"DEVELOPMENT OF MOLECULAR MARKERS FOR BLIGHT DISEASE RESISTANCE IN TARO USING BIOINFORMATICS TOOLS"

by

ATHUL V. S.

(2013-09-109)

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 **"DEVELOPMENT OF MOLECULAR MARKERS FOR BLIGHT DISEASE RESISTANCE IN TARO USING BIOINFORMATICS TOOLS"** 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 to me of any degree, diploma, associateship, fellowship or other similar title, of any other University or Society.

Vellayani

Date: 07.12.2018

ATHUL V. S.

(2013-09-109)

स्वच्छ भारत
एक कदम स्वच्छता की ओर

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

भाकृ अनुप ICAR

केकृफअनं CTCRI
ISO 9001:2008

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 **"DEVELOPMENT OF MOLECULAR MARKERS FOR BLIGHT DISEASE RESISTANCE IN TARO USING BIOINFORMATICS TOOLS"** is a record of research work done by **Mr. ATHUL V. S. (2013-09-109)** under my guidance and supervision and that this is not previously formed the basis for the award of any degree, diploma, fellowship or associateship to him.

Place: Sreekariyam
Date: 07.12.2018

Dr. J. Sreekumar
(Major Advisor, Advisory Committee)
Principal Scientist, (Agricultural Statistics)
ICAR-CTCRI

डॉ. जे. श्रीकुमार/ Dr. J. SREEKUMAR
प्रधान वैज्ञानिक (कृषि सांख्यिकी)
Principal Scientist (Agricultural Statistics)
एक्सटेंशन और सामाजिक विज्ञान अनुभाग
Section of Extension and Social Sciences
भा कृ अनु प - केंद्रीय कंद फसल अनुसंधान संस्थान
I C A R - Central Tuber Crops Research Institute
श्रीकार्यम / Sreekariyam
तिरुवनंतपुरम / Thiruvananthapuram - 695 017

Phone : +91-471-2598551 to 2598554
Director (Per) : +91-471-2598431
(Res) : +91-471-2597211
Sr. Admn. Officer : +91-471-2598193

RIGHT TO INFORMATION

Fax : +91-471-2590063
E-mail : director.ctcri@icar.gov.in
ctcritvm@gmail.com
Web : http://www.ctcri.org

CERTIFICATE

We, the undersigned members of the advisory committee of Mr. Athul V. S. (2013-09-109), a candidate for the degree of B. Sc. - M. Sc. (Integrated) Biotechnology, agree that the thesis entitled "**DEVELOPMENT OF MOLECULAR MARKERS FOR BLIGHT DISEASE RESISTANCE IN TARO USING BIOINFORMATICS TOOLS**" may be submitted by **Mr. ATHUL V. S.** in partial fulfillment of the requirement for the degree.

Dr. J. Sreekumar
(Chairman, Advisory Committee)
Principal Scientist,
Section of Extension and Social Sciences
(ICAR-CTCRI)
Sreekariyam, Thiruvananthapuram

Dr. Swapna Alex
(Member, Advisory Committee)
Professor and Head
Department of Plant Biotechnology,
College of Agriculture, Vellayani
Thiruvananthapuram

Dr. A. Asha Devi
(Member, Advisory Committee)
Principal scientist (Genetics)
Division of Crop Improvement
(ICAR-CTCRI)
Sreekariyam,Thiruvananthapuram

Dr. K. B. Soni
(Member, Advisory Committee)
Professor
Department of Plant Biotechnology,
College of Agriculture, Vellayani
Thiruvananthapuram

Dr. M. K. Rajesh
(External examiner)
Principal Scientist (Biotechnology)
Central Plantation Crops Research Institute
Indian Council of Agricultural Research
Kasaragod, Kerala

ACKNOWLEDGEMENT

With no creation in the world being a solo effort, it's my privilege to look over the journey past and thank everyone who supported me in making this thesis to a good shape. I blissfully take this opportunity to express my heartfelt gratitude to:

Almighty GOD, for the blessings and presence, providing me the perseverance and mental support without which I would not have completed the work.

Dr. Archana Mukherjee, Director, ICAR-CTCRI, for giving me an opportunity to work at the institute for my M. Sc. project and supporting me.

Dr. Sheela Immanuel, Head, Division of Extension and Social Sciences, for extending the facilities to perform my work at the department.

Dr. J. Sreekumar, my cherished advisor, who gave me the freedom to work on my own way, supporting my participation at various seminars and conferences. His patience, guidance and advice were of great encouragement throughout my project.

Dr. A. Asha Devi, my advisory committee member who assisted me with great enthusiasm in performing the wet lab part of my work at Division of Crop Improvement.

Dr. Mohan who provide me valuable insights on wet lab experiments and assisting me to get in contact with IDT technology.

Dr. Senthil for helping me in using primer designing tools and designing the primers.

Mr. Prakash Krishnan B. S. for sparing his busy schedule at the department for being with me for assisting in PCR and DNA isolation.

Dr. Anil Kumar A., Dean, COA, Vellayani for providing all the necessary help and facilities provided.

Dr. Swapna Alex, Professor and Head, Department of Plant Biotechnology for her talented guidance and valuable suggestions.

Dr. Soni K. B. for her reasoned criticism and inspiring support from the college. Even with her busy schedule she takes her time to spare for our academic matters and being strict.

Ambu chetan for giving valuable lessons during BSL life and guiding me. He always helped kindheartedly with open suggestions.

BSL members - Reshma, Sahla, Aswathy, Achuth, Shilpa, Gayathri, Haritha, Sruthy, Akshay, Rekha chechi and Priya chechi for all the fun, unparallel affection and care and making BSL more lively.

Special thanks to members of Crop improvement lab - Bimal, Sabari, Arya and Anjitha who spared their valuable time for me at the end phase.

Jithu Sanghi for being with me at CTCRI during the project and supporting me.

Archnaleksmi, for being with me and supporting during my hardtimes through thoughtful and enjoyable discussions.

To all my seniors and juniors at College of Agriculture, who have been the real inspiration for me.

All my colleagues of 2013 kidilams batch for being with me during all my ups and downs for the past 5 years.

My parents Valsalakumar C. and Sreekala D. for being supporting all the times.

Athira V. S. for helping me during the project and mentoring me.

I convey my wholehearted thanks to all my well wishers who were far too numerous to have been mentioned here.

Athul V. S.

Dedicated to My Parents

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PLATES

11

LIST OF APPENDICES

LIST OF ABBREVIATIONS

%	Percentage
°C	Degree Celsius
A_{260}	Absorbance at 260nm wavelength
A_{280}	Absorbance at 280nm wavelength
AFLP	Amplified Fragment Length Polymorphisms
AGE	Agarose Gel Electrophoresis
ABVC	Alomae–Bobone virus complex
BC	Before Christ
CTCRI	Central Tuber Crops Research Institute
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
EST	Expressed Sequence Tag
E-value	Expect value
FAOSTAT	Food And Agricultural organziation Database
GC	Guanine-Cytosine
GBS	Genotyping by sequencing
g	gram
h	hour
ha	hectare
HTML	Hypertext Markup Language
ie	that is
kb	Kilobase

kg	kilogram
mg	milligrams
MAS	Marker Assisted Selection
$MgCl_2$	Magnesium chloride
min	Minute
ml	Milli Litre
mm	Milli Metre
mM	Milli Molar
NCBI	National Centre for Biotechnology Information
ng	Nanogram
NIH	National Institutes of Health
nm	Nano Metre
OD	Optical Density
PCR	Polymerase Chain Reaction
RNA	Ribose Nucleic Acid
RNase	RiboNuclease
rpm	Rotation per minute
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
SSR	Simple Sequence Repeats
TANSAO	Taro Network for South East Asia and Oceania
Taq	*Thermus aquaticus*
TBE	Tris-Borate-EDTA
TLB	Taro Leaf Blight
Tm	Melting Temperature

14

UV	UltraViolet
V	Volt
v/v	volume/volume
w/v	weight/volume
µL	Microlitre
µM	MicroMolar

INTRODUCTION

1. INTRODUCTION

Colocasia esculenta (L.) Schott. a member of the Araceae family, is a widely distributed tropical tuber crop in the world with a global production of 10 million tonnes and a yield of 6,066 kg/ha (FAOSTAT, 2016). Being a tuber, it is the staple food crop of many Pacific countries. With uncertainties existing regarding the origin of taro, the crop is believed to be originated in the south-east Asian regions by ethno botanical evidence and introduced to other countries (Lebot *et al.*, 2004). Taro has a chromosome number of 14 and two cytotypes, a diploid one with 28 chromosomes and a triploid one with 42 chromosomes (Chaïr *et al.*, 2016). With more than 200 cultivars, the crop is mainly classified into wetland taro and upland taro. The ability to propagate vegetatively (by corms) and to adapt to a wide variety of substrate and climate make them an attractive crop globally.

Taro posses a high calorific value of 112 calories/100 grams and serves to be a major source for carbohydrates, dietary fibers, Pyridoxine, Riboflavin, Copper, Zinc and a minor source for fats and proteins (USDA, National Nutrient Database, 2018).

Colocasia esculenta suffers great damage due to the taro leaf blight caused by *Phytophthora colocasiae* apart from the attack by taro beetles which significantly lowered the yield globally (Singh *et al.*, 2012). As chemical controls are harmful and less effective with evolving pathogens, a genetic basis should be adopted for controlling plant-pathogen interactions. Molecular basis of pathogen attacks and crop resistance hold a key role in developing resistant varieties.

In the absence of a reference genome sequence, transcriptome sequencing has proved to be an efficient tool for discovery of molecular markers, gene expression profiling and mapping (Mutz *et al.*, 2013). The *in-silico* approaches for the discovery of molecular markers mainly revolve around the information gathered from expressed sequence tags (EST) using Sanger sequencing. Recent trend focuses on next-generation sequencing (NGS) for the molecular marker and gene discovery which bypasses the expensive and time-consuming nature of the EST-based method and generates significant output data with quality, robustness

and low noise with the aid of powerful computers and complex algorithms (Buermans *et al.*, 2014).

Single nucleotide polymorphisms (SNPs) are the type of genetic markers with high abundance and slow mutation rate within the genome. SNP discovery is crucial to determine the genetic variability of an organism and the *in-silico* approach is based upon the sequence information available in public databases, in most cases as EST and NGS and are considered to be faster and cheaper than experimental procedures (Tang *et al.*, 2006).

SSRs (simple sequence repeat) originally designated as STRs (short tandem repeats) are the class of molecular markers with repeats of 2 - 6 nucleotides with genetic co-dominance, abundance, high level of polymorphism, multi-allelic variation, high reproducibility and dispersal throughout the genome, make them ideal for molecular mapping and plant breeding studies (Li *et al.,* 2002 : Eujayl *et al.*, 2004).

Crop research is a gradually expanding field of science with significant achievements being made in the past decade (Bilsborough, 2013). Data sharing, integration, and annotation are crucial for validating the findings made experimentally. Bioinformaticians and computer scientists with little or no help from biologist could perform these. On the contrary, biologists are crucial as they are the major producers and penultimate users of the data. Sharing, integration, and annotation, however, depends on the adoption of standards, submission mechanism, shared formats etc. which enables the convenience for other research purposes. Successful data integration from a computational viewpoint and its application in the field of biological research contributes to new discovery and scope in the future (Laptas *et al.*, 2015).

With the arrival of new sequencing platforms, identification of genome wide distribution of SNPs, SSRs, etc. was possible, which in turn helped in identifying the disease-resistance genes. The genome sequences of organisms are fundamentally important for discerning the genes, their functions, evolutionary relationships and unknown regulatory mechanisms. The approach not only has a

weighty impact on human disease and diagnostics but also aids in crop improvement. Sequential information comes handy for breeding, identifying challenges and to utilize the variation present within a genome (Bevan *et al.*, 2013).

The present study was undertaken with the following objectives to computationally develop SNPs and SSRs for taro leaf blight disease resistance, and to validate them for understanding their effectiveness.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

Taro [*Colocasia esculenta* (L.) Schott], referred to as "potato of the tropics" or "elephant ears" is a member of Araceae family with wide adaptability and large-scale acceptability. It is grown primarily for its edible corms, leaves, and petioles. The taro plant as such is useful with the stem being used as salads, the tubers as a source of digested starch, leaves as a green vegetable and for wrapping food.

The crop is known by different names all over the country such as arvi (Hindi), chempu (Malayalam), seppan kizhangu (Tamil), kachchi (Kannada), chamadumpa (Telugu), alu (Marathi) and kachu (Bengali) (Edison *et al.,* 2003).

Apart from being a backyard crop, its commercial cultivation accounts for about 16,69,708 ha globally (FAOSTAT, 2016). Taro grows with an average annual precipitation of 2500 mm or more (Weightman, 1989). Survival in waterlogged conditions utilizing the hydromorphic soil makes it more acceptable where other tuber crops fail (Onwueme, 1978).

Aroids, often known as "orphan crops" are not extensively traded and studied by researchers and constitute to be a minor crop globally. Even being a minor crop it is quite essential for the food security with their unique nutrient profile. *Colocasia* and *Xanthosoma* represent the major class of aroids with the former known as taro/dasheen and the latter as cocoyam or tannia.

The narrow genetic bases available are the major limitations faced by taro breeding programmes (Banjaw, 2017), however, exchange of genotypes could broaden up the bases of breeding (Lebot and Aradhya, 1991). Lebot *et al.* (2004) suggested a breeding strategy using wide genetic bases composing of parents from diverse regions. The diversification allows for gene pools among different cross cultivars as crosses from one country are not desirable.

Phytophthora colocasiae Rac., a foliar pathogen causes TLB which accounts for a decrease in taro production. TLB occurrence is highly related to the climatic condition of a region (Edison *et al.,* 2003). The deadly disease affects taro globally with serious outbreaks being reported in Samoa in 1993 and in the Cameroon, Ghana, and Nigeria during the past few years (Singh *et al.,* 2012). Leaf blight caused by *Phytophthora colocasiae* Raciborski limited the production of the crop in Nagaland with expression being reported around monsoon and continues throughout the rainy season (Pongener *et al.,* 2016).

Many breeding programmes target either resistance against disease or increasing yield, achieved by means of molecular markers (Scholten *et al.,* 2005). Molecular marker improves the efficiency of plant breeding by carrying out the selection of traits linked on to it (Mohan *et al.,* 1997). Being unaffected by environmental conditions in which plants are grown and detectable in all plant growth stages makes marker-assisted selection (MAS) more practical.

It has been predicted that a combination of changing dietary habits and prospering human population growth will result in an increased demand for agricultural production of 60-110% by 2050 (Alexandratos and Bruinsma, 2012). Increasing production demands for the practice of cultivation of high yielding and disease resistant plants (Godfray *et al.*, 2010). Improvement demands the better understanding of the genetic mechanisms controlling traits of interest, and genomics approaches (Bilsborough *et al.*, 2013).

In this chapter, literature concerning the leaf blight disease, *in-silico* development of molecular markers (SSR and SNP), and their validation have been presented.

2.1 CENTRE OF ORIGIN

Taro (*Colocasia esculenta*), a vegetable and starchy tuber cultivated all over the world, is believed to have originated in South Central Asia, probably in India or the Malay Peninsula with Nigeria, Cameroon, and Ghana account for more than 50% of global production (FAOSTAT, 2017). The absence of written records, linguistic records, archeological evidence and descriptional confusion with *Xanthosoma* species make it difficult to support the exact view of origin (Leon, 1977).

Even before human used planting, harvesting cycles, and conventional agricultural techniques, the collection of starch from the sago palm (*Metroxylon sagu*) and taro (*Colocasia esculenta*) was in practice around marshy areas, lakes, swamp forests, and rivers (Goltenboth *et al.*, 2006). With significant citations in the Classical (Greek and Latin) texts that record the name *Colocasia* from the 3rd century BC onwards, there is a possibility for the crop to be originated in the Mediterranean region also (Grimaldi *et al.*, 2018).

The diversity and number of private alleles were observed more in Asian accessions, mainly from India. Bayesian clustering revealed the origin of diploids around Asia-Pacific region and a second diploid-triploid group to India (Chair *et al.*, 2016).

Being a crop significant for production and trade due to their medicinal and edible qualities may also contribute to their worldwide dispersal all over the world through maritime and terrestrial trading routes.

2.2 TARO NUTRITION PROFILE

Njintang *et al.* (2008) found out that taro starch has high solubility index and water holding capacity than other starch synthesizing counterparts. With low fat and protein, 70–80 % starch, minerals, vitamins and rich in anthocyanins such as cyanidin-3-chemnoside, pelargonidin-3-glucoside, and cyanidin-3-glucoside which were revealed to possess anti-inflammatory and antioxidative property

makes taro more preferable (Kaushal *et al.*, 2015). The presence of resistant starch and mucilage in taro peculiarized with slower digestion leads to the slower release of glucose and aids in treating diabetes, obesity and several diseases (Liu *et al.*, 2006).

Several studies reveal the presence of several macro and micro minerals in taro with potassium being the abundant one along with magnesium, calcium, phosphorous etc. (Mwenye *et al.*, 2011). Huang *et al.* (2007) investigated the role of cultivars and field preparations and observed taro to be rich in thiamin, riboflavin, and ascorbic acid. Lewu *et al.* (2010) carried out the comparative assessment of taro and observed fewer concentrations of zinc, manganese, and iron. The composition of minerals, however, was influenced by the interaction of the genotype and climatic conditions (Mwenye *et al.*, 2011). The nutrient profile comprising high vitamin E, fiber, potassium, and other macro and micronutrients makes taro unique over other tuber counterparts (USDA, 2018).

2.3 PLANT MORPHOLOGY

In the book "Species Plantarum" by Carl Linnaeus, taro was classified into two types - *Arum colocasia* and *Arum esculentum*. However, in 1832, Schott established the genus *Colocasia* and renamed them as *Colocasia esculenta* and *Colocasia antiquorum* respectively. Purseglove in 1972 morphologically identified two varieties of taro: eddoe and dasheen. Eddoe characterized with a central corm surrounded by many small cormels, and dasheen, with one main large corm (Plucknett 1983). O'Sullivan *et al.* (1996) described eight polymorphic variants in *Colocasia esculenta* of which *Colocasia (L.)* Schott *var. esculenta* and *Colocasia* (L.) Schott var. *antiquorum* being the widely cultivated ones.

A monocotyledonous herbaceous plant with 1- 2 cm height, apically growing large heart-shaped leaves from the top of corms composed of a multi-layered palisade and air-filled spongy mesophyll, abaxial and adaxial stomata, highly vacuolated epidermal cells, variable morphology, peltate structure and

laterally growing underground corms (Stein *et al.*, 1983). The name taro now accounts for about 3 aroid species *Alocasia macrorrhiza* (L.) G. Don (giant taro), *Colocasia esculenta* (true taro), and *Cyrtosperma merkusii* (Hassk.) Schott (swamp taro). Among them, true taro is further classified into two as *C. esculenta* var. *esculenta* and *C. esculenta* var. *antiquorum* (Ivancic and Lebot, 2000).

Onwueme in 1978 reported chromosome numbers as, 2n = 22, 26, 28, 38, and 42 for taros from various regions. Chromosomal variation occurs in the plant depending on their origin with 2n = 24 and 4n = 48 for clones from India, 2n = 28 for clones from Polynesia, while 2n = 28 is found directionally distributed from India to Japan and to New Caledonia, and 3n = 42 in New Zealand (Yen *et al.*, 1968). However, two chromosome numbers are commonly reported for taro, 2n = 28 and 3n = 42 (Kuruvilla *et al.*, 1981). In India both triploid and diploids are reported, diploids dominate in the southern region while triploids dominate in the north (Sreekumari and Mathew, 1991).

2.4 TARO LEAF BLIGHT (TLB)

Attacks on plants represent a global threat to food security. Due to the local consumption and lack of entry to the international trade and market, taro blight has gone unnoticed over the past (Gregory, 1983). One of the important destructive disease of taro accounting for 20-50 yield loss, caused by *Phytophthora colocasiae* Rac. The pathogen also caused serious post-harvest loss to the species (Misra *et al.*, 2008). Trujillo (1965) observed the higher frequency of TLB in areas with high humidity and rainfall whereas lower in areas with a warmer climate.

Wagih *et al.* (1994) reported declining production of taro in Papua New Guinea by the attack of *Phytophthora colocasiae.* Along with taro leaf blight (TLB), declining soil fertility, attacks by taro beetles, and the Alomae – Bobone virus complex (ABVC) together add to the declining production globally (Singh *et al.*, 2008). Sharma *et al.* (2009) identified the genes which conferred blast disease resistance. Sharma *et al.* (2008) used virulent *P. colocasiae* to inoculate

compatible and incompatible varieties to characterize the host-pathogen interactions using Suppressive Subtractive Hybridization (SSH), Northern blot analysis and high throughput DNA sequencing.

Phytophthora colocasiae with a limited host range, primarily infecting the *Colocasia* species is believed to reduce the corm yield by 50%, leaf yield by 95% and also possess significant threats during the storage periods (Singh *et al.*, 2012). Genetic analysis of plant pathogen is crucial to determine the evolution and resistance for an efficient leaf blight management (Milgroom *et al.*, 1997; Lebot *et al.*, 2003).

2.5 MOLECULAR MARKERS

Development of molecular marker technology in the 1980s had revolutionized plant breeding and achieved significant improvements. Morphological, cytological and biochemical markers constitute the major classes of markers and DNA markers such as AFLP, RAPD, SNP, SSR, and ISSR are the widely used ones. Depending on the types of repeats and purity, the efficiency of marker development varies (Vieira *et al.*, 2016). Molecular markers serve as the ideal candidates for detection and screening of mutations, insertion-deletions, and duplications (Hayward *et al.*, 2015).

2.6 SNP

Single nucleotide polymorphism (SNP) refers to an alteration in a single nucleotide -A-T-C or G- between members of a species (Ching *et al.*, 2002). SNPs can be categorized into 3

- Transition (C/T or G/A)
- Transversion (C/G, A/T, C/A, or T/G)
- InDels (small insertions/deletions)

Doveri *et al.* (2008) found SNPs to be bi-, tri- or tetra-allelic, with bi-allelic being common and tetra being rarest. The detection of SNPs has a great role in determining the relation between allelic forms of a gene and their phenotypes (Jorde, 2000). Recent developments in sequencing technology eased the

discovery of SNP and insertion-deletions. With high frequencies of one per ~100–500 base pairs (bp) SNPs are widely used choice to exploit the linkage disequilibrium and obtain high-resolution genetic mapping (Rafalski, 2002).

With high abundance and amenability for high throughput detection, computational-based approaches dominate the SNP discovery methods (Batley *et al.*, 2003). Increasing sequential information in the database and complexity of genomes posses a great challenge in the identification of SNPs. SNP assays with accurate phenotyping have accelerated marker-assisted selection to create salt-tolerant soybean cultivars (Patil *et al.*, 2016). SNPs are crucial for pathogen analysis, phylogenetic analysis and correlation of genotype with phenotype.

2.7 SSR

Microsatellites often referred to as SSR (simple sequence repeats) or STR (short tandem repeats) are short 2- 6 bp DNA motifs repeated within the genome of an organism. SSR markers are being widely exploited to study the functional genomics of an organism. Its occurrence results from either addition or deletion of repeating motifs. With the difference in the number and type of repeats, variation occurs in the genome.

Being found in both prokaryotes and eukaryotes with wide distribution found in coding and non-coding DNA, SSRs are widely used for genotyping plants over last decades (Taheri *et al.*, 2018). Temnykh *et al.* (2001) found out SSRs with longer repeats to be highly polymorphic and shorter repeats to be less polymorphic while studying the rice genome. Qu *et al.* (2013) observed the distribution of SSR across the maize genome to be non-random, with UTR region accounting for the most. Various researches and findings by researchers propose that longer and purer repeats posses higher mutation frequency whereas shorter repeats have lower frequencies.

2.8 SNP AND SSR MARKERS IN PLANTS

SSR markers with high polymorphism and SNPs with high abundance are essential in plant breeding programmes (Gonzaga *et al.*, 2015). With significant

achievements being made in the field of molecular genetics, the co-dominant markers such as SNP and SSR are being exploited more and more to achieve progress. By surviving innovation and possessing technical advances, these markers remain as the prime target of the research community (Vieira *et al.*, 2016).

2.9 MOLECULAR ASPECTS OF TARO

The major constraint in the field of research in taro is the narrow genetic base and the lack of exotic collections. Genetic improvement for taro could be achieved with the acquisition of pathogen-free varieties from Pacific and other regions (Edison *et al.*, 2003).

22 ESTs, 144 genes, 88 UniGenes, 2,088 protein sequences, 2,138 DNA and RNA sequences, six experimentally-determined biomolecular structures, 117 sequence sets from phylogenetic and population studies and one functional genomics study have been so far reported for taro in NCBI, which clearly highlights the lack of research in the crop.

In the absence of a well-sequenced genome and EST information, the molecular marker development provides sufficient information for obtaining a genetic linkage map, to study the genetic basis of phenotypic traits of interest and other genotypic information (Helmkampf *et al.*, 2017).

Segregation of traits could be better understood by employing techniques to develop molecular marker and linkage maps. Isozyme studies conducted by Lebot and Aradhya in 1991 showed greater variation in accessions from Indonesia, Hawaii, and Melanesia. However, of the 1,417 accessions, 343 accessions from the Hawaiian region doesn't constitute any variation. Matthews *et al.* (1992) analyzed ribosomal DNA to separate a few taro accessions from Japan. Irwin *et al.* (1998) used random amplified polymorphic DNA (RAPD) primers for evaluating genetic diversity in *Colocasia* from Hawaiian and

Indonesian accessions. The study also reported triploid and diploid accessions to be useful in parental selection for crop improvement.

Quero-García *et al.* (2006) recommended for the inclusion of a large number of SSR markers, progenies and important traits for an effective mapping analysis in taro. Eleven microsatellite markers were isolated from a population of 30 for germplasm management and population evolution in China (Hu *et al.*, 2009). A simple sequence repeat-sequence characterized amplified region (SSR-SCAR) was developed by Dai *et al.* (2016) for facilitating the conservation and utilization of *Colocasia esculenta* cv. Xinmaoyu which clearly distinguished between cultivars of Jiangsu Province and Fujian Province. Wang *et al.* (2017) sequenced the transcriptome of Jingjiang Xiangsha variety to develop 127 pathways in the Kyoto Encyclopedia of genes and genomes (KEGG). With high polymorphism value which ranged from 0.042 to 0.778, the 65,878 unigenes could be used up for gene analysis and other discoveries.

Kreike *et al.* (2004) used a combination of three AFLP primers to group 255 accessions from Vietnam, Thailand, Malaysia, Indonesia, Philippines, Papua New Guinea, and Vanuatu based on gene distance and genetic diversity measured. Similarly, Noyer *et al.* (2003) made use of AFLP primers to study genetic diversity within the accessions of TANSAO.

DarT (diversity arrays technology) markers were used to analyze the somaclonal variation in taro along with greater yam (*Dioscorea alata*) in the islands of Vanuatu (Vandenbroucke *et al.*, 2016). A low, 3 % polymorphic clones were detected against 13% in yam on the DArT arrays and somaclonal variants were selected as the new varieties.

Mace *et al.* (2002) used microsatellites as a tool for genome mapping and marker-assisted selection for the genotypes from Southeast Asia and Oceania region. Lu *et al.* (2011) opted SSR markers for distinguishing and studying the evolutionary history of taro species in southwestern China.

Inter-Simple Sequence Repeat (ISSR) markers were used for distinguishing *Xanthosoma sagittifolium* (L.) Schott (Taioba) and *Colocasia esculenta* (L.) Schott (Taro) (Sepúlveda-Nieto *et al.*, 2017).

Matsuda *et al.* (2002) discovered Restriction fragment length polymorphism (RFLP) while investigating ribosomal DNA (rDNA) polymorphism in 227 accessions of taro from China, Japan, Taiwan, and Vietnam. Sharma *et al.* (2008) used AFLP markers for analyzing geographical differentiation and for identifying markers linked to taro leaf blight disease.

Tahara *et al.* (1999) studied the SNPs in 13 accessions of taro for distinguishing *Colocasia* and *Alocasia*. Of the two loci, only *trnL* - *trnF* loci showed variations which were not sufficient to classify them. Soulard *et al.* (2017) constructed two genetic linkage maps of taro using SNPs identified using GBS to develop a reliable SNP set in taro.

2.10 NEXT-GENERATION SEQUENCING (NGS).

Sanger and Coulson's sequencing proved to be effective in *Arabidopsis thaliana*, however, the complexity of genomes, time factor and cost made the research community to pull out of it to move towards NGS platforms (Arabidopsis Genome Initiative, 2000). Advances in NGS have made a new plot for detection of markers, especially SSR and SNP.

Different platforms are present in NGS analysis such as 454 Roche (http://www.my454.com) for bacterial and viral genomes, Illumina genome analyzer (http://www.Illumina.com) for plants, humans, and mouse, ABI SOLID (http://www.thermofisher.com), Ion Torrent (http://www.thermofisher.com), and Qiagen GeneReader (http://www.genereaderngs.com) for other microbes and prokaryotes.

Being huge in size NGS data provide solutions to overcome issues related to origin, external contamination, and degradation of samples. The advances being

made in the field further promotes and boosts research interest among scientific community (Di Donato *et al.*, 2018).

More and more sequencing of plant genomes is being done with the onset of the NGS. Genome assembly generation in plants having polyploid genomes with high levels of repetitive sequences is confronting (Bevan *et al.*, 2013).

Gimode *et al.* (2016) used Next Generation Sequencing (NGS) for developing SSR and SNP markers. 10,327 SSRs and 23,285 non-homologous SNPs were found out and validated which significantly contributed to the finger millet genetic information. Wang *et al.* (2013) used NGS for the discovery of SSR markers and assembling of unigenes in *Chrysanthemum nankingense*, which yielded 70,895 unigenes and 1,788 primer pairs.

With the combination of genomics and NGS technology, SNP and SSR markers have accelerated the pace of plant breeding programmes (Mammadov *et al.*, 2012). NGS technology provides powerful methods to breeders for high accurate analysis of genomes. With the higher accuracy and reproducibility they are being widely accepted for marker development and genotyping (Torkamaneh *et al.*, 2018). Illumina, 454 pyrosequencing are being widely used for developing SSR and SNP among plant species (Taheri *et al.*, 2018).

NGS technology as a whole got applications among pathogen detection and data management also. It bridges the gap among genome data and breeding programmes via marker development and utilization of the raw data (Choe *et al.*, 2018). Genome assembly of many crops has been accomplished by combined approaches of bioinformatics and next-generation sequencing which opened up new frontiers for developing and improving new varieties.

2.11 BIOINFORMATICS TOOLS FOR MOLECULAR MARKER DEVELOPMENT

Being faster and cheaper, bioinformatic approaches are effective for molecular marker development. With various tools written in different scripts assigned to different functions, a combined approach among breeders and

researchers will foster improved crop production. A few tools are being described below.

2.11.1 Trimmomatic

Developed by Bolger *et al.* (2014), it is a faster-multithreaded command line tool which trims and crops the paired or single end data according to the parameters users provide and also assists in removing adaptors. Trimmomatic performs trimming and clipping in 2 different steps, in the first step the java programme finds for matches between adapters and reads based on input parameters and gives an alignment score based on which the second sliding window step trims with a threshold score.

Trimmomatic over the past few years has cited several applications, analyzing lncRNAs in CD4+ T cell differentiation (Ranzani *et al.*, 2017), drafting genome sequence of *Pythium periplocum* (Kushwaha *et al.*, 2017), characterization of species among juniper forests (Wahid *et al.*, 2016), enhancing structural annotation of yeast genome (Devillers *et al.*, 2016), for identifying differential expression in CHO cells (Monger *et al.*, 2017), for assembly of cucumber somaclones (Skarzynska *et al.*, 2017), for identifying gene regulation in maize during root emergence and initial growth (Hwang *et al.*, 2018) etc.

2.11.2 Trinity

Trinity serves as the platform for *de novo* reconstruction of transcriptomes from RNA-Seq data without a reference genome. Inchworm, Chrysalis, and Butterfly serve as the three different software modules for Trinity. The 3 step process begins with assembling the datasets into transcript sequences by inchworm, construction of de Bruijn graphs and partitioning of the reads to produce transcripts by Chrysalis and synthesis of transcripts by Butterfly. The runtime of the protocol depends on the size and complexity of data (Grabherr *et al.*, 2011).

Several researches had used Trinity as the *de novo* assembly and transcriptome analysis tool such as in expression analysis of *Diuraphis noxia* for

selecting reference genome (Sinha *et al.*, 2014), genome annotation of *Colletotrichum acutatum* (Han *et al.,* 2016), *De novo* assembly and transcriptome analysis of *Rubus idaeus* (Ward *et al.*, 2012), *Oryza officinalis* (Bao *et al.,* 2015), Chili Pepper (Liu *et al.*, 2013), *Camelina sativa* (Liang *et al.*, 2013) *Monotropa hypopitys* (Beletsky *et al.*, 2017) and *Petunia hybrida* (Villarino *et al.*, 2014).

2.11.3 CAP3

CAP3 refers to the sequence assembly program for clipping 5′ and 3′ low-quality regions of reads is the third successor to CAP (Contig Assembly Program) developed by Huang in 1992. It generates consensus sequences based on multiple sequence alignment of the reads based on quality values (Huang *et al.*, 1999). CAP3 on comparison with PHRAP produces smaller contigs with few or nill error. He *et al.*, 2015 observed CDTA (Combined *De novo* Transcriptome Assembly) strategy and SAMP (Single-Assembler Multiple-Parameter) strategy to be better for transcriptome assembly.

CAP3 is widely used in molecular marker development studies such as EST-derived SSRs in *Epimedium sagittatum* (Zeng *et al.*, 2010), common bean (Hanai *et al.*, 2007), *Vicia faba* (Ma *et al.*, 2011), *Vaccinium corymbosum* (Boches *et al.*, 2005), study of molecular chaperones in sugarcane (Borges *et al.*, 2007) *and* annotation of cDNAs in *Thellungiella halophila (*Taji *et al.*, 2008).

2.11.4 SNP Identification Tools

With the experimental methods highly expensive and unavailable to all, computational approach holds the potential for the discovery of SNPs (Schlotterer, 2004). Different tool are being used for identification of SNP such as SNAP (Johnson *et al.*, 2008), kSNP3.0 (Gardner *et al.*, 2015), PolyPhred (Nickerson *et al.*, 1997), POLYBAYES (Marth *et al.*, 1999), *Consed* (Gordon *et al.*, 1998; 2013) Phred (Ewing *et al.*, 1998), SNPServer (Savage *et al.*, 2005) AutoSNP (Barker *et al.*, 2003) and QualitySNP (Tang *et al.*, 2006) being a few among them. Unfortunately, many of them are outdated due to lack of funding and are not publicly available to the research community.

2.11.4.1 AutoSNP

A freely available perl script programme for detection of SNPs from sequence data using redundancy-based approach. d2cluster and cap3 are being used by AutoSNP for aligning the sequences and differentiating the candidate SNPs (Barker *et al.*, 2003). Batley *et al.*, 2003 used AutoSNP for identifying SNPs in maize and found out them to be of true genetic variation.

2.11.4.2 Quality SNP

An algorithm developed for the detection of reliable SNPs in the presence or absence of quality files. It runs on UNIX/ LINUX and Windows platform using 3 filters for SNP detection from polyploid and diploid species. The filters screens for potential SNPs, reliable SNPs and calls non-synonymous SNPs (Tang *et al.*, 2006). It also hosts for an SNP database with SNPs developed from apple, potato and other species using ESTs. It outperforms almost all SNP prediction pipelines by identifying haplotypes and examining the gene cluster.

2.11.5 SSR Identification Tools

Conventional methods for SSR detection seems to be expensive and time-consuming (Powell *et al.*, 1996) whereas the advent of sequencing technologies, increased potential and less expensiveness makes computational approaches good to go. Microsatellite identification tools like WebSat (Martins *et al.*, 2009), GMATo (Wang *et al.*, 2013), SSR Locator(Da Maia *et al.*, 2008), FullSSR (Metz *et al.*, 2016), SciRoKo (Kofler *et al.*, 2007) and SSRIT (Temnykh *et al.*, 2001) are being employed. Unfortunately, many of them are outdated due to lack of funding, the complexity of organisms and increased sequential information.

2.11.5.1 MISA

A platform-independent perl script programme for the identification of SSRs. It serves to be an offline tool capable of handling large sequences (Thiel *et al.*, 2003). With additional supplementary scripts MISA can also design primers

and perform statistical analysis. However, acceptance of input data only in fasta format and inappropriate clustering are some of the disadvantages. MISA has been employed up for detecting SSRs in eukaryotic organisms (Sharma *et al.*, 2007), eucalyptus (Ceresini *et al.*, 2005) and coffee (Aggarwal *et al.*, 2007).

2.11.5.2 SSRIT

A platform independent program for finding SSRs (2-6 bp) available in both online and stand-alone version. SSRIT accepts only perfect repeats and statistical analysis needs to be done separately (Temnykh *et al.*, 2001). SSRIT has been successfully employed for identification of SSRs in *Gossypium raimondii* (Wang *et al.*, 2006), wheat (Li *et al.*, 2008), barley, maize, rice, sorghum and wheat (Kantety *et al.*, 2002) and *Jatropha curcas* (Yuanzhen *et al.*, 2010).

2.11.5.3 GMATo

Genome-wide Microsatellite Analyzing Tool (GMATo), an SSR mining programme for data of any length (Wang *et al.*, 2013). Being accessible on Windows, Linux, and Mac and written on both perl and java scripts, GMATo serves to be better in characterizing huge genome. Wang *et al.* (2013) found out GMATo to be more effective in processing large datasets within a short time. Zhang *et al.* (2017) used GMATo for characterization of the chloroplast genome of *Primula chrysochlora*.

2.11.6 Primer3plus

A web-based interface to the primer design program primer3, in Perl script instead of CGI scripts with an open architecture. With Polymerase chain reaction (PCR) becoming more vital in modern science, the need for reliable primer design is also of utmost importance (Untergasser *et al.*, 2007; 2012). A successful molecular biological experiments crucial part lies in designing of oligonucleotide primers (Hung *et al.*, 2016). With general settings and advanced settings, Primer3Plus let users define parameters such as Product Size Ranges, Primer Size,

Primer Tm, Max Tm Difference, Primer GC%, Concentration of monovalent cations and dNTPs with minimum, optimum and maximum values.

2.11.7 ClustalW

Clustal programs, in general, are used for aligning nucleotide or protein sequences. ClustalX corresponds to a simple text system whereas ClustalW provides a graphical interface system (Thompson *et al.*, 2003). ClustalW is a tool for carrying out multiple sequence alignment via a three-step process - pairwise alignment, tree generation and progressive alignment (Li, 2003).

2.12 VALIDATION TECHNIQUES OF *IN SILICO* DATA

2.12.1 Gel electrophoresis

Obtained from seaweeds, agar can be classified into agaropectin, with high sulphate and carboxyl groups and agarose, with a neutral fraction of components (Jeppson *et al.*, 1979). Separation (0.5 to 25 kb DNA fragments) and visualization of DNA can be done by agarose gel electrophoresis with varying gel concentrations (0.3-3%). With submarine gel system being universally used, it is run either horizontally or vertically (Smith, 1996). It is a 3 stage process starting with gel preparation followed by loading of samples and staining of the gel (Voytas, 2000).

2.12.2 PCR

A technique developed for *in vitro* amplification of DNA or RNA using repeated cycles of denaturation, annealing, and polymerase extension (Mullis *et al.*, 1986). PCR makes use of polymerase enzymes that use a defined segment in DNA or RNA as a template and synthesize a complementary strand (Schochetman *et al.*, 1988). Thermostable DNA polymerase isolated from *Thermus aquaticus* is being used for the amplification, at higher temperatures for greater specificity, yield, and products (Saiki *et al.*, 1985). New types of PCR are being developed such as Droplet Digital Polymerase Chain Reaction (PCR) which surpasses the real-time PCR (Doi *et al.*, 2015). PCR has got applications in a

wide area ranging from smartphone-assisted molecular diagnostics (Jiang *et al.*, 2014) to microfluidic devices (Ahrberg *et al.*, 2016).

Materials and Methods

3. MATERIALS AND METHODS

The study entitled "Development of molecular markers for blight disease resistance in taro using bioinformatics tools" was conducted at the Central Tuber Crop Research Institute (CTCRI) during 2017-2018. In this chapter, details regarding the experimental materials used and methodology adopted are disclosed.

3.1 TARO SEQUENCE DATA SET

The preliminary data for marker development was obtained from SRA section of NCBI (https://www.ncbi.nlm.nih.gov/sra). Sequence Read Archive (SRA) comprises of biological sequence data information collected from sequencing platforms such as Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Being the primary archive for high throughput sequencing data of NIH (National Institutes of Health), it makes the data available to the research community for new discoveries and addresses the challenges faced by massive sequencing technologies. Being the central repository of NGS data, it also provides a link to other related data sets and facilitates easy data retrieval.

SRA data with the accession number SRX290678 submitted by the College of Life Sciences, Wuhan University was used (Wang *et al.*, 2017). The data was obtained from the leaf sample of a general taro variety named - "HBTARO No. 1". The sequences were obtained in paired fastq format using high-throughput Illumina HiSeq 2000 sequencing technology.

Workflow for identifying SSR and SNP from the above data set is given in Figure 1.

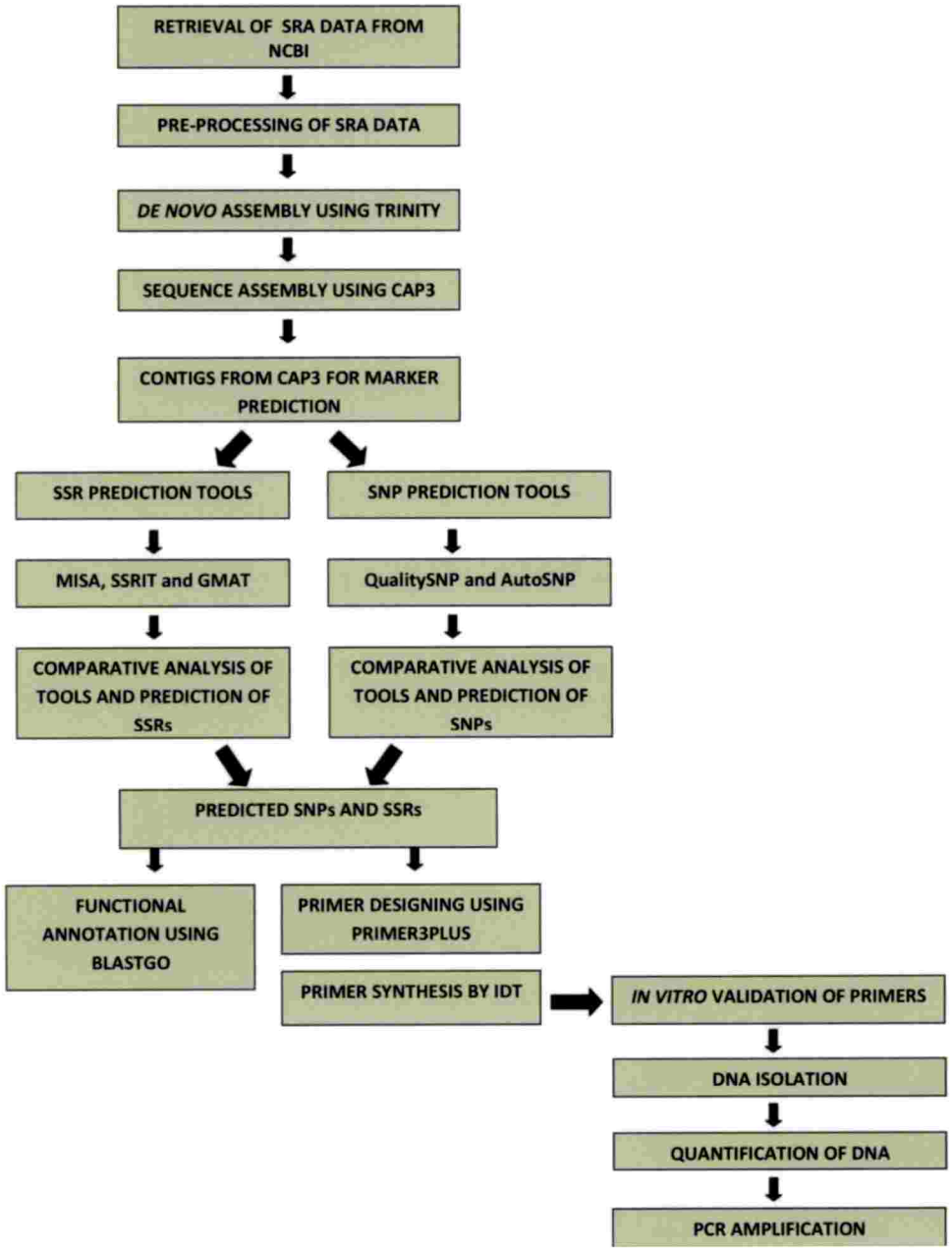

Figure 1. Workflow for the identification of SNP and SSR markers for blight disease resistance in taro.

3.2 PREPROCESSING OF SEQUENCES

Trimmomatic was used for preprocessing the taro sequences to remove sequences of lower quality. The program works by trimming the input paired sequence based on parameters provided in the command.

Since two individual reads are needed for preprocessing in trimmomatic, the given SRA file was split to the left and right reads using the command -

fastq-dump --split-files SRR873449.sra

where SSR873449 was the run ID of the accession number SRX290678.

The important parameters which were given to trimmomatic were,

ILLUMINACLIP - for cutting adapters and illumina specific sequences from the input sequence given.

SLIDINGWINDOW- for trimming within the window for below average sequences.

LEADING - for cutting bases from the start of sequence which fails to meet the threshold quality.

TRAILING - for cutting bases from the end of sequence which fails to meet the threshold quality.

CROP - for trimming the read to a desired length

HEADCROP - for removing certain bases from the start of a read

MINLEN - for eliminating a read, if it fails to meet the desired length.

TOPHRED33 - for converting the quality scores to Phred-33

TOPHRED64 - for converting the quality scores to Phred-64.

Default value set is Phred-64, *ie* if no conversion parameters are given, sequences quality file would be converted to Phred-64.

For a paired data the workflow of trimmomatic is as given in Figure 2. With default parameters (Bolger *et al.*, 2014) trimmomatic was run in terminal using the command -

```
java -jar trimmomatic-0.30.jar PE -phred64 R1.fastq R2.fastq R1_paired.fq.gz
R1_unpaired.fq.gz R2_paired.fq.gz R2_unpaired.fq.gz
ILLUMINACLIP:contams_forward_rev.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

After the terminal operation gets over a log file was generated indicating the name of the read, the length of the sequence after trimming, the location of first and last base present after leading and trailing cut, which indicates amount of reads trimmed from the start and end. Depending upon the reads multiple commands could be added up.

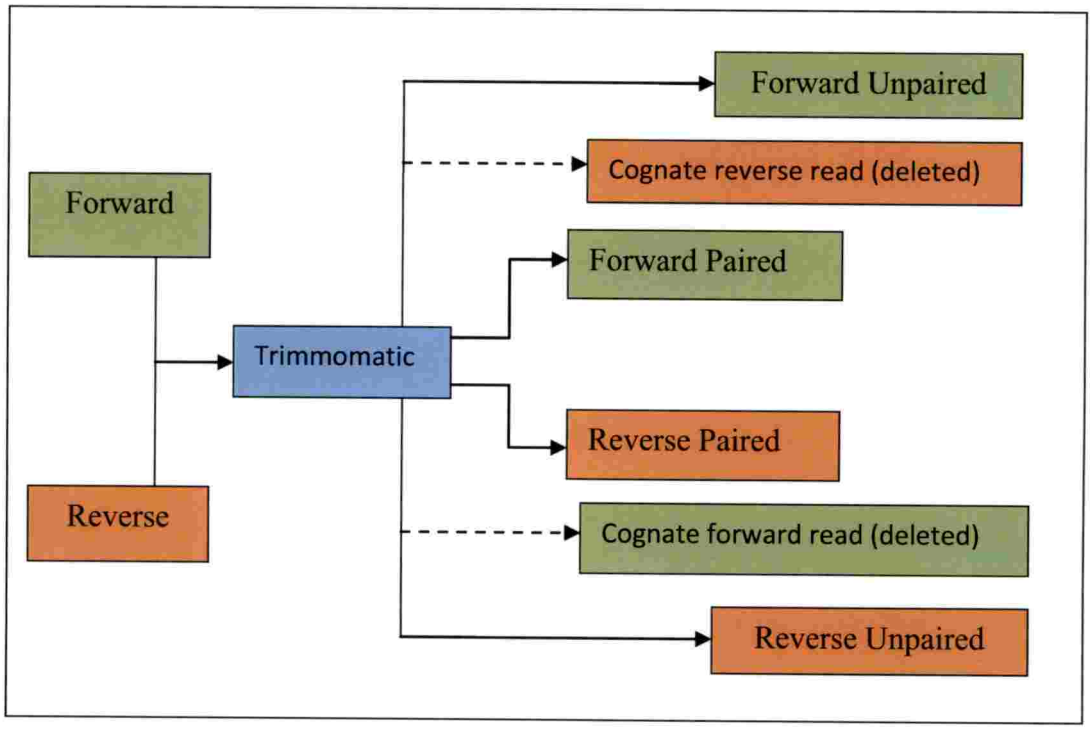

Figure: 2 Trimmomatic workflow for paired reads

3.3 *DE NOVO* ASSEMBLY USING TRINITY

For *de novo* assembly Trinity (version Trinity- v2.4.0) was used (Haas *et al.*, 2011). The Perl script program consists of 3 steps Inchworm, Chrysalis, Butterfly. Trinity exports the final output in fasta format after assessing the quality of the reads. Trinity was downloaded from https://github.com/trinityrnaseq/trinityrnaseq/releases.

Trinity normally performs assembling at a single k-mer size, hence no merging was done. Based on the length and number of reads, the time for *de novo* assembly varies. Trinity was run with initial parameters set to :

--seqType fq --left SRR873449_TRIM1.fastq --right SRR873449_TRIM2.fastq --CPU 8 --max_memory 100G

where SRR873449_TRIM1.fastq and SRR873449_TRIM2.fastq where the two trimmed reads.

3.4 CAP3

Single-Assembler Multiple-Parameter (SAMP) strategy (Iorizzo *et al.*, 2011) was employed which uses raw input data assembled with different parameters and assembled with CAP3. It was used to reduce the number of *de novo* assembled transcripts.

CAP3 is a 3 step sequence assembly and clustering program (Figure 3). It starts by clipping 5′ and 3′ low-quality regions, merges two overlapping sequences to make contigs and finally aligns the reads with the base quality values (Huang and Madan, 1999).

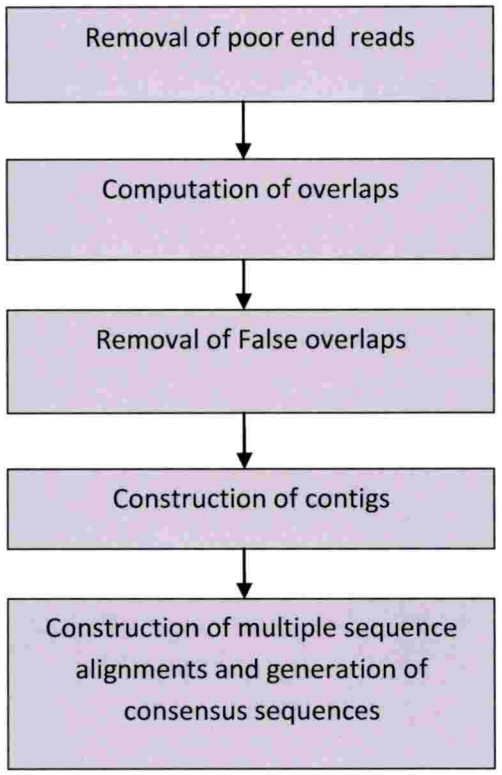

Figure: 3 Steps in CAP3 assembly

A standalone version of CAP3 compatible for Linux was downloaded from http://seq.cs.iastate.edu/cap3.html. The downloaded file was extracted and input file for assembling was copied into it. The command for CAP3 was given as -

./cap3 Trinity.fasta

where Trinity.fasta was the output of *de novo* assembly using Trinity.

Options in CAP3 (default values) (Huang and Madan, 1999) :

-a N specify band expansion size N > 10 (20)

-b N specify base quality cutoff for differences N > 15 (20)

-c N specify base quality cutoff for clipping N > 5 (12)

-d N specify max qscore sum at differences N > 20 (200)

-e N specify clearance between no. of diff N > 10 (30)

-f N specify max gap length in any overlap N > 1 (20)

-g N specify gap penalty factor N > 0 (6)

-h N specify max overhang percent length N > 2 (20)

-i N specify segment pair score cutoff N > 20 (40)

-j N specify chain score cutoff N > 30 (80)

-k N specify end clipping flag N >= 0 (1)

-m N specify match score factor N > 0 (2)

-n N specify mismatch score factor N < 0 (-5)

-o N specify overlap length cutoff > 15 (40)

-p N specify overlap percent identity cutoff N > 65 (90)

-r N specify reverse orientation value N >= 0 (1)

-s N specify overlap similarity score cutoff N > 250 (900)

-t N specify max number of word matches N > 30 (300)

-u N specify min number of constraints for correction N > 0 (3)

-v N specify min number of constraints for linking N > 0 (2)

-w N specify file name for clipping information (none)

-x N specify prefix string for output file names (cap)

-y N specify clipping range N > 5 (100)

-z N specify min no. of good reads at clip pos N > 0 (3)

3.5 MARKER PREDICTION

For the obtained contigs SSR and SNP marker prediction were done using various tools.

3.5.1 QualitySNP

The standalone version for QualitySNP was downloaded from http://www.bioinformatics.nl/tools/snpweb/download2.html. It is an efficient tool for discovering SNPs particularly insertions/deletions (indels). The QualitySNP detects SNPs in 4 steps - Assembly of sequences using CAP3 clustering, analyzing the alignment information, detecting SNP and haplotype and finally the discovery of non-synonymous SNP.

The file named QualitySNP11102007.tar.gz was downloaded and extracted and compiled using

% make all

After making QualitySNP, the assembled 8547 contigs were run with the following commands-

% Getalignmentinfo testseq.cap 4, (4- default minimal cluster size)

After getting alignment information, these steps were done simultaneously

% Getavailcontigseq filename.cap

% Getavailcontigqual filename.cap

% QualitySNP filename.cap min-allelesize lowqual5side similarity1 similarity2 lowqual3side weightlowqual min-confidencescore

where Min-allelesize is the minimum size of alleles of SNP (default - 2), lowqual5side - the length of the low quality region at the 5' end of sequence (default -30) similarity1 is the similarity on one polymorphic site (default - 0.75) similarity2 is the similarity on all polymorphic sites (default - 0.8) lowqual3side is the low quality region of 3' side (default - 0.2) weightlowqual is the weight value of the low quality region (default - 0.5) min-confidence score is the minimal confidence score (default - 2).

Next step was the most crucial one, *ie* identification of non-synonymous SNPs and was done using Fasty34.

```
% fasty34_t allavailcontigseqwithSNP Viridiplantae –b 6 –d 6 –Q >
  allavailcontigseqwithSNP.fasty
% GetnonsySNPfasty availcontigseq allavailcontigseqwithSNP
  allavailcontigseqwithSNP.fasty
```

where Viridiplantae is the protein database, "availcontigseq" contains the consensus sequences of contigs with SNPs, As these sequences are not curated, they may contain padding symbols ("*"), which may indicate either insertions and/or deletions in the sequences, but in many cases these may be caused by sequencing errors and "allavailcontigseqwithSNP" contains the consensus sequences of SNP-containing contigs which did not contain any insertions or deletions.

Results obtained were classified into allavailSNP- total SNPs detected, Ssnpcodingdata- corresponding to list of synonymous SNPs, Nssnpcodingdata - list of Non-Synonymous SNPs, Ssnpfastydata - list showing the transcribed sequence of the SNPs, Nssnpfastydata - list showing the transcribed sequence of the SNPs, Indelsnpdata – list of Indels.

3.5.2 AutoSNP

AutoSNP is an online tool for detecting SNPs based on the frequency of occurrence of polymorphisms and co-segregation of multiple SNPs. It uses the d2 cluster and cap3 for clustering and aligning the input data. SNP detection is being carried out using redundancy score and co-segregation score. Co-segregation score corresponds to the percentage of other SNPs with an identical segregation and redundancy score refers to the minimum number of reads per allele.

AutoSNP takes input either in the form of fasta sequences or ace file. Command for running AutoSNP is-

```
perl cap3SNP (-f <fasta name> | -a <ace name>)
```

It also provides option to create tab delimited text files and zip files.

3.5.3 MISA

MISA (MIcroSAtellite identification tool) was downloaded from http://pgrc.ipk-gatersleben.de/misa/download/misa.pl. The command given for executing MISA was

```
perl misa.pl <FASTAfile>
```

where <FASTAfile> corresponds to contig files containing DNA sequences in FASTA format.

Default unit size / minimum number of repeats condition set for identifying microsatellites in MISA is (1/10) (2/6) (3/5) (4/5) (5/5) (6/5). If a sequence fails to achieve the minimum number of repeats, then it will go undetected.

3.5.4 SSRIT

Simple Sequence Repeat Identification Tool was downloaded from ftp://ftp.gramene.org/pub/gramene/archives/software/scripts/ssr.pl. The command given for executing MISA was

```
perl ssr.pl <FASTAfile> >SSRIT_OUTPUT
```

where FASTAfile corresponded to the contig sequences. The default unit size / minimum number of repeats condition set for identifying microsatellites in SSRIT is (2/6) (3/5) (4/5) (5/5) (6/5).

3.6 RESISTANT VIRUS GENE DATABASE

A leaf blight resistant database was constructed for screening the molecular markers predicted. A database was constructed manually from protein sequences obtained from different leaf blight resistant genes from different plants. The sequences were retrieved from the UniProt Knowledgebase (UniProtKB) (https://www.uniprot.org/help/uniprotkb) which accounts for the protein information.

The sequence duplication within the resistant gene sequences was removed using the command-

```
awk '/^>/{f=!d[$1];d[$1]=1}f' in.fa > out.fa
```

After removing duplication a blight disease resistant database was constructed using the command-

makeblastdb -in UNIPROT_SEQ -out leafblightdatabase -dbtype prot -parse_seqids

where UNIPROT_SEQ was the set of protein sequences corresponding to leaf blight resistant genes.

The desired sequence with the contig ID was retrieved from the CAP3 output using the seqretrieve command.

```
perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$_}=1 if @ARGV'
CONTIGLIST CAP3OUTPUT > retrieved_output
```

where CONTIGLIST contained the set of contig IDs. The seqretrieve command was done for both SSR and SNP and the sequences were retrieved and further processed.

The sequence for primer designing was chosen based on the percentage identity and e-value obtained on blastx against the resistant database created. The command given was-

```
blastx -query INPUT -out OUTPUT -outfmt 6 -db leafblightdatabase
```

where INPUT file refers to the set of contig sequences that contain the SNP/SSR.

3.7 PRIMER DESIGNING

Primer Designing for the predicted SNPs and SSRs using QualitySNP and MISA was done using Primer3plus. 5 contigs each for SNP and SSR were taken and primers were designed using the web interface of Primer3plus tool. The primer designing takes into account certain criteria such as Product Size Range (ranging from 150 - 1000 bp), Primer Size, Primer Tm, Max Tm Difference, Primer GC%, Concentration of monovalent cations, Concentration of divalent cations and Concentration of dNTPs where user could give a minimum, optimum and maximum values.

The primer design was done with SSR and SNP site serving to be the target site. The primer length was set between 20-22 bp, Primer Tm between 55-60 °C, GC content between 55-60%, product size between 200-600 bp, Max Tm

difference 5°C and remaining conditions were set to default (Untergasser *et al.*, 2007; 2012).

3.8 PRIMER SYNTHESIS

The 20 designed primers sequences (both forward and reverse) were sent to IDT technologies for synthesizing (Figure 4).

3.9 VALIDATION OF SNP AND SSR MARKERS FOR TLB RESISTANCE

The *in silico* predicted markers need to be validated for assuring their ability to differentiate susceptible and tolerant varieties. The validation was done using PCR with the designed primers using resistant and susceptible DNA samples in Agarose Gel Electrophoresis.

3.9.1 Genomic DNA isolation

A total of six taro varieties were taken which included 3 TLB resistant and 3 TLB susceptible varieties based on field trials at Central Tuber Crop Research Institute (CTCRI), Thiruvananthapuram.

Fresh young leaves from the plants were collected in small plastic bags and were bought to the lab. CTAB method proposed by Doyle and Doyle (1987), and modified by Sharma *et al.* (2008) was used for the isolation. 160 mg of leaf tissue was weighed and grounded into a fine powder using liquid nitrogen in an autoclaved mortar and pestle. 2 ml of freshly prepared extraction buffer (Appendix I) was added to mortar before sample get thawed up. The contents were transferred to a sterile 2 ml Eppendorf tubes and 5 µl of proteinase K (10mg/ml) was added to the tubes. The tubes were then incubated at 37 °C with intermittent shaking for 30 minutes. The tubes were then again incubated at 65 °C for 30 minutes followed by centrifugation at 12,000 rpm for 15 minutes. The supernatant obtained was transferred to a fresh tube. An equal volume of chloroform: isoamyl alcohol (24:1) was added to it and mixed thoroughly by inversion. The tubes were then allowed to stand at room temperature for 5

IDT®
INTEGRATED DNA TECHNOLOGIES THE CUSTOM BIOLOGY COMPANY

NORMALIZED TUBE ORDER FORM

Name of Primer	Primer Sequence (5' - 3)	No of Base pair	Concentration	
CeSNP1F	TCTCCACCACTTCCTCCTCT	20	25 nmole DNA Oligo	
CeSNP1R	GAGTCTTCCACGTCACTTGC	20	25 nmole DNA Oligo	
CeSNP2F	CTGACCTTGCCTTTGGACTC	20	25 nmole DNA Oligo	
CeSNP2R	ACTCGTCCAGCCTTCTTCAC	20	25 nmole DNA Oligo	
CeSNP3F	GGTACACCAGTTGCTCACGA	20	25 nmole DNA Oligo	
CeSNP3R	GCGAGCGAGACGTACAAGAT	20	25 nmole DNA Oligo	
CeSNP4F	GCACTCTTCACTCGTGTTGC	20	25 nmole DNA Oligo	
CeSNP4R	CCTTCCTTCACCAGAACTGC	20	25 nmole DNA Oligo	
CeSNP5F	CGAGAAGGGTCCCAGGTACT	20	25 nmole DNA Oligo	
CeSNP5R	GCCAGCCACCACTATCTCTC	20	25 nmole DNA Oligo	
CeSSR1F	CAGGGTTTCCATTACCTCCTC	21	25 nmole DNA Oligo	
CeSSR1R	GAGCTTTGTGAGGTCCAGATG	21	25 nmole DNA Oligo	
CeSSR2F	CTAGTCAGTCCTGGCAAAGC	20	25 nmole DNA Oligo	
CeSSR2R	GCTCAGAGGTTAGAGCATCG	20	25 nmole DNA Oligo	
CeSSR3F	CTGTGTGAAGGAAGCGAAGAG	21	25 nmole DNA Oligo	
CeSSR3R	CCAATCAGGTCAGAACACCAC	21	25 nmole DNA Oligo	
CeSSR4F	CCACCAGAACAACACTCTTCG	21	25 nmole DNA Oligo	
CeSSR4R	CGCTCCCTCTCTTTCTGTTCT	21	25 nmole DNA Oligo	
CeSSR5F	CAGCAACCCTCAGGTGTAGAG	21	25 nmole DNA Oligo	
CeSSR5R	CTGCGTTTCCTTGATGATCC	20	25 nmole DNA Oligo	
	Total number of bases	**407**		

Figure 4. Primer Order Form

minutes to ensure phase separation. The tubes were then again centrifuged at 12,000 rpm for 15 minutes at room temperature. The upper aqueous phase of the tubes was transferred to fresh tubes using cut tips. An equal volume of chloroform: isoamyl alcohol (24:1) was again added to the tubes and mixed gently by inversion. After inversion, the tubes were centrifuged at 12,000 rpm for 15 minutes at room temperature. The resultant upper aqueous phase was transferred to new tubes and an equal volume of isopropanol was added to it. The tubes were then gently mixed until DNA threads get formed. The threads formed were then centrifuged at 10,000 rpm for 10 minutes. The precipitated DNA was then washed using 70 % ethanol for 2-3 times. The pellets were then air dried to remove the traces of ethanol and was finally dissolved in 100 µl TE buffer (Appendix II). RNase 5 µl (10ng/µl) was added to the tubes and incubated at 37 °C for 1 hour. After RNase treatment the DNA was properly labeled and stored at -20 °C freezer.

Table 1. List of taro varieties selected for DNA isolation

Sl No.	Susceptible varieties	Tolerant varieties
1	Sree Rashmi	Muktakeshi
2	Sree Kiran	Bhu Kripa (Field tolerant)
3	Telia	Bhu Sree (Field tolerant)

3.9.1.1 Analysis of DNA using Agarose Gel Electrophoresis

Agarose gel electrophoresis (0.8%) was used for checking the quality of the DNA obtained. The casting tray and comb was cleaned and assembled to make a mold on a plane surface. 0.8% agarose (Sigma Aldrich) was dissolved in 1X TBE (Appendix III) and melted by boiling for 1-3 minutes. 0.4 µl of EtBr was added to the conical flask after the temperature gets lowered and mixed well. The molten gel was then poured onto the casting tray and allowed to solidify. The combs were removed after 10-15 minutes and the gel was transferred to the electrophoretic system containing TBE. Sufficient buffer was added to the tank to

ensure gel get immersed completely. 5 µl of DNA along with 3 µl loading dye was mixed and loaded into the wells using a micropipette. The gel was then allowed to run for 40 minutes at 100V. The gel was then visualized under UV light for visualizing the DNA using the gel documentation system (G: Box, M/S Syngene).

3.9.1.2 Quantification of DNA

The quantification of DNA was done using Nanodrop® ND-100 by taking 1 µL of each DNA sample with TE buffer as blank. For each sample information regarding concentration of DNA(ng/µL), A260/230 and A260/280 ratio were noted down.

3.9.2 Dilution of DNA

The DNA samples were diluted to obtain a uniform concentration. The dilution was done using sterile distilled water based on the concentration of DNA present in the sample.

3.9.3 Dilution of the primer

The primers synthesized by IDT were centrifuged and dissolved in sterile distilled water for preparing master stock inside a Laminar Air Flow chamber. The primers were dissolved according to the specification sheet provided. The master stock was prepared for obtaining a concentration of 100µM. The master stock was again diluted to get a working stock for PCR reactions.

3.10 PCR AMPLIFICATION

The annealing temperature for the PCR reaction was calculated using the formula

$$Ta = Tm-5$$

where Ta and Tm corresponds to annealing temperature and melting temperature respectively.

For determining the efficiency of primers, the amplified PCR products were checked by AGE. The PCR products were resolved in 3% AGE with 100 bp ladder. The gel was then visualized under UV light of G: Box gel documentation system using GeneSyS software (M/s. Syngene). Band quality was observed and scored to validate the primers. PCR master mix was prepared for a volume of 15 µl with DNA sample, forward and reverse primer, $MgCl_2$, dNTPs, *Taq* Buffer, *Taq* polymerase and autoclaved distilled water (Appendix V).

3.11 VALIDATION OF SNP

For validation, two samples (one TLB resistant and one TLB susceptible variety) were taken against the five primer sets and PCR was done. A total of 15 µl reaction with 40ng/ µl genomic DNA, 0.25 µM of each forward and reverse primer (CeSNP1, CeSNP2, CeSNP3, CeSNP4 and CeSNP5), 1U *Taq* DNA polymerase, 0.25 mM of dNTP, 1X *Taq* buffer, 1.5 mM $MgCl_2$ and autoclaved ultrapure water. Amplifications were done in a BioRad C1000^{TM} thermal Cycler programmed with an initial denaturation of 3 min. at 94°C then 30 cycles of 45-second denaturation at 94°C, 1-minute annealing (different Ta for different primers), 1-minute extension at 72°C and a final extension of 10-minutes at 72°C. The amplification of PCR products was then analyzed in 3% agarose gel electrophoresis. Based on the prominent single band appearance at desired product size, primers were selected. The selected primers were again amplified and the PCR products were sequenced.

3.11.1 Clustal Omega

Clustal is a graphical interface for performing multiple sequence alignment of nucleotide and protein sequences. Varying versions were found for Clustal program with Clustal Omega (ClustalO) being the latest one. ClustalX is an offline interface for multiple sequence alignment whereas Clustal Omega, on the other hand, is a command line interface. It provides multiple sequence alignment of hundreds of sequences within a shorter time span. Alignment scores can be

calculated and desired sequences could be highlighted. Clustal Omega can be run online at http://www.ebi.ac.uk/Tools/msa/clustalo/.

The multiple sequence alignment was done using Clustal Omega with the sequenced PCR products and contig sequences to validate the predicted SNPs.

3.12 VALIDATION AND SCREENING OF SSR

For screening, two samples (one TLB resistant and one TLB susceptible) were taken against the five primer sets and PCR was done. A total of 15 µl reaction with 40ng/ µl genomic DNA, 0.25 µM of each forward and reverse primer (CeSSR1, CeSSR2, CeSSR3, CeSSR4, CeSSR5), 1U *Taq* DNA polymerase, 0.25 mM of dNTP, 1X *Taq* buffer, 1.5 mM $MgCl_2$ and autoclaved ultrapure water. Amplifications were done in a BioRad C1000^{TM} thermal Cycler programmed with an initial denaturation of 3-minute at 94°C then 30 cycles of 45-second denaturation at 94°C, 1-minute annealing (various temperatures for different primers), 1-minute extension at 72°C and a final extension of 10-minutes at 72°C. The amplification of PCR products was then analyzed in 3% agarose gel electrophoresis.

Based on product size and banding pattern one among the primer was selected for further screening of the six DNA samples and PCR was done.

RESULTS

4. RESULTS

The results of the study entitled "Development of molecular markers for blight disease resistance in taro using bioinformatics tools" carried out at ICAR - CTCRI are presented in this chapter.

4.1 TARO SEQUENCE DATASET

The preliminary data set was obtained from NCBI with accession number SRX290678 in .sra format and was split into two reads -left/forward and right/reverse. About 6,479,882 sequences in fastq format were present and split into R1.fastq and R2.fastq. The splitted sequences were then taken up for further processing.

4.2 PRE-PROCESSING OF SEQUENCES

The taro sequence dataset obtained from NCBI was split into two reads and were processed by Trimmomatic. The sequences were checked for adaptors, bases with lower threshold quality, and length. The sequences which failed for the given parameters were trimmed off. The pre-processing step minimized the number of sequences and only good quality sequences were further taken up for *de novo* assembly.

A total of 160,048 sequences were removed from 6,479,882 sequences, minimizing the total sequences to be 6,319,834. The trimmed files were - SRR873449_TRIM1 and SRR873449_TRIM2.

4.3 *DE NOVO* ASSEMBLY OF SEQUENCES

De novo assembly of the trimmed fastq sequence was carried out with Trinity to give output as Trinity.fasta. The assembly generated about 79,608 sequences.

4.4 ASSEMBLY OF SEQUENCES USING CAP3

After *de novo* assembly of the sequences, CAP3 was run to obtain the assembled reads and singlets. It also computed the overlaps among the reads and removed false reads. A total of 8,547 contigs and 59,242 singlets were obtained with the default parameters set. The contigs were then taken up for marker prediction and development. Apart from contigs and singlets, a links file, an ace file, a quality file, info file and con file were also produced.

4.5 MARKER PREDICTION

Molecular marker prediction for the obtained 8547 contigs was done successfully using different pipelines. MISA and SSRIT were chosen for predicting SSRs and QualitySNP and AutoSNP were chosen for predicting SNPs.

4.5.1 Identification of SNP using QualitySNP

562 SNPs identified using QualitySNP are summarized in the table below (Table 2).

Table 2. Distribution of transition and transversion of SNPs from QualitySNP

Characterization	Type of nucleotide substitution	Number of SNPs	Total
TRANSITION	C/T	81	**180**
	G/A	99	
TRANSVERSION	A/C	30	**138**
	A/T	30	
	C/G	39	
	T/G	39	

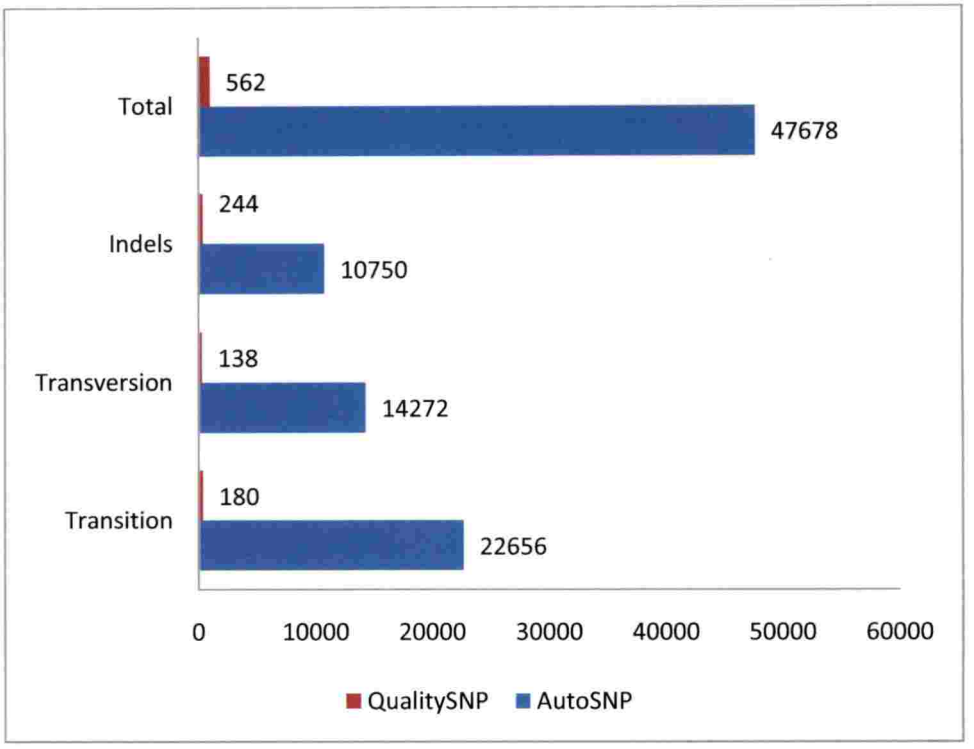

Figure 5. Distribution of SNP polymorphisms in QualitySNP and AutoSNP

Of the identified SNPs 518 were non-synonymous which indicated a change in translational product and 44 were synonymous. 180 transitions, 138 transversions, and 244 indels were obtained among the 562 SNPs identified. Both C/T and A/G transitions were observed to be same, however, C/G transversion dominated A/C, A/T, and T/G.

4.5.2 Identification of SNP using AutoSNP

AutoSNP detected a massive total of 47,678 SNPs. The output was displayed in HTML format with summary and list of contigs. The detected SNPs consist of 22656 transitions, 14272 transversions, and 10750 InDels. The SNP occurrence frequency was found out to be 0.52/100 bp. The list of SNPs detected using AutoSNP is shown in Table 3.

Table 3. Distribution of transition and transversion of SNPs from AutoSNP

Characterization	Type of nucleotide substitution	Number of SNPs
Transition	C/T + G/A	22656
Transversion	A/C + A/T+ C/G+ T/G	14272

4.5.3 Comparative evaluation of SNP prediction tools

Both QualitySNP and AutoSNP were executed in a stand-alone mode. With the difference in the programme and parameters, a varying number of SNPs were produced. In comparison, AutoSNP produced 47,678 SNPs whereas QualitySNP identified 562 SNPs (Figure 5). The results are summarized in table 4.

Of the two, AutoSNP has a polymorphism ratio of 1.58 which is quite higher comparing to a healthy ratio of 1.30 by QualitySNP. QualitySNP doesn't need trace/quality files or genomic sequences for identifying SNPs whereas

AutoSNP cannot distinguish paralogs, leading to false detection of SNPs. QualitySNP was also capable of distinguishing between synonymous and non-synonymous SNPs. Hence contigs containing SNPs detected using QualitySNP were taken for primer designing.

Table 4. Comparison of AutoSNP and QualitySNP

SNP Tools	Number of SNPs	Transition to Tansversion Ratio
AutoSNP	47,678	1.58
QualitySNP	562	1.30

4.5.4 Identification of SSR using MISA

Two output files were created, "<FASTAfile>.misa which corresponds to a tablewise distribution of identified microsatellites and "<FASTAfile>.statistics" which summarizes the frequency of SSR according to their size (Table 5).

Using MISA 3034 SSRs were identified from 8547 contig sequences (Table 3). Dinucleotide repeats were the abundant ones accounting for 48.91%. SSRs with repeat motifs of 1–3 bp (mono-, di- and tri-)accounted for 99.28% of total SSRs detected. The distribution of different SSRs is being shown in Table 6.

Table 5. Summary of MISA based prediction of SSR

MISA - Result summary	
Total number of assembled transcripts examined	8547
Total size of assembled transcripts sequences (bp)	9121567
Total number of identified SSRs	3034
Number of SSR containing transcript sequences	2113
Number of sequences containing more than 1 SSR	610
Number of SSRs present in compound formation	393

Table 6. Category wise distribution of SSRs predicted using MISA

Type of SSR identified	No: of SSRs	Percentage (%)
Mono	967	31.87
Di	1484	48.91
Tri	558	18.30
Tetra	14	0.46
Penta	2	0.06
Hexa	9	0.20
Poly	0	0
Total	**3034**	**100**

4.5.5 Identification of SSR using SSRIT

An output file containing sequence ID, motif (repeat) type, no. of repeats, SSR start, SSR end and length of the sequence was displayed (Table 7).

Dinucleotide repeats were the abundant ones accounting for 75.13 %. SSRs with repeat motifs of 2-4 bp (di- , tri- and tetra-) accounted for 100 % of SSRs detected (Table 8). However, the algorithm doesn't detect any mono repeats.

Table 7. Summary of SSRIT based prediction of SSR

SSRIT - Result summary	
Total number of assembled transcripts examined	8547
Total size of assembled transcripts sequences (bp)	9121567
Total number of identified SSRs	1078
Number of SSR containing transcript sequences	916
Number of sequences containing more than 1 SSR	134

Table 8. Distribution of different classes of repeats identified in SSRIT

Type of SSR identified	No: of SSR	Percentage (%)
Mono	0	0
Di	810	75.13
Tri	254	23.56
Tetra	14	1.29
Penta	0	0
Hexa	0	0
Poly	0	0
Total	**1078**	**100**

4.5.6 Comparative evaluation of SSR prediction tools

MISA and SSRIT were used for identifying SSRs in the contigs. Both tools produced significant results with more number of SSR being reported by MISA, 3034 comparing to 1078 by SSRIT (Figure 6). MISA identified mono-, penta- and hexa- repeat in addition to di-, tri-, and tetra- repeats identified by SSRIT. In both the tools di- repeats were found out to be more in number, however, the type of repeats and their distribution varies among species.

The output generated from SSRIT needs to be inter-converted for better understanding, which would be difficult in larger datasets. MISA produced more types of repeats in a shorter duration of time comparing to SSRIT. Hence contigs containing SSRs detected using MISA were chosen up for primer designing.

4.6 LEAF BLIGHT RESISTANT DATABASE

The leaf blight resistant database was constructed from 42 different genes. The database comprised of 1199 sequences, both reviewed and unreviewed. The

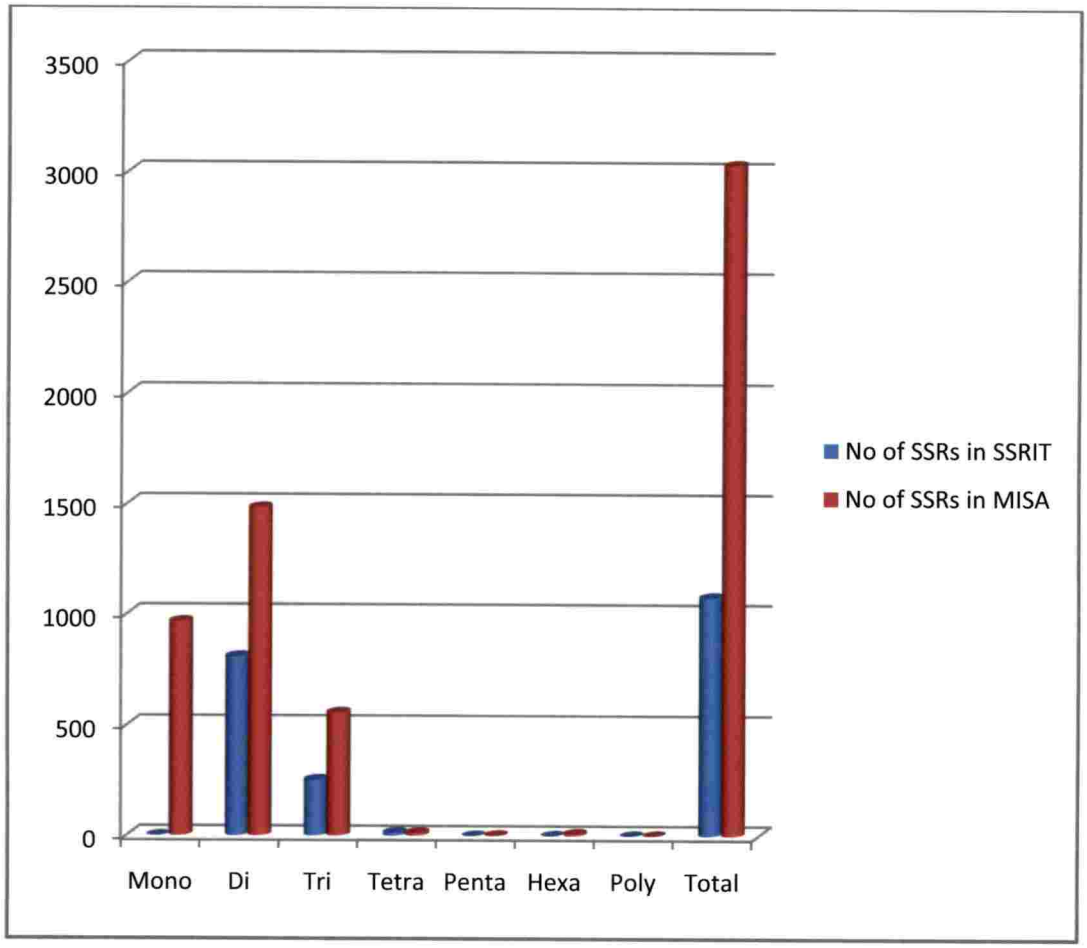

Figure 6. Distribution of SSR in MISA and SSRIT

number of sequences was reduced to 1012 after removing duplication. Leaf blight resistant database was constructed using this 1012 sequences and a file with six different extensions - .pin .phr .psq .pog .psi .psd were produced.

The desired contigs were selected from BLAST result based on higher percentage identity (90-100%) and lower E-values (≤ 0) and were retrieved using the seqretrieve command. Five contigs were selected for both SNP and SSR (Table 9).

Table 9. Predicted markers and selected markers for primer synthesis

Type of marker	No of sequences with polymorphism	No of sequences selected for primer synthesis
SNP	996	5
SSR	3034	5

4.7 PRIMER DESIGNING

Primer designing was done using Primer3plus. Of the 5 primer combinations displayed, one was selected for synthesizing based on GC (> 50%) content and Tm (55-60°C) values for each contig.

5 pairs of primers were designed for each contig of SNP (Table 10) and SSR (Table 11). Based on adequate product size, Tm and GC content a single primer pair was selected from the combinations and send for synthesis. A total of five forward and reverse primers for both SNP & SSR was sent (Table 12 & 13).

4.8 PRIMER SYNTHESIS

Primers were synthesized and delivered by a company named Integrated DNA Technologies, Inc. (IDT) in lyophilized form.

Table 10 --- List of SNP primers designed using Primer3Plus

Sl No.	Contig	Left Primer 5'- 3'				Right Primer 5'- 3'				Product Size(bp)
		Left Primer	Length(bp)	Tm(°C)	GC(%)	Right Primer	Length(bp)	Tm(°C)	GC(%)	
1	2330	CTCTCTCTCCACCACTTCCTC	21	58	57.1	GAGTCTTCCACGTCACTTGC	20	58.4	55	310
		CTCTCTCTCTCCACCACTTCC	21	58	57.1	GAGTCTTCCACGTCACTTGC	20	58.4	55	312
		CCACCACTTCCTCCTCTTCT	20	58.3	55	GAGTCTTCCACGTCACTTGC	20	58.4	55	302
		CTCCACCACTTCCTCCTCTT	20	58.3	55	GAGTCTTCCACGTCACTTGC	20	58.4	55	304
		TCTCCACCACTTCCTCCTCT	20	58.8	55	GAGTCTTCCACGTCACTTGC	20	58.4	55	305
2	3289	CTGACCTTGCCTTTGGACTC	20	59.8	55	AGGTACTTGGGAGCATACCG	20	59.1	55	381
		GCGTTACTGGTTCTCGGAAG	20	59.9	55	AGGTACTTGGGAGCATACCG	20	59.1	55	430
		CTGACCTTGCCTTTGGACTC	20	59.8	55	GTGTGGAAAGAGCAGCTGTG	20	59.8	55	490
		CTGACCTTGCCTTTGGACTC	20	59.8	55	ACTCGTCCAGCCTTCTTCAC	20	59.5	55	600
		GCGTTACTGGTTCTCGGAAG	20	59.9	55	GTGTGGAAAGAGCAGCTGTG	20	59.6	55	539
3	3577	ACGAGCTGGTGAACTTGGTG	20	61.3	55	GCGAGCGAGACGTACAAGAT	20	60.6	55	277
		GGTACACCAGTTGCTCACGA	20	59.8	55	GCGAGCGAGACGTACAAGAT	20	60.6	55	293
		GGGGTACACCAGTTGCTCAC	20	60.4	60	GCGAGCGAGACGTACAAGAT	20	60.6	55	295
		ATCCACCAGTGCACACTTCC	20	61	55	GACATCTCCTCCTCCCTTCC	20	60	60	250
		ATCCACCAGTGCACACTTCC	20	61	55	CCCACTGACATCTCCTCCTC	20	59.6	60	256
4	5624	AGAGAGAGAGAGAGGGGAGGA	21	58.7	57.1	CCCCAGAAGCCAACATCTAC	20	59.5	55	526
		AGAGAGAGAGAGGGGAGGACA	21	59.5	57.1	CCCCAGAAGCCAACATCTAC	20	59.5	55	524
		GGGTGGAGAGAGAGAGAGAGAG	22	58.8	59.1	CCCCAGAAGCCAACATCTAC	20	59.5	55	536
		AGAGAGAGAGAGAGGGGAGGAC	22	59.6	59.1	CCCCAGAAGCCAACATCTAC	20	59.5	55	526
		GCACTCTTCACTCGTGTTGC	20	59.6	55	CCTTCCTTCACCAGAACTGC	20	59.8	55	524
5	7006	CGAGAAGGGTCCCAGGTACT	20	60.5	60	GCCAGCCACCACTATCTCTC	20	59.8	60	252
		CGAGAAGGGTCCCAGGTACT	20	60.5	60	AGAAGCCTCCTTTTCCATCC	20	59.7	50	279
		CGAGAAGGGTCCCAGGTACT	20	60.5	60	GAAGCCTCCTTTTCCATCCT	20	59.7	50	278
		CGAGAAGGGTCCCAGGTACT	20	60.5	60	TCCTCTCTCCTTGGCATTTC	20	59.4	50	231
		CGAGAAGGGTCCCAGGTACT	20	60.5	60	TCTCTCCTTGGCATTTCTCC	20	59.4	50	228

Table 11 --- List of SSR primers designed using Primer3Plus

Sl No.	Contig	SSR	Left Primer 5'- 3'				Right Primer 5'- 3'				Product Size (bp)
			Left Primer	Length (bp)	Tm (°C)	GC (%)	Right Primer	Length (bp)	Tm (°C)	GC (%)	
1	1315(SSR1)	(cgg)6	CAGGGTTTCCATTACCTCCTC	21	59.8	52.4	GAGCTTTGTGAGGTCCAGATG	21	59.9	52.4	231
			CAGGGTTTCCATTACCTCCTC	21	59.8	52.4	GAGCCTCTTCAGGTGCTTCTT	21	60.1	52.4	152
			GGTTTCCATTACCTCCTCCAC	21	59.7	52.4	GCTTTGTGAGGTCCAGATGAG	21	59.9	52.4	226
			GGTTTCCATTACCTCCTCCAC	21	59.7	52.4	GAGCTTTGTGAGGTCCAGATG	21	59.9	52.4	228
			CAGGGTTTCCATTACCTCCTC	21	59.8	52.4	GTGAGGTCCAGATGAGGGTTT	21	60.4	52.4	224
2	6412(SSR2)	(ta)15	CTAGTCAGTCCTGGCAAAGC	20	57.7	55	CTTATGCCGTGGTAACTTCC	20	57.2	50	474
			GTCGCTCTGTCAGACCATAA	20	56.4	50	CTTATGCCGTGGTAACTTCC	20	57.2	50	596
			GGTCCTCGGTAACGAGACATA	21	59.1	52.4	AGCTCAGAGGTTAGAGCATCG	21	58.9	52.4	556
			ACTAGTCAGTCCTGGCAAAGC	21	58.6	52.4	CTCCAAATGTCGAGTTGCTC	20	58.4	50	509
			CTAGTCAGTCCTGGCAAAGC	20	57.7	55	GCTCAGAGGTTAGAGCATCG	20	57.8	55	603
3	6734(SSR3)	(ga)14	CTGTGTGAAGGAAGCGAAGAG	21	60.2	52.4	ATCAGGTCAGAACACCACCAG	21	60	52.4	194
			CTGTGTGAAGGAAGCGAAGAG	21	60.2	52.4	CAGGTCAGAACACCACCAGTT	21	60.1	52.4	192
			CTGTGTGAAGGAAGCGAAGAG	21	60.2	52.4	CCAATCAGGTCAGAACACCAC	21	60.4	52.4	197
			CTGTGTGAAGGAAGCGAAGAG	21	60.2	52.4	TCAGGTCAGAACACCACCAGT	21	60.6	52.4	193
			GTGTGAAGGAAGCGAAGAGG	20	60	55	ATCAGGTCAGAACACCACCAG	21	60	52.4	192
4	7825(SSR4)	(ga)11	CTCTTCGCGGCTTTTCTCTAC	21	60.6	52.4	CGCTCCCTCTCTTTCTGTTCT	21	60.1	52.4	182
			GCGGCTTTTCTCTACTTCTGC	21	60.6	52.4	CGCTCCCTCTCTTTCTGTTCT	21	60.1	52.4	176
			CCACCAGAACAACACTCTTCG	21	60.7	52.4	CGCTCCCTCTCTTTCTGTTCT	21	60.1	52.4	196
			CTCTTCGCGGCTTTTCTCTA	20	59.9	50	CGCTCCCTCTCTTTCTGTTCT	21	60.1	52.4	182
			TCGCGGCTTTTCTCTACTTC	20	59.7	50	CGCTCCCTCTCTTTCTGTTCT	21	60.1	52.4	178
5	8428(SSR5)	(ag)12	CAGCAACCCTCAGGTGTAGAG	21	59.9	57.1	CTGCGTTTCCTTGATGATCC	20	60.6	50	226
			GAACAGCAACCCTCAGGTGTA	21	60.2	52.4	CCCCAGTTAGGGTTTCCTCT	20	59.4	55	168
			GAACAGCAACCCTCAGGTGTA	21	60.2	52.4	CTGCGTTTCCTTGATGATCC	20	60.6	50	229
			ACAGCAACCCTCAGGTGTAGA	21	59.8	52.4	CCCCAGTTAGGGTTTCCTCT	20	59.4	55	166
			ACAGCAACCCTCAGGTGTAGA	21	59.8	52.4	CTGCGTTTCCTTGATGATCC	20	60.6	50	227

Table 12. Selected SNP primers for Synthesizing

Sl. No.	Name of Primer	Forward Primer (5'-3')	Tm(°C)	Reverese primer (5' - 3')	Tm(°C)	Product Size (bp)
1	**CeSNP1**	TCTCCACCACTTCCTCCTCT	58.8	GAGTCTTCCACGTCACTTGC	58.4	305
2	**CeSNP2**	CTGACCTTGCCTTTGGACTC	59.8	ACTCGTCCAGCCTTCTTCAC	59.5	600
3	**CeSNP3**	GGTACACCAGTTGCTCACGA	59.8	GCGAGCGAGACGTACAAGAT	60.6	293
4	**CeSNP4**	GCACTCTTCACTCGTGTTGC	59.6	CCTTCCTTCACCAGAACTGC	59.8	524
5	**CeSNP5**	CGAGAAGGGTCCCAGGTACT	60.5	GCCAGCCACCACTATCTCTC	59.8	252

Table 13. Selected SSR primers for Synthesizing

Sl. No.	Name of Primer	Forward Primer (5' - 3')	Tm(°C)	Reverese primer (5' - 3')	Tm(°C)	Product Size (bp)
1	**CeSSR1**	CAGGGTTTCCATTACCTCCTC	59.8	GAGCTTTGTGAGGTCCAGATG	59.9	231
2	**CeSSR2**	CTAGTCAGTCCTGGCAAAGC	57.7	GCTCAGAGGTTAGAGCATCG	57.8	603
3	**CeSSR3**	CTGTGTGAAGGAAGCGAAGAG	60.2	CCAATCAGGTCAGAACACCAC	60.4	197
4	**CeSSR4**	CCACCAGAACAACACTCTTCG	60.7	CGCTCCCTCTCTTTCTGTTCT	60.1	196
5	**CeSSR5**	CAGCAACCCTCAGGTGTAGAG	59.9	CTGCGTTTCCTTGATGATCC	60.6	226

4.9 VALIDATION OF SNP AND SSR MARKERS FOR TLB RESISTANCE

The *in-silico* predicted markers were validated using the designed primers against TLB resistant and susceptible varieties.

4.9.1 ISOLATION OF DNA

DNA isolation of 6 taro leaf samples were done using the CTAB method and were stored at -20 °C.

4.9.1.1 Analysis of DNA

The DNA samples isolated using the CTAB method were analyzed using 0.8% agarose gel electrophoresis (Plate 1). Although some shearing were present the samples showed clear bands.

4.9.1.2 Quantification of DNA

Quantification of DNA was done using NanoDrop® ND-100. The concentration of DNA(ng/µL), $A_{260/230}$, $A_{260/280}$ obtained are shown below (Table 14).

Table 14. Quantification of DNA

Sl. No.	Sample Name	Concentration of DNA(ng/µL)	$A_{260/230}$	$A_{260/280}$
1	Muktakeshi	363.116	1.28	2.08
2	Bhu Kripa	777.059	1.68	2.20
3	Bhu Sree	1180.209	1.62	2.09
4	Sree Rashmi	3028.352	2.03	2.19
5	Sree Kiran	2028.846	1.85	2.19
6	Telia	173.613	0.69	1.80

Plate 1: 0.8% EtBr stained agarose gel showing DNA of 6 taro samples after electrophoresis (5 µl DNA sample + 1 µl 1X loading dye)

A- 100 bp ladder, B- Muktakeshi , C- Bhu Kripa , D- Bhu Sree,
E- Sree Rashmi, F- Sree Kiran and G- Telia

4.9.3 Dilution of The DNA

Based on the stock concentration a working stock of 10 ng/µL was prepared using the dilution volume obtained. Sterile distilled water was used for dilution and the samples were stored at -20°C.

4.9.4 Dilution of the primer

A working stock of 10 µM was prepared. The master stock of primers with 100 µM concentration was properly labelled and stored at -20°C. The working stock was taken for preparing PCR cocktail.

4.10 PCR

PCR reaction for the designed primers was carried out using the designed primers and the calculated annealing temperatures (Table 15).

Table 15. Annealing Temperature for the synthesized primers

Sl No.	Name of the primer	Annealing temperature - Ta (°C)
1	CeSNP1	56
2	CeSNP2	56
3	CeSNP3	56
4	CeSNP4	56
5	CeSNP5	56
6	CeSSR1	56
7	CeSSR2	54
8	CeSSR3	56
9	CeSSR4	56
10	CeSSR5	56

4.11 VALIDATION AND SCREENING OF SNP

The diluted DNA samples of one resistant and one susceptible taro variety was screened against the five SNP primers- CeSNP1, CeSNP2, CeSNP3, CeSNP4 and CeSNP5 using PCR in AGE. Banding pattern in resistant and susceptible varieties were looked upon.

CeSNP3 produced a prominent thick band at the desired product size (293 bp) (Plate 2). The prominent single band in both resistant and susceptible varieties confirmed the markers ability to distinguish resistant and susceptible varieties.

The PCR products of CeSNP3 were sequenced using Genei Laboratories Pvt Ltd., Bangalore using 3500 capillary DNA Genetic Analyzer (Applied Biosystem). Replicates were also sent in order to avoid sequencing errors. The sequences obtained were then aligned against corresponding contigs using Clustal Omega (ClustalO) (Figure 7).

Sequence bands from the resistant variety Muktakeshi were aligned against sequence from Contig 3577 from which the primer CeSNP3 was designed. The results showed that sequence with CeSNP3 showed SNP at positions 359, 377, 402, 452 as predicted using QualitySNP. The predicted SNPs were to be G/A at 359th position, A/G at 377th position, G/C at 401st position, G/A at 452th position.

4.12 VALIDATION AND SCREENING OF SSR

The diluted DNA samples of one resistant and one susceptible taro variety were screened against five SSR primers - CeSSR1, CeSSR2, CeSSR3, CeSSR4, CeSSR5. The PCR products were validated using AGE and banding pattern between resistant and susceptible varieties was looked upon.

The primer CeSSR4 produced some banding at the desired product size of 196 bp only among the resistant variety (Plate 3) and was selected to screen the remaining samples.

The presence of bands which could clearly distinguish between resistant and susceptible varieties were looked upon. Banding was observed at the desired

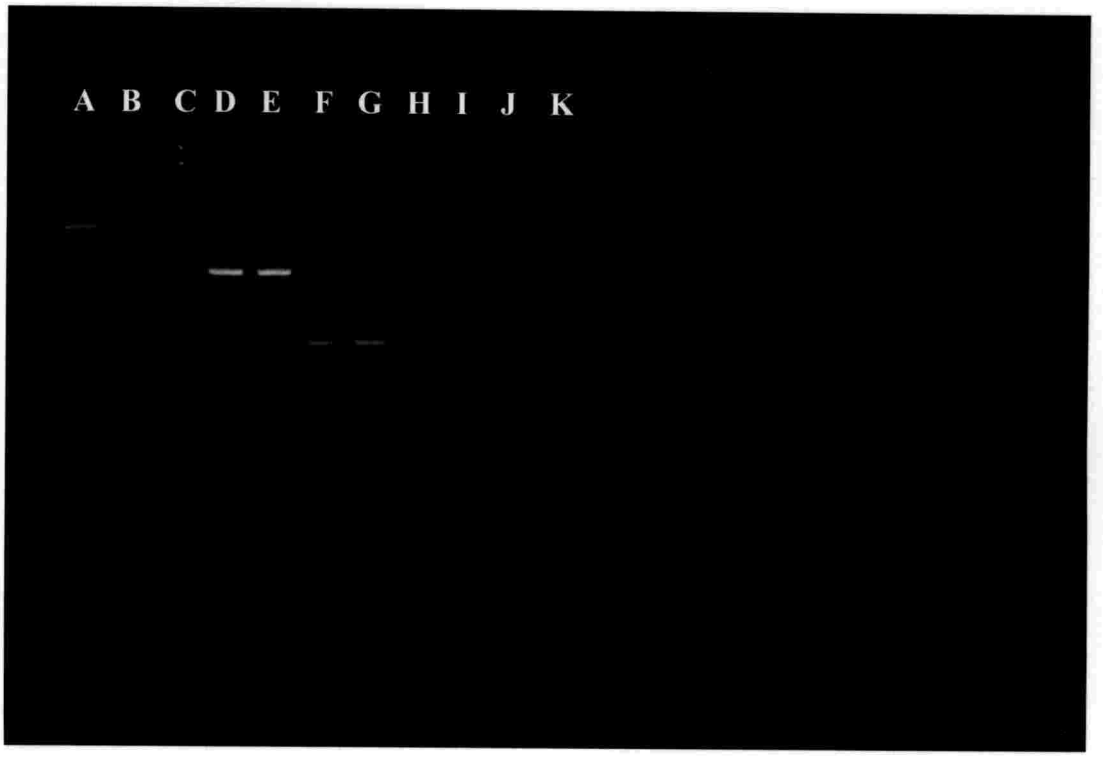

Plate 2. Screening of CeSNP1, CeSNP2, CeSNP3, CeSNP4 and CeSNP5 in 3% agarose gel.

Expected product Size: CeSNP1 - 305 bp , CeSNP2- 600bp, CeSNP3- 293bp, CeSNP4-524bp and CeSNP5- 252bp

A- 100bp ladder, **B-** Muktakeshi CeSNP1, **C-** Sree Rashmi CeSNP1, **D-** Muktakeshi CeSNP2, **E-** Sree Rashmi CeSNP2, **F-** Muktakeshi CeSNP3, **G-** Sree Rashmi CeSNP3, **H-** Muktakeshi CeSNP4, **I-** Sree Rashmi CeSNP4, **J-** Muktakeshi CeSNP5, **K-** Sree Rashmi CeSNP5

```
CLUSTAL O(1.2.4) multiple sequence alignment


Contig3577      CTTGGAATCCACCAGTGCACACTTCCGAACCAAAAACAATGAAACCCACACGGCACAGAC    60
R1              ------------------------------------------------------------    0


Contig3577      AACACCATTTAATCAGCCAAGAGTAGAAAATTTGATCCACAAAGAAAACCGGGCATTTCT    120
R1              ------------------------------------------------------------    0


Contig3577      CTTTTACATGTCAAAGCAGCCTCCTTTTTTCCATGTAACTGCGAGAAAAACAGAAGAGGG    180
R1              ------------------------------------------------------------    0


Contig3577      ATGGGGGCAACAACGCCTGCAGATTCCGACATCTACAAGGTTTTACAGCAGTAAAGGGAA    240
R1              ------------------------------------------------------------    0


Contig3577      GGGAGGAGGAGATGTCAGTGGGAAATTTGGGAACACTCTAAACGGGGGAATTGAGCGGGG    300
R1              ------------------------------------------------------------    0


Contig3577      GTACACCAGTTGCTCACGAGCTGGTGAACTTGGTGACGGCCTTGGTGCCCTCGGAGACAG    360
R1              ------------------------TGAACTTGGTGACCGCCTTGGTGCCCTCGGAGACGG    36
                                        ************* ******************** *

Contig3577      CGTGCTTGGCGAGCTCACCGGGGAGGACGAGGCGGACGGACGTCTGGATCTCCCGGGAGG    420
R1              CGTGCTTGGCGAGCTCGCCGGGGAGGACGAGGCGGACGGAGGTCTGGATCTCCCGGGAGG    96
                **************** ********************** *******************

Contig3577      TGATGGTGGGCTTCTTGTTGTAGCGGGCGAGACGGGATGCCTCCTGGGCGAGCTTCTCGA    480
R1              TGATGGTGGGCTTCTTGTTGTAGCGGGCGAGGCGGGATGCCTCCTGGGCGAGCTTCTCGA    156
                ****************************** ***************************

Contig3577      AGATGTCGTTGATGAAGCTGTTCATGATGACCATGGCCTTGCTGGAGATGCCGATGTCCG    540
R1              AGATGTCGTTGATGAAGCTGTTCATGATGCCCATGGCCTTGCTGGAGATGCCCAATGTCC    216
                ***************************** ********************** *    *

Contig3577      GG--TGCACCTGCTTCAGCACCTTGAAGATGTAGATCTTGTACGTCTCGCTCGCCTTCTT    598
R1              CGGGTGCACCTGCTTCAGCCCCTGGAG--------------------------------    243
                 *  *************** *** **

Contig3577      CTTCATCTTCTTCTTCTTCTTGTCCCCG    626
R1              ----------------------------    243
```

Figure 7. ClustalX alignment of CeSNP3 with Muktakeshi

Contig3577- Contig sequence containing predicted SNP

R1- sequenced PCR product of Mukthakeshi with CeSNP3

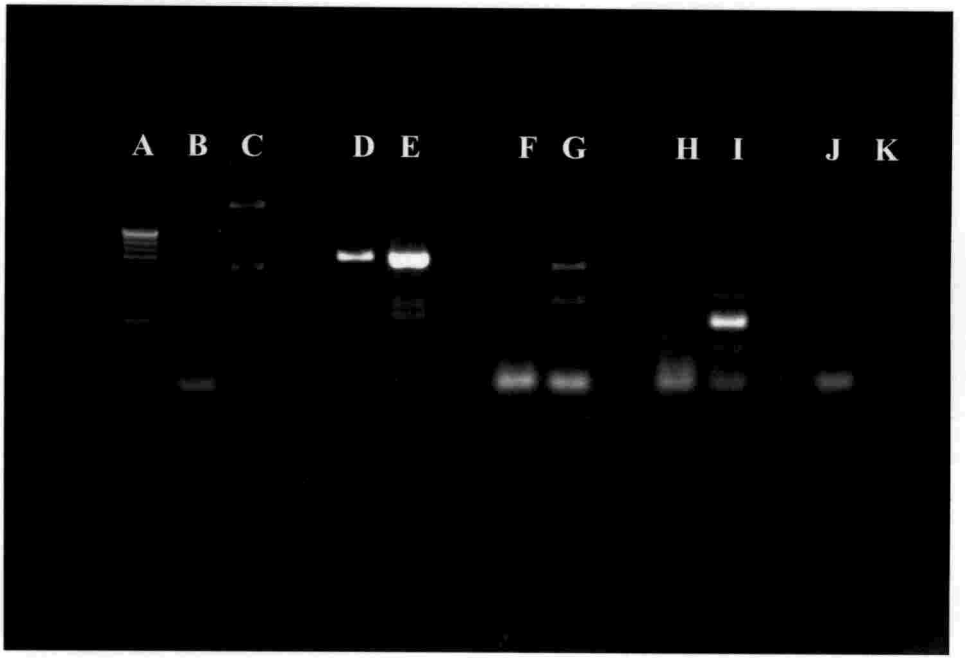

Plate 3. SSR screening against CeSSR1, CeSSR2, CeSSR3, CeSSR4, CeSSR5

Expected product Size: CeSSR1 -231bp, CeSSR2 - 603bp, CeSSR3 - 197bp, CeSSR4 - 196bp, CeSSR5 - 226bp

A- 100bp ladder, B-Muktakeshi CeSSR1, C-Sree Rashmi CeSSR1, D- Muktakeshi CeSSR2, E- Sree Rashmi CeSSR2, F- Muktakeshi CeSSR3, G- Sree Rashmi CeSSR3, H- Muktakeshi CeSSR4, I- Sree Rashmi CeSSR4, J- Muktakeshi CeSSR 5, K- Sree Rashmi CeSSR5

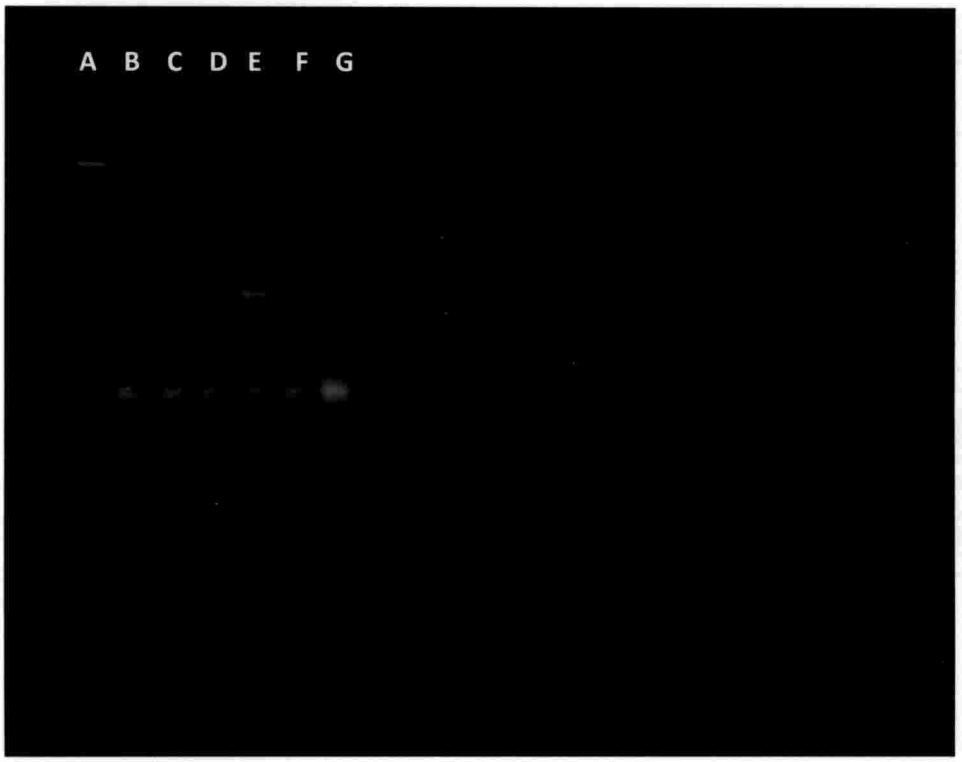

Plate 4. Gel image of CeSSR4

Expected Product Size - 196 bp

A - 100 bp ladder, **B** -Bhu Sree, **C**- Bhu Kripa, **D** - Muktakeshi,

E- Sree Rashmi, **F**- Sree Kiran, **G** - Telia

product size in Muktakeshi and were absent among others. Hence the designed SSR marker CeSSR4 was capable of differentiating between resistant and susceptible varieties.

Discussion

5. DISCUSSION

The results of the study entitled "Development of molecular markers for blight disease resistance in taro using bioinformatics tools" carried out at ICAR - CTCRI are discussed in this chapter.

Molecular markers have got wider acceptance globally in spite of its type. With emerging technologies and innovations, there is a trend to overcome the traditional methods and techniques. Development of molecular markers using information publicly available in the biological databases has been attributed with enhanced credibility over the years. The advent of molecular markers made biologists to exploit the unseen potential in breeding endeavors. The markers could be used to accelerate agricultural productivity through better techniques (Paterson *et al.*, 1991). With GBS and NGS platforms dominating the sequence availability, a comprehensive understanding of markers could complement breeding programmes (Nadeem *et al.*, 2018). Molecular markers are considered to be efficient in detecting heritable variations or polymorphisms and exploits them. They could deploy favourable gene combinations to achieve disease control in plants (Kumar *et al.*, 1999). With greater amplification and cost-effective nature, *in silico* molecular markers are being widely exploited.

The utility and approach of molecular marker varies with the context of the crop. SNP and SSR markers have gained importance in plant breeding programmes over the years. SNP markers serve to be efficient in characterizing an organism whereas, SSR seems to be more suitable in diversity analysis and fingerprinting (Varshney *et al.*, 2007). However, combination of SNP and SSR markers were efficiently demonstrated in cowpea, capable of identifying resistant locus within the genome (Kusi *et al.*, 2018). In taro RAPD and SSR were widely used either for evaluating genetic diversity (Irwin *et al.*, 1998) or germplasm management (Mace *et al.*, 2002). Little thrust has been given to SNP and SSR, as a marker against leaf blight or any disease. This could be the first report on developing markers on blight disease resistance using the information available in public databases. Lack of adequate EST and other genetic information on

databases limits the developmental procedures. However, with the transcriptomic data available it opens up new fronts in marker development.

For several organisms computational strategies for marker prediction revolved around EST information available in databases (Nagaraj *et al.*, 2006). However, many crops faced the barrier for marker prediction with fewer ESTs available. In taro, with 22 ESTs reported so far were not enough to develop SSRs or SNPs. Transcriptome information (Wang *et al.*, 2017) on taro was used here to develop the molecular marker- SNP and SSR, which served to be a reliable option even with a complex methodology and processing.

The molecular marker discovery not only helps in achieving better yields but also in identifying gene functions and genetic diversity, the relation between the polymorphism detected and molecular breeding (Semagn *et al.*, 2005).

Taro leaf blight continues to remain a major threat for the farming community with chemical controls quiet unsuccessful. Being a staple food crop in many countries, the decreasing production seems to worsen the condition. However, not as a prominent contributor and competitor in the international market, TLB hasn't achieved significant attention yet. With marker-assisted selection and breeding being an efficient tool for enhanced disease resistance, it could pave the way to substitute fungicides and other harmful chemicals.

Marker-assisted selection always seems to be superior to conventional breeding techniques where there is increased risk or presence of harmful organisms. Marker-assisted selection enables a breeder to eliminate susceptible varieties and concentrate on resistant varieties. MAS could be more beneficial in the case of TLB, enabling breeder to concentrate on fewer lines of varieties.

In this work, about 562 SNPs and 3034 SSRs were predicted form a generalized taro transcriptome data. Of the detected SNPs, 518 were nonsynonymous which resulted in a change in the translational product with a change in the nucleotide. Among the SSRs identified using MISA, 49% corresponded to dinucleotide repeats.

The *in-silico* predicted markers were validated against TLB resistant and susceptible varieties to determine their efficacy.

5.1 COMPARATIVE EVALUATION OF SNP PREDICTION TOOLS

QualitySNP and AutoSNP were used to predict SNPs from the assembled contigs. On comparative evaluation, QualitySNP produced more reliable results with a fewer number of SNPs and classified them to Synonymous and Non-synonymous. AutoSNP, on the other hand, produced more SNPs which were not reliable. The major highlight of SNPs detected by QualitySNP was that they were classified based on the translational product produced with the change in nucleotide sequence.

5.2 COMPARATIVE EVALUATION OF SSR PREDICTION TOOLS

MISA and SSRIT were used to predict SSR among the assembled contigs. On comparative evaluation, MISA showed the higher number of SSR and polymorphism among the detected SSR, whereas in SSRIT the repeats were confined within di-, tri-, and tetra repeats. Increase diversity among the type of repeats and the higher number make MISA more preferable.

5.3 VALIDATION OF THE PREDICTED SNP AND SSR

In-silico developed markers were screened on resistant and susceptible varieties to validate them. The validation confirms the credibility of the developed markers. However, the primers designed for the predicted SNP and SSR maybe hypothetical, as all designed primers may not work well to distinguish between resistant and susceptible varieties. It could be influenced and inhibited by many external factors.

With prediction tools, we could develop markers for plants targetted with a specific function. The transcriptomic data served to be an excellent choice for marker prediction with fewer EST available in the database. The markers designed could be of great use in breeding programmes once it is validated in

larger sample sizes. It could help breeders to opt out the resistant varieties as the designed markers were once screened with a leaf blight resistant database.

SUMMARY

6. SUMMARY

The study entitled "Development of Molecular markers for blight disease resistance in taro using bioinformatics tools" was conducted at the Central Tuber Crop Research Institute (CTCRI) during 2017 - 2018. The main objectives of the study were to develop and evaluate marker prediction pipelines of SNP and SSR, computational prediction, and validation of the markers. The study was divided into two phases, *in silico* prediction of molecular markers and their validation. The notable observations of the study are stated below.

The raw data for identifying SSR and SNP marker were obtained from the SRA section of NCBI (https://www.ncbi.nlm.nih.gov/sra). The NGS data served to be raw dataset in absence of adequate number of EST. The transcript corresponded to about 6,479,882 paired reads which were trimmed to 6,319,834 reads. The reads were then assembled *de novo* by Trinity and aligned using CAP3 to produce 8547 contigs which served to be the input for marker prediction.

QualitySNP and AutoSNP were the SNP prediction tools used for detecting SNP, whereas SSRIT and MISA were employed to predict SSRs for the dataset obtained.

QualitySNP with better algorithm proved to be more useful and reliable, as it clearly distinguished between synonymous and nonsynonymous SNPs. Nonsynonymous SNPs produced a precise change in the translational product with the change in single nucleotide sequence. With the huge number of SNPs detected by AutoSNP, it is quite untrustworthy. MISA, on the other hand, serves to be more reliable even with the increased number comparing to lower repeats identified by SSRIT. With a better algorithm, it predicted more types of repeats and compound SSRs. With the SSR/SNP containing contigs crosschecked via BLAST against a leaf blight resistant database enhanced the decisiveness of the markers.

QualitySNP identified about 562 SNPs of which 518 were nonsynonymous and 44 were synonymous which corresponded to 238 contigs. In MISA 967 mono, 1484 di, 558 tri, 14 tetra, 2 penta and 9 hexa repeats were detected which together add to a total of 3034 SSRs. Five sequences from each with lower e-value and good percentage identity on BLAST with resistant database were chosen for primer designing to validate the *in silico* data. The primers were validated against 3 susceptible and 3 tolerant varieties. Among the primers designed, CeSSR4 in the case of SSR and CeSNP2 and CeSNP3 in SNP were capable of distinguishing resistant and susceptible varieties.

Scope for future work

With only 5 SSR and SNP being validated, the remaining markers could be validated in future. With CeSSR4 and CeSNP3 being able to differentiate susceptible and resistant lines among the five selected, validation of remaining could add up the resources. The designed markers could also prove to be beneficial in marker-assisted selection and other breeding programmes for taro.

REFERENCES

7. REFERENCES

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): 359p.

Ahrberg, C.D., Manz, A., and Chung, B.G. 2016. Polymerase chain reaction in microfluidic devices. *Lab on a Chip* 16(20): 3866-3884.

Alexandratos, N. and Bruinsma, J. 2012. *World agriculture towards 2030/2050: the 2012 revision.* ESA Working paper No. 12-03. Rome, FAO.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nat.* 408(6814): 796-815.

Banjaw, D.T. 2017. Review of Taro (*Colocasia esculenta*) Genetics and Breeding *J. Hortic.* 4(1): 196. Available: http://DOI:10.4172/2376-0354.1000196 [30 January 2018].

Bao, Y., Xu, S., Jing, X., Meng, L., and Qin, Z. 2015. De novo assembly and characterization of Oryza officinalis leaf transcriptome by using RNA-Seq. *BioMed. Res. Int.* 2015: 7. Available: http://dx.doi.org/10.1155/2015/982065 [15 March 2018].

Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J., and Edwards, D. 2003. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinforma.* 19(3): 421-422.

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

Beletsky, A.V., Filyushin, M.A., Gruzdev, E.V., Mazur, A.M., Prokhortchouk, E.B., Kochieva, E.Z., Mardanov, A.V., Ravin, N.V., and Skryabin, K.G. 2017. *De novo* transcriptome assembly of the mycoheterotrophic plant *Monotropa hypopitys*. *Genomics data*, 11: 60-61. Available: https://doi.org/10.1016/j.gdata.2016.11.020 [22 December 2017].

Bevan, M.W. and Uauy, C. 2013. Genomics reveals new landscapes for crop improvement. *Genome Biol.* 14(6): 206.

Bilsborough, G.D., 2013. Plant genomics: sowing the seeds of success. *Genome Biol.* 14: 404. Available: https://doi.org/10.1186/gb-2013-14-6-404 [10 December 2017].

Boches, P.S., Bassil, N.V., and Rowland, L.J. 2005. Microsatellite markers for Vaccinium from EST and genomic libraries. *Mol. Ecol. Notes*, 5(3), pp.657-660.

Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinforma.* 30(15): 2114-2120.

Borges, J.C., Cagliari, T.C., and Ramos, C.H. 2007. Expression and variability of molecular chaperones in the sugarcane expressome. *J. plant physiol.* 164(4): 505-513.

Buermans, H.P.J. and Den Dunnen, J.T. 2014. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Mol. Basis Dis.* 1842(10): 1932-1941.

Ceresini, P.C., Silva, C.L.S.P., Missio, R.F., Souza, E.C., Fischer, C.N., Guillherme, I.R., Gregorio, I., Silva, E.H.T.D., Cicarelli, R.M.B., Silva, M.T.A.D., and Garcia, J.F. 2005. Satellyptus: Analysis and database of microsatellites from ESTs of Eucalyptus. *Genet. and Mol. Biol.* 28(3): 589-600.

Chair, H., Traore, R.E., Duval, M.F., Rivallan, R., Mukherjee, A., Aboagye, L.M., Van Rensburg, W.J., Andrianavalona, V., de Carvalho, M.P., Saborio, F., Prana, M.S., and Komolong, B. 2016. Genetic diversification and dispersal of taro (*Colocasia esculenta* (L.) Schott). *PloS one.* 11(6), p.e0157712. Available: https://doi.org/10.1371/journal.pone.0157712 [10 January, 2018].

Ching, A.D.A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O., Tingey, S., Morgante, M., and Rafalski, A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC genet.* 3(1): 19.

Choe, J., Kim, J.E., Lee, B.W., Lee, J.H., Nam, M., Park, Y.I., and Jo, S.H. 2018. A comparative synteny analysis tool for target-gene SNP marker discovery: connecting genomics data to breeding in Solanaceae. *Database*, 2018: 47.

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. Plant Genomics*, 2008: 194-200. Available: https://doi.org/10.1155/2008/412696. [18 March, 2018].

Devillers H., Morin N., and Neuveglise C. 2016. *Methods in Molecular Biology* (volume 1361). Humana Press, New York, 41-56.

Doi, H., Takahara, T., Minamoto, T., Matsuhashi, S., Uchii, K., and Yamanaka, H. 2015. Droplet digital polymerase chain reaction (PCR) outperforms real-time PCR in the detection of environmental DNA from an invasive fish species. *Environ. Sci. Technol.* 49(9): 5601-5608.

Doveri S., Lee D., Maheswaran M. and Powell W. (2008). *Principles and Practices of Plant Genomics* (Volume 1) Science Publishers, USA, pp.23-68.

Doyle J.J. and J.L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull* 19: 11–15.

Edison, S., Sreekumari, M.T., Pillai, S.V., and Sheela, M. N. 2003. Diversity and genetic resources of taro in India. In: *3rd Taro Symposium*; 21-23, May, 2003, Fiji islands. Secretariat of the Pacific Community and the International Plant Genetic Resources Institute, Rome, 85-88. Paper 1.7.

Eujayl, I., Sledge, M.K., Wang, L., May, G.D., Chekhovskiy, K., Zwonitzer, J.C., and Mian, M.A.R. 2004. Medicago truncatula EST-SSRs reveal cross-species genetic markers for Medicago species. *Theor. and Appl. Genet.* 108(3): 414-422.

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

FAOSTAT [Food and Agriculture Organization Corporate Statistical Database] 2016 Available: http://www.fao.org/faostat/en/#data/qc [2 July, 2018].

Gardner, S.N., Slezak, T., and Hall, B.G. 2015. kSNP3. 0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinforma.* 31(17): 2877-2878.

Gimode, D., Odeny, D.A., de Villiers, E.P., Wanyonyi, S., Dida, M.M., Mneney, E.E., Muchugi, A., Machuka, J., and de Villiers, S.M. 2016. Identification of SNP and SSR markers in finger millet using next generation sequencing technologies. *PLoS one*, 11(7): p.e0159437. Available: https://doi.org/10.1371/journal.pone.0159437. [20 February, 2018].

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., and Toulmin, C. 2010. Food security: the challenge of feeding 9 billion people. *Sci.* Available: DOI: 10.1126/science.1185383. [15 November, 2017].

Goltenboth, F., Timotius, K. H., Milan, P.P., and Margraf, J. 2006. *Ecology of Insular Southeast Asia.* Elsevier, Amsterdam, 568p.

Gonzaga, Z.J., Aslam, K., Septiningsih, E.M., and Collard, B.C. 2015. Evaluation of SSR and SNP markers for molecular breeding in rice. *Plant Breed. Biotechnol.* 3(2): 139-152.

Gordon, D. and Green, P. 2013. Consed: a graphical editor for next-generation sequencing. *Bioinforma.* 29(22): 2936-2937.

Gordon, D., Abajian, C. and Green, P., 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8(3): 195-202.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., and Chen, Z. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29(7): 644.

Gregory, P.H. 1983. Some major epidemics caused by Phytophthora. In *Phytophthora: its biology, taxonomy, ecology and pathology.* pp.271-278.

Grimaldi, I.M., Muthukumaran, S., Tozzi, G., Nastasi, A., Boivin, N., Matthews, P.J., and van Andel, T. 2018. Literary evidence for taro in the ancient Mediterranean: A chronology of names and uses in a multilingual world. *PloS one* 13(6): p.e0198333. Available: https://doi.org/10.1371/journal.pone.0198333 [20 March, 2018].

Haas, B.J., Grabherr, M.G., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., and Chen, Z. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29(7): 644.

Han, J.H., Chon, J.K., Ahn, J.H., Choi, I.Y., Lee, Y.H., and Kim, K.S. 2016. Whole genome sequence and genome annotation of *Colletotrichum acutatum*, causal agent of anthracnose in pepper plants in South Korea. *Genomics data* 8: 45-46.

Hanai, L.R., De Campos, T., Camargo, L.E.A., Benchimol, L.L., De Souza, A.P., Melotto, M., Carbonell, S.A.M., Chioratto, A.F., Consoli, L., Formighieri, E.F., and Siqueira, M.V.B.M. 2007. Development, characterization, and comparative analysis of polymorphism at common bean SSR loci isolated from genic and genomic sources. *Genome* 50(3): 266-277.

Hayward, A.C., Tollenaere, R., Dalton-Morgan, J., and Batley, J. 2015. *Plant Genotyping* Humana Press, New York, 1245:13-27.

He, B., Zhao, S., Chen, Y., Cao, Q., Wei, C., Cheng, X., and Zhang, Y. 2015. Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. *BMC genomics* 16(1): 65.

Helmkampf, M., Wolfgruber, T.K., Bellinger, M.R., Paudel, R., Kantar, M.B., Miyasaka, S.C., Kimball, H.L., Brown, A., Veillet, A., Read, A., and Shintaku, M. 2017. Phylogenetic relationships, breeding implications, and cultivation history of Hawaiian taro (*Colocasia esculenta*) through genome-wide SNP genotyping. *J. Heredity*, 109(3): 272-282.

Hu, K.A.N., Huang, X.F., Ke, W.D., and Ding, Y.I. 2009. Characterization of 11 new microsatellite loci in taro (*Colocasia esculenta*). *Mol. Ecol. Resour.* 9(2): 582-584.

Huang, C.C., Chen, W.C., and Wang, C.C. 2007. Comparison of Taiwan paddy- and upland-cultivated taro [*Colocasia esculenta* (L.)] cultivars for nutritive values. *Food Chem.* 102(1): 250-256.

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9(9): 868-877.

Hung, J.H. and Weng, Z. 2016. Designing polymerase chain reaction primers using Primer3Plus. *Cold Spring Harbor Protocols* [e-journal] 2016(9). Available: http://cshprotocols.cshlp.org/content/2016/9/pdb.prot093096.short [1 April 2018].

Hwang, S.G., Kim, K.H., Lee, B.M., and Moon, J.C. 2018. Transcriptome analysis for identifying possible gene regulations during maize root emergence and formation at the initial growth stage. *Genes genomics*, 40(7): 755-766.

Iorizzo, M., Senalik, D.A., Grzebelus, D., Bowman, M., Cavagnaro, P.F., Matvienko, M., Ashrafi, H., Van Deynze, A., and Simon, P.W. 2011. *De novo* assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC genomics* 12(1): 389.

Irwin, S.V., Kaufusi, P., Banks, K., De la Pena, R., and Cho, J.J. 1998. Molecular characterization of taro (*Colocasia esculenta*) using RAPD markers. *Euphytica*, 99(3): 183.

Ivancic, A. and Lebot, V. 2000. *The genetics and breeding of taro*. French Agricultural Research Centre for International Development (CIRAD), Paris, 194p.

Jeppson, J.O., Laurell, C.B., and Franzén, B. 1979. Agarose gel electrophoresis. *Clin. Chem.* 25(4): 629-638.

Jiang, L., Mancuso, M., Lu, Z., Akar, G., Cesarman, E., and Erickson, D. 2014. Solar thermal polymerase chain reaction for smartphone-assisted molecular diagnostics. *Sci. Rep.* 4: 4137.

Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'donnell, C.J., and De Bakker, P.I. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinforma*. 24(24): 2938-2939.

Jorde, L.B. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* 10(10): 1435-1444.

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(5-6): 501-510.

Kaushal, P., Kumar, V., and Sharma, H.K. 2015. Utilization of taro (*Colocasia esculenta*): a review. *J. Food Sci. Technol.* 52(1): 27-40.

Kofler, R., Schlötterer, C., and Lelley, T. 2007. SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinforma.*, 23(13): 1683-1685.

Kreike, C.M., Van Eck, H.J., and Lebot, V. 2004. Genetic diversity of taro, *Colocasia esculenta* (L.) Schott, in Southeast Asia and the Pacific. *Theor. and Appl. genetics* 109(4): 761-768.

Kumar, L.S. 1999. DNA markers in plant improvement: an overview. *Biotechnol. Adv.* 17(2-3): 143-182.

Kuruvilla, K.M., and Singh, A. 1981. Karyotypic and electrophoretic studies on taro and its origin. *Euphytica*, 30(2): 405-413.

Kushwaha, S.K., Vetukuri, R.R., and Grenville-Briggs, L.J. 2017. Draft genome sequence of the mycoparasitic oomycete Pythium periplocum strain CBS 532.74. *Genome announcements*, 5(12): e00057-17. Available: https://doi.org/10.1128/genomeA.00057-17 [26 January 2018].

Kusi, F., Padi, F.K., Obeng-Ofori, D., Asante, S.K., Agyare, R.Y., Sugri, I., Timko, M.P., Koebner, R., Huynh, B.L., Santos, J.R., and Close, T.J. 2018. A novel aphid resistance locus in cowpea identified by combining SSR and SNP markers. *Plant Breed.* 137(2): 203-209.

Lapatas, V., Stefanidakis, M., Jimenez, R.C., Via, A., and Schneider, M.V. 2015. Data integration in biological research: an overview. *J. Biol. Res.* 22(1): 9.

Lebot, V. and Aradhya, K.M. 1991. Isozyme variation in taro (*Colocasia esculenta* (L.) Schott) from Asia and Oceania. *Euphytica*, 56(1): 55-66.

Lebot, V., Herail, C., Gunua, T., Pardales, J., Prana, M., Thongjiem, M., and Viet, N. 2003. Isozyme and RAPD variation among *Phytophthora colocasiae* isolates from South-east Asia and the Pacific. *Plant Pathol.* 52(3): 303-313.

Lebot, V., Prana, M.S., Kreike, N., Van Heck, H., Pardales, J., Okpul, T., Gendua, T., Thongjiem, M., Hue, H., Viet, N., and Yap, T.C. 2004. Characterisation of taro (*Colocasia esculenta* (L.) Schott) genetic resources in Southeast Asia and Oceania. *Genetic Resour. Crop Evol.*, 51(4): 381-392.

Lewu, M.N., Adebola, P.O., and Afolayan, A.J. 2010. Comparative assessment of the nutritional value of commercially available cocoyam and potato tubers in South Africa. *J. food Qual.* 33(4): 461-476.

Li, K. B., 2003. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinforma.* 19(12): 1585-1586.

Li, L., Wang, J., Guo, Y., Jiang, F., Xu, Y., Wang, Y., Pan, H., Han, G., Li, R., and Li, S. 2008. Development of SSR markers from ESTs of gramineous species and their chromosome location on wheat. *Prog. Nat. Sci.* 18(12): 1485-1490.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A., and Nevo, E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11(12): 2453-2465.

Liang, C., Liu, X., Yiu, S.M., and Lim, B.L. 2013. *De novo* assembly and characterization of *Camelina sativa* transcriptome by paired-end sequencing. *BMC genomics* 14(1): 146.

Liu, Q., Donner, E., Yin, Y., Huang, R.L., and Fan, M.Z. 2006. The physicochemical properties and in vitro digestibility of selected cereals, tubers and legumes grown in China. *Food Chem.* 99(3): 470-477.

Liu, S., Li, W., Wu, Y., Chen, C., and Lei, J. 2013. *De novo* transcriptome assembly in chili pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of capsaicinoids. *PloS one*, 8(1): p.e48156. Available: https://doi.org/10.1371/journal.pone.0048156 [14 November 2017].

Lu, Z., Li, W., Yang, Y., and Hu, X. 2011. Isolation and characterization of 19 new microsatellite loci in *Colocasia esculenta* (Araceae). *Am. J. Bot.* 98(9): 239-241.

Ma, Y., Yang, T., Guan, J., Wang, S., Wang, H., Sun, X., and Zong, X. 2011. Development and characterization of 21 EST-derived microsatellite markers in *Vicia faba* (fava bean). *Am. J. Bot.* 98(2): 22-24.

Mace, E.S. and Godwin, I.D. 2002. Development and characterization of polymorphic microsatellite markers in taro (*Colocasia esculenta*). *Genome* 45(5): 823-832.

Mammadov, J., Aggarwal, R., Buyyarapu, R., and Kumpatla, S. 2012. SNP markers and their impact on plant breeding. *Int. J. Plant Genomics*, 2012: 1-11 Available: http://dx.doi.org/10.1155/2012/728398 [20 February 2018].

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

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. *Bioinformation*, 3(6): 282.

Matsuda, M. and Nawata, E. 2002. Geographical distribution of ribosomal DNA variation in taro, *Colocasia esculenta* (L.) Schott, in eastern Asia. *Euphytica*, 128(2): 165.

Matthews, P., Matsushita, Y., Sato, T., and Hirai, M., 1992. Ribosomal and mitochondrial DNA variation in Japanese taro (*Colocasia esculenta* L. Schott). *Jpn. J. Breed.* 42(4): 825-833.

Metz, S., Cabrera, J.M., Rueda, E., Giri, F., and Amavet, P. 2016. FullSSR: microsatellite finder and primer designer. *Adv. Bioinforma.*, 2016: 1-4. Available: https://doi.org/10.1155/2016/6040124 [10 March 2018].

Milgroom, M.G. and Fry, W.E. 1997. Contributions of population genetics to plant disease epidemiology and management. In: *Advances in Botanical Research* (Vol. 24) Academic Press, Cambridge. pp 1-30.

Misra, R.S., Sharma, K., and Mishra, A.K. 2008. Phytophthora leaf blight of Taro (*Colocasia esculenta*)–a review. *Asian Aust. J. Plant Sci. Biotechnol.* 2: 55-63.

Mohan, M., Nair, S., Bhagwat, A., Krishna, T.G., Yano, M., Bhatia, C.R., and Sasaki, T. 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol. Breed.* 3(2): 87-103.

Monger, C., Motheramgari, K., McSharry, J., Barron, N., and Clarke C. 2017. A Bioinformatics Pipeline for the Identification of CHO Cell Differential Gene Expression from RNA-Seq Data. *Methods Mol. Biol.* 1603: 169-186.

Moss, P. and Thein, S.L. 1998. *Clinical Applications of PCR*. Humana Press, New York City, pp.145-148.

Mullis, K., Faloona, F., Scharf, S., Saiki, R.K., Horn, G.T., and Erlich, H., 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In: *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press, New York. 51: 263-273

Mutz, K.O., Heilkenbrinker, A., Lonne, M., Walter, J.G., and Stahl, F. 2013. Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.* 24: 22–30.

Mwenye, O.J., Labuschagne, M.T., Herselman, L., and Benesi, I.R.M. 2011. Mineral composition of Malawian cocoyam (*Colocasia esculenta* and *Xanthosoma sagittifolium*) genotypes. *J. Biol. Sci.* 11(4): 331-335.

Nadeem, M.A., Nawaz, M.A., Shahid, M.Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., and Özkan, H. 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnological Equip.* 32(2): 261-285.

Nagaraj, S.H., Gasser, R.B., and Ranganathan, S. 2006. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings Bioinforma.* 8(1): 6-21.

Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic acids Res.* 25(14): 2745-2751.

Njintang, Y.N., Scher, J., and Mbofung, C.M.F. 2008. Physicochemical, thermal properties and microstructure of six varieties of taro (*Colocasia esculenta* L. Schott) flours and starches. *J. Food Eng.* 86(2): 294-305.

Noyer, J.L., Billot, C., Weber, A., Brottier, P., Quero-Garcia, J., and Lebot, V. 2003, May. Genetic diversity of taro (*Colocasia esculenta* (L.) Schott) assessed by SSR markers. In: *Proceedings of the Third Taro Symposium,* 21 – 23 May 2003, Nadi, Fiji Islands. pp.21-23.

O'Sullivan, J.N., Asher, C.J., Blamey, F.P.C., Asher, C.J., and Blamey, F.P.C. 1996. *Nutritional disorders of taro.* Department of Agriculture, The University of Queensland. 80-81.

Onwueme, I.C., 1978. *The tropical tuber crops: yams, cassava, sweet potato, and cocoyams*. John Wiley and Sons, Chichester, 3-97.

Paterson, A.H., Tanksley, S.D., and Sorrells, M.E. 1991. DNA markers in plant improvement. In: *Adv. Agron.* Academic Press, United States, 39-90

Patil, G., Do, T., Vuong, T.D., Valliyodan, B., Lee, J.D., Chaudhary, J., Shannon, J.G., and Nguyen, H.T. 2016. Genomic-assisted haplotype analysis and the development of high-throughput SNP markers for salinity tolerance in soybean. *Sci. Rep.* [e-journal] 6(19199). Available: https://www.nature.com/articles/srep19199 [21 March 2018].

Plucknett, D.L. 1983. Taxonomy of the genus *Colocasia*. In: Wang J.K. (Ed.), A Review of *Colocasia esculenta* and its Potentials, Univ Hawaii Press, Honolulu. pp. 14–19.

Pongener, N. and Daiho, L. (2016). Survey on leaf blight of taro (*Phytophthora colocasiae* Raciborski) in Nagaland. *Acta Hortic.* 1118: 125-130 DOI:10.17660/ActaHortic.2016.1118.18 Available: https://doi.org/10.17660/ActaHortic.2016.1118.18

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

Purseglove, J.W. 1972. *Tropical crops: monocotyledons* Halsted Press, Division of John Wiley and Sons, London. 334p.

Qu, J. and Liu, J. 2013. A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. *BMC Res. Notes* 6(1): 403.

Rafalski, A., 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opinion Plant Biol.* 5(2): 94-100.

Ranzani, V., Arrigoni, A., Rossetti, G., Panzeri, I., Abrignani, S., Bonnal, R.J., and Pagani, M. 2017. Next-generation sequencing analysis of long noncoding RNAs in CD4+ T cell differentiation. In: *T-Cell Differentiation* Humana Press, New York. pp.173-185.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K., Hoorn, G.T., and Arnheim, N. 1985. Polymerase chain reaction. *Sci.* 230: 1350-1354.

Savage, D., Batley, J., Erwin, T., Logan, E., Love, C.G., Lim, G.A., Mongin, E., Barker, G., Spangenberg, G.C., and Edwards, D. 2005. SNPServer: a real-time SNP discovery tool. *Nucleic Acids Res.* 33(2). Available: https://doi.org/10.1093/nar/gki462 [January 2018].

Schlotterer C. (2004) The evolution of molecular markers - just a matter of fashion? *Nat. Rev. Genet.* 5(1): 63-69.

Schochetman, G., Ou, C.Y. and Jones, W.K., 1988. Polymerase chain reaction. *J. infectious Dis.* 158(6): 1154-1157.

Scholten, O.E., van der Linden, C.G., and Vosman, B. 2005. Molecular markers for disease resistance in various crops obtained by NBS profiling. In: Scholten O.E. (ed.), *COST SUSVAR/ECO-PB Workshop on organic plant breeding strategies and the use of molecular markers*. Proceedings of an international workshop, Driebergen, The Netherlands. Louis Bolk Institute, Driebergen, The Netherlands, 95p.

Semagn, K., Bjørnstad, Å., and Ndjiondjop, M.N. 2006. An overview of molecular marker methods for plants. *Afr. J. Biotechnol.* 5(25): 2540-2568

Sepúlveda-Nieto, M.D.P., Bonifacio-Anacleto, F., Faleiros de Figueiredo, C., de Moraes-Filho, R.M., and Alzate-Marin, A.L. 2017. Accessible Morphological and Genetic Markers for Identification of Taioba and Taro, T wo Forgotten Human Foods. *Horticulturae*, 3(4): 49.

Sharma, K., Mishra, A.K., and Misra, R.S. 2008. A simple and efficient method for extraction of genomic DNA from tropical tuber crops. *Afr. J. Biotechnol.*, 7(8): 1018-1022.

Sharma, K., Mishra, A.K., and Misra, R.S. 2008. Analysis of AFLP variation of taro population and markers associated with leaf blight resistance gene. *Academic J. Plant Sci.* 1(3): 42-48.

Sharma, K., Mishra, A.K., and Misra, R.S. 2009. Identification and characterization of differentially expressed genes in the resistance reaction in taro infected with *Phytophthora colocasiae*. *Mol. Biol. Rep.* 36(6): 1291-1297.

Sharma, P.C., Grover, A., and Kahl, G. 2007. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.* 25(11): 490-498.

Singh, D., Jackson, G., Hunter, D., Fullerton, R., Lebot, V., Taylor, M., Iosefa, T., Okpul, T., and Tyson, J. 2012. Taro leaf blight—a threat to food security. *Agric.* 2(3): 182-203.

Singh, D., Mace, E.S., Godwin, I.D., Mathur, P.N., Okpul, T., Taylor, M., Hunter, D., Kambuou, R., Rao, V.R., and Jackson, G. 2008. Assessment and rationalization of genetic diversity of Papua New Guinea taro (*Colocasia esculenta*) using SSR DNA fingerprinting. *Genetic Resour. Crop Evol.* 55(6): 811-822.

Sinha, D.K. and Smith, C.M., 2014. Selection of Reference Genes for Expression Analysis in *Diuraphis noxia* (Hemiptera: Aphididae) Fed on Resistant and Susceptible Wheat Plants. *Sci. Rep.* 4: 5059.

Skarzynska, A., Kusmirek, W., Pawełkowicz, M., Pląder, W., and Nowak, R.M. 2017. Assembly of cucumber (*Cucumis sativus* L.) somaclones. In: *Photonics Appl. Astron. Commun. Ind. High Energy Phys. Exp.* 2017. International Society for Optics and Photonics. 10445: 1044534.

Smith D.R. (1996) Agarose Gel Electrophoresis. In: Harwood A.J. (ed.) Basic DNA and RNA Protocols. Methods in Molecular Biology, vol 58. Humana Press, , New York, pp.17-21.

Soulard, L., Mournet, P., and Guitton, B. 2017. Construction of two genetic linkage maps of taro using single nucleotide polymorphism and microsatellite markers. *Mol. Breed.* 37(3): 37.

Sreekumari, M.T. and Mathew, P.M. 1991. Karyotypically Distinct Morphotypes in Taro (*Colocasia esculenta* (L.) Schott). *Cytologia*, 56(3): 399-402.

Stein, B.D., Strauss, M.S., and Scheirer, D.C. 1983. Anatomy and Histochemistry of Taro, *Colocasia esculenta* (L.) Schott, Leaves[abstract]. In Abstracts, Sixth Symposium of International Society for Tropical Root Crops; 21-26 February, 1983, Peru. International Potato Centre Lima, Peru, p.12. Abstract No. 5.1.5.

Tahara, M., 1999. Phylogenetic Relationships of taro, *Colocasia esculenta* (L.) Schott and related taxa by non-coding chloroplast DNA seguence analysis. *Aroideana*, 22: 79-89.

Taheri, S., Lee Abdullah, T., Yusop, M.R., Hanafi, M.M., Sahebi, M., Azizi, P., and Shamshiri, R.R. 2018. Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. *Molecules*, 23(2): 399.

Taji, T., Sakurai, T., Mochida, K., Ishiwata, A., Kurotani, A., Totoki, Y., Toyoda, A., Sakaki, Y., Seki, M., Ono, H., and Sakata, Y. 2008. Large-scale collection and annotation of full-length enriched cDNAs from a model halophyte, *Thellungiella halophila*. *BMC Plant Biol.* 8(1): 115.

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. *Genome Res.* 11(8): 1441-1452.

Thiel, T. 2003. MISA—Microsatellite identification tool [on-line]. Available: http://pgrc. ipk-gatersleben. de/misa/ [17 December 2017].

Torkamaneh, D., Boyle, B., and Belzile, F. 2018. Efficient genome-wide genotyping strategies and data integration in crop plants. *Theor. Appl. Genet.* 131: 499-511.

Trujillo, E.E. 1965. Effects of humidity and temperature on Phytophthora blight of taro. *Phytopathol.* 55(2): 183.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. 2012. Primer3—new capabilities and interfaces. *Nucleic acids Res.* 40(15): 115.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., and Leunissen, J.A. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids Res.* 35(2): 71-74.

USDA (United States Department of Agriculture) 2018 Available: https://ndb.nal.usda.gov/ndb/search/list/?manu=&fgcd=&ds=&qlookup=Taro,%20raw [July 15, 2018].

Vandenbroucke, H., Mournet, P., Vignes, H., Malapa, R., Duval, M.F., and Lebot, V. 2016. Somaclonal variants of taro (*Colocasia esculenta* L. Schott) and

yam (*Dioscorea alata* L.) are incorporated into farmers varietal portfolios in Vanuatu. *Genet. Resour. crop Evol.* 63(3): 495-511.

Varshney, R.K., Chabane, K., Hendre, P.S., Aggarwal, R.K., and Graner, A. 2007. Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Sci.* 173(6): 638-649

Vieira, M.L.C., Santini, L., Diniz, A.L., and Munhoz, C.D.F. 2016. Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39(3): 312-328.

Villarino, G.H., Bombarely, A., Giovannoni, J.J., Scanlon, M.J., and Mattson, N.S. 2014. Transcriptomic analysis of *Petunia hybrida* in response to salt stress using high throughput RNA sequencing. *PLoS One*, [e-journal] 9(4). Available: https://doi.org/10.1371/journal.pone.0094651. [21 March 2018].

Voytas, D. 2000. Agarose gel electrophoresis. *Curr. protocols Neuroscience*, [e-journal] 11(1): Available: https://doi.org/10.1002/0471142301.nsa01ns11.

Wagih, M.E. 1994. Seed rescue culture: a technique for the regeneration of Taro (*Colocasia esculenta*). *Afr. Crop Sci. J.* 2(1): 9 - 15.

Wahid, H.A., Barozai, M.Y.K., and Din, M. 2016. Functional characterization of fifteen hundred transcripts from *Ziarat juniper* (*Juniperus excelsa* M. Bieb). *Advmt. Life Sci.* 4(1): 20-26.

Wang, C., Guo, W., Cai, C., and Zhang, T. 2006. Characterization, development and exploitation of EST-derived microsatellites in *Gossypium raimondii* Ulbrich. *Chinese Sci. Bull.* 51(5): 557-561.

Wang, H., Jiang, J., Chen, S., Qi, X., Peng, H., Li, P., Song, A., Guan, Z., Fang, W., Liao, Y., and Chen, F. 2013. Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome permits large-scale

unigene assembly and SSR marker discovery. *PloS one*, [e-journal] 8(4): Available: https://doi.org/10.1371/journal.pone.0062293.

Wang, L., Yin, J., Zhang, P., Han, X., Guo, W. and Li, C., 2017. De novo assembly and characterization of transcriptome and microsatellite marker development for Taro (*Colocasia esculenta* (L.) Schott.). *Int. J. Genet. Mol. Biol.* 9(5): 26-36.

Wang, X., Lu, P., and Luo, Z. 2013. GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformation*, 9(10): 541.

Ward, J.A., Ponnala, L., and Weber, C.A. 2012. Strategies for transcriptome analysis in nonmodel plants. *Am. J. Bot.* 99(2): 267-276.

Weightman, B. 1989. Agriculture in Vanuatu, A Historical Review. Grosvenor Press, Portsmouth. 54-78.

Yen, D.E., and Wheeler, J.M. 1968. Introduction of taro into the Pacific: the indications of the chromosome numbers. *Ethnol.* 7(3): 259-267.

Yuanzhen, L., Hai, G., and Chunxin, L. 2010. Screening and analysis of EST-SSR in *Jatropha curcas*_. BMC Res. Notes [e-journal] 3(42). Available: https://doi.org/10.1186/1756-0500-3-42 [23 March 2018].

Zeng, S., Xiao, G., Guo, J., Fei, Z., Xu, Y., Roe, B.A., and Wang, Y. 2010. Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics*, 11(1): 94.

Zhang, C.Y., Liu, T.J., Xu, Y. and Yan, H.F., 2017. Characterization of the whole chloroplast genome of a rare candelabra primrose *Primula chrysochlora* (Primulaceae). *Conserv. Genet. Res.* 9(3): 361-363.

APPENDICES

APPENDIX I

Preparation of DNA extraction buffer (Sharma *et al.* 2008)

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

APPENDIX II

Preparation of TE buffer (10X)

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

Final volume made upto 100ml with distilled water.

APPENDIX III

TBE buffer (10X)

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

APPENDIX IV

100bp marker

1. 100bp marker : 5µl
2. Loading dye : 40µl
3. Sterile distilled water: 55µl

APPENDIX V

PCR Mastermix

PCR Cocktail	Stock concentration	Final concentration	Volume taken (µL)	
DNA	100 ng/ µL	40 ng/ µL	4	15 µL
Forward Primer	10 µM	0.25 µM	0.375	
Reverse Primer	10 µM	0.25 µM	0.375	
dNTPS	2.5 mM	0.25 mM	1.5	
Taq Buffer	10X	1X	1.5	
Taq polymerase	5U/ µL	1U/ µL	0.2	
$MgCl_2$	25mM	1mM	0.6	
Sterile Water			6.45	

APPENDIX VI						
List of Synonymous SNP coding data identified by QualitySNP						
Sl. No.	Contig Name	Position	SNP	Sequence	Sequence with Base change	Transcribed Proteins
1	Contig507	157	TC	ATCGTTTTGTTTTTGAAGGGCGTAC	ATCGTTTTGTTTCTGAAGGGCGTAC	IVLFLKGVX
2	Contig507	210	CT	GCAACACCTGATCTACAGGTTTGGC	GCAACACCTGATTTACAGGTTTGGC	ATPDLQVWX
3	Contig507	465	CT	CGTACGGGCGTACTGGACGCAATCC	CGTACGGGCGTATTGGACGCAATCC	RTGVLDAIX
4	Contig679	256	TC	CAACTTTATCTGTTGCCGGACACCT	CAACTTTATCTGCTGCCGGACACCT	QLYLLPDTX
5	Contig866	760	TC	AAGGACGAGAGCTTAGAAAAGTTTG	AAGGACGAGAGCCTAGAAAAGTTTG	KDESLEKFX
6	Contig1284	123	CT	AATCTGACCGATCTAAGGTCATTTC	AATCTGACCGATTTAAGGTCATTTC	NLTDLRSFX
7	Contig1423	780	AC	CAGCCACTGCTCAGACGACGGATCG	CAGCCACTGCTCCGACGACGGATCG	QPLLRRRIX
8	Contig1471	1973	TC	GGCCAGCCTTCCTTGGAGAGATTTC	GGCCAGCCTTCCCTGGAGAGATTTC	GQPSLERFX
9	Contig1634	203	TC	CCATTCTTGCTCTTGACAAGGGACA	CCATTCTTGCTCCTGACAAGGGACA	PFLLLTRDX
10	Contig1684	285	CT	GGAGTTTATTAGCTGGATCGAGAAC	GGAGTTTATTAGTTGGATCGAGAAC	GVY*LDREX
11	Contig1722	407	CT	AAAGACTGTATACTGTTTCCCTTCT	AAAGACTGTATATTGTTTCCCTTCT	KDCILFPFX
12	Contig1725	2126	CT	TGCAGTAATTATCTGCCGCAGAGGC	TGCAGTAATTATTTGCCGCAGAGGC	CSNYLPQRX
13	Contig2580	492	CT	AACTTTCTAGCTCTAATTGCAGAAA	AACTTTCTAGCTTTAATTGCAGAAA	NFLALIAEX
14	Contig2705	446	CT	TATCTGCAAGAGCTGGTGTACAAGC	TATCTGCAAGAGTTGGTGTACAAGC	YLQELVYKX
15	Contig2762	565	CT	GGCATCTGCCTCCTGATATCCTCCG	GGCATCTGCCTCTTGATATCCTCCG	GICLLISSX
16	Contig2902	392	CT	CGGCACAACCACCTATTTTGATGGA	CGGCACAACCACTTATTTTGATGGA	RHNHLF*WX
17	Contig2902	580	TC	CTAAAGTACAGATTGGACGACATGG	CTAAAGTACAGACTGGACGACATGG	LKYRLDDMX
18	Contig2930	659	CT	GCCTGGGTCCTGCTGCACGCCTGAG	GCCTGGGTCCTGTTGCACGCCTGAG	AWVLLHA*X
19	Contig2949	194	TC	AAACAACTGACCTTGAAAAGATCGA	AAACAACTGACCCTGAAAAGATCGA	KQLTLKRSX
20	Contig2975	345	TC	ACGACATCATCTTTATTGGGTCCCT	ACGACATCATCTCTATTGGGTCCCT	TTSSLLGPX
21	Contig3068	1004	TC	TCCACAGGGCAGTTGAATGGCCTCT	TCCACAGGGCAGCTGAATGGCCTCT	STGQLNGLX
22	Contig3178	518	TC	GAGTATTCTCATTTACAAGTCTTGG	GAGTATTCTCATCTACAAGTCTTGG	EYSHLQVLX
23	Contig3178	547	TC	GCCTTTTAACTTTTGAACAATGCTA	GCCTTTTAACTTCTGAACAATGCTA	AF*LLNNAX
24	Contig3416	217	AC	CGGCTATGCCTGAGATCAGGATATA	CGGCTATGCCTGCGATCAGGATATA	RLCLRSGYX
25	Contig3430	990	AC	GTAAGGCCTTGGAGACATCAAGCCT	GTAAGGCCTTGGCGACATCAAGCCT	VRPWRHQAX
26	Contig3657	268	CT	AGGAGGTGGCGTCTGTCCAGGGAGC	AGGAGGTGGCGTTTGTCCAGGGAGC	RRWRLSREX
27	Contig4255	327	TC	AGGGTCGACCTGTTGAAAACCCTAA	AGGGTCGACCTGCTGAAAACCCTAA	RVDLLKTLX
28	Contig4301	182	CT	AAGACAATGAAACTGGCGAGCTTTT	AAGACAATGAAATTGGCGAGCTTTT	KTMKLASFX
29	Contig4402	419	TC	TCATTGACTGTCTTAGGCTCTATTT	TCATTGACTGTCCTAGGCTCTATTT	SLTVLGSIX
30	Contig4702	210	CT	CCTTGAGTACCACTGGAAGGCGGTA	CCTTGAGTACCATTGGAAGGCGGTA	P*VPLEGGX
31	Contig5563	748	CT	ACCTCTAGGTGGCTATGTATCAACC	ACCTCTAGGTGGTTATGTATCAACC	TSRWLCINX
32	Contig6405	288	TC	GAAAAATTATGCTTGGCATTGGTTT	GAAAAATTATGCCTGGCATTGGTTT	EKLCLALVX
33	Contig6542	450	CT	AGTGAACGTTAACTGGAAGCTGTAT	AGTGAACGTTAATTGGAAGCTGTAT	SER*LEAVX
34	Contig6571	300	CT	CCCTTTCCCCTGCTGCCCCTCGGCA	CCCTTTCCCCTGTTGCCCCTCGGCA	PFPLLPLGX
35	Contig7067	1195	CA	ATCACCCAGTACCGGCCCTCCTGGG	ATCACCCAGTACAGGCCCTCCTGGG	ITQYRPSWX
36	Contig7096	711	CT	GAAGTTGAAGAACTGACAGTGAGGA	GAAGTTGAAGAATTGACAGTGAGGA	EVEELTVRX
37	Contig7480	1280	CT	AGACCGACACTACTAGTATTAAGAG	AGACCGACACTATTAGTATTAAGAG	RPTLLVLRX
38	Contig7843	207	AC	CCGCGAAGGAGCAGAGAAGGGCGAG	CCGCGAAGGAGCCGAGAAGGGCGAG	PRRSREGRX
39	Contig7843	716	CT	GTAACGATCTTCCTGTTCAACTTCT	GTAACGATCTTCTTGTTCAACTTCT	VTIFLFNFX
40	Contig7874	195	TC	TGATATGAGTGTTTGCGTGGAGTTC	TGATATGAGTGTCTGCGTGGAGTTC	*YECLRGVX
41	Contig8067	1232	CT	ATGAGTGTCTGTCTGTTAACTTTTT	ATGAGTGTCTGTTTGTTAACTTTTT	MSVCLLTFX
42	Contig8309	176	CT	TGCCAAAGGAGTCTGAAGCCAGAAA	TGCCAAAGGAGTTTGAAGCCAGAAA	CQRSLKPEX
43	Contig8316	126	TC	AAAAGAGAACATTTGTATTGGGTTC	AAAAGAGAACATCTGTATTGGGTTC	KREHLYWVX
44	Contig8482	155	CT	CGCCATGCCCAACTGTTGCATAAGG	CGCCATGCCCAATTGTTGCATAAGG	RHAQLLHKX

APPENDIX VII						
List of Nonsynonymous SNP identified by QualitySNP						
Contig No	Position	SNP	Normal Sequence	Sequence with Base change	Transcribed Proteins	
Contig46	348	CT	AGAGGGGTTGAACGGCAGCACTGCG	AGAGGGGTTGAATGGCAGCACTGCG	RGVERQHCX	RGVEWQHCX
Contig286	267	GC	AATCATCCGTTCGTTGTAGAGTAAC	AATCATCCGTTCCTTGTAGAGTAAC	NHPFVVE*X	NHPFLVE*X
Contig297	2816	TC	CTCAGATATTAGTAGCTCACCTCTT	CTCAGATATTAGCAGCTCACCTCTT	LRY**LTSX	LRY*QLTSX
Contig381	287	AC	GTTGTATGCATCATCAATCCTAGTC	GTTGTATGCATCCTCAATCCTAGTC	VVCIINPSX	VVCILNPSX
Contig499	408	GT	GGGTCACAGGTAGAATCTGCCTGTT	GGGTCACAGGTATAATCTGCCTGTT	GSQVESACX	GSQV*SACX
Contig499	627	TC	GGGAGGAGGATTTGTCATGAAAAGA	GGGAGGAGGATTCGTCATGAAAAGA	GRRICHEKX	GRRIRHEKX
Contig507	157	AG	GTACGCCCTTCAAAAACAAAACGAT	GTACGCCCTTCAGAAACAAAACGAT	VRPSKTKRX	VRPSETKRX
Contig507	210	GA	GCCAAACCTGTAGATCAGGTGTTGC	GCCAAACCTGTAAATCAGGTGTTGC	AKPVDQVLX	AKPVNQVLX
Contig507	255	CT	TACCCCGGTGACCTCCCGGAGAGGG	TACCCCGGTGACTTCCCGGAGAGGG	YPGDLPERX	YPGDFPERX
Contig507	297	AG	GGAGAAATAGTCAGTGGCTGCCAGG	GGAGAAATAGTCGGTGGCTGCCAGG	GEIVSGCQX	GEIVGGCQX
Contig507	465	GA	GGATTGCGTCCAGTACGCCCGTACG	GGATTGCGTCCAATACGCCCGTACG	GLRPVRPYX	GLRPIRPYX
Contig507	480	AG	GACGCAATCCGTATACATGGTCGGC	GACGCAATCCGTGTACATGGTCGGC	DAIRIHGRX	DAIRVHGRX
Contig630	63	TA	TCTCTCTCTCTCTCTCTCATGGGGTTC	TCTCTCTCTCTCACTCATGGGGTTC	SLSLSHGVX	SLSLTHGVX
Contig653	133	TC	TGACCCCATTTTTGGAGGGCGTGTG	TGACCCCATTTTCGGAGGGCGTGTG	*PHFWRACX	*PHFRRACX
Contig653	265	CT	TCGAACGCCATACTCCGGGTTGACA	TCGAACGCCATATTCCGGGTTGACA	SNAILRVDX	SNAIFRVDX
Contig677	360	CT	AGAAGAGGCTATCACTTGGTCGGAA	AGAAGAGGCTATTACTTGGTCGGAA	RRGYHLVGX	RRGYYLVGX
Contig677	378	TC	GTCGGAATTTCTTGATGTATTTAAC	GTCGGAATTTCTCGATGTATTTAAC	VGIS*CI*X	VGISRCI*X
Contig677	696	GA	AGGTAGTTTTCGGCTACCGCAGCAG	AGGTAGTTTTCGACTACCGCAGCAG	R*FSATAAX	R*FSTTAAX
Contig677	759	TC	TTCTTCATCAGGTTTTGGTAAATGG	TTCTTCATCAGGCTTTGGTAAATGG	FFIRFW*MX	FFIRLW*MX
Contig679	256	AG	AGGTGTCCGGCAACAGATAAAGTTG	AGGTGTCCGGCAGCAGATAAAGTTG	RCPATDKVX	RCPAADKVX
Contig679	486	AG	GTACCAGATGAGATTTATGGCACTT	GTACCAGATGAGGTTTATGGCACTT	VPDEIYGTX	VPDEVYGTX
Contig732	1015	GA	TTTTACAGGAACGCCGGATAAACAA	TTTTACAGGAACACCGGATAAACAA	FYRNAG*TX	FYRNTG*TX
Contig737	143	AT	CAAGTCCATCTCATCGCCCTCTTCT	CAAGTCCATCTCTTCGCCCTCTTCT	QVHLIALFX	QVHLFALFX
Contig737	174	GA	GTCAGAAATCTCGACAGGATGTCCT	GTCAGAAATCTCAACAGGATGTCCT	VRNLDRMSX	VRNLNRMSX
Contig737	191	TC	GATGTCCTCCCTTCTGATCGTCACC	GATGTCCTCCCTCCTGATCGTCACC	DVLPSDRHX	DVLPPDRHX
Contig737	243	CT	TCCGTCTTCCTCCGAATCAGCTCGT	TCCGTCTTCCTCTGAATCAGCTCGT	SVFLRISSX	SVFL*ISSX
Contig737	265	TC	CGTACACGAACTTATCCATCACTTT	CGTACACGAACTCATCCATCACTTT	RTRTYPSLX	RTRTHPSLX
Contig768	204	-G	GCCGGAGTGTGGAAGGGCCTCTCCG	GCCGGAGTGTGGGAAGGGCCTCTCC	AGVWKGLSX	AGVWEGPLX
Contig866	760	AG	CAAACTTTTCTAAGCTCTCGTCCTT	CAAACTTTTCTAGGCTCTCGTCCTT	QTFLSSRPX	QTFLGSRPX
Contig987	260	C-	GGAAGCATCCTCCAACACTACTTTG	GGAAGCATCCTCAACACTACTTTGG	GSILQHYFX	GSILNTTLX
Contig1092	1398	AG	GATCTTTCTTCTAGGTGCTGCCGTA	GATCTTTCTTCTGGGTGCTGCCGTA	DLSSRCCRX	DLSSGCCRX
Contig1092	1415	TC	CTGCCGTAACAGTGGAGTTGGTGTC	CTGCCGTAACAGCGGAGTTGGTGTC	LP*QWSWCX	LP*QRSWCX
Contig1104	237	G-	CAGGGAGGGGGGGAGCTCCTAGCTC	CAGGGAGGGGGGAGCTCCTAGCTCG	QGGGELLAX	QGGGSS*LX
Contig1145	315	GA	GGTTTGAATTTCGCCCATCCCTGAT	GGTTTGAATTTCACCCATCCCTGAT	GLNFAHP*X	GLNFTHP*X
Contig1234	125	GC	CTGGTCGTCCTCGTCCTGGCCCTGG	CTGGTCGTCCTCCTCCTGGCCCTGG	LVVLVLALX	LVVLLLALX
Contig1270	420	GA	GAGGAAGATGTTGAGTAGCGTGTTC	GAGGAAGATGTTAAGTAGCGTGTTC	EEDVE*RVX	EEDVK*RVX
Contig1284	103	TC	AATATAGGCTAGTTTCCCTTGAAAT	AATATAGGCTAGCTTCCCTTGAAAT	NIG*FPLKX	NIG*LPLKX
Contig1284	123	GA	GAAATGACCTTAGATCGGTCAGATT	GAAATGACCTTAAATCGGTCAGATT	EMTLDRSDX	EMTLNRSDX
Contig1284	160	TC	CGATTGCTAGATTGCCTTGATTCTG	CGATTGCTAGATCGCCTTGATTCTG	RLLDCLDSX	RLLDRLDSX
Contig1347	499	-C	GATCGGAAAACCGCCTCACTGCACC	GATCGGAAAACCCGCCTCACTGCAC	DRKTASLHX	DRKTRLTAX
Contig1347	688	CT	AAGTAGTGGATCCGACGACGAGGCC	AAGTAGTGGATCTGACGACGAGGCC	K*WIRRRGX	K*WI*RRGX
Contig1347	703	GC	CGACGAGGCCAGGGTGACCGATGTT	CGACGAGGCCAGCGTGACCGATGTT	RRGQGDRCX	RRGQRDRCX
Contig1407	156	TC	CGATGAGGACGTTGACTGTTGAACT	CGATGAGGACGTCGACTGTTGAACT	R*GR*LLNX	R*GRRLLNX
Contig1407	239	AG	GACGTTAACCTGACATGCCTGTGTT	GACGTTAACCTGGCATGCCTGTGTT	DVNLTCLCX	DVNLACLCX

APPENDIX VIII						
List of SSRs identified by MISA						
Contig ID	Number of SSR	SSR type	SSR	size	start	end
Contig4	1	p2	(AG)8	16	1506	1521
Contig16	1	p2	(GA)7	14	353	366
Contig23	1	p2	(GA)11	22	289	310
Contig27	1	c	(CA)8ccaggccaggtactctctccttc(CT)7	53	89	141
Contig46	1	p1	(A)10	10	525	534
Contig46	2	p1	(T)10	10	1246	1255
Contig53	1	p2	(GA)8	16	57	72
Contig53	2	p2	(GA)7	14	206	219
Contig65	1	p2	(GA)16	32	12	43
Contig74	1	p2	(AT)9	18	145	162
Contig77	1	p1	(T)11	11	289	299
Contig83	1	p2	(TC)6	12	234	245
Contig91	1	p2	(GA)12	24	149	172
Contig98	1	p2	(TC)17	34	426	459
Contig110	1	p1	(T)10	10	322	331
Contig120	1	p1	(T)13	13	1	13
Contig125	1	c	(AG)19aacaagttgcctcaacaacagtgaccaagtgatgggtagcataccc(T)10	94	1027	1120
Contig145	1	p1	(A)10	10	110	119
Contig153	1	p2	(CG)6	12	168	179
Contig164	1	p1	(T)19	19	246	264
Contig176	1	p3	(TCT)5	15	242	256
Contig180	1	p3	(GCC)5	15	431	445
Contig182	1	p2	(GA)7	14	799	812
Contig189	1	p1	(G)12	12	384	395
Contig196	1	p2	(AG)8	16	234	249
Contig197	1	p2	(AG)17	34	324	357
Contig198	1	p3	(GAG)5	15	505	519
Contig211	1	p2	(TC)16	32	45	76
Contig219	1	p1	(A)10	10	790	799
Contig223	1	p1	(A)10	10	19	28
Contig224	1	p2	(AG)7	14	724	737
Contig229	1	p2	(AG)13	26	527	552
Contig234	1	p2	(CT)6	12	1587	1598
Contig241	1	p1	(T)14	14	617	630

Contig241	2	p1	(A)13	13	1143	1155
Contig244	1	p1	(T)10	10	477	486
Contig253	1	p3	(CCT)6	18	71	88
Contig260	1	p1	(A)10	10	19	28
Contig260	2	p2	(CT)10	20	1261	1280
Contig265	1	p1	(A)12	12	3420	3431
Contig267	1	p2	(CT)9	18	9	26
Contig267	2	p3	(TCC)5	15	193	207
Contig275	1	p2	(GA)17	34	919	952
Contig276	1	p2	(GA)12	24	927	950
Contig291	1	p3	(CTT)7	21	146	166
Contig292	1	p2	(GA)10	20	1028	1047
Contig293	1	p3	(CCA)5	15	842	856
Contig294	1	p2	(CT)14	28	1	28
Contig294	2	c	(CT)11(CA)8	38	923	960
Contig294	3	p1	(A)14	14	1647	1660
Contig301	1	p3	(AAT)7	21	342	362
Contig305	1	p1	(G)10	10	139	148
Contig325	1	p3	(GCG)5	15	987	1001
Contig332	1	p5	(CTTCC)5	25	402	426
Contig337	1	p3	(GAG)5	15	1208	1222
Contig338	1	p2	(CT)7	14	386	399
Contig339	1	p2	(CT)7	14	383	396
Contig345	1	p2	(CT)19	38	1	38
Contig353	1	p2	(AG)13	26	1083	1108
Contig354	1	p1	(A)10	10	754	763
Contig354	2	p3	(GGC)5	15	973	987
Contig357	1	p1	(A)15	15	1040	1054
Contig362	1	p2	(AG)21	42	1	42
Contig365	1	p3	(CGA)6	18	877	894
Contig373	1	p3	(CGT)5	15	137	151
Contig374	1	c	(AAG)6agcagaagaaatcgaaaccct(AG)16	71	56	126
Contig375	1	p2	(GA)11	22	325	346
Contig391	1	c	(TC)7ta(TC)7	30	1055	1084
Contig392	1	c	(T)15agccaaaacgggacaaataattttttgtattgagaatgtaggtctgcatt(A)12	77	1	77
Contig393	1	c	(T)15agccaaaacgggacaaataattttttgtattgagaatgtaggtctgcatt(A)12	77	1	77
Contig398	1	p3	(CTG)5	15	1147	1161
Contig404	1	p2	(AG)6	12	2818	2829
Contig408	1	p2	(AT)12	24	771	794

Contig414	1	p2	(CT)10	20	75	94
Contig417	1	p2	(CA)6	12	861	872
Contig418	1	p3	(ATG)5	15	615	629
Contig420	1	p3	(ATG)5	15	1380	1394
Contig429	1	c	gaggccggatatcgccggggagcgcaggaaggaagaaggggagggaggggagatgagag	113	2558	2670
Contig430	1	p1	(G)10	10	21	30
Contig441	1	p2	(CT)16	32	758	789
Contig449	1	p1	(G)10	10	1	10
Contig451	1	p2	(TC)12	24	451	474
Contig458	1	p3	(CAG)5	15	357	371
Contig469	1	p3	(TAT)5	15	174	188
Contig474	1	p1	(A)19	19	12	30
Contig484	1	p2	(GA)7	14	1098	1111
Contig492	1	p3	(AGG)5	15	269	283
Contig511	1	p1	(T)11	11	903	913
Contig516	1	c	(AC)6gagcacaacggctcaac(CA)6	41	692	732
Contig539	1	p3	(GGA)5	15	261	275
Contig540	1	p3	(GGA)5	15	261	275
Contig542	1	p2	(GA)6	12	449	460
Contig543	1	p2	(GA)6	12	449	460
Contig544	1	c	ctggtactcgccgccaccgacagagctcggcggcgccgcgtaccggcggcccatcggcgtcag	134	320	453
Contig545	1	p3	(GCC)6	18	279	296
Contig550	1	p3	(CAG)7	21	32	52
Contig551	1	p2	(GA)9	18	567	584
Contig551	2	c	aagagggggagaccaaatgatagagaaagtacgtatgtaagggaagaagaagaaagacga	106	721	826
Contig555	1	p3	(GGA)5	15	67	81
Contig558	1	p2	(GA)7	14	3	16
Contig562	1	p3	(AGC)5	15	286	300
Contig563	1	c	(GA)16aggatgggtttgtagggaacaacgctaaggggtgggggttctctggaaggg(T)10	94	224	317
Contig564	1	p1	(T)12	12	28	39
Contig565	1	p1	(T)15	15	17	31
Contig566	1	c*	(CATA)6(AT)12*(TGTA)5*	64	314	377
Contig575	1	p3	(GCT)6	18	2141	2158
Contig581	1	c	atgcatatgcgatgccagacgaatccatctagacatgagtcgtaaggagggagtggggttcggg	125	741	865
Contig583	1	p3	(CGC)10	30	540	569
Contig585	1	p2	(CT)8	16	152	167
Contig587	1	p2	(GA)7	14	769	782
Contig588	1	p2	(GA)7	14	1520	1533
Contig589	1	p1	(T)10	10	177	186

APPENDIX IX

List of SSRs identified by SSRIT

Contig ID	SSR number	SSR type	SSR	size	start	end
Contig4	1	p2	(AG)8	16	1506	1521
Contig16	1	p2	(GA)7	14	353	366
Contig23	1	p2	(GA)11	22	289	310
Contig27	1	c	(CA)8ccaggccaggtactctctccttc(CT)7	53	89	141
Contig46	1	p1	(A)10	10	525	534
Contig46	2	p1	(T)10	10	1246	1255
Contig53	1	p2	(GA)8	16	57	72
Contig53	2	p2	(GA)7	14	206	219
Contig65	1	p2	(GA)16	32	12	43
Contig74	1	p2	(AT)9	18	145	162
Contig77	1	p1	(T)11	11	289	299
Contig83	1	p2	(TC)6	12	234	245
Contig91	1	p2	(GA)12	24	149	172
Contig98	1	p2	(TC)17	34	426	459
Contig110	1	p1	(T)10	10	322	331
Contig120	1	p1	(T)13	13	1	13
Contig125	1	c	caagttgcctcaacaacagtgaccaagtgatgggtagcata	94	1027	1120
Contig145	1	p1	(A)10	10	110	119
Contig153	1	p2	(CG)6	12	168	179
Contig164	1	p1	(T)19	19	246	264
Contig176	1	p3	(TCT)5	15	242	256
Contig180	1	p3	(GCC)5	15	431	445
Contig182	1	p2	(GA)7	14	799	812
Contig189	1	p1	(G)12	12	384	395
Contig196	1	p2	(AG)8	16	234	249
Contig197	1	p2	(AG)17	34	324	357
Contig198	1	p3	(GAG)5	15	505	519
Contig211	1	p2	(TC)16	32	45	76
Contig219	1	p1	(A)10	10	790	799
Contig223	1	p1	(A)10	10	19	28
Contig224	1	p2	(AG)7	14	724	737
Contig229	1	p2	(AG)13	26	527	552
Contig234	1	p2	(CT)6	12	1587	1598
Contig241	1	p1	(T)14	14	617	630
Contig241	2	p1	(A)13	13	1143	1155
Contig244	1	p1	(T)10	10	477	486
Contig253	1	p3	(CCT)6	18	71	88
Contig260	1	p1	(A)10	10	19	28
Contig260	2	p2	(CT)10	20	1261	1280
Contig265	1	p1	(A)12	12	3420	3431
Contig267	1	p2	(CT)9	18	9	26
Contig267	2	p3	(TCC)5	15	193	207
Contig275	1	p2	(GA)17	34	919	952
Contig276	1	p2	(GA)12	24	927	950
Contig291	1	p3	(CTT)7	21	146	166
Contig292	1	p2	(GA)10	20	1028	1047
Contig293	1	p3	(CCA)5	15	842	856
Contig294	1	p2	(CT)14	28	1	28
Contig294	2	c	(CT)11(CA)8	38	923	960
Contig294	3	p1	(A)14	14	1647	1660
Contig301	1	p3	(AAT)7	21	342	362
Contig305	1	p1	(G)10	10	139	148
Contig325	1	p3	(GCG)5	15	987	1001
Contig332	1	p5	(CTTCC)5	25	402	426
Contig337	1	p3	(GAG)5	15	1208	1222
Contig338	1	p2	(CT)7	14	386	399
Contig339	1	p2	(CT)7	14	383	396
Contig345	1	p2	(CT)19	38	1	38
Contig353	1	p2	(AG)13	26	1083	1108
Contig354	1	p1	(A)10	10	754	763
Contig354	2	p3	(GGC)5	15	973	987
Contig357	1	p1	(A)15	15	1040	1054
Contig362	1	p2	(AG)21	42	1	42
Contig365	1	p3	(CGA)6	18	877	894
Contig373	1	p3	(CGT)5	15	137	151
Contig373	2	c	gctgctgctccgtagcctccgtctctctgtgggccgcagcgat	147	253	399
Contig374	1	c	(AAG)6agcagaagaaatcgaaaccct(AG)16	71	56	126
Contig375	1	p2	(GA)11	22	325	346

Contig391	1	c	(TC)7ta(TC)7	30	1055	1084
Contig392	1	c	caaaacgggacaaataattttttgtattgagaatgtaggtctg	77	1	77
Contig393	1	c	caaaacgggacaaataattttttgtattgagaatgtaggtctg	77	1	77
Contig398	1	p3	(CTG)5	15	1147	1161
Contig404	1	p2	(AG)6	12	2818	2829
Contig408	1	p2	(AT)12	24	771	794
Contig414	1	p2	(CT)10	20	75	94
Contig417	1	p2	(CA)6	12	861	872
Contig418	1	p3	(ATG)5	15	615	629
Contig420	1	p3	(ATG)5	15	1380	1394
Contig429	1	c	cgccggggagcgcaggaaggaagaaggggagggagg	113	2558	2670
Contig430	1	p1	(G)10	10	21	30
Contig441	1	p2	(CT)16	32	758	789
Contig449	1	p1	(G)10	10	1	10
Contig451	1	p2	(TC)12	24	451	474
Contig458	1	p3	(CAG)5	15	357	371
Contig469	1	p3	(TAT)5	15	174	188
Contig474	1	p1	(A)19	19	12	30
Contig484	1	p2	(GA)7	14	1098	1111
Contig492	1	p3	(AGG)5	15	269	283
Contig511	1	p1	(T)11	11	903	913
Contig516	1	c	(AC)6gagcacaacggctcaac(CA)6	41	692	732
Contig539	1	p3	(GGA)5	15	261	275
Contig540	1	p3	(GGA)5	15	261	275
Contig542	1	p2	(GA)6	12	449	460
Contig543	1	p2	(GA)6	12	449	460
Contig544	1	c	ccaccgacagagctcggcggcgccgcgtaccggcggcc	134	320	453
Contig545	1	p3	(GCC)6	18	279	296
Contig550	1	p3	(CAG)7	21	32	52
Contig551	1	p2	(GA)9	18	567	584
Contig551	2	c	gaccaaatgatagagaaagtacgtatgtaagggaagaag	106	721	826
Contig555	1	p3	(GGA)5	15	67	81
Contig558	1	p2	(GA)7	14	3	16
Contig562	1	p3	(AGC)5	15	286	300
Contig563	1	c	gggtttgtagggaacaacgctaaggggtgggggttctctg	94	224	317
Contig564	1	p1	(T)12	12	28	39
Contig565	1	p1	(T)15	15	17	31
Contig566	1	c*	(CATA)6(AT)12*(TGTA)5*	64	314	377
Contig575	1	p3	(GCT)6	18	2141	2158
Contig581	1	c	tgccagacgaatccatctagacatgagtcgtaaggaggga	125	741	865
Contig583	1	p3	(CGC)10	30	540	569
Contig585	1	p2	(CT)8	16	152	167
Contig587	1	p2	(GA)7	14	769	782
Contig588	1	p2	(GA)7	14	1520	1533
Contig589	1	p1	(T)10	10	177	186
Contig590	1	p1	(T)10	10	177	186
Contig591	1	p1	(T)10	10	1358	1367
Contig603	1	p3	(CTG)5	15	342	356
Contig630	1	p2	(CT)15	30	30	59
Contig631	1	c	tctccttctctgtttgct(TC)13ctctttcttcactctgcttccact	123	10	132
Contig632	1	p2	(TC)15	30	1422	1451
Contig633	1	c	atatataacatctgtatgtacacatatatacatatactttatatcg	125	58	182
Contig636	1	p2	(TC)9	18	55	72
Contig637	1	p2	(TC)9	18	55	72
Contig645	1	p1	(A)12	12	202	213
Contig649	1	p3	(CTG)5	15	1756	1770
Contig651	1	p2	(GA)9	18	287	304
Contig651	2	p2	(GA)13	26	595	620
Contig652	1	c	(CA)9(GA)14	46	902	947
Contig657	1	p1	(A)11	11	219	229
Contig673	1	p2	(GA)14	28	238	265
Contig674	1	c	caagaattcataggaagctcgatcacacttggcagtaatatt	114	523	636
Contig677	1	p3	(GCA)5	15	993	1007
Contig679	1	p3	(GCA)5	15	993	1007
Contig691	1	p1	(A)11	11	1	11
Contig691	2	p3	(CTC)6	18	271	288
Contig691	3	p3	(CAG)5	15	553	567
Contig692	1	p3	(CCT)6	18	32	49
Contig692	2	p3	(TCC)5	15	158	172
Contig693	1	p1	(T)11	11	322	332
Contig693	2	p1	(T)10	10	2132	2141
Contig700	1	p2	(CT)10	20	1061	1080

Contig700	2	p2	(GA)20	40	2335	2374
Contig725	1	p2	(TC)8	16	216	231
Contig726	1	p2	(AG)13	26	454	479
Contig728	1	p2	(CT)6	12	94	105
Contig730	1	p2	(GA)6	12	435	446
Contig730	2	p2	(AG)6	12	646	657
Contig732	1	p2	(CT)11	22	1488	1509
Contig733	1	c	0ttctgtcccctcacta(AATT)6tagtgaggggacagaa(	96	272	367
Contig733	2	p3	(CTC)5	15	1144	1158
Contig739	1	p1	(A)10	10	2223	2232
Contig742	1	p2	(GA)10	20	599	618
Contig744	1	p2	(CT)26	52	1	52
Contig745	1	p2	(TC)13	26	114	139
Contig745	2	c	(AGC)5(AAC)7	36	342	377
Contig747	1	p1	(T)14	14	34	47
Contig747	2	p1	(C)14	14	1010	1023
Contig747	3	p1	(T)11	11	1173	1183
Contig749	1	p1	(T)14	14	34	47
Contig749	2	p1	(T)11	11	1166	1176
Contig751	1	p1	(T)11	11	3083	3093
Contig755	1	p3	(GCC)8	24	212	235
Contig756	1	p3	(GGA)6	18	1100	1117
Contig761	1	c	AG)9aagaaa(AGACG)5ggacgggagggagagag	182	928	1109
Contig788	1	p3	(GCT)6	18	944	961
Contig800	1	p3	(GGC)5	15	227	241
Contig800	2	p2	(AT)6	12	567	578
Contig804	1	c	(TC)8tatgtaatctgtgtgtgtggatgtgcggcgt(G)17	64	378	441
Contig807	1	c	aaggggagagagggagtgtgacatagcagagaacaga(	138	53	190
Contig819	1	p4	(TCAC)6	24	566	589
Contig822	1	p2	(GA)20	40	922	961
Contig825	1	p1	(T)10	10	553	562
Contig827	1	p2	(AG)8	16	199	214
Contig827	2	p3	(AGC)5	15	724	738
Contig828	1	p1	(A)10	10	1414	1423
Contig829	1	p1	(T)10	10	847	856
Contig829	2	p1	(A)17	17	982	998
Contig830	1	p1	(T)11	11	986	996
Contig831	1	p3	(CGC)7	21	600	620
Contig832	1	p3	(CGG)7	21	317	337
Contig839	1	p2	(GA)6	12	1220	1231
Contig840	1	p2	(GA)7	14	122	135
Contig842	1	p2	(TC)6	12	59	70
Contig843	1	p2	(TC)6	12	59	70
Contig844	1	c	(CT)11(CA)6	34	1	34
Contig855	1	p1	(G)12	12	1867	1878
Contig863	1	p2	(TC)7	14	392	405
Contig863	2	p2	(AT)7	14	593	606
Contig863	3	p2	(TA)6	12	932	943
Contig864	1	p2	(TC)7	14	404	417
Contig864	2	p2	(AT)7	14	605	618
Contig864	3	p2	(TA)6	12	944	955
Contig865	1	p3	(CCT)5	15	87	101
Contig873	1	p3	(GGA)5	15	290	304
Contig885	1	p1	(T)10	10	473	482
Contig890	1	p3	(CCG)5	15	818	832
Contig893	1	p1	(A)10	10	244	253
Contig907	1	p3	(TTC)5	15	831	845
Contig914	1	p2	(AG)8	16	20	35
Contig917	1	p1	(A)16	16	10	25
Contig923	1	c	(TC)8g(CT)11ttcatacgcagaa(AC)6	64	57	120
Contig924	1	p1	(G)11	11	1	11
Contig925	1	c	(AG)8atgattggcttgtccttggggtcggggt(AG)16	76	579	654
Contig930	1	p2	(TC)6	12	2723	2734
Contig932	1	p1	(A)20	20	810	829
Contig944	1	p2	(TC)8	16	3651	3666
Contig965	1	p2	(AG)8	16	1220	1235
Contig970	1	p2	(AG)8	16	162	177
Contig971	1	p3	(TTC)5	15	431	445
Contig991	1	p1	(T)28	28	1773	1800
Contig993	1	p3	(GCA)5	15	148	162
Contig996	1	p1	(A)10	10	160	169
Contig1001	1	p3	(CTG)5	15	683	697

Abstract

"DEVELOPMENT OF MOLECULAR MARKERS FOR BLIGHT DISEASE RESISTANCE IN TARO USING BIOINFORMATICS TOOLS"

ATHUL V. S.

(2013-09-109)

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

119

ABSTRACT

Development of molecular markers using sequential information publicly available in the biological databases has enhanced their credibility over the years. The study entitled "Development of Molecular markers for blight disease resistance in taro using bioinformatics tools" was conducted at the Central Tuber Crop Research Institute (CTCRI) during 2017-2018. The objectives of the study included the development and evaluation of various Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeats (SSR) prediction pipelines, computational prediction and validation of the molecular markers for blight disease resistance in taro.

The preliminary data set for the study was obtained from the Sequence Read Archive (SRA) section of NCBI. A total of 6,479,882 sequences obtained initially were reduced to 6,319,834 after pre-processing. The processed sequences were reduced to 79,608 sequences after *de novo* assembly and were finally assembled to 8547 contigs and 59,242 singlets. The contigs were then processed with various prediction pipelines to predict SSRs and SNPs.

The tools, QualitySNP and AutoSNP were employed to detect the SNPs present within the contig sequences. The efficiency of these tools in determining the number of synonymous and non-synonymous SNPs was also analyzed.

The tools, MISA and SSRIT were used to detect the SSRs within the sequences. The efficiency in predicting more number and types of reliable repeats were considered. The analysis was done with a wide range of repeats such as mono-, di-, tri-, tetra-, penta-, hexa-, and poly repeats and their numbers.

QualitySNP identified 518 synonymous and 44 non-synonymous SNPs from the 8547 contigs. MISA identified 967 mono-, 1484 di-, 558 tri-, 14 tetra-, 2 penta-, 9 hexa-, and 393 compound SSRs. Five SNP and SSR primers were designed and synthesized from the contigs containing SSRs and SNPs. The synthesized SNP and SSR primers were then validated against tolerant and susceptible varieties of taro leaf blight.

Among the primers synthesized the SSR primer CeSSR4 and SNP primer CeSNP3 were capable of differentiating leaf blight resistant and susceptible varieties. The markers need to be analyzed further with a large number of samples to develop them as a marker for taro leaf blight. Once analyzed, they could be used in marker-assisted selection and breeding programmes of taro.