Patterns and Relative Rates of Nucleotide and Insertion/Deletion Evolution at Six Chloroplast Intergenic Regions in New World Species of the Lecythidaceae

Matthew B. Hamilton,*† John M. Braverman,‡ and David F. Soria-Hernanz*

*Department of Biology, Georgetown University; †Biological Dynamics of Forest Fragments Project,
National Institute for Research in the Amazon, Manaus, Brazil; ‡Department of Biology, Loyola University, Chicago

Insertions and deletions (indels) in chloroplast noncoding regions are common genetic markers to estimate population structure and gene flow, although relatively little is known about indel evolution among recently diverged lineages such as within plant families. Because indel events tend to occur nonrandomly along DNA sequences, recurrent mutations may generate homoplasy for indel haplotypes. This is a potential problem for population studies, because indel haplotypes may be shared among populations after recurrent mutation as well as gene flow. Furthermore, indel haplotypes may differ in fitness and therefore be subject to natural selection detectable as rate heterogeneity among lineages. Such selection could contribute to the spatial patterning of cpDNA haplotypes, greatly complicating the interpretation of cpDNA population structure. This study examined both nucleotide and indel cpDNA variation and divergence at six noncoding regions (*psb*B-*psb*H, *atp*B-*rbc*L, *trn*L-*trn*H, *rpl*20-5′*rps*12, *trn*S-*trn*G, and *trn*H-*psb*A) in 16 individuals from eight species in the Lecythidaceae and a Sapotaceae outgroup. We described patterns of cpDNA changes, assessed the level of indel homoplasy, and tested for rate heterogeneity among lineages and regions. Although regression analysis of branch lengths suggested some degree of indel homoplasy among the most divergent lineages, there was little evidence for indel homoplasy within the Lecythidaceae. Likelihood ratio tests applied to the entire phylogenetic tree revealed a consistent pattern rejecting a molecular clock. Tajima's 1D and 2D tests revealed two taxa with consistent rate heterogeneity, one showing relatively more and one relatively fewer changes than other taxa. In general, nucleotide changes showed more evidence of rate heterogeneity than did indel changes. The rate of evolution was highly variable among the six cpDNA regions examined, with the *trn*S-*trn*G and *trn*H-*psb*A regions showing as much as 10% and 15% divergence within the Lecythidaceae. Deviations from rate homogeneity in the two taxa were constant across cpDNA regions, consistent with lineage-specific rates of evolution rather than cpDNA region-specific natural selection. There is no evidence that indels are more likely than nucleotide changes to experience homoplasy within the Lecythidaceae. These results support a neutral interpretation of cpDNA indel and nucleotide variation in population studies within species such as *Corythophora alta*.

Introduction

Chloroplast DNA (cpDNA) sequence comparisons have been a widely used tool in studies of plant phylogenetics and genome evolution (Curtis and Clegg 1984; Palmer 1990; Clegg et al. 1994). More recently, chloroplast DNA has been employed as a genetic marker for studies focused on intraspecific evolution, particularly estimates of population structure. The use of chloroplast genetic markers in population studies has been motivated by the same features of haploid genomes that have made mitochondrial DNA such a widely used tool in animal studies (e.g., Avise 1991; Ennos et al. 1999). In plants, the lack of cpDNA recombination and frequent uniparental inheritance permits the spatial pattern of cpDNA haplotypes to be interpreted as an estimate of past seed or pollen dispersal (McCauley 1995). Furthermore, because the chloroplast genome is haploid, it has a smaller effective population size than the diploid nuclear genome. This accelerates the process of drift, causing the chloroplast genome to exhibit neutral differentiation among diverged lineages or populations sooner than nuclear alleles show substantial differentiation (Birky, Fuerst, and Maruyama 1989; Petit, Kremer, and Wagner 1993; Hamilton and Miller 2002). Thus, if divergence has occurred relatively recently, such as between populations with limited gene

flow, neutral cpDNA polymorphisms will have more power to detect such differentiation compared to neutral nuclear polymorphisms.

Chloroplast DNA has blossomed as an intraspecific genetic marker for at least two practical reasons. First, complete genome sequences and the development of conserved oligonucleotide primers for both coding and noncoding cpDNA regions (e.g., Taberlet et al. 1991; Demesure, Sodzi, and Petit 1995; Dumolin-Lapegue, Pemonge, and Petit 1997; Hamilton 1999*a*) have provided a ready means to assay cpDNA sequence variation. Second, cpDNA intergenic regions have exhibited substantial intraspecific insertion/deletion (indel) polymorphism within and among plant populations (e.g., McCauley 1994, 1998; Powell et al. 1995; Dumolin-Lapégue, Pemonge, and Petit 1998; Hamilton 1999*b*; Caron et al. 2000; Desplanque et al. 2000; Dutech, Maggia, and Joly 2000; Muloko-Ntoutoume et al. 2000; Oddou-Muratorio et al. 2001). Although the traditional view that plant cpDNA changes very slowly in nucleotide sequence still holds to some extent (Palmer 1990; Provan, Powell, and Hollingsworth 2001), indels appear to evolve more rapidly and are therefore useful for studies both at the intraspecific and shallow interspecific levels. Despite the availability of intraspecific cpDNA indel variation, employing polymorphic indel haplotypes to make evolutionary inferences about population divergence and gene flow faces several challenges.

Indels clearly do not occur at random locations within organelle genomes, but they are often associated with specific features of DNA sequences. Regions containing repeats that lead to slipped-strand mispairing (e.g.,

mononucleotide and microsatellite repeats), stem-loop secondary structure, and intramolecular recombination are thought to cause the majority of insertion/deletion mutations (reviewed by Kelchner 2000). Thus, recurrent mutations ("multiple hits") may occur at the same sites and generate homoplasy for organelle indel haplotypes (Clegg et al. 1994). Haplotypes defined by indels then have increased probabilities of being identical in state without being identical by descent. Such homoplastic characters cause character state reversals in phylogenetic trees and violate the basic assumptions of measures of population subdivision (e.g., F_{ST} or G_{ST}). Indel homoplasy will tend to reduce inferred population subdivision, because populations can share haplotypes as a result of both gene flow and recurrent mutation. This problem has been widely recognized for nuclear microsatellite loci (e.g., Jarne and Lagoda 1996; Goldstein and Pollock 1997; Hedrick 1999) and for the use of indel polymorphism in phylogenetic reconstruction (e.g., Golenberg et al. 1993; Graham et al. 2000; Kelchner 2000). Fewer studies have evaluated the extent of cpDNA homoplasy among recently diverged species (e.g., Doyle et al. 1998) or among populations of a single species (e.g., Desplanque, et al. 2000).

Both indels and nucleotides exhibit homoplasy. Many models of nucleotide substitution are available to estimate actual sequence divergence from observed sequence changes (reviewed in Swofford et al. 1996). At present, there are no generally employed methods to adjust observed numbers of indel changes for multiple unobserved evolutionary events at an indel site, although a few indel mutation models do exist that could form the basis of a correction method (e.g., Tajima and Nei 1984; Thorne, Kishinmo, and Felsenstein 1992; McGuire, Denham, and Balding 2001). Without estimates of rates of indel mutational homoplasy, it is potentially difficult to distinguish the action of evolutionary processes that homogenize populations (e.g., gene flow, selection) from convergence because of saturation of changes as divergence among populations increases.

Nucleotide substitution rate estimates are available for coding regions of the chloroplast genome compared among relatively deeply diverged taxa (reviewed by Clegg et al. 1994; Muse 2000) and more recently for intergenic regions compared at shallow levels of divergence such as within families (e.g., Gaut et al. 1997; reviewed in Kelchner 2000). Because indels found in multiple regions of the cpDNA genome may exhibit absolute rate differences, estimates of rate variation among cpDNA intergenic regions would be helpful when selecting regions for population studies because genetic markers sought for a specific hypothesis must be selected based on the level of lineage resolution required (Avise 1994; Parker et al. 1998). Furthermore, estimates of indel evolutionary rates would be valuable for interpreting the spatial patterns of indel haplotypes observed in population structure studies, providing context on the relative degree of divergence (e.g., numbers of indel changes among populations compared to among closely related species).

Divergence estimates can help identify cpDNA regions exhibiting disproportionately slow or fast rates of evolution, a result consistent with some types of selection

(see animal mtDNA review by Ballard and Kreitman 1995). Indel mutations may directly alter fitness and therefore be subject to natural selection. For example, deletion mutations could be favored if smaller genomes have higher fitness in a "race for replication" (reviewed by Rand 1993), causing the rate of indel evolution to differ from a neutral rate. The action of selection on indel haplotypes could be diagnosed as rate heterogeneity among lineages (Sarich and Wilson 1973). Chloroplast regions where mutational changes do not approximate a molecular clock have the potential to greatly complicate the interpretation of spatial patterns of haplotypes observed in population structure studies. Direct selection on haplotypes can result in patterns of population subdivision that are the result of local selection pressures and not historical gene flow—for example, locally adapted haplotypes where the fitness of a given haplotype varies among populations. Direct selection on haplotypes can also result in reduced population subdivision of haplotypes as a result of uniform selection pressures among populations, even though gene flow may be limited (e.g., Stephan et al. 1998).

In this study we examined patterns of inter- and intraspecific sequence variation at six chloroplast intergenic sites for 16 individuals from eight species in the Brazil nut tree family (Lecythidaceae). Our analyses were designed to examine the evolution of indels used previously to estimate population structure in the tropical tree *Corythophora alta* (Hamilton 1999a). That study found cpDNA indel variation partitioned almost entirely among populations. The pattern is consistent with either very limited seed dispersal (if indels are neutral with low levels of homoplasy), or with regional selection pressures that have influenced the spatial pattern of cpDNA haplotypes. Our goals here were to describe the evolutionary history of cpDNA variation in close relatives of *C. alta* and to test the assumptions of the population structure study to the extent possible with a phylogenetic analysis. We employed a comparative analysis of cpDNA variation using nucleotide changes as a standard, because they are well understood and thoroughly modeled, to examine the evolution of indel changes, which are much less well understood or modeled. We compared the relative evolutionary rates of cpDNA indel and nucleotide changes to determine if indel homoplasy occurs frequently enough to confound substantially the effects of gene flow with the effects of recurrent mutation. We also sought to determine if the relative rate of cpDNA evolutionary change in *C. alta* is excessively fast or slow compared to other closely related species. Such relative rate variation could be consistent with homoplasy or with the action of selection on cpDNA haplotypes and would indicate that the distribution of cpDNA variation among populations may not be due exclusively to drift and gene flow. We also sought to determine if there was evidence of rate heterogeneity among the six cpDNA regions examined.

Materials and Methods

Leaf tissues from eight species of Lecythidaceae (*Bertholletia excelsa*, *Cariniana micrantha*, *Corythophora alta*, *Corythophora rimosa*, *Couratari guianensis*,

Table 1
List of Species, Population and Individual Identification Numbers, and Taxonomic Abbreviation for the cpDNA Lineages Sampled

Family	Species	Population[a]	Individual[a]	Abbreviation
Lecythidaceae	*Bertholletia excelsa*	n/a	n/a	Bex
	Cariniana micrantha	2107	122	Cmi
	Corythophora alta	1202	2058	Cal
		1202	2829	
		1202	5022	
		1501	4133	
		3114	216	
		3209	1752	
	Corythophora rimosa	1202	5614	Cri
	Couratari guianensis	1501		Cgu
	Couratari multiflora	1501	23	Cmu
		1501	26	
	Eschweilera romeucardosoi	1501	4514	Ero
	Lecythis zabucajo	1202	1267	Lza
		2108	86	
Sapotaceae	*Chrysophylum cainito*	n/a	n/a	Cca

[a] Populations are numbered according to BDFFP nomenclature, and individual numbers are tree tag numbers if present.

Couratari multiflora, *Eschweilera romeucardosoi*, *Lecythis zabucajo*) were collected at the Biological Dynamics of Forest Fragments Project (BDFFP) in Manaus, Brazil (Gascon and Bierregaard 2001). Consistent with BDFFP botanical inventory plot nomenclature, individual trees were identified by a four-digit population number and a one- to four-digit individual number (e.g., 1202–2829; table 1). Trees sampled in population 1501, which has no BDFFP botanical inventory plot, were from the Lecythidaceae plot established by S. Mori and P. Becker (S. Mori, personal communication; see Mori, Becker, and Kincaid 2001). *Chrysophylum cainito* (Sapotaceae) was collected in Panama and used as an outgroup. A member of the Sapotaceae was chosen as the outgroup because the family was shown to be closely related to Lecythidaceae (Morton et al. 1997; Savolainen et al. 2000) and therefore provided character state polarization with minimal divergence.

Six different noncoding regions of the chloroplast genome were amplified by polymerase chain reaction (PCR) and sequenced. The *trn*L-*trn*H region was amplified with primers c and f and sequenced with primers c, d, e, and f of Taberlet et al. (1991). The four regions *psb*B-*psb*H, *rpl*20-5′*rps*12, *trn*S-*trn*G and *psb*A-*trn*H were amplified and sequenced using the primers described in Hamilton (1999*b*). The *atp*β-*rbc*L region was amplified using primers 2 and 5 described by Savolainen et al. (1994), with two modified internal sequencing primers (7sub: 5′ TCACAACAACAAGGTCTACTCG 3′; 9sub: 5′ GAATTTGAAAATTCAACCAACCC 3′) designed to substitute for primers 7 and 9. The alternate internal primer 5′ GAATTTGAAAATTCAACCACCC 3′ was used in lieu of primer 9sub for *C. micrantha*.

Genomic DNA was extracted by grinding frozen leaf tissue in liquid nitrogen and using a DNeasy plant kit (QiaGen, Valencia, Calif.) according to the manufacturer's instructions. The PCR reactions contained 2 μL of DNA template (DNA concentration was not determined), 2 μL of 10× Thermopol buffer (containing 20 mM MgSO₄), 0.2 mM each dNTP, 0.4 μm of each primer and 0.4 units of Vent exo-polymerase (New England Biolabs, Cambridge, Mass.) in

a total volume of 50 μL. The thermal cycling profiles were 5 min at 96°C followed by 30–40 cycles of 96°C for 45 s, annealing temperature for 1 min and 72°C extension for 30 s (see Hamilton 1999*b*). For *trn*H-*trn*L we used 30 cycles and a 55°C annealing temperature. For *atp*β-*rbc*L we used 5 min at 94°C followed by 35 cycles of 94°, 55° and 72°C for 1 min each. The PCR products were purified with QiaQuick spin columns (QiaGen), and both strands were sequenced in reactions containing 4.6 μL water, 2 μL template, 2.4 μL primer (1 μM) and 6 μL dRhodamine or Big Dye version 2 terminator reaction ready mix (Applied Biosystems, Foster City, Calif.). Sequence reactions were purified with Centrisep spin columns (Princeton Separations, Adelphia, N.J.) and electrophoresed on a model 377 sequencer (Applied Biosystems, Foster City, Calif.). In a few cases the PCR product repeatedly yielded poor sequence data for one strand, and the PCR product was cloned with the Zero Blunt TOPO PCR Cloning Kit Sequencing Version H (Invitrogen, Carlsbad, Calif.). With cloned PCR products, sequencing reactions used T7 and T3 vector primers instead of region-specific cpDNA primers.

The resulting sequences for each cpDNA region were aligned into contigs for each individual and edited using Sequencher 3.1.1 and 4.1.2 (GeneCodes, Ann Arbor, Mich.). Multiple sequence alignments for each cpDNA region were made manually with the aid of Sequencher from the consensus sequences from each individual. Sequences were trimmed to exclude terminal coding regions and the *psb*N gene within the *psb*B-*psb*H region in all analyses. Gaps in the multiple sequence alignments were positioned to minimize the number of nucleotide differences among sequences where possible. Insertion/deletion characters and states were scored following the "complex" coding method of Simmons and Ochoterena (2000), including the use of ordered step matrices to encode the minimum number of events that separate indel character states where necessary. Indel character states and step matrices were coded using MacClade 4.03. Indel character states were based only on length and did not include nucleotide differences within regions exhibiting sequence length variation. This conser-

vative approach was taken because such nucleotide changes may not be independent of indel events. Single base indels and indels associated with mononucleotide runs were all subject to an additional check against original sequence chromatograms after multiple sequence alignment to verify that they were not a result of base calling error. Sequences were deposited in GenBank individually (accession numbers AY172691 to AY172799) and as multiple sequence alignments for each region or PopSets.

Statistical Analysis

The numbers of indel and nucleotide changes on each branch of the phylogenetic tree were compared using a regression approach (Golenberg et al. 1993; Saitou and Ueda 1994; Graham et al. 2000) that tested two hypotheses: (1) that rates of nucleotide and indel changes were equal and (2) that indel changes did not experience saturation because of mutational homoplasy. The first hypothesis was evaluated by testing for a slope of one between indel and nucleotide changes, and the second hypothesis was tested by comparing coefficients of determination (R^2) of linear and quadratic fits. First, PAUP* (version 4.0b10; Swofford 2002) was used to estimate a maximum parsimony, 1,000 iteration bootstrap consensus tree with the outgroup specified while excluding all sites with gaps for any taxa and all missing and ambiguous sites. Parsimony character state optimization was set to Deltran to avoid a known bug with Arctran branch length estimates (http://paup.csit.fsu.edu/problems.html). This tree was then saved, all data except indel characters were excluded, and a second maximum parsimony, bootstrap consensus tree was estimated under the topological constraints of the saved nucleotide-only tree. This provided an estimate of the number of indel and nucleotide changes on each branch of the tree estimated from nucleotide data, which can be compared by regression without non-independent comparisons. Terminal clades with branch lengths of zero (two *C. alta* and one *C. multiflora*) for both nucleotide and indel changes were counted only once in the regression analysis. Regression analyses were carried out with JMP (version 5.0.1a, SAS Institute, Cary, N.C.).

To choose the best supported model of nucleotide evolution for the aligned sequences with hierarchical likelihood ratio tests, we employed the program Modeltest 3.06 (Posada and Crandall 1998). The best-fitting model and the resulting parameter estimates were then used in PAUP* to estimate maximum likelihood pairwise distances between sequences for nucleotide changes only (not indels). When distances were estimated by maximum likelihood for cpDNA regions separately, base frequencies were estimated by maximum likelihood for each region independently.

A likelihood ratio test of the molecular clock for the full five- and six-region data sets as well as for each data set while excluding one region was conducted along the lines of Posada and Crandall (2001). Employing the substitution model selected with Modeltest, we obtained in PAUP* likelihood scores and topologies of phylogenies not assuming a molecular clock. This topology was then loaded as a constraint, and likelihood scores were obtained for phylogenies assuming a molecular clock. The log

likelihood scores were used to compute the statistic delta $= 2(\ln L_1 - \ln L_0)$, which is distributed as a χ^2 with the number of taxa minus 2 degrees of freedom (see Huelsenbeck and Crandall 1997).

We used Tajima's (1993) 1D and 2D nonparametric tests for deviations from the constant rate expectation that the number of unique changes along one lineage equals that along its sister lineage. Because these tests do not depend on a model of sequence change, they can be applied directly to indel characters as well as to nucleotide substitutions. We used only nucleotide changes that occurred outside of indel sites, consistent with our scoring of indel haplotypes explained above. The 1D was conducted with either nucleotide or indel changes separately using a χ^2 distribution with 1 degree of freedom. The 2D was applied to nucleotide substitutions and indels to test rate constancy using the joint outcome of the two types of changes using a χ^2 distribution with 2 degrees of freedom. To obtain the results of the 1D and 2D tests, a program called T1Dand2D (version 6.4) was written in C and used in conjunction with an Excel spreadsheet.

Not all pairs of sequences had sufficient numbers of nucleotide or indel changes to apply the χ^2 approximation for the 1D, which requires that each lineage have at least six observed changes (Tajima 1993; Nei and Kumar 2000). Thus, for 1D tests where at least one lineage had five or fewer changes we applied an exact binomial test to calculate the probability of observing an outcome as likely or less likely than the one observed under the null hypothesis that 50% of the changes occur on each lineage (using a binomial calculator at http://home.clara.net/sisa/binomial.htm).

We observed that the χ^2 approximation has a high type I error (null rejected when it is actually true) when at least one lineage had five or fewer changes. The χ^2 approximation is especially poor for small numbers of changes on both lineages. For example, we observed cases in the 1D test of four changes on one lineage and zero changes on the other lineage. Such an outcome has a probability of 1/16 or 0.0625 when the null hypothesis of equal rates is true, but a probability of 0.0455 under the χ^2 approximation. Because the 2D test uses the sum of the χ^2 values for each class of change to determine the overall probability of having observed the joint outcome of changes under the null hypothesis, the exact binomial test cannot be employed. Thus, the 2D tests are not adjusted for the high type I error rate of the χ^2 approximation and likely overstate the occurrence of deviations from rate constancy when numbers of changes are small for one or both categories of changes. The results indicate cases of the 2D test that have 5 or fewer changes on one lineage for one or both categories of mutational change.

Results and Discussion

The alignment of five cpDNA regions (less *trnH-psbA*) for all taxa and individuals produced 4,209 nucleotide sites (table 2). Sequences were impossible to align for all taxa for the *trnH-psbA* region, but a 527 site alignment of all haplotypes for *B. excelsa*, *C. alta*, *C. rimosa*, *E. romeucardosoi*, and *L. zabucajo* was possible. Therefore, two data sets were produced: one that contained

Table 2
Numbers of Types of Indels Observed in Six Different Non-Coding cpDNA Regions in Lecythidaceae

Region	Length in Base Pairs[a] (length gaps excluded)	Mononucleotide Repeats	Perfect Repeats	Imperfect Repeats	Palindromic or Inversion	Other[b]
psbB-psbH	539 (437)	0	3	4	0	3
atpB-rbcL	837 (716)	3 A/T, 1 G/C	7	1	0	4
trnL-trnH	998 (876)	3 A/T, 1 G/C	8	2	0	6
rpl20-5'rps12	830 (737)	4 A/T, 1 G/C	4	1	3	6
trnS-trnG	1005 (551)	6 A/T, 2 G/C	5	11	1	15
trnH-psbA[c]	527 (273)	2 A/T	5	1	0	5
Total	4736 (3590)	18 A/T, 5 G/C	32	20	4	39

[a] Total length of aligned intergenic sequence including or excluding sites with gaps.

[b] Insertion or deletion of sequence that is not obviously repetitive, structured, or similar to nearby sequence.

[c] Based on alignment of 11 haplotypes from five species (*B. excelsa*, *C. alta*, *C. rimosa*, *E. romeucardosoi*, and *L. zabucajo*) for this region.

sequences from all haplotypes for all cpDNA regions except *trnH-psbA* and one that contained 11 haplotypes from *B. excelsa*, *C. alta*, *C. rimosa*, *E. romeucardosoi*, and *L. zabucajo* for all six cpDNA regions. Hereafter, we refer to them as the "five region" or "six region" data sets.

The aligned sequences exhibited substantial amounts of both intra- and interspecific indel and nucleotide variation. Table 2 summarizes the nature of the DNA sequence in the 95 observed indels, which were classified into six types. Indel classes were: (1) A/T or C/G mononucleotide repeats, (2) perfect repeats of a two base pair or greater motif, (3) imperfect repeats of a two base pair or greater motif, (4) palindromic sequences, (5) reverse-compliment indels or indels abutting such repeats, and (6) other sequences that were not apparently repetitive, structured, or related to abutting sequence. Overall, indels associated with nonrepetitive sequence were most common, followed by perfect repeats, mononucleotide repeats, and then imperfect repeats. Only four of 95 indels were associated with palindromic or reverse-compliment repeat sequences. Nucleotide changes were not frequently observed within indel regions, occurring mostly in *trnS-trnG* and *trnH-psbA* cpDNA regions.

Figure 1 shows the frequency distribution of the lengths of indels, which ranged in length from one to 68 base pairs. Shorter indels were more common, with only 11 of 95 (11.6%) indels longer than 20 base pairs. The distribution has a modal indel length of five base pairs. The indel length distribution observed here is somewhat less skewed to the left than a distribution observed for a broad comparison of angiosperms, which had a modal length of one and average length of almost five base pairs (Graham et al. 2000).

Multiple haplotypes sequenced for *C. alta*, *L. zabucajo*, and *C. multiflora* showed different levels of intraspecific polymorphism. The six *C. alta* haplotypes formed two distinct groups, with individuals from the 1202 and 1501 populations sharing identical sequences and individuals from the 3114 and 3209 populations sharing identical sequences. These two groups of *C. alta* haplotypes were distinguished by seven nucleotide changes and 23 indel changes, a pattern consistent with a previous survey of *C. alta* cpDNA indel variation (Hamilton 1999a). The *L. zabucajo* haplotypes, sampled from two populations about 30 km apart, showed 17 nucleotide and 11 indel changes. Two *C. multiflora*

haplotypes, sampled within 200 m of each other at the same location, showed neither nucleotide nor indel changes. In analyses involving pairwise sequence comparisons, the multiple haplotypes of *C. alta* were represented by one sequence from the 1202 population and one sequence from the 3114 population, whereas *C. multiflora* was represented by one sequence.

The bootstrap consensus parsimony trees based only on nucleotide data from the five and six region data sets are shown in figure 2. *C. cainito* was specified as the outgroup for the five region data (fig. 2A), and *B. excelsa* was used as the outgroup for the six region data (fig. 2B) based on its position in the five region tree. For each region there was a single most parsimonious tree with topology identical to the bootstrap consensus, maximum likelihood, and Neighbor-Joining trees (results not shown). Each branch of these trees shows the inferred nucleotide and indel character state changes used for regression analyses. Increased character state reversals (decreased consistency indices) for trees with combined nucleotide and gap data suggested the possibility of modest homoplasy for either nucleotide or indel data. Consistency indices were higher for phylogenies estimated with nucleotide data alone (0.954 and 0.988) than when estimated using combined data constrained to nucleotide-estimated topology (0.869 and 0.849) for the five and six region data, respectively.

Using the five region data, the quadratic model estimate for nucleotide branch length on indel branch length was # indel changes = $2.500 + 0.543*$(# nucleotide changes) $- (0.001*$ (# nucleotide changes $- 17.541))^2$ (fig. 3). Both the intercept ($t = 2.37$, $P < 0.029$) and linear slope ($t = 6.73$, $P < 0.0001$) were significantly different than zero. The linear model estimate # indel changes = $3.836 + 0.324*$(# nucleotide changes) had an intercept significantly different than zero ($t = 2.37$, $P < 0.029$) and slope significantly different than both zero and one ($t = 13.78$, $P < 0.001$; $t = -28.73$, $P < 0.001$, respectively). A quadratic regression provided a better fit to the data than a linear model, indicated by a slightly higher R^2 value (0.935 and 0.906, respectively) and a quadratic term significantly different than zero ($t = -2.83$, $P = 0.011$). Regression results for the six region data supported a linear model, # indel changes = $3.969 + 0.502 *$ (# nucleotide changes) with an intercept significantly different than zero ($t = 2.48$, $P < 0.030$), a slope significantly different than both zero and one ($t = 6.12$, $P < 0.0001$; $t = -6.06$, $P <$
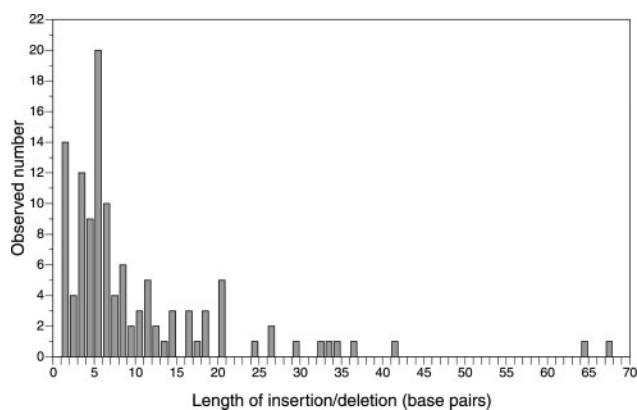
FIG. 1.—Frequency of the lengths of each indel scored as a character for all six cpDNA regions. Indel length was defined as the difference between the length of the longest and shortest haplotype at each indel site.
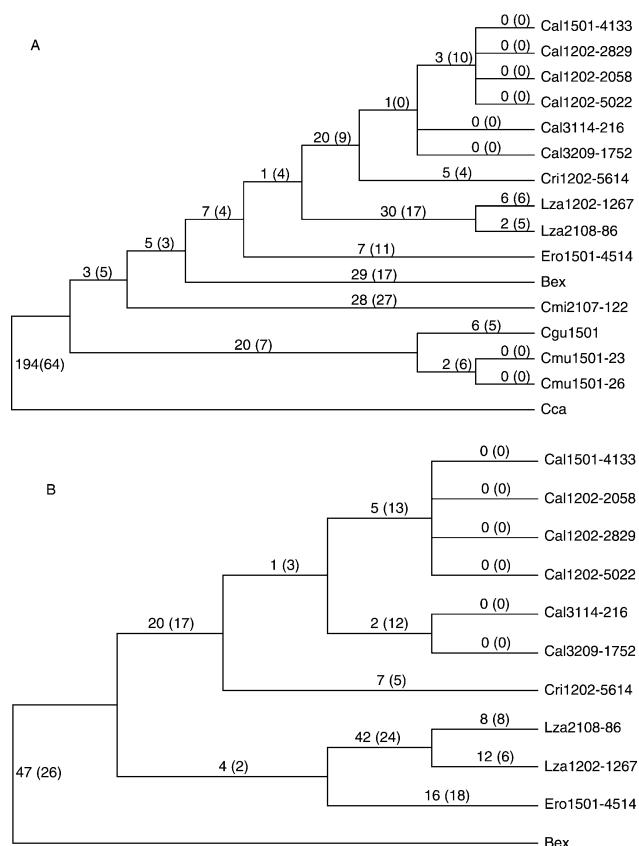

FIG. 2.—Maximum parsimony trees estimated from nucleotide changes showing inferred number of nucleotide and indel changes on each branch (indel changes in parentheses). Tree A was estimated from the five region data with *Chrysophylum cainito* specified as the outgroup (CI = 0.954, CI = 0.869 for combined data constrained to nucleotide-estimated topology). Tree B was estimated from the six region data with *Bertholletia excelsa* specified as the outgroup (CI = 0.988, CI = 0.849 for combined data constrained to nucleotide-estimated topology).

0.0001, respectively) and the variance explained was high ($R^2 = 0.77$). A quadratic regression for the six region data did explain slightly more variance ($R^2 = 0.80$), but the quadratic term was not significantly different than zero ($t = -1.15$, $P = 0.275$). Because of the low level of nucleotide divergence, there was little effect on the regression results when using inferred nucleotide changes after adjusting nucleotide changes with the substitution model specified by Modeltest (results not shown).

The regression results provided an approximate apparent rate comparison for indel and nucleotide changes. The regression y-intercepts predicted about 2.5 and 3.9 indel changes when there had been no nucleotide changes for the five and six region data, respectively. If indel and nucleotide changes occurred at the same rate, the expected slope of the regression would be 1. The estimated regression slopes were less than 1, suggesting that indel changes apparently occurred about half as frequently as nucleotide changes. This is evidence for homoplasy of indel changes because the intercepts indicate a higher rate for indel changes, but the slope indicates a lower rate for indel changes.

The quadratic regression terms also provided insight into indel homoplasy, because multiple hits are expected to lead to saturation of changes and a rate of apparent change that declines as divergence increases. There was evidence for weak saturation of apparent changes for the five region data because the quadratic term was significantly negative. It is important to note that the degree of curvature in the regression was slight and the estimate was heavily influenced by one branch with the most changes. The quadratic regression term for the six region data provided no evidence for indel saturation, perhaps because of more recent divergence among taxa and a smaller sample size of branch lengths.

For both the five and six region data sets, Modeltest selected the K81uf + G (also called K3P) model of nucleotide substitution (Kimura 1981). Selection of this model indicated it was most likely that nucleotide frequencies were unequal (estimated frequencies of A = 0.3339, C = 0.1578, G = 0.1455, T = 0.3627 were essentially identical for the two data sets), rates of transition and transversion were unequal, rates varied

among sites, and no sites were invariant. The substitution models for the five and six region data sets did differ in their rate matrices (1.0000 1.6091 0.2494 0.2494 1.6091 and 1.0000 1.0954 0.3555 0.3555 1.0954) and gamma parameters (1.1457 and 0.2094). Substitution models for individual regions were K81uf (*atp*B-*rbc*L, *trn*L-*trn*H and *trn*S-*trn*G), HKY (*psb*B-*psb*H) or F81 (*trn*H-*psb*A and *rpl*20-5′*rps*12). The former two models both indicate unequal transition and transversion rates, but the K81uf has two different rates for transversions. The F81 indicates no difference in rates of transition and transversion. The K81uf + G models served as an average model for combined-region data and were used to estimate pairwise distances between sequences (tables 3 and 4). Distances in intraspecific comparisons (within *C. alta* and *L. zabucajo*) were 0.4% divergence or less, whereas intergeneric divergences within the Lecythidaceae ranged from above 3% down to those observed at the intraspecific level. Divergences between the Lecythidaceae and the Sapotaceae outgroup were around 7% in all cases. We note that distances estimated under the K81uf + G model, a region-specific model (e.g., K81uf) and uncorrected distances ("p" distances) were very similar, differing by a few
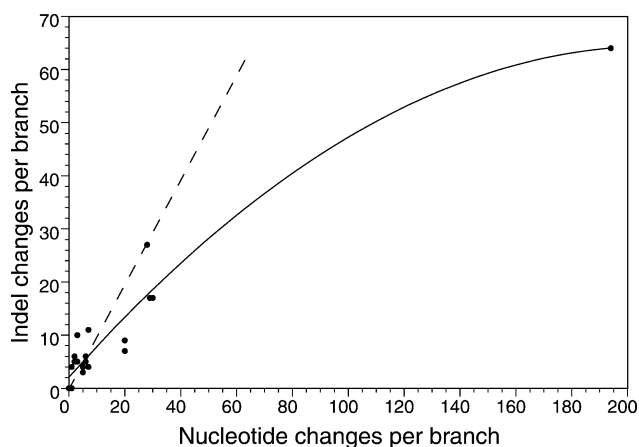
Fig. 3.—Plot of the number of nucleotide changes and number of indel changes on each branch of a parsimony tree estimated from nucleotide changes in the five region data set. The line is a quadratic regression fit with the equation # indel changes = 2.500 + 0.543* (# nucleotide changes) − (0.001*(# nucleotide changes − 17.541))². The dashed line has a slope of 1 as expected if numbers of indel and nucleotide changes were equal on each branch.

percent at most in comparisons among the most diverged lineages (results not shown).

Using the K81uf + G substitution model, the likelihood ratio tests of the molecular clock at the level of the entire phylogeny rejected the hypothesis of rate homogeneity for nucleotide changes among lineages for the five and six region data sets (results not shown). Based on these results, we sought to examine the role of specific taxa and specific cpDNA regions in the pattern of rate heterogeneity using a series of Tajima's 1D/2D tests. The 1D/2D tests also had the advantage of being applicable to both indels and nucleotide changes, whereas the likelihood tests ratio can be applied only to nucleotide changes.

For the five region data, the two Tajima 1D/2D tests did not in general support deviations from a molecular clock for indel or nucleotide changes (table 5). Both the 1D for nucleotide substitutions alone and the 2D showed that only the *E. romeucardosoi* lineage gave consistently significant deviations from rate constancy. *E. romeucardosoi* generally showed relatively fewer nucleotide changes per site and relatively fewer indel changes, suggesting a rate slowdown in that lineage (see table 3 and figure 2A). Most of the remaining lineages did not indicate significant deviations from rate constancy using indel changes alone, nucleotide changes alone, or both classes of changes in the 2D test.

The Tajima tests for the six region data set are presented in table 6. This test suggested that relative evolutionary rates were not constant in *E. romeucardosoi* and *L. zabucajo*. In particular, the *L. zabucajo* lineages exhibited high nucleotide divergences consistent with relative rate acceleration but relative rates of indel change consistent with a molecular clock (see table 4 and figure 2B). *E. romeucardosoi* exhibited an opposite pattern of relative rate slowdown for indel changes and a few instances of nucleotide changes (see table 4 and fig. 2B). The *C. alta* and *C. rimosa* lineages showed no evidence of deviations from a molecular clock for either nucleotide or indel changes.

Tajima's 1D/2D is potentially sensitive to the outgroup sequence employed (Tajima 1993; Bromham et al. 2000). However, we found results for Tajima's 1D/2D to be generally similar even when employing several different taxa as outgroups within the five and six region data sets. In the five region data set, changing the outgroup from *C. cainito* to *B. excelsa* resulted in frequent rejection of the molecular clock for nucleotides in *L. zabucajo* as well as *E. romeucardosoi*. This result was consistent with the pattern in the six region data where both the *L. zabucajo* and *E. romeucardosoi* lineages do not match the expectations of

Table 3
Number of Indel Changes (Above Diagonal) and Number of Nucleotide Substitutions and Nucleotide Changes per Site (K81uf + G Model, Below Diagonal) from Pairwise Comparisons of Combined DNA Sequence from Five cpDNA Regions (Excluding *trnH-psbA*)

	Cal1202	Cal3114	Cri	Cgu	Cmi	Cmu	Lza1202	Lza2108	Ero	Bex	Cca
Cal1202	—	19	10	35	44	35	38	34	30	38	78
Cal3114	3	—	14	35	46	35	38	36	29	37	79
	0.00091										
Cri	9	6	—	31	42	30	33	32	25	33	74
	0.00273	0.00182									
Cgu	64	63	66	—	36	9	35	34	25	29	73
	0.01995	0.01963	0.02059								
Cmi	65	62	66	53	—	34	45	43	36	43	85
	0.02037	0.01941	0.02070	0.01643							
Cmu	60	59	63	8	51	—	36	36	29	31	74
	0.01868	0.01836	0.01965	0.00243	0.01580						
Lza1202	59	56	60	77	75	74	—	8	29	37	81
	0.01834	0.01739	0.01867	0.02417	0.02362	0.02318					
Lza2108	55	52	56	73	71	70	8	—	28	35	80
	0.01705	0.01610	0.01738	0.02285	0.02230	0.02187	0.00243				
Ero	32	29	33	48	47	44	44	40	—	26	77
	0.00981	0.00888	0.01013	0.01483	0.01459	0.01357	0.01357	0.01230			
Bex	54	51	55	61	61	57	73	69	43	—	78
	0.01678	0.01583	0.01711	0.01893	0.01906	0.01766	0.02289	0.02157	0.01328		
Cca	226	223	225	220	225	215	232	228	206	227	—
	0.07562	0.07453	0.07530	0.07332	0.07541	0.07158	0.07789	0.07634	0.06823	0.07599	

Table 4
Number of Indel Changes (Above Diagonal) and Number of Nucleotide Substitutions and Nucleotide Changes per Site (K81uf + G Model, Below Diagonal) from Pairwise Comparisons of Combined DNA Sequence from Six cpDNA Regions (Including *trnH-psbA*)

	Cal1202	Cal3114	Cri	Lza1202	Lza2108	Ero	Bex
Cal1202	—	23	14	52	47	43	50
Cal3114	7	—	15	52	50	42	51
	0.00213						
Cri	13	10	—	46	45	38	47
	0.00374	0.00266					
Lza1202	79	77	81	—	11	43	50
	0.02277	0.02241	0.02367				
Lza2108	75	73	77	17	—	41	49
	0.02146	0.02110	0.02235	0.00457			
Ero	43	40	44	70	66	—	35
	0.01188	0.01129	0.01245	0.01998	0.01870		
Bex	72	69	73	104	100	66	—
	0.02071	0.02008	0.02133	0.03108	0.02967	0.01888	

a molecular clock using *B. excelsa* as the outgroup (table 6). This suggests that the relative acceleration of nucleotide changes observed in *L. zabucajo* may be a relatively recent event because it is apparent when compared to a more recent ancestor (*B. excelsa*) but not when compared to a more distant ancestor (*C. cainito*) where the relative rate has been averaged over a longer time period.

Differences in the rates of evolutionary change at the intergenic regions examined would cause variation in the levels of divergence among cpDNA regions. To test the hypothesis of regional rate heterogeneity, we estimated pairwise nucleotide distances within chloroplast intergenic regions in the five and six region data sets and then plotted these distances (fig. 4). These calculations show that the range of distances varies with cpDNA region. In the five region data set, *trn*S-*trn*G showed the greatest divergences (up to about 12%) as well as the widest variability, whereas *atp*β-*rbc*L showed the smallest divergences and variability. In the six region data where fewer, more closely related taxa were compared, *trn*H-*psb*A showed divergences over 15% with a broad range of values, and the other five regions showed modest divergence (up to about 4%) and low variability. For *trn*H-*psb*A, intraspecific comparisons were 1% to 3% divergence, whereas

intergeneric divergence was frequently above 6%. The loci other than *trn*H-*psb*A showed fairly uniform divergence in the six region data set. Nucleotide divergence within the *trn*H-*psb*A region was slightly greater than that observed for the *trn*S-*trn*G region, even though the former was estimated from a set of more closely related taxa. Thus, the *trn*H-*psb*A region showed the highest evolutionary rate among the regions examined. These results were consistent with our inability to produce a multiple sequence alignment for all taxa for the *trn*H-*psb*A region, as well as with previous observations that the *trn*H-*psb*A region evolves rapidly (e.g., Aldrich et al. 1988).

Some authors observed that regions containing sites with more frequent GC neighboring bases, and therefore a higher overall GC content, evolve at higher rates (Morton and Clegg 1995; Morton 1995, 2000). A positive correlation between frequency of transversions and AT content has also been observed (Morton 1995, 2000). In the regions we studied, no correlation between overall GC content and average distance between sequences was observed, nor was there a significant correlation between average pairwise transversion/transition ratio and average overall AT content (M. Hamilton, D. F. Soria-Hernanz, and J. Braverman, unpublished data).

Table 5
Results for Tajima's 1D for Nucleotide or Indel Changes (Below Diagonal) and Tajima's 2D with Nucleotide and Indel Changes Combined as Distinct Classes of Data (Above Diagonal) Using the Five Region Data and *C. cainito* as the Outgroup

	Cal1202	Cal3114	Cri	Cgu	Cmi	Cmu	Lza1202	Lza2108	Ero	Bex
Cal1202	—								2D**	
Cal3114		—							2D**	
Cri			—						2D**	
Cgu				—	2D*					
Cmi				G*	—	2D*			2D**	
Cmu					G*	—	2D*			
Lza1202				G*	N*		—		2D***	
Lza2108								—	2D**	
Ero	N***	N**	N***	N*	N**		N***	N***	—	
Bex									N***	—

NOTE.—N indicates a rejection of the null for nucleotides only, G indicates a rejection of the null for indel changes only, and 2D indicates a rejection of the null for the joint outcome.
* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Table 6
Results for Tajima's 1D for Nucleotide or Indel Changes (Below Diagonal) and Tajima's 2D with Nucleotide and Indel Changes Combined as Distinct Classes of Data (Above Diagonal) Using the Six Region Data and *B. excelsa* as the Outgroup

	Cal1202	Cal3114	Cri	Lza1202	Lza2108	Ero
Cal1202	—			2D***	2D***	2D**
Cal3114		—		2D***	2D***	2D*
Cri			—	2D**	2D***	2D*
Lza1202	N***	N***	N***	—		
Lza2108	N***	N***	N**		—	2D***
Ero	G*	G*		N*** G*	N*** G*	—

NOTE.—N indicates a rejection of the null for nucleotides only, G indicates a rejection of the null for indel changes only, and 2D indicates a rejection of the null for the joint outcome.

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

The Tajima's 1D/2D tests were also used to determine if individual cpDNA regions were associated with any relative rate deviations observed. This test was conducted by applying the Tajima 1D/2D tests to both the five and six region data sets where one region had been deleted (analogous to a delete-one-region jackknife). The general patterns of deviations from the null hypothesis observed in the full data set (tables 5 and 6) were also observed in the delete-one-region data sets (results not shown). The same results were obtained with likelihood ratio tests of a molecular clock conducted with delete-one-region data sets (results not shown). This outcome suggests that neither the *trnS-trnG* nor the *trnH-psbA* region was individually causing the relative rate heterogeneity observed in *E. romeucardosoi* and *L. zabucajo* despite having larger genetic distances among taxa than other regions (see fig. 4). These results are consistent with deviations from the molecular clock not caused by processes specific to any cpDNA region such as selection, but rather caused by phenomena affecting the entire chloroplast genome such as a generation time effect within specific lineages (Gaut 1998; but see Whittle and Johnston 2003). This finding is consistent with previous evidence that synonymous substitutions in the chloroplast genome routinely demonstrate a generation time effect (Muse and Gaut 1997), suggesting that noncoding regions and synonymous changes may evolve under similar evolutionary processes. A generation time effect would not affect the employment of cpDNA regions in intraspecific population studies because the spatial patterns of haplotypes observed should still reflect gene flow events. A generation time effect may, however, cause some taxa to have more or less divergence among haplotypes to employ as genetic markers compared to related taxa.

Several statistical issues in our employment of Tajima's 1D/2D deserve further comment. Tables 5 and 6 contain a large number of non-independent pairwise comparisons of taxa. Thus, the Tajima 1D/2D results should be interpreted with the qualification that some of the significant results may be spurious because of the large number of tests conducted. For example, in table 5 a total of 150 tests are presented (50 indel 1D, 50 nucleotide 1D, and 50 2D). At a significance level of 0.05 we would expect 5 out of 100 tests to reject the null hypothesis (rate
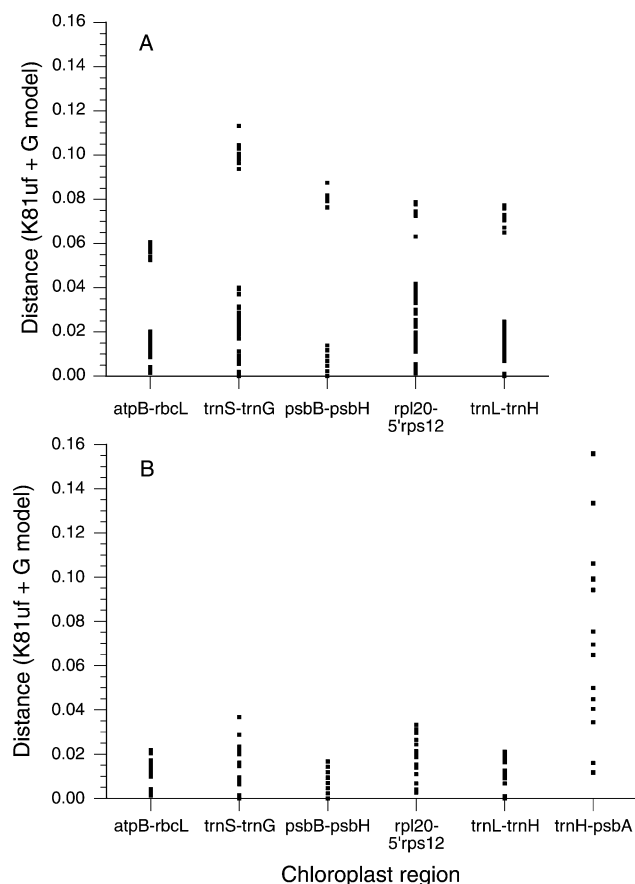

FIG. 4.—Plot of pairwise distances for nucleotide changes for all pairs of taxa estimated within chloroplast intergenic regions in the five (*A*) and six (*B*) region data sets. The distances were estimated by maximum likelihood based on the best-fit substitution model for the entire data set (F81uf + G) with base frequencies estimated for each region.

constancy) even when it is true if each comparison is independent (Rice 1989). Therefore, we expect at least 7 or 8 of the tests in table 5 to falsely reject the null hypothesis of rate constancy. Thus instead of placing a large amount of weight on particular significant cells in the table, we considered repeated rejection of the null hypothesis with data from particular lineages to be biologically significant. This is indicated by many significant cells along a row or down a column (e.g., see *E. romeucardosoi* [Ero] in table 5). A large number of comparisons was not as much an issue in table 6 because it had only 30 tests.

We find it unlikely that the statistical power of the Tajima 1D/2D was low, leading to an insensitive test of rate heterogeneity. Our use of a binomial exact test allowed inclusion of comparisons with few nucleotide or indel changes to achieve additional power. Tajima (1993) and Bromham et al. (2000) both looked at sequence length's influence on power, finding that longer sequences were critical in detection of rate variation. Our sequences were quite long, and therefore we expect high power, with the qualification that power obviously drops as regions are excluded to test for among-region rate variation. Both studies also noted that a distant outgroup lineage can reduce power. Because our *C. cainito* outgroup showed about 7% nucleotide divergence from the Lecythidaceae,

we also tested alternative outgroups with lower levels of nucleotide divergence. As described above, Tajima 1D/2D results were qualitatively identical regardless of outgroup. Although our tests suggest that these chloroplast indel and nucleotide changes are evolving in a neutral fashion, the lack of power might give false confidence that indels meet assumptions of population structure analyses. Yet the critical issue is expressed in the following question: is statistical power too low to detect rate variation which would bias estimates of population structure such as F_{ST}? In this case, the data that are available are not affected by non-neutral processes to a great enough extent to bias estimates of population structure. This view parallels the approach of Bromham et al. (2000), who focused specifically on the magnitude of error which would be introduced into the estimation of the date of a common ancestor by assuming the molecular clock when rates vary.

Our data were also collected to test the possibility that natural selection acts directly on observed indels (or nucleotide substitutions). We do not find strong evidence that rates of divergence for chloroplast regions among the Lecythidaceae sampled are heterogeneous with the exception of the *E. romeucardosoi* and *L. zabucajo* lineages. This is consistent with no net effect of selection on the rate of chloroplast indel or nucleotide evolution. However, these data cannot test the possibility that selection acting at linked sites in the genome influences the levels or geographic distribution of chloroplast polymorphism. This is so because selective sweeps are not expected to alter the absolute degree of neutral divergence between populations or species because even complete linkage to advantageous or deleterious mutations does not affect the neutral substitution rate (Birky and Walsh 1988; Charlesworth 1998). Testing for the influence of selection at linked sites on levels of indel or nucleotide polymorphism within *C. alta* will require additional haplotype polymorphism data.

Conclusions

These data suggest that rates of indel and nucleotide evolution vary among the six cpDNA regions studied and that nucleotide divergence for two of these regions exceeds 12% within a single plant family. Homoplasy was evident for indel changes, although the saturation effect was modest and seen most markedly in the divergence between the Lecythidaceae and Sapotaceae. There was no evidence for indel homoplasy within the Lecythidaceae. The data provide little evidence for relative rate differences consistent with natural selection acting on indel changes. What evidence we did find to reject the molecular clock for these cpDNA regions comes mostly from relative rates of nucleotide change and less frequently from relative rates of indel change. In the cases where the molecular clock was rejected, the relative rate differences were most consistent with a generation time effect because the relative rate heterogeneity is not associated with individual cpDNA regions, but rather was a constant across all regions. If the generation time effect observed here is found more broadly, it may help explain variability among closely related species in the amount of chloroplast haplotype variation observed in population studies. These results

suggest that the spatial pattern of cpDNA haplotypes in *C. alta* are not influenced by homoplasy or selection acting directly on the alternative haplotypes observed.

Supplementary Materials

Available online are the five and six region multiple sequence alignments as Nexus files containing either base pair data only or combined indel characters (with transition matrices) and base pair data. Also available are the nexus files used to specify the substitution models in PAUP*, and the program T1Dand2D and associated Excel files. These materials are also available at http://bioserver. georgetown.edu/faculty/hamilton under the "downloads" link.

Literature Cited

Aldrich, J. A., B. W. Cherney, E. Merlin, and L. Christopherson. 1988. The role of insertions/deletions in the evolution of the intergenic region between *psb*A and *trn*H in the chloroplast genome. Curr. Genet. **14**:137–146.

Avise, J. C. 1991. Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochrondrial DNA. Annu. Rev. Genet. **25**:45–69.

———. 1994. Molecular markers, natural history and evolution. Chapman and Hall, New York.

Ballard, J. W. O., and M. Kreitman. 1995. Is mitochondrial DNA a strictly neutral marker? Trends Ecol. Evol. **10**:485–488.

Birky, C. W., P. Fuerst, and T. Maruyama. 1989. Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. Genetics **121**:613–627.

Birky, C. W., and J. B. Walsh. 1988. Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85**:6414–6418.

Bromham, L. D., D. Penny, A. Rambaut, and M. Hendy. 2000. The power of relative rates tests depends on the data. J. Mol. Evol. **50**:296–301.

Caron, H., S. Dumas, G. Marque, C. Messier, E. Bandou, R. J. Petit, and A. Kremer. 2000. Spatial and temporal distribution of chloroplast DNA polymorphism in a tropical tree species. Mol. Ecol. **9**:1089–1098.

Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15**:538–543.

Clegg, M. T., B. S. Gaut, G. H. Learn, and B. R. Morton. 1994. Rates and patterns of chloroplast DNA evolution. Proc. Natl. Acad. Sci. USA **91**:6795–6801.

Curtis, S. E., and M. T. Clegg. 1984. Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. **1**:291–301.

Demesure, B., N. Sodzi, and R. J. Petit. 1995. A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol. Ecol. **4**:129–131.

Desplanque, B., F. Viard, J. Bernard, D. Forcioli, P. Saumitou-Laprade, J. Cuguen, and H. van Dijk. 2000. The linkage disequilibrium between chloroplast DNA and mitochondrial DNA haplotypes in *Beta vulgaris* ssp *maritima* (L.): the usefulness of both genomes for population genetic studies. Mol. Ecol. **9**:141–154.

Doyle, J. J., M. Morgante, S. V. Tingey, and W. Powell. 1998. Size homoplasy in chloroplast microsatellites of wild perennial relatives of soybean (*Glycine* subgenus *Glycine*). Mol. Biol. Evol. **15**:215–218.

Dumolin-Lapégue, S., M.-H. Pemonge, and R. J. Petit. 1997. An enlarged set of consensus primers for the study of organelle DNA in plants. Mol. Ecol. **6**:393–397.

———. 1998. Association between chloroplast and mitochondrial lineages in Oaks. Mol. Biol. Evol. **15**:1321–1331.

Dutech, C., L. Maggia, and H. I. Joly. 2000. Chloroplast diversity in *Vouacapoua americana* (Caesalpiniaceae), a neotropical forest tree. Mol. Ecol. **9**:1427–1432.

Ennos, R. A., W. T. Sinclair, X.-S. Hu, and A. Langdon. 1999. Using organelle markers to elucidate the history, ecology and evolution of plant populations. Pp. 1–19 *in* P. M. Hollingsworth, R. M. Bateman, and R. J. Gornall, eds. Molecular systematics and plant evolution. Taylor and Francis Ltd., London.

Gascon, C., and R. O. Bierregaard. 2001. The Biological Dynamics of Forest Fragments Project: the study site, experimental design, and research activity. Pp. 31–42 *in* R. O. Bierregaard, C. Gascon, T. E. Lovejoy, and R. Mesquita, eds. Lessons from Amazonia: the ecology and conservation of a fragmented forest. Yale University Press, New Haven.

Gaut, B. S. 1998. Molecular clocks and nucleotide substitution rates in higher plants. Evol. Biol. **30**:93–120.

Gaut, B. S., L. G. Clark, J. F. Wendel, and S. V. Muse. 1997. Comparisons of the molecular evolutionary process at *rbc*L and *ndh*F in the grass family (Poaceae). Mol. Biol. Evol. **14**:769–777.

Goldstein, D. B., and D. D. Pollock. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. J. Hered. **88**:335–342.

Golenberg, E. M., M. T. Clegg, M. L. Durbin, J. Doebley, et al. 1993. Evolution of a noncoding region of the chloroplast genome. Mol. Phylogenet. Evol. **2**:52–64.

Graham, S., W. Reeves, P. A. Burns, and A. C. E. Olmstead. 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. Int. J. Plant Sci. **161**:S83–S96.

Hamilton, M. B. 1999*a*. Tropical tree gene flow and seed dispersal. Nature **401**:129–130.

———. 1999*b*. Four primer pairs for the amplification of chloroplast intergenic regions with intraspecific variation. Mol. Ecol. **8**:521–523.

Hamilton, M. B., and J. M. Miller. 2002. Comparing relative rates of pollen and seed gene flow in the island model using nuclear and organelle measures of population structure. Genetics **162**:1897–1909.

Hedrick, P. W. 1999. Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution **53**:313–318.

Huelsenbeck, J. B., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.

Jarne, P., and P. J. L. Lagoda. 1996. Microsatellites, from molecules to populations and back. Trends Ecol. Evol. **11**:424–429.

Kelchner, S. A. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. Missouri Bot. Gard. **87**:482–498.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458

McCauley, D. E. 1994. Contrasting the distribution of chloroplast DNA and allozyme polymorphism among local populations of *Silene alba*: implications for studies of gene flow in plants. Proc. Natl. Acad. Sci. USA **91**:8127–8131.

———. 1995. The use of cloroplast DNA polymorphism in studies of gene flow in plants. Trends Ecol. Evol. **10**:198–202.

———. 1998. The genetic structure of a gynodioecious plant: nuclear and cytoplasmic genes. Evolution **52**:255–260.

McGuire, G., M. C. Denham, and D. J. Balding. 2001. Models of sequence evolution for DNA sequences containing gaps. Mol. Biol. Evol. **18**:481–490.

Mori, S. A., P. Becker, and D. Kincaid. 2001. Lecythidaceae of a central Amazonian lowland forest. Pp. 54–67 *in* R. O. Bierregaard, C. Gascon, T. E. Lovejoy, R. Mesquita, eds. Lessons from Amazonia: the ecology and conservation of a fragmented forest. Yale University Press, New Haven, Conn.

Morton, B. R. 1995. Neighboring base composition and transversion/transition bias in a comparison of rice and maize chloroplast noncoding regions. Proc. Natl. Acad. Sci. USA **92**:9717–9721.

———. 2000. Codon bias and the context dependency of nucleotide substitutions in the evolution of plastid DNA. Evol. Biol. **31**:55–103.

Morton, B. R., and M. T. Clegg. 1995. Neighboring base composition in strongly correlated with base substitution in a region of the chloroplast genome. J. Mol. Evol. **41**:597–603.

Morton, C. M., S. A. Mori, G. T. Prance, K. G. Karol, and M. W. Chase. 1997. Phylogenetic relationships of Lecythidaceae: a cladistic analysis using *rbc*L sequence and morphological data. Am. J. Bot. **84**:530–540.

Muloko-Ntoutoume, N., R. J. Petit, L. White, and K. Abernethy. 2000. Chloroplast DNA variation in a rainforest tree (*Aucoumea klaineana*, Burseraceae) in Gabon. Mol. Ecol. **9**:359–363.

Muse, S. V. 2000. Examining rates and patterns of nucleotide substitution in plants. Plant Mol. Biol. **42**:25–43.

Muse, S. V., and B. S. Gaut. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. Genetics **146**:393–399.

Muse, S. V., and B. S. Weir. 1992. Testing for equality of evolutionary rates. Genetics **132**:269–276.

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics. Oxford University Press, Oxford, U.K.

Oddou-Muratorio, S., R. J. Petit, B. Le Guerroue, D. Guesnet, and B. Demesure. 2001. Pollen- versus seed-mediated gene flow in a scattered forest tree species. Evolution **55**:1123–1135

Palmer, J. D. 1990. Contrasting modes and tempos of genome evolution in land plant organelles. Trends Genet. **6**:115–20.

Parker, P. G., A. A. Snow, M. D. Schug, G. C. Booton, and P. A. Fuerst. 1998. What molecules can tell us about populations: choosing and using a molecular marker. Ecology **79**:361–382.

Petit, R. J., A. Kremer, and D. B. Wagner. 1993. Finite island model for organelle and nuclear genes in plants. Heredity **71**:630–641.

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14**:817–818.

———. 2001. Selecting models of nucleotide substitution: an application to human immunodeficiency virus (HIV-1). Mol. Biol. Evol. **18**:897–906.

Powell, W., M. Morgante, R. McDevitt, G. G. Vendramin, and J. A. Rafalski. 1995. Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. Proc. Natl. Acad. Sci. USA **92**:7759–7763.

Provan J., W. Powell, and P. M. Hollingsworth. 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol. Evol. **16**:142–147.

Rand, D. M. 1993. Endotherms, ectotherms, and mitochondrial genome-size variation. J. Mol. Evol. **37**:281–295.

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution **43**:223–225.

Saitou, N., and S. Ueda. 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. **11**:504–512.

Sarich, V. M., and A. C. Wilson. 1973. Generation time and genomic evolution in primates. Science **179**:1144–1147.

Savolainen, V., M. W. Chase, S. B. Hoot, C. M. Morton, D. E. Soltis, C. Bayer, M. F. Fay, A. Y. de Bruijn, S. Sullivan, and Y.-L. Qiu. 2000. Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences. Systemat. Biol. **49**:306–362.

Savolainen, V., J. F. Manen, E. Douzery, and R. Spichiger. 1994. Molecular phylogeny of families related to Celastrales based on rbcL 5′ flanking sequences. Mol. Phylogenet. Evol. **3**: 27–37.

Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. **49**:369–381.

Stephan, W., L. Xing, D. A. Kirby, and J. M. Braverman. 1998. A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc. Natl. Acad. Sci. USA **95**:5649–5654.

Swofford, D. L. 2002. PAUP*. phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular systematics. 2nd Edition. Sinauer Associates, Sunderland, Mass.

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

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

Tajima, F., and M. Nei. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

Thorne, J. L., H. Kishinmo, and J. Felsenstein. 1992. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. **34**:3–16.

Whittle, C.-A., and M. O. Johnston. 2003. Broad-scale analysis contradicts the theory that generation time affects molecular evolutionary rates in plants. J. Mol. Evol. **56**: 223–233.