The Complete Chloroplast Genome Sequence of Korean Landrace “Subicho” Pepper (*Capsicum annuum* var. *annuum*)

Sebastin Raveendar[†], Young-Ah Jeon[†], Jung-Ro Lee, Gi-An Lee, Kyung Jun Lee, Gyu-Taek Cho, Kyung-Ho Ma, Sok-Young Lee, Jong-Wook Chung*

National Agrobiodiversity Centre, National Academy of Agricultural Science, Rural Development Administration, Jeonju 560-500, Republic of Korea

ABSTRACT Chloroplast DNA sequences are a versatile tool for species identification and phylogenetic reconstruction of land plants. Different chloroplast loci have been utilized for phylogenetic classification of plant species. However, there is no report for a short DNA sequence that can distinguish all plant species from each other. Molecular markers derived from the complete chloroplast genome can provide effective tools for species identification and phylogenetic resolution. Thus, the complete chloroplast genome sequence of Korean landrace “Subicho” pepper (*Capsicum annuum* var. *annuum*) has been determined here. The total length of the chloroplast genome is 156,878 bp, with 37.7% overall GC content. A pair of IRs (inverted repeats) of 25,801 bp was separated by a small single copy (SSC) region of 17,929 bp and a large single copy (LSC) region of 87,347 bp. The chloroplast genome harbors 132 known genes, including 87 protein-coding genes, 8 ribosomal RNA genes, and 37 tRNA genes. A total of seven of these genes are duplicated in the inverted repeat regions, nine genes and six tRNA genes contain one intron, while two genes and a *ycf* have two introns. Analysis revealed 144 simple sequence repeat (SSR) loci and 96 variants, mostly located in the intergenic regions. The types and abundances of repeat units in *Capsicum* species were relatively conserved and these loci will be useful for developing *C. annuum* cp genome vectors.

Keywords *Capsicum annuum*, Chloroplast, DNA sequencing, Subicho pepper

INTRODUCTION

Chloroplasts are the large organelles responsible for photosynthesis that converts CO₂ into carbohydrates. It is widely agreed that chloroplast was derived from a single ancestral cyanobacterium into the ancestral plant by one symbiogenetic event (Gray 1999). The chloroplast is unusual in containing its own genetic system, which replicates by division. DNA fragments of the chloroplast genome have been widely used for phylogenetic reconstruction and species-level identification in land plants (Dong *et al.* 2012; Dong *et al.* 2015; Hollingsworth *et al.* 2011; Group *et al.* 2009). Because the chloroplast genome has a simple and stable genetic structure, there are no or rare recombinations, and universal primers have been used to amplify target sequences. In general, the chloroplast gene structure

in angiosperms is conserved, with inverted repeat (IR) regions separated by small (SSC) and large (LSC) single copy regions (Palmer 1991).

Capsicum L. (pepper) is a member of the Solanaceae family and comprises about 32 recognized species (Moscone *et al.* 2007). *Capsicum* originated in the New World and is cultivated in temperate and tropical regions (Eshbaugh 1993; Pozzobon *et al.* 2005). The exact nature of when, where, and how the process of *Capsicum* domestication occurred is unknown. *Capsicum annuum* var. *annuum*, is a unique *Capsicum* species known as the aji Amarillo chilli or Peruvian hot pepper and is well-known as “Subicho” in Korea for its rich variation in flavors and aromas. Pepper has important roles in various aspects of the economy, in food and in pharmaceuticals (Kumar *et al.* 2006). Therefore, knowledge of the genetic diversity among the germplasm is

Received Jun 13, 2015; Revised Jun 19, 2015; Accepted Jun 21, 2015; Published June 30, 2015

*Corresponding author Jong-Wook Chung, jwchung73@korea.kr, Tel: +82-63-238-4872, Fax: +82-63-238-4859

[†] These authors contributed equally to this work

vital for strategic germplasm collection, maintenance, conservation and utilization. Molecular methods based on DNA sequence analysis provide useful information for taxonomy, species identification, and phylogenetics. In the last few decades molecular phylogenetics has rapidly developed, and is gaining increasing importance in resolving phylogenetic relationships.

Currently, the interspecies relationships within the genus *Capsicum* remain highly controversial. Molecular phylogenetic research on *Capsicum* has been extensively carried, but there is no apparent structure to reveal the true phylogenetic relationships between its species. The major reason for the lack of phylogenetic structure is because the genus *Capsicum* contains a wide variety of species. In addition, the lack of appropriate DNA sequences greatly limits the ability to perform adequate molecular phylogenetic research on *Capsicum*. Most of the phylogenetic studies performed to date have limited on availability of suitable DNA sequences. Consequently, achieving phylogenetic resolution and performing species identification have been almost impossible. Two chloroplast genome sequences from *Capsicum* species, the American bird pepper (*Capsicum annuum* var. *glabriusculum*) (Zeng *et al.* 2014) and cultivated pepper (*Capsicum annuum* L.) (Jo *et al.* 2011), were reported previously. The complete chloroplast genome sequence of the Korean landrace “Subicho” pepper *Capsicum annuum* var. *annuum* reported here enriches the gene information for *C. annuum* var. *annuum* and contributes to further study of population, phylogenetic and cp genetic engineering of this species.

MATERIALS AND METHODS

Sampling and DNA extraction

Subicho seeds (Accession No. IT216436) were collected from the National Agrobiodiversity Center, Rural Development Administration, Republic of Korea. Seeds were germinated and grown in a greenhouse, and fresh leaves were collected from 40-day-old seedlings, and DNA was extracted to construct cp DNA libraries.

Library preparation and sequencing

An Illumina paired-end cp DNA library (average insert size of 500 bp) was constructed using the Illumina TruSeq library preparation kit following the manufacturer’s instructions. The libraries were sequenced with 2 x 300 bp on the MiSeq instrument at LabGenomics (<http://www.Lab.genomics.com/kor/>).

Chloroplast genome assembly

Prior to cp de novo assembly, low quality sequences (quality score < 20; Q20) were filtered out, and the remaining high quality reads were assembled using the CLC Genome Assembler (version beta 4.6, CLC Inc. Aarhus, Denmark) with a 200-600-bp overlap size. Cp contigs were selected from the initial assembly by performing a BLAST (ver. 2.2.31) search against known cp sequences. The selected contigs were oriented to construct the complete cp genome structure. Ambiguous nucleotides or gaps were corrected manually to build the complete cp genome.

Gene annotation

The web-based program Dual OrganellarGenoMe Annotator (DOGMA, <http://dogma.cccb.utexas.edu/>) was used to annotate the assembled genome using default parameters to predict protein coding, tRNA and rRNA genes. Subsequently, BLASTN (ver. 2.2.31) was used to further identify intron-containing gene positions by searching a published cp genome database (GenBank accession NC_018552; Jo *et al.* 2011). A cp gene map was constructed using the OrganellarGenomeDRAW software (OGDRAW, <http://ogdraw.mpimp-golm.mpg.de>).

Discovery of SNPs and SSRs

Sputnik (<http://espressoftware.com/pages/sputnik.jsp>) software was used to find the SSR markers present in the cp genome of *C. annuum* var. *annuum*. It uses a recursive algorithm to search for repeats with length between 2 and 5, and finds perfect, compound and imperfect repeats. Sputnik has been applied for SSR identification in many species including *Arabidopsis* and barley (Cardle *et al.* 2000). To identify SNP and INDEL variants in *C. annuum* var. *annuum* cp genome, we used BWA (Li and Durbin

2009) and Samtools (Li *et al.* 2009) software. More detailed method and algorithm are described in Li (2012).

RESULTS

Chloroplast genome assembly

We sequenced the cp genome of Korean landrace

“Subicho” pepper *C. annuum* var. *annuum* using the Illumina genome analyzer platform. Illumina paired-end (2×300 bp) sequencing produced a total of 7,100,754 paired-end reads, with an average fragment length of 256 bp, which were then analyzed to generate 1,536,526,037 bp of sequence. Low quality reads (Q20) were filtered out, and the remaining high quality reads were mapped to the reference cp genome of *C. annuum* L. (GenBank accession

Fig. 1. Complete genome map of the Korean landrace “Subicho” pepper *C. annuum* var. *annuum* chloroplast. Genes drawn inside the circle are transcribed clockwise, while those outside are counterclockwise and marked with two arrows. Differential functional gene groups are color-coded. The GC content variation is shown in the middle circle.

NC_018552), which contains 20,161,616 mapped nucleotides with an average coverage of 128× on the cp genome. The cp reads extracted from the Illumina dataset were assembled into a total of four contigs. Contig alignment and scaffolding based on paired-end data resulted in a complete circular *C. annuum* var. *annuum* cp genome sequence (Fig. 1). We have successfully applied a whole genome sequencing approach to determine the complete chloroplast genome sequence of Korean landrace “Subicho” pepper *C. annuum* var. *annuum*. This is the basis for this whole cp genome sequencing strategy, as it allows micro-reads to be assembled correctly using a reference-guided method. The fully annotated chloroplast genome sequence of “Subicho” pepper *C. annuum* var. *annuum* has been deposited in the GenBank database under accession number KR078313.

Features of the chloroplast genome

The chloroplast genome of Korean landrace “Subicho” pepper *C. annuum* var. *annuum* reported here has a total length of 156,878 bp (Figure 1) and is composed of a LSC region of 87,347 bp, two IR copies (IRA and IRB) totalling 25,801 bp and a SSC region of 17,929 bp. The whole chloroplast genome of “Subicho” pepper *C. annuum* var. *annuum* was 97 bp longer than the reported *C. annuum* L. chloroplast genome (GenBank accession NC_018552) and

266 bp longer than the *C. annuum* var. *glabriusculum* chloroplast genome (GenBank accession KJ619462). Also, the SSC and IR regions of *C. annuum* var. *annuum* were 80 and 78 bp longer, respectively, and its LSC region was 19 and 131 bp, respectively, shorter than the previously reported chloroplast genomes. The overall GC content of the chloroplast genome was 37.7%, and in the LSC, SSC and IR regions it was 35.7%, 31.9%, and 43.0%, respectively. The chloroplast genome includes 132 unique genes, composed of 87 protein-coding genes, 37 tRNA genes and 8 rRNA genes (Table 1). A total of seven of these genes are duplicated in the inverted repeat regions, nine genes and six tRNA genes contain one intron, while two genes (*clpP* and *rps12*) and a *ycf* (*ycf3*) have two introns.

Discovery of SSRs and SNPs

A total of 144 potential SSRs motifs were identified which are located only in the intergenic region (Table S1), and the majority belonged to tetra-nucleotide (50%) and penta-nucleotide (21.5%) repeats. All other types of SSRs such as di and tri nucleotide motifs were relatively low (28.5%), and the majority of tetra-nucleotide SSRs had the ATAA/TAAA/AAAT motif, followed by those with the AAAT/AATA/ATAA motif, and the remaining those with the TTTG/TTGT/TGTT, TCTT/CTTT/TTTC, and AATT/ATTA/TTAA motifs were found with similar proportion (6.25%). Three different repeats those with the ATATT/TATTA/ATTAT, TTTTA/TTTAT/TTATT, and TTATT/TATTT/ATTTT motifs were identified among penta-nucleotide SSRs. The TTC/TCT/CTT and TTA/TAT/ATT motifs were identified among the tri-nucleotide SSRs. Only, the TA/AT motif was identified as the dinucleotide SSRs (Table S1). Comparison of *C. annuum* var. *annuum* cp genome sequence with the reference cp sequence of *C. annuum* revealed a total of 96 mutations (45 SNPs and 51 InDels) and 46 of these variants involving more than one nucleotide (Table S2 and S3). Amongst the detected variants, 13 SNPs and 3 InDels were observed in the coding region of the cp genome. Amongst these SNPs and InDels, there were 78, 17 and 1 mutations located in LSC, SSC, and IRa region, respectively.

Table 1. General features of the *C. annuum* var. *annuum* chloroplast genome

Features	Chloroplast
Genome size (bp)	156,878
GC content (%)	37.7
Total number of genes	132
Protein coding genes	87
No. of rRNA genes	8
No. of tRNA genes	37
No. of gene duplicated in IR regions	7
Total introns	12
Single intron (gene)	9
Double introns (gene)	3
Single intron (tRNA)	6

DISCUSSION

Here we report the re-sequencing and assembly of a cp genome using the Illumina sequencing platform in which we recovered five contigs comprising 156,878 bp covering the entire *C. annuum* var. *annuum* cp genome. Reported Capsicum cp genomes range in size from 156,612 to 156,781 bp, and the size of the *C. annuum* var. *annuum* cp genome identified here is consistent with those reported previously in plants of the same species (Zeng *et al.* 2014; Jo *et al.* 2011). The entire cp genome of *C. annuum* var. *annuum* was 97 bp longer than the reported *C. annuum* L. cp genome (GenBank accession NC_018552) and 266 bp longer than *C. annuum* var. *glabriusculum* cp genome (GenBank accession KJ619462). Also, the SSC and IR regions of *C. annuum* var. *annuum* were 80 and 77 bp longer, respectively, and the LSC region was 19 bp and 33 bp shorter, respectively, than those of the previously reported cp genomes.

The average GC content in the *C. annuum* cp genome is 37.7%, similar to other capsicum species. The data generated using the Illumina platform covered a greater depth (128 ×) of the cp genome whereas, in the previous studies cp genome sequence coverage was not reported and were able to resolve the ambiguities present in the GS-FLX pyrosequencing. Thus, the data from the cp assembly reported here supports previous findings that Illumina can produce high quality sequence assemblies covering a greater genome depth (Wu *et al.* 2014). For the cp genome assembly, small reads alignment to reference genome is an alternative approach to the strategy of *de novo* assembly. The read mapping approach is computationally less demanding and faster than *de novo* assembly. Moreover, it has the advantage that the read coverage information can be used for reliable detection of sequence variation.

Cp structural rearrangements and gene loss-and-gain events often occur in some angiosperms and are especially common in monocot cp genomes. The organization and

Table 2. Genes present in the *C. annuum* var. *annuum* chloroplast genome.

Gene products of <i>Capsicum annuum</i> var. <i>annuum</i>	
Photosystem I	psaA, B, C, I, J, ycf3 ²⁾ , ycf4
Photosystem II	psbA, B, C, D, E, F, H, I, J, K, L, M, N, T, Z
Cytochrome b6/f	petA, B ¹⁾ , D ¹⁾ , G, L, N
ATP synthase	atpA, B, E, F ¹⁾ , H, I
Rubisco	rbcL
NADH oxidoreductase	ndhA ¹⁾ , B ¹⁾³⁾ , C, D, E, F, G, H, I, J, K
Large subunit ribosomal proteins	rpl2 ¹⁾³⁾ , 14, 16 ¹⁾ , 20, 22, 23 ³⁾ , 32, 33, 36
Small subunit ribosomal proteins	rps2, 3, 4, 7 ³⁾ , 8, 11, 12 ²⁾³⁾⁴⁾ , 14, 15, 16 ¹⁾ , 18, 19
RNA polymerase	rpoA, B, C1 ¹⁾ , C2
Unknown function protein coding gene	ycf1 ³⁾ , 2 ³⁾ , 15 ³⁾
Other genes	accD, ccsA, cemA, clpP ²⁾ matK
Ribosomal RNAs	rrn16 ³⁾ , 23 ³⁾ , 4.5 ³⁾ , 5 ³⁾
Transfer RNAs	trnA-UGC ¹⁾³⁾ , trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-UCC ¹⁾ , trnG-GCC, trnH-GUG, trnI-CAU ³⁾ , trnI-GAU ¹⁾³⁾ , trnK-UUU ¹⁾ , trnL-UAA ¹⁾ , trnL-UAG, trnL-CAA ³⁾ , trnM-CAU, trnM-CAU, trnN-GUU ³⁾ , trnP-UGG, trnQ-UUG, trnR-ACG ³⁾ , trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-UAC ¹⁾ , trnV-GAC ³⁾ , trnW-CCA, trnY-GUA

¹⁾ Gene containing a single intron

²⁾ Gene containing two introns

³⁾ Two gene copies in IRs

⁴⁾ Trans-splicing gene

gene order of the *Capsicum* cp genome exhibited the general cp genome structure of angiosperms (Sugiura 1992). The *Capsicum* cp genome contained 132 genes (Table 2), of which there were 8 rRNA genes, 37 tRNA genes, 21 ribosomal subunit genes (12 small subunit and 9 large subunit) and 4 DNA-directed RNA polymerase genes. Forty-six genes were involved in photosynthesis, of which 11 encoded subunits of the NADH-oxidoreductase, 7 for photosystem I, 15 for photosystem II, 6 for the cytochrome b6/f complex, 6 for different subunits of ATP synthase and 1 for the large chain of ribulose biphosphate carboxylase. Five genes were involved in different functions, and three genes were of unknown function. As shown in Figure 1 and Table 2, genome organization appeared to be more conserved with unique gene sequences, as discovered previously in *Capsicum* species (Zeng *et al.* 2014; Jo *et al.* 2011). However, in this newly determined cp genome, we found 132 predicted genes and size variations were observed in the IR and LSC regions.

A total of 144 cpSSRs markers were identified in 156.8 kb sequence of the capsicum chloroplast genome. The observed frequency of SSRs was approximately 1/1.08 kb of chloroplast genome. This was higher than the SSR frequency previously observed for the *C. annuum* cp genome (1/2.90 kb) (Jo *et al.* 2011). More interestingly, the cpSSRs were only observed in the non-coding region of the cp genome. Similarly, most of the SNPs and InDels in the cp genome present in intergenic region, and only 16 variants were located in genic region (Table S2 and S3).

CONCLUSION

Solanaceae is an important ethnobotanical family of dicots comprising more than 3,000 species and is extensively utilized by humans and has recently become a model of comparative and evolutionary genomics research. The cp genome sequences of *Capsicum* species have been reported previously; however, information on cp gene content is limited. Here, we report the re-sequencing and assembly of a Korean landrace “Subicho” pepper *C. annuum* var. *annuum*. The cp genome is well-conserved in terms of size, gene arrangement, and coding sequences, within major

subgroups of the plant kingdom. This represents the cp genome sequence of pepper described herein is a valuable resource for studying *Capsicum* population, phylogenetic and cp genetic engineering of this genus.

ACKNOWLEDGMENTS

This study was carried out with the support of the “Research Program for Agricultural Science & Technology Development (Project No. PJ008623)” and was supported by the 2014 Postdoctoral Fellowship Program of the National Academy of Agricultural Science, Rural Development Administration, Republic of Korea.

REFERENCES

- Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156: 847-854.
- Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015. *ycf1*, the most promising plastid DNA barcode of land plants. *Sci Rep* 5: 8348.
- Dong WP, Liu J, Yu J, Wang L, Zhou SL. 2012. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. *PLoS One* 7: e35071.
- Eshbaugh WH. 1993. History and exploitation of a serendipitous new crop discovery. *New Crops*. Wiley, New York.
- Gray MW. 1999. Evolution of organellar genomes. *Current Opinion in Genetics & Development* 9: 678-687.
- Group CPW, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, *et al.* 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106: 12794-12797.
- Hollingsworth PM, Graham SW, Little DP. 2011. Choosing and using a plant DNA barcode. *PLoS One* 6: e19254.
- Jo YD, Park J, Kim J, Song W, Hur CG, Lee YH, Kang BC. 2011. Complete sequencing and comparative analyses of the pepper (*Capsicum annuum* L.) plastome revealed high frequency of tandem repeats and large insertion/

- deletions on pepper plastome. Plant cell reports 30: 217-229.
- Kumar S, Kumar R, Singh J. 2006. Cayenne/American pepper (*Capsicum* species). In: Peter KV (ed) Handbook of herbs and spices Woodhead, Cambridge: 299-312.
- Li H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28: 1838-1844.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
- Moscone EA, Scaldaferrro MA, Grabiele M, Cecchini NM, Sánchez García Y, Jarret R, Daviña JR, Ducasse DA, Barboza GE, Ehrendorfer F. 2007. The evolution of chili peppers (*Capsicum*-*Solanaceae*): A cytogenetic perspective. Acta Horti 745: 137-170.
- Palmer JD. 1991. Plastid chromosomes: Structure and evolution. Mol Biol Plastids 7: 5-53.
- Pozzobon MT, Schifino-Wittmann MT, Bianchetti LDB. 2005. Chromosome numbers in wild and semidomesticated *Brazilian Capsicum* L. (*Solanaceae*) species: do $x = 12$ and $x = 13$ represent two evolutionary lines? Botanical J of the Linnean Soc 151: 259-269.
- Sugiura M. 1992. The chloroplast genome. Plant Molecular Biology 19: 149-168.
- Wu ZH, Gui ST, Quan ZW, Pan L, Wang SZ, Ke WD, Liang DQ, Ding Y. 2014. A precise chloroplast genome of *Nelumbo nucifera* (*Nelumbonaceae*) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. BMC Plant Biol 14: 289.
- Zeng FC, Gao CW, Gao LZ. 2014. The complete chloroplast genome sequence of American bird pepper (*Capsicum annuum* var. *glabriusculum*). Mitochondrial DNA. doi: 10.3109/19401736.2014.913160.