American Journal of Botany 94(3): 275–288. 2007.

Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III[1]

Joey Shaw,[2,3] Edgar B. Lickey,[3] Edward E. Schilling,[3] and Randall L. Small[3]

[2]Department of Biological and Environmental Sciences, 615 McCallie Avenue, University of Tennessee, Chattanooga, Tennessee 37403 USA; and
[3]Department of Ecology and Evolutionary Biology, 442 Hesler Biology, University of Tennessee, Knoxville, Tennessee 37996 USA

Although the chloroplast genome contains many noncoding regions, relatively few have been exploited for interspecific phylogenetic and intraspecific phylogeographic studies. In our recent evaluation of the phylogenetic utility of 21 noncoding chloroplast regions, we found the most widely used noncoding regions are among the least variable, but the more variable regions have rarely been employed. That study led us to conclude that there may be unexplored regions of the chloroplast genome that have even higher relative levels of variability. To explore the potential variability of previously unexplored regions, we compared three pairs of single-copy chloroplast genome sequences in three disparate angiosperm lineages: *Atropa* vs. *Nicotiana* (asterids); *Lotus* vs. *Medicago* (rosids); and *Saccharum* vs. *Oryza* (monocots). These three separate sequence alignments highlighted 13 mutational hotspots that may be more variable than the best regions of our former study. These 13 regions were then selected for a more detailed analysis. Here we show that nine of these newly explored regions (*rpl32-trnL*$^{(\mathrm{UAG})}$, *trnQ*$^{(\mathrm{UUG})}$*-5′rps16*, *3′trnV*$^{(\mathrm{UAC})}$*-ndhC*, *ndhF-rpl32*, *psbD-trnT*$^{(\mathrm{GGU})}$, *psbJ-petA*, *3′rps16–5′trnK*$^{(\mathrm{UUU})}$, *atpI-atpH*, and *petL-psbE*) offer levels of variation better than the best regions identified in our earlier study and are therefore likely to be the best choices for molecular studies at low taxonomic levels.

Key words: chloroplast marker; cpDNA; molecular systematics; noncoding chloroplast DNA; phylogeny; phylogeography.

Noncoding sequences of the chloroplast genome are a primary source of data for molecular systematic, phylogeographic, and population genetic studies of plants, yet relatively little is known about levels of variation among different noncoding regions of the chloroplast genome. Because of the lack of a comprehensive comparison of the different noncoding portions of the chloroplast genome, little is known about the different utilities of the many potential chloroplast markers. In a previous study (Shaw et al., 2005), we compared the relative levels of variability among 21 commonly employed noncoding chloroplast DNA regions. In this study we assessed the relative levels of variability among *all* noncoding regions of the single-copy portions of the chloroplast genome and directly compared those regions that appear to be the most variable to the standard of our former study.

Chloroplast genomes typically range in size from 120 to 170 kilobase pairs (kb), and there is a relatively high degree of conservation in size, structure, gene content, and linear order of the genes in land plants (for a more detailed discussion, see Downie and Palmer, 1992). With few exceptions, the chloroplast genome contains two inverted repeats (approximately 25 kb each) that are mirror images of one another in terms of gene complement. The inverted repeats are separated from each other by one large and one small single-copy region (LSC and SSC, respectively). Previous studies have suggested that the inverted repeat regions accumulate point mutations slower than the single-copy regions (Curtis and Clegg, 1984; Wolfe et al., 1987; Wolfe, 1991; Gaut, 1998). Perry and Wolfe (2002) showed that the nucleotide substitution rate is 2.3 times higher in the single-copy regions relative to the inverted repeats. Because the inverted repeats evolve at a relatively slower rate and the first adopted chloroplast regions (e.g., *rbcL*, *atpB*, *trnL-trnL-trnF*) are located in the LSC, most plant researchers using molecular tools have focused on the single-copy regions.

The chloroplast genome can be divided into three functional categories including (1) protein-coding genes, (2) introns, and (3) intergenic spacers; the latter two do not encode proteins and are referred to as noncoding regions. According to the *Nicotiana* chloroplast map (Wakasugi et al., 1998), approximately 43% of the LSC and SSC is noncoding. Fifteen introns make up approximately 10.6% of the single-copy chloroplast DNA, while 92 intergenic spacers comprise 32.3%.

The use of noncoding chloroplast DNA sequences to generate plant phylogenies began in the early 1990s with the seminal publications of Taberlet et al. (1991), Clegg et al. (1994), Morton and Clegg (1993), and Gielly and Taberlet (1994). These studies were facilitated by the three chloroplast genomes that had been completely sequenced (*Marchantia polymorpha*, *Nicotiana tabacum*, and *Oryza sativa*). The field of plant molecular systematics has made great strides in the last several years with over 50 completely sequenced land plant chloroplast genomes and several more on the horizon; furthermore, there are scores of published plant phylogenies of all different ranks, and the line between phylogenetics, population genetics, and phylogeography has become increasingly thinner. However, despite more than 15 years of use of noncoding cpDNA sequences in molecular systematic research on plants (since Taberlet et al., 1991), relatively few noncoding chloroplast DNA regions have been directly compared in sequence-based investigations. To date, we really have not taken advantage of what Olmstead and Palmer (1994, p. 81) referred to as "the greatest advantage of DNA sequencing," which is the phylogenetic breadth to which sequence data can be applied—because of differing evolutionary rates among different portions of the genome.

[1] Manuscript received 5 July 2006; revision accepted 20 December 2006.

The authors acknowledge funding for this project from the Hesler Endowment Fund from the University of Tennessee Herbarium; the Department of Botany, University of Tennessee, Knoxville; the National Science Foundation; and start-up funds for the primary author from the University of Tennessee, Chattanooga.

[2] Author for correspondence (e-mail: joey-shaw@utc.edu)

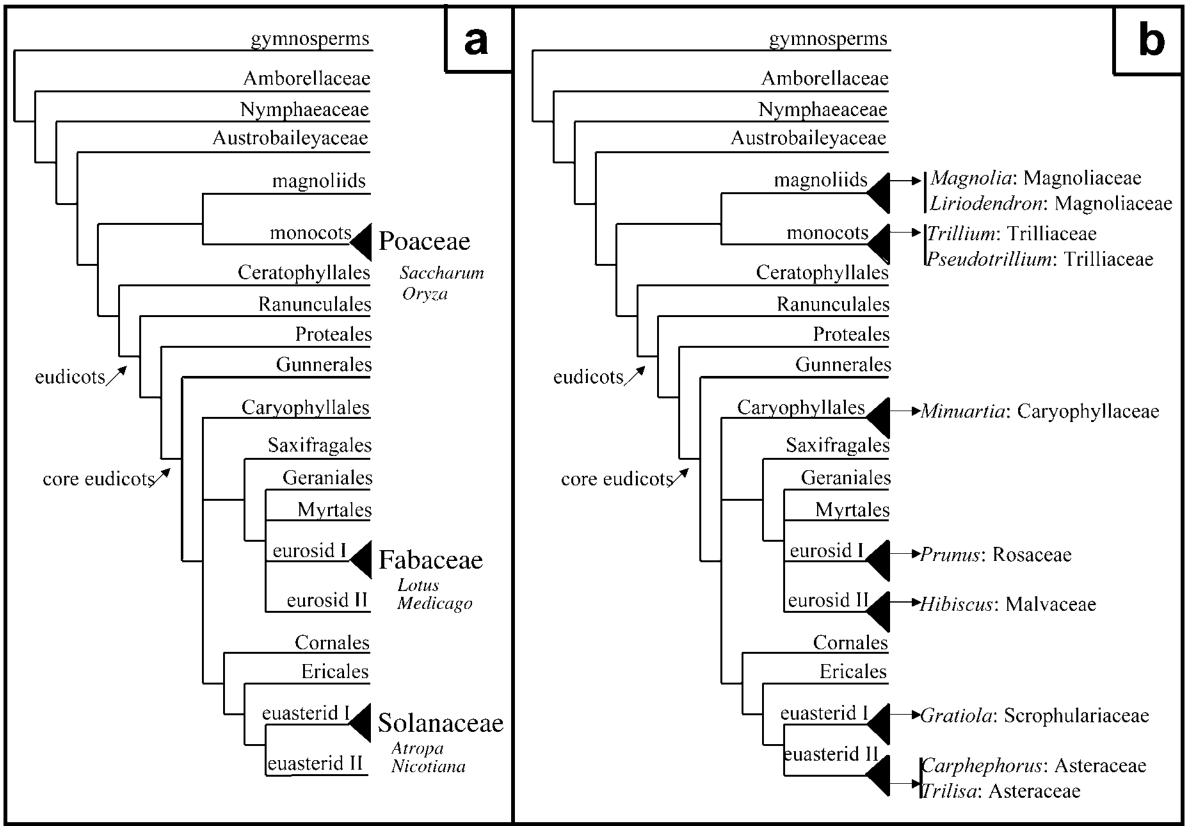


Fig. 1. Simplified phylogenetic representation, modified from APG II (2003), of (a) the three lineages whose complete sequences were directly compared in the initial screening of the genomes and (b) the seven lineages used in the direct comparison of 34 noncoding chloroplast regions in angiosperms.

In a previous study (Shaw et al., 2005), we evaluated the relative phylogenetic utility of 21 noncoding chloroplast regions using 10 lineages that span the phylogenetic breadth of seed plants, focusing on those regions that had previously been successfully used in species-level phylogenetic or population-level studies. We showed that (1) there is predictable variation in the levels of variability between the different noncoding regions, and (2) the more variable noncoding regions have rarely been employed, while the most widely used regions are among the least variable. That study led us to suspect that unexplored, highly variable regions of the chloroplast genome likely exist.

The discovery (or documentation) of additional noncoding plastid regions of predictably high variability is of great utility. Sequence data from such regions have numerous, important applications in systematics and evolutionary biology such as elucidating the origin of domesticated species (Wills and Burke, 2006), tracing biogeographic movements (Ickert-Bond and Wen, 2006; Schönswetter et al., 2006a, b), and clarifying complex relationships among species (Shaw and Small, 2005). Fast evolving plastid regions are also useful for species identification via molecular barcoding or microarray analysis. Databased sequences of these rapidly evolving regions are also lending themselves to studies of simulating sequence evolution (Cartwright, 2005), and they may eventually lead to a better understanding of the functions of noncoding DNA or the mechanisms for cpDNA evolution based on comparative frequency of various types of mutational events.

To target previously unexplored noncoding cpDNA regions, we compared three pairs of published whole chloroplast genome sequences from three disparate angiosperm lineages (Fig. 1a): *Atropa* vs. *Nicotiana* (Solanaceae, asterid); *Lotus* vs. *Medicago* (Fabaceae, rosid); and *Saccharum* vs. *Oryza* (Poaceae, monocot). The single-copy regions of each of these related pairs were aligned to aid in the detection of potential mutational hotspots. Using the best regions of Shaw et al. (2005) as a baseline, we identified 13 noncoding regions to evaluate more thoroughly. Here we show that at least nine newly explored regions offer levels of variation higher than those of the most variable regions identified in our earlier study (Shaw et al., 2005), thus providing the plant systematics, population genetics, and phylogeographic communities with several additional quickly evolving noncoding chloroplast markers from which to choose.

MATERIALS AND METHODS

Taxonomic sampling—Species and lineages sampled in this study are shown in Fig. 1b and listed in Appendix 1. They are the same accessions and DNA stocks used in our previous study (Shaw et al., 2005), minus one lineage each from the two asterid lineages and the gymnosperm representative *Taxodium*/*Glyptostrobus*/*Cryptomeria*. Sampling focused on representing all major angiosperm lineages sensu APG II (2003) (Fig. 1b) in addition to representing different habits and life strategies (e.g., woody perennials [*Magnolia* and *Prunus*], herbaceous perennials [*Carphephorus*, *Trillium*, and *Hibiscus*], and herbaceous annuals [*Gratiola* and *Minuartia*]). Three fairly closely related species (referred to as a

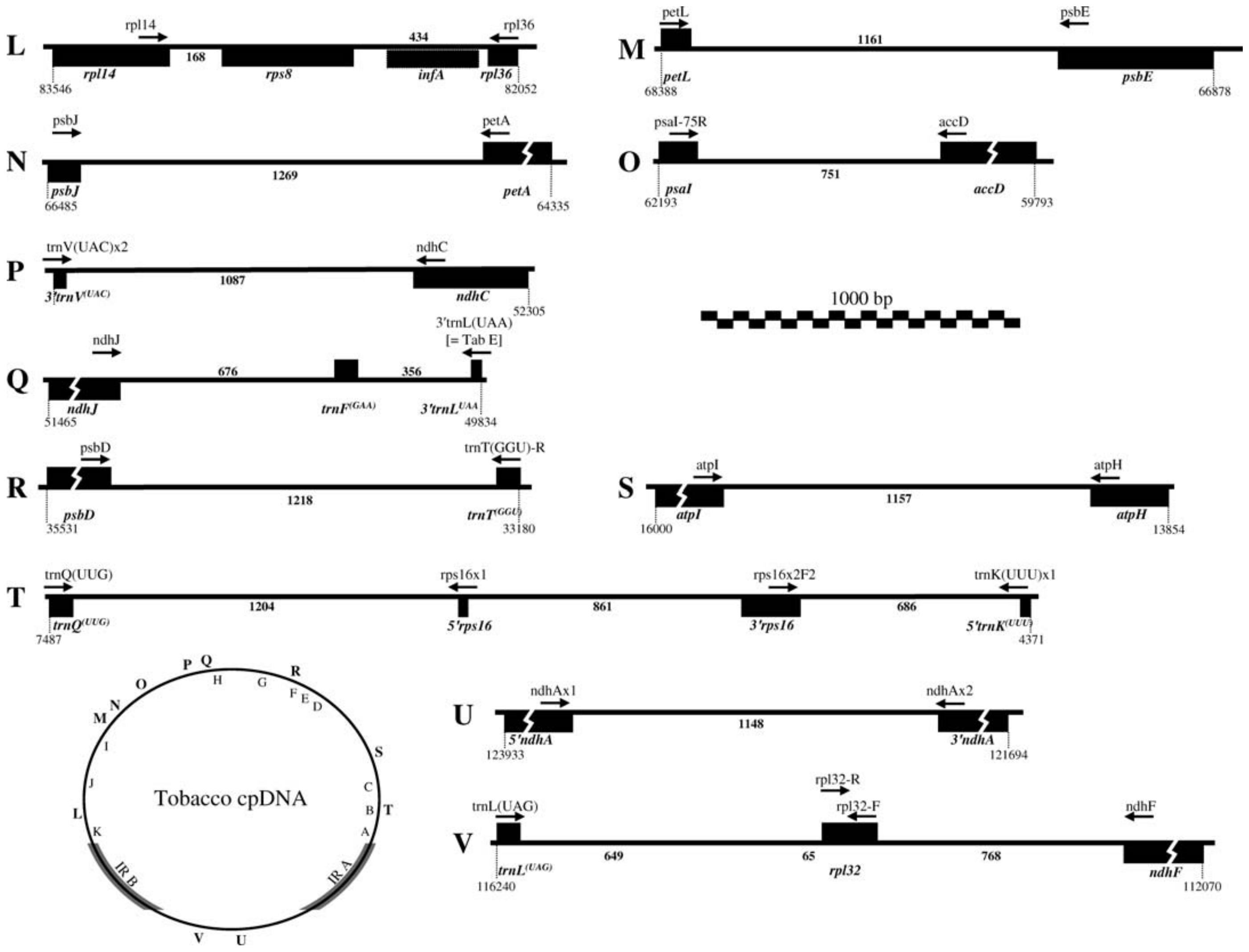


Fig. 2. Scaled map of the 13 noncoding cpDNA regions surveyed in this investigation (based on the *Nicotiana* chloroplast genome [Wakasugi et al., 1998]). The orientation and relative positions of the genes are identified (L–V); the relative positions of the 21 regions of Shaw et al. (2005) are identified (A–K). Specific positions of each of the 13 regions of this study are denoted by offset numbers at the beginning and end of each region. Gene names are italicized below and amplification and sequencing primer names are in roman typeface above with directional arrows. Lengths of noncoding regions are centered below each intergenic spacer and intron.

"three-species group") were chosen within each of the seven angiosperm lineages. Within each lineage, two species were chosen to represent ingroup taxa of different clades, while the third was chosen as a closely related outgroup taxon. Voucher information and GenBank accession numbers for each taxon–cpDNA region combination are listed in Appendix 1.

Identifying previously unexplored cpDNA regions—Because the focus of this investigation was to identify rapidly evolving regions of the chloroplast genome and the inverted repeats have been shown to accumulate mutations at a slower rate than the single-copy regions (Perry and Wolfe, 2002), our effort focused on the single-copy regions of the chloroplast genome. Both the LSC and the SSC portions of the genomes were compared between two taxa from each of three disparate angiosperm lineages (Fig. 1a): *Atropa* vs. *Nicotiana* (Solanaceae, asterid); *Lotus* vs. *Medicago* (Fabaceae, rosid); and *Saccharum* vs. *Oryza* (Poaceae, monocot). For each related pair of chloroplast sequences (e.g., *Atropa* vs. *Nicotiana*) ClustalX (Thompson et al., 2001) was used to align the LSC and SSC regions. Then the pairs of sequences were manually adjusted in MacClade version 4.06 (Sinauer, Sunderland, Massachusetts, USA), and the number of variable sites for each noncoding region were tabulated. Calculating the number of mutations observed within each noncoding region of a related species pair was done for all noncoding cpDNA regions, including those regions surveyed in our earlier investigation (Shaw et al., 2005). In so doing we were able to use the most variable regions from Shaw et al. (2005) to set a baseline to help us identify regions that may offer an even greater number of variable sites than the most variable regions identified in that earlier study. On the basis of the number of variable sites in these comparisons, we selected 13 noncoding regions to evaluate by adding to our existing data set. All regions from both this and our last study are listed here as they occur on the Wakasugi et al. (1998) *Nicotiana* cpDNA map starting at the junction of Inverted Repeat A (the newly explored regions of this study are shown in boldface type, and they and their primer sites are mapped in Fig. 2): *rpl16* intron, ***rpl14-rps8-infA-rpl36***, *psbB-psbH*, *5′rps12-rpl20*, ***petL-psbE***, ***psbJ-petA***, ***psaI-accD***, **3′*trnV*$^{(UAC)}$-*ndhC***, ***ndhJ-trnF*$^{(GAA)}$**, *trnL*$^{(UAA)}$-*trnF*$^{(GAA)}$, *trnL*$^{(UAA)}$ intron, *trnT*$^{(UGU)}$-*trnL*$^{(UAA)}$, *rps4-trnT*$^{(UGU)}$, *trnS*$^{(GGA)}$-*rps4*, *trnS*$^{(UGA)}$-*trnfM*$^{(CAU)}$, ***psbD-trnT*$^{(GGU)}$**, *trnD*$^{(GUC)}$-*trnT*$^{(GGU)}$, *psbM-trnD*$^{(GUC)}$, *ycf6-psbM*, *trnC*$^{(GCA)}$-*ycf6*, *rpoB-trnC*$^{(GCA)}$, ***atpI-atpH***, *trnG*$^{(UCC)}$ intron, *trnS*$^{(GCU)}$-*trnG*$^{(UCC)}$, ***trnQ*$^{(UUG)}$- 5′*rps16***, *rps16* intron, **3′*rps16*-5′*trnK*$^{(UUU)}$**, *matK*-5′*trnK*$^{(UUU)}$, 3′*trnK*$^{(UUU)}$-*matK*, *psbA*-3′*trnK*$^{(UUU)}$, *trnH*$^{(GUG)}$-*psbA*, ***ndhA* intron**, ***ndhF-rpl32***, and ***rpl32-trnL*$^{(UAG)}$**. The last three regions are from the SSC region.

TABLE 1. Sequences of primers used for PCR amplification and sequencing.

Fig. 2 code	Region	Primer name and sequence (5′-3′)
L	*rpl14-rps8-infA-rpl36*	**rpL14**: AAG GAA ATC CAA AAG GAA CTC G **rpL36**: GGR TTG GAA CAA ATT ACT ATA ATT CG
M	*petL-psbE*	**petL**: AGT AGA AAA CCG AAA TAA CTA GTT A **psbE**: TAT CGA ATA CTG GTA ATA ATA TCA GC
N	*psbJ-petA*	**psbJ**: ATA GGT ACT GTA RCY GGT ATT **petA**: AAC ART TYG ARA AGG TTC AAT T
O	*psaI-accD*	**accD**: AAT YGT ACC ACG TAA TCY TTT AAA **psaI-75R**: AGA AGC CAT TGC AAT TGC CGG AAA
P	*3′trnV-ndhC*	**trnV$^{(UAC)}$x2**: GTC TAC GGT TCG ART CCG TA **ndhC**: TAT TAT TAG AAA TGY CCA RAA AAT ATC ATA TTC
Q	*ndhJ-trnF*	**ndhJ**: ATG CCY GAA AGT TGG ATA GG **TabE**: GGT TCA AGT CCC TCT ATC CC (Taberlet et al., 1991)
R	*psbD-trnT*	**psbD**: CTC CGT ARC CAG TCA TCC ATA **trnT$^{(GGU)}$-R**: CCC TTT TAA CTC AGT GGT AG
S	*atpI-atpH*	**atpI**: TAT TTA CAA GYG GTA TTC AAG CT **atpH**: CCA AYC CAG CAG CAA TAA C
T	*trnQ-5′rps16*	**trnQ$^{(UUG)}$**: GCG TGG CCA AGY GGT AAG GC **rpS16x1**: GTT GCT TTY TAC CAC ATC GTT T
T	*3′rps16-5′trnK*	**rpS16x2F2**: AAA GTG GGT TTT TAT GAT CC **trnK$^{(UUU)}$x1**: TTA AAA GCC GAG TAC TCT ACC
U	*ndhA* intron	**ndhAx1**: GCY CAA TCW ATT AGT TAT GAA ATA CC **ndhAx2**: GGT TGA CGC CAM ARA TTC CA
V	*ndhF-rpl32*	**rpL32-R**: CCA ATA TCC CTT YYT TTT CCA A **ndhF**: GAA AGG TAT KAT CCA YGM ATA TT
V	*rpl32-trnL*	**trnL$^{(UAG)}$**: CTG CTT CCT AAG AGC AGC GT **rpL32-F**: CAG TTC CAA AA A AAC GTA CTT C
C	*trnS-trnG-trnG*	**trnG$^{(UUC)}$***: GAA TCG AAC CCG CAT CGT TAG **trnS$^{(GCU)}$***: AAC TCG TAC AAC GGA TTA GCA ATC

Molecular techniques—Because the genes surrounding noncoding regions are relatively conserved across seed plants (and especially within angiosperms), all the polymerase chain reaction (PCR) primers for amplification and sequencing can be used across the diverse taxonomic groups of this study. Nearly all the primers used here were created for this study, although a few of the regions have been tried elsewhere (see Discussion of the previously unexplored regions). Alignment of GenBank sequences from a wide array of angiosperm lineages was used to create angiosperm-universal primers, and this study is demonstrative of their "universality."

DNA was extracted from leaf tissue using either the DNeasy Plant Mini Kit (Qiagen, Valencia, California, USA) or the CTAB method (Doyle and Doyle, 1987). The polymerase chain reaction (PCR) was performed using Eppendorf (Westbury, New York, USA) Mastercycler gradient or Mastercycler gradient ep thermal cyclers in 25-μL volumes with the following reaction components: 1 μL template DNA (~10–100 ng), 1× *rTaq* buffer (PanVera/TaKaRa, Madison, Wisconsin, USA), 200 μmol/L each dNTP, 3.0 mmol/L $MgCl_2$, 0.2 μg/μL bovine serum albumin, 0.1 μmol/L each primer, and 1.25 units *rTaq* (PanVera/TaKaRa).

Primer sequences for each region are presented in Table 1, and their relative positions and orientations are illustrated in Fig. 2, which parallels a similar figure (fig. 3) from Shaw et al. (2005). In Fig. 2 the letters A–K around the inside of the chloroplast map represent regions from our earlier study, while letters L–T around the outside represent the newly surveyed regions. In an attempt to simplify amplification parameters throughout this study, and unless otherwise indicated later, a single PCR program, the "*rpl16*" program of Shaw et al. (2005), was used for all cpDNA regions surveyed here because it is "slow and cold" and has proven to be effective across a wide range of taxa and genomic regions. The PCR cycling conditions were template denaturation at 80°C for 5 min followed by 30 cycles of denaturation at 95°C for 1 min, primer annealing at 50°C for 1 min, followed by a ramp of 0.3°C/s to 65°C, and primer extension at 65°C for 4 min; followed by a final extension step of 5 min at 65°C.

PCR products were checked on 1% agarose gels before being cleaned with ExoSAP-IT (USB, Cleveland, Ohio, USA). All DNA sequencing was performed with the ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction kit, v. 3.1 (Perkin-Elmer/Applied Biosystems, Foster City, California, USA) and electrophoresed and detected on an ABI Prism 3100 automated sequencer (University of Tennessee Molecular Biology Resource Facility, Knoxville, Tennessee, USA).

As mentioned, we simplified the PCR amplification parameters by using the same "*rpl16*" PCR program for all regions of this study. However, this protocol yielded a multibanded product in the *Hibiscus* (eurosid II) lineage when the *ndhA* intron was amplified. Therefore, the PCR cycling conditions for the *ndhA* intron in the *Hibiscus* lineage were 35 cycles of denaturation at 94°C for 30 s, primer annealing at 55°C for 30 s, and primer extension at 72°C for 2 min.

trnS$^{(GCU)}$-trnG$^{(UCC)}$-trnG$^{(UCC)}$—The previously published primers (Shaw et al., 2005) for this region have proven to be troublesome for several groups (Asclepiadaceae: M. Fishbein, Portland State University, personal communication; Illiciaceae: A. Morris, University of Florida, personal communication; Oxalidaceae: E. Emshwiller, The Field Museum of Natural History, personal communication). In that study (Shaw et al., 2005), we had to develop a strict protocol for amplification. Therefore, we are here publishing a new set of primers for this region that have been tested across many angiosperm families. The new primers are listed in Table 1. Internal sequencing primers published in Shaw et al. (2005) can also be used on the amplicons generated from these primers.

Comparison of the previously unexplored cpDNA regions—The program Sequencher 4.2.1 (Gene Codes Corp., Ann Arbor, Michigan, USA) was used to compile contiguous sequences (contigs) of each accession from electropherograms generated on the automated sequencer. The DNA sequences of each three-species group were initially aligned with ClustalX (Thompson et al., 2001) and subsequently manually adjusted by eye in MacClade v. 4.06 (Sinauer, Sunderland, Massachusetts, USA). Variable positions in the three-species group data matrix were double checked against the original chromatogram files to make sure that all base calls were true at all variable positions. In a few cases, alignment of potentially informative positions was ambiguous owing to mononucleotide runs or repeated motifs; where we deemed appropriate, one or a few indels were inserted in an attempt to conservatively score these regions (rather than calling each potentially "misaligned" base a potentially informative character). Positions of coding and noncoding (gene, exon, and intron) borders were determined by sequence comparison with *Arabidopsis* (NC 000932), *Lotus* (NC 001874), or *Nicotiana* (NC 002694) entire cpDNA sequences in GenBank. Terminal coding regions and, in a few rare cases, short, unreadable ends of the noncoding portions of the PCR amplicons were excluded from the contigs. Alignments are available upon request from J.S. or R.L.S.

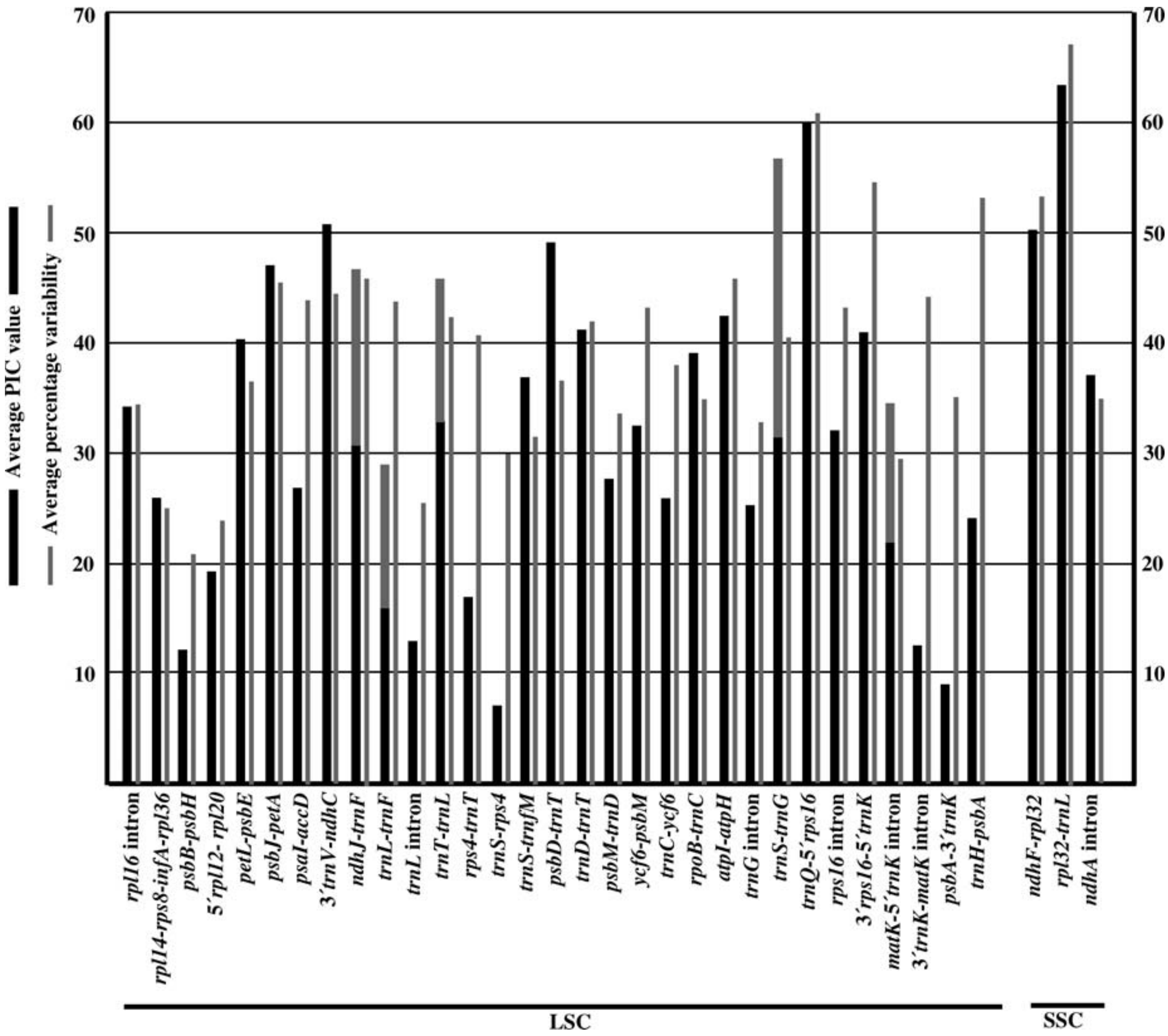


Fig. 3. The average PIC (potentially informative character) and percentage variability values of each of the 34 regions (21 are from Shaw et al., 2005, and 13 are from this study). The 34 regions are oriented relative to one another across the genome from inverted repeat B to inverted repeat A. Thick black lines represent average PIC values, and thin gray lines illustrate the percentage variability found within each region. Thick gray bars stacked on top of black bars indicate the average PIC value of those regions that are often combined (i.e., *trnT-trnL-trnL*, *ndhJ-trnF-trnL*, *trnS-trnG-trnG*, both halves of the *trnK* intron, and *trnL-trnL-trnF*). LSC, large single-copy region; SSC, small single-copy region.

The number of nucleotide substitutions, indels, and inversions (hereafter referred to collectively as potentially informative characters or PICs) between the two ingroup species and between either ingroup species and the outgroup species were tallied for each noncoding cpDNA region in each of the seven lineages. Because indels have been shown to be prevalent and often phylogenetically informative (Golenberg et al., 1993; Morton and Clegg, 1993; Gielly and Taberlet, 1994), they were scored in this study, as were inversions. Indels, any nucleotide substitutions within the indels, and inversions were scored as independent, single characters.

Three types of calculations were performed. First, we estimated the proportion of observed mutational events for each noncoding cpDNA region using a modified version of the formula used in O'Donnell (1992) and Gielly and Taberlet (1994). The proportion of mutational events (or % variability) = $[(NS + ID + IV) / L] \times 100$, where NS = the number of nucleotide substitutions, ID = the number of indels, IV = the number of inversions, and L = the aligned sequence length. Second, we calculated the average number of PICs found within each noncoding chloroplast region. Third, to ensure that lineages containing a greater number of mutational events between the three-species groups (i.e., older or faster evolving lineages) were not overrepresented (weighted) in the average PIC value for each region, we normalized the PICs within each lineage. The number of PICs was normalized for each region/lineage combination by dividing the number of PICs found within that region/lineage combination by the sum total of PICs found within a given lineage. For example, in the *Magnolia* lineage the 21 PICs that were tallied for *rpl16* were divided by the 1021 PICs found within that lineage for all 34 noncoding regions. By doing this, we hope to have reduced the influence of differing evolutionary rates or distances among the different taxa. These normalized values were then used to generate average normalized PIC values for each of the noncoding cpDNA regions so that they could be directly compared.

RESULTS

Primer universality—All primers amplified and sequenced easily across all of the angiosperm lineages used in this study. Additionally, the new *trnS*$^{(GCU)}$-*trnG*$^{(UCC)}$-*trnG*$^{(UUC)}$ primers worked well in all of the taxa of this study as well as in many other *Prunus* (Rosaceae) taxa, *Oxalis* (Oxalidaceae) (E. Emshwiller, The Field Museum of Natural History, personal communication), *Crataegus* (Rosaceae) (A. Dönmez, Hacettepe University, personal communication), *Aethionema* (Brassicaceae) (M. Menke, Washington University, personal communication), and *Mimulus* (Scrophulariaceae) (J. Beck, Washington University, personal communication). Furthermore, all of the primers used in this study were also tried in the gymnosperm lineage that was part of Shaw et al. (2005) (i.e., *Cryptomeria*, *Glyptostrobus*, and *Taxodium*; Cupressaceae). While most primer pairs failed to amplify in these gymnosperm taxa, the primers for the *atpI-atpH*, *trnQ*$^{(UUG)}$-5′*rps16*, *petL-psbE*, and the *ndhA* intron regions amplified fragments that were consistent in size with fragments amplified in the angiosperms of this study.

Amount of chloroplast genome surveyed—Between the two studies (Shaw et al., 2005, and this study), we have now generated >670 kb of sequence data from three species in each of 10 seed plant lineages in an attempt to compare the systematic utilities of the noncoding portions of the chloroplast genome. According to the *Nicotiana tabacum* chloroplast model (Wakasugi et al., 1998), combining this and our previous study (Shaw et al., 2005) we have now surveyed 26 753 out of 45 988 (58.2%) base pairs of the noncoding portions of the LSC and SSC regions. Many of the remaining noncoding regions are too small to be of use to molecular studies because they are less than 350 bp; excluding these very short regions from the realm of possible markers, we have surveyed 67% of the noncoding regions of the chloroplast genome (76% of the intergenic spacer regions).

Relevant to the 13 explored regions of this study, we have sequenced >90 kb from seven angiosperm lineages. Of that, we observed 2917 nucleotide substitutions, 1038 indels, and three inversions for a total of 3958 PICs.

Combining the numbers of this study and our previous work (Shaw et al., 2005), for the seven angiosperms lineages relevant here, we have amassed a data set consisting of >565 kb from 34 noncoding cpDNA regions and observed 5422 nucleotide substitutions, 2117 indels, and nine inversions for a total of 7548 PICs. In all, nucleotide substitutions account for 71.8% of the variable characters, while indels and inversions account for 28.1% and 0.1%, respectively.

Assessment of the noncoding cpDNA regions—Because of potentially different rates of evolution among the different lineages, the different within-lineage phylogenetic distances observed among the lineages, and the exclusion of some regions as a result of structural rearrangement of the cpDNA molecule or PCR amplification or sequencing difficulties, we did not apply statistical analyses to these data. The following discussion is based on our qualitative interpretation of the results, which are compiled in Appendix S1 (see Supplemental Data accompanying online version of this article) and Figs. 3 and 4.

Figure 3 shows both the average PIC value and the average percentage variability found within each of the 34 noncoding cpDNA regions (data in Appendix S1, see Supplemental Data with online version of this article). Because of the disparity among taxa with respect to the number of PICs found within a given region (owing to the different divergence times among taxa of a three-species group or differing evolutionary rates), we normalized the PICs for each region/taxon combination. These data are illustrated in Fig. 4 and are also recorded in Appendix S1 (see Supplemental Data with online version of article).

In our previous study (Shaw et al., 2005), we divided cpDNA regions up into three tiers because there were natural breaks in the data; with the addition of the 13 regions of this study, no such natural breaks exist in the data, and a tier system is now less meaningful. Figure 4 ranks the regions based on the normalized PIC value found within each region and shows that nine regions, namely *rpl32-trnL*$^{(UAG)}$, *trnQ*$^{(UUG)}$-5′*rps16*, 3′*trnV*$^{(UAC)}$-*ndhC*, *ndhF-rpl32*, *psbD-trnT*$^{(GGU)}$, *psbJ-petA*, 3′*rps16*-5′*trnK*$^{(UUU)}$, *atpI-atpH*, and *petL-psbE*, are likely to offer more PICs than the best regions of Shaw et al. (2005).

Assessment of a correlation between PICs and length—To address the question of whether or not longer regions provide more PICs simply because they are longer, we generated regression lines and calculated coefficients of determination for each of the seven lineages (Fig. 5). Coefficients of determination range from 0.52 in *Minuartia* (caryophyllid) and *Gratiola* (euasterid I) to 0.71 in *Magnolia* (magnoliid) (Fig. 5). The average of these coefficients of determination is 0.58, suggesting that the length of the region does explain some of the PIC values but that length does not account for all of the variability observed, consistent with the observations of Shaw et al. (2005).

Comparison of introns and intergenic spacers—Of the noncoding regions of this study, six are introns (excluding the *matK* reading frame of the *trnK*$^{(UUU)}$ intron) and 27 are intergenic spacers. In all we surveyed 63% of the bases found in intergenic spacer regions of the LSC and SSC and 45% of the bases found in introns within the LSC and SSC regions. The average percentage variability of the six introns is 3.09% with a standard deviation of 0.57, whereas the average percentage variability of the 27 intergenic spacer regions is 4.12% with a standard deviation of 1.10, suggesting that intergenic spacer regions are more variable than introns and have a broader range of variance.

DISCUSSION

The important results of this study are twofold. First, this study corroborates our earlier findings that a disparity exists in the relative evolutionary rates of different noncoding chloroplast regions. Second, there are many regions available to researchers that offer levels of variation much higher than those regions commonly employed in plant molecular systematics (i.e., *trnL-trnL-trnF*, *trnK-matK-trnK*, *rps16* intron, and *rpl16* intron). Comparison of the normalized PIC values in Fig. 4 poignantly illustrates this fact. Our results are also exciting because they present evidence that the chloroplast genome may be better suited for low-level inquiry than previously thought, when interpretations were based mostly on *trnL-trnL-trnF* and *trnK-matK-trnK* data. The findings of our two studies (Shaw et al., 2005, and this study, summarized here in Figs. 3 and 4) provide an index of the relative levels of variability of appropriately sized cpDNA markers for phylogenetic, phylogeographic, population genetic, and DNA barcoding studies.

In a recent paper, Hughes et al. (2006) highlighted the fact that lack of resolution is a widespread problem among many published phylogenies. They show that most phylogenetic studies end in imperfectly resolved phylogenetic hypotheses that still often serve as the basis for extrapolations of evolutionary history, biogeography, hybridization, polyploidy, and character evolution. Because molecular phylogenetic studies often serve as foundations for testing other biological hypotheses, it is crucial that the systematics community

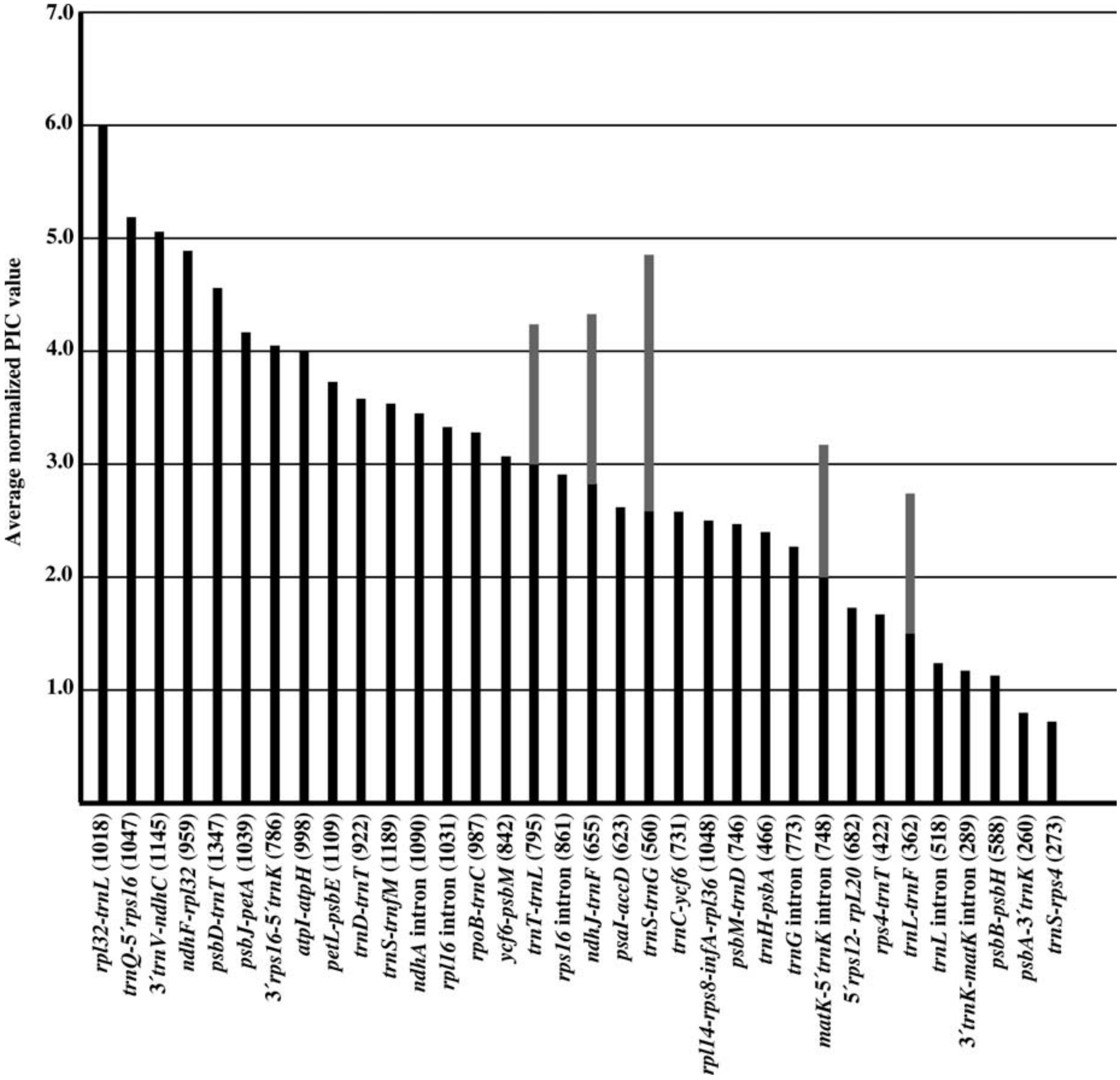
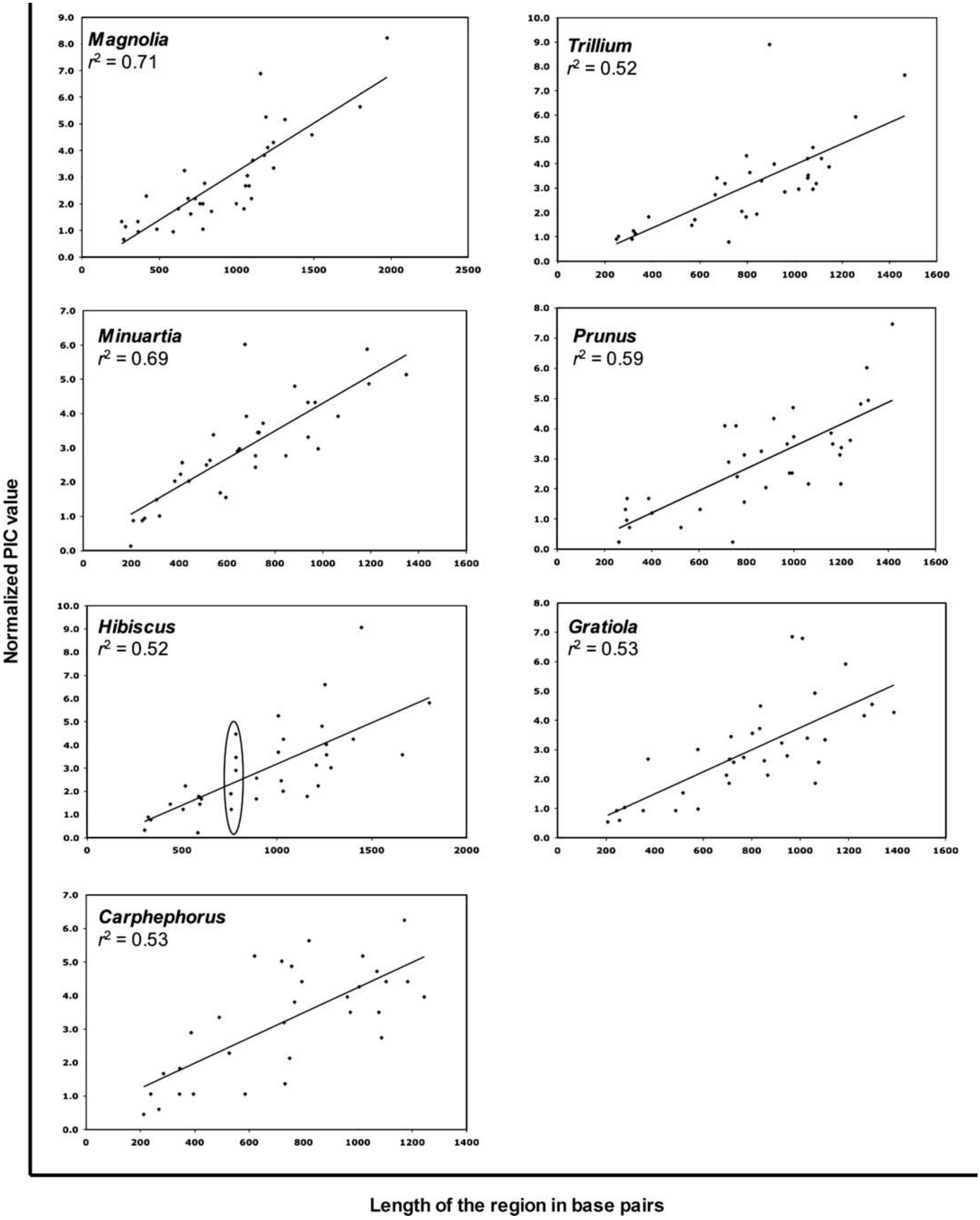


Fig. 4. The normalized PIC (potentially informative character) value of each of the 34 regions (21 are from Shaw et al., 2005, and 13 are from this study). The 34 regions are oriented from most to least number of PICs (left to right). Gray bars stacked on top of black bars indicate the normalized PIC value of often-combined regions. From left to right these are *trnT-trnL-trnL*, *ndhJ-trnF-trnL*, *trnS-trnG-trnG*, both halves of the *trnK* intron, and *trnL-trnL-trnF*. Numbers in parentheses indicate average length of the region.

incorporate more powerful, in other words more polymorphic, markers so that dependent hypotheses can be critically addressed. A potential solution is to utilize molecular markers that have an inherently higher level of variability and are capable of providing a higher level of phylogenetic resolution.

In our earlier study (Shaw et al., 2005), we did not conclude that a single noncoding cpDNA region is "the best" for low-level inquiry but rather that some regions are likely better choices than others and researchers should survey a few promising regions to determine the best one in a given study group. Shaw et al. (2005) showed that a three-species survey is a highly effective means of surveying promising regions. That being said, we have now screened the entire single-copy chloroplast genome and directly compared about 60% of the noncoding regions that are of an appropriate size for sequence-based query. Here we show that nine regions provide more PICs than the best regions of Shaw et al. (2005). This is an exciting conclusion because *rpl32-trnL*, *trnQ*-5′*rps16*, 3′*trnV-ndhC*, *ndhF-rpl32*, *psbD-trnT*, *psbJ-petA*, 3′*rps16*-5′*trnK*, *atpI-atpH*, and *petL-psbE* offer levels of variability previously unseen in the chloroplast genome, thus providing the molecular community with a list of the most informative noncoding cpDNA regions found within angiosperms. Having these better markers to survey prior to beginning an all out molecular sequencing study should ultimately lead to better resolved sequence-based studies and more accurate dependent hypotheses.

Screening the chloroplast genome for useful markers—To highlight noncoding cpDNA regions to study here, we compared related pairs of published whole chloroplast genome


Normalized PIC value
Length of the region in base pairs
Magnolia
r^2 = 0.71
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0
0 500 1000 1500 2000 2500
Trillium
r^2 = 0.52
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0
0 200 400 600 800 1000 1200 1400 1600
Minuartia
r^2 = 0.69
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0
0 200 400 600 800 1000 1200 1400 1600
Prunus
r^2 = 0.59
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0
0 200 400 600 800 1000 1200 1400 1600
Hibiscus
r^2 = 0.52
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0
0 500 1000 1500 2000
Gratiola
r^2 = 0.53
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0
0 200 400 600 800 1000 1200 1400 1600
Carphephorus
r^2 = 0.53
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0
0 200 400 600 800 1000 1200 1400

sequences. The success of this methodology is highlighted in the fact that nine of the 13 regions studied here contain a greater number of PICs than the best regions of Shaw et al. (2005) (Fig. 4). Using whole genome sequences from GenBank to survey for potentially highly informative regions has also proven to be an effective means of identifying informative regions as exemplified by other recent studies. Provan et al. (2004) recently used the entire chloroplast genome sequences of *Oryza*, *Triticum*, and *Zea* to screen the chloroplast genomes of grasses to identify SSRs. Similarly, Takahashi et al. (2005) compared *Saccharum* and *Zea* chloroplast genome sequences to search for regions of high variability for use in a phylogeny of *Saccharum*, Daniell et al. (2006) surveyed complete chloroplast genome sequences of four Solanaceae species, and Timme et al. (2007) compared complete chloroplast genome sequences of *Helianthus* and *Lactuca* (Asteraceae). Lastly, Kress et al. (2005) compared the complete chloroplast genomes of *Atropa* and *Nicotiana* to find an appropriate region for DNA barcoding in plants.

In the Kress et al. (2005), Daniell et al. (2006), and Timme et al. (2007) papers, the authors compiled a list of the most variable (on a percentage basis) noncoding regions of the chloroplast genome. In many cases, the regions identified as the most variable in these taxonomically limited analyses correspond to regions found to be highly variable in our study. Many of the regions listed by these authors, however, were not included in our study. This highlights an important difference between our study and previous studies that have used whole chloroplast genome comparisons to find highly variable regions. Specifically, there are two approaches to indexing variability. The mostly widely used is to calculate percentage variability as was done by Kress et al. (2005), Daniell et al. (2006), and Timme et al. (2007). Another approach is to tabulate the actual number of variable characters that may potentially be found in a given region (PICs) because a greater number of characters is what systematists are looking for in a marker, not necessarily a high percentage variability. Thus, in our screening of chloroplast genome sequences, we searched for regions that are both highly variable and of sufficient length to provide a reasonable number of PICs. We reasoned that with current sequencing technology, a region up to ca. 800 bp could be sequenced with a single sequencing primer, and regions up to ca. 1500 bp could be sequenced with two primers. Thus we excluded regions that were less than 500 bp from further consideration regardless of their percentage variability. Many of the regions listed as highly variable by Daniell et al. (2006) and Timme et al. (2007) are exceptionally short. For example, 15 of the 25 regions listed as most variable by Timme et al. (2007) are less than 500 bp (and many are less than 300 bp).

Discussion of the previously unexplored regions—We next summarize each of the 13 previously unexplored noncoding cpDNA regions that we have surveyed in this study including a brief history of their utility, if any, in previous studies and an assessment of their utility based on our results. Prior use of these regions was determined by searches on NCBI-GenBank, Web of Science, and Biological Abstracts. The regions are ordered beginning with those that yielded the highest normalized PIC values (Fig. 4).

It is worth noting here that several regions may be coamplified, sequenced, and successfully concatenated with the same two PCR primers, and from a cost perspective, they may be equal to amplifying and sequencing a portion of each alone. Some potentially concatenated regions include *psbA*-3′*trnK*-*matK*, *trnS*-*trnG*-*trnG*, *trnC*-*ycf6*-*psbM*, *ycf6*-*psbM*-*trnD*, *rps4*-*trnT*-*trnL*, and *trnL*-*trnL*-*trnF*, *ndhJ*-*trnF*-*trnL*, and possibly *ndhF*-*rpl32*-*trnL*. Some of the more common or potentially more informative coamplifiable regions are illustrated in Figs. 3 and 4, with the exception of *ndhF*-*rpl32*-*trnL*, because this concatenated region might be relatively long at 2 kb and the resulting bar would literally be off the chart.

rpl32-*trnL*$^{(UAG)}$—The *rpl32*-*trnL* intergenic spacer is in the SSC region of the chloroplast genome (Fig. 2V). To our knowledge, this region has not been used in any sequence-based investigations, although Timme et al. (2007) noted it as being highly variable. The average length of *rpl32*-*trnL* is 1018 bp, and it ranges from 543-1417 bp; this region was aberrantly small in *Minuartia* (caryophyllid) at 543 bp. Large indels were observed in *Hibiscus* (eurosid II) (70 bp), *Magnolia* (magnoliid) (50 bp), *Minuartia* (caryophyllid) (56 bp), and *Prunus* (eurosid I) (52 bp). Figures 3 and 4 suggest that this is the best region of the 34 regions surveyed for low-level molecular studies. In a few lineages *rpl32*-*trnL* was coamplified with the *ndhF*-*rpl32* intergenic spacer (next). Together they are about 2 kb and would then offer a PIC value additive to both bars in Fig. 4. The *rpl32* gene is less than 200 bp and is approximately in the center of this 2 kb region; so, if *ndhF*-*rpl32*-*trnL* is amplified as a single fragment, the *rpl32* primers can then serve as internal sequencing primers that will sequence through each other.

trnQ$^{(UUG)}$-5′*rps16*—The *trnQ*-5′*rps16* intergenic spacer is located in the LSC region (Fig. 2T) and was noted as highly variable by both Daniell et al. (2006) and Timme et al. (2007). Hahn (2002) used it in a study of Arecoid palms (Arecaceae) and showed that it yielded more characters than the *rbcL* or *atpB* genes. Although he showed *trnQ*-5′*rps16* to provide fewer characters than *trnD*-*trnT* (a Tier 1 region of Shaw et al., 2005), one Arecoid genus with highly divergent *trnQ*-5′*rps16* sequences was removed from the data set. The average length of *trnQ*-5′*rps16* is 1046 bp, and it ranges from 588-1975 bp. Large indels were observed in *Carphephorus* (euasterid II) (79 bp), *Gratiola* (euasterid I) (53 bp), *Prunus* (eurosid I) (178 bp), and *Trillium* (monocot) (112 bp, 54 bp).

3′*trnV*$^{(UAC)}$-*ndhC*—The 3′*trnV*-*ndhC* intergenic spacer lies within the LSC region (Fig. 2P). Aside from Takahashi et al. (2005) who used this region among *Saccharum* species (Poaceae) but made no comparison to other regions in their study, a search on GenBank revealed no other studies to have employed this region, although it was noted as highly variable by Timme et al. (2007). The average length of 3′*trnV*-*ndhC* is

←

Fig. 5. Scatterplots and regression lines of each of the seven lineages of this study showing the relationship of region length and its normalized PIC (potentially informative character) value. The normalized PIC value is represented on the *y*-axis, and the length of the region in base pairs is shown on the *x*-axis. Coefficients of determination (r^2 values) are shown in the upper left corner of each scatterplot. The circled values on the *Hibiscus* scatterplot highlight the fact that regions of a similar size, in this case ~800 bp, offer a significantly different number of PICs.

1146 bp, and it ranges from 318-1800 bp. This region appears to be especially prone to large indels. An indel of 735 bp was observed in *Gratiola* (euasterid I) and within that indel several other smaller indels (56 bp, 7 bp, 6 bp, 6 bp) were observed. Large indels were also observed in *Hibiscus* (eurosid II) (67 bp), *Prunus* (eurosid I) (129 bp, 95 bp), and *Trillium* (monocot) (310 bp). Last, 3′*trnV*-*ndhC* was far smaller (318 bp) in *Minuartia* (caryophyllid) than it was in the other taxa and still it contained an indel of 174 bp. Considering that these large indels create "missing data" in a three-species survey, this region may rank higher than our results indicate.

ndhF-*rpl32*—The *ndhF*-*rpl32* intergenic spacer is in the SSC region of the chloroplast genome (Fig. 2V) that is adjacent to *rpl32*-*trnL*$^{(UAG)}$. This region was noted as highly variable by Timme et al. (2007). Yamane and Kawahara (2005) utilized *ndhF*-*rpl32* in a study of *Triticum*-*Aegilops* (Poaceae), but because they pooled data from several regions, a comparison cannot be made. Its average length is 960 bp, and it ranges from 729-1254 bp. Large indels were observed in *Gratiola* (euasterid I) (260 bp), *Minuartia* (caryophyllid) (261 bp), and *Trillium* (monocot) (596 bp, 178 bp). According to Figs. 3 and 4, this region is among the best choices for low-level molecular investigation, especially if it is coamplified with the *rpl32*-*trnL*$^{(UAG)}$ intergenic spacer.

psbD-*trnT*$^{(GGU)}$—The *psbD*-*trnT* intergenic spacer is found in the LSC (Fig. 2R). To our knowledge, this is the first time that this intergenic spacer has been studied for sequence-based investigation, although it was noted as highly variable by Daniell et al. (2006). The average length of *psbD*-*trnT* is 1348 bp, and it ranges from 1057–1662 bp. A few short poly-A/T runs were observed in several of the lineages. Large indels were observed in *Minuartia* (caryophyllid) (94 bp, 77 bp, 68 bp, 57 bp).

psbJ-*petA*—The *psbJ*-*petA* intergenic spacer is located in the LSC region (Fig. 2N). This region was recently used in an intraspecific phylogeographic study of *Trochodendron aralioides* (Trochodendraceae) where several haplotypes were observed across Taiwan (Huang et al., 2004). In earlier studies, a cpSSR within *psbJ*-*petA* was polymorphic among closely related pines (Bucci et al., 1998) and was therefore used in a related study to identify 100-year-old herbarium specimens of *Pinus brutia* (Pinaceae) (DeCastro and Menale, 2004). The *psbJ*-*petA* intergenic spacer was also implicated as a potentially useful microsatellite region in *Castanea* (Fagaceae; Sebastiani et al., 2004). The average length of this intergenic spacer is 1040 bp, and it ranges from 734-1261 bp. Within the more quickly evolving lineages of this study (*Gratiola*, euasterid I and *Minuartia*, caryophyllid), there were regions that were difficult to align. Because we scored these regions conservatively by opening up gaps, this region may be more informative than shown here. Additionally, we observed several poly-A/T runs in all of the lineages confirming the likely presence of the cpSSR region in many other taxa besides *Pinus* and *Castanea*.

3′*rps16*–5′*trnK*$^{(UUU)}$—This intergenic spacer is located in the LSC (Fig. 2T). Aside from Takahashi et al. (2005) who used this region among *Saccharum* species (Poaceae) but made no comparison to other regions in their study, a search on GenBank revealed no other studies to have employed this region. This intergenic spacer was noted as highly variable by Daniell et al. (2006). The average length of *rps16*-5′*trnK* is 786 bp, and it ranges from 529-1008 bp. Large indels were found in this region in *Gratiola* (135 bp, 137 bp).

atpI-*atpH*—The *atpI*-*atpH* intergenic spacer is a region of the LSC that is approximately 518 bp (Fig. 2S). The average length of *atpI*-*atpH* is 998 bp, and it ranges from 514-1262 bp. Within this region, Provan et al. (2004) described cpSSR primers for grasses. We observed poly-A/T runs in all lineages studied here, but only one (24 bp) in *Carphephorus* (euasterid II) was long enough to cause problems during sequencing. One large indel (299 bp) was observed in *Magnolia* (magnoliid).

petL-*psbE*—The *petL*-*psbE* intergenic spacer is a region of the LSC (Fig. 2M) that averages 1109 bp (892-1315 bp). Popp et al. (2005) first employed this region during an investigation of closely related *Silene* (Caryophyllaceae), and they noted that *petL*-*psbE* offered information comparable to the *rps16* intron (P. Erixon, Uppsala University, and M. Popp, University of Oslo, personal communication). Several short poly-A/T runs were observed in each of the lineages, but none were long enough to negatively impact sequencing.

ndhA intron—The only gene of the SSC region with an intron is *ndhA* (Fig. 2U). Small et al. (1998) included the *ndhA* intron in their comparative study and showed that it yielded fewer potential characters than *rpl16*, *psaI*-*accD*, *trnT*-*trnL*, *trnL*-*trnF*, and *atpB*-*rbcL*. The average length of the *ndhA* intron is 1090 bp, and it ranges from 968-1160 bp. The *ndhA* intron ranked only slightly above the median in number of PICs compared to other regions that were surveyed. While it is the best-ranking intron of the six surveyed here, it is probably not statistically better than the *rpl16* intron.

ndhJ-*trnF*$^{(GAA)}$—The *ndhJ*-*trnF* region is located in the LSC region adjacent to the popular *trnL*-*trnF* intergenic spacer (Fig. 2Q). While a search of GenBank reveals that several researchers have recently used this region in molecular studies, few are published. Xu and Ban (2004) used this region in a study of closely related *Elymus* species (Poaceae) and showed that it provided twice the number of characters as the *trnL*-*trnF* intergenic spacer (6 vs. 3) and an equal number of characters as the *atpB*-*rbcL* intergenic spacer. Yamane and Kawahara (2005) combined the data from this region with that of several other regions in a study of *Triticum*-*Aegilops* (Poaceae) and did not directly compare the utility of the regions. The average length of *ndhJ*-*trnF* is 655 bp, and it ranges from 238–791 bp. Like *trnV*-*ndhC*, this region appears to be prone to large indels, though not to the same extent. Large indels were observed in *Gratiola* (euasterid II) (41 bp), *Minuartia* (caryophyllid) (50 bp), *Hibiscus* (eurosid II) (114 bp, 110 bp), and *Trillium* (monocot) (165 bp). In *Carphephorus* (euasterid II), this region is substantially shorter (238 bp.) than it is in other lineages. Because this intergenic spacer is adjacent to the *trnL*-*trnF* spacer, we coamplified the *trnL*-*trnF* intergenic spacer with this region (using the well-established Tab E primer of Taberlet et al. [1991]). Coamplification of *ndhJ*-*trnF*-*trnL* appears to be a much better choice than the common coamplification of *trnL*-*trnL*-*trnF* (Fig. 4).

psaI-*accD*—The *psaI*-*accD* intergenic spacer is a region of the LSC portion of the chloroplast genome (Fig. 2O). Takayama et al. (2005) recently showed *psaI*-*accD* to provide fewer characters than the *trnK*-*matK*-*trnK* or *atpB*-*rbcL* among closely related *Hibiscus* (Malvaceae) species. Loayza et al. (2005) suggested that this intergenic spacer is less informative than either *trnK-matK-trnK* or *trnL-trnF* in their study of *Phragmipedium* (Orchidaceae). In a study of *Cleistes* (Orchidaceae), *psaI-accD* was no more informative than *rps16* (Smith et al., 2004). In contrast to these reports, a preliminary survey of *Maxillaria* (Orchidaceae) showed that this region is too variable to be aligned confidently (M. Whitten, University of Florida, personal communication). Further, Small et al. (1998) showed that *psaI-accD* was more than twice as variable as *atpB-rbcL* and four times as variable as *trnL-trnF* in *Gossypium* (Malvaceae). That *psaI-accD* is more informative than *trnL-trnF* was also suggested by Kimura et al. (2003) in a study on *Pyrus* (Rosaceae), and this intergenic spacer successfully distinguished closely related orchid species of *Dendrochilum* (Orchidaceae) (Barkman and Simpson, 2002). Lastly, *psaI-accD* was identified as a potentially useful microsatellite region in *Castanea* (Fagaceae) (Sebastiani et al., 2004).

The average length of *psaI-accD* is 622 bp, and it ranges from 320–784 bp; *psaI-accD* was much shorter (320 bp) in *Trillium* (monocot) compared to the rest of the lineages. We also observed several small poly-A/T runs in several lineages.

rpl14-rps8-infA-rpl36—The *rpl14-rps8-infA-rpl36* region lies within the LSC (Fig. 2L). To our knowledge, this is the first study to compare this region to others of the chloroplast genome. Although mostly coding, this region was chosen because the intergenic spacers on either side of *rps8* had very high levels of variation in our initial screening of the complete single-copy regions of the genome. The average length of *rpl14-rps8-infA-rpl36* is 1049 bp, and it ranges from 981 to 1077 bp. A cpSSR of approximately 20–25 bp exists in all of the lineages in the *rpl14-rps8* intergenic spacer. It may have been this cpSSR that effectively highlighted this region during our initial screening process.

Indels vs. nucleotide substitutions—Overall, nucleotide substitutions account for 71.8% of the PIC value, while indels account for 28.1% and inversions only 0.1% of the surveyed chloroplast genome. In an attempt to be conservative in places we were calling PICs, we opened gaps in areas that were difficult to align; this may have slightly inflated the relative percentage of the total PIC value due to indels. Even still, this study represents the largest data set to compare the relative amounts of indels and nucleotide substitutions. Previous authors have addressed the issue of the relative frequencies of nucleotide substitutions and indels in noncoding cpDNA sequences, and conflicting hypotheses have been put forward. Clegg et al. (1994) wrote that indels may occur more frequently than nucleotide substitutions, and Golenberg et al. (1993) and Gielly and Taberlet (1994) both suggested that indels occur with nearly the same frequency as nucleotide substitutions. Our results agree with Small et al. (1998) and are much in line with our earlier work; all three of these studies suggest that nucleotide substitutions account for about 70% and indels account for about 30% of all mutations in the chloroplast genome (recognizable inversions accounting for a negligible <1.0%).

Mono- and polynucleotide repeats—Several of the regions were rich in strings of mononucleotide repeats and/or small tandem repeat units that are likely the result of slipped-strand mispairing (Levinson and Gutman, 1987). Mononucleotide (A/T) repeats and/or small tandem repeats (AT) were especially noted in *trnQ-5′rps16*, *psbD-trnT*, *psbJ-petA*, *atpI-atpH*, *psaI-accD*, *rpl14-rps8*, and from our last study the *trnH-psbA*, *psbA-3′trnK*, *matK-5′trnK*, *trnS-trnfM*, *trnS-trnG*, *trnD-trnT*, *trnT-trnL* intergenic spacers and in the *rps16* and *trnG* introns. Several of these cpDNA regions have mono- or polynucleotide portions that have been used in cpSSR studies, and other regions containing such repeats may also prove useful.

Introns vs. intergenic spacers—Herein we have compared six introns (combining both halves of *trnK*) and 27 intergenic spacers. The average percentage variability of the six introns was less than that of the intergenic spacers (3.09% and 4.12%, respectively), and the standard deviation of the introns was also less than that of the intergenic spacers (0.57 and 1.10, respectively). While the two sample sizes are not equivalent, these numbers do suggest that intergenic spacer regions are more variable than introns and have a broader range of variance. Figure 4 highlights this observation wherein the introns are scattered amidst the center and not to either extreme in contrast to the intergenic spacers that are spread out through the figure. Interestingly, the least variable intron of the six surveyed in this study is the *trnL* intron, the only group I intron of the chloroplast genome (all the rest are group II introns). Timme et al. (2007) also observed greater levels of variability in intergenic spacers than introns in a comparison of *Helianthus* and *Lactuca* where the average *p*-distance for spacers was 0.057, while for introns it was 0.032.

DNA barcoding—During the last few years, there has been much discussion on the topic of DNA barcoding (Hebert et al., 2003; Will and Rubinoff, 2004; Kress et al., 2005; Rubinoff et al., 2006). Kress et al. (2005) suggested that the *trnH-psbA* intergenic spacer is a useful marker for such an effort —and we agree that this may be about as good a marker as there is in the chloroplast genome for three reasons: (1) the *trnH-psbA* intergenic spacer is among the most variable (in terms of percentage variability, Fig. 3), (2) it is a relatively short region across angiosperms allowing for successful PCR amplification from degraded herbarium specimens, and (3) published primers appear to be especially “universal” so that one primer pair is likely to amplify nearly all angiosperm taxa. On the other hand, as mentioned earlier and in Shaw et al. (2005), this region is particularly short and therefore may not yield enough PICs to distinguish among closely related species (discounting the fact that closely related plants will be difficult to “barcode” because it has been shown that recent histories of hybridization can homogenize or even uncouple plastid genome phylogenies from species phylogenies [Shaw and Small, 2005]). Both Kress et al. (2005) and Rubinoff et al. (2006) suggest that in all likelihood we will need to employ more than one marker in a barcoding effort, and we feel that our work provides the necessary background to make an informed decision as to which cpDNA markers might be candidates for such an effort.

Conclusions—At the outset of this line of study, we were in search of “the hare,” or the chloroplast region that would provide the greatest number of characters for low-level molecular phylogenetic studies. Instead, we illuminated the noncoding chloroplast regions that are likely to provide the greatest number of characters for low-level molecular phylogenetic studies, while at the same time we highlighted the regions that may provide the least. As recommended in Shaw et al. (2005) and because there is no single region that is the best across all taxonomic lineages, we recommend that the top few choices be screened before committing to an all out sequencing effort in order to determine which of these regions is (are) the most suitable in a given lineage. In other words, instead of finding "the hare," the significant results of this line of study were to cull the many "tortoises"—the hares should still be screened to determine which might be the fastest in a given "race."

LITERATURE CITED

APG II. 2003. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141: 399–436.

BARKMAN, T. J., AND B. B. SIMPSON. 2002. Hybrid origin and parentage of *Dendrochilum acuiferum* (Orchidaceae) inferred in a phylogenetic context using nuclear and plastid DNA sequence data. *Systematic Botany* 27: 209–220.

BUCCI, G., M. ANZIDEI, A. MADAGHIELE, AND G. G. VENDRAMIN. 1998. Detection of haplotypic and natural hybridization in *halepensis*-complex pine species using chloroplast simple sequence repeat (SSR) markers. *Molecular Ecology* 7: 1633–1643.

CARTWRIGHT, R. A. 2005. DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics* 21: iii31–iii38.

CLEGG, M. T., B. S. GAUT, G. H. LEARN JR., AND B. R. MORTON. 1994. Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences, USA* 91: 6795–6801.

CURTIS, S. E., AND M. T. CLEGG. 1984. Molecular evolution of chloroplast DNA sequences. *Molecular Biology and Evolution* 1: 291–301.

DANIELL, H., S.-B. LEE, J. GREVICH, C. SASKI, T. QUESADA-VARGAS, C. GUDA, J. TOMKINS, AND R. K. JANSEN. 2006. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theoretical and Applied Genetics* 112: 1503–1518.

DECASTRO, O., AND B. MENALE. 2004. PCR amplification of Michele Tenore's historical specimens and facility to utilize an alternative approach to resolve taxonomic problems. *Taxon* 53: 147–151.

DOWNIE, S. R., AND J. D. PALMER. 1992. Use of chloroplast DNA rearrangements in reconstruction plant phylogeny. *In* Soltis et al. [eds.], Molecular systematics of plants, 1–13. Chapman and Hall, New York, New York, USA.

DOYLE, J. J., AND J. L. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

GAUT, B. S. 1998. Molecular clocks and nucleotide substitution rates in higher plants. *In* M. K. Hech, R. J. MacIntyre, and M. T. Clegg [eds.], Evolutionary biology, vol. 30, 93–120. Plenum Press, New York, New York, USA.

GIELLY, L., AND P. TABERLET. 1994. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus *rbcL* sequences. *Molecular Biology and Evolution* 11: 769–777.

GOLENBERG, E. M., M. T. CLEGG, M. L. DURBIN, J. DOEBLEY, AND D. P. MA. 1993. Evolution of a noncoding region of the chloroplast genome. *Molecular Phylogenetics and Evolution* 2: 52–64.

HAHN, W. J. 2002. A phylogenetic analysis of the arecoid line of palms based on plastid DNA sequence data. *Molecular Phylogenetics and Evolution* 23: 189–204.

HEBERT, P. D. N., A. CYWINSKA, S. L. BALL, AND J. R. DEWAARD. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, B, Biological Sciences* 270: 313–321.

HORNING, M. E., AND R. C. CRONN. 2006. Length polymorphism scanning is an efficient approach for revealing chloroplast DNA variation. *Genome* 49: 134–142.

HUANG, S.-F., S.-Y. HWANG, J.-C. WANG, AND T.-P. LIN. 2004. Phylogeography of *Trochodendron aralioides* (Trochodendraceae) in Taiwan and its adjacent areas. *Journal of Biogeography* 31: 1251–1259.

HUGHES, C. E., R. J. EASTWOOD, AND C. D. BAILEY. 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 361: 211–225.

ICKERT-BOND, S. M., AND J. WEN. 2006. Phylogeny and biogeography of Altingiaceae: evidence from combined analysis of five non-coding chloroplast regions. *Molecular Phylogenetics and Evolution* 39: 512–528.

KIMURA, T., H. IKETANI, K. KOTOBUKI, N. MATSUTA, Y. BAN, T. HAYASHI, AND T. YAMAMOTO. 2003. Genetic characterization of pear varieties revealed by chloroplast DNA sequences. *Journal of Horticultural Science and Biotechnology* 78: 241–247.

KRESS, W. J., K. J. WURDACK, E. A. ZIMMER, L. A. WEIGT, AND D. H. JANZEN. 2005. Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences, USA* 102: 8369–8374.

LEVINSON, G., AND G. A. GUTMAN. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* 4: 203–221.

LOAYZA, M. D., N. H. WILLIAMS, AND M. WHITTEN. 2005. *Phragmipedium kovachii*: molecular systematics of a new world orchid. *Orchids* 72: 132–137.

MORTON, B. R., AND M. T. CLEGG. 1993. A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near *rbcL* in the grass family (Poaceae). *Current Genetics* 24: 357–365.

O'DONNELL, K. 1992. Ribosomal DNA internal transcribed spacers are highly divergent in the phytopathogenic ascomycete *Fusarium sambucinum* (*Gibberella pulicaris*). *Current Genetics* 22: 213–220.

OLMSTEAD, R. G., AND J. D. PALMER. 1994. Chloroplast DNA systematics: a review of methods and data analysis. *American Journal of Botany* 81: 1205–1224.

PERRY, A. S., AND K. H. WOLFE. 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *Journal of Molecular Evolution* 55: 501–508.

POPP, M., P. ERIXON, F. EGGENS, AND B. OXELMAN. 2005. Origin and evolution of a circumpolar polyploidy species complex in *Silene* (Caryophyllaceae) inferred from low-copy nuclear RNA polymerase introns, rDNA, and chloroplast DNA. *Systematic Botany* 30: 302–313.

PROVAN, J., P. M. BISS, D. MCMEEL, AND S. MATTHEWS. 2004. Universal primers for the amplification of chloroplast microsatellites in grasses (Poaceae). *Molecular Ecology Notes* 4: 262–264.

RONNING, S. B., K. RUDI, K. G. BERDAL, AND A. HOLST-JENSEN. 2005. Differentiation of important and closely related cereal plant species (Poaceae) in food by hybridization to an oligonucleotide array. *Journal of Agricultural and Food Chemistry* 53: 8874–8880.

RUBINOFF, D., S. CAMERON, AND K. WILL. 2006. Are plant DNA barcodes a search for the Holy Grail? *Trends in Ecology and Evolution* 21: 1–2.

SCHÖNSWETTER, P., M. POPP, AND C. BROCHMANN. 2006a. Central Asian origin of and strong genetic differentiation among the populations of the rare and disjunct *Carex atrofusca* (Cyperaceae) in the Alps. *Journal of Biogeography* 33: 948–956.

SCHÖNSWETTER, P., M. POPP, AND C. BROCHMANN. 2006b. Rare arctic-alpine plants of the European Alps have different immigration histories: the snow bed species *Minuartia biflora* and *Ranunculus pygmaeus*. *Molecular Ecology* 15: 709–720.

SEBASTIANI, F., S. CARNEVALE, AND G. G. VENDRAMIN. 2004. A new set of mono- and dinucleotide chloroplast microsatellites in Fagaceae. *Molecular Ecology Notes* 4: 259–261.

SHAW, J., E. LICKEY, J. T. BECK, S. B. FARMER, W. LIU, J. MILLER, K. C. SIRIPUN, C. T. WINDER, E. E. SCHILLING, AND R. L. SMALL. 2005. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142–166.

SHAW, J., AND R. L. SMALL. 2005. Chloroplast DNA phylogeny and phylogeography of the North American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). *American Journal of Botany* 92: 2011–2030.

SMALL, R. L., J. A. RYBURN, R. C. CRONN, T. SEELANAN, AND J. F. WENDEL. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogenetic reconstruction in a recently diverged plant group. *American Journal of Botany* 85: 1301–1315.

Smith, S. D., R. S. Cowan, K. B. Gregg, M. W. Chase, N. Maxted, and M. F. Fay. 2004. Genetic discontinuities among populations of *Cleistes* (Orchidaceae, Vanilloideae) in North America. *Botanical Journal of the Linnaean Society* 145: 87–95.

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

Takahashi, S., T. Furukawa, T. Asano, Y. Terajima, H. Shimada, A. Sugimoto, and K. Kadowaki. 2005. Very close relationship of the chloroplast genomes among *Saccharum* species. *Theoretical and Applied Genetics* 110: 1523–1529.

Takayama, K., T. Ohi-Toma, H. Kudoh, and H. Kato. 2005. Origin and diversification of *Hibiscus glaber*, species endemic to the oceanic Bonin Islands, revealed by chloroplast DNA polymorphism. *Molecular Ecology* 14: 1059–1071.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 2001. ClustalX. Computer program available at ftp://ftp-igbmc.u-strasbg.fr/pub/clustalx/.

Timme, R., E. J. Kuehl, J. L. Boore, and R. K. Jansen. 2007. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *American Journal of Botany* 94: 302–313.

Wakasugi, T., M. Sugita, T. Tsudzuki, and M. Sugiura. 1998. Updated gene map of tobacco chloroplast DNA. *Plant Molecular Biology Reporter* 16: 231–241.

Will, K. W., and D. Rubinoff. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20: 47–55.

Wills, D. M., and J. M. Burk. 2006. Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.) *Journal of Heredity* 97: 403–408.

Wolfe, K. H. 1991. Protein-coding genes in chloroplast DNA: compilation of nucleotide sequences, data base entries, and rates of molecular evolution. *In* L. Bogorad and I. K. Vasil [eds.], Cell culture and somatic cell genetics of plants, vol. 7B, 467–482. Academic Press, New York, New York, USA.

Wolfe, K. H., W. H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054–9058.

Xu, D. H., and T. Ban. 2004. Phylogenetic and evolutionary relationships between *Elymus humidus* and other *Elymus* species based on sequencing of non-coding regions of cpDNA and AFLP of nuclear DNA. *Theoretical and Applied Genetics* 108: 1443–1448.

Yamane, K., and T. Kawahara. 2005. Intra- and interspecific phylogenetic relationships among diploid *Triticum-Aegilops* species (Poaceae) based on base-pair substitutions, indels, and microsatellites in chloroplast noncoding sequences. *American Journal of Botany* 92: 1887–1898.

Appendix 1. Voucher information and GenBank accession numbers for taxa used in this study. Voucher specimens are deposited in the following herbaria: TENN = The University of Tennessee, Knoxville, Tennessee, USA; CANB = The Centre for Plant Biodiversity Research, Canberra, Australia; EKY = Eastern Kentucky University, Richmond, Kentucky, USA; USCH = The University of South Carolina, Columbia, South Carolina, USA. O.G. = the taxa serving as outgroup taxa in the three-species surveys.

Angiosperm lineage

Taxon—GenBank accessions: *rpl14-rps8-infA-rpl36*, *petL-psbE*, *psbJ-petA*, *psaI-accD*, *3′trnV-ndhC*, *ndhJ-trnF*, *psbD-trnT*, *atpI-atpH*, *trnQ-5′rps16*, *3′rps16–5′trnK*, *ndhF-rpl32-trnL*, *ndhA* intron; *Voucher specimen*; Source.

Magnoliids

Magnolia acuminata L.—DQ826347, DQ826221, DQ826305, DQ826241, DQ826263, DQ826325, DQ813516, DQ826178, DQ826200, DQ826377, DQ826284, DQ826158; *J.T. Beck 6000*, USA, TN; TENN. ***Magnolia tripetala*** L.—DQ826346, DQ826220, DQ826304, DQ826240, DQ826262, DQ826324, DQ813515, DQ826179, DQ826199, DQ826376, DQ826283, DQ826157; *J.T. Beck 6001*, USA, TN; TENN. O.G. = ***Liriodendron tulipifera*** L.—DQ826345, DQ826219, DQ826303, DQ826242, DQ826261, DQ826326, DQ813517, DQ826177, DQ826198, DQ826375, DQ826282, DQ826156; *J.T. Beck 6002*, USA, TN; TENN.

Monocots

Trillium ovatum Pursh—DQ826350, DQ826223, DQ826306, DQ826243, DQ826264, DQ826329, DQ813518, DQ826180, DQ826203, DQ826368, DQ826286, DQ826161*; S. Farmer s.n.*, USA, OR; TENN. ***Trillium texanum*** Buckl.—DQ826349, DQ826222, DQ826307, DQ826245, DQ826266, DQ826328, DQ813520, DQ826182, DQ826202, DQ826367, DQ826287, DQ826159; *S. Farmer and Singhurst s.n.*, TX, USA; TENN. O.G. = ***Pseudotrillium rivale*** (S. Wats.) S.B. Farmer—DQ826348, DQ826224, DQ826308, DQ826244, DQ826265, DQ826327, DQ813519, DQ826181, DQ826201, DQ826366, DQ826285, DQ826160; *Graham s.n.*, cult. from USA, OR; TENN.

Caryophyllids

Minuartia cumberlandensis (B.E. Wofford & Kral) McNeill—DQ826352, DQ826226, DQ826310, DQ826246, DQ826268, DQ826332, DQ813521, DQ826183, DQ826206, DQ826370, DQ826290, DQ826163; *C.T. Winder s.n.*, USA, TN; TENN. ***Minuartia glabra*** (Michx.) Mattf.—DQ826353, DQ826227, DQ826309, DQ826247, DQ826269, DQ826331, DQ813522, DQ826184, DQ826205, DQ826371, DQ826289, DQ826164; *C.T. Winder s.n.*, USA, TN; TENN. O.G. = ***Minuartia uniflora*** (Walt.) Mattf.—DQ826351, DQ826225, DQ826311, DQ826248, DQ826267, DQ826330, DQ813523, DQ826185, DQ826204, DQ826369, DQ826288, DQ826162; *C.T. Winder s.n.*, USA, GA; TENN.

Eurosids I

Prunus hortulata Bailey—DQ826355, DQ826229, DQ826312, DQ826251, DQ826271, DQ826333, DQ813524, DQ826187, DQ826208, DQ826373, DQ826293, DQ826166; *J. Shaw JSh821–017*, USA, TN; TENN. ***Prunus nigra*** Ait.—DQ826356, DQ826230, DQ826313, DQ826250, DQ826272, DQ826334, DQ813525, DQ826188, DQ826209, DQ826374, DQ826292, DQ826167; *J. Shaw JSh979–125*, USA, VT; TENN. O.G. = ***Prunus virginiana*** L.—DQ826354, DQ826228, DQ826314, DQ826249, DQ826270, DQ826335, DQ813526, DQ826186, DQ826207, DQ826372, DQ826291, DQ826165; *J. Shaw JSh871–040*, USA, NH; TENN.

Eurosids II

Hibiscus cannabinus L.—DQ826358, DQ826232, DQ826316, DQ826252, DQ826274, DQ826337, DQ813529, DQ826190, DQ826211, DQ826378, DQ826295, DQ826169*; R.L. Small s.n.*, USA, FL (cultivar); TENN. ***Hibiscus mechowii*** Garcke—DQ826359, DQ826233, DQ826315, DQ826253, DQ826275, DQ826338, DQ813528, DQ826191, DQ826212, DQ826379, DQ826296, DQ826170; *R.L. Small s.n.*, Zambia; TENN. O.G. = ***Hibiscus macrophyllus*** Roxb.—DQ826357, DQ826231, DQ826317, DQ826254, DQ826273, DQ826336, DQ813527, DQ826189, DQ826210, DQ826380, DQ826294, DQ826168; *L. Craven 10202*; Indonesia; CANB.

Euasterids I

Gratiola brevifolia Raf.—DQ826362, DQ826235, DQ826318, DQ826256, DQ826278, DQ826341, DQ813531, DQ826193, DQ826215, DQ826386, DQ826298, DQ826173; *D. Estes 02513*, USA, TN; EKY. ***Gratiola virginiana*** L.—DQ826361, DQ826236, DQ826319, DQ826257, DQ826277, DQ826339, DQ813532, DQ826192, DQ826214, DQ826385, DQ826299, DQ826171; *D. Estes 04608*, USA, TN; TENN. O.G. = ***Gratiola neglecta*** Torr.—DQ826360, DQ826234, DQ826320, DQ826255, DQ826276, DQ826340, DQ813530, DQ826194, DQ826213, DQ826384, DQ826297, DQ826172; *D. Estes 04609*, USA, TN; TENN.

Euasterids II

Carphephorus corymbosus (Nutt.) Torr. & A. Gray—DQ826364, DQ826238, DQ826322, DQ826258, DQ826279, DQ826342, DQ813534, DQ826197, DQ826217, DQ826381, DQ826300, DQ826176; *E.E. Schilling 2036*, USA, GA; TENN. ***Trilisa paniculata*** (Willd.) Cass.—DQ826365, DQ826239, DQ826321, DQ826259, DQ826280, DQ826343, DQ813535, DQ826196, DQ826218, DQ826382, DQ826301, DQ826174; *J.B. Nelson 21688*, USA, SC; USCH. O.G. = ***Eupatorium capillifolium*** (Lamarck) Small—DQ826363, DQ826237, DQ826323, DQ826260, DQ826281, DQ826344, DQ813533, DQ826195, DQ826216, DQ826383, DQ826302, DQ826175; *K.C. Siripun 02-Eup-155*, USA, NC; TENN.