J Mol Evol (2004) 58:182–190
DOI: 10.1007/s00239-003-2540-4

JOURNAL OF MOLECULAR EVOLUTION

Causes of Size Homoplasy Among Chloroplast Microsatellites in Closely Related *Clusia* Species

Marie L. Hale,[1] Anne M. Borland,[1] Mats H.G. Gustafsson,[2] Kirsten Wolff[1]

[1] School of Biology, University of Newcastle-upon-Tyne, NE1 7RU, UK
[2] Department of Systematic Botany, Institute of Biological Sciences, University of Aarhus, Aarhus, Denmark

Received: 18 March 2003 / Accepted: 7 August 2003

Abstract. Chloroplast DNA sequences and microsatellites are useful tools for phylogenetic as well as population genetic analyses of plants. Chloroplast microsatellites tend to be less variable than nuclear microsatellites and therefore they may not be as powerful as nuclear microsatellites for within-species population analysis. However, chloroplast microsatellites may be useful for phylogenetic analysis between closely related taxa when more conventional loci, such as ITS or chloroplast sequence data, are not variable enough to resolve phylogenetic relationships in all clades. To determine the limits of chloroplast microsatellites as tools in phylogenetic analyses, we need to understand their evolution. Thus, we examined and compared phylogenetic relationships of species within the genus *Clusia*, using both chloroplast sequence data and variation at seven chloroplast microsatellite loci. Neither ITS nor chloroplast sequences were variable enough to resolve relationships within some sections of the genus, yet chloroplast microsatellite loci were too variable to provide any useful phylogenetic information. Size homoplasy was apparent, caused by base substitutions within the microsatellite, base substitutions in the flanking regions, indels in the flanking regions, multiple microsatellites within a fragment, and forward/reverse mutations of repeat length resulting in microsatellites of identical base composition that were not identical by descent.

Key words: *Clusia* — Chloroplast microsatellites — Size homoplasy

Introduction

Chloroplast microsatellites are emerging as very useful tools for population genetic and phylogenetic analysis within and between closely related taxa (Provan et al. 2001). Several sets of primers that amplify chloroplast microsatellite loci have recently been published, including some primers that are considered to be "universal" (e.g., Weising and Gardner 1999; Provan et al. 1999a; Grivet et al. 2001). Microsatellite loci are particularly useful for intraspecific genetic comparisons, but their high mutation rate makes them less useful for interspecific analyses. Size homoplasy, where alleles of identical size are not identical by descent, is a particular problem when analyzing interspecific microsatellite data sets (e.g., Doyle et al. 1998), yet the lower mutation rate of chloroplast microsatellites compared to nuclear microsatellites (Provan et al. 1999b) suggests that size homoplasy may not be a major problem for chloroplast microsatellite comparisons between closely related species. Chloroplast microsatellites have been used successfully for interspecific comparisons among several closely related taxa (e.g., *Hordeum* [Provan et al. 1999a], *Abies* [Parducci et al. 2001],

Correspondence to: M.L. Hale; *email:* m.l.hale@ncl.ac.uk

Triticum and *Aegilops* species [Ishii et al. 2001]) and may provide a useful tool for interspecific analyses when other markers, such as ITS and chloroplast DNA sequences, are simply not variable enough (Provan et al. 2001).

The genus *Clusia* comprises approximately 300 species of trees and shrubs divided into 14 sections, distributed throughout tropical and subtropical America (Pipoly et al. 2000). The species are predominantly dioecious and at least 85 are epiphytic stranglers with anastomizing aerial roots, found in habitats ranging from moist montaine rainforest to arid rocky coasts. In addition to a variety of morphological and ecological attributes, there is substantial physiological diversity within this genus. *Clusia* is the only genus of tropical trees known to exhibit CAM, a specialized mode of photosynthetic carbon acquisition that enables the uptake of CO_2 at night and thereby conserves water. Moreover, *Clusia* shows exceptional diversity in the range of CAM expression, from constitutive CAM to C_3–CAM intermediates to C_3 species (Borland et al. 1998). To better understand the evolution of this enormous diversity within this genus a phylogenetic study is needed. However, phylogenetic analysis of some sections in this group has been difficult due to low variation and homoplasy in both chloroplast and nuclear (ITS) DNA nucleotide substitutions. The low level of variation or homoplasy makes it particularly difficult to resolve relationships in the sections Anandrogyne and Retinostemon, respectively (Gustafsson and Bittrich 2002).

We examined seven chloroplast microsatellites to see if they are likely to be appropriate for determining phylogenetic relationships between closely related *Clusia* species. We sequenced approximately 3000 bp of chloroplast genome from four different regions in 17 *Clusia* species to compare information gained from base substitutions and indels with that from variation in number of repeats at chloroplast microsatellite loci. This large data set allowed us not only to test for the presence of size homoplasy in chloroplast microsatellites, but also to examine the causes of any size homoplasy detected. Size homoplasy can be assessed by comparing the pattern of variation at microsatellite loci with the phylogeny of the surrounding genes (Doyle et al. 1998). For chloroplast microsatellites, this can be achieved by comparison of microsatellite variation to the phylogeny based on base substitutions and indels throughout the chloroplast genome. However, in the genus *Clusia*, chloroplast variation is low, and a robust phylogeny for all sections is difficult to obtain from chloroplast sequence data. Therefore, to assess size homoplasy in *Clusia* chloroplast microsatellites we also assessed the level of linkage disequilibrium between chloroplast microsatellite loci. Because the chloroplast genome is nonrecombining, alleles at microsatellite loci within the chloroplast genome should be correlated, forming distinct haplotypes (Provan et al. 2001). Lack of linkage equilibrium suggests stepwise mutations and therewith homoplasy caused by increase and decrease of repeat numbers at these loci. Alternatively we have to assume frequent recombination in the chloroplast genome in the *Clusia* species to explain the absence of linkage disequilibrium. We also sequenced microsatellite regions, and the surrounding DNA, to determine the cause of any size homoplasies observed.

Materials and Methods

Sample Collection

Both fresh and dried leaf tissue samples were obtained from living and herbarium collections (Table 1). Seventeen *Clusia* species were examined: *C. aripoensis* ($n = 1$), *C. croatii* ($n = 1$), *C. ducu* ($n = 5$), *C. flava* ($n = 1$), *C. fluminensis* ($n = 1$), *C. grandiflora* ($n = 1$), *C. intertexta* ($n = 1$), *C. lanceolata* ($n = 1$), *C. major* ($n = 1$), *C. minor* ($n = 4$), *C.* cf. *multiflora* ($n = 3$), *C. nemorosa* ($n = 1$), *C. rosea* ($n = 1$), *C. stenophylla* ($n = 1$), *C. tocuchensis* ($n = 1$), *C. torresii* ($n = 1$), and *C. valerii* ($n = 1$). The determination of the *C.* cf. *multiflora* group is difficult, but for the present analysis they were grouped together.

DNA Extraction and Amplification

DNA was extracted using a DNeasy Plant Mini Kit (Qiagen). Four chloroplast fragments were amplified for each individual: *Trn*L intron using primers *Trn*L-c and *Trn*L-d (Taberlet et al. 1991), *Trn*L–*Trn*F intergenic spacer region using primers *Trn*L-e and *Trn*F-f (Taberlet et al. 1991), *Atp*B–*rbc*L intergenic spacer region using primers *Atp*B and *rbc*L (Hodges and Arnold 1994), and *acc*D–*psa*L intergenic spacer region using primers *acc*D-769F and *psa*L-75R (Small et al. 1998). All fragments were amplified in 25-µl reactions [1× *Taq* buffer (16 m*M* $(NH_4)_2SO_4$, 67 m*M* Tris–HCl, 0.01% Tween-20), 2.0 m*M* $MgCl_2$, 0.2 m*M* each dNTP, 0.2-µ*M* each primer, 1.0 U *Taq* (Bioline), and 0.5 µl template DNA]. The *Trn*L intron and *Trn*L–*Trn*F intergenic spacer region were amplified in 35 cycles at 93°C for 1 min, 50°C for 1 min, 72°C for 2 min. The *Atp*B–*rbc*L and *acc*D–*psa*L intergenic spacer regions were amplified as follows: 94°C for 5 min, followed by 30 cycles at 94°C for 30 s, 50°C for 30 s, 72°C for 2 min, with a final extension at 72°C for 4 min. All reactions were performed in a PTC-100 thermocycler (MJ Research).

Sequencing

PCR products were sequenced directly. All PCR products were purified using QIAquick PCR Purification Kits (Qiagen). Purified PCR products were then sequenced using BigDye Terminator Cycle Sequencing chemistry (Applied Biosystems), following the manufacturer's recommended conditions, and sequences detected on an ABI 310 Prism automated sequencer (Applied Biosystems). The *Trn*L intron was sequenced in both directions using both amplification primers. In addition, one internal sequencing primer (*Trn*L-int: TGAACTTGGGATTGATTCAAGA) was used. The internal primer was required because of sequence deterioration due to the presence of several large microsatellites within the *Trn*L intron fragment. The *Trn*L–*Trn*F intergenic spacer region was sequenced in a single direction, using the *Trn*L-e primer. The *Atp*B–

Table 1. Information on origin and herbarium location of the *Clusia* accessions

Specimen	Specimen location	Voucher number	Geographic origin
C. aripoensis	Moorbank	TR922	Trinidad
C. croatii	INBio	22394 (INB)	Costa Rica
C. ducu A	AAU	Gustafsson 348	S. Ecuador
C. ducu B	AAU	Gustafsson 349	S. Ecuador
C. ducu C	AAU	Gustafsson 360	S. Ecuador
C. ducu D	AAU	Gustafsson 363	S. Ecuador
C. *ducu* E	AAU	Gustafsson 376	S. Ecuador
C. flava	RBGE	19696452	Cultivated material
C. fluminensis	Moorbank	BR941	Brazil
C. grandiflora	RBGE	19471021	Cultivated material
C. intertexta	NHTT	TRIN31668	Trinidad
C. lanceolata	Moorbank	BR942	Brazil
C. major	Aarhus	Gustafsson 396	Martinique
C. minor A	Moorbank	TR921	Trinidad
C. minor B	NY	Gustafsson 291	Unknown
C. minor C	INBio	22418 (INB)	Costa Rica
C. minor D	AAU	Gustafsson 398	Venezuela
C. multiflora A	AAU	Gustafsson 350	S. Ecuador
C. multiflora B	AAU	Gustafsson 379	S. Ecuador
C. multiflora C	Moorbank	VZ931	Venezuela
C. nemorosa	Aarhus	—	French Guiana
C. rosea	Moorbank	TR923	Trinidad
C. stenophylla	INBio	22395 (INB)	Costa Rica
C. tocuchensis	Moorbank	TR924	Trinidad
C. torresii	INBio	22396 (INB)	Costa Rica
C. valerii	INBio	22393 (INB)	Costa Rica

Note. Moorbank is the botanical garden at the University of Newcastle, UK. INBio—Missouri Botanical Garden and Instituto Nacional de Biodiversidad, Costa Rica. AAU—Herbarium of the University of Aarhus, Denmark. RBGE—Royal Botanic Garden Edinburgh, UK. NHTT—National Herbarium of Trinidad and Tobago, West Indies. NY—Herbarium of the New York Botanical Garden. Aarhus—Botanical Garden University of Aarhus.

*rbc*L intergenic spacer was sequenced in both directions using both amplification primers, as well as with two internal primers (one in each direction). Sequencing with the two additional internal primers, *Atp*B-int (GTTCGATATCAAGTTTATCGG) and *rbc*L-int (GCTATAGGTGTAACTCAATATG), was needed to gain a complete sequence of this fragment because of sequence deterioration due to microsatellite regions. The *acc*D–*psa*L intergenic spacer region was sequenced in both directions using the two amplification primers.

Data Analysis

Sequences were aligned and edited manually using ProSequence (Filatov 2002). The sequence data set was split into two data sets: one containing all nonmicrosatellite sites (microsatellite regions simply deleted from sequences) and one containing all microsatellite regions. Sequences from all four chloroplast loci were combined and analyzed as a single locus for nonmicrosatellite DNA. Microsatellite regions were analyzed both as individual loci and as multilocus haplotypes with alleles defined by repeat length, ignoring single base substitutions within the microsatellite region. An unrooted maximum parsimony tree was constructed for nonmicrosatellite sequence data using PAUP* 4.0 (Swofford 2002), and a neighbor-joining tree based on sum of squared size differences was constructed using ARLEQUIN (Schneider et al. 2000) and PHYLIP 3.5c (Felsenstein 1993) for multilocus microsatellite haplotypes.

Linkage disequilibrium analysis, based on a likelihood-ratio test was conducted using ARLEQUIN (Schneider et al. 2000). Linkage disequilibrium among the microsatellite loci was compared to the level of linkage disequilibrium in both single-nucleotide polymorphisms (SNPs) and indels to determine whether size homoplasy exists in the microsatellite loci resulting from stepwise mutation. Because all microsatellite loci are on the same DNA molecule, we expect to see linkage disequilibrium between microsatellite regions if there is little homoplasy (Provan et al. 2001).

Results

A total of 2661 base pairs of aligned chloroplast sequence data was collected for each individual, comprising the *Trn*L intron (671 bp), *Trn*L–*Trn*F intergenic spacer (427 bp), *Atp*B–*rbc*L intergenic spacer (807 bp), and *acc*D–*psa*L intergenic spacer (756 bp) (GenBank accession numbers AY143996–AY144099). Of this, 2545 bp was nonmicrosatellite DNA and the remaining 116 bp microsatellite DNA. There were 108 polymorphic sites in the nonmicrosatellite DNA, counting each indel as a single polymorphic site; 82 of these were base substitutions, and 26 were indels ranging in size from 1 to 27 bp long. Chloroplast sequence variation occurred both within and between species and was used to construct a phylogenetic tree (Fig. 1). This tree confirmed species relationships found using rDNA ITS data (Gustafsson and Bittrich 2002) and groupings based on morphology, particulars of which will be discussed in

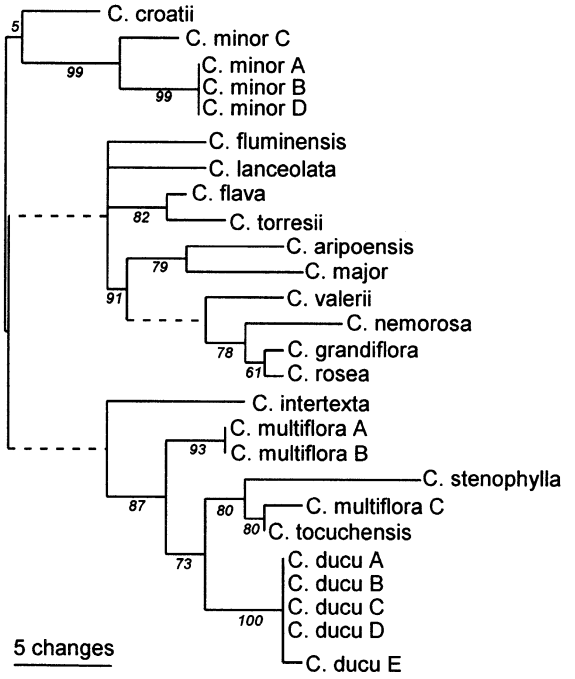

Fig. 1. Maximum parsimony tree of 17 *Clusia* species and individuals within species for 2545 bases of chloroplast DNA. Four regions were sequenced and sequences combined for the analysis: *Trn*L intron, *Trn*L–*Trn*F intergenic spacer region, *Atp*B–*rbc*L intergenic spacer region, and *acc*D–*psa*L intergenic spacer region. Microsatellites and indels have been removed from the sequences, therefore, variation represents single nucleotide polymorphisms only. Jackknife support values are indicated. Branches collapsed in the strict consensus are dashed.

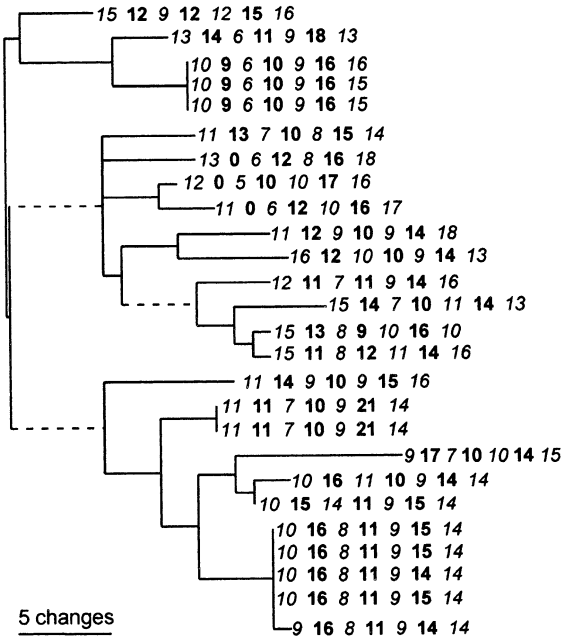

Fig. 2. Microsatellite repeat numbers for each of the seven loci have been overlaid on the maximum parsimony tree of 17 *Clusia* species and individuals within species (Fig. 1). The odd-numbered loci are in italics and the even-numbered loci in bold face to facilitate reading of the tree. For all seven microsatellite loci there is clear evidence of size homoplasy based on this phylogenetic tree. For example, 11 repeats for locus 1 appear to have evolved multiple times, as have 12 repeats.

a future publication. Within-species variation was low in both *C. minor* and *C. ducu*, with individuals within these two species closely grouped together. *C.* cf. *multiflora* individuals did not form an exclusive group (Fig. 1).

There were six mononucleotide and one dinucleotide microsatellites; four within the *Trn*L intron, two within the *Atp*B–*rbc*L intergenic spacer region, and one within the *Trn*L–*Trn*F intergenic spacer region. Variability of the microsatellites ranged from 4 to 12 alleles, with the dinucleotide microsatellite (AT) being the least variable. Size homoplasy can be detected by comparing the variation in microsatellite repeat length to a phylogeny based on the surrounding chloroplast genome. When the repeat lengths of each locus, determined from the sequence data, were superimposed on the maximum parsimony tree produced from the chloroplast base substitution data (Fig. 1), all seven loci show evidence of size homoplasy (Fig. 2), i.e., identical microsatellite repeat length has evolved more than once.

The degree of homoplasy in *Clusia* chloroplast simple base substitution data and nonmicrosatellite indels was investigated to determine whether the 100% level of homoplasy found in microsatellite loci is unusual in the chloroplast genome of this group. Overall, 18.4% of chloroplast single-nucleotide polymorphisms (SNPs) showed evidence of homoplasy (Table 2) compared to the chloroplast phylogenetic tree (Fig. 1). The level of homoplasy in indels was slightly higher, at 25% (12 informative indels in total). However, neither SNPs nor indels show as high a level of homoplasy as the microsatellite loci. Nucleotide diversity per site (ignoring indels) was lowest in the fragment with the greatest number of microsatellites. This fragment (*Trn*L intron) had considerably fewer informative SNPs compared to the other three chloroplast regions sequenced.

We analyzed the mutation patterns in each of the four chloroplast regions using a likelihood-based approach with the program CLUSTERM (Glazko et al. 1998) to determine whether homoplasy in the SNPs was the result of mutational "hotspots." Three hotspots were identified, one in each of the following regions: *Trn*L intron, *Atp*B–*rbc*L intergenic spacer region, and *acc*D–*psa*L intergenic spacer. Two of these hotspots were not homoplastic (those in the *Atp*B–*rbc*L intergenic spacer and *acc*D–*psa*L intergenic spacer regions). Thus only 14% of homoplastic SNPs were due to mutational hotspots.

It has been suggested that pooling the data from several microsatellite regions may reduce the impact of size homoplasy on the resulting phylogeny (Provan et al. 2001). However, although the phylogenetic tree

Table 2. Homoplasy and nucleotide diversity within each of the four chloroplast regions and over the total area of chloroplast DNA examined

Region	*Trn*L intron	*acc*D–*psa*L	*Atp*B–*rbc*L	*Trn*L–*Trn*F	Total
Size (bp)	699	756	807	427	2689
Size minus indels	480	718	744	416	2358
Variable sites	15	26	24	17	82
Informative sites	3	14	12	9	38
Homoplasies	2	2	2	1	7
% homoplastic	66.7	14.3	16.7	11.1	18.4
No. microsats	4	0	2	1	7
Nucleotide diversity	0.021	0.008	0.007	0.010	0.011
Diversity minus indels	0.003	0.008	0.006	0.007	0.006

Note. Diversity measures are per nucleotide (Pi). All sequenced regions represent noncoding DNA. All regions have a similar level of homoplasy except for the *Trn*L intron. The large proportion of homoplastic sites is most probably an artifact of the very low number of informative sites in this region.

constructed from multilocus microsatellite haplotypes (seven loci pooled) bears some resemblance to that constructed from the chloroplast sequence data, it is not similar enough to suggest that microsatellite haplotypes may be useful for phylogenetic reconstruction in this genus (compare Fig. 1 and Fig. 3). The fact that several strong groupings in the sequence phylogeny are not reflected in the microsatellite phylogeny (e.g., *C. aripoensis* and *C. major*, *C. flava* and *C. torresii*, the *C. minor* group) suggests that microsatellite fragment size may not be a good phylogenetic marker in this genus.

Within-species variation across all seven microsatellites was low in *C. ducu*. The average sum of squared size difference between *C. ducu* individuals was significantly lower than that between species (mean sum of squared size differences between *C. ducu* individuals = 1.0, compared to 83.89 between species; Mann–Whitney $U' = 400.0$, $p < 0.0001$). However, within-species multilocus microsatellite variation in *C. minor* and *C.* cf. *multiflora* was relatively high. Individuals within species did not form monophyletic groups using the multilocus microsatellite haplotypes for phylogenetic reconstruction (Fig. 3). The average sum of squared size differences between *C. minor* individuals was significantly lower than between species (22.67 and 83.89, respectively; $U' = 202.0$, $p = 0.036$), but the sum of squared size differences between *C. multiflora* individuals was not lower than that between species (60.67 and 83.89, respectively; $U' = 79.0$, $p = 0.728$).

Size homoplasy may result from a number of different mutational processes. Two of the microsatellite regions show size homoplasy due to base substitutions within the microsatellite. Within one of these regions (in the *Atp*B–*rbc*L intergenic spacer), there were three separate examples of size homoplasy due to base substitutions within the microsatellite: one group of four different sequences of the same length and two examples of two different sequences of the same length (Table 3), resulting from two variable

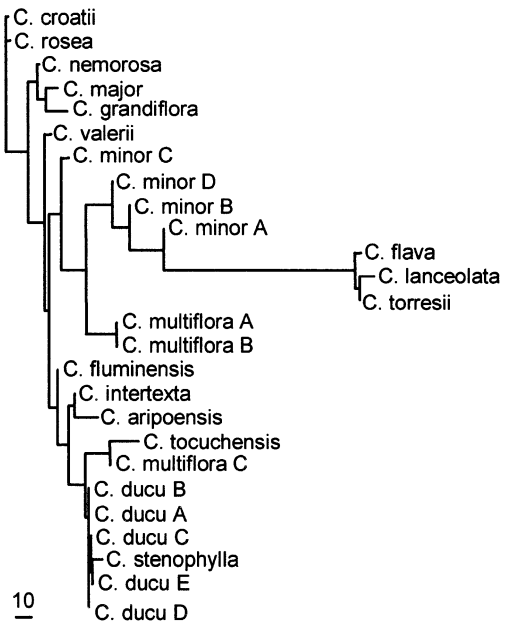

Fig. 3. Neighbor-joining tree of the sum of squared size differences between multilocus microsatellite haplotypes. Multilocus haplotypes were constructed from seven chloroplast microsatellite loci, four within the *Trn*L intron, two within the *Atp*B–*rbc*L intergenic spacer region, and one within the *Trn*L–*Trn*F intergenic spacer region. This distance measure assumes a stepwise mutation model. Jackknife values are unavailable for this type of data.

sites. The sequence data show 12 alleles at this locus, while fragment length alone shows only 7 alleles.

Size homoplasy can also be a result of gain and loss of repeat units resulting in fragments of equal size (and identical sequence) that are not identical by descent. This cause of size homoplasy is likely to occur often if microsatellites mutate in a strictly stepwise fashion. To determine whether homoplasy due to this "backward and forward" mutation is likely to exist in *Clusia*, we examined linkage disequilibrium between the seven microsatellite loci. Because all seven microsatellites are on the nonrecombining chloroplast genome, we would expect that

Table 3. Examples of size homoplasy due to base substitutions within a microsatellite

Individual	Sequence 5′ to 3′	Length
C. aripoensis	TAGACAAAATTAAAAAAAA-------GCTA	23
C. multiflora	TAGACAAAAAAAAAAAAA-------GCTA	23
C. nemorosa	TAGACAAAATCAAAAAAAA-------GCTA	23
C. stenophylla	TAGACAAAATAAAAAAAAA-------GCTA	23
C. intertexta	TAGACAAAATAAAAAAAAAA------GCTA	24
C. ducu B	TAGACAAAAAAAAAAAAAA------GCTA	24
C. minor D	TAGACAAAATAAAAAAAAAA-----GCTA	25
C. torresii	TAGACAAAATCAAAAAAAAA-----GCTA	25

Note. Sequence represents bases 369 to 398 of the aligned *Atp*B–*rbc*L intergenic spacer region sequences. The microsatellite region is shaded. The first four individuals have a fragment of identical length, yet different sequence composition, as do the following two pairs of sequences.

the seven loci should be linked if alleles of the same size are identical by descent. Identical haplotypes within species were removed from the multilocus haplotype data set, and linkage disequilibrium was calculated on the remaining 19 haplotypes using a likelihoood-ratio test. None of the seven microsatellite loci were linked to any other locus ($p > 0.05$), suggesting that size homoplasy due to backward and forward stepwise mutation exists in microsatellite loci between these *Clusia* species. It is possible that linkage was not found simply because the test did not have enough power with only 19 haplotypes. We therefore calculated linkage disequilibrium on both the informative base substitutions (SNPs) and indels within the chloroplast sequence data for these same 19 haplotypes. Of the SNP data (36 informative sites in total), 97% of informative sites were linked to at least one other site, and 50% of informative sites were linked to at least five other sites. Nonmicrosatellite indels also displayed large amounts of linkage; 92% (11 of 12) of the informative indels were linked to at least one other indel, with 58% linked to two or more indels. This suggests that the lack of linkage between the microsatellite loci is not due to lack of power in the linkage disequilibrium test.

The data set collected demonstrates three other potential sources of size homoplasy if microsatellite fragments were to be amplified and alleles determined by size alone. For the above analyses we calculated microsatellite repeat size directly from the sequence data. However, microsatellite repeat length variability is usually calculated from the size of fragments containing a microsatellite, amplified with PCR. This means that variability in the flanking regions as well as in the microsatellite itself may contribute to size homoplasy. For example, within the *Trn*L intron in *Clusia*, two microsatellite regions occurred, separated by only 24 bp of nonmicrosatellite DNA. Any primer pair designed to amplify this region would amplify both microsatellites in the same fragment. We found five examples where loss of repeat units in one microsatellite was compensated for by an equal gain in repeat units in the second microsatellite, resulting in an equivalent "fragment" size (Table 4). These five fragment sizes actually represent 12 two-locus haplotypes. Additionally, in three species the second microsatellite and intervening nonmicrosatellite DNA was completely missing (Table 4). While this combining of microsatellites will not cause size homoplasy, it will result in fragment size being unrelated to differences in repeat length at the microsatellite locus.

Indels close to the microsatellite may also result in size homoplasy, or simply in fragment size not being representative of microsatellite repeat length. The dinucleotide microsatellite within the *Trn*L intron had five indels directly adjacent to it (Table 5) and another four indels within 50 bp of the microsatellite. In one case, lack of indels in one area was compensated for by the presence of nearby indels, resulting in size homoplasy (*C. nemorosa* and *C. torresii*; Table 5). However, for most haplotype comparisons, the problem was not size homoplasy, but that fragment size differences reflected the presence/absence of indels more than differences in microsatellite repeat length.

Flanking sequence nucleotide variation provided another potential source of size homoplasy. The number of variable sites in flanking sequence up to 100 bp on either side of each microsatellite varied between 2 and 10 sites, with a mean of 4.7 ± 1.1 polymorphic sites. There were several instances of individuals with the same repeat length at a microsatellite possessing different flanking sequences. Flanking sequence variation within species is considerably lower than between species, so size homoplasy due to base substitutions within the flanking regions may not be a problem for intraspecific studies. Within-species nucleotide diversity (Watterson theta) for nonmicrosatellite chloroplast DNA was 0.48 ± 0.48 (*C. ducu*), 4.91 ± 2.97 (*C. minor*), and 7.33 ± 4.73 (*C.* cf. *multiflora*), compared to 28.04 ± 9.20 between species.

Discussion

Chloroplast microsatellites have been used as phylogenetic markers between closely related species in a

Table 4. Examples of size homoplasy due to compensating variation at closely situated microsatellites.

Individual	Sequence 5′ to 3′	Micro 1 length	Micro 2 length	Total length
C. valerii	TTTTCCGAAAAAAAAAAA----CAAAGATTTATAAAGAAAAAACATAAAAAAAAA-----GGATAGGTG	12	9	61
C. aripoensis	TTTTCCGAAAAAAAAAAA-----CAAAGATTTATAAAGAAAAAACATAAAAAAAAA-----GGATAGGTG	11	10	61
C. intertexta	TTTTCCGAAAAAAAAAAA-----CAAAGATTTATAAAGAAAAAACATAAAAAAAAA---GGATAGGTG	11	12	63
C. ducu E	TTTTCCGAAAAAAAAAA-------CAAAGATTTATAAAGAAAAAACATAAAAAAAAAAAA-GGATAGGTG	9	14	63
C. tocuchensis	TTTTCCGAAAAAAAAAA-------CAAAGATTTATAAAGAAAAAACATAAAAAAAAAAA--GGATAGGTG	10	13	63
C. stenophylla	TTTTCCGAAAAAAAAAA-------CAAAGATTTATAAAGAAAAAACATAAAAAAAAAAAAGGATAGGTG	9	15	64
C. rosea	TTTTCCGAAAAAAAAAAAAAA-CAAAGATTTATAAAGAAAAAACATAAAAAAAAA-----GGATAGGTG	15	9	64
C. multiflora C	TTTTCCGAAAAAAAAAA-------CAAAGATTTATAAAGAAAAAACATAAAAAAAAAAAA-GGATAGGTG	10	14	64
C. minor C	TTTTCCGAAAAAAAAAAAA---CAAAGATTTATAAAGAAAAAACATAAAAAAAAAAA---GGATAGGTG	13	12	65
C. croatii	TTTTCCGAAAAAAAAAAAAA-CAAAGATTTATAAAGAAAAAACATAAAAAAAAA-----GGATAGGTG	15	10	65
C. major	TTTTCCGAAAAAAAAAAAAAAACAAAGATTTATAAAGAAAAAACATAAAAAAAAA-----GGATAGGTG	16	10	66
C. grandiflora	TTTTCCGAAAAAAAAAAAAAAA-CAAAGATTTATAAAGAAAAAACATAAAAAAAAA---GGATAGGTG	15	11	66
C. lanceolata	TTTTCCGAAAAAAAAAAAAA---------------------------------------GGATAGGTG	13	0	29
C. flava	TTTTCCGAAAAAAAAAAAA--GGATAGGTG	12	0	28
C. torresii	TTTTCCGAAAAAAAAAAA---GGATAGGTG	11	0	27

Note. Sequence represents bases 92 to 162 of the aligned *Trn*L intron sequences. Two microsatellite regions (shaded) are separated by a 24-bp segment of nonmicrosatellite DNA. *C. valerii* and *C. aripoensis* have a total fragment length of 61 bp, yet the two species have different repeat lengths at both microsatellites—fewer repeats at one microsatellite are compensated for by more repeats at the second microsatellite. Size homoplasy due to this compensating effect was common within this genus (five examples found). In three species the second microsatellite was missing entirely.

number of studies (e.g., *Hordeum* [Provan et al. 1999a], *Abies* [Parducci et al. 2001], *Triticum* and *Aegilops* species [Ishii et al. 2001]). They have been suggested as a possible phylogenetic marker between closely related species where variation in other markers is too low to provide good phylogenetic resolution (Provan et al. 2001). While nuclear microsatellites are known to display size homoplasy in interspecific analyses (Doyle et al. 1998), the reported lower mutation rate of chloroplast compared to nuclear microsatellites (Provan et al. 1999b) suggests that size homoplasy may not be an issue in phylogenetic analyses of closely related species based on chloroplast microsatellite data. However, the results of this study show that size homoplasy is extensive in chloroplast microsatellite data collected from 17 *Clusia* species.

The use of more loci has been suggested as a means to avoid the problems of size homoplasy in chloroplast microsatellite data (Provan et al. 2001). However, combining the data from seven loci did not improve the accuracy of the phylogenetic signal in *Clusia*. The multilocus microsatellite data not only failed to produce an accurate phylogeny among all 17 species, it even failed to resolve the closely related groups, both interspecific and intraspecific, correctly. It is not unreasonable that the multilocus microsatellite phylogeny should differ from the chloroplast sequencing data because many groups are poorly resolved in the sequence data phylogeny. However, there are some well-resolved groups, and three interspecific pairs in particular (*C. flava* and *C. torresii*, *C. rosea* and *C. grandiflora, C. aripoensis* and *C. major*) always pair in both chloroplast sequence phylogenies and nuclear (ITS) phylogenies. Yet none of these three pairs was maintained in the multilocus microsatellite phylogeny. Perhaps even more disquieting, the intraspecific groups that are well resolved in the chloroplast sequence phylogeny (*C. ducu* and *C. minor*) do not form exclusive groups in the multilocus microsatellite phylogeny.

Intraspecific variation was present in both chloroplast sequence and microsatellite data sets in all three species where data were available but was much greater in the microsatellite data, particularly for *C. minor*. The difference in amount of intraspecific variation among the three species is probably a result of random sampling due to small sample size and may be particularly influenced by the variety in sources of the samples. Although the sample sizes of the intraspecific samples are too small to test for an association between genetic variation and geographic variation, it is clear that samples from different areas were more different genetically than samples from the same area. In addition, variation in the group accessions grouped under *C.* cf. *multiflora* could be caused by uncertain determination. Two of the three

Table 5. Examples of size homoplasy, and misleading phylogenetic information due to indels in the flanking sequences of a microsatellite

Individual	Sequence 5′ to 3′	Length
C. minor A	GTATATATATATA----------TTTTATA--TTTT----------ATTT---------	28
C. nemorosa	GTATATATATATATA----------TTATATA--TTTT----------ATTT---------	30
C. torresii	GTATATATATATATA----------TTAT----------ATTTATTTTATTT	30
C. ducu A	GTATATATATATATATA----------TTATATA--TTTT----------ATTT---------	32
C. croatii	GTATATATATATATATA----------TTATATATATTTA----------ATTT---------	36
C. lanceolata	GTATATATATATA----------TTATATA--TTTT----------ATTTATTTTTATTT	37
C. multiflora C	GTATATATATATATATATATATA----------TTATATA--TTTT----------ATTT---------	38
C. fluminensis	GTATATATATATATA----------TTATATA--TTTT----------ATTTATTTTTATTT	39
C. rosea	GTATATATATATATATA----------TTATATA--TTTT----------ATTTATTTTTATTT	41
C. aripoensis	GTATATATATATATATATA----------TTATATA--TTTT----------ATTTATTTTTATTT	43
C. tocuchensis	GTATATATATATATATATATATATATATA----------TTATATA--TTTT----------ATTT---------	44
C. major	GTATATATATATATATATA----------TTATATA--TTTT----------ATTTATTTTTATTT	45
C. flava	GTATATATATA----------TTATATA--TTTTTGTATATATATATTTATATTTTATTTTATTT	51

Note. Sequence represents bases 374 to 444 of the aligned *Trn*L intron sequences. The fragments from *C. torresii* and *C. nemorosa* are the same size (30 bp), yet they have substantially different sequences: variation in microsatellite repeat length, plus two compensating indels in the flanking region. Similarity in fragment size does not reflect similarity in microsatellite repeat length. *C. croatii*, *C. lanceolata*, *C. multiflora*, and *C. fluminensis* all have fragments differing by a single base in length, yet the microsatellites vary by as many as 10 bases. Variation in microsatellite length is masked by variation at indels in the flanking regions.

species with sample size greater than one (*C. minor*, *C. ducu*) formed monophyletic groups with high jackknife values in the sequence data phylogeny, yet none of these species formed a monophyletic group in the multilocus microsatellite phylogeny. The presence of fairly substantial intraspecific variation in the microsatellite haplotypes suggests that large samples of each taxon need to be genotyped for within-species and between-species variation to be calculated, allowing the microsatellite data to be used as a phylogenetic marker.

The lack of linkage between the seven chloroplast microsatellite regions suggests that there is either size homoplasy, probably due to the "backward and forward" stepwise mutation of microsatellite regions or recombination between the loci. As the microsatellite regions are on the uniparentally inherited, nonrecombinant chloroplast genome, it is generally assumed that recombination can be ignored (Provan et al. 2001). Yet there is some recent evidence for interchromosomal recombination of chloroplast DNA in lodgepole pine (Marshall et al. 2001) and in *Microseris* species (Vijverberg and Bachmann 1999), which suggests that the possibility of recombination being responsible for the lack of linkage between *Clusia* chloroplast microsatellite loci cannot be totally discounted. However, two of the *Clusia* microsatellite loci are separated by only 24 bases of nonmicrosatellite DNA, yet are not linked. The physical closeness of these two microsatellites makes it unlikely that recombination is responsible for this lack of linkage. In addition, the presence of linkage between single-base substitution sites and, also, between indels in the sequences surrounding the microsatellites suggests that recombination is not responsible for the lack of linkage between the microsatellites. If recombination were involved, we would expect a lack of linkage in all three types of variation (SNP, indels and microsatellites).

Of course we do not know whether the large number of homoplasies in the genus *Clusia* is exceptional in the plant kingdom as few other studies of this extent have been published yet. One could hypothesize that the relatively high number of microsatellites in this genus causes higher rates of molecular evolution in the flanking regions. Yet our data suggest that this is not the case. In contrast, the region with the most microsatellites (*Trn*L intron) had the lowest number of informative SNPs and nucleotide diversity of all four regions. It did, however, have the highest number of indels suggesting that the region may be particularly susceptible to polymerase "slippage" events—leading to a region rich in both microsatellites and indels.

There are six reasons why phylogenetic history may not be predicted by allele fragment size at microsatellite loci. (1) Base substitutions within a microsatel-

lite, resulting in interrupted microsatellites or creation of pure microsatellites from an interrupted one, may result in identical sized alleles with different sequences; (2) base substitutions in the microsatellite flanking sequences with the same result as above; (3) indels close to the microsatellite may result in large differences in allele size that are unrelated to repeat number; (4) backward and forward mutation in repeat number that result in identical alleles with different evolutionary histories; (5) more than one microsatellite within a fragment (either compound or separated by nonmicrosatellite DNA) with different mutation or evolutionary histories at each locus. This situation can result in size homoplasy where reduction of repeats at one locus is offset by increase in repeat number at the second locus, resulting in fragments of the same size with different microsatellite repeat lengths; and (6) true parallelism where the evolutionary process leading to the same DNA sequence or fragment length happens in separate lineages, independently from each other. All six 'types' of homoplasy occur in chloroplast microsatellites and flanking sequences within the genus *Clusia*.

Four of the six 'types' of size homoplasy can be detected by sequencing any microsatellite regions amplified. The results of our study suggest that sequencing of chloroplast microsatellites should be done for any interspecific study to correctly classify alleles and therefore reduce the incidence of size homoplasy in the dataset. The effects of the stepwise mutation, however, cannot be detected by sequencing. Thus, any interspecific study relying on chloroplast microsatellite data should do two things: 1) increase intraspecific sampling to allow estimates of both within and between species variation, and 2) sequence alleles of the same size in different species to detect size homoplasy from base substitutions or indels. Given the extra work this requires, it may be more feasible from a cost-benefit point of view to simply sequence more of the chloroplast genome to gain enough variation for phylogenetic analysis, rather than using microsatellite data.

Acknowledgments. We thank Barry Hammel (Missouri Botanic Garden), The Royal Botanic Garden Edinburgh, and The National Herbarium of Trinidad and Tobago for providing us with samples. This research was funded by Natural Environment Research Council Grant NERC B/S/2000/00147.

References

Borland AM, Tesci LI, Leegood RC, Walker RP (1998) Inducibility of crassulacean acid metabolism (CAM) in *Clusia* species; Physiological/biochemical characterisation and intercellular localization of carboxylation and decarboxylation processes in three species which exhibit different degrees of CAM. Planta 205:342–351

Doyle JJ, Morgante M, Tingey SV, Powell W (1998) Size homoplasy in chloroplast microsatellite of wild perennial relatives of soybean (*Glycine* subgenus *Glycine*). Mol Biol Evol 15:215–218

Felsenstein J (1993) PHYLIP (phylogeny inference package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle

Filatov DA (2002) ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets. Mol Ecol Notes 2:621–624

Glazko GB, Milanesi L, Rogozin IB (1998) The subclass approach for mutational spectrum analysis: application of the SEM algorithm. J Theor Biol 192:475–487

Grivet D, Heinze B, Vendramin GG, Petit RJ (2001) Genome walking with consensus primers: Application to the large single copy region of chloroplast DNA. Mol Ecol Notes 1:345–349

Gustafsson MHG, Bittrich V (2002) Evolution of morphological diversity and resin secretion in flowers of *Clusia* (Clusiaceae): Insights from ITS sequence variation. Nordic. J Bot 22:183–203

Hodges SA, Arnold ML (1994) Columbines: A geographically widespread species flock. Proc Natl Acad Sci USA 91:5129–5132

Ishii T, Mori N, Ogihara Y (2001) Evaluation of allelic diversity at chloroplast microsatellite loci among common wheat and its ancestral species. Theor Appl Genet 103:896–904

Marshall HD, Newton C, Ritland K (2001) Sequence-repeat polymorphisms exhibit the signature of recombination in lodgepole pine chloroplast DNA. Mol Biol Evol 18:2136–2138

Parducci L, Szmidt AE, Madaghiele A, Anzidei M, Vendramin GG (2001) Genetic variation at chloroplast microsatellites (cpSSRs) in *Abies nebrodensis* (Lojac.) Mattei and three neighboring *Abies* species. Theor Appl Genet 102:733–740

Pipoly JJ, Kearns DM, Berry PE (2000) Key to the species and sections of *Clusia* in the Venezuelan Guayana. http://www.mobot.org/MOBOT/research/ven-guayana/clusiaceae/clusia.html

Provan J, Russell JR, Booth A, Powell W (1999a) Polymorphic chloroplast simple sequence repeat primers for systematic and population studies in the genus *Hordeum*. Mol Ecol 8:505–511

Provan J, Soranzo N, Wilson NJ, Goldstein DB, Powell W (1999b) A low mutation rate for chloroplast microsatellites. Genetics 153:943–947

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: New tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147

Schneider S, Roessli D, Excoffier L (2000) Arlequin ver. 2000: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva, Switzerland

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (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

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, MA

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol Biol 17:1105–1109

Vijverberg K, Bachmann K (1999) Molecular evolution of a tandemly repeated *trn*F(GAA) gene in the chloroplast genomes of *Microseris* (Asteraceae) and the use of structural mutations in phylogenetic analyses. Mol Biol Evol 16:1329–1340

Weising K, Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42: 9–19