Gene 551 (2014) 39–48

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

De novo assembly and characterization of the complete chloroplast genome of radish (*Raphanus sativus* L.)

Young-Min Jeong [a,1], Won-Hyung Chung [b,1], Jeong-Hwan Mun [c], Namshin Kim [b,*], Hee-Ju Yu [a,*]

[a] Department of Life Science, The Catholic University of Korea, Bucheon 420-743, Republic of Korea
[b] Korean Bioinformation Center, Korea Research Institute of Bioscience & Biotechnology, Daejeon 305-806, Republic of Korea
[c] Department of Bioscience and Bioinformatics, Myongji University, Yongin 449-728, Republic of Korea

A R T I C L E I N F O

Article history:
Received 18 June 2014
Received in revised form 19 August 2014
Accepted 20 August 2014
Available online 21 August 2014

Keywords:
Chloroplast genome
Raphanus sativus
Brassicaceae family
Next-generation sequencing
Simple sequence repeat

A B S T R A C T

Radish (*Raphanus sativus* L.) is an edible root vegetable crop that is cultivated worldwide and whose genome has been sequenced. Here we report the complete nucleotide sequence of the radish cultivar *WK10039* chloroplast (cp) genome, along with a de novo assembly strategy using whole genome shotgun sequence reads obtained by next generation sequencing. The radish cp genome is 153,368 bp in length and has a typical quadripartite structure, composed of a pair of inverted repeat regions (26,217 bp each), a large single copy region (83,170 bp), and a small single copy region (17,764 bp). The radish cp genome contains 87 predicted protein-coding genes, 37 tRNA genes, and 8 rRNA genes. Sequence analysis revealed the presence of 91 simple sequence repeats (SSRs) in the radish cp genome.

Phylogenetic analysis of 62 protein-coding gene sequences from the 17 cp genomes of the Brassicaceae family suggested that the radish cp genome is most closely related to the cp genomes of *Brassica rapa* and *Brassica napus*. Comparisons with the *B. rapa* and *B. napus* cp genomes revealed highly divergent intergenic sequences and introns that can potentially be developed as diagnostic cp markers. Synonymous and nonsynonymous substitutions of cp genes suggested that nucleotide substitutions have occurred at similar rates in most genes. The complete sequence of the radish cp genome would serve as a valuable resource for the development of new molecular markers and the study of the phylogenetic relationships of *Raphanus* species in the Brassicaceae family.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The chloroplast (cp) is a plant-specific organelle that carries out photosynthesis as well as many other crucial roles such as synthesizing starch, fatty acids, DNA, and RNA (Saski et al., 2007). Since the sequencing of the *Nicotiana tabacum* (Shinozaki et al., 1986) cp genome, almost 500 plastid genomes have been deposited in GenBank as of February 2014. Although some variations are observed, most cp genomes are 120–180 kb in length and have similar genome structures (Zhang et al., 2012). Most cp genomes also exhibit a pair of exact inverted repeats (IRs). These genomes generally exhibit a quadripartite structure containing a large single copy (LSC) region, a small single copy (SSC) region, and the two IRs. The cp genome contains both protein-coding genes and also functional RNA genes, such as tRNA genes and rRNA genes. Some protein-coding genes and tRNA genes contain introns that predominantly belong to group II (Kim and Lee, 2004). Among these introns, *rps12* is known to be trans-spliced (Hildebrand et al., 1988; Koller et al., 1987). In most cases, the cp is inherited from the maternal lineage, and exhibits no or very little recombination. Since the genetic features of the cp genome are relatively simple, the cp genome has played essential roles in molecular systematics and DNA barcoding approaches that rely on sequence variations (Dong et al., 2012).

Radish (*Raphanus sativus* L.) is an annual root vegetable crop that is cultivated worldwide and is an important item of seed industry in East Asia. Radish is a member of the Brassicaceae family and is also a close relative of the model organism *Arabidopsis thaliana*, in addition to other *Brassica* species. The phylogenetic origin of radish and its relationship with other species in the Brassiceae tribe, especially other *Brassica* species, are some of the unsolved aspects of radish biology. Previous phylogenetic studies have shown that radish is closely related to other *Brassica* species in the Brassiceae tribe (Warwick and Black, 1991, 1997). However, analyses of cp and nuclear sequences from radish and other *Brassica* species have yielded discrepant results regarding their phylogenetic relationships. For example, restriction site polymorphism analysis of several cp genomes placed radish in the *B. rapa*/*Brassica oleracea* lineage of the subtribe Brassicinae, which itself consisted of *B. rapa*/*B. oleracea* and *Brassica nigra* lineages (Warwick and Black, 1991, 1997). In contrast, molecular phylogenetic studies using a nuclear DNA

Abbreviations: bp, base pair(s); cp, chloroplast; DNA, deoxyribonucleic acid; GC, guanine–cytosine; IR, inverted repeat; K_a, nonsynonymous; kb, kilobase(s); K_s, synonymous; LSC, large single-copy region; ML, maximum likelihood; NGS, next-generation sequencing; RNA, ribonucleic acid; rRNA, ribosomal RNA; SSC, small single-copy region; SSR, simple sequence repeat(s); tRNA, transfer RNA.

* Corresponding authors.
E-mail addresses: n@rna.kr (N. Kim), yuheeju@catholic.ac.kr (H.-J. Yu).

[1] Both authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2014.08.038
0378-1119/© 2014 Elsevier B.V. All rights reserved.

marker suggested that radish is more closely related to *B. nigra* than to *B. rapa*/*B. oleracea* (Song et al., 1988, 1990; Thormann et al., 1994; Yang et al., 1999). This conflict may arise from the hybrid origin of radish between the *B. rapa*/*B. oleracea* and *B. nigra* lineages (Song et al., 1990; Yang et al., 2002).

Comprehensive studies using diverse accessions of *Raphanus* species have suggested that cultivated radishes have multiple origins (Yamagishi et al., 2009; Yamane et al., 2005, 2009). Based on the sequences of conserved regions in the cp genome, several protein-coding genes or intergenic regions have been used for phylogenetic studies of *Raphanus* species. For example, the 5′-*matK* region of the cp genome was used to investigate the origin of cultivated radish (Yamane et al., 2005). Ridley et al. (2008) studied interspecific hybridization between cultivated radish and wild species (*Raphanus raphanistrum*) on the basis of polymorphisms in the *trnL*–*rpl32* intergenic region. Yamane et al. (2009) used 25 cp simple sequence repeats (SSRs) to study the phylogenetic relationships of 82 accessions of cultivated and wild *Raphanus* species. In addition, molecular markers in the cp sequence have been developed to distinguish radish cytotypes that are related to male sterility (Kim et al., 2009). However, the complete genome of radish would greatly facilitate research on the origin and phylogenetic relationships of radish.

Despite its importance in agriculture and evolutionary phylogenetics, information on the radish genome is still relatively limited compared with the wealth of information available for the *Arabidopsis* or *Brassica* genomes. Recently, draft genome sequences of Japanese radish (cv. Aokubi, Kitashiba et al., 2014) and wild radish (*R. raphanistrum*, Moghe et al., 2014) have been reported. However, the complete sequence of radish cp genome is not yet available. We have been sequencing the whole genome of a Korean radish cultivar, *WK10039* (2n = 18, 510 Mb), using the next-generation sequencing (NGS) strategy since 2012. The vast amount of sequence reads generated from the whole genome sequencing enabled the de novo assembly of the cp genome (Wolf et al., 2011; Wu et al., 2012; Zhang et al., 2013). In this study, we report the successful de novo assembly of a complete radish cp genome, obtained using different types of NGS reads. We also identified SSRs in the cp genome. The resulting data set allowed a comparative genomic analysis with the cp genomes of other Brassicaceae species, thereby demonstrating the close relationships of radish with *B. rapa* and *B. napus*.

2. Material and Methods

2.1. Plant Material and DNA Extraction

To sterilize seed surfaces, seeds of the *R. sativus* cv. *WK10039* were treated with 70% (v/v) ethanol for 2 min and 10% (v/v) commercial bleach for 10 min; seeds were then washed five times in sterile water. Sterilized seeds were plated on 0.5× Murashige and Skoog agar plates (0.7%) and grown 1 month in a growth chamber. Seeds were grown at 22 °C and 60% humidity, with a 16 h/8 h light/dark cycle. Genomic DNA was extracted from young plants using a DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's protocol.

2.2. DNA Sequences

Whole genome sequences of cv. *WK10039* were generated using three different NGS platforms: Illumina HiSeq1000, Roche/454 GS-FLX Plus, and PacBio RS II (our unpublished data). In brief, 500-bp libraries were constructed using the TruSeq™ DNA sample preparation kit (Illumina, USA) for Illumina HiSeq sequencing. After generating clusters with the TruSeq PE Cluster V3-cbot-hs kit and the cBot system, sequencing was performed using the TruSeq SBS V3-hs kit. For 454 sequencing, 1.5-kb shotgun libraries were constructed using a GS rapid library preparation kit (Roche, USA). To construct 3-kb, 8-kb, 20-kb, and 30-kb mate-paired libraries, a GS FLX Titanium Library Paired End Adaptors kit was used. Libraries were then amplified using the GS FLX Titanium LV emPCR kit, and sequencing was performed using the GS FLX Titanium XL+ sequencing kit. The 10-kb library was constructed using a PacBio RS II DNA Template Preparation Kit (Pacific Biosciences, USA); the DNA sequencing kit (v2) was used to sequence the PacBio Rs II samples. Sequencing reactions were performed according to the manufacturers' protocols.

2.3. cp Genome Assembly

The 454 shotgun sequencing reads were converted to fastq format using sffinfo, a format converter of Newbler 2.7 (Margulies et al., 2005). To isolate the cp reads, all 454 shotgun reads were subjected to BLAT (Kent, 2002) analysis (identity > 90%; alignment length > 100 bp) using the published *B. rapa* cp genome (NC_015139) as a reference. PacBio reads were mapped to the *B. rapa* cp genome based on alignments generated with BLASR (Chaisson and Tesler, 2012) (identity > 85%; alignment length > 500 bp). PacBio reads were error-corrected with the 454 shotgun reads by PacBioToCA (Koren et al., 2012). Selected 454 shotgun reads, generated from 56 runs, were individually assembled using Newbler 2.7 with its default options. The resulting contigs were integrated with CISA, version 1.2 (Lin and Liao, 2013). Scaffolds were constructed by mapping the contigs to the *B. rapa* cp genome using LASTZ (Harris, 2007) and a software program developed in-house. PBgelly (English et al., 2012) was used to gap-fill the scaffolds with PacBio reads, thereby generating a circular cp genome. To correct possible misassemblies and errors, Illumina short reads were mapped to the assembled cp genome using BWA (version 0.5.9) (Li and Durbin, 2009). All corrections were manually curated at the homozygous variation sites, which were identified by GATK UnifiedGenotyper (version 2.4.9) (McKenna et al., 2010). Frame-shift errors were also corrected during gene prediction. The complete radish cp genome was ultimately deposited in GenBank (accession number: NC_024469).

2.4. Gene Annotation and Sequence Analysis

Genes in the cp genome were identified using DOGMA (Dual Organellar GenoMe Annotator) (Wyman et al., 2004) and BLAST (Altschul et al., 1990) searches, with subsequent manual curation. Transfer RNA genes were identified with the tRNA scan-SE program (http://lowelab.ucsc.edu/tRNAscan-SE). All identified genes and tRNA genes were validated by comparison with the appropriate homologs in the *B. rapa* cp genome. A genomic alignment was then performed with LASTZ, using the *B. rapa* cp genome as a reference. A circular

Table 1
List of cpDNA of the Brassicaceae species used in phylogenetic analysis.

Name	GenBank accession	Reference
Aethionema cordifolium DC.	NC_009265	Direct deposit
Aethionema grandiflorum Boiss. et Hohen.	NC_009266	Direct deposit
Arabidopsis thaliana (L.) Heynh.	NC_000932	Sato et al. (1999)
Arabis hirsuta (L.) Scop.	NC_009268	Direct deposit
Barbarea verna (Mill.) Asch.	NC_009269	Direct deposit
Brassica napus L.	NC_016734	Hu et al. (2011)
Brassica rapa L.	NC_015139	Wu et al. (2012)
Capsella bursa-pastoris (L.) Medik.	NC_009270	Direct deposit
Carica papaya L.	NC_010323	Direct deposit
Crucihimalaya wallichii Al-Shehbaz	NC_009271	Direct deposit
Draba nemorosa L.	NC_009272	Direct deposit
Lepidium virginicum L.	NC_009273	Direct deposit
Lobularia maritima (L.) Desvaux	NC_009274	Direct deposit
Nasturtium officinale R. Br.	NC_009275	Direct deposit
Olimarabidopsis pumila Al-Shehbaz	NC_009267	Direct deposit
Pachycladon cheesemanii Heenan & A.D. Mitch	NC_021102	Direct deposit
Pachycladon enysii (Cheeseman ex Kirk) Heenan & A.D. Mitch	NC_018565	Direct deposit
Raphanus sativus L.	NC_024469	This study

gene map of the radish cp genome was generated using CGView (Stothard and Wishart, 2005). To calculate codon usage frequencies, CodonW (http://codonw.sourceforge.net/) was used.

2.5. Phylogenetic Analysis and Synonymous/Nonsynonymous Substitution Calculations

The amino acid sequences of 62 protein-coding genes (*psaA*, *psaB*, *psaC*, *psaI*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *petA*, *petB*, *petD*, *petG*, *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *rbcL*, *rpoA*, *rpoB*, *rpoC1*, *rpoC2*, *ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *rpl2*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps14*, *rps18*, *clpP*, and *ycf4*) from 18 cp genomes (Table 1) were used for phylogenetic analysis. Each sequence was aligned using MUSCLE v3.8.31 (Edgar, 2004), and then GBLOCKS 0.91b (Talavera and Castresana, 2007) was used to remove poorly aligned regions. The resultant conserved regions were concatenated into a single sequence for each cp genome. Phylogenetic trees were generated by the maximum likelihood (ML) method, under the JTT + G + I + F model, using MEGA5 (Tamura et al., 2011). The stability of each tree node was tested by bootstrap analysis with 100 replicates.

To calculate the rates of synonymous (K_s) and nonsynonymous (K_a) substitutions per site in the 62 protein-coding genes, all coding sequences were aligned with MUSCLE v3.8.31 (Edgar, 2004). The K_a and K_s values were obtained with the KaKs-calculator (Zhang et al., 2006), using the model averaging method. *Carica papaya* was used as an outgroup.

2.6. Whole cp Sequence Comparisons and Identification of SSRs

The cp genome sequences of *R. sativus*, *B. rapa*, and *B. napus* (NC_016734) were compared and visualized using mVISTA (Frazer et al., 2004). *R. sativus* cp genome annotation information was included in the visualization. SSRs were identified using SSRIT (Temnykh et al., 2001), and SSRs less than 10 bp were removed. The remaining SSRs in the cp genomes were then compared by performing a BLASTN search.

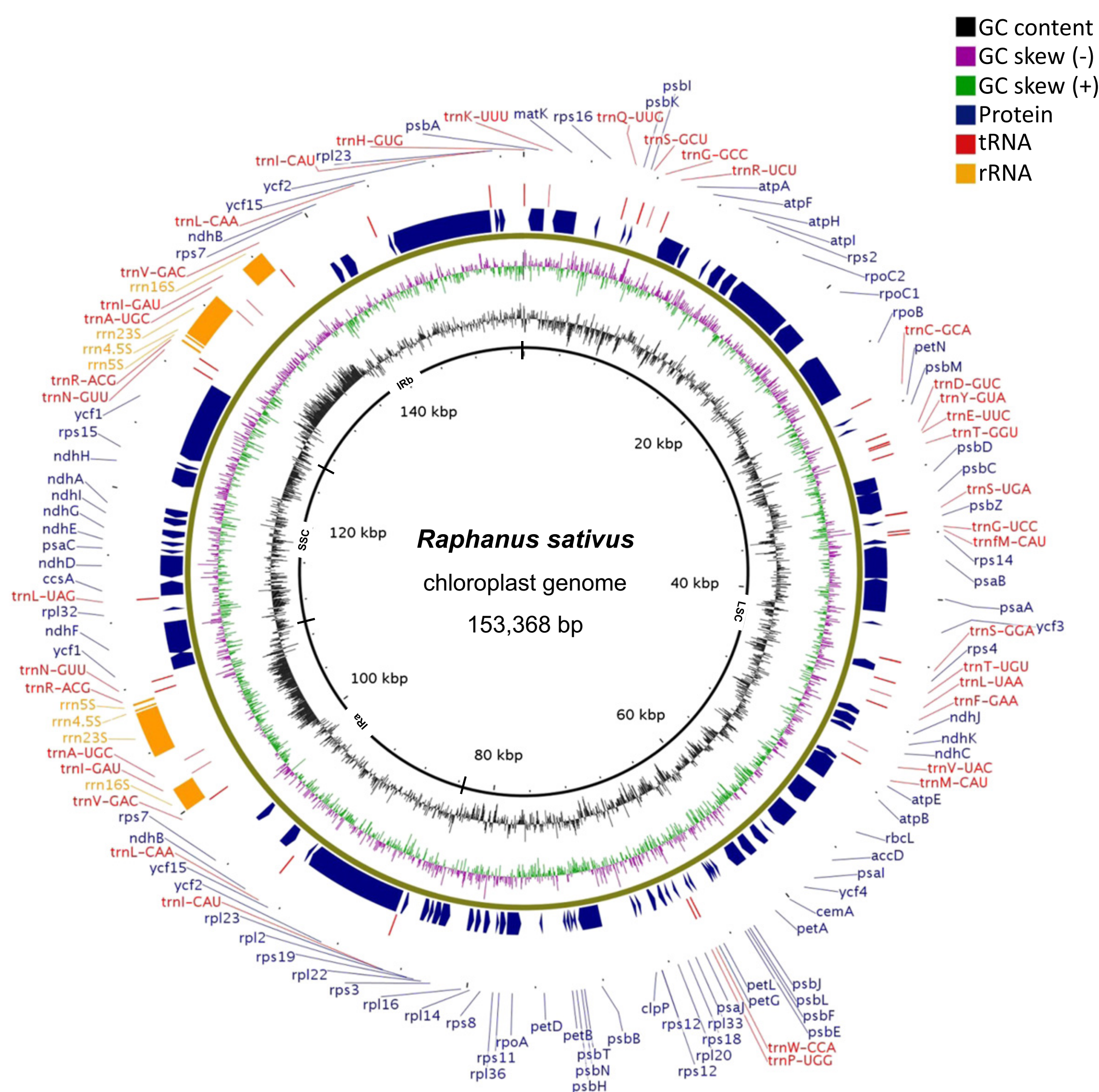

Fig. 1. Gene map of the *R. sativus* cv. *WK10039* cp genome. Blue arrows indicate the direction of transcription of protein-coding genes. Red and orange arrows denote tRNA and rRNA genes, respectively. Circle 1 (inside) indicates the coordination and structure of the cp. Circles 2 and 3 indicate the GC content and GC skew, respectively.

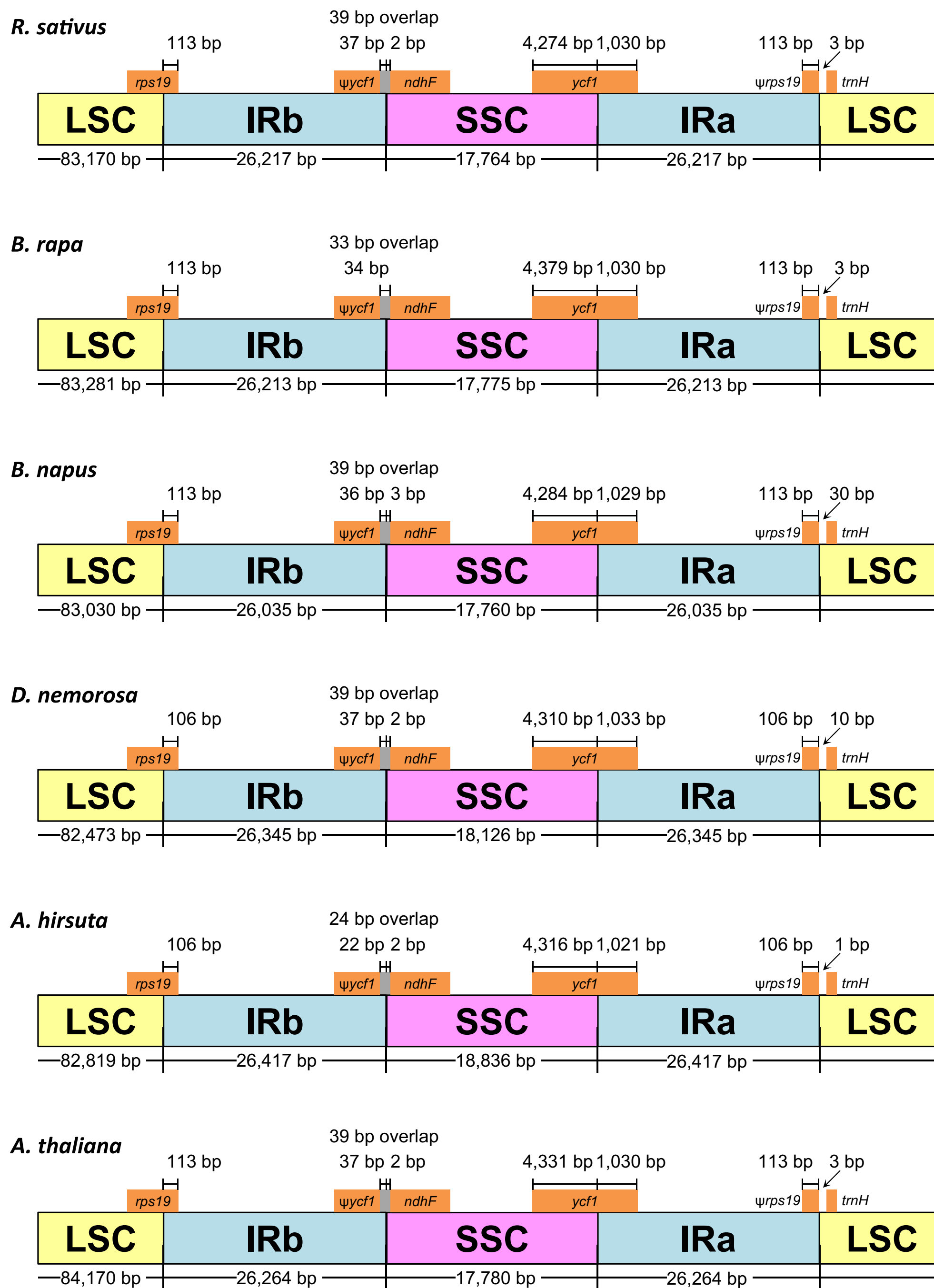

Fig. 2. Comparison of the border regions of the cp genomes from six species in the Brassicaceae family.

3. Results and Discussion

3.1. Radish cp Genome Assembly and Validation

Using a whole genome shotgun sequencing strategy comprising multiple NGS technologies, including Illumina, 454, and PacBio platforms, we obtained the entire radish cp genome sequence. Since each sequencing platform exhibits both advantages and disadvantages regarding sequence length, quantity, and quality, we used raw reads from multiple sequencing platforms to achieve the assembly of the cp genome. As an initial step in the cp genome assembly, we isolated radish reads exhibiting high similarity with the *B. rapa* cp genome, as assessed by a BLAT search. A total of 3.16 million (1.53 Gb; 9.8%) sequence reads

Table 2
General features of the *R. sativus* cp genome.

Size (bp)	153,368
G + C content (%)	36.3
Protein-coding genes	87
Coverage of the cp genome (%)	52.1
Average size (bp)	918
G + C content (%)	37.0
tRNA genes	37
rRNA genes	8
4.5S	2
5S	2
16S	2
23S	2

Table 3
Gene content of the *R. sativus* cp genome.

Group	Name of genes	Numbers
Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*	5
Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*	15
Cytochrome b/f complex	*petA*, *petB*, *petD*, *petG*, *petL*, *petN*	6
ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*	6
NADH dehydrogenase	*ndhA*, *ndhB*(×2), *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*	12
RubisCO large subunit	*rbcL*	1
RNA polymerase	*rpoA*, *rpoB*, *rpoC1*, *rpoC2*	4
Small subunits of ribosome	*rps2*, *rps3*, *rps4*, *rps7*(×2), *rps8*, *rps11*, *rps12*(×2), *rps14*, *rps15*, *rps16*, *rps18*,*rps19*	14
Large subunits of ribosome	*rpl2*(×2), *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*(×2), *rpl32*, *rpl33*, *rpl36*	11
Miscellaneous proteins	*accD*, *ccsA*, *cemA*, *clpP*, *matK*	5
Conserved open reading frames	*ycf1*(×2, part), *ycf2*(×2), *ycf3*, *ycf4*, *ycf15*(×2)	8
tRNAs	*trnA–UGC*(×2), *trnC–GCA*, *trnD–GUC*, *trnE–UUC*, *trnF–GAA*, *trnM–CAU*, *trnG–GCC*, *trnG–UCC*, *trnH–GUC*, *trnI–CAU*(×2), *trnI–GAU*(×2), *trnK–UUU*, *trnL–UAA*, *trnL–CAA*(×2), *trnL–UAG*, *trnM–CAU*, *trnN–GUU*(×2), *trnP–UGG*, *trnQ–UUG*, *trnR–UCU*, *trnR–ACG*(×2), *trnS–GCU*, *trnS–UGA*, *trnS–GGA*, *trnT–GGU*, *trnT–UGU*, *trnV–UAC*, *trnV–GAC*(×2), *trnW–CCA*, *trnY–GUA*	37
rRNAs	*rrn4.5*(×2), *rrn5*(×2), *rrn16*(×2), *rrn23*(×2)	8

were selected from 15.6 Gb of whole genome 454 shotgun sequences. These selected reads corresponded to 9976× coverage of the assembled radish cp genome. Isolated cp reads from the 454 shotgun sequences from each run were individually assembled, resulting in an average of 13.4 contigs generated per run. From each run, 348 contigs were assembled and integrated into super contigs, thus generating 55 super contigs. Since PacBio sequence reads are long and therefore suitable for scaffolding, we used the PacBio reads to generate a circular genome. The assembled sequence was finally validated by mapping the Illumina reads to the assembled cp genome, which resulted in the detection of 38 errors. Among these errors, 33 corresponded to deletions in A or T homopolymers; 454 sequencing technology is known to have a high error rate in detecting these homopolymers (Margulies et al., 2005; Wang et al., 2005; Wolfe et al., 1987). After correcting these errors, a circular gene map was generated (Fig. 1).

3.2. Structure and Features of the Radish cp Genome

The radish cp genome is 153,368 bp in length and contains a pair of IRs, each of which is 26,217 bp in length. These IRs separate 83,170 bp of LSC and 26,217 bp of SSC (Fig. 2). The radish cp genome contains 87 protein-coding genes, 8 rRNA genes, and 37 tRNA genes (Tables 2 and 3). Among the 61 possible codons (excluding stop codons), tRNAs for 27 exist in the *R. sativus* cp genome (Table 4). Moreover, 6 protein-coding genes (*rps7*, *ndhB*, *ycf15*, *ycf2*, *rpl23*, and *rpl2*), 7 tRNA genes, and all rRNA genes reside within the IRs. Eighteen protein-coding

Table 4
Codon usage of protein-coding genes in the *R. sativus* cp genome.

Amino acid	Codon	Count	Codon usage	tRNA	Amino acid	Codon	Count	Codon usage	tRNA
Phe	UUU	1091	1.35		Tyr	UAU	796	1.62	
Phe	UUC	523	0.65	*trnF–GAA*	Tyr	UAC	189	0.38	*trnY–GUA*
Leu	UUA	953	2.03	*trnL–UAA*	Stop	UAA	52	1.79	
Leu	UUG	527	1.12	*trnL–CAA*	Stop	UAG	22	0.76	
Leu	CUU	585	1.25		His	CAU	459	1.51	
Leu	CUC	182	0.39		His	CAC	150	0.49	
Leu	CUA	396	0.84	*trnL–UAG*	Gln	CAA	736	1.55	*trnQ–UUG*
Leu	CUG	171	0.36		Gln	CAG	214	0.45	
Ile	AUU	1142	1.49		Asn	AAU	1012	1.54	
Ile	AUC	431	0.56	*trnI–GAU*	Asn	AAC	305	0.46	*trnN–GUU*
Ile	AUA	728	0.95		Lys	AAA	1160	1.52	*trnK–UUU*
Met	AUG	599	1	*trn(f)M–CAU*	Lys	AAG	362	0.48	
Val	GUU	534	1.49		Asp	GAU	840	1.61	
Val	GUC	179	0.5	*trnV–GAC*	Asp	GAC	203	0.39	*trnD–GUC*
Val	GUA	508	1.42	*trnV–UAC*	Glu	GAA	1066	1.53	*trnE–UUC*
Val	GUG	209	0.58		Glu	GAG	331	0.47	
Ser	UCU	600	1.76		Cys	UGU	244	1.52	
Ser	UCC	295	0.86	*trnS–GGA*	Cys	UGC	78	0.48	*trnC–GCA*
Ser	UCA	418	1.22	*trnS–UGA*	Stop	UGA	13	0.45	
Ser	UCG	202	0.59		Trp	UGG	452	1	*trnW–CCA*
Pro	CCU	428	1.6		Arg	CGU	341	1.3	*trnR–ACG*
Pro	CCC	200	0.75		Arg	CGC	109	0.42	
Pro	CCA	307	1.15	*trnP–UGG*	Arg	CGA	364	1.39	
Pro	CCG	135	0.5		Arg	CGG	129	0.49	
Thr	ACU	562	1.63		Ser	AGU	408	1.19	
Thr	ACC	242	0.7	*trnT–GGU*	Ser	AGC	127	0.37	*trnS–GCU*
Thr	ACA	419	1.22	*trnT–UGU*	Arg	AGA	465	1.78	*trnR–UCU*
Thr	ACG	152	0.44		Arg	AGG	161	0.62	
Ala	GCU	630	1.84		Gly	GGU	576	1.3	
Ala	GCC	208	0.61		Gly	GGC	168	0.38	*trnG–GCC*
Ala	GCA	383	1.12	*trnA–UGC*	Gly	GGA	735	1.66	*trnG–UCC*
Ala	GCG	149	0.44		Gly	GGG	289	0.65	

Table 5
The predicted length of exons and introns in the *R. sativus* cp genome.

Gene	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
atpF	145	723	409		
clpP	71	933	292	564	228
ndhA	553	1086	530		
ndhB	777	679	762		
petB	6	788	642		
petD	8	682	475		
rpl16	9	1079	399		
rpl2	390	682	435		
rpoC1	432	779	1611		
rps12	114	–	232	537	26
rps16	40	859	227		
trnA-UGC	38	800	35		
trnG-UCC	23	716	49		
trnI-GAU	37	809	35		
trnK-UUU	37	2565	35		
trnL-UAA	35	313	50		
trnV-UAC	39	615	35		
ycf3	126	797	228	722	153

Table 6
Base composition of the *R. sativus* cp genome.

Regions	A (%)	T (%)	G (%)	C (%)	Length (bp)
SSC	35.4	35.4	14.1	15.1	17,764
IR	28.8	28.8	21.2	21.2	26,217
LSC	32.1	33.8	16.6	17.5	83,170
Total	31.3	32.3	17.8	18.5	153,368

genes or tRNA genes have introns, ranging from 26 to 2565 bp in length (Table 5). Out of the total cp genome, protein-coding regions consist of 52.1%. The GC content of the cp genome is 36.3%; this value is slightly higher (37.0%) in the protein-coding regions. In addition, IRs show a higher GC content (42.3%), whereas the SSC region has a lower GC content (29.2%). The nucleotide composition of each region is listed in Table 6.

3.3. Identification of SSRs in the Radish cp Genome

SSRs are repeated DNA sequences consisting of direct tandem repeats of short (1–10 bp) nucleotide motifs. Due to their highly polymorphic nature, SSRs are highly useful for marker development. Yamane et al. (2009) used 25 *Arabidopsis* cp SSR markers to examine 82 accessions of *Raphanus* species. Seven loci were polymorphic; moreover, most of them had less than 10 nucleotides. Using these markers, 20 different haplotypes were identified, suggesting that cp SSR markers are effective for genotype determination (Yamane et al., 2009). In this study, we identified 91 SSR loci, each longer than 10 nucleotides, in the radish cp genome. The longest SSR was 15 nucleotides in length; 63.7% (58 out of 91) of all SSRs were mononucleotide repeats (Table 7). In addition, the most abundant nucleotides in SSRs were A and T. These results are consistent with those of the *B. napus* cp genome (Hu et al., 2011), in which 84 SSRs were identified. Furthermore, BLAST analysis indicated that 66 radish SSRs are conserved in the *B. napus* cp genome (Supplementary Table S1). Therefore, the SSRs reported in this study will potentially be a useful resource for genetic and evolutionary studies of diverse *Raphanus* and *Brassica* species.

3.4. Comparison of the WK10039 cp With Known Radish cp Haplotypes

Markers derived from cpDNA have been used to study the phylogenetic relationships and diversity of both cultivated and wild radishes. To examine the maternal lineage of *WK10039*, we compared the *WK10039* cp genome sequence with published cp haplotype markers (Supplementary Table S2). Notably, the *rpoC3–trnC* region of *WK10039* is predicted to be haplotype D, which was the predominant haplotype detected in a study of Korean and Japanese cultivars (Yamane et al., 2005).

Table 7
Distributions of SSRs in the *R. sativus* cp genome.

Base	Length	Numbers of SSR	Coordination
A	10	16	1706-1715, 4091-4100, 4285–4294, 7400–7409, 13,633–13,642, 26,992–27,001, 41,546–41,555, 50,260–50,269, 55,438–55,447, 66,081–66,090, 67,820–67,829, 80,079–80,088, 109,368–109,377, 114,409–114,418, 119,800–119,809, 138,188–138,197
	11	2	47,320–47,330, 80,892–80,902
	12	1	8233–8244
	13	2	116,217–116,229, 126,021–126,033
	14	1	113,694–113,707
	16	1	137,399–137,414
C	10	1	53,465–53,474
T	10	12	7389–7398, 16,796–16,805, 25,287–25,296, 26,336–26,345, 28,296–28,305, 45,379–45,388, 45,766–45,775, 50,192–50,201, 81,166–81,175, 98,342–98,351, 125,961–125,970, 127,162–127,171
	11	6	8293–8303, 17,551–17,561, 41,300–41,310, 112,103–112,113, 123,228–123,238, 125,991–126,001
	12	4	42,683–42,694, 111,835–111,846, 123,116–123,127, 124,976–124,987
	13	6	4051–4063, 70,315–70,327, 77,049–77,061, 118,169–118,181, 124,793–124,805, 125,294–125,306
	14	4	12,497–12,510, 47,538–47,551, 49,013–49,026, 81,429–81,442
	16	2	7812–7827, 99,125–99,140
AT	10	6	7948–7957, 107,500–107,509, 112,662–112,671, 120,324–120,333, 129,030–129,039, 143,043–143,052
	12	1	35,534–35,545
	14	3	3751–3764, 4607–4620, 13,352–13,365
	16	1	30,694–30,709
TA	10	5	7853–7862, 9549–9558, 18,923–18,932, 62,097–62,106, 93,486–93,495
	12	4	4519–4530, 6290–6301, 111,651–111,662, 122,838–122,849
	14	1	26,530–26,543
AAT	12	1	12,645–12,656
ATT	12	1	45,832–45,843
AATA	12	1	112,730–112,741
ATAG	12	1	111,413–111,424
TAAA	12	2	45,654–45,665, 62,221–62,232
TTAA	12	2	45,674–45,685, 66,248–66,259
TTCT	12	1	34,414–34,425
TTTA	12	1	6866–6877
CAACA	15	1	137,901–137,915
TGTTG	15	1	98,624–98,638

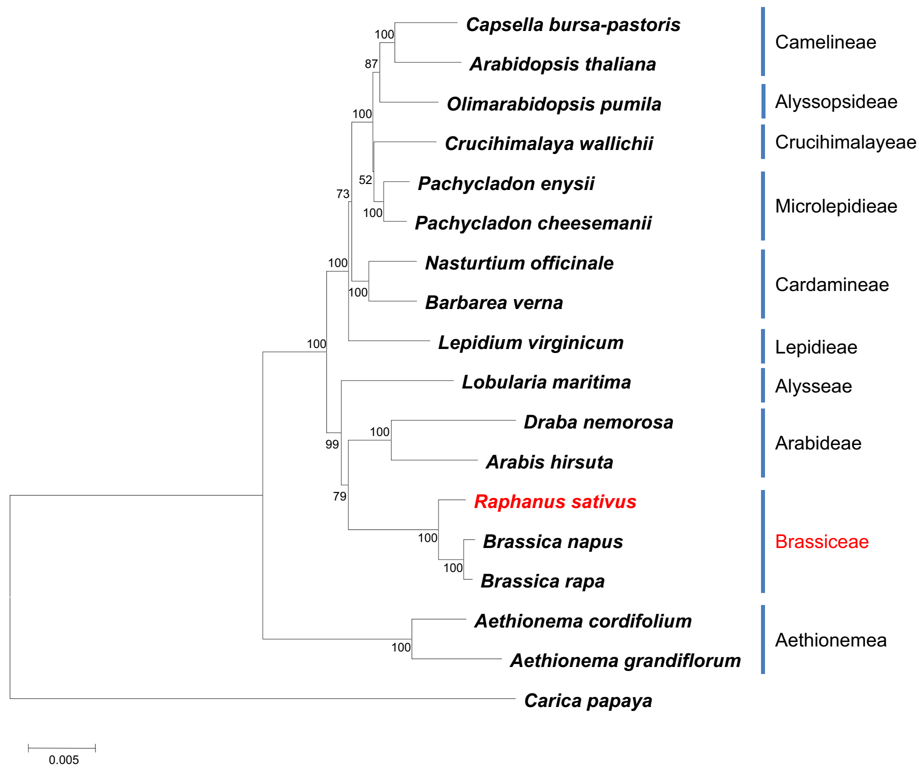

Fig. 3. Phylogenetic tree of 17 Brassicaceae species, based on 62 protein-coding genes in the cp genome. The tree was generated by maximum likelihood analysis of the conserved regions using the MEGA5 program. The stability of each tree node was tested by bootstrap analysis with 100 replicates. Bootstrap values are indicated on the branches; the branch length reflects the estimated number of substitutions per 100 sites. The tribe names of the corresponding species are indicated on the right margin (blue bars).

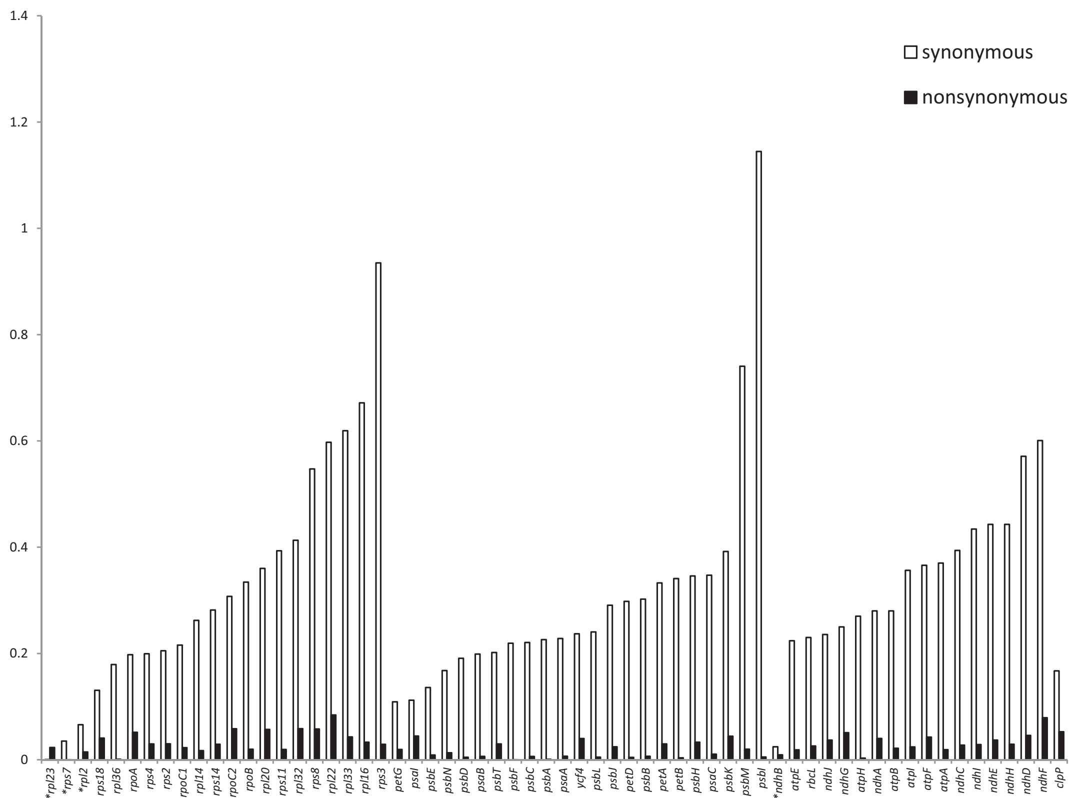

Fig. 4. Mean K_s and K_a values of the 62 protein-coding genes in various species of the tribe Brassicaceae and *C. papaya*. Blank and black bars indicate K_s and K_a values, respectively. Genes located within an IR are indicated with asterisks.

Among the three groups of the *Raphanus* genus, which are based on the *trnK*/*matK* sequence (Lü et al., 2008), the *WK10039* sequence is identical to that of *R. sativus* var. *sativus* (European small radish), which exclusively belongs to group A (Fig. S1). Ridley et al. (2008) identified eight haplotypes using multiple cp markers in *trnL–rpl32* (EU998961–EU998968), and found that the sequence of *WK10039* is highly similar to that of haplotype D (EU998964), which is only found in cultivated radish. Therefore, the maternal lineage of *WK10039* is likely to be neither a region-specific nor a unique cultivar.

3.5. Phylogenetic Analysis of the Radish cp Genome

Using 62 protein-coding genes in the cp genome, we performed a phylogenetic analysis of 17 cp genomes in the Brassicaceae family. We aligned the amino acid sequences of the 62 protein-coding genes, and used GBLOCKS software to identify conserved blocks. The conserved blocks were then merged into a single sequence, consisting of 15,295 amino acids, for each cp genome. The resultant ML tree was in good agreement with taxonomical classification data (Fig. 3). Furthermore, the topology of the tree was virtually identical to that of the tree reported in a study by Hu et al. (2011), which used the nucleotide sequence of 61 protein-coding genes in the cp genomes. However, we were unable to determine whether radish belongs to the B. *rapa*/*B. oleracea* or the *B. nigra* lineage, since the *B. nigra* cp genome is unavailable for analysis. Previous analysis of mitochondrial genomes suggested that radish is closer to *Brassica carinata* than to *B. rapa*, *B. oleracea*, *B. napus*, or *Brassica juncea* (Chaisson and Tesler, 2012). Since *B. carinata* is an allotetraploid hybrid between *B. oleracea* and *B. nigra*, mitochondrial genome analysis is unable to clearly determine whether *R. sativus* belongs to the *B. nigra* lineage. Therefore, the phylogenetic relationships and origin of *R. sativus* promise to be more evident when the complete nuclear and organellar genome sequences of *B. nigra* are available.

3.6. Nucleotide Substitutions in Protein-Coding Genes

Synonymous and nonsynonymous substitution patterns are two important indicators of gene evolution. To assess possible differences in gene evolution among the *Brassica* species, the nucleotide substitution rate of each species was calculated. Using the *C. papaya* cp gene set as an outgroup, the average K_s and K_a values for 62 protein-coding genes of 17 Brassicaceae species used in the phylogenetic analysis were calculated. The average K_s and K_a values for all genes were 0.32 ± 0.20 and 0.027 ± 0.020, respectively. Among the 62 genes, 37 (59.7%) exhibited a K_s value ranging from 0.2 to 0.4, suggesting that nucleotide substitutions occurred at similar rates in those genes. As reported in previous studies (Koren et al., 2012; Li and Durbin, 2009; McKenna et al., 2010), genes located in an IR (*rps7*, *ndhB*, *rpl2*, and *rpl23*) showed low nucleotide substitution rates (Fig. 4). In contrast, genes such as *psbI*, *psbM*, and *rps3* showed high K_s values compared with other genes. These results are consistent with those reported for *Phalaenopsis* and grasses (English et al., 2012), in addition to fern and seed plants (Harris, 2007). It is particularly noteworthy that these genes have very low K_a values (less than 0.1), suggesting that high selective constants or purifying selection act on these genes.

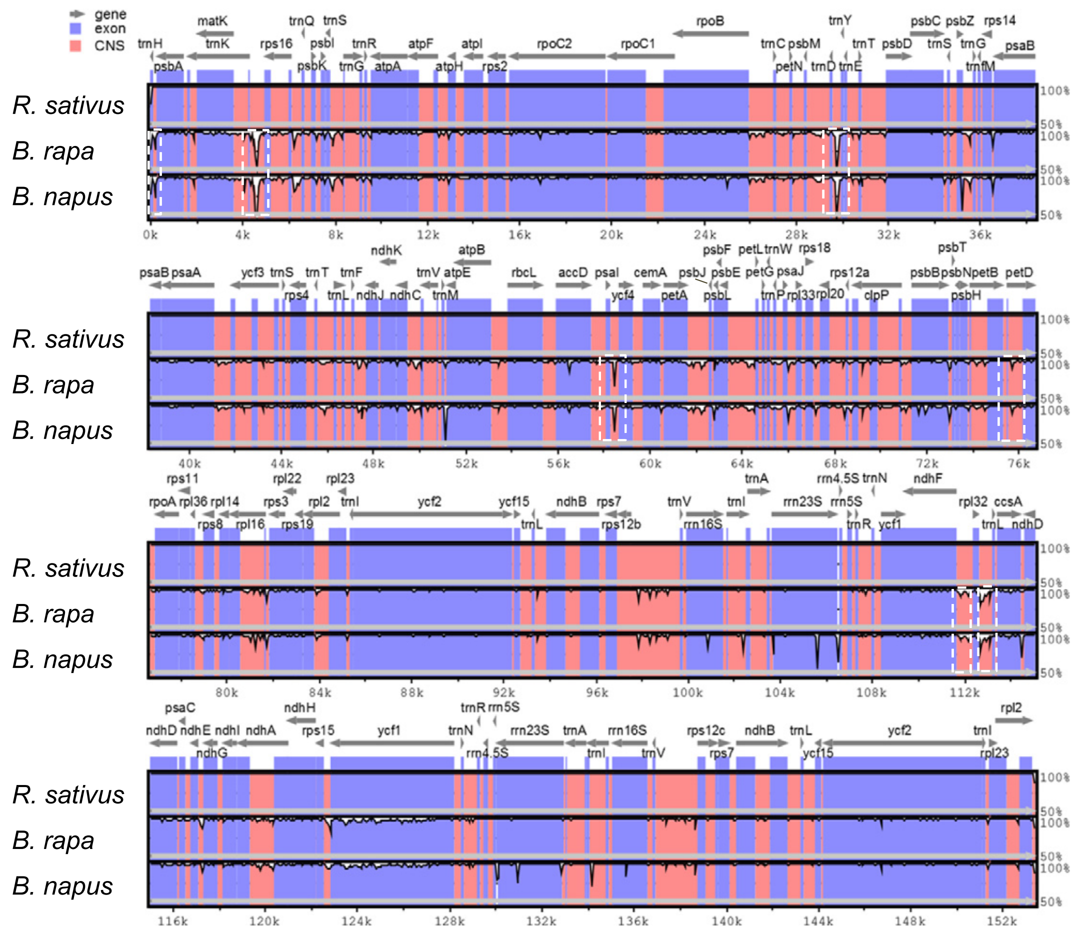

Fig. 5. Sequence comparison of the *R. sativus*, *B. rapa* and *B. napus* cp genomes generated with mVISTA. Gray arrows indicate the position and direction of each gene. Areas colored in blue and red indicate genic and intergenic regions, respectively. Black lines designate regions of sequence identity with *R. sativus*, using a 50% identity cutoff. Dashed rectangles indicate highly divergent regions of *R. sativus* compared to *B. rapa* and *B.napus*.

3.7. Comparison of the Radish cp Genome With the B. rapa and B. napus cp Genomes

Among the Brassicaceae species with sequenced cp genomes, *B. rapa* and *B. napus* were the closest relatives of radish in the Brassiceae tribe. The border structures of the *R. sativus*, *B. rapa*, and *B. napus* cp genomes were highly similar to one another. However, one important variation was the length of the *ycf1* gene in the SSC (Fig. 2). To examine the divergent regions of the cp genome, we analyzed the overall sequence identities between *R. sativus*, *B. rapa*, and *B. napus* using mVISTA (Frazer et al., 2004). Multiple sequence alignments showed that the coding regions are highly conserved, whereas the noncoding regions are divergent (Fig. 5). For instance, the intergenic sequences between the *trnH–GUG–psbA*, *trnK–UUU–rps16*, *trnD–GUC–trnY–GUA*, *psaI–ycf4*, *ndhF–rpl32*, *rpl32–trnL–UAG*, and *petD* introns were highly divergent. We note that parts of these regions have been reported as divergent sequences in previous studies. Ridley et al. (2008) used the sequence variation of the *trnL–rpl32* intergenic region to identify different haplotypes of California wild radish. A recent meta-analysis also indicated that the *trnH–psbA* sequence was more useful than the *matK* and *rbcL* sequences for this purpose (Pang et al., 2012). Cumulatively, these divergent regions are highly promising for the development of diagnostic cp markers in radish.

4. Conclusions

This is the first report of the complete radish cp genome sequence, and has thus increased the characterization of the radish cp genome and enabled its comparison to the cp genomes of other related species in the Brassicaceae family. Moreover, these data increase the genetic and genomic resources available in radish by adding a new strategy of organellar genome assembly. The cp genome reported here will thus provide valuable tools to aid in the genetic and systematic analysis of the *Raphanus* genomes.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2014.08.038.

Acknowledgements

This work was supported by a grant from the Next-Generation BioGreen21 program (PJ008019; Rural Development Administration, Korea) to NK and HJY.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13, 238.

Dong, W., Liu, J., Yu, J., Wang, L., Zhou, S., 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS One 7, e35071.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., Gibbs, R.A., 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7, e47768.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I., 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 32, W273–W279.

Harris, R.S., 2007. Improved Pairwise Alignment of Genomic DNA. Computer Science and Engineering, The Pennsylvania State University, State College.

Hildebrand, M., Hallick, R.B., Passavant, C.W., Bourque, D.P., 1988. Trans-splicing in chloroplasts: the *rps 12* loci of *Nicotiana tabacum*. Proc. Natl. Acad. Sci. U. S. A. 85, 372–376.

Hu, Z.-Y., Hua, W., Huang, S.-M., Wang, H.-Z., 2011. Complete chloroplast genome sequence of rapeseed (*Brassica napus* L.) and its evolutionary implications. Genet. Resour. Crop. Evol. 58, 875–887.

Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.

Kim, K.J., Lee, H.L., 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 11, 247–261.

Kim, S., Lee, Y.P., Lim, H., Ahn, Y., Sung, S.K., 2009. Identification of highly variable chloroplast sequences and development of cpDNA-based molecular markers that distinguish four cytoplasm types in radish (*Raphanus sativus* L.). Theor. Appl. Genet. 119, 189–198.

Kitashiba, H., Li, F., Hirakawa, H., Kawanabe, T., Zou, Z., Hasegawa, Y., Tonosaki, K., Shirasawa, S., Fukushima, A., Yokoi, S., Takahata, Y., Kakizaki, T., Ishida, M., Okamoto, S., Sakamoto, K., Shirasawa, K., Tabata, S., Nishio, T., 2014. Draft sequences of the radish (*Raphanus sativus* L.) genome. DNA Res. http://dx.doi.org/10.1093/dnares/dsu014.

Koller, B., Fromm, H., Galun, E., Edelman, M., 1987. Evidence for in vivo trans splicing of pre-mRNAs in tobacco chloroplasts. Cell 48, 111–119.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Adam, M.P., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. 30, 693–700.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

Lin, S.H., Liao, Y.C., 2013. CISA: contig integrator for sequence assembly of bacterial genomes. PLoS One 8, e60843.

Lü, N., Yamane, K., Ohnishi, O., 2008. Genetic diversity of cultivated and wild radish and phylogenetic relationships among *Raphanus* and *Brassica* species revealed by the analysis of *trnK/matK* sequence. Breed. Sci. 58, 15–22.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Moghe, G.D., Hufnagel, D.E., Tang, H., Xiao, Y., Dworkin, I., Town, C.D., Conner, J.K., Shiu, S. H., 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.

Pang, X., Liu, C., Shi, L., Liu, R., Liang, D., Li, H., Cherny, S.S., Chen, S., 2012. Utility of the *trnH–psbA* intergenic spacer region and its combinations as plant DNA barcodes: a meta-analysis. PLoS One 7, e48833.

Ridley, C.E., Kim, S.C., Ellstrand, N.C., 2008. Bidirectional history of hybridization in California wild radish, *Raphanus sativus* (Brassicaceae), as revealed by chloroplast DNA. Am. J. Bot. 95, 1437–1442.

Saski, C., Lee, S.B., Fjellheim, S., Guda, C., Jansen, R.K., Luo, H., Tomkins, J., Rognli, O.A., Daniell, H., Clarke, J.L., 2007. Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. Theor. Appl. Genet. 115, 571–590.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Tabata, S., 1999. Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res 6, 283–290.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., Sugiura, M., 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5, 2043–2049.

Song, K.M., Osborn, T.C., Williams, P.H., 1988. Brassica taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs): 2. Preliminary analysis of subspecies within *B. rapa* (syn. campestris) and *B. oleracea*. Theor. Appl. Genet. 76, 593–600.

Song, K., Osborn, T.C., Williams, P.H., 1990. Brassica taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs): 3. Genome relationships in Brassica and related genera and the origin of *B. oleracea* and *B. rapa* (syn. campestris). Theor. Appl. Genet. 79, 497–506.

Stothard, P., Wishart, D.S., 2005. Circular genome visualization and exploration using CGView. Bioinformatics 21, 537–539.

Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11, 1441–1452.

Thormann, C.E., Ferreira, M.E., Camargo, L.E., Tivang, J.G., Osborn, T.C., 1994. Comparison of RFLP and RAPD markers to estimating genetic relationships within and among cruciferous species. Theor. Appl. Genet. 88, 973–980.

Wang, A., Yang, M., Liu, J., 2005. Molecular phylogeny, recent radiation and evolution of gross morphology of the rhubarb genus *Rheum* (Polygonaceae) inferred from chloroplast DNA *trnL-F* sequences. Ann. Bot. 96, 489–498.

Warwick, S.I., Black, L.D., 1991. Molecular systematics of *Brassica* and allied genera (Subtribe Brassicinae, Brassiceae)—chloroplast genome and cytodeme congruence. Theor. Appl. Genet. 82, 81–92.

Warwick, S.I., Black, L.D., 1997. Phylogenetic implications of chloroplast DNA restriction site variation in subtribes Raphaninae and Cakilinae (Brassicaceae, tribe Brassiceae). Can. J. Bot. 75, 960–973.

Wolf, P.G., Der, J.P., Duffy, A.M., Davidson, J.B., Grusz, A.L., Pryer, K.M., 2011. The evolution of chloroplast genes and genomes in ferns. Plant Mol. Biol. 76, 251–261.

Wolfe, K.H., Li, W.H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84, 9054–9058.

Wu, J., Liu, B., Cheng, F., Ramchiary, N., Choi, S.R., Lim, Y.P., Wang, X.W., 2012. Sequencing of chloroplast genome using whole cellular DNA and solexa sequencing technology. Front. Plant Sci. 3, 243.

Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252–3255.

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.

Yamane, K., Lü, N., Ohnishi, O., 2005. Chloroplast DNA variations of cultivated radish and its wild relatives. Plant Sci. 168, 627–634.

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.

Yang, Y.W., Lai, K.N., Tai, P.Y., Ma, D.P., Li, W.H., 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.

Yang, Y.W., Tai, P.Y., Chen, Y., Li, W.H., 2002. A study of the phylogeny of *Brassica rapa*, *B. nigra*, *Raphanus sativus*, and their related genera using noncoding regions of chloroplast DNA. Mol. Phylogenet. Evol. 23, 268–275.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K., Yu, J., 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4, 259–263.

Zhang, T., Fang, Y., Wang, X., Deng, X., Zhang, X., Hu, S., Yu, J., 2012. The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. PLoS One 7, e30531.

Zhang, H., Li, C., Miao, H., Xiong, S., 2013. Insights from the complete chloroplast genome into the evolution of *Sesamum indicum* L. PLoS One 8, e80508.