Genetics and Molecular Research

Development of primer pairs from diverse chloroplast genomes for use in plant phylogenetic research

Y.C. Yang[1, 2], T.L. Kung[1, 3], C.Y. Hu[4] and S.F. Lin[1]

[1]Department of Agronomy, National Taiwan University, Taipei, Taiwan, Republic of China
[2]Central Region Branch, Agriculture and Food Agency, Changhwa, Taiwan, Republic of China
[3]Taoyuan District Agricultural Research and Extension Station, Council of Agriculture, Taoyuan, Taiwan, Republic of China
[4]Wenshan Branch, Tea Research and Extension Station, Taipei, Taiwan, Republic of China

Corresponding author: S.F. Lin
E-mail: shunfu@ntu.edu.tw

Genet. Mol. Res. 14 (4): 14857-14870 (2015)
Received May 20, 2015
Accepted July 22, 2015
Published November 18, 2015
DOI http://dx.doi.org/10.4238/2015.November.18.51

ABSTRACT. Variation in the chloroplast DNA sequence is useful for plant phylogenetic studies. However, the number of variable sequences provided by chloroplast DNA for suggested genes or genomic regions in plant phylogenetic analyses is often inadequate. To identify conserved regions that can be used to design primers and amplify variable sequences for use in plant phylogenetic studies, the complete chloroplast genomic sequences of six plant species (including *Oryza sativa*, *Arabidopsis thaliana*, *Glycine max*, *Lotus japonicus*, *Medicago truncatula*, and *Phaseolus vulgaris*), searched from the taxonomy database of NCBI were investigated. A total of 93 conserved regions, 32 in large single copy and 61 in inverted repeat regions, were identified. A set of five primer pairs were designed according to the conserved sequences located in the *psbA~trnK*, *psbB~psbH*, *rpl23~trnI*, *trnR~trnN,* and *trnY~trnD* regions to amplify variable DNA fragments. An additional 18 plant accessions from 14 species were used to validate

their utility. Each of the tested species could be distinguished by length polymorphisms of fragments amplified with the five primer pairs. *trnR~trnN* and *rpl23~trnI* amplified fragments specific to monocot and legume species, respectively. Three primer pairs located in the *psbA~trnK*, *psbB~psbH,* and *trnR~trnN* regions were applied to amplify variable DNA sequences for phylogenetic analysis using the maximum parsimony method. The consistent result between taxonomy and phylogenetic analysis on the variable sequences amplified with these three primer pairs was revealed. The five newly developed primer pairs are recommended as tools for use in the identification of plant species and in phylogenetic studies.

Key words: Chloroplast genome; Conserved sequence; Universal primer; Phylogenetic analysis; Species identification; Variable sequence

INTRODUCTION

In recent years, the rapid development of DNA sequencing technologies has led to an increase in the number of genomic studies performed (Edwards et al., 2013). Comparative genomics compares different genomes, investigates the implications of any differences or similarities, and provides explanations for species differences and evolutionary hypotheses (De Las Rivas et al., 2002; Jiao and Guo, 2014).

The evolution of chloroplast DNA is relatively conservative and slow, with an average evolution speed four times slower than that of nuclear DNA in plants (Wolfe et al., 1987). Therefore, chloroplast DNA is suitable for studies on comparative genomics. In addition, the speed of chloroplast evolution differs depending on its location within the genome. Typically, the evolution or mutation speed of large single copy (LSC) and small single copy (SSC) regions is faster than that of the inverted repeat (IR) regions (Curtis and Clegg, 1984; Wolfe et al., 1987). The nucleotide substitution speed in the LSC and SSC regions is approximately 2.3 times that in the IR regions (Perry and Wolfe, 2002). The least conservative region in the chloroplast genome is located in the LSC region (Ravi et al., 2008).

Many studies have indicated that chloroplast DNA contains a great deal of genetic variation between species and populations (Wojciechowski et al., 2004; Byrne and Hankinson, 2012; Yang et al., 2013). Non-coding areas of SC regions are frequently used in phylogenetic analysis of lower plants (Clegg et al., 1994). A representative study was reported by Saski et al. (2007), who compared variable sequences in chloroplast DNA from *Hordeum vulgare*, *Sorghum bicolor*, and *Agrostis stolonifera* to assess the phylogeny among species.

The chloroplast genome is substantially smaller than the nucleus genome. However, the number of publicly announced whole genome sequences of higher plants or crop species remains extremely limited (Gao et al., 2010). Although conducting whole genome sequencing for each species or individuals can provide sufficient and accurate information, a significant amount of time and funding is required. Because the amount of variation provided by the DNA sequences of ITS (internal transcribed spacer), IGS (intergenic spacer), *rbcL* (large subunit of ribulose 1,5-bisphosphate carboxylase), and *matK* (maturase K) in plant phylogeny analysis is often inadequate, these sequences cannot be used to reveal the relationships between some species (Soltis and Soltis, 1998; Selvaraj et al., 2013). Demesure et al. (1995) developed universal primers based on conservative

sequences of chloroplast genomes from three land plants. Subsequently, Dumolin-Lapègue et al. (1997) expanded a set of primers for use in the study of population genetics or low taxonomic levels of plant, and verified these primer pairs in five plant species (including three angiosperms). However, the available primer pairs remained insufficient for plant identification or phylogenetic analysis. In this study, we analyzed the complete chloroplast DNA sequences of six diverse plant species. A set of five primer pairs was designed according to two adjacent conserved regions. The effectiveness of these primer pairs was evaluated by testing on an additional 18 plant accessions.

MATERIAL AND METHODS

Chloroplast DNA sequence search and alignment

The complete chloroplast genomic sequences of six plant species including *Arabidopsis thaliana* (arabidopsis), *Lotus japonicus* (Japanese trefoil), *Medicago truncatula* (barrel medic), *Glycine max* (soybean), *Phaseolus vulgaris* (common bean), and *Oryza sativa* (rice) were searched in the public taxonomy database of the National Center for Biotechnology Information, U. S. National Library of Medicine) (NCBI, Table 1). Apart from *M. truncatula*, the referenced papers describing chloroplast genomes for all investigated species were available (Table 1). *O. sativa*, *L. japonicus,* and *A. thaliana*, represent monocot, dicot, and legume model plants, respectively. In the taxonomy database, *O. sativa* can be divided into *O. sativa* ssp. *indica* and *O. sativa* ssp. *japonica*. Because the DNA sequence of *O. sativa* ssp *japonica* was annotated to a greater extent, it was chosen for investigation.

Table 1. Coverage of coding region, conserved region length, and GC content in chloroplast genomes of six investigated species.

Species	GenBank accession No.	Gene No.	Percentage of coding region (%)	Whole genome		Conserved region			Non-conserved region			Reference
				Length (bp)	GC content (%)	Length (bp)	Genome coverage (%)	GC content (%)	Length (bp)	Genome coverage (%)	GC content (%)	
Arabidopsis thaliana	NC_000932	129	51	154,478	36.29	13,265	8.59	45.93	141,213	91.41	33.39	Sato et al. (1999)
Lotus japonicus	NC_002694	128	51	134,525	36.03	13,264	8.81	45.93	137,255	91.19	35.07	Kato et al. (2000)
Medicago truncatula	NC_003119	109	52	150,519	33.97	13,268	10.7	45.93	110,765	89.3	32.54	not available
Glycine max	NC_007942	128	50	124,033	35.37	13,264	8.71	45.93	138,954	91.29	34.37	Saski et al. (2005)
Phaseolus vulgaris	NC_009259	129	51	152,218	35.44	13,265	8.83	45.93	137,020	91.17	34.43	Guo et al. (2007)
Oryza sativa	NC_001320	159	48	150,285	38.99	13,268	9.86	45.94	121,257	90.14	38.23	Hiratsuka et al. (1989)
Average					36.02	13,266	9.25	45.93	131,077	90.81	34.67	

The complete chloroplast genome sequences of the six plant species were downloaded from the GenBank sequence database provided by the NCBI. The ClustalW Multiple alignment function in the version 7.0.5.2 of the BioEdit software (Hall, 1999) was used to perform pairwise alignment of the sequences.

Identification of conserved regions

The Find Conserved Regions function was used to differentiate the conserved regions

from the non-conserved regions in the sequence data. Settings in the BioEdit function adopted in this study were as follows: the minimum length of the conserved region was 60 bp, the maximum average entropy of the similarity matrix was 0.2, the maximum entropy of each position was 0.8, the maximum gap in the conserved region was 1 bp, and the maximum length of the fragment inserted into the conserved region causing gaps was 2 bp.

Design primers based on conserved sequences

In this study, primer pairs based on the conserved regions were designed and used to amplify DNA fragments of approximately 800 bp in one run of DNA sequencing. Because of the presence of unstable signals at both ends (about 50 bp in length) of the sequenced data, the reliable sequenced data of the central region was approximately 700 bp.

Application of designed primer pairs in other species

To test the transfer use efficiency of the designed primer pairs, a total of 14 plant species (18 accessions) were selected. The species analyzed in this study included *Pisum sativum, Arachis hypogaea, Vigna radiata, Glycine max, Vigna unguiculata, Vigna angularis, Vicia faba, Cucumis melo, Camellia sinensis, Solanum tuberosum, O. sativa, Zea mays, Euchlaena mexicana*, and *Coix lacryma-jobi* (Table 2). Both *G. max* and *O. sativa* included two and three varieties, respectively. Two subspecies of *V. unguiculata*, *cylindrica,* and *sesquipedalis*, were included. Seeds of the 18 plant accessions tested in this study were provided from the breeding stations or research institutes in Taiwan (Table 2).

Table 2. List of 18 accessions from 14 crop species used to test five primer pairs developed in this study.

Accession code	Accession	Source of accession
Ps	*Pisum sativum* cv. TC12	Taichung DARES, Taiwan
Ah	*Arachis hypogaea* cv. TN12	Tainan DARES, Taiwan
Vr	*Vigna radiata* cv. TN3	Tainan DARES, Taiwan
Gm1	*Glycine max* cv. KVS5	Kaohsiung DARES, Taiwan
Gm2	*Glycine max* cv. KVS8	Kaohsiung DARES, Taiwan
Vu1	*Vigna unguiculata* ssp *cylindrica* cv. White Bean	National Taiwan University, Taiwan
Vu2	*Vigna unguiculata* ssp *sesquipedalis* cv. KS Green Pod	Kaohsiung DARES, Taiwan
Va	*Vigna angularis* cv. KS5	Kaohsiung DARES, Taiwan
Vf	*Vicia faba* cv. NTU-VF03	National Taiwan University, Taiwan
Cm	*Cucumis melo* cv. TWAA	Tainan DARES, Taiwan
Cs	*Camellia sinensis* cv. Ching-Shin-Oolong	TRES, Taiwan
St	*Solanum tuberosum* cv. Kennebec	SIPS, Taiwan
Os1	*Oryza sativa* ssp *Japonica* cv. TN67	TARI, Taiwan
Os2	*Oryza sativa* ssp *Japonica* cv. TK9	TARI, Taiwan
Os3	*Oryza sativa* ssp *Indica* cv. TCS10	TARI, Taiwan
Zm	*Zea mays* cv. TN442a	Tainan DARES, Taiwan
Em	*Euchlaena mexicana* cv. Teo-3	National Taiwan University, Taiwan
Cl	*Coix lacryma-jobi* cv. Job-1	Taichung DARES, Taiwan

DARES = District Agricultural Research and Extension Station; TRES = Tea Research and Extension Station; SIPS = Seed Improvement and Propagation Station; TARI = Taiwan Agricultural Research Institute.

PCR and DNA sequencing

Total DNA was extracted from fresh and healthy leaves according to the CTAB method presented by Doyle and Doyle (1990). DNA samples from the 14 plant species were used as

templates for the polymerase chain reaction (PCR), and five designed primer pairs were employed to amplify the target variable regions. The PCR products were loaded on a 2% LE (low electroendosmosis) agarose gel and the results of electrophoresis were recorded as images. The bands on the electrophoretic gel were removed for DNA sequencing. The DNA fragments were sequenced using Applied Biosystems 96-capillary 3730xl DNA analyzer by Tri-I Biotech.

Phylogenetic analysis based on DNA sequence

After each targeted variable region was completely sequenced, the BioEdit program version 7.0.5.2 was employed to remove two unstable ends of the sequence. The Kimura 2-parameter model was used to calculate the relative genetic distance between species based on the DNA sequence data (Kimura, 1980). The maximum parsimony method was adopted to establish the phylogenetic tree, and the sample bootstrap method was used to perform repeated sampling, 1000 times, to estimate the reliability of the constructed phylogenetic tree.

RESULTS

Identification of chloroplast conserved regions

The chloroplast genome lengths of the six species investigated in this study ranged from 124,033 bp (*M. truncatula*) to 154,478 bp (*A. thaliana*). Excluding *japonica* rice (monocot) and *M. truncatula* (dicot), which had lengths of 134,525 and 124,033 bp, respectively, the remaining dicot plants had sequences that varied significantly, with lengths greater than 150 kb (Table 1). The number of genes in each genome ranged from 109 (*M. truncatula*) to 159 (*japonica* rice). Except for those, the remaining chloroplast genomes had 128 or 129 genes. The proportion of the entire genome comprising the coding region ranged from 48% (*O. sativa* ssp *Japonica*) to 52% (*M. truncatula*). Apart from *O. sativa* ssp *japonica* (monocot), the coding regions of the dicot plants all comprised more than 50% of the total sequence (Table 1).

As the maximum entropy values increased, the number of identified conserved regions, their length, the coverage of conserved regions, and the average length of single conserved regions also increased (Table 3), this study set the maximum entropy of each position as 0.8 (that is, at the same position of every base, two of six species differed from the others and were thus considered as not conservative) in order to design primers with better specificity than the conserved sequences. Under these settings, a total of 93 conserved regions were identified (Table 4). The lengths of the conserved regions accounted for approximately 9.25% of the entire chloroplast genome (Table 1).

Table 3. Number and average length of conserved regions in chloroplast genomes identified with different entropies.

	Maximum entropy per position							
	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
Number of conserved region	51	51	93	93	106	106	89	89
Coverage of conserved region (%)	3.5	3.5	9.2	9.2	11.7	11.7	14.2	14.2
Average length of conserved region (bp)	100	100	143	143	160	160	230	230

Table 4. Distribution and length of 93 conserved regions identified from six whole chloroplast genomes.

	Large single copy (LSC) region		Inverted repeat (IR) region			Total
	Within gene	Between gene and spacer	Within gene	Between gene and spacer	Intergenic spacer	
Number of conserved region	25	7	45	15	1	93
Average of length (bp)	113.1 ±71.9	79.7 ±12.7	159.0 ±108.8	177.3 ±109.6	64.0	142.6 ±103.7
Range of length (bp)	62-384	66-103	62-526	74-427	64	62-526
Average of GC content (%)	40.0 ±4.5	45.0 ±11.9	44.9 ±8.3	50.7 ±9.6	42.2	44.5 ±8.6
Range of GC content (%)	28.8-45.9	34.0-50.7	26.9-58.7	33.3-62.9	42.2	26.9-62.9
Location of conserved region	*psbA* (5), *trnK*, *trnY*, *trnD*, *petA* (2), *psbE*, *rps18*, *psbB* (4), *psbH*, *petB*, *petD* (3), *rpoA* (2)	*petA* ~ *cemA_petA*, *psbL* ~ *psbL_psbF*, *psbE* ~ *psbF_psbE* ~ *psbE*, *psbF* ~ *psbL_psbF*, *trnW* ~ *petG_trnW*, *trnP* ~ *trnP_psaJ*, *psbN* ~ *psbN_psbH*	*rps11*, *rpl2* (8), *rpl23*, *ycf2* (2), *ndhB* (11), *rps7* (2), *rps12*, *rrn16* (3), *trnI* (3), *trnA*, *rrn23* (8)	*rps8* ~ *rpl36_ps8*, *rpl23* ~ *rpl23_trnI*, *trnI* ~ *rpl23_trnI* ~ *trnI_ycf2*; *ndhB* exon2 ~ *ndhB* exon2_*rps7*, *rps12* exon ~ *rps12* exon_*trnV*, *trnV* ~ *trnV_rrn16*, *trnV_rrn16* ~ *rrn16*, *rrn16* ~ *rrn16_trnI* exon1, *rrn16_trnI* exon1 ~ *trnI* exon1 ~ *trnI* exon _*trnI* intron, *trnI* intron_ *trnI* exon2 ~ *trnI* exon2_*trnA* exon1, *trnA* intron ~ *trnA* exon2 ~ *trnA* exon2 ~ *trnA* exon2_*rrn23*, *trnA* exon2_*rrn23* ~ *rrn23*, *rrn23* ~ *rrn23_rrn4.5*, *rrn4.5_rrn5* ~ *rrn5*, *rrn5_trnR* ~ *trnR* ~ *trnR_trnN*		*rps12_trnV*

The GC contents of the complete chloroplast genomes of the six species analyzed in this study were between 33.97 and 38.99%. The GC content in the conserved sequences was approximately 45.93%, whereas that in the non-conserved regions was 32.54-38.23% (Table 1).

The average length of the 93 conserved regions identified in this study was 142.6 bp. Among these, the minimum length was 62 bp, and the maximum length was 526 bp. Thirty-two conserved sequences were located within the LSC region, 61 within the IR region, and none were found in the SSC region. If divided by the “within gene” and “intergenic spacer”, 70 conserved regions were located within gene (Table 4). Only one conserved region (a 64-bp interval between *rps12* and *trnV*) was located in the intergenic spacer. In addition, there were 22 conserved regions extending within the gene and intergenic spacer. The average lengths of conserved sequences within gene (159.0 bp) and between the gene and spacer (177.3 bp) in the IR region were greater than those within gene (113.1 bp) and between the gene and spacer (79.7 bp) in the LSC region. Similarly, the conserved sequences in the IR region had relatively higher GC contents within gene (44.9%) and between gene and spacer (50.7%) than those within (40.0%) and between gene and spacer (45.0%) in the LSC region (Table 4). This indicated that the conserved sequences in the IR region were longer and had a higher GC content. Consequently, these sequences are valuable for designing primer pairs.

Designing primer pairs for systematic analyses

Five primer pairs were designed according to the GC content of the conserved regions, and the variable sequence between neighboring conserved regions of six chloroplast genomes obtained from the database. Of these, cp101 (*psbA~trnK*), cp102 (*psbB~psbH*), and cp105 (*trnY~trnD*) are located within the LSC region, whereas cp103 (rpl23~trnI) and cp104 (*trnR~trnN*) are within the IR region (Table 5).

Table 5. Description of five pairs of chloroplast primers designed from this study.

Primer code	Location	Primer sequence	Primer length (bp)	GC content (%)
cp101F	*psbA~trnK*	5' ACAGAAGTTGCGGTCA 3'	16	50
cp101R	(LSC region)	5' CATAGGGAAAGCCGTGTGCA 3'	20	55
cp102F	*psbB~psbH*	5' CCATTGCAACACCCA 3'	17	53
cp102R	(LSC region)	5' TGGCATGGTGCTAGAAC 3'	15	53
cp103F	*rpl23~trnI*	5' GCATCCATGGCTGAATGG 3'	19	47
cp103R	(IR region)	5' CAAAGAAGAGTTCGACCCA 3'	18	56
cp104F	*trnR~trnN*	5' TCCTCAGTAGCTCAGTGGTAG 3'	20	55
cp104R	(IR region)	5' GGCCTGTAGCTCAGAGGATT 3'	21	52
cp105F	*trnY~trnD*	5' CAGCTTCCGCCTTGA 3'	15	60
cp105R	(LSC region)	5' GCCCGAGCGGTTAAT 3'	15	60

LSC = large single copy; IR = inverted repeat.

The five primer pairs were tested using 14 plant species (18 accessions) as materials. Except for *V. unguiculata* and *C. melo*, which had an extra weaker band, all tested species presented a single clear band amplified with a primer pair (Figure 1, cp104 as an example). The length polymorphisms of the amplified fragments were found when each of five primer pairs was used (Table 6). For most of the tested species, primer pairs cp101 and cp102 could amplify fragments approximately 500-700 and 800-900 bp, respectively. However, *E. mexicana* could not be successfully amplified by these two primer pairs. Approximately 550-750 bp fragments of legume species and 400-bp fragments of the other species were amplified by primer pair cp103.

One 400-bp fragment amplified by primer pair cp104 is specific to monocot species. Primer pair cp105 could also amplify polymorphic fragments ranging from 500 to 630 bp (Table 6). All 14 tested species can be discriminated according to the length polymorphisms of DNA fragments amplified by the five primer pairs designed in this study.

Table 6. Length of PCR products from 14 crop species (18 accessions) amplified with five primer pairs.

Accession	Primer pair				
	cp101	cp102	cp103	cp104	cp105
P. sativum cv. TC12	610	900	550	420	590
A. hypogaea cv. TN12	680	890	680	780	630
V. radiata cv. TN3	600	870	650	700	600
G. max cv. KVS5	580	850	700	750	580
G. max cv. KVS8	580	850	700	750	580
V. unguiculata ssp *cylindrica*	620	850	700	720	600
V. unguiculata ssp *sesquipedalis*	620	850	700	720	600
V. angularis cv. KS5	600	840	700	720	600
V. faba cv. NTU-VF03	620	860	750	430	580
C. melo cv. TWAA	580	880	400	770	450
C. sinensis cv. Ching-Shin-Oolong	500	850	400	750	450
S. tuberosum cv. Kennebec	510	890	400	720	480
O. sativa ssp *Japonica* cv. TN67	500	820	400	400	500
O. sativa ssp *Japonica* cv. TK9	500	820	400	400	500
O. sativa ssp *Indica* cv. TCS10	500	820	400	400	500
Z. mays cv. TN442a	520	830	400	400	490
E. mexicana cv. Teo-3	-	-	400	400	490
C. lacryma-jobi cv. Job-1	550	850	400	400	490

- = no PCR product.

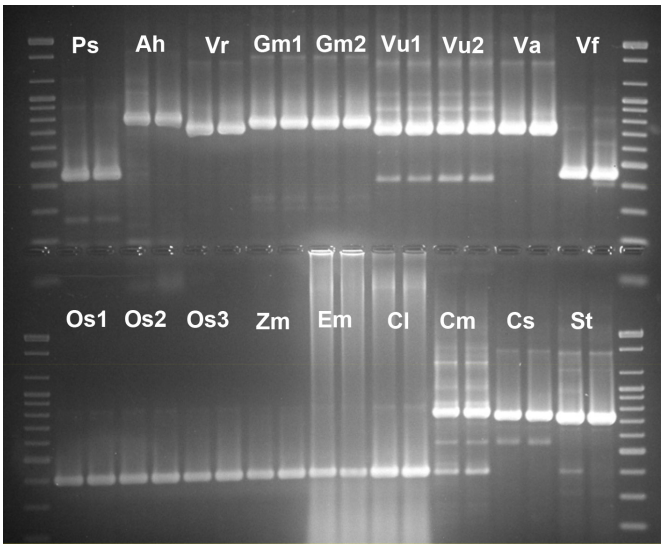

Figure 1. Electrophoresis patterns of PCR products for 14 crop species (18 accessions) amplified with primer pair cp104. Ps = *Pisum sativum TC12*; Ah = *Arachis hypogaea TN12*; Vr = *Vigna radiate TN3*, Gm1 = *Glycine max KVS5*; Gm2 = *G. max KVS8;* Vu1 = *Vigna unguiculata* ssp *Cylindrical;* Vu2 = *V. unguiculata* ssp *Sequipedalis;* Va = *Vigna angularis KS5*; Vf = *Vicia faba NTU-VF03;* Cm = *Cucumis melo TWAA;* Cs = *Camellia sinensis Ching-Sing-Wu-Long*; St = *Solanum tuberosum Kennebec;* Os1 = *Oryza sativa TN67;* Os2 = *O. sativa TK9*; Os3 = *O. sativa TCS10;* Zm = *Zea mays TN442a*; Em = *Euchlaena mexicana Teo-3*; and Cl = *Coix lacryma-jobi Job-1*.

Because a few unstable bands of fragments amplified by the primer pairs cp103 and cp105 were found, only the PCR products of primer pairs cp101, cp102, and cp104 were sequenced. The phylogenetic trees of the species tested were displayed according to the DNA sequences amplified by the three pairs of primers (Figure 2). Because the primer pairs cp101 and cp102 were unable to amplify DNA fragments from *E. mexicana,* sequence data from the other 13 species were investigated in the combined analysis. Phylogenetic analyses showed that *S. tuberosum* and *C. sinensis* were independently classified as an out-group, and *C. melo* was categorized between the legume and Poaceae groups. The Poaceae group included *Z. mays*, *C. lacryma-jobi*, and *O. sativa*. *Z. mays* and *C. lacryma-jobi* were divided into a small group. In the legume group, *A. hypogaea*, *P. sativum*, and *V. faba* were classified into the same cluster. *G. max* was grouped closely with *Vigna* species. The grouping result from the primer pair cp102 was similar to that from cp101. However, *A. hypogaea*, which was classified in the same subgroup as *P. sativum* and *V. faba* based on primer pair cp101, was classified in the outside group of legumes. The primer pair cp104 could amplify the DNA fragments of *E. mexicana*, and four Poaceae species (*O. sativa*, *C. lacryma-jobi*, *Z. mays*, and *E. mexicana*) were assigned in the same cluster. In addition, we found that the grouping result from the combined data of the three primer pairs was the same as that from the primer pairs cp101 and cp104. According to the grouping result of the sequence amplified by primer pair cp102, a difference only existed in the classification of *A. hypogaea*. The grouping results showed that adopting one or three primer pairs had a consistent result on systematic classification. Besides, different accessions of the same species discriminated by any one primer pair were found (Figure 2).

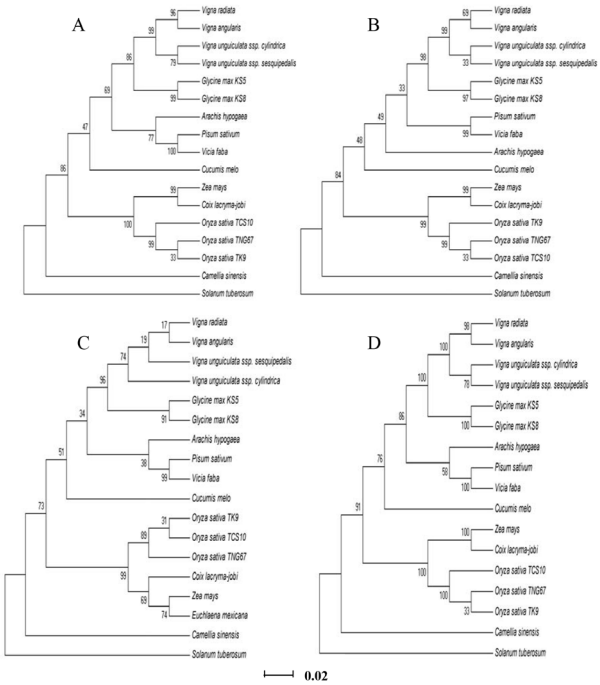

Figure 2. Phylogenetic trees of Kimura 2-parameter distances of 14 crop species (18 accessions) displayed by the maximum parsimony method based on chloroplast DNA sequences amplified with three primer pairs. **A.** primer pair cp101; **B.** primer pair cp102; **C.** primer pair cp104; and **D.** combined data of three primer pairs. Bootstrap values are shown at the nodes.

DISCUSSION

In this study, variation in sequence length (124-154 kb) of the six complete chloroplast genomes was found. This could be because the genomes of some legumes (such as *M. truncatula*) lack 1-2 IR regions, leading to shorter genome sequences than those of other species. Species that lack an IR region include six groups of the Papilionaceous tribes (Galegeae, Hedysarae, Carmichaelieae, Vicieae, Cicereae, and Trifolieae) (Strauss et al., 1988; Lavin et al., 1990). In addition, long-term evolution could have resulted in shorter genomes. For example, the genome size of Poaceae crops, such as *Triticum aestivum*, *Secale cereale*, *H. vulgare*, *Avena sativa*, *Setaria italica*, and *O. sativa*, is approximately 135 kb, which is shorter than that of legume species, which do not lack IR (Doyle et al., 1996; Saski et al., 2005). Alternatively, these length variations could be caused by large scale rearrangement of the genome (Cosner et al., 1997).

Because *M. truncatula* lacks one IR fragment, it has comparatively few genes. However, the size of the coding region increases as the complete genome size decreases (Saski et al., 2005). Although *O. sativa* has approximately 30 more genes than dicot plants do, among these genes, 24 are hypothetical genes and five are pseudogenes (http://www.ncbi.nlm.nih.gov/nuccore/NC_001320). This implies that most of the extra genes have yet to be verified, or that they do not have any function.

To design highly specific primers from the conserved sequences, this study set the maximum entropy of each position at 0.8, and the maximum average similarity matrix within the conserved region at 0.2. When using the Find Conserved Regions function of the BioEdit software, differences between species were considered as variable regions. For example, only one IR region was found in *M. truncatula*, but two IR regions were highly conserved in other species. In addition, the majority of legume plants, in comparison with the *A. thaliana* genome sequence, had a 51-kb inversion in the LSC region. This 51-kb inversion fragment was considered as a variable region, and the opportunity to design universal primers was eliminated. In addition, conserved sequences shorter than 60 bp were not included, thus decreasing the percentage of conserved regions. Therefore, with the maximum entropy of each position set to 0.8, the percentage of conserved regions in the six species was approximately 8.59-10.70%.

The GC content of the six analyzed chloroplast genomes was approximately 33.97-38.99%. The difference in the GC content is not obvious in the chloroplast genomes, but the GC content of nuclear genomes is approximately 60-70% in monocot plants, and approximately 46% in dicot plants (Salinas et al., 1988; Carels et al., 1998). The results of this study indicate that the GC content of conserved regions in chloroplast genomes is 45.93%, which is greater than that of the non-conserved regions (34.67%). This indicates that conserved regions are easier to search for sequences containing a GC content larger than 50% for regularly designing primers.

Because the purpose of this study was to design primers that could amplify variable regions of the chloroplast genome, the 51-kb inversion sequence was not reversed and the IR region of *M. truncatula* was not removed to perform sequence comparisons. In these 93 conserved regions, 32 (34.41%) and 61 (65.59%) were respectively distributed in LSC and IR regions. Previous studies indicated that the least and the most conservative regions of the chloroplast genome were separated in the LSC (Ravi et al., 2008; Li et al., 2014) and IR (Curtis and Clegg, 1984; Wolfe et al., 1987). The possible explanation for this result is that conserved sequences identified for designing primers and long variable sequences selected for detecting variation among plant species were

considered at the same time. In addition, 70 (25 + 45) of the 93 regions were located within gene regions and 22 (7 + 15) were located in sites covering gene and intergenic spacers (IGS), and only one was located in an intergenic spacer region (Table 4). This suggests that the sequences of gene regions are highly conservative, and that regions between neighboring genes are useful to detect sequence differences in the DNA sequence among species. Shaw et al. (2007) found that the non-coding regions of the chloroplast genome have higher DNA sequence variation than do coding regions. In our study, because only seven genes (*trnK*, *petB*, *petD*, *rpl2*, *rps12*, *trnI,* and *trnA*) were identified in the conserved regions and annotated with intron and exon areas, we could not provide sufficient evidence to support the result reported by Shaw et al. (2007).

Cai et al. (2006) found that the highest GC content in the IR region was influenced by the extremely high GC content of rRNA genes, whereas the lowest GC content in the SSC region was because of the low GC content in the NAD(P)H dehydrogenase genes (Cai et al., 2006). Among the six species analyzed in this study, the GC content of the conserved region (45.93%) was higher than that of the non-conserved region (34.67%) (Table 1). This is consistent with the findings of Cai et al. (2006).

Matsuoka et al. (2002) compared the chloroplast genome sequences of *O. sativa*, *Z. mays*, and *T. aestivum*. They found that there was high variation in the sequences of *matK*, *cemA*, and *clpP*, whereas the sequences of rRNA and tRNA genes were conservative. Similar results were observed in our study. We also identified more conserved regions in rRNA and tRNA genes and a variable region in the *matK* gene (amplified by cp101). Because the rRNA and tRNA gene regions have high DNA sequence similarity between species, two pairs of primers (cp104 and cp105) developed in this study targeted those regions.

Although six diverse chloroplast genomes of more than 190 completely sequenced plastid genomes (Gao et al., 2010) were selected to identify conserved regions of chloroplast genomes, it was necessary to test the effectiveness of these primer pairs with additional plant accessions. Castillo et al. (2010) successfully transferred the use of SSR markers in barley to different species and genera. In addition, chloroplast markers of variable lengths have been used in population genetic studies of two legume species (Wheeler et al., 2012). In our study, five primer pairs were designed according to the conserved chloroplast sequences of six species, and were tested with 14 species. The length polymorphism of amplified DNA fragments was found for each primer pair, indicating the potential of applying the five primer pairs to various plant species. Interestingly, the primer pair cp103 can be used to differentiate legume species (with a band >400 bp) and non-legume species (with a 400-bp band), and the primer pair cp104 can be applied to discriminate monocot plants (with a 400-bp band) and dicot plants (with a band >400 bp). This suggests that the variable regions in *rpl23~trnI* (amplified by cp103) and *trnR~trnN* (amplified by cp104) were possibly involved in divergent evolution between chloroplast genomes. The *psbA~trnK* sequence, which was amplified with primer pair cp101, partially overlapped the *psbA~trnH* sequence, which had the most intra-species variation within the chloroplast genome (Selvaraj et al., 2013). On the whole, the five primer pairs are useful for detecting amplified fragment length polymorphisms to discriminate between different plant species, and primer pairs cp101, cp102, and cp104 are valuable for identifying DNA sequence variation to enable variety identification and phylogenetic analysis.

The revised edition of the Angiosperm Phylogeny Group (APG) states that angiosperms are primarily composed of three branches: magnoliids, monocots, and eudicots. Under monocots, commelinids are also distinguished. Under eudicots, core eudicots are distinguished. Under core

eudicots, rosids and asterids are divided (APG, 2003). Legume plants and *C. melo* belong to eurosids I under rosids. *C. sinensis* is included in Theaceae and belongs to asterids. *S. tuberosum* is a member of Solanaceae and belongs to euasterids I under astrids (APG, 2003). Results of the phylogenetic analysis based on sequences amplified with single or combined primer pairs, are consistent with the classifications of the APG.

Shaw and Small (2005) and Shaw et al. (2007) showed that the most variable regions in the chloroplast genome included the *rpl14~rps8~infA~rpl36, petL~psbE, psbJ~petA, psaI~accD, trnV~ndhC, ndhJ~trnF, psbD~trnT, atpI~atpH, trnQ~rps16, rps16~trnK, ndhA intron, ndhF~rpl32, rpl32~trnL*, and *trnS~trnG~trnG* regions. Recently, use of the chloroplast intergenic spacers *trnE~trnT*, *trnT~psbD*, *ndhF~rpl32,* and *rpl14~rpl16* was suggested for discriminating between cultivar species (Selvaraj et al. 2013). However, the variable regions described above were not amplified by the five primer pairs designed in this study. Only one region (*psbA~trnH*) published by Byrne and Hankinson (2012) and Selvaraj et al. (2013) was partially overlapping with the *psbA~trnK* region (amplified by the primer pair cp101). One explanation is that when we designed primers, the identification of conserved regions was first considered, followed by the longer lengths (approximately 700-800 bp) of the variable regions. Next, the variation frequency of the amplified sequences was also considered. On the other hand, the region (*psbA~trnK*) amplified by the cp101 primer pair includes the whole *matK* gene and adjacent regions. This gene has already been proved to have excellent utility on specific classification (Soltis and Soltis, 1998). Furthermore, the amplified region (*trnY~trnD*) of the cp105 primer pair partially overlapped with a region amplified by one primer pair developed by Demesure et al. (1995).

In this study, six diverse chloroplast genomes were analyzed in order to identify conserved regions and design universal primer pairs. The variable DNA sequences amplified by the primers developed in this study were designed to identify genetic variation between species. The transfer use efficiency of these primer pairs has been demonstrated. These primer pairs are recommended for specific discrimination and phylogenetic analyses. Although these primer pairs could discriminate different varieties of *G. max*, *O. sativa*, and *V. unguiculata*, only a few varieties of each crop were investigated. Further studies are needed to test the feasibility of applying this tool to variety discrimination. In addition, due to the nature of the haploid chloroplast genome, the developed primer pairs may also have potential for application on mixed plant products, such as vegetable oils and ground plant powders.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank the researchers of the Taiwan Agricultural Research Institute, Seed Improvement and Propagation Station, Tea Research and Extension Station, Taichung District Agricultural Research and Extension Station (DARES), Tainan DARES, and Kaohsiung DARES in Taiwan for providing seeds for this study. We also thank Dr. Chi-Dong Liu (Department of Agronomy, National Chia-Yi University), Dr. Su-Hsian Wang (Tainan District Agricultural Research and Extension Station), and Dr. Mao-Sung Ye (Department of Agronomy, National Chung-Hsing University) for their comments on an earlier version of this manuscript.

REFERENCES

Angiosperm Phylogeny Group (APG II) (2003). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* 141: 399-436.

Byrne M and Hankinson M (2012). Testing the variability of chloroplast sequences for plant phylogeography. *Austr. J. Bot.* 60: 569-574.

Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, et al. (2006). Complete plastid genome sequences of *Drimys*, *Liriodendron*, and *Piper*: implications for the phylogenetic relationships of magnoliids. *BMC Evol. Biol.* 6**:** 77.

Carels N, Hatey P, Jabbari K and Bernardi G (1998). Compositional properties of homologous coding sequences from plants. *J. Mol. Evol.* 46**:** 45-53.

Castillo A, Budak H, Martin AC, Dorado G, et al. (2010). Interspecies and intergenus transferability of barley and wheat D-genome microsatellite markers. *Annal. Appl. Biol.* 156**:** 347-456.

Clegg MT, Gaut BS, Learn GH Jr and Morton BR (1994). Rates and patterns of chloroplast DNA evolution. *Proc. Natl. Acad. Sci. U.S.A.* 91**:** 6795-6801.

Cosner ME, Jansen RK, Palmer JD and Downie SR (1997). The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Curr. Genet.* 31: 419-429.

Curtis SE and Clegg MT (1984). Molecular evolution of chloroplast DNA sequences. *Mol. Biol. Evol.* 1**:** 291-301.

De Las Rivas J, Lozano JJ and Ortiz AR (2002). Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. *Genome Res.* 12**:** 567-583.

Demesure B, Sodzi N and Petit RJ (1995). A set of universal primers for amplication of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Mol. Ecol.* 4: 129-131.

Doyle JJ and Doyle JL (1990). Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Doyle JJ, Doyle JL, Ballenger JA and Palmer JD (1996). The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. *Mol. Phylogen. Evol*. 5: 429-438.

Dumolin-Lapègue S, Pemonge MH and Petit RJ (1997). An enlarged set of consensus primers for the study of organell DNA in plants. *Mol. Ecol.* 6: 393-397.

Edwards D, Batley J and Snowdon RJ (2013). Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* 126: 1-11.

Gao L, Su YJ and Wang T (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: current status and prospects. *J. Syst. Evol*. 48: 77-93.

Guo X, Castillo-Ramírez S, González V, Bustos P, et al. (2007). Rapid evolutionary change of common bean (*Phaseolus vulgaris* L.) plastome, and the genomic diversification of legume chloroplasts. *BMC Genomics* 8: 228.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acid. Symp. Ser.* 41**:** 95-98.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, et al. (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol. Gener. Genet.* 217**:** 185-194.

Jiao YN and Guo H (2014). Prehistory of the angiosperms: characterization of the ancient genomes. In: Advances in Botanical Research - Genomes on Herbaceous Land Plants (Paterson AH, ed.). Elsevier Ltd, Oxford, 223-245.

Kato T, Kaneko T, Sato S, Nakamura Y, et al. (2000). Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Res.* 7**:** 323-330.

Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.

Lavin M, Doyle JJ and Palmer JD (1990). Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. *Evolution* 44**:** 390-402.

Li PB, Li ZH, Liu HM and Hua JP (2014). Cytoplasmic diversity of the cotton genus as revealed by chloroplast microsatellite markers. *Res. Crop. Evol.* 61**:** 107-119.

Matsuoka Y, Yamazaki Y, Ogihara Y and Tsunewaki K (2002). Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals. *Mol. Biol. Evol.* 19**:** 2084-2091.

Perry AS and Wolfe KH (2002). Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J. Mol. Evol.* 55**:** 501-508.

Ravi V, Khurana JP, Tyagi AK and Khurana P (2008). An update on chloroplast genomes. *Plant Syst. Evol.* 271**:** 101-122.

Salinas J, Matassi G, Montero LM and Bernardi G (1988). Compositional compartmentalization and compositional patterns in

the nuclear genomes of plants. *Nucl. Acid. Res.* 16: 4269-4285.
Saski C, Lee SB, Daniell H, Wood TC, et al. (2005). Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant. Mol. Biol.* 59**:** 309-322.
Saski C, Lee SB, Fjellheim S, Guda C, et al. (2007). Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theor. Appl. Genet.* 115**:** 571-590.
Sato S, Nakamura Y, Kaneko T, Asamizu E, et al. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6: 283-290.
Selvaraj D, Park JI, Chung MY, Cho YG, et al. (2013). Utility of DNA barcoding for plant biodiversity conservation. *Plant Breed. Biotech*. 1**:** 320-332.
Shaw J and Small RL (2005). Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus subgenus Prunus* section *Prunocerasus*, Rosaceae). *Am. J. Bot.* 92**:** 2011-2030.
Shaw J, Lickey EB, Schilling EE and Small RL (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Amer. J. Bot.* 94**:** 275-288.
Soltis DE and Soltis PS (1998). Choosing an approach and an appropriate gene for phylogenetic analysis. In: Molecular Systematics of Plants II: DNA Sequencing (Soltis DE, Soltis PS and Doyle JJ, eds.). Kluwer Academic Publisher, Dordrecht, 1-42.
Strauss SH, Palmer JD, Howe GT and Doerksen AH (1988). Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proc. Nat. Acad. Sci. U.S.A.* 85: 3898-3902.
Wheeler GL, McGlaughlin ME and Wallace LE (2012). Variable length chloroplast markers for population genetic studies in Acmispon (Fabaceae). *Am. J. Bot.* 99**:** e408-e410.
Wojciechowski MF, Lavin M and Sanderson MJ (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *Am. J. Bot.* 91: 1846-1862.
Wolfe KH, Li WH and Sharp PM (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Nat. Acad. Sci. U.S.A.* 84: 9054-9058.
Yang JB, Tang M, Li HT, Zhang ZR, et al. (2013). Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol. Biol.* 13**:** 84.