

**PREDICTION OF SSR AND SNP MARKERS FOR  
ANTHRACNOSE RESISTANCE IN YAM USING  
BIOINFORMATICS TOOLS AND THEIR VALIDATION**

By

**SAHLA K.**  
(2013-09-102)

**THESIS**

**Submitted in partial fulfilment of the  
requirement for the degree of**

**B. Sc. - M. Sc. (INTEGRATED) BIOTECHNOLOGY**

**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**

**2018**

## DECLARATION

I hereby declare that this thesis entitled **“Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation”** 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.

Place: Vellayni

Date: 07.12.2018



SAHLA K.

(2013-09-102)



# भा.कृ.अनु.प- केंद्रीय कन्द फसल अनुसंधान संस्थान

(भारतीय कृषि अनुसंधान परिषद, कृषि और किसान कल्याण मंत्रालय, भारत सरकार)

श्रीकार्यम, तिरुवनन्तपुरम-695 017, केरल, भारत



## ICAR- CENTRAL TUBER CROPS RESEARCH INSTITUTE

(Indian Council of Agriculture Research, Ministry of Agriculture and Farmers Welfare, Govt. of India)  
Sreekariyam, Thiruvananthapuram-695 017, Kerala, India



### CERTIFICATE

Certified that this thesis entitled “**Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation**” is a record of research work done independently by Ms. Sahla K. (2013-09-102) 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 her.

**Dr. J. Sreekumar**

(Chairperson, Advisory Committee)  
Principal Scientist (Agrl. Statistics),  
Section of Extension and Social Sciences,  
ICAR-CTCRI, Sreekariyam,  
Thiruvananthapuram- 695 017

Place: Sreekariyam

Date: 07/12/2018

डॉ. जे. श्रीकृष्णमूर्ति (Dr. J. SREEKUMAR)  
प्रधान ऐकानिक (कृषि सांख्यिकी)  
Principal Scientist (Agricultural Statistics)  
एक्सटेंशन और सामाजिक विज्ञान अनुभाग  
Section of Extension and Social Sciences  
भा.कृ.अनु.प- केंद्रीय कन्द फसल अनुसंधान संस्थान  
ICAR-CTCRI, श्रीकार्यम, तिरुवनन्तपुरम-695 017  
तिरुवनन्तपुरम, केरल, भारत

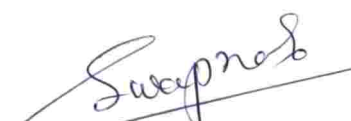
## CERTIFICATE

We, the undersigned members of the advisory committee of Ms. Sahla K. (2013-09-102), a candidate for the degree of B. Sc. – M. Sc. (Integrated) Biotechnology, agree that the thesis entitled “**Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation**” may be submitted by Ms. Sahla K. in partial fulfillment of the requirement for the degree.



**Dr. J. Sreekumar**

(Chairperson, Advisory Committee)  
Principal Scientist (Agrl. Statistics),  
Section of Extension and Social  
Sciences, ICAR-CTCRI. Sreekariyam,  
Thiruvananthapuram- 695 017



**Dr. Swapna Alex**

(Member, Advisory Committee)  
Professor & Head  
Department of Plant Biotechnology  
College of Agriculture, Vellayani  
Thiruvananthapuram – 695 522



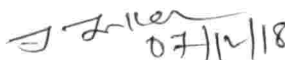
**Dr. M. N. Sheela**

(Member, Advisory Committee)  
Principal Scientist & Head  
Division of Crop improvement  
ICAR-CTCRI. Sreekariyam  
Thiruvananthapuram - 695 017



**Dr. K. B. Soni**

(Member, Advisory Committee)  
Professor & Course Director  
B. Sc. – M. Sc. (Integrated) Biotechnology  
Department of Plant Biotechnology  
College of Agriculture, Vellayani  
Thiruvananthapuram – 695 522



**Dr. J. Jayasankar**

(External Examiner)  
Principal Scientist  
ICAR- CMFRI  
Kochi – 682018, Kerala

## ACKNOWLEDGEMENT

It is with my heartfelt feelings, I wish to express my deep sense of gratitude and sincere thanks to my beloved advisor Dr. J. Sreekumar, Principal Scientist, ICAR-Central Tuber Crops Research Institute, for his valuable guidance, patience, and encouragement. In addition to his support for completing the project, his advice was valuable for me in every aspect.

My special thanks to Dr. Archana Mukherjee (Director, ICAR - CTCRI) and to Dr. Anilkumar A. (Dean, College Of Agriculture, Vellayani) for allowing me to do my project work and also for their support.

I take immense pleasure to express my deep sense of gratitude to Dr. M. N. Sheela, Dr. C. Mohan, Dr. Swapna Alex and Dr. Rajmohan not only for their insightful comments and motivation, but also for their valuable counselling and constructive suggestions that were much helpful throughout my research progress. My sincere thanks also goes to Dr. K. B. Soni, our Course Director for her encouragement and help rendered during the course of study. I would also like to put on record my sincere thanks and gratitude to Dr. Sheela Immanuel, Head, Section of Extension and Social Sciences. ICAR-CTCRI for permitting me and extending all facilities to complete my work.

My sincere gratitude towards Mr. Ambu Vijayan, who helped me with the technical aspects of my work using various bioinformatics tools and software. My heartfelt thanks are also to Mr. Prakash Krishnan, and Mr. Abhilash. My special thanks to my friends Athul V.S., Reshma Bhasker T., Aswathy M.B., Rekha Jayaraj and Priya for helping me in completing this work. My acknowledgement would be lacking if I don't mention my gratitude to my beloved friends Bimal Thomas, Sabarinath V. B., Arya R.S., Anjitha Nair U.M., Achuth P. Jayaraj, Aryalekshmi A.S., Jithu M.J., and all other classmates for their invaluable care, constant support, motivation and selfless help.

I wish to express my deep gratitude to all teaching non teaching staff members of ICAR-CTCRI, my seniors, juniors and teachers in college for their

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

I owe this achievement to my family who always stood along my side in all my happiness, difficulties and sorrow with their love, and prayers. Finally, I humbly thank the Almighty for showering his blessings, and bestowing the wisdom, perseverance and physical ability to accomplish this work.

**TABLE OF CONTENTS**

<b>Sl.No</b>	<b>Chapters</b>	<b>Page No.</b>
<b>1</b>	<b>INTRODUCTION</b>	<b>1-3</b>
<b>2</b>	<b>REVIEW OF LITERATURE</b>	<b>4-19</b>
<b>3</b>	<b>MATERIALS AND METHOD</b>	<b>20-36</b>
<b>4</b>	<b>RESULTS</b>	<b>37-47</b>
<b>5</b>	<b>DISCUSSION</b>	<b>48-52</b>
<b>6</b>	<b>SUMMARY</b>	<b>53-54</b>
<b>7</b>	<b>REFERENCES</b>	<b>55-67</b>
<b>8</b>	<b>APPENDICES</b>	<b>68-81</b>
<b>9</b>	<b>ABSTRACT</b>	<b>82-83</b>

## LIST OF TABLES

<b>Table No.</b>	<b>Title</b>	<b>Page No.</b>
1	List of accessions of greater yam used for the study	31
2	The reaction mixture used for SSR	34
3	Result of Pre-processing of primary dataset using SeqClean	38
4	Result of Screening of primary dataset against resistant gene database	38
5	Assembling of Sequences using CAP3	39
6	Result of identified SNPs using QualitySNP	40
7	Result of identified SNPs using AutoSNP	41
8	Comparative evaluation of SNP prediction tools	41
9	Distribution of different repeat classes in MISA	42
10	Distribution of different repeat classes in SSRIT	43
11	Comparative Evaluation of SSR Prediction Tools	43
12	SNP Primers	45
13	SSR Primer	46
14	Annealing Temperature (Ta) of designed SSR Primers	46
15	Annealing Temperature (Ta) of designed SNP Primers	47
16	Result of quantification of DNA	47



**LIST OF FIGURES**

<b>Figure No.</b>	<b>Title</b>	<b>Between pages</b>
Figure 1	Workflow	20-21
Figure 2	User window of Uniprotkb	38-39
Figure 3	Percentage of matching queries after BlastX	38-39
Figure 4	Comparative evaluation of SNP prediction tools	41-42
Figure 5	Comparative Evaluation of SSR Prediction Tools	43-44

**LIST OF APPENDICES**

<b>Sl. No.</b>	<b>Title</b>	<b>Page No.</b>
1	CTAB Extraction Buffer	68
2	TE Buffer (10X)	68
3	TBE Buffer (10 X)	68
4	Wash solution	69
5	Chloroform: Isoamyl alcohol	69
6	80% ethanol	69
7	SSR Predicted using MISA	70-75
8	SNP Predicted using QualitySNP	76-81

**LIST OF ABBREVIATIONS**

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
EMBL	European Molecular Biology Laboratory
MAS	Marker assisted selection
%	Per cent
mM	millimolar
μl	Micro litre
°C	Degree Celsius
bp	Base pair
et al.	And other co workers
Fig.	Figure

g	Gram
g-l	Per gram
mg	Milli gram
ml	Millilitre
sec	Seconds
min	Minutes
ng	Nanogram
SSRIT	Simple Sequence Repeat Identification Tool
MISA	MicroSAtalite identification tool
UniprotKB	Uniprot Knowledge Base
CAP3	Contig Assembly Programme
BLAST	Basic Local Alignment Search Tool

**LIST OF PLATES**

Sl.No	Title	Page No.
1	DNA bands in 0.8% agarose gel	i
2	SNP bands in 3% agarose gel	Ii
3	DaSSR1 and DaSSR2 in 3% agarose gel	Iii
4	DaSSR3 and DaSSR4 in 3% agarose gel	iv

# **INTRODUCTION**

## 1. INTRODUCTION

*D. alata* which is commonly known as greater yam/ water yam/ purple yam, is an important edible tuber crop that comes under the family Dioscoreaceae. The family comprises about 4–6 genera, 870 species that are distributed largely in the tropical and subtropical region of the world. *Dioscorea alata* is a dioecious species with several ploidy level. The crop is well appreciated for its high yield potential, ease of propagation through production of bulbils early vigor for weed suppression and long storage life of tubers. *D. alata* is inhabitant to south-eastern Asia (Acevedo *et al.*, 2005). The worldwide yam production is almost about 47 million metric tons and the tropical country like Africa alone accounts for about more than 96% of the global yam production (FAOSTAT, 2013). It is reported that about 80 to 90% decline in yam production occur due to a disease called, anthracnose or yam dieback which is caused by the fungus *Colletotrichum gloeosporioides*. This particular disease is a main problem faced by the farmers in the cultivation of the greater yam.

Molecular markers have imperative application in plant breeding and crop improvement strategies. They help to change and improve the traits of a plant on the basis of genotype analysis, that depict some modern breeding strategies like marker assisted selection (MAS), marker assisted backcrossing (MABC) etc (Rafalski, A., 2002). Up to now, only a few markers are available for the different types of studies and analysis of this tuber crop (Tamiru M *et al.*, 2015). So it is necessary to develop molecular markers against dieback, that will help to evolve resistant varieties, and eventually mitigate the problem of anthracnose.

Expressed sequence tags (ESTs) are short single read sequences from Genbank. EST is a subsequence of complementary DNA, and act as a resource for evaluating gene expression, to compare the potential variation among the species and also for annotating genes. EST data plays vital role in better understanding of the gene expression and also for the detection of SNPs among various species

(Batley *et al.*, 2003). Analysis of ESTs allows the disentanglement of the complications of gene expression and this method has sophisticated into an economical and capable gene discovery methods. About 74316793 million ESTs are available at the EST database of National Center for Biotechnology Information (NCBI).

Single-nucleotide polymorphism (SNP) and simple sequence repeats (SSR) markers are the two important markers that plays important role in yam breeding programmes (Rafalski, A., 2002). Many researchers have made studies on genetic diversity of yam to develop molecular markers against anthracnose (Arnau G *et al.*, 2017, Tamiru M *et al.*, 2015). These studies were mainly focused on the development of indels, and SNPs for a variety of applications in yam, like genomic selection, genome-wide association analysis, linkage mapping, and MAS. These development of markers and their use in various breeding programmes like MAS or MABC will help to evolve resistant varieties and thus help to ensure the food security in tropical and sub tropical region of the world (Tamiru M *et al.*, 2017). Due to co dominant nature and high abundance in the genome, Single nucleotide polymorphisms (SNPs) are markers of choice (Rafalski, A., 2002). Depending on species of plants, the SNPs occur at a rate of one per 100–500 bp. Due to the progression in sequencing and less expensive methods, the genome-wide discovery of SNPs became much popular. Simple 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, co dominant inheritance, high information rate about polymorphism, and reproducibility (Varshney *et al.*, 2005).

The present study was undertaken to computationally identify SNPs and SSRs for anthracnose resistance in yam and to validate the predicted markers using resistant and susceptible lines. Different SNP and SSR development tools were also evaluated to compare their performance.



# **REVIEW OF LITERATURE**

## 2. REVIEW OF LITERATURE

### 2.1 YAMS (*Dioscorea alata*)

*Dioscorea alata* also known as purple yam , water yam or winged yam is a species of yam and an imperative tuberous root vegetable. It is a dioecious plant under the family Dioscoreaceae that produces edible underground tubers which weigh up to 100 kilograms. *Dioscorea* genus is the only dioecious genus in the family Dioscoreaceae and comprises of about 870 species (Wilkin *et al.*, 2005). Out of the 870 known species only a few are edible (Poornima *et al.*, 2007). Yam is an important staple cash crop, which constitutes about 53% of total root and tuber consumption. It is a cheap source of carbohydrate in the diets of millions of people worldwide and especially in tropical West Africa, Asia, and South America (Asiedu *et al.*, 2014). The greater yam is one of the major cultivated species with wide geographical distribution. It is currently second to *D. rotundata* in production volumes. Several traits of *D. alata* makes it particularly valuable for commercial cultivation. These include high yield potential, ease of propagation, early growth vigour for weed suppression, and long storability of tubers (Sartie A *et al.*, 2014). The tubers possess a high nutritional content with an average crude protein content of 7.4%, starch content of 75–84%, and vitamin C content ranging from 13.0 to 24.7 mg/100g.

*Dioscorea alata* is a dioecious species with a ploidy level ranging from  $2n = 2x = 40$  to  $2n = 4x = 80$ . A study based on the heredity of microsatellite markers has shown that the basic chromosome number of this species is  $x = 20$  and not  $x = 10$  as previously assumed (Arnau G *et al.*, 2009). This species was considered to be highly polyploid with six levels of ploidy ( $2n = 30, 40, 50, 60, 70$  and  $80$ ). However, it is now accepted that it has only three cytotypes ( $2n = 40, 60$  and  $80$ ) and that the most common forms are diploids, followed by triploids and tetraploids are rare (Arnau G *et al.*, 2017) . The center of origin of *D. alata* is not known. Based on archaeological evidence, it is thought to have been domesticated

6000 years ago and is native to Asia-Pacific (Lebot V 2009). It has the largest global distribution of all the yams, and is grown throughout the tropics.

## 2.2 ANTHRACNOSE

Worldwide, yam consumption is 18 million tons. In 2014, yam production was 52 million tons worldwide, of which Africa produced 96%, and Nigeria is the major producer with more than 37 million tons (Palaniyandi S. A., *et al.*, 2017). The consumer demand for yam is very high in sub-Saharan region of Africa, but the yam production is declining in this region due to various factors (Abang M. M., *et al.*, 2002). Among these yam dieback, or anthracnose is regarded as the most serious disease which is caused by an important field pathogen, an airborne fungus *Colletotrichum gloeosporioides*. *Colletotrichum* is a large genus of ascomycete fungi, containing many species which cause anthracnose or blight on a wide range of important crops and ornamental plants. It is probably present in all the countries of the region and is often a major problem where yams (*Dioscorea* spp.) are grown intensively resulting in significant yield losses (Abang M. M., *et al.*, 2003).

Anthracnose disease causes leaf necrosis and shoot dieback of yams thus reducing the photosynthetic efficiency of the plant which result in yield losses of over 90% in susceptible genotypes. However, water yam (*D. alata*) is thought to be more susceptible to anthracnose than other yams. The genetic improvement of yam at IITA and CTCRI (India) concentrated on the development of disease resistant and high yielding varieties. Through classical breeding, it would be very difficult to develop a resistant cultivar due to constraints such as the long growth cycle (8-10 months), dioecious and poor flowering nature, polyploidy, vegetative propagation and heterozygous genetic background (Mignouna *et al.*, 2003). The importance of yams for food security has led to the establishment of several breeding programs for *D. alata*, in order to develop high-yielding cultivars with resistance to anthracnose, and tuber characteristics adapted to farmers'

requirements (Arnau G *et al.*, 2011) Nevertheless, the lack of knowledge on its origin and genetic diversity limits the efficacy of genetic improvement.

### 2.3 EXPRESSED SEQUENCE TAGS (EST)

Expressed sequence tags (ESTs) are fragments of mRNA sequences derived through single sequencing reactions performed on randomly selected clones from cDNA libraries. This unique stretch of DNA within a coding region of a gene is useful for identifying full-length genes and serves as a landmark for mapping. Over the past decade, ESTs have accumulated at an exponential rate and have become a major source of plant sequence data. Today more than 2 million plant derived ESTs from various species are available at public databases. These data have provided a rich resource for gene discovery and annotation (Rudd, 2016). 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).

dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>) a division of GenBank, is a central repository for all the publicly available EST sequences. Users can search ESTs from plants, as well as other kingdoms with NCBI's Entrez system. 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).

### 2.4 GENETIC VARIABILITY AND DIVERSITY IN YAM

Dioscorea species are the chief food security crops for millions of small-scale farmers in the tropical and subtropical regions of Africa, Asia, the Pacific, the Caribbean and Latin America (Ayensu *et al.*, 1972). It is one of the major

cultivated species with wide geographical distribution (Abraham *et al.*, 1990) and is currently second to *D. rotundata* in production volumes. A number of traits of *D. alata* make it particularly valuable for commercial cultivation. The center of origin of *D. alata* is not clear, but based on archaeological evidence, it is thought to have been domesticated ca. 6000 years ago and is native to Asia-Pacific (Lebot *et al.*, 2009). The greatest phenotypic variability in *D. alata* was observed in the southern part of Southeast Asia and in Melanesia, which may be the probable center of origin for this species (Malapa *et al.*, 2005). The South Pacific islands (Papua New Guinea, Fiji, New Caledonia, the Solomon and Vanuatu islands) have rich *ex situ* collections of *D. alata*, including more than 1000 cultivars (Lebot *et al.*, 2009). A wide diversity also exists in India and a rich genetic diversity of yams and the occurrence of about fifty different *Dioscorea* species was reported, largely in the west, east and northeastern regions. Among the *Dioscorea* species, greater yam (*Dioscorea alata* L.) is the most important species grown throughout India. The National Repository on Tuber Crops Germplasm at Central Tuber Crops Research Institute, India conserves 431 accessions of greater yam as field gene bank (Sheela *et al.*, 2016). In addition, several international collections have been assembled, including those of the CRB-PT (Centre de Ressources Biologiques Plantes Tropicales INRA-CIRAD, Guadeloupe, France) and the IITA (International Institute of Tropical Agriculture, Ibadan, Nigeria), with 181 and 772 accessions of *D. alata*, respectively.

## 2.5 MOLECULAR MARKERS

A molecular marker is a molecule contained within a sample taken from an organism or other matter. It can be used to divulge certain characteristics about the respective source. The dramatic development of molecular genetics has laid the groundwork for genomics. It has introduced new generations of molecular markers for use in the genetic improvement of various organisms. These markers provide more precise genetic information and better understanding of the genetic resources. Each marker has its own advantages and disadvantages. Since these molecular markers are numerous in a genome, it can be easily detected. When

they are mapped by linkage analysis, they fill the voids between genes of known phenotype.

Molecular markers are important tools for applications such as estimating genetic diversity and phylogenetic relationships, cultivar identification, mapping of major genes and QTLs, assessing population structure, selection of desirable genotypes in breeding programs, and for authentication of progenies obtained from genetic crosses (Tamiru *et al.*, 2015). Various molecular markers have been used to characterize the genetic diversity of the *D. alata* collections, including RAPDs (Mignouna *et al.* 2002), AFLPs (Egesi *et al.*, 2018) and SSRs, with each method differing in principle, in application, in the type and amount of polymorphism detected, and in cost and time requirements. Although a wide range of marker systems is currently available, marker comparison studies suggest that the choice of method may be dependent on the crop investigated. The use of random amplified polymorphic DNA is quick, easy and requires no prior sequence information. The technique has been used for cultivar identification in many crops, including yam (Dansi *et al.*, 2000). Amplified fragment length polymorphism is generally more reliable than the RAPD technique (Egesi *et al.*, 2018), but is also more laborious and time consuming. As with RAPDs, AFLPs are dominant markers but technical refinements to distinguish homozygous and heterozygous genotypes have recently been developed. The technique has been widely applied for DNA fingerprinting and genetic diversity studies of several crops including yams (Mignouna *et al.*, 2003). SSR markers (microsatellites) are considered to be the markers of choice for analyzing genetic diversity because of their co-dominance, high reproducibility, high global mutation rates and polymorphism (Otoo *et al.*, 2015).

### 2.5.1 Simple Sequence Repeats

Simple sequence repeats (SSRs), also called microsatellites and minisatellites are mutation-prone DNA tracts composed of tandem repetitions of relatively short motifs. SSRs are commonly regarded as 'junk' ,i.e. with no significant role as genomic information. But many molecular and phenotypic

effects of SSR repeat-number variation provide support to the hypothesis that SSRs could have a positive role in adaptive evolution (Thomas *et al.*, 2005). Simple sequence repeat markers are very popular because they are codominant and multiallelic and, thus, are more informative than dominant markers (Zalapa *et al.*, 2012). Microsatellites arose about 25 years ago, and still remain a commonly used genetic marker system in plant genetics and breeding (Miah *et al.*, 2013; Matthies *et al.*, 2012) and forensics (Butler, 2005), where they are commonly referred to as simple sequence repeats (SSRs) or short tandem repeats (STR), respectively. SSR markers are considered to be the markers of choice for analyzing genetic diversity because of their co-dominance, high reproducibility, high global mutation rates and polymorphism (Otoo *et al.*, 2015). There are many SSR markers reported in various crops. These predicted markers provide additional public domain genomic resources for economically important crops to serve as tools for yam genetic research, genetic diversity analysis, and selective breeding (Tamiru *et al.*, 2015). 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, 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).

Simple sequence repeats (SSRs) often serve to modify genes with which they are associated. The influence of SSRs on gene regulation, transcription and protein function typically depends on the number of repeats. SSRs thus provide a prolific source of quantitative and qualitative variation. Over the past decade, researchers have found that this spontaneous variation has been tapped by natural and artificial selection to adjust almost every aspect of gene function (Kashi & King, 2006). 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. 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). They have a high level of potential polymorphism, locus-specificity, multi-allelic and codominant nature, relative abundance and reproducibility.

To date, only a few genomic SSR markers have been developed for *D. cayenensis* and the other *Dioscorea* species (Tostain *et al.*, 2006). Tamiru *et al.*, 2015 developed genomic SSR markers for yellow Guinea yam (*D. cayenensis*) using the method of enriched microsatellite libraries and demonstrate their use in multiple *Dioscorea* species. Despite a growing interest in water yam, published data on molecular characterization and genetic diversity of this crop are scanty (Siqueira, 2011). Few studies on genetic diversity of water yam have been reported using isozymes (Bressan *et al.*, 2011), RAPDs (Random amplified polymorphic DNA) (Mignouna *et al.*, 2002; Zannou *et al.*, 2009), AFLPs (Amplified fragments length polymorphism) (Malapa *et al.*, 2005, Tamiru *et al.*, 2007) and Expressed Sequence Tags (Narina *et al.*, 2011), with each method differing in terms of principle, application, type and amount of polymorphism detected and time requirements (Agarwal *et al.*, 2008). Microsatellite primers have been developed for a few *Dioscorea* species (Mizuki *et al.*, 2005; Hochu *et al.*, 2006), including *D. alata* (Tostain *et al.*, 2006; Siqueira *et al.*, 2011), and have been used on segregation studies, and genetic characterization of *Dioscorea* species (Mignouna *et al.*, 2003; Scarcelli *et al.*, 2005; Bousalem *et al.*, 2006; Tostain *et al.*, 2007; Arnau *et al.*, 2009; Obidiegwu *et al.*, 2009b,c). 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)

## 2.5.2 SSR Prediction Tools

### 2.5.2.1 WebSat

It is a web software for microsatellite molecular marker prediction and development. WebSat is accessible through the Internet, requiring no program installation. WebSat makes use of Ajax techniques, providing a rich, responsive user interface, allowing the submission of sequences, visualization of microsatellites, design of primers suitable for their amplification, and exportation of the resulting data. These tools are very useful, providing a standalone version and, in some cases, a web online version as well. However, the web versions of the programs do not usually have a graphical representation for all the steps involved in the process. It is written in PHP and JavaScript, making use of Ajax techniques. Its input can be either individual sequences, in raw or FASTA format, or a group of sequences in a multi-FASTA format. The user can also choose to upload a file, with a maximum of 150,000 characters. The output generated by WebSat lists the sequences along with the SSRs found, colored yellow and underlined, in a table format. To help the user localize the SSR coordinates, the lines are numbered, and groups of ten bases are separated by a space. By moving the mouse over an SSR, the user can find out its motif and length. The user can then click on any SSR to invoke a primer design program to design a pair of primers flanking the SSR (Martins *et al.*, 2009). The web tool can be accessed at <http://purl.oclc.org/NET/websat/>.

### 2.5.2.2 GMATo

Genome-wide Microsatellite Analyzing Tool (GMATo) is a novel tool for SSR mining and statistics at genome aspects. It is faster and more accurate than existed tools SSR Locator and MISA. If a DNA sequence is too long, it is fragmented in to short segments at several Mb, followed by motifs generation and searching using Perl powerful pattern match function. Matched loci data from each fragment is then merged to produce final SSR loci information. Only one input file is required which contains raw FASTA DNA sequences and output files in tabular format, that list all SSR loci information and statistical distribution at four classifications. GMATo is programmed in Java and Perl with both graphic and command line interface, either executable alone in platform independent manner with full parameters control. Software GMATo is a powerful tool for complete SSR characterization in genomes at any size. It is also easy to mine SSR in the genome using a normal computer because processing one segment at a time in GMATo required less computing memory. Both graphic user and command line interface were provided in GMATo, either executable independently in Windows, Linux or Mac OS system. Only one input file containing DNA sequence in FASTA format is required to be chosen in graphic mode or typed in command mode if taken the default parameters. The output files generated by GMATo is one formatting report, one file containing SSR loci information and another file containing statistical distribution of SSR (Wang *et al.*, 2013). The tool can be accessed at <https://sourceforge.net/projects/gmato/files/>.

### 2.5.2.3 SSR Locator

It is a software for detection and characterization of SSRs. In addition to the SSR detection, it also finds minisatellite motifs between 1 and 10 base pairs, design primers for each locus that have found, simulate PCR (polymerase chain reaction), thus amplify the fragments with different primer pairs, from a given set of fasta files. Finally it also align the amplicons generated by the same primer

pair and estimate the global alignment scores. The algorithm used for perfect and imperfect micro-/minisatellite searches is written in Perl and consists of the generation of a matrix that mixes A(adenine), T(thymine), C(cytosine), and G(guanine) in all possible composite arrangements between 1 and 10 nucleotides. The script instructions perform readings on fasta files, searching all possible arrangements in each database sequence (Maia *et al.*, 2008). The tool is available at <http://microsatellite.org/ssr.php>.

#### **2.5.2.4 MISA**

MISA microsatellite finder (Thiel *et al.*, 2003) is a tool for finding microsatellites in nucleotide sequences. In addition to the detection of perfect microsatellites, MISA is also able to find perfect compound microsatellites that are composed multiple occurrences of more than one simple sequence motif (Beier *et al.*, 2017). A microsatellite analysis with the command line version of MISA requires two input files, a configuration file ('MISA.ini') with three input parameters: 'SSR search parameters', 'compound SSR search parameter' and 'output file type parameter'; and a FASTA file containing the nucleotide sequence that is to be mined for microsatellites. MISA-web runs on a standard Linux server and works in conjunction with several helper scripts and programs in addition to the core MISA PERL script. 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). The tool is available at <http://misaweb.ipk-gatersleben.de/>.

#### **2.5.2.5 SSRIT**

SSRIT finds all perfect simple sequence repeats (SSRs) in a given sequence. Even though the output does contain sequence ID, motif (repeat) type, no. of repeats, SSR start and end, it does have some limitations. The program currently is not capable of detecting mononucleotide repeats, and the output is not perfect 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. The tool is available at [www.gramene.org/db/markers/ssrtool](http://www.gramene.org/db/markers/ssrtool).

### 2.5.3 Single Nucleotide Polymorphism

Sequence variation in the genomic DNA of individuals of the same species or related species are typically single nucleotide polymorphisms (SNP) or small insertions/ deletions (indels) (Useche *et al.*, 2001). Because of their abundance and slow mutation rate within the genome, they are the most common type of genetic markers for studying complex genetic traits and genome evolution (Syvanen *et al.*, 2001). In addition SNPs in coding sequences can be used to directly study the genetics of expressed genes and to map functional traits (Rickert *et al.*, 2003 : Grivet *et al.*, 2003). Non-synonymous SNPs (nsSNPs) are of particular interest because they change the protein sequence, possibly affecting protein function (Kim *et al.*, 2003). There are several strategies, both experimental and computational for SNP discovery. Experimental SNP discovery often consists of a number of laborious steps that make this process complex and expensive (Useche *et al.*, 2001). The computational approach makes use of the large sequence datasets present in public databases. Over the last few years, a number of pipelines have been developed that automatically detect SNPs in such databases.

Single nucleotide polymorphism (SNP) analysis provides an important tool in applications as genetic linkage mapping, fine-mapping of candidate regions and to determine haplotypes associated with traits of interest, in order to understand the genetic basis of phenotypic diversity within and between populations. Recently, large-scale identification and characterization of SNPs has attracted much interest in connection with the sequencing projects of the human and vertebrate genomes (Guryev *et al.*, 2004). Due to the high abundance in the genome, thousands of potentially informative SNP markers can be identified for the development of high density SNP maps, which are an essential resource to

identify the underlying genes responsible for the variation of complex traits or QTLs (Vignal *et al.*, 2002).

SNPs may be considered the ultimate genetic marker as they represent the finest resolution of a DNA sequence. They are generally abundant in populations and have a low mutation rate (Syvanen *et al.*, 2001). Analysis of assembled EST sequence data provides a cost effective means to identify large numbers of SNPs associated with functional genes (Duran *et al.*, 2008). Molecular genetic markers describe genetic variations and provide a link between observed phenotypes and the underlying genotype. The development of high-through put methods for the detection of single nucleotide polymorphisms (SNPs) and small insertion/deletions (indels) has led to a revolution in their use as molecular markers. Single nucleotide polymorphisms (SNPs) are important tools in studying complex genetic traits and genome evolution.

#### **2.5.4 SNP Prediction Tools**

##### **2.5.4.1 AutoSNP**

It is a tool for SNP detection using SNP redundancy score and co-segregation. The frequency of occurrence of polymorphism at a particular locus provides a measure of confidence in the SNP, representing a true polymorphism and is referred to as the SNP redundancy score. In addition, true SNPs that represent divergence between homologous genes co-segregate to define a conserved haplotype. A co-segregation score based on whether a SNP position contributes to defining a haplotype is a further independent measure of SNP confidence. The SNP score and co-segregation score together provide a valuable means for estimating confidence in the validity of SNPs within aligned sequences independent of sequence trace files. AutoSNPdb has a flexible interface facilitating a variety of queries. Users may search for SNPs within genes of predicted function, and through sequence identity with known genes. In addition, it is possible to add additional levels of annotation and novel queries specific to areas of interest. This web interface allows users to query and visualize the SNP

and annotation data. Sequence annotations may be searched by gene keyword, sequence ID, GO term or through similarity to defined regions. A BLAST interface enables identification by sequence similarity. SNPs may be retrieved that differentiate between cultivars, providing a valuable resource for genetic mapping and association studies. To aid interpretation of the predicted SNP data, SNPs are viewed graphically as vertical bars, where the position of the bar along the x-axis reflects the relative position of the SNP in the consensus sequence; the height of the bar represents the SNP redundancy score; and the bar colour reflects the SNP-weighted co-segregation score. Information about each SNP is displayed by moving the cursor over the bar, while selecting a bar centres the sequence assembly at that position. The sequence assembly may be moved using the scroll bar and can be toggled between the full sequence assembly and a SNP summary. The AutoSNPdb system was developed for flexible use and permits extension to a broad range of annotation and species (Duran *et al.* , 2008). The tool is available at <http://autosnpdb.qfab.org.au/>.

#### **2.5.4.2 QualitySNP**

It is an efficient tool for SNP detection, storage and retrieval in diploid as well as polyploid species. It is available for running on Linux or UNIX systems. The program, test data, and user manual are available at <http://www.bioinformatics.nl/tools/snpweb/>. It uses a haplotype - based strategy to detect reliable synonymous and non-synonymous SNPs from public EST data without the requirement of trace/quality files or genomic sequence data. QualitySNP distinguishes itself from other programs mainly in the approach it takes for detecting sequencing errors and paralogous sequences. The source code and the manual of the program are freely available for academic use. Several steps in the QualitySNP pipeline are designed to improve the reliability of the SNP output of the program. These steps include

- The mihap/mahap calculations

- Use of High Confidence Scores which effectively eliminates most of the SNPs identified in presumably low quality sequence regions
- Haplotype reconstruction
- Use of D-value thresholds for filtering out paralog containing clusters

In QualitySNP, most of the settings can be adjusted according to the user's preference. It has a retrieval system that allows the user to extract additional useful information from the analysis. Information about the nature of the SNPs (synonymous or non-synonymous) can be made part of the output. The SNP output can be modified by changing the reference genotype, and the D-value setting can be used to adjust the stringency with which paralogous clusters are detected and excluded. This is very useful when focusing on a specific gene family where alleles of different paralogous sequences need to be identified. Statistics concerning the number of different types of SNPs and clusters can be included in the output. Searching parameters include the contig reference number, GenBank/EMBL/DDBJ accession number of ESTs, and UniGene ID. Output options include SNP information, alignment information, EST function annotation information and ORF information of the contig (Tang *et al.*, 2006).

#### **2.5.4.3 PredictSNP**

It is an accurate and robust tool for the prediction of SNPs. This user-friendly web interface enables easy access to eight prediction tools like MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP. It has improved prediction performance, and at the same time it also provides results for all mutations, confirming consensus prediction. Using this web server, a user can load an amino acid sequence of a query protein in FASTA format, and can select positions for mutations and desired mutations using the input page. In addition to this, the user can submit a list of mutations in a text format. After all desired mutations are specified, the user can select tools to be employed for the evaluation of selected mutations. A time estimate is provided for each tool and a number of mutations, based on an average evaluation time for individual tools. The server then runs the prediction using all selected tools. In the

cases where MAPP is included in the selection, the necessary multiple sequence alignment and phylogenetic tree are automatically calculated. The prediction is finalized by calculation of the PredictSNP confidence score. The PredictSNP web server is freely available to the community at <http://loschmidt.chemi.muni.cz/predictsnp>. The developed datasets, the user manual, and standalone version of PredictSNP consensus calculator are also available from the website. The standalone version represents an alternative to web server that is suitable for massive mutagenesis studies (Bendl *et al.*, 2014)

#### **2.5.4.4 SNAP**

It is a neural-network based method that uses *in silico* derived protein information (e.g. secondary structure, conservation, solvent accessibility, etc.) in order to make predictions regarding functionality of mutated proteins. The network takes protein sequences and lists of mutants as input, returning a score for each substitution. These scores can then be translated into binary predictions of effect. SNAP utilizes various biophysical characteristics of the substitution, as well as evolutionary information, some predicted structural features, and possibly annotations, to predict whether or not a mutation is likely to alter protein function. Although such predictions are already available from other methods, SNAP added important novelty. Amongst the novel aspects, the improved performance throughout the entire spectrum of accuracy/coverage thresholds and the provision of a reliability index that enables users to either zoom into very few very accurate predictions, or to knowingly broadcast less reliable ones (Bromberg *et al.*, 2007). The tool is available at <http://www.rostlab.org/services/SNAP>.

#### **2.5.4.5 PMut2017 PREDICTOR**

PMut portal is a novel approach for SNP prediction that largely improves previous 2005 PMut server. The new portal offers not only a generally trained predictor that performs in a competitive manner with current available methods, but allows the user to access an automatic procedure to train new predictors with specific datasets or features. The possibility of enriching the analysis with



alternative predictors, or training predictors with specific information of a single protein family, largely increases the scope of usability of the portal. Overall, the 2017 release of PMut is a powerful tool to approach the issue of predicting functional consequences of protein sequence variants. It has large acceptance in the field of predicting Mendelian pathological mutations. PMut internal engine has a fully featured standalone training and prediction engine that not only powers PMut web portal, but that can generate custom predictors with alternative training sets or validation. The default predictor performs with good quality scores. The PMut portal is freely accessible at <http://mmb.irbbarcelona.org/PMut>. A complete help and tutorial is available at <http://mmb.irbbarcelona.org/PMut/help>. PMut prediction engine (PyMut) is prepared as a Python 3 module (Ferrando *et al.*, 2017).

# **MATERIALS AND METHOD**

### 3. MATERIALS AND METHODS

The study entitled “Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation” was conducted at the Central Tuber Crop Research Institute (CTCRI) during 2017-2018. Details regarding the experimental materials used and methodology adopted for various experiments are presented in this chapter.

#### 3.1 YAM SEQUENCE DATASET

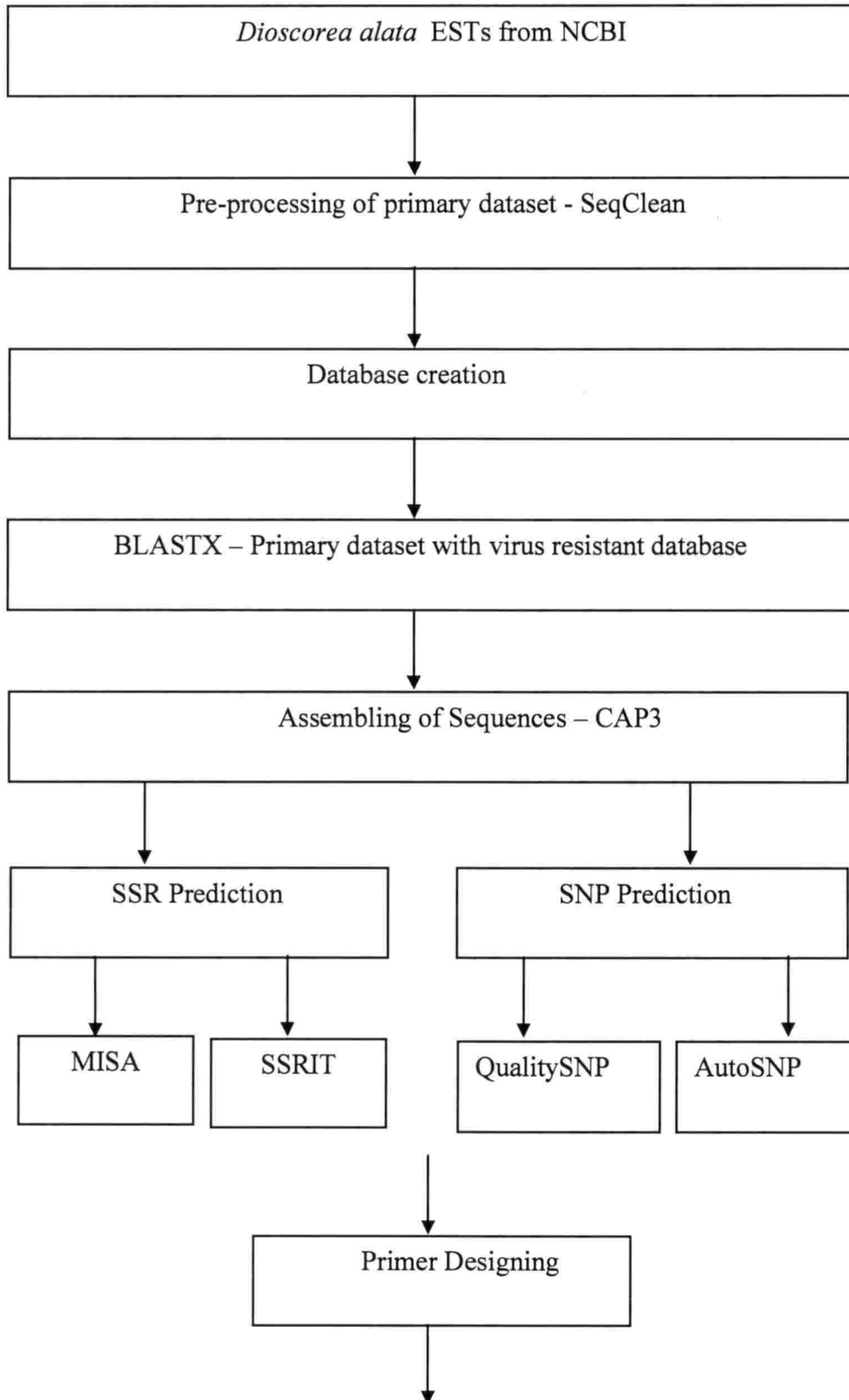
The preliminary data set for the work was obtained from the EST section of NCBI (<https://www.ncbi.nlm.nih.gov/>). The EST database is a collection of short single-read transcript sequences from Genbank. These sequences provide a resource to evaluate gene expression, find potential variation, and annotate genes. *Dioscorea alata* sequences were retrieved from the Genbank EST section on 15<sup>th</sup> November 2017. A total of 44134 ESTs of yam were downloaded from NCBI and this was taken as the primary dataset for research work. Work flow is given in Figure 1.

#### 3.2 PRE-PROCESSING OF SEQUENCES

The sequences were processed for removing 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 FASTA file. 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)

## WORKFLOW



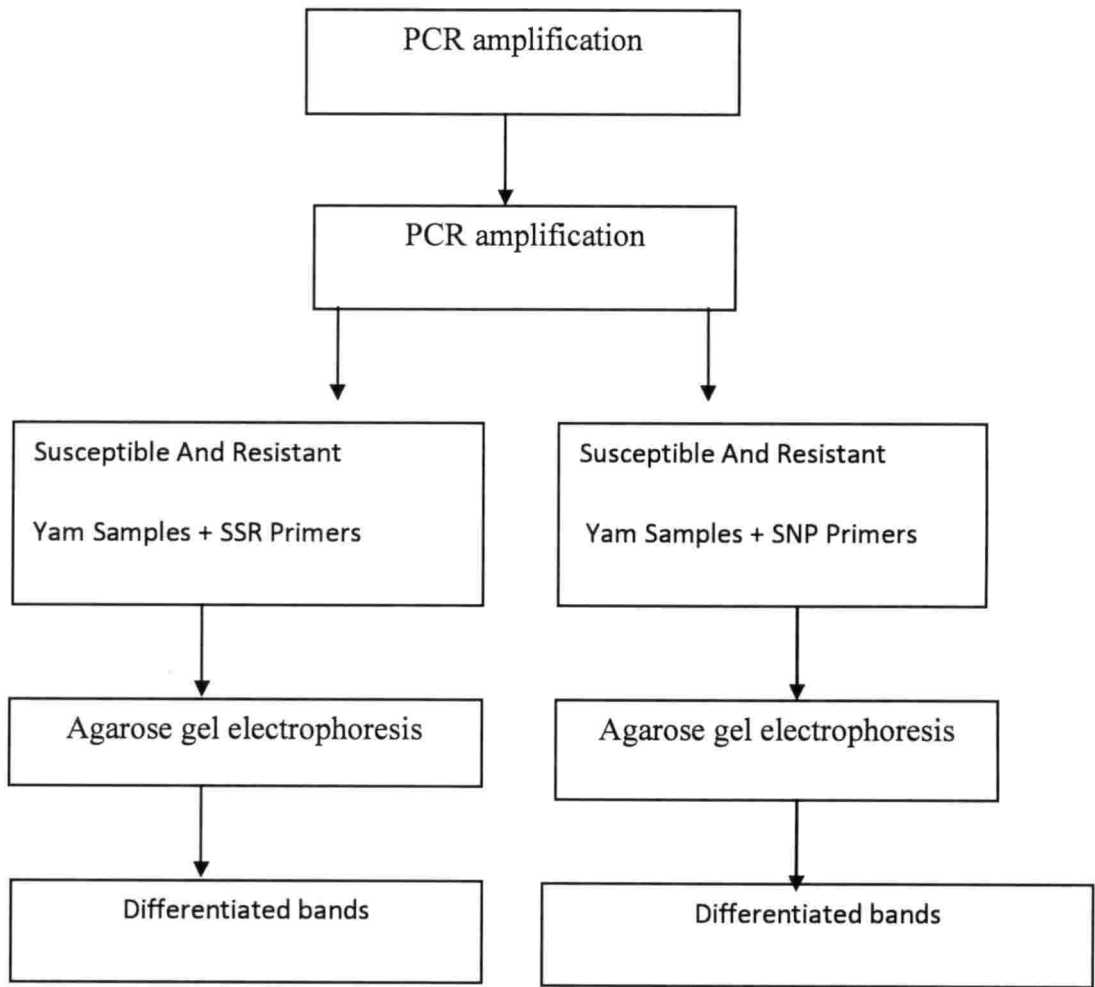


Figure 1: Workflow for the identification of SSRs and SNPs

- Strong matches with other contaminants or unwanted sequence (mitochondrial, ribosomal, bacterial, other species than the target organism etc.)

To clean the sequence using SeqClean, certain requirements are needed. The basic requirements are

- Perl version  $\geq 5.6$
- 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))

### 3.2.1 Installation of SeqClean

First of all we have to create a directory where we plan the package to reside. Then copy the compiled archive into that directory and unpack the archive in there. The command **tar xvzf seqclean.tar.gz** will unpack a few files in the current directory and will create a bin subdirectory with several files.

### 3.2.2 Usage and methods

A short usage message is displayed when seqclean script is launched without any parameters. The seqclean script takes an input sequence file in fasta format , and when the command **seqclean your\_est\_file** is given, it will produce two output files:

- 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.).

### 3.2.3 Cleaning report format

Each line in the cleaning report file (\*.cln) has 7 tab-delimited fields as follows:

- The name of the input sequence
- The percentage of undetermined bases in the clear range
- 5' coordinate after cleaning
- 3' coordinate after cleaning
- Initial length of the sequence
- Trash code
- Trimming comments (contaminant names, reasons for trimming/trashing)

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 GENE DATABASE

In order to develop the markers related to anthracnose, a plant specific database of virus resistant genes is required. Hence resistance virus gene database was created and compiled from uniprotKB (UniProt Knowledgebase) manually. The UniProtKB provides the freely accessible central database of protein sequences with accurate, consistent, rich sequence and functional annotation. It is 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:

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

For doing BlastX, the resistant genes should be strictly from plant related genes. This was achieved through the Plant protein database in uniprot called Viridiplantae ([tp://141.161.180.197/pub/databases/uniprot/current\\_release/knowledgebase/taxonomic\\_divisions/uniprot\\_sprot\\_plants.dat.gz](http://141.161.180.197/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_plants.dat.gz)). The R-gene or resistant genes downloaded through this way, which are related to anthracnose diseases were screened out and was used for database creation. The virus resistance protein database consisted of 290 resistant genes.

### 3.3.1 Processing of Resistant Gene Database

**BLAST** (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. The database is created using the command line **makeblastdb -in my\_reference.fa -out my\_reference -parse\_seqids -dbtype prot**. But the database thus created will contain numerous duplication. In order to make a strong database, it is necessary to remove these duplications. Duplications are removed using the awk command: **awk '/^>/{f=!d[\$1];d[\$1]=1}f' in.fa > out.fa**



### 3.4 SCREENING OF PRIMARY DATASET AGAINST RESISTANT GENE DATABASE USING BLASTX

For screening of primary dataset with virus resistant protein sequences 'BlastX Search protein database using a translated nucleotide query' was used. The retrieved EST sequences from NCBI is compared with the resistant gene database from Uniprot using BlastX by giving the command **blastx -query fasta.file -db database\_name -outfmt 6 -num\_alignments 1 -num\_descriptions 1 -out output\_file**. The resulting output is then imported to Excel for further sorting. The sorting process help to select the sequences with low Expectation value which will give more efficient and accurate result. It is done by giving the command line **sort -u list > -output** and the result will be displayed in tabular form. Thus *Dioscorea alata* ESTs were blasted against resistant genes and the sequences having high similarity were used for further analysis.

### 3.4 ASSEMBLING OF SEQUENCES

CAP3 (Contig Assembly Program Version 3) is a sequence assembly program which is available at (<http://seq.cs.iastate.edu/CAP3.html>). The input a file of sequence must be given 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. The programme is run by giving the command line CAP3 File\_of\_reads . Here the File\_of\_reads is a file of DNA reads in FASTA format. 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 was assembled using the CAP3 program with default runtime options.

### 3.5. SNP PREDICTION:

There are many tools for the prediction of SNPs (Single Nucleotide Polymorphism). Here in this study mainly two prediction tools are used.

- AutoSNP
- QualitySNP

#### 3.5.1 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.

### 3.5.2 QualitySNP

QualitySNP is an efficient tool for SNP detection, storage and retrieval. It implements a new algorithm 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 six steps:

1. The first step performs EST assembling using `cross_match` for removing vectors and `CAP3` for sequence clustering.
2. The second step is the analysis of the alignment information to select clusters with at least 4 EST members; this is done by the Perl script “`Getalignmentinfo`”. If sequences with quality information are available, another Perl script “`Getalignmentinfoqual`” is used instead of “`Getalignmentinfo`”.
3. The third step 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. In the case of sequences with quality information, the program “`QualitySNPqual`” should be used instead of “`QualitySNP`”. Before using “`QualitySNPqual`”, another program “`GetavailESTqual`” should be run to get quality score for each sequence in each cluster that contains at least 4 sequences.
4. Step four is the non-synonymous SNP discovery 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.

5. The fifth step transfers the final results into a SNP database. It includes two C programs: "Getsnpindexcontig" (for sequences without quality information) or "Getsnpindexcontigqual" (for sequences with quality information) formats all information of contigs and the location and types of SNPs for insertion into the MySQL database; the second program "Transfersnpfasta" is used to convert SNP-containing probes for microarray analysis into the format for the MySQL database. There are two SQL scripts to create the database and to load the data into the database. "dbcreator.sql" (for sequences without quality information) or "dbcreatorQ.sql" (for sequences with quality information) can create the database and tables that are used by the retrieval system, and "dataload.sql" puts the formatted data into database.
6. The final part is the retrieval system that is written in PHP. All PHP scripts and HTML pages are stored in the website's directory tree.

#### Commands to run Quality SNP:

1. % cap3 filename -p similarity -o 100. Here the 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, and min-clustersize is the minimum cluster size. The default minimal cluster size is 4.
3. % Getavailcontigseq filename.cap  
 % Getavailcontigqual filename.cap  
 % 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 (2 in our study)

- lowqual5side is the length of the low quality region at the 5' end of sequence (30 nucleotides in our study)
  - 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 (0.2, 20% of the whole sequence in our study).
  - weightlowqual is the weight value of the low quality region (0.5)
  - min-confidencscore is the minimal confidence score (2)
4. % fasty34\_t allavailcontigseqwithSNP Uniprot -b 6 -d 6 -Q > allavailcontigseqwithSNP.fasty

The parameters used in these commands are :

- Uniprot is the Uniprot (or any other) protein database. 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. File “allavailcontigseqwithSNP” contains the consensus sequences of SNP-containing contigs which did not contain any insertions or deletions.

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.

### 3.6 SSR PREDICTION

There are many tools for the prediction of SSRs(Simple Sequence Repeats). Here in this study mainly two prediction tools are used.

- SSRIT – Simple Sequence Repeat Identification Tool
- MISA - MicroSAtelite identification tool

#### 3.6.1 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.

The SSR were obtained by giving the command **perl ssr.pl input sequence**

#### 3.6.2 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 command used to run is **perl misa.pl input sequence** 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.

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

### 3.7 PRIMER DESIGNING FOR SNPs AND SSRs PREDICTED USING QUALITYSNP AND MISA

A primer is a short strand of RNA or DNA which generally have a size about 18-22 bases , that serves as a starting point for DNA synthesis. 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. For this the retrieved SNP and SSR are blasted against the resistant gene database and then contigs with hit percentage between 80% – 100% and lower Evalue was selected. Based on this, the best five contigs, satisfying above criteria was selected and primer pairs are designed using Primer3plus tool.

#### 3.7.1 Primer3plus

Primer3plus is a widely used program for designing PCR primers (PCR = "Polymerase Chain Reaction"). PCR is an essential and ubiquitous tool in genetics and molecular biology. Primer3 can also design hybridization probes and sequencing primers. Primer3 picks primers for PCR reactions, considering certain criteria such as 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. The parameters considered in primer designing:

- **Primer Length:**

It is generally accepted that the optimal length of PCR primers is 18-22 bp. This length is long enough for adequate specificity and short enough for primers to bind easily to the template at the annealing temperature

- **Primer Melting Temperature:**

Primer Melting Temperature ( $T_m$ ) by definition is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicates the duplex stability. Primers with melting temperatures in the range of 52-58 °C generally produce the best results.

- **GC Content:**

The GC content (the number of G's and C's in the primer as a percentage of the total bases) of primer should be 40-60%.

- **GC Clamp:**

The presence of G or C bases within the last five bases from the 3' end of primers (GC clamp) helps promote specific binding at the 3' end due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.

### 3.8 VALIDATION

#### 3.8.1 Sample collection for DNA extraction

Fresh, young leaves of 6 *Dioscorea alata* accessions were collected from germplasm collection of ICAR-CTCRI, Thiruvananthapuram. The collected samples comprised of 3 anthracnose resistant and 3 susceptible accessions. Young leaf tissues were collected in plastic sample collection bags from the field and brought to lab in an ice box.

Table 1: List of accessions of greater yam used for the study

Sample No:	Resistant Sample	Sample No:	Susceptible Sample
Da 1	Sree Karthika	Da 3	Sree Neelima
Da 2	Sree Keerthi	Da 4	Orissa Elite
Da 3	Sree Swathi	Da 5	Sree Roopa



### 3.8.2 DNA Extraction

The pre requisite for the validation is the isolation of good quality DNA from the plant tissue. DNA was extracted from fresh and young leaves of collected samples using modified protocol of Raj *et al.* (2014).

Young leaf tissues of *D. alata*, were collected during early hours in the morning and DNA was isolated from these leaves. Leaf tissues (200–250mg) were ground to a fine powder using liquid nitrogen. Prewarmed extraction buffer (1ml) was added to the samples and it was ground one more. The samples were transferred to 2.0 ml Eppendorf tubes and 10  $\mu$ l Proteinase K (10 mg/ml) was added. The tube was incubated in 37°C for 30 min and then at 65°C for another 30min with frequent swirling. Samples were centrifuged at 12,000rpm for 15min at RT and supernatant was transferred to fresh eppendorf tube. Equal volume of Chloroform: isoamyl alcohol (24:1) were added and mixed by gentle inversion for 30–40 times. The samples were centrifuged at 12,000g for 10 min at RT and the supernatant was transferred to a fresh tube. The above step was repeated again to remove any further proteins present. To the supernatant collected in a fresh tube, 150  $\mu$ L of 2 M NaCl solutions containing 4% PEG was added. The samples were centrifuged at 12,000 rpm for 10 min at RT. The supernatant was transferred to a fresh tube and precipitated with 200 $\mu$ l of ethanol. The nucleic acids was precipitated and collected by centrifuging at 12,000 rpm for 10 min. The nucleic acid pellet was washed twice with wash solution, air-dried until the ethanol was removed and dissolved in appropriate amount of TE buffer (100–150 $\mu$ l). The nucleic acid dissolved in TE buffer were treated with ribonuclease (RNase, 10mg/ml), incubated at 37°C for 30 min and stored at -20°C until use. All samples were checked in 1% agarose gel and confirmed.

### 3.8.3 Quantification of DNA

Using Nanospectrophotometer (DENOVIX), the isolated DNA was quantified. It was used to determine the yield and purity of the isolated DNA. TE buffer in which the DNA was dissolved was used to calibrate the machine. To measure the quantity and purity of DNA, 1 $\mu$ lof DNA sample was placed in the

sensor probe of the machine. It's the benefit of Nanospectrophotometer that no other dye is required. The quantity and quality of the DNA was displayed as the concentration of the DNA in  $\text{ng}/\mu\text{L}$ , the absorbance ratio  $\text{OD}_{260/280}$  and  $\text{OD}_{260/230}$ . At  $\text{OD}_{260}$ , the quantity of DNA was determined and the purity was determined by calculating the  $\text{OD}_{260/280}$  ratio. The samples were selected based on the better absorbance value or OD value.

### 3.8.4 Agarose gel electrophoresis

Agarose gel electrophoresis is widely used to separate biomolecules such as DNA, RNA and proteins based upon charge, size and shape. Agarose gel electrophoresis possesses great resolving power, yet it is simple to perform. The samples are mixed with loading buffer containing glycerol or sucrose and tracking dyes are loaded into wells in agarose gel. A direct current supply is connected to the electrophoresis apparatus. Molecules having a net negative charge (DNA) migrate towards the positive electrode (anode) while net positively charged molecules migrate towards the negative electrode (cathode). The 1X TBE or TAE buffer in the gel tank serves as a conductor of electricity and control pH. pH is important to the charge and stability of biomolecules. Smaller molecules move faster through the pores in the gel than larger ones. Molecules can have similar molecular weight and charge but different shape. Molecules that have a very compact shape can move faster through the pores.

Weighed 1g of agar powder was transferred to a conical flask. 100ml of 1X TBE buffer was added and heated for 2 min on a microwave oven to dissolve the agarose.  $1\mu\text{l}$  ethidium bromide ( $.5\mu\text{g}/\text{ml}$ ) was added to the pre-cooled agarose solution, mixed well and the solution was poured into 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  $2\mu\text{l}$  gel loading dye and loaded to the wells.  $2\mu\text{l}$  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.

### 3.8.5 Dilution of DNA Samples

Samples were diluted to a concentration of 10ng/ $\mu$ l using nuclease free water, irrespective of the varying concentrations calculated by spectrophotometrically.

### 3.8.6 Primer dilution and PCR amplification

Working stock of 10  $\mu$ M concentration were made out of the 100 $\mu$ M stock and stored in -20°C refrigerator. The diluted samples were amplified in thermal cycler using different primers of SSR and SNP at different conditions. All of the primers were screened and optimum amplifying conditions were used for validating all primers of SSR and SNP.

### 3.8.7 PCR using EMERALD Amp GT PCR master mix

EMERALD Amp GT PCR Master Mix by TAKARA BIO INC is a 2x premix composed of a DNA polymerase, optimized reaction buffer, dNTPs, and a density reagent. The premix also contains a vivid green dye that will separate into blue and yellow dye fronts when run on an agarose gel. The premix simplifies PCR assembly; simply add primers, template, and water and start the reaction. After PCR, the reaction mixture can be applied directly to a gel for analysis. For the sake of convenience and for saving time during the PCR mix preparation, the Emerald Amp GT PCR master mix was used for the work.

Table 2: The reaction mixture used for SSR

Components	Stock Concentration	Required Concentration	Volume For One Reaction (15 $\mu$ l)
Emerald MasterMix	2X	1X	7.5 $\mu$ L
Forward Primer	10 $\mu$ M	0.3 $\mu$ M	0.45 $\mu$ L
Reverse Primer	10 $\mu$ M	0.3 $\mu$ M	0.45 $\mu$ L
DNA	10ng / $\mu$ L	40ng	4.0 $\mu$ L
Sterile Distilled Water	.....	.....	2.6 $\mu$ L
<b>TOTAL</b>		<b>15<math>\mu</math>L</b>	

### 3.8.8 PCR Conditions

PCR was carried out in ProflexThermalcycler programmed for an initial denaturation at 95 °C for 3 minutes followed by 30 cycles of denaturation at 95 °C for 45 seconds, primer annealing (Ta) for 45 seconds and extension at 72 °C for 1 minute. The final extension was performed at 72 °C for 10 minutes followed by hold at 4 °C.

The amplified products were separated on 4% agarose gel along with 100bp ladder to compare the molecular weight of obtained bands.

### 3.8.9 Agarose gel electrophoresis

Weighed 1.7g of agarose in 250ml conical flask, added 85ml 1X TBE buffer (for 2% gel) and gently boiled the solution in a microwave oven with occasional mixing until the agarose gets completely dissolved in buffer. Allowed it to cool for 40° C and added 0.4µl ethidium bromide carefully without spilling. Get ready the gel plates and kept combs in position. Poured the warm gel to plate and cool for 20 minutes. Once the gel gets solidified, removed the comb and placed the plate with gel in to the tank containing 1X TBE buffer (Appendix III). 8uL of samples, which already contain the Emerald dye, were loaded into the wells and 3uL 1Kb or 100bp ladders were also loaded as reference. Run the gel at 85V and 220mA for 1 to 1.5 hour. Visualized and documented the gel on a gel documentation system. The images could be finally scored to detect polymorphism or to identify specific bands that can be linked to a particular trait.

### 3.8.10 Validation of SSR Markers

To confirm that the designed SSR markers are working, AGE was done. Validation of SSR markers was done by running the 5 designed SSR primers (DaSSR1, DaSSR2, DaSSR3, DaSSR4, DaSSR5) with all three resistant and susceptible yam varieties, and then examining the bands for any distinct variability in position which would confirm that the marker is working. The

annealing temperature for the primers were obtained from the Tm Calculator of Thermo Fisher Scientific.

### **3.8.11 Validation of SNP Markers**

To confirm that the designed SNP markers are working, AGE was done. Validation of SNP markers was done by running the 4 designed SNP primers (DaSNP1, DaSNP2, DaSNP3, DaSNP4, DaSNP5) with a resistant and susceptible yam varieties, and then examining the bands for any distinct variability in position which would confirm that the marker is working. The annealing temperature for the primers were obtained from the Tm Calculator of Thermo Fisher Scientific.

# RESULTS

## 4. RESULTS

### 4.1 COLLECTION OF YAM SEQUENCE DATASET

The preliminary data set for the work was obtained from the EST section of NCBI (<http://www.ncbi.nlm.nih.gov>). A total of 44134 ESTs of yam were downloaded from NCBI and this was taken as the primary dataset for research work.

### 4.2 PRE-PROCESSING OF PRIMARY DATASET

The primary dataset were processed for removing contamination or simple repeats using the SeqClean. The sequences are checked for sequence contamination and simple repeats by using the SeqClean, with the default runtime options. UniVec\_Core database of NCBI is used to clean the ESTs.

A total of 44134 sequences were analyzed, and the rest of 43114 sequences were used for further processing. The remaining 1020 sequences were removed during cleaning. Among the trashed sequences, 7 were trashed due to "low qual " and it is assigned when the percentage of undetermined bases is greater than 3% in the clear range; 87 were trashed due to short q, which is assigned when the sequence length decreases below the minimum accepted length (-1) after polyA or low quality ends trimming; 921 were removed due to short t and 5 were removed due to dust, which is assigned when less than 40nt of the sequence is left unmasked by the "dust" low-complexity filter (Table 3).

### 4.3 RESISTANT GENE DATABASE

The R-gene or resistant genes was retrieved through the Plant protein database in uniprot called Viridiplantae. The resistance protein database consisted of 290 resistant genes. But the database thus created contained numerous duplication. In order to make a strong database, it was necessary to remove these duplications. There were only 287 resistant genes left after duplication removal.

Table 3: Result of Pre-processing of primary dataset using SeqClean

Total sequence analyzed	44134
Valid sequences	43114
Trashed	1020
Trashing summary	
<ul style="list-style-type: none"> <li>• By low qual</li> <li>• By short q</li> <li>• By short t</li> <li>• Dust</li> </ul>	7 87 921 5

#### 4.4 SCREENING OF THE SEQUENCES AGAINST VIRUS RESISTANT GENES USING BLASTX

The retrieved EST sequences from NCBI is compared with the resistant gene database from Uniprot using BlastX. The primary sequence dataset with 43114 EST were blasted against resistant gene database with 287 resistant genes. There were about 38179 output sequences. Among these 97% of sequences showed similarity and remaining 3% of sequences were failed to show any similarity.

Table 4: Result of Screening of primary dataset against resistant gene database

Query	EST (43114)
Database	Resistant gene database (287)
Number of output sequence	38179



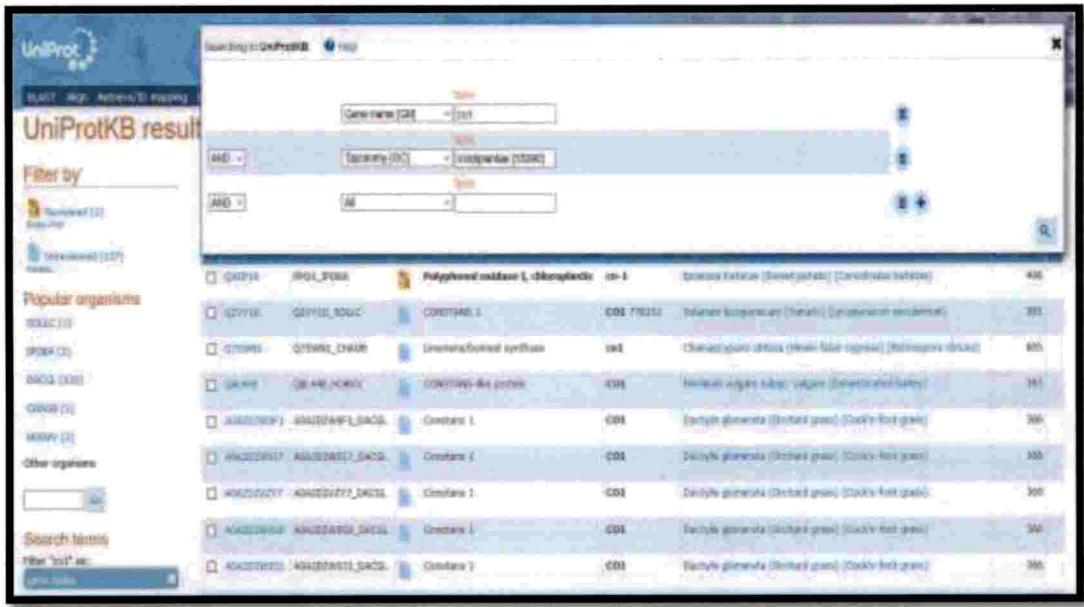


Figure 2: User window of uniprotKB

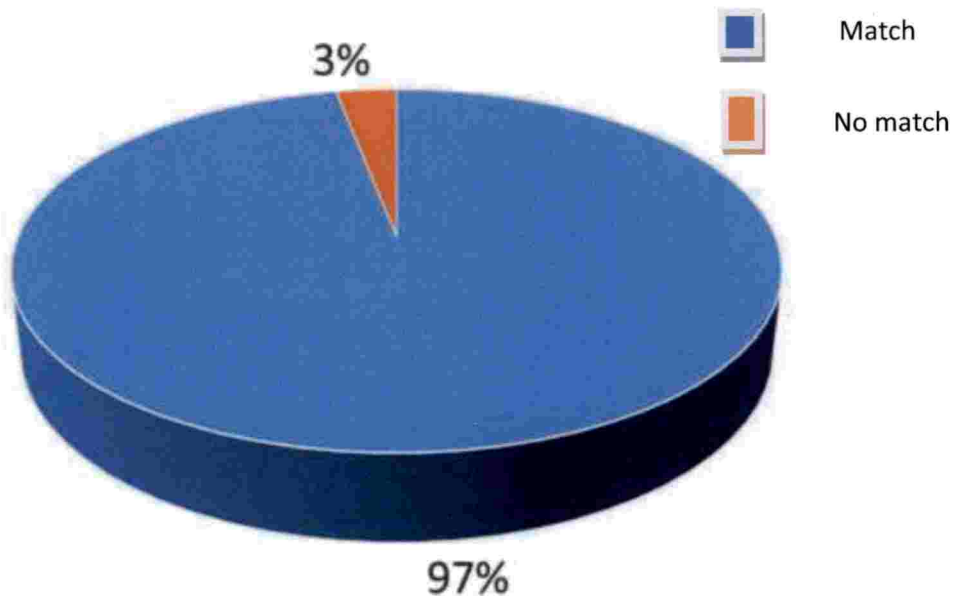


Figure 3: Percentage of matching queries after blastx

#### 4.5 ASSEMBLING OF SEQUENCES

CAP3 (Contig Assembly Program Version 3) is a sequence assembly program which is available at (<http://seq.cs.iastate.edu/CAP3.html>). The result of BLASTX was taken as the input file for the assembly. After assembling of the sequences 5940 contigs and 12431 were obtained.

Table 5: Assembling of Sequences using CAP3

Number of sequences analyzed	38179
Contigs	5940
Singlets	12431

#### 4.6 MARKER PREDICTION:

##### 4.6.1 Identification of SNPs Using QualitySNP

A total of 5940 contigs were analyzed, and about 1862 were predicted using the QualitySNP. Based on the type of SNPs these were further classified into synonymous SNPs and non-synonymous SNPs (Table 5). About 1789 SNPs were nonsynonymous SNPs. This means that all these SNPs will effect a change in the translated protein. About 73 SNPs were synonymous means that the mutations will not cause any change in the system. Again, based on the type of polymorphism these SNPs are further classified into Transitions, Transversions and InDels. About 650 Transitions, 540 Transversions and 672 Transversion were obtained. The total number of transitions - 650 was greater than the total number of transversions – 540 yielding a transition-to-transversion ratio of 1.20.

Table 6: Result of identified SNPs using QualitySNP

Total contigs analysed	5940
Total SNP detected	1862
Synonymous SNP	73
Non synonymous SNP	1789
Total transitions	650
Total transversion	540
Total indels	672

#### 4.6.2 Identification of SNPs Using AutoSNP

From 5940 contigs created from 19930 sequences, a total of 22707 SNPs were identified by AutoSNP. Based on the type of SNPs they are classified into Transitions, Transversions and finally InDels. A total of 8902 Transitions, 6952 Translations, and 2414 InDels were identified by AutoSNP. The Transition to Transversion ratio was 1.28

Table 7: Result of identified SNPs using AutoSNP

Total sequence analysed	19930
Total SNP detected	22707
Total transitions	8902
Total transversion	6952
Total indels	6852

#### 4.6.3 Comparative Evaluation of SNP Prediction Tools

SNP target prediction tools are implemented either in the form of a web server or as a standalone tool. Both AutoSNP and QualitySNP are offline prediction tools and both need Linux operating system to perform. The results of SNP target prediction tools: QualitySNP and AutoSNP are summarized in terms of types of polymorphism and their ability to predict SSRs (Table 9). Of the two tools considered for SNP prediction, the transition – transversion ratio in AutoSNP is 1.28 slightly higher when compared to QualitySNP which has a ratio of 1.20. Even though AutoSNP predict more SNPs, it fail to distinguish synonymous and non synonymous SNPs.

Table 8: Comparative evaluation of SNP prediction tools

Properties	AutoSNP	QualitySNP
No. of SNPs identified	22707	1862
No. of Synonymous SNP	-	1789
No. of Non Synonymous SNPs	-	73
Transition	8902	656
Transversion	6952	540
Indels	6853	672
Transition/Transversion	1.28	1.20

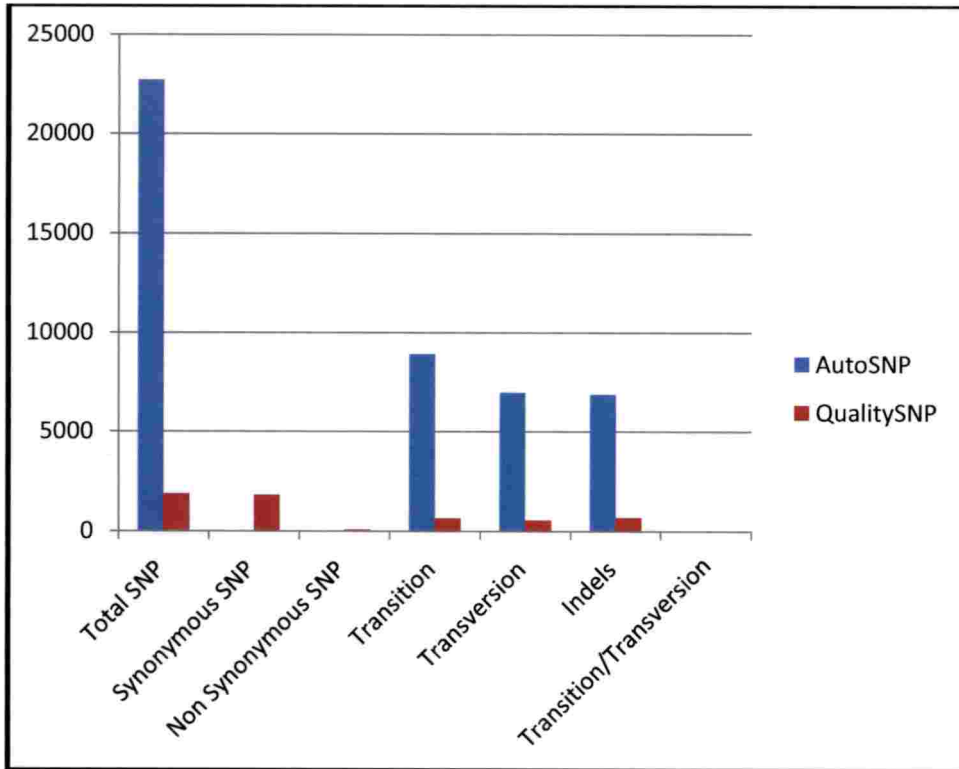


Figure 4: Comparative evaluation of SNP prediction tools

#### 4.6.3 Identification of SSRs Using MISA

From 5940 contigs created from 19930 sequences , 849 sequences contain SSRs. Total number of identified SSR is 1002. These identified SSRs were distributed in different types of repeat classes (Table 9).

Table 9: Distribution of different repeat classes in MISA

Unit Size	Number of SSR
Mono	359
Di	268
Tri	342
Tetra	17
Penta	7
Hexa	9

#### 4.6.4 Identification of SSRs Using SSRIT

From 5940 contigs created from 19930 sequences , 271 sequences contain SSRs. Total number of identified SSR is 295. These identified SSRs were distributed in different types of repeat classes (Table 10).

Table 10: Distribution of different repeat classes in SSRIT

Unit Size	Number of SSRIT
Mono	-
Di	92
Tri	186
Tetra	17
Penta	-
Hexa	-

#### 4.6.5 Comparative Evaluation of SSR Prediction Tools

SSR target prediction tools are implemented either in the form of a web server or as a standalone tool. MISA is an offline tool, but SSRIT is available in both online and offline mode. Also two tools need Linux operating system to carry out their performance. The results of SSR target prediction tools: MISA and SSRIT are summarized in terms of ability to predict SSRs (Table 10). The total number of SSRs predicted by MISA is very much higher than that of SSRIT. Also SSRIT fails to identify mono, penta, and hexa types of repeats.

Table 11: Comparative Evaluation of SSR Prediction Tools

SSR Types	SSRIT	MISA
Mono repeats	-	359
Di repeats	92	268
Tri repeats	186	342
Tetra repeats	17	17
Penta repeats	-	7
Hexa repeats	-	9
Total	295	1002

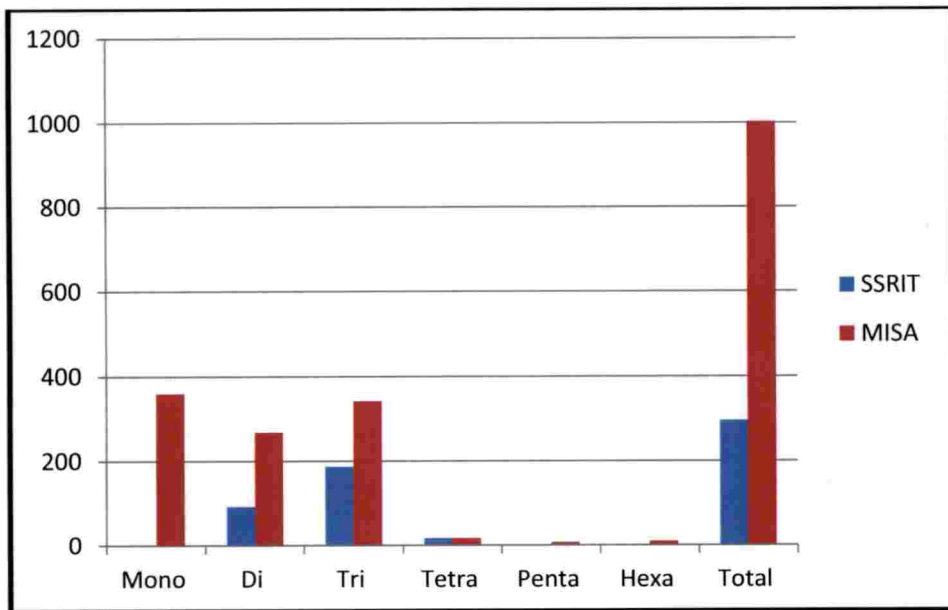


Figure 5: Comparative Evaluation of SSR Prediction Tools



#### 4.7 PRIMER DESIGNING

A total of 1789 SNPs and 1002 SSRs are predicted using QualitySNP and MISA respectively. But for validation only a few 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. Four SNP and SSR containing contigs were selected for primer designing using primer3 plus.

On the basis of GC content and melting temperature four primers have designed for SNP and SSR. All the selected SNPs and SSRs have a product ranging between 200 bp - 400 bp, and for SNPs have a product size ranging from 200 bp- 300 bp. (Table 12, Table 13).

#### 4.8 PRIMER SYNTHESIS

Primers were synthesized by a company called INTEGRATED DNA TECHNOLOGIES: THE CUSTOM BIOLOGY COMPANY. Forward and reverse primer of all 5 SNPs and 4 SSRs were synthesized and delivered. The annealing temperature for the primers were obtained from the Tm Calculator of Thermo Fisher Scientific.

Table 14: Annealing Temperature (Ta) of designed SSR Primers

Primer	Annealing Temperature (Ta)
DaSSR1	50 °C
DaSSR2	49 °C
DaSSR3	49 °C
DaSSR4	50 °C

Table 12: Designed primers for SNP

Sl.No	Primer	Forward Primer	Reverse Primer	Product Size
1	Da SNP1	CTCGTTGTTGTAAGGCAGAG	TCTGAGAGGTGGGAGCTTAT	303
2	Da SNP2	GAAAGAGGAGGTGAAAGTGG	CCCTGAAACACTCAAAGGAG	382
3	Da SNP3	CCTCAATACCCCTTGTCACCT	CCCCTGATCAGTTAGTGGAT	362
4	Da SNP4	CCGGAAGACTTCACTCAACA	GACTTGACGGTACATGACAGC	392
5	Da SNP5	GAGGTTCACGCAAAGGTCTA	GTTCCCTCAAGCTCTTCACCA	305

Table 13: Designed primers for SSR

Sl.No	Primer	Forward Primer	Reverse Primer	Product Size
1	Da SSR1	GAGTGATGAGGTACCGTGAG	AGAGCGTCGTAGATCAGAGA	223
2	Da SSR2	GCATGTCCAAGATGTCAGTC	TGCTAGACTAGACTGCTGCTG	277
3	Da SSR3	ATGGGACCATAGTGACAACC	GGCCAGATCATAACCACTTC	232
4	Da SSR4	GTCGCTAGGGTTAGGGTTTC	GAGATGCAAGACGATGAGGT	249

#### 4.10 DETERMINATION OF QUALITY OF DNA

Quality of DNA was determined by Agarose Gel Electrophoresis (Plate 1). Clear bands were observed in the gel. Using Nanospectrophotometer (DENOVIIX), the isolated DNA was quantified (Table 16). The samples were selected based on the better absorbance value or OD value.

Table 16: Result of quantification of DNA

Sample	Amount of DNA(Ng/ $\mu$ l)	OD at 260/230	OD at 260/280
Sree Karthika	805.417	1.24	2.13
Sree Keerthi	950.456	1.17	2.10
Sree Swathi	1440.069	1.80	2.25
Orissa Elite	1454.271	1.59	2.20
Sree Neelima	1433.603	1.68	2.26
Sree Roopa	1370.158	1.37	2.16

#### 4.11 VALIDATION OF SNPs

Validation of SNP was done in AGE using a susceptible and resistant samples. Of the five designed primers three primers (DaSNP1, DaSNP2 & DaSNP3 ) were succeed in giving prominent bands (Plate2. The rest two primers fail to give bands in the product size. On sequencing and sequence alignment using clustalw, there was SNP on the predicted sites.

Table 15: Annealing Temperature (Ta) of designed SNP Primers

Primer	Annealing Temperature (Ta)
DaSNP1	48
DaSNP2	48
DaSNP3	49
DaSNP4	50
DaSNP5	50

#### 4.9 DNA EXTRACTION

DNA was extracted from fresh and young leaves of collected samples using modified protocol of Raj *et al.* (2014).

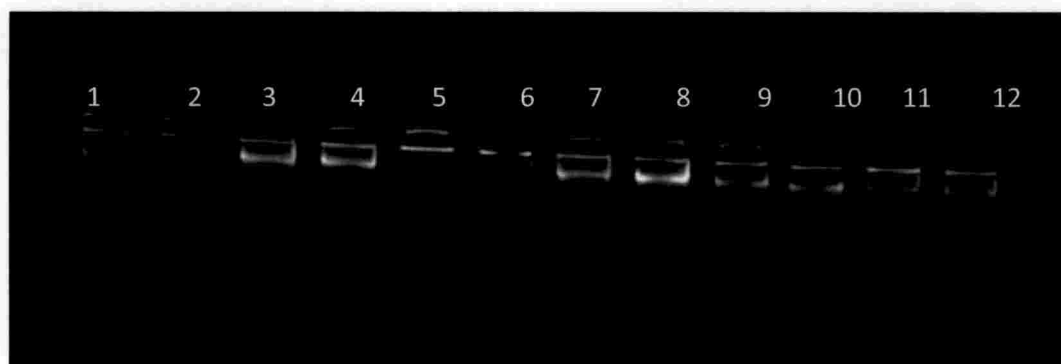


Plate 1. DNA bands in .8% agarose gel

Lane 1&2	Lane 3&4	Lane 5&6	Lane 7&8	Lane 9&10	Lane 11&12
Sree Karthika	Sree Keerthi	Sree Swathi	Orissa Elite	Sree Neelima	Sree Roopa



A: Sree Karthika (Resistant)

B: Orissa Elite (Susceptible)

Plate 2: SNP bands in 3% agarose gel

Lane 1	Lane 2&3	Lane 4&5	Lane 6&7	Lane 8&9	Lane 10&11
100bp ladder	DaSNP1	DaSNP2	DaSNP3	DaSNP4	DaSNP5



Figure 6: Clustalw result of DaSNP2 and DaSNP5

#### 4.12 VALIDATION OF SSRs

Validation of SSR was done in AGE using all three susceptible and three resistant samples. Of the four designed primers two primers (DaSSR1 & DaSSR2)

were discriminating primers (Plate3), ie those primers were able to differentiate the resistant and susceptible lines. The rest two primers gave thick bands. But fail to differentiate between resistant and susceptible varieties.



Plate3: DaSSR1 and DaSSR2 in 3% agarose gel

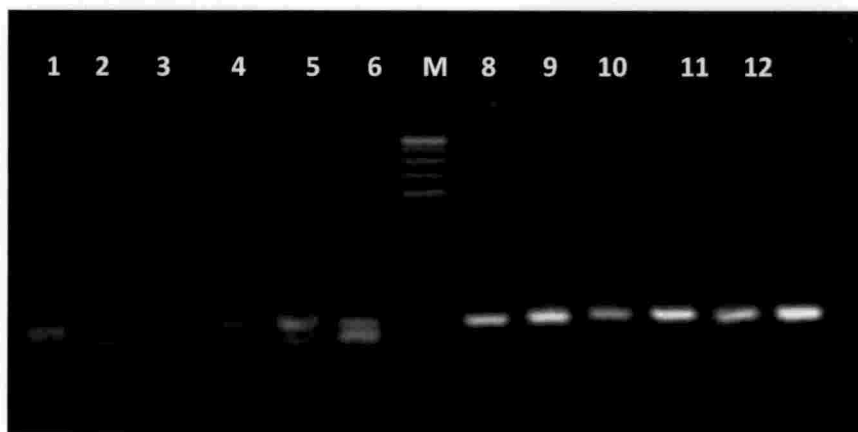


Plate4: DaSSR3 and DaSSR4 in 3% agarose gel

Lane1 & 8	Lane 2& 9	Lane 3& 10	Lane 4& 11	Lane 5& 12	Lane 6&1 3	Lane 7
Sree Karthika	Sree Keerthi	Sree Swathi	Orissa Elite	Sree Neelima	Sree Roopa	100bp ladder

# **DISCUSSION**



## 5. DISCUSSION

The study entitled “Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation” was conducted to to computationally identify SNPs and SSRs for anthracnose resistance in yam and to validate the predicted markers. The result of this study presented in chapter 4 are discussed here.

Yams (*Dioscorea* spp.) are a group of economically important multi-species crop that serve as a valuable source of food in tropical and sub-tropical countries across Africa, Asia, South America, the Caribbean and the Pacific (Coursey, 1976). Water yam (*Dioscorea alata*) is the most widely distributed species of the food yams globally (Onwueme, 1978). Among many advantages, it has a better agronomic flexibility than other popular yam species due to its ease of propagation and high multiplication ratio.

Yam anthracnose and constitute major pathological problems in *D. alata* production in all yam growing regions of the world. The disease is caused by *Colletotrichum gloesporioides*, and it has been identified as the most important biotic constraint to *D. alata* production worldwide (Emehute *et al.*, 1998). It can cause considerable damage in a large number of crops such as cereals, coffee, and legumes (Bailey *et al.*, 1992) and even in human subcutaneous hyalohyphomycosis (Guarro *et al.*, 1998). Anthracnose disease causes leaf necrosis and shoot die-back of yams thus reducing the photosynthetic efficiency of the plant with resultant yield losses of over 90% in susceptible genotypes (Winch *et al.*, 1984; Green, 1994). The rate of disease through sequence of spatial patterns was affected by the susceptibility of the cultivar, age of leaf, age of vine, stage of the epidemic, rainfall, and agronomic practices such as fungicide application (Sweetmore *et al.*, 1994). Anthracnose and virus diseases exert devastating impacts on yam production in many tropical regions of the world where the crop contributes to food security and income generation. The

complexities in their epidemiology necessitate the use of integrated approaches in their management (Egesi *et al.*, 2006).

The use of resistant yam genotypes as a component of an integrated disease management approach is a control measure for yam diseases in the field (Nwankiti *et al.*, 1987; Amusa *et al.*, 2003). Recent studies showed the existence of reliable sources of anthracnose resistance in the *D. alata* germplasm held at IITA, Nigeria and at the Institut National de la Recherche Agronomique (INRA), French West Indies (Mignouna *et al.*, 2001; Egesi *et al.*, 2004; Onyeka *et al.*, 2006). Combining host plant resistance with good cultural practices will provide inexpensive and easy-to-adopt disease management for yam farmers. Although some level of control of both anthracnose and virus diseases can be achieved by removal of infected plant materials, the role of cultural practices in the field management of yam diseases is generally not clear (Emehute *et al.*, 1998). However, initial results suggest that most cultural practices such as delayed planting tend to increase anthracnose symptoms in the field (IITA, 1983). The effect of planting date on disease development and associated yield loss has also been demonstrated for other plant diseases (Krell *et al.*, 2005; Matheron *et al.*, 2005).

Over the past few decades, plant genomics research has been studied extensively bringing about a revolution in the field of plant biotechnology. Molecular markers, useful for plant genome analysis, have now become an important tool in crop improvement. The development and use of molecular markers for the detection and exploitation of DNA polymorphism is one of the most significant developments in the field of molecular genetics. The presence of various types of molecular markers, and differences in their principles, methodologies and applications require careful consideration in choosing one or more of such methods. DNA-based molecular markers are a versatile tool in the fields of taxonomy, physiology, embryology, genetic engineering, etc. (Schlotterer 2004). They are no longer looked upon in simple DNA fingerprinting markers in variability studies or in mere forensic tools. Ever since the

development of molecular markers, these are constantly being modified to enhance the utility and to bring about automation in the process of genome analysis. The discovery of PCR (polymerase chain reaction) was a landmark in this effort and proved to be a unique process that brought about a new class of DNA profiling markers. This facilitated the development of marker based gene tags, genetic mapping, map-based cloning of agronomically important genes, genetic diversity studies, phylogenetic analysis, and marker-assisted selection of desirable genotypes etc. (Joshi *et al.* 2000). Thus, giving new dimensions to breeding and marker-aided selection, that can reduce the time span of developing new and better varieties and the dream of super varieties come true. These DNA markers offer several advantages over traditional phenotypic markers, as they provide data that can be analyzed objectively. The existence of various molecular techniques and differences in their principles and methodologies require careful consideration in choosing one or more of such marker types.

Molecular markers are important tools for applications such as estimating genetic diversity and phylogenetic relationships, cultivar identification, mapping of major genes and QTLs, assessing population structure, selection of desirable genotypes in breeding programs, and for authentication of progenies obtained from genetic crosses (Tamiru *et al.*, 2015). They play imperative role in plant breeding and crop improvement. They help to alter and improve plant traits on the basis of genotype assays, that describe several modern breeding strategies including marker assisted selection (MAS) ,marker assisted backcrossing (MABC) etc (Rafalski *et al.*, 2002). Availability of high quality sequence information is necessary for designing molecular markers associated with resistance.

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 diversity and molecular marker development in yam have been published (Arnau G *et al.*, 2017, Tamiru M *et al.*, 2015) and several studies have focused on the development of indels, and SNPs

for various applications in Guinea yam, including linkage mapping, genome-wide association analysis, genomic selection, and MAS. Development of DNA markers linked to agronomically important traits and their use for MAS increase the role yam plays in ensuring food security for resource-poor households in Africa and beyond (Tamiru M *et al.*, 2017). Thus SSR and SNP markers have important role in plant breeding and crop improvement when compared to other markers.

In this work about 1862 SNPs and 1002 SSRs are predicted which is exclusively related to anthracnose resistance in yam. These can be validated and screened for effective markers against anthracnose resistance. More than 560 SNPs are confirmed in the coding region which makes them candidate SNPs for screening for resistance against anthracnose. About 1789 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. Also QualitySNP have 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.

#### 5.2 COMPARATIVE EVALUATION OF SSR PREDICTION TOOLS

On comparative evaluation of MISA and SSRIT, MISA shows more promising SSRs 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.

### 5.3 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), in cassava against Anthracnose. Polymorphisms (SNPs and InDels) were discovered from the contig sequence alignment according to the certain criteria. As a result they were able to discover a total of 10546 SNPs and 674 InDels from the whole genome.

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 “Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation” was conducted at the ICAR-Central Tuber Crop Research Institute during 2017-2018. The objectives of the study is to computationally identify SNPs and SSRs for anthracnose resistance in Greater Yam and the verification of identified markers using resistant and susceptible varieties. The salient findings of the study are summarized below.

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>). After pre-processing and screening, the dataset was reduced from 44134 to 44114 sequences. Since the sequences were compared with the anthracnose resistant genes under the screening stage itself lead to the significant reduction in time taken for the identification of SSRs and SNPs. The resulting sequences were assembled and aligned using CAP3 and 5940 contigs were obtained. From these contigs using QualitySNP, about 1862 SNPs were identified. In that 1789 SNPs were nonsynonymous and 73 SNPs were synonymous SNPs. From that best five sequences were selected based on the percentage identity and E value, for primer designing. About 1002 SSR were identified using MISA. In that 359 were mono, 268 were di, 342 were tri, 17 were tetra, 7 was penta and 9 was hexa. Five sequences which have high hit percentage and low E value were selected for validation and primer designing. Primers were designed for both SNPs and SSRs. These primers were validated using 3 resistant and 3 susceptible yam varieties. Among the primers after validation in wet lab, three SNPs (DaSNP1, DaSNP2, DaSNP3) and two SSRs (DaSSR1 and DaSSR2) 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 anthracnose resistance in yam.



Different tools for the prediction of SNP and SSR were also compared as a part of the work. The result was like this. The SNP prediction tool QualitySNP was found to be a better tool compared to AutoSNP. Because QualitySNP had better SNP prediction algorithm and the ability for classification of the identified SNPs into various categories. Also it has the ability to annotate and identify nonsynonymous and synonymous SNPs which helps to select more precise SNPs for the research work. For the prediction of SSRs, the 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.

#### 6.1 SCOPE FOR FUTURE WORK

As the resources were limited only few predicted SSRs and SNPs were validated for differentiating susceptible and resistant genes in Greater yam. In future, the identified 1789 SNPs and 1002 SSRs can be validated in wet lab and the resulting potential markers can be utilized in the breeding program for screening anthracnose resistance in yam.



# REFERENCES

## 7. REFERENCES

- Abang, M.M., Winter, S., Green, K.R., Hoffmann, P., Mignouna, H.D. and Wolf, G.A., 2002. Molecular identification of *Colletotrichum gloeosporioides* causing yam anthracnose in Nigeria. *Plant Pathol.* 51(1): 63-71.
- Abang, M.M., Winter, S., Mignouna, H.D., Green, K.R. and Asiedu, R., 2003. Molecular taxonomic, epidemiological and population genetic approaches to understanding yam anthracnose disease. *Afr. J. of Biotechnol.* 2(12) : 486-496.
- Abraham, K. and Nair, P.G., 1990. Floral biology and artificial pollination in *Dioscorea alata* L. *Plant Breed Rev.* 48(1): 45-51.
- Agarwal, M., Shrivastava, N. and Padh, H., 2008. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell rep.* 27(4): 617-631.
- Aggarwal, R.K., Hendre, P.S., Varshney, R.K., Bhat, P.R., Krishnakumar, V. and Singh, L., 2007. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor. and Appl. Genet.* 114(2): 359-72.
- Al-Samarai, F.R. and Al-Kazaz, A.A., 2015. Molecular markers: An introduction and applications. *Eur. J. of Mol. Biotechnol.* (3): 118-130.
- Amusa, N.A., Adigbite, A.A., Muhammed, S. and Baiyewu, R.A., 2003. Yam diseases and its management in Nigeria. *Afr. J. of Biotechnol.*, 2(12): 497-502.
- Arnau, G., Bhattacharjee, R., Sheela, M.N., Malapa, R., Lebot, V., Abraham, K., Perrier, X., Petro, D., Penet, L. and Pavis, C., 2017. Understanding the genetic diversity and population structure of yam (*Dioscorea alata* L.) using microsatellite markers. *PLoS Biol.* 12: 3-29.

- Arnau, G., Némorin, A., Maledon, E. and Abraham, K., 2009. Revision of ploidy status of *Dioscorea alata* L.(Dioscoreaceae) by cytogenetic and microsatellite segregation analysis. *Theor. Appl. Genet.* 118(7) : 1239-1249.
- Arnau, G., Nemorin, A., Maledon, E. and Nudol, E., 2011. Advances on polyploid breeding in yam. *Plant Sci.* 211: 52 – 60.
- Ayensu, E.S. and Coursey, D.G., 1972. Guinea yams the botany, ethnobotany, use and possible future of yams in West Africa. *J. Biol. Chem.* 26(4): 301-318.
- Bailey, J.A., 1992. Colletotrichum; biology, pathology and control. *Plant Pathol.* 94:538-544.
- 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 physiol.* 132(1): 84-91.
- Beier, S., Thiel, T., Münch, T., Scholz, U. and Mascher, M., 2017. MISA-web: a web server for microsatellite prediction. *BMC Bioinforma.* 33(16): 2583-2585.
- Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J. and Damborsky, J., 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS comput. biol.* 10(1), p.e1003440.
- Bousalem, M., Arnau, G., Hochu, I., Arnolin, R., Viader, V., Santoni, S. and David, J., 2006. Microsatellite segregation analysis and cytogenetic evidence for tetrasomic inheritance in the American yam *Dioscorea trifida* and a new basic chromosome number in the Dioscoreae. *Theoret. Appl. Genet.* 113(3): 439-451.

- Bressan, E.D.A., Briner Neto, T., Zucchi, M.I., Rabello, R.J. and Veasey, E.A., 2011. Morphological variation and isozyme diversity in *Dioscorea alata* L. landraces from Vale do Ribeira, Brazil. *New Biotechnol.* 68(4): 494-502.
- Bromberg, Y. and Rost, B., 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 35(11): 3823-3835.
- Butler, J.M., 2005. *Forensic DNA typing: biology, technology, and genetics of STR markers.* *Plant J.* 43: 815- 830
- Collard, B.C. and Mackill, D.J., 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *J. Exp. Bot.* 363(1491): 557-572.
- Cordeiro, G.M., Casu, R., McIntyre, C.L., Manners, J.M. and Henry, R.J., 2001. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci.* 160(6): 1115-1123.
- Coursey, D.G., 1976. Yams: *Dioscorea* spp.(Dioscoreaceae). *Evol. Of Crop Plants. NW.* 278: 779-785.
- Da Maia, L.C., Palmieri, D.A., De Souza, V.Q., Kopp, M.M., de Carvalho, F.I.F. and Costa de Oliveira, A., 2008. SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *Int. J. of Plant Genomics.* 32: 54-65.
- Dansi, A., Mignouna, H.D., Zoundjihékpon, J., Sangare, A., Ahoussou, N. and Asiedu, R., 2000. Identification of some Benin Republic's Guinea yam (*Dioscorea cayenensis/Dioscorea rotundata* complex) cultivars using randomly amplified polymorphic DNA. *Genet. Res. Crop Evol.* 47(6): 619-625.

- Dufie, W.M.F., Oduro, I., Ellis, W.O., Asiedu, R. and Maziya-Dixon, B., 2013. Potential health benefits of water yam (*Dioscorea alata*). *Food & function*. 4(10): 1496-1501.
- Duran, C., Appleby, N., Clark, T., Wood, D., Imelfort, M., Batley, J. and Edwards, D., 2008. AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *New Biotechnol.* 37: D951-D953.
- Egesi, C.N., Asiedu, R., Ude, G., Ogunyemi, S. and Egunjobi, J.K., 2006. AFLP marker diversity in water yam (*Dioscorea alata* L.). *Plant Genet. Resour.* 4(3): 181-187.
- Egesi, C.N., Ogunyemi, S. and Asiedu, R., 2004, November. Evaluation of water yam (*Dioscorea alata* L.) genotypes for reaction to yam anthracnose disease. *Plant Pathol.* 54: 580-592
- Egesi, C.N., Onyeka, T.J. and Asiedu, R., 2007. Severity of anthracnose and virus diseases of water yam (*Dioscorea alata* L.) in Nigeria I: effects of yam genotype and date of planting. *Plant Pathol.* 26(8): 1259-1265.
- Green, K.R., 1994. Studies on the epidemiology and control of yam anthracnose. *Plant J.* 152: 1219-1250.
- Grivet, L., Glaszmann, J.C., Vincentz, M., Da Silva, F. and Arruda, P., 2003. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. *Theor. and Appl. Genet.* 106(2): 190-197.
- Guarro, J., Svidzinski, T.E., Zaror, L., Forjaz, M.H., Gené, J. and Fischman, O., 1998. Subcutaneous Hyalohyphomycosis Caused by *Colletotrichum gloeosporioides*. *J. of Clin. Microbiol.* 36(10): 3060-3065.
- Guryev, V., Berezikov, E., Malik, R., Plasterk, R.H. and Cuppen, E., 2004. Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res.* 14(7): 1438-1443.

- Hahn, S.K., Isoba, J.C. and Ikotun, T., 1989. Resistance breeding in root and tuber crops at the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. *Plant Pathol.* 8(3): 147-168.
- Heesacker, A., Kishore, V.K., Gao, W., Tang, S., Kolkman, J.M., Gingle, A., Matvienko, M., Kozik, A., Michelmore, R.M., Lai, Z. and Rieseberg, L.H., 2008. SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. *Theor. and Appl. Genet.* 117(7): 1021-1029.
- He, G., Meng, R., Newman, M., Gao, G., Pittman, R.N. and Prakash, C.S., 2003. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). *BMC plant boil.* 3(1): 3.
- Hochu, I., Santoni, S. and Bousalem, M., 2006. Isolation, characterization and cross- species amplification of microsatellite DNA loci in the tropical American yam *Dioscorea trifida*. *Mol. Ecol.* 6(1): 137-140.
- Joshi, S.P., Gupta, V.S., Aggarwal, R.K., Ranjekar, P.K. and Brar, D.S., 2000. Genetic diversity and phylogenetic relationship as revealed by inter simple sequence repeat (ISSR) polymorphism in the genus *Oryza*. *Theor. and Appl. Genet.* 100(8): 1311-1320
- Kantety, R.V., La Rota, M., Matthews, D.E. and Sorrells, M.E., 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.* 48(56): 501-510.
- Kashi, Y. and King, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. *Plant Cell.* 22(5): 253-259.
- Khlestkina, E.K., Than, M.H.M., Pestsova, E.G., Röder, M.S., Malyshev, S.V., Korzun, V. and Börner, A., 2004. Mapping of 99 new microsatellite-derived loci in rye (*Secale cereale* L.) including 39 expressed sequence tags. *Theoretical and Appl. Genet.* 109(4): 725-732.

- Kim, H., Schmidt, C.J., Decker, K.S. and Emara, M.G., 2003. A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. *Anim. Genet.* 34(4): 249-254.
- Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J. and Graner, A., 2001. Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). *Hereditas.* 135(23): 145-151.
- Krell, R.K., Pedigo, L.P., Rice, M.E., Westgate, M.E. and Hill, J.H., 2005. Using planting date to manage bean pod mottle virus in soybean. *Crop Prot.* 24(10): 909-914.
- Lebot, V., 2009. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. *Plant physiol.* 112: 255-264.
- Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D. and Hallauer, A., 2002. Expanding the genetic map of maize with the intermated B73× Mo17 (IBM) population. *Plant Mol. Biol.* 48(56): 453-461.
- Liang, X., Chen, X., Hong, Y., Liu, H., Zhou, G., Li, S. and Guo, B., 2009. Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. *BMC Plant Biol.* 9(1): 35.
- López-Ferrando, V., Gazzo, A., de la Cruz, X., Orozco, M. and Gelpí, J.L., 2017. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucl. Acids Res.* 45: 222-228.
- Malapa, R., 2005. *Description de la diversité de Dioscorea alata L. Du Vanouatou à l'aide de marqueurs agro-morphologiques et moléculaires (AFLP): Relations avec les autres espèces de la section Enanthiophyllum* (Doctoral dissertation, Rennes, Agrocampus Ouest).
- Martins, W.S., Lucas, D.C.S., de Souza Neves, K.F. and Bertioli, D.J., 2009. WebSat- A web software for microsatellite marker development. *BMC Biotechnol.* 3(6): 282.

- Matthies, I.E., van Hintum, T., Weise, S. and Röder, M.S., 2012. Population structure revealed by different marker types (SSR or DArT) has an impact on the results of genome-wide association mapping in European barley cultivars. *Mol Breed.* 30(2): 951-966.
- Matheron, M.E., McCreight, J.D. and Tickes, B.R., 2005. Effect of planting date, cultivar, and stage of plant development on incidence of Fusarium wilt of lettuce in desert production fields. *Plant Dis.* 89(6). 565-570.
- Miah, G., Rafii, M.Y., Ismail, M.R., Puteh, A.B., Rahim, H.A., Islam, K.N. and Latif, M.A., 2013. A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. *Int. J. of Mol. Biol.* 14(11): 22499-22528.
- Mignouna, H.D., Abang, M.M. and Asiedu, R., 2003. Harnessing modern biotechnology for tropical tuber crop improvement: Yam (*Dioscorea spp.*) molecular breeding. *Afr. J. of Biotechnol.* 2(12): 478-485.
- Mignouna, H.D., Abang, M.M. and Fagbemi, S.A., 2003. A comparative assessment of molecular marker assays (AFLP, RAPD and SSR) for white yam (*Dioscorea rotundata*) germplasm characterization. *Ann. of Appl. Biol.* 142(3): 269-276.
- Mignouna, H.D., Abang, M.M., Green, K.R. and Asiedu, R., 2001. Inheritance of resistance in water yam (*Dioscorea alata*) to anthracnose (*Colletotrichum gloeosporioides*). *Theor. and Appl. Genet.* 103(1): 52-55.
- Mignouna, H.D., Abang, M.M., Onasanya, A. and Asiedu, R., 2002. Identification and application of RAPD markers for anthracnose resistance in water yam (*Dioscorea alata*). *Ann. of Appl. Biol.* 141(1): 61-66.
- Mizuki, I., Tani, N., Ishida, K. and Tsumura, Y., 2005. Development and characterization of microsatellite markers in a clonal plant, *Dioscorea japonica* Thunb. *Mol. Ecol.* 5(4): 721-723.



- Narina, S.S., Buyyarapu, R., Kottapalli, K.R., Sartie, A.M., Ali, M.I., Robert, A., Hodeba, M.J., Sayre, B.L. and Scheffler, B.E., 2011. Generation and analysis of expressed sequence tags (ESTs) for marker development in yam (*Dioscorea alata* L.). *BMC genomics*. 12(1): 100.
- Nwankiti, A.O., Okoli, O.O. and Okpala, E.U., 1987. Screening of water yam (*Dioscorea alata*) cultivars for tolerance to anthracnose/blotch disease. *Plant Pathol.* 12(1): 36-39.
- Obidiegwu, J.E., Kolesnikova-Allen, M., Ene-Obong, E.E., Muoneke, C.O. and Asiedu, R., 2009. SSR markers reveal diversity in Guinea yam (*Dioscorea cayenensis*/D. *rotundata*) core set. *Afr. J. of Biotechnol.* 8(12): 2730-2739.
- Onwueme, I.C., 1978. The tropical tuber crops: yams, cassava, sweet potato, and cocoyams. *Plant Physiol.* 54: 569-575.
- Onyeka, T.J., Petro, D., Ano, G., Etienne, S. and Rubens, S., 2006. Resistance in water yam (*Dioscorea alata*) cultivars in the French West Indies to anthracnose disease based on tissue culture- derived whole- plant assay. *Plant pathol.* 55(5): 671-678.
- Orkwor, G.C., Asiedu, R. and Ekanayake, I.J., 1998. *J. Bio. Chem.* 26: 860-866.
- Otoo, E., Anokye, M., Asare, P.A. and Tetteh, J.P., 2015. Molecular categorization of some water yam (*Dioscorea alata* L.) germplasm in Ghana using microsatellite (SSR) markers. *J. of Agric.l Sci.* 7(10): 225.
- Palaniyandi, S.A., Yang, S.H., Cheng, J.H., Meng, L. and Suh, J.W., 2011. Biological control of anthracnose (*Colletotrichum gloeosporioides*) in yam by *Streptomyces* sp. *J. of Appl. Microbiol.* 11(2): 443-455.
- Pashley, C.H., Ellis, J.R., McCauley, D.E. and Burke, J.M., 2006. EST databases as a source for molecular markers: lessons from *Helianthus*. *J. of heredity.* 97(4): 381-388.

- Pinto, L.R., Oliveira, K.M., Ulian, E.C., Garcia, A.A.F. and De Souza, A.P., 2004. Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *BMC Biotechnol.* 47(5): 795-804.
- Poncet, V., Rondeau, M., Tranchant, C., Cayrel, A., Hamon, S., De Kochko, A. and Hamon, P., 2006. SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the *Coffea* genus. *Mol. Genet. and Genomics.* 276(5): 436-449.
- Poornima, G.N. and Ravishankar, R.V., 2007. In vitro propagation of wild yams, *Dioscorea oppositifolia* (Linn) and *Dioscorea pentaphylla* (Linn). *Afr. J. Biotechnol.* 6(20): 21-32.
- Rafalski, A., 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opi. in Plant Biol.* 5(2): 94-100.
- Raj, M., Nath, V.S., Senthil@ Sankar, M., Jeeva, M.L. and Hegde, V., 2014. Rapid isolation of DNA from *Dioscorea* species suitable for PCR, restriction digestion and pathogen screening. *Arch. Phytopathol. Plant Prot.* 47(6), pp.753-760.
- Rickert, A.M., Kim, J.H., Meyer, S., Nagel, A., Ballvora, A., Oefner, P.J. and Gebhardt, C., 2003. First- generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. *Plant Biotechnol. J.* 1(6): 399-410.
- Rudd, S., 2003. Expressed sequence tags: alternative or complement to whole genome sequences. *BMC Plant Biol.* 8(7): 321-329.
- Sakurai, T., Mochida, K., Yoshida, T., Akiyama, K., Ishitani, M., Seki, M. and Shinozaki, K., 2013. Genome-wide discovery and information resource development of DNA polymorphisms in cassava. *PloS one*, 8(9): 25-31.

- Sartie, A. and Asiedu, R., 2014. Segregation of vegetative and reproductive traits associated with tuber yield and quality in water yam (*Dioscorea alata* L.). *Afr. J. Biotechnol.* 13(28): 2807–2818.
- Sato, S., Isobe, S., Asamizu, E., Ohmido, N., Kataoka, R., Nakamura, Y., Kaneko, T., Sakurai, N., Okumura, K., Klimenko, I. and Sasamoto, S., 2005. Comprehensive structural analysis of the genome of red clover (*Trifolium pratense* L.). *DNA Res.* 12(5): 301-364.
- Scarcelli, N., Daïnou, O., Agbangla, C., Tostain, S. and Pham, J.L., 2005. Segregation patterns of isozyme loci and microsatellite markers show the diploidy of African yam *Dioscorea rotundata* ( $2n= 40$ ). *Theor. and Appl. Genet.* 111(2): 226-232.
- Schlötterer, C., 2004. The evolution of molecular markers—just a matter of fashion. *J. Nat.* 5(1): 63.
- Sheela, M.N., Abhilash, P.V., Asha, K.I. and Arnau, G., 2014, August. Genetic diversity analysis in greater yam (*Dioscorea alata* L.) native to India using morphological and molecular markers. In *XXIX Int. Hortic. Cong. on Hortic.* 1118: 51-58.
- Siqueira, M.V., 2011. Yam: a neglected and underutilized crop in Brazil. *Hortic. Br.* 29(1): 16-20.
- Sweetmore, A., Simons, S.A. and Kenward, M., 1994. Comparison of disease progress curves for yam anthracnose (*Colletotrichum gloeosporioides*). *Plant Pathol.* 43(1): 206-215.
- Syvänen, A.C., 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* 2(12): 930.
- Tamiru, M., Becker, H.C. and Maass, B.L., 2007. Genetic diversity in yam germplasm from Ethiopia and their relatedness to the main cultivated

- Dioscorea species assessed by AFLP markers. *Crop Sci.* 47(4): 1744-1753.
- Tamiru, M., Natsume, S., Takagi, H., White, B., Yaegashi, H., Shimizu, M., Yoshida, K., Uemura, A., Oikawa, K., Abe, A. and Urasaki, N., 2017. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC boil.* 15(1): 86.
- Tamiru, M., Yamanaka, S., Mitsuoka, C., Babil, P., Takagi, H., Lopez-Montes, A., Sartie, A., Asiedu, R. and Terauchi, R., 2015. Development of genomic simple sequence repeat markers for Yam. *Crop Sci.* 55(5): 2191-2200.
- Tang, J., Vosman, B., Voorrips, R.E., van der Linden, C.G. and Leunissen, J.A., 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinforma.* 7(1): 438.
- Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Proc. Natl. Acad. Sci.* 11(8): 1441-1452.
- Thiel, T., Michalek, W., Varshney, R. and Graner, A., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. and appl. genet.* 106(3): 411-422.
- Thomas, E.E., 2005. Short, local duplications in eukaryotic genomes. *Nucleic Acids Res.* 15(6): 640-644.
- Tilman, D. and Clark, M., 2015. Food, Agriculture & the environment: Can we feed the world & save the Earth. *Plant Pathol.* 144(4): 8-23.

- Tostain, S., Agbangla, C., Scarcelli, N., Mariac, C., Daïnou, O., Berthaud, J. and Pham, J.L., 2007. Genetic diversity analysis of yam cultivars (*Dioscorea rotundata* Poir.) in Benin using simple sequence repeat (SSR) markers. *Plant Genet. Res.* 5(2): 71-81.
- Useche, F.J., Gao, G., Hanafey, M. and Rafalski, A., 2001. High-throughput identification, database storage and analysis of SNPs in EST sequences. *BMC Bioinforma.* 12: 194-203.
- Varshney, R.K., Graner, A. and Sorrells, M.E., 2005. Genomics-assisted breeding for crop improvement. *Trends in plant Sci.* 10(12): 621-630.
- Vignal, A., Milan, D., SanCristobal, M. and Eggen, A., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *J. Biol. Chem.* 34(3): 275.
- Wang, X., Lu, P. and Luo, Z., 2013. GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. *BMC Biotechnol.* 9(10): 541.
- Wilkin, P., Schols, P., Chase, M.W., Chayamarit, K., Furness, C.A., Huysmans, S., Rakotonasolo, F., Smets, E. and Thapayai, C., 2005. A plastid gene phylogeny of the yam genus, *Dioscorea*: roots, fruits and Madagascar. *Syst. Bot.* 30(4): 736-749.
- Winch, J.E., Newhook, F.J., Jackson, G.V.H. and Cole, J.S., 1984. Studies of *Colletotrichum gloeosporioides* disease on yam, *Dioscorea alata*, in Solomon Islands. *Plant pathol.* 33(4): 467-477.
- Yu, J.K., La Rota, M., Kantety, R.V. and Sorrells, M.E., 2004. EST derived SSR markers for comparative mapping in wheat and rice. *Mol. Genet. Genomics.* 271(6): 742-751.
- Zalapa, J.E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., McCown, B., Harbut, R. and Simon, P., 2012. Using next- generation sequencing

approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. of bot.* 99(2): 193-208.

Zannou, A., Agbicodo, E., Zoundjihékpon, J., Struik, P.C., Ahanchédé, A., Kossou, D.K. and Sanni, A., 2009. Genetic variability in yam cultivars from the Guinea-Sudan zone of Benin assessed by random amplified polymorphic DNA. *Afr. J. of Biotechnol.*

# **APPENDICES**

**APPENDIX I****CTAB Extraction Buffer**

Tris HCl (pH 8.0) 100mM

EDTA 25mM

NaCl 1.5 M

CTAB 2.5%

$\beta$ - Mercaptoethanol 0.2% (v/v)

PVP 1% (w/v)

**APPENDIX II****TE Buffer (10X)**

Tris – HCl (pH 8.0) 10 mM

EDTA 1 mM

**APPENDIX III****TBE Buffer (10 X)**

Tris base 107g

Boric acid 55g

0.5 M EDTA (pH 8.0) 40ml

Final volume made up to 1000ml with distilled water and autoclave before use.



**APPENDIX IV****Wash solution**

Ammonium acetate 15mM

Ethanol 70%

**APPENDIX V****Chloroform: isoamyl alcohol**

Chloroform 24 ml

Isoamyl alcohol 1ml

Mix 24 parts of chloroform with 1 part of isoamyl alcohol and stor in tightly capped bottle

**APPENDIX VI****80% ethanol**

100% ethanol 80 ml

Distilled water 20 ml

## APPENDIX VII

## LIST OF SSR IDENTIFIED BY MISA

ID	SSR	SSR TYPE	SSR	Size	Start	End
Contig10	1	p1	(T)11	11	28	38
Contig13	1	p3	(CAC)6	18	39	56
Contig19	1	p2	(TA)8	16	106	121
Contig36	1	p1	(T)11	11	26	36
Contig38	1	p3	(CAA)5	15	47	61
Contig42	1	p1	(T)10	10	26	35
Contig42	2	p1	(A)12	12	206	217
Contig53	1	p1	(A)10	10	30	39
Contig53	2	p3	(TCC)5	15	314	328
Contig56	1	p2	(AT)15	30	776	805
Contig64	1	p1	(T)11	11	24	34
Contig71	1	p1	(A)11	11	456	466
Contig80	1	p3	(GAA)7	21	44	64
Contig81	1	p3	(ATC)8	24	252	275
Contig88	1	p1	(A)11	11	1450	1460
Contig93	1	p2	(TA)11	22	1219	1240
Contig96	1	p3	(TTG)5	15	830	844
Contig102	1	p3	(TCT)7	21	153	173
Contig106	1	p3	(TAT)7	21	219	239
Contig116	1	p4	(AAAG)5	20	562	581
Contig130	1	p2	(AT)7	14	395	408
Contig130	2	p1	(A)11	11	583	593
Contig134	1	p3	(AAG)7	21	105	125
Contig135	1	p2	(AG)6	12	69	80
Contig135	2	p1	(A)10	10	716	725
Contig141	1	p1	(A)11	11	1140	1150
Contig143	1	p1	(T)11	11	29	39
Contig146	1	p1	(A)10	10	511	520
Contig150	1	p1	(T)11	11	24	34
Contig153	1	p1	(A)10	10	1439	1448
Contig160	1	p3	(AAG)6	18	234	251
Contig166	1	p1	(A)11	11	527	537
Contig172	1	p3	(CTT)7	21	49	69
Contig174	1	p2	(GA)7	14	851	864
Contig181	1	p4	(CTCG)5	20	704	723

Contig192	1	p3	(CCG)7	21	341	361
Contig197	1	p3	(AAG)7	21	69	89
Contig197	2	p1	(A)14	14	637	650
Contig199	1	p1	(A)24	24	561	584
Contig203	1	p1	(A)12	12	502	513
Contig210	1	p1	(T)10	10	21	30
Contig210	2	p1	(A)13	13	242	254
Contig212	1	p1	(A)11	11	522	532
Contig219	1	p3	(ACA)5	15	467	481
Contig227	1	p3	(TAT)5	15	431	445
Contig235	1	p3	(GGA)5	15	19	33
Contig244	1	c	(A)13cttccc(T) 14	33	114	146
Contig244	2	p3	(CAG)5	15	296	310
Contig247	1	p3	(CAC)5	15	677	691
Contig250	1	p3	(TCT)6	18	24	41
Contig257	1	p1	(T)10	10	140	149
Contig257	2	p1	(A)10	10	274	283
Contig261	1	p3	(AGA)5	15	378	392
Contig272	1	p3	(TAA)11	33	50	82
Contig283	1	p2	(TC)7	14	22	35
Contig285	1	p1	(T)11	11	399	409
Contig294	1	p3	(ATG)5	15	614	628
Contig294	2	p1	(T)10	10	768	777
Contig297	1	p3	(CTC)7	21	120	140
Contig298	1	c	(TC)8(TA)6	28	58	85
Contig299	1	p2	(TC)6	12	66	77
Contig306	1	p3	(CAC)5	15	85	99
Contig308	1	p1	(T)13	13	197	209
Contig318	1	p3	(TCT)5	15	66	80
Contig326	1	p3	(GAA)6	18	34	51
Contig329	1	p3	(TTA)6	18	550	567
Contig332	1	p3	(GCA)5	15	330	344
Contig337	1	c	(AAG)5at(CA A)5gaacaagaa agagcttcatcaat gaacaagtttagaa gatcttagctaatta tataactcaagtgtaa ctgttccatttacac actttt(TA)8	137	373	509

Contig345	1	p2	(AG)10	20	516	535
Contig346	1	p1	(T)10	10	794	803
Contig351	1	p1	(A)10	10	90	99
Contig352	1	p1	(A)12	12	968	979
Contig357	1	p2	(CT)14	28	718	745
Contig359	1	p2	(TA)6	12	603	614
Contig387	1	p1	(T)11	11	565	575
Contig388	1	p1	(A)10	10	19	28
Contig391	1	p3	(GAA)5	15	573	587
Contig394	1	p3	(GCA)5	15	962	976
Contig399	1	p1	(A)10	10	749	758
Contig401	1	p2	(TA)25	50	1304	1353
Contig404	1	p1	(A)10	10	1118	1127
Contig409	1	p3	(CCG)5	15	25	39
Contig414	1	p2	(AG)6	12	47	58
Contig425	1	p2	(TA)9	18	115	132
Contig436	1	p3	(TCT)6	18	85	102
Contig438	1	p2	(TA)9	18	789	806
Contig441	1	p3	(GAT)12	36	442	477
Contig470	1	p1	(T)11	11	24	34
Contig473	1	p3	(AAC)11	33	98	130
Contig473	2	c	(C)11lacaaaaac aatccaagatcac ttattcccacctagtt ctagatcgcgatta attaac(T)10	80	248	327
Contig489	1	p1	(A)13	13	29	41
Contig495	1	p1	(T)10	10	24	33
Contig495	2	p1	(A)11	11	172	182
Contig496	1	p1	(T)11	11	30	40
Contig496	2	p3	(GTG)5	15	170	184
Contig497	1	p1	(A)11	11	1421	1431
Contig502	1	p1	(A)11	11	1135	1145
Contig530	1	p1	(T)10	10	535	544
Contig538	1	p3	(ATC)5	15	1355	1369
Contig544	1	p6	(AAGGCG)6	36	44	79
Contig549	1	p1	(T)12	12	923	934
Contig573	1	p3	(TCT)9	27	101	127
Contig575	1	p1	(T)11	11	56	66

Contig577	1	p1	(T)11	11	570	580
Contig577	2	p1	(A)10	10	683	692
Contig586	1	p1	(A)11	11	416	426
Contig594	1	p3	(CTC)6	18	195	212
Contig594	2	p1	(A)10	10	909	918
Contig601	1	p3	(TCA)5	15	523	537
Contig610	1	p1	(A)10	10	557	566
Contig611	1	p3	(GCA)5	15	131	145
Contig611	2	p1	(T)10	10	578	587
Contig612	1	p3	(CTC)5	15	114	128
Contig613	1	p3	(CGA)6	18	17	34
Contig615	1	p1	(A)11	11	497	507
Contig624	1	p1	(A)10	10	1033	1042
Contig625	1	p3	(GCG)6	18	62	79
Contig627	1	p1	(A)10	10	2025	2034
Contig631	1	p3	(CTC)8	24	54	77
Contig644	1	p1	(T)11	11	23	33
Contig646	1	p2	(AC)10	20	21	40
Contig657	1	p1	(T)10	10	12	21
Contig657	2	p1	(A)10	10	160	169
Contig659	1	p3	(GCG)5	15	25	39
Contig675	1	p1	(A)11	11	1126	1136
Contig677	1	p1	(A)11	11	848	858
Contig691	1	p1	(T)10	10	15	24
Contig691	2	p1	(A)10	10	166	175
Contig691	3	p3	(CGG)8	24	414	437
Contig691	4	p3	(GGC)6	18	1091	1108
Contig694	1	p3	(AGC)7	21	468	488
Contig695	1	p3	(TGG)7	21	216	236
Contig700	1	p3	(CCT)5	15	31	45
Contig700	2	p2	(CT)6	12	445	456
Contig701	1	p1	(A)11	11	618	628
Contig702	1	p1	(A)10	10	697	706
Contig706	1	p1	(T)11	11	24	34
Contig706	2	p3	(TAA)5	15	350	364
Contig711	1	p3	(ATT)8	24	1606	1629
Contig713	1	p1	(A)11	11	573	583
Contig716	1	p2	(GA)6	12	13	24

Contig720	1	p2	(AG)6	12	74	85
Contig723	1	p1	(A)10	10	827	836
Contig724	1	p1	(A)11	11	935	945
Contig735	1	p3	(GAT)7	21	322	342
Contig743	1	p3	(TTG)8	24	775	798
Contig747	1	p1	(A)10	10	935	944
Contig748	1	p1	(A)11	11	910	920
Contig754	1	p3	(ATA)8	24	644	667
Contig757	1	p1	(A)10	10	703	712
Contig758	1	p1	(A)10	10	814	823
Contig770	1	p2	(CT)6	12	611	622
Contig773	1	p2	(AG)6	12	35	46
Contig775	1	p1	(A)11	11	1460	1470
Contig780	1	p1	(A)11	11	116	126
Contig786	1	p3	(GCT)6	18	324	341
Contig791	1	p1	(T)10	10	29	38
Contig793	1	p2	(TA)6	12	4	15
Contig802	1	p1	(T)10	10	27	36
Contig811	1	p1	(A)11	11	759	769
Contig812	1	p1	(A)15	15	846	860
Contig815	1	p3	(TGT)6	18	251	268
Contig827	1	p3	(AAG)5	15	107	121
Contig828	1	p1	(T)11	11	1072	1082
Contig831	1	p1	(T)13	13	987	999
Contig834	1	p2	(CT)6	12	35	46
Contig834	2	p1	(A)10	10	1366	1375
Contig836	1	p1	(A)15	15	35	49
Contig836	2	p1	(T)12	12	802	813
Contig847	1	p3	(GAG)5	15	33	47
Contig853	1	p3	(TGA)6	18	1311	1328
Contig857	1	p3	(CTC)6	18	379	396
Contig858	1	p1	(T)12	12	175	186
Contig867	1	p1	(T)10	10	632	641
Contig870	1	p2	(AT)17	34	333	366
Contig885	1	p1	(A)10	10	93	102
Contig893	1	c	(CAG)6caacag catctacaacaaca ggttgagaaca( AG)6	62	502	563

Contig903	1	p3	(GTG)7	21	244	264
Contig906	1	p1	(T)11	11	30	40
Contig925	1	p1	(A)11	11	36	46
Contig930	1	p3	(AAT)7	21	672	692
Contig933	1	p3	(TTG)6	18	733	750
Contig941	1	p2	(AT)7	14	823	836
Contig945	1	p2	(GA)10	20	39	58
Contig948	1	p6	(AGGCAC)5	30	367	396
Contig967	1	p1	(A)10	10	368	377
Contig970	1	p3	(TCC)5	15	52	66
Contig972	1	p1	(T)10	10	1111	1120
Contig983	1	p1	(T)10	10	14	23
Contig988	1	p3	(GAT)5	15	525	539
Contig995	1	p2	(GA)13	26	209	234
Contig995	2	p3	(ATG)6	18	729	746
Contig998	1	p1	(A)11	11	683	693
Contig1000	1	p3	(CCG)5	15	63	77
Contig1008	1	p1	(T)13	13	51	63

## APPENDIX VIII

## List of Nonsynonymous SNP coding data identified by QualitySNP

>Contig1	239	senseTA
>Contig1	460	senseCT
>Contig16	118	senseCT
>Contig27	1615	sense-T
>Contig28	1011	senseAG
>Contig28	1053	senseGT
>Contig28	1092	senseTC
>Contig35	685	senseGA
>Contig35	705	senseAC
>Contig105	287	senseCA
>Contig105	505	sense-A
>Contig132	140	senseCA
>Contig132	195	senseTC
>Contig141	924	senseCT
>Contig145	127	senseTC
>Contig146	134	senseCT
>Contig146	176	senseGA
>Contig149	256	senseAG
>Contig149	317	senseTC
>Contig151	786	senseCT
>Contig151	861	senseGC
>Contig151	882	sense-C
>Contig151	973	senseGA
>Contig151	1012	senseCT
>Contig151	1036	senseGA
>Contig153	1066	senseTC
>Contig153	1112	senseA-
>Contig153	1230	senseG-
>Contig156	44	senseCT
>Contig156	159	senseCT
>Contig156	183	senseCT
>Contig156	286	senseGC
>Contig163	77	senseAG
>Contig163	177	senseAC
>Contig163	218	senseCT
>Contig163	261	senseTC
>Contig163	310	senseTG
>Contig164	94	senseTC
>Contig164	623	senseTC



>Contig164	650	senseTC
>Contig165	114	sense-G
>Contig169	576	senseAC
>Contig169	702	senseTC
>Contig169	771	senseAT
>Contig169	801	senseCT
>Contig173	335	senseTC
>Contig176	671	sense-A
>Contig194	398	senseTC
>Contig195	110	senseCT
>Contig195	211	senseGA
>Contig197	173	senseAT
>Contig197	220	senseAT
>Contig197	325	senseAG
>Contig198	127	senseAG
>Contig198	151	senseAG
>Contig198	212	senseTG
>Contig199	67	senseCT
>Contig199	116	senseCT
>Contig199	342	senseGA
>Contig203	66	senseGA
>Contig203	144	senseAG
>Contig203	315	senseCT
>Contig203	370	senseAT
>Contig206	849	senseTC
>Contig206	909	senseCT
>Contig207	552	senseT-
>Contig207	756	senseCT
>Contig210	417	senseTC
>Contig210	471	senseGA
>Contig251	233	senseAG
>Contig276	219	senseCT
>Contig276	248	senseCT
>Contig283	526	senseCT
>Contig283	1015	senseGT
>Contig289	1431	senseCT
>Contig294	223	senseC-
>Contig299	209	senseAG
>Contig344	148	senseA-
>Contig404	207	senseTG
>Contig404	862	sense-A
>Contig404	983	senseT-
>Contig412	326	sense-T

>Contig453	754	senseTA
>Contig489	260	senseTC
>Contig547	364	senseTC
>Contig594	502	senseTC
>Contig594	718	senseCA
>Contig601	399	senseTC
>Contig601	1382	senseAT
>Contig601	1469	senseCA
>Contig601	1511	senseGA
>Contig601	1794	senseAG
>Contig601	1834	senseCA
>Contig601	1850	senseAT
>Contig601	1966	senseGA
>Contig624	658	senseGA
>Contig624	758	senseAT
>Contig624	841	senseT-
>Contig624	949	senseTA
>Contig633	217	senseGA
>Contig703	200	senseCA
>Contig703	526	sense-C
>Contig746	1409	senseAG
>Contig746	1847	senseGA
>Contig746	1897	senseGC
>Contig746	1927	senseTA
>Contig746	1985	senseAG
>Contig746	2025	senseTC
>Contig782	544	senseCT
>Contig795	436	senseTG
>Contig796	83	senseCT
>Contig796	135	senseGA
>Contig813	591	senseCT
>Contig813	642	senseGA
>Contig813	657	senseCT
>Contig827	83	senseAT
>Contig827	466	senseCT
>Contig827	481	senseCT
>Contig827	645	senseAG
>Contig827	674	senseAT
>Contig829	822	senseTC
>Contig829	1136	sense-T
>Contig843	377	senseAG
>Contig843	404	senseTC
>Contig846	420	senseA-

>Contig851	395	senseGA
>Contig872	525	senseCT
>Contig872	586	senseAG
>Contig876	257	sense-G
>Contig906	128	senseTG
>Contig906	323	senseTC
>Contig906	347	senseCA
>Contig906	519	senseCT
>Contig906	550	senseTC
>Contig937	622	senseCG
>Contig965	675	senseGC
>Contig999	466	senseC-
>Contig999	541	senseTC
>Contig1023	195	senseTC
>Contig1038	577	senseGA
>Contig1038	645	senseAC
>Contig1039	311	senseTC
>Contig1042	422	senseCT
>Contig1042	543	senseCT
>Contig1042	616	senseAC
>Contig1042	632	senseGT
>Contig1044	1065	senseAG
>Contig1044	1128	senseGT
>Contig1074	411	senseTC
>Contig1091	251	senseTG
>Contig1102	427	senseCT
>Contig1111	482	senseA-
>Contig1130	110	senseTC
>Contig1138	630	senseCT
>Contig1138	675	senseCT
>Contig1138	742	senseCA
>Contig1160	612	senseCT
>Contig1160	675	senseGA
>Contig1165	143	senseTA
>Contig1165	169	senseCA
>Contig1204	106	senseAC
>Contig1204	181	senseTC
>Contig1248	436	senseAG
>Contig1248	611	senseAG
>Contig1414	108	senseGT
>Contig1416	97	senseTC
>Contig1416	153	senseAG
>Contig1416	225	senseAT

>Contig1418	857	senseGA
>Contig1418	1166	senseCG
>Contig1418	1199	senseTC
>Contig1419	119	senseTC
>Contig1420	63	senseCT
>Contig1420	100	senseAG
>Contig1420	318	senseCT
>Contig1420	483	senseAG
>Contig1421	500	senseTG
>Contig1422	489	senseTC
>Contig1422	504	senseTC
>Contig1429	514	senseGC
>Contig1429	772	senseCG
>Contig1432	256	senseTC
>Contig1433	585	sense-G
>Contig1434	65	senseGA
>Contig1434	128	senseAT
>Contig1434	287	senseAT
>Contig1434	377	senseGT
>Contig1434	401	senseGT
>Contig1434	440	senseTA
>Contig1435	1203	senseAG
>Contig1435	1483	senseTG
>Contig1435	1610	senseAG
>Contig1435	1685	senseCT
>Contig1435	1745	senseGT
>Contig1435	1850	senseA-
>Contig1437	1397	senseAG
>Contig1437	1420	senseAG
>Contig1439	336	senseAT
>Contig1439	416	senseCT
>Contig1443	159	senseAT
>Contig1443	213	senseTC
>Contig1443	331	senseTC
>Contig1444	336	senseTA
>Contig1450	361	sense-T
>Contig1452	449	senseAC
>Contig1452	599	senseGA
>Contig1452	670	senseTC
>Contig1452	745	senseAC
>Contig1452	775	senseTC
>Contig1454	228	senseGA
>Contig1454	570	senseG-

>Contig1454	641	senseCT
>Contig1455	524	senseTC
>Contig1456	144	senseAG
>Contig1458	245	senseTC
>Contig1458	299	senseCA
>Contig1458	736	senseAG
>Contig1458	821	senseAG
>Contig1458	972	senseTG
>Contig1459	301	senseTC
>Contig1459	739	senseGC
>Contig1459	844	senseCG
>Contig1461	62	senseCT
>Contig1465	172	senseAC
>Contig1465	307	senseCA
>Contig1465	523	senseA-
>Contig1469	767	senseCA
>Contig1474	267	senseGA
>Contig1482	179	senseTC
>Contig1482	251	senseGC
>Contig1482	272	senseTC
>Contig1494	382	senseCT
>Contig1494	494	senseTA
>Contig1495	708	senseTA
>Contig1495	727	senseTC

# **ABSTRACT**

**PREDICTION OF SSR AND SNP MARKERS FOR  
ANTHRACNOSE RESISTANCE IN YAM USING  
BIOINFORMATICS TOOLS AND THEIR VALIDATION**

By

**SAHLA K.**  
(2013-09-102)

**Abstract of the thesis**

**Submitted in partial fulfilment of the  
requirement for the degree of**

**B. Sc. - M. Sc. (INTEGRATED) BIOTECHNOLOGY**

**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**

**2018**

## ABSTRACT

The study entitled “Prediction of SSR and SNP markers for anthracnose resistance in yam using bioinformatics tools and their validation” was conducted at ICAR-Central Tuber Crop Research Institute, Sreekariyam, Thiruvananthapuram during October 2107 to August 2018. The objectives of the study is to computationally identify SNPs and SSRs for anthracnose resistance in Greater Yam and the verification of identified markers using resistant and susceptible varieties. The preliminary data set for the identification of SSR and SNP markers was obtained from the EST section of NCBI. A total of 44134 sequences was obtained. The dataset was reduced to 44114 sequences after several pre-processing and screening steps. The resulting sequences were assembled and aligned using CAP3 and 5940 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. For SSRs the SSR prediction tools such as MISA and SSRIT was compared and 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, 1789 nonsynonymous SNPs and 73 synonymous SNPs were identified. Using MISA, 359 mono SSRs, 268 di SSRs, 342 tri SSRs, 17 tetra SSRs, 7 penta SSRs, and 9 hexa SSRs were identified. Five sequences from identified SNPs and SSRs which having high hit percentage and low E value were selected for validation and primer designing for anthracnose resistant genes. These primers were validated using 3 resistant and 3 susceptible yam varieties. Among the primers after validation in wet lab, three SNPs (DaSNP1, DaSNP2, DaSNP3) and two SSRs (DaSSR1 and DaSSR2) 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 anthracnose resistance in yam.

174551



112