all types of effects at the protein level and can be applied to any human protein with known relatives (Yue *et al.*, 2005). Overall, such an approach has broader applicability as it does not require knowledge of three-dimensional (3D) structures (Yue & Moult, 2006) to predict the impact on functions of resulting proteins. However, sequence-based predictions (based on homology and evolutionary conservation) are unable to shed light on the underlying mechanisms of how SNPs result in changed protein phenotypes (Yue & Moult, 2006), which might have consequences for drug targets. Structure-based approaches are useful as they shed light on how a given amino acid structure can result in an altered protein phenotype by predicting its effect on the 3D structure. The main disadvantage of a structure-based approach is that the 3D structures of most proteins are unknown. Thus, this approach has limited applicability. Tools that integrate both approaches have the added advantage of being able to assess the reliability of the generated prediction results by cross-referencing the results from both approaches. Tools that combine these approaches use different algorithms and methodologies for prediction, thereby having a wider cover- age of the different aspects of SNP analysis.

It has been shown that select amplicons in the non- coding regions, such as introns, 3' untranslated regions (3'-UTRs) and BAC end sequences are a good source of data for SNP discovery and increase the frequency of detection of polymorphisms by up to threefold (Rafalski, 2002; Zhu *et al.*, 2003b)

Single nucleotide polymorphisms have been shown to be the most abundant type of molecular genetic markers in the genome (Cho *et al.*, 1999) and are quickly becoming the marker of choice in agricultural research, especially for use in high-throughput marker-assisted breeding (Rafalski, 2002). Based on different studies on animals and plants, it is necessary to sequence at least 200–500 bp of non-coding DNA on average to find a single non-coding SNP (ncSNP) and about 500– 1,000 bp to locate a coding SNP (cSNP) (Brumfield *et al.*, 2003). In plants, studies on the occurrence and nature of SNPs are beginning to receive considerable attention, particularly in Arabidopsis. In this plant more than 37,000 SNPs have been identified through the comparison of two accessions (Jander *et al.*, 2002). In soybean, the presence of 280 SNPs in 143 amplicons totaling about 76.3 kb of DNA sequence has been reported (Zhu *et al.*, 2003b). The frequency of one ncSNP per 31 bp and 1 cSNP per 124 bp in 18 maize genes assayed in 36 inbred lines (Ching *et al.*, 2002).

The inherent redundancy in EST data makes them a potentially significant resource for the detection of SNPs. This resource has recently been used in a large-scale identification of SNPs in Arabidopsis (Schmid *et al.*, 2003), maize (Batley *et al.*, 2003) and sugarcane (Grivet *et al.*, 2003).

There are three different categories of SNPs:

- TRANSITIONS - (C/T or G/A)
  change of purine to pyrimidine
- TRANSVERSIONS - (C/G, A/T, C/A, or T/G)
  change of purine to purine or pyrimidine to pyrimidine
- INSERTIONS/DELETIONS (indels)

SNPs at any particular site could be bi-allelic, tri-allelic or tetra-allelic. In these types tri-allelic and tetra-allelic SNPs are rare. SNPs are generally bi-allelic. SNPs may occur in the coding, non-coding and intergenic regions of the genome, thus enabling the discovery of genes as a result of the differences in the nucleotide sequences. SNPs are excellent markers for association mapping of polygenic traits with highest map resolution (Brookes, 1999; Bhattramakki *et al.*, 2002).

SNPs are the most frequent type of variation found in DNA (Brookes, 1999; Cho *et al.*, 1999), and their discovery together with insertions/deletions has formed the basis of most allele variations.

2.6.1 Strategies used for development of SNP markers

- Wet lab methods or experimental methods
- Computational methods or bioinformatics methods.

The experimental method of SNP discovery is expensive and time consuming (Schlotterer, 2004; Useche *et al.*, 2001). A computational approach to discover potential SNPs from publicly available sequences makes the development of SNP markers rapid and less expensive.

For computational SNP discovery:

- The program should be able to distinguish allelic variation from sequence variation between paralogous sequences (Batley *et al.*, 2003; Dantec *et al.*, 2004; Marth *et al.*, 1999).
- The program should be able to recognize sequencing errors, which are usually caused by poor quality sequences, especially for EST data (Batley *et al.*, 2003; Garg *et al.*, 1999b; Matukumalli *et al.*, 2006; Picoult-Newberg *et al.*, 1999).

2.6.2 Mining of SNPs from EST sequences marker development in plants.

The steps involved in SNP discovery from EST sequences include clustering, sequence assembly and SNP detection (Batley *et al.*, 2003). A number of methods used to identify SNPs in aligned sequence data rely on sequence trace file analysis to filter out sequence errors by their dubious trace quality (Marth *et al.*, 1999). Two complementary approaches have been adopted to differentiate between sequence errors and true polymorphisms:

- Assessing redundancy of the polymorphism in an alignment
- Assessing co-segregation of SNPs to define a haplotype.

The most important limitation for use of EST for SNP marker development is that EST data provides very limited polymorphisms (Matukumalli *et al.*, 2006). Also, other factors such as alternative splicing, reverse transcription errors and RNA editing interfere with the predictions even after including sequence quality scores. But SNP discovery from EST sequences was successfully implemented for maize (Rafalski, 2002) and pine (Dantec *et al.*, 2004) species by constructing a software data analysis pipeline. Hence, the selection of appropriate tool for SNP identification basically depends on the nature of input sequences.

2.7 SNP in Cassava

Cassava is considered as an allopolyploid, with a high level of heterozygosity and suffers from inbreeding depression. A molecular genetic linkage map of cassava has been constructed based principally on isoenzymes, RAPD and RFLP markers (Fregene *et al.* 1997).

A number of other new resources have been generated over the last few years to improve the efficiency of cassava breeding. They include the detection of QTLs associated with agronomic characteristics (Okogbenin & Fregene, 2003) and resistance to cassava bacterial blight (CBB) and the isolation of resistance gene candidates (RGCs) that can be used in marker-assisted selection as well as in map-based cloning of resistance genes (Lopez *et al.*, 2003).

Cassava ESTs are exploited to detect SNPs in the cultivars used to generate the EST collection. Further information on the frequency of SNPs in cassava was obtained by analysis of 33 amplicons from 3' EST and BAC end sequences in six cassava cultivars. This information helped to develop new strategies for the mapping of these ESTs and establish their association with phenotypic characteristics. ESTs represent a rich source of molecular information. C assava EST sequences (Lopez *et al.*, 2004) for the identification of cSNPs using data from the 1,875 contigs obtained after assembly using the StackPack software (Miller *et al.*, 1999). Among these, 964 contained four or more sequence reads. The sequences were inspected for the presence of polymorphisms using polyBAYES software (Marth *et al.*, 1999). Among the contigs analyzed, 111 contained sequence variants which could be divided into two types: those present within the same cultivar, intra-cultivar SNPs and inter-cultivar SNPs. (81 SNPs and 15 indels in the first category and 76 SNPs and five indels in the second category). Transitions (C/T or G/A and vice versa) were most common in both intra- and inter-cultivars (64 and 65% respectively, than transversions (A/C, A/T, G/C or G/T and vice versa). In total the number of transitions was significantly higher than transversions. A greater number of indels were detected within cultivars (15) than between them. Overall, 144 SNPs were detected, totaling 73,332 bp, thus giving a total of one SNP every 509 bp (Lopez *et al.*, 2005).

2.8 Simple Sequence Repeats

Molecular markers are powerful tools for marker assisted selection (MAS) in plant breeding (Collard & Mackill, 2008). MAS is more efficient, effective, reliable and cost-effective than conventional selection for many traits during plant breeding. They are ubiquitous in prokaryotes and eukaryotes, present even in the smallest bacterial genomes (Morgante & Olivieri, 1993; Toth *et al.*, 2000). DNA polymerase slippage causes errors and generate base pair insertions or deletions, resulting, respectively, in larger or smaller regions (Iyer *et al.*, 2000). The genomic abundance of microsatellites, and their ability to associate with many phenotypes, make this class of molecular markers a powerful tool for diverse application in plant

genetics. The identification of microsatellite markers derived from EST (or cDNAs), and described as functional markers, represent an even more useful possibility for these markers when compared to those based on assessing anonymous regions (Kashi & King, 2006; Varshney *et al.*, 2005; Varshney *et al.*, 2006).

The characterization of tandem repeats and their variation within and between different plant families, could facilitate their use as genetic markers and consequently allow plant-breeding strategies that focus on the transfer of markers from model to orphan species to be applied. EST-SSR also have a higher probability of being in linkage disequilibrium with genes/QTLs controlling economic traits, making them more useful in studies involving marker-trait association, QTL mapping and genetic diversity analysis (Gupta *et al.*, 2003).

Microsatellite markers are widely used to construct genetic maps, associate traits with underlying genomic regions and for MAS (Varshney *et al.*, 2005). Microsatellites are found in all eukaryotic genomes. They consist of 1–6 bp of nucleotide motifs repeated in 5–20 copies distributed throughout the genome both in coding and non-coding regions (Kashi *et al.*, 1997). The use of genomic DNA enriched for satellites to produce libraries for DNA sequencing is a common and reliable technique to develop markers in many plant species, including maize (Sharopova *et al.*, 2002), peanut (He *et al.*, 2003), and red clover (Sato *et al.*, 2005). Their polymorphisms consist of variations in the number of repeats, which was suggested to be due to slippage of the polymerase (Kruglyak *et al.*, 1998). They have a high level of potential polymorphism, locus-specificity, multi-allelic and codominant nature, relative abundance and reproducibility.

Conventional methods used for developing SSRs involve the construction of a genomic library and subsequent screening for the presence of SSR repeat motifs in the clones (Weber & May, 1989). This makes the approach laborious, time consuming and expensive (Schlotterer, 2004). Meanwhile, with the ever increasing number of sequences in publicly available databases, in silico approaches for screening SSRs from sequences have become an efficient and inexpensive

alternative for many species. Several software packages have been developed to detect SSRs in these sequences, especially from ESTs.

SSRs have been reported to be superior to other molecular markers because

- Multiple SSR alleles may be detected at a single locus using a simple PCR based screen
- SSRs are evenly distributed all over the genome
- They are co-dominant
- Very small quantities of DNA are required for screening
- Analysis may be semi- automated (Varshney *et al.*, 2005).

Sequence data for many fully characterized genes and full length cDNA clones have been generated for some plant species (Varshney *et al.*, 2005).

Genic SSRs or EST SSR have certain noticeable advantages over genomic SSRs.

- quickly obtained by electronic sorting
- represents functional region of the genome
- more transferable between related species (Cordeiro *et al.*, 2001; Varshney *et al.*, 2005; Yu *et al.*, 2004)

The presence of SSR in expressed region of genomes suggests that they may have a role in gene expression or function. The waxy gene in rice has been found to contain a poly(CT) microsatellite in the 5'-untranslated region (UTR) whose length polymorphisms is associated with amylase content. In general, approximately 5% of plant EST contain SSRs with a minimum length of 20 nucleotides (Kantety *et al.*, 2002; Poncet *et al.*, 2006; Varshney *et al.*, 2005).

2.9 SSR in Cassava

For cassava, SSRs had been developed and used in genetic linkage map construction (Fregene *et al.* 1997). The first genetic linkage map of cassava was constructed from F_1 intra-specific cross using SSR, RFLPs, RAPDs and isoenzymes (Fregene *et al.* 1997). SSRs (Chavarriaga-Aguirre *et al.* 1998) and EST-SSRs (Raji

et al., 2009; Tangphatsornruang *et al.*, 2008) were developed for germplasm evaluation in cassava and its related species. In addition, 172 SSR markers were developed from genomic DNA-derived satellite enriched library and mapped in an F_1 population. In 2006, a genetic map of an F_2 population was developed using SSR markers (Okogbenin *et al.* 2006). Composite map of an F_1 population that consisted of AFLP, SSR and EST markers (Kunkeaw *et al.*, 2010). However, none of these maps could completely encompass the genome of cassava. A recent genetic map of cassava was constructed using F_1 population. However, the map is mostly based on AFLPs (65%) and SSRs, EST-SSRs and sequence-related amplified polymorphisms (SRAPs) (Sraphet *et al.*, 2011). In the case of cassava, SSR markers have been utilized for the characterization of genetic resources (Fregene *et al.*, 2003; Raji *et al.*, 2009; Roa *et al.*, 2000) and an SSR-based molecular genetic map for cassava comprising 100 markers was described (Okogbenin *et al.*, 2006). For cassava, 14 different primer sequences were designed to amplify SSRs containing mostly perfect or imperfect GA repeats. The primers were tested on 522 accessions of the cultivated cassava core collection conserved at CIAT and showed heterozygosity values between 0.00 and 0.88, with as many as 15 different alleles at one locus (Chavarriaga-Aguirre *et al.*, 1998).

2.10 Bioinformatics tools for SNP prediction

AutoSNP

A program has been developed to detect SNPs and indels from EST sequences (Barker *et al.*, 2003; Batley *et al.*, 2003). The program uses d2cluster (Burke *et al.*, 1999) for clustering the ESTs and CAP3 (Huang & Madan, 1999) to align the sequences. It differentiates between SNPs and sequence errors using redundancy value. Polymorphisms are identified as occurring in multiple reads within an alignment. SNP redundancy score is referred to as the frequency of occurrence of a polymorphism at a particular locus providing a primary measure of confidence in the SNP representing a true polymorphism. Co-segregation score provides a second measure of confidence in SNP validity.

QualitySNP

QualitySNP is an efficient tool for SNP detection, retrieval and storage in diploid and polyploidy species. It can be run on Linux platforms. Haplotype-based strategy is used to detect reliable synonymous SNPs Synonymous SNPs are SNPs in protein-coding exons that don't change the amino acid due to the redundancy of the genetic code and non-synonymous SNPs are SNPs in protein-coding exons that cause a change in the amino acid. SNPs are detected from public EST data. Haplotypes represent the different alleles of a gene in a dataset. It uses three filters for the identification of reliable SNPs.

- Filter 1: screens for all potential SNPs and identifies variation between or within genotypes.
- Filter 2: is the core filter that uses a haplotype-based strategy to detect reliable SNPs. Clusters with potential para- logs as well as false SNPs caused by sequencing errors is identified.
- Filter 3: screens SNPs by calculating a confidence score, based upon sequence redundancy and quality.

Non-synonymous SNPs are identified by detecting ORFs of contigs with SNPs.

HaploSNPer

HaploSNPer is a web-based SNP discovery and allele detection tool based on QualitySNP (Tang *et al.*, 2008). It makes use of BLAST for finding homologous sequences. ESTs are used as input files. CAP3 or PHRAP are used for aligning, and QualitySNP for predicting possible allelic sequences and SNPs. HaploSNPer provides a user friendly interface for visualization of SNP and alleles. Singhal and his team in 2011 used HaploSNPer and found 40589 reliable SNPs in Sorghum bicolor genome. Limitation of HaploSNPer is that it does not extend to the analysis of diversity, linkage disequilibrium or haplotype network study.

SniPlay

SniPlay integrates pipeline which is freely accessible through the internet, combining existing software's with new tools to detect SNPs and to compute different types of statistical indices and graphical layouts for SNP data. It is able to detect SNPs and indels from standard sequence alignments, genotyping data or Sanger sequencing.

The pipeline allows the use of external data such as phenotype, geographic origin, taxa, and stratification to define groups and compare statistical indices. It also integrates database for storing polymorphisms, genotyping data and grapevine sequences released by public and private projects which allows the user to retrieve SNPs using various filters such as genomic position, missing data, polymorphism type, and allele frequency. It can be used to compare SNP patterns between populations (Dereeper *et al.*, 2011).

SNPServer

SNPServer (Savage *et al.*, 2005) is an online tool. It is the real time implementation of the SNP prediction method AutoSNP. It uses AutoSNP software by providing a web interface for sequence input, comparison and assembly and permits rapid discovery of SNPs. It uses BLAST to identify related sequences, and CAP3 to cluster and align these sequences. The alignments are parsed to the SNP discovery software AutoSNP.

InSNP

InSNP is windows based software package for SNP prediction.

SNPdetector

SNPdetector scripts work only on Unix/Linux platforms. It use the Smith-Waterman algorithm for aligning reads, as well as a modified version of the NQS (Altshuler *et al.*, 2000) method for detecting homozygous SNPs among different individuals. SNP detector requires a minimum of a 30 per cent threshold for secondary peak intensity for detecting heterozygous SNPs.

NovoSNP

NovoSNP works on windows as well as Unix/Linux based platforms. It uses BLAST (Altschul *et al.*, 1990) for aligning sequence reads and uses a series of filters to reduce false positive SNPs. This package is configured to work with a database, and, hence, it makes polymorphism discovery and data storage convenient.

2.11 Bioinformatics tools for SSR prediction

Sputnik

Sputnik is a program developed in C programme. It searches DNA sequence file in FASTA format for simple sequence repeats. It uses a recursive algorithm to search for repeated patterns of nucleotides of length between 2 and 5 and finds perfect, compound and imperfect repeats. The output of Sputnik is a file of SSRs in tabular format. Sputnik has already been applied for SSR identification in many species including Arabidopsis and barley (Cardle *et al.*, 2000).

Tandem Repeats finder

Tandem Repeats Finder (TRF) (Benson, 1999) can identify very large sized SSR repeats, up to a length of 2000 bp it uses a set of statistical tests for reporting SSRs, which is based on four distributions:

- pattern length
- matching probability
- indel probability
- tuple size

TRF finds perfect, imperfect and compound SSRs. TRF is available for Linux. TRF has been used for SSR identification in cowpea (Chen *et al.*, 2007).

Sequence Repeat Identification Tool

Simple Sequence Repeat Identification Tool (SSRIT), uses Perl script to execute the programme and find perfect SSR repeats (2 to 10 bp in length) within a sequence (Temnykh *et al.*, 2001). SSRIT was used to mine SSR in ESTs from Barley (Kantety *et al.*, 2002), maize, rice, sorghum and wheat. SSRIT was used to mine SSRs in wheat rust *Puccinia sp* (Singh *et al.*, 2011).

Tandem Repeat Occurrence Locator

TROLL, draws a keyword tree and matches it with a technique adapted from bibliographic searches, based on the Aho- Corasick algorithm (Castelo *et al.*, 2002). One of the major disadvantages of TROLL is that it cannot handle very large sequences and cannot process large batches of sequences as the tree takes up large amounts of memory.

MISA

The microsatellite (MISA) tool identifies perfect, compound and interrupted SSRs. It is a perl programme. It requires a set of sequences in FASTA format and a parameter file "misa configuration settings" that defines unit size and minimum repeat number of each SSR. The output includes a file containing the tables of repeats found, and a summary file.

MISA can also design PCR amplification primers either side of SSR. The tool is written in Perl and is therefore platform independent, but it requires as installation of Primer3 for primer search (Thiel *et al.*, 2003). MISA has been applied for SSR identification in coffee (Aggarwal *et al.*, 2007), barley (Kota *et al.*, 2001; Thiel *et al.*, 2003), wheat (Yu *et al.*, 2004), rye (Khlestkina *et al.*, 2004) and peanut (Liang *et al.*, 2009).

Repeat Finder

Repeat Finder (Volfovsky *et al.*, 2001) is a tool that works only in linux platforms.

It finds SSRs in four steps.

1. Find all exact repeats using Repeat Match or REPuter.
2. Merges repeats together into repeat classes
3. Merging all of the other repeats that match those already merged, into the same classes.
4. Matches all repeats and classes against each other in a non-exact manner using BLAST.

The input is a genome or set of sequences, and the output is a file containing the repeat classes and number of merged repeats found in each class. Repeat Finder can finds repeats of any length. Also it finds perfect, imperfect and compound repeats. It has been used to identify SSRs in peanut (Jayashree *et al.*, 2006).

SSRPoly

The only SSR identification tool which is capable of identifying polymorphic SSRs from DNA sequence data. The input is a file of FASTA format sequences. SSRPoly includes a set of Perl scripts and MySQL tables that can be implemented on UNIX, Linux and Windows platforms (Tang *et al.*, 2008).

MATERIALS AND METHODS

3. MATERIALS AND METHODS

The study entitled "Molecular marker development of cassava mosaic disease resistance using bioinformatics tools and its validation." was conducted at the Central Tuber Crop Research Institute (CTCRI) during 2014-2015. Details regarding the experimental materials used and methodology adopted for various experiments are presented in this chapter.

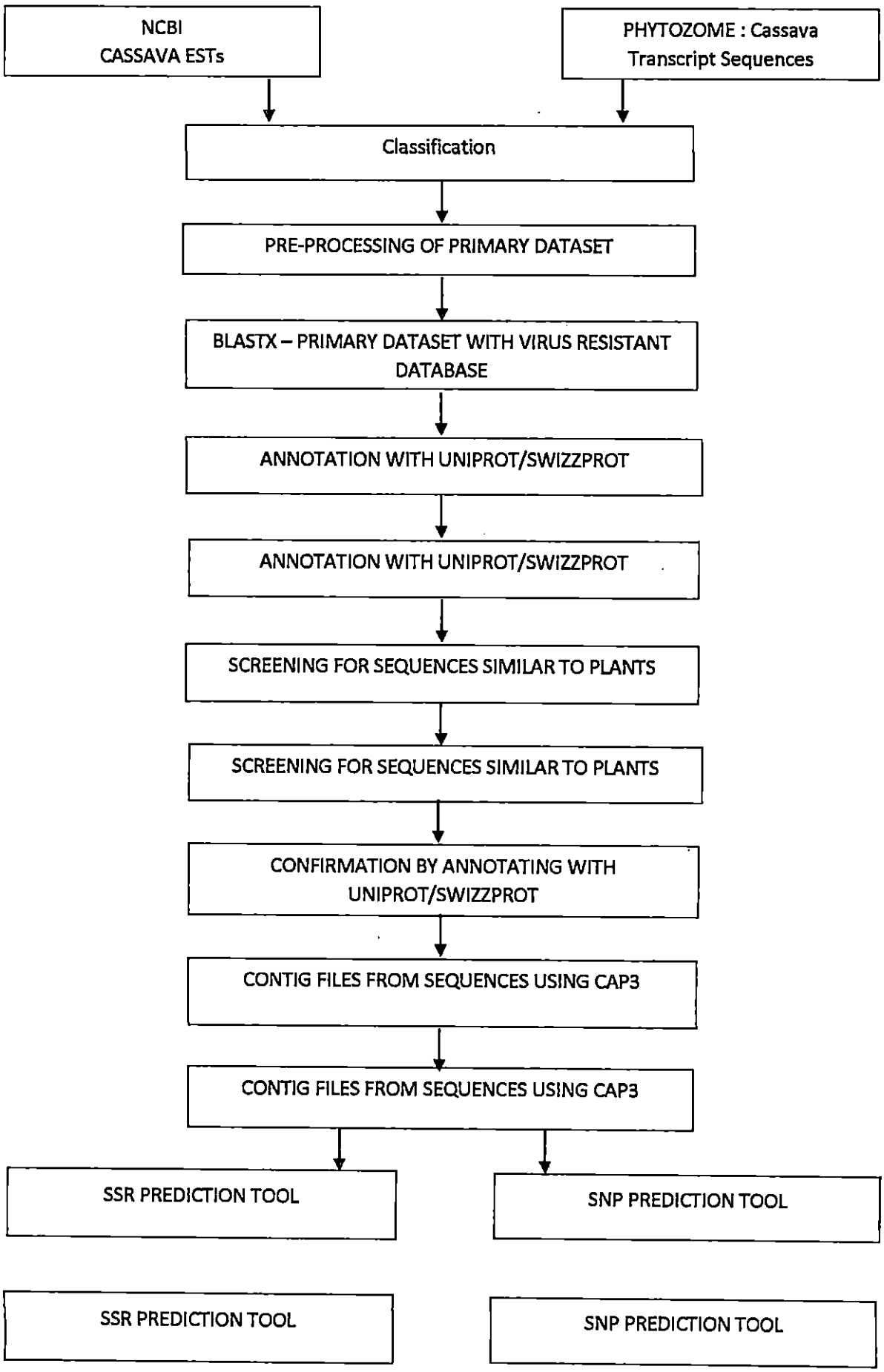
3.1 Cassava Sequence Dataset

The preliminary data set for the work was obtained from the EST section of NCBI (*http://www.ncbi.nlm.nih.gov/nucest*). The EST database is a collection of short single-read transcript sequences from Genbank. These sequences provide a resource to evaluate gene expression, find potential variation, and annotate genes.

To discover polymorphisms, the cassava draft genome sequence and transcript sequences (variety AM560-2, JGI annotation v4.1) from the Phytozome website (*http://phytozome.jgi.doe.gov/pz/portal.html*) were also downloaded. Phytozome is the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute. Families of related genes representing the modern descendants of ancestral genes are constructed at key phylogenetic nodes.

Cassava sequences were retrieved from the Genbank EST section on 11th November 2014. A total of 86310 ESTs of cassava were downloaded from NCBI. The transcript sequence of cassava obtained from phytozome had a total of 34151 transcript sequences.

Together, a total of 1,20,461 sequences were taken as the primary dataset for research work. Work flow is given in figure 1.


NCBI
CASSAVA ESTs
PHYTOZOME : Cassava
Transcript Sequences
Classification
PRE-PROCESSING OF PRIMARY DATASET
BLASTX – PRIMARY DATASET WITH VIRUS RESISTANT DATABASE
ANNOTATION WITH UNIPROT/SWIZZPROT
ANNOTATION WITH UNIPROT/SWIZZPROT
SCREENING FOR SEQUENCES SIMILAR TO PLANTS
SCREENING FOR SEQUENCES SIMILAR TO PLANTS
CONFIRMATION BY ANNOTATING WITH UNIPROT/SWIZZPROT
CONTIG FILES FROM SEQUENCES USING CAP3
CONTIG FILES FROM SEQUENCES USING CAP3
SSR PREDICTION TOOL
SNP PREDICTION TOOL
SSR PREDICTION TOOL
SNP PREDICTION TOOL

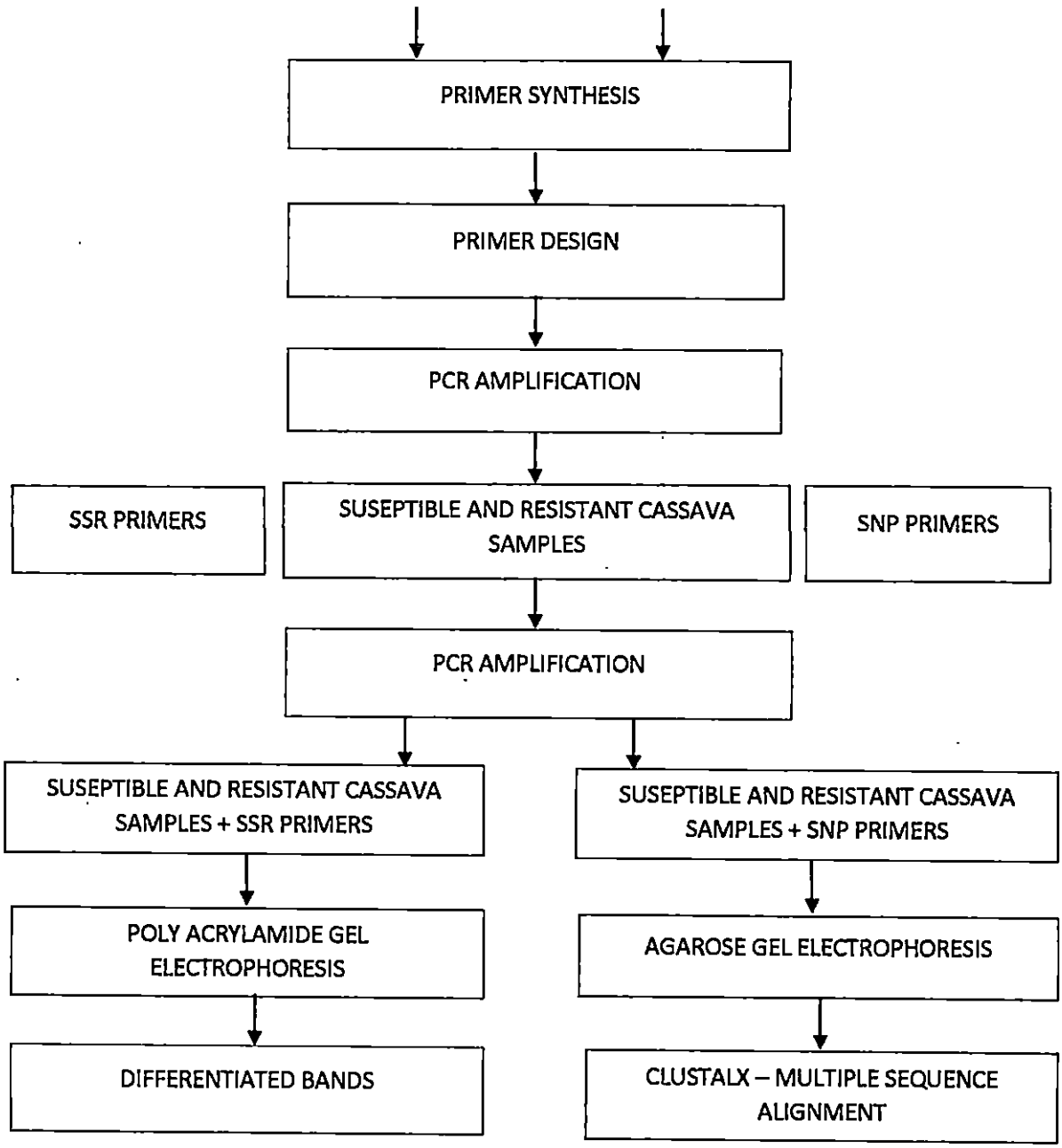


Figure 1. Workflow for the identification of SNP and SSR markers

3.2 Pre-processing of Sequences

The sequences are processed for contamination or simple repeats using the SeqClean script (*http://sourceforge.net/projects/seqclean/files/*).

SeqClean is a tool for validation and trimming of DNA sequences from a flat file database (FASTA format). SeqClean was designed primarily for "cleaning" of EST databases, when specific vector and splice site data are not available, or when screening for various contaminating sequences is desired.

The program works by processing the input sequence file and filtering its content according to a few criteria:

- Percentage of undetermined bases
- PolyA tail removal
- Overall low complexity analysis
- Short terminal matches with various sequences used during the sequencing process (vectors, adapters)
- Strong matches with other contaminants or unwanted sequence (mitochondrial, ribosomal, bacterial, other species than the target organism etc.)

Requirements to run SeqClean

- Perl version >= 5.6
- A working installation of recent versions of NCBI's blastall and megablast programs (one or more databases of potential contaminants (e.g. a vector database like NCBI's UniVec) properly formatted to work with NCBI's blastall (using formatdb))

Usage: seqclean your_est_file

Seqclean creates two output files of interest:

- The filtered FASTA file (your_est_file.clean for the example above) containing only valid (non-trashed) and trimmed ("clear range") sequences
- A "cleaning report" (your_est_file.cln) providing details about sequence trimming and trashing (coordinates, reasons for trashing, contaminant names etc. - see below for a detailed description).

The sequences are checked for sequence contamination and simple repeats by using the SeqClean script with the default runtime options. Vector sequences in these ESTs are then trimmed using the UniVec_Core database (*http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/)* of NCBI.

3.3 Resistant Virus Gene Database

In order to develop the markers related to CMD, a plant specific database of virus resistant genes is needed. Hence resistance virus gene database was created and compiled from uniprotKB manually. The UniProt Knowledgebase (UniProtKB) is the central access point for extensively curated protein information, including function, classification and cross-references.

The database acts as the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

The UniProt Knowledge base is a non-redundant and complete protein sequence database consisting of two components:

1. Swiss-Prot - section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis
2. TrEMBL- section with computationally analyzed records that await full manual annotation

R-gene or resistant genes related to cassava and mosaic diseases was screened from it and was used for database creation. The virus resistance protein database consisted of 730 resistant genes.

3.4 Processing the Primary Dataset

3.4.1 Screening for Resistant Genes in Primary Dataset

For screening of primary dataset with virus resistant protein sequences 'BlastX – Search protein database using a translated nucleotide query' was used. The software Klast was used for doing sequence comparison. Klast provides a graphical interface to blast. All the databases are downloaded through database manager of klast. Public and personal databases can be installed through the manager. The sequences to be searched is loaded in the query tab of klast. it provides options for blasting using online databases. It also supports annotation and retrieval of biological classification data and other related data directly from selected databases.

In order to introduce biological classification data within KLAST results:

- The biological classifications managed by KDMS; they are listed by the end of the Public databanks panel: Enzyme, Gene Ontology terms, InterPro terms, NCBI Taxonomy and Pfam terms
- A reference sequence databank that is annotated with such classifications; a very well-known example is Uniprot_Swissprot.

3.4.2 Annotation of Screened Sequences

Cassava ESTs and transcript sequences were blasted against resistant genes and the transcript sequences having high similarity were annotated with Uniprot/Swissprot database.

3.4.3 Screening for Plants from Annotated Data

Sequences after annotation may contain organisms other than plants as annotation was done using Uniprot/Swissprot database. These sequences were further screened for plant related genes only and there were 14336 related sequences.

3.4.4 Confirmation of Annoted Sequences

Sequences were again annotated with Swissprot/Uniprot database for confirmation. A 100% match was obtained in this step. This resulting sequence was taken as the input for the prediction of SNP and SSR.

3.5 DNA Polymorphism Discovery

3.5.1 Assembling of Sequences

Clustering of sequences was done using a perl script called CAP3 which is available at (http://seq.cs.iastate.edu/CAP3.html).

CAP3 takes as input a file of sequence reads in FASTA format. CAP3 takes two optional files: a file of quality values in FASTA format and a file of forward-reverse constraints. The file of quality values must be named "xyz.qual", and the file of forward-reverse constraints must be named "xyz.con", where "xyz" is the name of the sequence file. CAP3 uses the same format of a quality file as Phrap.

Usage: CAP3 File_of_reads [options]

File_of_reads is a file of DNA reads in FASTA format

Options (default values):

-a N specify band expansion size N > 10 (20)
-b N specify base quality cutoff for differences N > 15 (20)
-c N specify base quality cutoff for clipping N > 5 (12)
-d N specify max qscore sum at differences N > 100 (200)
-e N specify extra number of differences N > 10 (20)
-f N specify max gap length in any overlap N > 10 (300)
-g N specify gap penalty factor N > 0 (6)
-h N specify max overhang percent length N > 5 (20)
-i N specify segment pair score cutoff N > 20 (40)
-j N specify chain score cutoff N > 30 (80)
-k N specify end clipping flag N >= 0 (1)
-m N specify match score factor N > 0 (2)
-n N specify mismatch score factor N < 0 (-5)
-o N specify overlap length cutoff > 15 (40)
-p N specify overlap percent identity cutoff N > 65 (90)
-q N specify flag for reads of length >= 30 kb N >= 0 (0)
-r N specify reverse orientation value N >= 0 (1)
-s N specify overlap similarity score cutoff N > 250 (900)
-t N specify max number of word occurrences N > 30 (500)
-u N specify min number of constraints for correction N > 0 (4)
-v N specify min number of constraints for linking N > 0 (2)
-w N specify file name for clipping information (none)
-x N specify prefix string for output file names (cap)
-y N specify clipping range N > 5 (100)
-z N specify min no. of good reads at clip pos N > 0 (2)

If no quality file is given, then a default quality value of 10 was used for each base. To get assembly results in CAP format, first go to the standard output

and then direct it to a file. CAP3 also produces assembly results in ace file format (".ace"). This allows CAP3 output to be viewed in Consed. CAP3 saves consensus sequences in file ".contigs" and their quality values in file ".contigs.qual". Reads that are not used in assembly are put in file ".singlets". Additional information about assembly is given in file ".info". The CAP3 program reports whether each constraint is satisfied or not. The report is in file ".results".

The sequences obtained by the above process and the predicted transcript sequences from the cassava draft genome sequence were assembled using the CAP3 program with default runtime options.

3.5.2 SNP Prediction

Two tools were used for prediction of SNP

- QualitySNP
- AutoSNP

QualitySNP

QualitySNP is an efficient tool for SNP detection, storage and retrieval. It implements a new algorithm developed by us to reliably detect single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) in expressed sequence tag (EST) data, both with and without quality files. The new algorithm uses a haplotype based strategy on potential SNPs, which predicts reliable SNPs, as well as reliable haplotypes.

The pipeline consists of four steps:

1. EST assembling using cross_match for removing vectors and CAP3 for sequence clustering.
2. Analysis of the alignment information to select clusters with at least 4 EST members; this is done by the Perl script "Getalignmentinfo".
3. Performs SNP and haplotype detection, and distinguishes variations between or within genotypes. This is the core part of the pipeline, using the C program named "QualitySNP" that implements the algorithms for

prediction haplotypes and SNPs. The helper programs "Getavailcontigseq" and "Getavailcontigqual" extract the sequences from the contigs and get the quality information of contigs.

4. The non-synonymous SNP discovery was done using FASTY, from Pearson's FASTA package. A C program named "GetnonsySNPfasty" is used to analyze FASTY results, detect the ORFs and find non-synonymous SNPs.

Commands to run QualitySNP:

1. % CAP3 filename –p similarity –o 100 with the parameters in these commands are: filename is the file with sequences in FASTA format and similarity is the similarity of overlap for CAP3.
2. % Getalignmentinfo filename.cap min-clustersize

   where

   - filename is the sequence file
   - min-clustersize is the minimum cluster size.
   - The default minimal cluster size is 4.

QualitySNP is executed

1. % Getavailcontigseq filename.cap
2. % Getavailcontigqual filename.cap
3. % QualitySNP filename.cap min-allelesize lowqual5side similarity1 similarity2 lowqual3side weightlowqual min-confidencescore

   The parameters used in these commands are:

   - Min-allelesize is the minimum size of alleles of each SNP
   - lowqual5side is the length of the low quality region at the 5' end of sequence
   - similarity1 is the similarity on one polymorphic site (0.75)
   - similarity2 is the similarity on all polymorphic sites (0.8)
   - lowqual3side is the low quality region of 3' side.

- weightlowqual is the weight value of the low quality region (0.5)
- min-confidencscore is the minimal confidence score (2)

Analysis of non-synonymous SNP

% fasty34_t allavailcontigseqwithSNP Uniprot –b 6 –d 6 –Q > allavailcontigseq withSNP.fasty

Parameters used:

- Viridiplantae – Plant protein database in uniprot (tp://141.161.180.197/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_plants.dat.gz).
- This can be either the full path leading to a FASTA-formatted protein database, or a single letter to indicate the database, in case the FASTLIBS environment variable is used to specify databases in the FASTA suite.

The files "availcontigseq" and "allavailcontigseqwithSNP" are from the results of QualitySNP, File "availcontigseq" contains the consensus sequences of contigs with SNPs, as produced by CAP3. As these sequences are not curated, they may contain padding symbols ("*"), which may indicate either insertions and/or deletions in the ESTs, but in many cases these may be caused by sequencing errors. The file "allavailcontigseqwithSNP" contains the consensus sequences of SNP-containing contigs which did not contain any insertions or deletions.

2088 contigs from 14336 sequences with resistance against virus was taken as the input of QualitySNP.

- The result obtained was classified into:
  - Ssnpcodingdata - list of synonymous SNPs
  - Nssnpcodingdata - list of Non-Synonymous SNPs
  - Ssnpfastydata - list showing the transcribed sequence of the SNPs
  - Nssnpfastydata - list showing the transcribed sequence of the SNPs

- Indelsnpdata – list of Indels
- Contigorfdata – open reading frames of contigs
- Utrsnpdata – list of SNPs in untranslated region
- Snptagdata – list of SNP tags

AutoSNP

AutoSNP is one of the first tools for SNP discovery aimed at exploiting the large number of ESTs available in the public domain. The input consists of large sets of ESTs of often unknown gene origin and without trace files or base quality data. The ESTs are first clustered with d2cluster and additionally aligned and assembled with CAP3.

- Two parameters are used for putative SNP identification:
    - SNP redundancy score is the minimum number of reads per allele (two by two).
    - SNP cosegregation score is the percentage of other SNPs with an identical segregation pattern.

The AutoSNP computer program carries out automated analysis of EST sequence data and identifies SNPs as well as insertion/deletion (InDel) variations present in them. It aligns the EST sequences and distinguishes between predicted SNPs and sequencing errors on the basis of the redundancy criterion.

For each candidate SNP, redundancy-score and co-segregation score are estimated. The redundancy score of a predicted SNP locus is the frequency of polymorphism at this locus. Co-segregation score is the likelihood that the predicted SNP will be transmitted together with other SNPs present in the vicinity in the EST sequence.

The AutoSNP output includes the predicted SNPs and InDels along with their redundancy and co-segregation scores.

On comparative evaluation of QualitySNP and AutoSNP, QualitySNP shows more promising SNPs unlike AutoSNP where a huge number of SNPs are predicted which cannot be used practically.

3.5.3 SSR prediction

Two tools were used for the prediction of SSR

- MISA - MIcroSAtellite identification tool
- SSRIT - Simple Sequence Repeat Identification Tool

MISA

MISA allows the identification and localization of perfect microsatellites as well as compound microsatellites which are interrupted by a certain number of bases. In conjunction with a set of additional software programs (Primer 3, stackPACK, BlastX), the Microsatellite search module (MISA) identifies SSR-containing ESTs from an input database together with primer sequences for a non-redundant set of SSRs and data about putative functions.

The categorized results of the microsatellite searches are stored in two files:

- Localization and type of identified microsatellite(s) in a table wise manner
- Frequency of a specific microsatellite type according to the unit size or individual motifs.

SSRIT

SSRIT finds all *perfect* simple sequence repeats (SSRs) in a given sequence.

Although the output does contain sequence ID, motif (repeat) type, no. of repeats, SSR start and end, it does have the following limitations against criteria:

- The program currently is not capable of detecting mononucleotide repeats;

- The output is not perfected currently due to which it requires some additional work by the user which is especially cumbersome when dealing with medium-sized (hundreds of sequences) datasets.

On comparative evaluation of MISA and SSRIT, the number of SSRs were more in MISA and also the number of classes of SSRs were also high in MISA. The ability of MISA to predict complex SSRs unlike SSRIT was also considered.

3.6 Primer Designing for Predicted SNPs and SSR using QualitySNP and MISA

Primer pairs are designed to amplify the genomic region around each discovered SNP or SSR site. Sequences are selected for primer designing based on the hit percentage of contigs containing SNP and SSR with the resistant genes. SNP and SSR containing contigs with hit percentage between 80% – 100% was selected. Primer pairs are designed using Primer3plus tool.

3.5.4 Primer3plus

Primer3 picks primers for PCR reactions, considering as criteria: oligonucleotide melting temperature, size, GC content, primer-dimer possibilities, PCR product size, positional constraints within the source (template) sequence, possibilities for ectopic priming (amplifying the wrong sequence) and many other constraints.

For selection of sequences for primer sequencing, Primers are designed and selected for synthesis based on mainly 2 categories: GC content should be above 50% and Melting temperature should be between 55°C - 60°C.

3.7 Validation of SNP and SSR

3.7.2 Genomic DNA extraction

A total of 10 cassava varieties include 5 CMD resistant and 5 susceptible cassava were selected based on field trials conducted at Central Tuber Crop Research Institute (CTCRI), Thiruvananthapuram. Fresh young leaves were collected from 10 cassava varieties and DNA was isolated from these leaves

samples using the method described by Dellaporta *et al.*, (1983) with some modifications (Appendix I). About 0.1g of leaves was weighed and grinded to fine powder in liquid nitrogen using sterile pestle and mortar. 2% of PVP was added to the samples at the time of grinding to avoid mucus content. The powdered leaf sample was then transferred into sterile 2ml eppendorf tubes. To these samples, 15 ml of extraction buffer was added and mixed well which were then incubated at 4°C for 30 min. After incubation, 1ml of SDS (20%) was added to the samples and mixed well by inverting the tubes. It was again incubated on water bath (Memmert) at 65°C for 30 min. To the samples, 5ml of 5M potassium acetate was added and incubated on ice for 20 min at 4°C. The samples were then centrifuged (Sigma laborzentrifge) at 12000 rpm for 20 min at 4°C and the supernatant was collected. 2/3 volume of chilled isopropanol was added to the above collected supernatant, slowly inverting the tubes to precipitate the DNA and it was kept at a temperature of -20°C for 30 min. The precipitated DNA was centrifuged at 12000 rpm for 15 min at 4°C. The supernatant was discarded and pellets were resuspended in 500µl TE buffer (Appendix II). Then 5µl of RNase (10mg/ml) was added to suspended pellet and incubated at 37°C for 1 hr on water bath. To remove RNase, equal volume of chloroform: isoamyl alcohol (24:1) was added into the samples and centrifuged at 12000 rpm for 15 min at 4°C. The supernatants were collected and 10µl sodium acetate was added along with 200µl ice cold absolute ethanol to precipitate DNA. These were then mixed properly by gentle inversion, and incubated for 2hr at -20°C. After incubation, the samples were centrifuged at 1000 rpm for 15 min at 4°C and DNA pellets were collected. To the DNA pellet 500µl of 70% Ethanol was added and again centrifuged. Later 100µl TE buffer was added to the properly dried DNA pellet and stored either 4°C or -20°C.

3.7.3 Determination of quality of DNA

Agarose gel electrophoresis

Weighed 0.8g of agar powder (Sigma) was transferred it to a conical flask. 100ml of 1X TBE buffer (Appendix III) was added and heated for 2 min on a microwave oven to dissolve the agarose. 1µl ethidium bromide (.5µg/ml) was added

to the pre-cooled agarose solution, mixed well and the solution was poured in to gel casting tray fitted with comb. After the agarose was solidified, the gel tray was transferred into gel tank filled with 1X TBE running buffer and the comb was then carefully removed. One microlitre of DNA was properly mixed in 2 µl gel loading dye and loaded to the wells. 2µl of 100bp DNA ladder was added into first well or last well of the agarose gel as a base pair size indicator. The gel was then run at 80 volts for 30 min and was documented in gel documentation system to visualize the bands (Alpha Imager, USA).

Quantification of DNA

The concentration and purity of all 10 DNA samples was determined using UV spectrophotometer by taking absorbance at 260/280 nm. Firstly, the spectrophotometer was calibrated using TE buffer as blank at 260nm as well as 280nm (Systronics). After calibration, all the samples were individually read at 260nm and 280nm and OD values were recorded. The purity of the samples was checked by taking the OD value at 260nm and it should be in the range 1.8 to 2. The amount of DNA can be quantified using the formula:

$$\text{DNA concentration } (\mu g/ml) = \frac{\text{OD260 x 100 (dilution factor) x 50 } \mu g/ml}{1000}$$

According to the reading obtained after quantification, genomic DNA was diluted to a concentration of 50ng/µl and stored at 4°C. The stock DNA was then stored in -20°C.

PCR Amplification for SNP Primers

A total of 20µl reaction contained 5ng genomic DNA, 0.2µM of each forward and reverse primer, 100µM of dNTP, 1X buffers (10 mM Tris-Hcl (pH 8.3), 50 mM KCl, 1.5 mM MgCl), 3U Taq DNA polymerase and autoclaved ultra-pure water. Amplifications were done in a BioRad C1000™ thermal Cycler programmed for initial denaturation of 2 min. at 94°C then 30 cycles of 1 minute at 94°C, 1 minute at 55°C, 2 minute at 72°C, and a final extension of 30 minute at

72°C. The amplification of PCR products were analysed in 2% agarose gel electrophoresis.

PCR Amplification for SSR primers

A total of 5 SSR primer pairs were used for molecular characterization on the selected 10 cassava samples. The PCR amplifications were performed in a volume of 20µl reaction containing 5ng genomic DNA, 0.2µM of each forward and reverse primer, 100µM of dntp, 1X buffer (10 mM Tris-Hcl (pH 8.3), 50 mM KCl, 1.5 mM MgCl), 3U Taq DNA polymerase. The cocktail for PCR was prepared according to Appendix V. Amplifications were proceeded in a BioRad C1000^{TM} thermal Cycler programmed for initial denaturation of 5 minute at 95°C then 30 cycles of 1 minute at 95°C, 1.30 minute at 58°C, 2 minute at 72°C, and a final extension of 5 minute at 72°C. After amplification, a volume of 8µl of loading dye was added to each of the amplified product, and the products were run on 2% agarose gel, stained with ethidium bromide and visualized in a gel documentation system. The sizes of the amplified products were determined using an appropriate molecular ladder.

Polyacrylamide gel electrophoresis (PAGE)

Larger plate cleaning

The glass plate was cleaned using deionized water to remove all the contaminants, then wiped with kimwipe soaked in absolute ethanol. The plate was then air dried laboline was applied in an evenly manner using kimwipes.

Small plate cleaning

The glass plate was cleaned thoroughly with water and laboline. The glass plate was then rinsed with deionized water to remove detergent residues and wiped with kimwipes which was soaked in absolute ethanol. Glass plate was air dried and

bind silane was gently and evenly applied (Appendix VIII) on the inner surface of small plate.

Gel preparation and casting

The larger glass plate was laid flat on the bench and spacers were placed on its side's, then smaller plate was placed on it (coated side should be towards spacer). Edges were aligned. The unit was then assembled with side clamps and bottom caster assembly and was locked. Six percentage of polyacrylamide solution containing 7M urea was prepared for gel casting (Appendix VII). 60µl TEMED and 600µl APS was added at the time of gel casting. The above mixed solution was drawn into a 120ml syringe and the syringe is inverted to expel any trapped air that has entered the syringe. The nozzle of the syringe was introduced into the notched region on the caster base where both the glass plates were aligned. The mixed solution from the syringe is expelled, filling the space almost to the top. After the solution was filled up, the comb is inserted in gel to the edge of the plate. The apparatus is kept in an appropriate position and the unit was left undisturbed for ½ an hour for polymerization.

Gel running

The apparatus with casted gel was assembled, then the apparatus is dislodged from the precision caster base and fitted vertically into the universal base using a stabilizer bar. The temperature indicator was adhered to the surface of the outer plate to monitor the temperature during the run. The upper and the lower buffer chambers were filled with the required volume of 1X TBE buffer. The gel was pre-runned for 20 minutes at 100W. Following the completion of the pre-run, the power supply was stopped. The wells were thoroughly rinsed using a pipette to remove any deposited urea. The denatured PCR samples were prepared by denaturation of all 10 PCR amplified DNA samples along with gel loading dye (Appendix IV) at 95°C for 5 min in a thermal cycler. 3-4 µl of each denatured samples were loaded along with 100bp ladder and each of the empty wells are loaded with empty well

dye (Appendix IV). The samples were electrophoresed at 100 W for 35-40 minutes (specific according to each of the primers PCR product size). The power supply was turned off after the completion of the run and the upper buffer chamber was partially emptied by attaching the connector to the drain port on the gel unit. The unit was then disassembled and the larger plate was removed and small plate containing the gel was used for silver staining.

Silver staining

The glass plate containing gel was transferred into a large tray containing the fixer and placed on a shaker (RiVOTEK) for 20 min, ensuring that the gel surface faced upwards. Similarly, staining was performed using silver stain (Appendix X) after the gel was washed in another trough containing deionized water for 5 min, subsequent to a further wash with deionized water for few seconds, then the stained gel was developed by transferring into a trough containing the developer (Appendix XI) and gently rocking the trough in a to and fro motion. A white surface was placed under the gel to enable visualization during development. After the bands had visibly developed, the plate was immediately transferred into the fixer (Appendix IX) for few min to properly fix the bands. Following a final wash step, ensure that the wavy nature on the glass plate due to the fixer residues should wash out. Gel was allowed to dry on open air overnight. Clear and reproducible bands were only selected for scoring.

3.9 Validation of SNP Markers

Finally, validation of SNP markers was done by running the marker with all five resistant and five susceptible cassava varieties in agarose gel electrophoresis and then eluting the bands, sequencing it and comparing it with the reference genome of cassava.

Reference genome of cassava is available in the Phytozome database. ClustalX was used for aligning the sequences and to validate the SNP.

3.9.1 ClustalX

ClustalX is a windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analysing the results. Alignment quality analysis can be performed and low-scoring segments or exceptional residues can be highlighted. ClustalX has the ability to cut-and-paste sequences to change the order of the alignment, select a subset of the sequences to be realigned, and select a sub-range of the alignment to be realigned and inserted back into the original alignment. Alignment quality analysis can be performed and low-scoring segments or exceptional residues can be highlighted. Quality analysis and realignment of selected residue ranges provides a powerful tool to improve and refine difficult alignments and to trap errors in input sequences.

To do a multiple alignment on a set of sequences, multiple alignment mode is selected. A single sequence data area is then displayed. The alignment menu then allows to either produce a guide tree for the alignment, or do a multiple alignment following the guide tree, or to do a full multiple alignment.

Multiple sequence alignment is performed in clustalx to perform snp validation from the sequenced snps which have resistance against cmd.

3.10 Validation of SSR Markers

To confirm that the designed SSR markers are working, AGE was done. Clearly visible thick bands were obtained in the gel. These confirmed that the SSRs are working. Validation of SSR markers was done by running the marker with all five resistant and five susceptible cassava varieties in poly acrylamide gel electrophoresis and then examining the bands for any distinct variability in position which would confirm that the marker is working. PAGE will clearly differentiate the bands produced by the SSR primers which is not possible to view in AGE

RESULTS

4. RESULTS

4.1 Classification of Cassava Sequence Dataset

The preliminary data set for the work was obtained from the EST section of NCBI (*http://www.ncbi.nlm.nih.gov/nucest*) and the cassava draft genome sequence and transcript sequences from the Phytozome website (*http://phytozome.jgi.doe.gov /pz/portal.html*). The sequences were classified based on cultivars into 19 cultivars and one category with unclassified sequences. All the phytozome transcript sequences were from a cultivar of cassava named AM560-2. Most sequences in NCBI were from MTai-16 (35400) sequences and the number of sequences was least from H226 (21) (Figure 2).

4.2 Pre-processing of Primary Dataset

The primary dataset was processed for contamination or simple repeats using the SeqClean script. The sequences are checked for sequence contamination and simple repeats by using the SeqClean script with the default runtime options. UniVec_Core database of NCBI is used to clean the ESTs.

A total of 63 sequences were removed after cleaning the primary dataset with SeqClean (Table. 1). The primary dataset obtained from NCBI and Phytozome which consisted of 1,20,461 sequences was reduced to 1,20,398 sequences after removing the contamination using SeqClean.

4.3 Screening for Resistant Genes in Primary Dataset

The cassava ESTs and transcript sequences were screened against resistant genes using BLASTX. Total number of sequences from NCBI-EST and Phytozome were 120398. After screening using BLASTX, the virus resistant genes obtained from UniProtKB. The sequences similar to resistant gene were 16299 sequences (Figure 3). About 86% i.e., 104099 sequences were screened out by this process.

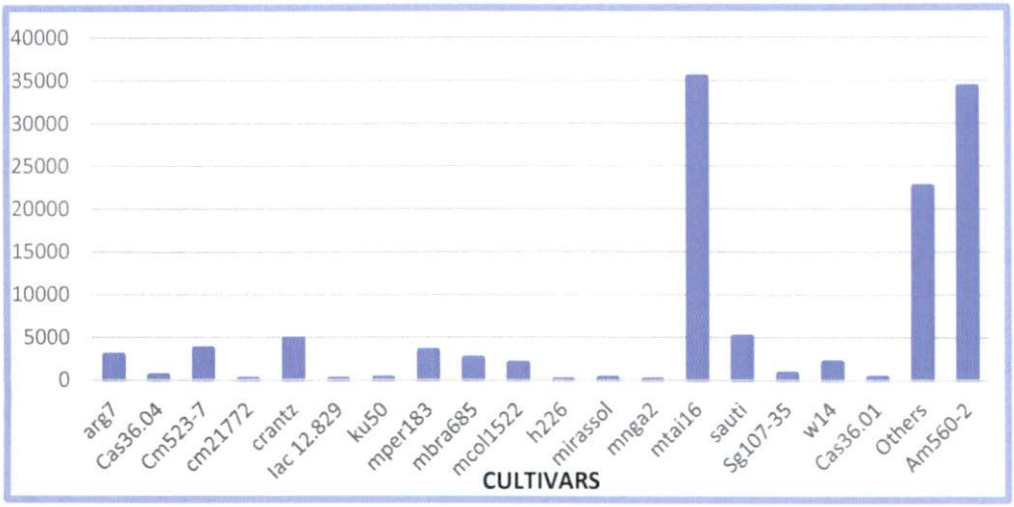


Figure 2. Distribution of sequences based on cassava cultivars

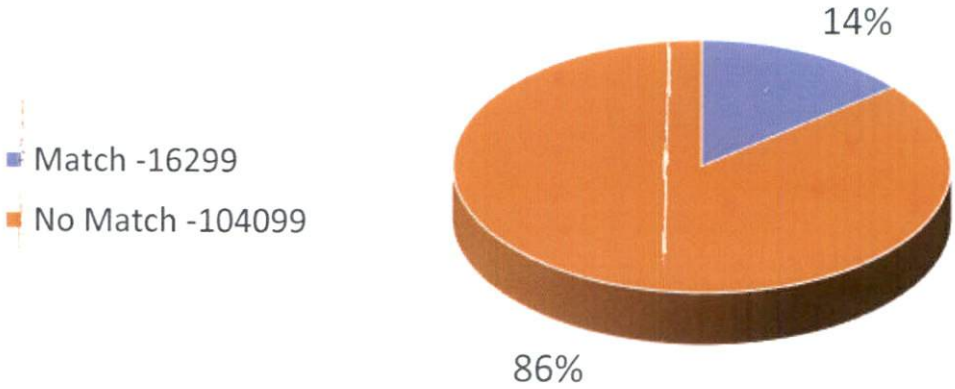


Figure 3. Distribution of primary dataset and screened sequences using virus resistant database

Table 1: Cassava sequences from genbank classified according to cultivars

SI NO	CULTIVARS	NO: 0F SEQUENCES	NO: 0F CLEANED SEQUENCES
1	arg7	2924	2913
2	Cas36.04	488	488
3	Cm523-7	3608	3607
4	cm21772	95	94
5	crantz	4764	4764
6	lac 12.829	63	63
7	ku50	172	172
8	mper183	3391	3391
9	mbra685	2506	2503
10	mcol1522	1979	1979
11	h226	21	21
12	mirassol	210	210
13	mnga2	40	32
14	mtai16	35400	35392
15	Sauti,Gomani,Mbundumali,TME1 and Mkondezi	5046	5027
16	Sg107-35	720	720
17	w14	2089	2086
18	Cas36.01	254	254
19	Others	22540	22531
20	Am560-2	34151	34151
	TOTAL	120461	120398

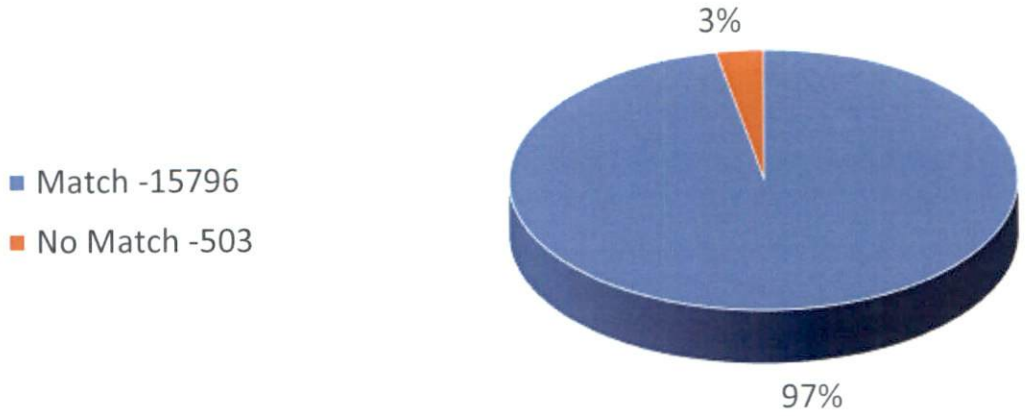


Figure 4. Distribution of annotated data

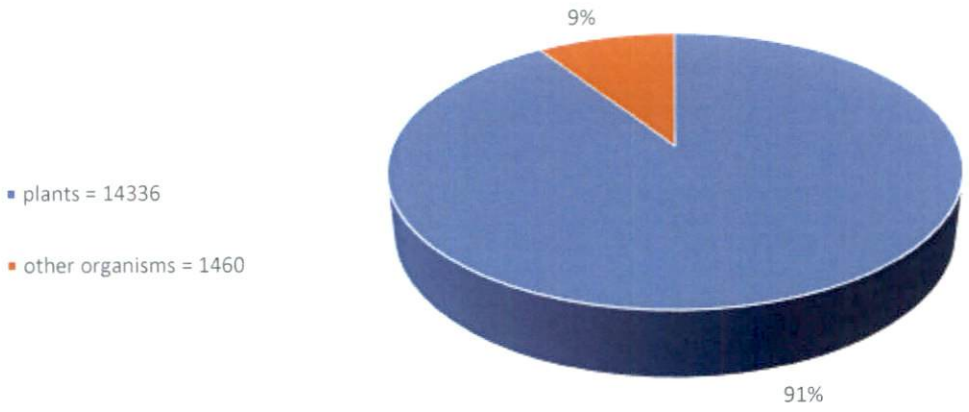


Figure 5. Distribution of sequences based on similarity to plants

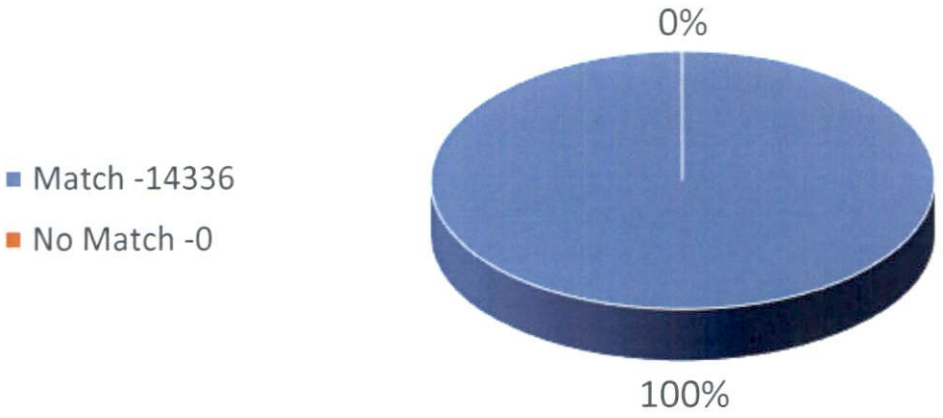


Figure 6. Percentage of matching queries after final annotation

4.4 Screening of Primary Dataset

Annotation of screened sequences was done using Uniprot/SwissProt database. cassava ESTs and transcript sequences were screened against resistant genes are annotated and only sequences which have functional annotations were retained. The number of sequences after annotation was reduced to 15796 sequences. (Figure 4). About 3% were screened out after this process i.e., about 503 sequences was removed. For screening for plants from annotated data, Uniprot/SwissProt database is used to screen sequences for presence of sequences with similarity to any organism other than plants. After screening 14336 sequences was left. About 1460 sequences were eliminated because of the presence of sequences similarity to organisms other than plants (Figure 5). About 9% were screened out after this step i.e., about 1460 sequences belonging to other organisms was removed. For confirmation of annoted sequences, SwissProt/Uniprot database was used for screening sequences again annotate to confirm. A 100% match was obtained in this step and the resulting sequence were taken as the input for the prediction of SNP and SSR (Figure 6). No sequence was left out after annotation.

4.4 Assembling of sequences using CAP3

The sequences after screening are aligned and assembled using cap3. 2088 contigs were obtained from 14336 sequences with similarity to virus resistant genes. Default runtime options were used for clustering and aligning the sequences to obtain contigs.

4.6 Identification of SNPs Using QualitySNP

A total of 128 SNPs were identified using QualitySNP. Based on the annotation data SNPs were classified as SNPs in coding region, non-coding region and SNPs in untranslated region. About 56 SNPs were found in coding region. These SNPS can alter proteins. About 65 SNPs are predicted from non-coding regions and 76 SNPS from untranslated region.

Based on the type of SNPs these were further classified into synonymous SNPs and non-synonymous SNPs (Table 2). About 30 SNPs are nonsynonymous SNPs. This means that all these SNPs will effect a change in the translated protein. About 26 SNPs were synonymous means that the mutations will not cause any change in the system. Again, based on the type of polymorphism these SNPS can be classified into SNPs and InDels. About 72 SNPs and 56 InDels are obtained. The total number of transitions - 67 is marginally greater than the total number of transversions - 54 yielding a transition-to-transversion ratio of 1.24.

Table 2. Distribution of transition and transversion SNPs from QualitySNP

CHARACTERIZATION	TYPE OF SNP	SNPs	TOTAL
TRANSITION	C/T	33	67
	G/A	34	
TRANSVERSION	A/C	14	54
	A/T	11	
	C/G	17	
	T/G	12	

4.7 Identification of SNPs Using AutoSNP

From 2088 contigs created from 14336 sequences, a total of 18081 SNPs were identified by AutoSNP. Based on the type of SNPs they can be classified into Transitions, Transversions and finally InDels. A total of 8827 Transitions was identified, 6840 Translations was identified, and 2414 InDels were identified by AutoSNP.

The Transition to Transversion ratio was 1.29 which was comparable to QualitySNP which had transition to transversion ratio of 1.24

4.5 Identification of SSRs Using MISA

From 2088 contigs created from 14336 sequences which had similarity to virus resistant genes, a total of 3297227 sequences were examined. About 582 SSRs were identified (Table 3). From 2088 contigs, only 461 contigs had SSRs and about 95 contigs had more than one SSR. A total of 45 compound SSRs were found.

Table 3. Distribution to different repeat type classes in MISA

Type of SSR	No: of SSR
Mono	217
Di	132
Tri	139
Tetra	3
Penta	1
Hexa	3
Poly	42
Total	537

4.6 Identification of SSRs Using SSRIT

From 2088 contigs created from 14336 sequences which had similarity to virus resistant genes, a total of 3297227 sequences were examined. About 133 SSRs were identified (Table 4). From 2088 contigs, only 125 contigs had SSRs and about 8 contigs had more than one SSR. No compound SSRs were found

Table 4. Distribution to different repeat type classes in SSRIT

Type of SSR	No: of SSR
Mono	0
Di	62
Tri	68
Tetra	3
Penta	0
Hexa	0
Poly	0
Total	**133**

4.9 Comparative Evaluation of SNP Target Prediction Tools

SNP target prediction tools are implemented either in the form of a web server or as a standalone tool. Out of the 12 published SNP target prediction tools, 8 are available for offline use with free license. 4 are available through web servers. A summary of all the SNP tools is presented in (Table 5).

The SNP prediction tools: QualitySNP and AutoSNP were compared. The results of SNP target prediction tools: QualitySNP and AutoSNP are summarized in terms of types of polymorphism (Table 6). Of the two SNP considered for SNP prediction, the ratio between the polymorphisms in AutoSNP is 1.29 slightly higher when compared to QualitySNP which has a ratio of 1.24.

SINGLE NUCLEOTIDE POLYMORPHISM IDENTIFICATION TOOLS					
SINO:	PROGRAM	WEBSITE	AUTHOURS	SOFTWARE	PLATFORM
1.	autoSNP	http://acpfg.imb.uq.edu.au	Zhang et al.	OFFLINE	LINUX
2.	HaploSNPer	http://www.bioinformatics.nl/tools/haplosnper/	Chang et al.	ONLINE	CLOUD
3.	InSNP	www.mucosa.de/insnp/	Weckx et al.	ONLINE	CLOUD
4.	novoSNP	http://www.molgen.ua.ac.be/bioinfo/novosnp	Savage et al.	OFFLINE	WINDOWS/LINUX
5.	PolyBayes	http://bioinformatics.bc.edu/marthlab/PolyBayes	Batley et al.	OFFLINE	LINUX
6.	PolyPhred	http://droog.mbt.washington.edu/	Marth et al.	OFFLINE	LINUX
7.	QualitySNP	http://www.bioinformatics.nl/tools/snpweb/	Tang et al.	OFFLINE	LINUX
8.	Seq-SNPing	(http://bio.kuas.edu.tw/Seq-SNPing	Dereeper et al.	OFFLINE	WINDOWS
9.	SNiPlay	http://sniplay.cirad.fr/	Marth et al.	ONLINE	CLOUD
10.	SNPdetector	http://lpg.nci.nih.gov	Manaster et al.	OFFLINE	LINUX
11.	SNP-PHAGE	http://bfgl.anri.barc.usda.gov/ML/snp-phage	Tang et al.	OFFLINE	LINUX
12.	SNPServer	http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html	Matukumalli et al.	ONLINE	CLOUD

TABLE 5. LIST OF SINGLE NUCLEOTIDE POLYMORPHISM IDENTIFICATION TOOLS

Table 6. Comparative study of SNPs from AutoSNP and QualitySNP

Type of polymorphism	No: of polymorphisms in AutoSNP	No: of polymorphisms in QualitySNP
Transition	8827	67
Transversion	6840	54
Indels	2414	72
Total	18081	193

4.10 Comparative Evaluation of SSR Target Prediction Tools

SSR target prediction tools are implemented either in the form of a web server or as a standalone tool. Out of the 8 published SSR target prediction tools, 8 are available for offline use with free license. Among these SSRIT and TRF has both online and offline interface. A summary of all the tools is presented in Table 7.

The SSR prediction tools: MISA and SSRIT were compared. The results of SSR target prediction tools: MISA and SSRIT are summarized in terms of number of SSRs in a category. Of the two SSR considered for SSR prediction, the number of SSRs in different categories is high for MISA when compared to SSRIT. SSRIT fails to identify polySSRs (Figure 7).

SIMPLE SEQUENCE REPEATS IDENTIFICATION TOOLS					
SINO:	Program	WebSite	AUTHOURS	SOFTWARE	PLATFORM
1.	MicroSAtellite (MISA)	http://pgrc.ipk-gatersleben.de/misa/	Thiel et al.	OFFLINE	LINUX
2.	Repeat Finder	http://www.cbcb.umd.edu/software/RepeatFinder/	Volfovsky et al.	OFFLINE	WINDOWS/LINUX
3.	Sputnik	http://espressosoftware.com/sputnik/	Abajian et al.	OFFLINE	WINDOWS/LINUX
4.	SSR identification Tool (SSRIT)	http://www.gramene.org/db/searches/ssrtool	Kantety et al.	BOTH	LINUX
5.	SSR Locator	http://www.ufpel.edu.br/	Maia et al.	OFFLINE	LINUX
6.	SSRPoly	http://acpfg.imb.uq.edu.au/ssrpoly.php	Tang et al.	OFFLINE	LINUX
7.	Tandem repeat Finder (TRF)	http://tandem.bu.edu/trf/trf.html	Benson et al.	BOTH	WINDOWS/LINUX
8.	Tandem repeat Occurrence Locator (TROLL)	http://wsmartins.net/webtroll/troll.html	Castelo et al.	OFFLINE	LINUX

TABLE 7. LIST OF SIMPLE SEQUENCE REPEATS IDENTIFICATION TOOLS

Figure 7. SSR distribution in MISA and SSRIT

4.13 Validation of SNP and SSR markers for CMD Resistance

A total of 204 SNP contigs and 537 SSR containing contigs are predicted using QualitySNP and MISA respectively. But for validation only a few contigs are selected. The selection was based on the percentage of hits in BLAST with resistant gene database. All contigs with SNP and SSR are blasted together against the resistant gene database and the contigs with hit percentage between 80%-100% were selected for primer designing. Nine SNP and nine SSR containing contigs were selected for primer designing using primer3 plus. A total of 48 SNP and 43 SSR primer pairs were designed from the selected SNP and SSR contigs. (Table 8 Table 9)

4.12 Selected Sequences for Primer Sequencing

On the basis of GC content and melting temperature five SNPs and five SSRs containing contigs were selected from a total of 204 SNPs and 537 SSRs for validation (Table 10, Table 11). All the selected SNPs and SSRs have a product ranging between 500 bp and 600 bp. (Table 12)

Table 12. Predicted markers and selected markers for primer synthesis

Type of polymorphism	**No: of sequences with polymorphism**	**Selected for primer synthesis**
SNP	204	5
SSR	537	5

SI.NO	CONTIG	LEFT PRIMER				RIGHT PRIMER				ELIGIBILITY
		L.PRIMER	LENGTH	TM	GC	R.PRIMER	LENGTH	TM	GC	
1	896	CACTGTGTGTGCATGGAAGC	20	60	55	GGAACCCAGTAAGCAGGCAT	20	60	55	TRUE
		CTTCCCTTACCTCGCGTTGT	20	60	55	ACTCTCCCCATCGCTCATCT	20	60.1	55	TRUE
		GCCTTTGCAAATGGGTGGTT	20	59.9	50	CTGCAGAAGCCTTGTGAGGA	20	60	55	TRUE
		CAGAGAAAGGCTTCCGGGTT	20	60	55	CAAAAGCACAGGGGGACTCT	20	59.9	55	TRUE
		ACCTCCCATCAAACAAGCCC	20	60.3	55	CTCCTTTGCCTTGGCCCTTA	20	60	55	TRUE
2	732	AGACCTTCAAGTCCTTGTAGCA	22	59	45.5	ATCAAGCGTACCATCGTGCA	20	60.1	50	FALSE
		CAAGTCCTTGTAGCATTCTTGCG	23	60.4	47.8	GGGACGATGCCTCTTCTGAC	20	60.2	60	FALSE
		GGCATACGATCAAGAAGACCT	21	57.5	47.6	GCCTCTTCTGACGAGCTCAA	20	59.8	55	FALSE
		CGATCAAGAAGACCTTCAAGTCC	23	59.1	47.8	GCCACAAGCTGAGGTAGTCC	20	60.4	60	FALSE
		ACCTTCAAGTCCTTGTAGCATTCT	24	60	41.7	ACGATGCCTCTTCTGACGAG	20	59.5	55	FALSE
3	1930	TCGTCAGCCTTTAGCCCAAG	20	60	55	CAGCAAATCTTACAAGAGGAAGTGA	25	59.3	40	FALSE
		ATTAGCGCATCCGTACCTGG	20	60	55	TCAATCTATTGCACACAATAAGCA	24	57.2	33.3	FALSE
		ATCCTCGTCAGCCTTTAGCC	20	59.5	55	GCAAATCTTACAAGAGGAAGTGAGAA	26	59.6	38.5	FALSE
		ATGAACCAGTGTGCCCAGTT	20	59.8	50	TGCACACAATAAGCATTAACTACA	24	57.2	33.3	FALSE
		CTGATGAACCAGTGTGCCCA	20	60.3	55	TGCACACAATAAGCATTAACTACATTG	27	59.2	33.3	FALSE
4	1889	ATTCGTAGGGGAGTTGGCAC	20	59.7	55	CGCTCTGGTTGGAGTTGGAT	20	60	55	TRUE
		TTGTTGTGCCCTAGCTCTGG	20	60	55	GTGCCAACTCCCCTACGAAT	20	59.7	55	TRUE
		GGCTGGAATGAATGTTGCCC	20	59.8	55	TGCCCCACAGCTTTGAAGAT	20	59.9	50	TRUE
		TCAACCAACACAAAGGGGGT	20	59.7	50	ATCTGCCCCACAGCTTTGAA	20	59.9	50	TRUE
		GGAGTGGTGACTTGCCTCAG	20	60.3	60	TCTGCACCAAGGTCTCCTCT	20	60.2	55	TRUE
5	1043	AAAGAGCTTGTCGGATCCGG	20	60.1	55	CTCTGGACCTTCTAGTCGCG	20	59.6	60	TRUE
		CCGTGGACTCTCTCTGCATC	20	59.9	60	TGGACCTTCTAGTCGCGGAT	20	60.4	55	TRUE
		CAATTCCGGCGTCAACCATG	20	60.2	55	CGCTCAAATGGTCCACTGGT	20	60.6	55	TRUE
		GATCCGACCCTTGTCAACCC	20	60.4	60	GGCAAGTTGGGCAGTTTTCA	20	59.5	50	TRUE
		TGCAACCAGGATAATCGCGT	20	60.1	50	GGTTTTAGGCAAGTTGGGCAG	21	60	52.4	TRUE
6	361	GGCCAGGATGAATCGTCGAT	20	60	55	AAGACCACCGGCTTTGAGAG	20	60	55	TRUE
		CGCGGAGACTTTGACCTCAT	20	60.1	55	TTGCTCGCTAAGGCTGACAA	20	60	50	TRUE
		CCGTTAATCAGGCAGGTGGT	20	60	55	GGATCGCACTCATGGTCACA	20	60.1	55	TRUE
		CATCGCGGAGACTTTGACCT	20	60.1	55	CATCAGCCCACCATTGCAAC	20	60.1	55	TRUE
		TTGGAAGGCGCTCATTTCCT	20	60	50	TCTCTCCCCAGTTCTCACCC	20	60.3	60	TRUE
7	463	CTCTCTCGCCTGCTGTCTTC	20	60.2	60	ACGTGCCTTGGAGTACTTGG	20	60	55	TRUE
		AAAAGGGCAGCTCTCTCTCG	20	59.8	55	GAGTACTTGGGAGTGGTGGC	20	60	60	TRUE
		GCCTGCTGTCTTCGACAAGT	20	60.6	55	GGTGGCATCCATCTTGTTGC	20	59.8	55	TRUE
		CTCTCTCTCGCCTGCTGTC	19	59.6	63.2	GGCATCCATCTTGTTGCAGC	20	60.2	55	TRUE
		CTCGCCTGCTGTCTTCGAC	19	60.8	63.2	GGAGTGGTGGCATCCATCTT	20	59.7	55	TRUE
8	567	TCTCTCTCTCTCTCGCCACG	20	60.5	60	ACACATTAGCAGTGGAACCCT	21	59.3	47.6	FALSE
		AAAACCCTAGGATCTGGCGC	20	60.1	55	CCTCCACAAACACAAAAGACTCC	23	59.9	47.8	FALSE
		TGCTTGAATGGCGGGTAGAG	20	60.1	55	ACACCTCCACAAACACAAAAGAC	23	59.8	43.5	FALSE
		GCGCTTTCTCTCTCCTTGGA	20	59.8	55	ACAAACACAAAAGACTCCAACCC	23	59.8	43.5	FALSE
		CGAAGAAGGCCACCTCTAGG	20	59.5	60	ACAACTACACCTCCACAAACAC	22	58.7	45.5	FALSE
9	1136	GCTGCTGCTAATCGGAGACT	20	59.9	55	ATTTGCTTGCAGTTGGCTGG	20	60	50	TRUE
		GCGACTGCCCTTTAACCTCT	20	60	55	GGTCAAGAAAGCCTGCTCCA	20	60.3	55	TRUE
		GCTGCTAATCGGAGACTCCTC	21	60	57.1	TTCTGCTGGATTGGCTGTCC	20	60.3	55	TRUE
		TCCTCCGTTGGAAAATCGTGT	21	59.9	47.6	CTGGATTGGCTGTCCCTTCA	20	59.7	55	FALSE
		TCTCTCTCTCTCACGCTTTGC	21	59.8	52.4	CCCCAGCCATGGTCAAGAAA	20	60.3	55	TRUE

Table 12. List of predicted SNP primers

173722

SI.NO	CONTIG	SSR	LEFT PRIMER				RIGHT PRIMER				ELIGIBILITY
			L.PRIMER	LENGTH	TM	GC	R.PRIMER	LENGTH	TM	GC	
1	Contig58	(CT)6	ACTCTTGGTCAAATTAATCTGGATCT	26	58.4	34.6	TAGCACAAACAGGGTCCTCC	20	59.3	55	FALSE
			TGGTCAAATTAATCTGGATCTCTCT	25	57.6	36	AGCCAGCAACGTATACAACCA	21	60	47.6	FALSE
			TGGATACTCTTGGTCAAATTAATCTGG	27	59.2	37	TCCAATCATAGCCAGCAACGT	21	60.1	47.6	FALSE
2	Contig254	(GCA)5	ATGCAGCAAACCAACGGTTC	20	60	50	CGGAAGGGCTGATCTGTGTT	20	60	55	TRUE
			AAGATGCAGCAAACCAACGG	20	59.7	50	CTTCCAATCGCCTGTCACCT	20	60	55	TRUE
			CAGCCACATCAGAATCAGCG	20	59.3	55	GCATCAGTCACATCTGCAGC	20	59.6	55	TRUE
			TCTCTGCGCGAAAATTGCAC	20	60.1	50	TGACCAGTCTGCTTATTGCGT	21	60	47.6	FALSE
			TCTGAAATCAGCACACCCGC	20	61	55	CAGAACCAGCTTCCAATCGC	20	59.6	55	TRUE
3	Contig1362	(AAG)8	ACTGCATCGCAACTTTCAGC	20	59.8	50	TGGGTAAGCTTTCCTGCTAGC	21	60.1	52.4	TRUE
			ACATCCACTGCATCGCAACT	20	60.3	50	AGCTTTCCTGCTAGCTCTTGA	21	59.1	47.6	FALSE
			GCTTTGGAGCTGCTAACTGG	20	59.2	55	AGATGGGTAAGCTTTCCTGCT	21	59.1	47.6	FALSE
			TCTACATCCACTGCATCGCA	20	59.2	50	GGTAAGCTTTCCTGCTAGCTCT	22	59.8	50	TRUE
			TCCACTGCATCGCAACTTTC	20	59.1	50	TCCTGCTAGCTCTTGATTTCGT	22	59.5	45.5	FALSE
4	Contig1053	(TCT)7	CTATGGTCCATCGGCCTGTC	20	60	60	ACCCAACACTCACAACCTGG	20	60.1	55	TRUE
			CATCGGCCTGTCCTTACTGG	20	60.2	60	CACTTCCCAAGTCGCTGGAT	20	60	55	TRUE
			GTCCATCGGCCTGTCCTTAC	20	60.2	60	GGCAAGCCATCGGAAAACTC	20	59.8	55	TRUE
			TTCCTATGGTCCATCGGCCT	20	60.4	55	ACAACCTGGCAAAACAACGG	20	59.8	50	TRUE
			AGCCCTTGGTTTCCAGAACT	20	59.1	50	TGAACCGGACACTTCCCAAG	20	59.9	55	TRUE
5	Contig2063	(GGA)7	GTTGGGGATTCATGGGTCGT	20	60	55	AGAGCGACAGAATGCTTCCA	20	59.4	50	TRUE
			GTGGTTGGGGATTCATGGGT	20	60	55	TGCTGGGAAGGTATCAAGGG	20	59.1	55	TRUE
			TGCCTGTTGATGAGGTGGAT	20	59	50	CCAGGCACAAATCCAATGGC	20	60.1	55	TRUE
			TAGAGGAGTGGGAGCTTGGT	20	59.6	55	CCTGGAAAGCACAGTTGTTGG	20	59.9	52.4	TRUE
			GTCGTGCCCTATTTGCAGGT	20	60.7	55	GCTACCGTGTTAATGCTGGG	20	59	55	TRUE
6	Contig2063	(TG)11	GTTGGGGATTCATGGGTCGT	20	60	55	AGAGCGACAGAATGCTTCCA	20	59.4	50	TRUE
			GGCAAACAGAGGGAGGAAGA	20	59.3	55	CCAGGCACAAATCCAATGGC	20	60.1	55	TRUE
			GTGGTTGGGGATTCATGGGT	20	60	55	TGCTGGGAAGGTATCAAGGG	20	59.1	55	TRUE
			TGCCTGTTGATGAGGTGGAT	20	59	50	TCCCCAGGCACAAATCCAAT	20	59.6	50	TRUE
			TAGAGGAGTGGGAGCTTGGT	20	59.6	55	CCTGGAAAGCACAGTTGTTGG	21	59.9	52.4	TRUE
7	Contig953	(A)11	ATCAGCAGAAACTCTAAACCCT	22	57.1	40.9	CAGCTTTGGTGTTGTCCTGC	20	60	55	FALSE
			TCAGCAGAAACTCTAAACCCTAATCA	26	60	38.5	ATGCTGACTGGAAAGGCTCC	20	60	55	FALSE
			ACCCTAATCAAAATGAACCCAAAA	24	57.2	33.3	CGTGCTCCTGAAGTGATCGA	20	59.8	55	FALSE
			AGCAGAAACTCTAAACCCTAATCAAAA	27	59.3	33.3	CTCCGCTTCAATCACCAGGT	20	60	55	FALSE
			TCTAAACCCTAATCAAAATGAACCCA	26	58.8	34.6	GGAAAGGCTCCGCTTCAATC	20	59.3	55	FALSE
8	Contig414	(T)10G(T)11	CAACCAAAGAAAGCGGAGGC	20	60	55	ACATGCAGATCTGTGATCTTCT	22	57.2	40.9	FALSE
			GCTGCCGGAACTGGAATTTAC	21	59.9	52.4	ACATGCAGATCTGTGATCTTCTAAA	25	58.4	36	FALSE
			GCGTTATCCAGAGCTCCTCA	20	59.3	55	ACATCACTCATTCTTTTGTACATGT	25	57.2	32	FALSE
			AAGGCACAAGTGGTAGGGTG	20	59.9	55	AATATTTTCACCTTGTATTCCCACAA	26	57.6	30.8	FALSE
			GGCTGCCGGAACTGGAATT	19	60.7	57.9	TGCAGATCTGTGATCTTCTAAAAATT	26	57.4	30.8	FALSE
9	Contig1246	(A)11	ATGGAAAGGGGAGAGGGACA	20	59.9	55	CACCGCTCTCCGAAACGATA	20	59.9	55	TRUE
			TCGGTCTTTCCATCGCATCC	20	60.2	55	CGAATTGTTTGGGATCGGGC	20	59.9	55	TRUE
			GCAGCTGCTCATCACTCTGT	20	60.4	55	GCGGTTGTCCTGGTTCTGTA	20	60	55	TRUE
			GGAGAGGGACAGGGGAAAGT	20	60.5	60	TGGTTCTGTAAGCGGCCTTT	20	59.9	50	TRUE
			TGGATGGAAAGGGGAGAGGG	20	60.6	60	GTTGCGGCTTTTAGGTCAGC	20	60.1	55	TRUE

Table 13. List of predicted SSR primers

Sl.No	Primer No.	Forward labelled primer (5'-3')	Reverse primer (5'-3')	Product size	Dye used
1.	SNP896	CACTGTGTGTGCATGGAAGC	GGAACCCAGTAAGCAGGCAT	543	6-FAM
2.	SNP1043	CAATTCCGGCGTCAACCATG	CGCTCAAATGGTCCACTGGT	589	NED
3.	SNP361	CCGTTAATCAGGCAGGTGGT	GGATCGCACTCATGGTCACA	509	VIC
4.	SNP463	GCCTGCTGTCTTCGACAAGT	GGTGGCATCCATCTTGTTGC	517	PRT
5.	SNP1136	GCGACTGCCCTTTAACCTCT	GGTCAAGAAAGCCTGCTCCA	595	6-FAM

Table 8. Selected SNPs for primer synthesis

Sl.No	Primer No.	Forward labelled primer (5'-3')	Reverse primer (5'-3')	Product size	Dye used
1.	SSR254	CAGCCACATCAGAATCAGCG	GCATCAGTCACATCTGCAGC	578	6-FAM
2.	SSR1362	ACTGCATCGCAACTTTCAGC	TGGGTAAGCTTTCCTGCTAGC	563	NED
3.	SSR1053	CTATGGTCCATCGGCCTGTC	ACCCAACACTCACAACCTGG	556	VIC
4.	SSR2063	TAGAGGAGTGGGAGCTTGGT	CCTGGAAAGCACAGTTGTTGG	519	PRT
5.	SSR414	CAACCAAAGAAAGCGGAGGC	ACATGCAGATCTGTGATCTTCT	565	6-FAM

Table 9. Selected SSRs for primer synthesis

Of the selected five SNP and SSR markers for primer synthesis, only forward primers are fluorescent labelled. Four different fluorescent dyes were used. 6-FAM, NED, VIC, PET. A total of 10 cassava varieties including five CMD resistant and five susceptible cassava were selected based on field trials (Table 13).

Table 13. Samples of CMD resistant and susceptible cassava for validation

SI:NO	RESISTANT	SUSCEPTIBLE
1	Albert	CI 732
2	96/1089A	CO2
3	Cr 11/43	Ambakadan
4	TME – 3	Sree vijaya
5	MNga-1	Sree jaya

4.16 Primer Synthesis

Primers were synthesized by a company called EUROFINS (Figure 8). Forward and reverse primer of all 5 SNPs and 5 SSRs were synthesized and delivered by EUROFINS. HSPA was used for the purification of all these primers.

4.17 Genomic DNA extraction

Fresh young leaves were collected from 10 cassava varieties and DNA was isolated from these leaf samples using the method described by (Dellaporta *et al.*, 1983).

4.17.1 Determination of quality of DNA

Quality of DNA was determined by Agarose Gel Electrophoresis (Plate 1). Clear bands were observed in the gel. Quantification of DNA was done using UV spectrophotometry at 260/280 nm.

eurofins mwg|operon

Oligonucleotide Synthesis Report

Mr Anup Kumar R
Biogene

Order ID 7063064
Customer ID 603391
Your Order ID (PO#) 86006 / 86006

Order Date 11/08/2015
Lab No 2025
No. of Oligos 20/20

Eurofins Genomics India Pvt Ltd
#54B/1, Doddanekundi
Industrial Area 2,
Hoodi, Whitefield,
Bangalore 560048,
Karnataka
India

No	Oligo Name	Sequence (5' -> 3')	Yield [OD]	Yield [µg]	Yield [nmol]	Concentration [pmol/µl]	Vol. for 100pmol/µl	Tm [°C]	MW [g/mol]	GC-Content	Synthesis Scale	Purification	Modification	Barcode IDO	QC Report
1	SSR254-F	CAGCCACATCAGAATCAGCG (20)	14.6	394	64.7	-	647	59.4	6079	55 %	0.01 µmol	HPSF	-	001778215	-
2	SSR254-R	GCATCAGTCACATCTGCAGC (20)	8.8	242	39.9	-	399	59.4	6061	55 %	0.01 µmol	HPSF	-	001778216	-
3	SSR1362-F	ACTGCATCGCAACTTTCAGC (20)	13.1	376	62.2	-	622	57.3	6036	50 %	0.01 µmol	HPSF	-	001778217	-
4	SSR1362-R	TGGGTAAGCTTTCCTGCTAGC (21)	12.3	362	56.3	-	563	59.8	6428	52.4 %	0.01 µmol	HPSF	-	001778218	-
5	SSR1063-F	CTATGGTCCATCGGCCTGTC (20)	12.4	381	62.9	-	629	61.4	6059	60 %	0.01 µmol	HPSF	-	001778219	-
6	SSR1053-R	ACCCAACACTCACAACCTGG (20)	9.5	263	43.8	-	438	59.4	5999	55 %	0.01 µmol	HPSF	-	001778220	-

Figure 8. oligonucleotide synthesis report

```
Contig896      TTAAAGAATCCGGAAGCTAACAAAGAAGCCTTTGCAAATGGGTGGTTTCATTCTGGTGAT 1380
2R             ------------------------------------ATGGGAGG-----YWCKG--WAT 474
                                                   *****:**       *.*   **

Contig896      CTCGCTGTGAAGCATCCAGATGGGTATATAGAAATCAAGGATAGAAGCAAGGACATTAGC 1440
2R             CCCACT--CAGGCTTCCA------------------------------------------ 490
               * *.**   *.**:****

Contig896      ATTTCAGGAGGTGAAAACATTAGTAGCTTGGAAGTAGAAAATGTGCTATATACGCACCCA 1500
2R             ------------------------------------------------------TGCACCAC 498
                                                                      *****..

Contig896      GCAGTGTATGAAGTATCTGTGGTAGCCAGAGAAGATGAGCGATGGGGAGAGTCCCCCTGT 1560
2R             ACAGTGA----------------------------------------------------- 505
               .*****:

Contig896      GCTTTTGTCACATTGAAACCAGGCATGGAGAAATCTAGTGAAGGAAGTTTGGCAGAAGAT 1620
2R             ------------------------------------------------------------

Contig896      ATAATAAAGTTTTGTCGGTCGAAAATGCCTGCTTACTGGGTTCCAAAATCTGTTGTATTT 1680
2R             ------------------------------------------------------------
```

Figure 9. ClustalX alignment of SNP896 and MNga showing predicted SNP at 1493th position

Plate 1: Quality of DNA was determined by Agarose Gel Electrophoresis

4.18 Screening of SNP

To confirm that the SNP is working AGE is done in all 5 susceptible and 5 resistant samples. AGE will give a straight single band that confirms that the SNPs are working. SNP 896, SNP 1043 gives a straight bright band in the gel (Plate 2). SNP 361 and SNP 1136 does not give bands in all the varieties, so they are not considered for the validation part as markers for CMD resistance (Plate 3).

4.19 Screening of SSR

To confirm that the designed SSR markers were working, AGE was done. Clear thick bands were obtained in the gel (Plate 4 and 5). These confirm that the SSRs were amplified.

4.20 Validation of SSR

Validation of SSR was done in PAGE. Validation of SSR is done in all five susceptible and five resistant samples using PAGE. PAGE will clearly differentiate the bands produced by the SSR primers which is not possible to view in AGE. Only SSR 2063 showed clear difference in separation of the bands between resistant and susceptible which indicate that SSR 2063 can distinguish between CMD resistant and CMD susceptible varieties (Plate 6).

4.21 Validation of SNPs

Validation of SNP was done by eluting the separated bands from the gel and then sequencing it to obtain the sequence. This sequence was aligned with the corresponding contig sequence from which the respective primer was designed. Multiple sequence alignment is done using ClustalX.

The bands were eluted from the gel using elution kit and these sequences were analyzed using 3500 capillary DNA Genetic Analyzer (Applied Biosystem). Three technical replicates were carried out to avoid sequencing errors. These sequences are then aligned against its respective contigs using ClustalX (Figure 9). Sequence bands from resistant variety MNga and susceptible variety CI732 which contains the designed primers SNP896 and SNP1043 was sequenced. These

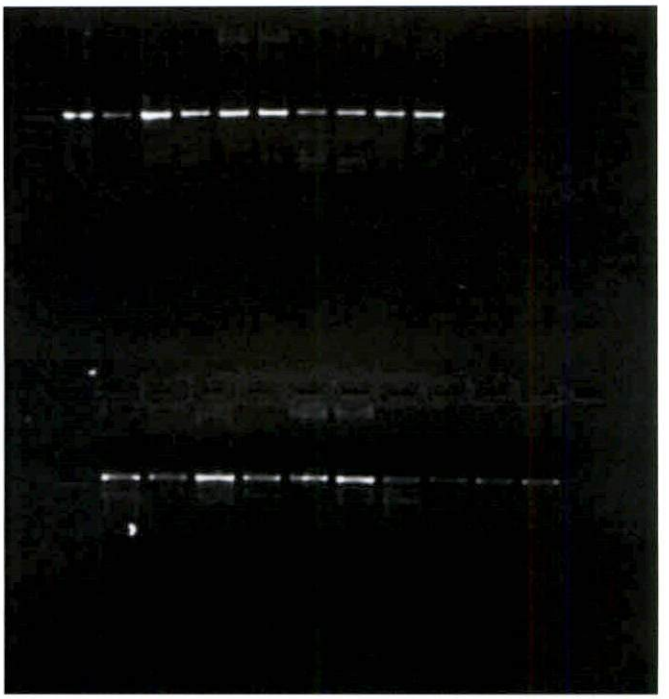


Plate 2. Gel image of SNP 896 and SNP 1136

Plate 3. Gel image of SNP 1043 and SNP 1136

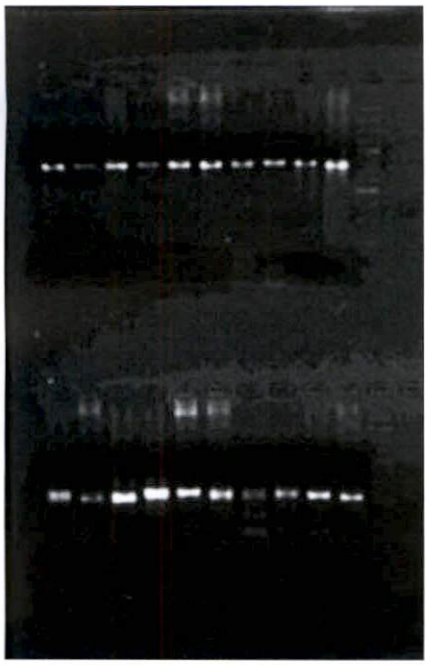


Plate 4: Gel image of SSR1362 and SSR2063

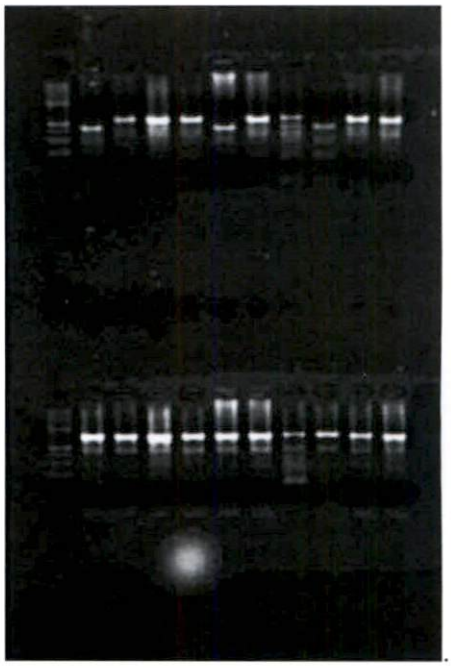


Plate 5. Gel image of SSR1053 and SSR414

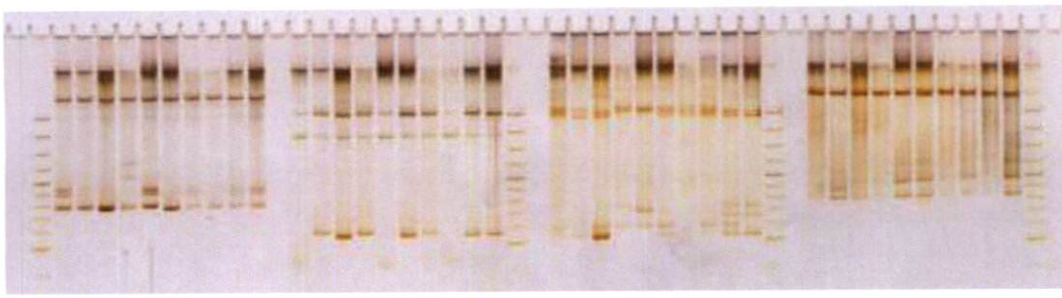


Plate 6. Gel image of SSR1362 SSR1053, SSR2063, SSR414

sequences are aligned against contig896 and contig1043 from which these primers were designed. ClustalX is used for multiple sequence alignment. The results showed that the sequence with SNP1043 did not showed any variation in predicted SNP site but SNP896 in MNga showed SNP at the 1493th position as designed but with a variation in the base. The predicted SNP was that Cytosine(C) will be replaced by Adenine(A), where a Cytosine (C) is replaced with Thymine (T), whereas the same SNP896 in CI732 showed no variance in that position.

DISCUSSION

5. DISCUSSION

Molecular markers are important for plant research and breeding, and are being applied to accelerate effective plant selection through marker-assisted selection, based on genome-level selection of chromosomal segments. In plant genetic research, molecular markers are also being used for the analysis of population structure, the study of evolutionary relationships, and, in sequenced model systems such as Arabidopsis, for studies on the genetic structure of individuals at the whole-genome level (Cao *et al.*, 2011).

Single-nucleotide polymorphism (SNP) and simple sequence repeats (SSR) markers have recently gained interest in the scientific and plant-breeding communities (Rafalski, 2002). Studies on genetic mapping and molecular marker development in cassava have been published (Akano *et al.*, 2002; Okogbenin & Fregene, 2002; Rabbi *et al.*, 2012), and several studies have focused on the analysis or discovery of simple sequence repeat loci and mapping of quantitative trait loci (Whankaew *et al.*, 2011). To further promote progress in genetics and breeding, higher-density markers, such as SNP markers, are required. SNPs and insertions and deletions (InDels) are common natural mutations in populations (Cho *et al.*, 1999; Syvanen *et al.*, 1999). The SNPs and InDels discovered in cassava (Ferguson *et al.*, 2012; Lopez *et al.*, 2005) are quite important for cassava breeding research; cassava is an outcrossing species and produces botanical seed in many environments, but is mainly propagated using stem cuttings. Thus, most cassava cultivars are considered heterozygous, which makes it more difficult to develop molecular markers (Prochnik *et al.*, 2012). Therefore, it is necessary to detect additional DNA polymorphisms using the available cassava genome and transcribed sequences to improve molecular marker development in cassava. DNA polymorphism discovery is important not only for molecular breeding but also for understanding gene function, elucidation of the relationship between polymorphisms, gene function, and gene duplication (Kalyna *et al.*, 2012; Yngvadottir *et al.*, 2009).

Cassava mosaic disease (CMD) is the most- important disease of cassava (Manihot esculenta) in Africa, and is a potential threat to Latin American (LA) cassava production. Although this viral disease is still unknown in LA, its vector – the whitefly – has recently been found. The disease is best controlled through host-plant resistance, which was first found in third backcross derivatives of an interspecific cross between cassava and Manihot glaziovii, and is thought to be polygenic. The marker, SSRY28, is located on linkage group R of the male-parent-derived molecular genetic map. The gene, designated as CMD2, is flanked by the SSR and RFLP marker GY1 at 9 and 8 cM, respectively. This is the first report of qualitative virus resistance in cassava, and of molecular markers that tag CMD resistance in cassava. We discuss the use of markers linked to CMD2 for marker-assisted breeding of CMD resistance in Latin America and for in- creasing the cost-effectiveness of resistance breeding in Africa.(Akano *et al.*, 2002).

A dominant gene for resistance to CMD has been found by conventional genetic analysis and molecular genetic mapping in a F1 cross between resistant and susceptible parents. The major gene nature also means that a genetic marker for marker assisted selection (MAS) can easily be identified. MAS would thus become an invaluable tool for breeding CMD resistance in Latin America where the disease is not found, but where the presence of the vector makes it a threat. Selecting for high levels of resistance with a marker may be more efficient than conventional breeding in Africa, where rapid deployment of high resistance genes into cassava gene pools is needed to protect cassava from the ravages of CMD. The advantage of MAS is that the breeder can, in early stages, eliminate CMD-susceptible genotypes. In the case of a heterozygous CMD-resistant donor parent, elimination would be 50%, reducing the costs of disease evaluation by half and increasing selection efficiency. The breeder can then concentrate on fewer genotypes at the seedling and crucial single-row trial stages where progenies are reduced by as much as 95%. Identification of markers for other traits in addition to CMD resistance can be used to choose parents more efficiently that combine the different traits The gene designated as CMD2 is different from the earlier found CMD1, which controls the currently deployed resistance. CMD2 is located on linkage group R, whereas

CMD1 is on linkage group D of the cassava molecular map (Fregene et al., 1997). The action of the two genes is also different: CMD2 is dominant, whereas CMD1 appears recessive in that its effect is detected only in backcross progeny, and not in the F1. The presence of two different sources of CMD resistance, and markers in tight linkage with them, provides a means of combining multiple sources of resistance. The recessive nature of the older source of resistance, however, makes it less attractive, given cassava's out-crossing and heterozygous nature.

In this work about 204 SNPs and 537 SSRs are predicted which is exclusievely related to CMD resistance in cassava. These can be validated and screened for effective markers against CMD resistance. More than 56 SNPs are confirmed in the coding region which makes them candidate SNPs for screening for resistance against CMD resistance. More than 30 SNPs are nonsynonymous which will result in change in the transcription product.

5.1 Comparative evaluation of SNP prediction tools

On comparative evaluation of QualitySNP and AutoSNP, QualitySNP shows more promising SNPs unlike AutoSNP where a huge number of SNPs including false positive SNPs are predicted. QualitySNP showed unique ability to annotate and classify SNPs based on their polymorphism, Based on the type of annotation data and based on the type of SNP. All these are not possible in AutoSNP where classification is entirely based on the type of SNPs. QualitySNP gave a more detailed and precise information whereas AutoSNP succeed in predicting thousands of SNPs but the viable ones are hard to find from the enormous list of SNPs identified by AutoSNP.

Comparative evaluation of SSR prediction tools

On comparative evaluation of MISA and SSRIT, MISA shows more promising SNPs unlike SSRIT where only di, tri and tetra SSRs are identified. MISA on the other hand scans for mono, di, tri, tetra, penta, hexa, and poly SSRs. SSRIT completely neglects complex SSRs. MISA has a more robust script for

identifying various types of SSRs. MISA even recognized double the number of SSRs found by SSRIT within the same time period.

5.2 Sequence summary of DNA polymorphism discovery

Sequences of cassava were classified into 20 cultivars including an uncategorized category. A total of 120461 sequences were obtained in total and pre-processed using SeqClean. After cleaning, exactly 120398 sequences were obtained and used as the primary dataset for DNA polymorphism discovery. This result is similar to the result obtained by the work done on genome-wide discovery and information development of DNA polymorphism in cassava (Sakurai *et al.*, 2013). Here they categorized the cassava sequence into 17 cultivars including a uncategorized category. This had a total of 114782 sequences and after preprocessing they were able to obtain 96885 sequences as their primary dataset for DNA polymorphism discovery.

Additional categories like ARG7, H226 and MTai16 were categorized. One category named MCol22 with only 7 sequences which they obtained was not obtained in this study.

5.3 DNA polymorphism discovery

SNPs and InDels were identified using AutoSNP and QualitySNP where SNPs were identified by the prebuilt categories defined in the tool, but users can change the default values according to needs. Contigs are aligned using CAP3 on both tools and contigs were used to find DNA polymorphisms.

A similar computational analysis of SNP was carried out by (*Sakurai et al., 2013*). Polymorphisms (SNPs and InDels) were discovered from the contig sequence alignment according to the following criteria: (i) The contig could be aligned with the cassava draft genome sequence (Prochnik *et al.*, 2012) (ii) The nucleotide at the polymorphism site was not N; (iii) The SNP consisted of 2 types of nucleotides (to avoid false SNP detection due to cross-contamination with other loci in the contig sequence alignment); (iv) The polymorphism was supported by at least 2 sequences

in a cassava variety; (v) The nucleotide at the polymorphism site was the same in the contig sequence alignment of each variety; (vi) There were fewer than 3 other discontinuous nucleotide polymorphisms around 5 bp of a SNP site .(Sakurai *et al.*, 2013).

As a result they were able to discover a total of 10546 SNPs and 674 InDels from the whole genome of cassava. Using SNP identification tools on sequences with similar CMD resistant genes, about 15667 SNPs and 2414 InDels were obtained using AutoSNP and 56 SNPs while 72 InDels were obtained by using QualitySNP. Since the work is restricted to one particular character the number of SNPs should be less compared to the work done in whole genome sequence. This is one of the reasons why the results of QualitySNP was used for validation of SNPs. Based on the annotation data from the results of QualitySNP, about 67 transitions and 54 transversions were obtained which were related to resistance for CMD which had a ratio of 1.24, whereas the whole genome polymorphism discovery gave a result of 5845 transitions and 4701 transversions. (Sakurai *et al.*, 2013). The transition to transversion ratio was also 1.24.

With the help of these prediction tools we will be able to develop novel markers which can be used for a lot of applications. The availability of large EST sequence data makes it an economical choice to develop SSR and SNP markers. EST SSR and EST SNP are gene specific and thus functional molecular markers. All these computational tools for DNA polymorphism discovery will help in identification of SNPs and SSRs in sequence data as well as for designing primers for these markers. These will help plant breeders, new to molecular breeding and marker assisted selection to opt for SSR and SNP markers to solve crop disease related problems. Since we have screened the whole sequences for similarity with virus resistance genes, the number of sequences for identification of SSR and SNP has been considerably reduced and the time taken for the identification of markers got significantly reduced.

SUMMARY

6. SUMMARY

The study entitled "Molecular marker development of cassava mosaic disease resistance using bioinformatics tools and its validation." was conducted at the Central Tuber Crop Research Institute during 2014-2015. The objectives of the study included development and evaluation of various SNP and SSR prediction pipelines, computational prediction and characterization of SNP and SSR in cassava, verification of SNP and SSR markers for cassava mosaic disease (CMD) resistant and susceptible breeding lines. The salient findings of the study are summarized below.

The SNP prediction tool QualitySNP was found to be a better tool compared to AutoSNP. QualitySNP had better SNP prediction algorithm and the ability for classification of the identified SNPs into various categories. It has the ability to annotate and identify nonsynonymous and synonymous SNPs which helps to select more precise SNPs for the research work. The SSR prediction tool MISA was found to be better compared to SSRIT. MISA had better SSR prediction algorithm and the ability for classification of SSRs based on the type of SSR. Mono, di, tri, tetra, penta, hexa and poly SSRs are identified in MISA.

The preliminary data set for the identification of SSR/SNP markers was obtained from the EST section of NCBI (*http://www.ncbi.nlm.nih.gov/nucest*) and the cassava transcript sequences (variety AM560-2, JGI annotation v4.1) from the Phytozome website (*http://phytozome.jgi.doe.gov/pz/portal.html*). The whole sequences were classified into 20 cultivars totaling to 120461 sequences. After preprocessing and screening, the dataset was reduced to 14336 sequences. Since the sequences were compared with virus resistant genes under the screening stage, significant reduction in time taken for the identification of SSRs and SNPs could be achieved. The resulting sequences were assembled and aligned using CAP3 and 2088 contigs were obtained.

From these contigs using QualitySNP, about 56 SNPs were identified. In that 30 SNPs were nonsynonymous and 26 SNPs were synonymous SNPs. From that 5 sequences were selected for primer designing. From the 2088 contigs using MISA, about 537 SSRs were identified. In that 217 were mono, 132 were di, 139 were tri, 3 were tetra, 1 was penta 3 was hexa and 42 complex SSRs. Five sequences which have high hit percentage were selected for validation and primer designing. Primers were designed for both SNPs and SSRs for CMD resistant genes. These primers were validated using 5 resistant and 5 susceptible cassava varieties. Among the 10 primers, after validation, one SNP (SNP896) and one SSR (SSR 2063) primer was able to clearly differentiate between the resistant and susceptible varieties. This is the first report of SNPs and SSRs computationally identified and verified in wet lab.

Scope for Future work

As the resources were limited only few predicted SSRs and SNPs were validated for differentiating susceptible and resistant genes in cassava. In future, the identified 56 SNPs and 537 SSRs can be validated in wet lab and the resulting potential markers can be utilized in the breeding program for screening CMD resistance in cassava.

REFERENCES

7. REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. 252(5013): 1651-6.

Aggarwal, R.K., Hendre, P.S., Varshney, R.K., Bhat, P.R., Krishnakumar, V., Singh, L. 2007. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 114(2): 359-72.

Aharoni, A., Keizer, L.C., Bouwmeester, H.J., Sun, Z., Alvarez-Huerta, M., Verhoeven, H.A., Blaas, J., van Houwelingen, A.M., De Vos, R.C., van der Voet, H., Jansen, R.C., Guis, M., Mol, J., Davis, R.W., Schena, M., van Tunen, A.J., O'Connell, A.P. 2000. Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. *The Plant cell*. 12(5): 647-62.

Akano, O., Dixon, O., Mba, C., Barrera, E., Fregene, M. 2002. Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 105(4): 521-525.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., Lander, E.S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*. 407(6803): 513-6.

Alves AAA (2002) Cassava botany and physiology. In: Hillocks RJ, Thresh MJ, Bellotti AC (eds) Cassava: biology, production and utilisation. CABI International, Oxford, pp 67–89

Alves, A.A., Setter, T.L. 2004. Response of cassava leaf area expansion to water deficit: cell proliferation, cell expansion and delayed development. *Annals of botany*. 94(4): 605-13.

An, D., Yang, J., Zhang, P. 2012. Transcriptome profiling of low temperature-treated cassava apical shoots showed dynamic responses of tropical plant to cold stress. *BMC genomics*. 13: 64.

Anderson, J.V., Delseny, M., Fregene, M.A., Jorge, V., Mba, C., Lopez, C., Restrepo, S., Soto, M., Piegu, B., Verdier, V., Cooke, R., Tohme, J., Horvath, D.P. 2004. An EST resource for cassava and other species of Euphorbiaceae. *Plant molecular biology*. 56(4): 527-39.

Awoleye F, Duren M, Dolezel J, Novak FJ (1994) Nuclear DNA content and in vitro induced somatic polyploidization cassava (Manihot esculenta Crantz) breeding. Euphytica 76: 195–202.

Barker, G., Batley, J., H, O.S., Edwards, K.J., Edwards, D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*. 19(3): 421-2.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant physiology. 132(1): 84-91.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant physiology*. 132(1): 84-91.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*. 27(2): 573-80.

Berrie, L.C., Rybicki, E.P., Rey, M.E. 2001. Complete nucleotide sequence and host range of South African cassava mosaic virus: further evidence for recombination amongst begomoviruses. *The Journal of general virology*. 82(Pt 1): 53-8.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J.C., 3rd, Tingey, S.V., Rafalski, A. 2002. Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant molecular biology*. 48(5-6): 539-47.

Bottcher, B., Unseld, S., Ceulemans, H., Russell, R.B., Jeske, H. 2004. Geminate structures of African cassava mosaic virus. *Journal of virology*. 78(13): 6758-65.

Brookes, A.J. 1999. The essence of SNPs. *Gene*. 234(2): 177-86.

Brumfield, R. T., Beerli, P., Nickerson, D. A., Edwards, S. V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol* **18**:249–256

Burke, J., Davison, D., Hide, W. 1999. d2_cluster: a validated method for clustering EST and full-length cDNAsequences. *Genome research*. 9(11): 1135-42.

Cao, J., Shi, F., Liu, X., Jia, J., Zeng, J., Huang, G. 2011. Genome-wide identification and evolutionary analysis of Arabidopsis sm genes family. *Journal of biomolecular structure & dynamics*. 28(4): 535-44.

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., Waugh, R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics*. 156(2): 847-54.

Castelo, A.T., Martins, W., Gao, G.R. 2002. TROLL--tandem repeat occurrence locator. *Bioinformatics*. 18(4): 634-6.

Cerda, J. 2009. Molecular pathways during marine fish egg hydration: the role of aquaporins. *Journal of fish biology*. 75(9): 2175-96.

Chavarriaga -a guirre , P., Maya, m. m., Onierbale, m. w. b., Resovich, s. k., Fregene, m., Ohme, j. t., Ochert, g. k., 1998. Microsatellites in cassava (Manihot esculenta Crantz): discovery, inheritance, and var- iability. Theoretical and Applied Genetics 97: 493–501

Chen, C., Zhou, P., Choi, Y.A., Huang, S., Gmitter, F.G., Jr. 2006. Mining and characterizing microsatellites from citrus ESTs. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 112(7): 1248-57.

Chen, X., Laudeman, T.W., Rushton, P.J., Spraggins, T.A., Timko, M.P. 2007. CGKB: an annotation knowledge base for cowpea (Vigna unguiculata L.) methylation filtered genomic genespace sequences. *BMC bioinformatics*. 8: 129.

Chepelev, I., Wei, G., Tang, Q., Zhao, K. 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic acids research. 37(16): e106.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., Rafalski, A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC genetics*. 3: 19.

Cho, R.J., Mindrinos, M., Richards, D.R., Sapolsky, R.J., Anderson, M., Drenkard, E., Dewdney, J., Reuber, T.L., Stammers, M., Federspiel, N., Theologis, A., Yang, W.H., Hubbell, E., Au, M., Chung, E.Y., Lashkari, D., Lemieux, B., Dean, C., Lipshutz, R.J., Ausubel, F.M., Davis, R.W., Oefner, P.J. 1999. Genome-wide mapping with biallelic markers in Arabidopsis thaliana. *Nature genetics*. 23(2): 203-7.

Collard, B.C., Mackill, D.J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. 363(1491): 557-72.

Cooke, R., Raynal, M., Laudie, M., Grellet, F., Delseny, M., Morris, P.C., Guerrier, D., Giraudat, J., Quigley, F., Clabault, G., Li, Y.F., Mache, R., Krivitzky, M., Gy, I.J., Kreis, M., Lecharny, A., Parmentier, Y., Marbach, J., Fleck, J., Clement, B., Philipps, G., Herve, C., Bardet, C., Tremousaygue, D., Hofte, H., et al. 1996. Further progress towards a catalogue of all Arabidopsis genes: analysis of a set of 5000 non-redundant ESTs. *The Plant journal : for cell and molecular biology*. 9(1): 101-24.

Cordeiro, G.M., Casu, R., McIntyre, C.L., Manners, J.M., Henry, R.J. 2001. Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. *Plant science : an international journal of experimental plant biology*. 160(6): 1115-1123.

Dantec, L.L., Chagne, D., Pot, D., Cantin, O., Garnier-Gere, P., Bedon, F., Frigerio, J.M., Chaumeil, P., Leger, P., Garcia, V., Laigret, F., De

Daruvar, A., Plomion, C. 2004. Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. *Plant molecular biology*. 54(3): 461-70.

De Carvalho, R., Guerra, M. 2002. Cytogenetics of Manihot esculenta Crantz (cassava) and eight related species. *Hereditas*. 136(2): 159-68.

Dereeper, A., Nicolas, S., Le Cunff, L., Bacilieri, R., Doligez, A., Peros, J.P., Ruiz, M., This, P. 2011. SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC bioinformatics*. 12: 134.

Dellaporta, S.L., Wood, J. and Hicks, J.R 1983. A plant DNA minipreparation: version II. Plant Mol. Biol.Rep., **1:** 19-21

Dickau, R., Ranere, A.J., Cooke, R.G. 2007. Starch grain evidence for the preceramic dispersals of maize and root crops into tropical dry and humid forests of Panama. *Proceedings of the National Academy of Sciences of the United States of America*. 104(9): 3651-6.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nature genetics*. 21(1 Suppl): 10-4.

Edwards, D., Batley, J. 2010. Plant genome sequencing: applications for crop improvement. *Plant biotechnology journal*. 8(1): 2-9.

El-Sharkawy, M.A. 2004. Cassava biology and physiology. *Plant molecular biology*. 56(4): 481-501.

FAOSTAT (2013) Food and agriculture organizations statistics database. FAO, Rome. http://faostat3.fao.org/home/E. Accessed Nov 2014

Fauquet, C. and Fargette, D. 1990. African cassava mosaic virus. Etiology, epidemiology and control. Plant Disease 74(6): 404–411.

Ferguson, M.E., Hearne, S.J., Close, T.J., Wanamaker, S., Moskal, W.A., Town, C.D., de Young, J., Marri, P.R., Rabbi, I.Y., de Villiers, E.P. 2012. Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik. 124(4): 685-95.

Ferguson, M.E., Hearne, S.J., Close, T.J., Wanamaker, S., Moskal, W.A., Town, C.D., de Young, J., Marri, P.R., Rabbi, I.Y., de Villiers, E.P. 2012. Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik.* 124(4): 685-95.

Fregene, M., Angel, F., Gomez, R., Rodriguez, F., Chavarriaga, P., Roca, W., Tohme, J., Bonierbale, M. (1997) A molecular genetic map of cassava (Manihot esculenta Crantz). *Theor Appl Genet* **95**:431– 441

Fregene, M.A., Suarez, M., Mkumbira, J., Kulembeka, H., Ndedya, E., Kulaya, A., Mitchel, S., Gullberg, U., Rosling, H., Dixon, A.G., Dean, R., Kresovich, S. 2003. Simple sequence repeat marker diversity in cassava landraces: genetic diversity and differentiation in an asexually propagated crop. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik.* 107(6): 1083-93.

Garg, A., Wilson, R., Barnes, R., Arioglu, E., Zaidi, Z., Gurakan, F., Kocak, N., O'Rahilly, S., Taylor, S.I., Patel, S.B., Bowcock, A.M. 1999a. A gene for congenital generalized lipodystrophy maps to human chromosome 9q34. *The Journal of clinical endocrinology and metabolism.* 84(9): 3390-4.

Garg, K., Green, P., Nickerson, D.A. 1999b. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome research.* 9(11): 1087-92.

Grivet, L., Glaszmann, J.C., Vincentz, M., da Silva, F., Arruda, P. 2003. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik.* 106(2): 190-7.

Gupta, P.K., Rustgi, S., Sharma, S., Singh, R., Kumar, N., Balyan, H.S. 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular genetics and genomics : MGG.* 270(4): 315-23.

Gutierrez, R.A., Ewing, R.M., Cherry, J.M., Green, P.J. 2002. Identification of unstable transcripts in Arabidopsis by cDNA microarray analysis: rapid

decay is associated with a group of touch- and specific clock-controlled genes. *Proceedings of the National Academy of Sciences of the United States of America*. 99(17): 11513-8.

Harrison, B., Robinson, D. 1999. Natural Genomic and Antigenic Variation in Whitefly-Transmitted Geminiviruses (Begomoviruses). *Annual review of phytopathology*. 37: 369-398.

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

Heesacker, A., Kishore, V.K., Gao, W., Tang, S., Kolkman, J.M., Gingle, A., Matvienko, M., Kozik, A., Michelmore, R.M., Lai, Z., Rieseberg, L.H., Knapp, S.J. 2008. SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 117(7): 1021-9.

Hong, Y.G., Robinson, D.J., Harrison, B.D. 1993. Nucleotide sequence evidence for the occurrence of three distinct whitefly-transmitted geminiviruses in cassava. *The Journal of general virology*. 74 (Pt 11): 2437-43.

Huang, X., Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome research*. 9(9): 868-77.

Ichikawa, T., Nakazawa, M., Kawashima, M., Iizumi, H., Kuroda, H., Kondou, Y., Tsuhara, Y., Suzuki, K., Ishikawa, A., Seki, M., Fujita, M., Motohashi, R., Nagata, N., Takagi, T., Shinozaki, K., Matsui, M. 2006. The FOX hunting system: an alternative gain-of-function gene hunting technique. *The Plant journal : for cell and molecular biology*. 48(6): 974-85.

Iyer, R.R., Pluciennik, A., Rosche, W.A., Sinden, R.R., Wells, R.D. 2000. DNA polymerase III proofreading mutants enhance the expansion and deletion of triplet repeat sequences in Escherichia coli. *The Journal of biological chemistry*. 275(3): 2174-84.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M., Last, R.L. 2002. Arabidopsis map-based cloning in the post-genome era. *Plant physiology*. 129(2): 440-50.

Jayashree, B., Punna, R., Prasad, P., Bantte, K., Hash, C.T., Chandra, S., Hoisington, D.A., Varshney, R.K. 2006. A database of simple sequence repeats from cereal and legume expressed sequence tags mined in silico: survey and evaluation. *In silico biology*. 6(6): 607-20.

Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B., Marshall, J., Fuller, J., Cardle, L., McNicol, J., Dinh, H.Q., Barta, A., Brown, J.W. 2012. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic acids research*. 40(6): 2454-69.

Kantety, R.V., La Rota, M., Matthews, D.E., Sorrells, M.E. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant molecular biology*. 48(5-6): 501-10.

Kashi, Y., King, D., Soller, M. 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends in genetics : TIG*. 13(2): 74-8.

Kashi, Y., King, D.G. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in genetics : TIG*. 22(5): 253-9.

Khlestkina, E.K., Than, M.H., Pestsova, E.G., Roder, M.S., Malyshev, S.V., Korzun, V., Borner, A. 2004. Mapping of 99 new microsatellite-derived loci in rye (Secale cereale L.) including 39 expressed sequence tags. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 109(4): 725-32.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, C.J., Ohtsuki, K., Shishiki, T., Foundation of Advancement of International Science Genome, S., Analysis, G., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuki, Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizuno,

K., Yokomizo, S., Niikura, J., Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Riken, Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konno, H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M., Hayashizaki, Y., Yasunishi, A. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*. 301(5631): 376-9.

Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J., Graner, A. 2001. Generation and comparison of EST-derived SSRs and SNPs in barley (Hordeum vulgare L.). *Hereditas*. 135(2-3): 145-51.

Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro, C.F. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America*. 95(18): 10774-8.

Kunkeaw, S., Tan, S., Coaker, G. 2010. Molecular and evolutionary analyses of Pseudomonas syringae pv. tomato race 1. *Molecular plant-microbe interactions : MPMI*. 23(4): 415-24.

Lee, J.M., Williams, M.E., Tingey, S.V., Rafalski, J.A. 2002. DNA array profiling of gene expression changes during maize embryo development. *Functional & integrative genomics*. 2(1-2): 13-27.

Li, K., Zhu, W., Zeng, K., Zhang, Z., Ye, J., Ou, W., Rehman, S., Heuer, B., Chen, S. 2010a. Proteome characterization of cassava (Manihot esculenta Crantz) somatic embryos, plantlets and tuberous roots. *Proteome science*. 8: 10.

Li, Y.Z., Pan, Y.H., Sun, C.B., Dong, H.T., Luo, X.L., Wang, Z.Q., Tang, J.L., Chen, B. 2010b. An ordered EST catalogue and gene expression profiles of cassava (Manihot esculenta) at key growth stages. *Plant molecular biology*. 74(6): 573-90.

Liang, X., Chen, X., Hong, Y., Liu, H., Zhou, G., Li, S., Guo, B. 2009. Utility of EST-derived SSR in cultivated peanut (Arachis hypogaea L.) and Arachis wild species. *BMC plant biology*. 9: 35.

Lokko, Y., Anderson, J.V., Rudd, S., Raji, A., Horvath, D., Mikel, M.A., Kim, R., Liu, L., Hernandez, A., Dixon, A.G., Ingelbrecht, I.L. 2007. Characterization of an 18,166 EST dataset for cassava (Manihot esculenta Crantz) enriched for drought-responsive genes. *Plant cell reports*. 26(9): 1605-18.

Lopez, C., Jorge, V., Piegu, B., Mba, C., Cortes, D., Restrepo, S., Soto, M., Laudie, M., Berger, C., Cooke, R., Delseny, M., Tohme, J., Verdier, V. 2004. A unigene catalogue of 5700 expressed genes in cassava. *Plant molecular biology*. 56(4): 541-54.

Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J., Verdier, V. 2005. Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (Manihot esculenta Crantz). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 110(3): 425-31.

Lopez, C.E., Zuluaga, A.P., Cooke, R., Delseny, M., Tohme, J., Verdier, V. 2003. Isolation of Resistance Gene Candidates (RGCs) and characterization of an RGC cluster in cassava. *Molecular genetics and genomics : MGG*. 269(5): 658-71.

Mah, J.T., Chia, K.S. 2007. A gentle introduction to SNP analysis: resources and tools. *Journal of bioinformatics and computational biology*. 5(5): 1123-38.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nature genetics*. 23(4): 452-6.

Matukumalli, L.K., Grefenstette, J.J., Hyten, D.L., Choi, I.Y., Cregan, P.B., Van Tassell, C.P. 2006. Application of machine learning in SNP discovery. *BMC bioinformatics*. 7: 4.

Miller, M.P., Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Human molecular genetics*. 10(21): 2319-28.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R., Hide, W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome research*. 9(11): 1143-55.

Mochida, K., Shinozaki, K. 2010. Genomics and bioinformatics resources for crop improvement. *Plant & cell physiology*. 51(4): 497-523.

Morgante, M., Olivieri, A.M. 1993. PCR-amplified microsatellites as markers in plant genetics. *The Plant journal : for cell and molecular biology*. 3(1): 175-82.

Nanjo, T., Sakurai, T., Totoki, Y., Toyoda, A., Nishiguchi, M., Kado, T., Igasaki, T., Futamura, N., Seki, M., Sakaki, Y., Shinozaki, K., Shinohara, K. 2007. Functional annotation of 19,841 Populus nigra full-length enriched cDNA clones. *BMC genomics*. 8: 448.

Ng, P.C., Henikoff, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*. 31(13): 3812-4.

Ohlrogge, J., Benning, C. 2000. Unraveling plant metabolism by EST analysis. Current opinion in plant biology. 3(3): 224-8.

Ohlrogge, J., Benning, C. 2000. Unraveling plant metabolism by EST analysis. *Current opinion in plant biology*. 3(3): 224-8.

Okogbenin, E., Fregene, M. 2002. Genetic analysis and QTL mapping of early root bulking in an F1 population of non-inbred parents in cassava (Manihot esculenta Crantz). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 106(1): 58-66.

Okogbenin, E., Fregene, M. 2003. Genetic mapping of QTLs affecting productivity and plant architecture in a full-sib cross from non-inbred parents in Cassava (Manihot esculenta Crantz). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 107(8): 1452-62.

Okogbenin, E., Marin, J., Fregene, M. A.: An SSR-based molecular genetic map of cassava. Euphytica 2006, 147:433-440.

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

Amazonian origin of domestication. *American journal of botany*. 88(1): 131-42.

Olsen, K.M., Schaal, B.A. 1999. Evidence on the origin of cassava: phylogeography of Manihot esculenta. *Proceedings of the National Academy of Sciences of the United States of America*. 96(10): 5586-91.

Oztur, Z.N., Talame, V., Deyholos, M., Michalowski, C.B., Galbraith, D.W., Gozukirmizi, N., Tuberosa, R., Bohnert, H.J. 2002. Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. *Plant molecular biology*. 48(5-6): 551-73.

Paritosh, K., Yadava, S.K., Gupta, V., Panjabi-Massand, P., Sodhi, Y.S., Pradhan, A.K., Pental, D. 2013. RNA-seq based SNPs in some agronomically important oleiferous lines of Brassica rapa and their use for genome-wide linkage mapping and specific-region fine mapping. BMC genomics. 14: 463.

Pashley, C.H., Ellis, J.R., McCauley, D.E., Burke, J.M. 2006. EST databases as a source for molecular markers: lessons from Helianthus. *The Journal of heredity*. 97(4): 381-8.

Patil, B.L., Rajasubramaniam, S., Bagchi, C., Dasgupta, I. 2005. Both Indian cassava mosaic virus and Sri Lankan cassava mosaic virus are found in India and exhibit high variability as assessed by PCR-RFLP. *Archives of virology*. 150(2): 389-97.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M. 1999. Mining SNPs from EST databases. *Genome research*. 9(2): 167-74.

Pinto, L.R., Oliveira, K.M., Ulian, E.C., Garcia, A.A., de Souza, A.P. 2004. Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*. 47(5): 795-804.

Poncet, V., Rondeau, M., Tranchant, C., Cayrel, A., Hamon, S., de Kochko, A., Hamon, P. 2006. SSR mining in coffee tree EST databases: potential use

of EST-SSRs as markers for the Coffea genus. *Molecular genetics and genomics : MGG*. 276(5): 436-49.

Potokina, E., Sreenivasulu, N., Altschmied, L., Michalek, W., Graner, A. 2002. Differential gene expression during seed germination in barley (Hordeum vulgare L.). *Functional & integrative genomics*. 2(1-2): 28-39.

Prochnik, S., Marri, P.R., Desany, B., Rabinowicz, P.D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T., Rokhsar, D.S., Rounsley, S. 2012. The Cassava Genome: Current Progress, Future Directions. Tropical plant biology. 5(1): 88-94.

Prochnik, S., Marri, P.R., Desany, B., Rabinowicz, P.D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T., Rokhsar, D.S., Rounsley, S. 2012. The Cassava Genome: Current Progress, Future Directions. *Tropical plant biology*. 5(1): 88-94.

Puonti-Kaerlas. J. (2001) Molecular biology of cassava. *Hort Rev* **26**: 85–159

Rabbi, I.Y., Kulembeka, H.P., Masumba, E., Marri, P.R., Ferguson, M. 2012. An EST-derived SNP and SSR genetic linkage map of cassava (Manihot esculenta Crantz). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*. 125(2): 329-42.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology. 5(2): 94-100.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Current opinion in plant biology*. 5(2): 94-100.

Raji, A.A., Anderson, J.V., Kolade, O.A., Ugwu, C.D., Dixon, A.G., Ingelbrecht, I.L. 2009. Gene-based microsatellites for cassava (Manihot esculenta Crantz): prevalence, polymorphisms, and cross-taxa utility. *BMC plant biology*. 9: 118.

Reilly, K., Bernal, D., Cortes, D.F., Gomez-Vasquez, R., Tohme, J., Beeching, J.R. 2007. Towards identifying the full set of genes expressed during cassava post-harvest physiological deterioration. *Plant molecular biology*. 64(1-2): 187-203.

Reymond, P., Weber, H., Damond, M., Farmer, E.E. 2000. Differential gene expression in response to mechanical wounding and insect feeding in Arabidopsis. *The Plant cell*. 12(5): 707-20.

Richmond, T., Somerville, S. 2000. Chasing the dream: plant EST microarrays. *Current opinion in plant biology*. 3(2): 108-16.

Riva, A., Kohane, I.S. 2002. SNPper: retrieval and analysis of human SNPs. *Bioinformatics*. 18(12): 1681-5.

Roa, A.C., Chavarriaga-Aguirre, P., Duque, M.C., Maya, M.M., Bonierbale, M.W., Iglesias, C., Tohme, J. 2000. Cross-species amplification of cassava (Manihot esculenta) (Euphorbiaceae) microsatellites: allelic polymorphism and degree of relationship. *American journal of botany*. 87(11): 1647-55.

Rothenstein, D., Haible, D., Dasgupta, I., Dutt, N., Patil, B.L., Jeske, H. 2006. Biodiversity and recombination of cassava-infecting begomoviruses from southern India. *Archives of virology*. 151(1): 55-69.

Sakurai, T., Kondou, Y., Akiyama, K., Kurotani, A., Higuchi, M., Ichikawa, T., Kuroda, H., Kusano, M., Mori, M., Saitou, T., Sakakibara, H., Sugano, S., Suzuki, M., Takahashi, H., Takahashi, S., Takatsuji, H., Yokotani, N., Yoshizumi, T., Saito, K., Shinozaki, K., Oda, K., Hirochika, H., Matsui, M. 2011. RiceFOX: a database of Arabidopsis mutant lines overexpressing rice full-length cDNA that contains a wide range of trait information to facilitate analysis of gene function. *Plant & cell physiology*. 52(2): 265-73.

Sakurai, T., Mochida, K., Yoshida, T., Akiyama, K., Ishitani, M., Seki, M., Shinozaki, K. 2013. Genome-wide discovery and information resource development of DNA polymorphisms in cassava. *PloS one*. 8(9): e74056.

Sakurai, T., Plata, G., Rodriguez-Zapata, F., Seki, M., Salcedo, A., Toyoda, A., Ishiwata, A., Tohme, J., Sakaki, Y., Shinozaki, K., Ishitani, M. 2007. Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response. *BMC plant biology*. 7: 66.

Sato, S., Isobe, S., Asamizu, E., Ohmido, N., Kataoka, R., Nakamura, Y., Kaneko, T., Sakurai, N., Okumura, K., Klimenko, I., Sasamoto, S., Wada, T.,

Watanabe, A., Kohara, M., Fujishiro, T., Tabata, S. 2005. Comprehensive structural analysis of the genome of red clover (Trifolium pratense L.). *DNA research : an international journal for rapid publication of reports on genes and genomes*. 12(5): 301-64.

Saunders, K., Salim, N., Mali, V.R., Malathi, V.G., Briddon, R., Markham, P.G., Stanley, J. 2002. Characterisation of Sri Lankan cassava mosaic virus and Indian cassava mosaic virus: evidence for acquisition of a DNA B component by a monopartite begomovirus. *Virology*. 293(1): 63-74.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 270(5235): 467-70.

Schlotterer, C. 2004. The evolution of molecular markers--just a matter of fashion? *Nature reviews. Genetics*. 5(1): 63-9.

Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T., Weisshaar, B. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. *Genome research*. 13(6A): 1250-7.

Schmitz, P.M., Kavallari, A. 2009. Crop plants versus energy plants--on the international food crisis. Bioorganic & medicinal chemistry. 17(12): 4020-1.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A., Shinozaki, K. 2002. Functional annotation of a full-length Arabidopsis cDNA collection. *Science*. 296(5565): 141-5.

Sharopova, N., McMullen, M.D., Schultz, L., Schroeder, S., Sanchez-Villeda, H., Gardiner, J., Bergstrom, D., Houchins, K., Melia-Hancock, S., Musket, T., Duru, N., Polacco, M., Edwards, K., Ruff, T., Register, J.C., Brouwer, C., Thompson, R., Velasco, R., Chin, E., Lee, M., Woodman-Clikeman, W., Long, M.J., Liscum, E., Cone, K., Davis, G., Coe, E.H., Jr. 2002.

Development and mapping of SSR markers for maize. *Plant molecular biology*. 48(5-6): 463-81.

Singh, R., Pandey, B., Danishuddin, M., Sheoran, S., Sharma, P., Chatrath, R. 2011. Mining and survey of simple sequence repeats in wheat rust Puccinia sp. *Bioinformation*. 7(6): 291-5.

Siqueira, M.V., Pinheiro, T.T., Borges, A., Valle, T.L., Zatarim, M., Veasey, E.A. 2010. Microsatellite polymorphisms in cassava landraces from the Cerrado biome, Mato Grosso do sul, Brazil. *Biochemical genetics*. 48(9-10): 879-95.

Soderlund, C., Descour, A., Kudrna, D., Bomhoff, M., Boyd, L., Currie, J., Angelova, A., Collura, K., Wissotski, M., Ashley, E., Morrow, D., Fernandes, J., Walbot, V., Yu, Y. 2009. Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. *PLoS genetics*. 5(11): e1000740.

Sojikul, P., Kongsawadworakul, P., Viboonjun, U., Thaiprasit, J., Intawong, B., Narangajavana, J., Svasti, M.R. 2010. AFLP-based transcript profiling for cassava genome-wide expression analysis in the onset of storage root formation. *Physiologia plantarum*. 140(2): 189-98.

Sraphet, S., Boonchanawiwat, A., Thanyasiriwat, T., Boonseng, O., Tabata, S., Sasamoto, S., Shirasawa, K., Isobe, S., Lightfoot, D.A., Tangphatsornruang, S., Triwitayakorn, K. 2011. SSR and EST-SSR-based genetic linkage map of cassava (Manihot esculenta Crantz). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik.* 122(6): 1161-70.

Sunyaev, S., Ramensky, V., Bork, P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in genetics : TIG*. 16(5): 198-200.

Syvanen, A.C., Landegren, U., Isaksson, A., Gyllensten, U., Brookes, A. 1999. First International SNP Meeting at Skokloster, Sweden, August 1998. Enthusiasm mixed with scepticism about single-nucleotide polymorphism

markers for dissecting complex disorders. *European journal of human genetics : EJHG.* 7(1): 98-101.

Taji, T., Sakurai, T., Mochida, K., Ishiwata, A., Kurotani, A., Totoki, Y., Toyoda, A., Sakaki, Y., Seki, M., Ono, H., Sakata, Y., Tanaka, S., Shinozaki, K. 2008. Large-scale collection and annotation of full-length enriched cDNAs from a model halophyte, Thellungiella halophila. *BMC plant biology.* 8: 115.

Tang, J., Leunissen, J.A., Voorrips, R.E., van der Linden, C.G., Vosman, B. 2008. HaploSNPer: a web-based allele and SNP detection tool. *BMC genetics.* 9: 23.

Tangphatsornruang, S., Sraphet, S., Singh, R., Okogbenin, E., Fregene, M., Triwitayakorn, K. 2008. Development of polymorphic markers from expressed sequence tags of Manihot esculenta Crantz. *Molecular ecology resources.* 8(3): 682-5.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome research.* 11(8): 1441-52.

Thiel, T., Michalek, W., Varshney, R.K., Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik.* 106(3): 411-22.

Toth, G., Gaspari, Z., Jurka, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome research.* 10(7): 967-81.

Umezawa, T., Sakurai, T., Totoki, Y., Toyoda, A., Seki, M., Ishiwata, A., Akiyama, K., Kurotani, A., Yoshida, T., Mochida, K., Kasuga, M., Todaka, D., Maruyama, K., Nakashima, K., Enju, A., Mizukado, S., Ahmed, S., Yoshiwara, K., Harada, K., Tsubokura, Y., Hayashi, M., Sato, S., Anai, T., Ishimoto, M., Funatsuki, H., Teraishi, M., Osaki, M., Shinano, T., Akashi, R., Sakaki, Y., Yamaguchi-Shinozaki, K., Shinozaki,

K. 2008. Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. *DNA research : an international journal for rapid publication of reports on genes and genomes*. 15(6): 333-46.

Useche, F.J., Gao, G., Harafey, M., Rafalski, A. 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome informatics. International Conference on Genome Informatics*. 12: 194-203.

Utsumi, Y., Tanaka, M., Morosawa, T., Kurotani, A., Yoshida, T., Mochida, K., Matsui, A., Umemura, Y., Ishitani, M., Shinozaki, K., Sakurai, T., Seki, M. 2012. Transcriptome analysis using a high-density oligomicroarray under drought stress in various genotypes of cassava: an important tropical crop. *DNA research : an international journal for rapid publication of reports on genes and genomes*. 19(4): 335-45.

van Hal, N.L., Vorst, O., van Houwelingen, A.M., Kok, E.J., Peijnenburg, A., Aharoni, A., van Tunen, A.J., Keijer, J. 2000. The application of DNA microarrays in gene expression analysis. *Journal of biotechnology*. 78(3): 271-80.

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

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

Varshney, R.K., Hoisington, D.A., Tyagi, A.K. 2006. Advances in cereal genomics and applications in crop breeding. *Trends in biotechnology*. 24(11): 490-9.

Volfovsky, N., Haas, B.J., Salzberg, S.L. 2001. A clustering method for repeat analysis in DNA sequences. *Genome biology*. 2(8): RESEARCH0027.

Wang, Z., Moult, J. 2001. SNPs, protein structure, and disease. *Human mutation*. 17(4): 263-70.

Weber, J.L., May, P.E. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American journal of human genetics*. 44(3): 388-96.

Whankaew, S., Poopear, S., Kanjanawattanawong, S., Tangphatsornruang, S., Boonseng, O., Lightfoot, D.A., Triwitayakorn, K. 2011. A genome scan for quantitative trait loci affecting cyanogenic potential of cassava root in an outbred population. *BMC genomics*. 12: 266.

Yngvadottir, B., Xue, Y., Searle, S., Hunt, S., Delgado, M., Morrison, J., Whittaker, P., Deloukas, P., Tyler-Smith, C. 2009. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *American journal of human genetics*. 84(2): 224-34.

Yu, J.K., Dake, T.M., Singh, S., Benscher, D., Li, W., Gill, B., Sorrells, M.E. 2004. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*. 47(5): 805-18.

Yue, P., Li, Z., Moult, J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of molecular biology*. 353(2): 459-73.

Yue, P., Moult, J. 2006. Identification and analysis of deleterious human SNPs. *Journal of molecular biology*. 356(5): 1263-74.

Zhang, W., Olson, N.H., Baker, T.S., Faulkner, L., Agbandje-McKenna, M., Boulton, M.I., Davies, J.W., McKenna, R. 2001. Structure of the Maize streak virus geminate particle. *Virology*. 279(2): 471-7.

Zhu, T., Budworth, P., Chen, W., Provart, N., Chang, H.S., Guimil, S., Su, W., Estes, B., Zou, G., Wang, X. 2003a. Transcriptional control of nutrient partitioning during rice grain filling. *Plant biotechnology journal*. 1(1): 59-70.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., Cregan, P.B. 2003b. Single-nucleotide polymorphisms in soybean. *Genetics*. 163(3): 1123-34.

APPENDICES

APPENDIX I

Preparation of DNA extraction buffer (Dellaporta *et al.*, 1983)

a. Tris- HCl (pH 8.0) : 100mM
b. EDTA (pH 8.0) : 50mM
c. NaCl : 500mM
d. β-mercaptoethanol : 0.2 % (v/v) freshly added prior to DNA extraction
e. PVP : 2% (w/v)
f. SDS : 20%
g. Potassium acetate : 500mM
h. Ice-cold Isopropanol
i. Sodium acetate(pH 5.2) : 300mM
j. RNase A
   i. 10 mg/ml (RNase A was dissolved in TE buffer and boiled for 15 minutes at 100^0 C to destroy DNase and stored at –20 o C).
k. Chloroform:Isoamyl alcohol : (24:1)
l. Ethanol : 70%

APPENDIX II

Preparation of TE buffer (10X)

1. Tris- HCl (pH 8.0) :10 mM
2. EDTA : 1 mM

Final volume made upto 100ml with distilled water.

APPENDIX III

TBE buffer (10X)

1. Tris base : 107 g
2. Boric acid : 55 g
3. 0.5 M EDTA (pH 8.0) : 40 ml
4. Final volume made up to 1000 ml with distilled water and autoclave before use.

APPENDIX IV

a. **Gel loading dye**

Glycerol

TBE(10X)

EDTA : 0.25M

SDS : 20%

Bromophenol blue : 10%

Sterile water

b. **Empty well dye**

Loading dye (Appendix IV): 50µl

Sterile distilled water : 50µl

c. **100bp marker**

100bp marker : 5µl

Loading dye : 40µl

Sterile distilled water: : 55µl

APPENDIX V

PCR Cocktail	Stock concentration	Volume taken	Final concentration	
DNA	50ng	2.0µl	5.0 ng	20 µl
Primer	2.0µM	2.0 µl	0.2 µM	
dNTPs	40µM	0 .2 µl	0.4 µM	
Taq buffer	10x	2.0 µl	1x	
Taq DNA polymerase	3.0 unit	2.0 µl	0.03 unit	
Sterile water		13.6 µl		

APPENDIX VI

Acrylamide

Acrylamide : 38g

Bis acrylamide : 2g

Made up the final volume up to 100 ml using distilled water.

APPENDIX VII

6% Polyacrylamide gel containing 7 M urea

42 g urea was dissolved in a beaker containing 10 ml TBE buffer (10X) and 15 ml distilled water by heating in a microwave oven for 30-40 s. 15 ml acrylamide solution() was filtered and added to a measuring cylinder followed by the urea solution. The final volume was made up to 100 ml using distilled water and stored in dark till use. 60 µl TEMED and 600 µl APS (100 mg/ml) was added and mixed just before casting the gel.

APPENDIX VIII

Bind silane

Absolute ethanol : 99.5%

Acetic acid : .5%

Bind silane : 1.0 µl

APPENDIX IX

Fixer

Acetic acid : 200 ml

Distilled water : 1800 ml

APPENDIX X

Silver stain

2 g silver nitrate dissolved in distilled water to a final volume of 2000 ml and 3 ml formaldehyde added.

APPENDIX XI

Developer

60 g sodium carbonate dissolved in distilled water to a final volume of 2000 ml and stored at -20°C. 3 ml formaldehyde and 4 ml sodium thiosulphate (10 mg/ml) was added and mixed thoroughly before use.

APPENDIX XII						
List of SSRs identified by MISA						
ID	SSR nr	SSR typ	SSR	size	start	end
Contig22	1	p3	(AGA)6	18	838	855
Contig24	1	p1	(T)13	13	105	117
Contig26	1	p1	(A)12	12	1078	1089
Contig26	2	p1	(T)10	10	1398	1407
Contig43	1	c	tctgctctctcagcttccctat	108	29	136
Contig44	1	p1	(T)11	11	1251	1261
Contig48	1	p2	(TC)8	16	55	70
Contig51	1	p1	(A)13	13	1011	1023
Contig54	1	p2	(AG)15	30	143	172
Contig54	2	p1	(T)14	14	1842	1855
Contig57	1	p2	(CT)6	12	38	49
Contig58	1	p2	(CT)6	12	38	49
Contig59	1	p3	(CAG)8	24	1913	1936
Contig61	1	c	(CT)9cgttttctccaa(T)11	41	70	110
Contig63	1	p1	(T)10	10	973	982
Contig65	1	p3	(TGG)5	15	2136	2150
Contig66	1	p2	(GA)7	14	28	41
Contig69	1	p3	(TAT)5	15	3760	3774
Contig70	1	p3	(TAT)5	15	3682	3696
Contig74	1	p2	(CT)7	14	19	32
Contig75	1	p2	(AG)6	12	26	37
Contig84	1	p1	(T)18	18	2548	2565
Contig87	1	p1	(C)10	10	176	185
Contig93	1	p2	(TC)6	12	53	64
Contig94	1	c	(TC)6ttcc(T)11	27	31	57
Contig94	2	p3	(TGA)6	18	576	593
Contig121	1	p2	(AG)6	12	525	536
Contig122	1	c	(CA)8gagagagat(AG)9	43	61	103
Contig123	1	p2	(CT)16	32	196	227
Contig125	1	p1	(T)13	13	492	504
Contig126	1	p1	(A)14	14	94	107
Contig141	1	p2	(TC)6	12	14	25
Contig145	1	p3	(GCT)5	15	1127	1141
Contig149	1	p1	(A)23	23	1095	1117
Contig150	1	p4	(ATCA)5	20	76	95
Contig153	1	c	(TC)13ttttt(A)10	41	1454	1494
Contig154	1	p2	(TC)10	20	46	65
Contig156	1	p3	(GGC)6	18	595	612
Contig157	1	p3	(GGC)6	18	472	489
Contig159	1	p1	(T)10	10	2253	2262
Contig160	1	p2	(AG)13	26	43	68
Contig171	1	p1	(A)10	10	1709	1718
Contig184	1	p2	(TA)10	20	1629	1648
Contig188	1	p3	(GCT)5	15	132	146
Contig194	1	c	)38gcttgacaaaacagt(C	103	1	103
Contig197	1	p2	(AT)7	14	68	81
Contig198	1	p3	(CAG)7	21	498	518

Contig200	1	p3	(ATA)6	18	473	490
Contig200	2	p2	(AT)11	22	751	772
Contig208	1	p1	(T)10	10	54	63
Contig212	1	c	gaaaataccaaatttttaac	90	1037	1126
Contig214	1	p3	(CAG)8	24	1321	1344
Contig217	1	p3	(AAG)5	15	85	99
Contig229	1	p1	(A)12	12	798	809
Contig236	1	p3	(TGA)5	15	1515	1529
Contig236	2	p1	(A)11	11	2002	2012
Contig238	1	p1	(T)10	10	261	270
Contig240	1	p2	(TC)6	12	256	267
Contig241	1	p1	(T)12	12	4555	4566
Contig248	1	p1	(T)16	16	863	878
Contig254	1	p3	(GCA)5	15	222	236
Contig255	1	p3	(CAG)5	15	94	108
Contig264	1	p1	(T)10	10	12	21
Contig271	1	p2	(TC)6	12	73	84
Contig271	2	p1	(A)12	12	1070	1081
Contig275	1	p2	(AT)11	22	1135	1156
Contig287	1	p3	(TCT)6	18	174	191
Contig289	1	p3	(TCT)7	21	616	636
Contig293	1	p3	(ATG)6	18	3069	3086
Contig302	1	p2	(CT)6	12	2787	2798
Contig303	1	p1	(T)10	10	809	818
Contig326	1	p2	(AT)7	14	1231	1244
Contig328	1	c	A)15gaagcacgct(TC)1	45	19	63
Contig335	1	p2	(CT)8	16	1175	1190
Contig338	1	p3	(CTG)5	15	3976	3990
Contig344	1	p2	(TC)6	12	69	80
Contig359	1	p3	(TCT)6	18	57	74
Contig367	1	p1	(A)10	10	1156	1165
Contig369	1	p3	(GAG)12	36	926	961
Contig370	1	p2	(CT)26	52	59	110
Contig378	1	p3	(GAT)5	15	736	750
Contig382	1	p2	(TC)6	12	57	68
Contig388	1	p2	(CT)7	14	56	69
Contig390	1	p2	(AT)12	24	576	599
Contig403	1	p1	(T)10	10	1308	1317
Contig407	1	p1	(T)12	12	913	924
Contig412	1	p1	(A)12	12	1727	1738
Contig414	1	p3	(TTC)5	15	17	31
Contig414	2	c	(T)10g(T)11	22	950	971
Contig414	3	p2	(AT)6	12	1178	1189
Contig416	1	p3	(TCT)5	15	1375	1389
Contig417	1	p1	(A)12	12	695	706
Contig424	1	p1	(T)12	12	402	413
Contig426	1	p2	(TA)6	12	1516	1527
Contig429	1	p2	(TA)6	12	38	49
Contig432	1	p1	(T)18	18	1100	1117
Contig433	1	p1	(T)13	13	992	1004

Contig436	1	p1	(T)13	13	300	312
Contig437	1	p1	(T)12	12	1762	1773
Contig441	1	p1	(T)12	12	404	415
Contig442	1	p2	(AG)15	30	1	30
Contig443	1	p2	(GA)11	22	3	24
Contig443	2	p1	(T)11	11	1611	1621
Contig456	1	p1	(A)15	15	270	284
Contig459	1	c	tctctctcgcggctagggtttt	86	3	88
Contig475	1	p3	(CTT)6	18	90	107
Contig476	1	p1	(T)10	10	1836	1845
Contig480	1	p2	(AG)7	14	17	30
Contig481	1	p3	(TTA)5	15	108	122
Contig482	1	p1	(A)12	12	1252	1263
Contig500	1	p2	(AG)8	16	39	54
Contig501	1	p2	(TC)11	22	233	254
Contig502	1	c	acaatcgtggaggcggtggc	60	396	455
Contig507	1	p2	(GT)7	14	166	179
Contig513	1	p1	(A)10	10	55	64
Contig532	1	p1	(A)12	12	9	20
Contig536	1	p3	(GAT)5	15	516	530
Contig542	1	p2	(TC)6	12	71	82
Contig543	1	p2	(GA)7	14	17	30
Contig545	1	c	tgtgaaaattaattaatggtt	51	868	918
Contig547	1	p1	(T)10	10	459	468
Contig548	1	p3	(AAC)5	15	189	203
Contig548	2	p1	(T)14	14	1113	1126
Contig552	1	p3	(TCT)5	15	96	110
Contig555	1	p3	(CTC)5	15	1	15
Contig565	1	p1	(T)13	13	911	923
Contig567	1	p2	(CT)7	14	1	14
Contig575	1	p2	(TC)6	12	217	228
Contig575	2	p1	(T)14	14	1141	1154
Contig592	1	p1	(A)10	10	1719	1728
Contig594	1	p3	(TTC)5	15	73	87
Contig594	2	p1	(T)14	14	337	350
Contig600	1	p1	(A)10	10	166	175
Contig600	2	p1	(A)10	10	332	341
Contig603	1	p1	(A)10	10	26	35
Contig606	1	p3	(CCT)5	15	127	141
Contig610	1	p1	(T)12	12	607	618
Contig615	1	p1	(A)13	13	27	39
Contig615	2	p1	(A)12	12	1110	1121
Contig616	1	p2	(TA)9	18	290	307
Contig627	1	p1	(A)10	10	1129	1138
Contig627	2	p2	(AG)9	18	1244	1261
Contig627	3	p1	(A)10	10	1540	1549
Contig628	1	p1	(A)11	11	57	67
Contig632	1	p1	(A)15	15	1067	1081
Contig639	1	p1	(A)12	12	114	125
Contig647	1	p1	(A)16	16	75	90

Contig651	1	p3	(TTG)5	15	242	256
Contig651	2	p3	(CAG)5	15	2166	2180
Contig657	1	p1	(T)10	10	282	291
Contig662	1	p1	(A)10	10	1885	1894
Contig663	1	p1	(T)11	11	2037	2047
Contig664	1	p2	(AT)7	14	22	35
Contig674	1	p2	(TA)6	12	827	838
Contig677	1	p1	(T)14	14	1450	1463
Contig680	1	p2	(TC)12	24	85	108
Contig681	1	p2	(CT)11	22	26	47
Contig681	2	p1	(A)11	11	1248	1258
Contig683	1	p1	(T)14	14	61	74
Contig688	1	p1	(T)10	10	1043	1052
Contig691	1	p2	(CT)6	12	218	229
Contig693	1	p3	(CAG)5	15	2324	2338
Contig693	2	p1	(T)14	14	3882	3895
Contig694	1	p2	(CT)6	12	46	57
Contig702	1	p2	(CT)22	44	13	56
Contig702	2	p3	(CGA)5	15	171	185
Contig702	3	p1	(A)10	10	1587	1596
Contig703	1	p3	(CTT)7	21	31	51
Contig704	1	p1	(T)10	10	381	390
Contig704	2	p3	(CAC)11	33	768	800
Contig708	1	p2	(AG)7	14	21	34
Contig709	1	p1	(A)11	11	1516	1526
Contig710	1	p1	(A)13	13	54	66
Contig715	1	p1	(A)15	15	64	78
Contig715	2	p3	(GAA)5	15	606	620
Contig718	1	p1	(A)13	13	102	114
Contig720	1	p1	(A)14	14	72	85
Contig722	1	p2	(CT)6	12	39	50
Contig722	2	p1	(C)10	10	154	163
Contig722	3	p1	(T)16	16	3646	3661
Contig724	1	c	agttgctcatcagcctgtgaa	116	544	659
Contig733	1	p3	(CAG)5	15	429	443
Contig734	1	p2	(TC)16	32	5	36
Contig738	1	p2	(TC)7	14	22	35
Contig749	1	c	ctttctcgggaaacaagcac	84	117	200
Contig759	1	p1	(T)10	10	1024	1033
Contig759	2	p1	(T)11	11	1168	1178
Contig768	1	p3	(ACC)7	21	145	165
Contig769	1	p1	(T)14	14	947	960
Contig770	1	p2	(TC)12	24	178	201
Contig772	1	p3	(GGC)6	18	673	690
Contig778	1	p2	(TA)9	18	1613	1630
Contig784	1	p3	(AAG)9	27	187	213
Contig785	1	p3	(ATC)9	27	932	958
Contig788	1	c	gaagtgttatgagtgtggtga	155	330	484
Contig792	1	p2	(TA)9	18	1546	1563
Contig798	1	p3	(AGG)8	24	1415	1438

Contig807	1	p1	(A)10	10	69	78
Contig814	1	p3	(CAC)5	15	62	76
Contig816	1	p2	(CT)14	28	42	69
Contig817	1	p3	(CAG)9	27	243	269
Contig817	2	p3	(CCA)8	24	555	578
Contig818	1	p3	(GAA)11	33	290	322
Contig818	2	p3	(CAG)6	18	1070	1087
Contig827	1	p1	(A)11	11	1762	1772
Contig828	1	c	)13tcgttccagcttgttt(C	54	15	68
Contig829	1	c	(CT)6(AT)6	24	1648	1671
Contig830	1	p2	(TC)6	12	257	268
Contig833	1	p6	(CATGGT)5	30	1368	1397
Contig847	1	p3	(GAA)10	30	139	168
Contig858	1	p3	(ATT)5	15	159	173
Contig861	1	p3	(CTC)5	15	1092	1106
Contig861	2	p3	(CAG)5	15	1366	1380
Contig864	1	p2	(TA)7	14	616	629
Contig875	1	p1	(A)11	11	916	926
Contig884	1	p1	(A)12	12	15	26
Contig891	1	c	ctaagaacgcgaagaacag	89	99	187
Contig893	1	p6	(CAGTCT)5	30	358	387
Contig894	1	p2	(TC)7	14	89	102
Contig894	2	p2	(TA)10	20	1605	1624
Contig896	1	c	(GA)6gt(GA)14	42	28	69
Contig897	1	p3	(AGA)5	15	34	48
Contig897	2	p1	(A)11	11	2426	2436
Contig898	1	p2	(TC)8	16	42	57
Contig898	2	p3	(GCT)5	15	634	648
Contig901	1	p1	(A)11	11	2228	2238
Contig901	2	p1	(T)10	10	2373	2382
Contig903	1	p3	(GAA)5	15	521	535
Contig909	1	p3	(GAT)5	15	1098	1112
Contig914	1	p1	(A)19	19	65	83
Contig915	1	p2	(TA)7	14	1459	1472
Contig921	1	p3	(GAT)5	15	555	569
Contig924	1	p1	(T)11	11	1446	1456
Contig927	1	p3	(AGC)5	15	238	252
Contig927	2	p2	(TA)7	14	1301	1314
Contig929	1	p3	(GAA)5	15	3343	3357
Contig933	1	p3	(CCT)5	15	268	282
Contig941	1	p2	(CT)9	18	45	62
Contig942	1	p3	(AAG)6	18	16	33
Contig943	1	p3	(AAC)5	15	185	199
Contig948	1	p3	(TGA)5	15	168	182
Contig949	1	p2	(CT)6	12	21	32
Contig953	1	p1	(A)11	11	112	122
Contig954	1	p1	(C)10	10	640	649
Contig961	1	p2	(AT)10	20	688	707
Contig965	1	p1	(A)12	12	876	887
Contig977	1	p1	(T)11	11	830	840

Contig980	1	c	ccaacatttgtcaacaggaa	80	148	227
Contig980	2	p3	(CAG)5	15	358	372
Contig980	3	c	(T)14(G)13	27	985	1011
Contig981	1	p1	(T)11	11	749	759
Contig984	1	p3	(GCA)5	15	152	166
Contig985	1	p2	(CT)7	14	1925	1938
Contig993	1	p1	(T)11	11	40	50
Contig997	1	p1	(A)11	11	289	299
Contig1001	1	p3	(CTC)8	24	9	32
Contig1029	1	c	tttagcagcgaagaattga	67	1605	1671
Contig1031	1	p3	(TCC)5	15	306	320
Contig1033	1	p2	(TC)6	12	19	30
Contig1042	1	p1	(T)14	14	739	752
Contig1043	1	p2	(CT)10	20	1	20
Contig1047	1	p2	(CT)8	16	95	110
Contig1052	1	p1	(A)11	11	2431	2441
Contig1053	1	p3	(TCT)7	21	191	211
Contig1054	1	p1	(A)11	11	39	49
Contig1061	1	p3	(ACC)9	27	1053	1079
Contig1067	1	p1	(T)10	10	190	199
Contig1069	1	p1	(T)11	11	121	131
Contig1069	2	p1	(A)18	18	1131	1148
Contig1073	1	p3	(AGA)5	15	512	526
Contig1078	1	p3	(TCT)10	30	70	99
Contig1084	1	p1	(A)16	16	2682	2697
Contig1085	1	p1	(T)10	10	1897	1906
Contig1088	1	p1	(A)10	10	2199	2208
Contig1094	1	p3	(GCT)7	21	263	283
Contig1097	1	p2	(TC)8	16	59	74
Contig1098	1	p2	(TC)7	14	59	72
Contig1107	1	p1	(A)10	10	101	110
Contig1109	1	p1	(T)15	15	935	949
Contig1117	1	c	ttttatttgggtttgtgtacag	82	263	344
Contig1118	1	p2	(AG)7	14	79	92
Contig1129	1	p2	(TA)9	18	846	863
Contig1130	1	p1	(T)10	10	672	681
Contig1131	1	p1	(T)14	14	1842	1855
Contig1136	1	p2	(CT)8	16	52	67
Contig1146	1	p3	(TAA)6	18	1282	1299
Contig1147	1	p1	(A)15	15	1609	1623
Contig1149	1	p1	(T)15	15	2194	2208
Contig1156	1	p1	(T)11	11	694	704
Contig1161	1	p3	(ATC)5	15	1931	1945
Contig1162	1	p1	(T)11	11	511	521
Contig1164	1	p2	(AT)6	12	3071	3082
Contig1175	1	p3	(TTA)5	15	2746	2760
Contig1188	1	p1	(T)10	10	943	952
Contig1199	1	p2	(CT)6	12	1	12
Contig1201	1	p2	(GA)11	22	136	157
Contig1205	1	p1	(T)13	13	1731	1743

Contig1211	1	p1	(A)10	10	5	14
Contig1213	1	p1	(T)10	10	1148	1157
Contig1214	1	p1	(T)10	10	1248	1257
Contig1221	1	p2	(CT)7	14	5	18
Contig1223	1	p2	(AG)8	16	144	159
Contig1230	1	p2	(CT)15	30	1	30
Contig1231	1	p1	(T)11	11	205	215
Contig1244	1	p1	(A)10	10	809	818
Contig1246	1	p1	(A)11	11	215	225
Contig1248	1	p1	(A)15	15	41	55
Contig1248	2	p3	(GAA)5	15	521	535
Contig1253	1	p2	(TA)7	14	525	538
Contig1256	1	c	(TCT)5gctt(CTG)5	34	76	109
Contig1259	1	p3	(GGC)5	15	42	56
Contig1268	1	p3	(CTT)5	15	281	295
Contig1276	1	p3	(AAG)5	15	120	134
Contig1281	1	p1	(A)14	14	1	14
Contig1281	2	p3	(TCA)6	18	561	578
Contig1285	1	p3	(GCA)5	15	403	417
Contig1285	2	p1	(T)14	14	1286	1299
Contig1286	1	c	aacagacatgctctgcaact	131	66	196
Contig1288	1	c	ttgaaaaggaggaagaatcc	54	62	115
Contig1290	1	p1	(T)13	13	1331	1343
Contig1292	1	p1	(A)10	10	2427	2436
Contig1294	1	p1	(T)15	15	1396	1410
Contig1295	1	p2	(TC)7	14	140	153
Contig1295	2	p3	(GCA)5	15	1057	1071
Contig1299	1	c	(TC)7g(CT)6	27	3539	3565
Contig1302	1	p2	(AG)6	12	1594	1605
Contig1306	1	p2	(TC)6	12	405	416
Contig1311	1	p3	(AAC)6	18	588	605
Contig1333	1	p1	(T)10	10	1083	1092
Contig1335	1	p3	(ATT)5	15	1298	1312
Contig1339	1	p3	(TGG)9	27	35	61
Contig1342	1	p1	(T)11	11	903	913
Contig1346	1	p1	(T)10	10	204	213
Contig1353	1	p1	(T)11	11	129	139
Contig1353	2	p1	(T)10	10	529	538
Contig1353	3	p1	(T)14	14	809	822
Contig1354	1	p2	(TA)7	14	1529	1542
Contig1357	1	p2	(AG)6	12	17	28
Contig1362	1	p2	(CT)12	24	22	45
Contig1362	2	p3	(AAG)8	24	1133	1156
Contig1362	3	c	ctatggatgaaaggctttgt	124	1340	1463
Contig1363	1	p3	(GCA)5	15	61	75
Contig1364	1	p1	(A)17	17	590	606
Contig1370	1	p2	(AG)11	22	67	88
Contig1375	1	p1	(A)12	12	70	81
Contig1375	2	p3	(AAG)6	18	340	357
Contig1384	1	p2	(TG)8	16	1248	1263

Contig1387	1	c	taaacagacagaaagttctt	88	10	97
Contig1388	1	p1	(T)10	10	1709	1718
Contig1389	1	p3	(CTT)7	21	253	273
Contig1389	2	p1	(T)10	10	1224	1233
Contig1391	1	p6	(CTCCTT)6	36	685	720
Contig1396	1	p3	(TCC)5	15	29	43
Contig1399	1	c	(T)11atttgg(A)10	27	1446	1472
Contig1400	1	c	(T)11atttgg(A)10	27	1426	1452
Contig1401	1	c	(T)11atttgg(A)10	27	1337	1363
Contig1404	1	p3	(TAT)6	18	328	345
Contig1416	1	p2	(TC)6	12	1	12
Contig1423	1	p4	(TTAA)5	20	1323	1342
Contig1423	2	p1	(T)10	10	1491	1500
Contig1427	1	p2	(AG)8	16	68	83
Contig1428	1	p1	(A)11	11	122	132
Contig1431	1	p1	(T)14	14	288	301
Contig1431	2	p2	(AG)6	12	737	748
Contig1439	1	p3	(TTA)6	18	2955	2972
Contig1441	1	p2	(TC)13	26	37	62
Contig1443	1	p1	(T)13	13	1719	1731
Contig1444	1	p3	(GGA)5	15	798	812
Contig1444	2	p3	(TTA)6	18	1639	1656
Contig1447	1	p2	(TC)6	12	72	83
Contig1448	1	p2	(TC)7	14	165	178
Contig1452	1	p1	(A)10	10	2171	2180
Contig1462	1	p2	(TC)13	26	33	58
Contig1467	1	p2	(AT)8	16	95	110
Contig1468	1	p2	(TA)8	16	3912	3927
Contig1469	1	p3	(CAG)5	15	276	290
Contig1469	2	p3	(ACC)5	15	800	814
Contig1471	1	p3	(TCT)5	15	315	329
Contig1474	1	p1	(A)13	13	1365	1377
Contig1476	1	p1	(T)11	11	1422	1432
Contig1478	1	p1	(T)10	10	2860	2869
Contig1479	1	p2	(TA)8	16	135	150
Contig1481	1	p3	(CTC)6	18	136	153
Contig1484	1	p1	(T)12	12	1041	1052
Contig1486	1	p3	(ACT)6	18	1	18
Contig1487	1	p1	(T)14	14	2162	2175
Contig1489	1	p1	(T)22	22	184	205
Contig1495	1	p1	(T)15	15	2572	2586
Contig1509	1	p1	(T)11	11	1334	1344
Contig1516	1	p1	(A)10	10	141	150
Contig1524	1	p2	(TC)6	12	66	77
Contig1529	1	p1	(T)10	10	837	846
Contig1535	1	p2	(TG)6	12	486	497
Contig1542	1	p3	(GCT)5	15	79	93
Contig1546	1	p2	(AG)11	22	1703	1724
Contig1552	1	p1	(T)11	11	3898	3908
Contig1557	1	p1	(A)10	10	68	77

Contig1561	1	c	atccctgagaactcctttcat	72	26	97
Contig1561	2	p3	(TTC)5	15	1090	1104
Contig1566	1	p1	(T)10	10	352	361
Contig1571	1	p2	(AC)7	14	12	25
Contig1573	1	p1	(T)11	11	885	895
Contig1576	1	p1	(T)11	11	1277	1287
Contig1577	1	c	gtgacttgtgctgctgttcctt	109	1	109
Contig1578	1	p1	(T)14	14	146	159
Contig1591	1	p3	(GAA)7	21	2	22
Contig1593	1	p3	(GCA)5	15	147	161
Contig1593	2	p1	(A)10	10	381	390
Contig1596	1	p3	(GAG)7	21	571	591
Contig1602	1	p2	(TC)9	18	115	132
Contig1605	1	p1	(T)10	10	1381	1390
Contig1606	1	p1	(T)10	10	281	290
Contig1607	1	p1	(A)11	11	1778	1788
Contig1610	1	p1	(T)16	16	1264	1279
Contig1612	1	p1	(A)13	13	40	52
Contig1612	2	p1	(A)15	15	314	328
Contig1612	3	p1	(T)14	14	469	482
Contig1612	4	p1	(A)14	14	1899	1912
Contig1612	5	p3	(AAT)5	15	2142	2156
Contig1613	1	p1	(A)13	13	40	52
Contig1613	2	p1	(T)14	14	463	476
Contig1613	3	p1	(A)12	12	1781	1792
Contig1613	4	p3	(AAT)5	15	2022	2036
Contig1614	1	p2	(TC)7	14	21	34
Contig1621	1	c	cctattgcagaggcatgc(G	114	77	190
Contig1621	2	c	AG)5aacagtaatcca(CA	54	1042	1095
Contig1622	1	p3	(TAA)5	15	3173	3187
Contig1626	1	c	ctctgcaaatccaattattct	82	383	464
Contig1626	2	p1	(T)16	16	2144	2159
Contig1631	1	c	gctcatatatatagagag	57	1594	1650
Contig1634	1	p2	(AC)6	12	168	179
Contig1638	1	p3	(ATG)7	21	953	973
Contig1644	1	p2	(CT)11	22	1259	1280
Contig1651	1	p1	(T)12	12	1540	1551
Contig1661	1	p5	(ACGAC)5	25	8	32
Contig1661	2	p1	(T)11	11	1786	1796
Contig1665	1	p3	(TCT)8	24	202	225
Contig1672	1	p1	(T)12	12	2093	2104
Contig1694	1	p2	(CT)10	20	1337	1356
Contig1704	1	p3	(CCA)7	21	164	184
Contig1710	1	p1	(T)10	10	1753	1762
Contig1711	1	p1	(T)10	10	1711	1720
Contig1715	1	p1	(A)13	13	2586	2598
Contig1717	1	p3	(TCT)7	21	46	66
Contig1719	1	c	10ccttttttcttttc(T)10(A	45	3064	3108
Contig1725	1	p2	(CT)15	30	25	54
Contig1725	2	p2	(CT)11	22	215	236

Contig1727	1	p2	(TC)13	26	32	57
Contig1728	1	p2	(TC)13	26	32	57
Contig1731	1	p3	(CTC)5	15	124	138
Contig1735	1	p1	(A)10	10	148	157
Contig1735	2	p2	(GT)7	14	2102	2115
Contig1737	1	p1	(A)11	11	226	236
Contig1738	1	p1	(T)11	11	80	90
Contig1740	1	p3	(GAA)5	15	1195	1209
Contig1747	1	p1	(T)17	17	2477	2493
Contig1748	1	p1	(T)29	29	1301	1329
Contig1752	1	p1	(A)11	11	1803	1813
Contig1759	1	p2	(TC)9	18	37	54
Contig1760	1	p1	(T)12	12	308	319
Contig1761	1	p3	(CTT)8	24	255	278
Contig1762	1	p1	(T)10	10	104	113
Contig1769	1	p3	(CCT)5	15	459	473
Contig1770	1	p3	(GCA)5	15	3708	3722
Contig1772	1	p3	(TGA)5	15	1456	1470
Contig1773	1	p3	(TTC)6	18	28	45
Contig1775	1	p1	(T)10	10	2108	2117
Contig1781	1	p1	(T)10	10	2520	2529
Contig1788	1	p1	(A)10	10	2557	2566
Contig1792	1	p1	(T)10	10	1445	1454
Contig1800	1	p1	(A)10	10	973	982
Contig1807	1	p1	(A)12	12	2598	2609
Contig1808	1	p1	(T)10	10	2093	2102
Contig1810	1	p1	(T)15	15	1876	1890
Contig1816	1	p3	(GAT)5	15	265	279
Contig1818	1	p2	(AT)12	24	1839	1862
Contig1819	1	p2	(TA)18	36	4	39
Contig1831	1	p1	(A)10	10	27	36
Contig1836	1	p1	(A)15	15	1245	1259
Contig1840	1	p2	(AG)12	24	1180	1203
Contig1848	1	p2	(CT)6	12	44	55
Contig1849	1	p2	(CT)6	12	44	55
Contig1852	1	p1	(A)10	10	147	156
Contig1861	1	p1	(T)10	10	622	631
Contig1863	1	p1	(T)15	15	1712	1726
Contig1883	1	p1	(C)11	11	1844	1854
Contig1884	1	p2	(TA)6	12	2074	2085
Contig1884	2	p1	(T)12	12	2221	2232
Contig1890	1	p3	(CAT)5	15	585	599
Contig1890	2	p1	(A)10	10	1000	1009
Contig1908	1	p1	(T)11	11	318	328
Contig1908	2	p2	(TC)10	20	2414	2433
Contig1917	1	p3	(GAA)10	30	138	167
Contig1922	1	c	(GA)7a(AG)8	31	2253	2283
Contig1923	1	p2	(TC)6	12	26	37
Contig1923	2	p3	(ATA)5	15	356	370
Contig1923	3	p1	(T)11	11	1859	1869

Contig1924	1	p3	(GAA)5	15	78	92
Contig1924	2	p1	(T)10	10	1832	1841
Contig1930	1	p1	(T)14	14	1455	1468
Contig1939	1	p2	(CT)6	12	27	38
Contig1952	1	p1	(A)10	10	190	199
Contig1957	1	p1	(G)12	12	1837	1848
Contig1958	1	p1	(T)10	10	93	102
Contig1959	1	p3	(TGC)7	21	1245	1265
Contig1959	2	c	tgatttgaaatttggttagga	121	1555	1675
Contig1960	1	p3	(AGA)5	15	281	295
Contig1969	1	p2	(CT)20	40	44	83
Contig1978	1	c	atgggagaaaaaataagtg	44	1040	1083
Contig1979	1	p1	(T)12	12	900	911
Contig1981	1	p3	(CTT)10	30	287	316
Contig1981	2	p1	(T)13	13	434	446
Contig1982	1	p1	(T)13	13	1015	1027
Contig1984	1	p1	(T)16	16	3117	3132
Contig1986	1	p2	(GA)7	14	8	21
Contig1986	2	p2	(AT)6	12	1793	1804
Contig1987	1	p2	(TC)9	18	77	94
Contig1988	1	p3	(ATC)6	18	1066	1083
Contig2013	1	p4	(TTTC)6	24	154	177
Contig2014	1	p1	(A)10	10	678	687
Contig2016	1	p1	(T)10	10	1128	1137
Contig2022	1	p1	(A)12	12	19	30
Contig2029	1	p2	(TC)6	12	239	250
Contig2048	1	p1	(T)14	14	1271	1284
Contig2049	1	p2	(CT)7	14	17	30
Contig2049	2	p3	(ACC)5	15	485	499
Contig2051	1	p3	(GCA)7	21	151	171
Contig2056	1	p1	(T)13	13	1008	1020
Contig2062	1	p1	(T)10	10	874	883
Contig2063	1	p3	(GAA)7	21	684	704
Contig2063	2	p2	(TG)11	22	959	980
Contig2064	1	p3	(GAT)5	15	520	534
Contig2065	1	p3	(GAA)10	30	472	501
Contig2066	1	p3	(GCA)8	24	1468	1491
Contig2067	1	p3	(TTA)5	15	134	148
Contig2073	1	p1	(T)16	16	1378	1393
Contig2079	1	p3	(CTG)5	15	60	74

APPENDIX XIII							
List of NonsynonymousSNP coding data identified by QualitySNP							
Contig no:	position	SNP	length	normal sequence	sequence with base change	transcribed protein	
260	388	TC	10	CACCAGAATTTATCATCAAGC	CACCAGAATTCATCATCAAGC	HQNLSSS	HQNSSSS
344	509	AT	10	AAATCAGCTTATGCATTGTGT	AAATCAGCTTTTGCATTGTGT	KSAYALC	KSAFALC
385	683	GC	10	AACAGTGAGAGCAAACAAGAG	AACAGTGAGACCAAACAAGAG	NSESKQE	NSETKQE
401	630	GT	11	TTGCGCAAGCAGTACGGACCT	TTGCGCAAGCATTACGGACCT	LRKQYGP	LRKHYGP
401	833	GC	10	CGGAATCCAAGGAAAAGGCTA	CGGAATCCAACGAAAAGGCTA	RNPRKRL	RNPTKRL
401	836	GA	10	AATCCAACGAGAAGGCTATCA	AATCCAACGAAAAGGCTATCA	NPTRRLS	NPTKRLS
468	1143	AG	9	GCTGCATTCAATATGCCACCC	GCTGCATTCGATATGCCACCC	AAFNMPP	AAFDMPP
732	82	CA	11	GTTCAATCTCACCCCAGAAGC	GTTCAATCTCAACCCAGAAGC	VQSHPRS	VQSQPRS
896	1495	CA	10	GTGCTATATACGCACCCAGCA	GTGCTATATAAGCACCCAGCA	VLYTHPA	VLYKHPA
1043	635	CT	10	TCTCAAACAACGATTTATGTG	TCTCAAACAATGATTTATGTG	SQTTIYV	SQTMIYV
1053	1044	TG	11	TGTCAGGGAGATTATGTGGTG	TGTCAGGGAGAGTATGTGGTG	CQGDYVV	CQGEYVV
1073	76	CT	10	CGTGAACAACCTCCCTCCATC	CGTGAACAACTTCCCTCCATC	REQPPSI	REQLPSI
1073	79	TC	10	GAACAACCTCTCTCCATCCTC	GAACAACCTCCCTCCATCCTC	EQPLSIL	EQPPSIL
1073	126	TC	9	TTTGGCTCTTTTTCTCCCTTG	TTTGGCTCTCTTTCTCCCTTG	FGSFSPL	FGSLSPL
1228	2528	AG	10	TACAGCATCGAACTTCCAAGC	TACAGCATCGGACTTCCAAGC	YSIELPS	YSIGLPS
1238	415	AG	9	TTTCTCGTGATTTTGCTTTTG	TTTCTCGTGGTTTTGCTTTTG	FLVILLL	FLVVLLL
1889	668	AG	10	ACACCCGGCCAGGAATTTACT	ACACCCGGCCGGGAATTTACT	TPGQEFT	TPGREFT
1889	685	AG	9	ACTTTTACAATTCGTAGGGGA	ACTTTTACAGTTCGTAGGGGA	TFTIRRG	TFTVRRG
1889	881	GA	10	CTAAATGTTAGAGGAAAAAGC	CTAAATGTTAAAGGAAAAAGC	LNVRGKS	LNVKGKS
1930	1379	AC	9	GAGGTTAGTAACCTTACAGCC	GAGGTTAGTCACCTTACAGCC	EVSNLTA	EVSHLTA
2023	574	GT	9	AGCTACACTGTGGCTTATGGA	AGCTACACTTTGGCTTATGGA	SYTVAYG	SYTLAYG
2023	602	CG	10	CCAGAACCTACTTGTCCTTGT	CCAGAACCTAGTTGTCCTTGT	PEPTCPC	PEPSCPC
2055	1540	CT	10	AAAAAATATGCTGAGGTTCTT	AAAAAATATGTTGAGGTTCTT	KKYAEVL	KKYVEVL
2055	1560	GC	9	AGACTGATAGGGAGACTTACG	AGACTGATACGGAGACTTACG	RLIGRLT	RLIRRLT
2055	1563	AG	9	CTGATAGGGAGACTTACGTTG	CTGATAGGGGGACTTACGTTG	LIGRLTL	LIGGLTL
2055	1617	GC	9	CAAGACTCCGAGCTAGACCAA	CAAGACTCCCAGCTAGACCAA	QDSELDQ	QDSQLDQ
2055	1680	GA	9	AGTCTGGTTGCTTTAGCACCA	AGTCTGGTTACTTTAGCACCA	SLVALAP	SLVTLAP
2055	1725	GA	9	ATCACGTTGGAAGTGTTGAAA	ATCACGTTGAAAGTGTTGAAA	ITLEVLK	ITLKVLK
2055	1987	TA	10	GTAACTGTGATGCAATGCCCC	GTAACTGTGAAGCAATGCCCC	VTVMQCP	VTVKQCP
2064	625	GC	9	TCAAATCAGGCTTCAGTTACT	TCAAATCAGCCTTCAGTTACT	SNQASVT	SNQPSVT

APPENDIX XIV						
List of SynonymousSNP coding data identified by QualitySNP						
Contig no:	position	SNP	length	normal sequence	sequence with base change	transcribed protein
361	358	GA	11	GCTAACCTGAGGCGCGCTGCT	GCTAACCTGAGACGCGCTGCT	ANLRRAA
361	454	CG	11	AGGCAGTTTCTCGGGCTGAGG	AGGCAGTTTCTGGGGCTGAGG	RQFLGLR
361	1053	AT	11	TACGGTTCGGCAGGCTATGCT	TACGGTTCGGCTGGCTATGCT	YGSAGYA
361	1189	TC	11	TCCATGGTGTCTACTGTTGCT	TCCATGGTGTCCACTGTTGCT	SMVSTVA
401	591	CA	11	GATGTTGTTGGCAGTCCATAC	GATGTTGTTGGAAGTCCATAC	DVVGSPY
401	609	AG	11	TACTATGTCGCACCAGAGGTG	TACTATGTCGCGCCAGAGGTG	YYVAPEV
401	618	GA	11	GCACCAGAGGTGTTGCGCAAG	GCACCAGAGGTATTGCGCAAG	APEVLRK
401	633	CT	11	CGCAAGCAGTACGGACCTGAA	CGCAAGCAGTATGGACCTGAA	RKQYGPE
401	678	TC	11	ATTTTGTATATTTTATTATCT	ATTTTGTATATCTTATTATCT	ILYILLS
401	699	AT	11	GGAGTGCCACCATTTTGGGCA	GGAGTGCCACCTTTTTGGGCA	GVPPFWA
401	837	GA	11	AATCCAACGAAGAGGCTATCA	AATCCAACGAAAAGGCTATCA	NPTKRLS
463	141	CT	11	GGAAAGTCGACCACTACTGGT	GGAAAGTCGACTACTACTGGT	GKSTTTG
468	1115	AT	11	ATTTCTACAGGAGCCTTCCTT	ATTTCTACAGGTGCCTTCCTT	ISTGAFL
567	427	CT	11	CCAAAGAAGACCGGCACCTCA	CCAAAGAAGACTGGCACCTCA	PKKTGTS
896	1490	AT	11	GAAAATGTGCTATATACGCAC	GAAAATGTGCTTTATACGCAC	ENVLYTH
899	1299	TC	11	CCCGAGCTTGTTAACAAGCTG	CCCGAGCTTGTCAACAAGCTG	PELVNKL
1136	285	CT	11	AAAATCAGAACCGTGGAGCTG	AAAATCAGAACTGTGGAGCTG	KIRTVEL
1228	2604	GA	11	AAGTCATTCACGTGTACTTTA	AAGTCATTCACATGTACTTTA	KSFTCTL
1233	636	TC	11	GTTTATAAGATTGAAGCTGAA	GTTTATAAGATCGAAGCTGAA	VYKIEAE
1889	608	CT	11	ATGCTTGACACCAAGGGTCCT	ATGCTTGACACTAAGGGTCCT	MLDTKGP
1889	813	AT	11	AAGTCCAAGACAGATGACTCT	AAGTCCAAGACTGATGACTCT	KSKTDDS
1889	912	TG	11	CCTTCCATCACTGAAAAGGAC	CCTTCCATCACGGAAAAGGAC	PSITEKD
2023	369	AG	11	GCTATGTTGTCACGCTCTGCG	GCTATGTTGTCGCGCTCTGCG	AMLSRSA
2023	378	GT	11	TCACGCTCTGCGGCAGGAATA	TCACGCTCTGCTGCAGGAATA	SRSAAGI
2055	1724	GA	11	CTAATCACGTTGGAAGTGTTG	CTAATCACGTTAGAAGTGTTG	LITLEVL
2055	1757	GA	11	TTACTTAGTCTGGTAACATCT	TTACTTAGTCTAGTAACATCT	LLSLVTS

APPENDIX XV

ClustalX alignment of SNP896 with MNga

```
CLUSTAL 2.1 multiple sequence alignment


Contig896       GATGGCTCGCCATTTGACAGATCGAGGGAGAGAGAGAGAGTGAGAGAGAGAGAGAGAGAG 60
2R              ------------------------------------------------------------


Contig896       AGAGAGAGAGGAGTGCTGATTATGGCGAGAGACATCGATGACCTACCAAAGCTCGAAGCT 120
2R              -------------TAATSGTTCTR----------TCAACKTCCTTCAYTA---------- 27
                             *..*..**.*           **.*  :***:*. :*

Contig896       AACCATATGGCCTTGACGCCGCTGTGGTTCCTCGAGAGAGCAGCTACGGTGCACCCCACC 180
2R              --CGATTTYKYCAYGTY----CTG-GYTTCAATGTSACAAAAGCACARGGGGACTCYCCC 80
                  * **:*   *: *:     *** * ***.: *:.* *..***:.. * * ** * .**

Contig896       AGAACAGCCGTTGTCCATGGATCAGTTCACTATACGTGGCAGCAGACCTACCAACGGTGC 240
2R              ------------------------------------------------------CAYYG---- 85
                                                                     **  *

Contig896       CGTCGATTGGCCTCTGCTCTTTCCAAGCGCAACATCGGCGCCGGAAGCACGGTAGCAACA 300
2R              -------------CTSATCTKSYCTR--GCYACMACMGAYAC------------------ 112
                             **..***.: *:   ** ** :* *. .*

Contig896       ATTGCTCCAAATGTCCCAGCCATGTATGAAGCTCATTTTGGAATTCCAATGGCTGGCGCA 360
2R              ----YTCATAYACTGCTGGRTGYGTRTAYAGYAYATTTTCTACYTCCARG---------- 158
                     **.:* : * * .*  . ** *. ** : *****  *. ****

Contig896       GTGTTAAATACTGTCAATATTCGTCTAAACGCATCAACCATCGCTTTCCTGCTGGGCCAT 420
2R              ----------------------------------CTACTWATGTTTTCWC---------- 174
                                                 *:**  : * ****

Contig896       TCAAAATCTGCAATTGTGATGGTGGACCAAGAGTTCTTTCCCTTAGCGGAGAATGCTTTG 480
2R              -------CTCCWGARATGATAATG-TCCTTGS-WTCTRTCCTTRATYT------------ 213
                      ** * .: .****..** :**::*:  *** *** * *

Contig896       AAAATTTTAGCAGAGAAAGATAGCCATTATAACCCCCCATTGTTGATCGTTATAGCTGAT 540
2R              ------------------------CTATATASCYRTCTG---------------GATGCT 234
                                       *::****:*   * .               *.**.*

Contig896       GAAAGTTGTGATCCCAAGTCGCTCAAGGACGCTTTGGGAAAAGGGGCCATTGAATATGAC 600
2R              -----------TCACAG-----------------------------------------C 241
                           **.**.                                          *

Contig896       AAGTTCCTGGAAAGTGGTGACCCCGACTTTGCTTGGAAACCACCAGAAGATGAGTGGCAG 660
2R              RAGATCW--------------CCAGAAT-------GAAACCACCC--------------- 265
                 **:**               **.**.*       *********.

Contig896       AGTATTGCTTTGGGTTATACTTCGGGCACAACAGCCAGTCCCAAGGGGGTAGTGCTCAGC 720
2R              --ATTTGCAAARG---CTTCTTYG------------------------------------ 284
                  ::****::: *   .*:*** *

Contig896       CATCGAGGGGCATATCTAATGGCTTTAAGTTGTGCTATGATATGGGGTCTCAATGAAGGA 780
2R              ------------------TTRGCTTYYGGAT----------------TCTTYAAGTAG-- 308
                                 :* ****  .*:*                 ***  *:*:**

Contig896       GCTATTTATCTGTGGACTTTGCCCATGTTCCATTGCAATGGTTGGTGTTACCCCTGGTCG 840
2R              ---------------------CCCTT----CATYACAAS-------ATTKCCYCKCATCA 336
                                     ***:*    *** .***:       .** ** *. .**.

Contig896       CTTGCGGCTCTTTGCGGGACAAATATCTGCTTGAGACAGGTCACAGCCAAGGCAGTCTAT 900
2R              CT--------------------ATTTCT-------------------------------- 344
                **                     **:***
```

```
Contig896    TCGGCCATTGCCAACCAAGGTGTGACTCACTTCTGCGCTGCACCTGTGGTGCTCAACACC 960
2R           ------------------------------------------------------------CCC 347
                                                                        .**

Contig896    ATAGTAAATGCTCCGAACGAAGAGACTATCCTTCCCTTACCTCGCGTTGTCCATGTAAAC 1020
2R           ATGGTSKTT------------------------------------------------------ 356
             **.**: :*

Contig896    ACAGCTGGTGCTGCCCCACCCCCCTCTGTTCTCTTTGCAATGTCAGAGAAAGGCTTCCGG 1080
2R           ---------------------CCATCTG---------------CAGGKACAGGTTT---- 376
                                 **.****               ***. *.*** **

Contig896    GTTACCCATACATATGGGCTCTCAGAAACTTACGGTCCATCCACTGTGTGTGCATGGAAG 1140
2R           ------CATAGTTTTAG--KGTCAACWACWTCTAGRCCYTCCAAG--------------- 413
                   **** :*:*.*  . ***.. ** *. .* ** ****.

Contig896    CCTGAGTGGGACTCACTACCTCCCATCAAACAAGCCCGCCTCAACGCACGCCAAGGCGTG 1200
2R           ----------------------------------CCWATRTATCGMACKCCYTGGCGTG 438
                                               ** .   *:** ** ** :******

Contig896    CGATATATAGGCTTGGAGGGCCTAGAAGTAGTTGACACTAAAACTATGAAACCTGTACCT 1260
2R           CRTTKAG-------------------------------------------------------- 445
             * :* :.

Contig896    GCAGATGGAAAGACCATGGGAGAAATAGTGATGCGAGGAAATCTTGTAATGAAGGGCTAC 1320
2R           --------------------------------ACGGG-----CTTGTATK---------- 458
                                            .**.*     ******:.

Contig896    TTAAAGAATCCGGAAGCTAACAAAGAAGCCTTTGCAAATGGGTGGTTTCATTCTGGTGAT 1380
2R           -------------------------------------ATGGGAGG-----YWCKG--WAT 474
                                                   *****:**       *.*   **

Contig896    CTCGCTGTGAAGCATCCAGATGGGTATATAGAAATCAAGGATAGAAGCAAGGACATTAGC 1440
2R           CCCACT--CAGGCTTCCA--------------------------------------------- 490
             * *.**   *.**:****

Contig896    ATTTCAGGAGGTGAAAACATTAGTAGCTTGGAAGTAGAAAATGTGCTATATACGCACCCA 1500
2R           ------------------------------------------------------TGCACCAC 498
                                                                   *****..

Contig896    GCAGTGTATGAAGTATCTGTGGTAGCCAGAGAAGATGAGCGATGGGGAGAGTCCCCCTGT 1560
2R           ACAGTGA-------------------------------------------------------- 505
             .*****:

Contig896    GCTTTTGTCACATTGAAACCAGGCATGGAGAAATCTAGTGAAGGAAGTTTGGCAGAAGAT 1620
2R           ------------------------------------------------------------

Contig896    ATAATAAAGTTTTGTCGGTCGAAAATGCCTGCTTACTGGGTTCCAAAATCTGTTGTATTT 1680
2R           ------------------------------------------------------------

Contig896    GGACCATTGCCAAAAACTGCTACTGGGAAGATTCAAAAGCATGTGTTAAGGGCCAAGGCA 1740
2R           ------------------------------------------------------------

Contig896    AAGGAGATGGGACCTGTCAAAAAGAGCAGGTTATAGAAAATAGTGTATTCTGATGGCCTG 1800
2R           ------------------------------------------------------------

Contig896    AATAAGGATAACTCCTTTTGAATAGCCAATGTGGTATGGGTTTAGTTCCTCACAAGGCTT 1860
2R           ------------------------------------------------------------
```

```
Contig896      CTGCAGAAGTAGGATTATCTATTATTGTTCGCTTATTTCTGTTGTAAATTAAGCCATTAA 1920
2R             ------------------------------------------------------------

Contig896      TATGGATTTCTCTATTTTATGTTGAAATAAAGGTAGTTTATAATACGT 1968
2R             ------------------------------------------------
```

ABSTRACT

MOLECULAR MARKER DEVELOPMENT FOR CASSAVA MOSAIC DISEASE RESISTANCE USING BIOINFORMATICS TOOLS

AMBU VIJAYAN

(2010-09-105)

Abstract of the thesis submitted in partial fulfilment of the requirement for the degree of

MASTER OF SCIENCE (INTEGRATED) INBIOTECHNOLOGY

Faculty of Agriculture
Kerala Agricultural University, Thrissur

B.Sc.-M.Sc. (INTEGRATED) BIOTECHNOLOGY
DEPARTMENT OF PLANT BIOTECHNOLOGY

COLLEGE OF AGRICULTURE

VELLAYANI, THIRUVANANTHAPURAM-695 522

KERALA, INDIA

2015

ABSTRACT

The study entitled "Molecular marker development for cassava mosaic disease resistance using bioinformatics tools" was conducted at ICAR-CTCRI, Sreekariyam, Thiruvananthapuram during October 2104 to October 2015. The objectives of the study included development and evaluation of various SNP and SSR prediction pipelines, computational prediction and characterization of SNP and SSR in cassava, verification of SNP and SSR markers for cassava mosaic disease (CMD) resistant and susceptible breeding lines. The preliminary data set for the identification of SSR and SNP markers was obtained from the EST section of NCBI and the cassava transcript sequences from the Phytozome. A total of 120461 sequences was classified into 20 cultivars. The dataset was reduced to 14336 sequences after several pre-processing and screening steps. The resulting sequences were assembled and aligned using CAP3 and 2088 contigs were obtained. SNPs and SSRs were predicted from these datasets using respective prediction tools.

The SNP prediction tools such as QualitySNP and AutoSNP were compared for their performance. Analysis was performed to identify the tool with the ability to annotate and identify more viable nonsynonymous and synonymous SNPs.

The SSR prediction tools such as MISA and SSRIT was compared for their performance. Analysis was performed to identify the tool having the ability to predict more viable SSRs and the ability to classify them as mono, di, tri, tetra, penta, hexa and poly SSRs.

Using QualitySNP, thirty nonsynonymous SNPs and twenty-six synonymous SNPs were identified. Using MISA, n 217 mono SSRs, 132 di SSRs, 139 tri SSRs, 3 tetra SSRs, 1 penta SSRs, 3 hexa SSRs and 42 complex SSRs were identified. Five sequences from identified SNPs and SSRs which have high hit percentage were selected for validation and primer designing for CMD resistant genes. These primers were validated using 5 resistant and 5 susceptible cassava varieties. Among

the 10 primers after validation in wet lab, one SNP (SNP896) and one SSR (SSR 2063) primer was able to clearly differentiate between the resistant and susceptible varieties which can be used as potential markers in the breeding program for screening CMD resistance in cassava.