Tree Genetics & Genomes
DOI 10.1007/s11295-013-0604-1

ORIGINAL PAPER

The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms

Huma Asif · Asifullah Khan · Asif Iqbal · Ishtiaq Ahmad Khan · Berthold Heinze · M. Kamran Azim

Received: 5 April 2012 / Revised: 28 November 2012 / Accepted: 20 January 2013

Abstract Only a few studies to date have conducted comparative genomics in the *Myrtaceae* family. Here, we report the complete sequence and bioinformatics analysis of the chloroplast genome of *Syzygium cumini* (L.), one of the family members. The size of *S. cumini* cp genome was within the range of reported angiosperm chloroplast genomes. Comparison of *S. cumini* cpDNA sequence with previously reported partial sequences of *S. cumini* revealed several SNPs that resulted in non-synonymous mutations in maturase K and NADH-plastoquinone oxidoreductase subunit-5. These polymorphic characters might serve as intra-specific markers to address whether lineage sorting from polymorphic ancestry has occurred. Comparison of the *S. cumini* chloroplast genome with related dicots revealed an expansion in the intergenic spacer located between IRA/large single copy (LSC) border and the first gene of LSC region, driven by sequence of 54 bp. This type of variation in the intergenic regions can be utilized in the development of species-specific vectors for chloroplast genetic engineering. Several of the longer (30–40 bp) repeats were found to be conserved in other dicot species, suggesting that they might be widespread in angiosperm chloroplast genomes.

Communicated by D. Grattapaglia

H. Asif · A. Khan · A. Iqbal · I. A. Khan · M. K. Azim (✉)
H.E.J. Research Institute of Chemistry, International Center for Chemical and Biological Sciences, University of Karachi, Karachi 75270, Pakistan
e-mail: kamran.azim@iccs.edu

M. K. Azim
e-mail: mkamranazim@yahoo.co.uk

B. Heinze
Department of Genetics, Federal Research Centre for Forests,
Hauptstraße 7,
1140 Vienna, Austria

Keywords *Myrtaceae* · Plastid · Next-generation sequencing · *Eucalyptus*

Introduction

Chloroplast, an essential organelle in plants, was first reported to have its own genome by Sager in 1963 (Sager and Ishida 1963). The chloroplast (cp) genome is a circular molecule of double-stranded DNA ranging in size from 72 to 217 kb, and contains ~130 genes depending on the plant species (Tangphatsornruang et al. 2010). A typical cp genome consists of four distinct regions, a large and a small single copy region (LSC and SSC, respectively) separated by two inverted repeat regions (IRA and IRB) (Sugiura 1995). The cp genomes of higher plants are highly conserved in gene content, organization and structure and have a full complement of the transcriptional and translational machinery to express their genetic information. Despite the high degree of conservation of the cp genome, mutations, duplications, losses and rearrangements of genes have been observed in plants and algae (Wolfe et al. 1991). Due to the conserved nature and extensive characterization at the molecular level, the cp genome has long been a focus of research in plant molecular evolution and systematics (Raubeson and Jansen 2005). Moreover, uniparental inheritance of the chloroplast genetic repository has made it an excellent tool for genetic engineering of economically important crops (Tangphatsornruang et al. 2010).

Syzygium cumini (L.) Skeels (also classified as *Eugenia jambolana* L.) belongs to family *Myrtaceae.* It is commonly known as Jambolan or Java plum in English and Jamun in the Indian subcontinent (Pepato et al. 2005). *S. cumini* is native to the Indian subcontinent (i.e. India, Pakistan, Bangladesh, Nepal and Sri Lanka), the Malay Peninsula (including parts of Burma, Malaysia, Singapore and Thailand) and Indonesia. It is widely cultivated in tropical America, and there are also records of cultivation and

Published online: 24 February 2013

naturalisation in Australia (Ross 2003). It is a large evergreen tree up to 30 m tall, with a diameter of up to 3.6 m and a circumference of up to 15 m (Jadhay et al. 2009). The fruit is oval to elliptic, 1.5–3.5 cm long, dark purple or nearly black, succulent, fleshy and edible. It contains a single large seed. Economic importance of this agro-forest tree is due to its medicinally and nutritionally valuable fruit. Leaves and flower buds of *S. cumini* have diuretic, the bark has anti-diarrheal and HIV-1 proteases inhibitory activity and seeds have anticonvulsant effects in humans (Teixera et al. 1997; Consolini et al. 1999). The bark, fruits, seeds and leaves of this plant have shown hypoglycemic effect and therefore are used for treatment of diabetes in folk medicine (Grover et al. 2000). *S. cumini* is a diploid cultivated species with the chromosome count of $2n=66$. Studies on pollination mechanisms in *S. cumini* indicated that pollinator activity is essential in this species, and gravity, insects and wind are involved in pollination (Jai and Singh 2007). The most common and simplest method of raising the *S. cumini* tree is from seed; however, for improved and selected true-to-true fruits, vegetative methods of propagation like budding, air layering and inarching are desirable (Rajan and Markose 2007).

Few reports have been published on genetic diversity of cpDNA from the *Myrtaceae* (van der Merwe et al. 2005; Craven and Biffin 2005; Biffin et al. 2006). The chloroplast genome sequences of two *Myrtaceae* family members, *Eucalyptus globulus* (cpDNA size 160,286 kb) (Steane 2005) and *Eucalyptus grandis* (cpDNA 160,137 kb) (Paiva et al. 2011), have been reported. Here, we report on the complete chloroplast genome sequence of *S. cumini*, the third member of *Myrtaceae* for which a complete chloroplast genome sequence is now available. We have analysed the structural organization, gene order, gene content, location and distribution of repetitive sequences. We have also compared our data with the partial *S. cumini* chloroplast DNA sequences available in GenBank. Due to the paucity of molecular biology data on *S. cumini*, this report would provide an impetus for further studies related to intra- and inter-specific genetic variation in *S. cumini* and, moreover, other members of the economically important genus as e.g. cloves (S*zygium aromaticum*).

Methods

Plant material

Leaves obtained from a *S. cumini* tree present in the botanical gardens of University of Karachi, Karachi, Pakistan were used for this study. The voucher specimen is kept at the Herbarium, Department of Botany, University of Karachi, under voucher specimen number: 01 and General Herbarium number: KUH81317.

DNA extraction and genome sequencing

DNA was extracted from the chloroplast-enriched organelle fractions (Triboush et al. 1998; Jansen et al. 2005) obtained from young leaves of *S. cumini* by a modified CTAB method (Porebski et al. 1997) and the commercially available AxyPrep Multisource Genomic DNA Miniprep Kit (Axygen Biosciences, USA). In order to sequence the complete cpDNA of *S. cumini*, we adopted both Sanger-based and next-generation sequencing (NGS) technologies. Initially, a primer walking strategy termed as 'ASAP: Amplification, sequencing and annotation of plastomes' was used for amplification and sequencing of the inverted repeat region (Dhingra and Folta 2005). For this, the enriched cpDNA was used as a template. First, five large amplicons with sizes of 3–7 kb were generated with the consensus set of primers described in the ASAP protocol (Dhingra and Folta 2005). These large amplicons were used as templates for generation of small amplicons ranging in size from 780 to 1,150 bp. Internal sets of primers described earlier were used for this purpose (Dhingra and Folta 2005).The small amplicons were sequenced using the CEQ8000 Genetic Analyzer (Beckman Coulter Inc., USA). The primers used in amplification procedures were utilized for cycle sequencing reactions with the DTCS kit (Beckman Coulter Inc., USA) as recommended by the manufacturer.

The complete cp genome was sequenced by NGS technology. For this purpose, 9.21 μg of *S. cumini* chloroplast-enriched DNA was used for the construction of a paired-end library with insert size of 250 bp followed by DNA sequencing using a HiSeq2000 system (Illumina Inc., San Diego, USA). The expected size of paired-end sequences was 75 bp.

Genome assembling, annotation and analysis

Nucleotide sequence reads obtained by the Sanger method and NGS were assembled by the Lasergene package (DNASTAR Inc. USA) and CLC Genomics Workbench (CLC bio, Denmark) using the *E. globulus* chloroplast genome (Steane 2005) as a reference. The *S. cumini* cp genome was annotated using the program Dual Organellar GenoMe Annotator (Wyman et al. 2004). The predicted annotations were confirmed using BLAST (Altschul et al. 1990) similarity searches and ORF Finder (http://www.ncbi.nlm.nih.gov/gorf). All genes, tRNAs and rRNAs were identified by BLASTN searches against Chloroplast DB (chloroplast database; Cui et al. 2006). The circular map of the *S. cumini* cp genome was drawn by the GenomeVx online tool followed by manual modification (Conant and Wolfe 2008).

Detection of intra-specific variation within chloroplast DNA of *S. cumini*

GenBank contains 13 entries of partial sequences of *S. cumini* cpDNA. The entries contain sequences of *matK* (DQ088575.1,

Table 1 Sequencing and assembly results for cpDNA of *S. cumini* (L.)

Total number of paired-end reads	3,750,000
Average read length	73.5 bp
Reads used in complete cp genome assembly	370,880 (0.37 M) (10 %)
Number of bases used in complete cp genome assembly	27,259,680
Fold coverage of chloroplast genome	170X
(Fold coverage = number of bases sequenced/estimated length of the cpDNA)	

GU134997.1, GU135062.1, AY525140.1), *rbcL* (GU135161.1, GU135224.1), *ndhF* (DQ088496.1, AY498814.1) and *psbA genes* (GU135395.1); and *trnK_UUU–matK* (AY525140.1, two entries) and *trnH_psbA* (GU135329.1; GU135395.1) intergenic spacers. To reveal any intra-specific variations (SNPs and Indels), we compared the *S. cumini* cpDNA sequence with these previously deposited short, partial sequences present in GenBank, using standard alignment methods.

Gene content analysis and comparative genomics

A comparative analysis of the *S. cumini* cp genome was carried out with *E. globulus* and *E. grandis* of the same family (i.e. Myrtaceae) and four species from gradually more distantly related dicot families (i.e. *Arabidopsis thaliana* from Brassicaceae, *Gossypium barbadense* from Malvaceae, *Vigna radiata* from Fabaceae and *Nicotiana tabacum* from Solanaceae) (Shinozaki et al. 1986; Ibrahim et al. 2006) using Lasergene package (DNASTAR Inc. USA). The gene content of all species was visually checked and compared with their cp genome sequences downloaded from NCBI.

Gene order and repeat analyses

Gene order among chloroplast genomes was analysed and visualized by using the online zPicture software (Ovcharenko et al. 2004). REPuter (Kurtz and Schleiermacher 1999) was used to identify and locate direct and inverted repeats in the chloroplast genomes of *S. cumini* (GenBank accession GQ870669), *E. globulus* (GenBank accession NC_008115) (Steane 2005), *E. grandis* (GenBank accession NC_014570) and *N. tabacum* (Shinozaki et al. 1986) (GenBank accession NC_001879) using repeat cutoff size of 30 bp and a Hamming distance of 3 (i.e. a sequence identity of 90 % or greater).

Results

Chloroplast genome assembly

Sequencing of the *S. cumini* cpDNA was carried out by a combination of Sanger-based and NGS methods. Initially, 21,751 bp of the inverted repeat region was sequenced using the ASAP protocol (Dhingra and Folta 2005) and submitted to GenBank under accession number GQ870669. The ASAP method resulted in 82 % coverage of the inverted repeat (IR). The remaining genome was sequenced by NGS technology (Table 1). The low-quality bases were trimmed using a minimum threshold value of $\leq Q_{20}$ bases. Assembly of the nucleotide sequence reads, using the *E. globulus* chloroplast genome (GenBank accession NC_008115) as a reference, resulted in an alignment of 370,880 reads (i.e. 10 % of the total reads). The unassembled reads (~90 %) were mostly from the nuclear genome due to nuclear DNA contamination during chloroplast DNA isolation. By integrating both Sanger and NGS reads in this assembly, an average of 170X coverage of the entire genome was achieved (Table 1). The complete DNA sequence of the *S. cumini* chloroplast genome has been deposited in GenBank under accession number GQ870669.

General features of the *S. cumini* cp genome

Table 2 contains the statistics of main characteristics of the cp genome of *S. cumini*. Gene organization is shown in Fig. 1. The GC content is comparable to that of *E. globulus* (Steane 2005) and *E. grandis* (Paiva et al. 2011). Forty-three percent of the genome comprises non-coding regions, including introns, pseudo genes and intergenic spacers, and 57 % comprises coding regions. The list of protein coding

Table 2 Summary of complete chloroplast genome of *S. cumini*

Total cpDNA size	160,373 bp
Size inverted repeat (IR) region	26,392 bp
Size of large single copy (LSC) region	89,081 bp
Size of small single copy (SSC) region	18,508 bp
Total number of genes	128 genes
Number protein coding genes	83
tRNA genes	37
rRNA genes	8
Pseudogenes	2
GC content	36.83 %
Coding regions (proportion of whole genome)	57 %

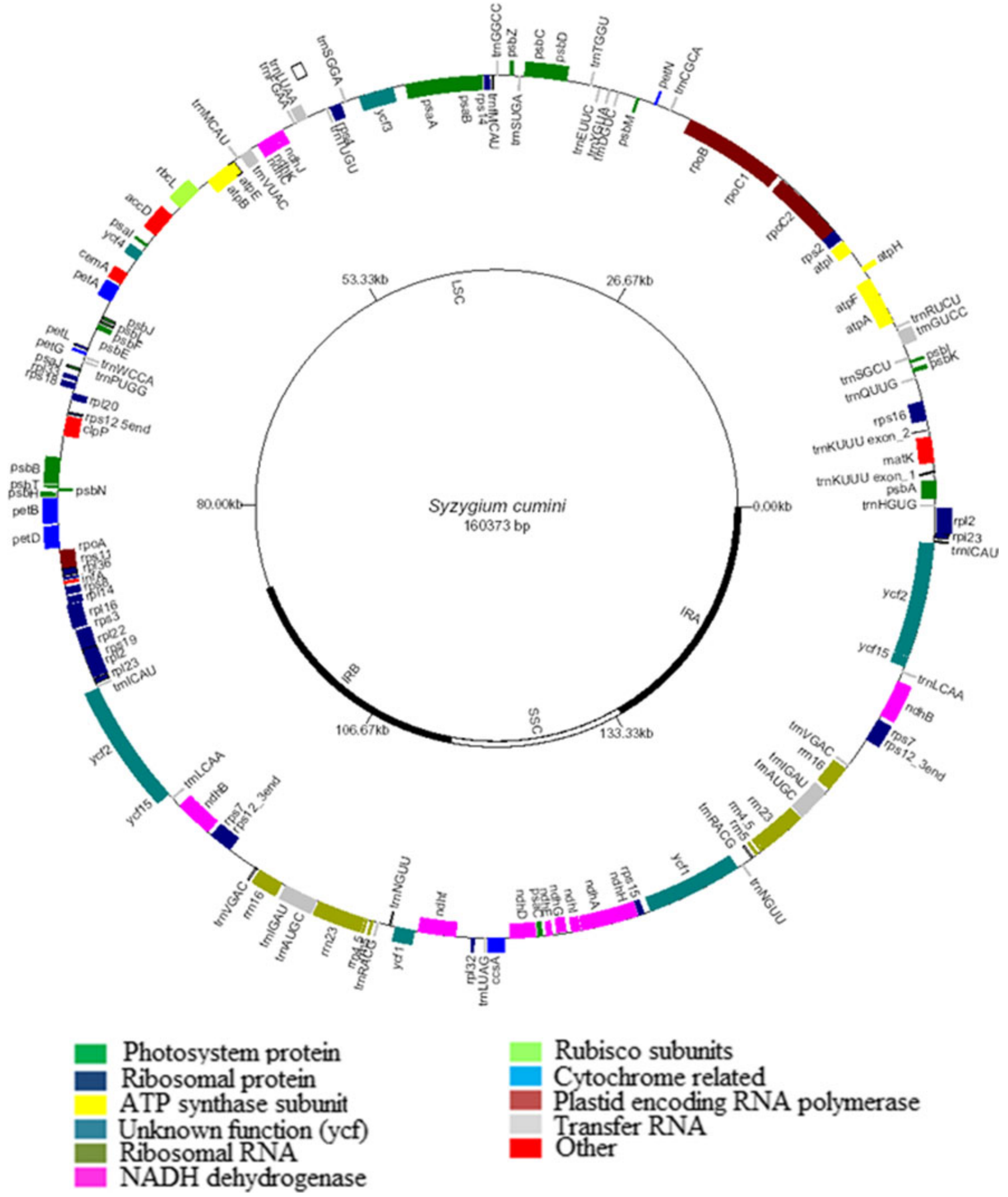

Fig. 1 Gene organization of the chloroplast genome from *S. cumini* L. *Genes outside the circle* are transcribed in the clockwise direction, and genes *inside the circle* are transcribed in the counterclockwise direction. The *central circle* represents the different regions (LSC, SSC and two IR) of the cp genome

genes annotated in *S. cumini* cp genome is given in Table 3. Of the 128 genes in the *S. cumini* cpDNA, 17 were in the IR regions (and were, therefore, duplicated). There were 30 distinct tRNAs and 4 ribosomal RNA genes; 7 of the tRNA genes and all rRNA genes occurred within the IR regions (Table 3). Twelve protein coding regions and six tRNA genes contained introns. The *clpP* and *ycf3* genes each contained 2 introns, resulting in a total of 20 introns across 18 genes in the cp genome.

Intra-specific variation of the chloroplast DNA of *S. cumini*

The comparison of our *S. cumini* cpDNA sequence with previously submitted entries in GenBank indicated eight indels

Table 3 Genes in the chloroplast genome of *S. cumini*

RNA genes	
Transfer RNA genes	*trnH-GUG*, *trnK-UUU*[a], *trnQ-UUG*, *trnS-GCU*, *trnG-UCC*[a], *trnR-UCU*, *trnC-GCA*, *trnD-GUC*, *trnY-GUA*, *trnE-UUC*, *trnT-GGU*, *trnS-UGA*, *trnG-GCC*, *trnfM-CAU*, *trnS-GGA*, *trnT-UGU*, *trnL-UAA*[a], *trnF-GAA*, *trnV-UAC*[a], *trnM-CAU*, *trnW-CCA*, *trnP-UGG*, *trnI-CAU*[c], *trnL-CAA*[c], *trnV-GAC*[c], *trnI-GAU*[a,c], *trnA-UGC*[a,c], *trnR-ACG*[c], *trnN-GUU*[c], *trnL-UAG*
Ribosomal RNAs genes	*rrn16S*[c], *rrn23S*[c], *rrn4.5S*[c], *rrn5S*[c]
Transcription and translation-related genes	
RNA polymerase and related genes	*rpoA*, *rpoB*, *rpoC1*[a], *rpoC2*
Ribosomal protein genes	
Large subunit	*rpl2*[a, c], *rpl14*, *rpl16*[a], *rpl20*, *rpl22*, *rpl23*[c], *rpl32*, *rpl33*, *rpl36*
Small subunit	*rps2*, *rps3*, *rps4*, *rps7*[c], *rps8*, *rps11*, *rps12*[a,c,d], *rps14*, *rps15*, *rps16*[a], *rps18*, *rps19*
Photosystem-related genes	
Rubisco large subunit gene	*rbcL*
Photosystem I genes	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
Photosystem II genes	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*(*ycf9*)
Assembly/stability of photosystem I	*ycf3*[b], *ycf4*
Cytochrome b/f complex genes	*petA*, *petB*[a], *petD*[a], *petG*, *petL*, *petN*
c-type Cytochrome genes	*ccsA* (*ycf5*)
Proteasome-like synthase genes	*atpA*, *atpB*, *atpE*, *atpF*[a], *atpH*, *atpI*
NADH dehydrogenase genes	*ndhA*[a], *ndhB*[a,c], *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Envelope membrane protein	*cemA* (*ycf10*)
Acetyl-CoA carboxylase gene	*accD*
ATP-dependent protease subunit	*clpP*[b]
Others	
Maturase	*matK*
Conserved reading frames (ycfs)	*ycf1*, *ycf2*[c]
Pseudogenes	*ycf15*[c], *infA*

[a] Gene containing one intron

[b] Genes containing two intron

[c] Gene in the IR regions

[d] Divided gene

(five one-base and three multi-base 2–6 bp). All one-base and two multi-base indels were present in the *'matK-trnK-UUU'* and *'trnH-psbA'* intergenic spacers. We detected a number of non-synonymous SNPs in *matK*, *rbcL* and *ndhF* genes that would result in amino acid mutations.

The *matK* protein coding region comprises 1,518 bp encoding 505 amino acid residues of Maturase K. Nucleotide sequences of the *matK* gene are present in GenBank from trees in Australia (two entries, complete sequence of 1,535 bp) and USA (two entries, partial sequence of 818 bp). Multiple alignments of these *matK* amino acid sequences show that the sequences from Australia and USA are identical, but differ from our Pakistan sample by variation at six amino acid positions (Fig. 2). Similarly, *S. cumini* from Pakistan has eight non-synonymous substitutions in the *ndhF* gene relative to identical *ndhF* sequences from Australian and American samples (Fig. 3). Comparison of partial nucleotide sequences of Ribulose-1,5-bis-phosphate carboxylase/oxygenase large subunit (*rbcL*) gene reported from USA (partial sequence of 567 bp) with that from Pakistan reveals mutations at two positions. These intra-specific cpDNA variations would be worth investigating for phylogeographic studies of *S. cumini* and for cultivar identification/characterization (Khan and Azim 2011; Khan et al. 2012; Lavin et al. 1991).

Gene content analysis and comparative genomics of *S. cumini* cpDNA and six other dicots

We conducted a comparative analysis of the cp genomes of *S. cumini*, *E. globulus*, *E. grandis* (Myrtaceae) and our species from progressively more distantly related dicot families (i.e. *A. thaliana* from Brassicaceae, *G. barbadense* from Malvaceae, *V. radiata* from Fabaceae and, *N. tabacum* from Solanaceae) (Shinozaki et al.

Fig. 2 Alignment of *S. cumini* Maturase K amino acid sequences (abbreviated as *MK*) reported from Australia (*Aus*), USA (*USA*) and Pakistan (*Pak*). GenBank accession numbers are mentioned in *parentheses*. Amino acid substitutions are highlighted with an *asterisk* (*)

```
MK-Aus(DQ088575)     -------------------K------------------------------ 200
MK-Aus(AY525140)     -------------------K------------------------------ 200
MK-USA(GU134997)     -------------------K------------------------------ 57
MK-USA(GU135062)     -------------------K------------------------------ 57
MK-Pak(GQ870669)     LIPYHIHLEILVQTLRYWVNDASSLHLLRFFLHEYWNSLITPKKHITLFS 200
                                        *

MK-Aus(DQ088575)     -------------------------------------------------- 250
MK-Aus(AY525140)     -------------------------------------------------- 250
MK-USA(GU134997)     -------------------------------------------------- 107
MK-USA(GU135062)     -------------------------------------------------- 107
MK-Pak(GQ870669)     KGNPRLFLFLYNSHICEYESTFLFLRNQSSHLRSTSSGIFFERIYFYVKI 250

MK-Aus(DQ088575)     ---A-------------------------------G--------N----- 300
MK-Aus(AY525140)     ---A-------------------------------G--------N----- 300
MK-USA(GU134997)     ---A-------------------------------G--------N----- 157
MK-USA(GU135062)     ---A-------------------------------G--------N----- 157
MK-Pak(GQ870669)     EHFVKVFFDNDFQCILWFFKDPFMHYVRYQGKSILASKDTPLLMKKWKYY 300
                        *                               *        *

MK-Aus(DQ088575)     --T----------------------------------P------------ 350
MK-Aus(AY525140)     --T----------------------------------P------------ 350
MK-USA(GU134997)     --T----------------------------------P------------ 207
MK-USA(GU135062)     --T----------------------------------P------------ 207
MK-Pak(GQ870669)     LVNLWQYHFYAWFQPGRIDINQLCKYSLDFLGYRSSVRLNSSVVRSQMLE 350
                       *                                  *

MK-Aus(DQ088575)     -------------------------------------------------- 400
MK-Aus(AY525140)     -------------------------------------------------- 400
MK-USA(GU134997)     -------------------------------------------------- 257
MK-USA(GU135062)     -------------------------------------------------- 257
MK-Pak(GQ870669)     NSFLINNAMKKFETIVPIIPLIGSLSKANFCNTLGHPISKPTRADSSDSD 400
```

1986; Ibrahim et al. 2006). *S. cumini* cpDNA exhibits the same gene order when compared to the above mentioned dicots, with the known exception of *V. radiata* (Tangphatsornruang et al. 2010). Gene content analysis of the *S. cumini*, *N. tabacum*, *E. globulus*, *E. grandis* and *G. barbadense* cp genomes revealed the presence of *infA* as a pseudogene (that encodes an initiation factor protein), though it is absent in *A. thaliana* (Sato et al. 1999). Likewise, *ycf15* gene was found as a pseudogene in *S. cumini*, *E. globulus*, *E. grandis* and *V. radiata* while in *N. tabacum* it encodes a specific protein. The *rps16* gene is a pseudogene in *V. radiata*, whereas in *S. cumini*, *E. globulus* and *G. barbadense*, it encodes a 16S ribosomal protein (Steane 2005; Ibrahim et al. 2006). Similarly, the *rpl33* and rpl22 genes are present in *S. cumini*, *E. globulus*, *E. grandis* and *G. barbadense* but are absent from *V. radiata* (Tangphatsornruang et al. 2010).

The analysis revealed that the cpDNA of *S. cumini* is the largest of these chloroplast genomes (Table 4) due, in part, to expansion in non-coding sequences (intergenic sequences and introns) in the LSC region. Comparison of intronic sequences showed that the total length of introns in *S. cumini* (14,469 bp) is about the same as in *E. globulus* and *E. grandis* but larger than *N. tabacum* (13,983 bp) and *A. thaliana* (14,235 bp;

Fig. 3 Alignment of *S. cumini* NADH-plastoquinone oxidoreductase subunit-5 amino acid sequences (encoded by *ndhF* gene; abbreviated as *NF*) reported from Australia (*Aus*), USA (*USA*) and Pakistan (*Pak*). GenBank accession numbers are mentioned in *parentheses*. Amino acid substitutions are highlighted with an *asterisk* (*)

```
NF-US(AY498814) LNA--------------------------------------------------------- 139
NF-AS(DQ088496) -ND--------------------------------------------------------- 59
NF-PK(GQ870669) LNASWLYSPIFAIIACSTAGLTAFYMFRIYLLTFEGHFNVHFQNYSGQKSRSYYSISLWG 480
                  *

NF-US(AY498814) --------N--------------------------------H------H-G--------- 199
NF-AS(DQ088496) --------N--------------------------------H------H-G--------- 119
NF-PK(GQ870669) KEVPKTIKKNFLSLLTMNNNERASFFSNKTYQIGGNGKNRMRPFITITNFVTKNTFSYPH 540
                        *                                *      * *

NF-US(AY498814) ---------------------------S-------------------------------- 259
NF-AS(DQ088496) ---------------------------S-------------------------------- 179
NF-PK(GQ870669) ESDNTMLFSMVILVLFTLFVGVVGIPFPFNQEGIHLDILSKLLNPSINLLHQNSNNSVDW 600
                                           *

NF-US(AY498814) ---------------------------------------------P-----D-------- 319
NF-AS(DQ088496) ---------------------------------------------P-----D-------- 239
NF-PK(GQ870669) YEFVTNASFSVGIAFFGIFIASFLYKPIYSSLQNLNLLNLFSKRGSNRILGDKIINVIYD 660
                                                             *     *
```

Table 4 Comparison of chloroplast genome of *S. cumini* with representative species of dicot families

Dicot family	*Myrtaceae*	*Myrtaceae*	*Malvaceae*	*Brassicaceae*	*Fabaceae*	*Solanaceae*
Species	*S. cumini*	*E. globulus*	*G. barbadense*	*A. thaliana*	*V. radiata*	*N. tabacum*
Total size (bp)	160,373	160,286	160,317	154,478	151,271	155,939
LSC (bp)	89,081	89,012	88,841	84,170	80,896	86,686
IR (bp)	26,392	26,393	25,591	26,264	26,474	25,341
SSC (bp)	18,508	18,488	20,294	17,780	17,427	18,571

LSC large single copy region, *IR* inverted repeat region, *SSC* small single copy region

Table 5). In contrast to these dicots, *G. barbadense* shows an even larger expansion of introns, i.e. 14,964 bp in the LSC region (Ibrahim et al. 2006). Therefore, we conclude that expansion of non-coding regions of DNA does not fully account for the relatively large cp genome of *S. cumini*.

Among various species, the junctions between the two inverted repeat regions (IRA and IRB) and the two single copy regions (LSC and SSC) show size variation. The size of the IR region in *S. cumini* cpDNA is 26,392 bp with 20 genes. We compared the sizes of the IR region and IR border positions in the cpDNA sequences of the six dicot species (Fig. 4). In all species, the IRA/SSC junction is situated in the coding region of the *ycf1* gene which results in the duplication of the 3′ end region of this gene. This duplication produces a pseudogene of variable length at the IRB/SSC border. At the IRB/SSC junction, the *ycf1* pseudogenes of *S. cumini*, *E. globulus*, *E. grandis* and *A. thaliana* are slightly shifted towards the SSC region, an expansion of the IR as relative to *G. barbadense* and *N. tabacum*. However, this situation is different in *A. thaliana*, where the *ycf1* pseudogene has an overlap of 37 bp with the *ndhF* gene. In contrast to the IR/SSC borders, only slight changes are seen at the IR/LSC borders. The border at IR/LSC is located in intergenic spacers in all the compared plants (*S. cumini*, *E. globulus*, *E. grandis*, *N. tabacum* and *G. barbadense*), except for *A. thaliana*, where it is located within the coding region of *rps19*. In *Eucalyptus*, intra- and inter-specific variations were observed in this region. Consequently, the IR/LSC border region has been used

Table 5 Distribution and length of introns in chloroplast genome of *S. cumini* and representative dicots (longest introns are highlighted in bold)

Genes containing introns	Intron length (bp)				
	Sc	**Eg**	**Nt**	**At**	**Gb**
ndhA	1,065	1,066	**1,148**	1,080	1,076
trnA	**803**	**803**	709	801	795
trnI	947	952	707	729	**959**
rps 12-3′end	**546**	**546**	536	537	536
ndhB	683	683	679	**685**	683
rpl2	664	664	666	682	**688**
rpl16	992	994	1,020	1,056	**1,135**
petD	766	**770**	742	709	754
petB	780	772	753	804	**821**
Clpp-1	621	621	807	**891**	890
Clpp-2	**883**	881	637	539	679
trnV	594	599	571	599	**609**
trnL	504	510	503	512	**582**
Ycf3-1	735	759	738	714	777
Ycf3-2	760	734	783	787	**789**
rpoc1	737	737	738	**791**	753
atpF	764	759	695	739	**805**
rps16	873	**875**	860	865	870
trnG	744	744	691	715	**763**
trnK	2,554	2,555	2,526	**2,559**	2,535
Total	14,461	14,469	13,983	14,235	14,964

Sc S. cumini, *Eg E. globules*, *Nt N. tabacum*, *Gb G. barbadense*, *At A. thaliana*

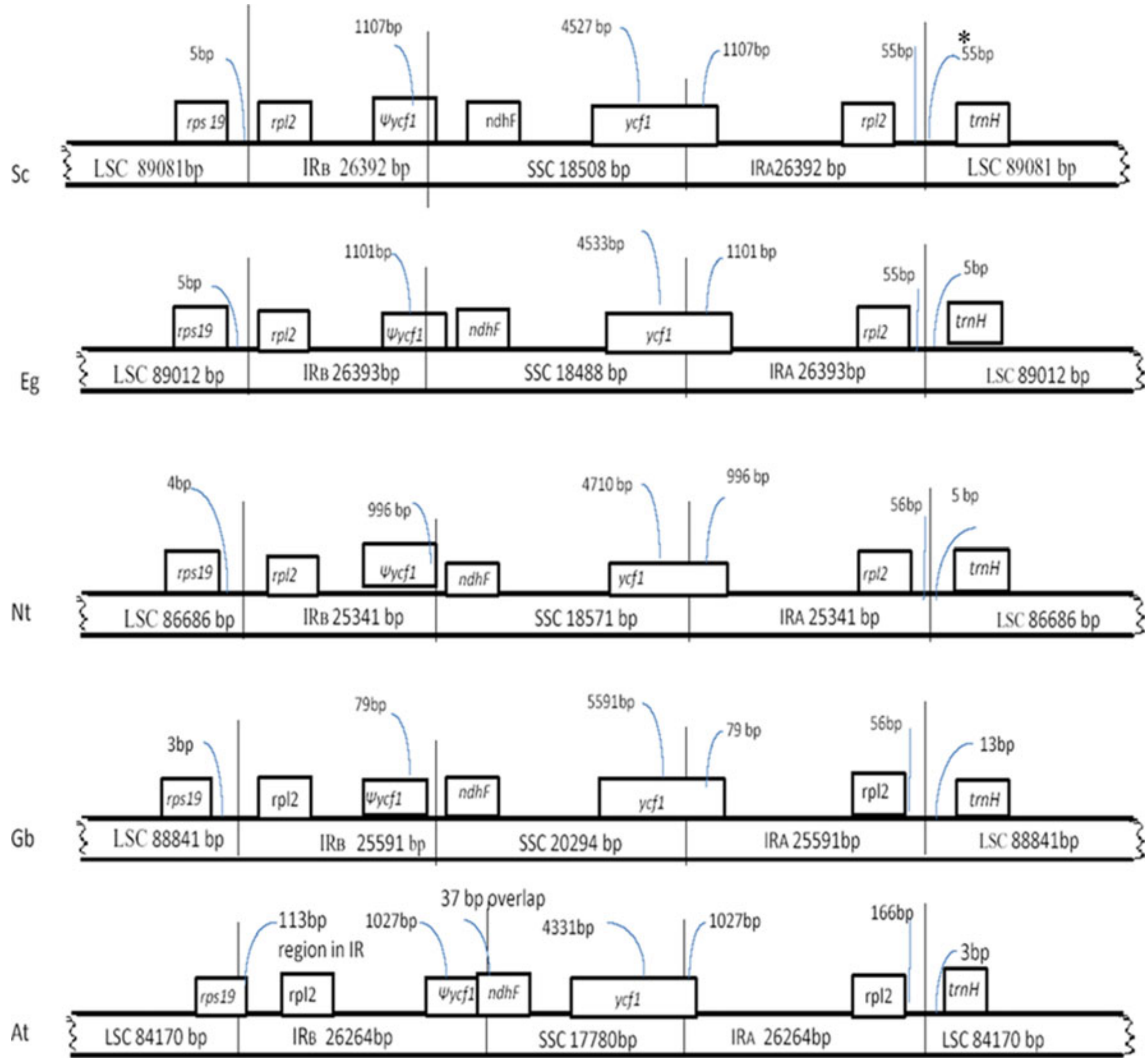

Fig. 4 Comparison of border positions of LSC, SSC and IR among *S. cumini* and related dicot species. *Boxes above the main line* indicate the predicted genes, while pseudogenes at the *borders* are shown by *Ψ*. Insertion of 55 bp in cpDNA of *S. cumini* located between the IRA/LSC border is indicated by *asterisk*. The figure is not to the scale and just shows relative changes at or near the IR-SC borders. *Sc S. cumini*, *Eg E. globulus*, *Nt N. tabacum*, *Gb G. barbadense*, *At A. thaliana* and *Egr E. grandis*

extensively for phylogeographic studies in *Eucalyptus* (Vaillancourt and Jackson 2000; Freeman et al. 2007; Foster et al. 2007) and has the potential to be equally useful for *Syzygium* and other species of *Myrtaceae*.

Beside these minor changes at the border positions, a major drift was observed in the *S. cumin* in the intergenic spacer located between the IRA/LSC border and the first gene of the LSC region. This intergenic spacer is 55 bp in *S. cumini* in contrast to the other dicots where it ranges in size of 2 (Shinozaki et al. 1986) to 12 bp (Ibrahim et al. 2006). The length of this region differs even between *S. cumini* and *E. globulus*, which belongs to the same family. Hence, this

Fig. 5 The alignment of *S. cumini* cpDNA sequence (positions 01–66) with corresponding sequences of *E. globulus*, *E. grandis*, *A. thaliana*, *N. tabacum* and *G. barbadense.* The 54-bp insertion in *S. cumini* cpDNA is highlighted

```
S.cumini      01 AAATGCTTTCTAGGTTTGTCTTGTGTCTCCGTAACAAATGGTCCATTTATTGATGGGCGAACGACG 66
E.globulus    01 TAATGG--------------------------------------------------GCGAACGACG 16
E.grandis     01 GG------------------------------------------------------GCGAACGACG 12
A.thaliana    01 ATGG----------------------------------------------------GCGAACGACG 14
N.tabacum     01 TTATGG--------------------------------------------------GCGAACGACG 16
G.barbadense  01 TTAAAAGAAAATGG------------------------------------------GCGAACGACG 24
```

Table 6 Size and location of repeated sequences in cpDNA of *S. cumini*

Number	Size (bp)	Repeat	Location
1	42	D	ycf3 (intron): ndhA (intron)
2	40	D	IGS between rps12_3′end-trnV-GAC: ndhA (intron)
3	39	D	ycf3 (intron): IGS rps12_3′end-trnV-GAC
4	41 and 31	D	psaB: psaA and psaB: psaA respectively
6	36	D	IGS between atpH and atp I: IGS between rps8 and rpl14
7	31	D	IGS between psaI and ycf4: IGS between psaI and ycf4
8	30	D	IGS between psbI and trnS-GCU: IGS between psbC and trnS-UGA
9	30	D	ycf3 (intron): IGS between rps12 3′end and trnV-GAC
10	30[a], 31[b]	D	ycf2: ycf2
11	30	D	IGS between rps12 3′end and trnV-GAC: ndhA (intron)
13	31	D	ycf3 (intron): IGS between rps12 3′end and trnV-GAC
16	30	D	IGS between rpl32 and trnL-UAG: IGS between rpl32 and trnL-UAG
17	40	IR	atpA: IGS between atpF and atp H
18	39	IR	ycf3 (intron): IGS between trnV-GAC and rps12 -3′end
19	30	IR	IGS between psb I and trnS-GCU: trn S-GGA
20	30	IR	psbD: IGS and ndhC & trn V-UAC
21	30	IR	ycf3 (intron): IGS between trnV-GAC and rps12 -3′end
22	30	IR	IGS between trnT-UGU and trn L-UAA: IGS between trnT-UGU and trnL-UAA
23	30[a], 31[b]	IR	ycf2: ycf2
25	30	IR	ndhA (intron): IGS between trnV-GAC and rps12-3′end
26	31	IR	IGS between trnE-UUC and trnT-GGU: IGS between trnE-UUC and trnT-GGU
27	31	IR	ycf3 (intron): IGS between trnV-GAC and rps12-3′end
28	31	IR	ycf3 (Intron): ycf3 (Intron)
31	30	IR	IGS between psbC and trnS-UGA: trnS-GGA
32	30	IR	IGS between psbZ and trnG-UCC: ndhAb (Intron)

D direct repeat, *IR* inverted repeat, *IGS* intergenic spacer

[a] Two different repeats appeared in separate locations

[b] Three different repeats appeared in separate locations

property can be considered as an important contributor to the expansion of the LSC region of *S. cumini* cpDNA compared to the other selected dicots (Fig. 5).

Repeat structure analysis

Repeat structure analysis using REPuter (Kurtz and Schleiermacher 1999) identified 32 repeats in the *S. cumini* cp genome under the conditions listed in "Methods" section; 16 repeats are direct and 16 are inverted repeats. Sizes of these repeats are in the range of 30–42 bp. The *E. globules* cp genome has a total of 32 repeats (17 direct and 15 inverted repeats), *E. grandis* has 31 (16 direct and 15 inverted repeats), *N. tabacum* has 34 (18 direct and 16 inverted repeats), *G. barbadense* has 50 (31 direct and 19 inverted repeats) and *A. thaliana* has 56 (30 direct and 26 inverted repeats). Most of the repeated sequences in *S. cumini* are located in the intergenic regions and introns, while some were in the protein-coding genes including *psaA*, *psaB*, *ycf2*, *atpA* and *psbD* (Table 6). Detailed comparison of the *S. cumini* repeats with the chloroplast genomes of *E. globulus*, *E. grandis*, *N. tabacum*, *G. barbadense* and *A. thaliana* revealed that many of the repeats occur in identical positions (Fig. 6). The repeats found in the *ndhA* intron, *ycf3* intron, *ycf2* gene and *rps7-trnV-GAC* intergenic spacer showed sequence identity greater than 98 %. The *ycf2* gene repeat has multiple copies, i.e. 4 identical copies in *S. cumini*, 4 identical copies in *E. globulus* and *E. grandis*, 5 identical copies in *N. tabacum*, 11 identical copies in *G. barbadense* and 6 identical copies in *A. thaliana.* The *ycf2* gene repeat was also identified previously in adzukibean, soybean and *Medicago* (Perry et al. 2002). The *psaB* and *psaA* repeats were 98 % identical in *S. cumini*, *N. tabacum*, *E. grandis*, *E. globulus* and *G. barbadense* and show less homology (30 %) in *A. thaliana*. However, the repeats present in the *psbC-trnS-TGA*, *psbI-trnS-GCU* and *psaI-ycf4* intergenic spacers are identical in *S. cumini*, *N. tabacum* and *E. globulus*

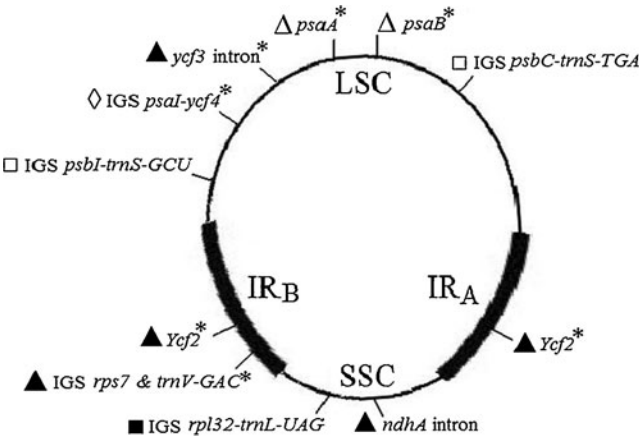

Fig. 6 Diagrammatic representation of repeat regions in *S. cumini*, *E. globulus*, *E. grandis*, *N. tabacum*, *G. barbadense* and *A. thaliana. Asterisk* multiple repeats with variable lengths (base pairs); *black triangles* repeats with high sequence homology in all analysed cp genomes; *white triangles* conserved repeats in *S. cumini*, *N. tabacum*, *E. globulus* and *G. barbadense* but low similarity to *A. thaliana*; *black squares* 98 % identical repeats in *S. cumini* and *E. globulus* and 70 % identical in *A. thaliana*; *white squares* 98 % identical repeats in *S. cumini*, *N. tabacum* and *E. globulus* while absent in *A. thaliana*; *white diamond* repeats with 98 % identity only in *S. cumini* and *E. globulus* while absent in other species

but are absent from *A. thaliana.* The longest repeats in *S. cumini*, other than the IR region, are found in the introns of the *ycf3* and *ndhA* genes (42 bp) and in the *psaA* and *psaB* genes (41 bp). The later repeat is also found in *E. globulus* and *E. grandis.* Two shorter 30-bp inverted repeats were found in a serine transfer RNA (*trnS-GGA*) gene. These repeats apparently play a role in the clover-leaf shape structure as analysed by the Vienna RNA web server (Hofacker 2003). These two repeats were present in all cpDNA sequences analysed.

Discussion

We sequenced *S. cumini* cpDNA followed by detailed bioinformatics analyses. The size of *S. cumini* cp genome was within the range of reported angiosperm chloroplast genomes (Tangphatsornruang et al. 2010). We compared *S. cumini* cpDNA sequence with previously reported partial sequences of *S. cumini* as present in GenBank, and found a range of sequence polymorphisms. As no other complete chloroplast cpDNA sequence from this species is currently available, these polymorphic characters might serve as intra-specific markers. They may also serve as starting points for investigations of polymorphisms in other *Syzygium* species. It is noted that American and Australian populations of *S. cumini* have identical cpDNA, while that from Pakistan is different. Due to the fact that cpDNA is maternally and clonally inherited, therefore, is it possible that the germplasm in Australia and America originated from the same population/region in Asia.

Comparison of the *S. cumini* chloroplast genome with a selection of other dicots suggests an expansion in the intergenic spacer located between the IRA/LSC border and the first gene of LSC region. This type of variation in the intergenic regions can be utilized in the development of species-specific vectors for chloroplast genetic engineering (Daniell et al. 2005). Gene order variations in chloroplast genomes are relatively uncommon. We confirmed the expected gene order in *S. cumini*, relative to *E. globulus*, *G. barbadense* and *A. thaliana.*

With the exception of the IR region, repeated sequences are usually rare in chloroplast genomes (Raubeson and Jansen 2005). The repeat analysis showed that the majority of the longer repeats (>30 bp) in *S. cumini*, *E. globulus*, *E. grandis*, *G. barbadense*, *A. thaliana* and *N. tabacum* chloroplast genomes are in the 30–40-bp length range. The repeats found in the coding and non-coding regions of *S. cumini* showed little variation. Several repeats in *S. cumini* were conserved in *E. globulus*, *E. grandis*, *N. tabacum* and *G. barbadense* suggesting these conserved repeats are widespread in angiosperm chloroplast genomes and might have functional roles.

Conclusion

The chloroplast genome structure and composition of *S. cumini* is similar to those reported for other dicots. However, compared to the selected dicots, this genome has an expansion in the intergenic spacer located between IRA/LSC border and the first gene of LSC region. The variation in this region can be useful for phylogeographic studies in *Syzygium* and other members of *Myrtaceae*. Several SNPs that resulted in non-synonymous mutations in maturase K and NADH-plastoquinone oxidoreductase subunit-5 were identified which could be useful as intra-specific markers to address whether lineage sorting from polymorphic ancestry has occurred within *Myrtaceae* family.

Acknowledgments The authors are thankful to Masahiro Sugiura and Yukawa Yasushi (Nagoya City University, Nagoya, Japan) for their suggestions on manuscript. Huma Asif, Asif Iqbal and Ishtiaq Ahmad Khan acknowledged the financial support of Higher Education Commission, Pakistan under Indigenous PhD Fellowship Program.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Mol Biol 215:403–410

Biffin E, Craven LA, Crisp MD, Gadek PA (2006) Molecular systematics of *Syzygium* and allied genera (*Myrtaceae*): evidence from the chloroplast genome. Taxon 55:79–94

Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24:861–862

Consolini AE, Baldinio AN, Amat AG (1999) Pharmacological basis for the empirical use of *Eugenia uniflora* L (*Myrtaceae*) as antihypertensive. J Ethnopharmacol 66:33–39

Craven LA, Biffin E (2005) *Anetholea anisata* transferred to, and two new Australian taxa of *Syzygium* (*Myrtaceae*). Blumea 50(1):157–162

Cui L, Veeraraghavan N, Richter A, Wall K, Jansen RK, Leebens-Mack J, Makalowska I, de Pamphilis CW (2006) ChloroplastDB: the chloroplast genome database. Nucleic Acids Res 34:D692–D696

Daniell H, Kumar S, Dufourmantel N (2005) Breakthrough in chloroplast genetic engineering of agronomically important crops. Trends Biotechnol 23:238–245

Dhingra A, Folta KM (2005) ASAP: amplification, sequencing and annotation of plastomes. BMC Genom 6:176

Foster SA, McKinnon GE, Steane DA, Potts BM, Vaillancourt RE (2007) Parallel evolution of dwarf ecotypes in the forest tree *Eucalyptus globulus*. New Phytol 175:370–380

Freeman JS, Marques CMP, Carocha V, Borralho NMG, Potts BM, Vaillancourt RE (2007) Origins and diversity of the Portuguese Landrace of *Eucalyptus globulus*. Ann For Sci 64(6):639–647

Grover JK, Vats V, Rathi SS (2000) Antihyperglycemic effect of *Eugenia jambolana* and *Tinospora cordifolia* in experimental diabetes and their effects on key metabolic enzymes involved in carbohydrate metabolism. J Ethnopharmacol 73:461–470

Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31(13):3429–3431

Ibrahim RI, Azuma J, Sakamoto M (2006) Complete nucleotide sequence of the cotton (*Gossypium barbadense* L) chloroplast genome with a comparative analysis of sequences among 9 dicots plants. Genes Genet Syst 81:311–321

Jadhay VM, Kamble SS, Kadam VJ (2009) Herbal medicine: *Syzygium cumini* (L.): a review. J Pharm Res 2(8):1212–1219

Jai P, Singh SP (2007) Mode of pollination, fruit set and fruit drop in Jamun (*Syzygium cumini* Skeels). Environ Ecol (Special 4): 1151–1153

Jansen RK, Raubeson LA, Boore LA, DePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol 395:348–384

Khan IA, Azim MK (2011) Variations in intergenic spacer rpl20-rps12 of mango (*Mangifera indica*) chloroplast DNA: implications in cultivar identification. Plant Syst Evol 292:249–255

Khan A, Khan IA, Heinze B, Azim MK (2012) The chloroplast genome sequence of date palm (*Phoenix dactylifera* L. cv. 'Aseel'). Plant Mol Biol Rep 30:666–678

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15(5):426–427

Lavin M, Mathews S, Hughes C (1991) Chloroplast DNA variation in *Gliricidia sepium* (*Leguminosae*): intraspecific phylogeny and tokogeny. Am J Bot 78:1576–1585

Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L (2004) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Res 14:472–477

Paiva JA, Prat E, Vautrin S, Santos MD, San-Clemente H, Brommonschenkel S, Fonseca PG, Grattapaglia D, Song X, Ammiraju JS, Kudrna D, Wing RA, Freitas AT, Bergès H, Grima-Pettenati J (2011) Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. BMC Genom 4(12):137

Pepato MT, Mori DM, Baviera AM, Harami JB, Vendramini RC, Brunett IL (2005) Fruit of the Jambolana tree (*Eugenia jambolana* Lam) and experimental diabetes. J Ethnopharmacol 96:43–48

Perry AS, Brennan S, Murphy DJ, Wolfe KH (2002) Evolutionary reorganisation of a large operon in Adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res 9:157–162

Porebski SL, Bailey G, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep 15(1):8–15

Rajan S, Markose BL (2007) Propagation of horticultural crops: vol. 6, pp. 74–75, Horticulture Science Series, New India Publishing, New Delhi, India

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants pages 45–68. In: Henry R (ed) Diversity and evolution of plants: genotypic and phenotypic variation in higher plants. CABI Publishing, London

Ross IA (2003) Medicinal plants of the world, Vol. 1, 2nd edn. Humana Press, Totowa, pp 445–451

Sager R, Ishida MR (1963) Chloroplast DNA in *Chlamydomonas*. Proc Natl Acad Sci USA 50:725–730

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6:283–290

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Steane DA (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum *Eucalyptus globulus* (*Myrtaceae*). DNA Res 12(3):215–220

Sugiura M (1995) The chloroplast genome. Essays Biochem 30:49–57

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

Teixera CC, Pinto LP, Kessler FHP, Knijnik L, Pinto CP, Gastaldo GJ, Fuchs FD (1997) The effect of *Syzygium cumini* (L.) skeels on post-prandial blood glucose levels in non-diabetic rats and rats with streptozotocin-induced diabetes mellitus. J Ethnopharmacol 56:209–213

Triboush SO, Danilenko NG, Davydenko OG (1998) A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. Plant Mol Biol Rep 16:183–189

Vaillancourt RE, Jackson HD (2000) A chloroplast DNA hypervariable region in eucalypts. Theor Appl Genet 101:473–477

van der Merwe MM, van Wyk AE, Botha AM (2005) Molecular phylogenetic analysis of *Eugenia* L. (*Myrtaceae*), with emphasis on southern African taxa. Plant Syst Evol 251(1):21–34

Wolfe KH, Morden CW, Palmer JD (1991) Ins and outs of plastid genome evolution. Curr Opin Genet Dev 4:523–529

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20(17):3252–3255