Comparative analysis of chloroplast DNA sequences of *Codonopsis lanceolata* and *Platycodon grandiflorus* and application in development of molecular markers

Sun-Goo Hwang 1 · Ju Hee Kim 1 · Jun-Cheol Moon 2 · Jin-Hyuk Kim 1 · Cheol Seong Jang 1

Received: 11 October 2016 / Accepted: 2 December 2016 / Published online: 23 December 2016
© The Korean Society for Applied Biological Chemistry 2016

**Abstract** *Codonopsis lanceolata* and *Platycodon grandiflorus* (order Asterales) originate from East Asia. Despite the high commercial availability of *C. lanceolata* and *P. grandiflorus*, limited genetic research has been performed on these plants. We applied a targeting enrichment method to detect genetic diversity in *C. lanceolata* and *P. grandiflorus* and recovered their chloroplast genomes from total DNA sequence data. Chloroplast DNAs (cpDNAs) were 61,154 bp (*C. lanceolata*) and 81,214 bp (*P. grandiflorus*) in length. Sixteen simple sequence repeats and 15 long repeat sequences were determined, which are useful as potential markers in both plant species. We surveyed the phylogenetic relationships with increased resolution in 14 plant species, including other 8 species from the order Asterales and 4 from the order Apiales. In addition, we demonstrated the availability of recovered chloroplast genomes through cpDNA marker development to determine the authenticity of food fraud at the DNA level of plant species.

**Keywords** Chloroplast genome · *Codonopsis lanceolata* · Marker · *Platycodon grandiflorus*

---

**Introduction**

Asterales is an order of dicotyledonous flowering plant species and includes 11 families evolved from one common ancestor. Many herbaceous plant species of Asterales are used as spices and traditional medicines. In previous phylogenetic research, Asterales have been included in the Campanulid clade, which is similar to the Euasterids II group of the Angiosperm Phylogeny Group III as one type of taxonomic classification (Bremer et al. 2002). The molecular evidence suggests that Asterales originated c. 100 Ma B.P. in the Cretaceous (Bremer and Gustafsson 1997).

In Korea, *Codonopsis lanceolata*, order Asterales, is known as *Deodeok*. Its roots are used in traditional food dishes. *Platycodon grandiflorus*, order Asterales, is a herbaceous flowering perennial plant, known as *Doraji* in Korea, and its roots are a recognized ingredient in salads and traditional cuisine. Both plants originate in East Asia and are used as natural and traditional medicine due to their own functional compounds. In addition, *Panax ginseng*, order Asterales, has been utilized in China as a traditional medicine for the treatment of ailments related to the central nervous system, and endocrine and adrenocortical systems, and to control blood pressure and diabetes, (Nah et al. 2007; Kim 2012). The complete chloroplast genome of Korean ginseng (*Panax schinseng* Nees) is 156,318 bp and North American ginseng (*Panax quinquefolius*) is 156,359 bp (Kim and Lee 2004; Han et al. 2015). Also, Zhao et al. (2015) reported the comparative analysis of chloroplast DNAs (cpDNAs) for four *P. ginseng* strains including Damaya, Ermaya, Gaolishen, and Yeshanshen, in which the complete chloroplast genome size (156,354 bp) of *P. ginseng* Damaya was equal to the total length of each
cpDNA of *P. ginseng* Ermaya and Gaolishen, unlike that of *P. ginseng* Yeshanshen. The different minor allele sites in the large single copy and inverted repeat regions of the chloroplast genome were also identified, suggesting an inferred evolutionary relationship among these four *P. ginseng* strains. However, high-resolution studies of evolutionary relationships using numerous sequences have not been reported for *C. lanceolata* and *P. grandiflorus* chloroplast genomes.

Recently, the consumption of functional foods for improving health and wellness has increased. However, commercial fraud related to food products has become a concern due to the differences in the cost of food materials, e.g., the price of ginseng is generally higher than that of *Deodeok* or *Doraji*; therefore, the latter cheaper options are being used in products and being marketed as original ginseng. Therefore, consumer concerns related to unoriginal ginseng food products, which are mixed with rather cheap foods such as *Deoduck* and *Doraji*, have increased. However, it is difficult to discriminate the ingredients of ginseng products, especially ginseng powders, visually. Molecular markers can be used to determine the authenticity of food at the DNA level of plant species. For example, molecular markers derived from *rpoB* and *rpoC2* successfully detected the cpDNA of rice grain flour in different mixed-flour samples using quantitative real-time PCR (qRT-PCR) (Hwang et al. 2015). Furthermore, Moon et al. (2016) determined unique species among five plants, including rice, barley, adlay, wheat, and maize, using a cpDNA marker developed from *rpoC2*, which was used for detecting the cpDNA for each plant species in seven commercial mixed-flour products. Molecular markers have been developed on the basis of sequence differences; however, the genetic information of *C. lanceolata* and *P. grandiflorus* mostly remains unknown, even though the chloroplast genome sequence of *P. ginseng* has been completed (Zhao et al. 2015).

Photosynthesis, which occurs in chloroplasts, is an important process for energy production of green plants. The chloroplasts contain a genome ranging in size from 120 to 170 kb (Clegg et al. 1994). The entire nucleotide sequence of cpDNA has been widely used for phylogenetic studies and molecular marker development due to its highly conserved genetic structure. Bremer et al. (2002) reported that the three coding and three non-coding cpDNA markers provided improved evidence for resolution among the Asterids order at higher taxonomic levels. Recently, the rapid development of the next generation sequencing technique provided a comprehensive methodology for monitoring genetic diversity in plant species. For example, the complete chloroplast genome sequences of *P. ginseng* obtained from DNA sequencing (DNA-seq) provided genome-wide information, such as the phylogenetic relationships and structural diversity of cpDNA among various plant species (Zhao et al. 2015).

Here, we performed the DNA-seq from *C. lanceolata* and *P. grandiflorus* and applied a target enrichment method for convenient cpDNA recovery from total DNA sequence data. Phylogenetic analysis was performed using 12 chloroplast genome sequences of Asterales, including *C. lanceolata* and *P. grandiflorus*, and 4 sequences of Apiales with *P. ginseng*. To develop molecular markers for discriminating three species, we selected two gene families, in which their segregating sites were 191 for *rpo* and 20 for *ndh*, and then designed gene-specific primers for quantitative real-time polymerase chain reaction (qRT-PCR). Finally, the specific primers were tested in total genomic DNA isolated from three plant species, *C. lanceolata*, *P. grandiflorus*, and *P. ginseng*.

### Materials and methods

**DNA sequencing of *C. lanceolata* and *P. grandiflorus***

Seeds of *C. lanceolata* and *P. grandiflorus* were obtained from the Gyeonggi-do Agricultural Research and Extension Services, and seedlings were grown in a growth chamber at 25 °C for 4 weeks. Total DNA was isolated from the leaves of 2-week-old seedlings of *C. lanceolata* and *P. grandiflorus* using the CTAB method (Saghai-Maroof et al. 1984), and the extracted DNA was used to conduct the DNA-seq by an Illumina HiSeq2000 following the manufacturer’s instructions.

**Recovery of the chloroplast genome from the total DNA sequence data**

To recover chloroplast genome sequences from DNA-seq data, we used the modified target enrichment method (Mandel et al. 2014). In summary, for quality control of raw data, the FASTQ files data obtained from DNA-seq were filtered using the PRINSEQ tool with a minimum read length of 40, minimum quality score of 15, minimum length of 3′-end poly-N tail of 10, and ambiguous base N percentage of 20 (http://prinseq.sourceforge.net/index.html). The filtered reads were annotated based on the complete chloroplast genome of ten plant species from the order Asterales obtained from NCBI (http://www.ncbi.nlm.nih.gov/) using BLAST search, and then the annotated reads identified as potential cpDNA sequences of *C. lanceolata* and *P. grandiflorus* were collected using PERL script. The collected cpDNA reads were used for de novo assembly using VelvetOptimiser (http://www.vicbioinformatics.com/software.velvetoptimiser.shtml). The assembled contigs were rearranged to produce one chromosome.
genome sequence using the ABACAS tool (Assefa et al. 2009). The Campanula takesimana chloroplast genome as a reference sequence was used for contig assembly due to the higher genetic similarity of cpDNA with C. lanceolata and P. grandiflorus in Asterales. The gene information of rearranged chloroplast genomes was annotated using the cpGAVAS tool with default parameters (Liu et al. 2012).

Identification of repeat sequences

Simple sequence repeats (SSRs) in C. lanceolata and P. grandiflorus chloroplast genomes were identified using a microsatellite identification tool (http://pgrc.ipk-gatersleben.de/misa/) with different minimum numbers of repeats, which were 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. Also, the long repeat sequences of each chloroplast genome were detected using the REPuter tool with a minimum length of repeats of 30 bp (Kurtz and Schleiermacher 1999). The genes of SSRs and long repeat sequences were manually identified using a general feature format file obtained from cpGAVAS (Liu et al. 2012).

Codon usage and phylogenetic analysis

We detected the codon frequency and relative synonymous codon usage in C. lanceolata and P. grandiflorus using the SeqinR R package with default parameters (Charif et al. 2005). A total of 14 different chloroplast genomes obtained from NCBI, as well as C. lanceolata and P. grandiflorus, were used to construct a phylogeny tree using the progressive alignment method of MAUVE software (Darling et al. 2004). The constructed tree was visualized by MEGA5 (Tamura et al. 2011), and the Tajima’s D of protein-encoding genes were calculated using MEGA5 (Tajima 1989).

Molecular marker development

To develop a molecular marker for C. lanceolata, P. grandiflorus, and P. ginseng, we selected the ndhF and rpoA genes as potential candidate genes, and designed the allele-specific PCR primers on the basis of sequence difference (Fig. 2). Total DNA was extracted from 2-week-old seedling leaves of C. lanceolata, P. grandiflorus, and P. ginseng using the CTAB method, respectively (Saghai-Maroof et al. 1984), and then used for qRT-PCR. qRT-PCR was performed using a Real-Time System with 20 μL reaction mixture containing 10 ng of template DNA, ten 1 L of SYBR® Green TOP real qPCR 2 × PremIX (Enzynomics™, Daejeon, Korea), and 10 pmol ndhF-targeted PCR primers. The qRT-PCR conditions were as follows: 10 min at 95 ºC, followed by 30 cycles of 10 s at 95 ºC, 30 s at 55–60 ºC, and 20 s at 72 ºC. The PCR products were loaded on a 2% agarose gel, and a photograph of the gel was taken using the Molecular Imager Gel DOC™ XR + System (Bio-Rad, Hercules, California, USA). The primer efficiency was determined using the formula of Bustin et al. (2009), and the slope value of the regression line was calculated by the linearity test (Ramakers et al. 2003).

Results and discussion

Recovery of C. lanceolata and P. grandiflorus chloroplast genomes from total DNA sequence data

To identify the genetic variation of cpDNAs between C. lanceolata and P. grandiflorus, we employed the modified target enrichment approach to determine the chloroplast genome sequence, which facilitates rapid collection of target sequences in the numerous DNA sequences (Mandel et al. 2014). The collected target chloroplast genome of C. lanceolata was 61,154 bp in length except a gap that included 59 different genes, which are composed of 51 protein-coding genes, as well as 7 tRNA and 1 rRNA (Table 1). A total of 73 different genes, including 62 protein-coding genes and 11 tRNA, were exhibited in the P. grandiflorus chloroplast genome of 81,214 bp (Table 1). There were 22 genes (7551 bp) in the intron fragment in C. lanceolata and 31 (11,598 bp) in P. grandiflorus.

Table 1 Gene characteristics mapped on each chloroplast genome

|                    | C. lanceolata         | P. grandiflorus       |
|--------------------|-----------------------|-----------------------|
| Size of cpDNA except gap | 61,154 bp             | 81,214 bp             |
| No. of different genes | 59                    | 73                    |
| No. of different protein-encoding genes | 51                    | 62                    |
| No. of different tRNA genes | 7                     | 11                    |
| No. of different rRNA genes | 1                     | Not found             |
| No. of different genes with introns | 22 (7551 bp)          | 31 (11,598 bp)        |
| Amino acid | Codon | C. lanceolata | P. grandiflorus | Amino acid | Codon | C. lanceolata | P. grandiflorus |
|------------|-------|---------------|----------------|------------|-------|---------------|----------------|
|            | Number | Frequency | RSCU | Number | Frequency | RSCU | Number | Frequency | RSCU |
| Ala (A)    | gca    | 55        | 26.96 | 1.08 | 92      | 26.06 | 1.04 | Leu (L)   | ctt    | 140 | 16.71 | 1.00 |
| Ala (A)    | gcc    | 55        | 26.96 | 1.08 | 88      | 24.93 | 1.00 | Leu (L)   | tta    | 166 | 19.81 | 1.19 |
| Ala (A)    | gcg    | 30        | 14.71 | 0.59 | 64      | 18.13 | 0.73 | Leu (L)   | tta    | 236 | 24.20 | 1.45 |
| Ala (A)    | gct    | 64        | 31.37 | 1.25 | 109     | 30.88 | 1.24 | Lys (K)   | aaa    | 180 | 55.21 | 1.10 |
| Arg (R)    | aga    | 116       | 28.78 | 1.73 | 214     | 26.07 | 1.56 | Lys (K)   | aag    | 146 | 44.79 | 0.90 |
| Arg (R)    | agg    | 69        | 17.12 | 1.03 | 186     | 22.66 | 1.36 | Met (M)   | atg    | 155 | 100.00| 1.00 |
| Arg (R)    | cga    | 71        | 17.62 | 1.06 | 137     | 16.69 | 1.00 | Phe (F)   | ttc    | 159 | 39.36 | 0.79 |
| Arg (R)    | cgc    | 40        | 9.93  | 0.60 | 70      | 8.53  | 0.51 | Phe (F)   | ttt    | 245 | 60.64 | 1.21 |
| Arg (R)    | cgg    | 58        | 14.39 | 0.86 | 111     | 13.52 | 0.81 | Pro (P)   | cca    | 68  | 26.15 | 1.05 |
| Arg (R)    | cgt    | 49        | 12.16 | 0.73 | 103     | 12.55 | 0.75 | Pro (P)   | ccc    | 52  | 20.00 | 0.80 |
| Asn (N)    | aac    | 92        | 38.49 | 0.77 | 135     | 28.18 | 0.56 | Pro (P)   | cgg    | 65  | 25.00 | 1.00 |
| Asn (N)    | aat    | 147       | 51.51 | 1.23 | 344     | 71.82 | 1.44 | Pro (P)   | cct    | 75  | 28.85 | 1.15 |
| Asp (D)    | gac    | 52        | 36.11 | 0.72 | 93      | 31.85 | 0.64 | Ser (S)   | agc    | 40  | 8.71  | 0.91 |
| Asp (D)    | gat    | 92        | 63.89 | 1.28 | 199     | 68.15 | 1.36 | Ser (S)   | agt    | 62  | 13.51 | 0.81 |
| Cys (C)    | tgc    | 56        | 49.56 | 0.99 | 176     | 46.19 | 0.92 | Ser (S)   | tca    | 89  | 19.39 | 1.16 |
| Cys (C)    | tgt    | 57        | 50.44 | 1.01 | 205     | 53.81 | 1.08 | Ser (S)   | tcc    | 69  | 15.03 | 0.90 |
| Gln (Q)    | caa    | 129       | 62.32 | 1.25 | 202     | 70.88 | 1.42 | Ser (S)   | tccg   | 76  | 16.56 | 0.99 |
| Gln (Q)    | cag    | 78        | 37.68 | 0.75 | 83      | 29.12 | 0.58 | Ser (S)   | tct    | 123 | 26.80 | 1.61 |
| Glu (E)    | gaa    | 132       | 58.67 | 1.17 | 250     | 70.42 | 1.41 | Thr (T)   | aca    | 77  | 29.50 | 0.88 |
| Glu (E)    | gag    | 93        | 41.33 | 0.83 | 105     | 29.58 | 0.59 | Thr (T)   | acc    | 61  | 23.37 | 1.12 |
| Gly (G)    | gga    | 95        | 31.88 | 1.28 | 196     | 28.61 | 1.14 | Thr (T)   | aca    | 65  | 24.90 | 0.95 |
| Gly (G)    | gcc    | 47        | 15.77 | 0.63 | 144     | 21.02 | 0.84 | Thr (T)   | act    | 58  | 22.22 | 1.18 |
| Gly (G)    | ggc    | 82        | 27.52 | 1.10 | 194     | 28.32 | 1.13 | Trp (W)   | tgg    | 129 | 100.00| 0.93 |
| Gly (G)    | ggt    | 74        | 24.83 | 0.99 | 151     | 22.04 | 0.88 | Tyr (Y)   | tac    | 113 | 40.65 | 1.00 |
| His (H)    | cac    | 52        | 44.83 | 0.90 | 65      | 32.34 | 0.65 | Tyr (Y)   | taq    | 165 | 59.35 | 0.89 |
| His (H)    | cat    | 64        | 55.17 | 1.10 | 136     | 67.66 | 1.35 | Val (V)   | gta    | 103 | 27.84 | 1.00 |
| Ile (I)    | ata    | 158       | 34.13 | 1.02 | 174     | 23.02 | 0.69 | Val (V)   | gtc    | 63  | 17.03 | 0.81 |
| Ile (I)    | atc    | 109       | 23.54 | 0.71 | 182     | 24.07 | 0.72 | Val (V)   | gtt    | 86  | 23.24 | 1.19 |
| Ile (I)    | att    | 196       | 42.33 | 1.27 | 400     | 52.91 | 1.59 | Val (V)   | gtt    | 118 | 31.89 | 1.11 |
| Leu (L)    | cta    | 133       | 15.87 | 0.95 | 156     | 14.35 | 0.86 | Stp (*)   | taa    | 107 | 29.40 | 0.68 |
| Leu (L)    | ctc    | 105       | 12.53 | 0.75 | 105     | 9.66  | 0.58 | Stp (*)   | tag    | 142 | 39.01 | 0.93 |
| Leu (L)    | ctg    | 102       | 12.17 | 0.73 | 101     | 9.29  | 0.56 | Stp (*)   | tga    | 115 | 31.59 | 1.28 |
in 838 amino acids (13.4%) in *C. lanceolata* and 1087 amino acids (9.9%) in *P. grandiflorus*. However, most the frequent codons encoding leucine were different in both species, i.e., TTG (22.91% within codons for leucine) in *C. lanceolata* and TTA (24.20% within codons for leucine) in *P. grandiflorus*. The least frequent codons were cysteine (113 amino acids, 1.81%) in *C. lanceolata* and methionine (172 amino acids, 1.56%) in *P. grandiflorus*, respectively. Overall, the tendency toward codon frequencies encoding different amino acids was similar between *C. lanceolata* and *P. grandiflorus*. The different frequency tendencies of leucine and cysteine in *C. lanceolata* and *P. grandiflorus* chloroplast genomes were similar in Banana and Omani lime chloroplast genomes (Martin et al. 2013; Su et al. 2014). Prosdocimi and Ortega (2007) suggested that leucine is a major component required to generate more stable DNA mutations in the 10,000 poli-codon sequences, indicating *C. lanceolata* and *P. grandiflorus* chloroplast genomes also exhibited the evolutionary stability from leucine codon usage. In addition, the frequencies of A or T at the third position in the codon usage were higher, suggesting a codon bias of genes toward A/T in *C. lanceolata* and *P. grandiflorus*.

Repetitive sequences on *C. lanceolata* and *P. grandiflorus* chloroplast genomes

A total of 16 SSR loci were detected in both the chloroplast genomes, in which the SSR loci accounted for 86 bp (7 SSRs) in *C. lanceolata* and 106 bp (9 SSRs) in *P. grandiflorus* (Table 3), respectively. Most SSRs were located in the intergenic region, whereas four SSRs were identified in the coding region of genes, i.e., (GAA)₄ was located in *psbC* of *C. lanceolata*, and (T)₁₃, (AT)₇, and (GAA)₄ were located in *ycf5*, *rpoC1*, and *psbC* of *P. grandiflorus*.

### Table 3 List of simple sequence repeats on *C. lanceolata* and *P. grandiflorus* chloroplast genomes

| Species      | SSR         | Number of nucleotide | Start position (bp) |
|--------------|-------------|----------------------|---------------------|
| *C. lanceolata* | (AC)⁵ᵇ      | 10                   | 12,768              |
|              | (TA)⁵       | 10                   | 44,790              |
|              | (GAA)⁴      | 12                   | 27,139 (psbC)       |
|              | (TTC)⁶      | 18                   | 69,752              |
|              | (TTAA)⁶     | 12                   | 3260                |
|              | (AGCA)³     | 12                   | 58,220              |
|              | (GAAG)³     | 12                   | 60,653              |
| *P. grandiflorus* | (T)₁₁      | 11                   | 17,637              |
|              | (A)₁₀       | 10                   | 37,005              |
|              | (T)₁₃       | 13                   | 78,678 (ycf5)       |
|              | (AC)⁶       | 12                   | 14,767              |
|              | (TA)⁵       | 10                   | 20,622              |
|              | (AT)⁷       | 14                   | 54,987 (rpoC1)      |
|              | (GAA)⁴      | 12                   | 34,182 (psbC)       |
|              | (TAA)⁴      | 12                   | 85,164              |
|              | (TTTC)³     | 12                   | 19,804              |

ᵃ SSR: Simple sequence repeats ⁰ Number of repeat sequence ³ Starting position of sequence in the chloroplast genomes

in 838 amino acids (13.4%) in *C. lanceolata* and 1087 amino acids (9.9%) in *P. grandiflorus*. However, most the frequent codons encoding leucine were different in both species, i.e., TTG (22.91% within codons for leucine) in *C. lanceolata* and TTA (24.20% within codons for leucine) in *P. grandiflorus*. The least frequent codons were cysteine (113 amino acids, 1.81%) in *C. lanceolata* and methionine (172 amino acids, 1.56%) in *P. grandiflorus*, respectively. Overall, the tendency toward codon frequencies encoding different amino acids was similar between *C. lanceolata* and *P. grandiflorus*. The different frequency tendencies of leucine and cysteine in *C. lanceolata* and *P. grandiflorus* chloroplast genomes were similar in Banana and Omani lime chloroplast genomes (Martin et al. 2013; Su et al. 2014). Prosdocimi and Ortega (2007) suggested that leucine is a major component required to generate more stable DNA mutations in the 10,000 poli-codon sequences, indicating *C. lanceolata* and *P. grandiflorus* chloroplast genomes also exhibited the evolutionary stability from leucine codon usage. In addition, the frequencies of A or T at the third position in the codon usage were higher, suggesting a codon bias of genes toward A/T in *C. lanceolata* and *P. grandiflorus*.

### Table 4 List of long repeat sequences on *C. lanceolata* and *P. grandiflorus* chloroplast genomes

| Species      | Repeat size | Type  | Start position of 1st repeat (bp) | Start position of the repeat found in other region (bp) | E-value  |
|--------------|-------------|-------|----------------------------------|--------------------------------------------------------|----------|
| *C. lanceolata* | 31 P        | 55,199| 72,065                           |                                                        | 3.24E−10 |
|              | 33 P        | 58,858 (rps12) | 69,604                          |                                                        | 2.02E−11 |
|              | 32 P        | 59,883 (rps12) | 67,745 (ndhA)                   |                                                        | 8.09E−11 |
| *P. grandiflorus* | 31 P       | 19,294 (psbB)   | 86,700                           |                                                        | 5.17E−10 |
|              | 31 P        | 28,933 (psaA)   | 77,363 (ndhF)                   |                                                        | 5.17E−10 |
|              | 30 P        | 67,480                     | 91,231                           |                                                        | 2.07E−09 |
|              | 30 P        | 69,680                     | 88,792                           |                                                        | 2.07E−09 |
|              | 30 P        | 70,901                     | 87,802                           |                                                        | 2.07E−09 |
|              | 31 P        | 71,692                     | 86,880                           |                                                        | 5.17E−10 |
|              | 32 P        | 73,320                     | 85,365                           |                                                        | 1.29E−10 |
|              | 30 P        | 73,390                     | 84,468 (ndhH)                    |                                                        | 2.07E−09 |
|              | 30 P        | 73,473 (ndhH)              | 83,998 (ndhH)                    |                                                        | 2.07E−09 |
|              | 32 P        | 74,061 (ndhH)              | 83,702 (ndhH)                    |                                                        | 1.29E−10 |
|              | 30 P        | 74,754 (ndhA)              | 82,624 (ndhA)                    |                                                        | 2.07E−09 |
|              | 30 P        | 75,920 (ndhG)              | 82,147 (ndhG)                    |                                                        | 2.07E−09 |
respectively. In addition, the 30 long repeat sequences over 30 bp in length were detected in both *C. lanceolata* (6 long repeat sequences) and *P. grandiflorus* (24 long repeat sequences) (Table 4). The 15 long repeat sequences were located in the genic region of genes, e.g., the eight long repeat sequences for *C. lanceolata* were found within six different genes, including *rps12*, *psbB*, *psaA*, *ndhH*, *ndhA*, and *ndhG*. Furthermore, four genes, including *ndhA*, *ndhF*, *ndhH*, and *ndhG*, were involved with seven long repeat sequences in the *P. grandiflorus* chloroplast genome. These repeat sequences might be useful in the development of molecular markers to determine the genetic diversity of *C. lanceolata* and *P. grandiflorus* populations. For example, Cmp3 as a SSR marker developed from cpDNA was used to evaluate the genetic similarity in twenty-five varieties of *Coffea arabica* (Vieira et al. 2010), evidencing the utilization of the SSR loci of *C. lanceolata* and *P. grandiflorus* for SSR markers.

### Table 5 Tajima’s neutrality test of coding sequences

| Gene families | m  | S   | ps | θ   | π     | D     |
|--------------|----|-----|----|-----|-------|-------|
| *atp*        | 11 | 168 | 1.00 | 0.34 | 0.67  | 4.64  |
| *ndh*        | 17 | 20  | 1.00 | 0.30 | 0.62  | 4.33  |
| *psa*        | 4  | 291 | 0.34 | 0.18 | 0.18  | -0.40 |
| *psb*        | 18 | 10  | 1.00 | 0.29 | 0.69  | 5.00  |
| *rbc*        | 5  | 157 | 0.93 | 0.45 | 0.59  | 2.45  |
| *rpl*        | 9  | 52  | 1.00 | 0.37 | 0.66  | 4.00  |
| *rpo*        | 4  | 191 | 0.72 | 0.39 | 0.48  | 2.23  |
| *rps*        | 13 | 17  | 1.00 | 0.32 | 0.70  | 4.95  |
| *ycf*        | 4  | 29  | 0.85 | 0.47 | 0.57  | 2.40  |

$m$ is the number of sequences; $S$ is the number of segregating sites; $p_s = S/m$; $θ = p_s/a_1$, where $p_s$ is the number of polymorphic sites, $a_1 = (1 + 1/2 + 1/3 + … + 1/n - 1)$, and $n$ is the number of sequences; $π$ is the nucleotide diversity; $D$ is the Tajima test statistic.

### Phylogenetic analysis

To study the evolutionary relationship of *C. lanceolata* and *P. grandiflorus* chloroplast genomes, 16 plant species from both orders of Asterales and Apiales, including the both genomes, were employed, and thus phylogeny tree was constructed by using a progressive alignment method of MAUVE software (Darling et al. 2004) (Fig. 1). The chloroplast genomes of 16 plant species were clearly divided into two groups: the Apiales clade of 4 plants and Asterales clade of 12 plants, including *C. lanceolata* and *P. grandiflorus*. The *C. lanceolata* and *P. grandiflorus* chloroplast genomes were remarkably similar and grouped with *Trachelium caeruleum* and *Campanula takesimana*. Similarly, the genetic distance of *C. lanceolata* and *P. grandiflorus* exhibited a close relationship on the basis of a maximum parsimony tree, which was constructed using the petD group II intron as a species level marker (Borsch et al. 2009). In addition, the strict consensus sequences of *atpB*, *atpB-rbcL*, and *atpF-H* between *C. lanceolata* and *P. grandiflorus* were more similar than other plant species of Campanulaceae (Kim and Yoo 2012). Therefore, the phylogenetic tree strongly evidenced a close genetic relationship between both chloroplast genomes of *C. lanceolata* and *P. grandiflorus*.

To identify the sequence diversity of protein-coding genes between *C. lanceolata* and *P. grandiflorus*, we...
performed the Tajima’s D test following multiple alignment (Table 5). The highest number of segregating sites was detected from the \textit{psa} gene family, which also exhibited a negative Tajima’s D value of \(-0.40\), indicating the low presence of rare alleles and a selective sweep. The \textit{psb} gene family showed the lowest number (\(S = 10\)) of segregating sites and positive Tajima’s D value (\(D = 5.00\)), which indicated the presence of multiple alleles and balancing selection. Many gene families exhibited a positive value (\(D > 1\)) of Tajima’s D test, suggesting that balancing selection maintains genetic variability of protein-coding genes between the \textit{C. lanceolata} and \textit{P. grandiflorus} chloroplast genome.

**Application of \textit{C. lanceolata} and \textit{P. grandiflorus} cpDNA for molecular marker development**

In the present study, we used the genetic information of \textit{C. lanceolata} and \textit{P. grandiflorus} chloroplast genomes for molecular marker development, and the three cpDNA-based markers, which are named \textit{Co\textsubscript{ndhF}}, \textit{Pl\textsubscript{rpoA}}, and \textit{Pa\textsubscript{rpoA}}, were developed from \textit{ndhF} and \textit{rpoA} genes using the qRT-PCR assay. The \textit{cooled} dots represent different plant species. Primer efficiency was calculated using the following formula: 

\[
%\text{Efficiency} = \left(\frac{10^{\text{1/slope}} - 1}{9}\right) \times 100
\]

The slope value of regression line was determined using the linearity test.
of the regression line and primer efficiency using qRT-PCR with three independent replicates (Fig. 2C) and identified appropriate primer efficiencies (83–97%) and obvious correlation over 0.99 of each cpDNA marker detected only in the targeted plant species, but not in the non-targeted plant species. In addition, the quantification cycle \( C_q \) values of each targeted plant species based on the concentration of the DNA template used in qRT-PCR, whereas the irregular \( C_q \) values were detected in non-targeted plant species regardless of the amount of DNA. The slope of the regression provided a useful tool to measure the cpDNA content of \( C. lanceolata \), \( P. grandiflorus \), and \( P. ginseng \) in commercial food products. Similarly, the three cpDNA markers developed from \( rpoB \) and \( rpoC2 \) genes were successfully applied in the detection of various rice flours in commercial mixed-flour products using qRT-PCR (Hwang et al. 2015), suggesting the utility of three cpDNA markers for \( C. lanceolata \). The accurate content estimation of food additives was possible using the slope of molecular markers detected by the linearity test. For example, de la Cruz et al. (2013) successfully estimated Brazil nut percentages in 19 commercial products by using a molecular marker developed in the 2S albumin DNA sequences of \( P. grandiflorus \) and \( P. ginseng \) in commercial mixed-flour products. Similarly, the three cpDNA and 73 genes for \( P. grandiflorus \). Also, we determined the repeat sequences for molecular marker development from cpDNA for each species. In the phylogeny tree, the chloroplast genome of \( C. lanceolata \) was closely linked to that of \( P. grandiflorus \) within Asterales, indicating the closest evolutionary relationship between \( C. lanceolata \) and \( P. grandiflorus \). In addition, we developed three cpDNA markers from \( ndhF \) and \( rpoA \) genes based on recovered chloroplast genome sequences for detecting the cpDNA of \( C. lanceolata \), \( P. grandiflorus \), and \( P. ginseng \). The chloroplast genome of \( C. lanceolata \) and \( P. grandiflorus \) recovered from total DNA sequence data will provide useful information to improve the phylogenetic resolution and efficiency of marker development; and the cpDNA markers developed in the present study are useful to distinguish specific plant species between \( C. lanceolata \), \( P. grandiflorus \), and \( P. ginseng \) in commercial mixed-flour products.

Acknowledgments This research was supported by a Grant (14162MFDS971) from the Korean Ministry of Food and Drug Safety in 2016.

References

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25:1968–1969. doi:10.1093/bioinformatics/btp347

Borsch T, Korotkova N, Raus T, Lobin W, Löhne C (2009) The petD group II intron as a species level marker: utility for tree inference and species identification in the diverse genus Campanula (Campanulaceae). Wildenowia 39:7–33

Bremer K, Gustafsson MH (1997) East Gondwanan ancestry of the sunflower alliance of families. Proc Natl Acad Sci USA 94:9188–9190

Bremer B, Karel B, Nahid H, Per E, Richard O, Arne A, Mari K, Barkhordarian E (2002) Phylogenetics of asters based on 3 coding and 3 non-coding chloroplast DNA markers and the utility of non-coding DNA at higher taxonomic levels. Mol Phylogenet Evol 24:274–301

Bustin SA, Stephen A, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfafli MW, Shipley GL, Vandesompele J, Wittwer CT (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55:611–622

Charif D, Thioulouse J, Lobry JR, Perriere G (2005) Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21:545–547

Clegg MT, Gaut BS, Lear GH Jr, Morton BR (1994) Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci USA 91:6795–6801

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403

De la Cruz S, Lopez-Calleja IM, Alcocer M, Gonzalez I, Martin R, Garcia T (2013) TaqMan real-time PCR assay for detection of traces of Brazil nut (Bertholletia excelsa) in food products. Food Control 33:105–113

Han ZJ, Li W, Liu Y, Gao LZ (2015) The complete chloroplast genome of North American ginseng, Panax quinquefolius. Mitochondrial DNA:1–2

Hwang SG, Kim JH, Moon JC, Jang CS (2015) Chloroplast markers for detecting rice grain-derived food ingredients in commercial mixed-flour products. Genes Genom 37:1027–1034

Kim JH (2012) Cardiovascular diseases and Panax ginseng; a review on molecular mechanisms and medical applications. J Ginseng Res 36:16–26

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

Kim K, Yoo K (2012) Phylogenetic relationships of North American campanulaceae based on chloroplast DNA sequences. Korean J Plant Taxon 42:282–293

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426–427

Lieu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genom 13:715. doi:10.1186/1471-2164-13-715

Mandel JR, Dokow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM (2014) A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl Plant Sci 2(2):1300085

© Springer
Martin G, Baurens FC, Cardi C, Aury JM, D’Hont A (2013) The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. PLoS ONE 8(6):e67350

Moon JC, Kim JH, Jang CS (2016) Development of multiplex PCR for species-specific identification of the Poaceae family based on chloroplast gene, rpoC2. Appl Biol Chem 59:201–207

Nah SY, Kim DH, Rhim H (2007) Ginsenosides: are any of them candidates for drugs acting on the central nervous system. CNS Drug Rev 13:381–404

Prosdocimi F, Ortega JM (2007) The codon usage of Leucine, Serine and Arginine reveals evolutionary stability of proteomes and protein coding genes. Braz Sympos Bioinform:149–159

Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339:62–66

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

Su HJ, Hogenhout SA, Al-Sadi AM, Kuo CH (2014) Complete chloroplast genome sequence of Omani lime (*Citrus aurantifolia*) and comparative analysis within the rosids. PLoS ONE 9:e113049. doi: 10.1371/journal.pone.0113049

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

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

Vieira ESN, Von Pinho EVD, Carvalho MGG, Esselink DG, Vosman B (2010) Development of microsatellite markers for identifying Brazilian Coffea arabica varieties. Genet Mol Biol 33:507–514

Zhao YB, Yin J, Guo HY, Zhang YY, Xiao W, Sun C, Wu JY, Qu XB, Yu I, Wang X, Xiao IF (2015) The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. Front Plant Sci. doi:10.3389/Fpls.2014.00696