Complete Chloroplast Genome Sequences from Korean Ginseng (*Panax schinseng* Nees) and Comparative Analysis of Sequence Evolution among 17 Vascular Plants

Ki-Joong Kim* and Hae-Lim Lee

School of Life Sciences and Biotechnology, Korea University, Seoul 136-701, Korea

(Received 31 May 2004; revised 14 June 2004)

Abstract

The nucleotide sequence of Korean ginseng (*Panax schinseng* Nees) chloroplast genome has been completed (AY582139). The circular double-stranded DNA, which consists of 156,318 bp, contains a pair of inverted repeat regions (IRa and IRb) with 26,071 bp each, which are separated by small and large single copy regions of 86,106 bp and 18,070 bp, respectively. The inverted repeat region is further extended into a large single copy region which includes the 5′ parts of the *rps19* gene. Four short inversions associated with short palindromic sequences that form stem-loop structures were also observed in the chloroplast genome of *P. schinseng* compared to that of *Nicotiana tabacum*. The genome content and the relative positions of 114 genes (75 peptide-encoding genes, 30 tRNA genes, 4 rRNA genes, and 5 conserved open reading frames *ycf8*), however, are identical with the chloroplast DNA of *N. tabacum*. Sixteen genes contain one intron while two genes have two introns. Of these introns, only one (*trnL-UAA*) belongs to the self-splicing group I; all remaining introns have the characteristics of six domains belonging to group II. Eighteen simple sequence repeats have been identified from the chloroplast genome of Korean ginseng. Several of these SSR loci show infra-specific variations. A detailed comparison of 17 known completed chloroplast genomes from the vascular plants allowed the identification of evolutionary modes of coding segments and intron sequences, as well as the evaluation of the phylogenetic utilities of chloroplast genes. Furthermore, through the detailed comparisons of several chloroplast genomes, evolutionary hotspots predominated by the inversion end points, indel mutation events, and high frequencies of base substitutions were identified. Large-sized indels were often associated with direct repeats at the end of the sequences facilitating intra-molecular recombination.

Key words: *Panax schinseng* Nees; chloroplast genome sequence; short inversion; intra-molecular recombination

1. Introduction

The complete nucleotide sequence of chloroplast genome from 17 vascular plant species has been established, which include: two ferns, *Psilotum nudum* (Psilotaceae, 138,829 bp, AP004638) and *Adiantum capillus-veneris* (Adiantaceae, 150,568 bp, NC004766); two gymnosperms, *Pinus thunbergii* (Pinaceae, 119,707 bp, NC001631) and *Pinus koraiensis* (Pinaceae, 116,866 bp, AY228468, unpublished); three monocots, *Oryza sativa* (Poaceae, 134,525 bp, NC001631), *Zea mays* (Poaceae, 140,384 bp, X86563), and *Triticum aestivum* (Poaceae, 134,545 bp, AB042240); and 10 dicots, *Amborella trichopoda* (Amborellaceae, 162,686 bp, NC005086), *Calycanthus floridus* (Calycanthaceae, 153,337 bp, NC004993), *Spinacia oleracea* (Chenopodiaceae, 150,725 bp, AJ400848), *Lotus japonicus* (Fabaceae, 150,519 bp, AP002983), *Medicago truncatula* (Fabaceae, 124,033 bp, AC093544, unpublished), *Oenothera elata* (Onagraceae, 163,935 bp, AJ271079), *Arabidopsis thaliana* (Brassicaceae, 154,478 bp, NC000932), *Atropa belladonna* (Solanaceae, 156,687 bp, AJ316582), *Nicotiana tabacum* (Solanaceae 155,939 bp, Z00045), *Epipogium virginiana* (Orobanchaceae, 70,028 bp, NC001568). In addition, the complete chloroplast genome sequences are also available from the three major nonvascular land plant lineages including a liverwort (*Marchantia polymorpha*, X04465), a moss (*Physcomitrella patens*, AP006572), and a hornwort (*Anthoceros formosae*, AB086179).

Available sequence data have revealed the relative conservative natures of chloroplast genomes with regard
to both structure and gene content. The presence of a large inverted repeat (IR), which ranges from 12 to 50 kb in length, is one of the conserved structural features and accounts for the length variation of the genomes. Two segments of IRs are separated by a large single-copy (LSC) and a small single-copy (SSC) region. The extreme contraction or loss of IR regions is only observed from Pinus and Medicago. The gene contents and the polycistronic transcription units of chloroplast genome are also largely conserved among most vascular plant species. The only exception is the non-photosynthetic parasitic plant, Epifagus virginiana, which lacks several genes related to photosynthesis. The gene order of the chloroplast genome is also relatively conserved. However, it is frequently reversed by inversion mutations that can be mediated by intra-molecular recombination events. The base substitution rate of chloroplast genes correlated to the position of the gene on the genome. For example, the genes in the IR regions diverge at a slower rate as compared to the genes located in the SSC and LSC. Most of the evolutionary hotspots that show high frequencies of indel mutations and base substitutions are concentrated on intergenic spacers that lack the polycistronic transcription units.

In this report, we present the complete sequence of the chloroplast genome from Korean ginseng (Panax schinseng Nees) and compared it with published sequences of vascular plants, with emphasis on the evolutionary modes of chloroplast genomes.

P. schinseng is one of the oldest and the most widely used herbal medicines for the oriental peoples. Recently, many different ginseng products are sold globally as alternative medicines for general health improvement. The primary active ingredients of P. schinseng are saponin triterpenoid glycosides, commonly called ginsenosides, that act on the central nervous system, cardiovascular system, endocrine system, and immune system. These active ingredients also increase endocrine secretion, promote immune function, and have anti-aging and stress-relieving effects. There are more than 30 ginsenosides reported from P. schinseng. In contrast to its high market demand, however, P. schinseng is almost extinct in wild habitats and is only found rarely in the mountains of Northeastern China, Korea, and the costal areas of Russian Far East. Fortunately, P. schinseng is widely cultivated not only in Korea but also in China, Japan, and several countries in North America and Europe under special shade conditions. P. schinseng is a member of the family Araliaceae, which also includes closely related Japanese ginseng (P. japonica), American ginseng (P. quinquefolius), Sanchi ginseng (P. notoginseng), Himalayan ginseng (P. pseudoginseng), and dwarf ginseng (P. trifoliatum).

The complete chloroplast genome data of P. schinseng will contribute to a better understanding of the photosynthetic mechanisms of shade plants that adapted under low light conditions and could also provide molecular phylogenetic information for commercially important ginseng species.

2. Materials and Methods

Seeds of wild P. schinseng were collected from a single plant in the central part of Korea (Mt. Sobek), and 11 individual plants were cultivated from the seeds for 3 years. Fifty grams of fresh leaves were harvested from 10 cultivated plants (voucher specimen kept at the Korea University Herbarium, Seoul, Korea, by Kim, K.-J., Yeo, J.-Y., and Kang, C.-W., May 25, 1997).

Chloroplasts of P. schinseng were isolated from the fresh leaves by the sucrose-gradient method. Chloroplast DNAs were isolated from the purified chloroplast by lysis and CsCl-EtBr gradient ultra-centrifugation, and further purified using a dialysis membrane (pore size 14,000 daltons). The chloroplast DNAs were digested with BamHI, Sac I, and Cla I restriction enzymes. The digested fragments were cloned into a pBluescript II vector. Vector-inserted cpDNA fragments were shotgun sequenced by the dyeoxy chain termination method (Big Dye 2.0™ Terminator Cycle Sequencing Kit, PE Applied Biosystems, Foster City, CA, USA) using an ABI 377 automatic sequencer. The cloned fragments from the three enzymes cover 98% of the whole genome after contig assemblages. The remaining three small gap regions were amplified by PCR and sequenced directly from the amplification products. The sequence fragments were assembled using Sequencher 4.1 (Gene Code Corporation, Ann Arbor, MI, USA). The same regions were sequenced 1–23 times (average of 4.7 times).

Seventeen complete chloroplast genome sequences were obtained from GenBank of the National Center for Biotechnology Information (NCBI). Gene annotations and comparative genome analyses were performed using the current versions of various BLAST (BLASTN, PHI-BLAST, BLASTX) and ORF finder programs from NCBI. Sequence alignments, base substitution analysis, and phylogenetic analysis were performed using the ClustalX, MEGA2, PAUP (Phylogenetic analysis using parsimony and other methods, by Swoford, D. L. 2002 Beta Version 4.0, Sinauer Asso. Inc., Sunderland, Massachusetts), and MacClade (version 4.0) programs. Pairwise sequence divergences were calculated by using Kimura’s two-parameter model. Repeating sequences were searched using REPutter. The locations and secondary structures of trn genes were evaluated using tRNAscan-SE (version 1.21) program. In addition, the secondary structures of rRNA, introns, and parts of DNA sequences were evaluated using MFOLD (version 3.0) program.
3. Results and Discussion

3.1. Gene content, arrangement, and codon usage of Panax schinseng chloroplast genome

The *P. schinseng* chloroplast genome has a pair of inverted repeat regions (IRa and IRb) consisting of 26,071 bp each. The two IR regions were separated by a large single copy (LSC) region of 86,106 bp and a small single copy (SSC) region of 18,070 bp. The total genome size is 156,318 bp in length (Acc. No. AY582139, Supplemental Fig. 1: http://www.dna-res.kazusa.or.jp/11/4/03/supplement/supplement.html). The positions of the 114 genes identified in the *P. schinseng* chloroplast genome are presented in Fig. 1 and Table 1. The major portion (58%) of the *P. schinseng* chloroplast genome consists of gene-coding regions (50% protein coding and 8% RNA regions), whereas, the intergenic spacers (including 27 introns) comprise 42% (intron, 13%; spacer, 29%). The overall A-T content of *P. schinseng* chloroplast genome is 62%. The A-T contents in the non-coding (65%) are higher than in the coding (59%) regions. The A-T contents of the IR region amounts to 57%, whereas, the A-T contents in the LSC and SSC regions are 64% and 68%, respectively. The low A-T contents of the IR region reflect the low A-T contents in the four rRNA (45%) genes in this region.

The gene contents of *P. schinseng* chloroplast genome
Table 1. Genes contained in *Panax* chloroplast genome (total 114 genes).

| Category for genes | Group of gene | Name of gene |
|--------------------|---------------|--------------|
| Self replication   | tRNA genes    | 16S (rrn16)(x2), 23S (rrn23)(x2), 4.5S (rrn4.5)(x2), 5S (rrn5)(x2) |
|                    | tRNA genes    | trnA-UGC* (x2), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnG-UGC*, trnH-GUG, trnI-CAU(x2), trnL-GAU* (x2), trnK-UUU*, trnL-CAAx(x2), trnL-UAA*, trnL-UAG trnM-CAU, trnM-CAU, trnN-GUUX(x2), trnP-UUG, trnQ-UGU, trnR-ACG(x2), trnR-UCA, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC(x2), trnV-UCAC*, trnW-CCA, trnY-GUA |
| Small subunit of ribosome | rps2, rps3, rps4, rps7(x2), rps8, rps11, rps12*(x2, part), rps14, rps15, rps16*, rps18, rps19(x2, part) |
| Large subunit of ribosome | rpl2*(x2), rpl14, rpl16*, rpl20, rpl22, rpl23(x2), rpl32, rpl33, rpl36 |
| DNA dependent RNA polymerase | rpoA, rpoB, rpoC1*, rpoC2 |
| Genes for photosynthesis | Subunits of NADH-dehydrogenase | ndhA*, ndhB* (x2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhK |
|  | Subunits of photosystem I | psaA, psaB, psaC, psal, psaI, ycf3** |
|  | Subunits of photosystem II | psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbl, psbI, psbK, psbl, psbm, psbn, psBT, psbZ |
|  | Subunits of cytochrome b/f complex | petA, petB*, petD*, petG, petL, petN |
|  | Subunits of ATP synthase | atpB, atpE, atpF*, atpH, atpI |
|  | Large subunit of rubisco | rbcL |
| Other genes | Translational initiation factor | infA |
|  | Maturase | matK |
|  | Protease | clpP** |
|  | Envelop membrane protein | cemA |
|  | Subunit of Acetyl-CoA-carboxylase c-type cytochrome synthesis gene | accD |
| Genes of Unknown function | Conserved (ORF, ycf) | ycf1(x2, part), ycf2(x2), ycf4, ycf15(x2) |

One and two asterisks indicate one- and two-intron containing genes, respectively. Genes located on the IR region indicate by the (x2) symbol after gene name.

are almost identical to that of tobacco,\textsuperscript{14} except for a few minor modifications in open reading frames (ORF or ycf) and IR/LSC and SSC junction regions (see IR contraction/expansion section).

Eighteen genes contain one or two introns. Table 2 summarizes the sizes of exons and introns for each gene. Five of these introns, *rps12*, *ndhB*, *rps12*, *trnI-CAU* and *trnA-UGC*, are located within the IR regions. Only the *trnL-UAA* gene intron belongs to the self-splicing group I, while all others belong to group II. Three genes, *clpP*, *rpoC1*, and *rps12*, have two introns. The *rps12* gene is the unique divided gene that the 5’ end exon locates in the large single copy region far away from its second and third exons, which are located as duplicates in the inverted repeat regions and which requires a trans-splicing mechanism between exon I and exon II in order to produce mature *rps12* transcripts.\textsuperscript{37}

A total of 30 tRNA genes were identified from the *P. schinseng* chloroplast genome. The numbers and kinds of tRNA genes from *P. schinseng* are identical to well-characterized vascular plant chloroplasts.\textsuperscript{5,10,14,20} The codon usage of the *P. schinseng* chloroplast genome and the anticodons present in the 30 tRNA species are summarized in Table 3. The codon usage has been deduced from all protein-coding genes and *ycf* genes as presented in Table 1. The high A-T content at the third codon position clearly reflects the codon usage bias to A or T. For example, twofold degenerated codons at the third position show the usage frequencies of 21–34% for C or G, while the usage frequencies range from 66% to 79% for A or T. Fourfold degenerated codons at the third position such as Gly, Ala, Thr, Pro, Ile and Val, also show a strong bias to A or T ranging from 66% to 79%. For the sixfold degenerated codons such as Leu, Ser, and Arg, the A-T bias is observed for both the first and the third codon positions. The usage frequencies of stop codons are also biased to A or T at both the second and third positions, showing 49% for TAA, 26% for TGA, and 21% for TAG. The same results are well documented in the chloroplast genomes of several other higher plants includ-
Table 2. The lengths of introns and exons for the splitting genes on *Panax* chloroplast genome.

| Gene     | Exon I | Intron I | Exon II | Intron II | Exon III |
|----------|--------|----------|---------|-----------|----------|
| *trnA-UGC* | 38     | 808      |         | 35        |          |
| *trnG-UCC* | 23     | 697      | 48      |           |          |
| *trnI-GAU* | 37     | 945      | 35      |           |          |
| *trnK-UUU* | 37     | 2524     | 35      |           |          |
| *trnL-UAA* | 35     | 507      | 49      |           |          |
| *trnV-UAC* | 39     | 578      | 35      |           |          |
| *rps12*  | 112    | -        | 232     | 536       | 26       |
| *rps16*  | 40     | 887      | 197     |           |          |
| *rpl2*   | 391    | 660      | 434     |           |          |
| *rpl16*  | 9      | 944      | 399     |           |          |
| *atpF*   | 145    | 730      | 410     |           |          |
| *petB*   | 6      | 783      | 642     |           |          |
| *petD*   | 8      | 751      | 475     |           |          |
| *ndhA*   | 553    | 1023     | 539     |           |          |
| *ndhB*   | 777    | 678      | 756     |           |          |
| *ycf3*   | 124    | 716      | 230     | 758       | 153      |
| *clpP*   | 71     | 771      | 292     | 632       | 228      |
| *rpoC1*  | 453    | 756      | 1617    |           |          |

Five genes with an asterisk locate on the IR regions. The *rps12* gene is divided. The 3′-*rps12* locates on the IR-region, while the 5′-*rps12* locate on the LSC region.

Table 3. Codon usage of the *Panax* chloroplast genome.

| Codon | Phe  | Ser  | Tyr  | Cys  | gcA |
|-------|------|------|------|------|-----|
| TTT   | Phe  | gaa  | 927  |      |     |
| TTC   | Phe  | gaa  | 535  |      |     |
| TTA   | Leu  | uaa  | 837  |      |     |
| TTG   | Leu  | caa  | 580  |      |     |
| CTT   | Leu  | uag  | 585  |      |     |
| CTC   | Leu  | uag  | 187  |      |     |
| CTG   | Leu  | uag  | 386  |      |     |
| ATT   | Ile  | gau  | 1,050|      |     |
| ATC   | Ile  | gau  | 471  |      |     |
| ATA   | Ile  | cau  | 696  |      |     |
| ATG   | Met  | cau  | 605  |      |     |
| GTT   | Val  | gac  | 513  |      |     |
| GTC   | Val  | gac  | 177  |      |     |
| GTA   | Val  | uac  | 511  |      |     |
| GTG   | Val  | uac  | 213  |      |     |

The codon usage has been determined from all genes listed in Table 1. Codons are given in capital letters and anticodons are in lower case letters. Asterisks indicate stop anticodons. A total numbers of each codon representing the same amino acid are provided in this table.

The strong AT bias of codon usages are universal for plant chloroplast genomes.

3.2. Short inversions and inverted sequences

Four short inversions, associated with inverted sequences, were identified in the *Panax* chloroplast genome as compared to the *Nicotiana* chloroplast genome...
These regions form distinct stem-loop hairpin structures and the orientation of the sequences at the loop regions are reversed between two chloroplast genomes. The recombination events between inverted sequences on the stem regions are responsible for the short inversions of loop sequences. The first short inversion (Fig. 2-1) is located one base downstream (genome coordinate 54,452 bp) of the trnM-CAU gene that also corresponds to 96 bp downstream (genome coordinate 54,490 bp) of the atpE gene. Both the second and third inversions (Figs. 2-2, 2-3) are located at the spacer region between the petA and psbJ genes, more closely downstream of petA and psbJ genes, respectively. The fourth inversion (Fig. 2-4) is also located closely downstream of trnL-UAA gene, which corresponds to the spacer region between ycf15 and trnL-UAA. Thus, all four regions correspond to the stem-loop forming regions downstream of the genes that function to stabilize the
mRNA molecules. Large inversion mutations have been reported frequently in several widely separated vascular plants. In contrast, the short inversions reported on the *P. schinseng* chloroplast genome are rarely reported in other plant groups. However, more cases will probably become apparent if the chloroplast genomes of other completely sequenced species are thoroughly investigated.

### 3.3. IR contraction and expansion

The border between the two IR/LSC and the two IR/SSC usually differs among various species. Large expansions and contractions of IR regions often create the large length variation of chloroplast genomes in different groups of plants. Detailed comparisons of IR boundaries among 16 vascular species suggest that wide ranges of expansions and contractions of IR are very common evolutionary events. As a result, various pseudogenes are created in genes such as *rps19* and *ycf1* at the boundaries between single-copy (SC) regions and IR regions. In Fig. 3, the detailed IR/SC border positions, with respect to the adjacent genes among *Panax*, *Nicotiana*, *Atropa* and *Arabidopsis* chloroplast genomes, are compared. This comparison shows that the border positions vary among chloroplast genomes even between closely related genera of the same family such as *Nicotiana* and *Atropa*. The IRA/SSC borders are located in the 3’ region of the *ycf1* gene and create the *ycf1* pseudogenes at the IRb/SSC border with lengths of 996 bp and 1,438 bp in *Nicotiana* and *Atropa*, respectively.

In *Panax*, IR is further extended deep into the *ycf1* gene and inserted into the *ycf1* pseudogene with 1,649-bp lengths at the IRb/SSC border. The IRA/SSC and IRb/SSC borders in *Arabidopsis* are located within the coding regions of the *ycf1* and *ndhF* genes, respectively. Thus, a portion (37 bp) of the 3’ end of the *ndhF* gene overlaps with the internal portion of *ycf1* in *Arabidopsis*. In contrast to the IR/SSC borders, only minor shifts were observed in the IR/LSC borders. For example, the IRA/LSC and IRb/LSC borders of *Nicotiana* are located downstream of the non-coding region of *trnH-GUG* and upstream of the non-coding region of *rps19*. Thus, no pseudogene is created at the border. However, the IR/LSC borders are extended into the 5’ portion of *rps19* gene in *Atropa*, *Panax*, and *Arabidopsis*. As a result, the various lengths of *rps19* pseudogenes are located on the IRA/LSC border in *Panax* (51 bp), *Atropa* (59 bp), and *Arabidopsis* (113 bp). The expansions/contractions of IR as observed in the IR/SSC borders are probably mediated by intra-molecular recombination between two short direct repeat sequences that frequently occur within the genes located at the borders.
3.4. Sequence divergence of major genes or regions

Completed chloroplast genome sequences are available from 18 vascular plants including two ferns, two gymnosperms, three monocots, and 10 dicots. We compared the gene contents and the levels of average pairwise nucleotide and amino acid divergence of 84 chloroplast genes among 16 vascular plant species, and their results are summarized in Table 4. The chloroplast genome sequences from Epifagus virginiana was not considered in this report because most chloroplast genes are lost in this non-green parasitic plant.\(^\text{15}\) Sequences from Pinus koraiensis was also not considered because the sequences are almost identical to that of P. thunbergii. Also, information on trn gene class was excluded in this summary table. The abnormal gene annotations including abnormal start and stop positions were adjusted after multiple sequence alignments during the comparative sequence analysis. Multiple internal stop codons or frame shift mutations were frequently encountered in Medicago (AC093544), Adiantum,\(^\text{2}\) and Oenothera\(^\text{11}\) sequences, suggesting that these problems are largely due to RNA editing or sequence errors.

Low levels (less than 8%) of average sequence divergence (ASD) among vascular plants were observed from four ribosomal RNA-coding genes (rrn16, rrn23, rrn5, and rrn4.5), which were all located in the IR regions (Table 4). In contrast, the relatively high divergence (more than 35%) were observed from rps15, rpl22, accD, infA, matK, ycf2, and ycf1 genes located in the LSC, IR, or SSC/IR junction regions. The most conserved rrn16 gene (ASD=3.3%) in the IR region is 16 times more conserved than the fastest evolving ycf1 gene (ASD=54.4%) in the LSC/IR junction region. These results support previous reports that the sequence contained in the IR regions diverge at slower rates compared to sequences located in the LSC or SSC regions.\(^\text{5,23}\) Some genes in the IR region, however, have faster divergence than those genes from the LSC or SSC regions. For example, rps7, rpl2, ycf15, rpl23, ndhB, and ycf2 in the IR regions show more variation than do several genes in the LSC or SSC regions. Furthermore, ycf2 in the IR regions is one of the fastest evolving genes from the chloroplast genomes. In contrast, the majority of the genes involved in photosystem I (psa gene class), photosystem II (psb gene class), and photoelectron transport system (pet gene class) between two photosystems located in the LSC or SSC regions show relatively slow evolutionary divergence, ranging from 9.8% to 15.0%. This data support the functional importance of the thylakoid membrane proteins in chloroplasts.\(^\text{24,26}\) However, the rps or rpl genes in the IR region clearly show slower evolutionary rates than the same rps or rpl gene classes in the SSC or LSC regions. Therefore, the divergence levels of chloroplast genes are influenced not only by the location of the genes on the chloroplast genome but also by the functional constraints of the genes. Similar divergence patterns were also observed at the protein level (Table 4).

The gene contents are relatively conserved among vascular plants except for the following examples. The accD is lost in the members of Poaceae (Zea and Triticum) or exists as a small pseudogene fragment (Oryza, ORF106).\(^\text{4,6}\) In addition, the accD gene is also lost independently in Medicago (Fabaceae) chloroplast genome and this gene shows wide ranges (from 933 in Adiantum to 1,623 bp in Amborella) of length variation when present. The short sequences of accD are primarily due to the truncation of the 5′ portion of the genes, such as that in Adiantum (933 bp), Pinus (966 bp), and Psilotum (933 bp).\(^\text{1−3}\) Three ribosomal proteins are lost from chloroplast genomes of some plant species. The first, rpl22, is lost among members of the Fabaceae (Lotus and Medicago) chloroplast genome,\(^\text{10}\) and this gene also shows a wide range of length variation when present, ranging from 351 bp in Psilotum to 600 pb in Spinacia. Most length variations occur at the beginning and the end of the sequences. The second gene, rpl32, is lost in the Oryza chloroplast genome.\(^\text{4}\) The third gene, rps16, is lost independently in the chloroplast genome of Medicago, Pinus, and Psilotum.\(^\text{1,3}\) The genes for the chloroplast-encoded NADH dehydrogenase protein components are also lost in some plant groups. Notably, all ndh gene components are lost in the Pinus (Gymnosperm) chloroplast genome.\(^\text{3}\) In addition, the psaM gene is only present in the Psilotum and Pinus chloroplast genomes.\(^\text{1,3}\)

Two recently identified genes, ycf3\(^\text{39}\) and ycf9 (psbZ),\(^\text{40}\) in the LSC region are present in all vascular plants and also have relatively low levels of variation (15.2% and 15.7%, respectively). In contrast, the ycf15, ycf1, and ycf2 genes are lost several times during the evolution of vascular plants (Table 4). The gene lengths and the sequence divergence of the ycf1 and ycf2 genes vary among plant species. The length variation of ycf1 is largely due to the large indels in the middle of the gene and to the extensions/contractions of the IR into the SSC regions (see detailed discussion in IR extension/contraction sections). The length variation of ycf2 is largely responsible for the internal indel mutations associated with direct repeat sequences. The infA gene is absent in the Arabidopsis, Lotus, and Medicago chloroplast genomes and is present as a truncated pseudogene in several chloroplast genomes. The rpcC2 gene also has a wide length variation mainly due to the internal duplicated insertion in Poaceae.\(^\text{4−6,41,42}\)

The conserved genes in the sequence also have generally less variation in length. For example, several genes related to photosystem I, photosystem II, and photoelectron transport systems show no length variation among widely separated vascular plants. In contrast, the functionally not well-defined ycf15, ycf4, cemA(ycf10), and accD(ycf11) genes show wide ranges of length variation. Except for the above few variable genes, the length vari-

---

\(\text{Vol. } 11,\)
| Order | Gene Name | Region | DNA (D) | DNA (SE) | Protein (D) | Protein (SE) | Average length | Range of length | Missing taxa | Functional groups of genes |
|-------|-----------|--------|---------|----------|-------------|--------------|----------------|----------------|--------------|---------------------------|
| 1     | rrn16     | IR     | 0.03342 | 0.00241  | 1493        | 1490-1504    | Ribosomal RNA  |
| 2     | rrn23     | IR     | 0.04196 | 0.00208  | 2824        | 2796-2888    | Ribosomal RNA  |
| 3     | rrn5      | IR     | 0.04716 | 0.01020  | 121         | 121-123      | Ribosomal RNA  |
| 4     | rrn4.5    | IR     | 0.07489 | 0.01654  | 102         | 95-104       | Ribosomal RNA  |
| 5     | psbA      | LSC    | 0.09771 | 0.00502  | 1055        | 954-1062     | Photosystem II |
| 6     | psbD      | LSC    | 0.11154 | 0.00592  | 1062        | 1062         | Photosystem II |
| 7     | rps12     | IR/LSC | 0.11189 | 0.01010  | 375         | 372-408      | Small ribosomal subunit |
| 8     | psbE      | LSC    | 0.11359 | 0.01198  | 252         | 252          | Photosystem II |
| 9     | psbL      | LSC    | 0.11825 | 0.01954  | 117         | 117          | Photosystem II |
| 10    | petB      | LSC    | 0.12389 | 0.00720  | 648         | 648          | Cytochrom b/f complex |
| 11    | petN      | LSC    | 0.12641 | 0.02443  | 90          | 90-93        | Cytochrom b/f complex |
| 12    | psbF      | LSC    | 0.12939 | 0.01894  | 1300        | 375          | Small ribosomal subunit |
| 13    | atpH      | LSC    | 0.12999 | 0.01453  | 120         | 120          | Photosystem II |
| 14    | psaB      | LSC    | 0.13068 | 0.00401  | 2205        | 2205-2208    | Photosystem I |
| 15    | psaA      | LSC    | 0.13116 | 0.00432  | 2258        | 2253-2325    | Photosystem I |
| 16    | rps7      | IR     | 0.13258 | 0.00942  | 467         | 443-474      | Small ribosomal subunit |
| 17    | petG      | LSC    | 0.13662 | 0.01091  | 487         | 483-528      | Cytochrom b/f complex |
| 18    | psbN      | LSC    | 0.14010 | 0.01425  | 132         | 132          | Photosystem II |
| 19    | ndhB      | IR     | 0.14049 | 0.00538  | 1521        | 1479-1533    | Pinus NADH-dehydrogenase |
| 20    | psbB      | LSC    | 0.14153 | 0.00683  | 1527        | 1527         | Photosystem II |
| 21    | psbI      | LSC    | 0.14644 | 0.02255  | 114         | 111-159      | Photosystem II |
| 22    | rbcL      | LSC    | 0.14737 | 0.00521  | 1428        | 1428-1431    | Large subunit of rubisco |
| 23    | ycf3      | LSC    | 0.15258 | 0.01054  | 509         | 477-552      | Photosystem I |
| 24    | psbZ      | LSC    | 0.15735 | 0.01705  | 189         | 189          | Photosystem II |
| 25    | atpB      | LSC    | 0.15871 | 0.00768  | 1494        | 1479-1503    | ATP synthase |
| 26    | psbJ      | LSC    | 0.16370 | 0.02443  | 123         | 123          | Photosystem II |
| 27    | rpl36     | LSC    | 0.16823 | 0.02564  | 114         | 114          | Large ribosomal subunit |
| 28    | atpA      | LSC    | 0.17323 | 0.00682  | 1508        | 1296-1536    | ATP synthase |
| 29    | apl1      | LSC    | 0.17469 | 0.00909  | 745         | 738-750      | ATP synthase |
| 30    | ndhC      | LSC    | 0.17715 | 0.01411  | 363         | 363          | Pinus NADH-dehydrogenase |
| 31    | psbT      | LSC    | 0.17840 | 0.02855  | 106         | 99-117       | Photosystem II |
| 32    | ndhH      | SSC    | 0.17846 | 0.00766  | 1182        | 1182-1185    | Pinus NADH-dehydrogenase |
| 33    | rpl2      | IR     | 0.18197 | 0.00842  | 822         | 792-834      | Large ribosomal subunit |
| 34    | petA      | LSC    | 0.18434 | 0.00893  | 963         | 957-966      | Cytochrom b/f complex |
| 35    | psbM      | LSC    | 0.18791 | 0.02693  | 107         | 105-114      | Photosystem II |
| 36    | ndhJ      | SSC    | 0.19092 | 0.01189  | 515         | 468-543      | Pinus NADH-dehydrogenase |
| 37    | rpl16     | LSC    | 0.19238 | 0.01391  | 408         | 405-414      | Large ribosomal subunit |
| 38    | ndhJ      | LSC    | 0.19317 | 0.01116  | 478         | 477-480      | Pinus NADH-dehydrogenase |
| 39    | ycf15     | IR     | 0.19652 | 0.02404  | 215         | 147-303      | Psilotum, Adiantum, Pinus, Oryza, Triticum, Zea, Lotus, Medicago, Arabidopsis |
| 40    | ndhK      | LSC    | 0.20777 | 0.01018  | 786         | 624-858      | Pinus NADH-dehydrogenase |
| 41    | rpl14     | LSC    | 0.20959 | 0.01557  | 370         | 366-372      | Large ribosomal subunit |
| 42    | psbH      | LSC    | 0.21464 | 0.02027  | 224         | 219-240      | Photosystem II |

D and SE indicate average percentage distances and standard errors, respectively, from 16 × 16/2 pairwise comparisons. DNA distances are corrected by Kimura’s 2-parameter model and protein distances are represented as a proportion of amino acid sites. All pairwise distances are calculated under the pairwise deletion of sequence alignments.
ations of most genes are usually limited and are confined at the 3' end of genes. If we consider sequence divergence levels together with the length variations of genes, the 16S, 23S, ndhB, psbA, psbD, psaA, psaC, psbB, and rbcL genes are probably good candidate genes for phylogenetic study of higher plants. In contrast, ycf2, accD, matK, rpoC2, and ndhF genes (if present) are good candidate genes for phylogenetic study among closely related families or infra-family levels of vascular plants.

Nine of 20 introns are lost in more than one taxon (Table 5). Among them, intron losses in ndhA, trnK-
UUU, and \textit{rps16} are the result of corresponding gene losses. Therefore, actual intron losses without gene losses are limited to 3′-\textit{rps12}, \textit{rpl2}, \textit{rpoC1}, and \textit{clpP}. The loss of the \textit{rpl2} intron is unique and is only seen in chloroplast genomes of \textit{Spinacia}.\textsuperscript{8} In addition, \textit{rpoC1} gene intron loss occurs at once and is confined to all members of Poaceae (\textit{Zea}, \textit{Oryza}, and \textit{Triticum}).\textsuperscript{4–6,41} In contrast, the 3′-\textit{rps12} intron on the IR region is independently lost on \textit{Medicago}, \textit{Adiantum}, and \textit{Psilotum} lineages.\textsuperscript{1,2} Both introns 1 and 2 in \textit{clpP} gene are lost in the chloroplast genomes of \textit{Pinus}, \textit{Oenothera}, and all members of Poaceae.\textsuperscript{3,4–6,11} However, only intron 1 is absent in the \textit{clpP} gene of \textit{Medicago}.

The sizes of introns are relatively conserved, mostly ranging from 500 to 1,300 bp. The only long intron is the \textit{trnK}-UUU intron (2,410–2,627 bp) that contains coding sequences of the mat\textit{K} gene within the intron. Each intron also shows relatively narrow ranges of variation in length. Small indels within introns are concentrated in specified regions of intron. Average sequence divergence of introns among 16 chloroplast genomes was low in the IR region and one to nine times slower than the introns in the LSC or SSC regions (Table 5). The five most conserved introns are all located within the IR region.

The sequence divergence of introns is neither closely related to their coding segments nor to the gene functions. For example, the relatively conserved \textit{petB} (SD=0.12389) gene contains a variable intron region (SD=0.45697) while the more variable \textit{rps16} (SD=0.21793) gene contains more conserved introns (SD=0.34797), both located in the LSC region. The 3′-\textit{rps12} intron sequences located in the IR region show a more conserved (5.9% differences) nature than their coding segments (11.2% differences), suggesting that the trans-splicing mechanism requires conserved intron sequence for this unique divided gene in chloroplast genomes.\textsuperscript{35} The introns on LSC and SSC show limited ranges of variation from 0.26073 to 0.50508 (1.9-fold differences), while the coding segments show up to 5.5-fold differences. Even if the 3′-\textit{rps12} intron was not considered, the intron regions change only 1.2–3.7 times faster than their coding segments. Some intron sequences in the IR regions are more conserved than the coding sequences of several genes in the LSC or SSC regions.

### 3.5. Intron divergence

The comparative intron sequence data clearly support the suggestion that the sequences in the IR regions diverged at slower rates compared to the sequences of the LSC and SSC regions. This is consistent with the idea that the stabilizing effect of the IR regions by genetic recombination is the main cause of their sequence conservation. Based on the sequence divergence and length variation levels, both \textit{ndhA} and \textit{rpl16} intron sequences are good candidate genes for phylogenetic studies within infra family levels. Meanwhile, \textit{clpP} gene introns may be useful for phylogenetic studies of closely related genera or infra generic levels.

### Table 5. Pairwise distances of nucleotide substitutions of 20 chloroplast introns among 16 fully sequenced chloroplast genomes of vascular plants.

| Order | Intron Name | Region | Average distances | Standard errors | Average lengths | Ranges of length | Missing taxa |
|-------|-------------|--------|-------------------|-----------------|----------------|-----------------|-------------|
| 1     | 3′-\textit{rps12} | IR     | 0.05922           | 0.00530         | 537            | 530-546         | \textit{Psilotum}, \textit{Adiantum}, \textit{Medicago} |
| 2     | \textit{trnA} | IR     | 0.11347           | 0.00594         | 788            | 681-819         |             |
| 3     | \textit{trnl} | IR     | 0.13844           | 0.00153         | 855            | 689-996         |             |
| 4     | \textit{ndhB} | IR     | 0.17663           | 0.00893         | 703            | 674-827         | \textit{Pinus} |
| 5     | \textit{rpl2} | IR     | 0.22500           | 0.01144         | 670            | 614-743         | \textit{Spinacia} |
| 6     | \textit{trnV} | LSC    | 0.26073           | 0.01286         | 586            | 541-632         |             |
| 7     | \textit{ycf3(II)} | LSC | 0.29742           | 0.01287         | 738            | 623-848         |             |
| 8     | \textit{trnl} | LSC    | 0.29864           | 0.01672         | 475            | 304-588         |             |
| 9     | \textit{ycf3(I)} | LSC | 0.30011           | 0.01258         | 729            | 677-778         |             |
| 10    | \textit{trnG} | LSC    | 0.33748           | 0.01458         | 731            | 677-937         |             |
| 11    | \textit{rps16} | LSC    | 0.34797           | 0.01454         | 820            | 423-891         | \textit{Psilotum}, \textit{Pinus}, \textit{Medicago} |
| 12    | \textit{rpoC1} | LSC    | 0.37017           | 0.01466         | 739            | 674-791         | \textit{Oryza}, \textit{Triticum}, \textit{Zea} |
| 13    | \textit{trnK} | LSC    | 0.38157           | 0.00775         | 2,509          | 2,410-2,627      | \textit{Adiantum} |
| 14    | \textit{atpF} | LSC    | 0.39065           | 0.01570         | 772            | 630-1,270       |             |
| 15    | \textit{ndhA} | SSC    | 0.40003           | 0.01252         | 1,068          | 804-1,274       | \textit{Pinus} |
| 16    | \textit{petD} | LSC    | 0.43026           | 0.01779         | 719            | 613-756         |             |
| 17    | \textit{rpl16} | LSC    | 0.44196           | 0.01533         | 977            | 678-1,110       |             |
| 18    | \textit{petB} | LSC    | 0.45697           | 0.01675         | 768            | 705-824         |             |
| 19    | \textit{clpP(II)} | LSC | 0.47848           | 0.01820         | 615            | 490-746         | \textit{Pinus}, \textit{Oryza}, \textit{Triticum}, \textit{Zea}, \textit{Oenothera} |
| 20    | \textit{clpP(I)} | LSC | 0.50508           | 0.02248         | 791            | 676-867         | \textit{Pinus}, \textit{Oryza}, \textit{Triticum}, \textit{Zea}, \textit{Medicago}, \textit{Oenothera} |

The distances are corrected by Kimura’s 2-parameter substitution models. The (I) and (II) symbol after intron name indicate intron I and intron II, respectively.
### 3.6. Mode of indel mutations and direct repeats

Several evolutionary hot spots in terms of indel mutations were identified in the chloroplast genomes.\(^5,23\) The indels can be classified into two groups. The first indel type is small indels and associated with short base repeating segments within short distance. It may have originated from a slipped-strand mispairing mechanism of surrounding sequences.\(^43\) Numerous small indels in non-coding intergenic and intron regions may correspond to this type. However, it is difficult to figure out the slipped-strand mispairing patterns from the non-coding intergenic spacer regions because numerous base substitution patterns and indels are often complicated in these fast-changing regions among remote related chloroplast genomes. However, slipped-strand mispairing patterns are relatively well preserved within several intron regions such as trnA-UGC and trnI-GAU introns (data not shown). Numerous short indels associated with one to eight base repeating units that duplicate less than five times more frequently occur in these introns.

The other distinctive indel type is relatively long in length and may have originated from illegitimate recombination events.\(^44-46\) This type of indel is usually associated with short direct repeats at remote distance and is often observed with direct repeating segments between short direct repeats. For example, the 228-bp unique insertion mutation in the *ycf1* gene of *Panax* is a four-time duplicated insertion of a 57-bp repeating unit that associates with direct repeat of six base (AGAAAC) at both ends (Fig. 4). The same 57-bp unit occurs in *Atropa* and *Nicotiana*. However, there is no duplicated pattern because the two terminal sequences, AGAACG and AGAAAC, are not identical. These data, therefore, suggest that the AGAACG to AGAAAC substitution at one end of the *Panax ycf1* gene generates the direct repeat at both ends and may trigger the duplicated insertions by illegitimate recombination. Another large deletion mutation in *Panax*, a 297-bp deletion in the *ycf2* gene, is also associated with direct repeats of GGATTCTAG (Fig. 5).

Similar 297-bp insertion mutations are also observed between the *petN* and *psbM* gene intergenic spacer in *Atropa* and *Nicotiana* (both are members of Solanaceae) (data not shown). Direct repeats of TCTTTGG are associated with this insertion mutation. Other 230-bp deletion mutations within the *trnI-GAU* intron are associated with the direct repeat of TCTTTAG in *Medicago*, *Nicotiana*, and *Atropa* or the direct repeat of TACATG in *Arabidopsis* and *Lotus*. These data support the idea that intra-molecular recombination mediated by these short direct repeat sequences are responsible for indel mutations that comprise up to several hundred base-pairs of chloroplast genome.

---

| Taxon\Position | 750 | 780 | 810 |
|---------------|-----|-----|-----|
| Nicotiana     |     |     |     |
| Atropa        |     |     |     |
| Arabidopsis   |     |     |     |
| Panax         |     |     |     |

Copy I

| Taxon\Position | 840 | 870 | 900 |
|---------------|-----|-----|-----|
| Nicotiana     |     |     |     |
| Atropa        |     |     |     |
| Arabidopsis   |     |     |     |
| Panax         |     |     |     |

Copy II

| Taxon\Position | 930 | 960 | 990 |
|---------------|-----|-----|-----|
| Nicotiana     |     |     |     |
| Atropa        |     |     |     |
| Arabidopsis   |     |     |     |
| Panax         |     |     |     |

Copy III

| Taxon\Position | 1020 | 1050 | 1078 |
|---------------|------|------|------|
| Nicotiana     |     |     |     |
| Atropa        |     |     |     |
| Arabidopsis   |     |     |     |
| Panax         |     |     |     |

Copy IV

**Figure 4.** Comparison of the nucleotide sequences in the middle of *ycf1* gene among four chloroplast genomes. The number indicates the coordination point of *ycf1* gene from the start codon. The 228-bp insertion, which duplicates four times of 57-bp repeating units, is limited in *Panax* and is associated with the six base (AGAAAC) direct repeats at both ends.
3.7. Simple sequence repeats

Chloroplast simple sequence repeats (SSR), which repeat the single nucleotide bases of As or Ts more than 10 times, were reported from *Pinus radiata* and *Oryza sativa*. In this study, we identified 18 SSR loci from *P. chinseng* chloroplast genome (Table 6). Fourteen of the 18 SSR loci occur in the intergenic spacers and were composed of multiple As or Ts. Only four of 18 SSR loci are multiple Gs in the *Nicotiana* chloroplast genome. Only five of the SSR loci occur on the same position between *Panax* and *Nicotiana* chloroplast genomes and can be aligned between two sequences. Some of the *Panax* SSR loci have systematic utilities in identifying the species or cultivars of this medically important plant species (Kim and Lee, unpublished data).

This research was supported by a grant (PF001102-02) from the Plant Diversity Re-
search Center of 21st Century Frontier Research Program, Ministry of Science and Technology of the Korean government and from the Plant Signaling Network Research Center, Korea Science and Engineering Foundation. We thank to Cristina Golia for her critical reading of the manuscript.

References

1. Wakasugi, T., Nishikawa, A., Yamada, K., et al. 1998, Complete nucleotide sequence of the plastid genome from a fern, *Psilotum nudum*, *Endocyt. Cell Res.*, 13 (Suppl.), 147.

2. Wolf, P. G., Rowe, C. A., Sinclair, R. B., and Hasebe, M. 2003, Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L., *DNA Res.*, 10, 59–65.

3. Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., and Sugiura, M. 1994, Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*, *Proc. Natl. Acad. Sci. USA*, 91, 9794–9798.

4. Hiratsuka, J., Shimada, H., Whittier, R., 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. Gen. Genet.*, 217, 185–194.

5. Maier, R. M., Neckermann, K., Igloi, G. L., and Kossel, H. 1995, Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing, *J. Mol. Biol.*, 251, 614–628.

6. Oghiara, Y., Isono, K., Kojima, T., et al. 2002, Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA, *Mol. Genet. Genomics*, 266, 740–746.

7. Goremykin, V. V., Hirsch-Ernst, K. I., Wölf, S., and Hellwig, F. H. 2003, Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm, *Mol. Biol. Evol.*, 20, 1499–1505.

8. Goremykin, V. V., Hirsch-Ernst, K. I., Wölf, S., and Hellwig, F. H. 2003, The chloroplast genome of the basal angiosperm *Calycanthus fertilis* — structural and phylogenetic analyses. *Plant Syst. Evol.*, 242, 119–135.

9. Schmitz-Linnneweber, C., Maier, R. M., Alcaraz, J. P., Cottet, A., Herrmann, R. G., and Mache, R. 2001, The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization, *Plant Mol. Biol.*, 45, 307–315.

10. Kato, T., Kaneko, T., Sato, S., Nakamura, Y., and Tabata, S. 2000, Complete structure of the chloroplast genome of a legume, *Lotus japonicus*, *DNA Res.*, 7, 323–330.

11. Hupfer, H., Swiatek, M., Hornung, S., et al. 2000, Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable *euoenothea* plastomestes, *Mol. Gen. Genet.*, 263, 581–585.

12. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. 1999, Complete structure of the chloroplast genome of *Arabidopsis thaliana*, *DNA Res.*, 6, 283–290.

13. Schmitz-Linnneweber, C., Regel, R., Du, T. G., Hupfer, H., Herrmann, R. G., and Maier, R. M. 2002, The Plastid Chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation, *Mol. Biol. Evol.*, 19, 1602–1612.

14. Shinozaki, K., Ohme, M., Tanaka, M., et al. 1986, The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression, *EMBO J.*, 5, 2043–2049.

15. Wolfe, K. H., Morden, C. W., and Palmer, J. D. 1992, Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant, *Proc. Natl. Acad. Sci. USA*, 89, 10648–10652.

16. Ohyama, K., Fukazawa, H., Kohchi, T., et al. 1986, Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA, *Nature*, 322, 572–574.

17. Sugiura, C., Kobayashi, Y., Aoki, S., et al. 2003, Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus, *Nucleic Acids Res.*, 31, 5324–5331.

18. Kugita, M., Kaneko, A., Yamamoto, Y., et al. 2003, The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants, *Nucleic Acids Res.*, 31, 716–721.
19. Sugiura, M. 1989, The chloroplast chromosomes in land plants, *Annu. Rev. Cell Biol.*, 5, 51–70.
20. Sugiura, M. 1992, The chloroplast genome, *Plant Mol. Biol.*, 19, 149–168.
21. Shimada, H., and Sugiura, M. 1989, Pseudogenes and short repeated sequences in the rice chloroplast genome, *Curr. Genet.*, 16, 293–301.
22. Palmer, J. D. 1990, Contrasting modes and tempos of genome evolution in land plant organelles, *Trends Genet.*, 6, 115–120.
23. Palmer, J. D. 1991, Plastid chromosomes: structure and evolution. In: Vasil, I. K., and Bogorad, L. (eds) *Cell Culture and Somatic Cell Genetics in Plants*, Vol. 7A, The Molecular Biology of Plastids, Academic Press, San Diego, pp. 5–53.
24. Shinozaki, K., Hayashida, N., and Sugiura, M. 1988, *The Molecular Biology of Plastids*, Academic Press, San Diego.
25. Kanno, A., and Hirai, A. 1993, A transcription map of the chloroplast genome from rice (*Oryza sativa*), *Curr. Genet.*, 23, 166–174.
26. Hermann, R. G., Oelmüller, R., Bichler, J., et al. 1991, The thylakoid membrane of higher plants: genes, their expression and interaction. In: Hermann, R. G., and Larkins, B. (eds) *Plant Molecular Biology 2*, Plenum Press, New York, pp. 411–427.
27. Hobbs, C. 1996, Ginseng: The Energy Herb, Botanica Press, Loveland, Connecticut, USA.
28. Mabberley, D. J. 1987, *The Plant-Book*. Cambridge Univ. Press, Cambridge.
29. Palmer, J. D. 1986, Isolation and structural analysis of the chloroplast DNA. In: Weissbach, A., and Weissbach, H. (eds) *Methods in Enzymology, Vol. 118*, Academic Press, New York, pp. 167–186.
30. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. 1997, The ClustalX windows interface: flexible strategies for multiple sequence alignment by using quality analysis tools, *Nucleic Acids Res.*, 24, 4876–4882.
31. Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. 2001, MEGA2: molecular evolutionary genetics analysis software, *Bioinformatics*, 17, 1244–1245.
32. Maddison, W. P. and Maddison, D. R. 2000, MacClade: analysis of phylogeny and character evolution. Version 4.0, Sinauer Asso. Inc., Sunderland, Massachusetts, USA.
33. Kimura, M. 1980, A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J. Mol. Evol.*, 16, 111–120.
34. Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. 2001, REPuter: the manifold applications of repeat analysis on a genomic scale, *Nucleic Acids Res.*, 29, 4633–4642.
35. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, 25, 955–964.
36. Zuker, M. 2003, Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res.*, 31, 3406–3415.
37. Sugiura, M., Shinozaki, K., Tanaka, M., et al. 1987, Split genes and cis/trans splicing in tobacco. In: von Wettstein, D., and Chua, N.-H. (eds), Plant Molecular Biology, Plenum Press, New York, pp. 65–76.
38. Kelchner, S. A. and Wendel, J. F. 1996, Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Curr. Genet.*, 30, 259–262.
39. Ruf, S., Kössel, H., and Bock, R. 1997, Targeted inactivation of a tobacco intron-containing open reading frame reveals a novel chloroplast-encoded photosystem I-related gene, *J. Cell Biol.*, 139, 95–102.
40. Swiatak, M., Kuras, R., Sokolenko, A., et al. 2001, The chloroplast gene ycf9 encodes a photosystem II (PSII) core subunit, psbZ, that participates in PSII supramolecular architecture. *Plant Cell*, 13, 1347–1367.
41. Cummings, M. P., King, L. M., and Kellogg, E. A. 1994, Slipped-strand mispairing in a plastid gene: rpoC2 in Grasses (Poaceae). *Mol. Biol. Evol.*, 11, 1–8.
42. Shimada, H., Fukuta, M., Ishikawa, M., and Sugiura, M. 1990, Rice chloroplast RNA polymerase genes: the absence of an intron in rpoC1 and the presence of an extra sequence in rpoC2, *Mol. Gen. Genet.*, 221, 395–402.
43. Levinson, G., and Gutman, G. A. 1987, Slipped-strand mispairing: a major mechanism for DNA sequence evolution, *Mol. Biol. Evol.*, 4, 203–221.
44. Ogihara, Y., Terachi, T., and Sasakuma, T. 1988, Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species, *Proc. Natl. Acad. Sci. USA*, 85, 8573–8577.
45. Shimada, H., and Sugiura, M. 1989, Pseudogenes and short repeated sequences in the rice chloroplast genome, *Curr. Genet.*, 16, 293–301.
46. Milligan, B. G., Hampton, J. N., and Palmer, J. D. 1989, Dispersed repeats and structural reorganization in subclover chloroplast DNA, *Mol. Biol. Evol.*, 6, 355–368.
47. Cato, S. A. and Richardson, T. E. 1996, Inter- and intraspecific polymorphism at chloroplast SSR loci and the inheritance of plastids in *Pinus radiata* D. Don., *Theor. Appl. Genet.*, 93, 587–592.
48. Provan, J., Corbett, G., Waugh, R., McNicol, J. W., Morgante, M., and Powell, W. 1996, DNA fingerprints of rice (*Oryza sativa*) obtained from hypervariable chloroplast simple sequence repeats, *Proc. R. Soc. Lond. B.*, 263, 1275–1281.
49. Powell, W., Morgante, M., McDevitt, R., Vendramin, G. G., and Rafalski, J. A. 1995, Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. *Proc. Natl. Acad. Sci. USA*, 92, 7759–7763.