

**Evolutionary Dynamics and Phylogeny of Family
Malvaceae**



By

Abdullah

**Department of Biochemistry
Faculty of Biological Sciences,
Quaid-i-Azam University, Islamabad**

2020

**Evolutionary Dynamics and phylogeny of Family
Malvaceae**



A dissertation in the partial fulfilment of the requirements of the degree of

Doctor of Philosophy

In

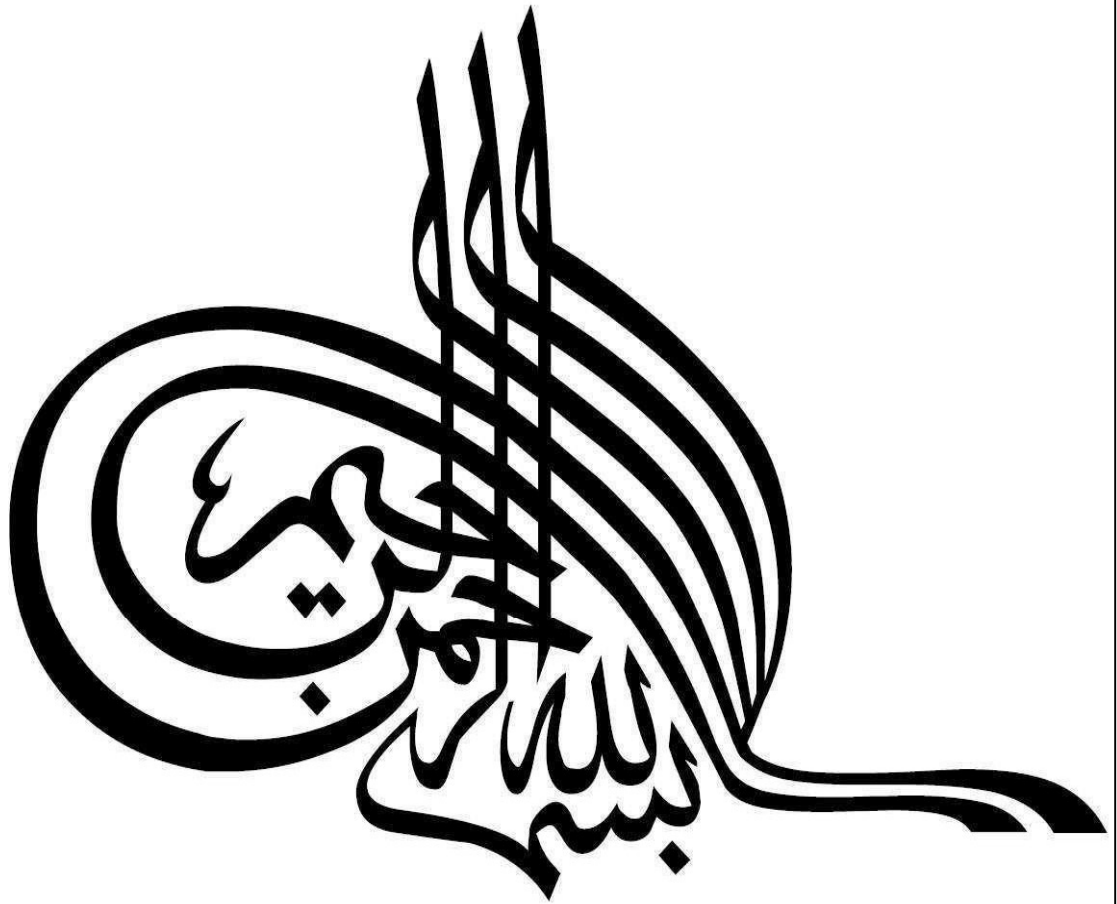
Biochemistry/Molecular Biology

Submitted by

Abdullah

**Department of Biochemistry,
Faculty of Biological Sciences,
Quaid-i-Azam University, Islamabad**

2020



*In the name of Allah,
the Most Beneficent,
the Most Merciful*

Author's Declaration

I **Abdullah** hereby state that my Ph.D thesis, *titled “Evolutionary dynamics and Phylogeny of family Malvaceae”* is my own work and has not been submitted previously by me for taking any degree from **Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan.**

Or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my Ph.D degree.

Abdullah
Date: July 23, 2020

Plagiarism Undertaking

I solemnly declare that research work presented in the Ph.D thesis, titled “**Evolutionary dynamics and Phylogeny of family Malvaceae**” is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and **Quaid-i-Azam University, Islamabad**, towards the plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the right to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student/Author Signature: _____

Abdullah

Date: July 23, 2020

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled “**Evolutionary dynamics and Phylogeny of family Malvaceae**” was conducted by **Mr. Abdullah** under the supervision of Dr. Mohammad Tahir Waheed.

No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan in partial fulfillment of the requirements for the **Degree of Doctor of Philosophy** in the field of Biochemistry from Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan.

Mr. Abdullah

Signature: _____

Examination Committee:

1. External Examiner:

Prof. Dr. Azra Yasmin

Department of Biotechnology

Fatima Jinnah Women University, Rawalpindi

Signature:  _____

2. External Examiner:

Prof. Dr. Ghazala Kaukab

University Institute of Biochemistry & Biotechnology

PMAS Arid Agriculture University, Rawalpindi

Signature:  _____

3. Supervisor:

Dr. Mohammad Tahir Waheed

Signature: _____

4. Co-Supervisor:

Dr. Ibrar Ahmed

Signature: _____

5. Chairman:

Prof. Dr. M. Rashid Khan

Signature: _____

Dated:

July 23, 2020

This thesis is dedicated to

My parents and teachers

I would like to acknowledge all the efforts of my loving parents who has provided all support and motivations for my education. I am also thankful to all my teachers whose teachings and encouragements motivated me to obtain the higher education degree.

CONTENTS

LIST OF FIGURES	i
LIST OF TABLES	iii
LIST OF ABBREVIATIONS	v
ACKNOWLEDGMENT	vii
ABSTRACT	ix

CHAPTER No. 1: Introduction and literature review

1.1	Introduction	1
1.2	Morphological features of Malvaceae	1
1.3	Importance of family Malvaceae	2
1.3.1	Fibre	2
1.3.2	Food	2
1.3.3	Horticulture	3
1.3.4	Medicinal properties	3
1.4	Taxonomic position of family Malvaceae	4
1.5	Nine subfamilies classifications of family Malvaceae	6
1.5.1	Subfamily Bombacoideae	6
1.5.2	Subfamily Brownlowioideae	6
1.5.3	Subfamily Byttnerioideae	6
1.5.4	Subfamily Dombeyoideae	7
1.5.5	Subfamily Grewioideae	7
1.5.6	Subfamily Helicteroideae	7
1.5.7	Subfamily Sterculioideae	7
1.5.8	Subfamily Tilioideae	8
1.5.9	Subfamily Malvoideae	8
1.5.10	Two major clades of Malvaceae	8
1.6	Genera of the family Malvaceae	10
1.6.1	Genus <i>Theobroma</i>	10
1.6.1.1	Genus <i>Theobroma</i> morphology	11
1.6.1.2	Molecular based studies in genus <i>Theobroma</i>	11

1.6.2	Genus <i>Firmiana</i>	11
1.6.2.1	Morphology of genus <i>Firmiana</i>	12
1.6.2.2	Molecular studies in genus <i>Firmiana</i>	12
1.6.3	Genus <i>Hibiscus</i>	12
1.6.3.1	Morphology of Genus <i>Hibiscus</i>	13
1.6.3.2	Plasticity in morphology of <i>Hibiscus</i>	13
1.6.3.3	Sections and segregates of <i>Hibiscus</i>	14
1.6.3.4	Inter-genus and intra-genus taxonomic discrepancies in <i>Hibiscus</i>	14
1.6.3.5	Molecular based studies of genus <i>Hibiscus</i>	15
1.7	General features of chloroplast genome	16
1.8	Chloroplast genome structure	16
1.9	Polymorphism in chloroplast genome sequences	17
1.10	Adaptive evolution and domestications	17
1.11	Transfer of chloroplast gene to nuclear or mitochondrial genomes and vice versa	19
1.12	Role of chloroplast genome in phylogenetic studies	20
1.13	Repeats in chloroplast genome	21
1.14	Comparisons of mutation rates among nuclear, mitochondrial and chloroplast genome	22
1.15	Hypotheses about origination of mutations and their co-occurrence	22
1.16	Status of complete chloroplast genome-based research in the family Malvaceae	23
1.17	Aims and objectives	24

CHAPTER No. 2: *De novo* assembly and characterisation of chloroplast genomes of eight family Malvaceae species

2.1	Introduction	25
2.1.1	<i>Hibiscus rosa-sinensis</i>	26
2.1.2	<i>Hibiscus mutabilis</i>	26
2.1.3	<i>Malva parviflora</i>	26
2.1.4	<i>Malvastrum coromandelianum</i>	27
2.1.5	<i>Urena procumbens</i>	27
2.1.6	<i>Firmiana colorata</i>	28
2.1.7	<i>Sterculia monosperma</i>	28

2.1.8	<i>Pterospermum truncatolobatum</i>	28
2.2	Materials and methods	29
2.2.1	Plant collection	29
2.2.2	DNA extraction	29
2.2.3	Sequencing	30
2.2.4	Downloading of SRA data	30
2.2.5	Chloroplast genomes assembly and annotations	31
2.3	Results	31
2.3.1	DNA extraction	31
2.3.2	Whole genome shotgun and chloroplast genome assembly	32
2.3.3	Chloroplast genome feature	32
2.3.4	Chloroplast genome features of <i>Hibiscus rosa-sinensis</i>	33
2.3.5	Chloroplast genome features of <i>Hibiscus mutabilis</i>	36
2.3.6	Chloroplast genome features of <i>Malva parviflora</i>	36
2.3.7	Chloroplast genome features of <i>Malvastrum coromandelianum</i>	39
2.3.8	Chloroplast genome features of <i>Urena Procumbens</i>	39
2.3.9	Chloroplast genome features of <i>Firmiana colorata</i>	42
2.3.10	Chloroplast genome features of <i>Sterculia monosperma</i>	42
2.3.11	Chloroplast genome features of <i>Pterospermum truncatolobatum</i>	45
2.4	Conclusion	46

CHAPTER No. 3: Comparative analyses of chloroplast genomes of family Malvaceae and rate of evolution in protein coding genes

3.1	Introduction	47
3.2	Materials and Methods	49
3.2.1	Genomics features of chloroplast genome in family Malvaceae	49
3.2.2	Rate of evolution of protein coding genes	49
3.3	Results	49
3.3.1	Genome structure and gene content	49
3.3.2	Length of the chloroplast genome and GC content	50
3.3.3	IR contraction and expansion	52
3.3.4	Comparative analyses of evolutionary rate of protein coding genes	55
3.4	Conclusion	71

CHAPTER No. 4: Correlations among oligonucleotide repeats, nucleotide substitutions and insertion – deletion mutations in chloroplast genomes of plant family Malvaceae

4.1	Introduction	72
4.2	Materials and methods	73
4.2.1	Correlations among substitutions, InDels and oligonucleotide repeats	73
4.3	Results	75
4.3.1	Correlations among substitutions, InDels and repeats	76
4.3.1.1	Correlations and regressions among mutational events at family level	76
4.3.1.2	Correlations and regressions among mutational events at genus level	78
4.3.2	Correlations and regression analyses of substitutions with SSR and non-SSR InDels	78
4.3.3	Correlations and regression analyses of repeats with SSR and non-SSR InDels	80
4.4	Conclusion	81

CHAPTER No. 5: Comparative analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*

5.1	Introduction	82
5.2	Materials and Methods	83
5.2.1	Reannotation and comparative analyses of chloroplast genomes	83
5.2.2	Repeats analysis in <i>Theobroma</i> chloroplast genomes	83
5.2.3	Substitutions and InDels analysis	84
5.3	Results	84
5.3.1	Genome organisation and features of <i>Theobroma</i> species	84
5.3.2	Amino acid frequency and codon usage	84
5.3.3	Putative RNA editing sites	87
5.3.4	Analyses of simple sequence repeats and oligonucleotide repeats	89
5.3.5	Substitutions and InDels analysis in chloroplast genomes of <i>Theobroma cacao</i> and <i>Theobroma grandiflorum</i>	97
5.3.6	Highly polymorphic regions between <i>Theobroma</i> species	97
5.4	Conclusion	99

CHAPTER No. 6: Comparative analyses of chloroplast genomes of *Firmiana colorata*, *Firmiana major* and *Firmiana pulcherrima*

6.1	Introduction	100
6.2	Materials and Methods	101
6.2.1	Chloroplast genome comparative analysis	101
6.2.2	Amino acids frequency, codon usage and putative RNA editing sites	101
6.2.3	Microsatellites and oligonucleotide repeats analysis	101
6.2.4	Identification of mutational hotspots	101
6.3	Results	102
6.3.1	Chloroplast genome organisation and features of <i>Firmiana</i>	102
6.3.2	Comparative analysis of amino acids frequencies and relative synonymous codon usage	102
6.3.3	Analyses of RNA editing sites in <i>Firmiana</i> genomes	105
6.3.4	SSRs and oligonucleotide repeats analyses	110
6.3.5	Substitutions and InDels analysis in <i>Firmiana</i>	123
6.3.6	Mutational hotspots in <i>Firmiana</i>	125
6.4	Conclusion	126

CHAPTER No. 7: Comparative analyses of *Hibiscus rosa-sinensis* and *Hibiscus syriacus* and screening of mutational hotspots

7.1	Introduction	127
7.2	Materials and Methods	128
7.2.1	Comparison among three <i>Hibiscus syriacus</i> chloroplast genomes	128
7.2.2	Codon usage, amino acid frequency and RNA editing sites	128
7.2.3	Analysis of InDels and substitution types	128
7.2.4	Analysis of repeats in <i>Hibiscus</i>	128
7.2.5	Screening of divergence regions	129
7.3	Results	129
7.3.1	Comparison among three available chloroplast genomes of <i>H. syriacus</i>	129
7.3.2	Comparative analyses of the genomic features between <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	129
7.3.3	Amino acids frequency and codon usage analyses	130
7.3.4	Analyses of RNA editing sites in <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	132
7.3.5	Analyses of repeats in <i>Hibiscus</i>	134

7.3.6	Analysis of substitutions and InDels in genus <i>Hibiscus</i>	145
7.3.7	Mutational hotspots in <i>Hibiscus</i> species	145
7.4	Conclusion	148

CHAPTER No. 8: Phylogenetic inference in the family Malvaceae

8.1	Introduction	149
8.2	Materials and methods	150
8.2.1	Reconstruction of phylogenetic tree	150
8.3	Results	151
8.3.1	Phylogenetic analyses based on protein coding genes	151
8.3.2	Role of IR contraction and expansion in phylogeny	156
8.3.3	Efficacy of complete chloroplast genome and LSC, SSC and IR regions in inferring of phylogeny	156
8.4	Conclusion	160

CHAPTER No. 9: Discussion

9.1	Discussion	161
9.2	<i>De novo</i> assembly of chloroplast genomes of eight Malvaceae species	162
9.3	Comparative analyses of chloroplast genomes among Malvaceae species	163
9.4	IRs contraction and expansion and its role in reduction and duplication of genes	164
9.5	Rate of synonymous and non-synonymous substitutions and adaptive evolution	164
9.6	Mutational dynamics in family Malvaceae and correlations among substitutions, InDels and repeats	165
9.7	Role of the contraction and expansion of inverted repeats in inference of phylogeny	167
9.8	Phylogenetic analysis of family Malvaceae	168
9.9	Comparison of species of three genera of Malvaceae	170
9.10	Codon usage analyses and its link to evolution of species	171
9.11	RNA editing sites	171
9.12	Simple sequence repeats	172
9.13	Oligonucleotide repeats	173
9.14	Substitutions and InDels in chloroplast genome	174
9.15	Mutational hotspots regions	174

9.15.1	Genus <i>Theobroma</i>	175
9.15.2	Genus <i>Firmiana</i>	176
9.15.3	Genus <i>Hibiscus</i>	176
	Conclusions	178
	Future perspectives	179
	References	180
	List of publications	212

LIST OF FIGURES

Sr. No.	Title	Page No.
Figure 1.1	Phylogenetic relationship among the subfamily of Malvaceae	9
Figure 2.1	Genomic DNA extraction	32
Figure 2.2	Circular map of chloroplast genome of <i>Hibiscus rosa-sinensis</i> with annotated genes	35
Figure 2.3	Circular map of chloroplast genome of <i>Hibiscus mutabilis</i> with annotated genes	37
Figure 2.4	Circular map of chloroplast genome of <i>Malva parviflora</i> with annotated genes	38
Figure 2.5	Circular map of chloroplast genome of <i>Malvastrum coromandelianum</i> with annotated genes	40
Figure 2.6	Circular map of chloroplast genome of <i>Urena procumbens</i> with annotated genes	41
Figure 2.7	Circular map of chloroplast genome of <i>Firmiana colorata</i> with annotated genes	43
Figure 2.8	Circular map of chloroplast genome of <i>Sterculia monosperma</i> with annotated genes	44
Figure 2.9	Circular map of chloroplast genome of <i>Pterospermum truncatolobatum</i> with annotated genes	45
Figure 3.1	Comparative analyses of boundary regions: inverted repeat regions (IR), small single copy (SSC), and large single copy (LSC) among 20 species of Malvaceae.	53
Figure 4.1	Methodology of SNPs, InDels and repeats counting for correlation	75
Figure 4.2	Represents the correlations and regression values at family level comparison	77
Figure 4.3	Represents the correlations and regression values at genus level comparison	79
Figure 5.1	Frequency of amino acids in <i>T. cacao</i> and <i>T. grandiflorum</i>	85
Figure 5.2	Comparison of repeats in <i>Theobroma cacao</i> and <i>Theobroma grandiflorum</i>	91
Figure 5.3	Types of substitutions between <i>T. cacao</i> and <i>T. grandiflorum</i>	97
Figure 5.4	Nucleotide diversity of the chloroplast genome regions	98
Figure 6.1	Comparison of amino acids frequency among <i>Firmiana</i> species	103
Figure 6.2	Comparison of SSRs and oligonucleotide repeats in <i>Firmiana</i>	115
Figure 6.3	Nucleotide diversity of chloroplast genome regions in genus <i>Firmiana</i>	124
Figure 7.1	Frequency of amino acids in <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	130
Figure 7.2	Comparison of microsatellites and oligonucleotide repeats between <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	136
Figure 7.3	Nucleotide diversity of chloroplast genome region between <i>Hibiscus</i> species	147

Figure 8.1	The phylogenetic analysis of four families of order Malvales based on protein coding genes	153
Figure 8.2	Phylogenetic analysis of family Malvaceae based on protein coding sequences	155
Figure 8.3	Phylogenetic tree based on complete chloroplast genome	157
Figure 8.4	Phylogenetic tree based the LSC regions	158
Figure 8.5	Phylogenetic tree based on SSC regions	159
Figure 8.6	Phylogenetic tree based on IR region	160

LIST OF TABLES

Sr. No.	Title	Page No.
Table 1.1	Families combined with Malvaceae s.s under different classification systems	4
Table 2.1	Detail of accessions and data downloaded from SRA database	30
Table 2.2	Chloroplast gene content and functional classification of Malvaceae species	34
Table 3.1	Comparative analyses of the general features of 20 species of family Malvaceae	51
Table 3.2	Comparison of evolutionary rates of 77 protein coding genes among Malvaceae species	56
Table 4.1	Correlations and regression of SNPs with SSR and non-SSR InDels	80
Table 4.2	Correlations and regression of SSR and non-SSR InDels with repeats	81
Table 5.1	General features and comparison of the chloroplast genomes of <i>Theobroma cacao</i> and <i>Theobroma grandiflorum</i>	85
Table 5.2	Relative synonymous codon usage comparison of <i>Theobroma cacao</i> and <i>Theobroma grandiflorum</i>	86
Table 5.3	Putative RNA editing site in <i>Theobroma cacao</i> and <i>Theobroma grandiflorum</i>	87
Table 5.4	Simple sequence repeats in <i>T. cacao</i> and <i>T. grandiflorum</i>	89
Table 5.5	Oligonucleotide repeats in <i>Theobroma</i> species	92
Table 5.6	Thirty high polymorphic regions among <i>T. cacao</i> and <i>T. grandiflorum</i>	99
Table 6.1	Comparison of chloroplast genomes of <i>F. colorata</i> , <i>F. major</i> and <i>F. pulcherrima</i> and their general features	103
Table 6.2	Relative synonymous codon usage of genus <i>Firmiana</i> species	104
Table 6.3	RNA editing sites in <i>Firmiana</i> genomes	105
Table 6.4	Microsatellites loci in <i>Firmiana</i> species	111
Table 6.5	Oligonucleotide repeats in <i>Firmiana</i> species	116
Table 6.6	Comparison of substitution in <i>Firmiana</i> species	123
Table 6.7	Distribution of InDels in <i>Firmiana</i> chloroplast genome	123
Table 6.8	Mutational hotspots among <i>Firmiana</i> species	125
Table 7.1	Comparison of chloroplast genome of <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	130
Table 7.2	Comparison of Relative synonymous codon usage (RSCU) between <i>H. rosa sinensis</i> and <i>H. syriacus</i>	131

Table 7.3	RNA editing sites in <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	132
Table 7.4	Microsatellites loci in <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	135
Table 7.5	Oligonucleotide repeats in <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	137
Table 7.6	Types and distribution of SNPs in <i>Hibiscus</i>	145
Table 7.7	Distribution of InDels in <i>Hibiscus</i>	145
Table 7.8	Mutational hotspots between <i>H. rosa-sinensis</i> and <i>H. syriacus</i>	146
Table 8.1	Accessions of species used in phylogenetic tree	152

LIST OF ABBREVIATIONS

BGI	Beijing genomics institute
BWA	Burrow wheel aligner
C	Complementary
CDS	Protein coding sequences
CNVs	Copy number variations
DOGMA	Dual Organellar genome annotator
F	Forward
FAO	Food and Agriculture organization
GBBSI	granule-bound starch synthase
GTR	Generalized time reversible
IGS	Intergenic spacer regions
InDels	Insertions and deletions
IR	Inverted repeat
ITS	Internal transcribed sequences
Ka	Non-synonymous substitution
Ks	Synonymous substitutions
LSC	Large single copy
MAFT	Multiple Alignment using Fast Fourier Transform
MISA	MIcroSAtellite identification tools
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
P	Palindromic
Poly A	Polyadenine
Poly C	Polycytosine
Poly G	Polyguanine
Poly T	Polythymine
PREP-cp	Predictive RNA editor for plant chloroplast genes
R	Reverse

RFLPs	Restriction fragment length polymorphisms
rRNA	Ribosomal RNA
RSCU	Relative synonymous codon usage
SDS	Sodium dodecyl sulphate
SNP	Single nucleotide polymorphism
SRA	Sequence Read Archive
SSC	Small single copy
SSR	Simple sequence repeat
tRNA	Transfer RNA
T _s	Transitions
T _v	Transversions
η	Total number of mutations
π	Nucleotide diversity

ACKNOWLEDGMENTS

I am grateful for all the blessing that **Allah Almighty** has showered on me which enabled me to complete this thesis. Countless blessings on the **Holy Prophet Syedena Muhammad** (Sallallahu Allaihe Waalae Wassalum), the sea of knowledge, for provision of arguments to faith in Allah and guidance to the true path of life.

My deepest gratitude to my supervisor, **Dr. Mohammad Tahir Waheed**, Assistant Professor, Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University Islamabad, Pakistan and Co-supervisor **Dr. Ibrar Ahmed**, Chief Executive Officer, Alpha Genomics, Private Limited, Islamabad. Their expertise, understanding and patience enabled me to complete this research. I am thankful to them for their inspiration, reassurance and counselling, specifically for providing friendly environment throughout my PhD study. I was lucky to have advisors like them. I am obliged to **Dr. Muhammad Rashid Khan**, Chairman, Department of Biochemistry, Faculty of Biological sciences, Quaid-i-Azam University, Islamabad, Pakistan for extending the research facilities of the Department to accomplish this task.

I also want to express my deep sense of gratitude to **Prof. Dr. Bushra Mirza**, Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan; **Dr. Shahid Waseem**, Alpha Genomics Private Limited, Islamabad; **Dr. Muhammad Naeem**, Federal Seed certification and Registration Department, Islamabad, Pakistan, for their guidance and valuable suggestions throughout the study. I would like to acknowledge Higher Education Commission, Pakistan (HEC) for providing the Indigenous scholarships and financial support during my research.

I would like to extend my deepest appreciation to those colleagues and staff, who helped me in one way or the other during my stay at Department of Biochemistry, Quaid-i-Azam University, Islamabad, Pakistan. Special thanks to **Irshad Khan, M. Amir, M. Rasheed** and **M. Ramzan** to assist me in various ways during my research work. I'll always remember the cooperation and help of Biochemistry office staff; **M. Tariq, M. Maqsood, M. Fayyaz** and **M. Shahzad**.

I would like to pay very special tribute to colleagues and friend Furrukh Mehmood, Muhammad Bilal, Muhammad Suleman Malik, Muhammad Nouman Malik, Irum Naz, Zain Ali, Iram Shahzadi, Muhammad Zahid, Sara Latif, Kiran Saba, Fatima Ijaz, Neelum Batool, Sumaira Sarwar and M. Usman Tareen for their inseparable support and prayers.

Most importantly, none of this would have been possible without the love and patience of my Family. I would like to express my heart-felt gratitude to my adorable father **Bakht Biland** and my sweetest mother **Jan Zari** for their unflagging love and unconditional support throughout my life and my studies. You made me live the most unique, magic and carefree childhood that has made me who I am now. Words fail to express my appreciation to my dearest brothers, my sisters, my cousins and kids **M. Yasir, M. Amar, Dua Abdullah** and **Fareeha Abdullah**. In last, I am also thankful to my beloved wife **Amria Bibi** for her love, prayers and support during my study.

ABDULLAH

ABSTRACT

The plant family Malvaceae (eudicot, angiosperms) belongs to order Malvales. This family comprises of 244 genera and 4225 species, which are distributed in tropical to temperate regions of the world. Malvaceae possesses plastic morphology which leads to taxonomic discrepancies in the classification. The subdivision of family Malvaceae into nine subfamilies is the most accepted classification. The complete chloroplast genome sequences are used to evaluate evolutionary dynamics and phylogeny of plant lineages that help to resolve taxonomic discrepancies. The current study aimed to evaluate the evolutionary dynamics and phylogenetic relationships within the family Malvaceae.

We sequenced and assembled chloroplast genome sequences of four species of family Malvaceae including *Hibiscus rosa-sinensis*, *Hibiscus mutabilis*, *Malvastrum coromandelianum* and *Malva parviflora*. The whole genomic DNA shotgun was generated through Illumina HiSeq2500 from pair end run with 150 bp short read and 350 bp insert size. Moreover, we also assembled chloroplast genome sequences of four species from the Sequence Read Archive (SRA) of National Center for Biotechnology Information (NCBI) including *Firmiana colorata*, *Sterculia monosperma*, *Pterospermum truncatolobatum*, and *Urena procumbens*. The quality of raw reads was accessed by fastQC, and chloroplast genomes were *de novo* assembled by Velvet 1.2.10 with various kmer values. The coverage analysis was performed with Burrow wheel aligner (BWA)/Bowtie and Tablet was used for visualisation. Chloroplast genomes were annotated with Dual Organellar genome annotator (DOGMA) and GeSeq. The genomic features, genes content, codon usage, and amino acids frequency were analysed using Geneious R8.1. RNA editing sites were analysed with Predictive RNA editor for plant chloroplast genes (PREP-cp), simple sequence repeats (SSRs) were analysed with MicroSATellite identification tools (MISA), and oligonucleotide repeats were analysed with REPuter program. The rate of synonymous and non-synonymous substitutions of protein coding genes was analysed in DnaSP 5.10 after pairwise alignment through Geneious R8.1 using *Theobroma cacao* as reference. Correlation among substitutions, InDels, and oligonucleotide repeats were analysed at family and genus level. At family level comparisons, one species, each from 13 genera was pairwise aligned to *Theobroma cacao*, a species basal to Malvaceae, to find substitutions and InDels across species in Malvaceae. Using coordinate positions of forward and reverse oligonucleotide repeats in *T. cacao* chloroplast genome, we evaluated correlations among direct and reverse repeats, substitutions and InDels in these Malvaceae chloroplast genomes. At the genus level, we investigated these correlations in five genera by taking one species of the respective genus as a reference and comparing it with

another species of the same genus. The phylogenetic tree was reconstructed based on chloroplast genome sequences using the IQ-tree program.

Our results showed that the sequencing of whole genomic DNA with low coverage depth provides quality genomic resources for the assembly of chloroplast genome with high coverage depth due to 100-1000 times higher chloroplast genomic DNA in the plant cell as compared to the nuclear genome. The comparative analyses of the chloroplast genomes of family Malvaceae from basal lineages to crown groups revealed high similarity in gene content, gene organisation, intron content, and GC content. We observed differences in the length of these chloroplast genomes due to variations in the length of intergenic spacer regions and contraction and expansion of IRs regions. The rate of synonymous and non-synonymous substitutions revealed about 95% similarities among Malvaceae species. However, the rate of synonymous substitutions was higher than non-synonymous substitutions. The IRs contraction and expansion not only lead to generation of pseudogenes at junctions of chloroplast genomes, but also lead to duplication or deletions of a single copy of some genes as observed in *Durio zibethinus* and *Abelmoschus esculentus*. The correlation analyses of substitutions with repeats, substitutions with InDels, and repeats with InDels revealed weak to strong correlations. High regression value was observed for substitutions on InDels, substitutions on repeats, and InDels on repeats. We hypothesize that such correlations are a common characteristic of chloroplast genomes in all plant lineages as these were also previously observed in the family Araceae (monocots, angiosperms) and Cephalotaxaceae (gymnosperms).

The comparative analyses of three genera including *Theobroma*, *Firmiana*, and *Hibiscus* revealed high similarities in inter-genus and intra-genus level comparison for codon usage, amino acid frequency, RNA editing sites, and simple sequence repeats. We found two times higher oligonucleotide repeats in the crown group of family Malvaceae (*Hibiscus*) than basal groups (*Theobroma* and *Firmiana*). Thirty mutational hotspots were identified in each genus. The phylogenetic inference of family Malvaceae based on complete chloroplast genome sequences attests the previous classification of family Malvaceae into nine subfamilies. Our results revealed high suitability of coding sequences and complete chloroplast genome in inferring of phylogeny and resolve species at the subfamilies level with 100 bootstrapping.

In conclusion, this study provides a broad insight into evolutionary dynamics and phylogeny of family Malvaceae. The identified mutational hotspots could be used for development of robust and cost-effective markers to infer phylogeny of these genera, specifically, in the genus *Hibiscus*.

Chapter 1

Introduction and literature review

1.1 Introduction

The plant family Malvaceae sensu lato (s.l) or Mallow is a eudicot family of order Malvales (Bayer *et al.*, 1999). Family Malvaceae consists of 244 genera and 4,225 species (Christenhusz and Byng, 2016). Plants of Malvaceae are widely distributed in tropical to temperate regions of the world (Xu and Deng, 2017) and possess plastic morphology which gives rise to taxonomical discrepancies both at family and genus level (Alverson *et al.*, 1999; Carvalho-Sobrinho *et al.*, 2016; Pfeil *et al.*, 2002; Tate *et al.*, 2005). Due to its plastic morphology, this family is divided into nine subfamilies which are Brownlowioideae, Bombacoideae, Byttnerioideae, Dombeyoideae, Grewioideae, Helicteroideae, Malvoideae, Tilioideae, and Sterculioideae (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Judd and Manchester, 1997; Xu and Deng, 2017).

1.2 Morphological features of Malvaceae

Although family Malvaceae possesses plastic morphology, many researchers describe its general morphological features (Bayer and Kubitzki, 2003; Heywood, 1993; Watson, 1992; Xu and Deng, 2017). According to these researchers, plants of family Malvaceae are herbs, shrubs, and trees usually with stellate hairs. Leaves are simple, rarely entire, and mostly exist in the form of dissected or digitately compound. Moreover, leaves are stipulate, spiral or non-sheathing and petiolate usually with palmate venation. Flowers are bisexual or unisexual, hypogynous, rarely zygomorphic, usually actinomorphic or asymmetrical, and commonly pollinated by insect due to floral nectar. The flowers are present solitary or forming compound cymes with regular or somewhat irregular shapes. Furthermore, flowers vary in size from small to large and cyclic with distinct corolla and calyx. Sepals are mostly 5 in numbers, fused or free, sometimes incompletely separated, mostly valvate in bud, with free, spreading tips, in some genera caducous, petaloid or persistent. Number of petals is mostly equal to sepals, sometimes reduced or completely lacked, present in various shapes, often twisted in bud, and free or fused with the bases of the stamen filaments. Epicalyx is absent or present, and the corolla is asymmetric, polypetalous, contorted or imbricate. Androecium or stamens range in number from 5 to more than 1000, usually in phalanges or antepetalous groups. The filaments of androecium are free or joined to petals and frequently forming conspicuous staminal tube. Anthers are dorsifixed or basifixed whereas spores are formed in di-, tetra- or polysporangiate. Staminodia is often present and fused with stamens usually antesealous and rarely petaloid. Gynaecium is superior, syncarpous, synovarious to synstyleovarious containing one to many carpels. Carpels are present in free or fused form (usually at fruit stage), sessile or on a distinct androgynophore. Ovary consists of one to many locules. Styles are apical, free or incompletely

joined ending with the stigmas that are dry papillate or non-papillate. Ovules are in axial placentation and anatropous to campylotropous in structure. Fruit is present in the form of berry or loculicidal capsules, fleshy or non-fleshy, schizocarp, dehiscent or indehiscent comprising nutlets or follicles. Seeds are with or without hairs and containing oily endosperm embryo with zigzag micropyle.

1.3 Importance of family Malvaceae

Family Malvaceae includes various economically important species. The species of Malvaceae are used as food, fodder, fibre, timber, ornamental, and medicines. A brief overview about the importance of family Malvaceae is provided below.

1.3.1 Fibre

Cotton is the most important crop and obtained from the four species of *Gossypium* including *Gossypium arboreum*, *Gossypium barbadense*, *Gossypium herbaceum*, and *Gossypium hirsutum* (Hinsley, 2008). According to report of Statista of 2017/2018, 6.21 million metric tons of cotton was produced (Statista, 2018; accessed 17 May 2019). Other fibre producing species of Malvaceae include Jute and Kenaf: Jute including species are *Corchorus olitorius* and *Corchorus capsularis* whereas Kenaf includes *Hibiscus cannabinus*. According to Food and Agriculture organization (FAO), the fibre produced from jute, kenaf and their allies were 2.3 million tons in 2010 (FAO, 2010; accessed on 17 May 2019). Kapok is produced by *Bombax ceiba* and used for insulation purposes in textiles industry due to its waterproof property (Hinsley, 2008).

1.3.2 Food

Many species of family Malvaceae are used as food. Durian is the fruit of *Durio zibethinus* with unique taste and aroma and used in Southeast Asia, including Thailand, Malaysia, Indonesia and Philippines (Leontowicz *et al.*, 2007). This fruit is considered as 'King of fruit' (Voon *et al.*, 2007). Cacao tree (*Theobroma cacao*) and Cupuassu tree (*Theobroma grandiflorum*) are the economically important and native to Brazil (Cuatrecasas, 1964). Cacao is grown in about 50 countries throughout the humid tropic regions (Motamayor *et al.*, 2013) and their enclosed seeds within the pods (fruits) are used in cosmetics, confectionary, and in chocolate production (Litz, 2005). Cacao is also vital for the livelihood of 40-50 million people around the globe, including smallholder farmers (Foundation TWC, <http://www.worldcocoafoundation.org/about-cocoa/>, accessed on 16 September 2018). Cupuassu is used to prepare chocolate-like product (cupulate), ice cream, juices, yogurt, liquor, candy, desserts and domestic jellies and jams (Cavalcante, 1991). Okra is the fruit of

Abelmoschus esculentus and used as vegetable in various parts of the world (Lamont, 1999). Some historian reported the use of Okra as fruit in Egypt in 1216 A.D (Lamont, 1999). Several other species are also used as food including *Abelmoschus manihot* and species of *Corchorus* and *Malva* as leafy vegetables (Hinsley, 2008). The vegetable oil is also produced as by-product from cotton and kenaf when processed for fibre production (Hinsley, 2008).

1.3.3 Horticulture

The species of family Malvaceae have showy flowers and about 100 species are used in horticulture (Hinsley, 2008). The popular genera of horticulture are *Alcea* and *Hibiscus* (Bayer and Kubitzki, 2003). Species of *Alcea*, commonly known as hollyhocks, are used in horticulture including *Alcea ficifolia* (Fig-Leaved Hollyhock), *Alcea rugosa* (Hairy Hollyhock) and *Alcea rosea* (Common Hollyhock). Many cultivars of *Alcea rosea* are available in different colours from white to black whereas pattern of flowers varies from single petal to double petals (<https://www.tropicos.org/>, accessed on 17 May 2019). Species of *Hibiscus* that are widely used in horticulture including *Hibiscus mutabilis* (Confederate Rose), *Hibiscus rosa-sinensis* (tropical *Hibiscus* or china rose), *Hibiscus syriacus* (Rose of Sharon), *Hibiscus moscheutos* (Swamp Mallow), and *Hibiscus trionum* (flower of an hour), *Hibiscus subdariffa* (Roselle, also used as food) (Pfeil and Crisp, 2005). Some other species that are used in horticulture include *Malva moschata* and *Lavatera trimestris* (Hinsley, 2008).

1.3.4 Medicinal properties

Family Malvaceae contains many medicinally important genera and species. The species of genus *Malva* include *Malva parviflora*, *Malva neglecta*, and *Malva sylvestris* has been reported with extensive medicinal properties including anti-diabetic, anti-oxidant, anti-inflammatory, free radical scavenger, metal chelating, and burn healing (Akbar *et al.*, 2014; Dalar *et al.*, 2012; DellaGreca *et al.*, 2009; Prudente *et al.*, 2013; Razavi *et al.*, 2011; Shale *et al.*, 2005). Some genera of subfamily Sterculioideae contain species with a broad range of medicinal activities. For instance, anti-cancer, anti-diabetic, anti-bacterial, anti-hypertensive, anti-constipation, anti-inflammatory, and wound healing properties have been reported (Al Muqarrabun and Ahmat, 2015). The species of genus *Sterculia* showed cytotoxic and anti-microbial activities (Atakpama *et al.*, 2015; Nanadagopalan *et al.*, 2015; Vital *et al.*, 2010) whereas the species of genus *Firmiana* showed antimicrobial (Ajaib *et al.*, 2014), anti-inflammatory (Lim *et al.*, 2017), anti-cancer (Woo *et al.*, 2015), neuroprotective (Lim *et al.*, 2017), and hepatoprotective properties (Kim *et al.*, 2015). *Firmiana colorata* is used for the intestinal dysfunction in some tribes of Bangladesh (Azam *et al.*, 2013). Moreover, the species of *Firmiana* are also used for the caffeine containing teas (Bayer and Kubitzki, 2003). Many species of *Hibiscus*, the genus

of tribe Hibisceae and subfamily Malvoideae have been shown to possess broad curative activities including anti-fungal, anti-bacterial (Vasudeva and Sharma, 2008), anti-viral (Baatartsogt *et al.*, 2016), anticancer, and apoptosis inducing properties (Alam *et al.*, 2018; Goldberg *et al.*, 2017). In some cases, species of the genus *Hibiscus* also showed activity against hypertension, inflammation, hyperlipidemia, obesity, and anaemia (Riaz and Chopra, 2018; Shen *et al.*, 2017). The *Malvastrum coromandelianum* also belonging to subfamily Malvoideae, possesses hypoglycaemic, analgesic, anti-inflammatory, anti-nociceptive, and anti-bacterial activities (Khonsung *et al.*, 2006; Sittiwet *et al.*, 2008).

1.4 Taxonomic position of family Malvaceae

Several previous studies focussed on taxonomical classification of family Malvaceae. Due to plastic morphology, different taxonomic classifications of Malvaceae were suggested in different classification systems. The former Malvaceae (recent Malvoideae) *sensu stricto* (s.s) was combined with different families and different classification for Malvaceae s.s was suggested (Table 1.1).

Table 1.1 Families combined with Malvaceae s.s under different classification systems

System	(Cronquist, 1981, 1988)	(Dahlgren, 1983)	(Thorne, 1992)	(Takhtadzhan and Takhtajan, 1997)	(APG, 1998, 2003)
Families	Malvaceae Sterculiaceae Tiliaceae Bombacaceae Lecythidaceae Elaeocarpaceae	Malvaceae Tiliaceae Sterculiaceae Bombacaceae Cistaceae Cochlospermaceae Huaceae Sphaerosepalaceae Sacrolaenaceae Dipterocarpaceae Plagiopteraceae Bixaceae	Malvaceae Sterculiaceae Bombacaceae Tiliaceae Diegodendraceae Dipterocarpaceae Gonystylaceae Sphaerosepalaceae Monotaceae Cistaceae Cochlospermaceae Sarcolaenaceae Thymelaeaceae Huaceae Plagiopteraceae Bixaceae	Malvaceae Bombacaceae Sterculiaceae Tiliaceae Dipterocarpaceae Sarcolaenaceae Diegodendraceae Plagiopteraceae Monotaceae Huaceae Sphaerosepalacea	Malvaceae Bombacaceae Sterculiaceae Tiliaceae

(The data for table 1.1 was taken from Taia 2009)

The closely related families including Tiliaceae, Sterculiaceae, Bombacaceae, and Malvaceae s.s have been merged into an extended family Malvaceae s.l. Different studies revealed monophyletic position of these four families as compared to paraphyletic position of the former

families (APG, 2003, 1998; Bayer *et al.*, 1999). Characters shared by family Malvaceae include trichomatous nectaries, valvate sepal aestivation (Vogel 2000), tile cells, and a common basic inflorescence structure (Bayer *et al.*, 1999). Other features that were used to illustrate the alliance are stratified phloem strands with triangular rays, mucilage cavities, cyclopropanoid acids, stellate hairs, and seed coats with exotegmic palisades (Bayer and Kubitzki, 2003).

Family Malvaceae s.l has been subdivided into nine subfamilies by phylogenetic analysis that was based on chloroplast genome sequences of *atpB* and *rbcL* (Bayer *et al.*, 1999). The similar classification was also suggested based on chloroplast genome sequence of *ndhF* (Alverson *et al.*, 1999). The nine subfamilies that were suggested within family Malvaceae include: Bombacoideae (traditionally was considered as family Bombacaceae), Brownlowioideae (including species from traditional family Tiliaceae), Byttnerioideae (includes species from traditional family Sterculiaceae), Dombeyoideae (members of this family previously belonged to family Sterculiaceae), Grewioideae (includes species of Tiliaceae), Helicteroideae (includes species of tribes Helictereae and Durioneae of families Sterculiaceae and Bombacaceae, respectively), Malvoideae (traditionally known as family Malvaceae s.s), Sterculioideae (traditionally known as family Sterculiaceae), and Tilioideae (includes species of traditional family Tiliaceae) (Bayer *et al.*, 1999; Bayer and Kubitzki, 2003; Taia, 2009). This broad circumscription of family Malvaceae is considered (Cronquist, 1988) as core Malvales. This classification has been accepted by various researchers (Baum *et al.*, 2004; Bayer and Kubitzki, 2003; Carvalho-Sobrinho *et al.*, 2016; Duarte *et al.*, 2011; Perveen *et al.*, 2004; Tate *et al.*, 2005). However, some researchers suggested different classifications. For instance, Thorne (2000) considered Bombacaceae and Sterculiaceae under Malvaceae s.l and Byttneriaceae and Tiliaceae (with some restrictions) as separate families. Hinsley (2006) suggested four alternative classifications of Malvaceae s.l (core Malvales) as follows.

1. Tiliaceae, Sterculiaceae, Bombacaceae and Malvaceae s.s (core Malvales) to be considered as single family.
2. To consider each of nine monophyletic clades as a separate family instead of subfamilies.
3. The seven subfamilies including Bombacoideae, Byttnerioideae, Dombeyoideae, Helicteroideae, Malvoideae, Tilioideae, and Sterculioideae to be considered as separate family Malvaceae whereas remaining two families Brownlowioideae and Grewioideae as another family.
4. If paraphyletic grouping is not ignored then 5 family classification is possible similar to traditional classification by adding Byttneriaceae with Tiliaceae, Sterculiaceae, Bombacaceae, and Malvaceae s.s but *Tilia* will be transferred to Grewiaceae from Tilioideae.

1.5 Nine subfamilies classification of family Malvaceae

Certain taxonomic discrepancies exist in family Malvaceae but most of the researchers accepted classification of Malvaceae into 9 subfamilies (Baum *et al.*, 2004; Bayer and Kubitzki, 2003; Carvalho-Sobrinho *et al.*, 2016; Duarte *et al.*, 2011; Perveen *et al.*, 2004; Tate *et al.*, 2005). Here, we will provide brief detail about each subfamily.

1.5.1 Subfamily Bombacoideae

This is the remainder of the Bombacaceae, after the inclusion of the tribes Durioneae in Helicteroideae and Matisieae in Malvoideae. This subfamily comprises of 17 genera and 160 species in which 90% are distributed in the Neotropics (Carvalho-Sobrinho *et al.*, 2016). Species of this subfamily are mostly present in the form of large trees or exceptionally in form of shrubs (Bayer and Kubitzki, 2003). The important genera that are included in this subfamily are *Adansonia*, *Bombax*, *Pseudobombax*, *Pachira* and *Eriotheca* (Bayer and Kubitzki, 2003). A new clade Malvatheca is comprised of Malvoideae and Bombacoideae (Nyffeler *et al.*, 2005) which can lead to a new classification. Recently, Carvalho-Sobrinho *et al.* (2016) suggested three tribes in Bombacoideae that are Adansonieae, Bernoullieae, and Bombaceae by using molecular markers and morphological characters.

1.5.2 Subfamily Brownlowioideae

The monophyly of this subfamily is undisputed since its classification by Burret (1926) and was also confirmed in various molecular markers-based studies (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Nyffeler *et al.*, 2005). This subfamily comprises of 8 genera with up to 80 species including most of the species from the Old World (Nyffeler *et al.*, 2005). Most of the species are present in the form of tall trees in South East Asia, whereas smaller trees are also present in drier forest including the African, Caribbean/Carpodiptera and the Asian forests (Bayer and Kubitzki, 2003). Some important genera including in this subfamily are *Brownlowia*, *Berrya*, *Carpodiptera*, *Christiana*, *Diplodiscus*, *Jarandersonia* and *Pentace*.

1.5.3 Subfamily Byttnerioideae

Subfamily Byttnerioideae comprises about 26 genera and 650 species (Bayer and Kubitzki, 2003). This family is pantropical and most of the species exist in form of trees or shrubs whereas herbs are present occasionally (Barbara *et al.*, 2001). Representative genera of this subfamily include *Theobroma*, *Herrania*, and *Guazuma* and exist in Neotropical forest whereas the *Glossostemon* exists in the form of perennial dessert herbs in North Africa (Bayer *et al.*, 1999; Bayer and Kubitzki, 2003). This subfamily contains 4 tribes that are Theobromateae, Byttnerieae, Lasiopetaleae, and Hermannieae (Alverson *et al.*, 1999). The important genera are

Theobroma, *Herrania*, *Glossostemon*, *Abroma*, *Scaphopetalum*, *Leptonychia*, *Byttneria*, *Guichenotia* and *Lasiopetalum*.

1.5.4 Subfamily Dombeyoideae

Subfamily Dombeyoideae consists of about 20 genera and 350 species (Bayer and Kubitzki, 2003). Dombeyoideae are primarily located in the Paleotropical regions of the world. The centres of diversity include Southeast Asia, Madagascar, and Mascarenes (Nyffeler *et al.*, 2005). This subfamily rarely includes herbs, mostly consists of shrubs and trees (Bayer and Kubitzki, 2003). The species of this family grow in xeric environmental conditions, gallery forests, and savannahs (Bayer and Kubitzki, 2003). The molecular studies confirmed the monophyletic position of this subfamily despite the morphological similarities with other subfamilies of Malvaceae (Alverson *et al.*, 1999; Bayer *et al.*, 1999). Some of the important genera that belongs to this subfamily are *Dombeya*, *Harmsia*, *Melhania*, *Nesogordonia*, *Pterospermum*, *Schoutenia*, and *Sicrea* (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Bayer and Kubitzki, 2003; Nyffeler *et al.*, 2005).

1.5.5 Subfamily Grewioideae

According to Bayer and Kubitzki (2003), this subfamily comprises 25 genera and up to 700 species in tropical parts of the New and Old World. Some of its genera contain trees of medium size in humid forests of South East Asia (*Colona*), Africa and South East Asia (*Microcos*), only Africa (several genera), the Old and the New World (*Trichospermum*), or only the New World (*Goethalsia* and *Mollia*). Some exceptional genera prefer sub-arid habitats (*Grewia*). *Corchorus* is the well-known genus of this subfamily and two of its species including *Corchorus olitorius* and *Corchorus capsularis* are cultivated for jute production in tropical regions of the world.

1.5.6 Subfamily Helicteroideae

The subfamily Helicteroideae comprises about 12 genera and 130 species and mostly occupied tropical Australasia (Nyffeler *et al.*, 2005). Some genera also extended into tropical Africa (*Triplochiton* and *Mansonia*) and the Neotropics (*Helicteres* and *Reevesia*) (Bayer and Kubitzki, 2003; Nyffeler *et al.*, 2005). The monophyly of subfamily Helicteroideae is well-supported in previous studies based on morphology and molecular markers of nuclear and chloroplast (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Nyffeler *et al.*, 2005).

1.5.7 Subfamily Sterculioideae

Sterculioideae consists of 12 genera and about 400 species (Nyffeler *et al.*, 2005). This subfamily usually comprises apetalous tropical trees from the Old World with unisexual

flowers, and androgynophores are prominent in the variable range (Nyffeler *et al.*, 2005). Some species tolerate aridity and develop swollen water-storing stems (e.g. the mostly Australian *Brachychiton*). Some of the important genera include *Sterculia*, *Firmiana*, *Heritiera*, *Coloa*, and *Pterocymbium* (Bayer and Kubitzki, 2003). Monophyly of this subfamily is well-supported by various molecular studies that support the existence of various unique morphological characters (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Bayer and Kubitzki, 2003; Nyffeler *et al.*, 2005).

1.5.8 Subfamily Tilioideae

This is a small family comprised about 2-3 genera and 40-52 species (Bayer *et al.*, 1999; Bayer and Kubitzki, 2003; Nyffeler *et al.*, 2005). Here, discrepancies exist on the number of genera that belong to Tilioideae. Some studies based on chloroplast markers of *atpB* and *rbcL* showed that two genera *Craigia* and *Tilia* belong to Tilioideae (both include 40 species) while the genus *Mortoniendron* (include 12 species) belongs to Brownlowioideae or Tilioideae but it lacks bootstrapping supports (Bayer *et al.*, 1999; Bayer and Kubitzki, 2003). Others molecular studies used molecular markers *matK* and *ndhF* along with internal transcribed sequences (ITS) that showed genus *Mortoniendron* as sister group to genera *Craigia* and *Tilia* (Alverson *et al.*, 1999; Judd and Manchester, 1997; Nyffeler *et al.*, 2005). The distribution of these three genera is unusual in subfamily Tilioideae such as *Mortoniendron* belongs to Central America, *Craigia* to China, and *Tilia* to America and Eurasia.

1.5.9 Subfamily Malvoideae

The subfamily Malvoideae is the largest subfamily in Malvaceae, comprising of 100 genera and 1,730 species (Bayer and Kubitzki, 2003). Species of this subfamily are distributed in temperate to tropical regions around the globe, but majority of the species are present in the New World (Bayer and Kubitzki, 2003; Fryxell, 1997; Heywood, 1993; Watson, 1992). Species of this subfamily are herbs, shrubs and trees (Bayer and Kubitzki, 2003; Fryxell, 1997; Heywood, 1993). The genera including in this subfamily are *Hibiscus*, *Gossypium*, *Malva*, *Malvastrum*, *Urena*, *Kokia*, *Abelmoschus*, *Talipariti* etc. In the current study, we will also focus on the genus *Hibiscus* due to taxonomic discrepancies that exist in the genus.

1.5.10 Two major clades of Malvaceae

The nine subfamilies of Malvaceae are divided into two major clades that are Byttneriina and Malvadendrina (Alverson *et al.*, 1999; Bayer and Kubitzki, 2003; Nyffeler *et al.*, 2005). The Byttneriina clade is divided into two subclades that include subfamilies Grewioideae and Byttnerioideae whereas the Malvadendrina clade is divided into 6 subclades that include seven

subfamilies Brownlowioideae, Dombeyoideae, Helicteroideae, Bombacoideae, Malvoideae, Tilioideae and Sterculioideae (Alverson *et al.*, 1999). The Bombacoideae and Malvoideae are combined in subclade Malvatheca. The complete detail is provided in Figure 1.1 (taken from the previous study of Alverson *et al.* 1999).

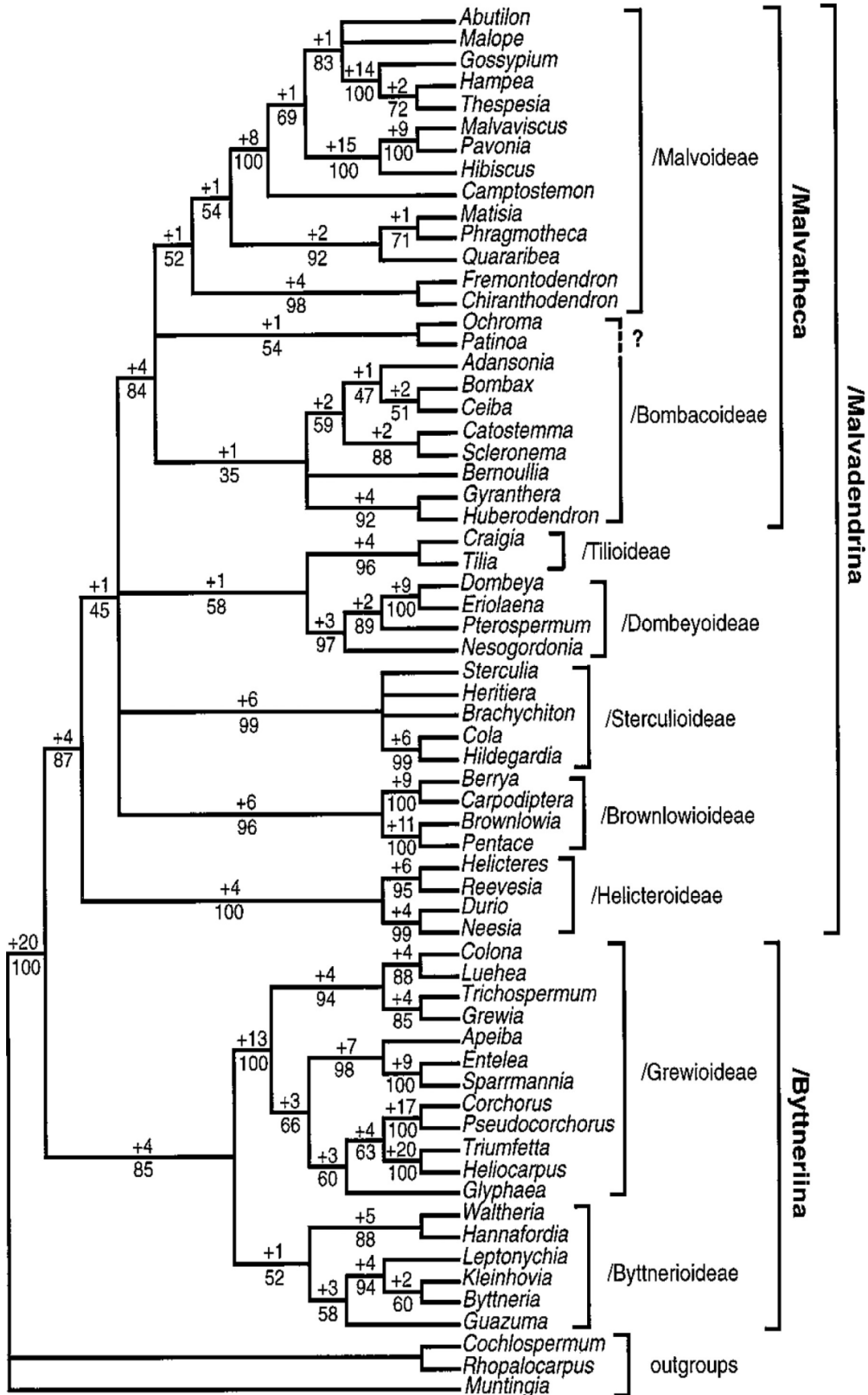


Figure 1.1 Phylogenetic relationship among the subfamily of Malvaceae

1.6 Genera of the family Malvaceae

The family Malveae includes 244 genera (Xu and Deng, 2017). The wide distribution and plastic morphology lead to taxonomic discrepancies (Bayer and Kubitzki, 2003; Pfeil *et al.*, 2002; Tate *et al.*, 2005). Certain taxonomic discrepancies exist at intra and inter-genus level classification of certain genera of family Malvaceae (Esteves, 2000; Patil *et al.*, 2015; Pfeil *et al.*, 2002; Pfeil and Crisp, 2005; Tate *et al.*, 2005; Wilkie *et al.*, 2006). Therefore, the use of new molecular markers were suggested for phylogeny inference (Patil *et al.*, 2015; Small, 2004; Tate *et al.*, 2005; Wilkie *et al.*, 2006). The use of lineage specific markers has been suggested for the phylogeny inference to resolve taxonomic discrepancies (Daniell *et al.*, 2016). The chloroplast genome polymorphism has been used in the studies of population genetics and phylogenetics (Ahmed, 2014; Henriquez *et al.*, 2014; Zhai *et al.*, 2019). The mutational hotspots have also been identified for designing of unique and robust markers for phylogeny inference of several genera (Cheon *et al.*, 2019; Li *et al.*, 2018; Yu *et al.*, 2019, 2017). Here, we will focus on the comparison of chloroplast genomes of three genera of the family Malvaceae which include *Theobroma*, *Firmiana* and *Hibiscus*.

1.6.1 Genus *Theobroma*

Genus *Theobroma* L. belongs to tribe Theobromeae and subfamily Byttnerioideae Burnett of family Malvaceae (Purseglove, 1968). According to “The plant list” (<http://www.theplantlist.org/>: accessed 10 Jan 2019), *Theobroma* consists of 91 plants: 25 are accepted species, 51 are synonym, and 15 are unresolved. *Theobroma cacao* is considered among the first evolved species of Malvaceae and this genus lies basal to Malvaceae (Richardson *et al.*, 2015). This evergreen crop is native of the South American rainforest (Bartley, 2005) with the centre of origin in the upper Amazon regions where the greatest diversity is observed (Bartley, 2005; Zhang *et al.*, 2011). Cacao tree (*Theobroma cacao*) and Cupuassu tree (*Theobroma grandiflorum*) are the economically important and native to Brazil (Cuatrecasas, 1964). Cacao is grown in up to 50 countries throughout the humid tropic region (Motamayor *et al.*, 2013) and their enclosed seeds within the pods (fruits) are used in cosmetics, confectionary, and in chocolate production (Litz, 2005). Cacao is also vital for the livelihood of 40-50 million people in the whole world including smallholder farmer (Foundation TWC, <http://www.worldcocoafoundation.org/about-cocoa/>, accessed on 16 September 2018). Cupuassu is used to prepare chocolate-like product (cupulate), ice cream, juices, yogurt, liquor, candy, desserts and domestic jellies and jams (Cavalcante, 1991).

1.6.1.1 Genus *Theobroma* morphology

The morphology of *Theobroma* has been described by Cuatrecasas (1964) and Bayer and Kubitzki (2003). Most of the species of genus *Theobroma* are trees with pseudo whorled plagiotropic branches. The leaves of the species are dimorphic, simple or weakly lobed, distichous on plagiotropic shoots, spiral on orthotropic, and exceptionally lack prominent basal veins. The inflorescences present in various forms from axillary to cauliflorous. The calyx split in 2-3 lobes and petals are deeply cucullate with apical appendage. The ovary is consisting of 5 locular with numerous ovules whereas the stamens after fusing form 2-3 groups and originating a short tube with the prominent staminodes. Fruits are carnose to lignified/non-lignified, indehiscent, and with many embedded seeds in the sweet endocarp pulp.

1.6.1.2 Molecular based studies in genus *Theobroma*

Several molecular studies evaluate genetic diversity, nuclear genome structure, domestication, phylogenetic, and ultra-barcoding of *Theobroma cacao* and *Theobroma grandiflorum* (Alves *et al.*, 2007; Kane *et al.*, 2012; Richardson *et al.*, 2015). However, none of the studies carried out detailed comparative analysis of chloroplast genome that can play a significant role in phylogeny to resolve taxonomic discrepancies. Kane *et al.* (2012) analysed chloroplast genomes of *Theobroma* for ultra-barcoding and suggested five barcoding regions but did not perform comparative chloroplast structure analysis. For instance, analyses of simple sequence repeats (SSRs), oligonucleotide repeats, putative RNA editing sites, codon usage, and amino acid frequency of the *Theobroma*, basal to Malvaceae, might be helpful to understand the evolutionary dynamics of family Malvaceae.

1.6.2 Genus *Firmiana*

Firmiana is a small genus that includes sixteen species in accepted state, whereas five species are categorised as unresolved (<http://www.theplantlist.org/>; accessed 29 December 2018). *Firmiana* genus includes deciduous trees (rarely shrubs) that are distributed in eastern Africa and eastern-south eastern Asia to Malaysia (Kostermans, 1957). Eight species of *Firmiana* inhabit China (Huang *et al.*, 2011). Some species of *Firmiana* are cultivated for their beautiful shape and lovely flowers (Fan *et al.*, 2013). With reference to medicinal perceptive, *Firmiana* species had shown antimicrobial (Ajaib *et al.*, 2014), anti-inflammatory (Lim *et al.*, 2017) and anti-cancer (Woo *et al.*, 2015) properties. Some species also showed neuroprotective (Lim *et al.*, 2017) and hepatoprotective effect (Kim *et al.*, 2015). *Firmiana colorata* is used for the intestinal dysfunction in some tribes of Bangladesh (Azam *et al.*, 2013). Moreover, the species of *Firmiana* are also used for the caffeine containing teas (Bayer and Kubitzki, 2003).

1.6.2.1 Morphology of genus *Firmiana*

The species of genus *Firmiana* are usually trees and rarely shrubs with simple, often cordate or lobed leaves. The sepals are five in number with whitish, reddish, orange or green yellow petaloid (Kostermans, 1957). Androgynophore is conspicuous with 10 or 15 stamens and 2-6 ovulate with 5 carpels. Follicles are papery and usually open before maturity with flat boat-shaped that disperse with seeds (Bayer and Kubitzki, 2003). The seeds are 2–6 situated on carpel margins with glabrous shape and abundant endosperms with straight embryo (Kostermans, 1957). *Firmiana* is categorised by distinguishing fruit that contains five stipitate follicles and up to 6 glabrous seeds that are present along the basal margins (Bayer and Kubitzki, 2003). Previously, this genus was included in the family Sterculiaceae but now this genus is included in subfamily Sterculioideae, after inclusion of the Sterculiaceae in the expanded Malvaceae s.l (Taia, 2009).

1.6.2.2 Molecular studies in genus *Firmiana*

Firmiana genus is ignored due to lack of efficient molecular markers (Fan *et al.*, 2013), and up to the best of our knowledge, till date, only one study focused on *Firmiana* genus in which four species of *Firmiana* were evaluated based on microsatellite markers developed by transcriptome assembly (Fan *et al.*, 2013). Another study based on *de novo* transcriptome assembly identified genes responsible for adaptation and protection of *Firmiana* species in various stresses (Chen *et al.*, 2015). Recently, with the advancement of next-generation sequencing technology, the chloroplast genome sequence of *Firmiana major* (Ya *et al.*, 2017) and *Firmiana purlcherrima* (Wang *et al.*, 2017) was reported but phylogenetic position of *Firmiana* in the family was not determined. Previous phylogenetic analyses of subfamily Sterculioideae showed certain discrepancies in phylogenetic of certain clades (Wilkie *et al.*, 2006). Therefore, the authors suggested the use of new molecular markers for the accurate resolution of Sterculioideae phylogeny.

1.6.3 Genus *Hibiscus*

Hibiscus is one of the most diverse and widely distributed genera of family Malvaceae (Pfeil and Crisp, 2005). This genus is included in tribe Hibisceae of subfamily Malvoideae (Rizk and Soliman, 2014). The taxonomic revision of the different sections/part of a section of *Hibiscus* revealed that the number of species of *Hibiscus* range from 300 to 350 (Pfeil and Crisp, 2005). These studies include Malesian species (Waalkes, 1966), Bombicella (Fryxell, 1980), Mexican species (Fryxell, 1997), and section Furcaria (Wilson, 1999). Members of this genus are used in industries, horticulture, agriculture, food, and medicines (Akpan, 2000). Some other

molecular based studies also focused on barcoding and resolving of phylogenetic relationships of *Hibiscus*, but these studies have been inconclusive (Koopman and Baum, 2008; Pfeil *et al.*, 2002; Poovitha *et al.*, 2016; Small, 2004; Tate *et al.*, 2005). Species of *Hibiscus* that are widely used in horticulture including *Hibiscus mutabilis* (Confederate Rose), *Hibiscus rosa-sinensis* (tropical hibiscus or china rose), *Hibiscus syriacus* (Rose of Sharon), *Hibiscus moscheutos* (swamp mallow), and *Hibiscus trionum* (flower of an hour), *Hibiscus subdariffa* (roselle, also used as food) (Pfeil and Crisp, 2005). The species of high medicinal values of *Hibiscus* include *H. cannabinus*, *H. elatus*, *H. macaranthus*, *H. taiwanensis*, *H. vitifolius*, *H. rosa-sinensis*, *H. syriacus*, and *H. subdariffa* etc (Vasudeva and Sharma, 2008). These species have been shown to possess broad curative activities including anti-bacterial, anti-fungal (Vasudeva and Sharma, 2008) and anti-viral activities (Baatartsogt *et al.*, 2016). Moreover, species of the genus *Hibiscus* also showed activity against hypertension, inflammation, hyperlipidemia, obesity and anaemia (Riaz and Chopra, 2018; Shen *et al.*, 2017). Anticancer and apoptosis-inducing properties of *Hibiscus* have been also reported (Alam *et al.*, 2018; Goldberg *et al.*, 2017).

1.6.3.1 Morphology of Genus *Hibiscus*

Species of the genus *Hibiscus* are herbs, subshrubs, shrubs or trees and distributed in tropical to temperate regions of the world (Ayanbami *et al.*, 2012). Number of morphological characters have been provided for *Hibiscus* by Bayer and Kubitzki (2003). Leaves are simple or sometimes parted or lobed, dentate or less commonly subentire, sometimes with abaxial foliar nectaries. The flowers are usually solitary or rarely fasciculate with long or short pedicel. If present in fasciculate form, then rarely aggregated apically. The epicalyx contains bracts (3–)8–10(–20), distinct or basally connate, rarely suppressed; calyx 5-lobed; petals sometimes large and showy with numerous anthers; branches of stylar are 5 in number, distally free and having capitate stigmas. Capsules are oblong or ovoid, 5-locular with several seeds, glabrescent to hirsute. According to Linnaeus (1753), the genus *Hibiscus* includes all the capsular-fruited mallows.

1.6.3.2 Plasticity in morphology of *Hibiscus*

Hibiscus shows a high level of plastic morphology, share ‘branching style’ and ‘gossypol synthesis absence’ with Decaschistieae, Malvaviceae, and Malveae; ‘staminal column covered by five teeth’, ‘carpel number’ and ‘style branch number’ are matching with Decaschistieae, Gossypieae and Malveae (Fryxell, 1997; Sivarajan and Pradeep, 1996). Therefore, this genus is regularly questioned with various genera often included or excluded. Moreover, the inter-genus classification has been unstable despite the existence of some

cohesive groups and present identical problems to the revisionary taxonomists (Pfeil *et al.*, 2002).

1.6.3.3 Sections and segregates of *Hibiscus*

The genus *Hibiscus* has been divided into sections based on flowers and fruits characters. The group of species that showed distinct characters of flowers and fruits were segregated from *Hibiscus* and placed in other genera or as new genera (Hinsley, 2009) such as *Hibiscus populneus* is placed in genus *Thespesia* which is closely related to *Gossypium* with name *Thespesia populnea*. Some genera are placed from the start of classification in other genera, but molecular studies revealed their placement in *Hibiscus* (Koopman and Baum, 2008; Pfeil *et al.*, 2002; Pfeil and Crisp, 2005). The classifications of sections are not accepted at universal level due to absence of recent treatment at molecular level. The recognised sections vary in nature, from species-rich groups like *Furcaria*, to diverse groups like *Ketmia*, and to small closely knit groups such as *Muenchhusia* (the North American rose-mallows) (Hinsley, 2009). The classification of these sections is questionable and therefore the views of researchers are different in this regard. Some other sections of *Hibiscus* have been suggested other than mentioned above including *Calyphylli*, *Lilibiscus*, *Spatula*, *Muenchhusia*, *Venusti*, *Striati*, and *Trionum* (Hinsley, 2009).

1.6.3.4 Inter-genus and intra-genus taxonomic discrepancies in *Hibiscus*

The inter and intra-genus classification of *Hibiscus* is also inconclusive, and many researchers presented different views regarding the classification of *Hibiscus*. The *Fioria* has been considered separate genus by Sivarajan and Pradeep (1996) but not by Waalkes (1966) and Thulin (1999). The section *Azanzae* of *Hibiscus* were recognized as section (Hochreutiner, 1900; Waalkes, 1966). However, its segregation from genus *Hibiscus* and inclusion in genus *Talipariti* was also suggested (Craven *et al.*, 2003). The intra-generic classification of *Hibiscus* has been unstable despite the existence of some cohesive groups. The composition of sections has been controversial due to which different opinions have been present regarding its classification and inclusion and removal of some species. The splitting of *Hibiscus* section *Trionum* was suggested into sections *Striati*, *Muenchhusia*, *Clypeati*, *Venusti*, and *Trionum* s.s (Joseph, 1977) but the section *Trionum* was also accepted previously (Fryxell, 1980). Furthermore, the merging of *Hibiscus* section *Trichospermum* was suggested within *Hibiscus* section *Ketmia* (Waalkes, 1966). These taxonomic discrepancies arise due to plastic morphology. Hence, due to this reason the taxonomic classification of the genus *Hibiscus* based on morphological characters can not be helpful (Pfeil *et al.*, 2002).

1.6.3.5 Molecular based studies of genus *Hibiscus*

Previously, the taxonomic classification of *Hibiscus* was done based on morphological characters that influenced by environmental factors (Vinutha *et al.*, 2015). The genus *Hibiscus* is a large group that can not be defined by unique morphological characteristics and can lead to discrepancies (Pfeil *et al.*, 2002). Moreover, a recent study based on pollen morphology also suggested that pollen morphology could not be only used for taxonomic grouping of *Hibiscus* (Bae *et al.*, 2015). Hence, a study based on authentic and effective molecular markers is needed for resolving of phylogeny in genus *Hibiscus*.

Some researchers used molecular markers to resolve taxonomic discrepancies and elucidate phylogeny of genus *Hibiscus* (Koopman and Baum, 2008; Pfeil *et al.*, 2002; Poovitha *et al.*, 2016; Small, 2004) but these studies have been inconclusive. Pfeil *et al.* (2002) used chloroplast genome sequences of *ndhF* and *rpl16* intron to describe the phylogeny of tribe Hibisceae and the genus *Hibiscus* with the aim to achieve a monophyletic position of genus *Hibiscus*. However, expected results were not achieved and many segregated genera from *Hibiscus* were embedded within its phylogenetic tree, i.e. *Fioria* and *Abelmoschus*. Moreover, some members of tribes Decaschistieae and Malvavisceae were also embedded within *Hibiscus*. Small (2004) used nuclear ribosomal internal transcribed sequences (ITS), non-coding part of chloroplast DNA (*rpl16* intron), and a nuclear coding gene granule-bound starch synthase (*GBSSI*) but insufficiently resolved the phylogeny of *Hibiscus*. Therefore, the author suggested additional data source beyond the commonly used markers.

The Pfeil and Crisp (2005) further advanced the previous studies by using the molecular markers of chloroplast *matK* and nuclear *rpb2* and gave different suggestions for resolution of taxonomic discrepancies.

1. The inclusion of all embedded genera into the genus *Hibiscus*.
2. The splitting of *Hibiscus* into different small genera.
3. All the species which were embedded in the phylogeny of genus *Hibiscus* should be placed in the respective genera and the genus *Hibiscus* should be considered as paraphyletic in nature.

According to Hinsley (2009), all the embedded genera should be included in the *Hibiscus*, and by doing this the number of species of *Hibiscus* will be raised to about 500. Poovitha *et al.* (2016) used *rbcL*, *matK*, ITS, and *trnH-psbA* to find best suitable loci for barcoding. They included sixteen species from nine sections of *Hibiscus* and revealed that *Trichospermum* and *Bombicella* were not monophyletic. Moreover, the discrimination power of these markers was also compromised and could not resolve *Hibiscus platanifolius* of section *Spatula* and *Hibiscus*

lunariifolius of section *Trichospermum* with the combination of all these markers. These all studies revealed the need of new molecular tools as also suggested by the Small (2004).

1.7 General features of chloroplast genome

The chloroplast is a self-replicating organelle which plays a vital role in photosynthesis (Cooper, 2000). This fundamental organelle in plants sustained life on earth by converting light energy into chemical energy in the form of carbohydrates (Daniell *et al.*, 2016). Despite its role in photosynthesis, chloroplast also participates in various important biochemical reactions including synthesis of nucleotides, amino acids, fatty acids, vitamins and phytohormones (Daniell *et al.*, 2016). Chloroplasts are inherited from one parent, maternally in most angiosperm species (Daniell, 2007) but paternally in some gymnosperms (Neale and Sederoff, 1989) as compared to the nuclear genome. Furthermore, chloroplast genome lacks meiotic recombination similar to the nuclear genome which takes place between homologous chromosomes (Palmer, 1985). Due to these properties, chloroplast genome polymorphism has been used for resolving of phylogenetic relationships and taxonomic discrepancies (Daniell *et al.*, 2016), species barcoding (Nguyen *et al.*, 2017), population genetics (Ahmed, 2014), endangered species conservation (Zhao *et al.*, 2018), enhancement of breeding (Wambugu *et al.*, 2015) and in transplastomic studies (Lössl and Waheed, 2011).

1.8 Chloroplast genome structure

The chloroplast is a double membrane bound organelle and contains its double strand DNA (Palmer, 1985) in up to 107 kb to 218 kb in size (Daniell *et al.*, 2016). The chloroplast genome is usually circular but there are reports about the existence of chloroplast genome in linear form (Daniell *et al.*, 2016). The chloroplast genome is mostly a quadripartite, consisting of one large single copy (LSC), one small single copy (SSC) and two inverted repeat (IRa and IRb) regions (Palmer, 1985). One copy of inverted repeats has been lost in some species, making entire genome as single copy (Wu *et al.*, 2011). The contraction and expansion of the inverted repeats are also considered an important phenomena and leads to duplication of complete gene or generation of pseudogene due to which number of genes sometime varies within the species of similar lineages (Menezes *et al.*, 2018). Many mutational events occur within chloroplast genomes such as structural rearrangements, insertions and deletions (InDels), inversions, translocations and copy number variations (CNVs) (Jheng *et al.*, 2012). Chloroplast genome is considered prokaryotic which lost many genes in the course of evolution (Krupinska *et al.*, 2013). Chloroplast genome of majority of plant species consists of about 120 conserved genes including transfer RNAs (tRNA), ribosomal RNAs (rRNA) and protein-coding genes (Daniell *et al.*, 2016). Several recent studies have identified that intergenic spacer regions are the most

polymorphic regions within the chloroplast genome (Chen *et al.*, 2018; Choi *et al.*, 2016; Menezes *et al.*, 2018; Zhang *et al.*, 2016).

Introns within the genes are also conserved like the number and types of genes in chloroplast genome (Daniell *et al.*, 2016). However, loss of introns is also observed in many species of angiosperms (monocot and eudicot) and gymnosperms (Downie *et al.*, 1991; Jansen *et al.*, 2007). Recent studies reported intron loss in *Astragalus membranaceus* (Lei *et al.*, 2016), *Lagerstroemia fauriei* (Gu *et al.*, 2016) and *Lonicera japonica* (He *et al.*, 2017). The protein-coding genes in which loss of intron/introns is/are reported include an RNA polymerase (*rpoC2*), ATP synthase (*atpF*), ribosomal proteins (*rpl2*, *rps12*, and *rps16*), and a clp protease (*clpP*) (Downie *et al.*, 1991; Gu *et al.*, 2016; He *et al.*, 2017; Jansen *et al.*, 2007; Lei *et al.*, 2016).

1.9 Polymorphism in chloroplast genome sequences

Polymorphisms in chloroplast genome are used for analysing deep divergence (at genus and family level) to study population genetics for inferring of phylogeny and domestication of plant lineages (Ahmed, 2014; Ahmed *et al.*, 2013; Amiryousefi *et al.*, 2018; Henriquez *et al.*, 2014; Jansen *et al.*, 2007; Menezes *et al.*, 2018; Pfeil *et al.*, 2002). The earlier studies of phylogeny inferring used partial sequences of the chloroplast genome for the resolution of phylogenetic relationships both at genus and family level (Alverson *et al.*, 1999; Baum *et al.*, 2004; Pfeil *et al.*, 2002). The use of multiple sequences in phylogeny can provide sufficient information for the inferring of phylogeny (Ahmed, 2014). However, previously available information was not enough to develop authentic and cost-effective markers to resolve taxonomic discrepancies of closely related taxa or taxa with complex taxonomy (Daniell *et al.*, 2016). Now, the availability of complete chloroplast genome sequencing provides us the opportunity to identify suitable loci for the development of authentic and cost-effective markers. Previously, Ahmed (2014) identified suitable loci by comparing two morphotypes of *Colocasia esculenta* and developed suitable markers for the study of population genetics and domestication of *Colocasia esculenta*. So, the polymorphism of chloroplast genome sequences can be used for the identification of cultivars and domestication studies. The determination of phylogenetic relationship among the domesticated species and their wild species is also important for the enhancement of breeding and chloroplast genome polymorphism can be helpful for the identification of compatible wild relatives for the transferring of desired genes through breeding (Daniell *et al.*, 2016).

1.10 Adaptive evolution and domestication

The information of chloroplast genome sequence is important for understanding the adaptive evolution and domestication of plant species (Daniell *et al.*, 2016; Menezes *et al.*, 2018). The

chloroplast genome structure of many species of legumes and *Citrus* sheds light on the domestication of these species with the interesting genome rearrangements and inversions.

The chloroplast genome of legumes is interesting due to multiple rearrangements including inversions of large inverted segments and loss of inverted repeats (Palmer *et al.*, 1988). For instance, a 51 kb inversion was reported in soybean (*Glycine max*) (Saski *et al.*, 2005) and other species of the subfamily Papilionoideae, family Fabaceae (Cai *et al.*, 2008; Magee *et al.*, 2010; Sabir *et al.*, 2014; Sherman-Broyles *et al.*, 2014); A 78 kb inversion was confirmed in *Phaseolus* and *Vigna* (Guo *et al.*, 2007; Tangphatsornruang *et al.*, 2010). Moreover, In the 51 kb inversion of soybean, 36 kb (Martin *et al.*, 2014) and 5.6 kb (Kazakoff *et al.*, 2012) inversions were also reported. The regions with these inversions contain many important genes, but none of the gene was disturbed due to the rearrangement of the chloroplast genome nor any adverse effect was noted on the survival of the species, which revealed the role of these inversions in adaptation of the species (Daniell *et al.*, 2016). These unique characteristics are useful in inferring of phylogeny (Schwarz *et al.*, 2015) as well as in chloroplast transformation of legumes (Daniell *et al.*, 2016). The insight into the chloroplast genome structure might be helpful for designing of suitable primers to amplify these sequences for further domestication and phylogenetic analysis.

Citrus is a genus of commercially important fruits. The chloroplast genome of first species of *Citrus*, sweet orange (*Citrus sinensis*), was published in 2006 (Bausher *et al.*, 2006) which provided basis for further research. Phylogenetic analysis of the 28 species of *Citrus* and 6 species of *Citrus*-related genera based on chloroplast genome revealed their origination from a common ancestor (Carbonell-Caballero *et al.*, 2015; Caspermeyer, 2015). The high rates of substitutions and InDels were observed in four genes (*ccsA*, *matK*, *ndhF*, and *ycf1*) of these species as compared to average that revealed existence of positive selection pressure on these genes. The sequence of *matK* gene is often used for inferring of phylogeny and encodes protein maturase that is involved in splicing of type II introns (Carbonell-Caballero *et al.*, 2015). The positive selection of *matK* in the 30 species of *Citrus* and the other related plants group indicated its role in the stresses that are faced by plants in different ecological niches (Chen and Xiao, 2010). The *ndhF* gene encodes a subunit of the chloroplast NAD(P)H dehydrogenase (NDH) complex. The NDH monomers of chloroplast genome are sensitive to high light stress. Hence, observation suggests the role of *ndhF* gene in the acclimation of stress (Peng *et al.*, 2011). Some other studies also relate the *ycf1* gene with the stresses and linked positive selection pressure of *matK*, *ndhF*, and *ycf1* to adaptation of species of hot and dry climate (Carbonell-Caballero *et al.*, 2015; Caspermeyer, 2015).

1.11 Transfer of chloroplast genes to nuclear or mitochondrial genomes and vice versa

Plant cell contains three distinct genomes: nuclear, mitochondrial and plastid. Mitochondria are believed to evolve from a single endosymbiotic event by the uptake of a proteobacterium whereas chloroplast evolved from endosymbiosis of a cyanobacterium after which there was a massive transfer of genes from the chloroplast to the nucleus (Timmis *et al.*, 2004). The translation system varies for these three genomes of plants. Many nuclear encoded genes are translated in cytosol and the protein product is then transported to various other organelles for functions including chloroplast (Li and Chiu, 2010). The protein encoded by chloroplast are directly synthesized within the chloroplast (Daniell *et al.*, 2016). Multi-subunit functional protein complexes that are involved in photosynthesis or protein synthesis are also assembled within chloroplasts (Daniell *et al.*, 2016). The gene content is mostly conserved in the chloroplast genome of angiosperm species (Barrett *et al.*, 2014; Kim *et al.*, 2018) but missing of some protein coding genes are also reported (Jansen *et al.*, 2007). However, further studies revealed transfer of these genes to the nuclear or mitochondrial genome. This transferring provides valuable information for evolutionary studies and phylogenetic analyses and can be identified by analyses of complete chloroplast genome sequences (Daniell *et al.*, 2016). The *infA* (translation initiation factor 1) of the chloroplast is a homolog of the *infA* gene of *Escherichia coli* (Cummings and Hershey, 1994; Millen *et al.*, 2001). This gene initiates translation along two nuclear encoded initiation factors to mediate interactions between mRNA, ribosomes, and initiator tRNA-Met (Millen *et al.*, 2001). The loss of *infA* gene has been reported from the chloroplast genome of many angiosperms during the course of evolution (Daniell *et al.*, 2016; Millen *et al.*, 2001). Furthermore, in many angiosperms' species the *infA* gene has been encoded by the nuclear genome include soybean (*Glycine max*), tomato (*Solanum lycopersicum*), *Arabidopsis thaliana*, and ice plant (*Mesembryanthemum crystallinum*) (Millen *et al.*, 2001). The transfer of essential gene *rpl22* was reported from the chloroplast genome into the nuclear genome in 57 species of 26 genera (Gantt *et al.*, 1991; Jansen *et al.*, 2011). The nuclear encoded *rpl22* also contains a transit peptide which is used for delivery of this protein from the cytosol of cell to the chloroplast. The transfer of *rpl32* gene is also reported into the nuclear genome (Cusack and Wolfe, 2007; Park *et al.*, 2015; Ueda *et al.*, 2007). The *ndh* subunits are involved in photosynthesis and mediate cyclic electron transport chain (Peltier and Cournac, 2002). These subunits are encoded by 11 *ndh* genes in chloroplast genome but the lack of functional copies of these essential genes was also reported in some plant species (Braukmann *et al.*, 2009; Chris Blazier *et al.*, 2011; McCoy *et al.*, 2008; Pan *et al.*, 2012; Sanderson *et al.*, 2015; Shahinnia and Sayed-Tabatabaei, 2009; Wakasugi *et al.*, 1994; Weng *et al.*, 2014; Wu *et al.*, 2010; Yang *et al.*, 2013), which supports the existence

of functional copies of these genes in nuclear or mitochondrial genome. In the genome of Orchid, deletion of the entire *ndhH* gene family has been reported instead of single gene (Lin *et al.*, 2015). Furthermore, DNA fragments of the some Orchid *ndh* genes were identified in the mitochondrial genome in form of pseudogene (Lin *et al.*, 2015). However, the observance of normal photosynthesis along with production of carbohydrate in these species revealed the transferring of these genes to other genomes in functional form (Burrows, 1998; Kofler *et al.*, 1998; Lin *et al.*, 2015; Ruhlman *et al.*, 2015; Shikanai *et al.*, 1998; Takabayashi *et al.*, 2002). The deletion of many other genes from chloroplast genome and its transferring to nuclear or mitochondrial genome was also reported in plant species include *accD*, *psaI*, *rpl20*, *rpl23*, *rpl33*, *rps16*, *rpoA*, and, *ycf1*, *ycf2*, and *ycf4* (Daniell *et al.*, 2016). Despite transferring of genes from the chloroplast genome to nuclear and mitochondrial genomes, the transferring of the gene of mitochondria or part of DNA of nuclear genome was also reported to the chloroplast genome (Ma *et al.*, 2015; Smith, 2014; Wysocki *et al.*, 2015).

1.12 Role of chloroplast genome in phylogenetic studies

The chloroplast genome has appropriate polymorphisms, lack meiotic recombination, and exhibits uniparental inheritance (Daniell, 2007; Palmer, 1985) which make it a valuable tool for inferring of phylogeny from population genetic to deep divergence (Ahmed *et al.*, 2013; Henriquez *et al.*, 2014; Nguyen *et al.*, 2018). Previously, complete gene or some part of gene or non-coding part of chloroplast genome was employed to infer phylogeny (Alverson *et al.*, 1999; Koopman and Baum, 2008; Tr *et al.*, 2016). The advancement of Next Generation Sequencing (NGS) technology makes it feasible to study the evolutionary dynamics of complete chloroplast genome, polyploidy events, domestication, phylogeny and development of conservation strategies. The *Gossypium* species have complex genome and comprise 8 genome groups from A to G, and K but their evolutionary dynamics, domestication, polyploidy events, and phylogeny were successfully analysed based on complete chloroplast genome in recent years (Chen *et al.*, 2017; Z. Chen *et al.*, 2016; Wu *et al.*, 2018; Xu *et al.*, 2012). The high resolution phylogeny tree of order Bryopsidales based on complete chloroplast genome sequences provided new insight into the evolution of many families of the Bryopsidales and new classification was proposed for families Rhipiliaceae, Udoteaceae, and Pseudocodiaceae (Cremen *et al.*, 2019). The *Dipteronia* and *Acer* are two sister genera of the family Sapindaceae, but their phylogenetic position was not confirmed. The phylogenetic analyses based on complete chloroplast genome data not only resolved the phylogeny of these two genera but also identified *Dipteronia* species among the East Asian flora (Feng *et al.*, 2019). Family Ranunculaceae is considered among the early diverging eudicots and consists of 14

tribes from which the phylogeny of 11 tribes were controversial (Zhai *et al.*, 2019). The phylogenetic analyses of 35 species from 31 genera of 14 tribes, based on complete chloroplast genomes, resolved the discrepancies in phylogeny and revealed the unsuitability of the previously used morphological characters for phylogeny due to parallel, convergent or even reversal evolution (Zhai *et al.*, 2019).

Certain taxonomic discrepancies also exist at genus level. Due to advancement of NGS technology, many researchers follow a new approach and identify mutational hotspots for the development of suitable markers to resolve taxonomic discrepancies by comparison of complete chloroplast genomes. Li *et al.* (2018) compared seven species of *Fritillaria* and suggested mutational hotspots for the development of suitable markers to resolve phylogeny of the genus *Fritillaria*. Cheon *et al.* (2019) made comparison of four species of *Violaceae* and suggested polymorphic loci for development of authentic and robust molecular markers for the resolution of phylogeny. Menezes *et al.* (2018) compared two species of genus *Byrsonima* of family Malphigeaceae and suggested mutational hotspots for resolution of taxonomic and phylogenetic relationships at genus level as well as at family level. The similar approach was also used in several other studies (Bi *et al.*, 2018; Choi *et al.*, 2016; Yu *et al.*, 2017). Here, we were also interested in the identification of mutational hotspots by comparing species of genera *Theobroma*, *Firmiana* and *Hibiscus* to provide a basis for the development of suitable markers to resolve taxonomic discrepancies at genus and/or family level.

1.13 Repeats in chloroplast genome

Repeats are present in chloroplast genome usually in the form of oligonucleotide repeats and simple sequence repeats (Bi *et al.*, 2018; Z. Chen *et al.*, 2016; Menezes *et al.*, 2018; Qian *et al.*, 2013; Wu *et al.*, 2018). The oligonucleotide repeats are usually present in four forms including forward, reverse, palindromic and complementary repeats, and are mostly evaluated with the minimum repeat size of 20-30 bp in the chloroplast genome (Asaf *et al.*, 2018; Kurtz *et al.*, 2001; Menezes *et al.*, 2018; Yu *et al.*, 2017). The role of moderate repeats 14-48 bp were suggested in the generation of InDels (Kawata *et al.*, 1997) and inversions (Kim and Lee, 2005; Whitlock *et al.*, 2010). Furthermore, based on statistical analyses, the oligonucleotide repeats were also suggested as proxy for identification of mutational hotspots (Ahmed *et al.*, 2012).

Simple sequence repeats (SSRs) are also present in the chloroplast genome. Recent studies of chloroplast genome revealed that mononucleotide SSRs are present mostly in 7-16 repeat units, dinucleotides in 4 repeat units, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide repeats are mostly present in 3 units. Moreover, the abundance of the repeats with composition of A/T having abundance due to the A/T rich chloroplast genome (Dong *et*

al., 2016; He *et al.*, 2016; Ma, 2018; Menezes *et al.*, 2018; Zhang *et al.*, 2016). The SSRs of chloroplast genome are used for the studies of population genetics and barcoding of species and cultivars (Huang *et al.*, 2018; Joh *et al.*, 2017; Nguyen *et al.*, 2018; Qiu *et al.*, 2013). Therefore, different researchers identified and reported SSRs loci in chloroplast genome of various species (Curci *et al.*, 2015; Huang *et al.*, 2018; Liu *et al.*, 2018; Qiu *et al.*, 2013; Yu *et al.*, 2017). The contraction and expansion of SSRs units are caused by slipped-strand mispairing during DNA replication (Levinson and Gutman, 1987).

1.14 Comparisons of mutation rates among nuclear, mitochondrial and chloroplast genome

Comparative studies of mutation rates among nuclear, chloroplast and mitochondrial DNA sequences revealed that rates of silent substitutions of mitochondrial DNA is less than one-third of chloroplast DNA and the rate of evolution of mitochondrial DNA is half as compared to nuclear DNA (Wolfe *et al.*, 1987). This might be due to slower rate of evolution in the mitochondrial DNA as compared to the chloroplast DNA (Wolfe *et al.*, 1987). The rate of synonymous substitutions is higher as compared to non-synonymous substitution in the protein coding sequences of chloroplast (Choi *et al.*, 2016; Menezes *et al.*, 2018). Furthermore, the higher substitution rate has been observed in non-coding sequences as compared to coding sequences that might be due to low selection pressure (Cai *et al.*, 2015; Choi *et al.*, 2016; Menezes *et al.*, 2018; Saina *et al.*, 2018a; Yang *et al.*, 2016). Lineage directed fluctuations have also been described based on comparative analyses of chloroplast genomes (Guo *et al.*, 2007). Studies showed reduction in the rate of evolution of chloroplast in dicots after monocots and dicot split and reduction in rate of evolution was more noticeable at non-synonymous sites instead of synonymous (Wolfe *et al.*, 1987). Many factors affect the mutation rates including: (i) how fast and accurately DNA repair machinery works, (ii) the fidelity of DNA replication; and (iii) intrinsic qualities of the nucleotides for being stable and their exposure and responses to mutagens. The overall correlation of these factors determines the rate of evolution throughout the genome (Baer *et al.*, 2007).

1.15 Hypotheses about the origination of mutations and their co-occurrence

Certain hypotheses have been presented about the generation of mutational events in the genomes. The time reversible substitution model has been commonly considered as an authentic model for the study of evolution in the sequences of chloroplast genome in which mutational events occur independently at each site (Drouin *et al.*, 2008). However, some studies revealed the complex process of evolution due to occurrence of non-random spatial patterns and lineage-specific substitutions (Gruenheit *et al.*, 2008; Liò and Goldman, 1998;

Zhong *et al.*, 2011). These studies revealed the significant limitation of the use of chloroplast genome sequences for phylogenetic analyses in deep divergence (Gruenheit *et al.*, 2008). In comparative analyses of closely related taxa, these studies help to understand the dynamics of mutational events taking place in the polymorphic regions of the chloroplast genome (Ahmed *et al.*, 2013; Choi *et al.*, 2016; Li *et al.*, 2018).

Analyses of DNA sequences in prokaryotes and eukaryotes show co-occurrence of substitutions and InDels, and several hypotheses have been proposed for its explanation. The first hypothesis suggests that some regions are specifically pre-disposed for mutations in the form of substitutions and InDels “the regional difference hypothesis” (Hardison *et al.*, 2003; Silva and Kondrashov, 2002). This hypothesis was suggested based on presence of higher rate of mutation in some regions of genome as compared to other regions. The second hypothesis suggests that large InDels induce substitutions during DNA repair process by recruiting error-prone DNA polymerase (Tian *et al.*, 2008; Zhu *et al.*, 2009). This hypothesis was suggested based on the existence of substitutions close to the InDels regions. The third hypothesis emphasize that oligonucleotide repeats originate substitutions and InDels by arresting replication fork that leads to recruitment of error-prone DNA polymerase (McDonald *et al.*, 2011). This hypothesis was presented based on the comparative analyses of prokaryotic and eukaryotic genomes. In these analyses, McDonald *et al.* (2011) identified 2/3 of the InDels in close association with oligonucleotide repeats. The fourth hypothesis was proposed based on statistical analysis and revealed co-occurrence among substitutions, InDels and oligonucleotide repeats in chloroplast genomes among species of family Araceae of angiosperms (Ahmed *et al.*, 2012) and genus *Cephalotaxus* of family Cephalotaxaceae, gymnosperms (Yi *et al.*, 2013). This fourth hypothesis has since been verified in the limited number of species in monocot family Araceae (angiosperms) and family Cephalotaxaceae (gymnosperms).

1.16 Status of complete chloroplast genome-based research in the family Malvaceae

The deep comparison of chloroplast genome among the species of the family Malvaceae was not reported. Till date (13 November 2019), the intra-genus level comparative analyses in two genera were reported. The extensive comparative analyses of about 40 *Gossypium* species were performed and the chloroplast genome structure, polyploidy level, domestication, and evolution of repeats were elucidated in different studies (Chen *et al.*, 2017; Z. Chen *et al.*, 2016; Ibrahim *et al.*, 2006; Lee *et al.*, 2006; Wu *et al.*, 2018; Xu *et al.*, 2012). The comparative analyses among four species of the genus *Tilia* was performed and an insight was provided about the chloroplast genome structure and their phylogeny (Cai *et al.*, 2015). The chloroplast genome characteristics of many other species were also reported without detailed analyses

including *Bombax ceiba* (Gao *et al.*, 2018), *Durio zibethinus* (Cheon *et al.*, 2017), *Heritiera parvifolia* (Xin *et al.*, 2018), *Heritiera angustata* (Zhao *et al.*, 2018), *Hibiscus syriacus* (Kim *et al.*, 2019; Kwon *et al.*, 2016) *Theobroma cacao* (Jansen *et al.*, 2011), *Theobroma grandiflorum* (Kane *et al.*, 2012), *Firmiana major* (Ya *et al.*, 2017), and *Firmiana pulcherrima* (Wang *et al.*, 2017). Furthermore, the chloroplast genome of five other species are also available in GenBank of National Center for Biotechnology information (NCBI) including *Talipariti hamabo* (KR259988.1), *Abelmoschus esculentus* (KY635876.1), *Firmiana simplex* (MH671308.1), *Althaea officinalis* (KY085914) and *Reevesia thyrsoidea* (MH939148.1).

This data revealed the need of comparative analyses of chloroplast genomes at family level. Hence, the detailed comparative analyses of chloroplast genomes along with some other species can provide insight into to the evolutionary dynamics of family Malvaceae. Our study will be helpful to understand the evolution and phylogeny of family Malvaceae. Furthermore, identification of suitable polymorphic loci for development of robust and cost-effective markers will provide a basis for the phylogeny inference and resolving of taxonomic discrepancies at genera and family levels.

1.17 Aims and objectives

The aims and objectives of the current research were to:

- sequence and/or *de novo* assemble the complete chloroplast genomes of eight species for the enhancement of chloroplast genome resources of the family Malvaceae.
- perform the comparative analyses of the chloroplast genome structure of the newly assembled and previously reported species to get insight into the evolutionary dynamics of chloroplast genome in the family Malvaceae.
- determine the rate of evolution of protein coding genes based on the rate of synonymous and non-synonymous substitutions in the studied species of family Malvaceae.
- evaluate correlations among substitutions, InDels and oligonucleotide repeats to understand the evolutionary relationships among these mutational events in the chloroplast genomes of family Malvaceae (eudicots).
- infer phylogeny of the family Malvaceae by reconstruction of high resolutions tree based on complete chloroplast genome.
- perform comparative analyses of three genera: *Theobroma*, *Firmiana* and *Hibiscus* to get insight into inter-genus and intra-genus level similarities and variations.
- identify suitable polymorphic loci for the development of robust, authentic and cost-effective markers to infer phylogeny and resolve taxonomic discrepancies at genus and family level.

Chapter 2
***De novo* assembly and**
characterisation of chloroplast
genomes of eight family Malvaceae
species

2.1 Introduction

The chloroplast is a self-replicating organelle which plays a vital role in photosynthesis (Cooper, 2000). Chloroplasts are inherited from single parent, maternally in most angiosperms species (Daniell, 2007). However, paternal inheritance in some gymnosperms is also reported (Neale and Sederoff, 1989). It is a double membrane bound organelle and contains its double stranded DNA in up to 107 kb to 218 kb in size. Chloroplast genome in majority of plant species consists of about 120 genes including transfer RNA (tRNA), ribosomal RNA (rRNA), and protein-coding genes (Daniell *et al.*, 2016). The chloroplast genome is mostly a quadripartite structure, consisting of a large single copy (LSC), a small single copy (SSC), and two inverted repeat regions (IRa and IRb) (Palmer, 1985). One copy of inverted repeats has been lost in some species, making entire genome as single copy (Wu *et al.*, 2011). Moreover, linear chloroplast genome structure was also reported (Oldenburg and Bendich, 2016).

The advancement of NGS technology makes it feasible to sequence and *de novo* assemble complete chloroplast genome in robust and cost-effective way. The sequencing and assembly of chloroplast genome were used for the study of evolutionary dynamics polyploidy events, domestication, for development of conservation strategies and for resolution of phylogeny of complex taxonomy (Chen *et al.*, 2017; Z. Chen *et al.*, 2016; Cremen *et al.*, 2019; Feng *et al.*, 2019; Wu *et al.*, 2018; Xu *et al.*, 2012; Zhai *et al.*, 2019).

Complete chloroplast genome of the species of few Malvaceae genera are available. The complete chloroplast genome of 40 species of genus *Gossypium* (Chen *et al.*, 2017; Z. Chen *et al.*, 2016; Ibrahim *et al.*, 2006; Lee *et al.*, 2006; Wu *et al.*, 2018; Xu *et al.*, 2012) and four species of genus *Tilia* are available (Cai *et al.*, 2015). Some other species for which complete chloroplast genome has been reported include *Bombax ceiba* (Gao *et al.*, 2018), *Durio zibethinus* (Cheon *et al.*, 2017), *Heritiera parvifolia* (Xin *et al.*, 2018), *Heritiera angustata* (Zhao *et al.*, 2018), *Hibiscus syriacus* (Kim *et al.*, 2019; Kwon *et al.*, 2016) *Theobroma cacao* (Kane *et al.*, 2012), *Theobroma grandiflorum* (Kane *et al.*, 2012), *Firmiana major* (Ya *et al.*, 2017), and *Firmiana pulcherrima* (Wang *et al.*, 2017). Furthermore, the chloroplast genome of five other species is also available in GenBank of NCBI including *Talipariti hamabo* (KR259988.1), *Abelmoschus esculentus* (KY635876.1), *Firmiana simplex* (MH671308.1), *Althaea officinalis* (KY085914) and *Reevesia thyrsoidea* (MH939148.1), as accessed on 08 April 2019. The broad level study of evolutionary dynamics of family Malvaceae and identification of mutational hotspots required more genetic resources. Here, we also sequenced and/or *de novo* assembled chloroplast genomes of eight species to increase genomic resources of family Malvaceae. The brief introduction of these species is given below.

2.1.1 *Hibiscus rosa-sinensis*

Hibiscus rosa-sinensis is commonly known as China rose (Pfeil and Crisp, 2005). This species belongs to tribe Hibisceae and subfamily Malvoideae (Pfeil *et al.*, 2002; Pfeil and Crisp, 2005). It is grown throughout the tropics and subtropics due to its ornamental and medicinal values (Prasad, 2014). Many of its varieties and cultivars are available with same morphology except flower colour that ranges from yellow or white to pink or red with single or double petals, but the flower is not available throughout the year which makes the identification of the cultivar almost impossible (Prasad, 2014). The extensive medicinal activities of *H. rosa-sinensis* have also been reported. For instance, antimicrobial, antioxidant, anti-tumour, anti-diabetic and wound healing (Mondal *et al.*, 2016; Prasad, 2014). Different cultivars and varieties varied in their antimicrobial and antioxidant activities toward different pathogenic species (Patel *et al.*, 2012; Prasad, 2014). Therefore, the identification of its cultivars is important in the deciding of the herbal medicines and the complete chloroplast genome sequence in this study might provide resources for the marker's development to identify different cultivars of *H. rosa-sinensis*.

2.1.2 *Hibiscus mutabilis*

Hibiscus mutabilis is commonly known as Confederate rose (Pfeil and Crisp, 2005). This species belongs to tribe Hibisceae and subfamily Malvoideae (Pfeil and Crisp, 2005). It is native to China and used as an ornamental due to large and showy flowers (Li *et al.*, 2015). Certain medicinal properties of *Hibiscus mutabilis* are also reported. In particular, its leaves are used as an anodyne, antidote, refrigerant, and expectorant (Xie *et al.*, 2011). Leaves of *H. mutabilis* contain the most important bioactive flavonoid glycosides such as Rutin (quercetin-3-rutinoside) (Yao *et al.*, 1994) which has a significant role in health benefits due to antihypertensive, antihyperglycemic, antifungal and antioxidative properties (Han, 2009; Lee *et al.*, 2007). Furthermore, it also reduces the risk of cardiovascular and kidney diseases (Hu *et al.*, 2009). The reduction in proliferation of HIV-1 virus is also reported due to its inhibitory effect on reverse transcriptase of the virus (Lam and Ng, 2009).

2.1.3 *Malva parviflora*

Malva is a genus of tribe Malveae of subfamily Malvoideae (Hinsley, 2004). According to plant list (<http://www.theplantlist.org/>; accessed on 19 May 2019), *Malva* genus consists of 31 accepted species. The species of this genus are distributed in a wide range in the world including North Africa, Southern Europe, and Southwest Asia (Bayer and Kubitzki, 2003). *Malva* are short-lived perennial, biennial or annual herbs (Bayer and Kubitzki, 2003). The

flowers are pink to purple or pink to bluish, or white (Bayer and Kubitzki, 2003). The stems and foliage are typically downy or hairy (Hinsley, 2004). The fruits consist of a divided capsule containing a ring of nutlets (Bayer and Kubitzki, 2003). Many species are used in horticulture and medicine (Hinsley, 2004). The species of genus *Malva* were reported with extensive medicinal properties including anti-diabetic, antioxidant, anti-inflammatory, radical scavenger, metal chelating and burn healing activities (Akbar *et al.*, 2014; Bouriche *et al.*, 2011; Dalar *et al.*, 2012; DellaGreca *et al.*, 2009; Prudente *et al.*, 2013; Razavi *et al.*, 2011; Shale *et al.*, 2005). The medicinal properties of *Malva parviflora* are similar to its genus, including anti-inflammatory, anti-diabetic, anti-bacterial, free radical scavenger and metal chelating activities (Akbar *et al.*, 2014; Bouriche *et al.*, 2011; Shale *et al.*, 2005). Furthermore, *Malva parviflora* is also used as a leafy vegetable (Hinsley, 2008).

2.1.4 *Malvastrum coromandelianum*

Malvastrum belongs to tribe Malveae and subfamily Malvoideae (Bayer and Kubitzki, 2003). According to plant list (<http://www.theplantlist.org/>: accessed on 19 May 2019), this genus contains 27 accepted species with distribution in a wide range of the world. The species of this genus are herb or subshrubs (Bayer and Kubitzki, 2003). *Malvastrum coromandelianum* (L.) commonly known as false mallow, broom weed and clock plant (Sanghai *et al.*, 2013). Various medicinal properties of *Malvastrum coromandelianum* are reported including hypoglycaemic, analgesic, anti-inflammatory, antinociceptive and anti-bacterial activities (Khonsung *et al.*, 2006; Sanghai *et al.*, 2013; Sittiwet *et al.*, 2008).

2.1.5 *Urena procumbens*

Urena is a genus of tribe Hibisceae and subfamily Malvoideae (Bayer and Kubitzki, 2003). According to the plant list (<http://www.theplantlist.org/>: accessed on 19 May 2019), this genus includes 12 accepted species, whereas the taxonomy of about 60 species are still inconclusive. The plant of this genus are shrubs and distributed in tropical to subtropical parts of the world (Bayer and Kubitzki, 2003). *Urena procumbens* is native to China and commonly known as Fan-tian-hua (L. Chen *et al.*, 2016). It is also grown in many tropical countries including America, Australia and South Africa. In traditional Chinese medicine, *Urena procumbens* is used as antipyretics, cough expectorant, diuretic and in the treatment of rheumatic disease (L. Chen *et al.*, 2016).

2.1.6 *Firmiana colorata*

Firmiana is a small genus that includes sixteen species in accepted state, whereas five species are categorised as unresolved (<http://www.theplantlist.org/>: accessed on December 29, 2018). *Firmiana* genus includes deciduous trees (rarely shrubs) that are distributed in Eastern Africa and Eastern-South and East Asia (Kostermans, 1957). Eight species of *Firmiana* inhabit China (Huang *et al.*, 2011). Some species of the genus *Firmiana* are cultivated for their beautiful shape and lovely flowers (Fan *et al.*, 2013). With reference to medicinal perspective, *Firmiana* species had shown antimicrobial (Ajaib *et al.*, 2014), anti-inflammatory (Lim *et al.*, 2017) and anti-cancer (Woo *et al.*, 2015) properties. Some species showed neuroprotective (Lim *et al.*, 2017) and hepatoprotective effects (Kim *et al.*, 2015). *Firmiana colorata* is used for the intestinal dysfunction in some tribes of Bangladesh (Azam *et al.*, 2013). Moreover, the species of *Firmiana* are also used for the caffeine containing teas (Bayer and Kubitzki, 2003).

2.1.7 *Sterculia monosperma*

Sterculia genus belongs to subfamily Sterculioideae (Wilkie *et al.*, 2006). This is a pantropical genus and consists of about 200-300 species (Bayer and Kubitzki, 2003). According to plant list (<http://www.theplantlist.org/>: accessed on 19 May 2019), 92 species are accepted so far whereas 287 species are categorised as unresolved. *Sterculia monosperma* is commonly known as China-chestnuts (Berry, 1982). This species is used for medicinal purposes in traditional Chinese medicines (Lim and Lim, 2012).

2.1.8 *Pterospermum truncatolobatum*

Pterospermum is the genus of subfamily Dombeyoideae and consists of about 18 species (Bayer and Kubitzki, 2003). The species of this genus are shrubs or trees and distributed in tropical Asia from India to Taiwan and Philippines (Bayer and Kubitzki, 2003). According to plant list (<http://www.theplantlist.org/>: accessed on 19 May 2019), the taxonomy of about 63 species of *Pterospermum* is unresolved yet. *Pterospermum truncatolobatum* is a tree growing up to 25 meters tall and reaches up to 80 cm in diameter. This species is distributed in East Asia, Southern China and in Southern Vietnam (Bayer and Kubitzki, 2003). In traditional medicine, it is used as poultice against itch and to treat wounds. The assemble chloroplast genome of this species will act as the first representative of the genus *Pterospermum* as well as for subfamily Dombeyoideae.

We aimed to assemble chloroplast genomes of eight species of Malvaceae to increase genomic resources for this family. We assembled chloroplast genomes for the first time for the species of five genera including *Malva*, *Malvastrum*, *Sterculia*, *Urena* and *Pterospermum*.

Furthermore, the genome of *Pterospermum* is also the first representative of subfamily Dombeyoideae. The sequencing and/or assembly of the complete chloroplast genomes of these eight species will further enhance our knowledge about the evolutionary dynamics of family Malvaceae at genera, tribes and subfamilies levels.

2.2 Materials and methods

2.2.1 Plant collection

The plants of four species which belong to three genera of family Malvaceae were collected from July-September from two different regions of Pakistan. *Hibiscus mutabilis* was collected from a nursery in Islamabad, Pakistan. *Malva parviflora* grows in wild and was collected from Nowshera, Khyber Pakhtunkhwa, Pakistan. *Malvastrum coromandelianum* also grows in wild and was collected from Quaid-i-Azam, University, Islamabad, Pakistan. The species of *Hibiscus rosa-sinensis* is cultivated for ornamental purposes and was collected from Quaid-i-Azam, University, Islamabad, Pakistan.

2.2.2 DNA extraction

Freshly growing leaves without any apparent disease symptoms were used for whole genomic DNA extraction. The leaves were washed with double distilled water and ethanol prior to DNA extraction. The DNA was extracted following Ahmed *et al.* (2009) with some modifications. The complete procedure as provided as follow:

1. Prior to DNA extraction, water bath was turned on and set at 65°C.
2. A small piece of leaf tissue of about 1 cm² was taken and grind in 2.5 ml extraction buffer.
3. After careful crushing, 750 µl of the homogeneous solution was taken in microcentrifuge tube. The 50 µl SDS (Sodium dodecyl sulphate) and 1 µl 2-Mercaptoethanol were added to microcentrifuge tube prior to vortex for 5-10 seconds.
4. The microcentrifuge tube was incubated for 30 minutes in water bath at 65°C and was agitated gently after 5 minutes.
5. After 30 minutes incubation, 750 µl phenol:chloroform:isoamyl alcohol (25:24:1) was added to each sample and centrifuged at 14,000 rpm for 8 minutes.
6. The supernatant (about 600-650 µl) was transferred to a new sterile microcentrifuge tube and incubated at 37°C for 20 minutes after addition of 10 µl RNase (10 mg/ml).
7. 600-650 µl chloroform:isoamyl alcohol (24:1) was added to the samples for removal of RNase.

8. 550-600 µl of supernatant was transferred to a new microcentrifuge and equal amount of chilled isopropanol was added for precipitation of DNA.
9. The sample was incubated for about 30 minutes at -20°C.
10. The sample was centrifuged at 14,000 rpm for 8 minutes to form pellet.
11. The solution was removed from the centrifuge and the pellet was washed twice with 70% alcohol at 10,000 rpm.
12. The pellet was air dried and then dissolved in 30-40 µl Tris-EDTA (Tris-HCl: 10 mM, EDTA: 1 mM).
13. The quality and quantity of DNA were confirmed with 1% agarose gel and nanodrop (Thermo Scientific).

2.2.3 Sequencing

Whole genomic DNA of four species including *Hibiscus rosa-sinensis*, *Hibiscus mutabilis*, *Malva parviflora*, and *Malvastrum coromandelianum* was sent in lyophilized form to Novogene, Hong Kong for sequencing after confirmation of the quality and quantity of genomic DNA by Nanodrop and 1% agarose gel. The sequencing of whole genomic DNA was performed at low coverage depth at Novogene, Hong Kong, where the Illumina HiSeq2500 was used to generate whole genome shot gun of 150 bp short reads with 350 bp insert size.

2.2.4 Downloading of SRA data

Short reads data was downloaded from the SRA (Sequence Read Archive) database of the National Center for Biotechnology Information (NCBI) for four species. These species included *Urena procumbens*, *Firmiana colorata*, *Sterculia monosperma* and *Pterospermum truncatolobatum*. This data was sequenced by Beijing genomics institute (BGI) using whole genome sequencing strategy with insert size of 200 bp and short reads length of 100 bp using BGISEQ-500. The data are available under BioProject accession number PRJNA438407 with various SRR accession numbers (Table 2.1).

Table 2.1 Detail of accessions and data downloaded from SRA database

Species	SRR Accessions	Experiment accession	Downloaded data	Short reads
<i>Urena procumbens</i>	SRR7121570	SRX4043119	11.4 GB	43.5 million
<i>Firmiana colorata</i>	SRR7121997	SRX4043546	12.9 GB	49.1 million
<i>Sterculia monosperma</i>	SRR7121864	SRX4043413	11.7 GB	44.6 million
<i>Pterospermum truncatolobatum</i>	SRR7122099	SRX4043648	10.3 GB	39.3 million

2.2.5 Chloroplast genomes assembly and annotations

The quality of the sequencing data (short reads) was assessed using FastQC. The short reads were used to assemble the chloroplast genome of eight species (mentioned below) using Velvet 1.2.10 (Zerbino and Birney, 2008) with kmer 71,81 and 121. The kmer values 71 and 81 were used for the *de novo* assembly of *Urena procumbens*, *Firmiana colorata*, *Sterculia monosperma* and *Pterospermum truncatolobatum*, whereas kmer value 121 was used for *Hibiscus mutabilis*, *Hibiscus rosa-sinensis*, *Malvastrum coromandelianum* and *Malva parviflora*. The contigs generated by Velvet were further combined by using *de novo* assembly option of Geneious R8.1 (Kearse *et al.*, 2012) that generated complete chloroplast genomes in 3-8 long contigs that belonged to large single copy (LSC), small single copy (SSC) and inverted repeat (IR) regions. Full length chloroplast genomes were assembled by carefully inspecting repeats at the boundary regions of LSC, IR and SSC. The *de novo* assembled chloroplast genomes were annotated by using GeSeq (Tillich *et al.*, 2017) and Dual Organellar Genome Annotator (DOGMA) (Wyman *et al.*, 2004) whereas five column table for NCBI submission was generated by GB2Sequin (Lehwark and Greiner, 2019). The coverage depth of *de novo* assembled chloroplast genomes were determined by mapping their respective short reads to each respective assembled genome using BWA (Li and Durbin, 2009) and visualised by Tablet (Milne *et al.*, 2009). The annotated chloroplast genomes sequences of each species were submitted to NCBI with specific accession as: *Hibiscus rosa-sinensis* (MK382984), *Hibiscus mutabilis* (MK820657), *Malvastrum coromandelianum* (MK860037), *Malva parviflora* (MK860036), *Urena procumbens* (BK010727), *Firmiana colorata* (BK010724), *Sterculia monosperma* (BK010726) and *Pterospermum truncatolobatum* (BK010725). For the last four species, short reads data were taken from NCBI and the *de novo* assembled chloroplast genome with full annotations were submitted as third party annotation to GenBank. The main features of newly assembled chloroplast genomes were determined by using Geneious R8.1 (Kearse *et al.*, 2012) whereas the circular structures of these genomes were drawn by using OrganellarGenomeDRAW (Lohse *et al.*, 2007).

2.3 Results

2.3.1 DNA extraction

The extracted DNA of all the samples was run on 1% agarose gel which showed a single intact band which showing the integrity of the extracted DNA. Whereas the nanodrop analyses revealed the amount of DNA up to 50 ng/μl (Figure 2.1).

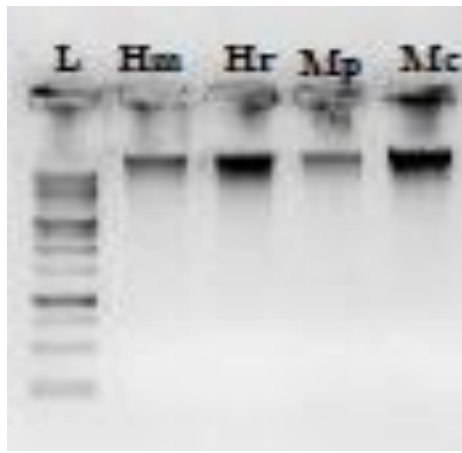


Figure 2.1 Genomic DNA extraction. L: ladder (10 kb), Hm: *Hibiscus mutabilis*, Hr: *Hibiscus rosa-sinensis*, Mp: *Malva parviflora*, and Mc: *Malva coromandelianum*

2.3.2 Whole genome shotgun and chloroplast genome assembly

The sequencing through Illumina Hiseq2500 produced about 13.1 GB (38.94 million reads) raw data of 150 bp short reads for *Hibiscus mutabilis*, 10 GB (28.93 million reads) for *Hibiscus rosa-sinensis*, 12.4 GB (36.85 million reads) for *Malva parviflora*, and 8.74 GB (25.9 million reads) for *Malvastrum coromandelianum*. The fastQC analyses revealed high quality of the data with an average Phred score of 34.2-37. The data was used for *de novo* assembly of chloroplast genomes of these species, and the high coverage depth was observed for *de novo* assembled chloroplast genomes after mapping of short reads. The number of mapped reads to assembled genome and average coverage depths observed for each species were: *Hibiscus mutabilis* (1.56 million, 1428X, respectively), *Hibiscus rosa-sinensis* (0.79 million, 725X, respectively), *Malva parviflora* (8.96 million, 8437X, respectively), and *Malvastrum coromandelianum* (0.97 million, 891X, respectively).

The genome assembled from the data downloaded from NCBI also showed high average coverage depth. For these four species the number of short reads mapped to each species and average coverage depths observed for each species were: *Firmiana colorata* (0.5 million, 310X, respectively), *Sterculia monosperma* (1.05 million, 650X, respectively), *Pterospermum truncatolobatum* (0.52 million, 323X, respectively), and *Urena procumbence* (1.03 million, 639X, respectively).

2.3.3 Chloroplast genome features

The *de novo* assembled chloroplast genome of all eight species exhibited similar quadripartite structure in which large single copy (LSC) region, and small single copy region (SSC) was separated by the Inverted repeat regions (IRa and IRb). The complete characteristics of the chloroplast genome of each species are given below.

2.3.4 Chloroplast genome features of *Hibiscus rosa-sinensis*

The chloroplast genome size of *H. rosa-sinensis* was 160,951 bp, comprising of two IRs (IRa and IRb, 25,598 bp each), separated by LSC (89,509 bp) and SSC (20,246 bp) that formed the quadripartite structure. The GC content of chloroplast genome varied among different regions of chloroplast genome such as LSC (34.9%), SSC (31.1%), and IR (42.9%). The GC content of the complete chloroplast genome was 37%.

The chloroplast genome of *Hibiscus rosa-sinensis* had 113 genes including 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes excluding truncated gene of *ycf1^ψ* (Table 2.2, Figure 2.2). Among these genes, 17 genes were duplicated in the IR regions. The duplicated genes in IR regions included 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. Eighteen genes contained introns that included 6 tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contained 5' part in the LSC region and 3' part in the IR region, so 3' part duplicated in the IR regions. Out of 18 genes, 16 genes had one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that contained introns were also duplicated in the IR regions. We found four partially overlapping genes *atpB/atpF* and *psbD/psbC*. The *ycf1* gene started from IR region and ended in SSC region, leaving a truncated copy of 123 bp in *H. rosa-sinensis*. The GC content of rRNAs (55.5%) was found higher as compared to tRNAs (53.2%) and protein coding genes (38.2%).

Table 2.2 Chloroplast gene content and functional classification of Malvaceae species

Category for gene	Group of gene	Name of gene					Number
Photosynthesis-related genes	Photosystem I	<i>psaA</i>	<i>psaB</i>	<i>psaC</i>	<i>psaI</i>	<i>psaJ</i>	5
	Photosystem II	<i>psbA</i>	<i>psbB</i>	<i>psbC</i>	<i>psbD</i>	<i>psbE</i>	15
		<i>psbF</i>	<i>psbH</i>	<i>psbI</i>	<i>psbJ</i>	<i>psbK</i>	
		<i>psbL</i>	<i>psbM</i>	<i>psbN</i>	<i>psbT</i>	<i>psbZ</i>	
	Cytochrome b/f complex	<i>petA</i>	<i>petB*</i>	<i>petD*</i>	<i>petG</i>	<i>petL</i>	6
		<i>petN</i>					
	ATP synthase	<i>atpI</i>	<i>atpH</i>	<i>atpA</i>	<i>atpF*</i>	<i>atpE</i>	6
		<i>atpB</i>					
	Cytochrome c-type synthesis	<i>ccsA</i>					1
	Assembly/stability of photosystem I	<i>ycf3**</i>	<i>ycf4</i>				2
NADPH dehydrogenase	<i>ndhB*,^a</i>	<i>ndhH</i>	<i>ndhA*</i>	<i>ndhI</i>	<i>ndhG</i>	12	
	<i>ndhJ</i>	<i>ndhE</i>	<i>ndhF</i>	<i>ndhC</i>	<i>ndhK</i>		
	<i>ndhD</i>						
Rubisco	<i>rbcL</i>					1	
Transcription and translation related genes RNA genes	Transcription Small subunit of ribosome	<i>rpoA</i>	<i>rpoC2</i>	<i>rpoC1*</i>	<i>rpoB</i>	<i>rps16*</i>	5
		<i>rps7^a</i>	<i>rps15</i>	<i>rps19</i>	<i>rps3</i>	<i>rps8</i>	13
		<i>rps14</i>	<i>rps11</i>	<i>rps12^{a,*}</i>	<i>rps18</i>	<i>rps4</i>	
		<i>rps2</i>					
	Large subunit of ribosome	<i>rpl2^{a,*}</i>	<i>rpl23^a</i>	<i>rpl32</i>	<i>rpl22,</i>	<i>rpl14</i>	11
		<i>rpl33</i>	<i>rpl36</i>	<i>rpl20</i>	<i>rpl16*</i>		
	Translational initiation factor	<i>infA</i>					1
	Ribosomal RNA	<i>rrn16^a,</i>	<i>rrn4.5^a,</i>	<i>rrn5^a,</i>	<i>rrn23^a</i>		8
	Transfer RNA	<i>trnV-GAC^a</i>	<i>trnI-CAU^a</i>	<i>trnA-UGC^{a,*}</i>	<i>trnN-GUU^a</i>	<i>trnP-UGG</i>	37
		<i>trnW-CCA</i>	<i>trnV-UAC*</i>	<i>trnL-UAA*</i>	<i>trnF-GAA</i>	<i>trnR-ACG^a</i>	
		<i>trnT-UGU</i>	<i>trnG-UCC*</i>	<i>trnT-GGU</i>	<i>trnR-UCU</i>	<i>trnE-UUC</i>	
		<i>trnY-GUA</i>	<i>trnD-GUC</i>	<i>trnC-GCA</i>	<i>trnS-GCU</i>	<i>trnH-GUG</i>	
		<i>trnK-UUU*</i>	<i>trnQ-UUG</i>	<i>trnM-CAU</i>	<i>trnG-GCC</i>	<i>trnS-UGA</i>	
		<i>trnS-GGA</i>	<i>trnL-UAG</i>	<i>trnM-CAU</i>	<i>trnL-CAA^a</i>	<i>trnI-GAU*,^a</i>	
	Other genes	RNA processing	<i>matK</i>				
Carbon metabolism		<i>cemA</i>					1
Fatty acid synthesis		<i>accD</i>					1
Proteolysis		<i>clpP**</i>					1
	Component of TIC complex	<i>ycf1</i>					1
	Hypothetical proteins	<i>ycf2^a</i>					2
Total							130

*contain on intron ** contain two introns ^a duplicated gene in the IR

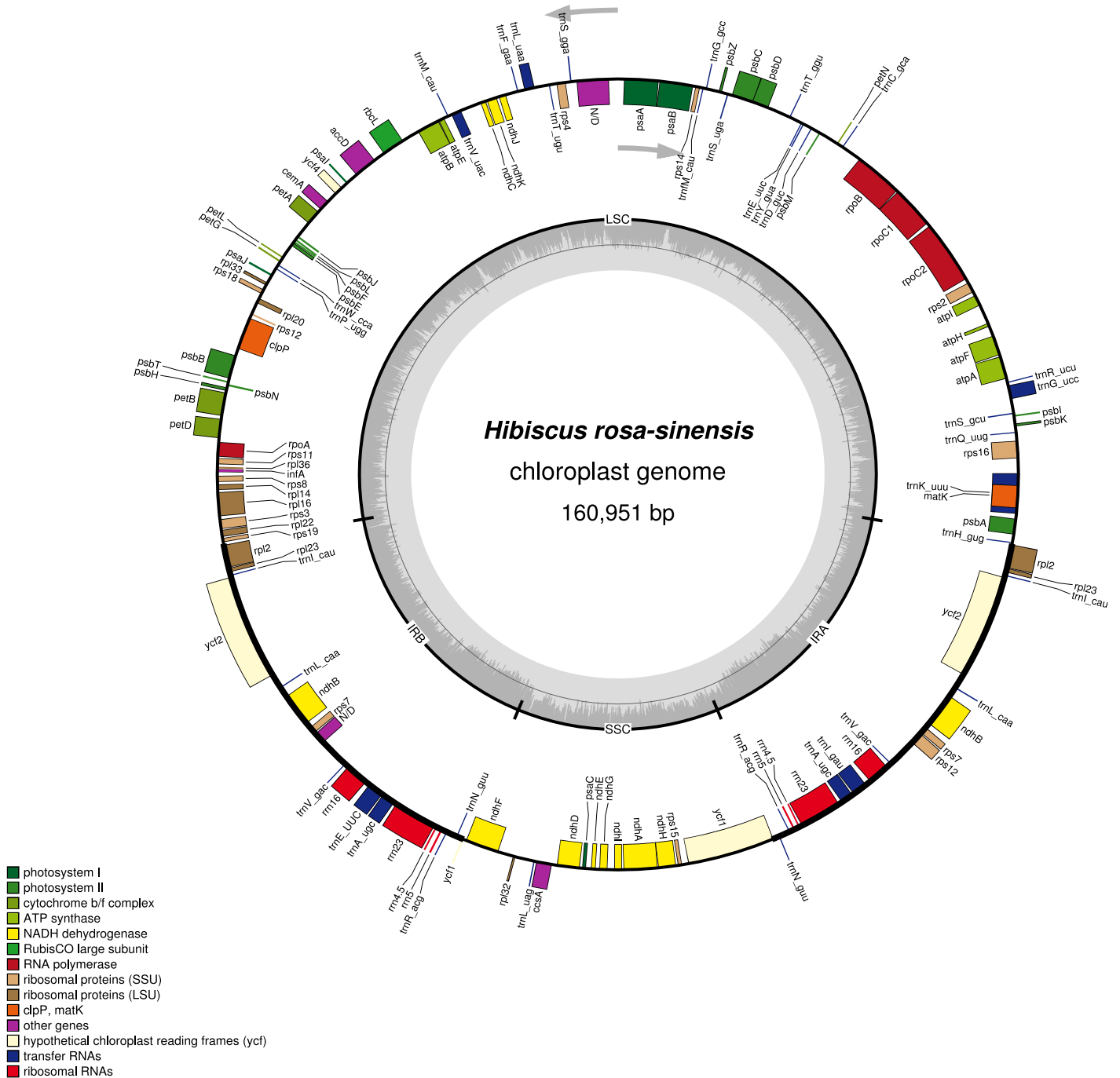


Figure 2.2 Circular map of chloroplast genome of *Hibiscus rosa-sinensis* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

2.3.5 Chloroplast genome features of *Hibiscus mutabilis*

Chloroplast genome size of *H. mutabilis* was 160,879 bp, comprising of two IR regions (IRa and IRb, 26,300 bp each), separated by LSC (89,353 bp) and SSC (18,926 bp) that formed the quadripartite structure. The GC content of chloroplast genome was variable among different regions of chloroplast genome such as LSC (34.7%), SSC (31.5%) and IR (42.6%). The average GC content of complete chloroplast genome was 36.9%.

The chloroplast genome of *H. mutabilis* had 113 genes included 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes except truncated gene of *ycf1*^ψ (Figure 2.3). Among these genes, 17 genes were duplicated in the IR regions except truncated gene of *ycf1*^ψ. The duplicated genes in IR regions included 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contain introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contain 5' part in the LSC region and 3' part in the IR regions, so 3' part was duplicated in the IRs. Out of these 18 genes, 16 genes had one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that contain introns were also duplicated in the IR regions. We found four partially overlapping genes *atpB/atpE* and *psbD/psbC*. The *ycf1* gene started from IR region and ended in SSC region, leaving a truncated copy of 1107 bp in *Hibiscus mutabilis*. The GC content of rRNAs (55.4%) was found higher as compared to tRNAs (53.1%) and protein coding genes (38.1%).

2.3.6 Chloroplast genome features of *Malva parviflora*

Chloroplast genome size of *Malva parviflora* was 158,412 bp, comprising of two IRs (IRa and IRb, 25,107 bp each), separated by LSC (87,086 bp) and SSC (21,112 bp) that formed the quadripartite structure. The GC content of chloroplast genome was varying among different regions of chloroplast genome such as LSC (34.9%), SSC (32.1%), and IR (43%). The GC content of complete chloroplast genome was 37.1%.

The chloroplast genome of *Malva parviflora* had 113 genes included 79 protein-coding genes, 30 tRNA genes and 4 ribosomal RNA genes (Figure 2.4). Among these genes, 17 genes were duplicated in the IR regions including 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contained introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, containing 5' part in the LSC region and 3' part in the IR regions, so 3' part duplicated in the IRs. Out of 18 genes, 16 genes had one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that contained introns were also duplicated in the IR regions. Two

partially overlapping genes *psbD/psbC* were found. The GC content of rRNAs (55.4%) were found higher as compared to tRNAs (52.9%) and protein coding genes (38.1%).

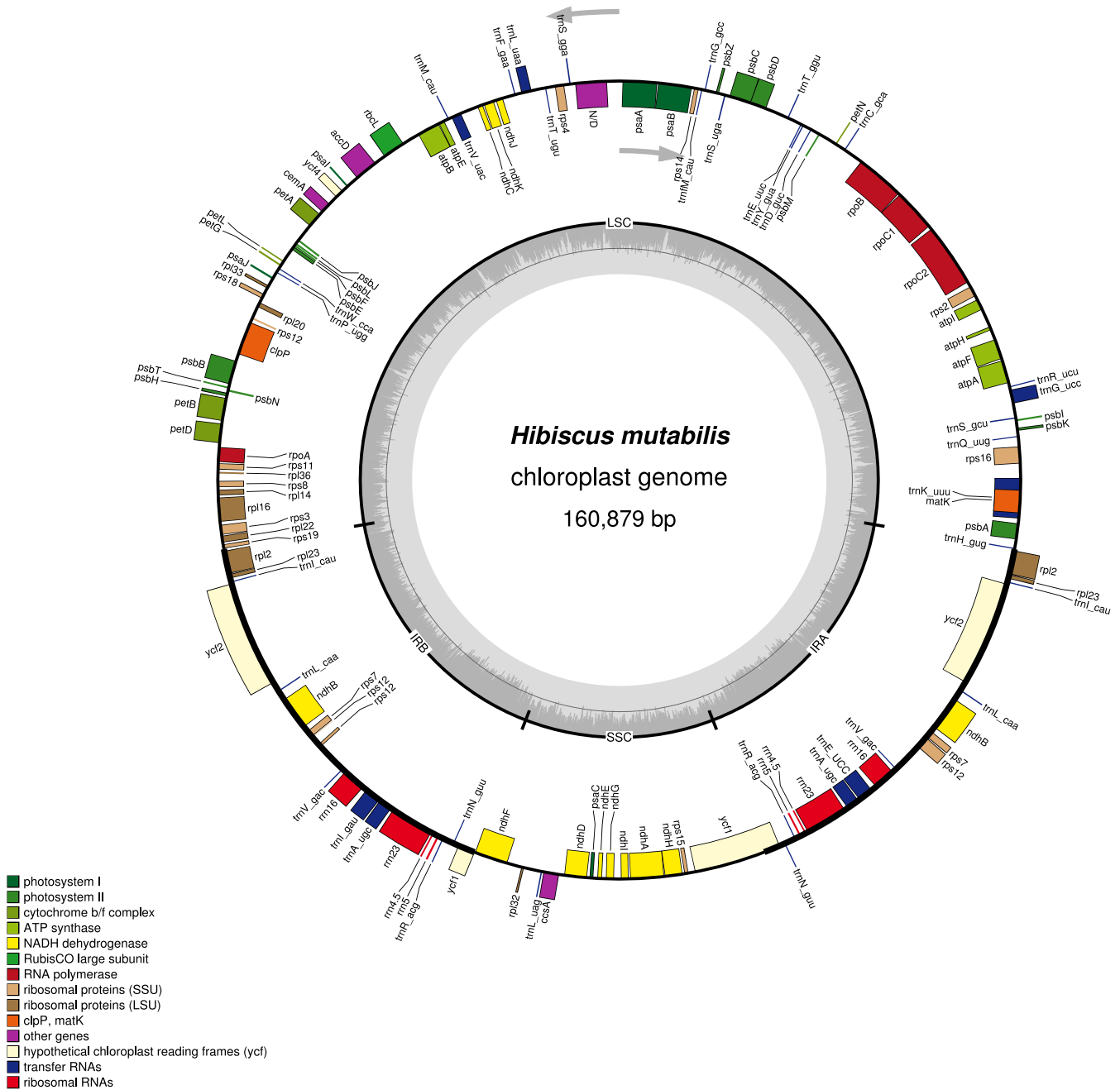


Figure 2.3 Circular map of chloroplast genome of *Hibiscus mutabilis* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

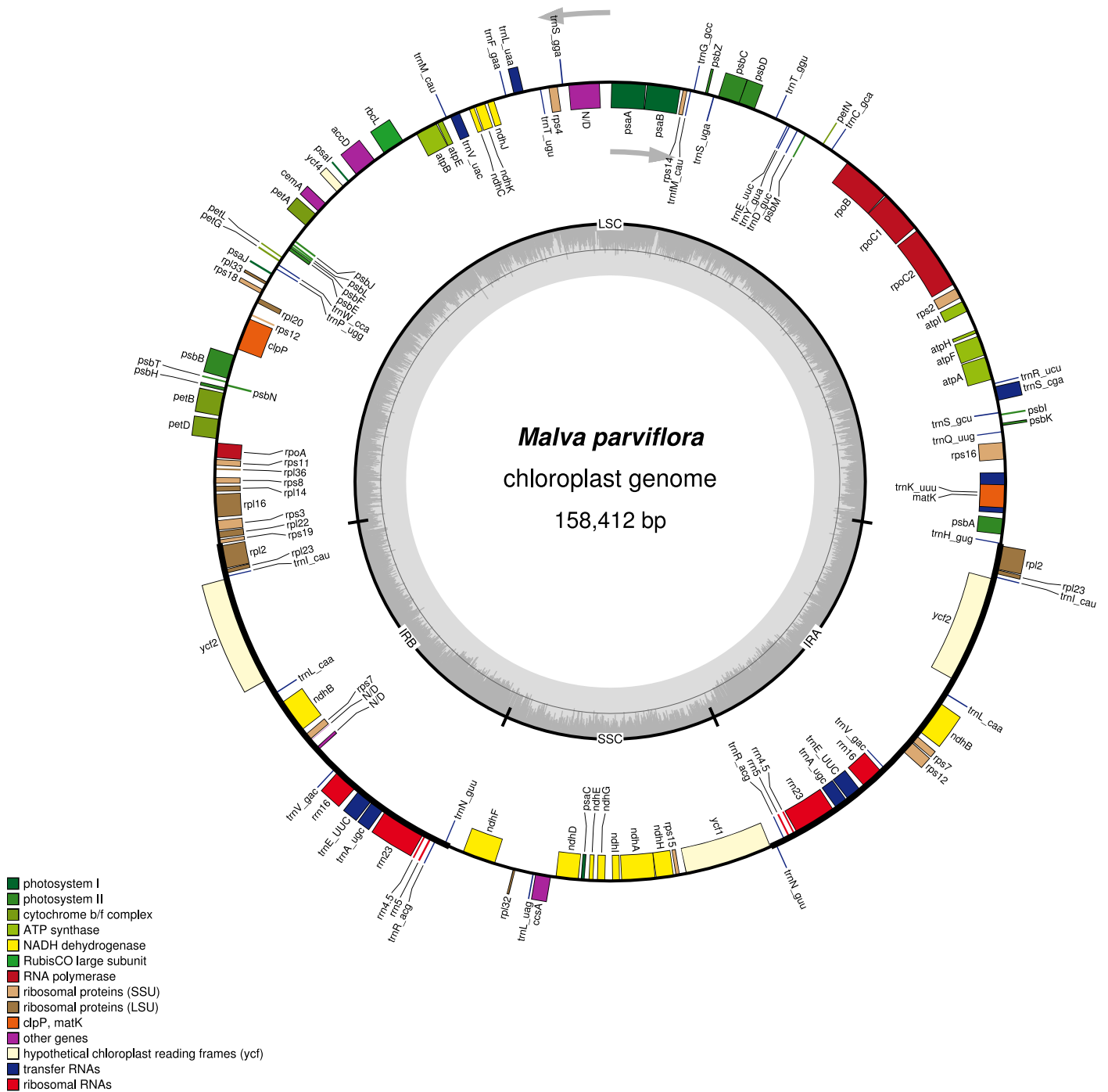


Figure 2.4 Circular map of chloroplast genome of *Malva parviflora* with annotated genes.

The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

2.3.7 Chloroplast genome features of *Malvastrum coromandelianum*

Chloroplast genome of *Malvastrum coromandelianum* was also quadripartite with genome size of 159,872 bp, comprising of two IRs (IRa and IRb, 25,506 bp each), separated by LSC (88,106 bp) and SSC (20,754 bp). The GC content of chloroplast genome was varying among different regions of chloroplast genome such as LSC (34.9%), SSC (32.2%), and IR (42.9%). The GC content of complete chloroplast genome was 37.1%.

The chloroplast genome of *Malvastrum coromandelianum* also consisted of 113 genes included 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes (Figure 2.5). Among these genes, 17 genes were duplicated in the IR regions including 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contained introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contained 5' part in the LSC region and 3' part in the IR regions, with 3' part duplicated in the IRs. Out of 18 genes, 16 genes had one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that contained introns were also duplicated in the IR regions. We found two partially overlapping genes *psbD/psbC*. The GC content of rRNAs (55.4%) were found higher as compared to tRNAs (53%) and protein coding genes (38.2%).

2.3.8 Chloroplast genome features of *Urena Procumbens*

Chloroplast genome size of *U. procumbens* was 161,541 bp, comprising of two IRs (IRa and IRb, 26,668 bp each), separated by LSC (88,991 bp) and SSC (19,214 bp) that formed the quadripartite structure. The GC content of chloroplast genome was varying among different regions of chloroplast genome such as LSC (34.7%), SSC (30.9%), and IR (42.5%). The GC content of complete chloroplast genome was 36.8%.

The chloroplast genome of *Urena procumbens* had 113 genes included 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes except truncated gene of *ycf1^ψ* (Figure 2.6). Among these genes, 17 genes were duplicated in the IR regions except truncated gene of *ycf1^ψ*. The duplicated genes in IRs regions including 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contain introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contain 5' part in the LSC region and 3' part in the IR regions, so 3' part duplicated in the IRs. Out of 18 genes, 16 genes have one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that contain introns were also duplicated in the IR regions. We found two partially overlapping genes *psbD/psbC*. The *ycf1* gene started from IR region and ended in SSC region, leaving a truncated copy of 1134 bp in *U. procumbens*. The

GC content of rRNAs (55.4%) was found higher as compared to tRNAs (53.1%) and protein coding genes (38%).

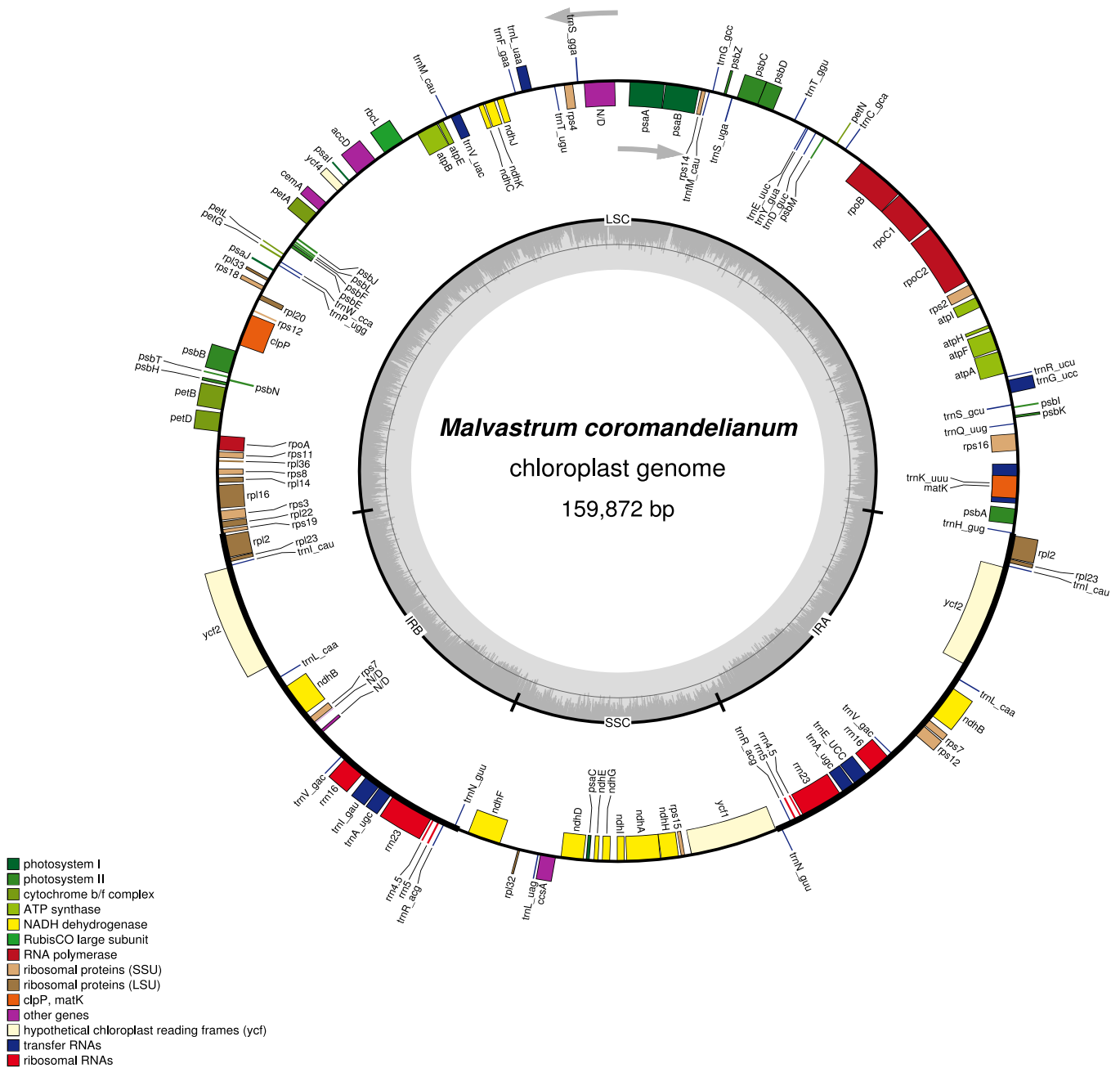


Figure 2.5 Circular map of chloroplast genome of *Malvastrum coromandelianum* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

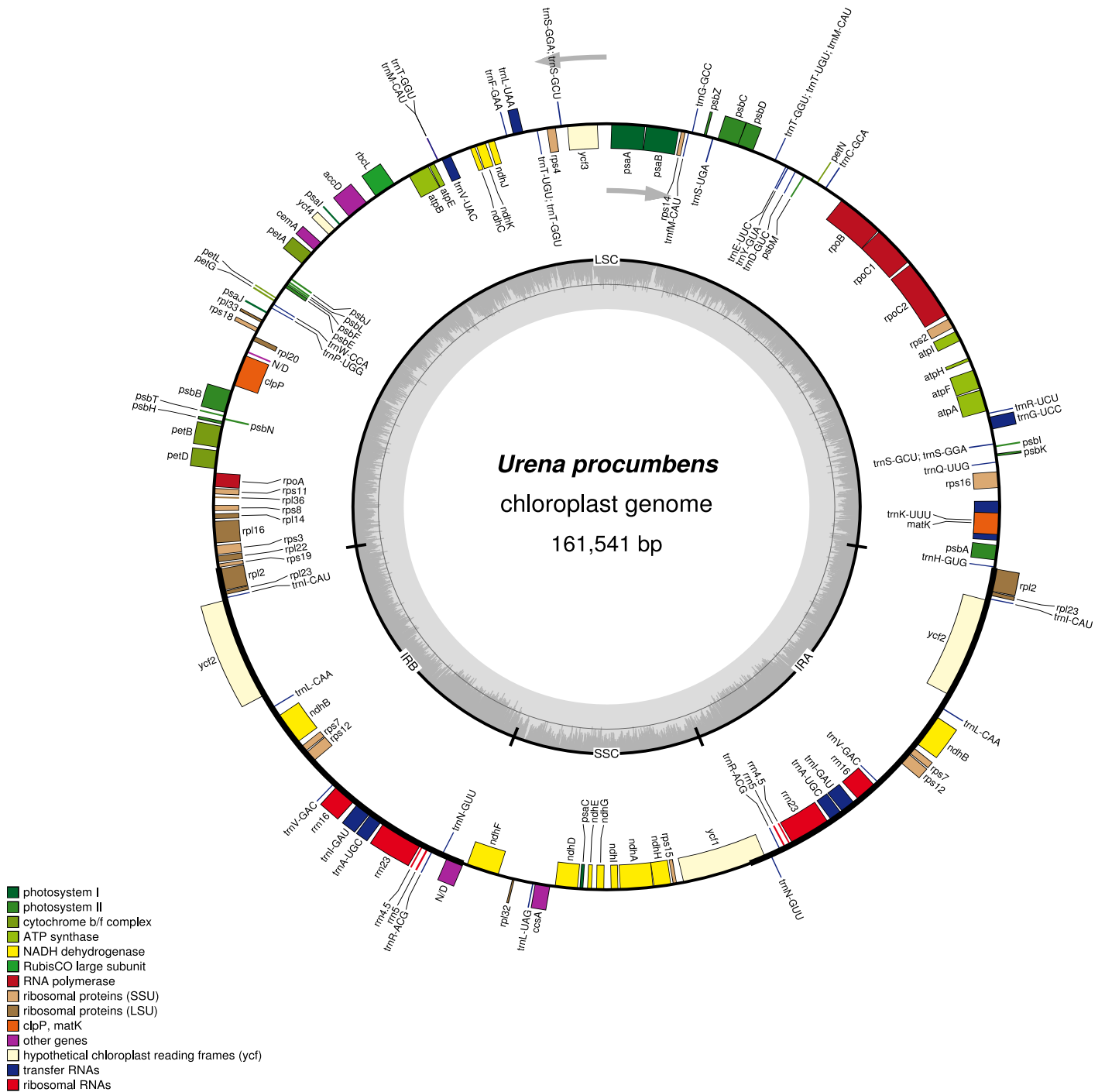


Figure 2.6 Circular map of chloroplast genome of *Urena procumbens* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

2.3.9 Chloroplast genome features of *Firmiana colorata*

Chloroplast genome size of *Firmiana colorata* was 160,700 bp, comprising of two IRs (IRa and IRb, 25,574 bp each), separated by LSC (89,551 bp) and SSC (20,001 bp) forming the quadripartite structure. The GC content of chloroplast genome was different among different regions of chloroplast genome such as LSC (35%), SSC (31.4%) and IR (42.9%). The GC content of complete chloroplast genome was 37.1%.

The chloroplast genome of *Firmiana colorata* had 113 genes including 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes except truncated gene of *ycf1^ψ* (Figure 2.7). Among these genes, 17 genes were duplicated in the IR regions except truncated gene of *ycf1^ψ*. The duplicated genes in IRs regions included 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contain introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contained 5' part in the LSC region and 3' part in the IR regions, so 3' part duplicated in the IRs. Out of 18 genes, 16 genes had one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that had introns were also duplicated in the IR regions. We found six partially overlapping genes *rps3/rpl22*, *atpB/atpE* and *psbD/psbC*. The *ycf1* gene started from IR region and ended in SSC region, leaving a truncated copy of 33 bp of this gene in *Firmiana colorata*. The GC content of rRNAs (55.5%) was found higher as compared to tRNAs (53.2%) and protein coding genes (38.2%).

2.3.10 Chloroplast genome features of *Sterculia monosperma*

Chloroplast genome size of *Sterculia monosperma* was 161,099 bp, comprising of two IRs (IRa and IRb, 25,546 bp each), separated by LSC (89,562 bp) and SSC (20,445 bp) that formed the quadripartite structure. The GC content of chloroplast genome differed among different regions of chloroplast genome such as LSC (34.8%), SSC (31.4%) and IR (42.9%). The GC content of complete chloroplast genome was 36.9%.

The chloroplast genome of *Sterculia monosperma* had 113 genes consisting of 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes (Figure 2.8). Among these genes, 17 genes were duplicated in the IR regions. These duplicated genes in IRs regions included 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contained introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contain 5' part in the LSC region and 3' part in the IR regions, so 3' part duplicated in the IRs. Out of 18 genes, 16 genes consisted of one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The 3 protein-coding genes and 2 tRNA genes that

contained introns were also duplicated in the IR regions. We found six partially overlapping genes *rps3/rpl22*, *atpB/atpE* and *psbD/psbC* for of *Sterculia monosperma*. The GC content of rRNAs (55.5%) was found higher than tRNAs (53.2) and protein coding genes (38.1%).

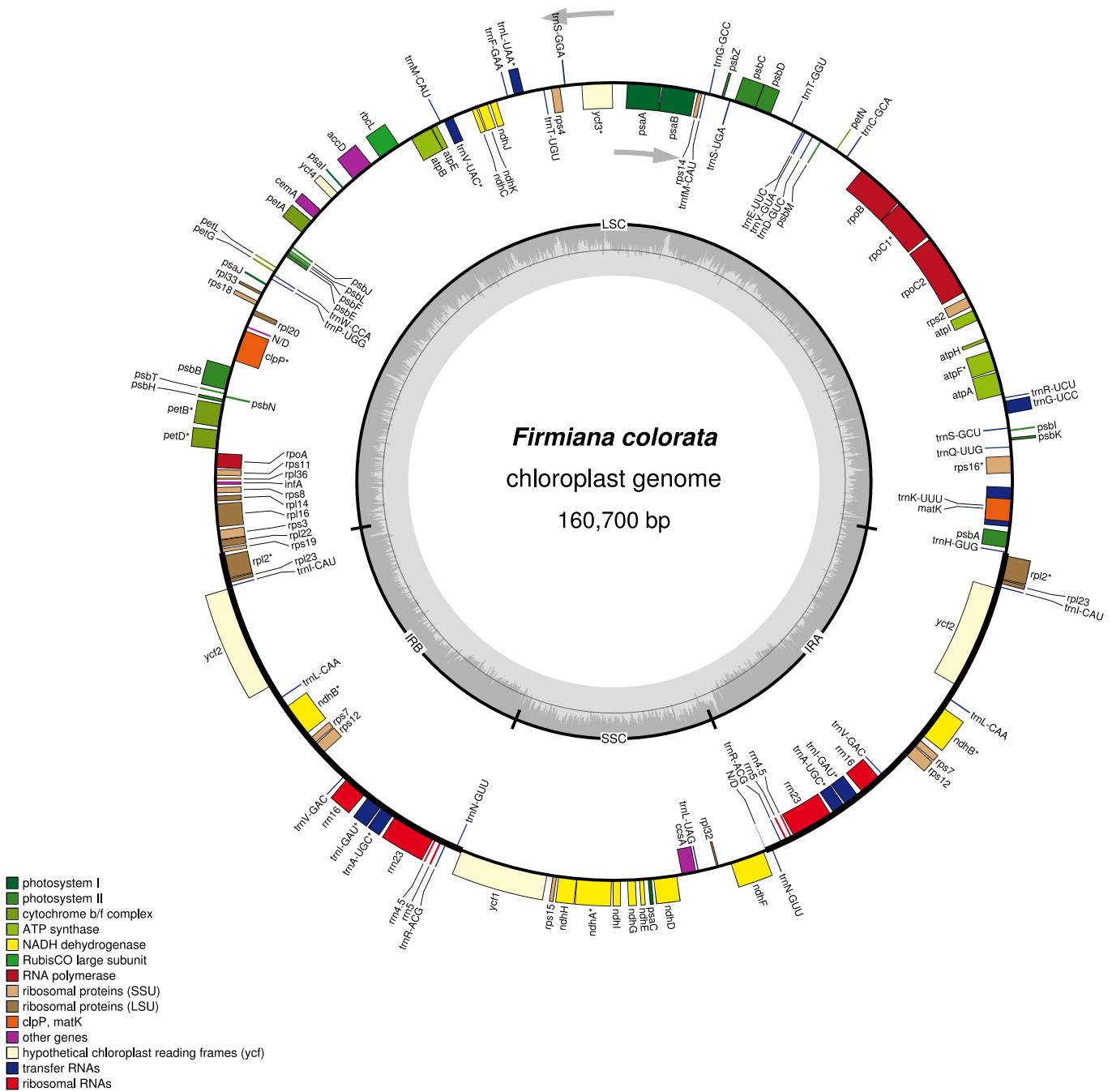


Figure 2.7 Circular map of chloroplast genome of *Firmiana colorata* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

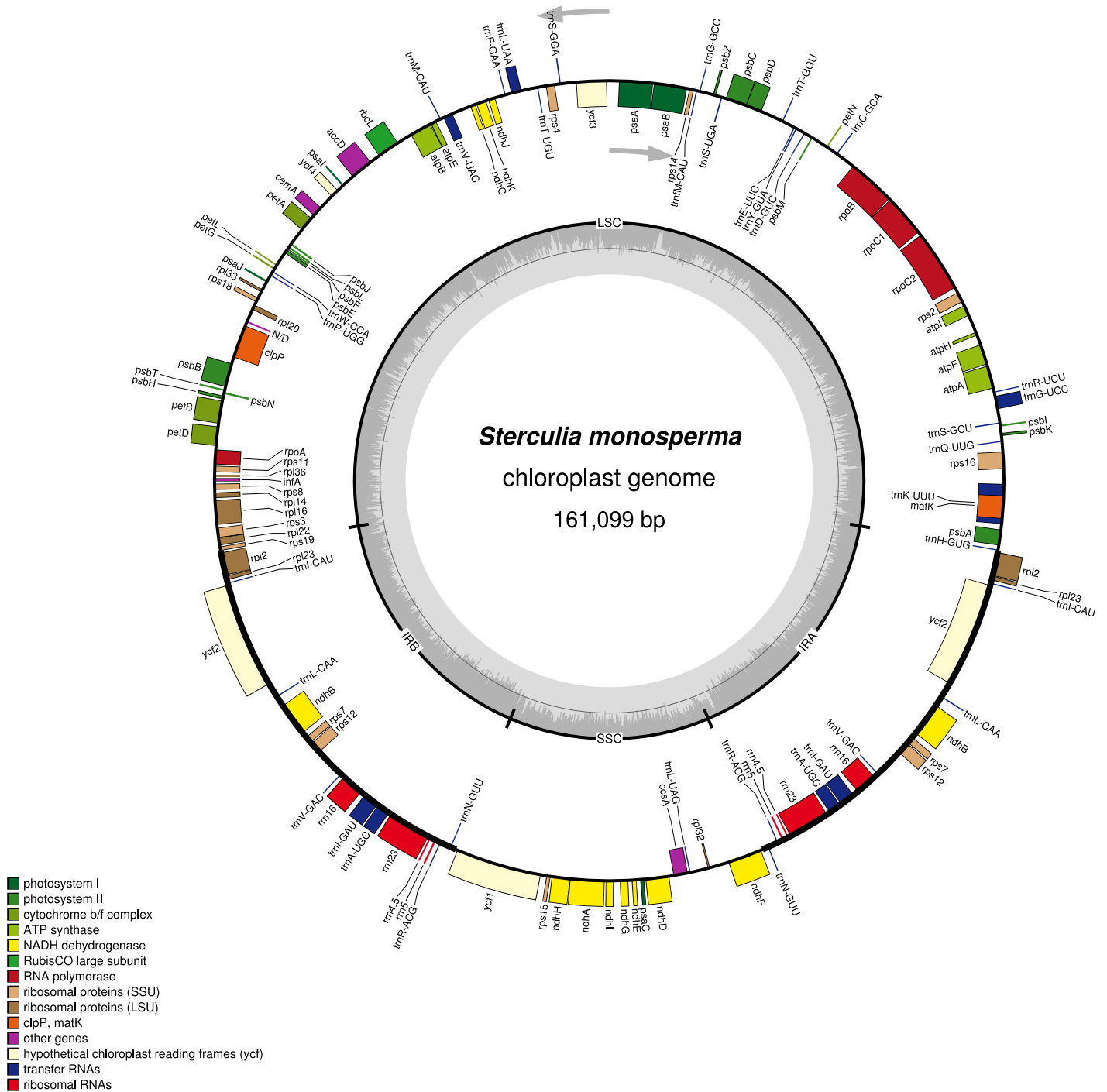


Figure 2.8 Circular map of chloroplast genome of *Sterculia monosperma* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

2.3.11 Chloroplast genome features of *Pterospermum truncatolobatum*

Chloroplast genome size of *Pterospermum truncatolobatum* was 162,320 bp, comprising of two IRs (IRa and IRb, 25,499 bp each), separated by LSC (91,506 bp) and SSC (19,816 bp) that formed the quadripartite structure. The GC content of chloroplast genome was varying among different regions of chloroplast genome such as LSC (33.9%), SSC (31.2%), and IR regions (42.9%). The GC content of complete chloroplast genome was 36.4%.

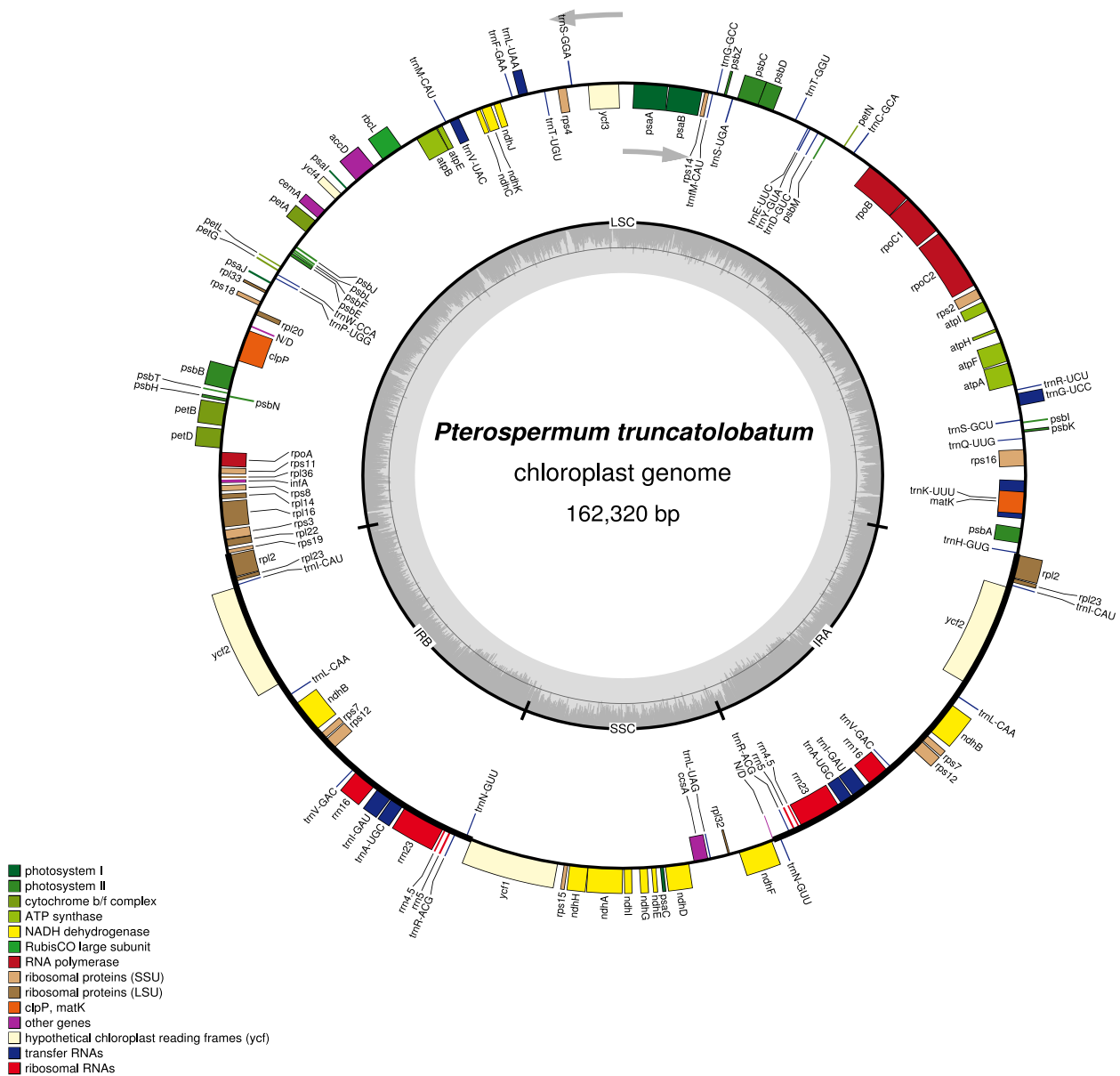


Figure 2.9 Circular map of chloroplast genome of *Pterospermum truncatolobatum* with annotated genes. The genes transcribed counter clockwise are shown inside of the circle, whereas the genes transcribed clockwise are shown outside of the circle. The borders of chloroplast genome are defined with LSC (large single copy), SSC (small single copy), IRa and IRb (inverted repeats). Every gene is coded according to its specific function. The dashed grey colour of inner circle shows the GC content, whereas the lighter grey colour shows AT content.

The chloroplast genome of *Pterospermum truncatolobatum* had 113 genes including 79 protein-coding genes, 30 tRNA genes, and 4 ribosomal RNA genes except truncated gene of *ycf1^ψ* (Figure 2.9). Among these genes, 17 genes were duplicated in the IR regions. The duplicated genes in IRs regions included 4 rRNA genes, 7 tRNA genes, and 6 protein coding genes. The 18 genes contained introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, containing 5' part in the LSC region and 3' part in the IR regions, thus 3' part was duplicated in the IRs. Out of 18 genes, 16 genes had one intron while two genes (*clpP* and *ycf3*) had two introns (Table 2.2). The three protein-coding genes and two tRNA genes that contained introns were also duplicated in the IR regions. We found six partially overlapping genes *rps3/rpl22*, *atpB/atpE* and *psbD/psbC*. The *ycf1* gene started from IR region and ended in SSC region, leaving a truncated copy of 60 bp in *Firmiana colorata*. The GC content of rRNAs (55.5%) was found higher than tRNAs (53.1%) and protein coding genes (38.1%).

2.4 Conclusion

We sequenced and/or *de novo* assembled chloroplast genomes of eight Malvaceae species. These genomic resources will broaden knowledge about the chloroplast genome structure of many genera of family Malvaceae and will be helpful for elucidation of evolutionary dynamics and phylogeny in family Malvaceae. Moreover, these resources will broaden information about the chloroplast genome structure of angiosperms.

Chapter 3

Comparative analyses of chloroplast genomes of family Malvaceae and rates of evolution in protein coding genes

3.1 Introduction

Chloroplast genome is double membrane organelle (Palmer, 1985), having genome size ranging from 107 kb to 218 kb with about 120 genes (Daniell *et al.*, 2016). The contraction and expansion of inverted repeat regions are considered as important phenomena, which lead to duplication of complete genes or generation of pseudogenes. Moreover, these also lead to variable number of genes within similar lineages (Menezes *et al.*, 2018). The intergenic spacer regions are the most polymorphic regions within chloroplast genome (Chen *et al.*, 2018; Choi *et al.*, 2016; Menezes *et al.*, 2018; Zhang *et al.*, 2016) and variations in length of intergenic spacer regions cause variations in the average length of complete chloroplast genome, LSC, SSC, and IRs within same lineages (Choi *et al.*, 2016; Li *et al.*, 2018; Sinn *et al.*, 2018; Zhang *et al.*, 2016). Many mutational events occur within chloroplast genome such as translocations, structural rearrangements, inversions, insertions and deletions (InDels) and copy number variations (CNVs) (Jheng *et al.*, 2012; Xu *et al.*, 2015). The studies of these mutational events are important to understand the evolution of chloroplast genome (Kim and Kim, 2014; Palmer, 1985; Sherman-Broyles *et al.*, 2014; Sinn *et al.*, 2018). Similar to the genes, the intron within the genes are also conserved in chloroplast genome (Daniell *et al.*, 2016). However, loss of introns were also observed in many species of angiosperms (monocot and eudicot) and gymnosperms (Downie *et al.*, 1991; Gu *et al.*, 2016; He *et al.*, 2017; Jansen *et al.*, 2007; Lei *et al.*, 2016).

Plant cells contain three distinct genomes including nuclear, mitochondrial, and plastid. The inter transfer of genes or part of genes has been reported in these genomes (Timmis *et al.*, 2004). Despite conserved genes content, some protein-coding genes of chloroplast genomes have been found missing in some species (Barrett *et al.*, 2014; Jansen *et al.*, 2007; Kim *et al.*, 2018) and further studies revealed transfer of those genes to nuclear and/or mitochondrial genomes. The inter-transfer of genes or part of genes in these genomes provides valuable information for the studies of evolution and phylogeny, and insights to the transferring events can be gained by analyses of complete chloroplast genome sequences (Daniell *et al.*, 2016). For instance, several genes of chloroplast genome have been found absent in chloroplast genomes of few species, and their transferring was confirmed to other genomes i.e. the *infA* (Daniell *et al.*, 2016; Millen *et al.*, 2001), *rpl22*, *rpl32*, (Cusack and Wolfe, 2007; Gantt *et al.*, 1991; Jansen *et al.*, 2011; Park *et al.*, 2015; Ueda *et al.*, 2007) and *ndhH* genes (Braukmann *et al.*, 2009; Chris Blazier *et al.*, 2011; McCoy *et al.*, 2008; Pan *et al.*, 2012; Sanderson *et al.*, 2015; Shahinnia and Sayed-Tabatabaei, 2009; Wakasugi *et al.*, 1994; Weng *et al.*, 2014; Wu *et al.*, 2010; Yang *et al.*, 2013). The deletion of many other genes from chloroplast genome and

its transferring to nuclear or mitochondrial genome was also reported in some plant species including *accD*, *psaI*, *rpl20*, *rpl23*, *rpl33*, *rps16*, *rpoA*, *ycf1*, *ycf2*, and *ycf4* (Daniell *et al.*, 2016).

The analyses of chloroplast genome sequence is important for understanding the adaptive evolution and domestication of plant species (Daniell *et al.*, 2016; Menezes *et al.*, 2018). The chloroplast genome structure of many species shed light on the domestication of the species with the interesting genome rearrangements and inversions. Furthermore, the analyses of the synonymous and non-synonymous substitutions of protein coding genes revealed the selection pressures on these genes along with their roles in adaptation to certain environmental conditions. The large scale genome rearrangements and inversions related to regions of essential genes of chloroplast genomes without any adverse effect in many species revealed their role in adaption and domestications (Cai *et al.*, 2008; Guo *et al.*, 2007; Kazakoff *et al.*, 2012; Magee *et al.*, 2010; Martin *et al.*, 2014; Sabir *et al.*, 2014; Saski *et al.*, 2005; Sherman-Broyles *et al.*, 2014; Tangphatsornruang *et al.*, 2010).

Instead of genome rearrangement and inversions, high rate of mutational events, substitutions and InDels, in protein coding genes of many species, were observed as compared to average rate of mutations (Choi *et al.*, 2016; Henriquez *et al.*, 2014; Jansen *et al.*, 2007; Menezes *et al.*, 2018; Saina *et al.*, 2018a). For instance, the positive selection of *matK* in the 30 species of *Citrus* and the other related plant groups indicate its role in the stresses that is faced by plants in different ecological niches (Chen and Xiao, 2010). The *ndhF* gene encodes a subunit of the chloroplast NAD(P)H dehydrogenase (NDH) complex. The NDH monomers of chloroplast genome are sensitive to high light stress. Hence, this observation also suggests the role of *ndhF* gene in the acclimation of stress (Peng *et al.*, 2011). Some studies also relate the *ycf1* gene with the stress. The studies in Australia also indicated the link of positive selection pressure of *matK*, *ndhF*, and *ycf1* to the adaptation of these species to a hot and dry climate of Australia (Carbonell-Caballero *et al.*, 2015; Caspermeyer, 2015).

In the current study, we aimed to compare the chloroplast genomes structure of family Malvaceae, the gene content, genes arrangements, and the effect of IR contraction and expansion. Furthermore, to get insight into the adaptation of Malvaceae, we determined evolutionary rate of 77 protein coding genes. Our study revealed high similarities in the chloroplast genomes of family Malvaceae in context of chloroplast genome structures, gene arrangements, gene content and intron content. The analyses of evolutionary rate of 77 protein coding genes also showed about 95% similarities.

3.2 Materials and Methods

3.2.1 Genomics features of chloroplast genome in family Malvaceae

In the current study, we assembled chloroplast genomes of 8 species of family Malvaceae as given in chapter 2. Along with these species, we further downloaded chloroplast genome sequences of 12 species from NCBI (Table 3.1). The features of chloroplast genomes of 20 species of family Malvaceae were determined and compared by visualisation in Geneious R8.1 (Kearse *et al.*, 2012). We took one species from 16 genera of Malvaceae whereas two species were included for genera of *Hibiscus* and *Firmiana* due to inclusion of our assembled genomes. Here, we compared genome structure, number of genes, intron containing genes, and GC content of these genomes. The IR contraction and expansion were also determined by visualisation of boundary regions of LSC/IRb, IRb/SSC, SSC/IRa and IRa/LSC, and the variations that lead to number of genes in different species were determined.

3.2.2 Rate of evolution of protein coding genes

We determined the evolutionary rate of the 77 protein coding genes among 19 species of family Malvaceae. For this purpose, we calculated the rate of non-synonymous substitution (Ka), synonymous substitutions (Ks) and their ratio (Ka/Ks). *Theobroma cacao*, a species that is basal to family Malvaceae, was used as a reference and protein-coding genes of all the species were aligned with *T. cacao* by Geneious pairwise alignment and analysed in DnaSP v.5.10 for Ka and Ks without stop codon (Rozas *et al.*, 2017).

3.3 Results

3.3.1 Genome structure and gene content

All the chloroplast genomes of family Malvaceae showed similarities in genomes structure, GC content, and gene content and orders. The chloroplast genome of the Malvaceae species contained 113 unique genes including 79 protein-coding genes, 30 tRNAs and 4 rRNAs (as describe in table 2.2) except *Gossypium herbaceum* which lacked *infA* gene. Seventeen genes were also duplicated in the IR regions. Hence, the total genes were 130 including 85 protein coding genes, 30 tRNAs, and 8 rRNAs genes in all species except *Durio zibethinus*, *Abelmoschus esculentus* and *Gossypium herbaceum*. The *Durio zibethinus* contained 83 protein coding genes due to contraction of IR regions due to which single copy of *rpl2* and *rpl23* is present in the genome instead of duplicate copy. In *Abelmoschus esculentus*, the expansion of IR regions leads to duplication of *rps19*, *rpl22*, and *rps3*. Therefore, except these two species, other species contained 17 duplicated genes in IR regions. Among 17 genes, 4 were ribosomal RNAs (rRNAs) genes, 7 transfer RNAs (tRNAs) and 6 were protein coding

genes. We also noted about half of the species including in our study had non-functional copy of *infA* gene (Table 3.1). We noted 18 intron containing genes included 12 protein coding genes and 6 tRNAs. The protein coding genes that contained introns included *petD*, *petB*, *atpF*, *ycf3*, *ndhB*, *ndhA*, *rpoC1*, *rps16*, *rps12*, *clpP*, *rpl2*, and *rpl16*. Among these genes, two genes *clpP* and *ycf3* contained 2 introns whereas all other genes contained 1 intron. Three intron containing protein coding genes also duplicated in the IR regions which included *rps12*, *rpl2*, and *ndhB*. The *rps12* gene was a trans-spliced gene, containing 5' part in the LSC region and 3' part in the IR regions. Hence, 3' part was duplicated in the IR.

3.3.2 Length of the chloroplast genome and GC content

The length of the complete chloroplast genomes also varied among the species and ranged from 158,412 (*Malva parviflora*) to 163,974 (*Durio zibethinus*). The length of the LSC showed high variation among the species of family Malvaceae and ranged from 87,086 bp (*Malva parviflora*) to 95,704 bp (*Durio zibethinus*). The variation in the length of SSC regions was inconsistent and interestingly, the length of SSC was found 18,926 bp in *Hibiscus mutabilis*, whereas the length of the SSC regions of *Malva parviflora* (the genome with smallest size) was highest among all the species with 21,112 bp. The length of IR regions was also showed interesting results. We noticed that the genome with the largest size, *Durio zibethinus*, showed smallest length of the IR regions with 23,726 bp, whereas the genome of *Abelmoschus esculentus* had highest length of the IR regions with 28,009 bp.

The GC content of the chloroplast genomes showed highest similarities among the 20 species of family Malvaceae (Table 3.1). The GC content of complete chloroplast genome ranged from 35.8% to 37.2%, LSC was 33.6% to 35.2%, SSC was 30.8% to 32.2%, and IR was 42% to 43%. The high GC content of IR regions belong to presence of rRNAs and tRNAs that having GC content of about 55% and 53%, respectively.

Table 3.1 Comparative analyses of the general features of 20 species of family Malvaceae

Species	Protein coding genes	SSC (bp)	GC content (%)	IR (bp)	GC content (%)	LSC (bp)	GC content (%)	Total Length (bp)	GC content (%)	Accession No.
<i>Hibiscus mutabilis</i>	85*	18,926	31.5	26,300	42.6	89,353	34.7	160,879	36.9	MK820657
<i>Hibiscus rosa-sinensis</i>	85*	20,246	31.3	25,598	42.9	89,509	34.9	160,951	37	MK382984
<i>Hibiscus syriacus</i>	85*	19,831	31.1	25,745	42.8	89,701	34.7	161,022	36.8	KR259989
<i>Talipariti hamabo</i>	85*	19,570	30.9	26,471	42.7	89,217	34.8	161,729	36.9	NC_030195
<i>Gossypium herbaceum</i>	84 ^a	20,385	31.7	25,570	42.9	88,779	35.2	160,304	37.2	HQ325742
<i>Althaea officinalis</i>	85*	21,057	32.2	25,445	43.0	88,040	34.7	159,987	37.0	NC_034701
<i>Abelmoschus esculentus</i>	88 ^c	19,032	31.5	28,009	42.0	88,071	34.5	163,121	36.7	NC_035234
<i>Malva parviflora</i>	85*	21,112	32.1	25,107	43	87,086	34.9	158,412	37.1	MK860036
<i>Malvastrum coromandelianum</i>	85*	20,754	32.2	25,506	42.9	88,106	34.9	159,872	37.1	MK860037
<i>Urena procumbens</i>	85*	19,214	30.9	26,668	42.5	88,991	34.7	161,541	36.8	BK010727
<i>Reevesia thyrsoidea</i>	85*	20,289	31.5	25,466	43	90,565	34.6	161,786	36.8	MH939148
<i>Tilia amurensis</i>	85	20,397	31.0	25,597	42.9	91,124	34.1	162,715	36.5	KT894772
<i>Theobroma cacao</i>	85	20,187	31.2	25,511	43.0	89,395	34.7	160,604	36.9	HQ336404
<i>Firmiana colorata</i>	85	20,001	31.4	25,574	42.9	89,551	35	160,700	37.1	BK010724
<i>Firmiana major</i>	85	20,038	31.3	25,543	42.9	90,178	34.7	161,302	36.9	NC_037242
<i>Heritiera parvifolia</i>	85	19,964	31.5	25,588	43.0	89,053	34.9	160,193	37.1	NC_038057
<i>Sterculia monosperma</i>	85	20,445	31.4	25,546	42.9	89,562	34.8	161,099	36.9	BK010726
<i>Durio zibethinus</i>	83 ^b	20,819	30.8	23,726	42.5	95,704	33.6	163,974	35.8	MG138151
<i>Bombax ceiba</i>	85	21,111	32.0	24,432	42.8	89,022	34.7	158,997	36.8	MG569974
<i>Pterospermum truncatolobatum</i>	85	19,816	31.2	25,499	42.9	91,506	33.9	162,320	36.4	BK010725

* *infA* gene is non-functional ^a *infA* gene is absent ^b contraction of IR regions leads to single copy of *rpl2* and *rpl23* ^c expansion of IR regions leads to duplication of *rps3*, *rpl22*, and *rps19*. 8 ribosomal RNAs and 37 tRNAs were present in chloroplast genome of each species, thus not mentioned.

3.3.3 IR contraction and expansion

We compared boundary regions of chloroplast genomes among the mentioned 20 species of Malvaceae (Figure 3.1). The gene *ycf1* was in junction of SSC/IR in 12 species due to which a pseudogene of *ycf1*^ψ was generated at 5' end at IRa/IRb junctions. The size of *ycf1*^ψ (pseudogene) ranged from 33 bp (in *Firmiana colorata*) to 1134 bp (in *Urena procumbens*). Both species of genus *Hibiscus* possessed large difference in size of *ycf1*^ψ. In *H. rosa-sinensis* *ycf1*^ψ gene was 123 bp whereas in *H. syriacus* the *ycf1*^ψ was 621 bp. The size of *ycf1*^ψ showed similarities. In *F. major* *ycf1*^ψ was 66 bp and in *F. colorata* the *ycf1*^ψ was 66 bp. The eight species in which *ycf1* gene was located within IR only, showed that *ycf1* gene was located as closed to boundary as a single base pair difference was not present (in *Malva parviflora*) to as far as 1225 bp (in *Bombax ceiba*). The *ndhF* gene that was present at the border of IR/SSC crossed SSC and entered IR junction in *Bombax ceiba* with 99 bp and in *Theobroma cacao* with 6 bp. In all other species, the *ndhF* gene was completely present in SSC and away from the junction: 6 bp in *Reevesia thyrsoidea* to 908 bp in *Althaea officinalis*. In both species of *Hibiscus*, *ndhF* gene was found 150 bp away from the junction whereas in *Firmiana* species differences existed. In *Firmiana major* the distance of *ndhF* was 139 bp from the border whereas in *Firmiana colorata* the distance was 177 bp. The junction of LSC/IRb showed variation to some extent. The *rps19* gene was mostly present in LSC integrated into LSC/IR junction in 14 species from 2-14 bp, while in three species found away from the junction: 2 bp in *Gossypium herbaceum* and 101 bp in *H. rosa-sinensis*.

Among these twenty species, two species showed other than *rps19* gene near the border of IR. The *rpl16* gene was present in *Abelmoschus esculentus* that was also integrated into IRb with 66 bp and the *rpl23* present in *Durio zibethinus* was far away from the junction with 64 bp. The IR expansion in *Abelmoschus esculentus* lead to complete duplication of *rps19*, *rpl22*, *rps3*, and a pseudogene of *rpl16* at 5' end in the IR region. On the other side, the contraction of IR and expansion of the LSC region lead to existence of single copy of *rpl2* and *rpl23* in *Durio zibenthinus*, whereas these genes were duplicated in all other plants. In species of genus *Hibiscus*, we noted that *rps19* gene was 101 bp away from the junction in *H. rosa-sinensis* whereas in *H. syriacus* the *rps19* integrated into IRb with 3 bp. In both species of *Firmiana*, the *rps19* was integrated into IRb with 6 bp. At the junction of IRa/LSC, *trnH-GUG* was present completely in LSC.

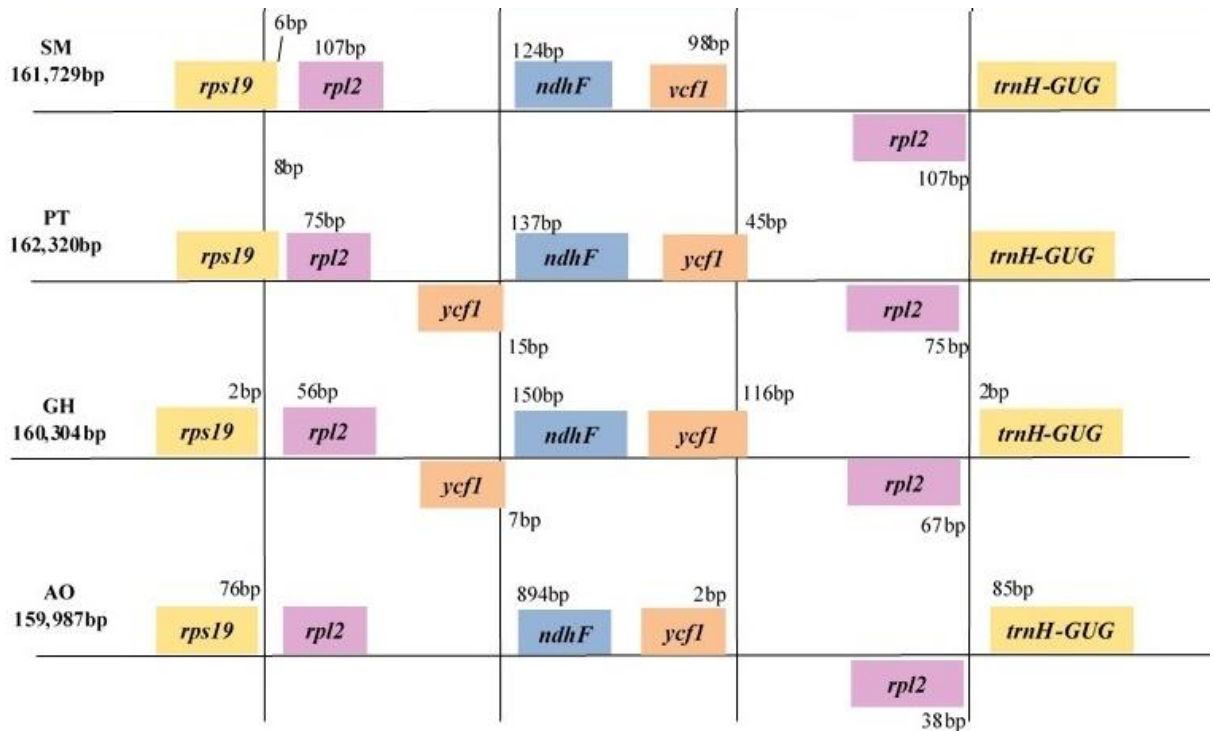


Figure 3.1 Comparative analyses of boundary regions: inverted repeat regions (IR), small single copy (SSC), and large single copy (LSC) among 20 species of Malvaceae. HR: *H. rosa-sinensis*, HS: *H. syriacus*, HM: *H. mutabilis*, AE: *A. esculentus*; TH: *T. hamabo*, BC: *B. ceiba*; DZ: *D. zibethinus*, FM: *F. major*, FC: *F. colorata*, HP: *H. parvifolia*; TC: *T. cacao*; TA: *T. amurensis*; MP: *M. parviflora*, MC: *M. coromandelianum*, UP: *U. procumbens*, RT: *R. thyrsoidea*, SM: *S. monosperma*, PT: *P. truncatolobatum*, GH: *G. herbaceum*, and AO: *A. officinalis*. Species of *Hibiscus* were presented at top of figure to show variations in both species. *Hibiscus* species revealed clear differences based on position of the genes from borders and in size of the *ycf1* pseudogene. All genes with reverse direction were given at the top of lines whereas genes with positive direction were given below line. *ycf1* present at the border of IRb and SSC represent truncated copy. The numbers of the bp indicate the distance from the boundary regions or part of the gene that cross the borderline among two locations of chloroplast. To get clear understanding of all species, we reversed direction of small single copy of some species.

3.3.4 Comparative analyses of evolutionary rate of protein coding genes

In the current study, we analysed non-synonymous and synonymous substitution rate of 77 protein-coding among the 19 species of family Malvaceae. This study revealed that evolution rate varied among the 19 species of family Malvaceae (Table 3.2). The Ka/Ks value was found less than 0.5 in about 95% of genes. The genes related to photosynthesis showed the lowest evolutionary rate whereby maximum genes showed only synonymous substitutions. The average values were $K_s = 0.048$, $K_a = 0.0023$, and $K_a/K_s = 0.047$. Eight genes (*rps16*, *rpoC2*, *rbcL*, *cemA*, *rps11*, and *rpl14*, *psbK*, and *psbJ*) had Ka/Ks rate higher than 0.5 and lower than 1 in one species, three genes (*rps14*, *rps4*, and *rpl20*) in two species, four genes (*rps2*, *rps18*, *ndhC*, and *ndhG*) in three species, two genes (*clpP* and *psbT*) in four species, three genes (*accD*, *ycf2*, and *ycf1*) in six species, one gene (*rpl22*) in seven species, and one gene (*matK*) in eight species. Five genes (*psbI*, *psbT*, *rpl23*, *matK* and *rps7*) showed Ka/Ks value greater than 1 in at least one species, *rpl22* in two species, and *clpP* was present at this higher rate in eight species. The evolutionary rate was similar in the 40 (52%) genes among two species of *Hibiscus* whereas 37 genes (about 48%) had the difference in Ka/Ks more than 5%. The 18 genes (*atpA*, *rpoC1*, *rpoB*, *psaA*, *rps4*, *ndhK*, *ndhC*, *atpB*, *accD*, *ycf4*, *cemA*, *rpl20*, *clpP*, *rps11*, *rps8*, *rpl22*, *ndhF*, *ndhI*) in *H. rosa-sinensis* evolved faster whereas 19 genes (*psbA*, *atpI*, *rps2*, *rpoC2*, *rps14*, *psaB*, *ycf3*, *rbcL*, *petA*, *rpl33*, *psbT*, *rpoA*, *rps3*, *rps19*, *ycf2*, *ycf1*, *ndhG*, *ndhA* and *ndhH*) evolved faster in *H. syriacus*. In genus *Firmiana*, evolutionary rate was similar in the 34 (44.2%) genes among two species of *Firmiana* whereas 43 genes (about 55.8%) had the differences in Ka/Ks more than 5%. We found 15 genes (*psaB*, *rbcL*, *accD*, *clpP*, *psbT*, *rpoA*, *rps8*, *rps3*, *rpl22*, *ycf1*, *ndhF*, *rpl32*, *ccsA*, *ndhH*, and *rps15*) fast evolving in *Firmiana major* whereas 28 genes (*psbA*, *matK*, *psbK*, *atpA*, *atpF*, *rps2*, *rpoC1*, *psbM*, *psaA*, *rps4*, *ndhK*, *ndhC*, *atpE*, *atpB*, *ycf4*, *cemA*, *petA*, *rps18*, *rpl20*, *psbB*, *petB*, *petD*, *rps11*, *rpl14*, *rpl2*, *ycf2*, *ndhD*, *ndhH*) were fast evolving in *Firmiana colorata*.

Table 3.2 Comparison of evolutionary rates of 77 protein coding genes among Malvaceae species

Gene	Species	Ks	Ka	Ka/Ks	Gene	Species	Ks	Ka	Ka/Ks
<i>PsbA</i>	HS	0.0411	0.0025	0.0608	<i>petG</i>	HS	0.0364	0	0
	RS	0.067	0.0025	0.0373		RS	0	0	0
	FM	0.0496	0.0025	0.0504		FM	0	0	0
	FC	0.0453	0.0025	0.0551		FC	0	0	0
	MP	0.067	0.0012	0.0179		MP	0	0	0
	MC	0.0627	0.0012	0.0191		MC	0	0	0
	RT	0.0327	0.0025	0.0764		RT	0	0	0
	UP	0.0497	0.0012	0.0241		UP	0	0	0
	SM	0.0496	0.0012	0.0241		SM	0.0364	0	0
	PT	0.0411	0.0012	0.0291		PT	0	0	0
	TA	0.0369	0.0012	0.0325		TA	0	0	0
	BC	0.0474	0.0026	0.0548		BC	0.0364	0	0
	GH	0.054	0.0012	0.0222		GH	0	0	0
	AO	0.0583	0.0012	0.0205		AO	0	0	0
	AE	0.054	0.0012	0.0222		AE	0	0	0
	HP	0.0539	0.0025	0.0463		HP	0	0	0
	DZ	0.0411	0.0012	0.0291		DZ	1	0.0122	0.0122
	TH	0.066	0.003	0.0454		TH	0.0365	0	0
<i>matK</i>	HS	0.0753	0.0368	0.4887	<i>psaJ</i>	HS	0.0667	0	0
	RS	0.0866	0.0444	0.5127		RS	0.0667	0	0
	FM	0.0688	0.0294	0.4273		FM	0.0667	0	0
	FC	0.0689	0.0329	0.4775		FC	0	0	0
	MP	0.0995	0.0435	0.4371		MP	0.0667	0	0
	MC	0.1099	0.0416	0.3785		MC	0.0667	0	0
	RT	0.0576	0.0316	0.5486		RT	0.1024	0	0
	UP	0.0837	0.0434	0.5185		UP	0.1024	0	0
	SM	0.0767	0.0371	0.4837		SM	0.1024	0	0
	PT	0.0287	0.0347	1.209		PT	0.0667	0	0
	TA	0.045	0.0236	0.5244		TA	0.0667	0	0
	BC	0.0609	0.0317	0.5205		BC	0.1024	0	0
	GH	0.0956	0.0508	0.5313		GH	0.0667	0	0
	AO	0.0995	0.0426	0.4281		AO	0.0667	0	0
	AE	0.0898	0.0397	0.442		AE	0.667	0	0
	HP	0.0608	0.0325	0.5345		HP	0.1024	0	0
	DZ	0.0559	0.0321	0.5742		DZ	0.0667	0	0
	TH	0.0766	0.0353	0.4608		AE	0.667	0	0
<i>rps16</i>	HS	0.0466	0	0	<i>rpl33</i>	HS	0.0709	0.0134	0.1889
	RS	0.0152	0	0		RS	0.0709	0	0
	FM	0	0	0		FM	0	0.0065	N/A
	FC	0.0152	0	0		FC	0	0.0065	N/A
	MP	0.0307	0	0		MP	0.0462	0.0132	0.2857
	MC	0.0307	0.0051	0.1661		MC	0.0462	0.0132	0.2857
	RT	0.0153	0.0051	0.3333		RT	0.0229	0	0

	UP	0.0466	0	0		UP	0.0465	0	0
	SM	0.0151	0.0102	0.6754		SM	0	0.0131	N/A
	PT	0	0.0051	N/A		PT	0	0.0131	N/A
	TA	0	0	0		TA	0.023	0.0065	0.2826
	BC	0.0308	0.0051	0.1655		BC	0.0229	0	0
	GH	0.0305	0.0154	0.5049		GH	0.0709	0	0
	AO	0.0307	0	0		AO	0.0462	0.0132	0.2857
	AE	0.0466	0	0		AE	0.0465	0	0
	HP	0.0307	0.0051	0.1661		HP	0	0.0065	N/A
	DZ	0.0632	0.0052	0.0822		DZ	0	0.0131	N/A
	TH	0.0308	0	0		TH	0.0711	0	0
<i>psbK</i>	HS	0.0524	0.007	0.1335	<i>rps18</i>	HS	0.0143	0	0
	RS	0.0524	0.007	0.1335		RS	0.0143	0	0
	FM	0.0524	0.007	0.1335		FM	0.0143	0.0044	0.3076
	FC	0.052	0.0141	0.2711		FC	0.0144	0.013	0.9027
	MP	0.0808	0.007	0.0866		MP	0.0435	0.0043	0.0988
	MC	0.0741	0	0		MC	0.0435	0.0043	0.0988
	RT	0.027	0.0218	0.8074		RT	0.0143	0.0087	0.6083
	UP	0.0801	0.0212	0.2646		UP	0.0143	0	0
	SM	0.0799	0.0141	0.1764		SM	0.0288	0.0043	0.1493
	PT	0.0801	0.007	0.0873		PT	0.0288	0.0087	0.302
	TA	0.0524	0	0		TA	0.0286	0	0
	BC	0.0524	0.007	0.1335		BC	0.0143	0	0
	GH	0.0524	0.007	0.1335		GH	0.0288	0.0043	0.1493
	AO	0.0801	0	0		AO	0.0435	0.0087	0.2
	AE	0.0524	0.007	0.1335		AE	0.0143	0	0
HP	0.0794	0.0141	0.1775	HP	0.0143	0.0043	0.3006		
DZ	0.0522	0.0141	0.2701	DZ	0.0471	0.019	0.4033		
TH	0.0801	0.007	0.0873	TH	0.0144	0	0		
<i>psbI</i>	HS	0.079	0	0	<i>rpl20</i>	HS	0.0601	0.0153	0.2545
	RS	0.079	0	0		RS	0.0909	0.0173	0.1903
	FM	0.0385	0	0		FM	0.0555	0.0244	0.4396
	FC	0.0385	0	0		FC	0.043	0.0245	0.5697
	MP	0.1219	0	0		MP	0.0733	0.0152	0.2073
	MC	0.1219	0	0		MC	0.0859	0.0114	0.1327
	RT	0.0385	0	0		RT	0.0998	0.0308	0.3086
	UP	0.1697	0	0		UP	0.0727	0.0192	0.264
	SM	0.079	0	0		SM	0.0683	0.0205	0.3001
	PT	0	0	0		PT	0.0481	0.023	0.4781
	TA	0	0	0		TA	0.0355	0.0114	0.3211
	BC	0	0	0		BC	0.048	0.0191	0.3979
	GH	0.0385	0	0		GH	0.1319	0.0172	0.1304
	AO	0.1219	0	0		AO	0.073	0.0114	0.1561
	AE	26.67	0	0		AE	0.0509	0.0159	0.3123
HP	0.0385	0	0	HP	0.0679	0.0244	0.3593		
DZ	0.0355	0.1085	3.0563	DZ	0.0646	0.0599	0.9272		

	TH	0.0423	0	0		TH	0.0603	0.0114	0.189
<i>atpA</i>	HS	0.0736	0.0044	0.0597	<i>rps12</i>	HS	0	0.0037	N/A
	RS	0.0617	0.0052	0.0842		RS	0	0.0037	N/A
	FM	0.0501	0.007	0.1397		FM	0	0	0
	FC	0.0416	0.007	0.1682		FC	0	0	0
	MP	0.0214	0	0		MP	0	0	0
	MC	0.0675	0.0079	0.117		MC	0	0.0037	N/A
	RT	0.0502	0.0025	0.0498		RT	0	0.0037	N/A
	UP	0.0826	0.0035	0.0423		UP	0	0.0037	N/A
	SM	0.0473	0.0044	0.093		SM	0.0208	0.0037	N/A
	PT	0.0502	0.0035	0.0697		PT	0	0	0
	TA	0.0388	0.0044	0.1134		TA	0	0	0
	BC	0.0473	0.0035	0.0739		BC	0	0	0
	GH	0.0588	0.0053	0.0901		GH	0	0.0037	N/A
	AO	0.0588	0.007	0.119		AO	0	0.0037	N/A
	AE	0.0619	0.0052	0.084		AE	0	0.0037	N/A
	HP	0.036	0.0053	0.1472		HP	0	0.0037	N/A
	DZ	0.0456	0.0036	0.0789		DZ	0	0	0
	TH	0.0765	0.0044	0.0575		TH	0	0.0037	N/A
	<i>atpF</i>	HS	0.058	0.0142		0.2448	<i>clpP</i>	HS	0.0518
RS		0.0669	0.0166	0.2481	RS	0.0518		0.0319	0.6158
FM		0.0412	0.0047	0.114	FM	0.0219		0.1637	7.4748
FC		0.0413	0.0071	0.1719	FC	0.0367		0.1446	3.94
MP		0.0512	0.0246	0.4804	MP	0.0677		0.0319	0.4711
MC		0.0582	0.019	0.3264	MC	0.0598		0.0295	0.4933
RT		0.0497	0.0023	0.0462	RT	0.0446		0.0534	1.1973
UP		0.0754	0.0142	0.1883	UP	0.0597		0.0226	0.3785
SM		0.0412	0.0047	0.114	SM	0.0325		0.1113	3.4246
PT		0.0672	0.0047	0.0699	PT	0.0295		0.0365	1.2372
TA		0.0502	0.0118	0.235	TA	0.0368		0.0203	0.5516
BC		0.0582	0.0094	0.1615	BC	0.0368		0.018	0.4891
GH		0.0579	0.0094	0.1623	GH	0.0597		0.018	0.3015
AO		0.0414	0.0194	0.4685	AO	0.0597		0.0296	0.4958
AE		0.0842	0.0166	0.1971	AE	0.0486		0.0565	1.1625
HP		0.0495	0.0142	0.2868	HP	0.0183		0.101	5.5191
DZ	0.0495	0.0118	0.2383	DZ	0.176	0.2107	1.1971		
TH	0.0667	0.0142	0.2128	TH	0.0367	0.0296	0.8065		
<i>atpH</i>	HS	0.0632	0	0	<i>psbB</i>	HS	0.0696	0.0009	0.0129
	RS	0.0469	0	0		RS	0.0724	0.0009	0.0124
	FM	0.0803	0	0		FM	0.0544	0.0009	0.0165
	FC	0.0632	0	0		FC	0.0694	0.0017	0.0244
	MP	0.0166	0	0		MP	0.0846	0.0017	0.02
	MC	0.0632	0	0		MC	0.0908	0.0017	0.0187
	RT	0.0632	0	0		RT	0.0543	0.0009	0.0165
	UP	0.0799	0	0		UP	0.0785	0.0009	0.0114
	SM	0.0799	0	0		SM	0.0574	0.0017	0.0296

	PT	0.0469	0	0		PT	0.0544	0.0009	0.0165
	TA	0.0469	0	0		TA	0.0397	0.0009	0.0226
	BC	0.0632	0	0		BC	0.0455	0.0009	0.0197
	GH	0.0799	0	0		GH	0.0574	0.0017	0.0296
	AO	0.0632	0	0		AO	0.0818	0.0017	0.0207
	AE	0.0471	0	0		AE	0.0755	0.0009	0.0119
	HP	0.062	0	0		HP	0.0484	0.0017	0.0351
	DZ	0.0468	0.0057	0.1217		DZ	0.0426	0.0026	0.061
	TH	0.0632	0	0		TH	0.0724	0.0009	0.0124
<i>atpI</i>	HS	0.0486	0.0063	0.1296	<i>psbT</i>	HS	0.042	0.0135	0.3214
	RS	0.0515	0.0036	0.0699		RS	0.0864	0.0135	0.1562
	FM	0.0516	0.0036	0.0697		FM	0.0406	0.0255	0.628
	FC	0.0516	0.0036	0.0697		FC	0.0834	0.0255	0.3057
	MP	0.0668	0.0135	0.202		MP	0.0883	0.0411	0.4654
	MC	0.0698	0.0126	0.1805		MC	0.0883	0.0411	0.4654
	RT	0.0282	0	0		RT	0.0864	0.0135	0.1562
	UP	0.0515	0.0036	0.0699		UP	0.042	0.0135	0.3214
	SM	0.0514	0.0072	0.14		SM	0.0406	0.0255	0.628
	PT	0.0225	0.0018	0.08		PT	0.042	0.0135	0.3214
	TA	0.0282	0.0036	0.1276		TA	0.0834	0.0255	0.3057
	BC	0.0397	0.0036	0.0906		BC	0.0406	0.0255	0.628
	GH	0.0827	0.0054	0.0652		GH	0.0406	0.052	1.2807
	AO	0.0637	0.0126	0.1978		AO	0.0883	0.0411	0.4654
	AE	0.0634	0.0054	0.0851		AE	0.042	0.0135	0.3214
	HP	0.0398	0.0036	0.0904		HP	0.0406	0.0255	0.628
DZ	0.0456	0.0036	0.0789	DZ	0.0406	0.0255	0.628		
TH	0.0456	0.0036	0.0789	TH	0.042	0.0135	0.3214		
<i>rps2</i>	HS	0.0244	0.0093	0.3811	<i>psbN</i>	HS	0.0663	0	0
	RS	0.0245	0.0074	0.302		RS	0.0663	0	0
	FM	0.0307	0.0074	0.241		FM	0.0324	0	0
	FC	0.0183	0.0074	0.4043		FC	0.0324	0	0
	MP	0.037	0.0056	0.1513		MP	0.0324	0	0
	MC	0.037	0.0056	0.1513		MC	0.0324	0	0
	RT	0.0182	0.0074	0.4065		RT	0.0324	0	0
	UP	0.0245	0.0074	0.302		UP	0.1019	0	0
	SM	0.0308	0.0093	0.3019		SM	0.0324	0	0
	PT	0.0306	0.0056	0.183		PT	0.0324	0	0
	TA	0.0244	0.0018	0.0737		TA	0.0324	0	0
	BC	0.0121	0.0074	0.6115		BC	0.0324	0	0
	GH	0.0497	0.0074	0.1488		GH	0.0663	0	0
	AO	0.037	0.0056	0.1513		AO	0.0324	0	0
	AE	0.0245	0.0093	0.3795		AE	0.1019	0	0
	HP	0.0183	0.0149	0.8142		HP	0.1019	0	0
DZ	0.0121	0.0112	0.9256	DZ	0.1021	0	0		
TH	0.0244	0.0112	0.459	TH	0.1019	0	0		
<i>rpoC2</i>	HS	0.0611	0.0165	0.27	<i>psbH</i>	HS	0.0568	0.0185	0.3257

	RS	0.0708	0.0166	0.2344		RS	0.0565	0.0185	0.3274
	FM	0.0332	0.0117	0.3524		FM	0	0.0123	N/A
	FC	0.0428	0.0145	0.3387		FC	0.0185	0.0123	0.6648
	MP	0.064	0.0147	0.2296		MP	0.0568	0.0123	0.2165
	MC	0.0724	0.0142	0.1961		MC	0.0767	0.0123	0.1603
	RT	0.0424	0.011	0.2594		RT	0	0.0123	N/A
	UP	0.0686	0.0182	0.2653		UP	0.0374	0.0185	0.4946
	SM	0.0334	0.012	0.3592		SM	0.0183	0.0248	1.3551
	PT	0.0502	0.0168	0.3346		PT	0.0279	0.0154	0.5519
	TA	0.029	0.0093	0.3206		TA	0	0.0123	N/A
	BC	0.0397	0.0099	0.2493		BC	0	0.0185	N/A
	GH	0.0588	0.0187	0.318		GH	0.0187	0.0061	0.3262
	AO	0.0629	0.0153	0.2432		AO	0.0568	0.0123	0.2165
	AE	0.0672	0.0179	0.2663		AE	0.0568	0.0185	0.3257
	HP	0.0395	0.0158	0.4		HP	0	0.0248	N/A
	DZ	0.0257	0.0154	0.5992		DZ	0.0374	0.0185	0.4946
	TH	0.0641	0.0176	0.2745		TH	0.0374	0.0185	0.4946
	<i>rpoC1</i>	HS	0.0681	0.01		0.1468	<i>petB</i>	HS	0.0799
RS		0.0579	0.0097	0.1675	RS	0.0941		0	0
FM		0.0514	0.0071	0.1381	FM	0.0729		0.0021	0.0288
FC		0.0493	0.0077	0.1561	FC	0.0591		0.0021	0.0355
MP		0.0447	0.009	0.2013	MP	0.087		0.0041	0.0471
MC		0.0513	0.009	0.1754	MC	0.0799		0.0021	0.0262
RT		0.0426	0.0071	0.1666	RT	0.066		0	0
UP		0.0536	0.0103	0.1921	UP	0.1157		0	0
SM		0.049	0.0045	0.0918	SM	0.0598		0	0
PT		0.058	0.0045	0.0775	PT	0.066		0	0
TA		0.036	0.0039	0.1083	TA	0.0591		0	0
BC		0.0426	0.0084	0.1971	BC	0.0524		0	0
GH		0.0512	0.0123	0.2402	GH	0.0799		0.0021	0.0262
AO		0.0468	0.0097	0.2072	AO	0.0799		0.0021	0.0262
AE		0.0535	0.0097	0.1813	AE	0.1084		0	0
HP		0.0469	0.009	0.1918	HP	0.0539		0.0041	0.076
DZ		0.0425	0.011	0.2588	DZ	0.0525		0.0041	0.078
TH		0.0491	0.0097	0.1975	TH	0.1231		0	0
<i>rpoB</i>	HS	0.0453	0.0049	0.1081	<i>petD</i>	HS	0.0525	0	0
	RS	0.0425	0.0057	0.1341		RS	0.0525	0	0
	FM	0.0326	0.0045	0.138		FM	0.0346	0	0
	FC	0.034	0.0045	0.1323		FC	0.0527	0.0028	0.0531
	MP	0.0502	0.0043	0.0856		MP	0.0525	0	0
	MC	0.0513	0.0047	0.0916		MC	0.0435	0	0
	RT	0.0466	0.0053	0.1137		RT	0.0435	0	0
	UP	0.0483	0.0078	0.1614		UP	0.0617	0.0055	0.0891
	SM	0.0298	0.009	0.302		SM	0.0258	0	0
	PT	0.0467	0.0061	0.1306		PT	0.0435	0	0
	TA	0.0257	0.0037	0.1439		TA	0.0258	0	0

	BC	0.0244	0.0041	0.168		BC	0.0346	0	0
	GH	0.0501	0.0059	0.1177		GH	0.0525	0	0
	AO	0.0474	0.0043	0.0907		AO	0.0525	0	0
	AE	0.0353	0.0061	0.1728		AE	0.0434	0	0
	HP	0.0313	0.0061	0.1948		HP	0.0346	0.0028	0.0809
	DZ	0.0299	0.0086	0.2876		DZ	0.0346	0	0
	TH	0.0424	0.0069	0.1627		TH	0.0614	0	0
<i>petN</i>	HS	0.0496	0	0	<i>rpoA</i>	HS	0.1097	0.0132	0.1203
	RS	0.0496	0	0		RS	0.1097	0.0119	0.1084
	FM	0.0496	0	0		FM	0.061	0.0232	0.3803
	FC	0	0	0		FC	0.0711	0.0218	0.3066
	MP	0.107	0	0		MP	0.0989	0.0172	0.1739
	MC	0.1599	0	0		MC	0.0962	0.0179	0.186
	RT	0	0	0		RT	0.0633	0.0105	0.1658
	UP	0.0496	0	0		UP	0.0988	0.0092	0.0931
	SM	0	0	0		SM	0.0782	0.0145	0.1854
	PT	0	0	0		PT	0.0634	0.0092	0.1451
	TA	0	0	0		TA	0.0482	0.0079	0.1639
	BC	0.0496	0	0		BC	0.0679	0.0132	0.1944
	GH	0.0496	0	0		GH	0.0912	0.0179	0.1962
	AO	0.1027	0	0		AO	0.0882	0.0172	0.195
	AE	0.0496	0	0		AE	0.099	0.0132	0.1333
	HP	0	0	0		HP	0.0784	0.0132	0.1683
DZ	0	0	0	DZ	0.0579	0.0079	0.1364		
TH	0.0496	0	0	TH	0.1043	0.0105	0.1006		
<i>psbM</i>	HS	0.0411	0	0	<i>rps11</i>	HS	0.0781	0.0066	0.0845
	RS	0.0411	0	0		RS	0.0679	0.0165	0.243
	FM	0.0411	0	0		FM	0.0578	0.0066	0.1141
	FC	0.0406	0.0132	0.3251		FC	0.0578	0.0099	0.1712
	MP	0.1298	0	0		MP	0.068	0.0066	0.097
	MC	0.1298	0	0		MC	0.0993	0.0066	0.0664
	RT	0.0411	0	0		RT	0.068	0.0099	0.1455
	UP	0.0406	0.0132	0.3251		UP	0.0781	0.0099	0.1267
	SM	0.1308	0	0		SM	0.0479	0.0099	0.2066
	PT	0.1327	0.013	0.0979		PT	0.0578	0.0066	0.1141
	TA	0.0846	0	0		TA	0.0478	0.0066	0.138
	BC	0.0406	0.0132	0.3251		BC	0.0478	0.0066	0.138
	GH	0.0411	0	0		GH	0.0991	0.0099	0.0998
	AO	0.1298	0	0		AO	0.0783	0.0099	0.1264
	AE	0.0411	0	0		AE	0.0782	0.0132	0.1687
	HP	0.411	0	0		HP	0.0383	0.0132	0.3446
DZ	0.0846	0	0	DZ	0.0725	0.0421	0.5806		
TH	0.0834	0.0132	0.1582	TH	0.0991	0.0099	0.0998		
<i>psbD</i>	HS	0.0379	0	0	<i>rpl36</i>	HS	0.0392	0.0119	0.3035
	RS	0.0498	0	0		RS	0.0392	0.0119	0.3035
	FM	0.0286	0	0		FM	0	0	0

	FC	0.0328	0	0		FC	0	0	0
	MP	0.0583	0	0		MP	0.0806	0.0119	0.1476
	MC	0.0584	0	0		MC	0.0806	0.0119	0.1476
	RT	0.0286	0	0		RT	0	0	0
	UP	0.0498	0.0012	0.024		UP	0.0392	0.0119	0.3035
	SM	0.037	0	0		SM	0.0795	0	0
	PT	0.0162	0	0		PT	0	0.0119	N/A
	TA	0.0162	0	0		TA	0.0806	0	0
	BC	0.0328	0	0		BC	0	0	0
	GH	0.0583	0	0		GH	0.0814	0.0118	0.1449
	AO	0.0627	0.0012	0.0191		AO	0.0806	0.0119	0.1476
	AE	0.0627	0.0012	0.0191		AE	0.0392	0.0119	0.3035
	HP	0.0286	0.0012	0.0419		HP	0	0	0
	DZ	0.0329	0	0		DZ	0.0385	0.012	0.3116
	TH	0.0498	0	0		TH	0.0392	0.0119	0.3035
<i>psbC</i>	HS	0.0721	0	0	<i>rps8</i>	HS	0.0618	0.01	0.1618
	RS	0.0809	0	0		RS	0.0619	0.0168	0.2714
	FM	0.0443	0	0		FM	0.01	0.0033	0.33
	FC	0.0413	0.0009	0.0217		FC	0.0201	0.0033	0.1641
	MP	0.0534	0.0009	0.0168		MP	0.0728	0.0168	0.2307
	MC	0.0473	0.0019	0.0401		MC	0.0728	0.0168	0.2307
	RT	0.0322	0	0		RT	0.02	0.0067	0.335
	UP	0.069	0	0		UP	0.0304	0.0134	0.4407
	SM	0.0414	0.0028	0.0676		SM	0.0201	0.01	0.4975
	PT	0.0535	0	0		PT	0.0406	0.01	0.2463
	TA	0.0263	0.0009	0.0342		TA	0.01	0.0033	0.33
	BC	0.0352	0	0		BC	0.0303	0.0067	0.2211
	GH	0.0534	0	0		GH	0.0619	0.0168	0.2714
	AO	0.0443	0.0009	0.0203		AO	0.0728	0.0168	0.2307
	AE	0.069	0.0009	0.013		AE	0.0303	0.01	0.33
HP	0.0383	0	0	HP	0.0199	0.0067	0.3366		
DZ	0.0412	0	0	DZ	0.02	0.01	0.5		
TH	0.0596	0	0	TH	0.0303	0.01	0.33		
<i>psbZ</i>	HS	0.0677	0	0	<i>rpl14</i>	HS	0.046	0.0036	0.0782
	RS	0.0914	0.0072	0.0787		RS	0.0702	0	0
	FM	0.0917	0	0		FM	0.0227	0.0036	0.1585
	FC	0.1165	0	0		FC	0.0226	0.0073	0.323
	MP	0.0219	0	0		MP	0.0277	0	0
	MC	0.0219	0	0		MC	0.0344	0	0
	RT	0.0445	0	0		RT	0.0462	0	0
	UP	0.0445	0	0		UP	0.0462	0.0036	0.0779
	SM	0.1165	0	0		SM	0.0299	0.0131	0.4381
	PT	0.0445	0	0		PT	0.022	0	0
	TA	0.0445	0	0		TA	0.0229	0.0036	0.1572
	BC	0.0218	0.0072	0.3302		BC	0.0344	0	0
GH	0.0219	0	0	GH	0.058	0.0036	0.062		

	AO	0.0219	0	0		AO	0.0113	0	0
	AE	0.0445	0	0		AE	0.0462	0	0
	HP	0.0914	0.0072	0.0787		HP	0.0344	0	0
	DZ	0.0219	0	0		DZ	0.0344	0.0183	0.5319
	TH	0.0445	0	0		TH	0.046	0.0073	0.1586
<i>rps14</i>	HS	0.03	0.0087	0.29	<i>rpl16</i>	HS	0.0517	0	0
	RS	0.03	0.0043	0.1433		RS	0.0411	0.0033	0.0802
	FM	0.0148	0.0043	0.2905		FM	0.0306	0	0
	FC	0.0148	0.0043	0.2905		FC	0.0411	0	0
	MP	0.0774	0.0043	0.0555		MP	0.0411	0.0066	0.1605
	MC	0.0774	0.0043	0.0555		MC	0.0306	0.0066	0.2156
	RT	0.0454	0.0043	0.0947		RT	0.0517	0	0
	UP	0.0454	0.0043	0.0947		UP	0.0847	0.0033	0.0389
	SM	0.0299	0.0131	0.4381		SM	0.0399	0	0
	PT	0.0149	0.0087	0.5838		PT	0.0502	0	0
	TA	0.03	0.0043	0.1433		TA	0.0306	0	0
	BC	0.03	0.0043	0.1433		BC	0.0411	0	0
	GH	0.0612	0.0043	0.0702		GH	0.0842	0	0
	AO	0.0774	0.0043	0.0555		AO	0.0306	0.0066	0.2156
	AE	0.0452	0.0087	0.1924		AE	0.0517	0	0
	HP	0.0148	0.0043	0.2905		HP	0.0516	0.0033	0.0639
DZ	0.0299	0.0175	0.5852	DZ	0.0411	0.0066	0.1605		
TH	0.0454	0.0043	0.0947	TH	0.0412	0.0033	0.08		
<i>psaB</i>	HS	0.0406	0.003	0.0738	<i>rps3</i>	HS	0.1254	0.0139	0.1108
	RS	0.0511	0.0035	0.0684		RS	0.1333	0.0139	0.1042
	FM	0.0395	0.0027	0.0683		FM	0.0905	0.0069	0.0762
	FC	0.0458	0.0027	0.0589		FC	0.0983	0.0069	0.0701
	MP	0.066	0.0041	0.0621		MP	0.0715	0.0099	0.1384
	MC	0.0703	0.0041	0.0583		MC	0.0641	0.0119	0.1856
	RT	0.0302	0.003	0.0993		RT	0.0869	0.0079	0.0909
	UP	0.049	0.003	0.0612		UP	0.1013	0.016	0.1579
	SM	0.05	0.005	0.1		SM	0.1355	0.0099	0.073
	PT	0.0427	0.0024	0.0562		PT	0.0866	0.0139	0.1605
	TA	0.0302	0.0024	0.0794		TA	0.0795	0.0079	0.0993
	BC	0.0323	0.0024	0.0743		BC	0.0641	0.0079	0.1232
	GH	0.0596	0.0024	0.0402		GH	0.1138	0.0169	0.1485
	AO	0.0681	0.0041	0.0602		AO	0.0717	0.0119	0.1659
	AE	0.0511	0.003	0.0587		AE	0.1054	0.015	0.1423
	HP	0.0406	0.0035	0.0862		HP	0.0869	0.0059	0.0678
DZ	0.0302	0.0024	0.0794	DZ	0.1098	0.0201	0.183		
TH	0.066	0.003	0.0454	TH	0.1013	0.016	0.1579		
<i>psaA</i>	HS	0.0531	0.0006	0.0112	<i>rpl22</i>	HS	0.093	0.0799	0.8591
	RS	0.0633	0.0012	0.0189		RS	0.0788	0.0754	0.9568
	FM	0.0329	0.0006	0.0182		FM	0.0517	0.0257	0.497
	FC	0.0429	0.0012	0.0279		FC	0.0739	0.0259	0.3504
	MP	0.0531	0.0006	0.0112		MP	0.0798	0.0509	0.6378

	MC	0.0592	0.0006	0.0101		MC	0.0797	0.092	1.1543
	RT	0.0449	0.0006	0.0133		RT	0.0763	0.0359	0.4705
	UP	0.0571	0.0006	0.0105		UP	0.0702	0.0845	1.2037
	SM	0.0499	0.0006	0.012		SM	0.0306	0.0359	1.1732
	PT	0.0389	0	0		PT	0.0625	0.0388	0.6208
	TA	0.0251	0.0006	0.0239		TA	0.0315	0.0164	0.5206
	BC	0.0389	0.0006	0.0154		BC	0.0603	0.0462	0.7661
	GH	0.053	0.0017	0.032		GH	0.0354	0.0326	0.9209
	AO	0.051	0	0		AO	0.0634	0.0554	0.8738
	AE	0.0551	0.0006	0.0108		AE	0.0809	0.0872	1.0778
	HP	0.031	0.0012	0.0387		HP	0.031	0.0226	0.729
	DZ	0.033	0	0		DZ	0.0702	0.0845	0.729
	TH	0.0551	0.0006	0.0108		TH	0.0701	0.0802	1.144
<i>ycf3</i>	HS	0.0531	0.0026	0.0489	<i>rps19</i>	HS	0.1059	0.0047	0.0443
	RS	0.0624	0.0026	0.0416		RS	0.1251	0.0047	0.0375
	FM	0.0436	0.0052	0.1192		FM	0.0511	0.0047	0.0919
	FC	0.0436	0.0052	0.1192		FC	0.0511	0.0047	0.0919
	MP	0.035	0.0026	0.0742		MP	0.0689	0.0047	0.0682
	MC	0.035	0.0026	0.0742		MC	0.0689	0.0047	0.0682
	RT	0.0348	0	0		RT	0.0872	0.0047	0.0538
	UP	0.0531	0.0052	0.0979		UP	0.1661	0.0093	0.0559
	SM	0.062	0.0026	0.0419		SM	0.0689	0.0093	0.1349
	PT	0.0529	0.0026	0.0491		PT	0.0689	0.0047	0.0682
	TA	0.0348	0	0		TA	0.0872	0.0047	0.0538
	BC	0.0348	0	0		BC	0.0689	0.0093	0.1349
	GH	0.0808	0.0026	0.0321		GH	0.1251	0.0093	0.0743
AO	0.044	0.0026	0.059	AO	0.0689	0.0047	0.0682		
AE	0.0624	0.0026	0.0416	AE	0.1452	0.0141	0.0971		
HP	0.0347	0.0026	0.0749	HP	0.0511	0.0188	0.3679		
DZ	0.0172	0	0	DZ	0.0693	0.0239	0.3448		
TH	0.0531	0.0052	0.0979	TH	0.1251	0.0047	0.0375		
<i>rps4</i>	HS	0.0499	0.0088	0.1763	<i>rpl2</i>	HS	0.0098	0	0
	RS	0.0429	0.0088	0.2051		RS	0.0092	0	0
	FM	0.0211	0.0132	0.6255		FM	0	0.0065	N/A
	FC	0.0281	0.0199	0.7081		FC	0.0098	0.0049	0.5
	MP	0.0875	0.0066	0.0754		MP	0.0098	0.0033	0.3367
	MC	0.0646	0.006	0.0928		MC	0.0098	0.0016	0.1632
	RT	0.0353	0.0088	0.2492		RT	0.0049	0	0
	UP	0.0461	0.0077	0.167		UP	0.0098	0	0
	SM	0.0281	0.0132	0.4697		SM	0	0.0033	N/A
	PT	0.0427	0.011	0.2576		PT	0.0049	0.0016	0.3265
	TA	0.021	0.0088	0.419		TA	0.0098	0	0
	BC	0.021	0.0088	0.419		BC	0.0049	0	0
	GH	0.0499	0.011	0.2204		GH	0.0297	0	0
AO	0.0722	0.0066	0.0914	AO	0.0098	0.0016	0.1632		
AE	0.0389	0.0077	0.1979	AE	0.0098	0	0		

	HP	0.0139	0.0066	0.4748		HP	0	0.0049	N/A
	DZ	0.0352	0.0132	0.375		DZ	0.0147	0.0016	0.1088
	TH	0.0461	0.0077	0.167		TH	0.0147	0	0
<i>ndhJ</i>	HS	0.1079	0	0	<i>rpl23</i>	HS	0	0.0047	N/A
	RS	0.1183	0	0		RS	0	0.0094	N/A
	FM	0.0869	0.0027	0.031		FM	0	0.0047	N/A
	FC	0.0869	0.0027	0.031		FC	0	0.0094	N/A
	MP	0.1077	0	0		MP	0	0.0047	N/A
	MC	0.0972	0	0		MC	0	0.0096	N/A
	RT	0.0767	0.0083	0.1082		RT	0	0.0047	N/A
	UP	0.1079	0.0027	0.025		UP	0	0.0142	N/A
	SM	0.0667	0	0		SM	0	0.0047	N/A
	PT	0.1077	0.0027	0.025		PT	0.0154	0.0047	0.3051
	TA	0.0667	0	0		TA	0	0.0047	N/A
	BC	0.0767	0	0		BC	0	0.0047	N/A
	GH	0.0972	0.0027	0.0277		GH	0	0	0
	AO	0.1077	0	0		AO	0	0.0047	N/A
	AE	0.1185	0	0		AE	0	0.0142	N/A
	HP	0.0767	0	0		HP	0	0.0047	N/A
	DZ	0.0869	0	0		DZ	0.0153	0.019	1.2418
TH	0.0974	0	0	TH	0	0.0047	N/A		
<i>ndhK</i>	HS	0.1179	0.0079	0.067	<i>yef2</i>	HS	0.0513	0.031	0.6042
	RS	0.1038	0.0099	0.0953		RS	0.0082	0.0042	0.5121
	FM	0.0571	0.002	0.035		FM	0.0088	0.0025	0.284
	FC	0.0772	0.0039	0.0505		FC	0.006	0.0032	0.5333
	MP	0.1108	0.0079	0.0712		MP	0.0079	0.0053	0.6708
	MC	0.111	0.0079	0.0711		MC	0.0086	0.0043	0.5
	RT	0.0838	0.0098	0.1169		RT	0.0054	0.0025	0.4629
	UP	0.125	0.0099	0.0792		UP	0.0106	0.0047	0.4433
	SM	0.0636	0.0079	0.1242		SM	0.0053	0.0021	0.3962
	PT	0.0634	0.0039	0.0615		PT	0.008	0.005	0.625
	TA	0.0571	0.0039	0.0683		TA	0.0026	0.0019	0.7307
	BC	0.0573	0.0059	0.1029		BC	0.006	0.0017	0.2833
	GH	0.104	0.0059	0.0567		GH	0.01	0.0047	0.47
	AO	0.1108	0.0079	0.0712		AO	0.0073	0.0039	0.5342
	AE	0.0969	0.0079	0.0815		AE	0.0097	0.0046	0.4742
	HP	0.0571	0.0039	0.0683		HP	0.004	0.0026	0.65
	DZ	0.0703	0	0		DZ	0.0435	0.0377	0.8666
TH	0.1108	0.0079	0.0712	TH	0.0086	0.0039	0.4534		
<i>ndhC</i>	HS	0.0235	0.0037	0.1574	<i>ndhB</i>	HS	0.0081	0.0009	0.1111
	RS	0.0116	0.0074	0.6379		RS	0.0081	0.0009	0.1111
	FM	0.0234	0.0037	0.1581		FM	0.0054	0	0
	FC	0.0166	0.0037	0.2228		FC	0.0054	0	0
	MP	0.0354	0.0037	0.1045		MP	0.0054	0.0017	0.3148
	MC	0.0117	0.0073	0.6239		MC	0.0054	0.0017	0.3148
	RT	0	0.0073	N/A		RT	0.0054	0.0009	0.1666

	UP	0.0117	0.0037	0.3162		UP	0.0054	0.0009	0.1666
	SM	0	0.0074	N/A		SM	0.0027	0	0
	PT	0	0.0111	N/A		PT	0.0081	0.009	1.1111
	TA	0.0116	0.0037	0.3189		TA	0.0027	0	0
	BC	0	0.0037	N/A		BC	0.0027	0	0
	GH	0	0.0074	N/A		GH	0.0027	0.0009	0.3333
	AO	0.0117	0.0073	0.6239		AO	0.0081	0.0017	0.2098
	AE	0	0.0074	N/A		AE	0.0054	0.0017	0.3148
	HP	0	0.0074	N/A		HP	0.0081	0	0
	DZ	0	0.0037	N/A		DZ	0.0054	0.0017	0.3148
	TH	0.0116	0.0037	0.3189		TH	0.0081	0.0009	0.1111
<i>atpE</i>	HS	0.1036	0	0	<i>rps7</i>	HS	0.0268	0	0
	RS	0.1156	0.0037	0.032		RS	0.027	0	0
	FM	0.091	0.0033	0.0362		FM	0.0088	0	0
	FC	0.079	0.0033	0.0417		FC	0.0088	0	0
	MP	0.0912	0.0033	0.0361		MP	0.0088	0	0
	MC	0.1034	0.0033	0.0319		MC	0.0177	0.0029	0.1638
	RT	0.0795	0.0099	0.1245		RT	0.0088	0	0
	UP	0.1291	0.0033	0.0255		UP	0.0178	0	0
	SM	0.0671	0.0099	0.1475		SM	0.0088	0	0
	PT	0.1036	0	0		PT	0.0088	0	0
	TA	0.0793	0.0033	0.0416		TA	0.0088	0	0
	BC	0.1292	0	0		BC	0.0088	0	0
	GH	0.1036	0	0		GH	0.0088	0	0
	AO	0.0912	0.0033	0.0361		AO	0.0088	0	0
	AE	0.1036	0	0		AE	0.0178	0	0
HP	0.0789	0.0066	0.0836	HP	0.0088	0	0		
DZ	0.0804	0.0265	0.3296	DZ	0.008	0.0261	3.2625		
TH	0.1286	0	0	TH	0.0178	0	0		
<i>atpB</i>	HS	0.065	0.0027	0.0415	<i>ycf1</i>	HS	0.0918	0.0616	0.671
	RS	0.0704	0.0045	0.0639		RS	0.1164	0.0596	0.512
	FM	0.0559	0.0036	0.0644		FM	0.0681	0.0538	0.79
	FC	0.0589	0.0063	0.1069		FC	0.0802	0.0533	0.6645
	MP	0.0768	0.0073	0.095		MP	0.1111	0.047	0.423
	MC	0.0768	0.0073	0.095		MC	0.1115	0.0502	0.4502
	RT	0.0501	0.0045	0.0898		RT	0.0827	0.0364	0.4401
	UP	0.0662	0.0059	0.0891		UP	0.1124	0.0521	0.4635
	SM	0.0736	0.0063	0.0855		SM	0.0909	0.0582	0.6402
	PT	0.0588	0.0027	0.0459		PT	0.0984	0.0597	0.6067
	TA	0.053	0.0036	0.0679		TA	0.0674	0.0318	0.4718
	BC	0.0443	0.0027	0.0609		BC	0.0649	0.0364	0.5608
	GH	0.0734	0.0054	0.0735		GH	0.1179	0.0484	0.4105
	AO	0.0798	0.0036	0.0451		AO	0.1115	0.0473	0.4242
	AE	0.0792	0.0027	0.034		AE	0.1111	0.0501	0.4509
HP	0.0558	0.0045	0.0806	HP	0.054	0.0527	0.9759		
DZ	0.05	0.0099	0.198	DZ	0.1609	0.1229	0.7638		

	TH	0.0792	0.0027	0.034		TH	0.1109	0.0468	0.422
<i>rbcL</i>	HS	0.0512	0.031	0.6054	<i>ndhF</i>	HS	0.0984	0.0245	0.2489
	RS	0.059	0.0257	0.4355		RS	0.1001	0.0336	0.3356
	FM	0.0328	0.0065	0.1981		FM	0.0713	0.0237	0.3323
	FC	0.039	0.0055	0.141		FC	0.0756	0.0231	0.3055
	MP	0.0658	0.0126	0.1914		MP	0.1138	0.0281	0.2469
	MC	0.0686	0.0191	0.2784		MC	0.1059	0.0241	0.2275
	RT	0.0359	0.0065	0.181		RT	0.0917	0.0189	0.2061
	UP	0.0627	0.0208	0.3317		UP	0.0928	0.0247	0.2661
	SM	0.0373	0.006	0.1608		SM	0.0821	0.0231	0.2813
	PT	0.0512	0.0112	0.2187		PT	0.0893	0.017	0.1903
	TA	0.0327	0.0074	0.2262		TA	0.0607	0.0141	0.2322
	BC	0.042	0.0056	0.1333		BC	0.0678	0.0171	0.2522
	GH	0.0561	0.0134	0.2388		GH	0.0931	0.0321	0.3447
	AO	0.0548	0.0131	0.239		AO	0.1037	0.0254	0.2449
	AE	0.0627	0.0237	0.3779		AE	0.0995	0.026	0.2613
	HP	0.0327	0.0074	0.2262		HP	0.0671	0.0276	0.4113
	DZ	0.0379	0.0114	0.3007		DZ	0.0651	0.0258	0.3963
	TH	0.0611	0.0222	0.3633		TH	0.1029	0.0243	0.2361
<i>accD</i>	HS	0.0714	0.0253	0.3543	<i>rpl32</i>	HS	0	0.0161	N/A
	RS	0.0572	0.0253	0.4423		RS	0	0.0161	N/A
	FM	0.0367	0.0183	0.4986		FM	0.0569	0.0243	0.427
	FC	0.0401	0.0184	0.4588		FC	0.0572	0.0165	0.2884
	MP	0.0575	0.0341	0.593		MP	0.0583	0.0082	0.1406
	MC	0.0613	0.0317	0.5171		MC	0.0282	0.0248	0.8794
	RT	0.0503	0.0192	0.3817		RT	0.0282	0.0165	0.5851
	UP	0.0472	0.0254	0.5381		UP	0.0277	0.025	0.9025
	SM	0.0482	0.0222	0.4605		SM	0.0305	0.0089	0.2918
	PT	0.0572	0.0271	0.4737		PT	0.0281	0.0242	0.8612
	TA	0.0331	0.0165	0.4984		TA	0	0.0082	N/A
	BC	0.0399	0.0165	0.4135		BC	0.0576	0.0082	0.1423
	GH	0.0644	0.0353	0.5481		GH	0.1046	0.0123	0.1175
	AO	0.0577	0.027	0.4676		AO	0.0282	0.0082	0.2907
	AE	0.054	0.0244	0.4518		AE	0	0	0
	HP	0.0346	0.0178	0.5144		HP	0.0285	0.0084	0.2947
	DZ	0.0544	0.0392	0.7205		DZ	0	0.0591	N/A
	TH	0.0504	0.0244	0.4841		TH	0	0.0161	N/A
<i>psaI</i>	HS	0	0	0	<i>ccsA</i>	HS	0.0769	0.0179	0.2327
	RS	0	0	0		RS	0.0673	0.015	0.2228
	FM	0.0385	0.0001	0.0025		FM	0.0471	0.0178	0.3779
	FC	0	0	0		FC	0.0619	0.015	0.2423
	MP	0	0	0		MP	0.1029	0.0137	0.1331
	MC	0	0	0		MC	0.0871	0.0164	0.1882
	RT	0	0	0		RT	0.0526	0.0193	0.3669
	UP	0	0	0		UP	0.1026	0.0191	0.1861
	SM	0	0	0		SM	0.0421	0.0219	0.5201

	PT	0.0395	0.0119	0.3012		PT	0.0669	0.0177	0.2645
	TA	0	0	0		TA	0.0375	0.015	0.4
	BC	0.0385	0	0		BC	0.0718	0.0109	0.1518
	GH	0.0385	0	0		GH	0.1052	0.0141	0.134
	AO	0	0	0		AO	0.0975	0.0137	0.1405
	AE	0	0.0242	N/A		AE	0.0874	0.0164	0.1876
	HP	0	0	0		HP	0.0427	0.0191	0.4473
	DZ	0	0	0		DZ	0.0472	0.0136	0.2881
	TH	0	0	0		TH	0.867	0.0108	0.0124
<i>ycf4</i>	HS	0.0832	0.0047	0.0564	<i>ndhD</i>	HS	0.086	0.012	0.1395
	RS	0.1012	0.0071	0.0701		RS	0.0844	0.0115	0.1362
	FM	0.0404	0.0047	0.1163		FM	0.0602	0.0053	0.088
	FC	0.0613	0.0083	0.1353		FC	0.0692	0.0071	0.1026
	MP	0.0744	0.0024	0.0322		MP	0.0876	0.0106	0.121
	MC	0.0744	0.0024	0.0322		MC	0.0967	0.0115	0.1189
	RT	0.0403	0.0071	0.1761		RT	0.0484	0.0062	0.128
	UP	0.0922	0.0071	0.077		UP	0.0945	0.0138	0.146
	SM	0.0486	0.0071	0.146		SM	0.0632	0.0124	0.1962
	PT	0.0572	0.0024	0.0419		PT	0.0784	0.0088	0.1122
	TA	0.0487	0.0047	0.0965		TA	0.0483	0.0044	0.091
	BC	0.0527	0.0024	0.0455		BC	0.0585	0.012	0.2051
	GH	0.1011	0	0		GH	0.0904	0.0115	0.1272
	AO	0.0744	0.0024	0.0322		AO	0.0845	0.0133	0.1573
	AE	0.092	0.0047	0.051		AE	0.0783	0.0106	0.1353
	HP	0.0572	0.0024	0.0419		HP	0.0662	0.0071	0.1072
DZ	0.0403	0	0	DZ	0.0665	0.0088	0.1323		
TH	0.0744	0.0047	0.0631	TH	0.0875	0.0097	0.1108		
<i>cemA</i>	HS	0.0704	0.0094	0.1335	<i>psaC</i>	HS	0.0736	0	0
	RS	0.0741	0.0122	0.1646		RS	0.0731	0.0054	0.0738
	FM	0.0487	0.0093	0.1909		FM	0.0736	0	0
	FC	0.0483	0.0131	0.2712		FC	0.0931	0	0
	MP	0.0858	0.0075	0.0874		MP	0.0736	0	0
	MC	0.0631	0.0056	0.0887		MC	0.0545	0	0
	RT	0.0343	0.0075	0.2186		RT	0.0359	0	0
	UP	0.0703	0.0112	0.1593		UP	0.0945	0.0138	0.146
	SM	0.0524	0.0065	0.124		SM	0.1132	0	0
	PT	0.0631	0.0019	0.0301		PT	0.0545	0	0
	TA	0.0415	0.0075	0.1807		TA	0.0177	0	0
	BC	0.0488	0.0056	0.1147		BC	0.0177	0	0
	GH	0.0926	0.015	0.1619		GH	0.0736	0	0
	AO	0.082	0.0103	0.1256		AO	0.0545	0	0
	AE	0.0704	0.0056	0.0795		AE	0.0545	0	0
	HP	0.0555	0.015	0.2702		HP	0.0731	0	0
DZ	0.0483	0.0284	0.5879	DZ	0.0177	0	0		
TH	0.0778	0.0056	0.0719	TH	0.0736	0	0		
<i>petA</i>	HS	0.0694	0.011	0.1585	<i>ndhG</i>	HS	0.0395	0.0076	0.1924

	RS	0.0693	0.0082	0.1183		RS	0.0558	0.005	0.0896
	FM	0.0408	0.0055	0.1348		FM	0.0395	0.0076	0.1924
	FC	0.0362	0.0068	0.1878		FC	0.0315	0.0076	0.2412
	MP	0.0454	0.0124	0.2731		MP	0.0392	0.0178	0.454
	MC	0.0549	0.011	0.2003		MC	0.0313	0.0178	0.5686
	RT	0.0647	0.011	0.17		RT	0.0234	0.005	0.2136
	UP	0.0693	0.0124	0.1789		UP	0.056	0.01	0.1785
	SM	0.055	0.011	0.2		SM	0.0234	0.005	0.2136
	PT	0.0695	0.0096	0.1381		PT	0.0395	0.0076	0.1924
	TA	0.0362	0.0082	0.2265		TA	0.0234	0.0076	0.3247
	BC	0.0549	0.0082	0.1493		BC	0.0155	0.0076	0.4903
	GH	0.079	0.0096	0.1215		GH	0.0476	0.005	0.105
	AO	0.0408	0.011	0.2696		AO	0.0313	0.0178	0.5686
	AE	0.0502	0.0082	0.1633		AE	0.0473	0.0101	0.2135
	HP	0.0408	0.0055	0.1348		HP	0.0155	0.0152	0.9806
	DZ	0.0311	0.011	0.3536		DZ	0.0234	0.0152	0.6495
	TH	0.0501	0.011	0.2195		TH	0.0557	0.0076	0.1364
	HS	0.0311	0	0		HS	0.0447	0.0052	0.1163
	RS	0.0311	0.0116	0.3729		RS	0.0356	0.0078	0.2191
	FM	0	0	0		FM	0.0541	0.0104	0.1922
	FC	0	0	0		FC	0.0542	0.0104	0.1918
	MP	0.0632	0.0166	0.2626		MP	0.0847	0.008	0.0944
	MC	0.0632	0.0166	0.2626		MC	0.0827	0.0131	0.1584
	RT	0.0309	0	0		RT	0.0266	0.0052	0.1954
	UP	0.0311	0	0		UP	0.0634	0.0052	0.082
	SM	0.0309	0	0		SM	0.0729	0.0026	0.0356
	PT	0.0308	0.0234	0.7597		PT	0.0447	0.0026	0.0581
	TA	0.0309	0	0		TA	0.0265	0.0026	0.0981
	BC	0	0	0		BC	0.0447	0.0026	0.0581
	GH	0	0	0		GH	0.073	0.0104	0.1424
	AO	0.0632	0.0116	0.1835		AO	0.0825	0.0157	0.1903
	AE	0.0311	0	0		AE	0.0635	0.0078	0.1228
	HP	0	0	0		HP	0.0634	0.0052	0.082
	DZ	0	0	0		DZ	0.0357	0.0052	0.1456
	TH	0.0311	0	0		TH	0.0729	0.0052	0.0713
	HS	0	0	0		HS	0.0969	0.008	0.0825
	RS	0	0	0		RS	0.0926	0.0092	0.0993
	FM	0.0423	0	0		FM	0.0663	0.0098	0.1478
	FC	0.0871	0	0		FC	0.0744	0.011	0.1478
	MP	0.0871	0	0		MP	0.0864	0.0086	0.0995
	MC	0.0423	0	0		MC	0.0865	0.0111	0.1283
	RT	0	0	0		RT	0.0581	0.0073	0.1256
	UP	0.0423	0	0		UP	0.0826	0.0148	0.1791
	SM	0.0871	0	0		SM	0.0785	0.0073	0.0929
	PT	0	0	0		PT	0.0784	0.0073	0.0931
	TA	0	0	0		TA	0.0661	0.0061	0.0922

	BC	0	0	0		BC	0.0621	0.0061	0.0982
	GH	0	0	0		GH	0.1118	0.0135	0.1207
	AO	0.0423	0	0		AO	0.0823	0.0086	0.1044
	AE	0.0423	0	0		AE	0.0863	0.0123	0.1425
	HP	0.0423	0	0		HP	0.0825	0.0073	0.0884
	DZ	0	0	0		DZ	0.0741	0.0086	0.116
	TH	0.0423	0	0		TH	0.0805	0.0142	0.1763
<i>psbF</i>	HS	0.0328	0	0	<i>ndhH</i>	HS	0.0717	0.0066	0.092
	RS	0	0	0		RS	0.0843	0.0044	0.0521
	FM	0.0328	0	0		FM	0.043	0.0077	0.179
	FC	0.0328	0	0		FC	0.0491	0.0082	0.167
	MP	0	0	0		MP	0.0633	0.0088	0.139
	MC	0	0	0		MC	0.051	0.0077	0.1509
	RT	0	0	0		RT	0.0591	0.0077	0.1302
	UP	0	0	0		UP	0.0716	0.0055	0.0768
	SM	0	0	0		SM	0.0676	0.0088	0.1301
	PT	0	0	0		PT	0.0593	0.0077	0.1298
	TA	0	0	0		TA	0.031	0.0077	0.2483
	BC	0	0	0		BC	0.0474	0.0044	0.0928
	GH	0	0	0		GH	0.0799	0.0055	0.0688
	AO	0.0328	0	0		AO	0.055	0.0077	0.14
	AE	0	0	0		AE	0.0633	0.0055	0.0868
	HP	0	0	0		HP	0.0389	0.0066	0.1696
	DZ	0	0	0		DZ	0.0591	0.0033	0.0558
TH	0.0328	0	0	TH	0.0714	0.0066	0.0924		
<i>psbE</i>	HS	0.0352	0	0	<i>rps15</i>	HS	0.0699	0.0096	0.1373
	RS	0.0534	0	0		RS	0.1078	0.0144	0.1335
	FM	0.0352	0	0		FM	0.1078	0.0144	0.1335
	FC	0.0534	0	0		FC	0.1272	0.0144	0.1132
	MP	0.0725	0	0		MP	0.1268	0.0144	0.1135
	MC	0.0534	0	0		MC	0.1073	0.0096	0.0894
	RT	0.0352	0	0		RT	0.0701	0.0144	0.2054
	UP	0.0352	0	0		UP	0.1073	0.0144	0.1342
	SM	0.0174	0	0		SM	0.0887	0.0341	0.3844
	PT	0.0534	0	0		PT	0.0342	0.0291	0.8508
	TA	0.0174	0	0		TA	0.1078	0.0193	0.179
	BC	0.0721	0	0		BC	0.0702	0.0193	0.2749
	GH	0.0352	0	0		GH	0.1253	0.0243	0.1939
	AO	0.0725	0	0		AO	0.1268	0.0096	0.0757
	AE	0.0352	0	0		AE	0.0887	0.0144	0.1623
	HP	0.0174	0	0		HP	0.0701	0.0144	0.2054
	DZ	0.0353	0.0053	0.1501		DZ	0.0703	0.0193	0.2745
TH	0.0352	0	0	TH	0.0886	0.0144	0.1625		
<i>petL</i>	HS	0.0414	0	0				0	
	RS	0.0414	0	0					
	FM	0	0	0					

FC	0	0	0				
MP	0.0414	0	0				
MC	0.0414	0	0				
RT	0.0408	0	0				
UP	0.0414	0	0				
SM	0	0	0				
PT	0	0	0				
TA	0	0	0				
BC	0.0414	0	0				
GH	0.0414	0	0				
AO	0.0414	0	0				
AE	0.0414	0.0148	0.3574				
HP	0	0	0				
DZ	0	0.015	N/A				
TH	0.0414	0	0				

RS: *H. rosa-sinensis*, HS: *H. syriacus*, HM: *H. mutabilis*, AE: *A. esculentus*; TH: *T. hamabo*, BC: *B. ceiba*; DZ: *D. zibethinus*, FM: *F. major*, FC: *F. colorata*, HP: *H. parvifolia*; TC: *T. cacao*; TA: *T. amurensis*; MP: *M. parviflora*, MC: *M. coromandelianum*, UP: *U. procumbens*, RT: *R. thyrsoides*, SM: *S. monosperma*, PT: *P. truncatolobatum*, GH: *G. herbaceum*, and AO: *A. officinalis*.

2.4 Conclusion

We conclude that chloroplast genome of Malvaceae is highly conserved in terms of genes content, genes order, GC content and introns content. The intergenic spacer regions showed variations in the length which thus caused variation in the length of chloroplast genomes of the family Malvaceae species. The IRs contraction and expansion lead to variable number of total genes. Moreover, origination of pseudogene was also observed at the junction of chloroplast genomes. The rate of synonymous and non-synonymous substitutions showed 95% similarities. However, the higher rate of synonymous substitutions than non-synonymous substitutions was observed. Certain genes were identified with positive selection pressure which revealed these genes might be important for these species in their ecological niches.

Chapter 4

**Correlations among oligonucleotide
repeats, nucleotide substitutions and
insertion – deletion mutations in
chloroplast genomes of plant family
Malvaceae**

4.1 Introduction

Several models have been proposed to explain the underlying mechanisms of molecular evolution: JC69 (Jukes and Cantor, 1969), HKY (Hasegawa *et al.*, 1985), K2P (Kimura, 1980), GTR (Tavaré, 1986). Among these, generalized time reversible (GTR) substitution model has been commonly considered as the best model to explain molecular evolution. A fundamental assumption of this model is the independent occurrence of mutational events at each site (Drouin *et al.*, 2008). However, some studies revealed complex processes of evolution due to occurrence of non-random spatial patterns and lineage-specific substitutions (Gruenheit *et al.*, 2008; Liò and Goldman, 1998; Zhong *et al.*, 2011). Certain parts of chloroplast genomes undergo slower rate of mutations (Palmer, 1985) compared to others (Ahmed *et al.*, 2013; Bi *et al.*, 2018; Menezes *et al.*, 2018; Zhang *et al.*, 2015).

The co-occurrence of mutational events has previously been reported in prokaryotic and eukaryotic genomes. Alternate hypotheses have been suggested to explain this co-occurrence. “The regional difference hypothesis” (Hardison *et al.*, 2003; Silva and Kondrashov, 2002) suggests that some genomic regions are specifically pre-disposed to be the mutational hotspots. A second hypothesis “the indel induced hypothesis” considers that error prone DNA polymerases are recruited to repair DNA damage in Insertion-Deletion (InDels) regions, inducing nucleotide substitutions (SNPs) during the repair process (Tian *et al.*, 2008; Zhu *et al.*, 2009). This hypothesis is supported by the observations of strong association between InDels and SNPs in eukaryotic genomes including mammals, fruit fly, primates, rodent, yeast, rice and *Arabidopsis* (Chen *et al.*, 2009; Longman-Jacobsen *et al.*, 2003; Tian *et al.*, 2008; Yang *et al.*, 2009; Zhang *et al.*, 2008) and prokaryotic genomes (Zhu *et al.*, 2009). McDonald *et al.* (2011) reported the associations among oligonucleotide repeats, InDels and substitutions in the genomes of prokaryotic and eukaryotic species (*Escherichia coli*, *Saccharomyces paradoxus* and *Drosophila*). They hypothesized that the oligonucleotide repeats are the reason for inducing the InDels and SNPs, rather than the InDels *per se*, putting more emphasis on the “regional difference hypothesis” rather than the “indel induced hypothesis”. The high-fidelity DNA polymerase are stalled in regions with higher repeat frequencies, thereby error-prone DNA polymerases are recruited to repair the DNA damage, and the adjacent sequences are replicated at higher than average error rate (McDonald *et al.*, 2011).

Chloroplast genome is prokaryotic in origin (Palmer, 1985). Associations among mutational events including repeats, InDels, substitutions and inversions have been reported for the specific genes or loci in different lineages (Graham *et al.*, 2000; Lockhart *et al.*, 2001; McLenachan *et al.*, 2000; Mes *et al.*, 2000). Some other studies also reported the role of repeats

in generation of InDels (Kawata *et al.*, 1997) and inversions (Kim and Lee, 2005; Whitlock *et al.*, 2010). All these findings were based on the analyses of some specific loci rather than based on the analyses of complete chloroplast genomes. Ahmed *et al.* (2012) systematically studied the extent of correlations among oligonucleotide repeats, InDels and SNPs in pair-wise alignments, as well as SNPs as a function of distance from InDel location points in multiple sequence alignment in aroid chloroplast genomes. They reported that mutually non-exclusive, multiple hypotheses explain the mutational dynamics in chloroplast genomes. Ahmed *et al.* (2012) also suggested that the distribution of oligonucleotide repeats could be used as a proxy for mutational hotspots. Following the methodology proposed by Ahmed *et al.* (2012), correlations were successively reported in genus *Cephalotaxus* of family Cephalotaxaceae, gymnosperm (Yi *et al.*, 2013). We aimed to explore such correlations in the widely distributed and diverse species of family Malvaceae, eudicot (angiosperms).

The aim was to investigate correlations among three types of mutational events in eudicot plant family Malvaceae. We investigated the correlations by selecting species from 14 genera across Malvaceae. Our findings revealed weak to strong correlations with high regression. Hence, we hypothesize that such correlations among the mutations might be a common characteristic of chloroplast genomes in all plant lineages and repeats could be used as proxy to identify mutational hotspots.

4.2 Materials and methods

4.2.1 Correlations among substitutions, InDels and oligonucleotide repeats

We determined correlations among substitutions, InDels, and oligonucleotide repeats in chloroplast genomes of 18 species of family Malvaceae. We evaluated these mutational events in far diverse species of family level as well as in closely related species of genus level. Details for all the species included in these comparisons have been provided in chapter 3. Here, for genus level comparison, we also downloaded chloroplast genome sequences of three other species including *Firmiana pulcherrima* (NC_036395), *Gossypium barbadense* (HQ901198), *Tilia mandshurica* (KT894773), and *Theobroma grandiflorum* (JQ228388) from NCBI.

The correlations among substitutions, InDels, and oligonucleotide repeats were determined by following Ahmed *et al.* (2012) with some modifications. We performed comparisons at genus and family levels. For the family level comparisons, one species from each genus was pair-wise aligned with *Theobroma cacao* chloroplast genome (basal to Malvaceae) using MAFFT alignment (Kato *et al.*, 2005). For the genus level comparison, the species of five genera were pairwise aligned by randomly selecting one of the two species of that genus as a reference

genome. In this way, *Firmiana major*, *Gossypium herbaceum*, *Hibiscus mutabilis*, *Tilia amurensis*, and *Theobroma cacao* were used as references for their comparisons with *Firmiana pulcherrima*, *Gossypium barbadense*, *Hibiscus syriacus*, *Tilia mandshurica*, and *Theobroma grandiflorum*, respectively. Genes located close to the junctions of LSC/IR and SSC/IR were removed from the alignment to avoid rate heterotachy (Wu *et al.*, 2011). In addition, all the inversions from the alignments were removed in order to reduce noise in the alignment, and to avoid false positive results.

All the deletions in the reference genome were removed by deleting that specific portion of the pairwise alignment after noting their positions, as these deletions were counted in the final analyses. The removal of deletions from reference genome enabled us to fix the coordinate positions of oligonucleotide repeats for further analyses. Alignment of complete chloroplast genome including regions of tRNAs, rRNAs, protein coding sequences, intronic regions as well as intergenic spacer regions were used in the correlation's analyses. Each pairwise alignment of complete chloroplast genome was divided into mutually exclusive bins of 250 bp each to count number of InDels and SNPs. Total number of bins ranged from 433–500 for all alignments.

We counted the InDels including the previously removed deletions from the reference genome and insertions in the reference genome. All the InDels were differentiated into SSRs and non-SSRs InDels and were manually allocated to each bin. The criteria that were considered for SSRs included: 7 repeat units for mono-, 4 for di-, and 3 for tri-, tetra-, penta-, and hexanucleotide SSRs.

We calculated and allocated substitutions to each bin by using a custom-made Perl script. Oligonucleotide repeats were counted only in the reference genomes for family and genus level comparisons using REPuter (Kurtz *et al.*, 2001). The repeats (forward and reverse) with minimum size of 14 bp and no mismatch between the two copies of a repeat were included in the comparisons. Oligonucleotide repeats were assigned to their respective bins using a custom-made python script. We also showed the procedure of SNPs, InDels, and repeats counting in the bins in Figure. 4.1. Normality tests on our data of substitutions, InDels and oligonucleotide repeats revealed that the data did not follow normal distribution. Therefore, Spearman's Rho correlations were calculated among substitutions, InDels, and oligonucleotide repeats using Minitab v. 18. We also determined the regression of “SNPs on InDels”, “SNPs on repeats”, “InDels on repeats” in these comparisons. Strengths of the correlations were expressed following Akoglu (2018), as follows: negligible or very weak (0.10-0.19), weak (0.20-0.29), moderate (0.30-0.39), strong (0.4-0.69), very strong (0.70-0.99), and perfect (1.0). The probability (*p*) of significance of correlations was tested at 0.05 α -level.

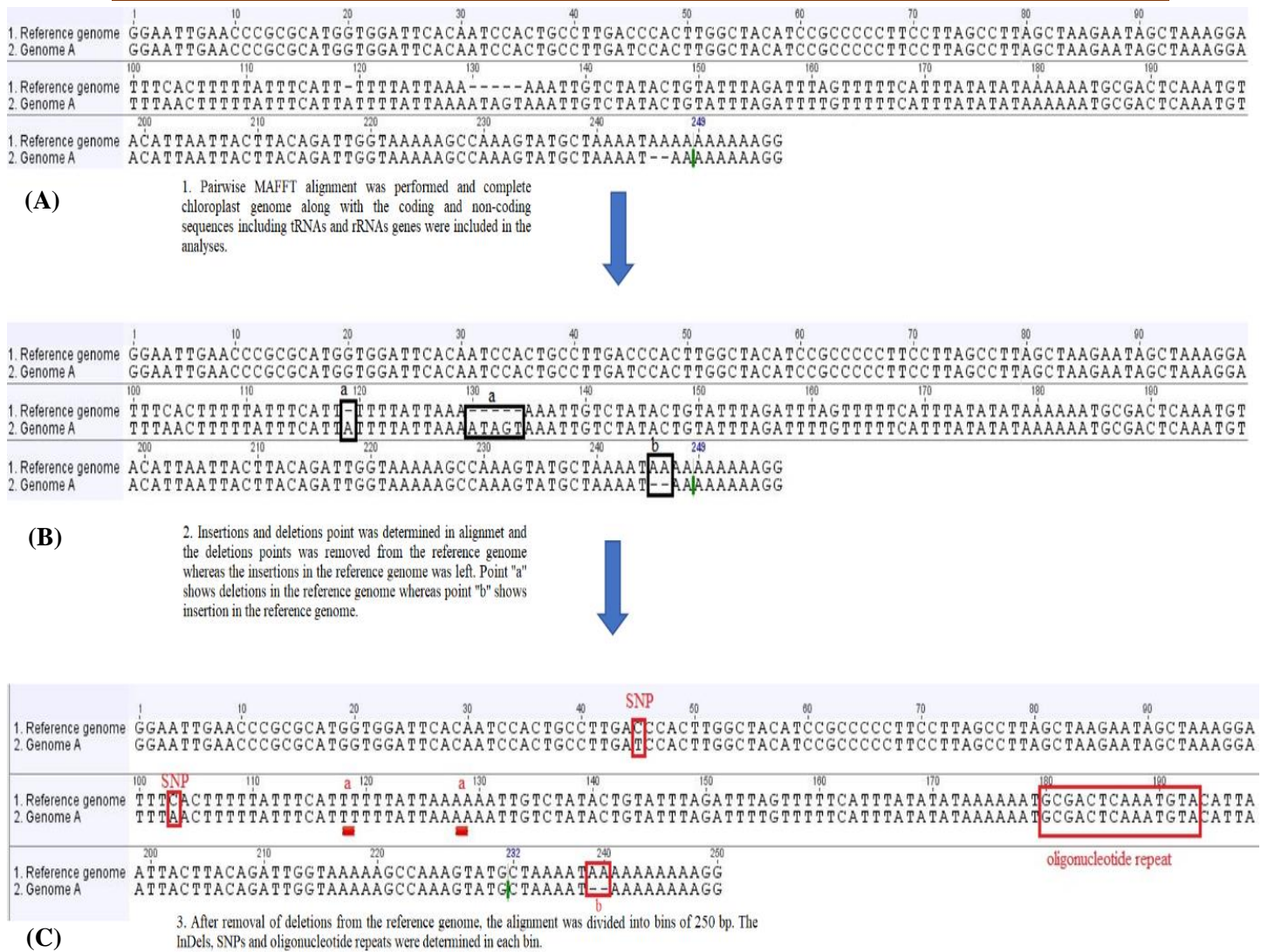


Figure 4.1 Methodology of SNPs, InDels and repeats counting for correlation. (A) Two genomes were aligned by considering one genome as reference. For family level comparison, *Theobroma cacao* was considered as reference whereas for genus level comparison the species that aligned to *T. cacao* was considered as reference for another member of the genus. (B) The deletions events that existed in reference genome were removed from the alignment as shown as point “a” whereas insertions events of reference genome were left in the alignment as shown as point “b”. (C) The complete chloroplast genomes were divided into multiple bins of 250 bp each after removal of deletions of reference genome. The number of InDels were counted manually whereas the number of substitutions were determined and allocated by custom made perl script and number of repeats were determined only in the reference genome and allocated to each bin by custom made Python script. The bins which has been shown in figure contains two SNPs, three InDels, and one oligonucleotide repeats. The same method was applied throughout the genome.

4.3 Results

4.3.1 Correlations among substitutions, InDels and repeats

In total, 18 pairwise alignments were used to find correlations among oligonucleotide repeats, InDels and substitutions. Among these, 13 alignments were used to find correlations at family level (using *Theobroma cacao* as reference), and five alignments were used to calculate correlations at genus level.

4.3.1.1 Correlations and regressions among mutational events at family level

At family level, our analyses revealed fluctuations in correlations at different levels of comparison. The SNPs with InDels revealed strong correlations in nine species, moderate correlations in two species, weak correlation in one species, and negligible or very weak correlations in one species. These all correlations values were statistically very significant ($p < 0.001$). The correlation between oligonucleotide repeats and InDels was strong in five species, moderate in six species, and weak in two species. The lowest correlations were found between SNPs and repeats as compared to “SNP and InDels” and “InDels and repeats”. So, substitutions and repeats revealed weak correlations in eight species and very weak or negligible correlations in five species (Figure 4.2A). All these correlations were observed with high significance. The average of correlations was stronger between “SNPs and InDels” followed by “InDels and repeats” and then by “SNPs and repeats”. The average value of correlations between SNPs and InDels was 0.435, between InDels and repeats was 0.359, and between SNPs and repeats was 0.212. At family level comparison, within the species, we observed highest correlations for “substitutions and InDels”, followed by “repeats and InDels” and then by “repeats and substitutions” except for *Tilia amurensis* and for *Gossypium herbaceum* for which correlations of “repeats and InDels” were stronger than the other two comparisons, as given in Figure 4.2A.

We determined regression value to evaluate role of InDels and repeats in generation of SNPs, and the role of repeats in generation of InDels (Fig. 4.2B). The average regression value was highest for “SNPs on InDels” with 20.87%, followed by “InDels on repeats” with 11.02% and then by “SNPs on repeats” with 5.19%. The regression values ranged from 1.58%-37.9% for “SNPs on InDels”, 1.93%-19.03% for “InDels on repeats”, and from 0.4%-12.13% for “SNPs on repeats”. The observation of high regression value confirms the role of InDels in generation of SNPs whereas the role of repeats was confirmed in the generation of SNPs and InDels. compared to “SNPs on SSR InDels”. The average value of regression was found 21.28% for “SNPs on non-SSR InDels” whereas the regression value of 2.66% was observed “SNPs on SSR InDels”.

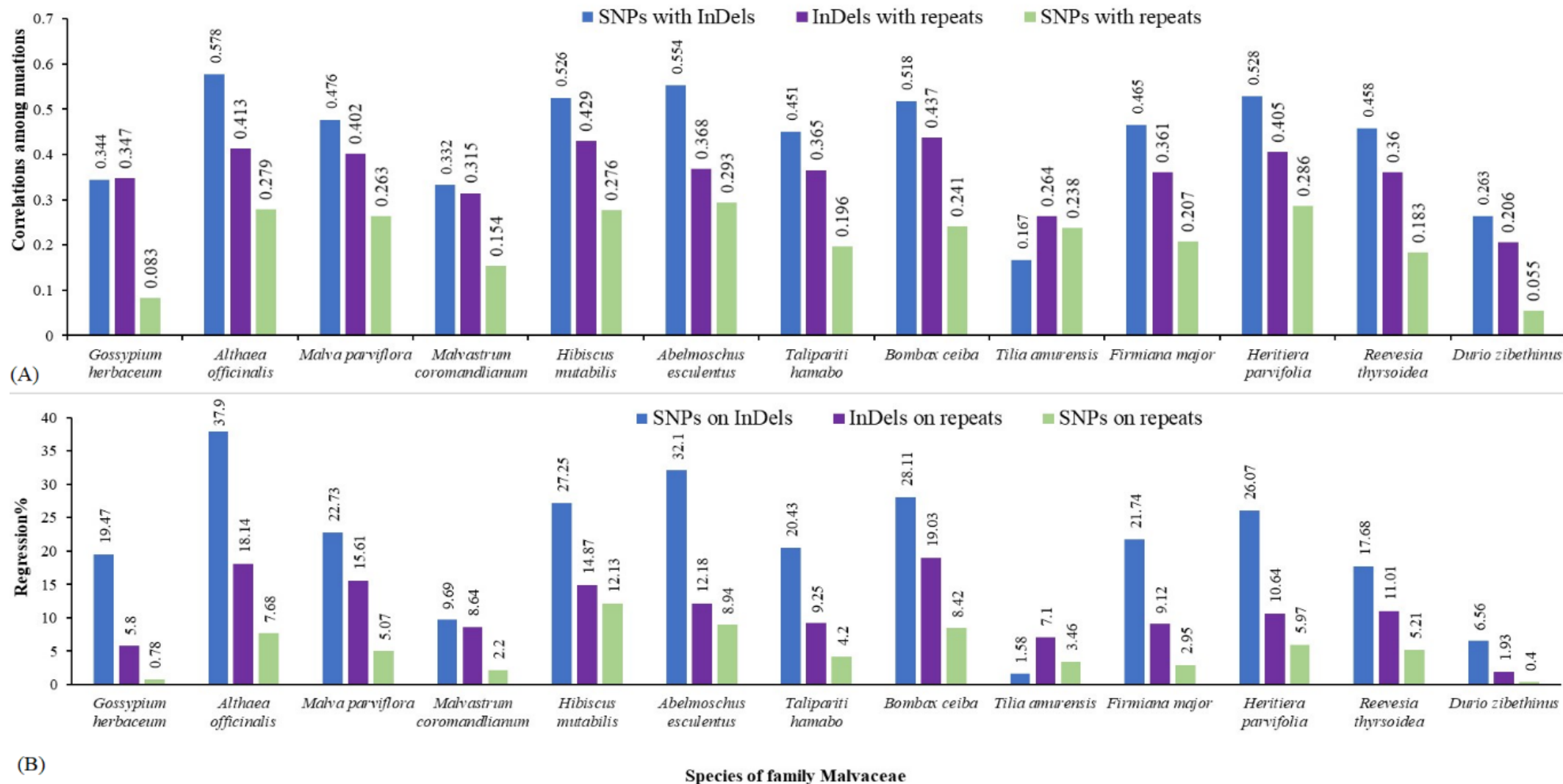


Figure 4.2 Represents the correlations and regression values at family level comparison. **(A)** Correlations among mutational events. The Y-axis shows the correlations whereas X-axis shows the species of family Malvaceae to which the correlations belong. **(B)** Regression up to which the InDels regress SNPs and repeats regress both InDels and SNPs.

4.3.1.2 Correlations and regressions among mutational events at genus level

Correlations at genus level also exhibited variations. We observed moderate correlations between SNPs and InDels in two species, and weak correlations in three species. The correlations between InDels and repeats were found strong in two species, moderate in two species and very weak in one species. The correlations between SNPs and repeats were found moderate in one species, weak in three species, and very weak or negligible in one species. Comparisons at genus level, within the species, showed differences in correlations among three types of mutational events. In *Firmiana*, *Gossypium* and *Hibiscus*, we observed strong correlations for “repeats and InDels” followed by “substitutions and InDels” and then by “repeats and substitutions”. The genus *Tilia* showed strong correlation for “repeats and InDels” followed by “repeats and substitutions” and “substitutions and InDels”. For the genus *Theobroma*, we observed strongest correlation for “repeat and SNPs” followed by “InDels and substitutions” and then by “InDels and repeats”. The average value of all genus level correlations revealed strong correlation between “InDels and repeats” with 0.357 followed by “SNPs and InDels” with 0.288, and then by “SNPs and repeats” with 0.254 (Figure 4.3).

The regression values also support the result of correlations. The average value of regression remained 17.27% for “InDels on repeats”, followed by 10.97% for “SNPs on InDels”, and 8.05% for “SNPs on repeats”. The regression value ranged from 3.3% to 35.89% for “InDels on repeats” 5.72%-15.32% for “SNPs on InDels”, and 0.22%-23.39% for “SNPs on repeats”. These analyses showed the role of InDels in generation of SNPs as well as the role of repeats in generation of both InDels and SNPs (Figure 4.3).

4.3.2 Correlations and regression analyses of substitutions with SSR and non-SSR InDels

Since SSR and non-SSR indels differ in their mode of generation. Therefore, we also separately studied correlations of substitutions and repeats with SSR and non-SSR InDels.

At family level, the analyses revealed weak correlations between SNPs and SSR InDels. We found weak correlations in four species, and very weak to negligible correlations in nine species. The average value of correlation was 0.172. Except three species, all these observations were highly significant having $p \leq 0.001$ (Table 4.1).

The analyses of correlations between SNPs and non-SSR InDels revealed strong correlation with the average correlation value of 0.431. The correlations were found strong in nine species, moderate in two species, weak in one species, and very weak or negligible in one species. These observations were strongly significant at $p < 0.001$ except one species, *Tilia amurensis*, for which $p = 0.001$. Regression values were also stronger in “SNP on non-SSR InDels” as

At genus level comparison, average value of correlation and regression between SNPs on non-SSRs InDels” remained 0.21% and 11.24%, respectively, about half of the values at the family level comparison. The correlation between SNPs and SSR InDels were found weak for 2 species and very weak for 3 species, whereas the correlation between SNPs and non-SSRs InDels were found moderate between 2 species and very weak or negligible in 3 species.

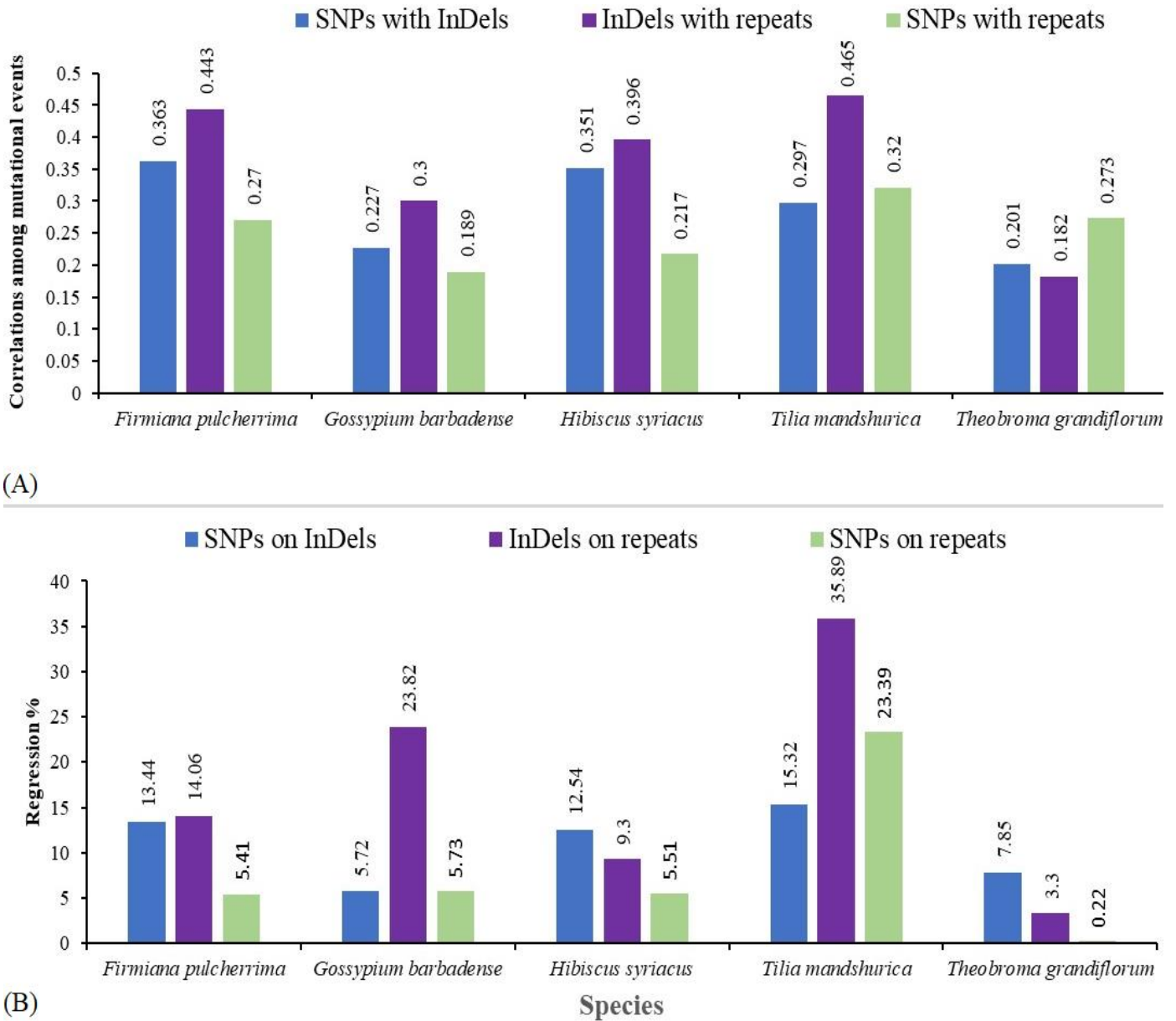


Figure 4.3 Represents the correlations and regression values at genus level comparison

(A) Represents the correlations among mutational events of two species of the same genus. The X-axis shows the species in which the correlations have been evaluated whereas the Y-axis shows the extent of correlations. (B) Represents the extent of regression up to which the InDels regress SNPs and repeats regress both InDels and SNPs at genus level comparison.

Table 4.1 Correlations and regression of SNPs with SSR and non-SSR InDels

Species	SNPs with SSRs InDels			SNPs with Non-SSR InDels		
	r	r ²	p	r	r ²	p
Comparison of correlations among mutational events at family level						
<i>Abelmoschus esculentus</i>	0.164	2.47	<0.001	0.569	33.30	<0.001
<i>Althaea officinalis</i>	0.125	1.75	0.007	0.595	39.15	<0.001
<i>Bombax ceiba</i>	0.2	2.98	<0.001	0.509	28.43	<0.001
<i>Durio zibethinus</i>	0.104	0.63	0.028	0.275	8.42	<0.001
<i>Firmiana major</i>	0.154	1.26	0.001	0.469	23.54	<0.001
<i>Gossypium herbaceum</i>	0.078	1.55	0.105	0.359	20.84	<0.001
<i>Heritiera parvifolia</i>	0.252	5.68	<0.001	0.510	24.93	<0.001
<i>Hibiscus mutabilis</i>	0.173	2.03	<0.001	0.497	29.25	<0.001
<i>Reevesia thyrsoides</i>	0.208	3.10	<0.001	0.444	17.11	<0.001
<i>Talipariti hamabo</i>	0.152	2.50	0.001	0.445	21.08	<0.001
<i>Tilia amurensis</i>	0.176	1.11	<0.001	0.154	0.98	0.001
<i>Malvastrum coromandelianum</i>	0.186	2.84	<0.001	0.304	7.65	<0.001
<i>Malva parviflora</i>	0.262	6.62	<0.001	0.473	22.0	<0.001
Average	0.172	2.66		0.431	21.28	
Comparison of correlations among mutational events at genus level						
<i>Firmiana pulcherrima</i>	0.228	4.14	<0.001	0.307	11.42	<0.001
<i>Gossypium barbadense</i>	0.166	1.52	<0.001	0.151	4.75	0.001
<i>Hibiscus syriacus</i>	0.177	2.99	<0.001	0.336	11.63	<0.001
<i>Tilia mandshurica</i>	0.240	2.86	<0.001	0.138	14.49	0.002
<i>Theobroma grandiflorum</i>	0.143	0.14	0.001	0.138	14.97	0.002
Average	0.191	2.33		0.214	11.45	

Strength of correlation: negligible or very weak (0.1-0.19), weak (0.20-0.29), moderate (0.30-0.39), strong (0.4-0.69)

4.3.3 Correlations and regression analyses of repeats with SSR and non-SSR InDels

The correlations between repeats and non-SSRs InDels were stronger as compared to repeats and SSR InDels. The correlations between repeats and non-SSR InDels were found moderate in eight species, weak in four species and very weak in one species (Table 4.2). The average correlation between repeats and non-SSR InDels remained moderate at 0.316. The correlation between repeats and SSR InDels were found moderate in three species, weak in six species and very weak in four species (Table 4.2). The average correlation between repeats and SSR InDels remained weak at 0.236. The average regression value between repeats and SSR InDels was 5.52%, whereas regression value between repeats and non-SSR InDels remained 8.4%. Except for one comparison, the observations were highly significant at $p < 0.001$.

At genus level comparison, between repeats and SSR InDels, three species showed moderate correlation, one species showed weak correlations and one species showed very weak

correlations. Average correlation between repeats and SSRs InDels remained 0.277, slightly higher than that in family level comparisons. On the other hand, for repeats and non-SSRs InDels, two species showed moderate correlations, two species showed weak correlations and one species showed very weak correlations (Table 4.2). Average of correlation in this case remained 0.259, slightly lower than that observed in family level comparisons. Interestingly, the analyses of regression showed inverse results in comparison to correlations and we found the higher regression value of 11.05% for “non-SSR InDels on repeats” as compared to that of 8.12% for “SSR InDels on repeats” (Table 4.2).

Table 4.2 Correlations and regression of SSR and non-SSR InDels with repeats

Species	SSR InDels with Repeats			Non-SSR InDels with Repeats		
	r	r ²	p	r	r ²	p
Comparison of correlations among mutational events at family level						
<i>Abelmoschus esculentus</i>	0.203	3.89	<0.001	0.354	10.52	<0.001
<i>Althaea officinalis</i>	0.161	1.64	<0.001	0.391	18.34	<0.001
<i>Bombax ceiba</i>	0.302	14.32	<0.001	0.39	12.32	<0.001
<i>Durio zibethinus</i>	0.175	1.83	<0.001	0.137	0.63	0.004
<i>Firmiana major</i>	0.234	4.61	<0.001	0.331	6.19	<0.001
<i>Gossypium herbaceum</i>	0.303	11.67	<0.001	0.265	1.79	<0.001
<i>Heritiera parvifolia</i>	0.275	3.76	<0.001	0.359	8.94	<0.001
<i>Hibiscus mutabilis</i>	0.276	5.09	<0.001	0.369	12.31	<0.001
<i>Reevesia thyrsoidea</i>	0.250	4.22	<0.001	0.277	8.19	<0.001
<i>Talipariti hamabo</i>	0.194	2.10	<0.001	0.342	8.30	<0.001
<i>Tilia amurensis</i>	0.171	3.17	<0.001	0.253	5.48	<0.001
<i>Malvastrum coromandelianum</i>	0.218	3.05	<0.001	0.279	6.44	<0.001
<i>Malva parviflora</i>	0.3	12.42	<0.001	0.355	9.80	<0.001
Average	0.236	5.52		0.316	8.40	
Comparison of correlations among mutational events at genus level						
<i>Firmiana pulcherrima</i>	0.363	12.96	<0.001	0.310	4.24	<0.001
<i>Gossypium barbadense</i>	0.224	5.76	<0.001	0.245	20.47	<0.001
<i>Hibiscus syriacus</i>	0.315	5.77	<0.001	0.331	5.78	<0.001
<i>Tilia mandshurica</i>	0.346	15.21	<0.001	0.295	21.98	<0.001
<i>Theobroma grandiflorum</i>	0.138	0.90	0.002	0.115	2.77	0.011
Average	0.277	8.12		0.259	11.05	

Strength of correlation: negligible or very weak (0.1-0.19), weak (0.20-0.29), moderate (0.30-0.39), strong (0.4-0.69)

4.4 Conclusion

We determined very weak to strong correlations among substitutions, InDels and repeats in the plant family Malvaceae eudicot, angiosperms. Since, such observations were also observed previously in monocots (angiosperms) and gymnosperms, we hypothesize that this might be the common phenomenon for all plant lineages. Further confirmation in other families could suggest the change of well-known model of phylogeny inferring, the GTR model.

Chapter 5

Comparative analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*

5.1 Introduction

The Genus *Theobroma* L. belongs to subfamily Byttnerioideae (Malvaceae). According to “The plant list” (<http://www.theplantlist.org/>: accessed on 10 January 2019), the genus *Theobroma* consists of 91 taxa in which 51 are considered synonymous, 25 are accepted as species and 15 are unresolved. Family Malvaceae is basal to genus *Theobroma* (Richardson *et al.*, 2015). This evergreen crop is native of the South American rainforest (Bartley, 2005). Cacao tree (*Theobroma cacao*) and Cupuassu tree (*Theobroma grandiflorum*) are the economically important species of genus *Theobroma* and native to Brazil (Cuatrecasas 1964). Cacao is grown in about fifty countries throughout the humid tropic region (Motamayor *et al.* 2013) and their enclosed seeds within pods (fruits) are used in confectionary, chocolate production and cosmetics. These are also used in ice creams, juices, chocolate-like product (cupulate), , candies, desserts, yogurt, liquor and domestic jellies and jams (Cavalcante 1991). Previously, several molecular studies evaluated genetic diversity, nuclear genome structure, domestication, phylogeny and ultra-barcoding of *Theobroma cacao* and *Theobroma grandiflorum* (Alves *et al.*, 2007; Kane *et al.*, 2012; Richardson *et al.*, 2015). The studies that accessed the genetic diversity mostly used the restriction fragment length polymorphisms (RFLPs), SNPs (single nucleotide polymorphisms) and simple sequence repeats (SSRs) of the nuclear genome (Gopaulchan *et al.*, 2019; Kim *et al.*, 2017; Motamayor *et al.*, 2008, 2002; Osorio-Guarín *et al.*, 2017; Thomas *et al.*, 2012). These studies aimed at conservation of the plants, identification of suitable cultivars for breeding purposes and to evaluate the domestication process of the Cacao.

Some studies also used chloroplast genome-based markers for the genetic diversity, population genetics and phylogenetic studies of Cacao. Yang *et al.* (2011) designed nine SSRs markers and assessed diversity of 95 hybrids of *Theobroma cacao*. In other studies, SNPs-based markers of chloroplast regions *trnH-psbA* has been used for identification of haplotypes of Cacao (Gutiérrez-López *et al.*, 2016), whereas the sequence of *ndhF* has been used for the phylogenetics and for the determination of time of divergence of genus *Theobroma*, specifically Cacao (Richardson *et al.*, 2015). Kane *et al.* (2012) analysed chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum* for ultra-barcoding and suggested five barcoding regions. However, the authors did not perform comparative chloroplast structure analysis such as SSRs, oligonucleotide repeats, putative RNA editing site analysis, codon usage and amino acid frequency. None of the studies carried out detailed comparative analysis of chloroplast genomes of *T. cacao* and *T. grandiflorum* to resolve taxonomic discrepancies that can play a significant role in inferring phylogeny.

Chloroplast genome exhibits uniparental inheritance and is slow evolving compared to nuclear genome (Palmer 1985), which makes it highly valuable for phylogenetics, plant barcoding and species identification (Daniell *et al.*, 2016). According to Daniell *et al.* (2016), success of the breeding program is dependent on the selection of genetically compatible species and chloroplast genome can serve as valuable tool for the identification of genetically compatible and closely related species. The efforts of understating of genetic relationships between cultivated crops and their wild relative help to introduce specific advantageous traits in the cultivated crops. Markers based on complete chloroplast genome sequence can be authentic, cost-effective and robust (Ahmed *et al.*, 2013; Nguyen *et al.*, 2018). Recently, many researchers focused on identification of suitable polymorphic loci based on genomic data for development of authentic, suitable and cost effective markers for inferring of phylogeny (Menezes *et al.*, 2018; Yu *et al.*, 2019).

In the current study, we aimed to carry out comparative structural analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum* to get broad insights into the evolutionary pattern of the species that exist at basal to Malvaceae. We identified polymorphic loci that might be helpful to design molecular markers for resolving phylogeny of the genus *Theobroma*. This could also help in selection of genetically compatible and closely related species for the breeding purposes to introduce beneficial characters in the cultivated crops of *Theobroma cacao* and *Theobroma grandiflorum* for high production and high resistance against disease causing pathogens of these species.

5.2 Materials and Methods

5.2.1 Reannotation and comparative analyses of chloroplast genomes

The chloroplast genome of *Theobroma cacao* (HQ336404) and *Theobroma grandiflorum* (JQ228388) were downloaded from NCBI. Dual Organellar Genome Annotator (DOGMA) (Wyman *et al.*, 2004) and GeSeq (Tillich *et al.*, 2017) were used to check and correct annotation errors, as published genomes are prone to errors (Amiryousefi *et al.*, 2018). Codon usage and amino acid frequency were analysed with Geneious R8.1 (Kearse *et al.*, 2012). The putative RNA editing sites in the coding sequences were predicted by predictive RNA editor for plants (PREP) suite (Mower, 2009).

5.2.2 Repeats analysis in *Theobroma* chloroplast genomes

SSRs were identified with the online available software MicroSATellite (MISA) (Thiel *et al.*, 2003) by setting the values 10, 5, 4, 3, 3, 3 for mono-, di-, tri-, tetra-, penta- and hexa- SSR motifs, respectively. The REPuter program (Kurtz *et al.*, 2001) was used to find forward (F),

palindromic (P), reverse (R) and complement (C) oligonucleotide repeats with the minimum repeat size of 30 bp, edit distance 3, and maximum computing repeats 500.

5.2.3 Substitutions and InDels analysis

Substitutions and InDels were determined among the two species by paired-end MAFFT (Multiple Alignment using Fast Fourier Transform) alignment (Kato *et al.*, 2005) extension in Geneious R8.1. The numbers of substitutions and InDels were counted in chloroplast genomes manually and exact location within the genome was noted.

5.3 Results

5.3.1 Genome organisation and features of *Theobroma* species

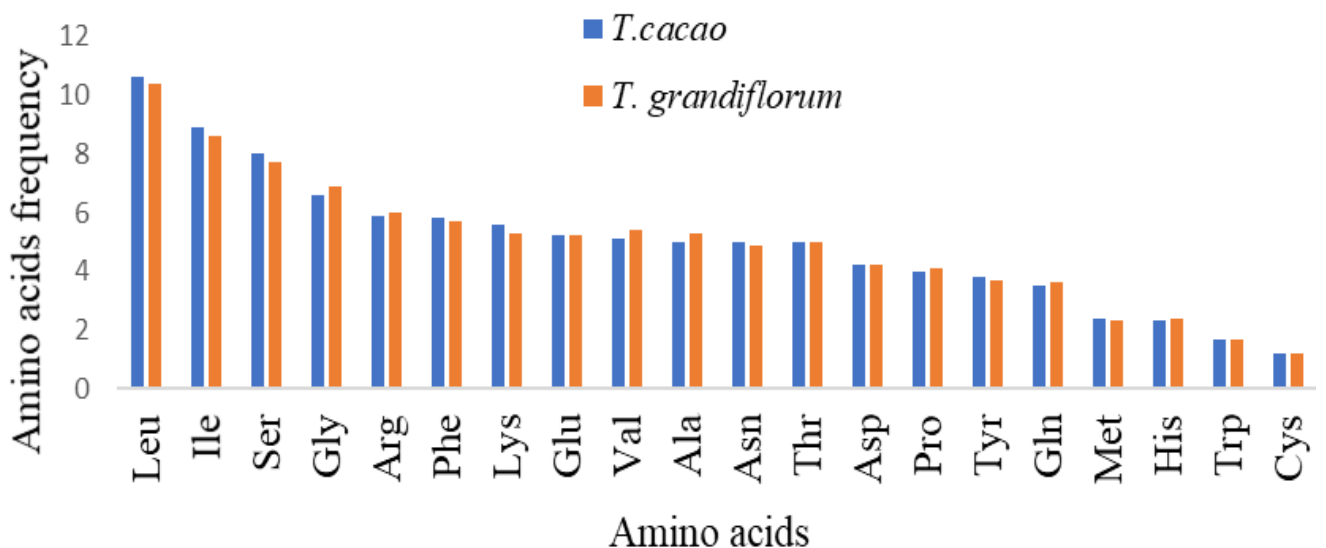
After the correction of errors in annotations, *T. cacao* and *T. grandiflorum* showed similar gene structure and organisation. Both exhibited the same gene content with 113 unique genes that included 30 tRNA, 4 rRNA, and 79 protein-coding genes. Among these, 17 genes were duplicated in the IR region of the genome (Table 5.1) comprising six protein-coding genes, four (rRNA) and seven tRNA. The only difference observed in the genome was for *infA* that was a pseudogene in *T. grandiflorum* while functional in *T. cacao*. The LSC, IR, and SSC had similar lengths with a slight difference in size of 62 bp, 35 bp, and 7 bp, respectively. The GC content was almost similar between both species but varied within chloroplast regions. The maximum GC content of IR was contributed by tRNA and rRNA that had GC content up to 55.5% and 53%, respectively (Table 5.1).

5.3.2 Amino acids frequency and codon usage

We analysed amino acid frequency and relative synonymous codon usage (RSCU). The protein-coding regions consisted of 78,852 bp (26,284 codons) in *T. cacao* and 78,651 bp (26,217 codons) in *T. grandiflorum*. Amino acid frequency and codon usage showed a high level of similarities. The leucine and isoleucine were the most abundant amino acids and had frequency of 10.6% and 8.9% in *T. cacao* and 10.4% and 8.6% in *T. grandiflorum*, respectively. The cysteine had the lowest frequency that was 1.2% (Figure 5.1). The RSCU (Relative synonymous codon usage) value of the codons revealed that codons having A/T at 3' end were in abundance. Except three codons, all other codons had RSCU value more than 1 whereas the codons having C/G at 3' end showed value less than 1 (Table 5.2).

Table 5.1 General features and comparison of the chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*

Characteristics		<i>Theobroma cacao</i>	<i>Theobroma grandiflorum</i>
Size (base pair; bp)		160,604	160,619
LSC length (bp)		89,395	89,333
SSC length (bp)		20,187	20,194
IR length (bp)		25,511	25,546
Number of genes		130	130
Protein-coding genes		85	85
tRNA genes		37	37
rRNA genes		8	8
Duplicate genes		17	17
GC content	Total (%)	36.9%	36.8%
	LSC (%)	34.7%	34.7%
	SSC (%)	31.2%	31.1%
	IR (%)	43%	42.9%
	CDS (%)	37.9%	37.9%
	rRNA (%)	55.5%	55.5%
	tRNA (%)	53%	52.9%
	All gene %	39.5%	39.5%
Protein coding part (CDS) (%bp)		49.08%	48.96%
All gene (%bp)		69.65%	69.19%
Non-coding region (%bp)		30.35%	30.81%

**Figure 5.1 Frequency of amino acids in *T. cacao* and *T. grandiflorum***

The figure shows the amino acids on X-axis and their frequencies on Y-axis.

Table 5.2 Relative synonymous codon usage comparison of *Theobroma cacao* and *Theobroma grandiflorum*

Relative synonymous codon usage				Relative synonymous codon usage			
Codon	Amino acid	<i>Tc</i>	<i>Tg</i>	Codon	Amino acid	<i>Tc</i>	<i>Tg</i>
GCA	A	1.09	1.08	CCA	P	1.108	1.16
GCC	A	0.68	0.68	CCC	P	0.788	0.73
GCG	A	0.46	0.48	CCG	P	0.568	0.54
GCT	A	1.77	1.76	CCT	P	1.536	1.57
TGC	C	0.5	0.49	CAA	Q	1.528	1.54
TGT	C	1.51	1.51	CAG	Q	0.472	0.46
GAC	D	0.39	0.38	AGA	R	1.806	1.83
GAT	D	1.61	1.62	AGG	R	0.666	0.65
GAA	E	1.48	1.48	CGA	R	1.368	1.36
GAG	E	0.52	0.52	CGC	R	0.456	0.45
TTC	F	0.7	0.71	CGG	R	0.426	0.42
TTT	F	1.3	1.29	CGT	R	1.254	1.29
GGA	G	1.56	1.58	AGC	S	0.348	0.35
GGC	G	0.39	0.39	AGT	S	1.182	1.19
GGG	G	0.73	0.74	TCA	S	1.236	1.22
GGT	G	1.29	1.29	TCC	S	0.96	0.97
CAC	H	0.52	0.52	TCG	S	0.53	0.53
CAT	H	1.48	1.48	TCT	S	1.72	1.72
ATA	I	0.94	0.95	ACA	T	1.24	1.25
ATC	I	0.59	0.59	ACC	T	0.75	0.74
ATT	I	1.46	1.46	ACG	T	0.44	0.43
AAA	K	1.49	1.49	ACT	T	1.57	1.57
AAG	K	0.51	0.51	GTA	V	1.49	1.5
CTA	L	0.83	0.82	GTC	V	0.49	0.49
CTC	L	0.4	0.4	GTG	V	0.6	0.59
CTG	L	0.41	0.41	GTT	V	1.43	1.42
CTT	L	1.25	1.26	TGG	W	1	1
TTA	L	1.85	1.84	TAC	Y	0.38	0.39
TTG	L	1.26	1.26	TAT	Y	1.61	1.61
ATG	M	1	1	TAA	*	1.62	1.64
AAC	N	0.46	0.45	TAG	*	0.74	0.71
AAT	N	1.54	1.55	TGA	*	0.64	0.64

Tc: *Theobroma cacao*, *Tg*: *Theobroma grandiflorum*, * stop codon

5.3.3 Putative RNA editing sites

We observed 100% identity in putative RNA editing sites among *T. cacao* and *T. grandiflorum*. PREP predicted 57 putative RNA editing sites in 23 protein-coding genes. A high number of editing sites existed in *ndhB* (11), *ndhD* (6), *rpoB* and *ndhF* (4), and *matK* (3) (Table 5.3). The high level of conversion was found in codons of serine (S) with 47.36%; serine (S) to leucine (L): 43.86% and serine (S) to phenylalanine (F): 3.5 %. On second and third number, codon conversion was observed for histidine 12.28% and proline 10.52%, respectively. The nucleotide substitutions occurred at first and second nucleotide of codons and did not occur at third nucleotide of codons. Out of 57 RNA editing sites, 43 codons (68%) were substituted at the second nucleotide and 14 codons (32%) were substituted at the first nucleotide. All RNA editing sites except three lead to hydrophobic product including leucine, isoleucine, methionine, phenylalanine, tyrosine, tryptophan, and valine. Other three RNA editing sites converted the apolar amino acid proline (P) to polar amino acid serine (S).

Table 5.3 Putative RNA editing site in *Theobroma cacao* and *Theobroma grandiflorum*

Gene	Nucleotide Position	Amino acid Position	Codon conversion	Amino acid conversion	Score
<i>matK</i>	457	153	CAC > TAC	H > Y	1
	634	212	CAT > TAT	H > Y	1
	1237	413	CAC > TAC	H > Y	1
<i>atpA</i>	914	305	TCA > TTA	S > L	1
	1148	383	TCA > TTA	S > L	1
<i>atpF</i>	92	31	CCA > CTA	P > L	0.86
<i>Atpl</i>	76	26	CTC > TTC	L > F	0.86
	629	210	TCA > TTA	S > L	1
<i>rps2</i>	248	83	TCA > TTA	S > L	1
	325	109	CCC > TCC	P > S	1
<i>rpoC2</i>	2287	763	CGG > TGG	R > W	1
	3155	1052	CCC > CTC	P > L	0.86
<i>rpoC1</i>	41	14	TCA > TTA	S > L	1
	1261	421	CCG > TCG	P > S	0.86
<i>rpoB</i>	338	113	TCT > TTT	S > F	1
	551	184	TCA > TTA	S > L	1
	566	189	TCG > TTG	S > L	1
	2426	809	TCA > TTA	S > L	0.86
<i>rps14</i>	80	27	TCA > TTA	S > L	1
	149	50	TCA > TTA	S > L	1
<i>atpB</i>	403	135	CCC > TCC	P > S	0.86
<i>accD</i>	97	266	TCG > TTG	S > L	0.8

	1406	469	CCG > CTG	P > L	1
<i>psaI</i>	83	28	TCT > TTT	S > F	0.86
<i>psbF</i>	77	26	TCT > TTT	S > F	1
<i>rpl20</i>	308	103	TCA > TTA	S > L	0.86
<i>clpP</i>	559	187	CAT > TAT	H > Y	1
<i>petB</i>	418	140	CGG > TGG	R > W	1
<i>rpoA</i>	329	110	GCC > GTC	A > V	0.86
	830	277	TCA > TTA	S > L	1
<i>ndhB</i>	149	50	TCA > TTA	S > L	1
	467	156	CCA > CTA	P > L	1
	542	181	ACG > ATG	T > M	1
	586	196	CAT > TAT	H > Y	1
	611	204	TCA > TTA	S > L	0.8
	737	246	CCA > CTA	P > L	1
	746	249	TCT > TTT	S > F	1
	830	277	TCG > TTG	S > L	1
	836	279	TCA > TTA	S > L	1
	1255	419	CAT > TAT	H > Y	1
1481	494	CCA > CTA	P > L	1	
<i>ndhF</i>	290	97	TCA > TTA	S > L	1
	1570	524	CTT > TTT	L > F	1
	1859	620	ACA > ATA	T > I	0.8
	1925	642	GCA > GTA	A > V	0.8
<i>ccsA</i>	383	128	GCG > GTG	A > V	1
	644	215	ACT > ATT	T > I	0.86
<i>ndhD</i>	2	1	ACG > ATG	T > M	1
	383	128	TCA > TTA	S > L	1
	674	225	TCG > TTG	S > L	1
	878	293	TCA > TTA	S > L	1
	1298	433	TCA > TTA	S > L	0.8
	1310	437	TCA > TTA	S > L	0.8
<i>ndhG</i>	166	56	CAT > TAT	H > Y	0.8
	314	105	ACA > ATA	T > I	0.8
<i>ndhA</i>	341	114	TCA > TTA	S > L	1
	566	189	TCA > TTA	S > L	1

5.3.4 Analyses of simple sequence repeats and oligonucleotide repeats

The MISA result of SSRs revealed a similar evolutionary pattern. *T. cacao* having 96 SSRs, and *T. grandiflorum* having 98 SSRs. The mononucleotide SSRs (A/T) was abundant with a frequency of up to 65% (Figure 5.2A) whereas C/G comprised 3.8% SSRs. Polyadenine (poly A) comprised 21.5%, and polythymine (poly T) comprised 41.5% of mononucleotide SSRs whereas as the polycytosine (poly C) comprised 0.2%, and polyguanine (poly G) contained 0% SSRs. The dinucleotide and trinucleotide SSRs were also abundant and comprised about 21% of total SSRs. The AT/TA motifs were abundant and comprised about 87% of dinucleotide SSRs whereas the TC/TG comprised only 13%. The trinucleotide SSRs existed in four types of motifs: AAT, TTA, ATT and TAA. All SSRs loci varied in number of repeat units (Table 5.4). LSC contained 70% of SSRs, SSC 21.5% SSRs and IR 6% (Figure 5.2B).

Table 5.4 Simple sequence repeats in *T. cacao* and *T. grandiflorum*

SSRs in <i>T. cacao</i>													
Repeats	3	4	5	6	7	8	9	10	11	12	13	14	Total
A/T	-	-	-	-	142	85	51	35	15	6	4	1	339
C/G	-	-	-	-	12	4	2	1					19
AC/GT	-		1										1
AG/CT	-	11	1										12
AT/AT	-	32	7	4	1	1							45
AAC/GTT	6												6
AAG/CTT	16												16
AAT/ATT	22	6		1									29
ACC/GGT	1												1
ACT/AGT	3												3
AGC/CTG	5												5
ATC/ATG	1												1
AAAT/ATTT	6												6
AACT/AGTT	1												1
AATC/ATTG	1												1
AATG/ATTC	1												1
AAAGT/ACTTT			1										1
AAATG/ATTTT	1												1
AAATT/AATTT	1												1
AATAT/ATATT	1												1
	Total												490
SSRs in <i>T. grandiflorum</i>													
A/T	-	-	-	-	141	84	51	33	18	6	3	2	338
C/G	-	-	-	-	11	4	1	2					18
AC/GT	-		1										1
AG/CT	-	10	1										11

AT/AT	-	33	6	4	1	1							45
AAC/GTT	6												6
AAG/CTT	16												16
AAT/ATT	23	6		1									30
ACC/GGT	1												1
ACT/AGT	2												2
AGC/CTG	5												5
ATC/ATG	1												1
AAAG/CTTT	1												1
AAAT/ATTT	5												5
AACT/AGTT	1												1
AATC/ATTG	1												1
AATG/ATTC	1												1
AAAGT/ACTTT								1				1	
AAATG/ATTTT	1												1
AAATT/AATTT	1												1
AATAT/ATATT	1												1
	Total												486

REPuter screening determined four types of oligonucleotide repeats: Forward (F), Reverse (R), Palindromic (P), and Complementary (C) (Table 5.5). The repeats distribution in chloroplast showed that LSC region had maximum oligonucleotide repeats, followed by SSC and then IR regions (Figure 5.2C). *Theobroma cacao* contained 46 oligonucleotide repeats (F=21; R=13; P=11; C=1) and *T. grandiflorum* contained 53 oligonucleotide repeats (F=22; R=13; P=13; C=5) (Figure 5.2D). Oligonucleotide repeats size ranged from 30-58, and maximum numbers of oligonucleotide repeats existed in range of 30-34 (Figure 5.2E). Most of repeat pairs were found in the intergenic spacer region, 31 in *T. cacao* and 42 in *T. grandiflorum* (Figure 5.2F). The protein-coding genes *rbcL*, *ycf2*, *psaA* and *psaB* genes had repeats motifs in the genic region, whereas the *rps16*, *clpP*, *rpoC1*, *ycf3*, *ndhA* and *rpl16* had repeats motif in the intronic region (Table 5.5).

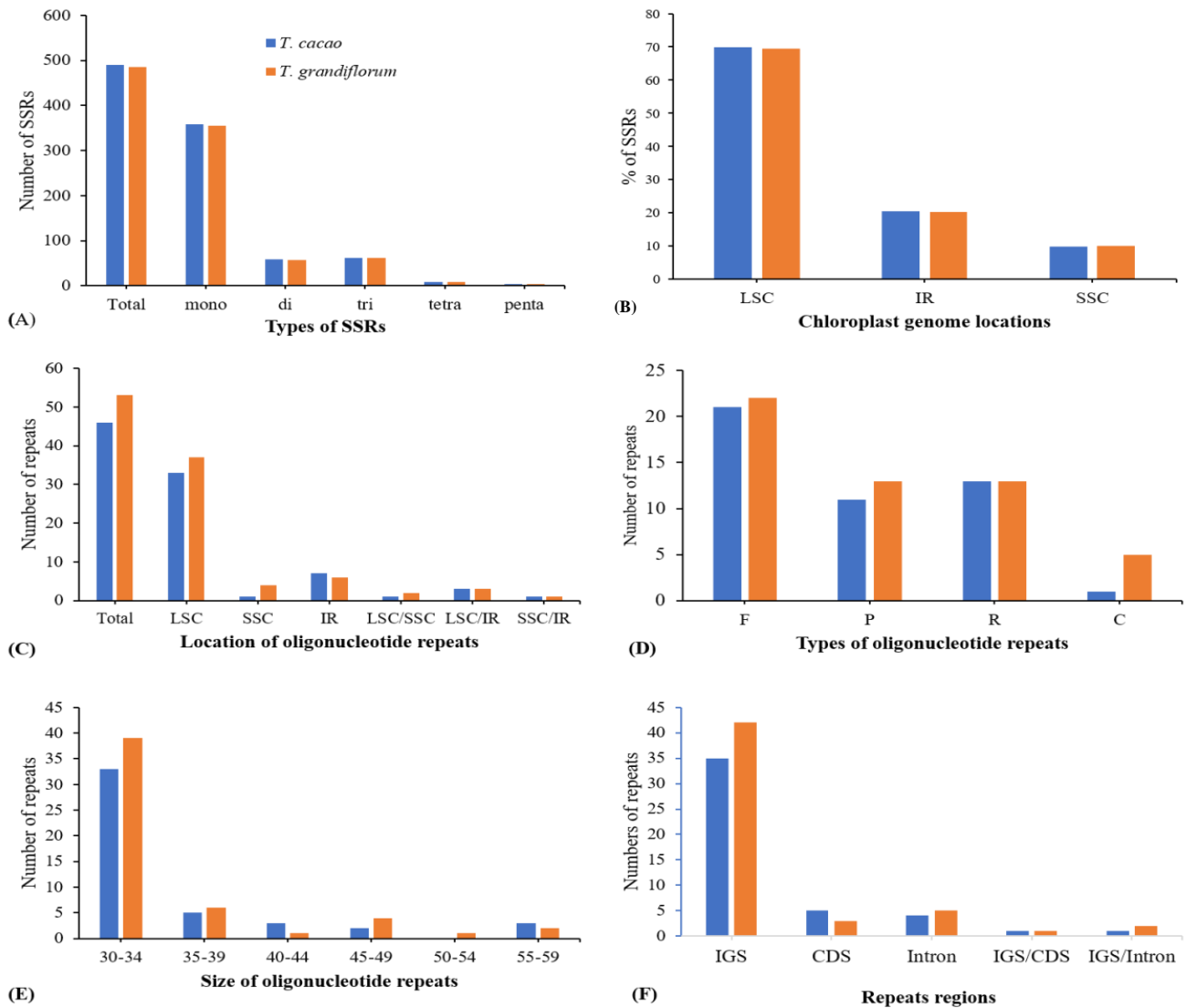


Figure 5.2 Comparison of repeats in *Theobroma cacao* and *Theobroma grandiflorum*

(A) Represent types of SSR by size. Total: all types of SSR present in the genome; mono: mononucleotide SSR; Di: dinucleotide SSR; tri: trinucleotide SSR; tetra: tetranucleotide SSR, and penta: pentanucleotide SSR. (B) Represents distribution of chloroplast genome in the large single copy (LSC), small single copy (SSC) and inverted repeat region (IR). (C) Distribution of oligonucleotide repeats motifs. LSC, SSC and IR stand for those oligonucleotides which are fully present in these regions whereas LSC/SSC, LSC/IR, and SSC/IR stand for those repeats which are dispersed in both locations, i.e. one copy of repeat existed in one location while the second copy was in another location. (D) Four types of oligonucleotide repeats exist in the genome. Forward (F), palindromic (P), reverse (R) and complementary (C). (E) The range of numbers shows the size of oligonucleotide repeats in that specific range. For instance, 30-34 stands for those oligonucleotide repeats which have the size from 30-34 bp. (F) Distribution of repeats based on function regions. IGS: Intergenic spacer regions, CDS: coding DNA sequences; Intron: Intronic regions.

Table 5.5 Oligonucleotide repeats in *Theobroma* species

S.No	Type	Location	Region	CDS ^a /IGS ^b / Intron ^c	Size	Sequence
Oligonucleotide repeats in <i>T. cacao</i>						
1.	P	LSC	<i>trnH-psbA/psbZ-trnG</i>	IGS	32	TTTTTTACTTTTTTTTTTATTTTCATTTTAT
2.	F	LSC	<i>trnS-trnG</i>	IGS	30	TTATATAGATATATACAACCTTTTATATAGATA
3.	P	LSC	<i>trnG-trnR</i>	IGS	58	TTTATTAATTCAATTCAATGATGCATTAATTAATGCATCATT GAATTGAATTAATAAA
4.	R	LSC	<i>trnR-atpA/psbZ-trnG</i>	IGS	31	TTATTATAATATAATTTTATTATTATTATTA
5.	p	LSC	<i>atpF-atpH</i>	IGS	38	TTTAATAATTAATAAATTATTAATTTATTATTATTA
6.	F	LSC	<i>atpF-atpH/psbZ-trnG</i>	IGS	31	TAATTAATAAATTATTAATTTATTATTATTA
7.	P	LSC	<i>rps2-rpoC2</i>	IGS	33	TCTCTTTTTTTTTTAAGTAAAAAAAAAAAAAGAGA
8.	F	LSC	<i>rpoC1/rpl16</i>	Intron	30	TCGGACATGAGAGTTTCCTCTCATCCGGCT
9.	P	LSC	<i>trnC-petN</i>	IGS	31	GTAGACACTCCACTACTAATGGAGTGTCTAC
10.	P	LSC/IR	<i>psbM-trnD/rps12-trnV</i>	IGS	31	AAAATAGAAAGGAAAAAAAAAGAAATAAAAAAAAA
11.	F	LSC	<i>trnT-psbD</i>	IGS	31	TATATGGATAATAATTATATGGATAATAATT
12.	R	LSC	<i>psbZ-trnG</i>	IGS		AAAGAAAATAATAAAATAGAAAATAATAAA
13.	R	LSC	<i>psbZ-trnG/rpl32-trnL</i>	IGS		ATTAATAAATAATATAGTAATAAATAATAT
14.	P	LSC	<i>psbZ-trnG</i>	IGS	34	AAATATATTAATATATATTTATATTAATATAAAT
15.	F	LSC	<i>psaB/psaA</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATATCAG TCAGCCATA
16.	F	LSC/SSC	<i>ycf3/ndhA</i>	Intron/	41	TCCAAAACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
17.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/IGS	36	AACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
18.	R	LSC	<i>trnT-trnL/atpB-rbcL</i>	IGS	30	TTTTAATATTTATATATTTATTTATATATT
19.	R	LSC	<i>trnT-trnL</i>	IGS	30	TTTAATATTTATATATTTATTTATATATTATA
20.	R	LSC	<i>trnT-trnL</i>	IGS	38	TTTAATATTTATATATTTATTTATATATTATATAATTT
21.	F	LSC	<i>trnT-trnL/ndhC-trnV</i>	IGS	31	ATATTTATATATTTATTTATATATTATATAA

22.	R	LSC	<i>trnT-trnL</i>	IGS	33	ATATTTATATATTTATTTATATATTATATAATT
23.	R	LSC	<i>trnT-trnL/atpB-rbcL</i>	IGS	31	TATTTATATATTTATTTATATATTATATAAT
24.	F	LSC	<i>trnT-trnR</i>	IGS	56	ATTTATATATTTATTTATATATTATATAATTTAATTTATATA ATTATTTATATATT
25.	F	LSC	<i>trnT-trnL/atpB-rbcL</i>	IGS	30	ATTTAATTTATATAATTATTTATATATTAT
26.	R	LSC	<i>trnF-ndhJ</i>	IGS	38	TTTATTTATTTTTATTTTTTATTTATTTTTATTTATTT
27.	P	LSC	<i>ndhC-trnV</i>	IGS	40	TATAGTAATAGTATAAATATATATTTATACTATTACTATA
28.	F	LSC/IR	<i>ndhC-trnV/rps12-trnV</i>	IGS	32	TTCTATTCTATTTTTTCTATTAGAATATATT
29.	F	LSC	<i>ndhC-trnV/atpB-rbcL</i>	IGS	30	TATATTTATATATTAATATATATATTATTT
30.	P	LSC	<i>ndhC-trnV</i>	IGS	30	ATATATATATTATTTAAATAATATATATAT
31.	C	LSC	<i>atpB-rbcL/petA-psbJ</i>	IGS	31	TTTATATTATAATATTTTATAATTTATATTA
32.	R	LSC	<i>atpB-rbcL</i>	IGS	30	ATATTATAATTATATATTTATTTATATATA
33.	F	LSC	<i>atpB-rbcL/rpl33- rps18</i>	IGS	30	ATATTTATTTATATATATTATTATATATAT
34.	R	LSC	<i>atpB-rbcL</i>	IGS	32	TTTATTTATATATATTATTATATATATTATTT
35.	F	LSC	<i>rbcL/rbcL-accD</i>	CDS/IGS	32	ATCAAATTTGAATTCGAAGCAATGGACTTTT
36.	R	LSC	<i>clpP</i>	Intron	30	TTTTTTTTCAAAAAAAAAAAGAAAGAAAAA
37.	F	IR	<i>ycf2</i>	CDS	31	TCTTTTTGTCCAAGTTACTTCTCTTTTTGTC
38.	F	IR	<i>ycf2</i>	CDS	56	CGATATTGATGATAGTGACGATATTGATGCTAGTGACGATA TTGATGCTAGTGACG
39.	F	IR	<i>ycf2</i>	CDS	34	TAGTGACGATATTGATGCTAGTGACGATATTGAT
40.	P	LSC	<i>psbT-psbN</i>	IGS	48	AATTGAAGTAATGAGCCTCCCAATATTGGGAGGCTCATTAC TTCAATT
41.	F	IR/SSC	<i>rps12-trnV/ndhA</i>	IGS/Intron	38	AACCGTACATGAGATTTTCACCTCATACGGCTCCTCGT
42.	R	IR	<i>rps12-16S rRNA/</i>	IGS		TATTATTTTTTTTCCCCCTCTTTTTTTTAT
43.	F	IR	<i>rps12-trnV</i>	IGS	39	TTCTATTATATTAGTATTAGATTAGTATTA
44.	F	IR	<i>rrn4.5-rrn5</i>	CDS	34	CATTGTTCAACTCTTTGACAACACGAAAAACCCA
45.	P	SSC	<i>ndhF-rpl32</i>	IGS	33	ACTAAAAAAAAAAAAAGAATTCTTTTTTTTTTAGT

46.	F	IR	<i>rpl32-trnL</i>	IGS	44	ATAATTAAATAAAAAGGATAATTAATAAAAAGGATAAATAA ATAA
Oligonucleotide repeats in <i>T. grandiflorum</i>						
1.	P	LSC	<i>trnK-rps16</i>	IGS	33	ATCTATTCTGAACATTCAATGTTTCAGATTAGAT
2.	R	LSC	<i>rps16/clpP</i>	Intron	31	TTTTTTCGAATAAAAAAAAAAAGAAGAAAAA
3.	F	LSC	<i>trnS-trnG</i>	IGS	30	TTATATAGATATATACAACCTTTATATAGA
4.	P	LSC	<i>trnG-trnR</i>	IGS	58	TTTATTAATTCAATTCAATGATGCATTAATTAATGCATCATT GAATTGAATTAATAAA
5.	R	LSC	<i>trnR-atpA/psbZ-trnG</i>	IGS	31	TTATTATAATATAATTTTATTATTATTATTA
6.	P	LSC	<i>atpF-atpH</i>	IGS	38	TTTAATAATTAATAAATTATTAATTTATTATTATTA
7.	F	LSC	<i>atpF-atpH/psbZ-trnG</i>	IGS	31	TAATTAATAAATTATTAATTTATTATTATTA
8.	P	LSC	<i>rps2-rpoC2</i>	IGS	33	TCTCTTTTTTTTTTAAGTAAAAAAAAAAGAGA
9.	F	LSC	<i>rpoC1/rpl16</i>	Intron	31	TCGGACATGAGAGTTTCCTCTCATCCGGCTC
10.	P	LSC	<i>trnC-petN</i>	IGS	31	GTAGACACTCCACTACTAATGGAGTGTCTAC
11.	P	LSC	<i>psbM-trnD/rps12-trnV</i>	IGS	34	TTTCCACTTCCACCGTTTACAAATAAACCCCAAC
12.	F	LSC	<i>trnT-psbD</i>	IGS	31	TATATGGATAATAATTATATGGATAATAATT
13.	P	LSC	<i>trnD-psbD</i>	IGS	32	GCAGTGCACGAGAAATCAAATCATAATAAAA
14.	R	LSC	<i>psbZ-trnG</i>	IGS	31	AAAGAAAATAATAAATAGAAAATAATAAAA
15.	R	LSC	<i>psbZ-trnG/rpl32-trnL</i>	IGS	31	ATTAATAAATAATATAGTAATAAATAATATA
16.	P	LSC	<i>psbZ-trnG/trnT-trnL</i>	IGS	32	AAATATATTAATATATATTTATATTAATATAA
17.	P	LSC	<i>psbZ-trnG</i>	IGS	36	AAATATATTAATATATATTTATATTAATATAAATAT
18.	C	LSC	<i>psbZ-trnG/trnT-trnL</i>	IGS	31	AATATATTAATATATATTTATATTAATATAA
19.	F	LSC	<i>psaB-psaA</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATATCAG TCAGCCATA
20.	F	LSC/SSC	<i>ycf3/ndhA</i>	Intron	41	TCCAAAACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
21.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/IGS	36	AACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
22.	R	LSC	<i>trnT-trnL/atpB-rbcL</i>	IGS	30	TTTTAATATTTATATATTTATTTATATATT
23.	R	LSC	<i>trnT-trnL</i>	IGS	32	TTTAATATTTATATATTTATTTATATATTATA

24.	R	LSC	<i>trnT-trnL</i>	IGS	38	TTTAATATTTATATATTTATTTATATATTATATAATTT
25.	R	LSC	<i>trnT-trnL</i>	IGS	30	ATATTTATATATTTATTTATATATTATATAATT
26.	F	LSC	<i>trnT-trnL/ndhC-trnV</i>	IGS	31	ATATTTATATATTTATTTATATATTATATAA
27.	R	LSC	<i>trnT-trnL</i>	IGS	32	TATTTATATATTTATTTATATATTATATAAATT
28.	F	LSC	<i>trnT-trnL</i>	IGS	54	ATTTATATATTTATTTATATATTATATAATTTAATTTATATA ATTATTTATATATT
29.	F	LSC	<i>trnT-trnL/atpB-rbcL</i>	IGS	30	ATTTAATTTATATAATTATTTATATATTATA
30.	F	LSC	<i>trnF-ndhJ/ndhK-ndhJ</i>	IGS	32	TTATTTATTTTATTTTATTTCTTTTTATT
31.	P	LSC	<i>ndhC-trnV</i>	IGS	34	AGTAATAGTATAAATATATATTTATACTATTACT
32.	F	LSC/IR	<i>ndhC-trnV/rps12-trnV</i>	IGS	31	TTCTATTCTATTTTCTATTAGAAATATATT
33.	F	LSC	<i>ndhC-trnV/atpB-rbcL</i>	IGS	30	TATATTTATATATTAATATATATATTATTT
34.	C	LSC	<i>atpB-rbcL/petA-psbJ</i>	IGS	34	TTTATATTATAAATTTTATAATTTATATTATAA
35.	R	LSC	<i>atpB-rbcL</i>	IGS	30	ATATTATAATTATATATTTATTTATATATA
36.	R	LSC	<i>atpB-rbcL</i>	IGS	32	TTTATTTATATATATTATTATATATATTATTT
37.	C	LSC/IR	<i>atpB-rbcL/trnA-rrn23</i>	IGS	31	ATTTCTATTTCAATTTATTTCAATTTCAATTT
38.	F	LSC	<i>rbcL/rbcL*-accD</i>	CDS/IGS	32	TCAAATTTGAATTCGAAGCAGTGGATACTTTA
39.	C	LSC	<i>petA-psbJ/rpl33-rps18</i>	IGS	30	ATATATTAAGTATAAAAATAAGTATAAATAA
40.	R	LSC	<i>clpP</i>	Intron	30	TTTTTTTCAAAAAAAAAAAGAAAGAAAAAAT
41.	P	LSC	<i>psbT-psbN*</i>	IGS	48	ATTGAAGTAATGAGCCTCCCAATATTTGGAGGCTCATTACT TCAATTA
42.	C	LSC/SSC	<i>petD-rpoA/ndhF- rpl32</i>	IGS	30	TTTTTTTTCTAAAGTAAAAAAAAAAATGAAA
43.	F	IR	<i>ycf2</i>	CDS	31	TCTTTTTGTCCAAGTTACTTCTTTTTTGTCT
44.	F	IR	<i>ycf2</i>	CDS	56	CGATATTGATGATAGTGACGATATTGATGCTAGTGACGATA TTGATGCTAGTGACG
45.	F	IR/SSC	<i>rps12-trnV/ndhA</i>	IGS/Intron	38	AACCGTACATGAGATTTTCACCTCATACGGCTCCTCGT
46.	R	IR	<i>rps12-trnV</i>	IGS	31	TATTATTTTTTTTCCCCCTTTTTTTTTATT
47.	F	IR	<i>rps12-trnV</i>	IGS	30	TTCTATTATATTAGTATTAGATTAGTATTA

5.3.5 Substitutions and InDels analysis in chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*

The divergence in chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum* were determined by pairwise MAFFT alignment. Chloroplast genomes of both species were identical (99.5%) with few divergences that were visible in protein-coding genes and intergenic spacer regions. We observed 99.5% identity with 444 substitutions and 23 InDels. The LSC region contained 329 (74.1%) SNPs and 16 (69.57%) InDels, the SSC region had 110 (24.77%) SNPs and 6 (26.09%) InDels, and the IRb region contained only 5 (1.13%) SNPs and 1 (4.35%) InDels. We determined 155 transition (Ts) substitutions and 289 transversion (Tv) and transversion substitution with a Ts/Tv ratio of 0.54. The six types of substitutions are shown in Figure 5.3.

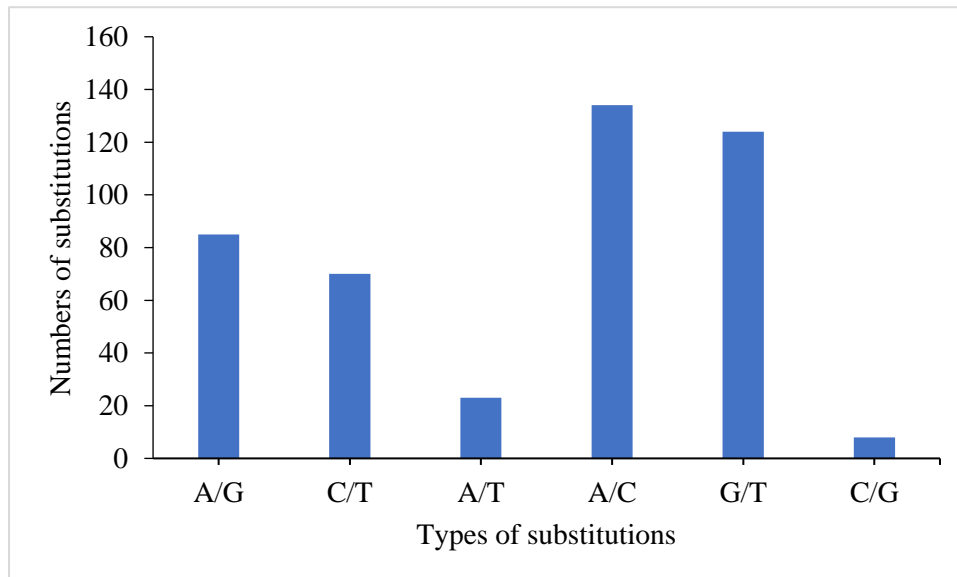


Figure 5.3 Types of substitutions between *T. cacao* and *T. grandiflorum*

5.3.6 Highly polymorphic regions between *Theobroma* species

The regions of chloroplast genomes were variable among the *Theobroma* species with nucleotide diversity (π) ranging from 0.000290 (*ycf2* intronic part of the gene) to 0.021277 (*ndhD-psaC* IGS region). The IGS showed higher genetic diversity than the protein coding and intronic regions (Figure 5.4) whereas tRNA regions proved to be remarkably conserved i.e. $\pi=0$. Among the 30 polymorphic regions, 22 were included from IGS, 5 from intronic regions, and only 1 from protein-coding genes (Table 5.6). The *ycf1* gene showed $\pi=0.007231$ whereas the *ndhF* gene showed $\pi=0.004032$. The protein coding regions were *rpl22*, *ycf1* and 2nd exon of *clpP*, whereas the IGS regions included *trnH-psbA*, *ndhE-ndhG*, *rpl32-trnL*, *trnQ-psbK*, *rpl33-rps18*, *trnP-psaJ* and *ndhF-rpl32*.

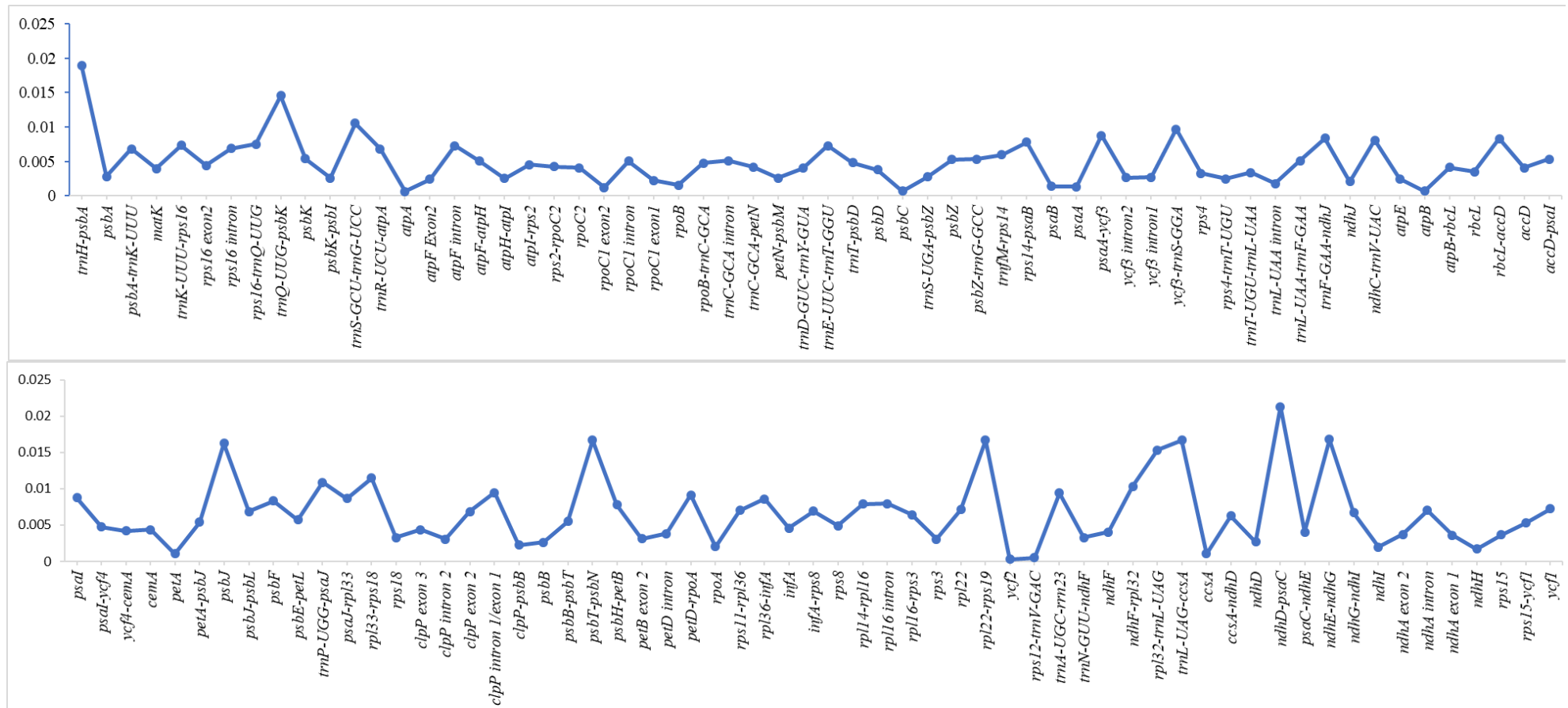


Figure 5.4 Nucleotide diversity of the chloroplast genome regions. X-axis represents chloroplast genome regions whereas Y-axis represents nucleotide diversity of the regions.

Table 5.6 Thirty high polymorphic regions among *T. cacao* and *T. grandiflorum*

S.No	Region	Nucleotide diversity	Mutational events	Region length
1.	<i>trnH-psbA</i>	0.018987	9	474
2.	<i>ndhE-ndhG</i>	0.016807	4	238
3.	<i>rpl32-trnL-UAG</i>	0.015345	18	1173
4.	<i>trnQ-UUG-psbK</i>	0.014577	5	343
5.	<i>rpl33-rps18</i>	0.011494	4	348
6.	<i>trnP-UGG-psaJ</i>	0.01087	4	368
7.	<i>trnS-GCU-trnG-UCC</i>	0.010526	9	855
8.	<i>ndhF-rpl32</i>	0.010358	11	1062
9.	<i>ycf3-trnS-GGA</i>	0.009709	9	927
10.	<i>clpP intron 1/exon 1</i>	0.009454	9	952
11.	<i>trnA-UGC-rrn23</i>	0.009434	2	212
12.	<i>petD-rpoA</i>	0.009132	2	219
13.	<i>psaA-ycf3</i>	0.008772	8	912
14.	<i>psaJ-rpl33</i>	0.008639	4	463
15.	<i>trnF-GAA-ndhJ</i>	0.00838	6	716
16.	<i>rbcL-accD</i>	0.008322	6	721
17.	<i>ndhC-trnV-UAC</i>	0.008071	10	1239
18.	<i>rpl16 intron</i>	0.007952	12	1509
19.	<i>rps16-trnQ-UUG</i>	0.007505	4	533
20.	<i>trnK-UUU-rps16</i>	0.007353	6	816
21.	<i>atpF intron</i>	0.007308	6	821
22.	<i>trnE-UUC-trnT-GGU</i>	0.007308	6	821
23.	<i>ycf1</i>	0.007231	41	5670
24.	<i>rpl22</i>	0.007143	3	420
25.	<i>ndhA intron</i>	0.007036	8	1137
26.	<i>rps16 intron</i>	0.006865	6	874
27.	<i>psbA-trnK-UUU</i>	0.006849	2	292
28.	<i>clpP exon 2</i>	0.006849	2	292
29.	<i>ndhG-ndhI</i>	0.006696	3	448
30.	<i>ccsA-ndhD</i>	0.006289	2	318

5.4 Conclusion

Theobroma cacao and *Theobroma grandiflorum* chloroplast genomes had similar genome structure, gene content, and organisation. The codon usage, amino acid frequency, putative RNA editing sites, oligonucleotide repeats, and microsatellite analyses showed similarities in genus *Theobroma*. Transition substitutions are fewer than transversions substitutions. Intergenic spacer regions and intronic sequences showed high divergent than coding sequencing. Thirty high polymorphic regions identified in current study might be used to develop suitable markers for phylogenetics inference of the genus *Theobroma*.

Chapter 6

Comparative analyses of chloroplast genomes of *Firmiana colorata*, *Firmiana major* and *Firmiana pulcherrima*

6.1 Introduction

The genus *Firmiana* belongs to subfamily Sterculioideae of family Malvaceae and includes deciduous trees (rarely shrubs) (Bayer *et al.*, 1999; Kostermans, 1957; Wilkie *et al.*, 2006). Sixteen taxa of the genus *Firmiana* has been categorised as accepted whereas five species are categorised as unresolved (<http://www.theplantlist.org/>; accessed 29 December 2018). Species of the genus *Firmiana* exist in Eastern Africa and Eastern-south Asia to Malaysia (Kostermans, 1957). Some species of this genus are cultivated for their beautiful shape and lovely flowers (Fan *et al.*, 2013). *Firmiana* species had also shown anti-inflammatory (Lim *et al.*, 2017), antimicrobial (Ajaib *et al.*, 2014) and anti-cancer (Woo *et al.*, 2015) properties. Some species have shown hepatoprotective (Kim *et al.*, 2015) and neuroprotective (Lim *et al.*, 2017) effects. *Firmiana colorata* is used for the intestinal dysfunction in some tribes of Bangladesh (Azam *et al.*, 2013).

The phylogenetic inference of genus *Firmiana* is ignored due to lack of efficient molecular markers (Fan *et al.*, 2013). Up to the best of our knowledge, one study focused on four species of *Firmiana* and evaluated the phylogenetic relationships based on microsatellite markers developed by transcriptome assembly (Fan *et al.*, 2013). The published chloroplast genome sequences of *Firmiana major* (Ya *et al.*, 2017) and *Firmiana pulcherrima* (Wang *et al.*, 2017) with the advent of NGS technology provide genomic resources for the molecular evolution study of the genus *Firmiana*. Previous phylogenetic analyses of subfamily Sterculioideae showed certain discrepancies in phylogenetics of certain clades (Wilkie *et al.*, 2006). Therefore, the authors suggested the use of new molecular tools for the accurate resolution of Sterculioideae phylogenetics.

As discussed in the chapter 5, lack of meiotic recombination and slow evolving nature of chloroplast genome as compared to nuclear genome (Daniell *et al.*, 2016; Palmer, 1985) make it a suitable source for development of markers for resolving of phylogenetics at genus and family level. Previously, robust and authentic molecular markers were developed on the basis of complete chloroplast genome for phylogenetics resolving as well as for barcoding purposes (Ahmed *et al.*, 2013; Nguyen *et al.*, 2018).

The *de novo* assembled chloroplast genome sequence of *Firmiana colorata* and its comparison with chloroplast genomes of *Firmiana major* (Ya *et al.*, 2017) and *Firmiana pulcherrima* (Wang *et al.*, 2017) will not only be helpful to elucidate chloroplast genome structure of genus *Firmiana*, but the broad comparison of these species will also enable us to identify the mutational hotspots regions for development of suitable and authentic markers for resolving of phylogenetics of genus *Firmiana* and subfamily Sterculioideae.

6.2 Materials and Methods

6.2.1 Chloroplast genome comparative analysis

Genomic features and organisation of chloroplast genomes of *Firmiana* species were compared using Geneious R8.1 (Kearse *et al.*, 2012). Different types of substitutions and rate of transition and transversion of substitutions were also determined from pairwise MAFFT alignment using *Firmiana major* chloroplast genome as reference for *Firmiana colorata* and *Firmiana purlcherrima* through Geneious R8.1 (Kearse *et al.*, 2012). For rate of substitutions and InDels in LSC, SSC and IR, we aligned each part separately and analyses was performed using DnaSP (Rozas *et al.*, 2017).

6.2.2 Amino acids frequency, codon usage, putative RNA editing sites

We determined amino acids frequency and codon usage using Geneious R8.1. The putative RNA editing sites were determined by PREP-cp (Putative RNA Editing Predictor of Chloroplast) with default parameters (Mower, 2009).

6.2.3 Microsatellites and oligonucleotide repeats analysis

MISA (Thiel *et al.*, 2003) was used to identify microsatellites loci within genomes of *Firmiana* using parameters: 7 for mononucleotide repeats, 4 for Di- and 3 each for tri-, tetra-, penta- and hexanucleotide repeats. REPuter (Kurtz *et al.*, 2001) was used to identify four types of repeats: forward (F), reverse (R), palindromic (P) and complementary (C), with parameters: oligonucleotide repeats size ≥ 30 bp and edit distance of 3 to get minimum sequence similarity 90% and compute repeats numbers was set to 500.

6.2.4 Identification of mutational hotspots

Chloroplast genome of *Firmiana colorata* was compared with *Firmiana major* available under accession NC_036395 (Ya *et al.*, 2017) and *Firmiana purlcherrima* available under accession NC_037242 (Wang *et al.*, 2017) by multiple alignment using MAFFT (Multiple Alignments using Fast Fourier Transform) (Kato *et al.*, 2005). We compared each region of genomes such as intergenic spacer regions (IGS), introns, tRNAs, ribosomal RNAs and protein-coding genes determination of nucleotide diversity in Geneious R8.1 (Kearse *et al.*, 2012).

6.3 Results

6.3.1 Chloroplast genome organisation and features of *Firmiana*

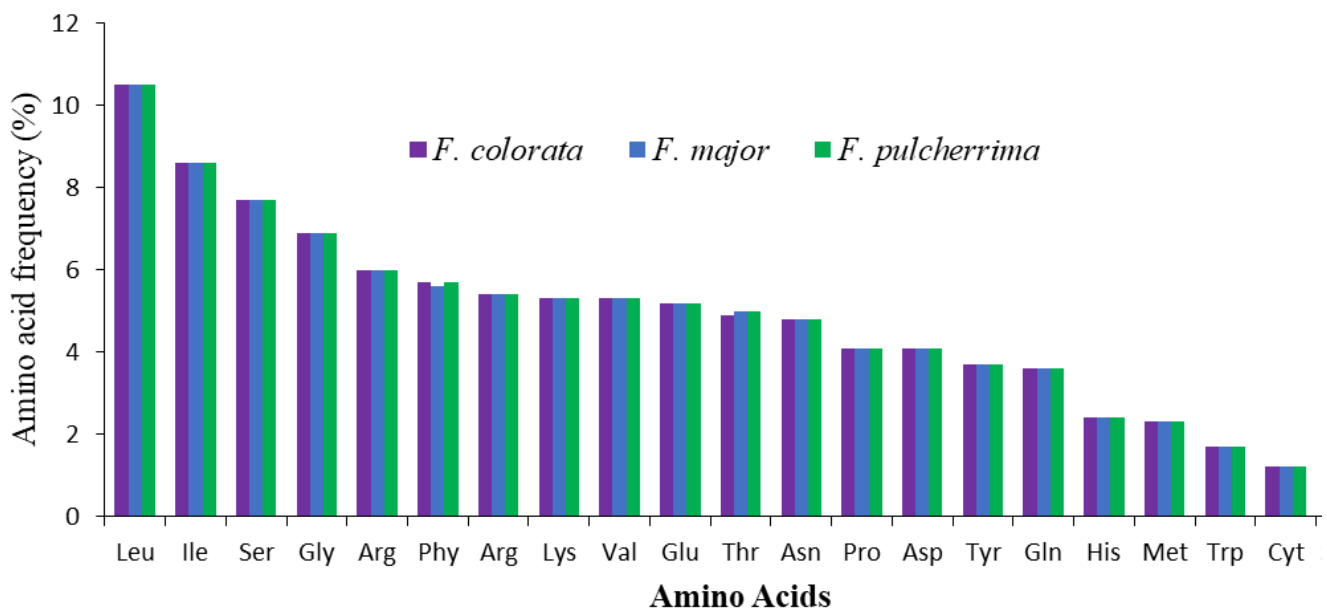
Chloroplast genome had 113 unique genes whereas seventeen genes were present as duplicates in IR regions. Among 113 genes, 79 were protein-coding genes, 30 were tRNA genes (in *Firmiana pulcherrima* 35 tRNAs existed), and 4 were rRNA genes. Among 17 duplicated genes in IR region, 6 were protein-coding genes, four rRNA genes and seven were tRNA genes. Among these genes, 18 genes had intron that included 12 protein-coding genes and 6 tRNA genes. The *rps12* gene was a trans-spliced gene, therefore, its 5' part was present in LSC whereas the 3' parts were present in IR and was duplicated. We found three partially overlapping genes *atpB/atpF*, *psbD/psbC*, and *rps3/rpl22*. So, the genome organisation of *F. colorata* was similar to *F. major* whereas for *F. pulcherrima* little bit difference was found in number of tRNA genes: 35 in *F. pulcherrima* whereas 37 in *F. colorata* and *F. major*. Moreover, SSC was inverse in *F. pulcherrima* as compared to *F. colorata* and *F. major*. The size of the complete chloroplast genome, LSC, SSC, and IR regions showed variations among *Firmiana* species. The complete chloroplast genome ranged in size from 159,556 bp (in *F. pulcherrima*) to 161,302 bp (in *F. major*), LSC regions was 88,444 bp (in *F. pulcherrima*) to 90,187 bp (in *F. major*), IR regions (each IRa and IRb) were 25,572 bp (in *F. major*) to 25,576 bp (in *F. pulcherrima*), SSC region was 19,960 bp (in *F. pulcherrima*) to 20,038 bp (in *F. major*). We also observed similarities among genomes features of *Firmiana* including GC content of LSC, SSC, IR, and complete chloroplast genome. The tRNAs (53.3%) and rRNAs (55.5%) had highest GC content. Consequently, IR regions possessed highest GC content (about 43%) followed by LSC (about 37%) and SSC (about 31.5%). The differences were found in the size of genomes. The complete detail has been provided in Table 6.1.

6.3.2 Comparative analysis of amino acids frequencies and relative synonymous codon usage

Amino acids frequencies and codon usage showed similarities among species of *Firmiana*. Amino acids leucine and isoleucine were the most abundant in species of *Firmiana* whereas tryptophan was least encoded amino acid (Figure 6.1). Relative synonymous codon usage (RSCU) analyses revealed high frequency for codons that had A/T at 3' end and having RSCU value about 1 which indicated that these codons encoded most of amino acids than codons having C/G at 3' end (Table 6.2).

Table 6.1 Comparison of chloroplast genomes of *F. colorata*, *F. major* and *F. pulcherrima* and their general features

Characteristics	<i>F. colorata</i>	<i>F. major</i>	<i>F. pulcherrima</i>	
Size (base pair; bp)	160,700	161,302	159,556	
LSC length (bp)	89,551	90,187	88,444	
SSC length (bp)	20,001	20,038	19,960	
IR length (bp)	25,574	25,532	25,576	
Number of genes	130	130	128	
Protein-coding genes	79	79	79	
tRNA genes	30	30	28	
rRNA genes	4	4	4	
Duplicate genes	18	18	18	
GC content	Total (%)	37.1%	36.9%	37.1%
	LSC (%)	35%	34.7%	35%
	SSC (%)	31.4%	31.3%	31.5%
	IR (%)	42.9%	42.9%	42.9%
	CDS (%)	38.2%	38.1%	38.2%
	rRNA (%)	55.5%	55.5%	53.3%
	tRNA (%)	53.3%	53.2%	55.5%
	All gene %	39.7%	39.6%	39.7%
Protein coding part (CDS) (%bp)	49.24%	49.37%	49.49%	
All gene (%bp)	69.76%	70%	69.72%	
Non-coding region (%bp)	30.34%	30%	30.28%	

**Figure 6.1 Comparison of amino acids frequency among *Firmiana* species.**

The figure shows the amino acids on X-axis and their frequencies on Y-axis.

Table 6.2 Relative synonymous codon usage of genus *Firmiana* species

Codon	Amino acid	Relative synonymous codon usage			Codon	Amino acid	Relative synonymous codon usage		
		<i>Fm</i>	<i>Fc</i>	<i>Fp</i>			<i>Fm</i>	<i>Fc</i>	<i>Fp</i>
GCA	A	1.08	1.1	1.09	CCA	P	1.13	1.15	1.14
GCC	A	0.68	0.68	0.68	CCC	P	0.76	0.74	0.75
GCG	A	0.48	0.47	0.48	CCG	P	0.56	0.56	0.56
GCT	A	1.77	1.76	1.76	CCT	P	1.55	1.55	1.55
TGC	C	0.49	0.49	0.49	CAA	Q	1.54	1.56	1.99
TGT	C	1.51	1.51	1.51	CAG	Q	0.46	0.44	0.57
GAC	D	0.40	0.41	0.41	AGA	R	1.81	1.80	1.78
GAT	D	1.60	1.59	1.60	AGG	R	0.65	0.66	0.64
GAA	E	1.47	1.47	1.48	CGA	R	1.41	1.42	1.41
GAG	E	0.52	0.52	0.52	CGC	R	0.45	0.44	0.45
TTC	F	0.72	0.73	0.72	CGG	R	0.41	0.44	0.42
TTT	F	1.28	1.14	1.27	CGT	R	1.27	1.25	1.24
GGA	G	1.60	1.60	1.59	AGC	S	0.36	0.39	0.39
GGC	G	0.40	0.41	0.41	AGT	S	1.21	1.20	1.18
GGG	G	0.70	0.72	0.73	TCA	S	1.22	1.20	1.19
GGT	G	1.30	1.28	1.27	TCC	S	0.98	0.99	0.99
CAC	H	0.50	0.50	0.51	TCG	S	0.56	0.55	0.55
CAT	H	1.51	1.50	1.50	TCT	S	1.67	1.68	1.69
ATA	I	0.94	0.93	0.93	ACA	T	1.23	1.23	1.23
ATC	I	0.59	0.60	0.6	ACC	T	0.77	0.78	0.79
ATT	I	1.47	1.47	1.47	ACG	T	0.45	0.42	0.43
AAA	K	1.48	1.48	1.47	ACT	T	1.55	1.53	1.55
AAG	K	0.52	0.52	0.53	GTA	V	1.45	1.44	1.44
CTA	L	0.83	0.83	0.83	GTC	V	0.49	0.49	0.50
CTC	L	0.42	0.51	0.41	GTG	V	0.61	0.60	0.60
CTG	L	0.41	0.51	0.42	GTT	V	1.45	1.47	1.46
CTT	L	1.25	1.52	1.24	TGG	W	1	1	1
TTA	L	1.86	2.26	1.86	TAC	Y	0.401	0.41	0.41
TTG	L	1.23	1.51	1.24	TAT	Y	1.59	1.59	1.59
ATG	M	1	1	1	TAA	*	1.66	1.72	1.74
AAC	N	0.47	0.47	0.48	TAG	*	0.66	0.66	0.66
AAT	N	1.53	1.52	1.52	TGA	*	0.68	0.62	0.59

**Fm*: *Firmiana major*, *Fc*: *Firmiana colorata*; and *FP*: *Firmiana pulcherrima*

6.3.3 Analyses of RNA editing sites in *Firmiana* genomes

PREP-cp predicted 65 putative RNA editing sites in 24 genes of *F. colorata*, 62 editing sites in 25 genes of *F. major* and 65 editing sites in 25 genes of *F. pulcherrima* (Table 6.3). High similarities were present among genes for RNA editing sites except: *petD* was unique to *F. colorata* and *rpl2* was unique to *F. major*. Among genes, most of RNA editing sites were found in *ndhB* (11), *ndhD* (6) and *rpoB* (5) in all *Firmiana* species. The *ndhF* gene showed little bit variation among species: In *F. colorata* and *F. pulcherrima* 7 RNA editing sites were found and in *F. major* 4 RNA editing sites were present. RNA editing sites changed A/T on 1st or 2nd nucleotide position of codons whereas position of second nucleotide change was about 4x higher than 1st nucleotide position. Most of conversions were observed for Serine to leucine. Change of RNA editing sites lead to hydrophobic amino acids, i.e. phenylalanine, leucine, and valine.

Table 6.3 RNA editing sites in *Firmiana* genomes

Gene	Nucleotide Position	Amino acid Position	Codon change	A. acid change	Score
RNA editing sites in <i>Firmiana colorata</i>					
<i>accD</i>	797	266	TCG > TTG	S > L	0.8
	1406	469	CCT > CTT	P > L	1
<i>atpB</i>	403	135	CCT > TCT	P > S	0.86
<i>atpA</i>	914	305	TCA > TTA	S > L	1
	1148	383	TCA > TTA	S > L	1
<i>atpF</i>	92	31	CCA > CTA	P > L	0.86
<i>atpI</i>	629	210	TCA > TTA	S > L	1
<i>ccsA</i>	383	128	ACA > ATA	T > I	0.86
	389	130	GCG > GTG	A > V	1
	650	217	ACT > ATT	T > I	0.86
<i>clpP</i>	196	66	CTT > TTT	L > F	1
	559	187	CAT > TAT	H > Y	1
<i>matK</i>	634	212	CAT > TAT	H > Y	1
	1237	413	CAC > TAC	H > Y	1
<i>ndhA</i>	25	42	ACA > ATA	T > I	0.8
	341	114	TCA > TTA	S > L	1
	566	189	TCA > TTA	S > L	1
<i>ndhB</i>	149	50	TCA > TTA	S > L	1
	467	156	CCA > CTA	P > L	1
	542	181	ACG > ATG	T > M	1
	586	196	CAT > TAT	H > Y	1
	611	204	TCA > TTA	S > L	0.8
	737	246	CCA > CTA	P > L	1
	746	249	TCT > TTT	S > F	1

	830	277	TCG > TTG	S > L	1
	836	279	TCA > TTA	S > L	1
	1255	419	CAT > TAT	H > Y	1
	1481	494	CCA > CTA	P > L	1
<i>ndhD</i>	2	1	ACG > ATG	T > M	1
	383	128	TCA > TTA	S > L	1
	674	225	TCG > TTG	S > L	1
	878	293	TCA > TTA	S > L	1
	1298	433	TCA > TTA	S > L	0.8
	1310	437	TCA > TTA	S > L	0.8
<i>ndhF</i>	253	85	CTT > TTT	L > F	1
	290	97	TCA > TTA	S > L	1
	743	248	GCA > GTA	A > V	1
	1549	517	CTT > TTT	L > F	1
	1753	585	CCA > TCA	P > S	1
	1832	611	ACA > ATA	T > I	0.8
	1898	633	GCG > GTG	A > V	0.8
<i>ndhG</i>	166	56	CAT > TAT	H > Y	0.8
	314	105	ACA > ATA	T > I	0.8
<i>petB</i>	418	140	CGG > TGG	R > W	1
<i>psaI</i>	83	28	TCT > TTT	S > F	0.86
<i>psbF</i>	77	26	TCT > TTT	S > F	1
<i>rpl20</i>	308	103	TCA > TTA	S > L	0.86
<i>rpoA</i>	329	110	GCC > GTC	A > V	0.86
	830	277	TCA > TTA	S > L	1
<i>rpoB</i>	338	113	TCT > TTT	S > F	1
	551	184	TCA > TTA	S > L	1
	566	189	TCG > TTG	S > L	1
	610	204	CTC > TTC	L > F	1
	2426	809	TCA > TTA	S > L	0.86
<i>rpoC1</i>	41	14	TCA > TTA	S > L	1
	1261	421	CCG > TCG	P > S	0.86
<i>rpoC2</i>	1972	658	CAC > TAC	H > Y	1
	2293	765	CGG > TGG	R > W	1
	3161	1054	CCC > CTC	P > L	0.86
	4087	1363	CTT > TTT	L > F	0.86
<i>rps2</i>	248	83	TCG > TTG	S > L	1
	325	109	CCC > TCC	P > S	1
<i>rps14</i>	80	27	TCA > TTA	S > L	1
	149	50	TCA > TTA	S > L	1
<i>ycf3</i>	28	10	CTT > TTT	L > F	1
RNA editing sites in <i>Firmiana major</i>					
<i>accD</i>	541	181	CAT > TAT	H > Y	0.8
	797	266	TCG > TTG	S > L	0.8

	1406	469	CCT > CTT	P > L	1
<i>atpB</i>	403	135	CCT > TCT	P > S	0.86
<i>atpA</i>	914	305	TCA > TTA	S > L	1
	1148	383	TCA > TTA	S > L	1
<i>atpF</i>	92	31	CCA > CTA	P > L	0.86
<i>atpI</i>	629	210	TCA > TTA	S > L	1
<i>ccsA</i>	43	15	CTT > TTT	L > F	1
	389	130	GCG > GTG	A > V	1
	650	217	ACT > ATT	T > I	0.86
<i>clpP</i>	196	66	CTT > TTT	L > F	1
	559	187	CAT > TAT	H > Y	1
<i>matK</i>	634	212	CAT > TAT	H > Y	1
	1237	413	CAC > TAC	H > Y	1
<i>ndhA</i>	341	114	TCA > TTA	S > L	1
	566	189	TCA > TTA	S > L	1
<i>ndhB</i>	149	50	TCA > TTA	S > L	1
	467	156	CCA > CTA	P > L	1
	542	181	ACG > ATG	T > M	1
	586	196	CAT > TAT	H > Y	1
	611	204	TCA > TTA	S > L	0.8
	737	246	CCA > CTA	P > L	1
	746	249	TCT > TTT	S > F	1
	830	277	TCG > TTG	S > L	1
	836	279	TCA > TTA	S > L	1
	1255	419	CAT > TAT	H > Y	1
	1481	494	CCA > CTA	P > L	1
<i>ndhD</i>	2	1	ACG > ATG	T > M	1
	383	128	TCA > TTA	S > L	1
	674	225	TCG > TTG	S > L	1
	878	293	TCA > TTA	S > L	1
	1298	433	TCA > TTA	S > L	0.8
	1310	437	TCA > TTA	S > L	0.8
<i>ndhF</i>	253	85	CTT > TTT	L > F	1
	290	97	TCA > TTA	S > L	1
	1549	517	CTT > TTT	L > F	1
	1898	633	GCA > GTA	A > V	0.8
<i>ndhG</i>	166	56	CAT > TAT	H > Y	0.8
	314	105	ACA > ATA	T > I	0.8
<i>petB</i>	418	140	CGG > TGG	R > W	1
<i>psaI</i>	83	28	TCT > TTT	S > F	0.86
<i>psbF</i>	77	26	TCT > TTT	S > F	1
<i>rpl2</i>	70	24	CCC > TCC	P > S	0.86
<i>rpl20</i>	308	103	TCA > TTA	S > L	0.86
<i>rpoA</i>	329	110	GCC > GTC	A > V	0.86

	830	277	TCA > TTA	S > L	1
<i>rpoB</i>	338	113	TCT > TTT	S > F	1
	551	184	TCA > TTA	S > L	1
	566	189	TCG > TTG	S > L	1
	610	204	CTC > TTC	L > F	1
	2426	809	TCA > TTA	S > L	0.86
<i>rpoC1</i>	41	14	TCA > TTA	S > L	1
	1261	421	CCG > TCG	P > S	0.86
<i>rpoC2</i>	2293	765	CGG > TGG	R > W	1
	3185	1062	CCC > CTC	P > L	0.86
	3728	1243	GCA > GTA	A > V	1
<i>rps2</i>	248	83	TCG > TTG	S > L	1
	325	109	CCC > TCC	P > S	1
<i>rps14</i>	80	27	TCA > TTA	S > L	1
	149	50	TCA > TTA	S > L	1
<i>ycf3</i>	28	10	CTT > TTT	L > F	1
RNA editing sites in <i>Firmiana pulcherrima</i>					
<i>accD</i>	797	266	TCG > TTG	S > L	0.8
	1406	469	CCT > CTT	P > L	1
<i>atpB</i>	403	135	CCT > TCT	P > S	0.86
<i>atpA</i>	914	305	TCA > TTA	S > L	1
	1148	383	TCA > TTA	S > L	1
<i>atpF</i>	92	31	CCA > CTA	P > L	0.86
<i>atpI</i>	629	210	TCA > TTA	S > L	1
<i>ccsA</i>	383	128	ACA > ATA	T > I	0.86
	389	130	GCG > GTG	A > V	1
	650	217	ACT > ATT	T > I	0.86
<i>clpP</i>	196	66	CTT > TTT	L > F	1
	559	187	CAT > TAT	H > Y	1
<i>matK</i>	634	212	CAT > TAT	H > Y	1
	1237	413	CAC > TAC	H > Y	1
<i>ndhA</i>	125	42	ACA > ATA	T > I	0.8
	341	114	TCA > TTA	S > L	1
	566	189	TCA > TTA	S > L	1
<i>ndhB</i>	149	50	TCA > TTA	S > L	1
	467	156	CCA > CTA	P > L	1
	542	181	ACG > ATG	T > M	1
	586	196	CAT > TAT	H > Y	1
	611	204	TCA > TTA	S > L	0.8
	737	246	CCA > CTA	P > L	1
	746	249	TCT > TTT	S > F	1
	830	277	TCG > TTG	S > L	1
	836	279	TCA > TTA	S > L	1
	1255	419	CAT > TAT	H > Y	1

	1481	494	CCA > CTA	P > L	1
<i>ndhD</i>	20	7	ACG > ATG	T > M	1
	401	134	TCA > TTA	S > L	1
	692	231	TCG > TTG	S > L	1
	896	299	TCA > TTA	S > L	1
	1316	439	TCA > TTA	S > L	0.8
	1328	443	TCA > TTA	S > L	0.8
<i>ndhF</i>	253	85	CTT > TTT	L > F	1
	290	97	TCA > TTA	S > L	1
	743	248	GCA > GTA	A > V	1
	1549	517	CTT > TTT	L > F	1
	1753	585	CCA > TCA	P > S	1
	1832	611	ACA > ATA	T > I	0.8
	1898	633	GCG > GTG	A > V	0.8
<i>ndhG</i>	166	56	CAT > TAT	H > Y	0.8
	314	105	ACA > ATA	T > I	0.8
<i>petB</i>	418	140	CGG > TGG	R > W	1
<i>petD</i>	395	132	GCT > GTT	A > V	1
<i>psaI</i>	83	28	TCT > TTT	S > F	0.86
<i>psbF</i>	77	26	TCT > TTT	S > F	1
<i>rpl20</i>	308	103	TCA > TTA	S > L	0.86
<i>rpoA</i>	329	110	GCC > GTC	A > V	0.86
	830	277	TCA > TTA	S > L	1
<i>rpoB</i>	338	113	TCT > TTT	S > F	1
	551	184	TCA > TTA	S > L	1
	566	189	TCG > TTG	S > L	1
	610	204	CTC > TTC	L > F	1
	2426	809	TCA > TTA	S > L	0.86
<i>rpoC1</i>	41	14	TCA > TTA	S > L	1
	1261	421	CCG > TCG	P > S	0.86
<i>rpoC2</i>	1972	658	CAC > TAC	H > Y	1
	2293	765	CGG > TGG	R > W	1
	3161	1054	CCC > CTC	P > L	0.86
<i>rps2</i>	248	83	TCG > TTG	S > L	1
	325	109	CCC > TCC	P > S	1
<i>rps14</i>	80	27	TCA > TTA	S > L	1
	149	50	TCA > TTA	S > L	1
<i>ycf3</i>	28	10	CTT > TTT	L > F	1

6.3.4 SSRs and oligonucleotide repeats analyses

SSRs were detected by MISA in all three species of *Firmiana*. *F. colorata* contained 467, *F. major* (472), and *F. pulcherrima* (459) SSRs. Most of SSRs were mononucleotide followed by trinucleotide and then by dinucleotide (Figure 6.2A). The A/T motifs were most abundant in mononucleotide SSRs and AT/TA motifs were most abundant in dinucleotide SSRs. Mononucleotide SSRs motifs varies from 7-22 units repeats, dinucleotide SSRs motif varies from 4-8 unit, repeats whereas other types of SSRs existed mostly in 3 unit repeats (Table 6.4). Most of SSRs were present in LSC, followed by SSC and IR (Figure 6.2B).

REPuter was used to analyse repeats of chloroplast genome in three species of *Firmiana*. We determined four types of repeats: F (forward), P (palindromic), R (reverse), and C (complementary). Total 163 oligonucleotide repeats were present in three *Firmiana* species: 49 in *F. colorata*, 65 in *F. major* and 49 in *F. pulcherrima*. Forward (F) repeats showed abundance in all species: 25 (51%) in *F. colorata*, 32 (49.23%) in *F. major*, and 27 (55.1%) in *F. pulcherrima* (Figure. 6.2C). Oligonucleotide repeats ranged in size from 30-49 bp (Figure 6.2D) and most of repeats existed in range of 30-34 bp. Most of oligonucleotide repeats were found in LSC as compared to SSC and IR (Figure 6.2E). LSC region comprised 24 (49%) repeats in *F. colorata*, 44 (68%) in *F. major* and 26 (53%) in *F. pulcherrima*. Some repeats were shared among LSC, SSC and IR. In chloroplast genome, most of repeats were present in IGS, followed by intronic regions and least number of repeats were found in coding regions (Figure 6.2F). Complete detail of repeats for each species is provided in Table 6.5.

Table 6.4 Microsatellites loci in *Firmiana* species

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	22	Total
A	-	-	-	-	65	26	17	17	9	4	3	1		2	1	2		146
C	-	-	-	-	3	3		1										7
G	-	-	-	-	9	1												10
T	-	-	-	-	75	34	16	14	15	3	2	2	1		1		1	164
AG	-	3																3
AT	-	18	6															24
CT	-	3																3
GA	-	5																5
TA	-	17	3	1	1													22
TC	-		1															1
AAC	2																	2
AAG	2																	2
AAT	5		1															6
AGA	2																	2
ATA	4																	4
ATT	4	1																5
CAA	2																	2
CTC	1																	1
CTG	1																	1
CTT	3																	3
GAA	3																	3
GAT	1																	1
GCA	1																	1
GCT	2																	2
GGT	1																	1
GTT	1																	1
TAA	4																	4
TAG	1																	1
TAT	5	2																7
TCT	1																	1
TGC	1																	1
TTA	8																	8
TTC	6																	6
TTG	2																	2
AAAT	2																	2
ATAA		1																1
ATCA	1																	1
ATCT	1																	1
GAAT	1																	1
TAGT	1																	1
TATT	2																	2
TCTA	1																	1

Firmiana major

	TTAT	1																	1
	AGAAA	1																	1
	TAAAG	1																	1
	TATTA	2																	2
	TGAAA	1																	1
	TTATA	1																	1
	TTCTA	2																	2
	TTTTTA	1																	1
	Total																		472
<i>Firmiana pulcherrima</i>	A	-	-	-	-	70	26	18	12	11	6	2	2	1					148
	C	-	-	-	-		1	3	1										5
	G	-	-	-	-	9				1					1				11
	T	-	-	-	-	65	38	27	11	8	6	2	1		1				159
	AG	-	3																3
	AT	-	16	5			1												22
	CT	-	4																4
	GA	-	5																5
	TA	-	17	2			1												20
	TC	-		1															1
	TG	-	1																1
	AAC	2																	2
	AAG	2																	2
	AAT	6																	6
	AGA	2																	2
	ATA	3																	3
	ATT	2	1																3
	CAA	1																	1
	CTC	1																	1
	CTG	1																	1
	CTT	4																	4
	GAA	3																	3
	GAT	1																	1
	GCA	1																	1
	GCT	2																	2
	GGT	1																	1
	GTT	1																	1
	TAA	3																	3
	TAG	1																	1
	TAT	6	3																9
	TCT	1																	1
	TGC	1																	1
TTA	6																	6	
TTC	6																	6	
TTG	2																	2	

	AAAT	2																	2
	AGGA	1																	1
	ATAA	1																	1
	ATCA	1																	1
	ATCT	1																	1
	GAAT	1																	1
	TAGT	1																	1
	TATT	2																	2
	TCTA	1																	1
	TTTA	2																	2
	TAAAG	1																	1
	TATTA	1																	1
	TGAAA	1																	1
	TTCTA	1																	1
	Total																		459
<i>Firmiana colorata</i>	A	-	-	-	-	68	29	21	10	11	5	2	2	1	1				150
	C	-	-	-	-	3	1	1		1	1								7
	G	-	-	-	-	8		1		1									10
	T	-	-	-	-	68	42	22	12	8	3	2	3		1				161
	AG	-	2																2
	AT	-	17	5															22
	CT	-	4																4
	GA	-	5																5
	TA	-	18	2			1												21
	TC	-		1															1
	TG	-	1																1
	AAC	2																	2
	AAG	2																	2
	AAT	6																	6
	AGA	2																	2
	ATA	2																	2
	ATT	5	2																7
	CAA	1																	1
	CTC	1																	1
	CTG	1																	1
	CTT	4																	4
	GAA	3																	3
	GAT	1																	1
GCT	2																	2	
GGT	1																	1	
GTT	1																	1	
TAA	5																	5	
TAG	1																	1	
TAT	5	2																7	

TCT	1																			1	
TGC	1																				1
TTA	6																				6
TTC	6																				6
TTG	2																				2
AAAT	2																				2
AGGA	1																				1
ATAA	1																				1
ATCA	1																				1
ATCT	1																				1
ATTT	1																				1
GAAT	1																				1
TAGT	1																				1
TATT	1																				1
TTAA	1																				1
TTAT	1																				1
TTTA	2																				2
AGAAA	1																				1
TAAAG	1																				1
TGAAA	1																				1
TTCTA	1																				1
Total																				467	

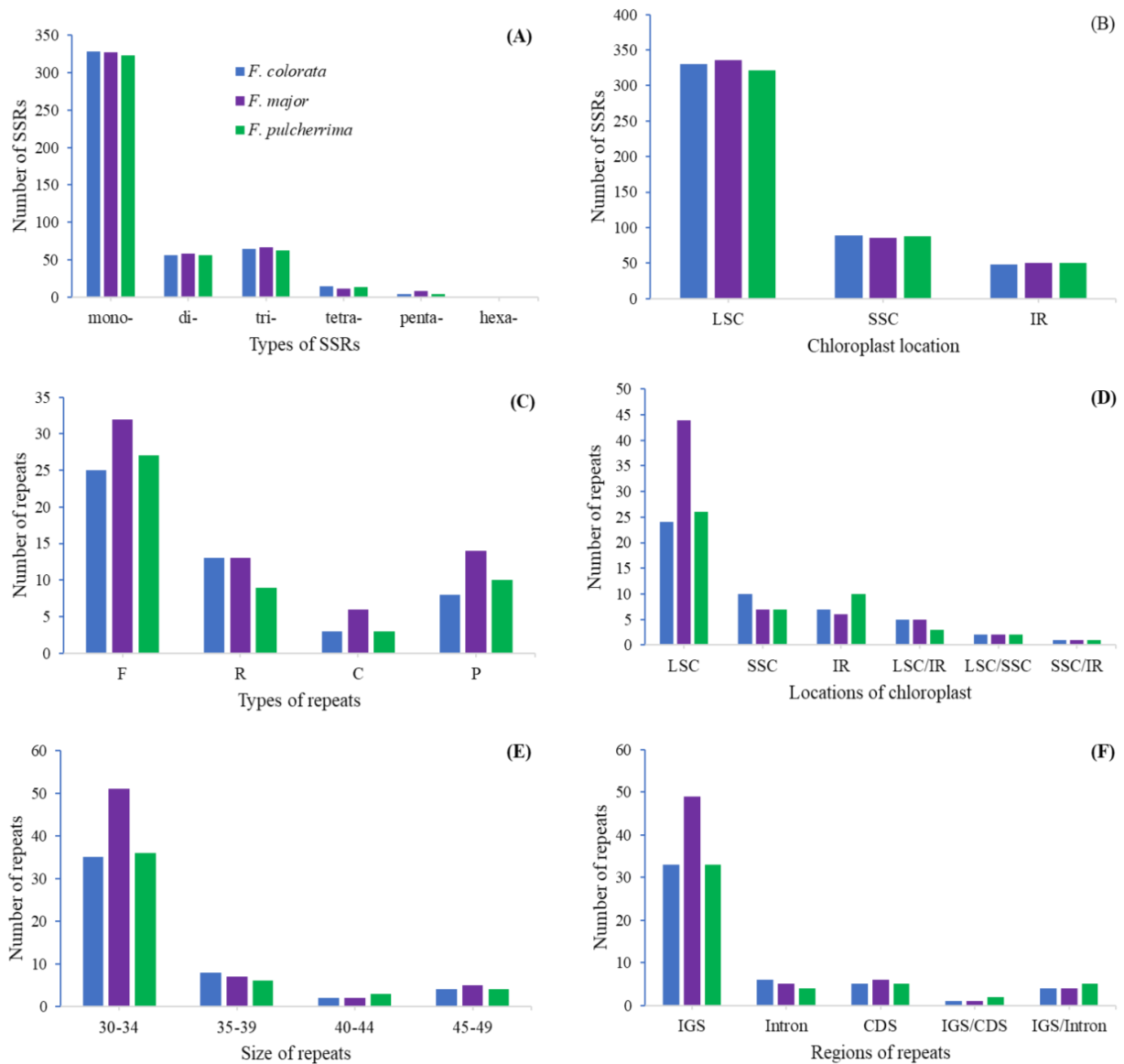


Figure 6.2 Comparison of SSRs and oligonucleotide repeats in *Firmiana*.

(A) Represent types of SSR by size. Total: all types of SSRs present in the genome; mono: mononucleotide; Di: dinucleotide; tri: trinucleotide; tetra: tetranucleotide; penta: pentanucleotide, and hexa: hexanucleotide. (B) Represents distribution of chloroplast genome in the large single copy (LSC), small single copy (SSC) and inverted repeat region (IR). (C) Distribution of oligonucleotide repeats motifs. LSC, SSC and IR stand for those oligonucleotides repeats which are fully present in these regions whereas LSC/SSC, LSC/IR, and SSC/IR stand for those repeats which are dispersed in both locations, i.e. one copy of repeat existed in one location while the second copy was in another location. (D) Four types of oligonucleotide repeats exist in the genome. Forward (F), palindromic (P), reverse (R) and complementary (C). (E) The range of numbers shows the size of oligonucleotide repeats in that specific range. For instance, 30-34 stands for those oligonucleotide repeats which have the size from 30-34 bp. (F) Distribution of repeats based on function regions. IGS: Intergenic spacer regions, CDS: coding DNA sequences; Intron: Intronic regions.

Table 6.5 Oligonucleotide repeats in *Firmiana* species

S.No	Type	Location	region	IGS/Intron/ CDS	size	Sequences
Oligonucleotide repeats in <i>Firmiana colorata</i>						
1.	P	LSC	<i>rps16-trnK</i>	IGS	45	ATTATAATCTAATCTGAACATTCAATGTTTCAGATTAGATTATAAT
2.	F	LSC	<i>rps16-trnK</i>	IGS	35	ATTATTTAATTATTAATTATTATTAATATTTTTTA
3.	R	LSC	<i>rps16-trnK</i>	IGS	36	TTAATTATAAATTATTATTAATTTTTTTTAATTATT
4.	C	LSC/SSC	<i>rps16-trnK/ycf1</i>	IGS/CDS	31	ATTATTAATTTTTTTTAATTATTTAATTATT
5.	P	LSC	<i>trnR-atpA/psbZ-trnG</i>	IGS	32	ATTTTATATATTTCTATTTTTATTCTAATTAT
6.	F	LSC	<i>trnR-atpA/ndhC-trnV</i>	IGS	32	TTCTATTTTTATTCTAATTATTCTAATTTCTA
7.	F	LSC	<i>atpA-atpH/rpl16</i> intron	IGS/Intron	31	TTTTTTTATTTAATAAATTAATATTAATAAAA
8.	P	LSC	<i>rps2-rpoC2</i>	IGS	36	TCTCTCTTTTTTTTTTACTAAAAAAGAGAGA
9.	F	LSC	<i>rpoC1/rpl16</i>	Intron/Intron	31	TCGGACATGAGAGTTTCCTCTCATCCGGCTC
10.	F	LSC	<i>psbZ-trnG/accD-psal</i>	IGS	30	AATAGAATAATAATATAATTAGAATAATAA
11.	C	LSC	<i>psbZ-trnG/rpl33-</i> <i>rps18</i>	IGS	33	AATTATAAAATAAATAATATAAATAATATA
12.	F	LSC	<i>psaB/psaA</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATATCAGTCAGCCATA
13.	R	LSC	<i>psaA/ycf3</i>	IGS	30	TCTATCTATCTATTTTTATCTATCTATTT
14.	P	LSC/SSC	<i>ycf3/ndhA</i>	Intron/Intron	41	TCCAAAACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
15.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/Intron	36	AACCGTACGTGAGATTTTCACCTCATACGGCTCCTCCTC
16.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/Intron	30	TGAGATTTTCACCTCATACGGCTCCTCCTT
17.	R	LSC/IR	<i>trnT-trnL/ndhF-ycf1*</i>	IGS	31	ATTTTAAAATTATTTAATTATTTTATTTT
18.	F	LSC	<i>trnT-trnL</i>	IGS	36	TATTTTATTTAATTATTTAATTCTATTTGTATTTT
19.	F	LSC	<i>ndhC-trnV</i>	IGS	31	TAATTTGATATAGTACTAATTTGTATAGTA
20.	F	LSC/IR	<i>ndhC-trnV/rps12-trnV</i>	IGS	31	TATTTCTATTCTATTTATTCTAATTCTATTA
21.	F	LSC	<i>ndhC-trnV</i>	IGS	31	CTATATATATTTCTATATATTATTTCTATATG
22.	R	LSC	<i>ndhC-trnV/rpl33-</i> <i>rps18</i>	IGS	32	TCTATATATTTATATATATTCTATTTTTTAAT

23.	F	LSC	<i>atpB-rbcL</i>	IGS	30	TTAATTTCTTATTATTTAATTTCTATTATA
24.	F	LSC	<i>rbcL/rbcL-accD</i>	IGS	32	ATCAAATTTGAATTCGAAGCAATGGATACTTT
25.	F	LSC	<i>accD-psaI</i>	IGS	31	ATCTCCCGAGAATTCTATTTTGACTAAAAAT
26.	R	LSC	<i>rpl33-rps18/rpl16</i>	IGS/Intron	31	AAAATATTTATATATTTAATATTTTATTAT
27.	R	LSC	<i>rpl33-rps18</i>	IGS	36	TTATATATTATTTATATATCTATATATTTAATATAT
28.	R	LSC	<i>rpl33-rps18</i>	IGS	33	TTAATATATTATTATTCTATATTATTATTATAT
29.	R	LSC	<i>clpP</i>	Intron	34	TATTTATATATATTTATAATTTATATATATTTAT
30.	C	LSC/IR	<i>clpP/rps19-rpl2</i>	Intron/IGS	33	TATATATATTTATAATTTATATATATTTATAAT
31.	P	LSC	<i>psbT-psbN</i>	IGS	48	AATTGAAGTAATGAGCCTCCCAATATTGGGAGGCTCA TACTTCAATT
32.	F	IR	<i>rps19-rpl2</i>	IGS	32	ATATTAATATTATATATTAATATTATATAT
33.	R	IR	<i>rps19-rpl2</i>	IGS	33	TTAAATATTATATATATATATATTATATATT
34.	F	IR	<i>ycf2</i>	CDS	31	TCTTTTTGTCCAAGTTACTTCTTTTTTGTG
35.	F	IR	<i>ycf2</i>	CDS	47	CGATATTGATGATAGTGACGATATTGATGCTAGTGAC GATATTGATG
36.	F	IR	<i>ycf2</i>	CDS	33	AGTGACGATATTGATGCTAGTGACGATATTGAT
37.	P	IR/SSC	<i>rps12-trnV/ndhA</i>	IGS/Intron	38	AACCGTACATGAGATTTTCACCTCATACGGCTCCTCGT
38.	F	IR	<i>rps12-trnV</i>	IGS	31	TTTTCTATTCTATTAGTATTAGATTAGTATT
39.	F	IR	<i>rrn4.5-rrn5</i>	IGS	34	CATTGTTCAACTCTTTGACAACACGAAAAACCCA
40.	F	SSC	<i>ycf1</i>	CDS	32	ATTAATAAAAATATTATTGAAATTAATAAAAA
41.	P	SSC	<i>ndhA</i>	Intron	34	AAAGAATAAAAAAGAAATTTTTTTTTTATTCTTT
42.	P	SSC	<i>psaC-ndhD</i>	IGS	43	AAAAACCCGTGCTCAAAAAAAGTCTTTTTGAGCACGG GTTTT
43.	F	SSC	<i>ndhD*-ccsA</i>	IGS	30	ATTTTTATAGATAGTTTTTCATGAATAAAAT
44.	F	SSC	<i>trnL-rpl32</i>	IGS	30	TAAAAATAACAAATTATTAACACTAACA
45.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	34	TTTTAATTTATTATTTTAGAATTTATTATTTAAA
46.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	31	TTTTAATTTATTATTTTAGAATTTATTATTT
47.	F	SSC	<i>ndhF-ycf1Ψ</i>	IGS	31	TTTATTATTTTAGAATTTATTATTTAAATTT
48.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	31	TTATTATTTTAGAATTTATTATTTAAATTTT
49.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	30	AATTTATTATTTAAATTTTTATTATTTTAG

Oligonucleotide repeats in *Firmiana major*

1.	P	LSC	<i>trnK-rps16</i>	IGS	45	ATTATAATCTAATCTGAACATTCAATGTTTCAGATTAGATTATAAT
2.	F	LSC	<i>trnS-trnG</i>	IGS	31	TATCATTTTTTTTTCTTCCTCTTTCTTTTT
3.	R	LSC	<i>trnG/trnT-trnL</i>	IGS	30	AATTTTATTTTATAATTATATTATAAATTT
4.	P	LSC	<i>trnG-trnR</i>	IGS	30	ATTTGAATATTTATTAATTC AATTC AATGAT
5.	P	LSC	<i>trnG-trnR</i>	IGS	47	TAATTC AATTC AATGATGCATTAATAATGCATCATTGAATTGAATTA
6.	P	LSC	<i>trnR-atpA</i>	IGS	33	ATATTCTATATTATATAGAATATATATTCTATATA
7.	P	LSC	<i>trnR-atpA</i>	IGS	30	ATAATTAATTAGATAATTCTAATTAATTAT
8.	R	LSC	<i>trnR-atpA/psbZ-trnG</i>	IGS	35	TAATTAATTATATTTTATATTTTTATTCTAATTATT
9.	P	LSC	<i>trnR-atpA/psbZ-trnG</i>	IGS	34	ATTAATTATATTTTATATTTTTATTCTAATTATT
10.	P	LSC	<i>trnR-atpA</i>	IGS	35	ATTCTAATTATTCTAATTCTAATTAGAAATTAGA
11.	F	LSC	<i>rpoC2</i>	CDS	30	CTTTTTAATTGATTGTTTTATCACTTTTT
12.	P	LSC	<i>trnE-trnT/rpl33-rps18</i>	IGS	30	TATATATATAAATTATATAAATAGATCTTA
13.	C	LSC	<i>psbZ-trnG/rpl33-rps18</i>	IGS	30	AATATAATAATAATTTTATAATATATAATA
14.	F	LSC	<i>psbZ-trnG</i>	IGS	31	TAATATATAATAAATAGAAGAATAAATAATAT
15.	C	LSC	<i>psbZ-trnG/rpl33-rps18</i>	IGS	34	ATAATAATATAATTAGAATAATAATAAATAATATATA
16.	P	LSC	<i>psbZ-trnG/trnT-trnL</i>	IGS	31	TAATAATATAATTAGAATAATAATAATAATA
17.	R	LSC	<i>psbZ-trnG</i>	IGS	30	ATAATAATAATATATAATTAATATATAATA
18.	C	LSC	<i>psbZ-trnG/rpl33-rps18</i>	IGS	32	ATAAAAATAAAGATATAATTCTAAAATAAATAA
19.	F	LSC	<i>psbZ-trnG</i>	IGS	34	TATATTAATATAATTTTATATTTTATATTAATT
20.	R	LSC/SSC	<i>psbZ-trnG/ndhF-ycf1*</i>	IGS	31	TTTTATATTTATATTAATTTTTTATATTTATA
21.	F	LSC	<i>psbZ-trnG</i>	IGS	31	TAATATTAATAACATTAATAACAATTAATAA
22.	F	LSC	<i>psaB/psaA</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATATCAGTCAGCCATA
23.	R	LSC	<i>psaA-ycf3</i>	IGS	30	TCTATCTATCTATTTTTTATCTATCTATTT
24.	P	LSC	<i>psaA-ycf3</i>	IGS	32	AAAATAAAATATAAATATATATATTTTTATTTT
25.	P	LSC/SSC	<i>ycf3/ndhA</i>	Intron/Intron	41	TCCAAAACCGTACGTGAGATTTTCACCTCATACGGCTCTC
26.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/IGS	36	AACCGTACGTGAGATTTTCACCTCATACGGCTCCTC

27.	C	LSC	<i>ycf3-trnS/clpP</i>	IGS/Intron	34	TTTTTTTTTTTTTTTTTTTTTTTTTAAATATAGC
28.	F	LSC	<i>trnT-trnL</i>	IGS	42	TATTTTAATTATTTTATTTTAATTATTTAATTCTATTGTAT
29.	F	LSC/IR	<i>trnT-trnL/rps19-rpl2</i>	IGS/Intron	31	TATTAATATTCTAATTAGAAATTTAAATATT
30.	F	LSC/IR	<i>trnT-trnL/rps12-trnV</i>	IGS	30	TATTATATATAGTAGTATATTATATTAGTAT
31.	F	LSC	<i>trnT-trnL</i>	IGS	32	AATAAAAATGAAATTTAAAGAAATAAAAATGAA
32.	F	LSC/IR	<i>ndhC-trnV/rps12-trnV</i>	IGS	31	TATTTCTATTCTATTTATTCTAATTCTATTA
33.	F	LSC	<i>ndhC-trnV</i>	IGS	31	TATATATTTATATATATATTTTTAATCTATTT
34.	F	LSC	<i>atpB-rbcL</i>	IGS	30	AATTTCTTATTATTGAATTTCTATTATTTAA
35.	F	LSC	<i>atpB-rbcL</i>	IGS	36	TATTATTGAATTTCTATTATTTAATTTCTATTATAT
36.	F	LSC	<i>atpB-rbcL</i>	IGS	31	AATTTCTATTATTTAATTTCTATTATATAAAA
37.	F	LSC	<i>rbcL/rbcL-accD</i>	CDS/IGS	32	ATCAAATTTGAATTCGAAGCAATGGATACTTT
38.	F	LSC	<i>petA-psbJ</i>	IGS	33	TATTAAGTATAAAAATAAGTATTAAGTATAAAAT
39.	F	LSC	<i>rpl33-rps18</i>	IGS	31	AATATTTTATTTATATATTTAATATTTTATT
40.	R	LSC	<i>rpl33-rps18</i>	IGS	32	ATATATTATTATATTATATTATTATTATATTA
41.	R	LSC/IR	<i>rpl33-rps21/rps12-trnV</i>	IGS	30	TATTATTATATTATATTATTATTATATTAT
42.	F	LSC	<i>rpl33-rps20</i>	IGS	32	TATTATTATATTATATTATTATTATATTATAT
43.	R	LSC	<i>rpl33-rps19</i>	IGS	39	TATTATTATATTATATTATTATTATATTATATATTAT
44.	R	LSC	<i>rpl33-rps22</i>	IGS	30	ATTATTATATTATATTATTATTATATTATAT
45.	R	LSC	<i>rpl33-rps24</i>	IGS	30	TTATTATATTATATATATTATTACATATAT
46.	R	LSC	<i>rpl33-rps23</i>	IGS	33	TTATTATATTATATATATTATTACATATATTTA
47.	F	LSC	<i>clpP</i>	Intron	31	CTTTTTTTGAAAAAAGAAAGAAAAAAAAG
48.	P	LSC	<i>psbT-psbN</i>	IGS	48	AATTGAAGTAATGAGCCTCCAATATTGGGAGGCTCA TTACTTCAATT
49.	C	LSC	<i>petD-rpoA/rpl32-ndhF</i>	IGS	33	TTTTTTTTTTTTTTACTTTAGAAAAAAA
50.	F	LSC	<i>rpl16</i>	Intron	30	CTTAATTTAATATTTATTTAATAATTTAAT
51.	R	LSC	<i>rpl16</i>	Intron	30	ATTATTTAATAATTTAATCTTAATTTAAT
52.	F	IR	<i>ycf2</i>	CDS	31	TCTTTTTGTCCAAGTTACTTCTTTTTGTG
53.	F	IR	<i>ycf2</i>	CDS	47	CGATATTGATGATAGTGACGATATTGATGCTAGTGAC GATATTGATG
54.	F	IR	<i>ycf2</i>	CDS	33	AGTGACGATATTGATGCTAGTGACGATATTGAT

55.	P	IR/SSC	<i>rps12-trnV/ndhA</i>	IGS/Intron	38	AACCGTACATGAGATTTTCACCTCATACGGCTCCTCGT
56.	F	IR	<i>rps12-trnV</i>	IGS	37	ATTAGTATTAGATCTATTAGTATTAGATTAGTATTA
57.	F	IR	<i>rrn4.5-rrn5</i>	IGS	34	CATTGTTCAACTCTTTGACAACACGAAAAACCCA
58.	F	IR	<i>trnR-trnN</i>	IGS	31	TATTTATATTTATCGCATATTTACTATTTAT
59.	F	SSC	<i>ycf1</i>	CDS	31	ATTAATAAAAAATTTTATTGAAATTAATAAAA
60.	P	SSC	<i>ndhA</i>	Intron	33	AAAGAATAAAAAAGAATTTTTTTTTTATTCTTT
61.	C	SSC	<i>trnL-rpl32/rpl32-ndhF</i>	IGS	30	ATTATTATTTAATTATTTAGTATTATAAAT
62.	F	SSC	<i>rpl32-ndhF</i>	IGS	31	TAAATAATAATTGATAATATTTAAATAATAA
63.	F	SSC	<i>ndhF-ycf1Ψ</i>	IGS	30	TTATTTTAATATTTTAATTCTATTTTAATT
64.	F	SSC	<i>ndhF-ycf1Ψ</i>	IGS	30	TTTTAATTTCTTATTTTAATTTTAAATTA
65.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	30	TTTAATTTCTTATTTTAATTTTAAATTTAT
Oligonucleotide repeats in <i>F. pulcherrima</i>						
1.	P	LSC	<i>trnK-rps16</i>	IGS	45	ATTATAATCTAATCTGAACATTCAATGTTTCAGATTAGA TTATAAT
2.	P	LSC	<i>trnR-atpA/psbZ-trnG</i>	IGS	32	ATTTTATATATTTCTATTTTATTCTAATTAT
3.	F	LSC	<i>trnR-atpA/ndhC-trnV</i>	IGS	32	TTCTATTTTATTCTAATTATTCTAATTTCTA
4.	F	LSC	<i>atpF-atpH/rpl16</i>	IGS/Intron	31	TTTTTTTATTTAATAAATTAATATTAATAAAA
5.	C	LSC/SSC	<i>atpF-atpH/trnL-rpl32</i>	IGS	30	TAATAAATTAATATTAATAAATTTCTAATT
6.	P	LSC	<i>rps2-rpoC2</i>	IGS	35	TCTCTCTTTTTTTTTTGGACTAAAAAAAAGAGAGA
7.	F	LSC	<i>rpoC1/rpl16</i>	Intron/Intron	31	TCGGACATGAGAGTTTCCTCTCATCCGGCTC
8.	F	LSC	<i>psbZ-trnG</i>	IGS	33	TAAATCATAAATGATAATATTTAAATAATATAA
9.	F	LSC	<i>psbZ-trnG/accD-psaI</i>	IGS	30	AATAGAATAATAATATAATTAGAATAATAA
10.	C	LSC	<i>psbZ-trnG/rpl33- rps18</i>	IGS	31	ATAGAATAATAATATAATTAGAATAATAATA
11.	P	LSC	<i>psbZ-trnG/rpl33- rps18</i>	IGS	31	AATAATAATATAGAATATATATAAAATAATA
12.	F	LSC	<i>psaB/psaA</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATAT CAGTCAGCCATA
13.	R	LSC	<i>psaA-ycf3</i>	IGS	30	TCTATCTATCTATTTTTATCTATCTATTT
14.	P	LSC	<i>psaA-ycf3</i>	IGS	32	AAAATAAAATATAAATATATATATTTTATTTT
15.	P	LSC/SSC	<i>ycf3/ndhA</i>	Intron/Intron	41	TCCAAAACCGTACGTGAGATTTTCACCTCATACGGCTC CTC

16.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/IGS	36	AACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
17.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/IGS	30	TGAGATTTTCACCTCATACGGCTCCTCCTT
18.	F	LSC	<i>trnT-trnL</i>	IGS	36	TATTTTAATTATTTTATTTTAATTATTTAATTCTAT
19.	F	LSC	<i>trnT-trnL</i>	IGS	36	TATTTTATTTTAATTATTTAATTCTATTTGTATTTT
20.	F	LSC	<i>ndhC-trnV</i>	IGS	31	TAATTTTCGATATAGTACTAATTTGTATAGTA
21.	F	LSC/IR	<i>ndhC-trnV/rps12-trnV</i>	IGS	31	TATTTCTATTCTATTTATTCTAATTCTATTA
22.	F	LSC	<i>rbcL/rbcL-accD</i>	CDS/IGS	32	ATCAAATTTGAATTCGAAGCAATGGATACTTT
23.	F	LSC	<i>accD-psal</i>	IGS	31	ATCTCCCGAGAATTCTATTTTGACTAAAAAT
24.	F	LSC	<i>petA-psbJ</i>	IGS	30	TATTAAGTATAAAATAAGTATTAAGTATAA
25.	R	LSC	<i>rpl33-rps18/rpl16</i>	IGS/Intron	31	AAAATATTTATATATTTAATATTTTATTTAT
26.	R	LSC	<i>rpl33-rps18</i>	IGS	33	TTAATATATTATTATTCTATATTATTATTATAT
27.	F	LSC	<i>rpl33-rps19</i>	IGS	33	TATATTATTATTCTATATTATTATTATTATTATT
28.	F	LSC	<i>rpl33-rps20</i>	IGS	41	TATTATTCTATATTATTATTATATTATTCTATATTATTA TT
29.	F	LSC	<i>rps12-clpP</i>	IGS	30	TTCTATTTAATAGATTAATAAAATATTA
30.	R	LSC	<i>clpP</i>	Intron	34	TATTTATATATATTTATAATTTATATATATTTAT
31.	P	LSC	<i>psbT-psbN</i>	IGS	48	AATTGAAGTAATGAGCCTCCAATATTGGGAGGCTCA TACTTCAATT
32.	R	IR	<i>rps19-rpl2</i>	IGS	34	TTATATTATTATATATATATATATATTATATATT
33.	C	IR	<i>rps19-rpl3</i>	IGS	31	TATATTATTATATATATATATATATTATATA
34.	F	IR	<i>rps19-rpl4</i>	IGS	31	TATATATATATATATATTATATATTTTTTAT
35.	R	IR	<i>rps19-rpl5</i>	IGS	32	TATATATTATATATTTTTTATATATTTTTTAT
36.	R	IR	<i>rps19-rpl6</i>	IGS	32	TTTTTATATATTTTTTATTTTTTATATTTTT
37.	F	IR	<i>ycf2</i>	CDS	31	TCTTTTTGTCCAAGTTACTTCTTTTTGTGTC
38.	F	IR	<i>ycf2</i>	CDS	47	CGATATTGATGATAGTGACGATATTGATGCTAGTGAC GATATTGATG
39.	F	IR	<i>ycf2</i>	CDS	33	AGTGACGATATTGATGCTAGTGACGATATTGAT
40.	P	IR/SSC	<i>rps12-trnV/ndhA</i>	IGS/Intron	38	AACCGTACATGAGATTTTCACCTCATACGGCTCCTCGT
41.	F	IR	<i>rps12-trnV</i>	IGS	30	TTCTATTCTATTAGTATTAGATTAGTATTA
42.	F	IR	<i>4.5 rrn-5 rrn</i>	IGS	34	CATTGTTCAACTCTTTGACAACACGAAAAACCCA
43.	F	SSC	<i>ycf1</i>	CDS	32	ATTAATCAAATATTATTGAAATTAATAAAAA

44.	P	SSC	<i>ndhA</i>	Intron	35	AAAGAATAAAAAAGAAATTTTTTTTTTTATTCTTT
45.	P	SSC	<i>psaC-ndhD</i>	IGS	43	AAAAACCCGTGCTCAAAAAAGTCTTTTTGAGCACGG GTTTT
46.	F	SSC	<i>ndhD*/ndhD-ccsA</i>	CDS/IGS	30	ATTTTTATAGATAGTTTTTCATGAATAAAAT
47.	F	SSC	<i>trnL-rpl32</i>	IGS	30	TAAAATAACAAATTATTAACACTAACA
48.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	32	TATTTTAATTCTTTTTTTTTTTATTAATTTTAT
49.	R	SSC	<i>ndhF-ycf1Ψ</i>	IGS	33	TTTTAATTTATTATTTTATAATTTATTATTAA

6.3.5 Substitutions and InDels analysis in *Firmiana*

Substitutions types were determined in *Firmiana* species using *F. major* as reference. *Firmiana colorata* and *F. pulcherrima* showed 1110 and 1051 SNPs, respectively, in complete chloroplast genome (one IR removed). Types of substitutions showed similarities among *Firmiana* species. A/G and C/T conversions were most abundant as compared to other SNPs (Table 6.6). The transition to transversion ratio (Ts/Tv) was 0.88 and 0.91 for *F. colorata* and *F. pulcherrima*, respectively. In *F. colorata*, LSC contained 816, IR 42, and SSC contained 252 SNPs. In *F. major*, LSC contained 768, IR contained 44 and SSC had 239 SNPs. InDels were also analysed using DnaSP in each part of chloroplast genome. In total, 241 InDels were found in *F. colorata* and 244 InDels were found in *F. pulcherrima*. Most InDels were found in LSC followed by SSC whereas IR contained least InDels (Table 6.7).

Table 6.6 Comparison of substitution in *Firmiana* species

Types	<i>F. colorata</i>	<i>F. pulcherrima</i>
A/G	260	248
C/T	260	253
A/C	200	185
C/G	59	65
G/T	192	187
A/T	139	113
Total	1,110	1,051
Location wise distribution		
LSC	816	768
IR	42	44
SSC	252	239

F. major was used as reference for SNPs detection in *F. colorata* and *F. pulcherrima*.

Table 6.7 Distribution of InDels in *Firmiana* chloroplast genome

	<i>F. colorata</i>	InDel length (bp)	InDel average length
LSC	190	1611	8.47
IR	10	62	6.2
SSC	41	299	7.29
	<i>F. pulcherrima</i>	InDel length (bp)	InDel average length
LSC	188	2786	14.82
IR	13	132	10.15
SSC	43	296	6.88

F. major was used as reference for SNPs detection in *F. colorata* and *F. pulcherrima*.

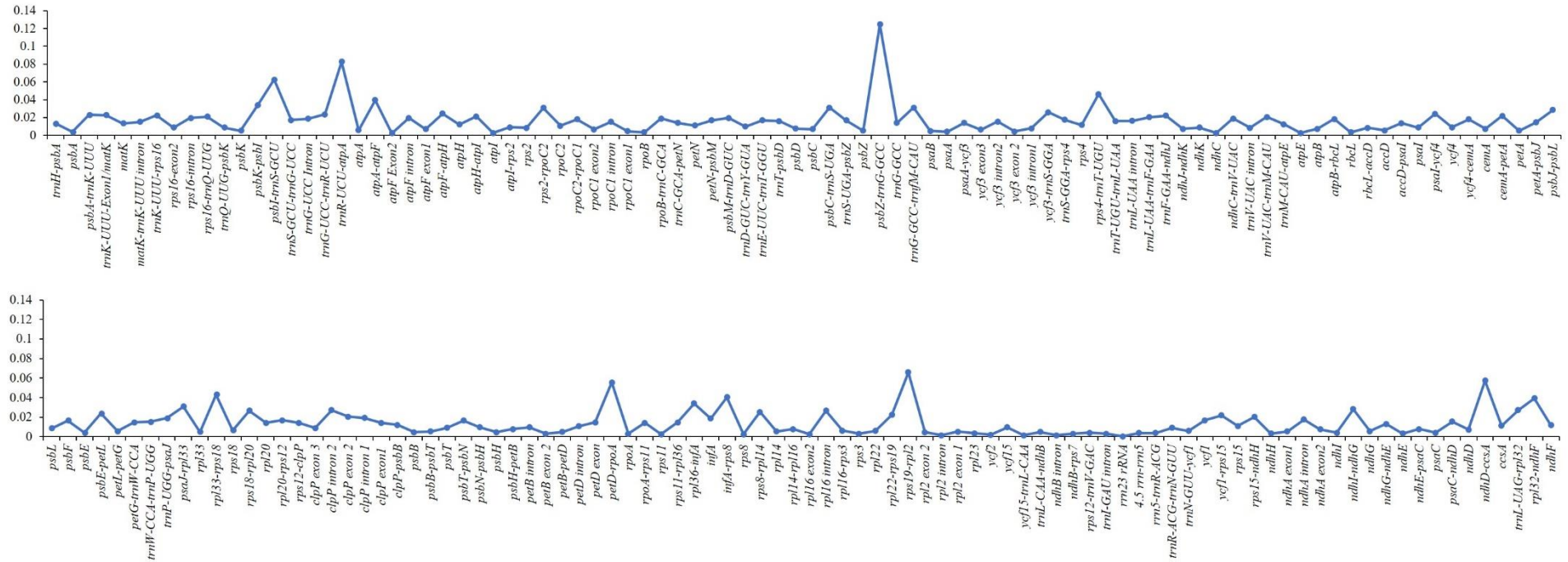


Figure 6.3 Nucleotide diversity of chloroplast genome regions in genus *Firmiana*.

The X-axis represent the chloroplast genome regions whereas the Y-axis represent the nucleotide diversity.

6.3.6 Mutational hotspots in *Firmiana*

We compared protein-coding genes, IGS and intron regions of complete chloroplast genome among three *Firmiana* species to identify mutational hotspots (polymorphic regions). The nucleotide diversity (π) was calculated for each part manually in Geneious R8.1 through MAFFT alignment. IGS regions were more polymorphic (average $\pi=0.018559$) as compared to intron regions ($\pi=0.013274$), and protein-coding regions (average $\pi=0.005249$). Among *Firmiana* three species (*F. colorata*, *F. major*, and *F. pulcherrima*) values ranged from 0.001435 (*rpl2* intron) to 0.084615 (*psbZ-trnG-GCC* region) (Figure 6.3). Here, we selected 30 highly polymorphic regions with alignment length of 200 or greater than to select suitable loci for markers development (Table 6.8).

Table 6.8 Mutational hotspots among *Firmiana* species

S.No	Region	Nucleotide diversity	No's of mutations	Region length
1.	<i>psbZ-trnG-GCC</i>	0.084615385	5	650
2.	<i>trnR-UCU-atpA</i>	0.082677165	21	254
3.	<i>ndhD-ccsA</i>	0.057324841	18	314
4.	<i>petD-rpoA</i>	0.055555556	14	252
5.	<i>rps4-trnT-UGU</i>	0.046189376	20	433
6.	<i>rpl33-rps18</i>	0.043103448	20	464
7.	<i>rpl32-ndhF</i>	0.039443155	34	862
8.	<i>psbK-psbI</i>	0.033653846	14	416
9.	<i>psbC-trnS-UGA</i>	0.031128405	8	257
10.	<i>psaJ-rpl33</i>	0.030888031	16	518
11.	<i>rps2-rpoC2</i>	0.030812325	11	357
12.	<i>ndhI-ndhG</i>	0.02832244	13	459
13.	<i>clpP intron 2</i>	0.027231467	18	661
14.	<i>trnL-UAG-rpl32</i>	0.027187766	32	1177
15.	<i>rps18-rpl20</i>	0.026755853	8	299
16.	<i>rpl16 intron</i>	0.026737968	30	1122
17.	<i>ycf3-trnS-GGA</i>	0.025974026	26	1001
18.	<i>atpF-atpH</i>	0.024509804	15	612
19.	<i>psaI-ycf4</i>	0.024154589	10	414
20.	<i>psbE-petL</i>	0.023595506	21	890
21.	<i>trnG-UCC-trnR-UCU</i>	0.023474178	5	213
22.	<i>psbA-trnK-UUU</i>	0.023026316	7	304
23.	<i>trnK-UUU-Exon1/matK</i>	0.022727273	7	308
24.	<i>trnK-UUU-rps16</i>	0.022346369	20	895
25.	<i>trnF-GAA-ndhJ</i>	0.022222222	16	720
26.	<i>cemA-petA</i>	0.021929825	5	228
27.	<i>ycf1-rps15</i>	0.02173913	8	368
28.	<i>atpH-atpI</i>	0.02124183	26	1224
29.	<i>rps16-trnQ-UUG</i>	0.021015762	12	571
30.	<i>clpP exon 2</i>	0.020547945	6	292

6.4 Conclusion

Our study provides insight into structure of chloroplast genome of *Firmiana* which can be helpful for understanding the pattern of evolution in this genus. The highest similarities among chloroplast genomes structures of *Firmiana* species revealed that close resemblance exists in these species. Mutational hotspots identified in current study might be helpful for development of suitable markers for inferring of phylogenetic in genus *Firmiana* and subfamily Sterculioideae.

Chapter 7

Comparative analyses of *Hibiscus rosa-sinensis* and *Hibiscus syriacus* and screening of mutational hotspots

7.1 Introduction

Hibiscus is one of the most diverse and widely distributed genera of family Malvaceae consisting of about 250-350 species (Prasad, 2014; Rizk and Soliman, 2014). This genus is included in tribe Hibisceae of subfamily Malvoideae (Rizk and Soliman, 2014). Species of this genus are herbs, shrubs or trees and distributed in tropical to temperate regions of the world (Ayanbamiji *et al.*, 2012). Members of this genus are used in industries, horticulture, agriculture, food, and medicines (Akpan, 2000). *Hibiscus* also includes species of high medicinal values that have been shown to possess broad curative activities including anti-bacterial, anti-fungal (Vasudeva and Sharma, 2008) and anti-viral activities (Baatartsogt *et al.*, 2016). In some cases, species of genus *Hibiscus* also showed activity against hypertension, inflammation, hyperlipidaemia, obesity, and anaemia (Riaz and Chopra, 2018; Shen *et al.*, 2017). Anticancer and apoptosis-inducing properties of *Hibiscus* have been also reported (Alam *et al.*, 2018; Goldberg *et al.*, 2017). Taxonomic discrepancies also exist in genus *Hibiscus* due to its plastic morphology (Fryxell, 1997; Pfeil *et al.*, 2002). Some researchers used molecular markers to resolve taxonomic discrepancies and elucidate phylogeny of family Malvaceae and *Hibiscus* (Pfeil *et al.*, 2002; Poovitha *et al.*, 2016; Small, 2004; Tate *et al.*, 2005) but these studies have been inconclusive.

Hibiscus rosa-sinensis is grown throughout the tropics and subtropics due to its ornamental and medicinal values (Prasad, 2014). Many of its varieties and cultivars are available with same morphology except flower colour that ranges from yellow or white to pink or red with single or double petals, but the flower is not available throughout the year which makes the identification of the cultivar almost impossible (Prasad, 2014). The extensive medicinal activity of *H. rosa-sinensis* has also been reported. For instance, antimicrobial, antioxidant, anti-tumour, anti-diabetic and wound healing (Mondal *et al.*, 2016; Prasad, 2014). Different cultivars and varieties varied in their antimicrobial and antioxidant activities toward different pathogenic species (Patel *et al.*, 2012; Prasad, 2014). Therefore, the identification of its cultivars is important in its judgement and use in the herbal medicines.

Owing to lack of meiotic recombination and uniparental inheritance, chloroplast genome evolves slower than the nuclear genome (Palmer, 1985). However, appropriate polymorphisms make it a suitable source for phylogenetics and population genetics studies. For the barcoding and phylogenetics of species, markers based on the chloroplast genome sequencing can be highly authentic, robust, and cost-effective (Ahmed *et al.*, 2013; Nguyen *et al.*, 2018).

The comparative analyses of sequencing of chloroplast genome of *H. rosa-sinensis* with *Hibiscus syriacus* in the current study will not only be useful in elucidating the structure of

chloroplast genome in genus *Hibiscus*, but it will also enable us to identify mutational hotspots that might be helpful for the development of suitable markers for species and cultivar identification, resolving taxonomical discrepancies, and inferring phylogenetics relationship of *Hibiscus*.

7.2 Materials and Methods

7.2.1 Comparison among three *Hibiscus syriacus* chloroplast genomes

In NCBI, chloroplast genome of *Hibiscus syriacus* was available under three accessions KR259989, MH330684 and NC_026909. We compared these three genomes for intra-polymorphism using MAFFT pairwise alignment. On the basis of this comparison, we further selected one genome (accession: KR259989) for comparison with *Hibiscus rosa-sinensis*.

7.2.2 Codon usage, amino acid frequency and RNA editing sites

Codon usage and amino acid frequency of *H. rosa-sinensis* and *H. syriacus* were calculated by Geneious R8.1. The putative RNA editing sites were determined in the coding gene of both species with predictive RNA editor for plants (PREP) suite (Mower, 2009).

7.2.3 Analysis of InDels and substitution types

For the analysis of InDels, LSC, SSC and IR regions of *H. syriacus* and *H. rosa-sinensis* were aligned separately through MAFFT (Katoh *et al.*, 2005) extension in Geneious R8.1 (Kearse *et al.*, 2012). The alignment was exported and analysed in DnaSP v5.10 (Rozas *et al.*, 2017) for the events of InDels. The number of substitutions and types were analysed through Geneious R8.1 (Kearse *et al.*, 2012).

7.2.4 Analysis of repeats in *Hibiscus*

The comparison of oligonucleotide repeats and microsatellites was made between *H. syriacus* and *H. rosa-sinensis*. The MISA was used (Thiel *et al.*, 2003) with a minimum threshold of 7 nucleotides for mononucleotide repeats, 4 for dinucleotide repeats, 3 for tri- tetra-, penta and hexanucleotide repeats. The REPuter program (Kurtz *et al.*, 2001) was used to find forward (F), palindromic (P), reverse (R) and complementary (C) oligonucleotide repeats with minimum repeat size of 30 bp and similarities of 90%. REPuter program over-estimated repeats and redundant repeats were found in large repeats as well as in duplicated tRNAs.

7.2.5 Screening of divergence regions

Genetic divergence between *Hibiscus* species was calculated using nucleotide diversity (π), and total number of mutation (η) for IGS, introns and coding regions of gene by manual visualising of MAFFT (Kato *et al.*, 2005) alignment through Geneious R8.1.

7.3 Results

7.3.1 Comparison among three available chloroplast genomes of *H. syriacus*

The intra-polymorphism analysis of *Hibiscus syriacus* accessions showed negligible variations as compared to variations between *Hibiscus rosa-sinensis* and *Hibiscus syriacus*. The polymorphism that was noted between KR259989 and NC_026909 included 20 small InDels and 3 SNPs, whereas between KR259989 and MH330684 we observed 1 SNP and 7 InDels. Since, most of the InDels belonged to SSRs that cannot be modelled for maximum likelihood analysis, we used only one accession KR259989 for comparative analysis with *Hibiscus rosa-sinensis* because of more accurate annotations.

7.3.2 Comparative analyses of the genomic features between *H. rosa-sinensis* and *H. syriacus*

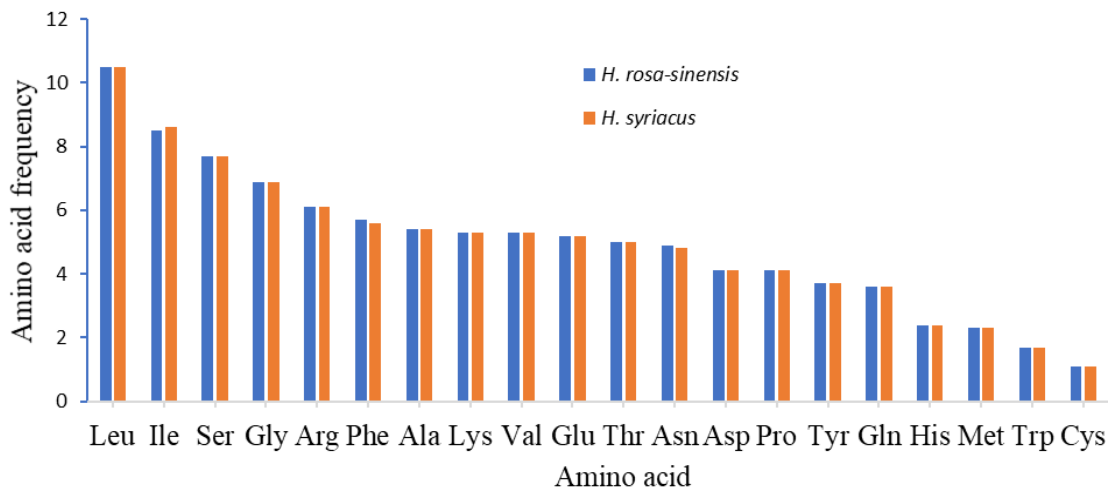
Chloroplast genomes of *H. rosa-sinensis* and *H. syriacus* had similar structure and organisation. Chloroplast genome size of *H. rosa-sinensis* was 160,951 bp, comprising of two inverted repeat regions (IRa and IRb, 25,598 bp each), separated by large single copy (LSC, 89,509 bp) and small single copy (SSC, 20,246 bp) that formed the quadripartite structure. The chloroplast genome of *H. syriacus* had 161,022 bp and possessed similar quadripartite structure with few differences in size of chloroplast genome and its regions. *H. rosa-sinensis* and *H. syriacus* had same gene content and order. The chloroplast genome of both had 113 genes including 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes each. Among these genes, 17 genes were duplicated in the IR regions except truncated gene of *ycf1^ψ*; 19 genes contained introns that included seven tRNA genes and 12 protein-coding genes. The *rps12* gene was a trans-spliced gene, therefore, contained 5' part in the LSC region and 3' part in the IR regions, so 3' part was duplicated in the IR. Out of 19 genes, 17 genes had one intron while two genes (*clpP* and *ycf3*) had two introns. The three protein-coding genes and three tRNA genes that contained introns also duplicated in the IR regions. We found two partially overlapping genes *atpB/atpF* and *psbD/psbC*. The *ycf1* gene started from IR and ended in SSC regions, leaving a truncated copy of 123 bp in *H. rosa-sinensis* and 621 bp in *H. syriacus* in the IRb. Complete detail of comparison has been summarized in Table 7.1.

Table 7.1 Comparison of chloroplast genomes of *H. rosa-sinensis* and *H. syriacus*

Characteristics		<i>H. rosa-sinensis</i>	<i>H. syriacus</i>
Size (base pair; bp)		160,951	161,022
LSC length (bp)		89,509	89,701
SSC length (bp)		20,246	19,831
IR length (bp)		25,598	25,745
Number of genes		130	130
Protein-coding genes		85	85
tRNA genes		37	37
rRNA genes		8	8
Duplicate genes		17	17
GC content	Total (%)	37%	36.8%
	LSC (%)	34.9%	34.7%
	SSC (%)	31.3%	31.1%
	IR (%)	42.9%	42.8%
	CDS (%)	38.2%	38.1%
	rRNA (%)	55.5%	55.4%
	tRNA (%)	53.2%	53.2%
	All gene (%)	39.6%	39.5%
Protein coding part (CDS) (%bp)		49.03%	48.97%
All gene (%bp)		69.3%	69.54%
Non-coding region (%bp)		30.7%	30.46%

7.3.3 Amino acids frequency and codon usage analyses

We determined amino acids frequency and codon usage for *H. rosa-sinensis* and *H. syriacus*. Both species had high similarities in amino acids frequency and codon usage (Table 7.2, Figure 7.1). Among amino acids, leucine was the most abundant amino acid followed by iso-leucine whereas the cysteine was the least coding amino acid. RSCU revealed that most of the amino acids were coded with codons having A/T at 3' end (Table 7.2). Codons with A/T at 3' end had RSCU > 1 except CTA and ATA encoding for leucine and isoleucine, respectively.

**Figure 7.1 Frequency of amino acids in *H. rosa-sinensis* and *H. syriacus***

The figure shows the amino acids on X-axis and their frequencies on Y-axis.

Table 7.2 Comparison of relative synonymous codon usage (RSCU) between *H. rosa sinensis* and *H. syriacus*

Relative synonymous codon usage				Relative synonymous codon usage			
Codon	Amino Acid	<i>Hr</i>	<i>Hs</i>	Codon	Amino Acid	<i>Hr</i>	<i>Hs</i>
GCA	A	1.064	1.068	CCA	P	1.108	1.14
GCC	A	0.676	0.672	CCC	P	0.788	0.748
GCG	A	0.516	0.508	CCG	P	0.568	0.556
GCT	A	1.748	1.752	CCT	P	1.536	1.56
TGC	C	0.514	0.506	CAA	Q	1.528	1.536
TGT	C	1.486	1.494	CAG	Q	0.472	0.464
GAC	D	0.398	0.406	AGA	R	1.806	1.77
GAT	D	1.602	1.594	AGG	R	0.666	0.69
GAA	E	1.47	1.474	CGA	R	1.368	1.398
GAG	E	0.53	0.526	CGC	R	0.456	0.468
TTC	F	0.712	0.706	CGG	R	0.426	0.396
TTT	F	1.288	1.294	CGT	R	1.254	1.278
GGA	G	1.552	1.568	AGC	S	0.348	0.36
GGC	G	0.416	0.424	AGT	S	1.182	1.17
GGG	G	0.748	0.74	TCA	S	1.236	1.248
GGT	G	1.284	1.268	TCC	S	0.96	0.954
CAC	H	0.504	0.51	TCG	S	0.564	0.546
CAT	H	1.496	1.49	TCT	S	1.71	1.722
ATA	I	0.936	0.951	ACA	T	1.188	1.196
ATC	I	0.591	0.588	ACC	T	0.764	0.764
ATT	I	1.473	1.464	ACG	T	0.504	0.48
AAA	K	1.478	1.478	ACT	T	1.544	1.56
AAG	K	0.522	0.522	GTA	V	1.472	1.48
CTA	L	0.828	0.822	GTC	V	0.516	0.52
CTC	L	0.414	0.414	GTG	V	0.556	0.552
CTG	L	0.402	0.39	GTT	V	1.452	1.448
CTT	L	1.242	1.248	TGG	W	1	1
TTA	L	1.884	1.872	TAC	Y	0.406	0.406
TTG	L	1.236	1.248	TAT	Y	1.592	1.594
ATG	M	1	1	TAA	*	1.695	1.695
AAC	N	0.49	0.476	TAG	*	0.705	0.672
AAT	N	1.51	1.524	TGA	*	0.6	0.636

Hr: *Hibiscus rosa-sinensis*, *Hs*: *Hibiscus syriacus*, *Stop codon

7.3.4 Analyses of RNA editing sites in *H. rosa-sinensis* and *H. syriacus*

We also determined putative RNA editing sites in both species of *Hibiscus*. PREP (Mower, 2009) detected 58 putative RNA editing sites in *H. rosa-sinensis* and 61 putative RNA editing sites in *H. syriacus* in 23 genes, which were same in both species. RNA editing sites were similar in both species that lead to same amino acid conversion except three RNA editing sites which were unique to *H. syriacus* (Table 7.3). The highest number of editing sites were determined in *ndhB* (11 in each), *ndhD* (7,8), *ndhF* (4,5) and *matK* (4 each) genes for *H. rosa-sinensis* and *H. syriacus*, respectively. The maximum conversion was predicted for leucine: 30 in *H. rosa-sinensis* and 32 in *H. syriacus*. All the editing sites changed from C/G to A/T at first or second nucleotide position in codons, whereby second position in codon showed 4x higher rate of conversions compared to the first position. Except three, all the RNA editing sites in both species lead to hydrophobic products comprising phenylalanine, methionine, tyrosine, tryptophan, leucine, valine and isoleucine.

Table 7.3 RNA editing sites in *H. rosa-sinensis* and *H. syriacus*

Gene	Nucleotide position	Amino acids position	Amino acids and codon conversion	Score
<i>accD</i>	815	272	TCG (S) => TTG (L)	0.8
	1424	475	CCT (P) => CTT (L)	1
<i>atpA</i>	914	305	TCA (S) => TTA (L)	1
	1148	383	TCA (S) => TTA (L)	1
<i>atpb</i>	403	135	CCT (P) => TCT (S)	0.86
<i>atpF</i>	92	31	CCA (P) => CTA (L)	0.86
<i>atpI</i>	629	210	TCA (S) => TTA (L)	1
<i>ccsA</i>	401	134	GCG (A) => GTG (V)	1
	668	223	ACT (T) => ATT (I)	0.86
	559	187	CAT (H) => TAT (Y)	1
<i>matK</i>	457	153	CAT (H) => TAT (Y)	1
	634	212	CAT (H) => TAT (Y)	1
	638	213	GCA (A) => GTA (V)	1
	1237	413	CAC (H) => TAC (Y)	1
<i>ndhA</i>	89	30	TCG (S) => TTG (L)	0.8
	341	114	TCA (S) => TTA (L)	1
	566	189	TCA (S) => TTA (L)	1
<i>ndhB</i>	149	50	TCA (S) => TTA (L)	1
	467	156	CCA (P) => CTA (L)	1
	542	181	ACG (T) => ATG (M)	1
	586	196	CAT (H) => TAT (Y)	1
	611	204	TCA (S) => TTA (L)	0.8
	737	246	CCA (P) => CTA (L)	1
	746	249	TCT (S) => TTT (F)	1

	830	277	TCG (S) => TTG (L)	1
	836	279	TCA (S) => TTA (L)	1
	1255	419	CAT (H) => TAT (Y)	1
	1481	494	CCA (P) => CTA (L)	1
<i>ndhD</i>	2	1	ACG (T) => ATG (M)	1
	47	16	TCT (S) => TTT (F)	0.8
	383	128	TCA (S) => TTA (L)	1
	674	225	TCG (S) => TTG (L)	1
	878	293	TCA (S) => TTA (L)	1
	1298	433	TCA (S) => TTA (L)	0.8
	1310	437	TCA (S) => TTA (L)	0.8
	1337*	446	CCA (P) => CTA (L)	1
<i>ndhF</i>	290	97	TCA (S) => TTA (L)	1
	1172	391	GCA (A) => GTA (V)	0.8
	1,555	519	CTT (L) => TTT (F)	1
	1,838	613	ACA (T) => ATA (I)	0.8
	1,892*	631	GCG (A) => GTG (V)	0.8
<i>ndhG</i>	166	56	CAT (H) => TAT (Y)	0.8
	314	105	ACA (T) => ATA (I)	0.8
<i>petB</i>	418	140	CGG (R) => TGG (W)	1
<i>psaI</i>	83	28	TCT (S) => TTT (F)	0.86
<i>psbF</i>	77	26	TCT (S) => TTT (F)	1
<i>rpoA</i>	329	110	GCC (A) => GTC (V)	0.86
	830	277	TCA (S) => TTA (L)	
<i>rpoB</i>	338	113	TCT (S) => TTT (F)	1
	551*	184	TCA (S) => TTA (L)	1
	2426	809	TCA (S) => TTA (L)	0.86
<i>rpl20</i>	308	103	TCA (S) => TTA (L)	0.86
<i>rpoC1</i>	41	14	TCA (S) => TTA (L)	1
	1261	421	CCA (P) => TCA (S)	0.86
<i>rpoC2</i>	2317	773	CGG (R) => TGG (W)	1
	3185	1062	CCC (P) => CTC (L)	0.86
<i>rps2</i>	248	83	TCG (S) => TTG (L)	1
	325	109	CCC (P) => TCC (S)	1
<i>rps8</i>	217	73	CAT (H) => TAT (Y)	1
<i>rps14</i>	80	27	TCA (S) => TTA (L)	1
	149	50	TCA (S) => TTA (L)	1

RNA editing sites are almost same except three sites indicated with (). Some differences are shown in position of nucleotide but lead to same type of amino acids change. The positions belong to *H. rosa-sinensis* except that three RNA editing sites that are specific to *H. syriacus*.

7.3.5 Analyses of repeats in *Hibiscus*

The MISA found 487 microsatellites or simple sequence repeats (SSRs) in *H. rosa-sinensis* and 511 in *H. syriacus*. The maximum SSRs were mononucleotide and comprised about 70% of total SSRs, varying in size from seven to fifteen nucleotides. Dinucleotide and trinucleotide SSRs were also abundant and comprised about 28% of the total SSRs (Figure 7.2A). In case of dinucleotide SSRs, AT/TA motifs were abundant whereas in case of trinucleotide SSRs AAT/TTA were abundant. The number of repeats units was also determined for all types of SSRs repeats (Table 7.4); mostly abundant repeat units were three for all SSRs. About 63% SSRs repeats were found in LSC, 27% in SSC and 10% in IR (Figure 7.2B).

We also analysed oligonucleotide repeats by REPuter and found four categories of oligonucleotide repeats: palindromic (P), forward (F), reverse (R), and complementary (C). In chloroplast genome of *H. rosa-sinensis*, REPuter revealed 100 repeats (F=55, R=9, P=33, and C=3) whereas in *Hibiscus syriacus* chloroplast genome it showed 130 repeats (F=79, R=21, P=22, and C=3) (Figure 7.2C). The size of repeats was 30-54 bp in *H. rosa-sinensis* and 30-78 bp in *H. syriacus* (Figure 7.2D). The Intergenic spacer regions (IGS) contained most of the oligonucleotide repeats followed by introns (Figure 7.2E). Most of the repeats were found in LSC (65, 57), followed by SSC (13,40) and lowest were in IR (5, 6) in *H. rosa-sinensis* and *H. syriacus*, respectively. We also found some shared sequences in LSC/SSC (12, 25), SSC/IR (2/1), and LSC/IR (3, 1) in *H. rosa-sinensis* and *H. syriacus*, respectively (Figure 7.2F). The complete detail of repeats position, location and regions of *H. rosa-sinensis* and *H. syriacus* were also provided in Table 7.5.

Table 7.4 Microsatellites loci in *H. rosa-sinensis* and *H. syriacus*

Species	Number of Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
<i>Hibiscus rosa-sinensis</i>	A/T	-	-	-	-	152	78	51	18	16	3	4	1		323
	C/G	-	-	-	-	15	5	2							22
	AC/GT	-	1												1
	AG/CT	-	15	1											16
	AT/AT	-	37	4	3	1									45
	AAC/GTT	6													6
	AAG/CTT	22													22
	AAT/ATT	28	4	1											33
	ACC/GGT	2													2
	AGC/CTG	5													5
	AGG/CCT	2													2
	ATC/ATG	3													3
	AAAG/CTTT	1													1
	AAAT/ATTT	2													2
	AACT/AGTT	1													1
	AATC/ATTG	1													1
	AATG/ATTC	1													1
	AAACT/AGTTT	1													1
	Total														
<i>Hibiscus syriacus</i>	A/T	-	-	-	-	152	71	37	43	19	10	5	3	1	341
	C/G	-	-	-	-	13	2	2	1						18
	AC/GT	-	1												1
	AG/CT	-	15	1											16
	AT/AT	-	39	9											48
	AAC/GTT	7													7
	AAG/CTT	25													25
	AAT/ATT	25	6		1										32
	ACC/GGT	2													2
	ACT/AGT	2													2
	AGC/CTG	5													5
	ATC/ATG	3													3
	AAAG/CTTT	1													1
	AAAT/ATTT	3													3
	AACT/AGTT	1													1
	AATC/ATTG	1													1
	AATG/ATTC	1													1
	AAAGT/ACTTT	2													2
	AATAT/ATATT	2													2
Total															511

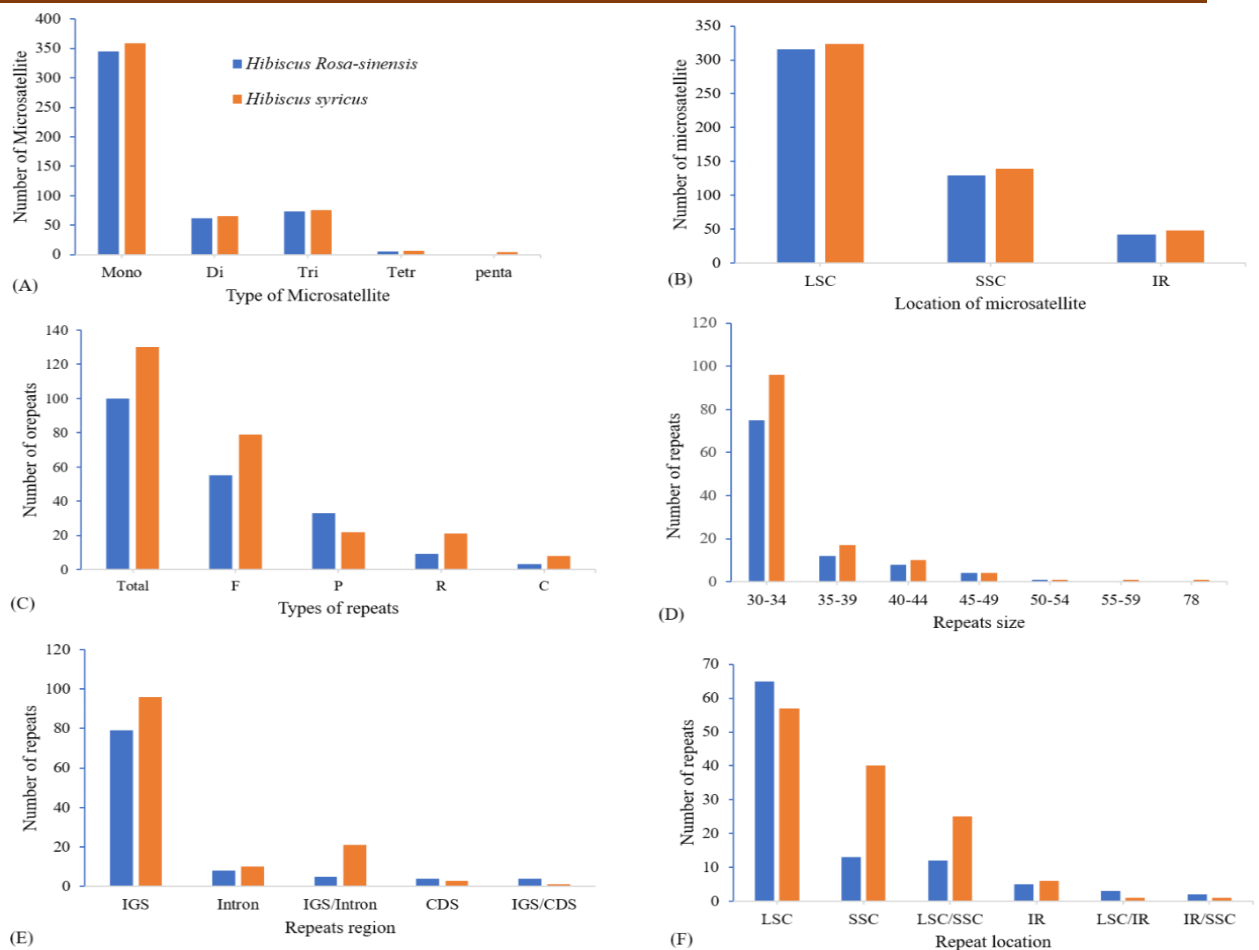


Figure 7.2 Comparison of microsatellites and oligonucleotide repeats between *H. rosa-sinensis* and *H. syriacus*

(A) Represents numbers of different types of microsatellites; Mono: mononucleotide, Di: dinucleotide, Tri: trinucleotide, Tetra: tetranucleotide, Penta: pentanucleotide. **(B)** Locations of microsatellites loci are shown in different regions of chloroplast genome. LSC: large single copy, SSC: small single copy, IR: inverted repeat region. **(C)** Describes different types of oligonucleotide repeats. Total: All types of repeats existing in genomes of *Hibiscus*, F: forward, P: palindromic, R: reverse, C: complementary. **(D)** Represents repeats present in specific range of size i.e 30-34 represents numbers of repeats within the size range of 30 and 34. **(E)** Represents number of repeats in different regions of chloroplast genome. IGS: Intergenic spacer region, CDS: coding DNA sequences, Intron: intronic regions, IGS/Intron: one copy of repeat present in intergenic spacer region and other in intronic regions. IGS/CDS: one copy of repeat present in intergenic spacer region and other in intronic region. **(F)** Represents number of repeats present in different locations of chloroplast genome. LSC: Large single copy, SSC: small single copy, IR: inverted repeat, LSC/SSC: one copy of repeat present in LSC and another in SSC, LSC/IR: one copy of repeat present in LSC and another in SSC, IR/SSC, one copy presents in IR and another in SSC.

Table 7.5 Oligonucleotide repeats in *H. rosa-sinensis* and *H. syriacus*

Oligonucleotide repeats in <i>H. rosa-sinensis</i>						
S.No	Type	Location	Region	CDS/IGS/Intron	Size	Sequence
1.	F	LSC	<i>trnH-psbA</i>	IGS	35	TATGGTAATAAGAAATCTATGGTAATAAGAAA
2.	F	LSC	<i>trnH-psbA</i>	IGS	32	ATGTCAAGAAAAGAAAGACTACTAAATGTAAAGAAA
3.	P	LSC	<i>trnK-rps16/atpB-rbcL</i>	IGS	30	AATATAGAATATAAAAAAGACTATAAAAAA
4.	F	LSC	<i>trnK-rps16</i>	IGS	30	TATATTTTTATATGTTATATATTTTTTATATG
5.	F	LSC	<i>trnK-rps16</i>	IGS	34	TATTCTATATTCTATAGTAGCAATATTCTATATT
6.	F	LSC	<i>rps16-trnQ</i>	IGS	32	TCTATATTTTTATTAGAATAGAATTCTATATA
7.	F	LSC	<i>rps16-trnQ</i>	IGS	30	TAGAATAGAATAATATATAATAATAATAGAA
8.	P	LSC	<i>rps16-trnQ/atpF-atpH</i>	IGS	32	TAGAATAATATATAATAATAATAGAATAGAATAA
9.	C	LSC/IR	<i>rps16-trnQ/rps12-trnV</i>	IGS	30	GAATAATATATAATAATAAAAAAAAAAAAAA
10.	F	LSC	<i>trnG</i>	Intron	30	TTCTTGTTTTTTGAATTTATCCATCTCCAT
11.	P	LSC	<i>trnG-trnR</i>	IGS	30	TTGAATATTTATTAATCAATTCAACGATG
12.	P	LSC	<i>trnR-atpA</i>	IGS	40	TAGAAATATTAATAATTAGAATTCTAATTTTTATATTTCTA
13.	P	LSC	<i>atpF-atpH</i>	IGS	31	AAAAAAAAAAGAAGACTCGGTTCTTTTTTTTTT
14.	F	LSC	<i>atpH-atpI/ycf4-cemA</i>	IGS	32	AAACAAAAACAATATCAATATAAAGATAATAT
15.	F	LSC	<i>rpoC2</i>	CDS	33	AATCAATTGAAATGGAATAATAAATCTATTT
16.	F	LSC	<i>rpoC1/rpl16</i>	Intron	30	TCGGACATGAGAGTTTCCCTCTCATCCGGCTC
17.	F	LSC	<i>petN-psbM</i>	IGS	30	TTAACTATTGAGTCTTCATTTCGAATTTA
18.	F	LSC	<i>trnD-trnY/clp</i>	IGS/Intron	30	AAAATGAAAAAAAAAAGAGTCTTTTTTGT
19.	P	LSC/IR	<i>trnE-trnT/trnR-trnN</i>	IGS	31	ATACTTATATATATATAAATAGTATATATAAA
20.	F	LSC	<i>trnT-psbD</i>	IGS	30	TGAACATTTGTTTATTATAATGAATAATAAT
21.	C	LSC/SSC	<i>trnT-psbD/psaC-ndhE*</i>	IGS	36	TATTATAATGAATAATAATTGTTTATTATAATGAAT
22.	P	LSC/SSC	<i>psbZ-trnG/rpl32-trnL</i>	IGS	31	TATTATTTAATTTAAATATTATTTAAATAT
23.	P	LSC	<i>psbZ-trnG/atpB-rbcL</i>	IGS	30	TTAATTTAAATATTATTTAAATATATAAAT
24.	P	LSC	<i>psbZ-trnG/atpB-rbcL</i>	IGS	32	TTAATTTAAATATTATTTAAATATATAAATAA
25.	F	LSC	<i>psbZ-trnG/rpl33-rps18</i>	IGS	31	ATTTAAATATTATTTAAATATATAAATAAATAA
26.	F	LSC	<i>psbZ-trnG</i>	IGS	32	AATATATAAATAATAAATAAGATATACTAATAT
27.	F	LSC	<i>psbZ-trnG</i>	IGS	31	TTAATATTATTCTTAATATTATTCTTAATAA
28.	F	LSC	<i>psbZ-trnG</i>	IGS	30	GAAATATATTAATATTATATTGAAATATAT

29.	F	LSC	<i>psbZ-trnG/atpB-rbcL</i>	IGS	31	AATATATTAATATTATATTTATATAAATATT
30.	F	LSC	<i>psbZ-trnG</i>	IGS	39	TAATTAATACTAATAACAAATAGTATAAATTAATAATAA
31.	F	LSC	<i>psab/psaC</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATATCAGTCAGC CATA
32.	F	LSC/SSC	<i>ycf3/ndhA</i>	Intron	41	TCCAAAACCGTACATGAGATTTTCACCTCATACGGCTCCTC
33.	F	LSC/IR	<i>ycf3/rps12-trnV</i>	Intron/IGS	36	AACCGTACATGAGATTTTCACCTCATACGGCTCCTC
34.	F	LSC	<i>ycf3</i>	Intron	33	ATAAAGTAATATATATAAAAGTAATATATATA
35.	P	LSC/SSC	<i>psaA/psaC-ndhE</i>	CDS/IGS	32	GAGCAACTTTTAATTTATTATGAGCCCAAACGA
36.	P	LSC/SSC	<i>psaA/psaC-ndhE</i>	CDS/IGS	33	GAGCAACTTTTAATTTATTATGAGCCCAAACGA
37.	P	LSC/SSC	<i>trnT-trnL/ycf1^ψ-ndhF</i>	IGS	32	TTCTATTTTATTATATTAATTTATTATATTAT
38.	F	LSC	<i>trnT-trnL</i>	IGS	31	TATGTTATATTTATTCTAATTATCTCTATT
39.	F	LSC	<i>trnT-trnL/ndhC-trnV</i>	IGS	30	ATTCTAATTATCTCTATTTAGTTTATTCTAA
40.	C	LSC	<i>trnT-trnL/rps12-trnV</i>	IGS	30	AGAAGAAAAAAAAAAAAAAAAATAAAATAGGGAAAG
41.	F	LSC	<i>trnL-trnF</i>	IGS	31	TTGTGATATATTGTGATATATTGTGATATA
42.	P	LSC/SSC	<i>trnF-ndhJ/ndhF</i>	IGS/CDS	31	ATTTCTTTTTTCTATTTTGTATTTATTTT
43.	P	LSC	<i>ndhC-trnV</i>	IGS	40	TATAGTAATAGTATAAATCTATATTTATACTATTACTATA
44.	F	LSC	<i>ndhC-trnV</i>	IGS	39	TTTTTTATTTTATAGCTATAATATTTATATATATTGAATA
45.	R	LSC	<i>ndhC-trnV/rpl33-rps18</i>	IGS	30	ATATATTTATATATATTGAATATATATTTAA
46.	R	LSC/SSC	<i>ndhC-trnV/ycf1^ψ</i>	IGS/CDS	30	TTGATTTTTTTTCTTTTTTTTTTTTACCTTT
47.	F	LSC	<i>ndhC-trnV</i>	IGS	30	TCAACAAAGAATACAATAATTAGCTCAACAAA
48.	F	LSC	<i>ndhC-trnV</i>	IGS	32	ATTAATTTAATGAATATTTTAAAATTAATTTAA
49.	F	LSC	<i>atpB-rbcL</i>	IGS	32	ATTTCTTATTTCTTATTATTGATTTCTTATTATT
50.	F	LSC	<i>atpB-rbcL</i>	IGS	32	TTTATATTATATATATAATTTATATTAGATGT
51.	F	LSC/SSC	<i>atpB-rbcL/ycf1^ψ-ndhF</i>	IGS	32	AATATATTAATATTTATATTTAATTAATATT
52.	F	LSC	<i>atpB-rbcL</i>	IGS	42	ATATTTAATTAATATTTAAATTAATATTTAAATTAATATT
53.	F	LSC	<i>atpB-rbcL</i>	IGS	34	TTAAATTAATATTTAAATTAATATTTAATATTAT
54.	F	LSC	<i>atpB-rbcL/trnP-psaJ</i>	IGS	32	AATTAATATTTAAATTAATATTTAATATTA
55.	F	LSC	<i>atpB-rbcL/rpl16</i>	IGS/Intron	31	AATATTTAAATTAATATTTAATATTATTATT
56.	R	LSC	<i>petG-trnW/rpl33-rps18</i>	IGS	30	TATCTTTAATTTTATATCTTTAATTTTATA
57.	R	LSC	<i>petG-trnW</i>	IGS	31	TAATTTAATAATTTAAAATTTAATAATTTAAT
58.	F	LSC	<i>petG-trnW/trnP-psaJ</i>	IGS	30	TAATTTAATAATTTAAAATTTAATAATTTAAT
59.	P	LSC	<i>trnP-psaJ</i>	IGS	39	AATTTAATATTTAATATTTAATATTTAATATTTAATATT

90.	P	SSC	<i>ycf1^ψ-ndhF</i>	IGS	41	ATTAATATTAATATATTAATTAATAATATATTAATATTAAT
91.	F	SSC	<i>ycf1^ψ-ndhF/rpl32-trnL</i>	IGS	33	TTAAAATTAATAATAATCTAATAAATAATAT
92.	F	SSC	<i>ndhF-rpl32/rpl32-trnL</i>	IGS	31	AATATTAATATAATAAAAATAATAAAATATT
93.	R	SSC	<i>rpl32-trnL</i>	IGS	31	TAAATTAATATTAATAATAAAAATAATAAAT
94.	F	SSC	<i>rpl32-trnL</i>	IGS	34	TTAAATATTAATAATAAAAATAATAATAATATT
95.	P	SSC	<i>ndhD-psaC</i>	IGS	43	AAAAACCCGTGCTCAGAAAGATTTTTCTGAGCACGGGTTTTT
96.	R	SSC	<i>psaC-ndhE</i>	IGS	30	ATATTATTATATTATTAATATATTATTATA
97.	F	SSC	<i>psaC-ndhE</i>	IGS	39	ATATTATTATATTATTAATATATTATTATTATTAATA
98.	R	SSC	<i>psaC-ndhE</i>	IGS	34	ATTATTATATTATTAATATATTATTATTATTATA
99.	F	SSC	<i>psaC-ndhE</i>	IGS	35	ATATTATTAATATATTATTATTATTATTAATATATA
100.	F	SSC	<i>psaC-ndhE</i>	IGS	35	TAATATATAAATATTATTAATATAATATTACTTATT
Oligonucleotide repeats in <i>H. syriacus</i>						
1.	P	LSC	<i>trnK-rps16/atpB-rbcL</i>	IGS	31	AATAAAGAATATAAAAAAGACTATAAAAAA
2.	P	LSC/SSC	<i>trnK-rps16/ycf1^ψ-ndhF</i>	IGS	32	ATTTATTCTTTATTTACTTTATTTATTAAT
3.	F	LSC	<i>trnK-rps16</i>	IGS	30	TTTATTTACTTTATTTATTAATAATATTTTAT
4.	C	LSC	<i>trnK-rps16/rpl16</i>	IGS/Intron	30	ATTAATTATTTAATTTATTTCTTTATTTATT
5.	R	LSC	<i>trnK-rps16/trnT-trnL</i>	IGS	31	TATGTTATTTTATCTTATCTGTTTTTTTTAT
6.	R	LSC/SSC	<i>trnS-trnG/PsaC-ndhE</i>	IGS	30	TTATATAGATTATTATTAATATATAATTAT
7.	R	LSC	<i>trnS-trnG/rpl33-rps18</i>	IGS	31	ATATAGATTATTATTAATATATAATTATAT
8.	F	LSC	<i>trnS-trnG/PsaC-ndhE</i>	IGS	32	TATTATTAATATATAATTATATTATTAA
9.	C	LSC	<i>trnS-trnG</i>	IGS	30	TATTAATATATAATTATATTATTAATATAT
10.	F	LSC	<i>trnS-trnG/PsbZ-trnG</i>	IGS	30	TAATATATAATTATATTATTAATATATAAATA
11.	R	LSC	<i>trnS-trnG</i>	IGS	31	AATATATAAATATATAGATATATAAATATATA
12.	P	LSC	<i>trnG/trnG-trnR</i>	Intron/IGS	40	TTTACGTTTTGACTTTTTTTTTCTTTTGTTTCGGTTTTTT
13.	P	LSC	<i>trnG-trnR</i>	IGS	56	TATTTATTAATTCAATTCAACGATGCATTAGCATCGTTGAATTGAA TTAATAAATA
14.	F	LSC	<i>trnR-atpA</i>	IGS	32	TATCTATATAAATTATCTATATATCTATATAA
15.	R	LSC	<i>trnR-atpA</i>	IGS	33	TATCTATATAAATTATCTATATATCTATATAAAT
16.	P	LSC	<i>trnR-atpA</i>	IGS	30	ATAATTAATAGAAATTCTAATTTAATTAT
17.	F	LSC	<i>atpF-atpH</i>	IGS	48	TTTAAACATAAATAATAAAAAATTTTAAACATAAATAAAAA ATA
18.	F	LSC	<i>atpH-atpI/ycf4-cemA</i>	IGS	32	AAACAAAACAATATCAATATAAAGATAATAT

19.	P	LSC	<i>atpI-rps2/ycf1^ψ-ndhF</i>	IGS	30	TAATTTATTTATCTTATTATTAATTTATTA
20.	P	LSC	<i>atpI-rps2/rpI16</i>	IGS/Intron	30	TAATTTATTTATCTTATTATTAATTTATTATT
21.	F	LSC	<i>rpoC1/rpI16</i>	Intron	30	TCGGACATGAGAGTTTCCTCTCATCCGGCTC
22.	F	LSC	<i>rpoB-trnC</i>	IGS	31	ATAATCATAAAAATATATTCATAAAAATATATATA
23.	C	LSC	<i>rpoB-trnC/trnR-trnN</i>	IGS	30	AATCATAAAAATATATTCATAAAAATATATATA
24.	P	LSC	<i>trnE-trnT/trnR-trnN</i>	IGS	32	ATACTTATATATATAAATAGTATATATATAAATA
25.	F	LSC	<i>trnE-trnT</i>	IGS	30	TATATATATAAATAGTATATATATAAATAGTAT
26.	P	LSC	<i>trnE-trnT/trnR-trnN</i>	IGS	30	AATAGTATATATATAAATAGTATATATAAAA
27.	F	LSC	<i>trnT-psbD</i>	IGS	32	TTATAATGAATAATAATAAGTCGTCTCTTGAA
28.	R	LSC/SSC	<i>PsbZ-trnG/ycf1^ψ-ndhF</i>	IGS	30	TAAATGAAATAAATAGAAATAAATAGAATT
29.	P	LSC	<i>PsbZ-trnG</i>	IGS	53	AATTATTATATATTATTATTTAAATATTATTTAAATAATAATATAT AATAATT
30.	P	LSC	<i>PsbZ-trnG</i>	IGS	30	TATTATATATTATTATTTAAATATTATTTA
31.	P	LSC	<i>PsbZ-trnG/atpB-rbcL</i>	IGS	30	TATTTAAATATTATTTAAATAATAATATAT
32.	F	LSC	<i>PsbZ-trnG</i>	IGS	34	TTAAATAATAATATATAATAAATTAATAATAATATA
33.	R	LSC	<i>PsbZ-trnG</i>	IGS	38	AATAATAATATATAATAAATTAATAATAATATATAATA
34.	P	LSC/SSC	<i>PsbZ-trnG/rpI32-trnL</i>	IGS	30	AATAATAATATATAATAAATTAATAATAAAT
35.	C	LSC/SSC	<i>PsbZ-trnG/rpI32-trnL</i>	IGS	31	AATAATAATATATAATAAATTAATAATAAATATA
36.	P	LSC	<i>PsbZ-trnG/atpB-rbcL</i>	IGS	31	AATATATAATAAATTAATAATAATATATAAA
37.	C	LSC/SSC	<i>PsbZ-trnG/rpI32-trnL</i>	IGS	31	TAATAATTAATAATAATATATAATATTAAT
38.	P	LSC/SSC	<i>PsbZ-trnG/rpI32-trnL</i>	IGS	30	AATAATAATATATAATATTAATTAATATAT
39.	C	LSC/SSC	<i>PsbZ-trnG/psaC-ndhE</i>	IGS	31	ATAATATTATATTATATTATAAATTATATTAT
40.	C	LSC/SSC	<i>PsbZ-trnG/psaC-ndhE</i>	IGS	31	ATAATATTATATTATATTATAAATTATATTAT
41.	R	LSC/SSC	<i>PsbZ-trnG/rpI32-trnL</i>	IGS	31	ATATTATATTATAAATTATATTATTCTTAAT
42.	F	LSC/SSC	<i>PsbZ-trnG/psaC-ndhE</i>	IGS	31	AAATATATTAATATAAATATTATTAATTTA
43.	F	LSC/SSC	<i>PsbZ-trnG/psaC-ndhE</i>	IGS	37	AAATATATTAATATAAATATTATTAATTTATATTA
44.	F	LSC	<i>psaB/psaA</i>	CDS	49	AAAAGAAATGCAATAGCTAAATGATGGTGTGCAATATCAGTCAGC CATA
45.	F	LSC	<i>ycf3/ndhA</i>	Intron	41	TCCAAAACCGTACATGAGATTTTCACCTCATACGGCTCCTC
46.	F	LSC/IRB	<i>ycf3/ndhA/rps12-trnV</i>	Intron/IGS	36	AACCGTACATGAGATTTTCACCTCATACGGCTCCTC
47.	F	LSC	<i>ycf3</i>	Intron	42	GTAAAGTAAATAAAGTAAAGTAAATAAAGTAAATAAAGTAA
48.	F	LSC	<i>ycf3</i>	Intron	37	TAAAGTAAATAAAGTAAAGTAAATAAAGTAAATAAA

49.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	31	AAATAAAGTAAAGTAAATAAAGTAAATAAA
50.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	34	AAATAAAGTAAAGTAAATAAAGTAAATAAAGTAA
51.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	34	AAATAAAGTAAAGTAAATAAAGTAAATAAAGTAA
52.	F	LSC	<i>ycf3</i>	Intron	36	TAAAGTAAAGTAAATAAAGTAAATAAAGTAAAGTAA
53.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	32	TAAAGTAAAGTAAATAAAGTAAATAAAGTAA
54.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	34	AAAGTAAAGTAAATAAAGTAAATAAAGTAAAGTAA
55.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	31	TAAAGTAAATAAAGTAAATAAAGTAAAGTAA
56.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	32	TAAAGTAAATAAAGTAAATAAAGTAAAGTAA
57.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	31	AAATAAAGTAAATAAAGTAAAGTAAATAAA
58.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	34	AAATAAAGTAAATAAAGTAAAGTAAATAAAGTAA
59.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	32	TAAAGTAAATAAAGTAAAGTAAATAAAGTAA
60.	F	LSC/SSC	<i>ycf3/ycf1^ψ-ndhF</i>	Intron/IGS	30	TAAATAAAGTAAAGTAAATAAAGTAAATATA
61.	R	LSC	<i>rps4-trnT</i>	IGS	30	ATAGATATAATAATAATATATAATATATAA
62.	F	LSC	<i>trnL-trnF</i>	IGS	31	TTGTGATATATTGTGATATATTGTGATATA
63.	F	LSC	<i>trnM-atpE</i>	IGS	30	TGATTCTATTTAATCGGCAGAATCAAATG
64.	F	LSC	<i>atpB-rbcL</i>	IGS	30	TTATATTAGATGTTAGAATATTATATTAGA
65.	P	LSC	<i>atpB-rbcL</i>	IGS	37	ATTAATATTTAAATTAATATTTAATATTAATTTAA
66.	P	LSC	<i>atpB-rbcL</i>	IGS	38	TAATATTTAAATTAATATTTAATATTAATTTAATATTA
67.	P	LSC	<i>atpB-rbcL/rpI16</i>	IGS/Intron	34	TATTTAAATTAATATTTAATATTAATTTAATATT
68.	F	LSC/SSC	<i>atpB-rbcL/rpI32-trnL</i>	IGS	31	TTAATATTATTAATTATTATTATTTTTTAT
69.	F	LSC	<i>rbcL-accD</i>	IGS	34	ATTCTATTGTATTAGTAGACGAGATTTTACGAAA
70.	F	LSC	<i>petA-psbJ</i>	IGS	33	TTTTTTTTTTTTGATTTACTTATAATAGATAATAGA
71.	F	LSC	<i>petA-psbJ</i>	IGS	32	TTATAATAGATAATAGATATTTTTTTTTTTAAT
72.	R	LSC	<i>petA-psbJ</i>	IGS	39	ATAGATAATAGATATTTTTTTTTTTAATAGATAATAGATA
73.	R	LSC	<i>petA-psbJ</i>	IGS	34	TTTTTTTTTTAATAGATAATAGATATTTTTTTTTT
74.	F	LSC	<i>rpI33-rps18</i>	IGS	34	AAACCAAATCCTGTTTATTATATTCATTCTATAA
75.	R	LSC	<i>clpP/ccsA-ndhD</i>	Intron/IGS	31	TTTTTTTTTTCTTTCATCGAAAAAAGAAAAA
76.	P	LSC	<i>clpP/ycf1^ψ-ndhF</i>	Intron/IGS	32	ATTTACTTTTTTATTTTATTTATTTAATT
77.	F	LSC	<i>clpP-psbB</i>	IGS	30	TTTATCATATTTTATACATAGATAGAATTTA
78.	P	LSC	<i>psbT-psbN</i>	IGS	46	ATTGAAGTAATGAGCCTCCCAATATTGGGAGGCTCATTACTTCAAT
79.	C	LSC	<i>rpI16</i>	Intron	30	TTAATTTATTATAATTTATTATTTAATATTA

80.	P	LSC	<i>rpI16</i>	Intron	32	TTTATTATAATTTATTATTTAATATTAATAAT
81.	F	LSC	<i>rpI16</i>	Intron	30	TTAATATTAATAATAAATTAATATTAATAAT
82.	R	LSC	<i>rpI16</i>	Intron	37	TAATATTAATAATAAATTAATATTAATAATAAATTA
83.	F	LSC/SSC	<i>rpI16/ycf1^ψ</i>	Intron	31	TAATAAATTAATATTAATAATAAATTTAAA
84.	F	IR	<i>ycf2</i>	CDS	31	TCTTTTTGTCCAAGTTACTTCTCTTTTTGTC
85.	F	IR	<i>ycf2</i>	CDS	38	CGATATTGATGCTAGTGACGATATTGATGCTAGTGACG
86.	F	IR/SSC	<i>rps12-trnV/ndhA</i>	IGS/Intron	38	AACCGTACATGAGATTTTCACCTCATACGGCTCCTCGT
87.	F	IR	<i>rps12-trnV</i>	IGS	33	TCTTTCTATTATATTAGTCTTTCTATTATATT
88.	F	IR	<i>rrn4.5-rrn5</i>	IGS	34	CATTGTTCAACTCTTTGACAACACGAAAAATCCA
89.	P	IR	<i>rrn5-trnR</i>	IGS	43	TCATTCTTATTACTTTTTCAATATGAAAAAGTAATAAGAATGA
90.	F	IR	<i>ycf1/ycf1^ψ-ndhF</i>	IGS	30	TAATAAATTAATAAATTAATAAATAAATTA
91.	F	SSC	<i>ycf1/ycf1^ψ-ndhF</i>	IGS	33	AATAAATTAATAAATTAATAAATAAATTAATA
92.	F	SSC	<i>ycf1/ycf1^ψ-ndhF</i>	IGS	46	AAATTAATAAATTAATAAATAAATTAATAAATTAATAAATAAATAA
93.	R	SSC	<i>ycf1^ψ-ndhF</i>	IGS	33	AAATTAATAAATTAATAAATTAATAAATAAAT
94.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	40	TAAATAAATAAATAAATAAATAAATAAATAAAGTAAAT
95.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	34	AAATAAATAAATAAATAAATAAATAAATAAAGTAA
96.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	36	AAATTAATAAATAAATAAATAAATAAATAAAGTAAATTA
97.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	41	AATAAATAAATAAATAAATAAATAAATAAAGTAAATTAATAAATAAAGTAA
98.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	40	TAAATTAATAAATAAATAAATAAATAAATAAAGTAAATTAATAAATAAAGTAAAT
99.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	36	AAATTAATAAATAAATAAATAAATAAATAAAGTAAATTAATAAATAAAGTAA
100.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	31	TAAATAAATAAATAAATAAATAAATAAATAAAGTAA
101.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	34	AAATAAATAAATAAATAAATAAATAAATAAAGTAA
102.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	34	TAAATTAATAAATAAATAAATAAATAAATAAAGTAAATAA
103.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	44	AAATTAATAAATAAATAAATAAATAAATAAAGTAAATAAATAAATAAATAAATAAATAAATAA
104.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	33	AATAAATAAATAAATAAATAAATAAATAAAGTAAATAAATAA
105.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	43	AAATAAATAAATAAATAAATAAATAAATAAAGTAAATAAATAAATAAATAAATAAATAAATAA
106.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	33	AAATTAATAAATAAATAAATAAATAAATAAAGTAAATAAATAAATAA
107.	F	SSC	<i>ycf1^ψ-ndhF</i>	IGS	30	AAATAAATAAATAAATAAATAAATAAATAAAGTAAATAA
108.	F	SSC	<i>ycf1-ndhF/ndhF</i>	IGS/CDS	31	AAATAAATAAATAAATAAATAAATAAATAA
109.	F	SSC	<i>ndhF-rpI32</i>	IGS	30	TCTATTTACATTTATATTTCCATATAATTA
110.	F	SSC	<i>rpI32-trnL</i>	IGS	41	TTAAAATTATTAATTATTATTAAAATTATTAATTATTATTATA

111.	R	SSC	<i>rpI32-trnL</i>	IGS	38	TAAAATTATTAATTATTATTATATTATTATTATTATTA
112.	R	SSC	<i>rpI32-trnL</i>	IGS	32	TAATTATTATTATATTATTATTATTATTATTATA
113.	F	SSC	<i>rpI32-trnL</i>	IGS	30	ATTATTATTATATTATTATTATTATTATTATAA
114.	F	SSC	<i>rpI32-trnL</i>	IGS	78	ATAATATCCTTTCTTGTGTTTTCAACCCACTTGAGAAAAACCCTTTC TAATTTAGTGGATATAGAAATAATGTATCAAT
115.	F	SSC	<i>ccsA-ndhD/ndhA</i>	IGS/Intron	31	AAGAAAAAAAAAAAAAAAAAGACTTCTTTCTTTTTT
116.	F	SSC	<i>PsaC-ndhE</i>	IGS	31	AATTAATATTATTAATATATAAATATTATTA
117.	F	SSC	<i>PsaC-ndhE</i>	IGS	30	TAAATATTATTAATATATAAATATTATTATA
118.	R	SSC	<i>PsaC-ndhE</i>	IGS	34	TAAATATTATTAATATATAAATATTATTATAAATAT
119.	F	SSC	<i>PsaC-ndhE</i>	IGS	36	ATATAAATATTATTATAAATATTATTAATATAA
120.	F	SSC	<i>PsaC-ndhE</i>	IGS	35	TATAAATATTATTATAAATATTATTAATATAA
121.	F	SSC	<i>PsaC-ndhE</i>	IGS	30	TATTATTATAAATATTATTAATATAAATATTA
122.	F	SSC	<i>PsaC-ndhE</i>	IGS	31	TATTATTATAAATATTATTAATATAAATATTA
123.	R	SSC	<i>PsaC-ndhE</i>	IGS	32	ATTATTATAAATATTATTAATATAAATATTAT
124.	R	SSC	<i>PsaC-ndhE</i>	IGS	34	ATTATTATAAATATTATTAATATAAATATTATTA
125.	F	SSC	<i>PsaC-ndhE</i>	IGS	38	TATAAATATTATTAATATAAATATTATTAATATAAATAT
126.	R	SSC	<i>PsaC-ndhE</i>	IGS	34	TATTATTAATATAAATATTATTAATATAAATATAA
127.	F	SSC	<i>PsaC-ndhE</i>	IGS	32	TAATATAAATATTATTAATATAAATATAAATATTA
128.	F	SSC	<i>PsaC-ndhE</i>	IGS	34	TATTAATATAAATATAAATATTATTAATATAAATATTA
129.	R	SSC	<i>PsaC-ndhE</i>	IGS	32	TATAATATAAATATTATTAATATAAATATTAT
130.	F	SSC	<i>ndhG-ndhI</i>	IGS	30	TTAATAGAATAGGAGGTTTACTTACTCAGT

7.3.6 Analysis of substitutions and InDels in genus *Hibiscus*

We determined 1612 Substitutions between *H. rosa-sinensis* and *H. syriacus* species. We found high conversion for A/G and C/T as compared to other SNPs (Table 7.6). The ratio of transition to transversion (Ts/Tv) was 0.74. LSC contained 1202, IR contained 35, and SSC contained 375 SNPs. Insertions and Deletions (InDels) were also analysed using DnaSP in each part of chloroplast genome. In total, 500 InDels were found in which most of InDels were found in LSC (413) followed by SSC (69) whereas IR contained (18) least InDels (Table 7.7).

Table 7.6 Types and distribution of SNPs in *Hibiscus*

Types	Number
A/G	326
C/T	357
A/C	260
C/G	140
G/T	300
A/T	229
Total	1612
Regions wise distribution	
LSC	1202
IR	35
SSC	375

Table 7.7 Distribution of InDels in *Hibiscus*

Regions	Number	InDel length (bp)	InDel average length
LSC	413	3437	8.32
IR	18	432	24
SSC	69	952	13.79

7.3.7 Mutational hotspots in *Hibiscus* species

The comparative analysis of the different regions in chloroplast genomes of genus *Hibiscus* revealed high level of variation (Figure 7.3). The nucleotide diversity (π) ranged from 0.001 (*trnL-GAU* intron) to 0.0933 (*ccsA-ndhD*). All the tRNA genes showed $\pi = 0$ except *trnR-UCU*, *trnT-UGU*, and *trnL-UAA* had π 0.0139, 0.0137, 0.0112, respectively; and the average value of nucleotide diversity was 0.00146 for tRNAs. We found IGS with high average nucleotide diversity ($\pi = 0.0284$), followed by intronic region ($\pi = 0.01591$), and lowest was found for coding sequences ($\pi = 0.006$). All thirty highly divergence regions of the chloroplast genome with alignment length greater than or equal to 200 bp belonged to IGS regions and none of the intronic or coding sequences were included in this list (Table 7.8). The larger regions identified with the highest polymorphism was *atpB-rbcL* with $\pi = 0.050$ and alignment

size 1,159 bp. The previously employed gene or region for resolving of taxonomical discrepancies and barcoding of species showed the nucleotide diversity: *matK* (0.0178), *rpl16* intron (0.0152), *rbcL* (0.0184) and *ndhF* (0.0185); that was quite lower than the suggested 30 highly divergence regions. The *ycf1* gene possessed nucleotide diversity ($\pi=0.011$) and contained 77 substitutions and 10 InDels.

Table 7.8 Mutational hotspots between *H. rosa-sinensis* and *H. syriacus*

S.No	Region	Nucleotide diversity	Total Number of mutations	Region length
1.	<i>psbZ-trnG-GCC</i>	0.110092	72	654
2.	<i>trnK-UUU-rps16</i>	0.065817	62	942
3.	<i>trnD-GUC-trnY-GUA</i>	0.063584	33	519
4.	<i>trnW-CCA-trnP-UGG</i>	0.062802	13	207
5.	<i>rpl33-rps18</i>	0.062201	13	209
6.	<i>petG-trnW-CCA</i>	0.054455	11	202
7.	<i>trnS-GCU-trnG-UCC</i>	0.053156	48	903
8.	<i>trnH-psbA</i>	0.050881	26	511
9.	<i>atpB-rbcL</i>	0.050043	58	1159
10.	<i>rpl32-trnL-UAG</i>	0.04918	57	1159
11.	<i>petD-rpoA</i>	0.04918	12	244
12.	<i>ccsA-ndhD</i>	0.046875	15	320
13.	<i>rps4-trnT-UGU</i>	0.046763	26	556
14.	<i>ndhF-rpl32</i>	0.045793	43	939
15.	<i>psbA-trnK-UUU</i>	0.044521	13	292
16.	<i>psaC-ndhE</i>	0.040936	14	342
17.	<i>trnF-GAA-ndhJ</i>	0.04047	31	766
18.	<i>petN-psbM</i>	0.038005	48	1263
19.	<i>trnP-UGG-psaJ</i>	0.035971	15	417
20.	<i>infA-rps8</i>	0.035714	8	224
21.	<i>ndhG-ndhI</i>	0.035011	16	457
22.	<i>trnT-UGU-trnL-UAA</i>	0.033508	32	955
23.	<i>rps16-trnQ-UUG</i>	0.033333	21	630
24.	<i>atpH-atpI</i>	0.032258	41	1271
25.	<i>rpoB-trnC-GCA</i>	0.031397	40	1274
26.	<i>ycf4-cemA</i>	0.030667	23	750
27.	<i>petA-psbJ</i>	0.030392	31	1020
28.	<i>atpF-atpH</i>	0.029008	19	655
29.	<i>rps15-ycf1</i>	0.028916	12	415
30.	<i>trnQ-UUG-psbK</i>	0.028736	10	348

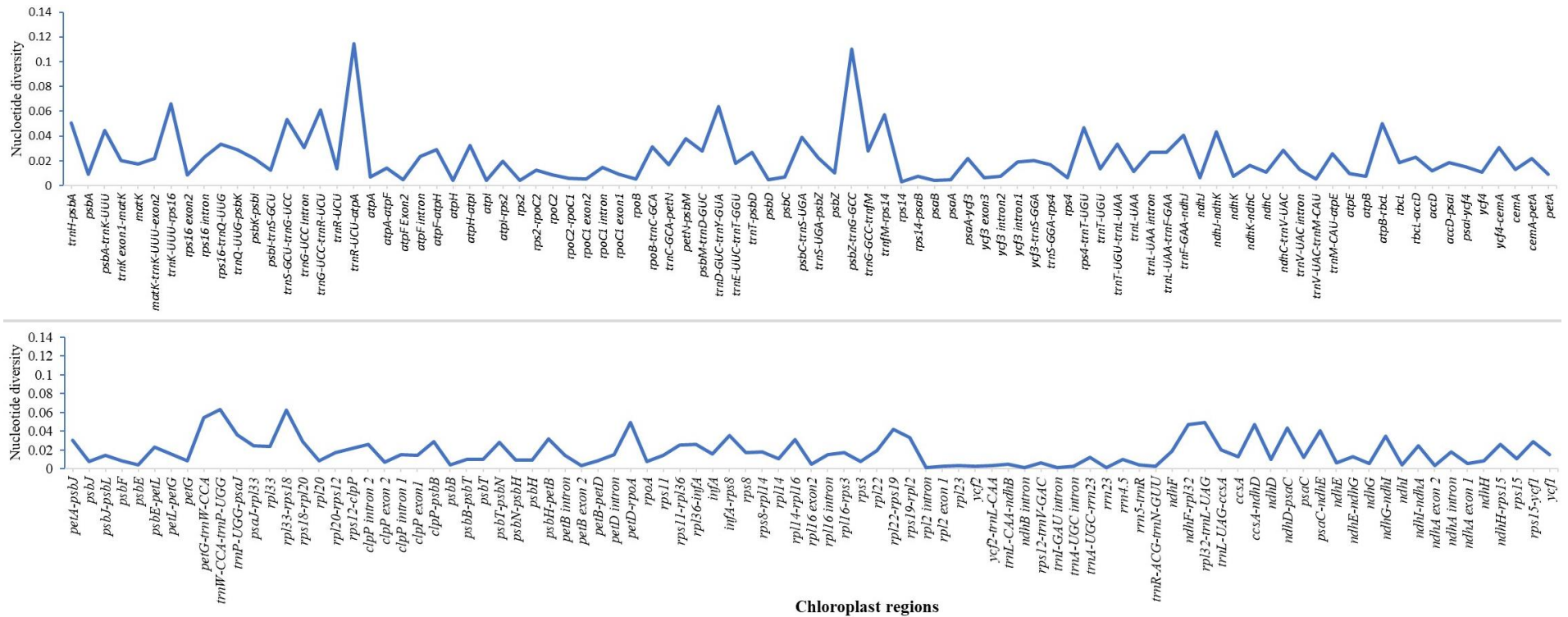


Figure 7.3 Nucleotide diversity of chloroplast genome region between *Hibiscus* species.

X-axis represents chloroplast genome regions whereas Y-axis represents nucleotide diversity of these regions.

7.4 Conclusion

Chloroplast genome of *Hibiscus rosa-sinensis* and *H. syriacus* showed similar genes content and organisation. The comparative analyses of both genomes revealed high similarities in codon usage, amino acid frequency, RNA editing sites, simple sequence repeats and oligonucleotide repeats. The analyses of substitutions and InDels revealed high divergence in LSC regions in comparison to SSC and IR regions. The mutational hotspots identified in the current study might be helpful for the development of authentic, robust, reproduceable and cost-effective markers to infer phylogeny in genus *Hibiscus* and resolve taxonomic discrepancies

Chapter 8

Phylogenetic inference in the family

Malvaceae

8.1 Introduction

The plant family Malvaceae is comprising 244 genera and 4225 species (Christenhusz and Byng, 2016). Plants of Malvaceae are widely distributed in tropical to temperate regions of the world and possess plastic morphology (Xu and Deng, 2017), which give rise to taxonomic discrepancies at genus and family levels (Alverson *et al.*, 1999; Carvalho-Sobrinho *et al.*, 2016; Pfeil *et al.*, 2002; Tate *et al.*, 2005). This family has been divided into nine subfamilies which are Brownlowioideae, Bombacoideae, Byttnerioideae, Dombeyoideae, Grewioideae, Helicteroideae, Malvoideae, Tilioideae and Sterculioideae (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Judd and Manchester, 1997; Xu and Deng, 2017).

Several studies have focused on taxonomical position of the family Malvaceae. Due to plastic morphology, different taxonomic classifications have been suggested for the family Malvaceae and the former Malvaceae s.s (recent Malvoideae) was combined with different other families. These classifications have been shown in table 1.1 (in Chapter 1). Four families have been considered core Malvales including Tiliaceae, Sterculiaceae, Bombacaceae and Malvaceae s.s, and merged into an extended family Malvaceae as molecular based studies showed the monophyletic position of the species of these families. (APG, 2003, 1998; Bayer *et al.*, 1999). The extended family Malvaceae was also subdivided into nine subfamilies based on phylogenetic inference of chloroplast genome sequences of *atpB*, *rbcL* and *ndhF* (Alverson *et al.*, 1999; Bayer *et al.*, 1999). This classification has been adopted by various researchers (Baum *et al.*, 2004; Bayer and Kubitzki, 2003; Carvalho-Sobrinho *et al.*, 2016; Duarte *et al.*, 2011; Perveen *et al.*, 2004; Tate *et al.*, 2005). A recent study also support the nine subfamilies classification (Richardson *et al.*, 2015). However, at subfamily level phylogenetic relationships among these subfamilies are still inconclusive and different classifications have been reported based on molecular markers from few loci of chloroplast genome (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Nyffeler *et al.*, 2005; Richardson *et al.*, 2015). Some researchers suggested different classification for the family Malvaceae. Thorne (2000) considered Bombacaceae and Sterculiaceae under Malvaceae s.l and Byttneriaceae and Tiliaceae (with some restriction) as separate families whereas Hinsley (2006) also suggested other four alternate classifications as mentioned in introduction of this thesis.

This data showed that still discrepancies exist in phylogenetic relationships of the family Malvaceae. The previous studies of family Malvaceae were based on few genes as mentioned in the details. Recently, the complete chloroplast genomes or coding sequences were used for reconstruction of high resolution phylogenetic tree (Cremen *et al.*, 2019; Du *et al.*, 2017; Henriquez *et al.*, 2014). This approach was applied for the phylogenetic inference in many

plant families. The phylogenetic inference of family Araceae based on 70 protein-coding genes of chloroplast provided new insight into phylogenetic relationship of many clades and subfamilies of the family Araceae (Henriquez *et al.*, 2014). The high resolution phylogenetic tree of order Bryopsidales was reconstructed based on complete chloroplast genome sequences, which provided insight into evolution of many families of the Bryopsidales and proposed a new classification for families Rhipiliaceae, Udoteaceae and Pseudocodiaceae (Cremen *et al.*, 2019). The *Dipteronia* and *Acer* are two sister genera of family Sapindaceae, but their phylogenetic position was not confirmed. The phylogenetic analyses based on complete chloroplast genome data not only well resolved the phylogeny of these two genera but also identified *Dipteronia* species among the East Asian flora (Feng *et al.*, 2019). Family Ranunculaceae is considered among the early diverging eudicots and consists of 14 tribes from which the phylogeny of 11 tribes was controversial (Zhai *et al.*, 2019). The phylogenetic analyses of 35 species from 31 genera of 14 tribes resolved discrepancies in the phylogeny. This data also revealed the inappropriateness of the previously used morphological characters in phylogenetic inference due to parallel, convergent or even reversal evolution (Zhai *et al.*, 2019).

The phenomenon of contraction and expansion of inverted repeat (IR) regions is given in detail in chapter 2. Some studies revealed the role of IRs contraction and in determination of the resemblance of plant species (Iram *et al.*, 2019; Liu *et al.*, 2018). These authors included closely related species in the comparison. Therefore, these authors suggested further evaluation of their results in the diverse species.

In current study, we aimed to infer phylogeny of family Malvaceae based on complete chloroplast genomes sequences to reconstruct high resolution phylogenetic tree. Moreover, comparison of the relationship among species determined from phylogenetic tree with the phenomena of IR contraction and expansion tested its role in determination of resemblance in plant species.

8.2 Materials and methods

8.2.1 Reconstruction of phylogenetic tree

We sequenced chloroplast genomes of four species of family Malvaceae. We also *de novo* assembled chloroplast genomes of four species of Malvaceae from SRA data which were downloaded from NCBI. Further, we downloaded chloroplast genomes sequences of thirty-two species of four families of order Malvales and used in the inferring of phylogeny of family Malvaceae (Table 8.1). Among these thirty-two species, twenty-two species belonged to family

Malvaceae whereas ten species belonged to three other families of order Malvales included Bixaceae, Dipterocarpaceae, and Thymelaeaceae.

We reconstructed maximum likelihood phylogenetic tree by different approaches to infer phylogeny of the family Malvaceae and to get insight into the efficacy of different regions of chloroplast genome in inferring of phylogeny.

1. The phylogenetic tree was reconstructed for order Malvales and family Malvaceae separately based on seventy protein coding genes. We extracted seventy protein coding genes and concatenated them in the Geneious R8.1 (Kearse *et al.*, 2012). The concatenated sequences were multiple align using MAFFT (Multiple Alignments using Fast Fourier Transform) (Kato *et al.*, 2005). All the InDels were removed from the alignment manually to reconstruct the phylogenetic tree based on only substitutions.

2. To check the efficacy of different regions of chloroplast genome, the phylogeny was inferred of family Malvaceae separately with complete chloroplast genome (IRa removed), LSC region, SSC region, and IR region. The sequence of each region of chloroplast genome was aligned through MAFFT and all InDels were removed from the sequence through Geneious R8.1 (Kearse *et al.*, 2012) to reconstruct phylogenetic tree based on only substitutions. The maximum likelihood tree was reconstructed based on all the regions of chloroplast by using the IQ-tree program (<http://iqtree.cibiv.univie.ac.at>; accessed 14 June 2019) by default parameters (Hoang *et al.*, 2018; Kalyaanamoorthy *et al.*, 2017; Nguyen *et al.*, 2015). The default parameters included 1000 iterations, 1000 replicates and best-fit model selection. TreeDyn was used for further enhancement of phylogenetic tree analyses (Chevenet *et al.*, 2006; Dereeper *et al.*, 2008).

8.3 Results

8.3.1 Phylogenetic analyses based on protein coding genes

Maximum likelihood phylogenetic tree based on protein coding sequences was reconstructed with best fit model TVM+F+I+G4 among four families of order Malvales: Malvaceae, Bixaceae, Dipterocarpaceae, and Thymelaeaceae. Phylogeny of these families was inferred based on 53,340 nucleotide sites in which 43,430 (81.42%) were similar in all species, whereas 9910 (13.29%) varied among species. In these variable sites, 6727 were parsimony informative sites. The phylogenetic tree showed that species of all families was well resolved with high bootstrapping support value and showed monophyletic position. The species of Thymelaeaceae including *Daphne kiusiana*, *Aquilaria sinensis*, *Aquilaria malaccensis* existed on the basal of

the tree among four Malvales species, whereas the species of the family Bixaceae acted as the first outgroup of family Malvaceae (Figure 8.1).

Table 8.1 Accessions of species used in phylogenetic tree

S.No	Species	Family	Accession
1.	<i>Firmiana pulcherrima</i>	Malvaceae	NC_036395
2.	<i>Firmiana colorata</i>	Malvaceae	BK010724
3.	<i>Firmiana major</i>	Malvaceae	NC_037242
4.	<i>Gossypium arboreum</i>	Malvaceae	HQ325740
5.	<i>Gossypium barbadense</i>	Malvaceae	HQ901198
6.	<i>Gossypium hirsutum</i>	Malvaceae	HQ901196
7.	<i>Gossypium thurberi</i>	Malvaceae	GU907100
8.	<i>Gossypium herbaceum</i>	Malvaceae	HQ325742
9.	<i>Heritiera angustata</i>	Malvaceae	NC_038057
10.	<i>Heritiera parvifolia</i>	Malvaceae	NC_038057
11.	<i>Talipariti hamabo</i>	Malvaceae	NC_030195
12.	<i>Tilia oliveri</i>	Malvaceae	KT894774
13.	<i>Tilia paucicostata</i>	Malvaceae	KT894775
14.	<i>Tilia amurensis</i>	Malvaceae	KT894772
15.	<i>Theobroma cacao</i>	Malvaceae	HQ336404
16.	<i>Theobroma grandiflorum</i>	Malvaceae	JQ228388
17.	<i>Pterospermum truncatolobatum</i>	Malvaceae	BK010725
18.	<i>Pterospermum kingtungense</i>	Malvaceae	MH606238
19.	<i>Malva parviflora</i>	Malvaceae	MK860036
20.	<i>Durio zibethinus</i>	Malvaceae	MG138151
21.	<i>Bombax ceiba</i>	Malvaceae	MG569974
22.	<i>Althaea officinalis</i>	Malvaceae	NC_034701
23.	<i>Abelmoschus esculentus</i>	Malvaceae	NC_035234
24.	<i>Urena procumbens</i>	Malvaceae	BK010727
25.	<i>Sterculia monosperma</i>	Malvaceae	BK010726
26.	<i>Reevesia thyrsoidea</i>	Malvaceae	MH939148
27.	<i>Hibiscus rosa-sinensis</i>	Malvaceae	MK382984
28.	<i>Hibiscus mutabilis</i>	Malvaceae	MK820657
29.	<i>Hibiscus syriacus</i>	Malvaceae	KR259989
30.	<i>Malvastrum coromandelianum</i>	Malvaceae	MK860037
31.	<i>Daphne kiusiana</i>	Thymelaeaceae	NC_035896
32.	<i>Bixa orellana</i>	Bixaceae	MH025909
33.	<i>Aquilaria malaccensis</i>	Thymelaeaceae	NC_041117
34.	<i>Aquilaria sinensis</i>	Thymelaeaceae	KT148967
35.	<i>Vatica mangachapoi</i>	Dipterocarpaceae	NC_041485
36.	<i>Shorea zeylanica</i>	Dipterocarpaceae	NC_040965
37.	<i>Shorea pachyphylla</i>	Dipterocarpaceae	NC_040966
38.	<i>Neobalanocarpus heimii</i>	Dipterocarpaceae	NC_041191
39.	<i>Parashorea macrophylla</i>	Dipterocarpaceae	MH791330
40.	<i>Hopea dryobalanoides</i>	Dipterocarpaceae	MH791329

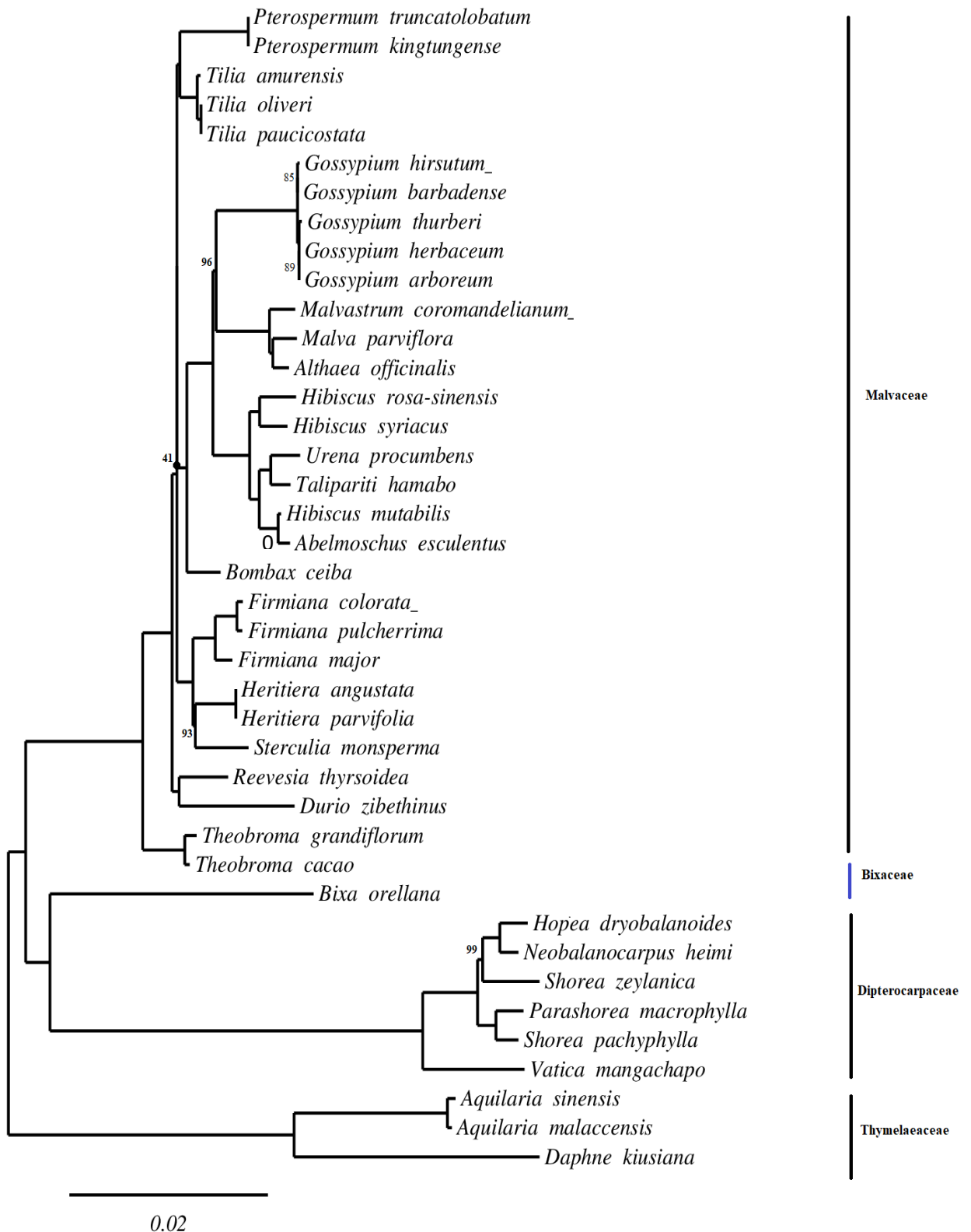


Figure 8.1 The phylogenetic analysis of four families of order Malvales based on protein coding genes. All the nodes with bootstrapping equal to 100 were not mentioned. Family Malvaceae is present on the top of the phylogenetic tree and family Bixaceae act as an outgroup to family Malvaceae. The species of family Thymelaeaceae are present at the basal of tree.

The phylogenetic tree showed the species of family Malvaceae present on the top of the tree and well resolved from the species of others Malvales families with high bootstrapping support of 100.

The maximum likelihood phylogenetic tree was reconstructed based on seventy protein coding genes among the species of family Malvaceae without including species from other families. We included thirty species of family Malvaceae from seventeen genera that belonged to seven subfamilies of family Malvaceae (Figure 8.2). The phylogenetic tree was reconstructed with best fit model TVM+F+I+G4. The alignment included 53,340 nucleotide sites in which 1181 were distinct sites and 2615 were parsimony informative sites. The phylogenetic tree revealed high bootstrapping support for the resolution of subfamilies of Malvaceae (Figure 8.2). Species of all the subfamilies were well resolved based on coding sequences with the bootstrapping support of 100. The subfamily Malvoideae included 14 species from 8 genera and was present at the top of the tree. All the species of 8 genera were well resolved with bootstrapping support of 100. However, the genus *Hibiscus* showed polyphyletic position, and we found *Hibiscus mutabilis* as sister taxa to *Abelmoschus esculentus* with 100 bootstrapping support. *Bombax ceiba* of subfamily Bombacoideae was well resolved from the other species of other subfamilies and appeared sister to subfamily Malvoideae with bootstrapping support of 100. The subfamilies Tilioideae and Dombeyoideae were also observed as sister groups with bootstrapping support of 100. We found Sterculioideae as sister group to the subfamilies Tilioideae and Dombeyoideae, but with the bootstrapping support 47. The subfamily Sterculioideae included five species from three genera: *Firmiana*, *Heritiera* and *Sterculia*. Among these genera, *Heritiera* was sister to genus *Sterculia* that shared a common node with genus *Firmiana*. The species of two subfamilies including Helicteroideae and Byttnerioideae were also well resolved with the bootstrapping value of 100. These two subfamilies also showed sister group relationship. The Helicteroideae included two species *Reevesia thyrsoidea* and *Durio zibethinus*, whereas Byttnerioideae included *Theobroma cacao* and *Theobroma grandiflorum* and existed at basal to phylogeny of family Malvaceae. The reconstructed phylogenetic tree of family Malvaceae based on complete chloroplast genomes sequences also support the phylogenetic relationship determined on the basis of coding sequences.

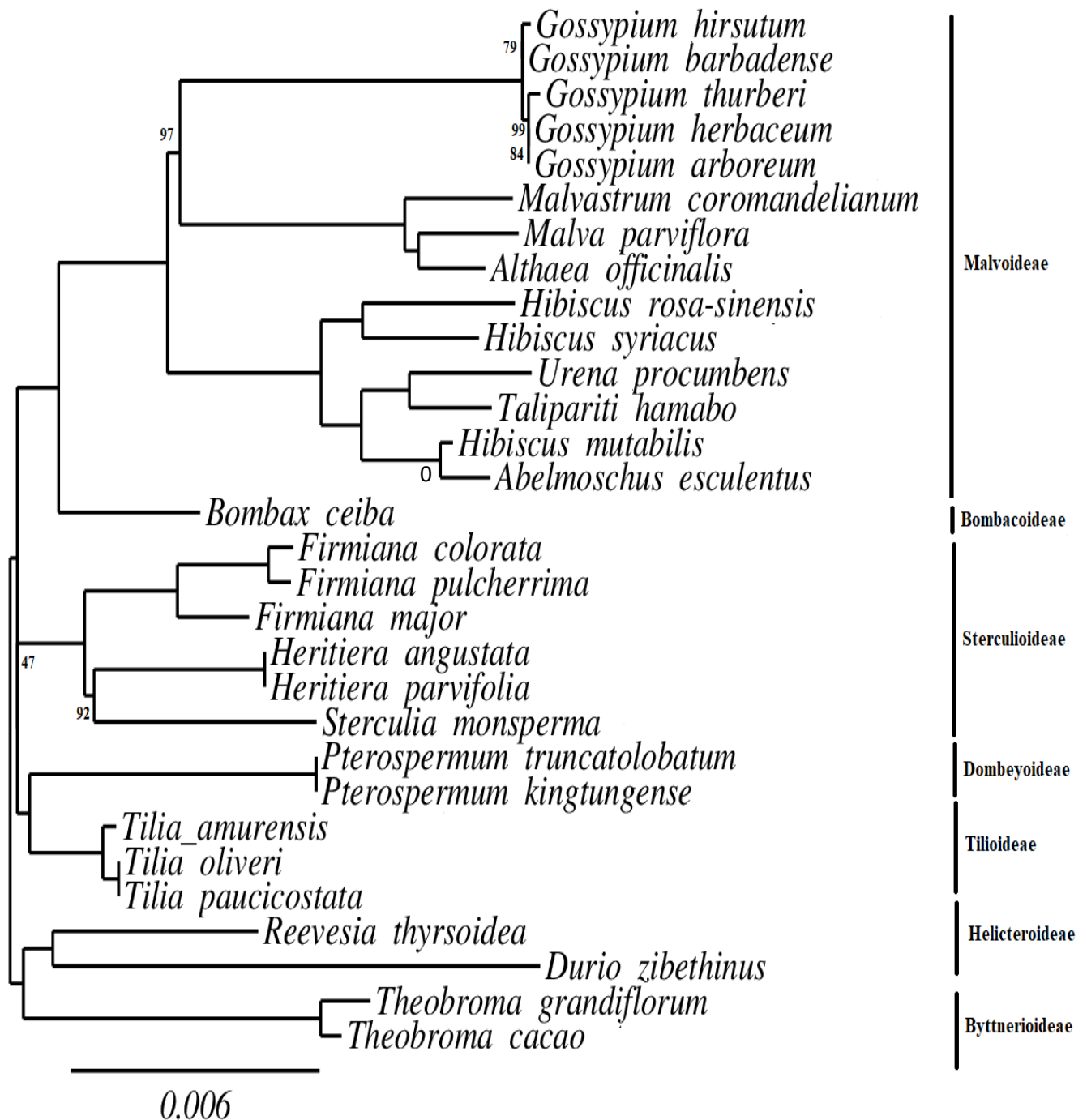


Figure 8.2 Phylogenetic analysis of family Malvaceae based on protein coding sequences.

All the nodes with bootstrapping equal to 100 were not mentioned. Species of subfamily Malvoideae were present at the top of the tree, whereas the family Byttnerioideae was present at the basal of tree. *Theobroma cacao* was present basal to all species.

8.3.2 Role of IRs contraction and expansion in determining phylogenetic relationship

In our study, we compared the IR contraction and expansion and positions of the genes at the borders of LSC, SSC and IRs regions (Figure 3.2). Here, we compared the IRs contraction and expansion with the phylogeny of the family Malvaceae. Our study in wide species revealed that this phenomenon can not be used for inferring of phylogeny. In phylogenetic tree (Figure 8.2), the *Abelmoschus esculentus* was closely related to other species of subfamily Malvoideae but IR expansion of *Abelmoschus esculentus* leads to complete duplication of *rps19*, *rpl22*, *rps3*, and a pseudogene of *rpl16* at 5' end in the IRs regions. In subfamily Helicteroideae, the contraction of IRs regions led to existence of single copy of *rpl2* and *rpl23* in *Durio zibenthus* but in another species, *Reevesia Thyrsoides* of the subfamily Helicteroideae, these genes were present in duplicate. Comparison of phylogenetic tree (Figure 8.2) and position of the gene at the borders of LSC, SSC and IRs regions (Figure 3.2 in chapter 3) also did not suggest the link of IR contraction and expansion in phylogeny. In case of the phenomena of IR contraction and expansion, the position of the genes at each border was unevenly related in species. The species that were found closely related in phylogenetic tree showed high divergence when compared based on the genes position at junctions, whereas species those showed high phylogenetic distance were found more closely related when analysed based on IR contraction and expansion. Moreover, at different junctions, we observed difference in positions of genes for which phylogenetic relationships could not be defined. The comparative analyses of figure 8.2 and figure 3.2 provide clear picture about the unsuitability of IR contraction and expansion in inferring of phylogeny.

8.3.3 Efficacy of complete chloroplast genome and LSC, SSC and IR regions in inferring of phylogeny

We also inferred phylogeny of family Malvaceae based on complete chloroplast genome, LSC, SSC, and IR regions. We investigated efficacy of all these parts in inferring of phylogeny. We reconstructed phylogenetic tree for 29 species of family Malvaceae based on complete chloroplast genome with best fit model TVM+F+I+G4. One species *Herritera angustata* was not included in the phylogeny due to uneven IR contraction and expansion which lead to problematic alignment. The phylogenetic tree was reconstructed based on 111,189 nucleotide sites in which about 97,887 (88.03%) sites were constant among species. In the variable sites, 7567 nucleotide sites were parsimony informative sites and 2971 nucleotide sites showed distinct pattern among these species. The result of complete chloroplast genome was similar to the phylogenetic tree of coding sequences. The phylogenetic tree based on complete chloroplast

genome showed high resolving efficacy and these species were well resolved in comparison to coding sequences (Figure 8.3).

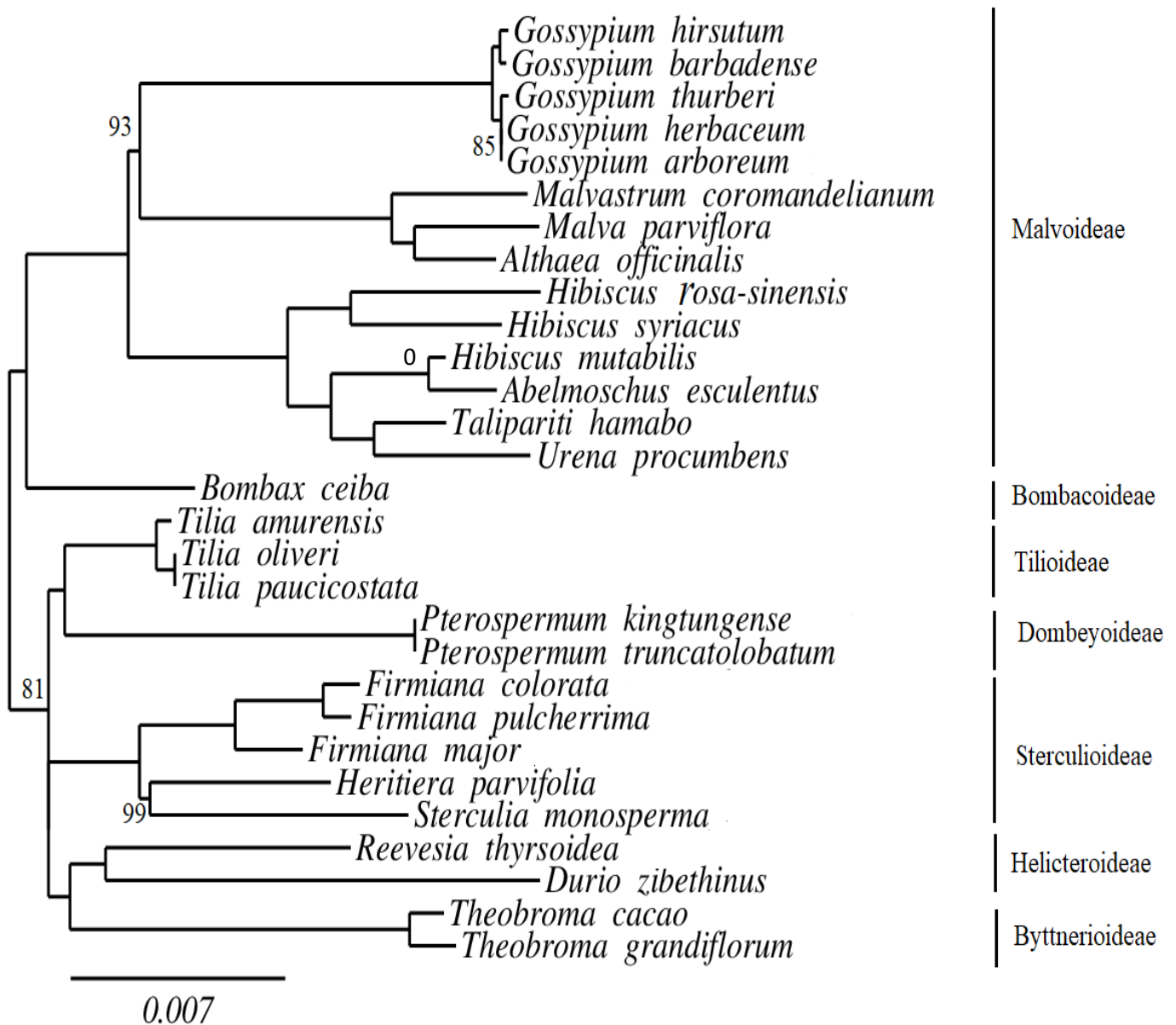


Figure 8.3 Phylogenetic tree based on complete chloroplast genome. The node of 100 bootstrapping value has been omitted.

The phylogenetic tree was also reconstructed based on the LSC region of chloroplast genome. The phylogenetic tree based on LSC region contained 73,586 nucleotide sites in which 64,371 (86.25%) sites were constant in all species. The 5910 nucleotide sites were parsimony informative whereas 2504 nucleotide sites showed distinct site pattern. The phylogenetic tree was reconstructed with best fit model TVM+F+I+G4. The phylogenetic tree of complete chloroplast genome and LSC region was reconstructed with same model but the reconstruction based on LSC regions could not well resolve the species of family Malvaceae as compared to

complete chloroplast genome. The phylogenetic relationship based on LSC regions is shown in figure 8.4.

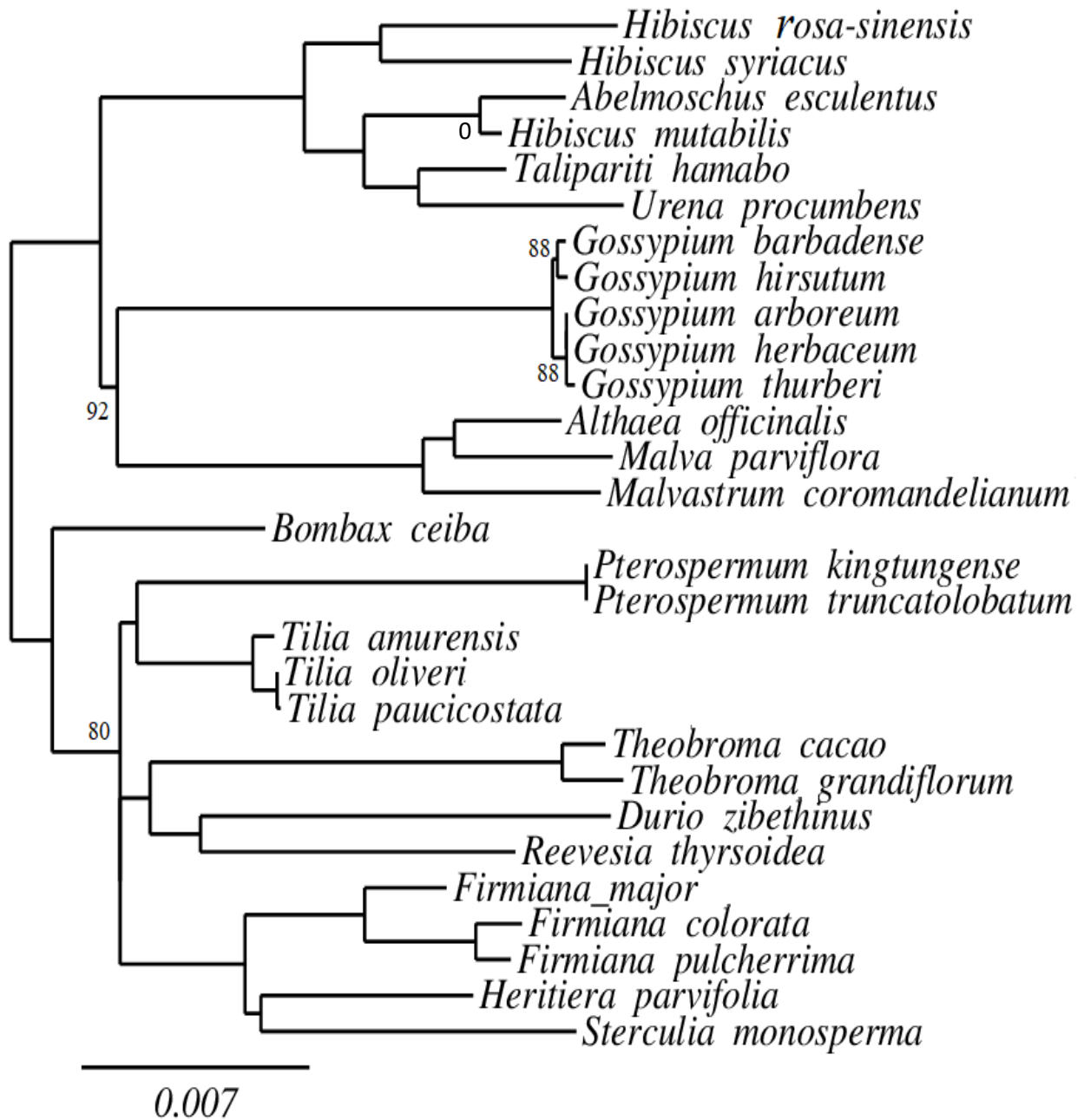


Figure 8.4 Phylogenetic tree based the LSC regions. Bootstrapping value from the node with 100 is omitted.

The phylogenetic tree was also reconstructed with the SSC region based on 16,592 nucleotide sites. Among these sites, 1898 sites were parsimony sites and 1182 sites showed distinct pattern, while 80.46% sites were found constant in the alignment. SSC region could not perform well in inferring of phylogeny and provided false result about the phylogenetic relationship. The result of this region deviates to larger extent from the results of coding

sequences and complete chloroplast genome. Furthermore, this tree shows *Durio zibethinus* as outgroup instead of *Theobroma* genus (Figure 8.5).

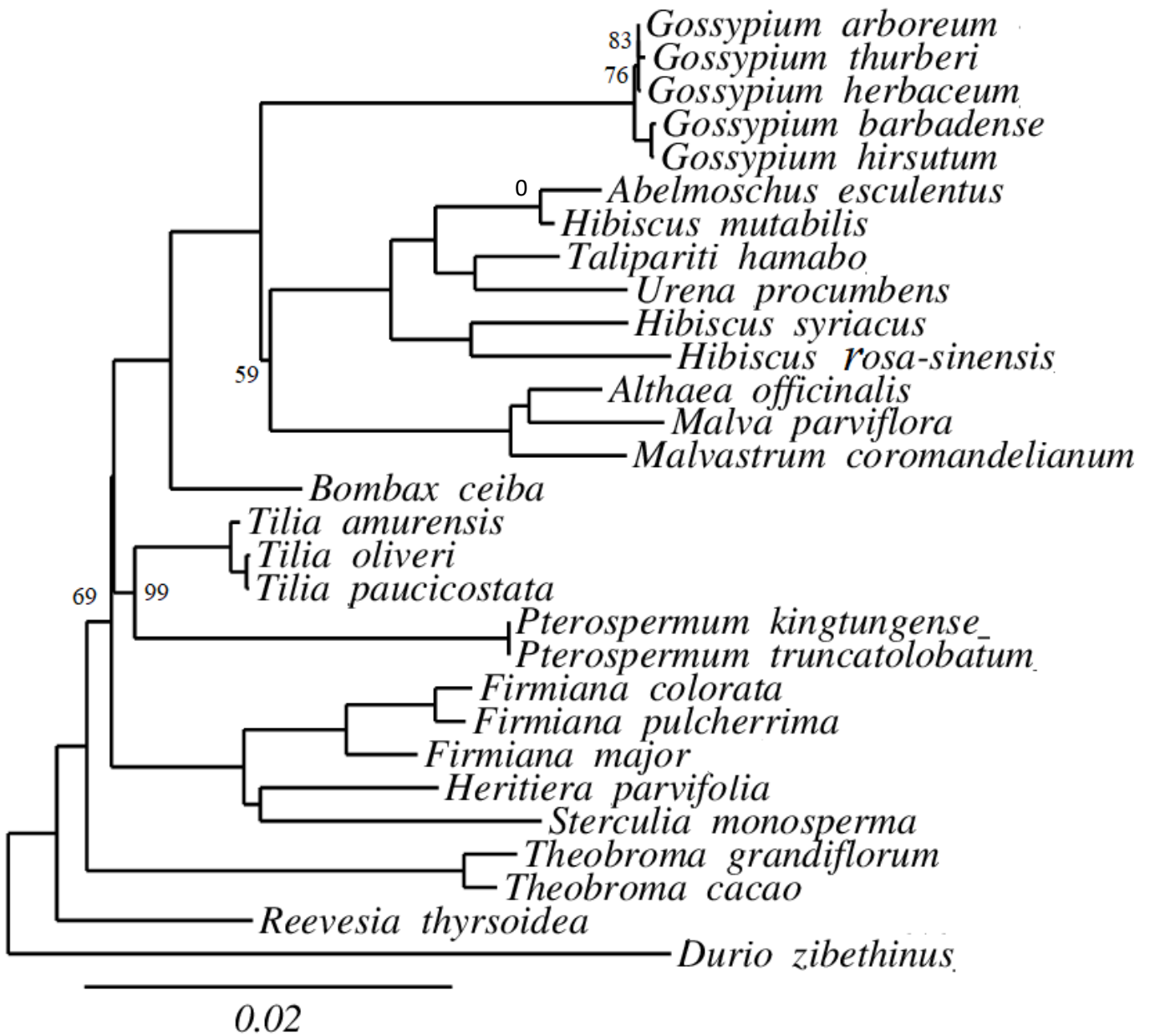


Figure 8.5 Phylogenetic tree based on SSC regions. The bootstrapping value of 100 was omitted from the nodes.

The phylogenetic tree was reconstructed based on the IR region of chloroplast genome containing 21,142 nucleotide sites in which 20,118 (95.16%) sites were constant in all species. The nucleotide sites which were parsimony informative were 269 whereas 262 sites were found with distinct pattern. The phylogenetic tree was reconstructed with best fit model TVM+F+I+G4 but could not well resolve the species due to low polymorphism of the IR regions and the bootstrapping value of the node also decreased significantly. The phylogenetic relationship based on IR region is shown in Figure 8.6.

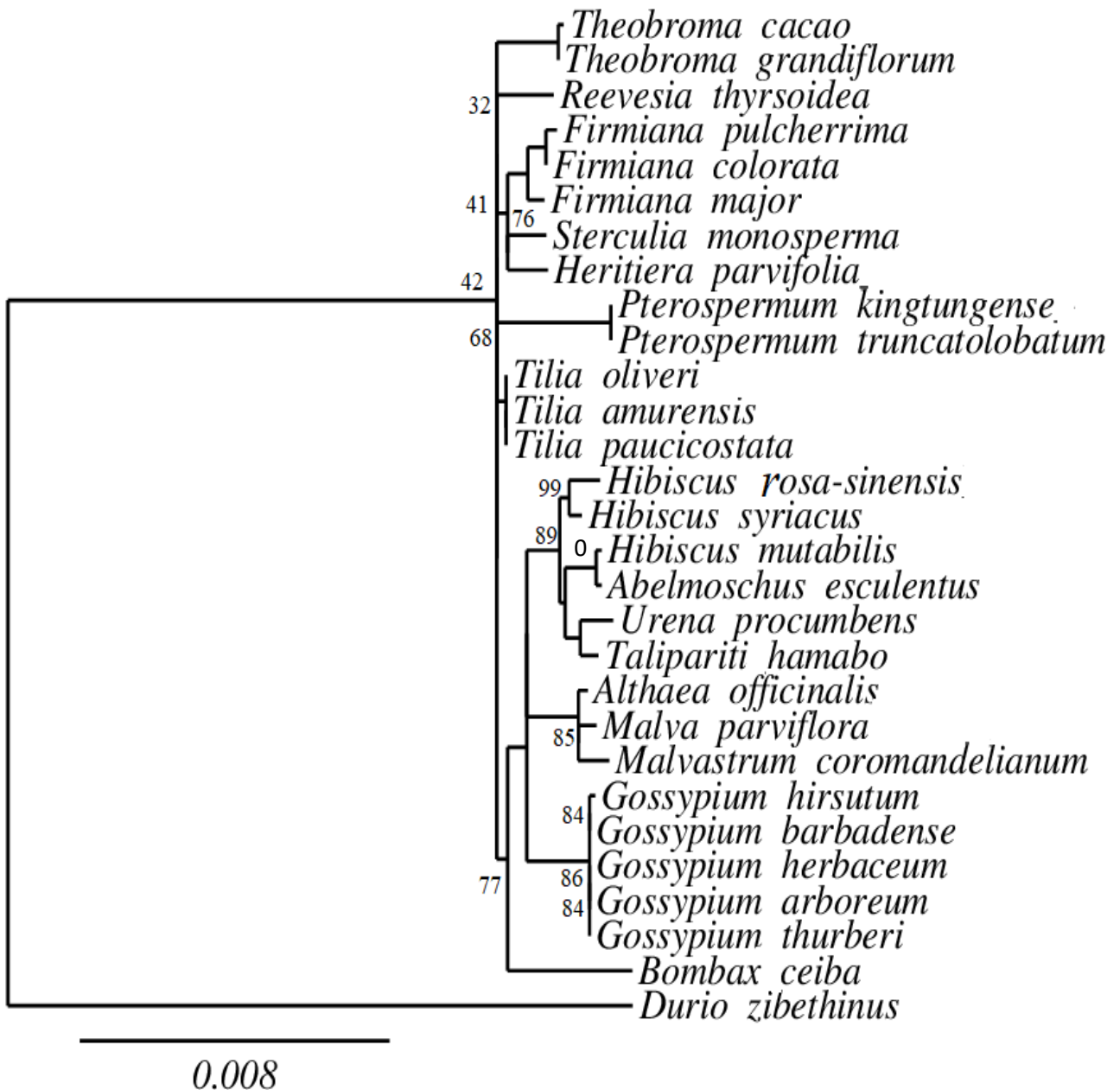


Figure 8.6 Phylogenetic tree based on IR region. All the nodes were shown with the bootstrapping supports.

8.4 Conclusion

Our study supports the classification of family Malvaceae into nine subfamilies. The IRs contraction and expansion can not be used for the determination of relationships between species. The coding sequences and complete chloroplast genome sequences showed high efficacy in inferring of phylogeny as compared to LSC, SSC and IR regions.

Chapter 9

Discussion

9.1 Discussion

Chloroplast genome in angiosperms is conserved regarding gene content and gene order (Amiryousefi *et al.*, 2018; Daniell *et al.*, 2016; Jiang *et al.*, 2018; Li *et al.*, 2018; Menezes *et al.*, 2018). In the chloroplast genome, some tRNA and protein coding genes contain introns that also exhibit conserved nature (Ahmed *et al.*, 2012; Daniell *et al.*, 2016; Menezes *et al.*, 2018). However, loss of some genes or introns from genes are also reported in some species (Daniell *et al.*, 2016; Jansen *et al.*, 2007; Menezes *et al.*, 2018). The chloroplast genome is mostly a quadripartite structure, in which one large single copy (LSC) region and one small single copy (SSC) region are separated by two inverted repeat regions (IRa and IRb), (Feng *et al.*, 2019; Jiang *et al.*, 2018; Sherman-Broyles *et al.*, 2014). In some species, loss of one copy of the inverted repeat regions made the entire genome a single copy (Wu *et al.*, 2011). Furthermore, in some species linear chloroplast genome is also reported (Daniell *et al.*, 2016; Oldenburg and Bendich, 2016). Many mutational events take place in chloroplast genome including substitutions, InDels, structural rearrangements, translocations, inversions, and copy number variations (CNVs) (Ahmed *et al.*, 2012; Cho *et al.*, 2015; Xu *et al.*, 2015). The polymorphic sequences of chloroplast genome have been used for phylogenetic inference to get insights into taxonomic affiliation ranging from population genetics (Ahmed, 2014; Li *et al.*, 2013; Yamane *et al.*, 2003) to deep divergences (Bayer *et al.*, 1999; Feng *et al.*, 2019; Henriquez *et al.*, 2014; Pfeil *et al.*, 2002; Zhai *et al.*, 2019) and barcoding of species (Li *et al.*, 2014). The variations in chloroplast genome has also been used for the studies of evolutionary dynamics of the different lineages of plant (Li and Zheng, 2018; Turmel *et al.*, 2015; Xu *et al.*, 2015).

The evolutionary dynamics of family Malvaceae is not well elucidated. The inter-genus comparative analyses of two genera was previously performed including *Tilia* and *Gossypium* based on complete chloroplast genome (Cai *et al.*, 2015; Z. Chen *et al.*, 2016; Wu *et al.*, 2018; Xu *et al.*, 2012). Despite the lack of data on evolutionary dynamics of family Malvaceae, certain taxonomic discrepancies also exist at family and genera levels. The chloroplast genomes of few species are available. To broaden the knowledge about the evolutionary dynamics of the family Malvaceae, the *de novo* assembly of chloroplast genomes of other species were required.

In current study, we *de novo* assembled chloroplast genomes of eight Malvaceae species to widen the genomic resources of the family Malvaceae. The comparative analyses of chloroplast genomes structure among wide Malvaceae species were performed to get insight into the structure of chloroplast genomes of Malvaceae. We inferred phylogeny of 30 species that

belongs to seven subfamilies of family Malvaceae based on chloroplast genomes to evaluate the previous classifications of family Malvaceae. The strong correlations among mutational events including substitutions, InDels and oligonucleotide repeats were determined which negative correlated with the GTR model for phylogeny inferring. The suitable polymorphic loci were determined for inferring the phylogeny of complex genera. Overall, our study will be helpful for understanding the evolutionary dynamics of family Malvaceae and will provide basis for resolution of taxonomic discrepancies of certain genera of family Malvaceae.

9.2 *De novo* assembly of chloroplast genomes of eight Malvaceae species

Conventionally, the sequencing of chloroplast genome is performed by isolation or enrichment of chloroplast genome by amplification with long range PCR followed by DNA extraction (Amiryousefi *et al.*, 2018; Cai *et al.*, 2015; Kwon *et al.*, 2016). The advancements NGS technologies and availability of massively parallel sequencing data made it now possible to extract and assemble complete chloroplast genome from total genomic DNA. This is because of hundred times higher copy number of chloroplast in leaves than nuclear genome that provide high coverage depth to chloroplast genome (Nock *et al.*, 2011). Recently, several other studies also followed the suit and assembled chloroplast genome from whole genomic DNA shotgun (Menezes *et al.*, 2018; Nguyen *et al.*, 2017; Saina *et al.*, 2018b; Wambugu *et al.*, 2015). We also extracted whole genomic DNA from the fresh leaves that did not show any apparent disease sign to avoid contamination and the whole genomic DNA was used for the sequencing. The short reads were *de novo* assembled by using Velvet 1.2.10 following Ahmed *et al.* (2012). The coverage depth analyses showed high coverage depth for the *de novo* assembled chloroplast genome. Hence, our study agrees with the previous studies in which the authors reported *de novo* assembled complete chloroplast genomes sequences from the whole genome shotgun. The sequencing and/or *de novo* assembly of chloroplast genomes of eight species widen the genomic resources for the understanding of evolutionary dynamics of family Malvaceae. Here, chloroplast genome of *Malva parviflora*, *Malvastrum coromandelianum*, *Urena procumbens*, *Sterculia monosperma*, *Pterospermum truncatolobatum* serve as the first representative of their genus. Therefore, this data provided insight into the chloroplast genome structure of many genera for the first time. The species of *Hibiscus rosa-sinensis*, *Hibiscus mutabilis*, and *Firmiana colorata* was sequenced and/or *de novo* assembled to widen the genomic resources for inter-genus comparison. Two species of genus *Firmiana* were available including *Firmiana major* and *Firmiana pulcherrima* (Wang *et al.*, 2017; Ya *et al.*, 2017). These two species of *Firmiana* were closely related. Therefore, we also assembled chloroplast genome of *Firmiana colorata* and used in the comparison for the identification of suitable

mutational hotspots for the development of authentic and robust markers. Only one species for genus *Hibiscus* was available, which was *Hibiscus syriacus*. Here, we assembled chloroplast genome of *Hibiscus rosa-sinensis* and *Hibiscus mutabilis* from sections Euhibiscus and Trionum, respectively (Pfeil and Crisp, 2005). Therefore, we broaden the genomic resources for genus *Hibiscus* to understand the evolutionary dynamics of this taxonomical complex genus that contains certain taxonomic discrepancies at inter-genus and intra-genus levels (Hinsley, 2009; Koopman and Baum, 2008; Pfeil *et al.*, 2002; Pfeil and Crisp, 2005).

9.3 Comparative analyses of chloroplast genomes among Malvaceae species

We compared chloroplast genome structure of 20 species from 17 genera for genomic features. All the species analysed in the current study possessed almost similar gene content, gene organisation and GC content, except *Abelmoschus esculentus* and *Durio zibethinus* that showed increase or decrease in gene content due to IRs contraction and expansion. The tRNA and protein coding genes that contained introns also showed similarities in reference to the existence of introns. The conserved structure of chloroplast genome along with the similar gene content and organisation has been reported in other lineage of angiosperms (Amiryousefi *et al.*, 2018; Bi *et al.*, 2018; Daniell *et al.*, 2016; Li *et al.*, 2018; Menezes *et al.*, 2018; Raman *et al.*, 2017). However, loss of introns was also observed in many species of angiosperms (monocot and eudicot) and gymnosperms (Downie *et al.*, 1991; Jansen *et al.*, 2007). Recent studies reported intron loss in *Astragalus membranaceus* (Lei *et al.*, 2016), *Lagerstroemia fauriei* (Gu *et al.*, 2016) and *Lonicera japonica* (He *et al.*, 2017). The protein-coding genes in which loss of intron is reported include an RNA polymerase (*rpoC2*), ATP synthase (*atpF*), ribosomal proteins (*rpl2*, *rps12*, and *rps16*), and a clp protease (*clpP*) (Downie *et al.*, 1991; Gu *et al.*, 2016; He *et al.*, 2017; Jansen *et al.*, 2007; Lei *et al.*, 2016). However, our analyses of species of the family Malvaceae revealed conserved genome regarding gene content and gene organisation as well as showed similar pattern regarding the existence of intron/introns in the genes. In the species of Malvaceae, we found the *infA* gene either functional or non-functional or completely missing from the chloroplast genome. This gene initiates translation along two nuclear encoded initiation factors to mediate interactions between mRNA, ribosomes and initiator tRNA-Met (Millen *et al.*, 2001). The loss of *infA* gene has been reported from chloroplast genome of many others angiosperms species during the course of evolution (Daniell *et al.*, 2016; Millen *et al.*, 2001). Furthermore, in many angiosperms' species the *infA* gene has been encoded by nuclear genome such as soybean, tomato etc. (Millen *et al.*, 2001). Here, the absence of *infA* gene or presence of its pseudogene in many Malvaceae species might indicate the transfer of this gene to nuclear genome or might be another functional copy exist

within the nuclear genome that performs important step of translation initiations in these species.

9.4 IRs contraction and expansion and its role in reduction and duplication of genes

The inverted repeats contraction and expansion is common phenomenon in the chloroplast genome (Lin *et al.*, 2012; Rabah *et al.*, 2017; Turmel *et al.*, 2015; Zhang *et al.*, 2016; Zhu *et al.*, 2016), and is considered important for understanding of evolutionary pattern (Menezes *et al.*, 2018). The locations of the LSC/IR and SSC/IR junctions are sometimes regarded as an index of chloroplast genome evolution (Zhang *et al.*, 2013). The contraction of IR regions leads to reduction of a copy of certain genes, whereas expansion of IR regions leads to either origination of pseudogenes or complete duplication of functional genes (Amiryousefi *et al.*, 2018; Li *et al.*, 2018; Menezes *et al.*, 2018; Nazareno *et al.*, 2015).

We also noted variations in the direction of SSC, duplication of few genes in IR (*rps19* and *rps3*, *rpl22* in *Abelmoschus esculentus*) or existence of a single copy of genes (*rpl2* and *rpl23* in *Durio zibethinus*) due to presence in the LSC region instead of IRs. Hence, the IR contraction and expansion lead to either increase or decrease in number of total genes. Therefore, in *Abelmoschus esculentus*, 88 protein coding genes were present whereas in the *Durio zibethinus* 82 genes were present. This data revealed that this common phenomenon of angiosperm chloroplast genome also exists in the family Malvaceae. The origination of pseudogenes at junctions of LSC/IRs and SSC/IRs were also noted in most of the species of Malvaceae. Moreover, some species of other families of order Malvales also showed increase or decrease in gene number due to IRs contraction and expansion. For instance, the species of family Thymelaeaceae showed increase in number of genes due to IRs expansion (Lee *et al.*, 2018) whereas the family Dipterocarpaceae showed decrease in number of genes due to IRs contraction (Heckenhauer *et al.*, 2019). Hence, the variations in the numbers of genes are possible in the chloroplast genome due to the IRs contraction and expansion.

9.5 Rate of synonymous and non-synonymous substitutions and adaptive evolution

The non-synonymous (K_a) and synonymous (K_s) pattern of nucleotide substitutions is important marker in evolution (Kimura, 1979). The K_a/K_s ratio indicates the selection pressure on protein coding genes. The K_a/K_s less than 1 indicates purifying selection, 1 shows neutral evolution, and more than 1 indicates positive selection pressure on those protein coding genes (Lawrie *et al.*, 2013). In current study, we observed higher synonymous substitutions (average = 0.053673) than non-synonymous substitutions (Average = 0.009346). The analysed species of Malvaceae exhibited slow evolutionary rate as expected for the chloroplast genome in

general. The lower Ka/Ks ratio (< 0.5) in most of the genes indicates purifying selection working on these genes. The genes which showed higher Ka/Ks value in some species might be under low purifying selection pressure and greater positive selection pressure due to certain environmental conditions such as *clpP* in *Abelmoschus esculentus* (1.16), *Durio zibethinus* (1.2), *Firmiana major* (7.47), *Firmiana colorata* (3.94), *Heritiera parvifolia* (5.52), *Reevesia thyrsoidea* (1.2), *Pterospermum truncatolobatum* (1.24), and *Sterculia monosperma* (3.42). This might be attributed to varying degrees of biotic or abiotic stresses faced by different species in their ecological niches. We found slow evolutionary rate of photosynthetic genes, which is common in plant cells and this is due to high purifying selection pressure on the genes (Choi *et al.*, 2018; Menezes *et al.*, 2018; Saina *et al.*, 2018a).

The synonymous and non-synonymous substitutions rate was also compared with the rate of transition and transversion substitutions in selected genes of chloroplast genome. We included six genes (*psbA*, *atpA*, *rps2*, *petD*, *rpoA*, and *rpoC1*) with purifying selection pressure, five genes (*ndhH*, *rbcL*, *clpP*, *rpl20*, *ycf1*) with neutral selection pressure and three genes (*accD*, *rpl20*, and *rps4*) with positive selection pressure of genus *Firmiana*. Our analyses revealed that most of the non-synonymous substitutions were produced by transversion substitutions in genes showing neutral selection pressure or positive selection pressure. Interestingly, genes with purifying selection pressure showed most of the non-synonymous substitutions (55.6%) linked to transition substitutions as compared to transversion substitutions. This observation shows that genes exhibiting purifying selection try to avoid transversion substitutions and allow transition substitutions. A previous study also reported non-synonymous substitutions as the cause of transversion substitutions in chloroplast genome of cereal crops (Matsuoka *et al.*, 2002).

9.6 Mutational dynamics in family Malvaceae and correlations among substitutions, InDels and repeats

Previous studies reported strong associations among mutational events including substitutions, InDels and oligonucleotide repeats in prokaryotes and eukaryote chloroplast genomes (McDonald *et al.*, 2011; Tian *et al.*, 2008; Zhu *et al.*, 2009). Ahmed *et al.* (2012) reported strong correlations among these mutational events in the chloroplast genomes of limited number of species of family Araceae (monocot, angiosperm) and Yi *et al.* (2013) in family Cephalotaxaceae (gymnosperm) by comparing to species of genus *Cephalotaxus*. In the current study, we also evaluated the correlations among substitutions, InDels, and repeats in diverse species of family Malvaceae (eudicot, angiosperms) and found highly significant correlations

among these mutational events which negatively correlated with the GTR model and suggest the interconnection or dependence among mutational events.

To calculate correct number of substitutions and InDels in pairwise alignments, we removed/corrected all the inversions from alignments. Inversions are frequently present in chloroplast genomes (Xu *et al.*, 2015) and cause noise in the alignment that may lead to false positive results in comparative analyses and in inferring of phylogeny (Menezes *et al.*, 2018). The genes that often shift between single copies and inverted repeat regions due to IRs contraction and expansion in chloroplast genomes lead to difference in rate of mutations (Zhu *et al.*, 2016), a phenomenon known as rate heterotachy (Lockhart *et al.*, 2006). Genes in IRs regions have slower rates of mutations compared to those in LSC and SSC (Ahmed *et al.*, 2012). Hence, we also removed all those genes from the alignment that were located at junction of LSC/IR and SSC/IR to avoid the effects of evolutionary rate heterotachy (Wu *et al.*, 2011) in different lineages. Following genes and regions were excluded from the comparisons: *rpl16* to second exon of *rpl2* at junction of LSC/IRb, up to 1,000 bp from each sites of IRb/SSC, also partially removing *ndhF* gene, and complete removal of *ycf1* gene from the junction of SSC/IRa.

Ahmed *et al.* (2012) displayed correlations among substitutions, InDels, and oligonucleotide repeats in complete chloroplast genomes of family Araceae (monocots, Angiosperms). Following this study, similar correlations were later reported in the complete chloroplast genome of genus *Cephalotaxus*, family Cephalotaxaceae of gymnosperm (Yi *et al.*, 2013). Here, we evaluated and confirmed the existence of such correlations in family Malvaceae (eudicots, angiosperms) by including species from basal lineages to crown groups. In current study, we also analysed regression of SNPs on InDels, SNPs on repeats, and InDels on repeats. The regression analyses also confirmed the role of InDels in generation of SNPs in genome whereas repeats were also determined as the cause of generation of SNPs and InDels in genomes. The first related study of Ahmed *et al.* (2012) was limited to comparative analyses of the chloroplast genomes of two morphotypes of one species, *Colocasia esculenta*, and four other closely related species of clade *Lemnoideae* including *Wolffiella lingulata*, *Lemna minor*, *Wolffia australiana* and *Spirodela polyrhiza*. *Colocasia esculenta* is among the crown groups and Lemnoideae clade is basal to other aroids in family Araceae (Henriquez *et al.*, 2014). In the successive study, Yi *et al.* (2013) used only two species from genus *Cephalotaxus* of family Cephalotaxaceae (Gymnosperms) to report the correlations. In the current study, the observed weak to strong correlations of all three types of mutational events in evolutionarily close comparisons (within a genus) as well as far comparisons (within family) is in agreement with

the previous reports (Ahmed *et al.*, 2012; Yi *et al.*, 2013) and supports the hypothesis that the distribution of oligonucleotide repeats can be used as a proxy for mutational hotspots (Ahmed *et al.*, 2012). Similar observations have also been reported in other prokaryotic and eukaryotic genomes (Tian *et al.* 2008, Zhu *et al.* 2009, McDONald *et al.* 2011), which suggest that such correlations might be a universal property in the genomes of all living organisms. However, there is further needed to empirically evaluate such genome-wide correlations in other plant families by including closely as well as distantly related species. If such observations gain further support in independent studies, this will suggest a need for correction/revision in the most advance model, i.e. the generalized time reversible (GTR) model of molecular evolution, which assumes that the mutations arise independent of mutations on other sites (Drouin *et al.*, 2008).

9.7 Role of the contraction and expansion of inverted repeats in inference of phylogeny

The contraction and expansion of inverted repeat regions is common phenomena in the chloroplast genome of angiosperms (Lin *et al.*, 2012; Rabah *et al.*, 2017; Turmel *et al.*, 2015; Zhang *et al.*, 2016; Zhu *et al.*, 2016), and is considered important for understanding of evolutionary pattern (Menezes *et al.*, 2018). The locations of the LSC/IRs and SSC/IRs junctions are sometimes regarded as an index of chloroplast genome evolution (Zhang *et al.*, 2013). Recently, role of the IRs contraction and expansion and position of genes at junctions of LSC, SSC and IR regions of chloroplast were also suggested in the determination of phylogenetic relationships based on comparative analyses of few closely related species (Liu *et al.*, 2018). Therefore, authors suggested its validation in the far diverse species.

In the current study, we evaluated the link between the IRs contraction and expansion to the phylogenetics relationship among diverse species in seven subfamilies of family Malvaceae. Our analyses revealed that phenomena of IR contraction and expansion can not be linked to the phylogenetic of these species. The closely related species in phylogeny showed more divergence when their resemblance was observed on the basis of IRs contraction and expansion, whereas in contrast to these results, the far related species in phylogenetic tree were found closely related on the basis of IRs contraction and expansion. For instance, *Abelmoschus esculentus* showed uneven IRs expansion as compared to other species of the subfamily Malvoideae. Similarly, we noted *Durio zibethinus* (Helicteroideae) showed uneven IRs contraction in comparison to *Reevesia thyrsoides*, another species of subfamily Helicteroideae. Therefore, the IRs contraction and expansion provide false information about the resemblance of the species.

The comparison of the distance of the genes that exists at the junctions also showed uneven distance from junctions of LSC, SSC, and IR and lack link to phylogeny of family. Species of different subfamilies showed closed relationship in comparison to the species of same subfamilies based on position of the genes such as *rps19* and *rpl2* at the junctions of LSC/IRb. For instance, many species of the subfamily Malvoideae could be found closely related to the species of other subfamilies as compared to the species of the same subfamily. The similar observation was renowned based on the positions of genes at junction of IRb/SSC and SSC/IRa. Moreover, it is also difficult to draw a line about the position of the genes and their link to phylogeny of the species. Therefore, based on our analyses, we suggest the IRs contraction and expansion and the position of the genes at the junctions may not have any link to the phylogeny. These regions might lack link to phylogeny due to fast evolution at the junction regions or these regions might be similar in evolution pattern to simple sequence repeats, in which the contraction and expansion take place due to insertions and deletions of the repeat sequences. If this is the case, then IR contraction and expansion might be occurring due to a single mutational event and can not be employed for inferring of phylogeny. Nevertheless, we recommend the broad studies in diverse plant lineages for further elucidation.

9.8 Phylogenetic relationship of family Malvaceae

Previously, phylogeny of the family Malvaceae was inferred based on limited number of chloroplast genes. Bayer *et al.* (1999) used chloroplast genes *atpB* and *rbcL*, whereas Alverson *et al.* (1999) used chloroplast gene *ndhF* for inferring of phylogeny in family Malvaceae. Based on these studies, four families Bombaceae, Malvaceae s.s, Sterculiaceae, and Tiliaceae (core Malvales) were classified as family Malvaceae s.l. This extended family Malvaceae s.l was further classified based on these studies into nine subfamilies comprising Helicteroideae, Malvoideae, Brownlowioideae, Bombacoideae, Dombeyoideae, Grewioideae, Byttnerioideae, Tilioideae and Sterculioideae. This classification got acceptance in research community and has been adopted by various researchers (Ate *et al.*, 2005; Baum *et al.*, 2004; Bayer and Kubitzki, 2003; Carvalho-Sobrinho *et al.*, 2016; Duarte *et al.*, 2011; Perveen *et al.*, 2004; Tate *et al.*, 2005). A recent study also supports the nine subfamilies classification (Richardson *et al.*, 2015). Although this classification resolved the taxonomic discrepancies to some extent, but the inter-subfamilies level relationships are still inconclusive. In contrast, some researchers also suggested different classification for family Malvaceae. Thorne (2000) also suggested three families classification and considered Bombacaceae and Sterculiaceae under Malvaceae s.l and Byttneriaceae and Tiliaceae (with some restriction) as separate families. Hinsley (2006) suggested four other alternate classifications of Malvaceae s.l (core Malvales) that included: 1)

core Malvales can be considered single family, 2) nine monophyletic clades determined by Bayer *et al.* (1999) can be considered separate families instead of subfamilies, 3) except two subfamilies Brownlowioideae and Grewioideae, other subfamilies can be considered as family Malvaceae, 4) five family's classification similar to traditional classification by accepting paraphyletic grouping such as Sterculiaceae, Bombacaceae, Malvaceae s.s and Grewiaceae. The fifth family was considered based on merging of Byttneriaceae with Tiliaceae.

These previous classifications were suggested based on either morphological characters or molecular analyses based on few markers of chloroplast genome. Recently, discrepancies in phylogeny were resolved based on the analyses of complete chloroplast genome such as in order Bryopsidales and family Ranunculaceae (Cremen *et al.*, 2019; Zhai *et al.*, 2019). Here, we reconstructed phylogenetic tree based on coding sequences including 10 species from three families Bixaceae, Dipterocarpaceae, and Thymelaeaceae of order Malvales and 30 species from 7 subfamilies of Malvaceae. We reconstructed two separate phylogenetic trees. In one tree, we included 40 species of four families of order Malvales, whereas another tree was reconstructed separately based on protein coding genes for the species of family Malvaceae. The species of order Malvales were included to analyse the phylogenetic differences among the four subfamilies of order Malvales and the differences existed among the seven subfamilies of family Malvaceae, following Cremen *et al.* (2019).

Our results showed that species of each family of order Malvales are well resolved and showed monophyletic position. All the species of family Malvaceae form monophyletic clade and well resolved from the other families of Malvales with high branch length. All the seven subfamilies of Malvaceae were also well resolved and were monophyletic. Despite the use of limited number of species of family Malvaceae, the branch length difference was found less among the seven subfamilies of family Malvaceae as compared to the differences that existed within the family of order Malvales. Therefore, our results support classification of Malvaceae into nine subfamilies (Bayer *et al.*, 1999), which is also adapted by various researchers (Baum *et al.*, 2004; Carvalho-Sobrinho *et al.*, 2016; Wilkie *et al.*, 2006) instead of considering each subfamily as separate family (Hinsley, 2006). *H. rosa-sinensis* and *H. syriacus* have a common node whereas *Hibiscus mutabilis* share node with the *Abelmoschus esculentus* and showed polyphyletic origin with 100 bootstrapping. The later species belong to genera that were segregated from genus *Hibiscus* (Pfeil and Crisp, 2005). This testifies a close relationship among these genera. Moreover, the genus *Hibiscus* might be paraphyletic instead of monophyletic as suggested previously, or all the species from other genera that were embedded in the phylogeny of genus *Hibiscus* could be included in genus *Hibiscus* by considering the

Hibiscus as monophyletic genus (Hinsley, 2009; Koopman and Baum, 2008; Pfeil *et al.*, 2002; Pfeil and Crisp, 2005). Moreover, the segregation of the *Abelmoschus esculentus* from genus *Hibiscus* is not accurate and it might be that *Hibiscus* genus accommodates all the species from other genera that were found embedded in its phylogenetic tree (Pfeil *et al.*, 2002), and lead to expansion of genus *Hibiscus* as suggested previously (Hinsley, 2009; Pfeil and Crisp, 2005). In the current study, 30 species were included from 17 genera of 7 subfamilies of Malvaceae. Consideration of extensive sampling can further elaborate the taxonomic position of Malvaceae and the inter-subfamilies level classification. In the reconstruction of phylogenetic tree based on protein coding genes, we ignored gene *ycf1* due to inversion which lead to false result in the analyses of phylogenetic relationship (Menezes *et al.* 2018).

Some studies also evaluated the phylogeny inferring ability of complete chloroplast genome, LSC, SSC and IR regions (Amiryousefi *et al.*, 2018; Li *et al.*, 2017; Xue *et al.*, 2019). Here, we also reconstructed phylogenetic tree of family Malvaceae based on complete chloroplast genome and based on the chloroplast genome regions including LSC, SSC and IR region separately to access the efficacy of these regions in phylogenetic inference. We also removed *ycf1* from the complete chloroplast genome to avoid false results due to inversions and rate heterotachy (Lockhart *et al.*, 2006). We found that phylogenetic tree based on complete chloroplast genome well resolves the phylogeny of Malvaceae and the results were similar to the tree constructed based on coding sequences whereas the LSC, SSC and IR regions gave compromised result in phylogeny. Therefore, these regions could not be recommended to infer phylogeny of family Malvaceae. The compromised result of the SSC region might be due to the presence of *ycf1* genes which shows inversions and comprised about 1/3 of the SSC region. The compromised result of the IR region might be due to the presence of low nucleotide diversity in this regions as revealed from the comparison of the three genera of Malvaceae in current study and previously reported in the genus *Tilia* (Cai *et al.*, 2015).

9.9 Comparison of species of three genera of Malvaceae

We obtained insight into the molecular evolution of three genera of family Malvaceae including *Theobroma*, *Firmiana* and *Hibiscus*. The genus *Theobroma* belongs to subfamily Byttnerioideae which lies basal to family Malvaceae (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Cornejo *et al.*, 2017; Richardson *et al.*, 2015). The two species of genus *Theobroma* including *Theobroma cacao* and *Theobroma grandiflorum* were compared to get insight into the chloroplast genomes of the Malvaceae species as these species lie basal to the family Malvaceae, and to identify unique loci for the phylogenetic inference of the genus *Theobroma*. Genus *Firmiana* belongs to subfamily Sterculioideae. The subfamily Sterculioideae shows

taxonomic discrepancies and in previous phylogeny analyses some tribes and genera of this subfamily could not be well resolved (Wilkie *et al.*, 2006). Here, we identified mutational hotspots for development of suitable molecular markers that might be important for inferring of phylogeny of the subfamily Sterculioideae with high efficacy and authenticity. Furthermore, this subfamily also lies on the basal side of the family Malvaceae, therefore, the comparison of its species provided further insight about the molecular evolution in the chloroplast genomes of the species that lies on the basal. The genus *Hibiscus* belongs to subfamily Malvoideae and due to its complex phylogeny shows certain taxonomic discrepancies (Bayer *et al.*, 1999; Pfeil *et al.*, 2002; Pfeil and Crisp, 2005). The deep comparison of *Hibiscus rosa-sinensis* and *Hibiscus syriacus* was not only helpful for the study of molecular evolution of these two species, but also provided insight into the molecular evolution of the subfamily Malvoideae and family Malvaceae. The comparative analyses of *Hibiscus rosa-sinensis* and *Hibiscus syriacus* revealed mutational hotspots, which may be helpful to develop molecular marker for resolving of taxonomic issues of genus *Hibiscus*. Moreover, these markers might be helpful for barcoding of *Hibiscus* species to avoid substitutions and adulterations issues in their medicinal use.

9.10 Codon usage analyses and its link to evolution of species

The codon usage is vital to understand the evolutionary process, selection pressure on the genes, and genome structure (Yang *et al.*, 2014). We analysed codon usage within three genera and compared at intra-genus level. We analysed the codon usage in terms of relative synonymous codon usage (RSCU) to get insight into preferred and non-preferred synonymous codons that code for a specific amino acid. This approach is commonly used in chloroplast genome studies to find out preferred synonymous codons (Amiryousefi *et al.*, 2018; Redwan *et al.*, 2015; Saina *et al.*, 2018b). The high level of similarities in codon usage between the species at inter-genus level and intra-genus level revealed that these species are closely related and passed through the similar types of environmental stresses and conditions during the process of evolution. The extent of similarities was higher in intra-genus comparison than inter-genus comparison. These results agreed with previous report which linked codon usage to environmental conditions and evolutionary history of the species as species of the same genus are evolutionary closely linked and face similar type of environmental stresses during course of evolution (Yang *et al.*, 2014). In the current study, codon usage analyses in terms of RSCU value illustrates codons ending with A/T nucleotide had RSCU value greater than 1 and encode most of the amino acids as compared to codons ending with C/G nucleotide and had RSCU value less than 1. The RSCU value of 1 or greater than 1 indicates the high preference of that

codon in encoding a specific amino acid (Saina *et al.*, 2018b). This trend might be due to A/T rich chloroplast genome, and has also been observed in chloroplast genomes of other species (Amiryousefi *et al.*, 2018; He *et al.*, 2017; Menezes *et al.*, 2018; Qian *et al.*, 2013; Saina *et al.*, 2018b).

9.11 RNA editing sites

RNA editing is a post-transcriptional process for converting cytidine (C) to uridine (U), or U to C, at specific sites within RNA molecules to alter the identity of nucleotides between RNA and genomic DNA, serving as a mechanism to correct missense mutations of genes at the RNA level, and also to enrich genetic information (Wang *et al.*, 2016, 2015). First editing event was reported in the *rpl2* gene of mRNA transcript of the maize chloroplast genome in 1991 (Hoch *et al.*, 1991). Most of the RNA editing sites took place on the second nucleotide of the codons and lead to conversion of serine to leucine (Saina *et al.*, 2018b). The RNA editing also changes initiation codons at the post transcriptional level (Hoch *et al.*, 1991). RNA editing sites have been also reported in the chloroplast genomes of several species (Huang *et al.*, 2017; Kazakoff *et al.*, 2012; Kugita *et al.*, 2003; Lin *et al.*, 2012; Mower, 2009; Saina *et al.*, 2018b). The PREP-cp is a tool that is developed to determine RNA editing sites in 35 protein coding genes of chloroplast genome (Mower, 2009). Here, we analysed RNA editing sites in the two species of genus *Theobroma*, three species of genus *Firmiana*, and two species of genus *Hibiscus*. We determined RNA editing sites within 23 genes of the species of genus *Theobroma* and genus *Hibiscus*, whereas 25 genes contained RNA editing sites in genus *Firmiana*. The genes that contained RNA editing sites showed similarities at inter-genus and intra-genus comparison. The RNA editing sites were 100% identical within the closely related species of genus *Theobroma*. This data revealed that closely related species also contained similar RNA editing sites. Moreover, high similarities in RNA editing sites at inter-genus level comparison revealed species of same lineage had high similarities in RNA editing sites. The RNA editing sites observed in the current study have also been observed in other plant lineages which also shows the conserved nature of RNA editing sites within the chloroplast genome of plant lineages (Huang *et al.*, 2017; Lin *et al.*, 2012; Saina *et al.*, 2018b).

9.12 Simple sequence repeats

Simple sequence repeats (SSRs) are present in the chloroplast genome. The contraction and expansion of SSRs units are caused by slipped strand mispairing during DNA replication (Levinson and Gutman, 1987). The SSRs of chloroplast genome are used for the population genetics studies and barcoding of species and cultivars (Huang *et al.*, 2018; Joh *et al.*, 2017; Nguyen *et al.*, 2018; Qiu *et al.*, 2013). Recent studies revealed the chloroplast genome showed

abundance of mononucleotide repeats and most of the repeat motifs contained A/T. The number and types of repeats varies in the plant lineages (Dong *et al.*, 2016; He *et al.*, 2016; Liu *et al.*, 2018; Ma, 2018; Menezes *et al.*, 2018; Vieira *et al.*, 2016; Zhang *et al.*, 2016). In the current study, we also analysed SSRs in the in three genera. Our analyses of all the three genera provided almost similar results in inter-genus as well as in intra-genus comparison related to the number of SSRs, types, and their existence in the three large regions of chloroplast genome. Mononucleotide SSRs (A/T motifs) and dinucleotide SSRs (AT/TA motifs) had abundance in chloroplast genomes of all genera. Moreover, LSC contained highest SSRs, followed by SSC whereas IR had lowest SSRs. These finding of our study are in agreement with previous reports (Amiryousefi *et al.*, 2018; Menezes *et al.*, 2018; Poczai and Hyvönen, 2017; Saina *et al.*, 2018b; Xu *et al.*, 2012). SSRs loci reported in current study might be helpful for the barcoding and population genetics studies in these three genera.

9.13 Oligonucleotide repeats

The oligonucleotide repeats usually exist in four forms: forward, reverse, palindromic, and complementary. These repeats are usually evaluated with the minimum repeat size of 20-30 bp in the chloroplast genome (Asaf *et al.*, 2018; Kurtz *et al.*, 2001; Menezes *et al.*, 2018; Yu *et al.*, 2017). The role of moderate repeats with repeat size of 14-48 bp was suggested in generation of inversion (Kim and Lee, 2005; Whitlock *et al.*, 2010) and InDels (Kawata *et al.*, 1997). Furthermore, these repeats can be used as proxy for identification of mutational hotspots (Ahmed *et al.*, 2012; McDonald *et al.*, 2011). Our results are also in agreement with these reports. The oligonucleotide repeats analyses revealed highest similarities in intra-genus comparison, whereas showed differences at inter-genus comparison. In genus *Theobroma*, *T. cacao* had 46 repeats and *T. grandiflorum* had 53 repeats that showed high similarities. In genus *Firmiana*, there were 49 repeats each in *F. colorata* and *F. pulcherrima*, whereas in *F. major* 65 repeats were present. Although, the similarities in repeats types existed, the significant divergent was observed for *F. major* from two other species. Moreover, the repeats of two species of *Firmiana* also showed similarities with species of *Theobroma*. In genus *Hibiscus*, we found 100 repeats in *Hibiscus rosa-sinensis* and 130 in *Hibiscus syriacus*. Here, about 2 times increase was noted in the repeat's numbers as compared to the two mentioned genera (*Theobroma* and *Firmiana*). Genus *Theobroma* and *Firmiana* existed at the basal of the family Malvaceae phylogeny, whereas the genus *Hibiscus* existed at the top of the phylogeny tree as previously reported (Alverson *et al.*, 1999; Bayer *et al.*, 1999; Nyffeler *et al.*, 2005). Our phylogeny inferring also supports these results. This analysis shows species at basal of phylogenetic tree (*Theobroma* and *Firmiana*) exhibit less repeats as compared to the species

that exist at the top of the phylogenetic (*Hibiscus*) tree. Our analyses revealed that most of the repeats arisen in species of genus *Hibiscus* near to the repeats that were also reported in *Firmiana* and *Theobroma*. Hence, the maximum number of repeats observed in genus *Hibiscus* were due to generation of new repeats or due to generation of substitutions in the repeats of the species that lies on the basal due to which the number of repeats decrease in those species.

Most of the oligonucleotide repeats were found in the intergenic spacer regions, followed by the intronic regions, while the lowest repeats were found in the coding sequences. Here, our results agreed with the study done on Bromeliaceae, in which authors reported the same pattern of repeats distribution (Poczai and Hyvönen, 2017). However, some studies reported repeats abundance in coding region also (Menezes *et al.*, 2018).

9.14 Substitutions and InDels in chloroplast genome

Previous studies showed that substitutions and InDels most commonly occur in LSC and SSC regions of chloroplast genome whereas IRs region is the most conserved part of chloroplast genome (Ahmed *et al.*, 2012). Our results also showed that most of substitutions occurred in LSC and SSC regions while IR was the most conserved part. Analysis of substations types revealed that transitions were less common as compared to transversions with $Ts/Tv < 1$ in all three genera including *Theobroma*, *Firmiana* and *Hibiscus*. Transition to transversion ratio $Ts/Tv < 1$ has also been observed in chloroplast genomes of *Tilia*, another genus of Malvaceae (Cai *et al.*, 2015). Our finding also agreed with previous reports that showed $Ts/Tv < 1$ in chloroplast genome of gymnosperms and angiosperms (Dong *et al.*, 2016; Kim and Kim, 2014; Mustafina *et al.*, 2019; Song *et al.*, 2015; Yang *et al.*, 2016). However, in contrast to these studies, Cao *et al.*, (2018) also reported $Ts/Tv = 1.6$ in chloroplast genome of *Dioscorea polystachya*. The lower Ts/Tv ratio might be due to the A/T rich content of chloroplast genome (Menezes *et al.*, 2018) as nuclear genome with high GC content shows high Ts/Tv and reach in some cases up to 2.0 (Alipour *et al.*, 2017).

9.15 Mutational hotspots regions

Chloroplast genome of angiosperms is conserved but some regions are high polymorphic within chloroplast genome in comparison to other regions (Amiryousefi *et al.*, 2018; Choi *et al.*, 2016; Li *et al.*, 2018; Menezes *et al.*, 2018). Therefore, screening of the chloroplast regions is required to identify suitable polymorphic loci. The maternal inheritance and lack of meiotic recombination (Palmaer, 1985) makes it suitable for population genetics studies to phylogenetic level studies (Ahmed, 2014; Henriquez *et al.*, 2014; Yang *et al.*, 2019; Zhai *et al.*, 2019). The regions of chloroplast genome show different level of polymorphisms within

different lineages and different levels of polymorphisms are required for different types of studies (Daniell *et al.*, 2016), such as regions used for population genetic could be different from the regions used for phylogenetic studies at genus or family level (Li *et al.*, 2018; Menezes *et al.*, 2018; Pfeil *et al.*, 2002). The importance of suitable polymorphic loci further increases for the phylogenetic inference of species that reveal complex taxonomy and shows taxonomic discrepancies (Daniell *et al.*, 2016). Recently, several studies identified specific polymorphic loci for the development of authentic and cost-effective markers to resolve phylogeny of closely related and taxonomically complex genera and families (Bi *et al.*, 2018; Choi *et al.*, 2016; Li *et al.*, 2018; Menezes *et al.*, 2018; Yu *et al.*, 2017). Here, we performed comparative analyses of three genera to identify suitable polymorphic loci for the development of authentic markers to resolve phylogenetic relationships. We identified 30 polymorphic loci within three genera including *Theobroma*, *Firmiana* and *Hibiscus*. The inter-genus comparison revealed that the identified regions in each genus showed significant variations in the extend of polymorphism, and we also found that some regions which were selected as high polymorphic regions in one genus were even not selected within the 30 high polymorphic loci. Hence, our study also supports the view that polymorphic regions show variations in different level comparison at genus/family level. Therefore, identification of suitable and specific polymorphic loci is required for the high-resolution phylogenetic inference at genus/family level.

9.15.1 Genus *Theobroma*

Genus *Theobroma* is a small genus and phylogeny of its species is unresolved yet. The screening for polymorphic regions identified most of the polymorphic sequences from the IGS regions whereas the two intronic and one protein-coding gene were also included in the 30 high polymorphic sequences. Our results are similar to Menezes *et al.* (2018), as they also suggested that IGS are more polymorphic than coding and intronic regions and can be used to develop polymorphic markers. Kane *et al.* (2012) suggested *trnH-psbA*, *rbcL*, *matK* and *ccsA* as barcode for *T. cacao* and *T. grandiflorum*. In current study, our identified polymorphic regions revealed high polymorphism than the above-mentioned regions except *trnH-psbA*. Most of our identified polymorphic sequences have suitable length for development of sanger sequencing base genotyping markers that might be useful for phylogenetic inference. These markers might be also helpful in barcoding of *Theobroma* species and identification of suitable relative taxa for the breeding for the enhancement of qualitative and quantitative characters.

9.15.2 Genus *Firmiana*

Firmiana is a small genus of subfamily Sterculioideae (Bayer *et al.*, 1999; Wilkie *et al.*, 2006). Information about the suitable polymorphic loci lacked for development of authentic markers to infer phylogeny of the genus (Fan *et al.*, 2013). Fan *et al.* (2013) developed SSRs markers from the transcriptome of *Firmiana* but the SSRs could not be modelled for the maximum likelihood tree reconstruction of genus *Firmiana*. Moreover, Wilkie *et al.* (2006) could not resolve the phylogenetic relationships of many clades of Sterculioideae. Therefore, they suggested future studies based on new molecular markers in addition to morphological characters for phylogeny inferring of subfamily Sterculioideae. Here, we screened the regions of chloroplast genome by comparison of three *Firmiana* species and identified suitable polymorphic loci for the development of markers. In the current study, we found thirty highly polymorphic regions as compared to commonly used markers *trnH-psbA*, *rbcL* and *matK*. This finding revealed that the low efficacy of these markers might be a cause of low resolution of the Sterculioideae species. Therefore, 30 high polymorphic loci identified in our study can be used for development of suitable markers in inferring of phylogeny and population genetics studies specifically within genus *Firmiana* and subfamily Sterculioideae.

9.15.3 Genus *Hibiscus*

Not all genes are phylogenetically useful in resolving taxonomic discrepancies. Pfeil *et al.* (2002) used chloroplast genome sequences of *ndhF* and *rpl16* intron to describe the phylogeny of tribe Hibisceae and the genus *Hibiscus* with the aim to achieve monophyletic position of genus *Hibiscus*. However, expected results were not achieved and many segregated genera from *Hibiscus* were embedded within its phylogenetic tree i.e. *Fioria* and *Abelmoschus*. Moreover, some members of tribes Decaschistieae and Malvavisceae were also embedded within *Hibiscus*. (Small. (2004) used nuclear ribosomal ITS (internal transcribed sequences), non-coding part of chloroplast DNA (*rpl16* intron), and a nuclear coding gene granule-bound starch synthase (*GBSSI*) but insufficiently resolved the phylogeny of *Hibiscus*. Therefore, the author suggested additional data source beyond the commonly used markers. Tate *et al.* (2005) employed ITS markers on another tribe Malveae of family Malvaceae with the aim to elucidate the phylogenetics, but their results were non-significant. Therefore, they also suggested the use of other nuclear and chloroplast genome-based markers to elucidate the phylogeny of Malvaceae. A recent study of Poovitha *et al.* (2016) used *rbcL*, *matK*, ITS, and *trnH-psbA* to find best suitable loci for barcoding. They included sixteen species from nine sections of *Hibiscus* and revealed that *Trichospermum* and *Bombicella* were not monophyletic. Moreover, the discrimination power of these markers was also compromised and could not resolve

Hibiscus platanifolius of section *Spatula* and *Hibiscus lunariifolius* of section *Trichospermum* with the combination of all these markers. In the current study, we suggest a set of thirty divergence regions (≥ 200 bp) by comparing nucleotide diversity among regions between *H. rosa-sinensis* and *H. syriacus* to solve taxonomic discrepancies and provision of barcode for genus *Hibiscus*. All the suggested regions belonged to IGS regions, and these might be helpful for the development of molecular markers for phylogenetic and phylogeographic studies. The nucleotide diversity estimated for *matK*, *trnH-psbA*, *rbcL*, *ndhF* and *rpl16* intron was 0.0178, 0.0553, 0.0184 and 0.0152, respectively. Moreover, the markers employed previously also showed polymorphism and could be used for the deep divergence in the family Malvaceae but not for *Hibiscus*. In the present study, the thirty sequences identified in the current study had nucleotide diversity 0.0933 to 0.033 from highest to lowest. So, the sequences identified in the current study are highly polymorphic as compared to the sequences that were used in the previous studies. Therefore, based on data reported in the current study, the robust and authentic markers can be developed for these regions and can be used for the phylogenetics, phylogeographic studies and barcoding of *Hibiscus*.

Conclusions
and
Future perspectives

Conclusions

The present study provides broad insight into the evolutionary dynamics of family Malvaceae based on comparative analyses of large number of species from basal lineages to crown groups. The high similarities in genes content, organisation, introns content and GC content in the seven subfamilies of Malvaceae revealed close resemblance in these species at molecular level. The variation in the length of the complete chloroplast genomes and in its different regions existed due to variation in the length of intergenic spacer regions and IRs contraction and expansion. The IRs contraction and expansion lead to generation of pseudogenes or caused complete duplication or reduction of single copy of gene in chloroplast genome of family Malvaceae. The rate of synonymous and non-synonymous substitutions revealed about 95% similarities which further confirmed the close resemblance of the species. The positive selection pressure has been observed in some species which revealed that these genes might be involved in the adaptation of the species and important to the species in their ecological niches.

The analyses of correlations and regression among substitutions, InDels, and oligonucleotide repeats confirm the existence of correlations in these mutational events in family Malvaceae (eudicot, angiosperm) which lead to hypothesis that this might be a common character of all plant lineages and if this hypothesis is confirmed in future independent studies of other families, then required changes might be suggested in the most acceptable model of phylogeny inferring, the General Time Reversible model (GTR model). The phylogeny inferring of family Malvaceae based on complete chloroplast genome attest the previous classification of the family into nine subfamilies. Moreover, our results showed the low efficacy of SSC and IR regions in the inferring of phylogeny and could not be used as region of choice in phylogeny inferring.

Our result of the comparative analyses of three genera including *Theobroma*, *Firmiana* and *Hibiscus* reveals high similarities in inter genus and intra genus comparison when analysed for codon usage, amino acids frequency, RNA editing sites, and simple sequence repeats. The number of oligonucleotide repeats was found two times higher in the crown group (*Hibiscus*) in comparison to basal groups (*Firmiana* and *Theobroma*). This indicates increase in repeats from basal lineages to crown groups. The screening of divergence regions reveals differences in the high polymorphic regions in intra-genus level comparison. This observation suggests the use of genus-specific mutational hotspots might be able to accurately resolve the phylogeny of complex genera with taxonomic discrepancies. The mutational hotspots identified in current study might be suitable for the development of authentic, robust, and cost-effective markers for inferring of phylogeny within these groups with complex taxonomy. Therefore, mutational

hotspots identified in current study specifically in genus *Hibiscus* might be helpful in resolving at inter-genus and intra-genus level taxonomic discrepancies.

Future perspectives

- The advancement of next generation sequencing technology made feasible chloroplast genome sequencing with low cost. So, the extensive sequencing from the several other genera of family Malvaceae can further enhance our knowledge about the evolutionary dynamics in family Malvaceae.
- Correlations analyses among substitutions, InDels, and repeats in independent studies of other plant families might confirm correlations in these mutational events as common character of all plant lineages, which might be suggestive for changes in the most acceptable model of phylogeny inferring, GTR model.
- The low coverage depth sequencing of the extensive and diverse species of Malvaceae for phylogeny inferring might provide more broad insight into phylogenetic relationship of Malvaceae.
- The mutational hotspots identified in this report could be used for development of authentic and robust markers and could be employed in inferring of phylogeny of respective genera specifically in genus *Hibiscus* in which certain taxonomic discrepancies exist in inter-genus and intra-genus levels classification. Moreover, these mutational hotspots can be used as barcodes for inter-species and intra-species discriminations.
- The mutational hotspots identified within genus *Firmiana* might be used for inferring of phylogeny of subfamily Sterculioideae in which certain clades could not be resolved in previous studies

References

- Ahmed, I., 2014. Evolutionary dynamics in taro. PhD Dissertation, Massey University, Palmerston North, New Zealand.
- Ahmed, I., Biggs, P.J., Matthews, P.J., Collins, L.J., Hendy, M.D., Lockhart, P.J., 2012. Mutational dynamics of aroid chloroplast genomes. *Genome Biol. Evol.* 4, 1316–1323. <https://doi.org/10.1093/gbe/evs110>
- Ahmed, I., Islam, M., Arshad, W., Mannan, A., Ahmad, W., Mirza, B., 2009. High-quality plant DNA extraction for PCR: an easy approach. *J. Appl. Genet.* 50, 105–107. <https://doi.org/10.1007/BF03195661>
- Ahmed, I., Matthews, P.J., Biggs, P.J., Naeem, M., Mclenachan, P.A., Lockhart, P.J., 2013. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Mol. Ecol. Resour.* 13, 929–937. <https://doi.org/10.1111/1755-0998.12128>
- Ajaib, M., Wahla, S.Q., Khan, K.M., 2014. *Firmiana Simplex*: A potential source of antimicrobials. *J. Chem. Soc. Pakistan* 36, 744–753.
- Akbar, S., Hanif, U., Ali, J., Ishtiaq, S., 2014. Pharmacognostic studies of stem, roots and leaves of *Malva parviflora* L. *Asian Pac. J. Trop. Biomed.* 4, 410–415. <https://doi.org/10.12980/APJTB.4.2014C1107>
- Akpan, G.A., 2000. Cytogenetic characteristics and the breeding system in six *Hibiscus* species. *Theor. Appl. Genet.* 100, 315–318. <https://doi.org/10.1007/s001220050041>
- Al Muqarrabun, L.M.R., Ahmat, N., 2015. Medicinal uses, phytochemistry and pharmacology of family Sterculiaceae: A review. *Eur. J. Med. Chem.* 92, 514–530. <https://doi.org/10.1016/J.EJMECH.2015.01.026>
- Alam, P., Al-Yousef, H.M., Siddiqui, N.A., Alhowiriny, T.A., Alqasoumi, S.I., Amina, M., Hassan, W.H.B., Abdelaziz, S., Abdalla, R.H., 2018. Anticancer activity and concurrent analysis of ursolic acid, β -sitosterol and lupeol in three different *Hibiscus* species (aerial parts) by validated HPTLC method. *Saudi Pharm.* 26 (7): 1060-1067. <https://doi.org/10.1016/j.jsps.2018.05.015>
- Alipour, H., Bihamta, M.R., Mohammadi, V., 2017. Genotyping-by-Sequencing (GBS) Revealed molecular genetic diversity of Iranian wheat landraces and cultivars. *Front. Plant Sci.* 8, 1–14. <https://doi.org/10.3389/fpls.2017.01293>
- Alverson, W.S., Whitlock, B.A., Nyffeler, R., Bayer, C., Baum, D.A., 1999. Phylogeny of the
-
- Evolutionary dynamics and phylogeny of family Malvaceae* 180

- core Malvales: evidence from *ndhF* sequence data. *Am. J. Bot.* 86, 1474–1486. <https://doi.org/10.2307/2656928>
- Alves, R.M., Sebbenn, A.M., Artero, A.S., Clement, C., Figueira, A., 2007. High levels of genetic divergence and inbreeding in populations of cupuassu (*Theobroma grandiflorum*). *Tree Genet. Genomes* 3, 289–298. <https://doi.org/10.1007/s11295-006-0066-9>
- Amiryousefi, A., Hyvönen, J., Poczai, P., 2018. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLoS One* 13, 1–23. <https://doi.org/10.1371/journal.pone.0196069>
- APG, 2003. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* 141, 399–436.
- APG, 1998. An ordinal classification for the families of flowering plants. *Ann. Missouri Bot. Gard.* 85, 531–553. <https://doi.org/10.2307/2992015>
- Asaf, S., Khan, A.L., Khan, M.A., Shahzad, R., Lubna, Kang, S.M., Al-Harrasi, A., Al-Rawahi, A., Lee, I.J., 2018. Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. *PLoS One* 13, 1–29. <https://doi.org/10.1371/journal.pone.0192966>
- Atakpama, W., Atakpama, W., Batawila, K., Gnamkoulaba, A., Akpagana, K., 2015. Quantitative approach of *Sterculia setigera* Del. (Sterculiaceae) ethnobotanical uses among rural communities in togo (West Africa). *Ethnobot. Res. Appl.* 14, 063–080. <http://dx.doi.org/10.17348/era.14.0.063-080>.
- Ayanbamiji, T.A., Ogundipe, O.T., Olowokudejo, J.D., 2012. Taxonomic significance of the epicalix in the genus *Hibiscus* (Malvaceae). *Phytol. Balc.* 18, 135–140.
- Azam, M.N.K., Ahmed, M.N., Rahman, M.M., Rahmatullah, M., 2013. Ethnomedicines used by the Oraon and Gor tribes of Sylhet district, Bangladesh. *Am. J. Sustain. Agric.* 7, 391–402.
- Baatartsogt, T., Bui, V.N., Trinh, D.Q., Yamaguchi, E., Gronsang, D., Thampaisarn, R., Ogawa, H., Imai, K., 2016. High antiviral effects of *Hibiscus* tea extract on the H5 subtypes of low and highly pathogenic avian influenza viruses. *J. Vet. Med. Sci.* 78, 1405–1411. <https://doi.org/10.1292/jvms.16-0124>
- Bae, S.H., Younis, A., Hwang, Y.-J., Lim, K.-B., 2015. Various pollen morphology in *Hibiscus syriacus*. *Flower Res. J.* 23, 125–130. <https://doi.org/10.11623/frj.2015.23.3.2>
- Baer, C.F., Miyamoto, M.M., Denver, D.R., 2007. Mutation rate variation in multicellular

References

- eukaryotes: causes and consequences. *Nat. Rev. Genet.* 8, 619–631. <https://doi.org/10.1038/nrg2158>
- Barbara, A.W., Bayer, C., David, A.B., 2001. Phylogenetic relationships and floral evolution of the Byttnerioideae ("Sterculiaceae" or Malvaceae s . l.) Based on sequences of the chloroplast gene *ndhF* . *American. Syst. Bot.* 26, 420–437. <https://doi.org/10.1043/0363-6445-26.2.420>
- Barrett, C.F., Freudenstein, J. V., Li, J., Mayfield-Jones, D.R., Perez, L., Pires, J.C., Santos, C., 2014. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. *Mol. Biol. Evol.* 31, 3095–3112. <https://doi.org/10.1093/molbev/msu252>
- Bartley, B.G.D., 2005. The genetic diversity of cacao and its utilization. CABI, Wallingford. <https://doi.org/10.1079/9780851996196.0000>
- Baum, D.A., Smith, S.D., Yen, A., Alverson, W.S., Nyffeler, T., Whitlock, B.A., Oldham, R.L., 2004. Phylogenetic relationships of Malvatheca (Bombacoideae and Malvoideae; Malvaceae sensu lato) as inferred from plastid DNA sequences. *Am. J. Bot.* 91, 1863–1871. <https://doi.org/10.3732/ajb.91.11.1863>
- Bausher, M.G., Singh, N.D., Lee, S.-B., Jansen, R.K., Daniell, H., 2006. The complete chloroplast genome sequence of *Citrus sinensis* (L.) Osbeck var “Ridge Pineapple”: organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* 6, 21. <https://doi.org/10.1186/1471-2229-6-21>
- Bayer, C., Fay, M.F., De Bruijn, A.Y., Savolainen, V., Morton, C.M., Kubitzki, K., Alverson, W.S., Chase, M.W., 1999. Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: A combined analysis of plastid *atpB* and *rbcL* DNA sequences. *Bot. J. Linn. Soc.* 129, 267–303. <https://doi.org/10.1006/bojl.1998.0226>
- Bayer, C., Kubitzki, K., 2003. Malvaceae, in: Flowering plants · Dicotyledons. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 225–311. https://doi.org/10.1007/978-3-662-07255-4_28
- Berry, S.K., 1982. Fatty acid composition and cyclopropene fatty acid content of China-chestnuts (*Sterculia monosperma* , ventenat). *J. Am. Oil Chem. Soc.* 59, 57–58. <https://doi.org/10.1007/BF02670070>
- Bi, Y., Zhang, M.F., Xue, J., Dong, R., Du, Y.P., Zhang, X.H., 2018. Chloroplast genomic resources for phylogeny and DNA barcoding: A case study on *Fritillaria*. *Sci. Rep.* 8, 1–

12. <https://doi.org/10.1038/s41598-018-19591-9>

Borssum Waalkes, J., 1966. Malesian Malvaceae revised. *Blumea* 14, 1–213.

Bouriche, H., Meziti, H., Senator, A., Arnhold, J., 2011. Anti-inflammatory, free radical-scavenging, and metal-chelating activities of *Malva parviflora*. *Pharm. Biol.* 49, 942–946. <https://doi.org/10.3109/13880209.2011.558102>

Braukmann, T.W.A., Kuzmina, M., Stefanović, S., 2009. Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. *Curr. Genet.* 55, 323–337. <https://doi.org/10.1007/s00294-009-0249-7>

Burret, M., 1926. Beiträge zur Kenntnis der Tiliaceen. *Notizblatt des Bot. Gartens und Museums zu Berlin-Dahlem* 592–880.

Burrows, P.A., 1998. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid *ndh* genes. *EMBO J.* 17, 868–876. <https://doi.org/10.1093/emboj/17.4.868>

Cai, J., Ma, P.F., Li, H.T., Li, D.Z., 2015. Complete plastid genome sequencing of four *Tilia* species (Malvaceae): A comparative analysis and phylogenetic implications. *PLoS One* 10, 1–13. <https://doi.org/10.1371/journal.pone.0142705>

Cai, Z., Guisinger, M., Kim, H.-G., Ruck, E., Blazier, J.C., McMurtry, V., Kuehl, J. V., Boore, J., Jansen, R.K., 2008. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J. Mol. Evol.* 67, 696–704. <https://doi.org/10.1007/s00239-008-9180-7>

Cao, J., Jiang, D., Zhao, Z., Yuan, S., Zhang, Y., Zhang, T., Zhong, W., Yuan, Q., Huang, L., 2018. Development of chloroplast genomic resources in Chinese Yam (*Dioscorea polystachya*). *Biomed Res. Int.* 2018, 1–11. <https://doi.org/10.1155/2018/6293847>

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M., Dopazo, J., 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. *Mol. Biol. Evol.* 32, 2015–2035. <https://doi.org/10.1093/molbev/msv082>

Carvalho-Sobrinho, J.G., Alverson, W.S., Alcantara, S., Queiroz, L.P., Mota, A.C., Baum, D.A., 2016. Revisiting the phylogeny of Bombacoideae (Malvaceae): Novel relationships, morphologically cohesive clades, and a new tribal classification based on multilocus phylogenetic analyses. *Mol. Phylogenet. Evol.* 101, 56–74. <https://doi.org/10.1016/j.ympev.2016.05.006>

- Caspermeyer, J., 2015. Most comprehensive study to date reveals evolutionary history of Citrus. *Mol. Biol. Evol.* 32, 2217–8. <https://doi.org/10.1093/molbev/msv101>
- Cavalcante, P.B., 1991. Frutas comestíveis da Amazônia. Edições CEJUP.
- Chen, J.-Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., Tian, D., 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* 26, 1523–1531. <https://doi.org/10.1093/molbev/msp063>
- Chen, L., Bao, S., Wang, S., Xiong, J., Lee, P., Leung, B.L., Lin, C., Hu, L., Yang, X., Lin, G., Tan, Y., 2016. Effect of *Urena procumbens* on *CYP450* isoforms activity of rats by cocktail method, *Int. Clin. Exp. Med.*, 9 (3):6681-6686.
- Chen, N., Sha, L.-N., Dong, Z.-Z., Tang, C., Wang, Y., Kang, H.-Y., Zhang, H.-Q., Yan, X.-B., Zhou, Y.-H., Fan, X., 2018. Complete structure and variation of the chloroplast genome of *Agropyron cristatum* (L.) Gaertn. *Gene* 640, 86–96. <https://doi.org/10.1016/J.GENE.2017.10.009>
- Chen, S., Xiao, P., 2010. Molecular evolution and positive Darwinian selection of the chloroplast maturase *matK*. *J. Plant Res.* 123, 241–247.
- Chen, S.F., Li, M.W., Jing, H.J., Zhou, R.C., Yang, G.L., Wu, W., Fan, Q., Liao, W.B., 2015. *De novo* transcriptome assembly in *Firmiana danxiaensis*, a tree species endemic to the danxia landform. *PLoS One* 10, e013973. <https://doi.org/10.1371/journal.pone.0139373>
- Chen, Z., Feng, K., Grover, C.E., Li, P., Liu, F., Wang, Y., Xu, Q., Shang, M., Zhou, Z., Cai, X., Wang, X., Wendel, J.F., Wang, K., Hua, J., 2016. Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton species. *PLoS One* 11, 1–16. <https://doi.org/10.1371/journal.pone.0157183>
- Chen, Z., Grover, C.E., Li, P., Wang, Y., Nie, H., Zhao, Y., Wang, M., Liu, F., Zhou, Z., Wang, X., Cai, X., Wang, K., Wendel, J.F., Hua, J., 2017. Molecular evolution of the plastid genome during diversification of the cotton genus. *Mol. Phylogenet. Evol.* 112, 268–276. <https://doi.org/10.1016/j.ympbev.2017.04.014>
- Cheon, K.-S., Kim, K.-A., Kwak, M., Lee, B., Yoo, K.-O., 2019. The complete chloroplast genome sequences of four *Viola* species (Violaceae) and comparative analyses with its congeneric species. *PLoS One* 14, e0214162. <https://doi.org/10.1371/journal.pone.0214162>
- Cheon, S.H., Jo, S., Kim, H.W., Kim, Y.K., Sohn, J.Y., Kim, K.J., 2017. The complete plastome sequence of Durian, *Durio zibethinus* L. (Malvaceae). *Mitochondrial DNA Part*

- B Resour. 2, 763–764. <https://doi.org/10.1080/23802359.2017.1398615>
- Chevenet, F., Brun, C., Bañuls, A.-L., Jacq, B., Christen, R., 2006. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* 7, 439. <https://doi.org/10.1186/1471-2105-7-439>
- Cho, K.S., Yun, B.K., Yoon, Y.H., Hong, S.Y., Mekapogu, M., Kim, K.H., Yang, T.J., 2015. Complete chloroplast genome sequence of tartary buckwheat (*Fagopyrum tataricum*) and comparative analysis with common buckwheat (*F. esculentum*). *PLoS One* 10, 1–14. <https://doi.org/10.1371/journal.pone.0125332>
- Choi, K.S., Chung, M.G., Park, S., 2016. The complete chloroplast genome sequences of three *Veroniceae* species (Plantaginaceae): Comparative analysis and highly divergent regions. *Front. Plant Sci.* 7, 1–8. <https://doi.org/10.3389/fpls.2016.00355>
- Choi, K.S., Kwak, M., Lee, B., Park, S.J., 2018. Complete chloroplast genome of *Tetragonia tetragonioides*: Molecular phylogenetic relationships and evolution in Caryophyllales. *PLoS One* 13, 1–11. <https://doi.org/10.1371/journal.pone.0199626>
- Chris Blazier, J., Guisinger, M.M., Jansen, R.K., 2011. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Mol. Biol.* 76, 263–272. <https://doi.org/10.1007/s11103-011-9753-5>
- Christenhusz, M.J.M., Byng, J.W., 2016. The number of known plants species in the world and its annual increase. *Phytotaxa* 261, 201–217. <https://doi.org/10.11646/phytotaxa.261.3.1>
- Cooper, G., 2000. Chloroplasts and other plastids in the cell: A molecular approach, 2nd ed. Sunderland (MA): Sinauer Associates.
- Cornejo, O.E., Yee, M.-C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., Livingstone III, D., Stack, C., Romero, A., Umaharan, P., Royaert, S., Tawari, N.R., Pauline, N., Schnell, R., Wilbert, P., Mockaitis, K., Bustamante, C.D., Motamayor, J.C., 2017. Genomic insights into the domestication of the chocolate tree, *Theobroma cacao* L. *bioRxiv* 1–25. <https://doi.org/10.1101/223438>
- Craven, L.A., Wilson, F.D., Fryxell, P.A., 2003. A taxonomic review of *Hibiscus* sect. *Furcaria* (Malvaceae) in Western Australia and the Northern Territory. *Aust. Syst. Bot.* 16, 185–218. <https://doi.org/10.1071/SB01046>
- Cremen, M.C.M., Leliaert, F., West, J., Lam, D.W., Shimada, S., Lopez-Bautista, J.M., Verbruggen, H., 2019. Reassessment of the classification of Bryopsidales (Chlorophyta) based on chloroplast phylogenomic analyses. *Mol. Phylogenet. Evol.* 130, 397–405.

- <https://doi.org/10.1016/J.YMPEV.2018.09.009>
- Cronquist, A., 1988. The evolution and classification of flowering plants. New York Botanical Garden.
- Cronquist, A., 1981. An integrated system of classification of flowering plants. Columbia University Press.
- Cuatrecasas, J., 1964. Cacao and its allies: A taxonomic revision of the genus *Theobroma*. In Systematic plant studies. 35, 379–614.
- Cummings, H.S., Hershey, J.W., 1994. Translation initiation factor IF1 is essential for cell viability in *Escherichia coli*. J. Bacteriol. 176, 198–205. <https://doi.org/10.1128/jb.176.1.198-205.1994>
- Curci, P.L., De Paola, D., Danzi, D., Vendramin, G.G., Sonnante, G., 2015. Complete chloroplast genome of the multifunctional crop *Globe artichoke* and comparison with other Asteraceae. PLoS One 10, 1–18. <https://doi.org/10.1371/journal.pone.0120589>
- Cusack, B.P., Wolfe, K.H., 2007. When gene marriages don't work out: divorce by subfunctionalization. Trends Genet. 23, 270–272. <https://doi.org/10.1016/J.TIG.2007.03.010>
- Dahlgren, R., 1983. General aspects of angiosperm evolution and macrosystematics. Nord. J. Bot. 3, 119–149. <https://doi.org/10.1111/j.1756-1051.1983.tb01448.x>
- Dalar, A., Türker, M., Konczak, I., 2012. Antioxidant capacity and phenolic constituents of *Malva neglecta* Wallr. and *Plantago lanceolata* L. from Eastern Anatolia region of Turkey. J. Herb. Med. 2, 42–51. <https://doi.org/10.1016/J.HERMED.2012.03.001>
- Daniell, H., 2007. Transgene containment by maternal inheritance: effective or elusive? Proc. Natl. Acad. Sci. U. S. A. 104, 6879–6880. <https://doi.org/10.1073/pnas.0702219104>
- Daniell, H., Lin, C.-S., Yu, M., Chang, W.-J., 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 17, 134. <https://doi.org/10.1186/s13059-016-1004-2>
- DellaGreca, M., Cutillo, F., Abrosca, B. D', Fiorentino, A., Pacifico, S., Zarrelli, A., 2009. Antioxidant and radical scavenging properties of *Malva sylvestris*. Nat. Prod. Commun. 4, 1934578X0900400. <https://doi.org/10.1177/1934578X0900400702>
- Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., Gascuel, O., 2008. Phylogeny.fr:

- robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36, W465–W469. <https://doi.org/10.1093/nar/gkn180>
- Dong, W., Xu, C., Li, D., Jin, X., Li, R., Lu, Q., Suo, Z., 2016. Comparative analysis of the complete chloroplast genome sequences in psammophytic *Haloxylon* species (Amaranthaceae). *PeerJ* 4, e2699. <https://doi.org/10.7717/peerj.2699>
- Downie, S.R., Olmstead, R.G., Zurawski, G., Soltis, D.E., Soltis, P.S., Watson, J.C., Palmer, J.D., 1991. Six independent losses of the chloroplast DNA *rpl 2* intron in dicotyledons: molecular and phylogenetic implications. *Evolution (N.Y.)*. 45, 1245–1259. <https://doi.org/10.1111/j.1558-5646.1991.tb04390.x>
- Drouin, G., Daoud, H., Xia, J., 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49, 827–831. <https://doi.org/10.1016/j.ympev.2008.09.009>
- Du, Y., Bi, Y., Yang, F., Zhang, M., Chen, X., Xue, J., Zhang, X., 2017. Complete chloroplast genome sequences of *Lilium*: insights into evolutionary dynamics and phylogenetic analyses. *Sci. Rep.* 7, 5751. <https://doi.org/10.1038/s41598-017-06210-2>
- Duarte, M.C., Esteves, G.L., Salatino, M.L.F., Walsh, K.C., David, A., Duarte, M.C., Esteves, G.L., Salatino, M.L.F., Walsh, K.C., Baum, D.A., 2011. Phylogenetic analyses of *Eriotheca* and related genera (Bombacoideae, Malvaceae). 36, 690–701. <https://doi.org/10.1600/036364411X583655>
- Esteves, G.L., 2000. Taxonomic characters of the staminal tube and epicalyx in *Brazilian pavonia* (Malvaceae). *Brittonia* 52, 256–264. <https://doi.org/10.2307/2666576>
- Fan, Q., Chen, S., Li, M., He, S., Zhou, R., Liao, W., 2013. Development and characterization of microsatellite markers from the transcriptome of *Firmiana danxiaensis* (Malvaceae s.l.). *Appl. Plant Sci.* 1, 1300047. <https://doi.org/10.3732/apps.1300047>
- Feng, Y., Comes, H.P., Zhou, X.P., Qiu, Y.X., 2019. Phylogenomics recovers monophyly and early tertiary diversification of *Dipteronia* (Sapindaceae). *Mol. Phylogenet. Evol.* 130, 9–17. <https://doi.org/10.1016/j.ympev.2018.09.012>
- Fryxell, P., 1980. A revision of the American species of *Hibiscus* section *Bombicella* (Malvaceae). The administration, 1624.
- Fryxell, P.A., 1997. The American genera of Malvaceae-II. *Brittonia* 49 (2), 204–269. <https://doi.org/10.2307/2807683>
- Gantt, J.S., Baldauf, S.L., Calie, P.J., Weeden, N.F., Palmer, J.D., 1991. Transfer of *rpl22* to

References

- the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J.* 10, 3073–3078. <https://doi.org/10.1002/j.1460-2075.1991.tb07859.x>
- Gao, Y., Wang, H., Liu, C., Chu, H., Yan, Y., Tang, L., 2018. Complete chloroplast genome sequence of the red silk cotton tree (*Bombax ceiba*). *Mitochondrial DNA Part B Resour.* 3, 315–316. <https://doi.org/10.1080/23802359.2017.1422399>
- Goldberg, K.H., Yin, A.C., Mupparapu, A., Retzbach, E.P., Goldberg, G.S., Yang, C.F., 2017. Components in aqueous *Hibiscus rosa-sinensis* flower extract inhibit in vitro melanoma cell growth. *J. Tradit. Complement. Med.* 7, 45–49. <https://doi.org/10.1016/J.JTCME.2016.01.005>
- Gopaulchan, D., Motilal, L.A., Bekele, F.L., Clause, S., Ariko, J.O., Ejang, H.P., Umaharan, P., 2019. Morphological and genetic diversity of cacao (*Theobroma cacao* L.) in Uganda. *Physiol. Mol. Biol. Plants* 25, 361–375. <https://doi.org/10.1007/s12298-018-0632-2>
- Graham, S.W., Reeves, P.A., Burns, A.C.E., Olmstead, R.G., 2000. Microstructural Changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *Int. J. Plant Sci.* 161, S83–S96. <https://doi.org/10.1086/317583>
- Gruenheit, N., Lockhart, P.J., Steel, M., Martin, W., 2008. Difficulties in Testing for Covarion-Like Properties of sequences under the confounding influence of changing proportions of variable sites. *Mol. Biol. Evol.* 25, 1512–1520. <https://doi.org/10.1093/molbev/msn098>
- Gu, C., Tembrock, L.R., Johnson, N.G., Simmons, M.P., Wu, Z., 2016. The complete plastid genome of *Lagerstroemia fauriei* and loss of *rpl2* intron from *Lagerstroemia* (Lythraceae). *PLoS One* 11, e0150752. <https://doi.org/10.1371/journal.pone.0150752>
- Guo, X., Castillo-Ramírez, S., González, V., Bustos, P., Luís Fernández-Vázquez, J., Santamaría, R., Arellano, J., Cevallos, M.A., Dávila, G., 2007. Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts. *BMC Genomics* 8, 228. <https://doi.org/10.1186/1471-2164-8-228>
- Gutiérrez-López, N., Ovando-Medina, I., Salvador-Figueroa, M., Molina-Freaner, F., Avendaño-Arrazate, C.H., Vázquez-Ovando, A., 2016. Unique haplotypes of cacao trees as revealed by *trnH-psbA* chloroplast DNA. *PeerJ* 4, e1855. <https://doi.org/10.7717/peerj.1855>
- Han, Y., 2009. Rutin has therapeutic effect on septic arthritis caused by *Candida albicans*. *Int. Immunopharmacol.* 9, 207–211. <https://doi.org/10.1016/J.INTIMP.2008.11.002>

References

- Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., Schwartz, S., Furey, T.S., Whelan, S., Goldman, N., Smit, A., Miller, W., Chiaromonte, F., Haussler, D., 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* 13, 13–26. <https://doi.org/10.1101/gr.844103>
- Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174. <https://doi.org/10.1007/BF02101694>
- He, L., Qian, J., Li, X., Sun, Z., Xu, X., Chen, S., McPhee, D.J., 2017. Complete chloroplast genome of medicinal plant *Lonicera japonica*: Genome rearrangement, intron gain and loss, and implications for phylogenetic studies. *Molecules* 22, 3–4. <https://doi.org/10.3390/molecules22020249>
- He, Y., Xiao, H., Deng, C., Xiong, L., Yang, J., Peng, C., 2016. The complete chloroplast genome sequences of the medicinal plant *Pogostemon Cablin*. *Int. J. Mol. Sci.* 17. <https://doi.org/10.3390/ijms17060820>
- Heckenhauer, J., Paun, O., Chase, M.W., Ashton, P.S., Kamariah, A.S., Samuel, R., 2019. Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastid genomes. *Ann. Bot.* 123, 857–865. <https://doi.org/10.1093/aob/mcy220>
- Henriquez, C.L., Arias, T., Pires, J.C., Croat, T.B., Schaal, B.A., 2014. Phylogenomics of the plant family Araceae. *Mol. Phylogenet. Evol.* 75, 91–102. <https://doi.org/10.1016/j.ympev.2014.02.017>
- Heywood, V.H., Moore, D.M., Dunkley, J., King, C., 1993. Flowering plants of the world. Vol. 336. Oxford: Oxford University Press.
- Hinsley, S.R., 2004. The *Malva* (Mallow) Pages: Contents and Overview. <http://www.malvaceae.info/Genera/Malva/Malva.html#Introduction> (accessed 5.18.19).
- Hinsley, S., 2006. Classification of Malvaceae: Overview. <http://www.malvaceae.info/Classification/overview.html> (accessed 3.1.19).
- Hinsley, S.R., 2008. Economic Uses of Malvaceae - Overview. <http://www.malvaceae.info/Economic/Overview.html> (accessed 5.11.19).
- Hinsley, S.R., 2009. Sections and Segregates of *Hibiscus*. <http://www.malvaceae.info/Genera/Hibiscus/sections.php> (accessed 3.4.19).
- Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2:

- Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. <https://doi.org/10.1093/molbev/msx281>
- Hoch, B., Maier, R.M., Appel, K., Igloi, G.L., Kössel, H., 1991. Editing of a chloroplast mRNA by creation of an initiation codon. *Nature* 353, 178–180. <https://doi.org/10.1038/353178a0>
- Hochreutiner, B., 1900. Revision du genre *Hibiscus*. Romet, Genève.
- Hu, Q.-H., Wang, C., Li, J.-M., Zhang, D.-M., Kong, L.-D., 2009. Allopurinol, rutin, and quercetin attenuate hyperuricemia and renal dysfunction in rats induced by fructose intake: renal organic ion transporter involvement. *Am. J. Physiol. Physiol.* 297, F1080–F1091. <https://doi.org/10.1152/ajprenal.90767.2008>
- Huang, L.-S., Sun, Y.-Q., Jin, Y., Gao, Q., Hu, X.-G., Gao, F.-L., Yang, X.-L., Zhu, J.-J., El-Kassaby, Y.A., Mao, J.-F., 2018. Development of high transferability cpSSR markers for individual identification and genetic investigation in Cupressaceae species. *Ecol. Evol.* 8, 4967–4977. <https://doi.org/10.1002/ece3.4053>
- Huang, Y.-S., Wu, W.-H., Xu, W.-B., Liu, Y., 2011. *Firmiana calcarea* sp. nov. (Malvaceae) from limestone areas in Guangxi, China. *Nord. J. Bot.* 29, 608–610. <https://doi.org/10.1111/j.1756-1051.2011.01278.x>
- Huang, Y.Y., Cho, S.T., Haryono, M., Kuo, C.H., 2017. Complete chloroplast genome sequence of common bermudagrass (*Cynodon dactylon* (L.) Pers.) and comparative analysis within the family Poaceae. *PLoS One* 12, 1–16. <https://doi.org/10.1371/journal.pone.0179055>
- Ibrahim, R.I.H., Azuma, J.-I., Sakamoto, M., 2006. Complete nucleotide sequence of the cotton (*Gossypium barbadense* L.) chloroplast genome with a comparative analysis of sequences among 9 dicot plants. *Genes Genet. Syst.* 81, 311–321. <https://doi.org/10.1266/ggs.81.311>
- Iram, S., Hayat, M.Q., Tahir, M., Gul, A., Abdullah, Ahmed, I., 2019. Chloroplast genome sequence of *Artemisia scoparia*: Comparative analyses and screening of mutational hotspots. *Plants* 8, 476. <https://doi.org/doi:10.3390/plants8110476>
- Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Muller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.-B., Peery, R., McNeal, J.R., Kuehl, J. V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci.* 104, 19369–19374. <https://doi.org/10.1073/pnas.0709121104>

References

- Jansen, R.K., Sasaki, C., Lee, S.-B., Hansen, A.K., Daniell, H., 2011. Complete plastid genome sequences of three Rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for at least two independent transfers of *rpl22* to the nucleus. *Mol. Biol. Evol.* 28, 835–847. <https://doi.org/10.1093/molbev/msq261>
- Jheng, C.-F., Chen, Tien-Chih, Lin, J.-Y., Chen, Ting-Chieh, Wu, W.-L., Chang, C.-C., 2012. The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis orchids*. *Plant Sci.* 190, 62–73. <https://doi.org/10.1016/j.plantsci.2012.04.001>
- Jiang, P., Shi, F.-X., Li, M.-R., Liu, B., Wen, J., Xiao, H.-X., Li, L.-F., 2018. Positive selection driving cytoplasmic genome evolution of the medicinally important Ginseng plant genus *Panax*. *Front. Plant Sci.* 9, 359. <https://doi.org/10.3389/fpls.2018.00359>
- Joh, H.J., Kim, N., Jayakodi, M., Jang, W., Park, J.Y., Kim, Y.C., 2017. Authentication of Golden-berry P. ginseng cultivar ‘Gumpoong’ from a Landrace ‘Hwangsook’ based on pooling method using chloroplast-derived markers 2017, 16–24.
- Joseph, B.O., 1977. A revision of species segregated from *Hibiscus* sect. *Trionum* (Medicus) de Candolle Sensu lato (Malvaceae). Cornell Univeristy.
- Judd, W.S., Manchester, S.R., 1997. Circumscription of Malvaceae (Malvales) as determined by a preliminary cladistic analysis of morphological, anatomical, palynological, and chemical characters. *Brittonia* 49, 384. <https://doi.org/10.2307/2807839>
- Jukes, T., Cantor, C., 1969. Evolution of protein molecules. *Mamm. protein Metab.* 3, 132.
- Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. <https://doi.org/10.1038/nmeth.4285>
- Kane, N., Sveinsson, S., Dempewolf, H., Yang, J.Y., Zhang, D., Engels, J.M.M., Cronk, Q., 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot.* 99, 320–329. <https://doi.org/10.3732/ajb.1100570>
- Katoh, K., Kuma, K.I., Toh, H., Miyata, T., 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. <https://doi.org/10.1093/nar/gki198>
- Kawata, M., Harada, T., Shimamoto, Y., Oono, K., Takaiwa, F., 1997. Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted

- plastid DNAs (ptDNAs). *Curr. Genet.* 31, 179–84.
- 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. <https://doi.org/10.1371/journal.pone.0051687>
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199>
- Khonsung, P., Nantsupawat, S., Nimmannit Jesadanont, S., Chantharateptawan, V., Panthong, A., 2006. Anti-inflammatory and analgesic activities of water extract of *Malvastrum coromandelianum* (L.) Garcke, *Thai J Pharmacol.*
- Kim, C.K., Seol, Y.J., Perumal, S., Lee, J., Waminal, N.E., Jayakodi, M., Lee, S.C., Jin, S., Choi, B.S., Yu, Y., Ko, H.C., Choi, J.W., Ryu, K.Y., Sohn, S.H., Parkin, I., Yang, T.J., 2018. Re-exploration of U's triangle *Brassica* species based on chloroplast genomes and 45S nrDNA sequences. *Sci. Rep.* 8, 1–11. <https://doi.org/10.1038/s41598-018-25585-4>
- Kim, H.T., Kim, K.-J., 2014. Chloroplast genome differences between Asian and American *Equisetum arvense* (Equisetaceae) and the origin of the hypervariable *trnY-trnE* intergenic spacer. *PLoS One* 9, e103898. <https://doi.org/10.1371/journal.pone.0103898>
- Kim, J., Yang, H., Cho, N., Kim, B., Kim, Y., Sung, S., 2015. Hepatoprotective constituents of *Firmiana simplex* stem bark against ethanol insult to primary rat hepatocytes. *Pharmacogn. Mag.* 11, 55. <https://doi.org/10.4103/0973-1296.149704>
- Kim, K.-J., Lee, H.-L., 2005. Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Mol. Cells* 19, 104–113.
- Kim, Y., Oh, Y.J., Han, K.Y., Kim, G.H., Ko, J., Park, J., 2019. The complete chloroplast genome sequence of *Hibiscus syriacus* L. 'Mamonde' (Malvaceae). *Mitochondrial DNA Part B* 4, 558–559. <https://doi.org/10.1080/23802359.2018.1553526>
- Kim, Y.M., Kim, S., Koo, N., Shin, A.Y., Yeom, S.I., Seo, E., Park, S.J., Kang, W.H., Kim, Myung Shin, Park, J., Jang, I., Kim, P.G., Byeon, I., Kim, Min Seo, Choi, J.H., Ko, G., Hwang, J.H., Yang, T.J., Choi, S.B., Lee, J.M., Lim, K.B., Lee, J., Choi, I.Y., Park, B.S., Kwon, S.Y., Choi, D., Kim, R.W., 2017. Genome analysis of *Hibiscus syriacus* provides

References

- insights of polyploidization and indeterminate flowering in woody plants. *DNA Res.* 24, 71–80. <https://doi.org/10.1093/dnares/dsw049>
- Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. <https://doi.org/10.1007/BF01731581>
- Kimura, M., 1979. Model of effectively neutral mutations in which selective constraint is incorporated 76, 3440–3444.
- Kofer, W., Koop, H.-U., Wanner, G., Steinmüller, K., 1998. Mutagenesis of the genes encoding subunits A, C, H, I, J and K of the plastid NAD(P)H-plastoquinone-oxidoreductase in tobacco by polyethylene glycol-mediated plastome transformation. *Mol. Gen. Genet.* 258, 166. <https://doi.org/10.1007/s004380050719>
- Koopman, M.M., Baum, D.A., 2008. Phylogeny and biogeography of tribe Hibisceae (Malvaceae) on Madagascar. *Syst. Bot.* 33, 364–374. <https://doi.org/10.1600/036364408784571653>
- Kostermans, A.J.G, 1957. The genus *Firmiana Marsili* (Sterculiaceae). *Reinwartia* 4, 281–310.
- Krupinska, K., Melonek, J., Krause, K., 2013. New insights into plastid nucleoid structure and functionality. *Planta* 237, 653–664. <https://doi.org/10.1007/s00425-012-1817-5>
- Kugita, M., Kaneko, A., Yamamoto, Y., Takeya, Y., Matsumoto, T., Yoshinaga, K., 2003. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. *Nucleic Acids Res.* 31, 716–21.
- Kurtz, S., Choudhuri, J. V, Ohlebusch, E., Schleiermacher, C., Stoye, J., Giegerich, R., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642.
- Kwon, H.-Y., Kim, J.-H., Kim, S.-H., Park, J.-M., Lee, H., 2016. The complete chloroplast genome sequence of *Hibiscus syriacus*. *Mitochondrial DNA Part A* 27, 3668–3669. <https://doi.org/10.3109/19401736.2015.1079847>
- Lam, S.K., Ng, T.B., 2009. Novel galactonic acid-binding hexameric lectin from *Hibiscus mutabilis* seeds with antiproliferative and potent HIV-1 reverse transcriptase inhibitory activities.
- Lamont, W.J., 1999. Okra - A versatile vegetable crop. *Horttechnology* 9, 179–184.
- Lawrie, D.S., Messer, P.W., Hershberg, R., Petrov, D.A., 2013. Strong purifying selection at

References

- synonymous sites in *D. melanogaster* 9, 33–40.
<https://doi.org/10.1371/journal.pgen.1003527>
- Lee, S.-B., Kaittanis, C., Jansen, R.K., Hostetler, J.B., Tallon, L.J., Town, C.D., Daniell, H., 2006. The complete chloroplast genome sequence of *Gossypium hirsutum*: organization and phylogenetic relationships to other angiosperms. *BMC Genomics* 7, 61.
<https://doi.org/10.1186/1471-2164-7-61>
- Lee, S., Cho, S., Park, M., Kim, Y., Choi, J., Park, S., 2007. Growth and rutin production in hairy root cultures of Buckwheat (*Fagopyrum esculentum* M.). *Prep. Biochem. Biotechnol.* 37, 239–246. <https://doi.org/10.1080/10826060701386729>
- Lee, S.Y., Ng, W.L., Mohamed, R., Terhem, R., 2018. The complete chloroplast genome of *Aquilaria malaccensis* Lam. (Thymelaeaceae), an important and threatened agarwood-producing tree species. *Mitochondrial DNA Part B* 3, 1120–1121.
<https://doi.org/10.1080/23802359.2018.1519382>
- Lehwark, P., Greiner, S., 2019. GB2sequin - A file converter preparing custom GenBank files for database submission. *Genomics* 111, 759–761.
<https://doi.org/10.1016/J.YGENO.2018.05.003>
- Lei, W., Ni, D., Wang, Y., Shao, J., Wang, X., Yang, D., Wang, J., Chen, H., Liu, C., 2016. Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. *Sci. Rep.* 6, 21669.
<https://doi.org/10.1038/srep21669>
- Leontowicz, M., Leontowicz, H., Jastrzebski, Z., Jesion, I., Haruenkit, R., Poovarodom, S., Katrich, E., Tashma, Z., Drzewiecki, J., Trakhtenberg, S., Gorinstein, S., 2007. The nutritional and metabolic indices in rats fed cholesterol-containing diets supplemented with durian at different stages of ripening. *BioFactors* 29, 123–136.
<https://doi.org/10.1002/biof.552029203>
- Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203–21.
<https://doi.org/10.1093/oxfordjournals.molbev.a040442>
- Li, B., Zheng, Y., 2018. Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. *Sci. Rep.* 8, 9285. <https://doi.org/10.1038/s41598-018-27453-7>
- Li, H., Chiu, C.-C., 2010. Protein transport into chloroplasts. *Annu. Rev. Plant Biol.* 61, 157–

180. <https://doi.org/10.1146/annurev-arplant-042809-112222>
- Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, L.-F., Wang, H.-Y., Zhang, C., Wang, X.-F., Shi, F.-X., Chen, W.-N., Ge, X.-J., 2013. Origins and domestication of cultivated Banana inferred from chloroplast and nuclear genes. *PLoS One* 8, e80502. <https://doi.org/10.1371/journal.pone.0080502>
- Li, P., Zhang, Shujiang, Li, F., Zhang, Shifan, Zhang, H., Wang, X., Sun, R., Bonnema, G., Borm, T.J.A., 2017. A phylogenetic analysis of chloroplast genomes elucidates the relationships of the six economically important *Brassica* species comprising the triangle of U. *front. Plant Sci.* 8, 111. <https://doi.org/10.3389/fpls.2017.00111>
- Li, X., Yang, Y., Henry, R.J., Rossetto, M., Wang, Y., Chen, S., 2014. Plant DNA barcoding: From gene to genome. *Biol. Rev.* <https://doi.org/10.1111/brv.12104>
- Li, Y.-P., Zhang, X., Wu, W.-T., Miao, S.-X., Chang, J.-L., 2015. Chromosome and karyotype analysis of *Hibiscus mutabilis* f. *mutabilis*. *Fron. Life. Sci.*, 8 (3), 300-304. <https://doi.org/10.1080/21553769.2015.1041166>
- Li, Y., Zhang, Z., Yang, J., Lv, G., 2018. Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. *PLoS One* 13. <https://doi.org/10.1371/journal.pone.0194613>
- Lim, S.Y., Subedi, L., Shin, D., Kim, C.S., Lee, K.R., Kim, S.Y., 2017. A new neolignan derivative, balanophonin isolated from *Firmiana simplex* delays the progress of neuronal cell death by inhibiting microglial activation. *Biomol. Ther.* 25, 519–527. <https://doi.org/10.4062/biomolther.2016.224>
- Lim, T.K., Lim, T.K., 2012. *Sterculia monosperma*, in: Edible medicinal and non medicinal plants. Springer Netherlands, pp. 198–200. https://doi.org/10.1007/978-94-007-2534-8_27
- Lin, C.-P., Wu, C.-S., Huang, Y.-Y., Chaw, S.-M., 2012. The complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction. *Genome Biol. Evol.* 4, 374–381. <https://doi.org/10.1093/gbe/evs021>
- Lin, C.-S., Chen, J.J.W., Huang, Y.-T., Chan, M.-T., Daniell, H., Chang, W.-J., Hsu, C.-T., Liao, D.-C., Wu, F.-H., Lin, S.-Y., Liao, C.-F., Deyholos, M.K., Wong, G.K.-S., Albert, V.A., Chou, M.-L., Chen, C.-Y., Shih, M.-C., 2015. The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. *Sci. Rep.* 5, 9040.

- <https://doi.org/10.1038/srep09040>
- Linnaeus, C. V., 1753. *Species Plantarum* 1, 1st ed. Laurentius Salvius, Stockholm.
- Liò, P., Goldman, N., 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8, 1233–44.
- Litz, R., 2005. *Theobroma cacao*. In *Biotechnology of fruit and nut crops*. Vol. 29, CABI.
- Liu, L., Wang, Y., He, P., Li, P., Lee, J., Soltis, D.E., Fu, C., 2018. Chloroplast genome analyses and genomic resource development for epilithic sister genera *Oresitrophe* and *Mukdenia* (Saxifragaceae), using genome skimming data. *BMC Genomics* 19, 235. <https://doi.org/10.1186/s12864-018-4633-x>
- Lockhart, P., Novis, P., Milligan, B.G., Riden, J., Rambaut, A., Larkum, T., 2006. Heterotachy and tree Building: A case study with Plastids and Eubacteria. *Mol. Biol. Evol.* 23, 40–45. <https://doi.org/10.1093/molbev/msj005>
- Lockhart, P.J., McLenachan, P.A., Havell, D., Glenney, D., Huson, D., Jensen, U., 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine Buttercups: Molecular evidence under split decomposition. *Ann. Missouri Bot. Gard.* 88, 458. <https://doi.org/10.2307/3298586>
- Lohse, M., Drechsel, O., Bock, R., 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274. <https://doi.org/10.1007/s00294-007-0161-y>
- Longman-Jacobsen, N., Williamson, J., Dawkins, R., Gaudieri, S., 2003. In polymorphic genomic regions indels cluster with nucleotide polymorphism: Quantum Genomics. Elsevier 312, 257–261.
- Lössl, A.G., Waheed, M.T., 2011. Chloroplast-derived vaccines against human diseases: achievements, challenges and scopes. *Plant Biotechnol. J.* 9, 527–539. <https://doi.org/10.1111/j.1467-7652.2011.00615.x>
- Ma, L., 2018. The complete chloroplast genome sequence of the fragrant plant *Lavandula angustifolia* (Lamiaceae). *Mitochondrial DNA Part B Resour.* 3, 135–136. <https://doi.org/10.1080/23802359.2018.1431067>
- Ma, P.-F., Zhang, Y.-X., Guo, Z.-H., Li, D.-Z., 2015. Evidence for horizontal transfer of mitochondrial DNA to the plastid genome in a *Bamboo* genus. *Sci. Rep.* 5, 11608. <https://doi.org/10.1038/srep11608>

References

- Magee, A.M., Aspinall, S., Rice, D.W., Cusack, B.P., Semon, M., Perry, A.S., Stefanovic, S., Milbourne, D., Barth, S., Palmer, J.D., Gray, J.C., Kavanagh, T.A., Wolfe, K.H., 2010. Localized hypermutation and associated gene losses in Legume chloroplast genomes. *Genome Res.* 20, 1700–1710. <https://doi.org/10.1101/gr.111955.110>
- Martin, G.E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., de Carvalho, J.F., Ainouche, M., Salmon, A., Ainouche, A., 2014. The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: Evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the Legume family. *Ann. Bot.* 113, 1197–1210. <https://doi.org/10.1093/aob/mcu050>
- Matsuoka, Y., Yamazaki, Y., Ogihara, Y., Tsunewaki, K., 2002. Whole chloroplast genome comparison of rice, maize, and wheat: Implications for chloroplast gene diversification and phylogeny of cereals. *Mol. Biol. Evol.* 19, 2084–2091.
- McCoy, S.R., Kuehl, J. V, Boore, J.L., Raubeson, L.A., 2008. The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. *BMC Evol. Biol.* 8, 130. <https://doi.org/10.1186/1471-2148-8-130>
- McDonald, M.J., Wang, W.C., Huang, H. Da, Leu, J.Y., 2011. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol.* 9. <https://doi.org/10.1371/journal.pbio.1000622>
- McLenachan, P.A., Stöckler, K., Winkworth, R.C., McBreen, K., Zauner, S., Lockhart, P.J., 2000. Markers derived from amplified fragment length polymorphism gels for plant ecology and evolution studies. *Mol. Ecol.* 9, 1899–903.
- Menezes, A.P.A., Resende-Moreira, L.C., Buzatti, R.S.O., Nazareno, A.G., Carlsen, M., Lobo, F.P., Kalapothakis, E., Lovato, M.B., 2018. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. *Sci. Rep.* 8, 1–12. <https://doi.org/10.1038/s41598-018-20189-4>
- Mes, T.H., Kuperus, P., Kirschner, J., Stepanek, J., Oosterveld, P., Storchova, H., den Nijs, J.C., 2000. Hairpins involving both inverted and direct repeats are associated with homoplasious indels in non-coding chloroplast DNA of *Taraxacum* (Lactuceae: Asteraceae). *Genome* 43, 634–641. <https://doi.org/10.1139/g99-135>
- Millen, R.S., Olmstead, R.G., Adams, K.L., Palmer, J.D., Lao, N.T., Heggie, L., Kavanagh, T.A., Hibberd, J.M., Gray, J.C., Morden, C.W., Calie, P.J., Jermini, L.S., Wolfe, K.H., 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* 13, 645–58.

References

- Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., Marshall, D., 2009. Tablet-next generation sequence assembly visualization. *Bioinformatics* 26, 401–402. <https://doi.org/10.1093/bioinformatics/btp666>
- Mondal, S., Ghosh, D., Sagar, N., Ganapaty, S., 2016. Evaluation of antioxidant, toxicological and wound healing properties of *Hibiscus rosa-sinensis* L. (Malvaceae) ethanolic leaves extract on different experimental animal models. *Indian J. Pharm. Educ. Res.* 50, 620–637. <https://doi.org/10.5530/ijper.50.4.15>
- Motamayor, J.C., Lachenaud, P., da Silva e Mota, J.W., Loor, R., Kuhn, D.N., Brown, J.S., Schnell, R.J., 2008. Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao* L.). *PLoS One* 3, e3311. <https://doi.org/10.1371/journal.pone.0003311>
- Motamayor, J.C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., Cornejo, O., Findley, S.D., Zheng, P., Utro, F., Royaert, S., Saski, C., Jenkins, J., Podicheti, R., Zhao, M., Scheffler, B.E., Stack, J.C., Feltus, F.A., Mustiga, G.M., Amores, F., Phillips, W., Marelli, J.P., May, G.D., Shapiro, H., Ma, J., Bustamante, C.D., Schnell, R.J., Main, D., Gilbert, D., Parida, L., Kuhn, D.N., 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* 14. <https://doi.org/10.1186/gb-2013-14-6-r53>
- Motamayor, J.C., Risterucci, A.M., Lopez, P.A., Ortiz, C.F., Moreno, A., Lanaud, C., 2002. Cacao domestication : The origin of the cacao cultivated by the Mayas. *Heredity (Edinb.)* 89, 380–386. <https://doi.org/10.1038/sj.hdy.6800156>
- Mower, J.P., 2009. The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 37, W253–W259. <https://doi.org/10.1093/nar/gkp337>
- Mustafina, F.U., Yi, D.-K., Choi, K., Shin, C.H., Tojibaev, K.S., Downie, S.R., 2019. A comparative analysis of complete plastid genomes from *Prangos fedtschenkoi* and *Prangos lipskyi* (Apiaceae). *Ecol. Evol.* 9, 364–377. <https://doi.org/10.1002/ece3.4753>
- Nanadagopalan, V., Gritto, M.J., Doss, A., 2015. GC-MS analysis of biomolecules on the leaves extract of *Sterculia urens* Roxb. *J. Pharmacogn. Phytochem.* 3, 193–196.
- Nazareno, A.G., Carlsen, M., Lohmann, L.G., 2015. Complete chloroplast genome of *Tanaecium tetragonolobum* : The first Bignoniaceae plastome. *PLoS One* 10, e0129930. <https://doi.org/10.1371/journal.pone.0129930>

References

- Neale, D.B., Sederoff, R.R., 1989. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in *Loblolly pine*. *Theor. Appl. Genet.* 77, 212–216. <https://doi.org/10.1007/BF00266189>
- Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating Maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. <https://doi.org/10.1093/molbev/msu300>
- Nguyen, V.B., Linh Giang, V.N., Waminal, N.E., Park, H.S., Kim, N.H., Jang, W., Lee, J., Yang, T.J., 2018. Comprehensive comparative analysis of chloroplast genomes from seven *Panax* species and development of an authentication system based on species-unique single nucleotide polymorphism markers. *J. Ginseng Res.* <https://doi.org/10.1016/j.jgr.2018.06.003>
- Nguyen, V.B., Park, H.-S., Lee, S.-C., Lee, J., Park, J.Y., Yang, T.-J., 2017. Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. *J. Agric. Food Chem.* 65, 6298–6306. <https://doi.org/10.1021/acs.jafc.7b00925>
- Nock, C.J., Waters, D.L.E., Edwards, M.A., Bowen, S.G., Rice, N., Cordeiro, G.M., Henry, R.J., 2011. Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* 9, 328–333. <https://doi.org/10.1111/j.1467-7652.2010.00558.x>
- Nyffeler, R., Bayer, C., Alverson, W.S., Yen, A., Whitlock, B.A., Chase, M.W., Baum, D.A., 2005. Phylogenetic analysis of the Malvaceae clade (Malvaceae s.l.) based on plastid DNA sequences. *Org. Divers. Evol.* 5, 109–123. <https://doi.org/10.1016/j.ode.2004.08.001>
- Oldenburg, D.J., Bendich, A.J., 2016. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. *Curr. Genet.* 62, 431–442. <https://doi.org/10.1007/s00294-015-0548-0>
- Osorio-Guarín, J.A., Berdugo-Cely, J., Coronado, R.A., Zapata, Y.P., Quintero, C., Gallego-Sánchez, G., Yockteng, R., 2017. Colombia a source of Cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. *Front. Plant Sci.* 8, 1994. <https://doi.org/10.3389/fpls.2017.01994>
- Palmer, J.D., 1985. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* 19, 325–354. <https://doi.org/10.1146/annurev.ge.19.120185.001545>
- Palmer, J.D., Osorio, B., Thompson, W.F., 1988. Evolutionary significance of inversions in

References

- Legume chloroplast DNAs. *Curr. Genet.* 14, 65–74. <https://doi.org/10.1007/BF00405856>
- Pan, I.-C., Liao, D.-C., Wu, F.-H., Daniell, H., Singh, N.D., Chang, C., Shih, M.-C., Chan, M.-T., Lin, C.-S., 2012. Complete chloroplast genome sequence of an Orchid model plant candidate: *Erycina pusilla* apply in tropical *Oncidium* breeding. *PLoS One* 7, e34738. <https://doi.org/10.1371/journal.pone.0034738>
- Park, Seongjun, Jansen, R.K., Park, SeonJoo, 2015. Complete plastome sequence of *Thalictrum coreanum* (Ranunculaceae) and transfer of the *rpl32* gene to the nucleus in the ancestor of the subfamily Thalictroideae. *BMC Plant Biol.* 15, 40. <https://doi.org/10.1186/s12870-015-0432-6>
- Patel, R., Patel, A., Desai, S., Nagee, A., 2012. Study of secondary metabolites and antioxidant properties of leaves , stem and root among *Hibiscus rosa-sinensis* cultivars. *Asian J. Exp. Biol. Sci.* 3, 719–725.
- Patil, P., Sutar, S., K, J.J., Malik, S., Rao, S., Bhat, K. V, 2015. A systematic review of the genus *Abelmoschus* (Malvaceae). *Rheedea* 25, 14–30.
- Peltier, G., Cournac, L., 2002. Chlororespiration. *Annu. Rev. Plant Biol.* 53, 523–550. <https://doi.org/10.1146/annurev.arplant.53.100301.135242>
- Peng, L., Yamamoto, H., Shikanai, T., 2011. Structure and biogenesis of the chloroplast NAD (P) H dehydrogenase complex. *Biochim. Biophys. Acta - Bioenerg.* 1807, 945–953.
- Perez, G.R.M., 2012. Evaluation of hypoglycemic activity of the leaves of *Malva parviflora* in streptozotocin-induced diabetic rats. *Food Funct.* 3, 420. <https://doi.org/10.1039/c2fo10153j>
- Perveen, A., Grafström, E., El-Ghazaly†, G., 2004. World pollen and spore flora 23. *Malvaceae adams*. P.p. Subfamilies: Grewioideae, Tilioideae, Brownlowioideae. *Grana* 43, 129–155. <https://doi.org/10.1080/00173130410000730>
- Pfeil, B.E., Brubaker, C.L., Craven, L. a, Crisp, M.D., 2002. Phylogeny of *Hibiscus* and the tribe Hibisceae (Malvaceae) using chloroplast DNA sequences of *ndhF* and the *rpl16* intron. *Syst. Bot.* 27, 333–350. <https://doi.org/10.1043/0363-6445-27.2.333>
- Pfeil, B.E., Crisp, M., 2005. What to do with *Hibiscus*? A proposed nomenclatural resolution for a large and well known genus of Malvaceae and comments on paraphyly *Aust. Syst. Bot.* 18, 49–60. <https://doi.org/10.1071/SB04024>
- Poczai, P., Hyvönen, J., 2017. The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides* , Bromeliaceae) and its comparative

References

- analysis. PLoS One 12, 1–25. <https://doi.org/10.1371/journal.pone.0187199>
- Poovitha, S., Stalin, N., Balaji, R., Parani, M., 2016. Multi-locus DNA barcoding identifies *matK* as suitable marker for species identification in *Hibiscus* L. Genome 59, 1150–1156. <https://doi.org/10.2134/agronj2015.0612>
- Prasad, M.P., 2014. In vitro phytochemical analysis and antioxidant studies of *Hibiscus* species. Int. J. Pure Appl. Biosci. 2, 83–88.
- Prudente, A.S., Loddi, A.M.V., Duarte, M.R., Santos, A.R.S., Pochapski, M.T., Pizzolatti, M.G., Hayashi, S.S., Campos, F.R., Pontarolo, R., Santos, F.A., Cabrini, D.A., Otuki, M.F., 2013. Pre-clinical anti-inflammatory aspects of a cuisine and medicinal millennial herb: *Malva sylvestris* L. Food Chem. Toxicol. 58, 324–331. <https://doi.org/10.1016/J.FCT.2013.04.042>
- Purseglove, J.W., 1968. Tropical crops: Dicotyledons 1 and 2. Trop. Crop. Dicotyledons 1 2.
- Qian, J., Song, J., Gao, H., Zhu, Y., Xu, J., Pang, X., Yao, H., Sun, C., Li, X., Li, C., Liu, J., Xu, H., Chen, S., 2013. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. PLoS One 8. <https://doi.org/10.1371/journal.pone.0057607>
- Qiu, Y., Liu, Y., Kang, M., Yi, G., Huang, H., 2013. Spatial and temporal population genetic variation and structure of *Nothotsuga longibracteata* (Pinaceae), a relic conifer species endemic to subtropical China. Genet. Mol. Biol. 36, 598–607. <https://doi.org/10.1590/S1415-47572013000400019>
- Rabah, S.O., Lee, C., Hajrah, N.H., Makki, R.M., Alharby, H.F., Alhebshi, A.M., Sabir, J.S.M., Jansen, R.K., Ruhlman, T.A., 2017. Plastome sequencing of ten Non model crop Species uncovers a large insertion of mitochondrial DNA in *Cashew*. Plant Genome 10, 0. <https://doi.org/10.3835/plantgenome2017.03.0020>
- Raman, G., Park, V., Kwak, M., Lee, B., Park, S.J., 2017. Characterization of the complete chloroplast genome of *Arabis stellari* and comparisons with related species. PLoS One 12, 1–18. <https://doi.org/10.1371/journal.pone.0183197>
- Razavi, S.M., Zarrini, G., Molavi, G., Ghasemi, G., 2011. Bioactivity of *Malva sylvestris* L., a medicinal plant from iran. Iran. J. Basic Med. Sci. 14, 574–9.
- Redwan, R.M., Saidin, A., Kumar, S. V., 2015. Complete chloroplast genome sequence of MD-2 Pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. BMC Plant Biol. 15, 1–20. <https://doi.org/10.1186/s12870-015-0587-1>
- Riaz, G., Chopra, R., 2018. A review on phytochemistry and therapeutic uses of *Hibiscus*

References

- sabdariffa* L. Biomed. Pharmacother. 102, 575–586.
<https://doi.org/10.1016/J.BIOPHA.2018.03.023>
- Richardson, J.E., Whitlock, B.A., Meerow, A.W., Madriñán, S., 2015. The age of chocolate: a diversification history of *Theobroma* and Malvaceae. *Front. Ecol. Evol.* 3, 1–14.
<https://doi.org/10.3389/fevo.2015.00120>
- Rizk, R.M., Soliman, M.I., 2014. Biochemical and molecular genetic characterization of some species of family Malvaceae, Egypt. *Egypt. J. Basic Appl. Sci.* 1, 167–176.
<https://doi.org/10.1016/j.ejbas.2014.06.002>
- Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sanchez-Gracia, A., 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302.
<https://doi.org/10.1093/molbev/msx248>
- Ruhlman, T.A., Chang, W.-J., Chen, J.J., Huang, Y.-T., Chan, M.-T., Zhang, J., Liao, D.-C., Blazier, J.C., Jin, X., Shih, M.-C., Jansen, R.K., Lin, C.-S., 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biol.* 15, 100. <https://doi.org/10.1186/s12870-015-0484-7>
- Sabir, J., Schwarz, E., Ellison, N., Zhang, J., Baeshen, N.A., Mutwakil, M., Jansen, R., Ruhlman, T., 2014. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnol. J.* 12, 743–754. <https://doi.org/10.1111/pbi.12179>
- Saina, J.K., Gichira, A.W., Li, Z.Z., Hu, G.W., Wang, Q.F., Liao, K., 2018a. The complete chloroplast genome sequence of *Dodonaea viscosa*: comparative and phylogenetic analyses. *Genetica* 146, 101–113. <https://doi.org/10.1007/s10709-017-0003-x>
- Saina, J.K., Li, Z.Z., Gichira, A.W., Liao, Y.Y., 2018b. The complete chloroplast genome sequence of tree of heaven (*Ailanthus altissima* (mill.) (sapindales: Simaroubaceae), an important pantropical tree. *Int. J. Mol. Sci.* 19. <https://doi.org/10.3390/ijms19040929>
- Sanderson, M.J., Copetti, D., Burquez, A., Bustamante, E., Charboneau, J.L.M., Eguiarte, L.E., Kumar, S., Lee, H.O., Lee, J., McMahon, M., Steele, K., Wing, R., Yang, T.-J., Zwickl, D., Wojciechowski, M.F., 2015. Exceptional reduction of the plastid genome of *Saguaro cactus* (*Carnegiea gigantea*): Loss of the *ndh* gene suite and inverted repeat. *Am. J. Bot.* 102, 1115–1127. <https://doi.org/10.3732/ajb.1500184>
- Sanghai, D.B., Kumar, S.V., Srinivasan, K.K., Aswatharam, H.N., Shreedhara, C.S., 2013.

References

- Pharmacognostic and phytochemical investigation of the leaves of *Malvastrum coromandelianum* (L.) Garcke. *Anc. Sci. Life* 33, 39–44. <https://doi.org/10.4103/0257-7941.134596>
- Saski, C., Lee, S.B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H. G., Jansen, R. K., 2005. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other Legume genomes. *Springer* 59, 309–322. <https://doi.org/10.1007/s11103-005-8882-0>
- Schwarz, E.N., Ruhlman, T.A., Sabir, J.S.M., Hajrah, N.H., Alharbi, N.S., Al-Malki, A.L., Bailey, C.D., Jansen, R.K., 2015. Plastid genome sequences of Legumes reveal parallel inversions and multiple losses of *rps16* in Papilionoids. *J. Syst. Evol.* 53, 458–468. <https://doi.org/10.1111/jse.12179>
- Shahinnia, F., Sayed-Tabatabaei, B.E., 2009. Conversion of barley SNPs into PCR-based markers using dCAPS method. *Genet. Mol. Biol.* 32, 564–567. <https://doi.org/10.1590/S1415-47572009005000047>
- Shale, T.L., Stirk, W.A., van Staden, J., 2005. Variation in antibacterial and anti-inflammatory activity of different growth forms of *Malva parviflora* and evidence for synergism of the anti-inflammatory compounds. *J. Ethnopharmacol.* 96, 325–330. <https://doi.org/10.1016/J.JEP.2004.09.032>
- Shen, H.M., Chen, C., Jiang, J.Y., Zheng, Y.L., Cai, W.F., Wang, B., Ling, Z., Tang, L., Wang, Y.H., Shi, G.G., 2017. The N-butyl alcohol extract from *Hibiscus rosa-sinensis* L. flowers enhances healing potential on rat excisional wounds. *J. Ethnopharmacol.* 198, 291–301. <https://doi.org/10.1016/j.jep.2017.01.016>
- Sherman-Broyles, S., Bombarely, A., Grimwood, J., Schmutz, J., Doyle, J., 2014. Complete plastome sequences from *Glycine syndetika* and Six additional perennial wild relatives of Soybean. *G3 Genes, Genomes, Genet.* 4, 2023–2033. <https://doi.org/10.1534/G3.114.012690>
- Shikanai, T., Endo, T., Hashimoto, T., Yamada, Y., Asada, K., Yokota, A., 1998. Directed disruption of the tobacco *ndhB* gene impairs cyclic electron flow around photosystem I. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9705–9. <https://doi.org/10.1073/pnas.95.16.9705>
- Silva, J.C., Kondrashov, A.S., 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.* 18, 544–7. [https://doi.org/10.1016/S0168-9525\(02\)02757-9](https://doi.org/10.1016/S0168-9525(02)02757-9)

References

- Sinn, B.T., Sedmak, D.D., Kelly, L.M., Freudenstein, J. V., 2018. Total duplication of the small single copy region in the angiosperm plastome: Rearrangement and inverted repeat instability in *Asarum*. *Am. J. Bot.* 105, 71–84. <https://doi.org/10.1002/ajb2.1001>
- Sittiwet, C., Jesadanont, S., Pongpech, P., Naenna, P., Pongsamart, S., 2008. Antibacterial activity of *Malvastrum coromandelianum* Garcke against methicillin-sensitive and methicillin-resistant strains of *Staphylococcus aureus*. *Curr. Res. Bacteriol.* 1, 42–45. <https://doi.org/10.3923/crb.2008.42.45>
- Sivarajan, V., Pradeep, A., 1996. Malvaceae of southern Peninsular India a taxonomic monograph. Daya Publ. house Delhi.
- Small, R.L., 2004. Phylogeny of *Hibiscus* sect. *Muenchhusia* (Malvaceae) based on chloroplast *rpL16* and *ndhF*, and nuclear ITS and GBSSI sequences. *Syst. Bot.* 29, 385–392. <https://doi.org/10.1600/036364404774195575>
- Smith, D.R., 2014. Mitochondrion-to-plastid DNA transfer: it happens. *New Phytol.* 202, 736–738. <https://doi.org/10.1111/nph.12704>
- Song, Y., Dong, W., Liu, B., Xu, C., Yao, X., Gao, J., Corlett, R.T., 2015. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Front. Plant Sci.* 6, 662. <https://doi.org/10.3389/fpls.2015.00662>
- Taia, W.K., 2009. General View of Malvaceae Juss . S . L . and Taxonomic revision of genus *Abutilon* Mill . in Saudi Arabia 21, 349–363.
- Takabayashi, A., Endo, T., Shikanai, T., Sato, F., 2002. Post-illumination reduction of the plastoquinone pool in chloroplast transformants in which chloroplastic NAD(P)H dehydrogenase was inactivated. *Biosci. Biotechnol. Biochem.* 66, 2107–2111. <https://doi.org/10.1271/bbb.66.2107>
- Takhtadzhian, A., Takhtajan, A., 1997. Diversity and classification of flowering plants. Columbia University Press.
- Tangphatsornruang, S., Sangsrakru, D., Chanprasert, J., Uthaipaisanwong, P., Yoocha, T., Jomchai, N., Tragoonrung, S., 2010. The chloroplast genome sequence of Mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. *DNA Res.* 17, 11–22. <https://doi.org/10.1093/dnares/dsp025>
- Tate, J.A., Aguilar, J.F., Wagstaff, S.J., La Duke, J.C., Bodo Slotta, T.A., Simpson, B.B., 2005.

References

- Phylogenetic relationships within the tribe Malveae (Malvaceae, subfamily Malvoideae) as inferred from ITS sequence data. *Am. J. Bot.* 92, 584–602. <https://doi.org/10.3732/ajb.92.4.584>
- Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. life Sci.* 17, 57–86.
- 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.). *Theor. Appl. Genet.* 106, 411–422. <https://doi.org/10.1007/s00122-002-1031-0>
- Thomas, E., van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., van Etten, J., 2012. Present spatial diversity patterns of *Theobroma cacao* L. in the neotropics reflect genetic differentiation in pleistocene refugia followed by human-influenced dispersal. *PLoS One* 7, e47676. <https://doi.org/10.1371/journal.pone.0047676>
- Thorne, R.F., 2000. The classification and geography of the flowering plants: Dicotyledons of the class angiospermae. *Bot. Rev.* 66, 441–647. <https://doi.org/10.1007/BF02869011>
- Thorne, R.F., 1992. Classification and geography of the flowering plants. *Bot. Rev.* 58, 225–327. <https://doi.org/10.1007/BF02858611>
- Thulin, M., 1999. *Flora of Somalia*. Royal Botanic Gardens, Kew.
- Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., Chen, J.-Q., 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455, 105–108. <https://doi.org/10.1038/nature07175>
- Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., Greiner, S., 2017. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. <https://doi.org/10.1093/nar/gkx391>
- Timmis, J., Ayliffe, M., Huang, C., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Genetics* 5, 123.
- TR. R., Murali, S., Ms, F., 2016. DNA Barcoding of the selected *Artemisia spp.* using the five universal barcodes. *Int. J. Herb. Med.* 4, 38–42.
- Turmel, M., Otis, C., Lemieux, C., 2015. Dynamic evolution of the chloroplast genome in the green algal classes Pedinophyceae and Trebouxiophyceae. *Genome Biol. Evol.* 7, 2062–2082. <https://doi.org/10.1093/gbe/evv130>

References

- Ueda, M., Fujimoto, M., Arimura, S., Murata, J., Tsutsumi, N., Kadowaki, K., 2007. Loss of the *rpl32* gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus*. *Gene* 402, 51–56. <https://doi.org/10.1016/J.GENE.2007.07.019>
- Vasudeva, N., Sharma, S.K., 2008. Biologically active compounds from the genus *Hibiscus*. *Pharm. Biol.* 46, 145–153. <https://doi.org/10.1080/13880200701575320>
- 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. *Genet. Mol. Biol.* 39, 312–328. <https://doi.org/10.1590/1678-4685-GMB-2016-0027>
- Vinutha, K.B., Devi, A.A., Sreekumar, J., 2015. Morphological characterization of above ground characters of Taro (*Colocasia esculenta* (L.) accessions from North East India 41, 3–11.
- Vital, P.G., Velasco, R.N., Demigillo, J.M., Rivera, W.L., 2010. Antimicrobial activity, cytotoxicity and phytochemical screening of *Ficus septica* Burm and *Sterculia foetida* L. leaf extracts. *J. Med. Plants Res.* 4, 58–63. <https://doi.org/10.5897/JMPR09.400>
- Voon, Y.Y., Abdul Hamid, N.S., Rusul, G., Osman, A., Quek, S.Y., 2007. Characterisation of Malaysian durian (*Durio zibethinus* Murr.) cultivars: Relationship of physicochemical and flavour properties with sensory properties. *Food Chem.* 103, 1217–1227. <https://doi.org/10.1016/J.FOODCHEM.2006.10.038>
- Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., Sugiura, M., 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc. Natl. Acad. Sci. U. S. A.* 91, 9794–8. <https://doi.org/10.1073/pnas.91.21.9794>
- Wambugu, P.W., Brozynska, M., Furtado, A., Waters, D.L., Henry, R.J., 2015. Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. *Sci. Rep.* 5, 13957–13966. <https://doi.org/10.1038/srep13957>
- Wang, C., Xu, J.-R., Liu, H., 2016. A-to-I RNA editing independent of ADARs in filamentous fungi. *RNA Biol.* 13, 940–945. <https://doi.org/10.1080/15476286.2016.1215796>
- Wang, J.-H., Cai, Y.-C., Zhao, K.-K., Zhu, Z.-X., Zhou, R.-C., Wang, H.-F., 2017. Characterization of the complete chloroplast genome sequence of *Firmiana pulcherrima* (Malvaceae). *Conserv. Genet. Resour.* <https://doi.org/10.1007/s12686-017-0880-4>
- Wang, W., Zhang, W., Wu, Y., Maliga, P., Messing, J., 2015. RNA editing in chloroplasts of

References

- Spirodela polyrhiza*, an aquatic monocotyledonous species. PLoS One 10, e0140285. <https://doi.org/10.1371/journal.pone.0140285>
- Watson, L., 1992. The families of flowering plants: descriptions, illustrations, identification and information retrieval. <http://biodiversity.uno.edu/delta.htm>.
- Weng, M.-L., Blazier, J.C., Govindu, M., Jansen, R.K., 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol. Biol. Evol. 31, 645–659. <https://doi.org/10.1093/molbev/mst257>
- Whitlock, B.A., Hale, A.M., Groff, P.A., 2010. Intraspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. PLoS One 5, e11533. <https://doi.org/10.1371/journal.pone.0011533>
- Wilkie, P., Clark, A., Pennington, R.T., Cheek, M., Wilkie, P., Clark, A., Pennington, R.T., Cheek, M., Bayer, C., Wilcock, C.C., Edinburgh, G., Row, I., Eh, E., 2006. Phylogenetic relationships within the subfamily Sterculioideae (Malvaceae/ Sterculiaceae- Sterculieae) using the chloroplast gene *ndhF*. Syst. Bot. 31, 160–170.
- Wilson, F.D., 1999. Revision of *Hibiscus* section Furcaria (Malvaceae) in Africa and Asia. Bull. of the Nat. Hist. Museum, London 29, 47–79.
- Wolfe, K.H., Li, W.H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84, 9054–9058.
- Woo, K.W., Choi, S.U., Kim, K.H., Lee, K.R., Woo, K.W., Choi, S.U., Kim, K.H., Lee, K.R., 2015. Ursane saponins from the stems of *Firmiana simplex* and their cytotoxic activity. J. Braz. Chem. Soc. 26, 1450–1456. <https://doi.org/10.5935/0103-5053.20150113>
- Wu, C.-S., Wang, Y.-N., Hsu, C.-Y., Lin, C.-P., Chaw, S.-M., 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol. Evol. 3, 1284–1295. <https://doi.org/10.1093/gbe/evr095>
- Wu, F.-H., Chan, M.-T., Liao, D.-C., Hsu, C.-T., Lee, Y.-W., Daniell, H., Duvall, M.R., Lin, C.-S., 2010. Complete chloroplast genome of *Oncidium gower* Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol. 10, 68. <https://doi.org/10.1186/1471-2229-10-68>
- Wu, Y., Liu, F., Yang, D.-G., Li, W., Zhou, X.-J., Pei, X.-Y., Liu, Y.-G., He, K.-L., Zhang,

References

- W.-S., Ren, Z.-Y., Zhou, K.-H., Ma, X.-F., Li, Z.-H., 2018. Comparative chloroplast genomics of *Gossypium* species: Insights into repeat sequence variations and phylogeny. *Front. Plant Sci.* 9, 1–14. <https://doi.org/10.3389/fpls.2018.00376>
- Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. <https://doi.org/10.1093/bioinformatics/bth352>
- Wysocki, W.P., Clark, L.G., Attigala, L., Ruiz-Sanchez, E., Duvall, M.R., 2015. Evolution of the *Bamboos* (Bambusoideae; Poaceae): a full plastome phylogenomic analysis. *BMC Evol. Biol.* 15, 50. <https://doi.org/10.1186/s12862-015-0321-5>
- Xie, J., Shi, L., Zhu, X., Wang, P., Zhao, Y., Su, W., 2011. Mechanochemical-assisted efficient extraction of rutin from *Hibiscus mutabilis* L. *Innov. Food Sci. Emerg. Technol.* 12, 146–152. <https://doi.org/10.1016/J.IFSET.2010.12.009>
- Xin, G.-L., Ren, X.-L., Liu, W.-Z., Jia, G.-L., Deng, C.-Y., 2018. The complete chloroplast genome of a rare species *Heritiera parvifolia* Merr. (Malvales: Sterculiaceae). *Conserv. Genet. Resour.* 10, 885–888. <https://doi.org/10.1007/s12686-017-0888-9>
- Xu, J.-H., Liu, Q., Hu, W., Wang, T., Xue, Q., Messing, J., 2015. Dynamics of chloroplast genomes in green plants. *Genomics* 106, 221–231. <https://doi.org/10.1016/J.YGENO.2015.07.004>
- Xu, Q., Xiong, G., Li, P., He, F., Huang, Y., Wang, K., Li, Z., Hua, J., 2012. Analysis of complete nucleotide sequences of 12 *Gossypium* chloroplast genomes: Origin and evolution of Allotetraploids. *PLoS One* 7. <https://doi.org/10.1371/journal.pone.0037128>
- Xu, Z., Deng, M., 2017. Malvaceae, In: Identification and control of common weeds. Springer Netherlands, pp. 717–735.
- Xue, S., Shi, T., Luo, W., Ni, X., Iqbal, S., Ni, Z., Huang, X., Yao, D., Shen, Z., Gao, Z., 2019. Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. *Hortic. Res.* 6, 89. <https://doi.org/10.1038/s41438-019-0171-1>
- Ya, J., Yu, Z., Yang, Y.-Q., Zhang, S., Zhang, Z., Cai, J., Yang, J., Yu, W., 2017. Correction to: Complete chloroplast genome of *Firmiana major* (Malvaceae), a critically endangered species endemic to southwest China. *Conserv. Genet. Resour.* <https://doi.org/10.1007/s12686-017-0915-x>
- Yamane, K., Yasui, Y., Ohnishi, O., 2003. Intraspecific cpDNA variations of diploid and

- tetraploid perennial buckwheat, *Fagopyrum cymosum* (Polygonaceae). *Am. J. Bot.* 90, 339–346. <https://doi.org/10.3732/ajb.90.3.339>
- Yang, H., Wu, Y., Feng, J., Yang, S., Tian, D., 2009. Evolutionary pattern of protein architecture in mammal and fruit fly genomes. *Genomics* 93, 90–97. <https://doi.org/10.1016/j.ygeno.2008.09.009>
- Yang, J.-B., Tang, M., Li, H.-T., Zhang, Z.-R., Li, D.-Z., 2013. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol. Biol.* 13, 84. <https://doi.org/10.1186/1471-2148-13-84>
- Yang, J., Feng, L., Yue, M., He, Y.L., Zhao, G.F., Li, Z.H., 2019. Species delimitation and interspecific relationships of the endangered herb genus *Notopterygium* inferred from multilocus variations. *Mol. Phylogenet. Evol.* 133, 142–151. <https://doi.org/10.1016/j.ympev.2019.01.002>
- Yang, J.Y., Motilal, L.A., Dempewolf, H., Maharaj, K., Cronk, Q.C.B., 2011. Chloroplast microsatellite primers for Cacao (*Theobroma cacao*) and other Malvaceae. *Am. J. Bot.* 98, e372–e374. <https://doi.org/10.3732/ajb.1100306>
- Yang, X., Luo, X., Cai, X., 2014. Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasit. Vectors* 7, 527:537. <https://doi.org/10.1186/s13071-014-0527-1>
- Yang, Y., Zhou, T., Duan, D., Yang, J., Feng, L., Zhao, G., 2016. Comparative analysis of the complete chloroplast genomes of five *Quercus* species. *Front. Plant Sci.* 7, 959:972. <https://doi.org/10.3389/fpls.2016.00959>
- Yao, L., Yang, L., Chen, Z., 1994. Studies on chemical constituents of *Hibiscus mutabilis*. *Chinese Tradit. Herb. Drugs* 03.
- Yi, X., Gao, L., Wang, B., Su, Y.-J., Wang, T., 2013. The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): Evolutionary comparison of *Cephalotaxus* chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. *Genome Biol. Evol.* 5, 688–698. <https://doi.org/10.1093/gbe/evt042>
- Yu, X., Zuo, L., Lu, D., Lu, B., Yang, M., Wang, J., 2019. Comparative analysis of chloroplast genomes of five *Robinia* species: Genome comparative and evolution analysis. *Gene* 689, 141–151. <https://doi.org/10.1016/J.GENE.2018.12.023>
- Yu, X.Q., Drew, B.T., Yang, J.B., Gao, L.M., Li, D.Z., 2017. Comparative chloroplast

References

- genomes of eleven *Schima* (Theaceae) species: Insights into DNA barcoding and phylogeny. *PLoS One* 12, 1–18. <https://doi.org/10.1371/journal.pone.0178026>
- Zerbino, D.R., Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. <https://doi.org/10.1101/gr.074492.107>
- Zhai, W., Duan, X., Zhang, R., Guo, C., Li, L., Xu, G., Shan, H., Kong, H., Ren, Y., 2019. Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the Ranunculaceae. *Mol. Phylogenet. Evol.* 135, 12–21. <https://doi.org/10.1016/j.ympev.2019.02.024>
- Zhang, D., Figueira, A., Motilal, L., Lachenaud, P., Meinhardt, L.W., 2011. *Theobroma*, In: wild crop relatives: genomic and breeding resources. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 277–296. https://doi.org/10.1007/978-3-642-21201-7_13
- Zhang, H., Li, C., Miao, H., Xiong, S., 2013. Insights from the complete chloroplast genome into the evolution of *Sesamum indicum* L. *PLoS One* 8, e80508. <https://doi.org/10.1371/journal.pone.0080508>
- Zhang, J., Xu, Y., Chen, W., Dell, B., Vergauwen, R., Biddulph, B., Khan, N., Luo, H., Appels, R., Van den Ende, W., 2015. A wheat 1-FEH w3 variant underlies enzyme activity for stem WSC remobilization to grain under drought. *New Phytol.* 205, 293–305. <https://doi.org/10.1111/nph.13030>
- Zhang, W., Sun, X., Yuan, H., Araki, H., Wang, J., Tian, D., 2008. The pattern of insertion/deletion polymorphism in *Arabidopsis thaliana*. *Mol. Genet. Genomics* 280, 351–361. <https://doi.org/10.1007/s00438-008-0370-1>
- Zhang, Y., Du, L., Liu, A., Chen, J., Wu, L., Hu, W., Zhang, W., Kim, K., Lee, S.-C., Yang, T.-J., Wang, Y., 2016. The complete chloroplast genome sequences of five *Epimedium* species: Lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7, 1–12. <https://doi.org/10.3389/fpls.2016.00306>
- Zhao, K.-K., Wang, J.-H., Cai, Y.-C., Zhu, Z.-X., López-Pujol, J., Wang, H.-F., 2018. Complete chloroplast genome sequence of *Heritiera angustata* (Malvaceae): an endangered plant species. *Mitochondrial DNA Part B Resour.* 3, 141–142. <https://doi.org/10.1080/23802359.2017.1422398>
- Zhong, B., Deusch, O., Goremykin, V. V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S. V., Lockhart, P.J., 2011. Systematic error in seed plant phylogenomics. *genome Biol. Evol.* 3, 1340–1348. <https://doi.org/10.1093/gbe/evr105>

References

- Zhu, A., Guo, W., Gupta, S., Fan, W., Mower, J.P., 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. <https://doi.org/10.1111/nph.13743>
- Zhu, L., Wang, Q., Tang, P., Araki, H., Tian, D., 2009. Genome wide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol. Biol. Evol.* 26, 2353–2361. <https://doi.org/10.1093/molbev/msp144>

List of publications

The research work presented in the current dissertation resulted in the publication of following articles.

Abdullah, Mehmood, F., Shahzadi, I., Waseem, S., Mirza, B., Ahmed, I., Waheed, M.T., 2020. Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics* 112: 581-591. <https://doi.org/10.1016/j.ygeno.2019.04.010>

Abdullah, Shahzadi, I., Mehmood, F., Ali, Z., Malik, M.S., Waseem, S., Mirza, B., Ahmed, I., Waheed, M.T., 2019. Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. *Plant Gene* 19, 100199. <https://doi.org/10.1016/J.PLGENE.2019.100199>

Abdullah, Waseem, S., Mirza, B., Ahmed, I., Waheed, M.T., 2020. Comparative analyses of chloroplast genome of *Theobroma cacao* and *Theobroma grandiflorum*. *Biologia* 75: 761-771. <https://doi.org/10.2478/s11756-019-00388-8>

Abdullah, Mehmood, F., Shahzadi, I., Ali, Z., Islam, M., Naeem, M., Mirza, B., Lockhart, P., Ahmed, I., Waheed, M.T., 2020. Correlations among oligonucleotide repeats, nucleotide substitutions and insertion – deletion mutations in chloroplast genomes of plant family Malvaceae. *Journal of Systematics and Evolution*. <https://doi.org/10.1111/jse.12585>

Turnitin Originality Report

Processed on: 24-Jul-2020 13:33 PKT
 ID: 1361532520
 Word Count: 55693
 Submitted: 1

Similarity Index

15%

Similarity by Source

Internet Sources: 8%
 Publications: 8%
 Student Papers: 5%

Evolutionary dynamics and phylogeny of family Malvaceae By Abdullah .

1% match (publications)

[Abdullah, Claudia L. Henriquez, Furrugh Mehmood, Monica M. Carlsen et al. "Complete chloroplast genomes of and \(Pothoideae, Araceae\): unique inverted repeat expansion and contraction affect rate of evolution", Cold Spring Harbor Laboratory, 2020](#)

1% match (Internet from 02-Jul-2018)

<http://mdpi.com/1422-0067/19/4/929/htm>

1% match (Internet from 08-Apr-2020)

<https://www.mdpi.com/1999-4907/10/11/1000/html>

1% match (publications)

[Matias Köhler, Marcelo Reginato, Tatiana T. Souza-Chies, Lucas C. Majure. "Insights into chloroplast genome variation across Opuntioideae \(Cactaceae\)", Cold Spring Harbor Laboratory, 2020](#)

1% match (publications)

[Furrugh Mehmood, Abdullah, Zartasha Ubaid, Iram Shahzadi, Ibrar Ahmed, Mohammad Tahir Waheed, Péter Poczai, Bushra Mirza. "Plastid genomics of \(Solanaceae\): insights into molecular evolution, positive selection and the origin of the maternal genome of Aztec tobacco \(.\)", Cold Spring Harbor Laboratory, 2020](#)

< 1% match (student papers from 02-Jul-2020)

[Submitted to University of Southampton on 2020-07-02](#)

< 1% match (publications)

[Claudia L. Henriquez, Abdullah, Ibrar Ahmed, Monica M. Carlsen, Alejandro Zuluaga, Thomas B. Croat, Michael R. McKain. "Molecular evolution of chloroplast genomes in Monsteroideae \(Araceae\)", Planta, 2020](#)

< 1% match (student papers from 27-Dec-2019)

[Submitted to Korea University on 2019-12-27](#)

< 1% match (Internet from 03-Apr-2016)

http://mro.massey.ac.nz/bitstream/handle/10179/5610/02_whole.pdf?isAllowed=y&sequence=2

< 1% match (publications)

[Shabina Iram, Muhammad Qasim Hayat, Muhammad Tahir, Alvina Gul, Abdullah, Ibrar Ahmed. "Chloroplast Genome Sequence of Artemisia scoparia: Comparative Analyses and Screening of Mutational Hotspots", Plants, 2019](#)

< 1% match (student papers from 09-Oct-2019)

[Submitted to Emmanuel College on 2019-10-09](#)

< 1% match (publications)

[Flowering Plants · Dicotyledons, 2003.](#)

< 1% match (publications)

[Abdullah, Claudia L. Henriquez, Furrugh Mehmood, Iram Shahzadi et al. "Comparison among the first representative chloroplast genomes of , and of the plant family Araceae: inverted repeat dynamics are not linked to phylogenetic signaling ", Cold Spring Harbor Laboratory, 2020](#)

< 1% match (Internet from 29-Apr-2019)

<http://www.locus.ufv.br/bitstream/handle/123456789/24757/texto%20completo.pdf?isAllowed=y&sequence=1>

< 1% match (publications)

[Claudia L. Henriquez, Abdullah, Ibrar Ahmed, Monica M. Carlsen, Alejandro Zuluaga, Thomas B. Croat, Michael R. McKain. "Evolutionary dynamics of chloroplast genomes in subfamily Aroideae \(Araceae\)", Genomics, 2020](#)

< 1% match (Internet from 08-Jan-2019)

<https://www.mdpi.com/1420-3049/23/10/2426/htm>

< 1% match (publications)

[Iram Shahzadi, Abdullah, Furrugh Mehmood, Zain Ali, Ibrar Ahmed, Bushra Mirza. "Chloroplast genome sequences of Artemisia maritima and Artemisia absinthium: Comparative analyses, mutational hotspots in genus Artemisia and phylogeny in family Asteraceae", Genomics, 2020](#)

< 1% match (Internet from 06-Apr-2020)

<https://www.mdpi.com/1422-0067/20/12/2886/html>

< 1% match (student papers from 11-Jul-2020)

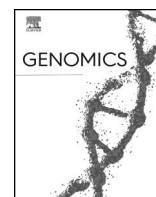
[Submitted to Ain Shams University on 2020-07-11](#)

< 1% match (Internet from 30-Apr-2019)

<https://peerj.com/articles/6663/>

<p>< 1% match (Internet from 08-Mar-2019) https://www.mdpi.com/1422-0067/19/8/2443</p>
<p>< 1% match (publications) Tang, Y.. "An embryological study of <i>Eriolaena candollei</i> Wallich (Malvaceae) and its systematic implications", <i>Flora</i>, 2009</p>
<p>< 1% match (Internet from 20-Apr-2020) https://www.hindawi.com/journals/bmri/2019/4370258/</p>
<p>< 1% match (Internet from 20-Jul-2020) https://link.springer.com/article/10.1186/s13059-016-1004-2?code=4b76afdd-7196-4adf-a9f8-f0867669fd2c&error=cookies_not_supported</p>
<p>< 1% match (Internet from 23-May-2010) http://gupea.ub.gu.se/dspace/bitstream/2077/9534/1/Avhandling%20final.pdf</p>
<p>< 1% match (Internet from 10-Mar-2019) https://peerj.com/articles/6320.pdf</p>
<p>< 1% match (publications) Nyffeler, R.. "Phylogenetic analysis of the Malvadendrina clade (Malvaceae s.l.) based on plastid DNA sequences", <i>Organisms Diversity & Evolution</i>, 20050610</p>
<p>< 1% match (publications) S.-M. Chaw. "Chloroplast Genome (cpDNA) of <i>Cycas taitungensis</i> and 56 cp Protein-Coding Genes of <i>Gnetum parvifolium</i>: Insights into cpDNA Evolution and Phylogeny of Extant Seed Plants", <i>Molecular Biology and Evolution</i>, 03/10/2007</p>
<p>< 1% match (Internet from 30-Jul-2019) https://e-sciencecentral.org/articles/pubreader/SC000033953</p>
<p>< 1% match (student papers from 31-May-2016) Submitted to Yeungnam University on 2016-05-31</p>
<p>< 1% match (publications) Rodrigues, Júlia Alves Marinho(Calmon, Paulo Carlos Du Pin). "Análise de redes e políticas de juventude", <i>RIUnB</i>, 2008.</p>
<p>< 1% match (Internet from 13-Jan-2017) http://www.mdpi.com/2073-4425/8/1/13/pdf</p>
<p>< 1% match (Internet from 20-Jul-2019) https://www.nature.com/articles/s41598-018-27453-7?code=f286df3f-2834-40b3-a596-ef7c84bc0caa&error=cookies_not_supported</p>
<p>< 1% match (publications) Fen Zhang, Wei Li, Cheng-wen Gao, Li-zhi Gao. "Deciphering tea tree chloroplast and mitochondrial genomes of var. ", <i>Cold Spring Harbor Laboratory</i>, 2019</p>
<p>< 1% match (student papers from 16-Aug-2016) Submitted to Vrije Universiteit Brussel on 2016-08-16</p>
<p>< 1% match (Internet from 22-Apr-2019) http://s-space.snu.ac.kr/bitstream/10371/140812/1/000000150082.pdf</p>
<p>< 1% match (publications) Furrukh Mehmood, Abdullah, Iram Shahzadi, Ibrar Ahmed, Mohammad Tahir Waheed, Bushra Mirza. "Characterization of <i>Withania somnifera</i> chloroplast genome and its comparison with other selected species of Solanaceae", <i>Genomics</i>, 2020</p>
<p>< 1% match (student papers from 16-Jan-2012) Submitted to Trinity College Dublin on 2012-01-16</p>
<p>< 1% match (Internet from 13-Apr-2016) http://ojsjournal.org/content/early/2014/08/21/g3.114.012690.full.pdf</p>
<p>< 1% match (publications) Furrukh Mehmood, Abdullah, Zartasha Ubaid, Iram Shahzadi, Ibrar Ahmed, Mohammad Tahir Waheed, Peter Poczaj, Bushra Mirza. "Plastid genomics of (Solanaceae): insights into molecular evolution, positive selection and the origin of the maternal genome of Aztec tobacco (.)", <i>PeerJ</i>, 2020</p>
<p>< 1% match (Internet from 10-Feb-2020) https://www.researchsquare.com/article/e3d4850a-9be0-4cd8-aeb4-358dd910737e/v1</p>
<p>< 1% match (Internet from 16-Dec-2010) http://www.biomedcentral.com/1471-2229/7/45</p>
<p>< 1% match (publications) Sithichoke Tangphatsornruang, Pichahpuk Uthapaisanwong, Duangjai Sangsrakru, Juntima Chanprasert et al. "Characterization of the complete chloroplast genome of <i>Hevea brasiliensis</i> reveals genome rearrangement, RNA editing sites and phylogenetic relationships", <i>Gene</i>, 2011</p>
<p>< 1% match (publications) Abdullah, Claudia L. Henriquez, Furrukh Mehmood, Monica M. Carlsen et al. "Complete Chloroplast Genomes of <i>Anthurium huixtense</i> and <i>Pothos scandens</i> (Pothoideae, Araceae): Unique Inverted Repeat Expansion and Contraction Affect Rate of Evolution", <i>Journal of Molecular Evolution</i>, 2020</p>
<p>< 1% match (Internet from 24-May-2016) http://www.amjbot.org/content/101/11/1987.full.pdf</p>

<p>< 1% match (Internet from 22-Sep-2017) https://repositories.lib.utexas.edu/bitstream/handle/2152/46750/WENG-DISSERTATION-2015.pdf?isAllowed=y&sequence=1</p>
<p>< 1% match (Internet from 28-Mar-2020) https://www.mdpi.com/1420-3049/23/5/1015/html</p>
<p>< 1% match (Internet from 11-Aug-2013) http://bioinf.uta.fi/WASbase/?content=base_table_2/IDbases</p>
<p>< 1% match (Internet from 03-Sep-2019) https://www.frontiersin.org/articles/10.3389/fpls.2018.01811/full</p>
<p>< 1% match (Internet from 28-Sep-2019) https://www.nature.com/articles/s41598-018-20189-4?code=91befdc6-be30-4860-9a04-2295689d181a&error=cookies_not_supported</p>
<p>< 1% match (publications) Danfeng Tang, Fan Wei, Muhammad Haneef Kashif, Fazal Munsif, Ruiyang Zhou. "Identification and analysis of RNA editing sites in chloroplast transcripts of kenaf (<i>Hibiscus cannabinus</i> L.)", <i>3 Biotech</i>, 2019</p>
<p>< 1% match (Internet from 10-Apr-2019) https://peerj.com/articles/4186/</p>
<p>< 1% match (publications) Nazareno, Alison Gonçalves, Monica Carlsen, and Lúcia Garcez Lohmann. "Complete Chloroplast Genome of <i>Tanaecium tetragonolobum</i>: The First Bignoniaceae Plastome", <i>PLoS ONE</i>, 2015.</p>
<p>< 1% match (student papers from 20-Jul-2020) Submitted to University of Huddersfield on 2020-07-20</p>
<p>< 1% match (Internet from 11-Jan-2019) https://epdf.tips/organelle-genetics-evolution-of-organelle-genomes-and-gene-expression.html</p>
<p>< 1% match (publications) Prabhudas, Sudheesh K., Sowjanya Prayaga, Parani Madasamy, and Purushothaman Natarajan. "Shallow Whole Genome Sequencing for the Assembly of Complete Chloroplast Genome Sequence of <i>Arachis hypogaea</i> L.", <i>Frontiers in Plant Science</i>, 2016.</p>
<p>< 1% match (Internet from 15-Mar-2020) https://worldwidescience.org/topicpages/c/chloroplast+dna+cpdna.html</p>
<p>< 1% match (publications) Yang, Yanji, Tao Zhou, Dong Duan, Jia Yang, Li Feng, and Guifang Zhao. "Comparative Analysis of the Complete Chloroplast Genomes of Five <i>Quercus</i> Species", <i>Frontiers in Plant Science</i>, 2016.</p>
<p>< 1% match (publications) Sang-Chul Kim, Jung Sung Kim, Joo-Hwan Kim. " Insight into infrageneric circumscription through complete chloroplast genome sequences of two species ", <i>AoB Plants</i>, 2016</p>
<p>< 1% match (student papers from 23-Jun-2016) Submitted to Gachon University on 2016-06-23</p>
<p>< 1% match (student papers from 13-Jul-2017) Submitted to Laureate Education Inc. on 2017-07-13</p>
<p>< 1% match (publications) Dhafer A. Alzahrani, Samaila S. Yaradua, Enas J. Albokhari, Abidina Abba. "Complete chloroplast genome sequence of <i>Barleria prionitis</i>, comparative chloroplast genomics and phylogenetic relationships among <i>Acanthoideae</i>", <i>BMC Genomics</i>, 2020</p>
<p>< 1% match (Internet from 29-Mar-2018) https://academic.oup.com/gbe/article/5/5/989/610733</p>
<p>< 1% match (Internet from 30-Oct-2018) http://www.locus.ufv.br/bitstream/handle/123456789/19640/artigo.pdf?isAllowed=y&sequence=1</p>
<p>< 1% match (publications) Josphat Saina, Zhi-Zhong Li, Andrew Gichira, Yi-Ying Liao. "The Complete Chloroplast Genome Sequence of <i>Tree of Heaven</i> (<i>Ailanthus altissima</i> (Mill.) (Sapindales: Simaroubaceae), an Important Pantropical Tree", <i>International Journal of Molecular Sciences</i>, 2018</p>
<p>< 1% match (Internet from 02-Jul-2020) https://www.mdpi.com/2223-7747/9/6/752/htm</p>
<p>< 1% match (student papers from 20-May-2015) Submitted to University of Melbourne on 2015-05-20</p>
<p>< 1% match (student papers from 30-Apr-2016) Submitted to University College London on 2016-04-30</p>
<p>< 1% match (Internet from 29-Sep-2012) http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1027&context=agronhortdiss</p>
<p>< 1% match (Internet from 13-Mar-2020) https://www.mdpi.com/1422-0067/19/8/2443/html</p>



Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): Comparative analyses and identification of mutational hotspots



Abdullah^a, Furrukh Mehmood^a, Iram Shahzadi^a, Shahid Waseem^b, Bushra Mirza^a, Ibrar Ahmed^{b,*},
 Mohammad Tahir Waheed^{a,*}

^a Department of Biochemistry, Quaid-i-Azam University, Islamabad, Pakistan

^b Alpha Genomics Private Limited, Islamabad, Pakistan

ARTICLE INFO

Keywords:

Malvaceae
Hibiscus
 Repeats
 Taxonomic discrepancies
 Evolution rate
 Phylogenetics

ABSTRACT

Previous studies to resolve phylogenetic and taxonomic discrepancies of *Hibiscus* remained inconclusive. Here, we report chloroplast genome sequence of *Hibiscus rosa-sinensis*. *Hibiscus rosa-sinensis* chloroplast genome was 160,951 bp, comprising of large single copy (89,509 bp) and small single copy (20,246 bp) regions, separated by IRA and IRB (25,598 bp each). The genome contained 130 genes including 85 protein-coding genes, 37 transfer RNAs and 8 ribosomal RNAs. Comparative analyses of chloroplast genomes revealed similar structure among 12 species within family Malvaceae. Evolutionary rates of 77 protein-coding genes showed 95% similarities. Analyses of codon usage, amino acid frequency, putative RNA editing sites, and repeats showed a great extent of similarities between *Hibiscus rosa-sinensis* and *Hibiscus syriacus*. We identified 30 mutational hotspots including *psbZ-trnG*, *trnK-rps16*, *trnD-trnY*, *trnW-trnP*, *rpl33-rps18*, *petG-trnW*, *trnS-trnG*, *trnH-psbA*, *atpB-rbcL*, and *rpl32-trnL* that might be used as polymorphic and robust markers to resolve phylogenetic discrepancies in genus *Hibiscus*.

1. Introduction

Family Malvaceae or Mallow belongs to order Malvales and consists of 244 genera and 4225 species [1,2]. Due to its plastic morphology, this family has been classified into nine subfamilies including Brownlowioideae, Bombacoideae, Byttnerioideae, Dombeyoideae, Grewioideae, Helicteroideae, Malvoideae, Tilioideae, and Sterculioideae [2,3]. *Hibiscus* is one of the most diverse and widely distributed genera of family Malvaceae consisting of about 250–350 species [4,5]. This genus is included in tribe Hibisceae of subfamily Malvoideae [4]. Species of this genus are herbs, shrubs or trees and distributed in tropical to temperate regions of the world [6]. Members of this genus are used in industries, horticulture, agriculture, food, and medicines [7]. *Hibiscus* also includes species of high medicinal values that have been shown to possess broad curative activities including anti-bacterial, anti-fungal [8] and anti-viral activities [9]. In some cases, species of genus *Hibiscus* also showed activity against hypertension, inflammation, hyperlipidaemia, obesity, and anaemia [10,11]. Anticancer and apoptosis-inducing properties of *Hibiscus* have been also reported [12,13]. Taxonomic discrepancies exist in genus *Hibiscus* due to its plastic morphology [14,15]. Some researchers used molecular markers to resolve taxonomic discrepancies and elucidate phylogeny of family

Malvaceae and *Hibiscus* [15–18] but these studies have been inconclusive.

Hibiscus rosa-sinensis is grown throughout the tropics and subtropics due to its ornamental and medicinal values [5]. Many of its varieties and cultivars are available with same morphology except flower colour that ranges from yellow or white to pink or red with single or double petals, but the flower is not available throughout the year which makes the identification of the cultivar almost impossible [5]. The extensive medicinal activity of *H. rosa-sinensis* has also been reported. For instance, antimicrobial, antioxidant, anti-tumour, anti-diabetic and wound healing [5,19]. Different cultivars and varieties varied in their antimicrobial and antioxidant activities toward different pathogenic species [5,20]. Therefore, the identification of its cultivars is important in judgement of the herbal medicines.



The chloroplast is a self-replicative organelle which plays a vital role in photosynthesis [21]. Chloroplasts are inherited from one parent, maternally in most angiosperm species [22] but paternally in some gymnosperms [23]. It is double membrane bound organelle and contains its double strand DNA in up to 75 kb to 250 kb in size. Chloroplast genome in majority of plant species consists of about 120 genes including transfer RNA (tRNA), ribosomal RNA (rRNA), and protein-coding genes [24]. The chloroplast genome is mostly a quadripartite

* Corresponding author.

E-mail addresses: iaqureshi_qau@yahoo.com (I. Ahmed), tahirwaheed@qau.edu.pk (M.T. Waheed).

Research Article

Correlations among oligonucleotide repeats, nucleotide substitutions, and insertion–deletion mutations in chloroplast genomes of plant family Malvaceae

Abdullah¹ , Furrukh Mehmood¹, Iram Shahzadi¹, Zain Ali^{1,2}, Madiha Islam³, Muhammad Naeem⁴, Bushra Mirza¹, Peter J. Lockhart⁵, Ibrar Ahmed^{2*}, and Mohammad Tahir Waheed^{1*} 

¹Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan

²Research and Development (R&D) Institute, Alpha Genomics Private Limited, Islamabad 45720, Pakistan

³Department of Genetics, Hazara University, Mansehra, Pakistan

⁴Federal Seed Certification and Registration Department, Research and Development (R&D) Institute, Islamabad, Pakistan

⁵Institute of Fundamental Sciences, Massey University, Palmerston North 4474, New Zealand

*Authors for correspondence. Ibrar Ahmed. E-mail: iaqureshi_qau@yahoo.com; Mohammad Tahir Waheed. E-mail: tahirwaheed@qau.edu.pk

Received 10 September 2019; Accepted 14 March 2020; Article first published online 19 March 2020

Abstract The co-occurrence of mutational events including substitutions and insertions–deletions (InDels) with oligonucleotide repeats has previously been reported for a limited number of prokaryotic, eukaryotic, and organelle genomes. In this study, the correlations among these mutational events in chloroplast genomes of species in the eudicot family Malvaceae were investigated. This study also reported chloroplast genome sequences of *Hibiscus mutabilis*, *Malva parviflora*, and *Malvastrum coromandelianum*. These three genomes and 16 other publicly available chloroplast genomes from 12 genera of Malvaceae were used to calculate the correlation coefficients among the mutational events at family, subfamily, and genus levels. In these comparisons, chloroplast genomes were pairwise aligned to record the substitutions and the InDels in mutually exclusive, 250 nucleotide long bins. Taking one among the two genomes as a reference, the coordinate positions of oligonucleotide repeats in the reference genome were recorded. The extent of correlations among repeats, substitutions, and InDels was calculated and categorized as follows: very weak (0.1–0.19), weak (0.20–0.29), moderate (0.30–0.39), and strong (0.4–0.69). The extent of correlations ranged 0.201–0.6 between “InDels and single-nucleotide polymorphism (SNP),” 0.182–0.513 between “InDels and repeat,” and 0.055–0.403 between “SNPs and repeats.” At family- and subfamily-level comparisons, 88%–96% of the repeats showed co-occurrence with SNPs, whereas at the genus level, 23%–86% of the repeats co-occurred with SNPs in same bins. Our findings support the previous hypothesis suggesting the use of oligonucleotide repeats as a proxy for finding the mutational hotspots.

Key words: Correlations, InDels, *Malva*, Malvaceae, *Malvastrum*, mutational dynamics, oligonucleotide repeats.

1 Introduction

The chloroplast is the fundamental organelle in plants that sustains life on earth by converting light energy into chemical energy in the form of carbohydrates (Daniell et al., 2016). Besides its role in photosynthesis, chloroplasts also participate in various important biochemical reactions, including synthesis of nucleotides, amino acids, fatty acids, vitamins, and phytohormones (Daniell et al., 2016). Chloroplast genomes exhibit uniparental inheritance, maternal inheritance in most species of angiosperm (Daniell, 2007) and paternal inheritance in some gymnosperms (Neale & Sederoff, 1989). The general size of the chloroplast genome of photosynthetic plants is independent of the nuclear genome and varies in size from 75 kb to 250 kb (Palmer, 1985). Usually, chloroplast genomes contain about 120 genes, including protein-coding genes,

ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs) (Daniell et al., 2016; Amiryousefi et al., 2018; Menezes et al., 2018). The chloroplast genome is mostly a quadripartite structure, consisting of two inverted repeat regions (IRa and IRb): one large-single copy (LSC) region and one small single-copy (SSC) region (Palmer, 1985). In some species, loss of one copy of the inverted repeat region make the entire genome a single copy (Wu et al., 2011). Moreover, linear chloroplast genomes have also been reported (Oldenburg & Bendich, 2016). Many mutational events take place in the chloroplast genome, including substitutions, insertions–deletions (InDels), structural rearrangements, translocations, inversions, and copy number variations (CNVs) (Jheng et al., 2012; Xu et al., 2015; Abdullah et al., 2019; Shahzadi et al., 2020). Oligonucleotide repeats are also reported among the mutational events in chloroplast genomes. They consist of small repeats that exist



Comparative analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*

Abdullah¹ · Shahid Waseem² · Bushra Mirza¹ · Ibrar Ahmed² · Mohammad Tahir Waheed¹

Received: 29 June 2019 / Accepted: 21 November 2019 / Published online: 10 December 2019
© Institute of Molecular Biology, Slovak Academy of Sciences 2019

Abstract

Theobroma is a plant genus included in tribe Theobromeae, subfamily Byttnerioideae Burnett and family Malvaceae. Discrepancies exist in taxonomy at family and genus level. Here, we compared structures of two previously reported chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum* after correction of annotation and search for highly polymorphic regions which could be used to design molecular markers to investigate phylogenetic relationships among the genus. Both chloroplast genomes showed similar structure, gene content and organization. The correctly annotated genomes contained 130 genes including 8 ribosomal RNA, 37 tRNA genes and 85 protein-coding genes. The amino acids frequencies, codon usage, putative RNA editing sites, microsatellites and oligonucleotide repeats were alike in two genomes. Polymorphic hotspot regions were mostly observed in intergenic regions followed by intronic regions, as well as in coding sequences of few genes. We identified 30 polymorphic loci including *trnH-psbA*, *ndhE-ndhG*, *rpl32-trnL*, *trnP-PsaJ*, *rpl33-rps18*, *trnQ-psbK*, *trnS-trnG*, and *ndhF-rpl32* that might be suitable for development of appropriate and suitable markers for inferring of genus *Theobroma* phylogeny. The inferring of phylogeny with the markers of these loci might be helpful to identify genetically compatible and closely related species to *T. cacao* and *T. grandiflorum* for breeding purposes to produce quality cultivars of these species for high quantity of fruit production and with high resistance towards the pathogens.

Keywords *Theobroma* · Malvaceae; microsatellites · Oligonucleotide repeats · InDels · Substitutions

Abbreviations

LSC	large single copy
SSC	small single copy
IR	inverted repeat regions
Ts/Tv	ratio of transition to transversion substitutions
SNPs	single nucleotide polymorphisms
InDels	Insertions and deletions
Ts	Transitions substitutions
Tv	Transversion substitutions

MAFFT	Multiple Alignment using Fast Fourier Transform
IGS	Intergenic spacer regions
SSR	Simple sequence repeats

Introduction

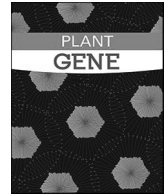
Chloroplast plays an important role in the growth and development of plants. It is not only vital for photosynthesis, but also involved in amino acids and fatty acids synthesis (Cooper 2000). Chloroplasts have 107 kb–218 kb circular and double stranded genome encoding about 120 genes including ribosomal RNA (rRNA), transfer RNA (tRNA) and protein-coding genes (Daniell et al. 2016). Chloroplast genome has quadripartite structure and consists of two inverted repeats regions (IRa and IRb), one large single copy (LSC) and another small single copy (SSC) (Palmer 1985). Several types of mutational events occur in chloroplast genome including structural rearrangements, translocations, inversions, InDels (insertions and deletions) and variations in the number of tandem repeats (Ahmed et al. 2012; Jheng et al. 2012; Xu et al. 2015;

Electronic supplementary material The online version of this article (<https://doi.org/10.2478/s11756-019-00388-8>) contains supplementary material, which is available to authorized users.

- ✉ Ibrar Ahmed
iaqureshi_qau@yahoo.com
- ✉ Mohammad Tahir Waheed
tahirwaheed@qau.edu.pk

¹ Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan

² Alpha Genomics (Pvt) Ltd, Islamabad, Pakistan



Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae

Abdullah^a, Iram Shahzadi^a, Furrukh Mehmood^a, Zain Ali^a, Muhammad Suleman Malik^a, Shahid Waseem^b, Bushra Mirza^a, Ibrar Ahmed^{b,*}, Mohammad Tahir Waheed^{a,*}

^a Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, 45320, Islamabad, Pakistan

^b Alpha Genomics Private Limited, Islamabad, 45710, Pakistan

ARTICLE INFO

Keywords:

Firmiana
Sterculioideae
Malvaceae
Divergence region
Evolutionary rate
Transition and transversion ratio
Substitutions types

ABSTRACT

Firmiana is a small genus of subfamily Sterculioideae of family Malvaceae. Previously, *Firmiana* genus was ignored, and its phylogenetics was not elucidated at genus level. Here, we assembled the chloroplast genomes of *Firmiana colorata* and compared with reported chloroplast genome of *F. major* and *F. pulcherrima*. *Firmiana colorata* had total length of 160,700 bp, contained large single copy (LSC) region of 89,551 bp and small single copy (SSC) region of 20,001 bp that were separated by two inverted repeats regions (IRA and IRB) of 25,574 bp each. The chloroplast genome of *Firmiana colorata* consisted of 113 unique genes including 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. *Firmiana colorata* and *Firmiana major* showed similar genes content and organisation except some variations. Comparative analyses revealed *Firmiana* species had similar GC content, codon usage, RNA editing sites, simple sequence repeats, oligonucleotide repeats, and substitutions. The substitution and InDels analyses showed that IR regions were more conserved as compared to LSC region and SSC region. Here, we observed low transition and transversion ratio (Ts/Tv) of about 0.90 among *Firmiana* species. The ratio of non-synonymous substitutions and synonymous substitutions (Ka/Ks) revealed similar evolutionary rate for protein-coding genes of *Firmiana*. Comparison of protein coding sequences, introns, and intergenic spacer regions (IGS) among *Firmiana* species revealed 30 mutational hotspots with maximum regions from IGS that including *psbZ-trnG-GCC*, *trnR-UCU-atpA*, *ndhD-ccsA*, *petD-rpoA*, *rps4-trnT-UGU*, *rpl33-rps18*, *rpl32-ndhF*, and *psbK-psbI*. Markers design from these mutational hotspots might be authentic, robust, and cost effective for inferring phylogeny of genus *Firmiana* and subfamily Sterculioideae of family Malvaceae.

1. Introduction

Family Malvaceae consists of 244 genera and 4,225 species (Christenhusz and Byng, 2016). Species of the family Malvaceae has plastic morphology and distributed in tropics and temperate regions of the world (Xu and Deng, 2017) and certain taxonomic discrepancies exist at the family level classification (Bayer et al., 1999). Therefore, Malvaceae is divided into nine subfamilies which are Brownlowioideae, Bombacoideae, Byttnerioideae, Dombeyoideae, Grewioideae, Helicteroideae, Malvoideae, Tilioideae, and Sterculioideae (Bayer et al., 1999; Xu and Deng, 2017).

Firmiana is a small genus belong to subfamily Sterculioideae of family Malvaceae and includes deciduous trees (rarely shrubs) (Bayer et al., 1999; Kostermans, 1957; Wilkie et al., 2006). This genus includes

sixteen species in accepted state whereas five species are categorised as unresolved (<http://www.theplantlist.org/>, accessed: December 29, 2018). Species of genus *Firmiana* are distributed in Eastern Africa and South East Asia to Malaysia (Kostermans, 1957). Eight species of *Firmiana* inhabit China (Huang et al., 2011). Some species of *Firmiana* are cultivated for their beautiful shape and lovely flowers (Fan et al., 2013). With reference to medicinal perceptiveness, *Firmiana* species had shown anti-microbial (Ajaib et al., 2014), anti-inflammatory (Lim et al., 2017) and anti-cancer (Woo et al., 2015) properties. Some species also showed neuroprotective (Lim et al., 2017) and hepatoprotective effects (Kim et al., 2015). *Firmiana colorata* is used for the intestinal dysfunction in some tribes of Bangladesh (Azam et al., 2013).

Firmiana genus is ignored due to lack of efficient molecular markers and up to the best of our knowledge, till date, only one study focused on

* Corresponding authors.

E-mail addresses: iaqureshi_qau@yahoo.com (I. Ahmed), tahirwaheed@qau.edu.pk (M.T. Waheed).

<https://doi.org/10.1016/j.plgene.2019.100199>

Received 23 April 2019; Received in revised form 4 July 2019; Accepted 9 July 2019

Available online 10 July 2019

2352-4073/ © 2019 Elsevier B.V. All rights reserved.