Mining of repeat elements from *Pongamia* for marker development

Thesis submitted by

RAHUL GUNVANTRAO SHELKE

For the award of the degree

of

Doctor of Philosophy

DEPARTMENT OF BIOSCIENCES AND BIOENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, ASSAM, INDIA

SEPTEMBER, 2018

Dedicated

To

My Beloved Parents

Department of Biosciences and Bioengineering

Indian Institute of Technology Guwahati

STATEMENT

I do hereby declare that the matter embodied in this thesis entitled "**Mining of repeat elements from *Pongamia* for marker development**" is the outcome of research work carried out by me in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, under the supervision of Prof. Latha Rangan.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made whenever work described here has been based on the findings of other investigators.

September, 2018

RAHUL G. SHELKE
(Roll No. 126106022)

Department of Biosciences and Bioengineering

Indian Institute of Technology Guwahati

Date: 10/09/2018

CERTIFICATE

This is to certify that the thesis entitled "**Mining of repeat elements from *Pongamia* for marker development**", being submitted by **Rahul G. Shelke (Roll No. 126106012)** for the award of degree of Doctor of Philosophy, is an authentic record of the results obtained from the research work carried out under my supervision in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree.

September, 2018

Prof. Latha Rangan

(Thesis Supervisor)

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Latha Rangan for the continuous support of my Ph.D study and related research, for her patience and support. Her guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. R. Swaminathan, Dr. Vishal Trivedi, and Dr. B. Anand, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also goes to Prof. Ajay Parida, MSSRF, Chennai, who provided me an opportunity to join their lab and gave access to the laboratory and research facilities. A special note of thanks Dr. G. Ganesan and Dr. Mohan Harikrishnan for helping and guiding me during my stint at MSSRF, without their precious support it would not be possible to complete my research objectives.

I would like to acknowledge present and past HODs of the Department, Prof. K. Pakshirajan, and Prof. Venkata Dasu Veeranki for providing an excellent instrumental facility and research environment. Organisation of regular Bio-Talk helped me to interact with the scientific community.

I express my sincere gratitude towards my lab seniors Dr. Ramesh Aadi Moolam and Dr. Supryo Basak for their suggestions and support. I am also thankful to Dr. Ganesh Thapa for his valuable advice and suggestions. I would also like to thank my fellow labmates Anuma, Reshmi, Ishani, Sanjana, Manish, Gaurav, Ashwitha, Shreekant and Alok for the stimulating discussions, cooperation and for all the fun we have had in the last four years. Also, I thank my friends Mahesh, Gaurav, Akash, Balajee, Karthikeyan and Shrikant for their wise advices, important feedback and friendship.

I extend special gratitude to IIT Guwahati for hostel accommodation and MHRD for institute fellowship and travel funds. I express my gratitude to the departmental and institutional staff for their cooperation and friendly behaviour throughout my Ph.D.

I am highly thankful to my M.Sc supervisor, Prof. Anath Bandhu Das for motivating me to do biological research. I also want to acknowledge all my mentors and teachers who taught me during my entire education journey.

Last but not the least, I would like to thank my family: my parents and to my brothers for supporting me throughout my entire education journey and my life in general.

September, 2018 **Rahul G. Shelke**

Content

Abbreviations

hAT	*hobo, Ac, Tam3*
µl	Microliter
AFLP	Amplified Fragment Length Polymorphism
AT/GC	Adenine-Thymine/ Guanine-Cytosine
BLASTN	Basic Local Alignment Search Tool For Nucleotide
BLASTX	Basic Local Alignment Search Tool For Nucleotide vs Protein
bp	Base Pair
$CaCl_2$	Calcium Chloride
CAP3	Contig Assembly Program
CD-HIT-EST	Cluster Database at High Identity With Tolerance- Express Sequence Tag
cDNA	Complementary Deoxyribonucleic Acid
CP	Chloroplast
CPSSR	Chloroplast Simple Sequence Repeat
CPT	Candidate Plus Tree
DArT	Diversity Arrays Technology
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
En/Spm	*Enhancer/ Suppressor Mutator*
EST	Expressed Sequence Tag
EtBr	Ethidium Bromide
FISH	Fluorescent *In Situ* Hybridization
Gag	Group-Specific Antigen
GBS	Genotyping By Sequencing
GBSS	*Granule-Bound Starch Synthase*
GO	Gene Ontology
GOI	Government of India
HARB	*Harbinger*
IITG	Indian Institute of Technology Guwahati
IN	Integrase
iPBS	inter-Primer Binding Site
IPC	Integral Plate Chamber
IRAP	Inter-Retrotransposons Amplified Polymorphism
ISBP	Insertion Site-Based Polymorphism
ISSR	Inter-Simple Sequence Repeat

JC	*Jatropha curcas*
Kbp	Kilo Base Pair
LB	Luria-Bertani
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MAS	Marker-Assisted Selection
Mbp	Mega Base Pair
MEGA	Molecular Evolutionary Genetics Analysis
MF	*Mesua ferrea*
$MgCl_2$	Magnesium Chloride
MI	Marker Index
MIN	Minute
miRNA	Micro Ribonucleic Acid
MISA	MIcroSAtellite
MITE	Miniature Inverted-Repeat Transposable Element
mM	Millimolar
MpLf	*Millettia pinnata* Leaf Freshwater
MpLs	*Millettia pinnata* Leaf Seawater
MpRf	*Millettia pinnata* Root Freshwater
MpRs	*Millettia pinnata* Root Seawater
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
MT	Mitochondria
MTSSR	Mitochondria Simple Sequence Repeat
NaOH	Sodium Hydroxide
NCBI	National Center for Biotechnology Information
NGPP	North Guwahati *Pongamia pinnata*
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
NJ	Neighbor-Joining
NR	Non-Redundant
NTSYS	Numerical Taxonomy System
NUMT	Nuclear Mitochondrial DNA
NUPT	Nuclear Plastid DNA
ORF	Open Reading Frame
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Primer Binding Site
PCR	Polymerase Chain Reaction
PCS	*Phytochelatin Synthase*

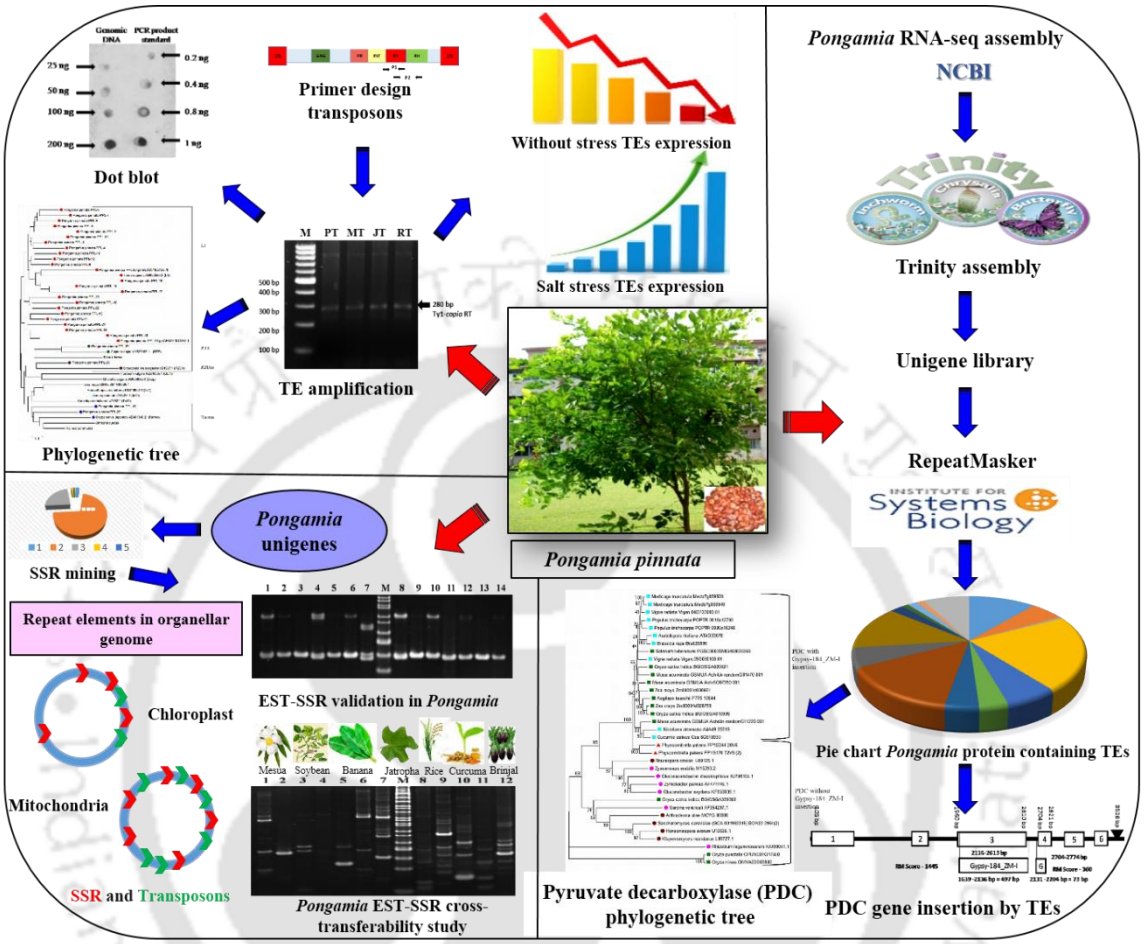
PDC	*Pyruvate Decarboxylase*
PEG	Polyethylene Glycol
PIC	Polymorphic Information Content
POL	Polymorphism
PP	*Pongamia pinnata*
PPGY	*Pongamia pinnata* Ty3-*gypsy*
PPL	*Pongamia pinnata* LINE
PPTY	*Pongamia pinnata* Ty1-*copia*
PVP	Polyvinylpyrrolidone
QA	Quality Analysis
QTL	Quantitative Traits Loci
RAPD	Random Amplification of Polymorphic DNA
RBIP	Retrotransposon-Based Insertion Polymorphism
RC	*Ricinus communis*
REMAP	Retrotransposon-Microsatellite Amplified Polymorphism
RFLP	Restriction Fragment Length Polymorphism
RJJMs	Repeat Junction-Junction Markers
RJMs	Repeat Junction Markers
RM	RepeatMasker
RNA	Ribonucleic Acid
RNase H	Ribonuclease H
RPM	Revolutions Per Minute
rpoC1	*DNA-directed RNA polymerase subunit beta*
RT	Reverse Transcriptase
RT	Room Temperature
SCAR	Sequence Characterized Amplified Region
SDS	Sodium Dodecyl Sulphate
Seq	Sequence
SINE	Short Interspersed Nuclear Element
SNP	Single Nucleotide Polymorphism
SRA	Short Read Archive
SSAP	Sequence-Specific Amplification Polymorphism
SSR	Simple Sequence Repeat
TAE	Tris Base, Acetic Acid and EDTA
TBE	Tris-Borate EDTA
TE	Transposable Element
TE	TRIS and Ethylenediaminetetraacetic Acid
TE-AFLP	Three Endonucleases- Amplified Fragment Length Polymorphisms

TEMED	Tetramethylethylenediamine
TF	Transcription Factor
TIR	Terminal Inverted Repeat
Tm	Temperature
Tris HCl	Tris Hydrochloride
trnK	*Maturase K*
TSD	Tandem Site Duplication
UNFCCC	United Nations Framework Convention on Climate Change
UPGMA	Unweighted Pair Group Method With Arithmetic Mean
UV	Ultraviolet
WGD	Whole-Genome Duplication

List of Tables

List of Figures

Graphical Abstract



Abstract

Pongamia pinnata is a species of tree in Leguminosae family that grows over the different agro-climatic condition of Indian subcontinent, southeast Asia, Australia and Pacific islands. The tree is a source of seeds containing non-edible oil for biodiesel preparation and industrial uses. However, very little is known about the genetic and genomic organisation of *Pongamia*. Repetitive elements in eukaryotes occupy a significant portion of the nuclear genome. They are mainly classified into three categories: interspersed repeats or transposable elements (TEs), tandem repeats and terminal repeats. Retrotransposons are a class of TEs occupying a significant portion of the nuclear genome. Due to the copy and paste mechanism of replication, they create permanent mutation and are responsible for genome size increment. In the present investigation, *reverse transcriptase* (RT) fragment of *copia, gypsy* and *long interspersed nuclear element (LINE)* was isolated and characterised from *Pongamia* genome. Interestingly, TEs were also found in the organellar genome of *Pongamia*. Copy number determination was conducted through dot-blot hybridisation. The copy number of RT gene of *copia, gypsy* and *LINE* in *Pongamia* was estimated to be around 14,653, 11,594 and 18,621 copies per haploid genome respectively. High heterogeneity among *RT* sequences was observed for both Ty1-*copia*, Ty3-*gypsy and LINE*, with *copia* more heterogeneous than *gypsy*. The annotation of *Pongamia* EST libraries yielded more than 400 TEs, confirming that some class of transposons are still transcriptionally active.

With a view to understand the contribution of TEs in distribution and insertional orientation in *Pongamia* transcriptome, ESTs were screened with the RepeatMasker program. Analysis revealed that most of the genes were found with TE insertions, with an average of 1.42 insertions per unigene. The significant population of identified TE insertions were from retrotransposons and less with DNA transposons. We showed with the three examples in which entire or part of genes are apparently derived from TEs. Among them, the insertion of Ty3-*gypsy* into a *Pongamia* unigene similar to the *pyruvate decarboxylase* (PDC) gene is analysed in detail. The high similarity of the *gypsy* sequence to the PDC protein and phylogenetic analysis strongly suggests the presence of Ty3-*gypsy* exaptation in the PDC gene.

Along with TEs, simple sequence repeats (SSR) were successfully mined from organellar genome and transcriptome libraries of *Pongamia*. Among isolated repeats, dinucleotide repeat SSRs were abundant in *Pongamia*. Primers were successfully designed for metabolic important genes. The 16 EST-SSRs amplified in 14 accessions of *Pongamia* produced clear amplicons of the expected size. All the amplified SSRs showed transferability among different families of plants. The newly designed SSR markers were found to be helpful in diversity analysis, as they successfully differentiated among accessions of *Pongamia*.

Keywords: Transposons, Retrotransposons, *Reverse Transcriptase*, *Pyruvate decarboxylase*, Express Sequence Tags, Simple Sequence Repeats.


भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 1

Chapter 1

Introduction

India is a developing country where constantly increasing population and booming industrial growth has a rising demand for energy. It is believed that by the year 2040, India is supposed to share the world's 11% of oil demand (bp.com/energyoutlook). Improper land use and a growing population have resulted in utilisation of large tract of agriculture land for urbanisation that in turn has resulted in a large tract of wasteland in India (over 129 lakh hector) (http://mospi.nic.in/statistical-year-book-india/2017/202). Declining global fossil fuel resources, increasing prices of petroleum products and the risk of global warming is likely to have severe consequences on the automobile industry and growth of the Indian economy. The answer to the above problem is to explore for an alternative substitute for existing fast depleting reserve of fossils and fuel by renewable natural resources like biofuel. Biodiesel could play a vital role in meeting the increased demand of fuel for the industrial and transport sector. India's biofuel share accounted for only 1% of the total global production, which includes 380 million litres of fuel ethanol and 45 million litres of biodiesel (Patni et al., 2011). The Government of India (GOI) is trying to encourage domestic production and greater use of alternative fuels to reduce the dependency on crude oil import by 10 % till the year 2022 (Press Information Bureau, Press Release, https://www.livemint.com/Industry/EsMwKmeQ7x21BWRzvfZG8K/India-aims-to-reduce-crude-import-10-by-2022-Dharmendra-Pr.html). India adopted the biofuel policy in 2009, a significant breakthrough in initiating biofuels strategy, which includes the use of 20% blending of both biodiesel and bioethanol by the year 2017. It is essential to consider the oil yield potential and adaptability of the various edible and non-edible crop, before opting the crop as a source of biodiesel. Non-edible plants are the ideal source for biodiesel production, considering the evident problem of food scarcity in developing countries. Hence, GOI has given top priority for the use and promotion of non-edible oil-yielding plants cultivation for the biodiesel production (Planning commission, GOI, 2003). The biodiesel is regarded as 'carbon neutral', due to no net output of carbon in CO_2 form (http://www.carbonneutralearth.com/biofuels.php). Moreover, the emphasis is on the utilisation of forest, marginal and wasteland for the cultivation of drought-resistant non-

edible oil-yielding plants. *Pongamia pinnata* is one such species that fall in such category of plants which can grow on marginal lands without any further maintenance.

Pongamia tree is a species in the legume family and is an important tree bearing seeds containing non-edible oil utilised for the preparation of biodiesel. The tree is commonly known by name such as Karanj, Pongam and Honge etc. The other botanical names are *Millettia pinnata*, *Derris indica* and *Pongamia glabra.* It is a diploid plant having a chromosome number 2n=22 and relative haploid nuclear genome size of 1198 Mbp (Choudhury et al., 2014). The tree is naturally distributed in tropical and subtropical regions; the centre of origin is regarded as India and China. In India, it grows throughout the region except for the Himalayan ranges. Besides India and China, it is widely distributed in South-east Asian countries, Australian continent, Egypt, some parts of African continent and America (Fig 1.1). It is well adapted in different agro-climatic conditions. Very little reports are published describing the information about the varieties of *Pongamia,* mainly their distribution and genetic relationship.

Figure 1.1. Distribution of *Pongamia* as indicated by orange colour circles.

It is a fast-growing, deciduous or evergreen, small multipurpose tree or large shrub of 15-25 m tall. The plant has raceme-like inflorescence which is axillary and pedant. Flowers are pea-shaped having a fragrance, bisexual, clustered white, purple, and pink coloured (http://www.worldagroforestry.org). The flowers are often pollinated by insects, including bees, flies, ant and thrips. Generally, 1-2 seeds are present inside the smooth and

flattened pods. Seeds are elliptical in shape like a bean, flattened and dark brown in colour. They are rich in oil content, which ranges from 28-40% (Halder et al., 2014).

The yield of *Pongamia* seed ranges from 900 to 9000 Kg/Hectare in different countries and regions, (Bobade and Khyade, 2012) and the oil yield is around 3.49 kilotons of oil from 1061501 plants (Halder et al., 2014). The oil content of seed ranges from 30% to 40%, depending upon the tree (Halder et al., 2014). The seed oil is non-edible due to the presence of flavonoids.

Pongamia has received considerable attention due to its immense role in the production of biodiesel as an eco-friendly fuel and is considered as one of the important sources of biofuel production among the various plant based fuel resources world over (Kesari et al., 2008). It is naturally grown, non-domesticated tree, popularised as one of the vital candidates for renewable energy sources due to features like drought tolerance, adaptability to wide range of environmental condition, easy propagation, rapid growth and higher oil yield than other oil crops. The tree is identified as an important third world feedback resource for biodiesel production and explored in hundreds of projects in India (Karmee and Chadha, 2005).

Alongside biodiesel aspect, *Pongamia* has some applications like potential reduction of environmental pollutants, reforestation, as an ornamental plant, phytoremediation, soil carbon sequestration, preventing soil erosion, biogas production and protein feed for animals. In Industry, the oil is used for soap making, tannery and lubricant industry. Since ancient time, *Pongamia* plant parts have been used for the preparation of extracts against different ailments. In Ayurveda, *Pongamia* is an important part of different medicines like Karanja tail, Maha manjistadi Kashayam and Vilwadi gulika, etc. The plant extracts are also useful against urinary tract disorder, diarrhoea, constipation, diabetes, skin diseases and worm infestation, etc. Several bioactive compounds were isolated from plant parts; viz. flavone and chalcone derivatives such as karanjin, pongone, galbone, pongalabol, pongagallone A and B, etc**.**

Previous scientific studies in *Pongamia* were mainly carried out in agroforestry and medicinal aspects. Molecular investigation in this tree has been significantly expanded for the last ten years. Besides the molecular aspect, investigations were also conducted on biodiesel quality, storage quality, production and stability etc. Selection of economically

important trait such as early maturating cultivars, seed size, high seed yield and oil content, fatty acid composition, abiotic and biotic stress resistance are pivotal prerequisites for making any species economically viable as a biodiesel crop. The understanding of genetic variation exist in seed oil content and morphology could be an important prospect in tree improvement programs. Despite numerous important features, the complete potential of *Pongamia* is yet to be explored.

Diversity study is important to understand the large gene pool present in the plants. The population of an organism which consists of larger gene pool may contain a high number of economically superior traits. These traits can be identified and utilised in plant breeding for the improvement of plants. Previously, the diversity in *Pongamia* population was assessed using morphological traits (Kaushik et al., 2007). The genetic diversity analysis based on phenotypic traits is not much reliable as they depend upon environmental changes. Different molecular markers have been successfully used for diversity studies in *Pongamia*, as molecular markers are independent of environmental condition (Kesari et al., 2010). Alongside molecular markers, organellar and transcriptome sequencing was successfully conducted in *Pongamia* (Kazakoff et al., 2012, Huang et al., 2012). Interestingly, proteomics and genomics studies are very limited. The tree is adaptable to a wide range of agro-climatic conditions, indicating the presence of a considerable amount of diversity. The existence of variation or diversity in genotypes is directly related to changes happened at the genomic level. The majority of adaptable changes are acquired through mutation caused by repetitive elements.

The repetitive elements are the largest fraction of the plant genome. They are mainly classified into three categories: interspersed repeats or transposable elements (TEs), tandem repeats and terminal repeats. TEs occupy a significant portion of the nuclear genome and are mainly divided into two types; DNA-transposons and retrotransposons based upon the mode of propagation (Wicker et al., 2007). DNA transposons propagate by cut and paste mechanism, while retrotransposons propagate by copy and paste mechanism. Further, DNA transposons are classified into two categories based on the presence of terminal inverted repeats (TIR). Retrotransposons are categorised into two types viz., long terminal repeats (LTR) retrotransposons and Non-LTR retrotransposons based upon the presence of LTR on either side of the element (Fig 1.2). Again LTR-retrotransposons are of two types Ty1-*copia* and Ty3-*gypsy* based upon the arrangement of *pol* gene domains.

Due to a copy and paste mechanism of replication, they create permanent mutation and are responsible for genome size increment. In the present research, we tried to isolate and characterise the different repetitive elements present in *Pongamia* organellar and nuclear genome. Sometimes, TEs get activated in response to different biotic and abiotic stimuli. This activation creates the mutations and joins in genome resulting in genetic diversity in plants. Generally, activation of transposons occurs due to stress conditions. The activation of TEs is often observed in cereals as they are less tolerant to different stress, unlike tree species. Hence, very few reports on active TEs on trees species are available.

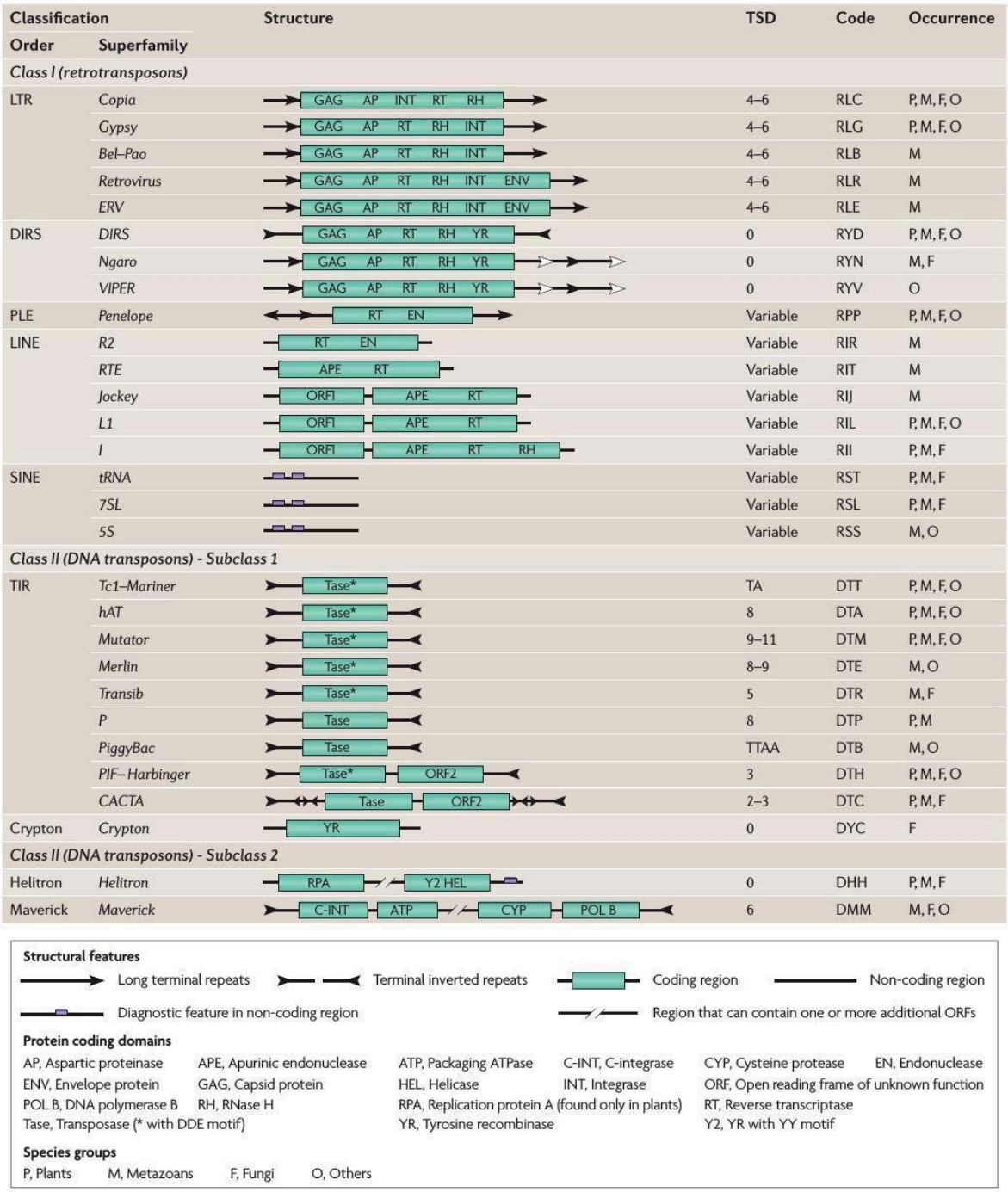


Figure 1.2. TE classification according to Wicker et al. (2007).

On some occasions, TEs are also observed in the genic region of the genome and lead to exon-intron shuffling, creation of new exons, donation of different regulatory sequences to genes, regulation of gene expression and protein diversity. Generally, multigene families are more prone to TEs insertion. These insertions often change the gene function and at times beneficial for organisms. Introns are a reservoir of large numbers of TEs and repetitive sequences, which encourage further mismatching and recombination of

non-homologous genes. It has been reported that introns are an important aspect for genetic recombination and exon arrangement has been a major factor responsible for protein evolution. Diversity created by TEs is the major contributor to adaptation occurring in organisms. In the current study, we investigated the transcriptome sequences of *Pongamia* for TEs insertions.

In the recent past, there is a surge of interest in identifying and developing different molecular markers for rapid assessment of genetic diversity and the selection of desired genotypes. SSRs are also called as microsatellite, categorised under the class of tandem repeat elements. These are the tract of 1-6 bp repeats, generally repeats themselves around 1-10 times. Microsatellites can be amplified by the polymerase chain reaction (PCR), which enables the researcher to detect the nucleotide length variation. This is the novel way of detecting polymorphism present in the variable region of the genome which is highly informative. In this context, in the current study, we have developed the EST-SRRs markers from transcriptome library. The EST-SSR validation was carried out in different *Pongamia* accessions collected from the different part of India. EST-SSRs derived from a transcribed portion of genes are involved in the variety of metabolic functions and unveil the biological significance. Genic microsatellite markers are sometimes less polymorphic than genomic SSRs due to their association with conserved coding regions, unlike non-coding ones. Hence, in most instances, they are transferable across species and serve as a useful tool for gene discovery, population genetics, gene tagging and genetic structure analysis. Molecular markers have great potential in plant breeding for enhancing the selection criteria for desirable traits through marker-assisted breeding, understanding the genetic relationships, evolutionary trends and fingerprinting of varieties.

Repetitive elements are an important part of the organellar genome. But, apart from microsatellites, other repetitive elements like TEs are not reported extensively in the organellar genome. To our knowledge, till date, no fragment of TEs are observed in protein-coding genes of the organellar genome. In plants, TEs were first reported in *Arabidopsis* mitochondrial genome (Knoop et al., 1996). The study of TEs in the organellar genome could be helpful in understanding the transfer of different organellar gene into the nuclear genome. Moreover, organellar tandem repeats like microsatellite are also used for population studies in different plants.

Pongamia remains as an undomesticated plant in most parts of the world. Due to its importance, *Pongamia* has been domesticated for its oil yield. Its seed yield and oil content are found to be variable with different agro-climatic conditions and zones. There are different problems associated with *Pongamia* such as (i) seed toxicity (ii) non-availability of high yielding varieties (iii) long gestation period etc. Different studies have been carried out in *Pongamia* to explore the genetic variation in different sets of genetic material. The majority of investigations were conducted on a morphological aspect, showed a low-level diversity present in a particular set of *Pongamia* population. However, recently many DNA markers were utilised for genetic diversity analysis and identification of polymorphic markers in *Pongamia*. The absence of inadequate genomic resources of co-dominant marker system and sequences related to an economically important trait is still the major constraints. Therefore, enriching the genomic resources like transposable elements and tandem repeats like SSR markers which represent the most abundant and diverse type genomic sequence is the most important need of the hour.

Keeping all these aspects in view, the present investigation was undertaken with the following objectives:

1. Isolation and characterisation of retrotransposons from *Pongamia* genome.
2. Role of transposable elements in *Pongamia* unigene diversity.
3. Development of EST-SSR marker in *Pongamia.*

Chapter 2

Chapter 2

Review of literature

Any nation's energy basket usually depends upon the availability of fossil and fuel reservoir. Enhancing self-reliance on energy is very pivotal for the overall economic development of any developing country. Uncertain price hikes in the international market and depleting fossil fuel reservoirs are causing a serious economic problem for developing countries. During United Nations Framework Convention on Climate Change held at Bonn, Germany (UNFCCC-2017), Government of India (GOI) has expressed a deep understanding of the need to curb down India's reliance on fossil fuels, and the severe effects of carbon emission on the environment and human health. Therefore, there is an urgent need to search for alternative safe, less polluting and renewable source energy. Amongst the different available non-edible oilseed crops, *Pongamia* and *Jatropha* plants are considered as important resources for the biodiesel preparation which can be utilised to mitigate India's over-reliance on fossils and fuels.

2.1. *Pongamia pinnata*: a multipurpose biofuel plant

Biodiesel is a clean-burning renewable source of energy produced from domestically available biomass. Biodiesel can increase energy security and improve air quality thereby reducing to some extent global warming. Moreover, it is a carbon neutral fuel, as it does not release fossil carbon dioxide into the atmosphere (http://www.carbonneutralearth.com/biofuels.php). *Pongamia* is an oleaginous seed-bearing tree. Generally, the seed contains non-edible oil which varies between 30-40 % (Halder et al., 2014). Various studies have been carried out to convert *Pongamia* oil into biodiesel (Karmee and Chadha, 2005, Naik et al., 2008, Bobade and Khyade, 2012, Halder et al., 2014, Mookan et al., 2014, Harreh et al., 2018). However, several laboratory reports suggest that the *Pongamia* biodiesel properties meet the specifications laid by ASTM and German biodiesel standard (Karmee and Chadha, 2005). The extracted non-edible oil is also used in different industries for soap making and tanning (Kesari et al., 2010, Kazakoff et al., 2012, Halder et al., 2014). Interestingly, the oil is suitable for the preparation of aviation biofuel and biodiesel for the cars (Kesari et al., 2008, Kesari et al.,

2010, Ahmad et al., 2009, Kazakoff et al., 2012). Due to its low cost and easy availability, farmers in Karnataka utilised this oil for running generators which irrigate their fields (Karmee and Chadha, 2005). *Pongamia* is being studied in hundreds of projects and regarded as an important feedstock tree. It is described in ancient sacred scriptures like Vedas, Samhita and in many Nighantu as an important medicinal plant (Sharma et al., 2013). Since ancient time, *Pongamia* plant parts have been used for the preparation of extracts and isolation of bioactive compounds like Karanjin, galbone etc. against different ailments like urinary tract disorder, diarrhoea, constipation, diabetes, skin disease etc. (Vismaya et al., 2011, Singh et al., 2011, Satish and Sunita, 2017). However, the plant parts are an important component of different Ayurvedic medicine preparations like Karanja Taila, Maha Manjishtadi Kashayam, Somaraji Tailam etc. Moreover, this tree is an ideal choice for cultivation over different agro-climatic zones and tolerant to stress (Saraswathi and Ezhilarasi, 2012, Kesari et al., 2013). It also fixes N_2 through rhizobium and not grazable by the animal (Kesari et al., 2013). Despite so many important applications, molecular studies on *Pongamia* is still insignificant.

Pongamia is a diploid plant having 22 chromosomes and a haploid genome size of about 1,198 Mb (Choudhury et al., 2014, Ramesh et al., 2014). According to a recently completed organellar genome sequencing reports, mitochondrial genome (425.7 kb) is 2.7 times larger than the chloroplast genome (152.9 kb) (Kazakoff et al., 2012). Moreover, the transcriptome profile of leaf, seed and root has been carried out (Huang et al., 2012, Huang et al., 2016, Sreeharsha et al., 2016, Wegrzyn et al., 2016, Huang et al., 2018). To date, the majority of work has been done on pharmacological aspects, but molecular studies are still lagging behind. The present lack of comprehensive genetic knowledge of *Pongamia* diversity is posing difficulties in producing commercial genotypes. The tree is naturally outcrossing and grows over in different agro-climatic conditions, hence a considerable amount of genetic diversity exists. *Pongamia* has a number of different varieties but very little information is available on them (https://www.daf.qld.gov.au/__data/assets/pdf_file/0003/67575/IPA-Pongamia-Risk-Assessment.pdf). The phenotypic based selection strategy of plants are not always reliable as the phenotypic characters are not stable in different environmental conditions. Hence, the global evaluation of genetic structure in current *Ponagmia* populations is inevitable for the improvement and breeding of new varieties. Earlier investigations revealed the genetic information among wild populations using molecular markers like RAPD, AFLP, SSR,

and ISSR (Kesari et al., 2010, Thudi et al., 2010, Kesari and Rangan, 2011, Pavithra et al., 2014, Sharma et al., 2017). But the vast number of studies were carried out on the limited number of accessions present in a particular area or region.

Understanding of genomics is a central theme to the domestication of any plant, particularly non-model plants. Plant genome is highly occupied with the repetitive DNA sequences than protein-coding genes. The amount of DNA present in the haploid cell of an organism is related to its genome complexity, often called as C value paradox. Generally, protein coding sequences are near about similar in different plants. The variations in genome size mainly are caused due to the variable population of repetitive DNA acquired during the path of evolution. Repetitive DNA are the sequences that are repeated in multiple copies in the genome. These elements occupy up to 90% portion of the plant genome (Foschetto et al., 2002, Yaakov and Kashkush, 2011, Brenchley et al., 2012). Repetitive sequences are mainly categorised into interspersed repeats and tandem repeats. Interspersed repeats are divided into two types; DNA transposons and retrotransposons based upon the strategy employed for the replication. Retrotransposons are the major component of the genome which are further classified into different groups based upon the structure of each element. Tandem repeats are present adjacent to each other in an inverted or directed manner. Depending upon the size of repeats they are grouped into microsatellite and minisatellite DNA. Repetitive elements are responsible for the generation of evolutionary variations and genome organisation in the plants. Since these elements are present as a small fragment in genes, now they no more regarded as junk DNA. Repetitive elements donate the different regulatory sequences to genes, responsible for the exon-intron shuffling, creation of new genes and phenotype. Knowledge of the distribution of different repetitive elements and genomic organisation with respect to evolutionary analysis is necessary in view to understand the behaviour and functional potential of repetitive sequences in the eukaryotic genome (Plohl, 2010). In the last few years, various repetitive elements have been isolated and characterised to ascertain more information about their impact on host genome, copy number, sequence diversity and lineages study. Recently evolution of next-generation sequencing (NGS) technology helped scientist to extract the wealth of genomic information for comparative genomics. In the next sections, we have briefly elaborated on the role of repetitive elements in genome evolution and crop improvement.

2.2. Isolation and characterisation of transposable elements

In recent time, a large amount of accurate genomic information has been available which provided a huge opportunity for understanding the patterns of sequence variation present in the plant retrotransposons. In addition, owing to revolution and decreasing the cost of NGS technologies, several genome and transcriptome sequence of the non-model plants are available. The recent development of computational biology program allowed us to detect the TE sequence patterns and prevalence of particular TE insertions in the host genome. Bioinformatics analysis of available databases has revealed that TEs occupy a major part of all eukaryotic genomes. There are mainly two methodologies often used for the isolation of TEs, first is amplification using PCR and second is a computational method.

Initially, due to the emergence of PCR methods for amplification, the majority of TEs were isolated from the plant genome through degenerate primers designed from conserved regions (Flavell et al., 1992, Ahmed et al., 2011, Barbaglia et al., 2012, Lee et al., 2013, Gao et al., 2016). However, TEs display extreme sequence heterogeneity and there for numerous different families of TEs that exist in plants (Ahmed et al., 2011, Wenke et al., 2011, Barbaglia et al., 2012). In the past number of TEs were isolated and characterised in different organisms (Gao et al., 2014). Among the TEs, retrotransposons are isolated regularly, due to their prominent conserved feature of *Pol* gene (Flavell et al., 1992). Not much data on DNA transposons are reported from the non-model plants. Class I elements were successfully amplified in *Zea mays*, rice and *Arabidopsis* (Yephremov and Saedler, 2001, Huang et al., 2009, Smith et al., 2012). Most of the DNA transposons were isolated from cereal plants in which prior genome information was available. *hAT* DNA transposons were isolated and characterised from *Petunia hybrida, Phaseolus vulgaris, Bambusa vulgaris, Brassica napus* and *Rhododendron simsii* using the degenerate primers (De Keukeleire et al., 2004). Similarly, *hAT* was also isolated through PCR assay from *Zea mays, Antirrhinum majus* and *Beta vulgaris* (Fedoroff et al., 1983, Hehl et al., 1991). However, *En/Spm* and *Helitrons* elements were isolated and characterised from plants (Staginnus et al., 2001, Altinkut et al., 2006). Due to the high level of sequence diversity that exists in DNA transposons families, isolating them though PCR assay is not an easy task. Hence, targeting retrotransposons are far easy than DNA transposons.

Among the transposons, LTR-retrotransposons are widely present in the plants and represent the huge class of TEs. Ty1-*copia* and Ty3-*gypsy* are the major groups of (long terminal repeat) LTR retrotransposons found in a variety of angiosperms. Flavell et al. (1992) isolated the fragments of Ty1-*copia* group of retrotransposons using polymerase chain reaction from higher plants and 56 out of 57 species successfully amplified the fragment of expected size for reverse transcriptase (RT) fragments of Ty1-*copia*. At present, the Ty1-*copia* group is the best-characterised LTR retrotransposon in plants. Investigations have been carried out for studying their sequence characteristic, transpositional activity and system evolution in various plants like tomato, jute, *Epimedium* species, lily, *Camellia sinensis*, *Excoecaria agallocha*, *Erianthus arundinaceus* (Cheng et al., 2009, Ahmed et al., 2011, Chen et al., 2012, Lee et al., 2013, Yao et al., 2017, Huang et al., 2017a, Huang et al., 2017b). Similarly, Ty3-*gypsy* elements were isolated from various plants such as *Prunus mume*, wheat, Poaceae, jute, *Chenopodium quinoa*, *Pyrus* (Wang et al., 2010, Salina et al., 2011, SteinbauerovC et al., 2011, Ahmed et al., 2011, Kolano et al., 2013, Jiang et al., 2013). Most of the isolated LTR retrotransposons are active or nonfunctional due to the presence of frameshift mutations or deletions and stop codons (Cheng et al., 2009, Wang et al., 2010, Muszewska et al., 2011, Huang et al., 2017a). Till date, different portions or domain of retrotransposons were amplified and isolated from plants including integrase (IN), RT, RNase H (RH) and LTR (Kalendar et al., 2004, Xiao et al., 2007, Kalendar et al., 2010, Woodrow et al., 2012, de Souza et al., 2018).

Phylogenetic analysis conducted among these transposons, isolated from dicot and monocot plant species, suggest that TEs existed early in plant during evolution and diverged to heterogeneous subgroups before modern plant orders arose (Flavell et al., 1992, Alipour et al., 2013, Kolano et al., 2013, Lee et al., 2013, Huang et al., 2017b). There are various lineages of Ty1-*copia* reported in plants like *SIRE-1* in *Glycine max*, *Retrofit* in *Oryzae australiensis*, *Tork* in mungbean, *Rider* in tomato, *Oryco* in *Oryza sativa etc.* (Laten et al., 2003, Piegu et al., 2006, Xiao et al., 2007, Cheng et al., 2009, Llorens et al., 2009). Similarly, Ty3-*gypsy* harbours lineages like *CRM*, *Galadriel*, *Del*, *Reina in Medicago truncatula*, *Lotus japonicus*, and *Zea mays* etc. (GorinsL ek et al., 2004). Retrotransposons display a broad pattern of insertion and heterogeneity (Flavell et al., 1992, Pearce et al., 1996, Ahmed et al., 2011, Wenke et al., 2011). Although the cause of presence of sequence heterogeneity is still unclear, there are various possible explanation

for heterogeneity such as (i) Absence of proof-reading activity and a high rate of error during replication of RT, which results in the accumulation of mutations with each replication cycle (Ma et al., 2004, Rajput and Upadhyaya, 2010, Ahmed et al., 2011) (ii) Some retrotransposons are truncated due to the illegitimate and unequal homologous recombination in host genome (Ma et al., 2004, Tian et al., 2009, Rajput and Upadhyaya, 2010) (iii) The heterogeneity may be also influenced by divergence that happened during both horizontal transmission and vertical transmission (Flavell et al., 1992). However, most of the retrotransposons are inactive in nature due to: (i) Defective ORFs (Flavell et al., 1992) (ii) Presence of stop codons and frameshifts (Ahmed et al., 2011) (iii) Presence of retrotransposons in the heterochromatic region (Gao et al., 2008).

In retrotransposons, the various lineages have been reported based upon the structural and length variations. The highest population of lineage are reported in the Ty1-*copia* and Ty3-*gypsy* elements followed by the LINEs and SINEs (Wicker et al., 2007). The availability of a large quantity of accurate genomic and transcriptomic information in various plant species will help in mining the different TEs in the plant. The development and advancement of computational biology have revolutionised our conventional mining strategies of TEs. However, since the last decades, several homologies and structure-based computational programs have been developed for the isolation of TEs from different plant databases. Among this programs, RepeatMasker, Repbase, Gypsy database, LTR finder and MGEScan etc. are the important tools in TEs identification and isolation (Xu and Wang, 2007, Rho and Tang, 2009, Tarailo-Graovac and Chen, 2009, Llorens et al., 2011, Bao et al., 2015).

2.3. Organellar transposable elements

Until the finding of transposons in the mitochondrial genome by Knoop et al. (1996), it was assumed that the TEs were only part of the nuclear genome. Previously on a number of occasions, nuclear-derived sequences were reported in the organellar genome (Knoop et al., 1996). Among the repetitive elements, microsatellite has been reported for organellar genome in numerous reports and utilised for the marker studies. Transposons in mitochondria (mt) commonly reside in intergenic regions (Knoop et al., 1996, Satoh et al., 2004). Not a signal report of transposon invasion of mt genes has been reported to date. Several nuclear derived TE sequences have been demonstrated in mitochondrial (mt) DNA of plant and yeast (Wu and Hao, 2015). Interestingly, all the reported TE-like sequences are fragmented and dispersed throughout the mt genome (Knoop et al., 1996, Alverson et

al., 2011, Islam et al., 2013). The mitochondrial TEs does not show any function as most of the fragments are not full length and contains stop codons (Knoop et al., 1996). The copy number and size of TEs in the mt genome varies in different plants. In *Citrullus* and *Cucurbita* mitochondrial genome, the population of TEs varies from 24 kb (6.4%) to 21 kb (2.1%) respectively, most of them resemble to *copia*- and *gypsy*-like retrotransposons (Alverson et al., 2010). Similarly, nine retrotransposons like fragments were observed, of which 6 covered 111 kb mitochondrial DNA sequences (Knoop et al., 1996). Nine and four fragments of TEs were reported in rice and maize mt genome (Notsu et al., 2002, Clifton et al., 2004). The majority of TEs fragments are found to belong to RT truncated region. Interestingly,Wang et al. (2012) could not observe any transposons in the *Spirodela* mt DNA. To our knowledge, no transposons are detected in the chloroplast genome. According to (Knoop et al., 1996), a massive influx of retrotransposon might have occurred into the organelle in a progenitor of *Arabidopsis* and later most of the nuclear copies were lost under the evolutionary pressure.

2.4. Transposons mediated genome size increment

It is very difficult to fathom and explain how the 1C value of nuclear genome of related species harbours near about the same number of genes but exhibit variation in the total amount of DNA. The study of genome size helps us to understand the presence of variability that occurs between species without any direct link to complexity. The 'C value paradox' exists due to the presence of numerous repeated sequences, among them transposable elements are a major contributor. TEs are an important source for genome size obesity and variation which contributes to genomic plasticity in eukaryotes. Along with TEs, whole-genome duplications (WGDs) i. e polyploidy is also important to the source of genome size increment. Although much of the genome size variation is due to TEs particularly LTR retrotransposons, their mechanism for proliferation is still not clear. Most of our recent knowledge is based on the genome estimation studies conducted by flow cytometry and large genomic sequencing projects.

Plant nuclear genome size occupy different magnitude ranging from 82-megabase genome of the carnivorous bladderwort plant *Utricularia gibba* to Japanese plant *Paris japonica* with a genome size of a 148.8 Gb (Pellicer et al., 2010, Ibarra-Laclette et al., 2013). Some TE families propagate at a different rate, as result some families may get an increase in copy number and diverge quickly in the genome, as it was observed in *Oryza*

and *Gossypium* (Hawkins et al., 2006, Piegu et al., 2006). Staggering genome size variations were observed in *Pisum sativum* L. where up to 1.29-fold genome size variation exists between intraspecific cultivars (Greilhuber and Ebert, 1994). In a previous study, Naito et al. (2009) reported an increase in copy number of DNA transposon *mPing* by ~40 per plant per generation in some rice strains like EG4, A119, and A123. *Eleocharis* genus revealed several interesting facts on genomic contraction and expansion. In younger *Eleocharis* species, an increase in density of Ty1-*copia* elements was responsible for evolution and observed the positive correlation of Ty1-*copia* population with $C_{/n}$-values in the genus (Zedek et al., 2010). The results also suggested the presence of polyploidy and agmatoploidy/symploidy. Similarly, in the *Oryza* genus, the genome size variation occurred due to both polyploidization and LTR-retrotransposon replication. The two families of Ty3-*gypsy* elements namely *RIRE2* and *Atlantis* are propagated to different numbers in various species and attributed for a 3-fold variation genome size range from 357 Mbp in *Oryza glaberrima* to 1283 Mbp in the polyploid *Oryza ridleyi* (Zuccolo et al., 2007). Around 3, 3.6 and 8.1-fold genome size variation was also reported within single genera such as cotton, rice and *Sorghum* respectively (Price et al., 2005, Ammiraju et al., 2006). *Oryza australiensis* is a wild relative of the Asian cultivated rice *O. sativa*, which is subjected to recent expansion through the proliferation of three LTR-retrotransposon families (Piegu et al., 2006). This led to the accumulation of 90,000 retrotransposon copies in the genome since the last three million years, resulting in a rapid 2-fold increase in genome size (Piegu et al., 2006).

LTR retrotransposons are the major fraction of TEs in all plant genomes, the occurrence of a particular family of retrotransposons may be highly variable among species and among varieties of the same species. In some cases, a few TE family may increase their copy number in one lineage (El Baidouri and Panaud, 2013). In case of *Eleocharis* genus, Ty1-*copia*/*Helos1* played a pivotal role in the evolution of both karyotype and genome size (Zedek et al., 2010). In addition, the rapid amplification of *PgDel* LTR retrotransposon has been proposed as an evolutionary driving force in *Panax quinquefolius* genome expansion (Lee et al., 2017).

Retrotransposons recombination are a source of chromosome, duplication and deletion. On some occasion, the genome of host species is actively involved in the removal of LTR retrotransposons from the genome, even though this mechanism is much slower than retrotransposon replication (Baucom et al., 2009). The low activity and purging of

LTR retrotransposons in a small genome can be done using Non-homologous end joining (NHEJ), which led to a massive genomic restructuring in *Oryza brachyanthais* genome (Chen et al., 2013). Hence, *O. brachyanthais* is 60% smaller than its close relative *Oryza sativa* (Chen et al., 2013). In species with a smaller genome, Hawkins et al. (2009) observed a faster rate of *Gorge3* sequence removal relative to the rate of accumulation that resulted in an overall reduction in genome size. Devos et al. (2002) predicted that the illegitimate recombination can act upon at least 5 times more of the genome compared to the process of unequal recombination between LTRs. Hence, illegitimate recombination deletes at least five times more DNA compared to unequal homologous recombination (Devos et al., 2002). All these reports suggested that the ability of TEs to invade genome may depend upon the element and the genome. On some occasions, TEs are able to escape the epigenetic control or genome more tolerant to TE proliferation.

2.5. Transcriptionally active transposable elements

TE is a fragment of genomic DNA that can autonomously jump into new chromosomal locations and in the process often make duplicate copies of themselves (Feschotte et al., 2002, Paterson et al., 2009, Schnable et al., 2009). TEs are the largest component of eukaryotes accounting 50-80% in some grass genome (Meyers et al., 2001). Although they have occupied abundant space, the majority of a TEs are inactive due to their short fragmented nature, the presence of stop codon or they are transcriptionally repressed by silencing (Fultz et al., 2015). For long, they were regarded as junk DNA or genomic parasite due to their deleterious effect on genes. Despite this, TEs have a complex type of interaction with host genes and influence the expression of many genes. However, on most of the occasions, plants are encountered by subtle or more significant environmental changes. To overcome these problems plants require to adapt rapidly to different environmental conditions. Since plants are sessile in nature and have no control over their offspring dispersal, sometimes their germination forces them to grow in habitat which is not conducive for growth and survival. In such conditions, some plants are able to adapt themselves against different abiotic and biotic stresses. Adaptation against these stresses is acquired by genetic changes, which can be achieved by numerous mechanisms including polyploidy, mutation, single nucleotide polymorphism (SNP), epigenetic changes and TEs. Mutation triggered by TEs is frequently associated with adaptation to the environment.

Among the above-listed mechanisms, TEs activity can have a considerable impact on genome structure and function through gene rearrangement and disruption, through up and down-regulation of genes. TEs are often activated due to stress like tissue culture, wounding, microbial elicitors and pathogen attack (Table 2.1) (Okamoto et al. 2000, Melayah et al., 2001). Much of the earlier account on the active TEs came from Barbara McClnitock (1940) study on maize transposons *Ac* and *Ds* element. It's already well-established fact that TEs, like insertional sequences in bacteria, are associated with environmental adaptations. Till now, the existence of a total of 11 transpositionally active LTR-retrotransposons has been demonstrated in rice alone (Finatto et al., 2015). However, in blood orange, significant alteration in the expression of adjacent genes was observed due to the transcriptional activation of LTR-retrotransposons (Butelli et al., 2012). In rice, DNA transposon like *mPing* elements was reported to activate in response to cold and salt stress (Naito et al., 2009, Yasuda et al., 2013). Furthermore, heat-activated *ONSEN* Ty1-*copia*-like retrotransposons are identified in most species of the family Brassicaceae (Pecinka et al., Tittel-Elmer et al., 2010, Ito et al., 2011, Ito et al., 2013, Masuda et al., 2016). Interestingly, Kimura et al. (2001) demonstrated that the jasmonic acid, UV light and salicylic acid treatment were responsible for the expression of *OARE-1* in oat. In tissue culture, callus is reported as the tissue having the highest number of active TEs. Tissue culture is suggested as the complex stress which is responsible for the activation of TEs in maize and rice. The transcriptome library derived from cell culture of maize was enriched with TE ESTs compared to other organ tissue (Smith et al., 2012). In *Arabidopsis*, transcript and extrachromosomal DNA of *ONSEN* was observed in callus of 30 days after heat stress (Matsunaga et al., 2012). Moreover, TEs activation in cell culture had been reported as a mechanism responsible for 'somaclonal variation' in cultured cells of many plant species, resulting in the regeneration of high frequency of mutant plants. The activation of latent transposons beckons the epigenetic changes that occur during the culture process.

Plants are not only challenged to abiotic stress, but also to biotic stress triggered by the interaction with pests, insects and various types of microbes. In protoplast preparation, the plant cell wall is digested by the fungal extract which activates the plant's defense response which results in the activation of the *Tnt1* retrotransposons (Wessler, 1996). Furthermore, crown rust fungus, *Puccina coronate* infection was demonstrated to be responsible for the activation of *OARE-1* elements in oat (Smith et al., 2012). Piya et

al. (2017) observed the increase in differential expression of TEs due to infection of *Heterodera schachtii* in *Arabidopsis* root. Multiple stress like external stresses, factors of microbial origin, bacterial, viral and fungal were reported for *Tnt1* retrotransposon activation (Grandbastien et al., 1997). The unique sensitivity of different TEs to specific stresses underlines the types and frequency of genetic variation induced in specific environments.

Table 2.1. List of some selected plants having transcriptionally active TEs

S.N	Plant	Stress type	Transposons	References
1	Rice	Tissue culture (TC)	*Tos17*, Class-I	(Hirochika et al., 1996)
2	Sugarcane	Stress/Normal	TEs,	(Rossi et al., 2001)
3	Sugarcane	Callus, TC	TEs	(Araujo et al., 2005)
4	Maize	TC	MITE, Class-II	(Barret et al., 2006)
5	Strawberry	Abiotic stresses	Ty1-*copia*, Class-I	(Ma et al., 2008)
6	*Citrus limon*	Salt stress/TC	Ty1-*copia*, Class-I	(De Felice et al., 2009)
7	Rice	TC	*hAT*, Class-II	(Huang et al., 2009)
8	strawberry	Naphthalene acetic acid, abscisic acid	Ty1-*copia*, Class-I	(He et al., 2010)
9	Chickpea	Desiccation	Ty1-*copia*, Class-I	(Rajput and Upadhyaya, 2010)
10	Maize	TC	*h*AT, Class-II	(Vicient, 2010)
11	Coffea	Abiotic and biotic stress	TEs	(Lopes et al., 2013)
12	Gossypium	Heat Stress	Ty1-*copia*, Class-I	(Cao et al., 2015)
13	Pyrus	Normal	LTR-retrotransposons, Class-I	(Jiang et al., 2016)
14	*Knadelia candel*	Nornal	Ty1-*copia*, Class-I	(Liu et al., 2016)

15	Triticeae	Normal	*PIF* and *Pong* like elements, Class-II	(Markova and Mason-Gamer, 2017)
16	*Excoecaria agallocha*	Nornal	Ty1-*copia*, Class-I	(Huang et al., 2017a)

2.6. Domestication of transposable elements in protein-coding genes

The phenomenon in which Junk DNA or TEs acquire new or novel function in genomes, the process is suggested as "exaptation' (Brosius and Gould, 1992). Transposons greatly influence the expression and function of the gene, when they mobilise in close vicinity to a gene or directly into coding and noncoding sequences. Retrotransposons recombination can bring a huge chromosomal rearrangement like translocation, inversion, duplication and deletion. However, TEs controls the gene expression either through the donation of regulatory sequences or inserting themselves near to gene. These regulatory interactions between TEs and genes help host in bringing the adaptive changes against different environmental conditions. Recent studies indicated that the plants use TE cassettes or insertion mechanisms for evolving the existing phenotypic traits (Table 2.2).

The TEs which are inserted inside the coding region either disrupt the gene function or creates a new phenotype. This was first reported by Barbara McClintock in the year 1950 using her experiment conducted on maize in which TE generated a null mutant allele by disrupting the coding region of the gene. In another example, TE was reported in the induction of anthocyanin variation in the blood orange (*Citrus sinensis*), where the expression of *Ruby* was modified due to the expression of a *copia* retrotransposon inserted in its upstream region (Butelli et al., 2012). In many plants, variations in flower and fruit colour pigmentation are caused by transposon activity. A truncated *DcGSTF2* gene was identified in carnations, affected by the insertion of a *CACTA*-type transposable element (Sasaki et al., 2012). The resulted mutant flower line was observed to be bearing deep pink sectors on pale pink petals (Sasaki et al., 2012).

TEs not only affect the colour phenotypes, but they can also alter the other phenotypic traits. TEs contain stress responsive elements (SRE) in their regulatory sequences. The newly propagated TEs having SREs could play the role of new regulatory elements and renders stress-responsiveness to close vicinity protein-coding genes, thereby modifying phenotypic character. A heat activated retrotransposons *ONSEN* was inserted

into genes and resulted in a mutation of abscisic acid (ABA) responsive gene responsible for the ABA-insensitive phenotype in *Arabidopsis* (Ito et al., 2016). Moreover, transposons can control drought tolerance in maize. A miniature inverted-repeat transposable element (MITE) was inserted in the promoter region of a *NAC* gene which is responsible for the natural variation in maize drought tolerance (Mao et al., 2015). Similarly, a MITE insertion ~70-kb upstream to the *ZmRAP*2.7 gene was associated with flowering time in maize (Salvi et al., 2007).

Some TE insertions born mutations are deleterious and are not helpful for plants. For example, in *Zea mays* allelic variation was observed in *knotted1* locus, resulting in mutations affecting leaf development due to the insertion of *Mu1* or *Mu8* TEs within a 310-bp region of the *kn1* third intron (Greene et al., 1994). Domestication of TEs can confer the resistance in crops against diseases. In this context, Hayashi and Yoshida (2009) demonstrated the role of LTR retrotransposon *Renovator*, which acts as a promoter and enhances the expression of *Pit* gene. This gene is responsible for the rice blast resistance, which confers a broad-spectrum fungal resistance (Hayashi and Yoshida, 2009). The TE insertion confers the resistance in rice cultivar K59, unlike susceptible cultivar Nipponbare (Hayashi and Yoshida, 2009). TEs can also regulate the gene expression from long-distance, thereby affecting plant morphology. In maize, the *Hopscotch* LTR retrotransposon inserted 60 kb upstream of '*teosinte branched 1*' (*tb1*) known to trigger overexpression of a gene that leads to apical dominance (Studer et al., 2011). To fulfil the important cellular functions sometimes host domesticates transposases. In *Arabidopsis thaliana*, *DAYSLEEPER*-like genes were identified which was domesticated from the *hAT*-superfamily encoded transposase (Bundock and Hooykaas, 2005, Knip et al., 2013). *DAYSLEEPER*- genes are unique to an angiosperm, essential for normal plant growth and can also regulate global gene expression (Bundock and Hooykaas, 2005, Knip et al., 2012).

Such studies can help to shade a considerable amount of light on the impact of TE domestication in plants and breeding. The recent analysis of many plants indicated that the TEs are responsible for the variability of different genic sequence in plants. The fast-growing sequencing technologies and new genomics tool will allow more global evaluation of TEs involved in crop domestication and evolution of new traits.

Table 2.2. List of some selected TEs responsible for phenotypic changes in plants

S.N	Plant	Gene/Locus	Transposons/ Class	Affected phenotype	Reference
1	Maize	*SBEI* gene	*Ac/Ds*, Class-II	Complex metabolic consequences on starch, lipid, and protein biosynthesis in the seed.	(Bhattacharyya et al., 1990)
2	*Sorghum bicolor*	*Y* gene	*Candystripe1 CACTA* like, Class-II	Variegated pericarp colour	(Chopra et al., 1999)
3	Grape	*Vvmby1A*	*Gypsy*, Gret1, Class-I	Changes in grape skin colour	(Kobayashi et al., 2004)
4	*Glycine max*	Wp locus	*Tgm-Express1* element, Class-II	Pink -flowered phenotype	(Zabala and Vodkin, 2005)
5	Rice	*GBSS*	*Dasheng*, Class-I	Glutinous rice seed in Oragamochi	(Hori et al., 2007)
6	*Ipomoea purpurea*	*bHLH2*	*Mutator*, Class-II	Pale flowers and ivory seeds	(Park et al., 2007)
7	*Arabidopsis*	*FHY3*, *FAR1*	*MULE*, Class II	Response to light signalling	(Lin et al., 2007)
8	Tomato	*SUN*	*Rider*, Class-I	Elongated fruit shape	(Xiao et al., 2008)
9	Rice	*Os02g0582900*	MITE, *mPing*, Class-II	Response to stress	(Naito et al., 2009)
10	*Oryzae sativa*	*Pit* gene	*Renovator*, Class-II	Disease resistance	(Hayashi and Yoshida, 2009)
11	*Brassica oleracea*	*Purple* (*Pr*) gene	*Harbinger*, Class-II	Purple coloration	(Chiu et al., 2010)
12	Maize	*tb1*	*hopscotch*, ,Class-I	Increased maize apical dominance (up insertion)	(Studer et al., 2011)
13	*Glycine soja*	*w1-m locus*	*CACTA*, Class-II	Flower variegation	(Takahashi et al., 2012)

14	*Antirrhinum*	*nivea* lucus	*Tam3*, Class-II	Petal color	(Uchiyama et al., 2012)
15	*Arabidopsis*	*RPP7*	*COPIA-R*, Class-I	Pathogen responses	(Tsuchiya and Eulgem, 2013)
16	Maize	*ZmCCT*	*CACTA*, Class-II	Photoperiod sensitivity	(Yang et al., 2013)

2.7. Microsatellite

Microsatellites are mono- to deca-nucleotide tandem repeats in DNA sequences. They are also referred as "simple sequence repeat" (SSR) or "short tandem repeat" (STR) DNA by Tautz (1989) and Tautz (1993). In a population, the number of repeats can be variable in DNA. On the basis of their location, they are grouped into genomic and genic microsatellites (Varshney et al., 2005). SSR sequences which are present outside the gene are regarded as "genomic SSRs", while "genic SSRs" are those which are present inside the coding region. Microsatellites are the category of tandem repeats which are grouped with the interspersed repeats together they make up genomic repetitive regions. Microsatellites are grouped according to the type of repeat sequence as imperfect, perfect, interrupted or composite (Oliveira et al., 2006). Perfect SSR repeat sequences are not interrupted by any non-repeat bases, unlike imperfect SSRs. Whereas interrupted SSR repeats are interrupted by small non-repeat sequences (Oliveira et al., 2006). Composite repeats are the two different repeats which are present adjacent to each other without any separation. Furthermore, microsatellite is divided into homozygous and heterozygous microsatellite. In homologous microsatellite, the number of repeats on both homologous chromosomes are same, unlike heterozygous microsatellite where the number of repeats is different for each allele *e.g.* one allele with 8 repeats and the other 9 (Li et al., 2017, Thompson and Salipante, 2009).

One of the important features of SSRs is its high mutation rate (Vieira et al., 2016). Different mechanisms have been reported to describe the reason behind the high rate of mutation in microsatellites, like polymerase slippage during DNA replication or repair, errors in recombination and unequal crossing-over (Li et al., 2004, Oliveira et al., 2006). According to Jarne and Lagoda (1996), microsatellites suffer higher rates of mutation than the rest of the genome. Several investigations have been conducted on the mutation rate of

microsatellites in plants, but the rate of mutation varies greatly from species to species, ranging from 0 to 5×10^{-3} per generation (Marriage et al., 2009). The mismatch-repair system corrects the majority of the primary mutations that occurred in microsatellites Ellegren (2004). The cells with deficient mismatch repair system display increase rates of microsatellite mutation Ellegren (2004). Microsatellites are associated with a particular type of mutation process in which the ancestral and the mutant alleles differ by a few repeats (Thuillet et al., 2002). Previously, microsatellites were regarded as evolutionary neutral, but recent studies revealed that they play a pivotal role in evolution and acts as a hotspot of recombination owing to the high rate of mutation (Oliveira et al., 2006, Brandstrom et al., 2008, Vieira et al., 2016).

Microsatellites are not only present in the nuclear genome but also in the organellar genome (Sablok et al., 2015). The existence of higher repeat motifs in the organellar genome may be absent as compared to the nuclear genome. The study of organellar SSRs allows us to comprehend their distribution and organisation in genic and intergenic regions. The important characteristic of the mt genome is that they are inherited maternally and stable even after recombination (Gyawali and Lin, 2013). However, the chloroplast genome shows uniparental inheritance and generally nonrecombining, sometimes display biparental inheritance leading to heteroplasmy (Wolfe and Randle, 2004, Greiner et al., 2015).

The allele length of repeat units may be pivotal in deciding the rate of mutation (Molnar et al., 2012). According to Schug et al. (1998), the mutation rate and mutational properties may also determine the type and nature of the repeat motif. Some reports suggest that the microsatellites mutation rate is greater in dinucleotide than trinucleotide and tetranucleotide microsatellites (Weber and Wong, 1993, Chakraborty et al., 1997). In contrast, Ellegren (2004) proposed that the mutation rate in microsatellite generally increases due to the repeat number. According to Vieira et al. (2016), ~80% of GC-rich trinucleotides is distributed in exons and AT-rich trinucleotides were present evenly throughout the genome. However, some reports mentioned the abundance of mononucleotide repeats in dicots while trinucleotide repeats being high in monocots (Lawson and Zhang, 2006, Kalia et al., 2011). The most abundant SSR motif (AT)*n* observed in plants, in case of the human genome (AC)*n* motif is most abundant (Miah et al., 2013). But the distribution of different SSR repeats may vary between different plant

species. Reports suggested that coding regions are predominant of tri- and hexanucleotide type SSRs (Zhang et al., 2004, Xu et al., 2013).

SSRs are present in transcribed regions or coding region of genomes, (Morgante et al., 2002). In contrast, according to Kantety et al. (2002) and Thiel et al. (2003), the number of repeats and lengths of SSRs may be relatively small in genic regions compared to genomic. The frequency of SSRs in genic regions differ in the different plant genome. Such as, in *Gossypium raimondii* 64.1% and 35.9% of SSRs were observed in the intergenic and genic regions (Zou et al., 2012), in *Phoenix dactylifera* around 75.6% and 24.4% intergenic and genic SSRs were reported (Mokhtar et al., 2016), around 21.46% of rice SSRs were observed in the coding region while fewer SSRs were located in wheat (10.62%) and its two progenitors namely *T. urartu* (5.95%) and *A. tauschii* (6.85%) of their total gene respectively (Deng et al., 2016). EST-SSRs possesses a wide range of functions such as transcription factors, metabolic enzymes, disease signalling, structural and storage proteins (Kalra et al., 2013). Furthermore, a functional analysis may further unearth their role in plant metabolism and gene evolution (Kalra et al., 2013, Joy et al., 2018).

2.8. Repetitive element derived markers

Genetic marker or DNA marker is a gene or fragment of DNA whose location on a chromosome is known. The DNA markers are basically divided into dominant and codominant markers. Codominant marker has the ability to differentiate between the heterozygous and homozygous individuals, unlike the dominant marker. Various parameters are mentioned for categorising marker system. Repetitive elements are the DNA markers derived from various repetitive elements present in the genome. There are two main groups of repeat elements such as interspersed nuclear elements i.e. TEs and microsatellites which have been regularly utilised for the preparation of DNA markers.

Markers development based on repetitive sequence renders an unprecedented opportunity to excess the diversity present in particular species. The repetitive sequences are present throughout the genome including genic and genomic region. Hence, it gives the researcher a choice to utilise each and every corner of the genome for the analysis. Till date, different markers have been developed and utilised for the different research purposes. TEs have been used to design DNA markers like Inter-Retrotransposon Amplified Polymorphism (IRAP), Repeat Junction Markers (RJMs), Repeat Junction-

Junction Markers (RJJMs), Insertion Site-Based Polymorphism (ISBP), Retrotransposon-Based Insertion Polymorphism (RBIP), Retrotransposon-Microsatellite Amplified Polymorphism (REMAP), inter-Primer Binding Site (iPBS) and Sequence-Specific Amplification Polymorphism (SSAP) (Teo et al., 2005, Syed et al., 2005, Chadha and Gopalakrishna, 2005, Paux et al., 2010, Kalendar et al., 2010, Yadav et al., 2015). The TEs based markers utilise different conserve domains or sequences of TEs for marker designing (Teo et al., 2005). For Example, IRAP marker uses RT and LTR sequences and iPBS marker utilises a primer binding site (PBS) of retrotransposons (Teo et al., 2005, Kalendar et al., 2010). But for the development of these markers, a prior idea of the genome sequence is needed. Moreover, for the preparation of these marker TEs library is also required. iPBS marker can be utilised in any species and without requiring prior knowledge of genomic sequences (Kalendar et al., 2010). But iPBS marker is dominant in nature and cannot be utilised in plant breeding for the gene tagging. However, the development of TE markers are quite cumbersome and requires technical skills.

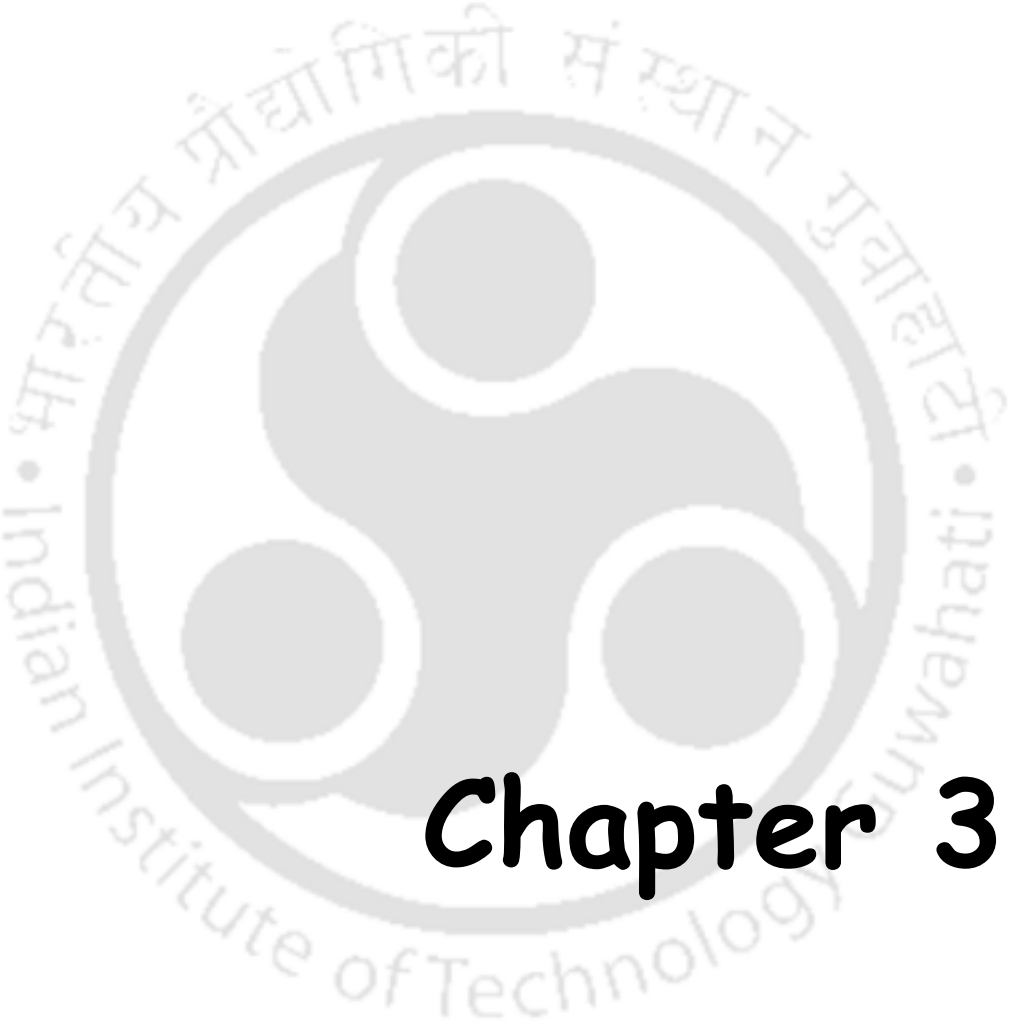
The development SSR markers are easy as compared to TEs DNA marker. The preparation of SSR library involves the processes of genomic and transcriptomic library construction, and sequencing of the clones (Zhang et al., 2017). Generally, the researcher uses the computational approach for the mining and designing of SSR markers from the DNA sequences which is a better choice than the conventional approach (Zhang et al., 2017). There are numerous computational tools available for the identification of microsatellites from DNA sequences. For example, TROLL, SSRIT, Microsatellite Finder, MISA, Tandem Repeat Finder (Benson, 1999, Castelo et al., 2002, da Maia et al., 2008, Metz et al., 2016, Beier et al., 2017). RepeatMasker and BatchPrimer3 are a user-friendly online software tool for the mining of SSR in DNA sequences (You et al., 2008, Tarailo-Graovac and Chen, 2009).

SSRs have been often utilised for the genotyping of plants over the last several years due to their codominant, highly informative and multi-allelic nature (Liu et al., 2013, Pan et al., 2009). EST-SSRs and organellar SSRs has the ability to reproduce and transfer the banding pattern among related species. These markers are particularly useful in wild species (1) for diversity studies to measure the genetic distance (Zong et al., 2015); (2) to determine the crossing over rates and gene flow (Vieira et al., 2016); and (3) in estimating intraspecific genetic relations for evolutionary studies (Jiang et al., 2016a). Furthermore,

a broad distribution, high polymorphism rate and high abundance throughout the genome have made SSRs one of the most widely utilised genetic markers in plant breeding programs for (*i*) quantitative traits loci (QTL) mapping (Kim et al., 2017); (*ii*) population genetics studies (Tsykun et al., 2017); (*iii*) constructing linkage maps (Zhao et al., 2016); (*iv*) marker-assisted selection (Ashkani et al., 2012); and (*v*) gene tagging (Ul Haq et al., 2016). Microsatellites are unstable and inherited in a Mendelian manner and hence they can be used for determining the genetic relationships (Kuntal et al., 2012). However, according to some reports the flanking region of some SSRs loci are conserved among related taxa, hence the designed primers can be utilised to PCR amplify homologous loci in related species (Ul Haq et al., 2016). The cross-species amplification of microsatellites was often reported in the case of organellar and EST-SSR primers (Zhang et al., 2011, Ul Haq et al., 2016). The conservation of microsatellite flanking regions have been reported in closely related species by several groups (Ul Haq et al., 2016, Zhou et al., 2016, Araya et al., 2017). Such an approach reduced down the extensive preliminary work required to design the PCR primers for non-model plants. These primers significantly reduced the cost of developing new primers and promote the quick and widespread application of microsatellite markers (Moges et al., 2016). Apart from the cost, the other disadvantages of SSR marker are the amplification of a shadow or stutter band, homoplasy and existence of null alleles or too many alleles at particular loci (Dettori et al., 2015, Tian et al., 2016, Meyer et al., 2017).

Pongamia is widely known as oil yielding crop, which can be used as feedstock for biodiesel production. Little information is known about the genetic makeup of *Pongamia* genome. Based upon the review of the literature, in the first objective, we are going to isolate the different retrotransposon elements from *Pongamia* genome. The study will help us to know whether the genetic diversity exists in retrotransposons and TEs transcriptional activity in different tissues. Furthermore, in the second objective, we will scout the genes having the TE insertions. TEs are responsible for the diversity of various genic sequences. In this context, we will try to find out the impact of TEs evolution on protein-coding sequences of *Pongamia*. Till date, many investigations have been conducted regarding the application of molecular markers in this plant for genetic diversity analysis using different markers. Microsatellite markers are the most powerful genetic markers for diversity analysis, genetic linkage study and marker-assisted selection. Moreover, mapping studies in *Pongamia* have not been conducted because of the

availability of a limited number of codominant DNA markers like SSRs. To determine the genetic make-up of *Pongamia*, development and application of microsatellite markers are of immense importance. In the third objective, we will design the gene-specific SSR markers for assessing the diversity present in metabolically important traits among the *Pongamia* accessions. The development of EST-SSRs in the present investigation will also shed light on the genotyping of *Pongamia* accessions. This study will lay the groundwork for designing a large number of SSR markers in *Pongamia*.



Chapter 3

Chapter 3

Isolation and characterisation of retrotransposons from *Pongamia* genome

3.1. Introduction

Transposable elements (TEs) occupy a significant portion of the plant nuclear genome. Of the total TEs, retrotransposons are the major component. TEs are classified into two main groups Class I and Class II, based on the structure and type of propagation in the genome. Class I elements are sub-grouped into LTR (long terminal repeat) and non-LTR retrotransposons depending on the presence and absence of LTR at either end. LTR-retrotransposons are sub-grouped into Ty1-*copia* and Ty3-*gypsy* element solely based on the difference in arrangement of *pol* genes. Maize genome is comprised of more than 85% TEs, in which over 75% belong to retrotransposons and can be mostly classified as LTR–retrotransposons (Tenaillon et al., 2010). They replicate with the help of reverse transcription and integrate through the resulting cDNA into another locus of the genome. Their mechanism of replication is almost like retroviruses except for the formation of infectious particle that travels out of the cell to infect other cells.

Non-LTR retrotransposons are mainly categorised into long interspersed elements (LINEs) and short interspersed elements (SINEs). Their copy numbers can also to be seen high in some plant species. These non-LTR retrotransposons are abundantly present in eukaryotic genomes. Of the non-LTR elements, SINEs are most common in primates. In SINEs, *Alu* elements are about 350 bp long, devoid of coding sequences and recognised by the name of restriction enzyme *AluI*. A recent investigation revealed the presence of both LINEs and SINEs in novel genes which are responsible for the functional diversity of gene. Due to difficulty in the mining of SINE elements, only a small number of SINE have been reported in a limited group of plant taxa, such as Poaceae, Solanaceae and Cruciferae.

Class II DNA transposons are of two types, subclass I that contains terminal inverted repeat (TIR) and subclass II as non-TIR transposons. However, subclass I are surrounded by the variable lengths of TIRs which can generate various lengths of tandem

site duplications (TSDs) after insertion. TIR TEs can move from one position to another through a transposase, which has endonuclease and ligase activity. In plants, TIR DNA transposons are mainly grouped into the following types based upon the sequence similarities: *hAT*, *CACTA*, *Mutator*, *P*, *Tc-Mariner* and *Harbinger* (Munoz-Lopez and Garcia-Perez, 2010).

The subclass II type of non-TIR DNA transposons are grouped into two non-TIR DNA transposon, *Maverick* and *Helitrons*. *Mavericks* are also known as *Polintons*. They are the largest and most complex DNA transposons which contain genes with homology to viral proteins. *Helitrons* are abundant and diverse in some plants. Identification of *Helitrons* is difficult due to the absence of a terminal repeat structure. As the TEs move from one point to another point in the genome, they increase the genome size and responsible for the sequence diversity which is an important aspect of the plant genome.

The present study deals with the identification and characterisation of TE elements from *Pongamia* genome. Recently, several studies reported the TE- mediated phenotypic changes in agronomically important plants like purple colouration in *Brassica oleracea*, Blood orange *in Citrus sinensis* and drought tolerance in maize (Chiu et al., 2010, Butelli et al., 2012, Mao et al., 2015). Presence of transcriptome libraries renders a strong basis for the development of *in silico* mining tools. However, not much is reported on the presence of TEs in the *Pongamia* genome. The identification of active retrotransposons and their distribution will be helpful to comprehend about *Pongamia* genome organisation and its evolution. *Pongamia* is also called a semi-mangrove plant due to its ability to sustain against high salts concentration (Huang et al., 2012). Despite its important agronomic features, genomics studies on *Pongamia* are still insignificant. In this context, we have isolated and characterised some TEs in the *Pongamia* genome. Furthermore, the identified TEs in present investigation can be helpful to design new molecular markers for elite genotype selection in future.

3.2. Review of literature

TEs are an important component of eukaryotic genomes. The activity of TEs within the genome responsible for the creation of genomic plasticity by inducing various chromosomal arrangement and functional diversity. Among the retrotransposons, Ty1-*copia* and Ty3-*gypsy* elements are the major components. Their *pol* gene made up of four domains: protease, integrase (IN), reverse transcriptase (RT) and ribonuclease H (RNase H). The identification between Ty1-*copia* and Ty3-*gypsy* elements is mainly based upon the arrangement of protease, IN and RT. LTR retrotransposons are well characterised in plants, unlike non-LTR retrotransposons. LINEs are characterised in insect, protozoa and mammals, but less investigated in plants. Like LTR-retrotransposons, RT is also found in LINEs. Eukaryotic RT shares some conserved amino acid domains with retrotransposons and in retroviruses. This conserved domain of *pol* gene was consistently used for primer designing. Flavell et al. (1992) isolated the *copia* elements from higher plants through degenerate primer designed from RT. Similarly, Ty3-*gypsy* and LINEs were isolated from different plants using RT degenerate primers (Friesen et al., 2001, Kubis et al., 1998). However, the identification of SINEs is still tricky due to their extreme heterogeneous nature. Moreover, their isolation by targeting *RNA polymerase III* promotors through PCR is not easy (Borodulina and Kramerov, 1999). Despite all these, a small population of SINEs are reported in some plants like Poaceae, Cruciferae, and Solanaceae (Wenke et al., 2011).

Among retrotransposons, *copia* element plays an integral part in genome diversification and evolution, but most of them are found to be inactive due to the presence of stop codon during the plant growth development (Flavell et al., 1992, Hirochika and Hirochika, 1993). Still, some active retrotransposon has been investigated in plants at different stages of plant growth and development. Their activity often exists in response to different biotic and abiotic stress due to the presence of stress-responsive regulatory elements in the LTR region (Finatto et al., 2015). However, the characterisation and understanding of TEs distribution with regards to copy number are helpful to comprehend the genome organisation and evolution. In *Oryza australiensis*, the activity of retrotransposons resulted in a rapid two-fold increment in genome size during the last 3 million years, suggesting that rapid amplification of LTR-retrotransposons has played a significant evolutionary role in genome expansion (Piegu et al., 2006). Moreover, their

activity is reported in cell and tissue culture, wounding and pathogen infection (Hirochika, 1997). Use of retrotransposons is more than a potentially important tool for studying genetic diversity, genome evolution and expression. Transcriptionally active retroelement produces new permanent insertion in the genome that leads to variation. High-level variation exists in retrotransposons lineages due to their RT, which does not have proofreading activity. This activity results in an extreme error-prone mode of replication leading to the introduction of point mutations. It is essential to notify that, activity and proliferation of retrotransposons through error-prone transcription is one of the critical factors in plant genome expansion and evolution.

Among the DNA transposons, *hAT* and *En/Spm* elements have often been isolated through the use of PCR assays. The transposase in *En/Spm* is highly conserved among plants. Hence, they have been isolated and characterised in Gramineae, Solanaceae, Leguminosae, Chenopodiaceae and Alliaceae species (Staginnus et al., 2001, Altinkut et al., 2006). Similarly, *hAT* transposons were predominately isolated from *Zea mays*, *Antirrhinum majus* and *Beta vulgaris* (Fedoroff et al., 1983, Hehl et al., 1991). Some of the DNA transposons like *Helitrons* were initially difficult to identify due to the absence of typical structural features (Du et al., 2008, Yang and Bennetzen, 2009). Given the level of sequence diversity present between DNA retrotransposons are far easy than DNA transposons. Due to the evolution of next- transposons, targeting them with PCR assay is not straightforward. Hence, the targeting generation sequencing (NGS) technologies, the different genome and transcriptome databases are available which could act as an important source for TEs mining.

Assessment of genetic diversity has been an essential component of plant breeding for crop improvement. High heterogeneity and dispersal of retrotransposon throughout plant genome have provided an excellent opportunity for designing molecular marker system to study the DNA fingerprinting and genetic linkage mapping. The present study deals with the identification and characterisation of retrotransposons elements from *Pongamia* genome. Furthermore, this investigation also attempts to examine their heterogeneity, abundance in the genome, phylogenetic relationships and transcriptional activity. The diversity observed in retrotransposons population can further be utilised for the study of molecular genetics in future as well.

3.3. Material and methods

3.3.1. Plant material

Pongamia accession NGPP-46 (North Guwahati *Pongamia pinnata), Mesua ferrea, Jatropha curcas and Ricinus communis* were used in the present study for DNA isolation (Kesari et al., 2008). Plants were raised using seeds in poly bags in Greenhouse at Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati (IITG), Assam, India.

3.3.2. Genomic DNA extraction

The total genomic DNA was isolated from young fresh leaves of *Pongamia, Mesua, Jatropha and Ricinus* using modified sodium dodecyl sulphate (SDS) method (Kesari et al., 2009). About 5 g of fresh and young leaves were collected for grinding using liquid nitrogen along with 2% PVP (Polyvinylpyrrolidone) in mortar and pestle to obtain a fine powder. The fine powder was immediately transferred to 50 ml polypropylene centrifuge tube and gently suspended in two volumes of preheated extraction buffer at 65°C, incubated for 30 min at 65°C in water-bath and mixed by gentle shaking after every 10 min interval. A double volume of chloroform: isoamyl alcohol (24:1) was added and the tubes were inverted gently shaken for 15 to 20 times and centrifuged for 20 min at 10,000 rpm at room temperature (RT). The upper aqueous phase was carefully transferred by wide-bore of tips to a fresh sterile 50 ml centrifuge tubes to avoid mechanical damage to DNA. Two volumes of ice-cold isopropanol were added to collect the upper aqueous phase, and the tube was gently shaken and kept at – 20°C for 1 hr to precipitate the DNA. The precipitate was centrifuged at 12,000 rpm for 15 min, and the supernatant was discarded. The pellet was washed with 70% chilled ethanol by centrifuging at 12,000 rpm for 15 min. The pellet was air-dried and suspended in 500 µl of TE buffer (pH- 8.0).

For purification of extracted genomic DNA, 3 µl RNase A (10 mg/ml) was added to the sample and the mixture was kept at 37°C for 30 min. An equal volume of chloroform: isoamyl alcohol was added in the sample followed by centrifugation at 10,000 rpm for 5 min. The aqueous phase was collected in a fresh vial; ethanol precipitation was carried out in the presence of 3 M sodium acetate (pH 5.2). The precipitated DNA was centrifuged to

a pellet and washed in 70% ethanol, air or vacuum dried. The final DNA pellet was dissolved in 30 to 50 µl (depending upon the pellet) of TE buffer.

3.3.3. Quantification and quality check of genomic DNA

The genomic DNA yield was determined using a Nanodrop spectrophotometer Tecan Infinite 200 PRO (Nanodrop Technologies, DE, USA) as per standard manufacturer's instructions. The ratio of absorbance at 260 nm and 280 nm was used to assess the purity of DNA and RNA. The purity of DNA was determined by calculating the ratio of absorbance at 260 nm and 280 nm. The concentration was recorded in ug/µl. In addition, the quality and concentration of genomic DNA was also determined by running 3 µl of DNA from each sample on a 0.8 % agarose gel containing 0.5 µg/ml of ethidium bromide (EtBr).

3.3.4. PCR validation

PCR amplification was carried out using isolated DNA from *Pongamia* accessions NGPP-46 in Mini Thermal Cycler (Applied Biosystems 9700, USA). Different retrotransposons primers were synthesised based on the earlier published report for PCR amplification. The list of each retrotransposons type and primer length sequence and annealing temperature are mentioned in Table 3.1. PCR amplification was conducted in 25 µl reaction volume containing 50 ng of DNA, 2X PCR master mix pH 8.5 (Promega, USA), 400 µM dNTP, 3 mM $MgCl_2$, nuclease-free water (Promega, USA) and 0.6 µl of 0.1–1.0µM of each retrotransposon forward and reverse primer. The reaction was performed in 0.2 ml microfuge tube (Dialabs, USA). The PCR cycling was as follows: 5 min at 95ºC initial denaturation followed by 35 cycles of 1 min at 94ºC, 1 min at annealing temperature (Tm), 72ºC for 40 s and the final extension of 5 min at 72°C. The genomic DNA amplified by retrotransposons primers were checked for amplification on 1.5% agarose gel. For conducting electrophoresis, agarose gels were prepared using agarose (Sigma, USA) in 1X Tris-Borate EDTA (TBE) buffer using horizontal agarose gel slab apparatus (Bio-Rad, USA). The PCR amplified samples and 100 bp size ladder (Himedia, India) as a reference marker with loading dye were loaded into the wells. Electrophoresis was carried out at 5V/cm for 1 hr. PCR amplified bands in the gel were observed under UV-transilluminator followed by gel documentation (Bio-Rad, USA).

Table 3.1. List of primers used for amplification of transposable elements in *Pongamia*.

S.N	Primer/(TE name)	Sequence (5′-3′)	Tm	References
1	Ty1-*copia* RT (Ty1-*copia*)	F: ACNGCNTTYYTNCAYGG R: ARCATRTCRTCNACRTA	46°C	(Flavell et al., 1992)
2	RT3/RNase H1 (Ty1-*copia*)	F: TATGTDGATGAYATGYTDATT R: CCTCACATCWATRTGYTTBGW	44°C	(Woodrow et al., 2012)
3	GyRT1/3 (Ty3-*gypsy*)	F: MRNATGTGYGTNGAYTAYMG R: YKNWSNGGNTAYCAYCCARAT	46°C	(Friesen et al., 2001)
4	DVO144/10712 (LINE)	F: GGGATCCNGGNCCNGAYGGNWT R: SWNARNGGRTCNCCYTG	47°C	(Wright et al., 1996)
5	BEL1MF/BEL2 (LINE)	F: RVNRANTTYCGNCCNATHAG R: TCYGTCCCCCTRGGRRACAG	44°C	(Alix and Heslop-harrison, 2004)
6	*Au* SINE	F: AGCTGCTGCCTTGTGACCAT R: GGGAAGGGTCCGACCACTT	60°C	(Ben-David et al., 2013)
7	AUFW2	F: TGGTAAAGYTGITGYCWTGTGA	52°C	(Fawcett et al., 2006)

	AURV2 (SINE)	R:STATWGTACGCAGCCTTWCCCT		
8	*hAT*	F: CA(C/T)GTI(A/C)GITG(C/T)IIITG (C/T)CA(C/T)AT(A/C/T)(C/T)T R: AAIGCI(C/G)I(C/T)TCI(C/G)(A/T)IGC (A/C/G/T)AC(A/C/G/T)GT	46°C	(De Keukeleire et al., 2004)
9	*En/Spm*	F: GGAAAACTAATATGATTCGACATAA TTTGAYITIATGCA R: ACCTACATRDASAAACTTTCTATACC TGTAGACAGATAC	58°C	(Staginnus et al., 2001)

3.3.5. Cloning

The whole PCR products were directly ligated into the high-quality ready-to-use TA cloning vector pTZ57R/T (Thermo Fisher Scientific, USA). The cloned product was sequenced by Macrogen sequencing service (South Korea). Before cloning competent cell were prepared for transformation. The details of the procedure are given as follows.

3.3.5.1. Competent cell preparation

1. On Day 1, a single colony of DH5-α strain inoculated in 25 ml of Luria-Bertani (LB) in 250 ml bottle and incubated at 37°C for 4-6 hrs at 250 rpm. 2. On Day 2, 100 ml LB medium was inoculated with 1ml of saturated overnight culture. 3. The flask kept for shaking at 37°C until OD at 600 increased to 0.4 (2-3 hrs). 4. Then the culture was transferred to two pre-chilled 50ml falcon tubes. 5. The tubes were centrifuged at 2700x g for 10 mins at 4°C. 6. The medium was removed after completion of centrifugation; the cell pellet was suspended in 1.6 ml ice-cold 100 mM $CaCl_2$ by swirling on ice gently. 7. The tubes were kept on ice for 30 mins followed by centrifugation at 2700x g for 10 minutes at 4°C. 8. Again the medium was removed and the cell pellet was suspended in 1.6 ml ice-cold 100 mM $CaCl_2$ by swirling on ice gently. 9. Tubes were kept on ice for 20 mins. 10. Then the cell pellet was distributed in Eppendorf tubes with an addition 0.5 ml of ice-cold 80% glycerol. 11. Eppendorf tubes were frozen in liquid nitrogen and store at -80°C for further use.

3.3.5.2. Transformation protocol

1. DH5- α competent cells were thawed on ice for 15-30 mins, same time PCR product was added in ligation mixture (10X ligation buffer, PCR product, TA vector 50ng/µl, T_4 ligase 1U) and kept for 30 mins at 22°C. 2. Around 4 µl of ligation mixture was added to the competent cells. 3. The ligation mixture and competent cells were kept on ice for 20-30 mins, mixing at every 5 min interval. 4. Heat shock treatment was given to competent cells containing ligation mixture by keeping it in a water bath at 42°C for 90 sec. 5. The competent cells were immediately kept on ice for 2 mins. 6. Then, 200 ul LB broth was added to heat-shocked cells. 7. Cells were incubated at 37°C on a shaker (180 rpm) for 1 hr. 8. After 1 hr, the cells were centrifuged at 5000 rpm for 1 min. 9. Around 600 µl of LB broth was removed and remaining broth mixed well with the pallet. 10. Around 100-150

µl broth containing transformed cells were spread on LB agar (Ampicillin 100 µg/mL). 11. The LB agar plates were kept for growth overnight at 37°C.

3.3.6. Colony PCR

Colony PCR was carried out in 6 µL of reaction mixture containing each transformed colony in separate PCR tube with M13/pUC primer: forward:- 5'- GTAAAACGACGGCCAGT-3' and reverse:- 3'-CA GTA TCG ACA AAG GAC-5'. PCR amplification was conducted with 2X Green GoTaq PCR master mix pH 8.5 (Promega, USA) containing 400 µM dNTP, 3 mM $MgCl_2$, 1U of Taq DNA and nuclease-free water (Promega, USA). Thermal cycling conditions were: 95 °C for 5 min; 35 cycles of 94°C for 30 sec; 60°C for 30 sec; 72°C for 1 min and the final extension at 72°C for 5 min. For colony PCR, multiple transformed colonies were selected from the LB agar plate. The amplification products were visualised on 1 % agarose gel containing 0.5 µg/mL of EtBr in 1xTAE buffer and documented under BIO-RAD UV transilluminator.

3.3.7. Plasmid isolation and sequencing

1. The positive colonies (transformed) were inoculated in 4 ul of LB broth and kept on shaker incubator for overnight at 37°C. 2. Cells were harvested from the overnight grown culture by centrifugation at 12,000 rpm at RT 3. The cell pellet was dissolved in 300 ul of P1 buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 100 µg/ml RNase A) thoroughly by vortexing. 4. In the same Eppendorf, 300 ul of P2 buffer (200 mM NaOH, 1% SDS) was added and incubated at RT for 5 min. 5. Next, 300 ul cold P3 buffer (3.0 M potassium acetate pH 5.5) was added and incubated on ice for 5 min. 6. After ice incubation, centrifugation was carried out at 14,000 rpm for 10 min at RT. 7. The supernatant was carefully taken out and mixed with 166 ul of 50% PEG 6000 and 118 ul of 5 M NaCl followed by centrifugation at 14,000 rpm for 10 min at RT. 8. Plasmid DNA was observed as a pellet in the tube. 9. The DNA pellet was washed with 500 ul of 70% ethanol at 10,000 rpm for 10 min. 10. The pellet was air dried and dissolved in 30 ul of TE buffer. 11. The dissolved plasmid was visualised on 1 % agarose gel containing 0.5 µg/mL of EtBr in 1xTAE buffer and documented under BIO-RAD UV transilluminator. 12. The cloned plasmid DNA was sequenced by Macrogen sequencing service (South Korea).

3.3.8. Mining of transposable elements from *Pongamia* organellar genome

Pongamia mitochondria and chloroplast complete genome sequences were retrieved from the National Center for Biotechnology Information (NCBI) GenBank database (GenBank accession no. JN673818.2 and JN872550.1). Repbase tool (http://www.girinst.org/censor/index.php) was employed for the screening of TEs present in *Pongamia* chloroplast and the mitochondrial genome.

3.3.9. Mining of transcriptionally active transposable elements from *Pongamia* unigene libraries

Transcriptome database was required to isolate the transcriptionally active transposable element. Before mining of active TEs, we assembled the transcriptome library from available RNA-seq reads at NCBI. The details of the procedure are given as follows.

3.3.9.1. Transcriptome cleaning and assembling

The high throughput Illumina 2x75 bp paired-end reads were downloaded from publically available *Pongamia* libraries SRR349650 (MpRs), SRR349651 (MpRf), SRR349652 (MpLs), SRR349653 (Mplf) using short read archive (SRA) toolkit (Huang et al., 2012) (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsof). Mp means *Melliatia pinnata,* R and L represent root and leaf tissue, s and f represent two types of treatments seawater or freshwater (Huang et al. 2012). Before transcriptome assembly, quality check of downloaded libraries was performed with a FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The raw data was cleaned with the Trimmomatic program (http://www.usadellab.org/cms/?page=trimmomatic) using following parameters: (i) read pairs were removed if the average quality scores equal or less than 30, (ii) adaptors contamination were eliminated from reads, (iii) reads length less than 20 bp were discarded. The cleaned paired-end reads were assembled separately with Trinity assembler (https://github.com/trinityrnaseq/trinityrnaseq/wiki) using default settings. To obtain longer and complete sequences, Trinity derived contigs were reassembled using a Cap3 tool (http://seq.cs.iastate.edu/cap3.html). CD-HIT-EST tool was employed to remove the redundant assemblies **(**http://www.bioinformatics.org/cd-hit/**)**.

3.3.9.2. Mining of active transposable elements

Four assembled unigene libraries and seed unigene library were retrieved from NCBI (GEOZ00000000.1). All libraries were screened to find the transcriptionally active TE fragments. The unigenes were scanned for TEs using the RepeatMasker (RM) version 3.1.5 against a database of 2,064 reference TE sequences from maize, sugarcane, sorghum and millet (*Panicoid*). To avoid the false positive results, RM cut off scores equal or higher than 250 were opted, without imposing additional length thresholds. The TE sequences orientations in unigenes were also determined by RepeatMasker. Furthermore, following parameters were used for analysis: (i) simple repeats and low complexity regions were removed, (ii) low complexity DNA sequences were not masked, (iii) high sensitivity/ low-speed search conditions.

The TEs annotated using Repbase were again employed for annotation using BLASTX search with a cut-off e-value = 1.0e-5 against protein databases such PLAZA 2.0 and 3.0, Swiss-Prot, NR (non-redundant) database at NCBI and *G. max* database at Ensemble plants. Unigene sequences annotated as TEs by BLAST and RepeatMasker tools; only those sequences were further selected for detail analysis. Thus, false positive identification of TE was avoided.

3.3.10. Sequence analysis

The sequences obtained from cloning were subjected to either BLASTN or BLASTX analysis **(**http://www.ncbi.nlm.nih.gov**)**. Edited sequences were further BLASTN against Repbase **(**http://www.girinst.org/censor/**)** and RepeatMasker program **(**http://www.repeatmasker.org/**)**. The nucleotide sequence alignment was conducted using MUSCLE program **(**http://www.ebi.ac.uk/Tools/msa/muscle/**)** for the generation of multiple sequence alignment followed by sequence annotation in Gene Doc V2.7 **(**http://genedoc.software.informer.com/2.7/**)**. The phylogenetic tree was constructed in MEGA 6 using Neighbor-Joining (NJ) method through 1000 bootstrap replicates **(**http://www.megasoftware.net/**)**.

3.3.11. Dot blotting

Genomic DNA and heterogeneous 0.9 kb PCR product were denatured in 0.4 M NaOH for 30 min followed by heating at 100ºC for 5 min and then quickly chilled. Denatured genomic DNA and PCR product were spotted on a positively charged nylon membrane (Hybond-N+ Amersham-Biosciences, England) in various concentrations. For the probe preparation, PCR product of retrotransposon was labelled by Biotin DecaLabel DNA Labeling Kit (Fermentas, Germany). Hybridisation was performed at 65ºC for 18-20 hrs (6x SSC, 5x Denhardt's solution, 0.5 % SDS and 100 µg/ml salmon sperm DNA). After washing, the signal was visualised immunologically using Biotin Chromogenic Detection Kit (Fermentas, Germany) according to the manufacturer's instruction. Analysis of dot blot was performed using ImageJ 1.48v software (http://rsbweb.nih.gov/ij/). The inverted image of dot blot was considered for measuring the hybridisation signals. Copy number was calculated using the equation given by Ma et al. (2008): Copy number = (the size of the haploid genome x average proportion of nuclear genomic DNA hybridising to the probe) / size of the probe element.

3.3.12. Synonymous and nonsynonymous substitution analysis

The synonymous and nonsynonymous substitutions pattern per site was determined for the RT sequences of LTR transposons. The transcribed sequences were selected from *Pongamia* transcriptome libraries. All the nucleotide sequences were submitted to Synonymous Non-synonymous Analysis Program (SNAP v2.1.1) to calculate synonymous and non-synonymous substitution rates based on a set of codon-aligned nucleotide sequences (https://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html).

3.4. Results and discussion

3.4.1. Isolation and confirmation of retrotransposons in *Pongamia* genome

Pongamia (Fabaceae) is a non-edible oil yielding tree grown in India. Despite its increasing popularity as a medicinal and oil-yielding tree in the Asian subcontinent, studies on molecular aspects are still lagging. Transposable elements are an important source of genetic variations which potentially causing variation in genome structure and gene expression responsible for evolution (Kidwell and Lisch, 1997, Zedek et al., 2010, Lisch, 2013). Considering that no TEs are reported in *Pongamia*, the present investigation was carried out to isolate different TEs.

In the present study, we amplified the partial *pol* gene domains of the retrotransposons in *Pongamia* genome using degenerate primers (Flavell et al., 1992). Among the retrotransposons, initially, we amplified the RT and RT-RH (RNase H) domains of the Ty1-*copia* like retrotransposons using degenerate primers. An expected amplicon size of 280 bp for RT and 800-850 bp size for RT-RH gene was amplified (Fig 3.1) and (Table 3.1). Along with *Pongamia*, we also tried to test these primers in *Mesua, Jatropha and Ricinus.* Similar size of PCR product was amplified in all four plants (Fig 3.1). This depicts that the conserved motifs of Ty1-*copia* are present in all plants. Isolated RT and RT-RH sequences were used for the cloning purpose. RT-RH fragment of *Pongamia* was further employed for the dot blot hybridisation.

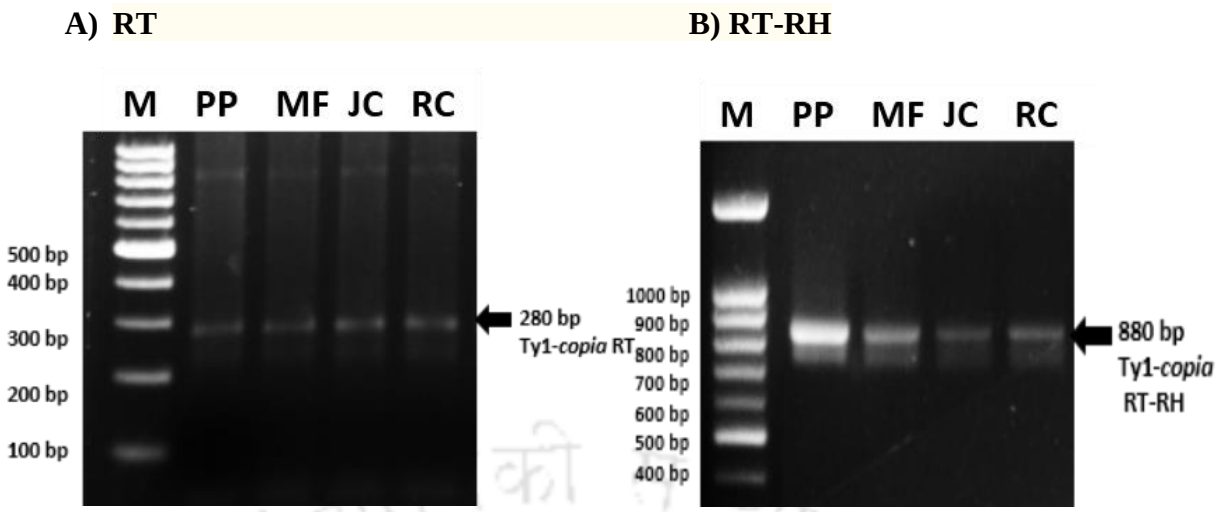


Figure 3.1. PCR amplified product: A) RT; B) RT-RH of Ty1-*copia* retrotransposons. M: 100 bp DNA ladder, PP: *P. pinnata*, MF: *M. ferrea*, JC: *J. curcas* and RC: *R. communis* Ty1-*copia* PCR product.

Similarly, RT domain of Ty3-*gypsy* of length 420 bp was successfully amplified using a degenerate primer in *Pongamia*, *Mesua*, *Jatropha and Ricinus* (Friesen et al., 2001) (Fig 3.2) and (Table 3.1). In previous studies, Ty3-*gypsy* isolation was carried out using degenerate primers in various plants like Japanese apricot, jute and *Chenopodium quinoa* (Wang et al., 2010, Ahmed et al., 2011, Kolano et al., 2013).

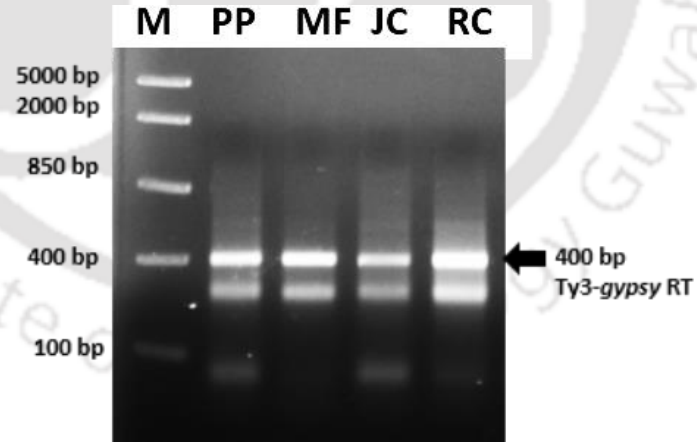


Figure 3.2. PCR amplified product: A) RT; B) RT-RH of Ty3-*gypsy* retrotransposons. M: 100 bp DNA ladder, PP: *P. pinnata*, MF: *M. ferrea*, JC: *J. curcas* and RC: *R. communis* Ty3-*gypsy* PCR product.

Non-LTR retrotransposons were successfully amplified in *Pongamia* genome using degenerate primers. To isolate LINEs, first we used BEL1MF/BEL2 degenerate primers, but we did not get any PCR amplification. Previously these primers were successfully amplified in *Hordeum*, *Allium*, *Oryza*, *Secale*, *Nicotiana* and *Antirrhinum* (https://www.le.ac.uk/bl/phh4/prretros.htm). Apart from above primers, we opted DVO144 and 10712 primers from the previous reports and amplified 600 bp of RT fragment from *Pongamia*, *Mesua*, *Jatropha* and *Ricinus* LINE (Fig 3.3) and (Table 3.1) (Wright et al., 1996). These primers were previously amplified in *Arabdopsis* and *Vicia* (Wright et al., 1996). Similarly, we tried to isolate SINEs in *Pongamia* but did not get any PCR amplification probably due to the absence of primer binding site. The SINEs are very difficult to isolate due to their heterogeneous nature. Hence, the identification and isolation of SINE is a tough task because of difficulty in targeting *RNA polymerase III* promoter regions through PCR (Borodulina and Kramerov, 1999). Nevertheless, SINEs have been well characterised in mammals and in some plants.

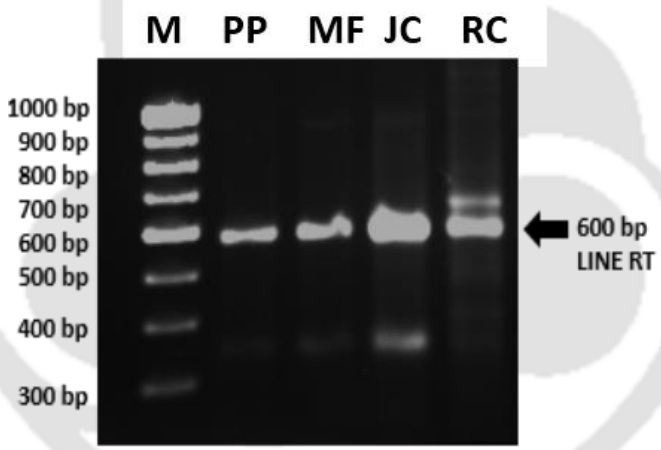


Figure 3.3. PCR amplified product of RT of LINE retrotransposons. M: 100 bp DNA ladder, PP: *P. pinnata*, MF: *M. ferrea*, JC: *J. curcas* and RC: *R. communis* LINE PCR product.

The whole PCR amplified product of RT from *copia*, *gypsy* and LINE were cloned into TA cloning vector as mentioned in methodology. Multiple transformed colonies of RT were opted for colony PCR. The positive colonies were selected for plasmid isolation and followed by Sanger sequencing. After sequencing, the homology-based search was carried out (BLASTN and BLASTX) for cloned retrotransposon sequences. Clones revealed the similarity (approximately 60-85% identity) to those of annotated retrotransposons from other plant species. All retrotransposons clones were again

reconfirmed using *Gypsy* database 2.0, Repbase and RepeatMasker programs. The annotated sequences were deposited to Genbank under the accession number: *Pongamia* Ty1-*copia* RT (KP202847.1- KP202834.1, MH397570- MH397584), *Ricinus* Ty1-*copia* RT-RH (MH397585), *Jatropha* Ty1-*copia* RT-RH (MH397586), *Mesua* Ty1-*copia* RT-RH (KU507530.1), *Pongamia* Ty3-*gypsy* RT (MH397537- MH397566), *Ricinus* Ty3-*gypsy* RT-RH (MH397569), *Jatropha* Ty3-*gypsy* RT (MH397568), *Mesua* Ty3-*gypsy* RT (MH397567), *Pongamia* LINE RT (MH397508- MH397533), *Ricinus* LINE RT (MH397534), *Jatropha* LINE RT (MH397535) and *Mesua* LINE RT (MH397536).

The *copia* clones were rich in AT bases, with an average AT/GC ratio of 1.47, which is similar to Ty1-*copia* identified in other species (Stergiou et al., 2002). The translation of these PCR amplified sequences implied that Ty1-*copia* RT sequences contained in-frame stop codon(s). Hence, the present sequences did not have potential functional RT fragment thereby depicting the transcriptional inactivity. Similarly, *gypsy* clones contained abundant AT bases, with an average AT/GC ratio of 1.45. In case of LINE, clones were AT-rich. The average AT/GC ratio of LINE RT was 1.31, slightly lower than *copia* and *gypsy* elements. Like *copia* clones, *gypsy* and LINE sequences harboured stop codon. This means that all clones investigated in this study were transcriptionally inactive in the *Pongamia* nuclear genome.

Beside retrotransposons, we also tried to isolate the DNA transposons in *Pongamia* genome using previously published degenerate primers, but we did not get any desirable results. For the isolation of DNA transposons, several methods have been employed in the past: EST and genomic library screening; *in silico* analysis and hybridisation through the heterogeneous probe (Rubin et al., 2001, MacRae et al., 1994). According to Hartings et al. (1998), cross-hybridisation of DNA transposons like *hAT* is difficult across taxa. Even the development of PCR assay is not straightforward considering the level of diversity present in between DNA transposons (De Keukeleire et al., 2004). Despite the high population in eukaryotic genomes, their isolation and annotations are difficult considering the rapid sequence evolution (Xiong et al., 2014). Due to unavailability of *Pongamia* genome database, we were not able to design the specific primers for DNA transposons. Therefore, we opted for *in silico* method to isolate active TEs from *Pongamia* transcriptome libraries. The details of the investigation are mentioned in Section No. 3.4.3.

3.4.2. Isolation of transposable elements in *Pongamia* organellar genome

With an aim to identify the TEs in the chloroplast, we searched chloroplast (CP) genome of length 152.9 kb against the repbase Panicoid database. Analysis revealed the presence of 12 fragments of TEs in CP genome (Fig 3.4). Among all TEs, LTR-retrotransposons were least present and only one copy of Ty3-*gypsy* was found having a length of 120 bp. DNA transposons were present in abundance. Investigation showed the presence of four copies *En/Spm* elements, highest among all CP DNA transposons. The individual fragments of TEs were located in different regions of the CP genome. Some TEs were found in the coding region, but most of them were present in the intergenic region.

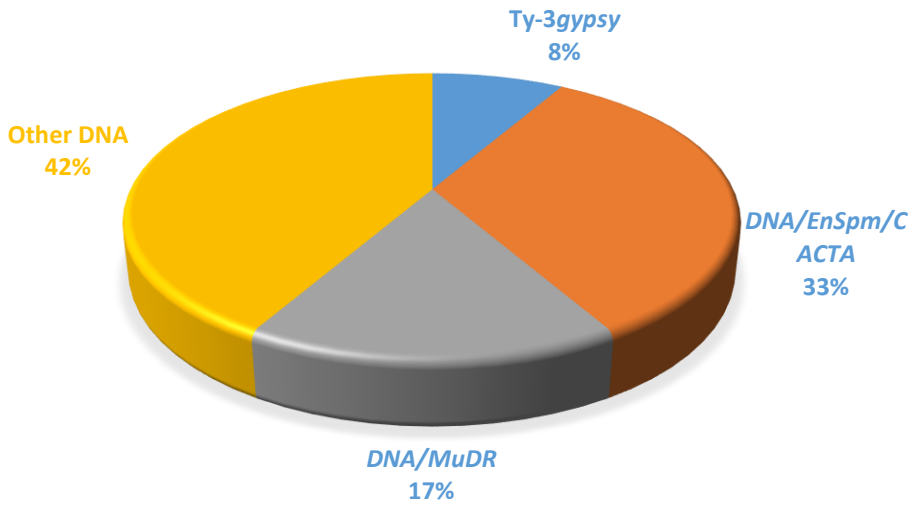


Figure 3.4. Pie chart showing the population of transposable elements in the chloroplast genome of *Pongamia*.

Mitochondrial (MT) genome (425.7 kb) was also searched against the Panicoid database of Repbase. A total of 24 TEs were identified in the MT genome. Around 20 fragments of retrotransposons were observed, in which Ty3-*gypsy* (12 copies) were found in an abundance (Fig 3.5). Like CP genome, no SINE elements were observed in the MT genome. The TE fragments were spread across the entire MT genome, including both the coding and intergenic regions. A total of 2.942 kb (0.66%) length of TE sequences were identified. An earlier study showed the presence of TEs in MT genome of perennial ryegrass (Islam et al., 2013). Similarly, Alverson et al. (2011) found the *copia* and *gypsy*-like retrotransposons in *Citrullus* and *Cucurbita* MT genomes having a total length 24 kb (6.4%) and 21 kb (2.1%) respectively. In contrast, not a single copy of transposon was found in the *Spirodela* MT genome (Wang et al., 2012). Unlike CP genome, *En/Spm*

elements were absent in present investigations. Homology search with BLASTX showed that some partial fragment of RT domain of retrotransposons were found dispersed in MT genome. BLASTX and Repbase results displayed the similarity of CP and MT genome TEs with nuclear TEs present in the database. This confirmed that the organellar TEs were probably derived from the nuclear genome.

All the observed fragments in organellar genomes were not complete or full-length TEs and contain the distorted reading frame. These fragments were regularly occupied by stop codons. Some of the fragments were present in coding sequences and probably translated as protein in the organelle.

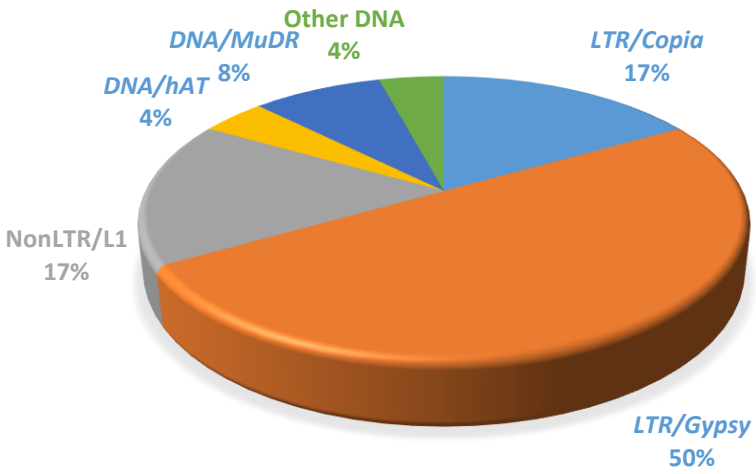


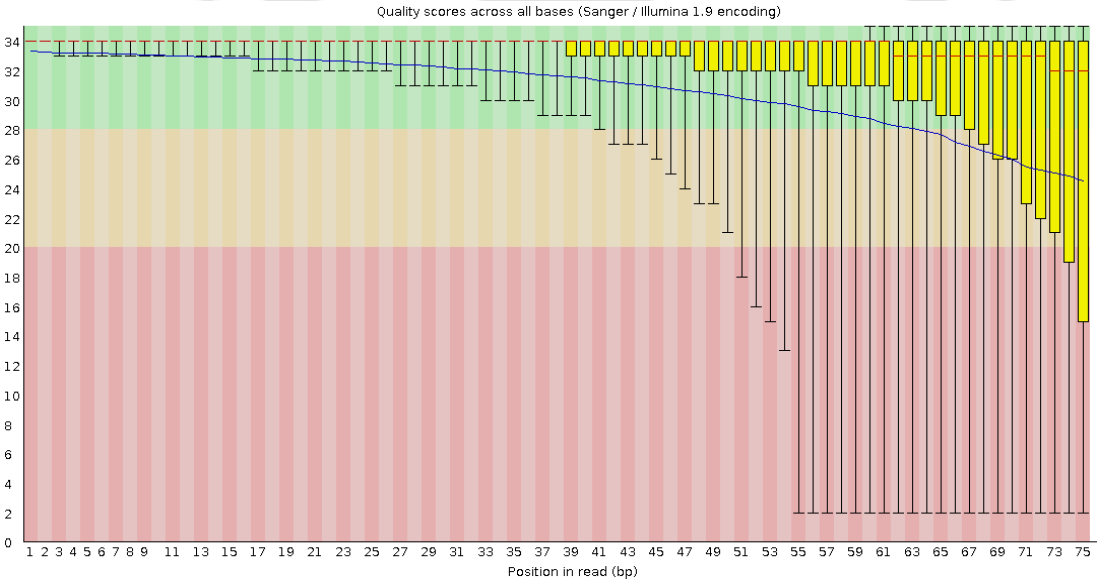
Figure 3.5. Pie chart showing the population of transposable elements in the mitochondrial genome of *Pongamia*.

3.4.3. Transcriptional activity of TEs in *Pongamia* unigene libraries

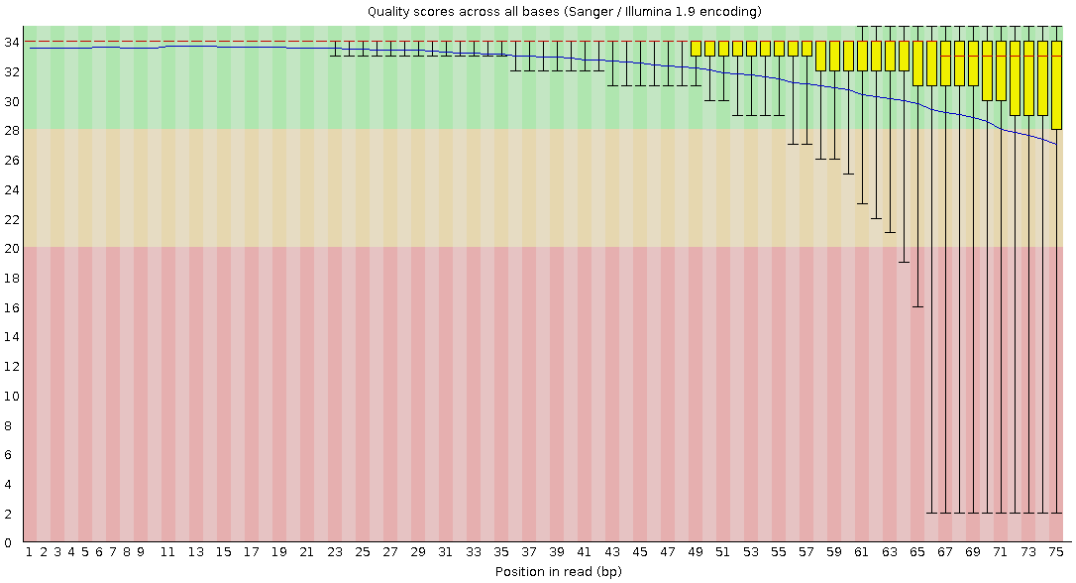
We assembled the transcriptome libraries before mining of transcriptionally active TEs. The details of the results are discussed below.

3.4.3.1. Illumina reads pre-processing and transcriptome assembling

To observe an overview of *Pongamia* transcriptionally active TEs, high throughput sequencing Illumina 2x75 bp paired-end reads were retrieved from NCBI under project accession number SRA046342.1. *Pongamia* libraries SRR349650 (MpRs), SRR349651 (MpRf), SRR349652 (MpLs), SRR349653 (MpLf) were downloaded using SRA toolkit **(**Huang et al., 2012). Before assembly, quality check of RNA-seq libraries was performed with a FastQC tool (Fig 3.6). The libraries were filtered for adapter sequences using the Trimmomatic program. The libraries were again cleaned for low quality reads with Q20 bases parameter through Trimmomatic. About ~87-90% of clean reads were obtained after cleaning and filtered all libraries (Table 3.2). Reads less than 35 bp length were removed from libraries before assembling. Around 10-13% of Illumina reads were removed before analysis. The observed GC content of all reads was present in between 42-44 %. The average QC of all reads was around 33, indicating the high quality of reads for transcriptome assembly (Fig 3.7). Unigenes less than 200 bp were removed from assembled libraries. Due to near similar results of FastQC, we only represented the quality analysis (QA) graph for MpRs library.

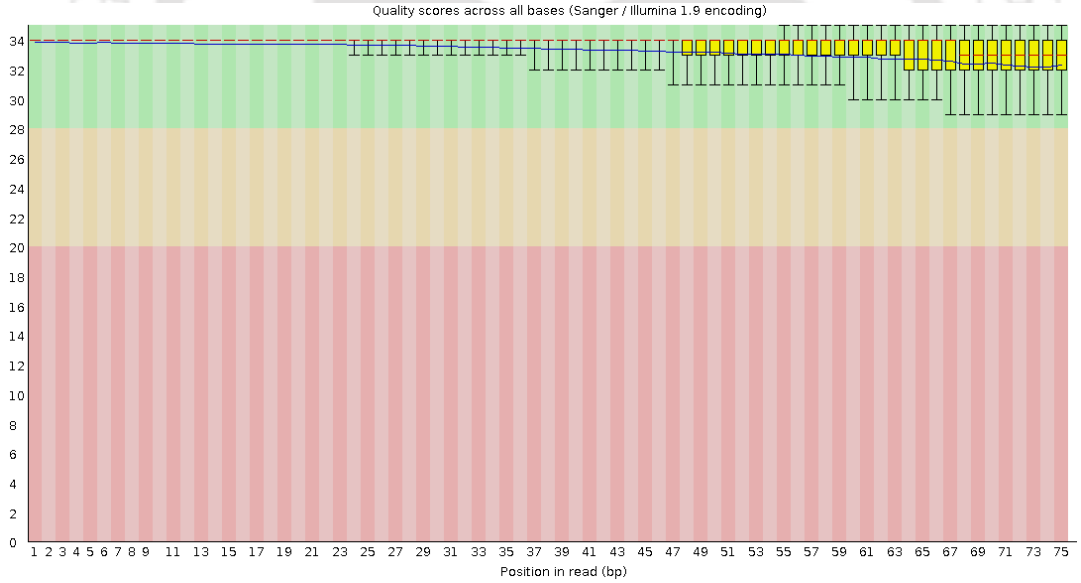


A) SSR349650.1 (MpRs)

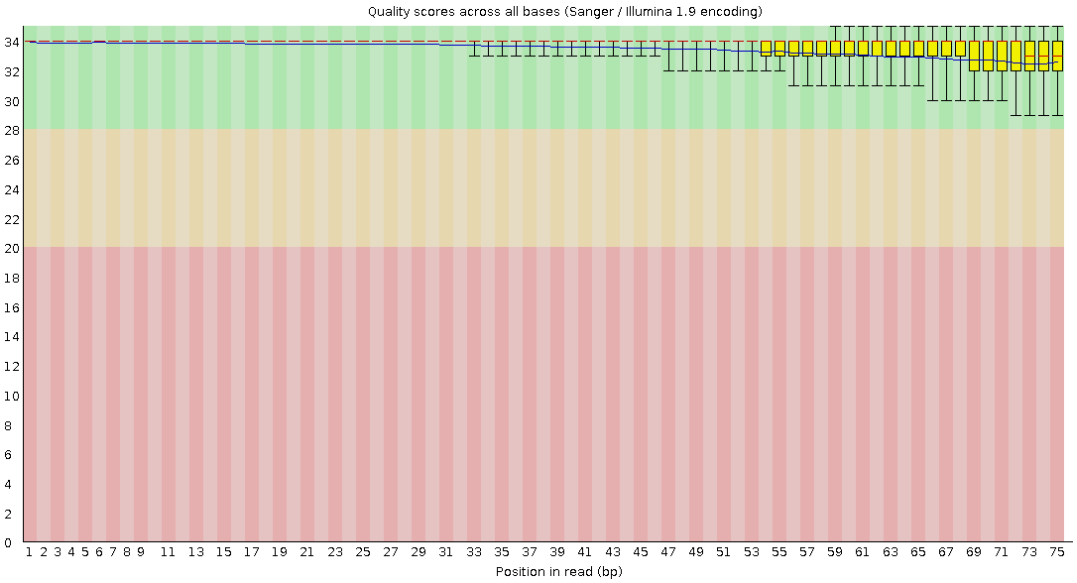


B) SSR349650.2 (MpRs)

Figure 3.6. QA graph extracted from the FastQC report before processing library A) SRR349650.1 (MpRs) B) SRR349650.2 (MpRs).



A) SSR349650.1 (MpRs)



B) SSR349650.2 (MpRs)

Figure 3.7. QA graph extracted from the FastQC report after processing with Trimmomatic program A) SRR349650.1 (MpRs) B) SRR349650.2 (MpRs).

The clean paired-end reads from all the libraries were assembled using Trinity assembler. The minimum length of assembled unigenes was 200 bp. A total 113009, 115955, 94826 and 98406 unigenes were assembled from MpRs, MpRf, MpLs and MpLf libraries respectively. Secondary assembling and clustering were carried out using the CAP3 assembly program with 50 bp overlap and 90% of identity. The unigenes were further used for removing the sequence duplicates. The redundant unigene assemblies were removed using CD-HIT-EST (v4.6.1) tool with 95 % identity threshold. A total of 58910, 60000, 49815 and 50556 unigenes were generated from MpRs, MpRf, MpLs and MpLf libraries, with a minimum size of 200 bp (Table 3.2). The highest number of unigenes were observed in root samples followed by the leaf samples using trinity assembler. These observations are corroborated with a previous study (Huang et al., 2012). However, the total number of unigenes assembled in existing transcriptome libraries were less compared to earlier assembled transcriptome by Huang et al. (2012). The difference in a number of unigenes in transcriptome libraries was due to the choice of different cleaning program and assembler used in transcriptome assembling.

Table 3.2. Statistics of unigenes assembled using Trinity assembler.

S.N	Content	MpRs	MpRf	MpLs	MpLf
1	Total sequences	48000000	48000000	48000000	48000000
2	Illumina sequences after Trimmomatic cleaning	43335652 (90%)	43132146 (89%)	42945130 (89%)	42190620 (87%)
3	Sequence length	33-75 bp	33-75 bp	33-75 bp	33-75 bp
4	Trinity assembly	113,009	115,955	94,826	98406
5	Cap3 + CD-HIT-EST assembly	58,910	60,600	49,815	50,556

3.4.3.2. Mining of active transposable elements

We further investigated the presence of TEs in different transcriptome libraries. The libraries required for this study were assembled using raw RNA-seq libraries submitted under the project name (GE0Z00000000.1). In addition to above, one seed assembled library was also included for TEs mining study. The TEs were mined in transcriptome data through two independent screenings. Initially, all the ESTs were screened against Repbase database of 2,064 reference TEs from maize, sugarcane, sorghum and millet (*Panicoid*). To remove false positive results, sequences less than RM score 250 were discarded from the further investigation. The sequences with RepetMasker annotation were utilised for functional annotation through BLASTX against protein databases. The EST sequences annotated as 'Transposable elements' both by RepeatMasker and BLASTX were further utilised for analysis.

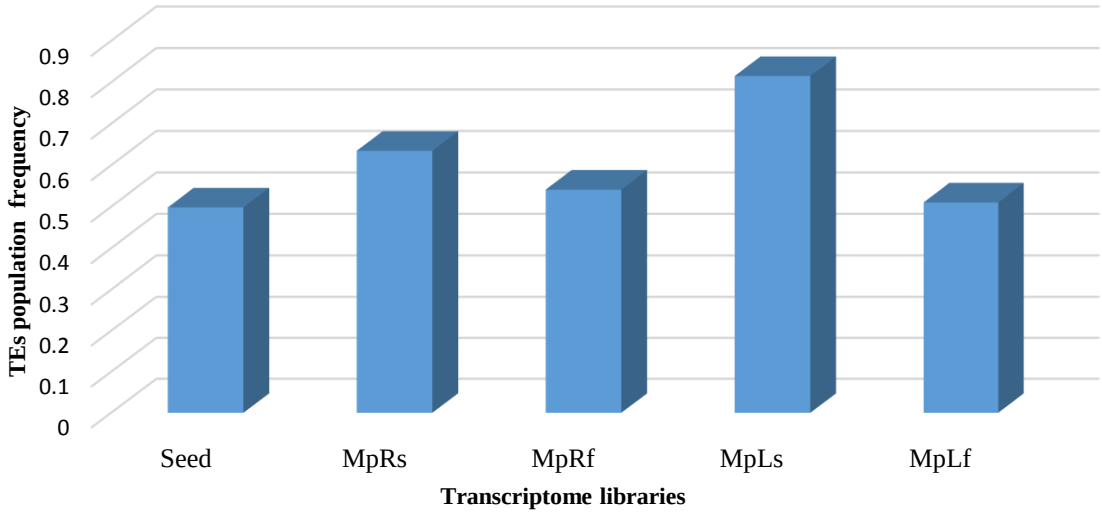


Figure 3.8. Bar chart showing the population of transposable elements in the EST libraries of *Pongamia*.

A total of 267 (seed), 374 (MpRs), 328 (MpRf), 406 (MpLs) and 258 ((MpLf) ESTs were found to have significant sequence similarity to TEs (Fig 3.8). In contrast, a higher number of active copies were reported in maize, common bean and sweet potato (Vicient et al., 2001, Gao et al., 2014, Yan et al., 2014). However, low copies of TEs were also detected in *Glycine max, Arabidopsis*, *Avena* and *Eucalyptus* (Vicient et al., 2001, Bacci Jr et al., 2005). SINEs were absent in all analysed EST *Pongamia* libraries. Of the total TE ESTs, around 77.5% to 89.5 % of ESTs were found to have significant similarity to retrotransposons. The presence of Ty1-*copia* ESTs was highest among the TE population. *Copia* ESTs were two to three-fold higher than Ty3-*gypsy* population. These results are consistent with a previous observation in *Oryza sativa* (285 *gypsy* in 1283 retrotransposons), sugarcane (19 *gypsy* in 128 active retrotransposons) and sweet potato (95 *gypsy* in 883 active retrotransposons) (Rossi et al., 2001, Yan et al., 2014). Unlike in plants, LINEs are abundantly present and often transcribed in mammals (Cordaux and Batzer, 2009, Gao et al., 2014). In case of non-LTR retrotransposons, LINEs were poorly represented (1-5 copies), which is consistent with a study in *O. sativa* (Yan et al., 2014). The highest (5) copies LINEs were detected in MpLs library.

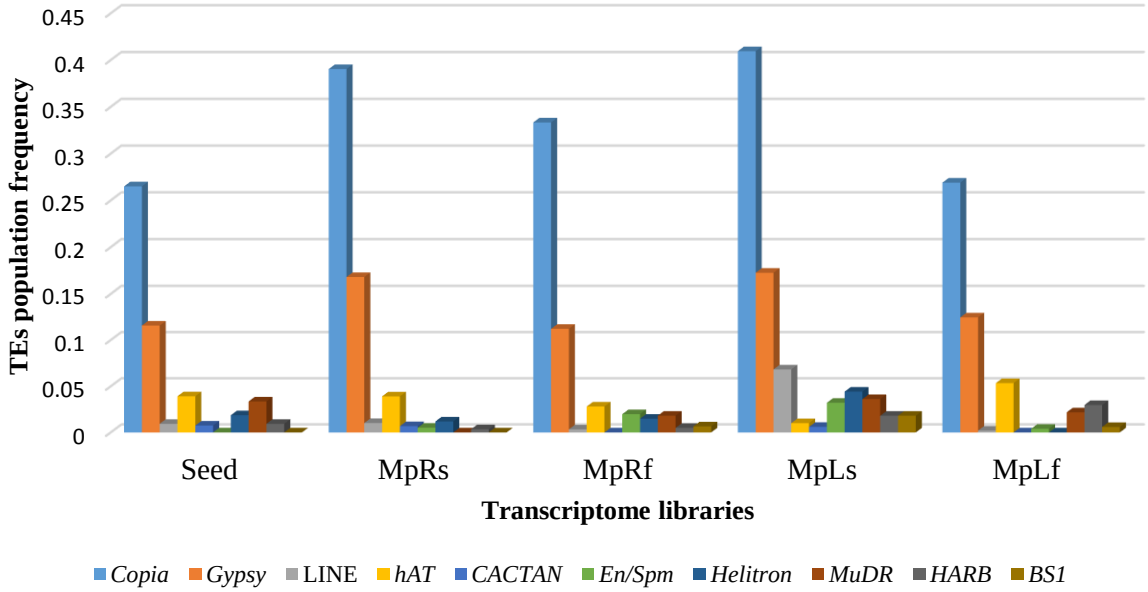


Figure 3.9. Bar chart showing the classification of different transposable elements population in the EST libraries of *Pongamia.*

Furthermore, the highest (82) and lowest (39) copies of EST DNA transposons were found in MpLs and MpRs libraries. Among the DNA transposons, *Helitrons* were present in abundant amount followed by *MuDR*, *En/Spm*, and *HARB* in different *Pongamia* libraries. Classification of TEs in each library is shown in Fig. 3.9. In all libraries, the frequency of TEs varied between 0 to 0.40%. Across all libraries, an average of 5.82 TEs was found per 1000 ESTs. This is a disagreement with the previous study where 1TE per 1000 ETSs was reported in *Eucalyptus* (Bacci Jr et al., 2005). Analysis revealed the low activity of TEs in *Pongamia* genome, considering the fact that TEs occupy the significant portion of the plant genome. Most of the identified TEs were partial sequences. This investigation confirmed that a tiny population of TEs were expressed in *Pongamia* genome.

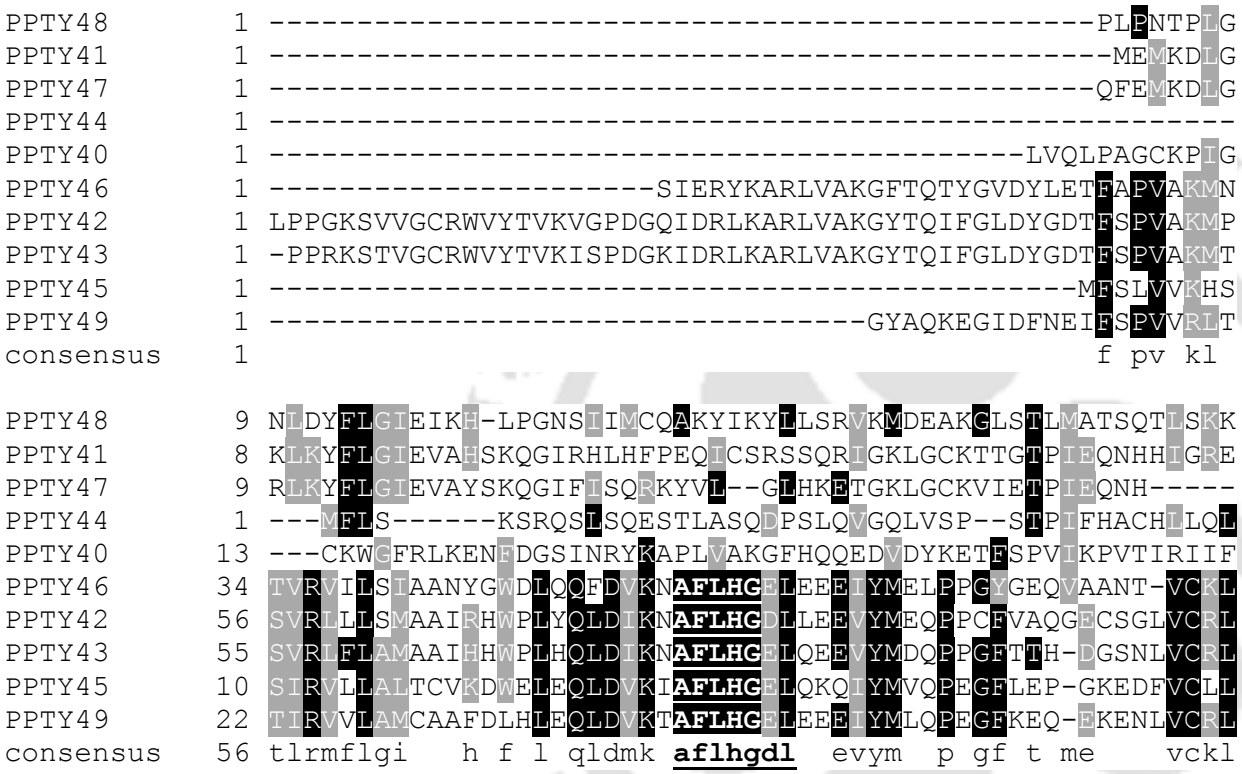
EST profiling of TEs has been conducted in the different plants such as a member of Gramineae, tomato, maize and common bean (Vicient et al., 2001, Cheng et al., 2009, Vicient, 2010, Gao et al., 2014). TEs identification in EST databases renders numerous advantages over other classic strategies. No prior knowledge of sequences is required for TEs identification which is essential for PCR primer designing. Furthermore, homology-based mining studies can identify the diverse superfamilies of TEs in quick succession. The application of TEs profiling in ESTs is helpful and appropriate in *Pongamia* as no TEs

have been identified to date. TEs accumulate many mutations, as a result, they are inactive in the genome. However, some TEs have the capacity to initiate the transcription. Organisms have mechanisms to protect their genome from the deleterious effect of TEs integration through methylation and transcriptional silencing. However, due to some biotic and abiotic stress conditions, some families of TEs transcribe and integrate into the genome. In the present study, we observed the increase in transcription activity of TEs against the salt stress in the root (MpRs) and leaf (MpLs) library (Fig. 3.8). This is not surprising because stress is responsible for TE activation. In contrast, TEs transcription was also detected in seed, leaf (MpLf) and root (MpRf) tissue in normal conditions. Similarly, Ahmed et al. (2011) detected the active Ty1-*copia* elements in jute leaf under normal conditions. Earlier, the activity of plant retrotransposons was often detected in leaf and root tissues (He et al., 2010, Ahmed et al., 2011). Recently, transposable elements like retrotransposns_gag, *MuDR, Ptta* and *En/Spm* were reported in *Pongamia* leaf transcriptome (Wegrzyn et al., 2016). Transcriptional activity of TEs was demonstrated in many plants, mostly for LTR-retrotransposons (He et al., 2010). Among the LTR-retrotransposons, *copia* elements were shown to be active in rice as *Tos* elements, *Tnt1* and *Tto1* in tobacco, *Rider* from tomato and BARE-1 in barley (Flavell et al., 1992, Cheng et al., 2009). In *Pongamia,* we have not classified the lineages of all active TE elements. However, the Repbase generated some primary annotation based on lineages for some active elements. Based on these primary annotations, in Ty1-*copia* family, we found *Maximus, Hopscotch and ivana* lineages in analysed EST libraries. Similarly, Ty3-*gypsy* family, we have detected *Reina, Del, Athila, scAle* and *tat* lineages. The transcriptional activity of *Hopscotch* was also reported in maize (Vicient, 2010). Some non-random pattern of distribution in TEs could well be due to the existence of different transcription mechanisms in TE families.

Our study demonstrated that the active TEs are present in the root, leaf and seed in normal and stress condition. However, it is worth mentioning that, the analysis was carried out using a limited number of plants reference TEs. Hence, there would be some chances where we might have missed the annotation of some key TEs. In addition, the stringent criteria used for TEs annotation through RepeatMasker and BLASTX might have lost some vital information.

3.4.4. Multiple sequence alignment

The multiple sequence alignment of LTR retrotransposons was carried out using the Muscle tool. For the analysis, ten and eleven RT partial EST sequences of Ty1-*copia* and Ty3-*gypsy* were opted from *Pongamia* transcriptome libraries. The nucleotide sequences were converted to amino acid for conducting MSA. We did not include the cloned PCR product for alignment as they contain stop codon in coding regions. Presence of stop codon is a general phenomenon in TEs, as most of TEs are nonfunctional in the plant genome. Still, some active elements are found in some organisms.



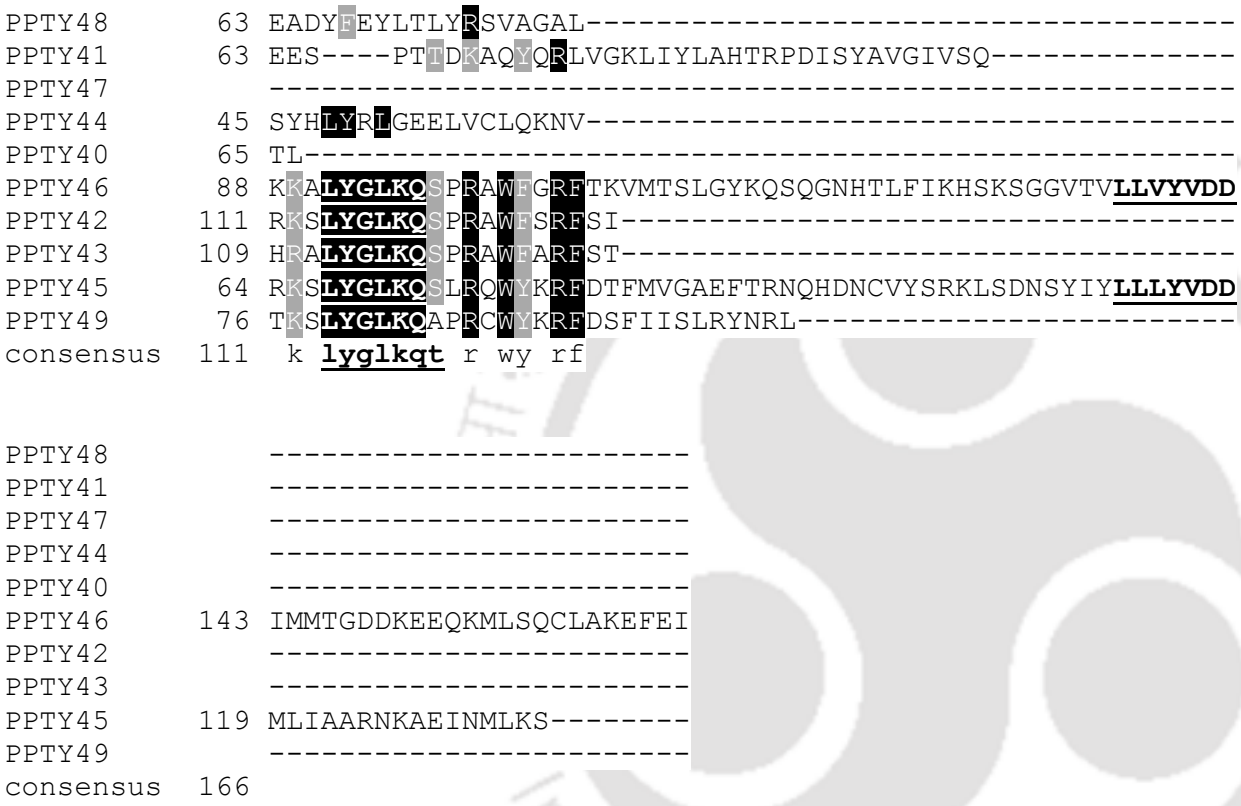


Figure 3.10. Multiple sequence alignment of deduced amino acid sequences of partial reverse transcriptase (RT) domain of Ty1-*copia* retrotransposons from *Pongamia* transcriptome library. Shaded letters represent the conserved residues.

For Ty1-*copia*, the highest sequence homology exists between sequence PPTY-42 (*P. pinnata* Ty1-*copia*) and PPTY-43 (80.47%). Some sequences showed conserved 'AFLHG' amino acid fragment at 5′ end which is characteristic of *copia* elements (Fig. 3.10). The same conserved amino acid fragment has been used in designing a PCR primer. Similarly, the second conserved amino acid fragment 'LLVYVDD' was found at 3′ end in PPTY-46 and PPTY-45. This motif was also mentioned in many literature for designing PCR degenerate primers. These observations are in agreement with previous studies conducted in datura Ty1-*copia* (Singh et al., 2011). Another conserved amino acid fragment 'YGLKQ' was present at the middle of the sequence. The similar conserved domain was reported in *copia* (Hirochika and Hirochika, 1993). As the studied sequences are not full length, hence most of the conserved motifs are absent in alignment.

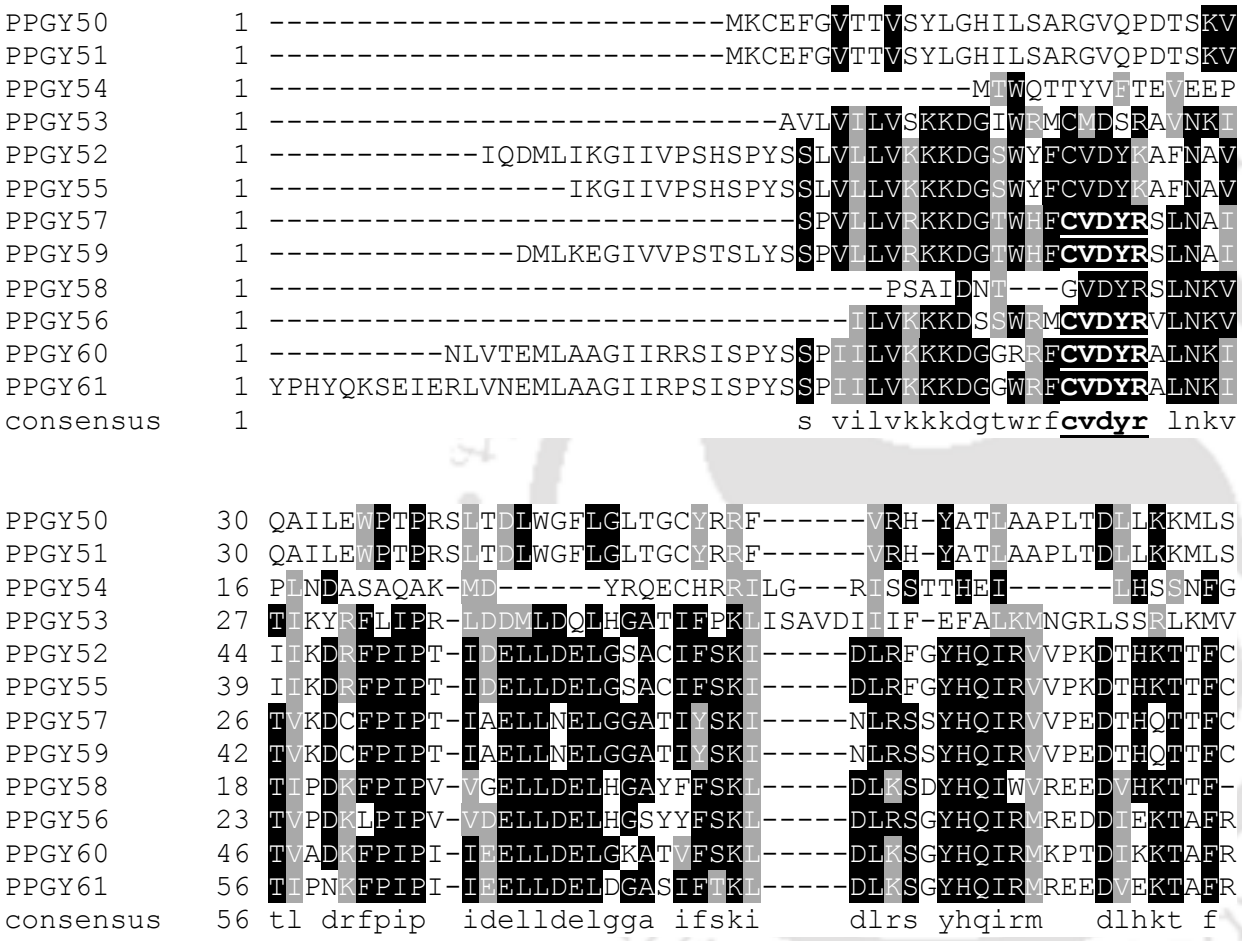


Figure 3.11. Multiple sequence alignment of deduced amino acid sequences of partial reverse transcriptase (RT) domain of Ty3-*gypsy* retrotransposons from *Pongamia* transcriptome library. Shaded letters represent the conserved residues.

For Ty3-*gypsy*, the sequence homology varied between 12.50 to 100%. The highest sequence homology exists between sequence PPGY-50 (*P. pinnata gypsy*) and PPTY-51 (100%). Similarly, the lowest sequence homology presents between sequence PPGY-50 and PPGY-61. Alignment of amino acid sequences showed that some sequences have the conserved 'CVDYR' fragment at 5′ end (Fig. 3.11). The same conserved amino acid fragment has been used in designing of PCR degenerate primer. These observations are in agreement with previous studies conducted in *Benincasa hispida* Ty3-*gypsy* (Jiang et al., 2013). Similarly, the second conserved amino acid fragment 'GHYEFLV' was found at 3′ end in *Pongamia gypsy* sequences.

MSA of LINE elements was not carried out due to unavailability of sufficient active elements in the library. The ESTs present in transcriptome libraries are partial sequences and do not regularly belong to a similar fragment (RT) of the gene. Therefore, targeting similar fragment (RT) of the gene in a limited number of sequences is difficult in EST library for multiple sequence alignment and phylogenetic analysis.

3.4.5. Phylogenetic study

The phylogenetic study was conducted using retrotransposon clones and some active retrotransposons elements from transcriptome library. The phylogenetic tree was constructed with 1000 bootstrap replicates using the NJ method with the aid of MEGA 6 program. The analysis revealed the classification of clones into different lineages of *copia*, *gypsy* and LINE elements.

3.4.5.1. Tyl-*copia*

To look for diversity between the Tyl-*copia* group of retrotransposons of different plant species and *Pongamia* clones, we conducted phylogenetic analysis using the neighbour-joining (NJ) method (Fig. 3.12). For diversity analysis, 28 genomic RT clones and 10 active RT sequences (EST library) were included for phylogenetic tree construction. In addition, single *copia* clone each from earlier characterised CPT-26, 28, 29 *Pongamia* tree were taken for analysis (Kesari et al., 2008). The phylogenetic tree revealed that the *Pongamia* RT clones separated into seven different groups. Though they belong to different plant family, still they shared sequence similarity for the RT domain of Ty1-*copia* element. There could be three reasons for the sequence homology of RT across the species:

1) there might be a horizontal transfer of a gene or the gene was conserved before the divergence; 2) they may contain variant sequences of RT due to its error-prone mechanisms of replication, so their RT were not grouped with same family plants; 3) the sequence similarity existed between RT across the species boundary probably due to the untranscribed or truncated sequences which existed before the species divergence. This phylogenetic tree also revealed the existence of different lineages of Ty1-*copia* family. In this study, we have determined seven groups based on *copia* lineages; group I belonged to *Tork*, which contained 17 numbers of *P. pinnata* Ty1-*copia* (PPTY) clones. Group II considered as *Bianca* and harboured 1 PPTY clone. The third group contained 2 PPTY clones and belonged to *Oryco* lineage. Group IV represented by *PyRE1G1* and contained 10 PPTY clone. The fifth group considered as *BARE-1*, which harboured highest 4 numbers of PPTY clones. The sixth group belonged to *Osser*, which contained 2 PPTY clone. The last group belonged to *Hopscotch*, which contained 2 PPTY clones. Analysis revealed the tendency of some clones to show close sequence similarity. This probably is due to the recent accumulation of replicated retrotransposons in the genome (Huang et al., 2012).

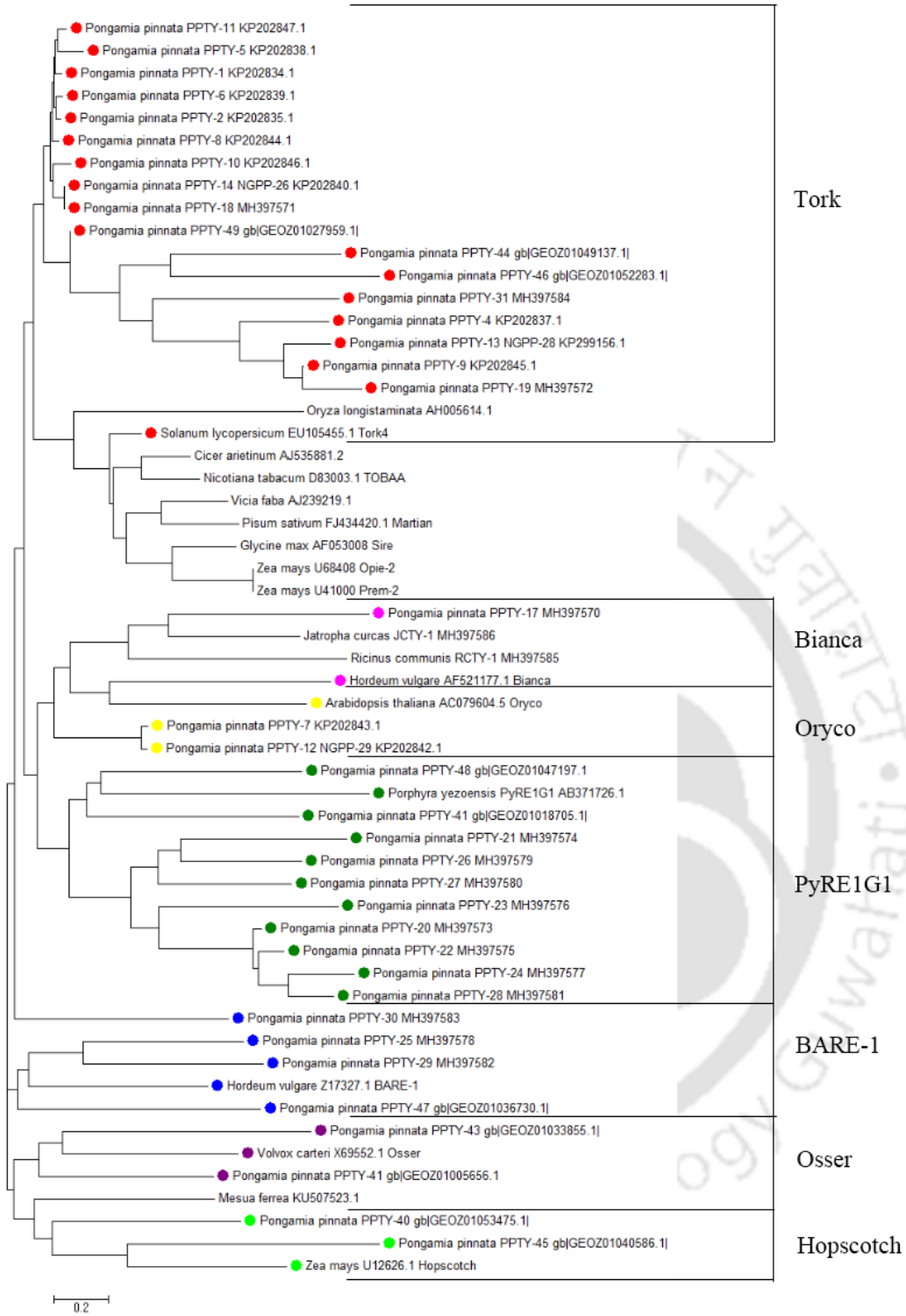


Figure 3.12. Phylogenetic analysis of the nucleotide sequences of RT domain of Ty1-*copia* clones from *Pongamia* and comparison with sequences from other species using the NJ method with 1000 bootstrap replicates.

3.4.5.2. Ty3-*gypsy*

To understand the diversity of Ty3-*gypsy* retrotransposons of *Pongamia*, we conducted phylogenetic analysis using the NJ method as described earlier. In the present investigation, a phylogenetic tree was constructed with the help of 31 *Pongamia* RT clones isolated from the nuclear genome. Along with nuclear RT, eleven active Ty3-*gypsy* were included from *Pongamia* transcriptome libraries. *Pongamia* RT clones were grouped with previously described Ty3-*gypsy* group retrotransposons from a different plant (Fig 3.13). The RT clones were mainly separated into seven groups with different species. These results suggest the existence of heterogeneity in RT sequences of Ty3-*gypsy* retrotransposons. Though they belonged to different plant families, still they share RT sequence similarity. The heterogeneity is probably due to 1) vertical, 2) horizontal transmission, 3) error-prone replication mechanism of RT with no proofreading activity, 4) presence of defective and truncated sequences and 5) the existence of lineages of Ty3-*gypsy*. *Pongamia* Ty3-*gypsy* divided into seven groups; the first group represented by *Athila metavirus* lineage and showed 21 numbers *P. pinnata gypsy* clones (PPGY), the second group belonged to *Calypso* 1-1 lineage and harboured around 5 numbers of PPGY clones. The third group harboured 2 PPGY clones and considered as *CRM/Del gypsy*. The smallest group IV and V belonged to *Maggy* and *Tat*, harboured each only 1 PPGY clone. The sixth group contained 9 PPGY clones and considered as *CRM/CR*. The last group showed a grouping of 2 PPGY clones and belonged to *Galadriel* lineage. The single clone PPGY-3 did not show homology with any characterised Ty3-*gypsy*, hence that clone remained ungrouped. This study is based on a limited number of clones. It is possible that we have not successfully amplified all lineages of retrotransposons due to the existence of sequence variation that sometimes hinders the binding of primers.

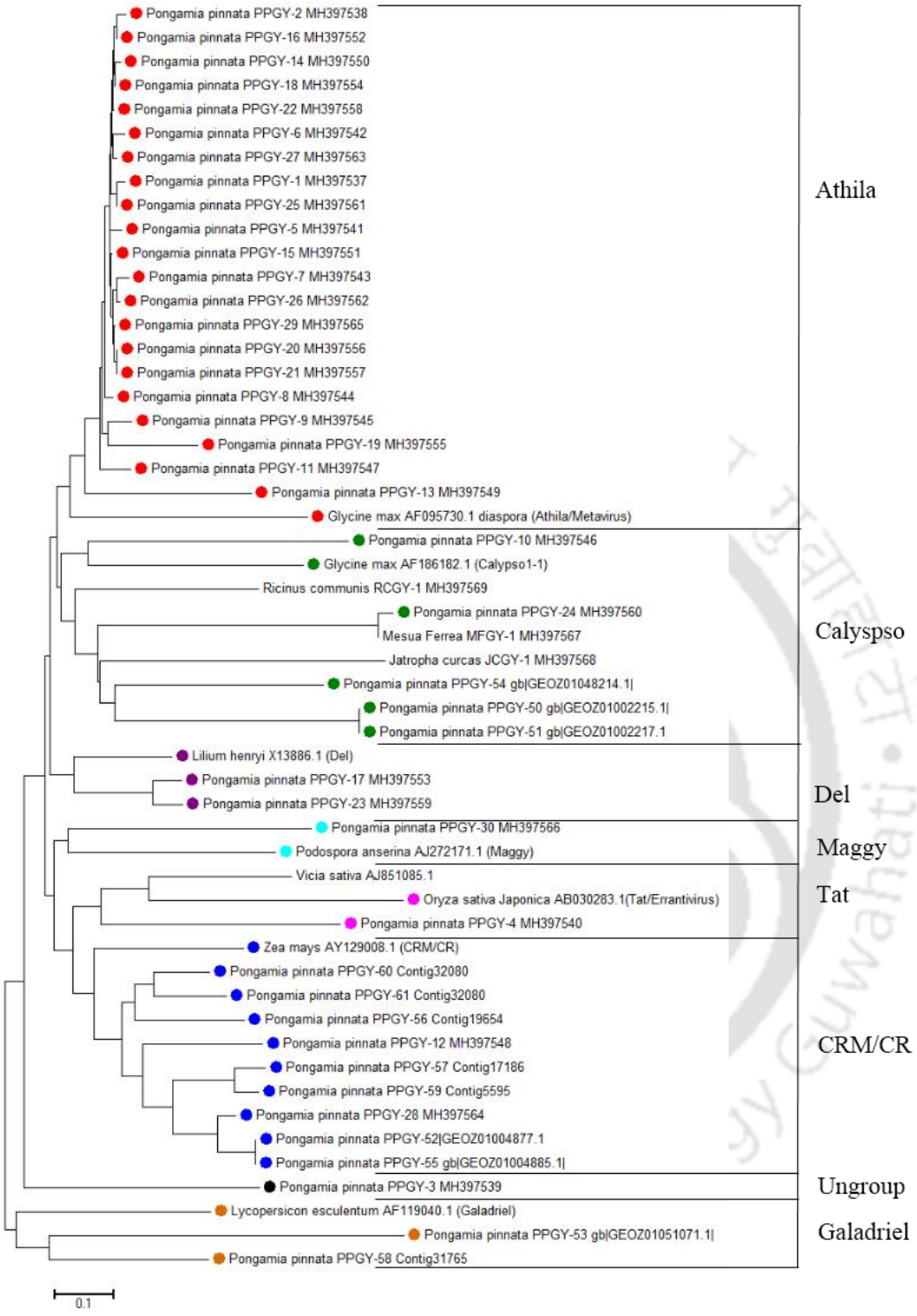


Figure 3.13. Phylogenetic analysis of the nucleotide sequences of RT domain of Ty3-*gypsy* from *Pongamia* and comparison with sequences from other species using the NJ method with 1000 bootstrap replicates.

3.4.5.3. LINE

Reverse transcriptase domain of LINE element isolated from *Pongamia* was subjected to phylogenetic study using the NJ method as described earlier. In the present study, for phylogenetic analysis, we included 25 clone isolated from the nuclear genome and 3 sequences from *Pongamia* transcriptome library. The study showed the grouping of *Pongamia* LINE elements separately with some exceptions where some RT formed a group with other organisms RT LINE (Fig 3.14). *Pongamia* LINEs mainly clustered into four different groups. The first group harboured twenty-four PPL clones that showed similarity with *Homo sapiens*. The second and third group contained PPL-29 (*P. pinnata* LINE) and PPL-26 RT clone respectively. The fourth group comprised of two clones PPL-18 and PPL-25. The fourth group is closely associated with the *Oryzae sativa* LINE *karma* lineage. Finally, this shows that the RT LINE elements are less heterogeneous than *copia* and *gypsy* elements. There have been very few reports available on LINE isolation and characterisation in plants (Leeton and Smyth, 1993, Wright et al., 1996, Kubis et al., 1998). In previous reports, LINE sequences were often shown as highly heterogeneous population (Kubis et al., 1998). A BLAST search with Repbase database showed that most of the isolated *Pongamia* LINE (PPL) clones are L1 types. This was also confirmed using NCBI-BLAST N and BLAST X search.

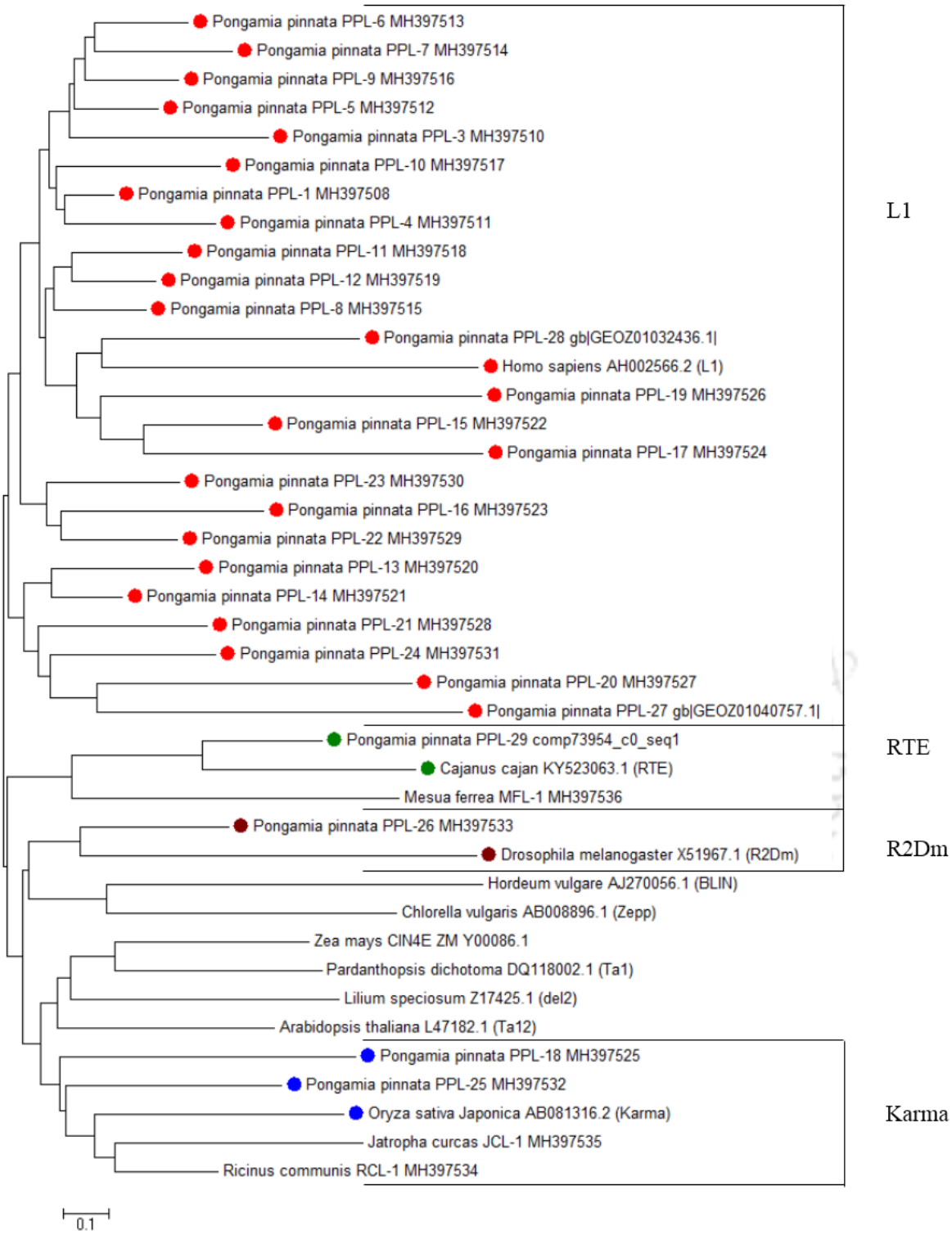


Figure 3.14. Phylogenetic analysis of the nucleotide sequences of RT domain of LINE from *Pongamia* and comparison with sequences from other species using NJ method with 1000 bootstrap replicates.

3.4.6. Estimation of copy number of retrotransposon in *Pongamia*

Dot blot hybridisation was performed to determine the copy number of retrotransposons in *Pongamia* genome. For all analysis, dot blot images were inverted for measuring the hybridisation signal intensities using ImageJ tool. Based on the hybridisation signal intensities of PCR product, the standard graph was prepared. Similarly, hybridisation signal intensities for genomic DNA was determined. The oversaturated hybridised dots were removed from further analysis. The selected values obtained from genomic DNA were compared with a standard graph. Further, these values were employed for the calculation of average proportion of nuclear genomic DNA hybridising to the probe.

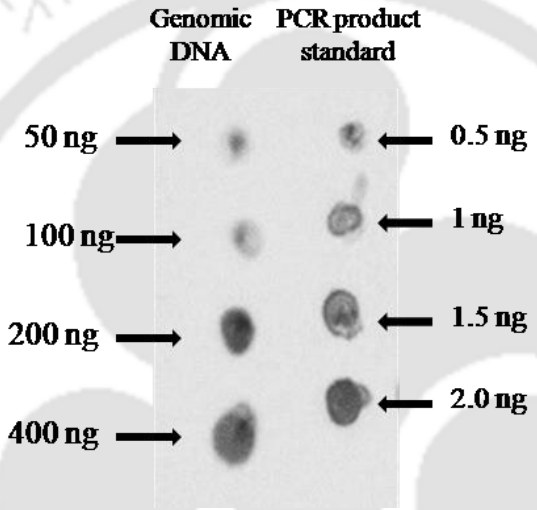
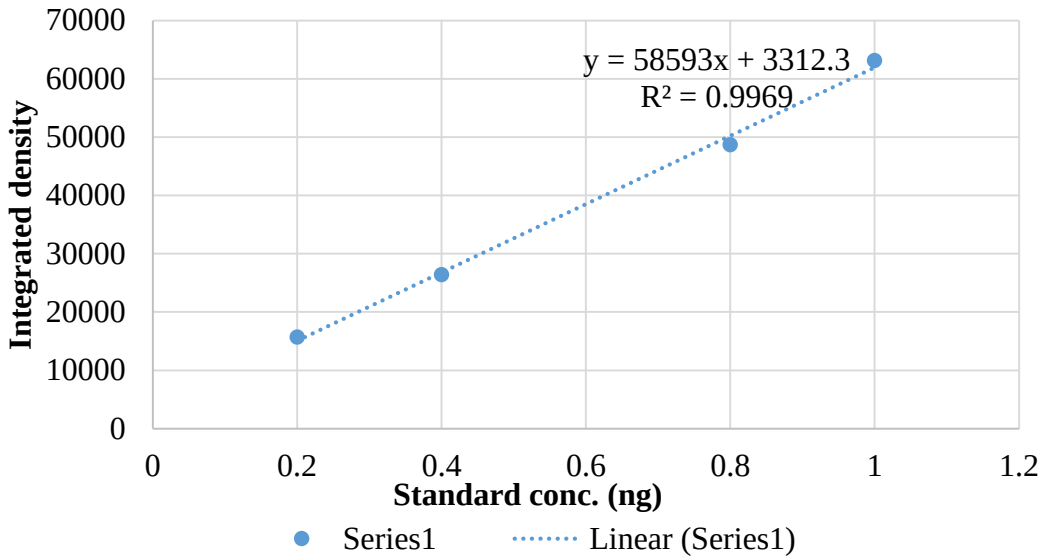


Figure 3.15. Dot blot hybridisation conducted for the determination of Tyl-*copia* like element copy number in *Pongamia*. Different concentrations of *Pongamia* genomic DNA serially diluted as 1. 50 ng, 2. 100 ng, 3. 200 ng, 4. 400 ng and RT-RH PCR product dilutions as 1. 0.5 ng, 2. 1 ng, 3. 1.5 ng, 4. 2 ng were dot spotted on a nylon membrane.

Initially, dot blot was carried out for Ty1-*copia* element. Whole PCR amplified product of RT-RH (~850 bp) of *Pongamia* was used as a probe for dot blot hybridisation. Serial dilutions of PCR product were used as standards against total *Pongamia* genomic DNA (NGPP-46) which is approximately 1,198 Mb in length as recorded in the previous study (Choudhury et al., 2014). After hybridisation, signal intensities of hybridised dots were determined by ImageJ tool. A standard graph was prepared based upon the corresponding intensities of each hybridised dots (PCR product) (Fig 3.15). The intensities of genomic DNA hybridisation were compared with a standard graph (Graph 3.1). The copy number of RT-RH gene in Ty1-*copia* of *Pongamia* was estimated to be

approximately 14,653 copies per haploid genome (Fig 3.15). Since the present study focuses only on partial RT-RH domains, assuming average size for Ty1-*copia* (7 kb) (Hill et al., 2005), this class of LTR retrotransposons may comprise up to 8.5% of total *Pongamia* haploid nuclear genome.



Graph 3.1. Standard graph prepared based upon the signal intensities of Tyl-*copia* (PCR product) dot blot hybridisation calculated by ImageJ program.

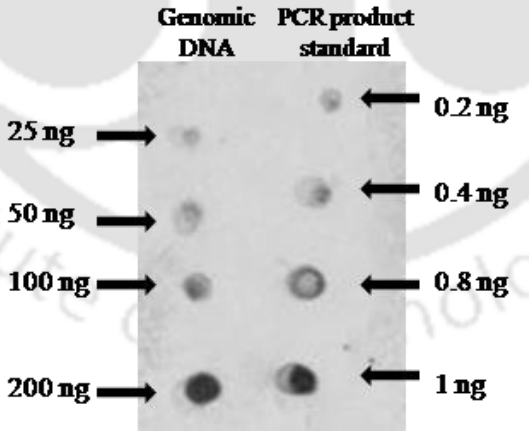
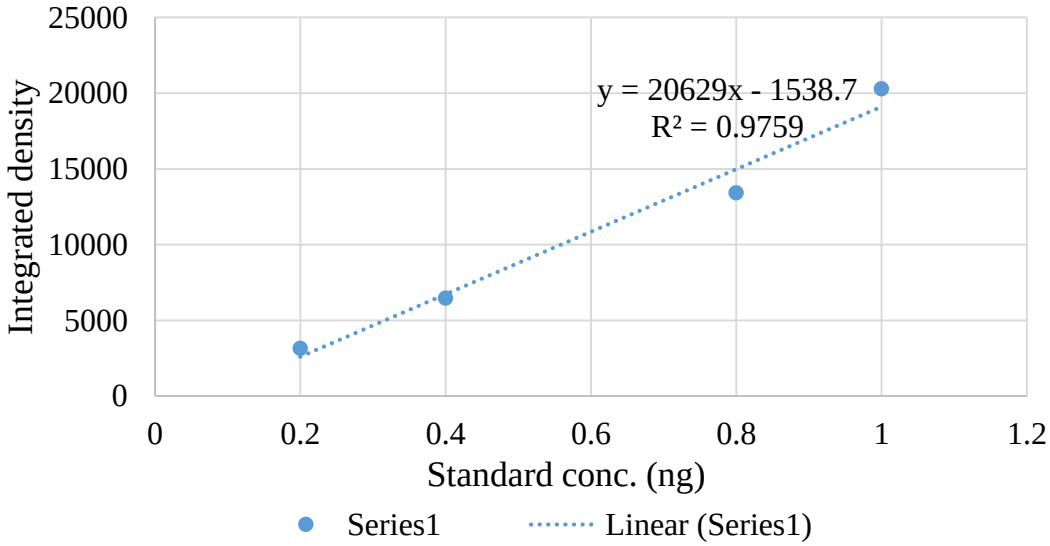


Figure 3.16. Dot blot hybridisation conducted for the determination of Ty3-*gypsy* element copy number in *Pongamia*. Different concentrations of *Pongamia* genomic DNA serially diluted as 1. 25 ng, 2. 50 ng, 3. 100 ng, 4. 200 ng and RT PCR product dilutions as 1. 0.2 ng, 2. 0.4 ng, 3. 0.8 ng, 4. 1 ng were dot spotted on a membrane.

Similarly, the copy number of Ty3-*gypsy* was estimated through dot blot analysis. Whole PCR amplified product of Ty3-*gypsy* RT (~420 bp) of *Pongamia* was used as a probe for dot blot hybridisation. Serial dilutions of PCR product were used as standards against total *Pongamia* genomic DNA (NGPP-46) which is approximately 1,198 Mb in length as recorded in the previous study (Choudhury et al., 2014).

Based on the signal intensities of hybridised dots, a standard graph was prepared (Graph 3.2). Signal intensities of genomic DNA dilutions were compared with a standard graph to calculate the average portion genomic DNA hybridised by a probe. The copy number of RT gene in Ty3-*gypsy* of *Pongamia* was estimated to be approximately 11,594 copies per haploid genome (Fig 3.16). Since the present study focuses only on partial RT domains, assuming average size for Ty3-*gypsy* (10 kb) (Hill et al., 2005), this class of LTR retrotransposons may comprise up to 9.67 % of total *Pongamia* haploid nuclear genome. An almost similar result was observed in Japanese apricot, where the reported population of Ty3-*gypsy* sequences were abundant and constitute at least 33.3% in the genome, while Ty1-*copia* content was 18.4% (Wang et al., 2010). The contribution of *copia* retrotransposons is usually higher than the *gypsy* elements in the genome (Wang et al., 2010). These results depict that the different types of LTR retrotransposons and their lineages proliferates at variable rates in different plant species. Due to this LTR-retrotransposons occupy a significant proportion of nuclear genomes in the plant kingdom.



Graph 3.2. Standard graph prepared based upon the signal intensities of Ty3-*gypsy* (PCR product) dot blot hybridisation calculated by ImageJ program.

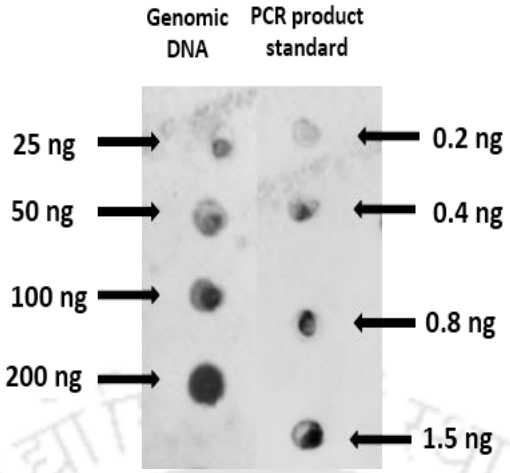
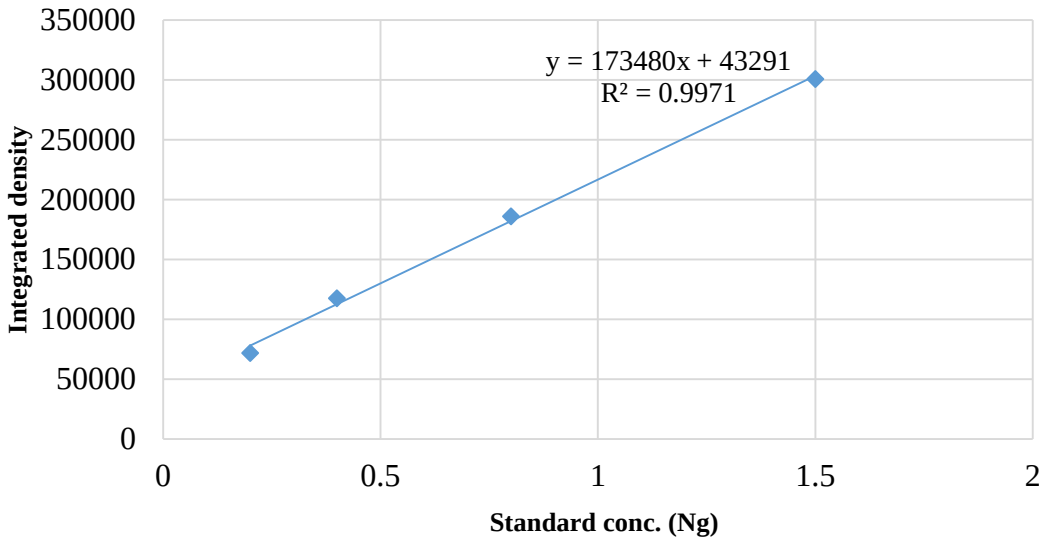


Figure 3.17. Dot blot hybridisation conducted for the determination of LINE-like element copy number in *Pongamia*. Different concentrations of *Pongamia* genomic DNA serially diluted as 1. 25 ng, 2. 50 ng, 3. 100 ng, 4. 200 ng and RT PCR product dilutions as 1. 0.2 ng, 2. 0.4 ng, 3. 0.8 ng, 4. 1.5 ng were dot spotted on a membrane.

Finally, the copy number of LINE was determined using dot blot. Whole PCR amplified product of LINE RT (~600 bp) of *Pongamia* was used as a probe for dot blot hybridisation. Serial dilutions of the PCR product were used as standards against total *Pongamia* genomic. DNA (NGPP-46) which is approximately 1,198 Mb in length as recorded in the previous study (Choudhury et al., 2014). Based on the signal intensities of the PCR product (standard) hybridised dots, a standard graph was prepared (Graph 3.3). The intensities of genomic DNA hybridisation were compared with a standard graph. Based upon the copy number formula, the copy number of RT gene in LINE of *Pongamia* was estimated to be approximately 18,621 copies per haploid genome (Fig. 3.17). Since the present study focused only on partial RT domains, considering the average size for LINE elements to be 4 kb (Becker et al., 2001), this class of LTR retrotransposons may comprise up to 6.0% of total *Pongamia* haploid nuclear genome.



Graph 3.3. Standard graph prepared based upon the signal intensities of LINE (PCR product) dot blot hybridisation calculated by ImageJ program.

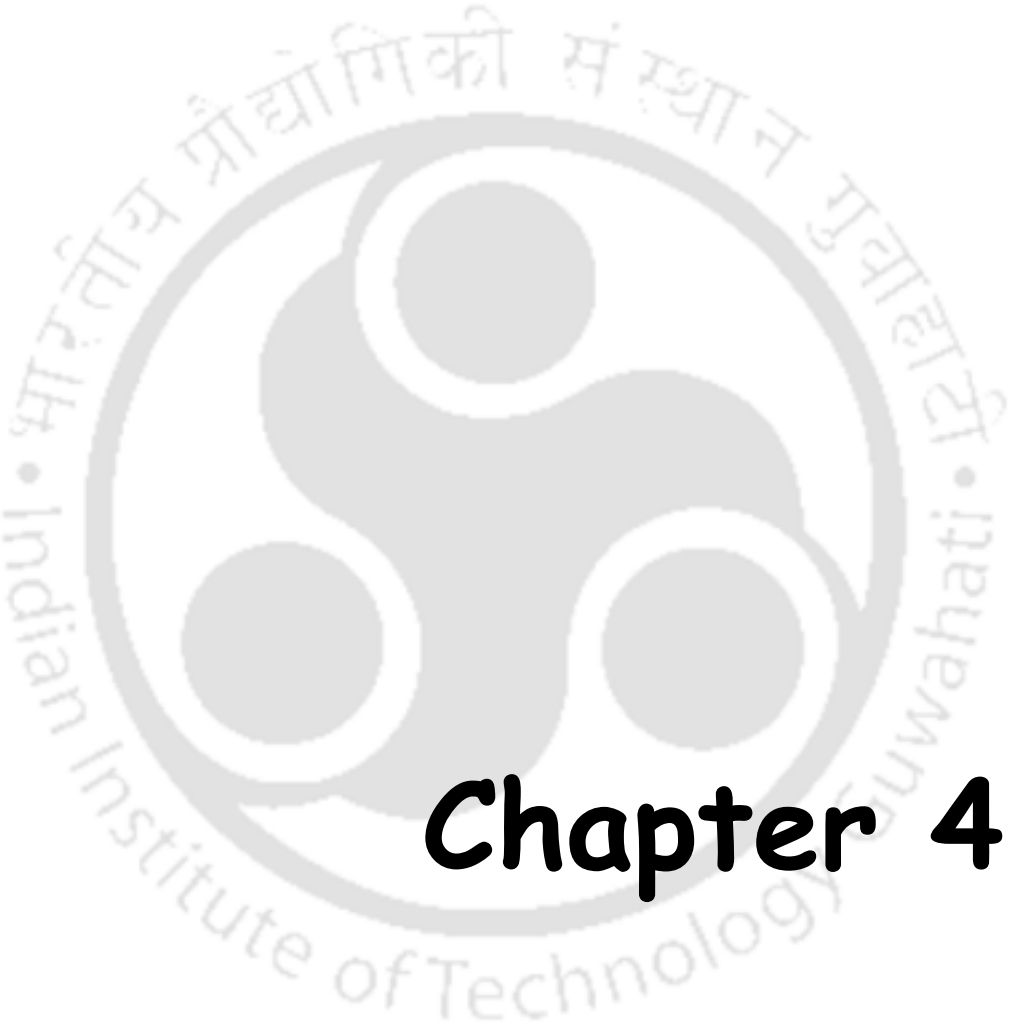
The estimated copy number of retrotransposons probably indicated the importance of retrotransposons in *Pongamia* genome evolution and its size. However, it is important to mention that limited numbers of retrotransposon were used to design the degenerate primers. Hence, some lineages of retrotransposons might not have been amplified or poorly presented in the PCR amplification using degenerate primers (Park et al., 2007). The whole PCR product was used for the preparation of a dot blot probe, which could also be responsible for non-specific hybridisation and results led to a little bit skewed copy number.

3.4.7. Synonymous and nonsynonymous substitution analysis

The analysis of the substitution pattern was conducted in *Pongamia* LTR retrotransposon (Ty1-*copia* and Ty3-*gypsy*) sequences. The potentially active functional RT domain were included from *Pongamia* EST library to estimate the synonymous and nonsynonymous substitution patterns. The accounted dN: dS ratio for Ty1-*copia* and Ty3-*gypsy* were 0.7866 and 1.0104 respectively. Substitution pattern ratio suggests the selective pressure acting on a protein-coding gene. A ratio higher than one suggests the positive or Darwinian selection. On the other hand, a ratio of less than one indicates the purifying selection and a ratio of one reflects the neutral selection. The purifying selection of RT sequences of Ty1-*copia* was reported in the yeast genome by (Jordan and McDonald, 1999). Our results for *copia* were also in agreement with previous studies in cultivated cotton (Zaki, 2005).

3.5. Conclusion

Transposable elements constitute a significant part of all plant genomes, and hence it is essential to understand their activity and systematic relationships within the genome. The current overview focuses on combined molecular genetic approach with a computer-based *in silico* study. Our investigation has identified the different LTR-retrotransposons in *Pongamia.* The phylogenetic analysis revealed the presence of different lineages of retrotransposons in the *Pongamia* genome, which are highly heterogeneous. However, the small fragmented population of TEs were also found in the organellar genome. The sequence similarity between retrotransposons from different species and the presence of sequence divergence in the same species point towards the horizontal transmission events that might have occurred during LTR-retrotransposons evolution (Woodrow et al., 2012). The results of TE ESTs profiling revealed the presence of different active superfamilies of TEs in *Pongamia* genome. In addition, *in silico* study showed the increase in transcriptional activity of TEs in response to salt stress. It is possible that the active TEs may have helped in alteration of *Pongamia* genome organisation, ultimately rendering a mechanism for biodiversity and genetic variation. Moreover, the results obtained from the present study can be used for the isolation of full-length active TEs. These active elements could be used to conduct a different epigenetic and molecular study in plant breeding programs.


भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Chapter 4

Chapter 4

Role of transposable elements in *Pongamia* unigene diversity

4.1. Introduction

Transposable elements (TEs) are a portion of the genome which can propagate or jump from one position of a genome to another. The abundance of TE in the genome differs significantly among different species. For example, *Saccharomyces cerevisiae* contains 3% TEs (Kim et al., 1998) whereas *Pinus* contains 90% TEs in a genome (Flavell, 1986, Pearce et al., 1996). For long, TEs were considered as 'selfish' or 'genomic parasite', due to their accumulation in a heterochromatic region and sometimes cause detrimental effects on gene function by the insertional disruption. Since the breakthrough discovery of TEs in maize, their actual role in genome evolution has been a theme of interest. However, polyploidisation and TE duplication are considered as responsible factors in reorganisation and expansion of the plant genome and evolution (Casacuberta et al., 2016, Wendel et al., 2016). The variation in genome size is greatly attributed to fluctuation in a population of TE elements in the genome, which may be due to TEs duplication or deletion in different species. This shows that the control of TEs could be varied between different organisms particularly between closely related plant species. According to Lee and Langley (2010), TEs present near or within a genic region will be deleted by natural selection due to their likely detrimental effect on the genes. In some cases, the accumulation of testing genes contributes to the evolution and adaptive characteristics. Moreover, TEs are also thought to be involved in the contribution of regulatory sequences that can control the function of close vicinity genes. Further, they are also responsible for genome evolution by contributing to the coding sequences. This happens when TE sequences are acquired as exon by host gene or mRNA, this process is called as exonisation. On some occasions, their insertion occurs in exon which is tolerated by natural selection without affecting the function of the host gene (DeBarry et al., 2006, West et al., 2014, Nystedt et al., 2013). Several recent studies have also provided information about TE insertion in functional genes of different plants like *Arabidopsis thaliana*, rice and coffee. However, up to what

magnitude TE exonisation contributes to protein-coding sequence variation is still unknown.

Genome and transcriptome sequences act as a valuable resource for the mining of different TE elements for expression studies (Gao et al., 2014). Genomics sequence analysis tools are important in understanding the possible adaptive role of TE in gene evolution and genome organisation (Shao et al., 2018). The introduction of next-generation sequencing (NGS) technologies has provided the new avenues for exploration of highly complex repetitive elements. With the advent of sequencing technologies, it has become easier to understand the TE mediated sequence variation through the application of comparative genomics (Yi et al., 2018). These sequence variations are probably caused due to mutation or integration of different repetitive elements. Availability of vast array of transcriptome data for non-model species has provided a vital insight of TE expression in different organisms. These databases are an important source for the study of transcriptome diversity resulted from TE activity. Despite the availability of a significant amount of transcriptomic data for many species, very few reports are available for active TEs. In *Pongamia* till date, no study has been conducted on the mining of transposable elements and their probable impact on protein-coding genes. The *Pongamia* transcriptome Illumina libraries available at NCBI provided us with the opportunity to assemble and analyse transcriptome sequences for TEs insertion. Moreover, the transcriptome sequences may also render us an opportunity to understand the possible impact of inserted TEs on the evolution and diversity of the host gene. In the current study, the following objectives are undertaken for the detailed study of TEs present in the coding region of genes.

1) Mining of TEs in unigenes and organellar genome.
2) Study of TEs in protein diversity and evolution.

4.2. Review of literature

Transposable elements are genetic units capable of moving within genomes and often making duplicate copies of themselves. As a consequence of this activity, they are mutagenic and can produce sequence variations and extensive genome rearrangements (O'Donnell and Burns, 2010, Ayarpadikannan and Kim, 2014). The activity of TEs can alter the pattern of gene expression and function (Feng et al., 2013, Oliver et al., 2013, Chuong et al., 2017). Moreover, TEs generate an enormous variability that can be used to create new genes, exons and new regulatory sequences (O'Donnell and Burns, 2010, Oliver et al., 2013, Bennetzen and Wang, 2014, Chuong et al., 2017, Pantzartzi et al., 2018). Besides their role in genome expansion, TEs contribute in the donation of transcription-regulating signals to genes (Conley and Jordan, 2012, Emera and Wagner, 2012, Bennetzen and Wang, 2014, Lynch et al., 2015). TEs are present in almost all organisms so far studied, but in some genomes likes *Zea mays*, they can represent about 60–80% of the nuclear genome (Meyers et al., 2001, Kidwell, 2002, von Sternberg and Shapiro, 2005).

TE-cassettes are fragments of TEs inserted into mRNA sequences (Lopes et al., 2013). The occurrence of TEs in intronic and intergenic regions has been widely reported (Zhang et al., 2011, Kannan et al., 2015, Guo et al., 2017). Further, it was demonstrated that these elements contributed substantially to the evolution of many genes at the transcriptional level through TE-cassettes (Sorek et al., 2002, Ganko et al., 2003, van de Lagemaat et al., 2003, Feschotte, 2008, Lopes et al., 2013). It has been proposed that TE-cassettes are generated after the activation of cryptic splice sites in an intron-residing TE sequence, or *de novo* through insertion into exons (Mitchell et al., 1991, Makalowski et al., 1994, Mao et al., 2015). Surprisingly, evidence also supports the translation of these cassettes that shows their contribution to proteome diversity (Hilgard et al., 2002, Hoenicka et al., 2002, Lin et al., 2016, Makałowski et al., 2017). The presence of TEs in the coding region is of great interest because they can change the function of the gene product. If this change is adaptive and conserved over evolutionary time, it is named as exaptation (Brosius and Gould, 1992, Brandt et al., 2005), molecular domestication (Miller et al., 1999), or co-opted events (Sarkar et al., 2003).

The presence of TEs in coding regions have been broadly studied in human genomes, but less studied in plant system (Turcotte et al., 2001, Meyers et al., 2001, Sakai et al., 2007). In human, Brownell et al. (1989) proposed that the rel proto-oncogene cDNA

contains an *Alu* fragment as a potential coding exon. Miniature inverted-repeat transposable elements (MITEs) are one of the DNA transposons which contributed to the evolution of novel genes by creating sequence diversity in rice genome and mRNAs (Oki et al., 2008). *Helitrons* are often reported to be a part of many sequence insertions, which led to the creation of novel genes through shuffling (Lopes et al., 2008, Barbaglia et al., 2012, Grabundzija et al., 2016). TEs were also reported for the diversity of protein in vertebrates (Chalopin et al., 2015, Makałowski et al., 2017). Along with exon donation and shuffling, TE cassettes contributed to the donation of regulatory sequences. These regulatory sequences are responsible for the expression of close vicinity genes. In some, cases TE-derived sequences are found to be responsible for the regulation of host genes (Shen et al., 2011, Trizzino et al., 2017, Chuong et al., 2017). Moreover, TE fragment insertions are associated with the changes in gene expression through the introduction of alternative splicing sites and polyadenylation sites into intronic or exonic regions (Davis et al., 1998, O'Donnell and Burns, 2010, Warnefors et al., 2010, Guo et al., 2016). The activity of TEs can change the expression of neighbouring genes by rendering their promoter and enhancer sequences (Bourque et al., 2008, Huda et al., 2011, Rebollo et al., 2012, de Souza et al., 2013). *Alu* elements contain several cryptic splicing sites present within its sequence which are responsible for alternative splicing sites in a host (Makaowski, 2000, Vorechovsky, 2010, Shen et al., 2011). Nigumann et al. (2002) reported the promoter sequences of L1 element responsible for transcription in several human genes.

Further, with regard to plant domestication, TEs are clearly involved in crop improvement and varietal diversification. Their activity results in a range of agronomically useful traits. Kawase et al. (2005) proposed the presence of diversity in waxy foxtail millet crops in East and Southeast Asia due to the insertion of multiple transposable elements. In sorghum, the presence of a MITE (*Tourist*) in the upstream region of organic acid efflux transporter locus (*AltSB*) responsible for enhanced expression of the *AltSB* gene at root apex, resulted in aluminium tolerance (Magalhaes et al., 2007). In soybean, *GmphyA2* is responsible for photoperiod sensitivity, the insertion of *SORE-1* was detected in *GmphyA2*. This insertion was only observed in soybean lines cultivated in the northern regions of Japan. This dysfunction of *GmphyA2* resulted in photoperiod insensitivity which allowed soybean cultivation at high latitudes (Kanazawa et al., 2009). The study of domestication

of different plant species provides help to comprehend the past and recent adaptive characters acquired through the activity of TEs.

Transposable elements are the major and most important part of the eukaryotic genome. In contrast to nuclear TEs, research on the mobile element of organellar genome is still lagging behind. Transposable elements are reported to be present in plant and yeast mitochondrial genome. Munoz-Lopez and Garcia-Perez (2010) investigated the presence of transposon-related sequences mainly LTR retrotransposons in the mitochondrial genome of melon. In *Arabidopsis,* several nuclear retrotransposons particularly Tyl-*copia*, Ty3-*gypsy* and non-LTR/LINE-families were reported to be part of the mitochondrial genome (Knoop et al., 1996). Four small size (50-277 bp) fragments similar to a known retrotransposon were detected in maize mitochondrial genome (Clifton et al., 2004). However, mitochondrial TE sequences are the fragments originated from a nuclear TEs (Knoop et al., 1996, Munoz-Lopez and Garcia-Perez, 2010). TEs are supposed to be involved in the transfer of sequences between organellar and nuclear genome in many organisms. Moreover, the distribution of TEs was positively correlated with the localisation of nuclear plastid DNA (NUPTs) and nuclear mitochondrial DNA (NUMTs) in *Arabidopsis* and sorghum (Michalovova et al., 2013). The insertion of TEs in the organellar gene has not been reported. In the *Pongamia* organellar genome, simple sequence repeats (SSRs) have been reported (Wang et al., 2017). But the population of transposable elements are still unknown.

Till date, no specific study is available on transposable elements in *Pongamia* plant. The study of insertion TEs in host coding sequences can be helpful to understand the impact of TEs on gene evolution. Computational methods can be useful in this regard to analyse the large data in a short time. Thus, *in silico* screening of TEs was carried out to identify the TEs fragments present in transcriptome and the organellar genome of *Pongamia*.

4.3. Material and methods

4.3.1. Presence of TE-cassettes in *Pongamia* transcripts

Four assembled unigene libraries from the previous chapter and seed unigene library retrieved from NCBI (GEOZ00000000.1) were screened to find the TE-derived fragments in protein-coding sequences (Huang et al., 2016). The unigenes were scanned for the occurrence of TEs using the RepeatMasker (RM) version 3.1.5 (http://www.repeatmasker.org) against a database of 2,064 reference TE sequences from maize, sugarcane, sorghum and millet (Panicoid). To avoid the false positive results, RM cut off scores equal or higher than 250 were opted, without imposing additional length thresholds. RepeatMasker was used for TE sequences orientations in unigenes. Furthermore, following parameters were used for analysis: (i) simple repeats and low complexity regions were removed, (ii) low complexity DNA sequences were not masked, (iii) high sensitivity/ low-speed search conditions.

4.3.2. Presence of TE-cassettes in organellar genome

Pongamia mitochondria and chloroplast complete genome sequences were retrieved from the National Center for Biotechnology Information (NCBI) GenBank database (GenBank accession no. JN673818.2 and JN872550.1). RepeatMasker tool was employed for the screening of TEs cassettes present in chloroplast and mitochondrial genome using the same parameters as described for *Pongamia* unigenes mining.

4.3.3. Annotation of unigenes containing TE-cassettes

The unigene sequences containing TE insertion were employed for annotation using BLASTX search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with a cut-off e-value = 1e-5 against protein databases such as PLAZA 2.0 and 3.0 (https://bioinformatics.psb.ugent.be/plaza/), Swiss-Prot (http://www.expasy.ch/sprot/), NR (non-redundant) database at NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.shtml#databases) and *G. max* database at Ensemble plants (plants.ensembl.org/). Unigene sequences annotated as TE were eliminated, as those sequences belonged to transcriptionally active TEs. Sequences which were annotated as protein and TE by BLAST and RepeatMasker tools were further selected for detail analysis. Thus, false positive identification of TE insertions in the

protein-coding genes of *Pongamia* was avoided. Additionally, unigenes containing TE-cassettes from all libraries were combined and further assigned to functional annotation with Panther online program (http://www.pantherdb.org/), gene ontology (GO) which classifies the sequences to molecular function, biological process, cellular component and protein class.

4.3.4. Gene structure prediction

Due to unavailability of *Pongamia* genome assemblies, the study of TE cassettes in coding region was conducted using genes from model plants. Thus, the gene sequences were retrieved from Ensemble, NCBI, SwissProt, *Vigna* genome and *Medicago* genome databases. Gene structure was determined using the FGENESH program. The structure prediction of LTR- retrotransposon was carried out through LTR Finder online tool (http://tlife.fudan.edu.cn/ltr_finder/).

4.3.5. Study of relation between TE-cassettes and host genes

The sequences required in phylogenetic analysis were obtained from the Ensemble, SwissProt, NCBI databases and genome databases. Nucleotide sequences were converted to amino acid sequences through FGENESH programme. The multiple sequence alignments of *Pyruvate decarboxylase* (*PDC*) protein isoforms of different plants were performed with Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) (Thompson et al., 1994). The evolutionary relationship between the isoforms of different PDC protein sequences was reconstructed through the Neighbor-joining (NJ) method using MEGA 6, with 1,000 bootstrap replicates analysis (https://www.megasoftware.net/).

4.4. Results and discussion

4.4.1. Mining of TE-cassettes in *Pongamia* unigenes

Before the discovery of TEs in maize by Barbara McClintock (1940's), mobile elements were regarded as useless and junk DNA. She proposed that these elements might play some regulatory role in genes after their activation (McClintock, 1965). The initial understanding about the role of TEs in the protein coding region or exon should be deleterious or harmful as they frequently disrupt open reading frame (ORF) and may lead to missense or nonsense mutations (Nekrutenko and Li, 2001, Sorek et al., 2002, Ayarpadikannan and Kim, 2014). However, in recent years, several reports have been published regarding the presence of TE fragment in the coding region of genes (Almeida et al., 2007, Lopes et al., 2008, Oki et al., 2008, Chiu et al., 2010, Butelli et al., 2012, Mao et al., 2015).

In the present study, screening of TEs was carried out in *Pongamia* unigenes. A total set of 53,586 of seed, 58,910 of (MpRs), 60,600 (MpRf), 49,815 (MpLs), 50,556 (MpLf) sequences were employed for TE mining with RepeatMasker using Repbase TE database. The unigenes showing RM score of less than 250 were discarded from the study. No penalty was kept for the length of TE cassettes present in protein coding sequences. A total of 2,064 sequences were included as RepeatMasker database from four plants viz. maize, millet, sorghum and sugarcane (Panicoid). RepeatMasker was also used in numerous investigations in the past for the analysis of repeat elements (Almeida et al., 2007, Sakai et al., 2007, Lopes et al., 2008, Lopes et al., 2013). A total of 894, 991, 1092, 831 and 705 candidate sequences with TE-cassettes were identified across the five *Pongamia* libraries. On average 1.9, 1.6, 1.7, 1.6, and 1.7 of TEs insertion per unigenes were observed.

Further, the unigenes harbouring TE cassettes were employed for functional annotation against a publically available protein database. Unigene sequences were annotated using BLASTX against databases such as PLAZA 2.0 and 3.0, Swiss-Prot, UniProt, NR database and *G. max* database at Ensemble plants. However, sequences with higher e-value than 1e-5 and sequence similarity less than 60% were discarded for further analysis. After annotation, a total of 339, 377, 352, 368 and 363 of unigenes were harboured around 480, 563, 475, 542 and 506 of TE cassettes in seed,

MpRs, MpRf, MpLs, and MpLf library respectively. Similar kind of study was conducted in coffee, where unigenes were annotated for the identification of TE-cassettes (Nekrutenko and Li, 2001). Nearly about 1.41, 1.49, 1.34, 1.47 and 1.39 of TEs insertion per unigenes was observed across five libraries after annotation. Previously, in *Bos taurus*, multiple insertions of TEs in nuclear genes were reported (18.41 insertions /gene), the insertion frequency is far higher than a present investigation in *Pongamia* (Almeida et al., 2007). The average frequency of TE present in unigenes varies between 1.44% to 1.83% in MpRs and MpLf respectively, which is higher compared to previous 0.18% frequency reported (35 LTR insertion) in *Caenorhabditis elegance* (Ganko et al., 2003). Similarly, low frequencies of TEs insertion were also reported in *Mus musculus* (0.14%) with 263 LTR insertions (DeBarry et al., 2006), *Drossophila melanogaster* (0.18%) with 25 LTR insertions and in coffee (0.18%) (Lopes et al., 2008). In case of *Oryza sativa*, around 2095 genes were identified containing 3508 copies of TEs in the exonic regions (Sakai et al., 2007), which is far higher than what is found in *Pongamia.*

4.4.2. Mining of TE-cassettes in *Pongamia* mitochondrial and chloroplast genome

Pongamia organellar genomes were also annotated using RepeatMasker with the same parameters which were used for unigenes. In mitochondrial genome, MuDR-N1_ZM DNA transposon was found to be inserted in coding genes. Two *trnK* genes were inserted by 58 bp length of the MuDR-N1_ZM element.

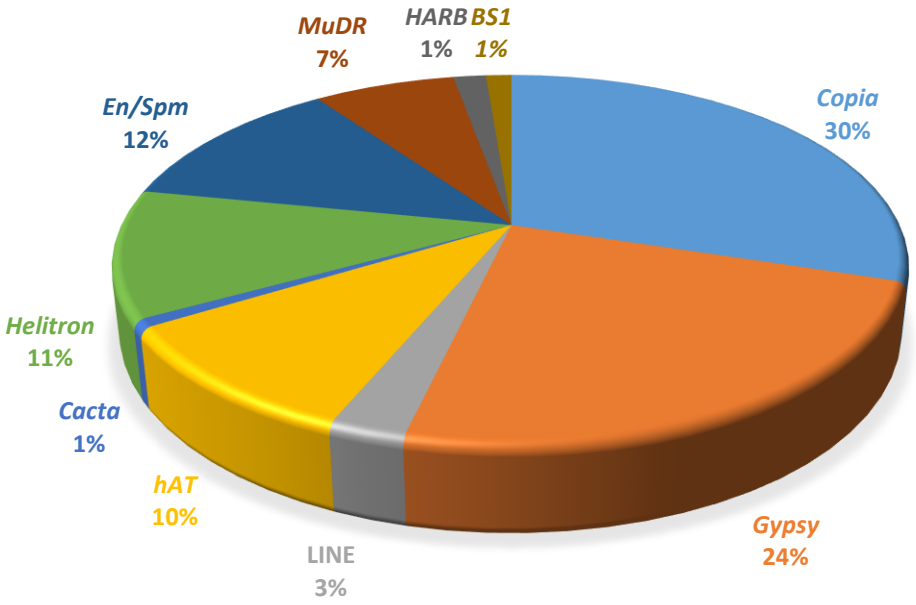
In case of the chloroplast genome, a single copy of MuDR-12_SBi of length 104 bp was found to have similarity with RNA polymerase gene (*rpoC1*). Interestingly, all insertion observed in the organellar genome by DNA transposons is in sense orientation. In the present investigation, *MuDR* elements were also reported in *Pongamia* unigenes. Previously, several studies have demonstrated the presence of transposable elements in the organellar genome (Losada et al., 2014). Generally, the prevalence of retrotransposons population is observed over DNA transposons in the mitochondrial genome (Hisano et al., 2016). It is interesting to note that no retrotransposons were observed in the coding region of mitochondria. TEs in the organellar genome often originated from the nuclear genome. To our knowledge, to date, no specific report of an association of TEs with organellar genes has been documented in the literature.

4.4.3. Transposable element-cassettes in the unigenes

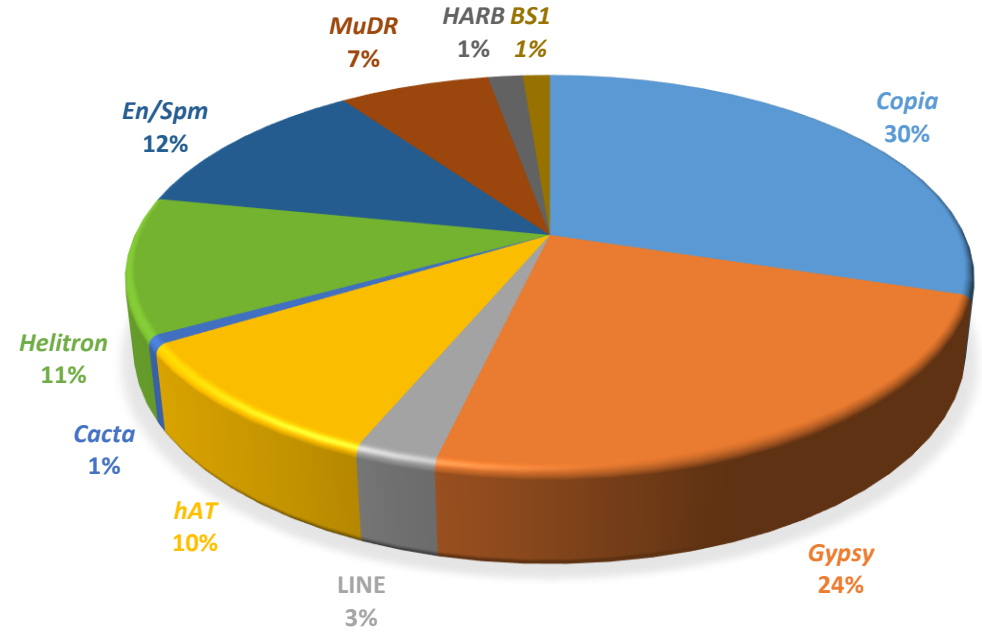
All the annotation results (RepeatMasker and BLASTX) from five libraries were merged and duplicates were removed. Finally, a total of 1290 protein-coding unigene sequences were found harbouring TE-cassettes across all libraries. During our analysis, the unigenes which showed similarity with TEs and not with proteins were discarded from libraries, as they were considered as putative transcriptionally active TE elements. This is the first instance in *Pongamia* where the mobile elements were detected in coding genes. Previously, Oki et al. (2008) proposed the role of transposable elements (MITEs) in the emergence of novel genes and the expansion of sequence diversity in rice genome. The occurrence of 1290 protein-coding unigenes with TE fragments indicate that the significant population of TE cassettes are present in unigenes of *Pongamia* when compared to 533 unigenes of *Homo sapiens*, 439 CDS of *O. sativa*, 140 unigenes in coffee (Nekrutenko and Li, 2001, Sakai et al., 2007, Lopes et al., 2008). The detail analyses of copy number, average RM score, the mean and maximum length of the TE sequences present in unigenes are mentioned in Table 4.1.

In the current investigation, the dominance of LTR-retrotransposons containing unigenes was observed in all libraries. LTR retrotransposons are an important and major component of a eukaryotic genome. Analysis showed the high prevalence of *copia* and *gypsy* element in protein-coding sequences of *Pongamia* (Fig. 4.1). LTR transposons were also reported to be an insignificant fraction in coffee ESTs (Lopes et al., 2008, Lopes et al., 2013). Among all TE cassettes, SINE element was the least present. A single copy of SINE was observed in seed and MpRf library. The absence of SINE elements was also reported in coffee (Lopes et al., 2008). Of the non-LTR TEs, LINE elements were present in ample amount. The mean length of the TE fragments varied from 34 bp (Non-LTR element) to 1466 bp (LTR element). In the previous report, the shortest length of TE fragment (31 bp) of L2 origin was reported in vertebrate protein; this observation corroborates with the present study (Lorenc and Makalowski, 2003). Earlier Makalowski et al. (1994) carried out a study on the mechanism of *Alu* element integration in human mRNA. Later, it was observed that the occurrence of a mobile element in protein-coding genes was not only confined to *Alu* elements but also to the other class of TEs (Makałowski, 1995). In the current investigation, we observed the presence of different kind of transposons in *Pongamia* unigenes (Fig. 4.1). Our results indicated the substantial

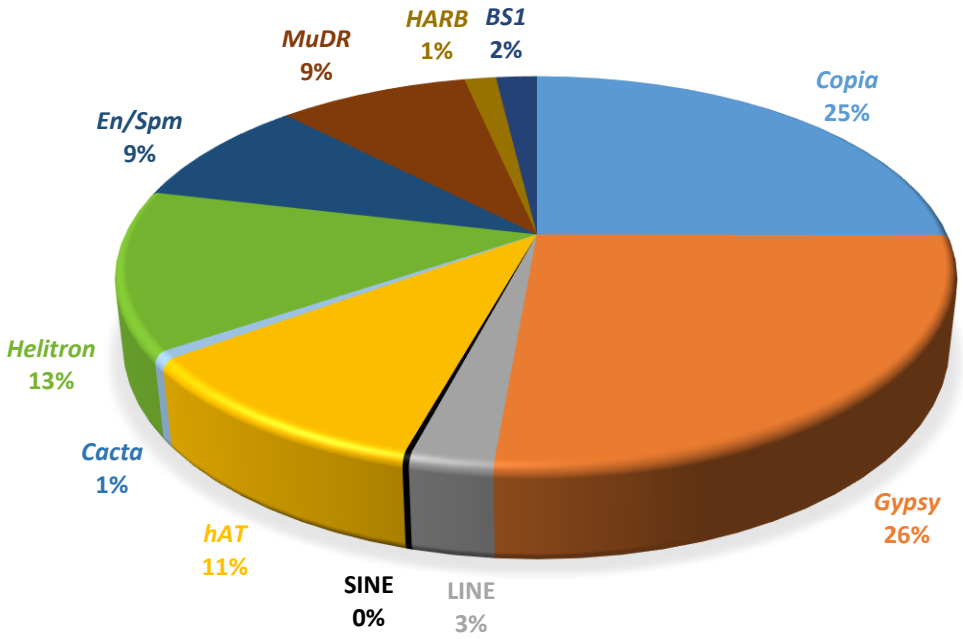
number of *Helitrons* DNA transposons in unigenes along with LTR-retrotransposons, which could also be responsible for plants protein diversity. *Helitrons* are suggested as a major driving force in gene evolution due to their ability to capture the different gene fragments and express them in chimeric form (Barbaglia et al., 2012). In addition, structurally *Helitrons* are highly polymorphic on either end (Du et al., 2009). Among DNA transposons, *hAT* (*hobo*, Ac and Tam3) and *En/Spm* (*Enhancer/Suppressor-mutator*) elements were also present in high amount. The *Ac/Ds* family made up of autonomous and non-autonomous group elements, which was earlier described by Barbara McClintock (1946) in maize as "controlling elements of the gene". According to Oki et al. (2008), *Alus* and MITES are a vital part of rice coding regions, which are absent in our investigation. The total population of retrotransposons in *Pongamia* unigenes varies between 50-57%, which are not that high as compared to DNA transposons. Vitte et al. (2014) proposed that out of the total TEs, class II elements are mainly inserted in genes or their close vicinity. In addition, Pogo superfamily (DNA transposons) has been described as a possible contributor in the domestication of genes in various eukaryotes. However, LTR-retrotransposons occupy a significant portion of the genome due to copy and paste mechanism of replication. Thus, class I elements have more opportunity to insert themselves into coding regions.



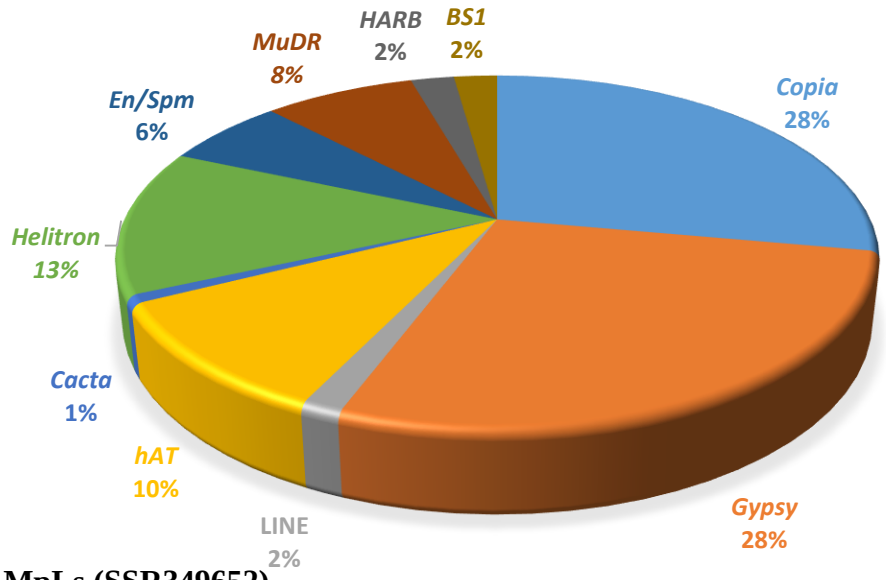
A) Seed (GEOZ00000000.1)



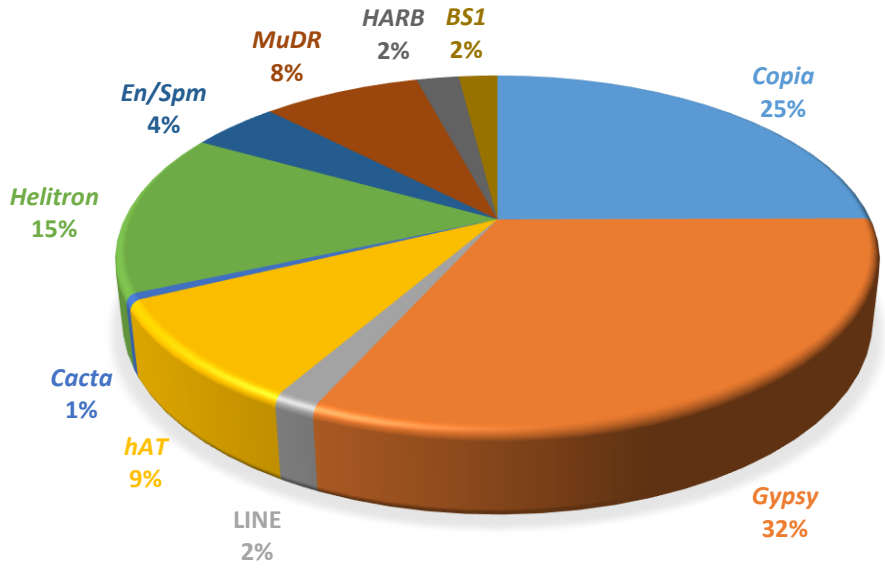
B) MpRs (SRR349650)



C) MpRf (SRR349651)



D) MpLs (SSR349652)



E) MpLf (SRR349653)

Figure 4.1. Pie chart showing the distribution of *Pongamia* unigenes containing TE cassettes in libraries A) Seed, B) MpRs, C) MpRf, D) MpLs, E) MpLf.

The mean highest RM score of 500.87 was observed for LTR retrotransposons and lowest 263.75 for a non-LTR element in library MpLs. The variation in frequency or number of TE cassettes present in unigenes were due to variations in a number of reads present across all libraries (Table 4.2). Some of the unigenes were found to harbour more than one TE cassette from the same or different family. The TEs were inserted in

unigenes at the same or different position with variable length. Hence, the variation in RM score was found for a different lineage of TEs of the same family inserted in unigenes. The variation in the number or frequency of TE cassettes in the protein-coding genes in existing and previous studies could be due to the following reasons: (1) Source and size of library from different tissue samples and conditions; (2) Number of sequences included in the study; (3) Selection of stringent criteria or different procedures; (4) The obtained results are species-specific; (5) Opting of different RepeatMasker database. Moreover, for annotation, we opted stringent criteria: i) 250 cut off RM score, ii) BLASTX E value and sequence similarity cutoff, which led to the removal of false positive results. However, opting of stringent parameters might have led to the loss of some important results. It is important to mention that all the above-reported analyses were done computationally and probably not represent the real occurrence of TEs to the proteome.

Table 4.1. Statistics of different TE cassettes present in *Pongamia* unigenes.

S.N	TE Classification	Number of TE-containing ESTs (%)	Average length of TE-cassettes (nt)	Minimum and maximum length of TE-cassettes (nt)	RM score average
1	**Seed (GEOZ00000000.1)**				
1.1	LTR	226 (47.08)	203.46	35-1424	458.88
1.2	Non-LTR	22 (4.8)	134.45	37-248	330
1.3	DNA	232 (48.33)	182.62	36-862	454
	Total	**480**			
2	**MpRs (SRR349650)**				
2.1	LTR	302 (53.64)	227.48	37-1466	482.56
2.2	Non-LTR	15 (0.017)	147.26	37-285	312
2.3	DNA	246 (43.69)	199.63	36-951	461.05
	Total	**563**			
3	**MpRf (SRR349651)**				
3.1	LTR	244 (51.36)	201.78	38-1425	463.95

3.2	Non-LTR	14 (2.94)	462.21	37-246	323
3.3	DNA	217(45.68)	191.62	36-1066	468.68
	Total	**475**			
4	**MpLs (SRR349652)**				
4.1	LTR	302 (55.71)	35-1466	228.32	500.87
4.2	Non-LTR	8 (1.47)	34-276	106.5	263.75
4.3	DNA	232 (42.80)	36-599	167.11	444.4
	Total	**542**			
5	**MpLf (SRR349653)**				
5.1	LTR	287 (56.71)	228.52	38-1443	491.63
5.2	Non-LTR	8 (1.58)	105.12	37-195	286
5.3	DNA	211 (41.69)	178.12	35-1247	462.67
	Total	**506**			

4.4.4. TE insertions orientation in relation to unigenes

TEs are inserted either in a 5' or 3' end orientation in the host genes sequence. The analysis was conducted to understand the pattern of insertion orientation of TEs in *Pongamia* unigenes. The study showed that the highest population of Ty1-*copia* superfamily was preferentially inserted in sense orientation in the three libraries viz. seed, MpRs and MpLs. On the other hand, the highest population of *copia* elements was inserted in antisense orientation in MpRf and MpLf, but very little difference was observed in a number of unigenes with sense and antisense insertion of LTRs and non-LTRs. According to Almeida et al., (2007), if the expected insertion frequencies of TEs are same for sense and antisense, then it would be considered as a random event. In addition, some studies observed that some population of TEs were preferentially inserted in host gene in the opposite orientation (Smit, 1999, Medstrand et al., 2002, Lorenc and Makalowski, 2003, Singer et al., 2004, Lopes et al., 2008).

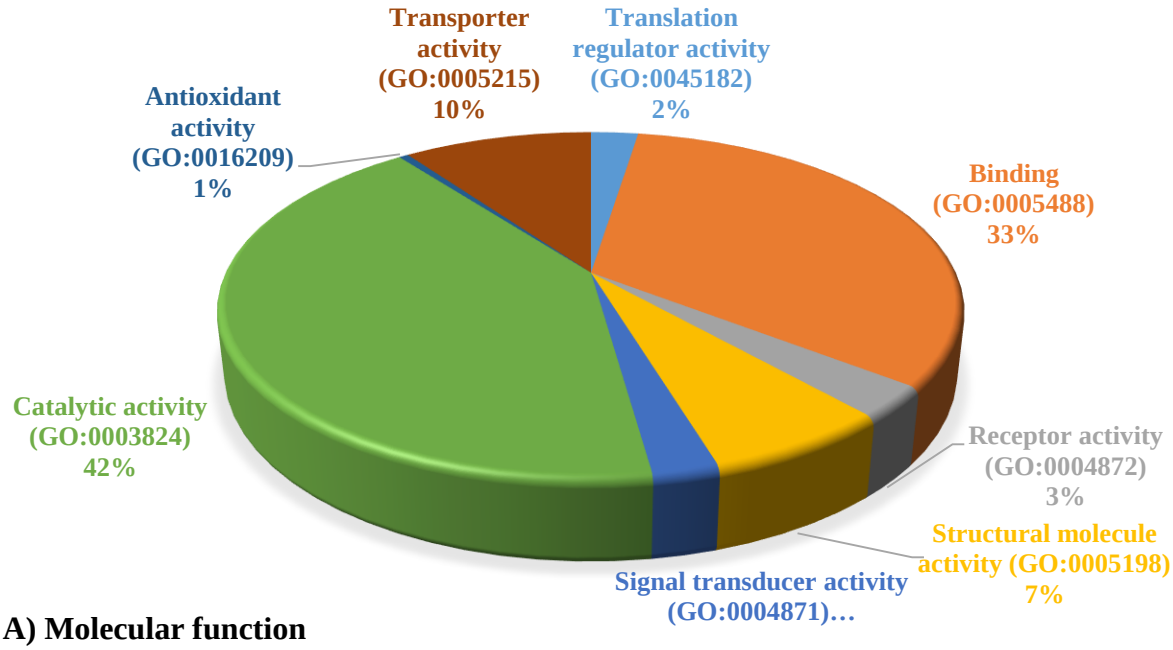
Table 4.2. Characterisation of the TE-cassettes according to their relative orientation to the host gene sequences.

S.N	TE Classification	Number of TE-containing unigenes	Frequency (%)	Sense *N* (%)	Antisense *N* (%)
1	**Seed (GEOZ00000000.1)**				
1.1	LTR	226	1.2	122	104
1.2	Non-LTR	22	0.046	14	8
1.3	DNA	232	0.69	124	108
	Total	**480**			
2	**MpRs (SRR349650)**				
2.1	LTR	302	1.2	165	137
2.2	Non-LTR	15	0.028	10	5
2.3	DNA	246	0.50	119	127
	Total	**563**			
3	**MpRf (SRR349651)**				
3.1	LTR	244	1.31	108	136
3.2	Non-LTR	14	0.03	8	6
3.3	DNA	217	0.54	120	97
	Total	**475**			
4	**MpLs (SRR349652)**				
4.1	LTR	302	1.17	165	137
4.2	Non-LTR	8	0.016	4	4
4.3	DNA	232	0.44	114	118
	Total	**542**			

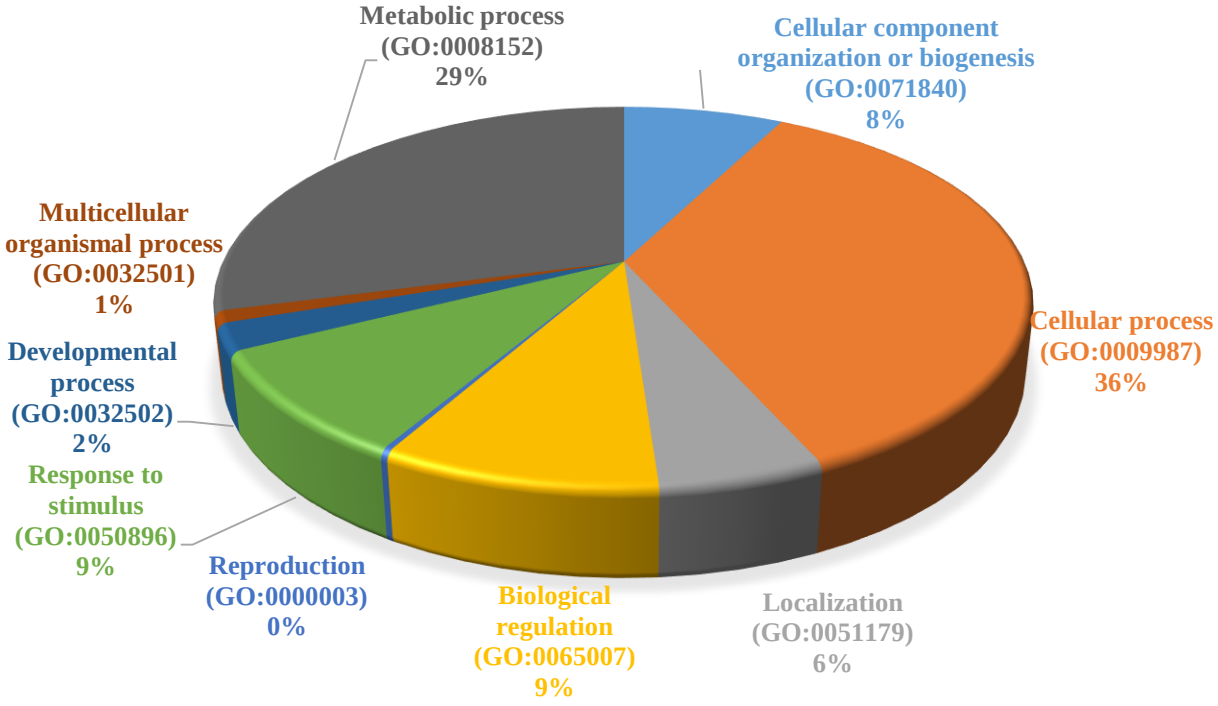
5	**MpLf (SRR349653)**				
5.1	LTR	287	0.95	139	148
5.2	Non-LTR	8	0.01	5	3
5.3	DNA	211	0.47	102	109
	Total	**506**			

4.4.5. Functional annotation of unigenes with TE cassettes

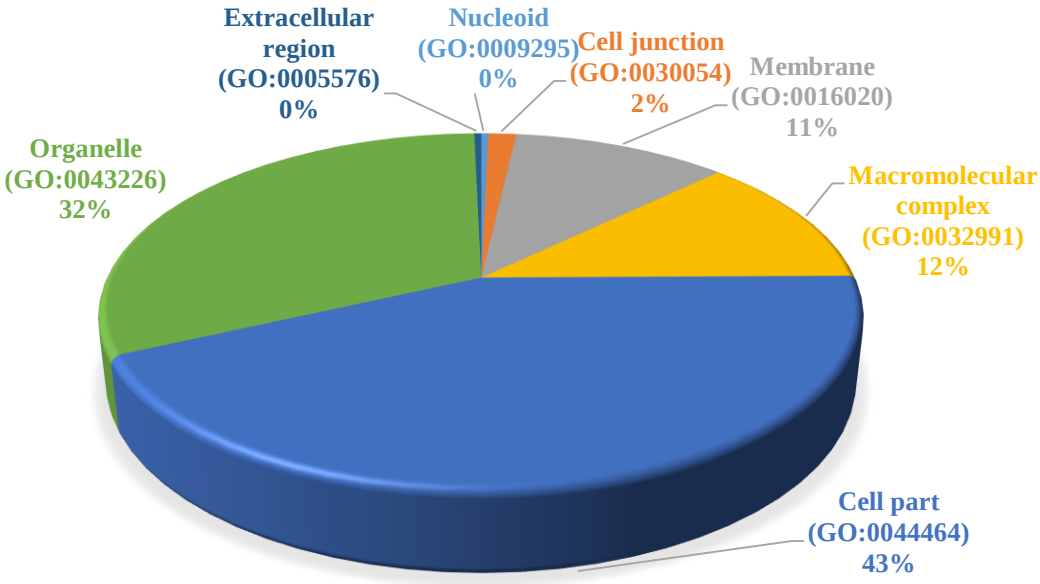
In order to assign the functional annotation, all unigenes having TE cassettes were examined using BLASTX, the details of which is mentioned in the previous section. All the annotation from five libraries were combined and further utilised for gene ontology. A total of 1290 unigene containing TE cassettes were identified as having significant similarity with known proteins for the above database. Similarly, Lopes et al. (2008) also carried out a functional annotation study for the validation of TE cassettes present in coffee genes. Based on annotation, unigenes were assigned to gene ontology (GO) terms using the Panther program. TE cassettes in unigene sequences were assigned to the molecular function, biological process and cellular component clusters respectively (Fig 4.2). *Pongamia* ESTs with TE cassettes were assigned to various pathways of metabolic process. In molecular function category, sequences related to the catalytic activity (GO: 0003824) were high in number followed by binding sequences (GO: 0005488) (Fig 4.2A). However, in the biological process section, cellular process (GO: 0009987) related sequences were highly abundant (Fig 4.2B). Furthermore, in the cellular component cluster, cell part (GO: 0044464) sequences were most abundant followed by organelle (GO: 0043226) component related sequences (Fig 4.2C). *Pongamia* is a tree which can grow on a wide range of agro-climatic conditions and show resistance to different categories of stress. It is said that transposable elements are more active in response to different biotic and abiotic stress for bringing the adaptable changes in plants. Hence, the study was focused on ESTs with particular relevance with protein diversity resulted from TEs insertion. The sequences were further categorised into protein class. Distribution of GO term in protein class revealed that the maximum sequences were associated with hydrolase (PC00121), transferase (PC00220) and least in chaperone (PC00072) (Fig 4.2D).



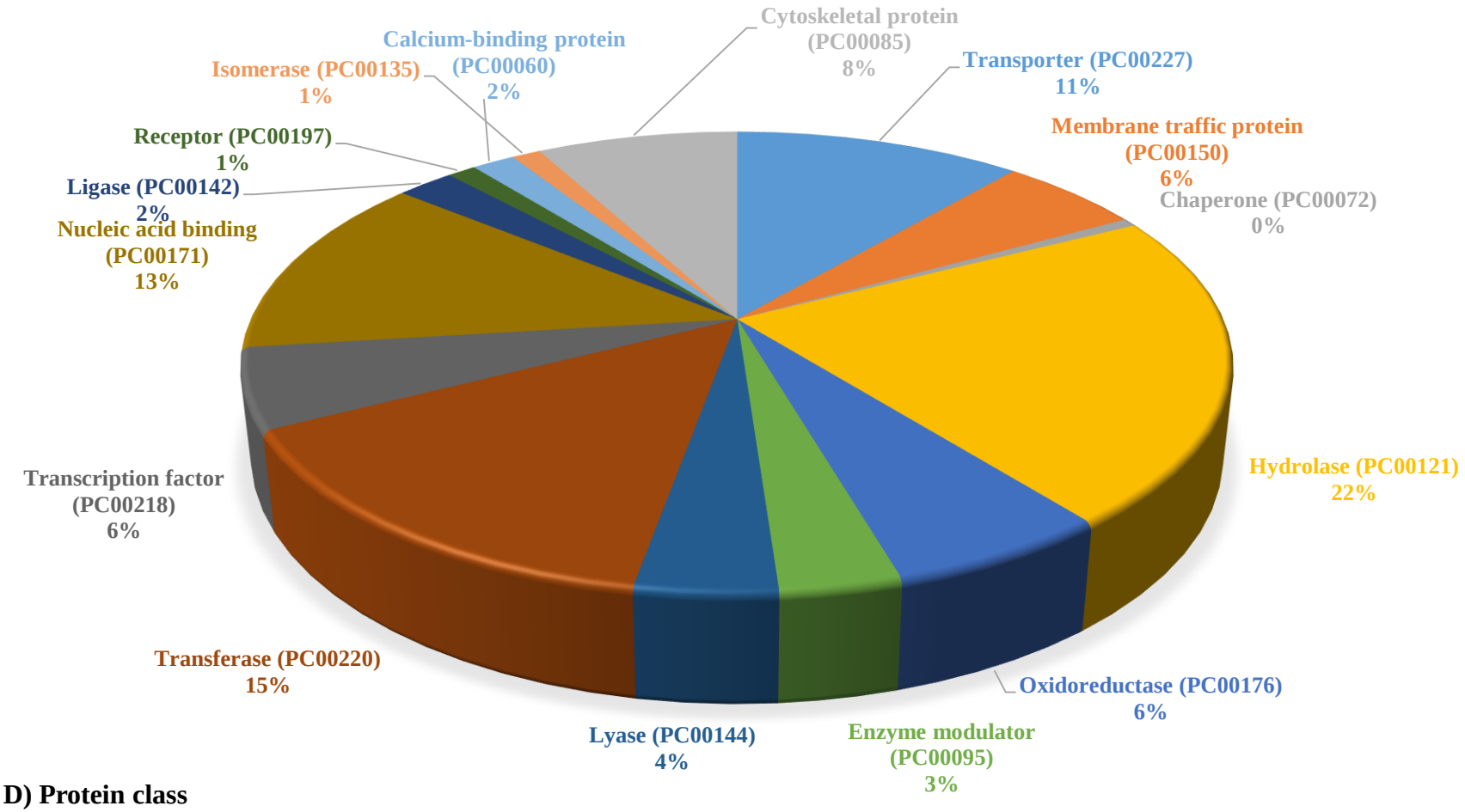
A) Molecular function



B) Biological process



C) Cellular component



D) Protein class

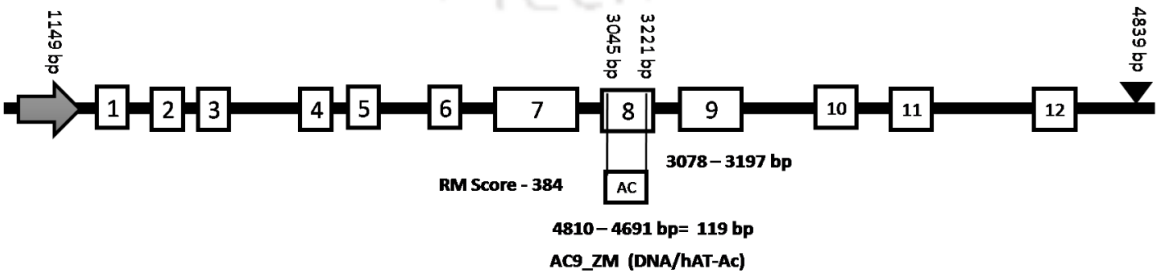
Figure 4.2. Details of GO terms (Pie chart) assigned to *Pongamia* unigenes containing TEs. These charts represent the distribution of GO classified as a molecular function (A), biological process (B), cellular component (C), protein class (D).

4.4.6. Exons origin from TE-cassettes

Three of the total 1270 identified proteins containing TE fragment in *Pongamia* unigenes were selected for the analysis. The opted unigenes for study were 1) *Granule-bound starch synthase* sequence id gb|GEOZ01007441.1| in seed library that matched to AC_9 ZM (*hAT* element) (Pohlman et al., 1984); 2) *Phytochelatin synthase* sequence id gb|GEOZ01022246.1| in seed library matched with *Gypsy*-184_ZM-I (Schnable et al., 2009); 3) *Pyruvate decarboxylase* sequence id gb|GEOZ01001322.1| in seed library that matched to *Gypsy*-184_ZM-I (Schnable et al., 2009).

Analysis showed the AC_9 ZM (*hAT* transposon) association with *Pongamia Granule-bound starch synthase* (GBSS) unigene. Likewise, the screening was carried out for TEs in *O. sativa* indica GBSS gene. The results showed the presence of same AC_9 ZM cassette of length 119 bp associated with exon number eight (Fig. 4.3A). To evaluate the presence of AC_9 ZM at the protein level, we analysed the protein sequence of GBSS-1 from *O. sativa* subsp. Japonica (sp|Q0DEV5|SSG1_ORYSJ) against Repbase. The small protein sequence of length 41 amino acid showed similarity with AC_9 ZM. This confirmed the translation of AC_9 ZM at the protein level. Another unigene, *Phytochelatin synthase* (PCS) of *Pongamia* showed the similarity with *Gypsy*-184_ZM-I retrotransposons. The same gene from the *Arabidopsis* was screened against RepeatMasker. Exon number four of length 133 bp showed association with *Gypsy*-184_ZM-I (Fig. 4.3B). Translation of *Gypsy*-184_ZM-I at protein level was confirmed using *A. thaliana Phytochelatin synthase* (PCS1_ARATH) protein retrieved from Swiss-Prot.

A) *O. sativa indica Granule-bound starch synthase* gene (GBSS) BGIOSGA024424



B) *A. thaliana Phytochelatin synthase* (PCS) AT5G44070

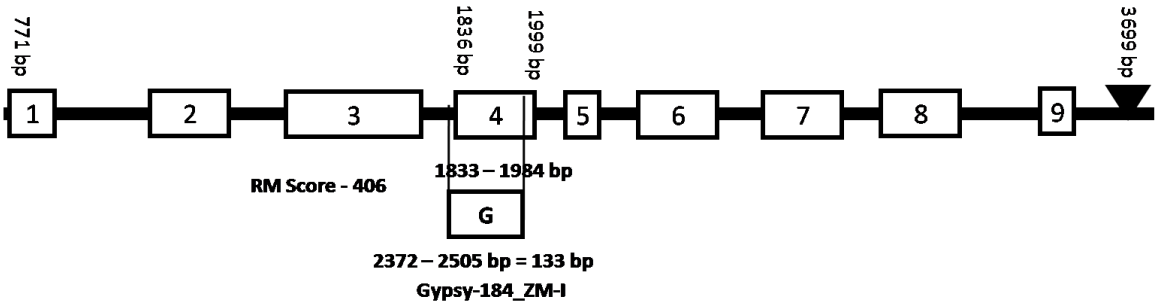
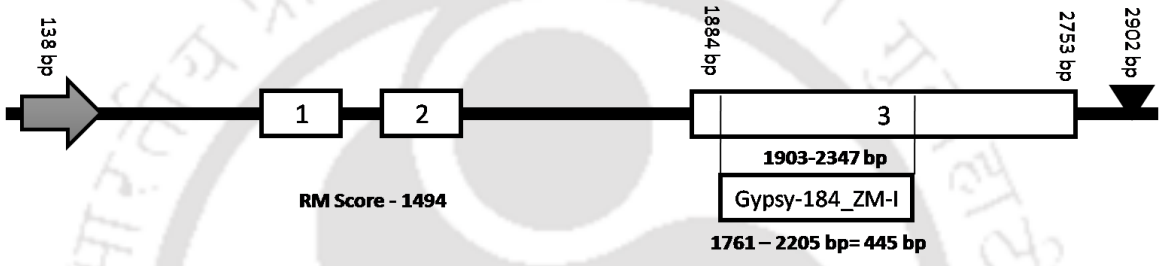


Figure 4.3. Diagrammatic representation of A) *Granule-bound starch synthase* gene, B) *Phytochelatin synthase* structure: arrow - transcription start site; boxes - exon; triangle – polyA tail; AC box belongs to AC_9 ZM, G box represents *Gypsy*-184_ZM-1. Gene structure shows the similarity of TE to an exon. The coordinates of the gene structure are mentioned along with the length.

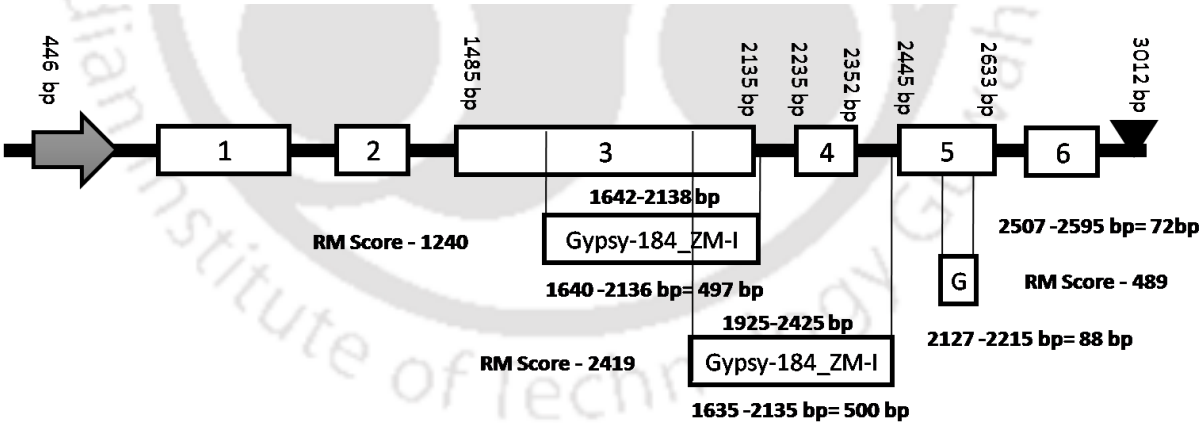
Out of three proteins, we opted for *PDC* for the study of exaptation event in *Pongamia.* The length of 565 bp region of *PDC* mRNA was found similar to *Gypsy*-184_ZM-I retrotransposon supported by high RM score 2057. The *PDC* genes containing TE fragments were investigated in details through the relationship between TE and host protein-coding genes with the help of model plants.

Besides plants, *PDC* gene from bacteria, fungi and bryophyte were also included for the annotation of TE cassettes and phylogenetic analysis. The respective genes were retrieved and screened against RepeatMasker using the same parameter which was used for screening *Pongamia* unigenes. Interestingly, of the total plant *PDC* genes analysed, three plant sequences did not show any TEs cassettes: *O. nivara* (ONIVA03G09800), *O. punctata* (OPUNC01G17650) and *O. sativa indica* (BGIOSGA039302). Similar results were also observed in bacterial, fungal and bryophyte *PDC* genes. Contrary to the above observations, the remaining plants showed the presence of TE cassettes in their *PDC* gene (Table 4.3). Primarily, one prominent TE fragment i.e. *Gypsy*-184_ZM-I (LTR-retrotransposons) of length around 400-500 bp was found associated with *PDC* genes. It was essential to understand the actual position on gene where the *gypsy* showed an association. FGENESH programme was used to determine the structure of the *PDC* gene. The relationship between *PDC* gene and TE cassettes have mentioned in detail in Fig. 4.4 and Table 4.3. Interestingly, all the events were detected with middle exons unlike in

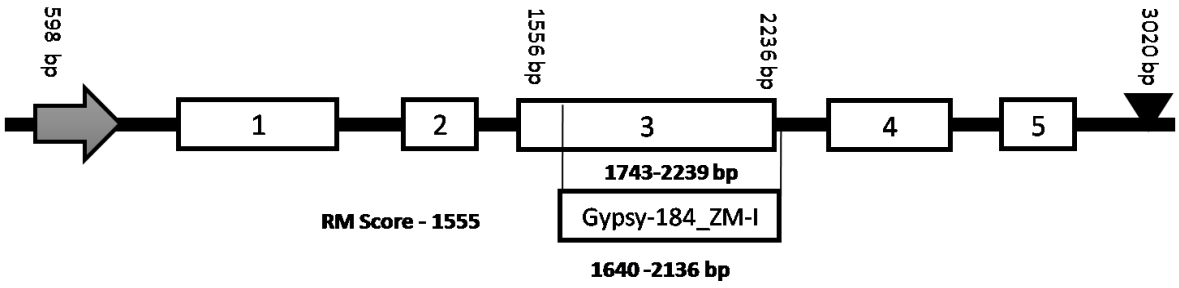
bovine genes where the TE association was found with last exon (Almeida et al., 2007). TE cassettes were either found associated with exon or with both exon and intron. There were other instances when multiple TE fragments were observed in genes: *O. sativa* indica *PDC1* (Gene id - BGIOSGA020021), *Solanum tuberosum PDC2* (Gene id - PGSC0003DMG400030369), *Z. mays PDC3* (Gene id - Zm00001d028759) (Fig 4.4 D, E, F). The association of TE with genes exists possibly due to either transposition of TE into exon (molecular domestication) or indirect insertion of an intronic TE (Lorenc and Makalowski, 2003, Lopes et al., 2008). From the analysis, it was confirmed that the *Gypsy*-184_ZM-I mobile element showed similarity with coding region i.e. exon sequence of the *PDC* gene.



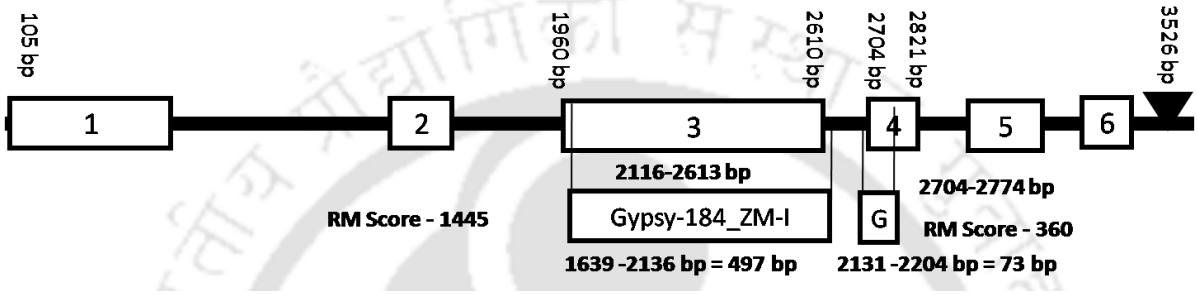
A) *Aegilops tauschii, PDC2* (Gene id - F775_10544)



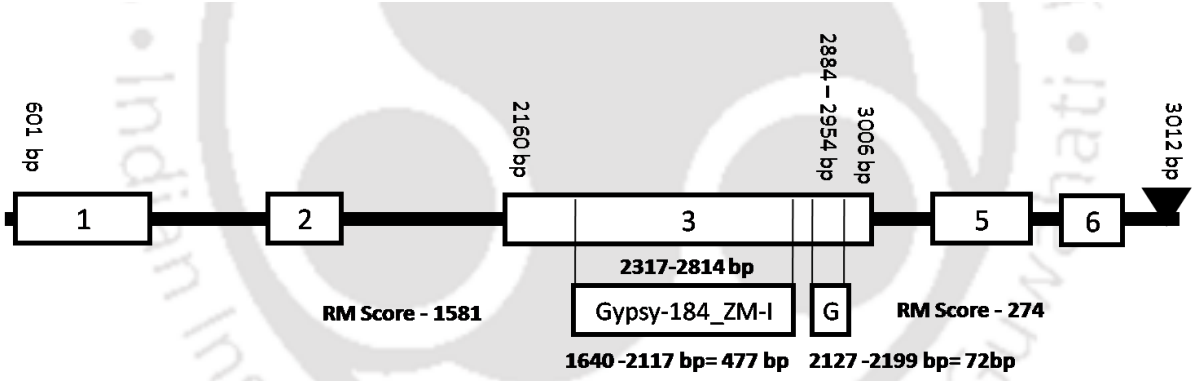
B) *A. thaliana, PDC1* (Gene id - AT4G33070)



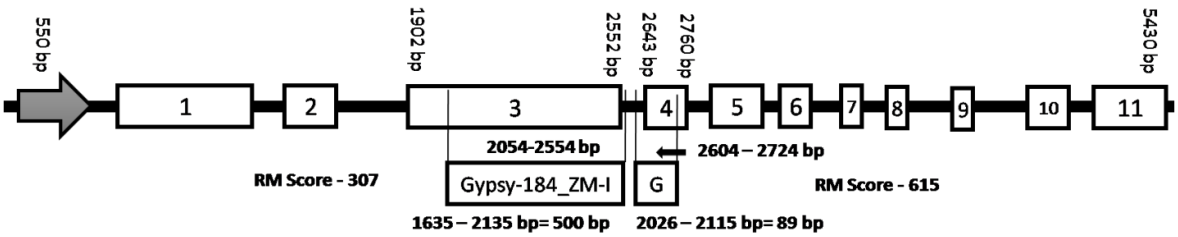
C) *Medicago truncatula, PDC* (Gene id - Medtr7g069540)



D) *O. sativa* Indica, *PDC1* (Gene id - BGIOSGA020021)



E) *S. tuberosum, PDC2* (Gene id - PGSC0003DMG400030369)



F) *Z. mays, PDC3* (Gene id - Zm00001d028759)

Figure 4.4. Diagrammatic representation of *pyruvate decarboxylase* gene structure: arrow - transcription start site; boxes - exon; triangle – polyA tail; G box represents *Gypsy*-184_ZM-1. Gene structure shows the similarity of TE to an exon. The coordinates of the gene structure are mentioned along with the length.

To confirm the structure of *Gypsy*-184_ZM-I element, maize B73 genome sequences (Accession number: - CM000777 to CM000786) were downloaded from the NCBI database. B73 genome sequences were screened against LTR finder with default parameter against maize genome LTR-retrotransposon. After analysis, we found around 3463 bp length of complete LTR retrotransposon. The two characteristic identical LTR region of length 205 bp were observed at either end of a retrotransposon. At the end of LTR region, tandem site duplication sequences were present, which are the characteristic feature of retrotransposons. The primer binding site (PBS) of length 15 bp is depicted in the yellow circle in the figure 4.5. This region plays an essential role in primer binding during transcription of retrotransposons. From the analysis, it is reconfirmed that the sequence which showed similarity with the *PDC* gene belongs to Ty3-*gypsy* family (Fig. 4.5).

Table 4.3. Details of *PDC* gene containing TE cassettes used for the construction of the phylogenetic tree.

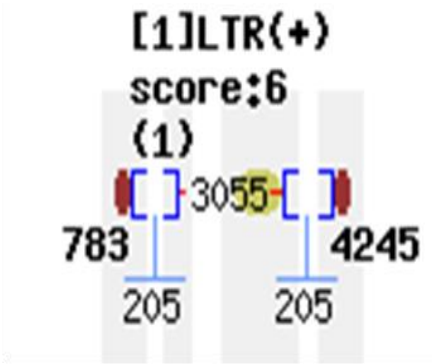
S.N	Plant name/Id	RM score	Gene start	Gene end	Length	Position	TEs type	TEs start	TEs end
1.1.	*Zea mays PDC3* Zm00001d028759	307	2054	2554	500	+	*Gypsy*-184_ZM-I	1635	2135
1.2.	*Zea mays PDC3* Zm00001d028759	615	2604	2724	120	+	*Gypsy*-184_ZM-I	2086	2215
2.1.	*Nicotiana attenuata PDC1* A4A49_28319	1132	2169	2573	404	+	*Gypsy*-184_ZM-I	1640	2044
2.1.	*Nicotiana attenuata PDC1* A4A49_28319	386	2991	3086	95	+	*Gypsy*-184_ZM-I	2139	2239
3.1.	*Arabidopsis thaliana PDC* AT4G33070	1555	1743	2239	496	+	*Gypsy*-184_ZM-I	1640	2136
4.1.	*Oryza sativa* Indica *PDC1* BGIOSGA020021	1240	1642	2138	496	+	*Gypsy*-184_ZM-I	1640	2136
4.2.	*Oryza sativa* Indica *PDC1* BGIOSGA020021	2419	1925	2425	500	+	*Gypsy*-184_ZM-I	1635	2135
4.3.	*Oryza sativa* Indica *PDC1* BGIOSGA020021	489	2507	2595	88	+	*Gypsy*-184_ZM-I	2127	2215

5.1.	*Zea mays PDC2* Zm00001d008651	305	199	255	56	+	*Helitron*-3N1_ZM	1550	1603
5.2.	*Zea mays PDC2* Zm00001d008651	1517	2282	2779	497	+	*Gypsy*-184_ZM-I	1639	2136
5.3.	*Zea mays PDC2* Zm00001d008651	347	3110	3211	101	+	*Gypsy*-184_ZM-I	2133	2239
5.4.	*Zea mays PDC2* Zm00001d008651	493	4122	4192	70	+	*En/Spm*-10_ZM	8289	8351
5.5.	*Zea mays PDC2* Zm00001d008651	3440	4193	4637	444	+	HUCK1-LTR_ZM	1	446
6.1.	*Brassica rapa PDC* Bra028896	1697	1622	2118	496	+	*Gypsy*-184_ZM-I	1640	2136
7.1.	*Aegilops tauschii PDC2* F775_10544	1494	1903	2347	444	+	*Gypsy*-184_ZM-I	1761	2205
8.1.	*Musa acuminata PDC1* GSMUA_AchrUn_randomG01470_001	1210	1644	2139	495	+	*Gypsy*-184_ZM-I	1640	2135
8.2.	*Musa acuminata PDC1* GSMUA_AchrUn_randomG01470_001	329	2218	2303	85	+	*Gypsy*-184_ZM-I	2130	2215

8.3.	*Musa acuminata PDC1* GSMUA_AchrUn_randomG01470_001	835	3176	3574	398	+	*Gypsy*-184_ZM-I	1646	2044
8.4.	*Musa acuminata PDC1* GSMUA_AchrUn_randomG01470_001	386	3948	4023	75	+	*Gypsy*-184_ZM-I	2133	2208
9.1.	*Cucumis sativus PDC* Csa_6G518930	748	1943	2347	404	+	*Gypsy*-184_ZM-I	1640	2044
9.2.	*Cucumis sativus PDC* Csa_6G518930	292	2442	2533	91	+	*Gypsy*-184_ZM-I	2042	2133
9.3.	*Cucumis sativus PDC* Csa_6G518930	343	2746	2826	80	+	*Gypsy*-184_ZM-I	2135	2215
10.1.	*Solanum tuberosum PDC2* PGSC0003DMG400030369	1581	2317	2814	497	+	*Gypsy*-184_ZM-I	1640	2137
10.2.	*Solanum tuberosum PDC2* PGSC0003DMG400030369	274	2882	2954	72	+	*Gypsy*-184_ZM-I	2127	2199
11.1.	*Musa acuminata PDC2* GSMUA_Achr5G07390_001	1787	1673	2169	496	+	*Gypsy*-184_ZM-I	1640	2136
11.2.	*Musa acuminata PDC2* GSMUA_Achr5G07390_001	297	2248	2330	82	+	*Gypsy*-184_ZM-I	2133	2215

12.1.	*Medicago truncatula* Medtr7g069500	1197	1590	2084	494	+	*Gypsy*-184_ZM-I	1639	2136
12.2.	*Medicago truncatula* Medtr7g069500	332	2165	2236	71	+	*Gypsy*-184_ZM-I	2133	2204
13.1.	*Medicago truncatula* Medtr2g009330	1445	2116	2613	497	+	*Gypsy*-184_ZM-I	1639	2136
13.2.	*Medicago truncatula* Medtr2g009330	360	2701	2774	73	+	*Gypsy*-184_ZM-I	2131	2204
14.1.	*Vigna radiata* Vigan.04G133000.01	1605	1717	2214	497	+	*Gypsy*-184_ZM-I	1639	2136
14.2.	*Vigna radiata* Vigan.04G133000.01	363	2297	2368	71	+	*Gypsy*-184_ZM-I	2133	2204
15.1	*Vigna radiata* Vigan.09G086100.01	1546	2568	3062	494	+	*Gypsy*-184_ZM-I	1639	2133
15.2.	*Vigna radiata* Vigan.09G086100.01	380	3161	3224	63	+	*Gypsy*-184_ZM-I	2141	2204
16.1.	*Populus trichocarpa* POPTR_0016s12760	1479	1976	2458	482	+	*Gypsy*-184_ZM-I	1639	2121
16.2.	*Populus trichocarpa* POPTR_0016s12760	328	2554	2631	77	+	*Gypsy*-184_ZM-I	2129	2206
17.1.	*Populus trichocarpa* POPTR_0006s10340	1357	1656	2152	496	+	*Gypsy*-184_ZM-I	1640	2136
17.2.	*Populus trichocarpa* POPTR_0006s10340	349	2233	2310	77	+	*Gypsy*-184_ZM-I	2129	2206

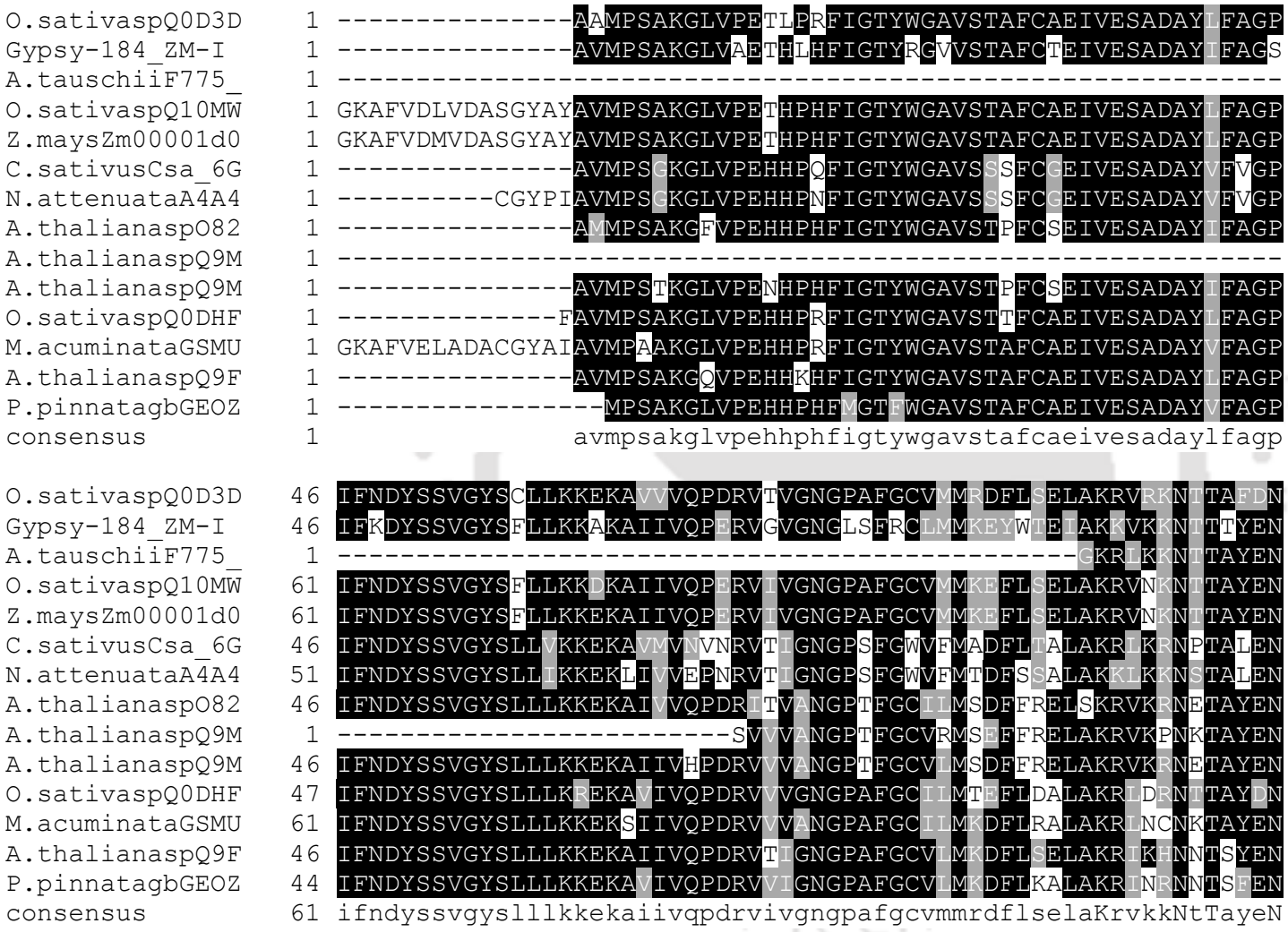
```
Program    : LTR_FINDER
Version    : 1.05

Load tRNA db [tRNAdb/zeaMay4-tRNAs.fa]  0.004 second
Predict protein Domains 0.005 second
>Sequence: Zea Len:12110
[1] Zea Len:12110
Location : 783 - 4245 Len: 3463 Strand:+
Score    : 6 [LTR region similarity:1]
Status   : 11111010000
5'-LTR   : 783 - 987 Len: 205
3'-LTR   : 4041 - 4245 Len: 205
5'-TG    : TG , TG
3'-CA    : CA , CA
TSR      : 778 - 782 , 4246 - 4250 [GACCA]
Sharpness: 0.5,0.5
Strand + :
PPT   : [12/15] 4026 - 4040
```




```
Details of exact match pairs:
4040-4245[206]
782-987[206]

Details of the LTR alignment(5'-end):
                                    |4041
-------ATTTCA---AGG-----A-GTGG-GGA--ATGTTAGATATATTAGGTGTAGTAGATGAGCTATGGGTTATGTGG
       | | ||   |||     | | || |||  |||||*|||||||||||||||||||||||||||||||||||||||
CCCCTCCACTCCACGCAGGGCGGCAGGCGGTGGACCATGTTAGATATATTAGGTGTAGTAGATGAGCTATGGGTTATGTGG
                                *****|783

Details of the LTR alignment(3'-end):
                                  4245|----*-****-
ATCTCTGATCTCCTTTCTTCTACCTTCTGTGGTGCTAACA----G-ACCA-ACTGGCGGAA-CGA--GA-ACGG-GCCGAG
|||||||||||||||||||||||||||||||||||||||*    | | || |  || |||| | |  || || | || |
ATCTCTGATCTCCTTTCTTCTACCTTCTGTGGTGCTAACATTTGGTATCAGA--GGAGGAATCCATGGATACTGTGCTGCA
                                  987|

Details of PPT(+):
TCAAGGAGTGGGGAA
|4026
```

Figure 4.5. Details of *Gypsy*-184_ZM-I retrotransposon structure and statistics generated using LTR Finder tool. Brown circle represents the target site repeats; the yellow circle represents the polyprotein tract region, box present at both ends depicts the LTR regions.

Multiple sequence alignment was carried out between *Gypsy*-184_ZM-I and protein sequences of PDC exon 3 and 4 of *Z. mays* (Zm00001d028759), *M. acuminata* (GSMUA_AchrUn_randomG01470_001), *C. sativus* (Csa_6G518930), *N. attenuata* (A4A49_28319), *A. tauschii* (F775_10544) and *Pongamia* unigene. The highest similarity was observed between *Gypsy*-184_ZM-I with *Z. mays* (81.25%) (Fig. 4.6). The contribution of TEs to the protein-coding region was the prime interest of this investigation, as their insertion brings diversity in protein sequences which leads to phenotypic changes. There were several reports available where TEs contributed in a normal function of genes in organisms. Almeida et al. (2007) reported the presence of TEs in six genes of bovine and their translation in protein. The insertion phenomenon of TEs are well reported on gene and transcript level, but very few descriptions available at the protein level (Almeida et al., 2007, Britten, 2006, Makałowski et al., 2017). However, the existence of TEs sequences on transcript level does not necessarily promise their translation to protein. This could happen due to i) possible deleterious effect of TEs insertion on encoded proteins (Nuzhdin, 1999); ii) Disruption of the cellular process through chromosome nicking by TE fragment containing proteins (Nuzhdin, 1999). Hence, there are several mechanisms by which TEs can be eliminated before translation (Gotea and Makalowski, 2006). These facts encouraged us to investigate the presence of TEs at the protein level, so the PDC protein sequences from the Swiss-Prot database were included for further study. We downloaded the different isoforms of PDC from UniProt : *A. thaliana* (sp|O82647|PDC1), *O. sativa* (sp|Q0D3D2|PDC3), *O. sativa* (sp|Q0DHF6|PDC1), *O. sativa* (sp|Q10MW3|PDC2), *A. thaliana* (sp|Q9FFT4|PDC2), *A. thaliana* (sp|Q9M039|PDC3), *A. thaliana* (sp|Q9M040|PDC4) for TE annotation using Repbase programme. From the analysis, it was confirmed that the insertion of the *Gypsy*-184_ZM-I element was also present at the protein level (Fig 4.6).



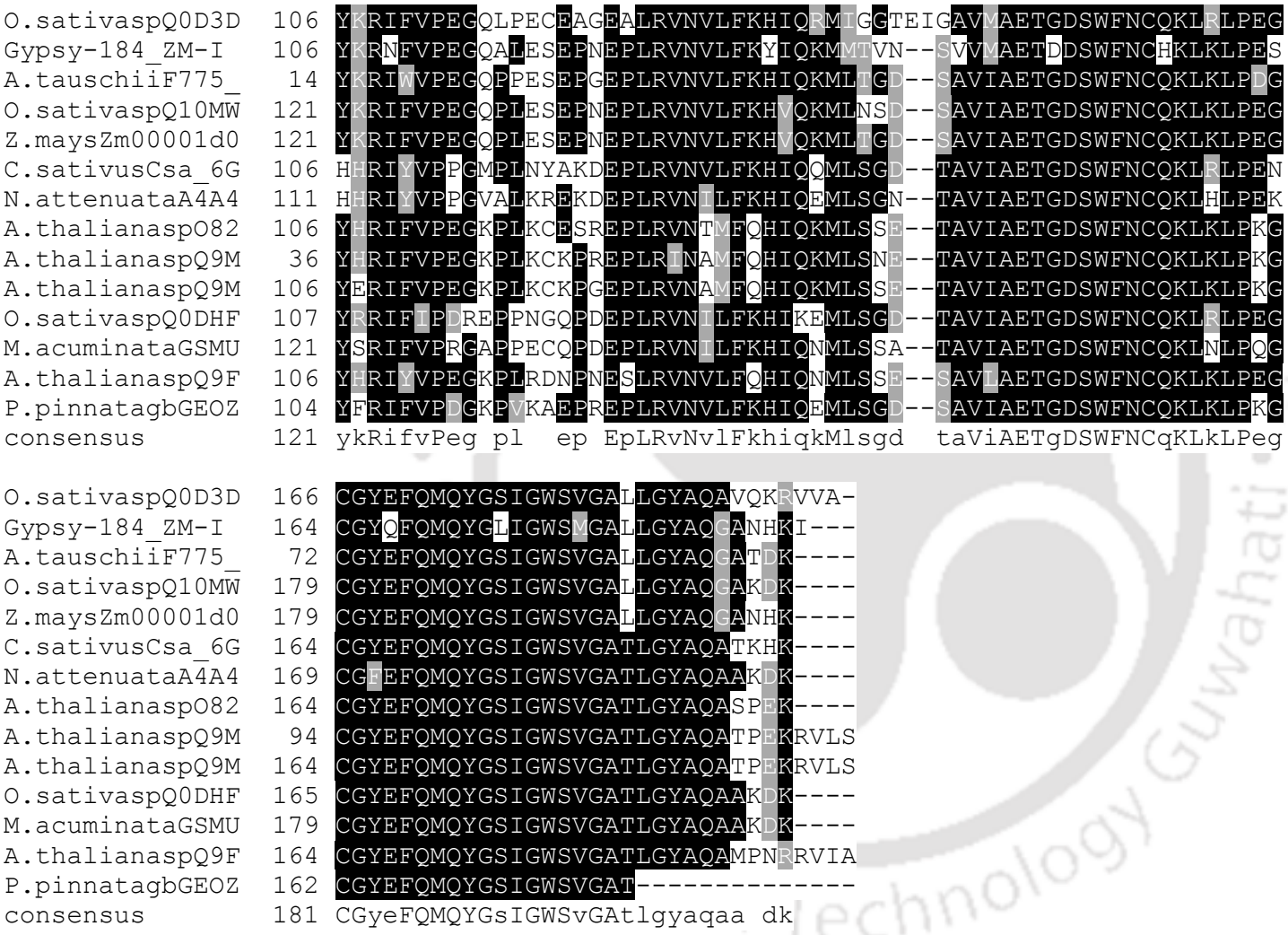


Figure 4.6. Alignment of the partial PDC protein with a *Gypsy*-184_ZM-I protein sequence. Highly conserved and similar amino acids are highlighted with black and grey colour. The names of plant PDC proteins are mentioned before every sequence.

Further, the insertion of the *gypsy* element was also proved using a phylogenetic tree. For the phylogenetic tree construction, different isoforms of PDC translated protein sequences from bacteria, fungi, bryophyte, monocot and dicot plants were analysed (Fig 4.7). The *PDC* sequence with *gypsy* insertion clustered separately, whereas the *PDC* with no *gypsy* insertion clustered at the end of the phylogenetic tree. *O. punctata* (OPUNC01G17650), *O. indica* sativa (BGIOSGA038476) and *O. nivara* (ONIVA03G09800) thus clustered with bacterial and fungal *PDC* genes. From the phylogenetic tree, it was cleared that *PDC* with *gypsy* insertion showed some sequence variation than the *PDC* sequence with no *gypsy* insertion. Interestingly, only three genes from the plants were found with no insertion of the *gypsy* element. It means plant does contain some original copies of *PDC* where no *gypsy* insertion are present.

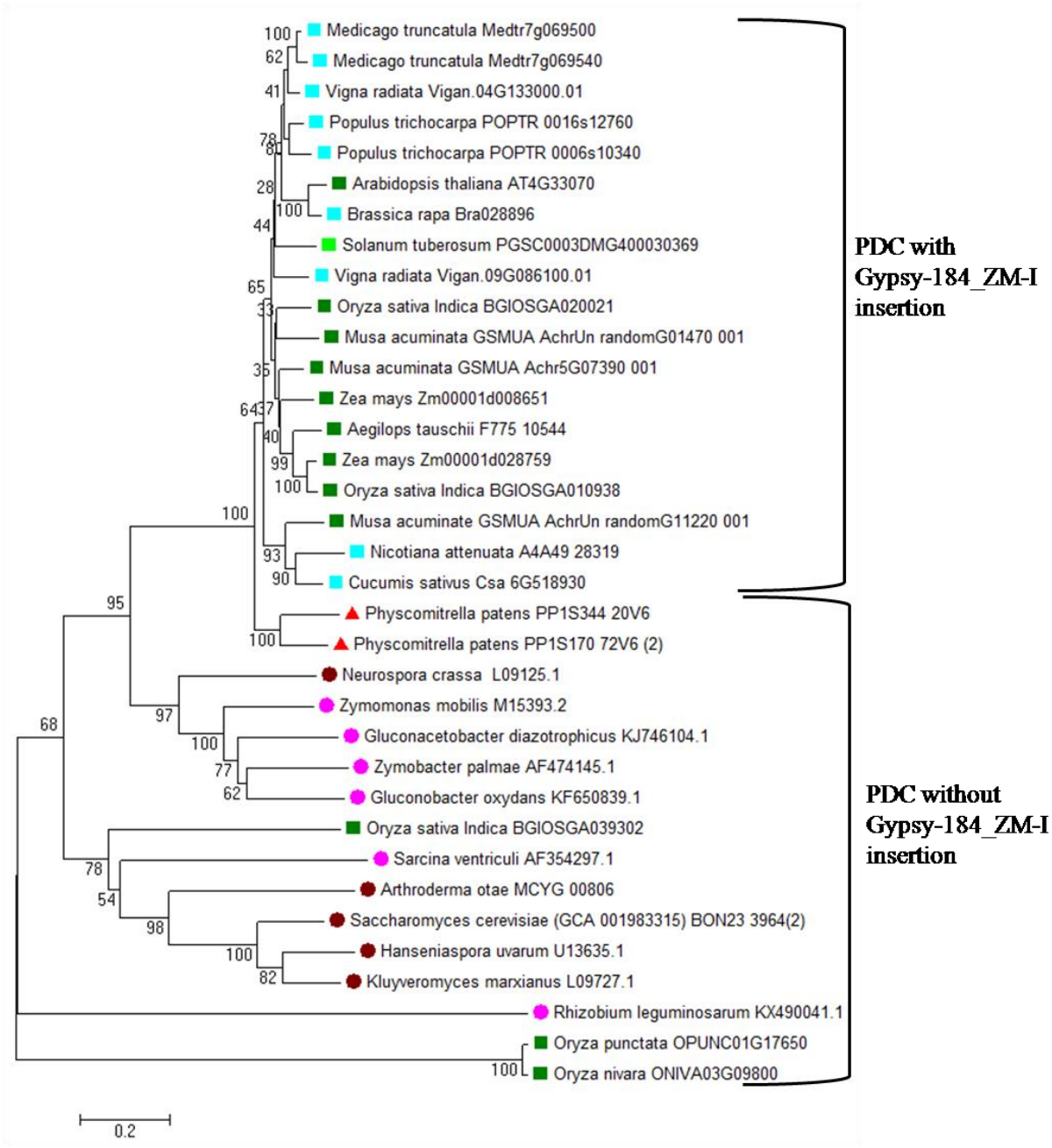


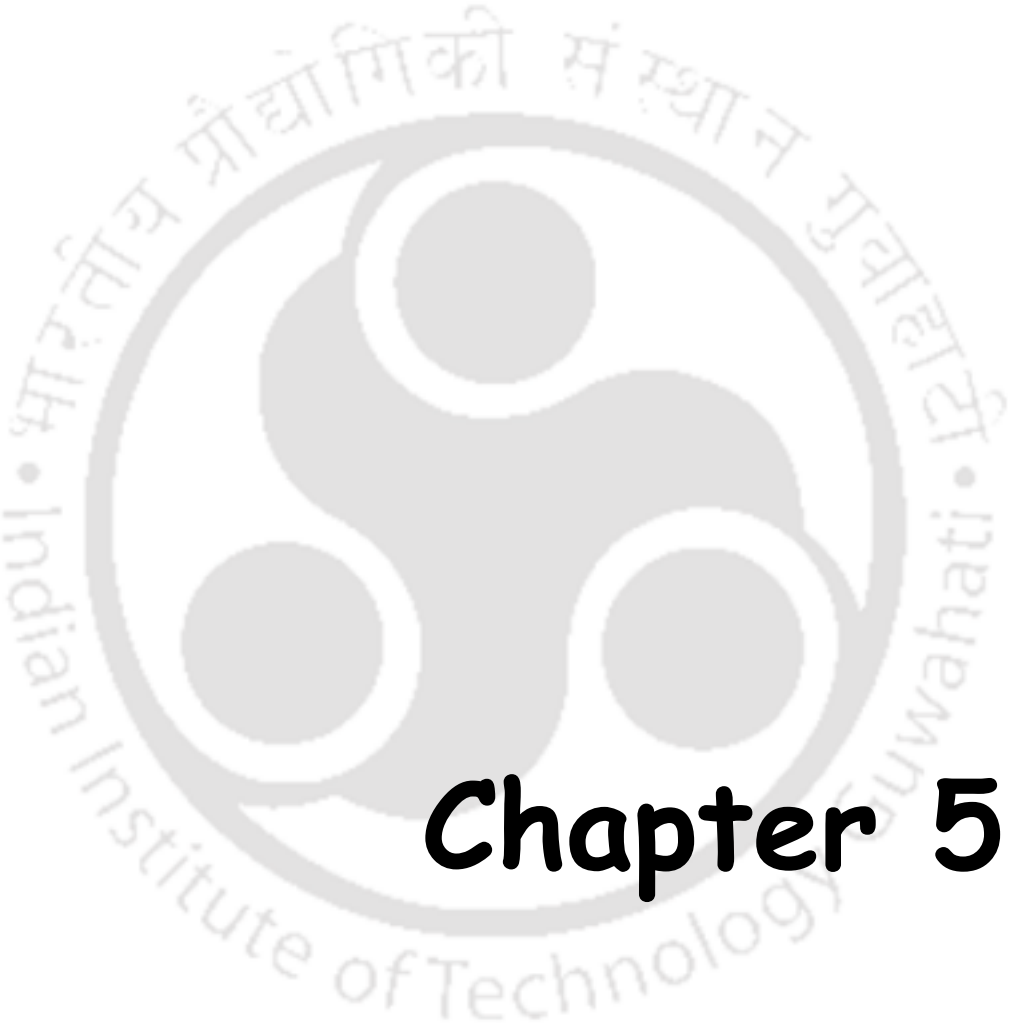
Figure 4.7. Phylogenetic tree of PDC protein sequences generated using the MEGA 6 program with 1000 bootstrap replicates. Latin descriptions are mentioned with their respective gene identification number corresponding to different databases. Green box represents monocot plant, blue box represents dicot plant, red triangle belongs to bryophyte, brown circle shows the fungal population, and pink circle represents the bacterial population.

The occurrence of TE fragment within protein indicates the expatation event. PDC is a homotetrameric enzyme primarily catalysing decarboxylation of pyruvic acid in cytoplasm and mitochondria. The enzyme consists of an identical subunit of alternating α-helices and β-sheets. Two domains exist in each 60 kDa subunit. We tried to explain the three arguments in view to support the exaptation phenomenon. First is explained by the fact that PDC is encoded by a multigene family which diverged very early during evolution (Hossain et al., 1996, Talarico et al., 2001). Around six types of *PDC* genes were observed in NCBI and Swiss-Prot database. In accordance with Ohno (1970), the occurrence of a duplicate copy of a gene would provide the new opportunities by allowing a copy of the duplicated gene to evolve for new functional properties. The other copy is conserved, resulting in alterations in regulation that does not cause loss of pre-existing specificities or functions (Robins and Samuelson, 1992). This could be the probable reason why multigene families tolerate the altered gene regulations through transposon-induced mutation (Robins and Samuelson, 1992). In addition, new duplicated copies are free from functional constraint and probably sustain significant changes until they acquire a new function. This phenomenon is well explained in the present study through the phylogenetic analysis of *PDC*, where the *PDC* with TEs and without TEs insertion separated into different groups. Results show that the insertion of the *gypsy* element might have happened before the divergence of dicots and monocot. Second is based on high similarity that existed between a fragment of *gypsy* and *PDC* gene exon sequence which was supported by multiple sequence alignment. *PDC* plays a diverse role in accordance with the environmental condition, tissue and organisms. In plants, besides a role in energy and metabolism, *PDC* also have a role in development activity, biotic and abiotic stresses such cold, salt, submergence, wounding and pathogen infection (Tadege et al., 1999, Ismond et al., 2003, Kursteiner et al., 2003, Mithran et al., 2014). Moreover, van de Lagemaat et al. (2003) observed the more likely existence of TE in genes with functions such as stress response, defence and response against external stimuli compared to other genes. TEs are activated due to different biotic and abiotic stresses like cold, salt, heat stress and responsible for TE-related adaptive mutations (Grandbastien et al., 2005, Naito et al., 2009, Cavrak et al., 2014, Ito et al., 2016). This stress inducible activation may allow them to insert their stress responsive elements to genes; resulting genome fluidity that confers stress related adaptation (Negi et al., 2016). This could be the probable reason for the presence of TEs in the *PDC* gene to bring functional diversity. The third argument is based on the fact that the bacteria do not contain Ty3-*gypsy* elements in their genome, this could be the reason

behind the absence of TEs in bacterial *PDC*. In case of fungus, they do have retrotransposons population. Still, the TE insertion is absent. This could be explained by the fact that the population of retrotransposons is much higher in plants compared to other organisms. Hence, plants have more chances of TEs insertion in their genic region for the adaption against environmental stress. This finding shows the insertion of TEs in the genic region of *Pongamia* and their contribution to host gene expression.

4.5. Conclusions

Transposable elements are well known for their mutagenic and parasitic activity. TE mobility could contribute to genome expansion and donation of regulatory sequences, which leads to evolutionary changes. In the current investigation, we tried to understand the potential of TEs in the evolution of *Pongamia* unigenes. As expected, the population of LTR retrotransposons were found to be high in unigenes, as they occupy a significant portion of the genome in eukaryotes. In this study, we have shown the contribution of the *gypsy* element in PDC protein diversity. From the findings, it was understood that the insertion of the *gypsy* retrotransposon was only present in higher plants and absent in bacterial, fungal and bryophyte genes. It is possible that the insertion of *gypsy* element in *PDC* might have happened before the divergence of monocot and dicot plants. Interestingly, some copies of *PDC* genes from *Oryza* were devoid of *gypsy* insertion. It shows that the original copies of *PDC* genes are still maintained in the plants. Their role varies from metabolic to stress-related activity in various tissue and environmental conditions. The duplicated and stress-related genes are more prone to transposable attack within and outside close vicinity to genes. The insertion of TEs was observed in middle exonic region of *PDC*, which sometimes occupy the whole exon and sometimes extends outside exon in the intron. The presence of *gypsy* insertion at exon and intron boundaries shows the possible role of TEs in alternative splicing. Even a small level of alternative splicing could bring functional diversity in protein. TEs can make considerable changes in coding sequences as compared to the slow process of evolution by nucleotide substitution. Moreover, during the investigation, we found TEs in protein coding regions of the organellar genome. To understand more about TEs and genes relationship, extensive studies and experiments are required to further validate this phenomenon.



Chapter 5

Chapter 5

Development of EST-SSR marker in *Pongamia*

5.1. Introduction

The Leguminosae family is one of the economically important, diverse and numerous clade of the flowering plants, including pigeon pea (*Cajanus cajan*), mungbean (*Vigna radiata*), groundnut (*Arachis hypogaea*), chickpea (*Cicer arietinum*), soybean (*Glycine max*), and non-edible oil yielding crop, karanj (*Pongamia pinnata*). *Pongamia* is a medium-sized fast-growing tree native to India, Malaysia, Northern Australia and Indonesia. It is an outcrossing trees species which starts fruiting after the fourth year. Generally considered tolerant to different biotic and abiotic stress, a plant is able to grow in a vast range of agro-climatic conditions, making it an attractive biofuel crop. *Pongamia* has emerged as one of the important source of renewable energy due to the availability of oil-rich seeds feedstock in abundance. Current energy crisis, fluctuating market prices and global warming has revived the interest in the promotion of biofuel from non-conventional sources. The potential of *Pongamia* has not been fully explored, mainly due to its variable and unpredictable oil yield that restricts large-scale plantation. Genetic improvement may alleviate this problem. However, recognition of genetic diversity and characterisation of the existing germplasm is needed for developing superior genotypes with desired traits (Kesari and Rangan, 2011). Collection, identification and conservation are essential components in recognising elite cultivars from existing plant genetic resources (Ramanatha Rao and Hodgkin, 2002, Tena Gashaw et al., 2016). A very little attempt has been made in the field of molecular genetics for the advancement of existing *Pongamia* germplasms. Hence the assessment of available genetic diversity in a naturally growing population is essential in releasing commercially important cultivars.

Genetic diversity primarily can be assessed with both conventional and molecular markers. Markers are important tools in genetics because of their ability to differentiate between various genotypes. They mainly consist of biochemical constituents (e.g. secondary metabolites, allozyme) and macromolecules, such as proteins and deoxyribonucleic acid (DNA). However, various issues with a biochemical marker such as environmental variation, tissue specificity and limited availability restrict their more extensive use. Hence, DNA markers are one of the most reliable and ubiquitous among the molecular markers used to date in most of the living organisms for the detection of a variation. Over the years, rapid advancement in molecular genetics field has led to the development of a range of DNA markers. DNA markers are the fragment of DNA which is used to distinguish polymorphism in genes or DNA sequence. The application of DNA markers has had a revolutionary impact on the investigation of genetic variations, species identification, genetic map construction and in plant breeding for marker-assisted selection (Collard and Mackill, 2008). The popular marker systems that are now regularly used include restriction fragment length polymorphism (RFLP) (Botstein et al., 1980), random amplification of polymorphic DNA (RAPD) (Williams et al. 1990), simple sequence repeat (SSR) (Tautz, 1989) and amplified fragment length polymorphisms (AFLP) (Vos et al., 1995). Furthermore, recently next generation markers are also employed for analysis, which includes diversity arrays technology (DArT), single nucleotide polymorphisms (SNPs) and genotyping by sequencing (GBS). RFLP and AFLP markers are time taking and cumbersome, RAPD results are often not reproducible between laboratories. Among the various existing markers, microsatellites have evolved as an important marker in plant breeding applications. In addition, microsatellites based markers are more rapidly automated, reproducible and can efficiently identify polymorphisms (Zhang et al., 2014). However, it remains a cumbersome task to identify highly polymorphic and tightly linked molecular markers for the important trait. The evolution of new gene-based molecular markers system has opened the doors for RNA-based (cDNA or EST or transcriptomic) markers (Xiao et al., 2014). The next-generation sequencing (NGS) has become the most powerful method for generating DNA markers within a short timeframe. Development of NGS based marker mainly involves the library preparation before sequencing

(Xiao et al., 2014, Zhang et al., 2016). Several markers are derived from NGS platforms which can be involved from the partial genome or whole genome sequence or transcriptome library.

Microsatellites are polymerase chain reaction (PCR) based markers, which occur as a tract of interspersed repetitive DNA ranging from length 1–6 base pairs motifs and are generally repeated 5–50 times. SSRs are favoured for a variety of analysis due to their simplicity, multiallelic nature, reproducibility, cross transferability, codominant inheritance, locus specificity and uniform distribution throughout the genome (Zhang et al., 2006, Agarwal et al., 2008, Parida et al., 2010). SSRs are distributed throughout the genome, including coding and non-coding region of the nuclear genome and plastid genome (Kuntal et al., 2012, Zhu et al., 2016). Mitochondrial and chloroplast genome are inherited in a maternal as well as uniparentally pattern (Birky, 1995). Due to the slow rate of mutation; organeller SSR have often been used in population genetics and phylogenetic studies (Pervaiz et al., 2015). However, genic regions of genome also harbour enormous microsatellites. Hence, expressed sequence tags (ESTs) and transcriptome libraries are often used for SSR marker preparation (Huang et al., 2016, Ul Haq et al., 2016b). Moreover, the marker derived from a transcribed portion of genes involved in the variety of metabolic functions and unveils the biological significance (Savadi et al., 2012, Zhang et al., 2016). Therefore, the present study was conducted to explore the genetic variability among the *Pongamia* accessions using EST-SSR markers for important traits. Transcriptome assembly was prerequisite for the mining of EST–SSRs; hence the raw libraries of *Pongamia* were downloaded from the NCBI for final transcriptome assembly (Huang et al., 2012). To evaluate the genetic diversity for the important traits in *Pongamia*, the broad objectives of the present chapter are

i) Identification and characterisation of EST and organellar SSRs
ii) Designing of EST-SSR primers and their validation to evaluate the level and pattern of genetic relatedness among the *Pongamia*
iii) Cross-species transferability of EST-SSR markers

5.2. Review of literature

Plant biodiversity is nothing but the total variability in genetic and phenotypic features of plants in their habitats. However, biological diversity is an important value in reorganisation and management of natural resources. The variation in traits shown by individual populations at morphological and genetic level creates the foundation for the evolutionary potential and might help in the face of changing environmental conditions. The knowledge of population structure and genetic variation in available germplasm is very crucial to ascertain the germplasm conservation and breeding programs (Zoratti et al., 2015). Broad gene pool or genetic diversity is required for the development of effective and successful plant breeding project (Hutchinson, 1940).

Traditionally, the genetic diversity assessment has been conducted using morphological features, especially those which can be monitored based on phenotypical traits of interest. Some of the important agronomic traits like plant height and size, branching pattern, palmate leaf number and colour, flower colour, seed and pod morphology, and oil content in diverse seed sources of *Pongamia* have been investigated (Kaushik et al., 2007, Kesari et al., 2008, Mukta et al., 2009, Sunil et al., 2010, Rao et al., 2011, Sahoo et al., 2011, Jiang et al., 2012). High level of polymorphism is expected in *Pongamia* due to its out-crossing reproductive nature and adaptability in different agro-ecological conditions. Improving oil yield and seed germination vigour are important breeding objectives. Morphological characterisation of pod and seed is one of the critical steps in ascertaining the genetic diversity in the wild accessions (Patil and Naik, 2016). Furthermore, seed polymorphism plays a pivotal role in seed germination, seedling survival and growth in *Pongamia* (Pathak et al., 1980, Manonmani et al., 1996). Patil and Naik (2016) showed significant variations in pod and seed features, oil content, and also reported the positive correlation in pod and seed traits during the progeny trial in North Karnataka, India. Jiang et al. (2012) reported that the content and seed oil composition varied between the trees and within the progeny of the single parent tree. The variation found in the *Pongamia* seed sources is mainly ascribed to the heterogeneity of the genotypes and their environment interactions (Raut et al., 2011). Selection of the plant material with superior phenotypes is an essential step

in any tree improvement strategies. However, the morphological marker traits are limited in number, susceptible to phenotypic plasticity and environmental condition. To reduce down the impact of environmental influence in the analysis, biochemical markers are often used as isozyme and protein electrophoresis was conducted.

Advances in molecular biology led to the emergence of numerous DNA markers like random amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), inter-simple sequence repeats (ISSRs), and simple sequence repeats (SSRs). These markers provide good reproducibility, avoid environmental influence and provide better resolution of genetic variation at an early stage of the plant. However, various types of DNA marker have been employed to study pattern and extent of genetic variation in *Pongamia* population. Despite many newly developed genetic markers, RAPD is widely used techniques to estimate genetic diversity quickly. Kesari et al. (2010) reported the variation amongst the *Pongamia* candidate plus trees (CPTs) based on morphometric features, especially pod and seed traits. Furthermore, genetic diversity was also assessed using RAPD primers to quantify the diversity of malapari in Java Island (Aminah et al., 2017). RAPD technique is quick, easy to perform, cost-effective and produces large quantities of polymorphic DNA bands (Tingey and del Tufo, 1993). However, this technique is prone to disadvantage like lack of reproducibility of amplification due to mismatch annealing (Jones et al., 1997). An ISSR marker uses repeat-anchored or non-anchored primers for the amplification of genomic sequences between two microsatellite regions (Zietkiewicz et al., 1994). The marker has been successfully employed for numerous species such as Wild rice (Qian et al., 2001), Cashew (Archak et al., 2003), Barley (Guasmi et al., 2012) and Durum wheat (Etminan et al., 2016), due its relative abundance, good reproducibility and discriminating information. (Sujatha et al., 2010) employed the ISSR marker to study the genetic variability among tissue culture raised *Pongamia* samples. AFLP analysis is powerful, and produces reproducible polymorphic bands and requires no sequence data for primer designing. Alongside, AFLP is the most time-consuming and labour intensive method. Due to its high marker index, higher polymorphic information and resolving power AFLP marker were found to be useful in detecting genetic diversity in 33 CPTs of *Pongamia* representing five agro-ecological zones of southern India (Pavithra et al., 2014). However, the bands per reaction in AFLP marker can

be “fine-tuned” during the selective amplification step either by increasing or reducing down the number of selective nucleotides (Vos et al., 1995). Sharma et al. (2011) investigated the efficacy of two molecular markers namely AFLP and three endonucleases (TE)-AFLP in twenty *Pongamia* individuals.

Meanwhile, the emergence in next-generation sequencing (NGS) method led to the generation of new generation DNA markers within a short timeframe. NGS is an accurate and cost-effective high throughput sequencing technology which can be used to reveal massive sequencing data for several non-model species. Development of NGS based marker consists of the preparation of the library before sequencing, which could be either a partial genome or whole genome sequence or transcriptome library. Transcriptome library offers a simple, rapid and economical way for mining a significant amount of genic or unigene-based SSR markers for gene tagging and MAS in many plants. Microsatellites are PCR based marker, which occurs as a tract of interspersed repetitive DNA ranging from length 1–6 base pairs motifs and are usually repeated 5–50 times. Microsatellites are favoured most in plant genetics and breeding applications for selections during backcross breeding programs. SSR possesses important attributes like simplicity, multiallelic nature, reproducibility, cross transferability, codominant inheritance, locus specificity and uniform distribution throughout the genome.

For long, genomic SSRs have been used as a co-dominant marker in plant breeding and phylogenetic studies. Meanwhile orgeneller SSRs gained popularity mainly due to the unipaternal pattern of inheritance and remain unperturbed by recombination (Provan et al., 1999). Recent revolutions in the sequencing of whole organellar genomes opened the door for the estimation of SSRs frequencies at the whole genome level. Due to a large number of availability of plastid sequences, a considerable number of markers has been designed to determine genetic diversity. However, genic regions of chromosome also harbour enormous microsatellites. Hence, expressed sequence tags (ESTs) and transcriptome libraries are often used for SSR marker preparation. Moreover, the marker derived from a transcribed portion of genes are involved in the variety of metabolic functions and unveils the biological significance. Genic microsatellite markers are sometimes less polymorphic than genomic

SSRs due to their association with conserved coding regions, unlike non-coding ones. Hence, in most instances, they are transferable across species and serve as a useful tool for gene discovery, population genetics, gene tagging and genetic structure analysis. Transferability study of SSR marker has been employed in several plant species like barrel medic, cotton, sorghum, peanut and Poaceae plants etc (Eujayl et al., 2004, Han et al., 2006, Nagaraja Reddy et al., 2012, Savadi et al., 2012, Ul Haq et al., 2016a). Several studies have been carried over on the feasibility of utilising EST-SSR marker between monocot and dicots plants, as some genes display a significant level of sequence conservation. Understating the pattern of genetic diversity within a species is critical to comprehend the population structure, local adaptation and diversity among the populations. SSRs have often been used for marker-assisted selection (MAS) which is a pivotal method for the improvement of many crops plants. Thus, one of the pressing needs of *Pongamia* genomics research is to develop molecular markers for superior traits such as high oil content and yield. More recently, genic-SSRs were designed for *Pongamia* using seed transcriptome assembly (Huang et al., 2016, Sreeharsha et al., 2016). Despite so many studies, to date, no genetic map of *Pongamia* has been reported. The present study describes the *in silico* mining and development of microsatellites (SSRs) using the *Pongamia* transcriptome assembly from leaf and root tissue libraries developed earlier (Huang et al., 2012).

5.3. Material and methods

5.3.1. Plant material

In the present study fourteen *Pongamia* accessions collected from different agro-climatic locations of six Indian states (Assam, Maharashtra, Orissa, Rajasthan, Telangana and Uttar Pradesh) were analysed (Fig 5.1). The different *Pongamia* accessions and their details of collection from various agro-climatic locations (latitude and longitude) are given in Table 5.1. Two *Pongamia* accessions CPT-29 and NGPP46 (North Guwahati *Pongamia pinnata)* from published literature were also included in the current study (Kesari et al., 2008). The rest of the accessions were named by initial PP i.e. *Pongamia pinnata*. Plants were raised using seeds in poly bags in Greenhouse at Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India.

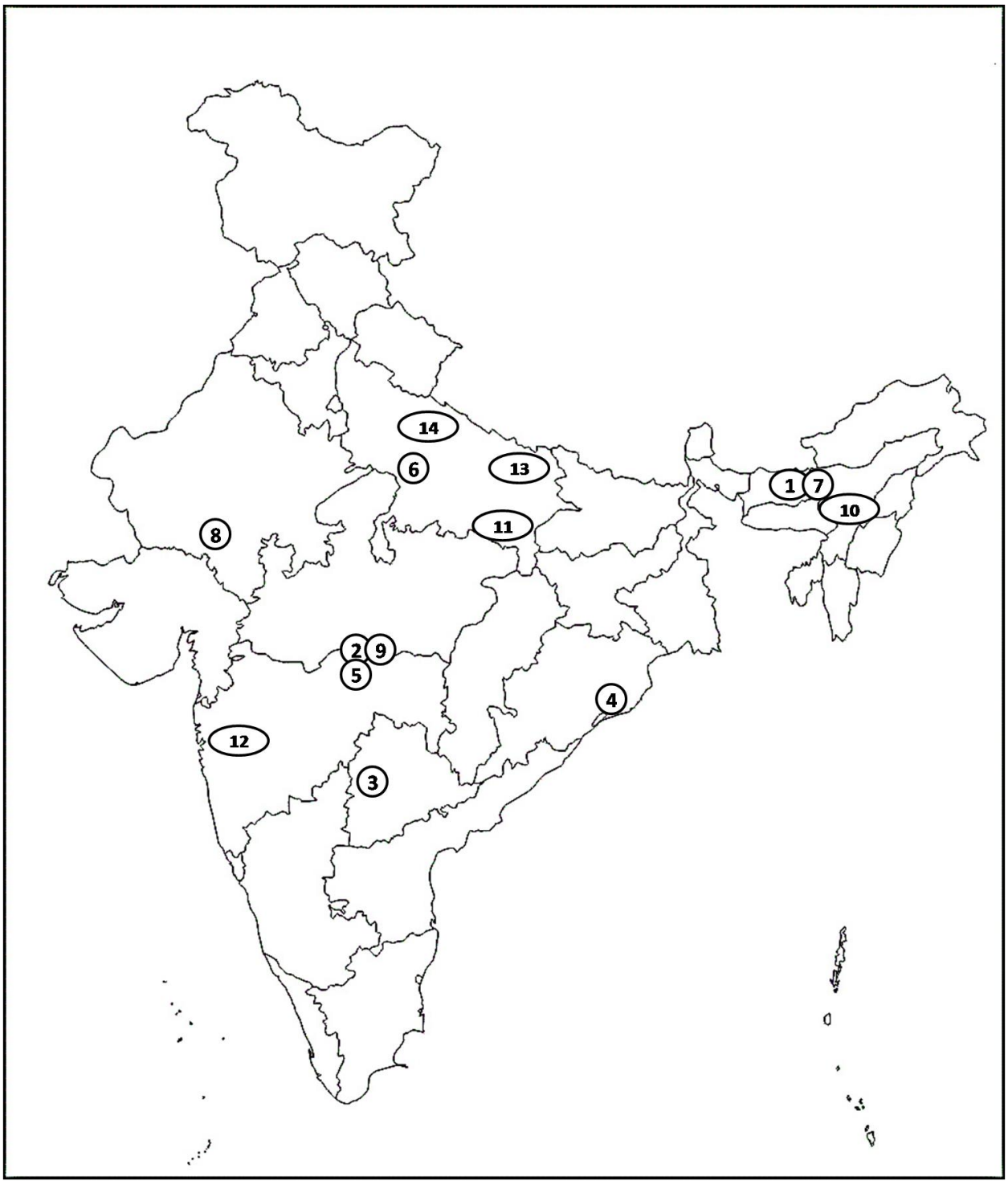


Figure 5.1. Map of India showing collection sites of *Pongamia* accessions from different states. Numbers mentioned on the map represents the serial number of *Pongamia* accession cited in Table no 5.

Table 5.1. Details of accessions of *Pongamia* used for diversity analysis.

S. N	Accessions	Source	State	Latitude	Longitude	Region
1	CPT-29	Guwahati	Assam	26.1445169	91.7362365	Bengal Assam plain
2	PP-1	Amravati	Maharashtra	20.9374238	77.7795513	Deccan plateau
3	PP-2	Hyderabad	Telangana	17.385044	78.486671	Deccan Telangana plateau and eastern ghat
4	PP-3	Bhubaneswar	Orissa	20.2356169	85.745673	Eastern coastal plain
5	PP-4	Akola	Maharashtra	20.7059345	77.0219019	Deccan plateau
6	PP-5	Kanpur	Uttar Pradesh	26.4148245	80.2321313	Northern plain
7	**NGPP-46**	Guwahati	Assam	26.1445169	91.7362365	Bengal Assam plain
8	PP-6	Udaipur	Rajasthan	24.585445	73.712479	Northern Plain and Central Highlands including Aravallis
9	PP-7	Melghat Tiger Reserve, Amravati	Maharashtra	21.4030119	77.3268121	Central Highlands
10	PP-8	Silchar	Assam	24.8332708	92.7789054	North eastern hills
11	PP-9	Banaras	Uttar Pradesh	25.3176452	82.9739144	Northern plain and central highlands including Aravali
12	PP-10	Pune	Maharashtra	18.5204303	73.8567437	Deccan plateau
13	PP-11	Gorakhpur	Uttar Pradesh	26.7605545	83.3731675	Eastern plain
14	PP-12	Lucknow	Uttar Pradesh	26.8466937	80.946166	Northern plain and central highlands including Aravali

Refference:-http://vikaspedia.in/agriculture/crop-production/weather-information/agro-climatic-zones-in-in

5.3.2. Genomic DNA extraction

The total genomic DNA was isolated from young fresh leaves of *Pongamia* using modified sodium dodecyl sulphate (SDS) method (Kesari et al., 2009). About 5 g fresh and young leaves were collected for grinding using liquid nitrogen along with 2% PVP (Polyvinylpyrrolidone) in mortar and pestle to obtain a fine powder. The fine powder was immediately transferred to 50 ml polypropylene centrifuge tube and gently suspended in two volumes of preheated extraction buffer at 65º C, incubated for 30 min at 65º C in water-bath and mixed by gentle shaking after every 10 min interval. A double volume of chloroform: isoamyl alcohol (24:1) was added and the tubes were inverted gently shaken for 15 to 20 times and centrifuged for 20 min. at 10,000 rpm at room temperature. The upper aqueous phase was carefully transferred by wide-bore of tips to a fresh sterile 50 ml centrifuge tubes to avoid mechanical damage to DNA. Two volumes of ice-cold isopropanol were added to collect the upper aqueous phase, and the tube was gently shaken and kept at – 20º C for 1 hr to precipitate the DNA. The precipitate was centrifuged at 12,000 rpm for 15 min, and the supernatant was discarded. The pellet was washed with 70% chilled ethanol by centrifuging at 12,000 rpm for 15 min. The pellet was air-dried and suspended in 500 µl of TE buffer (pH- 8.0).

For purification of extracted genomic DNA, 3 µl RNase A (10 mg/ml) was added to the sample and the mixture was kept at 37º C for 30 min. An equal volume of chloroform: isoamyl alcohol was added in the sample followed by centrifugation at 10,000 rpm for 5 min. The aqueous phase was collected in a fresh vial; ethanol precipitation was carried out in the presence of 3 M sodium acetate (pH 5.2). The precipitated DNA was centrifuged to a pallet and washed in 70% ethanol, air or vacuum dried. The final DNA pellet was dissolved in 30 to 50 µl (depending upon the pellet) of TE buffer.

5.3.3. Quantification and quality check of genomic DNA

The genomic DNA yield was determined using a Nanodrop spectrophotometer Tecan Infinite 200 PRO (Nanodrop Technologies, DE, USA) as per standard manufacturer's instructions. The ratio of absorbance at 260 nm and 280 nm was used to assess the purity of DNA and RNA. Purity DNA purity was determined by calculating the ratio of absorbance at A260 nm

and 280 nm. The concentration was recorded in ug/µl. In addition, the quality and concentration of genomic DNA was also determined by running 3 µl of DNA from each sample on a 0.8 % agarose gel containing 0.5 µg/ml of ethidium bromide (EtBr).

5.3.4. EST-SSRs identification

In the present study, total of four transcriptome nonredundant unigene datasets as described earlier in chapter 3 were investigated for mining of microsatellite population. Microsatellite identification was carried out using MicroSAtellite (MISA) software (http://pgrc.ipk-gatersleben.de/misa/misa.html), a Perl script, employed to detect perfect and compound SSRs in unigene sequences. Compound SSRs (two or more SSRs in the 50 bp interval) were considered for investigation. The SSRs were considered to contain mono- to deca-nucleotides motifs. The minimum repeat unit was set to 10 for mono- nucleotides, 3 for di- nucleotides to octa- nucleotides, 2 for nona-nucleotides, and 1 for deca- nucleotides motifs, respectively.

5.3.5. Organellar SSRs identification

Complete genome sequence of mitochondria and chloroplast of *Pongamia* were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database (GenBank accession no. JN673818.2 and JN872550.1). MIcroSAtellite identification (MISA) tool was employed for the screening of microsatellite motif in chloroplast and mitochondrial genome using the same parameters as described for EST-SSR mining.

5.3.6. EST-SSR sequences annotation

Unigene sequences having SSRs were employed for annotation using BLASTX search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with a cut-off e-value = 1e-5 against protein databases such PLAZA 2.0 and 3.0 (https://bioinformatics.psb.ugent.be/plaza/), Swiss-Prot (http://www.expasy.ch/sprot/) and *G. max* database at Ensemble plants (plants.ensembl.org/). Additionally, all EST libraries containing SSRs were combined and further assigned to functional annotation with Panther online program (http://www.pantherdb.org/), gene ontology (GO) which classifies the sequences to molecular function, biological process, cellular component and protein class.

5.3.7. EST-SSR primers designing

The EST-SSRs loci were used to design primer using PRIMER version 3.0 (http://bioinfo.ut.ee/primer3-0.4.0/). For designing primer pairs, simple and compound EST-SSR repeats were considered that contained motifs mono to deca-nucleotides in size. Parameters for designing the primers were set as follows:

(1) PCR product size ranging from 150 to 300 bp length;
(2) Primer length of 18–22 bp with an optimum of 20 bp;
(3) Melting temperature (Tm) of 55–65º C with 60º C as the optimum;
(4) GC content ranging from 40–60%

Of the total designed primers, twenty EST-SSR primers were selected from the current study. Four EST-SSR primers were selected from the previously published literature (Huang et al. 2016). Hence, a total of 24 primers were synthesised by Eurofins Scientific, India, for further analysis (Table 5.3).

5.3.8. PCR validation

PCR amplification was carried out using isolated DNA from *Pongamia* accessions in Mini Thermal Cycler (Applied Biosystems 9700, USA). Twenty-four EST-SSR primers were selected for PCR validation. The list of each marker, repeat type and length primer sequence and annealing temperature are mentioned in Table 5.3. PCR amplification was conducted in 25 µl reaction volume containing 50 ng of DNA, 2X PCR master mix pH 8.5 (Promega, USA), 400µM dNTP, 3mM $MgCl_2$, nuclease-free water (Promega, USA) and 0.6 µl of 0.1–1.0µM of each SSR forward and reverse primer. The reaction was performed in 0.2 ml microfuge tube (Dialabs, USA). The PCR cycling was as follows: 3 min at 94º C initial denaturation followed by 35 cycles of 40 s at 94º C, 40 s at annealing temperature (Tm), 72º C for 40 s and the final extension of 3 min at 72° C. The genomic DNA amplified by EST-SSR primers were initially checked for amplification on 1.7% agarose gel. For conducting electrophoresis, agarose gels were prepared using agarose (Sigma, USA) in 1X Tris-Borate EDTA (TBE) buffer using horizontal agarose gel slab apparatus (Bio-Rad, USA). Agarose powder was suspended in TBE buffer and dissolved by heating in a microwave oven. The molten agarose solution was

cooled down at 60º C followed by addition of 1.0 µg/ml ethidium bromide. Wells were made by keeping the comb in the gel, before the pouring of molten agarose. The solidified agarose gel was immersed in an electrophoresis tank containing 1X TBE buffer followed by removal of comb from the gel. The PCR amplified samples and 50 bp size ladder (Himedia, India) as a reference marker with loading dye were loaded into the wells. Electrophoresis was carried out at 5V/cm for 1 hr. PCR amplified bands in the gel were observed under UV-transilluminator followed by gel documentation (Bio-Rad, USA). After confirmation on the agarose gel, the amplified PCR products were finally separated on 8 % polyacrylamide gel.

5.3.9. Polyacrylamide gel electrophoresis (PAGE)

All the amplified EST-SSR PCR products were resolved using 8% polyacrylamide gel (29:1 acrylamide: N, N'-methylene bisacrylamide, 1X TBE buffer) on a Bio-Rad Sequi-Gen gel apparatus (Bio-Rad, USA). Following steps are involved in resolving the EST-SSR marker product on PAGE gel. (i) The glass plates and notched IPC (Integral Plate Chamber) were washed with labolene detergent, rinsed by distilled water and then allowed to air dry. (ii) Plates were again wiped with 70 % ethanol before use. (iii) The thin and thick glass plate pairs were assembled in the precision caster base with gasket using GT lever clamps. (iv)The required amount of TEMED was quickly added to the 8% polyacrylamide gel solution before the acrylamide polymerises, then the solution was poured between the plates using a 1ml pipette. (v) Immediate after the gel pouring, the comb was carefully inserted into the gel without allowing any air bubbles to trap under the teeth. (vi) The gel was allowed to polymerise for about 15 min. at room temperature. (vi) After the complete polymerisation, the gel was removed from gel caster; spilt gel was carefully cleaned from the back of white plates and inserted into Hoefer gelbox. (vii) After pouring of 0.5X fresh TBE running buffer, the comb was carefully removed from the polymerised gel. (vii) The wells were flushed out once more with 0.5X TBE using a Pasteur pipette. Total 6 µl of PCR samples and 50 bp DNA ladder (Himedia, India) with loading dye were loaded into the wells using a micropipette. (viii) After setting the assembly, the gel was run for about 3 hrs at a constant voltage of 40 V. The gel was run until the marker dyes migrated to the end of the gel. (ix) After electrophoresis, plates were detached, and the gels were removed carefully from the glass plates using a spacer.

The removed gel was transferred to a tray containing double distilled water for 2 min. for washing with gentle shaking. (xi) An acraymiade gel was kept in distilled water containing 0.5µg/ml EtBr for 15 min. with gentle shaking (xii) Furthermore, the excess of EtBr dye was rinsed away by washing the gels in distilled water for 10 min. with gentle shaking. (xiii) PCR amplified bands in the gel were observed under UV-transilluminator and documented using documentation with a gel documentation system (Bio-Rad, USA).

5.3.10. Transferability of *Pongamia* EST-SSR markers in different plants

Different plants such as *Jatropha curcas, Ricinus communis, Mesua ferrea, Glycine max, Cicer arietinum, Arachis* hypogeal*, Vigna radiata, Oryzae sativa*, *Musa acuminata, Curcuma longa, Solanum melongena* and *Phaseolus vulgaris* were raised using seeds in poly bags in Greenhouse at Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India. Total genomic DNA was isolated from fresh leaves using the above mentioned modified sodium dodecyl sulphate (SDS) method. Primer pairs that can amplify a clear band in *Pongamia* were selected for further transferability investigation. Hence, sixteen EST-SSR markers were tested for PCR amplification in different plants using the same PCR amplification protocol mentioned in the PCR validation section.

5.3.11. Statistical analysis of EST-SSR markers data

The EST-SSR amplified bands on PAGE were scored as present (+1) and absent or missing (0) for each primer. This binary matrix subjected to various statistical analysis using different software. For each marker, a duplicate sample from each plant was tested and only clear and reproducible bands were considered for further data analysis. The numbers of amplified polymorphic and monomorphic PCR products were determined for each primer against fourteen *Pongamia* accessions. Polymorphic information content (PIC) value was determined to compare the efficiency of primers PIC following Botstein et al. (1980). Marker index (MI) was also determined. The EST-SSR allelic data were converted into a binary matrix which was used to calculate the level of similarity among different accessions. Then level similarities among the accessions were established as polymorphic bands and the matrix of genetic generated by using Dice's coefficient (1945) using the SIMQUAL program of NTSYS.

Unweighted Pair Group Method with Arithmetic Mean (UPGMA) **(Sneath and Sokal, 1973)** method was employed on this matrix using the SHAN subroutine through the NTSYS- pc (Numerical taxonomy system, 2.2 version) (Numerical taxonomy system, Applied Biostatistics, N.Y.) (Rolf, 2012). A dendrogram was generated representing the genetic relationship among fourteen *Pongamia* accessions. The correlation between the original similarity indices and cophenetic values was performed using 300 permutations to check the goodness of fit of the *Pongamia* accessions to a specific cluster in the UPGMA cluster analysis.

5.4. Results and discussion

5.4.1. Isolation and characterisation of EST-SSRs

Recently, due to the accessibility of a significant amount of genomic information, a lot of EST datasets are available for many crop plants. These databases have been successfully utilised to develop novel molecular markers linked to genes or agronomically important traits (Zhang et al., 2016, Ul Haq et al., 2016a). Although, the *Pongamia* tree is an important source of non-edible-oil. Interestingly, very few EST-SSRs or gene-linked markers have been developed and tested in this crop. These facts provided the avenue to focus research on mining and development of EST-SSRs marker in *Pongamia*.

To design EST-SSR marker, we collected high throughput sequencing reads from publically available *Pongamia* libraries SRR349650 (Root Seawater treated), SRR349651 (Root freshwater treated), SRR349652 (Leaf Seawater treated), SRR349653 (Leaf freshwater treated) using short read archive (SRA) toolkit (Huang et al., 2012) (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsof). Four *Pongamia* transcriptome libraries were assembled using Trinity assembler as detailed in chapter 3.

A total of 219881 unigenes or EST sequences were examined across the four *Pongamia* libraries for simple sequence repeat (SSR) mining. Identification and characterisation of SSR were carried out using MIcroSAtellite identification tool (Table 5.2). The four libraries yielded around total 157802 EST-SSRs, of which 1,23,115 were simple SSRs and 34,687 SSRs were in compound formation. On average, one microsatellite was found in every 0.98 kb of *Pongamia* ESTs which is closed to some earlier reported plants like *Solanum lycopersicum* (SSRs per 1.3 Kb) and *Prunus* species (SSRs per 1.6 Kb) (Gupta et al., 2010, Sorkheh et al., 2016). In contrast, the reported density of SSRs in the present study is much higher than other legumes, such as chickpea (SSR per 8.54 kb) and *Medicago* (SSR per 7.47 kb) respectively (Agarwal et al., 2012, Wang et al., 2014). However, these variations in EST-SSRs frequency and abundance in different plant species may be due to the search criteria used, as the size of the dataset, type of SSR motif, and mining tools used. The highest amount of EST-SSRs was observed in SRR349650 library and lowest in an SRR349651

library. Analysis revealed that the most frequent number of EST-SSRs in *Pongamia* were dinucleotide repeats (67-72 %), followed by tri (23-27%), mono (2.2-2.6%), tetra (1.2-1.5%), penta and hexanucleotide (0.5-0.7%) (Fig. 5.2). The present report of high abundance of dinucleotide repeats corroborates with previously conducted studies on *Mentha piperita* (Kumar et al., 2015). Several earlier studies have also stated the high prevalence of dinucleotide repeats in different plant species such as coffee, Lactuca Species, *Jatropha* and Adzuki bean (Aggarwal et al., 2006, Riar et al., 2011, Yadav et al., 2011, Chen et al., 2015). The dominance of smaller SSRs repeats was observed among the analysed repeats. In contrast to our results, Sreeharsha et al. (2016) reported mononucleotide (36%) repeats as the largest fraction followed by trinucleotide (31.3%) repeat in *Pongamia* seed transcripts. In the current study, mononucleotide sequences were extracted, which was reported in very few plants like *P. vulgaris, V. radiate* and *Pongamia* (Garcia et al., 2011, Chen et al., 2015, Sreeharsha et al., 2016). Among the SSRs identified, nona nucleotide repeats were absent in SSR349652 and SSR349653 assemblies. The presence of repeat motifs decreases with the increase in the length of the repeat motif. This is in agreement with the fact that longer repeats are less stable due to higher mutation rates (Toth et al., 2000).

Among the dimeric motifs, the most frequent dinucleotide repeat motif was AG/CT (37%). Repeat motif, AC/GT was the second most abundant repeat among dinucleotide and accounted for 30% followed by AT/AT (17%). These repeat distributions are almost similar with results sited for *Iris*, *Brachypodium*, peanut and sugarcane (Tang et al., 2009, Sonah et al., 2011, Bosamia et al., 2015, Ul Haq et al., 2016a). AG/CT motifs have often been used for EST-SSR marker development due to their positive role in recombination, high abundance and polymorphic nature (Temnykh et al., 2001, Morgante et al., 2002, Guo et al., 2008). This motif was also present in both rice and human genomes (Guo et al., 2008).

Of the total trinucleotide motifs, AAG/CTT (16%) was the most prominent repeat, which is responsible for coding leucine and lysine. The second most trinucleotide abundant motifs was ATC/ATG (12%) followed by AGC/CTG (11%), respectively (Fig. 5.3). Repeat motif, AAG/CTT was also reported earlier in *Nelumbo nucifera*, faba bean, *M. piperita* (Pan et al., 2010, Akash and Myers, 2012, Kumar et al., 2015). The occurrence of trinucleotide

repeats motif in the coding region could result in forming a distinct group and encoded amino acid tracts in peptides sequences. However, the trinucleotide repeats in ESTs could play a vital role in various cellular, biological, and metabolic processes in plants (Kumari et al., 2013).

5.4.2. Isolation and characterisation of organeller SSRs

The pattern and distribution of SSR repeats in genomes vary significantly between different plants. The detailed analysis of frequency and distribution pattern of different repeats motifs from the *Pongamia* organellar genome was conducted using the MISA tool.

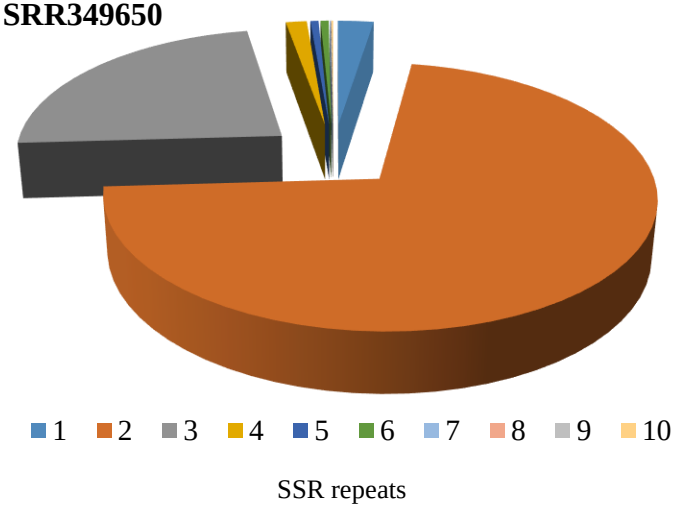
A total of 620 chloroplastic (cp) microsatellites or cpSSRs were identified in the chloroplast genome sequence. Septa-, octa- and nonanucleotide repeats were not detected in the analysis. The statistical distribution of identified cpSSRs presented in Fig. 5.2. Of the total 620 cpSSRs, 434 (70%) were simple cpSSRs and 186 (30%) cpSSRs were in compound formation. Among the identified repeats, dinucleotides (71.9%) were the most frequent repeat type, followed by mononucleotides (13.38%) and trinucleotide (12.90%) repeats. The abundance of dinucleotides in present investigation contradicts with previous cpSSRs studies in *Solanaceae* species, *Olea* species and *Anthoceros formosae* (Tambarussi et al., 2009, Filiz and Koc, 2012, Shanker, 2013). The density of SSRs in 152.9 kb of chloroplast genome was found to be 0.24 kb per cpSSRs, which is higher than *Olea* species (1.47 kb per SSR) and *Solanaceae* species (1.26 kb per SSR) respectively (Filiz and Koc, 2012, Tambarussi et al., 2009). Due to the choice of different parameters for SSR detection, e.g. minimum length of SSRs, repeat motifs, amount of data analysed and composition of genome sequence could be the cause of variations in SSR density. The dimeric motifs are abundantly present in *Pongamia* chloroplast genome. Of the total dimeric motifs, AT/AT cpSSR motif was present in ample amount (Fig. 5.3). Similar results were observed in rice, Brassicaceae family, Olive species and Glycine species (Rajendrakumar et al., 2007, Gandhi et al., 2010, Filiz and Koc, 2012, Ozyigit et al., 2015). AT repeats motifs are present in a huge amount in plants as AC repeats in animals, which could be the distinguishing feature for plant and animal genomes (Powell et al., 1996b). Mining of cpSSRs in the present study revealed the dominance of AT/AT type of repeat motif, which corroborates with current *Pongamia* EST-SSRs results.

The mitochondrial genome of *Pongamia* was analysed for mining of mitochondrial SSRs (mtSSRs), a total 1,327 mtSSRs were detected of which 1,123 (82%) were simple mtSSRs and 204 (18%) compound mtSSRs. Among the repeat types, dinucleotides were the most frequent repeats (50.7%) followed by trinucleotides (36.2%), and tetranucleotides (7.9%) (Fig. 5.2). Dinucleotide repeats were also significantly present in rice (Rajendrakumar et al., 2007). The current results are contradictory to the previous observations found in Brassica species and *Salix purpurea* (Filiz, 2013, Wei and Cao, 2016). Interestingly, there were no septa, hepta, octa, nona and deca nucleotide motifs found in the analysis. The SSRs density in 425 kb of *Pongamia* mitochondria was 0.32 kb per mtSSR, which was higher than Brassica species 1.03 kb per mtSSR (Filiz, 2013). Among the mononucleotides repeat motifs, A/T motif was abundantly present. In dinucleotide repeat motifs, AG/CT motif (47.2%) was present in the highest amount followed by AT/AT (15.5%) and CG/CG (5.4%). Of the trinucleotide motifs, AAG/CTT motif was present in ample (7.2%) (Fig. 5.3). According to Rajendrakumar et al. (2007) and Rajendrakumar et al. (2008) A/T, AT/TA, and AAG are most frequent motifs in the mitochondrial genome of higher plants. Dinucleotide AG/CT repeat motif was also abundantly found in Brassica species (Filiz, 2013). However, from the current study, it is evident that the dinucleotide SSRs occupied a major portion of the transcriptome and organeller genome of *Pongamia.*

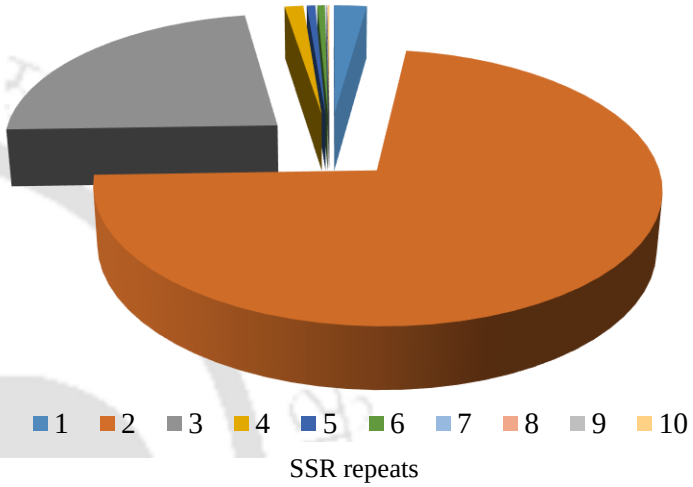
Table 5.2. Details of EST and organellar SSRs search statistics.

S.N	Contents	SRR349650	SRR349651	SRR349652	SRR349653	Chloroplast genome	Mitochondrial genome
1	Total number of transcriptome sequences /organellar genome	58,910	60,600	49,815	50,556	1 (152.96 kb)	1 (425.71 kb)
2	Total number of identified SSR after annotation	45,210	45,152	31,776	35,664	620	1327
3	Number of SSR containing sequences	6,840	8,263	4,454	4,980	--	--
4	Number of sequences containing more than 1 SSR	6,444	7,019	4,363	4,925	--	--
5	Number of SSRs present in compound formation	9,943	9,299	7,411	8,034	186	204

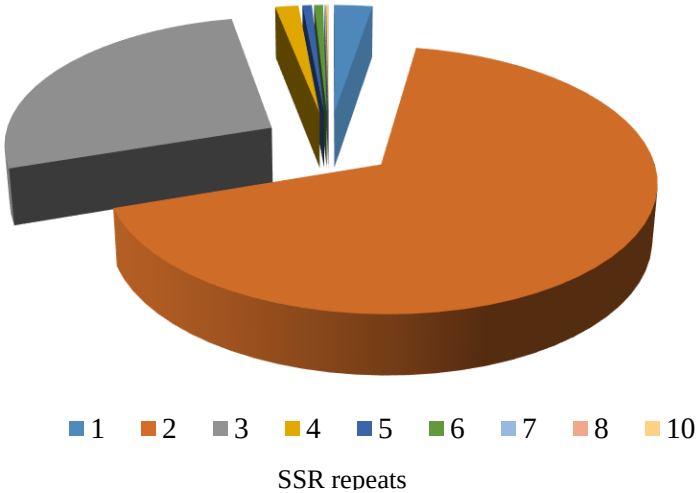
A) SRR349650



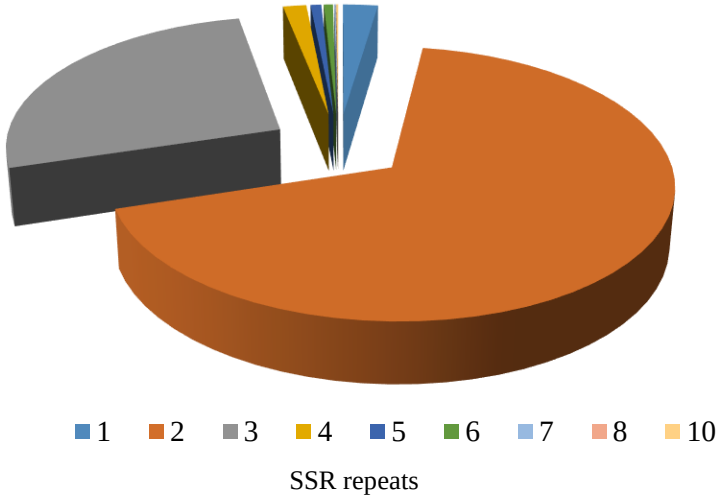
B) SRR349651



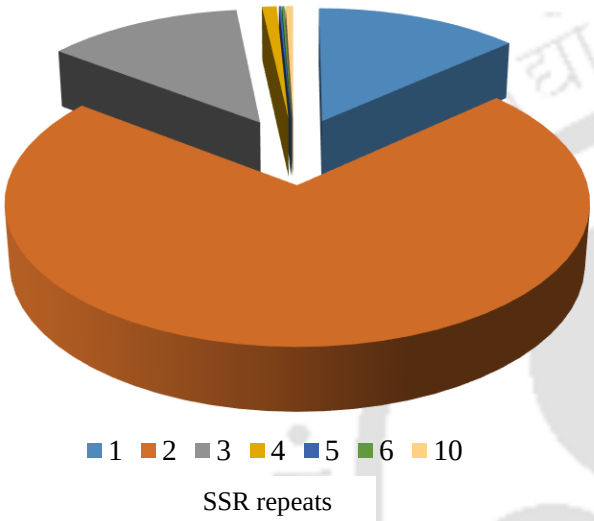
C) SRR349652



D) SRR349653



E) Chloroplast genome



F) Mitochondrial genome

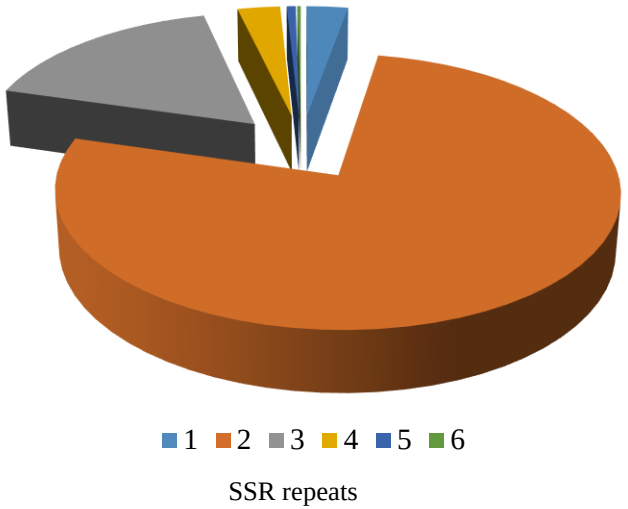
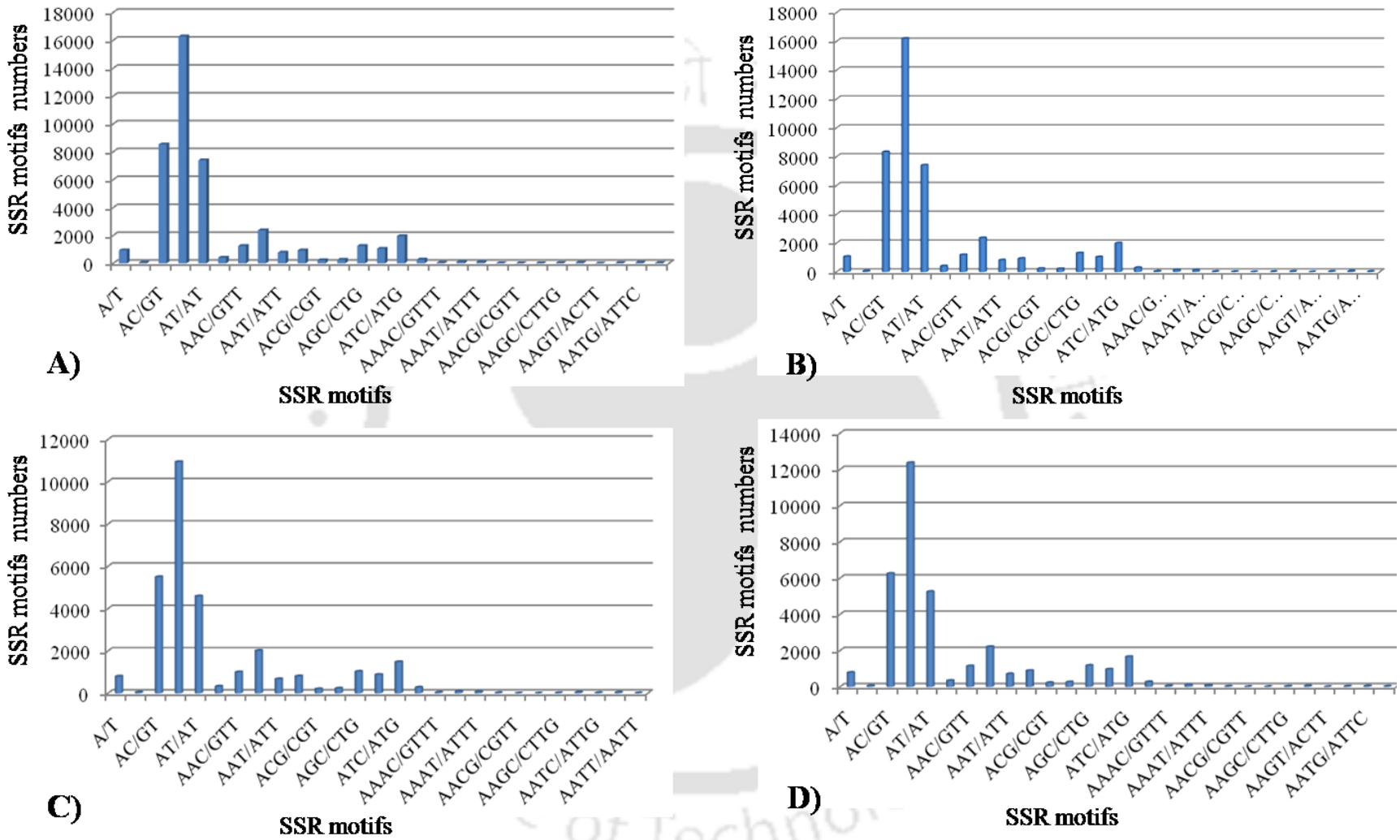


Figure 5.2. Characterisation of EST-SSR and their classification based on microsatellite repeats in selected four libraries and two organellar genome A) SRR349650, B) SRR349651, C) SRR349652, D) SRR349653, E) Chloroplast genome, F) Mitochondrial genome.



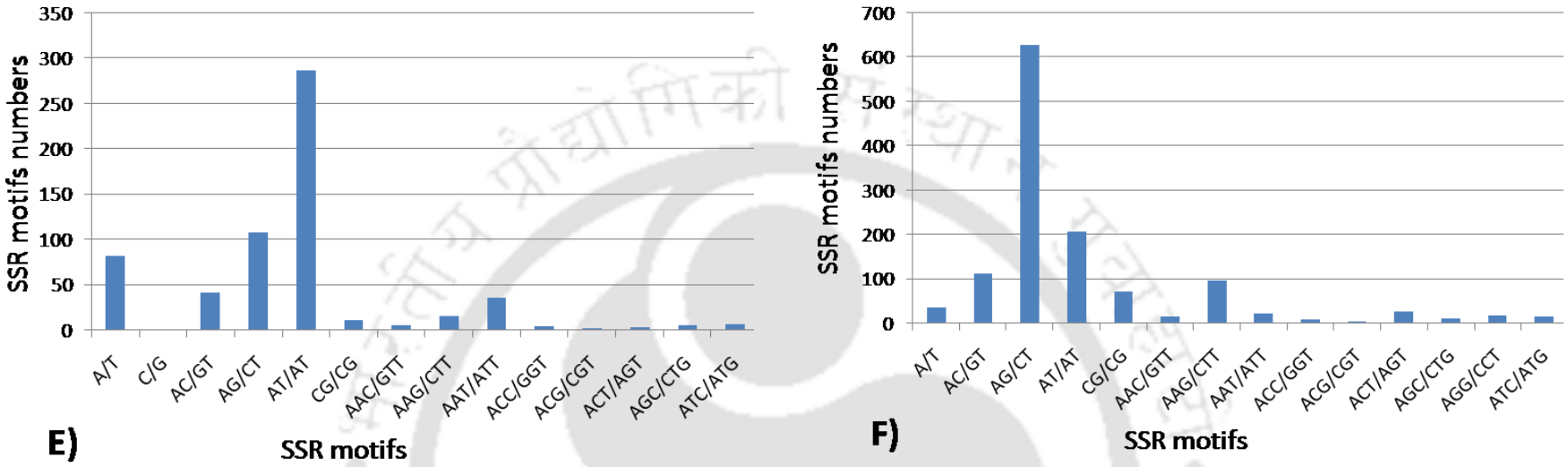
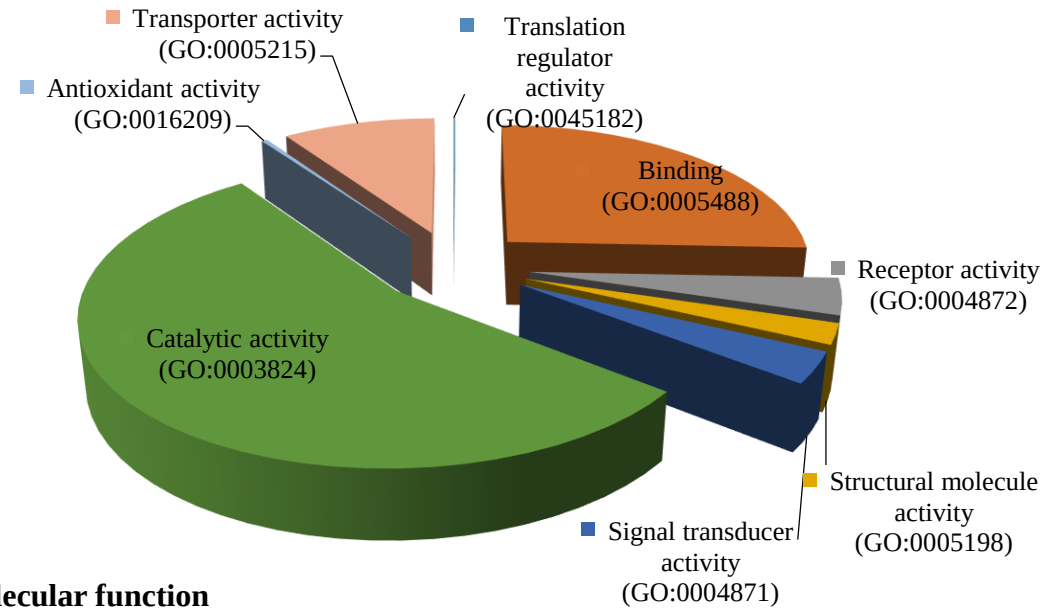


Figure 5.3. The frequency of identified EST-SSR motifs across four libraries and two organellar genomes A) SRR349650, B) SRR349651, C) SRR349652, D) SRR349653, E) Chloroplast genome, F) Mitochondrial genome.

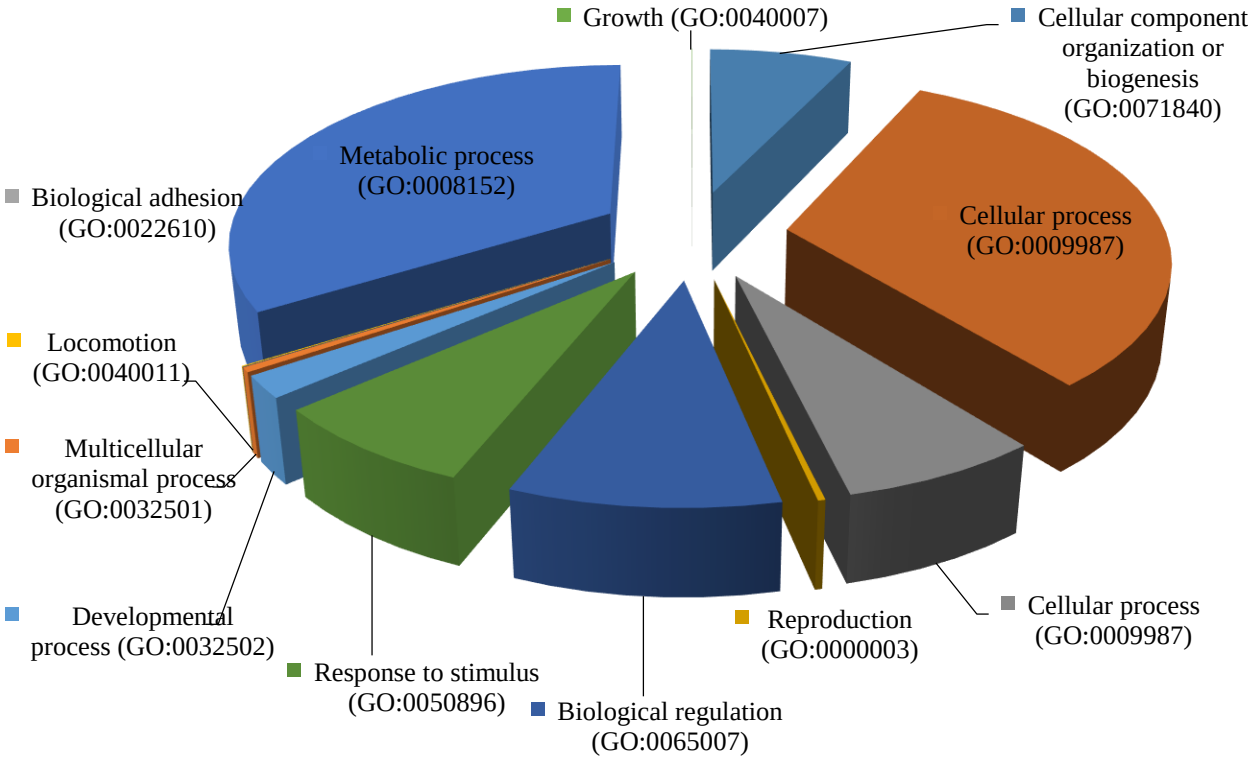
5.4.3. EST-SSR sequences annotation and marker validation

In order to assign the functional annotation, all unigene sequences having SSRs were examined by using Swissprot, Ensemble plants *G. max* database, Plaza 2.0 and Plaza 3.0. A total of 25,665 (12 %) unigene containing SSRs were identified as having significant similarity with known above mentioned databases. Similarly, Chen et al. (2015) also used functional annotation pipelines for the development and validation of EST-SSR molecular markers. Based on annotation, unigenes were assigned to gene ontology (GO) terms using the Panther program. SSR unigene sequences that assigned to the molecular function, biological process and cellular component clusters were classified into different terms (Fig 5.4). *Pongamia* EST-SSRs were assigned to various pathways of metabolic process. In molecular function category, sequences related to the catalytic activity (GO: 0003824) were high in number followed by Binding sequences (GO: 0005488) (Fig 5.4A). However, in the biological process section metabolic process (GO: 0008152) related sequences were highly abundant (Fig 5.4B). Furthermore, in the cellular component cluster, cell part (GO: 0044464) sequences were most abundant followed by organelle (GO: 0043226) component related sequences (Fig 5.4C). Since *Pongamia* is non-edible oil yielding crop, the research was focused on ESTs with particular relevance in seed metabolism. Particularly, unigenes involved in the biosynthesis of secondary metabolites and lipid including fatty acid and steroid biosynthesis. Further, the sequences were categorised into protein class. Distribution of GO term in protein class revealed that the maximum sequences were associated with hydrolase (PC00121), transferase (PC00220) and least in extracellular matrix protein (PC00102) (Fig 5.4D).

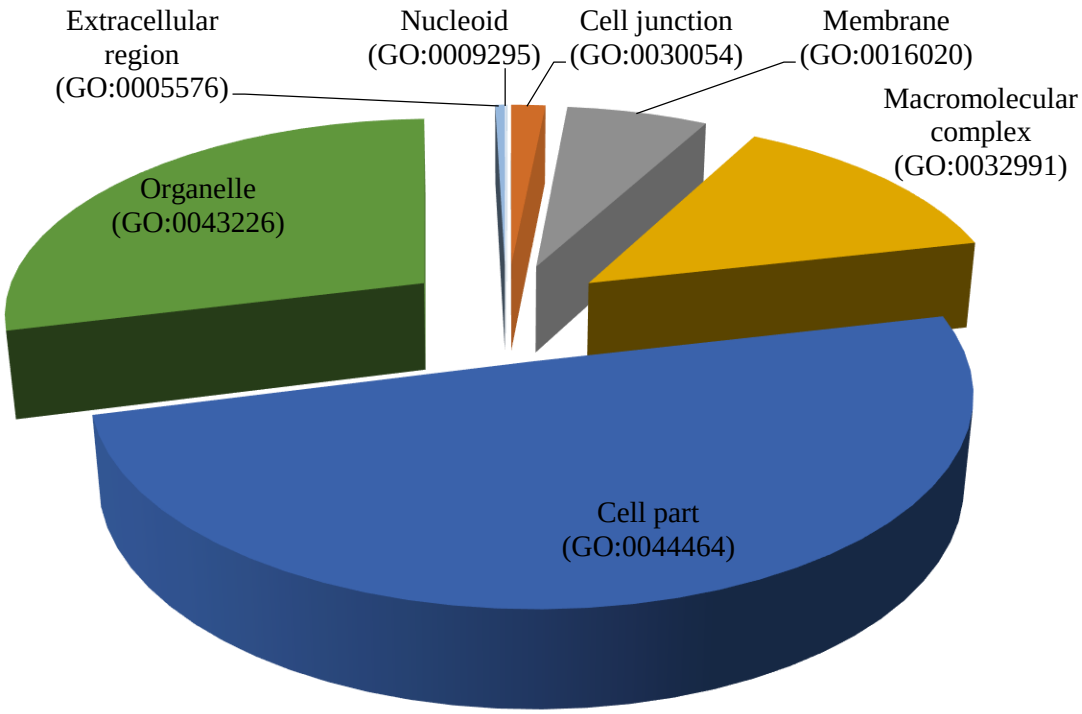
Unigene sequences containing EST-SSR motifs satisfying all the criteria were selected for marker development. Twenty ESTs having SSRs were chosen from the current study by their involvement in the various metabolic processes and secondary metabolites biosynthesis. Four primers were selected from an earlier publication (Huang et al., 2016). A total twenty-four primers synthesised for estimation of genetic diversity (Table 5.3). Sixteen out of twenty-four SSR primers showed amplification against fourteen accessions of *Pongamia.* Hence, sixteen primers were further employed for cross transferability study in different plants. The details of primer designing are mentioned in the methodology section.



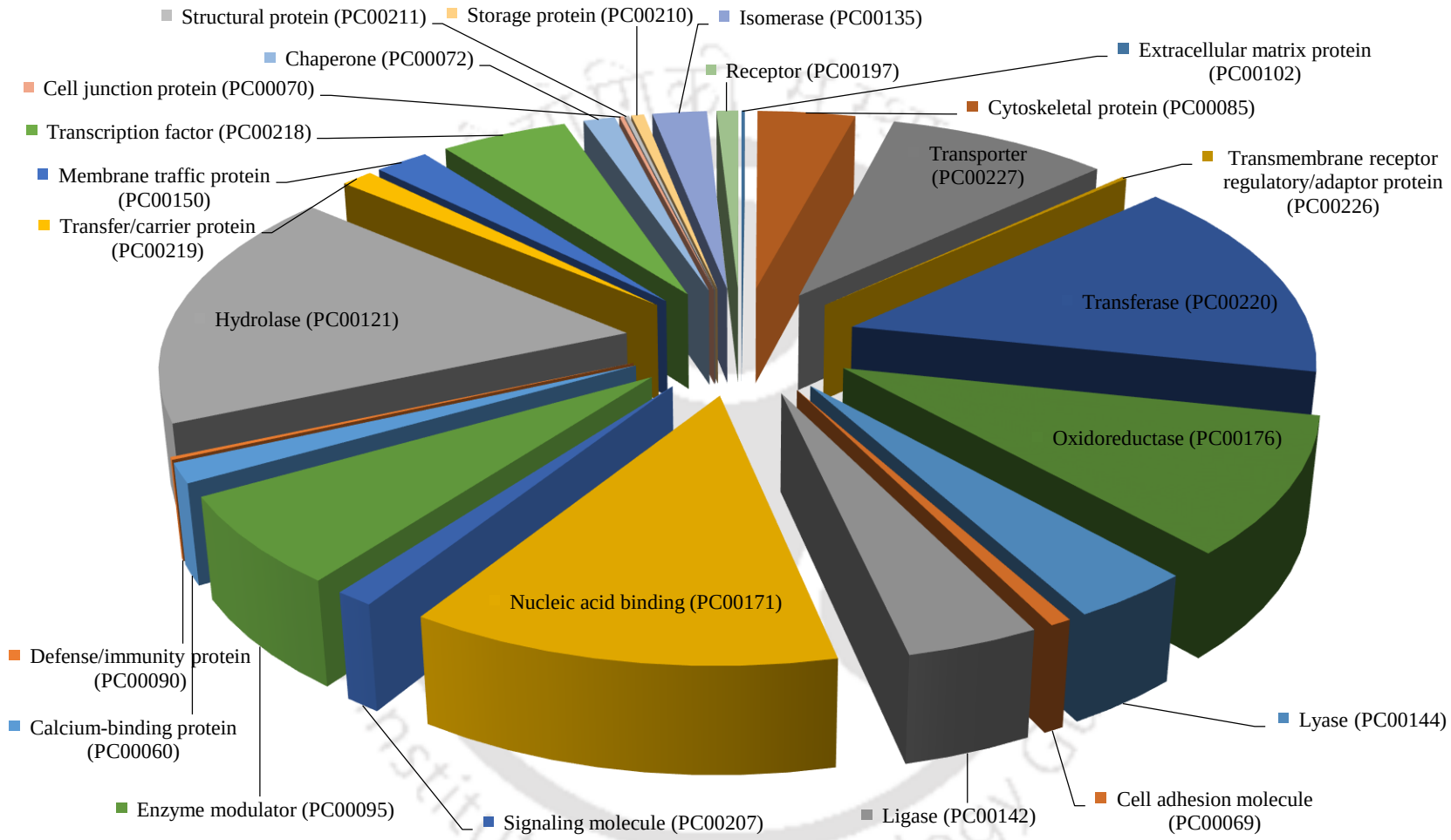
A) Molecular function



B) Biological process



C) Cellular component



D) Protein class

Figure 5.4. Details of GO terms (Pie chart) assigned to *Pongamia* EST-SSR. These charts represent the distribution of GO classified as a molecular function (A), biological process (B), cellular component (C), protein class (D).

Table 5.3. Description of EST-SSR markers developed from *Pongamia* unigenes.

S.N	Primer	Repeat motif	Primer sequence (5'-3')	Ta (^{0}C)	Product size (bp)	Description of putative function
1	MPM4	$(GCT)_5$	F: CTGTTATTGTTGGGGATGATTGT R:TGCAGCAACATCAGTAAGAGAAA	58	140–146	CDS (Huang et al. 2016)
2	MPM6	$(ATA)_8$	F: TTGCCAGCACTAGAGTTGTGTTA R: TCACCTGGACTAGAGATTTTCCA	59	137–143	CDS (Huang et al. 2016)
3	MPM46	$(AGA)_6$	R: CTTCCTCCCACTCTCTCATCTCT F:AACAACAGTGGAAGCAGACTCTC	60	159–163	3′UTR (Huang et al. 2016)
4	MPM51	$(GAATT)_4$	F: CTGACACAGCCTCTTCTTCATTT R: ACCAAACTCCATTCTTCAATCAA	58	144–154	5'UTR (Huang et al. 2016)
5	SSR11	$(TCT)_3$	F: TGCTGACTTGTTGTTTGGCG R: AGTGGCCTCAAGCTTGGTTT	60	282	*Probable sphingolipid transporter spinster homolog 2*
6	SSR12	$(GTT)_3$	F: ATCTTGCTGGAAGCTGGGAC	60	188	*Mevalonate pyrophosphate decarboxylase*

			R: AGGAGTCGCAGAATTGGGTG			
7	SSR13	$(TG)_5$	F: TGCGGAGAATGAAACGGAGA R: AGTGATTCTGGTCGCGGAAG	60	215	*CDP-diacylglycerol--inositol 3-phosphatidyltransferas*
8	SSR14	$(CTT)_4$	F: AAACCCTGGAGCAACTGCAT R: TTTCTTTGCACACAGGCTGC	59	207	*3-Oxoacyl-[acyl-carrier-protein reductase 4*
9	SSR15	$(TCA)_3$	F: AACCAAGAGCACTGGCAAGA R: CATGGCGGATCTCAAGTCCA	59	239	*Farnesyl diphosphate synthase*
10	SSR16	$(CA)_4$	F: TCATTTTTGTGCAGCACCGG R: GGAGGGGCCATAGGTTTCTG	59	196	*Acetate/butyrate-CoA ligase AAE7, peroxisomal*
11	SRR17	$(GAA)_3$	F: AGCGCAATTGCATTAAGGCG R: GCTGATTTAGTTGAGGCAGCG	59	260	*Putative peroxisomal acyl-coenzyme A oxidase 1.2*
12	SRR18	$(AAG)_3$	F: AATGGGAGGAGCTGCAATCC R: CTCTCCAGCAACAGCCATCA	60	193	*Probable 3-ketoacyl-CoA synthase 14*
13	SRR19	$(GA)_5$	F: TGCCTCCCCAGAAGATTGAG R: AAGGCTGGCACCTGTTTCAT	60	300	*Acyl-[acyl-carrier-protein] desaturase*

14	SRR20	$(TA)_6$	F: AACACTGGTTCTCTCGAGCG R: TGACAAGACGGAAAAGCCCA	59	253	*3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase*
15	SSR21	$(TC)_5$	F: AGCCTCCCCCTCCTTTCC R: AGAGTCTGGCGAGGGTGA	59	187	*Palmitoyl-acyl carrier protein thioesterase*
16	SRR22	$(CCA)_6$	F: TCCAGCCCCTCATAGCCC R: AGATCGGGTTCGCGACAC	59	203	*Oxysterol-binding protein-related protein 3C*
17	SRR23	$(GAGTCT)_4$	F: GCGTGGGGGCGGATATAT R: GGACTTCCCCATCCCTCCT	59	159	*Dehydrocholesterol reductase*
18	SRR24	$(CAAA)_3$	R: CCATGGACTCGCTCCCAC F: CACCCACCAACTGCTGCT	60	195	*15-cis-phytoene desaturase*
19	SRR25	$(CT)_{10}$	R: TGCCTTACCCAACTGCCA F: GGAAGAGGAGGGAGAGCCA	59	209	*Oxooacyl-coA reductase let-767*
20	SRR26	$(GAAGG)_3$	R: CGCGCTATCGGAGGAGAC F: CCTTTGTCTCTGTCGCTGC	58	250	*Lysophospholipid acyltransferase LPEAT1*
21	SRR27	$(TCC)_4$	F: TCCATGCTAAGCCCGCTG R: GTTCACGAGCCTGCTGGT	59	177	*O-acyltransferase WSD1*

22	SRR28	$(TTTA)_3$	F: TCAGCGACAATTGTGTGCT R: CCCCCAAGCCAGTGCATA	59	201	*Triacylglycerol lipase SDP1*
23	SRR29	$(TCA)_4$	F: CGGCGTTTGAAAGAGCGC R: GCTGGCTTGGAGGCTATGA	59	260	*Putative lipase*
24	SRR30	$(TA)_5$	F: GAGTGGTGTGGCACAGGG R: CTTCTCGCCGGTTTTCGC	59	274	*Flavonoid hydroxylase, Amorpha-4,11-diene 12-monooxygenase*

5.4.4. EST-SSR markers analysis

PCR amplification of EST-SSR marker was carried out using 24 primers as listed in Table 5.4. Of the 24 primers, 16 primers showed the reproducible and good quality banding patterns in *Pongamia* accessions. The sixteen EST-SRR primers produced a total of 650 fragments of size ranging from 50 bp to 700 bp and eight primers did not produce any band under different amplification conditions. This could be possible due to hybrid assembly, error in sequences or primer selection from the splice site at the exon-intron boundary (Dutta et al., 2011). Out of the 650 SSR loci, 238 (36.61%) bands were monomorphic, 388 (59.69%), bands were polymorphic, and 24 (3.69%) bands were unique in 14 *Pongamia* accessions. The highest number of bands were observed in SSR-30 (84 bands) followed by SSR-16 (69 bands) and 18 (68 bands). The average number of bands per primer was about 40, and the average number of polymorphic bands per primer was found to be 24.2. The highest number of polymorphic bands were found in SSR-18 (52 bands) and no polymorphic bands were observed in SSR-12, 26, 28. The highest number of monomorphic bands were found in SSR-19 (56 bands) followed by SSR-30 (42 bands). Of the total 24 unique bands, highest number (4) bands were observed in SSR-16 and SSR-23. The PIC values for EST-SSR markers varied between 0-0.90 for SSR-26 and SSR-21. The average PIC of EST-SSR markers, found to be 0.64 gives a high level of marker informativeness. Beside this, marker index ranged in between 0-88.8 (Table 5.4). PIC values are classified into three categories: high (PIC > 0.5), moderate (0.25 < PIC < 0.5) and low (PIC < 0.25) (Botstein et al., 1980), thus, the PIC for studied marker falls in high category (Botstein et al., 1980). The typical polymorphic and monomorphic EST-SSR fingerprinting using primer SSR-16 and SSR-23 are shown in Fig. 5.5 and 5.6.

SSR-13, 17, 24 and 29 primers did not produce any amplification product. This could be due to the presence of large intron sequence in the flanking region that leads to disruption of PCR extension. MPM-46, SSR-11, 14 and 15 primers produced faint bands with a smear. The occurrence of a smear in the gel is due to unspecific primer binding or redundancy of primer pair sequences to the target site. Some of the primers generated small size amplicon than expected size. These small size bands are nothing but the stutter bands or shadow bands. These bands are produced owing to the occurrence of replication slippage during PCR

amplification of microsatellite sequences. Stuttering can create confusion in the interpretation of bands profiling, leads to difficulty in size determination of PCR amplicon (Park et al. 2009).

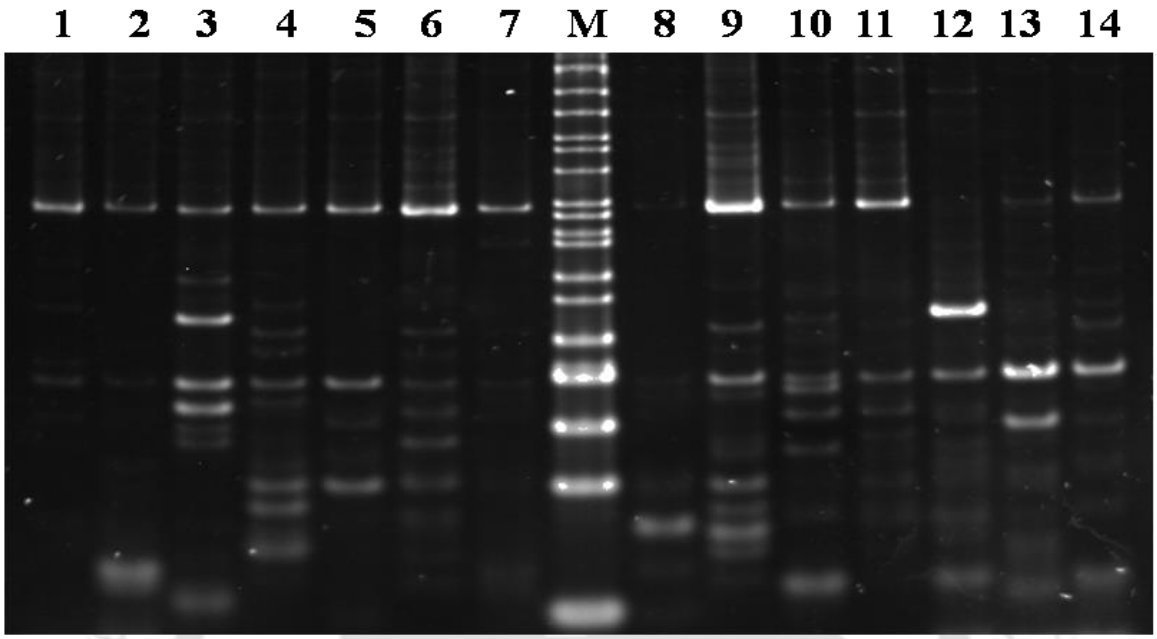


Figure 5.5. PCR amplification pattern with primer SSR-16 among different *Pongamia* accessions. M indicated the 50 bp ladder, the lower band - 50 bp. Orientation of *Pongamia* accessions on gel is according to Table 5.1.

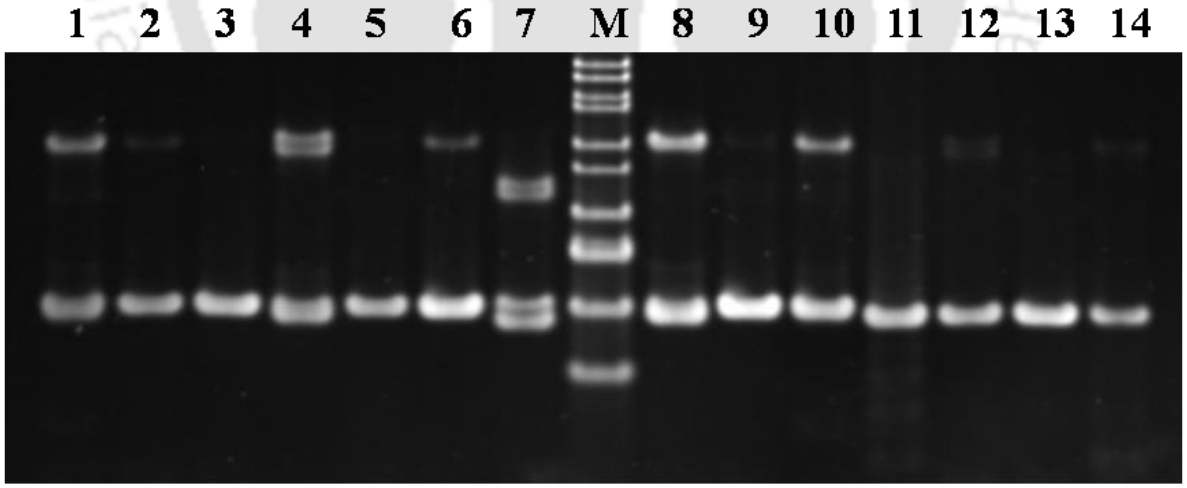


Figure 5.6. PCR amplification pattern with primer SSR-23 among different *Pongamia* accessions. M indicated the 50 bp ladder, the lower band - 100bp. Orientation of *Pongamia* accessions on gel is according to Table 4.1

Table 5.4. The degree of polymorphism and polymorphic information content (PIC) for EST-SSR primers applied to 14 accessions of *Pongamia.*

Primer code	Total number of bands	Number of polymorphic bands	POL (%)	PIC	MI
MPM-4	58	42	72.41379	0.8546	61.8848276
MPM-6	51	51	100	0.8004	80.04
MPM-51	44	29	65.90909	0.7324	48.2718182
SSR-12	28	0	0	0.5	0
SSR-16	69	51	73.91304	0.8905	65.8195652
SSR-18	68	52	76.47059	0.8762	67.0035294
SSR-19	61	4	6.557377	0.7868	5.15934426
SSR-20	14	13	92.85714	0.5408	50.2171429
SSR-21	51	50	98.03922	0.9065	88.872549
SSR-22	19	5	26.31579	0.3878	10.2052632
SSR-23	27	9	33.33333	0.6145	20.4833333
SSR-25	31	28	90.32258	0.8470	76.5032258
SSR-26	14	0	0	0	0
SSR-27	16	15	93.75	0.539	50.53125
SSR-28	15	0	0	0.1244	0
SSR-30	84	39	46.42857	0.8591	39.8867857
Total	**650**	**388**			
Mean	**40.625**	**24.25**	**54.769**	**0.6412**	**41.554**
Range	**70**	**52**	**100**	**0.9065**	**88.87**

POL - Polymorphism; PIC - Average polymorphic information content for polymorphic bands; MI – Marker index = POL (%) x PIC

5.4.5. Genetic diversity analysis by EST-SSRs marker

In order to evaluate the potential of EST-SSRs, the genetic analysis was done among 14 *Pongamia* accessions collected from a different geographical location of India. Sixteen EST-SSR primers were opted for cluster analysis through UPGMA dendrogram constructed using SHAAN neighbour-joining separately. The genetic distance among the fourteen accessions of *Pongamia* was determined by Dice genetic distance (Dice, 1945). The genetic similarity of coefficient varied from 0.43 to 0.76 with an average value of 0.58, which was almost similar to the earlier conducted study in *M. piperita* (Kumar et al., 2015). The UPGMA clustering fell into four major groups at the similarity index value of 0.43 (Fig 5.7). *Pongamia* accessions namely CPT-29 and PP-8 were at the extreme end of the dendrogram. Group-I consists of seven accessions, group-II encompassing five *Pongamia* accessions, group-III contain only one accession PP-3. Finally, group-IV comprising one accession namely PP-8 (Fig 5.7). Within group-I, two subgroups were present, subgroup-I containing CPT-29, PP-1, PP-2, PP-4 and NGPP-46, while other subgroup-II comprised of PP-5 and PP-7. Group-II also divided into two subclusters namely, subgroup-I having PP-6 and PP-9 while other subgroup consists of PP-10, PP-11 and PP-12. The present diversity study showed the presence of genetic variation among the *Pongamia* accessions collected from different locations of India. It is clear from the dendrogram that the clusters did not correctly form on the basis of a geographical location of samples. With an exception, some accessions belong to the same geographical region were grouped together by the EST-SSRs, such as PP-11 and P-12. The maximum similarity was found between the PP-11 and PP-12 (0.76). The lowest similarity or highest dissimilarity was found between PP-8 and PP-11 (0.34). However, the presence of a polymorphism in the EST–SSRs of *Pongamia* accessions suggested the importance of these markers in future for genetic studies.

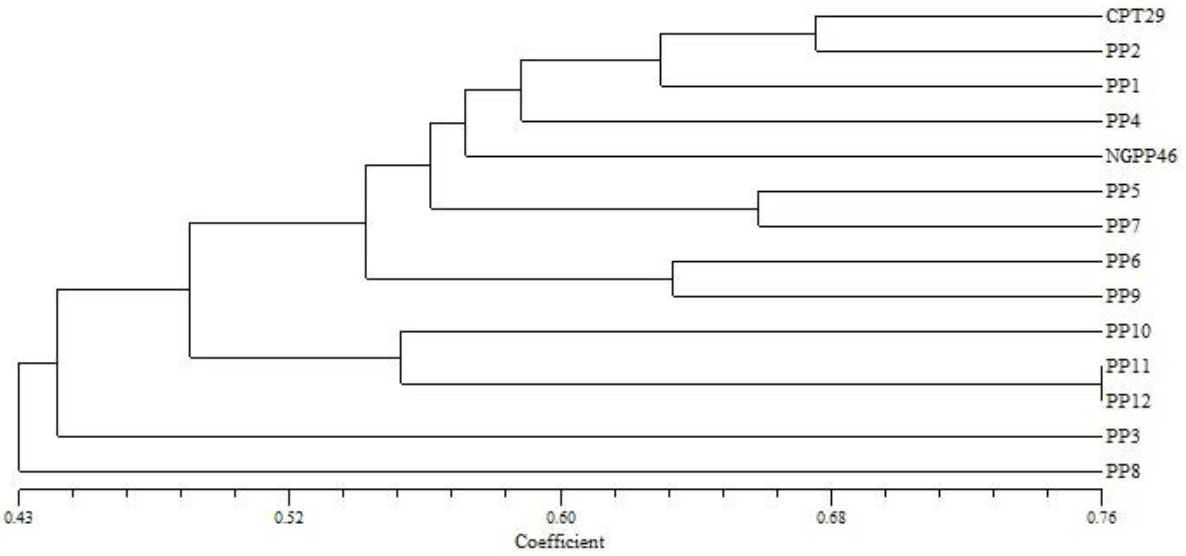


Figure 5.7. Phenogram representing the phylogenetic relationship among fourteen accessions of *Pongamia* determined by UPGMA cluster analysis. The genetic distances were estimated from the Dice similarity coefficient.

5.4.6. Transferability of *Pongamia* EST-SSR markers in different plants

The potentials of EST-SSR primers were examined for cross transferability among the twelve different monocot and dicot plants. All 16 EST-SSRs marker showed clear amplifications in almost all selected plants (Fig. 5.8). The estimated cross transferability was found to be 56.25 % in *J. curcas*, 87.5% in *R. communis*, 62.5% in *M. Ferrea*, 93.75% in *G. max*, 75% in *C. arietinum*, 93.75% in *A. hypogaea*, 100% in *V. radiate*, 81.25% in *O. sativa*, 81.25% in *M. acuminata*, 75% in *C. longa*, 81.25% in *S. melongena*, 75% in *P. vulgaris* (Table 5.5). The SSR markers namely SSR-16, 19, 20, 25, 27 and 30 primers were regarded as highly polymorphic among the tested plant species. These results illustrate the high transferability rate of EST-SSR across the species. This is in agreement with the previous investigation where a significant amount of cross-species transferability was reported (Thiel et al., 2003, Zhou et al., 2016). A high rate of transferability of EST-SSR between the species is due to their presence in genic regions which are regarded as conserved across the species. However, the transferability rate between monocot and dicot in the present study is more than 75%. Whereas in the previous report, Savadi et al. (2012) showed the transferability of SSR motifs was about 39% between peanut (dicot) and sorghum (monocot).

Transferability of EST-SSRs has been conducted in many plants, such as sugarcane, citrus, and cassava (Cordeiro et al., 2001, Luro et al., 2008, Raji et al., 2009). SSR-26 did not produce any band expect in *V. radiate*, which was least among all tested SSR marker. This happened probably due to the variation in target binding sequences. Some of the primers generated clear amplifications with both expected and unexpected sizes, due to the presence of intron sequence between forward and reverse primer. This indicates that transferability genic SSR markers can be helpful in phylogenetic studies in different genera and their possible utilisation in a comparative mapping of genes among closely related species. Nevertheless, the success of PCR amplification differed between various organism, higher or lower amplification rate depending upon the genetic similarity existing between the organisms. Additionally, genic SSR also have a higher level of probability of being linked with economically important traits which are controlled by quantitative trait loci (QTLs). Hence these markers could be highly useful in an investigation involving QTL mapping and marker-

assisted selection (MAS). Therefore, SSRs retrieved from genic of the genome are believed to be conserved and highly transferable across taxa.

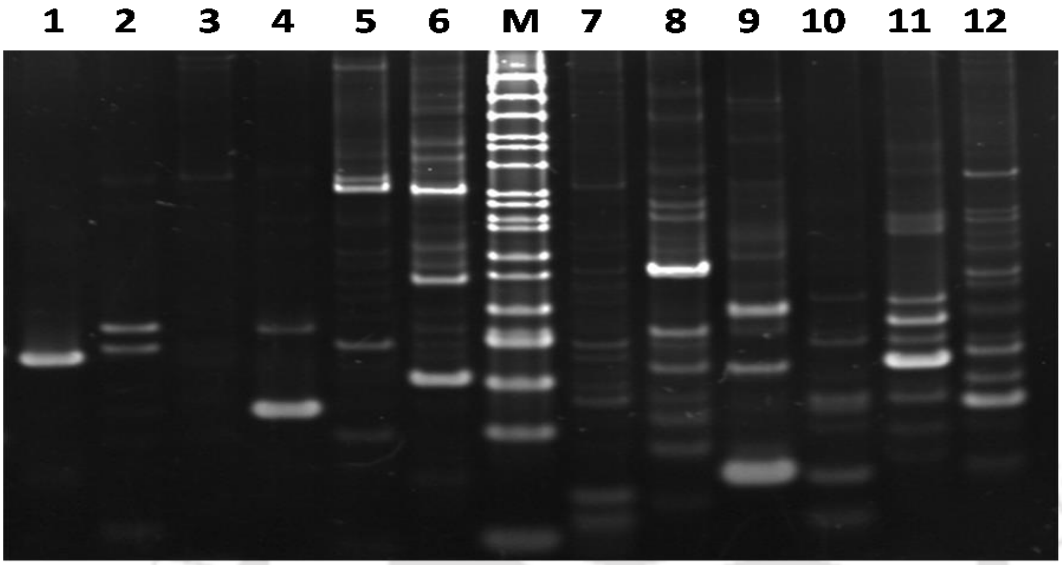


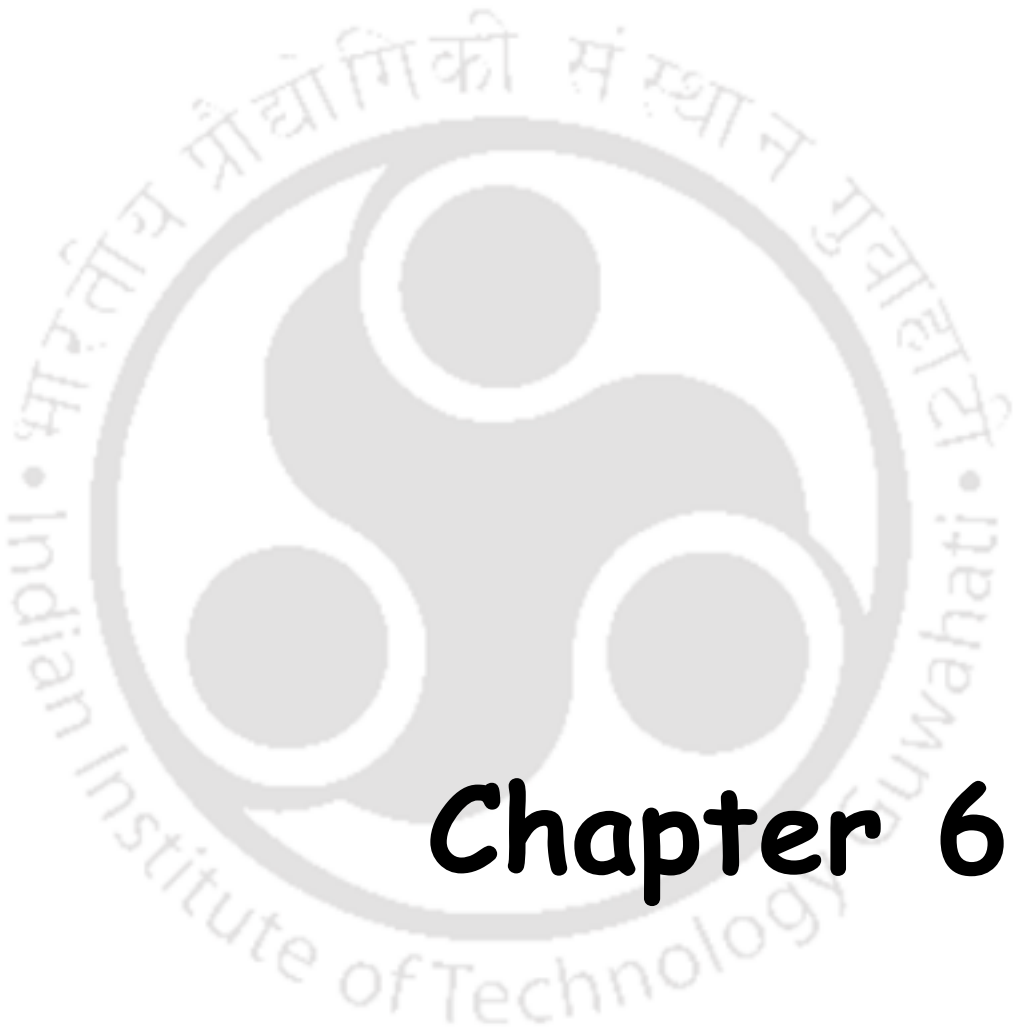
Figure 5.8. PCR amplification pattern with primer SSR-16 among different plant species 1. *Jatropha curcas,* 2. *Ricinus communis*, 3. *Mesua Ferrea,* 4. *Glycine max*, 5. *Cicer arietinum,* 6. *Arachis hypogaea*, 7. *Vigna radiata,* 8. *Oryza sativa*, 9. *Musa acuminata*, 10. *Curcuma longa*, 11. *Solanum melongena*, 12. *Phaseolus vulgaris.* M indicated the 50 bp ladder. Lower band - 50

Table 5.5. Estimation of cross transferability of 16 EST-SSR primers in twelve different plants. Numbers represent bands produced with primer in different plants mentioned at the upper row.

S.N /Lane	J. curcas	R. communis	M. Ferrea	G. max	C, arietinum	A. hypogaea	V. radiata	O. sativa	M. acuminata	C. longa	S. melongena	P. vulgaris	Primer Polymorphism (%)
SSR-1	1	1	1	2	2	6	2	3	1	0	5	2	91.66
SSR-2	1	1	0	1	2	1	1	1	1	1	2	1	91.66
SSR-4	2	1	1	3	1	2	2	2	5	4	0	3	91.66
SSR-12	0	1	0	1	0	2	2	0	0	0	1	2	50
SSR-16	1	3	1	3	5	7	7	9	9	6	9	9	100
SSR-18	0	0	3	6	4	7	5	9	5	4	3	1	83.33
SSR-19	7	5	6	5	4	6	4	7	6	7	6	7	100
SSR-20	3	3	2	2	4	3	2	3	4	2	3	2	100
SSR-21	0	5	0	4	6	4	2	4	0	2	2	0	66.66
SSR-22	0	1	0	1	0	1	1	1	1	1	0	0	58.33
SSR-23	0	2	3	3	2	3	2	3	2	1	2	3	91.66
SSR-25	3	3	2	2	2	3	4	5	5	3	2	3	100
SSR-26	0	0	0	0	0	0	1	0	0	0	0	0	8.33
SSR-27	4	4	3	6	4	7	1	8	9	9	9	9	100
SSR-28	0	1	0	1	0	1	2	0	4	0	2	0	50
SSR-30	6	6	5	9	7	6	4	5	7	3	4	5	100
Transferability (%)	56.25	87.5	62.5	93.2	75	93.25	100	81.25	81.25	75	81.25	75	

4.5. Conclusion

The EST-SSR developed in the present study will help to increase DNA sequence resources in *Pongamia*, which were previously very minimum**.** Most of the nucleotide sequences deposited in past at public databases are a DNA sequence which is commonly targeted in phylogenetic analyses, e.g., retrotransposons and matK etc. This chapter deals with *in silico* extraction of microsatellite from the assembled transcriptome data and publically available organellar genomes. The *in silico* or computational approach not only save the cost and time but also provides a sufficient amount of microsatellite for future marker development. The information obtained from the microsatellite mining helps to better understand the pattern and nature of SSRs in *Pongamia* genome. This could act as an important resource for designing future genomic studies in this crop. The functional classification of EST-SSRs sequences through ontology revealed the occurrence of microsatellites in protein-coding genes with different biology, cellular and molecular function. In search of a modern marker system, transcriptome sequences were employed for the development of economically important trait linked markers in *Pongamia*. The functional annotation of markers was allowed to identify the gene linked markers. The diversity study was carried out using EST- SSR markers in 14 diploid *Pongamia* accessions collected from the different geographical zone of India. High levels of allelic and genetic diversity were found in some EST-SSR markers. These results could be useful for a breeder for exploiting variation in a wild population. EST–SSR markers revealed a high level of polymorphism and were successfully transferable across the different plant species. Furthermore, transferability study provides an efficient way to conduct the genetic studies in a plant where the genomic or expression libraries are not available. In addition, transferability studies also help in reducing the cost and timing of genic primer development. The knowledge of the distribution and patterns of microsatellite in a genic region may help us to understand its role in gene expression, regulation and evolutionary properties. Further understanding of genic SSR motifs length and polymorphism distribution in *Pongamia* could be helpful to evaluate the possible mutational effects of SSR on genic regions. In conclusion, the genic SSR markers developed in the current study could be important in future for biodiversity, taxonomy, molecular breeding and genetic studies in *Pongamia.*



Chapter 6

Chapter 6

Summary

The present investigation entitled **"Mining of repeat elements from *Pongamia* for marker development."** was undertaken with the following objectives:

1. Isolation and characterisation of retrotransposons from *Pongamia* genome

2. Role of transposable elements in *Pongamia* unigene diversity

3. Development of EST-SSR marker in *Pongamia*

The summary and conclusion of the research work conducted during the present investigation are summarised below:

The *Pongamia* is a tree species belonging to the Fabaceae family. Due to depleting fossils fuel resources and increasing global warming, researchers are focusing on finding oil-bearing plants which can produce non-edible oils as the feedstock for biodiesel production. *Pongamia* is an oleaginous multipurpose nitrogen-fixing tree. It has caught public attention because of its high oil content seeds and adaptability to different agro-climatic conditions. For improvement of any plant species, phenotypic and genotypic studies are very important. Hence, in the present study, we worked on the mining of different repetitive element present in the *Pongamia* genome for marker development.

In the **first objective**, retrotransposons elements like Ty1-*copia*, Ty3-*gypsy* and LINEs were isolated from *Pongamia* genome using PCR methodology. TEs were also mined from the organelle genome. The distribution of TE superfamilies varied in chloroplast and the mitochondrial genome. A very little number of TEs were detected in organellar genes. The isolated retrotransposons harboured various heterogenous lineages. Retrotransposons copy number were estimated through dot blot hybridisation. Further, the transcriptome sequences were assembled through Trinity assembler using raw Illumina RNA-Seq libraries. However, the results of TE ESTs profiling revealed the presence of different transcriptionally active superfamilies of TEs in *Pongamia* genome. Moreover, *in silico* study showed the increase in transcriptional activity of TEs in response to salt stress.

In the **second objective**, we tried to understand the potential of TEs in the evolution of *Pongamia* unigenes. A total of 1290 *Pongamia* unigene containing TE cassettes were identified as having significant similarity with known protein databases. A majority of unigenes were harboured by LTR-retrotransposons fragments. Among these proteins, we have shown the contribution of the *gypsy* element in PDC protein diversity. Analysis revealed that the presence of the *gypsy*-like retrotransposon is only present in higher plants and absent in bacterial, fungal and bryophyte genes. It is possible that the insertion of a *gypsy* element in PDC might have happened before the divergence of monocot and dicot plants. The analysis of PDC protein sequences showed that the presence of *gypsy* at the protein level. We also tried to investigate the presence of TEs in organellar genes but did not find any insertion.

In the **third objective**, SSRs were successfully isolated from transcriptome databases. A total of 25,665 unigene containing SSRs were detected as having significant similarity with known protein databases. Further, the ESTs having SSRs were chosen based on their involvement in the various metabolic processes and secondary metabolite biosynthesis. Sixteen primers were successfully amplified in fourteen *Pongamia* accessions as well as plants belonging to different families. Similarly, SSRs were also isolated from the organelle genome. Among the SSRs, dinucleotides were abundantly present.

The existing work presented here and in the literature, has unearthed, rapid and substantial diversity present in repetitive elements in *Pongamia* genome. The recent work has combined both computational and wet lab studies. The results obtained from the present investigation can be used for isolation of full-length active TEs. Further, understanding of genic SSR motif length and polymorphism distribution in *Pongamia* could be helpful to evaluate the possible mutational effects of SSR in genic regions. The active elements and genic SSR markers detected in the current study could be important in future for biodiversity, epigenetic, molecular breeding and genetic studies in *Pongamia.* Although the outcome of some of the investigations is enlightening, we still have very little understanding of the underlying mechanisms. To comprehend more about TEs and genes relationship, extensive studies and experiments are required in the future to further validate this phenomenon. In long-term, such investigations will not be only helpful for

the development of a gene-specific marker for oil traits in biofuel crops but also assist in enhancing our capability to undertake larger breeding programs in biodiesel crops.

Future scope

Potential research that can be explored on the basis of present investigations are;

- Transcriptome libraries can be utilised for the preparation of SNP markers.
- Isolated RT can be used to isolate full-length TEs or LTR. These sequences will help in designing of inter-retrotransposons amplified polymorphisms (IRAPs) and retrotransposon-microsatellite amplified polymorphism (REMAP).
- Development of sequence characterised amplified region (SCAR) marker.
- The finding from this investigation provides an opportunity to determine the physical position of TEs on chromosome through fluorescent *in situ* hybridisation (FISH).
- ESTs can be employed to determine the biosynthetic pathways of important metabolites like karanjin, terpenoid pathway, alkaloid pathway and phenylpropanoid pathway.
- Identification of different transcription factors (TFs) from tissue-specific libraries.
- Comparative profiling of gene expression.
- Identification and characterisation of miRNA.
- Transposons mediated epigenetic studies.
- Transposons transformation studies in plant.
- Preparation of genomic libraries.
- Isolation of full-length active TEs.
- Development of Repeat-Join marker
- Identification of different TEs population in the genome.

References

References

AGARWAL, G., JHANWAR, S., PRIYA, P., SINGH, V. K., SAXENA, M. S., PARIDA, S. K., GARG, R., TYAGI, A. K. & JAIN, M. 2012. Comparative analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. ***PLoS One***, 7**,** e52443.

AGARWAL, M., SHRIVASTAVA, N. & PADH, H. 2008. Advances in molecular marker techniques and their applications in plant sciences. ***Plant Cell Reports***, 27**,** 617-31.

AGGARWAL, R. K., HENDRE, P. S., VARSHNEY, R. K., BHAT, P. R., KRISHNAKUMAR, V. & SINGH, L. 2006. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. ***Theoretical and Applied Genetics***, 114**,** 114:359.

AHMAD, M., ZAFAR, M., KHAN, M. A. & SULTANA, S. 2009. Biodiesel from *Pongamia pinnata* l. oil: A promising alternative bioenergy source. ***Energy Sources, Part A: Recovery, Utilization, and Environmental Effects***, 31**,** 1436-1442.

AHMED, S., SHAFIUDDIN, M. D., AZAM, M. S., ISLAM, M. S., GHOSH, A. & KHAN, H. 2011. Identification and characterization of jute LTR retrotransposons:: Their abundance, heterogeneity and transcriptional activity. ***Mobile Genetic Elements***, 1**,** 18-28.

AKASH, M. W. & MYERS, G. O. 2012. The development of faba bean expressed sequence tag–simple sequence repeats (EST-SSRs) and their validity in diversity analysis. ***Plant Breeding***, 131, 522-530.

ALIPOUR, A., TSUCHIMOTO, S., SAKAI, H., OHMIDO, N. & FUKUI, K. 2013. Structural characterization of copia-type retrotransposons leads to insights into the marker development in a biofuel crop, *Jatropha curcas* L. ***Biotechnology for Biofuels***, 6**,** 129-129.

ALIX, K. & HESLOP-HARRISON, J. S. 2004. The diversity of retroelements in diploid and allotetraploid Brassica species. ***Plant Molecular Biology***, 54**,** 895-909.

ALMEIDA, L. M., SILVA, I. T., SILVA, W. A., JR., CASTRO, J. P., RIGGS, P. K., CARARETO, C. M. & AMARAL, M. E. 2007. The contribution of transposable elements to *Bos taurus* gene structure. ***Gene***, 390**,** 180-9.

ALTINKUT, A., RASKINA, O., NEVO, E. & BELYAYEV, A. 2006. *En/Spm*-like transposons in Poaceae species: Transposase sequence variability and chromosomal distribution. ***Cellular & Molecular Biology Letters***, 11, 214-229.

ALVERSON, A. J., RICE, D. W., DICKINSON, S., BARRY, K. & PALMER, J. D. 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. ***Plant Cell***, 23**,** 2499-513.

ALVERSON, A. J., WEI, X., RICE, D. W., STERN, D. B., BARRY, K. & PALMER, J. D. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (cucurbitaceae). ***Molecular Biology and Evolution***, 27**,** 1436-1448.

AMINAH, A., SUPRIYANTO, SURYANI, A. & SIREGAR, I. Z. 2017. Genetic diversity of *Pongamia pinnata* (*Millettia pinnata*, aka. malapari) populations in Java Island, Indonesia. ***Biodiversitas***, 18**,** 6.

AMMIRAJU, J. S. S., LUO, M., GOICOECHEA, J. L., WANG, W., KUDRNA, D., MUELLER, C., TALAG, J., KIM, H., SISNEROS, N. B., BLACKMON, B., FANG, E., TOMKINS, J. B., BRAR, D., MACKILL, D., MCCOUCH, S., KURATA, N., LAMBERT, G., GALBRAITH, D. W., ARUMUGANATHAN, K., RAO, K., WALLING, J. G., GILL, N., YU, Y., SANMIGUEL, P., SODERLUND, C., JACKSON, S. & WING, R. A. 2006. The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. ***Genome Research***, 16, 140-7.

ARAUJO, P. G., ROSSI, M., JESUS, E. M., SACCARO, N. L., KAJIHARA, D., MASSA, R. R., FELIX, J. M., DRUMMOND, R. D., FALCO, M. C., CHABREGAS, S. M., ULIAN, E. N. C., MENOSSI, M. & SLUYS, M. A. V. 2005. Transcriptionally active transposable elements in recent hybrid sugarcane. ***The Plant Journal***, 44**,** 707-717.

ARAYA, S., MARTINS, A. M., JUNQUEIRA, N. T. V., COSTA, A. M., FALEIRO, F. C. B. G. & FERREIRA, M. C. R. E. 2017. Microsatellite marker development by partial sequencing of the sour passion fruit genome (*Passiflora edulis* Sims). ***BMC Genomics***, 18**,** 549.

ARCHAK, S., GAIKWAD, A. B., GAUTAM, D., RAO, E. V. V. B., SWAMY, K. R. M. & KARIHALOO, J. L. 2003. DNA fingerprinting of Indian cashew (*Anacardium occidentale* L.) varieties using RAPD and ISSR techniques. ***Euphytica***, 130**,** 397-404.

ASHKANI, S., RAFII, M. Y., RUSLI, I., SARIAH, M., ABDULLAH, S. N. A., ABDUL RAHIM, H. & LATIF, M. A. 2012. SSRs for marker-assisted selection for blast resistance in rice (Oryza sativa L.). ***Plant Molecular Biology Reporter***, 30**,** 79-86.

AYARPADIKANNAN, S. & KIM, H. S. 2014. The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. ***Genomics & Informatics***, 12**,** 98-104.

BACCI JR, M. C., SOARES, R. B. S., TAJARA, E. Z., AMBAR, G., FISCHER, C. N., GUILHERME, I. R., COSTA, E. P. & MIRANDA, V. F. O. 2005. Identification and frequency of transposable elements in *Eucalyptus*. ***Genetics and Molecular Biology***, 28**,** 634-639.

BAO, W., KOJIMA, K. K. & KOHANY, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. ***Mobile DNA***, 6, 11.

BARBAGLIA, A. M., KLUSMAN, K. M., HIGGINS, J., SHAW, J. R., HANNAH, L. C. & LAL, S. K. 2012. Gene capture by Helitron transposons reshuffles the transcriptome of maize. ***Genetics***, 190**,** 965-75.

BARRET, P., BRINKMAN, M. & BECKERT, M. 2006. A sequence related to rice Pong transposable element displays transcriptional activation by in vitro culture and reveals somaclonal variations in maize. ***Genome***, 49**,** 1399-1407.

BAUCOM, R. S., ESTILL, J. C., LEEBENS-MACK, J. & BENNETZEN, J. L. 2009. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. ***Genome Research***, 2, 243-54.

BECKER, H. A., SAEDLER, H. & LONNIG, W. E. 2001. Transposable Elements in Plants. *In:* MILLER, J. H. (ed.) ***Encyclopedia of Genetics***. New York: Academic Press.

BEIER, S., THIEL, T., MC, SCHOLZ, U. & MASCHER, M. 2017. MISA-web: A web server for microsatellite prediction. ***Bioinformatics***, 33**,** 2583-2585.

BEN-DAVID, S., YAAKOV, B. & KASHKUSH, K. 2013. Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat. ***The Plant Journal***, 76, 201-210.

BENNETZEN, J. L. & WANG, H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. ***Annual Review of Plant Biology***, 65**,** 505-530.

BENSON, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. ***Nucleic Acids Reserch***, 27**,** 573-80.

BHATTACHARYYA, M. K., SMITH, A. M., ELLIS, T. H. N., HEDLEY, C. & MARTIN, C. 1990. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. ***Cell***, 60**,** 115-122.

BIRKY, C. W. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. ***Proceedings of the National Academy of Sciences of the United States of America***, 92**,** 11331-8.

BOBADE, S. N. & KHYADE, V. B. 2012. Preparation of methyl ester (biodiesel) from karanja (*Pongamia pinnata*) oil. ***Research Journal of Chemical Sciences***, 2**,** 43-50.

BORODULINA, O. R. & KRAMEROV, D. A. 1999. Wide distribution of short interspersed elements among eukaryotic genomes. ***FEBS Letters***, 457**,** 409-13.

BOSAMIA, T. C., MISHRA, G. P., THANKAPPAN, R. & DOBARIA, J. R. 2015. Novel and stress relevant est derived ssr markers developed and validated in peanut. ***PLoS One***, 10**,** e0129127.

BOTSTEIN, D., WHITE, R. L., SKOLNICK, M. & DAVIS, R. W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. ***American Journal of Human Genetics***, 32**,** 314-31.

BOURQUE, G., LEONG, B., VEGA, V. B., CHEN, X., LEE, Y. L., SRINIVASAN, K. G., CHEW, J. L., RUAN, Y., WEI, C. L., NG, H. H. & LIU, E. T. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. ***Genome Research***, 18**,** 1752-1762.

BRANDSTROM, M., BAGSHAW, A. T., GEMMELL, N. J. & ELLEGREN, H. 2008. The relationship between microsatellite polymorphism and recombination hot spots in the human genome. ***Molecular Biology and Evolution***, 25**,** 2579-2587.

BRANDT, J., SCHRAUTH, S., VEITH, A. M., FROSCHAUER, A., HANEKE, T., SCHULTHEIS, C., GESSLER, M., LEIMEISTER, C. & VOLFF, J. N. 2005. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. ***Gene***, 345**,** 101-11.

BRITTEN, R. 2006. Transposable elements have contributed to thousands of human proteins. ***Proceedings of the National Academy of Sciences of the United States of America***, 103**,** 1798.

BROSIUS, J. & GOULD, S. J. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". ***Proceedings of the National Academy of Sciences of the United States of America***, 89**,** 10706.

BROWNELL, E., MITTEREDER, N. & RICE, N. R. 1989. A human rel proto-oncogene cDNA containing an Alu fragment as a potential coding exon. ***Oncogene***, 4**,** 935-942.

BUNDOCK, P. & HOOYKAAS, P. 2005. An Arabidopsis *hAT*-like transposase is essential for plant development. ***Nature***, 436**,** 282.

BUTELLI, E., LICCIARDELLO, C., ZHANG, Y., LIU, J., MACKAY, S., BAILEY, P., REFORGIATO-RECUPERO, G. & MARTIN, C. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. ***The Plant Cell***, 24**,** 1242-1255.

CAO, Y., JIANG, Y., DING, M., HE, S., ZHANG, H., LIN, L. & RONG, J. 2015. Molecular characterization of a transcriptionally active Ty1/*copia*-like retrotransposon in *Gossypium*. ***Plant Cell Reports.***, 6, 1037-47.

CASACUBERTA, J. M., JACKSON, S., PANAUD, O., PURUGGANAN, M. & WENDEL, J. 2016. Evolution of Plant Phenotypes, from Genomes to Traits. ***G3: Genes|Genomes|Genetics***, 6**,** 775.

CASTELO, A. T., MARTINS, W. & GAO, G. R. 2002. TROLL--tandem repeat occurrence locator. ***Bioinformatics***, 18, 634-6.

CAVRAK, V. V., LETTNER, N., JAMGE, S., KOSAREWICZ, A., BAYER, L. M. & SCHEID, O. M. 2014. How a retrotransposon exploits the plant's heat stress response for its activation. ***Plos Genetics***, 10, e1004115..

CHADHA, S. & GOPALAKRISHNA, T. 2005. Retrotransposon-microsatellite amplified polymorphism (REMAP) markers for genetic diversity assessment of the rice blast pathogen (*Magnaporthe grisea)*. ***Genome***, 48**,** 943-945.

CHALOPIN, D., NAVILLE, M., PLARD, F., GALIANA, D. & VOLFF, J. N. 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. ***Genome Biology and Evolution***, 7**,** 567-580.

CHEN, H., LIU, L., WANG, L., WANG, S., SOMTA, P. & CHENG, X. 2015. Development and validation of EST-SSR markers from the transcriptome of adzuki bean (*Vigna angularis*). ***PLoS One***, 10**,** e0131939.

CHEN, J., HUANG, Q., GAO, D., WANG, J., LANG, Y., LIU, T., LI, B., BAI, Z., LUIS GOICOECHEA, J., LIANG, C., CHEN, C., ZHANG, W., SUN, S., LIAO, Y., ZHANG, X., YANG, L., SONG, C., WANG, M., SHI, J., LIU, G., LIU, J., ZHOU, H., ZHOU, W., YU, Q., AN, N., CHEN, Y., CAI, Q., WANG, B., LIU, B., MIN, J., HUANG, Y., WU, H., LI, Z., ZHANG, Y., YIN, Y., SONG, W., JIANG, J., JACKSON, S. A., WING, R. A., WANG, J. & CHEN, M. 2013. Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying Oryza genome evolution. ***Nature Communications***, 4**,** 1595.

CHEN, J., LI, L. & WANG, Y. 2012. Diversity of genome size and Ty1-*copia* in *Epimedium* species used for traditional chinese medicines. ***HortScience***, 47**,** 979-984.

CHENG, X., ZHANG, D., CHENG, Z., KELLER, B. & LING, H. Q. 2009. A new family of Ty1-*copia*-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. ***Genetics***, 181**,** 1183-93.

CHIU, L. W., ZHOU, X., BURKE, S., WU, X., PRIOR, R. L. & LI, L. 2010. The purple cauliflower arises from activation of a MYB transcription factor. ***Plant Physiology***, 154**,** 1470.

CHOPRA, S., BRENDEL, V., ZHANG, J., AXTELL, J. D. & PETERSON, T. 1999. Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from *Sorghum bicolor*. ***Proceedings of the National Academy of Sciences of the United States of America***, 96**,** 15330.

CHOUDHURY, R. R., BASAK, S., RAMESH, A. M. & RANGAN, L. 2014. Nuclear DNA content of *Pongamia pinnata* L. and genome size stability of in vitro-regenerated plantlets. ***Protoplasma***, 251**,** 703-709.

CHUONG, E. B., ELDE, N. C. & FESCHOTTE, C. C. D. 2017. Regulatory activities of transposable elements: from conflicts to benefits. ***Nature Reviews Genetics***, 18**,** 71-86.

CLIFTON, S. W., MINX, P., FAURON, C. M. R., GIBSON, M., ALLEN, J. O., SUN, H., THOMPSON, M., BARBAZUK, W. B., KANUGANTI, S., TAYLOE, C., MEYER, L., WILSON, R. K. & NEWTON, K. J. 2004. Sequence and comparative analysis of the maize nb mitochondrial genome. ***Plant Physiology***, 136**,** 3486.

COLLARD, B. C. Y. & MACKILL, D. J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. ***Philosophical Transactions of the Royal Society of London B: Biological Sciences***, 363**,** 557-72.

CONLEY, A. B. & JORDAN, I. K. 2012. Cell type-specific termination of transcription by transposable element sequences. ***Mobile DNA***, 3**,** 15.

CORDAUX, R. & BATZER, M. A. 2009. The impact of retrotransposons on human genome evolution. ***Nature Reviews Genetics,*** 10**,** 691-703.

CORDEIRO, G. M., CASU, R., MCINTYRE, C. L., MANNERS, J. M. & HENRY, R. J. 2001. Microsatellite markers from sugarcane (*Saccharum spp.*) ESTs cross transferable to erianthus and sorghum. ***Plant Science***, 160**,** 1115-1123.

DA MAIA, L. C., PALMIERI, D. A., DE SOUZA, V. Q., KOPP, M. M., DE CARVALHO, F. I. F. C. L. & COSTA DE OLIVEIRA, A. 2008. SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. ***International Journal of Plant Genomics***, 2008**,** 412696.

DAVIS, M. B., DIETZ, J., STANDIFORD, D. M. & EMERSON, C. P., JR. 1998. Transposable element insertions respecify alternative exon splicing in three *Drosophila* myosin heavy chain mutants. ***Genetics***, 150**,** 1105-14.

DE FELICE, B., WILSON, R. R., ARGENZIANO, C., KAFANTARIS, I. & CONICELLA, C. 2009. A transcriptionally active *copia*-like retroelement in *Citrus limon*. ***Cellular & Molecular Biology Letters***, 14**,** 289-304.

DE KEUKELEIRE, P., DE SCHEPPER, S., GIELIS, J. & GERATS, T. 2004. A PCR-based assay to detect *hAT*-like transposon sequences in plants. ***Chromosome Research***, 12**,** 117-123.

DE SOUZA, F. V. S. J., FRANCHINI, L. A. F. & RUBINSTEIN, M. 2013. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? ***Molecular Biology and Evolution***, 6, 1239-51.

DE SOUZA, T. S. B., CHALUVADI, S. R., JOHNEN, L., MARQUES, A., GONZC!LEZ-ELIZONDO, M. S., BENNETZEN, J. L. & VANZELA, A. L. L. 2018. Analysis of retrotransposon abundance, diversity and distribution in holocentric *Eleocharis* (Cyperaceae) genomes. ***Annals of Botany***, 2, 279–290.

DEBARRY, J. D., GANKO, E. W., MCCARTHY, E. M. & MCDONALD, J. F. 2006. The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. ***Molecular Biology and Evolution***, 23**,** 479-481.

DENG, P., WANG, M., FENG, K., CUI, L., TONG, W., SONG, W. & NIE, X. 2016. Genome-wide characterization of microsatellites in Triticeae species: abundance, distribution and evolution. ***Scientific Reports***, 6**,** 32224.

DETTORI, M. T., MICALI, S., GIOVINAZZI, J., SCALABRIN, S., VERDE, I. & CIPRIANI, G. 2015. Mining microsatellites in the peach genome: development of new long-core SSR markers for genetic analyses in five *Prunus* species. ***SpringerPlus***, 4**,** 337.

DEVOS, K. M., BROWN, J. K. M. & BENNETZEN, J. L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. ***Genome Research***, 12**,** 1075-1079.

DICE, L. R. 1945. Measures of the amount of ecologic association between species. ***Ecology***, 26**,** 297-302.

DU, C., CARONNA, J., HE, L. & DOONER, H. K. 2008. Computational prediction and molecular confirmation of Helitron transposons in the maize genome. ***BMC Genomics***, 9**,** 51.

DU, C., FEFELOVA, N., CARONNA, J., HE, L. & DOONER, H. K. 2009. The polychromatic Helitron landscape of the maize genome. ***Proceedings of the National Academy of Sciences of the United States of America***, 106**,** 19916-21.

DUTTA, S., KUMAWAT, G., SINGH, B. P., GUPTA, D. K., SINGH, S., DOGRA, V., GAIKWAD, K., SHARMA, T. R., RAJE, R. S., BANDHOPADHYA, T. K., DATTA, S., SINGH, M. N., BASHASAB, F., KULWAL, P., WANJARI, K. B., K VARSHNEY, R., COOK, D. R. & SINGH, N. K. 2011. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* L.) Millspaugh]. ***BMC Plant Biology***, 11**,** 17.

EL BAIDOURI, M. & PANAUD, O. 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. ***Genome Biology and Evolution***, 5**,** 954-965.

ELLEGREN, H. 2004. Microsatellites: simple sequences with complex evolution. ***Nature Reviews Genetics***, 5**,** 435.

EMERA, D. & WAGNER, G. C. 2012. Transposable element recruitments in the mammalian placenta: impacts and mechanisms. ***Briefings in Functional Genomics***, 11**,** 267-276.

ETMINAN, A., POUR-ABOUGHADAREH, A., MOHAMMADI, R., AHMADI-RAD, A., NOORI, A., MAHDAVIAN, Z. & MORADI, Z. 2016. Applicability of start codon targeted (SCoT) and inter-simple sequence repeat (ISSR) markers for genetic diversity analysis in durum wheat genotypes. ***Biotechnology & Biotechnological Equipment***, 30**,** 1075-1081.

EUJAYL, I., SLEDGE, M. K., WANG, L., MAY, G. D., CHEKHOVSKIY, K., ZWONITZER, J. C. & MIAN, M. A. 2004. *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. ***Theoretical and Applied Genetics***, 108**,** 414-22.

FAWCETT, J. A., KAWAHARA, T., WATANABE, H. & YASUI, Y. 2006. A SINE family widely distributed in the plant kingdom and its evolutionary history. ***Plant Molecular Biology***, 61**,** 505-14.

FEDOROFF, N., WESSLER, S. & SHURE, M. 1983. Isolation of the transposable maize controlling elements *Ac* and *Ds*. ***Cell***, 35**,** 235-42.

FENG, G., LEEM, Y. E. & LEVIN, H. L. 2013. Transposon integration enhances expression of stress response genes. ***Nucleic Acids Research***, 41**,** 775-89.

FESCHOTTE, C. 2008. Transposable elements and the evolution of regulatory networks. ***Nature Reviews Genetics***, 9**,** 397-405.

FILIZ, E. & KOC, I. 2012. In Silico chloroplast SSRs mining of Olea species. ***Biodiversitas*** 13**,** 114-117.

FILIZ, E. 2013. SSRs mining of Brassica species in mitochondrial genomes: Bioinformatic approaches. ***Horticulture, Environment, and Biotechnology***, 54**,** 548-553.

FINATTO, T., DE OLIVEIRA, A. C., CHAPARRO, C., DA MAIA, L. C., FARIAS, D. R., WOYANN, L. G., MISTURA, C. C., SOARES-BRESOLIN, A. P., LLAURO, C., PANAUD, O. & PICAULT, N. 2015. Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice. ***Rice***, 8**,** 13.

FLAVELL, A. J., SMITH, D. B. & KUMAR, A. 1992. Extreme heterogeneity of Ty1-*copia* group retrotransposons in plants. ***Molecular and General Genetics MGG***, 231**,** 233-42.

FLAVELL, R. B. 1986. Repetitive DNA and chromosome evolution in plants. ***Philosophical transactions of the Royal Society of London. Series B, Biological sciences***, 312**,** 227-42.

FRIESEN, N., BRANDES, A. & HESLOP-HARRISON, J. S. 2001. Diversity, origin, and distribution of retrotransposons (*gypsy* and *copia*) in conifers. ***Molecular Biology and Evolution***, 18**,** 1176-88.

GANDHI, S. G., AWASTHI, P. & BEDI, Y. S. 2010. Analysis of SSR dynamics in chloroplast genomes of Brassicaceae family. ***Bioinformation***, 5**,** 16-20.

GANKO, E. W., BHATTACHARJEE, V., SCHLIEKELMAN, P. & MCDONALD, J. F. 2003. Evidence for the Contribution of LTR Retrotransposons to C. elegans Gene Evolution. ***Molecular Biology and Evolution***, 20**,** 1925-1931.

GAO, C., ZHOU, G., MA, C., ZHAI, W., ZHANG, T., LIU, Z., YANG, Y., WU, M., YUE, Y., DUAN, Z., LI, Y., LI, B., LI, J., SHEN, J., TU, J. & FU, T. 2016. Helitron-like transposons contributed to the mating system transition from out-crossing to self-fertilizing in polyploid Brassica napus L. ***Scientific Reports***, 6**,** 33785.

GAO, D., ABERNATHY, B., ROHKSAR, D., SCHMUTZ, J. & JACKSON, S. A. 2014. Annotation and sequence diversity of transposable elements in common bean (*Phaseolus vulgaris*). ***Frontiers in Plant Science***, 5**,** 339.

GARCIA, R. A., RANGEL, P. N., BRONDANI, C., MARTINS, W. S., MELO, L. C., CARNEIRO, M. S., BORBA, T. C. & BRONDANI, R. P. 2011. The characterization of a new set of EST-derived simple sequence repeat (SSR) markers as a resource for the genetic analysis of *Phaseolus vulgaris*. ***BMC Genetics***, 12**,** 41.

GORINSL EK, B., GUBENSL EK, F. & KORDISL, D. A. 2004. Evolutionary genomics of chromoviruses in eukaryotes. ***Molecular Biology and Evolution***, 21**,** 781-798.

GOTEA, V. & MAKALOWSKI, W. 2006. Do transposable elements really contribute to proteomes? ***Trends in Genetics***, 22**,** 260-7.

GRABUNDZIJA, I., MESSING, S. A., THOMAS, J., COSBY, R. L., BILIC, I., MISKEY, C., GOGOL-DORING, A., KAPITONOV, V., DIEM, T., DALDA, A., JURKA, J., PRITHAM, E. J., DYDA, F., IZSVC!K, Z. & IVICS, Z. N. 2016. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. ***Nature Communications*** , 7, 10716.

GRANDBASTIEN, M. A., AUDEON, C., BONNIVARD, E., CASACUBERTA, J. M., CHALHOUB, B., COSTA, A. P., LE, Q. H., MELAYAH, D., PETIT, M., PONCET, C., TAM, S. M., VAN SLUYS, M. A. & MHIRI, C. 2005. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. ***Cytogenet Genome Research***, 110**,** 229-41.

GRANDBASTIEN, M. A., LUCAS, H. C., MOREL, J. B. T., MHIRI, C., VERNHETTES, S. & CASACUBERTA, J. M. 1997. The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. ***Genetica***, 100**,** 241-252.

GREENE, B., WALKO, R. & HAKE, S. 1994. Mutator insertions in an intron of the maize knotted1 gene result in dominant suppressible mutations. ***Genetics***, 138**,** 1275.

GREILHUBER, J. & EBERT, I. 1994. Genome size variation in *Pisum sativum*. ***Genome***, 37**,** 646-655.

GREINER, S., SOBANSKI, J. & BOCK, R. 2015. Why are most organelle genomes transmitted maternally? ***Bioessays***, 37**,** 80-94.

GUASMI, F., ELFALLEH, W., HANNACHI, H., FERES, K., TOUIL, L., MARZOUGUI, N., TRIKI, T. & FERCHICHI, A. 2012. The use of ISSR and RAPD markers for genetic diversity among south tunisian barley. ***ISRN Agronomy***, 2012**,** 10.

GUO, B., CHEN, X., DANG, P., SCULLY, B. T., LIANG, X., HOLBROOK, C. C., YU, J. & CULBREATH, A. K. 2008. Peanut gene expression profiling in developing seeds at different reproduction stages during *Aspergillus parasiticus* infection. ***BMC Developmental Biology***, 8**,** 12-12.

GUO, C., SPINELLI, M., LIU, M., LI, Q. Q. & LIANG, C. 2016. A Genome-wide study of "non-3UTR" polyadenylation sites in *Arabidopsis thaliana*. ***Scientific Reports***, 6**,** 28060.

GUO, C., SPINELLI, M., YE, C., LI, Q. Q. & LIANG, C. 2017. Genome-wide comparative analysis of miniature inverted repeat transposable elements in 19 *Arabidopsis thaliana* ecotype accessions. ***Scientific Reports***, 7**,** 2634.

GUPTA, S., TRIPATHI, K. P., ROY, S. & SHARMA, A. 2010. Analysis of unigene derived microsatellite markers in family solanaceae. ***Bioinformation***, 5**,** 113-121.

GYAWALI, R. & LIN, X. 2013. Prezygotic and postzygotic control of uniparental mitochondrial DNA inheritance in *Cryptococcus neoformans*. ***MBio,*** 4, e00112-13.

HALDER, S., SAKTHIVEL, S., JAYARAJ, K. M. & GUPTA, P. D. 2014. Studies of transesterification of karanja (*Pongamia pinnata*) oil in a packed bed reactor. ***Chemical Engineering Communications***, 201**,** 88-101.

HAN, Z., WANG, C., SONG, X., GUO, W., GOU, J., LI, C., CHEN, X. & ZHANG, T. 2006. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. ***Theoretical and Applied Genetics***, 112**,** 430-9.

HARREH, D., SALEH, A. A., REDDY, A. N. R. & HAMDAN, S. 2018. An experimental investigation of karanja biodiesel production in Sarawak, Malaysia. ***Journal of Engineering***, 2018**,** 8.

HARTINGS, H., LAZZARONI, N., MOTTO, M. & FRIZZI, A. I. S. P. L. C., ROME (ITALY) 1998. Distribution of Bg-homologous sequences in Gramineae species. ***Maydica***, 43, 103-109.

HAWKINS, J. S., KIM, H., NASON, J. D., WING, R. A. & WENDEL, J. F. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. ***Genome Research***, 16, 1252-61

HAWKINS, J. S., PROULX, S. R., RAPP, R. A. & WENDEL, J. F. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. ***Proceedings of the National Academy of Sciences of the United States of America***, 106**,** 17811.

HAYASHI, K. & YOSHIDA, H. 2009. Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. ***The Plant Journal***, 57**,** 413-425.

HE, P., MA, Y., ZHAO, G., DAI, H., LI, H., CHANG, L. & ZHANG, Z. 2010. FaRE1: A transcriptionally active Ty1-*copia* retrotransposon in strawberry. ***Journal of Plant Research***, 123**,** 707-714.

HEHL, R., NACKEN, W. K., KRAUSE, A., SAEDLER, H. & SOMMER, H. 1991. Structural analysis of Tam3, a transposable element from *Antirrhinum majus*, reveals homologies to the *Ac* element from maize. ***Plant Molecular Biology***, 16**,** 369-71.

HILGARD, P., HUANG, T., WOLKOFF, A. W. & STOCKERT, R. J. 2002. Translated *Alu* sequence determines nuclear localization of a novel catalytic subunit of casein kinase 2. ***American Journal of Physiology Cell Physiology***, 283**,** C472-83.

HILL, P., BURFORD, D., MARTIN, D. M. A. & FLAVELL, A. J. 2005. Retrotransposon populations of Vicia species with varying genome size. ***Molecular Genetics and Genomics***, 273**,** 371-381.

HIROCHIKA, H. & HIROCHIKA, R. 1993. Ty1-*copia* group retrotransposons as ubiquitous components of plant genomes. ***The Japanese Journal of Genetics***, 68**,** 35-46.

HIROCHIKA, H. 1997. Retrotransposons of rice: their regulation and use for genome analysis. ***Plant Molecular Biology***, 35**,** 231-40.

HIROCHIKA, H., SUGIMOTO, K., OTSUKI, Y., TSUGAWA, H. & KANDA, M. 1996. Retrotransposons of rice involved in mutations induced by tissue culture. ***Proceedings of the National Academy of Sciences of the United States of America***, 93**,** 7783-7788.

HISANO, H., TSUJIMURA, M., YOSHIDA, H., TERACHI, T. & SATO, K. 2016. Mitochondrial genome sequences from wild and cultivated barley (*Hordeum vulgare*). ***BMC Genomics***, 17**,** 824.

HOENICKA, J., ARRASATE, M., DE YEBENES, J. G. & AVILA, J. 2002. A two-hybrid screening of human Tau protein: interactions with *Alu*-derived domain. ***Neuroreport***, 13**,** 343-9.

HORI, Y., FUJIMOTO, R., SATO, Y. & NISHIO, T. 2007. A novel wx mutation caused by insertion of a retrotransposon-like sequence in a glutinous cultivar of rice (*Oryza sativa*). ***Theoretical and Applied Genetics***, 115**,** 217-224.

HOSSAIN, M. A., HUQ, E., GROVER, A., DENNIS, E. S., PEACOCK, W. J. & HODGES, T. K. 1996. Characterization of pyruvate decarboxylase genes from rice. ***Plant Molecular Biology***, 31**,** 761-770.

HUANG, C. R. L., BURNS, K. H. & BOEKE, J. D. 2012. Active transposition in genomes. ***Annual Review of Genetics***, 46**,** 651-675.

HUANG, J., GUO, X., HAO, X., ZHANG, W., CHEN, S., HUANG, R., GRESSHOFF, P. M. & ZHENG, Y. 2016. De novo sequencing and characterization of seed transcriptome of the tree legume *Millettia pinnata* for gene discovery and SSR marker development. ***Molecular Breeding***, 36**,** 75.

HUANG, J., HAO, X., JIN, Y., GUO, X., SHAO, Q., KUMAR, K. S., AHLAWAT, Y. K., HARRY, D. E., JOSHI, C. P. & ZHENG, Y. 2018. Temporal transcriptome profiling of developing seeds reveals a concerted gene regulation in relation to oil accumulation in Pongamia (*Millettia pinnata*). ***BMC Plant Biology***, 18**,** 140.

HUANG, J., LU, X., YAN, H., CHEN, S., ZHANG, W., HUANG, R. & ZHENG, Y. 2012. Transcriptome characterization and sequencing-based identification of salt-responsive genes in *Millettia pinnata*, a semi-mangrove plant. ***DNA Research***, 19**,** 195-207.

HUANG, J., WANG, Y., LIU, W., SHEN, X., FAN, Q., JIAN, S. & TANG, T. 2017a. EARE-1, a Transcriptionally active Ty1/copia-like retrotransposon has colonized the genome of *Excoecaria agallocha* through horizontal transfer. ***Frontiers in Plant Science***, 8, 45.

HUANG, J., ZHANG, K., SHEN, Y., HUANG, Z., LI, M., TANG, D., GU, M. & CHENG, Z. 2009. Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the *hAT* family in rice. ***Genomics***, 93**,** 274-281.

HUANG, Y., LUO, L., HU, X., YU, F., YANG, Y., DENG, Z., WU, J., CHEN, R. & ZHANG, M. 2017b. Characterization, genomic organization, abundance, and chromosomal distribution of Ty1-*copia* retrotransposons in *Erianthus arundinaceus*. ***Frontiers in Plant Science***, 8**,** 924.

HUDA, A., TYAGI, E., MARINO-RAMIREZ, L., BOWEN, N. J., JJINGO, D. & JORDAN, I. K. 2011. Prediction of transposable element derived enhancers using chromatin modification profiles. ***PLoS One***, 6**,** e27513.

HUTCHINSON, J. B. 1940. The application of genetics to plant breeding. I. The genetic interpretation of plant breeding problems. ***Journal of Genetics***, 40**,** 271-82.

IBARRA-LACLETTE, E., LYONS, E., HERNANDEZ-GUZMAN, G., PEREZ-TORRES, C. A., CARRETERO-PAULET, L., CHANG, T. H., LAN, T., WELCH, A. J., JUAREZ, M. A. J. N. A., SIMPSON, J., FERNC!NDEZ-CORTES, A., ARTEAGA-VAZQUEZ, M., GONGORA-CASTILLO, E., ACEVEDO-HERNANDEZ, G., SCHUSTER, S. C., HIMMELBAUER, H., MINOCHE, A. E., XU, S., et al. 2013. Architecture and evolution of a minute plant genome. ***Nature***, 498**,** 94.

ISLAM, M. S., STUDER, B., BYRNE, S. L., FARRELL, J. D., PANITZ, F., BENDIXEN, C., MC8LLER, I. M. & ASP, T. 2013. The genome and transcriptome of perennial ryegrass mitochondria. ***BMC Genomics***, 14**,** 202.

ISLAM, M. S., STUDER, B., BYRNE, S. L., FARRELL, J. D., PANITZ, F., BENDIXEN, C., MC8LLER, I. M. & ASP, T. 2013. The genome and transcriptome of perennial ryegrass mitochondria. ***BMC Genomics***, 14**,** 202.

ISMOND, K. P., DOLFERUS, R., DE PAUW, M., DENNIS, E. S. & GOOD, A. G. 2003. Enhanced low oxygen survival in arabidopsis through increased metabolic flux in the fermentative pathway. ***Plant Physiology***, 132**,** 1292.

ITO, H., GAUBERT, H., BUCHER, E., MIROUZE, M., VAILLANT, I. & PASZKOWSKI, J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. ***Nature***, 472**,** 115.

ITO, H., KIM, J. M., MATSUNAGA, W., SAZE, H., MATSUI, A., ENDO, T. A., HARUKAWA, Y., TAKAGI, H., YAEGASHI, H., MASUTA, Y., MASUDA, S., ISHIDA, J., TANAKA, M., TAKAHASHI, S., MOROSAWA, T., TOYODA, T., KAKUTANI, T., KATO, A. & SEKI, M. 2016. A stress-activated transposon in arabidopsis induces transgenerational abscisic acid insensitivity. ***Scientific Reports***, 6**,** 23181.

ITO, H., YOSHIDA, T., TSUKAHARA, S. & KAWABE, A. 2013. Evolution of the ONSEN retrotransposon family activated upon heat stress in Brassicaceae. ***Gene***, 518**,** 256-261.

JARNE, P. & LAGODA, P. J. L. 1996. Microsatellites, from molecules to populations and back. ***Trends in Ecology & Evolution***, 11**,** 424-429.

JIANG, B. A., LIU, W. R., HE, X. M., PENG, Q. W. & XIE, D. S. 2013. Characterization and chromosomal distribution of Ty3-gypsy-like retrotransposons in wax gourd (*Benincasa hispida*). ***Scienceasia***, 39**,** 466-471.

JIANG, Q., YEN, S. H., STILLER, J., EDWARDS, D., SCOTT, P. T. & GRESSHOFF, P. M. 2012. Genetic, biochemical, and morphological diversity of the legume biofuel tree *Pongamia pinnata*. ***Plant Genetics, Genomics, and Biotechnology***, 1**,** 54-67.

JIANG, S., CAI, D., SUN, Y. & TENG, Y. 2016. Isolation and characterization of putative functional long terminal repeat retrotransposons in the Pyrus genome. ***Mobile DNA***, 7**,** 1.

JONES, C., EDWARDS, K., CASTAGLIONE, S., WINFIELD, M., SALA, F., VAN DE WIEL, C., BREDEMEIJER, G., VOSMAN, B., MATTHES, M. & DALY, A. 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. ***Molecular breeding***, 3**,** 381-390.

JORDAN, I. K. & MCDONALD, J. F. 1999. The role of interelement selection in *Saccharomyces cerevisiae* Ty element evolution. ***Journal of Molecular Evolution***, 49**,** 352-357.

JOY, N., MAIMOONATH BEEVI, Y. P. & SONIYA, E. V. 2018. A deeper view into the significance of simple sequence repeats in pre-miRNAs provides clues for its possible roles in determining the function of microRNAs. ***BMC Genetics***, 19, 29.

KALENDAR, R., ANTONIUS, K., SMYKAL, P. & SCHULMAN, A. H. 2010. iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. ***Theoretical and Applied Genetics***, 121**,** 1419-1430.

KALENDAR, R., VICIENT, C. M., PELEG, O., ANAMTHAWAT-JONSSON, K., BOLSHOY, A. & SCHULMAN, A. H. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. ***Genetics***, 166**,** 1437-50.

KALIA, R. K., RAI, M. K., KALIA, S., SINGH, R. & DHAWAN, A. K. 2011. Microsatellite markers: an overview of the recent progress in plants. ***Euphytica***, 177**,** 309-334.

KALRA, S., PUNIYA, B. L., KULSHRESHTHA, D., KUMAR, S., KAUR, J., RAMACHANDRAN, S. & SINGH, K. 2013. *De novo* transcriptome sequencing reveals important molecular networks and metabolic pathways of the plant, *Chlorophytum borivilianum*. ***PLoS One***, 8**,** e83336.

KANAZAWA, A., LIU, B., KONG, F., ARASE, S. & ABE, J. 2009. Adaptive evolution involving gene duplication and insertion of a novel Ty1/*copia*-like retrotransposon in soybean. ***Journal of Molecular Evolution***, 69**,** 164-75.

KANNAN, S., CHERNIKOVA, D., ROGOZIN, I. B., POLIAKOV, E., MANAGADZE, D., KOONIN, E. V. & MILANESI, L. 2015. Transposable element insertions in

long intergenic non-coding RNA genes. ***Frontiers in Bioengineering and Biotechnology***, 3**,** 71.

KANTETY, R. V., LA ROTA, M., MATTHEWS, D. E. & SORRELLS, M. E. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. ***Plant Molecular Biology***, 48**,** 501-510.

KARMEE, S. K. & CHADHA, A. 2005. Preparation of biodiesel from crude oil of *Pongamia pinnata*. ***Bioresource Technology***, 96, 1425-1429.

KAUSHIK, N., KUMAR, S., KUMAR, K., BENIWAL, R. S., KAUSHIK, N. & ROY, S. 2007. Genetic variability and association studies in pod and seed traits of *Pongamia pinnata* (L.) Pierre in Haryana, India. ***Genetic Resources and Crop Evolution***, 54**,** 1827-1832.

KAUSHIK, N., KUMAR, S., KUMAR, K., BENIWAL, R. S., KAUSHIK, N. & ROY, S. 2007. Genetic variability and association studies in pod and seed traits of *Pongamia pinnata* (L.) Pierre in Haryana, India. ***Genetic Resources and Crop Evolution***, 54**,** 1827-1832.

KAWASE, M., FUKUNAGA, K. & KATO, K. 2005. Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions. ***Molecular Genetics and Genomics***, 274**,** 131-40.

KAZAKOFF, S. H., IMELFORT, M., EDWARDS, D., KOEHORST, J., BISWAS, B., BATLEY, J., SCOTT, P. T. & GRESSHOFF, P. M. 2012. Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. ***PLoS One***, 7**,** e51687.

KESARI, V. & RANGAN, L. 2011. Coordinated changes in storage proteins during development and germination of elite seeds of *Pongamia pinnata*, a versatile biodiesel legume. ***AoB PLANTS***, 2011**,** plr026-plr026.

KESARI, V. & RANGAN, L. 2011. Genetic diversity analysis by RAPD markers in candidate plus trees of *Pongamia pinnata*, a promising source of bioenergy. ***Biomass and Bioenergy***, 35**,** 3123-3128.

KESARI, V., KRISHNAMACHARI, A. & RANGAN, L. 2008. Systematic characterisation and seed oil analysis in candidate plus trees of biodiesel plant, *Pongamia pinnata*. ***Annals of Applied Biology***, 152**,** 397-404.

KESARI, V., MADURAI SATHYANARAYANA, V., PARIDA, A. & RANGAN, L. 2010. Molecular marker-based characterization in candidate plus trees of *Pongamia pinnata*, a potential biodiesel legume. ***AoB Plants***, 2010**,** plq017.

KESARI, V., RAMESH, A. M. & RANGAN, L. 2013. Rhizobium pongamiae sp. nov. from Root Nodules of *Pongamia pinnata*. ***Biomed Research International***, 2013, 9.

KESARI, V., SUDARSHAN, M., DAS, A. & RANGAN, L. 2009. PCR amplification of the genomic DNA from the seeds of Ceylon ironwood, *Jatropha*, and *Pongamia*. ***Biomass and Bioenergy***, 33**,** 1724-1728.

KIDWELL, M. G. & LISCH, D. 1997. Transposable elements as sources of variation in animals and plants**.** ***Proceedings of the National Academy of Sciences of the United States of America***, 94**,** 7704-11.

KIDWELL, M. G. 2002. Transposable elements and the evolution of genome size in eukaryotes. ***Genetica***, 115**,** 49-63.

KIM, J. M., VANGURI, S., BOEKE, J. D., GABRIEL, A. & VOYTAS, D. F. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. ***Genome Research***, 8**,** 464-78.

KIM, J. H., CHUNG, I. K. & KIM, K. M. 2017. Construction of a genetic map using EST-SSR markers and QTL analysis of major agronomic characters in hexaploid sweet potato (*Ipomoea batatas* (L.) Lam). ***PLoS One***, 12**,** e0185073.

KIMURA, Y., TOSA, Y., SHIMADA, S., SOGO, R., KUSABA, M., SUNAGA, T., BETSUYAKU, S., ETO, Y., NAKAYASHIKI, H. & MAYAMA, S. 2001. OARE-1, a Ty1-copia retrotransposon in oat activated by abiotic and biotic stresses. ***Plant and Cell Physiology***, 42**,** 1345-1354.

KNIP, M., DE PATER, S. & HOOYKAAS, P. J. 2012. The SLEEPER genes: a transposase-derived angiosperm-specific gene family. ***BMC Plant Biology,*** 12, 192.

KNIP, M., HIEMSTRA, S., SIETSMA, A., CASTELEIN, M., DE PATER, S. & HOOYKAAS, P. 2013. DAYSLEEPER: a nuclear and vesicular-localized protein that is expressed in proliferating tissues. ***BMC Plant Biology***, 13**,** 211.

KNOOP, V., UNSELD, M., MARIENFELD, J., BRANDT, P., SUNKEL, S., ULLRICH, H. & BRENNICKE, A. 1996. *copia-*, *gypsy-* and LINE-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. ***Genetics***, 142**,** 579-85.

KOBAYASHI, S., GOTO-YAMAMOTO, N. & HIROCHIKA, H. 2004. Retrotransposon-induced mutations in grape skin color. ***Science***, 304**,** 982.

KOLANO, B., SARACKA, K., BRODA-CNOTA, A. & MALUSZYNSKA, J. 2013. Localization of ribosomal DNA and CMA3/DAPI heterochromatin in cultivated and wild *Amaranthus* species. ***Scientia Horticulturae***, 164**,** 249-255.

KUBIS, S. E., HESLOP-HARRISON, J. S., DESEL, C. & SCHMIDT, T. 1998. The genomic organization of non-LTR retrotransposons (LINEs) from three *Beta* species and five other angiosperms. ***Plant Molecular Biology***, 36**,** 821-831.

KUMAR, B., KUMAR, U. & YADAV HEMANT, K. 2015. Identification of EST–SSRs and molecular diversity analysis in *Mentha piperita*. ***The Crop Journal***, 3**,** 335-342.

KUMARI, K., MUTHAMILARASAN, M., MISRA, G., GUPTA, S., SUBRAMANIAN, A., PARIDA, S. K., CHATTOPADHYAY, D. & PRASAD, M. 2013. Development of eSSR-markers in *Setaria italica* and their applicability in

studying genetic diversity, cross-transferability and comparative mapping in millet and non-millet species. ***PLoS One***, 8**,** e67742.

KUNTAL, H., SHARMA, V. & DANIELL, H. 2012. Microsatellite analysis in organelle genomes of Chlorophyta. ***Bioinformation***, 8**,** 255-259.

KURSTEINER, O., DUPUIS, I. & KUHLEMEIER, C. 2003. The pyruvate decarboxylase1 gene of Arabidopsis is required during anoxia but not other environmental stresses. ***Plant Physiology***, 132**,** 968-78.

LATEN, H. M., HAVECKER, E. R., FARMER, L. M. & VOYTAS, D. F. 2003. SIRE1, an endogenous retrovirus family from *Glycine max*, is highly homogeneous and evolutionarily young. ***Molecular Biology and Evolution***, 20**,** 1222-1230.

LAWSON, M. J. & ZHANG, L. 2006. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. ***Genome Biology***, 7**,** R14-R14.

LEE, J., WAMINAL, N. E., CHOI, H. I., PERUMAL, S., LEE, S. C., NGUYEN, V. B., JANG, W., KIM, N. H., GAO, L. Z. & YANG, T. J. 2017. Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus *Panax*. ***Scientific Reports***, 7**,** 9045.

LEE, S. I., PARK, K. C., SON, J. H., HWANG, Y. J., LIM, K. B., SONG, Y. S., KIM, J. H. & KIM, N. S. 2013. Isolation and characterization of novel Ty1-*copia*-like retrotransposons from lily. ***Genome***, 56**,** 495-503.

LEE, Y. C. G. & LANGLEY, C. H. 2010. Transposable elements in natural populations of *Drosophila melanogaster*. ***Philosophical Transactions of the Royal Society B: Biological Sciences***, 365**,** 1219-1228.

LEETON, P. R. J. & SMYTH, D. R. 1993. An abundant LINE-like element amplified in the genome of *Lilium speciosum*. ***Molecular and General Genetics MGG***, 237**,** 97-104.

LI, L., FANG, Z., ZHOU, J., CHEN, H., HU, Z., GAO, L., CHEN, L., REN, S., MA, H., LU, L., ZHANG, W. & PENG, H. 2017. An accurate and efficient method for large-scale SSR genotyping and applications. ***Nucleic Acids Research***, 45**,** e88-e88.

LI, Y. C., KOROL, A. B., FAHIMA, T. & NEVO, E. 2004. Microsatellites Within Genes: Structure, Function, and Evolution. ***Molecular Biology and Evolution***, 21**,** 991-1007.

LIN, L., JIANG, P., PARK, J. W., WANG, J., LU, Z. X., LAM, M. P. Y., PING, P. & XING, Y. 2016. The contribution of *Alu* exons to the human proteome. ***Genome Biology***, 17**,** 15.

LIN, R., DING, L., CASOLA, C., RIPOLL, D. R., FESCHOTTE, C. C. D. & WANG, H. 2007. Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. ***Science***, 318**,** 1302.

LISCH, D. 2013. How important are transposons for plant evolution? ***Nature Reviews Genetics***, 14**,** 49-61.

LIU, S. R., LI, W. Y., LONG, D., HU, C. G. & ZHANG, J. Z. 2013. Development and characterization of genomic and expressed SSRs in *Citrus* by genome-wide analysis. ***PLoS One***, 8**,** e75149.

LIU, W., WANG, Y., SHEN, X. & TANG, T. 2016. Isolation, characterization, and marker utility of KCRE1, a transcriptionally active Ty1/*copia* retrotransposon from *Kandelia candel*. ***Molecular Genetics and Genomics***, 291**,** 2031-2042.

LLORENS, C., FUTAMI, R., COVELLI, L., DOMINGUEZ-ESCRIBA, L., VIU, J. M., TAMARIT, D., AGUILAR-RODRIGUEZ, J., VICENTE-RIPOLLES, M., FUSTER, G., BERNET, G. P., MAUMUS, F., MUNOZ-POMER, A., SEMPERE, J. M., LATORRE, A. & MOYA, A. 2011. The Gypsy database (GyDB) of mobile genetic elements: release 2.0. ***Nucleic Acids Research***, D70-4.

LLORENS, C., MUC1OZ-POMER, A., BERNAD, L., BOTELLA, H. & MOYA, A. S. 2009. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. ***Biology Direct***, 4**,** 41.

LOPES, F. R., CARAZZOLLE, M. F., PEREIRA, G. A. G., COLOMBO, C. A. & CARARETO, C. M. A. 2008. Transposable elements in Coffea (Gentianales : Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants. ***Molecular Genetics and Genomics***, 279**,** 385-401.

LOPES, F. R., JJINGO, D., DA SILVA, C. R., ANDRADE, A. C., MARRACCINI, P., TEIXEIRA, J. B., CARAZZOLLE, M. F., PEREIRA, G. A., PEREIRA, L. F., VANZELA, A. L., WANG, L., JORDAN, I. K. & CARARETO, C. M. 2013. Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. ***PLoS One***, 8**,** e78931.

LORENC, A. & MAKALOWSKI, W. 2003. Transposable elements and vertebrate protein diversity. ***Genetica***, 118**,** 183-91.

LOSADA, L., PAKALA, S. B., FEDOROVA, N. D., JOARDAR, V., SHABALINA, S. A., HOSTETLER, J., PAKALA, S. M., ZAFAR, N., THOMAS, E., RODRIGUEZ-CARRES, M., DEAN, R., VILGALYS, R., NIERMAN, W. C. & CUBETA, M. A. 2014. Mobile elements and mitochondrial genome expansion in the soil fungus and potato pathogen *Rhizoctonia solani* AG-3. ***FEMS Microbiol Letters***, 352**,** 165-73.

LURO, F. O. L., COSTANTINO, G., TEROL, J., ARGOUT, X., ALLARIO, T., WINCKER, P., TALON, M., OLLITRAULT, P. & MORILLON, R. 2008. Transferability of the EST-SSRs developed on *Nules clementine* (*Citrus clementina* Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. ***BMC Genomics***, 9**,** 287.

LYNCH, V. J., NNAMANI, M. C., KAPUSTA, A. L., BRAYER, K., PLAZA, S. L., MAZUR, E. C., EMERA, D., SHEIKH, S. Z., GRC, BAUERSACHS, S., GRAF,

A., YOUNG, S. L., LIEB, J. D., DEMAYO, F. J., FESCHOTTE, C. C. D. & WAGNER, G. C. 2015. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. ***Cell Reports***, 10**,** 551-561.

MA, J., DEVOS, K. M. & BENNETZEN, J. L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic dna loss in rice. ***Genome Research***, 14**,** 860-869.

MA, Y., SUN, H., ZHAO, G., DAI, H., GAO, X., LI, H. & ZHANG, Z. 2008. Isolation and characterization of genomic retrotransposon sequences from octoploid strawberry (*Fragaria x ananassa Duch.*). ***Plant Cell Reports***, 27**,** 499-507.

MACRAE, A. F., HUTTLEY, G. A. & CLEGG, M. T. 1994. Molecular evolutionary characterization of an *Activator* (*Ac*)-like transposable element sequence from pearl millet (*Pennisetum glaucum*) (Poaceae). ***Genetica***, 92**,** 77-89.

MAGALHAES, J. V., LIU, J., GUIMARAES, C. T., LANA, U. G., ALVES, V. M., WANG, Y. H., SCHAFFERT, R. E., HOEKENGA, O. A., PINEROS, M. A., SHAFF, J. E., KLEIN, P. E., CARNEIRO, N. P., COELHO, C. M., TRICK, H. N. & KOCHIAN, L. V. 2007. A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. ***Nature Genetics***, 39**,** 1156-61.

MAKAŁOWSKI, W., KISCHKA, T. & MAKAŁOWSKA, I. 2017. Contribution of transposable elements to human proteins. *eLS*. DOI: 10.1002/9780470015902.a0020793.pub2.

MAKALOWSKI, W., MITCHELL, G. A. & LABUDA, D. 1994. *Alu* sequences in the coding regions of mRNA: a source of protein variability. ***Trends in Genetics***, 10**,** 188-93.

MAKAOWSKI, W. 2000. Genomic scrap yard: how genomes utilize all that junk. ***Gene***, 259**,** 61-67.

MANONMANI, V., VANANGAMUDI, K. & RAI, R. V. 1996. Effect of seed size on seed germination and vigour in *Pongamia pinnata*. ***Journal of Tropical Forest Science*,** 9, 1-5.

MAO, H., WANG, H., LIU, S., LI, Z., YANG, X., YAN, J., LI, J., TRAN, L. S. P. & QIN, F. 2015. A transposable element in a *NAC* gene is associated with drought tolerance in maize seedlings. ***Nature Communications***, 6**,** 8326.

MARKOVA, D. N. & MASON-GAMER, R. J. 2017. Transcriptional activity of *PIF* and *Pong*-like Class II transposable elements in Triticeae. ***BMC Evolutionary Biology***, 17**,** 178.

MARRIAGE, T. N., HUDMAN, S., MORT, M. E., ORIVE, M. E., SHAW, R. G. & KELLY, J. K. 2009. Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). ***Heredity***, 103**,** 310-317.

MASUDA, S., NOZAWA, K., MATSUNAGA, W., MASUTA, Y., KAWABE, A., KATO, A. & ITO, H. 2016. Characterization of a heat-activated retrotransposon in natural accessions of *Arabidopsis thaliana*. ***Genes & Genetic Systems***, 91**,** 293-299.

MATSUNAGA, W., KOBAYASHI, A., KATO, A. & ITO, H. 2012. The effects of heat induction and the siRNA biogenesis pathway on the transgenerational transposition of ONSEN, a *copia*-like retrotransposon in *Arabidopsis thaliana*. ***Plant and Cell Physiology***, 53**,** 824-833.

MCCLINTOCK, B. 1946. Maize genetics. ***Carnegie Institution of Washington Year Book 45***, 176-186.

MCCLINTOCK, B. 1965. Components of action of the regulators *Spm* and *Ac*. ***Carnegie Institution of Washington Year Book 64***, 527-536.

MCCLINTOCK, B. 1950. The origin and behavior of mutable loci in maize. ***Proceedings of the National Academy of Sciences of the United States of America***, 36, 344–355.

MEDSTRAND, P., VAN DE LAGEMAAT, L. N. & MAGER, D. L. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. ***Genome Research***, 12**,** 1483-95.

METZ, S., CABRERA, J. M., RUEDA, E., GIRI, F. & AMAVET, P. 2016. FullSSR: microsatellite finder and primer designer. ***Advances in Bioinformatics***, 2016**,** 4.

MEYER, L., CAUSSE, R., PERNIN, F., SCALONE, R., BAILLY, G. C. R., CHAUVEL, B., DC)LYE, C. & LE CORRE, V. R. 2017. New gSSR and EST-SSR markers reveal high genetic diversity in the invasive plant *Ambrosia artemisiifolia* L. and can be transferred to other invasive Ambrosia species. ***PLoS One***, 12**,** e0176197.

MEYERS, B. C., TINGEY, S. V. & MORGANTE, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. ***Genome Research***, 11**,** 1660-1676.

MIAH, G., RAFII, M. Y., ISMAIL, M. R., PUTEH, A. B., RAHIM, H. A., ISLAM, K. N. & LATIF, M. A. 2013. A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. ***International Journal of Molecular Sciences***, 14, 22499–22528.

MICHALOVOVA, M., VYSKOT, B. & KEJNOVSKY, E. 2013. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. ***Heredity***, 111**,** 314-320.

MILLER, W. J., MCDONALD, J. F., NOUAUD, D. & ANXOLABEHERE, D. 1999. Molecular domestication - more than a sporadic episode in evolution. ***Genetica***, 107**,** 197-207.

MITCHELL, G. A., LABUDA, D., FONTAINE, G., SAUDUBRAY, J. M., BONNEFONT, J. P., LYONNET, S., BRODY, L. C., STEEL, G., OBIE, C. & VALLE, D. 1991. Splice-mediated insertion of an *Alu* sequence inactivates ornithine delta-aminotransferase: a role for *Alu* elements in human mutation.

Proceedings of the National Academy of Sciences of the United States of America, 88**,** 815-9.

MITHRAN, M., PAPARELLI, E., NOVI, G., PERATA, P. & LORETI, E. 2014. Analysis of the role of the *pyruvate decarboxylase* gene family in *Arabidopsis thaliana* under low-oxygen conditions. ***Plant Biology***, 16**,** 28-34.

MOGES, A. D., ADMASSU, B., BELEW, D., YESUF, M., NJUGUNA, J., KYALO, M. & GHIMIRE, S. R. 2016. Development of microsatellite markers and analysis of genetic diversity and population structure of *Colletotrichum gloeosporioides* from Ethiopia. ***PLoS One***, 11**,** e0151257.

MOKHTAR, M. M., ADAWY, S. S., EL-ASSAL, S. E. D. S. & HUSSEIN, E. H. A. 2016. Genic and intergenic SSR database generation, snps determination and pathway annotations, in date palm (*Phoenix dactylifera* L.). ***PLoS One***, 11**,** e0159268.

MOLNAR, R. I., WITTE, H., DINKELACKER, I., VILLATE, L. & SOMMER, R. J. 2012. Tandem-repeat patterns and mutation rates in microsatellites of the nematode model organism *Pristionchus pacificus*. ***Genes|Genomes|Genetics***, 2, 1027–1034.

MOOKAN, R., KRISHNASAMY, A., SUNDARESAN, M. & VELAN, P. 2014. Biodiesel production from *Pongamia pinnata* oil using synthesized iron nanocatalyst. ***International Journal of ChemTech Research***, 6**,** 4511-4516.

MORGANTE, M., HANAFEY, M. & POWELL, W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. ***Nature Genetics***, 30**,** 194.

MUKTA, N., MURTHY, I. Y. L. N. & SRIPAL, P. 2009. Variability assessment in *Pongamia pinnata* (L.) Pierre germplasm for biodiesel traits. ***Industrial Crops and Products***, 29**,** 536-540.

MUNOZ-LOPEZ, M. & GARCIA-PEREZ, J. L. 2010. DNA transposons: nature and applications in genomics. ***Current Genomics***, 11**,** 115-28.

MUSZEWSKA, A., HOFFMAN-SOMMER, M. & GRYNBERG, M. 2011. LTR retrotransposons in fungi. ***PLoS One***, 6**,** e29425.

NAGARAJA REDDY, R., MADHUSUDHANA, R., MURALI MOHAN, S., CHAKRAVARTHI, D. V. N. & SEETHARAMA, N. 2012. Characterization, development and mapping of unigene-derived microsatellite markers in sorghum [*Sorghum bicolor* (L.) Moench]. ***Molecular Breeding***, 29**,** 543-564.

NAIK, M., MEHER, L. C., NAIK, S. N. & DAS, L. M. 2008. Production of biodiesel from high free fatty acid karanja (*Pongamia pinnata*) oil. ***Biomass and Bioenergy***, 32**,** 354-357.

NAITO, K., ZHANG, F., TSUKIYAMA, T., SAITO, H., HANCOCK, C. N., RICHARDSON, A. O., OKUMOTO, Y., TANISAKA, T. & WESSLER, S. R. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. ***Nature***, 461**,** 1130-4.

NEGI, P., RAI, A. N. & SUPRASANNA, P. 2016. Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. ***Frontiers in Plant Science***, 7, 1448.

NEKRUTENKO, A. & LI, W. H. S. 2001. Transposable elements are found in a large number of human protein-coding genes. ***Trends in Genetics***, 17**,** 619-621.

NIGUMANN, P., REDIK, K., MATLIK, K. & SPEEK, M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. ***Genomics***, 79**,** 628-34.

NOTSU, Y., MASOOD, S., NISHIKAWA, T., KUBO, N., AKIDUKI, G., NAKAZONO, M., HIRAI, A. & KADOWAKI, K. 2002. The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. ***Molecular Genetics and Genomics***, 268**,** 434-445.

NUZHDIN, S. V. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. ***Genetica***, 107**,** 129-137.

NYSTEDT, B. R., STREET, N. R., WETTERBOM, A., ZUCCOLO, A., LIN, Y. C., SCOFIELD, D. G., VEZZI, F., DELHOMME, N., GIACOMELLO, S., ALEXEYENKO, A., VICEDOMINI, R., SAHLIN, K., SHERWOOD, E., ELFSTRAND, M., GRAMZOW, L., HOLMBERG, K., HC$LLMAN, J., KEECH, O., KLASSON, L., KORIABINE, M., KUCUKOGLU, M., KC$LLER, M., LUTHMAN, J., LYSHOLM, F., NIITTYLC, T., OLSON, C. K., RILAKOVIC, N., RITLAND, C., ROSSELLC, J. A., SENA, J., SVENSSON, T., TALAVERA-LOPEZ, C., THEIC EN, G. C., TUOMINEN, H., VANNESTE, K., WU, Z. Q., ZHANG, B., ZERBE, P., ARVESTAD, L., BHALERAO, R., BOHLMANN, J., BOUSQUET, J., GARCIA GIL, R., HVIDSTEN, T. R., DE JONG, P., MACKAY, J., MORGANTE, M., RITLAND, K., SUNDBERG, B. R., LEE THOMPSON, S., VAN DE PEER, Y., ANDERSSON, B. R., NILSSON, O., INGVARSSON, P. C. R. K., LUNDEBERG, J. & JANSSON, S. 2013. The Norway spruce genome sequence and conifer genome evolution. ***Nature***, 497**,** 579.

O'DONNELL, K. A. & BURNS, K. H. 2010. Mobilizing diversity: transposable element insertions in genetic variation and disease. ***Mobile DNA***, 1**,** 21-21.

OHNO, S. 1970. *Evolution by gene duplication*, London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

OKI, N., YANO, K., OKUMOTO, Y., TSUKIYAMA, T., TERAISHI, M. & TANISAKA, T. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, *Oryza sativa* ssp. japonica. ***Genes Genetics and Systems***, 83**,** 321-9.

OLIVEIRA, E. J., PC!DUA, J. G., ZUCCHI, M. I., VENCOVSKY, R. & VIEIRA, M. L. C. C. C. 2006. Origin, evolution and genome distribution of microsatellites. ***Genetics and Molecular Biology***, 29**,** 294-307.

OLIVER, K. R., MCCOMB, J. A. & GREENE, W. K. 2013. Transposable elements: Powerful contributors to angiosperm evolution and diversity. ***Genome Biology and Evolution***, 5**,** 1886-1901.

OZYIGIT, I. I., DOGAN, I. & FILIZ, E. 2015. In silico analysis of simple sequence repeats (SSRs) in chloroplast genomes of *Glycine* species. ***Plant Omics***, 8**,** 24-29.

PAN, L., XIA, Q., QUAN, Z., LIU, H., KE, W. & DING, Y. 2010. Development of novel EST-SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. ***Journal of Heredity***., 101**,** 71-82.

PANTZARTZI, C. N., PERGNER, J. & KOZMIK, Z. 2018. The role of transposable elements in functional evolution of amphioxus genome: the case of opsin gene family. ***Scientific Reports***, 8**,** 2506.

PARIDA, S. K., YADAVA, D. K. & MOHAPATRA, T. 2010. Microsatellites in *Brassica* unigenes: relative abundance, marker design, and use in comparative physical mapping and genome analysis. ***Genome***, 53**,** 55-67.

PARK, J. M., SCHNEEWEISS, G. M. & WEISS-SCHNEEWEISS, H. 2007. Diversity and evolution of Ty1-*copia* and Ty3-*gypsy* retroelements in the non-photosynthetic flowering plants Orobanche and Phelipanche (Orobanchaceae). ***Gene***, 387**,** 75-86.

PATHAK, S., GUPTA, K. & DEBROY, R. 1980. Studies on seed polymorphism, germination and seedling growth of *Pongamia pinnata*. ***Indian Journal of Forestry***, 2**,** 64-67.

PATIL, V. K. & NAIK, G. R. 2016. Variability in pod and seed traits of *Pongamia pinnata* Pierre ecotypes in North Karnataka, India. ***Journal of Forestry Research***, 27**,** 557-567.

PATNI, N., PILLAI, S. G. & DWIVEDI, A. H. 2011. Analysis of current scenario of biofuels in India specifically bio-diesel and bio-ethanol. ***Institute of Technology, Nirma University, Ahmedabad.***

PAUX, E., FAURE, S. C. B., CHOULET, F. D. R., ROGER, D., GAUTHIER, V. R., MARTINANT, J. P., SOURDILLE, P., BALFOURIER, F. O., LE PASLIER, M. C., CHAUVEAU, A. L., CAKIR, M., GANDON, B. C. A. & FEUILLET, C. 2010. Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. ***Plant Biotechnology Journal***, 8**,** 196-210.

PAVITHRA, H. R., SHIVANNA, M. B., CHANDRIKA, K., PRASANNA, K. T. & GOWDA, B. 2014. Genetic analysis of *Pongamia pinnata* (L.) Pierre populations using AFLP markers. ***Tree Genetics & Genomes***, 10**,** 173-188.

PEARCE, S. R., HARRISON, G., LI, D., HESLOP-HARRISON, J., KUMAR, A. & FLAVELL, A. J. 1996. The Ty1-copia group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosomal localisation. ***Molecular Genetics and Genomics***, 250**,** 305-15.

PECINKA, A., DINH, H. Q., BAUBEC, T., ROSA, M., LETTNER, N. & MITTELSTEN SCHEID, O. 2010. Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in *Arabidopsis*. ***Plant Cell***, 22, 3118-29.

PELLICER, J., FAY, M. F. & LEITCH, I. J. 2010. The largest eukaryotic genome of them all? ***Botanical Journal of the Linnean Society***, 164**,** 10-15.

PERVAIZ, T., SUN, X., ZHANG, Y., TAO, R., ZHANG, J. & FANG, J. 2015. Association between chloroplast and mitochondrial dna sequences in chinese prunus genotypes (*Prunus persica, Prunus domestica, and Prunus avium*). ***BMC Plant Biology***, 15, 4.

PIEGU, B., GUYOT, R., PICAULT, N., ROULIN, A., SANYAL, A., KIM, H., COLLURA, K., BRAR, D. S., JACKSON, S., WING, R. A. & PANAUD, O. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. ***Genome Research***, 16**,** 1262-9.

PIYA, S., BENNETT, M., RAMBANI, A. & HEWEZI, T. 2017. Transcriptional activity of transposable elements may contribute to gene expression changes in the syncytium formed by cyst nematode in arabidopsis roots. ***Plant Signaling & Behavior***, 12**,** e1362521.

PLOHL, M. 2010. Those mysterious sequences of satellite DNAs. ***Periodicum Biologorum***, 112**,** 403-410.

POHLMAN, R. F., FEDOROFF, N. V. & MESSING, J. 1984. The nucleotide sequence of the maize controlling element *Activator*. ***Cell***, 37**,** 635-643.

PRICE, H. J., DILLON, S. L., HODNETT, G., ROONEY, W. L., ROSS, L. & JOHNSTON, J. S. 2005. Genome evolution in the genus *Sorghum* (Poaceae). ***Annals of Botany***, 95**,** 219-227.

PROVAN, J., SORANZO, N., WILSON, N. J., GOLDSTEIN, D. B. & POWELL, W. 1999. A low mutation rate for chloroplast microsatellites. ***Genetics***, 153**,** 943-7.

QIAN, W., GE, S. & HONG, D. Y. 2001. Genetic variation within and among populations of a wild rice *Oryza granulata* from China detected by RAPD and ISSR markers. ***Theoretical and Applied Genetics***, 102**,** 440-449.

RAJENDRAKUMAR, P., BISWAL, A. K., BALACHANDRAN, S. M. & SUNDARAM, R. M. 2008. In silico analysis of microsatellites in organellar genomes of major cereals for understanding their phylogenetic relationships. ***In Silico Biology*** , 8**,** 87-104.

RAJENDRAKUMAR, P., BISWAL, A. K., BALACHANDRAN, S. M., SRINIVASARAO, K. & SUNDARAM, R. M. 2007. Simple sequence repeats in organellar genomes of rice: Frequency and distribution in genic and intergenic regions. ***Bioinformatics***, 23**,** 1-4.

RAJI, A. A. J., ANDERSON, J. V., KOLADE, O. A., UGWU, C. D., DIXON, A. G. O. & INGELBRECHT, I. L. 2009. Gene-based microsatellites for cassava (*Manihot*

esculenta Crantz): prevalence, polymorphisms, and cross-taxa utility. ***BMC Plant Biology***, 9**,** 118-118.

RAJPUT, M. K. & UPADHYAYA, K. C. 2010. Characterization of heterogeneity in Ty1-copia group retrotransposons in chickpea (*Cicer arietinum* L.). ***Molecular Biology***, 44**,** 529-535.

RAMANATHA RAO, V. & HODGKIN, T. 2002. Genetic diversity and conservation and utilization of plant genetic resources. ***Plant Cell, Tissue and Organ Culture***, 68**,** 1-19.

RAMESH, A. M., BASAK, S., CHOUDHURY, R. R. & RANGAN, L. 2014. Development of flow cytometric protocol for nuclear DNA content estimation and determination of chromosome number in *Pongamia pinnata* L., a valuable biodiesel plant. ***Applied Biochemestry and Biotechnology***, 172, 533-48.

RAO, G. R., SHANKER, A. K., SRINIVAS, I., KORWAR, G. R. & VENKATESWARLU, B. 2011. Diversity and variability in seed characters and growth of *Pongamia pinnata* (L.) Pierre accessions. ***Trees***, 25**,** 725-734.

RAUT, S. S., NARKHEDE, S. S., RANE, A. D. & GUNAGA, R. P. 2011. Seed and fruit variability in *Pongamia Pinnata* (L.) Pierre from Konkan region of Maharashtra. ***Journal of Biodiversity***, 2**,** 27-30.

REBOLLO, R., ROMANISH, M. T. & MAGER, D. L. 2012. Transposable elements: An abundant and natural source of regulatory sequences for host genes. ***Annual Review of Genetics***, 46**,** 21-42.

RHO, M. & TANG, H. 2009. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. ***Nucleic Acids Reserach***, 37, e-143.

RIAR, D. S., RUSTGI, S., BURKE, I. C., GILL, K. S. & YENISH, J. P. 2011. EST-SSR Development from 5 lactuca species and their use in studying genetic diversity among L. serriola Biotypes. ***Journal of Heredity***, 102**,** 17-28.

ROBINS, D. M. & SAMUELSON, L. C. 1992. Retrotransposons and the evolution of mammalian gene expression. ***Genetica***, 86**,** 191-201.

Rohlf, F. J. 1997. NTSYS-pc Version. 2.02i Numerical Taxonomy and Multivariate Analysis System. ***Applied Biostatistics Inc., Exeter Software, Setauket, New York***.

ROSSI, M., ARAUJO, P. G. A. & VAN SLUYS, M. A. 2001. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). ***Genetics and Molecular Biology***, 24**,** 141-146.

RUBIN, B. P., SINGER, S., TSAO, C., DUENSING, A., LUX, M. L., RUIZ, R., HIBBARD, M. K., CHEN, C. J., XIAO, S., TUVESON, D. A., DEMETRI, G. D., FLETCHER, C. D. & FLETCHER, J. A. 2001. KIT activation is a ubiquitous feature of gastrointestinal stromal tumors. ***Cancer Research***, 61**,** 8118-21.

SABLOK, G., PADMA RAJU, G. V., MUDUNURI, S. B., PRABHA, R., SINGH, D. P., BAEV, V., YAHUBYAN, G., RALPH, P. J. & PORTA, N. L. 2015. ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection. ***Database (Oxford)***, 2015, bav084. doi: 10.1093/database/bav084.

SAHOO, D. P., ROUT, G. R., DAS, S., APARAJITA, S. & MAHAPATRA, A. K. 2011. Genotypic variability and correlation studies in pod and seed characteristics of *Pongamia pinnata* (L.) Pierre in Orissa, India. ***International Journal of Forestry Research***, 2011**,** 6.

SAKAI, H., TANAKA, T. & ITOH, T. 2007. Birth and death of genes promoted by transposable elements in *Oryza sativa*. ***Gene***, 392**,** 59-63.

SALINA, E. A., SERGEEVA, E. M., ADONINA, I. G., SHCHERBAN, A. B., BELCRAM, H., HUNEAU, C. & CHALHOUB, B. 2011. The impact of Ty3-*gypsy* group LTR retrotransposons *Fatima* on B-genome specificity of polyploid wheats. ***BMC Plant Biology***, 11**,** 99.

SALVI, S., SPONZA, G., MORGANTE, M., TOMES, D., NIU, X., FENGLER, K. A., MEELEY, R., ANANIEV, E. V., SVITASHEV, S., BRUGGEMANN, E., LI, B., HAINEY, C. F., RADOVIC, S., ZAINA, G., RAFALSKI, J. A., TINGEY, S. V., MIAO, G. H., PHILLIPS, R. L. & TUBEROSA, R. 2007. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. ***Proceedings of the National Academy of Sciences of the United States of America***, 104**,** 11376.

SARASWATHI, S. G. & EZHILARASI, S. 2012. Comparative study on growth, yield and carbon content in *Pongamia pinnata* under water stress and urea supplementation. ***Journal of Environmental Biology***, 33, 579-84.

SARKAR, A., SIM, C., HONG, Y. S., HOGAN, J. R., FRASER, M. J., ROBERTSON, H. M. & COLLINS, F. H. 2003. Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. ***Molecular Genetics and Genomics***, 270**,** 173-80.

SASAKI, N., NISHIZAKI, Y., UCHIDA, Y., WAKAMATSU, E., UMEMOTO, N., MOMOSE, M., OKAMURA, M., YOSHIDA, H., YAMAGUCHI, M., NAKAYAMA, M., OZEKI, Y. & ITOH, Y. 2012. Identification of the glutathione S-transferase gene responsible for flower color intensity in carnations. ***Plant Biotechnology***, 29**,** 223-227.

SATISH, P. V. V. & SUNITA, K. 2017. Antimalarial efficacy of *Pongamia pinnata* (L) Pierre against *Plasmodium falciparum* (3D7 strain) and *Plasmodium berghei* (ANKA). ***BMC Complementary and Alternative Medicine***, 17**,** 458.

SATOH, M., KUBO, T., NISHIZAWA, S., ESTIATI, A., ITCHODA, N. & MIKAMI, T. 2004. The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. ***Molecular Genetics and Genomics***, 272**,** 247-256.

SAVADI, S. B., FAKRUDIN, B., NADAF, H. L. & GOWDA, M. V. C. 2012. Transferability of sorghum genic microsatellite markers to peanut. ***American Journal of Plant Sciences***, 3**,** 4.

SCHNABLE, P. S., WARE, D., FULTON, R. S., STEIN, J. C., WEI, F., PASTERNAK, S., LIANG, C., ZHANG, J., FULTON, L., GRAVES, T. A., MINX, P., REILY, A. D., COURTNEY, L., KRUCHOWSKI, S. S., TOMLINSON, C., STRONG, C., DELEHAUNTY, K., FRONICK, C., COURTNEY, B., ROCK, S. M., BELTER, E., DU, F., KIM, K., ABBOTT, R. M., COTTON, M., LEVY, A., MARCHETTO, P., OCHOA, K., JACKSON, S. M., GILLAM, B., CHEN, W., YAN, L., HIGGINBOTHAM, J., CARDENAS, M., WALIGORSKI, J., APPLEBAUM, E., PHELPS, L., FALCONE, J., KANCHI, K., THANE, T., et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. ***Science***, 326**,** 1112-5.

SCHUG, M. D., HUTTER, C. M., WETTERSTRAND, K. A., GAUDETTE, M. S., MACKAY, T. F. & AQUADRO, C. F. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. ***Molecular Biology and Evolution***, 15**,** 1751-1760.

SHANKER, A. 2013. Identification of microsatellites in chloroplast genome of *Anthoceros formosae*. *191*. Bonn: Universität Bonn, Arbeitsgruppe Bryologie.

SHARMA, S. S., ISLAM, M. A., NEGI, M. S. & TRIPATHI, S. B. 2017. Estimation of outcrossing rates in biodiesel species *Pongamia pinnata* based on AFLP and microsatellite markers. ***National Academy Science Letters***, 40**,** 105-108.

SHARMA, S. S., NEGI, M. S., SINHA, P., KUMAR, K. & TRIPATHI, S. B. 2011. Assessment of genetic diversity of biodiesel species *Pongamia pinnata* accessions using AFLP and three endonuclease-AFLP. ***Plant Molecular Biology Reporter***, 29**,** 12-18.

SHEN, S., LIN, L., CAI, J. J., JIANG, P., KENKEL, E. J., STROIK, M. R., SATO, S., DAVIDSON, B. L. & XING, Y. 2011. Widespread establishment and regulatory impact of *Alu* exons in human genes. ***Proceedings of the National Academy of Sciences of the United States of America***, 108**,** 2837-42.

SINGER, S. S., MANNEL, D. N., HEHLGANS, T., BROSIUS, J. & SCHMITZ, J. 2004. From "junk" to gene: curriculum vitae of a primate receptor isoform gene. ***Journal of Molecular Biology***, 341**,** 883-6.

SINGH, A., NIRALA, N. K., NARULA, A., DAS, S. & SRIVASTAVA, P. S. 2011. Isolation and characterization of Ty1-*copia* group of LTRs in genome of three species of *Datura: D. innoxia, D. stramonium* and *D. metel*. ***Physiology and Molecular Biology of Plants***, 17**,** 255-61.

SMIT, A. F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. ***Current Opinion in Genetics and Development***, 9**,** 657-63.

SMITH, A. M., HANSEY, C. N. & KAEPPLER, S. M. 2012. TCUP: A novel *hAT* transposon active in maize tissue culture. ***Frontiers in Plant Science***, 3, 6.

SNEATH, P. H. A. AND SOKAL, R. R. 1973. Numerical taxonomy: The principles and pratice of numerical classification. ***San Francisco: Freeman***, 573.

SONAH, H., DESHMUKH, R. K., SHARMA, A., SINGH, V. P., GUPTA, D. K., GACCHE, R. N., RANA, J. C., SINGH, N. K. & SHARMA, T. R. 2011. Genome-wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium*. ***PLoS One***, 6**,** e21298.

SOREK, R., AST, G. & GRAUR, D. 2002. *Alu*-Containing exons are alternatively spliced. ***Genome Research***, 12**,** 1060-1067.

SORKHEH, K., PRUDENCIO, A. S., GHEBINEJAD, A., DEHKORDI, M. K., EROGUL, D., RUBIO, M. & MARTINEZ-GOMEZ, P. 2016. In silico search, characterization and validation of new EST-SSR markers in the genus *Prunus*. ***BMC Research Notes***, 9**,** 336.

SREEHARSHA, R. V., MUDALKAR, S., SINGHA, K. T. & REDDY, A. R. 2016. Unravelling molecular mechanisms from floral initiation to lipid biosynthesis in a promising biofuel tree species, *Pongamia pinnata* using transcriptome analysis. ***Scientific Reports***, 6**,** 34315.

STAGINNUS, C., HUETTEL, B., DESEL, C., SCHMIDT, T. & KAHL, G. 2001. A PCR-based assay to detect *En/Spm*-like transposon sequences in plants. ***Chromosome Research***, 9**,** 591-605.

STEINBAUEROVC, V., NEUMANN, P., NOVC!K, P. & MACAS, J. C. 2011. A widespread occurrence of extra open reading frames in plant Ty3/*gypsy* retrotransposons. ***Genetica***, 139**,** 1543-1555.

STERGIOU, G., KATSIOTIS, A., HAGIDIMITRIOU, M. & LOUKAS, M. 2002. Genomic and chromosomal organization of Ty1-*copia*-like sequences in *Olea europaea* and evolutionary relationships of *Olea* retroelements. ***Theoretical and Applied Genetics***, 104**,** 926-933.

STUDER, A., ZHAO, Q., ROSS-IBARRA, J. & DOEBLEY, J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. ***Nature Genetics***, 43**,** 1160.

SUJATHA, K., RAJWADE, A. V., GUPTA, V. S. & SULEKHA, H. 2010. Assessment of *Pongamia pinnata* (L.) - a biodiesel producing tree species using ISSR markers. ***Current Science***, 99**,** 1327-1329.

SUNIL, N., KUMAR, V., SIVARAJ, N., LAVANYA, C., PRASAD R. B, N., RAO B. V. S, K. & VARAPRASAD K, S. 2010. Variability and divergence in *Pongamia pinnata* (L.) Pierre germplasm – a candidate tree for biodiesel. ***GCB Bioenergy***, 1**,** 382-391.

SYED, N. H., SURESHSUNDAR, S., WILKINSON, M. J., BHAU, B. S., CAVALCANTI, J. J. V. & FLAVELL, A. J. 2005. Ty1-*copia* retrotransposon-based SSAP marker development in cashew (*Anacardium occidentale* L.). ***Theoretical and Applied Genetics***, 110**,** 1195-1202.

TADEGE, M., DUPUIS, I. & KUHLEMEIER, C. 1999. Ethanolic fermentation: new functions for an old pathway. ***Trends in Plant Science***, 4**,** 320-325.

TAKAHASHI, R., MORITA, Y., NAKAYAMA, M., KANAZAWA, A. & ABE, J. 2012. An active CACTA-family transposable element is responsible for flower variegation in wild soybean *Glycine soja*. ***The Plant Genome,*** 62-70.

TALARICO, L. A., INGRAM, L. O. & MAUPIN-FURLOW, J. A. 2001. Production of the gram-positive Sarcina ventriculi pyruvate decarboxylase in *Escherichia coli*. ***Microbiology***, 147**,** 2425-35.

TAMBARUSSI, E. V., MELOTTO-PASSARIN, D. M., GUIDETTI GONZALEZ, S., BISSOLOTI BRIGATI, J., JESUS, A. D. F., BARBOSAANDRÉ, L., DRESSANO, K. & CARRER, H. 2009. In silico analysis of Simple Sequence Repeats from chloroplast genomes of Solanaceae species. ***Crop Breeding and Applied Biotechnology***, 9**,** 344-352.

TANG, S., OKASHAH, R. A., CORDONNIER-PRATT, M. M., PRATT, L. H., ED JOHNSON, V., TAYLOR, C. A., ARNOLD, M. L. & KNAPP, S. J. 2009. EST and EST-SSR marker resources for *Iris*. ***BMC Plant Biology***, 9**,** 72.

TARAILO-GRAOVAC, M. & CHEN, N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. ***Current Protocols in Bioinformatics***, 25**,** 4.10.1-4.10.14.

TAUTZ, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. ***Nucleic Acids Research***, 17**,** 6463-71.

TAUTZ, D. 1993. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. ***EXS.*** , 67**,** 21-8.

TEMNYKH, S., DECLERCK, G., LUKASHOVA, A., LIPOVICH, L., CARTINHOUR, S. & MCCOUCH, S. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. ***Genome Research*** , 11**,** 1441-52.

TENA GASHAW, E., MEKBIB, F. & AYANA, A. 2016. Genetic diversity among sugarcane genotypes based on qualitative traits. ***Advances in Agriculture***, 2016**,** 8.

TENAILLON, M. I., HOLLISTER, J. D. & GAUT, B. S. 2010. A triptych of the evolution of plant transposable elements. ***Trends Plant Science***, 15**,** 471-8.

TEO, C. H., TAN, S. H., HO, C. L., FARIDAH, Q. Z., OTHMAN, Y. R., HESLOP-HARRISON, J. S., KALENDAR, R. & SCHULMAN, A. H. 2005. Genome constitution and classification using retrotransposon-based markers in the orphan crop banana. ***Journal of Plant Biology***, 48**,** 96-105.

THIEL, T., MICHALEK, W., VARSHNEY, R. & GRANER, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). ***Theoretical and Applied Genetics***, 106**,** 411-422.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. 1994. Clustal-W improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. ***Nucleic Acids Research***, 22**,** 4673-4680.

THOMPSON, J. M. & SALIPANTE, S. J. 2009. PeakSeeker: a program for interpreting genotypes of mononucleotide repeats. ***BMC Research Notes***, 2**,** 17-17.

THUDI, M., MANTHENA, R., WANI, S. P., TATIKONDA, L., HOISINGTON, D. A. & VARSHNEY, R. K. 2010. Analysis of genetic diversity in Pongamia [*Pongamia pinnata* (L) Pierrre] using AFLP Markers. ***Journal of Plant Biochemistry and Biotechnology***, 19**,** 209-216.

THUILLET, A. C. C. L., BRU, D., DAVID, J., ROUMET, P., SANTONI, S., SOURDILLE, P. & BATAILLON, T. 2002. Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, *Triticum turgidum* (L.) Thell. ssp durum desf. ***Molecular Biology and Evolution***, 19**,** 122-125.

TIAN, Z., RIZZON, C., DU, J., ZHU, L., BENNETZEN, J. L., JACKSON, S. A., GAUT, B. S. & MA, J. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? ***Genome Research***, 19**,** 2221-2230.

TIAN, Z., ZHANG, F., LIU, H., GAO, Q. & CHEN, S. 2016. Development of SSR markers for a Tibetan medicinal plant, *Lancea tibetica* (Phrymaceae), based on RAD sequencing. ***Applications in Plant Scieinces***, 4, apps.1600076.

TINGEY, S. V. & DEL TUFO, J. P. 1993. Genetic analysis with random amplified polymorphic DNA markers. ***Plant Physiology***, 101**,** 349-352.

TITTEL-ELMER, M., BUCHER, E., BROGER, L., MATHIEU, O., PASZKOWSKI, J. & VAILLANT, I. 2010. Stress-induced activation of heterochromatic transcription. ***Plos Genetics***, 6**,** e1001175.

TOTH, G., GASPARI, Z. & JURKA, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. ***Genome Research***, 10**,** 967-981.

TRIZZINO, M., PARK, Y., HOLSBACH-BELTRAME, M., ARACENA, K., MIKA, K., CALISKAN, M., PERRY, G. H., LYNCH, V. J. & BROWN, C. D. 2017. Transposable elements are the primary source of novelty in primate gene regulation. ***Genome Research***, 27**,** 1623-1633.

TSUCHIYA, T. & EULGEM, T. 2013. An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. ***Proceedings of the National Academy of Sciences of the United States of America***, 110**,** E3535.

TSYKUN, T., RELLSTAB, C., DUTECH, C., SIPOS, G. & PROSPERO, S. 2017. Comparative assessment of SSR and SNP markers for inferring the population genetic structure of the common fungus *Armillaria cepistipes*. ***Heredity***, 119**,** 371.

TURCOTTE, K., SRINIVASAN, S. & BUREAU, T. 2001. Survey of transposable elements from rice genomic sequences. ***Plant Journal***, 25**,** 169-79.

UCHIYAMA, T., HIURA, S., EBINUMA, I., SENDA, M., MIKAMI, T., MARTIN, C. & KISHIMA, Y. 2012. A pair of transposons coordinately suppresses gene expression, independent of pathways mediated by siRNA in *Antirrhinum*. ***New Phytologist***, 197**,** 431-440.

UL HAQ, S., KUMAR, P., SINGH, R. K., VERMA, K. S., BHATT, R., SHARMA, M., KACHHWAHA, S. & KOTHARI, S. L. 2016. Assessment of Functional EST-SSR Markers (Sugarcane) in Cross-Species Transferability, Genetic Diversity among Poaceae Plants, and Bulk Segregation Analysis. ***Genetics Research Intertional***, 2016**,** 7052323.

VAN DE LAGEMAAT, L. N., LANDRY, J. R., MAGER, D. L. & MEDSTRAND, P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. ***Trends Genetics***, 19**,** 530-6.

VARSHNEY, R. K., GRANER, A. & SORRELLS, M. E. 2005. Genic microsatellite markers in plants: features and applications. ***Trends in Biotechnology***, 23**,** 48-55.

VICIENT, C. M. 2010. Transcriptional activity of transposable elements in maize. ***BMC Genomics***, 11**,** 601.

VICIENT, C. M., JAASKELAINEN, M. J., KALENDAR, R. & SCHULMAN, A. H. 2001. Active retrotransposons are a common feature of grass genomes. ***Plant Physiology***, 125**,** 1283-92.

VIEIRA, M. L. C., SANTINI, L., DINIZ, A. L. & MUNHOZ, C. D. F. 2016. Microsatellite markers: what they mean and why they are so useful. ***Genetics and Molecular Biology***, 39**,** 312-328.

VISMAYA, BELAGIHALLY, S. M., RAJASHEKHAR, S., JAYARAM, V. B., DHARMESH, S. M. & THIRUMAKUDALU, S. K. C. 2011. Gastroprotective properties of karanjin from karanja (*Pongamia pinnata*) seeds; role as antioxidant and H(+), K(+)-ATPase inhibitor. ***Evidence-Based Complementary and Alternative Medicine***, 2011, 747246.

VITTE, C. M., FUSTIER, M. A., ALIX, K. & TENAILLON, M. I. 2014. The bright side of transposons in crop evolution. ***Briefings in Functional Genomics***, 13**,** 276-295.

VON STERNBERG, R. & SHAPIRO, J. A. 2005. How repeated retroelements format genome function. ***Cytogenetic and Genome Research***, 110**,** 108-116.

VORECHOVSKY, I. 2010. Transposable elements in disease-associated cryptic exons. ***Human Genetics***, 127**,** 135-54.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., VAN DE LEE, T., HORNES, M., FRIJTERS, A., POT, J., PELEMAN, J., KUIPER, M. & ET AL. 1995. AFLP: a new technique for DNA fingerprinting. ***Nucleic Acids Research***, 23**,** 4407-14.

WANG, F., TONG, Z., SUN, J., SHEN, Y., ZHOU, J., GAO, Z. & ZHANG, Z. 2010. Genome-wide detection of Ty1-*copia* and Ty3-*gypsy* group retrotransposons in Japanese apricot (*Prunus mume* Sieb. et Zucc.). ***African Journal of Biotechnology***, 9**,** 8583-8596.

WANG, W., WU, Y. & MESSING, J. 2012. The Mitochondrial Genome of an Aquatic Plant, *Spirodela polyrhiza*. ***PLoS One***, 7**,** e46747.

WANG, Y., XIE, H., YANG, Y., HUANG, Y., WANG, J. & TAN, F. 2017. Chloroplast and mitochondrial microsatellites for *Millettia pinnata* (Fabaceae) and cross-amplification in related species. ***Applications in Plant Sciences***, 5**,** apps.1700034.

WANG, Z., YU, G., SHI, B., WANG, X., QIANG, H. & GAO, H. 2014. Development and characterization of simple sequence repeat (SSR) markers based on rna-sequencing of *Medicago sativa* and in silico mapping onto the *M. truncatula* genome. ***PLoS One***, 9**,** e92029.

WARNEFORS, M., PEREIRA, V. & EYRE-WALKER, A. 2010. Transposable elements: Insertion pattern and impact on gene expression evolution in hominids. ***Molecular Biology and Evolution***, 27**,** 1955-1962.

WEGRZYN, J. L., WHALEN, J., KINLAW, C. S., HARRY, D. E., PURYEAR, J., LOOPSTRA, C. A., GONZALEZ-IBEAS, D., VASQUEZ-GROSS, H. A., FAMULA, R. A. & NEALE, D. B. 2016. Transcriptomic profile of leaf tissue from the leguminous tree, *Millettia pinnata*. ***Tree Genetics & Genomes***, 12**,** 44.

WEI, L. & CAO, X. 2016. The effect of transposable elements on phenotypic variation: insights from plants to humans. ***Science China Life Sciences***, 59**,** 24-37.

WENDEL, J. F., JACKSON, S. A., MEYERS, B. C. & WING, R. A. 2016. Evolution of plant genome architecture. ***Genome Biology***, 17**,** 37.

WENKE, T., DOBEL, T., SORENSEN, T. R., JUNGHANS, H., WEISSHAAR, B. & SCHMIDT, T. 2011. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. ***Plant Cell***, 23**,** 3117-28.

WESSLER, S. R. 1996. Plant retrotransposons: Turned on by stress. ***Current Biology***, 6**,** 959-961.

WEST, P. T., LI, Q., JI, L., EICHTEN, S. R., SONG, J., VAUGHN, M. W., SCHMITZ, R. J. & SPRINGER, N. M. 2014. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. ***PLoS One***, 9**,** e105267.

WICKER, T., SABOT, F., HUA-VAN, A., BENNETZEN, J. L., CAPY, P., CHALHOUB, B., FLAVELL, A., LEROY, P., MORGANTE, M., PANAUD, O., PAUX, E., SANMIGUEL, P. & SCHULMAN, A. H. 2007. A unified classification system for eukaryotic transposable elements. ***Nature Reviews Genetics***, 8, 973-82.

WOLFE, A. D. & RANDLE, C. P. 2004. Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: implications for plant molecular systematics. ***Systematic Botany***, 29**,** 1011-1020.

WOODROW, P., CIARMIELLO, L. F., FANTACCIONE, S., ANNUNZIATA, M. G., PONTECORVO, G. & CARILLO, P. 2012. Ty1-*copia* group retrotransposons and the evolution of retroelements in several angiosperm plants: evidence of horizontal transmission. ***Bioinformation***, 8**,** 267-271.

WRIGHT, D. A., KE, N., SMALLE, J., HAUGE, B. M., GOODMAN, H. M. & VOYTAS, D. F. 1996. Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana*. ***Genetics***, 142**,** 569-78.

WU, B. & HAO, W. 2015. A dynamic mobile dna family in the yeast mitochondrial genome. ***G3: Genes|Genomes|Genetics***, 5**,** 1273.

XIAO, H., JIANG, N., SCHAFFNER, E., STOCKINGER, E. J. & VAN DER KNAAP, E. 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. ***Science***, 319**,** 1527.

XIAO, W., SU, Y., SAKAMOTO, W. & SODMERGEN 2007. Isolation and characterization of Ty1/*copia*-like retrotransposons in mung bean (*Vigna radiata*). ***Journal of Plant Research***, 120**,** 323-328.

XIAO, Y., ZHOU, L., XIA, W., MASON, A. S., YANG, Y., MA, Z. & PENG, M. 2014. Exploiting transcriptome data for the development and characterization of gene-based SSR markers related to cold tolerance in oil palm (*Elaeis guineensis*). ***BMC Plant Biology***, 14**,** 384.

XIONG, W., HE, L., LAI, J., DOONER, H. K. & DU, C. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. ***Proceedings of the National Academy of Sciences of the United States of America***, 111**,** 10263.

XU, J., LIU, L., XU, Y., CHEN, C., RONG, T., ALI, F., ZHOU, S., WU, F., LIU, Y., WANG, J., CAO, M. & LU, Y. 2013. Development and characterization of simple sequence repeat markers providing genome-wide coverage and high resolution in maize. ***DNA Research***, 20**,** 497-509.

XU, Z. & WANG, H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. ***Nucleic Acids Research***, 35**,** W265-W268.

YADAV, C. B., BONTHALA, V. S., MUTHAMILARASAN, M., PANDEY, G., KHAN, Y. & PRASAD, M. 2015. Genome-wide development of transposable elements-based markers in foxtail millet and construction of an integrated database. ***DNA Research***, 22**,** 79-90.

YADAV, H. K., ALOK, R., ASIF, M. H., SHRIKANT, M., SAWANT, S. V. & RAKESH, T. 2011. EST-derived SSR markers in *Jatropha curcas* L.: development, characterization, polymorphism, and transferability across the species/genera. ***Tree Genetics and Genomes***, 7**,** 207-219.

YAN, L., GU, Y. H., TAO, X., LAI, X. J., ZHANG, Y. Z., TAN, X. M. & WANG, H. 2014. Scanning of transposable elements and analyzing expression of transposase genes of sweet potato [*Ipomoea batatas*]. ***PLoS One,*** 9, e90895.

YANG, L. & BENNETZEN, J. L. 2009. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. ***Proceedings of the National Academy of Sciences of the United States of America***, 106**,** 19922-7.

YANG, Q., LI, Z., LI, W., KU, L., WANG, C., YE, J., LI, K., YANG, N., LI, Y., ZHONG, T., LI, J., CHEN, Y., YAN, J., YANG, X. & XU, M. 2013. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. ***Proceedings of the National Academy of Sciences of the United States of America***, 110**,** 16969-74.

YAO, J., XIAOYU, L., CHENG, P., YEYUN, L., JIAYUE, J. & CHANGJUN, J. 2017. Cloning and analysis of reverse transcriptases from Ty1-*copia* retrotransposons in *Camellia sinensis*. ***Biotechnology & Biotechnological Equipment***, 31**,** 663-669.

YASUDA, K., ITO, M., SUGITA, T., TSUKIYAMA, T., SAITO, H., NAITO, K., TERAISHI, M., TANISAKA, T. & OKUMOTO, Y. 2013. Utilization of transposable element mPing as a novel genetic tool for modification of the stress response in rice. ***Molecular Breeding***, 32**,** 505-516.

YEPHREMOV, A. & SAEDLER, H. 2001. Display and isolation of transposon-flanking sequences starting from genomic DNA or RNA. ***The Plant Journal***, 21**,** 495-505.

YOU, F. M., HUO, N., GU, Y. Q., LUO, M.-C., MA, Y., HANE, D., LAZO, G. R., DVORAK, J. & ANDERSON, O. D. 2008. BatchPrimer3: A high throughput web application for PCR and sequencing primer design. ***BMC Bioinformatics***., 9**,** 253.

ZABALA, G. & VODKIN, L. O. 2005. The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. ***The Plant Cell***, 17**,** 2619.

ZAKI, E. A. 2005. Ty1-copia group retrotransposon families in cultivated cottons *G. barbadense* L. identified by reverse transcriptase domain analysis. ***DNA Sequence***, 16**,** 288-294.

ZEDEK, F. E., E MERDA, J., E MARDA, P. & BUREE, P. 2010. Correlated evolution of LTR retrotransposons and genome size in the genus eleocharis. ***BMC Plant Biology***, 10**,** 265.

ZHANG, J. M., LIU J., SUN, H. L., SUN H. L., YU, J., WANG, J. X., ZHOU, S. L. & ZHOU, S. L. 2011. Nuclear and chloroplast SSR markers in *Paeonia delavayi* (Paeoniaceae) and cross-species amplification in *P. ludlowii*. ***American Journal Botany***, 98, e346-8.

ZHANG, K., WU, Z., TANG, D., LV, C., LUO, K., ZHAO, Y., LIU, X., HUANG, Y. & WANG, J. 2016. Development and identification of SSR markers associated with starch properties and N2-carotene content in the storage root of sweet potato (*Ipomoea batatas* L.). ***Frontiers in Plant Science***, 7, 223.

ZHANG, L., YUAN, D., YU, S., LI, Z., CAO, Y., MIAO, Z., QIAN, H. & TANG, K. 2004. Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. ***Bioinformatics***, 20**,** 1081-1086.

ZHANG, L., ZUO, K., ZHANG, F., CAO, Y., WANG, J., ZHANG, Y., SUN, X. & TANG, K. 2006. Conservation of noncoding microsatellites in plants: implication for gene regulation. ***BMC Genomics***, 7**,** 323.

ZHANG, S., TANG, C., ZHAO, Q., LI, J., YANG, L., QIE, L., FAN, X., LI, L., ZHANG, N., ZHAO, M., LIU, X., CHAI, Y., ZHANG, X., WANG, H., LI, Y., LI, W., ZHI, H., JIA, G. & DIAO, X. 2014. Development of highly polymorphic simple sequence repeat markers using genome-wide microsatellite variant analysis in Foxtail millet [*Setaria italica* (L.) P. Beauv.]. ***BMC Genomics***, 15**,** 78.

ZHANG, Y., ROMANISH, M. T. & MAGER, D. L. 2011. Distributions of transposable elements reveal hazardous zones in mammalian introns. ***PLOS Computional Biology***, 7**,** e1002046.

ZHANG, Y., ZHANG, X., WANG, Y. H. & SHEN, S. K. 2017. *De novo* assembly of transcriptome and development of novel EST-SSR markers in *Rhododendron rex* LC)vl. through Illumina sequencing. ***Frontiers in Plant Science***, 8**,** 1664.

ZHAO, X., HUANG, L., ZHANG, X., WANG, J., YAN, D., LI, J., TANG, L., LI, X. & SHI, T. 2016. Construction of high-density genetic linkage map and identification of flowering-time QTLs in orchardgrass using SSRs and SLAF-seq. ***Scientific Reports***, 6**,** 29345.

ZHOU, J., DUDASH, M. R., FENSTER, C. B. & ZIMMER, E. A. 2016. Development of highly variable microsatellite markers for the tetraploid *Silene stellata* (Caryophyllaceae). ***Applications in Plant Sciences***, 4**,** apps.1600117.

ZHU, H., SONG, P., KOO, D. H., GUO, L., LI, Y., SUN, S., WENG, Y. & YANG, L. 2016. Genome wide characterization of simple sequence repeats in watermelon genome and their application in comparative mapping and genetic diversity analysis. ***BMC Genomics***, 17**,** 557.

ZIETKIEWICZ, E., RAFALSKI, A. & LABUDA, D. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. ***Genomics***, 20**,** 176-83.

ZONG, J. W., ZHAO, T. T., MA, Q.H., LIANG, L. S. & WANG, G. X. 2015. Assessment of genetic diversity and population genetic structure of *Corylus mandshurica* in China using ssr markers. ***Frontiers in Plant Science***, 2, 223.

ZORATTI, L., PALMIERI, L., JAAKOLA, L. & HAGGMAN, H. 2015. Genetic diversity and population structure of an important wild berry crop. ***AoB Plants***, 7**,** plv117.

ZOU, C., LU, C., ZHANG, Y. & SONG, G. 2012. Distribution and characterization of simple sequence repeats in *Gossypium raimondii* genome. ***Bioinformation***, 8**,** 801-806.

ZUCCOLO, A., SEBASTIAN, A., TALAG, J., YU, Y., KIM, H., COLLURA, K., KUDRNA, D. & WING, R. A. 2007. Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. ***BMC Evolutionary Biology***, 7, 152.

Publications

Publications

Research Papers

1. DAS, R. [*], **SHELKE**, **R**. **G**[*]., RANGAN, L. & MITRA, S. 2018. Estimation of nuclear genome size and characterization of Ty1-*copia* like LTR retrotransposon in *Mesua ferrea* L. ***Journal of Plant Biochemistry and Biotechnology***, 1-10. doi.org/10.1007/s13562-018-0457-7 ([*] Equal contribution).

2. RAMESH, A. M., SINGH, A., **SHELKE**, **R**. **G**., SCOTT, P. T., GRESSHOFF, P. M. & RANGAN, L. 2016. Identification of two genes encoding microsomal oleate desaturases (FAD2) from the biodiesel plant *Pongamia pinnata* L. ***Trees-Structure and Function***; 30, 1351-1360.

Conferences

1. DAS, R., **SHELKE**, **R**. **G**. & RANGAN, L. 2018. Genome size and Ty1 *copia* retroelements in biofuel crops. ***24th ISCB International Conference (ISCBC-2018)***, held on 11th -13th January, 2018 at Manipal University Jaipur, India.

2. **SHELKE**, **R**. **G**. & RANGAN, L. 2018. Study of expression of repetitive elements and their application for gene linked marker development in *Pongamia pinnata*. ***Genomics Analysis & Technology Conference (GATC 2018)***, held on 8th -9th January, 2018 at Guwahati University, India.

3. **SHELKE**, **R**. **G**. & RANGAN, L. 2017. Identification and characterization of long interspersed nuclear elements (LINEs) in *Pongmia, Ricinus and Jatropha*. ***International Symposium on Plant Biotechnology for Crop Improvement (ISPBCI 2017)***, 20th-22nd January 2017, IIT Guwahati, India.

4. **SHELKE**, **R**. **G**. & RANGAN, L. 2017. Identification and characterization of long interspersed nuclear elements (LINEs) in *Pongamia, Ricinus and Jatropha*. ***Research Conclave'17***, 16th-19th March, 2017, IIT Guwahati, India.

5. DAS, R. & **SHELKE**, **R**. **G**. 2016. Genome mining in potential biofuel crops of North-East India- genome diversity and correlation with cell plasticity. ***Research Conclave'16***, 17th- 20th March 2016, IIT Guwahati, India.

6. SINGH, A., **SHELKE**, **R**. **G**., DAS, R., JAHAN, I., RANGAN, L., KHARE, A. & PANDA, A. N. 2015. Unravelling *Pongamia*- genomic and phylogenetic studies. ***Research Conclave'15***, 16th-19th March, 2015, IIT Guwahati, India.

7. **SHELKE**, **R**. **G**, DAS, R. & RANGAN, L. 2015. Evolution and distribution of LTR retrotransposons in *Pongamia pinnata* and its correlation with genome size. ***Second International Conference on Biotechnology and Bioinformatics (ICBB-2015)***, 5th-8th February 2015, Pune, India.

8. **SHELKE**, **R**. **G**., RAMESH, A. M., RANGAN, L. 2014. Identification and characterization of copia like retrotransposons from the genome of *Pongamia pinnata*. ***National Conference on Perspective and Trends in Plant Sciences and Biotechnology***, 21st-23rd February 2014, Chandigarh, India.

NCBI Genbank submissions

1. *Pongamia*, Ty1-*copia* RT (**KP202847.1- KP202834.1, MH397570- MH397584**).
2. *Ricinus*, Ty1-*copia* RT-RH (**MH397585**).
3. *Jatropha*, Ty1-*copia* RT-RH (**MH397586**).
4. *Mesua*, Ty1-*copia* RT-RH (**KU507530.1**).
5. *Pongamia*, Ty3-*gypsy* RT (**MH397537- MH397566**).
6. *Ricinus*, Ty3-*gypsy* RT-RH (**MH397569**).
7. *Jatropha*, Ty3-*gypsy* RT (**MH397568**).
8. *Mesua*, Ty3-*gypsy* RT (**MH397567**).
9. *Pongamia*, LINE RT (**MH397508- MH397533**).
10. *Ricinus*, LINE RT (**MH397534**).
11. *Jatropha*, LINE RT (**MH397535**).
12. *Mesua*, LINE RT (**MH397536**).