Genetica (2018) 146:101–113
https://doi.org/10.1007/s10709-017-0003-x

ORIGINAL PAPER

The complete chloroplast genome sequence of *Dodonaea viscosa*: comparative and phylogenetic analyses

Josphat K. Saina[1,2,3] · Andrew W. Gichira[1,2,3] · Zhi-Zhong Li[1,2] · Guang-Wan Hu[1,3] · Qing-Feng Wang[1,3] · Kuo Liao[1,3]

Received: 11 July 2017 / Accepted: 14 November 2017 / Published online: 23 November 2017

Abstract

The plant chloroplast (cp) genome is a highly conserved structure which is beneficial for evolution and systematic research. Currently, numerous complete cp genome sequences have been reported due to high throughput sequencing technology. However, there is no complete chloroplast genome of genus *Dodonaea* that has been reported before. To better understand the molecular basis of *Dodonaea viscosa* chloroplast, we used Illumina sequencing technology to sequence its complete genome. The whole length of the cp genome is 159,375 base pairs (bp), with a pair of inverted repeats (IRs) of 27,099 bp separated by a large single copy (LSC) 87,204 bp, and small single copy (SSC) 17,972 bp. The annotation analysis revealed a total of 115 unique genes of which 81 were protein coding, 30 tRNA, and four ribosomal RNA genes. Comparative genome analysis with other closely related Sapindaceae members showed conserved gene order in the inverted and single copy regions. Phylogenetic analysis clustered *D. viscosa* with other species of Sapindaceae with strong bootstrap support. Finally, a total of 249 SSRs were detected. Moreover, a comparison of the synonymous (Ks) and nonsynonymous (Ka) substitution rates in *D. viscosa* showed very low values. The availability of cp genome reported here provides a valuable genetic resource for comprehensive further studies in genetic variation, taxonomy and phylogenetic evolution of Sapindaceae family. In addition, SSR markers detected will be used in further phylogeographic and population structure studies of the species in this genus.

Keywords Chloroplast genome · *Dodonaea viscosa* · Phylogenetic analysis · Sapindaceae

Introduction

Consisting of woody and/or herbaceous shrubs, lianas and tropical trees, the family Sapindaceae has approximately 141 genera and 1900 species (Soltis et al. 2000; Li 2003; Harrington et al. 2005). This family is constituted of four major subfamilies i.e., Sapindoideae, Dodonaeoideae, Xanthoceroideae, and Hippocastanaceae (which includes Acereae and Hippocastaneae tribes), (Harrington et al. 2005; Buerki et al. 2009). *Dodonaea* (subfamily Dodonaeoideae) is the largest genus which consists 70 species with significant economical and ecological benefits (Wagner et al. 1999; Guerin et al. 2012). In particular, *Dodonaea viscosa* is a dense woody shrub with seven subspecies based on morphological and habitat characteristics (West 1984). *D. viscosa* has numerous medicinal properties which have led to increased research interests on its antimicrobial, (Getie et al. 2003; Thring et al. 2007), antibacterial, (Khurram et al. 2009), wound healing, (Habbu et al. 2007), antioxidant (Mothana et al. 2010), and anti-inflammatory (Alagarsamy et al. 2007; Salinas-Sánchez et al. 2012) properties.

The size of a typical chloroplast genome ranges between 72 and 217 kb and it consists of a small single-copy (SSC) and one large single-copy (LSC) separated by a pair of inverted repeats (IRs) (Sugiura 1992; Tangphatsornruang et al. 2010). The complete cp DNA has highly conserved gene order and is of significant interest to plant evolution

---

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10709-017-0003-x) contains supplementary material, which is available to authorized users.

✉ Qing-Feng Wang
qfwang@wbgcas.cn

✉ Kuo Liao
liaokuo@wbgcas.cn

[1] Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, Hubei, China

[2] University of Chinese Academy of Sciences, Beijing, China

[3] Sino-African Joint Research Center, Chinese Academy of Sciences, Wuhan, China

and systematic researchers (Wicke et al. 2011). Despite the next generation genome sequence assembly limitations, (Alkan et al. 2011) many chloroplast genome sequences have been recently documented. For instance, the use of complete cp genome in determining evolutionary patterns using genes from plastid genomes (Jansen et al. 2007), phylogenetic analysis of plastid genes (Moore et al. 2010) and the comparison of cp genomes representing angiosperm, gymnosperm, and fern families (Zhu et al. 2016) has been accomplished. In addition, there has been the successful usage of chloroplast DNA sequences to study the phylogenetics and phylogeography of angiosperms at lower taxonomic levels (Shaw et al. 2014).

Comparative cp genome analyses have been reported in several angiosperms e.g. in two *Dipteronia* species (Zhou et al. 2016), *Ampelopsis brevipedunculata* (Raman and Park 2016) and six *Rehmannia* species (Zeng et al. 2017). However, despite the rapidly advancing technology in genome sequencing, there is still limited Sapindaceae genomes available to date. Specifically, about eight complete chloroplast genomes have been reported across the Sapindaceae family. Previously, transcriptome sequencing, annotation, and polymorphism detection were performed by Christmas et al. (2015); while Harrington and Gadek (2009) analyzed the phylogeny of the species using nuclear ribosomal ITS and ETSf sequences. However, up to date there are no partial or complete sequences of chloroplast genome of the genus *Dodonaea*. Therefore, there is limited genetic knowledge regarding this genus. Our study aims were to unveil the cp genome of *D. viscosa* and compare it with eight other Sapindaceae cp genomes already reported, to detect the simple sequence repeats (SSRs), and to reconstruct its phylogenetic relationship with related species. The findings of this study will be valuable for further studies to understand the diversity, taxonomy, and phylogenetic relationships within Sapindaceae.

Methods

Plant materials and DNA isolation

Fresh leaves of *D. viscosa* were collected from Kenya (38°20′07″ E; 03°22′56″ S) and transported to Wuhan Botanical Garden, Chinese Academy of Sciences for further analyses. The duplicate voucher specimens were deposited at the East Africa herbarium (EA) and Wuhan Botanical Garden herbarium (HIB). The total genomic DNA was extracted from silica-dried leaves using modified cetyltrimethylammonium bromide (CTAB) method as described by Doyle (1991).

Genome assembly and annotation

A paired-end library was constructed using TruSeq DNA sample preparation kits according to the manufacturer's protocol (Illumina, San Diego, CA, USA), and then the whole genome sequencing was performed with an average read length of 150 bp pair-end using the Illumina Hiseq 2500 platform at NOVOgene company (Beijing, China). The PRINSEQ lite V0.20.4 (Schmieder and Edwards 2011); phredQ ≥ 20, Length ≥ 50 was used to filter raw reads and the plastid DNA extracted from the high-quality reads by mapping the available cp genome of *Sapindus mukorossi* (NC_025554) using a BLASTn (E-value: 10^{-6}). The retained reads were assembled by Velvet 1.2.10 (Zerbino and Birney 2008) with K-mer length of 99–119. Six sequence contigs (ranging from 7824 to 45,735 bp) were used for mapping against reference chloroplast genomes of (*S. mukorossi*, NC_025554; *Dipteronia sinensis*, NC_029338) using the software GENEIOUS R8.0.2 (Kearse et al. 2012).

Complete cp genome annotations, which included protein coding genes, tRNAs, and rRNAs was carried out using Dual Organellar GenoMe Annotator (DOGMA) according to default values as described by Wyman et al. (2004) with manual corrections for the start and stop codons, intron as well as exon boundaries. The positions of start and stop codons were further checked by BLAST homology searches. Similarly, the tRNA genes were verified by tRNAscan-SE 1.23 program (http://lowelab.ucsc.edu/tRNAscan-SE/, Schattner et al. 2005). A circular gene map was drawn in OGDraw v1.2 (Lohse et al. 2007). The fully annotated cp genome was deposited at the GenBank database (Accession no. MF155892).

Microsatellite detection analysis

SSR markers were detected using the Perl script tool Misa (Thiel et al. 2003) with minimal repeat numbers set at eight and five for mono- and di-nucleotides respectively and three for, tetra-, penta-, and hexa- nucleotides.

Comparative and phylogenetic analysis

The contraction/expansion regions of the inverted repeats (IRs) were compared among species in Sapindaceae family. Lastly, phylogenetic analysis was performed using maximum likelihood (ML) analysis in RAxML 8.0.20 following Stamatakis (2014) instructions. Subsequently, the best-fitting substitution model GTR + I + G model (p-inv = 0.47, and gamma shape = 0.93) were selected based on the Akaike information criterion as implemented in jModelTest 2.1.7 (Posada 2008). Finally, bootstrap analysis was performed with 1000 replications. The Bayesian inference (BI) analysis was performed by Mr Bayes version 3.2.6 (Ronquist and Huelsenbeck 2003) with default parameters. Bayesian analysis was performed for 20,000,000 generations with sampling every 5000 generations. The default setting was set for chain heating (temp = 0.2) and the first 10% of trees were removed as burn-in and the residual trees were combined to estimate posterior probabilities (PPs). Additionally, the genome rearrangement analyses of four Sapindaceae species with one outgroup species were performed using Mauve Alignment (Darling et al. 2004), and dot plots and gene identity plots between *D. viscosa* and *S. mukorossi* were created using PipMaker (Schwartz et al. 2000).

Synonymous (Ks) and non-synonymous (Ks) substitution rate analysis

The complete cp genomes of *D. viscosa* and its closely related species *S. mukorossi* were compared. We analyzed synonymous (Ks) and nonsynonymous (Ka) substitution rates and their ratio, Ka/Ks, using Model Averaging in the KaKs_Cal-culator program (Wang et al. 2011). 79 shared protein coding genes were aligned using Geneious Basic 5.6.4 (Kearseetal, 2012) to identify synonymous (Ks) and nonsynonymous (Ka) ratios using *S. mukorossi* as the reference in the alignment.

Results

Genome features

The cp genome of *D. viscosa* is a double-stranded DNA molecule with a length of 159,375 bp. Its LSC and SSC regions are 87,204 and 17,972 bp long, respectively, separated by a pair of inverted repeats (IRa and IRb) of 27,099 bp each. In addition, the overall GC content of this species was 37.9%, with the IR having a higher GC content probably due to the presence of all four Ribosomal RNA genes duplicated in these regions (42.8%) than SSC (32.0%) and the LSC (36.0%). The genome annotation analysis revealed 81 coded proteins, 30 tRNA and 4 ribosomal RNA genes (Table 1). A total of nine protein-coding genes including *rpl22, rps19, rpl2, rpl23, ycf2, ycf15, ndhB, rps7 and rps12*, had duplicates (Fig. 1). The LSC consisted of 60 protein-coding genes and 23 tRNAs, while the SSC comprised of 12 protein-coding genes and one tRNA with *ycf1* gene present within the IRa/SSC boundary. Nine protein coding genes, seven tRNAs, and all the rRNAs were repeated in the IR regions. Seventeen genes contained one intron while three genes harbored two (*ycf3, rps12* and *clpP*) (Table 1).

Cp genome comparison to other related Sapindaceae species

The cp genome of *D. viscosa* was similar to that of other Sapindaceae members in terms of gene content, and organization. A comparative analysis of the Sapindaceae family revealed that *D. viscosa* genome was 2,780 bp larger than the smallest genome. The total length of cp genome sequences of the eight species ranged from 156,595 to 160,481 bp. All species had a quadripartite structure with the length of the LSC region ranging between 85,227 and 87,204 bp, IR regions between 26,100 and 27,979 while the length of the SSC region varied between 17,972 and 18,873 bp (Table 2). The average GC content for all the available species was approximately 37.8%. The entire chloroplast genome structure, number, and order of genes among closely related species of *D. viscosa* appear to be well conserved (Table 2). Additionally, the alignment analysis revealed that the cp genomes of *A. davidii, D. sinensis, D. viscosa, B. sacra* (outgroup) and *S. mukorossi* were relatively conserved and no rearrangement occurred in the gene organization (Fig. 2). *D. viscosa* and *S. mukorossi* cp genome comparisons based on dot and identity plots analyses displayed conservation in terms of genome synteny (Fig. 3), however the percentage identities in the genomic sequences between the two species were relatively low in a region corresponding to *psaB* and *psaA* genes with > 50% identity, in contrast with other surrounding genic regions (Fig. 4).

Microsatellite detection analysis

In this study, a total of 249 SSRs repeat motifs were discovered in *D. viscosa* cp genome (Table 3). Among the SSRs identified, mononucleotide repeats were abundantly represented with 64.65%, followed by trinucleotide repeats with 29.72%. In addition, the di- and tetra-nucleotides occurred with less frequency of 2.81% each, while penta- and hexa-nucleotide repeats were not detected. The A/T repeat units were most common among the mononucleotides with proportions of 28.91 and 32.13%, respectively. Among the trinucleotides, AAG/CTT and AAT/ATT repeat motifs were the most abundant accounting for 12.05 and 8.83%, respectively, while only AG/CT and AT/AT dinucleotide motifs occurred with fewer frequencies accounting for 0.80 and 0.20%. The majority of tetranucleotide SSR motifs had AGAT/ATCT motif, followed by AAAG/CTTT, AAAT/ATTT, and AACT/AGTT motifs all of which had a similar low frequency (0.4%).

In all the genomes, mononucleotide repeats were the most abundant repeat type followed by the trinucleotides (Fig. 5). Penta- and hexa-nucleotides were not detected in most of the genomes and if present they occurred in low frequencies. In particular, penta- repeats were only observed in *S. mukorosi*,

Table 1 List of genes encoded by *Dodonaea viscosa* chloroplast genome

Category of genes	Groups of genes	Name of genes							
Self-replication	Ribosomal RNAs	*rrn16*	*rrn23*	*rrn4.5*	*rrn5*				
	Transfer RNAs	*trnA-UGC*[a]	*trnC-GCA*	*trnD-GUC*	*trnE-UUC*	*trnF-GAA*	*trnfM-CAU*	*trnG-UCC*	*trnH-GUG*
		trnI-CAU	*trnI-GAU*[a]	*trnK-UUU*[a]	*trnL-CAA*	*trnL-UAA*[a]	*trnL-UAG*	*trnM-CAU*	*trnN-GUU*
		trnS-GCU	*trnP-GGG*	*trnP-UGG*	*trnQ-UUG*	*trnR-ACG*	*trnR-UCU*	*trnW-CCA*	*trnY-GUA*
		trnS-GGA	*trnS-UGA*	*trnT-GGU*	*trnT-UGU*	*trnV-GAC*	*trnV-UAC*[a]		
	Small subunit of ribosome	*rps11*	*rps12*[b]	*rps14*	*rps15*	*rps16*	*rps18*	*rps19*[a]	*rps2*
		rps3	*rps4*	*rps7*[a]	*rps8*				
	Large subunit of ribosome	*rpl14*	*rpl16*	*rpl2*[a]	*rpl20*	*rpl22*[a]	*rpl23*[a]	*rpl32*	*rpl33*
		rpl36							
	DNA dependent RNA polymerase	*rpoA*	*rpoB*	*rpoC1*[a]	*rpoC2*				
Genes for photosynthesis	NADH dehydrogenase	*ndhA*[a]	*ndhB*[a]	*ndhC*	*ndhD*	*ndhE*	*ndhF*	*ndhG*	*ndhH*
		ndhI	*ndhJ*	*ndhK*					
	Photosystem 1	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*			
	photosystem ii	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*	*psbA*	*psbB*	*psbC*
		psbD	*psbE*	*psbF*	*psbH*	*psbI*	*psbJ*	*psbK*	*psbL*
		psbM	*psbN*	*psbT*	*psbZ*				
	Cytochrome b/f complex	*petB*	*petD*	*petG*	*petL*	*petN*			
	ATP synthase	*atpA*	*atpB*	*atpE*	*atpF*[a]	*atpH*	*atpI*		
	RubisCO large subunit	*rbcl*							
Other genes	Muterase	*matK*							
	Protease	*clpP*[b]							
	Envelope membrane protein	*cemA*							
	Subunit Acetyl-CoA-carboxylase	*accD*[a]							
	c-type cytochrome synthesis gene	*ccsA*							
Genes of uknown function	Open reading frames ycf	*ycf1*	*ycf15*[a]	*ycf2*	*ycf3*[b]	*ycf4*			

[a] Indicates genes containing one intron

[b] Genes containing two introns

while hexa repeats were present in both *Acer buergensis* and *S. mukorosi*. Moreover, all the cp genomes shared twelve SSR motifs (Fig. 5).

Comparison of IR and single-copy (SC) regions

The LSC/IRB/SSC/IRA region of *D. viscosa* was compared to the corresponding regions of its closely related cp genomes (Fig. 6). The duplicated *rpl22* gene was detected in the LSC/IRA and LSC/IRB junctions of seven genomes, but this gene was located in IRA region in *S. mukorossi* and LSC in *Acer miaotaiense* 754 and 302 bp respectively from LSC/IRA border. The *trnH* gene was found in the LSC region of all cp genomes, varying from 0 to 135 bp from the IRB/LSC border region. Complete duplication of *rps19* gene was observed in the IRA and IRB regions with 155–427 bp apart from the LSC/IRA and LSC/IRB boundary regions, however this gene was located at the LSC/IRA region in *A. miaotaiense* (Fig. 6). The *ycf1* gene located in IRB/SSC region had the size variation of 5477 (*D. dyeriana*) to 5642 (*S. mukorossi*) bp in all cp genomes. The *ycf1* pseudo gene crossed the IRA/SSC region in six genomes, whereas this gene was located in IRA region in *D. dyeriana, A. morrisonense* and *A. davidii*. Additionally, at the IRA/SSC junction the *ycf1* pseudo gene and *ndhF* genes were overlapped in all the eight genomes except in *D. dyeriana*, moreover this

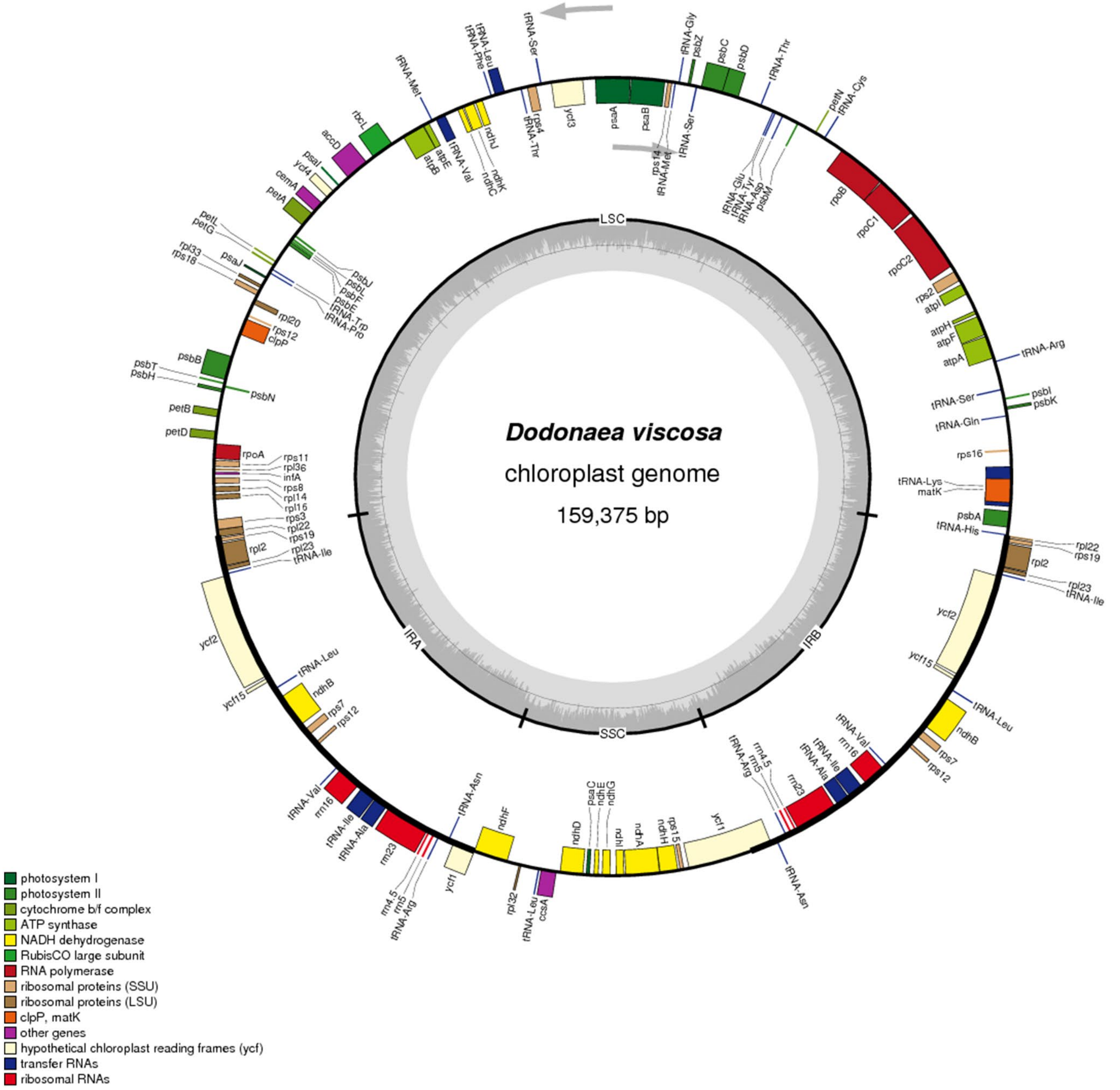


Fig. 1 Gene map of the *Dodonaea viscosa* chloroplast genome. Genes lying outside of the circle are transcribed clockwise, whereas genes inside the circle are transcribed counterclockwise. The colored bars indicate different functional groups. The dark gray area in the inner circle corresponds to GC content while the light gray corresponds to the AT content of the genome. *IR* inverted repeat, *LSC* large single copy, *SSC* small single copy are indicated

gene was shifted to the SSC region in *S. mukorossi, A. griseum* and *D. dyeriana* with 29, 17 and 4 bp gap respectively (Fig. 6).

Phylogenetic analysis

Phylogenetic analysis was performed using a dataset of the whole chloroplast complete sequences from eleven species inclusive of *D. viscosa* and two species as outgroup species (Fig. 7). The reconstructed phylogenetic tree group all the sapindales into two groups. *D. viscosa* and *S. mukorossi* were clustered together with the strong bootstrap value of 100%. All the *Acer* species and *Dipteronia* species formed another group with strong bootstrap support.

Table 2 Comparison of chloroplast genomes of Sapindaceae species

Species	GenBank accession	Genome (bp)	GC content (%)	Large single copy region	(LSC) GC content (%)	Short single copy region	(SSC) GC content (%)	Inverted repeat region	(IR) GC content (%)	Protein Coding Genes	tRNAs	rRNAs
Acer griseum	NC_034346	156,857	37.9	85,227	36.0	18,134	32.4	26,768	42.8	78	31	4
Acer davidii	NC_030331	157,044	37.9	85,410	36.1	18,112	32.3	26,761	42.7	78	31	4
Acer miaotaiense	NC_030343	156,595	37.9	86,327	36.1	18,068	32.2	26,100	42.6	80	30	4
Acer buergerianum	KF_753631	156,911	37.9	85,315	36.0	18,094	32.2	26,751	42.8	83	31	4
Acer morrisonense	NC_029371	157,197	37.8	85,655	36.0	18,086	32.1	26,728	42.7	78	31	4
Dipteronia sinensis	NC_029338	157,080	37.8	85,455	35.9	18,093	32.1	26,766	42.7	87	40	4
Dipteronia dyeriana	NC_031899	157,071	38.0	85,529	36.1	18,082	32.5	26,730	42.8	87	40	4
Sapindus mukorossi	NC_025554	160,481	37.7	85,650	35.8	18,873	31.5	27,979	42.6	78	30	4
Dodonaea viscosa	MF_155892	159,375	37.9	87,204	36.0	17,972	32.0	27,099	42.8	81	31	4
Boswellia sacra	NC_029420	160,543	37.6	88,055	35.6	18,962	32.2	26,763	42.8	83	27	4

Synonymous (Ks) and non-synonymous (Ks) substitution rate analysis

To estimate selection pressures of Chloroplast genes, a total of 79 protein coding genes in the cp genome of *D.viscosa* were used to compute synonymous and non-synonymous rates against *S. mukorossi* (Table S1). Genes with Ka/Ks ratios of not applicable (NA) were changed to 0 (Table S1). In this study most of the genes except *rps14* had the Ka/Ks ratio of significantly less than 1 in all the shared protein coding genes (Fig. 8). A comparison among the genes in each functional group showed that the substitution rates fluctuated widely among the 79 coding genes, with Ka and Ks ranging from 0 to 1.0209 and 0 to 1.0365 (Table S1), respectively. The genes related to photosystem I had the highest synonymous rate, with an average of 0.2338, while those related to the ATP synthase gene showed the lowest average rate of 0.1257. Genes associated with the small subunit of ribosome showed the highest non-synonymous (Ka) average rate of 0.1065, while those associated with the photosystem II contained the lowest average value of 0.0112. More remarkably, based on the Ka/Ks values 67 genes indicated purifying selection, while one gene *rps14* located in the LSC region had Ka/Ks ratio above 1.0, which indicates a positive selection, and no changes were observed in the Ks synonymous/nonsynonymous rate in *rps19, rps7, psaC, psbD, psbL, psbM, psbN, petN, atpH, rpl23*, and *psbI* genes.

Discussion

The length of *D .viscosa* cp genome reported here (157,375 bp) is similar but 1106 bp smaller than that of closely related species *S. mukorossi* (160,481) (Yang et al. 2016). Moreover, the genome organization, gene order and GC content of the cp genome are similar and conserved. The *D. viscosa* cp genome is AT-rich which is consistent with other species from Sapindaceae family, for example, *Acer miaotaiense* (62.12%) (Zhang et al. 2016), *A. davidii* (62.10%) (Jia et al. 2016) and other species in order Sapindales like *Citrus sinensis* (61.52%) (Bausher et al. 2006) and *Phellodendron amurense* (61.60%) (Chen et al. 2017) with the overall A + T content of 62.14%.

Among the 81 protein-coding genes, unusual start codons were observed in three genes (*rps19, rps12* and *ndhD*) which were GTG, ACT, and ACG, respectively. Previous studies have reported that due to RNA editing the translation initiation codons ACG and GUG may be restored to the standard start codon AUG (Kuroda et al. 2007; Takenaka et al. 2013), hence the same process might have occurred to the reported genes. The presence of several SSR motifs in the chloroplast genome provides useful sources for designing primers which could be used for phylogeography and population structure

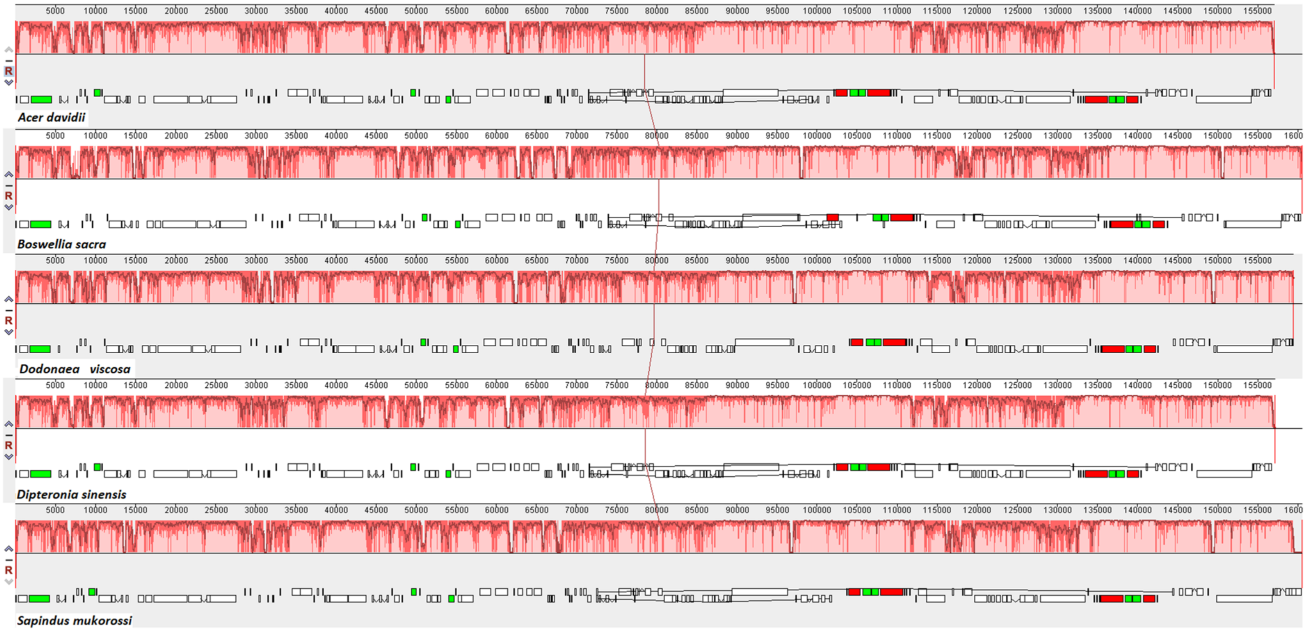


Fig. 2 MAUVE alignment of four Sapindaceae species chloroplast genomes and *Boswellia sacra* (outgroup). Within each of the alignments, local collinear blocks are represented by blocks of the same color connected by lines

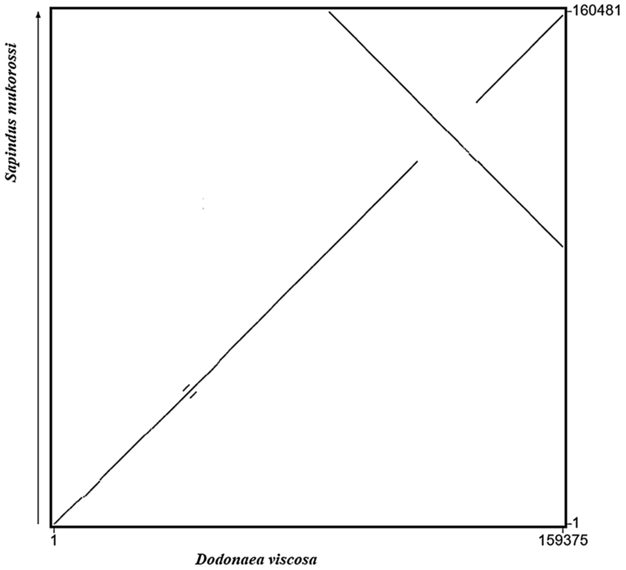


Fig. 3 Dot plot of chloroplast genome sequences shared by *D. viscosa* and *S. mukorossi*

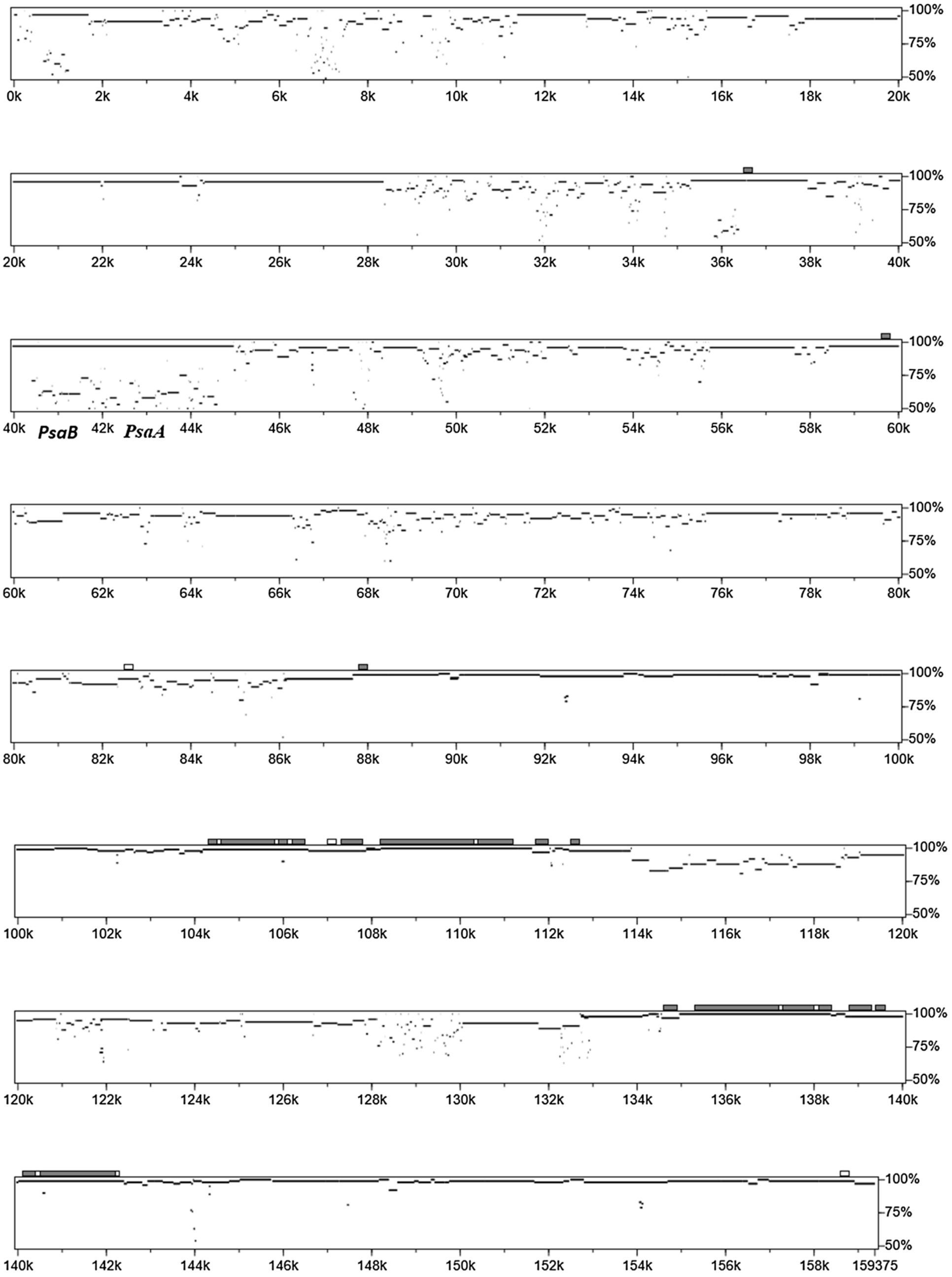
at the species level or provide useful genomic information for further population genetics and phylogenetic relationship in this genus.

The variable IR and the single-copy (SC) regions have been considered as a main mechanism causing the length variation of angiosperm cp genomes (Kim and Lee 2004). In this study nine species (Sapindaceae) were used to compare the exact IR border positions and their adjacent genes (Fig. 6). The LSC, IR and SSC regions of *D. viscosa* was found to have a slight size differences compared to other closely related genera of same family *S. mukorossi*. The IRb/SSC extended into *ycf1* gene leading to formation of *ycf1* pseudo gene at the IRa/SSC boundary with lengths of 951–1400 bp, however this gene was located at the IRa region in *A. davidii, A. morrisonense*, and *D. dyeriana*. In addition, the *ndhF* gene overlapped with the *ycf1* pseudo gene in most of the cp genomes but not in *S. mukorossi*, and *D. dyeriana*.

As shown in (Fig. 6) minor shifts occurred in IR/LSC borders for example the *rpl22* crossed the LSC/IRA border in *D. viscosa, A. griseum, A. davidii, A. buergerianum, A. morrisonense* and the two *Dipteronia* species, with the pseudo fragment duplicated at the IRb/LSC region however, it was located at IRa region in *S. mukorossi* and LSC in *A. miaotaiense*. Remarkably, the SSC region of *D. viscosa* was smaller than those of other species while the *trnH-GUG* sequence was located in LSC region of all the genomes. Recent studies have shown that the *rps19* gene is generally located in the IR regions in other angiosperm cp genomes (Li et al. 2017), our results also showed that this gene is located in IR region of all the eight species except in *A. miaotaiense*. Besides, the duplication of the genes is possibly due to IR region expansion which have been reported in other Sapindales (Bausher et al. 2006).

The phylogenetic tree showed a close relationship between *D. viscosa* and *S. mukorossi*, and formed a strong

Fig. 4 The percentage identities in gene identity plots of the genomic regions in *D. viscosa* and *S. mukorossi*



homogenous group with all the *Acer* and *Dipteronia* species (subfamily aceroideae) with maximum (100%) bootstrap value (Fig. 7). This observation is in line with previous studies which clustered all the *Acer* taxa forming a monophyletic clade (Zhang et al. 2016; Wang et al. 2017). Similarly, *Dipteronia* species formed a close relationship with *S. mukorossi* and *D. viscosa* as verified in our results with a strong bootstrap value (Fig. 7). Based on Bayesian inference the ML bootstrap supports (BS) the BI posterior probabilities (PP) strongly supported all the nodes, with values 100% and 0.9 respectively. All the nodes in the tree had high bootstrap support (Data not shown).

The selection pressure analysis showed that most genes exhibited a value less than 1, indicating the efficiency of purifying selection (Fig. 8). Recent studies (Machado et al. 2017; Raman and Park 2016) revealed that the Ka/Ks ratio favored purifying selection in many genes, this is consistent with our study. The average Ks values between the two closely related species were 0.1770, 0.0525 and 0.2191 for the LSC, IR and SSC regions respectively with an average Ks of 0.1662 across all regions (Table S1). Higher Ks values were observed in most of the genes across all the regions, with some genes revealing lower Ks values including *ycf3, rps11, rps4, rps2, rpl36, rpl33, clpP,atpF, psaA, petN,rbcL,*

Table 3 Simple sequence repeats (SSRs) in the *Dodonaea viscosa* chloroplast genome

Sequences	Number of repeats																Total
	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
A/T	–	–	–	–	–	58	43	24	12	5	6	2	–	1	–	1	152
C/G	–	–	–	–	–	4	4	1	–	–	–	–	–	–	–	–	9
AG/CT	–	–	2	–	–	–	–	–	–	–	–	–	–	–	–	–	2
AT/AT	–	–	4	1	–	–	–	–	–	–	–	–	–	–	–	–	5
AAC/GTT	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4
AAG/CTT	25	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	30
AAT/ATT	16	5	1	–	–	–	–	–	–	–	–	–	–	–	–	–	22
ACC/GGT	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
ACT/AGT	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
AGC/CTG	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5
AGG/CCT	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
ATC/ATG	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4
AAAG/CTTT	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1
AAAT/ATTT	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1
AACT/AGTT	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1
AGAT/ATCT	4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4
Total																	249

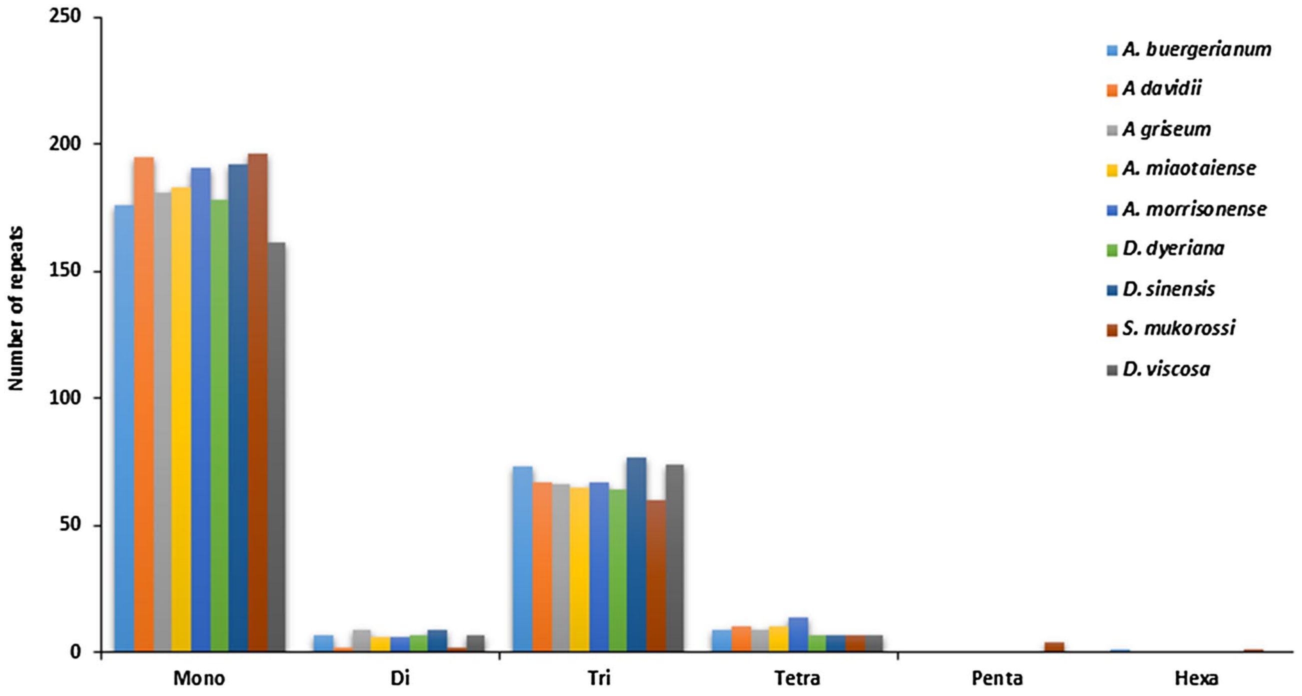


Fig. 5 Simple sequence repeats (SSRs) in the nine Sapindaceae chloroplast genomes

psbT,psbL, psbJ,psbI, psbF, psbE (LSC), only gene *rps15* (SSC) and all the genes in IR region except *ycf1* and *rpl22* (Table S1). The distribution of Ks indicated that most genes in SSC region have experienced higher selection pressures than the other cp genome regions, whereas the IR region is more conserved. Fu et al. (2016) found higher evolution rate in SSC compared to the LSC and IR, this is in agreement with the results in this study. Based on region comparison IR region is more conserved than the LSC and SSC regions. This is consistent with earlier reports that showed IR region

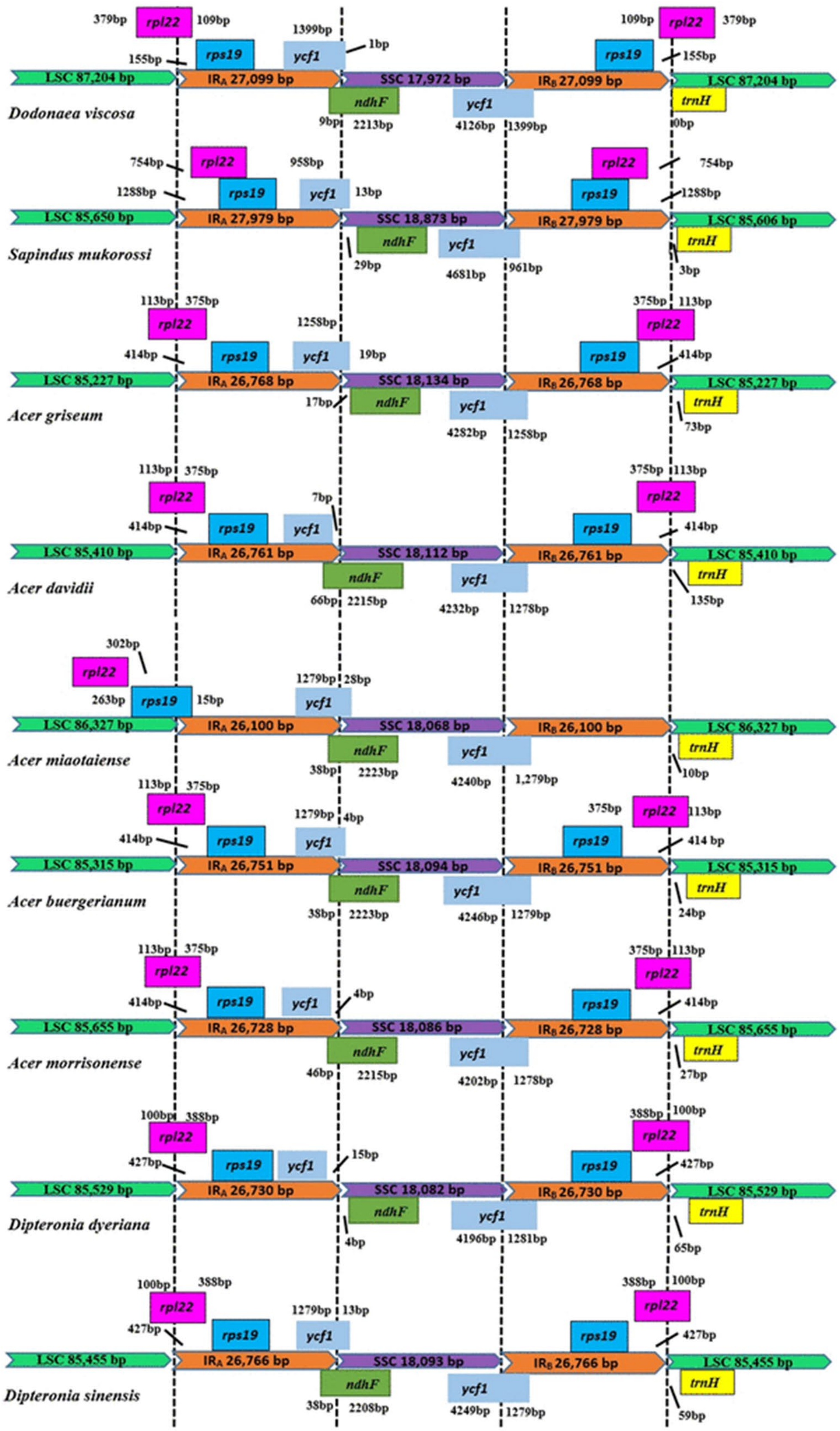


Fig. 6 Comparison of IR, LSC and SSC border regions among nine Sapindaceae cp genomes

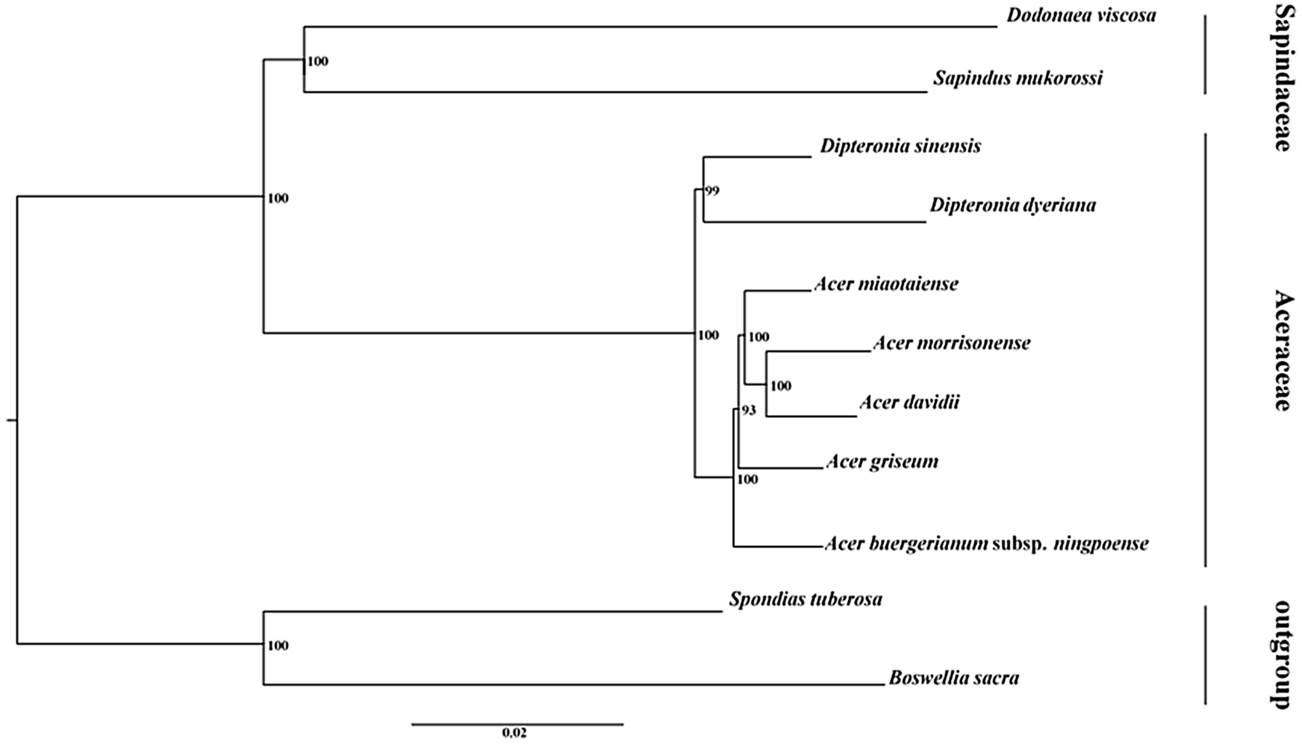


Fig. 7 Phylogenetic relationships based on complete chloroplast genome sequences of nine Sapindaceae species and two Outgroup species (*Spondias tuberosa* and *Boswellia sacra*) with maximum likelihood

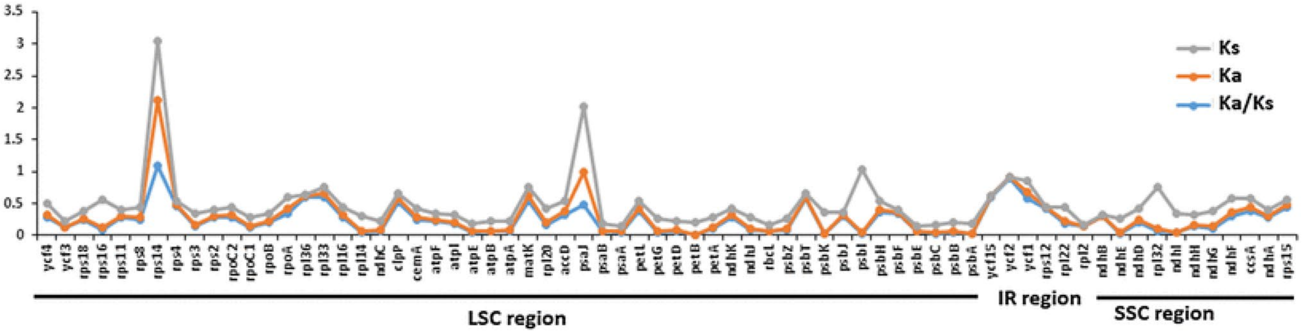


Fig. 8 Comparison of nonsynonymous (Ka) and synonymous (Ks) substitution rates and Ka/Ks ratio between *D. viscosa* and *S. mukorossi*

diverged at a slower rate compared to the LSC and SSC regions. (Cho et al. 2015; Fu et al. 2016).

This study has generated the complete cp genome sequence of *D. viscosa*, the first cp genome under subfamily Dodonaeoideae. Comparative analyses among species of Sapindaceae revealed that the available cp genomes of species under this family are well conserved in terms of overall structure. The phylogenetic analysis showed that *D. viscosa* and *S. mukorossi* are closely related. This novel *D. viscosa* cp genome could be a valuable resource that will aid in filling the gap in knowledge of genome evolution in Dodonoideae subfamily and related species.

Acknowledgements The authors thanks Jinming Chen for useful comments on the manuscript. This work was supported by the Sino-Africa Joint Research Center (#Y323771W07 and #SAJC201322) and Special Funds for the Young Scholars of Taxonomy of Chinese Academy of Sciences Grants to Liao Kuo (Grant no. ZSBR-013).

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

Alagarsamy V, Venket Narayanan R, Thangathirupathy A et al (2007) Anti-inflammatory activity of *Dodonaea viscosa* Linn leaf extracts. INDIAN DRUGS-BOMBAY 44:559

Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65. https://doi.org/10.1038/nmeth.1527

Bausher MG, Singh ND, Lee SB et al (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

Buerki S, Forest F, Acevedo-Rodríguez P et al (2009) Plastid and nuclear DNA markers reveal intricate relationships at sub familial and tribal levels in the soapberry family (Sapindaceae). Mol Phylogenet Evol 51:238–258. https://doi.org/10.1016/j.ympev.2009.01.012

Chen KK (2017) Characterization of the complete chloroplast genome of the Tertiary relict tree *Phellodendron amurense* (Sapindales: Rutaceae) using Illumina sequencing technology. Conserv Genetics Resour. https://doi.org/10.1007/s12686-017-0741-1

Cho KS, Yun BK, Yoon YH et al (2015) Complete chloroplast genome sequence of tartary buckwheat (*Fagopyrum tataricum*) and comparative analysis with common buckwheat (F. esculentum). PloS One, 10:e0125332. https://doi.org/10.1371/journal.pone.0125332

Christmas MJ, Biffin E, Lowe AJ (2015) Transcriptome sequencing, annotation and polymorphism detection in the hop bush, *Dodonaea viscosa*. BMC Genomics 16:1. https://doi.org/10.1186/s12864-015-1987-1

Darling AC, Mau B, Blattner FR et al (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403. https://doi.org/10.1101/gr.2289704

Doyle JJ (1991) DNA protocols for plants. Molecular techniques in taxonomy. Springer. https://doi.org/10.1007/978-3-642-83962-7_18

Fu PC, Zhang YZ, Geng HM et al (2016) The complete chloroplast genome sequence of *Gentiana lawrencei* var. farreri (Gentianaceae) and comparative analysis with its congeneric species. PeerJ 4:e2540. https://doi.org/10.7717/peerj.2540

Getie M, Gebre-Mariam T, Rietz R et al (2003) Evaluation of the anti-microbial and anti-inflammatory activities of the medicinal plants *Dodonaea viscosa, Rumex nervosus* and *Rumex abyssinicus*. Fitoterapia 74:139–143. https://doi.org/10.1016/S0367-326X(02)00315-5

Guerin GR, Wen H, Lowe AJ (2012) Leaf morphology shift linked to climate change. Biol Lett 8:882–886. https://doi.org/10.1098/rsbl.2012.0458

Habbu PV, Joshi H, Patil B (2007) Potential wound healers from plant origin. Pharmacognosy Rev 1:271

Harrington MG, Gadek PA (2009) A species well-travelled—the *Dodonaea viscosa* (Sapindaceae) complex based on phylogenetic analyses of nuclear ribosomal ITS and ETSf sequences. J Biogeogr 36:2313–2323. https://doi.org/10.1111/j.1365-2699.2009.02176.x

Harrington MG, Edwards KJ, Johnson SA et al (2005) Phylogenetic inference in Sapindaceae sensu lato using plastid *mat*K and *rbc*L DNA sequences. Syst Bot 30:366–382. https://doi.org/10.1600/0363644054223549

Jansen RK, Cai Z, Raubeson LA et al (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

Jia Y, Yang J, He YL et al (2016) Characterization of the whole chloroplast genome sequence of *Acer davidii* Franch (Aceraceae). Conserv Genet Resour 8:141–143. https://doi.org/10.1007/s12686-016-0530-2

Kearse M, Moir R, Wilson A et al (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

Khurram M, Khan MA, Hameed A et al (2009) Antibacterial activities of *Dodonaea* viscosa using contact bioautography technique. Molecules 14:1332–1341. https://doi.org/10.3390/molecules14031332

Kim KJ, LEE HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res 11:247–261. https://doi.org/10.1093/dnares/11.4.247

Kuroda H, Suzuki H, Kusumegi T et al (2007) Translation of *psbC* mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine–Dalgarno sequence in tobacco chloroplasts. Plant Cell Physiol 48:1374–1378. https://doi.org/10.1093/pcp/pcm097

Li A (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. https://doi.org/10.1046/j.1095-8339.2003.t01-1-00158.x

Li Z, Long H, Zhang L et al (2017) The complete chloroplast genome sequence of tung tree (*Vernicia fordii*): Organization and phylogenetic relationships with other angiosperms. Sci Rep. https://doi.org/10.1038/s41598-017-02076-6

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

Machado LO, Vieira LN, Stefenon VM et al (2017) Phylogenomic relationship of feijoa (*Acca sellowiana* (O. Berg) Burret) with other Myrtaceae based on complete chloroplast genome sequences. Genetica 145:163–174. https://doi.org/10.1007/s10709-017-9954-1

Moore MJ, Soltis PS, Bell CD et al (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci 107:4623–4628. https://doi.org/10.1073/pnas.0907801107

Mothana RA, Abdo SA, Hasson S et al (2010) Antimicrobial, antioxidant and cytotoxic activities and phytochemical screening of some yemeni medicinal plants. Evid Based Complement Alternat Med 7:323–330. https://doi.org/10.1093/ecam/nen004

Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256. https://doi.org/10.1093/molbev/msn083

Raman G, PARK S (2016) The complete chloroplast genome sequence of Ampelopsis: gene organization, comparative analysis, and phylogenetic relationships to other angiosperms. Front Plant Sci. https://doi.org/10.3389/fpls.2016.00341

Salinas-Sánchez DO, Herrera-Ruiz M, Pérez S et al (2012) Anti-inflammatory activity of hautriwaic acid isolated from *Dodonaea viscosa* leaves. Molecules 17:4292–4299. https://doi.org/10.3390/molecules17044292

Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686–W689. https://doi.org/10.1093/nar/gki366

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864. https://doi.org/10.1093/bioinformatics/btr026

Schwartz S, Zhang Z, Frazer KA et al (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10:577–586. https://doi.org/10.1101/gr.10.4.577

Shaw J, Shafer HL, Leonard OR et al (2014) Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. Am J Bot 101:1987–2004. https://doi.org/10.3732/ajb.1400398

Soltis DE, Soltis PS, Chase MW et al (2000) Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atp*B sequences. Bot J Linn Soc 133:381–461. https://doi.org/10.1111/j.1095-8339.2000.tb01588.x

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sugiura M (1992) The chloroplast genome. Plant Mol Biol 19:149–168. https://doi.org/10.1007/978-94-011-2656-4_10

Takenaka M, Zehrmann A, Verbitskiy D et al (2013) RNA editing in plants and its evolution. Annu Rev Genet 47:335–352. https://doi.org/10.1146/annurev-genet-111212-133519

Tangphatsornruang S, Sangsrakru D, Chanprasert J et al (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

Thiel T, Michalek W, Varshney R et al (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

Thring T, Springfield E, Weitz F (2007) Antimicrobial activities of four plant species from the Southern Overberg region of South Africa. Afr J Biotech 6:1779

Wagner WL, Herbst D, Sohmer S (1999) Manual of the flowering plants of Hawaii. University of Hawaii Press, Bishop Museum Press, Honolulu

Wang D, Liu F, Wang L et al (2011) Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. Biol Direct 6:13. https://doi.org/10.1186/1745-6150-6-13

Wang WC, Chen SY, Zhang XZ (2017) The complete chloroplast genome of the endangered Chinese paperbark maple, *Acer griseum* (Sapindaceae). Conserv Genet Resour 1–3. https://doi.org/10.1007/s12686-017-0715-3

West J (1984) A revision of Dodonaea Miller (Sapindaceae) in Australia. Aust Syst Bot 7:1–194. https://doi.org/10.1071/BRU9840001

Wicke S, Schneeweiss GM, Müller KF et al (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273–297. https://doi.org/10.1007/s11103-011-9762-4

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. doi.https://doi.org/10.1093/bioinformatics/bth352

Yang B, Li M, Ma J et al (2016) The complete chloroplast genome sequence of *Sapindus mukorossi*. Mitochondrial. DNA Part A 27:1825–1826. https://doi.org/10.3109/19401736.2014.971243

Zeng S, Zhou T, Han K et al (2017) The complete chloroplast genome sequences of six *Rehmannia* Species. Genes 8:103. https://doi.org/10.3390/genes8030103

Zerbino DR, 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

Zhang Y, Li B, Chen H et al (2016) Characterization of the complete chloroplast genome of *Acer miaotaiense* (Sapindales: Aceraceae), a rare and vulnerable tree species endemic to China. Conserv Genet Resour 8:383–385. https://doi.org/10.1007/s12686-016-0564-5

Zhou T, Chen C, Wei Y et al (2016) Comparative transcriptome and chloroplast genome analyses of two related *Dipteronia* Species. Front Plant Sci. https://doi.org/10.3389/fpls.2016.01512

Zhu A, Guo W, Gupta S et al (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