Analysis of complete chloroplast genome sequence of Korean landrace *Cymbidium goeringii*

Heng Wang^{1,2} · So-Yeon Park¹ · Su-Hyang Song¹ · Mar-Lar San¹ · Yong-Chul Kim¹ · Tae-Ho Ham^{3,4} · Dong-Yong Kim⁵ · Tae-Sung Kim⁴ · Joohyun Lee³ · Soon-Wook Kwon¹

Received: 28 June 2019 / Accepted: 18 December 2019 / Published online: 6 January 2020
© King Abdulaziz City for Science and Technology 2020

Abstract

The complete chloroplast genome sequence of Korean *Cymbidium goeringii* acc. smg222 was analyzed. Based on a comparison with Chinese *C. goeringii*, losses of nine *ndh* subunits (*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhH*, *ndhJ*, and *ndhK*), three protein-coding genes (*ycf* 1-like, *ycf* 15, and *ycf* 68), six transfer RNAs, and one conserved open reading frame (*orf* 42). In addition, 219 InDels (insertion or deletion) and 171 simple sequence repeats were observed. Twenty-Five of which InDel markers have been evaluated, that useful for distinguishing Korean and Chinese *Cymbidium* associations based on the polymorphisms of chloroplast genomes between Korean *Cymbidium goeringii* acc. smg222 and Chinese *C. goeringii* and evaluation of genetic diversity. Finally, the phylogenetic relationships of the 39 Korean and 22 Chinese species was constructed based on the five InDel markers of them and obtained high support, indicating that our data may be useful in resolving relationships in this genus. The information about chloroplast DNA structure and gene variants of *C. goeringii* acc. smg222 chloroplast genome will provide sufficient phylogenetic information for resolving evolutionary relationships. The molecular markers developed in here will contribute to further research of *Cymbidium* species and conservation of endemic *Cymbidium* species.

Keywords *C. goeringii* acc. smg222 · Chloroplast genome · InDel marker · Molecular evolution

Heng Wang and So-Yeon Park are contributed equally.

The complete chloroplast genome sequence of *C. goeringii* acc. smg222 has been deposited at GenBank under the accession number MF421552.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-019-2020-1>) contains supplementary material, which is available to authorized users.

✉ Soon-Wook Kwon
swkwon@pusan.ac.kr

- ¹ Department of Plant Bioscience, College of Natural Resources and Life Science, Pusan National University, Miryang 50463, Republic of Korea
- ² State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
- ³ Department of Applied Bioscience, Konkuk University, Seoul 05029, Republic of Korea
- ⁴ Department of Agricultural Science, Korea National Open University, Seoul 03087, Republic of Korea
- ⁵ Saemangeum BioCenter Co, 1044 Heomi-ro, Daeya-myeon, Gunsan 54061, Republic of Korea

Introduction

Chloroplasts contain their own genome and have similarities DNA structure (i.e., number, content, and order of genes) in green plants. The complete chloroplast genome usually consists of a conserved quadripartite structure, i.e., a pair of inverted repeats (IRs) separated by a large single copy (LSC) region and a small single copy (SSC) region. In typical flowering plants, most complete chloroplast genomes are consist of a complete circular molecule ranging from 120 to 160 kb in size (Wicke et al. 2011). In addition, the highly conserved chloroplast genomes typically contain 4 ribosomal RNAs, approximately 30 transfer RNAs, and up to 80 unique proteins (Guisinger et al. 2011). Although the structure and gene contents of chloroplast genomes are highly conserved with respect to their basic structure, comparative genomic studies have revealed some species have a unique chloroplast genome structure (i.e., gene losses and gains, inversions, and large-scale genomic rearrangements) (Asaf et al. 2017). The most obvious examples of several reduced chloroplast

genomes have been found in the parasitic plants *Cuscuta* (Funk et al. 2007), *Epifagus* (Wolfe et al. 1992), and *Rhizanthella* (Delannoy et al. 2011), which have lost some or all of their capacity for photosynthetic or the photosynthetic genes have become pseudo. The loss of chloroplast genes can occur if the gene is transferred and integrated into its hosts' nuclear genome or is functionally replaced by a nuclear gene (Dong et al. 2013). Another speculation is that some species have a compact chloroplast DNA to facilitate adaptation to harsh and competitive conditions (Wicke et al. 2011).

Cymbidium orchid (*Orchidaceae*) is one of the most popular orchids in the world. The 44 species included in genus *Cymbidium* are dispersed widely in tropical, subtropical Asia and northeastern Australia (Du Puy et al. 2007). Spring orchid (*C. goeringii*) is one of the most important species belonging to *Orchidaceae* owing to its aesthetic appeal, variegated leaves, scented flowers, typical characteristics, and highly decorative flower structures for use as a decorative pot flowers. Selection based on somatic mutations at the vegetative propagation and hybrids derived from the 44 *Cymbidium* species, Wang et al. (2009) and Liu et al. (2014) have produced the most commercially important *Cymbidiums* which are attractive, long lasting and larger-flowers. These currently recognized variants are often complex, involving several species in their ancestry (Obara-Okeyo and Kako 1998). Recently, various molecular methods, such as random amplified polymorphic DNA (RAPD) (Obara-Okeyo and Kako 1998; Wang et al. 2004; Choi et al. 2006), amplified fragment length polymorphism (AFLP) (Wang et al. 2004), inter-simple sequence repeat (ISSR) (Wang et al. 2009), simple sequence repeat (SSR) (Moe et al. 2010), and expressed sequence tag-simple sequence repeat markers (EST-SSR) (Liu et al. 2014), have been used to detect genetic diversity in *Cymbidium* resources. However, in general, information on genetic diversity within and between horticultural groups of *C. goeringii* is still quite limited, and the genetic relationships among major excellent lines remain unclear because of limited number of *Cymbidium* DNA sequences (Moe et al. 2010; Choi et al. 2006). Chloroplast as one of the essential cell organelles and its genome act as a single genome, it has become an effective method to analysis plant genetic population and phylogenetic at lower taxonomic levels (Parks et al. 2009).

The Korean landrace, *C. goeringii* acc. smg222 was used to analyze the complete chloroplast genome in current study. Special emphasis was placed on changes in genome structure, variation in the number, content, and order of genes and genome sizes, and the interrelationship among these factors. In addition, we developed polymorphic DNA markers that could efficiently distinguish Korean and Chinese *Cymbidium* accessions and apply for genetic diversity among *Cymbidium* species.

Materials and methods

Sampling and DNA extraction

Cymbidium goeringii acc. smg222 plants from the Saemangeum Bio Center, Republic of Korea were used. A total of 61 *Cymbidium* accessions were collected for validation of sequence variation (Table S1). DNA was extracted from young leaf using a DNeasy Plant Mini Kit (QIAGEN, Valencia, CA, USA) according to the manufacturer's instructions. The relative purity and concentration of extracted DNA were estimated using the NanoDrop ND-1000 (NanoDrop Technologies, Inc., Wilmington, DE, USA). The final concentration of the DNA sample was adjusted to 20 ng/ml.

Library preparation and sequencing

An Illumina paired-end DNA library (average insert size of 500 bp) was constructed using the Illumina TruSeq Library Preparation Kit (San Diego, CA, USA) following the manufacturer's instructions. The library was sequenced (2 × 300 bp) using the MiSeq instrument at LabGenomics (<https://www.labgenomics.co.kr/>).

Chloroplast genome assembly

Prior to de novo assembly of the chloroplast genome, low-quality sequences (quality score < 30; Q30) were filtered out. The remaining high-quality reads were assembled using the CLC Genome Assembler (version beta 4.6; CLC Bio, Aarhus, Denmark) with a minimum overlap size of 200 bp and maximum bubble size of 50 bp for the de Bruijn graph. Chloroplast contigs were selected from the initial assembly by performing a BLAST (version 2.2.31) search against the reference chloroplast genome of *C. kanran* (GenBank accession: KU179435) using CLC with the following parameters: 0.5 for length fraction, 0.8 for similarity, and 200–600 bp of overlap (Jo et al. 2011). The selected chloroplast contigs were merged into four contigs, and iterative contig extensions were performed to construct a complete *C. goeringii* acc. smg222 chloroplast genome by mapping raw reads to the contigs. Ambiguous nucleotides or gaps were corrected manually to build the complete cp genome.

Gene annotation and comparative analysis of genome structure

Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al. 2004) and CpGAVAS (Liu et al. 2012) were used to annotate the assembled chloroplast genome using default parameters to predict protein-coding, rRNA, and tRNA

genes. To further confirm the identified tRNA genes, the tRNA scan-SE search server was used to predict their corresponding structures (Schattner et al. 2005). Subsequently, BLASTN was used to further identify intron-containing genes by searching against a published cp genome of *C. kanran*. The genome map of *C. goeringii* acc. smg222 was drawn using Organellar Genome DRAW (OGDRAW, <https://ogdraw.mpimp-golm.mpg.de>). The mVISTA program was employed to plot the sequence identity by comparing the complete chloroplast genomes of *C. goeringii* acc. smg222 with Chinese *C. goeringii*.

Discovery of variations and primers design

SSR markers in the chloroplast genome of *C. goeringii* acc. smg222 were found using Sputnik (<https://espressosoftware.com/pages/sputnik.jsp>) software. It uses a recursive algorithm to search for repeats with lengths between 2 and 5 nucleotides, and finds perfect, compound, and imperfect repeats. To designate SSR in many species, including *Arabidopsis* and barley, Sputnik has been utilized (Cardle et al. 2000). BWA (Li and Durbin 2009) was used with the 'mem' command line options '-k19 -w100 -d100 -r1.5 -y20 -c500 -D0.5 -W0 -m50' and SAMtools was used with the 'mpileup' for identification of SNP and InDel variants in the chloroplast genome of *C. goeringii* acc. smg222. The detailed methods and algorithms are described in Li (2012). 25 InDels were selected to develop markers using Oligo7.0 software to screen for polymorphisms in Korean and Chinese *cymbidium* accessions. PCRs were performed in a 20- μ l volume containing 10 ng of DNA template, 10 pmol of each primer, 1X PCR buffer, 0.2 mM of dNTPs, and 1 unit of Taq DNA polymerase (Nurotics, Korea). PCRs were performed with an MJ Research PTC-100 thermocycler (Waltham, MA, USA) using the following conditions: initial denaturation at 94 °C for 5 min, followed by 36 cycles of denaturation at 94 °C for 30 s, annealing at 58 °C for 30 s, and extension at 72 °C for 1 min, with a final extension at 72 °C for 10 min. The PCR products were separated during 100 min using Fragment Analyzer™ (Advanced Analytical Technologies, Inc).

Genetic diversity and phylogenetic analysis

Genetic variability measured was based on the average number of alleles, gene diversity, major allele frequency and polymorphic information content (PIC) were calculated for each InDel markers using PowerMarker version 3.25 (Liu and Muse 2005, <https://brcwebportal.cos.ncsu.edu/powermarker/index.html>). Phylogenetic tree was constructed by the arithmetic average (UPGMA) method implemented in PowerMarker 3.0. Genetic distance was constructed using MEGA 7.0 software (Kumar et al. 2016, <https://www.megasoftware.net/>).

Results and discussion

Chloroplast genome sequencing

Illumina as one of the most widely employed next-generation sequence technology has been significantly improved the field of high throughput sequencing data collection. Illumina sequencing can generate high-quality sequence assemblies covering a greater genome depth with an average error rate of 1–1.5%. A previous research demonstrated that genome sequencing depths of greater than tenfold have little additional effect on genome coverage (Li et al. 2008). Using the Illumina platform, the coverage depth was high (356 $\times$) for the *C. goeringii* acc. smg222 chloroplast genome. Next-generation sequence technologies are subject to sequence dependent the GC content bias, in which the level of GC content effect the number of reads produced during high-throughput sequencing (Price et al. 2017). The complete chloroplast genome sequence of *C. goeringii* acc. smg222 was 37.1% overall GC content (Wang et al. 2018), similar to those of other *Cymbidium* species. These indicates that the whole-genome sequencing approach for the chloroplast of *C. goeringii* acc. smg222 was successful, yielding an reasonable sequencing depth and coverage. Using the Illumina platform, there were four contigs comprising 148,441 bp covering the whole chloroplast genome of *C. goeringii* acc. smg222 (Fig. 1). However, previously reported *Cymbidium* chloroplast genomes range in length from 154,769 to 156,904 bp (Yang et al. 2013), inconsistent with the size of the chloroplast genome of *C. goeringii* acc. smg222 in recent study. The *C. goeringii* acc. smg222 complete chloroplast genome was 8,751 bp shorter than the Chinese *C. goeringii* chloroplast genome (GenBank accession: NC_028524) and 6,328–8,463 bp shorter than *C. tortisepalum* 1, *C. tortisepalum* 2, *C. tortisepalum* 3, *C. mannii* 1, *C. mannii* 1, *C. aloifolium*, *C. sinense*, and *C. tracyanum* chloroplast genomes (GenBank accessions: KC876124, KC876128, KC876125, KC876129, KC876126, KC876122, KC876123, and KC876127). Additionally, the LSC, SSC, and IR regions of *C. goeringii* acc. smg222 were 1,609–2,330, 2,619–4,019, and 711–1,100 bp shorter than those of other genomes, respectively. Species with shorter life cycles tend to have shorter chloroplast genomes than those of species with longer life cycles (Dong et al. 2013). Differences in genome size might indicate chloroplast genome evolution in *C. goeringii* acc. smg222.

Fig. 1 Gene map of complete chloroplast genome of the *C. goeringii* acc. smg222. Gene on the outside of the map are transcribed in the clockwise direction, while other genes on the inside of the map are

transcribed in the counterclockwise direction. Differential functional gene groups are color-coded. The dark gray plot in the inner circle corresponds to GC content

Features of the *C. goeringii* acc. smg222 chloroplast genome

The 70 protein-coding, 30 transfer RNA, and 4 ribosomal RNA genes were composed as 104 unique genes in the *C. goeringii* acc. smg222 chloroplast genome (Wang et al. 2018). Among the 70 protein-coding genes, 21 ribosomal subunit genes consist of 12 small subunits and 9 large

subunits and DNA-directed RNA polymerase genes were 4. Thirty-seven genes were linked with photosynthetic process, of which two were encoded subunits of the NADH oxidoreductase, seven were in photosystem I, 15 were in photosystem II, six were in the cytochrome b6/f complex, six encoded different subunits of ATP synthase, and one encoded the large chain of ribulose biphosphate carboxylase. Six genes had functions unrelated to photosynthesis,

and two genes were of unknown function. Seven distinct genes [*petB*, *pteD*, *atpF*, *rpl16*, *rps16*, *rpoC1*, and *rpl2* (IR)] and six tRNA genes (*trnA*-UGC, *trnG*-GCC, *trnI*-GAU, *trnK*-UUU, *trnL*-UAA, and *trnV*-UAC) contained one intron, and three genes (*ycf3*, *rps12*, and *clpP*) contained two introns. We also detected some alternative start codons, for instance, ACG for *rpl2*, ACG for *rps12*, and GTG for *rps19*. Non-canonical start codons have been reported in other angiosperms (Raubeson et al. 2007) and ferns (Gao et al. 2009). From the above, the gene order of *C. goeringii* acc. smg222 was generally in agreement with that of previously reported chloroplast genomes of the genus *Cymbidium*.

Gene content differences

By aligning the chloroplast of *C. goeringii* acc. smg222, DNA with Chinese *C. goeringii* genome was shown (Fig. 2).

C. goeringii acc. smg222 exhibited a small genome size, gene losses and intron losses. One exception was that the number and content of genes were not similar to the sequenced chloroplast DNAs of *C. goeringii* with respect to 11 subunits of *ndh*; in *C. goeringii* acc. smg222, *ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhH*, *ndhJ*, and *ndhK* were loss-of-function. We only detected remnants of *ndhG* and *ndhI* sequences. In the two *C. goeringii* varieties chloroplast genome studied, *ndhA* and *ndhC* genes are truncated (partial sequence remained), of which *ndhA* has a truncation in the 3'-end and *ndhC* has a truncation occurring at the 5'-end; *ndhF*, *ndhH*, *ndhJ* and *ndhK* genes are absent (no sequence exists); in *ndhB*, two nucleotides substitution occurred, *ndhD* has a 17 bp deletion and *ndhE* has a 2 bp insertion, creating a frameshift mutation. The non-functionality of *ndh* genes also occur in other orchids (Chang et al. 2006; Wu et al. 2010). In Oncidiinae varieties, most *ndh* genes, with the exception of *ndhE* in some species, have no function;

Fig. 2 Visualization of alignment of the Korean and Chinese *Cymbidium goeringii* chloroplast genome sequences. The vertical scale indicates the percentage of identity, ranging from 50 to 100%. The hori-

zontal axis indicates the coordinates within the chloroplast genome. Genome regions are color coded as protein coding, rRNA coding, tRNA coding or conserved noncoding sequences

ndhB has a stop codon in the first exon, *ndhD* and *ndhJ* are truncated, *ndhK* and *ndhF* are absent, and *ndhC* has a frameshift mutation, creating a 17 bp deletion and a premature stop codon (Wu et al. 2010). *Phalaenopsis aphrodite* lacks *ndhA*, *ndhF*, and *ndhH*, and contains only remnants of the other eight subunit sequences, moreover, the 11 *ndh* genes are presumably nonfunctional based on either truncation or frameshift mutations (Chang et al. 2006). Here, we displayed that the truncation and/or absence of *ndh* genes in the chloroplast genome region is a general phenomenon in *Cymbidium*, not restricted to *Oncidiinae* and *Phalaenopsis*.

Compared with those of other *Cymbidium* chloroplast genomes, *C. goeringii* acc. smg222 also exhibited the losses of three protein-coding genes with unknown function (*ycf* 1-like, *ycf* 15, and *ycf* 68), six transfer RNAs, and one conserved open reading frame (*orf* 42), of which, *ycf* 1-like as an incomplete duplication of the normal *ycf* 1 is absent and the non-functionality of *ycf* 68 gene is located in the intron of *C. goeringii* acc. smg222 *trnI*-GAU gene. In the previous reports, *ycf* 1-like, *ycf* 15 and *ycf* 68 genes are non-functional pseudogenes in other *Cymbidium* genomes, but they are conserved in IRs among different species (Yang et al. 2013). We also detected the loss-of-function *ORF42* gene is located in the intron of *C. goeringii* acc. smg222 *trnA*-UGC gene.

ORF42 is a fragment of the *pvs-trnA* gene, which is related to a mitochondrial gene (Do et al. 2013); however, the characteristics and function of *ORF42* remain to be determined. Based on these observations, it has been hypothesized that selection for a reduced nuclear genome is beneficial owing to its lower maintenance cost when deletions are not deleterious or genes are dispensable Gray et al. (1998). suggested that an economic strategy can explain organellar genomes in which pseudogenes are usually rare and the non-coding content is low. In gnetophytes, the reduced chloroplast genome with deletions of genes and non-CDSs increases survivorship in harsh and competitive environments. It is reasonable to conclude that *C. goeringii* acc. smg222 might exist in resource-poor conditions and that a low-cost strategy is indispensable for survival.

Discovery of DNA variations

We analyzed the occurrence, type, and distribution of SSRs in the *C. goeringii* acc. smg222 chloroplast genome. In total, we identified 171 SSRs (Table S2). The majority of the SSRs in this chloroplast genome are mono-nucleotides 108 (63.2%) followed by di-nucleotides 53 (31.0%). Besides, tri-nucleotides 3 (1.8%), tetra-nucleotides 5 (2.9%),

Fig. 3 SSR candidates of *C. goeringii* acc. smg222 chloroplast genome. (A: A mono-nucleotides; T: T mono-nucleotides; C: C mono-nucleotides; G: G mono-nucleotides; di: di-nucleotides; tri: tri-nucleotides; tetra: tetra-nucleotides; com: compound nucleotides)

and compound SSRs 2 (1.2%) motifs were present in the *C. goeringii* chloroplast genome (Fig. 3). Among these, 113 were localized in intergenic spacers and 58 were in CDS region. Here, the detected frequency of SSR sites was about 1 per 868 bp of the chloroplast genome. Furthermore, these regions can be applied to intraspecific polymorphism and used for population structure and phylogeographic studies for *Cymbidium*.

In an equivalence of the chloroplast genome sequence of *C. goeringii* acc. smg222 with the Chinese *C. goeringii* one, we detected 219 InDels (Table S3). Among these InDels, we observed 31 InDels in coding regions and 152, 13, 41, and 13 mutations located in the LSC, IRa, SSC, and IRb regions, respectively. These InDel loci observed in this study is valuable for further studies in genetic diversity, population structure, and evolutionary analyses.

InDel markers diversity in *Cymbidium*

Complete chloroplasts genome sequences have been widely accepted as valuable and informative genomic resources for understanding evolutionary biology and practical application of DNA markers because of their relatively stable genome

structure and uniparental inherited features (Dong et al. 2012; Kim et al. 2015; Suo et al. 2012; Song et al. 2015).

Twenty-five InDel-based markers were developed by their sequence variations (Table 1) and were analyzed genetic diversity between 10 Korean and 10 Chinese *Cymbidium* accessions (Table 2). The mean major allele number in *Cymbidium* accessions was 2.32 and the average number of alleles in Korean *Cymbidium* accessions (Subgroup 1) and Chinese *Cymbidium* accessions (Subgroup 2) was 1.28 and 2.32, respectively. The overall PIC values ranged from 0 to 0.6324 with an average of 0.2360 (Table 2). These markers have a higher diversity in Subgroup 1 than Subgroup 2. In here, the InDel marker with one allele in a Subgroup and the frequency of the allele in another Subgroup (F_{sub1}) is 0 was selected as the specific marker for genetic diversity analyzing. Such as, the F_{sub1} values of InDel 001,520, InDel 058,556 and InDel 076,063 markers were 0 that means these InDel markers were specific for Subgroup 1. The F_{sub1} and F_{sub2} values of InDel 009,101 and InDel 116,414 markers were 0 that shows the two InDel markers were specific for Subgroup 1 and Subgroup 2. InDel 015,557, InDel 070,173, InDel 097,251, InDel 102,357, InDel 110,602, InDel 133,740 and InDel 147,362 had the specific alleles

Table 1 Detail of InDel markers developed in this study

No.	ID	Left primer	Right primer	Locus
1	ID001520	GGAAGTAGTCGGATGGAGTAGA	CATCATATTCGTGGTGAGATTG	<i>trnK-UUU~matK</i>
2	ID003480	TTTTGTGGCCGAAATCTATCTT	GATCTGTAGATTGGGCTCTCTG	<i>matK~trnK-UUU</i>
3	ID009101	CATTTTCATTGCCATTCCTAAT	TTTCATTGATGAATTTCCGAAT	<i>trnS-GCU~trnG-UCC</i>
4	ID012838	CCCATATTTTTGAGCCTATCCT	CGAATCGACGACCTATGTATTG	<i>atpF~atpF</i>
5	ID015557	ACTTGGGCATGAATTTGTAAAC	TATTTTGGTTTGCATCTTTTGG	<i>atpI~rps2</i>
6	ID037977	AAGATTCCGTCGGTATAGTTGA	AGTAGAGCAGTTTGGTAGCTCG	<i>trnG-GCC~trnfM-CAU</i>
7	ID044404	TATCTTTTCTCCACCTTCAGA	TAAAGTATTGAGCAGCGGTGTA	<i>ycf3~ycf3</i>
8	ID048067	CATTACAAATGCGATGCTCTAA	TTCCTCCTCCTTCATTTTACA	<i>trnT-UGU~trnL-UAA</i>
9	ID053814	CTTCATTGTGTTCAATTTGTGG	GGCGGTTTTGCTAGAATAAGTA	<i>trnM-CAU~atpE</i>
10	ID058556	CAGCAATTCCTTTTTGTCTTC	TCTATGGCCTGAACTAAGGAA	<i>rbcL~accD</i>
11	ID065011	TACGAGTCCAAGGTCTTCTGTT	AGGAGGAATCCATCGATTTTAT	<i>petA~psbJ</i>
12	ID070173	TTTTCTTCTATCTTCCCGGAGT	TTTGGTTTCTTCTCATCGAAAT	<i>rps18~rpl20</i>
13	ID073915	GACAAGTCGCACTATACGTCAA	ATCCCCCTTCGTTACAATCTTTC	<i>clpP~psbB</i>
14	ID076063	TACAAAATTCCTTTGCCATCTT	CCACACCTATTCATTTTGGATT	<i>psbB~psbT</i>
15	ID078919	TGTGGAATGATCCAATGTTCTA	CCTGTTCTTCCTTAGATCCCTT	<i>petD~petD</i>
16	ID096984	TCCAAACGGACTCCTATAAAGA	GACATGACCGATCGATAGAAAT	<i>ycf2~trnL-CAA</i>
17	ID097251	TCTATCGATCGGTCATGTCATA	TATTATGAGAAGGGGTCATTTCG	<i>trnL-CAA~rps7</i>
18	ID102357	ATCCTTTTCGATGACCTATGTTG	TCTCTCATGGTACAACCCTCTT	<i>rps12~trnV-GAC</i>
19	ID110602	GCGACAGAAGTATTGAGAATCC	TACAGTATCGTCACCGCAGTAG	<i>rrn4.5~rrn5</i>
20	ID116414	CCAAATTTCCATTTTGAATTG	GATTTTGAGACCCAACACCTTA	<i>rpl32~trnL-UAG</i>
21	ID116964	TATTATTTAAGGTAAGCCGCCA	TCACAATTGAGATGATCGAAAA	<i>trnL-UAG~ccsA</i>
22	ID120564	GCCAAGATACTTTGATTTCAT	TTAGTCATAATTCTTCTATTTTCCCA	<i>psaC~ndhG</i>
23	ID133740	TACAGTATCGTCACCGCAGTAG	GCGACAGAAGTATTGAGAATCC	<i>rrn5~rrn4.5</i>
24	ID147109	TATTATGAGAAGGGGTCATTTCG	TCTATCGATCGGTCATGTCATA	<i>rps7~trnL-CAA</i>
25	ID147362	GACATGACCGATCGATAGAAAT	TCCAAACGGACTCCTATAAAGA	<i>trnL-CAA</i>

Table 2 Summary statistics for the 25 InDel markers analyzed in this study

Marker	Subgroup1 (K01~K10)		Subgroup2 (C01~C10)		All		F _{sub1} ^a	F _{sub2}	Remark
	Allele no. (A)	PIC value	Allele no. (B)	PIC value	Allele no. (C)	PIC value			
ID001520	1	0.0000	3	0.4992	4	0.5812	0	—	Specific for sub1
ID003480	1	0.0000	2	0.3648	2	0.3318	0.4	—	
ID009101	1	0.0000	1	0.0000	2	0.3750	0	0	Specific for sub1 and sub2
ID012838	2	0.1638	1	0.0000	2	0.0905	—	0.9	
ID015557	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
ID037977	2	0.1638	2	0.2688	3	0.2469	—	—	
ID044404	1	0.0000	2	0.1638	2	0.0905	0.9	—	
ID048067	3	0.2688	3	0.1769	3	0.4064	—	—	
ID053814	2	0.1638	3	0.5478	4	0.6324	—	—	
ID058556	1	0.0000	3	0.4992	4	0.5812	0	—	Specific for sub1
ID065011	1	0.0000	2	0.1638	2	0.0905	0.9	—	
ID070173	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
ID073915	1	0.0000	3	0.4992	3	0.4662	0.2	—	
ID076063	1	0.0000	2	0.2688	3	0.4918	0	—	Specific for sub1
ID078919	1	0.0000	4	0.6035	4	0.4059	0.5	—	
ID096984	1	0.0000	2	0.0905	2	0.0476	0.9	—	
ID097251	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
ID102357	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
ID110602	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
ID116414	1	0.0000	1	0.0000	2	0.3750	0	0	Specific for sub1 and sub2
ID116964	1	0.0000	2	0.1638	2	0.0905	0.9	—	
ID120564	2	0.3648	5	0.6005	5	0.5054	—	—	
ID133740	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
ID147109	2	0.1638	1	0.0000	2	0.0905	—	0.9	
ID147362	1	0.0000	1	0.0000	1	0.0000	1	1	Monomorphic
Mean	1.28	0.0516	1.96	0.1964	2.32	0.2360			

PIC polymorphism information content

^aF_{sub1} frequency of subgroup1 specific allele in subgroup2, F_{sub2} frequency of subgroup2 specific allele in subgroup1

in Subgroup 1 and Subgroup 2, respectively, and the values of F_{sub1} and F_{sub2} were 1 that means the specific band were monomorphic (Table 2).

Finally, five InDel markers (ID001520, ID009101, ID058556, ID076063 and ID116414) were selected as Subgroup 1 specific markers and were tested among 61 *Cymbidium* accessions (Table S1). The PCR results of InDel 001,502 marker shown a special band in Korean *Cymbidium* accessions and polymorphism among Chinese *Cymbidium* accessions. The InDel 009,101 and ID116414 markers are specific for Korean and Chinese *Cymbidium* accessions, respectively, except one was not same as other. The InDel 058,556 and InDel 076,063 markers also shown specially in Korean *Cymbidium* accessions, but diversity in Chinese *Cymbidium* accessions (Table S1). These results almost same as above that demonstrated the sequence of *C. goeringii* acc. smg222 chloroplast was reliable.

The neighbor-joining tree of 61 *Cymbidium* accessions based on Nei's genetic distance revealed that Korean and

Chinese *Cymbidium* accessions are divided into two different groups (Fig. 4). The phylogenetic relationship of the *Cymbidiums* is associated with the geographic characteristics, indicating that these InDel markers may be useful for authentication of *Cymbidium* genetic resources across different regions and for improving germplasm conservation methods.

Conclusion

With the unique inherited nature of chloroplasts, where they are inherited maternally, the chloroplast genome sequence can provide useful information on plant evolution, systematics, and biogeography owing to its high conservation among species. The analysis of complete chloroplast genome sequence of Korean *Cymbidium goeringii* revealed genetic variation with Chinese *Cymbidium* and provided useful DNA makers such as 171 SSRs and 219 InDel. Among

Fig. 4 Neighbour-joining phylogenetic tree of the *Cymbidium* inferred from the five InDel markers of 61 *Cymbidium* accessions. A: Korean *Cymbidium* group; B: Chinese *Cymbidium* group

them 31 markers further evaluated by comparison among 10 Korean *Cymbidium goeringii* and 10 Chinese *Cymbidium*. In the validation analysis of the developed makers for the constructing the neighbor-joining tree among 61 *Cymbidium* accessions, 5 makers successfully distinguished Korean *Cymbidium* from Chinese *Cymbidium*. The results from this analysis can be used for further studies in the field of

biogeography, plant biotechnology, and population structure studies in *Cymbidium*, as well as phylogenetic studies, of species of this genus.

Acknowledgements Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) provided this research through Technology Commercialization Support

Program. The finance was contributed by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (815004–03).

Author contributions HW, SYP and SWK developed ideas, designed and performed all experiments, and wrote the manuscript. YCK and DYK produced all cymbidium materials. THH, JL, and TSK analyzed the sequencing dates. SHS and SML analyzed the InDel markers for experiments. All authors read and approved the final manuscript.

Funding This work was supported by grants from Ministry of Agriculture, Food and Rural Affairs of Korea (MAFRA).

Compliance with ethical standards

Conflict of interest All the authors declare that they have no conflict of interest in the publication.

References

- Asaf S, Khan AL, Khan MA, Imran QH, Kang SM, Al-Hosni K, Jeong EJ, Lee KE, Lee IJ (2017) Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species. *PLoS ONE* 12:e0182281
- Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156(2):847–854
- Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Cheng CH, Lin CY, Liu SM, Chang CC, Chaw SM (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol Biol Evol* 23(2):279–291
- Choi SH, Kim MJ, Lee JS, Ryu KH (2006) Genetic diversity and phylogenetic relationships among and within species of oriental cymbidiums based on RAPD analysis. *Sci Hortic* 108(1):79–85
- Delannoy E, Fujii S, des Francs-Small CC, Brundrett M, Small I (2011) Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Mol Biol Evol* 28(7):2077–2086
- Do HDK, Kim JS, Kim JH (2013) Comparative genomics of four *Liliales* families inferred from the complete chloroplast genome sequence of *Veratrum patulum* O. Loes. (Melanthiaceae). *Gene* 530(2):229–235
- Dong W, Liu J, Yu J, Wang L, Zhou S (2012) Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* 7(4):e35071
- Dong W, Xu C, Cheng T, Zhou S (2013) Complete chloroplast genome of *Sedum sarmentosum* and chloroplast genome evolution in Saxifragales. *PLoS ONE* 8(10):e77965
- Du Puy D, Cribb P, Tibbs M (2007) The genus *Cymbidium*. Royal Botanic Gardens, Kew
- Funk HT, Berg S, Krupinska K, Maier UG, Krause K (2007) Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta groenovii*. *BMC Plant Biol* 7(1):45
- Gao L, Yi X, Yang YX, Su YJ, Wang T (2009) Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. *BMC Evol Biol* 9(1):130
- Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Tim GL, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger G (1998) Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res* 26(4):865–878
- Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol Biol Evol* 28(1):583–600
- 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(2):217–229
- Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, Pakr HS, Yang TJ (2015) Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. *PLoS ONE* 10(6):e0117159
- Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33(7):1870–1874
- Li H (2012) Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* 28(14):1838–1844
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760
- Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24(5):713–714
- Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21(9):2128–2129
- Liu 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(1):715
- Liu X, Huang Y, Li F, Xu C, Chen K (2014) Genetic diversity of 129 spring orchid (*Cymbidium goeringii*) cultivars and its relationship to horticultural types as assessed by EST-SSR markers. *Sci Hortic* 174:178–184
- Moe KT, Zhao W, Song HS, Kim YH, Chung JW, Cho YI, Park PH, Park HS, Chae SC, Park YJ (2010) Development of SSR markers to study diversity in the genus *Cymbidium*. *Biochem Syst Ecol* 38(4):585–594
- Obara-Okeyo P, Kako S (1998) Genetic diversity and identification of *Cymbidium* cultivars as measured by random amplified polymorphic DNA (RAPD) markers. *Euphytica* 99(2):95–101
- Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol* 7(1):84
- Price A, Garhyan J, Gibas C (2017) The impact of RNA secondary structure on read start locations on the Illumina sequencing platform. *PLoS ONE* 12(2):e0173023
- Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genom* 8(1):174
- Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33(suppl 2):W686–W689
- Song Y, Dong W, Liu B, Xu C, Yao X, Gao J, Corlett RT (2015) Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Front Plant Sci* 6:662
- Suo Z, Zhang C, Zheng Y, He L, Jin X, Hou B, Li J (2012) Revealing genetic diversity of tree peonies at micro-evolution level with hyper-variable chloroplast markers and floral traits. *Plant Cell Rep* 31(12):2199–2213
- Wang HZ, Wang YD, Zhou XY, Ying QC, Zheng KL (2004) Analysis of genetic diversity of 14 species of *Cymbidium* based on RAPDs and AFLPs. *Shi Yan Sheng Wu Xue Bao* 37(6):482–486

- Wang HZ, Wu ZX, Lu JJ, Shi NN, Zhao Y, Zhang ZT, Liu JJ (2009) Molecular diversity and relationships among *Cymbidium goeringii* cultivars based on inter-simple sequence repeat (ISSR) markers. *Genetica* 136(3):391–399
- Wang H, Park SY, Lee AR, Jang SG, Im DE, Jun TH, Lee JH, Chung JW, Kwon SW (2018) Next-generation sequencing yields the complete chloroplast genome of *C. goeringii* acc. smg222 and phylogenetic analysis. *Mitochondr DNA B* 3(1):215–216
- Wicke S, Schneeweiss GM, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* 76(3–5):273–297
- Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci* 89(22):10648–10652
- Wu FH, Chan MT, Liao DC, Hsu CT, Lee YW, Daniell H, Duval MR, Lin CS (2010) Complete chloroplast genome of *Oncidium Gower Ramsey* and evaluation of molecular markers for identification and breeding in *Oncidiinae*. *BMC Plant Biol* 10(1):68
- Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20(17):3252–3255
- Yang JB, Tang M, Li HT, Zhang ZR, Li DZ (2013) Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol Biol* 13(1):84