

# Molecular Evolution of Insertions and Deletion in the Chloroplast Genome of *Silene*

Pär K. Ingvarsson,<sup>1</sup> Sarah Ribstein, and Douglas R. Taylor

Department of Biology, University of Virginia

Insertions, deletions, and inversions in the chloroplast genome of higher plants have been shown to be extremely useful for resolving phylogenetic relationships both between closely related taxa and among more basal lineages. Introns and intergenic spacers from the chloroplast genome are now increasingly used for phylogenetic and population genetic studies of populations from a single species, and it is therefore interesting to know whether indels can provide useful data and hence increase the power of intraspecific studies. Here, we show that indels in three cpDNA intergenic spacers and one cpDNA intron for two species of *Silene* evolve at slightly higher rates than base pair substitutions. Repeat indels appear to have the highest rate of evolution and are thus more prone to homoplasy. We show that coded indel data have high information content for phylogenetic analysis, and indels thus provide useful information to infer phylogenetic relationships at the intraspecific level.

## Introduction

Microstructural changes, such as insertions and deletions (indels) and inversions in the chloroplast genome of higher plants, can be extremely useful both for resolving phylogenetic relationships among basal lineages in the angiosperms (Graham et al. 2000) and for inferring relationships among more closely related taxa (Golenberg et al. 1993; Kelchner and Clark 1997; Kelchner 2000). Introns and intergenic spacers from the chloroplast genome are now increasingly used also for phylogenetic (Kelchner 2000) and population genetic (McCauley 1995) studies of either very closely related species or of populations from a single species. However, as the rate of base pair substitutions in the chloroplast is fairly low (Wolfe, Li, and Sharp 1987; Muse 2000) these studies often have very low resolution unless large stretches of DNA is sequenced. It would thus be valuable to know whether indels in these regions can also be of use for resolving relationships between closely related sequences and hence increase the power of intraspecific studies.

Several molecular processes are known to create indels. Polymerase slippages during DNA replication, so called slipped-strand mispairing (Levinson and Gutman 1987), add or subtract short repeat sequences, usually one or a few base pairs in length. Repeat structures in chloroplast DNA are primarily found in AT-rich regions and often involve long stretches of repeats of a single nucleotide (Kelchner 2000). Larger indels are often associated with the formation of hairpins (Kelchner and Wendel 1996) or stem-loop structures in DNA secondary structure (Kelchner 2000), and these indels may or may not show sequence similarity with the flanking region of the indel site. Different types of indels also show varying amounts of homoplasy. In general, in between-species studies, repeat indels seems to be more prone to homoplasy, simply because the rate at which they occur appears

to be higher than larger indels (Olsen 1999; Kelchner 2000). Another cause of homoplasy is multiple, overlapping indels within a single region of DNA (Kelchner 2000; Simmons and Ochoterena 2000).

Here, we present a study of the molecular evolution of indels in three intergenic spacers and one intron from the chloroplast genome of *Silene latifolia* and *S. vulgaris*. We show that indels in these four regions evolve at slightly higher rates than base pair substitutions. Repeat indels appear to evolve at higher rates than other types of indels and are thus more prone to homoplasy. We also show that the indel data have high information content for phylogenetic analysis, and coded indels can provide useful information to infer phylogenetic relationships at the intraspecific level.

## Material and Methods

*Silene latifolia* and *S. vulgaris* are weedy, short-lived perennial herbs, common throughout Europe. Both species are weeds of agriculture and have similar ecological and life history characteristics. The data used here were collected for a study on cytoplasmic genetic diversity in the two species (Ingvarsson and Taylor 2002). Seeds were collected from populations of both species scattered throughout Europe (from the Mediterranean northward into Scotland, and from the Atlantic Ocean eastward as far as Armenia), and DNA was isolated from one plant per population. Total genomic DNA was isolated from leaf tissue using DNeasy plant mini-prep kit (Qiagen Inc., Valencia, Calif.). We studied three chloroplast intergenic spacers, *trnL-trnF* (GenBank accession numbers AF518879 to AF518903 and AF519072 to AF519099), *trnH-psbA* (GenBank accession numbers AF518904 to AF5189028 and AF518957 to AF518984), and *trnG-trnS* (GenBank accession numbers AF518929 to AF518956 and AF519047 to AF519071), and one intron in *trnL* (GenBank accession numbers AF518985 to AF519012 and AF519013 to AF519037). The cpDNA regions were amplified using polymerase chain reaction (PCR) from 35 to 53 *S. latifolia* and 35 to 47 *S. vulgaris* plants (table 1). For all PCR amplifications, we used universal primers (Taberlet et al. 1991; Hamilton 1999). PCR products were cleaned using a QIAquick PCR Purification Kit (Qiagen Inc., Valencia, Calif.), and the cleaned PCR product was cycle

<sup>1</sup> Present address: Umeå Plant Science Centre, Department of Ecology and Environmental Science, University of Umeå, Umeå, Sweden.

Key words: insertions, deletions, chloroplast, *Silene*.

E-mail: pelle@eg.umu.se.

*Mol. Biol. Evol.* 20(11):1737–1740. 2003

DOI: 10.1093/molbev/msg163

*Molecular Biology and Evolution*, Vol. 20, No. 11,

© Society for Molecular Biology and Evolution 2003; all rights reserved.

**Table 1**  
**Results of cpDNA Analyses**

Species	Data	N	CI	RI
<i>Silene latifolia</i>	Base pair substitutions	18	0.854	0.838
	Insertion/deletions	11	0.933	0.989
	Repeat indels	9	0.614	0.899
<i>Silene vulgaris</i>	Base pair substitutions	10	1.000	1.000
	Insertion/deletions	5	0.765	0.917
	Repeat indels	12	0.653	0.775

NOTE.—CI = consistency index; RI = retention index.

sequenced directly using BigDye terminator ready reaction mix (Applied Biosystems Inc., Foster City, Calif.) and applied to an ABI377 automated sequencer (Applied Biosystems Inc.). To increase quality of the data, both forward and reverse strands were sequenced independently. Sequences were verified manually and contigs were assembled using the computer program Sequencher (Gene Codes Corporation, Inc.). Multiple sequence alignments were made using Pileup from the software package GCG (Wisconsin Package, v. 10.0, Accelrys Inc., Calif.) and adjusted manually. Insertions and deletions in the data were coded using presence/absence (0/1) for large indels and the number of repeat units (0 to 9) for mononucleotide and dinucleotide repeats (A, T, and AT repeats, respectively; see also Supplementary Material online).

In all analyses, we ignored singleton base pair substitutions and indels (i.e., characters that only occurred in a single individual in the data set and that hence were not parsimony informative). Phylogenetic trees were constructed in PAUP\* using maximum parsimony for two separate data sets, one containing only base pair substitutions and the other containing only the coded indels for each region. In all analyses, characters were weighted equally and were assumed to be unordered except when stated otherwise.

Repeat indels arise through replication slippage and are often assumed to evolve according to a stepwise mutation model (Levinson and Gutman 1987). We therefore performed all analyses of the indel data with repeat indels assigned as either unordered or ordered characters in PAUP\* (Swofford 1998). As a measure of the level of homoplasy in the data set, we scored the consistency index (CI) (Kluge and Farris 1969), the retention index (RI) and the rescaled consistency index (RC) (Farris 1989). Indels were then grouped by type (true indel or repeat indel) to determine whether the two types of indels provide the same phylogenetic resolution.

Next, we appended the two different data sets (base pair substitutions and indels) for the four chloroplast regions for individuals where all four regions had been sequenced. This combined data set consisted of 25 *S. latifolia* and 29 *S. vulgaris* individuals. We performed an incongruence length difference (ILD) test to determine whether the two types of data (base pair substitution or indels) produced phylogenetic trees that were congruent with each other (Farris et al. 1994). The data set containing base pair substitutions included only variable sites to avoid artificially inflating the results of ILD test (Cunningham 1997; Lee 2001).

## Results and Discussion

Several aspects of our data indicate that indels evolve at similar or slightly higher rates than base pair substitutions in the chloroplast regions studied. In *Silene latifolia*, there were a total of 50 segregating sites and 25 unique indels, of which 18 and 20, respectively, were parsimony informative. In *Silene vulgaris*, there were 27 segregating sites and 27 unique indels, of which 10 and 18, respectively, were parsimony informative (table 1). Both base pair substitutions and indels had low homoplasy, and indels can thus be an important complement to base pair substitutions for resolving phylogenetic patterns within species.

However, not all indels are equal. Single and dinucleotide repeat indels appeared to evolve at a faster rate than either base pair substitutions or other types of indels, as evidenced by the greater degree of homoplasy for repeat indels in both *S. latifolia* and *S. vulgaris* (table 1; see also Supplementary Material online). This was true regardless of whether repeat indels were scored as unordered or ordered characters in the parsimony analyses (data not shown).

The phylogenetic utility of indels is corroborated by the ILD test that show no evidence of phylogenetic incongruence between parsimony trees based on base pair substitution data or indel data alone (ILD test;  $P = 0.98$  for *S. latifolia* and  $P = 0.98$  for *S. vulgaris*). Not surprisingly, based on strict consensus trees, a combined data set of base pair substitutions and indels produced phylogenetic trees with higher resolution than trees based on base pair substitutions alone (fig. 1). Chloroplast DNA have lower substitution rate than nuclear DNA in plants (Wolfe, Li, and Sharp 1987; Muse 2000), and sequence diversity and phylogenetic resolution at the intraspecific levels is generally low for moderate amounts of sequence data (1 to 2 kb). However, our results indicate that coded indels have levels of homoplasy comparable with base pair substitutions, and including coded indels may therefore increase the resolution of phylogenetic studies. Moreover, some forms of indels have levels of homoplasy virtually identical to base pair substitutions, where other types, primarily repeat indels, are far less reliable. For example, we constructed phylogenetic trees for both *S. latifolia* and *S. vulgaris* using combined data from all four chloroplast regions, including both base pair substitutions and indels. In *S. latifolia*, the rescaled consistency index (RC) increased from 0.526 to 0.620 when repeat indels were excluded, whereas it is 0.832 for base pair substitutions alone. For *S. vulgaris*, RC equals 0.441 with all characters included in the analysis and 0.681 without repeat indels. For *S. vulgaris*, RC equals 0.877 for the base pairs substitutions alone.

In an earlier study (Ingvarsson and Taylor 2002), we showed that cpDNA sequence diversity was reduced by about 50% in *S. vulgaris* compared with *S. latifolia*, despite having equal levels of sequence diversity at a nuclear gene. Ingvarsson and Taylor (2002) argued that this was due to cytoplasmic male sterility causing a rapid turnover of cytoplasmic lineages in *S. vulgaris*. Somewhat surprisingly, the number of unique indels found in this

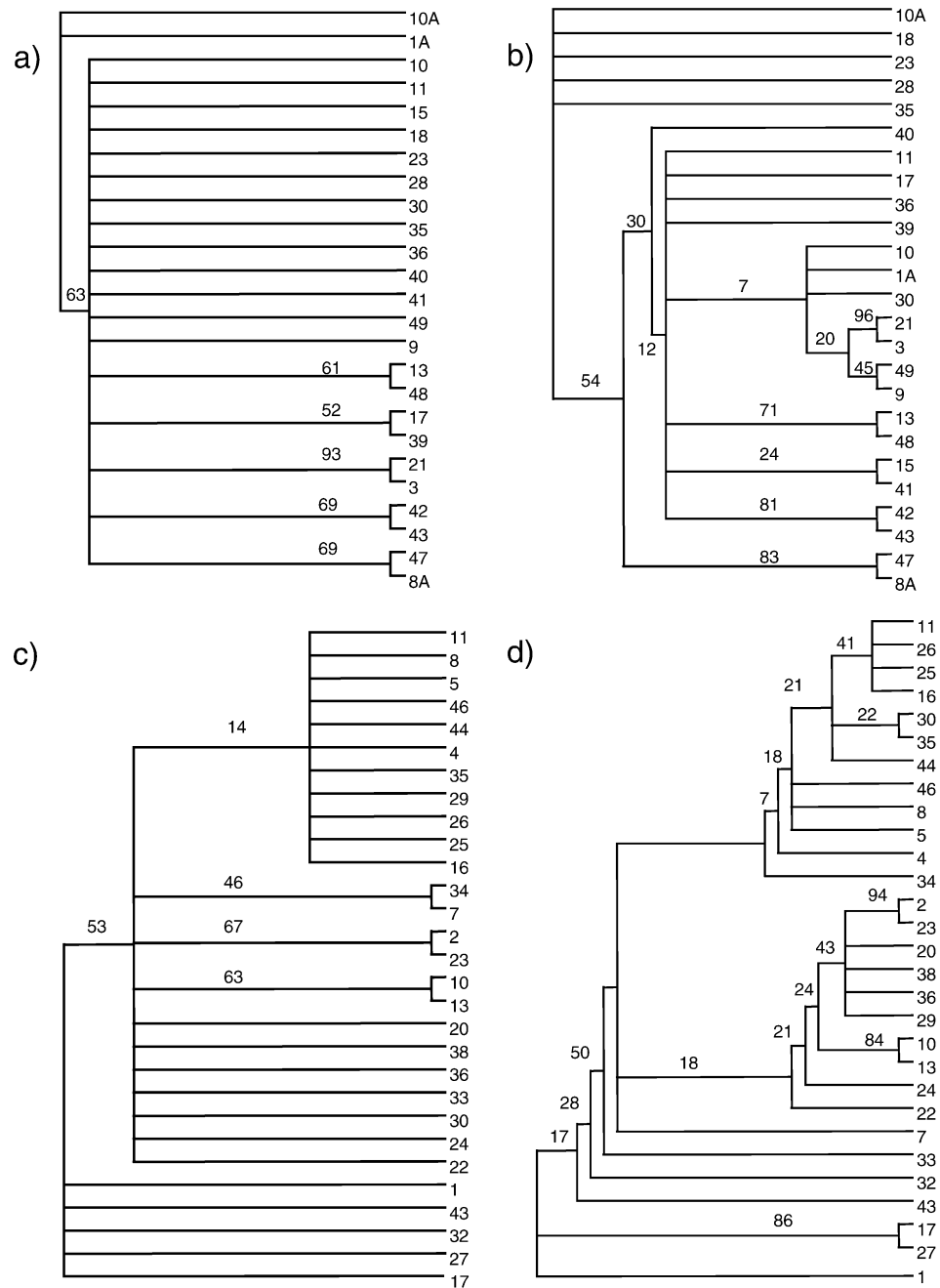


FIG. 1.—Strict consensus gene genealogies of cpDNA sequences in two *Silene* species with and without the use of coded indels. (a) *Silene latifolia* data set, including base pair substitutions only; consensus based on five equally parsimonious trees. (b) *S. latifolia* base pair substitutions and indels; consensus based on 24 equally parsimonious trees. (c) *S. vulgaris* data set, including base pair substitutions only; consensus based on the most parsimonious tree. (d) *S. vulgaris* base pair substitutions and indels; consensus based on 120 equally parsimonious trees. Numbers above the nodes are bootstrap support based on 100 bootstrap replicates for all branches were resolved on the strict consensus tree.

study did not differ between the two species (table 1). However the two species differed markedly in the types of indels they carry. Specifically, in *S. vulgaris*, most indels (71%) were rapidly evolving repeat indels, whereas in *S. latifolia*, only 47% of indels were repeat indels. The reduction in the number of slowly evolving indels (with low levels of homoplasy) in *S. vulgaris* relative to *S. latifolia* was roughly the same as the observed difference in base pair substitutions between the species (*S. vulgaris*

has 50% fewer slowly evolving indels and 46% fewer for base pair substitutions). This provides additional evidence that the molecular evolution of nonrepeat indels and base pair substitutions are roughly similar. The reduction in chloroplast phylogenetic diversity in *S. vulgaris* relative to *S. latifolia* at slowly evolving sites most likely reflects purifying selection in the chloroplast genome of *S. vulgaris* (Ingvarsson and Taylor 2002). The lack of a clear difference between the two species at rapidly evolving

repeat indels is more puzzling, although it is possible that the extensive homoplasy observed for repeat indels sites makes it more difficult to detect any differences in diversity at such sites.

We have shown that phylogenetic studies performed at the intraspecific levels may benefit from including coded indel data, although our data suggest that repeat indels are less reliable than other types of indels and should be avoided, if possible. At the present time, using coded indels has the drawback of restricting the analytical methods one can use. We have shown their utility in a parsimony-based analysis, but it would be more of a challenge to employ a maximum-likelihood-based approach without specific models of evolution, such as those available for base pair substitutions (see McGuire, Denham, and Balding [2001] for an example of such a model that includes indels). Nevertheless, we see the present study as a necessary first step in developing models of evolution for the different types of insertions and deletions in DNA sequence data.

### Acknowledgments

This research has been supported by grants from the National Science Foundation (to D.R.T.) and the Swedish Research Council (to P.K.I.).

### Literature Cited

- Cunningham, C. W. 1997. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* **14**:733–740.
- Farris, J. S. 1989. The retention index and the rescaled consistency index. *Cladistics* **5**:417–419.
- Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1994. Testing significance of congruence. *Cladistics* **10**:315–319.
- Golenberg, E. M., M. T. Clegg, M. L. Durbin, J. Doebley, and D. P. Ma. 1993. Evolution of a non-coding region of the chloroplast genome. *Mol. Phylogenet. Evol.* **2**:52–64.
- Graham, S. W., P. A. Reeves, A. C. E. Burns, and R. G. Olmstead. 2000. Microstructural changes in non-coding DNA: interpretation, evolution and utility of indels and inversions in basal angiosperm phylogenetic inference. *Int. J. Plant Sci.* **161**:S83–S96.
- Hamilton, M. B. 1999. Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. *Mol. Ecol.* **8**:521–523.
- Ingvarsson, P. K., and D. R. Taylor. 2002. Genealogical evidence for epidemics of selfish genes. *Proc. Natl. Acad. Sci. USA* **99**:11265–11269.
- Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Ann. MO Bot. Gard.* **87**:499–527.
- Kelchner, S. A., and L. G. Clark. 1997. Molecular evolution and phylogenetic utility of the chloroplast *rpl16* intron in *Chusquea* and the Bambusoideae (Poaceae). *Mol. Phylogenet. Evol.* **8**:385–397.
- Kelchner, S. A., and J. F. Wendel. 1996. Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Curr. Genet.* **30**:259–262.
- Kluge, A. G., and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* **18**:1–32.
- Lee, M. S. Y. 2001. Uninformative characters and apparent conflict between molecules and morphology. *Mol. Biol. Evol.* **18**:676–680.
- Levinson, G., and G. A. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**:203–221.
- McCauley, D. E. 1995. The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends Ecol. Evol.* **10**:190–202.
- McGuire, G., M. C. Denham, and D. J. Balding. 2001. Models of sequence evolution for DNA sequences containing gaps. *Mol. Biol. Evol.* **18**:481–490.
- Muse, S. V. 2000. Examining rates and patterns of nucleotide substitution in plants. *Plant Mol. Biol.* **42**:25–43.
- Olsen, K. M. 1999. Minisatellite variation in a single-copy nuclear gene: phylogenetic assessment of repeat length homoplasy and mutational mechanism. *Mol. Biol. Evol.* **16**:1406–1409.
- Simmons, M. P., and H. Ochoterena. 2000. Gaps and characters in sequence-based phylogenetic analysis. *Syst. Biol.* **42**:369–381.
- Swofford, D. L. 1998. PAUP\*: phylogenetic analysis using parsimony (\*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
- Taberlet, P., L. Gielly, G. Pautou, and J. Bouvet. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol. Biol.* **17**:1105–1109.
- Wolfe, K. H., W-H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* **84**:9054–9058.

Geoffrey McFadden, Associate Editor

Accepted May 5, 2003