

## MICROSTRUCTURAL CHANGES IN NONCODING CHLOROPLAST DNA: INTERPRETATION, EVOLUTION, AND UTILITY OF INDELS AND INVERSIONS IN BASAL ANGIOSPERM PHYLOGENETIC INFERENCE

Sean W. Graham,<sup>1</sup> Patrick A. Reeves, Analiese C. E. Burns, and Richard G. Olmstead

Department of Botany, Box 355325, University of Washington, Seattle, Washington 98195-5325, U.S.A.

Microstructural changes in several very slowly evolving chloroplast introns and intergenic spacers were characterized across a broad range of angiosperms, including most of the major basal lineages. Insertion/deletion events (indels) in the surveyed noncoding regions of the large inverted repeat (IR) region were shown to be rarer than nucleotide substitutions and thus constitute one of the slowest and least homoplastic types of data available to plant systematists. In our study we scored 180 indels in noncoding regions, of which 36 were parsimony informative within the angiosperms. Because they are relatively few in number, their general utility is currently limited. However, they provide support for specific major taxa, including the angiosperms as a whole, the water lilies, and Illiciaceae and relatives. Support for the basalmost angiosperm split is largely inconclusive, but a single indel supported a basal placement of the water lilies, not *Amborella*. We estimate that roughly double or triple the current amount (ca. 2.2 kb) of noncoding IR DNA would be required to obtain indel support for most of the deepest branches at the base of the angiosperms. A variety of molecular processes appear to be responsible for the observed indels. Indels are more frequently associated with tandem repeat sequences than not. Insertions are significantly more frequently associated with tandem repeats than are deletions. The latter finding may be, in part, a function of an ascertainment bias for insertions versus deletions. Single-base indels were the most common size class, but there was an unexplained deficit of some other small indel size classes. Coding indels can be problematical, particularly when they overlap among taxa in an alignment. We favor one simple scheme for coding overlapping indels but argue that no existing scheme for coding overlapping indels for phylogenetic analysis, or dealing with them in alignment, is ideal. Several small inversions were observed. These included the most homoplastic microstructural character in the current study. Each inversion was associated with short flanking inverted repeats.

*Keywords:* alignment, *Amborella*, ascertainment bias, basal angiosperms, chloroplast introns, DNA inversions, indels, indel coding, intergenic spacers, molecular evolution, noncoding DNA, tandem repeats.

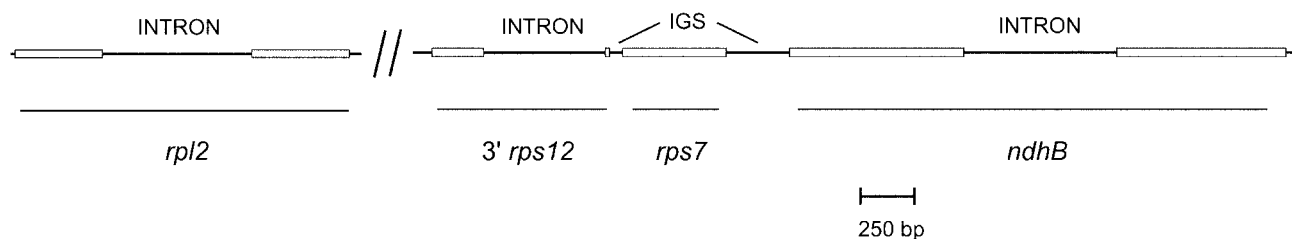
### Introduction

Introns and intergenic spacer regions of chloroplast DNA have become extremely important tools in the phylogenetic analysis of a broad range of plant groups, at a variety of taxonomic levels (reviewed in Kelchner 2000). Microstructural changes, primarily those involving insertion/deletion events, or “indels,” are especially frequent in these regions. This can make alignment a difficult and often subjective process. It is well recognized that there is significant room for improvement in existing DNA alignment algorithms. For example, too little attention has been paid to the molecular processes that generate structural changes in DNA (e.g., Kelchner 2000), and there is no firm statistical framework for performing alignment (Thorne et al. 1992). It has even been suggested that alignment programs developed for coding regions may be generally inadequate for noncoding data (Kelchner and Clark 1997). This is problematic since sequence alignment is the first and most

important assessment of nucleotide homology, the assignment of “primary homology” *sensu* De Pinna (1991). In addition, it has been argued that computer-generated alignments are more objective than those produced manually because they result from the application of explicit rules (Giribet and Wheeler 1999). However, any alignment can only be as good as the rules embodied in the alignment algorithm, and these are likely to be only rough approximations of the actual processes of molecular evolution responsible for generating DNA sequence diversity. In practice, therefore, most workers end up adjusting computer-generated alignments “by eye,” to a greater or lesser degree, prior to formal phylogenetic analysis.

Debate about how to proceed with computer-aided alignment has largely centered around whether and how to assess and incorporate information on, for example, (i) the phylogenetic relatedness of taxa; (ii) the local sequence context and DNA secondary structure, and (iii) the size and form of gap penalties (e.g., Hein 1989, 1990; Thorne et al. 1991, 1992; Gu and Li 1995; Hancock and Vogler 2000; Kelchner 2000; Simmons and Ochoterena 2000). No single alignment program deals satisfactorily with all (or perhaps any) of these factors. Indels are generated by a variety of molecular processes, and there is a growing recognition that this should be considered

<sup>1</sup> Author for correspondence. Current address: Department of Biological Sciences, Biological Sciences Centre, University of Alberta, Edmonton, Alberta T6G 2E9, Canada; e-mail swgraham@ualberta.ca.



**Fig. 1** Noncoding regions scored for indels. All regions are found in the inverted repeat (IR) region of the chloroplast genome. Three introns and two intergenic spacer regions (those between *3'-rps12*, *rps7*, and *ndhB*) were surveyed.

during alignment (e.g., Thorne et al. 1992; Gu and Li 1995; Benson 1997; Kelchner 2000).

For example, slippage processes during DNA replication or repair can result in the addition or deletion of short spans of sequence that repeat one side of the region flanking the indel (reviewed in Levinson and Gutman 1987). These tandem repeat regions can also expand, contract, and diversify among lineages in different ways (Benson and Dong 1999; Zhu et al. 2000). Alternatively, insertion or deletion may involve sequence that bears no obvious relationship to flanking regions (Golenberg et al. 1993; Hoot and Douglas 1998). These indels probably result from different mutational processes than those responsible for tandem repeats, such as deletion of loop regions of DNA secondary structure (e.g., Vom Stein and Hachtel 1988; Kelchner and Clark 1997). Indels that can be identified as resulting from tandem repeats should, therefore, be dealt with separately during alignment (Benson 1997). Finally, no existing alignment algorithm can detect or deal with the small DNA inversions that may be relatively common features of chloroplast genome evolution (Kelchner and Wendel 1996; Graham and Olmstead 2000a). Inversions and tandem repeats can cause problems during alignment and subsequent phylogenetic analysis. They can lead to erroneous nucleotide homology assessment across the entire affected region and may arise repeatedly through persistent “mutational triggers” (Kelchner and Clark 1997) both in the primary sequence or in the DNA secondary structure (Thorne et al. 1992; Morton and Clegg 1993; Graham and Olmstead 2000a; Kelchner 2000).

Empirical studies of noncoding DNA sequence variation are needed to provide insights into the type and frequencies of evolutionary processes that generate indels and other microstructural changes (Gu and Li 1995). Several recent studies supply basic quantitative data for relatively recently diverged plant taxa (e.g., Olmstead and Reeves 1995; Downie et al. 1996, 1998; Hoot and Douglas 1998; Olmstead et al. 2000), but in general there has been no systematic attempt to collect and characterize information on indels in noncoding regions. We attempt to fill part of that void here.

In addition to being essential for the construction of sequence alignments and subsequent assignment of nucleotide homology in noncoding regions, gaps representing indels can themselves provide information on phylogenetic relationships. Some recent studies suggest that indels in noncoding regions and relatively rapidly evolving coding regions of organellar genomes can act as valuable markers of deep evolutionary splits (Hilu and Alice 1999; Qiu et al. 1999). Plant systematists

generally recognize the potential of chloroplast indels as phylogenetic markers, a consequence of their low rate of occurrence compared to nucleotide substitutions (e.g., Downie et al. 1996, 1998; Kelchner and Clark 1997; Sang et al. 1997; Hoot and Douglas 1998; Small et al. 1998; Zhang 2000; but see Golenberg et al. 1993; Geilly and Taberlet 1994). This is reflected in the very low levels of homoplasy of noncoding chloroplast indels (e.g., Van Ham et al. 1994; Johnson and Soltis 1995). We were particularly interested in the utility of indels in marking basal splits among the extant angiosperms and in estimating how large a sampling of the chloroplast genome may be required to generate sufficiently numerous markers to address these deep events in flowering plant evolution.

We present here a database of very slowly evolving noncoding DNA from three introns and two intergenic spacers in the inverted repeat (IR) region of the chloroplast genome. Although divergence among the basal angiosperms is ancient, these noncoding regions were generally straightforward to align across all angiosperms examined. They are extremely slowly evolving (Downie et al. 1996; table 3), as is the rest of the IR region (Wolfe et al. 1987). Indeed, in our experience, noncoding IR regions evolve more slowly at the nucleotide level than conservatively evolving photosystem genes located in the single-copy regions of the chloroplast genome (S. W. Graham and R. G. Olmstead, unpublished data). We use this survey to characterize and quantify microstructural changes across a broad and phylogenetically deep range of angiosperm taxa, to examine some problematic issues concerned with the coding and characterization of indel characters in phylogenetic analysis, and to assess the utility of microstructural changes as markers of deep angiosperm phylogeny.

## Material and Methods

### *Regions and Taxa Sampled*

Methods of DNA isolation, PCR amplification, sequencing, and alignment are outlined in Graham and Olmstead (2000a, 2000b) and S. W. Graham, P. A. Reeves, A. C. E. Burns, and R. G. Olmstead (unpublished manuscript). Source and GenBank accession information are provided there for the 40 taxa (including 31 angiosperms) considered in this study. A chloroplast-DNA-based maximum parsimony tree was inferred using PAUP\*4, beta version 4.0b3 (Swofford 2000), for the 40 taxa (“rooted” analysis) or 31 angiosperms (“unrooted” analysis), from 17 chloroplast genes and a variety of

noncoding regions from the inverted repeat region. In total, ca. 13.7 kb (unaligned) of coding and intron sequence and 209 microstructural characters (described below) were considered per taxon, using the basic tree search settings employed in Graham and Olmstead (2000b).

These analyses included ca. 2.2 kb of noncoding sequence derived from three introns (fig. 1; unaligned sequence in *Nicotiana tabacum*: 666 bp from the *rpl2* intron, 536 bp from the 3'-*rps12* intron, 679 bp from the *ndhB* intron; see Graham et al. 2000b for details) and two short intergenic spacers (unaligned sequence in *N. tabacum*: 53 bp from the 3'-*rps12-rps7* spacer region and 332 bp from the *rps7-ndhB* spacer region). A minor difference from previously published analyses was the inclusion here of these two intergenic spacers and the exclusion of two short regions in the *rpl2* intron that align ambiguously in outgroup taxa. The lengths of noncoding DNA in these regions are generally well conserved across the seed plants (Graham and Olmstead 2000b). A mid-sized inversion (ca. 200 bp) affecting three taxa in the *rps7-ndhB* intergenic region was reinverted and included in analyses for the affected taxa (see Graham and Olmstead 2000a), thereby enabling nucleotide substitutions in this region to be scored for taxa having the inversion. Alignments are available from S. W. Graham. The noncoding regions include all those shown in figure 1. Indels and inversions were scored as described below and included as single binary (0 vs. 1) characters appended to the main matrix (see Swofford 1993). We arbitrarily scored taxa 1 if sequence was present in the region of a gap, but all microstructural changes were treated as unordered binary characters. Gap regions were treated as missing data for the taxa lacking nucleotides. A regression analysis of the relationship between frequency and length of alignment gaps was undertaken with both variables log-transformed (Gu and Li 1995; regressions performed using JMP version 4.0.1, SAS Institute 2000).

#### Alignment, and Indel Scoring and Characterization

Computer-generated alignments were obtained for each major region with Clustal W (Thompson et al. 1994), using default settings. These alignments were adjusted manually using Se-Al version 1.0 alpha 1 (Rambaut 1998) and were handled as described in Graham and Olmstead (2000b). The manual adjustments were performed using several rules of thumb outlined in Kelchner and Clark (1997), Hoot and Douglas (1998), and Simmons and Ochoterena (2000), with some additions noted here. We did not attempt to consider predicted secondary structures for placing gaps (cf. Kelchner and Clark 1997), for the reasons outlined in the "Discussion" section. Because it was often harder to score indels in outgroup taxa and our focus was on indel evolution in the angiosperms, we ignored unique (autapomorphic) or ambiguous indels in noncoding regions for the outgroups (see Graham and Olmstead 2000b). Some authors (e.g., Qiu et al. 1999) assume that indels of different length that occur in the same position are homologous. We agree with Simmons and Ochoterena (2000) and others that regions with gaps of different length must involve some nonhomologous indel events (independent indels in each taxon or an additional indel in one or more taxa). Scoring and characterizing indels that overlap can be difficult (Zhang 2000;

#### a) Gaps sharing a single common terminus

```
Taxon A  GATT--①-----TAGTCGTGACTG
Taxon B  GATTAGCT---②---TAGTCGTGACTG
Others   GATTAGCTAGTGGCTAGTCGTGACTG
```

```
Taxon A  GATT--①|-----TAGTCGTGACTG
Taxon B  GATTAGCT---②---TAGTCGTGACTG
Others   GATTAGCTAGTGGCTAGTCGTGACTG
```

```
Taxon A  GATT-----TAGTCGTGACTG
Taxon B  GATTAGCT---②---TAGTCGTGACTG
Others   GATTAGCTAGTGGCTAGTCGTGACTG
          ①
```

#### b) One gap completely nested in another

```
Taxon A  GATT--①-----GTGACTG
Taxon B  GATTAGCT---②---TAGTCGTGACTG
Others   GATTAGCTAGTGGCTAGTCGTGACTG
```

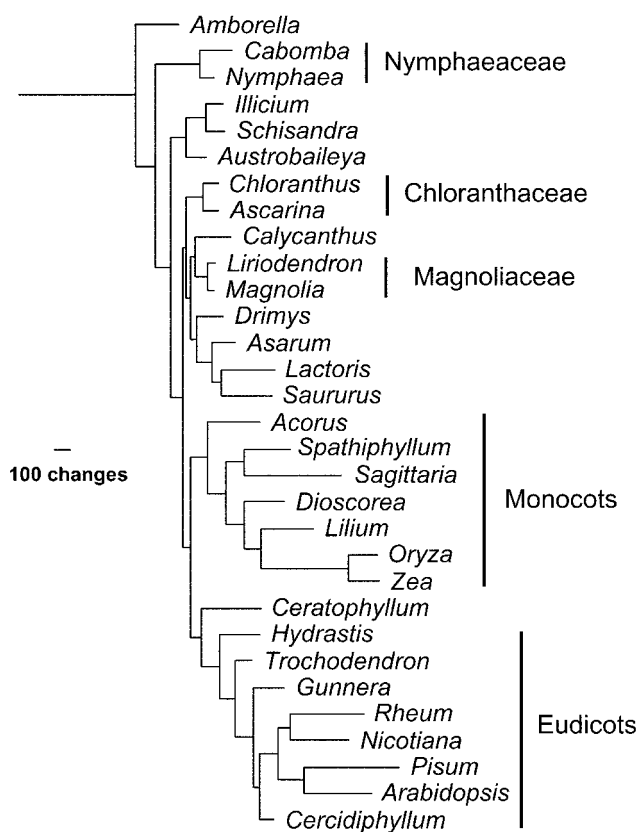
```
Taxon A  GATT--①|-----GTGACTG
Taxon B  GATTAGCT---②---TAGTCGTGACTG
Others   GATTAGCTAGTGGCTAGTCGTGACTG
```

#### c) Overlapping gaps with no shared termini

```
Taxon A  GATT--①---GGCTAGTCGTGACTG
Taxon B  GATTAGCT---②---TAGTCGTGACTG
Others   GATTAGCTAGTGGCTAGTCGTGACTG
```

Binary coding for gaps		
Type		① ②
(a)	Taxon A	0 ?
	Taxon B	1 0
	Others	1 1
(b)	Taxon A	0 ?
	Taxon B	1 0
	Others	1 1
(c)	Taxon A	0 1
	Taxon B	1 0
	Others	1 1

**Fig. 2** Scoring of indel events for regions with two overlapping gaps (1, 2). Gap 1 is found in taxon A only. The third sequence represents a majority-rule consensus for other taxa, including those more basal than taxon A or B. Scenarios that explain the observed patterns using the minimum number of indel events (two) are shown, with corresponding binary indel codings. Question marks indicate uncertainty in how to score the taxon with the overlapping gap among the different reconstructions (types a and b). a, Two gaps sharing a single endpoint. Three scenarios involving two indel events are shown. The third (bracketed) is much less plausible (see text). b, One gap completely nested in another. Two scenarios involving two indel events are shown. c, Two overlapping gaps with no shared edges. A single scenario involving two indel events is shown.



**Fig. 3** Portion of the single most parsimonious tree found using 17 chloroplast genes (*atpB*, *ndhB*, *ndhF*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbJ*, *psbL*, *psbN*, *psbT*, *rbcL*, *rpl2*, *3'-rps12*, *rps7*), including the noncoding regions shown in fig. 1. Indels were excluded from the analysis, but the same tree was found when they were included. Outgroups (*Marchantia polymorpha*, *Psilotum nudum*, *Ephedra nevadensis*, *Gnetum gnemon*, *Welwitschia mirabilis*, *Pinus thunbergii*, *Sciadopitys verticillata*, *Ginkgo biloba*, and *Zamia furfuracea*) were included in the analysis but excluded from the figure to emphasize the short branches near the base of the angiosperms. The complete tree can be found in S. W. Graham and R. G. Olmstead (unpublished manuscript). Branch lengths are calculated using ACCTRAN optimization. Length = 19,616 steps; CI = 0.476, CI\* = 0.393, RI = 0.499. The scale bar indicates 100 steps.

and see fig. 10 and “Discussion”) but may be facilitated by some prior phylogenetic information (see rule 6 below). Our basic rules for manual alignment and indel characterization are as follows:

1. We code each entire gap as a single binary character in phylogenetic analysis, regardless of length (see, e.g., Kelchner 2000 and Simmons and Ochoterena 2000 for more complete justifications).
2. Alignments were retained that maximized the matching nucleotides at sequence positions in the vicinity of a gap and minimized the total number of indel events across taxa and sequences.
3. Indels of equal length shared by more than one taxon were scored as homologous, unless they involved highly dissimilar sequences (e.g., in several cases indels of equivalent

length and position clearly represented different tandem repeats derived from each flanking region).

4. If the indel could be clearly identified as having arisen from a duplicated motif (either as an insertion, resulting in a duplicated motif, or as the deletion of one duplicate region present in other sequences), that information was used to position the indel.

5. In cases where multiple equally likely positions of an indel were possible, within or among taxa, we arbitrarily chose one position across all affected taxa (see also Gatesy et al. 1993; Davis et al. 1998; and Simmons and Ochoterena 2000).

6. When gaps of different length overlap in different taxa, the following approach was taken, described using the examples shown in figure 2. The minimum number of indel events is two, so only these scenarios are considered. The third sequence in each example represents a hypothetical consensus of sequence from other taxa in the study, including those more basal than the first two sequences. This provides limited knowledge of relationships and can aid in inferring the type and size of indel events. In practice, a majority-rule consensus is not guaranteed to represent the ancestral condition immediately prior to any particular indel event, and care should be exercised in using this approach (see fig. 10 for examples of how assumptions about plesiomorphic conditions can affect indel characterization). Boxes represent inferred indel events. The inferred binary codings are shown for each type of overlap. A question mark was used to account for uncertainty in scoring states in overlapping indels, when there were multiple possible scenarios involving the same number of steps (fig. 2a, 2b).

There are multiple ways in which two indels could have resulted in the observed pattern of two gapped regions sharing a single endpoint (a 5'- or 3'-terminus in common between gapped regions; fig. 2a). In the example shown, the plesiomorphic (primitive) state is inferred from the consensus sequence that includes the basal taxa. Only three possible pairs

**Table 1**

**Homoplasy Indices within the Angiosperms for Different Data Partitions**

Data partition	CI	CI*	RI
SC protein-coding genes (nt) <sup>a</sup> .....	0.461	0.390	0.429
IR protein-coding sequences (nt) <sup>b</sup> .....	0.679	0.512	0.601
IR noncoding sequences (nt) <sup>c</sup> .....	0.748	0.560	0.620
Indels in noncoding sequences <sup>d</sup> .....	0.907	0.679	0.761

Note. The angiosperm subtree shown in figure 3 was used to estimate homoplasy. Noninformative characters were excluded for calculation of CI\*; nt = nucleotide.

<sup>a</sup> Nucleotide sequences only for 13 protein-coding genes in the single copy region of the chloroplast genome (see text).

<sup>b</sup> Nucleotide sequences only for protein-coding regions of four genes in the inverted repeat region of the chloroplast genome (fig. 1).

<sup>c</sup> Nucleotide sequences only for two intergenic spacer regions and three introns in the inverted repeat (fig. 1). The *Ginkgo/Marchantia* start codon was used for *ndhB*.

<sup>d</sup> 180 indels (36 informative) across three introns and two intergenic spacers (fig. 1; table 2).

**Table 2**  
**Character Statistics for the IR Noncoding Regions within the Angiosperms**

Statistic	<i>rpl2</i> intron		<i>3'-rps12</i> intron		IGS regions <sup>a</sup>		<i>ndhB</i> intron	
	nt	Indel	nt	Indel	nt	Indel	nt	Indel
Total <sup>b</sup> .....	666	46	536	15 <sup>c</sup>	385	69	679	50
Variable .....	131	34	66	8	87	50	112	38
Informative .....	96	10	39	6	65	13	76	7
CI .....	0.736	0.863	0.729	0.933	0.750	0.875	0.764	1.00
RI .....	0.656	0.667	0.596	0.917	0.623	0.654	0.568	1.00

Note. The angiosperm subtree shown in figure 3 was used to estimate homoplasy. All statistics are for the angiosperms only, except that the total number of indels refers to those scored across the seed plants. Noninformative characters were excluded for calculation of CI\*; nt = nucleotide.

<sup>a</sup> The IGS regions between *3'-rps12* and *rps7*, and between *rps7* and *ndhB*. The *Ginkgo/Marchantia* start codon was used for *ndhB*.

<sup>b</sup> Number of nucleotides in *Nicotiana tabacum*.

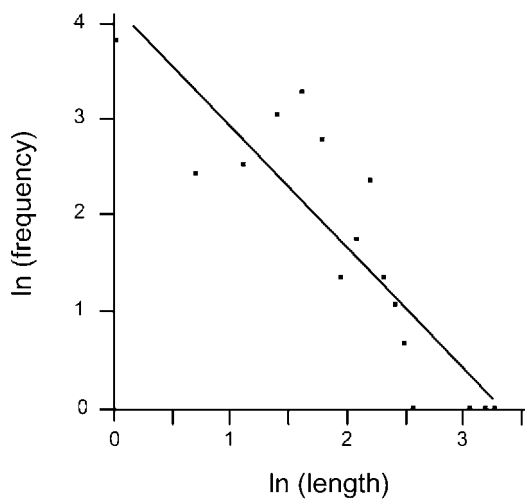
<sup>c</sup> Excludes intron loss in the branch subtending *Pisum sativum*.

of indels could generate the observed pattern, and the third (enclosed in parentheses) can be ruled out in the example shown, because of the low probability that a secondary insertion (bold box) could regenerate part of the original complex sequence (ruling the latter scenario out does not affect indel coding here, but it could affect how the indel was characterized with regard to its length or other characteristics). For the two remaining scenarios and in the absence of additional phylogenetic information, it cannot be determined whether taxon A originally shared a small deletion with taxon B and then experienced a second deletion, or whether these were two independent deletions of different length. This is reflected in the binary coding for gap 2 in taxon A, which allows for the possibility that there is a shared state.

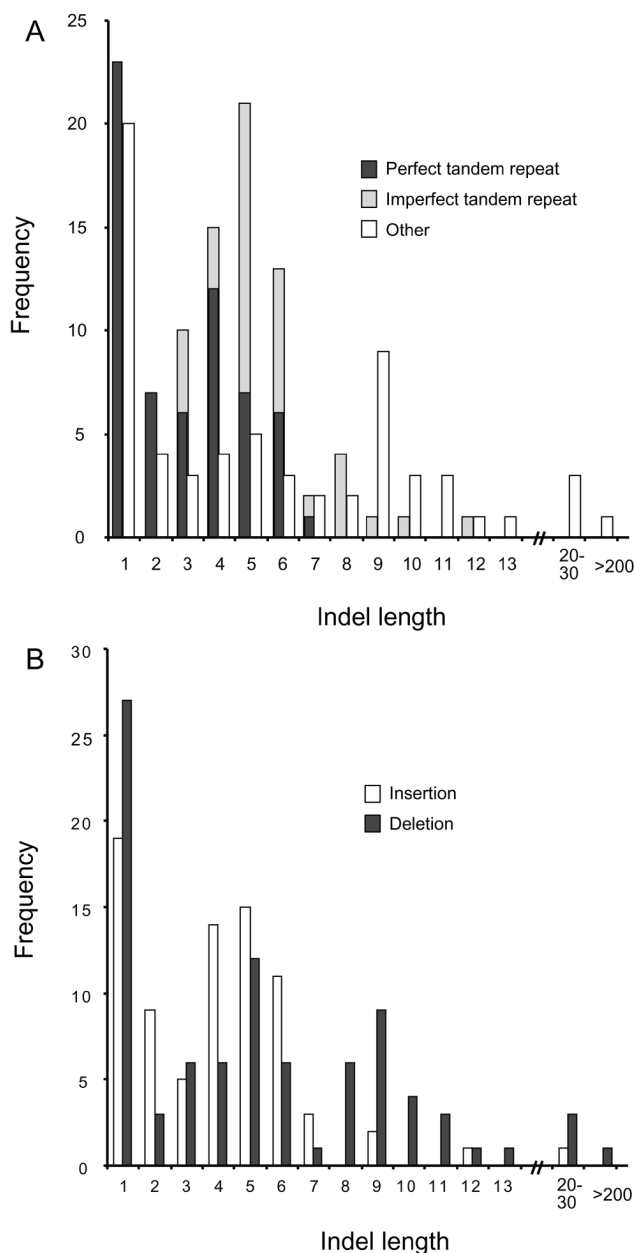
When a gap in one sequence is completely nested within a gap in another (fig. 2*b*), it also cannot be determined whether the first sequence originally shared the smaller deletion (bold

box), in the absence of additional phylogenetic information. For binary coding of these gaps, this uncertainty is reflected here in the use of “missing data” for gap 2 in taxon A. In the example shown, both are inferred to result from deletions, by comparison to the primitive condition. When gaps in two sequences overlap but have no shared endpoints (fig. 2*c*), the most likely explanation is that there were two independent (nonhomologous) indel events. However, in this case, there is no ambiguity in how to score each gap. Again, both are inferred to result from deletions, by comparison to the inferred primitive condition.

7. The length of each indel was determined, where this was not ambiguous (cf. fig. 2*a*, 2*b*). Single-base indels were cross-checked to the original chromatograms, to verify that they were not sequencing artifacts missed during base calling (*Nicotiana*, *Oryza*, *Zea*, and *Pinus* coding and noncoding sequences are from GenBank accessions and were not verified). Flanking sequences were examined in the same or closely related taxa to assess if the indel involved tandemly repeated sequence, although this could not always be determined unambiguously, particularly for overlapping gaps. Indels involving a single base or two bases were scored as tandem repeats only if they were identical to flanking sequence on either side. For longer indels, two classes of tandem repeats were scored, reflecting the possibility that nucleotide substitutions can partly obscure them after the duplication or deletion event. Perfect repeats were defined as those with 100% identity to a contiguous flanking region. Imperfect repeats were taken to be those that are longer than two bases and that share 60% or more sequence identity with a contiguous flanking region. In practice this corresponds to somewhat uneven percentage cutoffs for different classes of indel length (two out of three bases for three-base indels, three out of four bases for four-base indels). However, scoring by percentage identity is preferable to scoring on the basis of a set number of bases (Benson 1999). A 60% cutoff was used because we felt this was high enough that it would capture short tandem repeats (three to seven bases) that differ by only one or two additional nucleotide substitutions. The polarity of each indel was determined on the single tree found in the rooted analysis (see “Results”) using the Trace Character function in MacClade version 3.07 (Maddison and Maddison 1992). Each indel was characterized



**Fig. 4** Relationship between the length of different indel size classes in the noncoding regions and their frequency of occurrence (both log-transformed; see Gu and Li 1995). The distribution includes only those indels whose length was unambiguous and excludes one large indel (ca. 250 bp) and an intron that is entirely absent from the *rps12* gene of *Pisum sativum*.



**Fig. 5** Size distributions of indels in the noncoding regions. Indels are classified according to (A) whether or not they involve tandem repeats (perfect and imperfect repeats are differentiated) or (B) whether or not they are inferred to be insertions or deletions. Eighteen indels were excluded from the first distribution and 11 from the second due to ambiguity in length assessment or classification. The distributions also exclude an intron absent in the *3'-rps12* gene of *Pisum sativum*.

as an insertion or deletion or as “equivocal” if multiple parsimonious reconstructions of the indel indicated different inferred polarities.

8. Inversions can be particularly hard to spot, and if not recognized they can result in erroneous inference of one or more neighboring indels (Kelchner and Clark 1997). A hallmark of the inversions we encountered (see fig. 9) was that

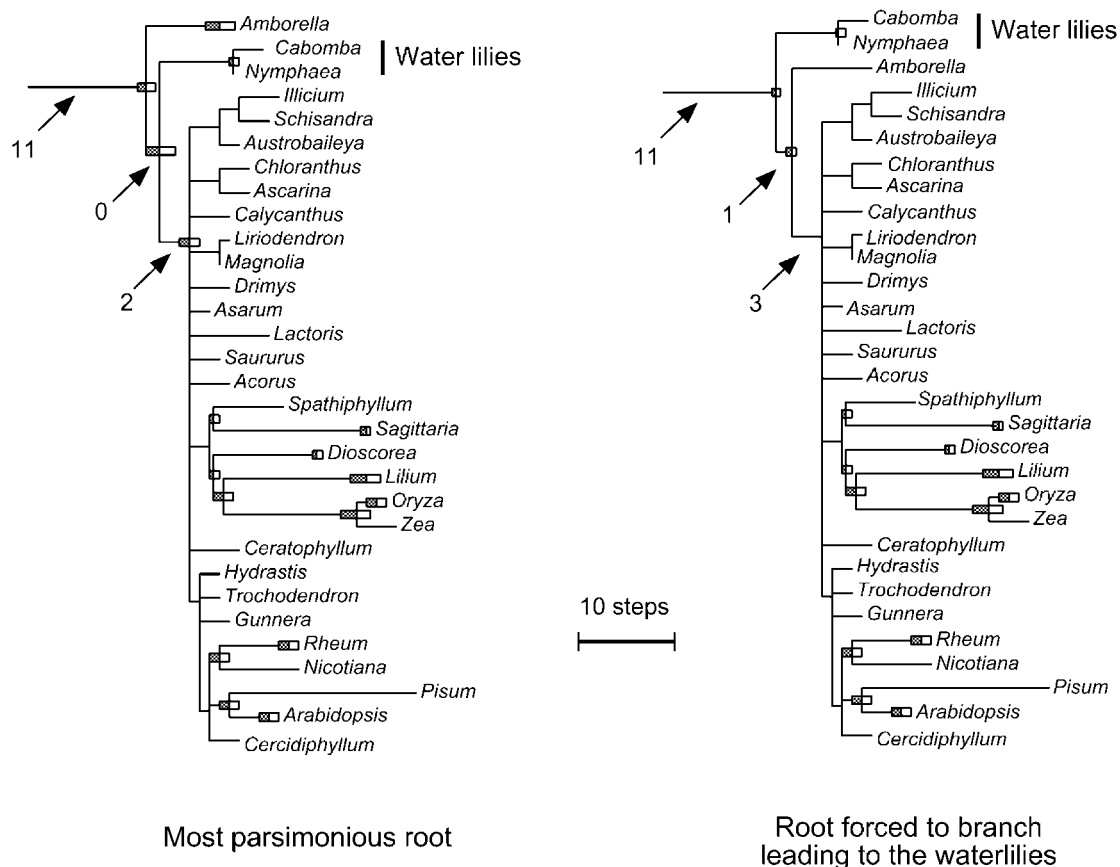
they were found in regions that were hard to align in a subset of taxa but that were otherwise well conserved. Each inversion had initially been treated as a pair of neighboring indels with opposite states (0 vs. 1). Closer examination revealed the presence of short inverted repeats flanking each region. The suspected inversions were examined for the stability of DNA secondary structure involving the inverted repeats, using the *mfold* 3.0 server (maintained by Michael Zucker, Washington University, <http://mfold.wustl.edu/~folder/dna/form1.cgi>).

## Results

Including gap characters in the matrix spanning 17 protein-coding genes and the IR noncoding regions did not affect phylogenetic inference. The same tree was inferred with or without gap characters, and the unrooted angiosperm tree (fig. 3) was the same, topologically, as the angiosperm subtree in the rooted analysis (S. W. Graham, P. A. Reeves, A. C. E. Burns, and R. G. Olmstead (unpublished manuscript), when indels were included or excluded. Indels were less homoplastic than nucleotides in any of the coding or noncoding regions examined in this study (table 1). Although some variation was evident between the noncoding regions examined, the level of homoplasy for the indels within a region was consistently lower than that seen for nucleotides (table 2). The IGS regions had the highest proportion of indels and highest homoplasy but were still inferred to have a lower probability of change and to exhibit less homoplasy than nucleotides there or elsewhere (table 1, 2). The *3'-rps12* intron was the most conservatively evolving region, with only eight variable indels and 66 nucleotide substitutions in ca. 540 bp across the broad range of angiosperms examined here. Collectively, the various noncoding gap characters had the lowest homoplasy of any class of data considered, surpassing that seen for coding or noncoding nucleotide substitutions in the IR, and substantially less homoplastic than nucleotide characters in protein-coding regions in the rest of the chloroplast genome (table 1).

Most indels were inferred to be 10 bp or shorter, and single-base indels were the most common size class. Longer size classes were generally less frequent than shorter ones (fig. 4). The regression analysis showed a significant negative relationship between natural log-transformed frequency and natural log-transformed length ( $R^2$  adjusted = 0.756,  $P < 0.0001$ ). However, there was some suggestion of a departure from the simple logarithmic relationships seen in other data sets (see Gu and Li 1995). Several size classes (lengths 2 bp, 3 bp, 7 bp) appeared to have a deficit of indel events (fig. 5), and fitting polynomial lines of degrees 2 and higher to the log-transformed data resulted in improved fit (e.g., polynomial fit of degree 2,  $R^2$  adjusted = 0.82; polynomial fit of degree 6,  $R^2$  adjusted = 0.92).

There was also evidence that different classes of indel events occurred at different frequencies (fig. 5). Of those that could be assessed as either insertions or deletions (169 indels), approximately the same number of deletions (89) and insertions (80) were observed across the noncoding regions. However, tandem repeats occurred more frequently than those that could not be related to tandem repeat sequences (fig. 5a). Among 162 indels that we classified, 98 (60%) were either perfect or imperfect tandem repeats. In addition, for those indels scorable



**Fig. 6** Microstructural changes mapped onto the single most parsimonious tree from the rooted analysis and onto the same tree but with the root node between water lilies and the remaining angiosperms (see text). Of the 209 microstructural changes considered, 180 are indels in the IR noncoding regions, one involves 3'-*rps12* intron loss in *Pisum*, three are inversions in these regions, and 25 are indels across 17 protein-coding genes (see text). Only the angiosperm subtree is shown here. Branch lengths are proportional to the average change across most parsimonious reconstructions and characters. Boxes on nodes indicate the minimum and maximum branch lengths possible across most parsimonious reconstructions and characters. Arrows indicate the number of changes involving unambiguously reconstructed microstructural characters on the branch supporting the angiosperms and several internal branches at the base of the angiosperms.

both for insertion/deletion status and whether or not they involved tandem repeats (155 indels total), tandem repeats were significantly more commonly associated with inferred insertions (78% of insertions vs. 43% of deletions;  $X^2 = 19.3$ ,  $df = 1$ ,  $P < 0.001$ ).

Single-base indels are the easiest size class to misinterpret as tandem repeats. Assuming equal base frequencies, for example, 50% of single-base nonduplicative indel events should match one or the other flanking base, by chance. However, when the 48 single-base indels were excluded from consideration (41 of which could be classified for insertion/deletion status and whether or not they involved tandem repeats), similar ratios of the various indel classifications were observed as when they were included. Of those that could be characterized, there were 62 deletions versus 61 insertions and 75 tandem repeats versus 44 that were nontandem. A total of 114 indels longer than a single base were scorable both for insertion/deletion status and whether or not they involved tandem repeats. Of these, tandem repeats were again significantly more commonly associated with inferred insertions (88% of insertions vs. 36% of dele-

tions were classified as tandem repeats;  $X^2 = 33.1$ ,  $df = 1$ ,  $P < 0.001$ ).

No substantial differences in levels of homoplasy were observed for different size classes. Indels with length 1 bp had a CI of 0.865 and an RI of 0.667 (across 48 indels total, 11 informative within the angiosperms). Homoplasy indices are not reported here for other individual size classes because most involved five or fewer informative characters. Indels ranging from 2 to 8 bp in length had a CI of 0.911 and an RI of 0.791 (102 indels total, 22 informative within the angiosperms). Only two indels of length 9 bp or longer were informative (both with no observed homoplasy) of 26 indels total.

All microstructural changes observed in this study (180 indels in the noncoding regions, 25 indels across 17 protein-coding genes, one intron loss, and three inversions) were mapped onto the shortest tree from the rooted analysis. Most nonterminal branches had no inferred support from indel synapomorphies. However, taxa supported by indel synapomorphies include the angiosperms as a whole; the angiosperms excluding the water lilies and *Amborella*; the water lilies; (*Il-*

(a)	<i>Welwitschia</i>	TTTATGATG---TCATGTTAAT	
	<i>Ephedra</i>	TTTATGATA---CCATGIGAAT	
	<i>Gnetum</i>	TTTATGATG---TCATGTGATT	
	<i>Sciadopitys</i>	TTTATGATGATACCATGTTAAT	
	<i>Pinus</i>	TTTATGATGATGCCATGIGAAT	Outgroups
	<i>Zamia</i>	TTTATGATGATGCCATGTTAAT	
	<i>Ginkgo</i>	<u>TTTATGATGATGCCATGTTAAT</u>	
	<i>Cabomba</i>	TTGATGATGATGCCATGIGAAT	
	<i>Nymphaea</i>	TTGATGATGATGCCATGIGAAT	Angiosperms
	<i>Amborella</i>	TTGATGATG---CCATGIGAAT	
	<b>Other angiosperms</b>	<b>TTGATGATG---CCATGIGAAT</b>	
(b)	<i>Pinus</i>	GATCTTATTTTCA-----TT-----GGAAC TATTA	Outgroups
	<i>Zamia</i>	GATCTTATTTCTAAAGAGATTT-----GGAAC TATTA	
	<i>Ginkgo</i>	<u>GATCTTCTTTATAAAGAGATTT-----GGAAC TATTA</u>	
	<i>Cabomba</i>	GATCTTCTTTCTAAAGAGATTC-----GGAAC TATTA	
	<i>Nymphaea</i>	GATCTTCTTTCTAAAGAGATTC-----GGAAC TATTA	Angiosperms
	<i>Amborella</i>	GATCTTCTTTCTAAAGAGATTC-----GGAAC TATTA	
	<b>Other angiosperms</b>	<b>GATCTTCTTTCTAAAGAGATTCGATTCGGAAC TATTA</b>	
(c)	<i>Welwitschia</i>	TTTATGA-----AAGAGATTTTAGA	
	<i>Ephedra</i>	TTTATGA-----AATAGAATTTAGA	
	<i>Gnetum</i>	TTTATGA-----AAGAGATTTTAGA	
	<i>Sciadopitys</i>	TTTATGA-----AATAAATTTGTGA	Outgroups
	<i>Pinus</i>	TTTATGA-----AATAGATTTCTGA	
	<i>Zamia</i>	TCTATGA-----AATGGATTATTGA	
	<i>Ginkgo</i>	<u>TTTATAA-----AATCGATTCCTGA</u>	
	<i>Cabomba</i>	CTTATAA-----AACC GATTCCTGA	
	<i>Nymphaea</i>	CTTATAA-----AACC GATTCCTGA	Angiosperms
	<i>Amborella</i>	CTTATAA-----AACTGATTCCTGA	
	<b>Other angiosperms</b>	<b>CTTATAAAAAGAAAAC T GATTCCTGA</b>	

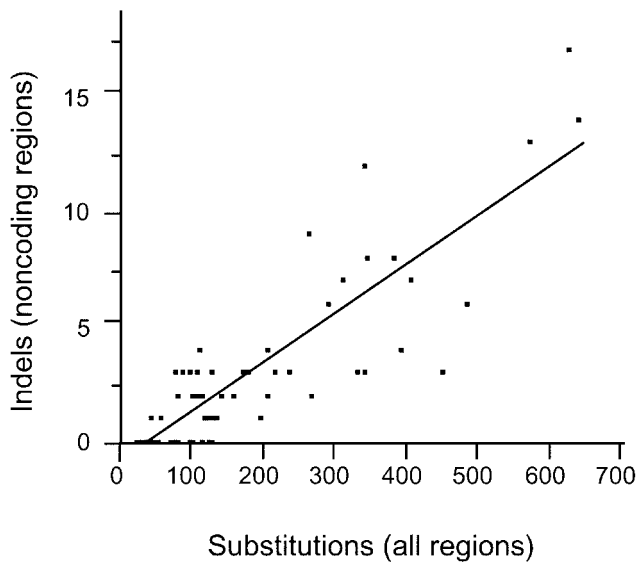
**Fig. 7** Three of the four indels that map unambiguously to the first two internal nodes at the base of the angiosperms if the root of the angiosperms is forced to the branch leading to the water lilies. The fourth (not shown) is a single-base indel shared only by *Ginkgo biloba* and *Amborella trichopoda* in an area with additional overlapping indels in *Zamia furfuracea*, the water lilies, and *Acorus calamus*. The first two (a, b) are in the 3'-*rps12* intron; the third (c) is in the intergenic region between 3'-*rps12* and *rps7*; and the fourth (not shown) is in the intergenic region between *rps7* and *ndhB*. Sequence for the region in (b) was not scored for *Psilotum nudum* and *Marchantia polymorpha*, while the three *Gnetales* taxa and *Sciadopitys verticillata* have indels spanning the region. Bold lines indicate the indel and any immediately neighboring perfect tandem repeat regions. Bold text indicates the angiosperm majority-rule consensus sequence. These "consensus" angiosperms possessed the same state for the three indels shown and virtually the same primary sequence throughout the regions shown, with the following exceptions (complete alignments are available from S. W. Graham): three angiosperms (*Pisum sativum*, *A. calamus*, and *Spathiphyllum wallisii*) each possess an additional novel indel in the region of the inferred insertion shown in c.

*licium* plus *Schisandra*); (*Illicium* plus *Schisandra* plus *Austrobaileya*); *Magnoliaceae*; *Chloranthaceae*; (*Zea* plus *Oryza*); the monocots excluding *Acorus*; the eudicots; and (*Pisum* plus *Arabidopsis*) (fig. 6). The most equivocal mapping of indels was right at the basal split of the angiosperms (fig. 6; note that boxes on nodes indicate minimum-average-maximum branch lengths across characters and most parsimonious reconstructions). No unambiguously reconstructed indel characters supported the arrangement with *Amborella* as sister to the remaining angiosperms.

If the root node of angiosperms is placed along the branch to water lilies with all other tree structure held constant, the ambiguity in reconstruction of indel evolution is reduced. A single indel with unambiguous reconstruction (a deletion in all angiosperms except the water lilies; figs. 6, 7a) supported this arrangement. Three of the four indel events that map to the first two branches in the angiosperms are shown in figure 7. If *Amborella trichopoda* is the sister group of the living angiosperms, as shown on the most parsimonious tree (fig. 3), the aforementioned indel (fig. 7a) has an equivocal optimization, with a CI of 0.33, the lowest of all indels examined.

However, if the water lilies are at the base, this indel can be inferred as a deletion (a secondary loss) within the angiosperms (CI = 0.50). Three indels (two are shown; fig. 7b, 7c) supported the branch that separates the majority of angiosperms from *Amborella* and the water lilies.

There was a significant relationship between the number of indels and the branch length inferred from all the available chloroplast evidence (nucleotide data from 17 genes and non-coding regions) across the 59 branches of the angiosperm subtree (fig. 8). Many of the branches at the base of the angiosperms are short relative to many of the terminal branches, on the order of 100 steps or less (fig. 3). At most, only a few indels per branch were observed for many of these short branches (figs. 6, 8). A regression was performed to estimate the relationships between branch length based on all the data and the number of indels inferred to fall on that branch. The line of best fit has the form  $y = 0.0211x - 0.912$  ( $R^2$  adjusted = 0.722,  $P < 0.0001$ ), and so many branches of the order of 100 steps should have an indel or two supporting them. However, the regression between the number of indels and branch length was barely significant for the shortest



**Fig. 8** Relationship between the number of indels per branch and the branch length inferred from the number of substitutions across 17 chloroplast genes, on the 59 branches in the angiosperm subtree shown in fig. 3.

branches ( $R^2$  adjusted = 0.134,  $P = 0.0286$  across the 29 shortest branches), suggesting that there is greater stochasticity in the observed indel support for shorter branches or less power to detect the relationship using the available evidence.

### Discussion

Noncoding regions make up a large proportion of nuclear genomes and a small but significant portion of the chloroplast genome. As genomic data emerge at an ever increasing rate and as the use of noncoding regions becomes widespread among systematists, it will become increasingly important to get a firm grasp on the evolutionary phenomena that mold these regions. From the perspective of plant phylogenetic studies, there are two main reasons for cataloging the rates and types of change in noncoding chloroplast DNA: (i) to provide more accurate criteria and parameters for alignment algorithms and (ii) to use indels as markers of history for the deepest branches of major clades, such as the flowering plants.

Placing gaps in an alignment can be made difficult by limited a priori knowledge of the processes and histories that generated the observed patterns of nucleotide variation, particularly when multiple events occurred in the same general region. Some alignment programs attempt to factor in the effect of history. Clustal W (Thompson et al. 1994), for example, uses a rough estimate of phylogeny, in the form of a distance-based tree, as an alignment aid. However, alignment algorithms may be biased by errors in the preliminary phylogeny estimate and are also currently limited in how they take into account the types and parameters of the molecular processes known to generate length variation. Such limitations mean that post hoc manual adjustment of computer-generated alignments is generally necessary to produce a satisfying alignment. Several factors that may contribute to the satisfaction of the human ob-

server include (i) observation and correction for inversions (Kelchner and Wendel 1996) and (ii) flexibility in gap penalties depending on local sequence context (see Thorne et al. 1992), such as the presence of tandem repeats or secondary structure.

Considering secondary structure during alignment has proven valuable for ribosomal RNA gene sequences (Kjer 1995; Hickson et al. 1996). However, there are several barriers that stand in the way of using this kind of information for aligning intron sequences. A major one concerns the complexity of the problem. Secondary structure at the nucleic acid level is usually based on predicted foldings, but in some cases comparative evidence has also aided in the recognition of conserved regions with functional importance. Chloroplast group 2 introns (such as the three considered in this study) are generally considered to be highly conserved in size and structure (Michel et al. 1989; Learn et al. 1992). However, comparative studies across introns and taxa also demonstrate fairly extensive variation in the length and presence of stems and even in the composition of major domains (Michel et al. 1989; Learn et al. 1992). Predicted foldings can vary across slightly suboptimal foldings (S. W. Graham, unpublished data), are sensitive to the length of sequence considered, and may vary across taxa and classes of intron. Second, even for a single taxon, the determination of intron secondary structure and its application to alignment is laborious and in need of automation. Finally, although secondary structure may be moderately conserved and predictable with some degree of confidence, the general principle of using it as a template to guide alignment may be questionable (Hancock and Vogler 2000). For variable domains of small subunit rRNA, Hancock and Vogler (2000) found that the inferred indels across taxa had minimal impact on predicted secondary structure, even in predicted stems. If a substantial proportion of indels in chloroplast group 2 introns have a similar effect, the determination of secondary structure may not provide substantial improvements in alignment.

Intergenic spacers may be substantially less constrained than introns in their rates and modes of evolution (although not completely so; Graham and Olmstead 2000a provide evidence for the selective maintenance of nonrandom secondary structure in the *rps7-ndhB* intergenic spacer region). They also do not appear to have any major conserved elements between chloroplast regions, apart from a tendency to have short inverted repeats flanking the 3'-ends of transcription units (Stern and Gruissem 1987). Secondary structural information may therefore be of no utility in aligning these regions. They may even prove to be positively misleading, for example, where persistent short inverted repeats are associated with highly homoplastic inversions (Kelchner 2000).

Various kinds of ascertainment bias (Zhu et al. 2000) may result in misleading characterizations of some of the indels. Noncontiguous tandem repeat sequences can generate some indels (fig. 5 in Levinson and Gutman 1987). Following slipped strand mispairing, one repeat sequence (plus intervening unique sequence) is deleted during the repair process (fig. 5 in Levinson and Gutman 1987); thus, it may be particularly hard to identify these following a deletion. Furthermore, the original noncontiguous repeat may not be readily apparent by inspection of the region in related taxa because the tandem repeats may be short and not in close proximity. However, when a

**Table 3**  
**Small to Mid-sized Inversions in Noncoding Chloroplast DNA**

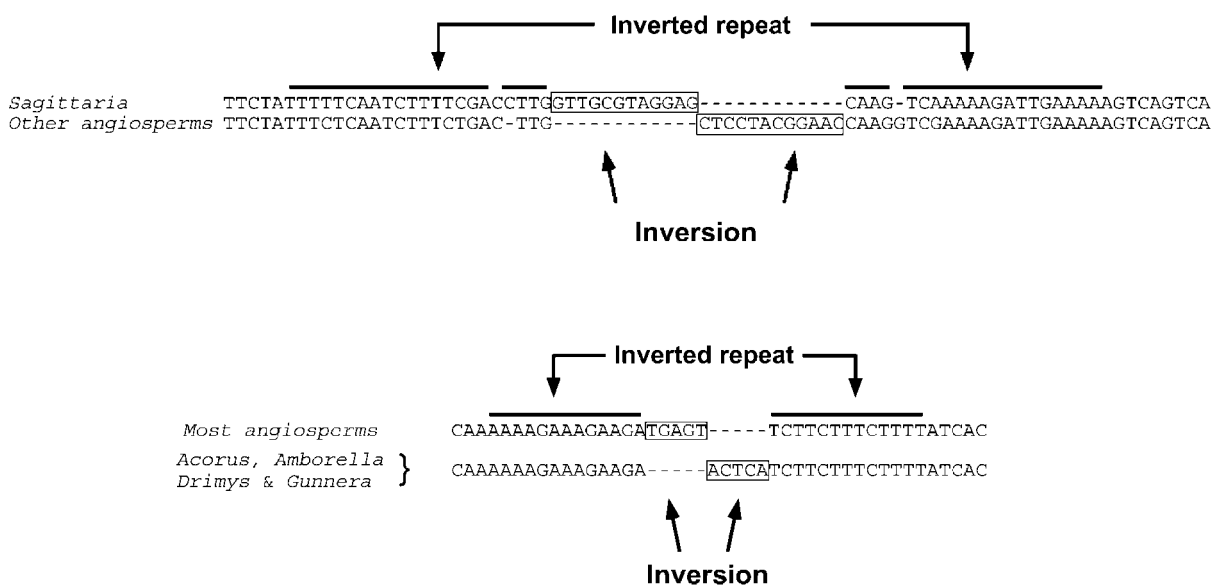
Taxon	Region	Length (bp)		Author(s)
		Inversion	Inverted repeat	
<i>Pennisetum glaucum</i> (Poaceae) .....	<i>atpB-rbcL</i> spacer	6	9	Golenberg et al. 1993; Kelchner and Wendel 1996
Bambusoideae (Poaceae) .....	<i>rpl16</i> intron	4	11	Kelchner and Wendel 1996
<i>Paeonia</i> (Paeoniaceae) .....	<i>psaA-trnH</i> spacer	21 or 6	20 or 27	Sang et al. 1997
Phoradendreae (Viscaceae) .....	<i>trnL-trnF</i> spacer	Up to 59	Up to 43–47	Ashworth 1999
<i>Wurmbea inframediana</i> (Colchicaceae) .....	<i>trnL-trnF</i> spacer	41	13	Case 2000
Various basal angiosperms .....	<i>rps7-ndhB</i> spacer	Ca. 200	14	Graham and Olmstead 2000a
<i>Fagopyrum</i> (Polygonaceae) .....	<i>trnK</i> intron	31–72	19	Ohsako and Ohnishi 2000
Various basal angiosperms .....	<i>trnC-rpoB</i> spacer	175–185	7–10	...
Various basal angiosperms .....	<i>ndhB</i> intron	5	13	This study
<i>Sagittaria latifolia</i> (Alismataceae) .....	<i>ndhB</i> intron	12	21–22	...

slipped strand mispair between noncontiguous tandem repeats is repaired such that there is duplication of intervening unique sequence plus a tandem unit, this results in readily scorable tandem repeats. The effect of this may be an asymmetry in how the insertions and deletions that result from noncontiguous tandem repeats are scored. This may explain, at least in part, our finding that significantly more tandem repeats are associated with insertions than with deletions. By the same reasoning, it would also mean that the observed overall bias in indels due to tandem repeats is an underestimate.

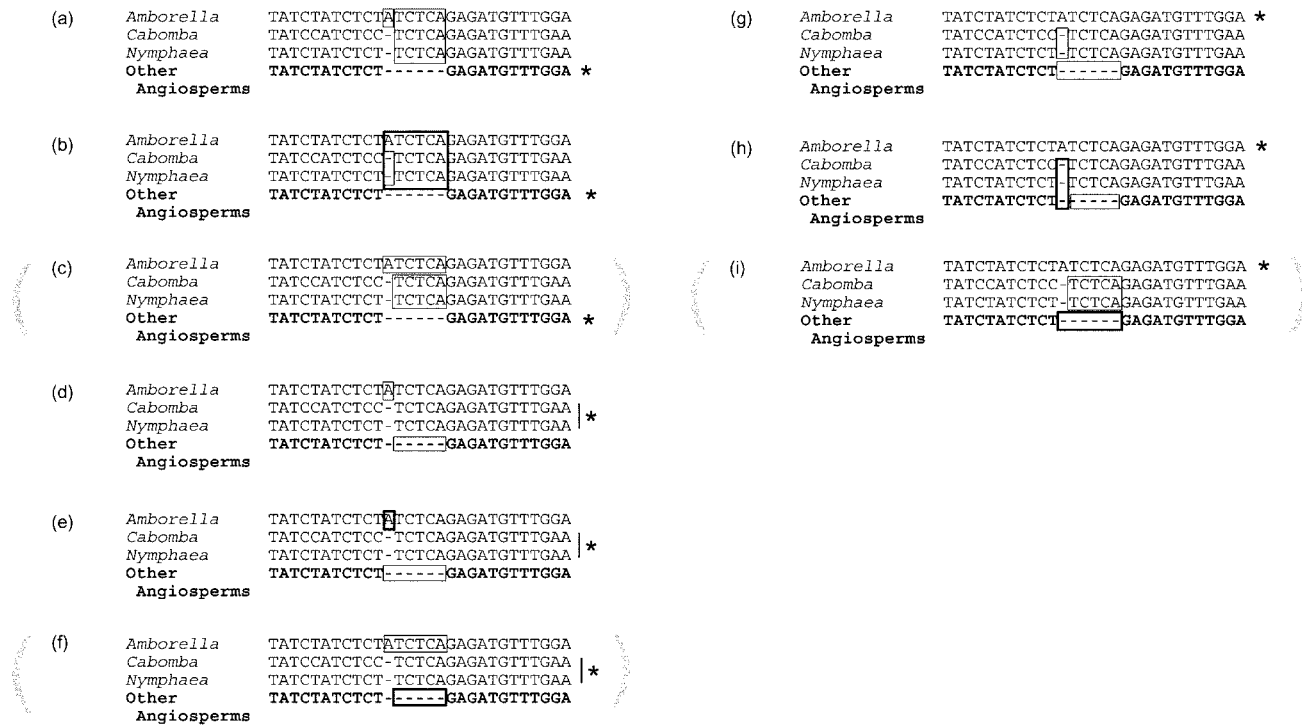
Another type of ascertainment bias derives from the possibility that tandem repeats may not be recognizable if sufficient nucleotide substitutions (or additional indels) after the duplication event act to camouflage them from visual inspection. This is unlikely for our data set because of the very low levels

of mutational change in these regions within the angiosperms (table 2). Nonetheless, we tried to take account of the former possibility by scoring both perfect and imperfect tandem repeats. A final type of ascertainment bias on estimates of indel size may result from undetected multiple indel hits. This may contribute, in part, to the deficit of several size classes (figs. 4, 5; and see “Results”). However, the total number of indel characters across the ca. 2.2 kb examined was relatively low (180 plus a few excluded indels across the noncoding regions) and the inferred average indel length was small (4.67 bp excluding the large *Pisum* deletions). Even assuming some unevenness in how the indels are distributed throughout these regions, it seems unlikely that multiple hits are responsible for a large fraction of these deficits.

Deep internal branches on the angiosperm subtree include



**Fig. 9** Two inversions inferred in the *ndhB* intron. Short bordering inverted repeat sequences are present in both cases. The first inversion is found only in *Sagittaria latifolia* (of taxa sampled). A stem loop structure derived from the inverted repeat has a predicted free energy ( $\Delta G$ ) of  $-13.15 \text{ kcal mol}^{-1}$ . The second inversion is seen in four unrelated angiosperms. The region including the inversion and short inverted repeat has a  $\Delta G$  of  $-11.09 \text{ kcal mol}^{-1}$ . The inversion is inferred to have occurred four times on the angiosperm subtree shown in fig. 3 (CI = 0.25), with a single inversion event inferred along the terminal branches leading to each of these four taxa.



**Fig. 10** Overlapping indels in part of the *ndhB* intron, with multiple alternative evolutionary explanations. *Amborella trichopoda* and the water lilies are at or near the root node of the angiosperms (the precise branching order is unclear; see text), and outgroup sequences for this region were not scored (*Psilotum nudum*; *Marchantia polymorpha*), unscorable (*Pinus thunbergii* lacks a comparable region), or of unclear homology (four bases [5'-TTTA-3'] span this region in *Ginkgo biloba* and *Zamia furfuracea*). Nine different scenarios (a – i) are shown that use the smallest possible number of indel events (two) resulting in this pattern of sequences. Asterisks mark presumed plesiomorphic sequences, and boxes indicate insertions or deletions (bold boxes must occur first). Part of the indel sequence likely results from tandem duplication of five or six flanking bases, with additional subsequent substitutions. Bracketed scenarios are less plausible because they require parallel substitution events subsequent to a tandem duplication.

some of the most interesting and poorly understood events in the diversification of the living groups of angiosperms. It would be particularly valuable to better define the position of, for example, Chloranthaceae, *Ceratophyllum*, and the monocots and eudicots among the basal angiosperms, since no current study resolves their position with a high degree of support. It is also not clear if *Amborella* is the sole candidate for being at the basalmost split in the living angiosperms (Graham and Olmstead 2000b; cf. Mathews and Donoghue 1999; Parkinson et al. 1999; Qiu et al. 1999; Soltis et al. 1999). It would be valuable to obtain indels to help address these and other aspects of basal angiosperm relationships, given the very low levels of homoplasy of these microstructural characters. To increase the probability of finding indels on these short internal branches, the amount of noncoding sequence examined would have to be increased substantially. We suggest that doubling or tripling the amount of noncoding IR sequence should provide us with indel support for most of these branches. This is based on the length of some of these deep internal branches (roughly 30–100 steps; fig. 3) and the observed relationship between branch length and number of observed indels (roughly one to two indels per hundred steps on the 17-gene tree; fig. 8; and see “Results”). Candidate regions to survey for more indels include other IR introns with equally low substitution

rates (*trnA*-UGC and *trnI*-GAU; Downie et al. 1996) and other intergenic spacer regions in the IR (see, e.g., fig. 2.1 in Downie and Palmer 1992).

An increasing number of small to midsized inversions have been recognized in chloroplast intron and intergenic spacer regions (table 3). In all cases the inversions are bordered by short (ca. 10–50 bp) inverted repeat regions. Stem loop formation at the inverted repeats may facilitate the inversion process, although a variety of mechanisms for this have been postulated (Kelchner and Wendel 1996; Graham and Olmstead 2000a). We report two additional inversions in the *ndhB* intron here (fig. 9; table 3). One is restricted to *Sagittaria latifolia* (of taxa sampled) and is bordered by a nearly perfect 21–22 bp inverted repeat. The other is bordered by a 13-bp inverted repeat region. The most remarkable thing about the latter inversion is that it is found in four distantly related angiosperm lineages, including *Amborella*, and thus has the highest homoplasy (CI = 0.25) of any microstructural change considered in this study. The high homoplasy of short inversions increases their potential for causing erroneous phylogenetic inference, particularly if they are associated with persistent features of the noncoding sequence (Kelchner and Wendel 1996; Sang et al. 1997; Kelchner 2000). This is because phylogenetic programs consider each position in the alignment as an indepen-

dent character, resulting in a series of characters all having the same pattern of homoplasy. Large inversions have been suggested to be highly reliable phylogenetic markers (e.g., Jansen and Palmer 1987; Raubeson and Jansen 1992). However, the utility of small to mid-sized inversions in phylogenetic inference may be poor because they are both few in number and more highly homoplastic.

Some knowledge of phylogeny is essential in characterizing indel parameters, such as their length. This knowledge can come from examining putative plesiomorphic sequences (fig. 2) or by mapping indels onto a phylogenetic tree, if that is already available. We used both approaches here because a primary goal of this study was to examine indel evolution in the context of what we now know about angiosperm phylogeny. However, assessment of plesiomorphic states (whether conscious or not) can affect how an indel is characterized. An example of how different assumptions of the plesiomorphic sequence condition affect indel characterization is shown in figure 10. The example involves overlapping indels between two basal angiosperm groups (*Amborella* and the water lilies), where outgroup comparisons are not possible and where the phylogenetic arrangement at the base of the angiosperms is unclear (Graham and Olmstead 2000b). Nine plausible reconstructions of the indel events that could have generated the observed pattern are shown, given some different assumptions about which sequence pattern is more plesiomorphic (three scenarios in parentheses are less plausible because they require an additional parallel substitution, subsequent to tandem duplication). The various interpretations shown also do not all have the same consequences for phylogenetic analysis. For example, interpretations *a* and *b* provide a five- or six-base indel synapomorphy linking *Amborella* and the water lilies or supporting the monophyly of the remaining angiosperms. Interpretation *c* does not support any particular arrangement at the base of the angiosperms, apart from potentially providing a five-base indel synapomorphy for the water lilies. Equivalent binary coding scenarios can be found for each of the remaining interpretations. This particular pair of indels were excluded from our study, but similar problems may be prevalent in other studies where indels are characterized.

We did not use the multistate “complex indel coding” proposed by Simmons and Ochoterena (2000, p. 375), for several reasons. Their scheme attempts to deal with partly or completely overlapping (nested) indels by coding them as a single multistate character and uses a step matrix to define the costs of interchange between the observed overlapping gap patterns.

One difficulty with this approach is that their scheme becomes quite elaborate when more than a few indels are nested within a larger deletion. Such is the case here, for example, with the inferred deletion of the 3'-*rps12* intron or the inferred loss of most of the *rps7-ndhB* intergenic spacer in *Pisum sativum*. Both deletions span a large number of smaller indels, which often in turn span more deeply nested indels. Their scheme also overlooks the fact that reversal of a nontandem indel may be difficult because it can require regeneration of substantial amounts of sequence that matches, in part or in whole, the original sequence that was deleted, prior to smaller secondary deletion events in the regenerated region (see figs. 3, 4 in Simmons and Ochoterena 2000). In figure 2a, for example, we ruled out the third scenario shown because it would require either insertion of complex nontandem sequence (that coincidentally matches the lost sequence in length and base sequence) or tandem duplication followed by multiple subsequent substitutions to match the original nucleotide states. Such regeneration is possible, but it is unlikely to result from the single simple intermediate indel event that would be assumed implicitly under their model. We therefore prefer Simmons and Ochoterena's (2000, p. 375) “simple indel coding” scheme, and that, essentially, is the scheme we used here.

Nonetheless, the difficulty of interchange occurring among some overlapping gap patterns suggests that unordered coding of the binary indel characters, and reversible costs of insertions and deletions during alignment, may not be entirely realistic. Ideally, we should probably consider the likelihood of putative intermediate “regeneration” events when making alignments or coding indels. Generating such schemes in turn depends on the completeness and reliability of our knowledge concerning the processes responsible for these patterns. That knowledge can only be derived from study of extensive sequence data.

### Acknowledgments

We thank Tandy Warnow for providing the germ of the idea for this study; Tatsuya Wakasugi, Mark Chase, and others for sharing DNA sequences and plant material; Andrea Case and Patrick Lorch for help with statistical interpretation; and Scot Kelchner for comments on the manuscript and access to his unpublished manuscript. This work was funded by NSF grant DEB 9727025.

### Literature Cited

- Ashworth VETM 1999 Phylogenetic relationships in Phoradendreae (Viscaceae) inferred from DNA sequence data. PhD diss. Claremont Graduate University.
- Benson G 1997 Sequence alignment with tandem duplication. *J Comp Biol* 4:351–367.
- 1999 Tandem repeats: a program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580.
- Benson G, L Dong 1999 Reconstructing the duplication history of a tandem repeat. Pages 44–53 in T Lengauer, R Schneider, P Bork, D Brutlag, J Glasgow, H-W Mewes, R Zimmer, eds. Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, Calif.
- Case A 2000 Evolution of combined versus separate sexes in *Wurmbia* (Colchicaceae). PhD diss. University of Toronto.
- Davis JL, MP Simmons, DW Stevenson, JF Wendel 1998 Data decisiveness, data quality and incongruence in phylogenetic analysis: an example from the monocotyledons using mitochondrial *atpA* sequences. *Syst Biol* 47:282–310.
- De Pinna MCC 1991 Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7:367–394.

- Downie SR, DS Katz-Downie, K-J Cho 1996 Phylogenetic analysis of Apiaceae subfamily Apioideae using nucleotide sequences from the chloroplast *rpoC1* intron. *Mol Phylogenet Evol* 6:1–18.
- Downie SR, JD Palmer 1992 Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. Pages 14–35 in PS Soltis, DE Soltis, JJ Doyle, eds. *Molecular systematics of plants*. Chapman & Hall, New York.
- Downie SR, S Ramanath, DS Katz-Downie, E Llanas 1998 Molecular systematics of Apiaceae subfamily Apioideae: phylogenetic analyses of nuclear ribosomal DNA internal transcribed spacer and plastid *rpoC1* intron sequences. *Am J Bot* 85:563–591.
- Gatesy J, R DeSalle, W Wheeler 1993 Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol Phylogenet Evol* 2:152–157.
- Geilly L, P Taberlet 1994 The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbcL* sequences. *Mol Biol Evol* 11:769–777.
- Giribet G, WC Wheeler 1999 On gaps. *Mol Phylogenet Evol* 13:132–143.
- Golenberg EM, MT Clegg, ML Durbin, J Doebley, DP Ma 1993 Evolution of a noncoding region of the chloroplast genome. *Mol Phylogenet Evol* 2:52–64.
- Graham SW, RG Olmstead 2000a Evolutionary significance of an unusual chloroplast DNA inversion in two basal angiosperm lineages. *Curr Genet* 37:183–188.
- 2000b Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am J Bot* (in press).
- Gu X, W-H Li 1995 The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J Mol Evol* 40:464–473.
- Hancock JM, AP Vogler 2000 How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Mol Phylogenet Evol* 14:366–374.
- Hein J 1989 A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol Biol Evol* 6:649–668.
- 1990 Unified approach to alignment and phylogenies. *Methods Enzymol* 183:626–645.
- Hickson RE, C Simon, A Cooper, GS Spicer, J Sullivan, D Penny 1996 Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Mol Biol Evol* 13:150–169.
- Hilu KW, LA Alice 1999 Evolutionary implications of *matK* indels in Poaceae. *Am J Bot* 86:1735–1741.
- Hoot SB, AW Douglas 1998 Phylogeny of the Proteaceae based on *atpB* and *atpB-rbcL* intergenic spacer region sequences. *Aust Syst Bot* 11:301–320.
- Jansen RK, JD Palmer 1987 A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc Natl Acad Sci USA* 84:5818–5822.
- Johnson LA, DE Soltis 1995 Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *matK* sequences. *Ann Mo Bot Gard* 82:149–175.
- Kelchner SA 2000 The evolution of noncoding chloroplast DNA and its application in plant systematics. *Ann Mo Bot Gard* (in press).
- Kelchner SA, LG 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 SA, JF Wendel 1996 Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Curr Genet* 30:259–262.
- Kjer KM 1995 Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Mol Phylogenet Evol* 4:314–330.
- Learn GH Jr, JS Shore, GR Fournier, G Zurawski, MT Clegg 1992 Constraints on the evolution of plastid introns in the gene encoding tRNA-Val(UAC). *Mol Biol Evol* 9:856–871.
- Levinson G, GA Gutman 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203–221.
- Maddison WP, DR Maddison 1992 MacClade 3.0: analysis of phylogeny and character evolution. Sinauer, Sunderland, Mass.
- Mathews S, MJ Donoghue 1999 The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286:947–950.
- Michel F, K Umeson, H Ozeki 1989 Comparative and functional anatomy of group II catalytic introns: a review. *Gene* 82:5–30.
- Morton BR, MT Clegg 1993 A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbcL* in the grass family (Poaceae). *Curr Genet* 24:357–365.
- Ohsako T, O Ohnishi 2000 Intra- and interspecific phylogeny of wild *Fagopyrum* (Polygonaceae) species based on nucleotide sequences of noncoding regions in chloroplast DNA. *Am J Bot* 87:573–582.
- Olmstead RG, RK Jansen, K-J Kim, SJ Wagstaff 2000 The phylogeny of the Asteridae s.l. based on chloroplast *ndbF* sequences. *Mol Phylogenet Evol* 16:96–112.
- Olmstead RG, PA Reeves 1995 Evidence for the polyphyly of the Scrophulariaceae based on chloroplast *rbcL* and *ndbF* sequences. *Ann Mo Bot Gard* 82:176–193.
- Parkinson CL, KL Adams, JD Palmer 1999 Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr Biol* 9:1485–1488.
- Qiu Y-L, J Lee, F Bernasconi-Quadroni, DE Soltis, PS Soltis, M Zanis, EA Zimmer, Z Chen, V Savolainen, MW Chase 1999 The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 410:404–407.
- Rambaut A 1998 Se-Al (sequence alignment editor version 1.0 alpha 1). Department of Zoology, University of Oxford, Oxford.
- Raubeson LA, RK Jansen 1992 Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science* 255:1697–1699.
- Sang T, DJ Crawford, TF Steussy 1997 Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). *Am J Bot* 84:1120–1136.
- SAS Institute 2000 JMP version 4.0.1. Statistical Discovery Software, Cary, N.C.
- Simmons MP, H Ochoterena 2000 Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol* 49:369–381.
- Small RL, JA Ryburn, RC Cronn, T Seelanan, JF Wendel 1998 The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. *Am J Bot* 85:1301–1315.
- Soltis PE, DE Soltis, MW Chase 1999 Angiosperm phylogeny inferred from multiple chloroplast genes as a tool for comparative biology. *Nature* 402:402–404.
- Stern DB, W Gruissem 1987 Control of plastid gene expression: 3' inverted repeats act as mRNA processing and stabilizing elements, but do not terminate transcription. *Cell* 51:1145–1157.
- Swofford DL 1993 PAUP: phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign.
- 2000 PAUP\*: phylogenetic analysis using parsimony (\*and other methods), version 4. Sinauer, Sunderland, Mass.
- Thompson JD, DG Higgins, TJ Gibson 1994 Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
- Thorne JL, H Kishino, J Felsenstein 1991 An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol* 33:114–124.
- 1992 Inching towards reality: an improved likelihood model of sequence evolution. *J Mol Evol* 34:3–16.
- Van Ham RCHJ, H Hart, THM Mes, JM Sandbrink 1994 Molecular

- evolution of noncoding regions of the chloroplast genome in the Crassulaceae and related species. *Curr Genet* 25:558–566.
- Vom Stein J, W Hachtel 1988 Deletions/insertions, short inverted repeats, sequences resembling *att*-lambda, and frame shift mutated open reading frames are involved in chloroplast differences in the genus *Oenothera* subsection *Munzia*. *Mol Gen Genet* 213:513–518.
- Wolfe KH, W-H Li, PM Sharp 1987 Rates of nucleotide substitution vary greatly across mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci USA* 84:9054–9058.
- Zhang W 2000 Phylogeny of the grass family (Poaceae) from *rpl16* intron sequence data. *Mol Phylogenet Evol* 15:135–146.
- Zhu Y, DC Queller, JE Strassmann 2000 A phylogenetic perspective on sequence evolution in microsatellite loci. *J Mol Evol* 50: 324–338.