ORIGINAL PAPER

Scot A. Kelchner · Jonathan F. Wendel

# Hairpins create minute inversions in non-coding regions of chloroplast DNA

Received: 5 March 1996 / 17 April 1996

**Abstract** Minute inversions (4 bp in length), associated with probable hairpin secondary structures, were inferred from comparative analysis of *rpl16* intron sequences from the chloroplast genomes of *Chusquea* species and related bamboos (Poaceae). The inverted sequences, which appear to have arisen independently on several occasions, comprise entire loops of the putative hairpins. The process of inversion seems dependent upon the stem length of the hairpin and its estimated free energy of formation. A similar inversion was uncovered for other plants in a previously published data set for a different non-coding region of the chloroplast genome, suggesting that the inversional process may be a common feature of non-coding DNA evolution. Several implications for phylogenetic analysis are noted.

**Key words** Inversions · Non-coding DNA · Phylogenetics · Molecular evolution · *rpl16* intron

## Introduction

Because non-coding regions of DNA evolve relatively rapidly, they have become popular for phylogenetic studies at lower taxonomic ranks. In plants, both nuclear sequences, such as the internal transcribed spacer (ITS) region of the 45*s* rDNA repeat (reviewed in Baldwin et al. 1995), and introns or intergenic spacers in the chloroplast genome (Golenberg et al. 1993; Gielly and Taberlet 1994; Johnson and Soltis 1994; vanHam et al. 1994) are often employed for such studies. One generality to emerge from this accumulating literature is that, in addition to nucleotide substitutions, sequences from related taxa are often distinguished by structural mutations. Most common in this respect are the insertion and deletion events ("indels") that are inferred to have taken place based on gaps in sequence alignments. In many cases the genesis of these length mutations is obscure, although in chloroplast DNA (cpDNA) the causative phenomena include slipped-strand mispairing (Takaiwa and Sugiura 1982; Zurawski et al. 1984; Golenberg et al. 1993; vanHam et al. 1994), intramolecular recombination between adjacent or nearby repeats (Palmer et al. 1985; Ogihara et al. 1988; Kanno and Hirai 1992; Kanno et al. 1993; Morton and Clegg 1993), and stem-loop secondary structure formations (Golenberg et al. 1993; Gielly and Taberlet 1994; vanHam et al. 1994; Ferris et al. 1995).

In addition to length mutations, the chloroplast genome is also subject to inversions although, in contrast to length mutations, inversions appear to be sufficiently rare that their distribution may be phylogenetically meaningful at higher taxonomic ranks (Downie and Palmer 1992). Also, unlike the many small indels commonly observed in non-coding cpDNA, which typically range in size from a single bp to perhaps several dozen bp in length, reported cpDNA inversions are considerably larger (approximately 1–62 kb; Downie and Palmer 1992). One mechanism underlying these mutations is thought to be intramolecular recombination between repeats in inverse orientation (Howe 1985; Palmer et al. 1987; Ogihara et al. 1988; Milligan et al. 1989; Hong et al. 1993; Knox et al. 1993; Hoot and Palmer 1994), a supposition supported by the positive correlation between the presence and frequency of dispersed repeats and inversional rearrangements in cpDNA (Palmer 1991). In principle, similar recombinational processes may operate on a finer scale, involving adjacent inverted repeats whose interaction would lead to small inversions associated with local secondary structural features.

The purpose of the present paper is to report what we believe is precisely this phenomenon. As part of an ongoing systematic study of a group of New World woody bamboos in the genus *Chusquea*, we have been generating sequence data for the rapidly evolving intron in the plastome gene *rpl16*, which encodes ribosomal protein L16 (Posno et al. 1986). In comparing sequences from differ-

S. A. Kelchner · J. F. Wendel (✉)
Department of Botany, Iowa State University, Ames, IA 50011, USA

Communicated by K. J. Newton

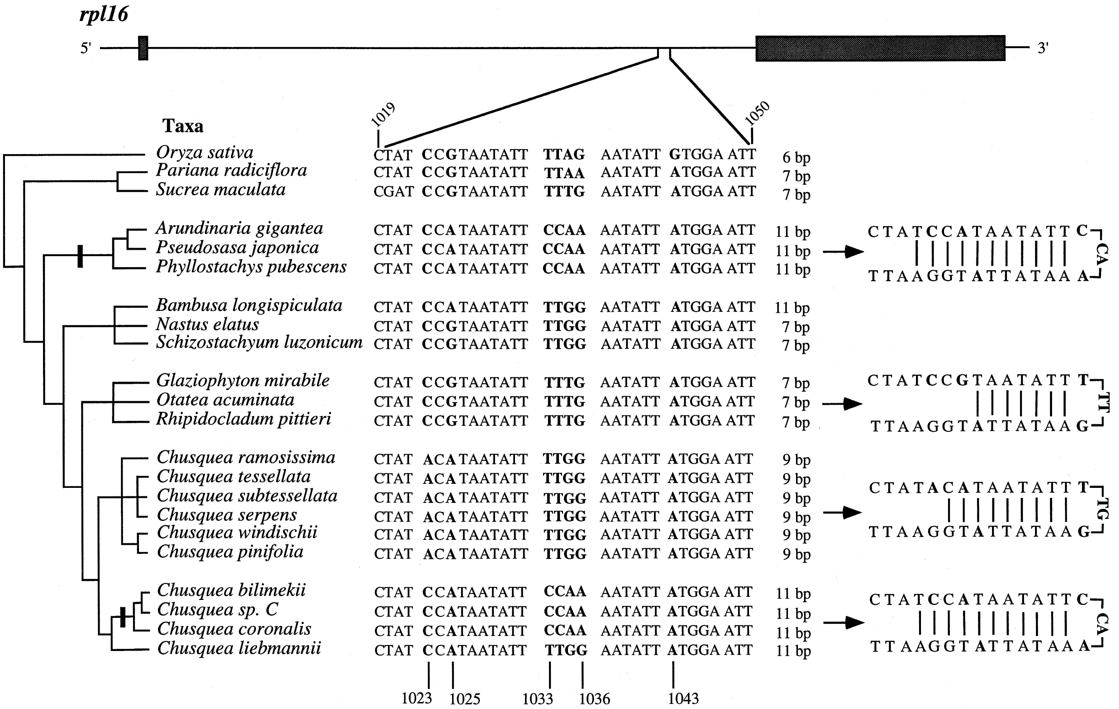
rpl16
5'
3'
1019
1050
Taxa
Oryza sativa CTAT CCGTAATATT TTAG AATATT GTGGA ATT 6 bp
Pariana radiciflora CTAT CCGTAATATT TTAA AATATT ATGGA ATT 7 bp
Sucrea maculata CGAT CCGTAATATT TTTG AATATT ATGGA ATT 7 bp
Arundinaria gigantea CTAT CCA TAATATT CCAA AATATT ATGGA ATT 11 bp
Pseudosasa japonica CTAT CCA TAATATT CCAA AATATT ATGGA ATT 11 bp
Phyllostachys pubescens CTAT CCA TAATATT CCAA AATATT ATGGA ATT 11 bp
Bambusa longispiculata CTAT CCA TAATATT TTGG AATATT ATGGA ATT 11 bp
Nastus elatus CTAT CCGTAATATT TTGG AATATT ATGGA ATT 7 bp
Schizostachyum luzonicum CTAT CCGTAATATT TTGG AATATT ATGGA ATT 7 bp
Glaziophyton mirabile CTAT CCGTAATATT TTTG AATATT ATGGA ATT 7 bp
Otatea acuminata CTAT CCGTAATATT TTTG AATATT ATGGA ATT 7 bp
Rhipidocladum pittieri CTAT CCGTAATATT TTTG AATATT ATGGA ATT 7 bp
Chusquea ramosissima CTAT ACA TAATATT TTGG AATATT ATGGA ATT 9 bp
Chusquea tessellata CTAT ACA TAATATT TTGG AATATT ATGGA ATT 9 bp
Chusquea subtessellata CTAT ACA TAATATT TTGG AATATT ATGGA ATT 9 bp
Chusquea serpens CTAT ACA TAATATT TTGG AATATT ATGGA ATT 9 bp
Chusquea windischii CTAT ACA TAATATT TTGG AATATT ATGGA ATT 9 bp
Chusquea pinifolia CTAT ACA TAATATT TTGG AATATT ATGGA ATT 9 bp
Chusquea bilimekii CTAT CCATAATATT CCAA AATATT ATGGA ATT 11 bp
Chusquea sp. C CTAT CCATAATATT CCAA AATATT ATGGA ATT 11 bp
Chusquea coronalis CTAT CCATAATATT CCAA AATATT ATGGA ATT 11 bp
Chusquea liebmannii CTAT CCATAATATT TTGG AATATT ATGGA ATT 11 bp
1023 1025 1033 1036 1043
CTATCCATAATATTC CA
TTAAGGTATTATAAA
CTATCCGTAATATTT TT
TTAAGGTATTATAAG
CTATACATAATATTT TG
TTAAGGTATTATAAG
CTATCCATAATATTC CA
TTAAGGTATTATAAA

**Fig. 1** Association of minute inversions with stem-loop hairpins in the intron of the chloroplast gene *rpl16*. Sequence data were generated (using standard PCR protocols and an ABI automated sequencer) for the entire *rpl16* intron (approximately 1.1 kb) in the grass species shown on the left. All but the top three taxa are in the grass subfamily Bambusoideae. The schematic at the top of the figure illustrates the position of the intron relative to the two exons (*shaded boxes*; approximately 8 bp and 600 bp, respectively). Phylogenetic relationships (tree on left) were hypothesized based on sequence variation for *rpl16* (data not shown) and the 2.1-kb chloroplast gene *ndhF* (Clark et al. 1995; and unpublished). Sequence variation from nucleotide 1019 to 1050 is shown, and particularly relevant nucleotides (see text) are highlighted in *bold* and *numbered* below. Putative hairpin structures are exemplified (right) and stem lengths are shown. The *rpl16* sequence from *O. sativa* is from Sugiura (1989); the remaining sequences may be located under GenBank accession numbers U54740–U54760 in the same order as shown in the figure from top to bottom

ent grasses, we observed polymorphisms that are most parsimoniously interpreted as having arisen from small (4 bp) inversions associated with probable hairpin structures.

## Results and discussion

The phylogenetic context and relevant sequence data are shown in Fig. 1. Interspecific comparisons demonstrate that the region of the *rpl16* intron corresponding to nucleotide positions 1023 through 1043 is characterized by sequence variation. Some of these polymorphisms presumably are due to point mutations; for example, those at positions 1023, 1025, and 1043. Other polymorphisms, however, are not so readily explained. Of particular relevance are four adjacent nucleotides (1033–1036) that are also variable. Given the overall high level of sequence similarity in the *rpl16* intron among the taxa studied (data not shown), and the fact that these are four consecutive nucleotide positions, it seems unlikely that the polymorphisms arose from independent substitutions. Equally improbable would be a scenario involving numerous indels, whereby both the alignment and the inference of nucleotide polymorphisms are incorrect. Instead, we suggest that four adjacent nucleotide polymorphisms are more parsimoniously explained by single mutational events, in this case small (4-bp) inversions. The organismal framework shown in Fig. 1 suggests that similar inversions have occurred twice, once in each of two disparate lineages of woody bamboos.

Assuming we are correct in concluding that inversions have occurred, we explored whether clues to their genesis are evident in the sequence data. While searching for this mechanistic explanation, we identified probable hairpin formations (using OLIGO 4.0; Rychlik and Rhoads 1989) involving bases flanking the inverted sequences. In all taxa exhibiting an inversion, the most probable single-stranded hairpin formation has a $\Delta G$ of –12.4 kcal/mol, indicating a high likelihood of spontaneous formation. In our data a $\Delta G$ of this magnitude exists only in those taxa that exhibit the longest hairpin stems (=11 bp; Fig. 1). All other taxa have shorter stems, correspondingly smaller $\Delta G$ values for hairpin formation, and uninverted nucleotides in positions 1033–1036. In *Chusquea ramosissima*, for example, the stem length is 9 instead of 11 bp due to an A instead of a C at position 1023; accordingly, $\Delta G$ for hairpin formation is only –7.7 kcal/mol (Fig. 1). Due to the decreased probability of hairpin formation, one might expect that the opportunity for inversions is also diminished. This appears to be the case, at least as judged by the absence of inversions in all taxa whose stem length is less than 11 bp.

At present little is known about the mechanism underlying the inversions. Given the fact that the inversions comprise entire loops of putative stem-loop hairpins, the mechanism may involve excision of four bases at the stem-loop interface and ligation in an inverted orientation, or perhaps recombination within the stemmed region itself, conceivably involving a four-stranded crossover event.

The phylogeny illustrated in Fig. 1 (Clark et al. 1995; unpublished data) suggests that the origin of the "long-stemmed" (11-bp) hairpins traces to point substitutions, at positions 1025 and 1043, that lengthened pre-existing "short" stems. In the phylogenetically basal *Oryza sativa* sequence, the hairpin stem (if formed) is 6 bp in length, and has a $\Delta$G of only –3.3 kcal/mol. If this is representative of the ancestral condition, two point mutations, both G to A transitions, are required to transform this secondary structure into a stem-loop hairpin with a stem of 11 bp, and, presumably, a favorability of formation that is high enough to promote inversional mutations. If instead, the sequences of *Pariana radiciflora* and *Sucrea maculata* (each with a 7-bp stem) are more representative of the ancestral condition, then only a single transition (at position 1025) is required to create a long-stemmed derivative. As illustrated by mapping stem-length onto the phylogeny (Fig. 1), there is considerable flux in this character within the woody bamboos. It is clear, though, that one or two nucleotide transitions occurred in parallel in two different lineages (depending on the ancestral condition assumed), once in the lineage that includes *Arundinaria gigantea, Pseudosasa japonica* and *Phyllostachys pubescens,* and a second time in the branch that includes *Chusquea bilimekii, Chusquea "sp. C"* (an unnamed new species; L. Clark, personal communication), and *C. coronalis*. We also note that having a long stem is insufficient to guarantee that an inversion will occur, as evidenced by the two woody bamboos, *Chusquea liebmanni* and *Bambusa longispiculata,* both of which have 11-bp stems and uninverted nucleotides at positions 1033–1036.

If small hairpin formation is associated with inversions, one might predict that similar secondary structures would be discovered in other taxa and for other sequences. To examine this possibility, we surveyed the cpDNA literature with a focus on non-coding regions. Despite the fact that relatively few data sets are available that meet the criteria necessary to infer the existence of inversions (these criteria include sequence data from at least two closely related taxa for relatively rapidly evolving spacers or introns), it appears that a similar phenomenon has occurred in the *atpB-rbcL* intergenic spacer in the grass *Pennisetum glaucum* (Golenberg et al. 1993). The authors noted a possible six-base inversion starting at position 527, which we analyzed using OLIGO 4.0 in conjunction with sequences flanking the putative inversion. This analysis led us to infer that the inverted nucleotides are associated with a stem-loop hairpin with a $\Delta$G of spontaneous formation of –11.0 kcal/mol and with a stem of nine paired bases. As in our own data, the entire loop of the hairpin structure has been inverted, although in this case it is 6 rather than 4 bp in length. At present, we are not aware of any additional examples of this phenomenon but, as data accumulate on other taxa and genomic regions, it will be of interest to assess the generality of the association of small inverted repeats with inversional mutations. It may also be that insights into the recombinational mechanism will derive from this comparative approach, particularly concerning the relationships between loop size, stem length, and the propensity to suffer inversional mutations.

The full significance of minute inversions with respect to the molecular evolution of non-coding DNA, as well as the details of the inversion mechanism, remain to be determined. There are, however, several clear phylogenetic implications. First, although the frequency of small inversions relative to other structural changes and point mutations is not known, they apparently are common enough that a similar or identical inversion may occur independently in two or more lineages. In our data this was evidenced by mapping the inversion onto an independently derived cpDNA phylogeny, which demonstrates that the inversion occurred at least twice, once in each of two distinct lineages of woody bamboos (Fig. 1). We conclude, therefore, that these mutational events need not be considered "signal" phylogenetic characters; indeed, small inversions of the type reported here may be just as subject to parallelisms and/or reversals (homoplasy) as nucleotide substitutions. In this respect, they provide a sharp contrast to the larger inversions (1–62 kb) previously reported in chloroplast genomes, many of which are phylogenetically meaningful even at higher taxonomic levels (Downie and Palmer 1992; see, however, Downie and Palmer 1994; Hoot and Palmer 1994 for recent examples of homoplasious inversions).

A second implication of the existence of small inversions in non-coding cpDNA (and possibly other genomes) is that their occurrence may influence sequence alignments and character interpretation. Specifically, if an inversion is not recognized in a data set, one of two situations will obtain: either gaps will be inserted to account for the positional polymorphisms, or multiple substitutions for independent characters will be inferred. Both introduce error into phylogenetic analysis, and as such we suggest that where possible the inversion hypothesis should be explored by scrutinizing the data for flanking inverted repeats. Because inversions are associated with paired stem bases, the possibility of compensatory mutations dictates that consideration be given to differential treatment of stemmed vs non-stemmed bases in phylogenetic analysis (e.g., Dixon and Hillis 1993). The ability to detect flanking inverted repeats, as well as inversions, will depend on many factors, including the frequency of inversions, their antiquity relative to subsequent and potentially confounding length mutations and point substitutions, and the range of sequence divergences included in the taxa under study. Our results suggest that it should be possible to detect small inversions within genera and perhaps even in family level analyses.

Finally, it seems likely that the mutational category reported here will turn out to be more widely distributed than in the *rpl16* intron or even the chloroplast genome. Given the fact that we were able to identify an additional example from the data of Golenberg et al. (1993), and the existence of at least the outlines of a plausible mechanism, it seems probable that additional examples of small inversions associated with hairpins will be found, in plastomes as well as in other genomes. These will be revealed as new data sets accumulate and as existing data are re-analyzed for the presence of local stem-loop structures.

**Acknowledgements** We thank W. Zhang for providing several *rpl16* sequences, R. Cronn, J. Davis, and R. Small for helpful discussion, S. Rodermel and J. Colbert for reviewing the manuscript, and L. Clark for her assistance with and direction of the *Chusquea* phylogeny project, from which this study derived. The work was supported by an award from the National Science Foundation.

## References

Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann Missouri Bot Gard 82:247–277

Clark LG, Zhang W, Wendel JF (1995) A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. Syst Bot 20:436–460

Dixon MT, Hillis DM (1993) Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol Biol Evol 10:256–267

Downie SR, Palmer JD (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis DE et al. (eds) Molecular systematics of plants. Chapman and Hall, New York, pp 14–35

Downie SR, Palmer JD (1994) A chloroplast DNA phylogeny of the Caryophyllales based on structural and inverted repeat restriction-site variation. Syst Bot 19:236–252

Ferris C, Oliver RP, Davy AJ, Hewitt GM (1995) Using chloroplast DNA to trace postglacial migration routes of oaks into Britain. Mol Ecol 4:731–738

Gielly L, Taberlet P (1994) The use of chloroplast DNA to resolve plant phylogenies: non-coding versus *rbcL* sequences. Mol Biol Evol 11:769–777

Golenberg EM, Clegg MT, Durbin ML, Doebley J, Ma DP (1993) Evolution of a non-coding region of the chloroplast genome. Mol Phyl Evol 2:52–64

Hong YP, Hipkins VD, Strauss SH (1993) Chloroplast DNA diversity among trees, populations and species in the California closed-cone pines (*Pinus radiata*, *Pinus muricata,* and *Pinus attenuata*). Genetics 135:1187–1196

Hoot SB, Palmer JD (1994) Structural rearrangements, including parallel inversions, within the chloroplast genome of *Anemone* and related genera. J Mol Evol 38:274–281

Howe CJ (1985) The endpoints of an inversion in wheat chloroplast DNA are associated with short repeated sequences containing homology to *att*-lambda. Curr Genet 10:139–145

Johnson LA, Soltis DE (1994) *matK* DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. Syst Bot 19:143–156

Kanno A, Hirai A (1992) Comparative studies of the structure of chloroplast DNA from four species of *Oryza*: cloning and physical maps. Theor Appl Genet 83:791–798

Kanno A, Watanabe N, Nakamura I, Hirai A (1993) Variations in chloroplast DNA from rice (*Oryza sativa*): differences between deletions mediated by short direct-repeat sequences within a single species. Theor Appl Genet 86:579–584

Knox EB, Downie SR, Palmer JD (1993) Chloroplast genome rearrangements and the evolution of giant lobelias from herbaceous ancestors. Mol Biol Evol 10:414–430

Milligan BG, Hampton JN, Palmer JD (1989) Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol Biol Evol 6:355–368

Morton BR, Clegg MT (1993) A chloroplast DNA mutational hotspot and gene conversion in a non-coding region near *rbcL* in the grass family (Poaceae). Curr Genet 24:357–365

Ogihara Y, Terachi T, 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

Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad L, Vasil IK (eds) Cell culture and somatic cell genetics of plants, vol 7A. Academic Press, San Diego, pp 5–53

Palmer JD, Jorgensen RA, Thompson WF (1985) Chloroplast DNA variation and evolution in *Pisum*: patterns of change and phylogenetic analysis. Genetics 109:195–213

Palmer JD, Nugent JM, Herbon LA (1987) Unusual structure of geranium chloroplast DNA: a triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. Proc Natl Acad Sci USA 84:769–773

Posno M, van Vliet A, Groot GSP (1986) The gene for *Spirodela oligorhiza* chloroplast ribosomal protein homologous to *E. coli* ribosomal protein L16 is split by a large intron near its 5′ end: structure and expression. Nucleic Acids Res 14:3181–3195

Rychlik W, Rhoads RE (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res 17:8543–8551

Sugiura M (1989) *Oryza sativa* chloroplast DNA 134 525 bp. Nagoya University Center for Gene Research, Nagoya, Japan

Takaiwa F, Sugiura M (1982) Nucleotide sequence of the 16*s*–23*s* spacer region in a rRNA gene cluster from tobacco chloroplast DNA. Nucleic Acids Res 10:2665–2676

vanHam RCHJ, t’Hart H, Mes THM, Sandbrink JM (1994) Molecular evolution of non-coding regions of the chloroplast genome in the Crassulaceae and related species. Curr Genet 25:558–566

Zurawski G, Clegg MT, Brown AHD (1984) The nature of nucleotide sequence divergence between barley and maize chloroplast DNA. Genetics 106:735–749