Whole chloroplast genome characterization and comparison of two sympatric species in genus *Hippophae* (Elaeagnaceae)

Luoyun Wang[1], Jing Wang[1], Caiyun He[1], Jianguo Zhang[1,2,*], Yanfei Zeng[1,*]

1 Key Laboratory of Tree Breeding and Cultivation, State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China; wangluoyun14@163.com (L.W.), 18810576992@163.com (J.W.), hecy@caf.ac.cn (C.H.),

2 Collaborative Innovation Center of Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China

* Correspondence: zhangjg@caf.ac.cn (J.Z.), zengyf@caf.ac.cn (Y.Z.);

Tel.: 86-10-62889601 (J.Z.), 86-10-62888786 (Y.Z.)

Abstract: *Hippophae* is a tree species with ecological, economic and social benefits. In this study, we assembled and annotated chloroplast genomes of sympatric *Hippophae gyantsensis* and *H. rhamnoides* subsp. *yunnanensis*. Their full-length are 155260 and 156415 bp, respectively. Each of them has 131 genes, comprising 85 protein-coding genes, 8 ribosomal RNA genes and 38 transfer RNA genes. After comparing the chloroplast genomes, we found 1302 base difference loci, and 63.29% are located in the intergenic region or intron sequences and 36.71% are located in the coding sequences. The SSC region has the highest mutation rate, followed by the LSC region; the IR regions have the lowest mutation rate. Among the protein-coding genes, three had a ratio of nonsynonymous to synonymous substitutions (Ka/Ks) >1 (but *P* values were non-significant) and 66 had Ka/Ks <1 (46 were significant). In general, the chloroplast protein-coding genes may be subject to purification selection. Among *H. gyantsensis* and *H. rhamnoides* subsp. *yunnanensis* chloroplast protein-coding genes, there are 20 and 16 optimal codons, respectively. Most of the optimal codons were ending with A or U, which indicates significant AT preference. It is an important reference for studies on the general characteristics and evolution of the *Hippophae* chloroplast genome.

 doi:10.20944/preprints201810.0638.v1

Key words: Chloroplast genome; *H. gyantsensis*; *H. rhamnoides* subsp. *yunnanensis*; Ka/ks; Optimal codons

1. Introduction:

Chloroplasts are vital sites for green plants and algae to conduct energy conversion and photosynthesis. A chloroplast is a semiautonomous cellular organelle that has relatively independent genetic material, which is called chloroplast DNA. As chloroplast DNA is maternally inherited in most angiosperm[1], and it also has a relatively stable genetic structure, it has attracted broad attention from biologists. Studies on chloroplast genomes have become increasingly important in recent years. Most chloroplast genomes of plants have a cyclic structure that includes one large single copy (LSC) region, two inverted repeat (IR) regions and one small single copy (SSC) region. The full length of a chloroplast genome is 120–2500 kb, with 110–130 genes[2, 3]. With the continuous development of molecular biology, especially the development of large-scale high-throughput genetic sequencing techniques, research on chloroplast genomes is gradually deepening[4].

Hippophae plants belong to the family Elaeagnaceae. They are usually dioecious deciduous shrubs or dungarunga and are widespread in many countries in Asia and Europe. The root system of *Hippophae* plants can form root nodules, which possess nitrogen fixation functions and can improve soil fertility. *Hippophae* plants also have strong environmental adaptability: they can resist low-temperature sandy environments and high-temperature, saline-alkaline, dry and humid environments. Thus, they can be utilized for ecological restoration and soil protection[5, 6]. According to studies, in the Loess Plateau region, *Hippophae* forests can reduce 80% of the direct surface runoff, 75% of the water erosion of the surface soil and 85% of the wind erosion[7]. Both *Hippophae* fruits and leaves contain multiple ingredients that have beneficial effects for humans, including regarding cardiovascular protection, immunity augmentation, and cancer and other tumor inhibition, thus functioning as health food and medicine[8]. Hence, many kinds of wide-ranging research is carried out on *Hippophae* plants. Currently, there are 7 recognized species and 11 recognized subspecies of *Hippophae*[9]. As the country with the most abundant *Hippophae*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

resources, China has 7 species and 5 subspecies of *Hippophae*. There are already a large amount of researches analyzing *Hippophae* interspecific differentiation and genetic diversity using traditional methods such as those involving amplified fragment length polymorphism (AFLP), simple sequence repeats (SSRs), internal transcribed spacers (ITSs) and chloroplast trnL-F and trnS-G sequences[10, 11]. However, there have been very few studies conducted at the *Hippophae* chloroplast genome level[12], and there are no published studies comparing the chloroplast genomes of sympatric species of *Hippophae*. *H. gyantsensis* and *H. rhamnoides* subsp. *yunnanensis* are found in the Qinghai Tibetan Plateau, which is an area that is sensitive to climate change[13]. Studies show that alpine environments may change the chloroplast microstructure of plants[14], so conducting research on the chloroplast genome of plants in this area is of great importance.

This study did complete genome random interrupt sequencing for H. *gyantsensis* and H. *rhamnoides* subsp. *yunnanensis* by High-throughput sequencing, de novo assembled , annotated and systematically compared the complete chloroplast genomes of this two *Hippophae.* Then we obtained some chloroplast genomes characteristics of genus *hippophae*. It was expected to provide an important reference for the future studies on the chloroplast genome of genus *hippophae* plants and even angiosperms in plateau region.

2. Results

2.1 Basic characteristics of the chloroplast genomes

The full length of the *H. gyantsensis* chloroplast genome is 155260 bp, which is a little shorter than the length of the *H. rhamnoides* subsp. *yunnanensis* chloroplast genome (156415 bp). The length of the *H. gyantsensis* LSC region is 83026 bp, the length of the SSC region is 18894 bp, and the length of the IR region is 26670 bp. The lengths of the *H. rhamnoides* subsp. *yunnanensis* LSC and SSC regions are 84078 bp and 19047 bp, respectively, which are both larger than those of *H. gyantsensis*. However, the length of the *H. rhamnoides* subsp. *yunnanensis* IR region (26648 bp) is smaller than that of *H. gyantsensis*. GC concentration of the total and each of regions in the two *Hippophae* species were similar,and the IR region has the highest GC

 doi:10.20944/preprints201810.0638.v1

concentration. The difference between the two *Hippophae* chloroplast genome lengths (not including introns) is 5 bp (Table 1).

Both of the chloroplast genomes have 131 functional genes, comprising 85 protein-coding genes, 38 transfer RNA (tRNA) genes and 8 ribosomal RNA (rRNA) genes (Table 1). Their specific classifications are shown in Table 2. Among them, 4 rRNA genes, 8 tRNA genes and 7 protein-coding genes are located in the IR regions (with two copies); 13 genes are located in the SSC region; and the rest are located in the LSC region (Figure 1). Notably, *rps12* is a trans-splicing gene, and its 5' end is located in the LSC region and there is one copy of the 3' end in each IR region. The two *ycf1* genes have different lengths; the shorter segment is only located in one IR region while the longer segment is located in the other IR region and the SSC region. There are 22 genes with introns, *clpP*, *rps12* and *ycf3* genes have two introns and the rest have one intron. The lengths of introns in the two *Hippophae* chloroplast genomes are similar. The *ndhA* gene has the intron with the largest difference in length between the two species: the intron length of *H. gyantsensis* is 20 bp larger than that of *H. rhamnoides* subsp. *yunnanensis*. Among all of the introns, the *trnk-UUU* gene has the longest intron; its intron length is 2485 and 2497 bp in *H. gyantsensis* and *H. rhamnoides* subsp. *yunnanensis*, respectively (Table 3).

We detected 100 SSR loci for the *H. gyantsensis* chloroplast genomes, comprising 75, 15, 4, 5 and 1 of one-, two-, three-, four-, five- and six-base SSR loci, respectively. We detected 80 SSR loci for *H. rhamnoides* subsp. *yunnanensis*, comprising 54, 16, 4 and 6 of one-, two-, three- and four-base SSR loci, respectively. The one-base SSR loci for both *Hippophae* species are A/T repeats. All the two-base SSR loci for the two species are AT repeats except for one CT repeat loci for *H. gyantsensis*. The SSR loci of the two chloroplast genomes are mainly located in the non-coding region. *H. rhamnoides* subsp. *yunnanensis* has 13 SSR loci and *H. gyantsensis* has 9 SSR loci located in the coding sequences, 6 of these occur at the same locations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

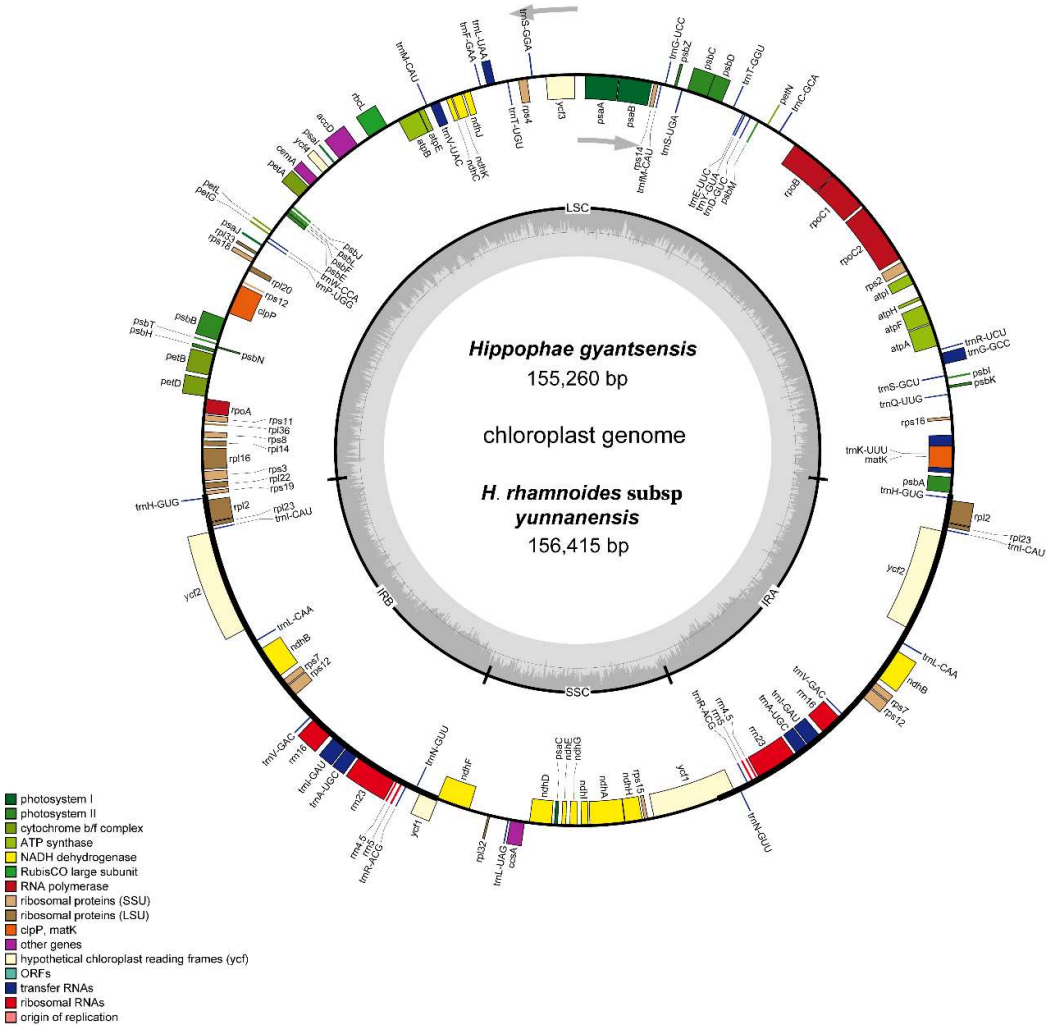


Fig 1 Physical map of the chloroplast genome for *H. gyantsensis* and *H.rhamnoides* subsp *yunnanensis*

Table1 Comparison of the Complete chloroplast genomes for *Hippophae gyantsensis* and *H. rhamnoides* subsp *yunnanensis*

Genome features	**H. gyantsensis**	**H. rhamnoides subsp yunnanensis**
Genome size (bp)/GC content (%)	155260/37	156415/37
LSC length (bp)/percent (%)/GC content (%)	83026/53.47/35	84072/53.75/35
SSC length (bp)/percent (%)/GC content (%)	18894/12.17/30	19047/12.17/30
IR length (bp)/percent (%)/GC content (%)	26670/17.18/42	26648/17.04/42
Gene size (bp) /percent (%)/GC content (%)	92455/59.55/40	92450/59.11/40
Number of genes	131	131
Number of protein-coding genes	85	85
Number of tRNA genes	38	38
Number of rRNA genes	8	8

IR: inverted repeat region ; LSC: large single copy region, SSC: small repeat region

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

Table2 Chloroplast Genome Coding Information of *H. gyantsensis* and *H. rhamnoides* subsp *yunnanensis*

Gene classification	Gene function	Gene name				
Photosynthesis	Photosystem I	*psaA*	*psaB*	*psaC*	*psaJ*	*psaI*
	Photosystem II	*psbA*	*psbE*	*psbJ*	*psbN*	*psbM*
		psbT	*psbC*	*psbH*	*psbL*	*psbK*
		psbB	*psbF*	*psbZ*	*psbD*	*psbI*
	Cytochrome b/f complex	*petA*	*petL*	*petB**	*petD**	*petG*
		petN				
	ATP synthase	*atpA*	*atpH*	*atpB*	*atpE*	*atpF**
		atpI				
	ATP Protease	*rbcL*				
Self-replication	Ribosomal RNA	*rrn4.5*(x2)	*rrn5*(x2)	*rrn16*(x2)	*rrn23*(x2)	
	Transfer RNA	*trnA-UGC*(x2)	*trnF-GAA*	*trnH-GUG*(x2)	*trnL-CAA*(x2)	*trnT-UGU*
		trnY-GUA	*trnD-GUC*	*trnG-GCC**	*trnI-GAU*(x2)	*trnV-UAC*
		trnW-CCA	*trnC-GCA*	*trnfM-CAU*	*trnI-CAU*(x2)	*trnS-UGA*
		trnV-GAC(x2)	*trnE-UUC*	*trnG-UCC*	*trnP-UGG*	*trnS-GCU*
		trnN-GUU(x2)	*trnR-UCU*	*trnT-GGU*	*trnR-ACG*(x2)	*trnK-UUU*
		trnL-UAG	*trnQ-UUG*	*trnS-GGA*	*trnL-UAA*	*trnM-CAU*
	Small subunit of ribosome	*rps2*	*rps8*	*rps15*	*rps3*	*rps7(x2)*
		rps4	*rps12****(x2)	*rps18*	*rps14*	*rps19*
		rps11	*rps16**			
	Large subunit of ribosome	*rpl2**(x2)	*rpl22*	*rpl36*	*rpl14*	*rpl33*
		rpl23(x2)	*rpl16**	*rpl32*	*rpl20*	
	RNA polymerase subunits	*rpoA*	*rpoB*	*rpoC1***	*rpoC2*	
	NADH dehydrogenase	*ndhA**	*ndhE*	*ndhI*	*ndhB**(x2)	*ndhH*
		ndhJ	*ndhC*	*ndhG*	*ndhK*	*ndhD*
		ndhF				
Other genes	Subunit of acetyl-CoA-carboxylase	*accD*				
	C-type cytochrome synthesis	*ccsA*				

 doi:10.20944/preprints201810.0638.v1

	Large subunit of rubisco	*matK*			
	Maturase	*clpP**			
	Envelope membrane protein	*cemA*			
Unknown function		*ycf1*(x2)	*ycf2*(x2)	*ycf3***	*ycf4*

*:gene has one intron; **：gene has two introns；***：gene is separated into two individual transcribed regions。
x2：gene has two copies.

Table 3 Characterization analysis of introns in *H. gyantsensis* and *H.rhamnoides* subsp *yunnanensis* chloroplast genomes

Gene	Location	Exon(bp)			*H. gyantsensis* Intron(bp)		*H. rhamnoides subsp yunnanensis* Intron(bp)	
		I	II	III	I	II	I	II
trnK-UUU	LSC	37	35		2485		2497	
trnG-GCC	LSC	23	48		697		698	
atpF	LSC	144	411		765		758	
rpoC1	LSC	435	1602		732		733	
ycf3	LSC	126	228	153	735	697	734	700
trnL-UAA	LSC	37	50		479		477	
trnV-UAC	LSC	39	37		588		588	
rps12(x2)	LSC	114	26	232		536		536
clpP	LSC	71	289	228	853	584	861	595
petB	LSC	6	642		788		794	
petD	LSC	8	475		717		718	
rpl16	LSC	9	399		1026		1023	
rpl2(x2)	IR	390	435		665		665	
ndhB(x2)	IR	777	756		681		681	
trnI-GAU(x2)	IR	42	35		947		947	
trnA-UGC(x2)	IR	38	35		802		802	
ndhA	SSC	552	540		1276		1256	

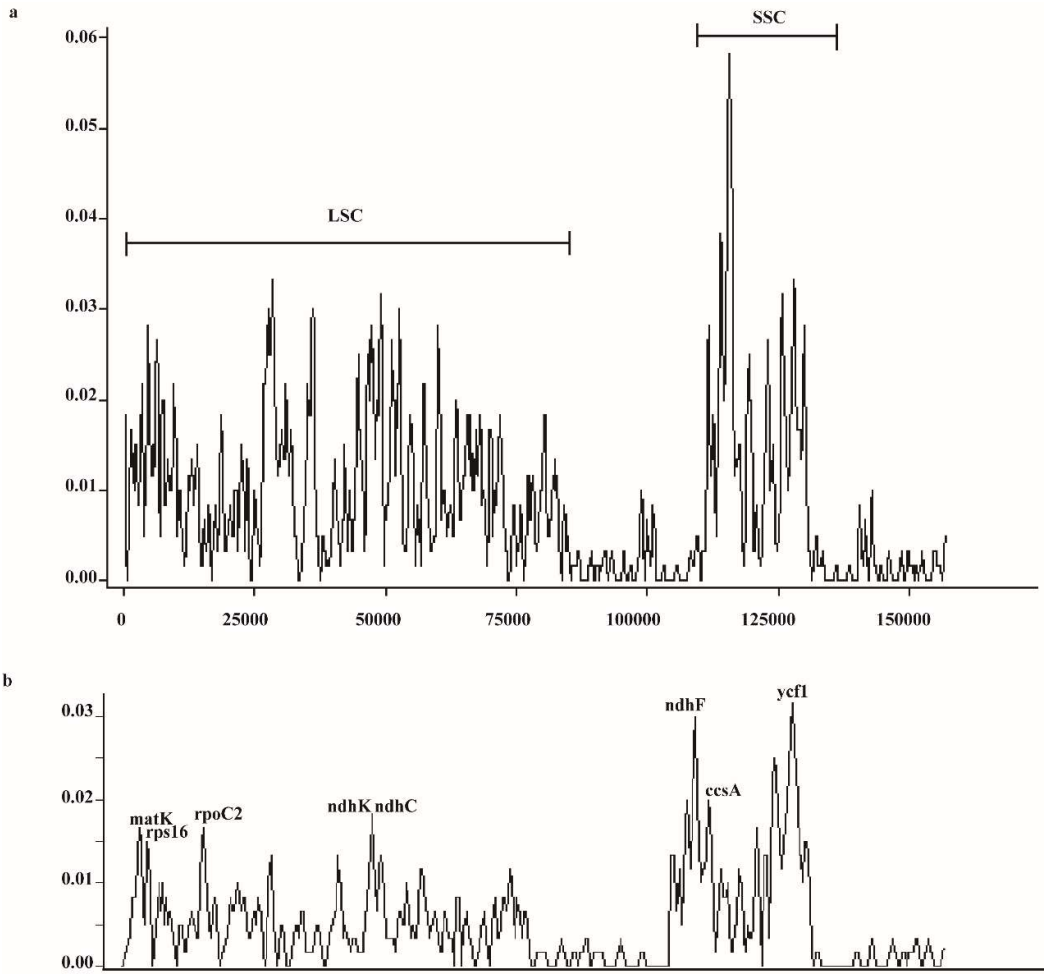
x2：gene has two copies

2.2 Base differences between chloroplast genomes

After comparing the complete chloroplast genomes of the two *Hippophae* species, we found 288 locations with base sequences insertions/deletions, 20 of which are >50 bp and mainly located in the LSC region. There are also 1302 base difference loci, 478 of which are in the coding region (36.71%) while 824 are in the intergenic region or are intron base difference loci (63.29%). The length of the coding region of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

each of the two species is 92455/92450 bp, and base difference loci amount to 5.17% . According to the results from the sliding-window analysis, regarding the complete chloroplast genomes, the largest difference between the two *Hippophae* species occurs in the SSC region, followed by the LSC region, and the IR regions have the smallest difference (Figure 2a). The coding region exhibited similar results; the longer segment *ycf1* gene located in the SSC region has the largest number of base difference loci per unit, followed by the *ndhF* gene in the SSC region; the *matK*, *rps16*, *rpoC2*, *ndhK* and *ndhC* genes in the LSC region have a relatively high number of base difference loci per unit (Pi value ≥0.15) (Figure 2b). The differences in the IR region are all small (Pi values <0.05). Comparing Figure 2a with 2b, the locations with large base differences are usually in the intergenic region.



X axes is the location of base，Y axes is the nucleotide polymorphism（Pi）；a：base difference of the complete chloroplast genomes，b：base difference of gene CDS region (not including introns)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

Fig 2 Sliding-window analysis of the chloroplast genomes of *H. gyantsensis* and *H.rhamnoides* subsp *yunnanensis*

2.3 Ka/Ks results of protein-coding genes in chloroplast genomes

Selection pressure on a gene can be identified by calculating the ratio of nonsynonymous substitutions (Ka) and synonymous substitutions (Ks) of protein-coding gene codons. According to Ka/Ks analysis of the protein-coding genes for the two *Hippophae* species, there are 472 substitute loci in total; 264 of them are Ks and 207 of them are Ka. Ka/Ks is 0.230 (<<1). The gene sequences of 19 protein-coding genes are completely the same between the two species, comprising 9 self-replicating-related genes and 10 photosynthesis-related genes. There are 66 protein-coding genes with Ka/Ks <1, and 46 of them have Ka/Ks <<1 (P value ≤0.05). After excluding genes with only 2 or 1 Ks or with only 1 Ka, there are 25 genes left, comprising 10 protein-coding genes that are related to photosynthesis, 11 protein-coding genes that are related to self-substitution and 4 genes with other functions. Ka/Ks >1 for the *matK*, *rpsl* and *rpoA* genes, though the statistic is non-significant(P value > 0.05).

Table 4. Kaks analysis of chloroplast genomes CDS region in *H. gyantsensis* and *H. rhamnoides* subsp *yunnanensis*

Sequence	Ka	Ks	Ka/Ks	P-Value(Fisher)	Substitutions	S-Substitutions	N-Substitutions
psbA	0	0.008164	0	0***	2	2	0
matK	0.012792	0.009082	1.40851	0.766264	18	3	15
rps16	0.033321	0.00945	3.52612	0.442746	6	0.5	5.5
atpA	0.002606	0.01927	0.13525	0.00276792**	10	7	3
atpF	0.002339	0.016329	0.143249	0.116775	3	2	1
atpH	0	0.030218	0	0***	2	2	0
atpI	0	0.017058	0	0***	3	3	0
rps2	0	0.006219	0	0***	1	1	0
rpoC2	0.00433	0.012759	0.339382	0.00668492**	26	12	14
rpoC1	0.001919	0.023829	0.0805486	0.0000162***	14	11	3
rpoB	0.003668	0.014811	0.247635	0.00215019**	20	11	9
petN	0	0.049127	0	0***	1	1	0
psbD	0	0.012173	0	0***	3	3	0
psbC	0	0.011665	0	0***	4	4	0
psbZ	0.007307	0.020917	0.349335	0.387491	2	1	1

 doi:10.20944/preprints201810.0638.v1

rps14	0.004344	0	NA	0***	1	0	1
psaB	0.001176	0.006027	0.195103	0.0799038	5	3	2
psaA	0.00116	0.02325	0.0498826	0.00000131***	14	12	2
ycf3	0	0.017916	0	0***	2	2	0
rps4	0.002167	0.007133	0.303798	0.358659	2	1	1
ndhJ	0	0.009797	0	0***	1	1	0
ndhK	0.007785	0.025179	0.309198	0.0969391	8	4	4
ndhC	0.007276	0.062188	0.116998	0.00865053**	7	5	2
atpE	0.003299	0.032204	0.102428	0.0406492*	4	3	1
atpB	0.002685	0.018911	0.141997	0.000406271***	10	7	3
rbcL	0.000923	0.02081	0.0443594	0.000262444***	8	7	1
accD	0.003277	0.009661	0.339173	0.153613	7	3	4
psaI	0	0.08097	0	0***	2	2	0
ycf4	0.009368	0.01613	0.58075	0.131858	6	2	4
cemA	0.001836	0.013656	0.134483	0.107091	3	2	1
petA	0.002738	0.013261	0.20648	0.0897403	5	3	2
psbE	0	0.073662	0	0***	4	4	0
psaJ	0	0.068657	0	0***	2	2	0
rpl20	0.011256	0.024653	0.456565	0.334622	5	2	3
clpP	0	0.021602	0	0***	3	3	0
psbB	0	0.008299	0	0***	3	3	0
psbH	0.00611	0	NA	0***	1	0	1
petB	0	0.013186	0	0***	2	2	0
petD	0	0.034702	0	0***	4	4	0
rpoA	0.005168	0.004638	1.11423	0.707553	5	1	4
rps11	0.003265	0.009404	0.347156	0.384262	2	1	1
rpl36	0	0.038688	0	0***	1	1	0
rps8	0.013219	0.020967	0.630478	0.444113	6	2	4
rpl16	0.006602	0.03039	0.217244	0.0999744	5	3	2
rps3	0	0.014104	0	0***	2	2	0
rpl22	0.003088	0.020584	0.150021	0.124411	3	2	1
rpl2(x2)	0	0.005007	0	0***	1	1	0
ycf2(x2)	0.001114	0.002022	0.550988	0.305029	9	3	6
ndhB(x2)	0	0.002754	0	0***	1	1	0
rps7(x2)	0	0.008877	0	0***	1	1	0
ycf1	0.007745	0.015987	0.484432	0.0945405	12	4.33333	7.66667
ndhF	0.010701	0.039575	0.270401	0.0000426***	38	19.5	18.5
ccsA	0.008078	0.034053	0.23723	0.0114163*	13	7	6
ndhD	0.003491	0.031685	0.110182	0.0000592***	15	11	4
psaC	0	0.018523	0	0***	1	1	0
ndhE	0	0.059331	0	0***	4	4	0

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

ndhG	0	0.01615	0	0***	2	2	0
ndhI	0	0.009306	0	0***	1	1	0
ndhA	0.002439	0.026624	0.0916031	0.000084***	9	7	2
ndhH	0.00109	0.02734	0.0398602	0.0001521658***	8	7	1
rps15	0.004963	0.052812	0.0939697	0.034479*	4	3	1
ycf1（longer）	0.01344	0.029824	0.450633	0.00053994***	93	32.8333	60.1667

x2: gene has two copies; *: 0.01 < p < 0.05; **: 0.001 < p < 0.01; ***: p < 0.001.

2.4 Comparison of the optimal codons of the chloroplast genomes

Based on the 53 genes of each of the two *Hippophae* species, and taking codons with RSCU >1 and ΔRSCU >0.08 to be the optimal codons, there are 20 optimal codons in the *H. gyantsensis* chloroplast genome and 16 in the *H. rhamnoides* subsp. *yunnanensis* chloroplast genome. Of these, 15 optimal codons are the same for the two species (Table 5). Regarding the amino acids coded by the codons, besides tryptophan (which has only one codon and no codon usage preference), only aspartic acid and cysteine in *H. gyantsensis,* and phenylalanine, aspartic acid, cysteine and glycine from *H. rhamnoides* subsp. *yunnanensis*, have no optimal codons. For all of the optimal codons for the two species, only UUG (coding for leucine) ends with G; the rest all end with A or U, which indicates a significant AT preference.

Table 5 Comparison of optimal codons of *H. gyantsensis* and *H.rhamnoides* subsp *yunnanensis* chloroplast genomes

Amino acid	Codon	*H. gyantsensis*		*H.rhamnoides* subsp *yunnanensis*	
		RSCU	ΔRSCU	RSCU	ΔRSCU
Phe	UUU*	1.39	0.65	—	—
Leu	UUA	1.98	0.74	1.97	0.55
	UUG	1.23	0.39	1.23	0.18
Ile	AUU	1.47	0.34	1.46	0.25
Val	GUU	1.49	0.46	1.48	0.5
	GUA	1.54	0.62	1.54	0.73
Ser	UCU	1.67	0.14	1.68	0.56
Pro	CCU	1.58	0.3	1.6	0.54
Thr	ACA	1.24	0.45	1.57	0.23
Ala	GCU*	—	—	1.8	0.21
	GCA	1.17	0.52	1.17	0.1
Tyr	UAU*	1.63	0.22	—	—
His	CAU*	1.54	0.29	—	—
Gln	CAA	1.54	0.22	1.08	0.2
Lys	AAA	1.52	0.55	1.27	0.55

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

Glu	GAA	1.52	0.11	1.17	0.23
Arg	CGU	1.36	0.67	1.6	0.62
	CGA	1.35	0.26	1.08	0.17
Ser	AGU	1.3	0.57	1.57	0.2
Arg	AGA*	1.89	0.25	—	—
Gly	GGU*	1.36	0.6	—	—

*: the optimal codon for one of the species.

2.5 Complete sequence cluster analysis of chloroplast genomes

We used the Bayes cluster, ML and MP methods to conduct a cluster analysis of the chloroplast genomes, with *Z. jujube* as the outgroup. Elaeagnaceae formed a single branch. Furthermore, there was a 100% agreement showing that *Elaeagnus macrophylla* Thunb. and *Elaeagnus mollis* formed a branch. *H. rhamnoides* subsp. *yunnanensis* first formed a small branch with *H. rhamnoides*, and then with *H. gyantsensis* (Figure 3). This reflects that although the two *Hippophae* species are located in the same geographical region, their chloroplast genomes indicate many interspecies differences.

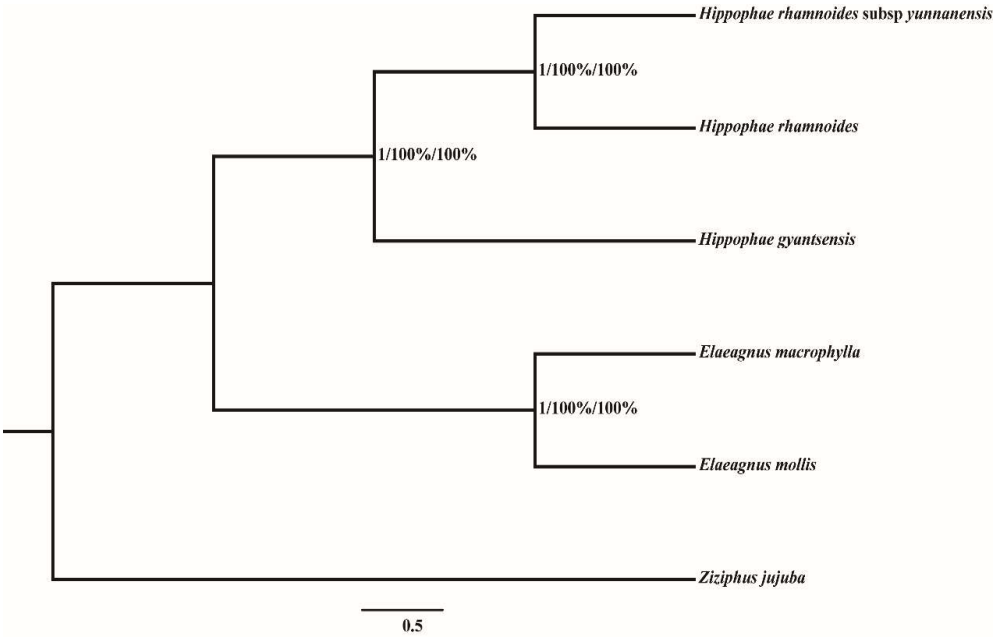


Fig 3 Cluster analysis of the chloroplast genomes

3 Discussion

3.1 Basic characteristics of the complete *Hippophae* chloroplast genomes

The full lengths of the chloroplast genomes of the two *Hippophae* species in this study and the published *H. rhamnoides* chloroplast genomes[12] are between 155260

 doi:10.20944/preprints201810.0638.v1

and 156415 bp, which are close to those of the model plant *Arabidopsis thaliana*, plants of the family Rosaceae (such as apple and pear) and plants of the family Salicaceae (such as *Populus trichocarpa* and *Cathay poplar*). Their genomes have the typical four-section structure of angiosperm chloroplast genomes[15-17]. The cluster analysis of the complete chloroplast genomes shows that *H. rhamnoides* subsp. *yunnanensis* formed a branch with *H. rhamnoides* with 100% support; it has a weaker genetic relationship with the same region of *H. gyantsensis*. Also, there is a complex pedigree sorting process between *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis*[18], so the chloroplast genomes of these two species represent the characteristics of *Hippophae* chloroplast genomes to a large degree.

Regarding the two *Hippophae* chloroplast genomes, the intergenic regions have the largest length difference, followed by the intron regions, while the length difference of the coding regions is only 5 bp. The total length difference is mainly caused by the LSC region, then the SSC region; the IR regions have the smallest length difference. SSR loci are mainly located in the intergenic regions, with AT repetition, probably as there is only a GC concentration of 37% in the *Hippophae* chloroplast genomes. The intron analysis indicated that the *trnk-UUU* gene has the longest intron, which is similar to results for many other species[19].

We found that the largest base difference between the *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis* chloroplast genomes is in the SSC region, followed by the LSC region; the IR regions have the smallest base difference. The GC concentration in the SSC region is 30%, while it is 35% in the LSC region and 42% in the IR regions. This shows that base differences in all genome regions is negatively correlated with GC concentration. A study by Zheng et al.[20] shows that for angiosperm chloroplast genomes, the IR region has a high GC concentration because the rRNA gene in this region has a high GC concentration, while the low concentration in the SSC region is related to the NADH genes in this region.

The IR regions of most angiosperm species are highly conserved[27] and have a relatively high GC concentration. This may be related to the presence of two copies. When mutations occur, the IR regions of angiosperm chloroplast genomes can adjust

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

via transposition in order to decrease the mutation ratio[21]. Additionally, we found that compared with the noncoding region in the IR region, the coding region has a lower mutation rate. There are probably two reasons for this. Firstly, genes in the coding region of the IR region experience a strong purification selection effect and a fast evolutionary rate; most of the mutations cannot be reserved and are directly eliminated, so there is a lower base difference than in the noncoding region. Secondly, mutations in the coding region of the IR region are only influenced by random drift, and the base mutation rate is lower than that of the noncoding region, so there are fewer base difference loci than in the noncoding region. Regarding protein-coding genes, we found that the purification effect is not significant for protein-coding genes in the IR region , which supports the second reason, which is that genes in the coding region of the IR region have a slower evolutionary rate.

3.2 Characteristics of *Hippophae* protein-coding genes

In the protein-coding genes of the two *Hippophae* chloroplast genomes, the gene that has the largest base difference is the longer segment *ycfl*. This gene is widespread in plant chloroplast genomes, but exhibits allele drop-out in herbaceous plants and cranberry plants[22]. Dong et al. found that among 420 species of *xylophyta*[23], 357 can be identified by the long segment *ycfl*, which is better than using *matK* and *rbcL*. The effectiveness of differentiation using the long segment *ycfl* is followed by differentiation using the *ndhF* gene (which encodes NADH dehydrogenase F) due to its high mutation ratio; many studies have used this gene to analyze interspecies genetic diversity in plants[24, 25]. Additionally, we found that in *Hippophae*, the *ccsA* gene also has a higher mutation rates, which conflicts with a former conclusion that the *ccsA* gene is widely conserved in photosynthetic plants[2]. The three genes above are completely or largely located in the SSC region. Additionally, the *matK*, *rps16*, *rpoC2*, *ndhC* and *ndh* genes in the LSC region also have relatively high mutation rates. All the abovementioned genes with high mutation rates are chloroplast self-replication-related genes or other functional genes, while the photosynthesis-related genes all have lower mutation rates because they need to be

 doi:10.20944/preprints201810.0638.v1

relatively strongly conserved to maintain photosynthesis as the main function of chloroplasts[2].

The Ka/Ks results shows that, in general, chloroplast genome protein-coding genes may be selected by purification effects. Among the 25 genes showing significant purification effect, 17 are located in the LSC region and 8 in the SSC region. As the number of protein-coding genes in the LSC region (60) is far greater than that in the SSC region (12), the purification effect of the SSC region protein-coding genes is more significant than in the LSC region. This illustrates the main reason why the SSC region has a larger base difference: compared with other chloroplast genome regions, the SSC region has the highest base mutation rate. Its significant purification effect can ensure that the normal functions of related genes are not disrupted.

3.3 Optimal codons

Codon usage preference analysis is an important method to explore genetic and evolutionary pathways of species[26]. The use of optimal codons is an important representation of codon preference. The research has shown that directional mutation and natural selection are two main factors that influence codon usage preference[27]. There are 15 common optimal codons for *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis* chloroplast genomes, which reflects the high commonality of codons in *Hippophae* plants. Except for one codon that ends with G, the rest all end with A or U, which reflects the significant AT preference. This is in accordance with many other analytical results on optimal codons in angiosperm[28, 29]. The utilization of certain optimal codons is matched with the nucleic acid content in the chloroplast genome intergenic region[29]. Species with high AT content in intergenic region also have optimal codons with high AT content. Through our calculations, the AT content in the intergenic regions of both *Hippophae* chloroplast genomes is 67.4%. Usually, the bases in the intergenic region will not be affected by selective action, so high AT content in the intergenic region probably occurs because the bases in this region are more likely to mutate into A/T. This also indicates that directional mutation may be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

the main factor for codon usage preference. Of course, natural selection may also have big effects, which needs to be further studied.

4. Materials and methodology

4.1 DNA extraction and sequencing

We collected leaves from a wild specimen of *H. gyantsensis* and *H. rhamnoides* subsp. *yunnanensis*, respectively, in Gongbo'gyamda County, Linzhi City of Tibet Autonomous Region (altitude: 3600 m) in August 2017. We treated them with silica gel desiccant and then brought them back to the lab. For each species, an appropriate amount of dry leaves was selected and the cetyltrimethylammonium bromide (CTAB) method was used to extract the total DNA. Agarose gel electrophoresis was used to assess the validity of the DNA extraction process. Then we sent the total DNA to Novogene Company for complete genome random interrupt sequencing, which involved HiSeqPE150 and a DNA-350 bp pool type. Finally, a 35-G sequence data set was obtained for each of the *Hippophae* species.

4.2 Assembling complete chloroplast genomes

After random interrupt sequencing, we used NOVOPlasty 2.7.0 to assemble the complete chloroplast genomes[30]. We downloaded the *H. gyantsensis* trnLF sequence (KU304417) and the *H. rhamnoides* subsp. *yunnanensis* trnLF sequence (KU304405) from the National Center for Biotechnology Information (NCBI) website, set them as the seed sequence, then set K-mer as 39, type as chloro and genome range as 120000–200000, inputted the forward reads and reverse reads sequence file location and set the other parameters as default.

To ensure assembly accuracy, after comparing the complete assembled chloroplast genome sequences of *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis*, we designed a primer for suspicious areas that had large differences, and used the total DNA of the two species as the templates to conduct PCR amplification. The PCR amplification products were sequenced by Beijing Tsingke Biological Technology Inc., and a comparative analysis was then conducted using these sequences and the assembled chloroplast genome sequences.

4.3 Chloroplast genome annotation and physical map drawing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

We used DOGMA (http://dogma.ccbb.utexas.edu/) to annotate the assembled chloroplast genomes, rectified the annotated results using Sequin15.10 software, submitted both annotations to the NCBI database and obtained a serial number for each submission. We uploaded two GenBank format chloroplast genome sequence files into OrganellarGenomeDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html), clearly labeled the locations of the IR, LSC and SSC regions in the chloroplast genomes and produced an annotated circular chloroplast genome map.

4.4 Comparative analysis of the two *Hippophae* chloroplast genomes

DnaSP 5.0 was utilized to count the sequence polymorphisms of the *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis* chloroplast genomes[31]. When screening for base differences between the chloroplast genomes, we used the sliding-window method for the calculations, and set the window length at 600 bp and the step size at 200 bp for the analysis.

4.5 SSR loci analysis

SSR loci of the two chloroplast genomes were searched by MISA software. The minimum number of repeats of mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide units were 10, 5, 4, 3, 3 and 3, respectively[32]; the smallest distance between two SSR loci was set as 100 bp.

4.6 Nonsynonymous substitution (Ka)/synonymous substitution (Ks) analysis

We aligned the coding regions for 85 protein-coding genes of *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis*, using the NG method in KaKs_Calculator 2.0 software[33] with default parameter values. Thus, we obtained Ka/Ks results for the two *Hippophae* species.

4.7 Optimal codon analysis

To ensure accuracy for the optimal codon analysis, we first removed the repetitive sequences in the protein-coding genes, then selected the sequences that ATG as the initiator codon and TAA, TAG and TGA as the terminator codons, and gene sequences >300 bp. There were 53 gene sequences for each of the two *Hippophae* species[34]. We used the effective number of codons (ENC) of each gene

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2018 doi:10.20944/preprints201810.0638.v1

as the criterion to order the selected gene sequences, then selected the 5 genes from both ends of the resulting sequence to construct high- and low-expression gene pools. We utilized CodonW to obtain the relative synonymous codon utilization (RSCU) value for the 53 genes and the ΔRSCU value for the high- and low-expression gene pools. Finally, we selected the optimal codons with RSCU>1 and ΔRSCU >0.08 for the *Hippophae* chloroplast genome[35].

4.8 Cluster analysis of chloroplast genomes

We searched for and downloaded the chloroplast genomes for *Ziziphus jujube* (NC_030299), *Elaeagnus macrophylla* (NC_028066), *Elaeagnus mollis* (NC_036932) and *Hippophae rhamnoides* (NC_035548.1), and analyzed them along with the two chloroplast genomes that we assembled in this study; *Z. jujube* was used as the outgroup. We compared the genomes using BioEdit software. We utilized the maximum likelihood (ML) method, maximum parsimony (MP) method with 1000 bootstrap replications and Bayes cluster method (mcmc ngen = 1000, sumt burnin = 2500) in MEGA7.0 software[36] to construct a phylogenetic tree for cluster analysis.

5. Conclusion

Through comparison and analysis of the chloroplast genome sequences of *H. rhamnoides* subsp. *yunnanensis* and *H. gyantsensis*, we found that the length of the *Hippophae* chloroplast genome is in accordance with most model plants and has the typical four-section structure of angiosperm chloroplast genomes. The evolution rates of the *Hippophae* chloroplast genome regions are different, and the rate is negatively related to GC content. The SSC region has the highest evolutionary rate, so it has the highest mutation rate and lowest GC content. The IR region has the lowest evolutionary rate, so it has the lowest mutation rate and highest GC content. All index values for the LSC region are in the middle of those for the former two regions. The self-replication-related genes or other functional genes of chloroplasts have relatively higher mutation rates, while the photosynthesis-related genes are more conserved. Protein-coding genes in the SSC region are significantly affected by purification selection to ensure normal operation of related protein-coding genes under the high mutation rates in the SSC region. These conclusions provide an important reference

for future studies on *Hippophae* plants, including regarding the characteristics of chloroplast genomes of angiosperm in high-altitude areas.

Acknowledgements: This work was financially supported by the National Natural Science Foundation of China (31670666) and the Fundamental Research Funds for the Central Non-profit Research Institution of Chinese Academy of Forestry (ZDRIF201706).

Authors contributions: LYW and YFZ developed the research idea and wrote the manuscript. LYW performed laboratory work. LYW and JW performed data analyses. CYH and JGZ obtained funding for the research and sampled the two species. All authors read and approved the final manuscript

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Corriveau, J. L.; Coleman, A. W. Rapid Screening Method to Detect Potential Biparental Inheritance of Plastid DNA and Results for Over 200 Angiosperm Species. *Am J. Bot.* **1988,** *75*, 1443-1458.
2. Susann, W.; Schneeweiss, G. M.; Depamphilis, C. W.; Müller, K. F.; Dietmar, Q. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* **2011,** *76*, 273-297.
3. Wang, L.; Dong, W. P.; Zhou, S. L. Structural mutations and reorganizations in chloroplast genomes of flowering plants. *Acta Bot. Boreal. -Occident. Sin.* **2012,** *32*, 1282-1288.
4. Ding, Y. Q.; Fang, Y.; Jin, Y. L.; Zhao, H.; He, K. Systematic evolution of *Lemnoideae* determined based on chloroplast genome analysis. *Chin J Appl Environ Biol* **2017,** *23*, 215-219.
5. Kortesniemi, M.; Sinkkonen, J.; Yang, B.; Kallio, H. NMR metabolomics demonstrates phenotypic plasticity of sea buckthorn (Hippophae rhamnoides) berries with respect to growth conditions in Finland and Canada. *Food Chem* **2017,** *219*, 139-147.
6. Stobdan, T.; Angchuk, D.; Singh, S. B. Seabuckthorn: An emerging storehouse for researchers in India. *Curr. Sci. India* **2008,** *94*, 1236-1237.
7. Zhou, D. S. Study on the genetic diversity of *Hippophae rhamnoides* subsp. *sinensis* and *Hippophae rhamnoides* subsp. *yunnanensis*. Chinese Academy of Forestry, 2005.
8. Qian, X. S.; Jin, J. H. Medical research and development of sea-buckthorn. *Chinese Wild Plant Res.* **2015,** *34*, 68-72.
9. Bartish, I. V. Phylogeny of Hippophae(Elaeagnaceae) inferred from parsimony analysis of chloroplast DNA and morphology. *Syst Bot* **2002,** *27*, 47-54.

10. Ma, Y. H.; Feng, G. S.; Xiang, Q. S.; Gao, Y.; Ynag, C. J.; Wei, G. L.; Song, W. X. Phylogenetic relationships of seabuckthorn based on ITS sequences. *Chinese J Appl Ecol* **2014,** *25*, 2985-2990.
11. Li, R.; Yu, Y. T. Application and prospect of DNA molecular markers in seabuckthorn heredity and breeding. *Glob Seabuckthorn Res. Dev.* **2008,** *6*, 19-25.
12. Chen, S. Y.; Zhang, X. Z. Characterization of the complete chloroplast genome of seabuckthorn (Hippophae rhamnoides L.). *Conservation Genetics Resources* **2017,** *9*, 623-626.
13. Cheng, K. Study on genetic diversity and protection of *Hipppophae gyantsensis* and *Hippophae rhamnenoide* subsp. *yunnanensis*. Sichuan Agricultural University Mster Dissertation, 2008.
14. Lutz, C.; Engel, L. Changes in chloroplast ultrastructure in some high-alpine plants: adaptation to metabolic demands and climate? *Protoplasma* **2007,** *231*, 183-192.
15. Li, Y.; Lv, G. H.; Zhang, X. N.; He, X. M. Chloroplast genome structure and variation analysis of Brassicaceae species. *Acta Bot. Boreal. -Occident. Sin.* **2017,** *37*, 1090-1101.
16. Cheng, H.; Ge, C. F.; Zhang, H.; Qiao, Y. Advances on chloroplast genome sequencing and phylogenetic analysis in fruit trees. *J Nucl Agr Sci* **2018,** *32*, 58-69.
17. Fan, L. Q.; Hu, H.; Zhang, H. L.; Wang, T. J.; Wang, Y. L.; Ma, T.; Mao, K. S. Complete sequence and comparative analysis of the chloroplast genome of the Chinese aspen (*Populus adenopoda*, Salicaceae). *J Sichuan Univ:Nat Sci Ed* **2018,** *55*, 165-171.
18. Cheng, K.; Sun, K.; Wen, H. Y.; Zhang, M.; Jia, D. R.; J.Q., L. Maternal divergence and phylogeographical relationships between *Hippophae gyantsensis* and *H. rhamnoides* subsp. *yunnanensis*. *Chinese J Plant Ecol* **2009,** *33*, 1-11.
19. Zhang, H. Y.; Li, C.; Miao, H.; Xiong, S. Insights from the Complete Chloroplast Genome into the Evolution of Sesamum indicum L. *Plos One* **2013,** *8* (11), e80508.
20. Zhengqiu, C.; Penaflor, C.; Kuehl, J. V.; Leebensmack, J.; Carlson, J.; Depamphilis, C. W.; Boore, J. L.; Jansen, R. K. Complete chloroplast genome sequences of Drimys, Liriodendron, andPiper: Implications for the phylogeny of magnoliids and the evolution of GC content. *Office of Scientific & Technical Information Technical Reports* **2006,** *6*.
21. Ruhlman, T. A.; Jansen, R. K. The plastid genomes of flowering plants. *Methods Mol Biol* **2014,** *1132*, 3-38.
22. Jan, D. V.; Sousa, F. L.; Bettina, B. L.; Jürgen, S.; Gould, S. B. YCF1: A Green TIC? *Plant Cell* **2015,** *27*, 1827-1833.
23. Dong, W.; Xu, C.; Li, C.; Sun, J.; Zuo, Y.; Shi, S.; Cheng, T.; Guo, J.; Zhou, S. ycf1, the most promising plastid DNA barcode of land plants. *Sci Rep-UK* **2015,** *5*, 1-5.
24. Datwyler, S. L.; Weiblen, G. D. On the origin of the fig: phylogenetic relationships of Moraceae from ndhF sequences. *Ame J Bot* **2004,** *91*, 767-777.
25. Steinmann, V. W.; Porter, J. M. Phylogenetic relationships in Euphorbieae (Euphorbiaceae) based on its and ndhF sequence data. *Ann Mo Botl Gard* **2002,** *89*, 453-490.
26. Zhao, Y.; Zheng, H.; Xu, A.; Yan, D.; Jiang, Z.; Qi, Q.; Sun, J. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution. *Bmc Genomics* **2016,** *17*, 677.

27. Wang, L.; Xing, H.; Yuan, Y.; Wang, X.; Saeed, M.; Tao, J.; Feng, W.; Zhang, G.; Song, X.; Sun, X. Genome-wide analysis of codon usage bias in four sequenced cotton species. *Plos One* **2018,** *13*, e0194372.

28. Hu, S. S.; Luo, H.; Wu, Q.; Yao, H. P. Analysis of codon bias of chloroplast genome of *Tartary buckwheat*. *Mol Plant Breeding* **2016,** *14*, 309-317.

29. Hershberg, R.; Petrov, D. A. General Rules for Optimal Codon Choice. *Plos Genetics* **2009,** *5*, e1000556.

30. Dierckxsens, N.; Mardulyn, P.; Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res* **2017,** *45*, e18.

31. Librado, P.; Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **2009,** *25*, 1451-1452.

32. Li, Q. L.; Yan, N.; Song, Q.; Guo, J. Z. Complete chloroplast genome sequence and characteristics analysis of *Morus multicaulis*. *Chinese Bull of Bot* **2018,** *53*, 94-103.

33. Zhang, Z.; Li, J.; Zhao, X. Q.; Wang, J.; Wong, G. K.; Yu, J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genom Proteom Bioinf* **2006,** *4*, 259-263.

34. Liu, H.; Wang, M. X.; Yue, W. J.; G.W., X.; L.Q., G.; X.J., N.; Song, W. N. Analysis of codon usage in the chloroplast genome of Broomcorn millet(*Panicum miliaceum* L.). *Plant Sci J* **2017,** *35*, 362-371.

35. Wang, P. L.; Yang, L. P.; Wu, H. Y.; Nong, Y. L.; Wu, S. C.; Xiao, Y. F.; Qin, Z. H.; Wang, H. Y.; Liu, H. L. Condon preference of chloroplast genome in *Camellia oleifera*. *Guihaia* **2018,** *38*, 135-144.

36. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **2016,** *33*, 1870-1874.