RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2020.063

ScienceAsia 46 (2020): **520–529**

The complete chloroplast genome sequence of carmine radish (*Raphanus sativus* L.) and its evolutionary implications

Jian Gao[a,b], Fabo Chen[a,b], Shumei Shang[a,b], Jun Shu[a], Yi Liu[a], Wenbo Li[a,b,*]

[a] Department of Life Sciences and Technology, Yangtze Normal University, Fuling 408100 China
[b] Centre for Green Development and Collaborative Innovation in Wuling Mountain Region, Yangtze Normal University, Fuling 408100 China

*Corresponding author, e-mail: liverb@163.com

Received 12 Jul 2019
Accepted 16 Jun 2020

ABSTRACT: Carmine radish is well known for its natural red pigment (red radish pigment) and is produced in Chongqing, China. Here, the chloroplast (cp) genome of the local carmine radish cultivar 'Hongxin 1' was identified through *de novo* assembly on a third-generation sequencing platform. The results showed that the Hongxin radish cp genome of 153 419 bp consists of four different regions: two inverted repeated regions (IRa and IRb, 26 986 bp each), a large single-copy region (LSC) of 88 448 bp, and a small single-copy region (SSC) of 17 814 bp. In addition, a number of genes (125) were identified, comprising 88 predicted protein-coding genes, 29 tRNA genes and 8 rRNA genes. Compared with the radish cultivar WK10039 cp genome, one predicted protein-coding gene was identified, but 8 tRNA genes were not found in the carmine radish chloroplast genome. Moreover, 12 forward and 14 inverted repeats were identified as well as 58 simple sequence repeats (SSRs). In addition, only slight differences were found between *Raphanus sativus* (Keroan) and carmine radish (Hongxin), except that ycf15 was only detected in *R. sativus* (Hongxin). Phylogenetic analysis of the 50 protein-coding genes from the cp genome sequences of 30 *Brassicaceae* species indicated that *R. sativus* (Hongxin) is most closely related to *Brassica napus* and *Brassica nigra*. This study is the first to generate a valuable resource for SSR marker development studies in carmine radish and provides a basis for further studies on genomics and functional genomics in this type of radish.

KEYWORDS: radish, chloroplast genome, simple sequence repeats, phylogenetic analysis

INTRODUCTION

Radish (*Raphanus sativus* L.) is an annual member of the Brassicaceae family and is produced around the world. Multiple colors of the flesh of the taproot have been identified, comprising white, yellow, red, black and purple. The carmine radish cultivar 'Hongxin' is well known for containing a natural red pigment (red radish pigment); this cultivar was produced in Chongqing Fuling and is named for three specialities of Fuling. Based on the phylogenetic origin and conserved relationships of radishes in the tribe Brassiceae, a previous study showed that radish presents a more highly conserved relationship with *Arabidopsis thaliana* than with other *Brassica* species [1]. However, subsequent cp and nuclear sequence analyses showed some discordant results regarding the phylogenetic relationships between different *Brassica* species. On the basis of the analysis of several cp genomes using restriction site polymorphism, Warwick and Black showed that radish belongs to the *Brassica rapa*/*Brassica oleracea* lineage of subtribe Brassicinae [1]. In contrast, radish was found to present a more conserved relationship with *Brassica nigra* than *B. rapa*/*B. oleracea* using a nuclear DNA marker [2]. Moreover, studies have demonstrated that cultivated radishes might exhibit multiple origins according to systematic research on diverse accessions of *Raphanus* species [3–5]. However, the cp genome of carmine radish has not been fully elucidated, and comparative analyses between the cp genomes of carmine radish and *R. sativus* (Keroan) are lacking.

In recent years, protein-coding genes as well as intergenic regions identified from conserved regions have been used to elucidate phylogenetic relationships; such regions include the 5′-matK region used to examine the conservation of cultivated radish [4], the trnL-rpl32 intergenic region used to investigate interspecific hybridization between the wild species

(*Raphanus raphanistrum*) and cultivated radish [6] and simple sequence repeats (SSRs) evaluated in 82 different *Raphanus* species [5]. Therefore, the complete carmine radish cp genome would be helpful for the assessment of phylogenetic origins and conserved relationships. However, compared with *Arabidopsis* or *Brassica* genomes, there is less reported information about the carmine radish cp genome. In recent years, draft cp genome sequences of wild radish [7] and Japanese radish [8] as well as a Korean radish cultivar [9] have been reported. Considering the differences between the local Hongxin radish cultivar and the Korean radish cultivar, the *de novo* assembly of the cp genome of carmine radish was conducted from high-quality data (37.67 Gb) extracted from raw sequence data (38.04 Gb) using third-generation sequencing in the present study. The dataset for the Hongxin radish cp genome will be helpful for conducting comparative genomic analyses with other *Brassicaceae* species, thereby elucidating the conserved relationships of radish with *B. nigra* or *B. Rapa*/*B. oleracea*.

MATERIALS AND METHODS

Plant materials and DNA extraction

Fresh leaves of Hongxin radish (which is well known for containing a natural red pigment and is produced in Chongqing Fuling; the radish is named for three specialties of Fuling) were collected at the FulingBreeding Base of Yangtze Normal University in China. Total genomic DNA was isolated from the fresh leaves using the MagicMag Genomic DNA Micro Kit following the manufacturer's protocol (Sangon Biotech Co., Shanghai, China). Subsequently, the integrity and quality of the DNA were checked through agarose gel electrophoresis and spectrophotometry using a Nanodrop 2000 instrument (Thermo Scientific, DE, USA), respectively.

Sequence assembly and annotation of Hongxin radish

Library construction was performed at the Breeding Company (Shanxi, China) on a third generation sequencing platform. After filtering by using PRINSEQ lite v0.20.4 software (parameters of phredQ, 20; length, 50), the high-quality data (37.67 Gb) were extracted from the raw sequence data (38.04 Gb); subsequently, *de novo* assembly of the cp genome was conducted using NOVOPlasty with the K value set to 31 [10]. Seeds and the reference plastome were obtained from *R. sativus* (NC_024469.1) (www.ncbi.nlm.nih.gov/nuccore/NC_024469.1). Finally, one contig of the Hongxin radish cp genome was generated, identified through mapping against the reference plastome (NC_024469.1) using GENEIOUS 8.1 [11] and annotated with CpGAVAS [12] software (http://47.96.249.172:16019/analyzer/home) and DOGMA software. Thereafter, we confirmed the tRNA genes at the tRNAscan-SE server (lowelab.ucsc.edu/tRNAscan-SE/) [13] and we drew the physical circular map of the cp genome using the OGDRAW program [14] with minor manual corrections.

Genome comparison and gene rearrangement

The chloroplast genome structure consisting of the large single copy (LSC) region, small single copy (SSC) region and two reversed duplicate regions (IRA and IRB) showed differences between different species. In this study, we compared the structure of the cp genome among eight representative species from the *Brassicaceae* order. Additionally, gene rearrangements were determined via the alignment of seven cp genomes with a single reference genome using Mauve v.4.0 [15]. Moreover, using the *A. thaliana* annotation as a reference, pairwise alignments among 8 cp *Brassicaceae* genomes were conducted with the mVISTA program [16] in LAGAN mode [17].

Repeat and SSR analysis of the Hongxin radish cp genome

The locations and sizes of long repetitive repeat sequences consisting of forward, palindromic, complementary and reverse repeats were analyzed by using the REPuter program (bibiserv.cebitec.uni-bielefeld.de/reputer) [18]. The parameter settings of a repeat size > 30 bp, sequence identity $\geqslant 90\%$ and Hamming distance (3) were used for the identification of long repetitive repeats. The online software MIcroSAtellite (MISA) [19] was employed to identify SSRs using the following parameter settings: $\geqslant 8$ mononucleotide SSR motifs; $\geqslant 5$ dinucleotide SSR motifs; $\geqslant 4$ trinucleotide SSR motifs; and $\geqslant 3$ tetranucleotide, pentanucleotide, and hexanucleotide SSR motifs.

Codon usage and RNA editing sites

To detect the deviation of the use of synonymous codons, codon W1.4.2 (downloads.fyxm.net/CodonW-76666.html) was selected and used to examine the effect of the amino acid composition according to relative synonymous codon usage (RSCU). Finally, possible RNA editing sites in

R. sativus (Hongxin) protein-coding genes were predicted using the Predictive RNA Editor for Plants (PREP) suite [20] with the cutoff value set to 0.8.

Phylogenetic analysis

Fifty protein-coding genes from the cp genome sequences of 30 *Brassicaceae* species and the *Vitis vinifera* (Vitales) cp genome (as an outgroup) obtained from GenBank were used for phylogenetic reconstruction (Table S1). We used GENEIOUS v8.0.2 for the alignment of their protein-coding sequences [11]. RAxML version 8.0.20 was selected for maximum likelihood (ML) analysis with 1000 replicates for bootstrap testing [21], and jModelTest v2.1.7 was used for the best substitution model (GTR+I+G) [22].

RESULTS

Genome content and organization of the cp genome in carmine radish

In this study, we generated approximately 37.6 Gb of clean data for Hongxin radish. The cp genome (153 419 bp) was assembled with a high mean coverage (almost 2174X) for carmine radish, and this genome was slightly longer than that of the radish cultivar WK10039 (153 368 bp). The structure and organization of the cp genomes are shown in Fig. 1 and Table S2. We found that the Hongxin radish cp genome of 153 419 bp in length is divided into 4 regions, including two inverted repeated regions (IRa and IRb, 26 986 bp each), a large single-copy region (LSC) of 88 448 bp and a small single-copy region (SSC) of 17 814 bp. The overall GC content is almost 36.3% for Hongxin radish. The analysis of the gene content revealed 125 genes (88 protein-coding genes, 29 tRNA genes and 8 ribosomal RNA genes). Compared with the radish cultivar WK10039 cp genome, there was one additional predicted protein-coding gene, but 8 tRNA genes were not found in the carmine radish chloroplast genome. Among the identified genes, 16 genes were duplicated in the IR regions, including 11 protein-coding genes (rpl2, rpl23, ycf2, ycf15, ndhB, rps7, rps12, rrn16S, rrn23S, rrn5S and rrn4.5S) and 5 tRNA genes (trnI-CAT, trnL-CAA, trnV-GAC, trnR-ACG and trnN-GTT). Among these genes, 9 genes and 5 tRNA genes contained one intron, while two genes (rpl2 and ndhB) contained two introns (Table 1 and Fig. 1). Subsequently, the frequency of codon usage was estimated for the Hongxin radish cp genome from the sequences of protein-coding and tRNA genes, and the results are summarized in Table S3. Taken together, the results showed that the genes of the Hongxin radish cp genome consisted of a total of 50 898 codons. Among these genes, leucine (Leu), encoded by 5345 codons, and Trp, encoded by 1067 codons, were the most and least frequent amino acids, respectively, encoded by the cp genome (Table S3). However, in the radish cultivar WK10039 cp genome, leucine (Leu) (2814 codons) and Met (599 codons) were identified as the most and least frequent amino acids, respectively.

Long repeat and SSR analysis

Based on repeat structure analysis, we identified 14 inverted and 12 forward repeats in the Hongxin radish cp genome (Table S4). The lengths of these repeats ranged from 24–35 bp, but one of the intergenic spacers (IGSs) was found to be the longest inverted repeat (72 bp). In the LSC region, two repeats were found to be related to the ycf1 and rps16 genes (no. 6 and 9), respectively, and 7 inverted and 8 forward repeats were found to exist in the intergenic spacers (IGSs) of the LSC region. In addition, 4 inverted repeats (no. 20, no. 24–26) in IGS were identified in the SSC region. Moreover, 2 inverted and 1 forward repeat related to trnS-TGA (intron) and 2 forward and 3 inverted repeats (no. 7 and 10, no. 16, no. 18 and 20) associated with intergenic spacers (IGSs) were also identified. In addition, a total of 58 SSRs were identified in the Hongxin radish cp genome. Most of these SSRs (30 SSRs) were found to be distributed in the LSC region, while 20 SSRs and 8 SSRs were identified as being distributed in the IR and SCC regions, respectively. These SSRs comprised 43 mononucleotide SSRs (74.14%), 7 dinucleotide SSRs (12.07%), and 8 other types of SSRs (15.09%) (Table S5). In contrast, in the radish cultivar WK10039 cp genome, 58 mononucleotide SSRs, 21 dinucleotide SSRs and 12 other types of SSRs were identified. In addition, only 13 SSRs were located in genes (ycf1, ccsA, rpoC2, rpoB, clpP, rpoA), and 45 SSR loci were found in intergenic regions. More interestingly, 41 of 43 mononucleotide SSRs were identified as belonging to the A/T type.

Comparative genomic analysis of cp genomes in *Brassicaceae*

To demonstrate the divergent sequences of the cp genome among related species in Brassicaceae, the pairwise comparison of cp genomes between Hongxin radish and the 7 other Brassicaceae cp genomes was conducted via comparative genome

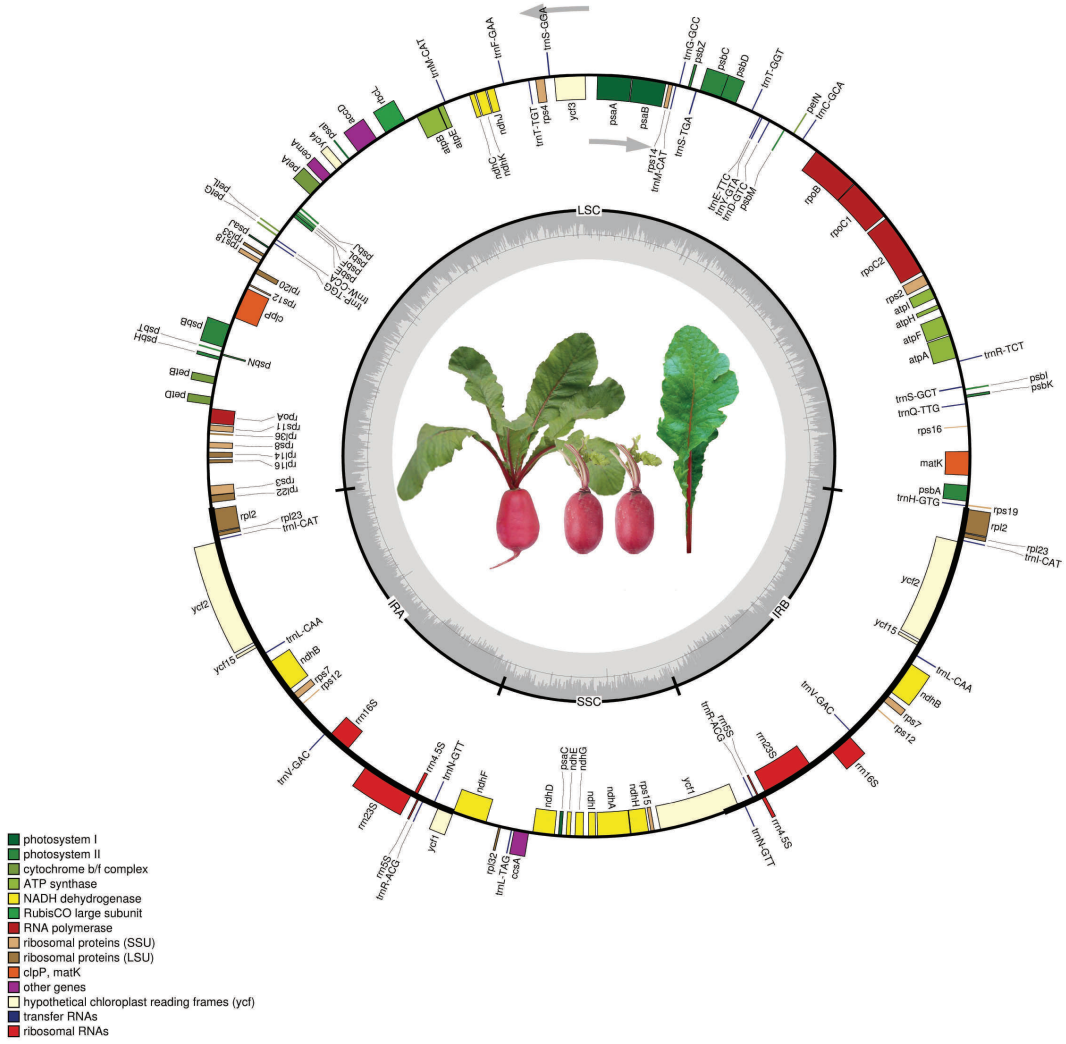

Fig. 1 Circular map of the cp genome of Hongxin radish with annotated genes. Genes inside and outside the circle are transcribed clockwise and counterclockwise, respectively. Genes are color coded following their functional groups. The boundaries of the small (SSC) and large (LSC) single-copy regions and inverted repeat (IRa and IRb) regions are noted in the inner circle. The photograph of Hongxin radish was taken in our lab.

analysis by using mVISTA software, with the *A. thaliana* annotation as a reference (Fig. 2). Compared with the LSC and SSC regions, two IR regions were identified as less divergent. In addition, the coding regions were found to be more conserved than the noncoding regions within the LSC and SSC regions. Moreover, we found that IR regions (gene order and number) were highly conserved in all 8 cp genomes from Brassicaceae except for the single-copy region junction. Intergenic spacers were contained within highly divergent regions among the 8 cp genomes, such as ycf1-rps32 and ndhI-ndhG in the SSC region and trnH-psbA, trnY-GUA-trnE-UUC, trnE-UUC-rpoB, trnV-UAC-ndhC, trnC-GCA-petN, psbM-petN, rpl32-trnL-UAG, rbcL-accD and accD-psaI in the LSC region. However, the coding regions of the ndhB, ycf15, ycf1 and ycf2 genes were identified as more divergent in 8 Brassicaceae cp

Table 1 Chloroplast genome gene content and functional classification for Hongxin *R. sativus* L.

Function	Family name	*R. sativus* L.
Genes for photosynthesis	Subunits of ATP synthase	atpA, atpB, atpE, atpF, atpH, atpI
	Subunits of NADH dehydrogenase	ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	Subunits of cytochrome	petA, petB, petD, petG, petL, petN
	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ
	Subunits of photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbk, psbL, psbM, psbN, psbT, psbZ
	Subunit of rubisco	rbcL
Self replication	Large subunit of ribosome	rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36
	DNA dependent RNA polymerase	rpoA, rpoB, rpoC1, rpoC2
	Small subunit of ribosome	rps18, rps19, rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps16
	rRNA Genes	rrn16S, rrn23S, rrn4.5S, rrn5S
	tRNA Genes	trnI-CAT, trnL-CAA, trnV-GAC, trnR-ACG, trnL-TAG trnN-GTT, trnR-ACG, trnV-GAC, trnI-CAT, trnH-GTG trnQ-TTG, trnS-GCT, trnC-GCA, trnD-GTC, trnY-GTA trnE-TTC, trnR-TCT, trnT-GGT, trnS-TGA, trnM-CAT trnN-GTT, trnW-CCA, trnL-CAA, trnP-TGG, trnG-GCC trnM-CAT, trnS-GGA, trnT-TGT, trnF-GAAv
Other genes	Subunit of Acetyl-CoA-carboxylase	accD
	c-type cytochrome synthesis gene	ccsA
	Envelop membrane protein	cemA
	Protease	clpP
	Maturase	matK
Genes of unkown function	Conserved open reading frames	ycf1, ycf15, ycf2, ycf3, ycf4

genomes. In addition, only slight differences were found between *R. Sativus* (Keroan) and *R. Sativus* (Hongxin) in this study, except that ycf15 was only detected in *R. Sativus* (Hongxin).

IR contraction and expansion in the *R. sativus* cp genome

In this study, the IR-SSC and IR-LSC boundaries among 8 Brassicaceae cp genomes (*A. thaliana*, *B. napus*, *B. rapa*, *Pugionium comutum*, *Cakile arabica*, *Cochlearia tridactylites*, *R. sativus* (Keroan) and *R. sativus* (Hongxin)) were compared and depicted in detail (Fig. 3). The ndhF gene was found to overlap the IRb/SSC border by 39 bp in all of the Brassicaceae cp genomes except for that of *B. rapa* (36 bp). More importantly, the LSC region was found to be longer in the *R. sativus* (Hongxin) cp genome than in the other Brassicaceae cp genomes. In addition, due to the expansion of the LSC region, the rps19 gene was also identified in the LSC region. Moreover, the comparison of cp genome size between different Brassicaceae cp genomes showed that the IR region (24 850 bp) was longer in the *R. sativus* (Hongxin) cp genome than in the other Brassicaceae cp genomes.

Phylogenetic analysis

In this research, based on the 50 protein-coding genes obtained from the cp genome sequences of 30 Brassicaceae species (Table S1), a phylogenetic tree was constructed, which indicated that *R. sativus* (Hongxin) is most closely related to *B. napus* and *B. nigra* (especially *B. napus*), and the *V. vinifera* (Vitales) cp genome was selected as an outgroup (Fig. 4). We propose that conflicts regarding phylogenetic relationships can be resolved by using

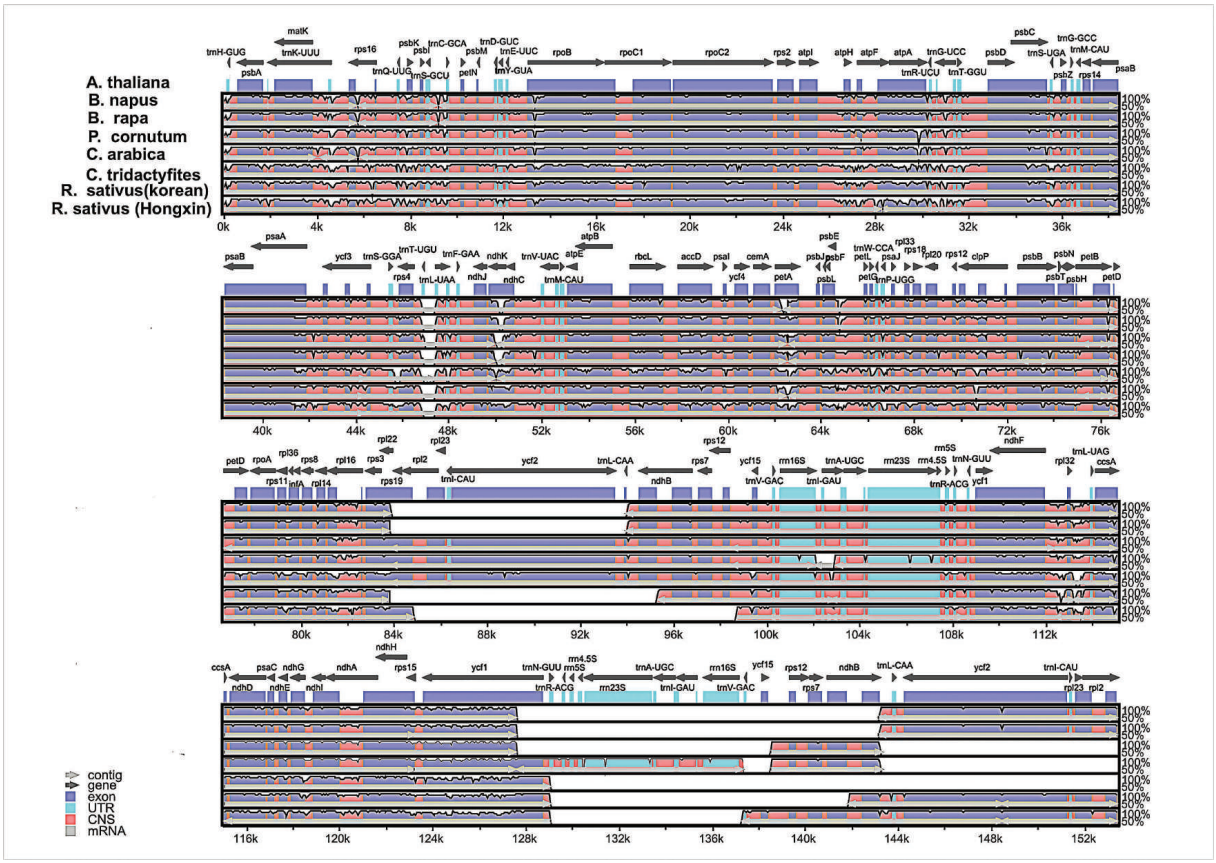

Fig. 2 Comparison of 8 cp genomes among related species of Brassicaceae using mVISTA. Gray arrows and thick black lines above the alignment indicate gene orientation. Purple bars, blue bars and pink bars represent exons, UTRs and noncoding sequences (CNS), respectively. The scale of the Y-axis represents the percent identity (50–100%). Genome regions are color coded as protein-coding exons, rRNAs, tRNAs or conserved noncoding sequences (CNS).

complete cp genome information, especially for cp coding genes, which could be useful for phylogenetic analyses in many closely related species and populations.

DISCUSSION

In this study, the complete nucleotide sequence of the Hongxin radish cp genome was identified through *de novo* assembly on a third-generation sequencing platform. Cp genome evolution was analyzed through comparative analysis of the Hongxin cp genome and other species of order *Brassicaceae*. Moreover, the sequences identified in this study will be helpful for further evolutionary studies in *Brassicaceae* species.

Although gene content and organization were generally found to be similar within the *Brassicaceae* species, coding regions were more conserved than noncoding regions within the LSC and SSC regions. However, the most highly differentiated regions corresponded to the ndhD, ndhF, trnH-psbA, ycf1, ndhK, matK, rpl32 and rps15 genes. The differences in these gene regions were also demonstrated in a previous study [23]. In addition, we found that IR regions were highly conserved (in terms of gene order and numbers) in all 8 cp genomes from Brassicaceae except at the single-copy region junction. Moreover, intergenic spacers were contained within highly divergent regions among the 8 cp genomes, as observed for ycf1-rps32 and ndhI-ndhG in the SSC region and trnH-psbA, trnY-GUA-trnE-UUC, trnE-UUC-rpoB, trnV-UAC-ndhC, trnC-GCA-petN, psbM-petN, rpl32-trnL-UAG, rbcL-accD and accD-psaI in the LSC region. Similar results have been obtained in the cp genomes of other plants [24, 25]. In addition, the coding regions of ndhB, ycf15, ycf1 and ycf2 genes were identified as more divergent in 8 Brassicaceae cp genomes. Only slight differences were found between *R. Sativus* (radish cultivar WK10039) and

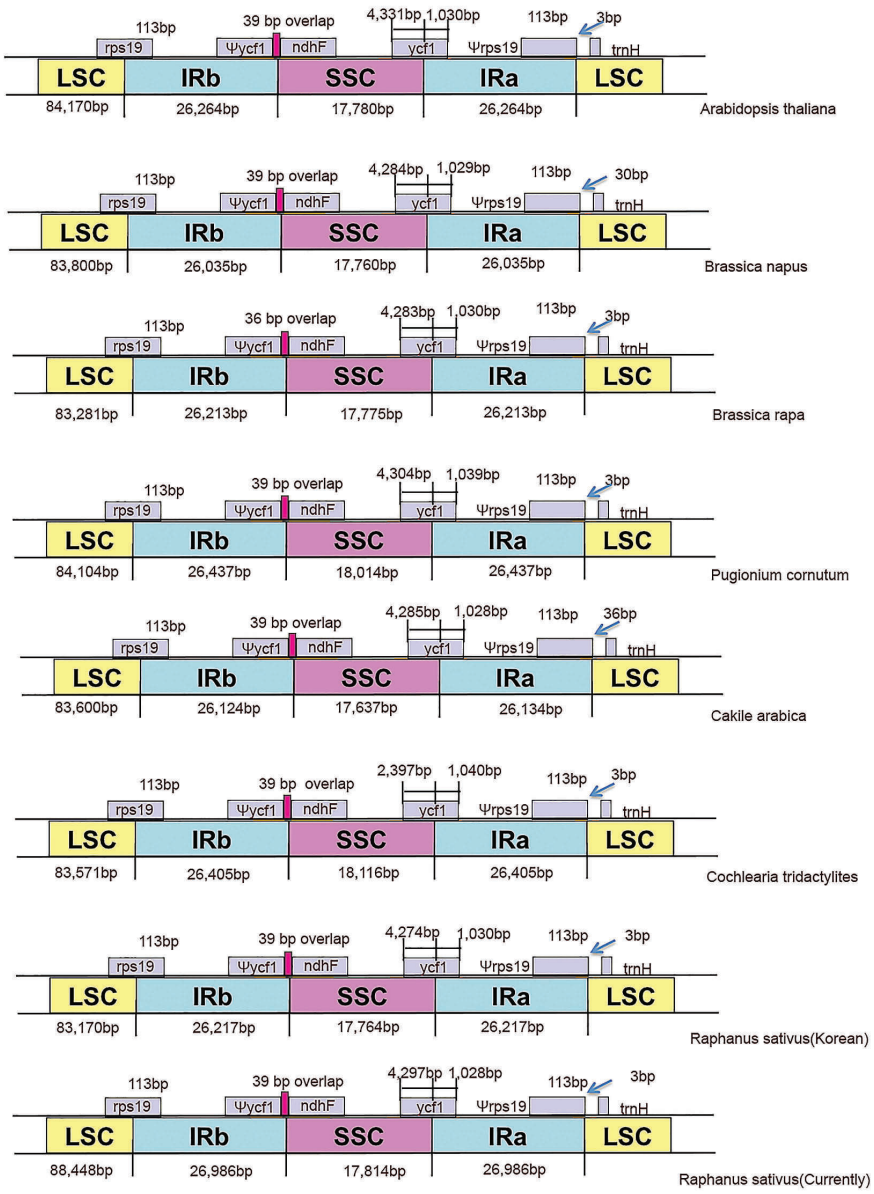

Fig. 3 Comparison of the borders of the LSC, SSC and IR regions among 8 cp genomes of related species of Brassicaceae. Y: pseudogenes, /: distance from the edge.

R. Sativus (Hongxin) in this study, except that ycf15 was only detected in *R. Sativus* (Hongxin).

The locations of the boundaries between cp regions can be used to further evaluate the cp genome. Previous studies have demonstrated that contractions and expansions are common at the borders of the intergenic regions (IRs) of cp genomes, which are used for identifying differences in size between cp genomes [26, 27]. In this study, the IR-SSC and IR-LSC boundaries among 8 Brassicaceae cp genomes (*A. thaliana*, *B. napus*, *B. rapa*, *P. comutum*, *C. arabica*, *C. tridactylites*, *R. sativus* (Keroan) and *R. sativus* (Hongxin)) were compared and depicted in detail. Among these boundaries, the IRb/SSC border (located between the ycf1 pseudogene and the ndhF gene) has been selected as an indicator for analyzing cp genome variation in higher plants and algae, and the ndhF gene was found to overlap the IRb/SSC border by 39 bp in all of the Brassicaceae cp genomes except for that of *B. rapa* (36 bp).

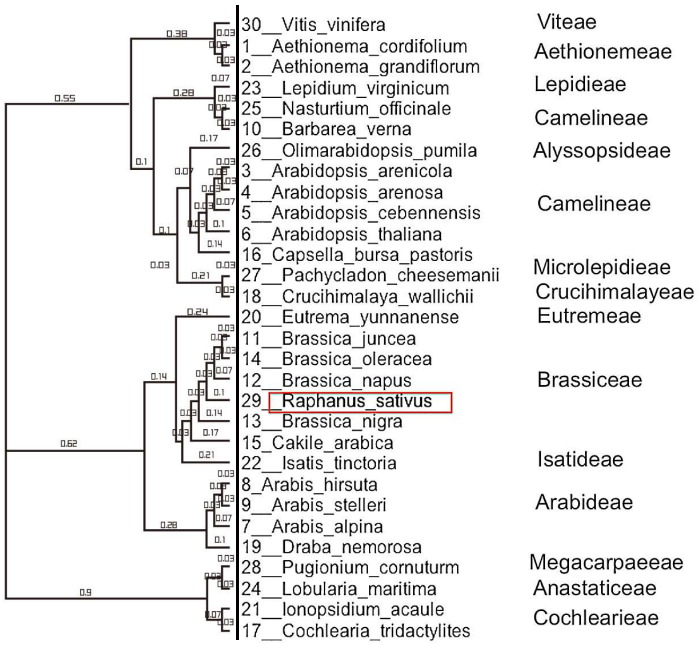

Fig. 4 ML phylogenetic tree reconstruction of 30 taxa of the Brassicaceae clade based on concatenated sequences from 50 cp protein-coding genes. The position of *R. sativus* (Hongxin) is indicated by the rectangular box. The *V. vinifera* (Vitales) cp genome was selected as an outgroup.

Repetitive sequences have been reported in evolutionary and population genetic studies of many plant lineages involving the cp genome. SSRs (1–6 bp) are widely distributed throughout the genome as tandemly repeated DNA sequences. cpSSRs have been widely used for analyses of the diversity, structure and differentiation of plant populations [28, 29]. Only 13 SSRs were found to be located in genes (ycf1, ccsA, rpoC2, rpoB, clpP, and rpoA), and 45 SSR loci were found in intergenic regions. More interestingly, 41 of 43 mononucleotide SSRs were identified as belonging to the A/T type. These results are in perfect accord with the previous hypothesis that cpSSRs rarely contain tandem guanine (G) or cytosine (C) repeats but are composed of short polyadenine (polyA) or polythymine (polyT) repeats. cpSSRs can be used for genetic diversity analysis, species identification, and analyses of species evolution [30]. In previous studies, phylogenetic analysis has been conducted using intergenic regions or protein-coding genes [31]. These regions include the 5′-matK region, used for the assessment of the conservation of cultivated radish [4]; the trnL-rpl32 intergenic region, used for investigating interspecific hybridization between the wild species (*R. raphanistrum*) and cultivated radish [6]; and simple sequence repeats (SSRs) evaluated in 82 different *Raphanus* species [5]. However, analyses of these regions cannot sufficiently represent phylogenetic relationships owing to their differences between groups. However, with the development of large-scale DNA sequencing methods, the entire cp genome can now be used as an indicator of plant phylogenetics and population genetics. In this study, based on 50 protein-coding genes collected from the cp genome sequences of 30 *Brassicaceae* species, a phylogenetic tree was constructed, which indicated that *R. sativus* (Hongxin) is most closely related to *B. napus* and *B. nigra* (especially *B. napus*). We propose that phylogenetic relationship conflicts can be resolved by using complete cp genome information, especially for cp coding genes, which could be useful for thorough phylogenetic analyses in closely related species and populations.

Appendix A. Supplementary data

Supplementary data associated with this article can be found at http://dx.doi.org/10.2306/scienceasia1513-1874.2020.063.

Acknowledgements: This study was funded by the Chongqing Research Program of Basic Research and Frontier Technology (KJ1601205), the Normal Program Frontiers and Applications Research of CSTC (No. cstc2016jcyjA0136), the Science and Technology Re-

search Program of the Chongqing Municipal Education Commission (KJ1712302), the Fuling District Science and Technology Plan Projects (FLKJ,2018BBB3009), and the Special Project of Science and Technology Innovation for Social Undertaking and People's Livelihood Guarantee of CSTC (No. cstc2017shms-xdny80074).

REFERENCES

1. Warwick SI, Black LD (1991) Molecular systematics of *Brassica* and allied genera (Subtribe Brassicinae, Brassiceae)–chloroplast genome and cytodeme congruence. *Theor Appl Genet* **82**, 81–92.
2. Yang YW, Lai KN, Tai PY, Ma DP, Li WH (1999) Molecular phylogenetic studies of Brassica, rorippa, arabidopsis and allied genera based on the internal transcribed spacer region of 18S–25S rDNA. *Mol Phylogenet Evol* **13**, 455–462.
3. Yamagishi H, Terachi T, Ozaki A, Ishibashi A (2009) Inter- and intraspecific sequence variations of the chloroplast genome in wild and cultivated Raphanus. *Plant Breed* **128**, 172–177.
4. Yamane K, Lü N, Ohnishi O (2005) Chloroplast DNA variations of cultivated radish and its wild relatives. *Plant Sci* **168(3)**, 627–634.
5. Yamane K, Lü N, Ohnishi O (2009) Multiple origins and high genetic diversity of cultivated radish inferred from polymorphism in chloroplast simple sequence repeats. *Breed Sci* **59**, 55–65.
6. Ridley CE, Kim SC, Ellstrand NC (2008) Bidirectional history of hybridization in California wild radish, *Raphanus sativus* (Brassicaceae), as revealed by chloroplast DNA. *Am J Bot* **95**, 1437–1442.
7. Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, Tonosaki K, Shirasawa S, et al (2014) Draft sequences of the radish (*Raphanus sativus* L.) genome. *DNA Res* **21**, 481–490.
8. Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu SH (2014) Consequences of whole–genome triplication as revealed by comparative genomic analyses of the wild radish *Raphanus raphanistrum* and three other brassicaceae species. *Plant Cell* **26**, 1925–1937.
9. Jeong YM, Chung WH, Mun JH, Kim N, Yu HJ (2014) *De novo* assembly and characterization of the complete chloroplast genome of radish (*Raphanus sativus* L.). *Gene* **551**, 39–48.
10. Dierckxsens N, Mardulyn P, Smits G (2017) NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res* **45**, ID e18.
11. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Alex Cooper A, et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649.
12. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, **13**, ID 715.
13. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, 686–689.
14. Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet* **52**, 267–274.
15. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**, ID R12.
16. Mayer C, Bredno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047.
17. Frazer,KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, 273–279.
18. Kurtz S (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* **29**, 4633–4642.
19. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585.
20. Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res* **37**, 253–259.
21. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.
22. Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, ID 772.
23. Kumar S, Hahn FM, McMahan CM, Cornish K, Whalen MC (2009) Comparative analysis of the complete sequence of the plastid genome of *Parthenium argentatum* and identification of DNA barcodes to differentiate *Parthenium* species and lines. *BMC Plant Biol* **9**, ID 131.
24. Nie X, Lv S, Zhang Y, Du X, Wang L, Biradar SS, Tan X, Wan F, et al (2012) Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). *PLoS One* **7**, ID e36869.
25. Ni L, Zhao Z, Xu H, Chen S, Dorje G (2016) The complete chloroplast genome of *Gentiana straminea* (Gentianaceae), an endemic species to the Sino-Himalayan subregion. *Gene* **577**, 281–288.
26. Raubeson LA, Peery R., Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new

sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* **8**, ID 174.
27. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol Biol* **8**, ID 36.
28. Bryan GJ, McNicoll J, Ramsay G, Meyer RC, De Jong WS (1999) Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants. *Theor Appl Genet* **99**, 859–867.
29. Provan J (2000) Novel chloroplast microsatellites reveal cytoplasmic variation in *Arabidopsis thaliana*. *Mol Ecol* **9**, 2183–2185.
30. Daniel E, Rod P (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Mol Ecol Resour* **9**, 673–690.
31. Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ (2011) Understanding angiosperm diversification using small and large phylogenetic trees. *Am J Bot* **98**, 404–414.

Appendix A. Supplementary data

Table S1 Taxa included in the phylogenetic analyses of cpDNA with Genbank accession numbers.

Taxon	Tribal	Family	Order	Accession no.
Aethionema cordifolium	Aethionemeae	Brassicaceae	Brassicales	NC_009265.1
Aethionema grandiflorum	Aethionemeae	Brassicaceae	Brassicales	NC_009266
Arabidopsis arenicola	Camelineae	Brassicaceae	Brassicales	NC_030346
Arabidopsis arenosa	Camelineae	Brassicaceae	Brassicales	NC_029334
Arabidopsis cebennensis	Camelineae	Brassicaceae	Brassicales	NC_029335
Arabidopsis thaliana	Camelineae	Brassicaceae	Brassicales	NC_000932.1
Capsella bursa pastoris	Camelineae	Brassicaceae	Brassicales	NC_009270.4
Olimarabidopsis pumila	Alyssopsideae	Brassicaceae	Brassicales	NC_009267.1
Pachycladon cheesemanii	Microlepidieae	Brassicaceae	Brassicales	NC_021102
Crucihimalaya wallichii	Crucihimalayeae	Brassicaceae	Brassicales	NC_009271.1
Barbarea verna	Cardamineae	Brassicaceae	Brassicales	NC_009269
Nasturium officinale	Cardamineae	Brassicaceae	Brassicales	NC_009275
Lepidium virginicum	Lepidieae	Brassicaceae	Brassicales	NC_023092
Eutrema yunnanense	Eutremeae	Brassicaceae	Brassicales	NC_008115
Brassica juncea	Brassiceae	Brassicaceae	Brassicales	NC_028272
Brassica oleracea	Brassiceae	Brassicaceae	Brassicales	KR233156
Brassica napus	Brassiceae	Brassicaceae	Brassicales	NC016734
Brassica nigra	Brassiceae	Brassicaceae	Brassicales	KT878383
Cakile arabica	Brassiceae	Brassicaceae	Brassicales	NC030775
Isatis tinctoria	Isatideae	Brassicaceae	Brassicales	NC028415
Arabis hirsuta	Arabideae	Brassicaceae	Brassicales	NC_009268
Arabis stelleri	Arabideae	Brassicaceae	Brassicales	KY126841
Arabis alpina	Arabideae	Brassicaceae	Brassicales	NC023367
Draba nemonrosa	Arabideae	Brassicaceae	Brassicales	NC_009273.1
Pugionium cornuturm	Megacarpaeeae	Brassicaceae	Brassicales	KT844941
Lobularia maritima	Anastaticeae	Brassicaceae	Brassicales	NC_009274.1
Ionopsidium acaule	Cochlearieae	Brassicaceae	Brassicales	NC_029333
Cochlearia tridactylites	Cochlearieae	Brassicaceae	Brassicales	NC029332
Vitis vinifera	Viteae	Vitaceae	Vitales	NC_007957.1

Table S2 General features of the *R. sativus* Hongxin chloroplast genome.

Feature	Chloroplast (Hongxin)	Chloroplast (WK10039)
Genome size (bp)	153 419	153 368
GC content (%)	36.30	36.30
Total number of genes	125	132
Protein coding genes	88	87
No. of rRNA genes	8	8
No. of tRNA genes	29	37
No. of gene duplications in IR regions	16	
Single intron (gene)	9	
Double intron (gene)	2	
Single intron (tRNA)	5	

Table S3 Codon-anticodon recognition patterns and codon usage of the *R. sativus* Hongxin chloroplast genome.

Codon-anticodon recognition patterns and codon usage of the hongxin chloroplast genome									
AA	Codon	No.	RSCU	tRNA	AA	Codon	No.	RSCU	tRNA
Phe	UUU(F)	2475	1.30		*Tyr*	UAU(Y)	1360	1.41	
	UUC(F)	1331	0.70	*trnF-GAA*		UAC(Y)	575	0.59	*trnY-GUA*
Leu	UUA(L)	1376	1.54	*trnL-UAA*	*Ter*	UAA(*)	1186	1.30	
	UUG(L)	1055	1.18	*trnL-CAA*		UAG(*)	651	0.71	
	CUU(L)	1060	1.19		*His*	CAU(H)	831	1.40	
	CUC(L)	566	0.64			CAC(H)	357	0.60	*trnH-GUG*
	CUA(L)	839	0.94		*Gln*	CAA(Q)	968	1.38	*trnQ-UUG*
	CUG(L)	449	0.50			CAG(Q)	432	0.62	
Ile	AUU(I)	1774	1.23		*Asn*	AAU(N)	1755	1.41	
	AUC(I)	951	0.66	*trnI-GAU*		AAC(N)	733	0.59	*trnN-GUU*
	AUA(I)	1608	1.11	*trnI-CAU*	*Lys*	AAA(K)	2409	1.43	*trnK-UUU*
Met	AUG(M)	814	1.00	*trn(f)M-CAU*		AAG(K)	970	0.57	
Val	GUU(V)	835	1.40		*Asp*	GAU(D)	1063	1.45	
	GUC(V)	392	0.66	*trnV-GAC*		GAC(D)	406	0.55	*trnD-GUC*
	GUA(V)	776	1.30	*trnV-UAC*	*Glu*	GAA(E)	1268	1.35	*trnE-UUC*
	GUG(V)	376	0.63			GAG(E)	607	0.65	
Ser	UCU(S)	1191	1.53		*Cys*	UGU(C)	681	1.28	
	UCC(S)	821	1.06	*trnS-GGA*		UGC(C)	386	0.72	*trnC-GCA*
	UCA(S)	1057	1.36	*trnS-UGA*	*Ter*	UGA(*)	903	0.99	
	UCG(S)	576	0.74		*Trp*	UGG(W)	647	1.00	*trnW-CCA*
Pro	CCU(P)	696	1.17		*Arg*	CGU(R)	429	0.80	*trnR-ACG*
	CCC(P)	580	0.98			CGC(R)	263	0.49	
	CCA(P)	706	1.19	*trnP-UGG*		CGA(R)	591	1.11	
	CCG(P)	390	0.66			CGG(R)	328	0.61	
Thr	ACU(T)	717	1.27		*Ser*	AGU(S)	633	0.81	
	ACC(T)	542	0.96	*trnT-GGU*		AGC(S)	387	0.50	*trnS-GCU*
	ACA(T)	690	1.22	*trnT-UGU*	*Arg*	AGA(R)	1038	1.95	*trnR-UCU*
	ACG(T)	314	0.56			AGG(R)	551	1.03	
Ala	GCU(A)	567	1.40		*Gly*	GGU(G)	572	1.06	
	GCC(A)	350	0.86			GGC(G)	311	0.58	*trnG-GCC*
	GCA(A)	464	1.15	*trnA-UGC*		GGA(G)	778	1.45	*trnG-UCC*
	GCG(A)	239	0.59			GGG(G)	491	0.91	
Codon-anticodon recognition patterns and codon usage of radish cultivar WK10039 chloroplast genome									
AA	Codon	No.	RSCU	tRNA	AA	Codon	No.	RSCU	tRNA
Phe	UUU	1091	1.35		*Tyr*	UAU	796	1.62	
	UUC	523	0.65	*trnF-GAA*		UAC	189	0.38	*trnY-GUA*
Leu	UUA	953	2.03	*trnL-UAA*	*Stop*	UAA	52	1.79	
	UUG	527	1.12	*trnL-CAA*		UAG	22	0.76	
	CUU	585	1.25		*His*	CAU	459	1.51	
	CUC	182	0.39			CAC	150	0.49	
	CUA	396	0.84	*trnL-UAG*	*Gln*	CAA	736	1.55	*trnQ-UUG*
	CUG	171	0.36			CAG	214	0.45	
Ile	AUU	1142	1.49		*Asn*	AAU	1012	1.54	
	AUC	431	0.56	*trnI-GAU*		AAC	305	0.46	*trnN-GUU*
	AUA	728	0.95		*Lys*	AAA	1160	1.52	*trnK-UUU*
Met	AUG	599	1.00	*trn(f)M-CAU*		AAG	362	0.48	
Val	GUU	534	1.49		*Asp*	GAU	840	1.61	
	GUC	179	0.50	*trnV-GAC*		GAC	203	0.39	*trnD-GUC*
	GUA	508	1.42	*trnV-UAC*	*Glu*	GAA	1066	1.53	*trnE-UUC*
	GUG	209	0.58			GAG	331	0.47	
Ser	UCU	600	1.76		*Cys*	UGU	244	1.52	
	UCC	295	0.86	*trnS-GGA*		UGC	78	0.48	*trnC-GCA*
	UCA	418	1.22	*trnS-UGA*	*Stop*	UGA	13	0.45	
	UCG	202	0.59		*Trp*	UGG	452	1.00	*trnW-CCA*
Pro	CCU	428	1.60		*Arg*	CGU	341	1.30	*trnR-ACG*
	CCC	200	0.75			CGC	109	0.42	
	CCA	307	1.15	*trnP-UGG*		CGA	364	1.39	
	CCG	135	0.50			CGG	129	0.49	
Thr	ACU	562	1.63		*Ser*	AGU	408	1.19	
	ACC	242	0.70	*trnT-GGU*		AGC	127	0.37	*trnS-GCU*
	ACA	419	1.22	*trnT-UGU*	*Arg*	AGA	465	1.78	*trnR-UCU*
	ACG	152	0.44			AGG	161	0.62	
Ala	GCU	630	1.84		*Gly*	GGU	576	1.30	
	GCC	208	0.61			GGC	168	0.38	*trnG-GCC*
	GCA	383	1.12	*trnA-UGC*		GGA	735	1.66	*trnG-UCC*
	GCG	149	0.44						

RSCU: relative synonymous codon usage.

Table S4 Forward and inverted repeats identified in the Hongxin radish cp genome using reputer.

Repeat start 1	Type	Size (bp)	Repeat start 2	Mismatch (bp)	E-value	Gene	Region
15044	F	35	15077	0	5.61E-12	IGS	LSC
15049	F	24	54964	0	2.35E-05	IGS	LSC
15049	F	24	54997	0	2.35E-05	IGS	LSC
15082	F	24	54964	0	2.35E-05	IGS	LSC
15082	F	24	54997	0	2.35E-05	IGS	LSC
25756	F	22	44292	0	3.76E-04	*ycf1*(CDS)	LSC
28614	F	22	140475	0	3.76E-04	IGS, *clpP*(intron)	LSC, IRA
54958	F	35	54991	0	5.61E-12	IGS	LSC
74778	F	22	83521	0	3.76E-04	*rps16*(CDS), IGS	LSC
77816	F	21	104727	0	1.51E-03	*trnS-GCT*(intron), *trnS-TGA*(intron)	LSC, IRB
106073	F	21	135551	0	1.51E-03	*trnM-CAT*(Intron), *trnP-TGG*(Intron)	IRB, SSC
108071	F	43	110295	0	8.56E-17	*psaB*(CDS), *psaA*(CDS)	IRB
15044	I	35	54958	0	5.61E-12	IGS	LSC
15049	I	24	15082	0	2.35E-05	IGS	LSC
15077	I	35	54991	0	5.61E-12	IGS	LSC
30469	I	21	111453	0	1.51E-03	*ccsA*(CDS), IGS	LSC, IRB
54964	I	24	54997	0	2.35E-05	IGS	LSC
70070	I	72	153347	0	2.97E-34	IGS	LSC, IRA
70366	I	23	70393	0	9.41E-05	IGS	LSC
76454	I	21	105709	0	1.51E-03	IGS	LSC, IRB
77812	I	28	114288	0	9.19E-08	*trnS-GCT*(Intron), *trnS-GGA*(Intron)	LSC
104727	I	21	114291	0	1.51E-03	*trnS-TGA*(Intron), *trnS-GGA*(Intron)	IRB
104793	I	24	114228	0	2.35E-05	*trnS-TGA*(Intron), *trnS-GGA*(Intron)	IRB
117472	I	24	119158	0	2.35E-05	IGS	SSC
119807	I	21	119834	0	1.51E-03	IGS	SSC
125494	I	21	125521	0	1.51E-03	IGS	SSC

Table S5 Simple sequence repeats in the *R. sativus* chloroplast genome.

cpSSR ID	SSR type	Repeat motif	Length (bp)	Start	End	Region	Annotation
1	p1	(T)10	10	15125	15134	LSC	
2	p1	(T)16	16	15908	15923	LSC	
3	p1	(A)10	10	26151	26160	LSC	*ycf1*
4	p1	(T)13	13	28618	28630	LSC	
5	p1	(T)10	10	28887	28896	LSC	
6	p1	(A)14	14	30477	30490	LSC	*ccsA*
7	p1	(A)11	11	33000	33010	LSC	
8	p1	(T)14	14	34947	34960	LSC	
9	p2	(TA)6	12	39587	39598	LSC	
10	c	(T)12[a]	123	39865	39987	LSC	*ycf1*
11	p1	(T)13	13	41542	41554	LSC	*ycf1*
12	p1	(T)12	12	41725	41736	LSC	*ycf1*
13	p1	(T)13	13	42043	42055	LSC	*ycf1*
14	c	(T)10[b]	73	42710	42782	LSC	*ycf1*
15	p1	(T)10	10	43911	43920	LSC	*ycf1*
16	p1	(A)16	16	54148	54163	LSC	
17	p1	(A)10	10	54937	54946	LSC	
18	p1	(A)10	10	71847	71856	LSC	
19	p2	(AT)8	16	73892	73907	LSC	
20	c	(T)14[c]	51	74194	74244	LSC	
21	c	(TA)6[d]	130	74672	74801	LSC	
22	p2	(TA)6	12	76461	76472	LSC	
23	c	(T)10c(A)10	21	77560	77580	LSC	
24	p1	(T)18	18	77983	78000	LSC	
25	c	(A)11[e]	70	78405	78474	LSC	
26	p1	(T)15	15	82669	82683	LSC	
27	p2	(AT)9	18	83525	83542	LSC	
28	p1	(A)10	10	83811	83820	LSC	
29	p1	(T)10	10	86973	86982	LSC	*rpoC2*
30	p1	(T)11	11	87728	87738	LSC	*rpoC2*
31	p1	(T)10	10	92248	92257	IRB	
32	p1	(T)10	10	95464	95473	IRB	*rpoB*
33	p2	(TA)7	14	96706	96719	IRB	
34	p1	(A)10	10	97168	97177	IRB	
35	p2	(AT)8	16	100873	100888	IRB	
36	p2	(AT)6	12	105713	105724	IRB	
37	p1	(T)14	14	111454	111467	IRB	
38	p1	(A)10	10	111703	111712	IRB	
39	p1	(A)11	11	111936	111946	IRB	
40	p1	(T)12	12	112842	112853	IRB	
41	p1	(T)10	10	115537	115546	SSC	
42	p1	(A)11	11	117476	117486	SSC	
43	p1	(T)14	14	117694	117707	SSC	
44	p1	(T)11	11	119169	119179	SSC	
45	c	(T)10[f]	78	120351	120428	SSC	
46	p1	(C)12	12	123624	123635	SSC	
47	p1	(A)11	11	125600	125610	SSC	
48	p1	(T)10	10	129458	129467	SSC	
49	p1	(G)10	10	135843	135852	IRA	
50	p1	(A)16	16	137353	137368	IRA	
51	p1	(A)10	10	137983	137992	IRA	
52	p1	(T)13	13	140479	140491	IRA	*clpP*
53	p1	(A)10	10	141038	141047	IRA	
54	p1	(T)13	13	147216	147228	IRA	*rpoA*
55	p1	(A)10	10	150256	150265	IRA	
56	c	(T)11[g]	59	151020	151078	IRA	
57	p1	(T)12	12	151342	151353	IRA	
58	p1	(T)10	10	151752	151761	IRA	

[a] (T)12aaattgaaacaaattagaattcgaaattctttacgtcgtttaggggatagaatagtttcagggacaaagaaatcataattatttttttttataaatatagc(T)11.

[b] (T)10ccacatgaaatttctaagaa(T)11agcccccatatatcaaacg(A)13.

[c] (T)14cgacaaggttgtaccgatcagaaaagc(A)10.

[d] (TA)6ctttttttagacctttttttatgacctttcattattcatataataatatatttatttagttattcatataaaataataatatatatttagtaaaataata(AT)10.

[e] (A)11gaatcctgctttgactaattttattaagtctacgctagaattttgtcg(T)11.

[f] (T)10gcattgggctctttcattaactgatagaaagatcagttagtctaccatatttttttctt(A)10.

[g] (T)11cattgtttttttcatcttttattacttttttttatttg(A)11.