Characterization of the complete chloroplast genome sequence and phylogenetic analysis of *B. oleracea* var. *italica*

Zhenchao Zhang

Zhenjiang institute of agriculture research

Zhongliang Dai (✉ zzc1981zzc@163.com)

Yuemei Yao

Zhenjiang institute of agriculture sciences

Yongfei Pan

Zhenjiang institute of agriculture sciences

Guosheng Sun

Zhenjiang institute of agriculture sciences

Xi Shan

Zhenjiang institute of agriculture sciences

Research article

Keywords: broccoli, chloroplast genome, phylogenetic relation ship, simple sequence repeat

Posted Date: January 15th, 2020

DOI: <https://doi.org/10.21203/rs.2.20976/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Backgrounds: Broccoli (*Brassica oleracea* var. *italica* L.) is known as one of the most nutritionally rich vegetables, as well as rich in functional components that benefit to health. The main purposes of this research were sequencing, assembling and annotation of chloroplast genome of broccoli based on Illumina HiSeq2500 sequencing platform.

Results: The size of the broccoli cp genome is 153,364 bp, including two inverted repeat (IR) regions of 26,197 bp each, separated by a small single copy (SSC) region of 17,834 bp and a large single copy (LSC) region of 83,136 bp. The GC content of the complete genome is 36.36%, while those of SSC, LSC, and IR are 29.1%, 34.15% and 42.35%, respectively. It harbors 134 functional genes, including 87 protein-coding genes, 39 tRNAs and 8 rRNAs, with 31 duplicates in the IRs. The most abundant amino acid in the protein-coding genes is leucine, while the least is cysteine. Codon usage frequency showed bias for A/T-ending codons in the cp genome. In the repeat structure analysis, a total of 34 repeat sequences and 291 simple sequence repeat (SSRs) were detected in the work. Although cp genomic structure and size are highly conserved, the SC-IR boundary regions are variable between the 7 cp genomes. The phylogenetic relationships based on complete cp genome from 9 species suggest that *B. oleracea* var. *italica* is closely related to *Brassica juncea*.

Conclusions: The complete cp genome sequence was obtained and annotated for broccoli for the first time. The information acquired from this research will be useful for further species identification, population genetics and biological research of broccoli.

Background

Chloroplasts (cp) are the most important and common plastid in plant cells, and are crucial organelle which are responsible for carbohydrate metabolism and photosynthesis in photoautotrophic plants [1–3]. Chloroplasts are considered to be “metabolic center” of cellular reactions and have been playing a critical role in various molecular processes, including synthesis of amino acid, nucleotides, phytohormones, vitamins, and other metabolites, as well as in the regulation of the physiology, growth, development and stress response [2, 4, 5]. The cp genome is mostly circular double-stranded structure, with multiple copies in cells and the size ranges from 120 to 160 kb [6–7]. The typical cp genome of angiosperms has a highly conserved quadripartite structure with a SSC region and a LSC region divided by two IR regions with similar sequence [4, 7]. Evolutionary studies suggested that cp genome is relatively conservative in genome organization and gene content, and few recombination or intraspecific variation occurs in angiosperms [8, 9]. Due to the characteristics of simple conserved structure and gene content, small compact genome size, cp genomes have been used for construction of phylogenetic classification [10–11] and DNA barcoding system [12, 13].

Broccoli is one of the major members among *B. oleracea* varieties. It is a highly nutritious vegetable, especially in vitamin A and K, antioxidants, β -carotene, calcium, riboflavin, iron [14–15]. It was also recognized as a functional vegetable. Consumption of broccoli is beneficial to human health, such as anti-inflammatory, anti-obesity, cholesterol-lowering and anti-carcinogenic, as well as increase antioxidant activity [16–19]. The activity properties are confirmed by health-promoting phytochemicals, such as phenols, flavonoids, glucosinolates, minerals and selenium [20–22], that accumulate in broccoli.

Broccoli originated in many areas of Europe along the eastern Mediterranean basin and has been cultivated throughout the world since the 15th century [23]. Broccoli is also cultivated for at least 30 years throughout China, with a cultivation area and total mass of 76,000 hm² and 3547,000 tons, respectively, in 2017, and the cultivation area will increase year by year [24]. However, being an economically important crop, genomic resources for broccoli are limited. Up to now, there are no published systematic researches of broccoli cp genome, although two assembly of the broccoli complete cp genomes have already been deposited in NCBI (accession number: MH388765.1 and MH388764.1). With the advances of high-throughput Illumina genome sequencing technologies, new methods have given us an opportunity to obtain and analyze the complete chloroplast genome of broccoli, then to better understand its molecular and genomic characteristics.

In the current research, we de novo sequenced and assembled the complete cp genome of broccoli with a Illumina HiSeq2500 sequencing platform, then analyzed its gene annotation and structure. Followed, we identified a number of SSR marks in our new assembly and reconstructed the evolutionary relationships with the family to investigate the phylogenetic position in several Brassicaceae species. It is expected that the results will be helpful to improve understanding of the cp genome and will also provide a theoretical basis for future scientific research of broccoli.

Results

Features of broccoli cp genome

The cp genome of broccoli was a typical quadripartite circular molecule 153,364 bp in length, and contained a pair of two IR (IRA and IRB) regions of 26,197 bp each, separated by a small single copy (SSC) region of 17,834 bp and a large single copy (LSC) region of 83,136 bp (Fig. 1 and Table 1). The overall cp genome AT and GC content is 63.64% and 36.36%, respectively. The percentage GC content of the IR regions (42.35%) is higher than that of the LSC and SSC regions (34.15% and 29.1%, respectively), which is similar to the cp genomes of other species (Fig. 2).

Table 1
Summary of cp genome of broccoli.

Features	Numerical value	Features	Numerical value
Genome size (bp)	153,364	GC content in SSC region (%)	29.1
LSC length (bp)	83,136	Gene number	134
SSC length (bp)	17,834	Protein-coding gene number	87
IR length (bp)	26,197	tRNA gene number	39
AT content (%)	63.64	rRNA gene number	8
GC content (%)	36.36	Gene number in LSC regions	82
GC content in IR region (%)	42.35	Gene number in SSC regions	21
GC content in LSC region (%)	34.15	Gene number in IR regions	31

The *B. oleracea* var. *italica* cp genome harbors 134 predicted functional genes and LSC, SSC and IR regions contain respectively 82, 21 and 31 genes (Fig. 1, Tables 1 and 2). Among of them, eight are ribosomal RNA (rRNA) genes (*rrn4.5*(× 2), *rrn5*(× 2), *rrn16*(× 2), *rrn23*(× 2)), 39 are transfer RNA (tRNA) genes, 14 genes are coding for small ribosomal subunit (SSU) and 11 for large ribosomal subunit (LSU), and 4 genes for the DNA-directed RNA polymerase. Forty-seven genes are dedicated to photosynthesis of which seven coded for the photosystem I and 15 for the photosystem II complex, 12 for subunits of the NADH dehydrogenase, six for the cytochrome b/f complex, six for different subunits of the ATP synthase and one for Calvin cycle. Five genes are dedicated to different functions except for self-replication and photosynthesis, and 6 genes are still unknown about their functions. Thirty-one genes, including 16 tRNA genes, two *rps7*, *rps19*, two *rpl2*, two *rpl23*, two *ycf1*, two *ycf2* and four *ycf15* are duplicated in the IR regions.

Table 2
Gene contents in the broccoli cp genome

Category for genes	Group of genes	Name of genes	Account
Self-replication	Ribosomal RNA genes	rrn4.5(× 2), rrn5(× 2), rrn16(× 2), rrn23(× 2)	8
	Transfer RNA genes	trnA-UGC ^{a,*} (× 2), trnI-GAU ^{a,*} (× 2), trnN-GUU ^a (× 2), trnV-GAC ^a (× 2), trnL-CAA ^a (× 2), trnI-GAU ^{a,*} (× 2) trnR-UCG ^{a,*} (× 2), trnR-ACG ^a (× 2), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnH-GUG, trnI-CAU ^a , trnK-UUU*, trnL-UAA*, trnL-UAG, trnP-UGG, trnQ-UUG, trnR-UCU, trnS-GGA, trnS-UGA, trnS-GCU, trnT-CGU*, trnT-UGU, trnT-GGU, trnY-GUA, trnM-CAU, trnV-UAC*, trnW-CCA	39
	Small subunit ribosomal proteins (SSU)	rps2, rps3, rps4, rps7 ^a (× 2), rps8, rps11, rps12 ^{**} (× 2), rps14, rps15, rps16*, rps18, rps19 ^a	14
	Large subunit ribosomal proteins (LSU)	rpl2 ^{a,*} (× 2), rpl14, rpl16*, rpl20, rpl22, rpl23 ^a (× 2), rpl32, rpl33,rpl36	11
	RNA polymerase	rpoA, rpoB, rpoC1*, rpoC2	4
Photosynthesis	Photosystem I	psaA,psaB, psaC, psal,psaJ,ycf3**, ycf4	7
	Photosystem II	psbA,psbB,psbC,psbD,psbE,psbF,psbH,psbI,psbJ,psbK,psbL,psbM,psbN, psbT, psbZ	15
	NADH dehydrogenase	ndhA*, ndhB ^{a,*} (× 2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK	12
	Cytochrome b/f complex	petA, petB*, petD*, petG ,petL, petN	6
	ATP synthase	atpA, atpB, atpE, atpF*, atpH, atpI	6
	Calvin cycle	rbcL	1
Other genes	Subunit of acetyl-CoA	accD	1
	Envelope membrane protein	cemA	1
	Maturase	matK	1
	Protease	clpP**	1
	C-type cytochrome synthesis gene	ccsA	1
Genes of unknown functions	Hypothetical chloroplast reading frames (ycf)	ycf1 ^a (× 2), ycf2 ^a (× 2), ycf15 ^a (× 2)	6
^a Two gene copies in IRs; * Gene contains one intron; ** gene contains two introns; (x2) indicates the number of the repeat unit is 2.			

Among the 134 distinct genes, a total of 21 genes contained one intron, including 11 coding genes (rps16, two rpl2, rpl16, rpoC1, ndhA, two ndhB, petB, petD, and atpF) and 10 tRNA genes (two trnR-UCG, two trnA-UGC, two trnI-GAU, trnK-UUU, trnL-UAA, trnT-CGU and trnV-UAC), while four genes (two rps12, ycf3 and clpP) contain two introns (Table 2 and Table 3). The trnR-UCG gene has the largest intron (2572 bp), followed by trnR-UCG gene (2570 bp), whereas the trnL-UAA has the smallest intron (313 bp). The length of intron in trnK-UUU gene is 2555 bp, and the matK gene is contained within the intron.

Table 3
The lengths of introns and exons in genes in the broccoli cp genome.

Gene	Location	Exon I(bp)	Intron II(bp)	Exon II(bp)	Intron II(bp)	Exon III(bp)
trnK-UUU	LSC	38	2555	37		
rps16	LSC	46	859	221		
trnT-CGU	LSC	34	712	43		
atpF	LSC	144	721	411		
rpoC1	LSC	432	778	1611		
ycf3	LSC	126	782	228	732	153
trnL-UAA	LSC	35	313	50		
trnV-UAC	LSC	40	599	36		
clpP	LSC	69	546	297	935	252
petB	LSC	6	775	648		
petD	LSC	8	733	475		
rpl16	LSC	9	1061	402		
rpl2	IRB	393	684	435		
ndhB	IRB	777	679	762		
rps12	IRB	232	-	27	539	101
trnI-GAU	IRB	32	809	40		
trnA-UGC	IRB	37	800	36		
trnR-UCG	IRB	37	2570	60		
ndhA	SSC	552	1098	531		
trnR-UCG	IRa	36	2572	59		
trnA-UGC	IRA	38	798	37		
trnI-GAU	IRA	33	807	41		
ndhB	IRA	777	679	762		
rpl2	IRA	393	684	435		
rps12	IRA	101	-	232	539	27

Codon Usage examination

According to the Coding sequence (CDS), the relative synonymous codon usage frequency (RSCU) and codon usage frequency were estimated in the study (Table 4, Fig. 3). All the protein-coding genes in the cp genome composed a total of 26,659 codons. Among these codons, the termination codons were UAA, UAG and UGA. The UUA encoded leucine had the highest RSCU value (2.025), whereas the GUG encoded methionine has the lowest value 0.0034. The most abundant amino acid in the protein-coding genes is leucine (2827 codons, about 10.6% of the total), while only 325 codons (1.22%) encoded cysteine.

Table 4
The codon usage for the broccoli cp genome.

AminoAcid	Codon	Number	RSCU	tRNA	AminoAcid	Codon	Number	RSCU	tRNA
Ter	UAA	52	1.7931		Met	GUG	1	0.0034	
Ter	UAG	22	0.7587		Asn	AAC	304	0.4624	trnN-GUU
Ter	UGA	13	0.4482		Asn	AAU	1011	1.5376	
Ala	GCA	382	1.116	trnA-UGC	Pro	CCA	310	1.1632	trnP-UGG
Ala	GCC	206	0.602		Pro	CCC	195	0.7316	
Ala	GCG	146	0.4264		Pro	CCG	141	0.5292	
Ala	GCU	635	1.8552		Pro	CCU	420	1.576	
Cys	UGC	80	0.4924	trnC-GCA	Gln	CAA	739	1.5508	trnQ-UUG
Cys	UGU	245	1.5076		Gln	CAG	214	0.4492	
Asp	GAC	201	0.384	trnD-GUC	Arg	AGA	472	1.8006	trnR-UCU
Asp	GAU	846	1.616		Arg	AGG	161	0.6144	
Glu	GAA	1061	1.5212	trnE-UUC	Arg	CGA	357	1.362	trnR-UCG
Glu	GAG	334	0.4788		Arg	CGC	109	0.4158	
Phe	UUC	519	0.6424	trnF-GAA	Arg	CGG	125	0.477	
Phe	UUU	1097	1.3576		Arg	CGU	349	1.3314	trnR-ACG
Gly	GGA	733	1.6592		Ser	AGC	125	0.366	trnS-GCU
Gly	GGC	168	0.3804	trnG-GCC	Ser	AGU	410	1.2006	
Gly	GGG	291	0.6588		Ser	UCA	419	1.227	trnS-UGA
Gly	GGU	575	1.3016		Ser	UCC	293	0.858	trnS-GGA
His	CAC	149	0.491	trnH-GUG	Ser	UCG	199	0.5826	
His	CAU	458	1.509		Ser	UCU	603	1.7658	
Ile	AUA	725	0.9462		Thr	ACA	428	1.244	trnT-UGU
Ile	AUC	434	0.5664	trnI-GAU	Thr	ACC	246	0.7152	trnT-GGU
Ile	AUU	1140	1.4877		Thr	ACG	146	0.4244	trnT-CGU
Lys	AAA	1168	1.5348	trnK-UUU	Thr	ACU	556	1.6164	
Lys	AAG	354	0.4652		Val	GUA	512	1.436	trnV-UAC
Leu	CUA	396	0.8406	trnL-UAG	Val	GUC	181	0.5076	trnV-GAC
Leu	CUC	190	0.4032		Val	GUG	200	0.5612	
Leu	CUG	172	0.3648		Val	GUU	533	1.4952	
Leu	CUU	587	1.2456		Trp	UGG	452	1	trnW-CCA
Leu	UUA	954	2.025	trnL-UAA	Tyr	UAC	188	0.3818	trnY-GUA

AminoAcid	Codon	Number	RSCU	tRNA	AminoAcid	Codon	Number	RSCU	tRNA
Leu	UUG	528	1.1208	trnL-CAA	Tyr	UAU	797	1.6182	
Met	AUG	602	1.9966	trnI-CAU					

The codon-anticodon recognition patterns of the cp genome showed that 29 tRNAs contained codons corresponding to 20 essential amino acids for protein biosynthesis. The AT content at the first, second and third codon positions were 55.3%, 62.53% and 71.21%, respectively. Moreover, of all 65 codons, the RSCU values of 31 codons were ≥ 1 , and most of those (13/15, 90.3%) end with base A or U, whereas 33 codons had the RSCU ≤ 1 , and most of those (16/14, 90.91%) end with base C or G. Trp is encoded by only a UGG codon, which means no codon biased usage (RSCU = 1).

Repeat sequences and SSR analysis

A total of 34 repeat sequences, including 12 forward (F), 20 palindromic (P) and 2 reverse (R) repeats, were detected by REPuter in the broccoli cp genome (Table 5). The length of most of these repeats were between 30 to 47 bp, and the longest repeat was 26,197 bp in length and was located in the IR region. LSC, SSC and IR regions harbored 20, 7 and 13 repeats separately. Most repeats were mainly located in the intergenic spacers (IGS), ycf gene and intron sequences, whereas 13 repeats were located in the psaA, psaB, rrn4.5, trnS-GGA, trnS-UGA, rps19 and trnS-UGA .

Table 5
Repeat sequences in the broccoli chloroplast genome.

ID	Repeat Start	Type	Size(bp)	Repeat Start2	Mismatch(bp)	E-Value	Gene	Region
1	3753	F	30	120284	-3	6.29E-04	trnK-UUU(intron);ndhA(intron)	LSC;SSC
2	7602	F	31	34396	-3	1.74E-04	IGS	LSC
3	37703	F	43	39927	-3	2.85E-11	psaB;psaA	LSC
4	37724	F	46	39948	-3	5.48E-13	psaB;psaA	LSC
5	42809	F	30	97786	-2	2.25E-05	ycf3(intron);IGS	LSC;IRb
6	61538	F	47	61582	0	3.34E-19	IGS	LSC
7	88069	F	32	88090	-3	4.80E-05	ycf2	IRb
8	97777	F	37	119317	-3	7.35E-08	IGS;ndhA(intron)	IRb;SSC
9	106663	F	34	106695	-2	1.13E-07	trnR-UCG(intron)/rrn4.5	IRb
10	124227	F	30	124254	-3	6.29E-04	ycf1	SSC
11	129771	F	34	129803	-2	1.13E-07	trnR-UCG(intron)/rrn4.5	IRa
12	148378	F	32	148399	-3	4.80E-05	ycf2	IRa
13	171	P	36	171	-2	7.94E-09	IGS	LSC
14	6222	P	32	6222	0	3.59E-10	IGS	LSC
15	7603	P	30	43912	-1	5.16E-07	IGS;trnS-GGA	LSC
16	9181	P	36	9181	0	1.40E-12	IGS	LSC
17	28144	P	40	28144	0	5.47E-15	IGS	LSC
18	34397	P	30	43912	-3	6.29E-04	IGS;trnS-GGA	LSC
19	34465	P	30	43850	-3	6.29E-04	trnS-UGA;IGS	LSC
20	42809	P	30	138684	-2	2.25E-05	ycf3(intron);IGS	LSC;IRa

ID	Repeat Start	Type	Size(bp)	Repeat Start2	Mismatch(bp)	E-Value	Gene	Region
21	61472	P	30	61472	0	5.74E-09	IGS	LSC
22	64604	P	30	64604	-2	2.25E-05	IGS	LSC
23	65896	P	30	65947	-3	6.29E-04	IGS	LSC
24	75673	P	45	75673	-1	7.21E-16	petD(intron)	LSC
25	88069	P	32	148378	-3	4.80E-05	ycf2	IRb;IRa
26	88090	P	32	148399	-3	4.80E-05	ycf2	IRb;IRa
27	83136	P	26197	127167	0	0.00E+00	rps19;ycf1	IRb;IRa
28	119317	P	37	138686	-3	7.35E-08	ndhA(intron);IGS	SSC;IRa
29	106663	P	34	129771	-2	1.13E-07	trnR-UCG(intron)/rrn4.5	IRb;IRa
30	106695	P	34	129803	-2	1.13E-07	trnR-UCG(intron)/rrn4.5	IRb;IRa
31	122595	P	30	123175	-2	2.25E-05	IGS;ycf1	SSC
32	124277	P	30	124277	-2	2.25E-05	ycf1	SSC
33	173	R	30	34490	-3	6.29E-04	IGS;trnS-UGA	LSC
34	184	R	30	112593	-3	6.29E-04	IGS	LSC;SSC

In the present research, a total of 291 SSRs, including 196 mononucleotides (P1), 15 dinucleotides (P2), 58 trinucleotides (P3), 4 tetranucleotides and 18 complex duplicate nucleotides, were explored, and most of them distributed in the SSC (189, 64.9%) and IR region (60, 20.6%) and partly distributed in the LSC region (42,14.4%) (Tables 6 and 7, Fig. 4). Among them, eight SSRs belonged to the C repeat units, and the others all belonged to the A and T types (96.94%), while TA and AT repeats composed all the dinucleotides. Trinucleotides are the second most prevalent and the number of repeat unit type is 24, higher than others. Meanwhile, 18 complex duplicate nucleotides were detected with lengths ranging from 17 to 47 bp. Most of the detected complex SSRs were within noncoding region; 15 were located in the IGS regions, 2 were located in the intron regions and one was contained in the ycf1 gene. Moreover, we also found 144 repeats were located in different genes, and the remaining were all harbored in intergenic regions.

Table 6
No. of SSRs distributed in the SSC, LSC and IR region.

Region	Exon	Intron	Intergenic	Number	Proportion
SSC	40	5	15	189	64.90%
LSC	35	35	119	42	14.40%
IR	23	6	13	60	20.60%

Table 7
Distribution of SSRs in the broccoli cp genome.

SSR type	Unit	Length	Number	Position on Genome(gene)
P1	A	8	40	221 – 218(IGS),1721–1728(trnK-UUU(intron)), 3436–3443(trnK-UUU(intron)/matK), 4100–4107(trnK-UUU(intron)), 4292–4299(IGS), 8202–8209(IGS), 14572–14579(IGS),17655–17662(rpoC2),21333–21340(rpoC1),21983–21990(rpoC1(intron)),28904–28911(IGS),30214–30221(IGS),30327–30334(IGS),44864–44871(IGS),45192–45199(IGS),45819–45826(IGS),45979–45986(IGS),47031–47038(IGS),50427–50434(trnV-UAC(intron)),53052–53059(IGS),60763–60770(petA),66057–66064(IGS),67871–67878(IGS),68627–68634(IGS),69779–69786(clpP),65230–65237(IGS),65961–65968(IGS),78805–78812(IGS),80051–80058(IGS),107026–107033(trnG-UCG(intron)),111087–111094(ndhF),111931–111938(IGS),112659–112666(IGS)114404–114411(IGS),114881–114888(ndhD),123903–123910(ycf1),126065–126072(ycf1),126529–126536(ycf1),127039–127046(ycf1),153228–153235(IGS)
		9	20	2842–2850(trnK-UUU(intron)),7742–7750(IGS),11106–11114(IGS),11574–11582(atpF(intron)),12713–12721(IGS),13601–13609(IGS),35230–35238(IGS),35681–35689(IGS),36335–36343(IGS),42030–42038(ycf3(intron)),55200–55208(IGS),67760–67768(IGS),70825–70833(IGS),80070–80078(IGS),80883–80891(rpl16(intron)),88691–88699(ycf2),116187–116195(IGS),119697–119705(ndhA(intron)),129019–129027(trnR-UCG(intron)),153266–153274(IGS)
		10	7	12446–12455(IGS),41568–41577(IGS),50050–50059(trnV-UAC(intron)),66012–66021(IGS),109314–109323(ycf1/ndhF),122605–122614(IGS),138192–138201(IGS)
		11	6	26937–26947(IGS),60220–60230(IGS),64103–64113(IGS),82925–82935(IGS),119762–119772(ndhA(intron)),137413–137423(IGS)
		13	3	67123–67135(IGS),124958–124970(ycf1),126037–126049(ycf1)
		14	1	113669–113682(ccsA)
		16	1	30260–30275(IGS)

Note: p1 means single base repetition; p2 means double base repetition; c means complex duplicate types.

SSR type	Unit	Length	Number	Position on Genome(gene)
	T	8	50	5715–5722(rps16(intron)),13390–13397(IGS),25143–25150(ropB),26288–26295(IGS),26565–26572(IGS),27828–27835(IGS),30357–30364(IGS),34514–34521(IGS),44198–44205(IGS),45162–45169(IGS),45548–45555(IGS),49491–49498(IGS),49766–49773(IGS),55111–55118(IGS),55396–55403(IGS),55476–55483(IGS),56056–56063(accD),58567–58574(ycf4),59448–59455(cemA),63934–63941(IGS),64197–64204(IGS),66201–66208(IGS),69220–69227(clpP(intron)),69261–69268(clpP(intron)),70403–70410(clpP(intron)),70888–70895(IGS),73109–73116(psbT),74029–74036(petB(intron)),74512–74519(petB(intron)),76831–76838(IGS),76852–76859(ropA),78375–78382(IGS),79473–79480(IGS),79602–79609(IGS),83266–83273(IGS),110130–110137(ndhF),112922–112929(IGS),120083–120090(ndhA(intron)),120227–120234(ndhA(intron)),122439–122446(rps15),123003–123010(ycf1),123749–123756(ycf1),124382–124389(ycf1),124861–124868(ycf1),125017–125024(ycf1),125189–125196(ycf1),125449–125456(ycf1),125572–125579(ycf1),126660–126667(ycf1),129468–129475(trnR-UCG(intron))
		9	24	2917–2925(trnK-UUU(intron)/matK),14417–14425(IGS),18047–18055(rpoC2),22032–22040(rpoC1(intron)),26088–26096(IGS),29021–29029(IGS),41145–41153(IGS),42465–42473(ycf3(intron)),49983–49991(trnV-UAC(intron)),59022–59030(IGS),59063–59071(IGS),64910–64918(IGS),65420–65428(IGS),69913–69921(clpP(intron)),70093–70101(clpP(intron)),81417–81425(rpl16(intron)),83227–83235(rps19),107474–107482(trnR-UCG(intron)),114563–114571(IGS),117892–117900(IGS),122272–122280(rps15),124755–124763(ycf1),126925–126933(ycf1),147802–147810(ycf2)
		10	17	15624–15633(ropC2),16763–16772(ropC2),25247–25256(ropB),28408–28417(IGS),29383–29392(IGS),53171–53180(IGS),55633–55642(IGS),64126–64135(IGS),70288–70297(clpP(intron)),98300–98309(IGS),81550–81559(rop16(intron)),87661–81670(IGS),80677–80686(rpl16(intron)),81154–81163(rpl16(intron)),123187–123196(ycf1),124444–124453(ycf1),127178–127187(ycf1)
		11	9	3978–3988(trnK-UUU(intron)),6815–6825(IGS),7777–7787(IGS),12467–12477(IGS),17512–17522(rpoC2),74110–74120(petB(intron)),123208–123218(ycf1),99078–99088(IGS),126007–126017(ycf1)
		12	4	4061–4072(trnK-UUU(intron)),70265–70276(clpP(intron)),63338–63349(IGS),123096–123107(ycf1)

Note: p1 means single base repetition; p2 means double base repetition; c means complex duplicate types.

SSR type	Unit	Length	Number	Position on Genome(gene)
		13	3	47324–47336(IGS),77038–77050(rpoA),125310–125322(ycf1)
		14	2	50869–50882(IGS),124990–125003(ycf1)
		15	1	12160–12174(atpF(intron)),
		16	1	7336–7351(IGS)
		19	1	124803–124821(ycf1)
	C	8	4	53228–53235(IGS),60689–60696(petA),63613–63620(IGS),78950–78957(IGS)
		9	1	5150–5158(rps16(intron))
		11	1	62109–62119(IGS)
P2	TA	10	8	4557–4566(IGS),6234–6243(IGS),7841–7850(IGS),18884–18893(rpoC2),26480–26489(IGS),61869–61878(IGS),93476–93485(IGS),122815–122824(ycf1)
		18	1	111597–111614(IGS)
	AT	10	4	7917–7926(IGS),107448–107457(trnR-UCG(intron)),129044–129053(trnR-UCG(intron)),143015–143024(IGS)
		12	1	13319–13330(IGS)
		14	1	30560–30573(IGS)
P3	AAT	9	3	16489–16497(ropC2),30852–30860(IGS),63724–63732(IGS),
		12	1	12612–12623(IGS)
	AAC	9	3	13748–13756(atpI),66874–66882(rps18),48656–48664(ndhK)
	AAG	9	2	43153–43161(ycf3(intron)),146995–147003(ycf2)
	ATT	9	3	111963–111971(IGS),112814–112822(IGS),113554–113562(ccsA)
		12	1	45612–45623(IGS)
	ATA	9	2	47401–47409(IGS),61935–61943(IGS)
	ATC	9	2	58154–58162(IGS),152669–152677(rpl2(intron))
	AGA	9	2	94531–94539(ndhB),147101–147109(ycf2)
	AGC	9	1	120428–120436(ndhA)
	ATG	9	1	39972–39980(psaA)
	AGT	9	1	34220–34228(IGS)
	CAA	9	2	108795–108803(ycf1),143853–143861(ycf15)

Note: p1 means single base repetition; p2 means double base repetition; c means complex duplicate types.

SSR type	Unit	Length	Number	Position on Genome(gene)
	CTG	9	1	103736–103744(rrn23)
	CAG	9	1	132757–132765(rrn23)
	CTT	9	1	89498–89506(ycf2)
	TTC	9	6	658–666(psbA),77128–77136(rpoA),83360–83368(rpl2),119111–119119(ndhA),126491–126499(ycf1),145265–145273(ycf2)
	TAA	9	6	4767–4775(IGS),41996–42004(ycf3(intron)),62142–62150(IGS),70034–70042(clpP(intron)),110965–110973(ndhF),122535–122543(IGS)
	TTA	9	3	14594–14602(IGS),49936–49944(trnV-UAV(intron)),53321–53329(IGS)
		12	1	26447–26458(IGS)
	TCT	9	5	36188–36196(rps14),82334–82342(rps3),89392–89400(ycf2),110087–11095(ndhF),141962–141970(ndhB)
	TGC	9	1	82710–82718(rpl22)
	TTG	9	2	92640–92648(ycf15),127698–127706(ycf1)
	GAA	9	2	91228–91236(ycf2),153133–153141(rpl2)
	GAT	9	3	37747–37755(psaB),14811–14819(rps2),83824–83832(rpl2(intron))
	GCA	9	1	39646–39654(psaA)
	GCT	9	1	54974–54982(rbcL)
P4	CAAA	12	1	27991–28002(IGS)
	TTCT	12	1	34268–34279(IGS)
	TAAA	12	1	45436–45447(IGS)
	ATAG	12	1	111359–111370(ndhF)
C	(A)8ctacg(A)10	23	1	12854–12876(IGS)
	(A)11gata(AT)6	27	1	112599–112625(IGS)
	(A)14tttgg(T)9	28	1	140343–140370(IGS)
	(T)8atc(A)9	20	1	136–155(IGS)
	(A)9ccaaa(T)14	28	1	96131–93158(IGS)
	(A)8cgatttgtc(T)16actact(A)8	47	1	111764–111810(IGS)
	(C)8(T)9	17	2	41080–41096(IGS),65897–65913(IGS)
	(T)9gatc(A)9	22	2	98259–98280(IGS),138221–238242(IGS)
	(T)10ac(T)9	21	1	48803–48823(IGS)

Note: p1 means single base repetition; p2 means double base repetition; c means complex duplicate types.

SSR type	Unit	Length	Number	Position on Genome(gene)
	(T)9caga(T)8	21	1	8259–8279(IGS)
	(AT)7tcc(T)11ca(T)11	41	1	120287–120327(ndhA(intron))
	(TAA)3c(TAA)3	19	1	30778–30796(IGS)
	(A)8gt(A)9gatcta(TATC)3	37	1	47095–47131(IGS)
	(T)8caagaaa(TAT)3aatcgata(T)9	41	1	112016–112056(IGS)
	(TTC)3gtttttatc(A)9(T)9	36	1	125951–125986(ycf1)
C*	(AT)7(T)14*	27	1	3756–3782(trnK-UUU(intron))
Note: p1 means single base repetition; p2 means double base repetition; c means complex duplicate types.				

Junction characteristics of IRs

The expansion and contraction of IR-SSC and IR-LSC boundaries of seven species, including *B. oleracea* var. *italica*, *Arabidopsis thaliana*, *Capsella bursa – pastoris*, *B. napus*, *B. juncea*, *B. nigra*, *Bunias orientalis*, were compared and the results were presented in Fig. 5. The IR sizes of the LSC, IR and SSC regions were similar in the cp genomes of the 7 species studied, and the lengths of IR varied from 26,035 bp in *B. napus* to 26,451 bp in *Capsella bursa – pastoris* (accession number: AP009371). The JLB border fell within the coding region of *rps19* gene in the above 7 species and left a fragment of the *rps19* gene different from 106 bp to 114 bp in the IRA region. Two genes *ycf1* and *ndhF* crossed the JSB junction. Most part of the *ycf1* gene in the 7 species were located in the IRB region and 0–3 bp located in the SSC region. Overlaps between the *ycf1* and *ndhF* were detected at the JSB boundary in all studied cp genomes, with lengths varied from 35 bp to 38 bp. The *ycf1* crossed the JSA region except in the cp genome of *B. juncea*, and its length reflected changes in the JSA region. The tRNA noncoding gene *trnH-GUG* in the 7 species were all fell in the LSC region, which at a distance of 3–30 bp from the JLA boundary. The results in this study suggested that the IR border shifts relatively minor, involving only a small number of genes, differences in gene overlap lengths and the distance of *trnH-GUG* gene located at the junction of JLA boundaries only presented irregular shifting.

Phylogenetic analysis

The cpDNA- based phylogenetic analysis has revealed better understanding of evolutionary relationship, population analysis and classification in rice species [25], Brassica genus [26], Myrtales species [27], and *Aristolochia* species [3]. To investigate the taxonomic status and evolutionary relationship of broccoli, alignment of 9 complete cp genomes downloaded from NCBI Gene Bank database was performed, and a ML tree was constructed by FastTree. The phylogenetic analysis corroborated the traditional taxonomy of the Brassicaceae with 100% bootstrap support (Figure 6). Specifically, 6 species *Alliaria grandifolia* (NC-034286.1), *Arabidopsis thaliana* (NC-000932.1), *Capsella bursa–pastoris* (NC-009270.1), *Braya humilis* (NC-035515.1), *Bunias orientalis* (NC-036111.1) and *Brassica nigra* (NC-030450.1) were clustered into a group, *Brassica napus* (NC-016734.1) was clusted into a group, and *Brassica. oleracea* var. *italica* was found to be closely related to *Brassica juncea* (NC-028272.1). The complete cp genome information reported in this paper would provide valuable data for future research to clarify the genomic information of broccoli chloroplasts and the chloroplast information could also be used on the phylogeny, molecular marker, DNA barcoding and conservation genetics.

Discussion

The typical organization cp genome of *Brassica oleracea* var. *italica* with two identical IRs separating the SSC and LSC regions exhibits identical gene content and order to land plant cp genomes. The size of the cp genome at 153,364 bp obtained in this study is exactly the same with the reference of *B. oleracea* var. *capitata* (NCBI accession: KX681654.1) (153364 bp), and also with the broccoli cp genomes deposited in NCBI (accession number: MH388765.1 (153,365 bp), MH388764.1 (153,365 bp), and KX681657.1 (153,363 bp)). The results obtained in this study indicated that the DNA GC content was not distributed evenly among the four regions. The GC content in the IR region is higher than that in other regions, this is possible because of the presence of higher GC content of the four rRNAs in this region and the DNA GC content was usually thought as an important indicator to distinguish species affinity [28–31].

The broccoli genome contains 134 genes, with highly conserved in composition and arrangement, including self-replication genes, photosynthetic genes, other functional genes, and some other genes of unknown functions, which is consistent with previous research [32]. Among the distinct genes, 25 genes contain one intron and two introns, and even the gene *trnR-UGG* has the largest intron. Introns play crucial roles in the regulation of some gene expression [33]. They might improve gene expression level, in the specific situation and on the special position [34–35]. Coding usage has key parts in shaping cp genome evolution. Among codons of broccoli cp genome, the most and the least used frequently amino acids is leucine and cysteine respectively, which is the same as reported in other angiosperm genomes, such as *Ananas comosus*, *Decaisnea insignis*, *Nasturtium officinale*, *M. zenii* [36–38]. The broccoli cp genome codons preferred AT over GC, especially at the second and the third position, 62.53% and 71.21%, respectively, which is consistent with results widely observed in many terrestrial species [39–40].

Repeat analysis revealed 12 forward, 20 palindromic and 2 reverse repeats in the broccoli cp genome. Most of these repeats were located in intron sequences, intergenic spacers and *ycf* gene, but several occurred in CDS and tRNAs. It was reported that repeat sequences took part in sequence variations, genome rearrangements and many rearrangement endpoints in rearranged algal and angiosperm genomes [41–44]. Because of highly conserved organization of cp genome sequences, and the SSR primer for cp genomes can be inherited across genera and species. SSRs are widely used as molecular markers genetic linkage map construction, population genetics, polymorphism research and plant breeding and also play an important role in plant taxonomy [45–47]. A total of 291 SSRs were obtained in this study, and 196 (67.4%) SSRs belonged to the P1 type, among of them, 190 (65.3%) belonged to A and T repeat units, while TA and AT repeats composed all the P2 type. These findings agree with the results in other researches [30, 31, 36]. Several complex SSRs were also detected, they might caused by two or more individual simple sequence repeats adjacent to each other and divided by a certain length of sequences [7, 48].

Plant cp genomes have been thought as highly reserved, but the sizes and LSC/IRb/SSC/IRa boundaries will change due to contraction/expansion at the borders of the IR region [49–50]. Our results indicated that the IR border variations between 7 species mainly because they crossed different position of 4 genes, *rps19*, *ycf1*, *ndhF* and *trnH-GUG*, this agree with the results of previous researches [36, 51–52].

Conclusions

In this research, we assembled the complete cp genome of broccoli with Illumina HiSeq platform. Annotation and comparison of the obtained data with reference helped us identify and verify 134 functional genes including 47 RNA and 87 protein-coding genes. The codon usage was biased toward A/T-ending. Repeat sequences and SSRs detected in the work could be used for molecular marker development and phylogenetic analysis. Phylogenetic reconstruction based on complete cp genome showed that *B. oleracea* var. *italica* is closely related to *Brassica juncea*. All the information presented in this paper will facilitate further biological research and genetic engineering of broccoli.

Methods

DNA extraction and sequencing

Fresh leaves were collected from broccoli plants grown in Zhenjiang Institute of Agricultural Sciences, Jiangsu Province, China, and immediately stored at -80°C until analysis. Total genomic

DNA was extracted from the leaves with the Plant DNA Isolation Reagent (Takara, USA) following the manufacturer's protocol. An Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) and NanoDrop 2000 Micro spectrophotometer were used to evaluate quality and integrity of the extracted DNA. After purification, the DNA was employed to build a sequencing library according to the manufacturer's instructions. An Illumina HiSeq2500 platform (San Diego, USA) was utilized to construct paired-end (PE) libraries with insert sizes of 150 bp and sequenced according to the standard protocols, including sample quality testing, library construction and quality testing and library sequencing.

Chloroplast genome assembly and annotation

High-quality clean reads were generated by trimming and filtering out the low-quality reads and sequencing adapters with Trimmomatic v 0.3649. The clean reads were mapped onto the available cp genome reference of *B. oleracea* var. *capitata* (NCBI accession: KX681654.1) using Bowtie2 software [53] with its default parameters and preset options. All of the cp-like reads were assembled into contigs by SPAdes [54]. Then the obtained contigs were align again on the reference of *L. capitata* using BLAST algorithm. The generated contigs and mate-pair reads were used to scaffold using SSPACE program [55] until a circular genome formed.

The annotation of tRNAs, rRNAs and protein-coding genes of the plastome was performed by CpGAVAS pipeline [56] and then corrected manually. The BLAST and DOGMA were used to check the results of the annotation [57]. The tRNAScanSE was applied to analyzed the tRNAs [58]. The physical circular cp genome map was generated using OrganellarGenomeDRAW (OGDRAW) program [59] with default settings and checked manually. The relative synonymous codon usage (RSCU) was performed with CodonW software [60]. The long repetitive sequences and simple sequences repeats (SSRs) were analyzed by REPuter [61] and MISA [62], respectively.

Cp genomes comparison

The complete cp genome of *B. italica* was compared with the cp genomes of *Arabidopsis thaliana*, *Capsella burse-pastoris*, *Brassica napus*, *Brassica juncea*, *Brassica nigra* and *Bunias orientalis* by mVISTA program in the shuffle-LAGAN mode [63], using the annotation of *B. oleracea* var. *capitata* as a reference. The comparison of IRB-LSC, IRB-SSC, IRA-SSC and IRA-LSC boundaries among the 7 species was carried out with the annotations of their cp genomes available in GenBank.

Phylogenetic analysis

To estimate phylogenetic relationships and analyze the phylogenetic position of broccoli within the Brassicaceae, the complete cp genomes of 9 taxa were compared. The cp genomes of other 8 species were downloaded from the NCBI database. The software MAFFTA [64] was used to align the sequences. The ML (maximum likelihood) method was employed to construct phylogenetic tree by FastTree version 2.1.10[65] with manual adjustment if necessary. Bootstrap analysis was calculated with 500 replications.

Abbreviations

NCBI:National Center for Biotechnology Information; JLB:LSC/IRB; JSB:IRA/SSC; JSA:SSC/IRA; JLA:IRA/LSC; BLAST:Basic Local Alignment Search Tool.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The annotated chloroplast genome of *Brassica oleracea* var. *italica* L. has been deposited in the NCBI GenBank with the accession number MN649876.1 (<https://www.ncbi.nlm.nih.gov/nuccore/MN649876.1>).

Competing interests

The authors declare that they have no competing interests.

Funding

ZCZ would like to gratefully thanks supported by National Natural Science Foundation of China (NSFC, grant No. 31401871) and Zhenjiang Science and Technology Innovation Fund Project (No. NY2019001). The funders did not take part in the design of the study and assembling, annotation and analysis of data and in writing the manuscript.

Authors' contributions

ZCZ and ZLD assembled and annotated the chloroplast genome, analysed the data and wrote the draft paper, YMY and YFP prepared and sequenced DNA libraries, SGS and XS designed and coordinated research and finalized the paper. All authors read and approved of the final manuscript.

Acknowledgements

Not applicable.

References

1. Bobik K, Burch-Smith TM. Chloroplast signaling within, between and beyond cells. *Front Plant Sci.* 2015;6:781.
2. Chung HY, Won SY, Kim YK, Kim JS. Development of the chloroplast genome-based InDel markers in Niitaka (*Pyrus pyrifolia*) and its application. *Plant Biotechnol Rep.* 2019;13:51-61.
3. Li XQ, Zuo YJ, Zhu XX, Liao S, Ma JS. Complete chloroplast genomes and comparative analysis of sequences evolution among seven *Aristolochia* (Aristolochiaceae) medicinal species. *Int. J Mol Sci.* 2019;20:1045.
4. Wicke S, Schneeweiss GM, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol.* 2011;76(3-5):273-297.
5. Spetea C, Hundal T, Lundin B, Heddad M, Adamska I, Andersson B. Multiple evidence for nucleotide metabolism in the chloroplast thylakoid lumen. *P Natl Acad Sci USA.* 2004; 101:1409-14.
6. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. The complete chloroplast genome sequence of *Pelargonium × hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol Biol Evol.* 2006;23(11):2175-2190.

7. Yan C, Du JC, Gao L, Li Y, Hou XL. The complete chloroplast genome sequence of watercress (*Nasturtium officinale* R. Br.): Genome organization, adaptive evolution and phylogenetic relationships in Cardamineae. *Gene*. 2019;699:24-36.
8. Raubeson LA, Jansen RK. Chloroplast genomes of plants. In: Henry R. (ed) Diversity and evolution of plants-genotypic and phenotypic variation in higher plants. CABI Publishing, Oxfordshire. 2005;45-68.
9. Asif H, Khan A, Iqbal A, Khan I, Heinze B, Azim M. The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. *Tree Genet. Genome*. 2013; 9: 867-877.
10. Hong SY, Cheon KS, Yoo KO, Lee HO, Cho KS, Suh JT, Kim SJ, Nam JH, Sohn HB, Kim YH. Complete chloroplast genome sequences and comparative analysis of *Chenopodium quinoa* and *C. album*. *Front Plant Sci*. 2017;8:1696.
11. Guo S, Guo LL, Zhao W, Xu, Li YY, Zhang XY, Shen XF, Wu ML, Hou XG. Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. *Molecules*. 2018;23:246.
12. Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL. A DNA barcode for land plants. *Proc Natl Acad Sci*. 2009;106(31):12794-12797.
13. Dong W, Liu J, Yu J, Wang L, Zhou S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One*. 2012;7(4):e35071.
14. Vallejo FC, Garcia-viguera C, Tomas-barberan FA. Changes in broccoli (*Brassica oleracea* var. *italica*) health-promoting compounds with inflorescence development. *J Agric Food Chem*. 2003;51:3776-3782.
15. Kumar P, Srivastava DK, Jang MW, Ha BJ. Effects of broccoli on anti-inflammation and anti-oxidation according to extraction solvents. *J Food Hyg Safety*. 2012;27:461-465.
16. Latté K P, Appel KE, Lampen A. Health benefits and possible risks of broccoli- an overview. *Food Chem Toxicol*. 2011;49:3287-3309.
17. Lee JJ, Shin HD, Lee YM, Kim AR, Lee MY. Effect of broccoli sprouts on cholesterol- lowering and anti-obesity effects in rats fed high fat diet. *J Korean Soc Food Sci Nutr*. 2009;38:309-318.
18. Finley JW. Reduction of cancer risk by consumption of selenium-enriched plants: enrichment of broccoli with selenium increases the anticarcinogenic properties of broccoli. *J Med Food*. 2003;6:19-26.
19. Kaur C, Kumar K, Anil D, Kapoor HC. Variations in antioxidant activity in broccoli (*Brassica oleracea* L.) cultivars. *J Food Biochem*. 2007;31:621-638.
20. Finley JW, Ip C, Lisk DJ, Davis CD, Hintze KG, Whanger PD. Cancer-protective properties of high-selenium broccoli. *J Agric Food Chem*. 2001;49:2679-2683.
21. Cartea ME, Francisco M, Soengas P, Velasco P. Phenolic compounds in *Brassica* vegetables. *Mol*. 2011;16:251-280.
22. Kumar P, Gaur A, Srivastava DK. Morphogenic response of leaf and petiole explants of broccoli using thidiazuron. *J Crop Improv*. 2015;29:432-446.
23. Jo JS, Bhandari SR, Kang GH, Lee JG. Comparative analysis of individual glucosinolates, phytochemicals, and antioxidant activities in broccoli breeding lines. *Hortic Environ Biotechnol*. 2016;57(4):392-403.
24. Li ZX, Liu YM, Fang ZY, Yang LM, Zhuang M, Zhang YY, Lv HH, Wang Y. Development status, existing problems and coping strategies of broccoli in China. *China vegetables*. 2019;(4):1 -5.
25. Kim K, Lee SC, Lee J, Yu Y, Yang K, Choi BS, Koh HJ, Waminal NE, Choi HI, Kim NH, et al. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Sci Rep*. 2015:515655.
26. Li PR, Zhang SJ, Li F, Zhang SF, Zhang H, Wang X, Sun RF, Bonnema G, Borm TJA. A phylogenetic analysis of chloroplast genomes elucidates the relationships of the six economically important *Brassica* species comprising the triangle of U. *Front Plant Sci*. 2017;8:11.
27. Gu CH, Ma L, Wu ZQ, Chen K, Wang YX. Comparative analyses of chloroplast genomes from 22 *Lythraceae* species: inferences for phylogenetic relationships and genome evolution within Myrtales. *BMC Plant Biol*. 2019;19:281.
28. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, L X, Zhang B, Xu J, Chen S. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. *Mol*. 2017;22:1330.

29. Guo S, Guo L, Zhao W, Xu J, Li Y, Zhang X, Shen X, Wu M, Hou X. Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. *Mol.* 2018;23:246.
30. Li X, Li YF, Zang MY, Li MZ, Fang YM. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *Int J Mol Sci.* 2018;19:2443.
31. Qian J, Song J, Gao H, Zhu Y, Xu J, Yao H, Sun C, Li X, Li C, et al. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. *PLoS ONE.* 2013;8(2): e57607.
32. Jansen RK, Raubeson LA, Boore JL, Depamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ. Methods for obtaining and analyzing whole chloroplast genome sequences. *Method Enzymol.* 2005; 395:348.
33. Xu J, Chu Y, Liao B, Xiao S, Yin Q, Bai R, Su H, Dong L, Li X, Qian J, et al. *Panax ginseng* genome examination for ginsenoside biosynthesis. *Gigasci.* 2017;6:1-15.
34. Niu D, Yang Y. Why eukaryotic cells use introns to enhance gene expression: Splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. *Biol Direct.* 2011;6:24.
35. Le H, Nott A, Moore MJ. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* 2003;28:215-220.
36. Li YF, Sylvester SP, Li M, Zhang C, Li X, Duan YF, Wang XR. The complete plastid genome of *Magnolia zenii* and genetic comparison to Magnoliaceae species. *Mol.* 2019;24:261.
37. Redwan RM, Saidin A, Kumar SV. Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. *BMC Plant Biol.* 2015;15:196.
38. Li B, Lin F, Huang P, Guo W, Zheng Y. Complete chloroplast genome sequence of *Decaisnea insignis*: genome organization, genomic resources and comparative analysis. *Sci Rep.* 2017;7:10073.
39. Wang W, Yu H, Wang J, Lei W, Gao J, Qiu X, Wang J. The complete chloroplast genome sequences of the medicinal plant *forsythia suspensa* (Oleaceae). *Int J Mol Sci.* 2017;18 (11):2288.
40. Yue F, Cui L, dePamphilis CW, Moret BM, Tang J. Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. *BMC Genomics* 2008;9 Suppl 1: S25.
41. Haberle RC, Fourcade HM, Boore JL, Jansen RK. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J Mol Evol* 2008;66: 350-361.
42. Pombert JF, Otis C, Lemieux C, Turmel M. The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. *Mol Biol Evol.* 2005;22: 1903-1918.
43. Doorduyn L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Res.* 2011;18(2):93-105.
44. He S, Wang Y, Volis S, Li D, Yi T. Genetic diversity and population structure: implications for conservation of wild soybean (*Glycine soja* Sieb. et Zucc) based on nuclear and chloroplast microsatellite variation. *Int J Mol Sci.* 2012;13:12608-12628.
45. Xue J, Wang S, Zhou SL. Polymorphic chloroplast microsatellite loci in *Nelumbo* (Nelumbonaceae). *Am J Bot.* 2012;99 (6):240-244.
46. Kofler R, Schlötterer C, Luschützky E, Lelley T. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. *BMC Genom.* 2008; 9:612.
47. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol Biol.* 2008;8:36.

48. Hu ZY, Hua W, Huang SM, Wang HZ. Complete chloroplast genome sequence of rapeseed (*Brassica napus* L.) and its evolutionary implications. *Genet Resour Crop Evol.* 2011;58:875-887.
49. Li X, Gao H, Wang Y, Song J, Henry R, Wu H, Hu Z, Yao H, Luo H, Luo K, et al. Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species. *Sci China Life Sci.* 2013;56:189-198.
50. Yang KW, Nath UK, Biswas MK, Kayum MA, Yi G, Lee JH, Yang TJ, Nou IS. Whole-genome sequencing of *Brassica oleracea* var. *capitata* reveals new diversity of the mitogenome. *PLOS ONE.* 2018; 13(3): e0194356.
51. Tong W, Kim TS, Park YJ. Rice chloroplast genome variation architecture and phylogenetic dissection in diverse *Oryza* species assessed by whole-genome resequencing. *Rice.* 2016;9:57.
52. Bolger A, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinform.* 2014;30:2114-2120.
53. Langmead B, Salzberg S. Fast gapped-read alignment with bowtie 2. *Nat Methods.* 2012;9:357 -359.
54. Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov AS, Lesin V, Nikolenko S, Pham S, Pribelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455-477.
55. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding preassembled contigs using SSPACE. *Bioinform.* 2011;27:578-579.
56. Liu C, Shi LC, Zhu Y, Chen HM, Zhang JH, Lin XH, Guan XJ. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genom.* 2012;13:715.
57. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organelle genomes with DOGMA. *Bioinform.* 2004;20:3252-3255.
58. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 2005;33:W686.
59. Lohse M, Drechsel O, Kahlau S, Bock R. Organellar genome DRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 2013;41:W575-W581.
60. Peden JF. Analysis of codon usage. *Biosystems.* 1999;5:45-50.
61. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 2001;29: 4633- 4642.
62. Mudunuri SB, Nagarajaram HA. IMEx: imperfect microsatellite extractor. *Bioinform.* 2007;23: 1181-1187.
63. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 2004;32 (suppl_2):W273-W279.
64. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005;33:511-518.
65. Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximumlikelihood trees for large alignments. *PloS One.* 2010;5:e9490.

Figures

Figure 1

Physical map of the *B. oleracea* var. *italica* cp genome. The quadripartite circular structure contains a pair of IRA and IRB, separated by a SSC region and a LSC region. Different color-coded suggest different functional gene groups. The darker gray in the inner circle corresponds to GC content, whereas the lighter grey indicates AT content. Genes drawn outside of the map are transcribed counter clockwise and inside of the circle are transcribed clockwise.

Figure 2

Blast result of comparison of cp genome structure and GC skew between broccoli and 8 species.

Figure 4

Statistics of length of repeat and the number of repeat sequences in the cp genome of broccoli.

Figure 6

Phylogenetic tree inferred by ML method based on the complete cp genomes from 9 species. The bootstrap support values are shown at the nodes.