Gene 530 (2013) 229–235

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of *Veratrum patulum* O. Loes. (Melanthiaceae)

Hoang Dang Khoa Do, Jung Sung Kim, Joo-Hwan Kim *

Department of Life Science, Gachon University, Seongnam, Gyeonggi-do 461-701, Republic of Korea

ARTICLE INFO

Article history:
Accepted 29 July 2013
Available online 23 August 2013

Keywords:
Veratrum patulum
Melanthiaceae
Plastid genome
Liliales

ABSTRACT

The sequence of the chloroplast genome, which is inherited maternally, contains useful information for many scientific fields such as plant systematics, biogeography and biotechnology because its characteristics are highly conserved among species. There is an increase in chloroplast genomes of angiosperms that have been sequenced in recent years. In this study, the nucleotide sequence of the chloroplast genome (cpDNA) of *Veratrum patulum* Loes. (Melanthiaceae, Liliales) was analyzed completely. The circular double-stranded DNA of 153,699 bp consists of two inverted repeat (IR) regions of 26,360 bp each, a large single copy of 83,372 bp, and a small single copy of 17,607 bp. This plastome contains 81 protein-coding genes, 30 distinct tRNA and four genes of rRNA. In addition, there are six hypothetical coding regions (*ycf1*, *ycf2*, *ycf3*, *ycf4*, *ycf15* and *ycf68*) and two open reading frames (*ORF42* and *ORF56*), which are also found in the chloroplast genomes of the other species. The gene orders and gene contents of the *V. patulum* plastid genome are similar to that of *Smilax china*, *Lilium longiflorum* and *Alstroemeria aurea*, members of the Smilacaceae, Liliaceae and Alstroemeriaceae (Liliales), respectively. However, the loss *rps16* exon 2 in *V. patulum* results in the difference in the large single copy regions in comparison with other species. The base substitution rate is quite similar among genes of these species. Additionally, the base substitution rate of inverted repeat region was smaller than that of single copy regions in all observed species of Liliales. The IR regions were expanded to *trnH_GUG* in *V. patulum*, a part of *rps19* in *L. longiflorum* and *A. aurea*, and whole sequence of *rps19* in *S. china*. Furthermore, the IGS lengths of *rbcL-accD-psaI* region were variable among Liliales species, suggesting that this region might be a hotspot of indel events and the informative site for phylogenetic studies in Liliales. In general, the whole chloroplast genome of *V. patulum*, a potential medicinal plant, will contribute to research on the genetic applications of this genus.

1. Introduction

The chloroplast, an important organelle that performs photosynthesis in green plants, originated from photosynthetic bacteria that interacted with non-photosynthetic hosts through endosymbiosis (Howe et al., 2003). Therefore, the chloroplast has its own genome structure, which is generally composed of four parts including large single copy (LSC), small single copy (SSC) and two inverted repeat (IR) regions. The size of the chloroplast genome ranges from less than 100 kb to more than 160 kb due to the evolutionary contraction and expansion of IR regions. The structure and gene contents of chloroplast DNA (cpDNA) are highly conserved. However, some species have a unique cpDNA structure (i.e., the number, content and order of genes) due to adaptations to specific conditions (Wicke et al., 2011).

Morphological data are necessary to identify the species and to discuss their relationships among the species in phylogenetic studies; however, homoplasy often interferes with the interpretation of phylogenetic relationships. Molecular data have been used to address this complication. Chloroplast genes are highly conserved among the species. For this reason, the structural characteristics of cpDNA are useful for phylogenetic studies. In recent decades, many studies of plant phylogenetics have been conducted using partial or whole genes of chloroplast genomes (Jansen et al., 2007; Moore et al., 2010). Since 1986, when the first complete chloroplast genome (that of the tobacco plant) was published (Shinozaki et al., 1986), the number of complete chloroplast genomes submitted to the National Centre for Biotechnology Information (NCBI, USA) has been increased significantly. Currently, there are more than 250 completely sequenced chloroplast genomes that can be downloaded from NCBI Organelle Genome Resources (http://www.ncbi.nlm.nih.gov/genomes). However, most of these complete cpDNA sequences are from dicots. Therefore, additional studies

Abbreviations: cpDNA, Chloroplast genome; LSC, Large single copy; SSC, Small single copy; IR(s), Inverted repeat region; IRA, Inverted repeat region A; IRB, Inverted repeat region B; ORF, Open reading frame; IGS, Intergenic spacer; PCR, Polymerase chain reaction; SSR, Single sequence repeat; bp, Base pairs; Ka, Non-synonymous substitution; Ks, Synonymous substitution.

* Corresponding author. Tel.: +82 31 750 8827; fax: +82 31 750 8738.
E-mail addresses: dohoangdangkhoa@yahoo.com (H.D.K. Do), jskim2010@gachon.ac.kr (J.S. Kim), kimjh2009@gachon.ac.kr (J.-H. Kim).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.07.100

examining the chloroplast genomes of monocots are required to represent the large biodiversity of this monophyletic group.

The genus *Veratrum*, also referred to as false hellebores or corn lilies, was first recognized by Linnaeus in 1753. Subsequently, this genus was described as a member of the family Colchicaceae by De Candolle, 1805. However, based on the molecular and morphological data, the Angiosperm Phylogeny Group (2003, 2009) placed *Veratrum* into the family Melanthiaceae, which consists of the five tribes of Melanthieae, Heloniadeae, Chionographideae, Xerophylleae and Parideae. *Veratrum*, a member of the tribe Melanthieae, comprises 26 species, which are distributed mainly in temperate regions of the northern hemisphere (WCSP, 2012). *Veratrum patulum* O. Loes. was placed in *Veratrum* genus based on the following characteristics: inflorescences decompounds racemose, stamen almost half as long as tepals, white spatulate tepals and elliptic leaves (Zomlefer et al., 2003). *Veratrum* are generally toxic because they contain high levels of steroid alkaloids, which cause symptoms such as nausea, bradycardia and hypotension if ingested (Schep et al., 2006). Conversely, extractions of *Veratrum* contain organic compounds that inhibit melanogenesis, so these natural compounds are used as potential skin whitening agents (Jin et al., 2002). In addition, a teratogenic alkaloid cyclopamine found in *Veratrum* is an effective treatment for human tumors (Incardona et al., 1998; Taipale et al., 2000). Although the potential uses of *Veratrum* extractions have been widely studied, genetic information about this genus is lacking (Zomlefer et al., 2003).

We analyzed the chloroplast genome sequence of *V. patulum*, providing the first report in the family Melanthiaceae, and compared it with the plastid genomes of *S. china* (Liu et al., 2012), *Lilium longiflorum* and *Alstroemeria aurea* (Kim and Kim, 2013), members of Smilacaceae, Liliaceae and Alstroemeriaceae, respectively, in the same order (Liliales). Our results provide essential information on the evolution of chloroplast genomes among these families. In addition, these data should be useful for future studies of chloroplast genomes and phylogenomic studies in Liliales.

2. Materials and methods

2.1. Plant samples, cpDNA extraction, sequencing and assembly of chloroplast genome sequences

Fresh leaves of *V. patulum* were collected from its habitat in Korea and a voucher specimen (KWU02202) was deposited in the Herbarium of Gachon University (GCU). To isolate the chloroplasts, we applied the revised Percoll gradient buffer method (Kim and Kim, 2013) using 50 g fresh leaves. The DNeasy Plant Mini Kit (Qiagen, Seoul, South Korea) was used to extract cpDNA from raw chloroplasts. The DNA was sequenced using the GS FLX Titanium platform (Roche Applied Science, Germany) after shearing by nebulization. The Geneious version 6.1, created by Biomatters (Biomatters Ltd. Auckland, New Zealand), was used to assemble the sequencing results. The chloroplast genome contigs were aligned to the cpDNA sequence of *L. longiflorum* (GenBank accession number: KC968977) (Kim and Kim, 2013) to identify gaps among the contigs. To fill the gaps, the purified DNA was used to perform polymerase chain reaction (PCR) with primers designed using Primer3 (Untergrasser et al., 2012). The cpDNA sequence of *L. longiflorum* was used as a reference for primer design. To identify the border of inverted repeat regions, large single copy and small single copy regions, two primer pairs of *ycf1–ndhF* (forward primer: 5′-CGATGTAGAAACAACTTCCGA-3′, reverse primer: 5′-TCGGACTATTCATAGCATCC-3′) and *trnH_GUG–psbA* (forward primer: 5′-CGCGCATGGTGGATTCACAATCC-3′, reverse primer: 5′-GTCATGCACGA ACGTAATGCTC-3′) were designed. The sequences received from these PCR products were used to compare with known sequences of *V. patulum* cpDNA to identify the junctions among regions. We sequenced all PCR products using a BigDye Terminator Kit (Ver 3.1, Life Technologies) based on the manufacturer's recommendations. Sequencher 5.0 (Gene Codes Corporation, USA) was used to assemble complete genomes from high-quality sequences.

2.2. Gene annotation and comparative analysis

The complete sequence of the chloroplast was annotated using DOGMA (Wyman et al., 2004). The tRNAs were confirmed using tRNAscan-SE (Schattner et al., 2005). Other protein-coding regions were checked based on data from the NCBI (http://blast.ncbi.nlm.nih.gov/) and manual corrections for start and stop codons were made. Illustrations of gene features of cpDNA were constructed using a Web-based tool, GenomeVx (Conant and Wolfe, 2008). For comparative analysis of cpDNA features among *V. patulum*, *S. china* (GenBank accession no. HM536959), *L. longiflorum* (GenBank accession no. KC968977), *A. aurea* (GenBank accession no. KC968976) and *Phoenix dactylifera* (GenBank accession no. NC013991), we used DnaSP (Librado and Rozas, 2009) for substitution rates. The codon usage of *V. patulum* was analyzed using MEGA5 (Tamura et al., 2011). Single sequence repeat (SSR) and large single repeats were detected by using Msatfinder program (Thurston and Field, 2005) and REPuter program (Kurtz et al., 2001), respectively.

3. Results

3.1. V. patulum chloroplast genome assembly and features

A total of 297,761 reads were generated with the maximum of 841 bp and the minimum of 14 bp in length by using GS FLX Titanium platform. These reads were assembled to *L. longiflorum* chloroplast genome and there were 10,350 reads (3.47%) which were assembled, with an average length of 478 bp and 32× coverage of the chloroplast genome. The cpDNA of *V. patulum* (GenBank accession no. KF437397) is 153,699 bp in length and composed of two IR regions of 26,360 bp, which are separated by LSC and SSC regions of 83,372 bp and 17,607 bp, respectively. About 58.6% of the sequence is made up of protein-coding regions, whereas 41.4% contains non-coding sequences, including introns and intergenic spacers. The AT and GC contents are 62.3% and 37.7%, respectively (Table 1). There are 115 unique genes in the chloroplast genome of *V. patulum* (Fig. 1 and Table 2), of which 21 are duplicated in IR regions. Among the 115 genes, there are 4, 30 and 81 rRNA, tRNA and protein-coding genes, respectively. Similar to other chloroplast genomes, 19 genes of *V. patulum* cpDNA contain introns. Most of them contain one intron, except *clpP*, *ycf3* and *rps12*, which are separated by two introns. In the case of the *rps12* gene, the first exon is located in LSC, while the second and third exons are duplicated in IR regions. Both *ycf15* and *ycf68* in *V. patulum* contain many internal stop codons, indicating that these sequences represent pseudogenes. Another pseudogene, *infA*, is truncated by one stop codons within its sequence. We also found two open reading frames (ORFs), ORF42 and ORF56, both of which are located in the intron region of *trnA-UGC*.

3.2. Codon usage, simple sequence repeats (SSRs) and large repeats

All genes of *V. patulum* are encoded by 26,094 codons. Among these, leucine and cysteine are the most and the least frequent amino acids in the genome, with 2695 and 312 codons, respectively (Table 3). We also examined the start and stop codons in protein-coding genes (data not shown). The results showed that ATG is the most common start codon, although one GTG start codon was found in the *rps19* gene, and ACG start codons were detected in the *rpl2* and *ndhD* genes.

In addition, we found 104 SSR loci, which are repeated 3–15 times in the *V. patulum* cpDNA (Table 4). Among these microsatellites, we identified 59 homopolymers, 7 dipolymers and 35 tripolymers. Homopolymers and tripolymers contain A, T, G and C nucleotides, whereas dipolymers comprise only A and T nucleotides. Most SSR loci are located in intergenic spacer regions. However, some of them occurred in coding