

**COMPARATIVE GENOME ANALYSIS IN COCONUT (*Cocos nucifera* Linn.)
AND MARKER DEVELOPMENT FOR DISTINGUISHING TALL AND
DWARF COCONUT TYPES**

By

SHRI HARI PRASAD

(2019-11-209)



DEPARTMENT OF PLANT BIOTECHNOLOGY

**CENTRE FOR PLANT BIOTECHNOLOGY AND MOLECULAR BIOLOGY
COLLEGE OF AGRICULTURE**

VELLANIKKARA, THRISSUR – 680656

KERALA, INDIA

2021

**COMPARATIVE GENOME ANALYSIS IN COCONUT (*Cocos nucifera* Linn.)
AND MARKER DEVELOPMENT FOR DISTINGUISHING TALL AND
DWARF COCONUT TYPES**

By

SHRI HARI PRASAD

(2019-11-209)

THESIS

Submitted in partial fulfilment of the requirements for the degree of

Master of Science in Agriculture

Faculty of Agriculture

Kerala Agricultural University, Thrissur



DEPARTMENT OF PLANT BIOTECHNOLOGY

**CENTRE FOR PLANT BIOTECHNOLOGY AND MOLECULAR BIOLOGY
COLLEGE OF AGRICULTURE**

VELLANIKKARA, THRISSUR – 680656

KERALA, INDIA

2021

DECLARATION

I, hereby declare that this thesis entitled “**Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types.**” is a bonafide record of research work done by me during the course of research and 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.

Vellanikkara,

Date: 22.02.2022



Shri Hari Prasad

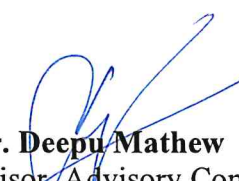
(2019-11-209)

CERTIFICATE

Certified that this thesis entitled “**Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types**” is a record of research work done independently by **Mr. Shri Hari Prasad (2019-11-209)** under my guidance and supervision and that it has not previously formed the basis for the award of any degree, diploma, fellowship or associateship to him.

Vellanikkara,

Date: 22. 2. 22



Dr. Deepu Mathew
(Major Advisor, Advisory Committee)
Associate Professor
Department of Plant Biotechnology
College of Agriculture
Vellanikkara

CERTIFICATE

We, the undersigned members of the advisory committee of **Mr. Shri Hari Prasad (2019-11-209)**, a candidate for the degree of **Master of Science in Agriculture** with major in **Plant Biotechnology**, agree that the thesis entitled “**Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types.**” may be submitted by **Mr. Shri Hari Prasad**, in partial fulfilment of the requirement for the degree.


Dr. Deepu Mathew

(Major Advisor, Advisory Committee)

Associate Professor

Department of Plant Biotechnology
College of Agriculture, Vellanikkara


Dr. Abida P.S.

(Member, Advisory Committee)

Professor and Head

Department of Plant Biotechnology
College of Agriculture, Vellanikkara


Dr. Rehna Augustine

(Member, Advisory Committee)

Assistant Professor

Department of Plant Biotechnology
College of Agriculture, Vellanikkara


Dr. Meera Manjusha A. V.

(Member, Advisory Committee)

Assistant Professor

Regional Agricultural Research Station, Pilicode

- ཨོཾ་མ་ཎི་པདྨེ་ཧཱུྃ་ -

Oṃ Maṇi Padme Hūṃ

I dedicate this thesis to my parents and my family.

ACKNOWLEDGEMENT

*First and foremost, I am forever indebted to my family **Amma, Acha Etta and Ettathiamma**, for their constant prayers, unfathomable love and boundless affection. Words can't express my soulful gratitude to them.*

*I would like to thank my major advisor **Dr. Deepu Mathew**, Associate Professor, CPBMB, College of Agriculture, Vellanikkara, for rendering continuous support and freedom throughout the research work. Not only the academic lessons but also the life lessons taught by him helped me a lot in various circumstances and I really find it fortunate in having him as my guide without whom this thesis would not have been possible.*

*With deep gratitude I thank **Dr. Abida P S**, Professor and Head, CPBMB, College of Agriculture, Vellanikkara, member of my advisory committee member for her support, guidance and blessings.*

*I am greatly indebted to **Mr. Ravi Shankar (DIC)** for his constant support, timely help and valuable guidance rendered in the conduct of the experiments carried out at Bioinformatics Centre, Kerala Agricultural University, Thrissur .*

*I wish to express my gratitude to **Dr. Rehna Augustine**, Assistant Professor, CPBMB, College of Agriculture, Vellanikkara, my advisory committee member for the valuable suggestions and expert advice rendered during the whole venture.*

*I am extremely thankful to **Dr. Meera Manjusha**, Assistant Professor, RARS Pilikkode, member of advisory committee for the generous help provided whole heartedly during this research programme.*

*I do take this opportunity to place on record my indebtedness to **Dr. Smita Nair, Dr. M. R. Shylaja**, Professor and former head of department and other faculties of dept., **Dr. Gibense Hinderson, Dr. Kiran and Dr. Preetha** and all other staff members*

of the Department of Plant Biotechnology for their invaluable help and unsolicited assistance throughout the course of my study.

*I wish to express special thanks to **Mullai Ramamoorthy and Vaisakh** for their help in finishing the research work. I also thank my classmates **Midhuna, Athira, Anjala, Gershome, Varsha, Nahla, Sanjay Sabu and Sanjay Sathiyam** for their affection and kind help.*

*I feel fortunate for having lovable seniors **Shivaji bhayya, Vipul bhayya and Feba chechi** for their jovial companionship, advice, timely help and encouragement. I'm grateful to my juniors **Mounica and Brindha** for their care and affection. My special thanks to all my friends **Shankar, Akhil, Ashish, Abin, Manjunath, Anoop**, for their affection, ever willing help and mental support without which this venture would not have completed this much smoothly.*

*I wish to express my sincere thanks to all the non-teaching staff members and labourers of **CPBMB** for their timely assistance and whole-hearted cooperation.*

*I express my gratitude to **DBT** for financial support and to **Kerala Agricultural University** for the financial and technical support for persuasion of my study and research work.*

It would be impossible to list out all those who have helped me in one way or another in the successful completion of this work. My word of apology to those I have not mentioned in person and a note of thanks one and all who worked for the successful completion of this endeavor.

Shri Hari Prasad

CONTENTS

Chapter	Title	Page No.
1	INTRODUCTION	1-2
2	REVIEW OF LITERATURE	3-23
3	MATERIALS AND METHODS	24-43
4	RESULTS	44-64
5	DISCUSSION	65-71
6	SUMMARY	72-73
7	REFERENCES	I-XVI
	ANNEXURE	
	ABSTRACT	

LIST OF TABLES

Table No.	Title	Page No.
3.1	Parental lines selected for leaf sample collection	38
3.2	Composition of PCR reaction mixture	42
3.3	Programme of thermal cycling	42
4.1	Coconut genome assemblies and raw reads identified	46
4.2	QUAST result Quality characteristics of the assembly	47
4.3	Repeat masking with Dfam+RepBase	51
4.4	Repeat masking with individual libraries in tall and dwarf coconut ecotypes	52
4.5	Repeat masking with combined <i>de novo</i> library in HT, CGD and CAGD	53
4.6	Repeat masking with combined <i>de novo</i> library in CNT and CND	54
4.7	Repeat masking with dwarf library in related palms	55
4.8	Repeat masking with tall library in related palms	56
4.9	Repeat masking with combined repeat library in related palms	57
4.10	Gene prediction results for coconut genome assemblies	58
4.11	Gene prediction results for other palm genomes	59
4.12	The number of unique sequences in each genome (BLAST+ output)	59
4.13	Quality of the DNA samples isolated	60
4.14	Amplification pattern of candidate markers using the primers designed	61
4.15	The number of unique sequences in each genome (BLAST+ output)	63

4.16	Microsatellite motif length and abundance	64
4.17	Simple and compound motifs and abundance in the coconut genomes	64

LIST OF FIGURES

Figure No.	Title	Between Page No.
3.1	Workflow displaying genome assembly retrieval and quality analysis using QUASt	27-28
3.2	Raw reads retrieval and assembly pipeline	29-30
3.3	RepeatModeler pipeline	31-32
3.4	Workflow depicting the combined repeat library development	31-32
3.5	RepeatMasker pipeline	31-32
3.6	BRAKER2 pipeline	33-34
3.7	AUGUSTUS pipeline	35-36
3.8	OmicsBox functional annotation pipeline	37-38

LIST OF PLATES

Plate No.	Title	Between Page No
4.1	Repeat library available at KAU webpage (http://www.kau.in/repeat-libraries)	51-52
4.2	Good quality of genomic DNA seen in agarose gel. L: 100 bp ladder, S1-S10: DNA from coconut accessions	61-62
4.3	Amplification pattern observed using the primer Cocos_21 with the samples	61-62

LIST OF APPENDICES

Appendix No.	Title
1	List of stock solutions used in experiment
2	List of chemicals and other consumables used for wet lab analysis
3	List of laboratory equipment used for the study
4	Online BLAST results
5	Primer sequences
6	Potential polymorphic microsatellite regions

INTRODUCTION

1. INTRODUCTION

The *Cocos* is a monotypic genus (Arecaceae) that accommodates coconut (*C. nucifera* L.). Coconut is one of the world's most valuable palms, with 93 countries growing it. In India's economy, coconut farming is quite important, with an annual production of 20,308.70 million nuts from 2.173 million ha (CDB, 2021). India is the world's third largest coconut producer, following Philippines and Indonesia. In terms of productivity, with 9,345 nuts/ha/year, India is the most productive country. In India, Kerala, Tamil Nadu, Karnataka, and Andhra Pradesh are the largest coconut-producing states, accounting for 84 per cent of the acreage and 87 percent of production (Kappil *et al.*, 2021).

Coconut palm is the state tree of Kerala. In 2019-20, Kerala had a production of 6980.30 million nuts from an area of 0.76 million ha, with the productivity of 9,175 nuts/ ha (CDB, 2021). Even with the highest production, productivity in Kerala is lesser compared to that of Andhra Pradesh (13,969), West Bengal (12,433) and Tamil Nadu (12,280).

Coconut is an important food and economic crop in the humid tropical parts of the globe. Copra and coconut oil are the primary traded products and a key source of foreign money for coconut producing countries. Because of its multiple applications as food, drink, fuel, construction materials, and so on, the coconut palm is frequently referred to as 'the tree of life or *Kalpavriksha*'. Coconut is a significant source of vitamins and minerals, making it a vital element of the human diet. Average-sized nuts that weigh 400 g have enough meat and water to satisfy most people's dietary needs. However, competition from other oilseeds and synthetic fibers have led to a rise in demand for the crop in recent decades. Despite a scarcity of inputs, notably in research and development, the coconut industry is facing the challenges such as the ageing plantations, scarcity of superior planting material, and a variety of pests and diseases.

There are two ecotypes in coconut, tall and dwarf. Even though the dwarf types start flowering in the third year itself and come to regular bearing in the ninth year, they have shorter life span of 40-50 years compared to the tall. The tall types attain a height of 15 to 18 m, long lived up to 80 to 90 years and are fairly resistant to diseases and pests.

Development of varieties with dwarf growth habit, high yield, higher life span, field tolerance to biotic and abiotic stresses, and good kernel and oil recovery is the most important breeding objective in this crop. Breeding attempts for dwarf palm stature are crippled with the non-availability of a precise methodology to identify the dwarf lines at the early plant stage itself. Development of molecular markers linked with this trait shall enable the marker assisted selection for dwarf palms.

Previous attempts for marker development for plant growth habit in coconut were on a single gene basis (Rajesh *et al.* 2016). Of late many dwarf and tall coconut ecotypes are whole genome sequenced and the sequences are made available in public databases (Xiao *et al.*, 2017; Lantican *et al.*, 2019; Muliya *et al.*, 2020; Wang *et al.*, 2021). The pathways for genome assembly, annotation and comparative analyses are also well established (Zhang *et al.*, 2014; Di Genova *et al.*, 2014; Zheng *et al.*, 2011). For this quantitative trait (show references), comparative whole genome analysis can reveal large number of genes which are differentially present in tall and dwarf genotypes. These differential regions shall be potent to develop genome wide markers for this quantitative trait, rather than single markers used in conventional methods.

Hence, the study entitled ‘Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types.’ was undertaken during 2019-2021 with the objective to identify the differential genes and genomic regions among the tall and dwarf coconut genotypes through comparative whole genome sequence analyses and to develop molecular markers for distinguishing tall and dwarf coconut types.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

“A review of pertinent works and thinking by others helps to enlarge, enrich and clarify one’s own work and thinking”

- Young (1996)

Extensive analysis of available literature is useful in gaining insight into and understanding the study problem. This chapter summarizes the systematic survey and review of the available literature relevant to the research entitled “Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types”. The review is organized under the following titles

- 2.1. Coconut
- 2.2. Coconut variability
- 2.3. Coconut genome
- 2.4. DNA isolation
- 2.5. Molecular markers and QTL analysis in coconut
- 2.6. Regulation of plant growth in coconut
- 2.7 Gene governing growth habit in plants
- 2.8. Genome and sequence comparative analyses for gene finding

2.1 COCONUT

The coconut (*Cocos nucifera* L.) is widely cultivated edible oil yielding crop in the tropical regions of the world. Write on the origin and distribution of coconut. Coconut is a small holders crop, with 96 per cent of the plantations having less than 4 ha (Batugal and Oliver, 2003). We need more info under this topic including area and production statistics in world, India and Kerala, nutritional qualities, decline in area and production, major reasons for decline in production etc.

There are two ecotypes of coconut, tall and dwarf. The hybrids of these ecotypes are reported to show promising hybrid vigour for many important traits (Patel, 1938).

2.2. VARIABILITY IN COCONUT

Due to the vast distribution and genetic mixing, coconut classification has yet to be standardized, resulting in the use of various terminologies by researchers to characterize the coconut types. The morphological descriptors for characterizing the coconut germplasm was issued by the then IPGRI (IPGRI, 1995). This crop is primarily classified into tall and dwarf ecotypes, based on their height and breeding habit (Narayana and John 1949).

In order to identify the superior genotypes and to decide on the breeding programmes, it will be critical to assess the nature and degree of diversity. For long-term coconut breeding programmes, an accurate evaluation of genetic connections between coconut types and a determination of genetic diversity are necessary (Perera *et al.*, 2003).

Details on the varieties used in this study are presented below.

2.2.1. TALL

Tall palms, which are also known as *var. typica* Nar., are widely grown commercially throughout the world's coconut growing regions. In coconut-growing locations, tall cultivars take up the majority of the land. Often reaching a height of 25-30 m, they have a pre-bearing age of 6-10 years, but being hardy they continue to produce for the nuts up to 80 years (Nair *et al.*, 2016). Since the male and female stages do not overlap, they are typically cross-pollinated. The fruit ranges in size from medium to big, and the nuts mature in about a year (Nair, 1992).

West Cost Tall (WCT)

This cultivar accounts for roughly 95% of the total coconut area in Kerala. WCT is a tall palm that produces an excellent yield up to 70 years of age. Many regional types in WCT have emerged as a result of continuous cultivation for many centuries and they are identified by the location where they are grown. Further, these regional types adapted to the local environment, provide a strong foundation for valuable alleles in coconut breeding programmes (Remany, 2003).

Javan Giant (JG)

Widely grown in Java, they are believed to be originated in the island. Palms are tall, strong with long leaves, fairly stout trunk and wide leaflets producing heavy big nuts, with an average yield of 95 nuts/ palm/ year and high oil content (66.0 %) (Menon and Pandali, 1958). Do you have a more recent reference for this?

Kappadam Tall (KT)

Widely grown in Thrissur district of Kerala, this hardy palm native to India's southwest coast is a selection from WCT. It is scientifically and economically significant due to its potential to be used in breeding programmes to increase the fruit size and to reduce the husk content. Also known as 'Chappadan' in some regions of Kerala, this cultivar produces the heaviest fruits, which are primarily green, oblong to spherical, and having a thinner husk than the other WCT Indian types. This cultivar is not generally chosen by growers because of the low quantity of nuts produced and the long pre-bearing time (Niral *et al.*, 2014a).

New Guinea Tall (NGT)

The plant is characterized by large, spherical or ellipsoid nuts with colours varying from green to brown. The nuts contain plenty of sweet water in the tender stage. The tree yield 65 nuts per year with high copra and an oil contents of 66 percent (Joseph, 2007).

2.2.2 DWARF

Dwarf palms, also known as var. *nana* (Griff.) Nar., despite their popularity as a household crop, are currently being grown on a large scale for tender nut purpose (Nair *et al.*, 2016). They yield fruits more quickly (3-4 years), have a shorter lifespan and lack the characteristic inflated 'bole' of tall types. Even though they are heavy bearers, bearing is irregular at times. Dwarfs are mostly distinguished by the colour of their nuts (green, yellow, red, and brown). The dwarf ecotypes are believed to be originated from the tall ecotypes due to mutation (Menon and Pandalai, 1958) or through inbreeding in tall palms (Swarninathan and Nambiar, 1961). According to Nair

et al. (2016), the presence of high total and reducing sugars, soluble solids, potassium, lower acidity and flavour-rich water, have led to preference of dwarf coconut over the tall types for tender-nut purpose.

Chawaghat Green Dwarf (CGD)

The dwarf variety is prevalent along Kerala's west coast. The variety is characterised by compact crown, dark green leaves and nut, and complete genetic purity due to self-pollination. Self-pollination occurs when the male and female phases overlap, resulting in homozygous offsprings (Joseph, 2007).

Chawaghat Orange Dwarf (COD)

This variety found in Kerala is mainly used for tender nut purpose. It is more robust than CGD and characterized by thin stem, short orange petioles, orange spathes and spherical small orange nuts. Although it is a self-pollinated variety, cross pollination does occur to some extent (Joseph, 2007).

Chawaghat Yellow Dwarf (CYD)

A unique and indigenous yellow dwarf coconut developed by CPCRI by screening the original population of Chowghat Orange Dwarf. The palms were further refined by inter-seed mating with original mother palms followed by selection for nut colour. The selection was made for the distinctive yellow fruits, rachis, flowers, and petiole, after evaluating the progenies for trait inheritance. The selection is regarded unusual because there is no indigenous yellow dwarf population in continental India. This variety is having economic significance as a parent for crossing with chosen tall types to generate hybrids (Niral *et al.*, 2014b).

Malayan Yellow Dwarf (MYD)

Malayan Yellow Dwarf (MYD) was originated in Java and then spread to Malaysia. The variety is characterized by yellow petioles, spadices and nuts. In Kerala, yellow dwarfs are widely utilized for hybrid seed nut production (Joseph, 2007) and

MYD is often chosen as the best male parent for tall cultivars because of its combining abilities (Ramachandran *et al.*, 1974).

2.3 COCONUT GENOME

The first coconut genome sequenced was that of Hainan Tall (HT) by Xiao *et al.* (2017). They have used Illumina HiSeq 2000 platform to generate 419.67 gigabases (Gb) of raw reads with sequencing depth of 173.32X, which was used to get a total scaffold length of 2.42 Gb constituting 90.91% of the genome. A total of 28,039 protein-coding genes were identified from the genome. It was also reported that, of the 2.42 Gb genome, 72.75% is comprised of transposable elements.

Lantican *et al.* (2019) published the genome draft of dwarf coconut Catigan Green Dwarf (CAGD). The reads obtained using PacBio SMRT sequencing platform, with a read depth of 15X, were error corrected using the assembled Illumina paired-end MiSeq reads with a read depth of 50X. This was further improved through Chicago sequencing to yield a scaffold level assembly of 2.1 Gb. The genome was reported to harbor 34,958 protein coding genes and nearly 78.3% was constituted by repetitive elements. By aligning the non-repetitive regions of HT and CAGD genome sequence reads, 58,503 variants were identified. CAGD genome possess a higher complete set of annotated genes (85.3%) compared to HT which has 81.2% annotated genes.

The genome of Chowghat Green Dwarf (CGD) was published by Muliyar *et al.* (2020). A hybrid sequencing strategy, employing the short reads from Illumina HiSeq 4000 platform (183.51 Gb raw sequence data with 70.6X read depth) and long reads from PacBio RSII platform (37.02 Gb with 14.3X read depth), was used to get a scaffold level assembly of 1.93 Gb representing 75% of the genome. The genome comprised of 13,707 protein coding genes and transposable elements accounted for about 77.29%.

The chloroplast genome of CGD has 154,628 bp size and accommodates 84 protein-coding genes, 38 tRNAs and 4 rRNAs whereas the mitochondrial genome has 744,799 bp size and hosts 123 genes, 26 tRNAs and 6rRNAs (Muliyar *et al.* 2020).

Wang *et al.*, 2021 has reported reference grade assemblies of tall (CNT) and dwarf (CND) coconut. The Nanopore PromethION platform was used to sequence the genomes with a read depth of 116X and 104X for CNT and CND respectively, to get a reference grade chromosomal level assembly of 2.40 Gb and 2.39 Gb respectively. They have also identified high similarity and collinearity between CNT and CND, even though they show large difference in morphological characters.

The coconut mitochondrial genome sequence (cv. Oman Local Tall) was reported by Aljohi *et al.* (2016). A hybrid sequencing strategy was used where the Roche/454 GS FLX sequence reads were error corrected using the Illumina HiSeq 2000 reads. The mitochondrial genome had 678,653 bp size and coded for 72 proteins, 23 tRNAs, 9 truncated proteins and 3 rRNAs. 17.26 % of the mitochondrial genome consisted of repetitive sequences.

The chloroplast genome sequence of dwarf coconut was reported by Huang *et al.* (2013). The 154,731 bp long genome sequenced on Illumina GAIIx platform, with 10X coverage, was predicted to contain 84 protein-coding, 38 tRNA and 8 rRNA genes.

The chloroplast genome of Kopyor Green Dwarf coconut, a unique coconut endemic to Indonesia was reported by Rahmawati *et al.* (2021). The chloroplast genome was 158,462 bp long and present as a quadripartite structure, harbouring 116 genes. Comparative genome analysis with other chloroplast genomes have revealed that gene duplication is responsible for larger sequence size in KGD.

To estimate the genome size of coconut ecotypes, Freitas *et al.* (2016) carried out flow cytometric analysis. The 2C DNA content of tall coconut was 5.72-5.48 pg (average 5.59 pg) whereas it was 5.58-5.52 pg (average 5.55 pg) in dwarf coconut.

2.4 DNA ISOLATION

Genomic DNA isolation is a tedious process in coconut, since the leaves have higher phenolic and polysaccharide content. Mechanical or physiological injuries to the tissues can trigger polyphenol release, which is responsible for tissue browning (Joslyn and Ponting, 1951). The polyphenols undergo rapid oxidation and binds irreversibly to

DNA and proteins (Katterman and Shattuck 1983). The brownish aggregates thus formed shall make the isolate unfit for further molecular analyses by inhibiting further enzymatic interactions.

A protocol for extraction of DNA from the spear leaf of coconut was proposed by Upadhyay *et al.* (1999), where the tissue was pulverised using liquid nitrogen, transferred to 10% sodium dodecyl sulphate (SDS) extraction buffer, pre-heated at 65 °C DNA was stabilized using equal volume of chloroform:isoamyl alcohol (24:1) and then 70% ethanol was used for precipitation.

In the analysis of genetic diversity as well as population structure of coconut germplasm in Florida, Meerow *et al.* (2003) performed DNA extraction from the freshly expanded leaves, which were silica gel dried. FastDNA Kit (BIO 101 Inc.) was used for extracting DNA.

Devakumar *et al.* (2010) have followed a system for plant DNA isolation developed by Upadhyay *et al.* (1999). The system involved homogenization of 5 g of spear leaf tissue using liquid nitrogen which was then transferred to extraction buffer containing 10 % SDS. Further, the mixture was incubated at 65 °C, cooled and DNA was extracted using an equal volume of 24:1 chloroform:isoamyl alcohol mixture.

Angeles *et al.* (2005) have used Dellaporta *et al.* (1983) method which was improvised by Datta *et al.* (1997) and found that the poor quality genomic DNA obtained from the endosperm is due to the high levels of lipid and galacto-mannan contaminants. The improved protocol used polyvinylpolypyrrolidone (PVPP) and a modification in the salt concentration in the extraction buffer (2 M instead of 0.5 M). The newly formed leaves from the fronds yielded good DNA.

In the method used by Manimekalai *et al.* (2006b), sprouting leaves were frozen in liquid nitrogen and pulverized with pestle and mortar. Before the addition of extraction buffer, Poly Vinyl Poly Pyrrolidone (PVPP) (0.50 g) was added to the ground powder. The DNA spool obtained was incubated with 25 ng/L of RNase for one hour at 37 °C. The protocol also used an equal mixture of ice-cold absolute ethanol and 3 M sodium acetate (1/10 volume) (pH 5.2), for precipitating DNA.

A protocol for DNA isolation from plants high in polyphenols, polysaccharides and tannins was developed by Porebski *et al.* (1997). The method used a modified CTAB extraction protocol in which polyphenols and polysaccharides were removed using a high salt concentration and polyvinyl pyrrolidone (PVP). Matured leaf tissues were used to isolate DNA from the wild and cultured octaploid and diploid *Fragaria* species, resulting an average yield of 20-84 µg high quality DNA per one-gram tissue.

Teulat *et al.* (2000) have isolated the genomic DNA from the lyophilized leaf samples by CTAB method and used for genetic diversity analysis in coconut at CIRAD (Montpellier, France).

Perera *et al.* (1998) isolated DNA from the frozen tender coconut leaves using a modified DNA miniprep protocol developed by Dellaporta *et al.* (1983). The method employed pre-mixed phenol/chloroform and isoamyl alcohol for the purification of DNA.

For verifying the homozygous status of the anther culture derived coconut lines, Perrera *et al.* (2008) have isolated good quality DNA by CTAB-method (Doyle and Doyle, 1987) with minor modifications. Concentration of DNA was estimated by comparing the DNA to the fluorescence intensity of a series of standard solutions.

An efficient protocol for the isolation of genomic DNA from plants with excessive polyphenolic contents was given by Couch and Fritz (1990). In this method, prior to lysis, the nuclei are concentrated away from the cytoplasmic constituents, and the synthesis of oxidized polyphenolic chemicals in the residual solution is strongly inhibited.

Plant DNA mini preparation protocol by Dellaporta *et al.* (1983) was based on the protocol described by Davis *et al.* (1980) for the isolation of DNA from yeast. The procedure involves grinding the leaf tissue into fine powder using liquid nitrogen in a mortar, addition of extraction buffer and β-mercaptoethanol, addition of SDS and vigorous stirring. Then potassium acetate will be added and the pellet redissolved in a solution of 50 mM TRIS and 10mM EDTA. The supernatant transferred to a centrifuge tube, to which sodium acetate and isopropanol will be added. The use of 0.3 M sodium

acetate treated samples along with relatively less quantity of isopropanol for precipitation, have yielded high quality DNA.

2.5. MOLECULAR MARKERS AND QTL ANALYSIS IN COCONUT

Molecular markers can be used as tags or probes to find a gene, chromosome, or an individual by identifying changes in gene sequences that are responsible for a trait (Kumar *et al.*, 2009). The identification of candidates carrying vital genes is an imperative function of molecular markers for crop improvement. In the presence of a tightly linked marker with a gene of interest, selection could be made based on scoring for the molecular marker rather than the gene. While in the case of polygenic characters, molecular markers could be used to identify regions in the genome possessing loci for such traits. This could further provide a direct method for the selection of individuals possessing desirable combinations of genes, by facilitate mapping of quantitative trait loci. Apart from diversity and population structure analysis, molecular markers are used in germplasm characterisation, genetic diagnostics, characterization of transformants, genome organisation, and phylogenetic analysis (Sing, 2008).

2.5.1 RAPD in coconut

Jayalakshmi (1996) has developed RAPD markers for distinguishing different coconut genotypes. Markers unique for each of the 17 distinct coconut populations from south pacific area were identified. OPC-4 primer gave the maximum polymorphic bands.

Duran *et al.* (1997) have screened RAPD, MP-PCR and ISTR marker systems to assess the variability among 48 East African tall coconut genotypes. Grouping and association studies have confirmed the predictions on the genetic relations based on known geographical origins and parental ties.

The diversity level of coconut populations from southern Pacific region and Indian ocean regions was studied using RAPD markers. Moderate diversity was seen within the Pacific population but it was low in Indian ocean population. Large

difference was observed between Indian ocean and Pacific ecotypes (Ashburner and Rohde 1994).

The genetic diversity and relationship among coconut accessions were studied using RAPD markers (Upadhyay *et al.*, 2004). Eight highly polymorphic primers were employed to amplify DNA from 81 palms representing Indian and exotic coconut accessions. A total of 77 markers were produced from the 8 primers. and tall accessions showed more genetic diversity than dwarfs. Tall accession showed higher number of polymorphic bands and thus displaying genetic diversity than dwarfs. When compared to exotic accession, indigenous accession revealed less variance, exotic accessions, on the other hand, showed considerable variation, thus displaying the narrow diversity in Indian coconut populations.

Manimekalai and Nagarajan (2006a) have used 199 ISSR markers generated with 19 primers to investigate the polymorphism of 33 coconut accessions from a global coconut collection preserved at the International Gene Bank in India. Tall accessions generated 137 ISSR markers, compared to 135 in dwarf and intermediate accessions. Coconut accessions from Southeast Asia, South Asia, and the South Pacific formed independent groups in the dendrogram and primary coordinate plots. This grouping was typically in agreement with their origin and pattern of coconut spread from their center of origin.

Manimekalai and Nagarajan, (2006b) have used 399 polymorphic RAPD markers generated using 45 random primers to assess the interrelationships among 33 coconut accessions gathered from South Asia (SA), South East Asia (SEA), South Pacific (SP), Atlantic and America, and Africa. The clustering pattern developed was consistent with previous results obtained using RFLP, SSR and AFLP marker systems by Lebrun *et al.* (1998), Perera *et al.* (2000) and Teulat *et al.* (2000), respectively.

Another genetic diversity analysis among 19 coconut populations using 127 polymorphic RAPD markers generated with 24 primers, has developed six clusters. Group 1 included the dwarf cultivars and the rest accommodated tall accessions.

Identification of markers for each population has suggested that they were genetically different (Daher *et al.*, 2002).

The genetic diversity and structure of East African Tall coconut accessions were estimated using RAPD markers. In the study of 120 accessions, ten primers were utilized. Jaccard's coefficient and Nei genetic distances were used for cluster analysis. The results revealed two main clusters, the first cluster with three sub-clusters and the second with two. The findings were able to distinguish among several regions and give evidence of various origins for coconut. Two primary clusters matched the history and distribution of coconuts in Tanzania's coastal zone (Masumbuko *et al.*, 2014).

2.5.2 Simple Sequence Repeat (SSR) markers in coconut

Perera *et al.* (2000) have used microsatellite markers to evaluate the extent of genetic diversity and population structure in 130 coconut accessions comprising of 75 tall and 55 dwarf plants, representing 94 distinct coconut ecotypes throughout the world. The eight sets of SSR primers were used to detect 51 alleles. The tall types displayed fifty alleles, compared to just 26 in dwarfs, and the average diversity value in tall ecotypes was much greater than that in dwarfs. The population was divided in to two groups, group I with only tall species and group II with subgroups (II a-d). Subgroups IIa and IIc consisted mainly of dwarfs and IIb and IId had tall types. All the plants used in the study had generated unique alleles.

Using sequence-tagged microsatellites (STMS) and amplified fragment length polymorphism (AFLP) marker systems, genetic diversity among 31 individuals from 14 coconut communities spanning the whole geographic range was examined (Teulat *et al.*, 2000). Across different communities, 37 SSR primer sets have generated 2-16 alleles per locus. With a total of 339 alleles, genetic diversity varied at 0.47 to 0.90. AFLP analysis with 12 primer combinations yielded 1106 bands, 303 of which were polymorphic. The similarity matrices, cluster and main co-ordinates analyses had given comparable linkages among the populations.

SSR analysis in fifteen selfed and reciprocally crossed progenies of Laccadive Tall and Gangobondam Dwarf, using 10 primer combinations has revealed their genetic

variation (Manimekalai *et al.*, 2005). A total of 42 alleles discovered and the number of alleles per locus varied between two and seven, with an average of 4.2 alleles per primer locus. The population was divided into two groups, Group I had LCT and LCT x GBD progenies while group II had GBD and GBD x LCT progenies.

Rajesh *et al.* (2008a) have employed 14 SSR markers to evaluate the pattern of diversity in 102 coconut trees, representing ten landraces from three coconut-growing communities in India. Ninety alleles were detected, with an average of 6.42 alleles per locus and a polymorphism information content of 0.61. UPGMA cluster analysis revealed two primary clusters, differentiating the tall and dwarf landraces. Within the tall landraces, two sub-clusters were seen and the clustering was based on their geographical regions and breeding habits.

Rajesh *et al.* (2008b) have assessed the genetic variability of coconut accessions from the Andaman and Nicobar islands. Fourteen microsatellite markers were used to screen 100 palms representing 26 native landraces. The SSRs were able to discover a total of 103 alleles, with an average of 7.35 alleles per locus. Tall accessions possessed highest heterozygosity and possessed majority of uncommon alleles. Clustering has grouped the bulk of tall and dwarf accessions individually.

To estimate the variability among the *ex situ* coconut germplasm in Sri Lanka, Dasanayaka *et al.* (2009) analysed 43 representative coconut accessions using 16 SSR markers. Common 'tall' and Pacific tall types were more diverse and possessed high polymorphism information content (PIC) than the autogamous dwarf coconuts. The marker analysis revealed genetic lineages based on evolutionary processes, thus indicating that coconut germplasm has a limited genetic base, with most of the variation restricted to 'tall' coconut.

To determine the genetic purity of coconut hybrids, Rajesh *et al.* (2012) have employed SSR markers. The parental lines CGD and WCT were screened with 50 hyper-polymorphic coconut SSR markers. Screening the DxT hybrids with the markers, had shown that 17 SSR markers generate complimentary alleles as in both parents,

indicating the potential of microsatellite markers for assessing the purity of coconut hybrids.

For the efficient genetic conservation and use of a species in plant breeding programmes, a clear understanding of its mating system is crucial. Since the genetic structure of the population is greatly influenced by the pattern of gene flow pollen, Rajesh *et al.* (2014a) carried out a study to use SSRs to estimate the rate of outcrossing in WCT. Two WCT mother palms and 88 progenies were examined using 15 highly polymorphic microsatellite primers. It was observed that the percentage similarity between the mother palm and the progenies varied from 55 to 74%. The study also used an RAPD primer capable of differentiating Tall palms from dwarf palms and the expected marker was present in all progenies, showing that the pollen came from Tall palms in every case.

Perera *et al.* (2016) have employed SSR markers to investigate the origin and domestication of dwarf coconuts. Allele frequencies were determined at 12 microsatellite loci for 51 Tall and 43 Dwarf coconut cultivars. Further, 246 individuals from 28 Dwarf types were screened using 13 microsatellite markers. Results have shown that the Dwarf types are mostly homozygous with a lower total allele richness. Geographic structure has been visible in the clustering and Dwarf types from different regions were distinguished at the country level. They also investigated the inheritance of height and bole by examining the distribution of the traits in 70 F₂ individuals from a D x T cross. The data implied that a single codominant locus was responsible for the occurrence of a bole and there was no clear link between having a bole and height. Height was also determined by a single codominant gene.

Mauro-Herrera *et al.* (2006) have isolated the WRKY sequences in coconut using degenerate primer pairs, grouped into WRKY groups, and used to develop ten markers. They were further tested in 15 genotypes that represented six different coconut cultivars. The number of alleles have varied between two and four and Single-Strand Conformation Polymorphism (SSCP) analysis has identified the SNP-containing alleles.

Rajes *et al.* (2015) have used 25 SCoT markers to analyze the genetic diversity among 23 coconut germplasm comprising of 10 tall and 13 dwarfs from various geographical locations. Fifteen primers chosen based on their consistent amplification patterns have yielded 102 scoreable bands, 88 percent of which were polymorphic. The similarity coefficient values ranged between 0.37 and 0.91 and the accessions were categorized using UPGMA cluster analysis. The extent of diversity detected using SCoT primers was comparable with the previous reports and coconut accessions from the same geographical location got grouped together. Tall and Dwarf coconut accessions were easily differentiated.

Perera *et al.* (1998) used AFLP profiling of 42 coconut genotypes native to Sri Lanka. Using eight primer pairs (*EcoRI* and *MseI*), 322 amplicons were generated. The tall (*Typica*) type had the maximum variability, followed by the intermediate (*Aurantiaca*) and dwarf (*Nana*) types. According to the hierarchical analysis of molecular variance, dwarf and intermediate types had the highest diversity across, rather than within. The tall types, on the other hand, showed just as much variability among them. *Aurantiaca* had more genetic proximity to the dwarf types than the tall types.

Pesik *et al.* (2017) have developed 16 SNP specific primer pairs from eight SNPs identified from the coconut WRKY genes, optimized the multiplex PCR technique and validated the effectiveness of single nucleotide amplified polymorphism (SNAP) marker for evaluating Kopyor coconut germplasm. For genotyping Kopyor coconut germplasm, duplex PCR utilizing two sets of primer pairs was proven more reliable than triplex PCR. The SNAP markers generated were simple alternative for codominant markers.

Perera *et al.* (2003) have studied the genetic relationships among 94 coconut varieties using 12 pairs of coconut microsatellite markers. Mean genetic diversity in Tall types was nearly double the diversity of Dwarf types. The phenic tree divided the population into two groups with the Tall types from southeast Asia, Pacific, west coast of Panama, and all Dwarf species first group and the tall types from South Asia, Africa,

and the Indian Ocean coast of Thailand in the second group. The allele distribution as well as the clustering have suggested that the dwarfs evolved from tall.

Using 48 SSR loci, Geethanjali *et al.* (2018) have assessed a world-wide coconut germplasm collection of 79 genotypes for genetic diversity and population structure. The genotypes displayed moderately high amount of genetic diversity, which was strongly structured according to the geographical origins. Number of SSR alleles ranged from 2 to 7 with an average of 4.1 per locus. Hierarchical clustering analysis grouped the genotypes into two major clusters with two sub-groups in each, corresponding with the geographic origins. SSR locus CnCir73 (chromosome 1), was putatively related with the fruit yield and it matched to a previously mapped QTL in coconut.

Using 32 SSR and 7 RAPD primers in single marker analysis (SMA), Shalini *et al.* (2007) have identified nine SSR and four RAPD markers linked with mite resistance in coconut. A stepwise multiple regression analysis had shown that a combination of six SSR markers represented 100 percent correlation with mite infestation and a combination of three markers accounted for 83.86 percent of mite resistance.

Through the Selectively Amplified Microsatellite (SAM) strategy, Wu *et al.* (2018) have developed 84 SSR markers for coconut, with 22.1% efficiency estimated from sequencing to polymorphism loci data. Twenty five SAM polymorphic markers were used to screen 42 accessions and the results had shown a genetic similarity coefficient of 0.6 to 0.8. Dwarf coconuts were found to have high similarity coefficient (0.783) compared to the Tall coconuts.

2.5.3 Molecular markers to distinguish tall/dwarf character

Initial attempts in the development of a marker to differentiate the tall and dwarf types were through bulk line analysis using RAPD markers. The DNA bulks of tall, dwarf and intermediate types were amplified with 30 primers and markers were found with OPM 02, OPM 06, and OPC 13 (Manimekalai and Nagarajan, 2010).

DNA pooling technique was also used by Rajesh *et al.* (2013) to identify the RAPD markers for distinguishing tall and dwarf coconut palms. When DNA bulks from tall and dwarf palms were analyzed using 200 RAPD primers, OPAU09 yielded a tall specific marker at 260 bp. The primer was further validated by screening in individual tall and dwarf coconut accessions representing different geographic regions, which revealed that the band was lacking in all dwarf accessions but present in all tall accessions. Further the tall specific band was used to create a Sequence Characterized Amplified Region (SCAR) marker, which was effectively employed to assess the hybrid quality of the dwarf tall crosses.

In a similar study, Rajesh *et al.* (2014b) have screened bulked DNA from tall and dwarf type coconuts using 200 RAPD primers and an OPBA3 marker was found to clearly discriminate the tall and dwarf bulks. The primer was further validated and was utilized to screen the parents of Dwarf x Tall crossings and for hybrid certification.

Rajesh *et al.* (2016) screened 24 SCoT primers to differentiate the tall and dwarf arecanut DNA pools. SCoT 11 has yielded a reproducible polymorphic marker at 1300 bp in tall DNA pool. The marker was also validated further with parental lines as well as crossed progenies.

2.6 QTL MAPPING IN COCONUT

Rohde *et al.* 1999 created the first genome map of coconut using ISTR markers for F₁ population derived from the cross between East African Tall and Laguna Tall . Herran *et al.* (2000) constructed a mapping population from a cross between Malayan Yellow Dwarf x Laguna Tall employing AFLP, ISSR, ISTR and RAPD markers. Sixteen linkage groups were generated using 382 markers and 6 QTLs were identified, which corresponded to early germination. The identification of the correlation between early germination with early flowering and higher yield created the opportunity for marker-assisted selection in coconut Ritter *et al.* (2000).

The mapping population in a set of coconut half-sib progenies from crosses between Cameroon Red Dwarf and Rennell Island Tall, was used to construct the linkage and QTL maps using AFLP and SSR markers. Two hundred and twenty seven

markers were arranged into 16 linkage groups and several QTLs were detected for the numbers of bunches and nuts (Lebrun *et al.*, 2001). Baudouin *et al.* (2006) added 52 new markers to the linkage map, resulting in a minor amendment to the existing map. For the 11 characters studied, a total of 52 putative QTLs were discovered. The QTLs for fruit component weight, endosperm humidity, and fruit production were detected at distinct sites in the genome, implying that picking QTLs for the individual components can result in effective marker-assisted selection for yield.

To construct a valid map, a mapping population should be big enough to incorporate enough genetic information from many segregating gametes. Because of its extended vegetative cycle and modest nut output within a set period, the fundamental challenge in coconut genome mapping is establishing an adequate mapping population. As a result, obtaining a segregating population of coconut of a decent size will take a long time, which was worsened by the poor success of artificial pollination in coconut (Bandaranayake and Kearsey, 2005). Using a simulation study, Bandaranayake (2006) has determined that the effective size of a mapping population for developing a map with consistent resolution in coconut is around 400 individuals. This finding has suggested that the linkage maps constructed by Herran *et al.* (2000) and Lebrun *et al.* (2001) using less than 65 individuals shall not be reliable.

Based on the previous relationship studies in coconut, Perera (2006) has suggested that best segregation of characters can only be achieved by crossing the genotypes from Southeast Asia and the Pacific group with variations from the Indo-Atlantic group. Thus, a large mapping population was created using 350 individuals resultant from a cross between Sri Lanka Red Dwarf and a single Sri Lanka Tall has been established in Sri Lanka, to acquire maximal trait segregation.

The tremendous advancement in crop improvement programs is a result of the use of DNA-based molecular markers. Production of linkage maps have accelerated the study of genetic loci influencing quantitative parameters of economic value, especially in plantation crops such as coconut. Even though several molecular markers have been utilised to build linkage maps and to describe the marker-trait associations in coconut, a genome-wide association or linkage disequilibrium analysis-based mapping has

received less attention. The availability of whole-genome sequences of coconut genotypes, as well as the introduction of next-generation sequencing methods, will encourage genome-wide trait-marker association studies as well as fine-mapping investigations (Rajesh *et al.*, 2021).

2.5. REGULATION OF PLANT GROWTH IN COCONUT

According to Peng *et al.* (1999), an anomalous response to gibberellin is the reason for its short stature. Mutant dwarfing alleles at one of two *Reduced height-1* (*Rht-B1* and *Rht-D1*) loci impart this reduced response to gibberellin. *Rht-B1/Rht-D1* and maize *dwarf-8* (*d8*) are orthologues of the *Arabidopsis Gibberellin Insensitive* (*GAI*) genes coding for proteins that look like nuclear transcription factors and include an SH2-like10 domain, indicating that phosphotyrosine is involved in gibberellin signalling. Since SH2 domains are generally coupled with phosphotyrosine signalling and bind tyrosine-phosphorylated polypeptides at an essential arginine residue. The *GAI/RGA/Rht-D1a/d8* transcription factors contain an SH2-like domain and exhibit features which are characteristic to STAT factors (signal transducers and activators of transcription). Thus implying that phosphotyrosine signalling may be involved in gibberellin-mediated plant growth regulation, similar to the STAT factors that mediate cytokine/ growth-factor control. Therefore mutations in the dominant dwarfing alleles of *d8* and *Rht-1*, like the mutation in the *GAI* allele, affect the N-terminal region of the proteins that they encode, thus making the plant dwarf.

Rht (reduced height) gene is a gain-of-function allele generated by a mutation in a transcription factor involved in the gibberellin signalling pathway (Sasaki *et al.*, 2002) or a point mutation in the *sd1* gene is responsible for a decrease in the GA production in rice (Hedden, 2003). Due to the hexaploid nature of wheat DNA, it lacks recessive alleles like *sd1* in rice, which could otherwise be utilized to create a semi-dwarf wheat strain. Although the rice *SD1* and wheat *RHT* proteins have completely distinct genetic and biochemical activities, their products are connected to gibberellin dysfunction. Thus, manipulating this growth hormonal production or signalling pathways could control the height of plants.

Proteolysis-mediated regulation has emerged as a prominent area in plant hormone signalling. All the phytohormone response pathways are characterised by ubiquitin-mediated degradation of important regulatory proteins (Smalle and Vierstra, 2004). In the auxin response model, the Aux/ IAA proteins form heterodimers in the absence of an auxin stimulus, thus inhibiting ARF transcriptional activity. The Aux/IAA proteins are targeted to the SCF^{TIR1} complex by an unknown receptor, resulting in their ubiquitination and destruction, de-repressing the ARF transcription factors. The ARF targets *Aux/IAA* genes, which create nascent Aux/IAA proteins that reinstate repression on the pathway in a negative feedback loop (Tiwari *et al.*, 2004).

The *GA20ox* gene, which codes for the enzyme involved in gibberellin (GA) production, was found to be mutated in numerous coconut plants, resulting in the short phenotype. The *CnGA20ox* was revealed to be multi-copy genes in coconut, and at least two groups, *CnGA20ox1* and *CnGA20ox2*, were discovered. The nucleotide sequences of the *CnGA20ox1* gene were same in both coconut kinds, but its expression was nearly three times greater in tall coconut leaves than in dwarf coconut leaves, which was in excellent accord with their height. Whereas the nucleotide sequences of the *CnGA20ox2* gene varied in tall and dwarf, but no expression of the *CnGA20ox2* gene was observed in either coconut types (Boonkaew *et al.*, 2018).

2.6. SEQUENCE COMPARISON STUDIES IN CROPS FOR GENE FINDING

Di Genova *et al.* (2014) have identified 240 unique genes in the genome of grape variety 'Sultanina', based on the whole genome comparison with the reference genome PN40024. Among them, unique genes consisted of 130 transposons and 88 putative genes. The biological function of the remaining 22 genes was linked with disease resistance/defense response. Proteolysis, embryo development, carbon-nitrogen bonding, methyltransferase, and anthocyanin production were among the other groups of novel genes identified in the 'Sultanina' genome.

The *Brassica rapa* genome was studied to identify chromosomal linkages, macro-synteny and micro-synteny within blocks, by comparing to the *Arabidopsis* genome. Based on genome comparisons, as a result of recent whole genome triplication

followed by a unique diploidization process, *B. rapa* possesses a distinctive arrangement of ancestral genome blocks. Some of the triplicated copies of *B. rapa* regions were deleted or regenerated, according to a genome-wide synteny comparison of *B. rapa* and *Arabidopsis* (Mun *et al.*, 2009).

The first report of genome-wide patterns of genetic variation in sorghum was by Zheng *et al.* (2011). The comparison revealed a collection of almost 1,500 genes that differentiate sweet and grain sorghum, which are involved in sugar and starch metabolism, nucleic acid metabolism, lignin and coumarin production, stress responses, and DNA damage repair. Furthermore, 1,057,018 SNPs and 99,948 InDels were also identified.

Hurwitz *et al.* (2010) performed the genome-wide analysis of structural variation among three closely related *Oryza* genomes (*O. nivara*, *O. rufipogon*, and *O. glaberrima*). The *O. sativa* genome was taken as reference genome, and upon comparison they were able to identify localised expansions, contractions, and inversions in the *Oryza* species genomes compared to *O. sativa*. Transposable elements (TEs) were discovered to be enriched in locations specific to *O. sativa*: long terminal repeats (LTRs) were randomly distributed across the chromosomes. Furthermore, single-copy genes associated to environmental protection mechanisms were overrepresented in rice-expanded areas.

Li *et al.* (2012) identified the genome-wide variations among three elite restorer lines of hybrid-rice to. While the genetic variations among these lines were smaller than those seen in the landrace population, they were higher than predicted, implying a complex genetic underpinning for the restorer lines' phenotypic variability.

Wei *et al.* (2016) have identified genome variations between sesame landrace and a variety. Two typical sesame landrace accessions, Baizhima and Mishuozhima, were chosen and re-sequenced and compared to variety Zhongzhi13. The comparison showed that in the coding areas of genes, there are 70,018 SNPs and 8,311 InDels. Agronomic factors including blooming time, plant height, and oil content are contributed/influenced by the variations. The chromosomal variants discovered were

effectively exploited in QTL mapping. The black pigment production gene, *PPO*, was identified as the candidate gene for sesame seed coat colour.

When compared to its parents, Golden Delicious and Indo, the apple cultivar Su Shuai has better disease resistance, shorter internodes, and a milder fruit flavour. To understand the variation in the genome between them, Zhang *et al.* (2014) compared to the genomes of Golden Delicious, Su Shuai and Indo. In the 'Indo' and 'Su Shuai' genomes, a vast range of genetic differences were discovered, comprising 2,454,406 and 18,749,349 SNPs and 59,547 and 50,143 structural variants, respectively. Seventeen genes connected to disease resistance, 10 genes related to gibberellin (GA), and 19 genes related to fruit taste were found among the structural variants in 'Su Shuai.'

In an investigation by Tabidze *et al.* (2017), genomes of four Georgian grape cultivars which are highly valuable for the wine industry, Chkhaveri, Saperavi, Meskhetian Green and Rkatsiteli were compared. Annotation has revealed 17,409, 17,021, 18,355 and 13,960 genes in Chkhaveri, Saperavi Meskhetian Green and Rkatsiteli, respectively. Further analysis revealed four new terpen synthase genes. Two of these have been discovered as probable full-length proteins, germacrene A synthase and (-) germacrene D synthase.

MATERIALS AND METHODS

3. MATERIALS AND METHODS

The present investigation entitled ‘Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types’ was undertaken at the Department of Plant Biotechnology, College of Agriculture, Thrissur during 2019-2021. The objective of this study was to identify the differential genes and genomic regions among the tall and dwarf coconut genotypes, through comparative whole-genome sequence analyses and to develop molecular markers for distinguishing the tall and dwarf coconut types. The materials utilised and the methods followed are outlined below.

3.1. Materials

3.1.1. System requirements - hardware

A high-performance cluster (HPC) consisting of a master node and three compute nodes was used to carry out the computational analysis.

Master node

DELL PowerEdge R740xd Rack Server

Processor : 2x Intel Xenon silver 4116 processor (@ 2.10GHz × 48)

: Model 85 Stepping 4

Hard Disk : 4.7 TB

RAM : 128 GB

Compute node

DELL PowerEdge R640 Rack Server

Processor : 2x Intel Xenon silver 4116 processor (@ 2.10GHz × 48)

: Model 85 Stepping 4

Hard Disk : 256 GB

RAM : 128 GB

Operating system: CentOS 7.5

Management Embedded/ At-the-Server: iDRAC9 (Integrated Dell Remote Access Controller 9)

Firmware version: 4.00.00

3.1.2. Software requirements*

The following software were used: SRA toolkit, ABySS, Velvet optimiser, SOAP2 denovo, RepeatModeller, RepeatMasker, AUGUSTUS, NCBI BLAST+, BRAKER2, GMATo and OmicsBox

3.1.3. Scripts used

The following scripts were used: Perl, awk, grep, HomeBrew and Cat

*individual software may require installation of dependency software, scripts and libraries.

3.2. Databases used

3.2.1. NCBI Assembly

The National Center for Biotechnology Information (NCBI) Assembly database contains information on assembled genome structure, assembly names and other meta-data, statistical reports, and connections to genomic sequencing data.

3.2.2. NCBI Sequence Read Archive (SRA)

SRA database of NCBI houses the raw sequencing data (read data) from the next-generation sequencing (NGS) systems. The database has several accessions including, SRR (run accession for actual sequencing data for the particular experiment), SRX (experiment accession representing the metadata for study, sample, library, and

runs), SRP (study accession representing the metadata for sequencing study and project abstract), SAMN/SRS (BioSample/SRA accession representing the metadata for biological sample).

3.2.3. Genome Warehouse (GWH)

GWH is a public repository under the National Genomics Data Center (NGDC), which is part of the China National Center for Bioinformation (CNCB) that houses genomic-scale data for a variety of species and provides a set of online services for submitting, storing, releasing, and sharing genome data.

3.2.4. Raw data and genome assembly retrieval

An exhaustive database survey was carried out. The assemblies of three coconut genomes (mention the acc. no, variety and tall/dwarf for each) and an unassembled genome (Laguna Tall, SRX1333617) were retrieved from NCBI SRA and the genome assemblies of Catigan Green Dwarf (GCA_006176705.1), Hainan Tall (GCA_008124465.1) and Chowghat Green Dwarf (GCA_003604295.1) from NCBI Assembly (**Fig. 3.1**), respectively. Two assemblies from the National Genomics Data Center (NGDC) Cn.tall (CNT) (GWHBEBT000000000) and Cn.dwarf (CND) (GWHBEBU000000000) submitted as of September 2021, were also used for the research later on. This was followed by analysing the quality of the assemblies using QUAST.

3.3. SRA Toolkit

The NCBI SRA Toolkit is a set of utilities for faster download, view and search a large volume of high-throughput sequencing data from SRA. The source code for SRA Toolkit ver. 2.9.1 was downloaded from GitHub (<https://github.com/ncbi/sra-tools/wiki/Downloads>). The binaries were compiled to get the executable programme. The fastq-dump tool of the program was used to convert the raw data from SRA to FASTQ format.

```
# download file: prefetch will download and save SRA file
related to SRR accession in
```

```

# the current directory under newly created SRA accession
directory
$ prefetch SRRxxxxxx # for a single file
$ prefetch SRRxxxxxx SRRxxxxxx # multiple files
# convert to FASTQ: fastq-dump will convert SRRxxxxxx.sra
to SRRxxxxxx.fastq
$ fastq-dump SRRxxxxxx # single file
$ fastq-dump SRRxxxxxx SRRxxxxxx # multiple files
# for paired-end data use --split-files (fastq-dump) and -
S or --split-files (fasterq-dump) option
$ fastq-dump --split-files SRRxxxxxx
$ fasterq-dump -S SRRxxxxxx
# download alignment files (SAM)
# make sure the corresponding accession has an alignment
file at SRA database
$ sam-dump --output-file SRRxxxxxx.sam SRRxxxxxx
To ensure the successful download and to validate the
downloaded SRA data integrity,
# download FASTQ file
$ prefetch SRRxxxxxx
# fastq-dump SRRxxxxxx
# check integrity of downloaded SRRxxxxxx.fastq file
# output from vdb-validate should report 'ok' and
'consistent' for all parameters
# Note: make sure you have .sra (not .cache) file for
corresponding accession in
# sra accession directory
$ vdb-validate SRRxxxxxx

```

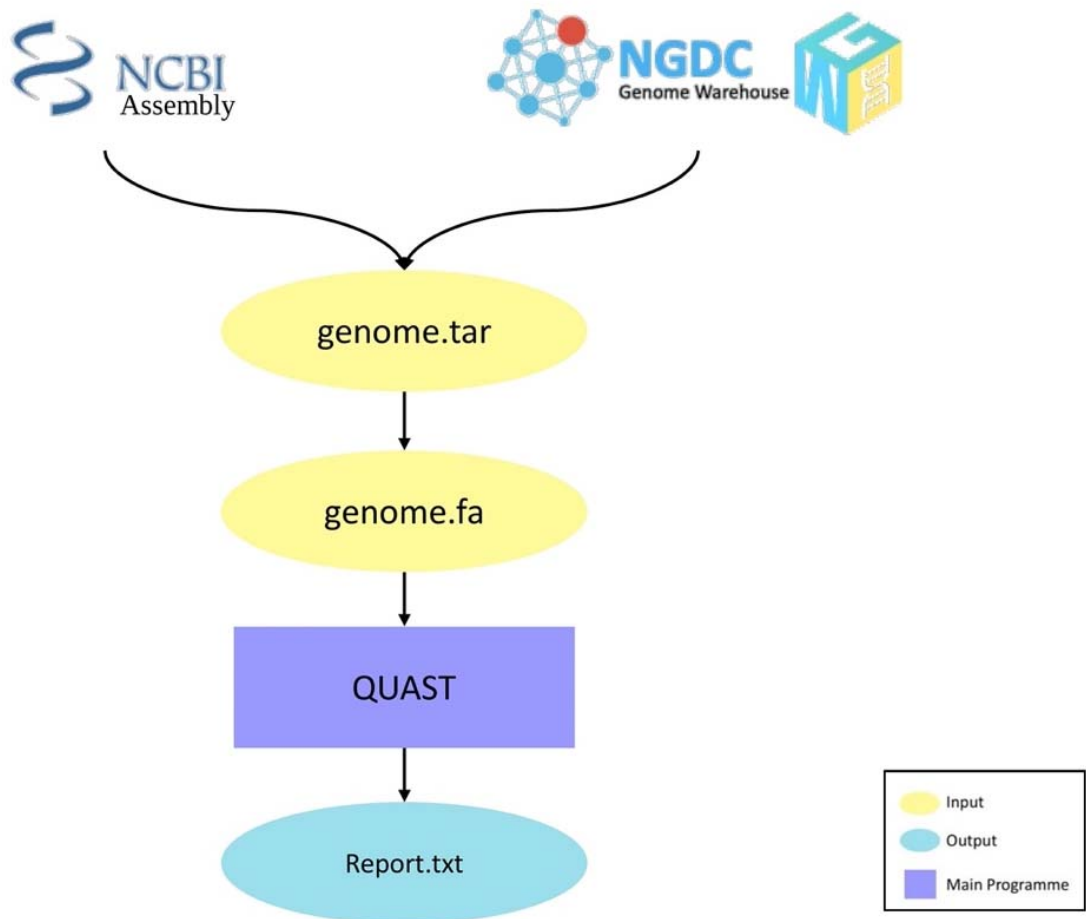


Fig. 3.1 Workflow displaying genome assembly retrieval and quality analysis using QUAST.

3.4. Genome Assembly

Since Laguna Tall assembly was unavailable, raw data were retrieved from SRA and the whole genome short reads from Illumina HiSeq 2000 were assembled using ABySS, SOAPdenovo2 and VELVET optimizer (**Fig. 3.2**).

3.4.1. ABySS

ABySS is a *de novo* sequence assembler that works with short paired-end reads and genomes of various sizes (Simpson *et al.*, 2009). The software was installed using the Homebrew package manager using the command,

```
#after successful installation of all the dependencies
$brew install abyss
```

ABySS was executed specifying the k-mer value path to the raw reads and the output file name using the command,

```
$abyss -pe k=kmer_value name=run_name in ='file1 file2'
#where pe is for paired-end sequencing
#name corresponds to the name of the output file
#file1 file2 corresponds to the path to the raw reads
```

3.4.2. SOAPdenovo2

SOAPdenovo2 is a novel short-read assembly method that can build a *de novo* draft assembly for large genomes. The program is specially designed to assemble Illumina GA short reads (Luo *et al.*, 2012).

After every dependency software and scripts were installed, the source code was downloaded (<https://github.com/aquaskyline/SOAPdenovo2>), unpacked to the destination folder and compiled by using GNU make with command

```
"make" at ${destination folder}/SOAPdenovo-V2.04.
```

Initially, the configuration file for the run was made. Since the raw reads were multiple fasta files generated from multiple libraries, the configuration file instructed

the assembler where to look for these files and what information they include. The information about the library, as well as the information about the sequencing data provided by the library, were grouped into the appropriate library section. After the configuration file was made available, the assembler was run using the command,

```
{bin} all -s config_file -K 63 -R -o graph_prefix  
1>ass.log 2>ass.err  
#config_file is the path to the configuration file  
#ass.log is assembly log  
#ass.error is assembly error
```

3.4.3. VelvetOptimiser

The VelvetOptimiser is an assembly optimization wrapper script for the Velvet assembler (Zerbino and Birney, 2008). It finds the best hash value range inside a specified range, calculates expected coverage, and then finds the optimal coverage cut-off. It uses Velvet's internal approach for anticipating paired-end library insert lengths and optimises the assemblies using the default or a user-supplied criterion.

Since VelvetOptimiser is a wraparound script of Velvet, Velvet Assembler was also downloaded and compiled prior to VelvetOptimiser in order to run the programme,

```
#after successful installation of all the dependencies  
$brew install homebrew/science/velvet  
#to install velvet  
$make  
#make is done in the velvet directory to make the software  
from the source code  
$brew install homebrew/science/velvetoptimiser  
#to install velvetoptimiser  
#to run velvetoptimizer  
$ velvetoptimizer.pl -short --t -fastq file  
# were -short for short reads  
#t is the number of threads for the maximum number of  
simultaneous velvet instances to run
```

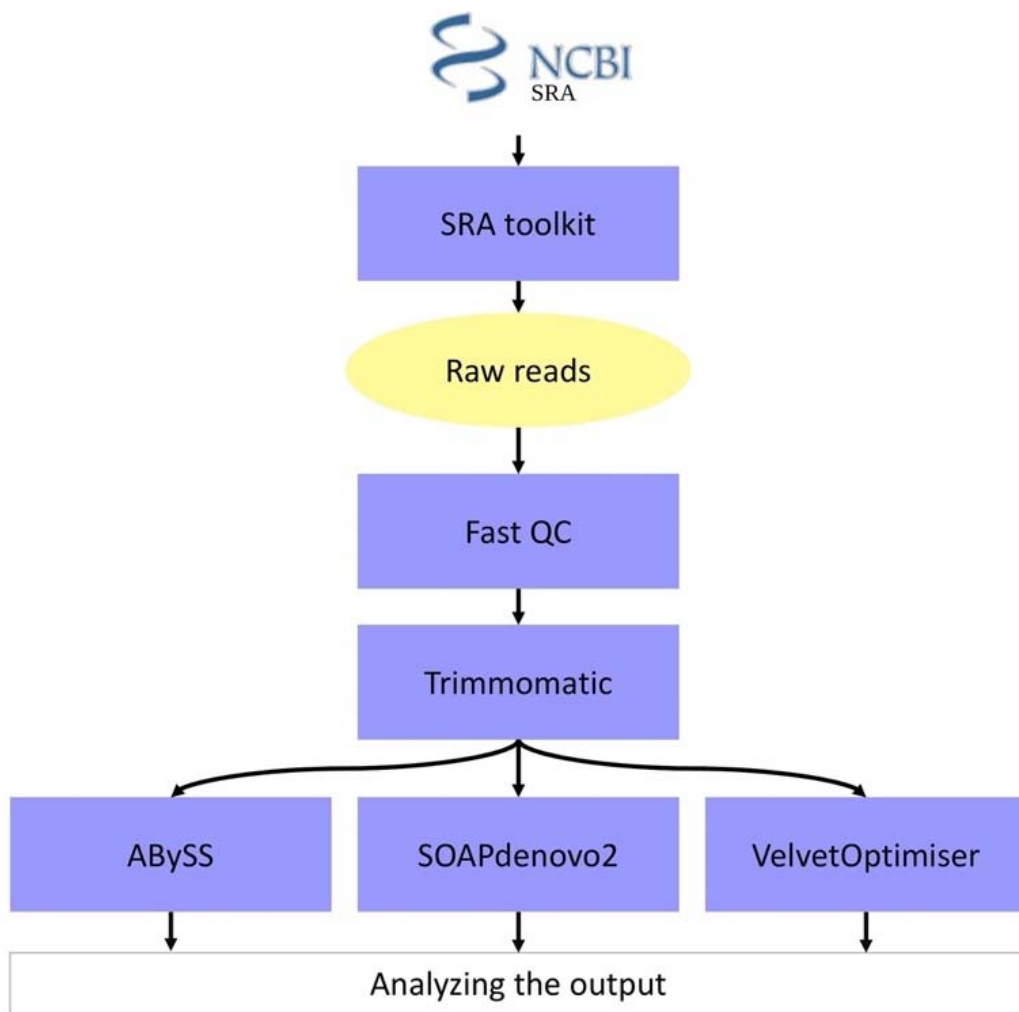



Fig. 3.2 Raw reads retrieval and assembly pipeline.

```
#fastq is the file format of the raw reads
#path to the raw reads file
```

3.5. Repeat modelling

RepeatModeler was used to prepare an exhaustive repeat library for coconut, through an integrative method. The *de novo* repeat identification by RepeatModeler is facilitated by two different discovery methods, RepeatScout (Price *et al.*, 2005) and RECON (Bao and Eddy, 2002), which is followed by consensus generation and classification (**Fig. 3.3**). The source code for RepeatModeler was obtained from <https://github.com/Dfam-consortium/RepeatModeler>, followed by decompressing and configuration. Modelling was performed on the three coconut genomes to get repeat libraries, and the repeat classification was carried out in these libraries using RepeatClassifier, a part of the RepeatModeler.

The core homology-based classification module of RepeatModeler (RepeatClassifier) compares TE families created by various *de novo* approaches to the RepeatMasker Repeat Protein Database (DB) and libraries. The Repeat Protein Database contains TE-derived coding sequences from a wide range of TE classes and species. By combining score and overlap filters, RepeatClassifier identifies and label the family using the RepeatMasker/Dfam classification scheme (Flynn *et al.*, 2020).

```
#after successful installation of all the dependencies
$git clone https://github.com/Dfam-consortium/RepeatModeler
#to get repeatmodeler
$ perl ./configure
#in the repeatmodeler directory to configure repeatmodeler
#Create a Database for RepeatModeler, choose your search engine and input file.
$./BuildDatabase -name -engine -dir
#Run RepeatModeler,select the database, choose your search engine, the number of threads
$./RepeatModeler -database -engine -pa
```

```
#if classification step is not carried out automatically,  
run RepeatClassifier  
$./RepeatClassifier -consensi -stockholm -engine -  
repeatmasker_dir
```

3.6. Combined repeat library construction

In order to create an exhaustive repeat library, the resultant repeat libraries were merged and the redundant sequences were removed (**Fig 3.4**).

```
$cat <file_1> <file_2> >outfile_name  
#to merge the libraries  
#file_1 and file_2 are the files to be merged
```

3.7. Repeat masking

RepeatMasker is a programme that scans for interspersed repetitions and low-complexity regions in DNA sequences (<https://www.repeatmasker.org>). The software generates a comprehensive annotation of the repetitions found in the query sequence, as well as a modified version of the query sequence that masks all of the annotated repeats. RepeatMasker uses one of several prominent search engines to compare sequences, including nhmmer, cross match, ABblast/WUblast, RMBlast, and Decypher, and uses curated repeat libraries such as Dfam (a profile HMM library built from Repbase sequences) and Repbase (by the Genetic Information Research Institute). RepeatMasker was obtained by cloning the master branch of the RepeatMasker repository and configuring the software. This paragraph may be shifted to Review of Literature chapter.

3.7.1 Development of a RepeatMasker library for coconut

Initially, repeat masking was performed using the combined Dfam and RepBase library and repeat masking was performed (**Fig 3.5**). Then the classified consensus sequences from tall and dwarf ecotype genomes were used as the repeat library and repeat masking was attempted in the corresponding genomes. Subsequently, the resultant repeat library files of the genomes were merged. The non-redundant library

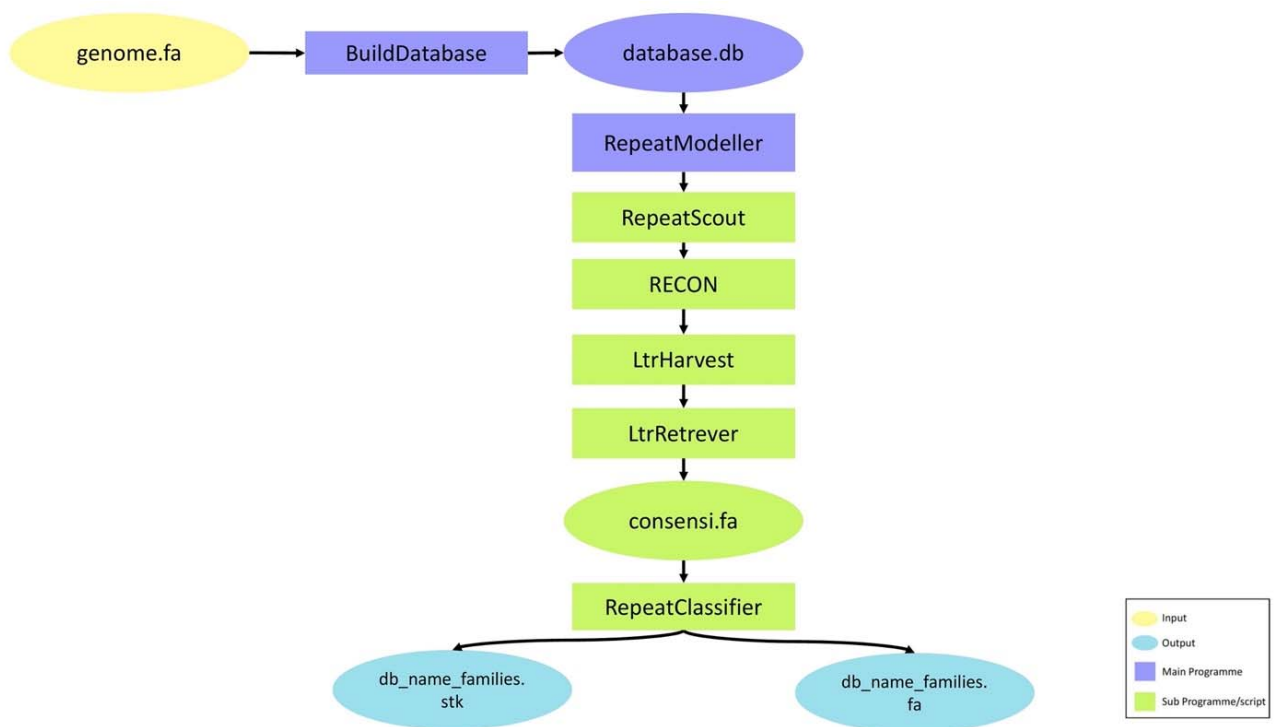


Fig. 3.3 RepeatModeler pipeline

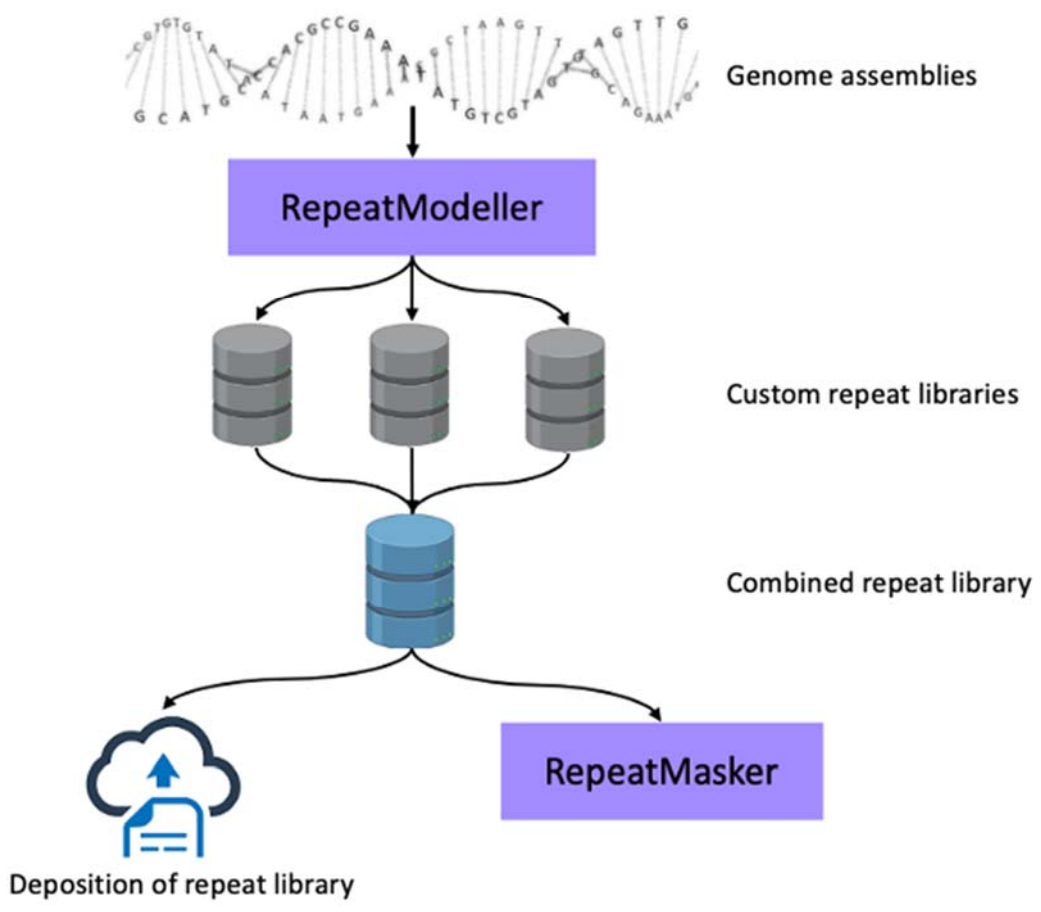


Fig 3.4 Workflow depicting the combined repeat library development.

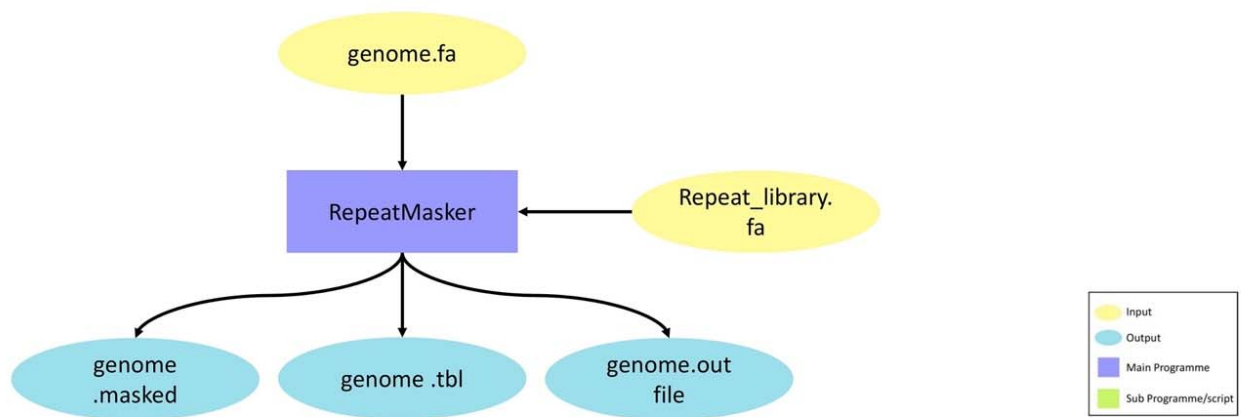


Fig. 3.5 RepeatMasker pipeline

thus generated was used to mask repeats in the genomes using RepeatMasker with the `-s` (sensitive) option, which allowed the masking of low complexity DNA sequences, simple repeats and transposable elements.

To install the software:

```
#after successful installation of all the dependencies
$ gitclone https://github.com/rmhubble/RepeatMasker.git
#to obtain the software package
$perl ./configure
#in the repeatmodeller directory to configure repeatmodeller
To run RepeatMasker
$./RepeatMasker -species name of specie -s -lib -dir
out_file path path_to genomefile.fna
```

3.7.2 Validation of repeat library

To check if the repeat library developed is effective for use in other palm crops, genome assemblies of date palm (PDK50-GCA_000181215.3), oil palm (EO8-GCA_000441515.1) and sago palm (GCA_017589505.1) were retrieved and subjected to repeat masking in RepeatMasker using the *de novo* developed tall, dwarf as well as the combined consensus sequence libraries, with same parameters mentioned above.

3.8. Gene Prediction

3.8.1. BRAKER2

BRAKER2 is a follow-up to BRAKER1 that enables completely automated training of the gene prediction tools GeneMark-EX and AUGUSTUS (Brůna *et al.*, 2021). Braker2 pipeline was used to obtain a training set for AUGUSTUS. After installation and configuration of the software (<https://github.com/Gaius-AUGUSTUS/BRAKER>) and its dependencies, the BRAKER analysis was performed using the genome assembly file and protein sequence as input files, specifying species name, the number of cores and the path to configuration files since BRAKER script

invokes various other software (**Fig 3.6**). The commands used in the analysis are given below,

```
./braker.pl --genome=path_to_genome_file --
species=species_name --prot_seq=path_to_prot_seq --cores
(number_of_cores) --skip_fixing_broken_genes --gff3 --
AUGUSTUS_CONFIG_PATH=path_to_config_file --
AUGUSTUS_BIN_PATH=path_to_AUGUSTUS_bin_file --
GENEMRK_PATH=path_to_genemark_file --
PROTHINT_PATH=path_to_prothint_file
```

BRAKER2 pipeline was performed to obtain training set for performing gene prediction using AUGUSTUS. The masked sequence was given as input along with the protein data from a closely related organism and the output of braker is used as the training set for AUGUSTUS.

3.8.2. AUGUSTUS

AUGUSTUS was used to carry out the gene prediction. AUGUSTUS is a eukaryotic gene-prediction tool that is based on the Hidden Markov Model incorporating a series of established methodologies and sub-models (Stanke & Waack, 2003; Stanke *et al.*, 2008).

AUGUSTUS training set obtained from BRAKER2 analysis was copied to the binary (bin) folder. AUGUSTUS was invoked by giving the species name along with the query sequence (which is the repeat masked sequence). The species name corresponds to the training set for running AUGUSTUS, which should be present in the binary folder.

```
./AUGUSTUS [parameters] --species=SPECIES query_file --
protein=on --introns=on --codingseq=on --gff3=on --
outfile=file_name --progress=true

#where species corresponds to the name of training set
```

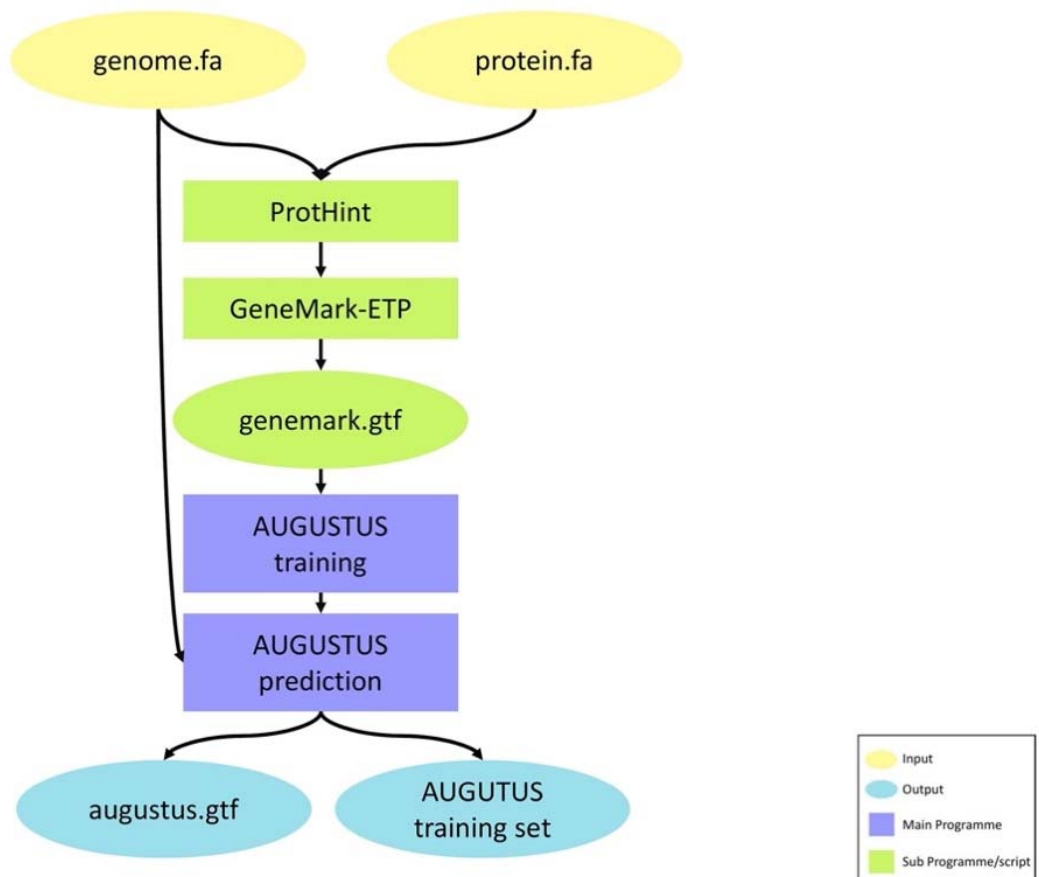



Fig 3.6 BRAKER2 pipeline

```
#query_file corresponds to the path to the query file
```

The resultant output is augustus.gtf file, which is further converted to protein (augustus.aa) and coding sequence (augustus.codingseq) using the script GetAnnoFasta.pl.

```
$/getAnnoFast.pl -outputfile  
#where output file corresponds to the path to the output  
file of AUGUSTUS
```

The script made a fasta file with protein sequences (AUGUSTUS.aa) and one with coding sequences (AUGUSTUS.codingseq) from the sequences provided in the comments of the AUGUSTUS output. These sequence comments were turned on with --protein=on and --codingseq=on, respectively. The AUGUSTUS pipeline is presented in **Fig. 3.7**.

Number of genes in the output was identified using the command

```
$grep -c "^>" file_name  
#where file chosen is either AUGUSTUS.aa or  
AUGUSTUS.codingseq
```

3.9. Stand-alone BLAST/ BLAST+

Basic Local Alignment Search Tool (BLAST) programme searches for areas of local similarity between protein or nucleotide sequences. The software compares nucleotide or protein sequences to database sequences and estimates their statistical significance (Altschul *et al.*, 1990). BLAST+ is a set of command-line tools provided by NCBI, to perform BLAST searches on servers with no constraints on size, volume, or database (Christian *et al.*, 2009). BLAST+ (standalone BLAST) was used to compare the tall and dwarf coconut genome assemblies. The HT amino acid sequence was taken as query and BLASTp analysis was carried out against the amino acid sequence of CAGD which was taken as database and vice versa. BLASTp analysis was carried out with HT amino acid sequence against the amino acid sequence of CGD taken as database and vice versa. Then the unique sequences were identified (which correspond

to '0 hits found' in the blast result) and the corresponding sequence was extracted from the coding sequence file.

The following commands were used for performing BLAST+

To create a custom database from a sequence FASTA file,

```
./makeblastdb -in mydb.fa -dbtype nucl -out
```

To perform protein BLAST,

```
./blastp -db prot -query nt.fa -out results.out
```

#where prot corresponds to the name of the protein database, nt.fa is the path to the query file and results.out is the name of the output file.

From the BLAST results, those showing zero hits were noted and corresponding sequences isolated from the coding sequences. To identify and filter the zero hits from the BLAST output, following command was used,

```
$awk -v N=2 -v pattern=" *pattern" '{i=(1+(i%N)); if  
(buffer[i]&& $0 ~ pattern) print buffer [i]; buffer[i]=$0;}'  
file >output_file.txt
```

#where N corresponds to the value to the Nth line before the pattern to print.

#*pattern is the regex to search (0 hits found)

#buffer is an array of N elements. It is used to store the lines. Each time the pattern is found, the Nth line before the pattern is printed.

#file corresponds to the path of the file from which the match is to be extracted

#output_file.txt is the name of the output file

To extract the sequence from the coding sequence file the following command was used

```
./seqkit grep -r -f input.txt sequence.fasta -o  
output_file.fasta
```

#where input.txt is the file that contains the list of sequence headers to be extracted,

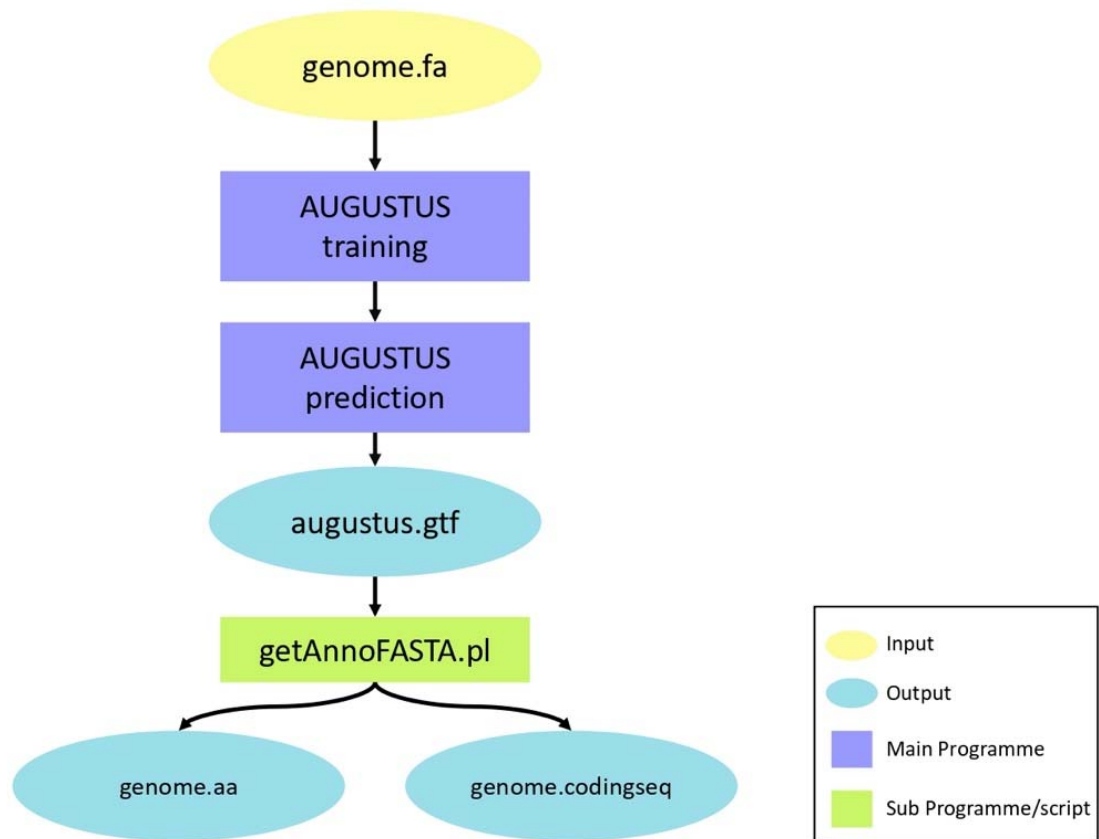


Fig 3.7 AUGUSTUS pipeline

#sequence.fasta is the file from which the sequence is to be extracted

#output_.fa is the name of the output file

3.10. Online BLAST

Sequences thus extracted were further identified by BLAST search against the non-redundant nucleotide database at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.

BLAST configuration

While configuring the BLAST search, *Viridiplantae* was selected from the taxonomy filter, the number of BLAST hits were set to 5 and the E-value was specified to $1.0E^{-5}$.

3.11. Gene Ontology (GO) and InterProScan

The Blast2GO annotation methodology in OmicsBox was used to perform a functional annotation (Götz *et al.*, 2008). The sequence was imported to the omics box and the online resource Blast2GO (<http://www.blast2go.com/b2ghome>) was used to assign GO terms. All the genes were analysed by performing BLASTx search against the NCBI non-redundant (nr) database with an Expect (E) value $\geq 1.0E^{-3}$ and a maximum of 5 hits for each gene (following the BLAST configuration given below). In the mapping step, default weights of the evidence codes were used. In the annotation step, only the gene hits with an E value $\geq 1.0E^{-6}$ were further analyzed (the filtering of annotations was done following the annotation configuration given below). After assigning a GO-Weight of 5 to mapped children words, an annotation score of 55 was utilised as the cut-off value. This was followed by an InterProScan search for conserved domains and motifs, which was followed by the use of the Annex function to augment the GO terms. The workflow is presented in **Fig. 3.8**.

Annotation configuration

1. Annotation cut-off (threshold): The annotation rule selects the lowest term per branch that lies over this threshold. Annotation cut-off value of 55 was chosen.

2. GO-Weight: This is the weight given to the contribution of mapped children terms to the annotation of a parent term. A GO-Weight of 5 was selected.
3. E-Value-Hit-Filter: an E-Value of $1.0E^{-6}$ was chosen.

3.12. Designing PCR primers for identifying the markers

Primer designing was carried out using Primer3 (<https://bioinfo.ut.ee/primer3-0.4.0/>, Untergasser *et al.*, 2012). The sequence was given as input and the product size range was specified to 300-500 bp. The primer sequences were validated using PCR primer stats (https://www.bioinformatics.org/sms2/pcr_primer_stats.html).

3.13. Microsatellite identification

Genome-wide Microsatellite Analyzing Tool or GMATo was used for identification of microsatellites. The perl based programme formats DNA sequences first, then segments lengthy DNA sequences into Mb-sized parts for processing. All microsatellite motifs are produced using Perl meta-characters and regular expression patterns. The pattern matching function in Perl searches each DNA segment for all motifs. SSR loci data are generated at each segment and at each chromosome after merging data from segments (Wang *et al.*, 2013). The preconfigured software was downloaded from (<https://sourceforge.net/projects/gmato/files/?source=navbar>) and was unzipped in the required directory. The programme takes the genome file in fasta format as input along with the parameters for filtering and searching microsatellites.

```
#to run the programme
$perl gmat.pl [parameters] -i /path_to_Sequence
#parameters:
-r: minimum repeated times of motif
-m: minimum length of motif
-x maximum length of motif
-i file in fasta
```

GMATo generates three files: a formatting report which summarise the input sequence (s), a file with SSR loci information, and a file with SSR statistical distribution. The SSR loci file contains the input sequence ID, length, microsatellite

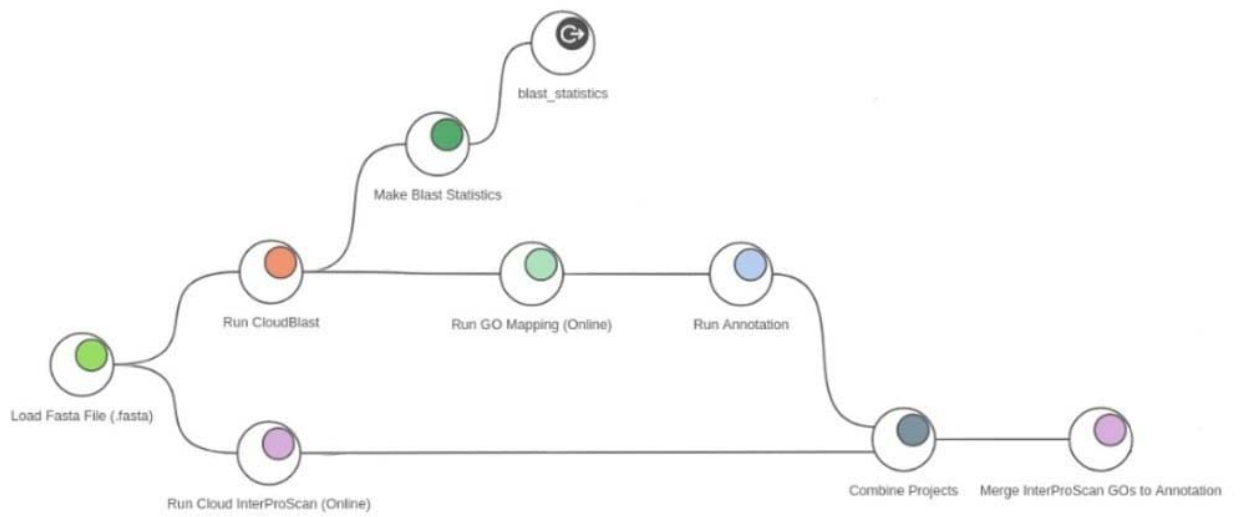


Fig. 3.8 OmicsBox functional annotation pipeline

start and end positions, repeated times, and motif sequence. The statistical distribution file contains data for four genomic classes. Each classification ends with a comprehensive summary. Classification I is the motif length statistics, presenting type, abundance in rank order. Classification II is motif statistics based on sequence composition, ranked occurrence. Classification III provides information on grouped complimentary motifs, such as TC/GA, and ranks their prevalence. The total occurrence of motif(s) and SSR frequency (loci/Mb) at each chromosome or super-scaffold are provided in Classification IV.

3.14. Collection of samples for analysis of the markers

Coconut leaf samples were collected from Regional Agricultural Research Station, Kerala Agricultural University, Pilicode. Ten parental lines, including five tall and five dwarf ecotypes (Table 3.1) were used in the marker analysis and validation. Spindle leaves were collected from the selected trees.

Table 3.1. Parental lines selected for leaf sample collection

Sl. No.	Cultivar	Tree number and block in RARS farm	
		Tree No.	Block
Tall ecotypes			
1.	Philippians Tall	134	G
2.	West Coast Tall	242	G
3.	Jawan Giant	61	G
4.	Kappadam	108	G
5.	New Guinea	174	G
Dwarf ecotype			
6.	Chowghat Orange Dwarf	66	J
7.	Chowghat Green Dwarf	61	J
8.	Chowghat Yellow Dwarf	35	J
9.	Malaysian Orange Dwarf	80	J

3.15. Isolation of total genomic DNA

The genomic DNA isolation was carried out by following a modified Cetyltrimethylammonium bromide (CTAB) method. DNA was extracted from all ten cultivars.

Modified protocol

- Midrib of the frozen leaf samples was removed and the young leaves were cut into small pieces and ground into a fine powder using liquid nitrogen, in autoclaved micro-pestle
- Contents were transferred to a 2.0 ml micro-centrifuge tube and 1.0 ml extraction buffer was added. The contents were mixed well and incubated at 65 °C for one hour with occasional mixing by gentle swirling.
- After incubation, the contents were centrifuged at 8,000 rpm for 5 minutes, 750 µl of supernatant was transferred to fresh 1.5 ml microcentrifuge tube and the remaining was discarded
- To the supernatant, 750 µl of chloroform: isoamyl alcohol (24:1) was added, contents were mixed thoroughly and centrifuged at 13,000 rpm for 10 minutes
- Aqueous phase was extracted and transferred to a fresh 1.5 ml micro-centrifuge tube, equal volume of chloroform: isoamyl alcohol (24:1) was added and contents were centrifuged for 10 minutes at 13,000 rpm
- This step was repeated twice
- The aqueous phase was extracted and transferred to a fresh 1.5 ml micro-centrifuge, equal volume of isopropanol was added, contents mixed by gentle inversion and incubated overnight at -20 °C
- After overnight incubation, the tubes were centrifuged for 10 minutes at 10,000 rpm and the supernatant was decanted
- DNA pellet was washed with 50 µl of 70 per cent ethanol and contents were centrifuged for 5 minutes at 8,000 rpm
- Solution was discarded and the pellet was air-dried.

- The air-dried pellet was dissolved in TE buffer (40-50 μ l), tubes were labelled and stored at -20 $^{\circ}$ C.

RNase treatment

The DNA was purified by following the steps,

- Two μ l RNase A solution (10 mg/ml) per 50 μ l TE was added to the DNA samples, followed by one hour incubation of the tubes at 37 $^{\circ}$ C in a water bath
- After one hour, to facilitate the denaturation of RNase A, the incubation temperature was increased to 65 $^{\circ}$ C for 10-15 minutes
- What was added to the tubes? Mention the centrifugation
- From the centrifuged tubes, the aqueous phase was extracted and transferred to a fresh 1.5 ml microcentrifuge tube. An equal volume of chloroform: isoamyl alcohol (24:1) was added and contents were centrifuged for 10 minutes at 13,000 rpm.
- Supernatant was extracted and transferred to a fresh sterile 1.5 ml microcentrifuge tube and an equal volume of isopropanol was added and content was mixed by gentle inversion and tubes were kept at -20 $^{\circ}$ C for two hours.
- Please check if the above para is a repetition of the previous one
- After incubation, the tubes were centrifuged for 10 minutes at 10,000 rpm and the supernatant was decanted
 - DNA pellet was washed with 50 μ l of 70 per cent ethanol and contents were centrifuged for 5 minutes at 8,000 rpm
- The air-dried pellet was dissolved in TE buffer (40-50 μ l), the tubes were labelled and stored at -20 $^{\circ}$ C.

3.16. Quantification of DNA

The quantity of DNA in each sample was determined using a NanoDrop spectrophotometer (ND-1000 v3.5.2, Nano Drop Technologies Inc., USA) by recording the absorbance at 260 and 280 nm. Following procedure was followed,

- The device was initialised using autoclaved distilled water
- 2.0 μ l TE buffer was used to set the blank value

- 1.0 µl DNA sample was loaded onto the pedestal to determine the amount of DNA
- For each sample, the amount of DNA in ng/l and the OD value was recorded.

Purity of the DNA samples was estimated using the ratio of readings at 260 and 280 nm (OD_{260}/OD_{280}). The 260 nm/ 280 nm OD ratio of pure DNA preparations is between 1.7 and 2.0. (Sambrook and Russel, 2001). The DNA samples were diluted to working concentrations of 100 ng/l using computed concentrations.

3.17. DNA quality check by agarose gel electrophoresis

The following procedure was followed

- Gel casting tray was cleaned with 70 per cent ethanol, kept in the casting tank and comb was placed parallel to the open sides of the tray
- To make 0.8% gel, 1.2 g agarose was dissolved by melting in 150 ml 1X TAE buffer.
- Agarose solution was left to cool down and when the temperature reached nearly 55 °C, 7.5µl of ethidium bromide was added (staining agent). Then the agarose solution was poured into the casting tray and allowed to solidify
- Gel transferred to the electrophoresis unit, with the wells facing the cathode. Gel tank was filled with 1X TAE buffer just enough to cover the surface of the gel.
- The DNA samples were mixed with 6X gel loading dye and loaded in individual wells.
- Electrophoresis was carried out at 80 volts for 45 minutes until the dye migrates to the end of the gel. The gel was documented and the DNA was visualized using a gel documentation system.

The intactness of the gel image, the clarity of the band, and the presence of contaminants such as proteins and RNA were checked.

3.18. PCR amplification

One µl each of the DNA template, along with forward and reverse primer pairs and the reaction mixture, was used in 0.2 ml tubes for PCR amplification (Table 3.2). After a quick spin, thermal cycling was started.

Table 3.2. Composition of PCR reaction mixture

Reagents	Volume (μl)
Taq assay buffer (10X)	2.0
dNTPs (2.5 mM)	1.5
Forward primer (10 pM)	1.0
Reverse primer (10 pM)	1.0
Taq DNA Polymerase (3U/ μ l)	0.3
Template (DNA 100 ng/ μ l)	1.0
Sterile distilled water	13.2
Total	20.0

[dNTPs, assay buffer and *Taq* DNA polymerase - GeNei Laboratories (India), primers -Sigma Aldrich (India)].

PCR program

Thermal cycling programme is presented in Table 3.3.

Table 3.3. Programme of thermal cycling

Sl. No.	Reaction step	Temperature ($^{\circ}$C)	Time
1	Initial denaturation	94.0	4 min
2	Denaturation	94.0	0:45 min
3	Annealing*	56.0-64.0	1:45 min
4	Primer extension	72.0	2 min
5	Repeat	36 cycles	
6	Final extension	72.0	10 min
7	Hold	4.0	∞

*Annealing temperature optimised for each primer

3.19. Agarose gel electrophoresis

Electrophoresis was carried out to resolve the PCR products on a 2.0 per cent agarose gel. Three μl of 6X bromophenol blue dye was added to the samples. The dye mixed DNA samples were loaded in the wells and electrophoresed at 60 V until the dye has diffused to the end of the gel. After electrophoresis, the gel was visualized and documented using a gel documentation system (BioRad XR⁺).

RESULTS

4. Results

The study entitled ‘Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types’ was undertaken at the Department of Plant Biotechnology, College of Agriculture, Kerala Agricultural University, Thrissur, India, during 2019-2021. The results obtained from the study are presented below.

4.1 Database survey

An exhaustive database survey was carried out and three coconut genome assemblies for Catigan Green Dwarf (GCA_006176705.1), Hainan Tall (GCA_008124465.1) and Chowghat Green Dwarf (GCA_003604295.1) were identified from the NCBI Assembly database. The raw sequence data of Laguna Tall (SRX1333617) was identified from the NCBI SRA database. The coconut mitochondrial, as well as chloroplast genomes, transcriptome profiles from leaves, inflorescences, and fruits, were also identified.

Subsequently, two genome assemblies Cn.tall (CNT) (GWHBEBT000000000) and Cn.dwarf (CND) (GWHBEBU000000000) submitted as of September 2021, were identified from NGDC (give full form). The details on genome assemblies and raw reads used in this study are presented in Table 4.1.

4.2 Data Retrieval

4.2.1 Raw data retrieval

SRA toolkit was used to obtain the raw reads of Laguna Tall. The source code of the software was first obtained from the repository and after installation of the dependency software, the source code was compiled to get an executable programme. The fastq-dump script was used to download SRA fastq files, which saved them to the working directory by default. During the download, a temporary directory was created in the path which was erased on completion of the download. Since the raw reads of Laguna Tall were paired-end, --split-files flag was given to obtain two files.

Note: The pre-fetch phase was unnecessary with fastq-dump and fasterq-dump. Using fastq-dump with no initial use of pre-fetch was slower than using pre-fetch and then fastq-dump.

4.2.2 Genome assembly retrieval

The genome assemblies were downloaded from the Assembly database of NCBI. The file was downloaded as genome_assemblies.tar. The resulted folder was named "genome_assemblies", which contained

- a report.txt file providing a summary of the downloaded file
- a folder named with the date of the download as "ncbi-genomes-YYYY-MM-DD", and contained:
 - README.txt file
 - md5checksums.txt file
 - data files with names such as *_genomic.fna.gz, where the assembly accession is the first part of the name, followed by the assembly name

Among the five genome assemblies, CGD and CAGD have been sequenced using the Illumina PacBio hybrid sequencing approach while HT genome was sequenced by Illumina HiSeq2000 sequencing technology. The CNT and CND genomes have been sequenced using Nanopore sequencing technology. Among the five coconut genome assemblies the CAGD and CGD assemblies were scaffold level assemblies, whereas HT, CNT and CND assemblies were chromosome level assemblies. The CGD had 7,998 contigs whereas CAGD had 59,328. HT, CNT and CND genomes had 16 chromosomes.

Further, the quality of the assembly was checked using QUASt. Results of the analysis are presented in Table 4.2. Out of the five genome assemblies three (HT, CNT and CND) are chromosomal level assembly while two (CAGD and CGD) are scaffold level assemblies. Comparing the N50 value among the chromosomal level assemblies the CNT and CND genome assemblies were found to be of better quality than HT. Comparing the number of contigs and N50 value among the scaffold level assemblies the quality was found to be better in CAGD than CGD genome assembly.

Table 4.1 Coconut genome assemblies and raw reads identified

Assembly characteristics	UPLB_dcnu_1.0	ASM812446v1	ASM360429v1	SRX1333617	CnT01	CnD01
Database	NCBI	NCBI	NCBI	NCBI	NGDC	NGDC
Type	Whole Genome Assembly	Whole Genome Assembly	Whole Genome Assembly	Raw reads	Whole Genome Assembly	Whole Genome Assembly
Organism name	Cocos nucifera (coconut palm)	Cocos nucifera (coconut palm)	Cocos nucifera (coconut palm)	Cocos nucifera (coconut palm)	Cocos nucifera (coconut palm)	Cocos nucifera (coconut palm)
Infraspecific name	Catigan Green Dwarf	Hainan Tall coconut	Chowghat Green Dwarf	Laguna Tall		
Ecotype	Dwarf	Tall	Dwarf	Tall	Tall	Dwarf
BioSample	SAMN09748030	SAMN06328965	SAMN07738886	SAMN04159271	SAMC393072	SAMC393073
BioProject	PRJNA483845	PRJNA374600	PRJNA413280	PRJNA298457	PRJCA005463	PRJCA005463
Submitter	The Coconut Genomics Program - Philippine Genome Center	Hainan Key Laboratory of Tropical Oil Crops Biology /Coconut Research Institute	CPCRI	Philippine Genome Center, University of the Philippines	National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University	National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University
Date	2019/06/10	2019/08/29	2018/10/01	2014/06/18	2021/09/30	2021/09/30
Assembly level	Scaffold	Chromosome	Scaffold	TruSeq synthetic long reads	Chromosome	Chromosome
Genome representation	Full	Full	Full	n/a	Full	Full
RefSeq category	Representative genome	Representative genome	Representative genome	n/a	n/a	n/a
GenBank assembly accession	GCA_006176705.1 (latest)	GCA_008124465.1 (latest)	GCA_003604295.1 (latest)	n/a	n/a	n/a
RefSeq assembly accession	n/a	n/a	n/a	n/a	n/a	n/a
RefSeq and GenBank assembly identical	n/a	n/a	n/a	n/a	n/a	n/a
WGS Project	QRFJ01	VOII01	PDMH01	n/a	n/a	n/a
Assembly method	SPARSE v. JUN-2016; DBG2OLC v. JUN-2016; HiRise v. MAY-2017	SOAP v. 2.2	SOAPdenovo v. 2.04	n/a	SMRTdenovo v1.0	SMRTdenovo v1.0
Expected final version	no	yes	yes	n/a	yes	yes
Genome coverage	50.0x	173.17x	52.48x	n/a	116x	104x
Sequencing technology	Illumina MiSeq; PacBio RSII; HiRise pipeline	Illumina HiSeq2000	Illumina; PacBio	Illumina HiSeq 2000	Nanopore	Nanopore

Table 4.2 QUAST result Quality characteristics of the assembly

Statistics without reference	UPLB_dcnu_1.0	ASM360429v1	ASM812446v1	CnT01	CnD01
Cultivar	Catigan Green Dwarf	Chowghat Green Dwarf	Hainan Tall	Tall	Dwarf
# contigs/chromosomes	7,998	59,328	16	16	16
# contigs (≥ 0 bp)	7,998	59,328	16	16	16
# contigs (≥ 1000 bp)	7,998	59,328	16	16	16
# contigs (≥ 5000 bp)	7,970	36,181	16	16	16
# contigs (≥ 10000 bp)	7,869	30,498	16	16	16
# contigs (≥ 25000 bp)	7,365	19,667	16	16	16
# contigs (≥ 50000 bp)	6,261	11,004	16	16	16
Largest contig/chromosome	87,79,653	8,26,246	11,99,63,539	21,44,44,701	21,78,08,934
Total length	2,10,24,17,611	1,83,91,72,334	2,20,23,88,255	2,39,41,38,060	2,39,95,73,174
Total length (≥ 0 bp)	2,10,24,17,611	1,83,91,72,334	2,20,24,55,121	2,39,41,38,060	2,39,95,73,174
Total length (≥ 1000 bp)	2,10,24,17,611	1,83,91,72,334	2,16,84,08,380	2,39,41,38,060	2,39,95,73,174
Total length (≥ 5000 bp)	2,10,23,18,704	1,79,74,31,649	2,07,61,46,081	2,39,41,38,060	2,39,95,73,174
Total length (≥ 10000 bp)	2,10,15,70,891	1,75,56,11,079	2,05,55,58,760	2,39,41,38,060	2,39,95,73,174
Total length (≥ 25000 bp)	2,09,23,00,676	1,57,50,47,929	2,00,67,01,319	2,39,41,38,060	2,39,95,73,174
Total length (≥ 50000 bp)	2,05,13,08,570	1,26,58,66,185	1,92,42,34,421	2,39,41,38,060	2,39,95,73,174
N50	5,70,487	85,564	12,17,559	17,19,31,881	17,16,39,811
N75	2,29,523	40,236	1,56,844	14,50,30,504	14,53,71,288
L50	771	5,682	59	7	7
L75	2,286	13,542	1,659	11	11
GC (%)	38	37	37	37	38
Mismatches					
# N's	59,45,968	27,23,02,417	5,52,48,642	2,41,700	38,500
# N's per 100 kbp	283	14,806	2,509	10	2

4.3. Genome assembly

As the whole genome assembly for Laguna Tall was unavailable, genome assembly was performed with the raw Illumina HiSeq 2000 reads retrieved from SRA database. Assembly was carried out using three assemblers, ABySS, SOAPdenovo2 and VELVETOptimizer. The output from all the three assemblers were analysed and it was found that the assembly accounted for only half of the genome, which meant that the assembly was incomplete. So, the sequence was not considered for further evaluation and the study was carried out with five genome assemblies available.

4.4 Repeat modelling

Repeat modelling was carried out in the genome assemblies using RepeatModeler which runs multiple algorithms in a given a genomic database, grouping redundant results, refine and categorise the families, and provide a high-quality library of TE families appropriate for use with RepeatMasker. The genome file was used as an input to create a database for RepeatModeler and repeat modelling was carried out for the database provided. The output of the RepeatModeler was a classified consensus sequence in FASTA format (species.consensi.classified.fna) and a seed alignment in Stockholm format (species.consensi.classified.stk). The classified consensus sequence was the repeat library which was used for RepeatMasking. Repeat modelling was carried out for the genome assemblies of HT, CGD and CAGD. The classified consensus sequences of the three runs were merged to get a combined repeat library. The curated Stockholm file was sent to Dfam, which is the largest open Transposable Elements database.

4.5 Repeat Masking

RepeatMasker was used to mask the interspersed repeats and low-complexity regions in the coconut genome assembly. The genome file and the repeat library was given as input to the RepeatMasker. The output of RepeatMasker was a masked genome (genome.masked.fna) file along with masking report (genome.masked.txt). Initially RepeatMasker was executed with Dfam+RepBase as library. Analysis with RepeatMasker employing the conventional Dfam+RepBase library identified and masked only 4.60, 4.91 and 1.04 % of the repeats in HT, CAGD and CGD (Table 4.3).

Hence, Dfam and RepBase libraries were found inefficient for masking the repeats in the coconut genome.

Hence, the repeat masking was carried out with the *de novo* repeat libraries for tall and dwarf (using RepeatModeler). With that 79, 81.17 and 65.05 % repeats were masked (Table 4.4). In order to further improve the masking percentage, repeat masking was carried out with the combined repeat library and it was found that using this combined library, 80.52, 82.19, 66.18, 83.55 and 84.00 % repetitive elements were identified in HT, CAGD, CGD CNT and CND coconut genome assemblies, respectively. Masking with the combined *de novo* library generated from the *de novo* libraries of tall and dwarf ecotypes was more efficient. Genomes of ‘Hainan Tall’, ‘Chowghat Green Dwarf’ and ‘Catigan Green Dwarf’ have shown enhanced masking by 1.52, 1.13 and 1.02 %, respectively. Of the 80.52 % repeats identified in HT, 62.89% belonged to LTR elements, whereas SINEs, LINEs, DNA elements, small RNA, satellites, simple repeats and low complexity sequences accounted for 0.02, 0.76, 4.76, 0.02, 0.03, 0.47 and 0.08 %, respectively (Table 4.5).

In CAGD genome, 82.19 % was comprised of repetitive elements, where LTR elements accounted for majority of the repeats (64.93 %). SINES, LINEs, DNA elements, small RNA, satellites, simple repeats and low complexity sequences accounted for 0.02, 0.75, 4.63, 0.02, 0.02, 0.6 and 0.08 %, respectively (Table 4.5).

CGD had 66.19 % repetitive elements, comprising LTRs, SINEs, LINEs, DNA elements, small RNA, satellites, simple repeats and low complexity sequences at 46.89, 0.02, 0.86, 5.03, 0.02, 0.03, 0.54 and 0.1 %, respectively (Table 4.5). Unclassified repeats accounted for a sizable fraction in all genomes, amounting to 12.02, 11.76 and 2.92 % in HT, CAGD and CGD, respectively.

The CNT genome assembly consisted of 84.61 % repetitive sequences, of which 68.55 % was LTR elements and 10.97 % unclassified repeats. SINEs, LINEs, simple repeats and DNA elements made up 0.02, 0.70, 0.54 and 4.27 %, respectively of the masked repeats (Table 4.6).

CND genome assembly displayed a similar masking result as CNT, with overall masking percentage of 84.89 %. More than half of the repeats (68.92 %) were LTRs

and the SINEs and LINEs accounted for 0.02 and 0.69 %, respectively while simple repeats and DNA elements made up to 0.48 and 4.26 % of the repeats. Similar to other genome assemblies, unclassified repeats accounted for a good fraction of 10.97 %.

4.6 Validation of repeat library


To analyse the feasibility of using the new combined library in masking the repetitive elements in the related genomes of Arecaceae, RepeatMasker was executed separately with repeat library of tall, dwarf and combined repeat library. The dwarf repeat library was able to mask 31.99, 40.62 and 15.35 %, respectively in date palm, oil palm and sago palm genome assemblies (Table 4.7) whereas the tall repeat library masked 31.38, 39.02 and 14.86 % respectively in date palm, oil palm and sago palm genome assemblies (Table 4.8). Using the combined library, the masking percentage was found to be 34.21, 42.48 and 16.59, respectively in date palm, oil palm and sago palm. Compared with the individual tall and dwarf libraries, the combined library had masked 3.46, 2.83 and 1.73 % more repeats, respectively in oil palm, date palm and sago palm (Table 4.9).

Since the combined repeat library developed was found to be more efficient than the currently available libraries and showed better masking percentage comparing to the original articles, we made the repeat libraries publicly available at <http://www.kau.in/repeat-libraries> (Plate 4.1).

Table 4.3 Repeat masking with Dfam+RepBase

	‘Hainan Tall’			‘Chowghat Green Dwarf’			‘Catigan Green Dwarf’		
	GCA_008124465.1_ASM812446v1			GCA_003604295.1_ASM360429v1			GCA_006176705.1_UPLB_denu_1.0		
Repeat class/family	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
Retroelements	84313	38453421 bp	1.75	47004	19090771 bp	1.04	78353	39080217 bp	1.86
SINEs	0	0	0	0	0	0	0	0	0
LINEs	0	0	0	0	0	0	0	0	0
CRE/SLACS	0	0	0	0	0	0	0	0	0
L2/CR1/Rex	0	0	0	0	0	0	0	0	0
L1/CIN4	0	0	0	0	0	0	0	0	0
LTR elements	84313	38453421 bp	1.75	47004	19090771 bp	1.04	78353	39080217 bp	1.86
Ty1/Copia	84313	38453421 bp	1.75	47004	19090771 bp	1.04	78353	39080217 bp	1.86
Gypsy/DIRS1	0	0	0	0	0	0	0	0	0
Retroviral	0	0	0	0	0	0	0	0	0
DNA transposons	0	0	0	0	0	0	0	0	0
En-Spm	0	0	0	0	0	0	0	0	0
MuDR-IS905	0	0	0	0	0	0	0	0	0
PiggyBac	0	0	0	0	0	0	0	0	0
Tourist/Harbinger	0	0	0	0	0	0	0	0	0
Rolling-circles	0	0	0	0	0	0	0	0	0
Unclassified	1	48 bp	0	5	264 bp	0	3	155 bp	0
Interspersed repeats		38453469 bp	1.75		19091035 bp	1.04		39080372 bp	1.86
Small RNA	1562	337291 bp	0.02	1033	100647 bp	0.01	1627	301443 bp	0.01
Satellites	0	0	0	0	0	0	0	0	0
Simple repeats	924079	46717679 bp	2.12	0	0	0	902536	48664348 bp	2.31
Low complexity	278445	15724714 bp	0.71	0	0	0	269160	15224915 bp	0.72
	Bases masked: 101233153 bp (4.60 %)			Bases masked: 19191682 bp (1.04 %)			Bases masked: 103271078 bp (4.91 %)		

കേരള കാർഷിക സർവ്വകലാശാല | ENGLISH Current Style: Standard


 **Kerala Agricultural University**
കേരള കാർഷിക സർവ്വകലാശാല Search

Home About KAU Education Research Outreach Library Services Resources Notice Board Ranking

Home » Resources

Repeat Libraries for Genome Analysis

Repeat Library for Genome Analysis of Coconut and Related Palms 📄 🖨

ATTACHMENT	SIZE
 DOWNLOAD (ZIP)	1.99 MB


KAU Main Websites

- College of Agriculture, Padannakkad
- College of Agriculture, Vellanikkara
- College of Agriculture, Vellayani
- College of Agriculture, Wayanad
- College of Climate Change and Environmental Science, Vellanikkara
- College of Co-operation, Banking & Management, Vellanikkara
- College of Forestry, Vellanikkara
- Institute of Agriculture Technology & RARS, Pattambi
- Kelappaji College of Agricultural Engineering & Technology, Tavanur


Switch Language


English | മലയാളം

Translations

 **Select Language**

Follow Us





Address

Kerala Agricultural University
KAU Main Campus
KAU P.O., Vellanikkara
Thrissur Kerala 680656
☎️ :+91-487-2438011
☎️ :+91-487-2438050
☎️ :+91-487-2370019

[Contact Us](#) | [Credits](#) | [Legal Notices](#)

Plate 4.1 Repeat library available at KAU webpage (<http://www.kau.in/repeat-libraries>)

Table 4.4 Repeat masking with individual libraries in tall and dwarf coconut ecotypes

Repeat class/family	'Hainan Tall'			'Chowghat Green Dwarf'			'Catigan Green Dwarf'		
	GCA_008124465.1_ASM812446v1			GCA_003604295.1_ASM360429v1			GCA_006176705.1_UPLB_denu_1.0		
	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
SINEs	3320	437647	0.02	3404	467454	0.03	3404	467888	0.02
ALUs	0	0	0	0	0	0	0	0	0
MIRs	0	0	0	0	0	0	0	0	0
LINEs	22623	11569901	0.53	30098	16407325	0.89	32953	18648779	0.89
LINE1	13688	9580238	0.43	16631	10536609	0.57	16307	10481805	0.5
LINE2	1472	359414	0.02	0	0	0	0	0	0
L3/CR1	0	0	0	0	0	0	0	0	0
LTR elements	931343	1334065267	60.57	938745	834787779	45.39	796650	1317926881	62.69
ERVL	0	0	0	0	0	0	0	0	0
ERVL-MaLRs	0	0	0	0	0	0	0	0	0
ERV_classI	24769	9943323	0.45	25540	12340922	0.67	25195	12136589	0.58
ERV_classII	0	0	0	0	0	0	0	0	0
DNA elements	169465	94595435	4.29	173108	83919732	4.56	161649	92230715	4.39
hAT-Charlie	0	0	0	0	0	0	0	0	0
TcMar-Tigger	0	0	0	0	0	0	0	0	0
Unclassified	830853	279328181	12.68	875740	259206695	14.09	834916	277725997	13.21
Small RNA	1713	357961	0.02	1396	280898	0.02	1417	282327	0.01
Satellites	6297	1831983	0.08	1206	261447	0.01	1186	258418	0.01
Simple repeats	251201	12143379	0.55	227867	10068335	0.55	234720	12356726	0.59
Low complexity	40418	2156850	0.1	38137	2005094	0.11	38508	2014117	0.1
	Bases masked: 1739899931 bp (79.00 %)			Bases masked: 1196380131 bp (65.05 %)			Bases masked: 1706438421 bp (81.17 %)		

Table 4.5 Repeat masking with combined *de novo* library in HT, CGD and CAGD

Repeat class/family	‘Hainan Tall’			‘Chowghat Green Dwarf’			‘Catigan Green Dwarf’		
	GCA 008124465.1 ASM812446v1			GCA 003604295.1 ASM360429v1			GCA 006176705.1 UPLB dcnu 1.0		
	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
SINEs	3477	462403	0.02	3327	444471	0.02	3331	445978	0.02
ALUs	0	0	0	0	0	0	0	0	0
MIRs	0	0	0	0	0	0	0	0	0
LINEs	31092	16716123	0.76	28930	15745867	0.86	28322	15751825	0.75
LINE1	18606	12166077	0.55	16988	11451439	0.62	16661	11439142	0.54
LINE2	1369	279711	0.01	1297	271183	0.01	1248	259837	0.01
L3/CR1	0	0	0	0	0	0	0	0	0
LTR elements:	933552	1385032240	62.89	912048	862435593	46.89	695368	1365029020	64.93
ERVL	0	0	0	0	0	0	0	0	0
ERVL-MaLRs	0	0	0	0	0	0	0	0	0
DNA elements	168210	102784668	4.67	172368	92442269	5.03	148559	97238099	4.63
hAT-Charlie	0	0	0	0	0	0	0	0	0
TcMar-Tigger	0	0	0	0	0	0	0	0	0
Unclassified	775801	264808027	12.02	731769	237683936	12.92	691593	247332093	11.76
Small RNA	1556	332660	0.02	1490	323309	0.02	1500	318743	0.02
Satellites	2262	645373	0.03	2038	593962	0.03	1797	522372	0.02
Simple repeats	227628	10374537	0.47	220025	9968378	0.54	228019	12691652	0.6
Low complexity	34310	1780734	0.08	34187	1783974	0.1	34196	1770392	0.08
	Bases masked: 1773400308 bp (80.52 %)			Bases masked: 1217111579 bp (66.18 %)			Bases masked: 1728031097 bp (82.19 %)		

Table 4.6 Repeat masking with combined *de novo* library in CNT and CND

	Cnut Tall			Cnut Dwarf		
	GWHBEBT00000000			GWHBEBU00000000		
Repeat class/family	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
SINEs:	3522	465860	0.02	3511	465044	0.02
ALUs	0	0	0.00	0	0	0.00
MIRs	0	0	0.00	0	0	0.00
LINEs:	30594	16737286	0.70	30431	16656339	0.69
LINE1	18077	12192568	0.51	18037	12144479	0.51
LINE2	1394	288246	0.01	1392	286194	0.01
L3/CR1	0	0	0.00	0	0	0.00
LTR elements:	780904	1641176658	73.41	771903	1653802802	74.15
ERV1	0	0	0.00	0	0	0.00
ERV1-MaLRs	0	0	0.00	0	0	0.00
DNA elements:	158418	102135770	4.27	156934	102126251	4.26
hAT-Charlie	0	0	0.00	0	0	0.00
TcMar-Tigger	0	0	0.00	0	0	0.00
Unclassified:	747536	262526686	6.11	744970	263339238	5.74
Small RNA:	1569	332225	0.01	1570	335134	0.01
Satellites:	2099	579151	0.02	2087	574409	0.02
Simple repeats:	221588	12894937	0.54	224241	11425600	0.48
Low complexity:	34231	1771279	0.07	34084	1872692	0.08
	Bases masked: 2025741462 bp (84.61 %)				Bases masked: 2037070884 bp (84.89 %)	

Table 4.7 Repeat masking with dwarf library in related palms

Repeat class/family	Date Palm			Oil palm			Sago palm		
	PDK50- GCA_000181215.3.			EO8- GCA_000441515.1			GCA_017589505.1		
	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
SINEs	3947	545699	0.06	3329	458742	0.03	2099	274576	0.05
ALUs	0	0	0	0	0	0	0	0	0
MIRs	0	0	0	0	0	0	0	0	0
LINEs	29860	13864794	1.62	36672	18766193	1.34	19616	6838692	1.19
LINE1	24469	12452189	1.46	29796	16439574	1.17	2799	1108835	0.19
LINE2	0	0	0	0	0	0	0	0	0
L3/CR1	0	0	0	0	0	0	0	0	0
LTR elements:	243879	175297867	20.51	473584	357279313	25.47	63201	24308947	4.22
ERV1	0	0	0	0	0	0	0	0	0
ERV1-MaLRs	0	0	0	0	0	0	0	0	0
DNA elements	47309	14252425	1.67	96996	35151458	2.51	25981	6217679	1.08
hAT-Charlie	0	0	0	0	0	0	0	0	0
TcMar-Tigger	0	0	0	0	0	0	0	0	0
Unclassified	288184	56613503	6.62	522877	148106437	10.56	79394	13706036	2.38
Small RNA	1871	261550	0.03	1759	403619	0.03	757	122790	0.02
Satellites	2	117	0	1799	364810	0.03	1	100	0
Simple repeats	227038	10142979	1.19	193309	8383558	0.6	299224	18132077	3.15
Low complexity	51142	2977635	0.35	43760	2340737	0.17	98298	18907111	3.28
	Bases masked: 273378184 bp (31.99 %)			Bases masked: 569806281 bp (40.62 %)			Bases masked: 88491121 bp (15.35 %)		

Table 4.8 Repeat masking with tall library in related palms

	Date Palm			Oil palm			Sago palm		
	PDK50- GCA_000181215.3			EO8- GCA_000441515.1			GCA_017589505.1		
Repeat class/ family	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
SINEs	3674	490278	0.06	3065	409968	0.03	2042	259727	0.05
ALUs	0	0	0	0	0	0	0	0	0
MIRs	0	0	0	0	0	0	0	0	0
LINEs	26233	11891693	1.39	25516	12205972	0.87	18339	5756413	1
LINE1	16552	9493229	1.11	12675	9294686	0.66	2500	1001366	0.17
LINE2	26	2447	0	5954	1356108	0.1	9	1107	0
L3/CR1	0	0	0	0	0	0	0	0	0
LTR elements:	231866	173701211	20.32	477498	364162493	25.96	63627	23696905	4.11
ERV1	0	0	0	0	0	0	0	0	0
ERV1-MaLRs	0	0	0	0	0	0	0	0	0
DNA elements	47794	14863665	1.74	83610	33937494	2.42	20581	4572618	0.79
hAT-Charlie	0	0	0	0	0	0	0	0	0
TcMar-Tigger	0	0	0	0	0	0	0	0	0
Unclassified	256973	52757885	6.17	504995	124641827	8.89	72098	12548038	2.18
Small RNA	4507	1010437	0.12	1377	325728	0.02	2220	688384	0.12
Satellites	1	96	0	2658	439433	0.03	1	121	0
Simple repeats	232047	10738714	1.26	212389	9784271	0.7	303673	18734238	3.25
Low complexity	53033	3098120	0.36	49379	2636860	0.19	98351	19366546	3.36
	Bases masked: 268216692 bp (31.38 %)			Bases masked: 547312702 bp (39.02 %)			Bases masked: 85628204 bp (14.86 %)		

Table 4.9 Repeat masking with combined repeat library in related palms

Repeat class/family	Date Palm			Oil palm			Sago palm		
	PDK50- GCA 000181215.3			EO8- GCA 000441515.1			GCA 017589505.1		
	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence	Number of elements	Length occupied	Percentage of sequence
SINEs	4109	549478	0.06	3452	460635	0.03	2321	294692	0.05
ALUs	0	0	0	0	0	0	0	0	0
MIRs	0	0	0	0	0	0	0	0	0
LINEs	38578	16924216	1.98	45966	21621699	1.54	22955	7676383	1.33
LINE1	28166	14368119	1.68	30834	17726796	1.26	3828	1418342	0.25
LINE2	25	2387	0	5897	1131022	0.08	9	1107	0
L3/CR1	0	0	0	0	0	0	0	0	0
LTR elements:	262883	183208100	21.44	486094	378618559	26.99	73468	27396605	4.75
ERV1	0	0	0	0	0	0	0	0	0
ERV1-MaLRs	0	0	0	0	0	0	0	0	0
DNA elements	55808	17040342	1.99	99392	39433035	2.81	30857	6858323	1.19
hAT-Charlie	0	0	0	0	0	0	0	0	0
TcMar-Tigger	0	0	0	0	0	0	0	0	0
Unclassified	319728	62394275	7.3	548412	147662936	10.53	95975	16021243	2.78
Small RNA	2730	534890	0.06	1740	380213	0.03	2053	525453	0.09
Satellites	1	96	0	2184	353465	0.03	2	221	0
Simple repeats	218327	9906031	1.16	186287	8291758	0.59	296798	18053030	3.13
Low complexity	48858	2849467	0.33	40055	2135240	0.15	97481	18862421	3.27
	Bases masked: 292342418 bp (34.21 %)			Bases masked: 595868230 bp (42.48 %)			Bases masked: 95623116 bp (16.59 %)		

4.7 Gene finding

AUGUSTUS was executed for the coconut genome assemblies prior to repeat masking and the results showed 214050, 222803, 142443, 189021 and 195600 genes for HT, CAGD, CGD, Coconut tall and Coconut dwarf genome assemblies, respectively. The gene finding results after repeat masking with Dfam+RepBase as library yielded 115587, 211886 and 140962 genes in HT, CAGD and CGD coconut genome assemblies. AUGUSTUS was also used for the *de novo* prediction of the repeat-masked sequence and the results obtained are 31997, 31119, 29952, 31094 and 31296 genes for HT, CAGD, CGD, CNT and CND respectively. The entire gene finding results are summarized in the Table 4.10.

Gene prediction with AUGUSTUS resulted 114174, 106886 and 93552 genes, respectively in date palm, oil palm and sago palm, compared to 78890, 47076 and 87755 obtained using the combined repeat library (Table 4.11)

Table 4.10 Gene prediction results for coconut genome assemblies

Cultivar	Unmasked genome	Masked genomes	
		Dfam+RepBase	Combined repeat library
Hainan Tall	214050	115587	31997
Chowghat Dwarf	Green 142443	140962	29952
Catigan Green Dwarf	222803	211886	31119
Coconut Tall	189021	-	31094
Coconut Dwarf	195600	-	31296

Table 4.11 Gene prediction results for other palm genomes

Crop	Unmasked genome	Masked genome (combined repeat library)
Date palm	114174	78890
Oil palm	106886	47076
Sago palm	93552	87755

4.8 Comparative genome analysis using BLAST+

The unique sequences identified by comparing the Tall and Dwarf genomes and the corresponding sequence extracted from the coding sequence file are presented in Table 4.12. Reverse BLAST has further confirmed that the unique sequences identified from tall genome are present only in the tall genome and vice versa. The identified unique sequences are made available online at https://drive.google.com/drive/folders/1cw9-63MeNRAJODfYnIT3uOhH2f_dDyaN?usp=sharing.

Table 4.12 The number of unique sequences in each genome (BLAST+ output)

Comparison	Number of unique sequence
HT vs CAGD	90
CAGD vs HT	175
HT vs CGD	77
CGD vs HT	88

4.9 Online BLAST

To assign functions to the unique genes identified (Tables 4.12 and 4.1_), online BLAST was carried. BLASTn was done with unique sequence as query against the non-redundant protein database of NCBI. Results showed the functions of the unique sequences and the enzyme family or pathway. Most of the sequences did not return any hit or found to be hypothetical proteins. The results are presented in Annexure VI.

4.10 Gene Ontology and InterProScan

GO and InterProScan analyses for the unique sequences have confirmed the BLAST results.

4.11 Validation of the *in silico* results

The primers were designed for the unique genes identified from the initial comparison.

4.11.1 Primer designing

The sequences of the primers designed for the unique regions are presented in Annexure V.

4.11.2 DNA isolation from the leaves

Genomic DNA was isolated from the spear leaves of five each of dwarf and tall varieties, modified CTAB method, and quality of DNA was checked on 0.8 per cent agarose gel (Plate 4.2). Quality analysis using NanoDrop spectrophotometer (Table 4.13) has shown that samples are free of RNA, protein and phenol contamination.

Table 4.13 Quality of the DNA samples isolated

Code	Sample	Conc. (ng/μl)	OD ₂₆₀ /OD ₂₈₀
S1	Philippians Tall	970	1.97
S2	West Coast Tall	1030	1.84
S3	Jawan Giant	990	1.83
S4	Kappadam	840	1.81
S5	New Guinea	880	1.47
S6	Chowghat Green Dwarf	1007	1.80
S7	Chowghat Orange Dwarf	1011	1.86
S8	Chowghat Yellow Dwarf	970	1.78
S9	Malaysian Orange Dwarf	963	1.79
S10	Malaysian Yellow Dwarf	863	1.96

4.11.3 Validation of primers

Ten primers among the 27 used to screen the samples have shown amplification. The primers Cocos_1, Cocos_2, Cocos_4, Cocos_6, Cocos_7, Cocos_9, Cocos_18, Cocos_24, and Cocos_25 showed amplification in sample 2 (West Cost Tall). The primer Cocos_22 showed amplification in sample 2 (West Cost Tall) and sample 3 (Jawan Giant).

The primer Cocos_21 amplified the samples S2, S3, S5, and S6 which is West Coast Tall, Jawan Giant, New Guinea and Chowghat Green Dwarf respectively. Among the four samples amplified three samples namely West Coast Tall, Jawan Giant and New Guinea belongs to the tall ecotype while Chowghat Green Dwarf belongs to the dwarf ecotype. (Plate 4.3). The results are presented in Table 4.14

Table 4.14 Amplification pattern of candidate markers using the primers designed

Primer	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Cocos_1	-	+	-	-	-	-	-	-	-	-
Cocos_2	-	+	-	-	-	-	-	-	-	-
Cocos_3	-	-	-	-	-	-	-	-	-	-
Cocos_4	-	+	-	-	-	-	-	-	-	-
Cocos_5	-	-	-	-	-	-	-	-	-	-
Cocos_6	-	+	-	-	-	-	-	-	-	-
Cocos_7	-	+	-	-	-	-	-	-	-	-
Cocos_8	-	-	-	-	-	-	-	-	-	-
Cocos_9	-	+	-	-	-	-	-	-	-	-
Cocos_10	-	-	-	-	-	-	-	-	-	-
Cocos_11	-	-	-	-	-	-	-	-	-	-
Cocos_12	-	-	-	-	-	-	-	-	-	-

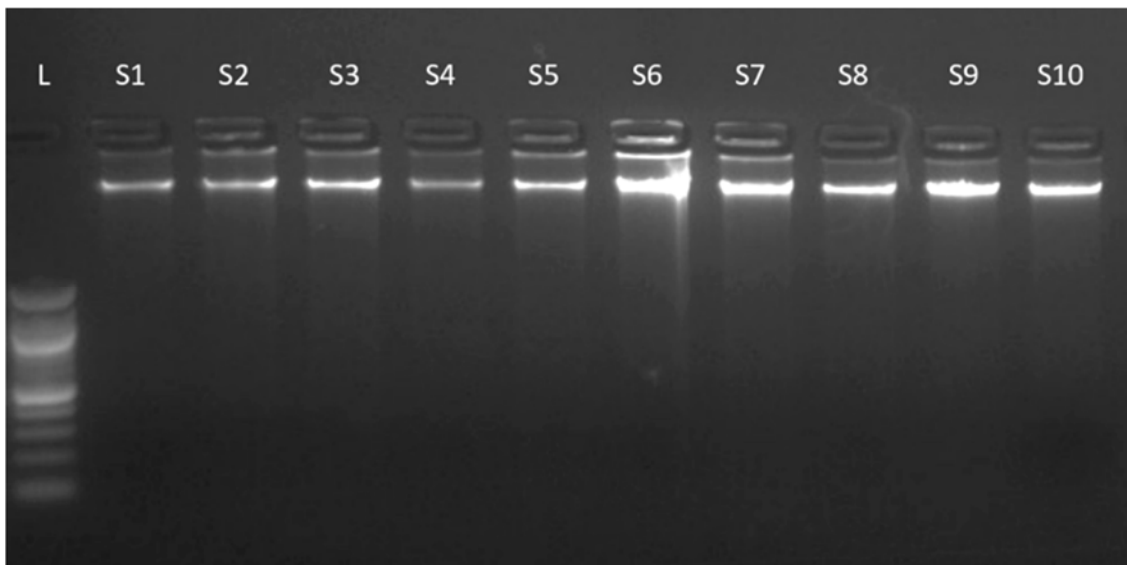


Plate 4.2 Good quality of genomic DNA seen in agarose gel.

(L: 100 bp ladder, S1-S10: DNA from coconut accessions)

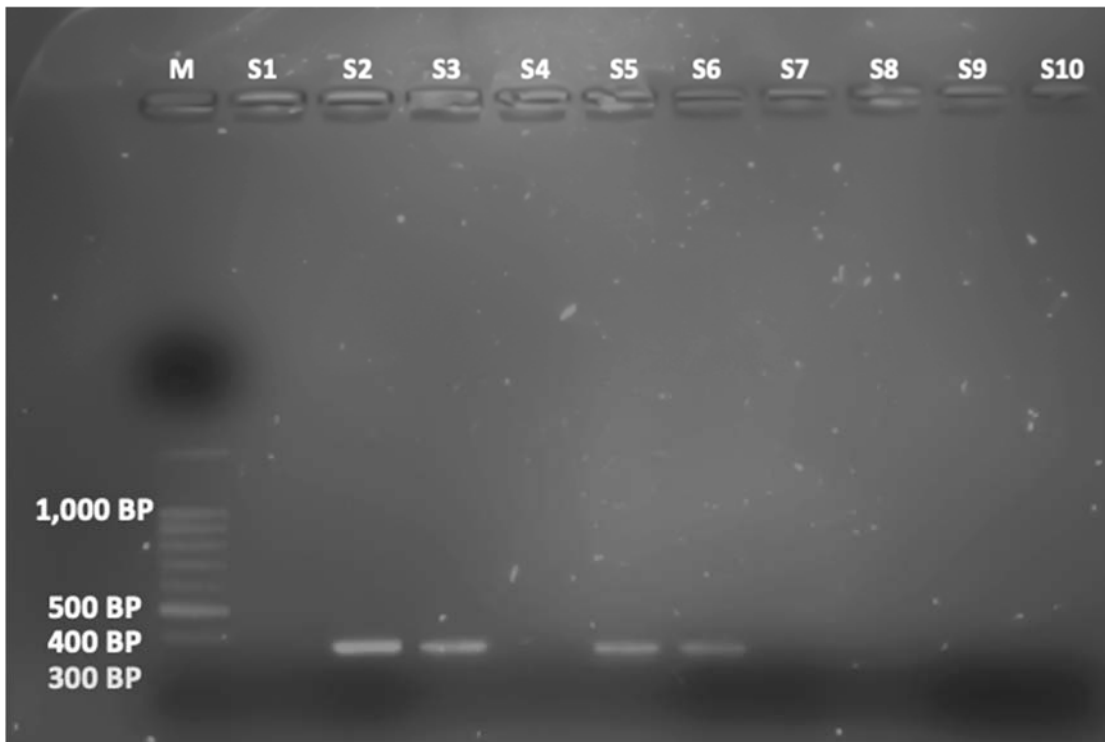


Plate 4.3 Amplification pattern observed using the primer Cocos_21 with the samples (M:100 bp ladder, S1 – S10 DNA from coconut accessions)

Cocos_13	-	-	-	-	-	-	-	-	-	-
Cocos_14	-	-	-	-	-	-	-	-	-	-
Cocos_15	-	-	-	-	-	-	-	-	-	-
Cocos_16	-	-	-	-	-	-	-	-	-	-
Cocos_17	-	-	-	-	-	-	-	-	-	-
Cocos_18	-	+	-	-	-	-	-	-	-	-
Cocos_19	-	-	-	-	-	-	-	-	-	-
Cocos_20	-	-	-	-	-	-	-	-	-	-
Cocos_21	-	+	+	-	+	+	-	-	-	-
Cocos_22	-	+	+	-	-	-	-	-	-	-
Cocos_23	-	-	-	-	-	-	-	-	-	-
Cocos_24	-	+	-	-	-	-	-	-	-	-
Cocos_25	-	+	-	-	-	-	-	-	-	-
Cocos_26	-	+	-	-	-	-	-	-	-	-
Cocos_27	-	-	-	-	-	-	-	-	-	-

4.12 Extended analysis with additional genomes

To search for more promising markers for the tall and dwarf growth of the palms, a second comparison was attempted, by additionally incorporating the CNT and CND genome assemblies reported as of September 2021.

The tall genome database was developed by individually performing BLASTp by taking the three dwarf coconut genome amino acid sequence as query against two tall coconut genome amino acid sequence. Check if this is what you meant. Similarly, The dwarf genome database was developed by individually performing BLASTp by taking the two tall coconut amino acid sequence as query against the dwarf genome

amino acid sequences. Further the unique sequences were identified (which correspond to ‘0 hits found’ in the blast result) and the corresponding sequence was extracted from the coding sequence file. From CAGD, 206 unique sequences were identified, which were absent in the tall genomes. From CGD and CNT, 106 and 139 unique sequences, respectively were identified. From HT and CNT, 141 and 154 unique sequences were identified (Table 4. 15). The unique sequences identified is made available online at https://drive.google.com/drive/folders/1cw9-63MeNRAJODfYnIT3uOhH2f_dDyaN?usp=sharing.

Table 4. 15 The number of unique sequences in each genome (BLAST+ output)

Genomes compared	Number of unique sequence
CAGD vs tall db	206
CGD vs tall db	106
CND vs tall db	139
HT vs dwarf db	141
CNT vs dwarf db	154

4.13 Microsatellite identification

Analysing the microsatellites identified using GMATo, it was found that the dinucleotide SSRs were most abundant class in all the genomes. A decrease in abundance was observed from dinucleotide to octanucleotide repeats. Details on the microsatellites are presented in Tables 4.16 and 4.17. From chromosomes 1 and 2, 37 and 23 potentially polymorphic SSRs, respectively were identified (Annexure VI).

Table 4.16 Microsatellite motif length and abundance

Motif (-mer)	CAGD	CGD	CND	CNT	HT
2	323105	236223	310447	350381	307588
3	37882	33003	41736	41148	37385
4	28957	14186	21279	30692	22528
5	8791	5022	10499	10714	7828
6	961	510	958	879	832
7	936	113	878	130	218
8	46	28	34	46	44
9	12	3	12	6	6
10	18	6	6	16	15

Table 4.17 Simple and compound motifs and abundance in the coconut genomes

	Simple motifs		Complimentary motifs	
	Number of motifs	Total number of occurrence	Number of motifs	Total number of occurrence
HT	773	376444	565	376444
CGD	746	289094	537	289094
CAGD	794	400708	574	400708
CNT	729	434012	533	434012
CND	804	385849	576	385849

DISCUSSION

5. DISCUSSION

Throughout the tropical parts of the world, coconut (*Cocos nucifera* L.) is a cash crop of subsistence. Kerala state of India is one among the leading coconut producers, and this crop plays a significant role in the state's economy and culture, by providing livelihoods for millions of small and marginal agricultural households. The crop also provides valuable service to a large number of people living in the vulnerable ecosystems of coastal and island locations. As a result, it is important to improve the yield of coconut. Hybridization followed by selection is the most practiced, yet most difficult breeding strategy in this crop. Prerequisites for successful coconut breeding include the idea on the genes and pathways governing important traits in coconut palms (Perera *et al.*, 2003).

Most coconut breeding programmes use either tall (*C. nucifera* L. var. *typica*) or dwarf (*C. nucifera* L. var. *nana*) ecotypes. This classification of coconut is based on their stature and breeding habits (Menon and Pandalai, 1958). Both the ecotypes differ in many characteristics, including height of the palms, shape, size, and color of the nuts, and the quality and yield of copra produced. In addition to the tall and dwarf ecotypes, intermediate types of palms have also been recorded. Due to the cross pollinating nature, tall palms exhibit a wide range of variation within a single variety whilst the self-pollinated dwarf palms are comparatively homogenous. It is hypothesized that the dwarf accessions originated from the inbreeding among tall accessions (Swaminathan and Nambiar, 1961).

Coconut being the state tree, Kerala is its largest producer, 6980.30 million nuts from an area of 0.76 million ha. But has a productivity of only 9,175 nuts/ ha in 2019-20, which is far lesser comparing to that of Andhra Pradesh (13,969), West Bengal (12,433) and Tamil Nadu (12,280) (CDB, 2021). Breeding of dwarf palms with higher yield, better copra and oil content and field tolerance to biotic and abiotic stresses shall boost the coconut production in Kerala.

Breeding attempts for dwarf palm stature are crippled with the non-availability of a precise methodology to identify the dwarf lines at the early plant growth stage

itself. Development of molecular markers linked with this trait will enable the marker assisted selection for dwarf palms with higher yield and other desirable traits. Previous attempts for marker development for plant growth habit in coconut were on a single gene basis. Since this is a quantitative trait, this kind of a marker cannot be recommended universally. Comparative whole genome analysis will be revealing large number of genes which are differentially present in tall and dwarf genotypes. Design based on this shall yield reliable and reproducible universal markers for quantitative traits, rather than single marker developed through conventional methods. Thus, this study has compared the whole genomes of dwarf and tall types of coconut to identify the differential genomic regions and to design many markers based on these regions.

5.1 Genome assembly

Flow cytometric studies to determine the genome size of coconut have reported a rough estimate of 2.72 Gbp (Gunn *et al.*, 2015). According to Neto *et al.* (2016), tall coconuts have a genome size of 2.733 Gbp and dwarfs have 2.723 Gbp. HT was the first coconut to be genome sequenced, with 2.20 Gbp size (Xiao *et al.*, 2017). The CAGD genome assembly is 2.1 Gbp while the size prediction by K-mer peak predicted 2.15 Gbp (Lantican *et al.*, 2019). In case of CGD, even though the flow cytometric analysis has predicted 2.59 Gbp, assembled genome was only 1.93 Gbp, 75% of the predicted genome size (Muliyar *et al.*, 2020).

Wang *et al.* (2021) has reported the reference grade assemblies for dwarf and tall genomes, where the genome assemblies of tall and dwarf were 2.40 and 2.39 Gbp, respectively. These values were comparable to the ~2.42 and ~2.44 Gb predicted using k-mer distribution analysis. Thus, by comparing the size of existing assemblies with those predicted by flow cytometry, it is found that HT assembly accounted 80.5 % of the genome, and similarly CAGD, CGD, CNT and CND assemblies accounting for 77.2, 70.9, 87.9, and 87.86 % of those predicted by flow cytometric analysis.

We have also performed genome assembly for Laguna Tall, with the raw sequence data available in the NCBI SRA database. It was identified that the assembled sequence length (879-900mb) was far less than that of the available genome assemblies.

The reason for this could be the shallow depth of sequencing and incomplete sequencing data made available. A prime reason for this difference in genome size may be because of the short read sequencing chemistry. The short-reads cannot span many of the lengthy repeat units', thus leading to a collapse of these regions during the assembly process. Further the chance of omission or fragmented repetitive regions containing TEs and tandem repeats are high due to reduced depth during sequencing. Thus due to these reasons Laguna Tall was not carried forward in the analysis.

5.2 Repeat Modelling

Sequence analysis in larger genomes is computationally demanding, as the coding sequences account for only a small proportion. Preparatory steps in the sequence analysis are mostly time consuming and repeat masking is one such process. Even though the draft genome assemblies such as *Triticum urartu* (Ling *et al.*, 2013, 2018), *Aegilops tauschii* (Jia *et al.*, 2013) and *Vigna mungo* (Pootakham *et al.*, 2021; Jagadeesan *et al.* 2021) have used custom made repeat libraries for repeat masking, the libraries are not made publicly available. Likewise, earlier genome analyses of coconut cultivars have used custom made repeat libraries, which were not made available in a public domain.

Using RepBase library, Mondal *et al.* (2018) were able to mask only 19.89 % of the repeats in *Oryza coarctata* genome whereas Bansal *et al.* (2020) have obtained 36.15 % masking in this species when *de novo* repeat library developed using RepeatModeler was employed. Custom repeat library developed by Shi *et al.* (2020) for *Oryza granulata* is reported to have masked 61.98 % of the repeats. Our studies have shown that the use of conventional or non-specific repeat library fails to mask the repeats, leading to erroneous gene finding results. Hence the need to develop a comprehensive repeat library was indispensable for coconut genome assemblies.

5.3 Repeat Masking and validation of repeat library

In coconut, Xiao *et al.* (2017) had identified and masked 74.48 % of repetitive elements in the genome assembly of 'Hainan Tall' which comprised of 2.64 % DNA elements, 0.87 % SINEs and 0.012 % LINEs. The combined library developed in this

study has enhanced the masking to 80.52 %, improving the identification of DNA elements, SINES and LINES to 4.67, 0.02 and 0.76 %, respectively.

Similarly, Lantican *et al.* (2019) have identified 78.33 % repetitive elements in the genome assembly of coconut cv. ‘Catigan Green Dwarf’, in which 60.26 % accounted for LTR elements. Compared to their results, our analysis showed 82.19 % overall masking, with 64.93 % LTR elements which shows the better efficiency of our newly developed repeat library.

The genome assembly of ‘Chowghat Green Dwarf’ (Mulyar *et al.*, 2020) had 77.29 % repeat masking, of which 58.85 % were LTRs. These results are comparable to the present results of 66.18 % overall masking, comprising 46.89 % LTRs.

Wang *et al.* (2021) has deposited the sequence of a tall and dwarf coconut cultivars, with 83.61 % masking in CNT and 83.83 % in CND. LTRs constituted much of the repeats accounting for 72.20 and 73.65 %, respectively. Even though SINES were present, their contribution was nearly zero, while LINES contributed 0.40 and 0.42 %, respectively. Our analysis reported an improved masking percentage of 84.61 and 84.89 % in CNT and CND, respectively. LTRs accounted for 73.41 and 74.5 % of the repeats. SINES contributed to 0.02 % percentage in both the genomes, while LINES contributed to 0.7 and 0.69 %. The combined library has been successful to mask a slightly more numbers of LTR elements, SINEs and LINES, compared those was reported.

Similar to other plant species, coconut genome was found to harbour considerable proportion of transposons. Class I retrotransposons were identified to be dominant, constituting more than half of the transposable elements identified in all the cases. LTRs were most abundant among the Class I retrotransposons.

Since the extent of masking or content of the repetitive elements in the palm genomes used in this study were not reported previously, a direct comparison of the masking efficiency was not possible and hence we have used the gene prediction results for a comparison. In oil palm, Singh *et al.* (2013) have identified 158,946 genes whereas our analysis has shown only 47,076 genes. On the contrary, number of genes in date palm reported by Al-Dous *et al.* (2011) was slightly lower than what we obtained,

suggesting that customising the libraries in each crop will be better. The draft genome report of sago palm is yet to be published.

5.4 Gene prediction

Since the plant genomes consist of a considerable amount of transposable elements, the creation of repeat libraries (TE libraries) followed by repeat masking is an important process. All the genomes displayed enormous number of genes during gene prediction without repeat masking. Whereas the repeat masked genomes displayed reasonable number of genes.

In the genome sequence assembly of HT, annotation had shown 28,039 protein coding genes (Xiao *et al.*, 2017) against a comparable number of 31,997 obtained in this study. Lantican *et al.* (2019) have identified 34,958 protein coding genes in the genome assembly of CAGD, compared to 29,952 genes obtained in this study.

From the draft genome of CGD, Mulyar *et al.* (2020) have annotated 51,953 coding genes while the results of this study showed 31,094 genes. Likewise Wang *et al.* (2021) have identified that the chromosome-level assemblies of CNT and CND harbour 29,897 and 28,111 protein coding genes, while this study has identified 31,094 and 31,296 genes, respectively. The slight difference in the gene numbers is due to the different analysis pipeline followed.

5.5 Identification and functional annotation of unique genes

Functional annotation of the unique genes was carried out using Online BLAST, InterProScan and GO. A similar strategy was followed by Bai *et al.*, 2013 for the comparative analysis of transcriptome for screening purple and white Nacre in Pearl Mussel, and Liu *et al.*, 2013 for the comparative analysis of the transcriptome of *Capsicum annuum* L. CMS line 121A and its near-isogenic restorer line 121C, where BLASTp search was performed to annotate the genes, which was followed by functional annotation by performing GO analysis.

While most of the sequences failed to give a BLAST hit, few of the BLAST hits were found to be either putative or uncharacterized or hypothetical genes. The

ambiguity in gene prediction accuracy still stands valid, since only 50 to 70 % of the genes in a genome can be predicted with reasonable certainty. The widely debated "70 percent hurdle" in protein prediction stays true (Brenner, 1999; Bork, 2000). These genes do not have a known homolog and are referred to as non-described/ unknown/ hypothetical, because they code protein or not, is unclear. Functional prediction becomes more challenging due to a lack of sequence similarity in the database for the gene/ protein of interest (Sivashankari and Shanmughavel, 2006).

5.6 Validation of markers

The easiest method for developing an inter-varietal primer is by designing the primers from a region of a genome for one variety and to empirically test the amplification in other varieties. Since the evolutionary history of the primer binding area across the species and varieties are unknown, forecasting the chance that any particular primer combination would operate in another species using this technique is impossible (Housley *et al.*, 2006). Twenty seven primers were designed from tall and dwarf unique genes and were validated by screening them with 10 coconut lines (five each of dwarf and tall). Among the twenty seven primers, ten primers displayed amplification for West Coast Tall. One primer amplified West Coast Tall and Jawan Giant. The primer Cocos_21 produced four amplifications where three were tall (West Coast Tall, Jawan Giant and New Guinea) and one was dwarf (Chowghat Green Dwarf). The absence of amplification in the two tall samples (Kappadam and Philippines tall) may be due to the presence of mismatches in the primer binding site. It might also be due to the fact that the marker represents a minor QTL. If the marker is representing a minor QTL, the absence of the QTL itself may not be reflected in the phenotype of a quantitative trait. Amplification of the marker designed for Tall types in Chowghat Green Dwarf is due to the fact that even though the name says dwarf, this variety belongs to the intermediate ecotype (KAU, 2020).

Screening the coconut accessions with the marker Cocos_21 was successful in amplifying three of the five tall parental lines. Except in case where tall nature is governed by other major QTL, it could be used to identify tall coconut types. Further, in order to establish a universal marker linked to the height of the coconut palm, more

whole genome sequences of tall, dwarf and intermediate ecotypes are required and insights from the sequencing data (whole genome and transcriptome) could help in more refined classification of the palms. After validation with other tall and dwarf cultivars, the marker identified in this present study may be used in breeding trials for marker assisted selection at a very early stage of plant development.

SUMMARY

6. SUMMARY

The study on “Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types” was carried out at Centre for Plant Biotechnology and Molecular Biology, Kerala Agricultural University, India, during 2019-2021, with the objective of identifying the differential genes and genomic regions among the tall and dwarf coconut genotypes through comparative whole genome sequence analyses and to develop molecular markers for distinguishing tall and dwarf coconut types.

The coconut genome assemblies and raw reads were retrieved from various databases, quality of the assemblies and raw reads analyzed, raw reads trimmed and assembled using SOAPdenovo2, ABySS and Velvetoptimiser. Since the assembled genome was found incomplete, it was not considered for further analysis.

Repeat masking was carried out on coconut genome assemblies, using RepeatMasker employing Dfam and RepBase as libraries. Since the library yielded insufficient masking percentage, *de novo* repeat library was prepared for the genome assemblies using RepeatModeller. The repeat libraries thus obtained have been merged to get a comprehensive and exhaustive repeat library for coconut and was used to perform repeat masking. Further, the efficiency of the combined repeat library for repeat masking in other palms was checked and found effective. The library was made publicly available at <https://kau.in/repeat-libraries>. Gene prediction was carried out for the repeat masked genomes using AUGUSTUS, a eukaryotic gene prediction tool. The gene prediction results were similar to the reported values.

Comparative analysis was carried out by NCBI BLAST+. The initial comparison was carried out using HT, CGD and CAGD genome assemblies. One to one comparison was carried out and the unique sequences obtained for dwarf and tall genomes were identified and extracted. Reverse BLAST was performed to ensure the sequences were unique and primers were designed from the sequences thus obtained. The genome assemblies of CNT and CND reported as of September 2021 were also considered for

the analysis and comparative analysis was performed by incorporating them along with the rest of the genomes.

Leaves from 10 coconut accessions, comprising five each of the tall and dwarf types, were collected from the parent palms at RARS Plicicode and the DNA was isolated using a modified CTAB method. Quality of the DNA was analyzed by electrophoresis and the quantity was analyzed using NanoDrop spectrophotometer, which ranged from 863 to 1030 ng/ μ l.

Twelve among the 27 primers screened have given PCR amplification. Ten primers produced amplification in West Coast Tall only while the primer Cocos_22 amplified West Cost Tall and Jawan Giant. The primer Cocos_21 has amplified the marker West Coast Tall, Jawan Giant, New Guinea and Chowghat Green Dwarf, thus amplifying three of five tall samples and one dwarf sample. Two tall samples didn't produce amplification which may be due to the presence of mismatches in the primer binding site or since the marker might be associated with a minor QTL. CGD produced an amplification since the variety belongs to the intermediate ecotype, even though its name is given as dwarf. The marker identified will be missing in the dwarf coconut types and thereby efficient to differentiate the tall and dwarf coconut genotypes.

REFERENCES

REFERENCES

- Al-Dous, E.K., George, B., Al-Mahmoud, M.E., Al-Jaber, M.Y., Wang, H., Salameh, Y.M., Al-Azwani, E.K., Chaluvadi, S., Pontaroli, A.C., DeBarry, J., and Arondel, V. 2011. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnol.* 29(6): 521-527.
- Aljohi, H.A., Liu, W., Lin, Q., Zhao, Y., Zeng, J., Alamer, A., Alanazi, I.O., Alawad, A.O., Al-Sadi, A.M., Hu, S., and Yu, J., 2016. Complete sequence and analysis of coconut palm (*Cocos nucifera*) mitochondrial genome. *PloS One* 11(10): e0163990.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215(3): 403-410.
- Angeles, J.G.C., Laurena, A.C., and Tecson-Mendoza, E.M., 2005. Extraction of genomic DNA from the lipid-, polysaccharide-, and polyphenol-rich coconut (*Cocos nucifera* L.). *Plant Mol. Biol. Rep.* 23(3): 297-298.
- Ashburner, G. R. and Rohde, W. 1994. Coconut germplasm characterisation using DNA marker technology. *Aust. Cent. Int. Agric. Res.* 96(4): 44-48.
- Bai, Z., Zheng, H., Lin, J., Wang, G., & Li, J., 2013. Comparative analysis of the transcriptome in tissues secreting purple and white nacre in the pearl mussel *Hyriopsis cumingii*. *PloS one* 8(1): e53617.
- Bandaranayake, C. K. 2006. An effective population size for reliable map resolution of Coconut (*Cocos nucifera* L.). *CORD* Give full name with abbreviation 22(2): 33-40.
- Bandaranayake, C. K. and Kearsey, M. J. 2005. Genome mapping, QTL analysis and MAS: Importance, principle, constraints and application in coconut. *Plant Genet. Resour. Newsl.* 142: 47-54.

- Bansal, J., Gupta, K., Rajkumar, M.S., Garg, R., and Jain, M., 2020. Draft genome and transcriptome analyses of halophyte rice *Oryza coarctata* provide resources for salinity and submergence stress response factors. *Physiol. Plant.*
- Bao, Z. and Eddy, S. R. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12: 1269-76. <https://doi.org/10.1101/gr.88502>
- Batugal, P. and Oliver, J., 2003. *Poverty reduction in coconut growing communities, Volume I: The framework and project plan*, IPGRI-APO, Serdang, Selangor, Malaysia, 337 p.
- Baudouin, L., Lebrun, P., Konan, J. L., Ritter, E., Berger, A., and Billotte, N. 2006. QTL analysis of fruit components in the progeny of a Rennell Island Tall coconut (*Cocos nucifera* L.) individual. *Theor. Appl. Genet.* 112(2): 258-268.
- Boonkaew, T., Mongkolsiriwatana, C., Vongvanrungruang, A., Srikulnath, K., and Peyachoknagul, S., 2018. Characterization of GA20ox genes in tall and dwarf types coconut (*Cocos nucifera* L.). *Genes Genomics* 40(7): 735-745.
- Bork, P., 2000. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res.* 10(4): 398-400.
- Brenner, S. E., 1999. Errors in genome annotation. *Trends Genet.* 15(4): 132-133.
- Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M., 2021. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinf.* 3(1): 108.
- CDB (Coconut Development Board, Government of India), 2021. All India final estimates of area and production of coconut 2019-20. Available: <https://coconutboard.gov.in/Statistics.aspx>, [Accessed 03 Jan. 2022].
- Chethana, S., 2016. Analysis of inbreeding depression in West Coast Tall coconut

(*Cocos nucifera* L.), M.Sc.(Ag.) thesis, Kerala Agricultural University, Thrissur, India, 77p.

Christian, C., George, C., Vahram, A., Ning, M., Jason, P., Kevin, B. and Thomas, L. M., 2009. BLAST+: architecture and applications. *BMC Bioinf.* 10(1): 421.

Couch, J. A. and Fritz, P. J., 1990. Isolation of DNA from plants high in polyphenolics. *Plant Mol. Biol. Report.* 8(1): 8-12.

Daher, R. F., Pereira, M. G., Tupinamba, E. A., Junior, A. A. T., Aragao, W. M., Ribeiro, F. E., Oliveira, L. O., and Sakiyama, N. S., 2002. Assessment of coconut tree genetic divergence by compound sample RAPD marker analysis. *Crop Breed. Appl. Biotechnol.* 2(3): 10–14.

Dasanayaka, P. N., Nandadasa, H. G., Everard, J. M. D. T., and Karunanayaka, E. H. 2009. Analysis of coconut (*Cocos nucifera* L.) diversity using microsatellite markers with emphasis on management and utilisation of genetic resources. *J. Natl. Sci. Found. Sri Lanka* 37(2): 99-109.

Datta, S. K., Torrizo, L. B., Tu, J., Oliva, N. P., and Datta, K., 1997. Production and molecular evaluation of transgenic rice plants. IRRI Discussion Paper Series No. 21, International Rice Research Institute, Manila, Philippines, pp. 26-27.

Davis, R.W., Thomas, M., Cameron, J., John, T. P. S., Scherer, S. and Padgett, R. A., 1980. Rapid DNA isolations for enzymatic and hybridization analysis. In: Grossman, L. and Moldave, K. (Eds.), *Methods in Enzymology*, 65: 404-411, Academic Press, Cambridge, USA.

Dellaporta, S. L., Wood, J., and Hicks, J. B., 1983. A plant DNA mini preparation: version II. *Plant Mol. Biol. Rep.* 1(4): 19-21.

Devakumar, K. V., Niral, B. A., Jerard, C., Jayabose, R., Chandramohan, and Jacob, P. M., 2010. Microsatellite analysis of distinct coconut accessions

- from Agatti and Kavaratti Islands, Lakshadweep, India. *Sci. Hortic.* 125: 309-315.
- Di Genova, A., Almeida, A. M., Munoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., Pinto, M., Hinrichsen, P., Orellana, A., and Maass, A., 2014. Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* 14(1): 7.
- Doyle, J. J. and Doyle, J. L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19: 11-15.
- Duran, Y., Rohde, W., Kullaya, A., Goikotxea, P., and Ritter, E., 1997. Molecular analysis of East African tall coconut genotypes by DNA marker technology. *J. Genet. Breed.* 57: 279-288.
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., and Smit, A. F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS.* 117(17): 9451-9457.
- Freitas Neto, M., Pereira, T. N. S., Geronimo, I. G. C., Azevedo, A. O. N., Ramos, S. R. R., and Pereira, M. G., 2016. Coconut genome size determined by flow cytometry: tall versus dwarf types. *Genet. Mol. Res.* 15(1): 1-9.
- Geethanjali, S., Rukmani, J. A., Rajakumar, D., Kadirvel, P., and Viswanathan, P. L., 2018. Genetic diversity, population structure and association analysis in coconut (*Cocos nucifera* L.) germplasm using SSR markers. *Plant Genet. Resour.* 16(2): 156-168.
- Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., and Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucl. Acids Res.* 36(10): 3420-3435.

- Gunn, B. F., Baudouin, L., Beulé, T., Ilbert, P., Duperray, C., Crisp, M., Issali, A., Konan, J. L. and Rival, A., 2015. Ploidy and domestication are associated with genome size variation in Palms. *Am. J. Bot.* 102(10): 1625-1633.
- Hedden, P., 2003. The genes of the Green Revolution. *Trends Genet.* 19(1): 5-9.
- Herran, A., Estioko, L., Becker, D., Rodriguez, M. J. B., Rohde, W., and Ritter, E., 2000. Linkage mapping and QTL analysis in coconut (*Cocos nucifera* L.). *Theor. Appl. Genet.* 101(2): 292-300.
- Housley, D. J., Zalewski, Z. A., Beckett, S. E. and Venta, P. J., 2006. Design factors that influence PCR amplification success of cross-species primers among 1147 mammalian primer pairs. *BMC Genomics* 7: 253.
- Huang, Y. Y., Matzke, A. J., and Matzke, M., 2013. Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). *PLoS One* 8(8): e74736.
- Hunter, H. and Leake, L. M., 1933. *Recent Advances in Agricultural Plant Breeding*, J. and A. Churchill, London, 326p.
- Hurwitz, B. L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S. A., Ware, D., Wing, R. A., and Stein, L., 2010. Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*. *Plant J.* 63(6): 990-1003.
- IPGRI. 1995. Descriptors for coconut (*Cocos nucifera* L.). International Plant Genetic Resources Institute, Rome, Italy, 68p.
- Jayalekshmy, V.G. 1996. Biochemical and molecular markers in coconut (*Cocos nucifera* L.), Ph.D. thesis, Tamil Nadu Agricultural University, Coimbatore, 153p.

- Jegadeesan, S., Raizada, A., Dhanasekar, P., and Suprasanna, P., 2021. Draft genome sequence of the pulse crop blackgram [*Vigna mungo* (L.) Hepper] reveals potential R-genes. *Sci. Rep.* 11(1): 1-10.
- Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R. et al., 2013. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496(7443): 91-95.
- Joseph, P. A., 2007. *Coconut in Kerala*, Kerala Agricultural University, Thrissur, India, 94p.
- Joslyn, M. A. and Ponting, J. D., 1951. Enzyme-catalyzed oxidative browning of fruit products. *Adv. Food Res.* 3: 1-44.
- Kappil, S. R., Aneja, R., and Rani, P., 2021. Decomposing the performance metrics of coconut cultivation in the South Indian States. *Humanit. Soc. Sci. Commun.* 8: 114.
- Katterman, F. R. H. and Shattuck, V. I., 1983. An effective method of DNA isolation from the mature leaves of *Gossypium* species that contain large amounts of phenolic terpenoids and tannins. *Prep. Biochem.* 13(4): 347-359.
- Kerala Agricultural University (KAU), 2020. Annual Research Report, Kerala Agricultural University, Thrissur, India.
- Kumar, P., Gupta, V. K., Misra, A. K., Modi, D. R., and Pandey, B. K. 2009. Potential of molecular markers in plant biotechnology. *Plant Omics J.* 2(4): 141-160.
- Kumar, S. P., Manimekalai, R., and Kumari, B. D. R., 2011. Microsatellite marker based characterization of south pacific coconut (*Cocos nucifera* L.) accessions. *Int. J. Plant Breed. Genet.* 5(1): 34-43.
- Lantican, D. V., Strickler, S. R., Canama, A. O., Gardoce, R. R., Mueller, L. A. and Galvez, H. F., 2019. *De novo* genome sequence assembly of dwarf coconut

- (*Cocos nucifera* L. ‘Catigan Green Dwarf’) provides insights into genomic variation between coconut types and related palm species. *G3: Genes, Genomes, Genet.* 9(8): 2377-2393.
- Lebrun, P., Baudouin, L., Bourdeix, R., Konan, J. L., Barker, J. H., Aldam, C., Herran, A., and Ritter, E., 2001. Construction of a linkage map of the Rennell Island Tall coconut type (*Cocos nucifera* L.) and QTL analysis for yield characters. *Genome* 44(6): 962-970.
- Lebrun, P., N'cho, Y. P., Seguin, M., Grivet, L., and Baudouin, L., 1998. Genetic diversity in coconut (*Cocos nucifera* L.) revealed by restriction fragment length polymorphism (RFLP) markers. *Euphytica* 101(1): 103-108.
- Li, S., Wang, S., Deng, Q., Zheng, A., Zhu, J., Liu, H., Wang, L., Gao, F., Zou, T., Huang, B., and Cao, X., 2012. Identification of genome-wide variations among three elite restorer lines for hybrid-rice. *PLoS One* 7(2): e30952.
- Ling, H.Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Y., and Yu, Y., 2018. Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature* 557(7705): 424-8.
- Ling, H.Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., and Gao, C., 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* 496: 87-90.
- Liu, C., Ma, N., Wang, P. Y., Fu, N., & Shen, H. L., 2013. Transcriptome sequencing and de novo analysis of a cytoplasmic male sterile line and its near-isogenic restorer line in chili pepper (*Capsicum annuum* L.). *PloS one*, 8(6): e65209.
- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., and Tang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1: 2047-217X.

- Manimekalai, R. and Nagarajan, P. 2006b. Interrelationships among coconut (*Cocos nucifera* L.) accessions using RAPD technique. *Genet. Resour. Crop Evol.* 53: 1137-1144.
- Manimekalai, R. and Nagarajan, P. 2010. Bulk line analysis in coconut (*Cocos nucifera* L.) for inferring relationship between Talls, Dwarfs and Niu Leka dwarf forms. *Indian J. Plant Genet. Resour.* 23(1): 77-81.
- Manimekalai, R. and Nagarajan, P., 2006a. Assessing genetic relationships among coconut (*Cocos nucifera* L.) accessions using Inter Simple Sequence Repeat markers. *Sci. Hortic.* 108: 49-54.
- Manimekalai, R., Nagarajan, P., Bharathi, M., Karun, A., Kumar, S. N., and Kumaran, P. M., 2005. Genetic variation of selected progeny lines of coconut (*Cocos nucifera* L.) based on simple sequence repeat markers. *Trop. Agri. Res. (Sri Lanka)* 17: 58–66.
- Masumbuko, L. I., Sinje, S., and Kullaya, A., 2014. Genetic diversity and structure of East African tall coconuts in Tanzania using RAPD markers. *Open J. Genet.* 4(2): 45335. doi: 10.4236/ojgen.2014.42018
- Mauro-Herrera, M., Meerow, A. W., Borrone, J. W., Kuhn, D. N., and Schnell, R. J., 2006. Ten informative markers developed from WRKY sequences in coconut (*Cocos nucifera*). *Mol. Ecol. Notes* 6(3): 904-906.
- Meerow, A. W., Wisser, R. J., Brown, S. J., Kuhn, D. N., Schnell, R. J., and Broschat, T. K., 2003. Analysis of genetic diversity and population structure within Florida coconut (*Cocos nucifera* L.) germplasm using microsatellite DNA, with special emphasis on the Fiji Dwarf cultivar. *Theor. Appl. Genet.* 106(4): 715-726.
- Menon, K. P. V. and Pandalai, K. M., 1958. *The Coconut Palm - A Monograph*, The Indian Coconut Committee, Ernakulam, Kerala, India, pp. 199-207.

- Mondal, T.K., Rawal, H.C., Chowrasia, S., Varshney, D., Panda, A.K., Mazumdar, A., Kaur, H., Gaikwad, K., Sharma, T.R., & Singh, N.K., 2018. Draft genome sequence of first monocot-halophytic species *Oryza coarctata* reveals stress-specific genes. *Sci. Rep.* 8(1): 13698.
- Muliyar, R. K., Chowdappa, P., Behera, S. K., Kasaragod, S., Gangaraj, K. P., Kotimoole, C. N., Nekrakalaya, B., Mohanty, V., Sampgod, R. B., Banerjee, G., and Das, A. J., 2020. Assembly and annotation of the nuclear and organellar genomes of a dwarf coconut (Chowghat Green Dwarf) possessing enhanced disease resistance. *OMICS: J. Integr. Biol.*, 24(12): 726-742.
- Mun, J. H., Kwon, S. J., Yang, T. J., Seol, Y. J., Jin, M., Kim, J. A., Lim, M. H., Kim, J. S., Baek, S., Choi, B. S., and Yu, H. J., 2009. Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biol.* 10: R111.
- Nair, M. K. and Ratnambal, M. J., 1994. Genetic resources of coconut. *Adv. Hortic.* 9: 51-63.
- Nair, M. K., 1992. Coconut genetic resources. *CORD* 8(1): 34.
- Nair, R. V., Jerard, B. A. and Thomas, R. J., 2016. Coconut breeding in India. In: Al-Khayri, J. M. Jain, M. S. and Johnson D. V. (eds.), *Advances in Plant Breeding Strategies: Agronomic, Abiotic and Biotic Stress Traits*, Springer, Cham Heidelberg New York, p. 257-279.
- Narayana, G. V. and John, C. M., 1949. Varieties and forms of the coconut. *Indian Coconut J.* 2: 209-226.
- Niral, V., Jerard, B. A., Samsudeen, K., Nair, R.V., Kumaran, P. M., and Thomas, G. V., 2014b. IND 414-Chowghat Yellow Dwarf (IC0598220; INGR13062), Distinct dwarf coconut (*Cocos nucifera*) germplasm with yellow coloured nuts and erect leaves. *Indian J. Plant Genetic Resour.* 27(1): 75-76.

- Niral, V., Jerard, B. A., Samsudeen, K., Ratnambal, M. J., Kumaran, P. M., Rao, E. V. V., and Thomas, G. V., 2014a. IND 001–Kappadam Tall (IC0430667; INGR13059), a coconut (*Cocos nucifera*) germplasm with low husk (33 to 36%) and high copra content (215 to 280 g). *Indian J. Plant Genet. Resour.* 27(1): 73.
- Patel, J. S., 1938. Coconut breeding. *Proc. Assoc. Evon. Biol.* 5: 1-16.
- Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., Beales, J., Fish, L. J., Worland, A. J., Pelica, F., and Sudhakar, D., 1999. ‘Green revolution’ genes encode mutant gibberellin response modulators. *Nature* 400(6741): 256-261.
- Perera, L., 2006. Report of the Genetics and Plant Breeding Division, Annual Report of the Coconut Research Institute, Sri Lanka.
- Perera, L., Baudouin, L., and Mackay, I., 2016. SSR markers indicate a common origin of self-pollinating dwarf coconut in South-East Asia under domestication. *Sci. Hortic.* 211: 255-262.
- Perera, L., Russell, J. R., Provan, J., and Powell, W., 2000. Use of microsatellite DNA markers to investigate the level of genetic diversity and population genetic structure of coconut (*Cocos nucifera* L.). *Genome* 43(1): 15-21.
- Perera, L., Russell, J. R., Provan, J., and Powell, W., 2003. Studying genetic relationships among coconut varieties/populations using microsatellite markers. *Euphytica* 132: 121-128.
- Perera, L., Russell, J. R., Provan, J., McNicol, J. W., and Powell, W., 1998. Evaluating genetic relationships between indigenous coconut (*Cocos nucifera* L.) accessions from Sri Lanka by means of AFLP profiling. *Theor. Appl. Genet.* 96 (3-4): 545-550.
- Perera, P. I., Perera, L., Hocher, V., Verdeil, J. L., Yakandawala, D. M. D., and

- Weerakoon, L. K., 2008. Use of SSR markers to determine the anther-derived homozygous lines in coconut. *Plant Cell Rep.* 27(11): 1697-1703.
- Pesik, A., Efendi, D., Novariantio, H., Dinarti, D., and Sudarsono, S., 2017. Development of SNAP markers based on nucleotide variability of WRKY genes in coconut and their validation using multiplex PCR. *Biodiversitas J. Biol. Diversity* 18(2): 465-475.
- Pootakham, W., Nawae, W., Naktang, C., Sonthirod, C., Yoocha, T., Kongkachana, W., Sangsrakru, D., Jomchai, N., U-thoomporn, S., Somta, P., Laosatit, K., and Tangphatsornruang, S., 2021. A chromosome-scale assembly of the black gram (*Vigna mungo*) genome. *Mol. Ecol. Resour.* 21, 238-50.
- Porebski, S., Bailey, L. G., and Baum, B. B., 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* 15(1): 8-15.
- Price, A. L., Jones, N. C., and Pevzner, P. A., 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics* 21: i351-i358. <https://doi.org/10.1093/bioinformatics/bti1018>
- Rahmawati, A., Volkaert, H.A., Dinarti, D., Maskromo, I., Hatta, A. N. N. L., and Sudarsono, S., 2021. Complete chloroplast genome sequences of coconut cv. Kopyor Green Dwarf and comparative genome analysis to oil palm, date palm, sago palm, and miniature sugar palm. In: Tombuloglu, H., Unver, T., Tombuloglu, G., and Hakeem, K. R. (eds.), *Oil Crop Genomics*, Springer Nature, Switzerland, pp.189-216.
- Rajesh, M. K., Arunachalam, V., Nagarajan, P., Lebrun, P., Samsudeen, K., and Thamban, C., 2008a. Genetic survey of 10 Indian coconut landraces by simple sequence repeats (SSRs). *Sci. Hortic.* 118(4): 282-287.
- Rajesh, M. K., Jerard, B. A., Preethi, P., Jacob, R., and Karun, A., 2014b. Application of RAPD markers in hybrid verification in coconut. *Crop Breed.*

Appl. Biotechnol. 14: 36-41.

Rajesh, M. K., Jerard, B. A., Preethi, P., Thomas, R. J., Fayas, T. P., Rachana, K. E., and Karun, A., 2013. Development of a RAPD-derived SCAR marker associated with tall-type palm trait in coconut. *Sci. Hortic.* 150: 312-316.

Rajesh, M. K., Nagarajan, P., Jerard, B. A., Arunachalam, V., and Dhanapal, R. 2008b. Microsatellite variability of coconut accessions (*Cocos nucifera* L.) from Andaman and Nicobar Islands. *Curr. Sci.* 94 (12): 1627-1631.

Rajesh, M. K., Ramesh, S. V., Perera, L., and Manickavelu, A., 2021. Quantitative Trait Loci (QTL) and Association Mapping for major agronomic traits. In: Rajesh, M. K., V. Ramesh, S. V., Perera L., and Kole, C. (Eds.), *The Coconut Genome*, Springer Nature, Switzerland, pp. 91-101.

Rajesh, M. K., Rijith, J., Rahman, S., Preethi, P., Rachana, K. E., Sajini, K. K., and Karun, A., 2014a. Estimation of out-crossing rates in populations of West Coast Tall cultivar of coconut using microsatellite markers. *J. Plant. Crops* 42(3): 277-288.

Rajesh, M. K., Sabana, A. A., Rachana, K. E., Rahman, S., Ananda, K. S., and Karun, A., 2016. Development of a SCoT-derived SCAR marker associated with tall-type palm trait in arecanut and its utilization in hybrid (dwarf x tall) authentication. *Indian J. Genet. Plant Breed.* 76: 119-122.

Rajesh, M. K., Sabana, A. A., Rachana, K. E., Rahman, S., Jerard, B. A., and Karun, A., 2015. Genetic relationship and diversity among coconut (*Cocos nucifera* L.) accessions revealed through SCoT analysis. *3 Biotech* 5(6): 999-1006. doi: 10.1007/s13205-015-0304-7.

Rajesh, M. K., Thomas, R. J., Rijith, J., Shareefa, M. and Jacob, P. M., 2012. Genetic purity assessment of D x T hybrids in coconut with SSR markers. *Indian J. Genet. Plant Breed.* 72(4): 472-last page.

- Ramachandran, M., Venkateswaran, A. N., Sridharan, C. S., and Balasubramanian, K., 1974. Performance of different hybrids-preliminary study. *Cocon. Bull.* 5: 2-7.
- Rasam, D. V., Gokhale, N. B., Sawardekar, S. V., and Patil, D. M., 2016. Molecular characterisation of coconut (*Cocos nucifera* L.) varieties using ISSR and SSR markers. *J. Hortic. Sci. Biotechnol.* 91(4): 347-352.
- Remany, C., 2003. Cataloguing and categorisation of line exploited ecotypes of coconut grown in Kerala. PhD thesis, Mahatma Gandhi University, India, 175 p.
- Ritter, E., Rodriguez, M. J. B., Herran, A., Estioko, L., Becker, D., and Rohde, W., 2000. Analysis of quantitative trait locis (QTL) based on linkage maps in coconut (*Cocos nucifera* L.). In: Bolwell, G. P. (.ed), *Plant Genetic Engineering: Towards the Third Millennium: Proceedings of the International Symposium on Plant Genetic Engineering*, Havana, Cuba, 6-10 December, 1999, Elsevier Science Publishers, pp. 42-48.
- Rohde, W., Becker, D., Kullaya, A., Rodriguez, J., Herran, A., and Ritter, E., 1999. Analysis of coconut germplasm biodiversity by DNA marker technologies and construction of a genetic linkage map. In: Oropeza, C., Verdeil, J. L., Ashburner, G. R., Cardena, R., and Santamaria, J. M., (Eds.), *Current Advances in Coconut Biotechnology*, Springer, Dordrecht, pp. 99-120.
- Sasaki, A., Ashikari, M., Ueguchi-Tanaka, M., Itoh, H., Nishimura, A., Swapan, D., Ishiyama, K., Saito, T., Kobayashi, M., Khush, G. S., and Kitano, H., 2002. A mutant gibberellin-synthesis gene in rice. *Nature*. 416(6882): 701-702.
- Shalini, K. V., Manjunatha, S., Lebrun, P., Berger, A., Baudouin, L., Pirany, N., Ranganath, R. M., and Prasad, D. T., 2007. Identification of molecular markers associated with mite resistance in coconut (*Cocos nucifera* L.). *Genome* 50(1): 35-42.

- Shi, C., Li, W., Zhang, Q. J., Zhang, Y., Tong, Y., Li, K., Liu, Y. L., and Gao, L. Z., 2020. The draft genome sequence of an upland wild rice species, *Oryza granulata*. *Sci. Data* 7(1): 131.
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19(6): 1117-1123.
- Sing, B. D., 2008. *Biotechnology Expanding Horizons* (2nd Ed.), Kalyani Publishers, New Delhi, 919 p.
- Singh, R., Ong-Abdullah, M., Low, E.T., Manaf, M.A., Rosli, R., Nookiah, R., Ooi, L.C., Ooi, S.E., Chan, K.L., Halim, M.A., and Azizi, N., 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* 500(7462): 335-9.
- Sivashankari, S. and Shanmughavel, P., 2006. Functional annotation of hypothetical proteins – A review. *Bioinformatics* 1(8): 335.
- Smalle, J. and Vierstra, R. D., 2004. The ubiquitin 26S proteasome proteolytic pathway. *Annu. Rev. Plant Biol.* 55: 555-590.
- Stanke, M., & Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(suppl_2): ii215-ii225.
- Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24(5): 637-644.
- Swaminathan, M. S. and Nambiar, M. C., 1961. Cytology and origin of the dwarf coconut palm. *Nature* 192(4797): 84-85.
- Tabidze, V., Pipia, I., Gogniashvili, M., Kunelauri, N., Ujmajuridze, L., Pirtskhalava, M., Vishnepolsky, B., Hernandez, A. G., Fields, C. J., and Beridze, T., 2017.

- Whole genome comparative analysis of four Georgian grape cultivars. *Mol. Genet. Genomics* 292(6): 1377-1389.
- Teulat, B., Aldam, C., Trehin, R., Lebrun, P., Barker, J. H., Arnold, G. M., Karp, A., Baudouin, L., and Rognon, F., 2000. An analysis of genetic diversity in coconut (*Cocos nucifera*) populations from across the geographic range using sequence-tagged microsatellites (SSRs) and AFLPs. *Theor. Appl. Genet.* 100(5): 764-771.
- Tingey, S. V. and Tufo, J. P., 1993. Genetic analysis with RAPD DNA markers. *Plant Physiol.* 101: 349-352.
- Tiwari, S. B., Hagen, G., and Guilfoyle, T. J., 2004. Aux/IAA proteins contain a potent transcriptional repression domain. *Plant Cell* 16(2): 533-543.
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G., 2012. Primer3 - new capabilities and interfaces. *Nuc. Acids Res.* 40(15): 115.
- Upadhyay, A., Jayadev, K., Manimekalai, R., and Parthasarathy, V. A., 2004. Genetic relationship and diversity in Indian coconut accessions based on RAPD markers. *Sci. Hortic.* 99: 353-362.
- Upadhyay, A., Parthasarathy, V. A., Seema, G., and Karun, A., 1999. An efficient method of DNA extraction from coconut. *Agrotropica* 11(1): 35-38.
- Wang, S., Xiao, Y., Zhou, Z.W., Yuan, J., Guo, H., Yang, Z., Yang, J., Sun, P., Sun, L., Deng, Y., and Xie, W.Z., 2021. High-quality reference genome sequences of two coconut cultivars provide insights into evolution of monocot chromosomes and differentiation of fiber content and plant height. *Genome Biol.* 22(1): 1-25.
- Wang, X., Lu, P., and Luo, Z., 2013. GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformatics* 9(10): 541.

- Wei, X., Zhu, X., Yu, J., Wang, L., Zhang, Y., Li, D., Zhou, R., and Zhang, X., 2016. Identification of sesame genomic variations from genome comparison of landrace and variety. *Front. Plant Sci.* 7: 1169.
- Wu, Y., Yang, Y., Qadri, R., Iqbal, A., Li, J., Fan, H., and Wu, Y., 2019. Development of SSR markers for coconut (*Cocos nucifera* L.) by Selectively Amplified Microsatellite (SAM) and its applications. *Trop. Plant Biol.* 12(1): 32-43.
- Xiao, Y., Xu, P., Fan, H., Baudouin, L., Xia, W., Bocs, S., Xu, J., Li, Q., Guo, A., Zhou, L., and Li, J., 2017. The genome draft of coconut (*Cocos nucifera*). *GigaScience* 6(11): 95.
- Young, N.D., 1994. Constructing a plant genetic linkage map with DNA markers. In: Phillips, R. L., and Vasil, I. K. (eds.), *DNA-based Markers in Plants* (Vol. 6), Springer, Dordrecht, pp. 39-57.
- Zerbino, D. R. and Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18(5): 821-829.
- Zhang, S., Chen, W., Xin, L., Gao, Z., Hou, Y., Yu, X., Zhang, Z., and Qu, S., 2014. Genomic variants of genes associated with three horticultural traits in apple revealed by genome re-sequencing. *Hortic. Res.* 1: 14045.
- Zheng, L. Y., Guo, X. S., He, B., Sun, L. J., Peng, Y., Dong, S. S., Liu, T. F., Jiang, S., Ramachandran, S., Liu, C. M., and Jing, H. C., 2011. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.* 12: R114.

ANNEXURE

Annexure I: List of stock solutions used in experiment

1. Extraction buffer: 100 mM Tris HCl, pH 8.0
50 mM EDTA, pH 8.0
500 mM NaCl
2. T10 E1 : Tris 10 mM containing 1 mM EDTA
3. RNase (10 mg/ml): Dissolve Rnase in water, place in a tube in a boiling water bath for 10 minutes. Allow this to cool on a bench and store at -20 0C
4. Chloroform : isoamyl : Chloroform 240 ml, Isoamyl alcohol 10 ml.
alcohol (24:1) Store in a dark room temperature. Make up and dispense the solution in a fumed cupboard.
5. Ethanol (70 %): Absolute alcohol (70 ml), Distill water (30 ml)
6. NaCl 5 M: Dissolve 292.2 g NaCl in 750 ml water.
Make up to 1 litre with water, filter and autoclave.
7. Tris HCl (1 M, pH 8.0): Dissolve 121.1 g Tris in 800 ml of water. Adjust pH 8.0 with conc. HCl. Make up the volume to 1 L and autoclave
8. Ethidium bromide: Dissolve 100 mg ethidium bromide in 10 ml of distil (10 mg/ml) water, wrap tube in aluminium foil and store at 4 0C.
Caution : Ethidium bromide is extremely mutagenic
9. 100 bp DNA ladder: 10 µl of 100 bp marker (genetics)
95 µl of loading dye
95 µl sterile dH2O
10. Loading dye (6Å~) : 0.25 Bromophenol blue
40 % (W/V) sucrose in water
Started at 4°C
11. 50Å~ TAE: 242 g/l Tris base and 57.1 ml Glacial acetic acid
(Tris-Acetate EDTA) 100 ml 0.5 M EDTA (pH 8.0)
1000 ml Distilled water

Annexure II: List of chemicals and other consumables used for wet lab analysis

Sl. No.	Chemical Name	Company Name
1	CTAB	SRL
2	2-Mercaptoethanol	SRL
3	Polyvinylpyrrolidone	SRL
4	EDTA Disodium Salt	SRL
5	Tris Buffer	SRL
6	Chloroform	SRL
7	Isopropanol	SRL
8	Agarose	SRL
9	dNTP Mix, 10 mM (2.5 mM each), 1000 µl	GeNei
10	Taq DNA Polymerase (3 U/µl) 1000 U	GeNei
11	RNaseA (DNase free) 50 mg	GeNei
12	PCR Tubes Flat Lid	Himedia
13	Pipette Tips, PP, Neutral colour	Himedia

Annexure III: List of laboratory equipment used for the study

Sl. No.	Equipment	Model	Manufacturer
1.	Centrifuge	5418 R – Micro Centrifuge	Eppendorf, Germany
2.	Thermal cycler	T100	Bio-Rad, USA
3.	Gel Doc	Gel Doc XR+	Bio-Rad, USA
4.	Electrophoresis unit	Wide Mini-Sub Cell GT Cell &PowerPac	Bio-Rad, USA
5.	Water bath	WATER BATH RECTANGULAR– THERMOSTATIC	ROTEK, India
6.	Vortex Mixer	Vortexer	GeNei, India
7.	Spinwin	MC-00 Micro Centrifuge	Tarsons, India
8.	Weighing balance	ATX	Shimadzu, Japan
9.	Nanodrop	ND1000	ThermoScientific

Annexure VI: Online BLAST results

HT					
Seqname	Description	Len gth	#h its	E- value	Sim mean
Cm017872.1.g16 83.tl	Unc93-like protein 1	522	5	8.28e -34	71.43
Cm017873.1.g44 67.tl	Protein hua2-like 2 isoform x1	483	5	2.95e -27	95.19
Cm017874.1.g53 81.tl	Putative pectinesterase-like	432	5	2.86e -32	78.16
Cm017874.1.g53 88.tl	F-box/kelch-repeat protein at1g57790	354	5	4.14e -24	74.49
Cm017874.1.g56 52.tl	Gtpase-activating protein pac-1	465	5	9.96e -24	84.09
Cm017874.1.g63 88.tl	Hypothetical protein b296_00053964	405	1	5.78e -04	75.86
Cm017875.1.g72 66.tl	Proline-rich extensin-like protein epr1	150 9	5	9.14e -17	43.83
Cm017877.1.g98 24.tl	Uncharacterized membrane protein c776.05	384	5	4.72e -79	89.93
Cm017877.1.g11 731.tl	Solute carrier family 25 member 44-like isoform x1	240	5	2.79e -14	77.29
Cm017878.1.g12 618.tl	Hypothetical protein cocnu_03g006860	336	1	8.83e -07	86.21
Cm017878.1.g12 907.tl	Mitochondrial inner membrane protein oxa1-like	651	5	1.56e -60	89.52
Cm017878.1.g13 656.tl	Protein executer 1, chloroplastic isoform x1	621	5	5.31e -62	71.59
Cm017878.1.g13 772.tl	E3 ubiquitin-protein ligase cip8-like	321	5	7.90e -13	55.14
Cm017879.1.g14 064.tl	Probable glutathione s-transferase dhar2, chloroplastic isoform x2	240	5	2.96e -23	88.09
Cm017881.1.g17 195.tl	Crocetin glucosyltransferase 3	534	5	7.93e -73	67.54
Cm017882.1.g18 311.tl	Gpi-anchored protein llg1-like	279	2	8.63e -08	90.78
Cm017882.1.g19 076.tl	Metal tolerance protein c4	417	5	1.21e -33	84.68
Cm017883.1.g19 887.tl	Dual specificity protein kinase clk3	441	2	1.33e -09	73.87
Cm017884.1.g20 164.tl	Putative phosphatidylinositol/phosphatidylcholine transfer protein sfh8	237	1	6.63e -13	96.88
Cm017884.1.g20 257.tl	Nuclear transcription factor y subunit a-3 isoform x2	222	5	9.84e -26	70.41
Cm017884.1.g20 436.tl	Lysine-specific histone demethylase 1	282	1	8.73e -06	100
Cm017885.1.g21 486.tl	Scopoletin glucosyltransferase-like	756	5	8.43e -102	89.92
Cm017885.1.g21 778.tl	Hypothetical protein es288_d10g149600v1	231	5	2.31e -06	77.96
Cm017886.1.g22 856.tl	Glycolipid transfer protein 3	351	5	5.14e -16	94.56
Cm017887.1.g23 255.tl	Protein hua2-like 3-like	657	5	4.97e -11	80.24

Voi01003786.1.g24951.t1	Lysine-specific histone demethylase 1 homolog 1	528	5	1.59e-69	82.82
Voi01014032.1.g25157.t1	Ubiquitin-like modifier-activating enzyme 5 isoform x2	264	5	3.21e-31	94.42
Voi01012763.1.g26490.t1	Uncharacterized protein loc103697680	110	5	8.97e-17	97.22
Voi01007875.1.g26538.t1	Hypothetical protein	753	5	9.86e-81	69.3
Voi01008680.1.g28527.t1	Brassinazole-resistant 1-like	393	5	6.42e-09	76.49
Voi01022983.1.g28657.t1	Proline-rich protein 36-like	832	5	1.34e-38	49.71
Voi01113112.1.g30207.t1	Cytochrome b561, dm13 and domon domain-containing protein at5g54830	876	5	2.59e-14	50.44
Voi01113087.1.g30213.t1	Mitogen-activated protein kinase kinase 10	140	5	2.48e-13	73.37
Voi01111989.1.g30502.t1	Protein what's this factor 9, mitochondrial	804	5	1.24e-121	79.93
Voi01110342.1.g30712.t1	Putative methyltransferase at1g22800, mitochondrial	373	5	9.88e-12	61.56
Voi01108478.1.g30874.t1	Udp-glycosyltransferase 71k1	261	5	1.55e-45	89.89
Voi01056850.1.g30992.t1	Gag-pol polyprotein-like protein	199	5	2.95e-06	71.37
Voi01101267.1.g31261.t1	Cytochrome c oxidase subunit 6a, mitochondrial	165	1	1.27e-05	83.33
Voi01094838.1.g31506.t1	Peroxidase 52	143	5	1.71e-18	91.49
Voi01089699.1.g31631.t1	Stress enhanced protein 1, chloroplastic	70	1	8.50e-04	90.91
Voi01085963.1.g31705.t1	Probable tetraacyldisaccharide 4'-kinase, mitochondrial isoform x4	451	5	5.16e-14	86.16
Voi01076373.1.g31860.t1	Phd finger protein alfin-like 6	154	5	1.10e-09	97.04
Voi01072228.1.g31909.t1	Uncharacterized protein loc105059337 isoform x1	99	5	2.61e-14	93.75
Voi01066753.1.g31966.t1	Uncharacterized protein at5g49945-like	294	3	4.21e-08	65.96
Voi01058983.1.g31976.t1	Nucleoporin nsp1-like	257	5	6.80e-09	84.86
Voi01065267.1.g31981.t1	Receptor-like serine/threonine-protein kinase sd1-8 isoform x1	71	5	1.91e-06	96.36
CNT					
Gwhbebt00000001.g641.t1	Putative sodium/metabolite cotransporter bass3, chloroplastic	426	5	6.36e-57	87.99
Gwhbebt00000001.g2402.t1	(s)-coclaurine n-methyltransferase	543	5	9.52e-08	87.71
Gwhbebt00000002.g3032.t1	E3 ubiquitin-protein ligase cip8-like	321	5	7.90e-13	55.14
Gwhbebt00000002.g3080.t1	Small heat shock protein, chloroplastic	531	5	2.41e-33	83.33
Gwhbebt00000002.g4131.t1	Uncharacterized protein pam68-like	411	2	7.85e-13	91.29

Gwhbebt0000000 3.g5384.t1	Uncharacterized membrane protein c776.05	384	5	4.72e -79	89.93
Gwhbebt0000000 3.g5780.t1	Hypothetical protein cocnu_06g003960	441	5	9.24e -32	97.14
Gwhbebt0000000 3.g7797.t1	3-oxo-5-alpha-steroid 4-dehydrogenase 1	255	5	9.52e -22	92.28
Gwhbebt0000000 4.g8958.t1	Unnamed protein product	582	1	5.35e -36	84.15
Gwhbebt0000000 4.g9059.t1	Gibberellin 2-beta-dioxygenase-like	684	2	1.76e -12	84.44
Gwhbebt0000000 5.g12332.t1	Putative heat repeat-containing protein	414	5	7.81e -31	95.95
Gwhbebt0000000 6.g15059.t1	Uncharacterized protein loc115667491	321	5	4.77e -11	57.27
Gwhbebt0000000 6.g15309.t1	Hypothetical protein	864	5	8.11e -100	82.83
Gwhbebt0000000 7.g17658.t1	Hypothetical protein cocnu_04g008430	573	1	6.50e -06	72.92
Gwhbebt0000000 7.g17726.t1	Ribulose-phosphate 3-epimerase, cytoplasmic isoform	252	5	7.23e -12	97.33
Gwhbebt0000000 7.g17845.t1	Glycosyltransferase family 92 protein	468	4	2.18e -24	82.35
Gwhbebt0000000 9.g20953.t1	Lysine-specific histone demethylase 1 homolog 1	528	5	1.59e -69	82.82
Gwhbebt0000001 3.g26269.t1	Probable glutathione s-transferase dhar2, chloroplastic isoform x2	240	5	2.96e -23	88.09
Gwhbebt0000001 5.g28998.t1	Nuclear transcription factor y subunit a-3 isoform x2	222	5	9.84e -26	70.41
Gwhbebt0000001 5.g29254.t1	Lysine-specific histone demethylase 1	282	1	8.73e -06	100
CND					
Gwhbebu000000 01.g193.t1	Putative myb family transcription factor at1g14600	276	5	2.29e -19	93.12
Gwhbebu000000 02.g3176.t1	Hypothetical protein	312	1	9.13e -05	60
Gwhbebu000000 02.g3959.t1	Hypothetical protein cocnu_11g002520	609	4	2.05e -11	51.84
Gwhbebu000000 02.g4715.t1	Transducin beta-like protein 3	435	5	1.29e -68	76.24
Gwhbebu000000 03.g6367.t1	Probable nad(p)h dehydrogenase subunit crr3, chloroplastic	312	5	3.29e -26	83.25
Gwhbebu000000 03.g6883.t1	Hypothetical protein cocnu_06g012000	618	1	6.27e -29	100
Gwhbebu000000 03.g7799.t1	Alpha-1,6-mannosyl-glycoprotein 2-beta-n-acetylglucosaminyltransferase	921	5	0	84.96
Gwhbebu000000 04.g9344.t1	Non-specific lipid-transfer protein-like protein	309	5	1.10e -36	69.93
Gwhbebu000000 04.g10323.t1	Hypothetical protein m569_17280, partial	102 9	1	2.98e -04	95.65
Gwhbebu000000 05.g12810.t1	Uncharacterized protein loc105061501 isoform x1	516	5	6.95e -25	87.98
Gwhbebu000000 05.g13300.t1	Abc transporter c family member 14	345	5	6.26e -32	58.79

Gwhbebu000000 05.g13449.t1	Proline-rich receptor-like protein kinase perk10	101	5	4.23e -58	90.2
Gwhbebu000000 06.g13725.t1	Protein exordium-like 3	312	5	9.24e -16	59.39
Gwhbebu000000 06.g14572.t1	Hypothetical protein cocnu_14g000340	861	1	2.10e -04	75
Gwhbebu000000 09.g20228.t1	Hypothetical protein	360	1	6.13e -04	81.25
Gwhbebu000000 13.g26514.t1	Putative serine/threonine-protein kinase	618	1	5.37e -04	68.29
Gwhbebu000000 13.g27039.t1	Uncharacterized protein loc105059678	279	5	2.05e -06	90.4
Gwhbebu000000 13.g27290.t1	Conserved oligomeric golgi complex subunit 4	246	5	1.51e -16	81.65
Gwhbebu000000 15.g28959.t1	3-oxoacyl-[acyl-carrier-protein] synthase iii, chloroplastic	348	5	2.91e -14	92.1
CAGD					
Qrfj01000014.1.g 2936.t1	Pentatricopeptide repeat-containing protein ogr1, mitochondrial	249	5	2.70e -39	85.01
Qrfj01000017.1.g 3150.t1	Glutathione hydrolase 1	237	5	1.19e -23	79.11
Qrfj01000021.1.g 3611.t1	Protein yeez isoform x2	219	4	5.22e -09	97.73
Qrfj01000028.1.g 4466.t1	Hypothetical protein m569_17280, partial	123	1	4.62e -04	95.65
Qrfj01000033.1.g 5139.t1	Protein exordium-like 3	312	5	9.24e -16	59.39
Qrfj01000073.1.g 8623.t1	Non-specific lipid-transfer protein-like protein	309	5	1.10e -36	69.93
Qrfj01000090.1.g 9710.t1	110 kda u5 small nuclear ribonucleoprotein component clo-like	747	5	5.10e -11	56.6
Qrfj01000111.1.g 10975.t1	Uncharacterized protein loc105055396	384	5	1.06e -61	75.08
Qrfj01000116.1.g 11329.t1	Golgi snap receptor complex member 1-2	555	5	2.29e -18	90.85
Qrfj01000133.1.g 12242.t1	Zinc finger rna-binding protein-like	438	5	1.07e -18	71.14
Qrfj01000171.1.g 13729.t1	Putative udp-rhamnose:rhamnosyltransferase 1	372	3	3.16e -17	69.8
Qrfj01000193.1.g 14354.t1	Probable methyltransferase pmt19	339	5	1.09e -10	73.3
Qrfj01000318.1.g 17517.t1	Thiosulfate sulfurtransferase 16, chloroplastic isoform x1	276	5	2.25e -56	74.44
Qrfj01000324.1.g 17731.t1	Hypothetical protein cocnu_02g016640	258	3	2.82e -20	87.38
Qrfj01000343.1.g 17942.t1	Putative myb family transcription factor at1g14600	279	5	2.38e -19	93.12
Qrfj01000612.1.g 20931.t1	Pentatricopeptide repeat-containing protein -like	309	5	4.17e -45	83.46
Qrfj01000628.1.g 21059.t1	Multiple myeloma tumor-associated protein 2	567	5	1.67e -69	98.02
Qrfj01000685.1.g 21752.t1	Protein ctr9 homolog	225	5	4.67e -14	95.15

Qrfj01000823.1.g 22580.t1	Novel plant snare 13	297	5	2.45e -07	96.12
Qrfj01000863.1.g 22831.t1	Uncharacterized protein loc105061501 isoform x1	516	5	6.95e -25	87.98
Qrfj01000879.1.g 22998.t1	Probable nad(p)h dehydrogenase subunit crr3, chloroplastic	312	5	3.29e -26	83.25
Qrfj01001072.1.g 23736.t1	Pentatricopeptide repeat-containing protein	630	5	2.76e -48	87.92
Qrfj01001144.1.g 24008.t1	Nitrate regulatory gene2 protein	297	5	9.18e -10	94.81
Qrfj01001929.1.g 26265.t1	Uncharacterized protein loc103703430	603	5	1.00e -15	63.3
Qrfj01002054.1.g 26460.t1	Chaperone protein clpd2, chloroplastic	336	2	4.91e -07	54.76
Qrfj01002298.1.g 26910.t1	Abc transporter c family member 14	345	5	6.26e -32	58.79
Qrfj01002537.1.g 27284.t1	Rna-dependent rna polymerase	291	5	2.64e -10	88.48
Qrfj01002668.1.g 27453.t1	Leaf rust 10 disease-resistance locus receptor-like protein kinase-like 1.2 isoform x2	411	5	3.11e -28	77.68
Qrfj01002950.1.g 28139.t1	Probable e3 ubiquitin-protein ligase rhc1a isoform x2	432	5	8.09e -39	82.65
Qrfj01003393.1.g 28603.t1	Nicotinamidase 1	498	5	3.12e -09	73.59
Qrfj01004717.1.g 29683.t1	Alpha-1,6-mannosyl-glycoprotein 2-beta-n- acetylglucosaminyltransferase	921	5	0	84.96
Qrfj01005752.1.g 30288.t1	Lysine n-methyltransferase eef2kmt	519	5	2.51e -12	94.68

CGD

Pdmh01000073.1 .g1460.t1	3-oxoacyl-[acyl-carrier-protein] synthase iii, chloroplastic	402	5	9.53e -22	93.33
Pdmh01000381.1 .g5293.t1	Hypothetical protein cocnu_11g002520	609	4	2.05e -11	51.84
Pdmh01000468.1 .g6126.t1	Non-specific lipid-transfer protein-like protein	309	5	1.10e -36	69.93
Pdmh01000649.1 .g7619.t1	Aspartic proteinase nana, chloroplast-like	300	5	3.08e -12	68.83
Pdmh01000703.1 .g8035.t1	Probable nad(p)h dehydrogenase subunit crr3, chloroplastic	312	5	3.29e -26	83.25
Pdmh01000933.1 .g9704.t1	Thaumatococcus-like protein 1	624	5	2.80e -111	74.98
Pdmh01001092.1 .g10578.t1	Chaperone protein clpd2, chloroplastic	336	2	4.91e -07	54.76
Pdmh01001176.1 .g11090.t1	Crocetin glucosyltransferase 3	585	5	3.06e -60	87.69
Pdmh01001337.1 .g11984.t1	Putative glycine-rich cell wall structural protein 1	348	5	2.99e -07	92.57
Pdmh01001394.1 .g12216.t1	Protein ctr9 homolog	225	5	4.67e -14	95.15
Pdmh01001548.1 .g12954.t1	Ribosome maturation factor rimm	531	3	3.35e -15	55.95
Pdmh01002283.1 .g15759.t1	Thiosulfate sulfurtransferase 16, chloroplastic isoform x1	276	5	2.25e -56	74.44

Pdmh01002363.1 .g16002.tl	-lysine n-methyltransferase eef2kmt	813	5	4.17e -11	96.63
Pdmh01003202.1 .g18317.tl	Hypothetical protein cocnu_08g002050	591	5	4.32e -66	84.63
Pdmh01003596.1 .g19249.tl	Transducin beta-like protein 3	435	5	1.29e -68	76.24
Pdmh01003777.1 .g19621.tl	Uncharacterized protein loc105055396	384	5	1.06e -61	75.08
Pdmh01004420.1 .g20786.tl	Protein haiku1	858	5	8.91e -18	71.97
Pdmh01005060.1 .g21787.tl	Translation initiation factor eif-2b subunit epsilon	225	5	5.16e -17	94.21
Pdmh01005286.1 .g22115.tl	Protein exordium-like 3	312	5	9.24e -16	59.39
Pdmh01005445.1 .g22310.tl	Multiple myeloma tumor-associated protein 2	567	5	1.67e -69	98.02
Pdmh01007041.1 .g24023.tl	Leaf rust 10 disease-resistance locus receptor-like protein kinase-like 1.2 isoform x2	411	5	3.11e -28	77.68
Pdmh01007661.1 .g24542.tl	Pc-esterase	423	5	1.39e -26	97.74
Pdmh01009935.1 .g26049.tl	Burp domain-containing protein 6	216	5	3.92e -45	97.68
Pdmh01010803.1 .g26480.tl	Alpha-1,6-mannosyl-glycoprotein 2-beta-n-acetylglucosaminyltransferase	921	5	0	84.96
Pdmh01015911.1 .g28162.tl	Aspartate--trna ligase, chloroplastic/mitochondrial isoform x1	316	5	1.69e -09	77.68
Pdmh01016699.1 .g28343.tl	Uncharacterized protein loc105042776	237	1	5.59e -05	82.76
Pdmh01018890.1 .g28756.tl	Hypothetical protein c4d60_mb05t14830	225	5	6.21e -09	97.78
Pdmh01045520.1 .g29893.tl	Polyadenylation and cleavage factor	127	5	9.67e -10	88.94

Annexure V: Primer sequences

PRIMER	SEQUENCE
cocos_d 1	5' ATCCGACCAGCCTGCATC 3' 5' GTATCGCCCAGCGGAATG 3'
cocos_d 2	5' GGAACGCGATGAACATGATA 3' 5' CTGCTGCACCCAGTTATCC 3'
cocos_d 3	5' ACCAACGTGTGGGAAAAGCTA 3' 5' CCAGCAGTTTGCCAGTTT 3'
cocos_d 4	5' CGAAGTGGTGGTGGAAACC 3' 5' CCGCGAAAGCCAATATCC 3'
cocos_d 5	5' GCTTTTGCACCCTGATTAGC 3' 5' TTAATCAGCGGATGATGCTG 3'
cocos_d 6	5' ATGGAACTGACCGGCATTAG 3' 5' GCCCCAGGTTTCTCCTG 3'
cocos_d 7	5' CGAATATGTGGTGTGGATG 3' 5' ATCATGGGTCAGGCTCTGG 3'
cocos_d 8	5' GTTTAGCAACCGCCTGGATA 3' 5' GGCAGATCGGTGGTAATCAG 3'
cocos_d 9	5' GAAGAATGGGAAGCAAAGA 3' 5' CTTTCGGTTCGCCCAGAT 3'
cocos_d 10	5' GAAAAGCAGCAAAGCGAAG 3' 5' GTGCTCGCCTGGTTATGATG 3'
cocos_t 11	5' ACCGTGCCGTATAGCACCTA 3' 5' CTGCAGAAAGCCATGTTTCA 3'
cocos_t 12	5' GCGAAGAAATTGGCAAAGAA 3' 5' CAATGGTGTTCAGCTATGG 3'
cocos_t 13	5' GAAAAGCCGCGTGGAAAGAA 3' 5' GCCGCAATGGTCACTTC 3'
cocos_t 14	5' CACCTTTTATGGCAACGTGA 3' 5' GTCGCATCCACCACTTCTTT 3'
cocos_t 15	5' CCGAAGAAGAAGAAGAAATGGA 3' 5' GAATCAGGCCTTTCATCTGG 3'
cocos_t 16	5' ACCAAATATGCGACCCTGAG 3' 5' CTCACCACCCACGGTTCTT 3'
cocos_t 17	5' AGACCATTCTGCCGCATC 3' 5' GCGGAAAGCAGCGTTAAT 3'
cocos_t 18	5' GAGCGGAGCCTGTTTCT 3' 5' AGCCACATAAAGCTGCTGGT 3'
cocos_t 19	5' GGCGGCGATGTTTACCAC 3' 5' GTTCGGCAGAATGCTGGTAT 3'
cocos_t 20	5' GCCGTATAGCTGCAACACC 3' 5' CAGCACCTGATAGCGAAAGC 3'
cocos_21	5' GTCTCTCGACCGCTGTCTCT 3' 5' TAGATCTGCCACGCTTTTCC 3'

cocos_22	5' GTCTCGCTGTCCCTTTTGG 3' 5' CTAGGGCCGAGGGAGGTAG 3'
cocos_23	5' AGGGAGATGTCGCCTGGA 3' 5' ACACCACACCCCTCCTGTC 3'
cocos_24	5' AGAGGAGGGACGCACGAT 3' 5' AGGATATAGGCGGGAGAGGA 3'
cocos_25	5' AGCGGAGGAGAAGGCTGAG 3' 5' CAAATCTCCGGTGAATAGGG 3'
cocos_26	5' GTTTGGGGTTGCGGTCAT 3' 5' CTATTCTCCCTGCCCTTGG 3'
cocos_27	5' ATGGCTTCCCCTGCACTCT 3' 5' GGTTGGTTTTCATCGAAGC 3'

Annexure VI: Potential polymorphic microsatellite regions

Sl.No.	Ecotype	Chromosome No.	StartPos	EndPos	Motif	Repetitions
1	Tall	1	2693056	2693091	TA	18
	Dwarf	1	2693120	2693175	TA	28
2	Tall	1	2974791	2974840	TA	25
	Dwarf	1	2975729	2975796	AT	34
3	Dwarf	1	3957627	3957810	AT	92
	Tall	1	3957807	3957836	AT	15
	Dwarf	1	3957835	3957910	AT	38
4	Dwarf	1	3957981	3958030	AT	25
	Dwarf	1	4801405	4801436	AT	16
	Tall	1	4801819	4801878	AT	30
5	Dwarf	1	7720908	7720939	GT	16
	Tall	1	7721171	7721214	TC	22
6	Tall	1	8207303	8207346	AT	22
	Dwarf	1	8207877	8207906	CT	15
7	Dwarf	1	9174293	9174396	TA	52
	Tall	1	9174464	9174501	AT	19
8	Tall	1	13139853	13139886	TA	17
	Dwarf	1	13139857	13139892	TC	18
9	Tall	1	13908859	13908892	AT	17
	Dwarf	1	13908863	13908912	AG	25
10	Dwarf	1	20837023	20837082	TGTA	15
	Tall	1	20837248	20837289	AT	21
11	Tall	1	20837497	20837590	TA	47
	Dwarf	1	20837560	20837609	TG	25
12	Tall	1	22482926	22482957	TC	16
	Dwarf	1	22483032	22483269	AT	119
13	Dwarf	1	22579377	22579432	AT	28
	Tall	1	22579933	22579966	TA	17
14	Dwarf	1	39819242	39819289	AT	24
	Tall	1	39819386	39819421	AG	18
15	Tall	1	40715352	40715449	AT	49
	Dwarf	1	40715591	40715628	AT	19
16	Tall	1	44103469	44103508	AT	20
	Dwarf	1	44103622	44103651	TA	15
	Dwarf	1	50784043	50784078	TG	18
17	Tall	1	50784081	50784118	TA	19
	Dwarf	1	50784127	50784190	TA	32
18	Dwarf	1	53704344	53704401	AT	29
	Tall	1	53704888	53704917	AT	15
19	Tall	1	54149878	54149947	AT	35

	Dwarf	1	54150100	54150133	GT	17
20	Tall	1	58172413	58172450	AG	19
	Dwarf	1	58172840	58172869	AT	15
21	Tall	1	159659839	159659932	AT	47
	Dwarf	1	159660153	159660212	ACAT	15
22	Dwarf	1	179201848	179201891	AT	22
	Tall	1	179202065	179202116	TA	26
	Dwarf	1	179202346	179202375	TA	15
23	Tall	1	184670210	184670247	TC	19
	Dwarf	1	184670504	184670577	AT	37
24	Dwarf	1	193396899	193396966	TA	34
	Tall	1	193397245	193397284	TA	20
25	Dwarf	1	193934983	193935046	AT	32
	Tall	1	193935470	193935517	TA	24
26	Dwarf	1	194685992	194686027	TC	18
	Tall	1	194686065	194686124	AC	30
27	Dwarf	1	195665134	195665201	AT	34
	Tall	1	195665384	195665473	AT	45
28	Tall	1	199102561	199102592	TC	16
	Dwarf	1	199102710	199102745	TC	18
29	Dwarf	1	204830212	204830253	AT	21
	Tall	1	204830296	204830337	TA	21
30	Tall	1	211825412	211825443	GA	16
	Dwarf	1	211825413	211825448	TC	18
31	Dwarf	1	211578001	211578056	TA	28
	Tall	1	211578437	211578466	AG	15
32	Dwarf	2	4860006	4860129	AT	62
	Tall	2	4860032	4860063	TC	16
	Dwarf	2	4860142	4860173	AT	16
33	Dwarf	2	6769813	6769842	TC	15
	Tall	2	6770015	6770060	GT	23
34	Dwarf	2	6817087	6817144	AT	29
	Tall	2	6817326	6817387	TA	31
35	Dwarf	2	7807898	7807951	TC	27
	Tall	2	7807979	7808018	TC	20
	Dwarf	2	20796708	20796827	AT	60
36	Tall	2	20796799	20796848	AC	25
	Dwarf	2	20796896	20796925	TA	15
	Dwarf	2	20797012	20797097	TA	43
	Tall	2	20797023	20797056	AC	17
37	Dwarf	2	20797109	20797246	AT	69
	Tall	2	20797169	20797200	AG	16
	Dwarf	2	20797249	20797356	AT	54

38	Tall	2	27386532	27386585	CT	27
	Dwarf	2	27386546	27386595	AT	25
39	Dwarf	2	27518237	27518278	AT	21
	Tall	2	27518378	27518437	CT	30
40	Dwarf	2	34911771	34911838	AT	34
	Tall	2	34911774	34911803	CT	15
41	Tall	2	43230376	43230407	TC	16
	Dwarf	2	43230487	43230528	AT	21
42	Dwarf	2	126817500	126817543	TA	22
	Tall	2	126817778	126817813	TC	18
43	Tall	2	154272500	154272557	TC	29
	Dwarf	2	154272532	154272629	AT	49
44	Dwarf	2	154700267	154700334	AT	34
	Tall	2	154700759	154700796	AG	19
45	Tall	2	159851173	159851218	AC	23
	Dwarf	2	159851205	159851268	AT	32
46	Dwarf	2	161663703	161663742	TA	20
	Tall	2	161664069	161664132	AT	32
47	Tall	2	166454520	166454549	AT	15
	Dwarf	2	166454925	166454970	GA	23
48	Dwarf	2	166739404	166739447	TA	22
	Tall	2	166739575	166739656	AT	41
49	Dwarf	2	166739735	166739818	TA	42
	Dwarf	2	167076882	167076931	AG	25
50	Tall	2	167077290	167077321	TA	16
	Dwarf	2	167077649	167077684	CT	18
51	Dwarf	2	167215269	167215304	AG	18
	Tall	2	167215743	167215772	CT	15
52	Tall	2	173808313	173808376	AT	32
	Dwarf	2	173808578	173808617	AG	20
53	Dwarf	2	184971783	184971814	TA	16
	Tall	2	184971811	184971856	TC	23
54	Tall	2	185665208	185665241	AG	17
	Dwarf	2	185665384	185665433	AT	25
54	Dwarf	2	186514990	186515019	AG	15
	Tall	2	186515485	186515518	AT	17

**COMPARATIVE GENOME ANALYSIS IN COCONUT (*Cocos nucifera* Linn.)
AND MARKER DEVELOPMENT FOR DISTINGUISHING TALL AND
DWARF COCONUT TYPES**

By

SHRI HARI PRASAD

(2019-11-209)

ABSTRACT OF THE THESIS

Submitted in partial fulfilment of the requirements for the degree of

Master of Science in Agriculture

Faculty of Agriculture

Kerala Agricultural University, Thrissur



DEPARTMENT OF PLANT BIOTECHNOLOGY

**CENTRE FOR PLANT BIOTECHNOLOGY AND MOLECULAR BIOLOGY
COLLEGE OF AGRICULTURE**

VELLANIKKARA, THRISSUR – 680656

KERALA, INDIA

2021

ABSTRACT

Coconut is an important oil nut crop in the humid tropics of the world. Because of its inevitable uses as food, drink, fuel, and so on, this palm is known as ‘the tree of life or *Kalpavriksha*’. There are two ecotypes in coconut, tall and dwarf. Development of dwarf varieties with high yield, longer life span, field tolerance to biotic and abiotic stresses and good kernel and oil recovery, is the major breeding objective in this crop. Breeding attempts for dwarf palm stature are crippled with the non-availability of a precise methodology to identify the dwarf lines at the early plant stage itself. Development of molecular markers linked with this trait will enable the marker assisted selection for dwarf palms.

The present study entitled “Comparative genome analysis in coconut (*Cocos nucifera* Linn.) and marker development for distinguishing tall and dwarf coconut types” was undertaken at Centre for Plant Biotechnology and Molecular Biology, College of Agriculture, Thrissur, during 2019 to 2022, with the objective to identify the differential genes and genomic regions among the tall and dwarf coconut genotypes through comparative whole genome sequence analyses and to develop molecular markers for distinguishing tall and dwarf coconut types.

The coconut genome assemblies and raw reads were retrieved from various databases, quality of the assemblies and raw reads analyzed, raw reads trimmed and assembled using SOAPdenovo2, ABySS and Velvetoptimiser. Repeat masking was carried out on coconut genome assemblies, using RepeatMasker employing Dfam and RepBase as libraries. Since the library yielded insufficient masking percentage, *de novo* repeat library was prepared for the genome assemblies using RepeatModeller. The repeat libraries thus obtained have been merged to get a comprehensive and exhaustive repeat library for coconut and was used to perform repeat masking. Further, the efficiency of the combined repeat library for repeat masking in other palms was checked and found effective. The library was made publicly available at <https://kau.in/repeat-libraries>. Gene prediction was carried out for the repeat masked genomes using AUGUSTUS, a eukaryotic gene prediction tool.

Comparative genome analyses were carried out by NCBI BLAST+, unique sequences obtained for dwarf and tall genomes were identified and extracted. The

extracted sequences were annotated using online BLAST and functional annotation was carried out using BLAST2GO. Reverse BLAST was performed to ensure that the sequences thus obtained are unique and PCR primers were designed for the sequences.

Leaf samples were collected from 10 coconut accessions, five each of the tall and dwarf types, from the parent palms at RARS Plicicode. DNA extracted from these accessions were screened using 27 primer combinations. Ten primer sets have amplified the markers in West Coast Tall only while the primer set Cocos_22 amplified the marker in West Cost Tall and Jawan Giant. Primer set Cocos_21 has amplified the markers in three tall samples, West Coast Tall, Jawan Giant, New Guinea and one dwarf sample Chowghat Green Dwarf. It is already recorded that Chowghat Green Dwarf is not a true dwarf ecotype but belongs to the semi-tall types.

Marker Cocos_21 was successful in marker generation in tall accessions, except where this highly quantitative trait is governed by other major QTL. Further, in order to establish a universal marker linked to the height of the coconut palm, more whole genome sequences of tall, dwarf and intermediate ecotypes are required and insights from the sequence data (whole genome and transcriptome) could help in more refined classification of the palms. After validation with other tall and dwarf cultivars, the marker identified in this study may be used in marker assisted selection of coconut.