The complete chloroplast genome sequence of *Morus cathayana* and *Morus multicaulis*, and comparative analysis within genus *Morus* L

Wei Qing Kong and Jin Hong Yang

Shaanxi Key Laboratory of Sericulture, Ankang University, Ankang, Shaanxi, China

ABSTRACT

Trees in the Morus genera belong to the Moraceae family. To better understand the species status of genus *Morus* and to provide information for studies on evolutionary biology within the genus, the complete chloroplast (cp) genomes of *M. cathayana* and *M. multicaulis* were sequenced. The plastomes of the two species are 159,265 bp and 159,103 bp, respectively, with corresponding 83 and 82 simple sequence repeats (SSRs). Similar to the SSRs of *M. mongolica* and *M. indica* cp genomes, more than 70% are mononucleotides, ten are in coding regions, and one exhibits nucleotide content polymorphism. Results for codon usage and relative synonymous codon usage show a strong bias towards NNA and NNT codons in the two cp genomes. Analysis of a plot of the effective number of codons (ENc) for five *Morus* spp. cp genomes showed that most genes follow the standard curve, but several genes have ENc values below the expected curve. The results indicate that both natural selection and mutational bias have contributed to the codon bias. Ten highly variable regions were identified among the five *Morus* spp. cp genomes, and 154 single-nucleotide polymorphism mutation events were accurately located in the gene coding region.

Subjects Genomics, Plant Science

Keywords *Morus cathayana*, *Morus multicaulis*, Mutation, Chloroplast genome, Codon usage

Submitted 8 July 2016
Accepted 27 January 2017
Published 8 March 2017

Corresponding author
Wei Qing Kong,
weiqing_kongwq@126.com

Academic editor
Abhishek Kumar

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj.3037**

INTRODUCTION

Mulberry (genus *Morus*, family Moraceae) is widely distributed in Asia, Europe, North and South America, and Africa. The trees are cultivated in East, Central, and South Asia for domesticated silkworm and silk production, which is of economic importance. Some mulberry species are also valued for their hard wood, delicious fruit, bark for paper production, and multiple uses in traditional oriental medicine (*Asano et al., 1994*; *Kim et al., 1999*). Linnaeus classified the Moraceae family into order Urticales in 1753 using morphological characteristics. However, the Moraceae family has been repositioned into order Rosales on the basis of the phylogenetic relationships among some nuclear genes and chloroplast (cp) loci or genomes (*Kong & Yang, 2016*; *Su et al., 2014*; *Zhang et al., 2011*). *Morus* is a genus of trees in family Moraceae, and species classification within *Morus* is the subject of ongoing controversy. In the Annals of Mulberry Varieties in China (*The Sericultural Research Institute Chinese Academy of Agricultural Sciences, 1993*), more than

3,000 mulberry germplasm resources were classified into 15 species and four varieties. In the *Flora of China* (*Zhou & Gilbert Michael, 2003*), *Morus* is classified into two groups and 11 species.

With the development of sequencing technology in recent years, in addition to nuclear genome sequences, cp genes (*matK* and *rbcl*), gene spacer regions (*trnH–psbA*, *atpB–rbcL*, *trnC–ycf6*, and *trnL*–F), and cp genome information have been used to study plant molecular systematics (*CBOL Plant Working Group, 2009*; *China Plant BOL Group et al., 2011*; *Kress & Erickson, 2007*; *Roy et al., 2010*). A nuclear gene internal transcribed spacer (ITS) and three cp DNA fragments in a total of four sequences are generally recommended for investigation of near-edge species. In *Morus*, *Zhao et al. (2005)* used the ITS and *trnL*-F sequence to cluster 13 mulberry genotypes, representing nine species and three varieties, into five branches. There are no other reports of molecular systematics applicable to the differentiation and classification of *Morus* species.

Chloroplasts are photosynthetic organelles that occur widely in algae and plants. Although the cp genome is more conserved than the nuclear genome, many mutation events such as indels, substitutions, and inversions have been identified in cp DNA sequences (*Ingvarsson, Ribstein & Taylor, 2003*). As a DNA polymorphism index, indels and single-nucleotide polymorphisms (SNPs) can be used at a low taxonomic level. The DNA polymorphism rate between cp genomes was 0.15% for *M. yunnanensis* and *M. balansae*, 0.3% for *M. indica* and *M. mongolica* (*Kong & Yang, 2016*; *Song et al., 2015*), and 0.2% among four different Chinese ginseng strains (*Zhao et al., 2015*). These results indicate the presence of variable characteristics among the cp genomes of different species.

Complete cp sequences can provide insight into evolution and natural selection, and have been used in significant contributions concerning evolutionary mechanisms for species and chloroplasts (*Asheesh & Vinay, 2012*). To date, cp genomes have been reported for three *Morus* species: *M. mongolica*, *M. indica*, and *M. notabilis* (*Chen et al., 2016*; *Kong & Yang, 2016*; *Ravi et al., 2006*). However, there has been no research on codon biology or genome evolution. Here, we report the cp genomes for another two *Morus* species (*M. cathayana* and *M. multicaulis*) and compare the codon usage bias (CUB), sequence divergence, and mutation events among the five *Morus* spp. cp genomes (Table 1).

MATERIALS & METHODS

Sample collection, DNA extraction, and sequencing

Samples of *M. cathayana* were collected from the Qinba Mountain Area and *M. multicaulis* was acquired from Shandong Sericultural Research Institute. Both plants used in the study are now maintained in the mulberry field of Shaanxi Key Laboratory of Sericulture, Ankang University, China. The cp DNA was extracted from 10 g of fresh leaves using a modified high-salt method (*Shi et al., 2012*) and treated according to a standard procedure (*Bartram et al., 2011*) for sequencing on an Illumina Hiseq 2000 platform (2 * 125 bp).

Chloroplast genome assembly and annotation

Both reads were assembled using SOAP *de novo* software (*Luo et al., 2012*) with the Kmer and maximal read length set to 60 and 50, respectively. The resulted contigs were aligned to

Table 1 Summary of the features of *Morus* chloroplast genomes.

Species	GenBank No.	Genome size/GC content	LSC size/GC content	SSC size /GC content	IR size/GC content
M. cathayana	KU981118	159,265/36.16	88,143/33.77	19,844/29.20	25,639/42.95
M. multicaulis	KU981119	159,103/36.19	87,940/33.82	19,809/29.26	25,677/42.91
M. indica	NC_008359	158,484/36.37	87,386/34.12	19,742/29.35	25,678/42.92
M. mongolica	KM491711	158,459/36.29	87,367/33.97	19,736/29.33	25,678/42.92
M. notabilis	NC_027110	158,680/36.36	87,470/34.11	19,776/29.34	25,717/42.89

the *M. mongolica* cp genome as the reference genome. Ambiguous sequences were manually trimmed. The mean fold coverage of the final assembled *M. cathayana* and *M. multicaulis* cpDNA reached approximately 690- and 770-fold, respectively. Transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and protein-coding genes (PCGs) in the two cp genomes were annotated using Dual Organellar GenoMe Annotator (DOGMA) software (*Wyman, Jansen & Boore, 2004*). The tRNAs and their corresponding structures were further verified and predicted using the tRNAscan-SE 1.21 program (*Schattner, Brooks & Lowe, 2005*). The physical maps were drawn using the web tool Organellar Genome DRAW (OGDraw) v1.2 (*Lohse, Drechsel & Bock, 2007*).

Analysis of simple sequence repeats (SSRs) and long repeat sequences

SSRs in *M. cathayana* and *M. multicaulis* cp genomes were detected using MISA (*Thiel et al., 2003*) with the minimal repeat number set to 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. The distribution, nucleotide composition, and polymorphism among SSRs were investigated. For long repeat sequences in five *Morus* species, the program REPuter was used to assess the number and location of all four type of forward match (F), reverse match (R), complement match (C) and palindromic match (P) (*Kurtz et al., 2001*). The identity of the repeat was limited to greater than 90% and the size was no less than 25 bp, respectively.

Indices of codon usage

The amino acid composition and relative synonymous codon usage (RSCU) values for *M. cathayana* and *M. multicaulis* cp genomes were calculated using Mega 5.05 (*Tamura et al., 2011*). Then, the GC content of the first, second, and third codon positions (GC1, GC2, and GC3, respectively) and the overall GC content of the coding regions (GCc) were calculated manually. GC3S is the GC content at the third synonymously variable coding position excluding Met, Trp, and three stop codons. The CodonW program on the Mobyle server (http://www.mobyle.fr) was used to calculate the GC3s and the effective number of codons (ENc). An ENc plot (ENc vs GC3s) was also analyzed.

Sequence divergence and mutation events analysis

A sliding window analysis was conducted using DnaSP 5.0 software (*Librado & Rozas, 2009*) for comparative analysis of the sequence divergence (Pi) among the cp genomes of *M. cathayana*, *M. multicaulis*, *M. mongolica*, *M. indica*, and *M. notabilis*. The window length was 600 bp and the step size was 200 bp.

To identify differences in coding sequences and corresponding amino acids among the five *Morus* cp genomes, pairwise alignment of the nucleotide sequence in coding region of five sequences was performed using Clustal X 1.83 (Supplemental Information 1). SNP events and transition (Ts) or transversion (Tv) in the plastomes were counted and positioned. In addition, SNPs were further classified into synonymous (S) and non-synonymous (N) substitutions using Mega 5.05 (*Tamura et al., 2011*). Then, a Z test ($P < 0.05$) was carried out by bootstrap method (1,000 replicates). The ancestral states were inferred using the ML method. The gene classification was according to *Chang et al. (2006)*.

RESULTS & DISCUSSION

Assembly and features of cp genomes

The complete cp genomes of *M. cathayana* and *M. multicaulis* are both closed circular molecules of 159,265 and 159,103 bp (GenBank accession number: KU981118, KU981119), respectively (Fig. 1). Both cp genomes show the typical quadripartite structure of most angiosperms, and comprise a pair of IRs (25,639 bp for *M. cathayana* and 25,677 bp for *M. multicaulis*) separated by the LSC (88,143 bp for *M. cathayana* and 87,940 bp for *M. multicaulis*) and SSC (19,844 bp for *M. cathayana* and 19,809 bp for *M. multicaulis*) regions. The GC content of *M. cathayana* and *M. multicaulis* cp genomes is 36.16% and 36.19%, respectively. Similar to values reported for Rosaceae cp genomes (*Su et al., 2014*; *Wang, Shi & Gao, 2013*), the GC content of the five *Morus* spp. cp genomes is 36.16–36.37%, with uneven distribution within the cp genome: the GC content is highest in the IR (42.89–42.95%), intermediate in the LSC (33.77–34.12%), and lowest in the SSC region (29.20–29.35%) (Table 1).

The *M. cathayana* and *M. multicaulis* cp genomes both encode 132 predicted functional genes, of which 112 are unique genes, including 78 protein-coding genes (PCGs), 30 tRNA and four rRNA genes with 63,873 bp, 2,208 and 4,524 bp, respectively; 18 were duplicated in the IR region, including seven PCGs and seven tRNA and all rRNA genes (Fig. 1). Fifteen genes (10 PCGs and five tRNA genes) contain one intron, and three PCGs (*clpP*, *ycf3*, and *rps12*) have two introns. As found in most other green terrestrial plants, the maturase K (*matK*) gene in the *M. cathayana* and *M. multicaulis* cp genomes is located within the *trnK* intron; and *rps12* is a trans-spliced gene; the 5′-end exon is in the LSC region and the other two reside in the IR region separated by an intron. The four rRNA genes and two tRNA genes of *trnI* and *trnA* are clustered as 16S–*trnI*–*trnA*–23S–4.5S–5S in the IR region, the same as in the cp genome of *M. mongolica* and *M. indica* (*Kong & Yang, 2016*; *Ravi et al., 2006*) and most of other plants (*Mardanov et al., 2008*; *Wang, Shi & Gao, 2013*; *Wu et al., 2014*).

Analyses of repetitive sequences

A total of 83 SSR loci, accounting for 973 bp, and 82 SSR loci, representing 949 bp in length, were detected in *M. cathayana* and *M. multicaulis* cp genomes (Table 2). Among these, there are 59 mono-, 7 di-, 3 tri-, 11 tetra-, and 3 penta-nucleotide repeats, respectively, in the *M. cathayana* cp genome; the corresponding numbers of these repeats in *M. multicaulis* are 63, 6, 2, 9, and 2. Mono-nucleotide repeats accounted for 71.1% and 76.8% of total SSRs in *M. cathayana* and *M. multicaulis*, respectively, similar to the levels found for *M.*

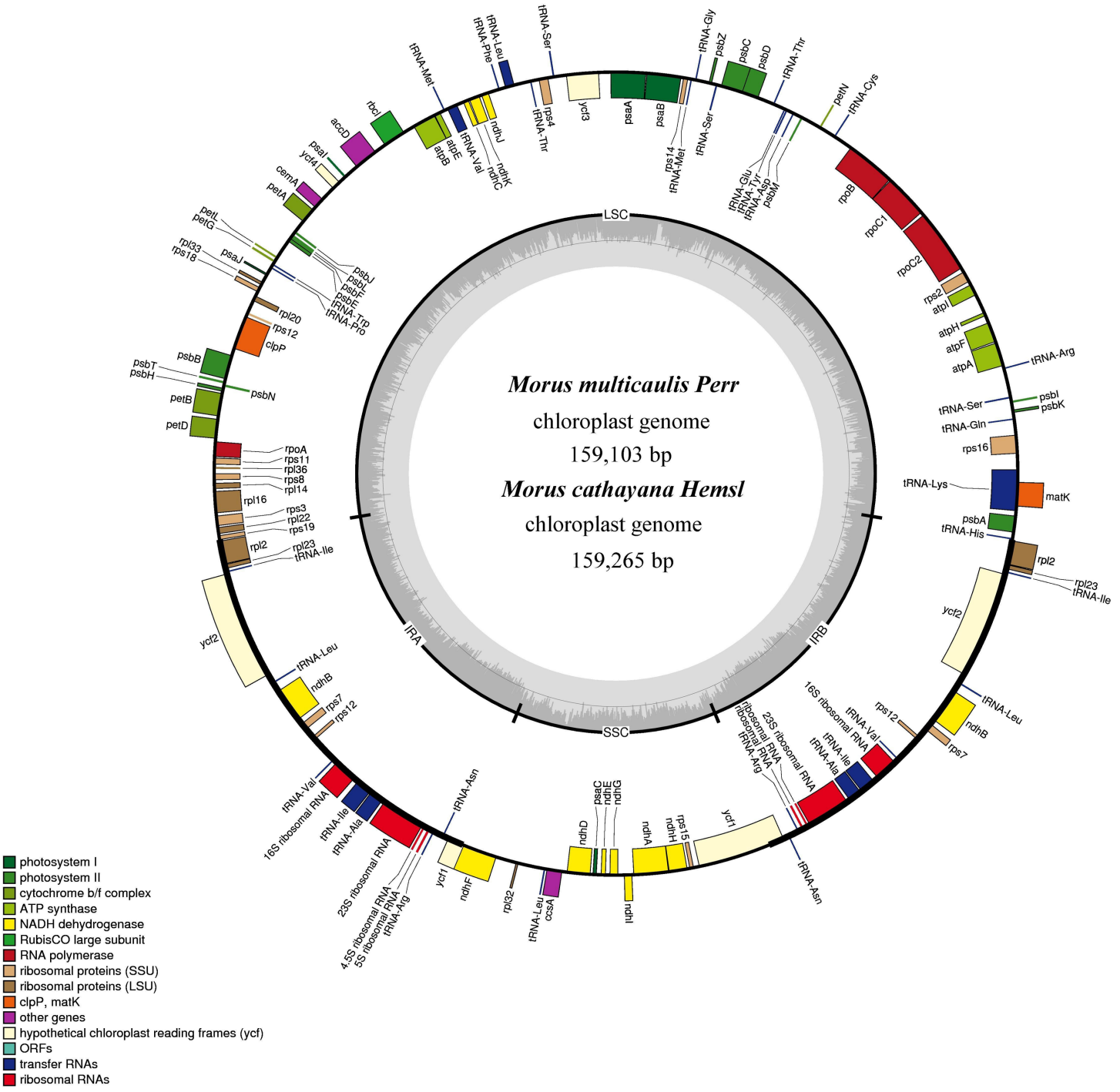

Figure 1 Gene map for *M. cathayana* and *M. multicaulis* plastomes. Genes lying outside the circle are transcribed in a clockwise direction, whereas genes inside are transcribed in a counterclockwise direction. Different colors denote known functional groups. The GC and AT contents of the genome are denoted by dashed darker and lighter gray in the inner circle. LSC, SSC, and IR indicate large single-copy, small single-copy, and inverted repeat regions, respectively.

Table 2 SSRs in *M. cathayana* and *M. multicaulis* chloroplast genomes.

SSR	Size	NUMBER	LOCI
a	10	8	**3950, 5052,590,29073, 50058, 68954, 68987, 114495(*ndhF*)**
	11	4	2057, 38564,54282^{a12}, **63153, 87796,** 116614^{a10}
	12	2	13584^{a13}, **84902**
	13	1	128314^{a14}
	14	1	**74502**
	15	1	9565^{a11}
	16	2	4799^{a13} 8984^{a15}
t	10	17	**5231, 9782,** 14830**,24363, 30678, 30944, 54323,** 55220, **57416(*atpB*), 62926, 67243, 69081, 71234, 74300,** 83303**, 117117,122510, 30637(*ycf1*), 132394(*ycf1*)**
	11	8	**479,** 8575^{t10}**,34221,** 57867^{t12}**,59882,** 75017^{t13}**,79022,131496(*ycf1*)**
	12	8	**12693,** 13275^{t13}, 14182^{t10}**,27623(*rpoB*),** 68830^{t13}, 69893^{t11}**,72813,86135**
	13	3	9207^{t14}, 52130^{t12}**,128735**
	14	1	64181^{t13}
	16	1	81690^{t13}
	17	1	**49756**
	18(20)	1	87252^{t14}, 116789
at	5	2	69153,**116007(*ndhF*)**
	6	1	**10796**
ta	6	3	**5495, 21235(*rpoC2*), 119074**
tc	6	1	**64908(*cemA*)**
aat	4	1	**128715**
ttc	4	1	**71261**
tat	4	1	50114
aaag	3	1	**135481**
aaat	3	3	**24057,** 38611, **47006**
atta	3	2	**33937, 116796**
attt	3	2	**14173, 62456**
tatt	3	1	**24394**
tctt	3	1	**111916**
ttat	3	1	118221
aagga	3	1	**14008(*atpF*)**
atata	3	1	117818
atttc	3	1	24297[tttct3]

Notes.

Parentheses, containing coding regions; boxed type, absent in *M. cathayana* but present in *M. multicaulis*; bolded type, absent in *M. multicaulis* but present in *M. cathayana*; underline and superscript, nucleotide length polymorphism; bracket, nucleotide content polymorphism, others are identical.

Table 3 Long repeat sequences in the chloroplast genome of five *Morus* species.

Type	Repeat size(bp)					Location	Region
	M. cathayana	*M. multicaulis*	*M. mongolica*	*M. indica*	*M. notabilis*		
P	51	–	–	–	–	IGS(*rpl2-trnH*; *rps19-rpl22*)	LSC
F	39	39	39	39	39	Intron(*ycf3*); IGS(*rps7-trnV*)	LSC; IRB
P	39	39	39	39	39	Intron(*ycf3*); IGS(*trnV-rps7*)	LSC; IRA
F	–	–	–	32	32	IGS(*trnT-trnL*)	LSC
F	–	31	31	31	31	IGS(*trnE-trnT*)	LSC
F	–	–	–	–	31	Intron(*clpP*)	LSC
P	30	30	30	30	30	*trnS*; *trnS*	LSC
P		30	–	30	–	Intron(*ndhA*)	SSC
R	28	–	26	26	27	IGS(*rps3-rpl22*)	LSC
P	28	28	28	–	28	IGS(*ccsA-ndhD*)	SSC
F	27	–	–	–	–	IGS(*rps7-trnV*)	IRA(IRB)
P	27	–	–	–	–	IGS(*rps7-trnV*; *trnV-rps7*)	IRA; IRB
P	26	26	26	26	26	IGS(*trnH-psbA*)	LSC
F	–	25	26	–	–	IGS(*ycf4-cemA*; *petD-rpoA*)	LSC

mongolica and *M. indica* cp genomes, and the study on lower and higher plants (*George et al., 2015*). Six SSRs, including one di-nucleotide of $(TC)_6$, one tri-nucleotide of $(TTC)_4$, two tetra-nucleotides of $(AAAG)_3$ and $(TTCT)_3$, and two penta-nucleotides of $(AAGGA)_3$ and $(TTTCT)_3$ in *M. multicaulis* or $(ATTTC)_3$ in *M. cathayana*, contain at least one C or G nucleotide, and the SSRs have a high AT content (97.5%). Most of the repeats are located in noncoding regions—i.e., intergenic spacers and introns—except for 10, which were found in seven coding genes of *ycf1*(3), *ndhF*(2), *rpoC2*, *rpoB*, *atpB*, *atpF* and *cemA*. The 10 SSRs in coding genes also occur in the *M. indica* and *M. mongolica* cp genomes. Comparison between *M. cathayana* and *M. multicaulis* revealed that 57 loci are identical, 19 exhibit length polymorphisms, six SSRs loci only exist in *M. cathayana*, and five SSRs loci only exist in *M. multicaulis*. One nucleotide content polymorphic of $(TTTCT)_3$ and $(ATTTC)_3$ exhibits in *M. multicaulis* and *M. cathayana*, respectively (Table 2).

A total of 14 long repeat sequences that are longer than 25 bp were identified in the five *Morus* spp. cp genomes, including one reverse, six forward and seven palindromic matches. Most of these repeats are located in intergenic spacers or introns, except for one 30 bp palindromic repeat, which is in the *trnS* genes (Table 3). Among all the long repeats, only four was detected in all the five cp genomes, which means that these repeats might create a diversity of cp genomes, and provide valuable information for phylogeny of genus *Morus* (*Yang et al., 2014*).

Codon usage patterns in *M. cathayana* and *M. multicaulis* cp genomes, and comparison with other three *Morus* spp.

As an important indicator of CUB, the RSCU value is the frequency observed for a codon divided by the expected frequency (*Sharp & Li, 1986*). RSCU is close to 1.0 when all synonymous codons are used equally without any bias; RSCU is greater or less than 1 when synonymous codons are more or less frequent than expected (*Gupta, Bhattacharyya*

Table 4 GC content of coding regions in the chloroplast genome of five *Morus* species.

Organism	CDS	Codons	GCc%	GC1%	GC2%	GC3%	GC3s%	ENc
M. cathayana	85	26599	37.29	45.13	37.42	29.3	26.28	49.28
M. multicaulis	85	26599	37.3	45.13	37.44	29.3	26.28	49.27
M. indica	84	26301	37.32	45.17	37.54	29.25	26.23	49.24
M. mongolica	84	26544	37.33	45.19	37.56	29.24	26.23	49.22
M. notabilis	84	26301	37.32	45.11	37.55	29.31	26.29	49.25

Notes.
CDS, Coding sequence; GCc, GC content in coding regions; GC1, GC2, GC3, GC3S, GC content at first, second, third, and synonymous third codon positions, respectively; ENc, effective number of codons.

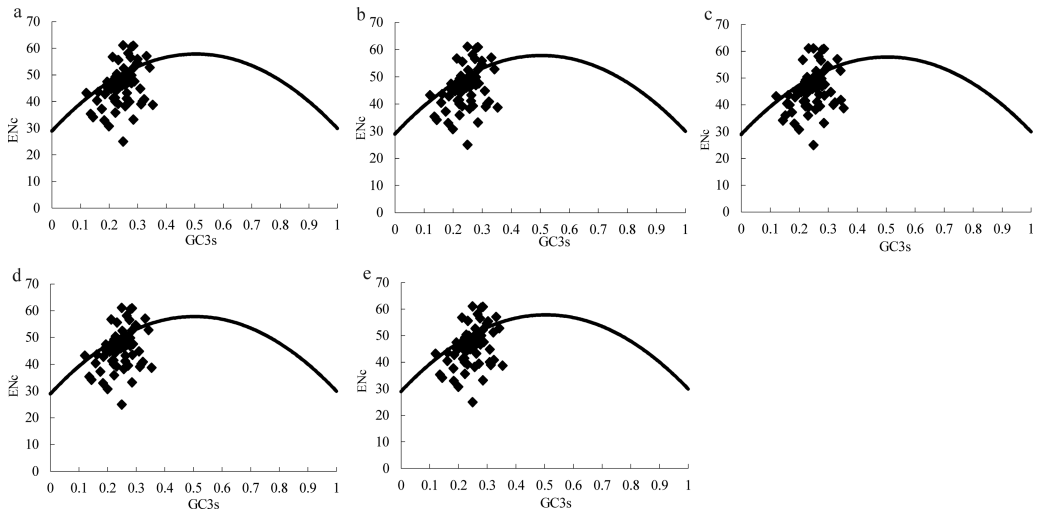

Figure 2 **ENc plots for the chloroplast genome of five *Morus* species.** Solid lines are expected ENc from GC3. (A) *M. cathayana*; (B) *M. multicaulis*; (C) *M. indica*; (D) *M. mongolica*; (E) *M. notabilis*.

& Ghosh, 2004). Codons ending with A and T have RSCU > 1 for *M. cathayana* and *M. multicaulis* cp genomes, indicating that they are used more frequently than synonymous codons, and may play major roles in the A+T bias of the entire cp genome. In terms of nucleotide composition, the most frequent codons are composed of T or A or their combination (i.e., ATT for Ile, AAA for Lys, AAT for Asn, and TTT for Phe); the least frequent codons have a high GC content (i.e., UGC for Cys, CGC and CGG for Arg, ACG for Thr, GCG for Ala, and CCG and CCC for Pro) (Table S1). Further analysis of the base composition at each codon position in the coding region revealed that the average GC content for GC1, GC2, and GC3 is lower than the AT content in the five *Morus* spp. cp genomes, and that the GC3 content is lower than the GC2 and GC1 content significantly (Table 4). The regular occurrence of A/T in the third codon position supports the previous finding that mutation towards A+T is a strong driving force for the cp genome.

The effective number of codons (ENc) is widely used as a measure of CUB, with range in 20–61, for genes with extreme bias using only one codon per amino acid or no bias using synonymous codons equally (*Wright, 1990*). GC3s is indicator of the level of nucleotide composition bias (*Ahmad et al., 2013*; *Wright, 1990*), and ENc plots (ENc vs $GC3_S$) are a

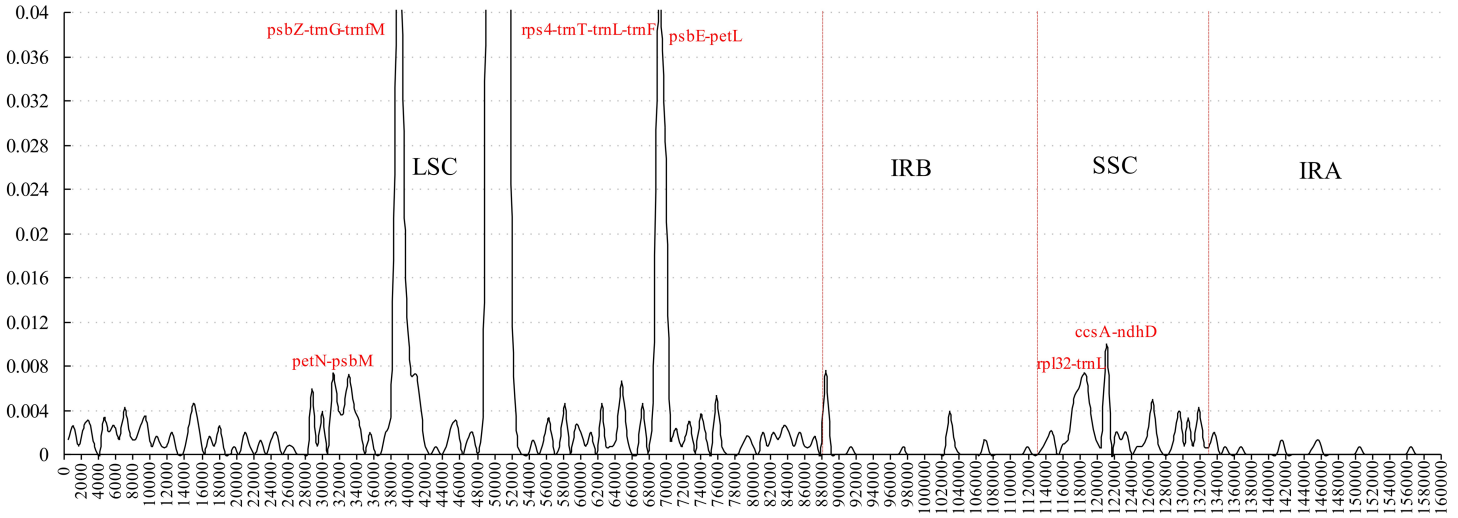

Figure 3 Sliding window analysis of the whole plastome for five *Morus* species (window length 600 bp, step size 200 bp). *X*-axis: position of the midpoint of a window; *Y*-axis: nucleotide diversity of each window.

Table 5 Transitions (Ts) and transversions (Tv) in the chloroplast genome of five *Morus* species.

	Ts		Tv			
	a < – >g	t < – >c	a < – >t	a < – >c	t < – >g	g < – >c
M. notabilis	13	22	4	9	18	4
M. indica	16	28	2	8	6	1
M. mongolica	–	–	–	1	–	1
M. cathayana	4	2	–	2	1	1
M. multicaulis	7	3	1	–	–	–
sum	40	55	7	20	25	7

useful indicator of the factors affecting codon usage, as well as the force of mutation or other factors. Predicted values lie on or just below the expected curve when a gene's codon usage is constrained only by the G+C mutation bias, and values are considerably below the curve when codon usage is subject to selection for codon optimization (*Wright, 1990*). Similar to the ENc plot for the Asteraceae family (*Nie et al., 2013*), the majority of genes among the five *Morus* spp. plastomes follow the standard curve, which indicates that the codon bias was mainly caused by nucleotide composition bias at the third position (GC3s) (Table 4, Fig. 2). In addition, there are several genes lying below the expected curve; this suggests that in addition to mutational bias, codon usage variation among genes can also be influenced by selection force (*Sablok et al., 2011*; *Wright, 1990*).

Sequence divergence among five *Morus* spp. cp genomes

Comparative analysis of sequence divergence among the five *Morus* spp. cp genomes revealed nucleotide variability (*Pi*) values in the range 0–0.28533 with average of 0.00432, indicating that the differences among the five cp genomes is small. However, nine intergenic regions (*trnT–trnL*, *petN–psbM*, *psbZ–trnG*, *trnG–trnfM*, *rps4–trnT*, *psbE–petL*, *rpl32–trnL*,

Table 6 Synonymous (S) and non-synonymous (N) substitutions in chloroplast coding genes for five *Morus* species.

	Gene	*M. notabilis*		*M. mongolica*		*M. indica*		*M. cathayana*		*M. multicaulis*	
		S	N	S	N	S	N	S	N	S	N
Photosynthetic apparatus	*psaA*	1				1					
	psaB	1	1			21 $^{-}$	2		1		
	psaI									2	
	psbA								3	1	
	psbC	1									
	psbD					2					
	psbE					1					
	psbK		1								
	psbZ					1	2				
	petD					1					
	petG					1					
	petL						2				
	ycf4		1								
	Total	**3**	**3**	**0**	**0**	**28**	**6**	**0**	**4**	**3**	**0**
Photosynthetic metabolism	*atpA*	2									
	atpB					2	2			1	
	atpE	1								1	
	ndhA						1				1
	ndhC									1	
	ndhD	1	2				1				
	ndhE								1		
	ndhF	1	3	1			1	1			1
	ndhI	1									
	ndhJ		1								
	rbcL		6 $^{+}$							1	
	Total	**6**	**12**	**1**	**0**	**2**	**5**	**1**	**1**	**4**	**2**
Gene expression	*rpl14*	1	1								
	rpl16	1									
	rpl20	1				1					
	rpl22	1									
	rpoA	1							1		
	rpoB	1					1				
	rpoC1		2								
	rpoC2	2	3								
	rps2	1									
	rps4					2					
	rps8		2			2					
	rps11		1			1					
	rps14					3	5				
	rps15		1								

(*continued on next page*)

Table 6 (*continued*)

	Gene	*M. notabilis*		*M. mongolica*		*M. indica*		*M. cathayana*		*M. multicaulis*	
		S	N	S	N	S	N	S	N	S	N
	rps16	1									
	rps19		1			2					
	Total	**10**	**11**	**0**	**0**	**11**	**6**	**0**	**1**	**0**	**0**
Other genes	*accD*		1	1					1		
	cemA		1								
	ccsA	1									
	matK	2	3			1				1	
	ycf1	5	11 +				2		2		1
	ycf2		1								
	Total	8	17	1	0	1	2	0	3	1	1
	Sum	**27**	**43**	**2**	**0**	**42**	**19**	**1**	**9**	**8**	**3**

Notes.
+, positive selection; −, negative selection.

trnL–trnF and *ccsA–ndhD*) and one intron (trnL) are highly variable, with much higher values ($Pi > 0.008$) than other regions (Fig. 3). All the 10 loci are in single-copy regions, not IR regions. These regions with highly variable loci are not random but are clustered in 'hot spots' (*Shaw et al., 2007*; *Worberg et al., 2007*). It has been reported that *rpl32–trnL* is particularly highly variable in *Machilus* and *Solanum* plastomes, and in spermatophyte (*Dong et al., 2012*; *Dong et al., 2015*; *Sarkinen & George, 2013*; *Song et al., 2015*), and that *trnL–trnF*, *trnT–trnL*, *rps4–trnT*, and the *trnL* intron are highly variable between two species of genus *Morus* L. (*Kong & Yang, 2016*). The *petN–psbM* intergenic region was reported as part of the *trnC–trnD* intergenic region, and has been used in lower-level phylogenetic studies in flowering plants (*Lee & Wen, 2004*). The *petN–psbM* region and the *clpP* intron were successfully used in a study of phylogenetic relationships in peach species (*Dong et al., 2012*). In the present study, four rarely reported highly variable loci, *psbZ–trnG*, *trnG–trnfM*, *psbE–petL*, and *ccsA–ndhD*, were found in *Morus* plastomes. Phylogenetic studies at the species level in *Morus* have not been carried out using highly variable regions, and our analysis provides a basis for further phylogenetic study of *Morus* using these highly variable regions.

Numbers and pattern of SNP mutations

As the most abundant type of mutation, we investigated SNPs in gene coding regions of five *Morus* spp. cp genomes. There were 154 SNPs detected, including 95 Ts, 59 Tv, and 74 N and 80 S substitutions (Tables 5 and 6). Among the Tv, 45 are GC content changes, 14 are between A and T, and between G and C (Table 5). The Kn, Ks and their ratio are indicators of the rates of evolution and natural selection (*Yang & Nielsen, 2000*). Relative to the reference ancestral states, 74 non-synonymous substitutions was observed in 28 of 79 protein-coding regions in five *Morus* spp. (Table 6). The photosynthetic apparatus gene *psaB* of *M. indica* share extra synonymous substitution sites, while *ycf1*, photosynthetic metabolism gene *rbcL* of *M. notabilis* and gene expression gene *rps14* of *M. indica* has more non-synonymous than synonymous substitution sites, suggesting a relatively high evolution

rate for these genes, as for *M. yunnanensis* and *M. balansae* (*Song et al., 2015*). A *Z* test can be used to detect positive selection by comparing the relative abundance of synonymous and non-synonymous substitutions within the gene sequences, the results of the *Z* test for each gene showed that *rbcL* and *ycf1* genes of *M. notabilis* exist positive selection, and *psaB* of *M. indica* exists negative selection, which means that the genes of *Morus* chloroplast might exert different evolutionary pressures. For all of these substitutions, 74.03% and 25.32% of the SNP sites are in LSC and SSC regions, respectively, while only one SNP site is in *M. indica* IR region, consistent with the more conservative characteristics of IR compared to LSC and SSC regions (*Dong et al., 2014*; *Jo et al., 2011*).

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Youth Science and Technology New Star Program of Shaanxi Province, China (No. 2013KJXX-96). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare there are no competing interests

Author Contributions

- Wei Qing Kong and Jin Hong Yang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

GenBank accession number: KU981118, KU981119.

Data Availability

The following information was supplied regarding data availability:

BioProject ID: PRJNA326788; PRJNA326736

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3037#supplemental-information.

REFERENCES

Ahmad T, Sablok G, Tatarinova TV, Xu Q, Deng XX, Guo WW. 2013. Evaluation of codon biology in citrus and *Poncirus trifoliate* based on genomic features and frame corrected expressed sequence tag. *DNA Research* **20**:135–150 DOI 10.1093/dnares/dss039.

Asano N, Tomioka E, Kizu H, Matsui K. 1994. Sugars with nitrogen in the ring isolated from the leaves of *Morus bombycis*. *Carbohydrate Research* **253**:235–245 DOI 10.1016/0008-6215(94)80068-5.

Asheesh S, Vinay S. 2012. *Evolutionary analysis of plants using chloroplast*. German: LAP Lambert Academic Publishing.

Bartram AK, Lynch MD, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Applied and Environmental Microbiology* **77**:3846–3852 DOI 10.1128/AEM.02772-10.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* **106(31)**:12794–12797 DOI 10.1073/pnas.0905845106.

Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM. 2006. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Molecular Biology and Evolution* **23**:279–291 DOI 10.1093/molbev/msj029.

Chen C, Zhou W, Huang Y, Wang Z. 2016. The complete chloroplast genome sequence of the mulberry *Morus notabilis* (Moreae). *Mitochondrial DNA* **27**:2856–2857 DOI 10.3109/19401736.2015.1053127.

China Plant BOL Group, Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL, Yang JB, Fu CX, Zeng CX, Yan HF, Zhu YJ, Sun YS, Chen SY, Zhao L, Wang K, Yang T, Duan GW. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America* **108**:19641–19646 DOI 10.1073/pnas.1104551108.

Dong W, Liu H, Xu C, Zuo Y, Chen Z, Zhou S. 2014. A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: a case study onginsengs. *BMC Genetics* **15**:138 DOI 10.1186/s12863-014-0138-z.

Dong W, Liu J, Yu J, Wang L, Zhou S. 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLOS ONE* **7(4)**:e35071 DOI 10.1371/journal.pone.0035071.

Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015. *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific Report* **5**:Article 8348 DOI 10.1038/srep08348.

George B, Bhatt BS, Awasthi M, George B, Singh AK. 2015. Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Current Genetics* **61**:665–677 DOI 10.1007/s00294-015-0495-9.

Gupta SK, Bhattacharyya TK, Ghosh TC. 2004. Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *Journal of Biomolecular Structure and Dynamics* **21**:527–535 DOI 10.1080/07391102.2004.10506946.

Ingvarsson PK, Ribstein S, Taylor DR. 2003. Molecular evolution of insertions and deletion in the chloroplast genome of Silene. *Molecular Biology and Evolution* **20**:1737–1740 DOI 10.1093/molbev/msg163.

Jo YD, Park J, Kim J, Song W, Hur C-G, Lee Y-H, Kang B-C. 2011. Complete sequencing and comparative analyses of the pepper (*Capsicum annuum* L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome. *Plant Cell Reports* **30**:217–229 DOI 10.1007/s00299-010-0929-2.

Kim SY, Gao JJ, Lee WC, Ryu KS, Lee KR, Kim YC. 1999. Antioxidative flavonoids from the leaves of *Morus alba*. *Archives of Pharmacal Research* **22**:81–85 DOI 10.1007/BF02976442.

Kong W, Yang J. 2016. The complete chloroplast genome sequence of *Morus mongolica* and a comparative analysis within the Fabidae clade. *Current Genetics* **62**:165–172 DOI 10.1007/s00294-015-0507-9.

Kress WJ, Erickson DL. 2007. A two-locus global DNA barcode for land plants: the coding *rbc*L gene complements the non-coding *trn*H-*psb* A spacer region. *PLOS ONE* **2(6)**:e508 DOI 10.1371/journal.pone.0000508.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye, Giegerich JR. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**:4633–4642 DOI 10.1093/nar/29.22.4633.

Lee C, Wen J. 2004. Phylogeny of Panax using chloroplast *trn*C-*trn*D intergenic region and the utility of *trn*C-*trn*D in interspecific studies of plants. *Molecular Phylogenetics and Evolution* **31**:894–903 DOI 10.1016/j.ympev.2003.10.009.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**:1451–1452 DOI 10.1093/bioinformatics/btp187.

Lohse M, Drechsel O, Bock R. 2007. Organellar Genome DRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52**:267–274 DOI 10.1007/s00294-007-0161-y.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Zhu X, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**:Article 18 DOI 10.1186/2047-217X-1-18.

Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV, Skyabin KG. 2008. Complete sequence of the duckweed (lemna minor) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. *Journal of Molecular Evolution* **66**:555–564 DOI 10.1007/s00239-008-9091-7.

Nie X, Deng P, Feng K, Liu P, Du X, You FM, Song W. 2013. Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family. *Plant Molecular Biology Reporter* **32**:828–840 DOI 10.1007/s11105-013-0691-z.

Ravi V, Khurana JP, Tyagi AK, Khurana P. 2006. The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis. *Tree Genetics & Genomes* **3**:49–59 DOI 10.1007/s11295-006-0051-3.

Roy S, Tyagi A, Shukla V, Kumar A, Singh UM, Chaudhary LB, Datt B, Bag SK, Singh PK, Nair NK, Husain T, Tuli R. 2010. Universal plant DNA barcode loci may not work in complex groups: a case study with Indian berberis species. *PLOS ONE* **5(10)**:e13674 DOI 10.1371/journal.pone.0013674.

Sablok G, Nayak KC, Vazquez F, Tatarinova TV. 2011. Synonymous codon usage, GC3, and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. *Molecular Biotechnology* **49**:116–128 DOI 10.1007/s12033-011-9383-9.

Sarkinen T, George M. 2013. Predicting plastid marker variation: can complete plastid genomes from closely related species help? *PLOS ONE* **8(11)**:e82266 DOI 10.1371/journal.pone.0082266.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**:W686–W689 DOI 10.1093/nar/gki366.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **24**:28–38 DOI 10.1007/BF02099948.

Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**:275–288 DOI 10.3732/ajb.94.3.275.

Shi C, Hu N, Huang H, Gao J, Zhao YJ, Gao LZ. 2012. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLOS ONE* **7(2)**:e31468 DOI 10.1371/journal.pone.0031468.

Song Y, Dong W, Liu B, Xu C, Yao X, Gao J, Corlett RT. 2015. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Frontiers in Plant Science* **6**:Article 662 DOI 10.3389/fpls.2015.00662.

Su HJ, Hogenhout SA, Ai-Sadi AM, Kuo CH. 2014. Complete chloroplast genome sequence of Omani lime (*Citrus aurantiifolia*) and comparative analysis within the rosids. *PLOS ONE* **9(11)**:e113049 DOI 10.1371/journal.pone.0113049.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**:2731–2739 DOI 10.1093/molbev/msr121.

The Sericultural Research Institute Chinese Academy of Agricultural Sciences. 1993. *Annals of mulberry varieties in China*. Beijing: China Agriculture Press, 7–8.

Thiel T, Michalek W, Varshney RK, Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106**:411–422 DOI 10.1007/s00122-002-1031-0.

Wang S, Shi C, Gao L. 2013. Plastid genome sequence of a wild woody oil species, *Prinsepia utilis*, provides insights into evolutionary and mutational patterns of Rosaceae chloroplast genomes. *PLOS ONE* **8(9)**:e73946 DOI 10.1371/journal.pone.0073946.

Worberg A, Quandt D, Barniske AM, Lohne C, Hilu KW, Borsch T. 2007. Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. *Organisms Diversity & Evolution* **7**:55–77 DOI 10.1016/j.ode.2006.08.001.

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87**:23–29 DOI 10.1016/0378-1119(90)90491-9.

Wu Z, Gui S, Quan Z, Pan L, Wang S, Ke W, Liang D, Ding Y. 2014. A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MipSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biology* **14**:289 DOI 10.1186/s12870-014-0289-0.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**:3252–3255 DOI 10.1093/bioinformatics/bth352.

Yang Y, Yuanye D, Qing L, Jinjian L, Xiwen L, Wang Y. 2014. Complete chloroplast genome sequence of poisonous and medicinal plant datura stramonium: organizations and implications for genetic engineering. *PLOS ONE* **9(11)**:e110656 DOI 10.1371/journal.pone.0110656.

Yang ZH, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**:32–43 DOI 10.1093/oxfordjournals.molbev.a026236.

Zhang SD, Soltis DE, Yang Y, Li DZ, Yi TS. 2011. Multi-gene analysis provides a well-supported phylogeny of Rosales. *Molecular Phylogenetics & Evolution* **60**:21–28 DOI 10.1016/j.ympev.2011.04.008.

Zhao W, Pan Y, Zhang, Jia Z, Miao S, Huang XY. 2005. Phylogeny of the genus Morus (Urticales: Moraceae) inferred from ITS and trnL-F sequences. *African Journal of Biotechnology* **4**:563–569 DOI 10.5897/AJB2005.000-3103.

Zhao Y, Yin J, Guo H, Zhang Y, Xiao, Sun C, Wu J, Qu X, Yu J, Wang X, Xiao J. 2015. The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. *Frontiers in Plant Science* **5**:Article 696 DOI 10.3389/fpls.2014.00696.

Zhou ZK, Gilbert Michael G. 2003. *Flora of China, morus*. Vol 5. Beijing, St. Louis: Science Press & Missouri Botanical Garden Press, 22–26.