11

Biotechnology and Genomics in Palms

R. Manimekalai, M.K. Rajesh and K.P. Manju

1. Introduction

The *Arecaceae* (Palmae or palm tree), a family of ~ 202 genera and ~ 3000 species, is one of the best symbolised angiosperm families in the fossil record. It is the first and sole modern family of monocots of flowering plant in the order Arecales. The leaf and stem fossils appeared in the fossil record as early as 80-85 million years ago. This family is divided into six subfamilies, further into 14 tribes and 36 subtribes and finally into around 200 genera. Palms are the largest and diverse family of monophyletic group of plants and are among the best known and extensively cultivated plant families and also are perennials and distinguished by having woody stems. It is distributed throughout tropical and subtropical areas of the world, but it is most diverse in highly threatened moist tropical forest habitats.

1.1. Current phylogenetical status of palm family

Phylogeny implies the evolutionary relationship among a set/group of organisms; the pattern of lineage branching produced by their evolutionary history. The phylogenetic tree is a structure in which species are arranged on branches that link them according to their relationship and/or evolutionary descent. A phylogenetic tree is composed of ***nodes***, each representing a taxonomic unit (species, populations, individuals), and ***branches***, which define the relationship between the taxonomic units in terms of descent and ancestry. The number of changes that have occurred in the branch is represented by the branch length. The time of divergence between two organisms since they last shared a common ancestor can also be recognized via the phylogenetic tree. With the aid of sequences, the genealogical ties between organisms can be established as a phylogenetic tree. Thus the closely related organisms have similar sequences and more distantly related organisms have more dissimilar sequences.

The palm family is resolved as monophyletic with a bootstrap support of 100%. The use of DNA sequence data as a source of character information to resolve evolutionary relationships has led to many refinements in palm classifications. All current groups are monophyletic. Evolutionary trees based on molecular data also provide an independent framework for the reconstruction of the ancestral states of structural characters.

1.2. Range of chromosome number and genome size

The term "genom(e)" is derived by joining 'gene' and 'chromosome' and genome is defined as the complete gene compliment or total DNA amount per haploid chromosome set. Genome is the genetic blue print of the hereditary information of an organism, which comprises many functional regions, non-coding sections and the vast unexplored areas of the nucleic acids. Revealing the genetic knowledge entitled in the genomes can increase our understanding of species, which could substantially improve their health and yield. In the case of crops, with such enhanced productivity, it could be sustainable solution to fulfilling the world's need for a wide variety of products.

Different palms have undergone very diverse patterns of genome size and chromosome evolution, giving rise to the genomic diversity. Genome size and total chromosome number (2n) are not at all interrelated with each other, while chromosomes can vary in size without any change in DNA depending on the nutrient status of the plant. Still, several studies have highlighted the potential for using chromosome data as proxies for genome size. It is clear that different palms species follow characteristic modes of genome size and chromosome data. The highest chromosome number so far recorded is 2n=596, for the palm *Voanioala gerardii*.

In the family *Arecaceae*, genome size data are available for 89 species in 57 of the 200 recognized genera, representing all five subfamilies (Table11.1). C-values range from 0.95 pg in the diploid *Phoenix* (*Coryphoideae*) (2*n* = 36) to 30.0 pg in the highly polyploid *Voanioala gerardii* (*Arecoideae*) with 2*n* = 596. This reflects cytological data showing that polyploidy is rare in palms with just four polyploid species reported to date, two tetraploids (*Arenga caudata*, 2*n* = 64, and *Rhapis humilis*, 2*n* = 72) and two rare, monotypic genera of high ploidy, *Jubaeopsis caffra* from South Africa (2*n = 160–200)* and *Voanioala gerardii* from Madagascar.

Studies have proven that different chromosome numbers have evolved mainly through dysploidy (an abnormal ploidy level) due to the broadly similar DNA amounts in the related genera differing in chromosome number. It is, however, clear that changes in genome size can occur with no alteration of chromosome number leading to related species having significantly different sized chromosomes.

Coconut (*Cocos nucifera*), arecanut (*Areca catechu*), oil palm (*Elaeis guineensis*) and date palm (*Phoenix dactlifera*) are the major economically important crop members planted in the *Arecaceae* family. Two of these, date palm and oil palm got their genetic package revealed paving the path to a vast area of genetic research. In this chapter the biotechnology of coconut and oil palm are covered along with genomics in date palm and oil palm.

Table 11.1: The known chromosome number and ploidy level of some of the palms

Palm	*Chromosome number*	*Ploidy level*	*C value*	*Genome size (Mbp)*
Acrocomia	2n=30	2	3.38	-
Archontophoenix	2n=32	2	4.75	-
Areca catechu	2n=32	2	6.1	-
Arenga	2n=64	2		-
Bactris	2n=30	2	4.08	-
Beccariophoenix	2n=36	2	1.8	-
Bentinckia	2n=32	2	2.78	-
Bismarckia	2n=36	2	2.03	-
Borassus	2n=36	2	8.6	-
Brahea	-	2	1.06	-
Calamus	-	-	1.45	-
Calyptrogyne	2n=28	2	3.43	-
Caryota	2n=34	2	6.6	-
Ceroxylon	-	-	3.85	-
Chamaedorea	-		4.08	-
Chuniophoenix	2n=36	2	1.55	-
Cocos nucifera	2n=32	2	3.55	3600
Coccothrinax	2n=37	3	7.45	-
Daemonorops	-	-	1.83	-
Desmoncus	2n=30	2	3	-
Dypsis	-	-	1.53	-
Elaeis	2n=32	2	1.88	1800
Euterpe	2n=36	2	5.3	-
Geonoma	2n=28	2	3.63	-
Guihaia	2n=36	2	5.95	-
Hyphaene	-	-	3.4	-
Iriartea	-	-	12.28	-
Johannesteijsmannia	2n=34	2	1.63	-
Jubaea	2n=32	2	2.55	-
Jubaeopsis	2n=40	2	-	
Latania	2n=28	2	3.5	-
Lepidocaryum	2n=30	2	4.1	-
Licuala	2n=28	2	1.5	-
Livistona	2n=36	2	1.65	-
Loxococcus	2n=32	2	3.58	-
Masoala	2n=32	2	2.68	-
Mauritia	2n=30	2	4.73	-
Mauritiella	2n=30	2	7.28	-
Medemia	-	-	3.63	-
Normanbya	-	-	4.85	-
Nypa	2n=34	2	1.18	-
Oenocarpus	2n=36	2	3.93	-
Phytelephas	2n=36	2	0.98	-
Phoenix dactlifera	2n=36	2	0.95	658

(*Contd.*)

Palm	*Chromosome number*	*Ploidy level*	*C value*	*Genome size (Mbp)*
Pseudophoenix	2n=34	2	2.83	-
Pinanga	2n=32	2	6.7	-
Ravenea	2n=30	2	2.18	-
Rhapis	2n=72	2	4.8	-
Roystonea	2n=36	2	4.8	-
Sabal	2n=36	2	2.25	-
Salacca	2n=28	2	1.68	-
Socratea	2n=36	2	4.55	-
Sommieria	2n=34	2	5.8	-
Syagrus	2n=32	2	2.05	-
Trachycarpus	2n=37	3	5.55	-
Voanioala	2n=596	2	30	-
Washingtonia	2n=36	2	1.55	-
Wendlandiella	2n=28	2	2.85	-

2. Introduction to genomics

Genomics is the sub-discipline of genetics pertaining to the study of the structure and function of the entire genome of a living organism, and the genome- scale technologies and their applications in the major field of life sciences. It is aimed to understand the role of genes, their expressions, and their product characteristics and thus which is devoted to mapping, sequencing and functional analysis of genomes. The word 'genomics' was first used by Thomas Roderick in 1986. This branch is focused on sequencing the DNA in an organism to obtain a total picture, and then figuring out specific genes in that sequence that could be of interest. Worldwide, efforts to sequence genomes had begun in 1990s, and through sequencing, the entire DNA pattern in every chromosome of an organism are revealed. Genomes are the entire set of hereditary information of an organism and the genomes of various species are distinctly different and for many species from bacteria to humans, genomes were sequenced. Within a species, genetic variation may be minimal, but yet fascinating, it can explain certain traits or tendencies. The position of genomics in the field of scientific research is found to be demanding over decades because to carry out high speed genome level analysis, data about its genetic information must be large, contiguous and flanking to chromosomes or in turn genetic information should be completely sequenced.

3. The two explored palm genomes – date palm and oil palm

At the genomic level, palms are remarkably diverse, with striking variation at many levels ranging from gene sequences through to the number of chromosomes per genome, and the amount of DNA per genome (genome size). Since palms are having relatively long generations and times-to production, the unlocking of genomic information will be useful for the researchers to identify the genetic mechanism mainly to 1) improve crop yield, 2) disease resistance and 3) growth responses to environmental factors. The chloroplast genome has long been a focus of

research in plant molecular evolution because of its small size, high copy number, and conservation and extensive characterization at the molecular level. Chloroplast genome size ranges from 120 kb to 210 kb, with most being 150 kb in size. Currently there are 132 completed chloroplast genomes available at National Centre for Biotechnology Information (NCBI), www.ncbi.nlm.nih.gov.

As decoding the nuclear genome is more tedious, currently two palm genomes are only available, i.e., *Phoenix dactylifera* (date palm) and *Elaeis guineensis* (oil palm). Yet another economically important palm is coconut (*Cocos nucifera*), whose genome still remains unexplored.

3.1. Date palm (*Phoenix dactylifera* L.)

Date palm (*Phoenix dactylifera* L.), also called the "tree of life" in the bible, is an evergreen, arborescent, dioecious, perennial, highly heterozygous, monocotyledonous, diploid (2n=36) plant with long generation time and an erect tall palm with pinnate and fronds held upright in "feather duster"-like arrangement; usually single trunked. Date palm is probably the second most well-known palm after the coconut. It is one of the most important fruit trees growing in the Arabian world and some neighboring counties and represents a good cash crop for many farmers. The Eastern region of Saudi Arabia has been found as a predominant growing area of date palm for centuries with more than 4 million trees located. The date palm is believed to have originated in the lands around the Persian Gulf and in ancient times was especially abundant between the Nile and Euphrates rivers. The date plam is the most important agricultural crop in the area and provides food and income to the majority of the inhabitants. It ranks first among all the crops due to its high nutritional and economic value to oasis agriculture and creates favorable conditions for improving secondary crops.

3.1.1. Date palm genome sequencing

The sequencing of the date palm nuclear genome was done by the research team at Weill Cornell Medical College in Qatar. The predicted genome size is ~650Mbp, a scaffold N50 of ~30kbp with most ordered gaps being extremely short, ~57,000 scaffolds, 3.5 million novel high quality SNPs between 9 genomes and the reference Khalas, ~25,000 gene predictions (excluding transposable elements), 38% GC in the nuclear genome, 381Mb of assembled sequence representing ~90% of genes and 60% of the genome sequence (remaining unassembled sequence is mostly highly repetitive) and draft chloroplast gene sequences.

3.2. Oil palm (Elaeis guineensis and *Elaeis oleifera)*

Oil palm is a tropical palm tree, versatile as oil seed next to soybean and thus it is the highest oil yielding tropical palm tree in the world. It belongs to the species *Elaeis guineensis* from tropical western Africa and a related species *E. oleifera* from South America. Biotechnological approaches play an enhancing role in breeding strategies for oil palm.

Research into the oil palm complete genome began in 2003. This step towards ultimate profiling of the oil palm by complete sequencing of genome was put forth by Asiatic Centre for Genome Technology Sdn Bhd (ACGT), and Synthetic Genomics Inc. (SGI), a privately held company. The completion of a first draft assembly and annotation of the oil palm genome was announced

in 2007. The genome size of oil palm is 1.8 billion bp, about four times the size of the rice genome and two thirds the size of the maize genome. But the whole genome sequence is not yet available in the public domain. Presently, National Genome Project of Oil Palm is also under progress under the aegis of MPOB, Malaysia.

4. Coconut (*Cocos nucifera* L.)

The coconut palm (*Cocos nucifera* L.; 2n=32), one of the ten most important plants on the planet Earth, is a perennial oleaginous plant, cultivated in all the tropical coastal regions of the world and it is a sole monotypic species in the genus *Cocos* belonging to the subfamily Cocoideae which includes 27 genera and 600 species. The two main ecotypes that exists in this species are, the Tall palms, *C. nucifera* typica (a cross pollinating crop) and the Dwarf palms, *C. nucifera* nana (a self pollinating crop). Due to the various aspects of by- product utilization, coconut is further more an invaluable plant for the rural community and it has been termed as "tree of life''. The centre of diversity of coconut is in the South-East Asia and Malaysia. According to FAO statistics, 58.1% of the six million hectares under this crop in the world is in the South-East Asia and another 24.6% of the area occurs in the South Asian countries.

4.1. In vitro propagation in coconut

As coconut is a highly heterozygous, long juvenile phased palm, breeding for crop improvement is a tedious and long term task. Tissue culture, a promising technique for vegetative propagation of coconut palms, offers a means of cloning improved plant material within a short period of time. Although coconut is the most recalcitrant species to *in vitro* culture, several tissue culture studies were begun in an effort to grow seedlings *in vitro* and to propagate coconuts vegetatively. The production of high quality and uniform planting material of coconut plants propagated from vegetative parts that can be multiplied on a year-round basis under disease-free conditions creates new opportunities in global trading, farmers, nursery owners, and improved rural employment.

The responses of zygotic tissues are more consistent than other inflorescence and leaf tissues. Thus the use of suitable explant and the fine tuned culture medium can diminish the genotypic effect on *in vitro* responses of coconut and to accelerate the growth of clonal plants *in vitro* and early *ex vitro*. Thus out of the various explants tested, plumules have shown the best response to *in vitro* culture.

Clonal plantlets cultured from plumular tissues excised from mature zygotic embryos are a significant system for developing a clonal propagation method for coconut. The efficiency of somatic embryogenesis in coconut can be increased through secondary somatic embryogenesis and multiplication of embryogenic callus and result in drastic increase in the yield of somatic embryos.

4.2. Use of molecular markers for characterization of coconut

The genetic diversity information about a crop would help in many breeding programmes. Coconut hybrids, involving talls and dwarfs, showed 130% increase in yield compared to the best parents. In coconut, naturally occurring super palms (high yielding pre potent palms) are

also available. But the improvement of this crop is hindered by several issues such as narrow genetic base, long juvenile phase, reproductive biology, highly heterogeneous nature of the phenotypes involved and long interval between the generations. Also, the monotypic nature of the coconut will, however, be a limiting factor in breeding coconut varieties suitable for cultivation in diverse ecological situations. The need for alternatives to overcome these genetic constraints for the coconut genetic improvement has long been recognized. The field gene bank of CPCRI holds the largest collection of germplasm accession of coconut. Several approaches have been considered *viz.* morphological, physiological, biochemical and molecular markers for the characterization of germplasm.

A tool that has proved very valuable for the genetic improvement of many crops is the use of molecular markers. The approach to identify genetic markers assists conventional breeding. Thus, the incorporation and the use of molecular marker techniques may improve breeding efficiency, in genetic diversity studies, to categorize genetic groups of different accessions, characterization and organization of germplasm, linkage mapping and identification of QTLs for marker-assisted selection (MAS). These markers would also help in trimming down the selection of crosses to be tested in a coconut hybrid breeding programmes. Identification of molecular markers linked to useful traits will strengthen and hasten breeding programmes, reduce costs, improve the efficiency and reduce the length of the choice cycles.

To decode genetic outline by detecting polymorphisms, many techniques have been used widely, comprising biochemical markers and DNA markers which include proteins (mainly isozymes), polyphenols, restriction fragment length polymorphism (RFLP), randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), inverse sequence-tagged repeat (ISTR), simple sequence repeats (SSR) and inter simple sequence repeats (ISSR).

5. Markers

5.1. Enzyme markers

Isozymes are neutral and are often expressed codominantly making the discrimination possible between homozygote and heterozygote. Early plant molecular marker systems focused on gene products such as isozymes, which are common enzymes expressed in plants. Characterization of different coconut populations were performed using biochemical markers like isozymes, polyphenols, carotenoids and leaf proteins. In order to carry out these investigations various parts of the coconut palm like leaf, endosperm, pollen, inflorescence etc. have been utilized for the study of isozymes. The isozymes analysis measures the presence or absence of different forms of an enzyme that differs from allelic variation at the same locus. The technique is simple, inexpensive and well characterized and the results are presented in a co-dominant fashion. But the technique is delimited by the requirement of fresh samples and the analysis only resolves a small number of loci per assay. The characterization of different coconut cultivars and hybrids showed the highest intrapopulation variation in coconut.

5.2. DNA markers

Despite of the wide usage of protein markers it was diminished by the fact that the information derived from isozyme markers has been inadequate by the number of loci for which assays are available. This is because of the fact that most plant cultivars are genetically very similar that isozymes in marker assisted selection (MAS) do not produce a great amount of polymorphism. This lead to the outperformance of DNA markers.

5.2.1. Restriction Fragment Length Polymorphism (RFLP)

RFLP analysis is an application of the Southern hybridization procedure. This marker is specific to a single clone/restriction enzyme combination and it exploits variations in homologous DNA sequences. RFLP outperforms the other traditional marker systems as 1) they are codominant and unaffected by the environment; 2) any source DNA can be used for the analysis; and 3) many markers can be mapped in a population that is not stressed by the effects of phenotypic mutations. Lebrun *et al.* (1998) applied the RFLP technique for the first time in coconut to assay the genetic diversity in 100 palms located in diverse geographical conditions. The results from the RFLP studies could furnish the historical dispersion reports of coconut. Considerable ecotype diversity was recorded via many RFLP studies from Far East and Pacific regions, which are considered putative area of origin of the coconut palm. The West African ecotypes were related to the Indian and Sri Lankan ecotypes suggesting recent extension of the palm along the length of the Atlantic Coasts of Africa via nuts originating from the Indian Ocean. The tall ecotypes exhibited the higher polymorphism compared to dwarfs.

5.2.2. Random Amplification of Polymorphic DNA (RAPD)

Randomly Amplified Polymorphic DNA is marker technique based on the differential PCR amplification of a sample of DNAs from short oligonucleotide sequences. They are the amplification products of anonymous DNA sequences using single, short and arbitrary oligonucleotide primers, and thus do not require prior knowledge of a DNA sequence. Low expense, efficiency in developing a large number of DNA markers in a short time and requirement for less sophisticated equipment has made the RAPD technique valuable although reproducibility of the RAPD profile is still the centre of debate. The main RAPD application covers the identification of cultivars and clones, genetic mapping, marker-assisted selection, population genetics, and molecular systematics at the species level.

RAPD studies were initiated in 1990s. A few RAPD markers unique to specific populations were identified. Several other studies also have stated the utility of RAPD techniques to assess the genetic diversity among coconut accessions. It was observed that the coconut populations collected from seeds of open pollinated plants serves as a basis for the occurrence of greater variability of coconut population. Cardena *et al.* (2003) detected markers that are potentially linked to lethal yellowing diseases in their study to identify RAPDs associated with resistance to lethal yellowing disease of the coconut palm using three coconut populations *viz.*, susceptible West African Tall, resistant Malayan Yellow Dwarf and a resistant population of Atlantic Tall palms.

The genetic diversity and relatedness among three dwarf coconut accessions *viz*., Malayan Yellow Dwarf (MYD), Kulashekaram Yellow Dwarf (KYD) and Andaman Yellow Dwarf (AYD) encountered using RAPD markers showed a close association of the two Indian yellow dwarfs (AYD and KYD) with the exotic yellow dwarf (MYD) indicating the likelihood of AYD and KYD evolving from a common progenitor, the MYD.

A major drawback of RAPD markers in population genetic studies of out breeding organisms is that they are dominant. Thus gene frequency estimates for such loci are necessarily less accurate than those obtained with codominant markers such as allozymes and RFLPs. RAPD technique is found out to be sensitive to slight changes to reaction conditions, which may inversely affect the banding pattern output in different experiments.

5.2.3. Inverse Sequence-Tagged Repeat (ISTR)

Inverse sequence-tagged repeat (ISTR) is used to detect DNA polymorphism and this technique was first started in coconut and later on extended to other plant and then to animal genomes. ISTR methodology is a PCR -based DNA marker technology developed on the basis of mobile genetic elements. The ISTR technique uses the ubiquitous presence of reverse transcriptase sequences of these elements. ISTR is found to be a powerful tool for genotype identification

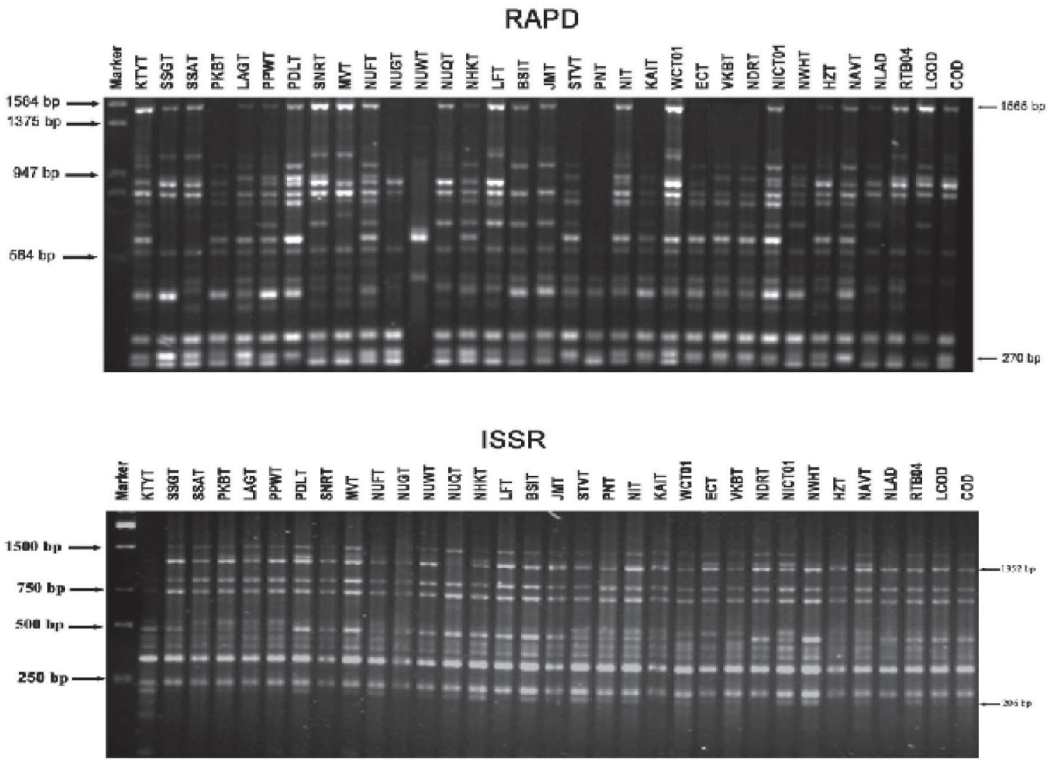

Fig. 11.1: RAPD and ISSR marker profile of coconut germplasm accessions

and differentiation. ISTR analysis proposed by Rohde *et al*. (1995) is a vital technique used for the detection of DNA polymorphism in which primers complementary to repetitive, copia-like sequences in the coconut genome were used to amplify a large number of genetic loci with an abundance of polymorphisms occurring among a set of selected coconut genotypes from various regions of the world. These studies also provided evidence for the existence of truncated, copia-like repetitive sequences in the coconut genome indicating that retro-elements may have played a role in the generation of genetic diversity in coconut.

5.2.4. Amplified Fragment Length Polymorphism (AFLP)

Amplified fragment length polymorphism (AFLP) technology represents a resourceful combination of RFLP analysis and PCR. This high-multiplex method for DNA profiling (Vos *et al*., 1995) intends to generate a large number of polymorphic genetic loci. AFLP technology is applicable to all organisms without previous sequence information, and generally results in highly informative fingerprints. It rapidly became one of the most popular and powerful approaches to detect DNA polymorphisms.

Many features are found to be unique for AFLP technology *viz*., 1) PCRs can be adoptable with random primers, as no prior sequence information is needed, 2) high robustness and high reproducibility, and 3) a limited set of AFLP primers can be combined to yield a large set of primer combinations, each producing a unique set of amplified fragments. AFLP has been used in coconut for analysis and diversity and shown high variability in tall forms (*typica)*, than intermediate (*aurantica*) and dwarf (*nana*) forms. *Aurantiaca* group was more similar to the dwarf than the tall group. In addition, putative duplicate accessions were identified in the *Aurantiaca* group.

5.2.5. Simple Sequence Repeats (SSR)

Simple sequence repeats (SSRs), or short tandem repeats (STRs) also known as microsatellites consists of tandemly reiterated (2-8 bps), size polymorphic short DNA sequences. They are co-dominant markers and so can be used to study gene duplication, deletion, genetic diversity, genetic and comparative mapping. Different genotypes differ in number of repeats present and in turn in the size of the repeated region. Flanking primers are designed targeting the repeat regions and used in PCR reactions for site specific amplification of the microsatellites, thereby producing sequence-tagged microsatellite markers.

In plants, the presence of microsatellites was first revealed by RFLP fingerprinting with oligonucleotide probes. Initial studies suggested a lower frequency of microsatellites in plants compared to animals, later on more surveys based on large data sets explored the greater abundance of microsatellites in plants than previously anticipated. Many studies reported that -tri, -tetra and -penta nucleotide motifs are found lesser compared to -mono and -dinucleotide repeats. Generally, tri nucleotide repeat types are found predominantly occurring in exonic regions whereas mono, di, tetra, penta and hexa are rare in genic regions.

SSRs were used in coconut for analyzing levels of variability among tall populations in India, Sri Lanka, Florida, Philippines, Africa and Mexico. Based on SSR marker analysis, the Red

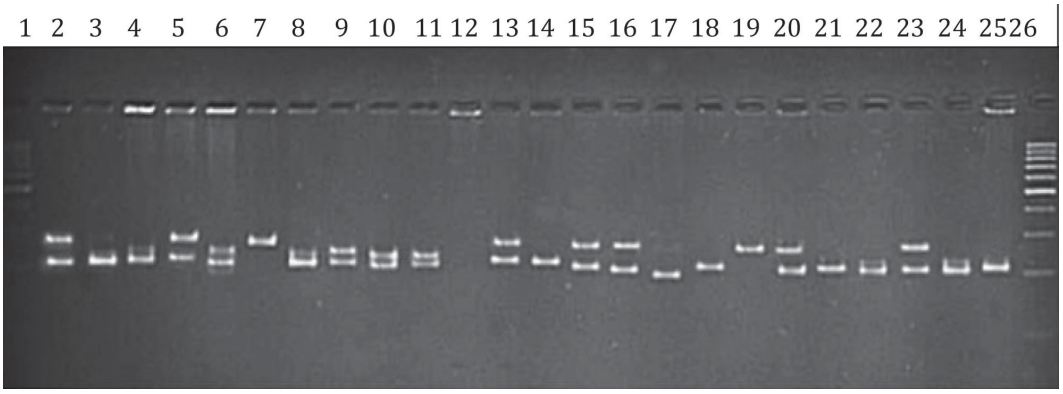

Fig. 11.2: SSR marker profile of coconut samples

Malayan Dwarfs were found to be genetically distinct from Green and Yellow ones. Also, genetic identity of 'Red Spicata' was found to be more to "Fiji Dwarf'. An indigenous dwarf, Kulasekharam Orange Dwarf, showed high genetic diversity higher than many talls. Within population variation (58%) was found to be higher than among population variation (42%). The genetic diversity of Brazilian Tall coconut populations suggested that a spatial structuring of the genetic variability.

5.2.6. Inter-Simple Sequence Repeats (ISSR)

Inter-simple sequence repeats (ISSR), is a fast inexpensive genotyping technique based on the variation in the region between microsatellites. ISSR markers are DNA sequences delimited by two inverted SSR composed of the same units which are amplified by a single PCR primer; composed of few SSR units with or without anchored end. Unlike other traditional markers, neither prior sequence information nor prior genetic studies were required for ISSR analysis, which resulted in this marker type being rapidly used by the research community in various fields of plant improvement. This technique is extremely useful for the characterization of genetic relatedness among populations, genetic fingerprinting, gene tagging, detection of clonal variation, cultivar identification, phylogenetic analysis, detection of genomic instability, and assessment of hybridization. ISSR markers produce multilocus patterns which are known to be abundant, highly reproducible, highly polymorphic, highly informative and quick to use. ISSR make use of the abundant, polymorphic and ubiquitous nature of SSRs present throughout the genome. The technique merges most of the benefits of AFLP and microsatellite analysis with the universality of RAPD. ISSR markers can be used in population genetic studies of plant species as they effectively detect very low levels of genetic variation. Based on SSR markers, it was shown that, coconut accessions from Southeast Asia, South Asia and South Pacific formed separate groups.

5.2.7. Single-Nucleotide Polymorphisms (SNP)

SNPs are the polymorphic events resultant by transitions/transversions in which the sequence variation characterized by a single base substitution at a particular position, at which different

sequence alternatives (alleles) exists in populations. Within a population, SNPs can be assigned a minor allele frequency and the least frequent allele should have an abundance of at least 1%. SNPs are widely popular as a molecular DNA marker system from the late 1990s onwards. The relatively low mutation rates, even distribution across their genome and relative ease of detection impart special attention to SNPs than any other marker systems. SNPs are found to be highly abundant and frequent, but their frequency varies significantly from different portions of the genome, genome to genome in any species and finally from species to species. But development of SNPs requires a great deal of prior sequencing work and validation.

Majority of SNPs are located in non-coding regions of the genome and such are called non-coding SNPs (ncSNPs) of which a subset resides in intronic regions. SNPs that are materialize in exons and the corresponding cDNAs are called coding SNPs, exonic SNPs, or cDNA SNPs, respectively. Within the exons, SNPs either can change the amino acid composition of the encoded protein (nonsynonymous SNP (nsSNP)) or does not cause any major effects (synonymous SNP (synSNP). Promoter SNPs (pSNPs) are the ones that are residing in promoter regions, thereby recommending the high influenze of SNPs in the activity of associated genes.

SNPs are increasingly becoming the marker of choice in genetic analysis and are used routinely as markers in agricultural breeding programmes. They also have many uses in human genetics, such as for the detection of alleles associated with genetic diseases and the identification of individuals. The average SNP density is found to be high in plant genomes. SNP-containing alleles were detected by single-strand conformation polymorphism (SSCP) analysis. SNP markers are being evaluated to serve as anchor markers for the integration of the available individual genotype-specific maps into general reference maps for coconut and oil palm.

6. Genetic diversity in date palm (*Phoenix dactylifera* L.)

There is an apparent diversity of date palms (fruit shape, side leaf structure, and morphological stages) and many studies have been conducted in this palm. RAPD technology appears very effective for identifying accessions of date palm. Amplified fragment length polymorphisms (AFLP) and SSR markers were used to evaluate the genetic diversity between date palm varieties and revealed high genetic diversity in accessions of Sudan date palm germplasm. The genetic diversity of Sudan date palms revealed more variation within groups rather than between groups and similar results have been obtained for Moroccan, Algerian and Tunisian date palm cultivars.

7. Genetic diversity studies in oil palm

Many molecular genomic programmes in oil palm have been conducted towards the development of specific DNA profiles. Initially many genetic variation studies using RAPD and RFLP markers among different accessions of oil palm germplasm were conducted to assess the level of variability in oil palm varieties.

The high genetic diversity observed in oil palm natural populations, specifically in the germplasm collections, indicates these materials are sufficient-to-good sources of new genes for introgression into the current breeding materials for oil palm and palm oil improvement. Similar

to coconut, most of the genetic variation is found within populations, than the between populations as expected for an allogamous and long-lived perennial species. But unlike the coconut, the cluster analysis did not show correlation of geographical distribution and genetic relatedness.

Now a days many research programmes are conducted on ESTs in many crops, which are the short subset of cDNA sequences. They are prominent in identifying gene transcripts, gene discovery and gene sequence determination. Many bioinformatic work flows are developed to analyze and mine sequence information, gene transcripts and polymorphisms in daily accumulated EST sequences.

EST resources have been created for oil palm from male and female inflorescences, shoot apices, zygotic embryos, callus and embryoid cultures. The cluster analysis showed that lipid transfer proteins were highly expressed in embryogenic tissues. Similar reports on six cDNA libraries from oil palm zygotic embryos, suspension cells, shoot apical meristems, young flowers, mature flowers and roots, were constructed (Ho *et al*., 2007).

8. Genetic diversity studies in arecanut (*Areca catechu* L.)

The arecanut palm (*Areca catechu.* L) is one of the most important commercial crops in India. The economic product is the fruit called 'arecanut', which is the most popular chewing substance in South East Asia. Areca palm, a monocot tall, unbranched, slender stem reaches a height of 30-60 meters.

Arecanut cultivation is threatened by a number of diseases during different stages of its growth and development. Among these, the Yellow Leaf Disease (YLD) is the main problem faced by the arecanut cultivators of Kerala and Karnataka where the crop is cultivated extensively. Phytoplasmal etiology of arecanut YLD in India was proved by electron microscopy. Considering the long life cycle of areca palm, selection and characterization of resistant and susceptible varieties through conventional method will be time consuming and laborious. Many individual and combined assays of RAPD and ISSRs were conducted and could prove the genetic discrimination among genotypes.

9. Comparative genomics of palms

Comparative genetic analyses have begun to show that different plant species often use homologous genes for very similar functions. Comparative genomics involves analyzing and comparing genetic material from different species to study evolution, gene function, and inherited disease. Herein, complete genome sequences of different organisms can be compared to check out for regions of similarity and differences and also to understand the uniqueness between different species. Also the gene location, the length and number of coding regions within genes, the amount of non-coding DNA in each genome, conserved linkage groups, conservation in patterns or markers, chromosomal rearrangements, etc. could be considered to impart significant insights to research. Since the comparison of genome sequences which may contain million or billion base pairs via dot plots against one another found a tedious task, lead to the computer based analysis tool for studying evolutionary changes among organisms. By lining

up various genome sequences using computer programs and technologies, apart from identifying conserved genes among species, dissimilar genes that may impart unique characteristics to a particular organism can also be found out. This could help in better understanding of structure, evolutionary process and function of genes which would ultimately pave paths to develop strategies to increase the yield and productivity of crops.

Since two palm genomes only are revealed for survey and most of the palm members are not yet undergone any research progress, the comparative genomic studies in palms have not been taken up. Many marker studies have been carried out for the past several years, mainly in coconut, oil palm and recently in date palm also. In a recent report, the transcriptome and metabolite content of oil palm and date palm mesocarp were compared (Bourgis *et al.*, 2011). The study proposed the higher content of oil in oil palm compared to date palm is due to the much higher transcript levels for all fatty acid synthesis enzymes, specific plastid transporters, and key enzymes of plastidial carbon metabolism. This comparative study illustrates how deep sequencing can provide insights into gene expression patterns of two species that lack genome sequence information.

Date palm whole genome is the first publicly available resource of its type for *Arecaceae*. From the 650Mb whole genome, ~380Mb sequence spanning mainly in the gene rich regions covering ~90% genes in the ~60% of the genome were explored. Also >3.5 million polymorphic sites, including >10,000 genic copy number variations were identified. In the same study the genome assembly was compared with 109,244 contigs of assembled date palm expressed sequence tags (ESTs) thereby 72% of EST contigs were obtained having match of at least 90% of their length, whereas 86% of high-quality EST bases could be aligned to the reference sequence with a minimum of 98% sequence identity. Thus 94% of core eukaryotic genes were found in the assembly, and 71% of these were recovered as full-length gene models. The uncaptured regions of the genome are likely to be highly repetitive and thus intractable to the assembly approach used.

Many marker mining studies have been reported mainly in oil palm and date palm. In an earlier SSR mining study in oil palm, the abundant occurrence of dinucleotide repeats (45.6%) than the other repeat types were reported. A recent study on palm genomics revealed dinucleotides (49%) are found to be predominant in most of the palms followed by mononucleotides (30%) and then tri nucleotides (19%). Also within dinucleotides, AG/GA/ TC/CT motif (55.8%) was observed as dominant repeats (Manju *et al.*, 2011). The predicted SSR frequency was also found in agreement with the earlier oil palm-EST SSR mining. This constant data discloses the conservative and ubiquitous character of SSRs over palms.

Apart from SSR identification, EST-derived SNPs and indels have also found attention in research. There is high frequency of SNP occurrence in oil palm and that will be sufficient to make them appropriate markers for any kind of genetic studies. SNPs occurred at a frequency of 16.8/1 kbp in *E. guineensis* species and 17.5/1 kbp in *E. oleifera* species. In coconut WRKY transcription factor genes, SNPs were detected at a frequency of 2.84/l00 bp (Manju and Arunachalam, 2011).

References

Bourgis, F., Kilaru, A., Cao., Ngando - Ebongue, G., Drira, N., Ohlrogge, J.B., Arondel, V. 2011. Comparative transecriptome and metabolite analysis of oilpalm and date palm mesocarp that differ drarnatically in carbon partitioning. PNAS, **108:** 12527-12532.

Cardena, R., Ashburner, G.R., Oropeza, C. 2003. Identification of RAPDs associated with resistance to lethal yellowing of the coconut (*Cocos nucifera* L.) palm. *Scientia Horticulturae,* **48**:257-263.

Ho, C.L., Kwan, Y.Y., Choi, M.C., Tee, S.S., Ng, W.H., Lim, K.A., Lee, Y.P., Ooi, S.E., Lee, W.W., Tee, J.M., Tan, S.H., Kulaveerasingam, H., Alwee, S.S., Abdullah, M.O. 2007. Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (*Elaeis guineensis* Jacq.). *BMC Genomics*, **8:**381-393.

Lebrun, P., Cho, N., Seguin, Y.P., Grivet, M.L. and Baudouin, L. 1998. Genetic diversity in coconut (*Cocos nucifera* L.) revealed by restriction fragment length polymorphism (RFLP) markers. *Euphytica,* **101**:103-108.

Manju, K. P and Arunachalam, V. 2011. Bioinformatic prediction of SNP markers in WRKY sequences of palms. *Cord,* **27**:17-25.

Manju, K. P., Manimekalai, R and Arunachalam, V. 2011. Microsatellite in palm (*Arecaceae*) sequences. *Bioinformation,* **7**: 347-351.

Rohde, W., Kullaya, A., Rodriguez, J. and Ritter, E. 1995. Genome analysis of *Cocos nucifera* L. by PCR amplification of spacer sequences separating a subset of copia-like *Eco* RI repetitive elements. *J. Genet. Breed.*, **49**:179-186.

Vos P., Hogers, R., Bleeker, M., Reijans, M., Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. and Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research,* **23**:4407-4414.