A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Molecular Genetic Analysis of Cytoplasmic Male Sterility by Mitochondrial Genome Sequencing and Isolation of a *Restorer-of-fertility* Candidate Gene in Pepper (*Capsicum annuum* L.)

미토콘드리아 유전체 분석 및 웅성불임 회복 후보 유전자 동정을 통한 고추 세포질 웅성불임 기작 연구

AUGUST, 2013

Yeong Deuk Jo

MAJOR IN HORTICULTURAL SCIENCE

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Molecular Genetic Analysis of Cytoplasmic Male Sterility by Mitochondrial Genome Sequencing and Isolation of a *Restorer-of-fertility* Candidate Gene in Pepper (*Capsicum annuum* L.)

UNDER THE DIRECTION OF DR. BYOUNG-CHEORL KANG
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF
SEOUL NATIONAL UNIVERSITY

BY
YEONG DEUK JO

MAJOR IN HORTICULTURAL SCIENCE
DEPARTMENT OF PLANT SCIENCE

AUGUST, 2013

APPROVED AS A QUALIFIED DISSERTATION OF YEONG DEUK JO
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
BY THE COMMITTEE MEMBERS

CHAIRMAN	**Doil Choi, Ph.D.**
VICE-CHAIRMAN	**Byoung-Cheorl Kang, Ph.D.**
MEMBER	**Tae-Jin Yang, Ph.D.**
MEMBER	**Sunggil Kim**
MEMBER	**Peter van Dijk, Ph.D.**

Molecular Genetic Analysis of Cytoplasmic Male Sterility by Mitochondrial Genome Sequencing and Isolation of a *Restorer-of-fertility* Candidate Gene in Pepper (*Capsicum annuum* L.)

YEONG DEUK JO

Department of Plant Science, Seoul National University

ABSTRACT

Cytoplasmic-genic male sterility (CGMS), which involves the interactions and conflicts between mitochondrial and nuclear genomes, has been widely used for hybrid seed production in crops including chili peppers. Although the candidates for CMS-associated mitochondrial gene were isolated and molecular markers linked to *Restorer-of-fertility* (*Rf*) have been developed in pepper, the understanding of CGMS mechanism and the utilization of reliable molecular breeding system have been limited. Therefore, comparative analysis of CMS and

normal mitochondrial genomes, isolation of an *Rf* candidate gene, and deduction of the origin of CMS cytoplasm were performed in this study.

In the first chapter, comparative analysis between mitochondrial genomes from a CMS pepper line, FS4401 and a fertile pepper line, Jeju was conducted. The complete mitochondrial genomes which were 507,450 and 493,911bp in length were assembled in FS4401 and Jeju, respectively. Although most of gene contents were conserved between two genomes, extensive rearrangement of genome structure resulted in the generation of eighteen blocks of sequences which were syntenic between two mitochondrial genomes and unaligned sequence segments between them. The CMS candidate genes, *orf507* and *ψatp6-2*, were located on the edges of the largest sequence segments which were specific to FS4401. Although severe rearrangements and lack of any similarity with reported sequences in this region hampered the elucidation of detailed mechanism, presence of repeated and overlap sequences on DNA segments implied that extensive rearrangements by nonhomologous end-joining followed by substoichiometric shift due to recombination through repeated sequence might be involved in generation and integration of this region on the master DNA molecule. Further analysis using mtDNA pairs of CMS-normal cytoplasms in other plant species showed common features of DNA regions around CMS-associated genes.

In the second chapter, three kinds of mapping strategies were used to isolate pepper *Rf* gene. Firstly, pepper BAC clones which contain sequences

homologous to petunia *Rf* gene were screened and mapped. Secondly, AFLP analysis was performed using more than one thousand primer combinations. Finally, comparative mapping was conducted using tomato genome sequence. As the result, a group of selected BAC clones and AFLP markers were mapped on pepper DNA region that was co-segregated with the *Rf* gene. By six times of chromosome walking started from this region, the sequence which spanned DNA region containing *Rf* was obtained. Prediction of expressed genes in this sequence using transcriptome analysis screened an *Rf*-candidate gene, PPR6. The PPR6 gene encoded a pentatricopeptide repeat protein in which degenerative 35 amino acid motif was repeated for fourteen times. Specific expression of this gene in restorer lines showed that PPR6 is a strong candidate for *Rf* in pepper.

In the third chapter, the origin of CMS cytoplasm was deduced by using plastid DNA markers. The complete sequence of plastid genome in a pepper line 'FS4401' was assembled and used as the source for marker development. Two plastid sequences, *trnH-psbA* and *rpl16-rpl18* intergenic sequences, were used to analyze cytoplasm types of pepper germplasms which include six *Capsicum* species. Plastid barcode analysis revealed that cytoplasm types can be divided into six types and four types for *trnH-psbA* and *rpl16-rpl18*, respectively. The sequences on these two regions in CMS pepper lines were identical to the sequences of a cytoplasm type of a particular clade in *C. annuum*. For further investigation, two molecular markers were designed from *TrnL-TrnF* and *rpl16-*

rpl18 intergenic regions, respectively. Application of these markers to a larger number of germplasm confirmed that the cytoplasm type of CMS is identical to the cytoplasm type of the particular *C. annuum* clade. These results suggest that the CMS cytoplasm of pepper may originate from an interspecific cross in which the seed parent belonged to the *C. annuum* clade.

The results of this study are expected to contribute to reliable and efficient molecular breeding for CGMS system as well as provide insights in evolution of CMS cytoplasm and *Rf* gene in pepper.

Keywords: *Capsicum annuum*, Cytoplasmic male steilirity (CMS), *Restorer-of-fertility* (*Rf*), Mitochondria, Chloroplast

Student Number: 2007-30299

CONTENTS

CHAPTER II

Isolation of the *Restorer-of-fertility* Gene in Pepper (*Capsicum annuum* L.)

CHAPTER III

Utilization of Chloroplast Genome Sequences for the Determination of the Origin of CMS Cytoplasm and Development of a Reliable Marker Associated with CMS

LIST OF ABBREVIATIONS

BAC	Bacterial artificial chromosome
BLAST	Basic local alignment sequence tool
CAPS	Cleaved amplified polymorphic sequence
cDNA	Complimentary deoxyribonucleic acid
cM	Centimorgan
CMS	Cytoplasmic male sterility
EST	Expressed sequence tag
LRR	Leucine rich repeats
mtDNA	Mitochondrial DNA
NBS	Nucleotide binding site
ORF	Open reading frame
PPR	Pentatricopeptide repeat protein
Rf	*Restorer-of-fertility*
RFL	*Rf*-like gene
SCAR	Sequence characterized amplified region
SNP	Single nucleotide polymorphism
SSS	Substoichiometric shift

GENERAL INTRODUCTION

Cytoplasmic male sterility (CMS), which is a maternally inherited characteristics, causes production of flowers with nonfunctional pollen. CMS has been reported in more than 150 plant species and is used to eliminate the laborious procedures of emasculation and hand-pollination in F_1 hybrid seed production (Schnable and Wise, 1998). Along with CMS, the nuclear fertility restoration gene, which suppresses the expression of CMS in mitochondria, has been used for hybrid seed production in both cereal and horticultural crops such as rice, radish, pepper, and petunia (Hanson and Bentolila, 2004). In addition to agricultural applications, the CMS and fertility-restoration phenomena provide excellent model systems for studying the interactions between mitochondrial and nuclear genomes at the molecular level (Hanson and Bentolila, 2004).

CMS has been known to be caused by chimearic genes which originated from novel rearrangements on mitochondrial genome. Plant mitochondrial genomes are highly dynamic in that structural variations including changes in gene orders, rearrangements, genome expansion and shrinkage, and incorporation of foreign DNAs very extensively occur (Palmer et al, 1988; Palmer, 1990; Wolfe et al, 1987). These unique characteristics of plant mitochondrial genome have been explained by existence of reservoir of subgenomic mtDNA molecules under

copy number suppression and dispersed repeated sequences which have potential to mediate frequent or rare recombinations (Arrieta-Montiel et al, 2001; Small et al, 1989).

Complete sequences were reported for more than fifty mitochondrial genomes so far. Especially, complete sequencing and comparative analysis between normal and CMS cytoplasm have been performed in several crop species including sugar beet, maize, wheat, rice, rapeseed, and radish to analyze the structural variation of mtDNA and identify candidate *orf*s associated with CMS (Allen et al., 2007; Chen et al., 2011; Liu et al., 2011; Park et al., 2013; Satoh et al., 2004; Tanaka et al., 2012). These studies showed that mitochondrial genome structure in CMS cytoplasm were extensively different from normal cytoplasm although gene contents were mostly conserved. For example, in sugar beet, normal and CMS mitochondrial genomes were composed by different arrays of fourteen sequence blocks which are syntenic between two genomes (Satoh et al., 2004). Recently, Tanaka et al. (2012) showed that a radish CMS mitochondrial genome, Ogura cytoplasm, contained large CMS-specific region in addition to syntenic block sequences. However, the origin of CMS-associated gene sequence and other CMS-specific sequence in this region was still remained to be unknown.

Rf genes have been identified in maize (Cui et al, 1996), petunia (Bentolila et al., 2002), rice (Komori et al., 2004; Hu et al., 2012; Fujii and Toriyama, 2009; Itabashi et al., 2011), radish (Brown et al., 2003; Desloire et al., 2003; Koizuka et

al., 2003), and sugar beet (Matsuhira et al., 2012). Most of *Rf* or genetically defined *Rf* candidate genes encoding pentatricopeptide repeat (PPR) proteins have a repeated motif composed of a degenerative array of 35 amino acids that may bind RNA through its superhelix structure (Small and Peeters, 2000). Supporting the idea that PPR protein binds RNA, many of PPR type *Rf* genes were shown to be involved in the processing or degradation of transcripts of CMS-associated genes (Gillman et al., 2007; Hu et al., 2012; Wang et al., 2006)

The large PPR gene family includes a total of 441 genes in Arabidopsis (Lurin et al., 2004). Although most of them are evenly distributed along the five Arabidopsis chromosomes, a total of nineteen genes are clustered into a region of less than 1 Mb on chromosome 1. Interestingly, the genes in this cluster included the closest homologs to the PPR encoding *Rf* genes, and similarity between genes was also high. Fujii et al. (2011) defined this kind of PPR genes among plant taxa as *Rf*-like (*RFL*) genes and showed that *RFL*s have undergone diversifying selection during the co-evolution with CMS cytoplasm. Characteristics of *Rf* genes supports the hypothesis that *Rf* genes have evolved by birth-and-death process which is usually found in disease resistance genes to cope with the appearance of new CMS genes (Touzet and Budar, 2004).

In pepper (*Capsicum annuum* L.), CMS was first discovered in an Indian *C. annuum* accession (USDA accession PI 164835), whose cytoplasm has been used as the sole source for CMS (Shifriss, 1997). Various models have been proposed

for how fertility restoration of CMS is inherited, including control by single dominant gene (Zhang et al., 2000b; Gulyas et al. 2006; Kim et al. 2006; Jo et al. 2010; Min et al. 2008; Min et al. 2009), two complementary genes (Novak et al., 1971) and QTL (Wang et al., 2004). In addition, partial restoration of fertility has been reported in some cases (Lee, 2001; Lee et al., 2008)

Candidates for pepper CMS-associated gene have been cloned (Kim et al., 2006; Kim et al., 2007), and several molecular markers linked to the pepper *Rf* gene have been developed (Zhang et al., 2000; Kim et al., 2006; Gulyas et al., 2006; Lee et al., 2008). A chimeric mitochondrial gene, *orf506*, was identified in the mitochondrial genome of CMS peppers, and male sterility could be induced when a portion of this gene was expressed in Arabidopsis (Kim et al., 2007). The another candidate, *ψatp6-2* gene, was generated by novel rearrangement on 3' region of normal *atp6-2*. Transcription patterns of *ψatp6-2* were different between male sterile and restorer lines showing possible association of this gene with CMS (Kim et al., 2006). However, the characterization of the unique rearrangement pattern between CMS and normal pepper cytoplasms has not been performed in mitochondrial genome scale yet. The molecular markers linked to *Rf* include OPP13-CAPS (1.1cM from *Rf*), AFRF8-CAPS (1.8cM from *Rf*), PR-CAPS (1.8cM from *Rf*), AFRF4 (0.1cM from *Rf*), and an STS marker named CRF-SCAR (5.3cM from *Rf*) (Kim, 2005; Kim et al., 2006; Lee et al., 2008; Min et al., 2009; Gulyas et al., 2006). These markers have limited applications in pepper lines,

however, because of failure in PCR amplification, the existence of third haplotypes and lack of agreement between marker genotype and phenotype (Min et al., 2008).

In this study, the complete mitochondrial genome sequence of pepper was firstly reported in fertile and CMS pepper lines. In addition, cloning of the candidate for *Rf* gene by combinational mapping strategies were performed. Finally, the origin of the CMS cytoplasm was studied with plastid DNA markers using diverse pepper germplasm as materials. The results of this study are expected to contribute to reliable and efficient molecular breeding for CGMS system as well as provide insights in evolution of CMS cytoplasm and an *Rf* gene in pepper.

LITERATURE REVIEW

1. Plant mitochondrial genome

Endosymbiont hypothesis suggested that the mitochondria originated from a free-living eubacterial ancestor resembling α-proteobacterium (Gray, 1999). The majority of mitochondrial genes were transferred to nucleus in the course of evolution and this process is still ongoing in flowering plants. (Palmer et al. 2000). The plant mitochondrial genomes are much larger than counterparts in other organisms such as animals, fungi, and insects. For example, even the the smallest plant mitochondrial genomes found in *Brassica* (Palmer, 1986) are 200kb in size, which is more than ten times bigger than animal mitochondrial genome that are 16-20kb in size (Fauron et al., 1995). The size of the biggest plant mitochondrial genome is up to 11.3Mb in *Silene conica* (Sloan et al., 2012).

Complete sequences of mitochondrial genomes have been reported in more than fifty mastercircles so far. After Fujii et al. (2010) started to use 454 GS-FLX system for assembly of mitochondrial genome sequences in rice, next generation sequencing techniques have greatly promoted mtDNA sequencing in many species. However, special care is required in these cases because the sequences derived from plastid or nuclear DNA hamper the discrimination of original plastid or nuclear DNA contaminated during mitochondria isolation and the reference genomes cannot be used in the assembly of newly generated

sequences due to high rearrangement on mtDNA (Bentolila at al., 2012; Kubo et al., 2011).

The higher plant mitochondrial genome include genes which encode subunits of protein complexes in electron transport chain (Complex I to V), large and small ribosomal proteins, cytochrome c maturation-related proteins, ribosomal RNAs, and transfer RNAs. The numbers of genes are from 50 to 69 including protein genes from 30 to 37 among plant species (Kubo et al., 2011). Although cucumber mitochondrial genome (1,685kb), which is one of the largest genome sequences so far, is more than seven times bigger than the smallest one in rapeseed (222kb), it contains only two more genes than rapeseed mtDNA (Handa, 2003; Alverson et al., 2011). The gene coding sequences of plant mitochondrial genomes are highly conserved. The rate of synonymous substitutions in plant mtDNA is known to be 50-100 times and three times lower than in vertebrate mitochondria and chloroplast which is another endosybiont organelle in plant cell (Palmer et al, 1988; Palmer, 1990; Wolfe et al, 1987). However, extensive DNA rearrangements and integrations of foreign sequences were detected on intergenic regions which resulted in lack of synteny between different mitochondrial genomes (Palmer et al., 2000).

Although both chloroplast and the mitochondria have been known to originate endosymbiotically, the patterns and tempos of the evolution of their genomes are strikingly different. In plastid, the overall structures and gene orders

of genomes have been well conserved during speciation. However, mitochondrial genomes of higher plants are not only highly variable in gene order and intergenic sequences between species, but also highly complex in the stoichiometry of subgenomic molecules (Mackenzie and McIntosh, 1999).

The investigation of plant mitochondrial genome using electron microscope showed that various sizes of circular and linear DNA molecules, and even catenane-like or rosette-like structures are present in plant mitochondria (Backert et al. 1996a; Barckert et al. 1996b). On the other hand, less intensity bands other than the bands which represents master circle were detected in restriction enzyme analysis of plant mitochondrial DNA. These ‘sublimons’ implied that plant mitochondrial genome might not be composed of a master circle DNA molecule (Borck and Walbot, 1982). Although some conformations were thought to be intermediates of replications (Backert et al., 1996b), many of variations could be explained by recombinations via repeat sequences.

The large repeats which are usually longer than 1kb in length undergo frequent and reversible recombinations that give result in the generation of subgenomic DNAs (Andre, 1992). For example, mitochondrial DNA of *Brassica campestris* in which the master circle is 218 kb in size can be recombined via 2 kb direct repeat to generate 135 and 83 kb subgenomic molecules. (Palmer and Shields, 1984) In this kind of recombination, the various recombination products were detected to be in almost the same stoichiometry by Southern blot analysis

(Palmer and Shields, 1984; Folkerts and Hanson, 1989; Siculella et al., 2001; Sloan et al., 2010). However, recent studies based on paired end sequencing analysis showed that the stoichiometry of each recombination products are regulated to be different in monkey flower and cucumber (Mower et al., 2012; Alverson et al., 2011).

The recombinations via smaller repeated sequences are rarly occur and irreversible (Andre, 1992). Ectopic recombination by intermediate sized repeats (50-556kb in Arabidopsis; Davila et al., 2011) were shown to be associated with the generation of novel DNA structure and change in the stoichiometry of subgenomic molecules which is called as ‘substoichiometric shift’ (SSS) (Shedge et al., 2007; Zaegel et al., 2006). Janska et al. (1998) reported that the proliferation of mitochondrial subgenomic molecule which contains the gene responsible for cytoplasmic male sterility (CMS) is related to origination of CMS in common bean. The CMS-associated subgenomic molecule was shown to be present even in normal common bean lines, but maintained at very low level (Arrieta-Montiel et al., 2001). On the other hand, variegated phenotype in Arabidopsis is also related to SSS (Martinez-Zapater *et al.,* 1992). The gene responsible for this phenotype was isolated and named as *MSH1*. This gene encodes a protein homologous to MutS which is a mismatch repair protein in *E.coli* (Abdelnoor *et al,*, 2003). Recent studies using *MSH1* mutants of Arabidopsis suggested the possible mechanism for novel DNA rearrangements and SSS in plant mtDNA (Shedge et

al., 2007; Arrieta-Montiel et al, 2009; Davila et al., 2011). In normal plants, *MSH1* suppresses the recombination through intermediate size repeats. However, in mutants of *MSH1,* recombinations were detectable in all of the near-perfectly repeated sequences longer than 50bp by high-depth sequencing analysis (Davila et al., 2011). If the original function of *MSH1* is regarded, these recombinations are likely to include double strand break of DNA, heteroduplex formation on repeat sequences, gene conversion, and mismatch repair (Shedge et al., 2007). The strand-invasion events at this process may initiate the recombination dependent replication (RDR) which promotes the amplification of recombination products, thus result in substoichiometric shift (Shcherbakov et al., 2006; Shedge et al., 2007; Stohr and Kreuzer, 2002). In this process, asymmetric recombination occurs in which only one kind of recombination product is detected (Davila et al., 2011). This pattern is also consistent with the characteristics of RDR (Shcherbakov et al., 2006; Stohr and Kreuzer, 2002). Davila et al. (2011) suggested that intensive rearrangements on plant mitochondrial genome are mediated by ectopic recombinations and nonhomologous end joining after double strand break of mtDNA.

2. Cytoplasmic male sterility

Cytoplasmic male sterility (CMS) is defined as maternally inherited inability to produce functional pollen (Hanson and Bentolila, 2004). CMS has

been suggested to occur by wide crosses, the interspecific exchange of nuclear and cytoplasmic genome or cell fusions (Carlsson et al., 2007; Dubreucq et al., 1999; Schnable, 1998). The CMS-associated genes were cloned in diverse plant species and all of them were novel chimearic genes which were composed by the fusions of sequences from other part of mitochondrial genome or sequences with unknown origin (Hanson and Bentolila, 2004) These chimearic genes includes *pvs-orf239* in common bean (Johns et al., 1992), *orf522* in sunflower (Kohler et al., 1991), *orf224* in canola (Singh et al., 1991), *orf77* in CMS-S maize (Levings and Sederoff, 1983), *T-urf13* in CMS-T maize (Dewey et al., 1986.), *S-pcf* in petunia (Young and Hanson, 1987), *orf79* in Boro-CMS rice (Kadowaki *et al.*, 1986), *orf107* in carrot (Tang et al., 1996), *orf256* in wheat (Rathburn and Hedgcot, 1993), *WA352* in CMS-WA rice (Luo et al., 2013), and *orf507* in chili pepper (Kim, 2007). Although CMS-associated genes cloned in diverse plant species do not show sequence similarity between them, they share several characteristics. Many of CMS-associated genes are associated with known genes which encode for ATP synthase subunits while *Brassica orf222* and petunia *pcf* are near to genes which encode subunits of Complex I (Hanson and Bentolila, 2004; Young and Hanson, 1987; L'Homme et al., 1997; Kubo et al., 2011). Close localization of CMS-associated genes with known mitochondrial genes enables the co-transcription of CMS genes with regular genes (Hanson and Bentolila, 2004). In addition, CMS-associated genes contain transmembrane domain in

common except for common bean ORF239 which is detected outside of mitochondria (Abad et al., 1995)

Various hypotheses have been proposed for how the mitochondrial chimeric genes induce CMS. These includes disruption of mitochondrial membranes (Sabar et al., 2000), dysfunction of ATP synthase (Bergman et al., 2000), programmed cell death of tapetal cells (Balk and Leaver, 2001) and alteration of the expression patterns of floral development genes (Carlsson et al., 2008). Recently, interaction between products of CMS-associated genes and mitochondrial proteins were reported in pepper and rice (Li et al., 2012; Luo et al., 2013). ORF507 in pepper interacted with nuclear-encoded ATP synthase 6kDa subunit and the ATP:ADP ratio was decrease. This result implied that interaction between ORF507 and ATP synthase 6kDa subunit may hamper the normal function of ATP synthase complex (Li et al., 2012). On the other hand, WA352 in CMS-WA rice interacted with a nuclear-encoded COX11. WA352 inhibited the function of COX11 in peroxide metabolism and this resulted in premature programmed cell death of tapetal cells (Luo et al., 2013)

3. Restoration of fertility

Restorer-of-fertility (*Rf*) is a nucleus-encoded gene which suppresses the induction of cytoplasmic male sterility (CMS) caused by CMS-associated genes located on mitochondrial genome. *Rf* genes have been identified in maize (Cui et

al, 1996), petunia (Bentolila et al., 2002), rice (Komori et al., 2004; Hu et al., 2012; Fujii and Toriyama, 2009; Itabashi et al., 2011), radish (Brown et al., 2003; Desloire et al., 2003; Koizuka et al., 2003), and sugar beet (Matsuhira et al., 2012). Most of the cloned *Rf* genes were the members of pentatricopeptide repeat (PPR) gene family. In addition, association of PPR genes and *Rf* loci was proved genetically in several other species including sorghum, *Mimulus*, and Maize (CMS-S) (Klein et al., 2005; Barr and Fishman, 2010; Xu et al., 2010). However, non-PPR type *Rf* genes were also cloned which encoded an aldehyde dehydrogenase (Rf2a), a glycine-rich protein (Rf17), a putative retrograde signaling control-related protein (Rf2), and a putative mitochondrial protein quality control-related protein (Rf1) in CMS-T maize, CW-CMS rice, LD-CMS rice, and sugar beet, respectively (Cui et al., 1996, Fujii and Toriyama, 2009, Itabashi et al., 2011, Matsuhira et al., 2012). None of special characteristics could be shared between non-PPR type *Rf* gene.

PPR proteins have a repeated motif composed of a degenerative array of 35 amino acids that may bind RNA through its superhelix structure (Small and Peeters, 2000). Supporting the idea that PPR protein binds RNA, PPR proteins have been reported to edit chloroplast genes in Arabidopsis (Kotera et al., 2005). In addition, *Rf1a* and *Rf1b* in the rice with Boro II cytoplasm encode proteins cleaving or degrading mRNA of the CMS-associated gene (Wang et al., 2006). Recent study showed that each 35 amino acid array of PPR genes determined the

specificity of the protein to one nucleotide of target RNA (Barkan et al., 2012). Combination of the first and sixth amino acids in PPR protein was crucial for the determination of specificity implying direct interaction of these amino acids with RNA (Barkan et al., 2012).

PPR-type RF proteins were shown to be a member in a large protein complex in several cases. For example, PPR592 of petunia form a 400kDa complex with other proteins and the transcript of CMS-associated gene (Gillman et al., 2007). Hu et al. (2012) also reported that RF5 in Hong-Lian CMS rice makes 400-500kDa complex with other proteins and CMS-associated gene transcript. However, instead of Rf5, a glycine rich protein named as GRP162 interacted with CMS-associated gene transcript in this case (Hu et al., 2012).

PPR genes constitute large gene family only in land plants (e.g. 450 in Arabidopsis, 477 in rice) although only a few are detected in other species including yeasts and algae (Fujii and Small, 2011). Cloned *Rf* genes shared several characteristics among PPR genes. First of all, they form a cluster with closely located PPR genes while other PPR genes are dispersed on entire genome sequences. For example, in rice, nine PPR genes including two restorer genes are clustered in ~150 kb-long region on chromosome 10 (Wang et al., 2006). In addition, *Rf* and clustered PPR genes show high sequence similarity. Fujii et al. (2011) classified these genes as *Rf*-like genes (*RFL*) based on phylogenetic analysis using PPR genes from all sequence database available. *RFL*s from

diverse plant species formed a clade separated from clades containing other PPR genes implying that *RFL*s have originated from the same ancestral gene which had existed before the speciation of land plants. Finally, *RFL*s show much higher rate of nonsynonymous to synonymous substitutions than other PPR genes (Fujii et al., 2011). The rate of diversifying selection was shown to be the highest on the first, third and sixth amino acid of RFL protein, which may be involved in the interaction with RNA ligand (Fujii et al., Barkan et al., 2012). Altogether, these characteristics of *Rf* genes supports the hypothesis that *Rf* genes has evolved by birth-and-death process which is usually found in disease resistance genes to cope with the appearance of new CMS genes (Touzet and Budar, 2004).

4. Cytoplasmic-genic male sterility (CGMS) in pepper

Cytoplasmic-genic male sterility (CGMS) system has been widely used for efficient hybrid seed production in chili pepper. CMS in pepper was first discovered in an Indian *C. annuum* accession (USDA accession PI 164835), whose cytoplasm has been used as the sole source for CMS (Shifriss, 1997). In the case of restoration of fertility, although many studies reported pepper *Rf* gene as a single dominant gene (Zhang et al. 2000; Gulyas et al. 2006; Kim et al. 2006; Jo et al. 2009; Min et al. 2008; Min et al. 2009), different inheritance patterns such as inheritance by two independent dominant genes (Peterson. 1958), interaction between two complementary genes (Novak et al. 1971) and QTLs have been

suggested. In the QTL analysis, one major QTL was located on upper region of chromosome 6 where the single dominant *Rf* gene was localized in other study (Jo et al., 2010) and four additional minor QTLs were also determined (Wang et al. 2004).

Candidate genes for CMS-associated gene were isolated in pepper. A chimeric mitochondrial gene, *orf456*, was identified in the mitochondrial genome of CMS peppers. The *orf456* gene construct fused with mitochondrial target sequence induced male sterility in Arabidopsis by transformation experiment implying that this gene was strong candidate for CMS-associated gene in pepper (Kim et al., 2007). In later studies, the *orf456* gene was shown to exist as longer *orf* (*orf507*) indicating there was a sequencing error on the 3' end region of the *orf* in the previous study (Gulyas et al., 2010). Another candidate gene named as *ψatp6-2* gene was generated by novel rearrangement on 3' region of normal *atp6-2*. Transcription pattern of *ψatp6-2* were different between male sterile and restorer lines showing possible association of this gene with CMS (Kim et al., 2006). Sequencing of the CMS mitochondrial region containing these genes and comparative analysis with the counterpart region in normal cytoplasm revealed that numerous DNA rearrangements had occurred between sequences from each genome and a pair of repeat sequences near 3' end of the genes may be involved in rearrangement process (Jo, 2007).

Several molecular markers linked to the *Rf* gene have been developed

although the *Rf* gene itself was not cloned yet. The molecular markers linked to *Rf* include OPP13-CAPS (1.1cM from *Rf*), AFRF8-CAPS (1.8cM from *Rf*), PR-CAPS (1.8cM from *Rf*) and an STS marker named CRF-SCAR (5.3cM from *Rf*) (Kim, 2005; Kim et al., 2006; Lee et al., 2008; Gulyas et al., 2006). These markers have limited applications in diverse pepper lines, however, because of failure in PCR amplification, the existence of third haplotypes and lack of agreement between marker genotype and phenotype (Min et al., 2008; Jo et al., 2010).

Recently, elucidation of the mechanism for CMS and restoration of fertility has been attempted in several studies. The CMS-associated protein, Orf507 was shown to interact with ATP synthase 6kDa subunit protein (Li et al., 2012). Because decrease in the ATP synthase activity was detected in a CMS pepper line, binding of Orf507 to ATP synthase 6kDa subunit protein to hamper the normal function of ATP synthase was suggested as a possible mechanism for CMS in pepper (Li et al., 2012). Meanwhile, comparative analysis on the transcriptome between a CMS line and its near-isogenic restorer line showed that many of possible fertility-related genes including ATP synthesis-related genes, MADS-box genes, and genes encoding active oxygen scavenger were up-regulated or down-regulated in anthers of each line. Especially, nine of PPR genes were highly up-regulated in restorer lines indicating that those genes can be good candidates for *Rf* (Liu et al., 2013).

Although CGMS system in pepper is stable in certain range of chili pepper

cultivars which includes hot dry type pepper in Korea (Lee 2001; Shifriss 1997), instability of sterility or restoration has been detected in other type of hot peppers or sweet peppers (Shifriss 1997). Instability in CGMS can be divided mainly by two phenotypes; partial restoration in which the anther contains intermediate amount of pollen between that of fully sterile and fully restored pepper, and unstable sterility in which sterility is maintained in normal condition, but collapsed in certain environmental conditions such as low temperature (Lee et al., 2008). The genetic analysis was performed for partial restoration and the allele for this phenotype was mapped on the same or adjacent position of *Rf* gene (Lee et al., 2008). However, the genetics of unstable sterility is controversial and molecular markers have not been developed although this phenotype is detected in many of pepper lines including sweet peppers (Min et al., 2009; Shifriss, 1997).

REFERENCES

Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang C, Du F, Xu D, Gibson M, Cifrese J, Clifton SW, Newton KJ (2007) Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 177: 1173-92

Abad AR, Mehrtens BJ, Mackenzie SA (1995) Specific expression in reproductive tissues and fate of a mitochondrial sterility-associated protein in cytoplasmic male-sterile bean. Plant Cell 7: 271-85

Abdelnoor RV, Yule R, Elo A, Christensen AC, Meyer-Gauen G, Mackenzie SA (2003) Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. Proc Natl Acad Sci U S A 100: 5968-73

Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD (2011) Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. Plant Cell 23: 2499-513.

Andre C, Levy A, Walbot V (1992) Small repeated sequences and the structure of plant mitochondrial genomes. Trends Genet 8: 128-3.

Arrieta-Montiel M, Lyznik A, Woloszynska M, Janska H, Tohme J, Mackenzie S (2001) Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean. Genetics 158: 851-64

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA (2009) Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics 183: 1261-8

Backert S, Dorfel P, Lurz R, Borner T (1996) Rolling-circle replication of mitochondrial DNA in the higher plant *Chenopodium album* (L.). Mol Cell

Biol 16: 6285-94

Backert S, Lurz R, Borner T (1996) Electron microscopic investigation of mitochondrial DNA from *Chenopodium album* (L.). Curr Genet 29: 427-36

Balk J, Leaver CJ (2001) The PET1-CMS mitochondrial mutation in sunflower is associated with premature programmed cell death and cytochrome c release. Plant Cell 13: 1803-18

Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. PLoS Genet 8: e1002910

Barr CM, Fishman L (2010) Cytoplasmic male sterility in *Mimulus* hybrids has pleiotropic effects on corolla and pistil traits. Heredity (Edinb) 106: 886-93

Bentolila S, Alfonso AA, Hanson MR (2002) A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci USA 99:10887-92

Bentolila S, Stefanov S (2012) A reevaluation of rice mitochondrial evolution based on the complete sequence of male-fertile and male-sterile mitochondrial genomes. Plant Physiol 158: 996-1017

Bergman P, Edqvist J, Farbos I, Glimelius K (2000) Male-sterile tobacco displays abnormal mitochondrial atp1 transcript accumulation and reduced floral ATP/ADP ratio. Plant Mol Biol 42: 531-44

Borck KS, Walbot V (1982) Comparison of the restriction endonuclease digestion patterns of mitochondrial DNA from normal and male sterile cytoplasm of *Zea mays* L. Genet 1982 102: 109-28

Brown GG, Formanova N, Jin H, Wargachuk R, Dendy C, Patil P, Laforest M, Zhang J, Cheung WY, Landry BS (2003) The radish Rfo restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. Plant J 35: 262-72

Carlsson J, Leino M, Glimelius K (2007) Mitochondrial genotypes with variable parts of Arabidopsis thaliana DNA affect development in Brassica napus lines. Theor Appl Genet 115: 627-41

Carlsson J, Leino M, Sohlberg J, Sundstrom JF, Glimelius K (2008) Mitochondrial regulation of flower development. Mitochondrion 8: 74-86

Chen J, Guan R, Chang S, Du T, Zhang H, Xing H (2011) Substoichiometrically different mitotypes coexist in mitochondrial genomes of *Brassica napus* L. PLoS One 6: e17662

Cui X, Wise RP, Schnable PS (1996) The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. Science 272: 1334-6

Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ, Weigel D, Mackenzie SA (2011) Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol 9: 64

Desloire S, Gherbi H, Laloui W, Marhadour S, Clouet V, Cattolico L, Falentin C, Giancola S, Renard M, Budar F, Small I, Caboche M, Delourme R, Bendahmane A (2003) Identification of the fertility restoration locus, Rfo, in radish, as a member of the pentatricopeptide-repeat protein family. EMBO Rep 4: 588-94

Dewey RE, Levings CS, 3rd, Timothy DH (1986) Novel recombinations in the maize mitochondrial genome produce a unique transcriptional unit in the Texas male-sterile cytoplasm. Cell 44: 439-49

Dubreucq A, Berthe B, Asset JF, Boulidard L, Budar F, Vasseur J, Rambaud C (1999) Analyses of mitochondrial DNA structure and expression in three cytoplasmic male-sterile chicories originating from somatic hybridization between fertile chicory and CMS sunflower protoplasts Theor Appl Genet 99: 1094-105

Fauron C, Casper M, Gao Y, Moore B (1995) The maize mitochondrial genome: dynamic, yet functional. Trends Genet 11: 228-35

Folkerts O, Hanson MR (1989) Three copies of a single recombination repeat occur on the 443 kb master circle of the Petunia hybrida 3704 mitochondrial genome. Nucleic Acids Res 17: 7345-57

Fujii S, Bond CS, Small ID (2011) Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. Proc Natl Acad Sci U S A 108: 1723-8

Fujii S, Kazama T, Yamada M, Toriyama K (2010) Discovery of global genomic re-organization based on comparison of two newly sequenced rice mitochondrial genomes with cytoplasmic male sterility-related genes. BMC Genomics 11: 209

Fujii S, Toriyama K (2009) Suppressed expression of Retrograde-Regulated Male Sterility restores pollen fertility in cytoplasmic male sterile rice plants. Proc Natl Acad Sci U S A 106: 9513-8

Gillman JD, Bentolila S, Hanson MR (2007) The petunia restorer of fertility protein is part of a large mitochondrial complex that interacts with transcripts of the CMS-associated locus. Plant J 49: 217-27

Gray MW (1999) Evolution of organellar genomes. Curr Opin Genet Dev 9: 678-87

Gulyas G, Pakozdi K, Lee JS, Hirata Y (2006) Analysis of fertility restoration by using cytoplasmic male-sterile red pepper (*Capsicum annuum* L.) lines. Breed Sci 56:331-334

Gulyas G, Shin Y, Kim H, Lee JS, Hirata Y (2010) Altered transcript reveals an *orf507* sterility-related gene in chili pepper (*Capsicum annuum* L.). Plant Mol Bio Rep 28: 605-12

Handa H (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res 31: 5907-16

Hanson MR, Bentolila S (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell 16 Suppl: S154-69

Hu J, Wang K, Huang W, Liu G, Gao Y, Wang J, Huang Q, Ji Y, Qin X, Wan L, Zhu R, Li S, Yang D, Zhu Y (2012) The rice pentatricopeptide repeat protein RF5 restores fertility in Hong-Lian cytoplasmic male-sterile lines via a complex with the glycine-rich protein GRP162. Plant Cell 24: 109-22

Itabashi E, Iwata N, Fujii S, Kazama T, Toriyama K (2011) The fertility restorer gene, Rf2, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein. Plant J 65: 359-67

Janska H, Sarria R, Woloszynska M, Arrieta-Montiel M, Mackenzie SA (1998) Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. Plant Cell 10: 1163-80

Johns C, Lu M, Lyznik A, Mackenzie S (1992) A mitochondrial DNA sequence is associated with abnormal pollen development in cytoplasmic male sterile bean plants. Plant Cell 4: 435-49

Jo YD, Kim YM, Park MN, Yoo JH, Park M, Kim BD, Kang BC (2010) Development and evaluation of broadly applicable markers for Restorer-of-fertility in pepper. Mol Breed 25:187-201

Kim DH, Kang JG, Kim BD (2007) Isolation and characterization of the cytoplasmic male sterility-associated *orf456* gene of chili pepper (*Capsicum annuum* L.). Plant Mol Biol 63:519-532

Kim DH, Kim BD (2006) The organization of mitochondrial *atp6* gene region in male fertile and CMS lines of pepper (*Capsicum annuum* L.). Curr Genet 49: 59-67

Kim DS (2005) Development of RAPD and AFLP markers linked to fertility restorer (*Rf*) gene in chili pepper (*Capsicum annuum* L.). Thesis, Seoul National University

Kim DS, Kim DH, Yoo JH, Kim BD (2006) Cleaved amplified polymorphic sequence and amplified fragment length polymorphism markers linked to the fertility restorer gene in chili pepper (*Capsicum annuum* L.). Mol Cells 21:135-140

Klein RR, Klein PE, Mullet JE, Minx P, Rooney WL, Schertz KF (2005) Fertility restorer locus Rf1 [corrected] of sorghum (*Sorghum bicolor* L.) encodes a pentatricopeptide repeat protein not present in the colinear region of rice chromosome 12. Theor Appl Genet 111: 994-1012

Kohler RH, Horn R, Lossl A, Zetsche K (1991) Cytoplasmic male sterility in sunflower is correlated with the co-transcription of a new open reading frame with the atpA gene. Mol Gen Genet 227: 369-76

Koizuka N, Imai R, Fujimoto H, Hayakawa T, Kimura Y, Kohno-Murase J, Sakai T, Kawasaki S, Imamura J (2003) Genetic characterization of a pentatricopeptide repeat protein gene, orf687, that restores fertility in the cytoplasmic male-sterile Kosena radish. Plant J 34: 407-15

Komori T, Ohta S, Murai N, Takakura Y, Kuraya Y, Suzuki S, Hiei Y, Imaseki H, Nitta N (2004) Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa* L.). Plant J 37: 315-25

Kotera E, Tasaka M, Shikanai T (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433: 326-30

Kubo T, Kitazaki K, Matsunaga M, Kagami H, Mikami T (2011) Male sterility-inducing mitochondrial genomes: how do they differ? Crit Rev. Plant Sci 30: 378-400

Levings CS, Sederoff RR (1983) Nucleotide sequence of the S-2 mitochondrial DNA from the S cytoplasm of maize. Proc Natl Acad Sci U S A 80: 4055-9

Li J, Pandeya D, Jo YD, Liu WY, Kang BC (2012) Reduced activity of ATP synthase in mitochondria causes cytoplasmic male sterility in chili pepper. Planta 237: 1097-109

Liu C, Ma N, Wang PY, Fu N, Shen HL (2013) Transcriptome Sequencing and De Novo Analysis of a Cytoplasmic Male Sterile Line and Its Near-Isogenic Restorer Line in Chili Pepper (*Capsicum annuum* L.). PloS one 8: e65209

Lee DH (2001) Studies on unstable fertility of CGMS (cytoplasmic-genic male sterility) in *Capsicum annuum* L. Dissertation, Seoul National University

Lee J, Yoon JB, Park HG (2008) Linkage analysis between the partial restoration (*pr*) and the *restorer-of-fertility* (*Rf*) loci in pepper cytoplasmic male sterility. Theor Appl Genet 117:383-9

L'Homme, Y., and Brown, G.G. (1993). Organizational differences between cytoplasmic male sterile and male fertile Brassica mitochon mitochondrial genomes are confined to a single transposed locus. Nucleic Acids Res. 21, 1903–9

Liu H, Cui P, Zhan K, Lin Q, Zhuo G, Guo X, Ding F, Yang W, Liu D, Hu S, Yu J, Zhang A (2011) Comparative analysis of mitochondrial genomes between a wheat K-type cytoplasmic male sterility (CMS) line and its maintainer line. BMC Genomics 12: 163

Luo D, Xu H, Liu Z, Guo J, Li H, Chen L, Fang C, Zhang Q, Bai M, Yao N, Wu H, Wu H, Ji C, Zheng H, Chen Y, Ye S, Li X, Zhao X, Li R, Liu YG (2013) A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. Nat Genet 45: 573-7

Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089-103

Mackenzie S, McIntosh L (1999) Higher plant mitochondria. Plant Cell 11: 571-86

Martinez-Zapater JM, Gil P, Capel J, Somerville CR (1992) Mutations at the

Arabidopsis CHM locus promote rearrangements of the mitochondrial genome. Plant Cell 4: 889-99

Matsuhira H, Kagami H, Kurata M, Kitazaki K, Matsunaga M, Hamaguchi Y, Hagihara E, Ueda M, Harada M, Muramatsu A, Yui-Kurino R, Taguchi K, Tamagake H, Mikami T, Kubo T (2012) Unusual and typical features of a novel *restorer-of-fertility* gene of sugar beet (*Beta vulgaris* L.). Genetics 192: 1347-58

Min WK, Lim H, Lee YP, Sung SK, Kim BD, Kim S (2008) Identification of a third haplotype of the sequence linked to the *Restorer-of-fertility* (*Rf*) gene and its implications for male-sterility phenotypes in peppers *(Capsicum annuum* L.). Mol Cells 25:20-9

Min WK, Kim S, Sung SK, Kim BD, Lee S (2009) Allelic discrimination of the *Restorer-of-fertility* gene and its inheritance in peppers (*Capsicum annuum* L.). Theor Appl Genet 119:1289-99

Mower JP, Case AL, Floro ER, Willis JH (2012) Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (Mimulus guttatus) lineage with cryptic CMS. Genome Biol Evol 4: 670-86

Novak F, Betlach J, Dubovsky J (1971) Cytoplasmic male sterility in sweet pepper (*Casicum annuum* L.). I. Phenotype and inheritance of male sterile character. Z Pflanzenzucht 65:129-40

Palmer JD (1990) Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet 6: 115-20

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97: 6960-6

Palmer JD, Herbon LA (1986) Tricircular mitochondrial genomes of Brassica and Raphanus: reversal of repeat configurations by inversion. Nucleic Acids Res 14: 9755-64

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28: 87-97

Palmer JD, Shields CR (1984) Tripartite structure of the *Brassica* campestris mitochondrial genome. Nat 307: 437-40

Park JY, Lee YP, Lee J, Choi BS, Kim S, Yang TJ (2013) Complete mitochondrial genome sequence and identification of a candidate gene responsible for cytoplasmic male sterility in radish (*Raphanus sativus* L.) containing DCGMS cytoplasm. Theor Appl Genet 126: 1763-74

Peterson PA (1958) Cytoplasmically inherited male sterility in *Capsicum*. Amer Nat 92:111-9

Rathburn H, Song J, Hedgcoth C (1993) Cytoplasmic male sterility and fertility restoration in wheat are not associated with rearrangements of mitochondrial DNA in the gene regions for cob, coxII, or coxI. Plant Mol Biol 21: 195-201.

Sabar M, De Paepe R, de Kouchkovsky Y (2000) Complex I impairment, respiratory compensations, and photosynthetic decrease in nuclear and mitochondrial male sterile mutants of Nicotiana sylvestris. Plant Physiol 124: 1239-50.

Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, Mikami T (2004) The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. Mol Genet Genomics 272: 247-56

Schnable PS, Wise RP (1998) The molecular basis of cytoplasmic male sterility and fertility restoration. Trends Plant Sci 3:175-80

Shcherbakov VP, Kudryashova EA, Shcherbakova TS, Sizova ST, Plugina LA (2006) Double-strand break repair in bacteriophage T4: recombination effects of 3'-5' exonuclease mutations. Genetics 174: 1729-36

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA (2007) Plant

mitochondrial recombination surveillance requires unusual RecA and MutS homologs. Plant Cell 19: 1251-64.

Shedge V, Davila J, Arrieta-Montiel MP, Mohammed S, Mackenzie SA Extensive rearrangement of the Arabidopsis mitochondrial genome elicits cellular conditions for thermotolerance. Plant Physiol 152: 1960-70

Shifriss C (1997) Male sterility in pepper (*Capsicum annuum* L.). Euphytica 93:83-8

Siculella L, Damiano F, Cortese MR, Dassisti E, Rainaldi G, Gallerani R, Benedetto C (2001) Gene content and organization of the oat mitochondrial genome. Theor Appl Genet 103: 359-65

Singh M, Brown GG (1991) Suppression of cytoplasmic male sterility by nuclear genes alters expression of a novel mitochondrial gene region. Plant Cell 3: 1349-62

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10:e1001241

Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR (2010) Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm Silene latifolia. BMC Evol Biol 10: 274

Small ID, Peeters N (2000) The PPR motif - a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25: 46-7

Small ID, suffolk R, Leaver CJ (1989) Evolution of plant mitochondrial genomes via substoichiometric intermediates. Cell 58: 69-76

Stohr BA, Kreuzer KN (2002) Coordination of DNA ends during double-strand-break repair in bacteriophage T4. Genetics 162: 1019-30

Tanaka Y, Tsuda M, Yasumoto K, Yamagishi H, Terachi T (2012) A complete

mitochondrial genome sequence of Ogura-type male-sterile cytoplasm and its comparative analysis with that of normal cytoplasm in radish (*Raphanus sativus* L.). BMC Genomics 13: 352

Tang HV, Pring DR, Shaw LC, Salazar RA, Muza FR, Yan B, Schertz KF (1996) Transcript processing internal to a mitochondrial open reading frame is correlated with fertility restoration in male-sterile sorghum. Plant J 10: 123-33

Touzet P, Budar F (2004) Unveiling the molecular arms race between two conflicting genomes in cytoplasmic male sterility? Trends Plant Sci 9: 568-70

Wang LH, Zhang BX, Lefebvre V, Huang SW, Daubeze AM, Palloix A (2004) QTL analysis of fertility restoration in cytoplasmic male sterile pepper. Theor Appl Genet 109:1058-1063

Wang Z, Zou Y, Li X, Zhang Q, Chen L, Wu H, Su D, Chen Y, Guo J, Luo D, Long Y, Zhong Y, Liu YG (2006) Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. Plant Cell 18: 676-87

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84: 9054-8

Young EG, Hanson MR (1987) A fused mitochondrial gene associated with cytoplasmic male sterility is developmentally regulated. Cell 50: 41-9

Zaegel V, Guermann B, Le Ret M, Andres C, Meyer D, Erhardt M, Canaday J, Gualberto JM, Imbault P (2006) The plant-specific ssDNA binding protein OSB1 is involved in the stoichiometric transmission of mitochondrial DNA in Arabidopsis. Plant Cell 18: 3548-63

Zhang BX, Huang SW, Yang GM, Guo JZ (2000) Two RAPD markers linked to a major fertility restorer gene in pepper. Euphytica 113:155-61

CHAPTER I

Comparative Analysis of Mitochondrial Genomes between CMS and Fertile Pepper (*Capsicum annuum* L.) Lines

ABSTRACT

Cytoplasmic male sterility (CMS) is an inability to produce functional pollen which is caused by mutation on mitochondrial genome. Comparative analyses on CMS and normal mitochondrial genomes in several species have shown structural differences between genomes which include extensive rearrangements caused by illegitimate recombinations. However, the overall mitochondrial genome structure and the unique DNA rearrangement patterns that specify CMS cytoplasm have not been characterized completely in chili pepper. In this study, we obtained the mitochondrial genome sequences of a pepper CMS line,

FS4401 and a fertile line, Jeju, by next-generation sequencing against mtDNAs. The mitochondrial genome of FS4401 and Jeju were 507,450 and 493,911bp in length, respectively. Both mitochondrial genomes shared the same contents of protein coding genes except for one more copy of *atp6* gene in FS4401. Comparative analysis between pepper and tobacco mitochondrial genomes revealed that only 55% of pepper mtDNA could be aligned with the counterpart in tobacco indicating extensive DNA rearrangements between them. Although gene sequences were highly conserved, structural alterations detected in several gene clusters implied the possible modification of co-transcription units during evolution. Comparative alignment of FS4401 and Jeju mtDNAs revealed eighteen syntenic between two mitochondrial genomes (>2kb, >95% similarity) were obtained which were rearranged by both nonhomologous end-joining (NHEJ) and asymmetric recombination through repeated sequences. On the other hand, sequences located between these syntenic blocks, which were specific to each line, constituted 30,380 and 17,847bp of mtDNA in FS4401 and Jeju, respectively. The CMS candidate genes, *orf507* and *atp6-2*, were located on the edges of largest sequence segments which were specific to FS4401. Most of the DNA region around these genes could not be aligned with any known sequences. Although severe rearrangements in this region hampered the elucidation of detailed mechanism, presence of repeated and overlap sequences on DNA segments implied that extensive rearrangement by NHEJ followed by substoichiometric

shift because recombination through repeated sequence might involved in generation and integration of this region on the master DNA molecule. Further analysis using mtDNA pairs of CMS-normal cytoplasms in other plant species showed common features of DNA region around CMS-associated genes.

INTRODUCTION

Mitochondrial genomes of higher plants are clearly different from its animal counterparts or plastid genomes in evolutionary dynamics of genome structure (Andre et al, 1992; Palmer et al, 2000). Although the rate of synonymous substitutions in plant mtDNAs is 50-100 times and three times lower than in vertebrate mitochondria and plastid mitochondria, respectively, structural variations including changes in gene orders, rearrangements, genome expansion and shrinkage, and incorporation of foreign DNAs are highly extensive in plant mitochondria than others (Palmer and Herbon, 1988; Palmer, 1990; Wolfe et al, 1987). Existence of reservoir of subgenomic mtDNA molecules under copy number suppression by nuclear control and repeated sequences dispersed around the genome have been pointed as the features that may explain the complexity of plant mtDNA structure (Arrieta-Montiel et al, 2001; Small et al, 1989).

More detailed mechanisms for evolution of plant mtDNA structures have come from recent studies using mutants of mitochondrial-targeted mismatch repair-related genes including MSH1, RACA3, and OSB1 (Arrieta-Montiel et al, 2009; Davila et al., 2011; Shedge et al., 2007; Zaegel et al., 2006). Mutant plant showed significantly increased frequency of recombination events via intermediate-sized repeat sequences which are very rarely occur in normal plants. These recombinations resulted in the formation of asymmetric chimera molecules

and change in the ratio of subgenomic molecules (substoichiometric shifts; SSS) which were inherited to the next generation. Involvement of mismatch repair-related genes and unique features of rearranged products implied that double strand break repair process which consist heteroduplex formation on repeat sequences, gene conversion, and mismatch repair may be responsible for rapid evolution in mtDNA structure and low rate of sequence substitution (Arrieta-Montiel et al, 2009; Davila et al., 2011).

Structural variation in mtDNA is related with several mutant phenotypes such as cytoplasmic male sterility (CMS) and variegated phenotypes (Hanson and Bentolila, 2004; Abdelnoor et al, 2003; Zaegel et al, 2006). CMS have been studied in many crop species because of its agronomical importance in hybrid seed production. Most of identified CMS-associated genes were novel chimearic *orfs* generated from the fusions of several sequence segments by rearrangement of mitochondrial genome. Although CMS-associated genes in different crop species did not show significant similarity in sequence, most of them shared several features in common such as possession of transmembrane domain and co-transcription with normal mitochondrial genes which often encode ATP-synthase or cytochrome C oxidase subunits (Ashutosh et al., 2008; Hanson and Bentolila, 2004; Kim et al., 2007; Wang et al., 2006). The detailed mechanism how these genes originated is remained mostly unknown yet. In common bean and radish, extremely small number of CMS-associated gene copies were detected even in

mitochondria of normal plants suggesting that CMS may occur due to rapid increase in copy number CMS-associated subgenomic molecules which had already existed in normal cytoplasm (Arrieta-Montiel et al., 2001; Janska et al., 1998; Kim et al., 2007).

Complete sequencing and comparative analysis of normal and CMS cytoplasm have been performed in several crop species including sugar beet, maize, wheat, rice, rapeseed, and radish to analyze the structural variation of mtDNA and identify candidate *orf*s associated with CMS (Allen et al., 2007; Chen et al., 2011; Liu et al., 2011; Park et al., 2013; Tanaka et al., 2012; Satoh et al., 2004). These studies showed that mitochondrial genome structures were extensively rearranged in CMS cytoplasm although gene contents were mostly conserved. For example, in sugar beet, normal and CMS mitochondrial genomes were composed by different arrays of fourteen sequence blocks which are syntenic between two genomes (Satoh et al., 2004). Recently, Tanaka et al. (2012) showed that a radish CMS mitochondrial genome, Ogura cytoplasm, contained large CMS-specific region in addition to syntenic block sequences. This region contained CMS-associated *orf* and was postulated to be inserted to normal mitochondrial genome by recombination using inverted repeat sequences located on borders. However, the origin of CMS-associated gene sequence and other CMS-specific sequence in this region was still remained to be unknown.

CMS has been widely used in hybrid seed production in chili peppers. Only a single source of cytoplasm has been reported to be responsible for CMS (Peterson, 1958). Kim et al (2006, 2007) found two candidate genes which were *orf456* and *ψatp6-2*. The *orf456* gene construct fused with mitochondrial target sequence induced male sterility in transgenic Arabidopsis implying that this gene was strong candidate for CMS-associated gene in pepper (Kim et al., 2007). In later studies, the *orf456* gene was shown to exist as a longer *orf* (*orf507*) indicating there was a sequencing error on the 3' end region of the *orf* in previous study (Gulyas et al., 2010). The *ψatp6-2* gene was generated by novel rearrangement on 3' region of normal *atp6-2*. Transcription patterns of *ψatp6-2* were different between male sterile and restorer lines showing possible association of this gene with CMS (Kim et al., 2006). Sequencing of the CMS mitochondrial region containing these genes and comparative analysis with the counterpart region in normal cytoplasm revealed that numerous DNA rearrangements had occurred in these genomes and a pair of repeat sequence near 3' end of the genes might be involved in rearrangement process (Jo, 2007). However, the characterization of the unique rearrangement pattern between CMS and normal pepper cytoplasms has not been performed in mitochondrial genome scale.

In this study, the complete mitochondrial genome sequence of pepper was firstly reported in fertile and CMS pepper lines. Comparative analysis between two mitochondrial genome provided insights in evolution of CMS cytoplasm structure in pepper.

MATERIALS AND METHODS

Plant materials

A pepper CMS line, ‘FS4401’, and a restorer line, ‘Jeju’, which is known to contain normal cytoplasm were provided by Monsanto Korea. For each pepper lines, approximately 3,000 seedling were grown in a dark condition for twenty days and harvested to isolate mitochondria.

Mitochondrial DNA extraction

The method described by Millar et al. (2001) and modified by Kim (2004) was used for mitochondrial DNA extraction. Seedlings were homogenized using mortar with isolation buffer containing 0.3 M mannitol, 50 mM Tris-HCl, 3 mM EDTA, 1 mM 2-mercaptoethanol, 0.1% BSA, 1% PVP, and protease inhibitor cocktail (Roche Applied Science, Indianapolis, USA) and adjusted to pH 7.5 with KOH. Homogenized tissue was filtered with one layer of miracloth and four layers of cheesecloth. Two times of centrifugation at 2,000g for 10 min were followed to remove cell debris and larger organelles in cells. The supernatant was centrifuged at 15,000g for 10 min to obtain crude mitochondrial pellet. The pellet gently resuspended with a painter’s brush in isolation buffer without PVP, adjusted to 10 mM $MgCl_2$ and treated with DNase I (50 μg/ml) for one hour to

degrade nuclear DNA. The sample was adjusted to 2 0mM EDTA and centrifuged at 15,000 g for 10 min. The pellet was gently resuspended in 500 μl of buffer II (0.3 M sucrose, 0.05 M Tris-HCl, 0.02 M EDTA, 0.1% BSA, pH 7.5) with a painter's brush. After resuspension, the sample was layered above 30 ml of Percoll cushion (28% Percoll; 0.3 M sucrose; 0.05 M Tris-HCl; 0.02 M EDTA; pH 7.5) and centrifuged at 40,000g for 90 min. Yellowish mitochondrial ring in the middle of the cushion was collected. Mitochondrial fraction was rinsed with washing buffer (0.3 M mannitol; 50 mM Tris-HCl; 1 mM EDTA; pH7.5) by three times of centrifugation at 15,000g for 10 min.

Mitochondrial DNA was extracted following a method described by Kim (2004). Isolated mitochondria were resuspended in lysis buffer (50 mM Tris-HCL; 10 mM EDTA; 2% sarkosyl; 25 μl of proteinase K (10 mg/ml); pH 8.0) and incubated at 65^{o}C for one hour. After incubation, 200μl of 3M ammonium acetate and equal volume of phenol:chloroform (1:1) were added and centrifugated at 1,000g for 10 min. Two volume of ethanol was added to supernatant and centrifuged at 15,000g for 10 min. The pellet was rinsed with 70% ethanol by centrifugation at 15,000g for 10 min. The dried pellet was dissolved in 100 μl of TE buffer.

DNA sequencing

mtDNA sequencing was performed by GS-FLX system (Roche Applied

Science, Indianapolis, USA) and produced sequences were assembled by Newbler Assembler Software Version 2.0 (454 Life Sciences, Branford, USA) in National Instrumentation Center for Environmental Management (NICEM, Seoul, Republic of Korea) (Table 1).

Sequence assembly

Generated contigs were further assembled as following strategies. First of all, analysis using the Basic Local Alignment Search Tool was performed for contig sequences longer 1kb against Genbank nucleotide database (http://www.ncbi.nlm.nih.gov). The contig sequences which contained significant matches with known mtDNA sequences from other species were used for the next analyses. In the second step, a DNA library in which the average size of inserts was about 3kb in length was constructed and the end sequences of inserts were analyzed using ABI3700 sequencing system (Applied Biosystems, Foster city, USA) in NICEM (Seoul, Republic of Korea) (Table 1). The mate-pair information of insert end sequences was used to order contig sequences. In the third step, primers were designed from end sequences of each contig and all of the possible combinations of primers were used in PCR analysis to identify connected contigs. If a gap sequence obtained from PCRs contained only plastid-derived sequence, the gap was considered to be obtained due to contamination of the plastid during mitochondrial DNA isolation, not because of the existence of subgenomic mtDNA

molecules. Through this step, gaps could be filled with the sequences of PCR products to reduce number of contigs. In the fourth step, genome walking from the ends of each assembled and connected scaffold sequences was conducted using GenomeWalker^{TM} universal kit (Clontech, Mountain View, USA) according to manufacturer's instruction. In the fifth step, LA-PCRs were performed to fill the large gaps between remained scaffolds using TaKaRa LA Taq^{TM} (TaKaRa, Shiga, Japan). Finally, all of information obtained by the stepwise approach was used to construct a master circle model which contains at least one copy of every mtDNA contigs. The original contig sequences which could be connected to multiple contig sequences may be due to containing of repeated sequences. These sequences were inserted in master circle more than two times according to the obtained information on the relationship with other contigs. The validity of insertions of repeated sequence was further analyzed by PCRs with primers designed from flanking regions of repeated sequences.

Gene annotation and identification of *orf*s

The protein and rRNA genes on mtDNA sequences were identified by Basic Local Alignment Search Tool using nucleotide and protein database of Genbank (http://www.ncbi.nlm.nih.gov). The tRNA genes were identified using tRNA scan-SE program (http://lowelab.ucsc.edu/tRNAscan-SE/). *Orf*s which were predicted to encode hypothetical proteins longer than 100 amino acids were

predicted using custom-made Perl scripts.

Sequences comparison and repeat sequence analysis

Alignment between target sequences was performed using BLASTN algorism of Basic Local Alignment Search Tool (http://blast.ncbi.nlm.nih.gov/). Pools of repeated sequences were obtained by analysis using BLASTN algorism of Basic Local Alignment Search Tool (http://blast.ncbi.nlm.nih.gov/) in which a given target sequence was used as both query and subject sequence. The alignments which meet the given criteria for syntenic sequence blocks or repeated sequences in length and similarity were isolated and visualized as Scalable Vector Graphics using custom-made Perl scripts.

RESULTS

Assembly of complete mitochondrial genome sequence

As a result of NGS sequencing using 454 GS-FLX system, contigs containing most of mitochondrial genome information were obtained for a pepper CMS line, FS4401 and a fertile line, Jeju, respectively. However, the total number of mtDNA contigs (>1kb) for each line were much more than expected when the high coverage of sequencing (>100 X) was considered (Table 1). The reasons for this were revealed in the process of further assembly of contigs and analysis of gap sequences. Firstly, the contamination of plastid DNA during sample preparation hampered the sequence assembly at the positions of mitochondrial genomes where plastid-derived sequences were located. Secondly, ends of large repeated sequences remained to be unconnected due to the short length of individual reads in 454 GS-FLX system (Table 1). Therefore, DNA library containing inserts which were 3 kb in average length was constructed and the end sequences of inserts were analyzed by ABI3700 for the ordering of contigs based on mate-pair information (Table 1). In addition, PCR analysis and genome walking from contig ends were performed to screen all of the possible combinations of large repeated sequences with other sequences. The final circular molecules for complete mitochondrial genomes included every mtDNA contig

longer than 1kb at least one time and was consistent with all of the results produced for sequence assembly.

Table 1. Sequencing and contig assembly results

Categories	FS4401	Jeju
Total length of sequences analyzed by 454 GS-FLX (bp)	58,113,817	64,752,391
Average length of sequences analyzed by 454 GS-FLX (bp)	247	249
No. mtDNA contigs assembled by Newbler2.0 (>1kb)	33	40
Total length of mtDNA from contigs (>1kb) (bp)	439,940	451,255
No. mtDNA contigs assembled by ABI3700 sequencing (mate-pair) and PCR analysis	12	12
Length of complete mtDNA contig	507,450	493,911

Comparative analysis of general features and sequence contents between mitochondrial genomes

General features of mitochondrial genomes were compared between FS4401, Jeju, and tobacco. Tobacco is the only species in Solanaceae whose complete mtDNA sequence was reported (Sugiyama et al., 2005). Complete mitochondrial genome of FS4401 and Jeju were 507,450 and 493,911 bp in length,

respectively, which were significantly larger than tobacco (Table 2). Proportion of gene-coding sequences was similar between pepper mtDNAs which were 8.0% and 8.1%, respectively. Ratio of chloroplast-derived sequences was slightly higher in FS4401 (12.2%) than Jeju (11.0%). These values were higher about 2.5 times than in tobacco (4.5%). In repeated sequence contents, tobacco had larger proportion of repeated sequences mainly due to containment of longer large repeated sequences. The total length of repeated sequences was longer in Jeju than FS4401 among pepper lines. Contents of genes encoding proteins and rRNAs were the same between Jeju and tobacco, whereas FS4401 had a duplicated copy of the *atp6* gene named as *ψatp6-2* (Kim et al., 2006), additionally. Number of genes encoding tRNAs were different between mtDNAs. FS4401 had the largest number of tRNAs (28) followed by Jeju and Tobacco, which were 24 and 21, respectively.

Sequence alignment analysis showed that most of sequence contents were shared between two mtDNAs of pepper. The 93.8% of sequence of FS4401 could be aligned with Jeju and 96.6% of Jeju with FS4401 (Table 3). In comparative analysis with tobacco sequence, only about 55% of each genome could be aligned with tobacco mtDNA.

Table 2. General features of FS4401 and Jeju mitochondrial genome and comparison with tobacco mitochondrial genome

Features	FS4401	Jeju	*Nicotiana tabacum*
Genome size (bp)	507,450	493,911	430,597
GC content (%)	44.5	44.6	45
Coding sequences (bp) [a]	40,696 (8.0%)	40,134 (8.1%)	43,425 (10.1%)
cp-derived sequences (bp) [b]	61,737 (12.2%)	54,459 (11.0%)	19,492 (4.5%)
Repeated sequences (bp) [c]	42,834 (8.4%)	55,290 (11.1%)	73,511 (17.1%)
Gene content (number)	68	63	60
Protein coding genes	37[e]	36	36
rRNAs	3	3	3
tRNAs	28 (14)	24 (12)	21 (9)

[a] All of the copies of duplicated genes were included.

[b] BLASTN algorithm was used to screen mtDNA sequences that could be aligned with chloroplast DNA sequences. Complete sequences of chloroplast genomes of FS4401 (Jo et al., 2011) and *Nicotiana tabacum* (Shinozaki et al., 1986) were used to screen chloroplast(cp)-derived sequences in FS4401/Jeju and *N.tabacum*, respectively. The value for *N.tabacum* was different from the Sugiyama et al. (2005) probably due to difference in methodology to isolate cp-derived sequence.

[c] Repeated sequences longer than 100bp and showing similarity higher than 95% between copies were considered. Nucleotides which were included in repeated sequences at least one time were counted without repetition.

[d] Genes that encode *mttB* and *tatC* were not included because of absence of a start codon for these genes although these genes showed high similarity with orthologs in other species.

[e] The number of tRNA genes which were included in cp-derived sequences was written in parenthesis.

Table 3. Ratio of mtDNA sequences which could be aligned between two mitochondrial genome of pepper and one of tobacco by BLASTN algorithm. The genomes in left column were used as references.

	FS4401	Jeju	*Nicotiana tabacum*
FS4401 (bp)	-	476,064 (93.8%)	234,583 (46.2%)
Jeju	477,070 (96.6%)	-	228,159 (44.9%)
Nicotiana tabacum	238,383 (55.4%)	236,737 (55.0%)	-

Gene contents and localization on mitochondrial genomes

Mitochondrial genes with known functions were annotated and localized on FS4401 and Jeju mitochondrial genomes, respectively (Fig.1). Protein coding genes were classified according to the functions of proteins. Thirty six protein encoding genes in Jeju included nine genes for complex I proteins (*nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9*), two for complexII (*sdh3, sdh4*), one for complexIII (*cob*), three for complexIV (*cox1, cox2, cox3*), three for ATP synthase subunits (*atp1, atp6, atp4, atp8, atp9*), eleven for ribosomal proteins (*rpl2, rpl5, rpl16, rps1, rps3, rps4, rps10, rps12, rps13, rps14, rps19*), four for proteins involved in biogenesis of cytochrome c biogenesis (*ccmB, ccmC, ccmFc, ccmFN*), and one for maturase (*matR*). FS4401 had one additional copy of the *atp6* gene (*ψatp6-2*). Both FS4401 and Jeju contained three of rRNAs (*rrn5, rrn18, rrn26*) while FS4401 contained four additional tRNA genes than Jeju.

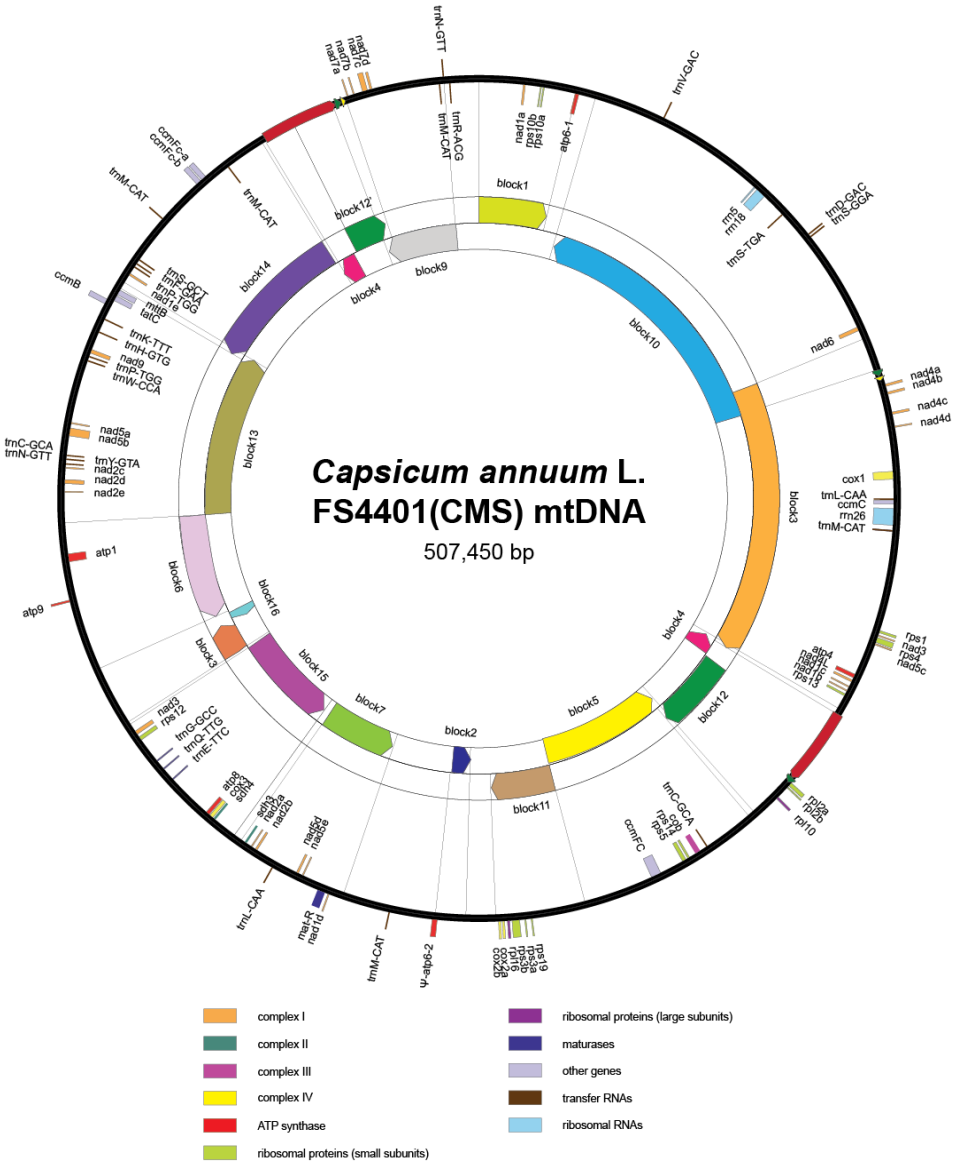
Comparison of protein encoding gene sequences between FS4401 and Jeju revealed six genes containing polymorphism on their sequences (Table 4). Sequence polymorphisms resulted in the change of protein sequence in *atp4*, *atp8*, *rpl2*, *sdh3*, and *atp6* whereas a synonymous substitution was detected on *matR*. Change in the length of gene product was predicted in *atp4*, *atp8*, and *rpl2* due to in-frame indels while length polymorphism in *atp6* involved a structural rearrangement (Fig. 2). The *atp6* gene in Jeju showed perfect match with *atp6-1* of FS4401 on the region which could be aligned between each other while upstream of conserved region could be aligned with a additional copy of *atp6* in FS4401, *ψatp6-2*. The gene sequence and the structure of gene flanking regions indicated that the *atp6* gene of Jeju is the *atp6-1* gene of male-fertile pepper reported in previous research (Kim el al., 2006) where two copies of *atp6* genes were present even in normal cytoplasm.

Comparison of genes carrying polymorphism with tobacco sequence showed that genotype on polymorphic site of *matR* and *sdh3* in FS4401 was the same with tobacco indicating that nucleotide substitutions on these were probably not related with sequence alteration during evolution of CMS cytoplasm.

Many of genes were located closely to each other forming gene clusters which may be co-transcription units. In total, fifteen clusters were detected in FS4401 and Jeju, which include *rpl5-rps14-cob-trnC*, *rps13-nad1bc-nad4L-atp4*, *rps1-nad3-rps4-nad5c*, *trnM-rrn26-ccmC-trnL*, *trnD-trnS*, *rrn18-rrn5*, *rps10a-*

rps10b-nad1a, *trnS-trnF-trnP-nad1e*, *nad9-trnP-trnW*, *trnC-trnN-trnY-nad2cde*, *nad3-rps12*, *atp8-cox3-sdh4*, *sdh3-nad2ab*, *nad1d-matR*, *rps19-rps3ab-rpl16-cox2ab* (Fig.1). Although numerous rearrangements between two mitochondrial genomes were detected, organization of gene cluster was highly conserved.

(a)



(b)

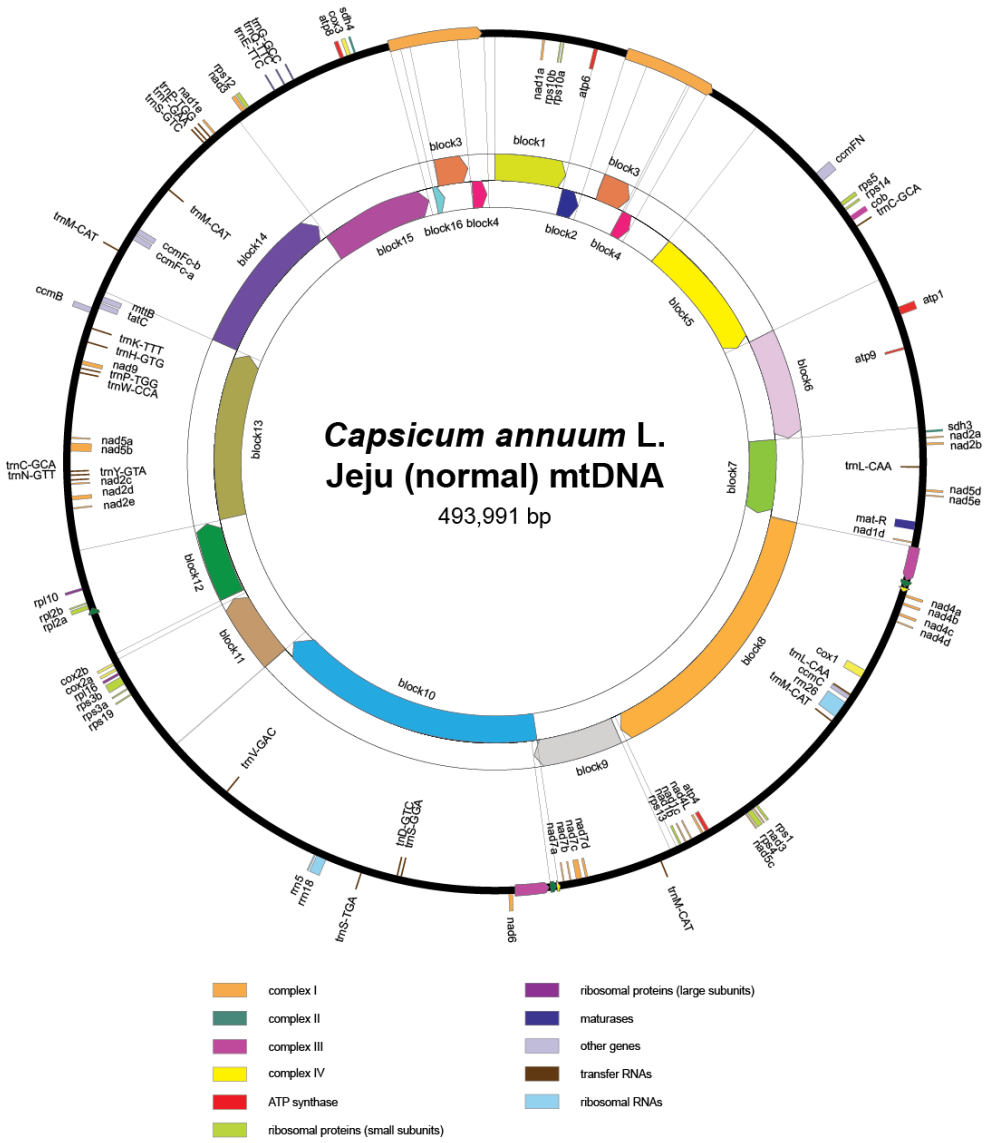


Fig. 1. Gene maps of the mitochondrial genomes of FS4401 (a) and Jeju (b). The genes drawn outside of the circle are transcribed clockwise and inside counterclockwise. The colors of the genes were classified according to the functions of the gene products. Large repeat sequences (>1kb) were drown as colored-arrows on circumference. Sequence blocks which were syntenic between genomes (>2kb; >95% similarity) were depicted on inner circle.

Table 4. Differences between Jeju and FS4401 in sequences of known genes. The pattern of base changes and corresponding amino acid changes was described. Left side of arrow shows the polymorphic site (numbers indicate the position of the first nucleotide written behind), polymorphic nucleotide (in capital letter), and corresponding amino acid on the genes in Jeju and right side in FS4401.

Genes in Jeju	Gene length in Jeju	Polymorphism in FS4401	Corresponding sequence in tobacco
matR	1977	904 gcA (A)→904 gcG (A)	904 gcG (A)
atp4	597	16 acGAATATGCAg (TNMQ) → 16 acg (T)	16 acGAATATGCAg (TNMQ)
atp8	462	178 ccCAACAGTTTg (PNSL) → 178 ccg (P)	178 ccCAACTGTTTg (PNCL)
rpl2	999	337 ccCGGGAAGGGgat (PGKGD) → 337 ccggat (PD)	337 ccCGGGAAGGGgat (PGKGD)
sdh3	333	178 tTCttc (FF) → 178 tCTttc (SF)	178 tCTttc (SF)
atp6	1296	*Ψ*atp6-2 283 gGt (G) → gCt (A) 316 ACa (T) →CAa (Q) 454 aaAGaa (KE) → aaCCaa (NQ) no similarity in downstream of 931th bp due to DNA rearrangement *atp6-1* no similarity in upstream of 497th bp due to DNA rearrangement	Higher similarity with *atp6-1* in FS4401

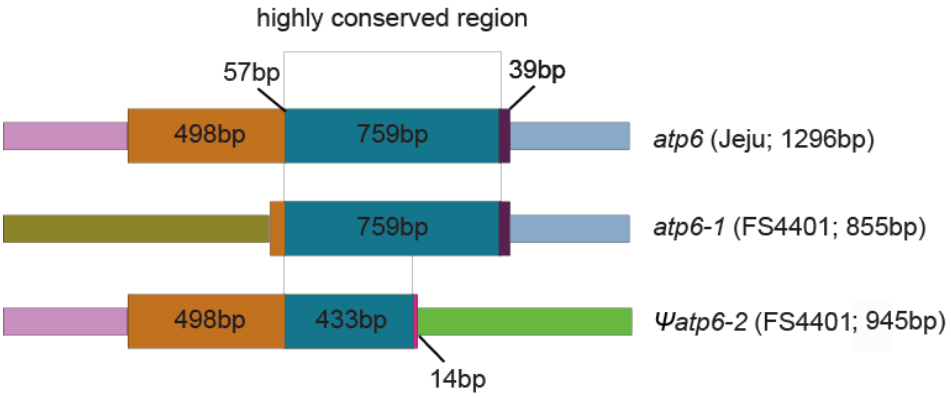


Fig.2. Structure of the *atp6* gene copies in Jeju and FS4401. The same structural units of sequences were depicted using the same color. Rectangles with sequence length indicate the *atp6* gene copies. Overall skeme of figure was adopted from Kim et al. (2006).

_orf_s unique to each mitochondrial genome

Open reading frames which encode proteins longer than 100 amino acids in length were screened in mitochondrial genomes (Table 5). A total of 162 and 148 *orfs* were detected in FS4401 and Jeju, respectively. Comparative analysis of these *orf*s showed that 38 and 28 *orf*s had polymorphisms with *orf* counterparts or were present only on each genome in FS4401 and Jeju, respectively. FS4401 mtDNA contained 17 *orfs* having sequence or length polymorphism and 21 *orfs* which were present only in FS4401 mtDNA whereas Jeju mtDNA included 16 *orfs* with polymorphism and 12 specific *orfs*. When the localization of *orfs* specific to FS4401 on mtDNA was investigated to screen candidates of CMS

associated gene, twelve *orfs* were shown to be close (< 2kb) from the edge of sequence blocks which are syntenic between two mitochondrial genomes (>2kb; >95% similarity) (Fig.3). These included *orf132c, orf133b, orf168, orf102e, orf126, orf141, orf300, orf104c, orf126, orf140, orf115b,* and *orf166*. Among these, five *orf*s including *orf132c, orf168, orf102e, orf300,* and *orf115b* were predicted to contain transmembrane domain by *in silico* analysis. Four *orf*s of these which were *orf132c, orf168, orf102e* and *orf300* were located closely with known genes in the same direction for possible transcription. The *orf507*, a strong candidate of CMS-associated gene reported in previous research (Kim et al., 2007), was indicated as *orf168* in this analysis (named according to polypeptide length instead of nucleotide length) and was included among finally selected *orf*s.

Table 5. *orf*s with polymorphism or unique to each mtDNA

Category	*orf*s specific in FS4401	*orf*s specific in Jeju
In frame Indel or SNPs	*orf133a, orf127, orf126*	*orf133a, orf129, orf130*
Length polymorphism due to rearrangement, frame shift or sequence variations on stop codon	*orf108a, orf109a, orf147, orf190, orf337, orf132c, orf133b, orf338, orf244, orf204b, orf300, orf140, orf115b, orf109b*	*orf160, orf126, orf370, orf277, orf109b, orf109c, orf171, orf117c, orf122b, orf110c, orf130, orf261, orf111b*
Specific presence on one of mitochondrial genome	*orf110a, orf119a, orf107b, orf131a, orf168, orf102i, orf132d, orf262, orf165, orf100c, orf141, orf107d, orf119b, orf100d, orf152c, orf104c, orf111b, orf115d, orf353, orf473, orf132e*	*orf103a, orf481a, orf104a, orf101b, orf112b, orf117b, orf148, orf217, orf102j, orf111c, orf103g, orf481b*
Total number of *orf*s	38	28

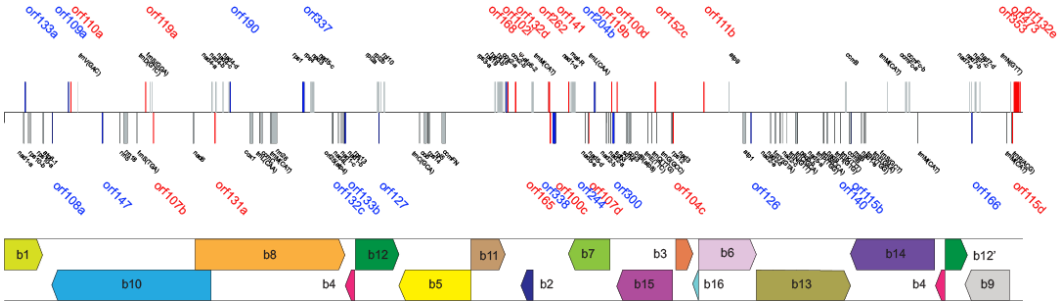


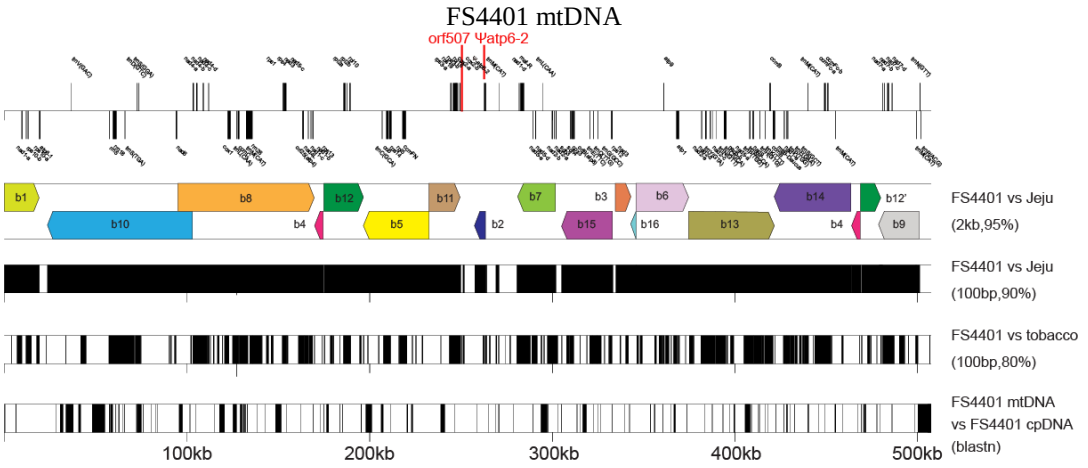
Fig.3. Distribution of *orf*s (longer than 300 bp) which were specific to FS4401 on FS4401 mtDNA. Red-colored *orf*s were present only in FS4401. Blue-colored *orf*s shows polymorphism in length or sequence with its counterpart in Jeju. The known genes were depicted in grey color.

Rearrangements of genome structure between CMS and normal mtDNA

A total of sixteen blocks of sequences which could be aligned between two pepper mitochondrial genomes by similarity higher than 95% and in sequence range broader than 2kb were defined and localized on each genomes (Fig.1 and 4). The sizes of blocks were from 2.9kb (block 16) to 79.5kb (block 10) (Table 6 and 7). Sequences in a part of block 12 were duplicated in FS4401 while block 3 and 4 were present as two copies in Jeju (Fig.4.). Both two mitochondrial genomes had a total of seventeen junctions between blocks, respectively. In FS4401, overlaps of sequences between blocks were detected in seven junctions while no-matching sequences between block were on ten junctions (Table 6). Among sequences on the no-matching sequences, the sequences between block 11 and 2, and between block 2 and 7 were noticeable and constituted 60.2% of total sequences in no-

matching sequences which were unique to FS4401 mtDNA. The *orf507* and *Ψatp6-2* genes which were characterized in previous researches to have possible relationship with CMS were localized on the large gap sequence between block 11 and 2, and one of the junctions of gap between block 2 and 7 (Fig.4a). In Jeju, a total of eight overlapping sequences and nine gap sequences were detected (Table 7). The gap sequence between block 4 and 5 contained the largest portion (43.5%) of sequences unique to Jeju. Most portion of large gap sequences were remained to be specific to each mtDNA in alignment analysis with less strict condition (>100bp; >90%) and could not be aligned with tobacco mtDNA sequence (Fig.4). A gap between block 9 and block 1 in FS4401 contained a chloroplast-derived sequence specific to FS4401 (Fig.4a).

(a)



(b)

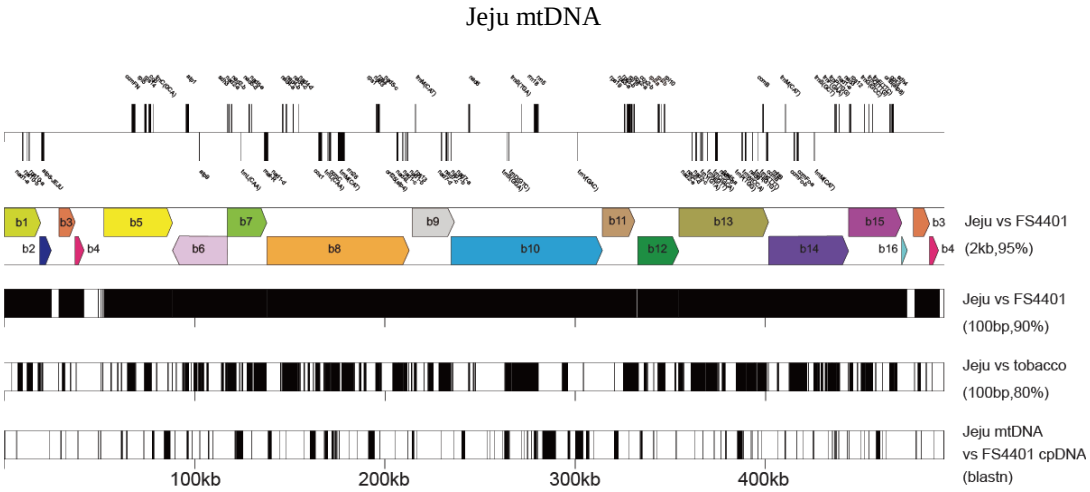


Fig. 4. Distribution of sequences showing similarity with the other pepper mtDNA, tobacco mtDNA and FS4401 plastid genome, respectively, in FS4401 mtDNA (a) and Jeju mtDNA (b). Localization of genes and syntenic sequence blocks (the linearized form of Fig.1.) were depicted on top of each figure. The aligned regions were described by black rectangles or bars. The criteria for alignments (minimum alignment length, minimum similarity, or algorithm) were indicated in parenthesis. The location of genes which was reported to be related with CMS (*orf507*, *Ψatp6-2*) was on FS4401 sequence.

Table 6. Localization of syntenic sequences blocks (>2kb; >95%) and size of gap or overlapping sequences between blocks on FS4401 mtDNA

syntenic block	length (bp)	direction	start site	end site	gap size (bp; %)[a]	overlaping sequence size (bp)
b 1	19,109	+	1	19,109	4,579 [b] (10.8)	-
b 10	79,476	-	23,687	103,162		8,009 [c]
b 8	74,694	+	95,154	169,847	214 (0.5)	-
b 4	4,692	-	170,060	174,751	377 (0.9)	-
b 12	21,668	+	175,127	196,794		22
b 5	35,959	-	196,773	232,731		8
b 11	17,024	+	232,724	249,747	7,910 (18.6)	-
b 2	6,016	-	257,656	263,671	17,652 (41.6)	-
b 7	20,756	-	281,322	302,077	3,325 (7.8)	-
b 15	27,874	-	305,401	333,274	1,597 (3.8)	-
b 3	8,478	+	334,870	343,347		76
b 16	2,906	-	343,272	346,177	39 (0.1)	-
b 6	28,796	+	346,215	375,010		19
b 13	46,985	+	374,992	421,976		65
b 14	42,021	-	421,912	463,932	207 (0.5)	-
b 4	4,738	-	464,138	468,875	377 (0.9)	-
b 12'	10,998	+	469,251	480,248		1,182
b 9	22,199	-	479,067	501,265	6,187 (14.6)	-

[a] The ratio of the given gap sequence to the total size of gap sequences were shown in percentage.

[b] The length of gap between the block in next line

[c] The sequence overlap with the block in next line

Table 7. Localization of syntenic sequences blocks (>2kb; >95%) and size of gap or overlap sequences between blocks on Jeju mtDNA

syntenic block	length (bp)	direction	start site	end site	gap size (bp; %) [a]	overlap sequence size (bp)
b 1	19,126	+	1	19,126	-	438 [b]
b 2	6,018	+	18,689	24,706	4,088 [c] (16.9)	-
b 3	8,483	+	28,795	37,277	-	32
b 4 [d]	4,693	+	37,246	41,938	10,467 (43.5)	-
b 5	35,987	+	52,406	88,392	2 (0.0)	-
b 6	28,836	+	88,395	117,230	-	51
b 7	20,757	+	117,180	137,936	109 (0.4)	-
b 8	74,723	+	138,046	212,768	1,550 (6.4)	-
b 9	22,206	+	214,319	236,524	-	1,777
b 10	79,532	+	234,748	314,279	-	40
b 11	17,043	+	314,240	331,282	1,502 (6.2)	-
b 12	21,638	+	332,785	354,422	54 (0.2)	-
b 13	47,076	+	354,477	401,552	-	18
b 14	42,071	+	401,535	443,605	30 (0.1)	-
b 15	27,932	+	443,636	471,567	-	16
b 16	2,905	+	471,552	474,456	3,931 (16.3)	-
b 3	8,483	+	478,388	486,870	-	32
b 4 [d]	4,739	+	486,839	491,577	2,334 (9.7)	-

[a] The ratio of the given gap sequence to the total size of gap sequences were shown in percentage.

[b] The sequence overlap with the block in next line

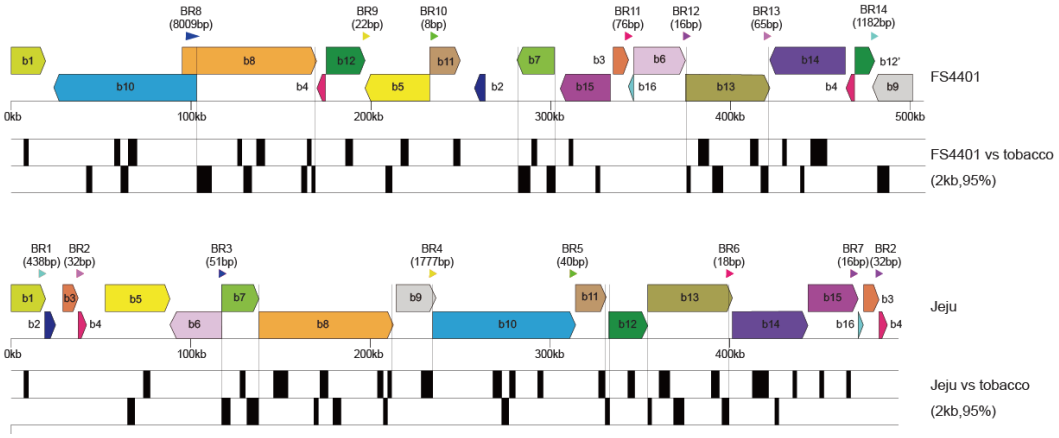
[c] The length of gap between the block in next line

[d] The length of the second copy of b 4 is longer than the other due to an addition of repeated sequence

The sequence overlaps between syntenic blocks in one mitochondrial genome corresponds to repeated sequences located at the edges of blocks in the other genome. The sizes of repeated sequences were varied from 8 to 8009 bp. All of the rearrangements that might occur via these sequences were depicted in Figure 5. For example, the BR1, which is the invert repeat sequence present at the edges of block 1 and block 2 in FS4401 might be involved in the recombination to result in the connection of block 1 to block 2 in Jeju (Fig.5b).

Analysis of alignment between tobacco mtDNA and each pepper mtDNA with the same condition (>2kb, >95%) used for alignments between pepper mtDNAs showed that several end points of the sequence blocks syntenic between tobacco and pepper were located very closely (<1kb) to the edges of blocks shared by FS4401 and Jeju (Fig.5). The number of block end regions included in this category was seven and eight in FS4401 and Jeju, respectively. At least two rearrangement events which include one during speciation of tobacco and pepper and the other during evolution of pepper mitochondrial genomes were expected in these regions. No single sequence blocks which could cover adjacent ends of blocks syntenic between FS4401 and Jeju were detected in the alignment with tobacco mtDNA (Fig.5).

(a)



(b)

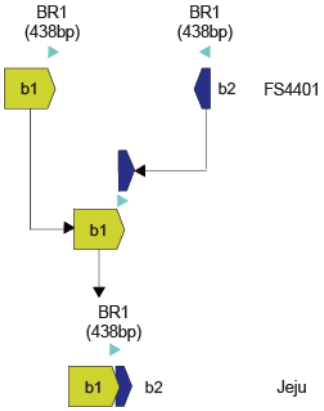


Fig. 5. Possible rearrangements via long or short repeated sequences at the end of syntenic blocks. BRs are the repeat sequences located on the eages of synthenic blocks. (a) Representation of sequences which were repeated in FS4401 and Jeju, respectively. (b) Schematic representation of possible rearrangement via repeated sequence at the end of syntenic blocks.

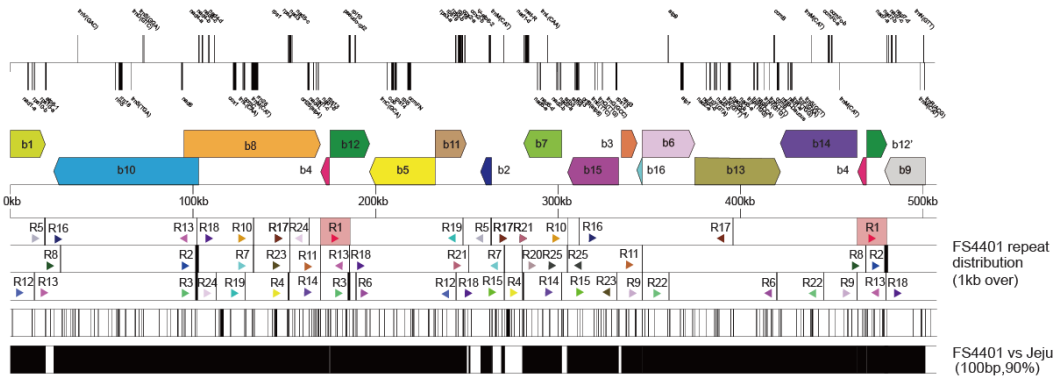
Distribution of repeated sequences on mitochondrial genomes

The repeated sequences that are longer than 100bp and show similarity higher than 95% between copies were screened and localized on mitochondrial genomes of FS4401 and Jeju, respectively (Fig 6). In FS4401, three large repeated sequences longer than 1kb were detected. The longest repeat was 16,278 bp in length and shared 4,670 bp with the longest one in Jeju. Other two large repeats which were 1,777 and 1,182 bp in length were included in or identical to large repeated sequences in Jeju, respectively. Intermediate-sized repeats (100bp-1kb) were 22 in number and spread around genome, but showed noticeable high density in the regions nearby ends of syntenic blocks or large repeated sequences, and the region containing large gap sequences unique to FS4401 (block 11-block 7) (Fig. 6a). Jeju also had three large repeated sequences longer than 1kb. The largest one was 17,234bp in length. The second largest repeat included the second largest one in FS4401 as a portion and 8,013bp in length. The third largest repeat was identical with the third largest on e in FS4401. In the case of intermediate-sized repeats, a total of sixteen were found in Jeju. These repeat were densely localized nearby ends of syntenic blocks or the second and third largest repeats (Fig. 6b). Although many intermediate-sized repeats existed nearby junctions of syntenic sequence blocks, most of the repeats could not be used to explain the rearrangement of sequence blocks by single recombination events which seemed to occur via some of the overlap sequences described in Fig.5. Some of the repeats

located nearby the ends of repeat blocks in one genome existed as single copies or absent in the other.

Other small repeated sequences which were screened with very low stringency (BLASTN) showed tendency to be clustered around or one side of large or intermediate-sized repeats (Fig.6). The region from one end of block 11 to block 2 in FS4401, where *orf507* and *Ψatp6-2* genes were localized, was noticeable in absence of any repeated sequences although high density of repeated sequences were detected around this region (Fig.6a).

(a)



(b)

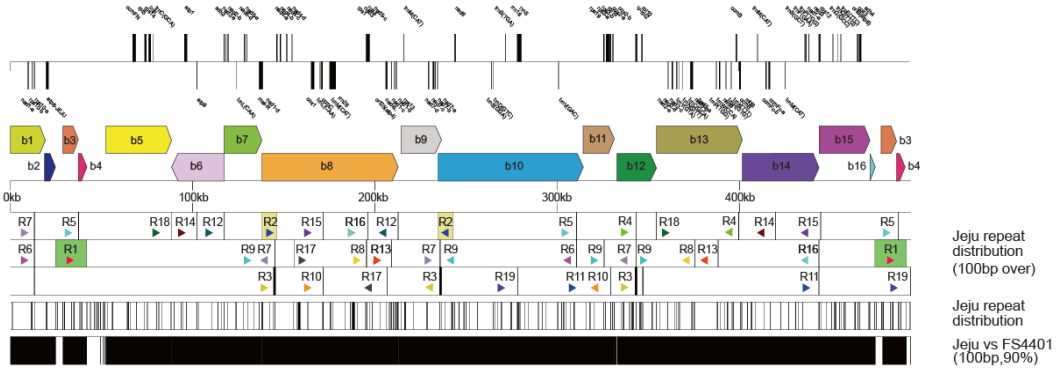


Fig.6. Distribution of repeated sequences which are longer than 100bp and show identity higher than 95% between copies in FS4401 (a) and Jeju (b). The names of repeats common between two mitochondrial genomes start with ‘C’ (eg. C7). The repeats in which ‘R’ is written in name were exist as repeats in one genome, but were single copies or absent in the other genome. Other repeated sequences obtained from analysis by BLASTN algorithm were shown below. Sequences which could be aligned between two mitochondrial genomes (>100bp, >90%) were described in bottom two lanes.

Structure of sequences around *orf507* and *ψatp6-2* gene

Mitochondrial DNA region containing *orf507* and *ψatp6-2* gene was shown to be unique to CMS mitochondrial genome in the comparative analysis with normal mtDNA performed in this study as well as the previous researches (Kim et al., 2006; Kim et al., 2007). Therefore, investigation on the sequence structure of this region was performed in detail (Fig.7). The *orf507* gene was located on the downstream of *cox2* and *ψatp6-2* was about 12kb apart from *orf507*in FS4401. However, not only *orf507* was absent, but also *cox2* and *atp6* were distantly located from each other in Jeju implying extensive rearrangements were occurred between two genes. Although the *cox2* gene was identical between two mtDNAs, sequences were divergent from 40 bp downstream of *cox2* 3' ends. In FS4401, a 132 bp repeat sequence (R19; Table 8) was started from 24 bp downstream of *cox2* and overlap with sequence in Jeju by 16 bp. The other repeat named as Ra was followed by R19 overlap with R19 by 11bp. A portion of this sequence was included in *orf507*. Sequences showing high similarity with R19 were detected in FS4401, Jeju and tobacco on 5' upstream region of the *nad9* gene. The part of *orf507* gene which was not covered by R19 did not show any similarity not only with any other sequences in FS4401, but also with Jeju and tobacco mtDNA, and any of the sequences registered in Genbank (http://www.ncbi.nlm.nih.gov/genbank). The sequence conserved between two genomes started from the 573bp downstream of *orf507* gene. Sequences named as

R21 and CS1 were conserved on 367 bp downstream of *cox2* in Jeju. Among sequence elements, R21 was duplicated in FS4401 whereas it was single copy in Jeju (Table 8). Sequence following CS1 was highly specific to FS4401 and could not be aligned with any sequences in Genbank until it reached to *ψatp6-2* gene. Although 5' region of *ψatp6-2* was highly conserved with *atp6* gene in Jeju, sequence divergence started again near the 3' end of *ψatp6-2* gene resulting in generation of premature stop codon. The repeated sequence named Rb followed by *ψatp6-2* sequence overlapping with sequences conserved between *ψatp6-2* and *atp6* gene by 12bp. Again, the following sequence units named CS2 and R21 was overlap with upstream sequence unit by 5bp. CS2 and R21 sequences were conserved in Jeju on the upstream of sequence unit, C1. The largest gap sequence between syntenic blocks (b 2-b 7) was located on the downstream of C1.

In this comparative analysis, multiple incorporation of short sequence elements via very short sequences overlap between connected elements were detected downstream of *cox2* and *ψatp6-2* in FS4401. These sequence elements were present as repeated sequence only in FS4401 implying that these sequences were duplicated during rearrangements (Table 8). Especially, R21 on Jeju was duplicated on the downstream of *ψatp6-2* in FS4401 resulting in generation of a repeat pair around the *ψatp6-2* gene. The *orf507* gene and other related sequence element were seemed to be inserted between *cox2* and *R21* by multiple DNA rearrangements.

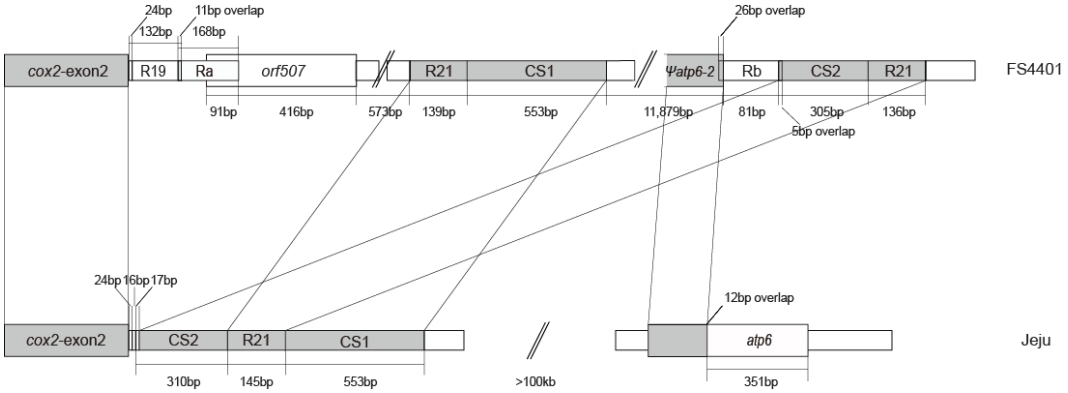


Fig.7. Comparison of sequence structure around *orf507* and *ψatp6-2* between FS4401 and Jeju. The sequence blocks which were conserved between two lines were described in gray color.

Table 8. Repeated sequences around *orf507* and *ψatp6-2*.

Repeat name[a]	Repeat length	Similarity between repeats in FS4401 (%)	Presence (+) or absence (-) in Jeju[b]
R19	132	98	+
Ra	168	91	+
R21	139	95	+
Rb	81	96	-

[a] Repeat names were followed those in Fig.7.
[b] All of the repeated sequences defined in FS4401 were single copies in Jeju if present.

DNA rearrangement pattern and localization of CMS-associated genes in CMS mitochondrial genomes of other crop species

Rearrangement pattern of CMS mitochondrial genomes were investigated in crop species where complete sequencing of CMS mitochondrial genome and at least one of normal mtDNA were finished and genes responsible for CMS were identified (Fig.8). Alignment of CMS mtDNA with normal mtDNA in seven mitochondrial genomes resulted in numerous sequence blocks (>2kb; >95%) as already reported by many of other researches implying intensive DNA rearrangements are common during the origination of CMS cytoplasm (Allen et al., 2007; Chen et al., 2011; Liu et al., 2011; Park et al., 2013; Tanaka et al., 2012; Satoh et al., 2004). All of the genes known to be associated with CMS were localized closely to the edge of syntenic sequence blocks. Especially, CMS-associated genes in pepper and radish were localized near the end of long sequences on the gap between blocks. Localization of repeated sequences (>100bp; >95%) showed that CMS genes were always located near repeated sequences.

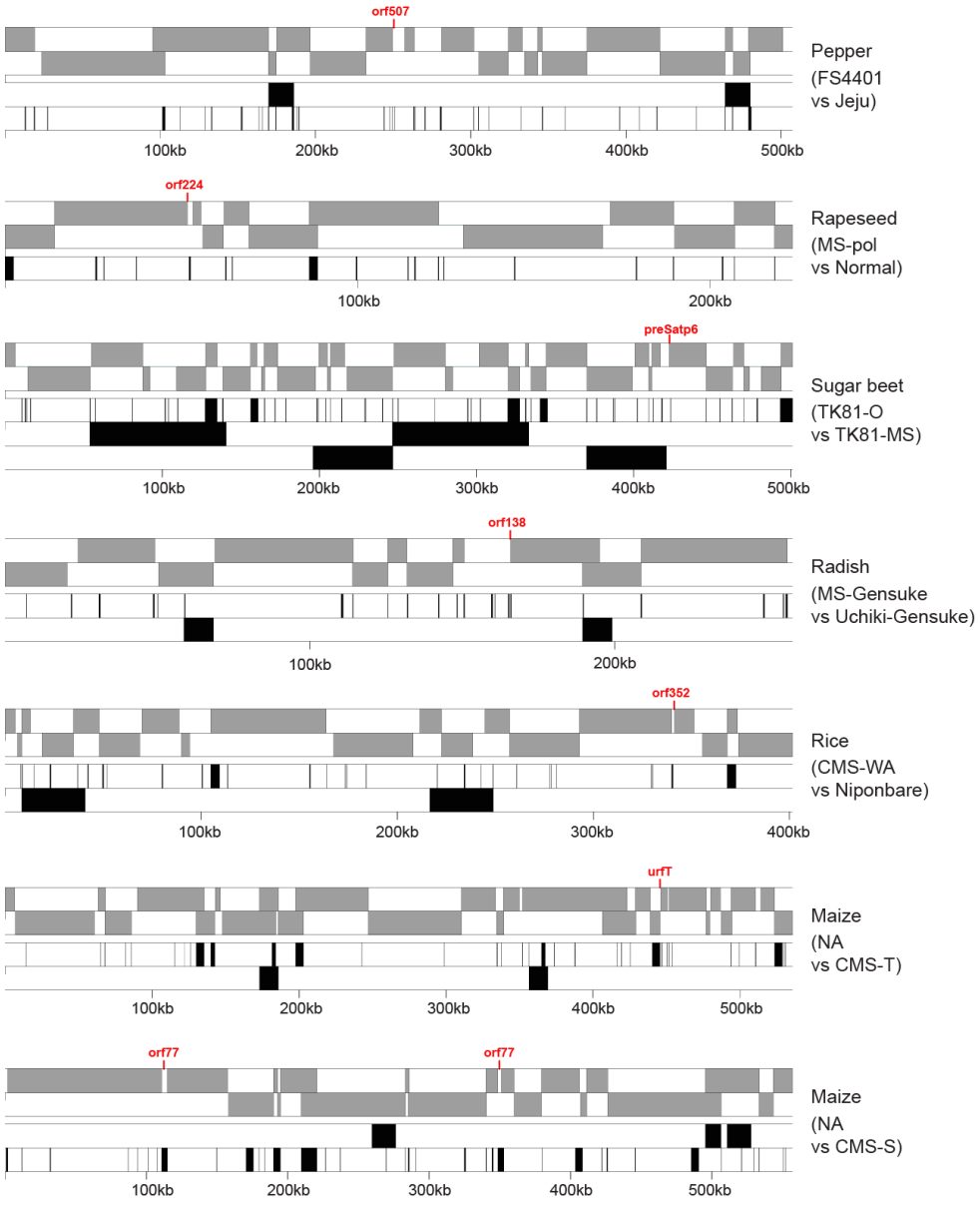


Fig.8. Sequence blocks (gray color) which were syntenic (>2kb, >95%) between a CMS and a normal cytoplasm in different crops. mtDNAs of CMS lines were used as reference always. Distribution of repeated sequences (>100bp, >95%) in CMS lines were described with black-colored bars and boxes. The CMS-associated genes in each crop were indicated above the alignments. Sequence blocks and repeated sequences were depicted in two or three layers to show the overlapped sequence units clearly.

DISCUSSION

In this study, the complete mitochondrial genome sequence of pepper (*C.annuum* L.) was firstly reported. It was the second mitochondrial genome fully sequenced in Solanaceae where only tobacco mitochondrial genome has been published so far (Sugiyama et al., 2005). Therefore, the mtDNA sequence of pepper can provide valuable source to study the evolution of mitochondrial genomes in Solanaceae. The contents and sequences of protein coding genes were highly conserved between tobacco and pepper mtDNA. However, overall structure and contents of non-coding sequences was extensively changed resulting in less than 50% coverage of pepper mitochondrial genome by tobacco mtDNA. Similar heavy rearrangements within a plant family have been reported in the comparative analysis of Arabidopsis and rapeseed mitochondrial genomes (Handa, 2003). Although gene sequences were protected from rearrangement events, clustering patterns of genes were changed in several cases. The *atp9-rps13-nad1bc*, *nad4-rps1-nad5ab*, *nad3-nad1a*, and *rps4-nad6* clusters which were reported in tobacco did not conserved in pepper (Sugiyama et al., 2005). Co-transcriptions of gene clusters including *nad3-nad1a* have been reported in tobacco while *nad3* and *nad1a* are incorporated in different cluster in pepper (Lelandais et al., 1996; Gutierres et al., 1997). Therefore, change in co-

transcription pattern might be resulted from DNA rearrangement during speciation of tobacco and pepper.

Numerous rearrangements of mitochondrial DNA were also detected even in the comparison of CMS and normal mtDNA within *C.annuum* species. High conservation of gene coding sequences and clustering pattern of genes indicated that mtDNA molecules maintaining normal function for gene expression was selected during rearrangement events or sequences in transcribed regions have characteristics which efficiently suppress rearrangements. However, multiple rearrangements occurred outside of gene clusters resulted in the fragmentation of alignment unit between two genomes. Several sequence blocks which were syntenic between genomes contained overlapping sequences that might explain rearrangement events directly. However, many of sequence blocks were connected with sequences unique to each genome or the overlapping sequences were too short to mediate homologous recombination. Several studies using *msh1*mutants explained the DNA rearrangement in Arabidopsis mtDNA as the result of nonhomologous end-joining (NHEJ) and asymmetric recombination via intermediate-sized repeats followed by randomly occurring double strand breaks (Davila et al., 2011; Shedge et al., 2007). Sequences having micro homology or without any homology are joined by NHEJ while asymmetric recombination is accompanied with repeat sequences longer than a given length (Davila et al., 2011). Therefore, rearrangement events at the end of repeat blocks in pepper

seemed to involve both NHEJ and asymmetric recombination after double strand breaks.

Large portion of repeated sequences were detected nearby structural rearrangement points. Many of them did not seem to be involved in the rearrangements between syntenic blocks directly because the sequence blocks were connected to other sequences by NHEJ or other repeat sequences nearby were responsible for rearrangements. If rearrangements and integration of repeated sequences were occurred independently in these cases, the regions might be prone to be rearranged acting as hotspots for double strand breaks. In fact, significant number of ends of syntenic blocks which could be aligned to tobacco mitochondrial genome were located closely to the junctions of syntenic blocks defined in the alignment of mtDNA of CMS and normal pepper lines. This implies those regions experienced at least two rearrangement events in a very short distance which includes the one between pepper and tobacco and the other between CMS and normal pepper lines. Clustering of repeated sequences was also reported in Arabidopsis (Forner et al., 2005; Shedge et al, 2009). Why recombinations are frequently occur in specific regions is still unknown although localization of DNA cruciforms, localized melting due to high transcriptional activity, and stalling replication folks were suggested as possible explanations (Shedge et al., 2007). Alternatively, there might be possibility that the asymmetric recombination or NHEJ on specific region in earlier stage reinforced the potential

for rearrangement in this region by generating broken ends of DNA or resulting in integration of already existing DNA segment to make new repeated sequences. In asymmetric recombination on intermediate-sized repeat sequence, one parental molecule might be remained to have broken ends after recombination (Shedge et al., 2007). One of the possible fates of this broken end is connection with other molecule by NHEJ (Kreuzer and Stohr, 2002). If a DNA sequence already existed in genome is incorporated by NHEJ, a new repeat pair can be generated in turn. In pepper, the region around *orf507* and *ψatp6-2* shows combination of multiple repeated sequences and other single copy sequences possibly undergone NHEJ. In pepper mtDNA level, large portion of small repeated sequences were found nearby larger repeated sequences. If we consider this scenario, the sites that already experienced rearrangements during divergence of pepper and tobacco might get more potential to drive rearrangements in the evolution within pepper species.

The *orf507* and *ψatp6-2* gene have been known to be associated with CMS in pepper (Kim et al., 2006; Kim et al., 2007). The genomic regions around these genes were highly specific in CMS line and not matched with any other known sequences. Recently, the similar result was reported in radish *Ogura* type cytoplasm in which the CMS-associated gene, *orf138*, was located on an edge of the largest genomic region unique to CMS line (Tanaka et al., 2012). Although the insertion of the DNA region containing *orf138* could be explained simply as the

result of the homologous recombination using a pair of inverted repeat sequences on the ends, the region around *orf507* and *ψatp6-2* in pepper contained more complicated structure to predict the mechanism by which the genomic structure of this region originated. The notable features in this structure were the presence of DNA fragments that seemed to be joined by NHEJ and a pair of intermediate-sized repeats (R21) of which a copy is located on the downstream of *orf507* and the other on the downstream of *ψatp6-2*. Several cycles of NHEJ on the downstream of *cox2* gene might result in incorporation of R19, Ra, and *orf507* specific sequence and deletion of CS2 on the downstream of *cox2* gene. As the next stage, two cycles of NHEJ using microsynteny of end sequences might result in the incorporation of a sequence segment located on the downstream of *cox2* gene (CS2+R21) to the downstream of *ψatp6-2* making this segment as repeated sequence. The newly obtained repeated sequences (R21, R19, Ra) acquired by NHEJ process might further facilitated rearrangements in this region by the asymmetric recombination process to result in incorporation of known sequences in this region.

The origin of 3' part of *orf 507* gene and large portion of the region around *orf507* and *ψatp6-2* was still remained to be unknown. One possible answer for how these sequences are specifically present in CMS line might be substoichiometric shift (SSS) model which has been reported in several plant

species (Arrieta-Montiel et al., 2001; Feng et al., 2009; Janska et al., 1998; Kim et al., 2007). According to this model, some of subgenomic molecules of mitochondrial DNA are present in very small copy number in normal condition in which recombination in intermediate-sized repeated sequences are suppressed, but they can be efficiently amplified and incorporated in the dominant form of subgenomic molecule due to occur of recombination dependent replication if the recombination in intermediate-sized repeated sequences is activated in certain condition (Shedge et al., 2007). In fact, small amount of *orf507* and *ψatp6-2* were detected by PCR even in fertile lines in pepper (Jo et al., 2009). Therefore, the subgenomic molecule containing CMS-associated gene might be maintained in low copy number even in normal pepper lines and integrated to autonomously-replicated master circle molecule in CMS line by asymmetric recombination via intermediate-sized repeat sequences around these genes when suppression of it was release in special conditions, for example, in interspecific crosses.

We performed analysis on the organization syntenic sequence blocks and repeat distribution to find the relationship with the location of CMS-associated genes on CMS mitochondrial genomes for seven cytoplasm types of six plant species in which complete mitochondrial sequence were analyzed in CMS line and a least one normal of line, respectively, and the CMS-associated genes were identified. In all of the cases, CMS genes were located at the edge of CMS-specific sequences located between syntenic blocks and close to intermediate-

sized repeat sequence or on the repeat sequence itself (CMS-S in maize). These findings fit well with the expected manner of origination of CMS genes which involves multiple rearrangement by NHEJ create novel DNA sequence region and copy number increase by recombination via adjacent repeat sequence. The close localization of CMS gene to syntenic sequence blocks might be due to the needs for sequence elements required in transcription of chimearic *orf*. In Arabidopsis, majority of chimearic *orfs* were shown not to be transcribed (Giege et al., 1998). It implies that utilization of promotors on conserved region might be requisite for the transcription of *orf*s. These common features of CMS-associated genes are expected to provide strategy to screen unknown CMS-gene candidates by comparative genomics approach which was exemplified in a new CMS cytoplasm type of radish (Park et al., 2013).

REFERENCES

Abdelnoor RV, Yule R, Elo A, Christensen AC, Meyer-Gauen G, Mackenzie SA (2003) Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. Proc Natl Acad Sci U S A 100: 5968-73

Allen JO, Fauron CM, Minx P, Roark L, Oddiraju S, Lin GN, Meyer L, Sun H, Kim K, Wang C, Du F, Xu D, Gibson M, Cifrese J, Clifton SW, Newton KJ (2007) Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 177: 1173-92

Andre C, Levy A, Walbot V (1992) Small repeated sequences and the structure of plant mitochondrial genomes. Trends Genet 8: 128-32

Arrieta-Montiel M, Lyznik A, Woloszynska M, Janska H, Tohme J, Mackenzie S (2001) Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean. Genetics 158: 851-64

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA (2009) Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics 183: 1261-8

Ashutosh, Kumar P, Dinesh Kumar V, Sharma PC, Prakash S, Bhat SR (2008) A novel *orf108* co-transcribed with the *atpA* gene is associated with cytoplasmic male sterility in *Brassica juncea* carrying Moricandia arvensis cytoplasm. Plant Cell Physiol 49: 284-9

Chen J, Guan R, Chang S, Du T, Zhang H, Xing H (2011) Substoichiometrically different mitotypes coexist in mitochondrial genomes of *Brassica napus* L. PLoS One 6: e17662

Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ, Weigel D, Mackenzie SA (2011) Double-strand break repair processes

drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol 9: 64

Feng X, Kaur AP, Mackenzie SA, Dweikat IM (2009) Substoichiometric shifting in the fertility reversion of cytoplasmic male sterile pearl millet. Theor Appl Genet 118: 1361-70

Forner J, Weber B, Wietholter C, Meyer RC, Binder S (2005) Distant sequences determine 5' end formation of cox3 transcripts in *Arabidopsis thaliana* ecotype C24. Nucleic Acids Res 33: 4673-82

Giege P, Konthur Z, Walter G, Brennicke A (1998) An ordered *Arabidopsis thaliana* mitochondrial cDNA library on high-density filters allows rapid systematic analysis of plant gene expression: a pilot study. Plant J 15: 721-6.

Gulyas G, Pakozdi K, Lee JS, Hirata Y (2006) Analysis of fertility restoration by using cytoplasmic male-sterile red pepper (*Capsicum annuum* L.) lines. Breed Sci 56:331-334

Gulyas G, Shin Y, Kim H, Lee JS, Hirata Y (2010) Altered transcript reveals an *orf507* sterility-related gene in chili pepper (*Capsicum annuum* L.). Plant Mol Bio Rep 28: 605-12

Gutierres S, Lelandais C, Paepe RD, Vedel F, Chetrit P (1997) A mitochondrial sub-stoichiometric orf87-nad3-nad1 exonA co-transcription unit present in solanaceae was amplified in the genus Nicotiana. Curr Genet 31: 55-62

Handa H (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res 31: 5907-16

Janska H, Sarria R, Woloszynska M, Arrieta-Montiel M, Mackenzie SA (1998) Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. Plant Cell 10: 1163-80

Jo YD, Jung HJ, and Kang BC (2009) Development of a CMS specific marker based on chloroplast-derived mitochondrial sequence in pepper. Plant Biotechnol Rep 3: 309-315

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 Rep 30: 217-29

Kim DH, Kang JG, Kim BD (2007) Isolation and characterization of the cytoplasmic male sterility-associated *orf456* gene of chili pepper (*Capsicum annuum* L.). Plant Mol Biol 63: 519-32

Kim DH, Kim BD (2006) The organization of mitochondrial *atp6* gene region in male fertile and CMS lines of pepper (*Capsicum annuum* L.). Curr Genet 49: 59-67

Kim S, Lim H, Park S, Cho KH, Sung SK, Oh DG, Kim KT (2007) Identification of a novel mitochondrial genome type and development of molecular markers for cytoplasm classification in radish (*Raphanus sativus* L.). Theor Appl Genet 115: 1137-45

Lee YP, Kim S, Lim H, Ahn Y, Sung SK (2009) Identification of mitochondrial genome rearrangements unique to novel cytoplasmic male sterility in radish (*Raphanus sativus* L.). Theor Appl Genet 118: 719-28

Lelandais C, Gutierres S, Mathieu C, Vedel F, Remacle C, Marechal-Drouard L, Brennicke A, Binder S, Chetrit P (1996) A promoter element active in run-off transcription controls the expression of two cistrons of *nad* and *rps* genes in *Nicotiana sylvestris* mitochondria. Nucleic Acids Res 24: 4798-804

Liu H, Cui P, Zhan K, Lin Q, Zhuo G, Guo X, Ding F, Yang W, Liu D, Hu S, Yu J, Zhang A (2011) Comparative analysis of mitochondrial genomes between a wheat K-type cytoplasmic male sterility (CMS) line and its maintainer line. BMC Genomics 12: 163

Millar AH, Sweetlove LJ, Giege P, Leaver CJ (2001) Analysis of the Arabidopsis mitochondrial proteome. Plant Physiol 127: 1711-27

Palmer JD (1990) Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet 6: 115-20

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97: 6960-6

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28: 87-97

Park JY, Lee YP, Lee J, Choi BS, Kim S, Yang TJ (2013) Complete mitochondrial genome sequence and identification of a candidate gene responsible for cytoplasmic male sterility in radish (*Raphanus sativus* L.) containing DCGMS cytoplasm. Theor Appl Genet 128: 1763-74

Peterson PA (1958) Cytoplasmically inherited male sterility in *Capsicum*. Amer Nat 92:111-9

Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, Mikami T (2004) The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. Mol Genet Genomics 272: 247-56

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA (2007) Plant mitochondrial recombination surveillance requires unusual *RecA* and *MutS* homologs. Plant Cell 19: 1251-64

Shedge V, Davila J, Arrieta-Montiel MP, Mohammed S, Mackenzie SA (2010) Extensive rearrangement of the Arabidopsis mitochondrial genome elicits cellular conditions for thermotolerance. Plant Physiol 152: 1960-70

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 BY, 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: 2043-9

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10:e1001241

Small I, Suffolk R, Leaver CJ (1989) Evolution of plant mitochondrial genomes via substoichiometric intermediates. Cell 58: 69-76

Stohr BA, Kreuzer KN (2002) Coordination of DNA ends during double-strand-break repair in bacteriophage T4. Genetics 162: 1019-30

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M (2005) The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genomics 272: 603-15

Tanaka Y, Tsuda M, Yasumoto K, Yamagishi H, Terachi T (2012) A complete mitochondrial genome sequence of Ogura-type male-sterile cytoplasm and its comparative analysis with that of normal cytoplasm in radish (Raphanus sativus L.). BMC Genomics 13: 352

Wang Z, Zou Y, Li X, Zhang Q, Chen L, Wu H, Su D, Chen Y, Guo J, Luo D, Long Y, Zhong Y, Liu YG (2006) Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. Plant Cell 18: 676-87

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84: 9054-8

Zaegel V, Guermann B, Le Ret M, Andres C, Meyer D, Erhardt M, Canaday J, Gualberto JM, Imbault P (2006) The plant-specific ssDNA binding protein

OSB1 is involved in the stoichiometric transmission of mitochondrial DNA in Arabidopsis. Plant Cell 18: 3548-63

CHAPTER II

Isolation of the *Restorer-of-fertility* Candidate Gene in Pepper (*Capsicum annuum* L.)

ABSTRACT

Cytoplasmic-genic male sterility (CGMS) has been used for efficient production of hybrid seeds in peppers (*Capsicum annuum* L.). Although mitochondrial candidate genes which may responsible for cytoplasmic male sterility (CMS) were identified, the nuclear *Rf* gene has not been cloned. *Rf*-linked molecular markers which have been developed by recent researches were not applicable to broad range of pepper germplasm implying that extensive DNA rearrangements might occur around *Rf* gene during evolution. Therefore, cloning of *Rf* gene is required to develop reliable markers for breeding as well as to understand the mechanism of interaction of nuclear *Rf* and mitochondrial CMS genes. In this study, we designed three strategies to perform fine mapping for isolation of pepper *Rf* gene. Firstly, pepper BAC clones which contain sequences

homologous to petunia *Rf* gene were screened and mapped based on the assumption that pepper *Rf* gene may be one of the homologous pentatricopeptide repeat (PPR) genes as other *Rf* genes cloned in petunia, radish and rice. Secondly, AFLP analysis was performed using more than one thousand primer combinations. Finally, comparative mapping was conducted using tomato genome sequence. As the result, a group of selected BAC clones and AFLP markers were mapped on pepper DNA region that was co-segregated with *Rf* gene and corresponded to 24.7kb-long sequence on upper region of tomato chromosome 6. By six times of chromosome walking started from the co-segregating marker, the DNA sequence which spanned *Rf*-containing DNA region was obtained. Prediction of expressed genes in this seuquence using transcriptome analysis screened an *Rf*-candidate gene, *PPR6*. The *PPR6* gene encoded a pentatricopeptide repeat protein in which degenerative 35 amino acid motif was repeated for fourteen times. Specific expression of this gene in restorer lines strengthened the hypothesis that *PPR6* is a strong candidate for *Rf* in pepper.

INTRODUCTION

Restorer-of-fertility (*Rf*) is a nucleus-encoded gene which suppresses the induction of cytoplasmic male sterility (CMS) caused by CMS-associated genes located on mitochondrial genome. Cloning or characterization of *Rf* genes have been attempted in many crop species because of its economical value in efficient hybrid seed production as well as scientific importance as a model for the analysis of nuclear control of mitochondrial gene expression or co-evolution of nuclear genome with mitochondrial DNA (Hanson and Bentolila, 2004). As the result, *Rf* genes have been successfully cloned in five crop species including maize (Cui et al, 1996), petunia (Bentolila et al., 2002), rice (Komori et al., 2004), radish (Brown et al., 2003; Desloire et al., 2003; Koizuka et al., 2003), sugar beet (Matsuhira et al., 2012). Most of the cloned *Rf* genes were the members of pentatricopeptide repeat (PPR) gene family. In addition, association of PPR genes and *Rf* loci was proved genetically in several other species including sorghum, *Mimulus*, and Maize (CMS-S) (Klein et al., 2005; Barr and Fishman, 2010; Xu et al., 2010). However, non-PPR type *Rf* genes were also cloned which encoded an aldehyde dehydrogenase (Rf2a), a glycine-rich protein (Rf17), a putative retrograde signaling control-related protein (Rf2), and a putative mitochondrial protein quality control-related protein (Rf1) in CMS-T maize, CW-CMS rice, LD-

CMS rice, and sugar beet, respectively (Cui et al., 1996, Fujii and Toriyama, 2009, Itabashi et al., 2011, Matsuhira et al., 2012).

PPR genes encode proteins which have a repeated motif composed of degenerative array of 35 of amino acids motif or slightly different length of motifs which can bind RNA through its superhelix structure (Small and Peeters, 2000). They were shown to be involved in RNA editing, splicing, processing and degradation according to their subclassifed structure which includes P subfamily and PLS subfamily (Lurin et al., 2004). All of PPR-type *Rf* genes were included in P subfamily that is characterized as array of regular 35 amino acid motifs (Bentolila et al., 2002; Brown et al., 2003; Komori et al., 2004). The products of these type of *Rf* genes were known to be involved in processing or degradation of transcripts of CMS genes as exemplified in *Rf1* and *Rf2* of CMS rice with Boro II cytoplasm which encode proteins cleaving or degrading mRNA (*atp6-orf79*) of the CMS-associated gene (Wang et al., 2006).

PPR genes constitute large gene family only in land plants (e.g. 441 in Arabidopsis, 477 in rice) although only a few are detected in other species including yeasts and algae (Fujii and Small, 2011). Cloned *Rf* genes shared several characteristics among PPR genes. First of all, they form a cluster with closely located PPR genes while other PPR genes are dispersed on entire genome sequences (Lurin et al., 2004). For example, in rice, nine PPR genes including two restorer genes are clustered in ~150 kb-long region on chromosome 10 (Wang et

al., 2006). In addition, *Rf* and clustered PPR genes show high sequence similarity. Fujii et al. (2011) classified these genes as *Rf*-like genes (*RFL*) based on phylogenetic analysis using known PPR genes from broad range of plant species. *RFL*s from diverse plant species formed a clade separated from clades containing other PPR genes implying that *RFL*s have originated from the same ancestral gene which had existed before the speciation of land plants. Finally, *RFL*s show much higher rate of nonsynonymous to synonymous substitutions than other PPR genes (Fujii et al., 2011). The rate of diversifying selection was shown to be the highest on the first, third and sixth amino acid of RFL protein, which may be involved in the interaction with RNA ligand. Recent study showed that each 35 amino acid array of PPR genes determined the specificity of the protein to one nucleotide of target RNA (Barkan et al., 2012). Combination of the first and sixth amino acids in PPR protein was crucial for the determination of specificity implying direct interaction of these amino acids with RNA (Barkan et al., 2012). Altogether, these characteristics of *Rf* genes supports the hypothesis that *Rf* genes has been evolved by birth-and-death process which is usually found in disease resistance genes to cope with the appearance of new CMS genes (Touzet and Budar, 2004).

In pepper, a CMS cytoplasm originated from an Indian accession (USDA accession PI 164835) and the *Rf* gene corresponds to this cytoplasm have been used for hybrid seed production (Peterson, 1958). Although one chimearic gene (*orf507*) and one pseudo gene (*Ψatp6-2*) were cloned as strong candidates of

CMS-associated gene and characterized (Kim et al., 2007), only development of molecular markers have been performed for *Rf* gene. These molecular markers include OPP13, AFRF8CAPS, CRF and AFRF4 (Kim, 2005; Kim et al., 2006; Lee et al., 2008; Gulyas et al., 2006; Min, 2009). Most of markers were developed by RAPD or AFLP methods using random primers.

Although several markers were highly closely linked to *Rf* gene (e.g. OPP13: 0.6cM, AFRF4: 0.1cM; Min, 2009) in a given population, application of these markers to pepper lines showed several limitations. Most of the markers could not detect genotypes correctly for wide range of pepper lines implying that intensive rearrangements have been occurred around *Rf* gene. Discrimination of lines with stable restorer from those with unstable restorers was neither successful using these markers. Application of cytoplasmic genic male sterility (CGMS) system is hampered in many pepper lines including sweet peppers because of presence of partial restoration in which anthers shed small amount of pollen compared to fully fertile anthers or unstable sterility which is affected by environmental conditions (Shifriss, 1997). Genetics of these phenotypes, especially the relationship or linkage with *Rf* genes, have been largely controversial. None of the previously developed markers were useful for this analysis because of incredibility in genotyping of *Rf* gene. To overcome these limitations, cloning of *Rf* gene itself or development of markers which are highly close to *Rf* gene showing high linkage disequilibrium with *Rf* gene are required.

In this study, cloning of the candidate for *Rf* gene by combinational mapping strategies were performed to facilitate the reliable and efficient molecular breeding for *Rf* gene as well as to provide the basis for understanding the interaction between *Rf* and CMS cytoplasm in pepper.

MATERIALS AND METHODS

Plant materials

Two different F_2 populations that segregated for the *Rf* gene were used for molecular marker linkage analysis. A total of 1,155 F_2 plants derived from a cross between a seed parent (S/*rfrf*) and a pollen parent (N/*rfrf*) of Chungyang commercial cultivar (Monsanto Korea, Chungju, Korea) was used for fine mapping. Another F_2 population developed from a cross in which 'TCMS' (S/*rfrf*) was the seed parent and 'MilyangK' (N or S/*RfRf*) was the pollen parent. This population consists of 160 F_2 plants and used for AFLP analysis. A population (AC99) consists of 92 F_2 plants was developed from a interspecific cross between RNaky (*C.annuum*) and CA4 (*C.chinense*) by a previous study (Livingstone et al., 1999) and was used to determine the location of the markers relative to pepper linkage groups. Leaf samples of a total of 51 breeding lines provided by Monsanto Korea and 50 lines from Enza Zaden (Enkhuizen, The Netherlands) were used to test marker applicability. For transcriptome analysis, anthers of Bukang seed parent (Bukang A; a CMS line) and pollen parent (Bukang C; a restorer line) were used.

Test crosses

To determine the *Rf* genotype of CM334, two kinds of crosses which were

CM334 X Miyang A (a CMS line) and CM334 X Bukang A were performed. The fertility was investigated in five F_1 plants of each cross.

Phenotyping scoring

A total of 1,155 F_2 individuals were grown in the greenhouse at the Seoul National University farm for four months, and the fertility of plants was evaluated at least three times by careful inspection for pollen on anthers. For breeding lines, phenotypes were scored by Monsanto Korea and Enza Zaden.

Identification and sequence analysis of *Rf* homologs

Pepper EST sequences that had high levels of homology to the petunia *Rf* gene were obtained from a pepper EST database at the Plant Diversity Research Center (PDRC) in The Korea Research Institute of Bioscience and Biotechnology (KRIBB). A full-length clone of the EST (ID : KS26037G12) that had the highest nucleotide and protein sequence similarity to petunia *Rf* was obtained using a SMART^{TM} RACE cDNA Amplification kit (Clontech, USA) with cDNA from the Chungyang cultivar. This cDNA sequence are reported in the GenBank nucleotide sequence database (accession number: GQ365708.1) Subcellular localization of PePPR1 was predicted by the iPSORT program (Bannai et al., 2002). Analysis by MEME software (http://meme.sdsc.edu/meme/ meme.html) was carried out to identify PPR motifs in the protein sequence of the cloned gene according to

Bentolila et al., (2002), and sequence alignment of PePPR1 and *Rf* genes from other plants was performed by ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/). The Basic Local Alignment Search Tool (BLAST) was used to find homologous PPR proteins and investigate similarities between protein sequences. The information on sequences of Arabidopsis PPR proteins and localization of PPR protein-encoding genes in Arabidopsis genome were obtained from The Arabidopsis Information Resource (TAIR) (www.arabidopsis.org). The sequences of the protein region containing PPR motifs in each PPR proteins were predicted by MEME software. Phylogenic analysis was performed for PPR motif-containing region sequences by a neighbor-joining algorithm using the Mega 3.1 program (The Biodesign Institute, Tempe, USA).

BAC library screening and grouping of BAC clones

A PCR product of 301bp from the PePPR3 clone was generated from an EST (ID: KS11010C08) and used as a probe for BAC library screening. The BAC library was screened using the procedure described in Yoo et al. (2003). Ten BAC clones were randomly selected from the 74 positives, and end sequences were obtained for each clone. From a total of 20 end sequences, three containing putative repeated sequences were discarded, and the remained 17 were used to classify the BAC clones. Primers sets were designed to amplify each of the 17 end sequences. BAC clones were subjected to PCR with the end sequence primers,

and grouped by the primer sets that gave a product.

Development of molecular markers using HRM analysis and AS-PCR

High resolution melting (HRM) analysis was used to develop molecular markers from BAC end sequences in each contig group. For HRM, PCR reactions were done in 20μl with 50 ng of template DNA, 2 μl of 10X *Taq* DNA polymerase buffer, 5 pmol of each primer, 200 μM dNTPs, 1.25 μM of SYTO 9 dye and 1 unit of DNA *Taq* DNA polymerase. Using a Rotor-gene^{TM} 6000 (Corbette, Australia), real-time PCR amplification (94^{o}C 10 min followed by 50 cycles of 94^{o}C 20 sec, 53^{o}C 20 sec and 72^{o}C 30 sec) and HRM analysis (increasing 0.1^{o}C every 1 minute from 72 ^{o}C to 90^{o}C) were performed. Fluorescence was graphed versus temperature with the highest point set to 100. To map PePPR1 in the Chungyang F_2 population, the original HRM marker was converted into an allele-specific PCR (AS-PCR) marker, by designing a forward PCR primer at a polymorphic difference between the parental alleles at the 3'distal end of primer sequence. AS-PCR amplifications using 50 ng of DNA were carried out as follows: 94^{o}C for 5 min; 35 cycles of 94^{o}C for 30 sec, 51 ^{o}C for 30 sec and 72 ^{o}C for 1 min; followed by a final extension of 72^{o}C for 10 min.

Linkage analysis

Linkage analysis was performed using 89 AC99 plants. Linkage analysis

of developed markers was performed using CarthaGene software (Givry et al., 2005) with a LOD-score threshold of 4.0 and a maximum distance of 30 cM.

BSA-AFLP

Three rounds of BSA-AFLP were performed to select the candidates of 0 cM markers. Firstly, each ten individuals from TCMS X MilyangK F2 population were selected to compose the pools for *RfRf* and *rfrf*, respectively. The genotype of F_2 individuals were determined by *Rf* phenotype and marker genotypes for G05G1 and OPP13CAPS. Only fertile plants showing the same genotype with that of MilyangK for G05G1 and OPP13CAPS were selected as the plants having *RfRf* genotype. Among the individual plants with *RfRf* and *rfrf*, each ten plants were selected randomly for the pools for *RfRf* and *rfrf*, respectively. AFLP analysis were performed for all of the primer combinations (1,024 combinations) which can be driven from primers containing two more selective nucleotide attached to pre-amplification primers. Secondly, BSA-AFLP was performed using *RfRf* and *rfrf* pools composed by each ten individuals which were not used for the first round of BSA-AFLP. BSA-AFLP was performed only for the primer combinations which showed clear polymorphism in the first round of BSA-AFLP. Finally, BSA-AFLP was performed using DNA pools composed by recombinants between four markers and *Rf* phenotype. To select recombinants, AFLPs were performed for all of the F_2 individuals using eight markers which were randomly

selected among the primer combinations selected in second round of BSA-AFLP. By comparison with the genotypes for OPP13CAPS and G05G1, the directions of markers from *Rf* were determined in case which application of the marker resulted in at least one recombinants. For the markers in which *Rf* was scored by the presence of DNA band, the sterile individuals in which DNA band was detected were determined as recombinants. Selected recombinants were classified as two groups according to the direction of markers from *Rf* gene. For the markers in which *rf* is scored by the presence of DNA band, the *RfRf*-genotyped individuals in which DNA band is detected were determined as recombinants. Selected recombinants were also classified as two groups according to the direction of markers from which the recombinants originated. In this way, all the collected recombinants were classified as four groups according to direction and target allele of the markers for which the recombinants originated. DNAs for recombinants in the same group were bulked together. BSA-AFLP was performed for four bulks using the primer combinations which were selected in second round of BSA-AFLP.

In all round of BSA-AFLP, AFLP was performed with labeled EcoRI primers and normal MseI primers following the protocol developed by KeyGene (Wageningen, The Nethelands) company.

AFLP for F_2 individuals

Only the markers selected in three round of BSA-AFLP were applied to all of TCMS X MilyangK F_2 individuals. PCR condition for AFLP and gel running procedure were the same with that used in BSA-AFLP. AFLP results were analyzed by Xtractor and data file editor programs developed by KeyGene (Wageningen, The Nethelands) company.

Sequencing of amplicons of markers

The gel fragments containing amplicons for each selected markers were picked with pipette tips and amplicons were eluted in PCR buffer. The amplicons were re-amplified and sequence of PCR products were determined by Sanger sequencing.

Marker development based on tomato gene sequences

Thirteen pepper ESTs which correspond to thirteen tomato genes located on 1.3Mb ~ 2.2Mb region of tomato chromosome 6, respectively, were selected. The putative positions of introns for pepper ESTs were predicted by the comparison with the DNA sequences of tomato genes. Primers were designed from EST sequences to amplify predicted intron sequences. PCRs were performed using designed primers for Chungyang A and Chungyang B. The primer combinations which generated amplicons only in one of two lines were regarded

as sequence characterized amplified region (SCAR) markers. When the amplification occurred in both lines, sequences of amplicons were compared to search for polymorphisms for cleaved amplified polymorphic sequence (CAPS) marker development. Designed SCAR or CAPS markers were applied to ten recombinants of Chungyang F_2 population

Screening of pepper scaffold sequences containing sequences of developed markers

Contigs or scaffolds for pepper full genome sequences assembled during pepper genome project were kindly provided from the Horticultural Crop Genomics Lab (Seoul National University, Seoul, Republic of Korea). The contigs or scaffolds containing marker sequences were selected by BLAST search in a website (http://cab.pepper.snu.ac.kr) developed by Horticultural Crop Genomics Lab. The 0.83 version contigs and 0.9 version scaffolds were used for development of new markers and anchoring of markers, respectively. The 1.1 version scaffolds were used in gap filling for contig sequences generated by sequencing of selected BAC clones.

Development of strategic pools for PCR-based BAC screening

Strategic bulking of CM334 BAC library was performed for efficient screening of BAC clones based on PCR analysis. The BAC clone cell cultures

located on the same columns of a total of 576 384-well plates in BAC library was bulked to generate the ‘1D column bulks’. These bulks in a plate were pooled again to form the ‘plate bulk’. Finally, two dimensional bulking to make ‘2D plate bulks’ was performed for each 96 ‘plate bulks’. Therefore, final set of BAC library pools consisted of six sets of the ‘2D plate bulks’, 576 tubes of the ‘plate bulks’, and 576 sets of ‘1D column bulk’. These strategic pools enabled the selection of a BAC clone among 221,184 clones by three rounds of PCRs which consists a total of 160 PCR reactions if the target clone exists as a single copy in BAC library.

Sequencing of selected BAC clones

Sequencing of the selected BAC clones was performed by GS-FLX system (Roche Applied Science, Indianapolis, USA). Each BAC clones were labeled differently during library construction for separated assembly. Generated sequences were assembled by Newbler Assembler Software Version 2.0 (454 Life Sciences, Branford, USA) in National Instrumentation Center for Environmental Management (Seoul, Republic of Korea)

Transcriptome analysis

RNAs were isolated from anther tissues of a restorer line, ‘Bukang C’ and a CMS line, ‘Bukang A’ using Hybrid-RTM RNA extraction kit (GeneAll

biotechnology, Seoul, Republic of Korea) according to manufacturer's description. Transcriptome sequences were produced by Illumina DNA sequenicing system (Illumina, San Diego, USA) in National Instrumentation Center for Environmental Management (NICEM, Seoul, Republic of Korea). The sequence reads were assembled by CLC workbench 5 (CLC bio, Aarhus, Denmark). Alignment of short reads on PPR6 gene was visualized using Tablet program (http://bioinf. scri.ac.uk/tablet/).

RT-PCR analysis

Total RNA was isolated from stems, leaves, ovules, and anthers (obtained from floral buds which are 3-5mm in size) from an individual with *RfRf* genotype and an individual with *rfrf* genotype using Hybrid-R^{TM} RNA extraction kit (GeneAll biotechnology, Seoul, Republic of Korea) according to manufacturer's description. cDNA was synthesized from 2 μg of total RNA using MMLV reverse transcription kit (Promega, WI, USA). The reverse-transcriptase PCR (RT-PCR) was performed using PPR6 specific primer set in 50 μl with 10 mM Tris-HCl (pH8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM of each dNTP, 10 pmole of each primer, 1 μl of reverse transcription products, and one unit of rTaq polymerase (Takara, Shiga, Japan).

RESULTS

Marker analysis and generation of segregants for new marker development

We tested the markers OPP13-CAPS (Kim et al., 2005), AFRF8-CAPS (Kim et al., 2006) and PR-CAPS (Lee et al., 2008), which were previously shown to be closely linked to the *Rf* gene, and found they did not show polymorphisms in the parents of the Chungyang cultivar, which is one of the most popular F_1 hybrids in Korea. Marker genotype which had been associated with *rf* allele in other research (Kim, 2006; Lee et al., 2008) were detected in both parents of Chungyang (data not shown). Analysis of sequences of OPP13-CAPS, AFRF8-CAPS and PR-CAPS in Chungyang parental lines also showed that there are no polymorphisms between two lines on these sequences. To develop new markers, an F_2 population was developed by self-pollination of Chungyang. Fertility of individual F_2 plants was evaluated by the presence or absence of pollen on anthers, and by anther size. Plant phenotypes could be unambiguously classified as either fertile or sterile because fertile plants had big anthers and abundant pollen while sterile plants had small anthers and no pollen. From 1,155 individuals, 855 were fertile and 300 were sterile. A goodness of fit $\chi 2$ test for a 3:1 segregation model with P values was determined (Table 1). The data confirmed that one dominant gene controls restoration of fertility in the Chungyang population.

Table 1. Segregation of fertile and sterile phenotypes in the Chungyang F_2 population.

No. individuals			Expected ratio	χ^2	Probability
Total	Fertile	Sterile			
1,155	855	300	3:1	0.584	0.445

Phylogenic analysis of petunia *Rf* gene homologs from pepper

Pepper sequences with a high predicted amino acid similarity to the petunia *Rf* gene were found by screening 122,503 pepper ESTs. The five with the highest similarity were selected as candidates (Table 2). The EST (ID: KS26037G12) with the highest nucleotide (78%) and protein (65% identity/76% similarity) similarity was named PePPR1 and used for further sequence analysis. The complete sequence of *PePPR1* was obtained by 5' and 3' RACE and it was found to be 1,734 bp with no introns. The predicted protein sequence contained an array of PPR motifs flanked by N- and C- terminal sequences. After the signal peptide, 14 degenerative repeats of 35 amino acids were found (Fig. 1a). The successive array of motifs indicated that PePPR1 is in the P subfamily of PPR proteins, according to the classification of Lurin et al. (2004). Like the petunia Rf protein, the array of PPR motifs in PePPR1 could be divided into two parts with one intervening amino acid: two motifs from amino acid 75 to 144, followed by

twelve motifs from amino acid 146 to 565.

Alignment of PePPR1 with three Rf proteins from other plants indicated the PPR-containing region was more highly conserved than the N- or C- terminal sequences (Fig. 1b). When the four different sequences were aligned, PePPR1 and petunia Rf were clearly distinguished from the others. PePPR1 and petunia Rf had deletions at six sites throughout the protein that were absent in the Rf protein sequences of rice or radish. PePPR1 also contained additional deletions of seven and three amino acids at two positions in the region N-terminal to the PPR motifs, and a five-amino acid deletion in the C terminus that were absent in petunia Rf.

A BLAST search for PePPR1 and the three other Rf proteins in Arabidopsis found Arabidopsis genes that code for proteins with high similarity to the four proteins is located on the 23 Mb, and 4.1-4.3 Mb regions of chromosome one. PePPR1 showed 32~39% amino acid identities and 56~63% similarities with the *Arabidopsis* PPR proteins encoded in 23 Mb and 4.1-4.3 Mb regions. When PePPR1 was compared to ten of P-subfamily PPR genes randomly selected from other locations on chromosome one or from other chromosomes, however, the identity and similarity decreased to much lower level (19~28% identity, 39~51% similarity). Dense localization of *Rf* homologs in a specific region of the Arabidopsis genome was also evident in phylogenic analysis (Fig. 1c). All predicted Arabidopsis proteins in the 23 Mb and 4.1-4.3 Mb regions were evolutionarily closer to PePPR1 and Rf proteins than other randomly selected

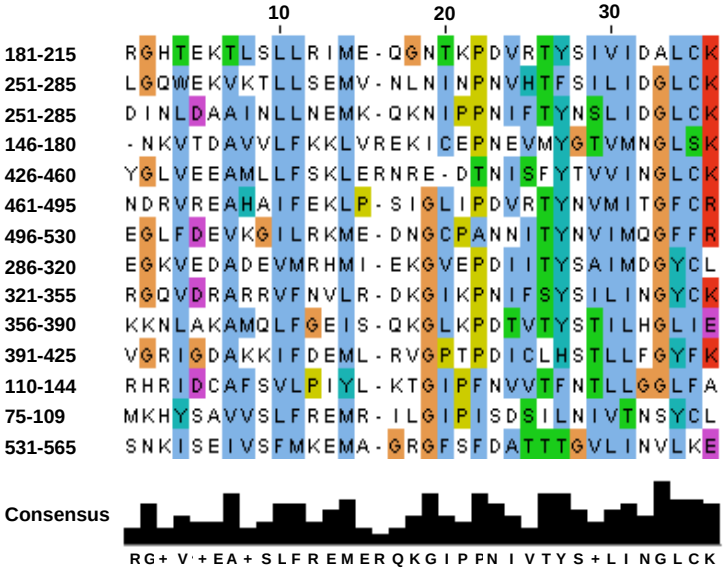
Arabidopsis genes for P-subfamily PPR proteins.

Table 2. Pepper homologs to the petunia *Rf* gene.

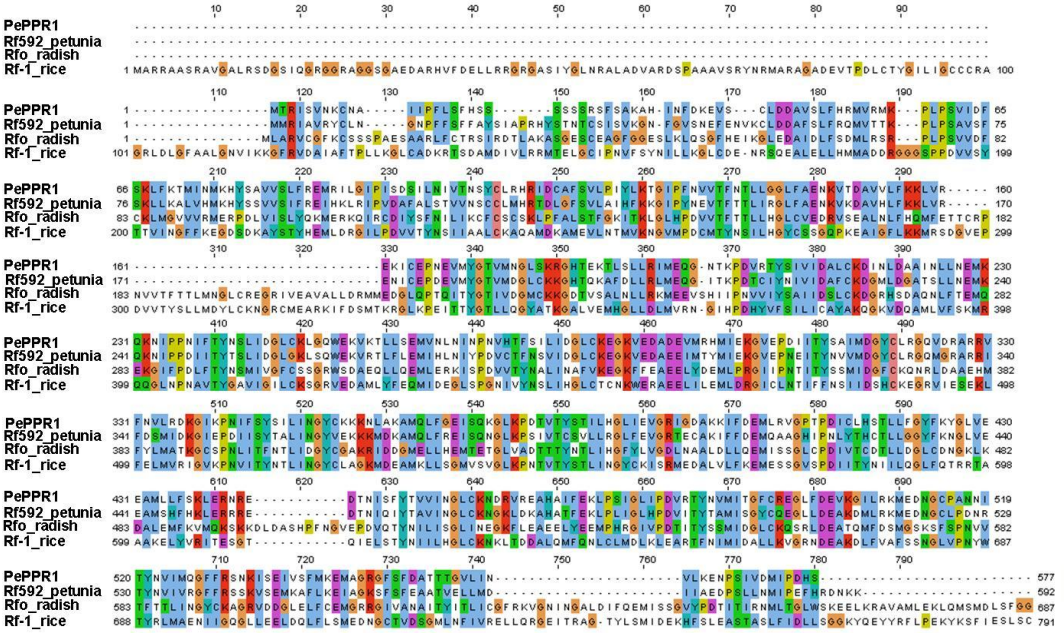
EST name	EST ID	Length (bp)	Amino acid similarity to petunia *Rf* (% identity/% similarity)	Mapping Group*
PePPR1	KS26037G12	609	65 / 76	Group 3
PePPR2	KS12062F04	586	64 / 76	Group 2
PePPR3	KS11010C08	374	59 / 76	Group 3
PePPR4	KS12009D04	685	37 / 57	Not mapped
PePPR5	KS26046E06	684	33 / 51	Not mapped

* The ‘mapping group’ refers to BAC contig groups which are co-segregated with pepper homologs to the petunia *Rf* gene.

(a)


10
20
30
181-215 RGHTEKTLSLLRIME-QGNTKPDVRTYSIVIDALCK
251-285 LGQWEKVKTLLSEMV-NLNINPNVHTFSILIDGLCK
251-285 DINLDAAINLLNEMK-QKNIPPNIFTYNSLIDGLCK
146-180 -NKVTDAVVLFKKLVREKICEPNEVMYGTVMNGLSK
426-460 YGLVEEAMLLFSKLERNRE-DTNISFYTVVINGLCK
461-495 NDRVREAHAIFEKLP-SIGLIPDVRTYNVMITGFCR
496-530 EGLFDEVKGILRKME-DNGCPANNITYNVIMQGFFR
286-320 EGKVEDADEVMRHMI-EKGVEPDIITYSAIMDGYCL
321-355 RGQVDRARRVFNVLR-DKGIKPNIFSYSILINGYCK
356-390 KKNLAKAMQLFGEIS-QKGLKPDTVTYSTILHGLIE
391-425 VGRIGDAKKIFDEML-RVGPTPDICLHSTLLFGYFK
110-144 RHRIDCAFSVLPIYL-KTGIPFNVVTFNTLLGGLFA
75-109 MKHYSAVVSLFREMR-ILGIPISDSILNIVTNSYCL
531-565 SNKISEIVSFMKEMA-GRGFSFDATTTGVLINVLKE
Consensus
RG+ V'+EA+SLFREMERQKGIPPNIVTYS+LINGLCK

(b)


PePPR1
Rf592_petunia
Rfo_radish
Rf-1_rice

(c)

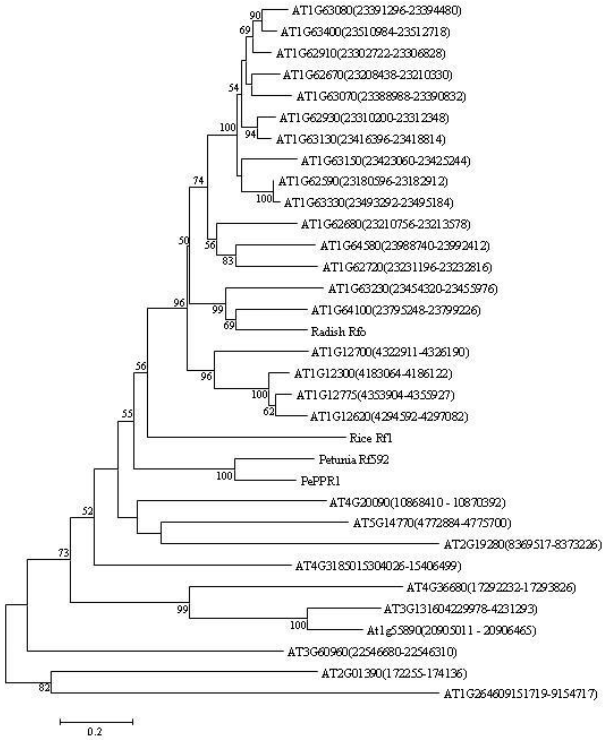


Fig. 1. Analysis of PePPR1 protein sequence. (a) Alignment of 14 PPR motifs present in PePPR1. Conserved amino acids are depicted by colors. (b) Alignment of the PePPR1 protein sequence with three Rf proteins from other crop species. (c) Phylogenic tree of PePPR1, three Rf proteins and 28 PPR proteins in Arabidopsis. The number followed by ‘at’ in the *Arabidopsis* gene IDs is the chromosome number on which the gene is located. The exact locations of genes are to the right of the gene IDs

BAC library screening and classification

To isolate BAC clones with pepper homologs of the petunia *Rf* gene, a BAC library was screened with a probe from internal sequences of *PePPR3*, which had an intermediate homology to the petunia *Rf*. A total of 74 BAC clones were obtained in the initial screening. To confirm if these BAC clones contained the *PePPR3* sequence, a systematic PCR analysis was performed using *PePPR3* primer sequences. PCR products were obtained from most of the BAC clones. Internal sequences were varied, however, and contained either *PePPR3* sequence or other *Rf* homologs. To group the BAC clones, 17 primer sets were designed from the end sequences of 10 randomly selected BAC clones. PCR analysis distributed a total of 52 BAC clones into three groups (Table 3). Twenty-two BAC clones from which PCR products were not amplified remained ungrouped.

Table 3. Classification and marker development using BAC clones selected with *PePPR3* probe.

Classification	Number of end sequences used to group clones	Number of clones in each group	Markers used for AC99	Markers used for Chungyang F_2
Group 1	6	22	BAC2T7	BAC13T7 SCAR
Group 2	7	19	BAC15SP6	BAC17T7 HRM
Group 3	4	11	BAC54SP6	PePPR1 AS-PCR

Screening of tomato BAC sequence containing petunia *Rf* homologs

Available tomato BAC sequences were screened using petunia *Rf* sequence as query. Only one BAC clone was detected to contain genes homologous to petunia *Rf* gene. A total of three PPR genes were detected among expected 15 genes in the selected sequence. All of three PPR genes showed more than 77% similarity with petunia *Rf* gene in nucleotide sequence. A molecular marker named G05G1 was developed from the pepper EST sequence that is homologous to sequence of the first gene in the BAC clone. The sequences of PPR genes were excluded in marker development because determination of the corresponding EST of pepper to a tomato PPR gene is difficult due to redundancy in the sequences of PPR genes.

Anchoring BAC contigs and G05G1 to a linkage map

BAC end sequences from three BAC groups and a tomato BAC sequence-originated marker, G05G1 were anchored on an AC99 map (Livingstone et al., 1999, Fig. 3). All of the markers were developed as codominant markers based on HRM analysis. The BAC end sequence-derived markers (BAC2T7, BAC15SP6, BAC54SP6; Table 3) were mapped to upper region on chromosome 6. BAC54SP6 from the Group 3 BAC clones co-localized with *PePPR1* and *PePPR3*, while BAC15SP6 in Group 2 co-segregated completely with PePPR2. BAC2T7 from group 1 was located between the BAC54SP6 and BAC15SP6 markers.(Fig. 3).

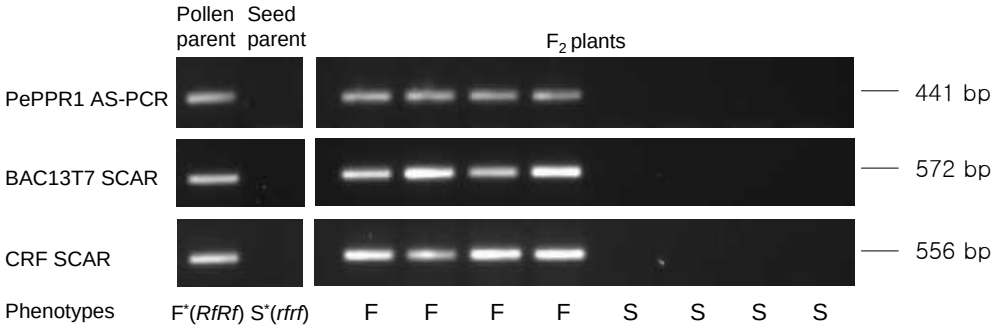
G05G1 was also localized on chromosome 6 between BAC2T7 and BAC15SP6, closest to BAC2T7 (Fig. 3)

Development of markers linked to the *Rf* gene

Two codominant and two dominant markers were developed from EST or BAC end sequences using the Chungyang F_2 population (Tables 3). A SCAR marker (BAC13T7 SCAR) was generated using the T7 end sequence of BAC clone 13 from Group 1 (Fig. 2a). PCR yielded a 572 bp product from fertile plants and no product from sterile plants. Of 244 plants tested for this marker, four exceptions were found. BAC17T7 HRM was developed using the T7 end sequence of BAC clone 17 in Group 2 (Fig. 2b). When HRM analysis was carried out for F_2 segregants, three distinguishable melting curves indicating *Rf* genotypes were obtained. A total of eight recombinants were detected for this marker. An AS-PCR marker was developed from *PePPR1* in Group 3 (Fig. 2 a). No polymorphisms were detected in the *PePPR1* EST sequence of the Chungyang parents. However, full sequencing of *PePPR1* revealed one SNP, located 54 bp downstream of the 5' end of the gene, between the Chungyang parents. A forward primer containing the polymorphic nucleotide of the Chungyang pollen parental type at the 3' end nucleotide was designed and paired with a reverse primer in the conserved sequence. As expected, PCR products were obtained only in plants with the Chungyang pollen parental genotype. A total of 32 recombinants were

detected for this marker, indicating considerable genetic distance from the *Rf* gene. Finally, a codominant HRM marker named as G05G1 was developed from a pepper EST sequence that is homologous to a sequence in tomato BAC clone containing petunia *Rf* homologs (Fig. 2b). A total of four recombinants were detected and these recombinants were identical to those detected in the application of 13T7 SCAR. This implies that G05G1 is located closely to 13T7 SCAR although some genetic distance was detected in mapping on AC99 map

(a)



(b)

Fig.2. Molecular markers applied to the Chungyang F_2 population. (a) Application of three PCR-based markers to the Chungyang parental lines and F_2 individuals. (b) Application of the HRM-based marker BAC17T7 HRM and G05G1 to F_2 plants. F and S represent fertile and sterile phenotypes, respectively.

Relative locations of *Rf*-markers

The new molecular markers developed from BAC ends or EST sequences and the previously isolated CRF-SCAR marker (Gulyas et al., 2006; Lee et al., 2008) were mapped relative to the *Rf* locus, using 244 Chungyang F_2 plants (Fig. 3). The CRF-SCAR was closest to the *Rf* gene with genetic distance of 0.4 cM, followed by BAC13T7 SCAR and G05F1 at 1.4 cM from *Rf*. The PePPR1 AS-PCR marker was farthest from the *Rf* gene at 14 cM. On the opposite side of these three markers, BAC17T7 HRM was 3.2 cM away from the *Rf* gene. Another *Rf*-segregating F_2 population developed by Kim et al. (2006) was also used to map the markers. OPP13-CAPS and CRF-SCAR segregated in this population, whereas BAC13T7 SCAR did not. When OPP13-CAPS and CRF-SCAR were mapped with other previously developed *Rf*-linked markers, CRF-SCAR was located on the opposite side of OPP13-CAPS. The genetic distance between CRF-SCAR and *Rf* was 1.4 cM in this population.

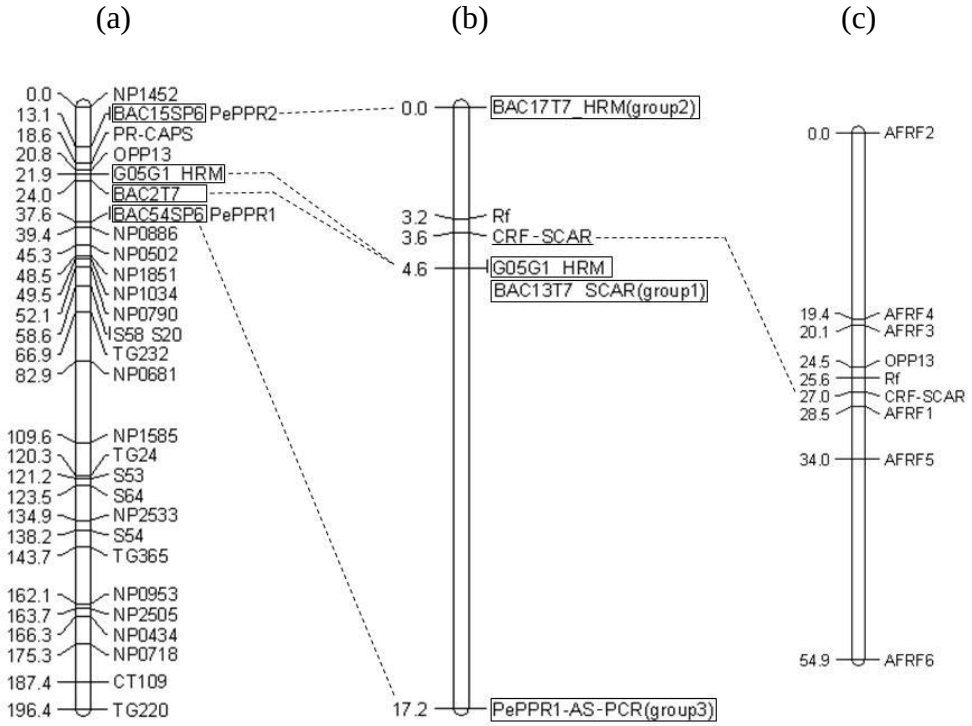


Fig.3. Linkage maps with genetic distances between markers linked to *Rf*. The three maps are for the AC99 (a), Chungyang (b), and TS502 X HK6T F_2 populations (c), respectively. The names of newly developed markers are enclosed in rectangles.

Development of AFLP markers which are closely linked to pepper *Rf* gene

Three round of BSA-AFLP was performed to select AFLP markers which are closely linked to pepper *Rf* gene from 1,024 primer combinations using an *Rf*-segregation population derived from a cross between TCMS (seed parent) and Milyang K (pollen parent). As the result of first two round of BSA-AFLP with two different set of DNA pools for *RfRf* and *rfrf*, 92 primer combinations were selected. Application of eight markers among selected primer combinations to F_2 individuals enabled detection of three markers, E38/M62, E60/M45 and E54/M51, which showed several recombinations with *Rf* phenotype. By the final round of BSA-AFLP with the pools of recombinants which were detected in the application of E38/M62, E60/M45, E54/M51 and G05G1, seventeen primer combinations were selected to detect genotypes clearly. Genotype scoring for fourteen out of seventeen primer combinations was performed for 160 F_2 individuals while three primer combinations did not result in clearly detectable amplification for individuals. In twelve among fourteen primer combinations, conflict between *Rf* phenotype and marker genotype was not detected although four of them were dominantly scored and contain unclear genotype for several individuals (Table 4). Among twelve markers, two markers were bi-allelic while others generated bands specific for *Rf* or *rf*. The eight AFLP markers for which clear genotyping result was obtained were defined as 0 cM markers of *Rf* in this population. They were classified as four groups according to the recombination pattern between markers.

The markers developed by previous researches were not included in these groups (Table 5, Fig.4).

Linkage analysis for *Rf*-linked markers

Genetic distance between molecular markers was determined based on recombination pattern and frequency between markers. Although no recombination was detected when genotypes for eight 0 cM markers were compared with dominantly scored phenotypes, several recombinations between markers could be found because co-dominant scoring was performed for most of the markers. At the borders of the linkage group of AFLP markers, E59/M59 and E51/M33 were located, respectively, and five recombinations were detected between these markers. Two groups of markers in which three markers were included, respectively, were located inside these borders. On the other hand, OPP13 and G05G1 were localized outside of the E59/M59. OPP13 and G05G1 were shown to be apart from *Rf* gene by 0.64 and 1.92 cM, respectively (Fig.6).

Table 4. Information of markers which do not show recombination with dominantly-scored phenotypes. For markers designated in shadow, genotyping could be performed dominantly or marker genotypes for several individuals were unclear. These markers did not show recombination with phenotype at least in the individuals for which genotyping result was clear.

Marker Name	Genotype which amplicon represents for	Fragment size (bp)
E32/M44	*Rf*	339 (303)[z]
E33/M36	*Rf/rf*	271 (235)[y]
		269 (233)
E41/M62	*rf*	260 (224)
E48/M38	*rf*	84 (48)
E48/M50	*rf*	443 (407)
E48/M53	*Rf*	~600 (~564)[z]
E49/M40	*rf*	107 (71)
E50/M55	*rf*	236 (200)
E51/M36	*Rf/rf*	266 (230)[y]
		264 (228)
E54/M36	*rf*	249 (213)
E59/M59	*Rf*	283 (247)
E60/M45	*Rf*	249 (213)

z) the number in parenthesis represents the size of marker-specific sequences excluding adaptor sequences

y) The size of allele for *Rf* and the allele for *rf* were determined respectively for bi-allelic markers. In all cases, longer size represent the size of allele for *Rf*.

z) The size of this fragment was not determined accurately because the sequencing for this marker was failed and the approximate size of the marker was out of size range which can be determined by size marker.

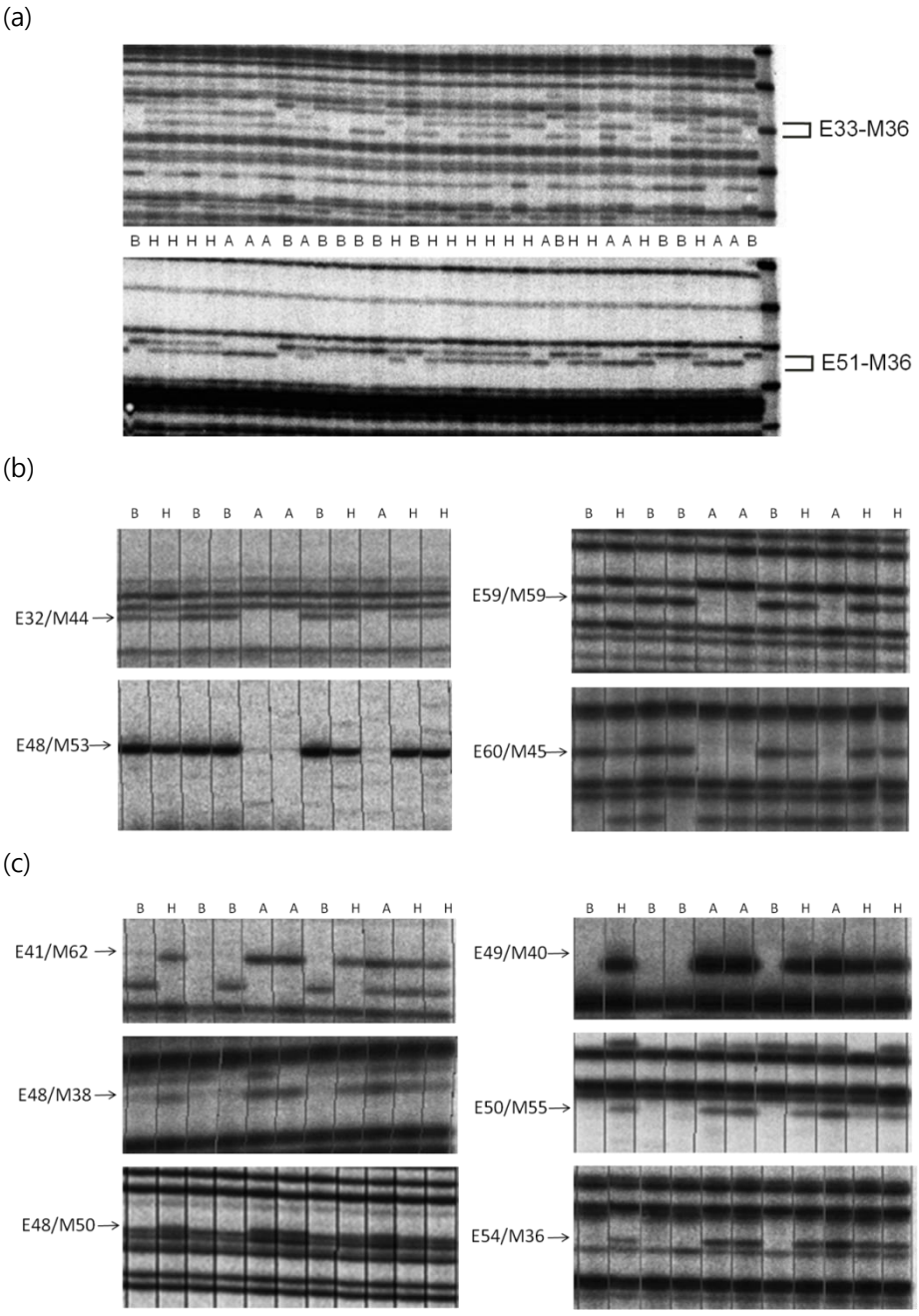


Fig.4. Markers which do not show recombination of Rf phenotypes of TCMS X MilyangK F_2 population. (a) Bi-allelic markers (b) markers in which amplicon represent *Rf* allele (c) markers in which amplicon represents *rf* allele

Table 5. Genotyping results of markers for individuals in which recombination between markers are detected. The genotypes identical to G05G1 genotypes were gray- or pink-colored. Pink color indicates the conflict between genotypes and marker genotypes. Markers classified as same groups were indicated with the same color. A: *rfrf*, B: *RfRf*, C: *RfRf* or *Rfrf*, D: *rfrf* or *Rfrf*, H: *Rfrf*

Markers	T003	T012	T013	T027	T039	T087	T088	T095	T108	T116	T147
Rf phenotype	C	C	C	C	C	A	A	C	C	A	C
G05G1	B	H	H	B	H	H	H	B	B	H	H
OPP13CAPS	B	H	H	B	H	A	H	H	B	A	B
E59/M59	B	H	B	B	H	D	A	H	B	D	B
E49/M40	B	H	B	B	B	A	A	H	D	D	B
E33/M36	B	H	B	B	B	A	A	H	H	A	B
E48/M50	B	D	B	B	B	D	D	D	D	D	B
E32/M44	H	B	B	B	B	A	A	H	H	A	U
E41/M62	H	B	B	B	B	A	A	H	H	A	B
E54/M36	D	B	B	B	B	D	D	U	H	D	B
E51/M36	H	B	B	H	B	A	A	H	C	A	B

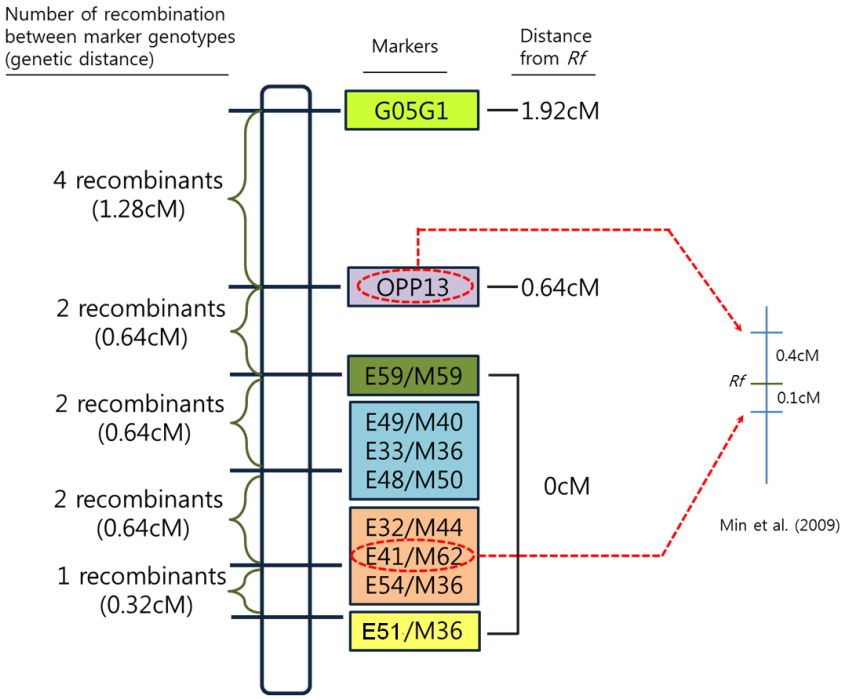
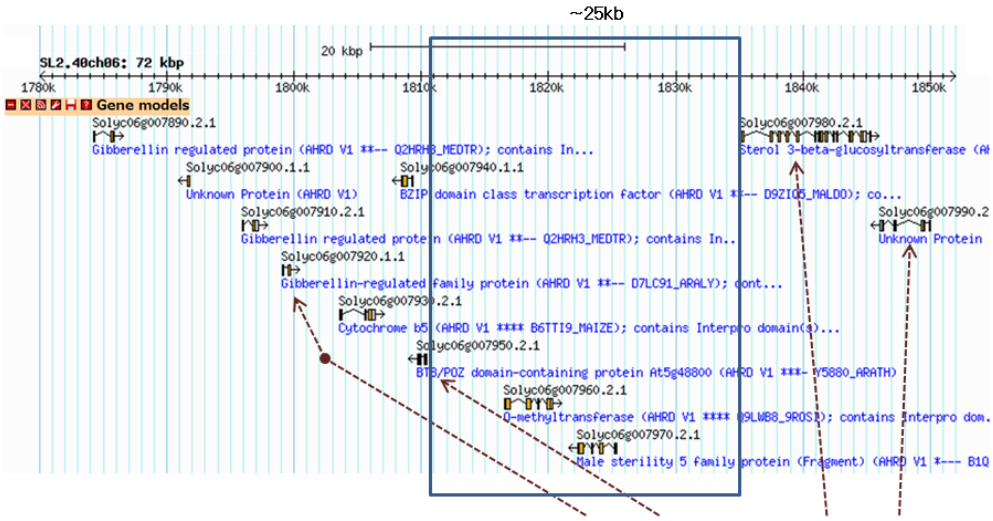


Fig. 6. Genetic map of the eight AFLP markers developed in this study and two markers previously published. The genetic distance between markers was designated on the left side of markers and genetic distance from Rf phenotypes were designated on the right side. Two markers in the circles of red dots were mapped in a previous research (Min et al., 2009).

Development of markers based on tomato sequences and application to recombinants of Chungyang F_2 population

Because G05G1 was developed based on tomato sequence, more markers could be designed from the tomato sequences around G05G1 based on synteny between tomato and pepper genome. Three EST sequences which are homologous to tomato genes located on chromosome 6 were selected to develop SCAR or CAPS markers. The first SCAR marker (1.31Mb-SCAR; Fig.7) was developed from a pepper EST (KS17062D04) which could be aligned with a tomato gene located on 1,307,093-1,310,260 bp region of tomato chromosome 6. The second CAPS marker (1.51Mb-CAPS; Fig.7) was developed from a pepper EST (KS17062D04) which was homologous to a tomato gene located on 1,501,436-1,506,288 bp region. Finally, an HRM-based codominant marker (1.85Mb-HRM) was developed from a pepper cDNA sequence (CAW15S1_Contig032733) which could be aligned with a tomato gene in 1,846,027-1,849,951 bp region. These markers were applied to Chungyang F_2 recombinants that were selected from the application of 13T7 SCAR and 17T7-HRM which were localized on the different side of *Rf* gene to a total of 244 Chungyang F_2 plants (Fig.7). The results showed that pepper *Rf* gene is located within the pepper genomic region which corresponds to 1.51-1.85 Mb region on tomato chromosome 6. Because G05G1 was shown to correspond to 1.71 Mb region of tomato, the candidate region was shrunken to be 1.71-1.85 Mb region. A total of 23 tomato gene were located in

this region. Three markers named as G16-CAPS, G20-SNP, and G23-SNP were further developed based on tomato genes located on 1,795,940-1,797,216, 1,808,475-1,809,248, and 1,835,086-1,845,250 bp regions, respectively. Application of these markers to recombinants revealed that *Rf* gene is located between G20-SNP and G23-SNP which correspond to 1.81-1.84Mb region in tomato (Fig. 7). Two tomato genes were located in this region and only the gene that encoded a homolog of male sterility 5 (*ms5*) family protein in Arabidopsis could be aligned with pepper genomic DNA sequence. However, application of two markers (ABHD1.5-SCAR, 4940-CAPS; Fig.8, Fig.9) which surrounds the pepper ortholog of tomato *ms5* resulted recombinants when it was applied to 1,068 individuals of Chungyang F_2 (Fig.7). These results implied that there might be no ortholog for pepper *Rf* gene in tomato genome and the *Rf* containing DNA region was uniquely inserted within this region in pepper.



Individual number in Chungyang F_2 population	Genotype	1.31Mb-SCAR (tomato-based)	1.51Mb-CAPS (tomato-based)	13T7 (BAC end)	G05G1 (tomato-based)	G16-CAPS (tomato-based)	G20 (tomato-based)	E33-M36 (AFLP)	G23 (tomato-based)	1.85Mb (tomato-based)	17T7 (BAC end)
827	*Rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*RfRf* or *Rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*Rfrf*	*Rfrf*	*Rfrf*	*Rfrf*	*rfrf*
687	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*RfRf* or *Rfrf*
435	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*RfRf* or *Rfrf*
998	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*RfRf* or *Rfrf*
887	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*Rfrf*	*RfRf* or *Rfrf*
830	*Rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*RfRf* or *Rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*Rfrf*	*Rfrf*	*rfrf*	*rfrf*	*rfrf*
520	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*Rfrf*	*Rfrf*	*RfRf* or *Rfrf*
970	*rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*RfRf* or *Rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*Rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*
378	*rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*RfRf* or *Rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*
628	*rfrf*	*RfRf* or *Rfrf*	*Rfrf*	*RfRf* or *Rfrf*	*RfRf* or *Rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*	*rfrf*

Fig. 7. The application of markers developed by strategies based on candidate gene approach, AFLP analysis, and comparative mapping with tomato to Chungyang F_2 recombinants. Red color indicates the conflict between genotypes of individuals determined by phenotyping and marker genotypes.

Anchoring of developed markers to pepper genomic DNA scaffolds and integration of mapping information

Markers developed by strategies based on candidate gene approach, AFLP analysis, and comparative mapping with tomato genome were anchored on pepper genomic DNA scaffold sequences produced during pepper genome project (Horticultural Crop Genomics Lab, Seoul National University, Republic of Korea). The sequences were from a pepper line 'CM334' which were shown to contain *Rf* gene by test cross with two CMS lines, 'Bukang A' and 'Milyang A' (see materials and methods). A total of four new markers were developed from the scaffold sequences which were localized on both side of *Rf* gene (scaffold1281, scaffold7480; Fig.8, Fig.9). A CAPS marker (4940 CAPS; Fig.8, Fig.9) was located on about 640kb region of scaffold1281 resulted in one recombinant in the application to 1,068 individuals in Chungyang F_2 population. The markers (3336-last2, 0.9-10k SCAR; Fig.8, Fig.9) developed on the end region of this scaffold were co-segregated with *Rf* gene. On the different side of *Rf* gene, a SCAR marker named 4162-2 SCAR was 0.3cM and located on the end of scaffold7480 (Fig.8, Fig.9). Therefore, *Rf* gene was thought to be located in the 0.0cM region defined by two markers which were 4940 CAPS and 4162-2 SCAR (Fig.9). The localization of markers developed by this study and previous researches on genetic maps and pepper/tomato gDNA sequences showed clearly that the newly developed markers were closer to *Rf* when they were compared with *Rf*-linked

markers which have been widely used (AFRF8 CAPS, OPP13 CAPS) and the *Rf* containing region was unique to pepper or not syntenic to tomato genome (Fig.9).

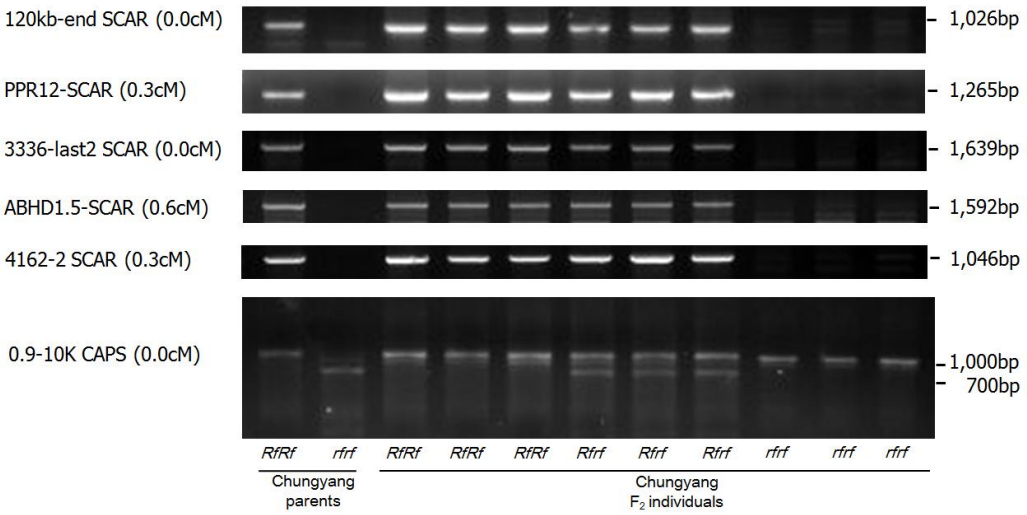


Fig 8. Performance of markers developed from pepper scaffolds closely located to *Rf* gene or the BAC clone sequence selected by the first round of genome walking.

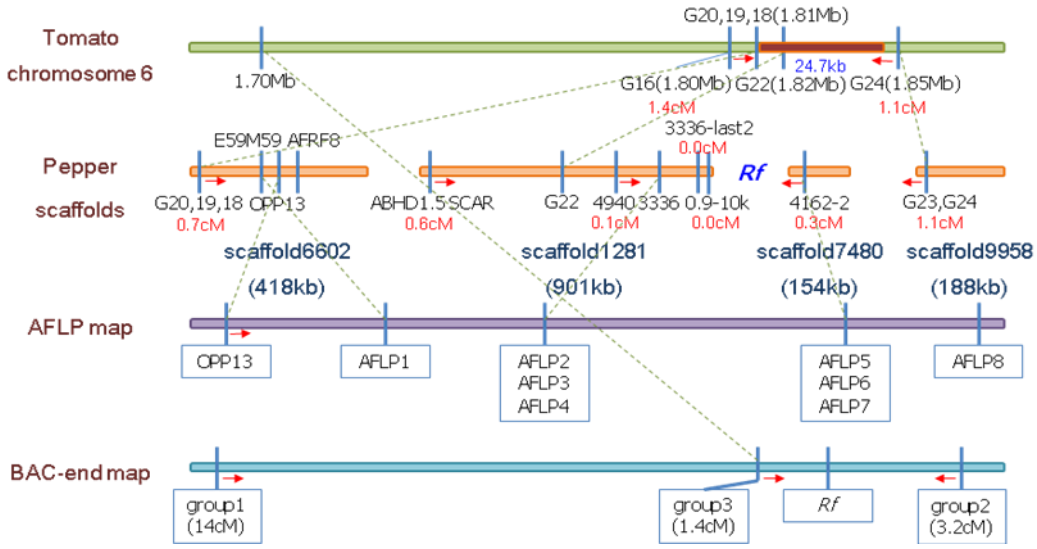


Fig 9. Integration of mapping information and comparative analysis between tomoto genome, pepper scaffolds, AFLP genetic map, and the genetic map derived by markers developed from BAC end sequences.

Chromosome walking to define DNA region which co-segregates with *Rf*

A total of six times of chromosome walkings were performed from the end of scaffold1282 (Fig.9) which co-segregated with *Rf* to obtain the full length sequence of pepper genomic DNA region which co-segregates with *Rf* in Chungyang F_2 population (1,068 individuals) (Fig.10). SNP or SCAR markers (120kb-F2R3, PPR12-SCAR; Fig.8, Fig.10) were developed from pepper whole genome contig sequences which were screened from the end sequences of each BAC clones. If the developed marker in a given BAC contig is co-segregate with pepper *Rf*, then the next round of BAC screening was performed using the most extended end of existing BAC clones. The primer sequences were designed from BAC end sequences. The BAC clone numbered as '415' contained two markers which resulted in three recombinants in Chungyang F_2 population (Fig. 10). Therefore, the region co-segregates with pepper *Rf* was defined to be from 640kb region of scaffold1281 (4940 CAPS; Fig.9) to PPR12-SCAR marker in BAC clone '415' (Fig.10). Among the BAC clones, the clones named as 'PPR5-70' and 'PPR5-11' were identical with two BAC clones selected using PePPR1 as probe (Table 3). These BAC clones were remained to be ungrouped in the grouping of selected BAC clones (Table 3). PPR 5-70 included both ends of PPR 5-11 (Fig.10).

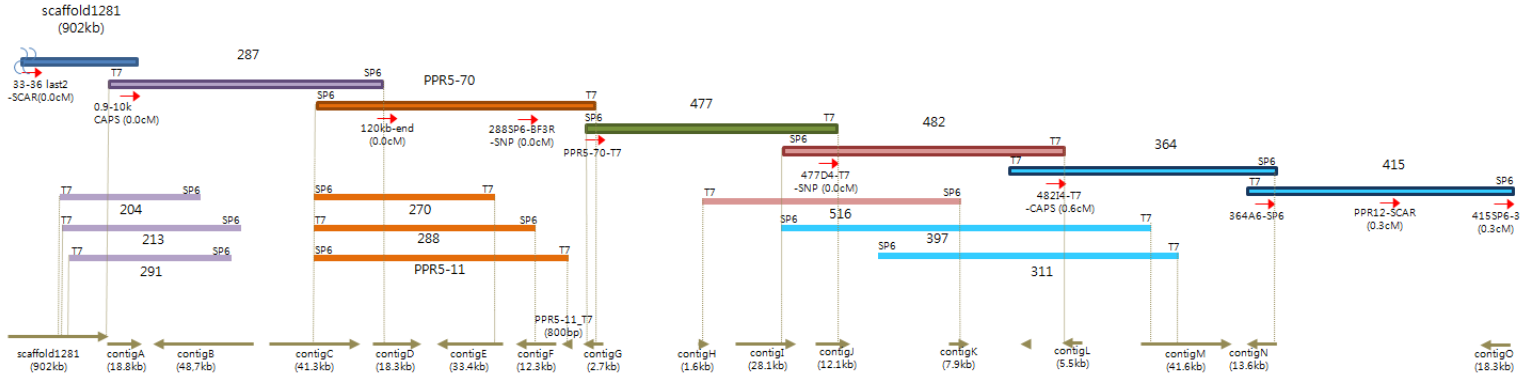


Fig.10. Schematic diagram of chromosome walking to define the DNA region which co-segregates with pepper *Rf*. The selected BAC clones were described as colored bars. The BAC clones drawn upside was used for chromosome walking and marker development and the BAC clones drawn lower side was used only for marker development. The markers developed during chromosome walking was depicted as red arrows. The pepper whole genome contigs which were selected using BAC end sequences and used for marker development were shown on the bottom of the figure as arrows.

Application of *Rf*-linked markers in breeding lines

Newly developed and previously developed *Rf*-linked markers were applied to panels of breeding lines provided by Monsanto Korea and Enza Zaden (Table 6, Table 7). In the application to 51 lines provided by Monsanto Korea, only the markers developed from scaffold1281 (Fig.9) which was the scaffold located on one side of co-segregating region showed high ratio of successful genotyping that was over 90% (Table 6). OPP13-CAPS and PR-CAPS, which were developed by previous researches (Kim et al., 2005; Lee et al., 2008) could not determine the genotypes of many lines due to presence of marker haplotypes (haplotype 3) which could not be classified as *Rf* or *rf*. Although 4940 CAPS was developed as co-dominant marker in Chungyang population, it could not generate amplicons in several lines which mainly have *rfrf* genotype. In the application to 50 lines provided by Enza Zaden, the markers developed from scaffold1281 showed much higher ratio of successful genotyping than OPP13-CAPS as in the application to the lines from Monsanto Korea (Table 7). However, the performance of the markers developed from scaffold1281 was clearly distinguished according to the relative location of markers on scaffold1281. The 3336 last2 SCAR marker, which is located on the end of scaffold1281, showed significantly higher accuracy than the markers developed from inner position on scaffold1281 (CRF-SCAR, 4940 CAPS) which was farther from *Rf* locus according to mapping results (Fig.9). The marker developed from the end sequence of the BAC clone ‘287’ (Fig.10) showed the same accuracy with 3336 last2 SCAR marker. The markers developed from the BAC clone ‘415’ or from scaffold 7480, which were located on the opposite side of *Rf* locus, showed very low accuracy in both panel of breeding lines. The 0.9-10k CAPS marker which was inside the

DNA region co-segregating with *Rf* gene could not be applied to diverse breeding lines due to incomplete restriction in many breeding lines which implies the possible duplication of the sequence of marker in those lines.

Table 6. Application of ten markers to pepper breeding lines from Monsanto

Markers	Marker haplotypes	Number of lines classified as homozygous for *Rf* (total 21 lines)	Number of lines classified as homozygous for *rf* (total 30 lines)	Ratio of successful genotyping (%)
OPP13-CAPS	OPP-haplotype 1	[z]6	-	27.5
	OPP-haplotype 2	1	8	
	OPP-haplotype 3	14	22	
PR-CAPS	PR-haplotype 1	8	9	33.3
	PR-haplotype 2	9	9	
	PR-haplotype 3	4	12	
CRF-SCAR	*RfRf* or *Rfrf*	19	3	90.2
	rfrf	2	27	
BAC13T7 SCAR	*RfRf* or *Rfrf*	5	9	49.0
	rfrf	16	20	
G05G1	[y] *RfRf* or *Rfrf*	4	4	58.8
	rfrf	17	26	
17T7 HRM	[x]17T7-haplotype 1	15	17	49.0
	17T7-haplotype 2	4	10	
	17T7-haplotype 3	2	3	
4940 CAPS	*RfRf*	19	2	92.2
	rfrf	1	17	
	No amplification	1	11[w]	
3336 last2 SCAR	*RfRf* or *Rfrf*	19	3	90.2
	rfrf	2	27	
4162-2 SCAR	*RfRf* or *Rfrf*	19	22	52.9
	rfrf	2	8	
PPR12 SCAR	*RfRf* or *Rfrf*	19	28	41.2
	rfrf	2	2	

[z] Underlines indicate the breeding lines whose phenotype was predicted correctly by each marker.

[y] Haplotypes for G05G1 designated in this table is opposite to the haplotypes in Chungyang population which are showed in Fig.2b.

[x] 17T7-haplotype 1 and 17T7-haplotype 2 refer to the haplotypes for *RfRf* and *rfrf* in Chungyang population (Fig.2b), respectively. 17T7-haplotype 3 is the haplotype which is not detected in Chungyang population.

[W] 4940 CAPS could not generate amplicons in several breeding lines. The absence of amplicon was clearly associated with *rfrf*. Therefore, absence of amplicon in 4940 CAPS was counted as *rfrf*.

Table 7. Application of eight markers to pepper breeding lines from Enza Zaden

Markers	Marker haplotypes	Number of lines classified as *RfRf* (total 24 lines)	Number of lines classified as *Rfrf* (total 12 lines)	Number of lines classified as *rfrf* (total 14 lines)	Ratio of successful genotyping (%)
OPP13-CAPS	OPP-haplotype 1	[z]6	0	1	24.0
	OPP-haplotype 2	0	0	0	
	OPP-haplotype 3	8	6	13	
	Heterotype	0	6	0	
CRF-SCAR	*RfRf* or *Rfrf*	15	8	2	70.0
	rfrf	9	4	12	
4940 CAPS[y]	*RfRf* or *Rfrf*	14	6	1	66.0
	rfrf	10	6	13	
3336 last2 SCAR	*RfRf* or *Rfrf*	21	12	1	92.0
	rfrf	3	0	13	
120kb F2R3 SCAR	*RfRf* or *Rfrf*	21	12	1	92.0
	rfrf	3	0	13	
415SP6-3 SCAR	*RfRf* or *Rfrf*	16	12	11	62.0
	rfrf	8	0	3	

[z] Underlines indicate the breeding lines whose phenotype was predicted correctly by each marker.

[y] Although 4940 CAPS was developed as co-dominant marker in Chungyang population, it was regarded as dominant marker in this application because it could not generate amplicons in many of breeding lines in this panel.

[z] Underlines indicate the breeding lines whose phenotype was predicted correctly by each marker.

Analysis of the DNA sequence which co-segregate with *Rf* gene

The sequences of six BAC clones selected during genome walking process were analyzed. The sequences in the end region of BAC clones showed perfect overlap with the connected clones as expected in genome walking (Fig.10). The assembly of sequences of six BAC clone provided a 631kb scaffold sequence which contains all of the BAC clone sequences. Using this sequence and the sequence of scaffold1281, the total sequence of DNA region which co-segregate with *Rf* gene, which was from the end of 4940 CAPS marker to end of PPR12 SCAR marker could be obtained. The length of this sequence was 818kb. The assembled transcriptome sequences from anther tissue of 'Bukang C', which is a restorer line, was aligned to the co-segregation DNA sequence to find the possible gene sequences that is expressed in anther. When the transcriptome contig sequences which were longer than 100bp and showed similarity higher than 98% with DNA sequence co-segregating with *Rf* was screened, a total of nineteen contigs were selected (Table 8). BlastX analysis result showed that the proteins coded by these genes include transposable element related proteins, NAC domain proteins, NBS-LRR proteins, P-selectin-glycoprotein, dihydroorotase, mitochondrial TOM40-1 protein, PPR protein, and proteins with unknown function. The transcribed sequences were densely located on 400-460kb region of the co-segregating sequence. When the possible *orfs* were searched for each transcribed sequence, all of the aligned sequences except for a transcript which encoded a PPR protein were shown to contain fragmented small *orf*s or no *orf*. This pattern was similar between transcriptome contig sequences and the corresponding BAC clone DNA sequences implying that the transcribed sequences except for the sequence which encodes PPR protein may not be translated to intact protein.

Table 8. List of contigs assembled from tanscriptome of anther tissue of Bukang C which can be aligned on the DNA sequence which co-segregates with *Rf* gene.

Name of contigs	Contig length	Function of gene product expected by BLASTX	Location on 0.0cM region		Coverage of contigs by 0.0cM region sequence (%)	Similarity (%)	Length of the longest *orf*	
			Start	End			on BAC clone sequence	on trancriptome contig seqeunce
contig_135691	498	NAC domain-containing protein	107390	107887	100	98.8	213	213
contig_90919	718	P-selectin-glycoprotein	168207	168412	100	100	365	362
			168619	168682		98.65		
			170818	171265		96.65		
contig_56827	499	transposable element	173685	173187	100	100	120	120
contig_133674	429	transposable element	175961	176389	100	99.77	-	-
contig_130916	637	unknown	404350	404727	97	100	228	228
			405587	405744	97	100		
			445115	445204	97	100		
contig_115973	393	Protein binding protein	405115	404723	100	100	260	266
contig_137829	479	unknown	406210	406689	100	99.38	244	222
contig_132622	666	unknown	409366	408702	100	99.85	330	337
contig_36222	738	transposable element	410038	410775	100	99.86	297	297
contig_133779	383	transposable element	412183	411801	100	100	351	357
contig_84137	584	transposable element	414086	413503	100	99.66	168	168
contig_81669	2108	NBS-LRR root knot nematode resistance	442820	441345	100	100	444	444
			440995	440901		98.11		
			440797	440645		100		
			440486	440366		100		
			440286	440024		99.27		
contig_112857	1298	NBS-LRR root knot nematode resistance	443460	444757	100	100	606	606
contig_64819	436	disease resistance RPP13-like	446306	445873	100	99.31	114	114
contig_90053	364	dihydroorotase	447536	447173	100	100	129	129
contig_110311	726	NBS-LRR root knot nematode resistance	450064	450157	100	95.74	357	357
			450634	451263		99.37		
contig_107563	986	TOM40-1	456816	457344	91	99.05	237	207
			457339	457700		98.08		
contig_116115	441	PPR1	459290	458850	100	100	441	435
contig_112011	678	transposable element	806082	805524	100	98.21	453	575

Sequence analysis of PPR gene located on DNA region which co-segregates with *Rf*

Primers were designed to amplify the complete length of gene sequence which included the transcriptome contig sequence that encodes PPR protein. The complete gene, which was 1,770bp in length could be amplified when PCR using the selected BAC clone DNA (BAC clone named as 'PPR5-70' in Fig.10) as template was performed. The gene that encoded a P-subclass PPR protein in which fourteen repeats of PPR motifs were detected (Fig.11). This gene was named as '*PPR6*'. The sequence of *PPR6* gene was not found in pepper genome database. The PPR6 protein showed 63% identity/76% similarity when aligned with petunia *Rf* and 73% identity/83% similarity with PePPR1. When the same primer set was used to amplify PPR6 in pepper lines includes Chungyang parental lines, multiple copies of genes were obtained implying presence of homologs in pepper genome. Therefore, clonings of PCR products were performed in pepper lines to isolate homologs of this gene. Four of PPR6 homolog sequences were obtained by the sequencing of clones from three lines which includes Chungyang parental lines and a CMS line provided by Enza Zaden (named as ENZA-11). Two copies of homologs were found in Chungyang restorer parent. The first copy was identical with *PPR6* in CM334 except for synonymous substitution of a nucleotide. The second copy named as *PPR6-2* encoded a protein which showed 93% identity/96% similarity when aligned with PPR6. In Chungyang CMS parent,

one different copy of *PPR6* homologs named as *PPR6-3* was detected. PPR6-3 showed 88% identity/94% similarity, 91% identity/95% similarity when aligned with PPR6 and PPR6-2. Finally, one homolog named as PPR6-4 was obtained from ENZA-11. The protein product of this homolog was identical with PPR6 except for three amino acids and the similarity between PPR6-4 and PPR6 was 100%.

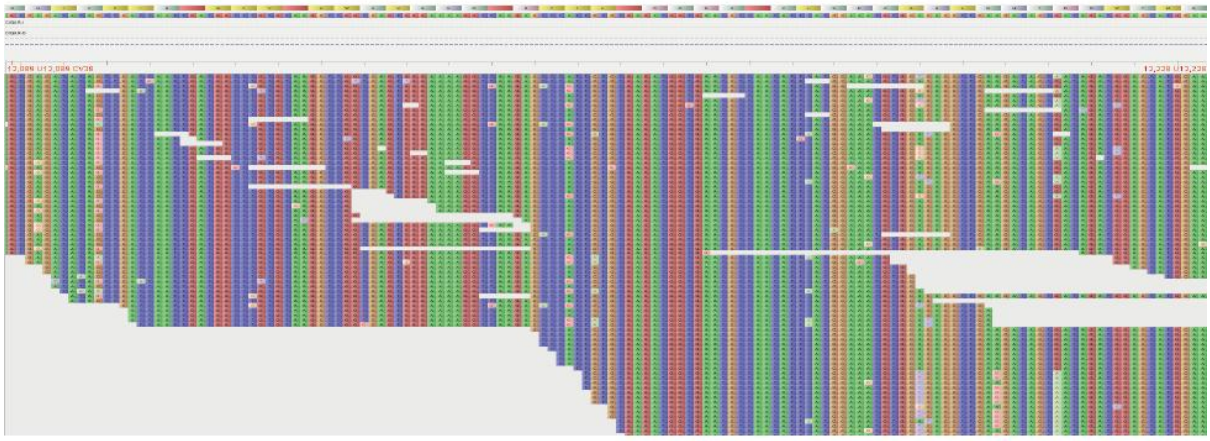
Fig.11. Alignment of protein sequences encoded by homologs of *PPR6* obtained from three different pepper lines. Repeat 1-14 indicates the predicted PPR motifs.

Expression of *PPR6* in CMS and restorer lines

The sequence reads produced from the transcriptome analysis for anthers of Bukang C (a restorer line) and Bukang A (a CMS line) were aligned to *PPR6* (Fig.12a). Although *PPR6* could be covered by sequence reads from Bukang C, only the sequence reads showing significant differences with *PPR6* sequence were aligned in Bukang A indicating that DNA sequence highly similar to PPR6 is not present or not expressed in Bukang A. To investigate the expression of *PPR6* in plant tissues, a primer set which can amplify a portion of *PPR6* gene in Chungyang and Bukang restorer parent, but cannot in Chungyang and Bukang CMS parent was designed using the polymorphism between *PPR6* and *PPR6-2* (Fig.12b). When RT-PCR was performed using this primer sets for four different plant tissues in Chungyang F_2 individuals carrying *RfRf* and *rfrf* genotypes, respectively, amplifications were detected in all of the tissues of individuals carrying *RfRf*.

(a)

Bukang C



Bukang A

(b)

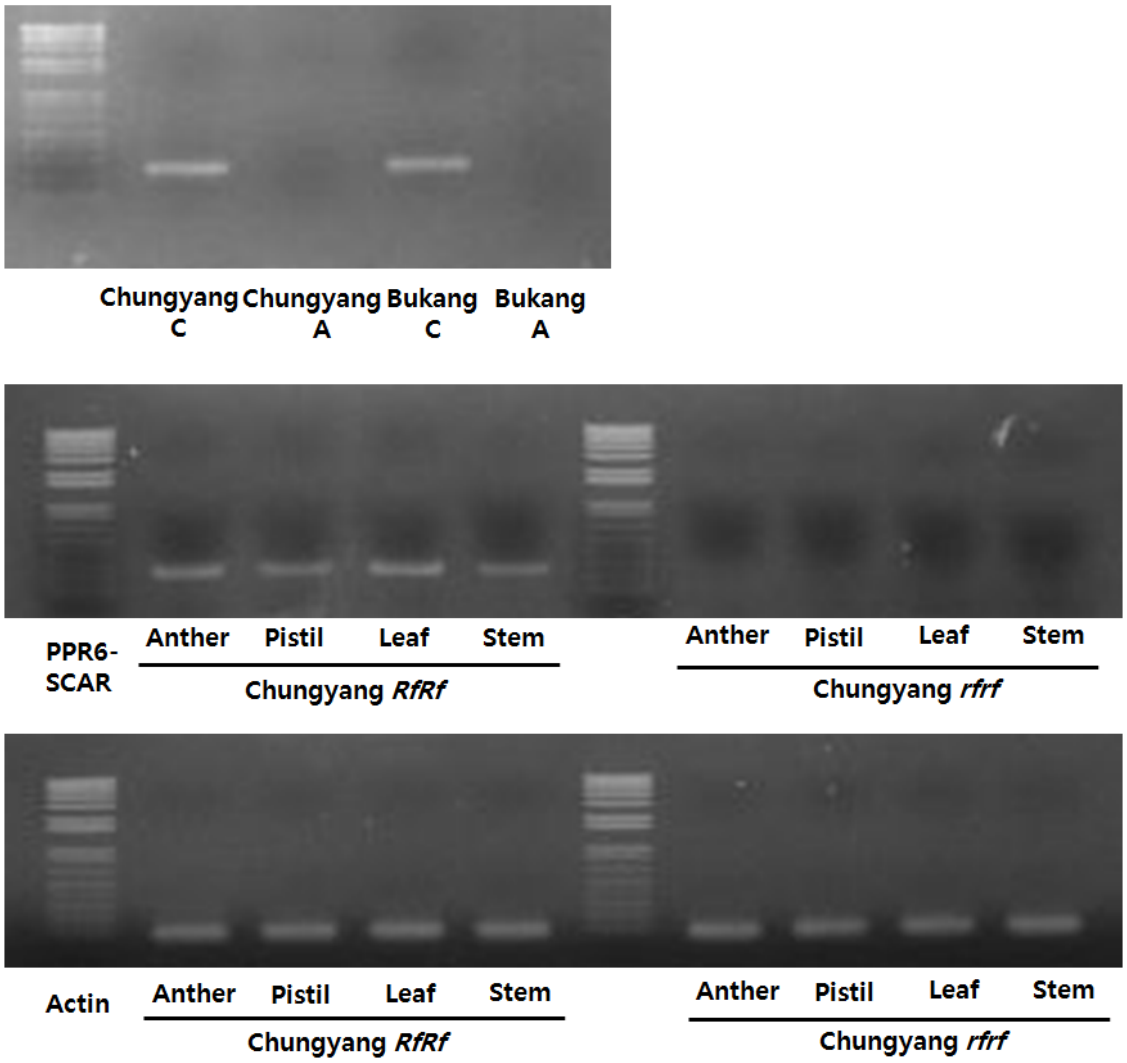


Fig.12. (a) Alignment of transcriptome short read sequences on *PPR6* in Bukang C and Bukang A. The pink colored boxes with red colored letters indicate the nucleotides which are not matched to *PPR6* sequence. (b) PCRs and RT-PCRs using the primer set which generate amplicon in Chungyang C and Bukang C, but not in Chungyang A and Bukang A. The figure upside shows the result of PCRs for gDNAs from each parental lines. The figure in the middle represents RT-PCR result for different tissues in Chungyang F_2 individuals with *RfRf* and *rfrf* genotypes, respectively. The figure in bottom shows results of RT-PCRs for actins.

DISCUSSION

We used petunia *Rf* (Bentolila et al., 2002) to develop molecular markers closely linked to pepper *Rf*. PPR-encoding *Rf* genes and paralog sequences are known to be clustered, similar to R gene clusters (Wang et al., 2006; Geddy and Brown, 2007). Fujii et al. (2011) showed these *Rf* paralogs which called as *Rf*-like genes (*RFL*) originated from the same ancestor and clearly distinguished from other PPR genes. We postulated that the structural similarity of the PPR-encoding *Rf* loci could be employed to clone the pepper *Rf* gene, and develop *Rf*-linked markers. The feasibility of this approach was tested by phylogenic analysis, using the complete sequence of *PePPR1*, the pepper EST with the highest similarity to petunia *Rf*. We demonstrated that *PePPR1* and other *Rf* genes were more closely related to a cluster of previously known Arabidopsis *RFL*s (Lurin et al., 2004; Geddy and Brown, 2007), than to other Arabidopsis PPR genes. This demonstrated that the PPR genes in the *Rf* gene cluster could be distinguished from other PPR genes, and suggested that the pepper homologs of petunia *Rf* was located near the pepper *Rf* locus and one of them might be *Rf* itself. Linkage of all of the mapped pepper *RFL*s supported this hyppothesis. Therefore, BAC clone screening by hybridization-based method was performed using a pepper EST which showed high similarity with petunia *Rf* to isolate BAC clones containing *RFL*s in pepper.

Although none of the markers based on the candidate gene approach co-segregated completely with *Rf* in our earlier attempts, a PPR gene that was on a BAC clone already selected by this approach was determined as the strongest candidate of *Rf* gene by integrating all of the fine mapping results from other strategies. The BAC clone which contained that PPR gene was missed during classification of BAC clones because of failure in efficient and accurate grouping of selected BAC clones due to difficulties in marker development from highly repeated BAC end sequences. However, the final result showed that the strategy which includes hybridization-based BAC screening with *RFL*s can be one of the most straightforward and fast method to isolate *Rf* gene in following two aspects. Firstly, the PPR-type *Rf* genes usually have paralogs which shows very high similarity in nucleotide sequence. The presence of sequences highly similar to *Rf* candidate gene was also detected in this study. If whole genome sequencing was performed based on next generation sequencing methods in which short read sequences are used, the *Rf* gene sequence is easy to be unassembled or wrongly assembled due to the redundancy and complexity of sequence. In fact, none of the contig sequence assembled in whole genome sequencing was perfectly matched with *Rf* candidate gene sequence in pepper. Therefore, the separation of *Rf* paralogs in each BAC clones is helpful to supplement the limitation of gene cloning method based on whole genome sequences obtained from next generation sequencing. Secondly, BAC clones containing *Rf* gene candidate can provide

more sequence information than the candidate gene itself to facilitate development of markers applicable to segregation population because marker development from candidate gene sequence is difficult due to existence of highly similar sequences. If whole genome sequence is available, the sequence information from parts of BAC sequences can be greatly extended to provide more chances for marker development. The marker developed in this way can be used to test the association of the candidate gene with *Rf* gene or accurate classification of selected BAC clones.

We screened a PPR gene named as *PPR6* in genomic DNA region which were co-segregated with *Rf*. We assumed this gene as a strong candidate for *Rf* gene because this was the only gene that matched to anther transcriptome sequence which contain intact *orf* structure. The agreement with the speculated function of gene product and the specific expression in fertile plants of Chungyang F_2 population supported this assumption. However, further analysis on DNA region which co-segregates with *Rf* is required to screen any possible functional transcripts in this region and to confirm the status of PPR6 as an *Rf* candidate because we cannot exclude the possibility that some functional transcripts were not screened or not assembled in transcriptome analysis.

The cloned *PPR6* gene showed typical characteristics of P-subfamily PPR genes. In PPR6 protein sequence, fourteen repeats of PPR motif were detected while one of the repeats had one amino acid insertion. General characteristics of

PPR6 which includes the number of motif repetition, total length of peptide, and insertion of one amino acid in one of the repeats were highly similar with the pattern in petunia Rf592 (Bentolila et al., 2002). Three of homologs of *PPR6* cloned in this study also revealed representative characteristics of *RFL*s. Fujii et al. (2011) showed that the first, third, and sixth amino acid in each repeats of PPR gene are under strong positive selection. In addition, Barkan et al. (2012) revealed that the combination of first and sixth amino acids in repeats mainly specify one nucleotide in RNA binding site for PPR gene. High rate of amino acid substitution in the three hypervariable amino acids was also detected in the comparison of *PPR6* homologs. For example, about one third (32%) of amino acid change on the total length of repeated region was detected on these three amino acids in the alignment of PPR6 and PPR6-2 which are paralogs in Chungyang restorer parent. These results implied that PPR6 paralogs might be under positive selection during the co-evolution of CMS cytoplasm and *Rf*.

A *PPR6* homolog cloned in a CMS line (*PPR6-3*) showed very high similarity with *PPR6* gene. In protein sequence, only three amino acids were different between PPR6 and PPR6-3. Interestingly, one of these was a sixth amino acid of a repeated motif which may be under positive selection. Analysis on the function of PPR6-3 or the expression of *PPR6-3* should be performed to test whether *PPR6* can remain to be candidate gene of *Rf*.

One of the goals in this study was to develop markers that can be broadly

applied to breeding lines. Lee et al. (2008) and Min et al. (2008) claimed that there are three haplotypes of OPP13-CAPS and PR-CAPS in Korean breeding lines. Min et al. (2008) identified the third haplotype of the OPP13-linked sequence in pepper germplasm and showed that a large number of unstable male sterile lines have this haplotype. They proposed that the third haplotype is significantly related to the stability of male sterility in pepper. Lee et al. (2008) revealed that the internal sequence of the PR-CAPS marker, which is linked to the partial restoration phenotype, also has the third haplotype. A clear relationship between the marker haplotype and the *Rf* genotype has not yet been demonstrated, however. In fact, when we used these markers to screen panels of pepper breeding lines from seed companies, we found many discrepancies between the *Rf* phenotype and the marker haplotype. Since the breeding lines were fully fertile or sterile, we concluded that the third haplotype is not predictive of *Rf* genotype in these lines. As noted previously (Lee et al., 2008), discrepancies between marker genotypes and *Rf* phenotypes may be due to frequent recombinations between marker loci and the *Rf* gene during introgression of the gene into elite lines.

To develop more broadly applicable *Rf*-linked markers, we used a Chungyang F_2 population for which previously developed markers (OPP13-CAPS, AFRF8-CAPS, PR-CAPS) cannot be used because of lack of polymorphism. We developed new molecular markers that can be used with Chungyang cultivars. In the application of most of newly-developed markers to panels of breeding lines,

high level of discrepancies between marker genotypes and *Rf* phenotypes were detected. The *Rf* genes, like the disease resistance genes, have been found in clusters of PPR genes which are proposed to be evolved by duplication (Geddy and Brown, 2007). This may lead to genome rearrangement over generations and thus contributed to discrepancy in marker location among different populations. However, two markers (3336 last2 SCAR, 120kb F2R3 SCAR) developed from the DNA region which were co-segregated with *Rf* gene showed highly accurate genotyping results that were clearly distinguished from those of other markers. This result indicated that the closest localization of two markers to *Rf* gene among developed markers resulted in high level of association between markers and *Rf* gene during evolution. We expect that these markers will increase the utility of molecular markers in molecular breeding for CGMS system of pepper.

REFERENCES

Barr CM, Fishman L (2010) Cytoplasmic male sterility in Mimulus hybrids has pleiotropic effects on corolla and pistil traits. Heredity (Edinb) 106: 886-93

Bentolila S, Alfonso AA, Hanson MR (2002) A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci USA 99: 10887-92

Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, Small I (2012) A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. PLoS Genet 8: e1002910.

Brown GG, Formanova N, Jin H, Wargachuk R, Dendy C, Patil P, Laforest M, Zhang J, Cheung WY, Landry BS (2003) The radish *Rfo* restorer gene of Ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. Plant J 35: 262-72

Carlsson J, Leino M, Sohlberg J, Sundstrom JF, Glimelius K (2008) Mitochondrial regulation of flower development. Mitochondrion 8: 74-86

Cui X, Wise RP, Schnable PS (1996) The *rf2* nuclear restorer gene of male-sterile T-cytoplasm maize. Science 272: 1334-6

Desloire S, Gherbi H, Laloui W, Marhadour S, Clouet V, Cattolico L, Falentin C, Giancola S, Renard M, Budar F, Small I, Caboche M, Delourme R, Bendahmane A (2003) Identification of the fertility restoration locus, *Rfo*, in radish, as a member of the pentatricopeptide-repeat protein family. EMBO Rep 4: 588-94

Fujii S, Toriyama K (2009) Suppressed expression of Retrograde-Regulated Male Sterility restores pollen fertility in cytoplasmic male sterile rice plants. Proc Natl Acad Sci U S A 106: 9513-8.

Fujii S, Bond CS, Small ID (2011) Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. Proc Natl Acad Sci U S A 108: 1723-8

Geddy R, Brown GG (2007) Genes encoding pentatricopeptide repeat (PPR) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. BMC Genomics 8: 130

Givry SD, Bouchez M, Chabrier P, Milan D, Schiex T (2005) CARTHAGENE: multi-population integrated genetic and radiation hybrid mapping. Bioinformatics 21: 1703-1704

Gulyas G, Pakozdi K, Lee JS, Hirata Y (2006) Analysis of fertility restoration by using cytoplasmic male-sterile red pepper (*Capsicum annuum* L.) lines. Breed Sci 56: 331-334

Hanson MR, Bentolila S (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell 16 Suppl S154-69

Itabashi E, Iwata N, Fujii S, Kazama T, Toriyama K The fertility restorer gene, *Rf2*, for Lead Rice-type cytoplasmic male sterility of rice encodes a mitochondrial glycine-rich protein. Plant J 65: 359-67

Kim DH, Kang JG, Kim BD (2007) Isolation and characterization of the cytoplasmic male sterility-associated *orf456* gene of chili pepper (*Capsicum annuum* L.). Plant Mol Biol 63: 519-32

Kim DS (2005) Development of RAPD and AFLP markers linked to fertility restorer (*Rf*) gene in chili pepper (*Capsicum annuum* L.). Thesis, Seoul National University

Kim DS, Kim DH, Yoo JH, Kim BD (2006) Cleaved amplified polymorphic sequence and amplified fragment length polymorphism markers linked to the fertility restorer gene in chili pepper (*Capsicum annuum* L.). Mol Cells 21: 135-40

Koizuka N, Imai R, Fujimoto H, Hayakawa T, Kimura Y, Kohno-Murase J, Sakai

T, Kawasaki S, Imamura J (2003) Genetic characterization of a pentatricopeptide repeat protein gene, *orf687*, that restores fertility in the cytoplasmic male-sterile Kosena radish. Plant J 34: 407-15

Komori T, Ohta S, Murai N, Takakura Y, Kuraya Y, Suzuki S, Hiei Y, Imaseki H, Nitta N (2004) Map-based cloning of a fertility restorer gene, *Rf-1*, in rice (*Oryza sativa* L.). Plant J 37: 315-25

Kotera E, Tasaka M, Shikanai T (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433: 326-30

Lee DH (2001) Studies on unstable fertility of CGMS (cytoplasmic-genic male sterility) in *Capsicum annuum* L. Dissertation, Seoul National University

Lee J, Yoon JB, Park HG (2008) Linkage analysis between the partial restoration (*pr*) and the restorer-of-fertility (*Rf*) loci in pepper cytoplasmic male sterility. Theor Appl Genet 117: 383-9

Liu F, Cui X, Horner HT, Weiner H, Schnable PS (2001) Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. Plant Cell 13: 1063-78

Livingstone KD, Lackney VK, Blauth JR, van Wijk R, Jahn MK (1999) Genome mapping in *Capsicum* and the evolution of genome structure in the *Solanaceae*. Genetics 152: 1183-202

Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette ML, Mireau H, Peeters N, Renou JP, Szurek B, Taconnat L, Small I (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16 : 2089-103

Matsuhira H, Kagami H, Kurata M, Kitazaki K, Matsunaga M, Hamaguchi Y, Hagihara E, Ueda M, Harada M, Muramatsu A, Yui-Kurino R, Taguchi K, Tamagake H, Mikami T, Kubo T (2012) Unusual and typical features of a novel restorer-of-fertility gene of sugar beet (*Beta vulgaris* L.). Genetics

192: 1347-58

Min WK, Lim H, Lee YP, Sung SK, Kim BD, Kim S (2008) Identification of a third haplotype of the sequence linked to the Restorer-of-fertility (*Rf*) gene and its implications for male-sterility phenotypes in peppers *(Capsicum annuum* L.). Mol Cells 25: 20-29

Novak F, Betlach J, Dubovsky J (1971) Cytoplasmic male sterility in sweet pepper (*Casicum annuum* L.). I. Phenotype and inheritance of male sterile character. Z Pflanzenzucht 65: 129-40

Peterson PA (1958) Cytoplasmically inherited male sterility in *Capsicum*. Amer Nat 92: 111-9

Schnable PS, Wise RP (1998) The molecular basis of cytoplasmic male sterility and fertility restoration. Trends Plant Sci 3: 175-80

Small ID, Peeters N (2000) The PPR motif - a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25: 46-7

Shifriss C (1997) Male sterility in pepper (*Capsicum annuum* L.). Euphytica 93: 83-8

Yoo EY, Kim S, Kim YH, Lee CJ, Kim BD (2003) Construction of a deep coverage BAC library from *Capsicum annuum*, 'CM334'. Theor Appl Genet 107: 540-3

Wang LH, Zhang BX, Lefebvre V, Huang SW, Daubeze AM, Palloix A (2004) QTL analysis of fertility restoration in cytoplasmic male sterile pepper. Theor Appl Genet 109: 1058-63

Wang Z, Zou Y, Li X, Zhang Q, Chen L, Wu H, Su D, Chen Y, Guo J, Luo D, Long Y, Zhong Y, Liu YG (2006) Cytoplasmic male sterility of rice with boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing. Plant Cell 18: 676-87

Zhang BX, Huang SW, Yang GM, Guo JZ (2000) Two RAPD markers linked to a major fertility restorer gene in pepper. Euphytica 113: 155-61

CHAPTER III

Utilization of Chloroplast Genome Sequences for the Determination of the Origin of CMS Cytoplasm in Pepper (*Capsicum annuum* L.)

ABSTRACT

Cytoplasmic male sterility (CMS) is a maternally inherited inability to produce functional pollen, which may be caused by incompatibility between nuclear and mitochondrial genome. Although the CMS cytoplasm from a pepper accession 'PI164835' has been widely used to produce F_1 hybrid seeds in pepper, the origin of this cytoplasm has not been determined in DNA sequence level. Because plastid genome is co-transmitted with mitochondrial genome and highly stable compared to mitochondria genome, we used plastid barcode sequences to deduce the origin of the pepper CMS cytoplasm. The complete sequence of

plastid genome in a pepper line ‘FS4401’ was assembled and used as the source for marker development. Two plastid sequences, *trnH-psbA* and *rpl16-rpl18* intergenic sequences, were used to analyze cytoplasm types of pepper germplasms which include six *Capsicum* species. Plastid barcode analysis revealed that cytoplasm types can be divided into six types and four types for *trnH-psbA* and *rpl16-rpl18*, respectively. The sequences in these two regions of CMS pepper lines were identical to the sequences of a cytoplasm type of a particular clade of *C. annuum*. For further investigation, two molecular markers were designed from *TrnL-TrnF* and *rpl16-rpl18* intergenic regions, respectively. Application of these markers to a larger number of germplasm confirmed that the cytoplasm type of CMS is identical to the cytoplasm type of the particular *C. annuum* clade. These results suggest that the CMS cytoplasm of pepper may originated from a cross in which the seed parent belonged to the *C. annuum* clade.

INTRODUCTION

Cytoplasmic male sterility (CMS) is defined as the maternally inherited inability to produce functional pollen (Hanson and Bentolila, 2004). In many crop plants, CMS is the result of mitochondrial chimeric genes generated by mitochondrial genome rearrangements (Hanson and Bentolila, 2004). Most rearrangements that induce CMS were shown to occur as a result of wide crosses, or interspecific exchange of nuclear and cytoplasmic genomes (Schnable and Wise, 1998).

In pepper, CMS cytoplasm was first isolated from an Indian *C.annuum* accession (PI164835) and has been used commercially to produce F1 hybrid seeds at seed companies (Peterson, 1958; Kumar et al, 2008). Two candidate genes for the male sterility, *orf507* and *ψatp6-2*, have been identified and studied. *orf507*, a chimeric gene fused with unknown sequences, induced male sterility in Arabidopsis when a portion of this gene was overexpressed. Moreover, expression of ORF507 was suppressed in pepper lines containing the restoration-of-fertility (*Rf*) gene in the nuclear genome (Kim et al, 2007). The *ψatp6-2* is a 3'-truncated form of wild type *atp6-2* present in the normal cytoplasm. Presence of these genes specifically in male sterile lines and regulation of their expression by the *Rf* gene support the idea that these genes may be involved in CMS of pepper (Kim et al, 2006).

To obtain a novel source of CMS cytoplasm and to unravel the origin of the present CMS cytoplasm, interspecific crosses were made between *C. annuum* and other species in *Capsicum* genus, including *C. baccatum*, *C. chinense*, *C. frutescens* and *C. chacoense* (Shifriss, 1997). Among these crosses, male sterility was observed in the crosses of *C. baccatum* X *C. annuum* (Andrasfalvy and Csillery, 1983), *C. chacoense* X *C. annuum* (Kumar et al., 2009) and *C. frutescens* X *C. annuum* (Yu, 1990). In the *C. baccatum* X *C. annuum* and *C. chacoense* X *C. annuum* combinations, antherless flowers were obtained; however, restorer genes were not identified. Meanwhile, the cytoplasm obtained by a *C. frutescens* X *C. annuum* cross combination resulted in male sterility in the maintainer nuclear background and fertility were restored by the same *Rf* gene from PI164835 (Yu, 1990). This implied that the CMS cytoplasm obtained by the interspecific cross and the cytoplasm of PI164835 may have similar features. However, the CMS organellar genome and those of artificial crosses have never been directly compared.

Mitochondrial genomes of higher plants are not only highly variable in gene order and intergenic sequences between species, but also highly complex in the stoichiometry of subgenomic molecules (Mackenzie and McIntosh, 1999). For exemple, a research in radish showed that a CMS-associated ORF exists even in a normal cytoplasm although the copy number of this ORF is maintained at a very low level (Kim et al., 2007). The substoichiometric copy number of mitochondrial

DNA molecules can be stably maintained through generations by unknown mechanisms (Kim et al., 2007). In contrast, the overall structures and orders of plastid genomes have been well conserved during speciation. However, several DNA regions called barcode sequences contain variations that are relevant for evolutionary study (Palmer, 1990; Taberlet et al., 1991; Kress and Erickson., 2005). These features enable plastid sequences to be used to determine species origin (Yukawa et al., 2006) and for the classification of cytoplasm type at the intraspecific level (Kim et al., 2009).

In pepper, several intergenic sequences on plastid genome have been used to classify accessions. For exemple, *trnH-psbA* (Jarret, 2008) sequences were shown to be highly polymorphic among species and be able to identify most of cultivated species in *Capsicum* genus. Recently, complete sequence of plastid genome containing 113 unique genes was analyzed and this could be utilized in evolutionary studies for pepper (Jo et al., 2011). In contrast, analysis of mitochondrial DNA sequences was performed only for the restricted DNA regions around the CMS-associated gene in pepper (Kim and Kim, 2005; Kim and Kim, 2006; Kim et al., 2007).

In this study, we assembled the complete sequence of pepper plastid DNA and developed markers which can classify the cytoplasm types among *Capsicum* species. These markers were applied to deduce the origin of the CMS cytoplasm in pepper.

MATERIALS AND METHODS

Plant materials

Plastid sequence analysis, a chili pepper cultivar (*C. annuum* L.) 'FS4401' was provided by Monsanto Korea and used.

A total of 72 *Capsicum* lines including 34 *C. annuum*, 10 *C. frutescens*, 4 *C. chinense*, 11 *C. baccatum*, 8 *C. chacoense* and 5 *C. pubescens* were randomly selected from the germplasm collections at the Seoul National University Germplasm Center (Seoul, Korea) and used for marker analysis (Table 3). Young leaves from the 16 lines containing cytoplasmic male sterile (CMS) cytoplasm and 25 lines containing wild-type *C. annuum* cytoplasm were provided by Monsanto Korea (Chochiwon, Korea; Table 4). All breeding lines used in this study were included in breeding lines which were used in a previous study (Jo et al., 2009).

Plastid genome assembly

The plastid sequences were obtained during mtDNA sequencing as byproduct (Chapter I) The sequences were assembled using the CAP3 program (Huang and Madan, 1999). Although the major portion of the analyzed DNA was from the mitochondria, extremely high sequence coverage enabled us to assemble contigs covering most of the plastid genome. A total of 12 contigs longer than 2

kb were shown to contain plastid DNA sequences using Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/) and these contigs covered 137 kb of the plastid genome in total. Gaps between contigs were filled by direct sequencing of PCR product amplified from primers designed using the end sequences of each contig. An additional 17 PCR reactions, amplifying contig regions that were highly divergent from other solanaceous chloroplast genomes, were performed using DNA from green leaves. The sequences of PCR products were directly analyzed to confirm validity of the contig assembly.

Gene annotation, sequence alignment, and repeat prediction

Chloroplast genes were annotated on the chloroplast genome sequence of FS4401 using the Dual Organellar GenoMe Annotator (DOGMA; Wyman et al., 2004). This program uses BLASTX against 16 chloroplast genomes of plants to identify chloroplast genes in query sequences. The complete chloroplast sequences of seven other solanaceous species were obtained from GenBank for comparative plastome analysis: *Nicotiana tabacum* [Z00044.2], *N. sylvestris* [AB237912.1], *N. tomentosiformis* [AB240139.1], *Atropa belladonna* [AJ316582.1], *Solanum bulbocastanum* [DQ347958.1], *S. tuberosum* [DQ386163.1], and *S. lycopersicum* [AM087200.3].

DNA isolation and sequence analysis

Total genomic DNA was isolated using the method of Prince et al. (1997) from young green leaves of germplasm and breeding lines. Twenty-six lines were randomly selected and the *trnH-psbA* and *rpl14-rpl16* intergenic sequences were analyzed. The primers for each sequence were designed using the complete FS4401 pepper plastid sequence. For *trnH-psbA*, the forward primer sequence was 5'-GATCCACTTGGCTACATCC-3' and the reverse primer sequence was 5'-GCTATCGAAGCTCCATCTAC-3'. For *rpl14-rpl16*, the forward primer sequence was 5'-CAGCCCTGACTACTTCTGATC-3' and the reverse primer sequence was 5'-GTTAAACCGGGGCGAATAC-3'. The amplification reaction was conducted as follows: an initial cycle at 94^{o}C for 5 min; 35 cycles of 94^{o}C for 30 sec, 55^{o}C for 30 sec, 72^{o}C for 2 min; and a final extension of 72^{o}C for 10 min. The PCR products were eluted and directly sequenced. Intergenic sequences were selected and aligned using the ClustalW2 program (http://www.ebi.ac.uk/Tools/msa /clustalw2/).

DNA marker analysis

Three sequence-characterized amplified regions (SCARs; PepRpl, *orf507* and *ψatp6-2*) and one marker based on high-resolution melting analysis (HRM; PepTrn) were applied to the germplasm and breeding lines as previously described (Table 1). The primer sequences of these markers are listed in Table 1. For PepRpl, PCR reactions were performed in 25 μl with 50 ng of template DNA, 2.5 μl of

10X *Taq* DNA polymerase buffer, 5 pmol of each primer, 200 μM dNTPs and 1 unit of *Taq* DNA polymerase. Amplification was conducted as follows: an initial cycle at 94^{o}C for 5 min; 35 cycles of 94^{o}C for 30 sec, 59^{o}C for 30 sec, 72^{o}C for 1 min; and a final extension of 72^{o}C for 10 min. The PCR conditions for *orf507* SCAR and *ψatp6-2* were the same with that of PepRpl except for annealing temperature, which was 58^{o}C and 55^{o}C, respectively.

High resolution melting analysis

High resolution melting (HRM) analysis was performed for PepTrn. PCR reactions were done in 20μl with 50 ng of template DNA, 2 μl of 10X Taq DNA polymerase buffer, 5 pmol of each primer, 200 μM dNTPs, 1.25 μM of SYTO 9 dye and 1 unit of DNA Taq DNA polymerase. The sequence of forward primer was 5'- GAGCAAGGAATCCCTAGTTG - 3' and sequence of reverse primer was 5'- GGATTTTCAGGGGTATACCAA- 3'. Using a Rotor-gene^{TM} 6000 (Corbette, Australia), real-time PCR amplification (95^{o}C 10 min; 50 cycles of 94^{o}C 20 sec, 53^{o}C 20 sec and 72^{o}C 30 sec; 95^{o}C 60 sec; 72^{o}C 60 sec) and HRM analysis (increasing 0.1^{o}C every 1 minute from 70^{o}C to 90^{o}C) were performed.

Table 1. Molecular markers and primer sequences used in this study

Marker	Primer	Sequence (5'to 3')	Annealing temperature (^{o}C)	Reference
PepRpl	PepRpl F	AATCCGTTATTTGAATGCATTT	59	This study
	PepRpl R	GTTAAACCGGGGCGAATAC		
PepTrn	PepTrn F	GAGCAAGGAATCCCTAGTTG	55	This study
	PepTrn R	GGATTTTCAGGGGTATACCAA		
orf507	orf456 F	ATGCCCAAAAGTCCCATGTA	58	Jo et al., 2009
SCAR	orf456 R	TTACTCGGTTGCATTGTTT		
ψatp6-2	atp6-2(S) F	TGGATCTCGCTATTAACCAC	55	Jo et al., 2009
SCAR	atp6-2(S) R	GTAGTTCATTCGGACCTAGTAG		

RESULTS

Assembly of *C. annuum* plastid genome

Plasid genome sequences were obtained as a byproduct of the sequencing of DNA isolated from a fraction of cell orgnelles of *C.annuum*. The majority of the *C. annuum* plastid genome sequence, constituting 137kb in total, was successfully assembled. Although plastid DNA is often integrated into the mitochondrial genome in plant species (Cliffton et al., 2004; Sugiyama et al., 2005), the contig sequences we used in the assembly of the pepper plastome likely originated from plastid rather than mitochondria for two reasons. First, no redundant contigs were obtained for any of the selected contigs. Second, we did not detect any significant mutations resulting in frame shifts or abrupt appearance of stop codons in any of the contig sequences; plastid sequences that are integrated into mitochondrial genomes are usually associated with gene sequences that are non-functional and prone to mutation. Meanwhile, gap sequences between contigs were expected to correspond to plastid genome portions that were integrated into the mitochondrial genome. For instance, a gap was detected on the upstream region of the *accD* gene during contig assembly, and a sequence with high similarity to this region is also found in the mitochondrial genome (Jo et al., 2009). The redundancy of these sequences might prevent the assembly of contigs

containing plastid sequences. The complete plastome sequence was obtained through sequence analysis of PCR products that were amplified by primers designed using end sequences of plastid-specific contigs and shown to contain single sequences by chromatogram analysis.

Organization and gene contents of pepper chloroplast genome

The estimated size of the pepper plastid genome is 156,781 bp, which is the largest among known solanaceous plastomes. The quardripartite structure includes 87,366 bp of LSC and 25,783 bp of SSC that are separated by a pair of 17,849 bp of IR copies (Fig. 1; Table 2). The GC content is 37.7%, which is consistent with other solanaceous plastomes. Coding sequences constitute 58.5% of the pepper plastome sequence. There are 113 unique genes among which 20 are duplicated in IR sequences. A total of 79 unique genes (six duplicated) encode proteins including photosynthesis-related proteins (46 genes), genetic system related proteins (27 genes), proteins with unique function such as acetyl-CoA carboxylase subunit (*accD*) and heme attachment to cytochrome C (*ccsA*), and proteins with unknown functions (*ycf*s). In addition, 30 unique genes (seven duplicated) and four unique duplicated genes encode for tRNAs and rRNAs, respectively. The gene contents and gene order of *C. annuum* were identical to those of the seven previously-known solanaceous plastomes (Table 2).

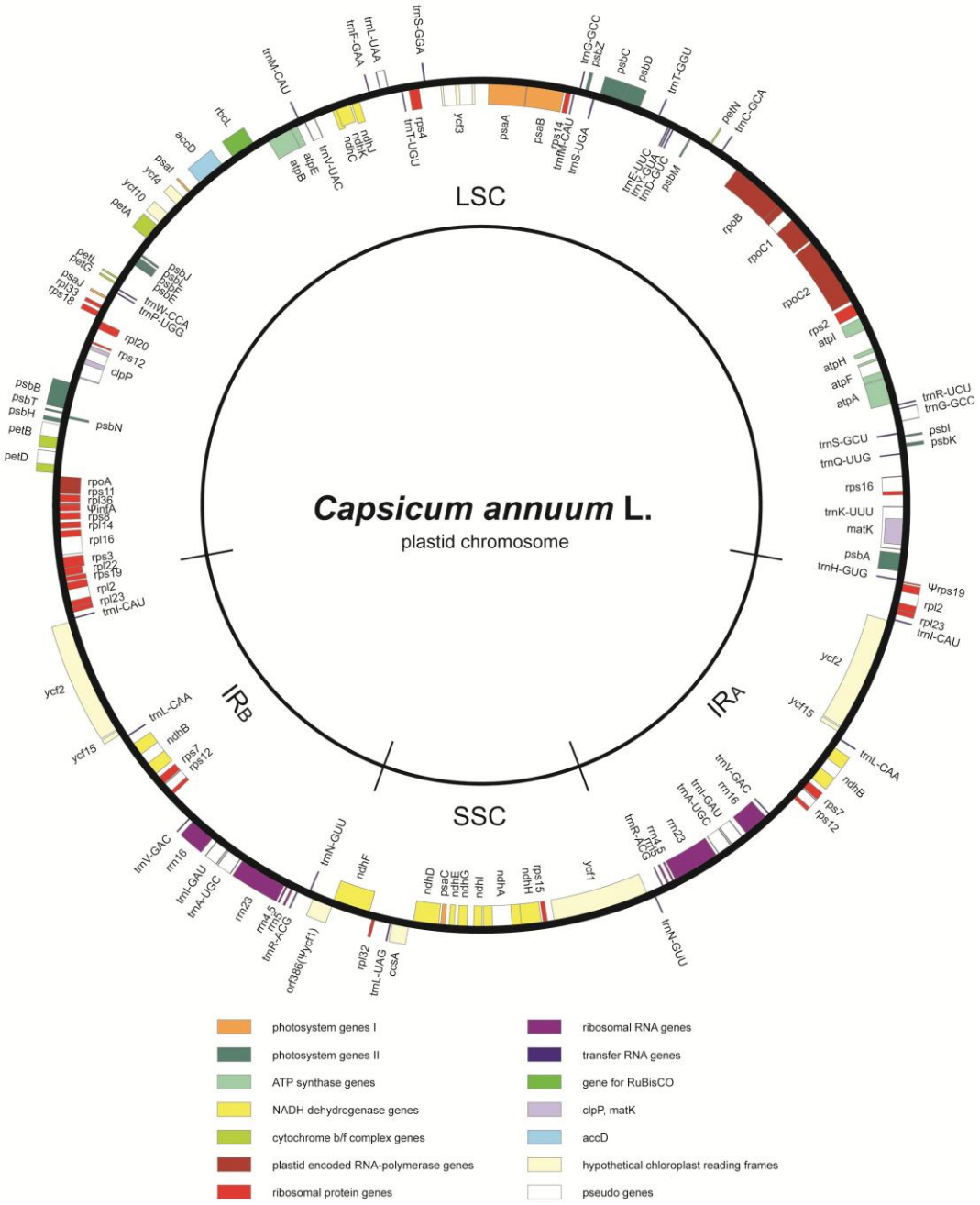


Fig.1. Gene map of the chloroplast genome of *Capsicum annuum* L. The genes drawn outside of the circle are transcribed clockwise and inside are transcribed counterclockwise. The colors of the genes are classified according to the functions of the gene products.

Table 2. Summary and comparison of solanaceous plastomes.

	Atropa belladonna	*Capsicum annuum*	*Nicotiana tabacum*	*Nicotiana sylvestris*	*Nicotiana Tomentosiformis*	*Solanum Bulbocastanum*	*Solanum lycopersicum*	*Solanum tuberosum*
Genome size(bp)	156,687	156,781	155,943	155,941	155,745	155,371	155,461	155,298
LSC(bp)	86,868	87,366	86,686	86,684	86,392	85,814	85,882	85,749
SSC(bp)	18,008	17,849	18,573	18,573	18,485	18,381	18,363	18,373
IR (bp)	25,906	25,783	25,342	25,342	25,429	25,588	25,611	25,595
GC content	37.6 (%)	37.7	37.9	37.9	37.8	37.9	37.9	37.9

Analyses of *trnH-psbA* and *rpl14-rpl16* intergenic sequences

The *trnH-psbA* and *rpl14-rpl16* intergenic sequences were analyzed for twenty-six pepper germplasms including six *Capsicum* species and three *C. annuum* CMS lines (Table 3; Fig. 2). Sequence comparison results revealed six and four haplotypes for *trnH-psbA* and *rpl14-rpl16* intergenic sequences, respectively. Based on *trnH-psbA* sequences, *C. annuum* sequences were sub-classified into two distinct haplotypes, Type 1 and Type 2. Most of *C. chinense* and *C. frutescens* lines showed Type 3 sequence while one of *C. frutescens* lines (C00276) contained Type 4 sequence. Other accessions included in three species, *C. baccatum*, *C. chacoense* and *C. pubescens*, represented three haplotypes, Type 4, 5 and 6, respectively. The CMS lines (Bukang A, Chungyang A, and FS4401) had the same haplotype as one of *C. annuum* line groups, represented as Type 2 (Table 3). The *trnH-psbA* intergenic sequence in this haplotype was clearly

distinguished from all other haplotype sequences by the presence of a seven base pair insertion (TAAATG) at the nucleotide 278 (Fig. 2a). As for the *rpl14-rpl16* intergenic sequences, *C. annuum, C. chinense, C. baccatum* and *C. frutescens* lines had either Type 1 or Type 2 haplotype and could not be grouped according haplotypes whereas *C. chacoense* and *C. pubescens* lines had Type 3 and Type 4, respectively. *C. annuum* lines having Type 2 haploptype was different from the other three haplotypes due to a ten base pair insertion (TGCATTTGAA) (Fig 2b). The lines classified as the Type 2 haplotype for *rpl14-rpl16* intergenic sequence were identical to the lines containing the Type 2 haplotype for the *trnH-psbA* intergenic sequence (Table 3). The *rpl14-rpl16* intergenic sequence haplotype for the CMS *C. annuum* lines was also Type 2.

Table 3. Classification of *trnH-psbA* and *rpl14-rpl16* haplotype sequences of cultivars or accessions in six *Capsicum* species

Species	Cultivar or accession	Origin	*trnH-psbA* haplotype	*rpl14-rpl16* haplotype
C. annuum	Twilight	America	Type 1	Type 1
	Poinsettia	America		
	Carolina Wonder	America	Type 2	Type 2
	Chitawn	India		
	Sweet banana	America		
C. annuum (CMS)	Bukang A	South Korea	Type 2	Type 2
	Chungyang A	South Korea		
	FS4401	South Korea		
C. chinense	Jalapeno	Mexico	Type 3	Type 1
	PI159234	America		
	Early red sweet	America		
	Habanero	Mexico		
C. frutescens	C00050	Costa Rica	Type 3	Type 1
	C00966	Peru		
	C00276	Netherlands	Type 4	
C. baccatum	C00044	Costa Rica	Type 4	Type 1
	C01686	America		
	C01742	Netherlands		
C. chacoense	C04388	Argentina	Type 5	Type 3
	C04389	Argentina		
	C04390	Bolivia		
	C04391	Bolivia		
C. pubescens	C01323	Guatemala	Type 6	Type 4
	C01374	Peru		
	C01572	Guatemala		
	C04895	Ecuador		

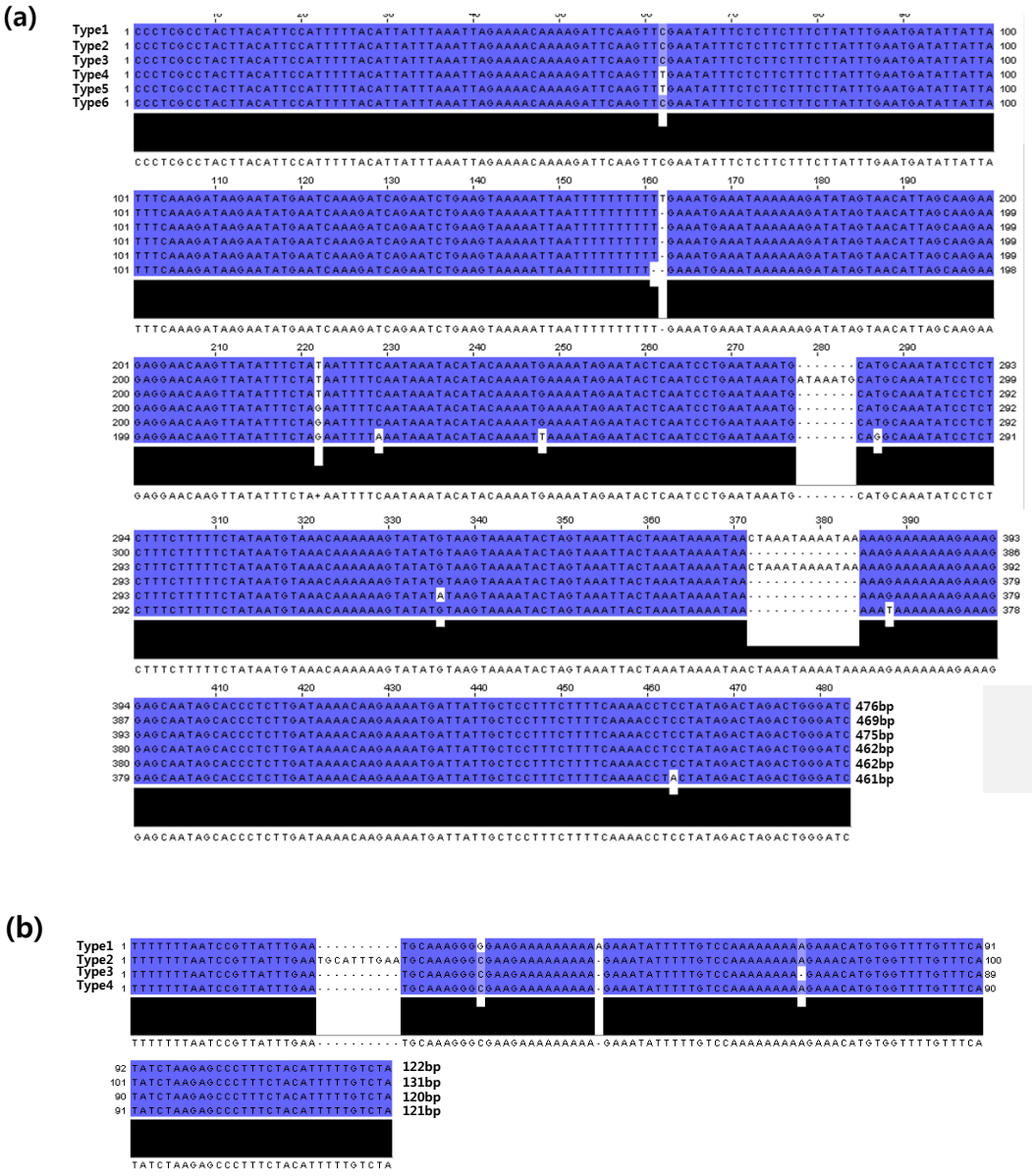
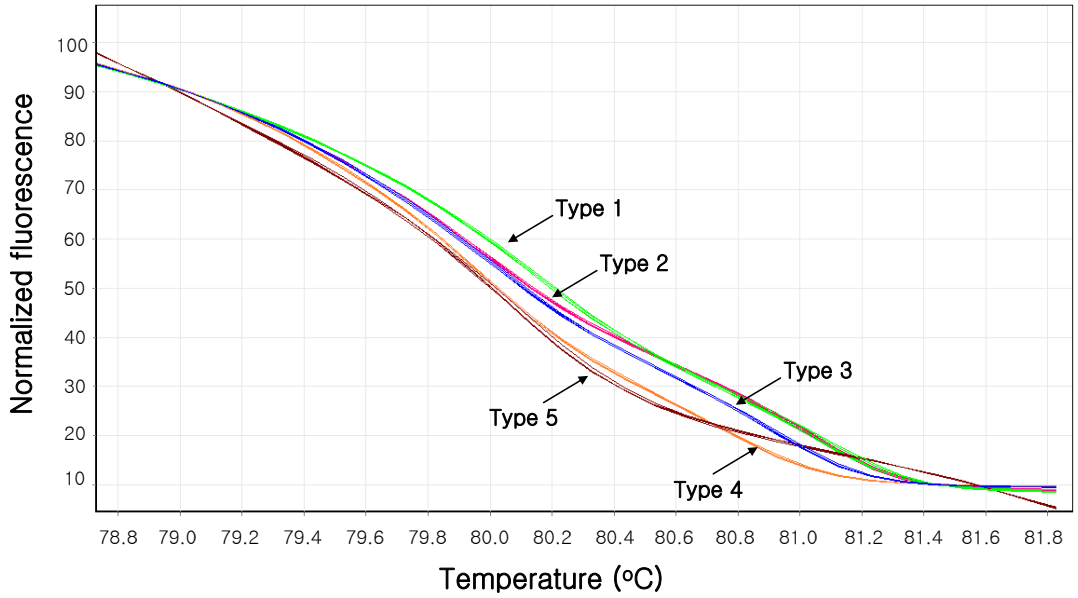


Figure 2. Nucleotide sequence alignments. (a) Alignment of six types of *trnH-psbA* intergenic sequences detected in six *Capsicum* species. (b) Alignment of four types of *rpl14-rpl16* intergenic sequences detected in six *Capsicum* species.

Development of markers to classify pepper species

An HRM marker (PepTrn) for the determination of species in *Capsicum* was designed from the intergenic sequence between *TrnL* and *TrnF*. Because previous report indicated that short length of amplicon is important factor which influence the resolution of HRM analysis (Park et al., 2009), the primers for this marker were designed to amplify a portion of the intergenic region instead of full length of this region. A total of five types of melting curves were detected as the result of high resolution melting analysis using DNAs from 72 germplasm (Fig 3a, b). The sequences of amplicons from representative germplasm for each type were analyzed. The sequence of each amplicon and the alignment between sequences were represented in Fig 4. Compared to type 1 sequence which is 376 bp in length, type 2 (362 bp) and type 3 (362 bp) sequences have a 14 bp deletion in common. Type 4 sequence (343 bp) contains a 19 bp deletion which makes this sequence as the shortest one and type 5 sequence (480 bp) have a large insertion which is distinguished from other sequences. In addition, type3 and type5 sequences have SNPs at 212th and 330th base pair, respectively, compared to other type sequences.

(a)



(b)

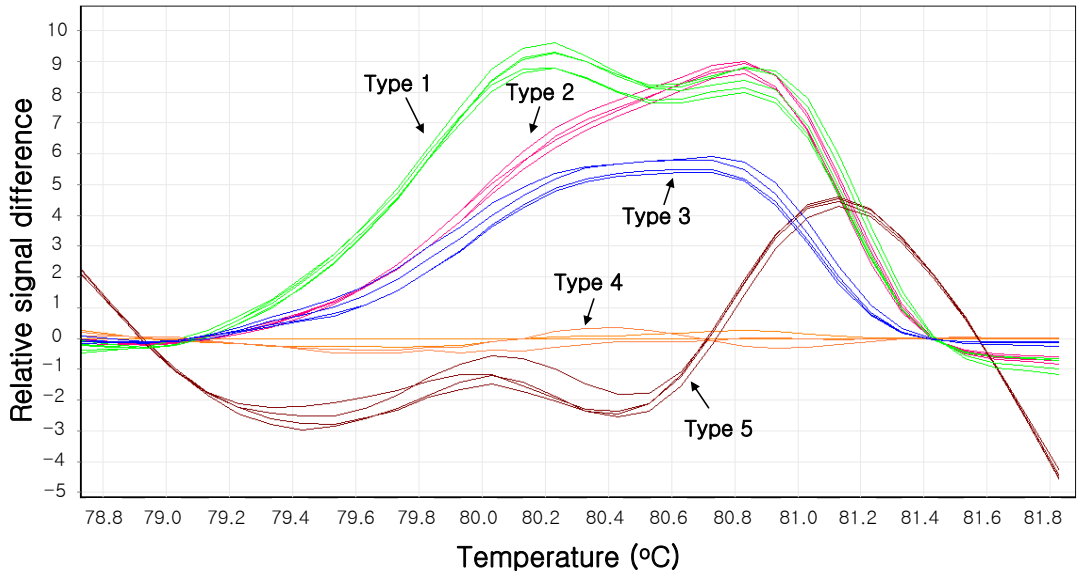


Fig.3. Five types of melting curves detected for PepTrn. (a) Normalized melting curves. (b) Difference plots.

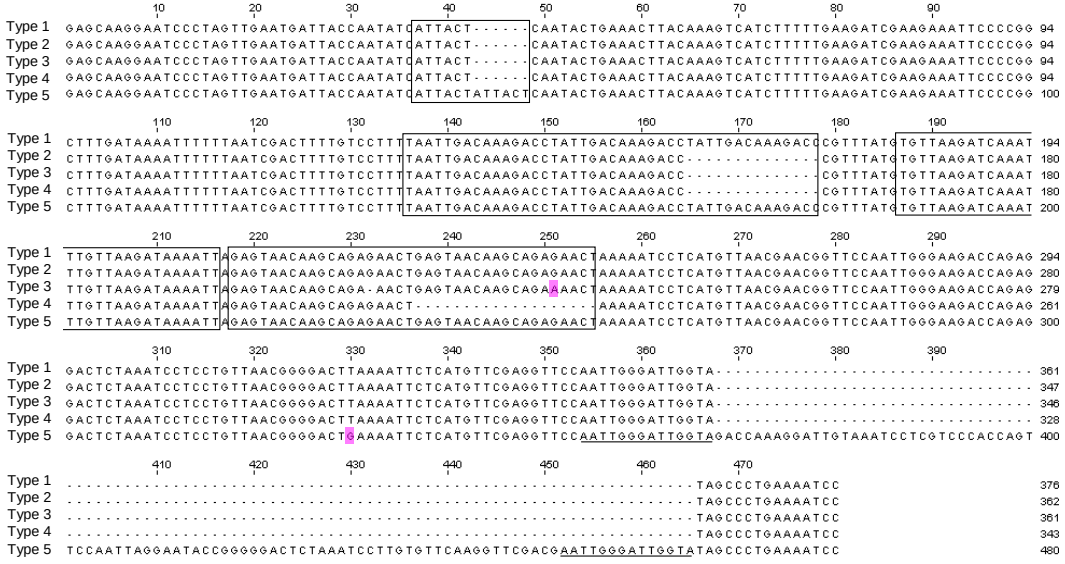


Fig.4. Alignment of five type sequences for PeChloro2. Array of repeated sequences are enclosed in boxes and repeated sequences in the borders of type 5 sequence are underlined. SNPs are indicated by shadows.

A SCAR marker, PepRpl, was designed to amplify a Type 2-specific *rpl14-rpl16* intergenic sequence. This marker was designed based on an alignment of sequences around the 3' end of *rpl16* (Fig. 5a). The forward primer of this marker was designed to include a part of the Type 2-specific insertion while the reverse primer was located in a conserved region of the *rpl16* sequence (Fig. 5a). Application to pepper lines (Fig. 5b) resulted in the specific amplification of a 246 bp DNA fragment (8-253 bp region in Fig. 5a) in lines that contained the Type 2 haplotype of the *rpl14-rpl16* intergenic sequence. (Fig. 5b and Table 3).

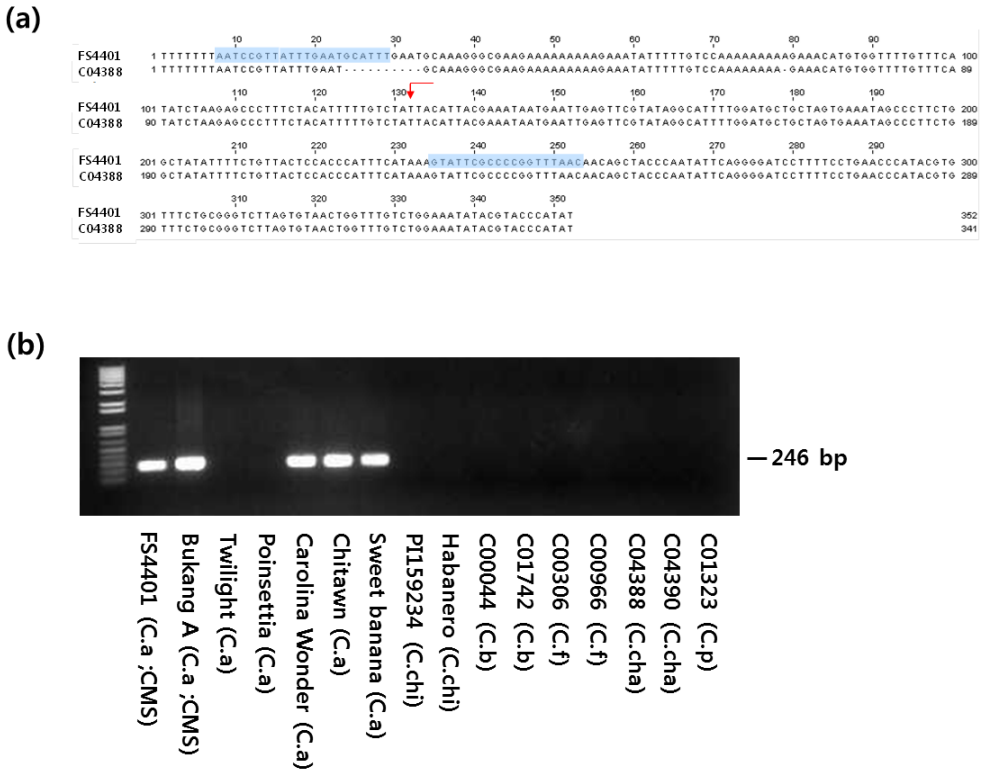


Figure 5. PepRpl marker development. (a) Comparison of FS4401 (*Capsicum annuum*) and C04388 (*Capsicum chacoense*) sequences that include the PepRpl sequence. Forward and reverse primer sequences are depicted by shadowboxes. Arrow indicates the 3' end of *rpl16*. (b) Amplification of PepRpl in pepper cultivars or accessions included in six *Capsicum* species (C.a CMS; *Capsicum annuum* CMS line, C.a; *Capsicum annuum*, C.chi; *Capsicum chinense*, C.b; *Capsicum baccatum*, C.f; *Capsicum frutescens*, C.cha; *Capsicum chacoense*, C.p; *Capsicum pubescens*).

Application of markers derived from plastid and mitochondria sequences to *Capsicum* germplasm

Two plastid genome sequence markers (PepTrn and PepRpl) and two markers from the mitochondrial genome sequence of CMS lines (*orf507* SCAR and *ψatp6-2* SCAR) were applied to 72 cultivars or accessions containing six *Capsicum* species (Table 4). PepTrn is a molecular marker based on the sequence variation on *trnL-trnF* intergenic region of pepper plastid genome. When this marker was applied to six pepper species (five cultivated species of *Capsicum* and *Capsicum chacoense*) in previous research, *Capcicum* species other than species in *C. annuum* complex (*C. annuum*, *C. chinense*, *C. frutescens*) could be identified according to species-specific haplotypes. In *C. annuum* complex, one subgroup of *C. annuum* accessions showed specific haplotype while most of others grouped together (Jeong et al., 2010). When this marker was applied to accessions in this study, 25 *C. annuum* accessions out of 34 *C. annuum* accessions showed Type 1 haplotype while other accessions showed different haplotypes except for one *C. frutescens* accession, C00657. When plastid-derived marker was applied (Jeong et al., 2010), 25 *C. annuum* lines that produced the PepRpl amplicon also showed the Type 1 haplotype for this marker (Table 4). Application of PepRpl resulted in the same classification pattern of *C. annuum* subgroups. All of the accessions in which PepRpl amplicon was generated were the same with the accessions which showed Type 1 haplotype for PepTrn. Amplicons were not detected in other

accessions except for C00657 which was also classified as Type 1 haplotype by PepTrn. To determine the relationship between the plastid-derived marker haplotypes and evolution of the CMS cytoplasm, two mitochondria-derived markers, *orf507* SCAR and *ψatp6-2* SCAR, which were known to be associated with CMS (Kim et al., 2007; Kim and Kim, 2006; Jo et al., 2009), were applied to the same panel of lines. For the *orf507* SCAR, amplicons were detected in some of lines in which amplicons were generated for PepRpl and the Type 1 haplotype was detected for PepTrn. The *ψatp6-2* marker gave rise to amplicons in only three lines that were also positive for the *orf507* SCAR marker. In the *C. frutescens* C00657 line, for which the PepRpl sequence was amplified, amplicons were also detected for *orf507* SCAR and *ψatp6-2* SCAR (Table 4).

Table 4. Summary of molecular marker analysis for pepper germplasm

Species	Cultivar or accession	Origin	PepTrn	PepRpl	*orf507*	*ψatp6-2*	Fruit type[y]
Capsicum	387-ONG	Malaysia	Type 2	-[z]	-	-	1
annuum	411 7593	China	Type 1	+	+	-	3
	Azeth	Brazil	Type 1	+	-	-	2
	Bird&quot's eye	Unknown	Type 2	-	-	-	1
	Calatuco	Argentina	Type 1	+	+	-	3
	Califonia Wonder	America	Type 1	+	+	-	3
	Carolina Wonder	America	Type 1	+	+	-	3
	Cassa dura ikeda	Unknown	Type 1	+	-	-	3
	Cayenne Chile	America	Type 1	+	-	-	2
	CHINDA2	Thailand	Type 1	+	-	-	2
	Chitawn	India	Type 1	+	+	-	3
	Chocolate Cherry	America	Type 1	+	-	-	3
	Cipanas	Indonesia	Type 1	+	-	-	1
	CO0799	Unknown	Type 2	-	-	-	1
	Criollo de Morelos 334	Mexico	Type 2	-	-	-	1
	EU016	Unknown	Type 1	+	+	+	2
	JMAV-C	Unknown	Type 1	+	-	-	1
	Lueng6	Thailand	Type 1	+	-	-	3
	Montego	Unknown	Type 1	+	+	-	3
	NARC-4	Pakistan	Type 1	+	+	-	1
	PBC458	Taiwan	Type 1	+	-	-	3
	PBC725	Papua New Guinea	Type 1	+	-	-	3
	Perennial HDV	India	Type 2	-	-	-	1
	PI 342948	America	Type 1	+	-	-	1
	PI 464	Israel	Type 1	+	-	-	3
	PI 467	Israel	Type 1	+	-	-	3
	Poinsettia	America	Type 2	-	-	-	1
	PP1993	Unknown	Type 2	-	-	-	1
	Starburst	America	Type 2	-	-	-	1
	Sweet Banana	America	Type 1	+	+	-	2
	Szeged1 csereszny e	Hungary	Type 1	+	-	-	3
	TCO6254	Uganda	Type 1	+	+	+	1
	Twilight	America	Type 2	-	-	-	1
	Yolo Wonder	America	Type 1	+	+	-	3
Capsicum	C00050	Costa Rica	Type 2	-	-	-	1
frutescens	C00065	Unknown	Type 2	-	-	-	1
	C00087	Costa Rica	Type 2	-	-	-	1
	C00088	Costa Rica	Type 2	-	-	-	1
	C00098	Thailand	Type 2	-	-	-	1

	C00309	Unknown	Type 3	-	-	-	1
	C00306	Unknown	Type 2	-	-	-	1
	C00657	Honduras	Type 1	+	+	+	1
	C01884	Unknown	Type 2	-	-	-	1
	C00966	Peru	Type 2	-	-	-	1
Capcicum	PI159234	USA	Type 2	-	-	-	1
chinense	Early red sweet	USA	Type 2	-	-	-	3
	Habanero	Mexico	Type 2	-	-	-	1
	Jalapeno	Mexico	Type 2	-	-	-	1
Capsicum	C00044	Costa Rica	Type 3	-	-	-	ND
baccatum	C01172	France	Type 3	-	-	-	ND
	C01218	Egypt	Type 3	-	-	-	ND
	C01220	Germany	Type 3	-	-	-	ND
	C01248	USA	Type 3	-	-	-	ND
	C01662	Paraguay	Type 3	-	-	-	ND
	C01686	USA	Type 3	-	-	-	ND
	C01692	Argentina	Type 3	-	-	-	ND
	C01742	Netherlands	Type 3	-	-	-	ND
	C02432	Unknown	Type 3	-	-	-	ND
	C02433	Peru	Type 3	-	-	-	ND
Capsicum	C04388	Argentina	Type 4	-	-	-	ND
chacoense	C04389	Argentina	Type 4	-	-	-	ND
	C04390	Bolivia	Type 4	-	-	-	ND
	C04391	Bolivia	Type 4	-	-	-	ND
	C04392	Bolivia	Type 4	-	-	-	ND
	C04395	Bolivia	Type 4	-	-	-	ND
	C04399	Argentina	Type 4	-	-	-	ND
	C04400	Bolivia	Type 4	-	-	-	ND
Capsicum	C01323	Guatemala	Type 5	-	-	-	ND
pubescens	C01324	Guatemala	Type 5	-	-	-	ND
	C01374	Peru	Type 5	-	-	-	ND
	C01572	Guatemala	Type 5	-	-	-	ND
	C04895	Ecuador	Type 5	-	-	-	ND

[z] Amplification (+) or non-amplification (-) of DNA fragment
[y] Fruit types were determined according to fruit shapes (1: small and short, 2: long and narrow, 3: blocky, ND: not determined

Application of markers derived from plastid and mitochondria sequences to CMS breeding lines

PepRpl, PepTrn, *orf507* SCAR and *ψatp6-2* SCAR were applied to breeding lines that are being used for hybrid production through the CGMS system (Table 5). In all investigated lines, PepRpl amplicons were generated and the PepTrn Type 1 haplotype was detected. The *orf507* SCAR and *ψatp6-2* SCAR amplifications were closely associated with the CMS cytoplasm. Except for one line, all CMS lines generated *orf507* SCAR and *ψatp6-2* SCAR amplicons (Table 5). In breeding lines with wild-type cytoplasm, the *orf507* SCAR marker generated amplicons for four lines while *ψatp6-2* SCAR generated amplicons in two lines that was positive for *orf507* SCAR.

Table 5. Summary of molecular marker analysis using *C. annuum* breeding lines

Line	Cytoplasm type [z]	PepTrn	PepRpl	*orf507* SCAR	*ψatp6-2* SCAR
S1	S	Type 1	+ [y]	+	+
S2	S	Type 1	+	+	+
S3	S	Type 1	+	+	+
S4	S	Type 1	+	+	+
S5	S	Type 1	+	+	+
S6	S	Type 1	+	+	+
L5	S	Type 1	+	+	+
L11	S	Type 1	+	+	+
L13	S	Type 1	+	+	+
L19	S	Type 1	+	-	-
L21	S	Type 1	+	+	+
L23	S	Type 1	+	+	+
L25	S	Type 1	+	+	+
L29	S	Type 1	+	+	+
L31	S	Type 1	+	+	+
L2	N	Type 1	+	-	-
L3	N	Type 1	+	-	-
L4	N	Type 1	+	-	-
L6	N	Type 1	+	+	+
L8	N	Type 1	+	-	-
L9	N	Type 1	+	+	-
L10	N	Type 1	+	-	-
L12	N	Type 1	+	-	-
L17	N	Type 1	+	-	-
L18	N	Type 1	+	-	-
L20	N	Type 1	+	-	-
L22	N	Type 1	+	-	-
L24	N	Type 1	+	-	-
L26	N	Type 1	+	-	-
L28	N	Type 1	+	-	-
L30	N	Type 1	+	-	-
C1	N	Type 1	+	-	-
C2	N	Type 1	+	-	-
C3	N	Type 1	+	+	-
C4	N	Type 1	+	-	-
C5	N	Type 1	+	-	-
C6	N	Type 1	+	-	-
C7	N	Type 1	+	-	-
C8	N	Type 1	+	-	-
C9	N	Type 1	+	-	-
C10	N	Type 1	+	+	+

[z] S: sterile cytoplasm, N: fertile cytoplasm

[Y] Amplification (+) or non-amplification (-) of DNA fragment

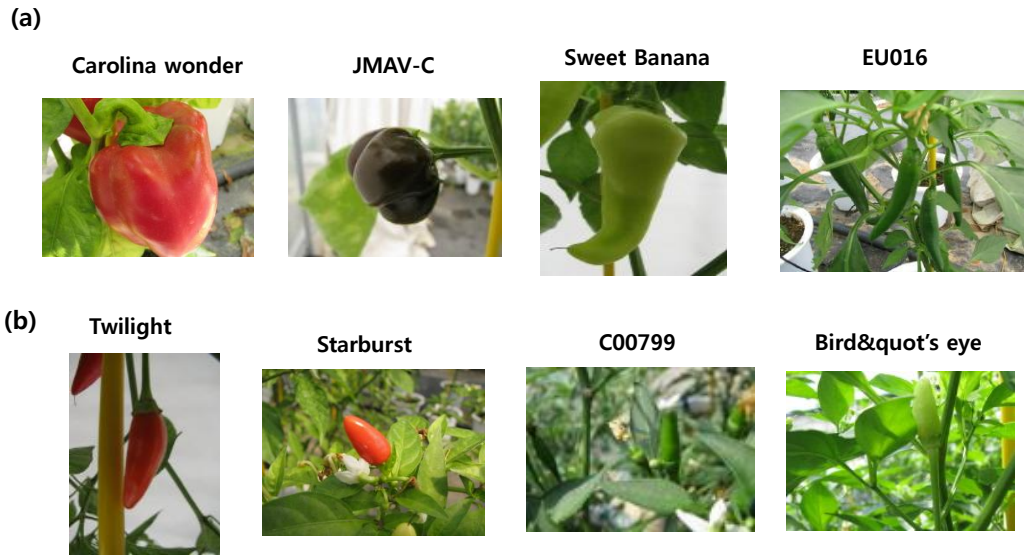


Figure 6. Correlation between fruit shape and genotype for the PepRpl marker. Fruit shapes were investigated for groups in which amplicons were (a) generated or (b) not generated when PepRpl was applied.

DISCUSSION

CMS has been known to be caused by wide crosses (Hanson and Bentolila, 2004). In pepper, one CMS cytoplasm has been reported and introgressed *C. annuum* lines, but the species from which this cytoplasm originated has not been determined. To deduce the candidates for species of cytoplasm donor, we compared plastid DNA sequences between CMS lines and germplasm in six species of *Capsicum* based on assumptions that plastid genome does not undergo intensive modification during interspecific cross (Palmer, 1990) and contains sequences which are specific to a given species (below).

Plastid DNA regions defined as barcode sequences have been widely exploited for the classification of plant species because gene sequences and gene order are highly conserved and provide common priming sites, while some gene sequences and intergenic regions contain sufficient variation to distinguish species (Taberlet et al., 2007). The candidate DNA barcode sequences include *rbcL* (Kress and Erickson et al., 2007), *rpoc1* (Chase et al., 2007), *matK* (Lahaye et al., 2007) as well as intergenic or intron sequences including *trnH-psbA* (Kress and Erickson et al., 2007), *trnL-trnF* (Taberlet et al., 1991) and *trnL* intron sequences (Taberlet at al., 2007). These barcode sequences have been shown to contain different levels of variation, thereby each sequence has its own power and limitations in terms of species classification (Taberlet at al., 1991).

In *Capsicum*, Jarret (2008) investigated variations in eight regions of plastid DNA and proposed that *trnH-psbA* or *trnL-trnT* could be used as barcode sequences in addition to the nuclear *waxy* gene. In this study, *trnH-psbA* and *rpl14-rpl16* intergenic sequences were investigated in a panel of *Capsicum* germplasms including six *Capsicum* species. The *trnH-*

psbA and *rpl14-rpl16* intergenic regions showed different levels of variation among accessions (Table 3). The *trnH-psbA* sequence discriminated *C. annuum*, *C. chacoense* and *C. pubescens* from other species. However, accessions in *C. frutescens* and *C. chinense* could not be separated from each other and *C. baccatum* could not also be precisely distinguished from *C. frutescens*. The *rpl14-rpl16* showed less variation among *Capsicum* species; only *C. chacoense* and *C. pubescens* had species-specific haplotype while a clade of *C. annuum*, *C. frutescens*, *C. chinense* and *C. baccatum* shared the same haplotype (Type 1). Among these, *C. annuum*, *C. frutescens* and *C. chinense* are taxonomically close and constitute the *C. annuum* complex (Pickersgill, 1997). Although the power of species classification was different between the two sequences, both sequences subdivided *C. annuum* into two clades (Table 3).

Jarret (2008) also reported that a group of *C. annuum* lines did not contain a 7 bp insertion in the *trnH-psbA* sequence, which is specific for *C. annuum*, while these lines contain a 13 bp insertion that is usually detected in *C. frutescens* or *C. chinense* in the *trnH-psbA* sequence. Many *C. annuum* lines in this group were included in the *C. annuum* var. *glabriuseulum* group instead of *C. annuum* var. *annuum*. *C. annuum* var. *glabriuseulum* is regarded as an ancestor of domesticated *C. annuum* lines and usually bear upright, small, short fruits. PepTrn and PepRpl also designated a similar pattern in the classification of germplasm in this study (Table 4; Fig.6). Many *C. annuum* lines grouped with *C. frutescens* or *C. chinense* accessions due to the absence of the PepRpl amplicon, which also bear upright, small, short fruits (Fig. 6). In contrast, none of the breeding lines belonged to this clade (Table 5). Therefore, the *C. annuum*-specific plastid sequences detected in a clade of *C.*

annuum (Type 2) may have evolved from corresponding sequences in another clade during domestication of *C. annuum* (Type 1).

Although several interspecific crosses between *Capsicum* species induced CMS in their progeny (Andrasfalvy and Csillery, 1983; Kumar et al., 2009; Yoo, 1990), restoration by the introgression of the *Rf* gene for wild-type CMS cytoplasm was detected only in one cross combination; *C. frutescens* X *C. annuum* (Yoo, 1990). Therefore, the cytoplasm that originated from this interspecific cross was considered to be identical to that of PI164835. In other words, the cytoplasm of PI164835 may have originated from *C. frutescens* by the *C. frutescens* X *C. annuum* cross.

However, the application of cpDNA-derived markers in CMS lines and diverse pepper germplasm showed that the present CMS cytoplasm appears to have originated from a cross in which a *C. annuum* accession served as the female parent when we regard maternal inheritance of mitochondria and chloroplast because plastid-derived haplotype markers in all CMS lines were identical to lines that belonged to one of the *C. annuum* clades (Table 4, Table 5). This indicates that the CMS-specific mitochondrial gene might have arisen by the structural alteration of the *C. annuum* mitochondrial genome rather than introgression of cytoplasm from another species. Recent studies on Arabidopsis mitochondrial DNA showed that species in the genus Arabidopsis might share common mitochondrial sequences although the substoichiometric ratio between subgenomic molecules was very different (Arrieta-Montiel et al., 2009). In this research, the pattern of DNA rearrangement changed remarkably when a mutation was induced in a nuclear gene related to recombination (*msh1*). Other studies of the CMS cytoplasm in radish showed that DNA sequences specific to CMS

cytoplasm exist even in maintainer lines, although the stoichiometry of the sequence was maintained in a very low level. Similarly in pepper, DNA structures specific for CMS cytoplasm (*orf507* and *ψatp6-2*) were detected even in maintainer lines with very low copy number (Jo et al., 2009; Min, 2009). This implies that subgenomic DNA molecules responsible for CMS-specific structures may be maintained in very low copy number in wild-type cytoplasm. Therefore, the CMS cytoplasm of pepper might have originated from a failure in the copy number suppression of CMS-specific subgenomic structures, possibly due to the alteration of nuclear control caused by mutation or interspecific cross. If the change in nuclear control of mitochondrial genome is more important than the specificity of the original mitochondrial DNA for induction of CMS, the speculation that the current CMS cytoplasm may have originated from *C. annuum* may not contradict a previous report that CMS cytoplasm was induced by a *C. frutescens* X *C. annuum* cross (Yoo, 1990). We cannot rule out the possibility that the current CMS cytoplasm originated from a specific clade of *C. frutescens* because *C. frutescens* C00657 showed the same plastid marker haplotypes as those of CMS lines (Table 4). Crosses between C00657 and CMS (or maintainer lines) should be performed for future study. Also, the possible paternal leakage in the inheritance of mitochondria or chloroplast may result in the exceptional case in which a plant contains organellar DNAs from both parental lines. Although functional mitochondria and chloroplast from paternal parent are excluded or degenerated during generation of generative cells and sperm cells or fusion of gametes, small portion of organelles may be descended to progeny from paternal parent if paternal leakage occurs (Maliga et al., 2007; Mogensen, 1996). Therefore, the possibility that mitochondria and chloroplast originated from different parental

lines during the evolution of pepper CMS cytoplasm cannot be excluded completely.

Application of two plastid and two mitochondrial markers to germplasm and breeding lines showed that the cytoplasm of *C. annuum* can be classified into at least four groups; group 1 containing *orf507, ψatp6-2* and PepRpl amplicon sequences, group 2 containing *orf507* and PepRpl amplicon sequences, group 3 containing only the PepRpl amplicon sequence, and group 4 containing none of *orf507, ψatp6-2* and PepRpl amplicon sequences (Table 3). Most CMS lines contained all *orf507, ψatp6-2* and PepRpl amplicon sequences while the other clade of *C. annuum* lines and most of the *C. frutescens* and *C. chinense* lines contained none of these sequences. Therefore, the evolution of CMS pepper lines involved at least two steps: 1) evolution of the *C. annuum* clade receiving the PepRpl sequence from the wild-type *C. annuum* cytoplasm and 2) proliferation of subgenomic mitochondrial DNA molecules containing *orf507* and *ψatp6-2* sequences from the *C. annuum* clade containing the PepRpl sequence. Crosses of *C. annuum* accessions at different evolutionary stages in diverse germplasms may confirm the origin of the CMS cytoplasm.

REFERENCES

Andrasfalvy A, Csillery G (1983) Cytoplasmic systems of interspecific hybrids in *Capsicum*, reconsidered. Eucarpia Vth Meeting on Genetics and Breeding of Capsicum and Eggplant (Plovdiv Bulgaria). P18-20

Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA (2009) Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics 183: 1261-8

Azhagiri AK, Maliga P (2007) Exceptional paternal inheritance of plastids in Arabidopsis suggests that low-frequency leakage of plastids via pollen may be universal in plants. Plant J 52: 817-23

Hanson MR, Bentolila S (2004) Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell 16 Suppl: S154-69

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-77

Jarret RL (2008) DNA barcoding in a crop genebank: the *Capsicum annuum* species complex. Open Biol J 1: 35-42

Janska H, Sarria R, Woloszynska M, Arrieta-Montiel M, Mackenzie SA (1998) Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. Plant Cell 10: 1163-80

Jeong HJ, Jo YD, Kang BC (2010) Identification of Capsicum species using SNP markers based on high resolution melting analysis. Genome 53: 1029-40

Jo YD, Jeong HJ, Kang BC (2009) Development of a CMS-specific marker based on chloroplast-derived mitochondrial sequence in pepper. Plant Biotechnol Rep 3: 309-15

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 Rep 30: 217-29

Kim DH, Kang JG, Kim BD (2007) Isolation and characterization of the cytoplasmic male sterility-associated orf456 gene of chili pepper (*Capsicum annuum* L.). Plant Mol Biol 63: 519-32

Kim DH, Kang JG, Kim BD (2007) Isolation and characterization of the cytoplasmic male sterility-associated orf456 gene of chili pepper (*Capsicum annuum* L.). Plant Mol Biol 63: 519-32

Kim DH, Kim BD (2006) The organization of mitochondrial *atp6* gene region in male fertile and CMS lines of pepper (Capsicum annuum L.). Curr Genet 49: 59-67

Kim DH, Kim BD (2005) Development of SCAR markers for early identification of cytoplasmic male sterility genotype in chili pepper (*Capsicum annuum* L.). Mol Cells 20: 416-22

Kim S, Lee YP, Lim H, Ahn Y, Sung SK (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: 189-98

Kim S, Lim H, Park S, Cho KH, Sung SK, Oh DG, Kim KT (2007) Identification of a novel mitochondrial genome type and development of molecular markers for cytoplasm classification in radish (*Raphanus sativus* L.). Theor Appl Genet 115: 1137-45

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding *trnH-psbA* spacer region. PLoS One 2: e508

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA

barcodes to identify flowering plants. Proc Natl Acad Sci U S A 102: 8369-74

Kumar R, Kumar S, Dwivedi N, Kumar S, Rai A, Singh M, Yadav DS and Rai M (2009) Genetics and distribution of fertility restoration associated RAPD markers in inbreds of pepper (*Capsicum annuum L*.) Scientia Hort 120: 167-172

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci U S A 105: 2923-8

Mackenzie S, McIntosh L (1999) Higher plant mitochondria. Plant Cell 11: 571-86

Min W (2009) Molecular genetic analysis and allelic discrimination of the Restorer-of-fertility (*Rf*) gene in peppers (*Capsicum annuum* L.). Ph.D. thesis, Seoul National University

Mogensen HL (1996) The hows and whys of cytoplasmic inheritance in seed plants. Amer J Bot 83: 383–404

Palmer JD (1990) Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet 6: 115-20

Peterson PA (1958) Cytoplasmically inherited male sterility in *Capsicum*. Amer Nat 92: 111-9

Pickersgill B (1997) Genetic resources and breeding of *Capsicum* spp. Euphytica 96: 129-33

Prince JP, Zhang W, Radwanski ER, Kyle MM (1997) A versatile and high-yielding protocol for the preparation of genomic DNA from *Capsicum* spp. (pepper). Hortiscience 32: 937-9

Schnable PS, Wise RP (1998) The molecular basis of cytoplasmic male sterility

and fertility restoration. Trends Plant Sci 3: 175-80

Shifriss C (1997) Male sterility in pepper (*Capsicum annuum* L.). Euphytica 93: 83-8

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M (2005) The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genomics 272: 603-15

Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35:e14

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol Biol 17: 1105-9

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-5

Yu IW (1990) The inheritance of male sterility and its utilization for breeding in pepper (*Capsicum annuum* L.). Kyung Hee University, South Korea. PhD thesis, p.1-70

Yukawa M, Tsudzuki T, Sugiura M (2006) The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of *Nicotiana tabacum*. Mol Genet Genomics 275: 367-73

초 록

세포질 유전자적 웅성불임(CGMS)은 핵-미토콘드리아 유전체 간의 상호 작용에 기인하는 현상으로서 고추를 포함한 다양한 작물의 잡종 종자 생산에 널리 이용되어 왔다. 고추에서는 웅성불임 후보 유전자가 동정되었고 회복 유전자와 가깝게 연간되어 있는 분자표지들이 개발되어 왔으나 세포질 유전자적 웅성불임의 기작을 이해하고 신뢰할 수 있는 분자 육종 시스템을 구축하는 데에는 한계가 있었다. 따라서 본 연구에서는 고추 웅성불임 및 가임 미토콘드리아 유전체의 비교 분석, 웅성불임 회복 후보 유전자 동정, 엽록체 유전체 서열 기반 분자 표지를 이용한 웅성불임 세포질의 기원에 대한 추론을 수행하였다.

첫 번째 장에서는 웅성불임 계통인 FS4401과 가임 계통인 Jeju의 미토콘드리아 유전체 서열을 비교 분석하였다. 각각 507,450bp 및 493,911bp의 길이를 지닌 미토콘드리아 유전체 전체 서열이 FS4401과 Jeju에서 분석되었다. 유전체 상의 대부분의 유전자 서열은 양 계통 간 잘 보존되어 있었으나 유전체 간 많은 재배열이 일어나 양 계통 간 상호 배열될 수 있는 18개의 서열 단위 (>2kb, >95% 상동성)로 유전체가 나뉘어졌으며 각 단위 사이에는 양 계통에 특이적인 서열이 위치하는 것으로 확인되었다. 웅성불임 후보 유전자인 *orf507* 과 *ψatp6-2*의 경우 FS4401 특이 서열 중 가장 큰 서열 단위들의 가장 자리에 위치하였다. *orf507* 과 *ψatp6-2* 인근 서열에서 극심한 재배열이 일어났고 검색이 가능한 서열 중 이 부위에 상응될 수 있는 서열이 없었으므로 구체적인 생성 기작을 논하기는 어려웠으나 해당 부위에 존재하는

DNA 조각들에서 나타나는 반복 서열 및 상호 중첩 서열의 존재로 미루어 볼 때 비상동적 말단 결합 과정(nonhomologous end-joining)을 통해 심한 재배열이 일어나고 그 이후 반복서열을 통해 재조합이 일어나며 웅성불임 후보 유전자를 포함하는 부분단위 유전체 분자(subgenomic molecule)의 양적 변화가 일어나는 과정을 통해 이 부위가 생성되고 주염색체 안으로 삽입되었을 것으로 추정되었다. 웅성불임 및 가임 계통의 미토콘드리아 서열이 확보된 다른 작물에서의 추가 분석을 통하여 미토콘드리아 유전체 내에서 웅성불임 유전자가 갖는 위치적 특성의 공통성이 도출될 수 있었다.

두 번째 장에서는 웅성불임 회복 유전자를 동정하기 위해 세 가지의 유전자 지도 작성 방법이 이용되었다. 첫째, 페튜니아의 회복 유전자와 상동성이 높은 고추 서열을 탐침으로 사용하여 BAC 클론을 선별하고 분류 및 분자 표지 개발을 수행하였다. 둘째, 1,000가지 이상의 프라이머 조합을 이용하여 AFLP를 수행하였다. 마지막으로 토마토 유전체 서열과의 비교 분석을 통하여 유전자 지도를 작성하였다. 그 결과 개발된 일부 분자 표지들이 회복 유전자와 공동 분리되는 DNA 서열상에 위치함이 확인되었다. 이 서열로부터 여섯 차례의 염색체 워킹(chromosome walking)을 수행하여 회복 유전자좌를 포괄하는 서열을 확보하였다. 전사체 분석 결과에 기반하여 해당 서열 중 약에서 발현될 가능성이 있는 유전자를 판별해 본 결과 회복 후보 유전자로 *PPR6* 유전자가 선별되었다. *PPR6* 유전자는 35개의 아미노산으로 이루어진 모티프가 14회 반복되는 PPR (pentatricopeptide repeat) 단백질을 암호화하는 것으로 나타났다. 회복 계통에서 특이적으로 나타나는 발현 양상으로 미루어 볼 때 *PPR6*는 유력한 회복 후보 유전자로 볼 수 있었다.

세 번째 장에서는 엽록체 유전체에서 개발된 분자 표지를 활용하여 웅성불임 세포질의 기원을 추론하였다. FS4401 계통의 엽록체 유전체의 전체 염기 서열이 분석되었으며 이는 분자 표지 개발을 위해 활용되었다. *trnH-psbA* 및 *rpl16-rpl18* 각각의 유전자 간 서열이 *Capsicum* 속 내의 6가지 종에 속하는 고추 유전자원에서 분석되었으며 각 서열은 각각 6가지, 4가지 타입으로 분류되었다. CMS 계통에서 분석된 해당 부위 서열은 *Capsicum annuum* 종의 특정 분류군에서 나타나는 서열과 일치하였다. 추가 분석을 위해 *trnL-trnF* 및 *rpl16-rpl18* 각각의 유전자 간 서열에서 분자표지를 개발하여 다수의 유전자원에 적용하여 보았다. 그 결과 웅성불임의 세포질형은 *Capsicum annuum* 종의 특정 분류군에서 나타나는 서열과 역시 일치하는 것으로 확인되었다. 이상의 결과는 고추의 웅성불임 세포질이 *Capsicum annuum* 종을 모친으로 하는 교잡에서 유래하였을 가능성이 높음을 시사한다.

본 연구의 이상의 성과는 고추 세포질 유전자적 웅성불임 체계에서 신뢰도가 높은 분자 육종 시스템을 구축하는데 기여할 뿐 아니라 고추의 웅성불임 세포질 및 회복 유전자의 진화를 연구하기 위한 기초 자료로서 가치가 있을 것으로 판단된다.

주요어: 고추(*Capsicum annuum*), 세포질적 웅성불임, 웅성불임 회복 유전자, 미토콘드리아, 엽록체

감사의 글

박사 과정 중에 깊은 관심과 애정으로 지도해 주시고 늘 열정적인 모습으로 본이 되어주셨던 강병철 지도 교수님께 깊은 감사를 드립니다. 바쁘신 와중에도 논문을 검토해 주시고 조언과 격려를 통해 나아갈 방향을 제시해 주셨던 최도일 교수님, 양태진 교수님, 김성길 교수님, Peter van Dijk 박사님 등 심사의원 선생님들과 학부 입학 때부터 박사 졸업 때까지 지켜보시며 지도해 주셨던 박효근 교수님, 김병동 교수님, 이승구 교수님, 김기선 교수님, 손정익 교수님, 이희재 교수님, 전창후 교수님, 허진회 교수님, 이은진 교수님 등 원예과학 전공 교수님들께도 감사의 마음을 전해드립니다. 특히 학자로서 큰 뜻을 품도록 일깨워주시고 더 넓은 시각을 갖도록 늘 지도해주시는 김병동 교수님께 존경과 감사를 드립니다.

연구에서나 학교 생활에서나 큰 힘이 되어주었던 선배님들과 동료분들께도 감사의 인사를 드립니다. 항상 마음이 편해지는 밝은 웃음을 주고 진심어린 격려를 해 주었던 원예작물육종학 연구실 식구들이 있었기에 여기까지 올 수 있었습니다.

늘 곁에서 힘이 되어 주었던 가족들과 사랑하는 분들께도 인사를 드립니다. 어떤 상황에서도 믿고 기댈 수 있는 버팀목이 되어 주셨던 아버지, 제대로 한 번 챙겨드리지도 못했는데도 저를 항상 먼저 걱정하시고 보살펴주셨던 어머니, 해준 것 없는 오빠를 챙겨주려 애쓰는 속깊은 동생 수연이에게도 미안한 마음과 함께 깊은 감사를 드립니다. 어렸을 때 저를 사랑과 희생으로 보살펴 주신 이모와 제가 학문의 기초를 닦을 수 있도록 스승이 되어 지도해 주신 외삼촌께도 고개 숙여 감사드립니다. 힘든 기간 동안 곁에서 늘 함께 하며 의지가 되어주고 긍정의 힘을 주었던 지현이에게도 따뜻한 고마움을 전합니다. 마지막으로 어린 시절의 저를 정성으로 보살펴 주시고 이후에도 매일 새벽 저를 위해 간절히 기도해 주셨던 외할아버지께 뒤늦게 이 논문을 드립니다.